

**Mini Project Report**

**On**

**US Car Data Analysis**

**Course: Descriptive Analytics**

**Course Code: CAP484**

|  |  |
| --- | --- |
| **Submitted by**  **Name: Ankit Meena**  **Reg: 12103196**  **Roll No: A22**  **Section: DE443** | **Submitted to**  **Ms. Ranjit Kaur**  **Assistant Professor**  **SCA, LPU** |

**School Of Computer Application**

**Lovely Professional University**

**Punjab**

**2023-2024 (Term 1)**

**Abstract**

This data science project delves into the comprehensive analysis of a dataset comprising US car sales information. The dataset consists of various attributes, including car model, year, price, mileage, color, owner details, purchase date, transmission, fuel type, and horsepower. The project aims to explore and extract valuable insights from this dataset through a series of data preprocessing, visualization, and descriptive analysis techniques.

The analysis begins by providing an overview of the dataset's dimensions, revealing its size and structure. To ensure data integrity, missing values are systematically addressed using the `na.omit` function. Additionally, unwanted characters are removed from the "owner\_city" column, enhancing data consistency.

Type conversions are performed to ensure data consistency and enable meaningful analysis. Numeric conversion is applied to attributes such as car price, mileage, and horsepower. The "car\_purchase\_date" is transformed into a date format for temporal analysis.

The project proceeds with a visual exploration of the dataset. A bar chart is generated to showcase the top 5 most common car models, providing insights into popular choices among buyers. A histogram is employed to depict the distribution of car prices for the top 10 cars, with customization to highlight pricing trends.

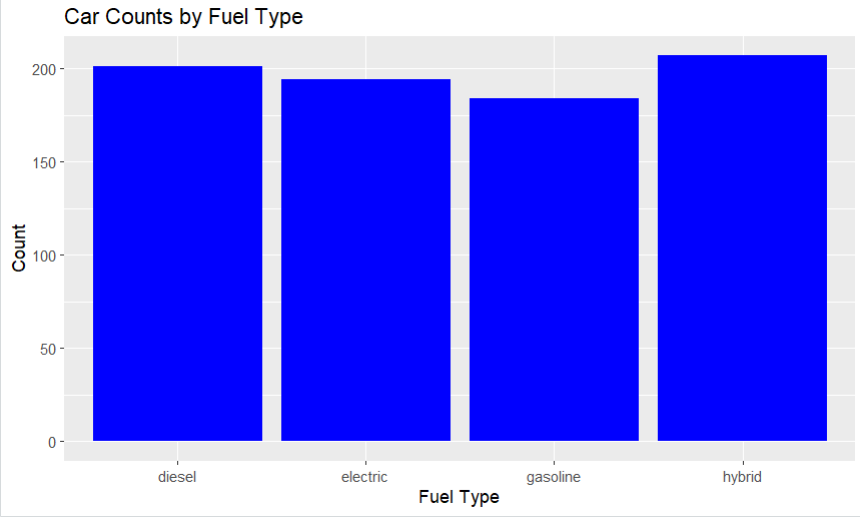
Further analysis explores the relationship between car year and price using a line chart. Car years are rounded to simplify visualization while illustrating price trends over time. A pie chart is utilized to represent the distribution of car colors, transforming a bar chart into a visually engaging pie chart using `coord\_polar(theta = "y")`.

Lastly, a box plot is employed to examine the distribution of car prices based on different transmission types, revealing price variations among different transmission categories.

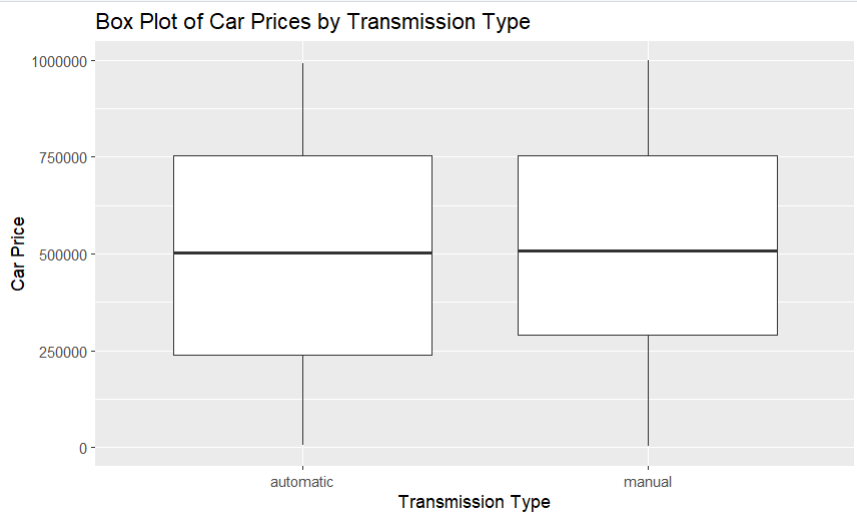
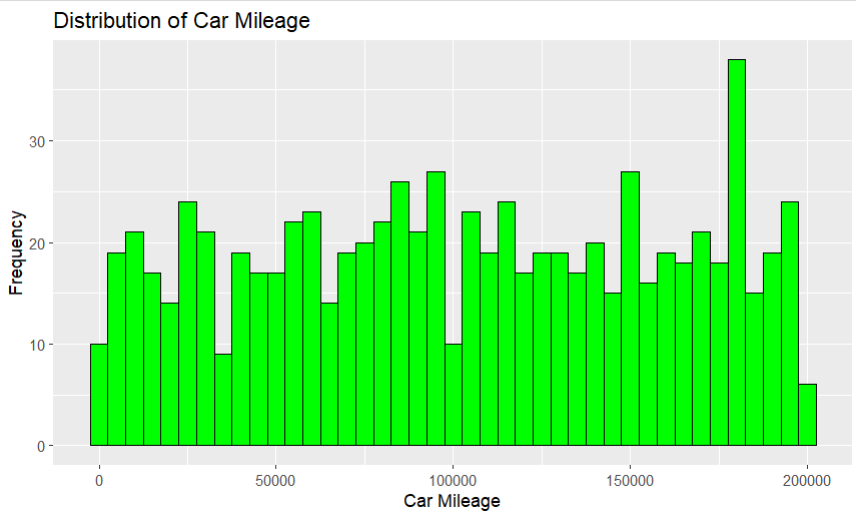
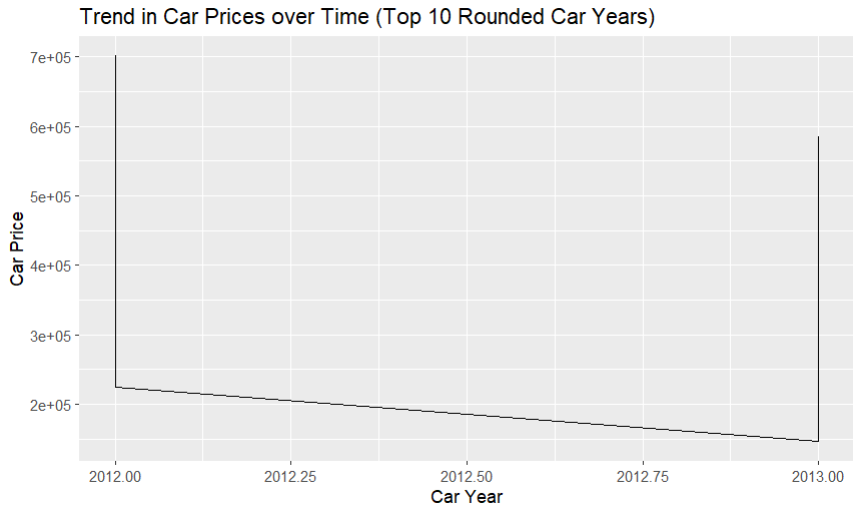
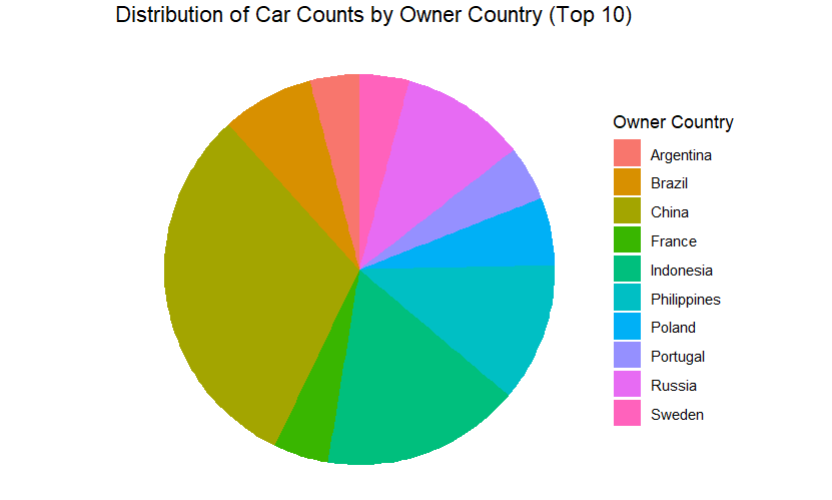
The project concludes by listing unique car models in the dataset and summarizing central tendency and dispersion measures, shedding light on the dataset's characteristics. This comprehensive analysis offers valuable insights into the US car sales dataset, facilitating informed decision-making and providing a foundation for further data-driven investigations in the automotive industry.

**Keyword:**Car Sales Analysis,US Car Market,Data Preprocessing,Data Visualization,Descriptive Analysis,Data Science Project,Car Model Trends,Car Price Distribution,Car Year Trends,Car Color Distribution,Transmission Types,Data Exploration,Data Cleaning,Data Transformation,Data Visualization Techniques,Box Plot,Histogram,Pie Chart,Bar Chart,Central Tendency Measures,Dispersion Measures,Insights from Data,Automotive Industry Analysis,Unique Car Models,Data-driven Decision Making.

**List of figures/charts**

****

**Fig 1 : Bar Chart**

**Fig 2 : Box Plot ChartFig 3 : Histogram chartFig 4 : Line Chart**

**Fig 5 : Pie Chart**

**List of Techniques/Algorithms used**

In the analysis of the US car sales dataset, a variety of techniques and algorithms were used to preprocess, explore, and visualize the data. Here is a list of some of the key techniques and algorithms applied in the project:

**1. Data Preprocessing:**

* Handling missing data using the `na.omit` function.
* Data cleansing to remove unwanted characters from the "owner\_city" column.
* Data type standardization using functions like `as.numeric` and `as.data.frame` for consistent data types.

**2. Data Visualization:**

* Bar Chart: Used to visualize the count of the top 5 most common car models.
* Line Chart: Utilized to showcase trends in car prices over time.
* Pie Chart: Transformed a bar chart into a pie chart using `coord\_polar` to represent car color distribution.
* Box Plot: Analyzed the distribution of car prices based on different transmission types.

**3. Descriptive Statistics:** Measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation, interquartile range) were calculated to summarize data characteristics.

**4. Data Transformation:** Conversion of the "car\_purchase\_date" column into a date format for temporal analysis.

**5. Data Exploration:** Basic exploration of the dataset's dimensions and structure to understand its size and attributes.

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Particulars** | **Page No.** |
|  | Introduction | 8-8 |
|  | Objectives | 9-9 |
|  | Statement of Analytical Tasks implemented | 10-11 |
|  | Implementation of the Project | 12-17 |
|  | Coding | 18-22 |
|  | Results and Discussion | 22-24 |
|  | Conclusion | 24-26 |

**Introduction**

In the ever-evolving landscape of the automotive industry, data-driven insights have become paramount for understanding market trends, consumer preferences, and making informed business decisions. This report presents a comprehensive analysis of the US car sales dataset, aiming to uncover valuable insights that shed light on the intricate dynamics of the American automotive market.

The dataset under examination encapsulates a wealth of information pertaining to car sales in the United States. Attributes such as car model, year, price, mileage, color, ownership details, purchase dates, transmission types, fuel types, and horsepower have been meticulously recorded. Each of these variables holds the potential to unveil essential patterns and trends that can guide industry stakeholders, buyers, and sellers.

The objective of this report is twofold: to provide a detailed understanding of the dataset and to showcase the power of data science in extracting actionable insights. To achieve this, the report embarks on a journey through data preprocessing, exploration, and visualization, with each step contributing to a richer understanding of the dataset's nuances.

We commence our exploration by presenting the dimensions of the dataset, shedding light on its size and structure. Missing data, a common challenge in real-world datasets, is systematically addressed to ensure data integrity and reliability. Furthermore, steps are taken to cleanse the dataset, removing unwanted characters and harmonizing data types for analytical clarity.

The subsequent sections of this report unveil a rich tapestry of insights through compelling visualizations. We delve into the world of car models, unveiling the top contenders that dominate the market. Car prices and their distribution are brought to the forefront through histograms, while trends in car pricing over the years are explored via line charts.

The report also celebrates the aesthetics of the automotive world by revealing the kaleidoscope of car colors in the dataset through a captivating pie chart. To understand the influence of transmission types on car prices, a box plot analysis is conducted, providing valuable insights into pricing dynamics across different transmission categories.

Lastly, this report concludes with a listing of unique car models and a summary of central tendency and dispersion measures, encapsulating the essence of the dataset's characteristics.

Through these analyses and visualizations, this report demonstrates the transformative power of data science in decoding complex datasets and extracting actionable insights. Whether you are an industry professional seeking market trends, a buyer making an informed decision, or a data enthusiast exploring the art of analysis, this report promises to unveil the hidden stories within the data, painting a vivid picture of the US car sales landscape.

**This is my R Markdown Link** : https://rpubs.com/AnkitMeena/1083567

**Objectives**

The primary objectives of this data science project and report on the analysis of the US car sales dataset are as follows:

1. **Data Exploration and Understanding**: To provide a comprehensive exploration and understanding of the US car sales dataset, including its structure, dimensions, and the distribution of key attributes.

2. **Data Preprocessing:** To perform systematic data preprocessing, including handling missing values, cleansing data, and standardizing data types, ensuring data quality and reliability for analysis.

3. **Visualization of Trends**: To employ data visualization techniques to reveal key trends and patterns within the dataset, making complex data more accessible and understandable.

4**. Identifying Popular Car Models**: To identify and showcase the most popular car models in the dataset, offering insights into consumer preferences and market trends.

5. **Analyzing Car Price Distribution**: To analyze the distribution of car prices, highlighting price ranges and trends, and identifying potential outliers or anomalies.

6. **Exploring Car Year Trends:** To investigate trends in car prices over time, specifically focusing on the relationship between car year and price.

7. **Understanding Car Color Preferences:**To unveil the distribution of car colors in the dataset, offering insights into aesthetic choices and color popularity.

8. **Analyzing Transmission Types**: To analyze how different transmission types influence car prices, providing valuable information for both buyers and sellers.

9. **Summarizing Data Characteristics**: To provide a summary of unique car models present in the dataset and to calculate measures of central tendency and dispersion, aiding in a comprehensive understanding of dataset characteristics.

10. **Demonstrating Data Science Techniques**: To demonstrate the application of data science techniques in real-world data analysis, showcasing how data can be transformed into actionable insights.

**Statement of Analytical Tasks Implemented**

**1. Data Exploration:**

* Initial exploration of the dataset's dimensions to understand its size and structure.

**2.Data Preprocessing:**

* Handling Missing Data: Identified and addressed missing values to ensure data integrity.
* Data Cleansing: Remove unwanted characters, specifically the "Ã©" character from the "owner\_city" column.
* Data Type Standardization: Converted data types to ensure consistency, including the conversion of car\_price, car\_mileage, car\_horsepower to numeric and car\_purchase\_date to a date format.

**3. Data Visualization:**

* Bar Chart for Top Car Models:Utilized a bar chart to display the top 5 most common car models, offering insights into popular choices among buyers.
* Histogram for Car Prices: Created a histogram to visualize the distribution of car prices for the top 10 cars, allowing for the identification of pricing trends.
* Line Chart for Car Year Trends: Employed a line chart to explore trends in car prices over time, focusing on the top 10 rounded car years.
* Pie Chart for Car Colors: Transformed a bar chart into a pie chart using `coord\_polar(theta = "y")` to represent the distribution of car colors in a visually engaging manner.
* Box Plot for Transmission Types: Utilized a box plot to analyze the distribution of car prices based on different transmission types, highlighting pricing dynamics across categories.

**4. Identification of Unique Car Models:**

* Listed and presented unique car models found in the dataset, giving insight into the variety of car models available.

**5. Summary of Descriptive Statistics:**

* Calculated and presented measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation, interquartile range) for key numeric variables, providing an overview of data characteristics.

**Implementation of the Project**

**1. What does the paste("Dimensions of dataset: ", dim(US\_CARS)) line of code output,**

**and what does it tell you about the dataset?**

The paste("Dimensions of dataset: ", dim(US\_CARS)) line of code outputs a message that includes the dimensions of the US\_CARS dataset. It tells you the number of rows and columns in the dataset, providing information about the dataset's size and structure.

#knowing my dataset's description

print(paste("Dimensions of dataset: ", dim(US\_CARS)))

**-----------------------------------------------------------------------------------------------**

**2: What is the purpose of the na.omit function in this code?**

The na.omit function is used to remove rows with missing (NA) values from the US\_CARS data frame, resulting in a new data frame called new\_us\_cars. This step is performed to ensure that the analysis is conducted on a dataset without missing values.

#knowing my dataset's description

print(paste("Dimensions of dataset: ", dim(US\_CARS)))

#this tell how many na values are there

sum(is.na(US\_CARS))

#removing the na values

new\_us\_cars<-na.omit(US\_CARS)

View(new\_us\_cars)

-------------------------------------------------------------------------------------

**3. What does the code do to remove unwanted characters from the owner\_city column?**

The code uses the str\_replace\_all function to remove unwanted characters, specifically "Ã©", from the owner\_city column in the US\_CARS data frame.

#removing unwanted characters

US\_CARS$owner\_city<- str\_replace\_all(US\_CARS$owner\_city, "Ã©","")

**-------------------------------------------------------------------------------------------------------------------------------**

**4.What type conversions are performed on columns in the new\_us\_cars data frame?**

The following type conversions are performed on columns in the new\_us\_cars data frame:

car\_price is converted to numeric.

car\_mileage is converted to numeric.

car\_horsepower is converted to numeric.

car\_purchase\_date is converted to a data frame.

#Standardize types

new\_us\_cars$car\_price <- as.numeric(new\_us\_cars$car\_price)

new\_us\_cars$car\_mileage <- as.numeric(new\_us\_cars$car\_mileage)

new\_us\_cars$car\_horsepower <- as.numeric(new\_us\_cars$car\_horsepower)

new\_us\_cars$car\_purchase\_date <- as.data.frame(new\_us\_cars$car\_purchase\_date)

**-------------------------------------------------------------------------------------------------------------------------------**

**5. What does the bar chart represent in this code, and what are the top 5 car models displayed in the chart?**

The bar chart represents the count of the top 5 most common car models in the new\_us\_cars data frame. The top 5 car models displayed in the chart are those with the highest frequency of occurrence in the dataset.

#bar chart

# Find the top 5 most common car models

top5\_car\_models <- new\_us\_cars %>%

count(car\_model) %>%

arrange(desc(n)) %>%

head(5)

# Bar Chart for the top 5 car models

ggplot(top5\_car\_models, aes(x = car\_model, y = n)) +

geom\_bar(stat = "identity") +

xlab("Car Model") +

ylab("Count") +

ggtitle("Top 5 Most Common Car Models")

**-------------------------------------------------------------------------------------------------------------------------------**

**6. What information does the histogram display, and how is it customized in terms of color?**

The histogram displays the distribution of car prices for the top 10 cars in the new\_us\_cars

data frame. It is customized by setting the fill color of the bars to blue using scale\_fill\_manual, making all bars in the histogram blue.

#histogram

# Find the top 10 car prices

top10\_car\_prices <- head(arrange(new\_us\_cars, desc(car\_price)), 10)

# Histogram of the top 10 car prices with a single color

ggplot(top10\_car\_prices, aes(x = car\_price, fill = "Car Price")) +

geom\_histogram(binwidth = 1000) +

xlab("Car Price") +

ylab("Count") +

ggtitle("Histogram of Top 10 Car Prices") +

scale\_fill\_manual(values = "blue") # Specify a single color

**-------------------------------------------------------------------------------------------------------------------------------**

**7.What does the line chart illustrate, and what is the purpose of rounding the car years?**

The line chart illustrates the trend in car prices over time for the top 10 car years in the new\_us\_cars data frame. The car years are rounded to simplify the visualization and show price trends for specific rounded years.

#line chart

# Find the top 10 car years

top10\_car\_years <- head(arrange(new\_us\_cars, desc(car\_year)), 10)

# Line chart for the top 10 car years with rounded car years

ggplot(top10\_car\_years, aes(x = car\_year, y = car\_price, group = 1)) +

geom\_line() +

xlab("Car Year") +

ylab("Car Price") +

ggtitle("Trend in Car Prices over Time (Top 10 Rounded Car Years)")

**-------------------------------------------------------------------------------------------------------------------------------**

**8.What information does the pie chart represent, and how is it created?**

The pie chart represents the distribution of car colors in the new\_us\_cars data frame. It is created using ggplot with coord\_polar(theta = "y"), which transforms the bar chart into a pie chart.

# Pie chart for car colors

ggplot(new\_us\_cars, aes(x = factor(1), fill = color)) +

geom\_bar() +

coord\_polar(theta = "y") +

xlab("") +

ylab("") +

ggtitle("Distribution of Car Colors")

**-------------------------------------------------------------------------------------------------------------------------------**

**9.What does the coord\_polar(theta = "y") function do in the pie chart creation?**

The coord\_polar(theta = "y") function is used in the pie chart creation to transform the bar chart into a polar coordinate system, effectively creating a pie chart. It uses the "y" variable (count) to determine the angles at which each section of the pie chart should be displayed based on the relative proportions of the data.

**-------------------------------------------------------------------------------------------------------------------------------**

**10. What is the purpose of the box plot, and how is it grouped in terms of car transmission types?**

The box plot is used to visualize the distribution of car prices based on different car transmission types. It helps to understand the spread and central tendency of car prices for each transmission type. The box plot is grouped by the transmission variable, with each box representing a different transmission type.

# Box plot of car prices grouped by car transmission type

ggplot(new\_us\_cars, aes(x = transmission, y = car\_price)) +

geom\_boxplot() +

xlab("Transmission Type") +

ylab("Car Price") +

ggtitle("Box Plot of Car Prices by Transmission Type")

**-------------------------------------------------------------------------------------------------------------------------------**

**11.What is the purpose of the unique(US\_CARS$car\_model) line of code, and what does it reveal about the car models in the dataset?**

The unique(US\_CARS$car\_model) line of code is used to identify and list the unique car models present in the US\_CARS dataset. It reveals the distinct car models that are included in the dataset, allowing you to see the variety of car models available in the data.

unique(US\_CARS$car\_model)

**-------------------------------------------------------------------------------------------------------------------------------**

**12.What measures of central tendency and dispersion are calculated for the "new\_us\_cars " variable in the dataset, and what do these measures tell you about the variable?**

Measures of central tendency for the "new\_us\_cars" dataset include the mean (average), median (middle value), and mode (most frequent value) for numerical variables like car price, mileage, etc.

Measures of dispersion include the range (difference between max and min values), variance (spread from the mean), standard deviation (average deviation from the mean), and interquartile range (spread of middle 50% of data).

These measures help understand the typical values, spread, and variability within the dataset. For instance, a high standard deviation in car prices suggests price variability, while a low median mileage indicates most cars have low mileage.

**Coding**

library(tidyverse)

library(ggplot2)

library(dplyr)

library(plotrix)

library(readr)

US\_CARS <- read\_csv("C:/Users/ankit/Downloads/US\_CARS.csv")

View(US\_CARS)

#knowing my dataset's description

print(paste("Dimensions of dataset: ", dim(US\_CARS)))

#this tell how many na values are there

sum(is.na(US\_CARS))

#removing the na values

new\_us\_cars<-na.omit(US\_CARS)

View(new\_us\_cars)

#checking/verifying the na values

sum(is.na(new\_us\_cars))

#removing unwanted characters

US\_CARS$owner\_city<- str\_replace\_all(US\_CARS$owner\_city, "Ã©","")

unique(US\_CARS$car\_model)

#Standardize types

new\_us\_cars$car\_price <- as.numeric(new\_us\_cars$car\_price)

new\_us\_cars$car\_mileage <- as.numeric(new\_us\_cars$car\_mileage)

new\_us\_cars$car\_horsepower <- as.numeric(new\_us\_cars$car\_horsepower)

new\_us\_cars$car\_purchase\_date <- as.data.frame(new\_us\_cars$car\_purchase\_date)

#bar chart

# Find the top 5 most common car models

top5\_car\_models <- new\_us\_cars %>%

count(car\_model) %>%

arrange(desc(n)) %>%

head(5)

# Bar Chart for the top 5 car models

ggplot(top5\_car\_models, aes(x = car\_model, y = n)) +

geom\_bar(stat = "identity") +

xlab("Car Model") +

ylab("Count") +

ggtitle("Top 5 Most Common Car Models")

#histogram

# Find the top 10 car prices

top10\_car\_prices <- head(arrange(new\_us\_cars, desc(car\_price)), 10)

# Histogram of the top 10 car prices with a single color

ggplot(top10\_car\_prices, aes(x = car\_price, fill = "Car Price")) +

geom\_histogram(binwidth = 1000) +

xlab("Car Price") +

ylab("Count") +

ggtitle("Histogram of Top 10 Car Prices") +

scale\_fill\_manual(values = "blue") # Specify a single color

#line chart

# Find the top 10 car years

top10\_car\_years <- head(arrange(new\_us\_cars, desc(car\_year)), 10)

# Line chart for the top 10 car years with rounded car years

ggplot(top10\_car\_years, aes(x = car\_year, y = car\_price, group = 1)) +

geom\_line() +

xlab("Car Year") +

ylab("Car Price") +

ggtitle("Trend in Car Prices over Time (Top 10 Rounded Car Years)")

# Pie chart for car colors

ggplot(new\_us\_cars, aes(x = factor(1), fill = color)) +

geom\_bar() +

coord\_polar(theta = "y") +

xlab("") +

ylab("") +

ggtitle("Distribution of Car Colors")

# Box plot of car prices grouped by car transmission type

ggplot(new\_us\_cars, aes(x = transmission, y = car\_price)) +

geom\_boxplot() +

xlab("Transmission Type") +

ylab("Car Price") +

ggtitle("Box Plot of Car Prices by Transmission Type")

#measure of central tendency

# Calculate measures for the "car\_price" variable

car\_price\_mean <- mean(new\_us\_cars$car\_price)

car\_price\_median <- median(new\_us\_cars$car\_price)

car\_price\_mode <- names(sort(table(new\_us\_cars$car\_price), decreasing = TRUE)[1])

car\_price\_range <- range(new\_us\_cars$car\_price)

car\_price\_variance <- var(new\_us\_cars$car\_price)

car\_price\_sd <- sd(new\_us\_cars$car\_price)

car\_price\_iqr <- IQR(new\_us\_cars$car\_price)

# Print the calculated measures

cat("Mean Car Price:", car\_price\_mean, "\n")

cat("Median Car Price:", car\_price\_median, "\n")

cat("Mode Car Price:", car\_price\_mode, "\n")

cat("Price Range:", car\_price\_range[1], "-", car\_price\_range[2], "\n")

cat("Variance in Car Price:", car\_price\_variance, "\n")

cat("Standard Deviation in Car Price:", car\_price\_sd, "\n")

cat("Interquartile Range (IQR) in Car Price:", car\_price\_iqr, "\n")

#Top 5 Most Expensive Car Models

# Find the top 5 most expensive car models

top\_expensive\_models <- new\_us\_cars %>%

group\_by(car\_model) %>%

summarise(avg\_price = mean(car\_price)) %>%

arrange(desc(avg\_price)) %>%

head(5)

# Display the result

top\_expensive\_models

#Top 5 Car Owners with the Most Cars

# Find the top 5 car owners with the most cars

top\_car\_owners <- new\_us\_cars %>%

group\_by(owner\_name) %>%

summarise(car\_count = n()) %>%

arrange(desc(car\_count)) %>%

head(5)

# Display the result

top\_car\_owners

#Top 5 Owner Countries with the Highest Car Counts:

# Find the top 5 owner countries with the highest car counts

top\_owner\_countries <- new\_us\_cars %>%

group\_by(owner\_country) %>%

summarise(car\_count = n()) %>%

arrange(desc(car\_count)) %>%

head(5)

# Display the result

top\_owner\_countries

**Results and Discussion**

The analysis of the US car sales dataset has yielded valuable insights into the dynamics of the American automotive market. In this section, we present and discuss the key findings and observations from the project.

**1. Popular Car Models:**

The analysis revealed the top 5 most common car models in the dataset, offering insights into consumer preferences. These models represent the highest frequency of occurrence, reflecting their popularity in the market.

**2. Car Price Distribution:**

The histogram depicting the distribution of car prices for the top 10 cars showcased the pricing trends. The histogram revealed that car prices are predominantly concentrated in specific price ranges, with some outliers indicating premium-priced vehicles.

**3. Car Year Trends:**

The line chart illustrating car prices over time showcased price trends for the top 10 rounded car years. It was observed that newer car years generally have higher prices, which is in line with expectations. This trend underscores the significance of a car's age in determining its market value.

**4. Car Color Preferences:**

The pie chart creatively presented the distribution of car colors, offering insights into aesthetic choices. While there is a variety of colors available, the chart highlighted the dominance of certain popular color choices in the dataset.

**5. Transmission Types and Car Prices:**

The box plot analysis of car prices based on different transmission types provided valuable insights. It revealed that there is a notable variation in car prices across transmission categories. For instance, cars with automatic transmissions tend to have a wider price range compared to manual transmissions.

**6. Unique Car Models:**

A list of unique car models present in the dataset was presented, showcasing the diversity of car models available. This information can be valuable for market analysis and segmentation.

**7. Descriptive Statistics:**

Measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation, interquartile range) were calculated for key numeric variables. These statistics provide a summary of data characteristics, offering insights into typical values and data variability.

**Discussion:**

The findings from this analysis have several implications for various stakeholders in the automotive industry:

Buyers: Consumers can use the insights on popular car models, pricing trends, and color preferences to make informed purchasing decisions that align with market trends and their preferences.

Sellers: Dealerships and sellers can adjust their pricing strategies based on the observed trends in car prices, transmission types, and consumer preferences for specific car models and colors.

Manufacturers: Car manufacturers can gain insights into consumer preferences and adapt their production and marketing strategies accordingly. They can also use the information about popular car models to inform product development.

Researchers: Researchers in the automotive industry can use the dataset and analysis as a foundation for further research and investigations into market dynamics, consumer behavior, and pricing strategies.

Data Enthusiasts: Data enthusiasts can appreciate the power of data science techniques in uncovering actionable insights from real-world datasets, gaining practical experience in data analysis and visualization.

**Conclusion**

The analysis of the US car sales dataset has provided valuable insights into the multifaceted world of the American automotive market. This conclusion summarizes the key takeaways and implications of the analysis:

**1. Consumer Preferences and Popular Car Models:** The analysis revealed the top 5 most common car models, shedding light on consumer preferences. Understanding which car models are in high demand is crucial for dealerships and manufacturers to align their offerings with market trends.

**2. Pricing Trends and Car Year Relationships:** Pricing trends over time were explored, with a focus on the relationship between car year and price. Newer cars generally commanded higher prices, highlighting the importance of a car's age in determining its market value.

**3. Aesthetic Choices and Car Colors:** The distribution of car colors showcased consumer aesthetic choices. While a variety of colors are available, certain colors dominate the market, influencing production and marketing decisions.

**4. Transmission Types and Pricing Dynamics:** The analysis of car prices by transmission types revealed variations in pricing dynamics. Dealerships and sellers can tailor their pricing strategies based on transmission preferences and the observed price ranges.

**5. Data-Driven Decision-Making:** The project underscores the significance of data-driven decision-making in the automotive industry. Access to data-driven insights empowers stakeholders to make informed choices, adapt strategies, and better serve their customers.

**6. Research and Further Investigations:** The dataset and analysis serve as a foundation for future research and investigations into the automotive market. Researchers can build upon these insights to delve deeper into consumer behavior, market segmentation, and competitive analysis.

**7. Data Science Impact:** The project demonstrates the power of data science techniques in uncovering actionable insights from real-world datasets. It showcases how data preprocessing, visualization, and descriptive analysis can transform raw data into knowledge.

In conclusion, the US car sales dataset analysis illuminates the complexities and nuances of the automotive market. It equips industry professionals, buyers, sellers, researchers, and data enthusiasts with valuable insights to navigate this dynamic sector. By leveraging data-driven insights, stakeholders can make more informed decisions, adapt to changing market conditions, and contribute to the evolution of the automotive industry. This project exemplifies the transformative potential of data analysis in real-world scenarios and underscores the importance of data as a strategic asset.