Flight Price R.Practice

Module2: descriptive statistics & plot, chart, abline

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi

Date: Nov 13th, 2022

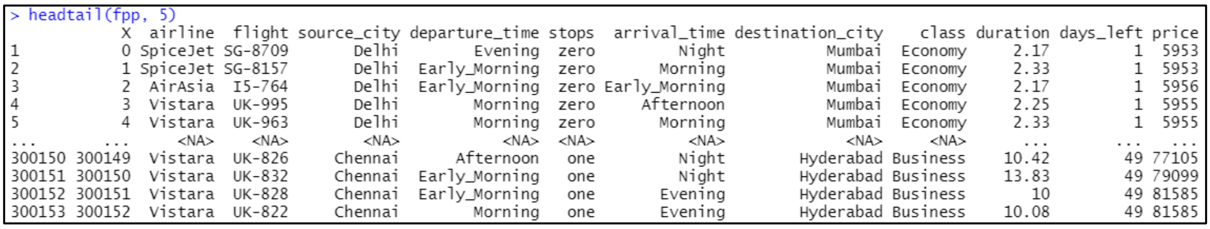
[Introduction]

Every data has its story. What analyst do is finding a story from data. An analyst must know the process of dealing with data to make story of data. Basically, analyst starts with what the data is. In this part, an analyst sees what variables is and what is the number and text in it. And then, analyzing statistical analysis of descriptive. If data cleaning is needed, analyst can do that with explanation. And make questions and hypothesis.

In this module 2 R practice, I will produce several descriptive statistics tables thinking about three-line tables. And using par(), and abline() I will produce scatter chart, jitter chart, and boxplot chart which can be said one of the basic plots. And then, talk about what is specific usage of these visualization charts.

[Statistical Analysis of Descriptive]

Table 1. The Three-line table of headtail 5



```
> 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 about outline or overview of data what’s in the data. I choose target varible is price. I will predict price according to other variables. Duration, days_left and price are numberic data. There are different airlines, class, stops, city, and time.

Table 2. pysco::describe of numeric data

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
days_left	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

- ✓ Statistical analysis of descriptive. Mean of days_left is 26 and range is 48. Lowest and highest price is 1105 and 123071. Sd of price is 22697.7. It looks dispersion is high comparing with others.

Table 3. Price by Airline (Group data 1)

airline	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Air_India	1	80892	23507.02	20905.12	11520	21636.28	10892.66	1525	90970	89444	0.58	-1.36	73.5
AirAsia	1	16098	4091.07	2824.06	3276	3568.31	1553.76	1105	31917	30812	3.3	16.83	22.26
Go_FIRST	1	23173	5652.01	2513.87	5336	5368.09	1570.07	1105	32803	31698	2.68	14.53	16.51
Indigo	1	43120	5324.22	3268.89	4453	4815.09	2226.87	1105	31952	30847	1.98	5.79	15.74
SpiceJet	1	9011	6179.28	2999.63	5654	5757.33	2351.4	1106	34158	33052	2.13	7.28	31.6
Vistara	1	127859	30396.54	25637.16	15543	28428.6	16986.15	1714	123071	121357	0.42	-1.39	71.7

- ✓ Table3: Air_India and Vistara has high mean of price over 20,000. Other arilines’ mean is around 5,000. And min price of other airlines is 1105 and max is 34158. There is huge difference between categorized airlines Air_India & Vistara vs Others.

Table 4. Price by Class (Group data 2)

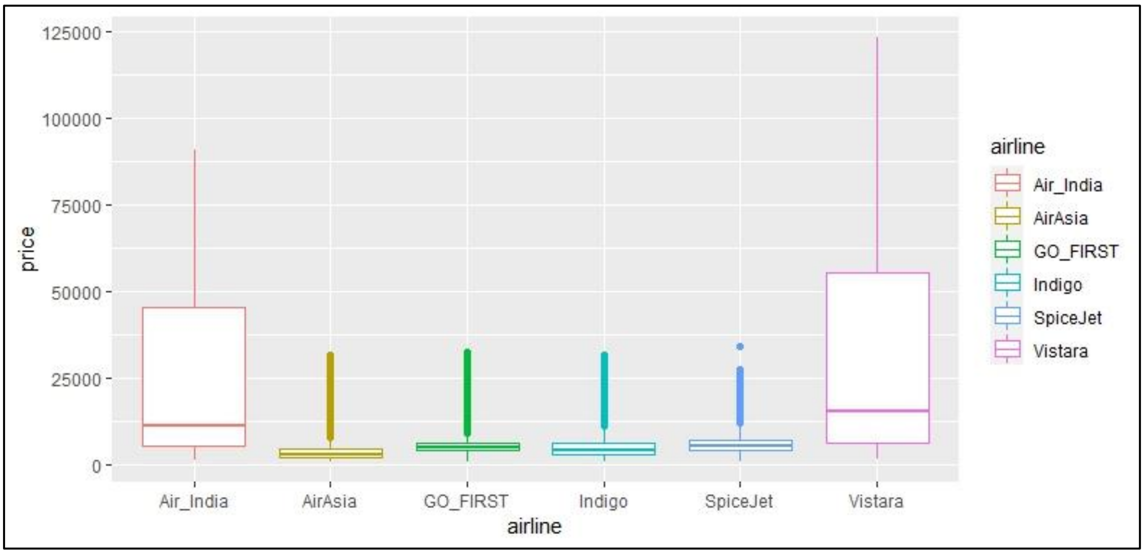
x	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Business	1	93487	52540.08	12969.31	53164	52827.86	10957.9	12000	123071	111071	-0.1	0.81	42.42
Economy	1	206666	6572.34	3743.52	5772	6050.55	2521.9	1105	42349	41244	1.7	4.17	8.23

- ✓ Table4: According to class, difference between mean of business and economy is 45,965.74 and sd difference is also huge. Sd of business is 12969.31 and sd of economy is 3743.52.

Q What is a three-line table format that is commonly used in white papers?
Three-line table is all tables should have the following three horizontal lines: one under the title, above the column headings. One between the column headings and the body of the table. One at the the bottom of the table (Shadi, Bartsch-Zimmer, n.d). Three-line tables (above, table 2~4) looks clearer than the normal table (table 1).

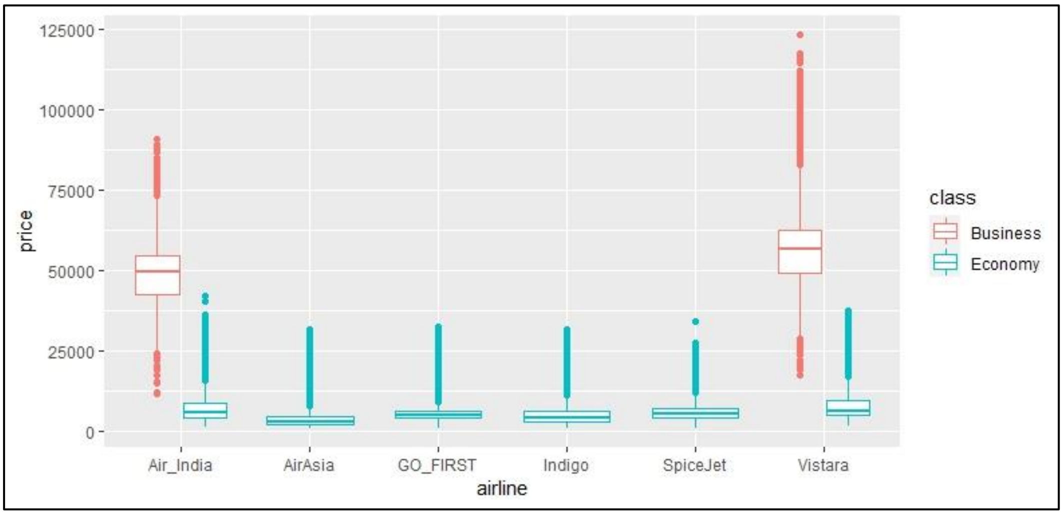
[Boxplot]

Boxplot 1. Price by Airline



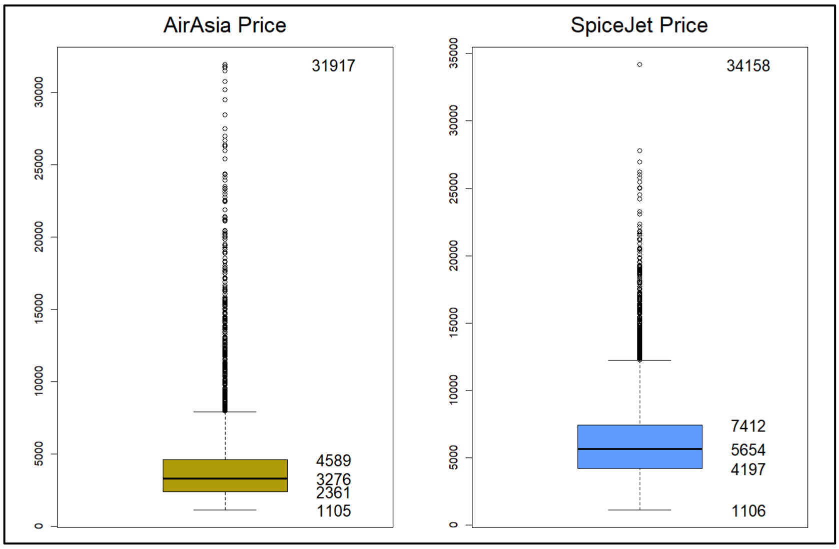
- ✓ Plot 1: As saw in the descriptive analysis, we can see huge difference between categorized airlines (Air_India & Vistara vs AirAsia, Go_FIRST, Indigo and SpiceJet).

Boxplot 2. Price by Airline adding class



- ✓ Plot 2: 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.

Boxplot 3. More detail of 2 airline about outliers



- ✓ Plot 3: Using par(), we can compare the difference between two airlines which seems having outlier in boxplot 2. And I decide the price of 34,158 from SpiceJet as outlier, because it is too far from other prices. However, the price 31,917 from AirAsia isn't too far from other prices, although they are out of boxplot range, I will leave this.

[Boxplot to detect outliers]

Q <u>What is outlier?</u>

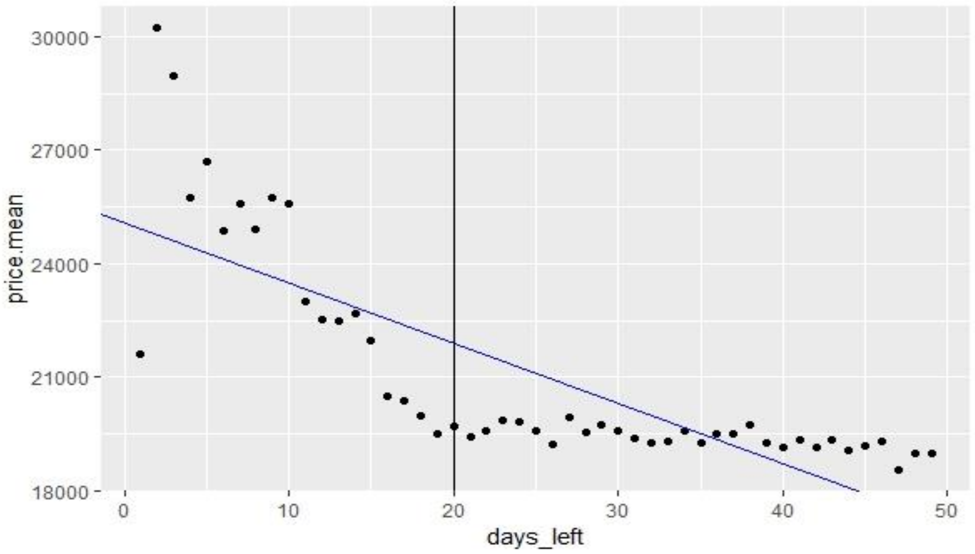
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population (NIST, n.d.). Summarizing what we have learned so far, the researcher can decide what is the outliers in specific dataset according to goal of analyzing dataset. In this analysis I want to know what factors affect on flight price, so I tried to include almost every data from dataset excepting abnormally far away from other data.

Q <u>How can you use boxplots to detect outliers?</u>

First draw a boxplot of graph. Second see the five numbers of boxplot. We can see interquartile range and Q1- 1.5*IQR & Q3+1.5*IQR, and this can be comfort zone. And I see the out of box and see the continuity. If data is continuous until the end of outliers, it would be not outliers. However, as mentioned above answer, every decision is according to goal of analysis. After deleting the outliers, the researchers must add explanations about that to prevent bias.

[Scatter plot]

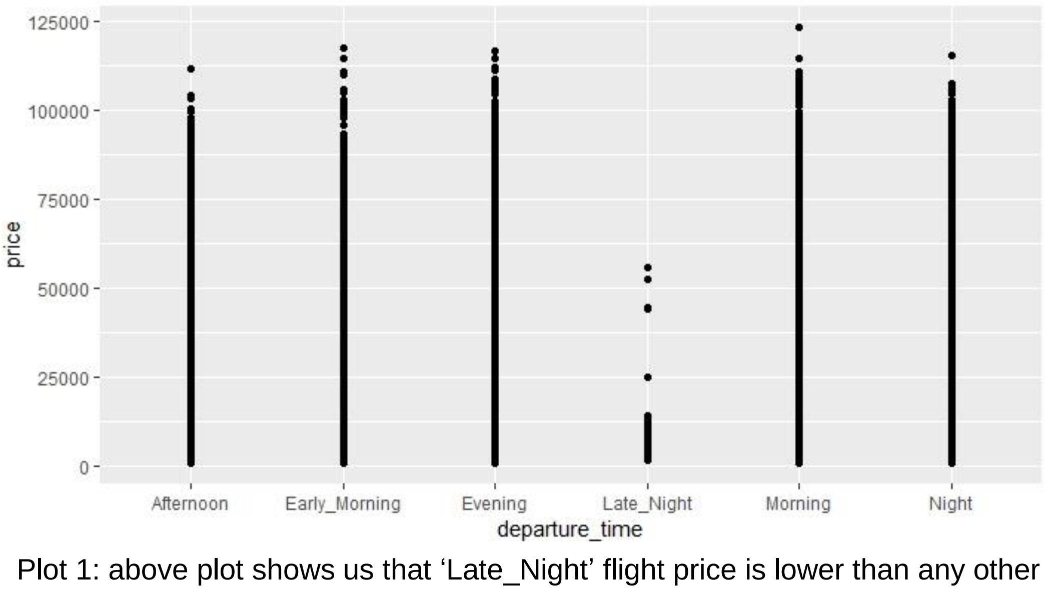
Scatter plot 1. Price.mean by days_left with abline



- ✓ Plot 1: Using abline(), we can find the correlation between days_left and price.mean by daysleft. There are 300,152 data in this dataset. It’s too many, so it’s clearer to see in the mean by day_lefts. 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.

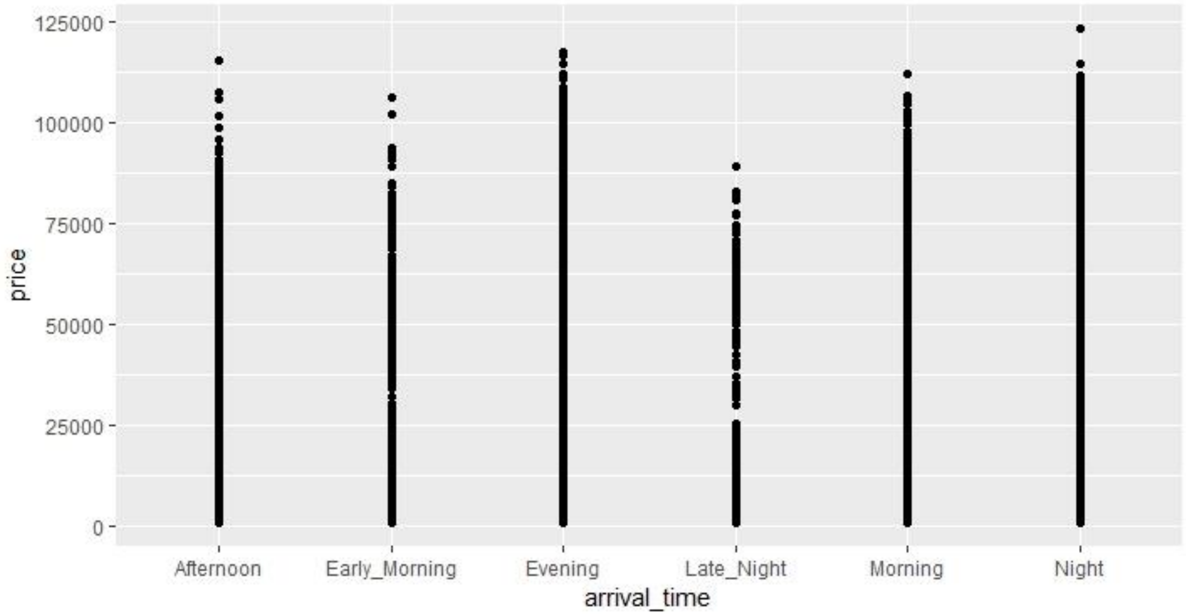
[Jitter plot]

Jitter plot 1. Price by departure_time



- ✓ Plot 1: above plot shows us that ‘Late_Night’ flight price is lower than any other departure time. But there is just little difference of highest price between other departure time.

Jitter plot2. Price by arrival_time



✓ Plot 2: above plot shows us that also ‘Late_Night’ flight price is lower than any other arrival time, but the difference looks smaller than departure time. We must check this difference between price according to departure time and arrival time.

Q <u>When do you use jitter chart?</u>
Primarily used to look at categorical variables. I think it's a good way to see how the dots are grouped according to a category. In the data above we can see price difference by departure time and arrival time.

[interpretation]

1. The above data is flight ticket prices in India.
2. With statistical descriptive analysis we can see what the important data is and what is in the data. And we can decide what we will going to do with this data.
3. There are a total of 6 airlines, 2 airlines have high flight prices, and the other 4 airlines are lower and similar. The reason why the 2 airlines have high flight prices is that only two airlines operate business class. We can see this with categorical data analysis.
4. Into more detail part, all six airlines have similar prices for seats except business class, but the two airlines which have business class has have higher prices on average even in the economy class.
5. We worked to find outliers through boxplots. In this case continuous data were excluded from outliers. Because I want to include as much as possible.
6. The correlation between days_left and the price had a negative correlation, but the price did not change significantly when more than 20 days were left.
7. According to departure time and arrival time, the price was the lowest when arriving or departing late night, but there was a significant difference between the arrival and departure times even in the late evening.

[Summary]

From this data I learned basic data analysis with description and charts. I learned how to visualize the numerical data in a clearly visible form. The method is to use a Three-line table. I've seen it in advanced reports before, but this is the first time I've used it.

Regarding data visualization, I was able to confirm the above-mentioned interpretation. I was able to figure out what kind of graph to draw and how to analyze through numerical analysis, and I think about how to identify outliers through box plots. In addition, through the scatter plot, it was possible to confirm what is the overview of data. And through the jitter plot, I confirmed what kind of data was categorized. I am constantly learning which type of graph to use depending on the data.

[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

Shadi, Bartsch-Zimmer. (n.d.). Know: a Journal on the formation of knowledge. The university of chicago press journals. Retrieved from https://www.journals.uchicago.edu/journals/know/prep-table

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

NIST. (n.d.). What are outliers in the data?
. Retrieved from https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

[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("Flight Price of India.csv", stringsAsFactors = T,
                header=T)

#Checking dataset & data structure
fpp
str(fpp)
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)

#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()

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

```
fpp.air
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)

#making boxplot3
fpp.vis <- subset(fpp, airline == "Vistara",
                  select = c(airline, price))

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

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)
par(opar)

#checking days_left scatter
p <- ggplot(fpp, aes(duration, price))
p + geom_jitter()

# days_left mean.price scatter
setDT(fpp)
df_dl <- fpp[ ,list(price.mean=mean(price)), by=days_left]
df_dl
coef(lm(price.mean ~ days_left, data = df_dl))
p <- ggplot(df_dl, aes(x=days_left, y=price.mean)) + geom_point()+abline()
p + geom_abline(intercept = 25100.1626, slope = -159.3565, color="blue")+
geom_vline(xintercept = 20)
str(fpp)

# jitter Plot
ggplot(fpp, aes(x=departure_time, y=price)) +geom_point()
ggplot(fpp, aes(x=arrival_time, y=price)) +geom_point()
```