Universidade Estadual de Londrina

LARYSSA RIBEIRO CALCAGNOTO

# UMA NOVA PROPOSTA METODOLÓGICA DE ANÁLISE PARA DADOS MULTIVARIADOS SOBRE ABSENTEÍSMO

Londrina
2021

LARYSSA RIBEIRO CALCAGNOTO

# UMA NOVA PROPOSTA METODOLÓGICA DE ANÁLISE PARA DADOS MULTIVARIADOS SOBRE ABSENTEÍSMO

Dissertação de mestrado apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

Orientador: Prof. Dr. Tiago Viana Flor Santana

Londrina
2021

**Catalogação elaborada pela Divisão de Processos Técnicos da Biblioteca Central da Universidade Estadual de Londrina**

**Dados Internacionais de Catalogação -na-Publicação (CIP)**

C144n Calcagnoto, Laryssa Ribeiro.
Uma Nova Proposta Metodológica De Análise Para Dados Multivariados Sobre Absenteísmo /
Laryssa Ribeiro Calcagnoto. – Londrina, 2021.
73 f. : il.

Orientador: Tiago Viana Flor Santana.
Dissertação (Mestrado em Matemática Aplicada e Computacional) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Matemática Aplicada e Computacional, 2021.

Inclui Bibliografia.

1. Absenteísmo - Tese. 2. Análise de clusters - Tese. 3. Biplot - Tese. 4. Componentes principais - Tese. I. Santana, Tiago Viana Flor. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Matemática Aplicada e Computacional. III. Título.

CDU 31

LARYSSA RIBEIRO CALCAGNOTO

# UMA NOVA PROPOSTA METODOLÓGICA DE ANÁLISE PARA DADOS MULTIVARIADOS SOBRE ABSENTEÍSMO

Dissertação de mestrado apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

**BANCA EXAMINADORA**

Orientador: Prof. Dr. Tiago Viana Flor de Santana
Universidade Estadual de Londrina - UEL

Prof. Dr. Rodrigo Rossetto Pescim
Universidade Estadual de Londrina - UEL

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Santa Catarina - UFSC

Londrina, 15 de dezembro de 2021.

CALCAGNOTO, Laryssa Ribeiro. **Uma nova proposta metodológica de análise para dados multivariados sobre absenteísmo.** 2021. 73 f. Dissertação (Mestrado em Matemática Aplicada e Computacional) – Universidade Estadual de Londrina, Londrina, 2021.

## RESUMO

O absenteísmo é a prática ou costume de um colaborador se ausentar de seu local de trabalho. Suas causas são diversas e afetam a renda do trabalhador, provoca transtornos operacionais, estressa a administração e causa prejuízos financeiros para empresa. A análise de clusters é uma ferramenta multivariada que pode ser utilizada para determinar grupos de modo que cada grupo apresente características próprias de acordo com as variáveis observadas. Assim, podese utilizar essa técnica como suporte para determinar as características que contribuem para o absenteísmo. O método para construção dos clusters utilizado foi o algorítimo hierárquico de Ward e para comparação dos grupos o teste não paramétrico de Kruskal-Wallis foi adotado. Por fim, um estudo sobre a força de associação entre as variáveis foi desenvolvido utilizando-se a correlação de Spearman e para a relação entre variáveis relacionadas à ausência e os aspectos sociais, utilizou-se a análise de componentes principais, assim como a construção de um biplot para resumir os resultados da correlação e componentes principais. Por meio desse estudo foi possível determinar três grupos heterogêneos na empresa e evidenciar características que são potenciais fatores causadores do absenteísmo em maior ou menor grau.

**Palavras-chave:** absenteísmo; análise de clusters; biplot; componentes principais; correlação de spearman; teste de kruskal-wallis.

CALCAGNOTO, Laryssa Ribeiro. **A new methodological proposal for analyzing multivariate data on absenteeism.** 2021. 73 p. Dissertação (Mestrado em Matemática Aplicada e Computacional) – Universidade Estadual de Londrina, Londrina, 2021.

## ABSTRACT

Absenteeism is the practice or custom of an employee to be absent from their workplace. Its causes are diverse and affect the worker's income, causing operational disturbances, stressing the administration and causing financial losses for the company. clusters analysis is a multivariate tool that can be used to determine groups so that each group presents its own characteristics according to the observed variables. Thus, this technique can be used as support to determine the characteristics that contribute to absenteeism. The method for constructing the clusters used was Ward's hierarchical algorithm and the nonparametric Kruskal-Wallis test was adopted for comparing the groups. Finally, a study on the strength of association between the variables was developed using Spearman's correlation and for the relationship among variables related to absence and social aspects, principal component analysis was used, as well as the construction of a biplot to summarize the correlation results and principal components. Through this study, it was possible to determine three heterogeneous groups in the company and to highlight characteristics that are potential factors causing absenteeism to a greater or lesser extent.

**Key words:** absenteeism; clusters analysis; biplot; principal componentes; spearman's correlation; kruskal-wallis test.

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

O termo absenteísmo foi introduzido durante a Revolução Industrial para se referir aos trabalhadores faltosos ao longo de seu turno de serviço. Por se tratar de um problema nascido na modernidade, começaram a surgir diversos questionamentos, como, por exemplo: quais variáveis estão associadas, como representar e como controlar o absenteísmo?

Para responder essa questão, alguns pesquisadores sugerem que o absenteísmo represente todo e qualquer tipo de ausência na empresa e seja apresentado por meio de um índice, dessa maneira podendo classificar os funcionários e o empreendimento. A Associação Brasileira de Controle de Qualidade (ABCQ) [1] indica que um índice aceitável para as empresas seria em torno de $1,5\%$.

Contudo, no ano de 2019, a revista Exame [21] realizou um estudo para verificar a taxa de absenteísmo nas empresas brasileiras em dois setores e ambos apresentaram uma taxa maior do que a aceitável pela ABCQ. A primeira área analisada foi o setor de serviços, que obteve um índice de $5\%$, e a segunda, o varejista, cujo índice variava entre $7\%$ e $10\%$.

Assim como o setor varejista e de serviços, no Paraná as empresas dos setores de agroindústria, alimentos e bebidas apresentam uma taxa de absenteísmo elevada, cuja média é de $5,1\%$, conforme exposto pelo $3^o$ Benchmarking Paranaense [22]. Com a alta desse índice, alguns autores começaram a estudar o tema, como apresentaremos adiante.

Na literatura pode-se observar muitos trabalhos que estudam o absenteísmo em áreas como recursos humanos, administração, psicologia e enfermagem. Entretanto, esses trabalhos dividem-se em dois grupos: o primeiro busca verificar a eficácia de métodos implantados nas empresas e o segundo, identificar as causas e consequências das ausências dos funcionários.

Calais [4] realizou uma pesquisa na área de psicologia em uma empresa de transporte urbano com uma amostra de $38$ motoristas, em que aplicaram-se questionários e inventários para um estudo de delineamento "quase experimental". Buscando-se verificar se os métodos adotados pelas empresas eram eficazes na redução do absenteísmo, a autora separou a amostra em um grupo experimental (GE) e um grupo controle (GC) com $19$ colaboradores, ambos com as mesmas características. Dessa forma, foram realizados testes intragrupal e intergrupal utilizando-se os testes não paramétricos de qui-quadrado para as variáveis relacionadas ao estresse e Wilcoxom as demais variáveis. Já para a análise intergrupal aplicou-se o teste de Mann- Whitney para duas amostras independentes. Concluindo-se que o programa de intervenção adotado na empresa reduz os níveis de estresse dos funcionários, apesar de não ter sido apresentada diferença significativa no estudo de resiliência.

Penatti [19] estudou o absenteísmo em uma indústria multinacional automobilística, na área de administração, coletando informações sobre o tema junto à empresa e entrevistando médicos e fisioterapeutas para desenvolvimento de um questionário com $21$ perguntas,

que depois aplicou a $10\%$ da população entre os anos de $2000$ e $2005$. O autor relata uma queda do absenteísmo na empresa com a adoção de algumas práticas para controle e um certo padrão no comportamento das abstenções, sendo quase sempre maior no setor produtivo comparado ao setor administrativo. Além disso, observou-se ainda uma correlação entre o clima organizacional (compreensão coletiva que os colaboradores têm da companhia) e o não comparecimento por motivo de doença, sendo este inversamente proporcional ao contentamento com o trabalho; assim, quanto maior a satisfação dos funcionários, menor a taxa de ausências.

Buscando identificar as causas e motivos do absenteísmo, Almeida [2] aplicou um questionário de $12$ questões abertas e fechadas a $22$ pessoas de uma empresa. A partir da coleta das informações e realização de uma análise descritiva dos dados, a autora concluiu que a principal causa das abstenções são ausências de $2$ a $3$ horas por motivos de consultas, problemas domésticos e idas ao banco.

Em sua dissertação de mestrado, Silva [24] realizou um estudo para estimar a prevalência do absenteísmo-doença na área de enfermagem, tomando como referência o CID-10 (Classificação Internacional de Doenças). O autor utiliza uma amostra de $294$ trabalhadores da saúde, sendo: $58$ enfermeiros, $212$ técnicos e $24$ auxiliares de enfermagem. A partir disso, realizou-se uma seleção de variáveis na regressão logística múltipla pelo método Stepwise Forward e uma análise do índice de absenteísmo utilizado o Teste de Poisson. Em sua análise descritiva, Silva [24] mostra que $203$ trabalhadores apresentaram pelo menos um afastamento no ano do estudo, sendo o maior causador doenças respiratórias. Do número total de entrevistados, $79\%$ são do sexo feminino e $55\%$ possuem um único vínculo empregatício, trabalham $30$ horas semanais e no pronto atendimento.

Silva [24] utilizou uma primeira análise simples, teste exato de Fisher e o teste não paramétrico de Mann-Whitney para identificar os fatores associados (p-valor $< 0.20$) ao absenteísmo-doença. O autor realiza a regressão logística para as variáveis associadas significativamente; desse modo, os únicos grupos não inclusos são o dos enfermeiros e o dos técnicos de enfermagem.

Landgraf [13], assim como os autores acima, buscou identificar o causador de absenteísmo em uma empresa, desta vez do setor alimentício. Para isso, em sua dissertação de mestrado, realizou um questionário fechado e coletou um relatório de faltas de $30$ empregados, todos do departamento operacional. A análise descritiva da autora apresenta motivos de saúde como o maior causador de abstenções.

Dentre os artigos pesquisados, verificou-se que o maior causador de absenteísmo são doenças, podendo demonstrar um padrão nas justificativas de ausências dos brasileiros, ou ainda uma facilidade em justificar as faltas. Outro fator comum são as ações sugeridas para reduzir o absenteísmo e os prejuízos por ele causados. São recomendadas melhorias no local de trabalho e a implementação de programas de segurança que aumentem a satisfação do funcionário, chequem as ausências e distribuam prêmios por assiduidade.

Apesar de ter se originado no século XVIII, o estudo do absenteísmo ainda

hoje se justifica, visto serem poucas as publicações a seu respeito e suas causas variarem de acordo com o contexto de cada empresa. Ao todo, sabe-se que o fenômeno afeta negativamente tanto a renda dos empreendimentos quanto os absentes e o fornecimento dos serviços conforme apresentado por Almeida e Silva [2, 23]. O objetivo desta dissertação é propor uma metodologia nova para a análise de dados multivariados relacionado ao absenteísmo. O fato dos trabalhos mencionados estarem associados a áreas de atuação distintas e realizarem análises descritivas, mostra a importância de seu estudo, uma vez que afeta todas as áreas de trabalho e, apesar de as empresas adotarem medidas para reduzi-lo, até hoje não se sabe como evitá-lo.

O corpo do trabalho divide-se em três partes, sendo o Capítulo 2 uma revisão bibliográfica sobre o absenteísmo, Seções 2.1 e 2.2, e as metodologias que serão utilizadas na proposta metodológica estão presentes entre as Seções 2.3 e 2.9, sendo estas a análise de componentes principais, seguida do *biplot*, do coeficiente de correlação de Spearman, técnicas para determinação de *clusters*, determinação do número $g$ de *clusters* e, ao final, o método de comparação de grupos de Kruskal Wallis.

O Capítulo 3 apresenta a proposta de análise multivariada aos dados sobre o absenteísmo e as informações sobre os dados coletados. O Capítulo 4 traz a aplicação dos dados a metodologia proposta na Seção 3.1. A primeira análise desse estudo foi realizada sobre os dados como um todo e a segunda, após uma separação, utilizou-se a técnica de Ward. Uma vez formados tais grupos, foram realizadas comparações, por meio de técnicas não paramétricas, buscando determinar diferenças em suas formações.

A Seção 5, que apresenta a conclusão da pesquisa, é seguida pelo anexo A, que contém o código utilizado na Seção 4, uma vez que todo o estudo foi realizado com a utilização do software R [20] com o auxílio do Rstudio [26].

# 2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta, uma breve revisão da origem do termo absenteísmo, quais variáveis estão contidas neste termo e o porque do desenvolvimento do índice de absenteísmo. Na sequência apresenta-se uma revisão das seguintes técnicas: análise de componentes principais (PCA), *biplot*, coeficiente de correlação de Spearman, análise de *clusters* e comparação pelo teste de Kruskal Wallis.

## 2.1 ABSENTEÍSMO

A prática de um colaborador ausentar-se de seu local de serviço é denominada absenteísmo. Apesar de ser conhecida e gerar prejuízos e/ou transtornos operacionais, ela ainda é pouco estudada fora das áreas de administração e saúde. Trabalhos como os expostos a seguir ajudam no entendimento do absenteísmo.

O termo absenteísmo, como conhecido na atualidade, é derivado da palavra "absentismo", que, segundo Penatti [19] apud Quick e Laperlosa, apenas no período industrial passou a ser utilizada para caracterizar trabalhadores faltosos. Antes disso, seu uso era sempre relacionado ao êxodo rural.

Apesar de o termo ser utilizado desde o século XIX para caracterizar esse comportamento absente, não se tem uma descrição concreta de quais as causas e variáveis do absenteísmo, havendo ainda hoje inúmeras definições para se estudar.

Chiavenato [6] define o absenteísmo como um termo utilizado para caracterizar o tempo de ausência do funcionário durante o seu período de trabalho. Dessa maneira, abrange faltas, atrasos, licenças ou outros motivos que mantenham o trabalhador afastado.

Para Landgraf [13] apud Couto, o absenteísmo é dividido em: "Absenteísmo voluntário, absenteísmo por doença, absenteísmo por patologia profissional, absenteísmo legal e absenteísmo compulsório".

Calais [4], por sua vez, afirma que o absenteísmo está relacionado diretamente ao estresse do funcionário, que pode ou não ser causado pela empresa. Ou ainda, como Lee e Eriksen [15], resumem o absenteísmo como inversamente proporcional à satisfação no trabalho, ou seja,

$$\text{Absenteísmo} = \frac{1}{\text{Satisfação}}.$$

A partir dessas definições, pode-se chegar à conclusão de que o termo absenteísmo é utilizado para caracterizar afastamentos do serviço, independentemente do motivo. Desse modo, pode-se classificar as principais causas do absenteísmo em dois grandes grupos: as ausências amparadas pela lei, como faltas justificadas, licenças, óbitos e casamentos, ou ainda as não amparadas, como faltas não justificadas, atrasos e suspensões.

Algumas ausências são esperadas pela companhia, de modo que podem ser tratadas com antecedência, como é o caso de férias e licenças. Contudo, a grande maioria das abstenções não são informadas às empresas até o dia em questão.

Partindo-se dessas definições, desenvolveu-se um indicador denominado índice de absenteísmo.

## 2.2 Índice de Absenteísmo

Segundo Penatti [19], esse índice é obtido pelo somatório dos períodos de ausência e atrasos do funcionário dentro de sua jornada de trabalho, isto é,

$$\text{Índice Absenteísmo} = \frac{\text{Períodos de ausências e atrasos}}{\text{jornada de trabalho}}. \tag{2.1}$$

Chiavenato [6] propõe um índice similar ao de Penatti, conforme exposto na igualdade (2.2). Desse modo, o índice de absenteísmo é obtido de uma proporção de horas perdidas por horas trabalhadas.

$$\text{Índice Absenteísmo} = \frac{\left(\dfrac{\text{Total de pessoas}}{\text{horas perdidas}}\right)}{\left(\dfrac{\text{Total de pessoas}}{\text{horas de trabalho}}\right)}. \tag{2.2}$$

O índice resultante da equação (2.2) ainda pode ser expresso de forma mensal ou anual, dependendo do interesse do pesquisador. Chiavenato [6] ainda afirma que o número de funcionários atuantes no período em questão pode ser obtido calculando-se

$$\text{Força de trabalho atuante} \;=\; 1 - \big(\text{ Índice Absenteísmo }\big).$$

Dessa forma, a empresa detém a informação do número de funcionários ausentes e ainda consegue repor ou redistribuir os colaboradores presentes para os principais locais de demanda.

## 2.3 Análise de componentes principais

A análise de componentes principais (PCA) é uma técnica multivariada desenvolvida por Karl Pearson, em 1901. Manly [16] a descreveu como "uma das mais simples", uma vez que seu objetivo é diminuir o número de variáveis, transformando-as em combinações lineares.

**Definição 2.1.** *Seja* $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]^T$ *um vetor aleatório, com* $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]^T$ *sendo o vetor de médias e* $\boldsymbol{O}_{p\times p}$ *a matriz ortogonal. Então, a componente principal é definida como* $\boldsymbol{Y} = \boldsymbol{O}^T\boldsymbol{X}$.

A matiz $\boldsymbol{O}_{p\times p}$ é construída por meio dos autovetores normalizados da matriz de covariância ($\boldsymbol{S}_{p\times p}$) ou correlação ($\boldsymbol{R}_{p\times p}$) enquanto, a matriz $\boldsymbol{\Lambda}_{p\times p}$ é construída por meio dos autovalores, ou seja,

$$\boldsymbol{O}_{p\times p} = \begin{bmatrix} e_{11} & e_{21} & \dots & e_{p1} \\ e_{12} & e_{22} & \dots & e_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1p} & e_{2p} & \dots & e_{pp} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Lambda}_{p\times p} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix},$$

em que, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. O vetor $[e_1\ e_2\ \ldots\ e_p]$ ainda satisfaz as seguintes condições, segundo Mingoti [17]:

i. $e_i^t e_j = 0, \quad \forall i \neq j,$

ii. $e_i^t e_i = 1, \quad \forall i = 1, 2, \ldots, p,$

iii. $S_{p\times p} e_i = \lambda_i e_i, \quad \forall i = 1, 2, \ldots, p,$

na qual, $S_{p\times p}$ é a matriz de covariância amostral. Sendo a esperança e a variância das componentes principais dadas pelas expressões

$$E[\boldsymbol{Y}_j] = \boldsymbol{e}_j^t \boldsymbol{\mu} \qquad \text{e} \qquad Var[\boldsymbol{Y}_j] = \lambda_j,$$

em que, $e_j$ é o autovetor normalizado correspondente a $j$-ésima variável, $\boldsymbol{\mu}$ é o vetor de médias e $\lambda_j$ é o $j$-ésimo autovalor.

Como o objetivo da análise de componentes principais é diminuir as $p$ combinações lineares e analisar apenas as $k$ primeiras $(k < p)$ componentes principais, utiliza-se a variabilidade total explicada para determinar o número $k$ de componentes principais estudadas.

**Definição 2.2.** *A variabilidade total é dada por*

$$tr(\boldsymbol{S}_{p\times p}) = \sum_{i=1}^{p} \lambda_i,$$

*ou ainda, como a frequência da variabilidade total explicada, segundo Ferreira [8]:*

$$\textit{Variabilidade Total Explicada}(Y_k) = \frac{\lambda_k}{\sum_{k=1}^{p} \lambda_k} \times 100.$$

Segundo Manly [16] usualmente utiliza-se duas componentes principais desde de que apresente uma variabilidade total considerada alta, acima de $80\%$. Essas definições são

válidas para as estimações das componentes principais para a matrizes de covariância e correlação amostral, sendo estas duas, respectivamente,

$$\boldsymbol{S}_{p\times p} = \sum_{j=1}^{p} \lambda_j e_j e_j^t \qquad \text{e} \qquad \boldsymbol{r}_{\boldsymbol{Y}_j \boldsymbol{Z}_i} = e_{ji}\sqrt{\lambda_j},$$

em que, $e_j$ é o $j$-ésimo autovetor, $\lambda_j$ o $j$-ésimo autovalor e o elemento $e_{ij}$ correspondente a matriz ortogonal $\boldsymbol{O}_{p\times p}$. Observa-se que a variável $\mathbf{Z}$ é a variável $\mathbf{X}$ normalizada, ou seja, é o mesmo processo de obter as componentes principais utilizando a matriz de correlação.

Segundo Mingoti [17] a utilização das componentes principais obtidas a partir da matriz de covariância apresenta um domínio da variável com maior variância, sendo assim utilizada apenas em casos que as variâncias não são muito discrepantes. Buscando-se diminuir essa influencia pode-se utilizar as componentes principais obtidas por meio da matriz de correlação, para os dados que apresentarem essas discrepâncias.

**Exemplo 1.** Calcule a componente principal para a matriz $\boldsymbol{X}_{4\times 3}$ de dados. Seja $\boldsymbol{X}_{4\times 3}$ dada por

$$\mathbf{X_{4\times 3}} = \begin{bmatrix} 20 & -9 & 6 \\ 6 & 12 & -15 \\ -10 & -6 & 9 \\ 8 & -12 & 12 \\ 6 & -9 & 20 \end{bmatrix}.$$

Calculando a componente principal por meio da matriz de correlação, tem-se

$$\mathbf{R_{3\times 3}} = \begin{bmatrix} 1,00 & -0,13 & -0,06 \\ -0,13 & 1,00 & -0,91 \\ -0,06 & -0,91 & 1,00 \end{bmatrix},$$

então seus autovalores e vetores são, respectivamente,

$$\mathbf{\Lambda_{3\times 3}} = \begin{bmatrix} 1,91 & 0,00 & 0,00 \\ 0,00 & 1,02 & 0,00 \\ 0,00 & 0,00 & 0,07 \end{bmatrix} \quad \text{e} \quad \mathbf{O_{3\times 3}} = \begin{bmatrix} 0,05 & 0,99 & -0,15 \\ -0,71 & -0,07 & -0,70 \\ 0,70 & -0,14 & -0,70 \end{bmatrix}.$$

Dessa forma, a componente principal é dada por $\boldsymbol{Y}^T = \boldsymbol{O}^T\boldsymbol{X}^T$.

$$\boldsymbol{Y}^T_{4\times 3} = \begin{bmatrix} -11,59 & -19,59 & 0,90 \\ 18,72 & -7,20 & -1,20 \\ -10,06 & 10,74 & 0,60 \\ -17,32 & -7,08 & 1,20 \\ -20,69 & -3,77 & 8,60 \end{bmatrix}$$

Assim, cada coluna representa uma componente principal. A variabilidade total explicada das duas primeiras componentes são $63,82\%$ e $33,94\%$, ou seja, pode-se utilizar apenas as duas primeiras componentes para explicar os dados, pois representam $97,76\%$ dos dados.

## 2.4 *Biplot*

*Biplot* é um gráfico utilizado para representar simultaneamente as observações e as variáveis em um espaço bidimensional, de forma que sejam expressas algumas informações em um "espaço reduzido", como explica Ferreira [9]. Tais informações são: a distância entre as observações, a importância das variáveis e o fato de o cosseno entre as variáveis ser aproximadamente igual à correlação.

Para que esses requisitos sejam satisfeitos, é necessário realizar a decomposição em valores singulares da matriz de dados $\boldsymbol{Y_{n\times p}}$.

**Definição 2.3.** *Seja $\boldsymbol{Y}_{n\times p}$ a matriz de dados em que $n$ representa o número de observações e $p$ o número de variáveis. Desse modo, $\boldsymbol{Y}_{n\times p}$ é escrita da seguinte forma:*

$$\boldsymbol{Y}_{n\times p} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix},$$

representando o elemento $y_{np}$ da matriz a informação da $n$-ésima observação correspondente à variável $p$.

**Definição 2.4.** *Decomposição de $\boldsymbol{Y}_{n\times p}$ em valores singulares (DVS), conforme Souza [25].*

*Seja $\boldsymbol{Y}_{n\times p} \in \mathbb{R}^{n\times p}$ e rank$(\boldsymbol{Y}) \leq r < min\{n, p\}$, desse modo a DVS é dada por*

$$\boldsymbol{Y}_{(r)n\times p=} = \boldsymbol{U}_{(r)n\times r}\boldsymbol{\Lambda}_{(r)r\times r}\boldsymbol{V}^T_{(r)r\times p},$$

*em que*

*i. As colunas das matrizes $\boldsymbol{U}_{(r)n\times r}$ e $\boldsymbol{V}_{r(p\times r)}$ são ortogonais, isto é,*

$$\boldsymbol{U}^T_{(r)r\times n}\boldsymbol{U}_{(r)n\times r} = \boldsymbol{V}^T_{(r)r\times p}\boldsymbol{V}_{(r)p\times r} = \boldsymbol{I}_{(r)};$$

*ii. $\boldsymbol{U}$ é uma matriz de autovetores da matriz $\boldsymbol{YY}^T$;*

*iii. $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$ formada pelos $r$ maiores autovalores de $\boldsymbol{YY^T}$ ou $\boldsymbol{Y^TY}$;*

*iv. $\boldsymbol{V}$ é uma matriz de autovetores da matriz $\boldsymbol{Y}^T\boldsymbol{Y}$.*

*Assim, a decomposição em valores singulares também pode ser expressa como*

$$\boldsymbol{Y}_{(r)} = \sum_{i=1}^{r} \lambda_i u_i v_i^t, \tag{2.3}$$

*sendo $\lambda_i$, $u_i$ e $v_i$ o i-ésimo elemento de suas respectivas matrizes.*

O *biplot* da matriz $\boldsymbol{Y}_{(r)n\times p}$ é uma representação gráfica realizada por meio de vetores. Esses vetores são chamados de marcadores para linhas, $a_1, a_2, \ldots, a_n$ e $b_1, b_2, \ldots, b_n$ para colunas de $\boldsymbol{Y}_{(r)n\times p}$. Eles podem ser representados, de acordo com Souza [25], como

$$a_i^t b_j^t \cong y_{ij} \qquad \text{com} \quad i = 1, 2, \ldots, n \quad \text{e} \quad j = 1, 2, \ldots, p$$

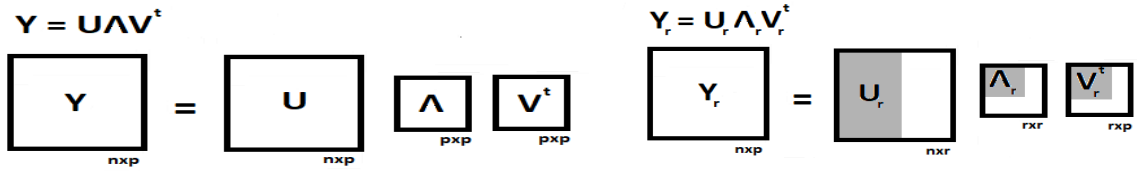
ou, ainda,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{B}^T. \tag{2.4}$$

De acordo com Souza [25], ao aproximar duas matrizes $\boldsymbol{Y}_{n\times p}$, uma com rank(r) e outra com rank(q), sendo $q < r$, equações (2.3) e (2.4), verifica-se:

$$\boldsymbol{Y}_{(r)} \approx \boldsymbol{Y}_{(q)} \Rightarrow \boldsymbol{U}_{(q)n\times q}\boldsymbol{\Lambda}_{(q)q\times q}\boldsymbol{V}^T_{(q)q\times p} = \boldsymbol{A}_{(q)n\times q}\boldsymbol{B}^T_{(q)q\times p}. \tag{2.5}$$

A Figura 2.1 ilustra essa aproximação: (a) apresenta a DVS para o posto original e (b) expõe a DVS com a aproximação do posto.




(a) (b)

Figura 2.1: Ilustra a mudança de posto da matriz $\boldsymbol{Y}_{n\times p}$.
Fonte: Reproduzido a partir de dos Santos [7], 2012, p.33.

Deste modo, a partir da equação (2.5), verificam-se os seguintes resultados:

$$\boldsymbol{A}_{(q)n\times q} = \boldsymbol{U}_{(q)n\times q}\boldsymbol{\Lambda}^s_{(q)q\times q} \quad \text{e} \quad \boldsymbol{B}^T_{(q)q\times p} = \boldsymbol{\Lambda}^{1-s}_{(q)q\times q}\boldsymbol{V}^T_{(q)q\times p},$$

sendo $s$ uma constante entre $0$ e $1$.

Assim, a escolha dos marcadores de linha e coluna depende de qual importância de representação o pesquisador deseja priorizar. Essa importância é exposta pela qualidade da representação.

Caso se deseje priorizar as colunas, a máxima qualidade de representação delas será obtida utilizando-se o GH-Biplot. Por outro lado, ao se utilizar o JK-Biplot, será

priorizada a qualidade de representação das linhas, enquanto a das colunas será minimizada, conforme Ferreira [10].

Se a intenção for manter a mesma importância para as linhas e as colunas, o pesquisador poderá escolher entre os métodos SQRT-Biplot e HJ-Biplot, que buscam a máxima qualidade de representação para ambas.

Uma outra maneira de verificar a qualidade de representação do *biplot* é utilizando a qualidade global de representação, que demonstra o quão distante está a aproximação de postos das matrizes $\boldsymbol{Y}_{(r)n\times p}$ por $\boldsymbol{Y}_{(q)n\times p}$, conforme apresentado por Souza [25], sendo possível, deste modo, verificar a precisão do *biplot*, o erro de aproximação.

De acordo com Souza [25], a qualidade global da representação é definida como:

$$\text{Qualidade global de representação} = \frac{(\lambda_i)^2}{\sum_{k=1}^{q}(\lambda_k)^2},$$

obtida de maneira similar ao coeficiente de determinação $R^2$. Caso esta qualidade não seja ainda o suficiente para avaliar os ajustes individuais, pode-se utilizar a qualidade de representação para linhas e colunas.

A escolha dos marcadores de linha e coluna para $\boldsymbol{Y}_{(q)n\times p}$ pode ser realizada de diversas maneiras pela aproximação de postos como mostrado na equação (2.5). Com isso apresenta-se na Tabela 2.1 o resumo dos métodos e qualidades de representação.

**Exemplo 2.** Seja $\boldsymbol{X}$ a matriz de dados e $s$ o valor que determina a representação em função da decomposição em valores singulares (DVS). Desse modo, $\boldsymbol{X}$ é expressa como:

$$\mathbf{X} = \begin{bmatrix} 20 & -9 & 6 \\ 6 & 12 & -15 \\ -10 & -6 & 9 \\ 8 & -12 & 12 \\ 6 & -9 & 20 \end{bmatrix}$$

e, utilizando-se o software R [20] para $s = 0$, obtém-se a representação GH-Biplot com as seguintes coordenadas cartesianas referentes às duas componentes da decomposição em valores singulares,

$$\mathbf{U} = \begin{bmatrix} -0,26 & 1,32 \\ 1,72 & 0,11 \\ -0,11 & -1-48 \\ -0,61 & 0,17 \\ -0,75 & -0,12 \end{bmatrix} \quad \text{e} \quad \mathbf{\Lambda V} = \begin{bmatrix} -0,07 & 1,00 \\ 0,98 & -0,07 \\ -0,97 & -0,14 \end{bmatrix}$$

ou, ainda, sua representação no *biplot* (Figura 2.2).

Dessa forma, utilizando $s = 0,5$ (priorizando igualmente as observações e

Tabela 2.1: Alguns valores de s

| | **Representação simultânea** | **Coordenadas linhas** | **Coordenadas colunas** | **Qualidade de representação linhas** | **Qualidade de representação colunas** |
|---|---|---|---|---|---|
| $s = 0$ | GH-Biplot | $\boldsymbol{U}$ | $\boldsymbol{\Lambda V}$ | $\frac{q}{r}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ |
| $s = 1$ | JK-Biplot | $\boldsymbol{U\Lambda}$ | $\boldsymbol{V}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ | $\frac{q}{r}$ |
| $s = 0,5$ | SQRT-Biplot | $\boldsymbol{U\Lambda}^{\frac{1}{2}}$ | $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ |
| | HJ-Biplot | $\boldsymbol{U\Lambda}$ | $\boldsymbol{\Lambda V}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ | $\frac{\sum_{k=1}^{q}\lambda_k^4}{\sum_{k=1}^{r}\lambda_k^4}$ |

Fonte: Reproduzido a partir de Souza [25], 2010, p.41.

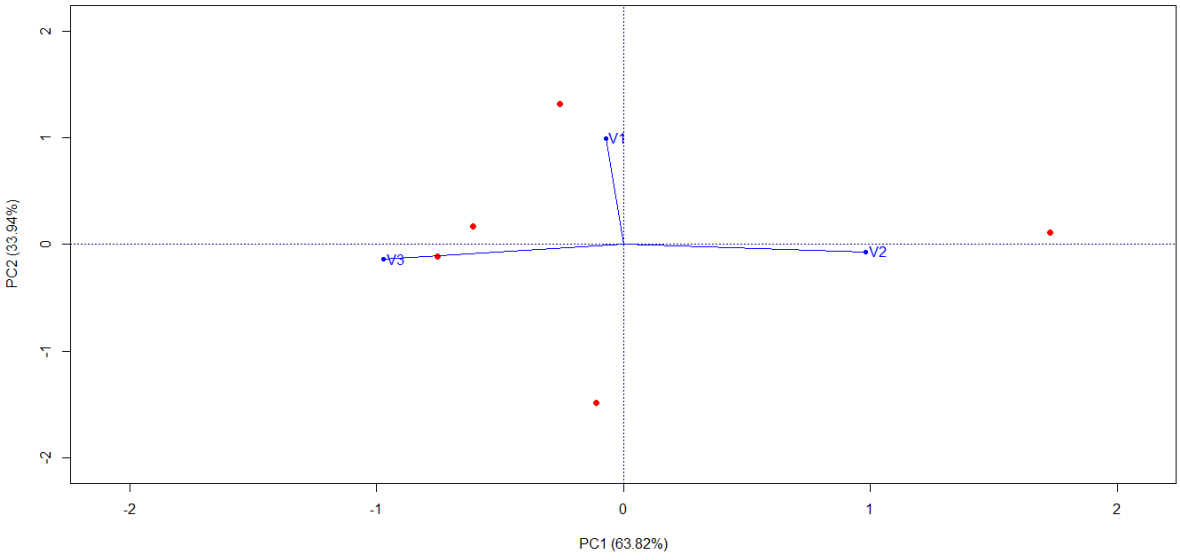



Figura 2.2: Ilustração do *Biplot* para $s = 0$, GH-Biplot. Os pontos em vermelho representam as observações e as setas em azul, as variáveis, sendo priorizada a representação destas.
Fonte: Os autores, 2021.

variáveis), obtém-se a representação SQRT-Biplot com as seguintes componentes do *biplot*.

$$\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}} = \begin{bmatrix} -0,21 & 0,94 \\ 1,44 & 0,08 \\ -0,09 & -1,06 \\ -0,51 & 0,12 \\ -0,63 & -0,08 \end{bmatrix} \quad \text{e} \quad \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V} = \begin{bmatrix} -0,09 & 1,40 \\ 1,18 & -0,10 \\ -1,17 & -0,20 \end{bmatrix},$$

ou expressa no *biplot* a seguir (Figura 2.3).

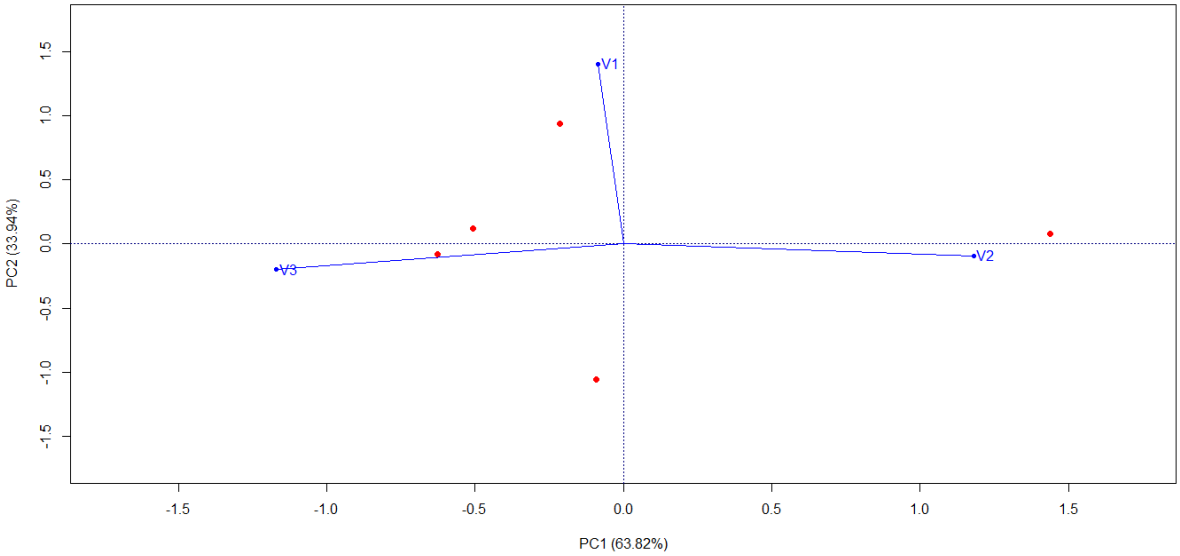



Figura 2.3: Ilustração do *Biplot* para $s = 0,5$, SQRT-Biplot. Os pontos em vermelho representam as observações e as setas em azul, as variáveis, ambas igualmente priorizadas.
Fonte: Os autores, 2021.

Nota-se que, apesar das coordenadas cartesianas apresentarem alterações, a interpretação é a mesma, com as variáveis $V2$ e $V3$ correlacionadas forte e negativamente; como $V1$ com $V3$ e $V2$ estão quase perpendiculares, suas correlações são quase nulas. Pode-se observar que o marcador de $V2$ indica maior peso à primeira componente e o $V1$, para a segunda. O *biplot* ainda mostra a dispersão dos pontos com um ponto, denominado 2, distante dos demais.

## 2.5 Coeficiente de correlação de postos de Spearman

A correlação de Spearman é o estudo e quantificação do relacionamento entre as variáveis duas a duas $(X, Y)$ para variáveis que não são normalmente distribuídas. É também denominada frequentemente como rho de Spearman ou $\rho$ de Spearman, em homenagem ao psicólogo e estatístico Charles Spearman (1863 - 1945).

O coeficiente é uma modificação do coeficiente de Pearson, no qual os valores observados de cada variável são codificados utilizando postos. A expressão matemática para o coeficiente de Spearman é dada por

$$r_s = 1 - \frac{6\sum_{i=1} d_i^2}{n\left(n^2-1\right)}, \tag{2.6}$$

em que $d_i$ é a diferença entre os postos obtidos em $x_i$ e $y_i$, e $n$ é o número de observações. O coeficiente $r_s$ é limitado entre -1 e 1. Valores próximos de 1 para $r_s$ indicam uma forte associação positiva entre $X$ e $Y$; valores de $r_s$ próximos de -1 representam uma forte associação negativa entre as variáveis comparadas. Por outro lado, se $r_s$ é identicamente igual a zero ou assume valores próximos de zero, diz-se que a correlação entre as variáveis é fraca ou inexistente.

Entretanto, um valor observado de $r_s$ pode ser resultado do acaso, devido à aleatoriedade da amostra e, portanto, faz-se necessária a realização de teste de hipótese para correlação. As hipóteses do teste são:

$$\begin{cases} H_0 : \text{“não existe associação entre X e Y”} \\ H_1 : \text{“Existe associação entre X e Y”} \end{cases}.$$

A estatística do teste é dada pela expressão

$$T = r_s\sqrt{\frac{n-2}{1-r_s^2}}, \tag{2.7}$$

em que a variável aleatória $T$ tem distribuição de probabilidade $t$-Student com $n-2$ graus de liberdade. Assim, $H_0$ é rejeitada se $p$-valor $= \mathbb{P}(T \geq t) < \alpha$ e conclui-se que as variáveis são correlacionadas.

**Exemplo 3.** Utiliza-se a matriz de dados $\boldsymbol{X}_{5\times 3}$ para calcular a matriz de correlação de Spearman. Seja

$$\mathbf{X_{5\times 3}} = \begin{bmatrix} 20 & -9 & 6 \\ 6 & 12 & -15 \\ -10 & -6 & 9 \\ 8 & -12 & 12 \\ 6 & -9 & 20 \end{bmatrix}$$

a matriz de dados, cujas linhas representam as observações e as colunas, as variáveis. Então, aplicando-a na fórmula do cálculo de correlação, (2.6), tem-se

$$\mathbf{R_{3\times 3}} = \begin{bmatrix} 1,00 & -0,61 & -0,10 \\ -0,61 & 1,00 & -0,67 \\ -0,10 & -0,67 & 1,00 \end{bmatrix}$$

e, pelo teste de correlação, (2.7), os seguintes p-valores:

$$\mathbf{p-valor} = \begin{bmatrix} \approx 0 & 0,2794 & 0,8696 \\ - & \approx 0 & 0,2189 \\ - & - & \approx 0 \end{bmatrix}.$$

Dessa forma, ao nível de significância de $5\%$ verifica-se que as variáveis da matriz de dados $\boldsymbol{X}$ não são correlacionadas entre si.

## 2.6 TÉCNICAS PARA CONSTRUÇÃO DE *clusters*

Para Chatfield [5], Mingoti [17] e Johnson [12], a análise de *clusters* é uma ferramenta usada na estatística multivariada para formação e classificação de grupos de acordo com características presentes em toda observação, tendo cada um deles uma especificidade. Assim, os agrupamentos são heterogêneos entre si, e as anotações particulares de cada um são homogêneas.

As técnicas para construção dos *clusters* são divididas em hierárquicas e não hierárquicas. As primeiras são do tipo aglomerativo ou divisivo. As aglomerativas partem de observações únicas, e são realizadas fusões até todas estarem em um mesmo grupo; as divisivas fazem o processo contrário. Essas técnicas buscam determinar um número $g$ de grupos com características remotas, segundo Mingoti [17]. A Figura 2.4 apresentada na sequência ilustra as técnicas hierárquicas aglomerativas e divisivas.

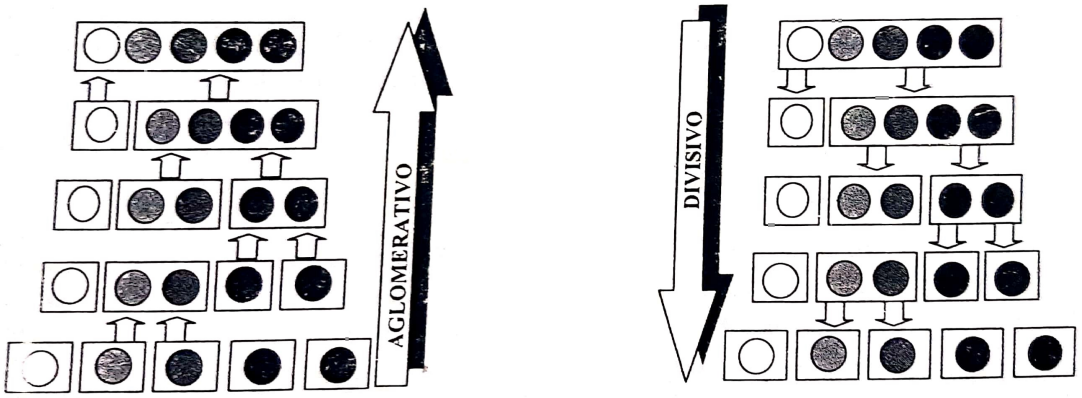



Figura 2.4: Imagem ilustrativa dos métodos hierárquicos - Aglomerativo e divisivo.
Fonte: Sueli Aparecida Mingoti, 2005, p.164.

A técnica não hierárquica, por sua vez, parte do princípio inverso, ou seja, o número de grupos $g$ já está determinado. Assim, a técnica não hierárquica não apresenta a formação de todos os conglomerados como a técnica hierárquica apenas apresenta o resultado para o número $g$ determinado de grupos. Um exemplo de seu uso é o método das k-médias, que divide as observações em g grupos e na sequência utiliza um algoritmo para determinar a qual

deles pertence a observação, conforme exposto por Ferreira [8]. A Figura 2.5 apresenta uma representação do método das k-médias.

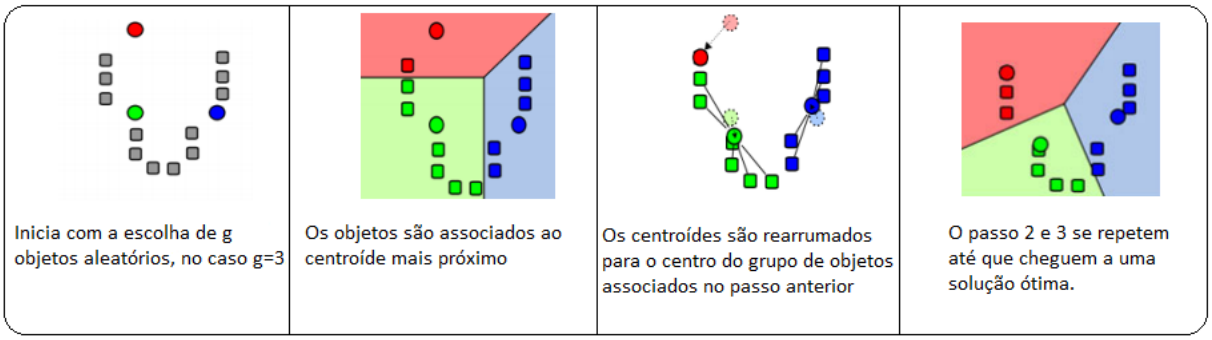



Figura 2.5: Ilustração do método das k-médias para 12 observações e 3 *clusters*.
Fonte: H. Di, M. Shafiq, G. AlRegib*, 2017.

### 2.6.1 Técnica hierárquica aglomerativa

As técnicas aglomerativas unem os *clusters* por medidas de similaridade ou dissimilaridade. Esses valores são calculados por meio de distâncias, que podem ser euclidianas, de Mahalanobis, entre outras. Os métodos mais comuns para a formação de *clusters* são:

**Definição 2.5.** ***- Método de Ligação Simples (Single Linkage):****Também conhecido como método de vizinhos mais próximos, consiste em analisar a similaridade de dois conglomerados e "juntar" os mais próximos, conforme Mingoti[17].*

Esse método une os conglomerados, de modo que os conglomerado tenham a menor distância, ou seja, $C_{ij} = min\{d(C_i, C_j), \ \ i \neq j\}$. Por exemplo:

**Exemplo 4.** Seja a matriz $\boldsymbol{D}$ de distâncias e aplicando-se o método da ligação simples, tem-se

$$\boldsymbol{D} = \begin{bmatrix} & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 \\ C_1 & 0 & 6 & 19 & 13 & 5 & 7 & 2 \\ C_2 & 6 & 0 & 24 & 19 & 9 & 13 & 7 \\ C_3 & 19 & 24 & 0 & 6 & 18 & 13 & 19 \\ C_4 & 13 & 19 & 6 & 0 & 14 & 7 & 14 \\ C_5 & 5 & 9 & 18 & 14 & 0 & 8 & 5 \\ C_6 & 7 & 13 & 13 & 7 & 8 & 0 & 9 \\ C_7 & 2 & 7 & 19 & 14 & 5 & 9 & 0 \end{bmatrix};$$

note que a menor medida de distância pertence ao conglomerado $C_1$ e $C_7$. Esses conglomerados são então reunidos por meio da técnica da ligação simples, ou seja, o $C_{17}$ é expresso como:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $C_1$ | 0 | 6 | 19 | 13 | 5 | 7 | 2 |
| $C_7$ | 2 | 7 | 19 | 14 | 5 | 9 | 0 |
| $C_{17} = min\{d(C_1, C_7)\}$ | - | 6 | 19 | 13 | 5 | 7 | - |

Então,

$$\boldsymbol{D}_1 = \begin{bmatrix} & C_{17} & C_2 & C_3 & C_4 & C_5 & C_6 \\ C_{17} & 0 & 6 & 19 & 13 & 5 & 7 \\ C_2 & 6 & 0 & 24 & 19 & 9 & 13 \\ C_3 & 19 & 24 & 0 & 6 & 18 & 13 \\ C_4 & 13 & 19 & 6 & 0 & 14 & 7 \\ C_5 & 5 & 9 & 18 & 14 & 0 & 8 \\ C_6 & 7 & 13 & 13 & 7 & 8 & 0 \end{bmatrix}.$$

Realizando-se o mesmo processo, verifica-se que os conglomerados $C_{17}$ e $C_5$ tem a menor distância. Logo, aplicando novamente o método, obtém-se o novo conglomerado $C_{175}$ no lugar do conglomerado $C_{17}$ e $C_5$.

| | | | | | | |
|---|---|---|---|---|---|---|
| $C_{17}$ | 0 | 6 | 19 | 13 | 5 | 7 |
| $C_5$ | 5 | 9 | 18 | 14 | 0 | 8 |
| $C_{175} = min\{d(C_{17}, C_5)\}$ | - | 6 | 18 | 13 | - | 7 |

Desse modo, a nova matriz $\boldsymbol{D}$ contendo o conglomerado $C_{175}$ é expressa como

$$\boldsymbol{D}_2 = \begin{bmatrix} & C_{175} & C_2 & C_3 & C_4 & C_6 \\ C_{175} & 0 & 6 & 18 & 13 & 7 \\ C_2 & 6 & 0 & 24 & 19 & 13 \\ C_3 & 18 & 24 & 0 & 6 & 13 \\ C_4 & 13 & 19 & 6 & 0 & 7 \\ C_6 & 7 & 13 & 13 & 7 & 0 \end{bmatrix}.$$

Aplicando-se o método até que todos estejam no mesmo grupo, verificam-se as seguintes matrizes:

$$\boldsymbol{D}_3 = \begin{bmatrix} & C_{1752} & C_3 & C_4 & C_6 \\ C_{1752} & 0 & 18 & 13 & 7 \\ C_3 & 18 & 0 & 6 & 13 \\ C_4 & 13 & 6 & 0 & 7 \\ C_6 & 7 & 13 & 7 & 0 \end{bmatrix}, \quad \boldsymbol{D}_4 = \begin{bmatrix} & C_{1752} & C_{34} & C_6 \\ C_{1752} & 0 & 13 & 7 \\ C_{34} & 13 & 0 & 7 \\ C_6 & 7 & 7 & 0 \end{bmatrix}$$

$$\text{e} \quad \boldsymbol{D}_5 = \begin{bmatrix} & C_{17526} & C_{34} \\ C_{17526} & 0 & 7 \\ C_{34} & 7 & 0 \end{bmatrix}.$$

Desta forma, os grupos foram estabelecidos na seguinte ordem:

| Passos $k$ | Nº de grupos $g$ | Grupos |
|---|---|---|
| 1 | $g = 7$ | $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ |
| 2 | $g = 6$ | $C_{17}$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ |
| 3 | $g = 5$ | $C_{175}$, $C_2$, $C_3$, $C_4$, $C_6$ |
| 4 | $g = 4$ | $C_{1752}$, $C_3$, $C_4$, $C_6$ |
| 5 | $g = 3$ | $C_{1752}$, $C_{34}$, $C_6$ |
| 6 | $g = 2$ | $C_{17526}$, $C_{34}$ |
| 7 | $g = 1$ | $C_{1752634}$ |

**Definição 2.6.** ***- Método de Ligação Completa (Complete Linkage):*** *Ainda denominado como método de vizinhos distantes, consiste em analisar os conglomerados a cada nível de interação e combinar os mais dissimilares até formarem um único cluster, segundo Mingoti [17].*

Parte-se do mesmo princípio do Método de Ligação Simples, no qual unem-se os conglomerados com menores distâncias, mas agora selecionando-se o máximo deles. Ou seja, $C_{ij} = \text{máx}\{d(C_i, C_j), \ \ i \neq j\}$, como demonstra-se no exemplo 5:

**Exemplo 5.** Seja a matriz $\boldsymbol{D} =$

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ |
|---|---|---|---|---|---|---|---|
| $C_1$ | 0 | 6 | 19 | 13 | 5 | 7 | 2 |
| $C_2$ | 6 | 0 | 24 | 19 | 9 | 13 | 7 |
| $C_3$ | 19 | 24 | 0 | 6 | 18 | 13 | 19 |
| $C_4$ | 13 | 19 | 6 | 0 | 14 | 7 | 14 |
| $C_5$ | 5 | 9 | 18 | 14 | 0 | 8 | 5 |
| $C_6$ | 7 | 13 | 13 | 7 | 8 | 0 | 9 |
| $C_7$ | 2 | 7 | 19 | 14 | 5 | 9 | 0 |

;

Aplicando-se o método de ligação completa, verifica-se que a menor distância é o elemento $X_{17}$. Desse modo, retirando-se as linhas e colunas referentes a esses índices, obtém-se uma nova matriz **D**, na qual a linha e a coluna novas serão expressas como o máximo dessas linhas.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $C_1$ | 0 | 6 | 19 | 13 | 5 | 7 | 2 |
| $C_7$ | 2 | 7 | 19 | 14 | 5 | 9 | 0 |
| $C_{17} = \text{máx}\{d(C_1, C_7)\}$ | - | 7 | 19 | 14 | 5 | 9 | - |

Então,

$$\boldsymbol{D}_1 = \begin{bmatrix} & C_{17} & C_2 & C_3 & C_4 & C_5 & C_6 \\ C_{17} & 0 & 7 & 19 & 14 & 5 & 9 \\ C_2 & 7 & 0 & 24 & 19 & 9 & 13 \\ C_3 & 19 & 24 & 0 & 6 & 18 & 13 \\ C_4 & 14 & 19 & 6 & 0 & 14 & 7 \\ C_5 & 5 & 9 & 18 & 14 & 0 & 8 \\ C_6 & 9 & 13 & 13 & 7 & 8 & 0 \end{bmatrix}.$$

Dessa forma, repetindo-se os processos, verificam-se os seguintes grupos:

| Passos $k$ | Nº de grupos $g$ | Grupos |
|---|---|---|
| 1 | $g = 7$ | $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ |
| 2 | $g = 6$ | $C_{17}$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ |
| 3 | $g = 5$ | $C_{175}$, $C_2$, $C_3$, $C_4$, $C_6$ |
| 4 | $g = 4$ | $C_{175}$, $C_2$, $C_{34}$, $C_6$ |
| 5 | $g = 3$ | $C_{1752}$, $C_{34}$, $C_6$ |
| 6 | $g = 2$ | $C_{17526}$, $C_{34}$ |
| 7 | $g = 1$ | $C_{1752634}$ |

.

**Definição 2.7.** ***- Método da Média das Distâncias (Average):*** *Neste método, os conglomerados unem-se caso a distância média seja próxima o suficiente, conforme Laureto [14].*

Esse método busca, assim como os outros, unir pela similaridade, então seleciona-se a menor distância e, na sequência, aplica-se a média nos valores dos conglomerados, ou seja,

$$C_{ij} = \sum_{l \in C_i} \sum_{k \in C_j} \left( \frac{1}{n_i n_j} \right) d(C_i, C_j)$$

em que $n_i$ e $n_j$ representam o número de elementos no $i$-ésimo e $j$-ésimo conglomerado e $d_{ij}$ é a distância entre os elementos dos conglomerados $i$ e $j$.

Dessa forma, ao realizarem-se as iterações, obtém-se os seguintes conglomerados para o exemplo 4:

| Passos $k$ | Nº de grupos $g$ | Grupos |
|---|---|---|
| 1 | $g = 7$ | $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ |
| 2 | $g = 6$ | $C_{17}$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ |
| 3 | $g = 5$ | $C_{175}$, $C_2$, $C_3$, $C_4$, $C_6$ |
| 4 | $g = 4$ | $C_{175}$, $C_2$, $C_{34}$, $C_6$ |
| 5 | $g = 3$ | $C_{1752}$, $C_{34}$, $C_6$ |
| 6 | $g = 2$ | $C_{17526}$, $C_{34}$ |
| 7 | $g = 1$ | $C_{1752634}$ |

.

**Definição 2.8.** ***- Método Centróide (Centroid):*** *definido como a distância entre a média de dois conglomerados, conforme Mingoti [17].*

Esse método busca determinar as médias dos conglomerados e então calcular sua distância como $C_{ij} = d(C_i, C_j) = (\overline{C}_i - \overline{C}_j)^t(\overline{C}_i - \overline{C}_j)$, em que $\overline{C}_l$ é a média dos conglomerados, com $l = i, j$ e $i \neq j$.

Assim, ao aplicar o método Centróide ao exemplo 4, obtêm-se os seguintes conglomerados:

| Passos $k$ | Nº de grupos $g$ | Grupos |
|---|---|---|
| 1 | $g = 7$ | $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ |
| 2 | $g = 6$ | $C_{17}$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ |
| 3 | $g = 5$ | $C_{175}$, $C_2$, $C_3$, $C_4$, $C_6$ |
| 4 | $g = 4$ | $C_{1752}$, $C_3$, $C_4$, $C_6$ |
| 5 | $g = 3$ | $C_{1752}$, $C_{34}$, $C_6$ |
| 6 | $g = 2$ | $C_{17526}$, $C_{34}$ |
| 7 | $g = 1$ | $C_{1752634}$ |

**Definição 2.9.** ***- Método de Ward:*** *método de agrupamento baseado na mínima variância, segundo Mingoti [17].*

No primeiro estágio, cada observação é considerada um *cluster* de tamanho unitário, totalizando $n$ grupos, e ao final do processo tem-se um único *cluster* com todas as observações. O número de grupos desejado, descrito por $g$, representa a divisão natural das observações e, portanto, $1 < g < n$, Mingoti e Johnson [12, 17].

Em cada passo, o algoritmo de Ward combina os dois *clusters* que resultam na menor soma de quadrado residual, definida como:

$$SSR = \sum_{i=1}^{g_k} SS_i,$$

sendo $g_k$ o número de grupos no passo $k$, $SS_i = \sum_{j=1}^{n_i} \left(c_{ij} - \overline{\mathbf{C}}_{i\cdot}\right)^T \left(c_{ij} - \overline{\mathbf{C}}_{i\cdot}\right)$ a soma de quadrados do $i$-ésimo *cluster* no passo $k$, $\mathbf{c}_{ij}$ a $j$-ésima observação do $i$-ésimo *cluster* e $\overline{\boldsymbol{C}}_{i\cdot}$ a média do $i$-ésimo *cluster*. Ou então, calcula-se a distância dos conglomerados definida como:

$$d_{i,j} = \left[\frac{n_i n_j}{n_i + n_j}\right] \left(\bar{\mathbf{C}}_{i\cdot} - \bar{\mathbf{C}}_{j\cdot}\right)^T \left(\bar{\mathbf{C}}_{i\cdot} - \bar{\mathbf{C}}_{j\cdot}\right),$$

em que $\bar{\mathbf{C}}_{i\cdot}$ e $\bar{\mathbf{C}}_{j\cdot}$ são as médias e $n_i$ e $n_j$ são os tamanhos dos $i$-ésimos e $j$-ésimos *clusters*, respectivamente.

Aplicando-se o método ao exemplo 4, obtêm-se os seguintes conglomerados:

| Passos $k$ | Nº de grupos $g$ | Grupos |
|---|---|---|
| 1 | $g = 7$ | $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ |
| 2 | $g = 6$ | $C_{17}$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$ |
| 3 | $g = 5$ | $C_{175}$, $C_2$, $C_3$, $C_4$, $C_6$ |
| 4 | $g = 4$ | $C_{175}$, $C_2$, $C_{34}$, $C_6$ |
| 5 | $g = 3$ | $C_{1752}$, $C_{34}$, $C_6$ |
| 6 | $g = 2$ | $C_{1752}$, $C_{346}$ |
| 7 | $g = 1$ | $C_{1752346}$ |

Observando-se os conglomerados formados pelos métodos apresentados, nota-se que o método de Ward tende a produzir grupos com o mesmo número de variáveis e menor variância entre eles, conforme Mingoti [17].

Mingoti [17], ainda, explica que o método da ligação simples, completa e média das distâncias podem ser utilizados quando apresentam-se variáveis quantitativas e qualitativas. Já o método centróide e Ward são utilizados apenas para variáveis quantitativas, uma vez que necessita do calculo das médias.

## 2.7 DETERMINAÇÃO DO NÚMERO $g$ DE *clusters*

Existem vários métodos para determinar o número de grupos $g$, ficando a critério do pesquisador definir qual usará. Alguns tratam-se de análises do comportamento da variável em relação a distâncias ou similaridade (análise do comportamento do nível de distância, análise do comportamento do nível de similaridade). Outros são estatísticas que determinam o quão perto o valor observado está de sua referência (coeficiente $R^2$, estatística pseudo-F, correlação semiparcial, estatística pseudo $T^2$ e estatística CCC).

O método adotado será o comportamento do nível de distâncias (fusão) sendo considerado por Hair [18] um dos mais simples e apresenta uma boa precisão.

### 2.7.1 Comportamento do nível de fusão

A análise de comportamento do nível de fusão, segundo Mingoti [17] estuda o comportamento dos *clusters* nos $k$ passos, sendo que do passo $k$ para o $k + 1$ a similaridade dos conglomerados diminuí.

Assim, para visualizar as distâncias, plota-se um gráfico (passos) $\times$ (medida de dissimilaridade), então para obter-se o número $g$ de grupos utiliza-se:

$$g = n - k,$$

sendo k o valor associado ao maior salto do nível de fusão e $n$ o tamanho da amostra. A Figura 2.6 apresenta um exemplo do tipo $d_{q,r} \times G$.

**Exemplo 6.** A base, aqui, foi o exemplo 4, apresentado em técnicas hierárquicas aglomerativas.

Ao gerar essas informações em um gráfico do tipo $d_{q,r} \times G$, obtém-se a Figura 2.6.

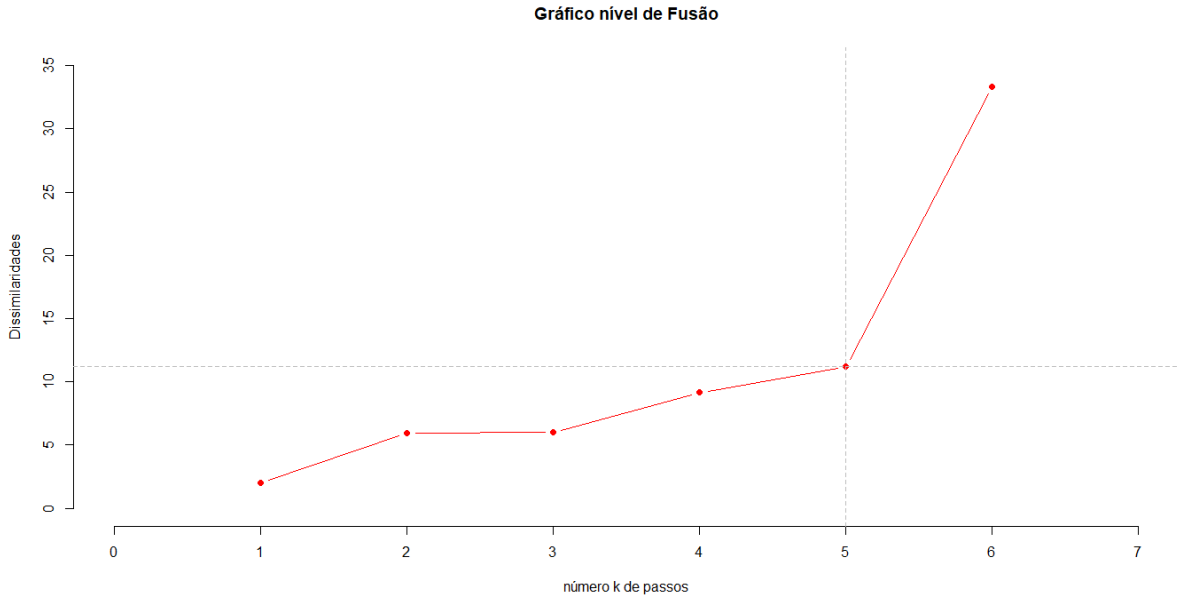



Figura 2.6: Gráfico do comportamento do nível de fusão
Fonte: Os autores, 2021

Ao observar o gráfico, nota-se que o primeiro salto significante encontra-se entre os passos $k = 5$ e $k = 6$. Assim, selecionado $k = 5$ e tendo $n = 7$, tem-se

$$\begin{aligned} g &= n - k \\ g &= 7 - 5 \\ g &= 2. \end{aligned}$$

## 2.8 DENDROGRAMA

Uma maneira de visualizar os resultados do agrupamento é por meio de um diagrama de árvore que exibe os agrupamentos combinados em cada passo do procedimento até que todos estejam contidos em um único *cluster*. Segundo Hair [18], o dendrograma é uma representação gráfica para o método hierárquico, de modo que apresente as medidas de dissimilaridades (ou a distância) no eixo vertical e as observações são expostas na horizontal.

A Figura 2.7 apresenta o dendrograma obtido para o exemplo 4 utilizando a técnica de agrupamento de Ward.

## 2.9 MÉTODO DE COMPARAÇÃO DE GRUPOS DE KRUSKAL WALLIS

O teste de Kruskal Wallis, descrito em 1952 no artigo "*Use of Ranks in One Criterion Variance Analysis*", é um teste não paramétrico utilizado para comparar três ou mais

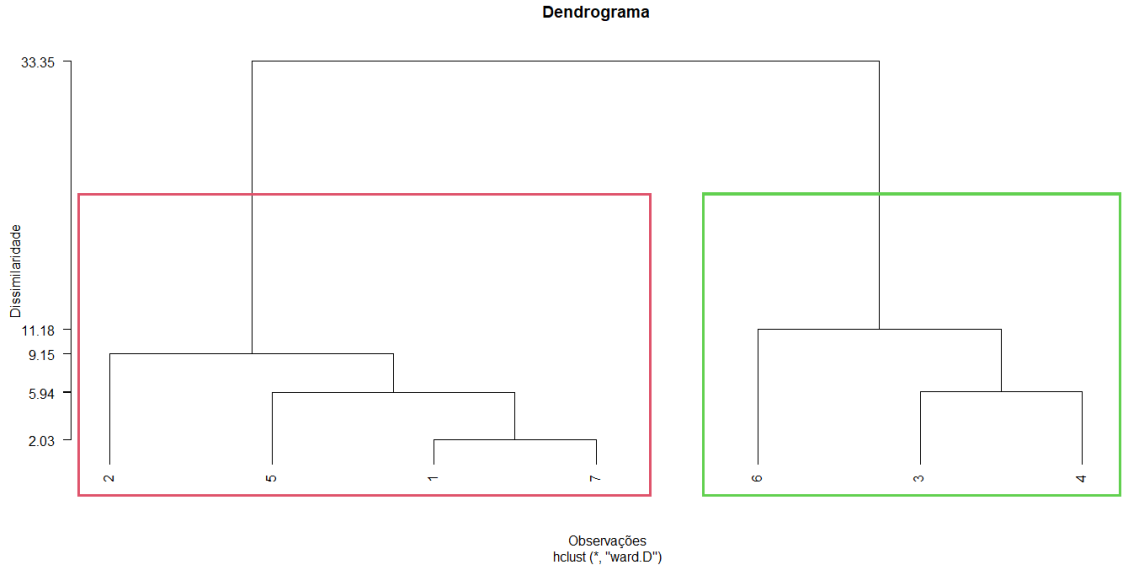



Figura 2.7: Representação do dendrograma pelo método de Ward para os dados do exemplo 4
Fonte: Os autores, 2021

grupos. Segundo Heeke [11] e Bewick [3], a técnica de Kruskal-Wallis verifica as hipóteses

$$\begin{cases} H_0 : \text{os grupos provém da mesma população} \\ H_1 : \text{os grupos são originados de populações distintas} \end{cases}.$$

Para executá-la, os valores conjuntos dos $g$ grupos devem ser ordenados e transformados em postos, atribuindo valor 1 para o menor observado, 2 para o segundo menor e assim por diante, até $N$, o maior valor observado na amostra conjunta. A estatística do teste é dada pela expressão:

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{g} n_j \bar{R}_j^2 - 3(N+1)}{1 - \sum_{i=1}^{l} (t_i^3 - t_i)/(N^3 - N)}, \tag{2.8}$$

em que $n_j$, $j = 1, 2, ..., g$, representa o número de observações do $j$-ésimo grupo, $N = \sum_{j=1}^{g} n_j$, $\bar{R}_j$ é a média dos postos no $j$-ésimo grupo, $l$ é o número de agrupamentos com postos empatados e $t_i$ é o número de empates no $i$-ésimo grupo.

Para $g > 3$, $n_j > 5$ e, supondo $H_0$ verdadeira, a estatística $H$ tem distribuição de probabilidade aproximada qui-quadrado com $k - 1$ grau de liberdade. Em todo teste de hipóteses, sempre existe um erro associado à decisão de rejeitar $H_0$, o nível de significância que, em geral, é fixado em $\alpha = 5\%$. Para se definir se será ou não rejeitado $H_0$, compara-se $\alpha$ com $p$-valor $= \mathbb{P}(H > h)$, em que $h$ é uma estimativa de (2.8). Se $p$-valor $< \alpha$, então despreza-se $H_0$; caso contrário, a hipótese nula não deve ser rejeitada.

Se a hipótese nula é rejeitada, então pelo menos um grupo difere-se dos demais, porém o teste de Kruskal-Wallis não identifica os grupos distintos. Ainda assim, pode-se

testar a diferença de grupos dois a dois, verificando-se a validade da desigualdade

$$|\bar{R}_u - \bar{R}_v| \geq z_{\alpha/g(g-1)} \sqrt{\frac{N(N+1)}{12} \left( \frac{1}{n_u} + \frac{1}{n_v} \right)} \, ,$$

em que os índices $u$ e $v$ identificam os grupos e $z_{\alpha/k(k-1)}$ é o quantil da distribuição de probabilidade Normal padrão tal que $\mathbb{P}(Z \geq z_{\alpha/k(k-1)}) = \alpha/k(k-1)$. Os resultados do teste de diferenças, em geral, são apresentados em uma tabela, cujos valores são seguidos por letras que os identificam.

# 3 MATERIAL E MÉTODOS

Este capítulo apresenta a metodologia proposta, os dados coletados e o índice de absenteísmo desenvolvido para o estudo.

## 3.1 PROPOSTA DE ANÁLISE MULTIVARIADA PARA DADOS SOBRE O ABSENTEÍSMO

Para sistematização da análise de dados multivariados sobre absenteísmo propõem-se um estudo dividido em 4 etapas, de forma a identificar fatores que possam potencializar a variação na ausência dos funcionários.

Basicamente o que se propõem é a divisão dos dados multivariados em grupos definidos pelas variáveis associadas a ausência do funcionário, como por exemplo: faltas não justificadas, atrasos, licenças médicas e etc, a análise de *clusters* é a ferramenta se indicada nessa situação. Após a definição dos grupos, um estudo estatístico utilizando apenas as variáveis associadas a aspectos sociais e/ou funcionais do funcionário pode ser realizado. Como por exemplo: pode ser feito análises de regressão, análise de comparação de grupos entre outros. Por fim, para construção de um gráfico de dispersão que informe a relação entre o primeiro conjunto de variáveis associadas a ausência do funcionário e entre o segundo conjunto associado a características sociais e/ou funcionais propõem-se a redução da dimensão dessas variáveis para apenas duas. Uma variável resultante da combinação das variáveis relacionadas a ausência e outra variável obtida da combinação dos aspectos sociais e/ou funcionaos do funcionário. O detalhamento de cada passo segue nos parágrafos seguintes:

Primeiramente, a base de dados é dividida em duas partes, a primeira parte cobre as variáveis de absenteísmo e a segunda parte diz respeito aos fatores sociais dos trabalhadores.

O segundo passo utiliza-se os dados relacionados ao absenteísmo para procurar características comuns na criação de *clusters*. Nesta fase, é utilizado o método hierárquico aglomerativo de Ward, que visa identificar grupos com um número de observações próximas e variância mínima. Por meio dos grupos, são identificados os colaboradores faltosos e uma comparação entre eles fornece informações a respeito dos maiores fatores que causam o absenteísmo.

Na sequência constrói-se as componentes principais dentro dos grupos para relacionar as variáveis relacionadas ao absenteísmo e aspecto social. Desse modo, ao estudar o gráfico de dispersão da primeira componente principal para as variáveis faltas e aspecto social é possível identificar o "comportamento"dessas variáveis.

O quarto, e último passo, é realizar um estudo descritivo dos aspectos sociais dentro de cada grupo, de modo identificar características distintas entre eles. Assim, essas características podem influenciar um aumento nos índice de absenteísmo. Dessa forma, a proposta

de análise multivariada para os dados sobre o absenteísmo, se resume em:

1. Dividir as variáveis em 2 grupos:
   (a) Primeiro grupo relacionada as ausências e
   (b) Segundo grupo sobre aspectos sociais.
2. Determinar os *clusters* a partir das variáveis relacionadas as ausências;
3. Construir as componentes principais para cada grupo e estudar o gráfico de dispersão;
4. Estudar os aspectos sociais dentro de cada grupo.

## 3.2 COLETA DE DADOS

Em 2016, foi realizado um estudo em uma companhia do ramo de transporte coletivo do estado do Paraná. A empresa originou-se no ano de 1958 e, em 2016, possuía um pouco mais de 1500 funcionários, sendo eles divididos nas mais diversas funções (motoristas, cobradores, coletores, mecânicos, entre outras).

A amostra conta com 82 observações (funcionários) e 13 variáveis identificadas, juntamente com o departamento de recursos humanos (RH). As variáveis coletadas são: faltas justificadas*, não justificadas*, atraso*, suspensão*, licença*, função, sexo, tempo, estado civil, idade, instrução, pensão e distância.

Essas variáveis foram divididas em dois grupos, sendo no primeiro apresentadas aquelas relacionadas ao perfil do colaborador: função, sexo, tempo, idade, instrução, pensão e distância. Já o segundo grupo contém as variáveis relacionadas ao absenteísmo (faltas justificadas, não justificadas, suspensão, atraso e licenças). A explicação dessas variáveis está exposta no Quadro 3.1.

### 3.2.1 Desenvolvimento do índice de absenteísmo

Este trabalho utiliza o índice de uma maneira diferente à citada por Chiavenato [6] e Penatti [19]; em vez de criar um único indicador geral para a empresa, conforme exposto nas equações (2.2) e (2.1), utiliza-se um índice para cada funcionário dessa forma, tornando-se possível encontrar funcionários que apresentem maiores índices. Na Equação (3.1) é apresentado o índice de absenteísmo utilizado.

$$\text{Índice}_i = \frac{\sum_{i=1}^{n} \big(\text{variáveis relacionadas à ausência}\big)_i}{\text{Tempo na empresa}} \tag{3.1}$$

Quadro 3.1: Apresenta o nome das variáveis coletadas durante o estudo na empresa e suas explicações.

| | |
|---|---|
| **Variáveis relacionadas ao absenteísmo** | |
| Justificadas$_i^*$: | Faltas justificadas por motivo de saúde ou lei |
| Não justificadas$^*$: | Faltas não justificadas por motivo de saúde ou lei |
| Atrasos$^*$: | Tempo de atraso contado a partir de seu horário inicial de serviço |
| Suspensão$^*$: | Tempo em que foi impedido de trabalhar pela empresa, por penalidade |
| Licença$^*$: | Permissões concedidas pela empresa ao colaborador |
| **Variáveis relacionada ao perfil dos colaboradores** | |
| Função: | Área de atuação |
| Sexo: | Masculino ou feminino |
| Tempo: | Há quanto tempo trabalha na empresa |
| Estado civil: | Casado ou solteiro |
| Idade: | Idade em anos |
| Instrução: | Grau de escolaridade |
| Pensão: | Se paga ou não pensão |
| Distância: | Distância em quilômetros entre sua casa e a empresa |

Fonte: Os autores, 2021.

O índice de absenteísmo, ainda pode ser obtido por meio da equação a seguir:

$$\text{Índice}_{\text{emp}} = \frac{\sum_{i=1}^{n} \text{Índice}_i}{n},$$

em que $n$ é o número de observações. Assim, uma variável denominada Faltas$^*$, que resume as informações relacionadas ao absenteísmo, foi criada a partir do cálculo da seguinte expressão:

$$\text{Faltas}_i^* = \text{Justificadas}_i^* + \text{Não justificadas}_i^* + \text{Atrasos}_i^* + \text{Suspensão}_i^* + \text{Licença}_i^*, \qquad (3.2)$$

com $i = 1,\ 2,\ \ldots, n$.

Note que a variável Faltas$^*$ equação (3.2), corresponde ao numerador do índice de absenteísmo da equação (3.1). Dessa maneira, o índice de absenteísmo é dado por:

$$\text{Índice}_i = \frac{\text{Faltas}_i^*}{\text{Tempo na empresa}}, \quad \text{com } \ i = 1,\ 2,\ \ldots, n.$$

# 4 RESULTADO E DISCUSSÕES

Este capitulo, apresenta os resultados obtidos pela aplicação da metodologia sugerida na seção 3.1. Partindo, inicialmente, por uma análise descritiva dos dados e na sequência é realizando a separação dos grupos e a análise dentro dos *clusters*.

## 4.1 ANÁLISE COM OS DADOS COMPLETOS

Após a coleta e a transformação de variável, foi realizada uma pesquisa preliminar dos dados relacionados ao perfil dos colaboradores, que evidenciam as características dos funcionários da empresa. Em sua maioria, tratam-se de homens $(96,3\%)$, com escolaridade média $(72\%)$, casados $(82,9\%)$, que não pagam pensão $(96,3\%)$, conforme apresentado no gráfico de barras (Figura 4.1). A análise exploratória ainda mostra que $63,4\%$ dos funcionários são motoristas que, em média, têm $45,3$ anos e residem a $6,2$ km da empresa.

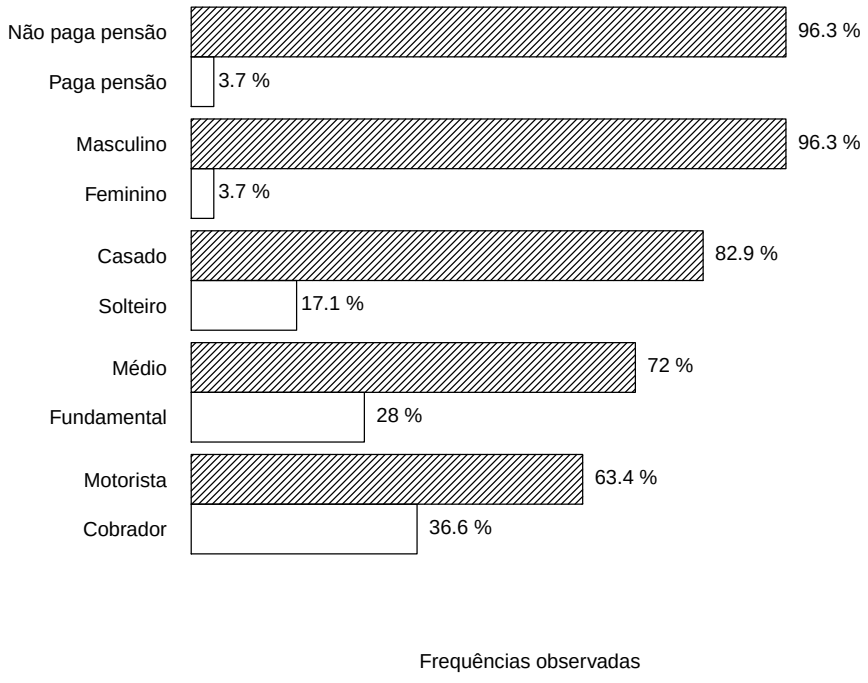



Figura 4.1: Apresenta a distribuição dos funcionários da amostra coletada.
Fonte: Os autores, 2021.

A Figura 4.2 mostra o histograma e o *boxplot* para as variáveis Idade, Distância, Faltas* e Tempo de contratação. Verifica-se que para a variável Idade, $54,88\%$ das observações têm entre $35,8$ e $53,4$ anos, enquanto que, para a Distância, a mesma proporção

fica entre $5$ e $7$ km da empresa.

Além disso, para as variáveis Faltas$^*$ e Tempo de serviço, verifica-se que $71,95\%$ dos funcionários tem até $18$ dias contínuos de faltas; considerando uma jornada de $8$ horas, essas ausências na empresa chegam a $54$ dias. A média de dias corridos ausentes é de $13$ dias ($42$ dias de uma jornada de $8$h).

Já para o tempo de contratação, $56\%$ dos funcionários estão na empresa no máximo oito anos. Entretanto, existe um grupo, correspondente a $23\%$ dos empregados, que trabalha na companhia há mais tempo, entre $15$ e $18$ anos.

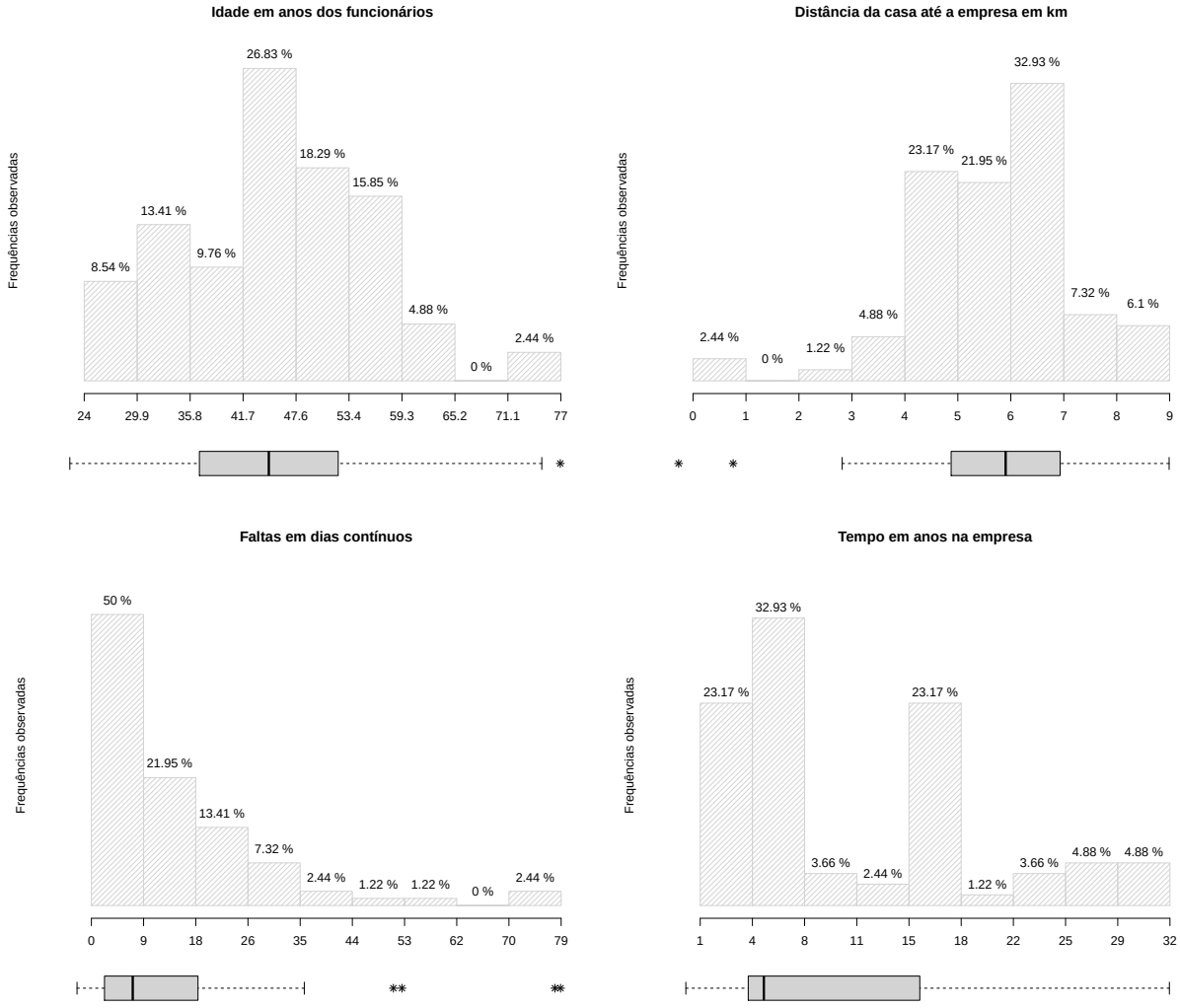



Figura 4.2: Apresenta o comportamento das variáveis Idade, Distância, Faltas$^*$ e Tempo por meio do histograma e do *boxplot*.

Fonte: Os autores, 2021.

O teste de correlação de Spearman entre as variáveis Faltas$^*$ e Tempo apresentou um $\hat{\rho} = 0,4375$ (p-valor $< 0,001$), logo, existe uma correlação moderada e positiva entre as variáveis. Esta, no entanto, era esperada, uma vez que, quanto maior o Tempo do funcionário na empresa, maior é o número de Faltas$^*$ esperado.

Assim, visando corrigir esse problema, foi novamente inserida uma variável

Falta$^*$, então normalizada pela variável Tempo, conforme apresentado no índice de absenteísmo, na Equação (4.1).

$$\begin{aligned} \text{Faltas}_i^* &= \left(\text{Justificadas}_i^* + \text{Não justificadas}_i^* + \text{Atrasos}_i^* + \text{Suspensão}_i^* + \text{Licença}_i^*\right)\frac{10000}{\text{Tempo}_i} \\ \text{Faltas}_i^* &= 10000 \times \text{Índice}_i \end{aligned} \tag{4.1}$$

com $i = 1,\ 2,\ \ldots, 82$. As variáveis relacionadas a ausências também foram padronizadas pela variável Tempo, como exposto nas Equações (4.2), (4.3), (4.4), (4.5) e (4.6).

$$\text{Justificadas}_i = \left(\frac{10000}{\text{Tempo}_i}\right) \times \text{Justificadas}_i^*; \tag{4.2}$$

$$\text{Não justificadas}_i = \left(\frac{10000}{\text{Tempo}_i}\right) \times \text{Não justificadas}_i^*; \tag{4.3}$$

$$\text{Atrasos}_i = \left(\frac{10000}{\text{Tempo}_i}\right) \times \text{Atrasos}_i^*; \tag{4.4}$$

$$\text{Suspensão}_i = \left(\frac{10000}{\text{Tempo}_i}\right) \times \text{Suspensão}_i^*; \tag{4.5}$$

$$\text{Licença}_i = \left(\frac{10000}{\text{Tempo}_i}\right) \times \text{Licença}_i^*. \tag{4.6}$$

O estudo descritivo das variáveis obtidas pela padronização pelo Tempo, Tabela 4.1 e Figura 4.3, mostrou que os dados são assimétricos à direita e possuem alguns valores extremos. Apesar disto, não há evidência de serem atípicos. Ainda observa-se que as Faltas$^*$ são, causadas, em sua maioria, pelas Justificadas$^*$ e, em menor proporção, Licença.

Tabela 4.1: Estudo exploratório para as variáveis Justificadas, Não justificadas, Atrasos, Suspensão e Licença.

| Variáveis | Mínimo | Mediana | Média | Máximo | Coef. Var. (%) |
|---|---|---|---|---|---|
| Tempo (horas) | 8766 | 52596 | 95463,89 | 280512 | 75,89 |
| Faltas | 0,00 | 32,28 | 42,58 | 167,28 | 93,80 |
| Justificadas | 0,00 | 18,99 | 32,50 | 133,47 | 112,44 |
| Não justificadas | 0,00 | 1,71 | 3,46 | 31,94 | 157,89 |
| Atrasos | 0,00 | 0,00 | 0,12 | 1,13 | 189,19 |
| Suspensão | 0,00 | 0,00 | 0,42 | 4,11 | 189,78 |
| Licença | 0,00 | 3,27 | 6,09 | 82,14 | 173,42 |

Fonte: Os autores, 2021.

A Tabela 4.2 apresenta o teste de correlação de Spearman para as variáveis padronizadas, no qual verifica-se uma correlação moderada entre as faltas Justificadas e Não justificadas; ou seja, quando os funcionários apresentam uma grande quantidade de faltas Justificadas, existe uma tendência de que apresente mais faltas Não justificadas.

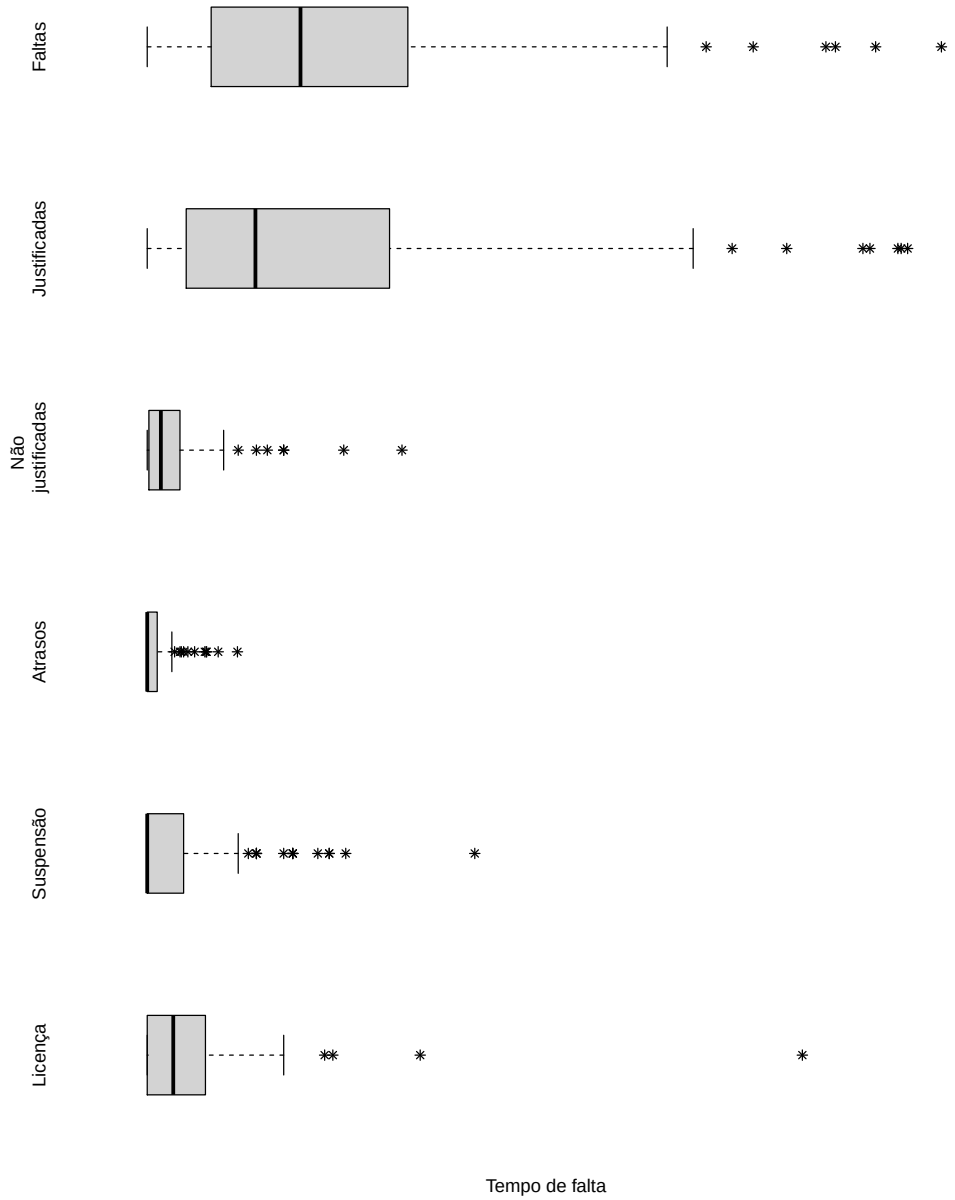



Figura 4.3: Gráfico *boxplot* para as variáveis Justificadas, Não justificadas, Atrasos, Suspensão e Licença. Cada *boxplot* está apresentado em escala diferente para fins de visualização.
Fonte: Os autores, 2021.

Outra correlação observada na Tabela 4.2 é entre faltas obtidas por Licença e Atrasos, que, assim como no caso anterior, é moderada e positiva: os colaboradores tendem a solicitar mais Licenças quando apresentam Atrasos com maior frequência.

Tabela 4.2: Correlação de Spearman entre as variáveis Justificadas, Não justificadas, Atrasos, Suspensão e Licença.

| | Não justificadas | Atrasos | Suspensão | Licença |
|---|---|---|---|---|
| Justificadas | 0,40*** | $0{,}14^{ns}$ | 0,24** | $0{,}15^{ns}$ |
| Não justificadas | | 0,25** | 0,24** | 0,31** |
| Atrasos | | | 0,21* | 0,42*** |
| Suspensão | | | | $0{,}17^{ns}$ |

$^{ns}$ p−valor $>$ 0,1; * p−valor $<$ 0,05; ** p−valor $<$ 0,01; *** p−valor $<$ 0,001

Fonte: Os autores, 2021.

Para estudar a associação das variáveis relacionadas com Faltas* (Justificadas, Não justificadas, Atrasos, Suspensão e Licença) com as variáveis de aspectos sociais (Função, Sexo, Estado civil, Idade, Instrução, Pensão e Distância), foram realizadas a análise de componentes principais e a construção do *biplot* com qualidade de representação igual para as observações e variáveis.

Assim, as componentes principais mostraram que, para as $82$ observações com as variáveis relacionadas a Faltas*, a primeira componente já seria suficiente para explicar o comportamento dos dados com uma variabilidade superior a $90\%$. E as variáveis de aspectos sociais também apresentaram uma variabilidade superior a $90\%$ para a primeira componente principal. As expressões obtidas para os componentes foram:

$$\begin{cases} CP_1 &= 0,998 \text{ Justificada } + 0,056 \text{ Não justificada } + 0,002 \text{ Atraso } + 0,005 \text{ Suspensão } - \\ & \quad -0,024 \text{ Licença} \\ CP_2 &= -0,005 \text{ Função } + 0,001 \text{ Sexo } + 0,009 \text{ Estado civil } + 1,000 \text{ Idade } - \\ & \quad -0,005 \text{ Instrução } - 0,001 \text{ Pensão } + 0,001 \text{ Distância} \end{cases},$$

sendo que a primeira componente principal para as variáveis relacionadas com ausência $CP_1$ explicou $90,7\%$ da variabilidade dos dados, enquanto a primeira componente principal para as variáveis associadas a aspectos sociais $CP_2$, explicou 97,5%. O coeficiente de maior magnitude está relacionado a faltas Justificadas para $CP_1$ e Idade para $CP_2$, indicando que esses fatores contribuem fortemente para a variação dos componentes. Pode-se interpretar a $CP_1$ como sendo um índice associado a faltas Justificadas e $CP_2$ um índice associado à idade dos funcionários, uma vez que os coeficientes restantes são proporcionalmente muito menores.

O gráfico de dispersão para as duas componentes principais é apresentado na Figura 4.4, na qual é possível verificar que existe um comportamento simultâneo para as duas componentes. Entretanto, sua obtenção pouco contribui para explicar o fenômeno do absenteísmo na empresa, uma vez que a mesma informação é obtida pelo gráfico de dispersão das variáveis Faltas Justificadas e Idade.

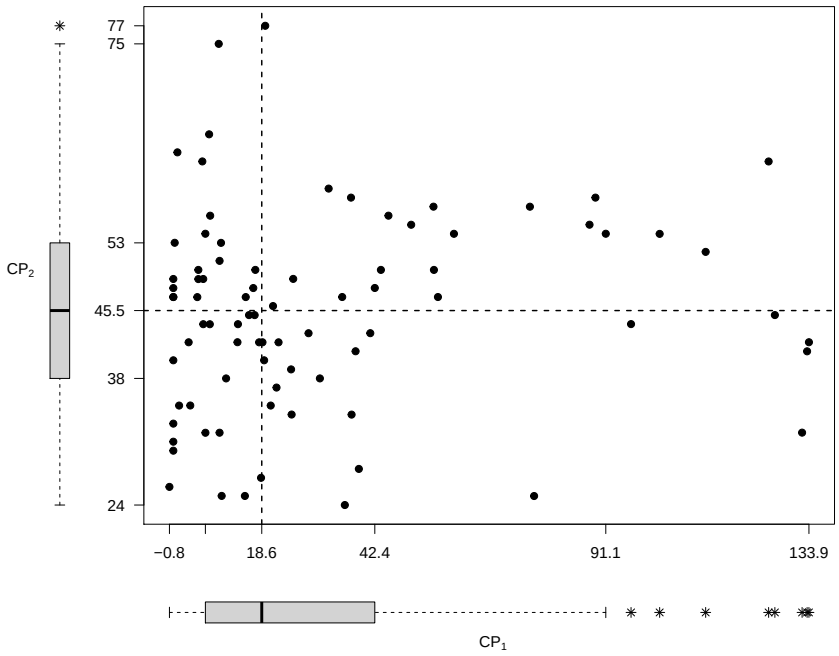



Figura 4.4: Gráfico de dispersão das componentes principais $CP_1$ e $CP_2$.
Fonte: Os autores, 2021.

Desse modo, assim como a análise de componentes principais, o *biplot* mostra que, para as variáveis associadas às ausências, o marcador que apresenta maior magnitude é o de faltas Justificadas, seguido pelo de Licença (Figura 4.5 a). O *biplot* ainda expõe a correlação entre faltas Justificadas e Licenças como sendo positiva. De maneira similar, a variável associada aos aspectos sociais pode ser explicada pela variável Idade (Figura 4.5 b).

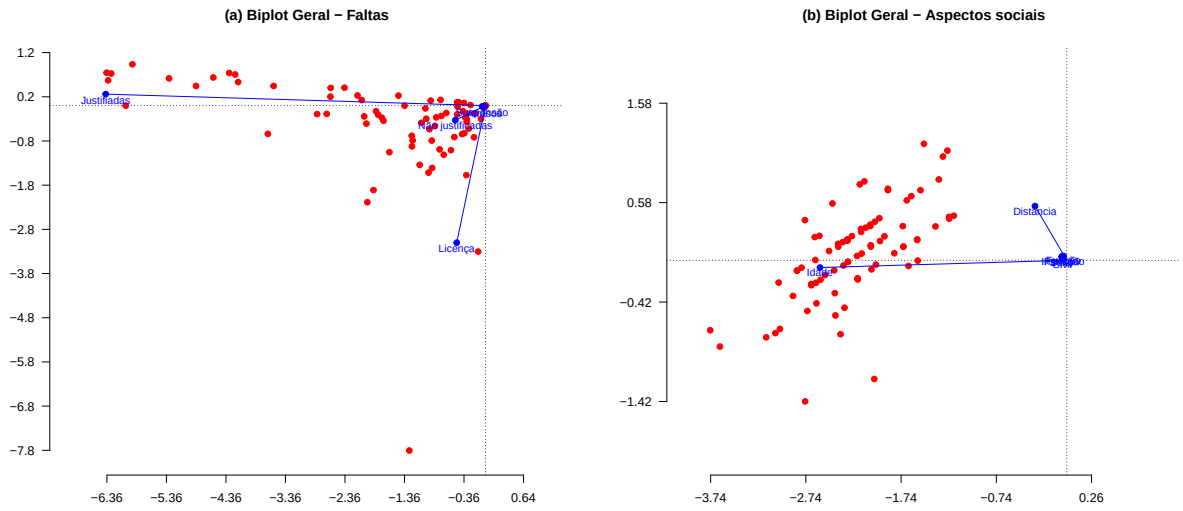



Figura 4.5: *Biplot* para as variáveis associadas a faltas e aspectos sociais.
Fonte: Os autores, 2021.

A análise com os dados completos apresentada até aqui pode não ser adequada, uma vez que não leva em consideração as variáveis em conjunto e, consequentemente, pode mascarar fatores relevantes, ou ainda indicar resultados de difícil interpretação. Além

disso, mesmo quando o conjunto das variáveis (estudo das componentes principais e *biplot*) é levado em consideração, os resultados não são satisfatórios. Isso talvez seja devido à amostra obtida de populações diversas, que mascara a informação do fenômeno sob estudo quando se analisa o conjunto completo, o que justificaria o coeficiente de variação alto.

A proposta, portanto, é utilizar as técnicas multivariadas, em particular das análises de *clusters* e de componentes principais, para separar o conjunto de dados completo em grupos mais informativos, levando-se em conta fatores associados à ausência dos trabalhadores.

Para a determinação dos grupos, foi considerado o algoritmo hierárquico aglomerativo de Ward aplicado aos dados relacionados as faltas, apresentado na seção anterior, e o número de grupos $g$ foi determinado por meio do comportamento do nível de fusão em cada passo do algoritmo de Ward.

O gráfico do nível de fusão e o dendrograma são expostos nas Figuras 4.6 e 4.7, respectivamente. No gráfico de nível de fusão, percebe-se salto mais acentuado do passo $79^o$ para o $80^o$ do algoritmo, indicando um grande aumento na dissimilaridade entre os grupos formados (de $237$ para $515$); portanto, o algoritmo foi finalizado no $79^o$ passo. Como o número de grupos é o complementar dos passos, obtém-se:

$$g = n - k = 82 - 79 = 3,$$

totalizando $g = 3$ grupos, daqui em diante denotados como G1, G2 e G3.

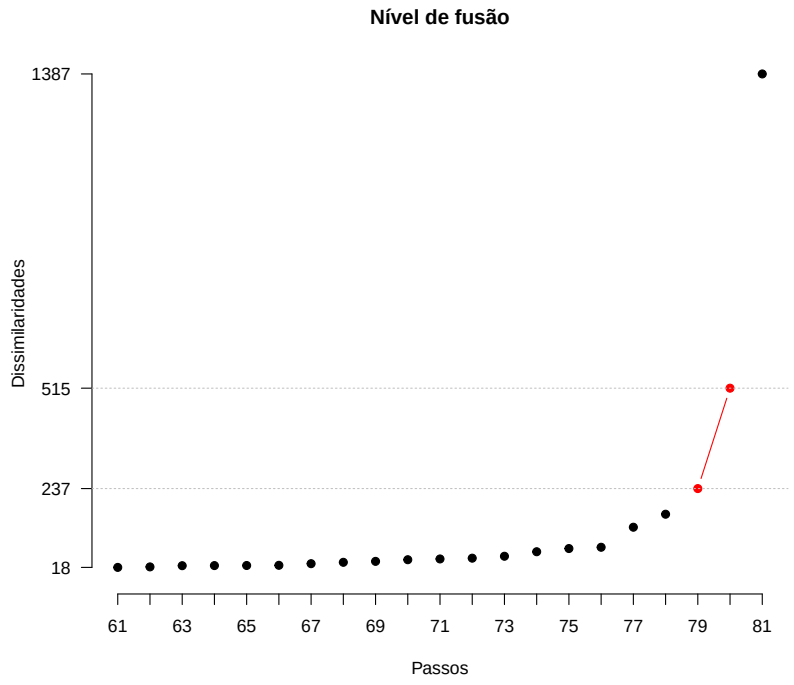



Figura 4.6: Gráfico do comportamento do nível de fusão do algoritmo de Ward.
Fonte: Os autores, 2021.

No dendrograma (Figura 4.7), estão em destaque os três grupos formados.

Sendo que $G1$, em azul, contém $53$ observações, $G2$ (grupo em verde) apresenta $16$ e o grupo $G3$, em vermelho, com apenas $13$ observações. Ainda, no dendrograma, nota-se que o corte dos grupos foi realizado entre o salto da dissimilaridade de $237$ para $515$, conforme apresentado na Figura 4.6.

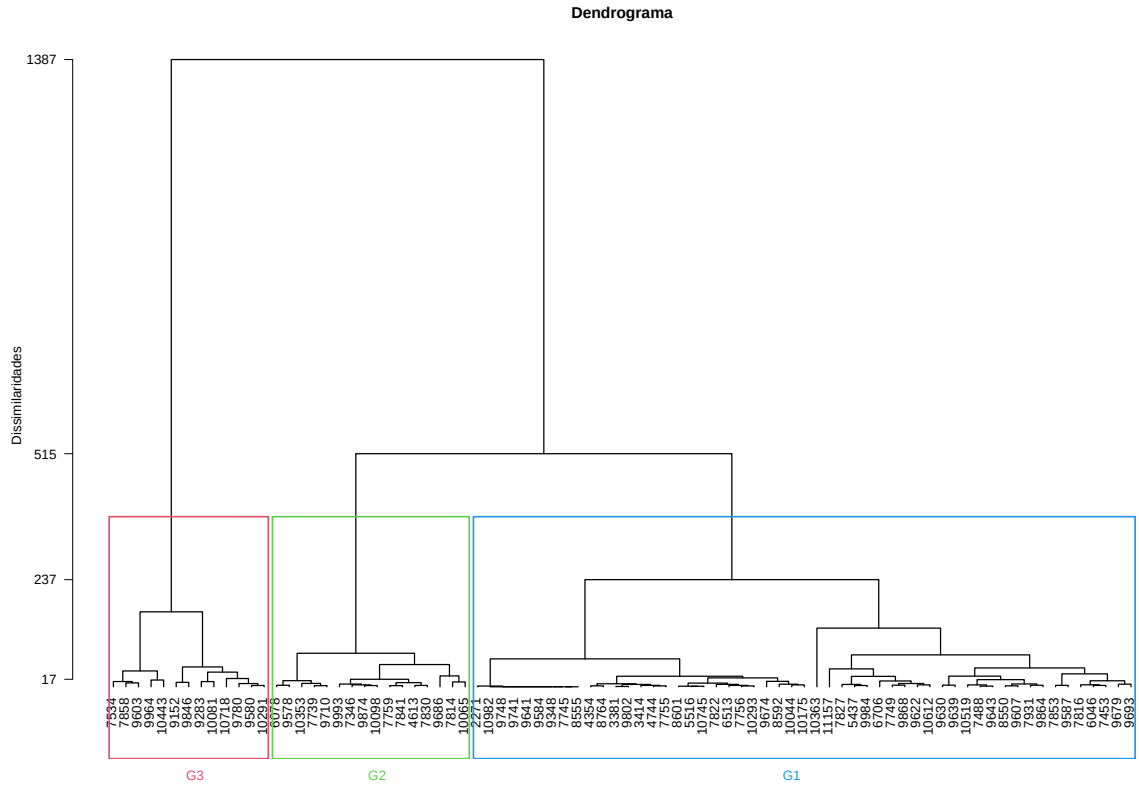



Figura 4.7: Diagrama dendrograma com o histórico hierárquico dos grupos formados.
Fonte: Os autores, 2021.

## 4.2 COMPARAÇÃO ENTRE OS GRUPOS FORMADOS

A comparação entre os grupos formados é demonstrada, na sequência, na Tabela 4.3, com os valores médios para as variáveis com os dados completos (n = $82$) e para os três grupos formados. A tabela também apresenta o teste de comparação não paramétrico de Kruskal-Wallis para as variáveis que diferem significativamente.

Tabela 4.3: Valores médios para as variáveis de cada grupo formado e o resultado do teste para comparação dos grupos.

| Grupos | n | Faltas (dias) | | Justificadas | | Não justificadas | |
|---|---|---|---|---|---|---|---|
| G | 82 | 14 | | 32,5 | | 3,5 | |
| G1 | 53 | 7 | c | 10,9 | c | 2,4 | b |
| G2 | 16 | 19 | b | 44,0 | b | 4,2 | ab |
| G3 | 13 | 39 | a | 106,2 | a | 6,8 | a |

Fonte: Os autores, 2021.

Desse modo, as variáveis que não apresentaram diferença entre os grupos foram retiradas da Tabela 4.3, como é o caso das variáveis Tempo de serviço, Idade dos funcionários, Distância até a empresa e ausências devido a Atraso, Suspensão e Licença. Entretanto, os grupos apresentam diferenças entre Faltas (dias), Justificadas e Não justificadas. O grupo G1, composto com $n = 53$ observações, é o que apresenta menos ausências em comparação aos demais, faltando em média 7 dias. O grupo G2 ($n = 16$) é o intermediário em faltas, e o G3 é o menor em observações, porém o mais faltoso, com 39 dias de ausências.

A variável Justificadas segue o mesmo comportamento de faltas com G1 sendo o grupo que menos falta, G2 o intermediário e G3 o que apresenta maior número de faltas Justificadas. Já para a variável Não justificadas, os grupos G3 e G1 se diferem, sendo que G3 apresenta uma maior média de faltas não explicadas, enquanto G1 possui menos ausências não justificadas. Para o caso da variável Não justificadas, o grupo G2 não se diferencia dos demais.

No geral, as Faltas são causadas pelas ausências Justificadas, sendo o grupo G3 o maior causador. Seus aspectos sociais são similares, com exceção do grupo G1, que é formado exclusivamente por funcionários com formação até o ensino médio. As Figura 4.8, 4.9 e 4.10 mostram a análise descritiva dos grupos e suas similaridades, como mencionado anteriormente.

Um estudo da correlação de Spearman dentro de cada um dos grupos (Tabelas 4.4, 4.5 e 4.6) mostrou que, para G1 e G3, as faltas Justificadas estão moderada e positivamente associadas às Não justificadas, com $\hat{\rho}_1 = 0,42$ e $\hat{\rho}_3 = 0,59$, respectivamente. Assim, é esperado que um aumento nas faltas Justificadas esteja associado ao aumento, ainda que menor, nas faltas Não justificadas para os funcionários desses grupos.

Ainda sobre o grupo G1, verifica-se que Licença foi significativa com todas as variáveis, indicando que pode ser o grande causador de ausências nesse grupo, apesar de o coeficiente de correlação ser menor que $0,5$. O grupo G2 apresentou associação moderada entre ausência devido a Atraso com faltas Não justificadas e Licenças, o que indica que funcionários que tendem a ter mais atrasos podem apresentar um número maior de faltas Não justificadas e de Licenças.

Tabela 4.4: Correlação de Spearman – Grupo G1.

| | Não justificadas | Atraso | Suspensão | Licença |
|---|---|---|---|---|
| Justificadas | 0,42** | $0,05^{ns}$ | 0,24* | 0,44** |
| Não justificadas | | $0,09^{ns}$ | 0,25* | 0,35** |
| Atraso | | | $0,11^{ns}$ | 0,43** |
| Suspensao | | | | 0,34* |

$^{ns}$ p−valor $>$ 0,1; * p−valor $<$ 0,1; ** p−valor $<$ 0,01; *** p−valor $<$ 0,001

Fonte: Os autores, 2021.

Por fim, foram realizadas uma análise de componentes principais obtidas pela matriz de covariância e a construção do SQRT-biplot para cada um dos grupos. O estudo ocor-

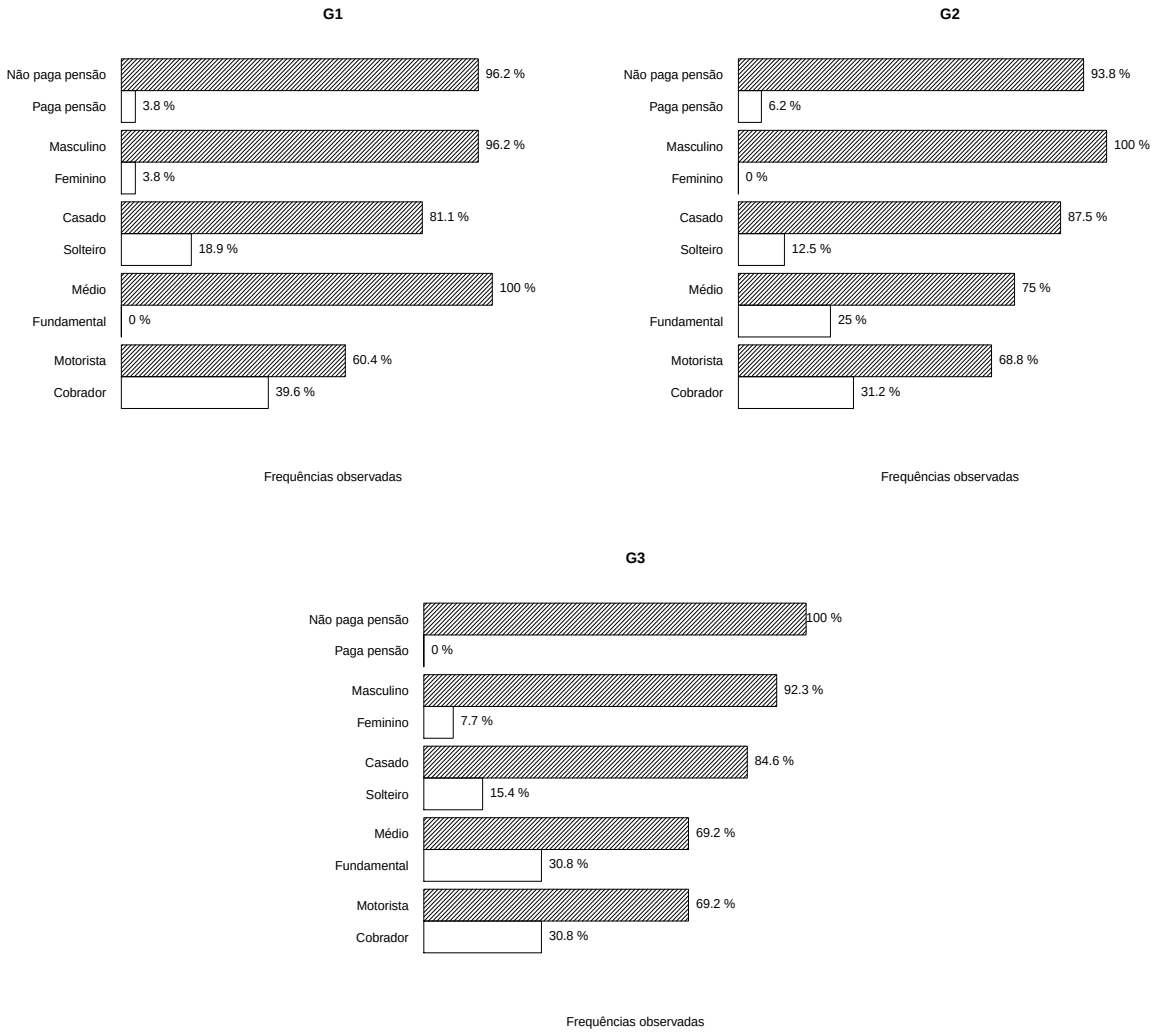



Figura 4.8: Gráfico de barras para variáveis sociais de cada grupo.
Fonte: Os autores, 2021.

Tabela 4.5: Correlação de Spearman – Grupo G2.

| | Não justificadas | Atraso | Suspensão | Licença |
|---|---|---|---|---|
| Justificadas | -0,41$^{ns}$ | -0,25$^{ns}$ | -0,07$^{ns}$ | -0,41$^{ns}$ |
| Não justificadas | | 0,51* | -0,07$^{ns}$ | 0,38$^{ns}$ |
| Atraso | | | 0,14$^{ns}$ | 0,49* |
| Suspensao | | | | -0,18$^{ns}$ |

$^{ns}$ p$-$valor $>$ 0,1; * p$-$valor $<$ 0,1; ** p$-$valor $<$ 0,01; *** p$-$valor $<$ 0,001

Fonte: Os autores, 2021.

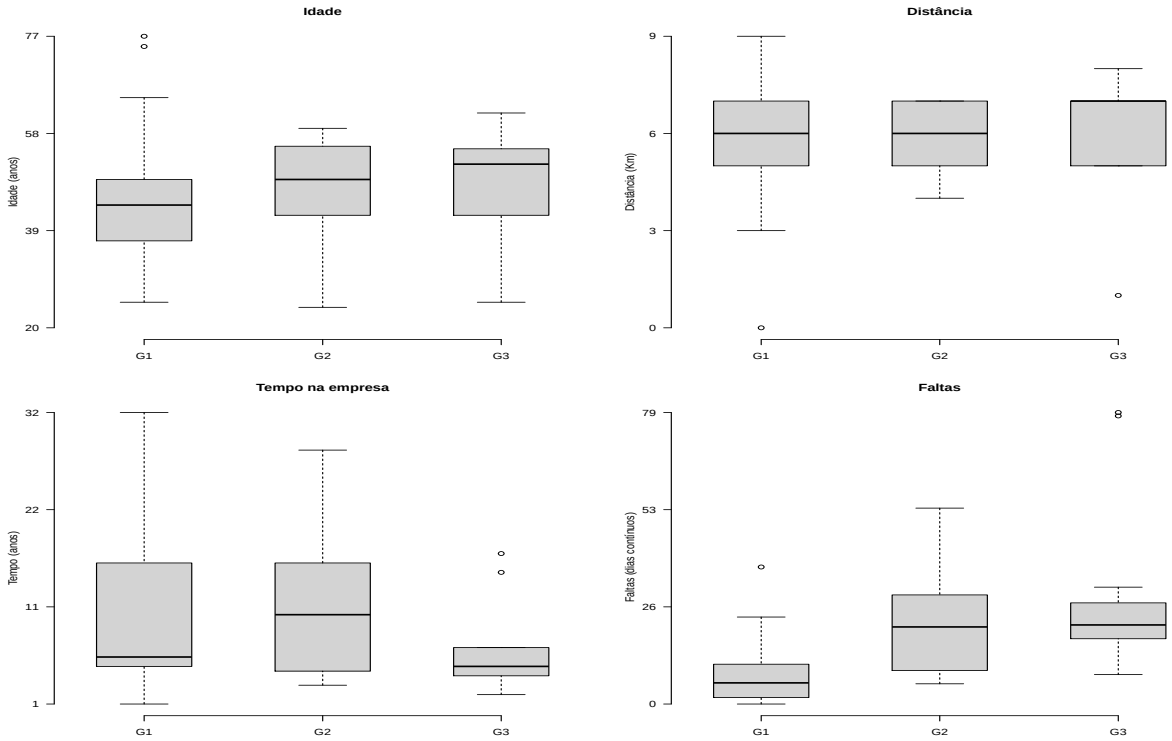



Figura 4.9: *Boxplot* para as variáveis Idade, Distância, Tempo e Faltas para cada um dos grupos.
Fonte: Os autores, 2021.

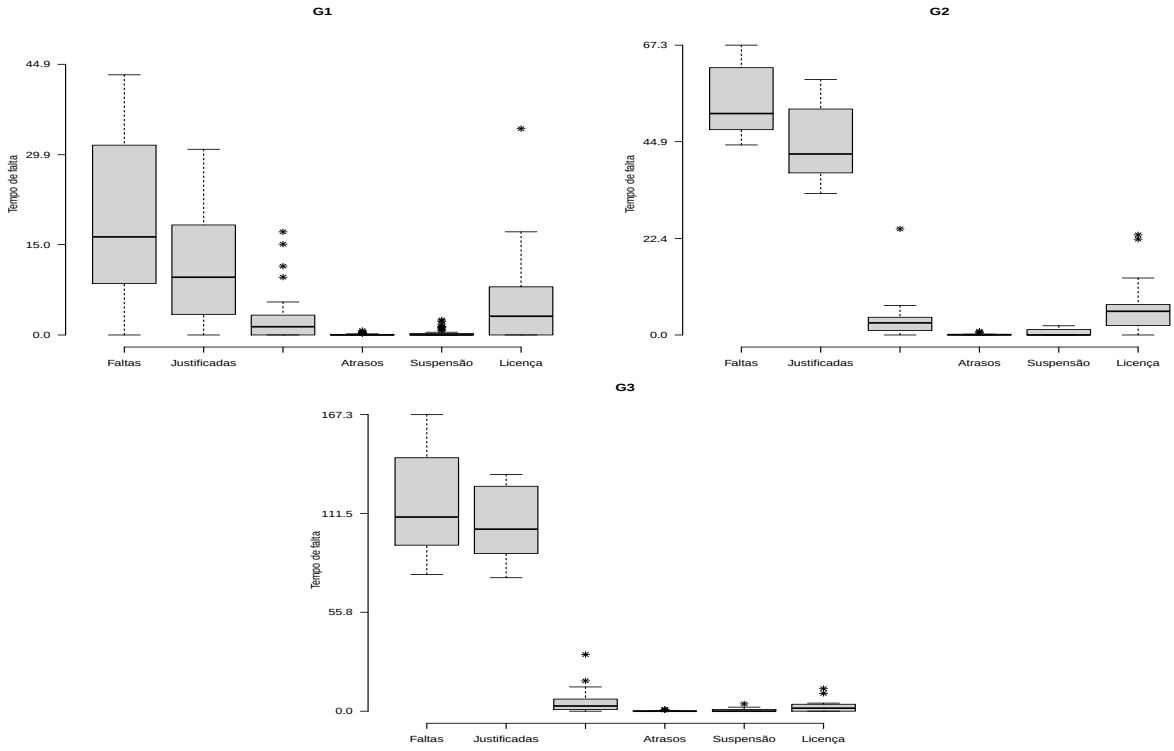



Figura 4.10: *Boxplot* para as variáveis associadas as ausências dentro de cada um dos grupos.
Fonte: Os autores, 2021.

Tabela 4.6: Correlação de Spearman – Grupo G3.

| | Não justificadas | Atraso | Suspensão | Licença |
|---|---|---|---|---|
| Justificadas | 0,59* | 0,20$^{ns}$ | 0,07$^{ns}$ | 0,08$^{ns}$ |
| Não justificadas | | 0,45$^{ns}$ | 0,43$^{ns}$ | 0,04$^{ns}$ |
| Atraso | | | 0,39$^{ns}$ | 0,39$^{ns}$ |
| Suspensão | | | | -0,10$^{ns}$ |

$^{ns}$ p−valor $>$ 0,1; * p−valor $<$ 0,1; ** p−valor $<$ 0,01; *** p−valor $<$ 0,001

Fonte: Os autores, 2021.

reu a partir das variáveis contidas nos grupos, subdivididas entre as relacionadas aos aspectos sociais e as associadas às ausências. As expressões para as componentes principais são apresentadas na sequência.

### 4.2.1 Grupo G1

$$\begin{cases} CP_1 &= 0,3 \text{ Justificadas } + 0,04 \text{ Não justificadas } + 0,0009 \text{ Atraso } + 0,005 \text{ Suspensão } + \\ & \quad +0,95 \text{ Licença} \\ CP_2 &= -0,01 \text{ Função } + 0,01 \text{ Sexo } + 0,005 \text{ Estado Civil } + 1 \text{ Idade } - 0,009 \text{ Instrução} - \\ & \quad -0,0009 \text{ Pensão } - 0,0003 \text{ Distância} \end{cases}$$

Ao serem estudadas as componentes do grupo G1, verificou-se que a variabilidade explicada dos dados é $66,2\%$ e a componente $CP_1$ é dominada por Licenças com o coeficiente de $0,95$ seguida de Justificadas com $0,3$. A componente $CP_2$, por sua vez, apresenta uma variação de $97,4\%$, tendo como maior coeficiente a variável Idade. Desse modo, o gráfico de dispersão (Figura 4.11) pode ser visto como a dispersão entre faltas 'documentadas' e Idade, na qual observa-se que funcionários com maior idade se ausentam menos que os mais jovens, devido a faltas justificadas e licenças.

Os *biplots* (Figura 4.12) mostram o mesmo comportamento exposto pelas primeiras componentes principais. Para o *biplot* associado às variáveis de Falta, destacam-se dois vetores, o primeiro relacionado às Justificadas e o segundo, às Licenças. No caso do *biplot* para aspectos sociais, ele apresenta um marcador maior que os demais, indicando que a variável mais relevante é a Idade.

Além dessas informações, o *biplot* ainda mostra a correlação entre os marcadores de faltas Justificadas e Licença por meio do cálculo do cosseno. A relação entre eles é positiva, pois não estão em posição perpendicular, nem em direções opostas.

### 4.2.2 Grupo G2

$$\begin{cases} CP_1 &= 0,7 \text{ Justificadas } - 0,04 \text{ Não justificadas } - 0,01 \text{ Atraso } + 0,003 \text{ Suspensao } - \\ & \quad -0,6 \text{ Licença} \\ CP_2 &= -0,001 \text{ Função } + 0,02 \text{ Estado civil } + 1 \text{ Idade } + 0,01 \text{ Instrução } - \\ & \quad -0,00006 \text{ Pensão } + 0,03 \text{ Distância} \end{cases}$$

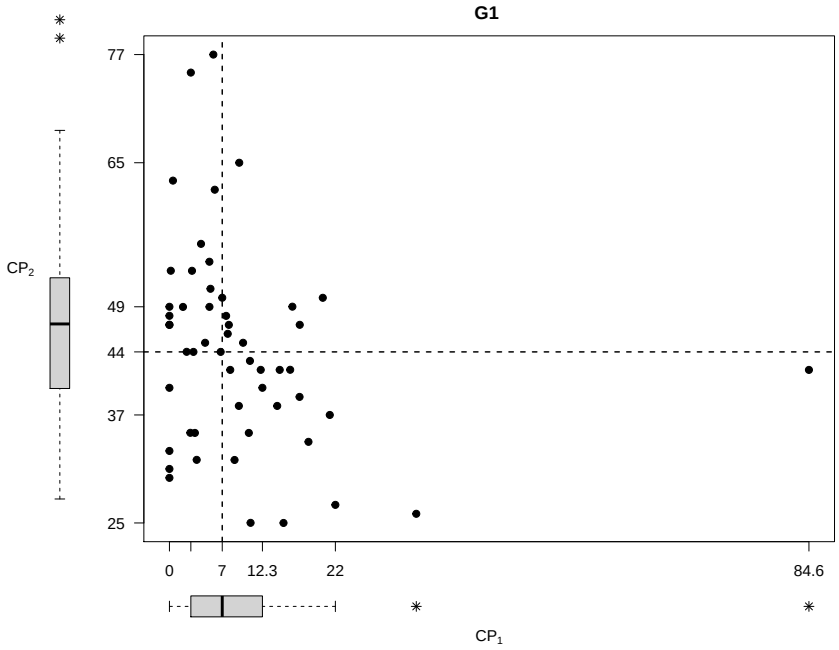



Figura 4.11: Gráfico de dispersão para as componentes $CP_1$ e $CP_2$ para o grupo G1.
Fonte: Os autores, 2021.

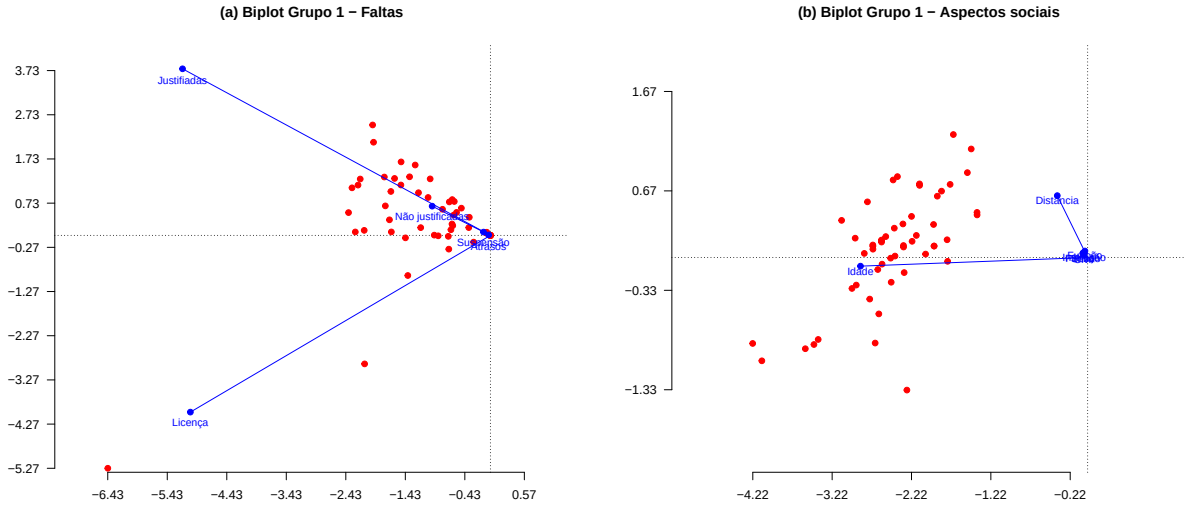



Figura 4.12: *Biplot* do grupo G1.
Fonte: Os autores, 2021.

No grupo 2, a $CP_1$ explica $65,8\%$ da variação dos dados e apresenta um contraste entre faltas Justificadas e Licenças. A $CP_2$ explica $98,6\%$ da variabilidade dos dados e é dominada pela Idade. Pelo gráfico da Figura 4.13, percebe-se prevalência de faltas Justificadas para funcionários mais velhos.

Destaca-se ainda nesse gráfico um funcionário que apresenta baixo valor para $CP_1$ e $CP_2$, ou seja, um jovem com alta frequência de faltas Justificadas e Licenças. A título de curiosidade, o colaborador tem 28 anos (a média de idade do grupo é de $46,9$ anos), está na empresa há 5 anos, é solteiro, não paga pensão, ocupa o cargo de motorista e tem 168 horas em faltas Justificadas, $108$ horas em faltas Não justificadas, $234$ horas em Atrasos e $102$ horas de Licença.

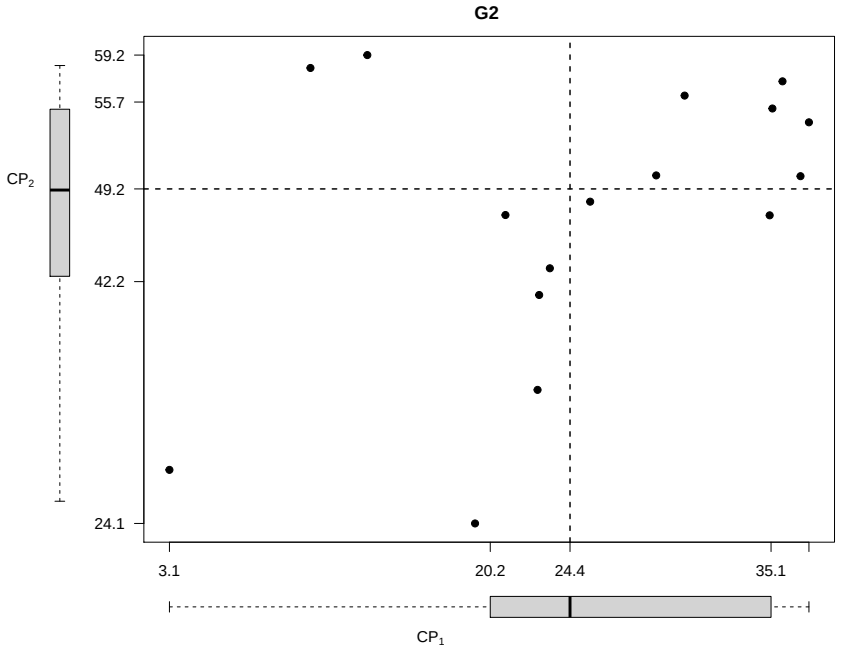



Figura 4.13: Gráfico de dispersão para as componentes $CP_1$ e $CP_2$ para o grupo G2.
Fonte: Os autores, 2021.

Os *biplots* apresentados na Figura 4.14 mostram a mesma dominância de variáveis apresentada para a primeira componente principal. Assim, para o *biplot* das variáveis relacionadas às faltas, verifica-se que o maior marcador é o das faltas Justificadas, seguido pelo de Licenças. Ele ainda apresenta uma correlação positiva entre Licença e faltas Não justificadas.

O *biplot* de aspectos sociais, por sua vez, indica que Idade é o marcador dominante para a primeira componente principal, com uma variabilidade explicada alta, marcador em paralelo ao eixo da primeira componente.

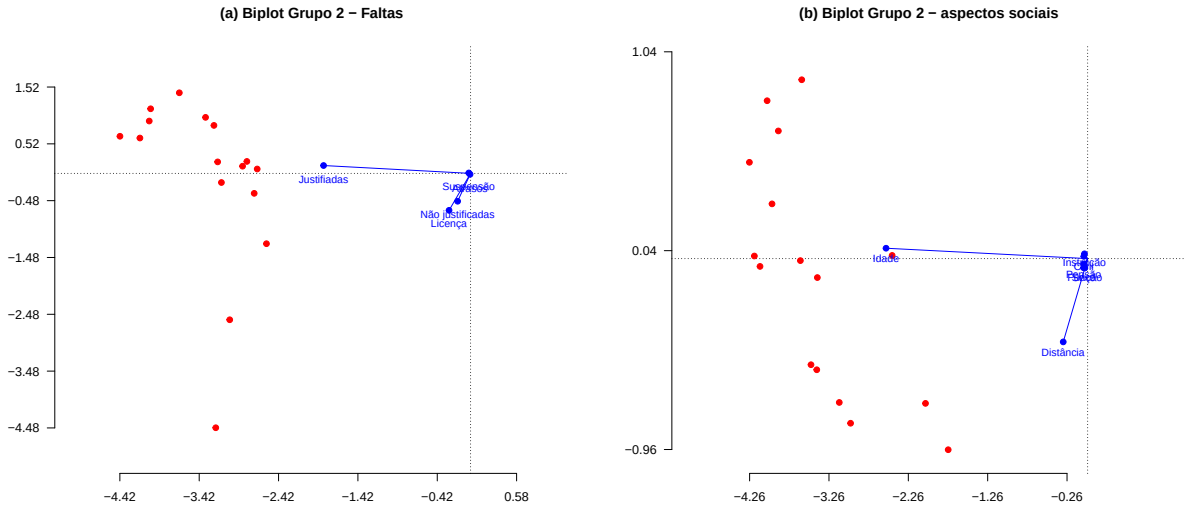



Figura 4.14: *Biplot* do grupo G2.
Fonte: Autoria dos autores

### 4.2.3 Grupo G3

$$\begin{cases} CP_1 &= 0,97 \text{ Justificadas } - 0,2 \text{ Não justificadas } - 0,005 \text{ Atraso } + 0,004 \text{ Suspensão } - \\ & \quad -0,01 \text{ Licença} \\ CP_2 &= 0,02 \text{ Função } + 0,004 \text{ Sexo } + 0,01 \text{ Estado Civil } + 1 \text{ Idade } - 0,005 \text{ Instrução } - \\ & \quad -0,02 \text{ Distância} \end{cases}$$

Por fim, para o grupo G3 a $CP_1$ explica $86,9\%$ da variação dos dados e é dominada por faltas Justificadas e Não justificadas, enquanto a $CP_2$ é novamente dominada pela Idade. No gráfico para $CP_1$ versus $CP_2$ (Figura 4.15), é possível observar que pessoas de mais idade ausentam-se menos por faltas Justificadas e Não justificadas que funcionários mais jovens.

Assim como nos demais grupos, a Figura 4.16 indica que os aspectos sociais do G3 são dominados pela Idade na primeira componente principal. Seu *biplot* de faltas apresenta dois marcadores destacados, sendo faltas Justificadas e Não justificadas; o maior marcador é mais importante para a primeira componente principal que o segundo, estando ambos ainda correlacionados de maneira positiva.

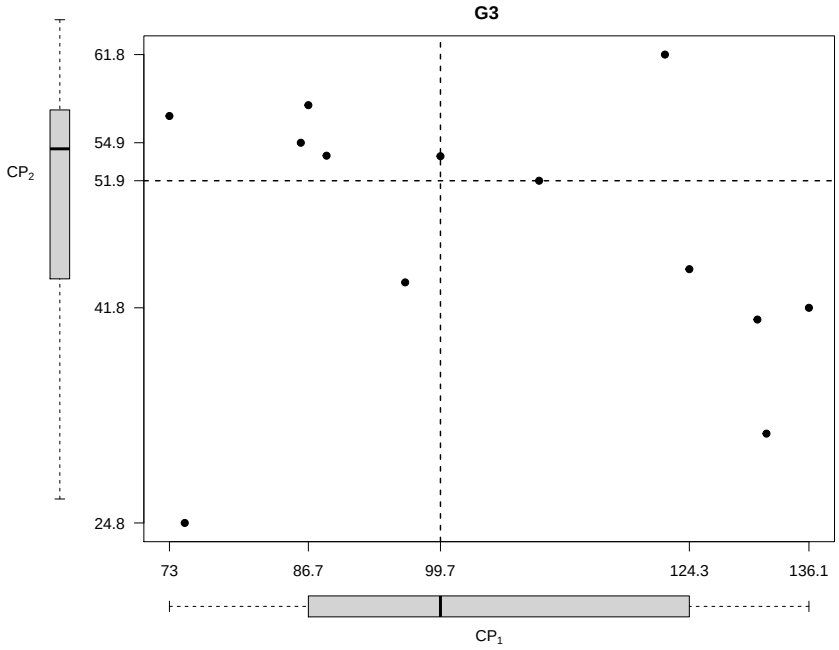



Figura 4.15: Gráfico de dispersão para as componentes $CP_1$ e $CP_2$ para o grupo G3.
Fonte: Os autores, 2021.

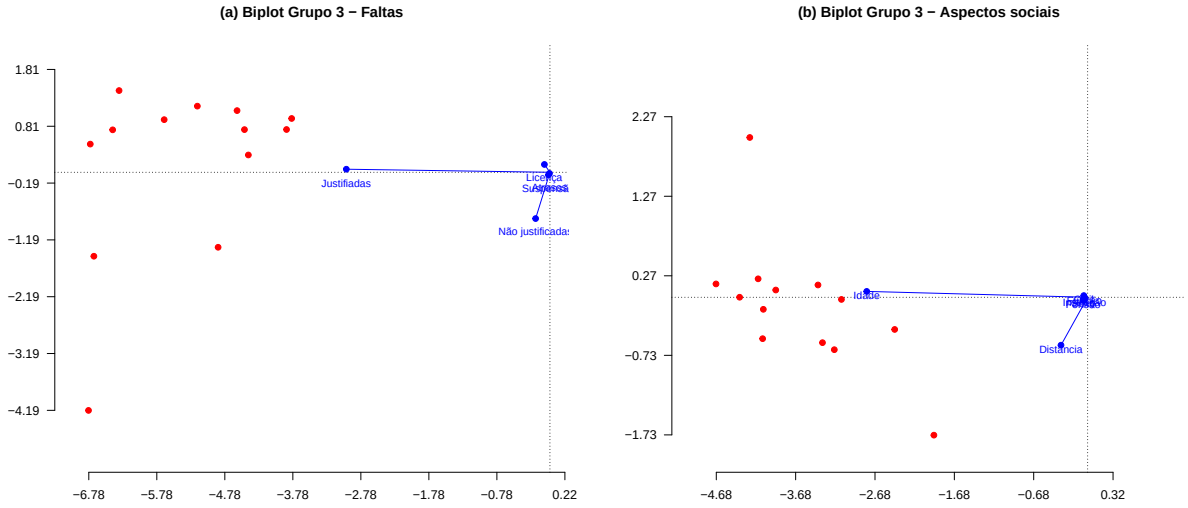



Figura 4.16: *Biplot* do grupo G3.
Fonte: Os autores, 2021.

# 5 CONCLUSÃO

Neste estudo, foi realizada uma análise do absenteísmo em uma empresa de transporte coletivo da região de Londrina, buscando classificar o perfil de ausências dos funcionários. Para isso, a pesquisa foi dividida em duas partes, sendo a primeira realizada com as $82$ observações coletadas e a segunda, com grupos determinados por meio da análise de *clusters*.

A primeira parte do estudo indicou que a maior causadora de ausências na empresa são as faltas justificadas, que apresentam uma correlação moderada com as faltas não justificadas. O aspecto social que apresenta maior importância é a idade, porém tal informação não se mostrou promissora para caracterizar faltosos. Desse modo, utilizou-se uma análise de *clusters* para determinar a existência de três grupos distintos, sendo estes G1, G2 e G3.

Com a organização dos grupos, obtiveram-se os seguintes resultados: G1 apresentou menos faltas em relação aos demais, sendo G2 o intermediário e G3 o detentor da maior frequência de faltas. Já em relação aos aspectos sociais, os grupos apresentaram características similares, com exceção do aspecto de formação escolar de G1, cujos funcionários possuem ensino médio completo.

Um fato curioso do grupo G2 é que apenas ele correlaciona negativamente as variáveis faltas não justificadas e licenças, o que pode indicar que funcionários com uma frequência maior de faltas não justificadas apresentam um menor número de licenças.

A análise de componentes principais e *biplots* ainda evidenciou os fatores que contribuem de maneira positiva ou negativa para as faltas na empresa, sendo a idade o principal fator no aspecto social para as faltas. Relacionado às ausências, verifica-se que os grupos G1 e G2 estão associados a justificadas e licença, enquanto o grupo G3 associa-se a justificadas e não justificadas.

Devido à pandemia da covid-19, não foi possível realizar uma análise mais complexa com outro banco de dados. Desse modo, é interessante, para trabalhos futuros, um maior número de variáveis explicativas a respeito da ausência, como, por exemplo, número de filhos, histórico de doenças, região habitacional, motivação e turno, para estudar o motivo pelo qual esses grupos se ausentam e tentar corrigi-los.

# REFERÊNCIAS


[1] ASSOCIAÇÃO BRASILEIRA DE CONTROLE DE QUALIDADE *Indicadores, Objetivos e Metas para Qualidade*. https://www.abcq.com.br/p/13/indicadores-objetivos-e-metas-para-qualidade.

[2] ALMEIDA, D. R., NASCIMENTO, I. G., NETO, J. M. S., AND ALMEIDA, A. G. B. *Causas e desvantagens do absenteísmo: O caso da empresa Auto Center 24 horas em Porto Velho*, 2015 (acessado 14-10-2019). http://www.inovarse.org/sites/default/files/T_15_497_6.pdf.

[3] BEWICK, V., CHEEK, L., AND BALL, J. Statistics review 10: Further nonparametric methods. *Crit Care 8* (2004), 196.

[4] CALAIS, S. L., AND ZANELATO, L. S. *Manejo de estresse e outros fatores em diferentes populações adultas*, 2015 (acessado 16-06-2016). http://books.scielo.org/id/sb6rs/pdf/valle-9788579831195-12.pdf.

[5] CHATFIELD, C. *Introduction to multivariate analysis*. Routledge, 2018.

[6] CHIAVENATO, I. *Gestão de pessoas: o novo papel dos recursos humanos nas organizações*, 3. ed. Elsevier, Rio de Janeiro, 2010.

[7] DOS SANTOS, E. C. *Mineração de dados usando álgebra linear para predição de alvos drogáveis*. PhD thesis, Belo Horizonte - MG, 2012.

[8] FERREIRA, D. F. *Estatástica multivariada*, 1. ed. UFLA, Lavras, 2008.

[9] FERREIRA, E. S. *Métodos Biplot aplicados a dados de biologia molecular*. PhD thesis, Universidade de Aveiro, 2010.

[10] FERREIRA, H. A. *Componentes de efeitos de safras representados em biplota corrigidos por predições de modelos GEE na classificação granulométrica de cafés*. PhD thesis, Universidade federal de Lavras, 2019.

[11] HECKE, T. V. Power study of anova versus kruskal-wallis test. *Journal of statistics and management systems 15(2–3)* (2012), 241–247.

[12] JOHNSON, R. A., AND WICHERN, D. W. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, 2007.

[13] LANDGRAF, R. V. *Fatores que impactam no absenteísmo dos empregados em dois setores de uma industria de alimentos - PB*. PhD thesis, João Pessoa, 2016.

[14] LAURETO, M. *Análise de Agrupamentos (Clusters)*, 2017(Acessado em 11/05/2018 às 21:13min). http://www.each.usp.br/lauretto/cursoR2017/04-AnaliseCluster.pdf.

[15] LEE, J. B., AND ERIKSEN, L. R. The effects of a policy change on three types of absence. *JONA 20* (1990), 37–40.

[16] MANLY, B. J. F. *Métodos estatísticos Multivariados: uma introdução*, 3 ed. Bookman, Porto Alegre, 2008.

[17] MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada*. Editora UFMG, Belo Horizonte, 2005.

[18] HAIR, J. F. *Análise multivariada de dados*. Bookman, Porto Alegre, 2009.

[19] PENATTI, IZIDRO; QUELHAS, O., AND ZAGO, J. S. A. *Absenteísmo: As consequências na gestão de pessoas*, 2006 (acessado 16-06-2016). http://www.aedb.br/seget/arquivos/artigos06/898_Seget_Izidro%20Penatti.pdf.

[20] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. https://www.R-project.org/.

[21] SALOMÃO, K. *Faltas, atrasos e trânsito: um ralo que custa milhões às grandes empresas. Revista Exame*. https://exame.com/negocios/faltas-atrasos-e-transito-um-ralo-que-custa-milhoes-as-grandes-empresas.

[22] 3º BENCHMARKING PARANAENSE DE RECURSOS HUMANOS 2011 DADOS 2010. https://www.indicadoresrh.com.br/benchmarking/download/?tipo=biblioteca&id=73&file=3748dc4a51bb3f29fae2c93701059baaa159b4a1.

[23] SILVA, M. M. *Absenteísmo: consequências e impactos na gestão de pessoas. Revista On-Line IPOG Especialize, Goiânia - GO v.1, n.7* (2014).

[24] SILVA, R. M. *Absenteísmo-doença entre profissionais de enfermagem atuantes na urgência e emergência:*. PhD thesis, Belo Horizonte - MG, 2018.

[25] SOUZA, E. C. *Os métodos Biplot e escalonamento multidimensional nos delineamentos experimentais*. PhD thesis, Esalq, 2010.

[26] TEAM, R. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2018. http://www.rstudio.com/.

# A CÓDIGO UTILIZADO NO SOFTWARE R

```
# Pacotes ----
library(agricolae)


# Entrada de dados ----
dados.ori<-read.table(file='dados.csv', header=TRUE,
sep=';', dec=',', na.strings='NA')
dados.ori<-na.omit(dados.ori)
dados.ori<-dados.ori[-c(1,2,3),] #Outliers tempo de serviço
dados.ori<-dados.ori[dados.ori$ï..COD!=9855,] #Outliers
dados.ori<-dados.ori[dados.ori$ï..COD!=9962,] #Outliers
dados.ori<-dados.ori[dados.ori$ï..COD!=9683,] #Outliers
dados.ori<-dados.ori[dados.ori$ï..COD!=11267,] #Outliers
dados.ori<-dados.ori[dados.ori$ï..COD!=11284,] #Outliers
attach(dados.ori)

# Conjuto de dados ---
total<-FALTAS.ATESTADAS.EM.HORAS + FALTAS.AVULSAS.EM.HORAS +
ATRASOS.EM.HORAS + SUSPENCAO + QTS.HORAS.DE.LICENCA
Indice<-total/TEMPO.DE.SERVICO.EM.HORAS

temp=10000*(cbind(FALTAS.ATESTADAS.EM.HORAS,
FALTAS.AVULSAS.EM.HORAS,
ATRASOS.EM.HORAS,
SUSPENCAO,
QTS.HORAS.DE.LICENCA)/TEMPO.DE.SERVICO.EM.HORAS)

dados_social<-dados.ori[,c(2,3,5,6,8,9,10)]
dados_social$FUNCAO.OPERACIONAL<-dados_social[,1]+1   #Motorista=1
dados_social[dados_social$FUNCAO.OPERACIONAL==2,1]<-0 #Cobrador=0
dados_social$PAGA.PENSAO<-dados_social[,6]+1   #Paga pensao=0
dados_social[dados_social$PAGA.PENSAO==2,6]<-0 #Nao paga pensao=1

dados<-cbind(temp,dados_social)
rownames(dados)<-dados.ori[,1]
dados.porc<-cbind(dados.ori[,1],(Indice)*10000,TEMPO.DE.SERVICO.EM.HORAS,
                  temp)
detach(dados.ori)
colnames(dados.porc)<-c('COD','Faltas','Tempo', 'Justificadas',
'Não justificadas', 'Atraso','suspensao','Licenca')
dados.porc<-na.omit(dados.porc)
dados.porc <- as.data.frame(dados.porc)
attach(dados.porc)
```

```
dados_aus<-dados.porc
colnames(dados.porc) <- c('COD','Faltas','Tempo', 'Justificadas',
'Não justificadas', 'Atraso','suspensao','Licenca')
detach(dados.porc)
rm(total,Indice,dados.porc)
dados_aus <- as.data.frame(dados_aus)
attach(dados_aus)

# Análise exploratória dados de ausências ----
bp<-boxplot(dados_aus[,-1])
apply(X=dados_aus[,-1],MARGIN=2,FUN=min)
apply(X=dados_aus[,-1],MARGIN=2,FUN=median)
apply(X=dados_aus[,-1],MARGIN=2,FUN=mean)
apply(X=dados_aus[,-1],MARGIN=2,FUN=max)
apply(X=dados_aus[,-1],MARGIN=2,FUN=sd)
coef.var=function(x)100*sd(x)/mean(x)
apply(X=dados_aus[,-1],MARGIN=2,FUN=coef.var)

# Gráfico boxplot Geral ---
x11()
par(mfrow=c(6,1), pch=8, mai=c(0.78, 0.78, 0, 0), cex=1, las=0)
boxplot(dados_aus[,2], ylim=c(0,168), axes=F, ylab='Faltas',
width=1, horizontal = T)
boxplot(dados_aus[,4], ylim=c(0,140), axes=F, ylab='Justificadas',
width=1, horizontal = T)
boxplot(dados_aus[,5], ylim=c(0,100), axes=F, ylab='Não justificadas',
width=2, horizontal = T)
boxplot(dados_aus[,6], ylim=c(0,10), axes=F, ylab='Atrasos',
width=2, horizontal = T)
boxplot(dados_aus[,7], ylim=c(0,10), axes=F, ylab='Suspensão',
width=2, horizontal = T)
boxplot(dados_aus[,8], ylim=c(0,100), axes=F, ylab='Licença',
xlab='Tempo de falta', width=2, horizontal = T)

# Histograma ---
x11()
nf<-layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(7,1))
par(mai=c(0.5, 1.2, 1, 0), cex.lab=1, cex=1)
k<-sqrt(dim(dados_aus)[1])
classes<-seq(from=min(dados_aus[,3]), to=max(dados_aus[,3]), length.out = 10)
h<-hist(x=dados_aus[,3], breaks = classes, axes=F,
        ylab='Frequências observadas', main='Tempo em anos na empresa',
        density = 20, ylim=c(0,30))
prop<-round(100*h$counts/82,2)
text(x=h$mids, y=h$counts+1, labels=paste(prop,'%'))
axis(side=1, at=classes, labels=round(classes/(24*365),0))
par(mar=c(0, 5, 0, 0))
```

```
boxplot(x=dados_aus[,3], horizontal = T, axes=F, pch=8)
rm(k,classes,h,prop)

x11()
par(mai=c(0.5, 1.2, 1, 0),cex.lab=1,cex=1)
k<-sqrt(dim(dados_aus)[1])
temp<-dados_aus[,3]*dados_aus[,2]/10000
classes<-seq(from=min(temp), to=max(temp), length.out = 10)
h<-hist(x=temp, breaks = classes, axes=F, ylab='Frequências observadas',
main='Faltas em dias contínuos', density = 20, ylim=c(0,45))
prop<-round(100*h$counts/82,2)
text(x=h$mids,y=h$counts+2,labels=paste(prop,'%'))
axis(side=1,at=classes,labels=round(classes/24,0))
par(mar=c(0, 5, 0, 0))
boxplot(x=temp,horizontal = T,axes=F,pch=8)
rm(k,temp,classes,h)

# Correlação ----
temp<-dados_aus[,3]*dados_aus[,2]/10000 #Faltas não padronizadas
cor.test(dados_aus[,3], temp, method='spearman')
temp<-cor(dados_aus[,4:8], method = 'spearman')
temp[lower.tri(temp)]<-NA
temp<-round(temp,2)
COR<-matrix( NA, nrow = 5, ncol = 5)
for(i in 1:5){
for (j in 1:5) {
ro = cor.test(dados_aus[,i+3], dados_aus[,j+3], method = 'spearman' )
COR[i,j] = round(ro$p.value,3)
}
}

# Análise de Clusters ----
D = dist(dados_aus[,-(1:3)], method = 'euclidean')
dendro = hclust(D, method='ward.D')
rm(D)

# Grafico do nível de fusão ---
x11()
par(mai=c(1, 1.1, 0.5, 0.125), cex.lab=1, cex=1)
plot(c(61,81),c(0,1387), type = 'n', frame.plot = FALSE, axes = FALSE,
ylab = 'Dissimilaridades', xlab='Passos', main = 'Nível de fusão')
points(dendro$height[-c(79,80,81)],pch=19,xlim=c(61,81),yaxt="n",xaxt="n")
axis(side=2, at=dendro$height[c(61,79:81)],
labels=round(dendro$height[c(61,79:81)],0),las=1)
axis(side=1, at=61:81) #3 grupos
points(c(79,80),dendro$height[79:80],col="red",type = 'b',pch=19)
points(81,dendro$height[81],pch=19)
```

```
abline(h=dendro$height[79:80], col='gray',lty=2,lwd=0.5)

# Gráfico dendrograma ---
x11()
par(mai=c(0.5, 1.25, 0.5, 0.125),cex.lab=1,cex=1,lwd=2)
plot(dendro, main ='Dendrograma', axes = F,ylab="Dissimilaridades",
lwd=1.5,hang=-1) # Dendrograma
axis(side=2,at=dendro$height[c(60,79:81)],
labels=round(dendro$height[c(60,79:81)],0),las=1)
rect.hclust(tree=dendro,h=238,border=2:4)
mtext(side=1,at=c(7.5,22,55), text=c("G3","G2","G1"),cex=1,line=0.5,col=2:4)

# Escrevendo os clusters ---
g<-3 #Grupos
grupos<-cutree(dendro, k = g)
table(grupos)

# Escrevendo o arquivo com os dados_aus de cada grupo ---
grupo1=dados[grupos==1,]
write.csv2(x=grupo1,file="grupo1.csv",row.names = F)
grupo2=dados[grupos==2,]
write.csv2(x=grupo2,file="grupo2.csv",row.names = F)
grupo3=dados[grupos==3,]
write.csv2(x=grupo3,file="grupo3.csv",row.names = F)
rm(g)

# Kruskal Wallis----
# Tabela 01 ---
gf<-as.factor(grupos)
n<-nlevels(gf)
grupo.temp<-vector(mode="list",length=n)

for(i in 1:n){
grupo.temp[[i]]=dados_aus[grupos==i,]
}

xbar.bd=round(apply(X=dados_aus, MARGIN = 2, FUN=mean),2)
xbar<-vector(mode="list",length=n)
for(i in 1:n){
xbar[[i]]=round(apply(X=dados_aus[grupos==i,], MARGIN = 2, FUN=mean),2)
}
temp<-matrix(nrow=n,ncol=ncol(dados_aus)+1)
for(i in 1:n){
temp1<-c(dim(grupo.temp[[i]])[1],xbar[[i]])
temp[i,]<-as.matrix(temp1,byrow=TRUE)
}
```

```
tabela<-rbind(c(dim(dados_aus)[1],xbar.bd),temp)
temp<-character(length=n)
for(i in 1:n){
temp[i]=paste0("Grupo",levels(gf)[i])
}
row.names(tabela)<-c("Geral",temp)
colnames(tabela)[1]<-"Obs"
tabela<-tabela[,-2]
tabela

# teste de medidas ---
grupo1<-dados_aus[grupos==1,]
grupo2<-dados_aus[grupos==2,]
grupo3<-dados_aus[grupos==3,]
dados_aus<-do.call(rbind, lapply(grupo.temp, as.data.frame))

temp1<-vector(mode="list",length = n)
for(i in 1:n){
temp=paste0("g",i)
temp1[i]=list(rep(temp,time=dim(grupo.temp[[i]])[1]))
}
grupos<-do.call(c,temp1)
rm(temp,temp1)
grupos<-as.factor(grupos)

# Variaveis relacionadas as ausencias ---
letras<-vector(mode="list",length=7)
for (i in 1:7) {
VAR<-dados_aus[,i+1]
mod = aov(VAR ~ grupos)
shapiro.test(residuals(mod))
bartlett.test(residuals(mod),grupos)

boxplot(VAR~grupos,main=colnames(dados_aus[,i+1]),pch=8)
a<-kruskal(VAR, grupos)
idx<-order(rownames(a$groups))
letras[[i]]<-as.data.frame(a$groups[idx,2])

}

# Tabela---
letras<-as.data.frame(letras)
tabela.letras=cbind(letras)
colnames(tabela.letras)=c("Faltas","Tempo","Justificadas",
"Não justificadas","Atraso",
"suspensao","Licenca")
linha0=rep(NA,times=7)
```

```
tabela.letras=rbind(linha0,tabela.letras)
tabela1=cbind(tabela,tabela.letras)
tabela2=tabela1[,c(1,2,9,3,10,4,11,5,12,6,13,7,14,8,15)]
colnames(tabela2)=c("n","Faltas","","Tempo","","Justificadas","",
"Não justificadas","","Atraso","","suspensao","",
"Licenca","")
rownames(tabela2) = c("G","G1","G2","G3");
tabela2
detach(dados_aus)

# Análise exploratória dados sociais ----
mean(dados_social[,4]) #Idade
coef.var=function(x)100*sd(x)/mean(x)
coef.var(dados_social[,4]) #Idade
summary(dados_social[,4])
mean(dados_social[,7]) #Distancia
coef.var=function(x)100*sd(x)/mean(x)
coef.var(dados_social[,7]) #Distancia
summary(dados_social[,7])

# Gráfico de barras ---
temp=apply(X=dados_social[,c(-4,-7)],MARGIN = 2,FUN = table)
temp=temp[,c(1,4,3,2,5)]
temp1=round(100*temp/82,1)
x11()
par(mai=c(1.3, 2.4, 0.82, 0),cex=1)
temp2=c("Cobrador", "Motorista", "Fundamental", "Médio",
"Solteiro", "Casado", "Feminino", "Masculino",
"Paga pensão", "Não paga pensão")
temp3=barplot(temp,beside=T,names.arg=temp2,horiz = T, las = 2,
density = c(0,25),col="black",xlim=c(0,90),axes = F,
space=c(0,0.25), axisnames= T, xlab = "Frequências observadas")
text(x=c(30,23,13,2,2),y=temp3[1,],labels = paste(temp1[1,],"%"),pos=4)
text(x=c(52,59,68,79,79),y=temp3[2,],labels =paste(temp1[2,],"%"),pos=4)
rm(temp,temp1,temp2,temp3)

# Histograma para idade e distância ---
x11()
nf<-layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(7,1))
par(mai=c(0.5, 1.1, 1, 0),cex.lab=1,cex=1)
k=sqrt(dim(dados_social)[1])
classes=seq(from=min(dados_social[,4]),to=max(dados_social[,4]),
            length.out = 10)
h=hist(x=dados_social[,4],breaks = classes,axes=F,xlab="",
main="Idade em anos dos funcionários",
density = 20,ylim=c(0,22.5), ylab = 'Frequências observadas')
prop=round(100*h$counts/82,2)
```

```
text(x=h$mids,y=h$counts+1,labels=paste(prop,"%"))
axis(side=1,at=classes,labels = round(classes,1))
par(mar=c(0, 4.5, 0, 0),cex=1)
boxplot(x=dados_social[,4],horizontal = T,axes=F,pch=8)
rm(k,classes,h,prop)

x11()
par(mai=c(0.5, 1.1, 1, 0),cex.lab=1,cex=1)
k=sqrt(dim(dados_social)[1])
classes=seq(from=min(dados_social[,7]),to=max(dados_social[,7]),
           length.out = 10)
h=hist(x=dados_social[,7],breaks = classes,axes=F,xlab="",
main="Distância da casa até a empresa em km",
density = 20,ylim=c(0,29), ylab = 'Frequências observadas')
prop=round(100*h$counts/82,2)
text(x=h$mids,y=h$counts+2,labels=paste(prop,"%"))
axis(side=1,at=classes,labels = round(classes,1))
par(mar=c(0, 4.5, 0, 0),cex.lab=1,cex=1)
boxplot(x=dados_social[,7],horizontal = T,axes=F,pch=8)
rm(k,classes,h,prop)

# Componente principal ----
# dados ausencias ---
(pca1 = prcomp(x=dados[,1:5]))
summary(pca1)
esc1=as.matrix(dados[,1:5])%*%pca1$rotation[,1]
summary(esc1)

# dados sociais ---
(pca2 = prcomp(x=dados[,6:12]))
summary(pca2)
esc2=as.matrix(dados[,6:12])%*%pca2$rotation[,1]
summary(esc2)

rm(pca1,pca2)

# Gráfico de componentes principais ---
bp1=boxplot(x=esc1,plot=F)
bp1.x=c(bp1$stats,max(bp1$out))
bp2=boxplot(x=-esc2,plot=F)
bp2.y=c(bp2$stats,max(bp2$out))

x11()
nf <- layout(mat=matrix(data=c(3,0,1,2),nrow=2),width=c(1,6),heights=c(7,1))
par(mai=c(0.5, 0.25, 0.125, 0.125),cex.lab=1,cex=1)
plot(esc1,-esc2,pch=19,axes=T,xaxt="n",yaxt="n")
axis(side=1,at=bp1.x,labels = round(bp1.x,1))
```

```
axis(side=2,at=bp2.y,labels = round(bp2.y,1),las=1)
abline(v=median(esc1),h=median(-esc2),lwd=1.5,lty=2)

par(mai=c(0, 0.25, 0, 0.125),cex.lab=1,cex=1)
boxplot(x=esc1,horizontal=T,pch=8,axes=F,boxwex=0.55)
text(x=diff(range(esc1))/2,y=0.60,labels = expression(CP[1]),
cex=1)

par(mai=c(0.5, 0, 0.125, 0),cex.lab=1,cex=1)
boxplot(x=-esc2,pch=8,axes=F,boxwex=0.35)
text(y=50,x=0.65,labels = expression(CP[2]),cex=1)
rm(bp1,bp2,bp1.x,bp2.y,nf,esc1,esc2)

# Biplot ----
# variáveis ausência ---
dvs <- svd(dados[,1:5])
rownames(dvs$u) <- rownames(dados[,1:5]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados[,1:5]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

# Gráfico ---
x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-7,1),c(-8,1), type = 'n', frame.plot = FALSE,
axes = FALSE, xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+1, by=1),2))
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1, by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.15,labels=rownames(variav),
col = "blue",cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = "blue")
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = "blue")
title(main = 'Biplot Geral - Faltas')

# variáveis sociais ---
dvs <- svd(dados[,6:12])
rownames(dvs$u) <- rownames(dados[,6:12]) # obj
U <- dvs$u
```

```
rownames(dvs$v) <- colnames(dados[,6:12]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

# Gráfico ---
x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-4,1),c(-2,2), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+2, by=1),2))
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1, by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.05,labels=rownames(variav),
col = 'blue',cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Geral - Aspectos sociais')

rm(grupo1)



# Grupo 1 ----
# leitura de dados ---
dados1<-read.table(file="grupo1.csv",header=TRUE,sep = ";",
dec =',')

# Gráfico de barras ---
n=dim(dados1)[1]
temp=apply(X=dados1[,c(-(1:5),-9,-10,-12)],MARGIN = 2,FUN = table)
temp=cbind(temp,c(0,n))
colnames(temp)[5]<-"Instrucao"
temp=temp[,c(1,5,3,2,4)]
temp1=round(100*temp/n,1)

x11()
par(mai=c(1.3, 2.4, 0.82, 0),cex=1)
temp2=c("Cobrador", "Motorista", "Fundamental", "Médio",
"Solteiro", "Masculino", "Feminino", "Casado",
"Paga pensão", "Não paga pensão")
```

```
temp3=barplot(temp,beside=T,names.arg=temp2,las=2,horiz = T,
density = c(0,25),col="black",axes=F,
xlim=c(0,max(temp)+7.5),main="G1",
space=c(0,0.25), xlab = 'Frequências observadas')
text(x=temp[1,],y=temp3[1,],labels = paste(temp1[1,],"%"),pos=4)
text(x=temp[2,],y=temp3[2,],labels =paste(temp1[2,],"%"),pos=4)
rm(temp,temp1,temp2,temp3)

# Grafico boxplot ausência ---
x11()
par(pch=8,mai=c(0.5, 1.2, 0.4, 0),cex.lab=1,cex=1)
Faltas=dados1[,1]+dados1[,2]+dados1[,3]+dados1[,4]+dados1[,5]
Temp13<-cbind(Faltas,dados1[,1:5])
name = c('Faltas', 'Justificadas', 'Não justificadas', 'Atrasos',
'Suspensão', 'Licença')
colnames(Temp13) <- name
temp=boxplot(Temp13,ylim=c(0,50),yaxt="n",
frame.plot=F,main="G1", ylab = 'Tempo de falta')
axis(side=2,at=round(seq(from=0,to=44.9,length.out = 4),1),las=1)

# Correlacao ---
temp=cor(dados1[,1:5], method = "spearman")
temp[lower.tri(temp)]<-NA
temp=round(temp,2)
COR = matrix( NA, nrow = 5, ncol = 5)
for(i in 1:5){
for (j in 1:5) {
ro = cor.test(dados1[,i],dados1[,j], method = 'spearman' )
COR[i,j]= round(ro$p.value,3)
}
}

# Componentes Principais ---
(pca1 = prcomp(x=dados1[,1:5]))
summary(pca1)
esc1=as.matrix(dados1[,1:5])%*%pca1$rotation[,1]
summary(esc1)
(pca2 = prcomp(x=dados1[,c(-(1:5))]))
summary(pca2)
esc2=as.matrix(dados1[,c(-(1:5))])%*%pca2$rotation[,1]
summary(esc2)
esc2=-esc2
plot(esc1,esc2)
rm(pca1,pca2)

# Gráfico de componentes principais ---
bp1=boxplot(x=esc1,plot=F)
```

```
bp1.x=c(bp1$stats,max(bp1$out))
bp2=boxplot(x=esc2,plot=F)
bp2.y=c(bp2$stats,max(bp2$out))

x11()
nf <- layout(mat=matrix(data=c(3,0,1,2),nrow=2),width=c(1,6),heights=c(7,1))
par(mai=c(0.25, 0.25, 0.5, 0.125),cex.lab=1,cex=1)
plot(esc1,esc2,pch=19,axes=T,xaxt="n",yaxt="n", main = 'G1 ')
axis(side=1,at=bp1.x,labels = round(bp1.x,1))
axis(side=2,at=bp2.y,labels = round(bp2.y,1),las=1)
abline(v=median(esc1),h=median(esc2),lwd=1.5,lty=2)

par(mai=c(0, 0.25, 0, 0.125),cex.lab=1,cex=1)
boxplot(x=esc1,horizontal=T,pch=8,axes=F,boxwex=0.55)
text(x=diff(range(esc1))/2,y=0.60,labels = expression(CP[1]),
cex=1)
par(mai=c(0.5, 0, 0.125, 0),cex.lab=1,cex=1)
boxplot(x=esc2,pch=8,axes=F,boxwex=0.35)
text(y=50,x=0.65,labels = expression(CP[2]),cex=1)
rm(bp1,bp2,bp1.x,bp2.y,nf,esc1,esc2)

# Biplot ---
dvs <- svd(dados1[,1:5])
rownames(dvs$u) <- rownames(dados1[,1:5]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados1[,1:5]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-7,1),c(-5,4), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+1,by=1),2),las=1)
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+2,by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.25,labels=rownames(variav),
col = 'blue',cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 1 - Faltas')
```

```
# variáveis de aspecto social ---
dvs <- svd(dados1[,6:12])
rownames(dvs$u) <- rownames(dados1[,6:12]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados1[,6:12]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

# Gráfico ---
x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-5,1),c(-2,2), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+2, by=1),2))
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1, by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.05,labels=rownames(variav),
col = 'blue',cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 1 - Aspectos sociais')

# Grupo 2 ----
# leitura de dados ---
dados2 = read.table(file="grupo2.csv",header=TRUE,sep = ";",
dec =',')

# Gráfico de barras ---
n=dim(dados2)[1]
temp=apply(X=dados2[,c(-(1:5),-7,-9,-12)],MARGIN = 2, FUN = table)
temp=cbind(temp,c(0,n))
colnames(temp)[5]<-"Sexo"
temp=temp[,c(1,3,2,5,4)]
temp1=round(100*temp/n,1)

x11()
par(mai=c(1.3, 2.4, 0.82, 0),cex=1)
temp2=c("Cobrador", "Motorista", "Fundamental", "Médio",
"Solteiro", "Casado", "Feminino", "Masculino",
```

```
"Paga pensão", "Não paga pensão")
temp3=barplot(temp,beside=T,names.arg=temp2,las=2,horiz = T,
density = c(0,25),col="black",axes=F,
xlim=c(0,max(temp)+2.4),
main="G2",space=c(0,0.25), xlab = 'Frequências observadas')
text(x=temp[1,],y=temp3[1,],labels = paste(temp1[1,],"%"),pos=4)
text(x=temp[2,],y=temp3[2,],labels =paste(temp1[2,],"%"),pos=4)
rm(temp,temp1,temp2,temp3)

# Grafico boxplot ausência ---
x11()
par(pch=8,mai=c(0.5, 1.2, 0.4, 0),cex.lab=1,cex=1)
Faltas=dados2[,1]+dados2[,2]+dados2[,3]+dados2[,4]+dados2[,5]
Temp13<-cbind(Faltas,dados2[,1:5])
name = c('Faltas', 'Justificadas', 'Não justificadas', 'Atrasos',
'Suspensão', 'Licença')
colnames(Temp13) <- name
temp=boxplot(Temp13,ylim=c(0,70),yaxt="n",
frame.plot=F,main="G2", ylab = 'Tempo de falta')
axis(side=2,at=round(seq(from=0,to=67.3,length.out = 4),1),las=1)

# Correlacao ---
temp=cor(dados2[,1:5], method = "spearman")
temp[lower.tri(temp)]<-NA
temp=round(temp,2)
COR = matrix( NA, nrow = 5, ncol = 5)
for(i in 1:5){
for (j in 1:5) {
ro = cor.test(dados2[,i],dados2[,j], method = 'spearman' )
COR[i,j]= round(ro$p.value,3)
}
}

# Componentes Principais ---
(pca1 = prcomp(x=dados2[,1:5]))
summary(pca1)
esc1=as.matrix(dados2[,1:5])%*%pca1$rotation[,1]
summary(esc1)
(pca2 = prcomp(x=dados2[,c(-(1:5))]))
summary(pca2)
esc2=as.matrix(dados2[,c(-(1:5))])%*%pca2$rotation[,1]
summary(esc2)
esc2=-esc2
plot(esc1,esc2)
rm(pca1,pca2)

# Gráfico de componentes principais ---
```

```
bp1=boxplot(x=esc1,plot=F)
bp1.x=c(bp1$stats)
bp2=boxplot(x=esc2,plot=F)
bp2.y=c(bp2$stats)

x11()
nf <- layout(mat=matrix(data=c(3,0,1,2),nrow=2),width=c(1,6),heights=c(7,1))
par(mai=c(0.25, 0.25, 0.5, 0.125),cex.lab=1,cex=1)
plot(esc1,esc2,pch=19,axes=T,xaxt="n",yaxt="n", main = 'G2 ')
axis(side=1,at=bp1.x,labels = round(bp1.x,1))
axis(side=2,at=bp2.y,labels = round(bp2.y,1),las=1)
abline(v=median(esc1),h=median(esc2),lwd=1.5,lty=2)

par(mai=c(0, 0.25, 0, 0.125),cex.lab=1,cex=1)
boxplot(x=esc1,horizontal=T,pch=8,axes=F,boxwex=0.55)
text(x=diff(range(esc1))/2,y=0.60,labels = expression(CP[1]),
cex=1)

par(mai=c(0.5, 0, 0.625, 0),cex.lab=1,cex=1)
boxplot(x=esc2,pch=8,axes=F,boxwex=0.35)
text(y=50,x=0.65,labels = expression(CP[2]),cex=1)
rm(bp1,bp2,bp1.x,bp2.y,nf,esc1,esc2)

# Biplot ---
dvs <- svd(dados2[,1:5])
rownames(dvs$u) <- rownames(dados2[,1:5]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados2[,1:5]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-5,1),c(-5,2), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+4,by=1),2),las=1)
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1,by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.25,labels=rownames(variav),
col = 'blue',cex = 0.8)
```

```
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 2 - Faltas')

# variáveis de aspecto social ---
dvs <- svd(dados2[,6:12])
rownames(dvs$u) <- rownames(dados2[,6:12]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados2[,6:12]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-5,1),c(-1,1), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+2, by=1),2))
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1, by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.05,labels=rownames(variav),
col = 'blue',cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 2 - aspectos sociais')

# Grupo 3 ----
# leitura de dados ---
dados3 = read.table(file="grupo3.csv",header=TRUE,sep = ";",
dec =',')

# Gráfico de barras ---
n=dim(dados3)[1]
temp=apply(X=dados3[,c(-(1:5),-9,-11,-12)],MARGIN = 2,FUN = table)
temp=cbind(temp,c(0,n))
colnames(temp)[5]<-"Pensao"
temp=temp[,c(1,4,3,2,5)]
temp1=round(100*temp/n,1)

x11()
par(mai=c(1.3, 2.4, 0.82, 0),cex=1)
```

```
temp2=c("Cobrador", "Motorista", "Fundamental", "Médio",
"Solteiro", "Casado", "Feminino", "Masculino",
"Paga pensão", "Não paga pensão")
temp3=barplot(temp,beside=T,names.arg=temp2,las=2,horiz = T,
density = c(0,25),col="black",axes=F,xlim=c(0,max(temp)+1.4),
main="G3",space=c(0,0.25),
xlab = 'Frequências observadas')
text(x=c(4,4,2,1,0),y=temp3[1,],labels = paste(temp1[1,],"%"),pos=4)
text(x=c(7.0,7.0,9,10,10.75)+2,y=temp3[2,],labels =paste(temp1[2,],"%"),
     pos=4)
rm(temp,temp1,temp2,temp3)

# Grafico boxplot Absenteismo ---
x11()
par(pch=8,mai=c(0.5, 1.2, 0.4, 0),cex.lab=1,cex=1)
Faltas=dados3[,1]+dados3[,2]+dados3[,3]+dados3[,4]+dados3[,5]
Temp13<-cbind(Faltas,dados3[,1:5])
name = c('Faltas', 'Justificadas', 'Não justificadas', 'Atrasos',
'Suspensão', 'Licença')
colnames(Temp13) <- name
temp=boxplot(Temp13,ylim=c(0,170),yaxt="n",
frame.plot=F,main="G3", ylab = 'Tempo de falta')
axis(side=2,at=round(seq(from=0,to=167.3,length.out = 4),1),las=1)

# Correlacao ---
temp=cor(dados3[,1:5], method = "spearman")
temp[lower.tri(temp)]<-NA
temp=round(temp,2)
COR = matrix( NA, nrow = 5, ncol = 5)
for(i in 1:5){
for (j in 1:5) {
ro = cor.test(dados3[,i],dados3[,j], method = 'spearman' )
COR[i,j]= round(ro$p.value,3)
}
}

# Componentes Principais ---
(pca1 = prcomp(x=dados3[,1:5]))
summary(pca1)
esc1=as.matrix(dados3[,1:5])%*%pca1$rotation[,1]
summary(esc1)
esc1=-esc1
(pca2 = prcomp(x=dados3[,c(-(1:5))]))
summary(pca2)
esc2=as.matrix(dados3[,c(-(1:5))])%*%pca2$rotation[,1]
summary(esc2)
esc2=-esc2
```

```
plot(esc1,esc2)
rm(pca1,pca2)

# Gráfico de componentes principais ---
bp1=boxplot(x=esc1,plot=F)
bp1.x=c(bp1$stats)
bp2=boxplot(x=esc2,plot=F)
bp2.y=c(bp2$stats)

x11()
nf <- layout(mat=matrix(data=c(3,0,1,2),nrow=2),width=c(1,6),heights=c(7,1))
par(mai=c(0.25, 0.25, 0.5, 0.125),cex.lab=1,cex=1)
plot(esc1,esc2,pch=19,axes=T,xaxt="n",yaxt="n", main = 'G3 ')
axis(side=1,at=bp1.x,labels = round(bp1.x,1))
axis(side=2,at=bp2.y,labels = round(bp2.y,1),las=1)
abline(v=median(esc1),h=median(esc2),lwd=1.5,lty=2)

par(mai=c(0, 0.25, 0, 0.125),cex.lab=1,cex=1)
boxplot(x=esc1,horizontal=T,pch=8,axes=F,boxwex=0.55)
text(x=73+diff(range(esc1))/2,y=0.60,labels = expression(CP[1]),
cex=1)

par(mai=c(0.5, 0, 0.125, 0),cex.lab=1,cex=1)
boxplot(x=esc2,pch=8,axes=F,boxwex=0.35)
text(y=50,x=0.65,labels = expression(CP[2]),cex=1)
rm(bp1,bp2,bp1.x,bp2.y,nf,esc1,esc2)

# Biplot ---
dvs <- svd(dados3[,1:5])
rownames(dvs$u) <- rownames(dados3[,1:5]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados3[,1:5]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-7,0),c(-5,2), type = 'n', frame.plot = FALSE, axes = FALSE,
xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),1,by=1),2),las=1)
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1,by=1),2),las=2)
```

```
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.25,labels=rownames(variav),
col = 'blue',cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 3 - Faltas')

# Aspectos sociais ---
dvs <- svd(dados3[,6:12])
rownames(dvs$u) <- rownames(dados3[,6:12]) # obj
U <- dvs$u
rownames(dvs$v) <- colnames(dados3[,6:12]) # var
V <- dvs$v
Lambda <- diag(dvs$d)
obs <- U %*% sqrt(Lambda)
variav <- t(sqrt(Lambda)%*%t(V))
d1 <- (max(variav[,1]) - min(variav[,1])) / (max(obs[,1]) - min(obs[,1]))
d2 <- (max(variav[,2]) - min(variav[,2])) / (max(obs[,2]) - min(obs[,2]))
d <- max(d1,d2)
variav <- variav/d

x11()
par(mai=c(0.5, 0.7, 0.82, 0.5),cex=1)
plot(c(-5,1),c(-2,3), type = 'n', frame.plot = FALSE,
axes = FALSE, xlab = " ", ylab = " " )
axis(side = 1, round(seq(min(obs[,1]),max(obs[,1])+3, by=1),2))
axis(side = 2, round(seq(min(obs[,2]),max(obs[,2])+1, by=1),2),las=2)
abline(h=0,v=0,lty=3)
points(obs[,1],obs[,2], pch = 19, col = "red")
text(x=variav[,1],y=variav[,2]-0.05,labels=rownames(variav),col = 'blue',
cex = 0.8)
points(x=variav[,1],y=variav[,2],pch=19,col = 'blue')
segments(0,0,variav[,1],variav[,2],lwd=0.8, col = 'blue')
title(main = 'Biplot Grupo 3 - Aspectos sociais')
```