Fundamentals of Data Analysis

Final project: Inferential Statistics Regarding Flight Prices in India.

Group 5:

Pilar Guerrero

Cloe Chapotot

Niccolò Pragliola

Tessa Correig

Alejandra Gómez

Moritz Goebbels

Miguel Vaquero

May 20, 2023

Abstract

In this study what we have done is analyze the main drivers of flight prices in India, for which we identified some goals. These goals were: Test if the variable price is dependent on the variables source and destination, test the population mean price for flights, study if the airline has an effect on flights with the same source and destination, understand if the airline and number of stops have a relationship, study the relationship between the number of stops and the duration, test the difference in prices between day flights and night flights, study if the variable price depends on the duration of the flight, understand the relationship between the day of the month (variable day) and the price, and finally, test the relationship between the variable month and the price of the flights.

We have obtained the dataset from the website “Ease my Trip” and after doing the data exploration and the data cleaning we started doing inferential statistics and we build a linear regression model.

We have some predictions such as: The average price will increase with higher standards of the classes, fewer stops will increase the price of the flight ticket for long distance flights, the prices for night flights are cheaper than the ones for day flights, etc.

After doing all the study of data some of our main conclusions are:

1. Business classes, as well as the Premium classes tend to have a higher average price.
2. In general fewer stops increase the price, however in this case fewer stops also indicate shorter flight distance, this means that especially Non- Stop flights are far cheaper since the distance is also shorter.
3. Flights at night are relatively cheaper than flights at day.
4. Month and day of the flight affect the price.

Table of Contents

1. Introduction and objectives

1.1. Justification

We are the director board of Namaste Journeys, a travel agency which mediates between airlines and clients in the sale of domestic flights tickets in India. We have been asked to create a report analyzing the main drivers of flight price.

As a matter of fact, the flight industry in India is one of the fastest-growing aviation markets in the world. The industry has seen tremendous growth in recent years, with several new airlines entering the market, increased passenger traffic, and the development of new airports in the main cities which allows for more growth.

Despite the high growth level, the industry has also faced many challenges in the past few years with high taxes and fuel prices, regulatory constraints and infrastructure issues. In order to face these challenges and aid the continued growth of the industry this report aims to give insight into flight prices in India and advise to both perspectives, airlines and clients, regarding flight price behavior and future expectations.

1.2. Objectives and hypotheses

Namaste Journeys is not any ordinary travel agency, we regard as important the improvement of our offerings, especially since this is a highly competitive industry. For this reason, we resort to innovation and take matters into our own hands by conducting an investigation. We have collected data on domestic flights for which we will run different analyses with the aim of identifying how these variables may affect the price of the flights. We are aware that for some of our clients, they are looking for comfortability, while others are looking for affordability when choosing an offer. In addition, we will consider certain myths about traveling to see if they have any factual correlation with the outcomes– such as price.

Having said that, our main objective is: “as a travel agency, to find the best option for our clients given their interests”.

After running certain analysis and models, we aspire to obtain the necessary insights to accomplish our objective and be considered the best air travel agency by our clients. For this, we have identified two main goals:

1) **Our first goal is related to clients traveling low-cost:**

We are a company specialized in advising travelers and we are willing to counsel them to get the flights that best adapt to their needs. For our first goal, a client who wants to travel low-cost, is looking for the cheapest option to fly from x to y, regardless of the duration or stops.

1. Test if the variable price is dependent on the variables Source and Destination.
    - ➢ Do Source and Destination significantly affect the price?

2. Test the population mean price for flights.
    - ➢ What's the average price for the flights per airline?

3. Study if the airline has an effect on flights with the same source and destination.
    - ➢ Does the airline affect the price level for the same flights (departing from the same source and arriving to the same destination)

2) Our second goal is related to high-budget clients :

We are a company specialized in advising travelers and we are willing to counsel them to get the flights that best adapt to their needs. For our second goal, a client who wants to travel comfortably and arrive as soon as possible to the destination, is looking for the best option to fly from x to y, regardless of the price.

1. Understand if the airline and number of stops have a relationship.
    - ➢ Does the airline affect the number of stops? Should we use the type of airline as a block to improve the accuracy of the model?

2. Study the relationship between the number of stops and the duration.
    - ➢ Does the number of stops significantly affect the duration of the flight?

3) Questions for both goals:

1. Test the difference in prices between day flights and night flights.
    - ➢ Is there a significant price difference between flights at night and during the day? Could you estimate it?

2. Study if the variable price depends on the duration of the flight.
    - ➢ Understand the relationship between duration and price?

3. Understand the relationship between the day of the month (variable day) and the price.
    - ➢ Does the date of the month (1-30) significantly affect price? Should we use the variable Day as a block?

4. Test the relationship between the variable month and the price of the flights.
    - ➢ Does the month have a significant relationship with price? Should we use the variable Month as a block?

1.3. Population of interest

Our population of interest are the flights that took off from India between March and June 2019. We are going to study it using a sample of around 10700 flights that has been randomly selected from this population.

2. Data Exploration

2.1. Dataset Description

To conduct this study we have gathered data about around 10700 flights, which has been extracted from the web page of "Ease my trip", an online travel agency in India that provides hotel bookings, air tickets, holiday packages, bus bookings, and white-label services. We are going to be focused only on the flight service.

In the initial dataset we had the following variables:

1. **Airlines:** Air Asia, Air India, GoAir, IndiGo, Jet Airways, Jet Airways Business, Multiple carriers, Multiple carriers Premium economy, SpiceJet, Trujet, Vistara, Vistara Premium economy (12 levels).
2. **Date of journey:** in day, month, and year. All of the flights were from 2019.
3. **Source:** Flight departure places among India.
4. **Destination:** Flight arrival places among India.
5. **Route:** With the departure and arrival place of the flight.
6. **Departure time:** in hours and minutes (24 hours clock).
7. **Arrival time:** in hours and minutes (24 hours clock).
8. **Duration:** in hours and minutes.
9. **Total stops:** number of stops. It was a categorical variable with 4 levels (non-stop, 1 stop, 2 stops, and 3 stops).
10. **Additional information:** It was a categorical variable with 9 levels (red-eye flight, no info, no check-in baggage included, in-flight meal not included, change airports, business class, 2 long layover, 1 short layover, and 1 long layover).
11. **Price:** In Rupias.

We have cleaned the dataset by using RStudio and Excel: firstly, we have eliminated the column Route since we already have that information (in the columns source and destination) and additional information since most of the flights didn't have extra information. Moreover, since we have the duration and the departure time, we have deleted the arrival one. Then we have divided the column date of journey into columns: one for the day and another for the month (we haven't included the year since all flights occurred in 2019). We have classified the different departure times into day and night in order to have a categorical independent variable with 2 levels. Furthermore, we have converted the duration to minutes (it was in hours and minutes) and, finally, we have made the stop's categorical column to a numerical one by changing the 4 different levels by their corresponding number. After that, we have a dataset that contains the following information:

1. **Airlines:** a categorical variable with the same 12 levels than at the beginning.
2. **Day** of the flight.
3. **Month:** with their respective number.
4. **Source:** flight departure places among India.
5. **Destination:** flight arrival places among India.

6. **Time:** a categorical variable with 2 levels: day and night.
7. **Duration** of the flight: in minutes.
8. **Price** of the flight in Rupias.
9. **Stops:** number of stops.

We will assume there are no more variables that can alter the results. This will help us to reduce the variability of the output and make them more accurate and reliable. Moreover, since our sample size is greater than 30, we will assume it follows a normal distribution. Furthermore, the assumption that the sample has been randomly selected in an independent manner from the population must also be done.

2.2. Data exploration: Graphical analysis and Numerical analysis

To start with, we have performed a data exploration of our data set in order to arrive at a better understanding of it, both numerically and graphically in order to be able to guide our future inferential tests.

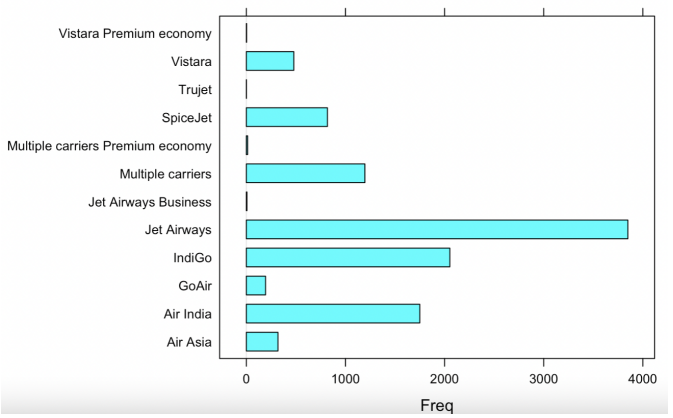
The str() function gave us a glimpse into the overall characteristics of our dataset. We had to alter some of the data types such as some which were initially classified as characters into factors.
Moreover, we also chose to change the names of the variables to ease the investigation, we did this using the command colnames().

```
> str(fda)
tibble [10,681 × 10] (S3: tbl_df/tbl/data.frame)
 $ Airline    : Factor w/ 12 levels "Air Asia","Air India",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ day        : Factor w/ 10 levels "1","3","6","9",..: 2 2 6 8 5 10 10 5 4 9 ...
 $ month      : Factor w/ 4 levels "3","4","5","6": 3 2 4 4 2 2 3 3 4 4 ...
 $ Source     : Factor w/ 5 levels "Banglore","Chennai",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Destination: Factor w/ 6 levels "Banglore","Cochin",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Time       : Factor w/ 2 levels "Day","Night": 2 2 2 2 2 2 2 2 2 2 ...
 $ Duration   : num [1:10681] 170 170 170 170 170 170 170 170 170 170 ...
 $ Price      : num [1:10681] 6181 4282 4282 4282 4282 ...
 $ Total_Stops: Factor w/ 5 levels "1 stop","2 stops",..: 5 5 5 5 5 1 1 1 1 1 ...
 $ STOPS      : num [1:10681] 0 0 0 0 0 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:2] 1554 1793
  ..- attr(*, "names")= chr [1:2] "1554" "1793"
```

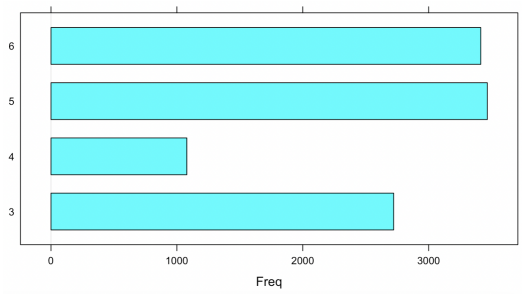
2.21 Data Exploration: Categorical Values

The exploration has been followed by the analysis of categorical variables using the describe () function and the barchart() function.

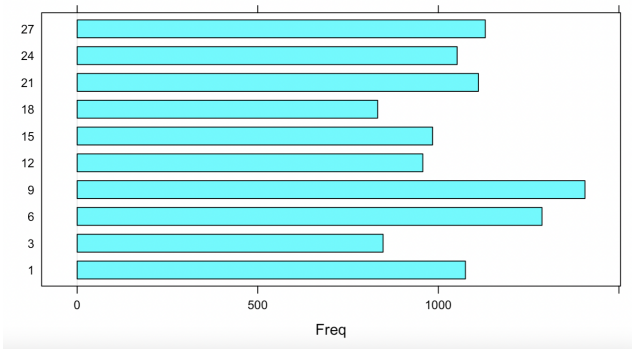
```
> fda$Airline = factor(fda$Airline)
> table(fda$Airline)

                     Air Asia                     Air India
                          319                          1750
                        GoAir                       IndiGo
                          194                          2053
                  Jet Airways          Jet Airways Business
                         3849                             6
           Multiple carriers Multiple carriers Premium economy
                         1196                            13
                     SpiceJet                       Trujet
                          818                             1
                      Vistara      Vistara Premium economy
                          479                             3
> options(scipen = 999)
> prop.table(table(fda$Airline))

                     Air Asia                     Air India
            0.02986611740474             0.16384233685984
                        GoAir                       IndiGo
            0.01816309334332             0.19221046718472
                  Jet Airways          Jet Airways Business
            0.36035951689917             0.00056174515495
           Multiple carriers Multiple carriers Premium economy
            0.11197453421964             0.00121711450239
                     SpiceJet                       Trujet
            0.07658458945792             0.00009362419249
                      Vistara      Vistara Premium economy
            0.04484598820335             0.00028087257747
> barchart(fda$Airline)
```



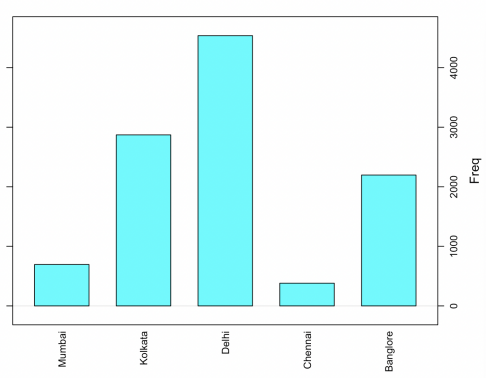

```
> fda$month = factor(fda$month)
> table(fda$month)

   3    4    5    6
2722 1079 3466 3414
> prop.table(table(fda$month))

           3            4            5            6
0.2548450520 0.1010205037 0.3245014512 0.3196329932
> barchart(fda$month)
```



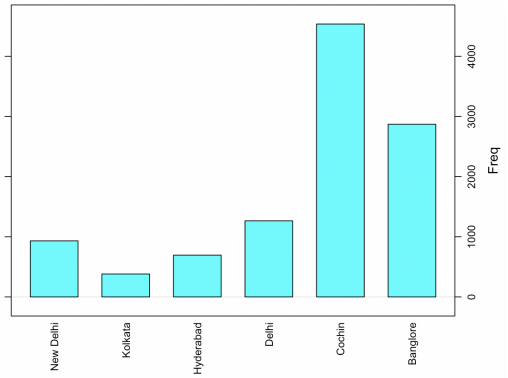

```
> fda$day = factor(fda$day)
> table(fda$day)

   1    3    6    9   12   15   18   21   24   27
1075  847 1287 1406  957  984  832 1111 1052 1130
> prop.table(table(fda$day))

            1            3            6            9           12
0.10064600693 0.07929969104 0.12049433574 0.13163561464 0.08959835221
           15           18           21           24           27
0.09212620541 0.07789532815 0.10401647786 0.09849265050 0.10579533752
> barchart(fda$day)
```



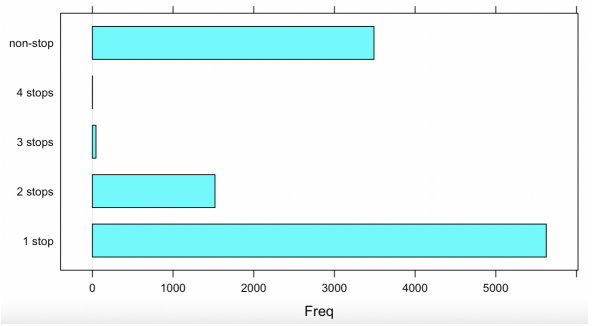
```
> fda$Source = factor(fda$Source )
> table(fda$Source)

Banglore  Chennai    Delhi  Kolkata   Mumbai
    2197      381     4537     2871      695
> prop.table(table(fda$Source ))

     Banglore       Chennai         Delhi       Kolkata        Mumbai
0.20569235090 0.03567081734 0.42477296133 0.26879505664 0.06506881378
> barchart(fda$Source )
```




```
> fda$Destination = factor(fda$Destination )
> table(fda$Destination)

 Banglore    Cochin     Delhi Hyderabad   Kolkata New Delhi
     2871      4537      1265       695       381       932
> prop.table(table(fda$Destination ))

     Banglore        Cochin         Delhi     Hyderabad       Kolkata     New Delhi
0.26879505664 0.42477296133 0.11843460350 0.06506881378 0.03567081734 0.08725774740
> barchart(fda$Destination )
```




```
> fda$Total_Stops = factor(fda$Total_Stops)
> table(fda$Total_Stops)

  1 stop  2 stops  3 stops  4 stops non-stop
    5625     1520       45        1     3490
> prop.table(table(fda$Total_Stops))

         1 stop          2 stops          3 stops          4 stops
0.52663608276379 0.14230877258684 0.00421308866211 0.00009362419249
        non-stop
0.32674843179478
> barchart(fda$Total_Stops)
```



As a general overview, we noticed some inconveniences:

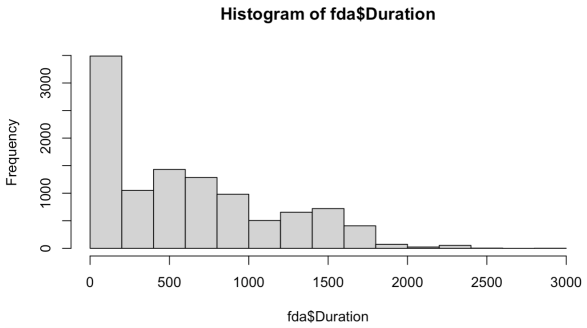
- there's some airlines for which we have very little information
- flights seem to always be on days which are multiples of three
- we only have data for march, april, june and july
- Source and destination cities are also not proportional.

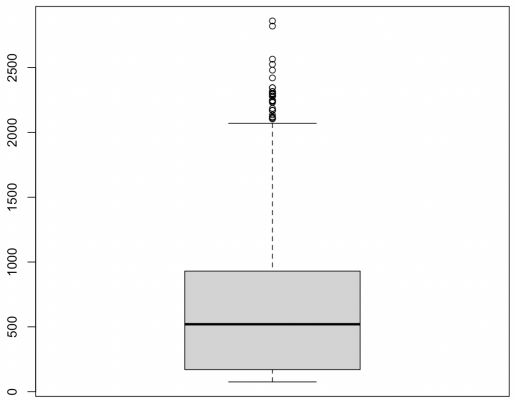
Nevertheless we still have lots of relevant data with around 10681 observations so we still think we can make meaningful inferences. Ofcourse, we will only make predictions regarding data in our sample range.

2.22 Data Exploration: Numerical Values

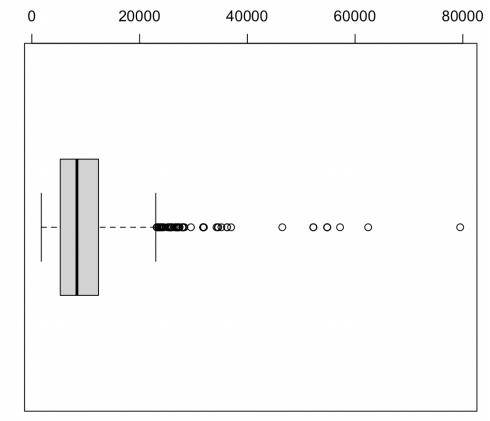
Next we have also used the describe() function, hist() function and boxplot() Function in order to analyze the numerical variables

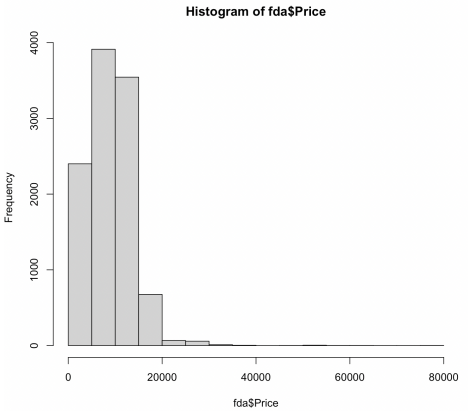
```
> hist(fda$Duration)
> boxplot(fda$Duration)
> outliers = boxplot(fda$Duration)$out
> outliers
 [1] 2295 2295 2295 2295 2295 2295 2295 2295 2295 2295 2295 2280 2280 2280 2280 2280 2280 2280 2280 2280
[21] 2280 2280 2280 2280 2420 2240 2240 2240 2240 2240 2480 2240 2245 2245 2120 2185 2170 2170 2170 2170
[41] 2170 2170 2315 2345 2315 2345 2245 2105 2105 2105 2115 2115 2315 2315 2245 2820 2300 2300 2300 2300
[61] 2300 2245 2245 2295 2240 2240 2230 2245 2245 2565 2525 2135 2860
> which(fda$Duration %in% outliers)
 [1]  339  341  342  343  344  345  346  350  352  353  354  392  393  394  395  397  398  404  407  409
[21]  410  411  412  416  483  490  492  493  494  497  510  514  519  522  602  615  688  693  695  696
[41]  701  704  775  777  779  785  952 1114 1116 1118 1169 1207 4441 4442 4446 4447 4448 4451 4455 4457
[61] 4458 4460 4461 4486 4530 4542 4594 4710 4718 5622 5751 6619 7200
>
```






```
> boxplot(fda$Price)
> hist(fda$Price)
> outliers = boxplot(fda$Price)$out
> outliers
 [1] 28322 23677 26480 26743 31783 25430 25430 25139 27282 24017 26092 25913 31945 25703 27430 26890
[17] 27992 26890 27992 26890 27992 28097 23528 26890 26890 27992 26890 26890 26890 27992 26890 27992
[33] 26890 23843 26890 26890 26890 26890 27992 26890 26890 25735 26890 27992 25735 25735 36235 36235
[49] 26890 27992 25735 24115 26890 26890 25735 25735 54826 26890 54826 35185 31825 31825 24210 31825
[65] 27210 25735 26890 54826 25735 31825 31825 25735 26890 26890 31825 25735 27210 52229 79512 62427
[81] 57209 46490 52285 29528 23170 36983 34273 34503 34608 24318 23583 24528 23533 23267
> which(fda$Price %in% outliers)
 [1]  337  610  611  764  765  768  775  903  939  943 1157 1588 1928 2005 2063 4599 4600 4616 4626 4639
[21] 4653 4660 4834 4886 4892 4902 4905 4906 4915 4916 4917 5079 5116 5148 5329 5332 5351 5357 5358 5365
[41] 5931 5936 5938 5947 5962 5966 6194 6243 6275 6291 6293 6329 6625 6629 6630 6639 6645 6652 6936 6937
[61] 6940 6944 7451 7715 7909 8006 8007 8024 8028 8031 8043 8051 8053 8067 8069 8070 8084 8166 8167 8168
[81] 8169 8170 8171 8174 8175 9032 9056 9219 9220 9267 9299 9307 9343 9811
```





As a general insight we can see how both of our numerical variables are skewed instead of following a normal distribution. This may be an inconvenience but we are guessing it is the nature of our data given that prices and duration of flights are generally within a lower range but there's some observations with higher values. We can also observe some outliers in the data which we will manage in following steps of the project

2.23 Data Exploration: Missing values

According to our results we don't have any important missing values for these variables so no need to recode data, we will just omit the one observation for the missing values in total stops.

```
> which(is.na(fda$Airline))
integer(0)
> which(is.na(fda$day))
integer(0)
> which(is.na(fda$month))
integer(0)
> which(is.na(fda$Source))
integer(0)
> which(is.na(fda$Destination))
integer(0)
> which(is.na(fda$Price))
integer(0)
> which(is.na(fda$Total_Stops))
[1] 1793
```

2.24 Data Exploration: Outliers

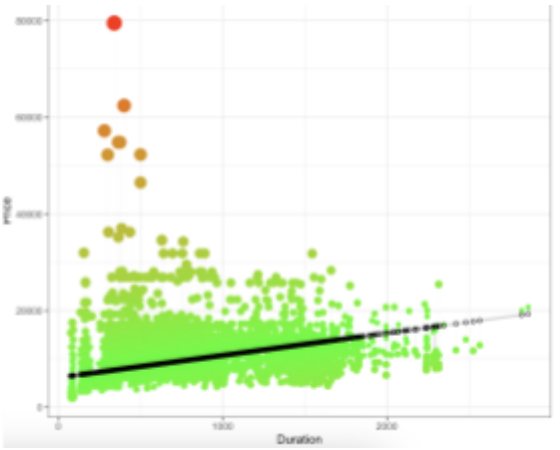
According to the plots above, we have different outliers (not only in terms of duration, but also in terms of price). After dueing with the NA values, we can confront this problem of outliers. To solve it, we have two options:

- The first one would be to exclude outliers. However, if we choose this option, we must take into consideration that there is a risk of deleting many valuable observations.
- The second option would be to recode outliers. Be that as it may, this is not very rigorous.

Having taken into account the previous disadvantages of the two options, we have decided to exclude outliers and do the whole statistical analysis with the two datasets (one which includes the outliers and the other within them), and then compare them. After doing the residual analysis of the model with the outliers and without them, we can see that the difference of the residual standard error is not too high so this supports our decision of deleting the outliers.

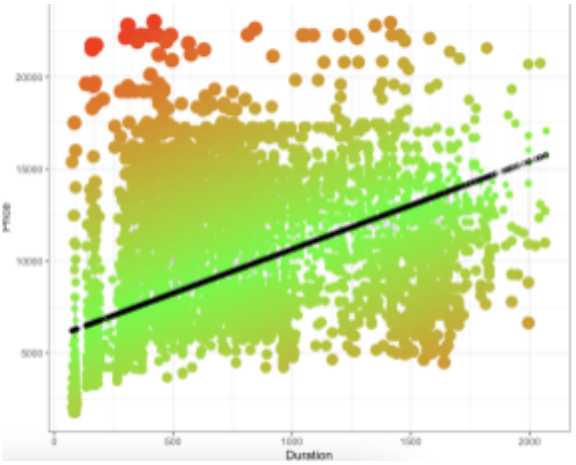
Study of duration and price with outliers:

```
Call:
lm(formula = Price ~ Duration, data = df)

Residuals:
   Min     1Q Median     3Q    Max
 -9186  -2520   -974   1719  71821

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.126e+03  6.207e+01   98.70   <2e-16 ***
Duration    4.601e+00  7.574e-02   60.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3975 on 10679 degrees of freedom
Multiple R-squared:  0.2568,    Adjusted R-squared:  0.2568
F-statistic:  3690 on 1 and 10679 DF,  p-value: < 2.2e-16
```



Study of duration and prices within outliers:

```
Call:
lm(formula = Price ~ Duration, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-9205.7 -2356.2  -865.3  1774.6 15452.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.861e+03  5.264e+01  111.33   <2e-16 ***
Duration    4.776e+00  6.583e-02   72.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3318 on 10513 degrees of freedom
Multiple R-squared:  0.3336,    Adjusted R-squared:  0.3335
F-statistic:  5262 on 1 and 10513 DF,  p-value: < 2.2e-16
```



As we can see from the outputs, the median of the residuals doesn't differ a lot. Moreover, we cannot see a big difference between the different residual standard error (being lower when we remove outliers). As a result, we conclude that we can delete outliers without having big changes in our outputs and conclusions because of that.

3. Inferential Statistics

3.1 Average price for flights per airline and their class:

Prediction:
Without prior knowledge we cannot say anything about the differences between airlines, however we would expect the average price to increase with higher standards of the classes.

R-Code:

```
df %>% group_by(Airline) %>% summarize(a = mean(Price)) %>% arrange(desc(a))
```

Output:

	Airline	a
	<chr>	<dbl>
1	Jet Airways Business	58359.
2	Jet Airways	11644.
3	Multiple carriers Premium economy	11419.
4	Multiple carriers	10903.
5	Air India	9610.
6	Vistara Premium economy	8962.
7	Vistara	7796.
8	GoAir	5861.
9	IndiGo	5674.
0	Air Asia	5590.
1	SpiceJet	4338.
2	Trujet	4140

Results:
As expected, Business classes, as well as the Premium classes tend to have a higher average price. Further, even more important, and clear to observe is that Jet Airways, followed by Multiple carriers and Air India tend to have the highest average prices, whereas Trujet and SpiceJet have the lowest.

3.1.5 Understanding the relationship between the Airlines and the price

Considering the means of the Airlines it is important to see if there are significant differences between them:

Hypothesis:

H0: μ1 =μ2 =...= μk
H_a : at least two means differ

R-Code:

```
df$Airline = factor(df$Airline)
anovatest = aov(Price ~ Airline, data = df)
anovasummary = summary (anovatest)
anovasummary
```

Output:

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
Airline        11 9.333e+10 8.485e+09   676.5 <2e-16 ***
Residuals   10670 1.338e+11 1.254e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results:

Considering the obtained p-value, we found out that there is a significant price difference between the airlines, meaning at least two means differ. This is caused by the different airlines and their classes targeting different customers and offering them either lower costs or higher comfort and support.

3.2. Understanding the relationship between the number of stops and the price

How does the number of stops affect the price per ticket?
→ We believe that fewer stops will increase the price of the flight ticket for long distance flights.

Hypothesis:

H0: μ1 =μ2 =...= μk
H_a : at least two means differ

R-Code:

```
df$Total_Stops = factor(df$Total_Stops)
anovatest = aov(Price ~ Total_Stops, data = df)
anovasummary = summary (anovatest)
anovasummary
```

Output:

```
              Df    Sum Sq   Mean Sq F value Pr(>F)
Total_Stops    4 2.034e+10 5.085e+09   262.5 <2e-16 ***
Residuals  10677 2.068e+11 1.937e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results:

From the output we obtain a p-value lower than 0.05, which means at least two means differ, meaning that the number of stops affect the price of the flight. To further analyze these differences in means we will proceed by undertaking a Multiple Comparison of Means using a "PostHocTest".

R- Code:

```
tukey = PostHocTest(anovatest, method = "hsd", conf.level = 0.95)
tukey
```

Output:

```
  Posthoc multiple comparisons of means : Tukey HSD
    95% family-wise confidence level

$Total_Stops
                        diff       lwr.ci     upr.ci    pval
2 stops-1 stop     645.7683    298.6661   992.8705 3.9e-06 ***
3 stops-1 stop     160.8916  -1636.1714  1957.9545  0.9992
4 stops-1 stop    -455.9751 -12464.1742 11552.2240  1.0000
non-stop-1 stop  -2768.0559  -3026.7611 -2509.3506 < 2e-16 ***
3 stops-2 stops   -484.8768  -2301.0965  1331.3430  0.9500
4 stops-2 stops  -1101.7434 -13112.8243 10909.3375  0.9991
non-stop-2 stops -3413.8242  -3782.8058 -3044.8426 < 2e-16 ***
4 stops-3 stops   -616.8667 -12756.6780 11522.9447  0.9999
non-stop-3 stops -2928.9474  -4730.3643 -1127.5306 9.1e-05 ***
non-stop-4 stops -2312.0808 -14320.9322  9696.7707  0.9848

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- ➢ **μ2 > μ1**
- ➢ Between μ3 and μ1 there is no significant difference, but since the middle of the interval is positive we can assume **μ3 > μ1**.
- ➢ Between μ4 and μ1 there is no significant difference, but since the middle of the interval is negative we can assume **μ1 > μ4**.
- ➢ **μ1 > μNon**
- ➢ Between μ3 and μ2 there is no significant difference, but since the middle of the interval is negative we can assume **μ2 > μ3**.
- ➢ Between μ4 and μ2 there is no significant difference, but since the middle of the interval is negative we can assume **μ2 > μ4**.
- ➢ **μ2 > μNon**
- ➢ Between μ4 and μ3 there is no significant difference, but since the middle of the interval is negative we can assume **μ3 > μ4**.
- ➢ **μ3 > μNon**
- ➢ Between μ4 and μNon there is no significant difference, but since the middle of the interval is negative we can assume **μ4 > μNon**.

Overall Results: μ2 > μ3 > μ1 > μ4 > μNon

Explanation:
In general fewer stops increase the price, however in this case fewer stops also indicate shorter flight distance, this means that especially Non- Stop flights are far cheaper since the distance is also shorter. Overall we can assume that 2-stop flights are the most expensive for long distance flights.

3.3. Understanding the relationship of prices for flights at day and night

Is there a significant price difference between the flights at night and during the day?
→ We expect the prices for night flights to be cheaper than the ones for day flights.

Hypothesis:

H0: μDay - μNight = 0
H_a : μ(Day) - μ(NIGHT) ≠ 0

R-Code:

```
t.test(Price ~ Time, data = df, var.equal = TRUE)
```

Output:

```
        Two Sample t-test

data:  Price by Time
t = 3.9629, df = 10680, p-value = 7.453e-05
alternative hypothesis: true difference in means between group Day and group Night is not equal to 0
95 percent confidence interval:
 247.5824 732.2334
sample estimates:
  mean in group Day mean in group Night
           9162.175            8672.267
```

Results:
From the output we obtain a p-value lower than 0.05, which means that there is a significant difference in the prices for flights at day and at night. In context this means flights at night are relatively cheaper with an average of 8672.27 than flights at day with an average of 9162.18.

3.4. Understanding the effect of Source and Destination on Price

Do Source and Destination significantly affect the price?
→ We believe there will be some sources and destinations which will increase the price more relatively compared to others.

R-Code:

```r
df$Source <- as.factor(df$Source)
df$Destination <- as.factor(df$Destination)
model <- lm(Price ~ Source + Destination, data = df)
summary(model)
```

Output:

```
Call:
lm(formula = Price ~ Source + Destination, data = df)

Residuals:
   Min     1Q Median     3Q    Max
 -8535  -3001   -321   2309  67594

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           11917.7      131.5  90.631   <2e-16 ***
SourceChennai         -7127.8      244.1 -29.199   <2e-16 ***
SourceDelhi           -1378.3      144.4  -9.547   <2e-16 ***
SourceKolkata         -2759.3      151.3 -18.232   <2e-16 ***
SourceMumbai          -6866.8      201.1 -34.144   <2e-16 ***
DestinationCochin          NA         NA      NA       NA
DestinationDelhi      -6773.8      173.3 -39.088   <2e-16 ***
DestinationHyderabad       NA         NA      NA       NA
DestinationKolkata         NA         NA      NA       NA
DestinationNew Delhi       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4014 on 10676 degrees of freedom
Multiple R-squared:  0.2426,    Adjusted R-squared:  0.2422
F-statistic: 683.8 on 5 and 10676 DF,  p-value: < 2.2e-16
```

Results:
From the output we can obtain that all the sources have a significant impact on the price of the flights, since the p-values are all significantly lower than 0.05. In this case Bangalore is taken as the base level and is included in the intercept, which means for example if a flight is not from Bangalore but rather from Chennai the price decreases by 7127. For the effect of the Destinations, we can just obtain the p-value for Delhi, which has a significant impact on the price. For the other Destinations we obtain NA values, which means we cannot interpret these results later. However, for Kolkata we could assume it has a significant impact on the price, since Kolkata Source is significant.

3.5. Understanding relationship between the price of a flight and its duration

Does flight duration significantly affect price?
→ Reasoning from a logical point of view, we expect flights to become more expensive as their duration increases but most importantly we expect the mean price of the flights to remain constant across all the levels of the Duration.

Hypothesis:

H0: $\mu 1 = \mu 2 = ... = \mu k$
H_a : at least two means differ

R-Code:

```
df$Duration = factor(df$Duration)
anovatest = aov(Price ~ Duration, data = df)
anovasummary1 = summary (anovatest)
```

Output:

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
Duration      366 1.141e+11 311843280   28.48 <2e-16 ***
Residuals   10314 1.129e+11  10950346
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

Results:
From this ANOVA test, we are able to conclude that we must reject the null hypothesis that the mean price of the flights is constant throughout all the levels of Duration since the p-value indicates that there is a statistically significant difference between Duration levels.

3.6. Understanding whether the day of the month affects Price.

Does the day of the month affect the Price?
→ Our null hypothesis is that the mean price of the flights will remain constant throughout the month, and won't be affected by the day.

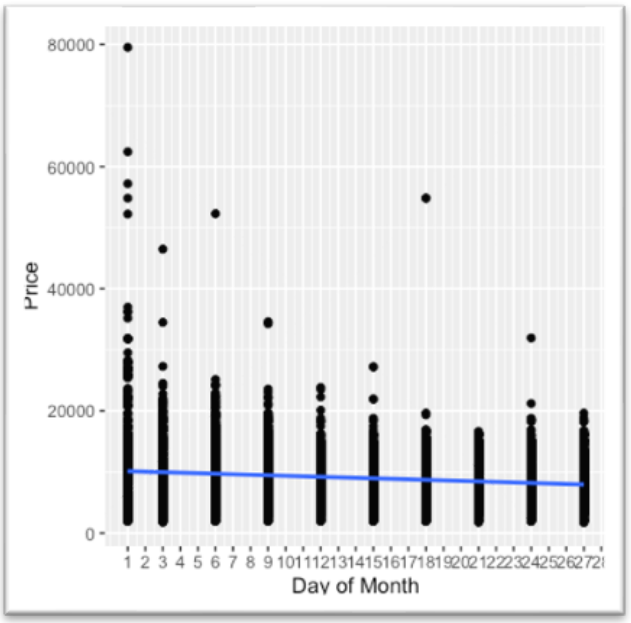
Hypothesis:

H0: $\mu 1 = \mu 2 = ... = \mu k$
H_a : at least two means differ

R-Code:

```
ggplot(df, aes(x = day, y = Price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Day of Month", y = "Price") +
  scale_x_continuous(breaks = 1:31)
```

Output:



Results:
As we can observe in the above graph, we must reject our null hypothesis as the days of the month appear to contribute to a fluctuation in the mean price of flights. Precisely, it appears that the mean prices of plane tickets gradually decrease as the month progresses. We believe that this is caused by the fact that at the beginning of the month consumers will be willing to spend more given that they have just received their income for the previous' month; instead, as the month progresses people tend to spend less as they will have less disposable income.

3.7. Understanding whether the month affects price.

Does the day of the month affect the Price?
→ Also in this case, our null hypothesis is that the different months available in the dataset won't affect the mean price of the flights.

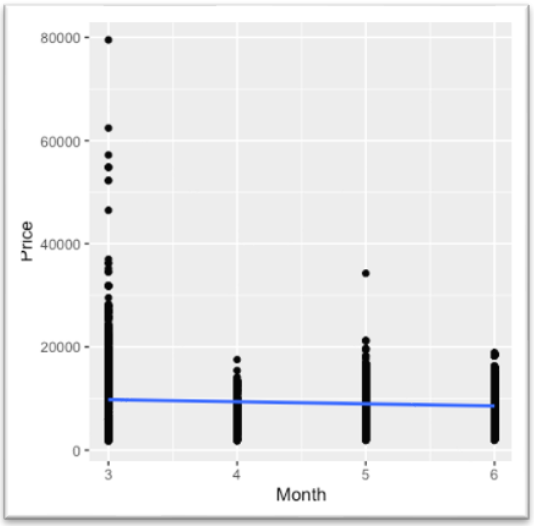
Hypothesis:

H0: $\mu 1 = \mu 2 = ... = \mu k$
H_a : at least two means differ

R-Code:

```
ggplot(df, aes(x = month, y = Price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Month", y = "Price") +
  scale_x_continuous(breaks = 3:6)
```

Output:



Results:

Judging from the graph, we must reject our null hypothesis once again and take into consideration that with the passage of the months, the mean price of flights decreases, with March being the month with the most expensive tickets on average while June being the cheapest one. We believe that this may be provoked by the fact that there are more flights within India during summer and in order to be price-competitive, most airlines reduce prices to attract lower-budget consumers. On the contrary, travelers who travel during the year may be more price elastic and airlines can charge higher prices.

3.8. Identifying the effect of airline type on stops .

Does the airline type and class significantly affect the number of stops ?
→ Our null hypothesis is that the coefficients of the airlines are equal to 0, meaning that all airlines operate in the same way when dealing with the number of stops they allocate in their flights.

R-Code:

```
lf <- df %>% mutate(airline_type = ifelse(Airline %in%
   c("IndiGo", "Air India", "Vistara", "Jet Airways"), "High-Cost", "Low-Cost"))

iodel <- lm(STOPS ~ Airline + airline_type, data = df)
;ummary(model)
```

Output:

```
lesiduals:
     Min       1Q   Median       3Q      Max
-1.23758 -0.40477 -0.01767  0.55115  2.76242

Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              0.46082    0.03223  14.298  < 2e-16 ***
\irlineAir India                         0.77676    0.03504  22.167  < 2e-16 ***
\irlineGoAir                             0.06496    0.05241   1.239 0.215199
\irlineIndiGo                           -0.05604    0.03464  -1.618 0.105749
\irlineJet Airways                       0.55685    0.03354  16.604  < 2e-16 ***
\irlineJet Airways Business              0.87252    0.23720   3.678 0.000236 ***
\irlineMultiple carriers                 0.58852    0.03627  16.225  < 2e-16 ***
\irlineMultiple carriers Premium economy 0.53918    0.16287   3.311 0.000934 ***
\irlineSpiceJet                         -0.27989    0.03800  -7.366 1.89e-13 ***
\irlineTrujet                            0.53918    0.57652   0.935 0.349686
\irlineVistara                          -0.01196    0.04160  -0.288 0.773665
\irlineVistara Premium economy          -0.46082    0.33389  -1.380 0.167576
;irline_typeLow-Cost                          NA         NA      NA       NA
---
;ignif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

lesidual standard error: 0.5756 on 10670 degrees of freedom
Multiple R-squared:  0.274,     Adjusted R-squared:  0.2733
:-statistic: 366.1 on 11 and 10670 DF,  p-value: < 2.2e-16
```

Results:

Given the extremely small p-value for some of them, we can conclude that for example Air India or GoAir have significant effects on the number of stops.

3. 9. To what extent the Airline affects the Price

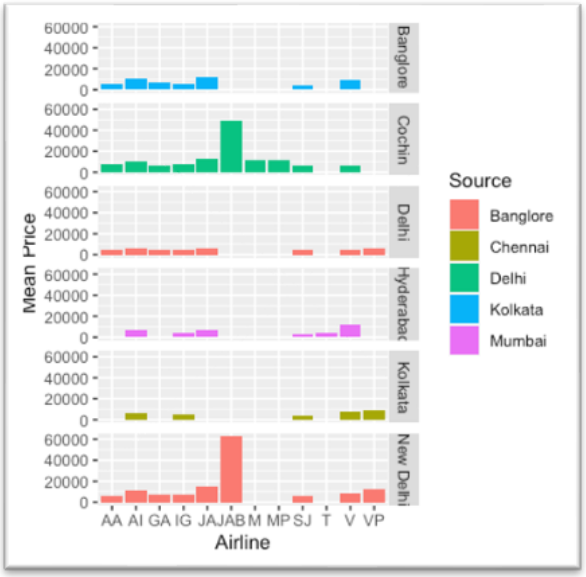
How do Airlines affect Price?

→ Given the vast number of airlines that operate in India, we predict that most airlines will try to offer low costs to appeal to the majority of travelers, while only a few will offer high prices to appeal to the travelers that want to fly luxuriously .

R-Code:

```
if %>%
  group_by(Source, Destination, Airline) %>%
  summarize(mean_price = mean(Price), n_flights = n()) %>%
  arrange(Source, Destination, mean_price) %>%
  mutate(acronym = str_extract_all(Airline, "[A-Z]") %>% sapply(paste, collapse = "")) %>%
  ggplot(aes(x = acronym, y = mean_price, fill = Source)) +
  geom_col() +
  facet_grid(rows = vars(Destination)) +
  labs(x = "Airline", y = "Mean Price", fill = "Source")
```

Output:



Results:

From the above graph it is clear that our prediction was correct and that low-budget travelers have a vast option of airlines whereas those who have higher budgets will most likely travel with Jet Airway Business.

4. Linear Regression Model

4.1 Linear regression model:

Moreover, one of the goals of this project is to help customers with a low budget get the best price for their flights. In order to see how all the variables affect the price, with the objective of optimizing it, we have decided to create a linear regression model.

Important note that for the purposes of the linear regression model we omitted any possible na values and didn't include identified outliers.

→ The first step in creating the model was to convert the necessary variables into factors.

```{r}
FDA = CLEAN_DATA_FDA


FDA$day = factor(FDA$day)
FDA$month = factor(FDA$month)
FDA$Source = factor(FDA$Source)
FDA$Destination = factor(FDA$Destination)
FDA$Time = factor(FDA$Time)
FDA$Total_Stops = factor(FDA$Total_Stops)

str(FDA)

```

```
tibble [10,683 × 10] (S3: tbl_df/tbl/data.frame)
 $ Airline    : chr [1:10683] "Air Asia" "Air Asia" "Air Asia" "Air Asia" ...
 $ day        : Factor w/ 10 levels "1","3","6","9",..: 2 2 6 8 5 10 10 5 4 9 ...
 $ month      : Factor w/ 4 levels "3","4","5","6": 3 2 4 4 2 2 3 3 4 4 ...
 $ Source     : Factor w/ 5 levels "Banglore","Chennai",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Destination: Factor w/ 6 levels "Banglore","Cochin",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Time       : Factor w/ 2 levels "Day","Night": 2 2 2 2 2 2 2 2 2 2 ...
 $ Duration   : num [1:10683] 170 170 170 170 170 170 170 170 170 170 ...
 $ Price      : num [1:10683] 6181 4282 4282 4282 4282 ...
 $ Total_Stops: Factor w/ 5 levels "1 stop","2 stops",..: 5 5 5 5 5 1 1 1 1 1 ...
 $ STOPS      : num [1:10683] 0 0 0 0 0 1 1 1 1 1 ...
```

→ Then, we split the dataset in order to create two random groups, one for training and one for validating our model.

1) Split Dataset - Model Validation - seed 2023

```{r}
set.seed(2023)
smp_size = round(0.75 * nrow(FDA))
check = sample(seq_len(nrow(FDA)), size = smp_size)
train = FDA[check, ]
test = FDA[-check, ]

```

4.2. Full Model

→ As a starting point, We created a linear regression model using "price" as the dependent variable and adding all the other variables as dependent variables to see if they have any significant effect on price.

```
Call:
lm(formula = Price ~ Airline + day + month + Source + Destination +
    Time + Duration + Total_Stops, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-12242  -1802   -261   1540  42035
```

```
Coefficients: (4 not defined because of singularities)
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                           9.390e+03  2.533e+02  37.066  < 2e-16 ***
AirlineAir India                      2.490e+03  2.227e+02  11.180  < 2e-16 ***
AirlineGoAir                          3.905e+02  3.139e+02   1.244    0.214
AirlineIndiGo                         1.999e+02  2.094e+02   0.955    0.340
AirlineJet Airways                    4.776e+03  2.099e+02  22.749  < 2e-16 ***
AirlineJet Airways Business           4.984e+04  1.351e+03  36.894  < 2e-16 ***
AirlineMultiple carriers              3.637e+03  2.330e+02  15.607  < 2e-16 ***
AirlineMultiple carriers Premium economy  4.666e+03  1.020e+03   4.573 4.88e-06 ***
AirlineSpiceJet                      -3.099e+02  2.322e+02  -1.335    0.182
AirlineTrujet                        -8.247e+02  2.986e+03  -0.276    0.782
AirlineVistara                        2.096e+03  2.536e+02   8.266  < 2e-16 ***
AirlineVistara Premium economy        2.561e+03  2.117e+03   1.210    0.226
day3                                 -9.872e+02  1.615e+02  -6.111 1.04e-09 ***
day6                                 -7.356e+02  1.451e+02  -5.068 4.11e-07 ***
day9                                 -1.588e+03  1.416e+02 -11.219  < 2e-16 ***
day12                                -1.584e+03  1.545e+02 -10.249  < 2e-16 ***
day15                                -2.246e+03  1.550e+02 -14.487  < 2e-16 ***
day18                                -1.915e+03  1.622e+02 -11.804  < 2e-16 ***
day21                                -2.951e+03  1.525e+02 -19.353  < 2e-16 ***
day24                                -2.192e+03  1.519e+02 -14.429  < 2e-16 ***
day27                                -2.742e+03  1.536e+02 -17.843  < 2e-16 ***
month4                               -2.885e+03  1.475e+02 -19.551  < 2e-16 ***
month5                               -1.033e+03  1.076e+02  -9.598  < 2e-16 ***
month6                               -1.837e+03  1.061e+02 -17.314  < 2e-16 ***
SourceChennai                        -2.338e+03  2.280e+02 -10.255  < 2e-16 ***
SourceDelhi                          -1.286e+02  1.518e+02  -0.847    0.397
SourceKolkata                        -1.100e+03  1.569e+02  -7.013 2.53e-12 ***
SourceMumbai                         -3.978e+03  1.861e+02 -21.375  < 2e-16 ***
DestinationCochin                            NA         NA      NA       NA
DestinationDelhi                     -2.749e+03  1.822e+02 -15.085  < 2e-16 ***
DestinationHyderabad                         NA         NA      NA       NA
DestinationKolkata                           NA         NA      NA       NA
DestinationNew Delhi                         NA         NA      NA       NA
TimeNight                             1.331e+02  9.607e+01   1.385    0.166
Duration                              1.477e+00  8.926e-02  16.550  < 2e-16 ***
Total_Stops2 stops                   -1.162e+02  1.078e+02  -1.078    0.281
Total_Stops3 stops                    8.255e+02  5.327e+02   1.550    0.121
Total_Stopsnon-stop                  -4.344e+01  8.576e+01  -0.506    0.613
---

Residual standard error: 2974 on 7977 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5947,    Adjusted R-squared:  0.593
F-statistic: 354.7 on 33 and 7977 DF,  p-value: < 2.2e-16
```

4.3 Reduced model

→ Now let's try to build a more reduced model with only those variables that have a relationship with the dependent variable y (Prices), which are the ones that are giving us useful information. For this purpose:

- We will only use those independent variables that have a relationship with the dependent variable (y) (STEP 1).
- We will not use redundant variables (STEP 2)

4.4 Step 1: study the relationship with independent variables.

1.1 Check if quantitative independent variables are related to the dependent variable:

Check p-values for the β coefficients to see if there is a linear relationship between each one of the quantitative independent variables and the dependent one

The results of our linear regression model illustrates the effect of each variable and their values. For example, we can see that the destination of the flight does not affect the price, except for the destination of Delhi. The NA coefficients convey that the impacts on the

independent variable are negligible, we also discussed this in step 4 where the impact of source in the Anova this was also represented. That being said, we can see that the source of the flight has a significant effect on the price while the destination doesn't. In addition, the different days also have an effect on the months.

If we evaluate significance with an alpha of 0.05, she can notice that some of the variables don't affect the price. That is the case with airline GoAir, IndiGo, SpiceJet, Trujet, and Vistara Premium economy. We understand that this happens because these companies may have a very similar price to the market demand. Since this market is very competitive, these don't affect the price much and most of the airlines are price takers. . That being said, other airlines do have a significant effect on the price if we observe their p-values. Consequently, the variable "Airline" will not be removed from the model.

Moreover, it appears that the number of stops does not have a significant effect on the price. This seems counterintuitive but we have concluded that the effect or variability of the number of stops is taken into account with the variable "Duration" whose p-value is very low. In addition the time of the flight (day or night) doesn't seem to have a major effect.

Lastly, the Time variable which categorizes whether the flight is during the day or during the night has a high p value of 0.166 which means at the level of confidence of 0.05, the time of the flight does not significantly impact the price of Indian Flights

4.42 Check if qualitative independent variables are related to the dependent variable:
Check results of the ANOVA TESTS; to remember: all significant according to anova

```{r}
summary(aov(Price~Airline, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Airline      11 9.332e+10 8.484e+09   676.7 <2e-16 ***
Residuals 10669 1.338e+11 1.254e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~day, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
day           9 7.629e+09 847687646   41.22 <2e-16 ***
Residuals 10671 2.194e+11  20564831
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~month, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
month         3 1.892e+10 6.308e+09   323.6 <2e-16 ***
Residuals 10677 2.082e+11 1.950e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~STOPS, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
STOPS         1 1.716e+10 1.716e+10   872.7 <2e-16 ***
Residuals 10679 2.099e+11 1.966e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~Source, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Source        4 3.056e+10 7.639e+09     415 <2e-16 ***
Residuals 10676 1.965e+11 1.841e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~Destination, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Destination   5 5.518e+10 1.104e+10   685.3 <2e-16 ***
Residuals 10675 1.719e+11 1.610e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~Time, FDA))
```

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
Time          1 3.323e+08 332282577   15.65 7.67e-05 ***
Residuals 10679 2.267e+11  21232721
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
summary(aov(Price~Duration, FDA))
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Duration      1 5.832e+10 5.832e+10    3690 <2e-16 ***
Residuals 10679 1.688e+11 1.580e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.5. Step 2: remove redundant variables

Redundant variables are those quantitative variables that are giving the same information, in other words, there is a high correlation between them. However because we have mostly categorical data and only one numerical variable (Duration) we will not be able to do a correlation matrix to check for dependencies. We will therefore check the impact of different combinations of variables on the adjusted R^2 which introduces a penalty for irrelevant variables.

4.6. Step 3: Reduce model

After this evaluation, despite all variables being significant as shown in the anova tests, due to the input of all variables in the full model, some variables might explain the same variation which means they lose significance when introduced in the overall model.

For this second model, we have eliminated the variables "Total_Stops" and "Time" since we identified previously that they were not significant.

```
Call:
lm(formula = Price ~ Airline + day + month + Destination + Source +
    Duration, data = train)

Coefficients: (4 not defined because of singularities)
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          8.292e+03  2.460e+02  33.710  < 2e-16 ***
AirlineAir India                     2.443e+03  2.164e+02  11.288  < 2e-16 ***
AirlineGoAir                         3.606e+02  3.126e+02   1.154    0.249
AirlineIndiGo                        1.873e+02  2.091e+02   0.896    0.370
AirlineJet Airways                   4.747e+03  2.057e+02  23.074  < 2e-16 ***
AirlineJet Airways Business          4.984e+04  1.350e+03  36.931  < 2e-16 ***
AirlineMultiple carriers             3.625e+03  2.254e+02  16.083  < 2e-16 ***
AirlineMultiple carriers Premium economy  4.644e+03  1.018e+03   4.560 5.18e-06 ***
AirlineSpiceJet                     -3.267e+02  2.313e+02  -1.412    0.158
AirlineTrujet                       -8.517e+02  2.986e+03  -0.285    0.775
AirlineVistara                       2.055e+03  2.515e+02   8.173 3.46e-16 ***
AirlineVistara Premium economy       2.499e+03  2.116e+03   1.181    0.238
day3                                -9.871e+02  1.616e+02  -6.110 1.04e-09 ***
day6                                -7.389e+02  1.451e+02  -5.091 3.65e-07 ***
day9                                -1.591e+03  1.416e+02 -11.240  < 2e-16 ***
day12                               -1.586e+03  1.545e+02 -10.266  < 2e-16 ***
day15                               -2.246e+03  1.550e+02 -14.487  < 2e-16 ***
day18                               -1.908e+03  1.622e+02 -11.764  < 2e-16 ***
day21                               -2.953e+03  1.525e+02 -19.361  < 2e-16 ***
day24                               -2.191e+03  1.519e+02 -14.427  < 2e-16 ***
day27                               -2.742e+03  1.536e+02 -17.846  < 2e-16 ***
month4                              -2.884e+03  1.475e+02 -19.550  < 2e-16 ***
month5                              -1.035e+03  1.076e+02  -9.623  < 2e-16 ***
month6                              -1.838e+03  1.061e+02 -17.328  < 2e-16 ***
DestinationCochin                    9.864e+02  9.636e+01  10.237  < 2e-16 ***
DestinationDelhi                    -1.635e+03  1.291e+02 -12.668  < 2e-16 ***
DestinationHyderabad                -2.851e+03  1.559e+02 -18.288  < 2e-16 ***
DestinationKolkata                  -1.229e+03  2.001e+02  -6.143 8.47e-10 ***
DestinationNew Delhi                 1.122e+03  1.561e+02   7.187 7.24e-13 ***
SourceChennai                               NA         NA      NA       NA
SourceDelhi                                 NA         NA      NA       NA
SourceKolkata                               NA         NA      NA       NA
SourceMumbai                                NA         NA      NA       NA
Duration                             1.486e+00  8.905e-02  16.692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2974 on 7981 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5944,    Adjusted R-squared:  0.5929
F-statistic: 403.3 on 29 and 7981 DF,  p-value: < 2.2e-16
```

In this case we see that “Destination” is now significant but “Source” isn’t. (Previously this was the other way around) Possibly, this could have happened because source and destination explain the same variability of price and with this new combination of variables destination’s explanation of variability is more relevant. As an alternative, we decided to test another model eliminating the variable “Source”.

```
Call:
lm(formula = Price ~ Airline + day + month + Destination + Duration,
    data = train)

Residuals:
   Min     1Q Median     3Q    Max
-12388  -1801   -263   1545  42031

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            8.292e+03  2.460e+02  33.710  < 2e-16 ***
AirlineAir India                       2.443e+03  2.164e+02  11.288  < 2e-16 ***
AirlineGoAir                           3.606e+02  3.126e+02   1.154    0.249
AirlineIndiGo                          1.873e+02  2.091e+02   0.896    0.370
AirlineJet Airways                     4.747e+03  2.057e+02  23.074  < 2e-16 ***
AirlineJet Airways Business            4.984e+04  1.350e+03  36.931  < 2e-16 ***
AirlineMultiple carriers               3.625e+03  2.254e+02  16.083  < 2e-16 ***
AirlineMultiple carriers Premium economy  4.644e+03  1.018e+03   4.560 5.18e-06 ***
AirlineSpiceJet                       -3.267e+02  2.313e+02  -1.412    0.158
AirlineTrujet                         -8.517e+02  2.986e+03  -0.285    0.775
AirlineVistara                         2.055e+03  2.515e+02   8.173 3.46e-16 ***
AirlineVistara Premium economy         2.499e+03  2.116e+03   1.181    0.238
day3                                  -9.871e+02  1.616e+02  -6.110 1.04e-09 ***
day6                                  -7.389e+02  1.451e+02  -5.091 3.65e-07 ***
day9                                  -1.591e+03  1.416e+02 -11.240  < 2e-16 ***
day12                                 -1.586e+03  1.545e+02 -10.266  < 2e-16 ***
day15                                 -2.246e+03  1.550e+02 -14.487  < 2e-16 ***
day18                                 -1.908e+03  1.622e+02 -11.764  < 2e-16 ***
day21                                 -2.953e+03  1.525e+02 -19.361  < 2e-16 ***
day24                                 -2.191e+03  1.519e+02 -14.427  < 2e-16 ***
day27                                 -2.742e+03  1.536e+02 -17.846  < 2e-16 ***
month4                                -2.884e+03  1.475e+02 -19.550  < 2e-16 ***
month5                                -1.035e+03  1.076e+02  -9.623  < 2e-16 ***
month6                                -1.838e+03  1.061e+02 -17.328  < 2e-16 ***
DestinationCochin                      9.864e+02  9.636e+01  10.237  < 2e-16 ***
DestinationDelhi                      -1.635e+03  1.291e+02 -12.668  < 2e-16 ***
DestinationHyderabad                  -2.851e+03  1.559e+02 -18.288  < 2e-16 ***
DestinationKolkata                    -1.229e+03  2.001e+02  -6.143 8.47e-10 ***
DestinationNew Delhi                   1.122e+03  1.561e+02   7.187 7.24e-13 ***
Duration                               1.486e+00  8.905e-02  16.692  < 2e-16 ***

Residual standard error: 2974 on 7981 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5944,    Adjusted R-squared:  0.5929
F-statistic: 403.3 on 29 and 7981 DF,  p-value: < 2.2e-16
```

In this model all of our variables result to be significant.

4.7. Step 4. Model Selection

	Full Model	Reduced Model1	Reduced Model3
R^2	0.593	0.5929	0.5929
RSE	2974	2974	2974
P-value F stat	2.2e-16	2.2e-16	2.2e-16

When comparing our different models, we can see that they don’t perform significantly differently than their R^2, RSE and p value determine a similar usefulness, accuracy and error. Therefore for the purposes of parsimonia, the preference of a simpler model, we will

use the reduced model 3 where betas have significant hypothesis tests rendering them as adequate, there's less values so more interpretation and it performs nearly equally to other models.

To ensure this we carried out an anova test in which the null hypothesis states that the full model and model2 have the same explainability. Because the p value of this test is very low, 0.2115 we assume that they don't differ in capturing the variability in price, we can't reject the null hypothesis that they are equally useful . This supports our previous conclusion that we will choose the reduced model 2.

```{r}
Fullmodel= lm(Price ~ Airline + day + month+ Source + Destination + Time+ Duration +
Total_Stops, data = train)
model1= lm(Price ~ Airline + day + month+ Destination + Source + Duration, data = train)
model2= lm(Price ~ Airline + day + month+ Destination + Duration, data = train)
anova(Fullmodel, model2)

```

```
Analysis of Variance Table

Model 1: Price ~ Airline + day + month + Source + Destination + Time +
    Duration + Total_Stops
Model 2: Price ~ Airline + day + month + Destination + Duration
  Res.Df        RSS Df Sum of Sq    F Pr(>F)
1   7977 7.0558e+10
2   7981 7.0610e+10 -4 -51654475 1.46 0.2115
```

4.8. Step 5. Ultimate model Interpretation and Evaluation

1) Hypothetical model

Price =

```
Air India*2.443e+03 +GoAir*3.606e+02 + IndiGo*1.873e+02 + Jet Airways*4.747e+03 +
Jet Airways Business*4.984e+04 + Multiple carriers*3.625e+03  +Multiple carriers
Premium economy* 4.644e+03 + SpiceJet*-3.267e+02 + Trujet*-8.517e+02 + Vistara*
2.055e+03 + Vistara Premium economy*2.499e+03 + day6 * -7.389e+02 + day9*
-1.591e+03 +1day12* -1.586e+03 +day15*-2.246e+03 +day18* -1.908e+03 + day21*
-2.953e+03 +day24*-2.191e+03 + day27 *-2.742e+03 + month4*-2.884e+03 + month5*
-1.035e+03 +month6* -1.838e+03 + DestinationCochin* 9.864e+02 + DestinationDelhi*
-1.635e+03 +DestinationHyderabad* -2.851e+03 +DestinationKolkata * -1.229e+03 +
DestinationNew Delhi* 1.122e+03 + Duration* 1.486e+00
```

2) Adequacy of the model

After running the model and obtaining and interpreting the p value of the f statistic (403.3) , we reject the null hypothesis that all the betas are equal due to a very low p value (2.2e-16), so we can conclude that overall there is a significant and positive effect of the model.

Moreover the t statistics and low p values of the individual Betas mean they are different from 0 and so and therefore the model is adequate and the impact of every variable on Price is correct.

For interpreting the numerical value of duration, an increase by one unit in duration will convey an increase in price of 1.486e+00.

Categorical values Betas however are not slopes, given that the variables are dummy's and store either 0s or 1s they will just have an impact on the intercept from the base level creating therefore alternative linear models for each combination of categorical values. For example Airline India will increase the price by 2.443e+03 compared to the base airline which is included in the intercept.

3) Usefulness of the model

Adjusted R^2 is 59.29 %. So our model, and our independent variables explain around 60% of the variability of Price of flights in India. This is not very high but in the real world explaining all of the variability of a specific variable such as price is very hard. Several other factors might affect prices that are not included in our model, moreover polynomial effects or interactions which are not considered in our model could increase the utility of our model.

Given that this R^2 is the highest that we got with our available data we will conclude that although the model could be improved it explains a considerable amount of price's variability.

4) RSE explains accuracy

The RSE of our model is 2947, although this might appear as a large error given that our prices range from 1759 to 79512, with a mean of 9087.0, in the appropriate units this is not as big of an error.

We can calculate a confidence interval for where 95% of our prices will lie.

```{r}
mean(FDA$Price) - 1.92*2947
mean(FDA$Price) + 1.92*2947
```

```
[1] 3427.856
[1] 14744.34
```

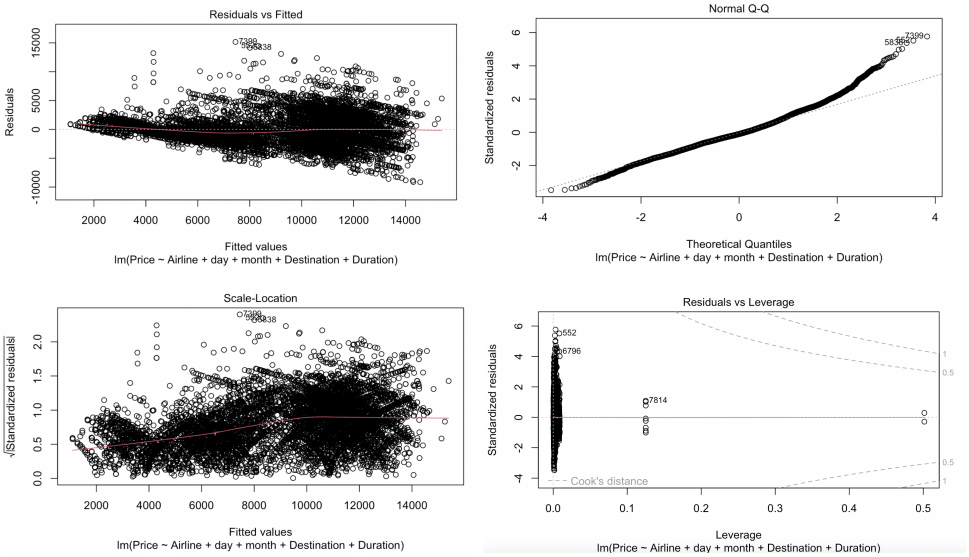
In this case prices lie between 3427.856 and 14744.34. This is a large range but of course by inherent nature flight price is a variable with variability.

6) Assumptions to take into account

Assumptions:

1. The relationship between the independent and dependent variables are linear
2. The residuals are normally distributed
3. All residuals have the same variance
4. The observations are independent of one another

Visualization of Residual analysis



1) Linearity of the model: To get a reliable model, we want the red line to be as horizontal as possible and to match the dotted line. We want the residuals to be equally dispersed above and below the dotted line. Our plot one matches this so this assumption is met.
2) The Normal Q-Q plot is used to check if our residuals follow Normal distribution or not. The residuals are normally distributed if the points follow the dotted line closely. Our model mostly also follows this assumption except.
3) Regarding the Homogeneity of the variance, the red line is not completely horizontal but is still nearly horizontal and values are spread.
4) In plot 4 no values are within the red dotted line, we can state that there are no important outliers in our model.
5) Ultimately, According to the independence of the error, we make the following hypothesis:

-H0: errors are uncorrelated
-Ha: errors are correlated

```
        Durbin-Watson test

data:  model3
DW = 1.9921, p-value = 0.3616
alternative hypothesis: true autocorrelation is greater than 0
```

Because the p value is 0.3616 we maintain the null hypothesis that errors are uncorrelated.
The overall conclusion is that assumptions are met, so there is no need to improve our model.

4.9 . Step 5. Validation

We used K- fold validation which involves the following steps:

1. Randomly split the data set into k-subsets (or k-fold) (for example 5 subsets)- predetermined number
2. Reserve one subset and train the model on all other subsets
3. Test the **model** on the reserved subset and record the prediction error
4. Repeat this process until each of the k subsets has served as the test set.
5. Compute the average of the k recorded errors. This is called the cross-validation error serving as the performance metric for the model.

```{r}
library(caret)
#Define training control
train = na.omit(train)
trainControl = trainControl(method = "cv", number = 10)
# Train the model
lm = train(Price ~ Airline + day + month + Destination + Duration, data = train, method ="lm", trControl =
trainControl)
# Summarize the results
lm
```

```
Warning: prediction from a rank-deficient fit may be misleadingLinear Regression

8011 samples
   5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7210, 7209, 7211, 7211, 7209, 7210, ...
Resampling results:

  RMSE      Rsquared   MAE
  2983.193  0.5879878  2142.75

Tuning parameter 'intercept' was held constant at a value of TRUE
```

The R^2 is very similar to what we already had and the RMSE is also similar to our RSE. So these reinforce the same ideas argued when interpreting the previous values.

4.91. Step 6. Model Prediction

Finally, we used the test data set we created at the start of the creation of our linear regression to test how efficient our model is at predicting price for different observations. We then calculated the sum of squared errors between the empirical and the predicted values by our model to calculate the residual standard error and then the average mean error of our predictions. The relative mean error is 0.3179. Although this is a considerable error, as mentioned due to the real life nature of our data set and the lack of enough predictor variables, some error is a reality. Nevertheless it still generalizes relatively well with an unknown data set, and does not suffer any overfitting problem.

```{r}
test$pred = predict(object = model, newdata = test)
SSE = sum((test$Price-test$pred)^2)
RSE = sqrt(SSE/(nrow(test)-3))
mean_error = RSE/(mean(test$Price))
mean_error
```

```
[1] 0.3178908
```

Finally, we also used our model to make individual predictions of price for specific scenarios, and calculate the expected value. Shown below.

```{r}
model3= lm(Price ~ Airline + day + month + Destination + Duration, data = train)
new = data.frame(Airline = "Air Asia", day = "3", month = "6", Destination = "Banglore", Duration = 300)
#mean expected value (mean of price)
predict(object = model3, newdata = new, interval = "confidence", level = 0.95)
#predict specific value of y, more variability.
predict(object = model3, newdata = new, interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 5912.356 5446.814 6377.898
       fit     lwr      upr
1 5912.356 63.1257 11761.59
```

In this case, for a flight with the Air Asia Airline, on day three of the month and on june with destination to Bangalore and duration of 300 minutes, the mean expected value of the price lies between 5446.814 and 6377.898, this can be considered a low range and so our model is good at predicting the price for this general case. However when predicting the specific value there's much more variability in place and the confidence interval ranges from 63.1257 to 11761.59 which is a very big range and hardly insightful. As a result due to our lack of explainability, afterall our model only explains no more than 59% of the variability in price, our model is not very confident when predicting specific prices which are more variable. In conclusion we can use our model to confidently make predictions of expected price but not of specific.

#Prediction

```{r}
model3= lm(Price ~ Airline + day + month + Destination + Duration, data = train)
new = data.frame(Airline = "Air India", day = "9", month = "4", Destination = "Cochin", Duration =
1045)
#mean expected value (mean of price)
predict(object = model3, newdata = new, interval = "confidence", level = 0.95)
#predict specific value of y, more variability.
predict(object = model3, newdata = new, interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 8799.464 8467.819 9131.109
       fit      lwr      upr
1 8799.464 2959.365 14639.56
```

As an alternative, this second prediction with different independent variables shows a similar prediction offering an interval for the expected price of 8647.819 to 9131 and for the specific value of 2959 to 14639.

5. Conclusions

This data exploration started off by looking at a dataset with information about the flights in India. As a travel agency, our objective is to to find the best option for our clients given their interests. For this, we identified two goals: offer the best route to a client looking for a low-cost option, and the best route for someone looking for the most comfortable flight experience. In order to do so, we came up with certain questions and predictions.

We predicted that fewer stops would increase the price per ticket especially for long distance flights, which turned out to be correct. Further we were analyzing the differences of prices between day and night flights. In this case we found out that the ones at night are significantly cheaper. Another point we were able to predict and identify was that the source as well as the destination cause differences in the prices.
After this, we wanted to learn whether the flight duration was a factor that significantly affected the ticket prices. Yet, the Anova test outlined that this isn't necessarily the case and that there must be several other factors that contribute to the price of flights.
Consequently, we thought about whether the days and the months contributed to fluctuations in price level. Unfortunately, we couldn't make a year-round analysis of these two factors as our dataset only included information regarding the months of March through June. Yet, for the available time frame we had, we concluded that as time progressed there was a slight decrease in price levels.
In order to differentiate between airlines that would provide comfort to our customers, we found the trend of airlines providing stops, to then identify those who mostly provide direct flights which we think is key for customers who are looking for comfort.
Finally, we outlined the different airlines that customers can rely on based on their budgets and where they want to fly from within India. Following an analysis, we were able to determine that customers that want to travel on a budget will have more options than those who prefer the luxurious route, as well as which airlines such customers should rely on.

Regarding the linear regression model, after reducing our model on the basis of statistical significance and taking the coefficient of determination as our main indicator, our regression model proved to be relatively effective. Despite only explaining less than 60% of the variability of the price variable, given it is a real case scenario and flight prices are highly variable we identify this as a success. Moreover, we were able to successfully predict mean expected prices.

Overall, we were able to successfully identify and comprehend each variable that contributes to the price of flights in India, which allows us to create a model that accurately predicts within a 95% interval the expected budget needed for travelers who are either looking to travel low-cost or as comfortably as possible.