Project 1 : Statistics Data Analysis Project: Real-World Data Exploration

Deadline: Before 11:59 pm on Friday, Feb. 28

Phillip Korolev

2025-03-03 14:29:16.835966

Contents

1 Project Overview

This project will involve:

1. Finding and downloading a dataset from an open-source platform.
2. Exploring the dataset (checking structure, missing values, data types).
3. Visualizing categorical and continuous variables.
4. Analyzing central tendency, skewness, and correlations.
5. Interpreting insights and telling a story using statistical measures.

1.1 Step 1: Introduction and Report Goal

Uber Technologies is a ride-sharing and ride-provider application developed in 2019 that lets users facilitate their travels from point A to point B by utilizing regular Uber-employed (contracted) drivers instead of taxi services, which was an outdated and less friendly experience. Since the launch of the application, Uber has been faced with challenges ranging from funding to country-wide lock down measures implemented during health pandemics. Alternatively, Uber has also seen extreme success during two major periods for the company.

Nevertheless, anything to do with a company can be generally measured alongside their stock performance, which tends to dip during hardship and rise during success. This report aims to visualize Uber's stock price, as well as other stock metrics, from IPO launch (2019) to February of this year.

The dataset includes some of the more general metrics when analyzing stock data, such as **Open** and **Close Price**, **Daily High** and **Low Price**, and **Daily Volume**. It might have already become apparent, but this dataset includes only continuous variables, except for its **Date** variable, which can follow both a categorical and continuous structure.

Our set of data allows us to draw a number of different questions and insights. For instance - and one that we will explore later on - "*How can daily trading volume effect a stock's daily high price?*". Other questions based on the data might look like:

- Can a relatively larger open price effect trading volume for a particular day?
- Does a stock's open price strongly indicate its price at close?
- Does a stock's close price indicate anything about its next day open price?

1.2 Step 2: Data Import and Cleaning

```
# Load dataset into R
uber_data <- read.csv('C:/Users/pkoro/OneDrive/Desktop/stock-data-uber.csv')

# Inspect the structure of the data
head(uber_data)
```

```
##         Date Adj.Close Close  High   Low  Open    Volume
## 1 2019-05-10     41.57 41.57 45.00 41.06 42.00 186322500
## 2 2019-05-13     37.10 37.10 39.24 36.08 38.79  79442400
## 3 2019-05-14     39.96 39.96 39.96 36.85 38.31  46661100
## 4 2019-05-15     41.29 41.29 41.88 38.95 39.37  36086100
## 5 2019-05-16     43.00 43.00 44.06 41.25 41.48  38115500
## 6 2019-05-17     41.91 41.91 43.29 41.27 41.98  20225700
```

```
str(uber_data)
```

```
## 'data.frame':    1444 obs. of  7 variables:
##  $ Date     : chr  "2019-05-10" "2019-05-13" "2019-05-14" "2019-05-15" ...
##  $ Adj.Close: num  41.6 37.1 40 41.3 43 ...
##  $ Close    : num  41.6 37.1 40 41.3 43 ...
##  $ High     : num  45 39.2 40 41.9 44.1 ...
##  $ Low      : num  41.1 36.1 36.8 39 41.2 ...
##  $ Open     : num  42 38.8 38.3 39.4 41.5 ...
##  $ Volume   : int  186322500 79442400 46661100 36086100 38115500 20225700 29222300 10802900
```

```
dim(uber_data)
```

```
## [1] 1444    7
```

```
sum(is.na(uber_data))
```

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

```
# Handle missing values (if any)
uber_data <- na.omit(uber_data)  # Remove rows with missing data

# Convert Date to Date format
uber_data$Date <- as.Date(uber_data$Date, format="%Y-%m-%d")
```

Our data set never included null rows, or even null cells inside the *csv* file, therefore our data omitting step is relatively useless; it serves as a safety net. During this step, we also change our date variable to include formal **Date** formatting throughout each entry.

1.3 Step 3: Exploratory Data Analysis

Generating summary statistics for some of our key variables:

```
# Generating summary statistics
cat("Open Price Summary:", "\n")
```

```
## Open Price Summary:
```

```
summary(uber_data$Open)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.96   31.87   41.22   44.49   54.66   85.64
```

```
cat("\n")
```

```
cat("Volume Summary:", "\n")
```

```
## Volume Summary:
```

```
summary(uber_data$Volume)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   3380000  14989050  20369650  24298003  28432800 364231800
```

```
cat("\n")
```

```
cat("Price High Summary:", "\n")
```

```
## Price High Summary:
```

```
summary(uber_data$High)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.80   32.65   41.91   45.29   55.63   87.00
```

```
cat("\n")
```

```
cat("Price Low Summary:", "\n")
```

```
## Price Low Summary:
```

```
summary(uber_data$Low)
```

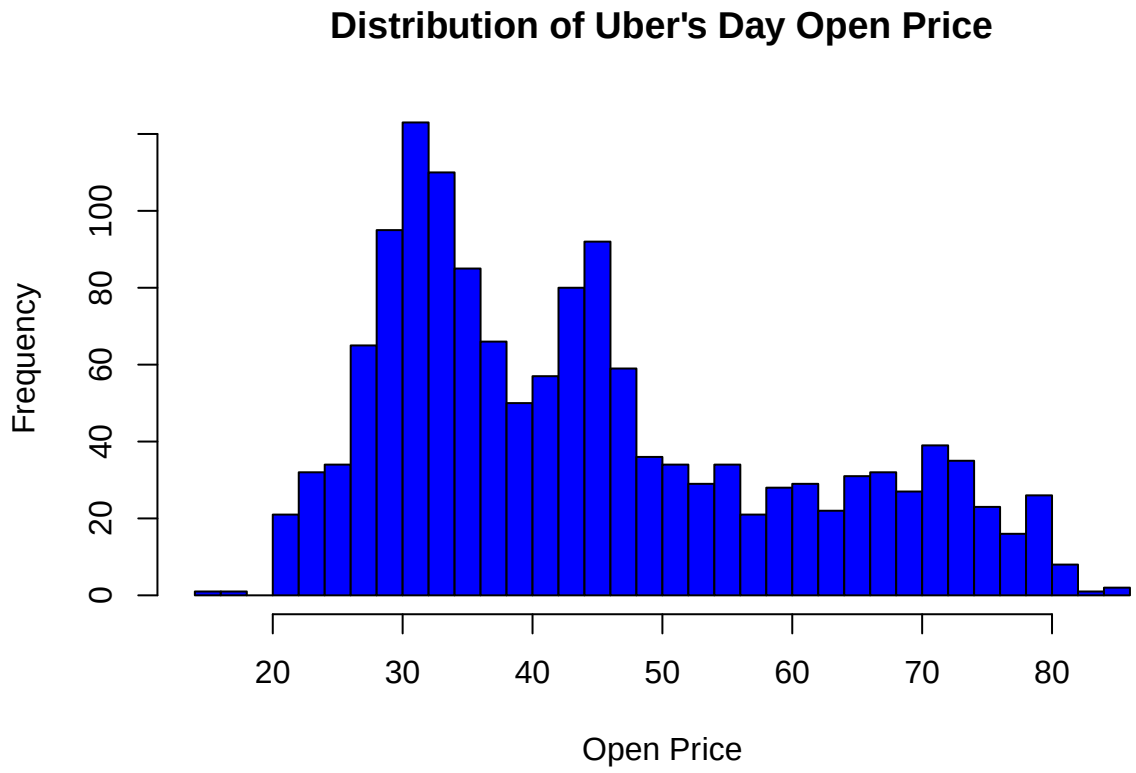
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.71   31.18   40.52   43.64   53.60   84.18
```

```
cat("\n")
```

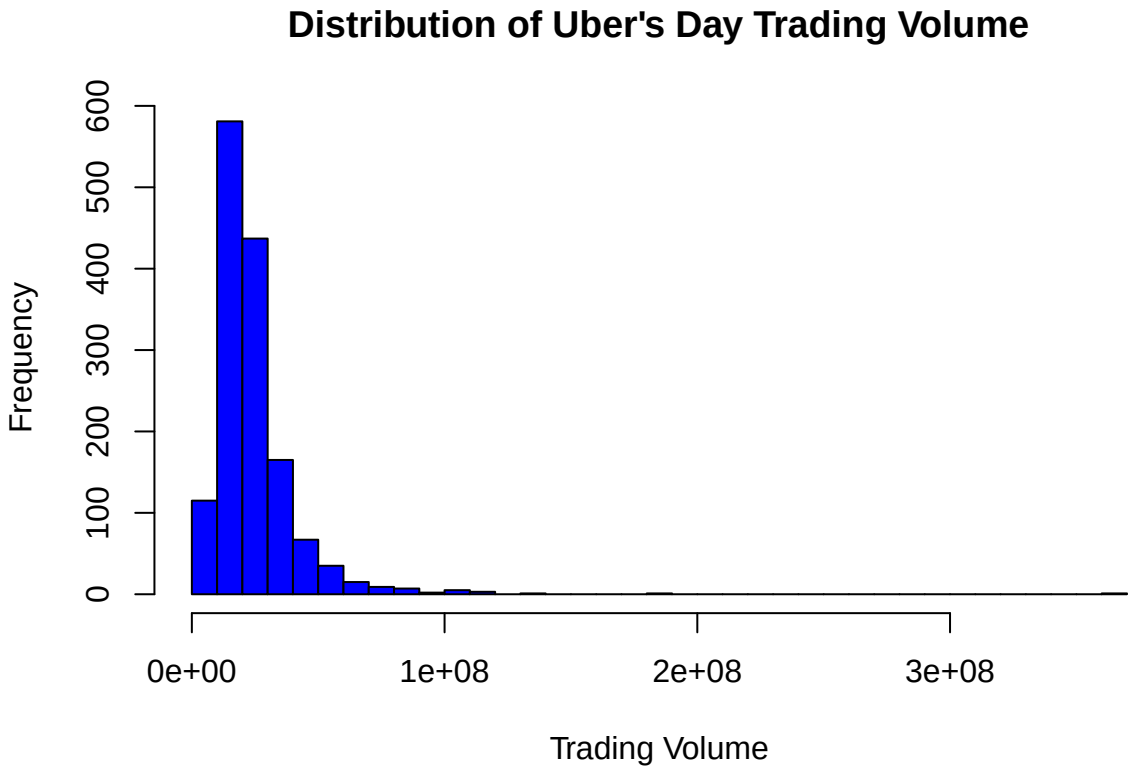
Numerically speaking, our Price values tend to fall into a smaller range, while Trading Volume values range from a minimum three million, to almost four-hundred million, a jump of over **13,000%**. The reason for this extreme range is primarily self-explanatory: a dramatic increase in trading volume. Though, we will re-cover this topic in greater detail when drawing conclusions based on generated box plots.

Next, we generate frequency distribution charts for **Open Price** and **Trading Volume**:

```r
# Create a histogram for the Open price
hist(uber_data$Open,
     breaks = 30,
     col = "blue",
     border = "black",
     main = "Distribution of Uber's Day Open Price",
     xlab = "Open Price",
     ylab = "Frequency")
```




```r
# Creating a histogram for the Volume
hist(uber_data$Volume,
     breaks = 30,
     col = "blue",
     border = "black",
     main = "Distribution of Uber's Day Trading Volume",
     xlab = "Trading Volume",
     ylab = "Frequency")
```



In both distribution graphs, we can see that the stock primarily sat within the 30-55 dollar range during the 2019-2025 period, with the trading volume summarized by this trend as well. Though as we know, Uber exploded in popularity and investment post-covid, touching higher stock prices and daily trading volume towards the latter two years of the above date range. Since these increased numbers only started to shape largely during 2024, the numbers show up less frequent than those reflected in earlier years of this date range (larger portion of the date range).

Generating Correlation values between **Open/High Price**, **Open/Low Price, High/Volume, High/Low Price** values:

```
# Creating correlation value variables
cor_open_high <- cor(uber_data$Open, uber_data$High)
cor_open_low <- cor(uber_data$Open, uber_data$Low)
cor_volume_high <- cor(uber_data$Volume, uber_data$High)
cor_high_low <- cor(uber_data$High, uber_data$Low)

#Display results
cat("Correlation between day Open/High values:", cor_open_high, "\n")

## Correlation between day Open/High values: 0.9989435

cat("Correlation between day Open/Low values:", cor_open_low, "\n")

## Correlation between day Open/Low values: 0.9987952
```

```
cat("Correlation between day High/Low values:", cor_high_low, "\n")
```

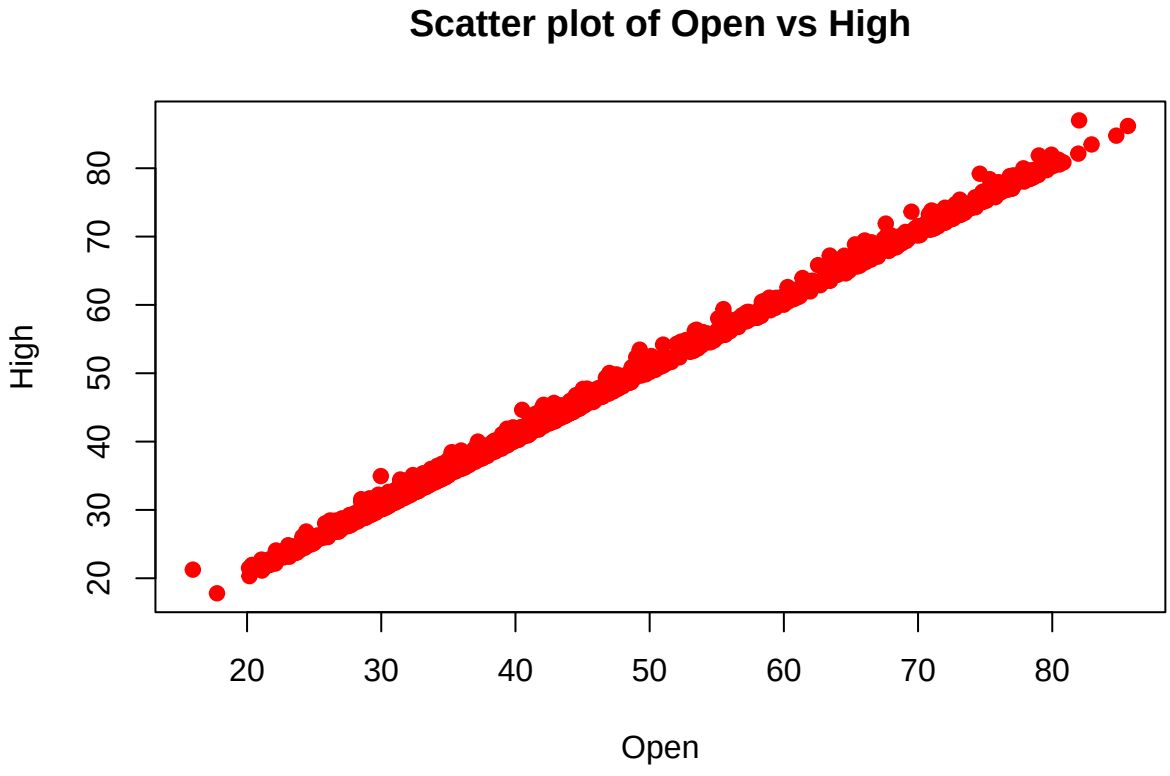
```
## Correlation between day High/Low values: 0.9987323
```

```
cat("Correlation between day High/Volume values:", cor_volume_high, "\n")
```

```
## Correlation between day High/Volume values: -0.1600539
```

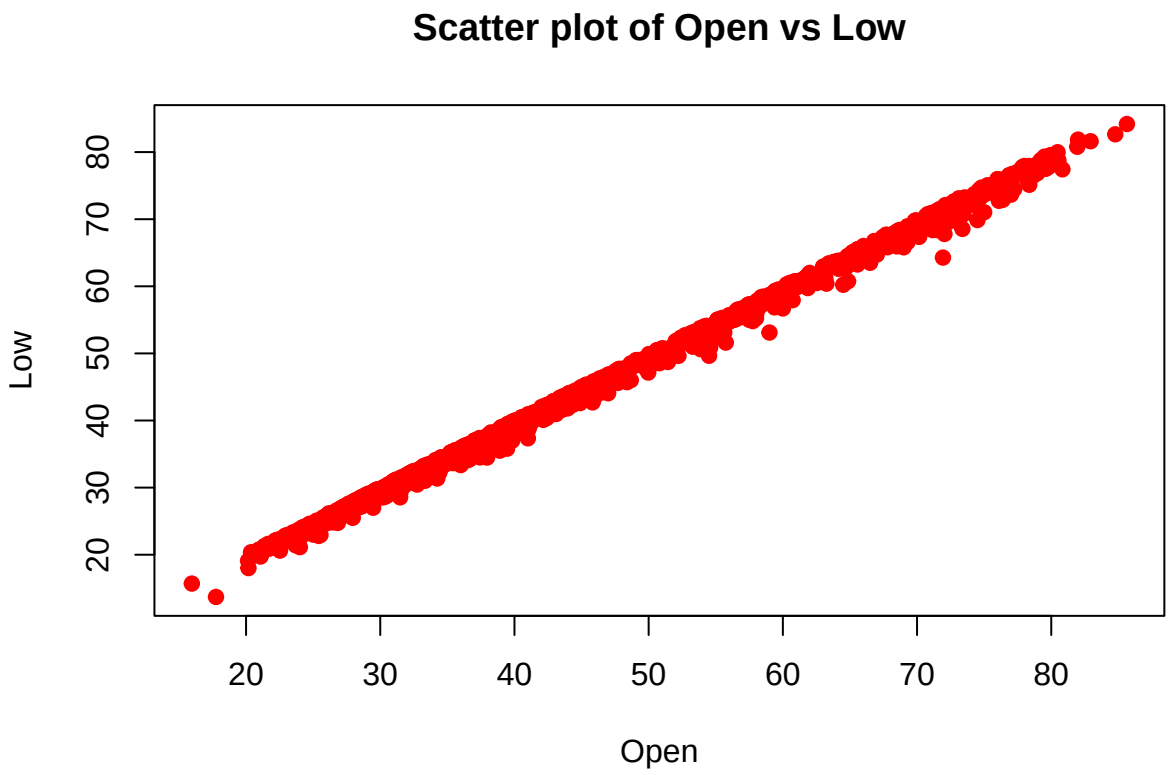
Generating scatter plots to visualize **Open/High Price**, **Open/Low Price**, and **Volume/High** relationships. In the fourth plot, we can also see a general price graph throughout our date range.

```
# Scatter plot for Open vs High
plot(uber_data$Open, uber_data$High,
     main = "Scatter plot of Open vs High",
     xlab = "Open",
     ylab = "High",
     col = "red",
     pch = 19)
```

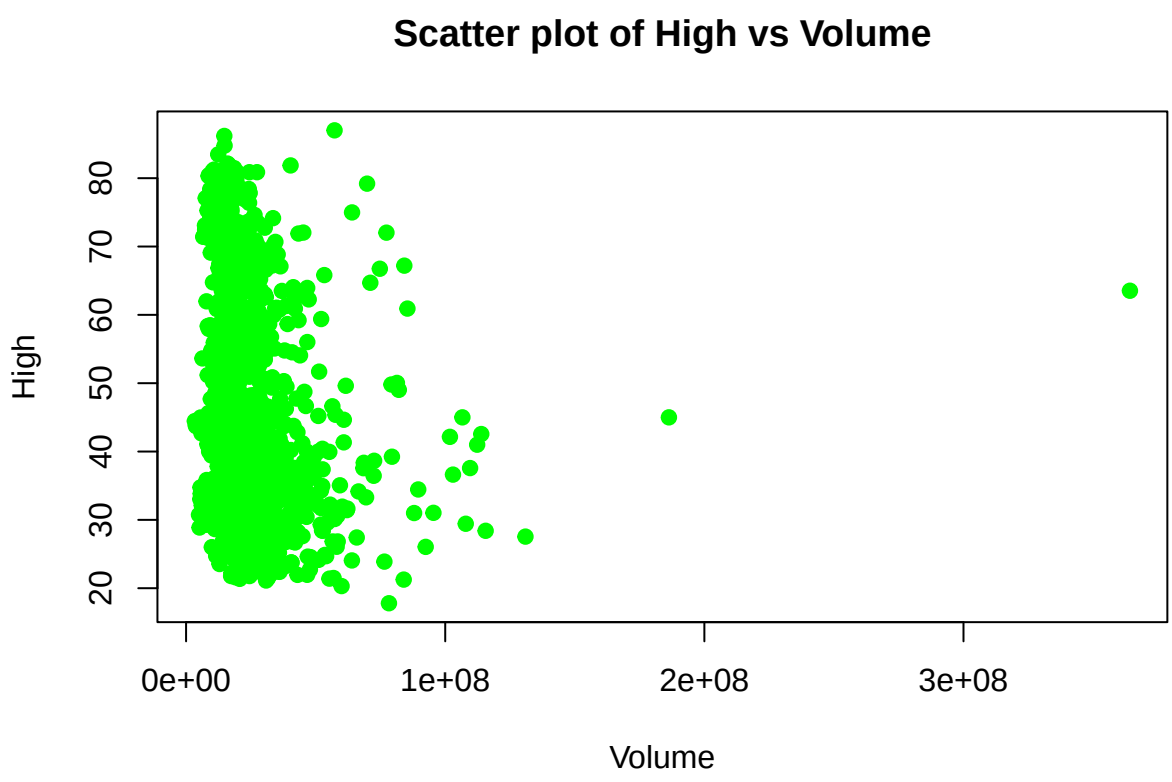



```
# Scatter plot for Open vs Low
plot(uber_data$Open, uber_data$Low,
```

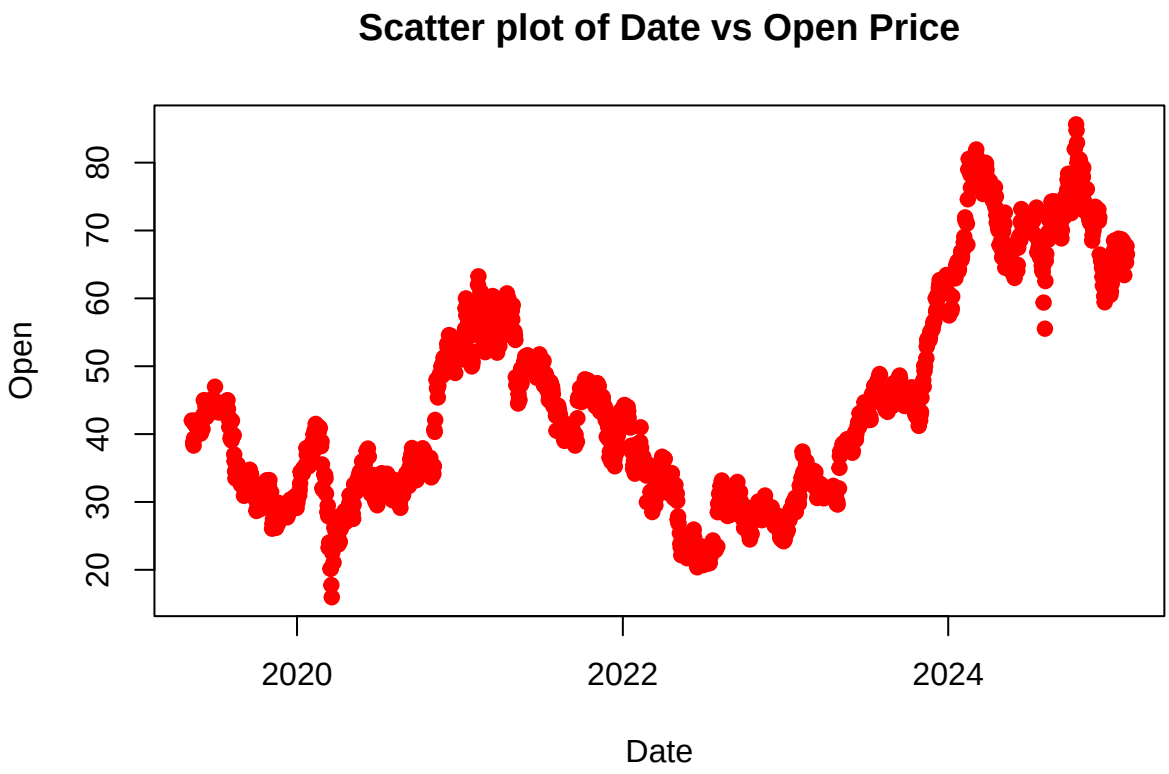
```
     main = "Scatter plot of Open vs Low",
     xlab = "Open",
     ylab = "Low",
     col = "red",
     pch = 19)
```




```
# Scatter plot for High vs Volume
plot(uber_data$Volume, uber_data$High,
     main = "Scatter plot of High vs Volume",
     xlab = "Volume",
     ylab = "High",
     col = "green",
     pch = 19)
```




```r
# General plot for stock price throughout our selected time range
plot(uber_data$Date, uber_data$Open,
     main = "Scatter plot of Date vs Open Price",
     xlab = "Date",
     ylab = "Open",
     col = "red",
     pch = 19)
```



We have an almost perfect correlation between Open Price and High/Low Price as expected. This tells us that Uber is not such a volatile or risky stock:

- Non-volatile stocks are stocks who's prices do not significantly increase or decrease throughout a short period of time. In Uber's case, the stock generally sits around the Open Price during a particular day, experiencing almost no swings, but rather very minor shifts in price.
- In the long haul, as seen in the fourth graph, price changes will look more consistent and will tend to follow a pattern for a select period of time.

1.4 Step 4: Statistical Analysis

Displaying key statistics For **Open Price** and **Volume**

```
# Compute mean, median, variance, and sd for Open Price
mean_open <- mean(uber_data$Open)
median_open <- median(uber_data$Open)
variance_open <- var(uber_data$Open)
sd_open <- sd(uber_data$Open)

# Compute mean, median, variance, and sd for Volume
mean_vol <- mean(uber_data$Volume)
```

```r
median_vol <- median(uber_data$Volume)
variance_vol <- var(uber_data$Volume)
sd_vol <- sd(uber_data$Volume)

# Display Open Price stats
cat("Mean of Open Price:", mean_open, "\n")

## Mean of Open Price: 44.49302

cat("Median of Open Price:", median_open, "\n")

## Median of Open Price: 41.215

cat("Variance of Open Price:", variance_open, "\n")

## Variance of Open Price: 244.7337

cat("Standard Deviation of Open Price:", sd_open, "\n")

## Standard Deviation of Open Price: 15.64397

cat("\n")

#Display Volume stats
cat("Mean of Daily Trading Volume:", mean_vol, "\n")

## Mean of Daily Trading Volume: 24298003

cat("Median of Daily Trading Volume:", median_vol, "\n")

## Median of Daily Trading Volume: 20369650

cat("Variance of Daily Trading Volume:", variance_vol, "\n")

## Variance of Daily Trading Volume: 3.147372e+14

cat("Standard Deviation of Daily Trading Volume:", sd_vol, "\n")

## Standard Deviation of Daily Trading Volume: 17740835
```

Computing Skewness of daily **Open Price** and **Trading Volume**:

```r
# Compute skewness using formula
n_open <- length(uber_data$Open)
skewness_open <- sum((uber_data$Open - mean_open)^3) / (n_open*(sd_open^3))

n_vol <- length(uber_data$Volume)
skewness_vol <- sum((uber_data$Volume - mean_vol)^3) / (n_vol*(sd_vol^3))

# Display skewness results
cat("Skewness of Open Price:", skewness_open, "\n")
```

```
## Skewness of Open Price: 0.6749594
```

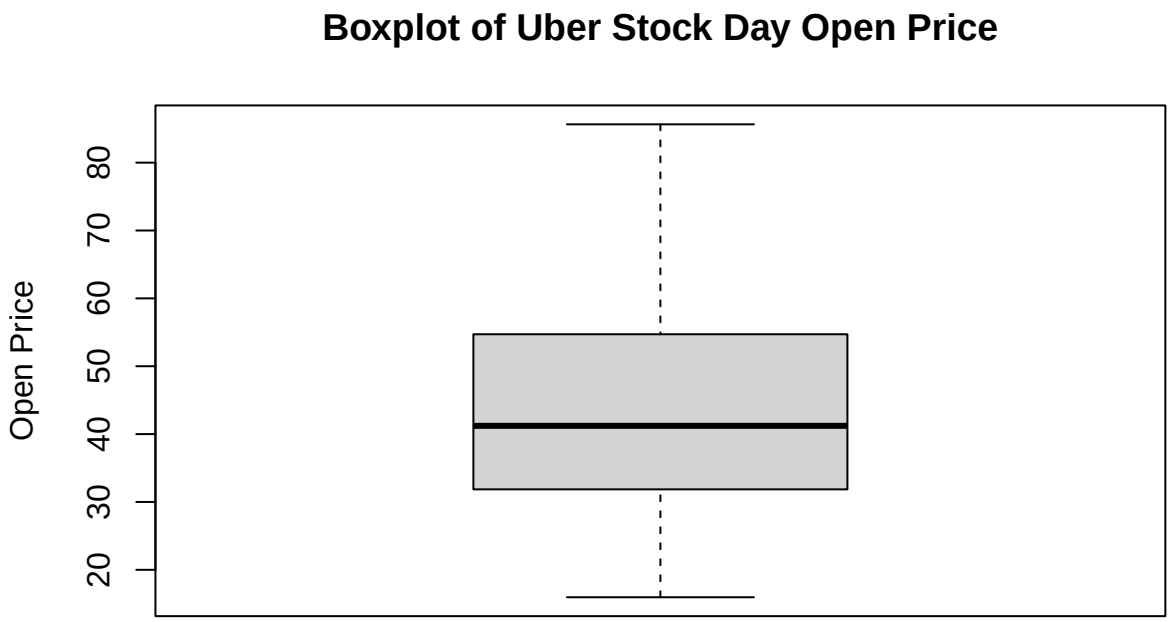
```r
cat("Skewness of Trading Volume:", skewness_vol, "\n")
```

```
## Skewness of Trading Volume: 6.794444
```

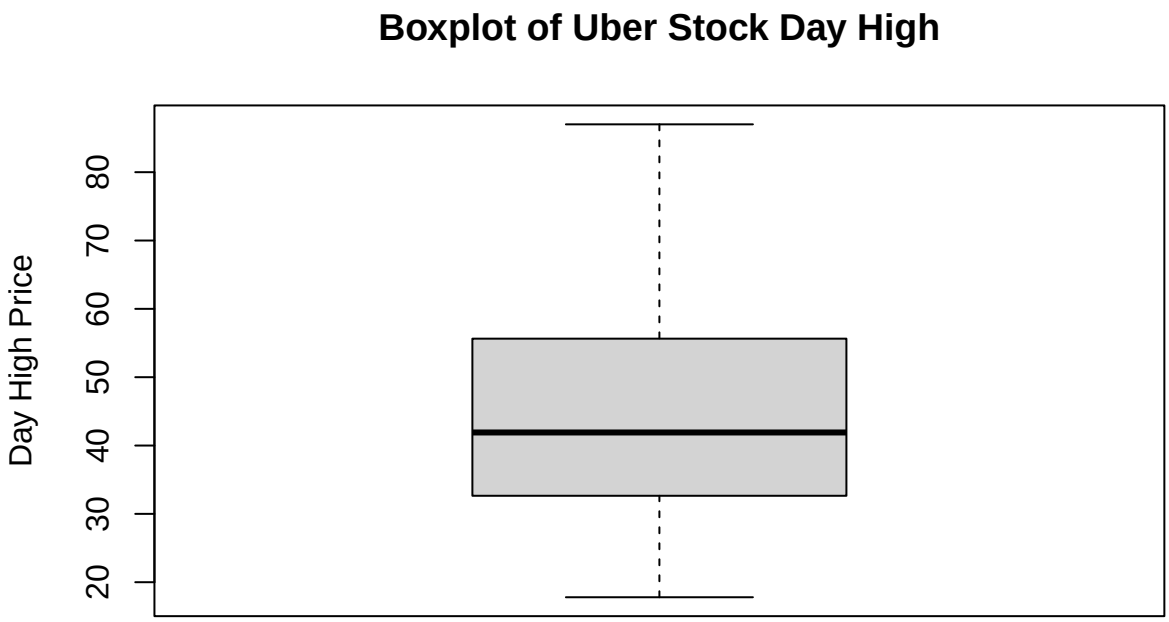
From the skewness data above, we can draw the conclusion that the **Open Price** follows (more or less) a normal distribution as its skewness is extremely close to zero. On the other, a variable like **Trading Volume** is more right-skewed, with an evident $skewness > 0$.

Checking for daily **Open Price, High/Low, Volume** outliers using box plots:

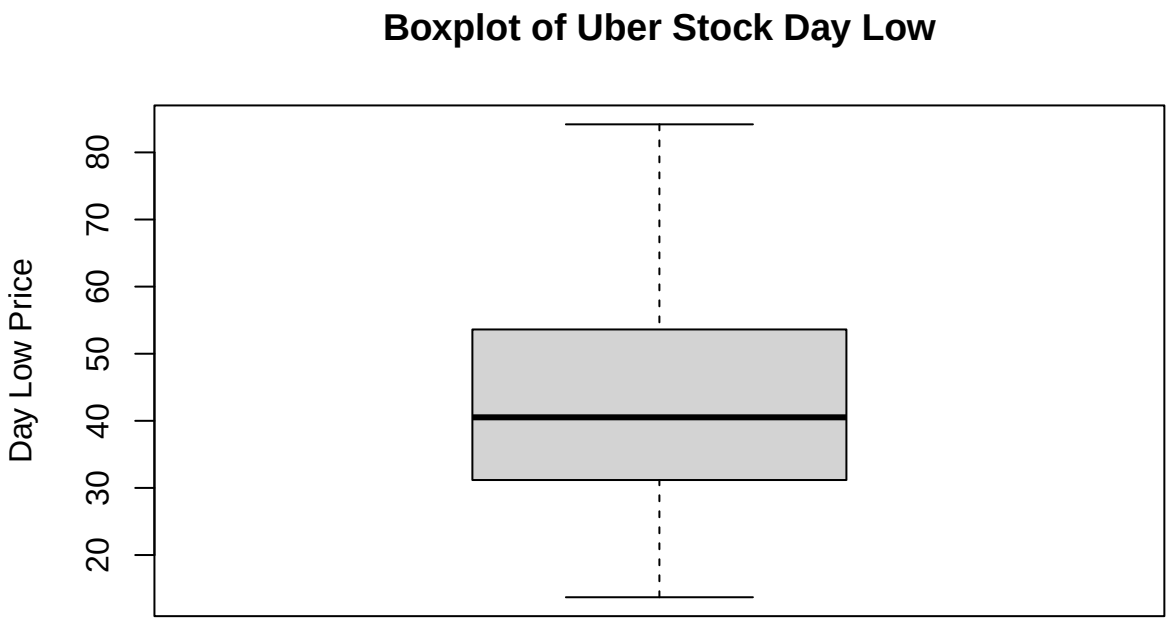
```r
# Creating day Open Price box plot
boxplot(uber_data$Open,
        main = "Boxplot of Uber Stock Day Open Price",
        ylab = "Open Price")
```



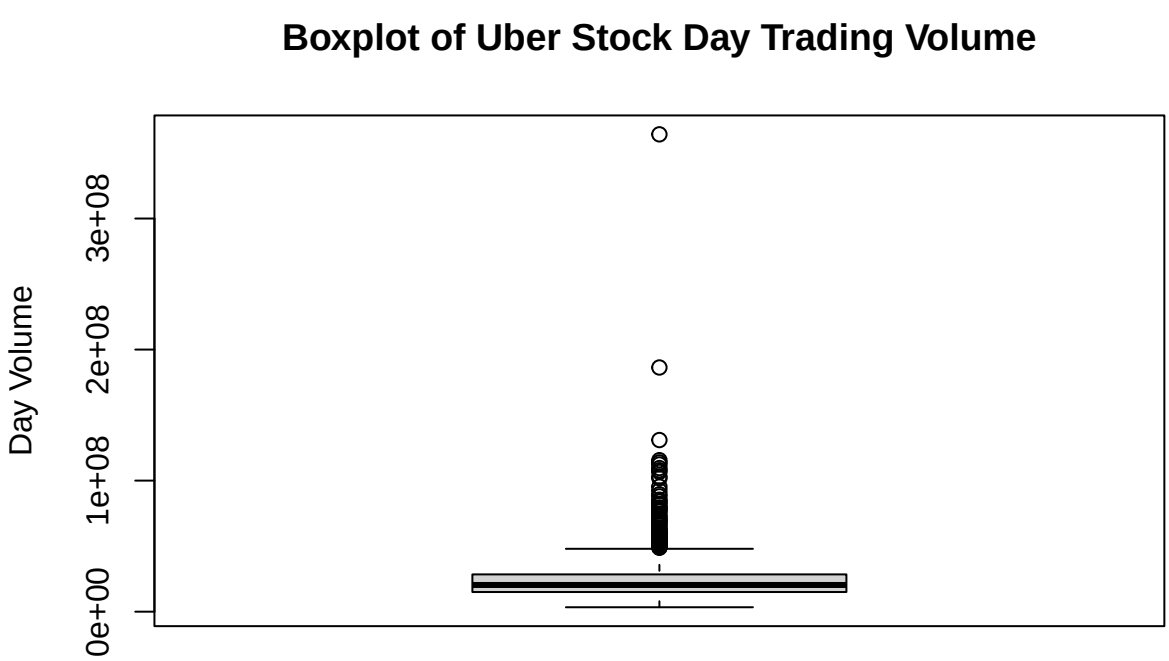

```
# Creating day High box plot
boxplot(uber_data$High,
        main = "Boxplot of Uber Stock Day High",
        ylab = "Day High Price")
```




```
# Creating day Low box plot
boxplot(uber_data$Low,
        main = "Boxplot of Uber Stock Day Low",
        ylab = "Day Low Price")
```




```
# Creating day Volume box plot
boxplot(uber_data$Volume,
        main = "Boxplot of Uber Stock Day Trading Volume",
        ylab = "Day Volume")
```



We see from the above generated box plots that Uber stock prices had no general outliers from our analyzed date range. Throughout all three price plots (Open, High, Low), we see that all data points lie within our whisker range. This lets us draw the following conclusion about the three box plots:

- Daily Open, High, Low prices generally follow a normal distribution, as their box plot is symmetric

On the other hand, the Volume box plot suggests an extreme amount of outliers, which can be partially explained by Uber's astronomical rise in popularity and company success during the post-covid period (2023-2025). In this period, the company attracted significantly more investors, - *including both retail and professional investors/formal funds* - exploding trading volume on certain days and weeks, causing extreme outliers.