Relatório AB1.1 - Probabilidade e Estatística

Alunos:

Jean Felipe Duarte Tenório

Jayme Vinicius Esteves Pedroza Melo

Professor: Petrúcio Antônio Medeiros Barros

10-08-2025

Contexto Geral do Projeto de acordo com a base de dados Escolhida:

Introdução e Contexto do Trabalho

Este trabalho tem como objetivo realizar uma **Análise Exploratória de Dados (AED)** e aplicar técnicas estatísticas descritivas e bivariadas para compreender padrões e relações presentes na base de dados escolhida. Todas as análises serão desenvolvidas no **R*.

Contexto Geral da Base de Dados

O conjunto de dados **"Estimation of Obesity Levels Based on Eating Habits and Physical Condition"** foi disponibilizado no **UCI Machine Learning Repository** em 26 de agosto de 2019.
Ele contém informações sobre **hábitos alimentares**, **condições físicas** e **características demográficas** de **2 111 indivíduos** de **México, Peru e Colômbia**, com idades entre **14 e 61 anos**.

O objetivo principal é **estimar o nível de obesidade** com base nas variáveis coletadas.
A variável-alvo `NObesity` (Nível de Obesidade) possui sete categorias: - Abaixo do Peso - Peso Normal - Sobrepeso Nível 1 - Sobrepeso Nível 2 - Obesidade Nível 1 - Obesidade Nível 2 - Obesidade Nível 3

Composição dos Dados

Tipo de Variável	Atributos
Demográficas	`Gender` , `Age` , `Height` , `Weight` , `family_history_with_overweight`
Hábitos alimentares	`FAVC` (alta caloria), `FCVC` (vegetais), `NCP` (refeições), `CAEC` (entre refeições), `CH2O` (água), `CALC` (álcool)
Condição física	`SCC` (monitoramento de calorias), `FAF` (atividade física), `TUE` (uso de tecnologia), `MTRANS` (transporte)
Variável alvo	`NObesity` (nível de obesidade)

Todos os registros são **completos**, sem valores ausentes.

Origem dos Dados

- **77% dos registros** foram gerados sinteticamente usando **WEKA** e o filtro **SMOTE** para balanceamento de classes.
- **23% dos registros** foram coletados diretamente por meio de uma plataforma web, com respostas fornecidas pelos próprios participantes.

Estrutura do Trabalho

1. Preparação e Análise Exploratória de Dados

- Importar a base para o R e identificar número de linhas e colunas.
- Criar um dicionário de dados com a descrição de cada variável.
- Padronizar nomes das colunas e valores categóricos.
- Verificar e tratar valores inconsistentes ou outliers.
- Salvar a base limpa em formato `.xlsx`.

2. Estatísticas Descritivas de uma Variável Numérica

- Selecionar uma variável numérica contínua.
- Calcular mínimo, máximo, média, mediana, desvio padrão e quartis.
- Construir histograma e boxplot, interpretando os resultados.

3. Análise Bivariada de Duas Variáveis Categóricas

- Escolher duas variáveis categóricas.
- Criar gráficos de barras comparativos.
- Interpretar padrões e relações encontradas.

4. Boxplot entre Variável Numérica e Categórica

- Relacionar uma variável numérica contínua e uma categórica.
- Construir boxplot e justificar a escolha das variáveis.
- Interpretar diferenças entre os grupos.

5. Matriz de Correlação

- Calcular e visualizar a matriz de correlação entre todas as variáveis numéricas.
- Explicar as correlações mais relevantes encontradas.

6. Análise Bivariada de Duas Variáveis Numéricas

- Escolher duas variáveis numéricas.
- Criar gráfico de dispersão com linha de tendência.
- Calcular a equação da reta ajustada, a correlação e o coeficiente de determinação (R^2).

7. Formulação e Teste de Três Perguntas

- Criar três perguntas envolvendo análise bivariada.
- Utilizar gráficos ou análises estatísticas para confirmar ou refutar as hipóteses.
- Interpretar cada resultado obtido.

Entregas

- Código-fonte em R (`.R`) e `.Rmd`
- Arquivo Excel com base tratada
- Relatório em HTML ou PDF
- Apresentação de até 7 minutos com principais descobertas

1. Preparação e Análise Exploratória de Dados

Nesta etapa, realizamos a preparação inicial da base escolhida, contemplando:

- Importação e inspeção inicial;
- Seleção e renomeação de variáveis;
- Padronização e tradução de valores categóricos;
- Tratamento de tipos de dados;
- Arredondamento de valores numéricos;
- Exportação da base limpa.

1.1. Limpeza do ambiente

1.1.1. Remove todos os objetos carregados na sessão

```
rm(list = ls())
```

1.2. Definir Diretório dos Arquivos

```
setwd("C:/Users/Jean/Desktop/Trabalho_Probabilidade")
```

1.3. Importação da Base Original

```
Obesidade_raw <- read.csv(
  "ObesityDataSet_raw_and_data_sinthetic.csv",
  header = TRUE, sep = ",", dec = "."
)
```

1.3.1. Primeiras linhas

```
head(Obesidade_raw)
```

	Gender <chr>	Age <dbl>	Height <dbl>	Weight <dbl>	family_history_with_overweight <chr>	FAVC <chr>	FCVC <dbl>	N... <dbl>	CAEC <chr>
1	Female	21	1.62	64.0	yes	no	2	3	Sometimes
2	Female	21	1.52	56.0	yes	no	3	3	Sometimes
3	Male	23	1.80	77.0	yes	no	2	3	Sometimes
4	Male	27	1.80	87.0	no	no	3	3	Sometimes
5	Male	22	1.78	89.8	no	no	2	1	Sometimes
6	Male	29	1.62	53.0	no	yes	2	3	Sometimes

6 rows | 1-10 of 18 columns

1.3.2. Resumo estatístico inicial

```
summary(Obesidade_raw)
```

```
##     Gender               Age            Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00
##  Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47
##  Mode  :character   Median :22.78   Median :1.700   Median : 83.00
##                     Mean   :24.31   Mean   :1.702   Mean   : 86.59
##                     3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
##                     Max.   :61.00   Max.   :1.980   Max.   :173.00
##  family_history_with_overweight     FAVC                FCVC
##  Length:2111                    Length:2111        Min.   :1.000
##  Class :character               Class :character   1st Qu.:2.000
##  Mode  :character               Mode  :character   Median :2.386
##                                                    Mean   :2.419
##                                                    3rd Qu.:3.000
##                                                    Max.   :3.000
##       NCP            CAEC              SMOKE                CH2O
##  Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
##  1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585
##  Median :3.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.686                                         Mean   :2.008
##  3rd Qu.:3.000                                         3rd Qu.:2.477
##  Max.   :4.000                                         Max.   :3.000
##      SCC                 FAF              TUE             CALC
##  Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111
##  Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :1.0000   Median :0.6253   Mode  :character
##                     Mean   :1.0103   Mean   :0.6579
##                     3rd Qu.:1.6667   3rd Qu.:1.0000
##                     Max.   :3.0000   Max.   :2.0000
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

1.3.3. Nomes originais das colunas

```
names(Obesidade_raw)
```

```
##  [1] "Gender"                         "Age"
##  [3] "Height"                         "Weight"
##  [5] "family_history_with_overweight" "FAVC"
##  [7] "FCVC"                           "NCP"
##  [9] "CAEC"                           "SMOKE"
## [11] "CH2O"                           "SCC"
## [13] "FAF"                            "TUE"
## [15] "CALC"                           "MTRANS"
## [17] "NObeyesdad"
```

1.4. Seleção das Colunas Relevantes

```
Obesidade_clean <- Obesidade_raw[, c(
  "Gender", "Age", "Weight", "Height", "family_history_with_overweight",
  "FAVC", "FCVC", "NCP", "CAEC", "SMOKE", "CH2O", "SCC",
  "FAF", "TUE", "CALC", "MTRANS", "NObeyesdad"
)]
```

1.5. Renomeação de Variáveis

```
names(Obesidade_clean) <- c(
  "Genero", "Idade", "Peso", "Altura", "Historico_Familiar_Obesidade",
  "Comida_Alta_Caloria", "Consumo_Vegetais", "Refeicoes_Diarias", "Comer_Entre_Refeicoes",
  "Fumante", "Agua_Diaria", "Monitora_Calorias", "Atividade_Fisica",
  "Uso_Tecnologia", "Consumo_Alcool", "Meio_Transporte", "Nivel_Obesidade"
)
```

1.6. Padronização e Tradução de Valores Categóricos

1.6.1. Gênero

```
Obesidade_clean$Genero <- c("Masculino", "Feminino")[match(Obesidade_clean$Genero, c("Male", "Female"))]
```

1.6.2. Variáveis binárias

```
binarios_en <- c("yes", "no")
binarios_pt <- c("Sim", "Não")
Obesidade_clean$Historico_Familiar_Obesidade <- binarios_pt[match(Obesidade_clean$Historico_Familiar_Obesidade, bin
        arios_en)]
Obesidade_clean$Comida_Alta_Caloria <- binarios_pt[match(Obesidade_clean$Comida_Alta_Caloria, binarios_en)]
Obesidade_clean$Fumante <- binarios_pt[match(Obesidade_clean$Fumante, binarios_en)]
Obesidade_clean$Monitora_Calorias <- binarios_pt[match(Obesidade_clean$Monitora_Calorias, binarios_en)]
```

1.6.3. Categorias de frequência

```
freq_en <- c("no","Sometimes","Frequently","Always")
freq_pt <- c("Não","Às Vezes","Frequentemente","Sempre")
Obesidade_clean$Comer_Entre_Refeicoes <- freq_pt[match(Obesidade_clean$Comer_Entre_Refeicoes, freq_en)]
Obesidade_clean$Consumo_Alcool <- freq_pt[match(Obesidade_clean$Consumo_Alcool, freq_en)]
```

1.6.4. Meio de transporte

```
transport_en <- c("Automobile","Bike","Motorbike","Public_Transportation","Walking")
transport_pt <- c("Carro","Bicicleta","Moto","Transporte Público","À Pé")
Obesidade_clean$Meio_Transporte <- transport_pt[match(Obesidade_clean$Meio_Transporte, transport_en)]
```

1.6.5. Níveis de obesidade

```
nivel_en <- c("Insufficient_Weight","Normal_Weight","Overweight_Level_I","Overweight_Level_II","Obesity_Type_I","Ob
        esity_Type_II","Obesity_Type_III")
nivel_pt <- c("Abaixo do Peso","Peso Normal","Sobrepeso Nível 1","Sobrepeso Nível 2","Obesidade Nível 1","Obesidade
        Nível 2","Obesidade Nível 3")
Obesidade_clean$Nivel_Obesidade <- nivel_pt[match(Obesidade_clean$Nivel_Obesidade, nivel_en)]
```

1.7. Tratamento de valores numéricos

```
Obesidade_clean$Peso <- round(Obesidade_clean$Peso, 1)
Obesidade_clean$Altura <- round(Obesidade_clean$Altura, 2)
Obesidade_clean$Idade <- as.integer(round(Obesidade_clean$Idade))
Obesidade_clean$Agua_Diaria <- round(Obesidade_clean$Agua_Diaria, 2)
Obesidade_clean$Uso_Tecnologia <- as.integer(Obesidade_clean$Uso_Tecnologia)
Obesidade_clean$Atividade_Fisica <- round(Obesidade_clean$Atividade_Fisica, 1)
```

1.7.1. USO de Tecnologia Conversão

```
valores_antigos5 <- c("0","1","2")
valores_novos5 <- c("Pouco","Moderado","Prolongado")
Obesidade_clean$Uso_Tecnologia <- valores_novos5[match(Obesidade_clean$Uso_Tecnologia, valores_antigos5)]
```

1.7.2. Consumo de vegetais e refeições diárias

```
Obesidade_clean$Consumo_Vegetais <- as.integer(Obesidade_clean$Consumo_Vegetais)
Obesidade_clean$Refeicoes_Diarias <- as.integer(Obesidade_clean$Refeicoes_Diarias)
valores_antigos5 <- c("1","2","3")
valores_novos5 <- c("Pouco","Frequentemente","Sempre")
Obesidade_clean$Consumo_Vegetais <- valores_novos5[match(Obesidade_clean$Consumo_Vegetais, valores_antigos5)]
```

1.8. Estrutura e tipos de variáveis

```
typeof(Obesidade_clean$Consumo_Vegetais)
```

```
## [1] "character"
```

```
str(Obesidade_clean)
```

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Genero                       : chr  "Feminino" "Feminino" "Masculino" "Masculino" ...
##  $ Idade                        : int  21 21 23 27 22 29 23 22 24 22 ...
##  $ Peso                         : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ Altura                       : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Historico_Familiar_Obesidade: chr  "Sim" "Sim" "Sim" "Não" ...
##  $ Comida_Alta_Caloria          : chr  "Não" "Não" "Não" "Não" ...
##  $ Consumo_Vegetais             : chr  "Frequentemente" "Sempre" "Frequentemente" "Sempre" ...
##  $ Refeicoes_Diarias            : int  3 3 3 3 1 3 3 3 3 3 ...
##  $ Comer_Entre_Refeicoes        : chr  "Às Vezes" "Às Vezes" "Às Vezes" "Às Vezes" ...
##  $ Fumante                      : chr  "Não" "Sim" "Não" "Não" ...
##  $ Agua_Diaria                  : num  2 3 2 2 2 2 2 2 2 2 ...
##  $ Monitora_Calorias            : chr  "Não" "Sim" "Não" "Não" ...
##  $ Atividade_Fisica             : num  0 3 2 2 0 0 1 3 1 1 ...
##  $ Uso_Tecnologia               : chr  "Moderado" "Pouco" "Moderado" "Pouco" ...
##  $ Consumo_Alcool               : chr  "Não" "Às Vezes" "Frequentemente" "Frequentemente" ...
##  $ Meio_Transporte              : chr  "Transporte Público" "Transporte Público" "Transporte Público" "À Pé" ...
##  $ Nivel_Obesidade              : chr  "Peso Normal" "Peso Normal" "Peso Normal" "Sobrepeso Nível 1" ...
```

1.9. Conversão para fator

```
dados <- c(
  "Genero",
  "Historico_Familiar_Obesidade",
  "Comida_Alta_Caloria",
  "Consumo_Vegetais",
  "Comer_Entre_Refeicoes",
  "Fumante",
  "Monitora_Calorias",
  "Uso_Tecnologia",
  "Consumo_Alcool",
  "Meio_Transporte",
  "Nivel_Obesidade"
)

Obesidade_clean[dados] <- lapply(Obesidade_clean[dados], factor)
```

1.10. Exportação da Base em formato XLSX

```
write.csv(
  Obesidade_clean,
  file = "Obesidade_clean.csv",
  row.names = FALSE,
  fileEncoding = "UTF-8"
)

library(openxlsx)
```

```
## Warning: pacote 'openxlsx' foi compilado no R versão 4.5.1
```

```
write.xlsx(Obesidade_clean, file = "Obesidade_clean.xlsx")
```

2. Estatísticas Descritivas de uma Variável Numérica (Peso)

2.1. Estatísticas Básicas

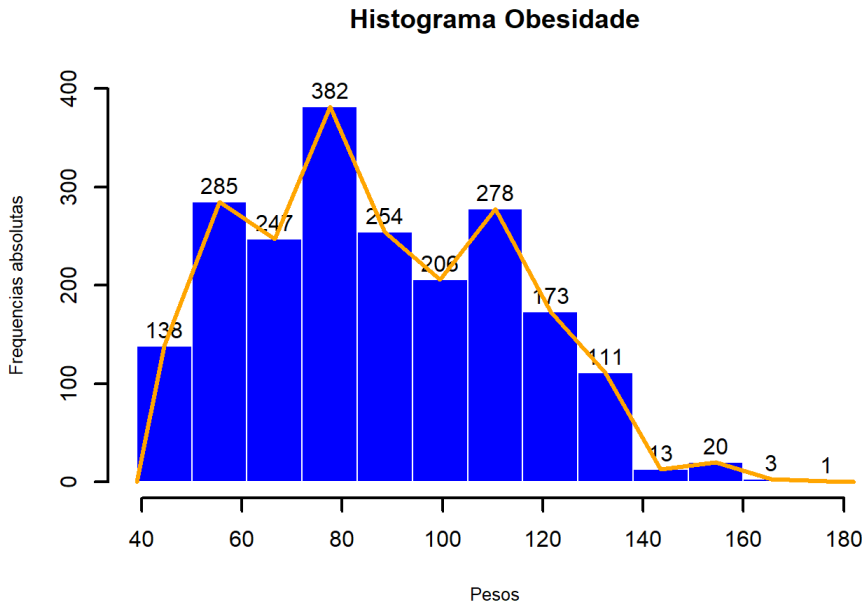
```
setwd("C:/Users/Jean/Desktop/Trabalho_Probabilidade")
obesidade <- read.csv("Obesidade_clean.csv")
obesidade<- within(obesidade, {
  Genero                       <- factor(Genero)
  Historico_Familiar_Obesidade<- factor(Historico_Familiar_Obesidade)
  Comida_Alta_Caloria          <- factor(Comida_Alta_Caloria)
  Consumo_Vegetais             <- factor(Consumo_Vegetais)
  Comer_Entre_Refeicoes        <- factor(Comer_Entre_Refeicoes)
  Fumante                      <- factor(Fumante)
  Monitora_Calorias            <- factor(Monitora_Calorias)
  Uso_Tecnologia               <- factor(Uso_Tecnologia)
  Consumo_Alcool               <- factor(Consumo_Alcool)
  Meio_Transporte              <- factor(Meio_Transporte)
  Nivel_Obesidade              <- factor(Nivel_Obesidade)
})
maximo_obesidade <- max(obesidade$Peso)
minimo_obesidade <- min(obesidade$Peso)
media_obesidade <- mean(obesidade$Peso)
mediana_obesidade <- median(obesidade$Peso)
desvio_padr_obesidade <- sd(obesidade$Peso)
quartis_obesidade <- quantile(obesidade$Peso)
Amplitude_obesidade <- maximo_obesidade - minimo_obesidade

Nclasses <- nclass.Sturges(obesidade$Peso); Nclasses
```

```
## [1] 13
```

```
amplitude_classes <- ceiling(Amplitude_obesidade / Nclasses)

limiteclas <- seq(minimo_obesidade,maximo_obesidade + amplitude_classes,by = amplitude_classes)

classes <- c(
  "39 |- 50",
  "50 |- 61",
  "61 |- 72",
  "72 |- 83",
  "83 |- 94",
  "94 |- 105",
  "105 |- 116",
  "116 |- 127",
  "127 |- 138",
  "138 |- 149",
  "149 |- 160",
  "160 |- 171",
  "171 |- 182"
)

Freq = table(cut(obesidade$Peso, breaks = limiteclas, right=FALSE,
                 labels=classes))
FreqAc <- cumsum(Freq);
FreqRel <- prop.table(Freq);
FreqRelAc <- cumsum(FreqRel)

TabResul = cbind(Freq,
                 FreqAc,
                 FreqRel = round(FreqRel*100,digits = 2),
                 FreqRelAc= round(FreqRelAc*100,digits = 2))
```

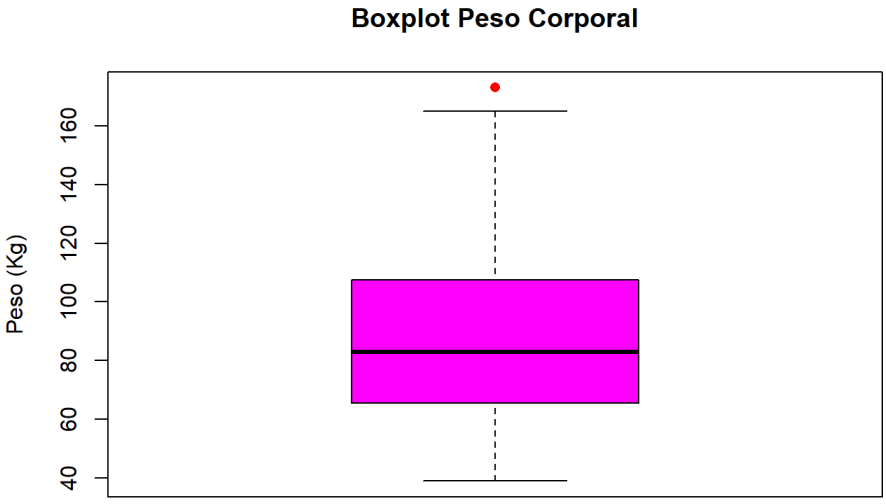
2.2. Exibição do Histograma

```
Pesos <- obesidade$Peso

h <- hist(Pesos, breaks=limiteclas,
         ylab="Frequencias absolutas",  labels=TRUE,cex.lab = 0.8,main="Histograma Obesidade",
         xlim=c(39,182), ylim = c (0,400), col="blue",right=FALSE,border="white",lwd=2)
lines(c(min(h$breaks), h$mids, max(h$breaks)),
      c(0,h$counts, 0), type = "l", col="orange",lwd=3)
```



2.3. Exibição do Boxplot da Variável Numérica Peso

```
boxplot(obesidade$Peso,  col = "magenta",main="Boxplot Peso Corporal",ylab="Peso (Kg)",outcol="red",outpch = 16)
```



3.Análise Bivariada de Duas Variáveis

3.1. Convertendo para factor Nivel de Obesidade para garantir a classificação Ordinal

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Genero                     : Factor w/ 2 levels "Feminino","Masculino": 1 1 2 2 2 2 1 2 2 2 ...
##  $ Idade                      : int  21 21 23 27 22 29 23 22 24 22 ...
##  $ Peso                       : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ Altura                     : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Historico_Familiar_Obesidade: Factor w/ 2 levels "Não","Sim": 2 2 2 1 1 1 2 1 2 2 ...
##  $ Comida_Alta_Caloria        : Factor w/ 2 levels "Não","Sim": 1 1 1 1 1 2 2 1 2 2 ...
##  $ Consumo_Vegetais           : Factor w/ 3 levels "Frequentemente",..: 1 3 1 3 1 1 3 1 3 1 ...
##  $ Refeicoes_Diarias          : int  3 3 3 3 1 3 3 3 3 3 ...
##  $ Comer_Entre_Refeicoes      : Factor w/ 4 levels "Às Vezes","Frequentemente",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Fumante                    : Factor w/ 2 levels "Não","Sim": 1 2 1 1 1 1 1 1 1 1 ...
##  $ Agua_Diaria                : num  2 3 2 2 2 2 2 2 2 2 ...
##  $ Monitora_Calorias          : Factor w/ 2 levels "Não","Sim": 1 2 1 1 1 1 1 1 1 1 ...
##  $ Atividade_Fisica           : num  0 3 2 2 0 0 1 3 1 1 ...
##  $ Uso_Tecnologia             : Factor w/ 3 levels "Moderado","Pouco",..: 1 2 1 2 2 2 2 2 1 1 ...
##  $ Consumo_Alcool             : Factor w/ 4 levels "Às Vezes","Frequentemente",..: 3 1 2 2 1 1 1 1 2 3 ...
##  $ Meio_Transporte            : Factor w/ 5 levels "À Pé","Bicicleta",..: 5 5 5 1 5 3 4 5 5 5 ...
##  $ Nivel_Obesidade            : Factor w/ 7 levels "Abaixo do Peso",..: 2 2 2 3 4 2 2 2 2 2 ...
```

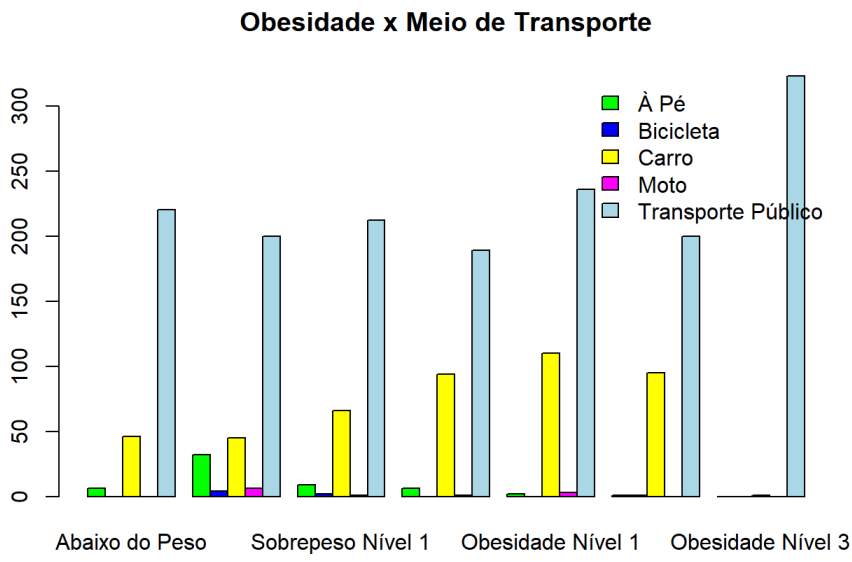
3.2. Analisando a distribuição de Níveis de Obesidade por Meio de Transporte

```
##                    Nivel_Obesidade
## Meio_Transporte     Abaixo do Peso  Peso Normal Sobrepeso Nível 1
##   À Pé                  0.28422549   1.51586926        0.42633823
##   Bicicleta             0.00000000   0.18948366        0.09474183
##   Carro                 2.17906206   2.13169114        3.12648034
##   Moto                  0.00000000   0.28422549        0.04737091
##   Transporte Público   10.42160114   9.47418285       10.04263382
##   Sum                  12.88488868  13.59545239       13.73756514
##                    Nivel_Obesidade
## Meio_Transporte     Sobrepeso Nível 2 Obesidade Nível 1 Obesidade Nível 2
##   À Pé                     0.28422549        0.09474183        0.04737091
##   Bicicleta                0.00000000        0.00000000        0.04737091
##   Carro                    4.45286594        5.21080057        4.50023685
##   Moto                     0.04737091        0.14211274        0.00000000
##   Transporte Público       8.95310279       11.17953577        9.47418285
##   Sum                     13.73756514       16.62719090       14.06916153
##                    Nivel_Obesidade
## Meio_Transporte     Obesidade Nível 3          Sum
##   À Pé                     0.00000000   2.65277120
##   Bicicleta                0.00000000   0.33159640
##   Carro                    0.04737091  21.64850782
##   Moto                     0.00000000   0.52108006
##   Transporte Público      15.30080531  74.84604453
##   Sum                     15.34817622 100.00000000
```

Meio_Transporte	Nivel_Obesidade	Frequência	Proporcao_Geral	Proporcao_por_Transporte
<fct>	<fct>	<int>	<dbl>	<dbl>
À Pé	Abaixo do Peso	6	0.28	10.71
Bicicleta	Abaixo do Peso	0	0.00	0.00
Carro	Abaixo do Peso	46	2.18	10.07
Moto	Abaixo do Peso	0	0.00	0.00
Transporte Público	Abaixo do Peso	220	10.42	13.92
À Pé	Peso Normal	32	1.52	57.14
Bicicleta	Peso Normal	4	0.19	57.14
Carro	Peso Normal	45	2.13	9.85
Moto	Peso Normal	6	0.28	54.55
Transporte Público	Peso Normal	200	9.47	12.66

1-10 of 35 rows Previous **1** 2 3 4 Next

3.3. Exibição do Gráfico Meio de Transporte x Nível de Obesidade



3.4. Interpretação dos Resultados

Os dados analisados revelam um padrão claro de associação entre o tipo de transporte utilizado e o perfil de obesidade dos indivíduos. A partir das frequências absolutas e proporções relativas à amostra total e por modalidade de transporte, destacam-se os seguintes achados:

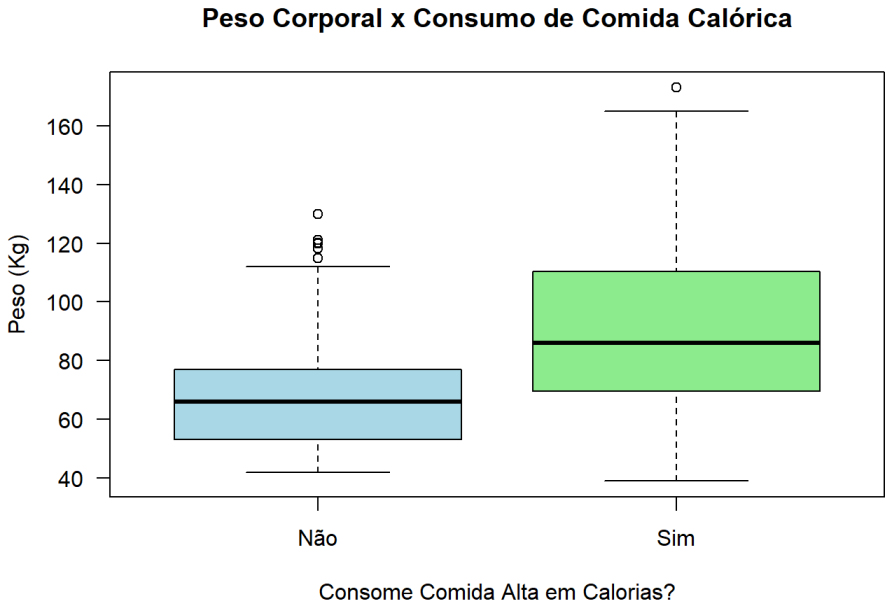
- **Transporte ativo (À Pé e Bicicleta)** apresenta as menores proporções nos níveis mais elevados de obesidade. Por exemplo, quem anda a pé concentra-se majoritariamente nos níveis “Peso Normal” e “Abaixo do Peso”, totalizando apenas **2.65%** da amostra geral, com proporção interna de **Obesidade Nível 3 igual a 0%**.
- **Carro** exibe uma distribuição mais ampla e expressiva em todos os níveis de obesidade, sendo responsável por aproximadamente **21.6%** da amostra. Destaca-se nos níveis “Obesidade Nível 2”, com **4.5%** da amostra geral e **20.74%** de seus próprios usuários nesse nível — indicando um perfil de maior risco.
- **Transporte Público** representa o grupo mais numeroso (**74.85%** da amostra). A distribuição interna revela uma proporção elevada de indivíduos em “Obesidade Nível 3” — **15.30%** da amostra total — e aproximadamente **25.4%** dos usuários desse transporte pertencem a essa categoria extrema. Isso evidencia uma possível relação entre maior sedentarismo urbano e risco obesogênico.
- **Moto e Bicicleta**, por sua baixa representatividade e concentrações tímidas nos níveis elevados, apresentam menor impacto estatístico isolado. No entanto, os padrões são consistentes com os demais meios ativos.

3.4.1. Conclusão

Os resultados sugerem que o tipo de transporte utilizado está vinculado ao nível de obesidade dos indivíduos. Meios ativos de locomoção parecem associados a perfis mais saudáveis, enquanto o uso frequente de transporte público ou automóvel se relaciona com maior prevalência de sobrepeso e obesidade severa.

Esses achados levantam hipóteses importantes sobre o papel da mobilidade urbana na saúde coletiva e podem subsidiar ações voltadas à promoção do transporte ativo como estratégia preventiva contra obesidade.

4. Boxplot entre Variável Numérica e Categórica



4.1. Interpretação dos Resultados do Boxplot

O boxplot mostra uma diferença marcante no peso corporal entre os grupos que consomem ou não comida calórica com frequência. O grupo consumidor apresenta uma mediana de peso consideravelmente maior (~=90 kg), além de uma maior dispersão e presença de valores extremos. Em contraste, os não consumidores têm peso mais concentrado em torno de 65 kg, com menor variabilidade.

4.1.1. Conclusão

Podemos perceber por meio desta análise inicial via boxplot que há indícios de uma boa relação entre a variável numérica contínua peso corporal(kg) com a variável categórica binária Consumo de comidas calóricas, indicando posssivelmente que indíviduos cujo consomem com mais frequências este tipo de alimentação tende a apresentar pesos relativaemnte maiores aqueles que não consomem, sendo um potencial fator de risco para a obesidade.

5. Matriz de Correlação entre Variáveis Numéricas

5.1. Seleção de variáveis numéricas

```
numerica_colunas <- sapply(obesidade, is.numeric)
obesidade_numerica <- obesidade[, numerica_colunas]
```

Selecionamos apenas as colunas numéricas da base para análise de correlação entre elas.

5.2. Matriz de Correlação de Pearson

```
##                   Idade  Peso Altura Refeicoes_Diarias Agua_Diaria
## Idade              1.00  0.20  -0.03             -0.07       -0.05
## Peso               0.20  1.00   0.46              0.13        0.20
## Altura            -0.03  0.46   1.00              0.21        0.21
## Refeicoes_Diarias -0.07  0.13   0.21              1.00        0.07
## Agua_Diaria       -0.05  0.20   0.21              0.07        1.00
## Atividade_Fisica  -0.14 -0.05   0.30              0.11        0.17
##                   Atividade_Fisica
## Idade                        -0.14
## Peso                         -0.05
## Altura                        0.30
## Refeicoes_Diarias             0.11
## Agua_Diaria                   0.17
## Atividade_Fisica              1.00
```

Apresentamos a matriz de correlação entre todas as variáveis numéricas, arredondada para duas casas decimais. Valores positivos indicam associação direta; negativos indicam inversa.

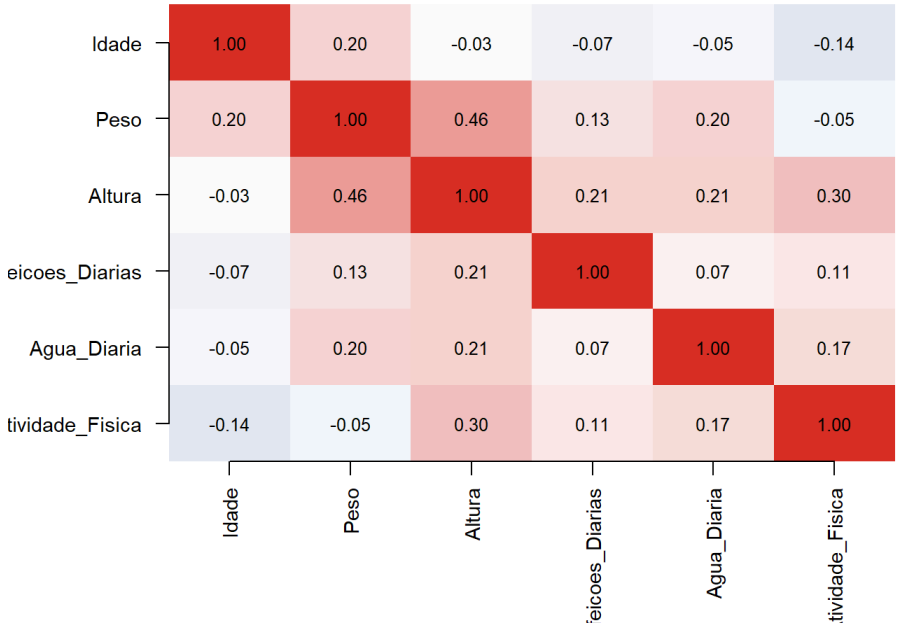
5.3. Principais correlações(Maiores em Módulo)

	var1 <chr>	var2 <chr>	cor <dbl>
3	Peso	Altura	0.4622833
13	Altura	Atividade_Fisica	0.2953816
6	Altura	Refeicoes_Diarias	0.2143986
9	Altura	Agua_Diaria	0.2134689
1	Idade	Peso	0.2034432
8	Peso	Agua_Diaria	0.2005400
15	Agua_Diaria	Atividade_Fisica	0.1673441
11	Idade	Atividade_Fisica	-0.1442263
5	Peso	Refeicoes_Diarias	0.1260511
14	Refeicoes_Diarias	Atividade_Fisica	0.1065681

1-10 of 10 rows

Lista das 10 correlações mais fortes (em módulo), para destacar os pares de variáveis com maior associação linear.

5.4.Visualização Gráfica da Matrix



5.6.Interpretação dos Resultados da Matrix de Correlação numérica

- Em geral, as correlações são fracas a moderadas. Só há uma correlação moderada (Peso–Altura). Isso sugere relações lineares sutis entre as variáveis, sem indícios de multicolinearidade forte.
- Destaques:
- Peso–Altura (r≈0.46): correlação moderada positiva; pessoas mais altas tendem a pesar mais — plausível biologicamente.
- Idade–Atividade_Fisica (r≈-0.14): leve tendência de menor atividade com maior idade, mas o efeito é fraco.
- As demais correlações são fracas; úteis para sinalizar direções, mas não bastam sozinhas para conclusões fortes.

6. Análise Bivariada de Duas Variáveis

6.1. Seleção de 2 Variáveis Numéricas com Maior Correlação

```
x <- obesidade_numerica$Altura
y <- obesidade_numerica$Peso
```

Selecionamos as variáveis Peso e Altura por terem maior correlação entre si e geralmente associadas remetem a distribuição de peso corporal por altura o que pode ser útil para avaliar a sua influência posterior no nível de obesidade.

6.2. Cálculo da Correlação entre Peso e Altura

```
cor_xy <- cor(x, y, use = "complete.obs", method = "pearson")
cof_det <- cor_xy^2
cof_det
```

```
## [1] 0.2137058
```

Aqui fizemos a correlação e o teste de correlação, além do cálculo do coeficiente de determinação para construção em seguida do gráfico de dispersão com reta ajustada

6.3. Cálculo do Modelo de Regressão linear entre os dois dados numéricos

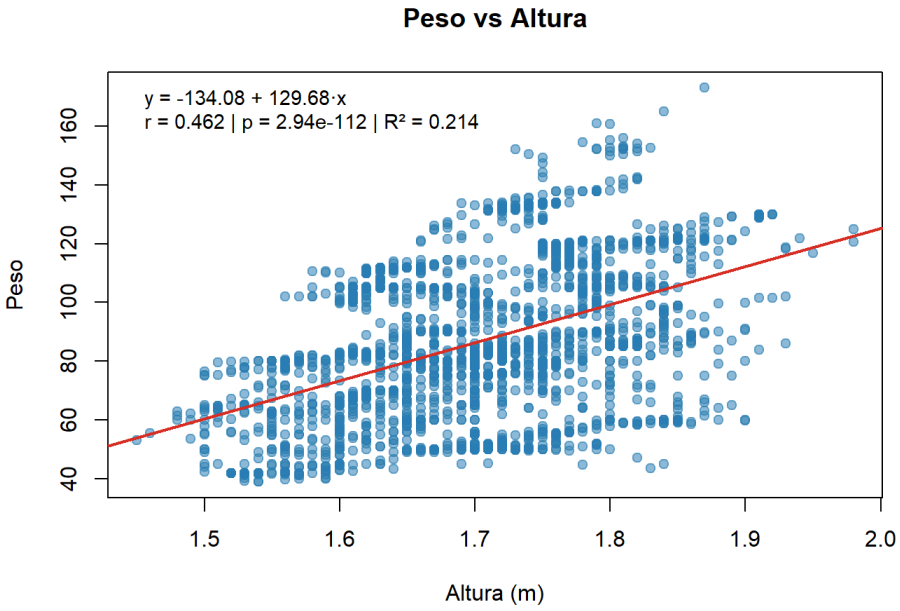
```
modelo <- lm(y ~ x)
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.734 -16.376  -1.931  17.434  64.579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -134.075      9.230  -14.53   <2e-16 ***
## x            129.677      5.416   23.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.23 on 2109 degrees of freedom
## Multiple R-squared:  0.2137, Adjusted R-squared:  0.2133
## F-statistic: 573.2 on 1 and 2109 DF,  p-value: < 2.2e-16
```

6.4. Definição dos coeficiente da reta e do coeficiente de Determinação após regressão linear para verificar consistência com a correlação de pearson ao quadrado

```
b0 <- coef(modelo)[1]
b1 <- coef(modelo)[2]
r2 <- summary(modelo)$r.squared
```

6.5. Plotagem do Gráfico de Dispersão com reta Ajustada



6.6. Interpretação do Gráfico de Dispersão de Altura vs Peso

6.6.1.Interpretação dos Coeficientes

- Intercepto: -134.075 Representa o valor estimado do peso quando a altura é igual a zero. Embora esse ponto não tenha significado prático, o intercepto é necessário para definir a reta de regressão.
- Coeficiente da Altura (x): 129.677 Indica que, para cada aumento de uma unidade na altura, o peso aumenta em média 129.68 kg. Isso sugere uma relação positiva entre as variáveis.
- Ambos os coeficientes apresentam p-valores inferiores a 2e-16, demonstrando alta significância estatística. Ou seja, é extremamente improvável que essa relação tenha surgido ao acaso.

6.6.2. Qualidade do Ajuste

- Coeficiente de Determinação (R^2): 0.2137 Aproximadamente 21.4% da variação no peso é explicada pela altura. Esse valor revela uma associação moderada, sugerindo que outros fatores também influenciam significativamente o peso dos indivíduos.
- Erro Padrão Residual: 23.23 Representa a média da diferença entre os valores observados e os valores ajustados pela regressão. Um erro padrão mais baixo indicaria maior precisão nas estimativas.

6.6.3 Conclusão desta Análise

Com base na análise estatística e no modelo ajustado, é possível concluir que existe uma relação positiva e estatisticamente significativa entre altura e peso. O coeficiente da altura demonstra que, em média, indivíduos mais altos tendem a apresentar maior peso corporal. Contudo, essa relação, embora clara, é apenas moderada: o valor de R^2 indica que menos de um quarto da variação no peso pode ser explicada pela altura. Isso sugere que outros fatores — como idade, sexo, composição corporal ou hábitos de vida — também desempenham papel relevante. Portanto, o modelo é útil para identificar tendências gerais, mas não é suficiente para realizar previsões precisas de peso com base apenas na altura.

7. Formulação e Teste de Três Perguntas

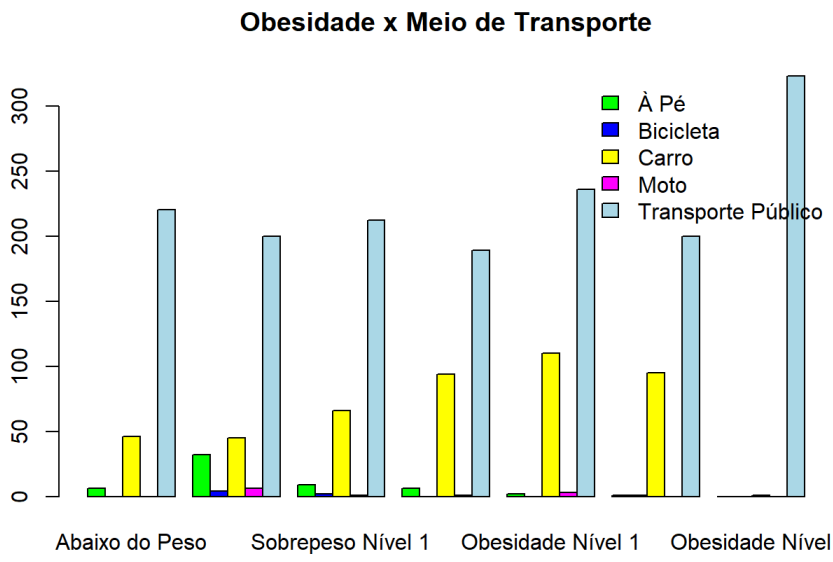
- *As perguntas a serem analisadas e respondidas com base na análise exploratória de dados serão:*

1) Existe associação entre o tipo de transporte utilizado e o nível de obesidade?

2) Indivíduos que consomem comida calórica frequentemente tendem a ter maior peso corporal?

3)A relação entre altura e peso corporal apresenta diferenças significativas entre indivíduos com e sem histórico familiar de obesidade?

7.1. Visualização Gráfica para ajudar a responder a questão 1.



Note que o gráfico plotado aqui foi utilizado na análise entre duas variáveis categóricas aqui daremos aprofundamento da interpretação complementando com outros testes.

7.2. Teste do Qui-Quadrado com simulação com Monte Carlo

```
set.seed(100)
B <- 10000
teste_chi <- chisq.test(meio_transporte_obesidade_tb, simulate.p.value = TRUE, B = B)
```

7.3. Resultado do Teste Qui-Quadrado

Resumo do Teste Qui-quadrado com Simulação Monte Carlo

	Estatistica	Graus_de_Liberdade	P_valor	Conclusao
X-squared	292.59	NA	< 1e-04	Associação significativa entre as variáveis

7.4. Análise e Resposta a primeira Pergunta:

Para responder a pergunta, avaliou-se a existência de associação entre o meio de transporte habitual e o nível de obesidade na amostra por meio do teste de independência qui-quadrado, com simulação Monte Carlo (B = 10 000). O resultado apontou estatística de X^2 = 292,59 e valor-p inferior a 0,0001. Diante desse valor-p muito menor que o nível de significância de 0,05, rejeita-se a hipótese nula de independência entre as variáveis, indicando associação estatisticamente significativa entre o tipo de transporte utilizado e o grau de obesidade dos participantes.

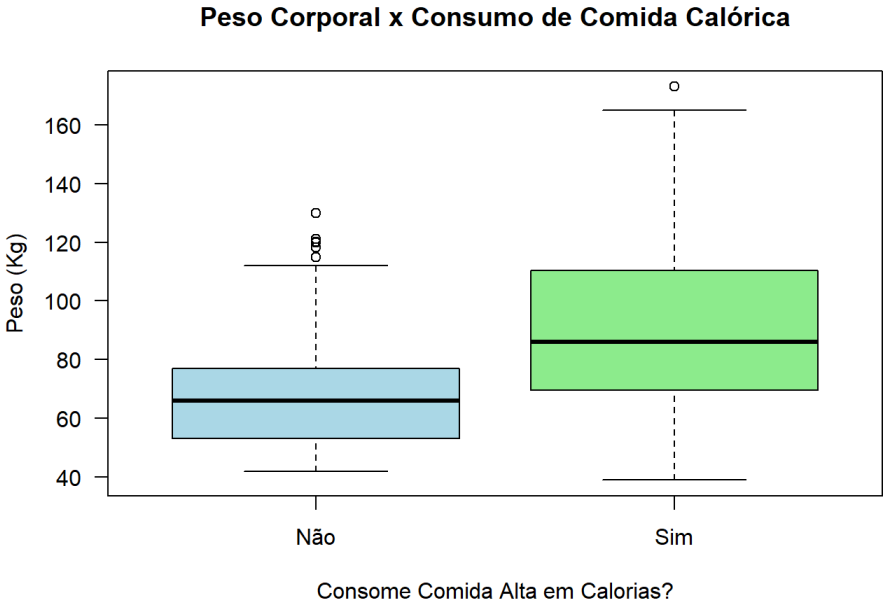
O gráfico de barras empilhadas (Figura X) ilustra com clareza o padrão de distribuição dos níveis nutricionais conforme a modalidade de transporte. Observa-se que:

- Entre os usuários de transporte motorizado privado (automóvel e motocicleta), a proporção de indivíduos classificados como obesos é substancialmente maior, representando a maior fatia da barra correspondente.
- Nos modos ativos (caminhada e bicicleta), predominam participantes com peso normal e sobrepeso leve, e a fração de obesos é marcadamente menor.

Em transportes coletivos (ônibus e metrô), nota-se perfil intermediário, com distribuição mais balanceada entre sobrepeso e obesidade leve.

Portanto, esses padrões visuais corroboram o resultado do teste qui-quadrado, sugerindo que as escolhas de deslocamento estão diretamente relacionadas ao status de obesidade. Assim, conclui-se que há uma relação inversa de transportes ativos a medida que a níveis de obesidade crescem e o uso em grande parte de meios de transporte menos ativos como o transporte público, sendo portanto um fator importante para previsão da obesidade.

7.5. Visualização Gráfica para ajudar a responder a questão 2.



7.6. Análise de Distribuição de Peso Corporal por Consumo de Comida Calórica

Estatísticas de Peso por Consumo de Comida Calórica

Comida_Alta_Caloria	Peso.n	Peso.mean	Peso.sd	Peso.se
Não	245	66.91	17.10	1.09
Sim	1866	89.17	26.08	0.60

7.7. Visualização do T test para confirmação das diferenças nas médias entre Peso Corporal e o Consumo Calórico.

Resumo completo do Teste t

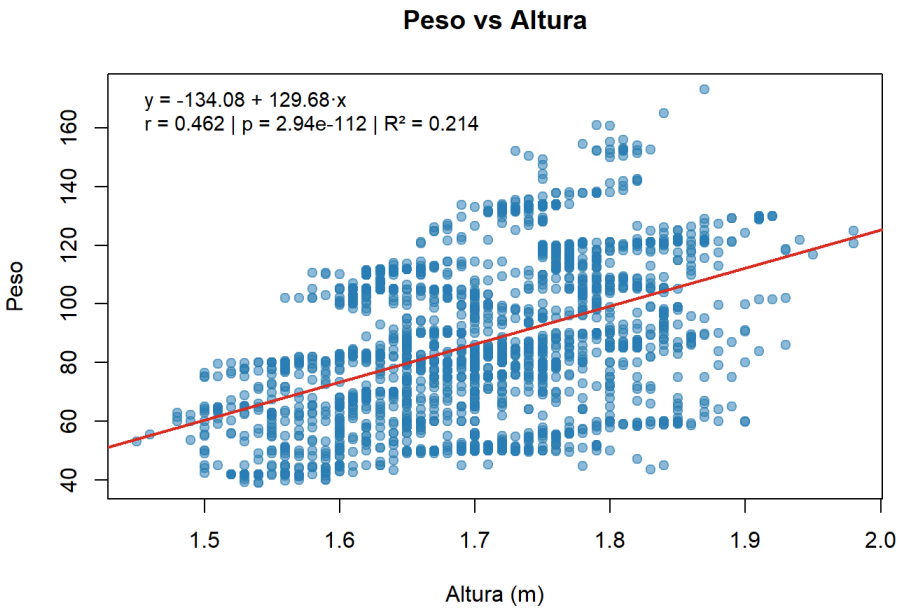
	Estatística_t	Graus_de_Liberdade	P_valor	Média_Grupo1	Média_Grupo2	IC_95_Lower	IC_95_Upper	Hipótese_Alternativa	Método	Concl
t	-17.832	411	0	66.91	89.17	-24.72	-19.81	two.sided	Welch Two Sample t-test	Difere signif

7.8.Análise e Resposta a segunda pergunta:

A análise descritiva evidenciou que indivíduos que não consomem comida alta em calorias apresentaram peso corporal médio de 66,91 kg (DP = 17,10; n = 245), enquanto aqueles que consomem apresentaram média de 89,17 kg (DP = 26,08; n = 1866). A diferença entre as médias foi de aproximadamente 22,26 kg. Para verificar a significância estatística dessa diferença, foi realizado o teste t de Welch para amostras independentes, cujo resultado indicou valor t = -17,832, com aproximadamente 411 graus de liberdade e valor-p < 0,001. O intervalo de confiança de 95% para a diferença de médias variou de -24,72 kg a -19,81 kg, não incluindo o valor nulo, o que confirma a robustez do resultado. Assim, conclui-se que indivíduos que consomem comida alta em calorias frequentemente tendem a apresentar maior peso corporal em comparação àqueles que não consomem, sendo essa diferença estatisticamente significativa. ´ Portanto, sim há uma relação muito significativo como observado pelo uso de test t pelas diferenças entre médias e desvio padrão e pelo próprio padrão observado no gráfico há uma maior peso associado a consumo de comida de alta calorias, havendo portanto uma boa relação com a obesidade em si.

7.9. Visualização Gráfica para ajudar a responder a questão 2.

Inicialmente será plotado aqui novamente o gráfico que mostra apenas a distribuição e relação entre peso e altura.



7.10. Separação entre grupos com histórico familiar e aqueles sem histórico de obesidade e definição do modelo de regressão linear.

Coeficientes do Modelo - Grupo com Histórico Familiar 'Sim'

Termo	Estimativa	Erro_Padrao	t_valor	p_valor
(Intercept)	-88.5938	10.2890	-8.6106	1.61e-17
Altura	105.8813	5.9999	17.6472	3.35e-64

Resumo dos Resíduos - Grupo com Histórico Familiar 'Sim'

Estatistica	Valor
Min.	-61.2278
1º Quartil	-15.2193
Mediana	0.1219
3º Quartil	16.5720
Máx.	63.5958

Métricas Gerais do Modelo - Grupo com Histórico Familiar 'Sim'

Estatistica	Valor
R-quadrado	0.153
R-quadrado Ajustado	0.1525
Erro Padrão Residual	22.308
F-Statistic	311.4225

Estatistica	Valor
Graus de Liberdade 1	1
Graus de Liberdade 2	1724
p-valor F	3.35e-64

Coeficientes do Modelo - Grupo com Histórico Familiar 'Não'

Termo	Estimativa	Erro_Padrao	t_valor	p_valor
(Intercept)	-78.9991	10.5101	-7.5165	4.02e-13
Altura	83.5215	6.3487	13.1556	7.10e-33

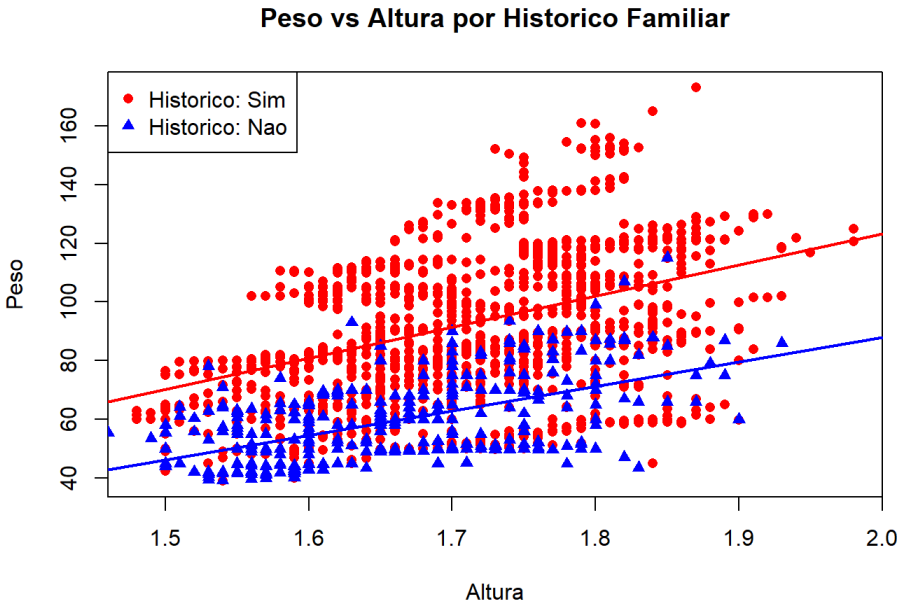
Resumo dos Resíduos - Grupo com Histórico Familiar 'Não'

Estatistica	Valor
Min.	-30.3452
1º Quartil	-9.6353
Mediana	-1.4987
3º Quartil	10.1647
Máx.	39.4844

Métricas Gerais do Modelo - Grupo com Histórico Familiar 'Não'

Estatistica	Valor
R-quadrado	0.3112
R-quadrado Ajustado	0.3094
Erro Padrão Residual	11.7852
F-Statistic	173.0705
Graus de Liberdade 1	1
Graus de Liberdade 2	383
p-valor F	7.10e-33

7.11. Plotagem de Gráfico para visualização dos dois grupos em relação ao peso e altura.



7.12. Utilização de Regressão linear com interação para ver se a influencia altura peso é

diferente ao interagir com o histórico familiar de obesidade

Coeficientes do Modelo

Termo	Estimativa	Erro_Padrao	t_valor	p_valor
(Intercept)	-78.9991	18.5450	-4.2599	2.14e-05
Altura	83.5215	11.2024	7.4557	1.30e-13
Historico_Familiar_ObesidadeSim	-9.5946	20.8784	-0.4595	0.6459
Altura:Historico_Familiar_ObesidadeSim	22.3598	12.5209	1.7858	0.0743

Resumo dos Resíduos

Estatistica	Valor
Min.	-61.2278
1º Quartil	-13.4044
Mediana	-0.3456
3º Quartil	15.2839
Máx.	63.5958

Métricas Gerais do Modelo

Estatistica	Valor
R-quadrado	0.370506421165722
R-quadrado Ajustado	0.369610132254235
Erro Padrão Residual	20.795023792418
F-Statistic	413.3783
Graus de Liberdade 1	3
Graus de Liberdade 2	2107
p-valor F	3.85e-211

7.13. Uso de IMC já que peso e altura tem a maior correlação entre si utilizando IMC que relaciona as duas conseguimos ao mesmo tempo relacionar com o historico familiar obesidade além de uso de t test para validar a possível hipótese.

Resultados do Teste t para IMC

Estatistica	Valor
Estatística t	-35.7944
Graus de Liberdade	1009.163
p-valor	8.31e-182
Média Grupo ‘Sim’	21.4997
Média Grupo ‘Não’	31.5331

Coeficientes do Modelo Linear para IMC

Termo	Estimativa	Erro_Padrao	t_valor	p_valor
(Intercept)	21.4997	0.3577	60.1060	0.00e+00
Historico_Familiar_ObesidadeSim	10.0333	0.3956	25.3633	4.31e-124

Resumo dos Resíduos do Modelo IMC

Estatistica	Valor
Min.	-18.2415
1º Quartil	-4.6408
Mediana	-0.0315

Estatistica	Valor
3º Quartil	4.9265
Máx.	19.2872

Métricas Gerais do Modelo Linear para IMC

Estatistica	Valor
R-quadrado	0.2337
R-quadrado Ajustado	0.2334
Erro Padrão Residual	7.0185
F-Statistic	643.2993
Graus de Liberdade 1	1
Graus de Liberdade 2	2109
p-valor F	4.31e-124

7.14. Análise e Resposta a terceira pergunta:

Para investigar se a relação entre altura e peso corporal apresenta diferenças significativas entre indivíduos com e sem histórico familiar de obesidade, realizamos uma análise de regressão linear separada para cada grupo, bem como uma análise de regressão com termo de interação.

Modelos de regressão separados:

Para o grupo com histórico familiar de obesidade, o modelo de regressão linear mostrou que a altura é um preditor significativo do peso corporal ($p < 0.001$), com coeficiente estimado de 105.88 (Erro padrão = 6.00). O modelo apresentou um R^2 de 0.153, indicando que aproximadamente 15,3% da variação do peso é explicada pela altura neste grupo.

Para o grupo sem histórico familiar, a altura também foi um preditor significativo do peso ($p < 0.001$), com coeficiente estimado de 83.52 (Erro padrão = 6.35). Neste caso, o R^2 foi de 0.311, indicando maior explicação da variação do peso pela altura nesse grupo (31,1%).

Modelo de regressão com interação:

Ao incluir o termo de interação entre altura e histórico familiar de obesidade no modelo, verificamos que o coeficiente da interação (Altura:Histórico_Familiar_ObesidadeSim) foi positivo (22.36) e próximo do limiar de significância estatística ($p = 0.0743$), sugerindo uma tendência para que a relação entre altura e peso seja mais acentuada para indivíduos com histórico familiar de obesidade, embora não tenha atingido significância estatística ao nível de 5%.

Visualização gráfica:

Os gráficos de dispersão e as linhas de regressão ajustadas para cada grupo corroboram os resultados, mostrando inclinações maiores para o grupo com histórico familiar, evidenciando uma relação mais forte entre altura e peso nesse grupo.

Conclusão:

Os dados indicam que existe uma diferença na relação entre altura e peso corporal entre indivíduos com e sem histórico familiar de obesidade, sendo a relação mais forte (maior coeficiente) para aqueles com histórico familiar. Contudo, essa diferença não atingiu significância estatística convencional no modelo de interação ($p > 0.05$), sugerindo que, embora haja uma tendência, não podemos afirmar com plena confiança estatística que a influência da altura no peso difere de forma significativa entre os grupos.