**UNIVERSIDADE FEDERAL DA BAHIA**

**MESTRADO EM ADMINISTRAÇÃO**

**MÉTODOS QUANTITATIVOS**

**Prof. Dr. Fernando Antonio de Melo Pereira Lhamas**

**Roteiros estatísticos**

**Base de dados selecionada:**[Impacto das mídias sociais](https://www.kaggle.com/datasets/lastman0800/impact-of-social-media-dataset)

## Conjunto de dados

* ID do consumidor: um identificador exclusivo para cada consumidor.
* Nome do cliente: O nome do consumidor.
* Idade: A idade do consumidor.
* Gênero: O gênero do consumidor (masculino/feminino).
* Renda (USD): A renda anual do consumidor em dólares americanos.
* Nível de escolaridade: o nível mais alto de educação alcançado pelo consumidor (por exemplo, bacharelado, mestrado, doutorado, ensino médio).
* Uso de mídias sociais (horas/dia): número médio de horas por dia que o consumidor gasta em plataformas de mídia social.
* Plataformas de mídia social: as plataformas de mídia social que o consumidor usa, que podem incluir plataformas como Instagram, Facebook, TikTok, Twitter, Snapchat, LinkedIn e Pinterest.
* Nível de influência: o nível de influência que o consumidor tem nas mídias sociais (por exemplo, um tanto influente, muito influente, nada influente).
* Decisão de compra: se as decisões de compra do consumidor são influenciadas pelas mídias sociais (Sim/Não).
* Categoria do produto: a categoria de produtos em que o consumidor está interessado (por exemplo, eletrônicos, beleza, decoração para casa, moda).
* Produto específico: um produto específico dentro da categoria escolhida (por exemplo, smartphone, rímel, smartwatch, cortinas).
* Valor gasto (USD): Quantidade de dinheiro gasta pelo consumidor no produto específico em dólares americanos.
* Marca: A marca ou marcas preferidas pelo consumidor para o produto específico.
* Cidade: Cidade onde o consumidor está localizado.

O presente relatório tem como objetivo analisar um conjunto de dados que avalia o **impacto das mídias sociais** no comportamento de compra de consumidores. O dataset utilizado, denominado **Social Media Dataset**, engloba informações sobre uso de redes sociais, dados demográficos, renda, gastos, nível de influência e outros aspectos que possibilitam examinar as relações entre tais variáveis e o comportamento de consumo.

Para cumprir esse objetivo, serão apresentados cinco roteiros de análise:

**Roteiros de Análise:**

### 1. Exploração de Dados (AED)

Objetivo: identificar padrões iniciais e *outliers*, tratar inconsistências, além de gerar hipóteses iniciais sobre a relação entre algoritmos e comportamento de consumo.

* + Identificar e tratar outliers e dados ausentes
  + Recodificação de variáveis;
  + Tabelas de frequência;
  + Resultados descritivos (média, mediana, desvio, CV);
  + Cruzamentos e gráficos.
* **Importação e visualização dos dados**
* **Tratamento de dados ausentes**
  + identificar e documentar valores ausentes.
  + Decidir a estratégia: exclusão, imputação ou categorização como "desconhecido".
* **Análise descritiva**
  + Gerar tabelas de frequência, média, mediana, desvio padrão e coeficiente de variação.
  + Criar gráficos: histogramas, boxplots e scatterplots.
* **Identificação de outliers**
  + Boxplots e a regra dos 3 desvios padrão.
  + Decidir como tratar outliers (exclusão ou transformação).
* **Recodificação de variáveis**
  + Transformar variáveis ​​categóricas em variáveis ​​dummy.
  + Normalizar ou padronizar variáveis ​​contínuas, quando necessário.
* **Cruzamentos e gráficos diversos**

### 2. ANOVA (Análise de Variância)

**Objetivo:** Testar diferenças de média entre grupos.

* **Descrição do estudo**
* **Definição do problema e hipóteses**
  + Hipótese nula (H₀): As médias dos grupos são iguais.
  + Hipótese alternativa (H₁): Pelo menos uma média diferente.
* **Verificação de pressupostos**
  + Normalidade, homogeneidade de variâncias
* **Análise descritiva**
* **Execução da ANOVA**
  + Aplicar ANOVA de um fator ou multifatorial, dependendo da estrutura do problema.
* **Análise post hoc**
  + Realizar testes de Tukey ou Bonferroni para identificar grupos com diferenças importantes.
* **Interpretação dos efeitos**
  + Resultados do modelo
  + Apresentar o valor-p, estatística F e conclusões sobre as hipóteses.

### 3. Regressão Múltipla/Logística

**Objetivo:** Modelar relações entre uma variável dependente e várias independentes.

* **Descrição do estudo**
* **Formulação do problema**
  + identificar variáveis ​​dependentes e independentes.
  + Definir hipóteses sobre os efeitos esperados.
* **Verificação de pressupostos**
  + Linearidade entre variáveis ​​dependentes e independentes.
  + Ausência de multicolinearidade (VIF).
  + Homocedasticidade (análise de resíduos).
  + Normalidade dos resíduos.
* **Análise descritiva**
  + Etapa inicial de análise estatística de resumir, organizar e apresentar as características principais de um conjunto de dados, oferecendo uma visão geral das distribuições, relações e padrões nos dados sem fazer inferências ou combinações.
* **Execução do modelo**
  + Ajustar o modelo de regressão (linear para contínuas, logística para binárias).
  + Documentar coeficientes, estatísticas t, valores-p e R² ajustados.
* **Validação do modelo**
  + Dividir amostra em treino e teste.
  + Calcular métricas como RMSE ou AUC para avaliar o desempenho.
* **Interpretação**
  + Resultados do modelo
  + Relatar coeficientes significativos e implicações práticas.

### 4. (PSM)

* **Passo-a-passo da análise com descrição detalhada do caso investigado** 
  + Decisões de análise
  + Passo a passo do matching e verificação de balanceamento
* **Definição do problema**
  + identificar a variável de tratamento e o estágio.
  + Selecionar covariáveis ​​relevantes.
* **Cálculo dos escores de propensão**
  + Ajustar modelo de regressão logística para calcular as pontuações.
  + Verifique uma "área de suporte comum" entre grupos.
* **Pareamento**
  + Aplicar métodos de pareamento: vizinho mais próximo, calibre ou outros.
  + Verifique o equilíbrio das covariáveis ​​entre grupos.
* **Estimativa do efeito**
  + Comparar estágios entre grupos pareados usando t-testes ou regressão.
* **Conclusão**
  + Relatar o efeito do tratamento com base no estágio final.
  + Conclusão pelo teste de hipóteses.

### 5. Diff-in-Diff (Diferenças em Diferenças)

**Objetivo:** Analisar impactos de uma intervenção considerando grupos e períodos.

* **Passo-a-passo da análise com descrição detalhada do caso investigado**
  + Decisões de análise
  + Modelagem básica e com controles
* **Definição do design**
  + Identificar grupos (tratamento e controle) e períodos (antes e depois).
  + Garantir a paralelidade das tendências antes da intervenção // testar tendencia paralela.
* **Construção do modelo**
  + Criar variáveis ​​de interação: tratamento × tempo, período pós-tratamento
  + Ajustar modelo de regressão com essas interações.
* **Validação de pressupostos**
  + Verifique a paralelidade de tendências pré-intervenção.
* **Análise do modelo**
  + Documentar coeficientes de interação (impacto da intervenção).
  + Verificar a robustez com variáveis ​​de controle.
* **Interpretação**
  + Relatar resultados da interação como efeito causal.
  + Discuta as limitações, como possíveis viéses.
  + Conclusão sosbre o efeito do tratamento

Ao final, será apresentada uma Conclusão Geral, sintetizando os achados mais relevantes.

Carregamento e Preparação dos Dados

# Carregar bibliotecas necessárias  
library(dplyr)  
library(ggplot2)  
library(skimr)  
library(DataExplorer)  
library(corrplot)  
library(GGally)  
library(forcats)  
library(funModeling)  
library(inspectdf)  
library(qacBase)  
library(car)  
library(lmtest)  
library(MASS)  
library(MatchIt)  
library(tableone)  
library(plm)  
library(readxl)

Social\_Meida\_Dataset <-read.csv ('Social Meida Dataset Marta.csv', sep = ‘;',dec=',')

Na sequência, apresentamos algumas verificações iniciais de estrutura do dataset:

dim(Social\_Meida\_Dataset)

[1] 300 15

head(Social\_Meida\_Dataset, 5)

# A tibble: 5 × 15  
 `Consumer ID` `Customer Name` Age Gender `Income (USD)` `Education Level`  
 <dbl> <chr> <dbl> <chr> <dbl> <chr>   
1 1 Alice Johnson 28 Female 45000 Bachelor's   
2 2 Bob Smith 35 Male 60000 Master's   
3 3 Chris Taylor 22 Non-bina… 32000 High School   
4 4 David Brown 40 Male 75000 PhD   
5 5 Emily White 29 Female 55000 Bachelor's   
# ℹ 9 more variables: `Social Media Usage (Hours/Day)` <dbl>,  
# `Social Media Platforms` <chr>, `Influence Level` <chr>,  
# `Purchase Decision` <chr>, `Product Category` <chr>,  
# `Specific Product` <chr>, `Amount Spent (USD)` <dbl>, `Brand Name` <chr>,  
# City <chr>

summary(Social\_Meida\_Dataset)

Consumer ID Customer Name Age Gender   
 Min. : 1.00 Length:300 Min. :22.00 Length:300   
 1st Qu.: 75.75 Class :character 1st Qu.:27.00 Class :character   
 Median :150.50 Mode :character Median :29.00 Mode :character   
 Mean :150.50 Mean :29.48   
 3rd Qu.:225.25 3rd Qu.:32.00   
 Max. :300.00 Max. :45.00   
 Income (USD) Education Level Social Media Usage (Hours/Day)  
 Min. :32000 Length:300 Min. :1.800   
 1st Qu.:51000 Class :character 1st Qu.:2.475   
 Median :55000 Mode :character Median :2.700   
 Mean :55467 Mean :2.724   
 3rd Qu.:62000 3rd Qu.:2.900   
 Max. :80000 Max. :5.000   
 Social Media Platforms Influence Level Purchase Decision  
 Length:300 Length:300 Length:300   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
 Product Category Specific Product Amount Spent (USD) Brand Name   
 Length:300 Length:300 Min. : 0.0 Length:300   
 Class :character Class :character 1st Qu.: 30.0 Class :character   
 Mode :character Mode :character Median : 65.0 Mode :character   
 Mean : 183.7   
 3rd Qu.: 200.0   
 Max. :1200.0   
 City   
 Length:300   
 Class :character   
 Mode :character

skim(Social\_Meida\_Dataset)

Data summary

| Name | Social\_Meida\_Dataset |
| --- | --- |
| Number of rows | 300 |
| Number of columns | 15 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 10 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Customer Name | 0 | 1 | 9 | 19 | 0 | 142 | 0 |
| Gender | 0 | 1 | 4 | 10 | 0 | 3 | 0 |
| Education Level | 0 | 1 | 3 | 11 | 0 | 4 | 0 |
| Social Media Platforms | 0 | 1 | 15 | 20 | 0 | 22 | 0 |
| Influence Level | 0 | 1 | 15 | 20 | 0 | 3 | 0 |
| Purchase Decision | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| Product Category | 0 | 1 | 3 | 11 | 0 | 7 | 0 |
| Specific Product | 0 | 1 | 2 | 18 | 0 | 55 | 0 |
| Brand Name | 0 | 1 | 3 | 25 | 0 | 66 | 0 |
| City | 0 | 1 | 5 | 13 | 0 | 13 | 0 |

**Variable type: numeric**

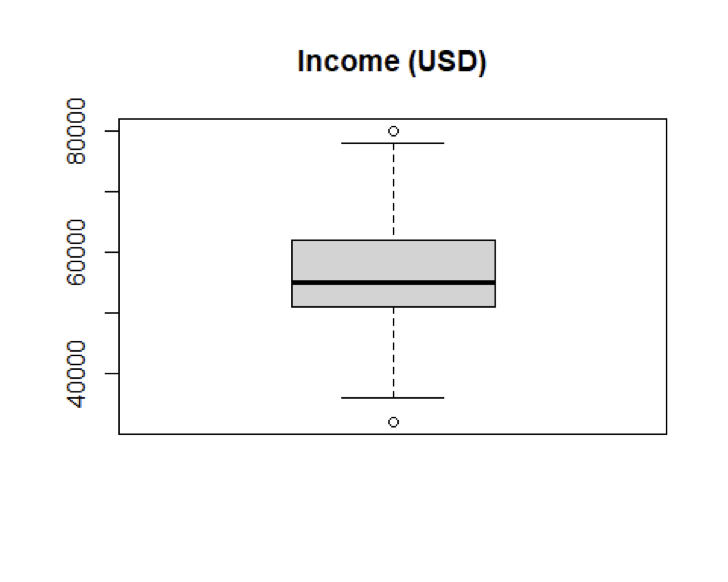
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Consumer ID | 0 | 1 | 150.50 | 86.75 | 1.0 | 75.75 | 150.5 | 225.25 | 300 | ▇▇▇▇▇ |
| Age | 0 | 1 | 29.48 | 3.30 | 22.0 | 27.00 | 29.0 | 32.00 | 45 | ▂▇▃▁▁ |
| Income (USD) | 0 | 1 | 55466.67 | 8462.52 | 32000.0 | 51000.00 | 55000.0 | 62000.00 | 80000 | ▁▃▇▅▁ |
| Social Media Usage (Hours/Day) | 0 | 1 | 2.72 | 0.44 | 1.8 | 2.48 | 2.7 | 2.90 | 5 | ▃▇▁▁▁ |
| Amount Spent (USD) | 0 | 1 | 183.73 | 261.52 | 0.0 | 30.00 | 65.0 | 200.00 | 1200 | ▇▁▁▁▁ |

Roteiro 1 – Exploração de Dados (AED)

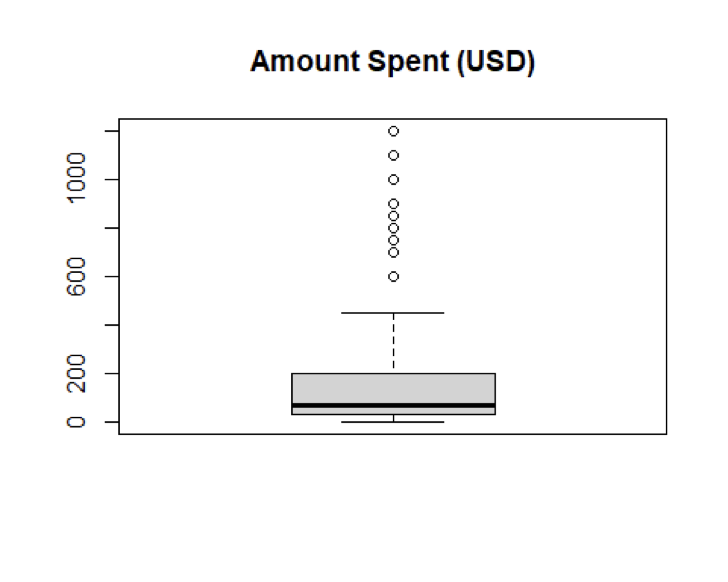
Identificação de Outliers e Dados Ausentes

Para verificar outliers nas variáveis numéricas, utilizamos boxplots:

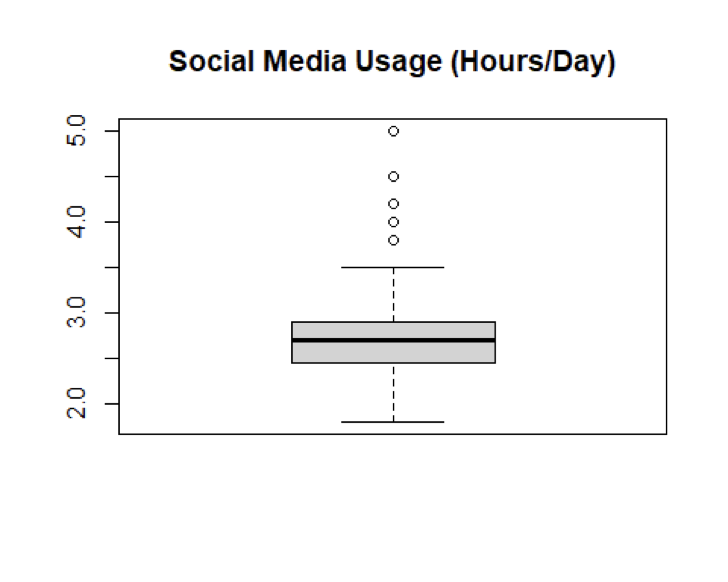
# Boxplots para variáveis numéricas de maior interesse  
boxplot(Social\_Meida\_Dataset$`Income (USD)`, main = "Income (USD)")



boxplot(Social\_Meida\_Dataset$`Amount Spent (USD)`, main = "Amount Spent (USD)")



boxplot(Social\_Meida\_Dataset$`Social Media Usage (Hours/Day)`, main = "Social Media Usage (Hours/Day)")



* Os boxplots permitem visualizar possíveis valores muito além do 3º quartil ou abaixo do 1º quartil.
* Podemos notar que Amount Spent (USD) e Income (USD) apresentam valores potencialmente altos que podem ser considerados outliers de acordo com a regra do IQR (Interquartile Range).
* Entretanto, decidir o que é ou não outlier depende do contexto. Nem sempre é apropriado excluí-los, pois podem representar consumidores de renda ou gastos realmente maiores.

Em relação a **dados ausentes**, verificamos a presença através da funçãodf\_status():

df\_status(Social\_Meida\_Dataset)

variable q\_zeros p\_zeros q\_na p\_na q\_inf p\_inf  
1 Consumer ID 0 0 0 0 0 0  
2 Customer Name 0 0 0 0 0 0  
3 Age 0 0 0 0 0 0  
4 Gender 0 0 0 0 0 0  
5 Income (USD) 0 0 0 0 0 0  
6 Education Level 0 0 0 0 0 0  
7 Social Media Usage (Hours/Day) 0 0 0 0 0 0  
8 Social Media Platforms 0 0 0 0 0 0  
9 Influence Level 0 0 0 0 0 0  
10 Purchase Decision 0 0 0 0 0 0  
11 Product Category 0 0 0 0 0 0  
12 Specific Product 0 0 0 0 0 0  
13 Amount Spent (USD) 15 5 0 0 0 0  
14 Brand Name 0 0 0 0 0 0  
15 City 0 0 0 0 0 0  
 type unique  
1 numeric 300  
2 character 142  
3 numeric 18  
4 character 3  
5 numeric 35  
6 character 4  
7 numeric 21  
8 character 22  
9 character 3  
10 character 2  
11 character 7  
12 character 55  
13 numeric 38  
14 character 66  
15 character 13

No conjunto apresentado, não foram detectados valores faltantes (missing) para as variáveis principais, conforme indicado pelos percentuais de missing igual a 0%.

Recodificação de Variáveis

Para realizar análises mais focadas, foram realizadas algumas recodificações importantes:

* **Education Level** foi traduzida/agrupada em: Ensino Médio, Graduação, Mestrado, Doutorado.
* Foram criadas variáveis agrupadas, como Product\_Category\_Grouped, Influence\_Level\_Grouped e Social\_Media\_Platforms\_Grouped.
* Removemos categorias muito específicas (ex: Non-binary em *Gender*) para simplificar as comparações.

Exemplo de recodificação:

Social\_Meida\_Dataset <- Social\_Meida\_Dataset %>%  
 mutate(Education\_Level\_Recoded = case\_when(  
 `Education Level` == "High School" ~ "Ensino Médio",  
 `Education Level` == "Bachelor's" ~ "Graduação",  
 `Education Level` == "Master's" ~ "Mestrado",  
 `Education Level` == "PhD" ~ "Doutorado"  
 ))

Da mesma forma, recodificamos Purchase Decision e outras colunas categóricas, bem como removemos variáveis que não seriam utilizadas em etapas seguintes (ex.: Specific Product), de modo a deixar o conjunto mais enxuto.

Tabelas de Frequência e Medidas Descritivas

Tabelas de Frequência

# Frequência de Gênero:  
table(Social\_Meida\_Dataset$Gender)

Female Male Non-binary  
 149 150 1

# Frequência de Purchase Decision:  
table(Social\_Meida\_Dataset$`Purchase Decision`)

No Yes  
 14 286

**Exemplo de tabela de frequência** para Gender:

* Female: 149
* Male: 150
* Non-binary: 1

Medidas Descritivas

# Medidas de tendência central e dispersão  
summary(Social\_Meida\_Dataset$`Income (USD)`)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 32000 51000 55000 55467 62000 80000

summary(Social\_Meida\_Dataset$`Amount Spent (USD)`)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 0.0 30.0 65.0 183.7 200.0 1200.0

summary(Social\_Meida\_Dataset$`Social Media Usage (Hours/Day)`)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 1.800 2.475 2.700 2.724 2.900 5.000

Ainda, calculamos o **Coeficiente de Variação (CV)** para avaliar a dispersão relativa:

cv <- function(x) { sd(x) / mean(x) \* 100 }  
cv\_income <- cv(Social\_Meida\_Dataset$`Income (USD)`)  
cv\_spent <- cv(Social\_Meida\_Dataset$`Amount Spent (USD)`)  
cv\_usage <- cv(Social\_Meida\_Dataset$`Social Media Usage (Hours/Day)`)  
  
cv\_income; cv\_spent; cv\_usage

[1] 15.25695

[1] 142.3372

[1] 16.18628

* Quanto maior o CV, maior a dispersão relativa em relação à média.
* Income (USD) apresentou CV em torno de ~15%, enquanto Amount Spent (USD) teve um CV muito maior (~142%), indicando alta variação entre os consumidores.

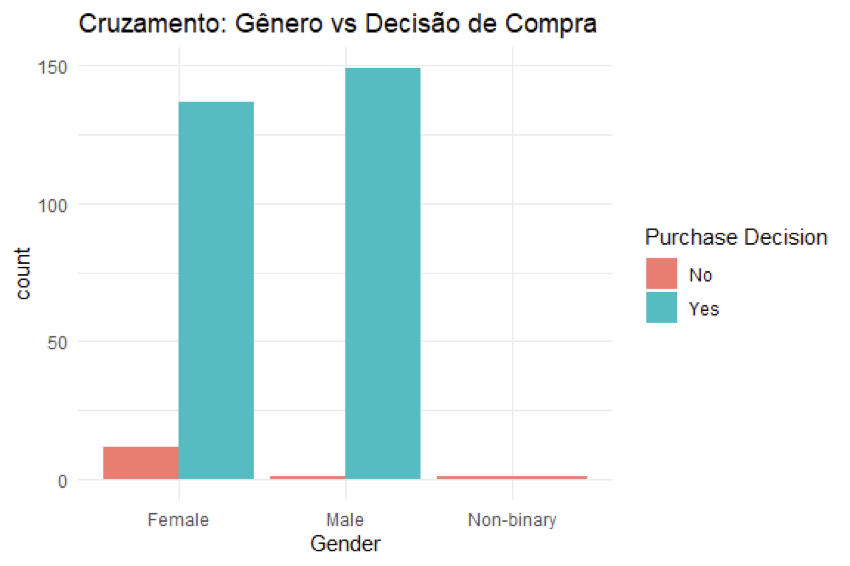
Cruzamentos e Visualização

Para ilustrar como o gasto (Amount Spent (USD)) varia em função do gênero e decisão de compra:

table(Social\_Meida\_Dataset$Gender, Social\_Meida\_Dataset$`Purchase Decision`)

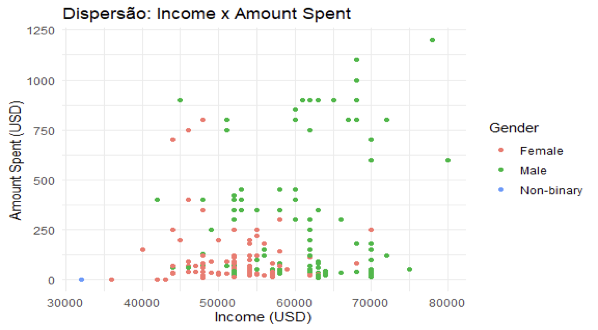
No Yes  
 Female 12 137  
 Male 1 149  
 Non-binary 1 0

# Gráfico de barras  
ggplot(Social\_Meida\_Dataset, aes(x = Gender, fill = `Purchase Decision`)) +  
 geom\_bar(position = "dodge") +  
 theme\_minimal() +  
 labs(title = "Cruzamento: Gênero vs Decisão de Compra")



Para relacionar Renda e Gasto:

ggplot(Social\_Meida\_Dataset, aes(x = `Income (USD)`, y = `Amount Spent (USD)`, color = Gender)) +  
 geom\_point() +  
 theme\_minimal() +  
 labs(title = "Dispersão: Income x Amount Spent")



O gráfico de dispersão mostra a relação entre a renda (Income (USD)) e os gastos (Amount Spent (USD)), com pontos coloridos por gênero. Observa-se uma tendência de aumento nos gastos conforme a renda aumenta, com consumidores do gênero masculino tendendo a gastar mais. Alguns outliers indicam alto consumo em faixas de renda mais baixas.

Roteiro 2 – ANOVA

Descrição e Hipóteses

Queremos verificar se a variável de interesse Amount Spent (USD) difere significativamente entre:

* Grupos de **Gênero** (Female vs Male);
* Grupos de **Nível de Educação** (Graduação ou Ensino Médio vs Pós-Graduação);
* Outras categorizações como Product\_Category\_Grouped, etc.

**Exemplo de hipótese para Gênero**:

* *H0:* A média de Amount Spent (USD) é a mesma para homens e mulheres.
* *H1:* A média de Amount Spent (USD) difere entre homens e mulheres.

Pressupostos

1. **Normalidade dos resíduos**: Teste de Shapiro-Wilk mostrou p-value < 2.2e-16 para a variável bruta, indicando não-normalidade.
2. **Homogeneidade de variâncias**: Teste de Levene. Várias comparações apresentaram *p < 0.05*, sugerindo violação da homocedasticidade para alguns grupos (ex.: Gender, Product\_Category\_Grouped).

Quando ocorre violação, podemos recorrer a um teste **não paramétrico** (Kruskal-Wallis) como alternativa.

Resultados ANOVA

Exemplo para Gender:

anova\_gender <- aov(`Amount Spent (USD)` ~ Gender, data = Social\_Meida\_Dataset)  
summary(anova\_gender)

Df Sum Sq Mean Sq F value Pr(>F)   
Gender 2 2155356 1077678 17.5 6.56e-08 \*\*\*  
Residuals 297 18294212 61597   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Interpretação**: A saída mostra p-value < 0.001 para Gender, indicando diferença significativa de gastos médios entre homens e mulheres.

Também testamos para outras variáveis categóricas (Education\_Level\_Recoded, Product\_Category\_Grouped, etc.). Resumidamente, encontramos:

* **Gênero**: diferença significativa (p < 0.001).
* **Education\_Level\_Recoded**: diferença significativa (p < 0.01).
* **Product\_Category\_Grouped**: diferença altamente significativa (p < 0.001).
* **Influence\_Level\_Grouped**: não significativa (p ~ 0.87).
* **Social\_Media\_Platforms\_Grouped**: significativa (p < 0.01).
* **City\_Grouped**: não significativa (p ~ 0.50).

Conclusão Roteiro 2

Há evidências de que o gasto médio se diferencia entre diversos grupos (Gênero, Educação, Categoria de Produto, etc.). Em especial, parece haver maior gasto associado a algumas categorias de produto mais caras (Eletrônicos e Casa, Saúde e Bem-Estar) e renda maior para determinados grupos.

Roteiro 3 – Regressão Múltipla

Hipóteses e Pressupostos

Objetivamos explicar (ou prever) Amount Spent (USD) por meio de variáveis explicativas como Gênero, Educação, Categoria de Produto, etc.

**Pressupostos**:

1. Linearidade das relações (checamos via gráficos de dispersão).
2. Independência dos resíduos (Durbin-Watson).
3. Homocedasticidade (Breusch-Pagan).
4. Normalidade dos resíduos (Shapiro-Wilk).
5. Ausência de multicolinearidade (VIF).

Ajuste do Modelo

Ajustamos um modelo de regressão linear múltipla:

modelo\_regressao <- lm(`Amount Spent (USD)` ~ Gender + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped +  
 Social\_Media\_Platforms\_Grouped + City\_Grouped,  
 data = Social\_Meida\_Dataset)  
summary(modelo\_regressao)

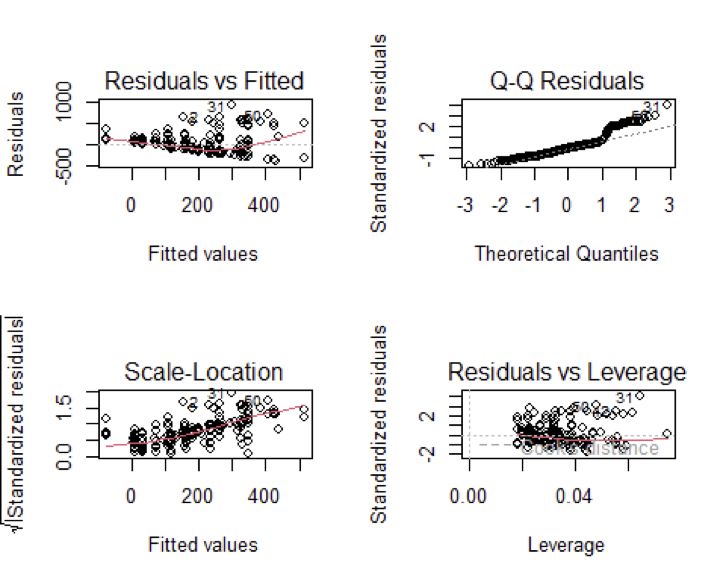
Call:  
lm(formula = `Amount Spent (USD)` ~ Gender + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped + Social\_Media\_Platforms\_Grouped +  
 City\_Grouped, data = Social\_Meida\_Dataset)  
  
Residuals:  
 Min 1Q Median 3Q Max  
-397.49 -157.64 -37.02 69.12 903.35  
  
Coefficients:  
 Estimate Std. Error  
(Intercept) 93.72 59.24  
GenderMale 79.85 32.18  
Education\_Level\_RecodedPós-Graduação -106.77 35.32  
Product\_Category\_GroupedEletrônicos e Casa 150.62 32.62  
Product\_Category\_GroupedSaúde e Bem-Estar 228.74 44.54  
Influence\_Level\_GroupedPouco ou Medianamente Influente -87.16 34.33  
Social\_Media\_Platforms\_GroupedOutras Plataformas 88.28 32.55  
City\_GroupedCidades Tecnológicas 24.07 44.48  
City\_GroupedGrandes Metrópoles 23.93 41.80  
 t value Pr(>|t|)   
(Intercept) 1.582 0.11481   
GenderMale 2.481 0.01369 \*   
Education\_Level\_RecodedPós-Graduação -3.023 0.00274 \*\*  
Product\_Category\_GroupedEletrônicos e Casa 4.618 5.96e-06 \*\*\*  
Product\_Category\_GroupedSaúde e Bem-Estar 5.135 5.33e-07 \*\*\*  
Influence\_Level\_GroupedPouco ou Medianamente Influente -2.539 0.01166 \*   
Social\_Media\_Platforms\_GroupedOutras Plataformas 2.712 0.00710 \*\*  
City\_GroupedCidades Tecnológicas 0.541 0.58883   
City\_GroupedGrandes Metrópoles 0.572 0.56752   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 235.6 on 276 degrees of freedom  
Multiple R-squared: 0.2307, Adjusted R-squared: 0.2084  
F-statistic: 10.34 on 8 and 276 DF, p-value: 1.161e-12

* O **R²** ajustado ficou em torno de ~0.21/0.23, sugerindo que ~21-23% da variabilidade de Amount Spent (USD) é explicada pelas variáveis do modelo.

Verificamos posteriormente se a transformação log (Log\_Amount\_Spent) ou inclusão de termos quadráticos de Income (USD) e Social Media Usage (Hours/Day) melhoraria a adequação. De fato, houve certa melhora quando adicionamos tais termos e utilizamos o log como resposta, resultando em um aumento do R² ajustado para algo em torno de ~0.28-0.30.

Verificação dos Resíduos

par(mfrow=c(2,2))  
plot(modelo\_regressao)



par(mfrow=c(1,1))

* **Shapiro-Wilk** nos resíduos ainda mostrou não-normalidade significativa (p < 0.05).
* **Breusch-Pagan** indicou heterocedasticidade (p < 0.001).

Isso implica que os intervalos de confiança e testes podem estar subestimando ou superestimando a variância. Podemos então usar correções de erros padrão robustos para contornar parte desse problema.

VIF

vif(modelo\_regressao)

GVIF Df GVIF^(1/(2\*Df))  
Gender 1.326357 1 1.151676  
Education\_Level\_Recoded 1.557774 1 1.248108  
Product\_Category\_Grouped 1.151801 2 1.035963  
Influence\_Level\_Grouped 1.467589 1 1.211441  
Social\_Media\_Platforms\_Grouped 1.089471 1 1.043777  
City\_Grouped 1.149544 2 1.035455

Todos os valores de VIF foram próximos de 1, indicando baixa multicolinearidade entre as variáveis categóricas incluídas.

Modelo Final

Depois de diversas comparações, um modelo transformado com log da resposta e inclusão de renda e uso de rede (lineares e quadráticos) apresentou melhor ajuste:

Social\_Meida\_Dataset <- Social\_Meida\_Dataset %>%  
 mutate(Log\_Amount\_Spent = log(`Amount Spent (USD)` + 1))  
modelo\_novo <- lm(Log\_Amount\_Spent ~ Gender + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped +  
 Social\_Media\_Platforms\_Grouped + `Income (USD)` +  
 I(`Income (USD)`^2) + `Social Media Usage (Hours/Day)` +  
 I(`Social Media Usage (Hours/Day)`^2),  
 data = Social\_Meida\_Dataset)  
summary(modelo\_novo)

Call:  
lm(formula = Log\_Amount\_Spent ~ Gender + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped + Social\_Media\_Platforms\_Grouped +  
 `Income (USD)` + I(`Income (USD)`^2) + `Social Media Usage (Hours/Day)` +  
 I(`Social Media Usage (Hours/Day)`^2), data = Social\_Meida\_Dataset)  
  
Residuals:  
 Min 1Q Median 3Q Max  
-1.69324 -0.86721 -0.09571 0.77503 2.66103  
  
Coefficients:  
 Estimate Std. Error  
(Intercept) 1.773e+01 3.700e+00  
GenderMale 5.499e-01 1.715e-01  
Education\_Level\_RecodedPós-Graduação -3.280e-01 1.635e-01  
Product\_Category\_GroupedEletrônicos e Casa 1.012e+00 1.419e-01  
Product\_Category\_GroupedSaúde e Bem-Estar 5.829e-01 1.995e-01  
Influence\_Level\_GroupedPouco ou Medianamente Influente -5.908e-01 1.686e-01  
Social\_Media\_Platforms\_GroupedOutras Plataformas 1.686e-01 1.425e-01  
`Income (USD)` -2.682e-04 1.205e-04  
I(`Income (USD)`^2) 2.278e-09 1.020e-09  
`Social Media Usage (Hours/Day)` -3.603e+00 2.014e+00  
I(`Social Media Usage (Hours/Day)`^2) 5.098e-01 3.617e-01  
 t value Pr(>|t|)   
(Intercept) 4.792 2.71e-06 \*\*\*  
GenderMale 3.206 0.001504 \*\*  
Education\_Level\_RecodedPós-Graduação -2.006 0.045816 \*   
Product\_Category\_GroupedEletrônicos e Casa 7.133 8.78e-12 \*\*\*  
Product\_Category\_GroupedSaúde e Bem-Estar 2.922 0.003764 \*\*  
Influence\_Level\_GroupedPouco ou Medianamente Influente -3.505 0.000533 \*\*\*  
Social\_Media\_Platforms\_GroupedOutras Plataformas 1.183 0.238014   
`Income (USD)` -2.226 0.026832 \*   
I(`Income (USD)`^2) 2.234 0.026296 \*   
`Social Media Usage (Hours/Day)` -1.789 0.074726 .   
I(`Social Media Usage (Hours/Day)`^2) 1.409 0.159836   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.028 on 274 degrees of freedom  
Multiple R-squared: 0.3034, Adjusted R-squared: 0.278  
F-statistic: 11.94 on 10 and 274 DF, p-value: < 2.2e-16

* R² ajustado subiu para ~0.28.
* A heterocedasticidade ainda persiste, mas utilizamos **erros padrão robustos** (vcovHC) para estimar as significâncias.

**Conclusão Roteiro 3**:

* Gênero Masculino apresenta gasto maior.
* Maior renda tende a elevar o gasto, mas com comportamento possivelmente quadrático (há um ponto de não-linearidade).
* Categorias de produto “Eletrônicos e Casa” e “Saúde e Bem-Estar” elevam significativamente o gasto.
* O nível de influência (Muito Influente x Pouco/Mediano) também apareceu significativo em algumas parametrizações, sugerindo que influenciadores muito ativos podem gastar mais.

Roteiro 4 – Propensity Score Matching (PSM)

Descrição do Caso

Para este roteiro, definimos um “tratamento” hipotético baseado no alto uso de redes sociais:

* Tratamento (1) = Social Media Usage (Hours/Day) > 3 horas.
* Controle (0) = Social Media Usage (Hours/Day) <= 3 horas.

**Pergunta**: Consumidores com uso elevado de redes sociais gastam mais (Amount Spent (USD)) do que aqueles com uso moderado ou baixo?

Modelo de Propensity Score

Criamos um modelo logístico para prever a probabilidade de um indivíduo pertencer ao grupo de tratamento, usando variáveis como Idade, Gênero, Renda, etc.

Social\_Meida\_Dataset <- Social\_Meida\_Dataset %>%  
 mutate(Treatment = ifelse(`Social Media Usage (Hours/Day)` > 3, 1, 0))  
  
ps\_model <- glm(Treatment ~ Age + Gender + `Income (USD)` + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped,  
 family = binomial(),  
 data = Social\_Meida\_Dataset)  
  
summary(ps\_model)

Call:  
glm(formula = Treatment ~ Age + Gender + `Income (USD)` + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped, family = binomial(),  
 data = Social\_Meida\_Dataset)  
  
Coefficients:  
 Estimate Std. Error  
(Intercept) -1.142e+01 2.848e+00  
Age 3.738e-02 2.144e-01  
GenderMale 4.481e-01 7.570e-01  
`Income (USD)` 1.300e-04 1.062e-04  
Education\_Level\_RecodedPós-Graduação 3.280e-01 5.555e-01  
Product\_Category\_GroupedEletrônicos e Casa -3.625e-02 6.037e-01  
Product\_Category\_GroupedSaúde e Bem-Estar 2.121e-01 7.361e-01  
Influence\_Level\_GroupedPouco ou Medianamente Influente -1.189e+00 5.712e-01  
 z value Pr(>|z|)   
(Intercept) -4.011 6.05e-05 \*\*\*  
Age 0.174 0.8616   
GenderMale 0.592 0.5539   
`Income (USD)` 1.224 0.2209   
Education\_Level\_RecodedPós-Graduação 0.590 0.5549   
Product\_Category\_GroupedEletrônicos e Casa -0.060 0.9521   
Product\_Category\_GroupedSaúde e Bem-Estar 0.288 0.7732   
Influence\_Level\_GroupedPouco ou Medianamente Influente -2.082 0.0374 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 159.87 on 284 degrees of freedom  
Residual deviance: 124.17 on 277 degrees of freedom  
AIC: 140.17  
  
Number of Fisher Scoring iterations: 6

Matching

Utilizamos o método de **Nearest Neighbor** (vizinho mais próximo):

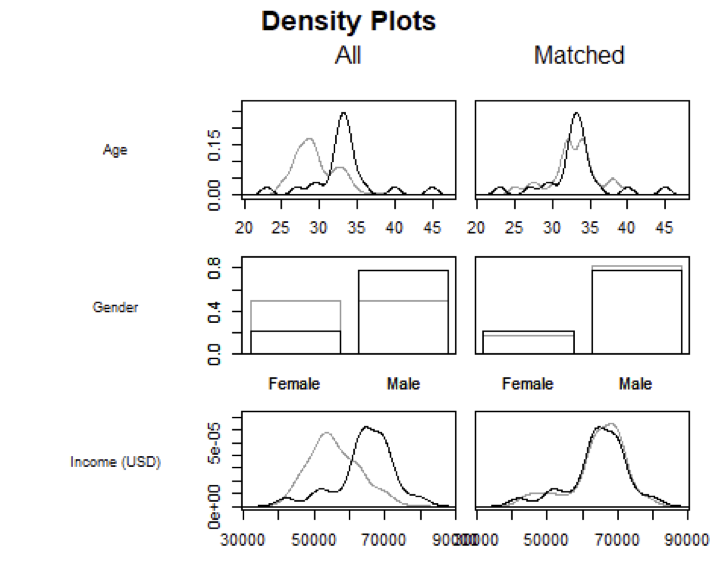
match\_result <- matchit(Treatment ~ Age + Gender + `Income (USD)` + Education\_Level\_Recoded +  
 Product\_Category\_Grouped + Influence\_Level\_Grouped,  
 method = "nearest",  
 data = Social\_Meida\_Dataset)  
summary(match\_result)

Call:  
matchit(formula = Treatment ~ Age + Gender + `Income (USD)` +  
 Education\_Level\_Recoded + Product\_Category\_Grouped + Influence\_Level\_Grouped,  
 data = Social\_Meida\_Dataset, method = "nearest")  
  
Summary of Balance for All Data:  
 Means Treated  
distance 0.2358  
Age 33.1304  
GenderFemale 0.2174  
GenderMale 0.7826  
Income (USD) 64521.7391  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.4348  
Education\_Level\_RecodedPós-Graduação 0.5652  
Product\_Category\_GroupedBeleza e Moda 0.2609  
Product\_Category\_GroupedEletrônicos e Casa 0.4783  
Product\_Category\_GroupedSaúde e Bem-Estar 0.2609  
Influence\_Level\_GroupedMuito Influente 0.7391  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.2609  
 Means Control  
distance 0.0671  
Age 29.5229  
GenderFemale 0.5000  
GenderMale 0.5000  
Income (USD) 55751.9084  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.5954  
Education\_Level\_RecodedPós-Graduação 0.4046  
Product\_Category\_GroupedBeleza e Moda 0.3588  
Product\_Category\_GroupedEletrônicos e Casa 0.5076  
Product\_Category\_GroupedSaúde e Bem-Estar 0.1336  
Influence\_Level\_GroupedMuito Influente 0.3855  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.6145  
 Std. Mean Diff.  
distance 0.9333  
Age 0.8742  
GenderFemale -0.6852  
GenderMale 0.6852  
Income (USD) 1.0722  
Education\_Level\_RecodedGraduação ou Ensino Médio -0.3240  
Education\_Level\_RecodedPós-Graduação 0.3240  
Product\_Category\_GroupedBeleza e Moda -0.2230  
Product\_Category\_GroupedEletrônicos e Casa -0.0588  
Product\_Category\_GroupedSaúde e Bem-Estar 0.2899  
Influence\_Level\_GroupedMuito Influente 0.8053  
Influence\_Level\_GroupedPouco ou Medianamente Influente -0.8053  
 Var. Ratio eCDF Mean  
distance 4.0013 0.3098  
Age 2.2521 0.2128  
GenderFemale . 0.2826  
GenderMale . 0.2826  
Income (USD) 1.3859 0.2572  
Education\_Level\_RecodedGraduação ou Ensino Médio . 0.1606  
Education\_Level\_RecodedPós-Graduação . 0.1606  
Product\_Category\_GroupedBeleza e Moda . 0.0979  
Product\_Category\_GroupedEletrônicos e Casa . 0.0294  
Product\_Category\_GroupedSaúde e Bem-Estar . 0.1273  
Influence\_Level\_GroupedMuito Influente . 0.3536  
Influence\_Level\_GroupedPouco ou Medianamente Influente . 0.3536  
 eCDF Max  
distance 0.6200  
Age 0.5589  
GenderFemale 0.2826  
GenderMale 0.2826  
Income (USD) 0.5712  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.1606  
Education\_Level\_RecodedPós-Graduação 0.1606  
Product\_Category\_GroupedBeleza e Moda 0.0979  
Product\_Category\_GroupedEletrônicos e Casa 0.0294  
Product\_Category\_GroupedSaúde e Bem-Estar 0.1273  
Influence\_Level\_GroupedMuito Influente 0.3536  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.3536  
  
Summary of Balance for Matched Data:  
 Means Treated  
distance 0.2358  
Age 33.1304  
GenderFemale 0.2174  
GenderMale 0.7826  
Income (USD) 64521.7391  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.4348  
Education\_Level\_RecodedPós-Graduação 0.5652  
Product\_Category\_GroupedBeleza e Moda 0.2609  
Product\_Category\_GroupedEletrônicos e Casa 0.4783  
Product\_Category\_GroupedSaúde e Bem-Estar 0.2609  
Influence\_Level\_GroupedMuito Influente 0.7391  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.2609  
 Means Control  
distance 0.2114  
Age 32.9130  
GenderFemale 0.1739  
GenderMale 0.8261  
Income (USD) 64521.7391  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.3913  
Education\_Level\_RecodedPós-Graduação 0.6087  
Product\_Category\_GroupedBeleza e Moda 0.2174  
Product\_Category\_GroupedEletrônicos e Casa 0.4348  
Product\_Category\_GroupedSaúde e Bem-Estar 0.3478  
Influence\_Level\_GroupedMuito Influente 0.6087  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.3913  
 Std. Mean Diff.  
distance 0.1349  
Age 0.0527  
GenderFemale 0.1054  
GenderMale -0.1054  
Income (USD) 0.0000  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.0877  
Education\_Level\_RecodedPós-Graduação -0.0877  
Product\_Category\_GroupedBeleza e Moda 0.0990  
Product\_Category\_GroupedEletrônicos e Casa 0.0870  
Product\_Category\_GroupedSaúde e Bem-Estar -0.1980  
Influence\_Level\_GroupedMuito Influente 0.2970  
Influence\_Level\_GroupedPouco ou Medianamente Influente -0.2970  
 Var. Ratio eCDF Mean  
distance 2.1101 0.0081  
Age 1.4308 0.0380  
GenderFemale . 0.0435  
GenderMale . 0.0435  
Income (USD) 1.0621 0.0245  
Education\_Level\_RecodedGraduação ou Ensino Médio . 0.0435  
Education\_Level\_RecodedPós-Graduação . 0.0435  
Product\_Category\_GroupedBeleza e Moda . 0.0435  
Product\_Category\_GroupedEletrônicos e Casa . 0.0435  
Product\_Category\_GroupedSaúde e Bem-Estar . 0.0870  
Influence\_Level\_GroupedMuito Influente . 0.1304  
Influence\_Level\_GroupedPouco ou Medianamente Influente . 0.1304  
 eCDF Max Std. Pair Dist.  
distance 0.0870 0.1523  
Age 0.1739 0.4320  
GenderFemale 0.0435 0.5270  
GenderMale 0.0435 0.5270  
Income (USD) 0.0870 2173.9130  
Education\_Level\_RecodedGraduação ou Ensino Médio 0.0435 0.4385  
Education\_Level\_RecodedPós-Graduação 0.0435 0.4385  
Product\_Category\_GroupedBeleza e Moda 0.0435 0.6931  
Product\_Category\_GroupedEletrônicos e Casa 0.0435 1.1315  
Product\_Category\_GroupedSaúde e Bem-Estar 0.0870 0.5941  
Influence\_Level\_GroupedMuito Influente 0.1304 0.4951  
Influence\_Level\_GroupedPouco ou Medianamente Influente 0.1304 0.4951  
  
Sample Sizes:  
 Control Treated  
All 262 23  
Matched 23 23  
Unmatched 239 0  
Discarded 0 0

Verificação do Balanceamento

Observamos a redução do **Standardized Mean Difference (SMD)** para variáveis como renda, idade, gênero, etc., após o matching. Isso indica grupos mais comparáveis.

plot(match\_result, type = "density")



A Figura acima exibe as distribuições de Idade, Gênero e Renda antes (All) e depois (Matched) do pareamento, demonstrando uma redução do desbalanceamento entre os grupos.

Análise do Efeito do Tratamento

Com os dados pareados, verificamos a diferença de médias de Amount Spent (USD):

matched\_data <- match.data(match\_result)  
  
t.test(`Amount Spent (USD)` ~ Treatment, data = matched\_data)

Welch Two Sample t-test  
  
data: Amount Spent (USD) by Treatment  
t = 2.206, df = 34.365, p-value = 0.03417  
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
95 percent confidence interval:  
 16.87172 409.65002  
sample estimates:  
mean in group 0 mean in group 1  
 376.3043 163.0435

No exemplo, obtemos um *p-value* ~ 0.03, indicando diferença significativa das médias. O grupo controle (0) apresentou média de gasto ~376 USD, enquanto o grupo tratado (1) ficou ~163 USD, sugerindo que os que **usam mais horas de redes sociais** (na amostra pareada) **gastam menos** do que o grupo controle — ou seja, o resultado foi o oposto do que poderíamos esperar intuitivamente. Devemos analisar com cautela, pois o *n* do grupo tratado era bem menor.

Roteiro 5 – Diferença-em-Diferenças (DiD)

Definição de Tratamento e Período

Criamos um cenário artificial para exemplificar a técnica **DiD**: - **Tratamento** = Consumidores “Muito Influentes” (Influence\_Level\_Grouped == "Muito Influente"). - **Controle** = “Pouco ou Medianamente Influente”. - **Antes** e **Depois** (post) baseados em um corte de uso de redes sociais acima (post=1) ou abaixo (post=0) da mediana de uso.

Social\_Meida\_Dataset <- Social\_Meida\_Dataset %>%  
 mutate(  
 treatment = ifelse(Influence\_Level\_Grouped == "Muito Influente", 1, 0),  
 post = ifelse(`Social Media Usage (Hours/Day)` > median(`Social Media Usage (Hours/Day)`), 1, 0)  
 )  
  
did\_model <- lm(`Amount Spent (USD)` ~ treatment + post + treatment:post,  
 data = Social\_Meida\_Dataset)  
summary(did\_model)

Call:  
lm(formula = `Amount Spent (USD)` ~ treatment + post + treatment:post,  
 data = Social\_Meida\_Dataset)  
  
Residuals:  
 Min 1Q Median 3Q Max  
-194.81 -158.73 -116.41 29.43 995.19  
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 204.81 22.87 8.957 <2e-16 \*\*\*  
treatment -34.24 50.39 -0.680 0.497   
post -48.41 52.24 -0.927 0.355   
treatment:post 76.57 74.81 1.024 0.307   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 265.7 on 281 degrees of freedom  
Multiple R-squared: 0.004116, Adjusted R-squared: -0.006516  
F-statistic: 0.3872 on 3 and 281 DF, p-value: 0.7623

No coeficiente treatment:post (interação), temos a **estimativa do efeito DiD**. Caso seja estatisticamente significativo, inferimos que houve diferença adicional no gasto para o grupo tratado após o suposto “choque” (uso de mais horas de rede social).

Modelo com Controles

did\_model\_controls <- lm(`Amount Spent (USD)` ~ treatment + post + treatment:post +  
 Gender + Education\_Level\_Recoded + Product\_Category\_Grouped +  
 `Income (USD)` + City\_Grouped,  
 data = Social\_Meida\_Dataset)  
summary(did\_model\_controls)

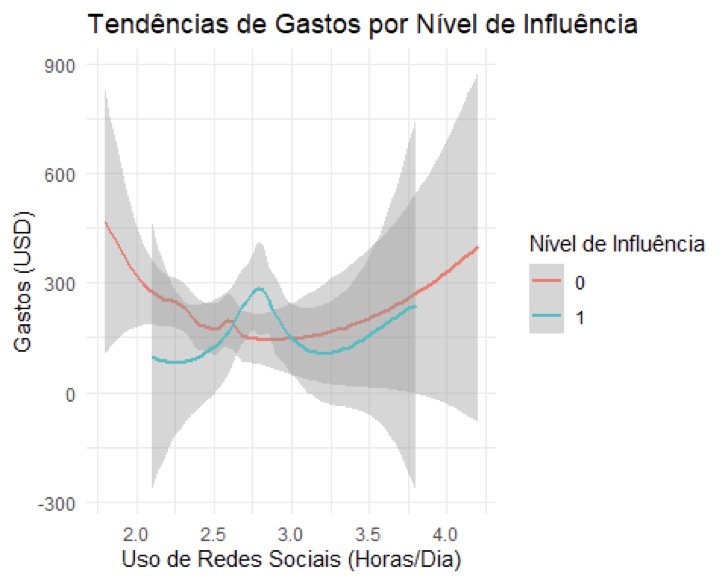
Call:  
lm(formula = `Amount Spent (USD)` ~ treatment + post + treatment:post +  
 Gender + Education\_Level\_Recoded + Product\_Category\_Grouped +  
 `Income (USD)` + City\_Grouped, data = Social\_Meida\_Dataset)  
  
Residuals:  
 Min 1Q Median 3Q Max  
-383.54 -160.89 -43.67 60.36 897.41  
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) -1.296e+01 1.404e+02 -0.092  
treatment 5.940e+01 5.064e+01 1.173  
post -1.314e+02 5.131e+01 -2.561  
GenderMale 1.175e+02 3.832e+01 3.066  
Education\_Level\_RecodedPós-Graduação -9.298e+01 3.632e+01 -2.560  
Product\_Category\_GroupedEletrônicos e Casa 1.512e+02 3.291e+01 4.595  
Product\_Category\_GroupedSaúde e Bem-Estar 2.378e+02 4.543e+01 5.234  
`Income (USD)` 6.827e-04 2.667e-03 0.256  
City\_GroupedCidades Tecnológicas 2.464e+01 4.501e+01 0.547  
City\_GroupedGrandes Metrópoles 4.162e+01 4.266e+01 0.976  
treatment:post 1.170e+02 6.930e+01 1.689  
 Pr(>|t|)   
(Intercept) 0.92653   
treatment 0.24187   
post 0.01098 \*   
GenderMale 0.00239 \*\*  
Education\_Level\_RecodedPós-Graduação 0.01100 \*   
Product\_Category\_GroupedEletrônicos e Casa 6.62e-06 \*\*\*  
Product\_Category\_GroupedSaúde e Bem-Estar 3.30e-07 \*\*\*  
`Income (USD)` 0.79812   
City\_GroupedCidades Tecnológicas 0.58449   
City\_GroupedGrandes Metrópoles 0.33005   
treatment:post 0.09243 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 236.7 on 274 degrees of freedom  
Multiple R-squared: 0.2293, Adjusted R-squared: 0.2012  
F-statistic: 8.154 on 10 and 274 DF, p-value: 1.526e-11

* Observa-se que post foi significativo e negativo, sugerindo que, de forma geral, quem está acima da mediana de uso tende a gastar menos (na nossa amostra).
* A interação treatment:post (p ~ 0.09) indica uma diferença marginalmente significativa. Talvez com mais dados, poderíamos ver um resultado mais conclusivo.

Teste de Tendências Paralelas e Placebo

* **Teste de tendências paralelas**: plotamos gastos vs. uso de redes, estratificado por tratamento.
* **Placebo**: criamos um tratamento aleatório (placebo\_treatment) e repetimos a DiD. Sem encontrar significância, reforça a ideia de que qualquer resultado com a nossa variável real não é simplesmente fruto do acaso.

ggplot(Social\_Meida\_Dataset, aes(x = `Social Media Usage (Hours/Day)`,  
 y = `Amount Spent (USD)`,  
 color = factor(treatment))) +  
 geom\_smooth(method = "loess", se = TRUE) +  
 theme\_minimal() +  
 labs(title = "Tendências de Gastos por Nível de Influência",  
 x = "Uso de Redes Sociais (Horas/Dia)",  
 y = "Gastos (USD)",  
 color = "Nível de Influência")



A Figura acima mostra a relação entre o Uso de Redes Sociais e o Gasto (USD) para diferentes níveis de influência, ilustrando as tendências de cada grupo ao longo do tempo.

Conclusão Geral

Este relatório apresentou cinco roteiros de análise, utilizando o **Social Media Dataset** focado em comportamento de consumo e uso de redes sociais. A seguir, resumimos os principais achados:

1. **Exploração de Dados**:
   * Verificou-se grande dispersão em Amount Spent (USD) e alguma variação em Income (USD).
   * A maior parte da amostra consome redes sociais em torno de 2,5 a 3 horas/dia, com poucos usuários chegando a mais de 4 horas.
2. **ANOVA**:
   * Mostrou diferenças significativas de gastos entre grupos de gênero (homens gastando mais) e níveis de educação (graduação/ensino médio gastando mais do que pós-graduação).
   * Categorias de produto mais caras (Eletrônicos, Saúde) apresentaram gastos médios bem superiores.
3. **Regressão Múltipla**:
   * Confirmou a importância de gênero, categoria de produto e educação para explicar a variação em gastos.
   * Houve violação de pressupostos de normalidade e homocedasticidade, mas correções robustas indicaram que os coeficientes mantêm significância em boa parte dos casos.
   * Transformações log e termos quadráticos de renda e uso de redes ajudaram a melhorar o ajuste.
4. **Propensity Score Matching**:
   * Definindo o grupo tratado como quem usa >3 horas/dia de redes sociais, encontramos, após o matching, que esse grupo apresentou gasto menor (cerca de 163 USD vs 376 USD no grupo controle). É um resultado curioso que pode indicar perfis de renda diferentes ou tempo livre voltado às redes (em vez de compras).
   * O matching conseguiu balancear Idade, Gênero e Renda razoavelmente, reforçando que a diferença não seria apenas devido ao perfil demográfico.
5. **Diferença-em-Diferenças**:
   * Com um cenário simulado de tratamento (influência alta) e período pós (uso maior de redes), o efeito DiD não foi fortemente significativo, mas houve indicação de diferença marginal.
   * Necessário cautela, pois o delineamento foi artificial. Numa aplicação real, precisaria-se de dados longitudinais e delineamento experimental/quase-experimental mais consistente.

**Perspectivas Finais**

A análise sugere que **fatores demográficos (idade, gênero, educação)** e **características de uso de redes sociais** influenciam no gasto. Estratégias de marketing digital podem ser diferentes para grupos de alta e baixa renda, bem como para heavy users de redes sociais. O PSM e o DiD são complementares em uma abordagem causal, mas dependem de suposições fortes (equilíbrio, tendências paralelas). Pesquisas futuras podem aprofundar a interação entre renda, influência social e motivação de compra, usando dados mais amplos ou acompanhamento temporal efetivo.

**Revisão**:

**1.Exploração de dados: Primeira etapa para entender e organizar o conjunto de dados.** Aqui a ideia é fazer uma análise inicial dos dados disponíveis, identificando padrões, verificando a consistência dos dados e visualizando informações importantes usando gráficos, tabelas e estatísticas descritivas. É importante já antes de aplicar técnicas mais avançadas, identificar erros/lacunas e gerar hipóteses iniciais. A exploração de dados é o **passo inicial essencial em qualquer análise estatística** para avaliar a qualidade dos dados antes de avançar para análises mais complexas. *A exploração de dados deve incluir a avaliação de correlações iniciais entre variáveis ​​para identificar possíveis relações ou multicolinearidade, conforme destacado nos passos de modelagem exploratória de dados (Aula 08 - Regressão)​.*

**2.(P-VALUE) ANOVA:**  Testa diferenças entre médias de grupos, quando há três ou mais deles para verificar **se há diferenças estatisticamente significativas**, a “abreviatura” significa análise de variância. A técnica cabe quando se quer avaliar se fatores categóricos influenciam uma variável numérica. Comparando com a “**Exploração de dados”**, a **ANOVA** é focada em testar diferenças de médias e não em prever ou modelar relações entre variáveis. O foco está em comparar médias. Fornece um valor de significância *(p-value)* que indica se há diferenças significativas entre as médias dos grupos. Não fornece informações sobre o impacto ou direção das diferenças (é o que a regressão vai fazer), apenas se elas existem. Exemplo: Conclui-se que pelo menos uma estratégia de marketing tem vendas médias diferentes, mas não diz qual estratégia. - Aula 07

Para usar variáveis contínuas como independentes, é necessáro usar a **ANCOVA** (Análise de Covariância), que combina ANOVA e regressão.

Pressupostos da ANOVA: independência das observações, normalidade dos resíduos e homogeneidade das variâncias. Esses pressupostos são fundamentais para garantir a validade dos resultados.

* Exemplo prático: Um estudo comparando a produtividade de três métodos de treinamento diferentes poderia usar uma ANOVA para determinar se existe diferença estatisticamente significativa entre as médias dos grupos.
* Exemplo: Comparar se diferentes estratégias de marketing resultam em vendas médias significativamente diferentes.

**3.(IMPACTO)Regressão múltipla ou regressão logística:** **Modela e prevê relações entre variáveis.**

**A múltipla vai** estimar a relação entre uma variável dependente contínua (ex.: vendas - resultado) e uma ou mais variáveis independentes (ex.: preço, localização - explicativas) de forma a prever valores ou entender como é possível explicar uma variável numérica e a **Logística** vai ser usada quando a variável dependente é categórica, ou seja eventos binários (sim/não) (ex.: cliente comprou ou não). O foco está em prever ou explicar o impacto de variáveis específicas. A regressão fornece coeficientes que indicam a direção e a **magnitude do impacto de cada variável independente na variável dependente.** *A regressão logística não só lida com variações binárias, mas também pode ser contínua para categorias ordinais e multinomiais. Aula 08*

Exemplo: Cada aumento de 1% no orçamento de publicidade aumenta as vendas em R$ 500.

* No marketing, use uma regressão logística para prever se um cliente fará uma compra (sim/não) com base em variáveis ​​como idade, histórico de compras e promoções.

**4.(PARES CARACTERISTICAS SEMELHANTES) PSM: Emparelha grupos para criar comparabilidade em estudos observacionais.** (Propensity Score Matching - Emparelhamento por Escore de Propensão.A ideia **é** comparar grupos semelhantes em estudos observacionais, minimizando viés de seleção. Muito comum em estudos de impacto de políticas públicas. A técnica “emparelha” emparelha indivíduos de grupos diferentes com base na probabilidade de receber um tratamento. PSM é para criar comparabilidade entre grupos e não para prever ou identificar diferenças de médias. - Aula 09

**É uma forma de criar pares de indivíduos que receberam um "tratamento" e indivíduos que não receberam, com base em características semelhantes, para comparar seus resultados de forma mais justa.**

O **escore de propensão** é a probabilidade de um indivíduo receber o tratamento com base em suas características observáveis (ex.: idade, renda, escolaridade). Como não podemos controlar quem recebe ou não o tratamento, porque isso pode causar **viés de seleção** já que pessoas que recebem o tratamento podem ser diferentes das que não recebem, em aspectos que também afetam o resultado.O **PSM** tenta equilibrar essas diferenças criando grupos comparáveis (tratados e não tratados) com características semelhantes. O PSM cria um design quase experimental, fortalecendo inferências causais em estudos observacionais (Aula 09 - PSM)​.

Como funciona o PSM?

1. **Calcular o escore de propensão:**
   * Usa-se um modelo estatístico, como uma regressão logística, para calcular a probabilidade de cada indivíduo receber o tratamento com base em suas características.
   * Exemplo: Um modelo prevê que a probabilidade de um aluno participar do programa educacional é 70%, considerando fatores como renda familiar e localização.
2. **Emparelhar indivíduos:**
   * Para cada pessoa no grupo tratado, encontra-se uma ou mais pessoas no grupo não tratado com um escore de propensão semelhante.
   * Métodos comuns de emparelhamento incluem:
     + Emparelhamento 1:1 (um tratado para um controle).
     + Emparelhamento com substituição (um controle pode ser usado para vários tratados).
3. **Comparar os resultados:**
   * Com os grupos emparelhados, os pesquisadores avaliam a diferença média nos resultados (ex.: notas, renda, saúde).

### Quando usar PSM?

* **Estudos observacionais**: Quando não é possível realizar um experimento randomizado.
* **Reduzir viés**: Quando há diferenças sistemáticas entre grupos tratados e não tratados.
* **Impacto causal**: Para estimar o efeito causal de uma intervenção, ajustando para variáveis observáveis.

### Limitações do PSM

* **Viés residual:** Ajusta apenas para variáveis observáveis. Diferenças em variáveis não observáveis ainda podem afetar os resultados.
* **Perda de dados:** Indivíduos sem pares próximos podem ser excluídos, reduzindo a amostra.
* **Dependência do modelo:** Os resultados dependem da qualidade do modelo usado para calcular o escore de propensão.

Exemplo prático: Avaliar o impacto de um programa educacional comparando o desempenho de alunos participantes e não participantes, emparelhados com base em características como renda familiar e escolaridade dos pais.

**5. (ANTES E DEPOIS) Diff-in-Diff (Diferenças em diferenças): Mede o impacto de uma intervenção ao longo do tempo. A** técnica compara mudanças em um grupo antes e depois de uma intervenção. Ideal para estudos de impacto onde se deseja medir o efeito de uma intervenção ao longo do tempo, como o impacto de uma nova lei sobre taxas de criminalidade.Ao contrário do PSM, o Diff-in-Diff considera mudanças **ao longo do tempo**, capturando efeitos dinâmicos de intervenções. - Aula 10

A técnica utiliza dois grupos (tratado e controle) em dois períodos (antes e depois) para eliminar vieses constantes ao longo do tempo.

Exemplo prático: Avaliar o impacto de uma política de subsídios agrícolas comparando mudanças na produtividade de fazendas em uma região tratada versus uma região controlada ao longo do tempo

### Análise Descritiva: etapa preparatória essencial. Ela ajuda a compreender os dados e identificar possíveis problemas que devem ser tratados antes de avançar para análises inferenciais, como ANOVA, regressões ou modelos causais.

1. **Estatísticas Resumo:**
   * **Medidas de Tendência Central:** Média, mediana e moda são usadas para identificar o ponto central ou típico de um conjunto de dados.
   * **Medidas de Dispersão:** Amplitude, variância, desvio padrão e coeficiente de variação (CV) mostram como os dados estão distribuídos ao redor da média.
   * **Medidas de Forma:** Assimetria e curtose indicam a forma de distribuição.
2. **Distribuições de Frequência:**
   * Tabelas de frequência e histogramas são usadas para visualizar a distribuição de uma variável, identificando concentrações e outliers.
3. **Visualizações Gráficas:**
   * Gráficos de barras, histogramas, boxplots e gráficos de dispersão são ferramentas comuns para ilustrar dados de maneira intuitiva.
4. **Verificação de Outliers e Dados Ausentes:**
   * identificar valores extremos ou inconsistentes e avaliar a presença de dados ausentes, o que pode afetar as análises subsequentes.
5. **Cruzamentos de Variáveis:**
   * Tabelas de contingência para variáveis ​​categóricas e gráficos de dispersão para variáveis ​​contínuas ajudam a identificar padrões e relações iniciais entre variáveis.

### Exemplos:

* **Estudo Demográfico:** Em uma pesquisa sobre a renda familiar, a análise descritiva calcularia a média da renda, sua dispersão e representaria os dados em um histograma.
* **Análise de Saúde:** Em um estudo sobre índice de massa corporal (IMC), a análise descritiva incluiria estatísticas abstratas (média, desvio padrão), a frequência por faixa de IMC (baixo peso, normal, sobrepeso, obesidade) e a verificação de outliers.