Analysis for delivery picking time

Module1: Importing & Cleaning & Rudimentary analysis

ALY6010:

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi

Date: Nov 6th, 2022

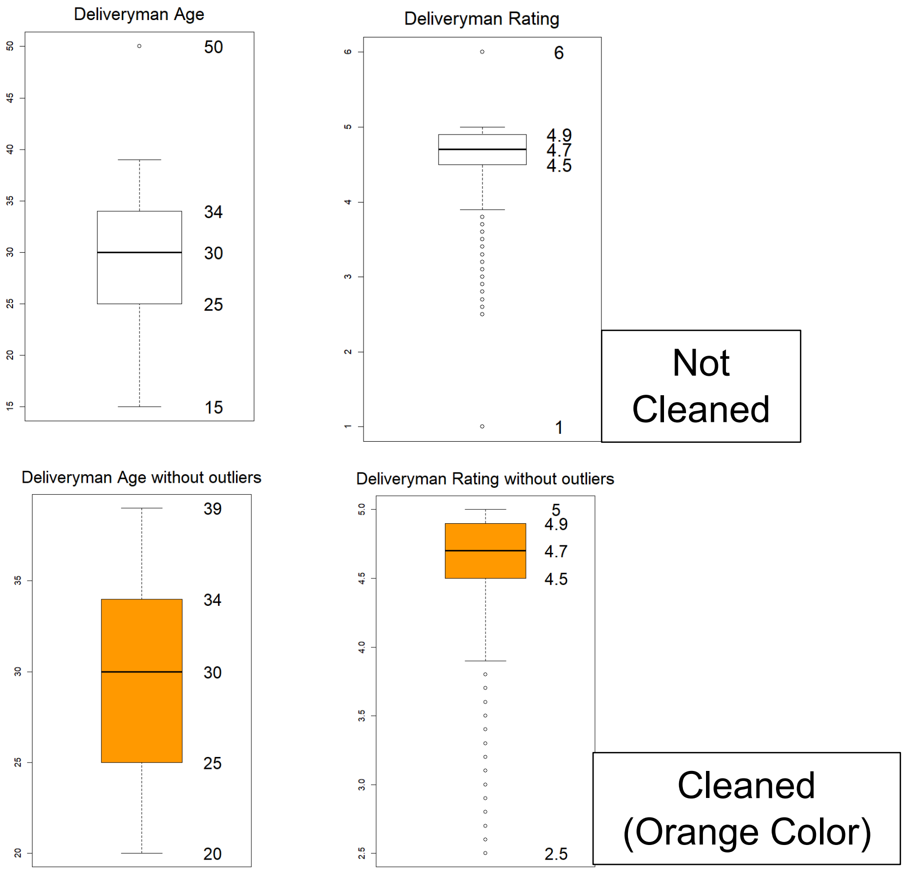
[Introduction]

This is about analysis using Amazon's delivery man & condition data from the Kaggle(https://www.kaggle.com/datasets/vikramxd/amazon-business-research-analyst-dataset/code.)
The question I'm looking for in this data is picking time (ordered time-order_picked_time) changing according to other factors such as the driver's rating and external conditions such as weather, traffic conditions, and distance to the delivery destination. Finally, I want to decide 'what is important factor for fast pick up'

[Cleaning data]
- changing table names to simple one (ex. 'Delivery_person_Age' to 'Deliveryman_Age')
- 'NaN' value to NA (as missing value in R)
- Finding extreme outliers or incomplete data or just wrong data in the all column.
- Delete latitude & longitude data which is 0 or below 0 before calculating accurate distance.
- Finding Drivers' age outliers in the data. [age=50 or age=15], because rating is abnorminal. Changing rating value of age=50 or 15 as NA value
- Making factor (10:50) as time data(%H:%M) which express time.

[Basic of data]
- Basic information of data: 11399 observations of 20 variables
- Deliveryman information: ID (num & factor), age, ratings
- Location of restaurant and delivery place as longitude & latitude.
- Time of ordered and time of picked each as time (5 minute increments)
- Condition of others: Weather, Traffic, Vehicle condition, type of Vehicle, type of order (snacks or drinks), multiple deliveries, city type, etc.

[Rudimentary Analysis1: Boxplots of Drivers' age & Rating]

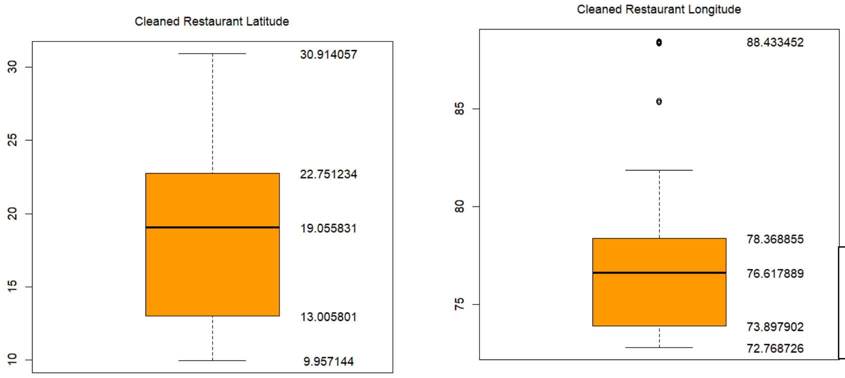


[Picked Time – Ordered Time]

Minutes	Counts
5	2961
10	2986
15	2933

Using as.POSIXct(amazon$Time_Orderd, format="%H:%M"), Make 10:50 as time and Calculate 'Picked time – Order Time' to know how much time take to piked up by delivery man after order made

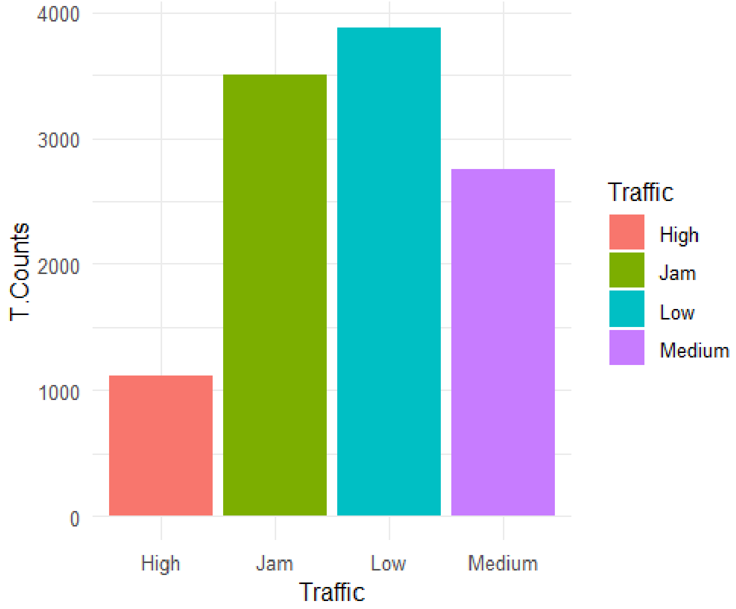
[Rudimentary Analysis2: Boxplots of Cleaned Restaurant Longitude & Latitude]



Cleaned
(Orange Color)

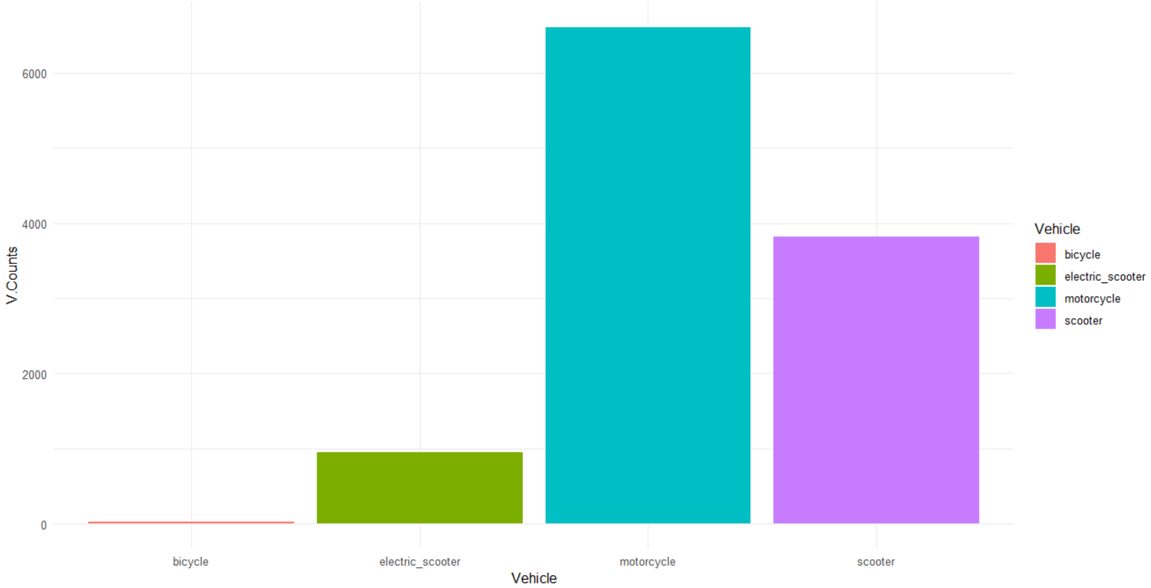
<u>Distance Formula</u>= *[Restaurant(Latitude) – Delivery Location(Latitude)]^2 +*
[Restaurant(Longitude) – Delivery Location(Longitude)]^2
Mean of not cleaned Distance^2 is 117, but after cleaning minus values and 0 values
Mean of cleaned Distance^2 is 127.

[Rudimentary Analysis3: Barplot of Traffic]

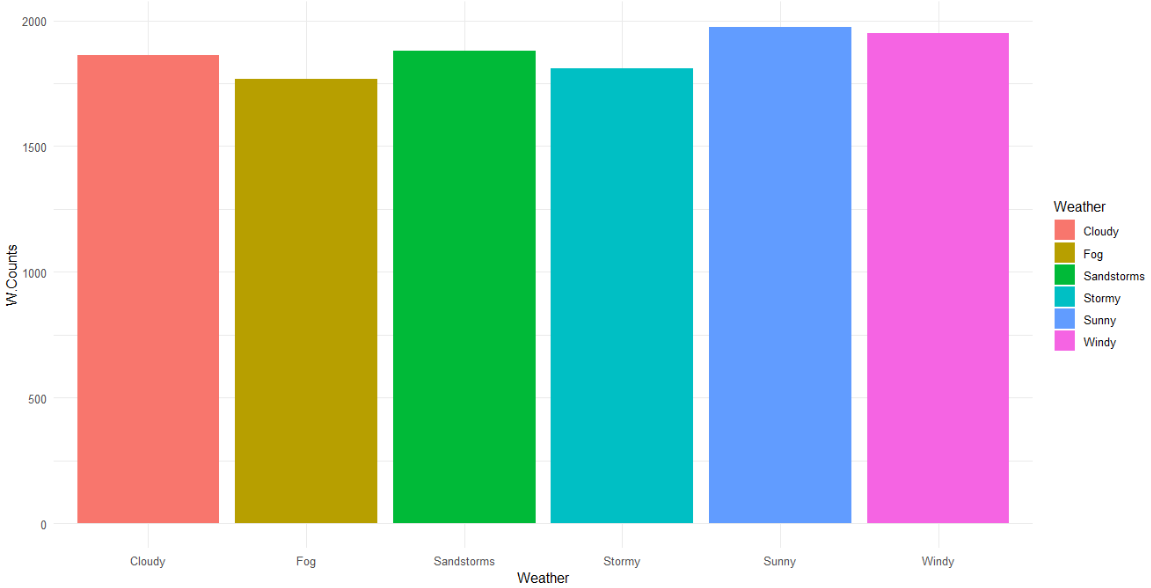


Traffic: Low(3881) > Jam(3503) > Medium(2751) > High(1110)

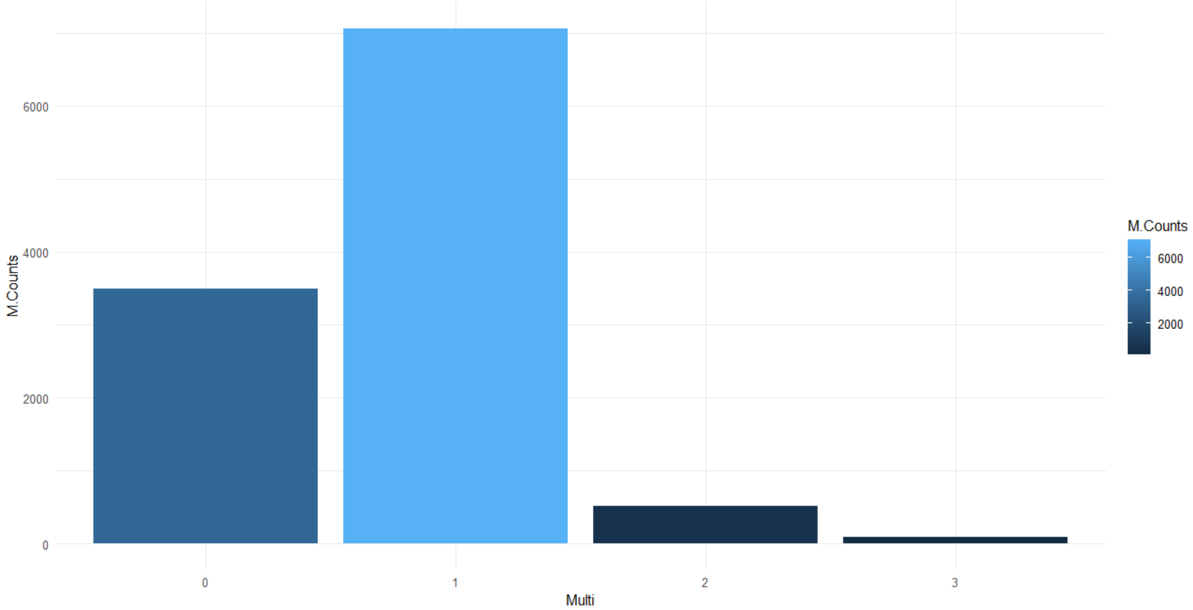
[Rudimentary Analysis3: Barplot of Vehicle, Weather and Multi Counts]



Vehicle: motorcycle(6609) > scooter(3817) > electric_scooter(950) > bicycle(23)



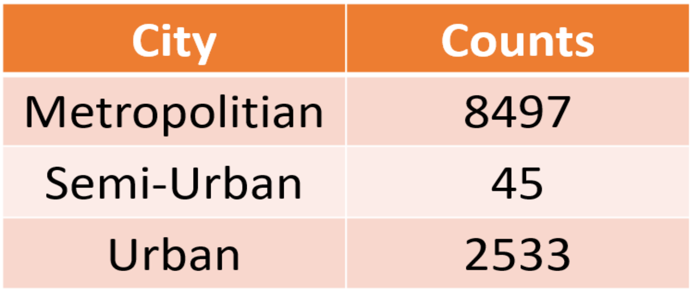
Weather(Most): Sunny(1975), Windy(1948)
Weather(Least): Fog(1768), Stormy(1811)



Multi Counts: 1(7065) > 0(3491) > 2(513) > 3(92)

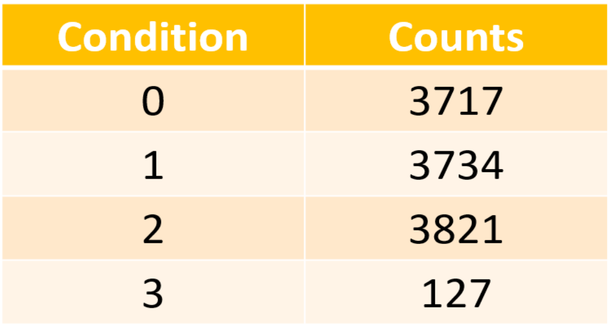
[Rudimentary Analysis4: Tables]

[Type of City]



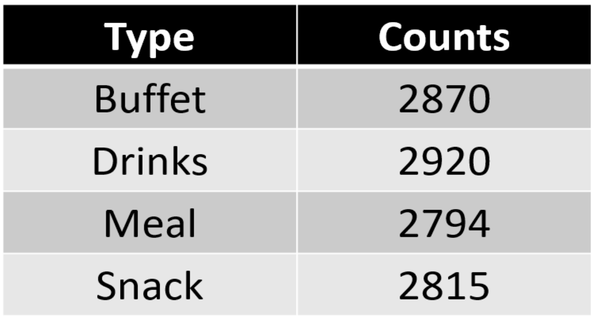
City	Counts
Metropolitian	8497
Semi-Urban	45
Urban	2533

[Vehicle Condition]



Condition	Counts
0	3717
1	3734
2	3821
3	127

[Type of Order]



Type	Counts
Buffet	2870
Drinks	2920
Meal	2794
Snack	2815

[Question: How do you show and compare the results across multiple classes?]

I would like to test which factors affect the pickup time on a differtime(picked time-ordered time) basis. Therefore, I display different impacting results based on differtime (picked time-ordered time) times and will analyze this data in the future. For example, I will assume that higher-rated deliveryman will have shorter pick-up times, and see where the pick-up times are distributed with rating.

[Summary]

This data is about the deliveryman(age & rating) and the factors affecting delivery pick up time. I think most important data is the rating of the deliveryman, and the differtime(picked time-ordered time). I performed a cleaning data that removes age 50 and 15 by setting them as outliers. The results of cleaning data are shown below in rudimentary analysis1. All data came from Amazon, so the color used for the boxplot referenced the Amazon orange color ('#FF9900') on the Amazon website(https://usbrandcolors.com/amazon-colors/).

First, the differtime(picked time-ordered time) is counted in units of 5 minutes, 10 minutes, and 15 minutes. Since all elements have similar counts, I thought it would be appropriate to compare them with other factors.

The second is the restaurant's Longititude and Latitude. I want to calculate the distance by comparing this to the Longititude and Latitude of the delivery location. For this purpose, cleaning was performed to remove data with Longititude and Latitude of 0 and minus. Based on this, I want to create a new attribute by calculating the Distance with **distance formula** in page2.

The third is a summary of factors that are likely to affect delivery pick up time. The bar plots are extracted using ggplot2. In the case of traffic, it appeared in the order of Low > Jam > Medium > High. In the case of Vehicle, motocycle > scooter > electric_scooter > bicycle appeared in the order. As for the weather, Sunny and Windy had the most, and Fog and Stormy had the least. In the case of multi delivery, the order was 1 > 0 > 2 > 3. It is found that multi delivery was the most common with 1.

Finally, the tables. In the case of city types, most of them appeared as Metropolitian and Urban. The condition of vehicle is mostly 0, 1, and 2, and the number is similar. Type of order like buffet, drinks, etc. also showed similar numbers.

Dataset: https://www.kaggle.com/datasets/vikramxd/amazon-business-research-analyst-dataset/code

REFERENCE

```
#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)
```