UNIVERSIDADE PRESBITERIANA MACKENZIE

Curso: Tecnologia de Ciência de Dados
Polo de apoio presencial: Higienópolis
Componente curricular / TEMA: Projeto Aplicado I
Nome completo do aluno: Natália Albuquerque

Walmart Sales project

Sumário

1. Introdução

Breve contextualização do estudo, justificativa para escolha do tema e relevância da análise de dados na tomada de decisão empresarial ou acadêmica.

O estudo analisa o conjunto de dados de vendas semanais de uma rede de lojas Walmart, contendo informações sobre vendas, temperatura, preços do combustível, CPI, desemprego e feriados. O objetivo é entender como esses fatores influenciam as vendas no varejo.

O tema foi escolhido devido à importância de compreender os fatores que afetam as vendas no setor varejista. A análise dos dados históricos ajuda as empresas a otimizar estratégias de marketing, precificação e estoque, além de prever o comportamento de consumo.

A análise de dados é crucial para tomar decisões informadas, tanto no setor empresarial quanto acadêmico. Ela permite que as empresas melhorem sua competitividade e estratégias, enquanto no ambiente acadêmico, contribui para a formulação de teorias e validação de modelos baseados em dados.

2. Objetivo do Estudo

Definir os objetivos gerais e específicos do projeto. Explicar o que se pretende extrair dos dados e como a análise pode ajudar na resolução do problema proposto.

Objetivo Geral:

Analisar o impacto de variáveis externas, como temperatura, preço do combustível, índice de preços ao consumidor (CPI), taxa de desemprego e feriados, sobre as vendas semanais de uma rede de lojas Walmart, com o objetivo de identificar padrões e insights que possam otimizar a estratégia de vendas e tomada de decisões no varejo.

Objetivos Específicos:

1. Analisar a distribuição das vendas semanais para identificar tendências sazonais e flutuações significativas.

2. Investigar a relação entre variáveis externas (temperatura, preço do combustível, CPI, desemprego, feriados) e as vendas semanais, verificando se existe correlação entre esses fatores e o desempenho das vendas.

3. Avaliar o impacto dos feriados nas vendas semanais, verificando se há aumento ou diminuição nas vendas durante períodos de feriados.

4. Identificar padrões de comportamento de compra que possam ser usados para otimizar estratégias de marketing, estoque e precificação.

5. Gerar recomendações práticas para a empresa com base nos insights obtidos, ajudando a melhorar a previsão de vendas e a estratégia de operações.

A análise dos dados visa identificar como fatores externos influenciam as vendas, permitindo à empresa otimizar suas decisões em várias áreas, como:

- Marketing e Promoções: Melhor planejamento de campanhas publicitárias com base nas variáveis que impactam as vendas.

- Gestão de Estoque: Antecipação da demanda durante períodos específicos, como feriados ou variações sazonais de temperatura.

- Precificação: Ajustes no preço do produto com base em fatores econômicos, como o CPI e preço do combustível, para manter a competitividade e aumentar a margem de lucro.

Essa análise ajudará a empresa a adotar uma abordagem mais assertiva e informada, com base em dados reais, o que pode resultar em maior eficiência e lucratividade.

3. Apresentação da Empresa e do Problema

O Walmart é uma das maiores redes de varejo do mundo, com presença em diversos países e uma ampla gama de produtos, desde alimentos até eletrônicos e roupas. Fundada em 1962, a empresa se destaca por sua estratégia de preços baixos, o que atrai milhões de consumidores todos os dias. No contexto do setor varejista, empresas como o Walmart enfrentam uma série de desafios, como a sazonalidade das vendas, flutuações econômicas, mudanças nos hábitos de consumo e a necessidade constante de otimizar estoques e estratégias de marketing. O setor varejista é altamente competitivo e sensível a variáveis externas, como clima, preços de combustíveis e condições econômicas, o que torna a análise de dados uma ferramenta essencial para a tomada de decisões estratégicas.

Descrição do Problema Identificado:

O principal problema identificado é a dificuldade de prever com precisão as vendas semanais, devido à influência de múltiplos fatores externos, como variações sazonais, preços do combustível, temperatura, feriados e condições econômicas. A falta de previsibilidade nas vendas pode resultar em problemas no planejamento de estoque, campanhas de marketing ineficazes e perda de oportunidades de venda. Além disso, os períodos de feriados podem causar picos inesperados de vendas, o que dificulta a preparação adequada de recursos e estratégias de vendas.

Hipóteses a Serem Testadas por Meio dos Dados:

1. Hipótese 1: Existe uma correlação positiva entre a temperatura e as vendas semanais, com vendas mais altas em semanas mais quentes.

2. Hipótese 2: Os feriados têm um impacto significativo nas vendas semanais, aumentando as vendas em períodos de feriados em comparação com semanas normais.

3. Hipótese 3: O preço do combustível afeta negativamente as vendas, com vendas mais baixas em semanas em que o preço do combustível está mais alto.

4. Hipótese 4: A taxa de desemprego tem uma correlação negativa com as vendas, com vendas mais baixas em períodos de maior desemprego.

5. Hipótese 5: O Índice de Preços ao Consumidor (CPI) influencia as vendas, com vendas mais baixas durante períodos de aumento no CPI, devido à maior pressão inflacionária sobre os consumidores.

4. Apresentação dos Metadados

- Fonte dos dados (ex: Kaggle, bases públicas, dados empresariais).
- Descrição dos atributos (variáveis do dataset).
- Período da coleta dos dados.
- Possíveis limitações dos dados.

Os dados utilizados no estudo foram extraídos do Kaggle, uma plataforma amplamente utilizada para compartilhamento e análise de datasets públicos. O conjunto de dados denominado “Walmart Sales", contém informações sobre vendas semanais de lojas do Walmart, além de variáveis externas que podem influenciar o desempenho das vendas.

O dataset contém as seguintes variáveis:

Store: Identificador da loja.

Date: Data correspondente à semana de vendas.

Weekly_Sales: Total de vendas semanais registradas para a loja correspondente.

Holiday_Flag: Indicador binário (1 = semana com feriado, 0 = semana sem feriado).

Temperature: Temperatura média registrada na semana.

Fuel_Price: Preço médio do combustível na região da loja.

CPI (Consumer Price Index): Índice de preços ao consumidor, refletindo o nível médio de preços de bens e serviços.

Unemployment: Taxa de desemprego na região da loja.

Período da Coleta dos Dados:

Os dados foram coletados entre fevereiro de 2010 e outubro de 2012, cobrindo aproximadamente dois anos e oito meses de vendas semanais.

Possíveis Limitações dos Dados:

1. Falta de Detalhamento sobre as Lojas: O dataset não fornece informações adicionais sobre o perfil das lojas (ex: localização exata, tamanho, volume de estoque), o que poderia influenciar nas análises.

2. Ausência de Dados sobre Promoções: Não há informações sobre descontos ou campanhas promocionais, que podem impactar significativamente as vendas.

3. Agregação Semanal: As vendas são reportadas em nível semanal, o que pode ocultar variações diárias importantes no comportamento do consumidor.

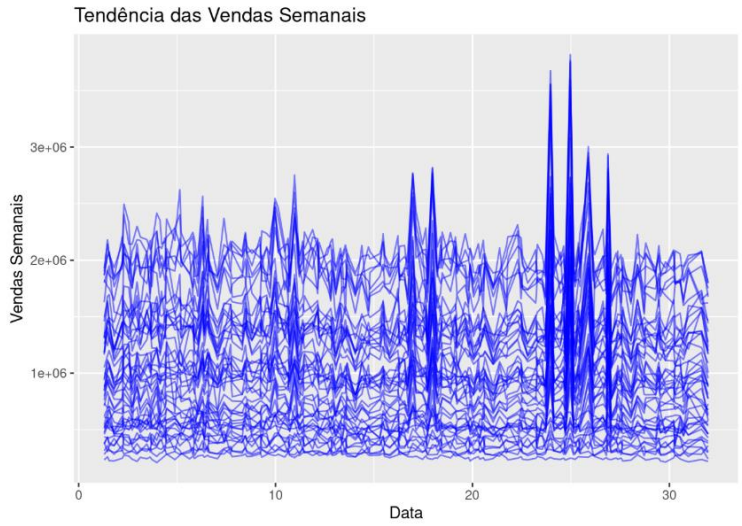
4. Fatores Externos Não Considerados: Elementos como eventos sazonais (ex: Natal, Black Friday), concorrência local e mudanças na economia global não estão explicitamente representados nos dados.

5. Análise Exploratória de Dados

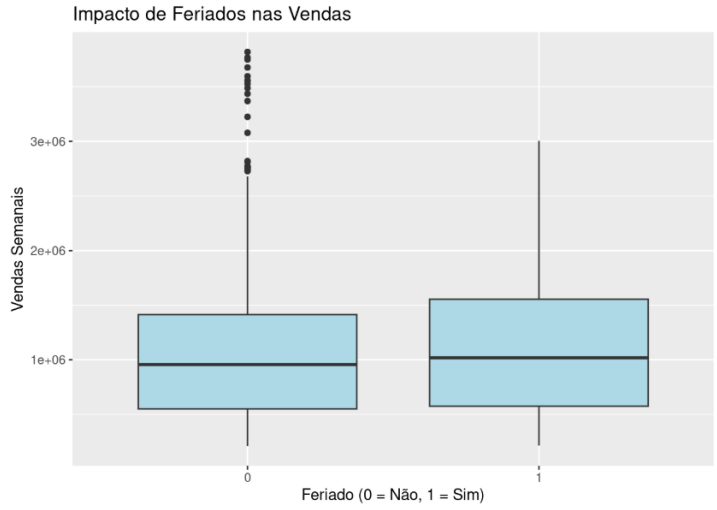
O arquivo do relatório está anexado junto ao projeto.

Mas aqui vou colocar os gráficos que gerei diante dos dados estudados.

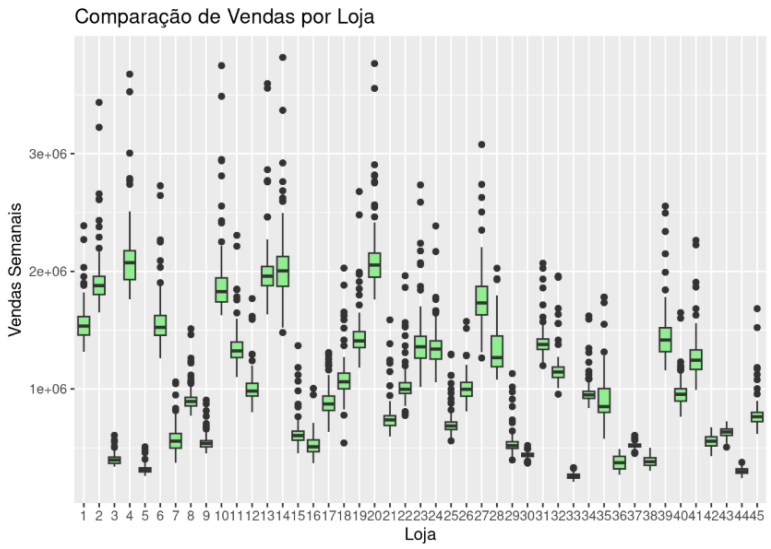
```
ggplot(df, aes(x = Date, y = Weekly_Sales, group = Store)) +
  geom_line(alpha = 0.5, color = "blue") +
  labs(title = "Tendência das Vendas Semanais", x = "Data", y = "Vendas Semanais")
```



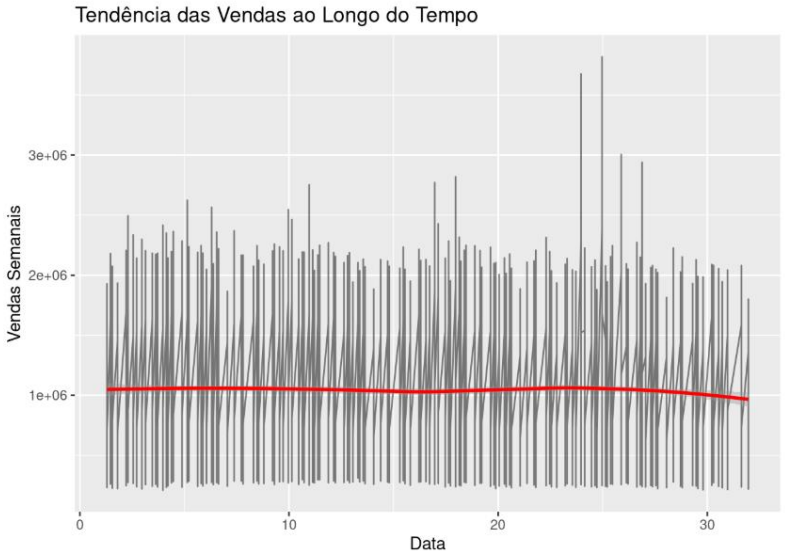

```
ggplot(df, aes(x = as.factor(Holiday_Flag), y = Weekly_Sales)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Impacto de Feriados nas Vendas", x = "Feriado (0 = Não, 1 = Sim)", y = "Vendas Semanais")
```



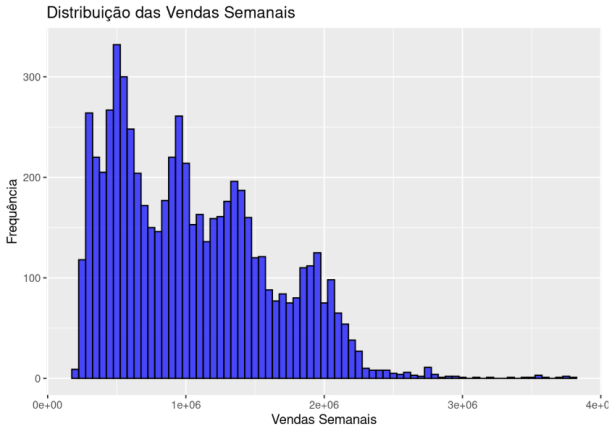
```
ggplot(df, aes(x = Store, y = Weekly_Sales)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Comparação de Vendas por Loja", x = "Loja", y = "Vendas Semanais")
```



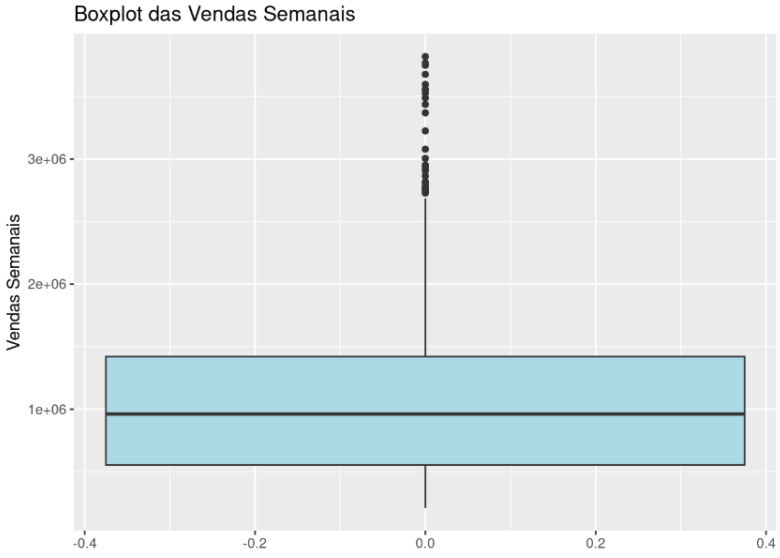

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(df, aes(x = Weekly_Sales)) +
  geom_histogram(binwidth = 50000, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribuição das Vendas Semanais", x = "Vendas Semanais", y = "Frequência")
```




```
ggplot(df, aes(y = Weekly_Sales)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot das Vendas Semanais", y = "Vendas Semanais")
```



6. Síntese da Resolução Analítica

- Metodologia utilizada na análise.
- Principais resultados obtidos.
- Implicações da análise para o problema identificado

7. Referências

Listagem das fontes utilizadas para embasar o estudo (artigos, livros, bases de dados, etc.).

Untitled

2025-04-01

```
df <- read.csv("Walmart_Sales.csv")
head(df)
```

```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI
## 1     1 05-02-2010      1643691            0       42.31      2.572 211.0964
## 2     1 12-02-2010      1641957            1       38.51      2.548 211.2422
## 3     1 19-02-2010      1611968            0       39.93      2.514 211.2891
## 4     1 26-02-2010      1409728            0       46.63      2.561 211.3196
## 5     1 05-03-2010      1554807            0       46.50      2.625 211.3501
## 6     1 12-03-2010      1439542            0       57.79      2.667 211.3806
##   Unemployment
## 1        8.106
## 2        8.106
## 3        8.106
## 4        8.106
## 5        8.106
## 6        8.106
```

```
# Verificar valores ausentes
sum(is.na(df))
```

```
## [1] 0
```

```
# Contar valores ausentes por coluna
colSums(is.na(df))
```

```
##        Store         Date Weekly_Sales Holiday_Flag  Temperature   Fuel_Price
##            0            0            0            0            0            0
##          CPI Unemployment
##            0            0
```

```
# Remover linhas com valores ausentes
df_clean <- na.omit(df)
# Verificar duplicatas
sum(duplicated(df))
```

```
## [1] 0
```

```
# Converter a coluna "Date" para o formato Date
df$Date <- as.Date(df$Date, format="%Y-%m-%d")
# Verificar se a conversão foi bem-sucedida
head(df$Date)
```

```
## [1] "5-02-20"  "12-02-20" "19-02-20" "26-02-20" "5-03-20"  "12-03-20"
```

```
df$Date <- as.Date(df$Date, format="%d-%m-%Y")
head(df$Date)
```

```
## [1] "5-02-20"  "12-02-20" "19-02-20" "26-02-20" "5-03-20"  "12-03-20"
```

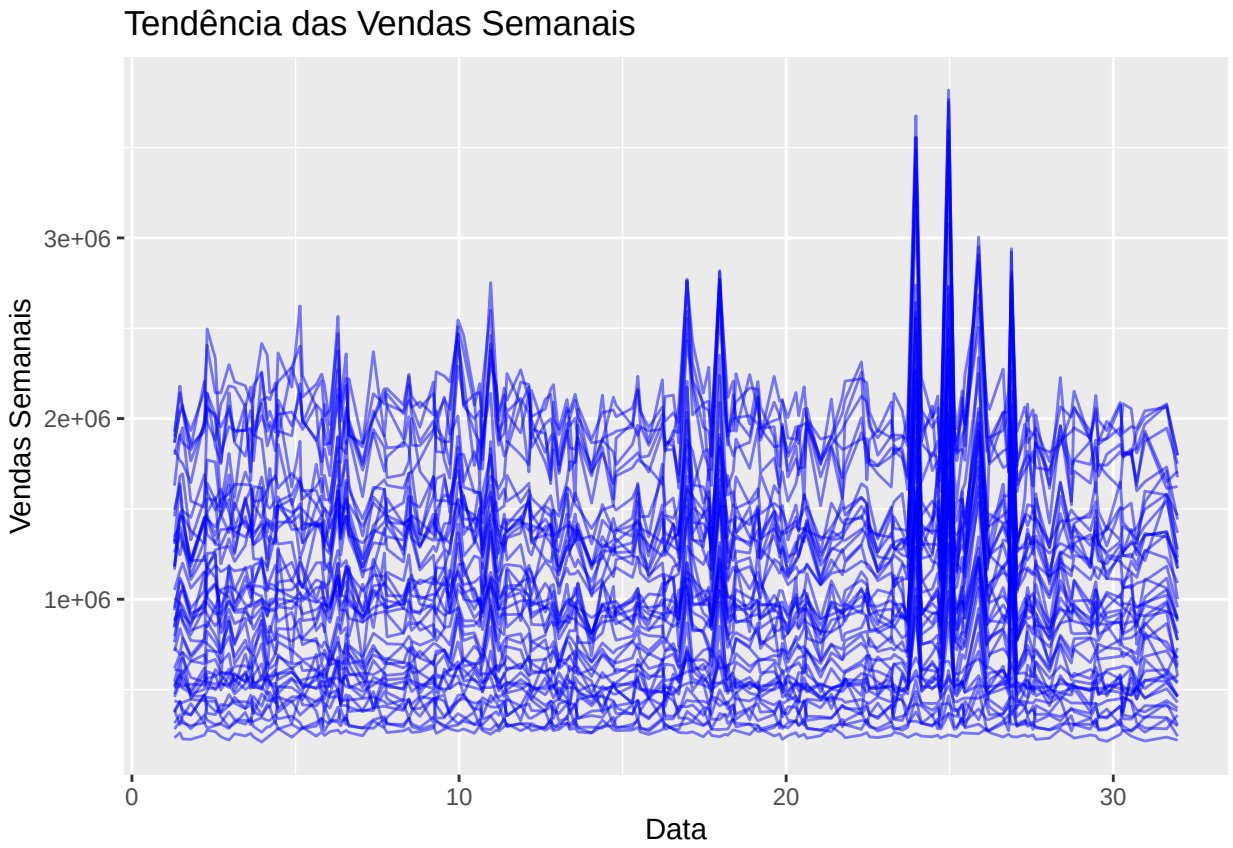
```r
df$Store <- as.factor(df$Store)
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

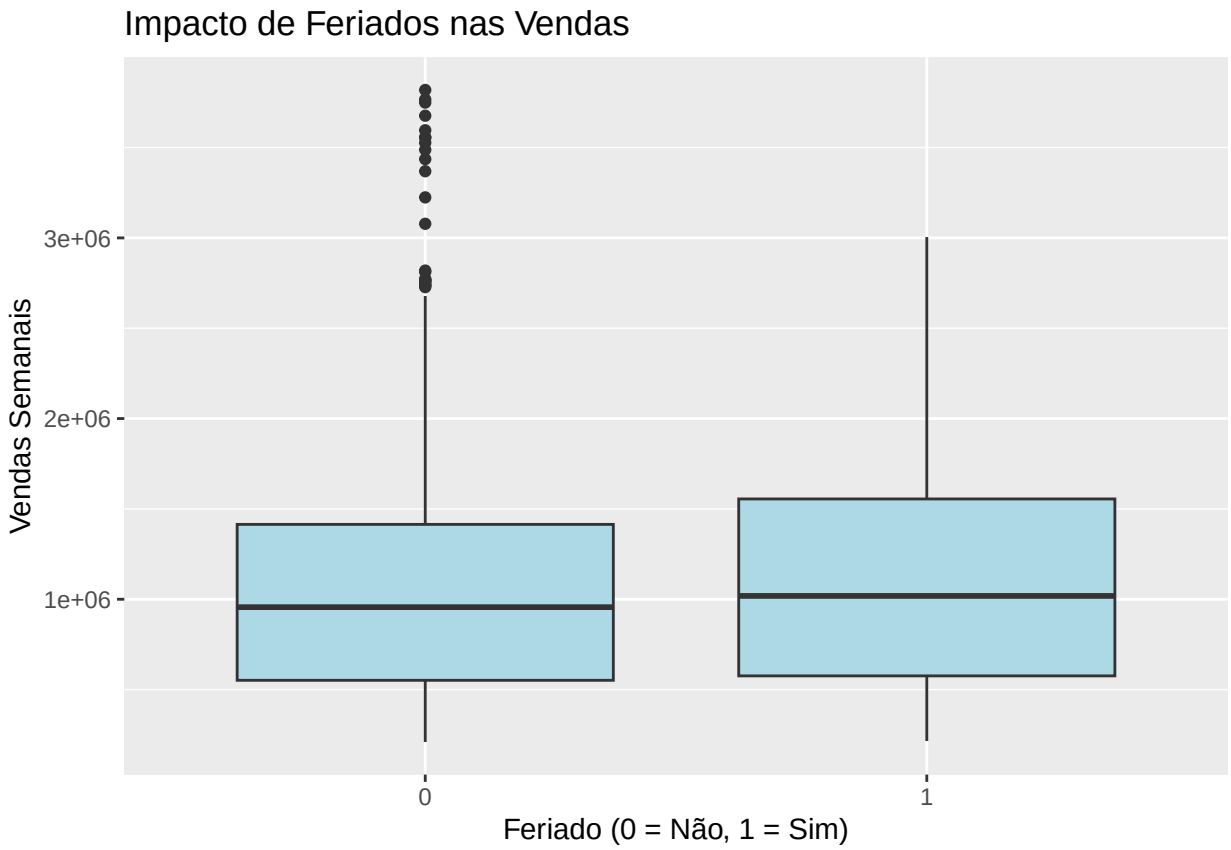
```r
library(ggplot2)
summary(df)  # Resumo geral
```

```
##      Store           Date           Weekly_Sales      Holiday_Flag
##  1      : 143   Min.   :1-04-20    Min.   : 209986   Min.   :0.00000
##  2      : 143   1st Qu.:8-07-20    1st Qu.: 553350   1st Qu.:0.00000
##  3      : 143   Median :16-04-20   Median : 960746   Median :0.00000
##  4      : 143   Mean   :16-03-07   Mean   :1046965   Mean   :0.06993
##  5      : 143   3rd Qu.:23-12-20   3rd Qu.:1420159   3rd Qu.:0.00000
##  6      : 143   Max.   :31-12-20   Max.   :3818686   Max.   :1.00000
##  (Other):5577
##   Temperature       Fuel_Price         CPI         Unemployment
##  Min.   : -2.06   Min.   :2.472   Min.   :126.1   Min.   : 3.879
##  1st Qu.: 47.46   1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891
##  Median : 62.67   Median :3.445   Median :182.6   Median : 7.874
##  Mean   : 60.66   Mean   :3.359   Mean   :171.6   Mean   : 7.999
##  3rd Qu.: 74.94   3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622
##  Max.   :100.14   Max.   :4.468   Max.   :227.2   Max.   :14.313
##
```

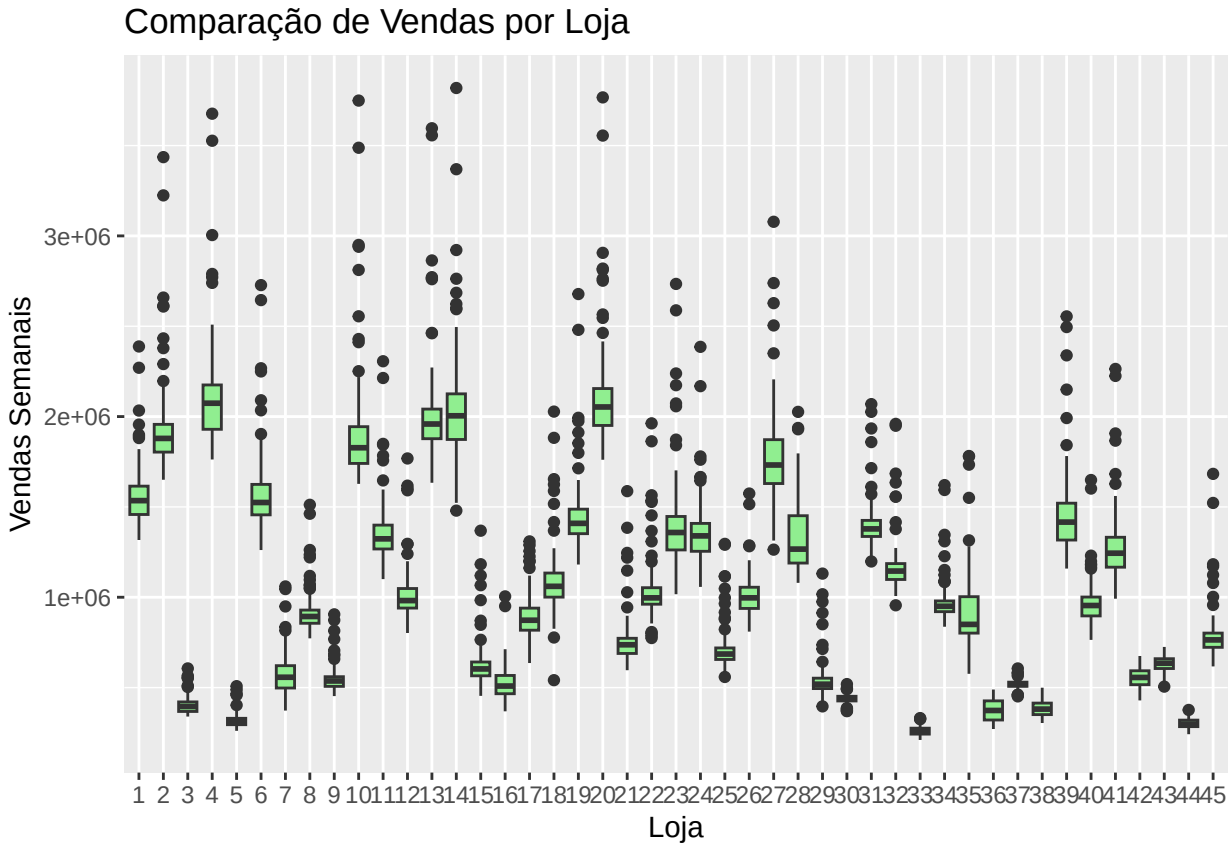
```r
ggplot(df, aes(x = Date, y = Weekly_Sales, group = Store)) +
  geom_line(alpha = 0.5, color = "blue") +
  labs(title = "Tendência das Vendas Semanais", x = "Data", y = "Vendas Semanais")
```




```
ggplot(df, aes(x = as.factor(Holiday_Flag), y = Weekly_Sales)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Impacto de Feriados nas Vendas", x = "Feriado (0 = Não, 1 = Sim)", y = "Vendas Semanais
```

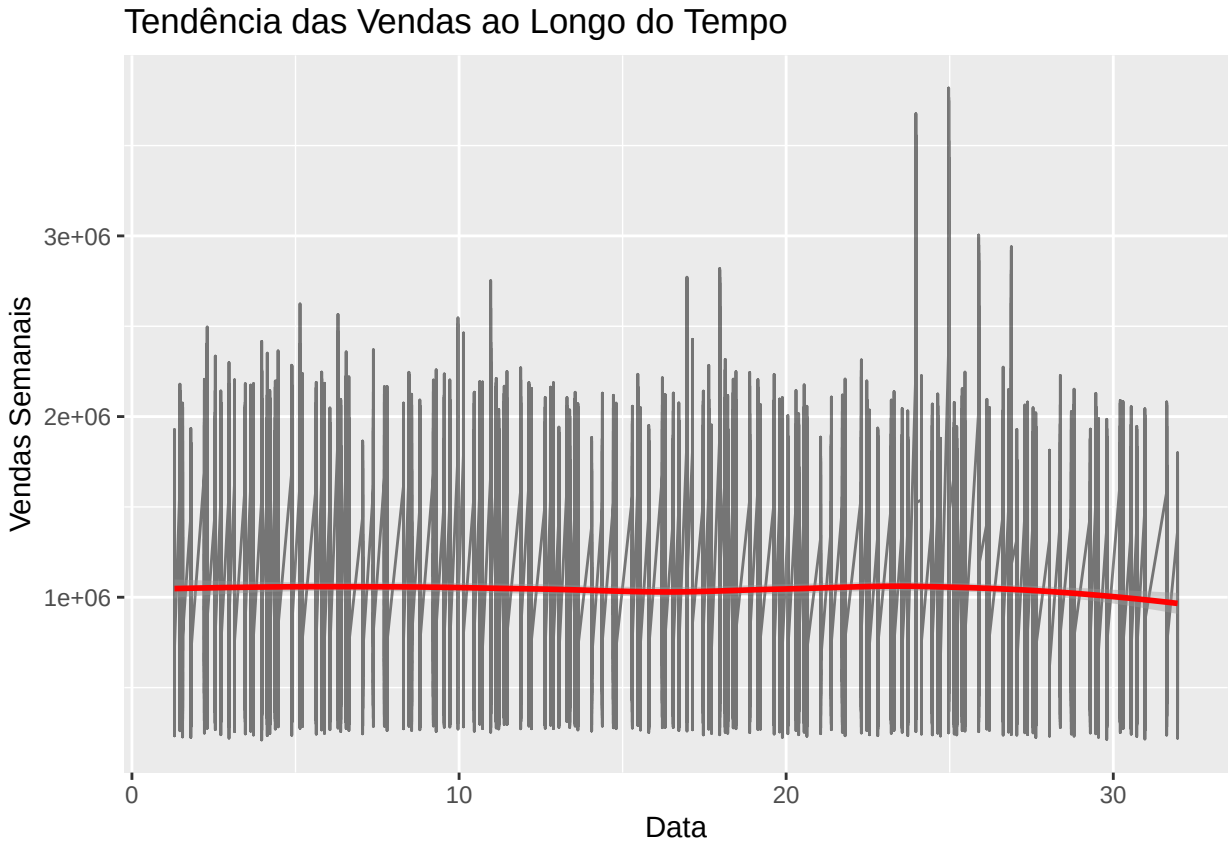



```
ggplot(df, aes(x = Store, y = Weekly_Sales)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Comparação de Vendas por Loja", x = "Loja", y = "Vendas Semanais")
```




```
ggplot(df, aes(x = Date, y = Weekly_Sales)) +
  geom_line(alpha = 0.5) +
  geom_smooth(method = "loess", color = "red") +
  labs(title = "Tendência das Vendas ao Longo do Tempo", x = "Data", y = "Vendas Semanais")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



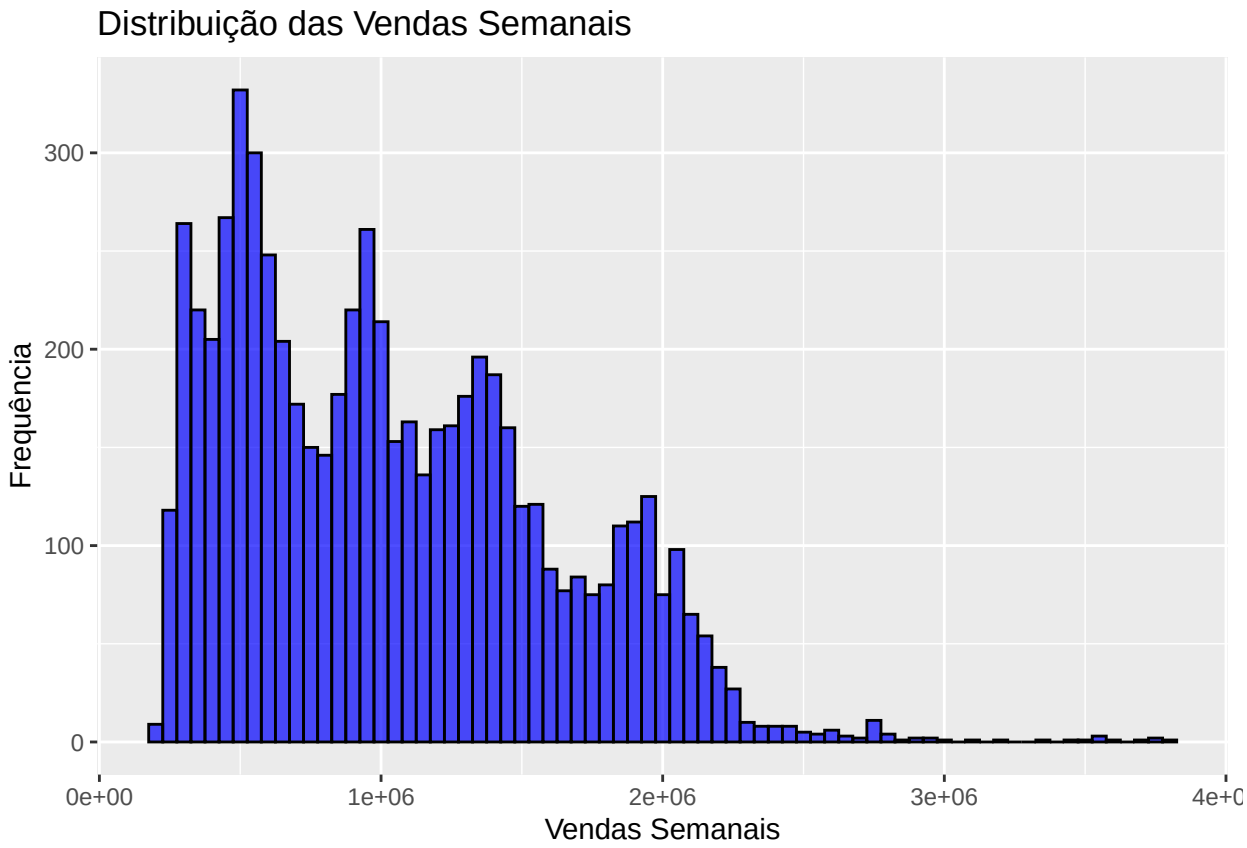

```
mean_sales <- mean(df$Weekly_Sales)  # Média
median_sales <- median(df$Weekly_Sales)  # Mediana

mean_sales
```

```
## [1] 1046965
```

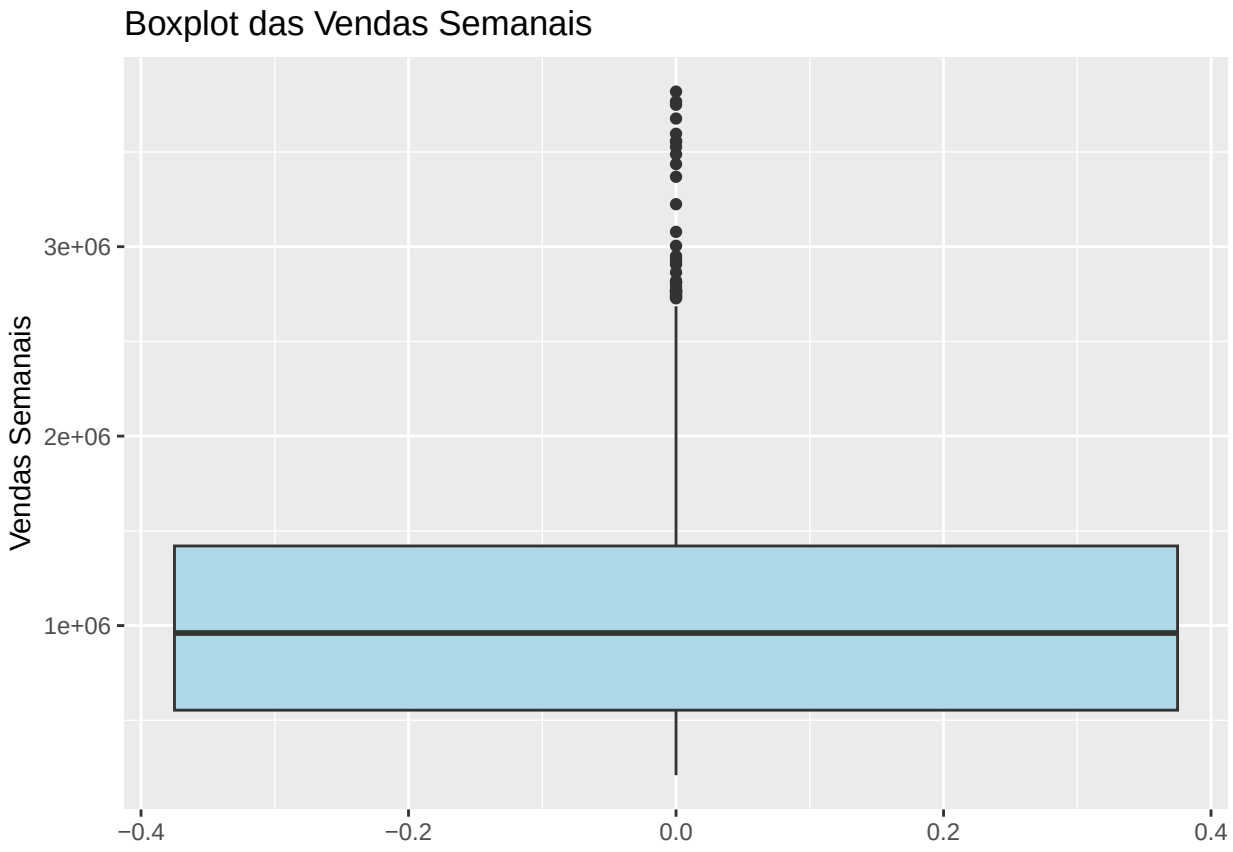
```
median_sales
```

```
## [1] 960746
```

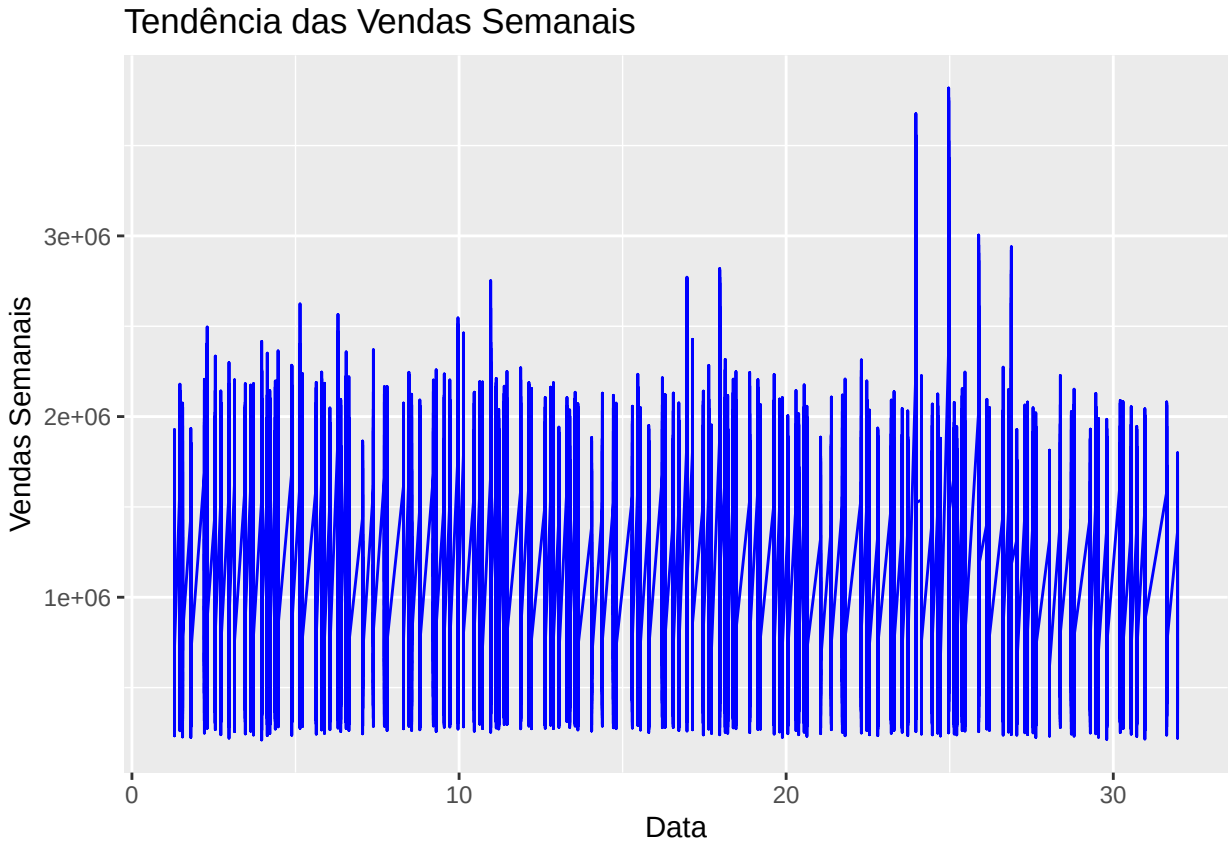
```
ggplot(df, aes(x = Weekly_Sales)) +
  geom_histogram(binwidth = 50000, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribuição das Vendas Semanais", x = "Vendas Semanais", y = "Frequência")
```




```
ggplot(df, aes(y = Weekly_Sales)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot das Vendas Semanais", y = "Vendas Semanais")
```

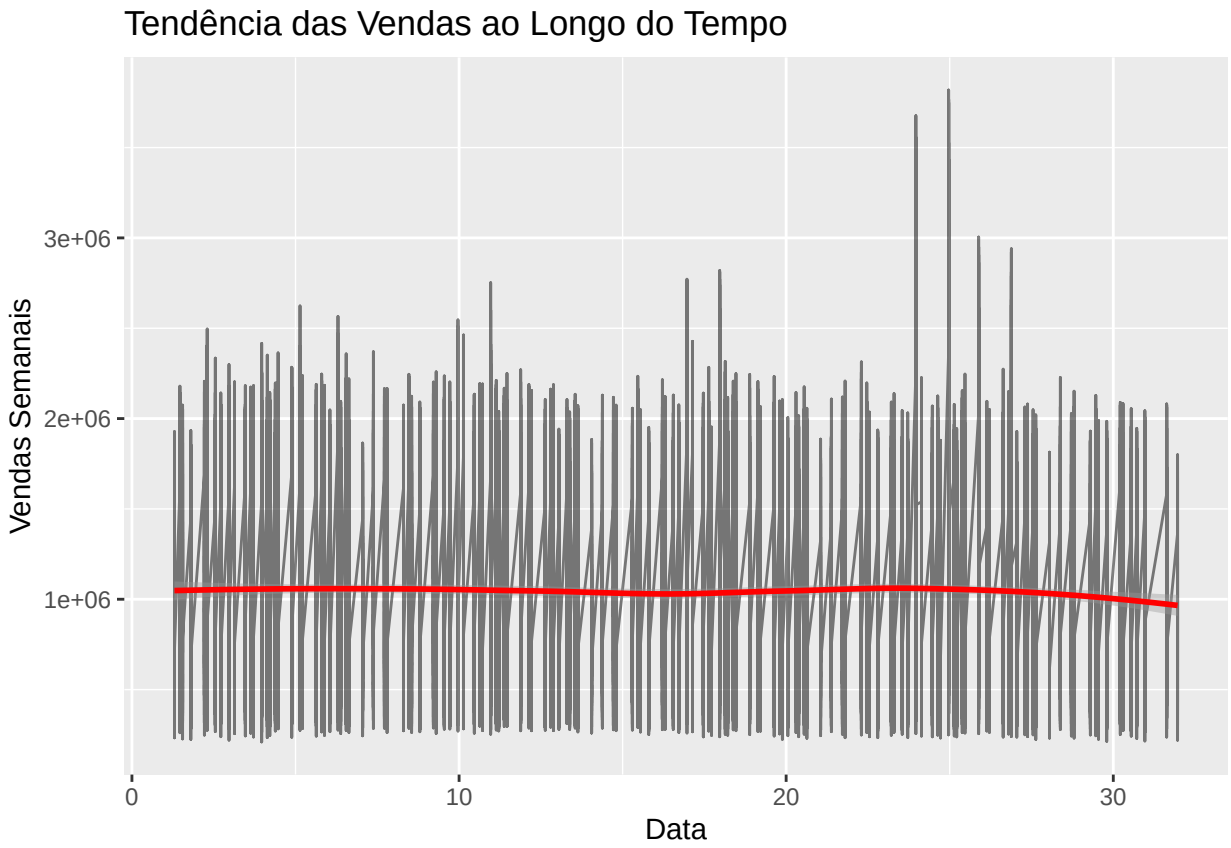



```
ggplot(df, aes(x = Date, y = Weekly_Sales)) +
  geom_line(color = "blue") +
  labs(title = "Tendência das Vendas Semanais", x = "Data", y = "Vendas Semanais")
```

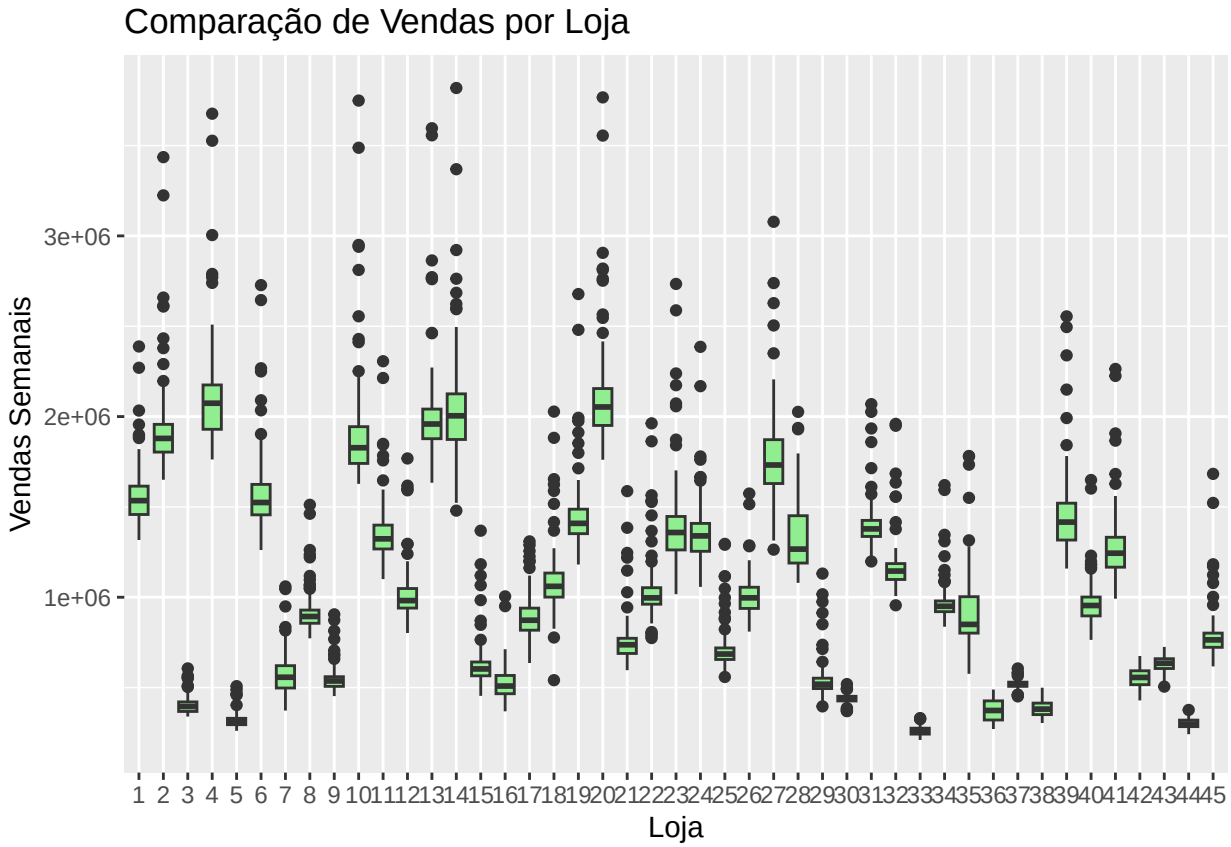



```
ggplot(df, aes(x = Date, y = Weekly_Sales)) +
  geom_line(alpha = 0.5) +
  geom_smooth(method = "loess", color = "red") +
  labs(title = "Tendência das Vendas ao Longo do Tempo", x = "Data", y = "Vendas Semanais")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
ggplot(df, aes(x = Store, y = Weekly_Sales)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Comparação de Vendas por Loja", x = "Loja", y = "Vendas Semanais")
```

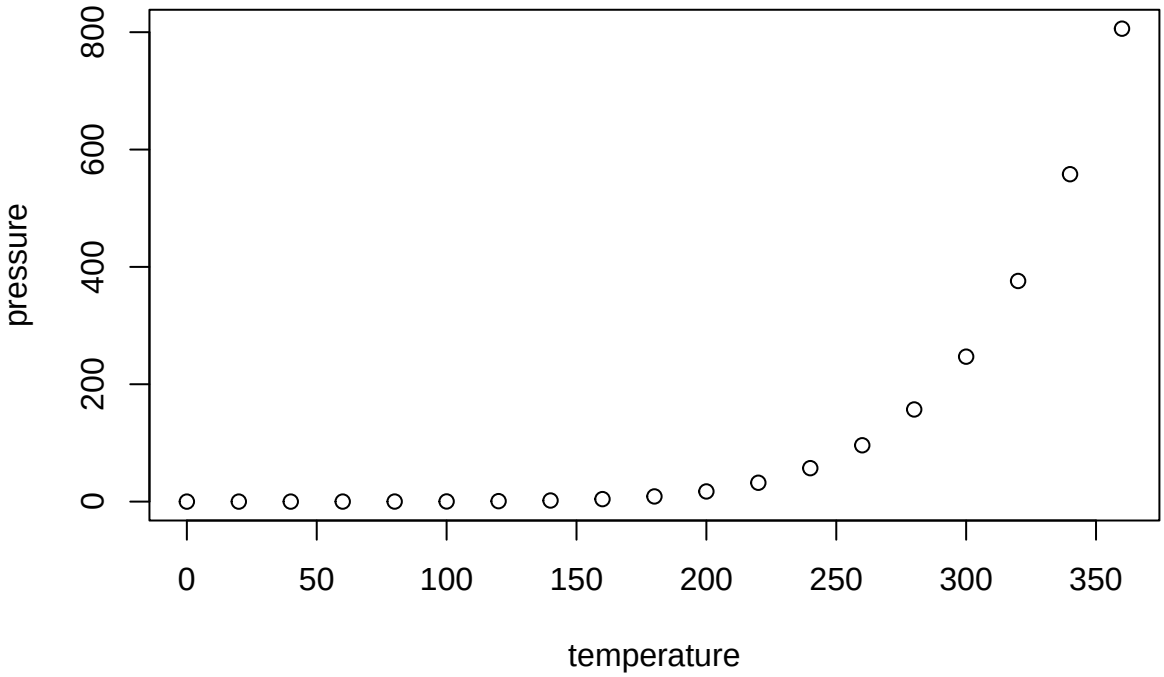



```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.