3 Exploratory Data Analysis

Mohan Khanal

2025-08-03

Table of contents


```
# Chapter 3: Exploratory Data Analysis
# Load required 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)
library(here)
```

```
here() starts at /home/mk/Documents/khanalmohan.github.io
```

```r
# Import dataset
car_data <- read.csv("data/car data.csv")
```

1 —- Part 1: Descriptive Statistics —-

1.1 Using Base Package ——

```r
# Summary of entire dataset
summary(car_data)
```

```
   Car_Name              Year      Selling_Price    Present_Price
 Length:301         Min.   :2003   Min.   : 0.100   Min.   : 0.320
 Class :character   1st Qu.:2012   1st Qu.: 0.900   1st Qu.: 1.200
 Mode  :character   Median :2014   Median : 3.600   Median : 6.400
                    Mean   :2014   Mean   : 4.661   Mean   : 7.628
                    3rd Qu.:2016   3rd Qu.: 6.000   3rd Qu.: 9.900
                    Max.   :2018   Max.   :35.000   Max.   :92.600
  Kms_Driven      Fuel_Type         Seller_Type        Transmission
 Min.   :   500   Length:301         Length:301         Length:301
 1st Qu.: 15000   Class :character   Class :character   Class :character
 Median : 32000   Mode  :character   Mode  :character   Mode  :character
 Mean   : 36947
 3rd Qu.: 48767
 Max.   :500000
     Owner
 Min.   :0.00000
 1st Qu.:0.00000
 Median :0.00000
 Mean   :0.04319
 3rd Qu.:0.00000
 Max.   :3.00000
```

```r
# Base R Descriptive Stats
mean(car_data$Selling_Price, na.rm = TRUE)
```

```
[1] 4.661296
```

```r
median(car_data$Selling_Price, na.rm = TRUE)
```

```
[1] 3.6
```

```r
sd(car_data$Selling_Price, na.rm = TRUE)
```

```
[1] 5.082812
```

```r
var(car_data$Selling_Price, na.rm = TRUE)
```

```
[1] 25.83497
```

```r
range(car_data$Selling_Price, na.rm = TRUE)
```

```
[1]  0.1 35.0
```

```r
quantile(car_data$Selling_Price, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
```

```
25% 50% 75%
0.9 3.6 6.0
```

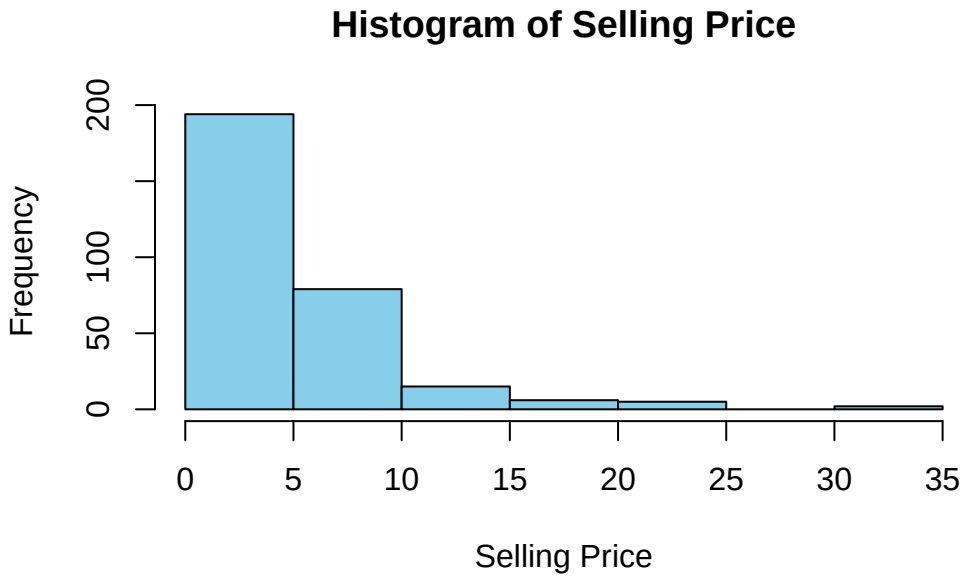
1.2 Using dplyr function —-

```r
# dplyr group-wise summary
car_data %>%
  group_by(Fuel_Type) %>%
  summarize(
    count = n(),
    avg_price = mean(Selling_Price),
    sd_price = sd(Selling_Price),
    max_price = max(Selling_Price)
  )
```

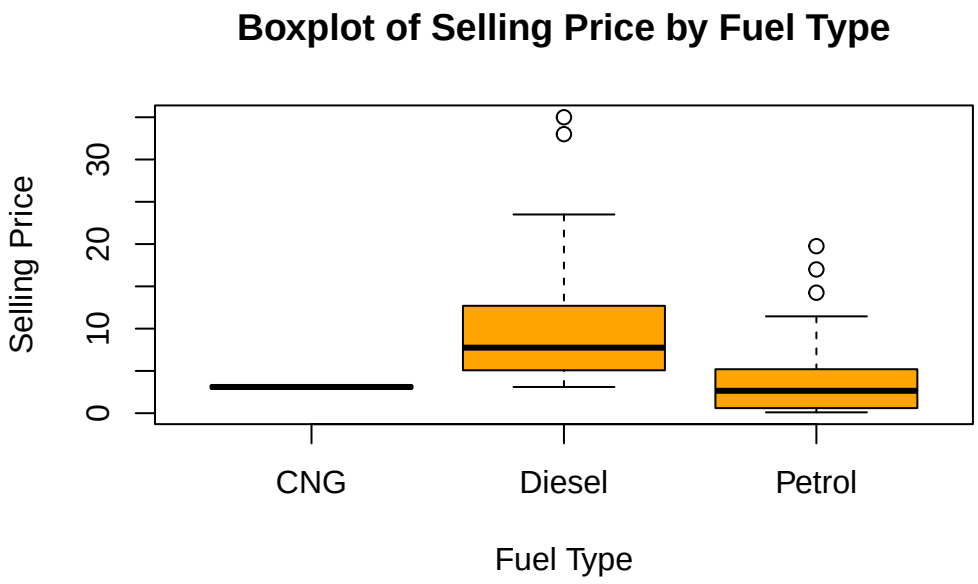
```
# A tibble: 3 x 5
  Fuel_Type count avg_price sd_price max_price
  <chr>     <int>     <dbl>    <dbl>     <dbl>
1 CNG           2      3.1     0.212      3.25
2 Diesel       60     10.3     7.19      35
3 Petrol      239      3.26    3.14      19.8
```

2 —- Part 2: Data Visualization using Base R —-

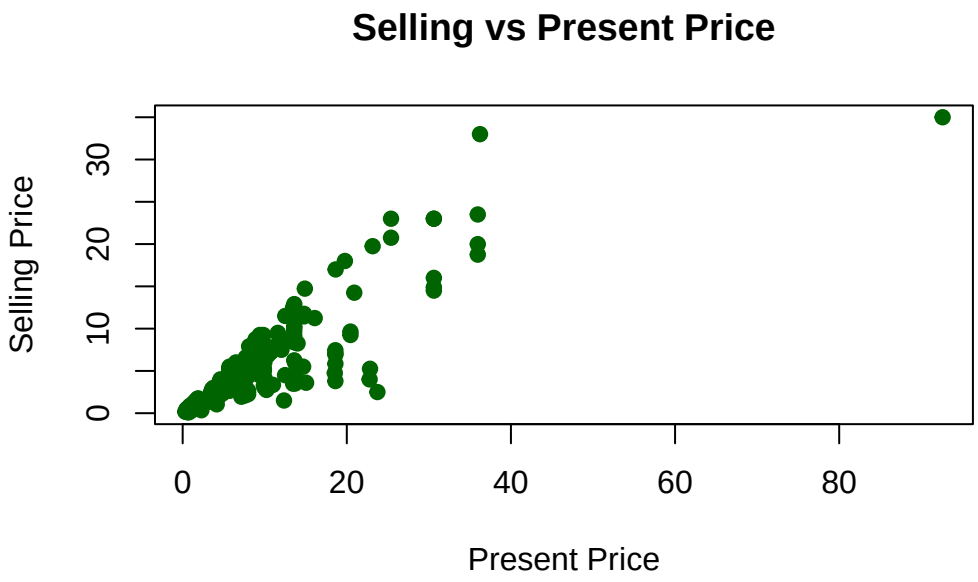
```r
# Histogram
hist(car_data$Selling_Price,
     main = "Histogram of Selling Price",
     xlab = "Selling Price",
     col = "skyblue",
     border = "black")
```



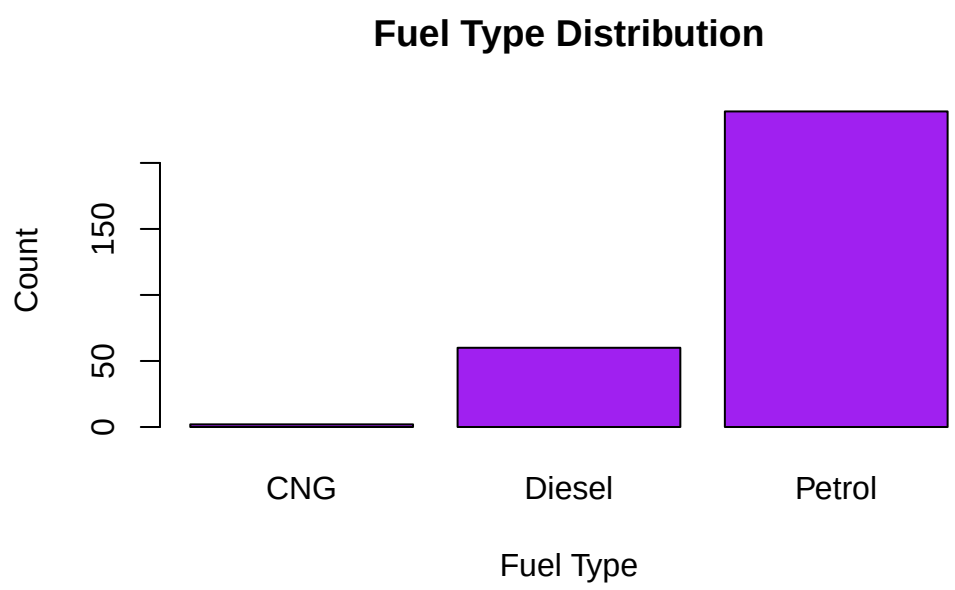

```r
# Boxplot
boxplot(Selling_Price ~ Fuel_Type, data = car_data,
        main = "Boxplot of Selling Price by Fuel Type",
        xlab = "Fuel Type", ylab = "Selling Price",
        col = "orange")
```




```
# Scatter Plot
plot(car_data$Present_Price, car_data$Selling_Price,
     main = "Selling vs Present Price",
     xlab = "Present Price", ylab = "Selling Price",
     pch = 19, col = "darkgreen")
```

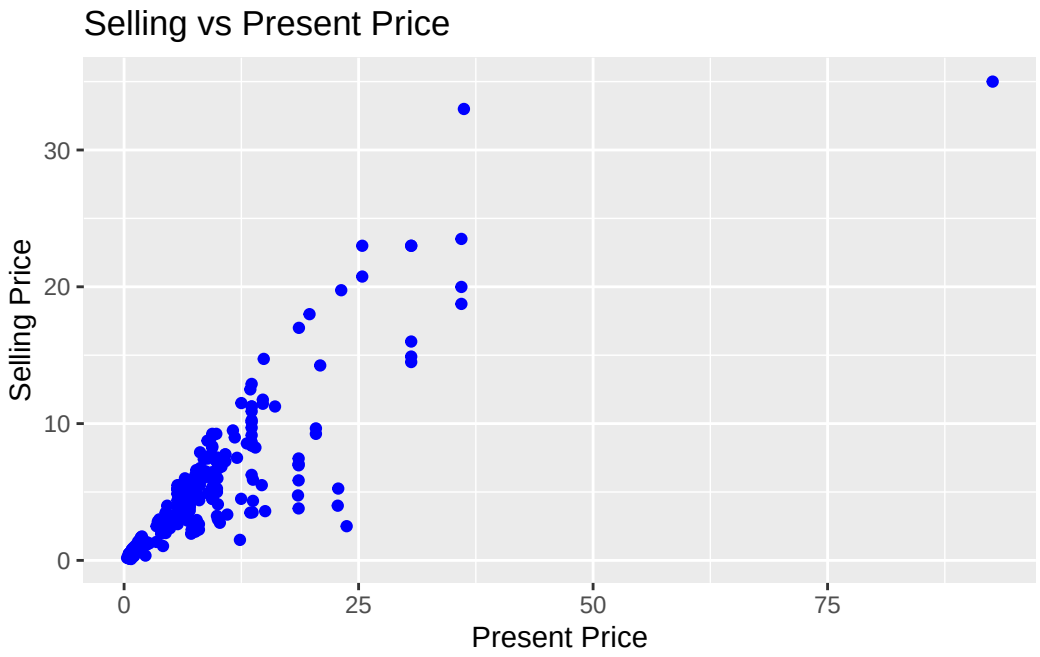



```
# Bar Plot
fuel_counts <- table(car_data$Fuel_Type)
barplot(fuel_counts,
        main = "Fuel Type Distribution",
        col = "purple",
        xlab = "Fuel Type",
        ylab = "Count")
```

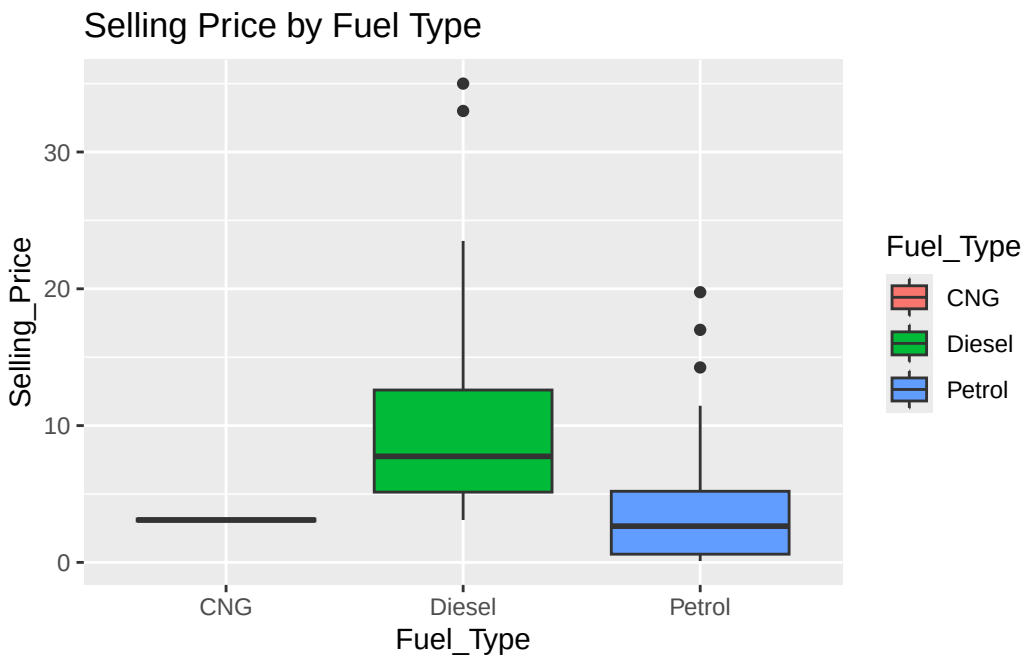


3 —- Part 3: Data Visualization using ggplot2 (Optional) —-

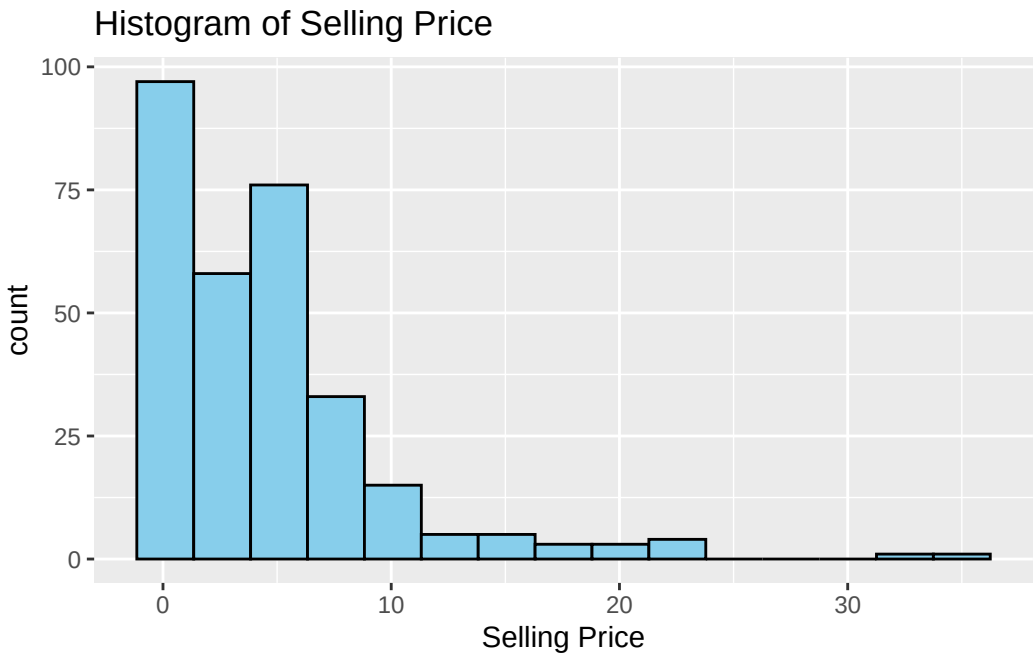
```
# Scatter Plot
ggplot(car_data, aes(x = Present_Price, y = Selling_Price)) +
  geom_point(color = "blue") +
  labs(title = "Selling vs Present Price", x = "Present Price", y = "Selling Price")
```




```
# Boxplot
ggplot(car_data, aes(x = Fuel_Type, y = Selling_Price, fill = Fuel_Type)) +
  geom_boxplot() +
  labs(title = "Selling Price by Fuel Type")
```




```
# Histogram
ggplot(car_data, aes(x = Selling_Price)) +
  geom_histogram(fill = "skyblue", color = "black", bins = 15) +
  labs(title = "Histogram of Selling Price", x = "Selling Price")
```



4 —- Part 4: Practice Tasks —-

4.1 Statistics —-

```
# Mean, median, SD of Present_Price
mean(car_data$Present_Price, na.rm = TRUE)
```

```
[1] 7.628472
```

```
median(car_data$Present_Price, na.rm = TRUE)
```

```
[1] 6.4
```

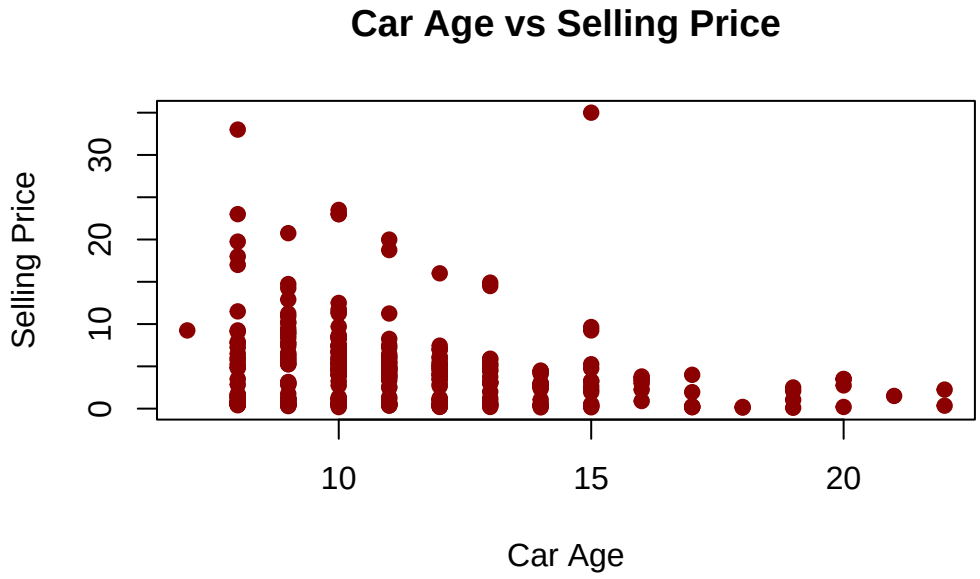
```
sd(car_data$Present_Price, na.rm = TRUE)
```

```
[1] 8.644115
```

5 Create Car Age and Scatter plot vs selling price —-

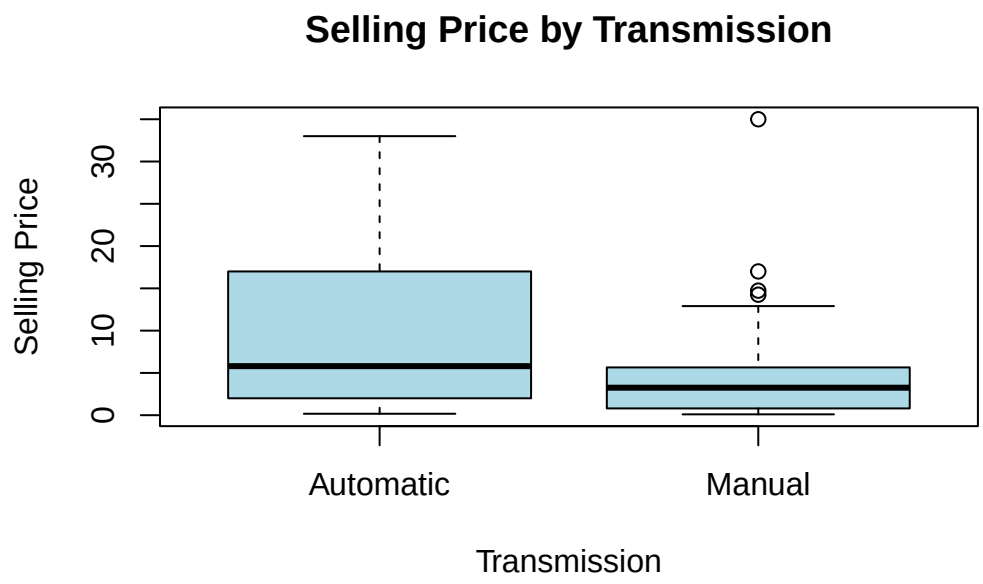
```
# Create Car_Age variable
car_data <- car_data %>%
  mutate(Car_Age = 2025 - Year)

# Scatter plot of Car_Age vs Selling_Price
plot(car_data$Car_Age, car_data$Selling_Price,
     main = "Car Age vs Selling Price",
     xlab = "Car Age", ylab = "Selling Price",
     pch = 19, col = "darkred")
```



6 Boxplot across Transmission —-

```
boxplot(Selling_Price ~ Transmission, data = car_data,
        main = "Selling Price by Transmission",
        xlab = "Transmission", ylab = "Selling Price",
        col = "lightblue")
```



7 Count by Seller_Type —-

```
car_data %>%
  group_by(Seller_Type) %>%
  summarize(count = n())
```

```
# A tibble: 2 x 2
  Seller_Type count
  <chr>       <int>
1 Dealer        195
2 Individual    106
```

8 Feedback —-