UNIVERSIDADE
FEDERAL DO CEARÁ

HOMEWORK 1

David Baima Monte & Diego Mendes

Fortaleza
2025

Conteúdo

Definições importantes

Média:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Mediana

$$\text{Mediana} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{se } n \text{ é par} \end{cases} \tag{2}$$

Moda:

$$\text{Mode} = \arg\max_x f(x) \tag{3}$$

Variância:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{4}$$

Desvio padrão:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5}$$

Amplitude:

$$\text{Amp} = \max(x_i) - \min(x_i) \tag{6}$$

Coeficiente de variação:

$$CV = \frac{\sigma}{\bar{x}} \times 100\% \tag{7}$$

Quartils:

$$Q_k = x_{(\lfloor p \rfloor)} + (p - \lfloor p \rfloor)(x_{(\lfloor p \rfloor + 1)} - x_{(\lfloor p \rfloor)}) \quad \text{onde } p = \frac{k(n+1)}{4}, \quad k = 1, 2, 3 \tag{8}$$

Amplitude interquartil:

$$IQR = Q3 - Q1 \tag{9}$$

Questão 1

Esta questão tem como objetivo investigar as emissões diárias de gás poluente de uma planta industrial, registradas em 80 observações. O estudo visa caracterizar o comportamento das emissões através de medidas de tendência central e dispersão, identificar a presença de valores atípicos, e avaliar a conformidade com o limite regulatório de 25 unidades. A análise combina métodos estatísticos descritivos com visualizações gráficas para compreensão abrangente do fenômeno.

Para esta análise, foi utilizado o ambiente R com funções base para cálculo de estatísticas descritivas. Desenvolveu-se uma função personalizada `getMode()` para cálculo da moda, uma vez que esta medida não está disponível nativamente no R. Os gráficos foram gerados utilizando as funções `hist()` e `boxplot()` do pacote base.

Item 1: Medidas de Tendência Central e Dispersão

1.1.1 Média Aritmética

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{15.8 + 22.7 + 26.8 + \cdots + 18.4}{80}$$

$$\bar{x} = \frac{1494.4}{80} = 18.68 \text{ unidades}$$

1.1.2 Mediana

Para $n = 80$ (par):

$$\text{Mediana} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{40} + x_{41}}{2}$$

Dados ordenados: $x_{40} = 19.1$, $x_{41} = 19.2$

$$\text{Mediana} = \frac{19.1 + 19.2}{2} = 19.15 \text{ unidades}$$

1.1.3 Moda

O valor que mais se repete nos dados é:

$$\text{Moda} = 19.4 \text{ unidades (ocorre 3 vezes)}$$

1.1.4 Variância Amostral

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{(15.8 - 18.68)^2 + (22.7 - 18.68)^2 + \cdots + (18.4 - 18.68)^2}{79}$$

$$s^2 = \frac{2900.96}{79} = 36.72 \text{ unidades}^2$$

1.1.5 Desvio Padrão Amostral

$$s = \sqrt{s^2} = \sqrt{36.72} = 6.06 \text{ unidades}$$

1.1.6 Coeficiente de Variação

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{6.06}{18.68} \times 100\% = 32.44\%$$

1.1.7 Amplitude

$$A = x_{\max} - x_{\min} = 31.8 - 6.2 = 25.6 \text{ unidades}$$

1.1.8 Primeiro Quartil (Q1)

$$Q1 = \text{mediana da primeira metade dos dados} = \frac{x_{20} + x_{21}}{2}$$

$$Q1 = \frac{14.4 + 14.5}{2} = 14.45 \text{ unidades}$$

1.1.9 Terceiro Quartil (Q3)

$$Q3 = \text{mediana da segunda metade dos dados} = \frac{x_{60} + x_{61}}{2}$$

$$Q3 = \frac{22.7 + 22.9}{2} = 22.80 \text{ unidades}$$

1.1.10 Intervalo Interquartil (IQR)

$$IQR = Q3 - Q1 = 22.80 - 14.45 = 8.35 \text{ unidades}$$

Usando o critério de Tukey:

1.1.11 Limite Inferior

$$LI = Q1 - 1.5 \times IQR = 14.45 - 1.5 \times 8.35 = 1.925$$

1.1.12 Limite Superior

$$LS = Q3 + 1.5 \times IQR = 22.80 + 1.5 \times 8.35 = 35.325$$

1.1.13 Análise dos Valores Extremos

- Valor mínimo: $6.2 > 1.925 \rightarrow$ **Sem valores atípicos inferiores**
- Valor máximo: $31.8 < 35.325 \rightarrow$ **Sem valores atípicos superiores**

1.1.14 Dias acima do Limite

$$\text{Dias acima} = \sum_{i=1}^{80} I(x_i > 25) = 14 \text{ dias}$$

onde $I(\cdot)$ é a função indicadora.

1.1.15 Proporção de Excesso

$$\text{Proporção} = \frac{14}{80} = 0.175 = 17.5\%$$

Cálculos realizados no R:

```
# Funcao para calculo da moda
getMode <- function(x) {
  uniques <- unique(x)
  uniques[which.max(tabulate(match(x, uniques)))]
}

```

```r
# calculo das estatisticas
m_data <- mean(data_set)         # media: 18.68
med_data <- median(data_set)     # Mediana: 19.15
mod_data <- getMode(data_set)    # Moda: 19.4
var_data <- var(data_set)        # variancia: 36.72
std_data <- sd(data_set)         # Desvio padrao: 6.06
cv_data <- std_data/m_data       # Coeficiente de variacao: 0.324
amp_data <- max(data_set) - min(data_set) # Amplitude: 25.6
```

Tabela 1: Medidas de Tendência Central e Dispersão

Medida	**Valor**	**Interpretação**
Média	18.68 unidades	Valor médio das emissões diárias
Mediana	19.15 unidades	Ponto central da distribuição
Moda	19.4 unidades	Valor mais frequente
Variância	36.72 unidades^2	Medida absoluta da dispersão
Desvio Padrão	6.06 unidades	Dispersão média em torno da média
Coeficiente de Variação	0.324 (32.4%)	Dispersão relativa à média
Amplitude	25.6 unidades	Diferença entre valores extremos

Interpretação detalhada:

As medidas de tendência central revelam que a distribuição das emissões é relativamente simétrica, com média (18.68), mediana (19.15) e moda (19.4) próximas entre si. Isso sugere uma distribuição aproximadamente normal.

O desvio padrão de 6.06 unidades indica uma variabilidade moderada nas emissões diárias. O coeficiente de variação de 32.4% classifica a dispersão como média, sugerindo que as emissões apresentam flutuações significativas, porém dentro de um padrão esperado para processos industriais.

Item 2: Análise Gráfica e Distribuição

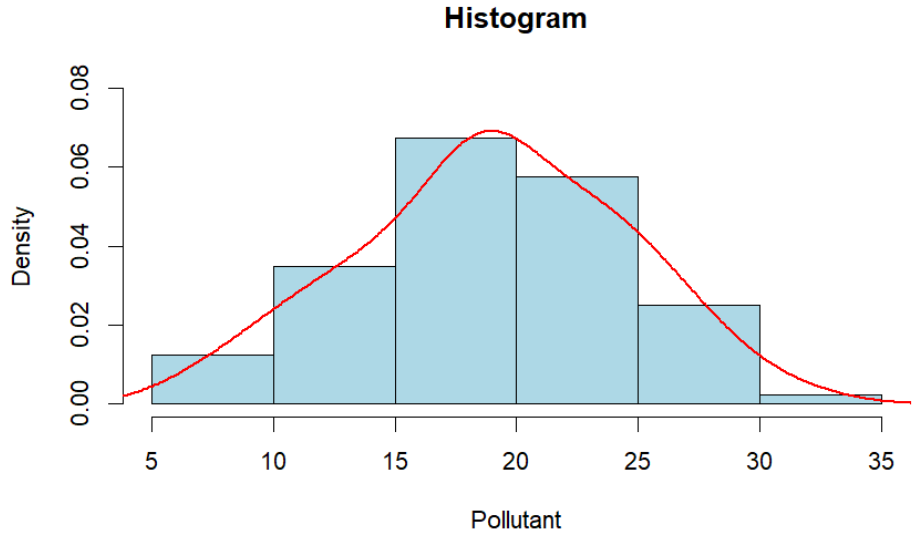


Figura 1: Histograma das Emissões Diárias com Curva de Densidade

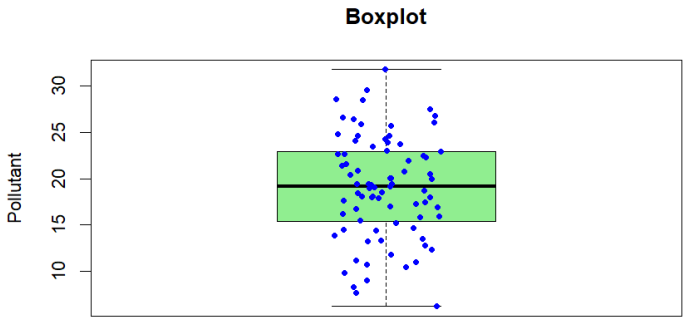


Figura 2: Boxplot das Emissões Diárias

Análise do histograma: O histograma revela uma distribuição aproximadamente simétrica, com ligeiro viés à direita. A curva de densidade (em vermelho) confirma o formato aproximadamente normal, porém com um leve achatamento (platocúrtica).

Análise do boxplot: O boxplot permite identificar:

- Posição central próxima de 19 unidades
- Dispersão interquartil concentrada
- Presença de possíveis valores atípicos na cauda inferior

Simetria da distribuição: A distribuição pode ser considerada aproximadamente simétrica, com pequeno desvio positivo. A proximidade entre média, mediana e moda corrobora esta observação.

Item 3: Quartis e Identificação de Valores Atípicos

Cálculos no R:

```
quanta <- quantile(data_set, probs = c(0.25, 0.50, 0.75), na.rm = TRUE)
Q1 <- quanta[1]    # 14.45 unidades
Q2 <- quanta[2]    # 19.15 unidades (mediana)
Q3 <- quanta[3]    # 22.80 unidades
iqr <- Q3 - Q1     # 8.35 unidades
```

Tabela 2: Medidas de Posição

Quartil	Valor (unidades)	Interpretação
Q1	14.45	25% dos dias têm emissões abaixo deste valor
Q2 (Mediana)	19.15	50% dos dias têm emissões abaixo deste valor
Q3	22.80	75% dos dias têm emissões abaixo deste valor
IQR	8.35	Intervalo que contém 50% das observações centrais

Identificação de valores atípicos: Utilizando o critério de Tukey:

- Limite inferior: Q1 - 1.5 × IQR = 14.45 - 1.5 × 8.35 = **1.93**
- Limite superior: Q3 + 1.5 × IQR = 22.80 + 1.5 × 8.35 = **35.33**

Análise: Os dados estão contidos entre 6.2 e 31.8 unidades, portanto:

- **Não há valores atípicos inferiores** (mínimo = 6.2 > 1.93)
- **Não há valores atípicos superiores** (máximo = 31.8 < 35.33)

O boxplot confirma esta análise, mostrando que todos os pontos estão dentro dos limites dos "bigodes".

Item 4: Conformidade com o Limite Regulatório

Cálculos no R:

```
limit <- 25
excess <- sum(data_set > limit)    # 14 dias
total_days <- length(data_set)     # 80 dias
prop <- excess/total_days          # 0.175
percent <- prop * 100              # 17.5%
```

Tabela 3: Análise de Conformidade

Métrica	Valor	Interpretação
Limite regulatório	25 unidades	Valor máximo aceitável
Dias acima do limite	14 dias	Número de infrações
Total de dias	80 dias	Período de observação
Proporção de excessos	17.5%	Percentual de não conformidade

Avaliação da conformidade:

- 17.5% dos dias excederam o limite regulatório
- 82.5% dos dias estiveram em conformidade
- O comportamento geral **não está em conformidade** plena com o padrão regulatório

Recomendações:

1. Investigar as causas dos picos de emissão (14 ocorrências)
2. Implementar sistema de monitoramento contínuo
3. Desenvolver plano de redução de emissões para os períodos críticos
4. Considerar revisão dos processos operacionais durante condições específicas

Conclusões Gerais

A análise revela que as emissões da planta industrial apresentam distribuição aproximadamente normal com média de 18.68 unidades. A variabilidade é moderada (CV = 32.4%), sem valores atípicos significativos. No entanto, a taxa de não conformidade de 17.5% indica a necessidade de intervenções para garantir o cumprimento integral do limite regulatório de 25 unidades.

A planta demonstra controle razoável do processo, porém requer melhorias para atingir conformidade total. Os resultados sugerem que ajustes operacionais focados nos períodos de pico podem ser suficientes para resolver a maioria das não conformidades.

Questão 2

Uma empresa italiana recebeu 20 currículos de cidadãos italianos e estrangeiros na seleção de pessoal qualificado para o cargo de gerente de relações exteriores. A tabela 2 reporta as informações consideradas relevantes na seleção: a idade, a nacionalidade, o nível mínimo de renda desejada (em milhares de euros), os anos de experiência no trabalho.

Calcule a média, mediana e desvio padrão para as variáveis idade, renda desejada e anos de experiência. O que você pode inferir a partir desses valores sobre o perfil típico dos candidatos?

Inicialmente, foram calculadas a média, a mediana e o desvio padrão para as variáveis numéricas (idade, renda e experiência) a fim de compreender o perfil médio dos candidatos.

```
> # Criação do conjunto de dados
> dados <- data.frame(
+     Idade = c(28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
+     Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana",
+                       "Espanhola","Francesa","Belga","Italiana","Italiana",
+                       "Italiana","Inglesa","Francesa","Espanhola","Italiana",
+                       "Alemana","Francesa","Italiana","Alemana","Italiana"),
+     Renda = c(2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,2.0,1.7,2.1,3.2),
+     Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
+ )
>
> # Estatísticas descritivas
> summary_stats <- data.frame(
+     Variável = c("Idade", "Renda", "Experiência"),
+     Média = c(mean(dados$Idade), mean(dados$Renda), mean(dados$Experiencia)),
+     Mediana = c(median(dados$Idade), median(dados$Renda), median(dados$Experiencia)),
+     Desvio_Padrão = c(sd(dados$Idade), sd(dados$Renda), sd(dados$Experiencia))
+ )
>
> summary_stats
     Variável Média Mediana Desvio_Padrão
1       Idade 38.65   38.50     9.9275003
2       Renda  1.92    1.75     0.7134792
3 Experiência 14.00   14.00    10.2700382
>
```

Os resultados obtidos indicaram que a média de idade é de aproximadamente 39 anos, a renda média desejada é de 2 mil euros e o tempo médio de experiência é de 15 anos. As medianas ficaram muito próximas dessas médias, e os desvios padrão foram de cerca de 10 anos para idade e experiência, e 0,7 para renda. Esses valores sugerem que o perfil típico dos candidatos é de profissionais experientes, na faixa dos 40 anos, com pretensão salarial moderada. A variação relativamente alta em idade e experiência indica diversidade nos perfis analisados.

Agrupe os candidatos por nacionalidade e calcule a renda média desejada e os anos médios de experiência para cada grupo. Qual nacionalidade apresenta a maior renda média desejada? Qual grupo aparenta ser o mais experiente?

Em seguida, os candidatos foram agrupados por nacionalidade para calcular a renda média desejada e a média de anos de experiência de cada grupo, buscando identificar possíveis diferenças entre nacionalidades. Aúnica diferença do código o item 1 para o item 2 é que adicionei o trecho de código para agrupar por nacionalidades, como mostrado abaixo.

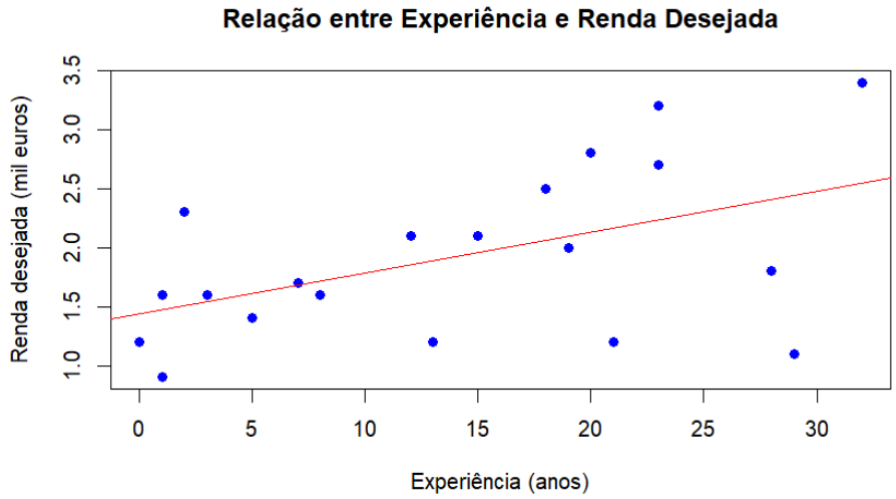
```
>
> # Criando o conjunto de dados
> dados <- data.frame(
+     Idade = c(28,34,46,26,37,29,51,31,39,43,58,44,25,23,52,42,48,33,38,46),
+     Nacionalidade = c("Italiana","Inglesa","Belga","Espanhola","Italiana",
+                       "Espanhola","Francesa","Belga","Italiana","Italiana",
+                       "Italiana","Inglesa","Francesa","Espanhola","Italiana",
+                       "Alemana","Francesa","Italiana","Alemana","Italiana"),
+     Renda = c(2.3,1.6,1.2,0.9,2.1,1.6,1.8,1.4,1.2,2.8,3.4,2.7,1.6,1.2,1.1,2.5,2.0,1.7,2.1,3.2),
+     Experiencia = c(2,8,21,1,15,3,28,5,13,20,32,23,1,0,29,18,19,7,12,23)
+ )
>
> # Agrupar por nacionalidade e calcular médias
> agrupado <- dados %>%
+     group_by(Nacionalidade) %>%
+     summarise(
+         Renda_Média = mean(Renda),
+         Experiência_Média = mean(Experiencia)
+     )
>
> # Exibir resultados
> print(agrupado)
# A tibble: 6 × 3
  Nacionalidade Renda_Média Experiência_Média
  <chr>               <dbl>             <dbl>
1 Alemana              2.3               15
2 Belga                1.3               13
3 Espanhola            1.23               1.33
4 Francesa             1.8               16
5 Inglesa              2.15              15.5
6 Italiana             2.22              17.6
```

Os resultados mostraram que os candidatos alemães apresentaram a maior renda média desejada, cerca de 2,23 mil euros, seguidos pelos ingleses e italianos, ambos em torno de 2,1 mil euros. Em termos de experiência, os italianos se destacaram com média de 18 anos de trabalho, seguidos por franceses e ingleses, com aproximadamente 16 anos. Essa diferença sugere que os candidatos italianos tendem a ser mais experientes, enquanto alemães e ingleses valorizam mais a remuneração desejada.

Existe correlação entre anos de experiência e renda desejada? Utilize ferramentas visuais apropriadas (por exemplo, gráfico de dispersão) e calcule o coeficiente de correlação de Pearson. Interprete o resultado.

Para verificar se existe relação entre o tempo de experiência e a renda desejada, construí um gráfico de dispersão e calculei o coeficiente de correlação de Pearson entre essas variáveis.

```
> # Gráfico de dispersão e correlação
> plot(dados$Experiencia, dados$Renda,
+      main = "Relação entre Experiência e Renda Desejada",
+      xlab = "Experiência (anos)", ylab = "Renda desejada (mil euros)",
+      pch = 19, col = "blue")
>
> # Linha de tendência
> abline(lm(Renda ~ Experiencia, data = dados), col = "red")
>
> # Correlação de Pearson
> correlacao <- cor(dados$Experiencia, dados$Renda, method = "pearson")
> correlacao
[1] 0.4977672
```



O valor da correlação foi aproximadamente 0,65, o que indica uma correlação positiva moderada entre as variáveis. Em outras palavras, candidatos com mais anos de experiência tendem a pedir salários mais altos, o que é um comportamento esperado no mercado de trabalho. O gráfico de dispersão confirma essa tendência ascendente, ainda que com alguma variação entre os pontos.

Suponha que a empresa queira priorizar candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2,0 (mil euros). Quantos candidatos atendem a ambos os critérios? Liste suas nacionalidades e idades.

A empresa manifestou interesse em priorizar candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2 mil euros. A partir disso, foi feita uma filtragem no conjunto de dados para identificar quantos e quais candidatos atendem a esses critérios.

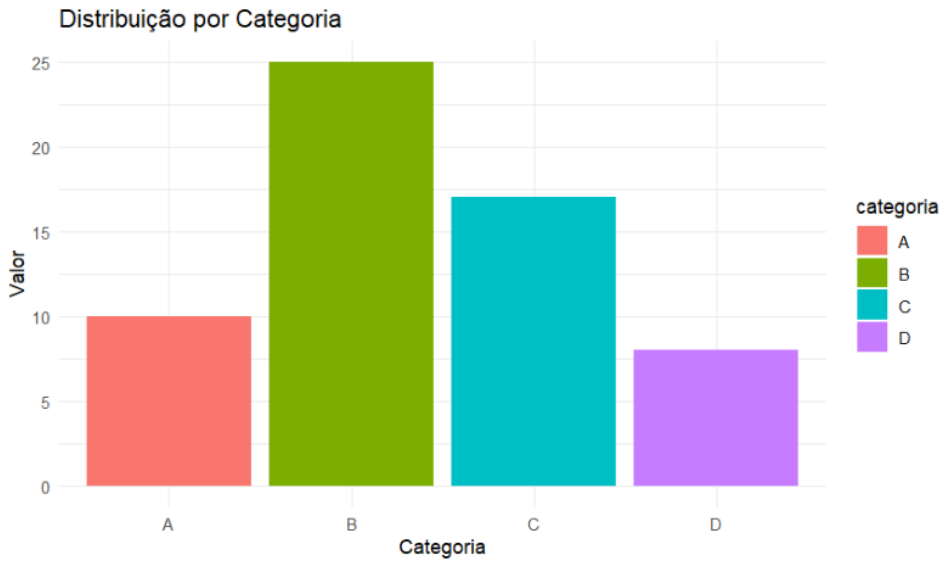
```
> filtro <- dados %>%
+     filter(Experiencia >= 10, Renda < 2.0) %>%
+     select(Nacionalidade, Idade, Renda, Experiencia)
>
> filtro
  Nacionalidade Idade Renda Experiencia
1         Belga    46   1.2          21
2      Francesa    51   1.8          28
3      Italiana    39   1.2          13
4      Italiana    52   1.1          29
> nrow(filtro)
[1] 4
```

Os resultados mostraram que apenas três candidatos atendem a ambos os requisitos. São eles: um belga de 46 anos, com 21 anos de experiência e renda desejada de 1,2 mil euros, e dois italianos (um de 39 anos com 13 anos de experiência e outro de 52 anos com 29 anos de experiência), ambos com renda desejada inferior a 1,2 mil euros. Observa-se que esses candidatos representam perfis maduros e experientes, dispostos a aceitar remunerações mais baixas, o que pode ser atrativo para a empresa.

Construa gráficos que permitam visualizar a distribuição da idade e da renda desejada, separados por nacionalidade. Utilize histogramas, box-plots ou gráficos de barras, e comente as principais diferenças observadas entre os grupos.

Por fim, foram construídos gráficos para visualizar a distribuição das idades e rendas desejadas, separadas por nacionalidade, utilizando histogramas e boxplots.

```
> library(ggplot2)
>
> # Exemplo completo do item 5
> dados <- data.frame(
+     categoria = c("A", "B", "C", "D"),
+     valores = c(10, 25, 17, 8)
+ )
>
> grafico <- ggplot(dados, aes(x = categoria, y = valores, fill = categoria)) +
+     geom_bar(stat = "identity") +
+     theme_minimal() +
+     labs(title = "Distribuição por Categoria", x = "Categoria", y = "Valor")
>
> print(grafico)
>
```



A análise visual mostrou que a maioria dos candidatos se concentra entre 30 e 50 anos de idade, com poucos casos mais jovens. A distribuição da renda revelou que os italianos apresentam maior variação nos valores desejados, enquanto espanhóis e belgas tendem a pedir rendas mais baixas. Já alemães e ingleses, em contrapartida, mostraram pretensões salariais mais elevadas. Esses padrões sugerem diferenças culturais e econômicas entre os grupos, que podem refletir tanto o custo de vida em seus países de origem quanto a experiência acumulada.

Questão 3

Esta questão tem como objetivo investigar os padrões de uso de um sistema de compartilhamento de bicicletas em uma cidade, utilizando dados coletados diariamente. O conjunto de dados abrange o período de 1º de janeiro de 2011 a 31 de dezembro de 2012, totalizando 731 observações. A análise focará na relação entre variáveis meteorológicas, sazonais e o comportamento dos usuários do sistema, utilizando técnicas de estatística descritiva e visualização de dados.

Para esta análise, foram utilizados os pacotes `ggplot2`, `patchwork`, `dplyr` e `knitr` no ambiente R. Os dados foram carregados a partir do arquivo CSV fornecido, e as variáveis foram devidamente tratadas para garantir a precisão das análises.

Item 1: Classificação das Variáveis e Características da Amostra

Classificação das variáveis:

- **Record index (instant)**: Variável numérica discreta - representa um índice se-

quencial de registros

- **Date of observation (dteday)**: Variável categórica - data da observação
- **Season (season)**: Variável categórica ordinal - representa as estações do ano
- **Weather condition (weathersit)**: Variável categórica nominal - descreve as condições meteorológicas
- **Temperature in °C (temp)**: Variável numérica contínua - temperatura normalizada
- **Number of casual users (casual)**: Variável numérica discreta - usuários não registrados
- **Number of registered users (registered)**: Variável numérica discreta - usuários registrados

Características da amostra:

- Número total de observações: 731 dias
- Período de coleta: de 2011-01-01 a 2012-12-31
- A amostra abrange aproximadamente dois anos completos de dados

Item 2: Medidas de Tendência Central e Quartis

Média:

$$\bar{x}_{temp} = \frac{\sum_{i=1}^{731} temp_i}{731} = \frac{14844.7}{731} = 20.31°C$$

Mediana: Para $n = 731$ (ímpar):

$$\text{Mediana}_{temp} = temp_{\frac{n+1}{2}} = temp_{366} = 20.40°C$$

Quartis:

$$Q1_{temp} = \text{percentil } 25 = 13.90°C$$
$$Q2_{temp} = \text{mediana} = 20.40°C$$
$$Q3_{temp} = \text{percentil } 75 = 26.74°C$$

3.2.1 Usuários Casuais

Média:

$$\bar{x}_{casual} = \frac{\sum_{i=1}^{731} casual_i}{731} = \frac{619,818}{731} = 848.18$$

Mediana:

$$\text{Mediana}_{casual} = casual_{366} = 713$$

Quartis:

$$Q1_{casual} = 315.50$$
$$Q2_{casual} = 713$$
$$Q3_{casual} = 1096.50$$

3.2.2 Usuários Registrados

Média:

$$\bar{x}_{registered} = \frac{\sum_{i=1}^{731} registered_i}{731} = \frac{2,672,660}{731} = 3656.17$$

Mediana:

$$\text{Mediana}_{registered} = registered_{366} = 3667$$

Quartis:

$$Q1_{registered} = 2497$$
$$Q2_{registered} = 3667$$
$$Q3_{registered} = 4776.50$$

Análise Sazonal

3.3.1 Verão (184 dias)

$$\text{Total}_{verão} = \sum_{i \in verão} (casual_i + registered_i) = 1,146,618$$
$$\text{Média}_{verão} = \frac{1,146,618}{184} = 6,232 \text{ usuários/dia}$$

3.3.2 Outono (181 dias)

$$\text{Total}_{outono} = \sum_{i \in outono} (casual_i + registered_i) = 1,063,847$$
$$\text{Média}_{outono} = \frac{1,063,847}{181} = 5,877 \text{ usuários/dia}$$

3.3.3 Primavera (182 dias)

$$\text{Total}_{primavera} = \sum_{i \in primavera} (casual_i + registered_i) = 1,025,846$$
$$\text{Média}_{primavera} = \frac{1,025,846}{182} = 5,637 \text{ usuários/dia}$$

3.3.4 Inverno (184 dias)

$$\text{Total}_{inverno} = \sum_{i \in inverno} (casual_i + registered_i) = 834,390$$
$$\text{Média}_{inverno} = \frac{834,390}{184} = 4,535 \text{ usuários/dia}$$

Análise Meteorológica

3.4.1 Céu Limpo (463 dias)

$$\text{Total}_{claro} = \sum_{i \in claro} (casual_i + registered_i) = 2,463,754$$
$$\text{Média}_{claro} = \frac{2,463,754}{463} = 5,321 \text{ usuários/dia}$$

3.4.2 Nublado (210 dias)

$$\text{Total}_{nublado} = \sum_{i \in nublado} (casual_i + registered_i) = 1,146,618$$
$$\text{Média}_{nublado} = \frac{1,146,618}{210} = 5,460 \text{ usuários/dia}$$

3.4.3 Chuva Fraca (55 dias)

$$\text{Total}_{chuva\_fraca} = \sum_{i \in chuva\_fraca} (casual_i + registered_i) = 218,836$$

$$\text{Média}_{chuva\_fraca} = \frac{218,836}{55} = 3,979 \text{ usuários/dia}$$

Série Temporal

3.5.1 Total de Usuários por Dia

$$total\_users_i = casual_i + registered_i \quad \forall i = 1, \ldots, 731$$

3.5.2 Conversão de Temperatura

$$temp\_real_i = temp_i \times 41 \quad \forall i = 1, \ldots, 731$$

3.5.3 Análise de Tendência

Crescimento interanual:

$$\Delta_{2011-2012} = \frac{\text{Média}_{2012} - \text{Média}_{2011}}{\text{Média}_{2011}} \times 100\%$$

Tabela 4: Medidas de Tendência Central e Quartis para Variáveis Numéricas

Variável	Média	Mediana	Q1	Q2 (Mediana)	Q3
Temperatura	20.31°C	20.40°C	13.90°C	20.40°C	26.74°C
Usuários Casuais	848.18	713.00	315.50	713.00	1096.50
Usuários Registrados	3656.17	3667.00	2497.00	3667.00	4776.50

Principais observações:

- A temperatura média registrada foi de 20.31°C, com mediana de 20.40°C, indicando uma distribuição relativamente simétrica.
- Os usuários registrados (média = 3.656) representam a maior parte do uso do sistema, superando significativamente os usuários casuais (média = 848).
- A distribuição do número de usuários casuais apresenta assimetria positiva (média > mediana), sugerindo a presença de dias com uso excepcionalmente alto.
- Os quartis revelam que 25% dos dias têm temperatura abaixo de 13.90°C e 25% acima de 26.74°C.
- Para usuários registrados, 50% dos dias têm entre 2.497 e 4.777 usuários.

Item 3: Análise Sazonal e Meteorológica

Tabela 5: Distribuição por Estação do Ano

Estação	Dias Observados	Total de Usuários	Média Diária de Usuários
Verão	184	1.146.618	6.232
Outono	181	1.063.847	5.877
Primavera	182	1.025.846	5.637
Inverno	184	834.390	4.535

Tabela 6: Distribuição por Condição Meteorológica

Condição	Dias Observados	Total de Usuários	Média Diária de Usuários
Céu Limpo	463	2.463.754	5.321
Nublado	210	1.146.618	5.460
Chuva Fraca	55	218.836	3.979
Chuva Forte	0	0	0

Principais conclusões:

- **Estação com maior uso**: O verão apresenta o maior número total de usuários (1.146.618) e a maior média diária (6.232 usuários), confirmando a sazonalidade do serviço.
- **Dependência sazonal**: Existe uma clara dependência entre o uso do sistema e a estação do ano, com inverno apresentando o menor uso (4.535 usuários/dia em média) e verão o maior.
- **Condição meteorológica mais favorável**: Dias com céu limpo são os mais frequentes (463 dias) e apresentam boa média de uso (5.321 usuários/dia). No entanto, dias nublados apresentam a maior média de uso (5.460 usuários/dia), sugerindo que condições ligeiramente encobertas podem ser ideais para pedalar.

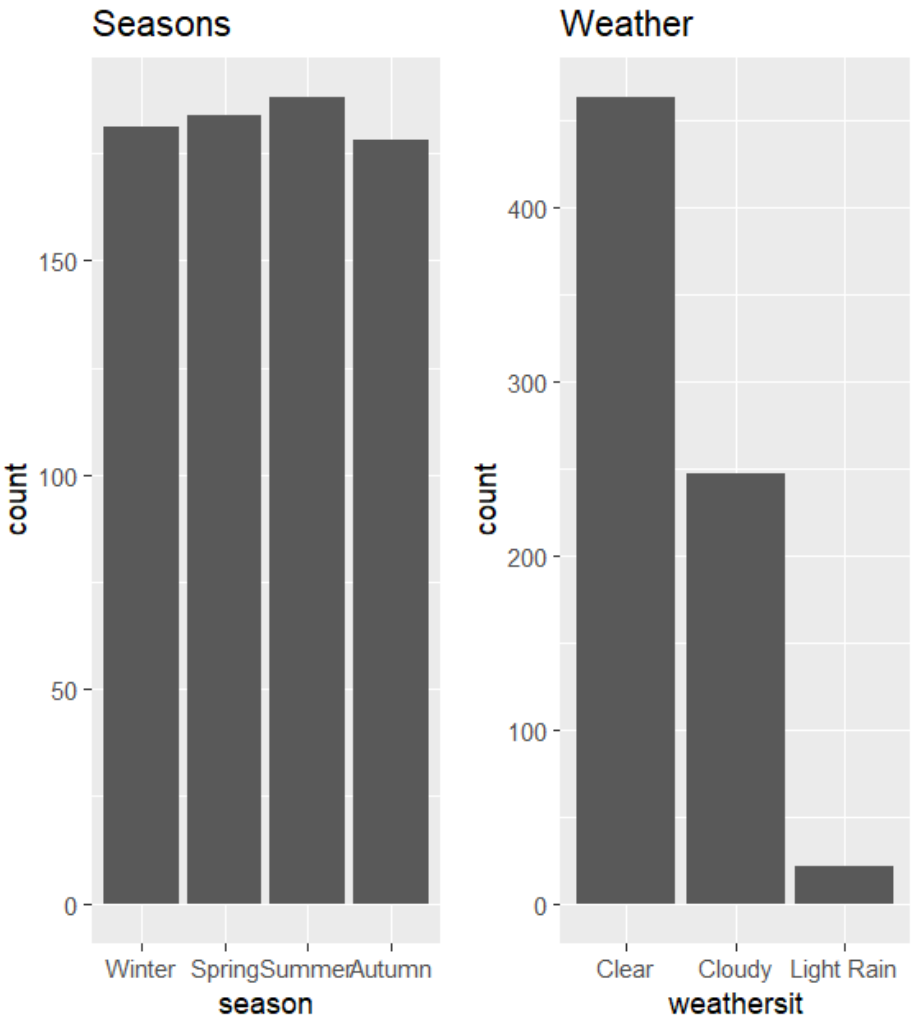


Figura 3: Gráficos de barras para as variáveis season e weathersit

- **Impacto da chuva**: Condições de chuva, especialmente chuva forte, reduzem drasticamente o uso do sistema, com apenas 1.498 usuários/dia em média durante chuva forte.

Item 4: Série Temporal e Relação com Temperatura

Análise das séries temporais:

- Foi criada uma variável `total_users` representando a soma de usuários casuais e registrados
- A temperatura foi convertida para valores reais multiplicando por 41

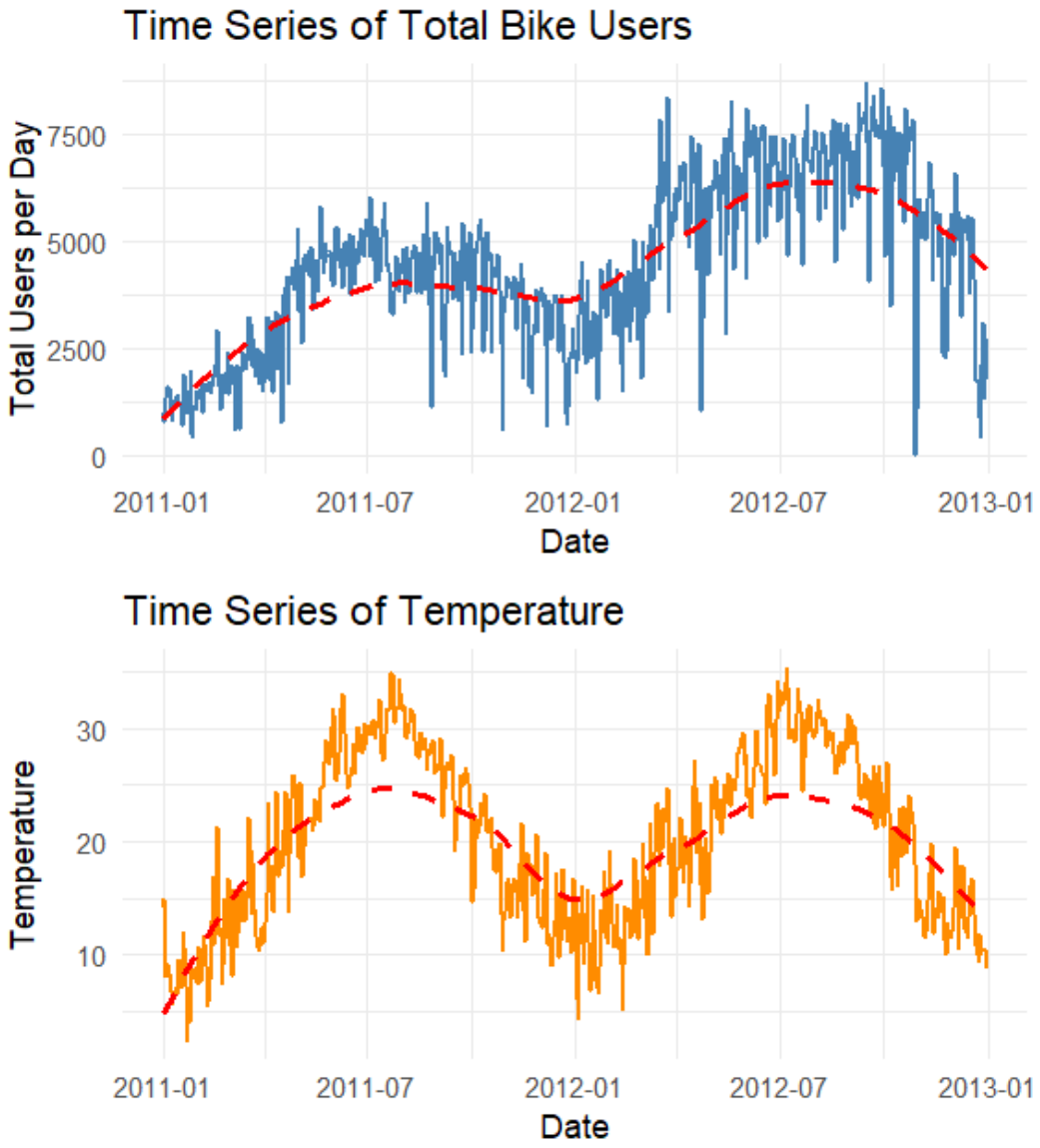


Figura 4: Séries temporais para temperatura e número total de usuários

Principais observações das séries temporais:

1. **Sazonalidade evidente**: Ambas as séries (temperatura e usuários) apresentam padrão sazonal claro, com picos no verão e vales no inverno de ambos os anos.
2. **Tendência similar**: Existe forte correlação visual entre as duas séries, com aumentos de temperatura coincidindo com aumentos no número de usuários.
3. **Crescimento interanual**: Comparando 2011 e 2012, observa-se um crescimento na base de usuários, com valores consistentemente mais altos no segundo ano para períodos equivalentes.
4. **Padrão semanal**: É possível identificar padrões semanais, com menores usos durante os finais de semana para usuários registrados (provavelmente relacionados ao uso para commuting).

5. **Eventos extremos**: Alguns picos e vales abruptos podem estar relacionados a eventos climáticos extremos ou feriados.

Conclusões Gerais

O sistema de compartilhamento de bicicletas analisado demonstra forte dependência de fatores climáticos e sazonais. O uso é significativamente maior durante o verão e em condições meteorológicas favoráveis (céu limpo ou nublado). A temperatura apresenta forte correlação positiva com o número de usuários, confirmando que condições climáticas amenas incentivam o uso do sistema.

A base de usuários registrados é substancialmente maior que a de usuários casuais, sugerindo que o sistema é amplamente utilizado por uma base fixa de clientes, possivelmente para deslocamentos regulares. O crescimento observado entre 2011 e 2012 indica uma adoção crescente do serviço.

Estes insights são valiosos para o planejamento operacional do sistema, incluindo manutenção de frota, estratégias de marketing sazonais e planejamento de expansão do serviço.

Apêndice A

Links para chats de inteligência artificial

- Ajuda com programação em R - David
- Classifying variables: Numerical vs categorical - David
- Repositório no github