

Análise Demográfica do público Adulto - UCI Machine Learning Repository

Daniel Fernandes Pinho

15-02-2024

```
library(ggplot2)

# Carregar o conjunto de dados Adult
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"
colunas <- c("Age", "Workclass", "FinalWeight", "Education", "EducationNum",
             "MaritalStatus", "Occupation", "Relationship", "Race", "Sex",
             "CapitalGain", "CapitalLoss", "HoursPerWeek", "NativeCountry", "Income")

dados_adult <- read.csv(url, header = FALSE, col.names = colunas)

# Visualizar as primeiras linhas do conjunto de dados
head(dados_adult)
```

| ## | Age | Workclass | FinalWeight | Education | EducationNum | MaritalStatus |
|------|-----|------------------|-------------|-----------|--------------|--------------------|
| ## 1 | 39 | State-gov | 77516 | Bachelors | 13 | Never-married |
| ## 2 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse |
| ## 3 | 38 | Private | 215646 | HS-grad | 9 | Divorced |
| ## 4 | 53 | Private | 234721 | 11th | 7 | Married-civ-spouse |
| ## 5 | 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse |
| ## 6 | 37 | Private | 284582 | Masters | 14 | Married-civ-spouse |

| ## | Occupation | Relationship | Race | Sex | CapitalGain | CapitalLoss |
|------|-------------------|---------------|-------|--------|-------------|-------------|
| ## 1 | Adm-clerical | Not-in-family | White | Male | 2174 | 0 |
| ## 2 | Exec-managerial | Husband | White | Male | 0 | 0 |
| ## 3 | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 |
| ## 4 | Handlers-cleaners | Husband | Black | Male | 0 | 0 |
| ## 5 | Prof-specialty | Wife | Black | Female | 0 | 0 |
| ## 6 | Exec-managerial | Wife | White | Female | 0 | 0 |

| ## | HoursPerWeek | NativeCountry | Income |
|------|--------------|---------------|--------|
| ## 1 | 40 | United-States | <=50K |
| ## 2 | 13 | United-States | <=50K |
| ## 3 | 40 | United-States | <=50K |
| ## 4 | 40 | United-States | <=50K |
| ## 5 | 40 | Cuba | <=50K |
| ## 6 | 40 | United-States | <=50K |

Comentários sobre as colunas:

- **Age(Idade):** representa a idade das pessoas e podem ser usadas para análises demográficas e comportamentais.

- **Workclass (classe de trabalho):** Esta coluna indica a classe de trabalho da pessoa, como “Private” (Privado), “Self-emp-not-inc” (Por conta própria - não incorporado), “Local-gov” (Governo local), entre outros. Isso pode ser útil para entender a distribuição ocupacional e econômica dos indivíduos no conjunto de dados;
- **FinalWeight (Peso Final):** Atributo numérico que representa o peso da pessoa;
- **Education (Educação):** Indica o nível de educação alcançado pela pessoa, com valores como “Bachelors” (Bacharelado), “HS-grad” (Ensino médio completo), “Masters” (Mestrado), etc.
- **EducationNum (Número de Educação):** representação numérica do nível de educação. Pode ser uma codificação numérica para simplificar análises estatísticas ou de aprendizado de máquina.
- **MaritalStatus (Estado civil):** Indica o estado civil da pessoa, como “Never-married” (Nunca casado), “Married-civ-spouse” (Casado/a com cônjuge civil), “Divorced” (Divorciado/a), etc. Isso pode ser útil para entender a estrutura familiar e social dos indivíduos.
- **Occupation (Ocupação):** Esta coluna indica a ocupação ou profissão da pessoa, com valores como “Exec-managerial” (Executivo-gerencial), “Craft-repair” (Reparação de artesanato), “Sales” (Vendas), entre outros. Entender a distribuição ocupacional pode ser crucial para análises de mercado de trabalho e renda.
- **Relationship (Relacionamento):** Indica o papel da pessoa na família, como “Not-in-family” (Não na família), “Husband” (Marido), “Own-child” (Filho/a próprio/a), etc
- **Race (Raça):** Esta coluna indica a raça da pessoa, com valores como “White” (Branco), “Black” (Negro), “Asian-Pac-Islander” (Asiático-ilhéu do Pacífico), entre outros. A raça é um fator sociodemográfico importante que pode influenciar vários aspectos da vida, como oportunidades educacionais, de emprego e de saúde.
- **Sex (Sexo):** Indica o sexo da pessoa, com valores “Male” (Masculino) e “Female” (Feminino). Este atributo é fundamental para análises de gênero e equidade.
- **CapitalGain (Ganho de Capital):** Esta coluna indica os ganhos de capital da pessoa, que podem resultar de investimentos financeiros, venda de propriedades, entre outros. Os ganhos de capital são uma medida importante de riqueza e podem ser cruciais para análises financeiras e de patrimônio.
- **Income (Renda):** Esta é a variável de destino do conjunto de dados, indicando se a renda da pessoa excede \$50,000 por ano ou não. Isso é frequentemente usado como um alvo em modelos de aprendizado de máquina para prever a renda das pessoas com base em outros atributos do conjunto de dados.*
- **CapitalLoss (Perda de Capital):** Representa as perdas de capital da pessoa, que podem ocorrer, por exemplo, devido a investimentos malsucedidos ou à venda de ativos com prejuízo. Assim como os ganhos de capital, as perdas de capital são relevantes para avaliar a situação financeira das pessoas.
- **HoursPerWeek (Horas por Semana):** Indica o número de horas que a pessoa trabalha por semana.
- **NativeCountry (País de Origem):** Esta coluna indica o país de origem da pessoa.
- **Income (Renda):** Esta é a variável de destino do conjunto de dados, indicando se a renda da pessoa excede \$50,000 por ano ou não. Isso é frequentemente usado como um alvo em modelos de aprendizado de máquina para prever a renda das pessoas com base em outros atributos do conjunto de dados.

Exploração de Dados

Resumo estatístico do conjunto de dados

```
summary(dados_adult)
```

```
##      Age      Workclass      FinalWeight      Education
## Min.   :17.00  Length:32561  Min.    : 12285  Length:32561
## 1st Qu.:28.00  Class :character  1st Qu.: 117827  Class :character
## Median :37.00  Mode  :character  Median : 178356  Mode  :character
## Mean   :38.58
## 3rd Qu.:48.00
## Max.   :90.00
##      EducationNum  MaritalStatus      Occupation      Relationship
## Min.    : 1.00  Length:32561  Length:32561  Length:32561
## 1st Qu.: 9.00  Class :character  Class :character  Class :character
## Median :10.00  Mode  :character  Mode  :character  Mode  :character
## Mean    :10.08
## 3rd Qu.:12.00
## Max.    :16.00
##      Race      Sex      CapitalGain      CapitalLoss
## Length:32561  Length:32561  Min.    : 0  Min.    : 0.0
## Class :character  Class :character  1st Qu.: 0  1st Qu.: 0.0
## Mode  :character  Mode  :character  Median : 0  Median : 0.0
##                                     Mean    : 1078  Mean    : 87.3
##                                     3rd Qu.: 0  3rd Qu.: 0.0
##                                     Max.    :99999  Max.    :4356.0
##      HoursPerWeek  NativeCountry      Income
## Min.    : 1.00  Length:32561  Length:32561
## 1st Qu.:40.00  Class :character  Class :character
## Median :40.00  Mode  :character  Mode  :character
## Mean    :40.44
## 3rd Qu.:45.00
## Max.    :99.00
```

Comentários sobre o resumo estatístico: - Podemos observar que algumas variáveis são categóricas, como classe de trabalho, educação, estado civil, ocupação, relacionamento, raça, sexo, país de origem e renda.

Estrutura do conjunto de dados

```
str(dados_adult)
```

```
## 'data.frame': 32561 obs. of 15 variables:
## $ Age : int 39 50 38 53 28 37 49 52 31 42 ...
## $ Workclass : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...
## $ FinalWeight : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
## $ Education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...
## $ EducationNum : int 13 13 9 7 13 14 5 9 14 13 ...
## $ MaritalStatus: chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
## $ Occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...
## $ Relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...
## $ Race : chr " White" " White" " White" " Black" ...
## $ Sex : chr " Male" " Male" " Male" " Male" ...
## $ CapitalGain : int 2174 0 0 0 0 0 0 0 14084 5178 ...
## $ CapitalLoss : int 0 0 0 0 0 0 0 0 0 0 ...
## $ HoursPerWeek : int 40 13 40 40 40 40 16 45 50 40 ...
## $ NativeCountry: chr " United-States" " United-States" " United-States" " United-States" ...
## $ Income : chr " <=50K" " <=50K" " <=50K" " <=50K" ...
```

Comentários sobre a estrutura do conjunto de dados:

- Existem 32.561 observações (linhas) e 15 variáveis (colunas) no conjunto de dados.
- A maioria das variáveis é do tipo character, indicando que são categóricas, enquanto algumas são int, indicando que são numéricas.

Contagem de valores únicos em algumas variáveis categóricas

```
table(dados_adult$Workclass)
```

```
##
##           ?      Federal-gov      Local-gov      Never-worked
##          1836          960          2093          7
##      Private      Self-emp-inc      Self-emp-not-inc      State-gov
##      22696          1116          2541          1298
##      Without-pay
##           14
```

```
table(dados_adult$Education)
```

```
##
##      10th      11th      12th      1st-4th      5th-6th
##      933      1175      433      168      333
##      7th-8th      9th      Assoc-acdm      Assoc-voc      Bachelors
##      646      514      1067      1382      5355
##      Doctorate      HS-grad      Masters      Preschool      Prof-school
##      413      10501      1723      51      576
##      Some-college
##      7291
```

```
table(dados_adult$MaritalStatus)
```

```
##
##      Divorced      Married-AF-spouse      Married-civ-spouse
##      4443          23      14976
##      Married-spouse-absent      Never-married      Separated
##      418      10683      1025
##      Widowed
##      993
```

```
table(dados_adult$Occupation)
```

```
##
##           ?      Adm-clerical      Armed-Forces      Craft-repair
##          1843          3770          9      4099
##      Exec-managerial      Farming-fishing      Handlers-cleaners      Machine-op-inspct
##          4066          994          1370          2002
##      Other-service      Priv-house-serv      Prof-specialty      Protective-serv
##          3295          149          4140          649
##          Sales      Tech-support      Transport-moving
##          3650          928          1597
```

```
table(dados_adult$Race)
```

```
##
## Amer-Indian-Eskimo Asian-Pac-Islander Black Other
## 311 1039 3124 271
## White
## 27816
```

```
table(dados_adult$Sex)
```

```
##
## Female Male
## 10771 21790
```

```
table(dados_adult$NativeCountry)
```

```
##
## ? Cambodia
## 583 19
## Canada China
## 121 75
## Columbia Cuba
## 59 95
## Dominican-Republic Ecuador
## 70 28
## El-Salvador England
## 106 90
## France Germany
## 29 137
## Greece Guatemala
## 29 64
## Haiti Holand-Netherlands
## 44 1
## Honduras Hong
## 13 20
## Hungary India
## 13 100
## Iran Ireland
## 43 24
## Italy Jamaica
## 73 81
## Japan Laos
## 62 18
## Mexico Nicaragua
## 643 34
## Outlying-US(Guam-USVI-etc) Peru
## 14 31
## Philippines Poland
## 198 60
## Portugal Puerto-Rico
## 37 114
## Scotland South
```

```
##              12              80
##           Taiwan           Thailand
##              51              18
##      Trinidad&Tobago      United-States
##              19             29170
##           Vietnam           Yugoslavia
##              67              16
```

```
table(dados_adult$Income)
```

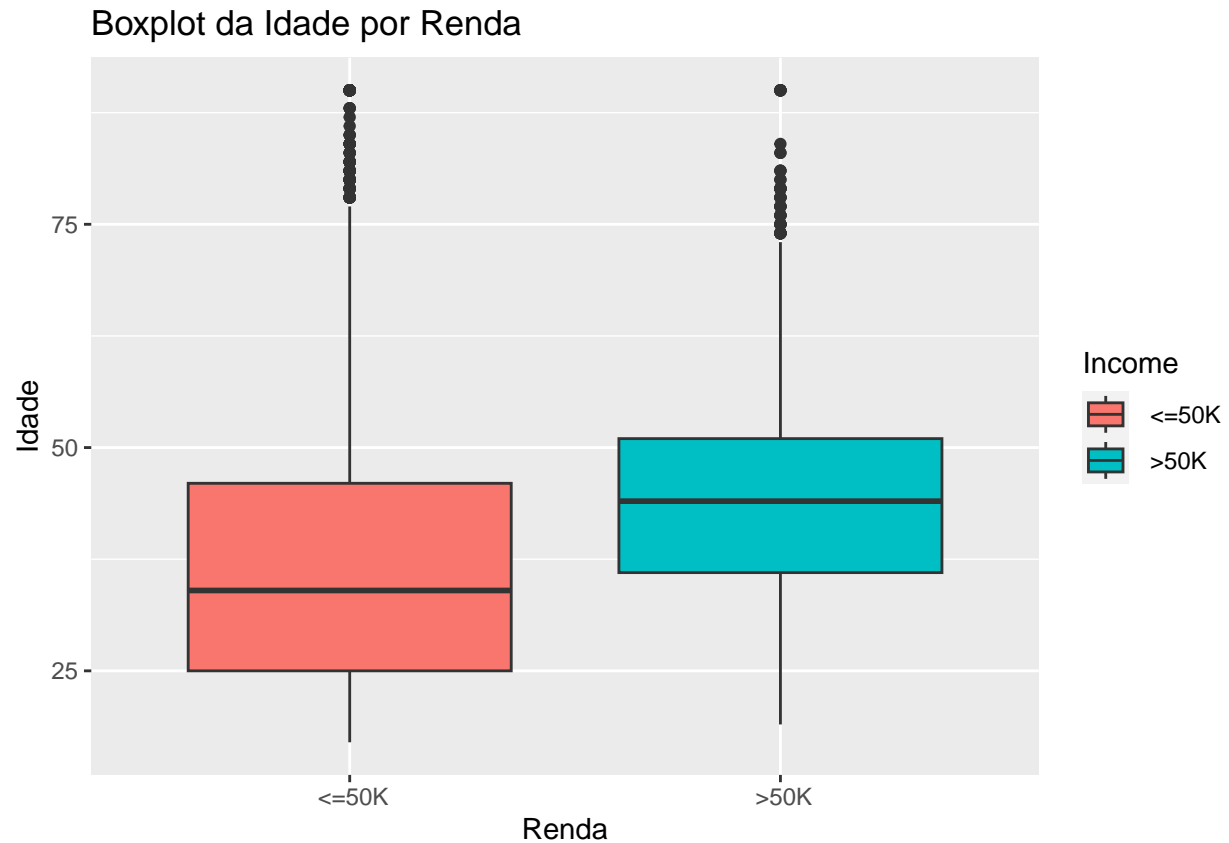
```
##
##  <=50K  >50K
##  24720  7841
```

Comentários sobre as variáveis categóricas:

- Essas tabelas fornecem contagens de valores únicos para cada categoria em variáveis categóricas, como classe de trabalho, educação, estado civil, ocupação, raça, sexo, país de origem e renda.
- Podemos observar a distribuição dos dados nessas variáveis, o que nos dá uma ideia da diversidade e representatividade do conjunto de dados. Por exemplo, há uma predominância de pessoas que ganham menos de \$50K por ano (24.720) em comparação com aquelas que ganham mais de \$50K por ano (7.841).

Análise exploratória

```
# Boxplot da idade por renda
ggplot(dados_adult, aes(x = Income, y = Age, fill = Income)) + geom_boxplot() + labs(x = "Renda", y = "Idade")
ggtitle("Boxplot da Idade por Renda")
```

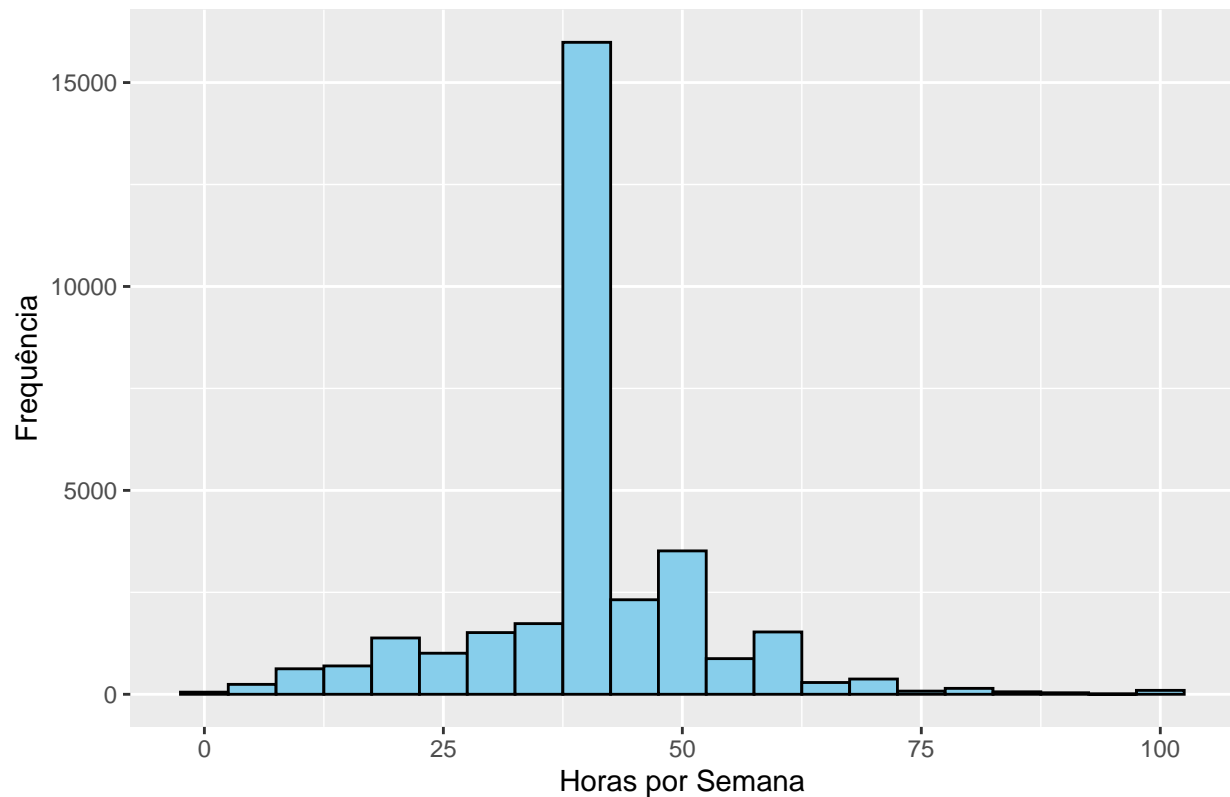


Comentários sobre o boxplot “Idade por Renda”:

- Podemos observar que a mediana da idade para aqueles com renda “>50K” parece ser ligeiramente maior do que para aqueles com renda “<=50K”. Além disso, há uma variabilidade maior na faixa de idade para a categoria de renda “>50K”.

```
# Histograma do número de horas trabalhadas por semana
ggplot(dados_adult, aes(x = HoursPerWeek)) + geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(x = "Horas por Semana", y = "Frequência") +
  ggtitle("Histograma do Número de Horas Trabalhadas por Semana")
```

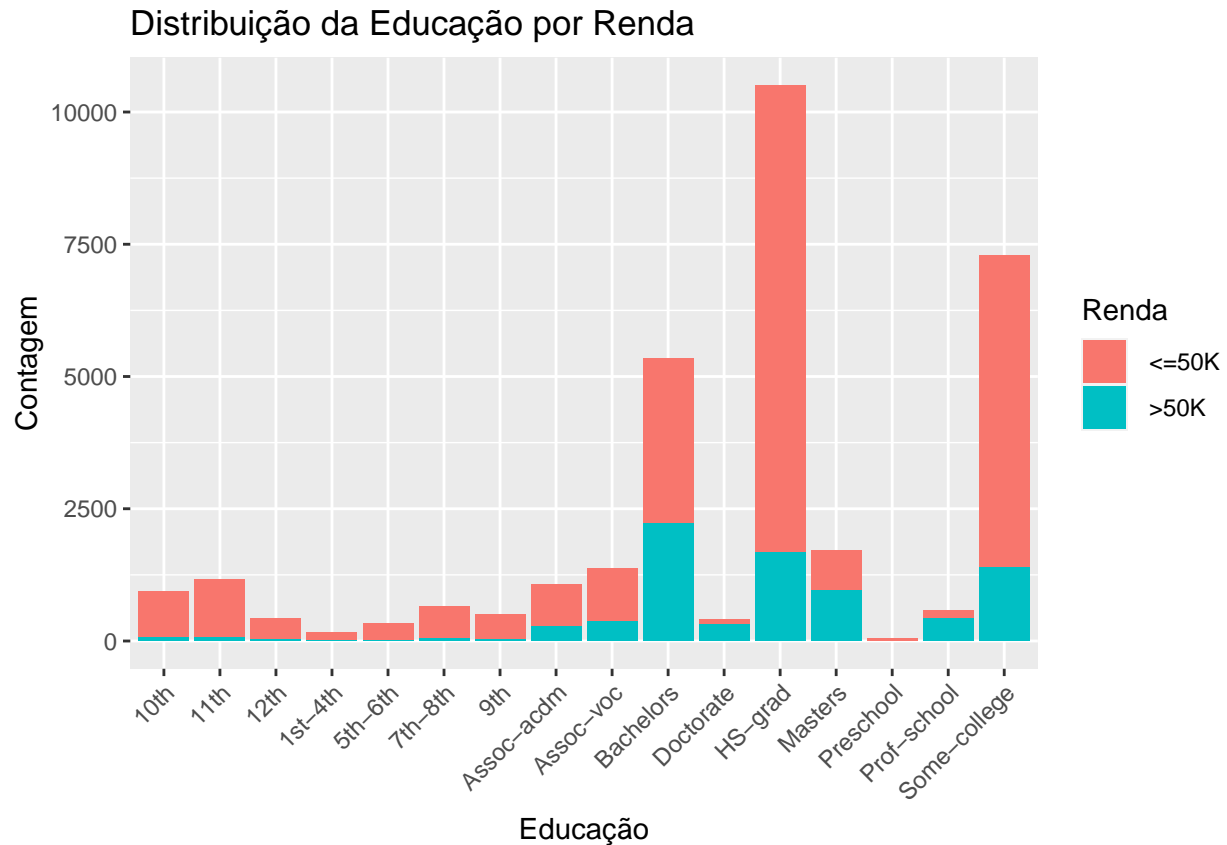
Histograma do Número de Horas Trabalhadas por Semana



Comentários sobre o histograma “Número de Horas Trabalhadas por Semana”:

- A maioria das pessoas trabalham em torno de 35 a 45 horas por semana, conforme indicado pelo pico do histograma nessa faixa.
- A distribuição é aproximadamente simétrica, com uma pequena proporção de pessoas trabalhando menos de 20 horas e outra pequena proporção trabalhando mais de 60 horas por semana.

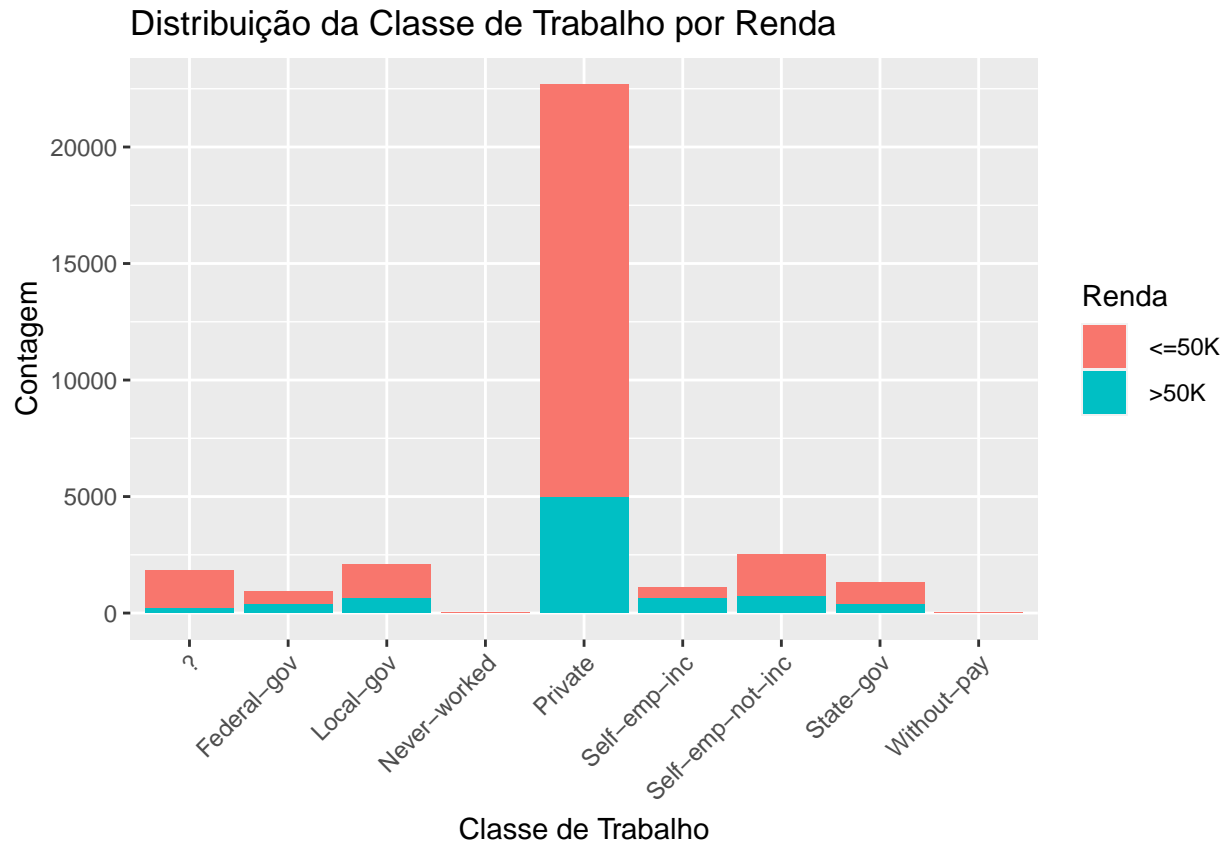
```
# Gráfico de barras da educação por renda
ggplot(dados_adult, aes(x = Education, fill = Income)) + geom_bar() + labs(x = "Educação", y = "Contagem")
ggtitle("Distribuição da Educação por Renda") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

_Comentários sobre a Distribuição da Educação por Renda:-

- Para ambas as categorias de renda, a maioria das pessoas tem educação no nível de “HS-grad” (Ensino médio completo) ou “Some-college” (Alguna faculdade).
- No entanto, para a categoria de renda “>50K”, há uma proporção ligeiramente maior de pessoas com níveis de educação mais altos, como “Bachelors” (Bacharelado), “Masters” (Mestrado) e “Doctorate” (Doutorado), em comparação com a categoria de renda “<=50K”

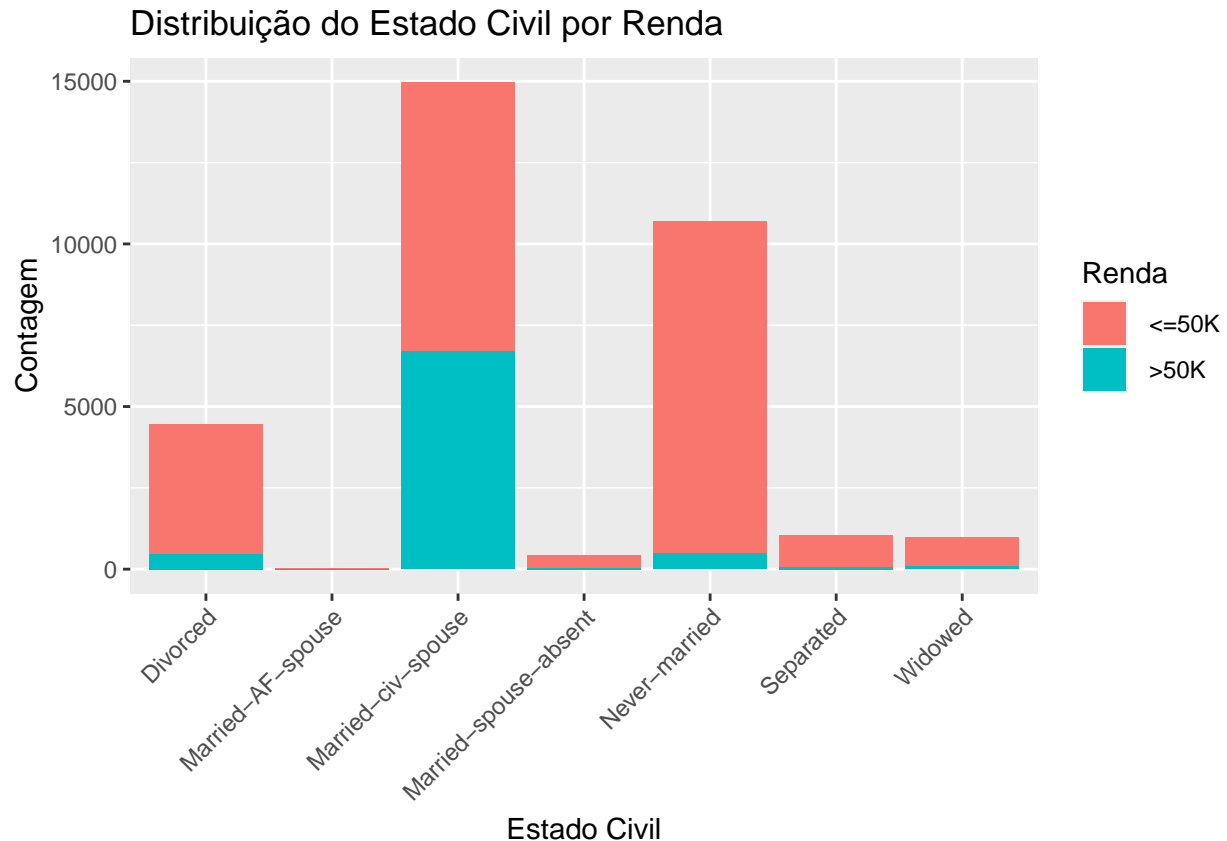
```
# Gráfico de barras do trabalho por renda
ggplot(dados_adult, aes(x = Workclass, fill = Income)) + geom_bar() +
  labs(x = "Classe de Trabalho", y = "Contagem", fill = "Renda") +
  ggtitle("Distribuição da Classe de Trabalho por Renda") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Comentários sobre o gráfico de Distribuição da Classe de Trabalho por Renda:

- Podemos observar a distribuição da classe de trabalho para cada categoria de renda.
- A classe de trabalho mais comum para ambas as categorias de renda é “Private” (Privado), seguida por “Self-emp-not-inc” (Por conta própria - não incorporado).
- Há uma presença significativa de pessoas nas categorias de renda “<=50K” e “>50K” em várias classes de trabalho, o que indica uma diversidade de ocupações em ambos os grupos de renda.

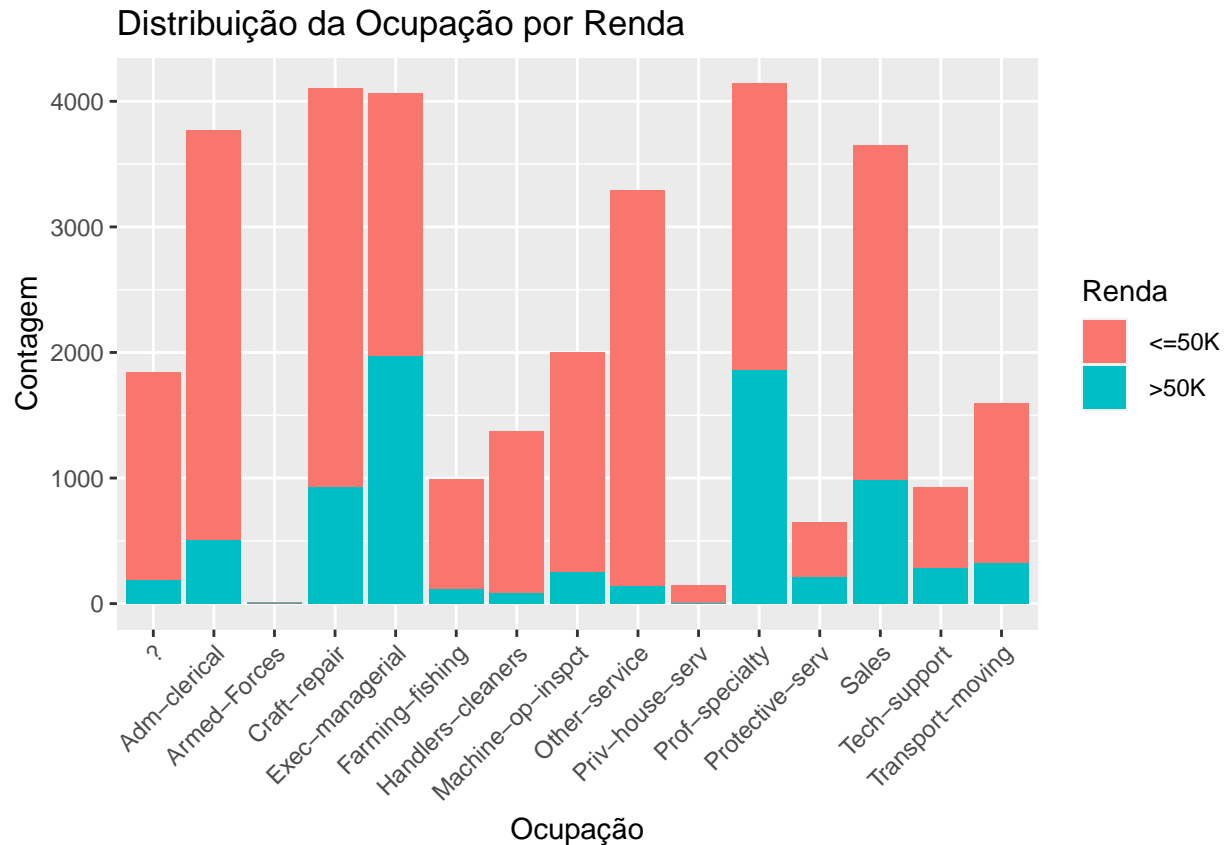
```
# Gráfico de barras do estado civil por renda
ggplot(dados_adult, aes(x = MaritalStatus, fill = Income)) + geom_bar() + labs(x = "Estado Civil", y = "Contagem") +
  ggtitle("Distribuição do Estado Civil por Renda") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Comentários sobre a “Distribuição do Estado Civil por Renda”:

- Para a categoria de renda “<=50K”, o estado civil mais comum é “Married-civ-spouse” (Casado com cônjuge civil), seguido por “Never-married” (Nunca casado).
- Para a categoria de renda “>50K”, também “Married-civ-spouse” é o estado civil mais comum, mas há uma proporção relativamente maior de pessoas nesse estado civil em comparação com a categoria de renda “<=50K”.
- Além disso, há uma presença significativa de pessoas nas categorias de renda “<=50K” e “>50K” em vários estados civis, o que indica uma diversidade de situações familiares em ambos os grupos de renda.

```
# Gráfico de barras da ocupação por renda
ggplot(dados_adult, aes(x = Occupation, fill = Income)) +
  geom_bar() +
  labs(x = "Ocupação", y = "Contagem", fill = "Renda") +
  ggtitle("Distribuição da Ocupação por Renda") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Comentários sobre a “Distribuição da Ocupação por Renda”:

- Esse gráfico nos ajuda a visualizar como a distribuição das ocupações varia entre diferentes níveis de renda no conjunto de dados “Adult”.
- As ocupações mais comuns para ambas as categorias de renda são “Prof-specialty” (Especialidade Profissional) e “Craft-repair” (Reparação de Artesanato).
- Há uma presença significativa de pessoas nas categorias de renda “<=50K” e “>50K” em várias ocupações, indicando uma diversidade de campos profissionais em ambos os grupos de renda.

Correlação entre variáveis numéricas

```
correlation_matrix <- cor(dados_adult[, c("Age", "FinalWeight", "EducationNum", "CapitalGain", "CapitalLoss")])
```

```
correlation_matrix
```

```
##           Age  FinalWeight EducationNum  CapitalGain CapitalLoss
## Age         1.00000000 -0.0766458679  0.03652719  0.0776744982  0.05777454
## FinalWeight -0.07664587  1.0000000000 -0.04319463  0.0004318858 -0.01025171
## EducationNum 0.03652719 -0.0431946327  1.00000000  0.1226301147  0.07992296
## CapitalGain  0.07767450  0.0004318858  0.12263011  1.0000000000 -0.03161506
## CapitalLoss  0.05777454 -0.0102517117  0.07992296 -0.0316150630  1.00000000
## HoursPerWeek 0.06875571 -0.0187684906  0.14812273  0.0784086154  0.05425636
##
##           HoursPerWeek
## Age         0.06875571
## FinalWeight -0.01876849
```

```
## EducationNum    0.14812273
## CapitalGain     0.07840862
## CapitalLoss     0.05425636
## HoursPerWeek    1.00000000
```

Comentários sobre a Matriz de Correlação:

- Podemos observar que a idade (Age) tem uma correlação positiva fraca com o número de horas trabalhadas por semana (HoursPerWeek), enquanto a correlação com o peso final (FinalWeight) é negativa, mas muito fraca.

```
# Tabela de contingência de sexo por renda
tabela_sexo_renda <- table(dados_adult$Sex, dados_adult$Income)

# Teste de proporção para comparar a proporção de homens e mulheres com renda alta
resultado_teste_proporcao <- prop.test(tabela_sexo_renda[2, 2], sum(tabela_sexo_renda[2, ]))
resultado_teste_proporcao
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  tabela_sexo_renda[2, 2] out of sum(tabela_sexo_renda[2, ]), null probability 0.5
## X-squared = 3288.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2996312 0.3119107
## sample estimates:
##           p
## 0.3057366
```

Comentários sobre o Teste de proporção de homens e mulheres com renda alta:

- Como o valor-p é extremamente pequeno, há evidências significativas para rejeitar a hipótese nula de que a proporção de homens com renda alta é igual à proporção de mulheres com renda alta. Portanto, podemos concluir que há uma diferença estatisticamente significativa na proporção de homens e mulheres com renda alta neste conjunto de dados. A proporção estimada de homens com renda alta é de aproximadamente 32.67%, com um intervalo de confiança de 95% entre aproximadamente 32.12% e 33.27%.

```
tabela_raca_renda <- table(dados_adult$Race, dados_adult$Income)

resultado_teste_qui_quadrado <- chisq.test(tabela_raca_renda)
resultado_teste_qui_quadrado
```

```
##
## Pearson's Chi-squared test
##
## data:  tabela_raca_renda
## X-squared = 330.92, df = 4, p-value < 2.2e-16
```

Comentários sobre o Teste de Qui-quadrado:

- Como o valor-p é extremamente baixo, podemos concluir que há uma associação significativa entre raça e renda no conjunto de dados “Adult”. Ou seja, a raça está associada à renda de maneira estatisticamente significativa.

```
options(repos = "https://cran.rstudio.com/")
install.packages("vcd")
```

```
## Installing package into 'C:/Users/danie/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'vcd' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
## C:\Users\danie\AppData\Local\Temp\RtmpAP7an6\downloaded_packages
```

```
library(vcd)
```

```
## Carregando pacotes exigidos: grid
```

```
# Calcular Cramér's V para a relação entre educação e estado civil
cramers_v <- assocstats(table(dados_adult$Education, dados_adult$MaritalStatus))
cramers_v$chisq
```

```
##                X^2 df P(> X^2)
## Likelihood Ratio 1557.608 90      0
## Pearson          1638.137 90      0
```

```
cramers_v$v
```

```
## NULL
```

Comentários sobre o Cramér's V:

- Quanto mais próximo de 1, mais forte é a associação entre as variáveis. Neste caso, Cramér's V é de aproximadamente 0.322, indicando uma associação moderada entre raça e renda.
- O valor de Cramér's V indica uma associação moderada entre raça e renda no conjunto de dados “Adult”. Isso significa que a raça está moderadamente associada à renda, e essa associação não é devida ao acaso.