Assignment_1

shrikunj

2025-03-23

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

R Markdown

```
# Load dataset
cars_dataset <- read.csv("E:/Project/Masters/BDA/Assignment1/cars_data_10K.csv")

# Check for missing values and remove them
cars_cleandata <- na.omit(cars_dataset)

# Install necessary packages
if (!require(outliers)) install.packages("outliers", dependencies = TRUE)
```

```
## Loading required package: outliers
```

```
library(outliers)

if (!require(tinytex)) install.packages("tinytex", dependencies = TRUE)
```

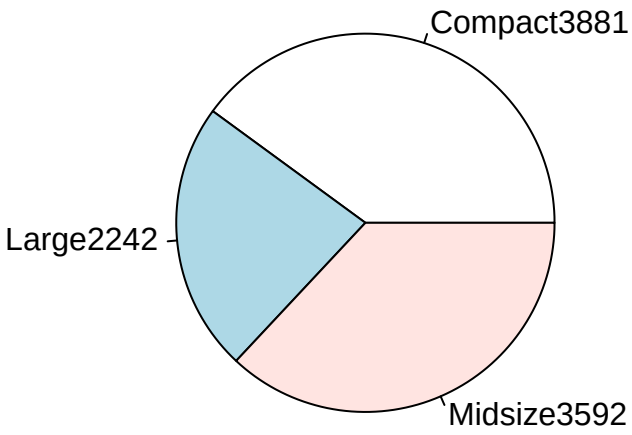
```
## Loading required package: tinytex
```

```
library(tinytex)

# Remove outliers in MSRP
outlier_scores <- scores(cars_cleandata$MSRP, type = "z")
is_outlier <- outlier_scores > 3 | outlier_scores < -3
cars_cleandata <- cars_cleandata[!is_outlier, ]

# Convert MSRP to numeric and Vehicle.Size to factor
cars_cleandata$MSRP <- as.numeric(cars_cleandata$MSRP)
cars_cleandata$Vehicle.Size <- as.factor(cars_cleandata$Vehicle.Size)

# Display summary of cleaned data
summary(cars_cleandata)
```

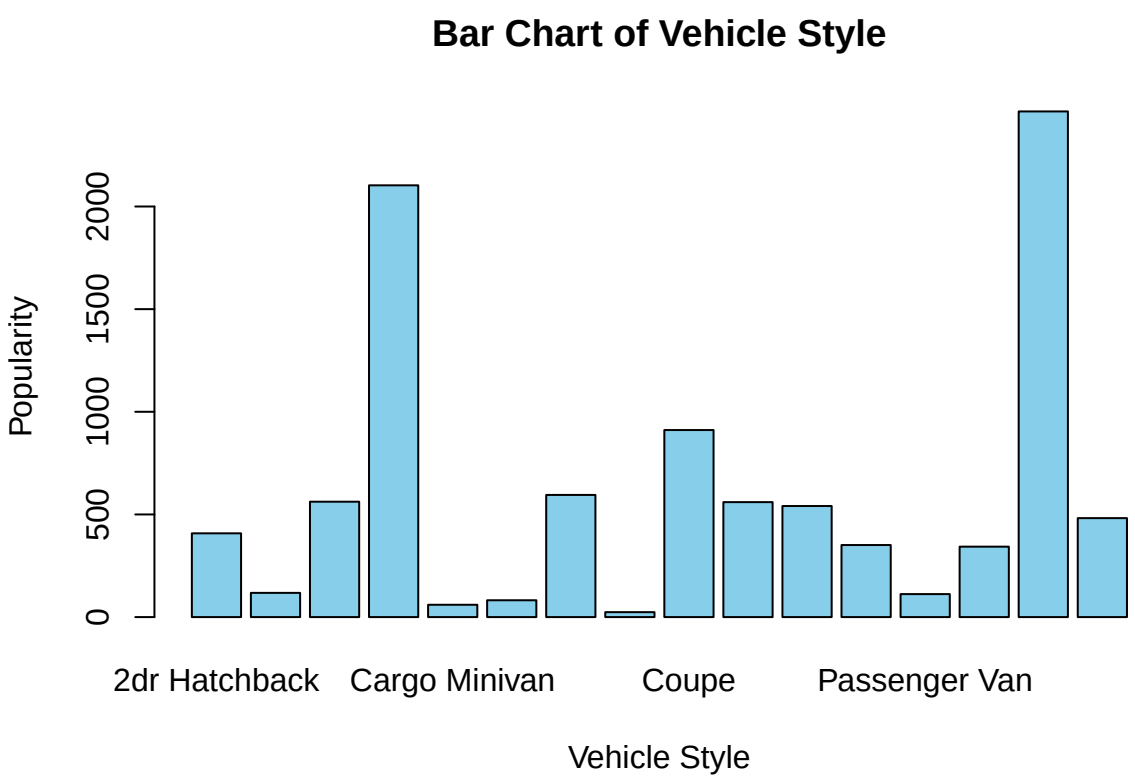
```
 Make               Model                Year       Engine.Fuel.Type
```

```
Length:9715 Length:9715 Min. :1990 Length:9715
Class :character Class :character 1st Qu.:2006 Class :character
Mode :character Mode :character Median :2015 Mode :character
Mean :2010
3rd Qu.:2016
Max. :2017
Engine.HP Engine.Cylinders Transmission.Type Driven_Wheels
Min. : 55.0 Min. : 0.000 Length:9715 Length:9715
1st Qu.:170.0 1st Qu.: 4.000 Class :character Class :character
Median :220.0 Median : 6.000 Mode :character Mode :character
Mean :242.7 Mean : 5.546
3rd Qu.:296.0 3rd Qu.: 6.000
Max. :707.0 Max. :12.000
Number.of.Doors Market.Category Vehicle.Size Vehicle.Style
Min. :2.000 Length:9715 Compact:3881 Length:9715
1st Qu.:2.000 Class :character Large :2242 Class :character
Median :4.000 Mode :character Midsize:3592 Mode :character
Mean :3.449
3rd Qu.:4.000
Max. :4.000
highway.MPG city.mpg Popularity MSRP
Min. : 12.00 Min. : 8.00 Min. : 2 Min. : 2000
1st Qu.: 22.00 1st Qu.: 16.00 1st Qu.: 549 1st Qu.: 20555
Median : 26.00 Median : 18.00 Median :1385 Median : 29510
Mean : 26.45 Mean : 19.45 Mean :1569 Mean : 34372
3rd Qu.: 30.00 3rd Qu.: 22.00 3rd Qu.:2009 3rd Qu.: 41150
Max. :354.00 Max. :137.00 Max. :5657 Max. :211000
```

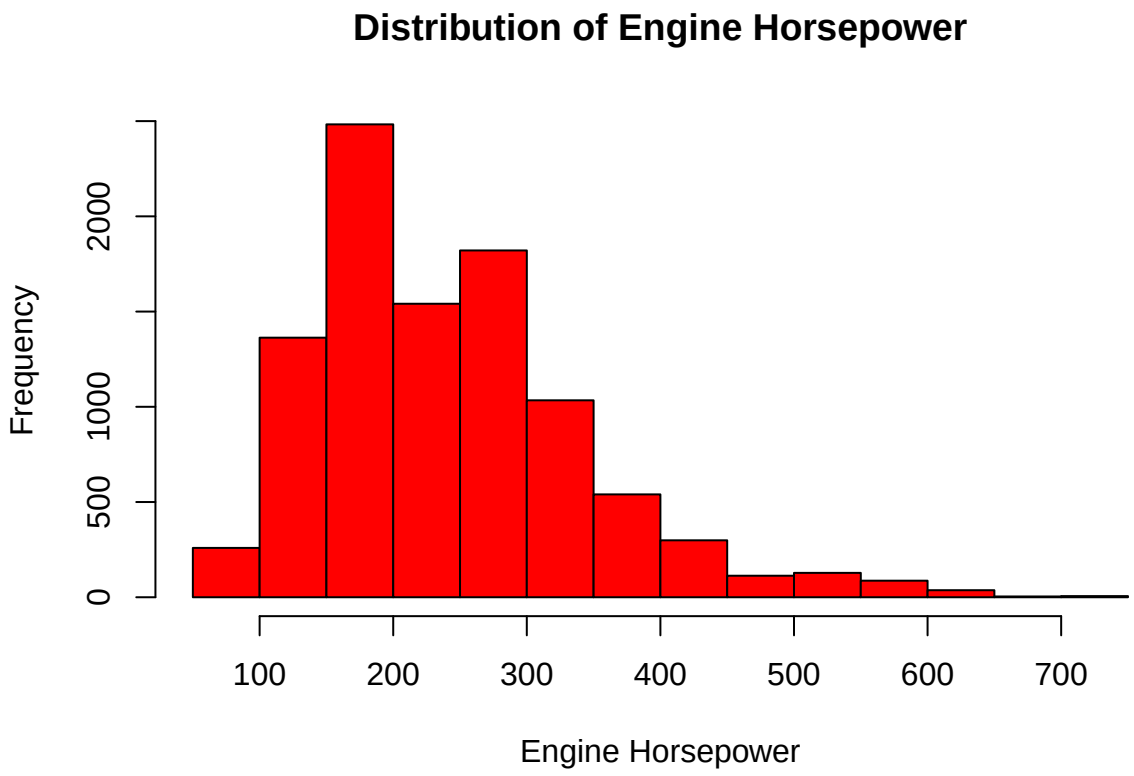
```r
## **Summary of Variables and Visualizations: Analyzing Car Dataset with Statistical Methods and Plots**

# Pie chart of vehicle sizes
 slices <- table(cars_cleandata$Vehicle.Size)
 lbls <- paste(names(slices), "", slices, sep = "")
 pie(slices, labels = lbls, main = "Pie Chart of Vehicle Sizes")
```



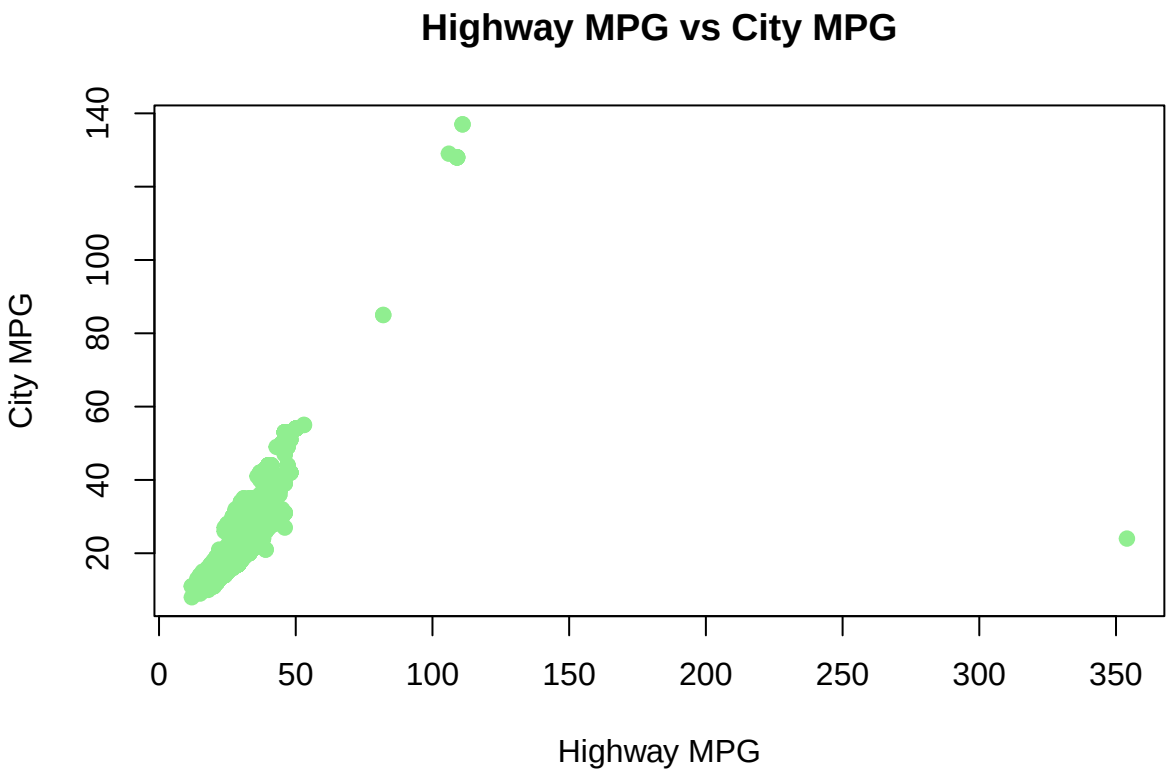

```r
# Bar chart of vehicle styles
 popularity_by_style <- table(cars_cleandata$Vehicle.Style)
 barplot(popularity_by_style, main = "Bar Chart of Vehicle Style",
 xlab = "Vehicle Style", ylab = "Popularity", col= "skyblue")
```



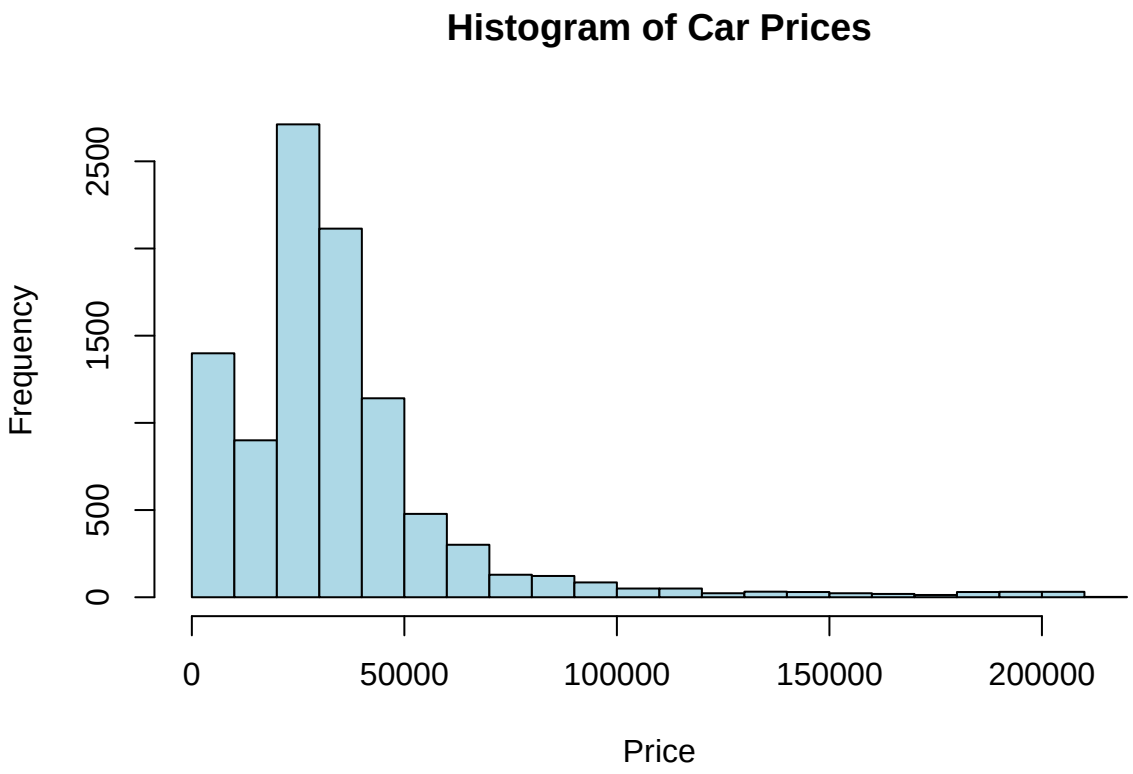

```
# Histogram of Engine HP
 hist(cars_cleandata$Engine.HP, main= "Distribution of Engine Horsepower",xlab= "Engine Horsepower", yla
```




```
# Scatter plot of Highway MPG vs City MPG
plot(cars_cleandata$highway.MPG, cars_cleandata$city.mpg,
     main = "Highway MPG vs City MPG", xlab = "Highway MPG", ylab = "City MPG",
     pch = 19, col = "lightgreen")
```




```
## **4(a) Histogram of Car Prices**
## The histogram provides insights into the distribution of car prices.
## If the histogram skews right, most cars are budget-friendly with a few luxury models.
## If the histogram skews left, it suggests that a majority of cars are expensive, with fewer affordabl
hist(cars_cleandata$MSRP, main = "Histogram of Car Prices",
     xlab = "Price", col = "lightblue", border = "black")
```




```
## **4(b) Summary Statistics and Price Grouping**
## Summary statistics help understand central tendencies and variations in car prices.
summary(cars_cleandata$MSRP)
```

Min. 1st Qu. Median Mean 3rd Qu. Max. 2000 20555 29510 34372 41150 211000

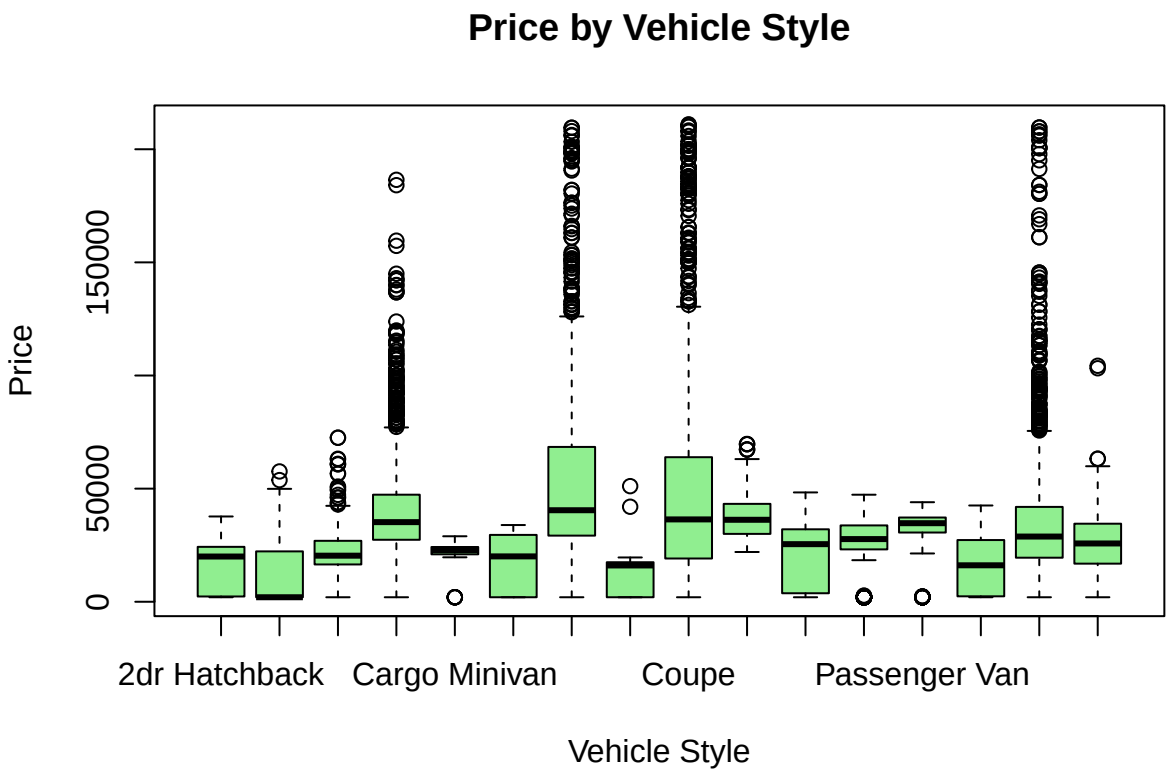
```
var(cars_cleandata$MSRP)
```

[1] 845775217

```
# Grouping prices into categories
price_groups <- cut(cars_cleandata$MSRP, breaks = c(0, 10000, 20000, Inf),
                    labels = c("Low", "Medium", "High"))
cars_cleandata$price_group <- price_groups
aggregate(MSRP ~ price_group, data = cars_cleandata, FUN = summary)
```

price_group MSRP.Min. MSRP.1st Qu. MSRP.Median MSRP.Mean MSRP.3rd Qu. 1 Low 2000.000 2000.000 2000.000 2369.651 2356.000 2 Medium 10135.000 15499.000 17192.500 16994.788 18810.000 3 High 20015.000 26875.000 34062.500 42518.498 45492.500 MSRP.Max. 1 9949.000 2 20000.000 3 211000.000

```
## **4(c) Boxplot of MSRP by Vehicle Style**
## The boxplot visualizes price variations across different vehicle styles.
## It helps identify trends in pricing, showing whether luxury and performance cars have a higher MSRP
boxplot(MSRP ~ Vehicle.Style, data = cars_cleandata,
        main = "Price by Vehicle Style", xlab = "Vehicle Style",
        ylab = "Price", col = "lightgreen")
```




```
## **4(d) Correlation Analysis**
## This analysis helps determine which factors influence car prices the most.
## High correlations suggest strong relationships between MSRP and variables like engine power and popu

# Calculate correlation matrix
cor_matrix <- cor(cars_cleandata[, c("Year", "Engine.HP", "Engine.Cylinders",
                                     "Number.of.Doors", "highway.MPG", "city.mpg",
                                     "Popularity", "MSRP")], use = "complete.obs")

# Extract correlation with MSRP
cor_with_msrp <- cor_matrix[, "MSRP"]

# Sort correlation values
sorted_cor <- sort(cor_with_msrp, decreasing = TRUE)

# Identify top 3 variables
top_3_variables <- names(sorted_cor[2:4])  # Exclude MSRP itself

# Print top 3 variables
cat("Top 3 variables most correlated with MSRP:", top_3_variables, "")
```

Top 3 variables most correlated with MSRP: Engine.HP Engine.Cylinders Year

```
## **How Does the Brand Affect Popularity and Price?**
## Brands impact car pricing based on reputation, quality, and features.
```

```
## - Popular brands like Toyota and Honda attract buyers with affordability and reliability.
## - Luxury brands such as BMW and Audi command higher prices for premium features and prestige.
## - Mass-market brands like Ford and Chevrolet cater to a wide range of customers.
## - Innovative brands like Tesla leverage advanced technology to justify premium pricing.
```