CS174 FA2.2

Darryl Babar, Christian Henry Miguel Caruz, Miguel Soniel & Jan Edgar Tupas BM1

In this project, we will explore on the given Flight_Data dataset. The dataset includes multiple features about a certain flight. These include Airline, Date of Journey, Source, Destination, Route, Departure Time, Arrival Time, Duration, Totals Stops, Additional Info, and Price. We will try to predict the dependent variable price using the remaining features as independent variables with a multiple regression model.

We will first unpack all the necessary libraries for the project. The libraries we will use are data.table, ggplot2, rpart, and rpart.plot.

```r
library(data.table)
library(ggplot2)
library(rpart)
library(rpart.plot)
```

Next, initialize and load our dataset into a dataframe named flight_data. We can show the first five samples in the dataframe, as well as the structure and summary of the dataset.

```r
flight_data <- fread(file = "Data_Train.csv", head = TRUE)

head(flight_data)

str(flight_data)

summary(flight_data)
```

```
> head(flight_data)
       Airline Date_of_Journey  Source Destination                 Route Dep_Time Arrival_Time
        <char>          <char>  <char>      <char>                <char>   <char>       <char>
1:      IndiGo      24/03/2019 Banglore   New Delhi            BLR ? DEL    22:20 01:10 22 Mar
2:   Air India       1/05/2019  Kolkata    Banglore CCU ? IXR ? BBI ? BLR   05:50        13:15
3: Jet Airways       9/06/2019    Delhi      Cochin DEL ? LKO ? BOM ? COK   09:25 04:25 10 Jun
4:      IndiGo      12/05/2019  Kolkata    Banglore       CCU ? NAG ? BLR   18:05        23:30
5:      IndiGo      01/03/2019 Banglore   New Delhi       BLR ? NAG ? DEL   16:50        21:35
6:    SpiceJet      24/06/2019  Kolkata    Banglore             CCU ? BLR   09:00        11:25
   Duration Total_Stops Additional_Info Price
     <char>      <char>          <char> <int>
1:   2h 50m    non-stop         No info  3897
2:   7h 25m     2 stops         No info  7662
3:      19h     2 stops         No info 13882
4:   5h 25m      1 stop         No info  6218
5:   4h 45m      1 stop         No info 13302
6:   2h 25m    non-stop         No info  3873
```

```
> str(flight_data)
Classes 'data.table' and 'data.frame':  10683 obs. of  11 variables:
 $ Airline        : chr  "IndiGo" "Air India" "Jet Airways" "IndiGo" ...
 $ Date_of_Journey: chr  "24/03/2019" "1/05/2019" "9/06/2019" "12/05/2019" ...
 $ Source         : chr  "Banglore" "Kolkata" "Delhi" "Kolkata" ...
 $ Destination    : chr  "New Delhi" "Banglore" "Cochin" "Banglore" ...
 $ Route          : chr  "BLR ? DEL" "CCU ? IXR ? BBI ? BLR" "DEL ? LKO ? BOM ? COK" "CCU ? NAG ? BLR" ...
 $ Dep_Time       : chr  "22:20" "05:50" "09:25" "18:05" ...
 $ Arrival_Time   : chr  "01:10 22 Mar" "13:15" "04:25 10 Jun" "23:30" ...
 $ Duration       : chr  "2h 50m" "7h 25m" "19h" "5h 25m" ...
 $ Total_Stops    : chr  "non-stop" "2 stops" "2 stops" "1 stop" ...
 $ Additional_Info: chr  "No info" "No info" "No info" "No info" ...
 $ Price          : int  3897 7662 13882 6218 13302 3873 11087 22270 11087 8625 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```
> summary(flight_data)
   Airline          Date_of_Journey       Source           Destination           Route          
 Length:10683       Length:10683       Length:10683       Length:10683       Length:10683      
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                               
                                                                                               
                                                                                               
   Dep_Time         Arrival_Time         Duration         Total_Stops        Additional_Info   
 Length:10683       Length:10683       Length:10683       Length:10683       Length:10683      
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                               
                                                                                               
                                                                                               
     Price      
 Min.   : 1759  
 1st Qu.: 5277  
 Median : 8372  
 Mean   : 9087  
 3rd Qu.:12373  
 Max.   :79512  
```

We will carry out exploratory analysis with our data. We will use histograms to visualize the frequencies of some features and boxplots to show their relationship with the price.

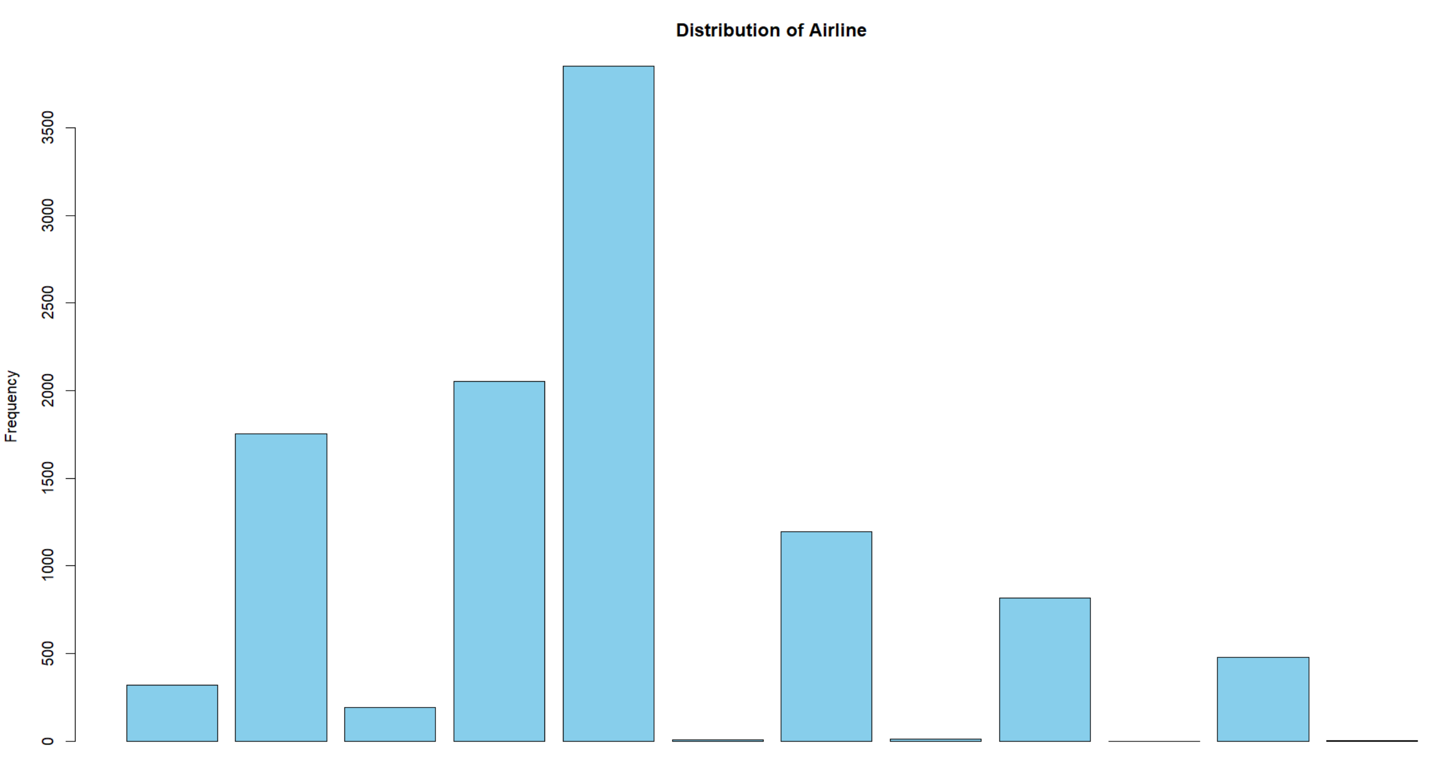
```
#Visualization
options(repr.plot.width = 8, repr.plot.height = 4)

par(mar = c(3, 4, 4, 0.5))

airline_counts <- table(flight_data$Airline)
print(airline_counts)
barplot(airline_counts,
        main = "Distribution of Airline",
        xlab = "Airline",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(airline_counts)),
        cex.names = 0.4)
```

```
info_counts <- table(flight_data$Additional_Info)
print(info_counts)
barplot(info_counts,
        main = "Distribution of Additional Info",
        xlab = "Additional Info",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(info_counts)),
        cex.names = 0.4)

ggplot(flight_data, aes(x = Airline, y = Price)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(title = "Boxplot of Ticket Prices by Airline",
       x = "Airline",
       y = "Ticket Price") +
        theme(axis.text = element_text(size = 7))

ggplot(flight_data, aes(x = Source, y = Price)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(title = "Boxplot of Ticket Prices by Source",
       x = "Source",
       y = "Ticket Price") +
  theme(axis.text = element_text(size = 7))

ggplot(flight_data, aes(x = Destination, y = Price)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(title = "Boxplot of Ticket Prices by Destination",
       x = "Destination",
       y = "Ticket Price") +
  theme(axis.text = element_text(size = 7))
```

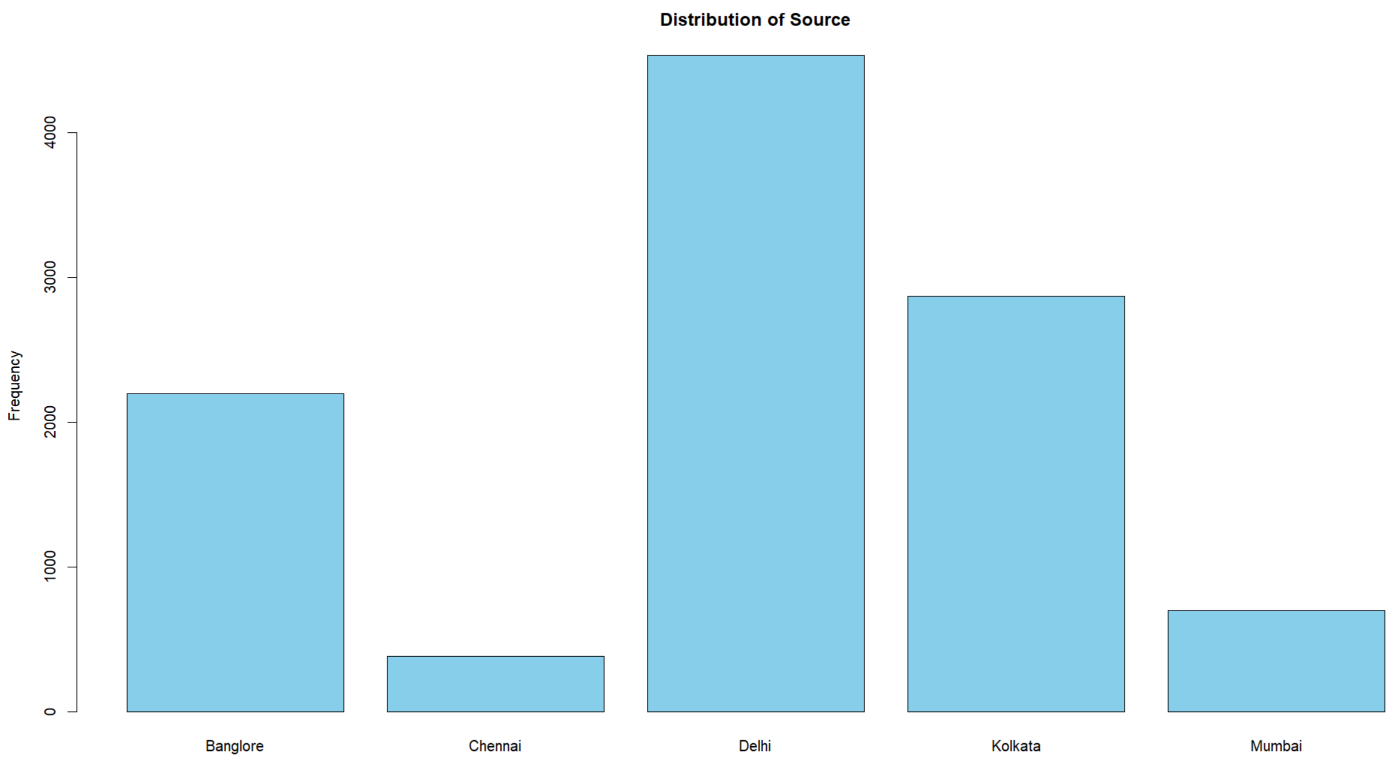
```
source_counts <- table(flight_data$Source)
print(source_counts)
barplot(source_counts,
        main = "Distribution of Source",
        xlab = "Source",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(source_counts)),
        cex.names = 1)

dest_counts <- table(flight_data$Destination)
print(dest_counts)
barplot(dest_counts,
        main = "Distribution of Destination",
        xlab = "Destination",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(dest_counts)),
        cex.names = 1)

route_counts <- table(flight_data$Route)
print(route_counts)
unique_route <- unique(flight_data$Route)
print(unique_route)
len_route <- length(unique_route)
print(len_route)

stop_counts <- table(flight_data$Total_Stops)
print(stop_counts)
barplot(stop_counts,
        main = "Distribution of Total Stops",
        xlab = "Total Stops",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(stop_counts)),
        cex.names = 1)
```

```
ggplot(flight_data, aes(x = Total_Stops, y = Price)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(title = "Boxplot of Ticket Prices by Total Stops",
       x = "Total Stops",
       y = "Ticket Price") +
  theme(axis.text = element_text(size = 7))

ggplot(flight_data, aes(x = Additional_Info, y = Price)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(title = "Boxplot of Ticket Prices by Additional Info",
       x = "Additional Info",
       y = "Ticket Price") +
  theme(axis.text = element_text(size = 7))
```

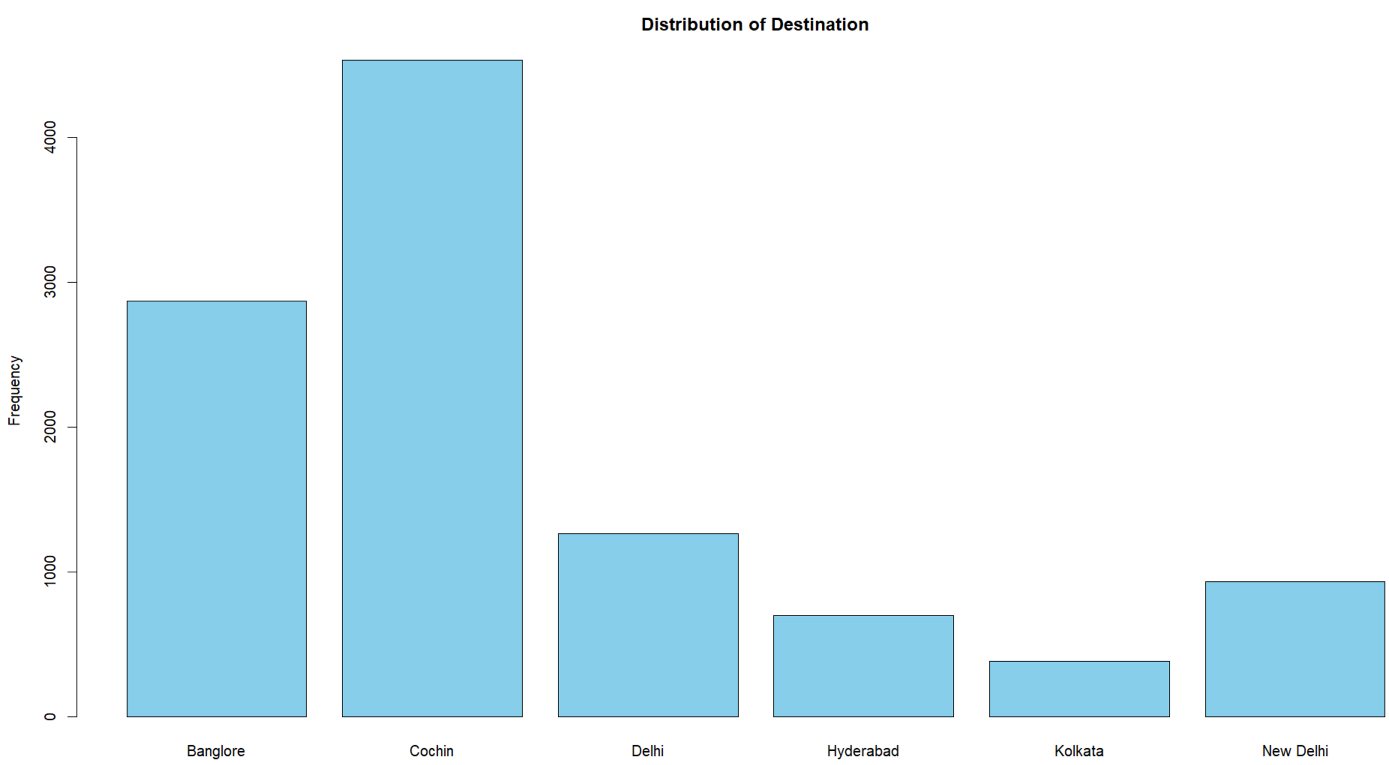
The distribution of airline categories is shown. We can observe that Jet Airways is selected most. It is followed by IndiGo and Air India. We can also observe that Jet Airways Business, Multiple carriers Premium economy, Trujet, and Vistara Premium Economy are the least selected.



Next is the distribution of Source. We can see that most flights are from Delhi, followed by Kolkata and Banglore. Meanwhile, the least number of flights are from Chennai.



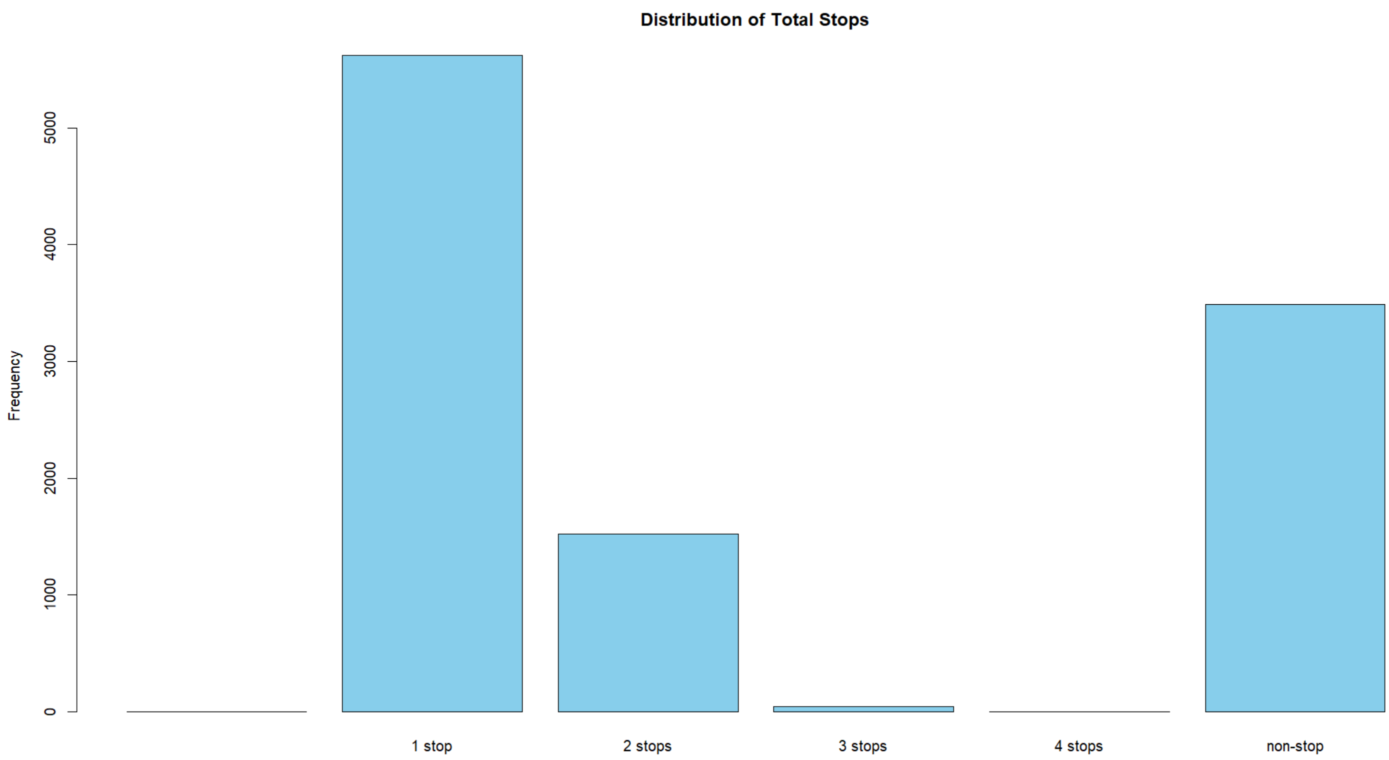
The figure below shows the distribution of Destination. Here, Cochin is seen to be the most selected destination for the flights. This is followed by Banglore and Delhi. Kolkata is the least selected destination for the flights.



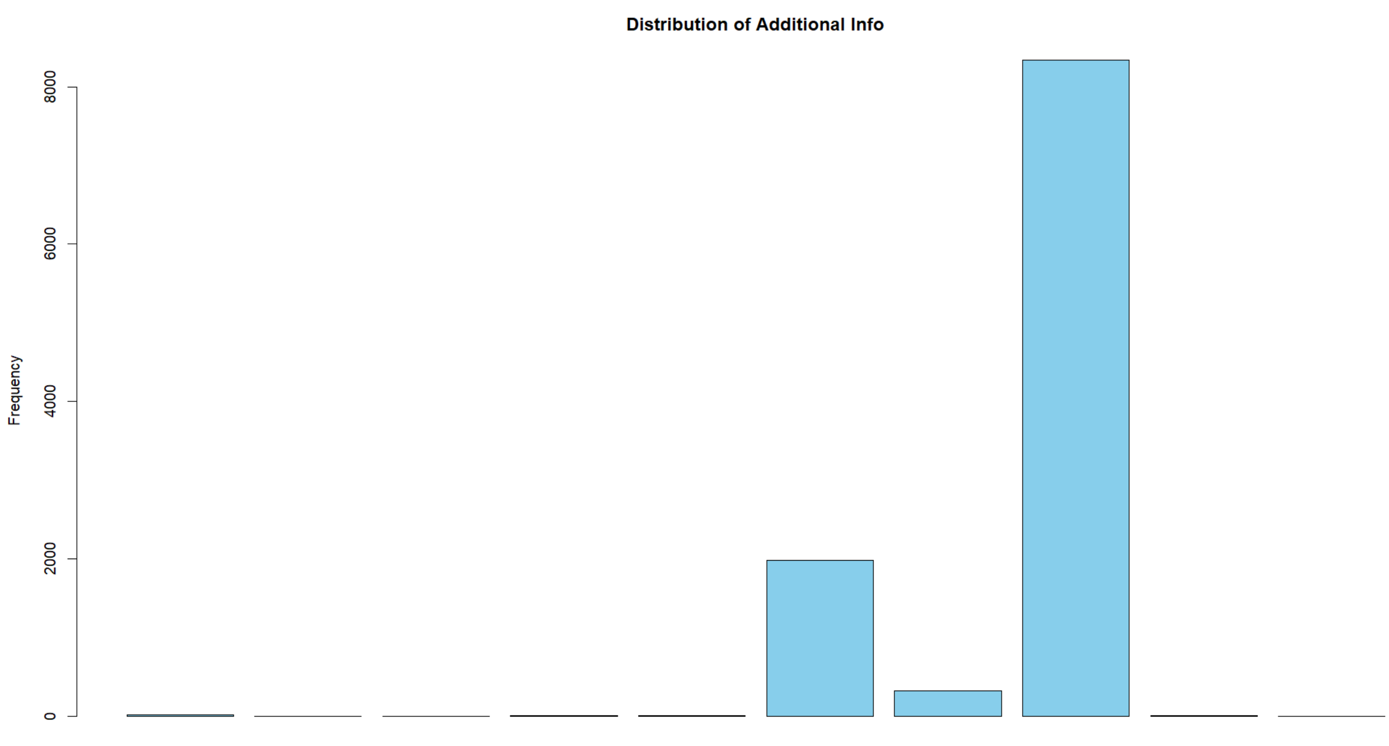
For the routes, there are a total of 129 unique routes available in the dataset.

```
[105] "BLR ? TRV ? COK ? DEL"          "BLR ? IDR ? DEL"
[107] "CCU ? IXZ ? MAA ? BLR"          "CCU ? GAU ? IMF ? DEL ? BLR"
[109] "BOM ? GOI ? PNQ ? HYD"          "BOM ? BLR ? CCU ? BBI ? HYD"
[111] "BOM ? MAA ? HYD"                "BLR ? BOM ? UDR ? DEL"
[113] "BOM ? UDR ? DEL ? HYD"          "BLR ? VGA ? VTZ ? DEL"
[115] "BLR ? HBX ? BOM ? BHO ? DEL"    "CCU ? IXA ? BLR"
[117] "BOM ? RPR ? VTZ ? HYD"          "BLR ? HBX ? BOM ? AMD ? DEL"
[119] "BOM ? IDR ? DEL ? HYD"          "BOM ? BLR ? HYD"
[121] "BLR ? STV ? DEL"                "CCU ? IXB ? DEL ? BLR"
[123] "BOM ? JAI ? DEL ? HYD"          "BOM ? VNS ? DEL ? HYD"
[125] "BLR ? HBX ? BOM ? NAG ? DEL"    ""
[127] "BLR ? BOM ? IXC ? DEL"          "BLR ? CCU ? BBI ? HYD ? VGA ? DEL"
[129] "BOM ? BBI ? HYD"
> len_route <- length(unique_route)
> print(len_route)
[1] 129
```

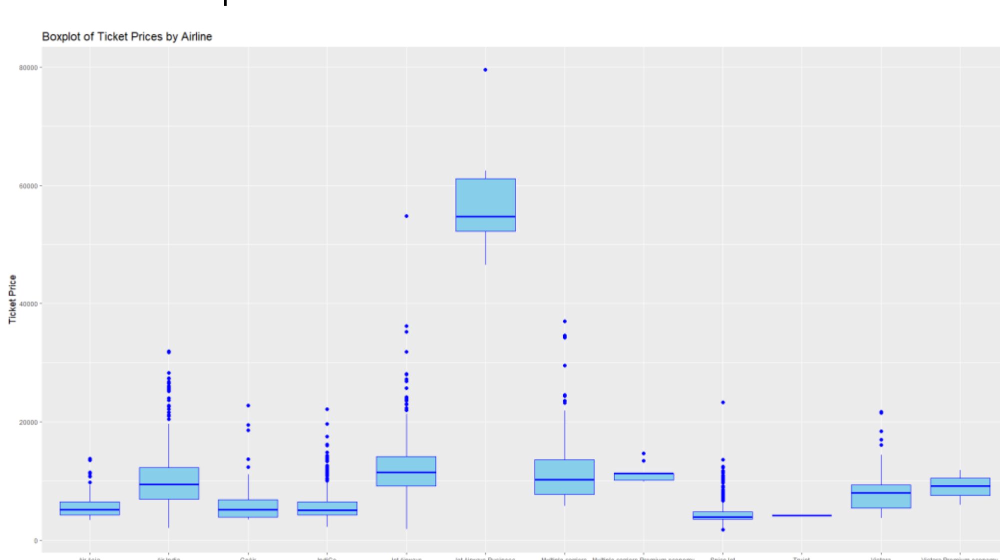
Now we have the distribution of Total Stops. We can see that one stop flights have the highest number. This is followed by non-stop flights. Flights with four and three stops are the least.



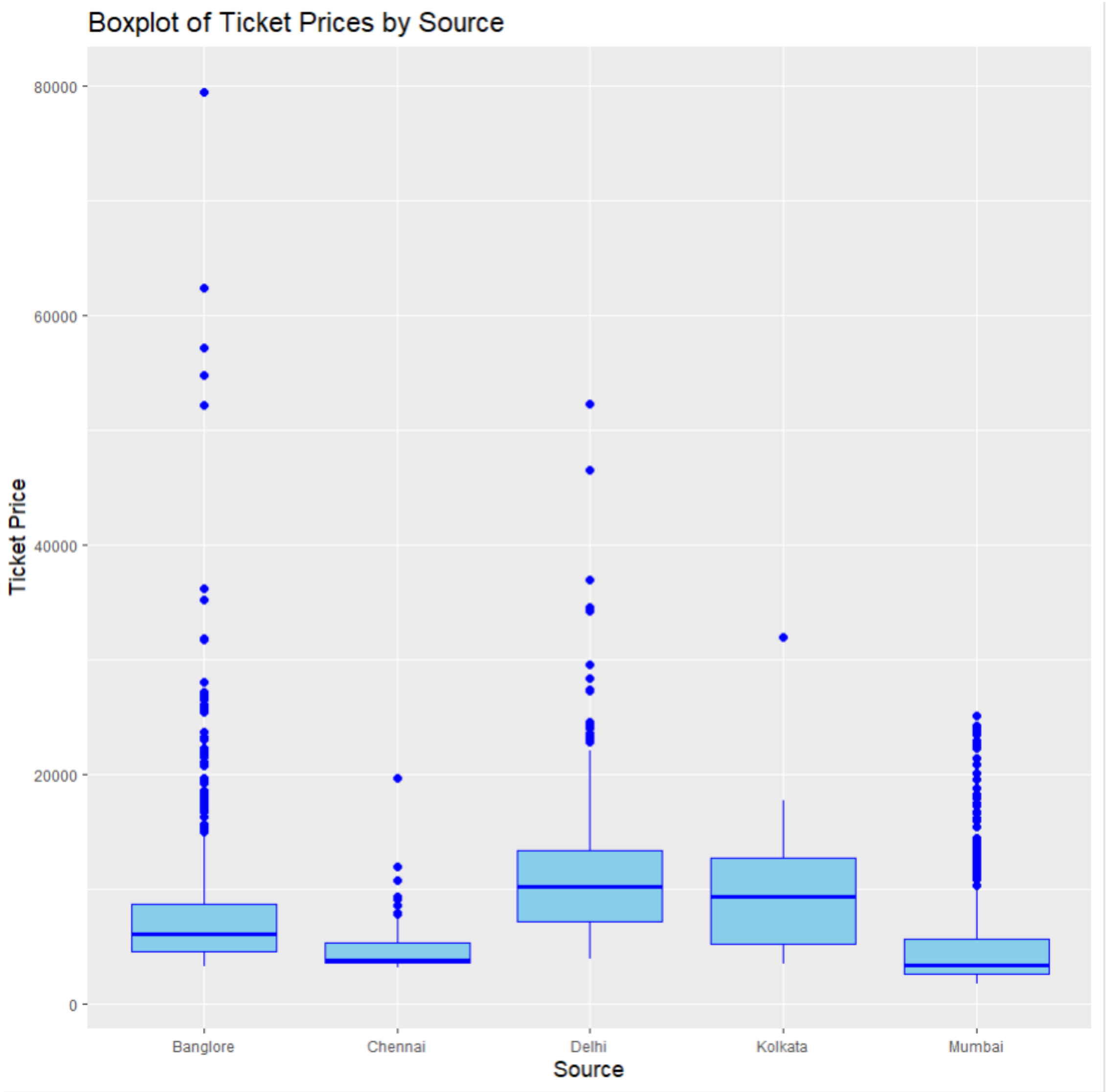
With the Additional Info, most flights do not have additional info. There are a significant number of flights that have no meal included. There are also flights that have no check-in baggage included.



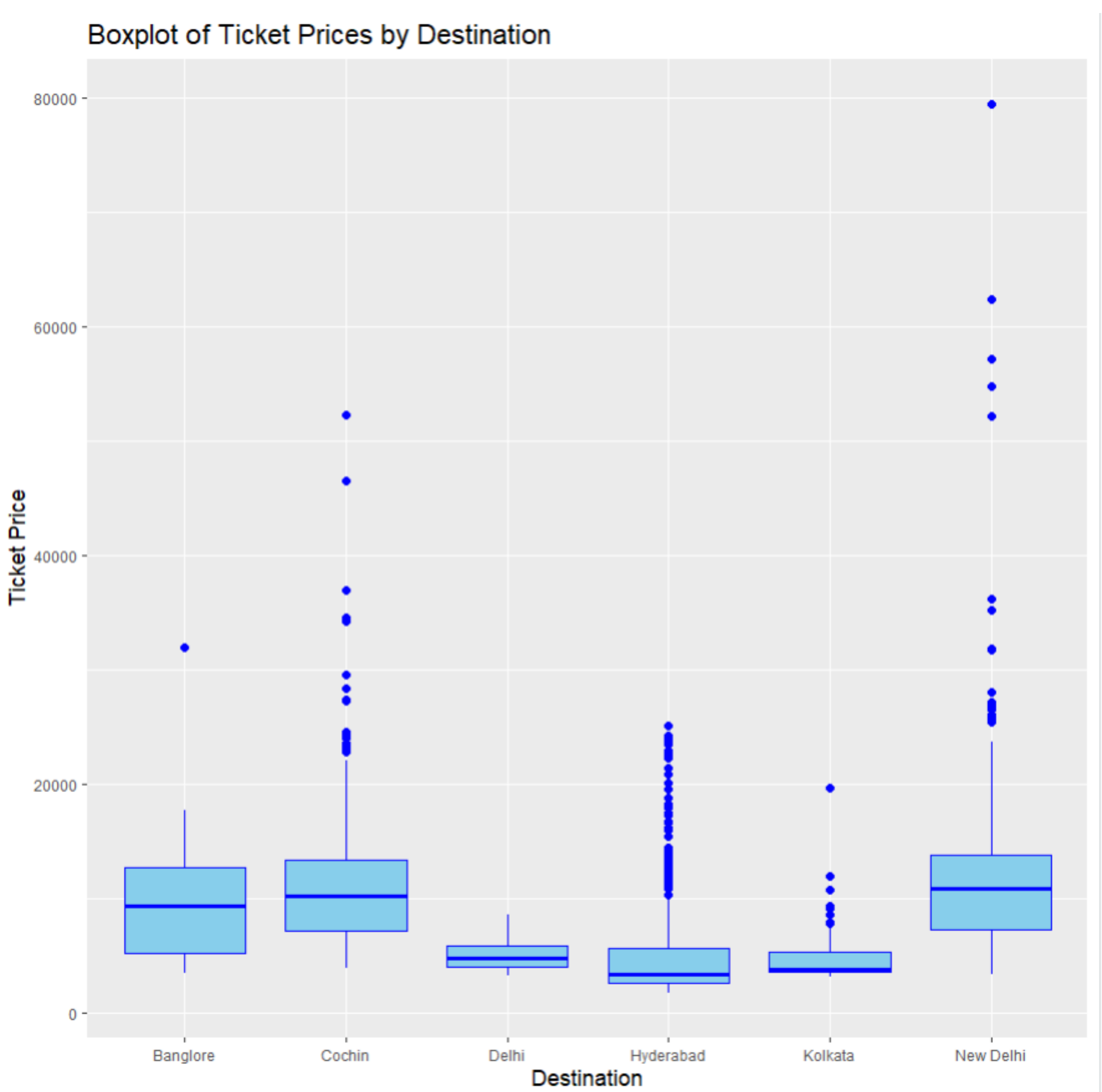
Now we show the boxplot of Prices by Airline. We can observe that there are outliers, most seen with Air India, IndiGo, and Spice Jet. Meanwhile, Trujet and Vistara Premium economy have no outliers, but these also have the lowest frequencies.



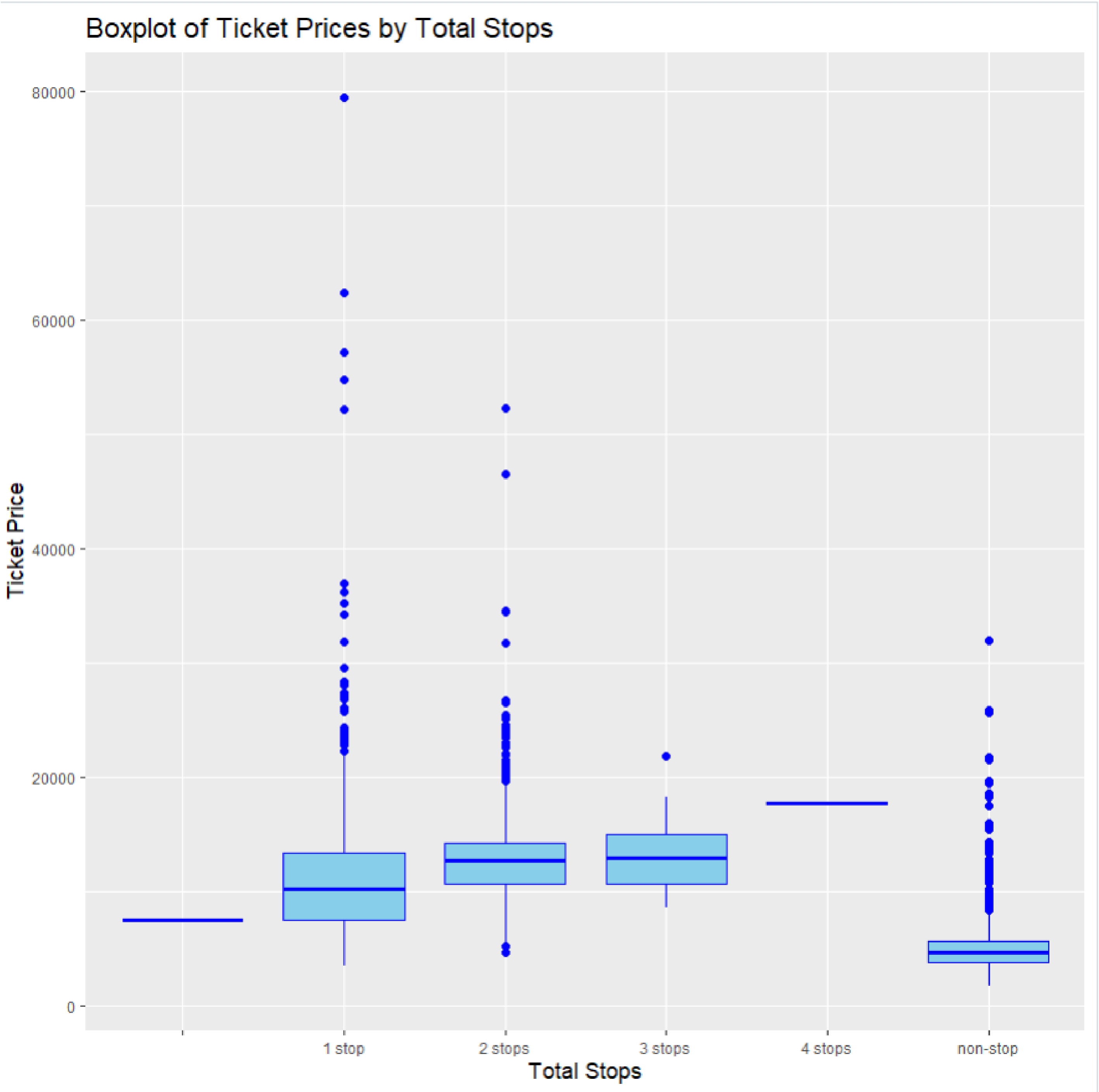
We show the boxplot of Prices by Source. We also observe outliers, where Banglore has the most. Kolkata on the other hand has the least amount of outliers.



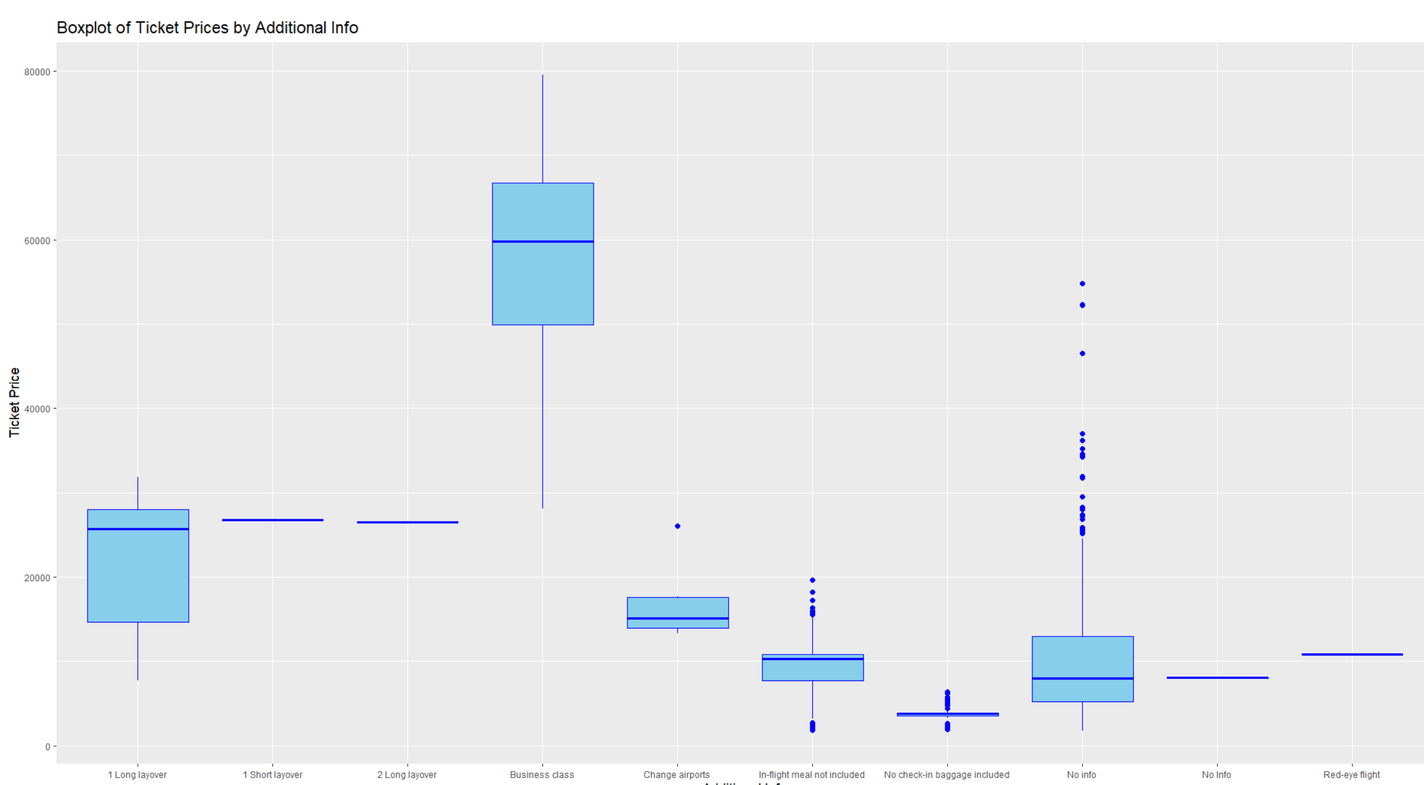
Next we have the boxplot of Prices by Destination. We can observe the most amount of outliers with Hyderabad and New Delhi. Meanwhile, Banglore and Delhi have little to no outliers.



We also have the Boxplot of Prices by Total Stops. We can see that one stop and non stop flights have a significant amount of outliers, but these are also the flights that have the most frequencies.



We show the boxplot of Prices by Additional Info. We can so outliers with No Info flights and no check-in baggage included flights.



Now we prepare our data for modelling. First we check if there are missing values. Additionally, our features are in char datatype. To be able to fit the data into a multiple regression model, we need to convert it to a datatype that the model can understand. We use the as.factor() method to convert all char data into of type factor.

```r
# Check for missing values
print(sum(is.na(flight_data)))

# Convert categorical variables to factors
flight_data$Airline <- as.factor(flight_data$Airline)
flight_data$Source <- as.factor(flight_data$Source)
flight_data$Destination <- as.factor(flight_data$Destination)
flight_data$Total_Stops <- as.factor(flight_data$Total_Stops)
flight_data$Additional_Info <- as.factor(flight_data$Additional_Info)
flight_data$Route <- as.factor(flight_data$Route)
flight_data$Duration <- as.factor(flight_data$Duration)
```

Next, we build a simple multiple regression model. We use Airline, Source, Destination, Total Stops, Additional Info, Route, and Duration as our independent variables. Our dependent variable is Price. We show the summary of the model. We also show the regression graph.

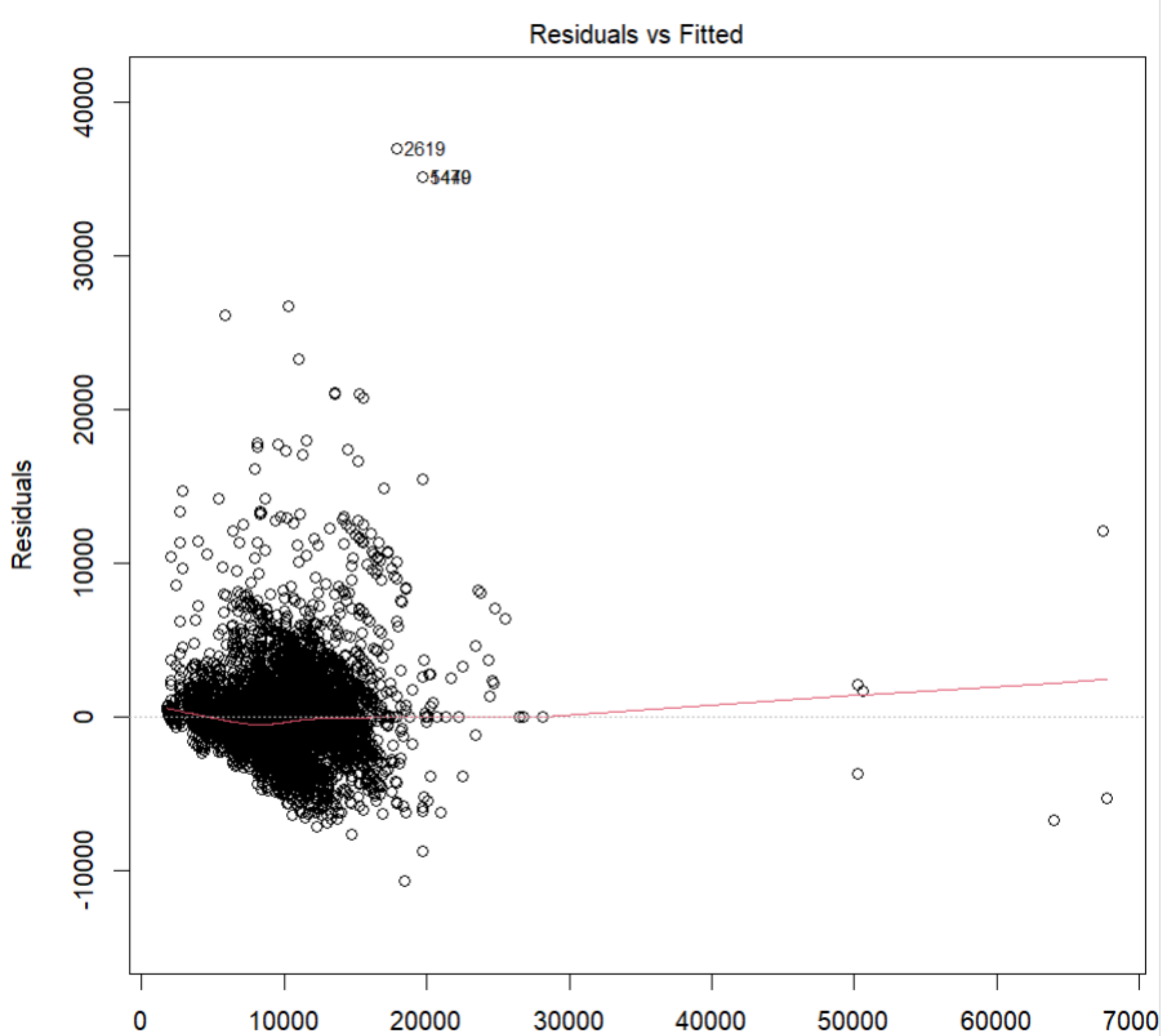
```r
# Build a multiple regression model
model <- lm(Price ~ Airline + Source + Destination + Total_Stops +
              Additional_Info + Route + Duration + Arrival_Time +
              Dep_Time + Date_of_Journey, data = flight_data)

summary(model)
```

Model summary including multiple r-squared score, adjusted r-squared score, and residuals.

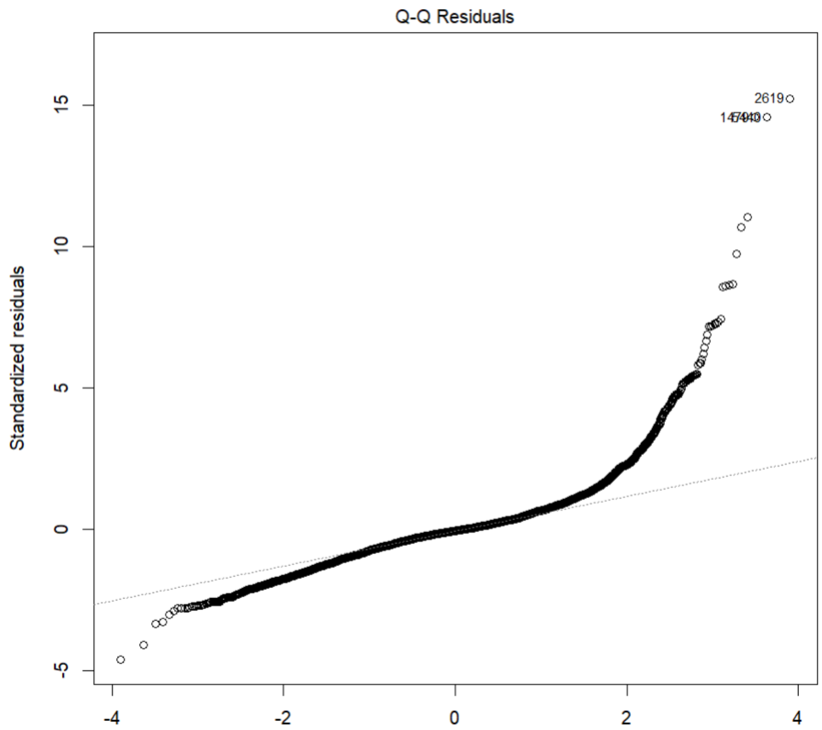
```
Residual standard error: 1459 on 8623 degrees of freedom
Multiple R-squared:  0.9192,    Adjusted R-squared:  0.8999
F-statistic: 47.62 on 2059 and 8623 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9086.7  -486.2     0.0   467.8 28548.1
```

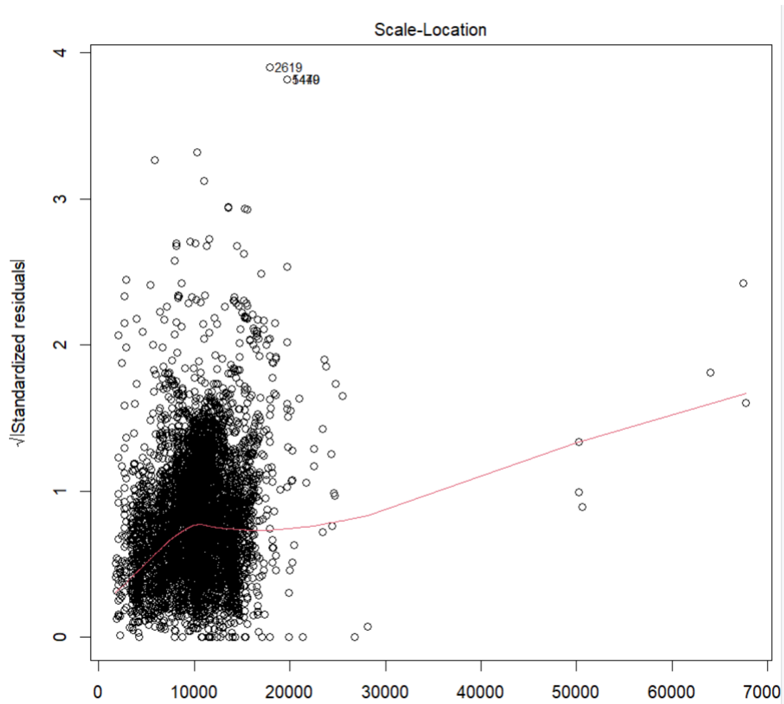
Plot of the model comparing Residuals against Fitted.



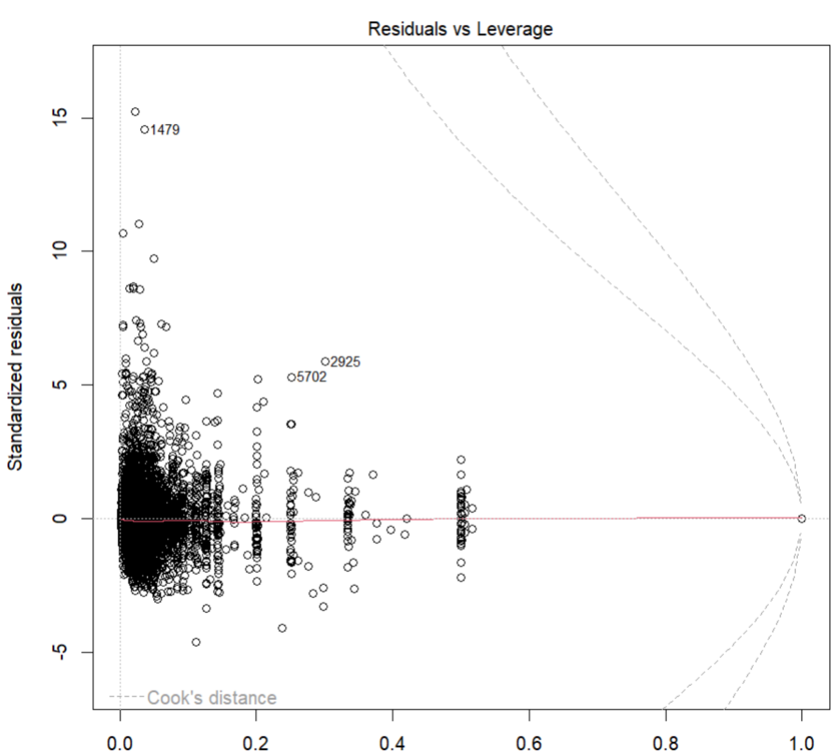
Plot of the model showing Q-Q residuals



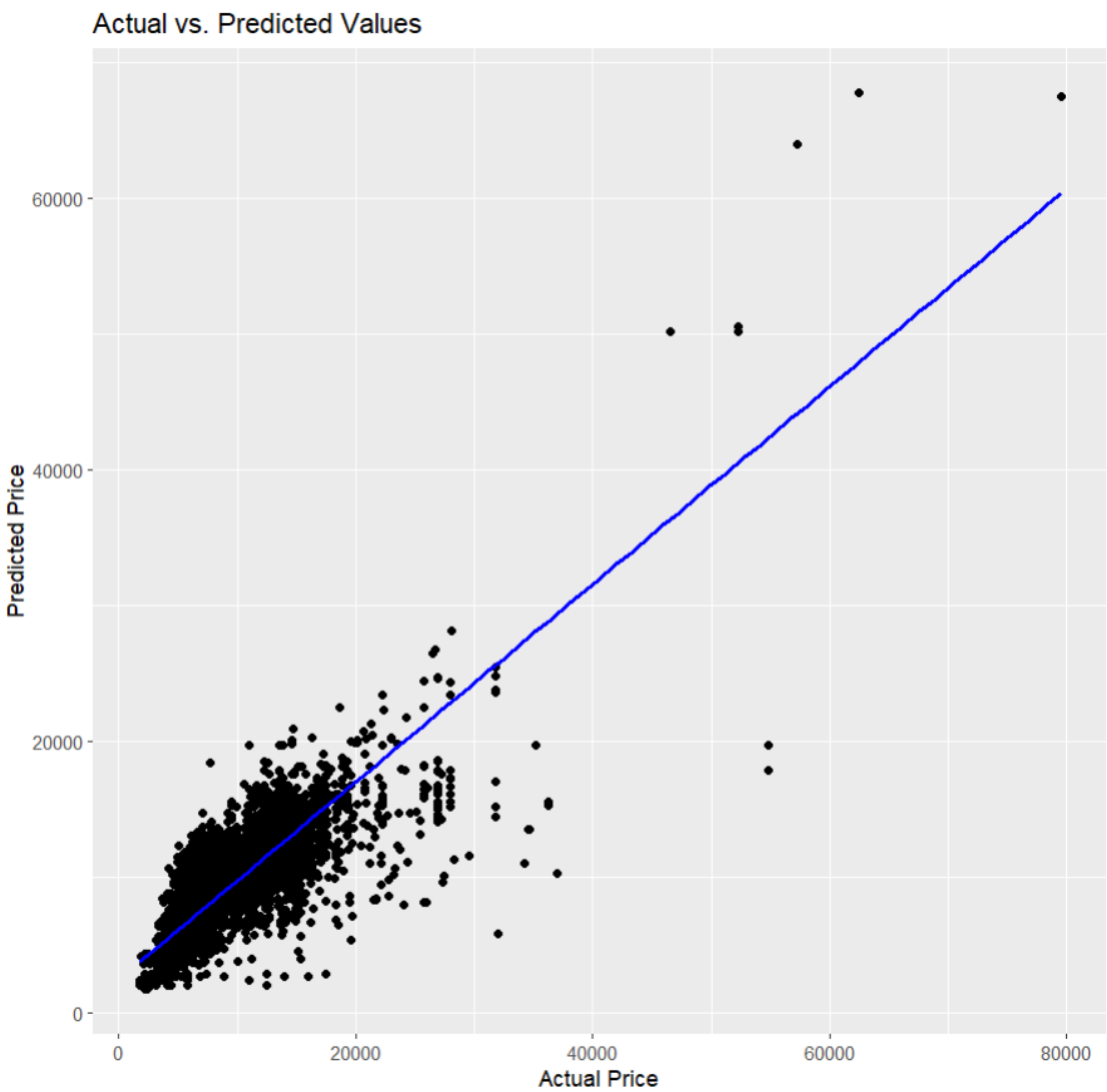
Plot of the model showing scale-location.



Plot of the model showing residuals vs leverage.



Regression Graph showing the actual vs predicted values.



We can analyze the performance of the model using the results we attained. Multiple r-squared and adjusted r-squared is used to identify the model can explain any variance with the data. Residuals are the differences that are identified between the observed and predicted values. Based on the summary of the model, we attain a multiple r-squared score of 0.7275 or 72.75%. Additionally, we attain an adjusted r-squared score of 0.717 or 71.7%. This means that the independent variables explain around 72.75% of the variance in the dependent variable which is the Price. The model does perform quite well in explaining the variability, but there is still room for improvement. Additionally, it also means that around 71.7% of the variance in the dependent variable Price is explained after adjusting the number of predictors. The residuals are also shown with a minimum of -10675, 1st quartile of -1153, median of -134, 3rd quartile of 837, and maximum of 36933. Furthermore, we can observe in the regression graph the prediction of the model compared to its actual price. The graph shows multiple points lie on the regression line, but there are still multiple points not among it, signifying incorrect predictions. We can also observe some outliers whose prediction are quite far from the actual price. We can also observe in the graph that most points are among the price range of 0 to 40000. Thus, the model is capable of predicting the price of flight tickets using the specified independent variables but there is still room for improvement. This could be improved by including more independent variables or by delving deeper into the data values of the chosen independent variables.

REFERENCES

Chee Hua Chew. (2021). *Artificial intelligence, analytics and data science. Volume 1, Core concepts and models*. Cengage Learning Asia Pte Ltd.

freeCodeCamp.org. (2019). R Programming Tutorial - Learn the Basics of Statistical Computing [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=_V8eKsto3Ug

Grolemund, G. (n.d.). A Installing R and RStudio | Hands-On Programming with R. In *rstudio-education.github.io*. https://rstudio-education.github.io/hopr/starting.html

RPubs - Applying linear regression to study flights delay. (n.d.). Rpubs.com. Retrieved March 6, 2024, from https://rpubs.com/salmaeng/linear_regression

Tutorial - Data with R. (n.d.). Www.mit.edu. Retrieved March 6, 2024, from https://www.mit.edu/~amidi/teaching/data-science-tools/tutorial/data-manipulation-with-r/