My Quarto Report

Rahma Aroua

Introduction

Context of the project

Carvana is a prominent online platform revolutionizing the way people buy and sell used cars. With its user-friendly interface and innovative business model, Carvana offers customers a hassle-free experience by allowing them to browse, purchase, finance, and even trade in their vehicles entirely online

Analyzing Carvana's dataset reveals market trends, informs predictive sales modeling, guides targeted marketing through customer segmentation, aids competitive analysis for strategic positioning, and drives operational efficiency improvements. In essence, it empowers businesses to make data-driven decisions and enhance customer experiences in the dynamic automotive market.

Data sources

Dataset came from Carvana company found at Kaaagle.com .

Data analysis

Importing Data

```
df <- read.csv("C:/Users/rahma/OneDrive/Documents/GL/R/vehicles.csv")
```

Cleaning and processing Data

Remove columns containing only missing values

```r
df <- df[, colSums(is.na(df)) < nrow(df)]
```

Remove rows with missing values

```r
df <- na.omit(df)
```

Remove rows with empty spaces

```r
df <- subset(df, !apply(df, 1, function(x) all(trimws(x) == "")))
```

Remove the “size” column

```r
df <- df[, !(names(df) == "size")]
```

Analysis and Interpretations

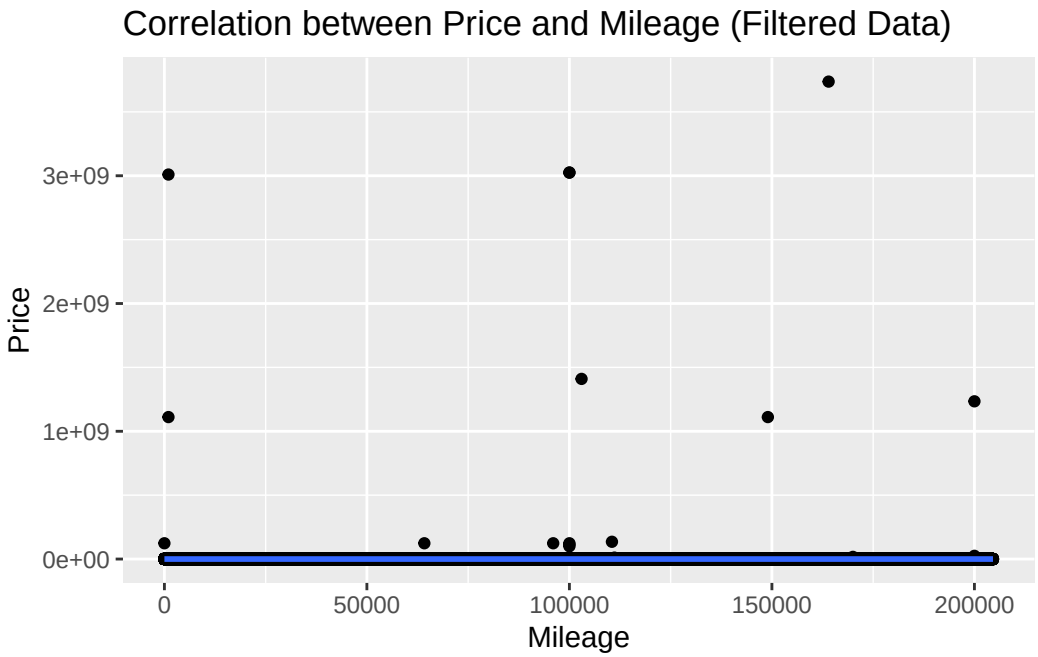
Analyzing the correlation between price and mileage (odometer)

```r
data <- read.csv("C:/Users/rahma/OneDrive/Documents/GL/R/vehicles_cleaned.csv")

# Load necessary libraries
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(ggplot2)

# Assuming 'data' is your dataset, replace it with your actual dataset
# For example:
# data <- read.csv("your_data.csv")



# Defining the threshold to filter the data
threshold <- 0.95 # For example, you can choose the 95th percentile

# Filtering the data to exclude outliers
filtered_data <- data %>%
  dplyr::filter(!is.na(odometer)) %>%
  dplyr::filter(odometer < quantile(odometer, threshold, na.rm = TRUE))

# Calculating the average prices for each mileage
average_prices <- filtered_data %>%
  dplyr::group_by(odometer) %>%
  dplyr::summarize(average_price = mean(price, na.rm = TRUE))

# Visualizing the scatter plot with filtered data and the average price curve
ggplot() +
  geom_point(data = filtered_data, aes(x = odometer, y = price)) +
  geom_smooth(data = average_prices, aes(x = odometer, y = average_price), method = "loess
  labs(title = "Correlation between Price and Mileage (Filtered Data)", x = "Mileage", y =
```

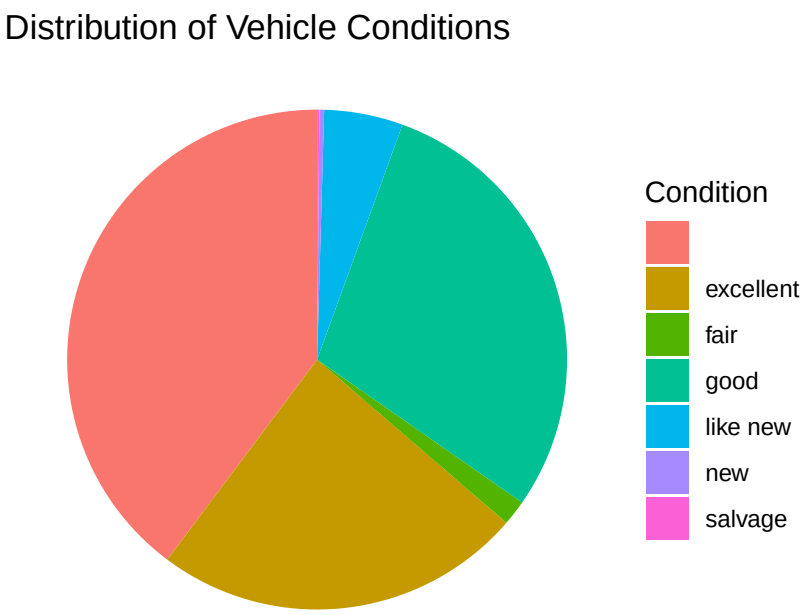
```
`geom_smooth()` using formula = 'y ~ x'
```



Vehicles with higher mileage tend to be priced lower, reflecting the depreciation of automotive assets. This insight can be leveraged by sellers to justify competitive pricing and attract buyers by emphasizing the economic benefits of high-mileage vehicles.

Pie chart of vehicle conditions

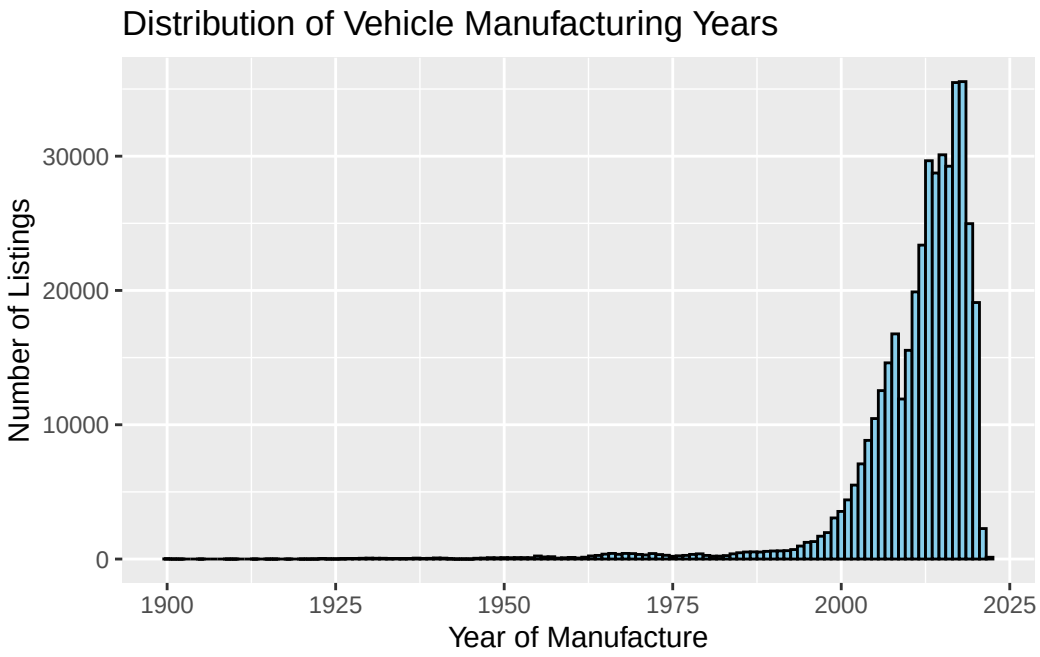
```
ggplot(data, aes(x = "", fill = condition)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Vehicle Conditions", fill = "Condition") +
  theme_void()
```



A prevalence of vehicles in good condition may indicate increased demand for quality vehicles, potentially leading to higher prices for well-maintained vehicles. Dealerships can use this information to justify higher prices by emphasizing the quality and reliability of the vehicle.

Histogram of manufacturing years

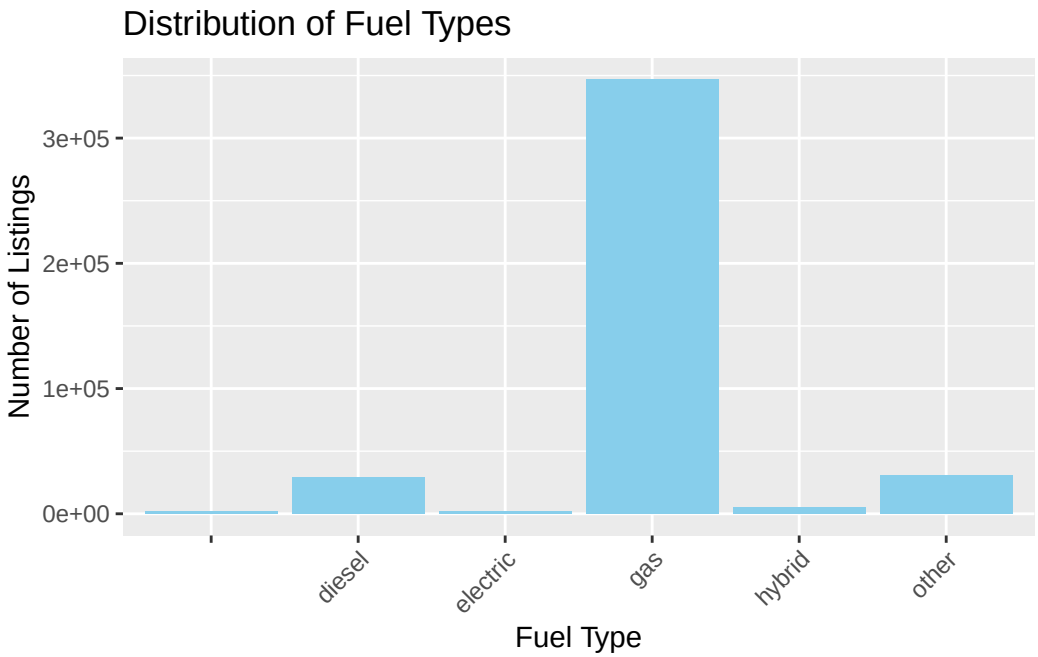
```
ggplot(data, aes(x = year)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Vehicle Manufacturing Years", x = "Year of Manufacture", y
```



Peaks in the histogram may indicate years of high production for popular models or periods of economic recession. This information can be used to plan model availability and guide the promotion of new models or the sale of used vehicles.

Bar chart of fuel types

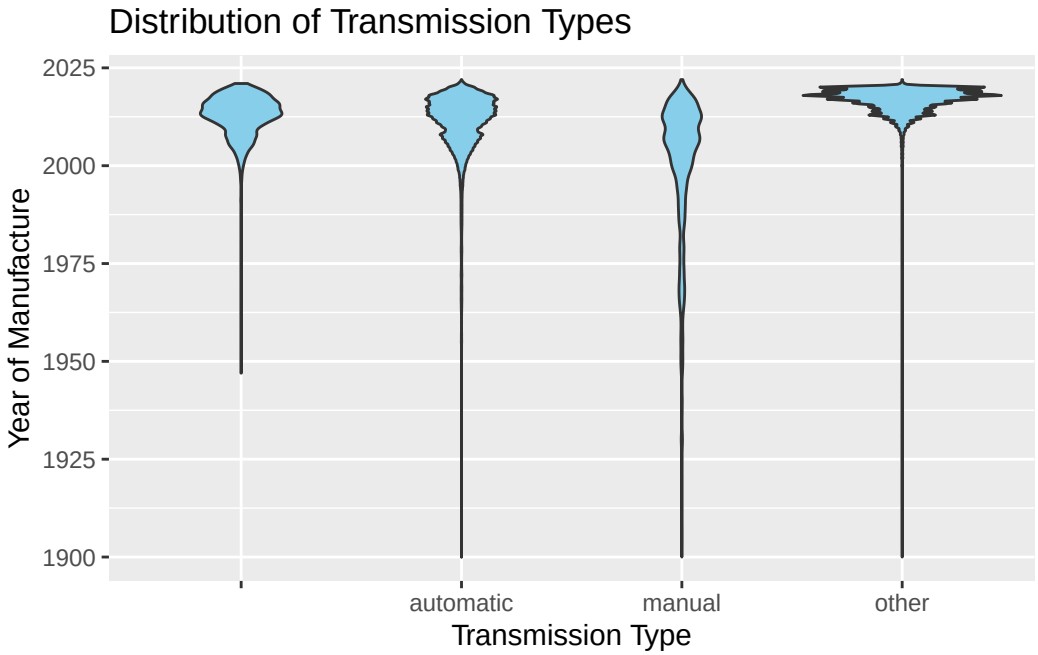
```
ggplot(data, aes(x = fuel)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Fuel Types", x = "Fuel Type", y = "Number of Listings") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The distribution of fuel types can reflect consumer preferences and trends in the adoption of alternative propulsion technologies. Manufacturers can use this information to guide the development of new models, and dealerships can offer a diverse range of vehicles powered by different fuel types.

Violin plot of transmission types

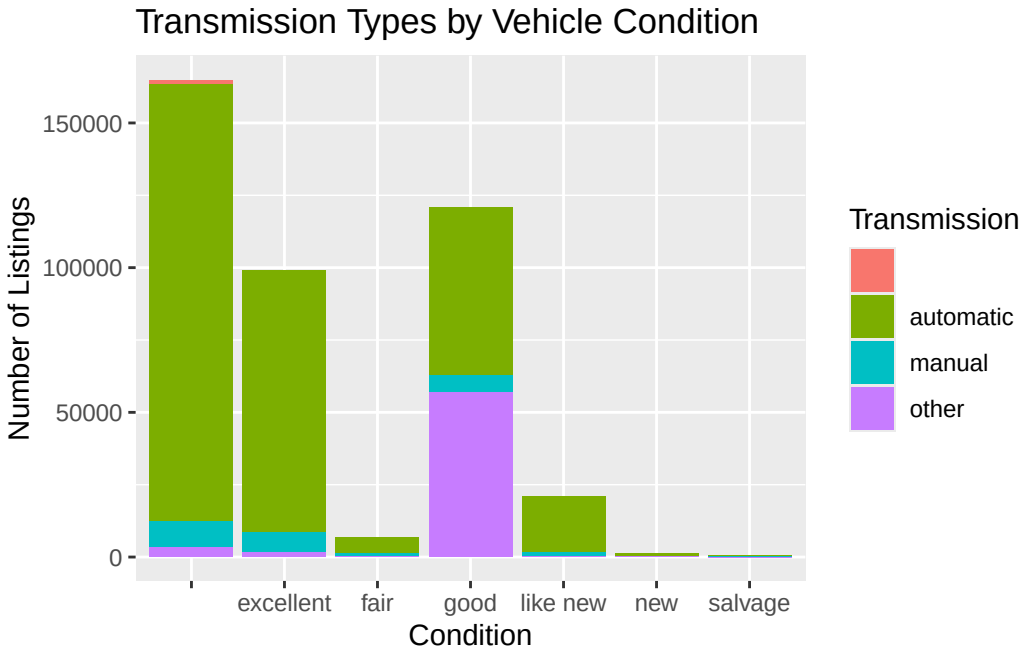
```
ggplot(data, aes(x = transmission, y = year)) +
  geom_violin(fill = "skyblue") +
  labs(title = "Distribution of Transmission Types", x = "Transmission Type", y = "Year of
```



Violin plots highlight the distribution of transmission types, revealing consumer preferences in driving and performance. This information can be used to guide the design of new models and offer transmission options based on regional preferences and market characteristics.

Stacked bar chart of transmission types by vehicle condition

```
ggplot(data, aes(x = condition, fill = transmission)) +
  geom_bar() +
  labs(title = "Transmission Types by Vehicle Condition", x = "Condition", y = "Number of
```



Conclusion

In summary, the analysis of Carvana's dataset provides actionable insights for optimizing sales, marketing, and operational strategies in the automotive market, facilitating informed decision-making and driving business growth.