Flight Price Prediction

Milestone1: exploratory data analysis

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi

Date: Nov 13th, 2022

[Introduction]

This is about analysis using flight price data about India airlines from the Kaggle. I choose this data to predict flight price. The price is target data. By using other sources like daysleft, duration, class, stops, and departure & arrival time, I will predict price. I'm interested in change of flight price, because I spend a lot of money in booking airplane. I want to analyze USA data and other country from analyzing india flight data first. I hope to learn what is the important factors which affect flight price and draw out key business question about flight industry. I cleaned data by renaming columns and delete outliers.

Q Describe your dataset

This data is all about flight price and information. Target data is price which is numeric data. There are airlines, cities, stops and departure & arrival time which can be categorized. There are duration and days left which is numeric data.

[Statistical Analysis of Descriptive]

Table 1. The 5 headtail before cleaning

```
> headtail(fpp, 5)
            X  airline  flight source_city departure_time stops  arrival_time destination_city    class duration days_left price
1           0 SpiceJet SG-8709       Delhi        Evening  zero         Night           Mumbai  Economy     2.17         1  5953
2           1 SpiceJet SG-8157       Delhi  Early_Morning  zero       Morning           Mumbai  Economy     2.33         1  5953
3           2  AirAsia  I5-764       Delhi  Early_Morning  zero Early_Morning           Mumbai  Economy     2.17         1  5956
4           3  Vistara  UK-995       Delhi        Morning  zero     Afternoon           Mumbai  Economy     2.25         1  5955
5           4  Vistara  UK-963       Delhi        Morning  zero       Morning           Mumbai  Economy     2.33         1  5955
...       ...     <NA>    <NA>        <NA>           <NA>  <NA>          <NA>             <NA>     <NA>      ...       ...   ...
300150 300149  Vistara  UK-826     Chennai      Afternoon   one         Night        Hyderabad Business    10.42        49 77105
300151 300150  Vistara  UK-832     Chennai  Early_Morning   one         Night        Hyderabad Business    13.83        49 79099
300152 300151  Vistara  UK-828     Chennai  Early_Morning   one       Evening        Hyderabad Business       10        49 81585
300153 300152  Vistara  UK-822     Chennai        Morning   one       Evening        Hyderabad Business    10.08        49 81585
```

✓ First look of outline or overview of data what's in the data exactly.

Q Describe any data cleaning you did

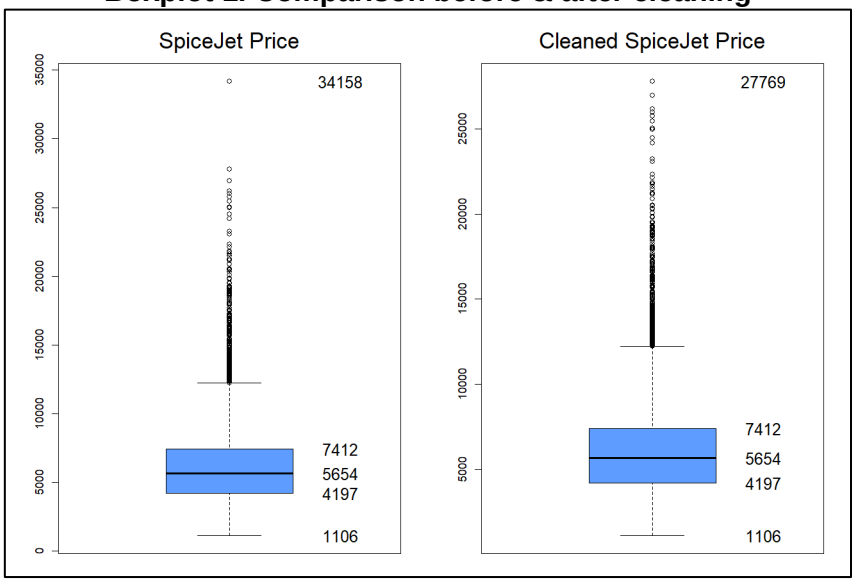
1. Change column name "source_city" into "from" and "destination_city" into "to" and "days_left" into "daysleft".
2. Check all boxplots for prices by airline and remove price outliers by treating them as NAs. I delete the price data which is too far from other prices. I confirm that other numeric data, which are duration and daysleft, is almost continuos
3. Added longitude and latitude data for each airport that is not in this data. The distance is calculated through this. The relationship between distance and price will also be explored.

Table 2. The 5 headtail after cleaning and editing

```
            X  airline  flight    from departure_time stops  arrival_time       to    class duration daysleft price from_lati from_long to_long to_lati distance
1           0 SpiceJet SG-8709   Delhi        Evening  zero         Night   Mumbai  Economy     2.17        1  5953     28.56      77.1   72.87    19.1   107.33
2           1 SpiceJet SG-8157   Delhi  Early_Morning  zero       Morning   Mumbai  Economy     2.33        1  5953     28.56      77.1   72.87    19.1   107.33
3           2  AirAsia  I5-764   Delhi  Early_Morning  zero Early_Morning   Mumbai  Economy     2.17        1  5956     28.56      77.1   72.87    19.1   107.33
4           3  Vistara  UK-995   Delhi        Morning  zero     Afternoon   Mumbai  Economy     2.25        1  5955     28.56      77.1   72.87    19.1   107.33
5           4  Vistara  UK-963   Delhi        Morning  zero       Morning   Mumbai  Economy     2.33        1  5955     28.56      77.1   72.87    19.1   107.33
...       ...     <NA>    <NA>    <NA>           <NA>  <NA>          <NA>     <NA>     <NA>      ...      ...   ...       ...       ...     ...      ...      ...
300150 300149  Vistara  UK-826 Chennai      Afternoon   one         Night Hyderabad Business    10.42       49 77105     12.99     80.18   78.42    17.24    21.13
300151 300150  Vistara  UK-832 Chennai  Early_Morning   one         Night Hyderabad Business    13.83       49 79099     12.99     80.18   78.42    17.24    21.13
300152 300151  Vistara  UK-828 Chennai  Early_Morning   one       Evening Hyderabad Business       10       49 81585     12.99     80.18   78.42    17.24    21.13
300153 300152  Vistara  UK-822 Chennai        Morning   one       Evening Hyderabad Business    10.08       49 81585     12.99     80.18   78.42    17.24    21.13
```

✓ Change column names as "source_city" into "from" and "destination_city" into "to" and "days_left" into "daysleft" to make it easier to see and manipulate. Cleaning outliers while checking in boxplots of price by airlines. Finding airport longitude and latitude from google and add it on data. Calculate the distance^2 between arrival and departure.

Boxplot 1. Comparison before & after cleaning



Q What is the purpose of the dataset? What is your data source?

The goal of my data analysis is to predict flight price by other factors. The main business question that can be derived from this is **what is best pricing policy that the company might choose and how to optimize it for better profit**. As a consumer, I know **what strategy I should take when booking a flight.**

Q How many rows of data are there? how many fields?

The data contains a total of **300,153 data and has 12 variables**. The main numeric data are duration, daysleft, and price, I found it using describe(). After cleaning and editing there are 5 more variables which are distance information with longitude & latitude.

Q What kind of data is included? Is it all text data, is it numerical?

My target data is price and other two numeric data which is duration and daysleft. The category data are airlines, cities, stops and departure & arrival time. There is 'from' and 'to' which indicate city, and I want to add it as a factor by adding longitude and latitude and calculating the distances between cities. Distance is numeric data.

Q Describe the data fields including the title, the data type.

The 'airline' column contains 6 airlines in inda. The 'stops' column has 'zero', 'one', 'two or more'. The 'departure & arrival time' contains 'night', 'morning', 'afternoon', and 'Late_night'. The 'flight' column has flight number.

Table 3. pysco::describe

X	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Duration	10	300153	12.22	7.19	11.25	11.72	6.92	0.83	49.83	49	0.60	-0.27	0.01
daysleft	11	300153	26.00	13.56	26.00	26.09	17.79	1.00	49.00	48	-0.04	-1.16	0.02
Price	12	300153	20889.6	22697.7	7425	17547.69	5825.14	1105	123071	121966	1.06	-0.40	41.43

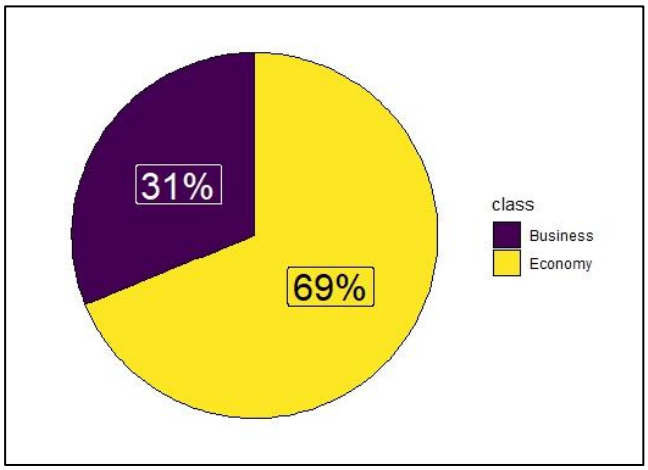
Q Provide descriptive statistical tables for key data fields of interest.

1. First, the most important data is the price data. Mean of price is 20889.6 and lowest and highest price is 1105 and 123071. Sd of price is 22697.7. It looks highly distributed comparing with other variables.

2. If you look at this as a class category, you can see that airlines with business have higher prices. Mean of daysleft is 26 and range is 48 The minimum value is 1 to a maximum of 49 days. Flight duration is a minimum of 0.83 hour and a maximum of 49.83 hours.

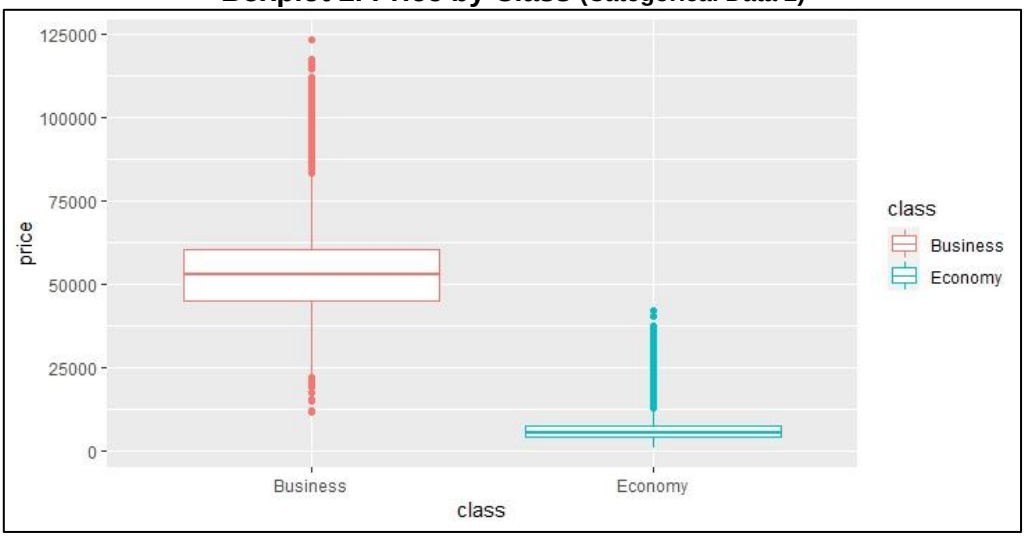
[Analysis by plot and charts]

Pie chart 1. Price by Class (Categorical Data 1)



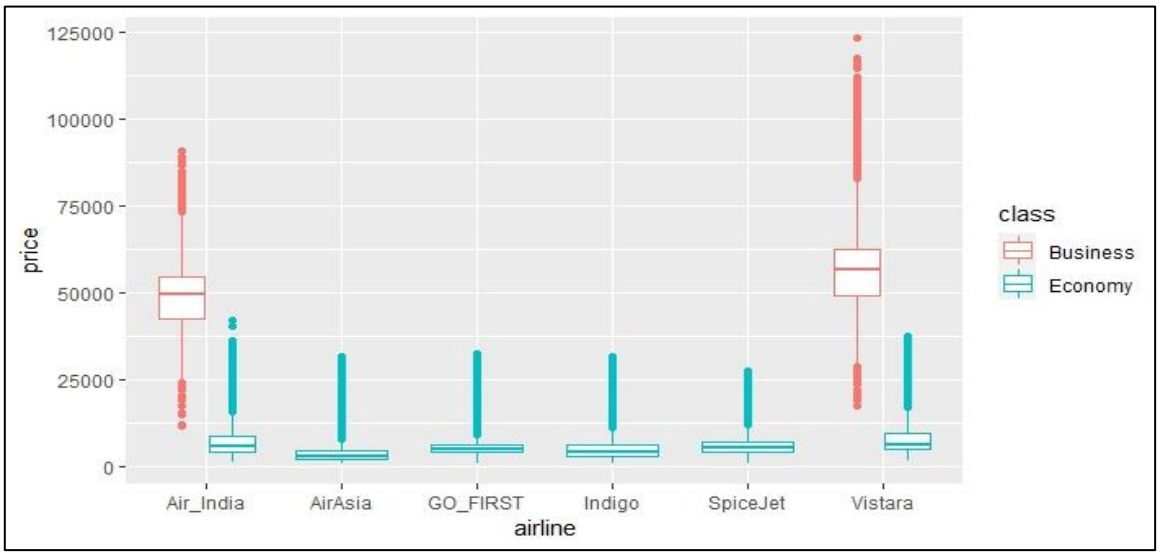
✓ Pie chart 1: In the common sense, class is very important factor that decide flight price, so I will start with the analysis of class.

Boxplot 2. Price by Class (Categorical Data 2)



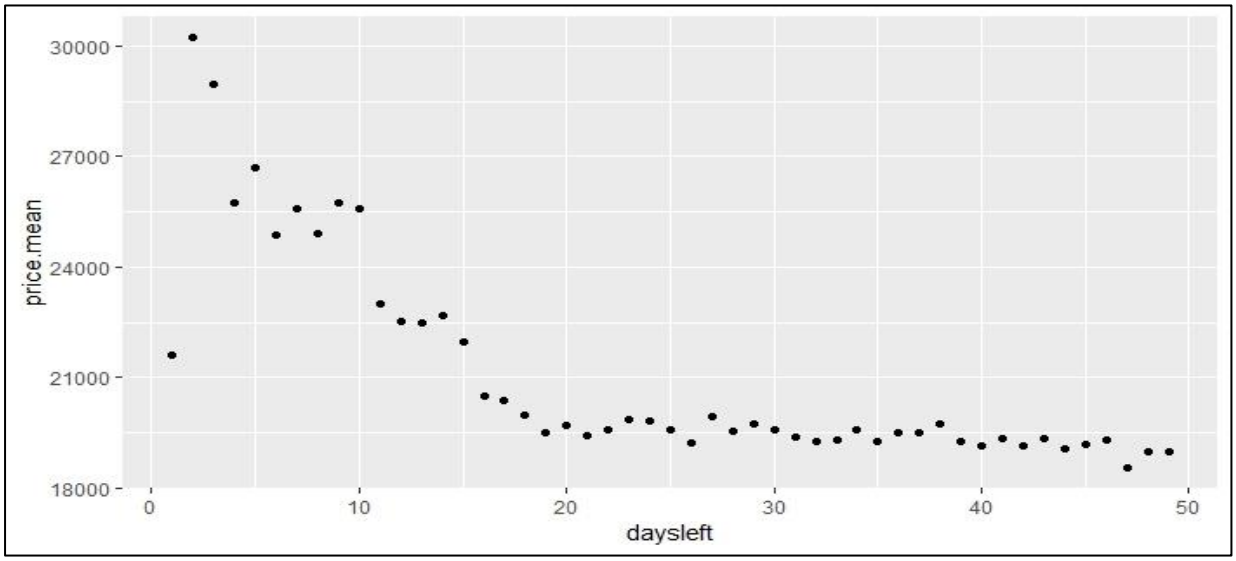
✓ Boxplot 2: According to class, difference between mean of business and economy is 45,965.74 and sd is 22,697.7. Max is 123071 and min is 1105, and price of economy is very narrow in the interquartile range.

Boxplot 3. Price by Airline adding class (Categorical Data 3)



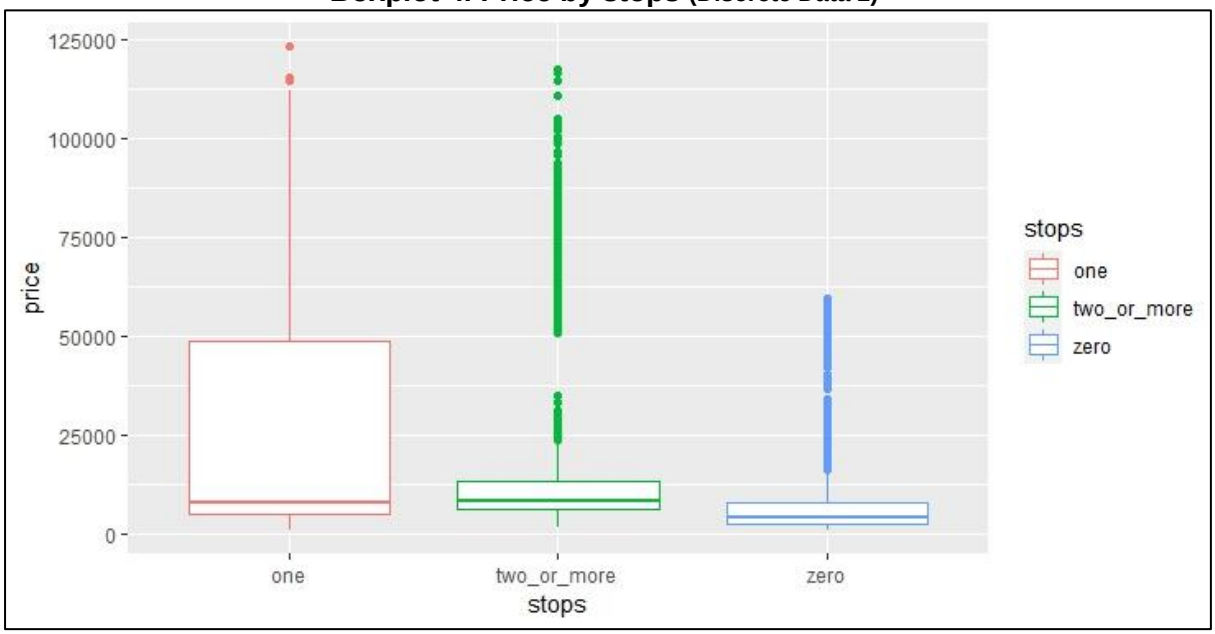
✓ Boxplot3: In the two expensive airlines, class affects huge impact on price. In economy class the price of all airlines is almost same but Air_India and Vistara is little higher than others.

Scatter plot 1. Price.mean by days_left (Discrete Data 1)

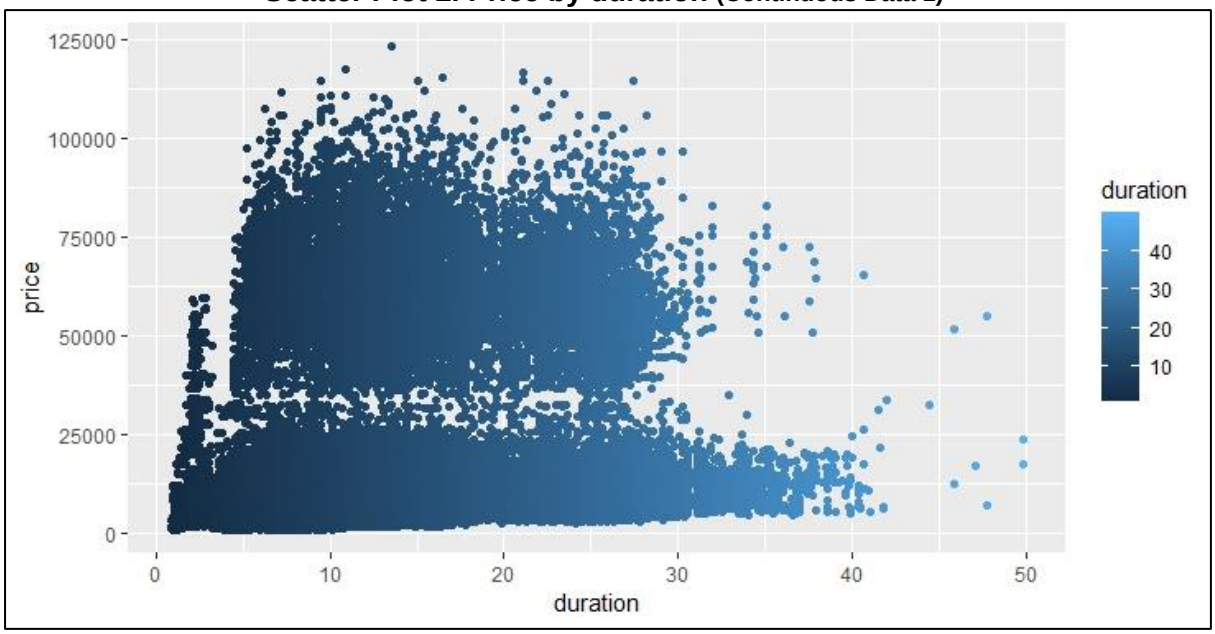


✓ Scatter plot 1: There are 300,153 data in this dataset, so it's clearer to see in the mean by daylefts. There is negative correlation between two. It means that if daysleft is high, the price will go down, and in over 20 daysleft, the price is not hugely changed.

Boxplot 4. Price by stops (Discrete Data 1)



Scatter Plot 2. Price by duration (Continuous Data 1)



Q <u>Provide visualizations of the key data and subset data of interest. This should be done for categorical data, discrete data and continuous data.</u>

1. To see what factor affects plane prices, I first looked at the class data that I thought would have the greatest impact. Through the pie chart 1, it was found that the class occupied 69% and 31%. After that, I compared the prices by class first.
2. Draw a boxplot that explain prices variation by airline. And add class data to the boxplot. I can see which airlines operate business class and what is the price range.
3. As the main discrete chart daysleft it has 1 day to 48 days. I can guess correlation between price and daysleft. I draw boxplot by stops which is ‘zero’, ‘one’, ‘two_or_more’ also.

4. As continuous data I draw ‘price by duration (scatter plot 2)’. It was found that the expensive ticket disappears when the duration is extremely long. In the future, we will see if this number is affected by the class.

Q Provide analysis above and beyond the graphs and tables. Explain what the tables and visualizations tell you about the data.

1. Looking at the daysleft data, the earlier the reservation date, the cheaper the price. However, after 20 days, the number hardly changes.
2. Depending on the ‘stops’, the ones that did not stop at any time had a lower price than expected. Two or more stops were unexpectedly higher than the price distribution never stopped. I guess it’s related to the distance. I will find out later analysis.

[Summary]

In this analysis, flight price is target data and I will analyze the change of price by other factors. I set a business question that what is the best price strategy for airline company, and I will analyze ‘what is the best reservation strategy to get reasonable flight ticket’.

With an exploratory data analysis, mean of price is 20889.6 and lowest and highest price is 1105 and 123071. Sd of price is 22697.7. I confirm that ‘class’ is very important factor to decide flight price. I can see the price difference by airlines with flight class (boxplot 3). Daysleft has negative correlation with price (scatter plot1). When duration is extremely long, the expensive ticket disappears.

[Bibliography]

SHUBHAM BATHWAL. (2022). Flight price prediction. kaggle. Retrieved from https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

RDocumentation. (n.d.). Describe: Basic descriptive statistics useful for psychometrics. Retrieved from https://www.rdocumentation.org/packages/psych/versions/1.0-17/topics/describe

Roxanne Miller. (2018). 3 line table. Youtube. Retrieved from https://www.youtube.com/watch?v=pOwATBtIoCQ

Data Viz with Python and R. (2021, Feburary 5). Learn to make plots in python and r. Retrieved from https://datavizpyr.com/how-to-make-grouped-boxplot-with-jittered-data-points-in-ggplot2/

ggplot2. (n.d.). Reference lines: horizontal, vertical, and diagonal. Retrieved from https://ggplot2.tidyverse.org/reference/geom_abline.html

ggplot2. (n.d.). Jittered points. Retrieved from https://ggplot2.tidyverse.org/reference/geom_jitter.html

codersgram9. (2021, March 16). Change column name of a given DataFrame in R. geeksforgeeks Retrieved from https://www.geeksforgeeks.org/change-column-name-of-a-given-dataframe-in-r/

[Appendix]

```
#Install.Packages
install.packages(c("easypackages", "reshape", "reshape2",
"FSA","FSAdata","magrittr","dplyr","tidyr","plyr","tidyverse", "psych", "opticskxi", "car"))
library(easypackages)
libraries("FSA", "FSAdata", "magrittr","dplyr","tidyr","plyr","tidyverse", "ggplot2", "reshape",
"reshape2", "psych", "opticskxi", "car", "data.table")

#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop\\6010\\Module 2")

#Import dataset using read.csv()
fpp <- read.csv("Clean_Dataset.csv", stringsAsFactors = T,
                            header=T)
#Checking dataset & data structure
fpp
str(fpp)
headtail(fpp, 5)

#Change name of column
colnames(fpp)[which(names(fpp) == "source_city")] <- "from"
colnames(fpp)[which(names(fpp) == "destination_city")] <- "to"
colnames(fpp)[which(names(fpp) == "days_left")] <- "daysleft"

headtail(fpp,5)


opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
fpp.spi <- subset(fpp, airline == "SpiceJet",
                          select = c(airline, price))
boxplot(fpp.spi$price, col='#5E9DFF')
fivenum(fpp.spi$price)
text(y=fivenum(fpp.spi$price), labels=fivenum(fpp.spi$price), x=1.35, cex=1.5)
mtext("SpiceJet Price", side=3, line=1, cex=2)


options(max.print=999999)
table(fpp.spi$price)


#Cleaning data1 num.stops as nums
fpp$price[fpp$price == 34158 & fpp$airline == "SpiceJet"] <- NA

#Check it again
```

```
fpp.spi <- subset(fpp, airline == "SpiceJet",
                          select = c(airline, price))
table(fpp.spi$price)

#confirming cleaning in boxplot
opar <- par(no.readonly = TRUE)
par(fig=c(0.5, 1, 0, 1), new=TRUE)
boxplot(fpp.spi$price, col='#5E9DFF')
fivenum(fpp.spi$price)
text(y=fivenum(fpp.spi$price), labels=fivenum(fpp.spi$price), x=1.35, cex=1.5)
mtext("Cleaned SpiceJet Price", side=3, line=1, cex=2)
par(opar)

#add longitude data of source
table(fpp$from)
fpp$from_long[fpp$from == "Bangalore"] <- 77.710136
fpp$from_long[fpp$from == "Chennai"] <- 80.176506
fpp$from_long[fpp$from == "Delhi"] <- 77.100281
fpp$from_long[fpp$from == "Hyderabad"] <- 78.4235
fpp$from_long[fpp$from == "Kolkata"] <- 88.4379
fpp$from_long[fpp$from == "Mumbai"] <- 72.874245

#add longitude data of source
fpp$from_lati[fpp$from == "Bangalore"] <- 13.199379
fpp$from_lati[fpp$from == "Chennai"] <- 12.988166
fpp$from_lati[fpp$from == "Delhi"] <- 28.556160
fpp$from_lati[fpp$from == "Hyderabad"] <- 17.2373
fpp$from_lati[fpp$from == "Kolkata"] <- 22.6332
fpp$from_lati[fpp$from == "Mumbai"] <- 19.097403

#add longitude data of destination
fpp$to_long[fpp$to == "Bangalore"] <- 77.710136
fpp$to_long[fpp$to == "Chennai"] <- 80.176506
fpp$to_long[fpp$to == "Delhi"] <- 77.100281
fpp$to_long[fpp$to == "Hyderabad"] <- 78.4235
fpp$to_long[fpp$to == "Kolkata"] <- 88.4379
fpp$to_long[fpp$to == "Mumbai"] <- 72.874245

#add longitude data of source
fpp$to_lati[fpp$to == "Bangalore"] <- 13.199379
fpp$to_lati[fpp$to == "Chennai"] <- 12.988166
fpp$to_lati[fpp$to == "Delhi"] <- 28.556160
fpp$to_lati[fpp$to == "Hyderabad"] <- 17.2373
fpp$to_lati[fpp$to == "Kolkata"] <- 22.6332
fpp$to_lati[fpp$to == "Mumbai"] <- 19.097403

#calculate distance
```

```
fpp$distance <- (fpp$from_long-fpp$to_long)^2 + (fpp$from_lati-fpp$to_lati)^2
headtail(fpp,5)


# descriptive statistics tables with describe()
describe(fpp, skew=TRUE,range=TRUE)
unique(fpp$stops)
unique(fpp$class)
describe(price ~ airline, data=fpp)
describe(price ~ class, data=fpp)
describe(price ~ stops, data=fpp)
describe(price ~ days_left, data=fpp)

# pie chart
t <- count(fpp$class)
c.df <- t %>%
   group_by(x) %>% # Variable to be transformed
   count() %>%
   ungroup() %>%
   mutate(perc = `freq` / sum(`freq`)) %>%
   arrange(perc) %>%
   mutate(labels = scales::percent(perc))

colnames(c.df)[which(names(c.df) == "x")] <- "class"

c.df
ggplot(c.df, aes(x = "", y = perc, fill = class)) +
   geom_col(color = "black") +
   geom_label(size=8, aes(label = labels), color = c("white", "black"),
                    position = position_stack(vjust = 0.5),
                    show.legend = FALSE) +
   guides(fill = guide_legend(title = "class")) +
   scale_fill_viridis_d() +
   coord_polar(theta = "y") +
   theme_void()

#ggplot2 boxplot1
ggplot(fpp, aes(x=airline, y=price, color=airline)) +
   geom_boxplot()


fpp.air <- subset(fpp, airline == "AirAsia",
                           select = c(airline, price))
fpp.air

#boxplot 2
ggplot(fpp, aes(x=airline, y=price, color=class)) + geom_boxplot()
```

```
ggplot(fpp, aes(x=stops, y=price, color=stops)) + geom_boxplot()

ggplot(fpp, aes(x=class, y=price, color=class)) + geom_boxplot()


# additional analysis
ggplot(fpp, aes(x=duration, y=price, color=duration)) + geom_point()


# explatory analysis of duration
setDT(fpp)
df_dr <- fpp[ ,list(price.mean=mean(price)), by=duration]
df_dr
ggplot(df_dr, aes(x=duration, y=price.mean, color=duration)) + geom_point()


opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(fpp.air$price, col='#AF9A0A')
fivenum(fpp.air$price)
text(y=fivenum(fpp.air$price), labels=fivenum(fpp.air$price), x=1.35, cex=1.5)
mtext("AirAsia Price", side=3, line=1, cex=2)

opar <- par(no.readonly = TRUE)
par(fig=c(0.5, 1, 0, 1), new=TRUE)

# daysleft mean.price scatter
setDT(fpp)
df_dl <- fpp[ ,list(price.mean=mean(price)), by=daysleft]
df_dl
ggplot(df_dl, aes(x=daysleft, y=price.mean)) + geom_point()
str(fpp)
```