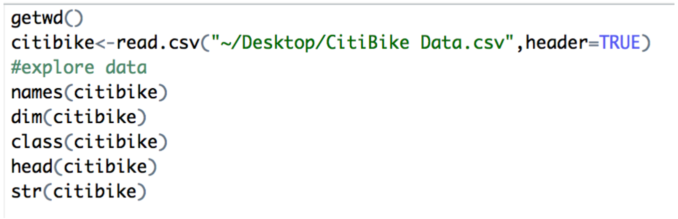
Part 1: Citibike Descriptive Analytics

Before doing analysis, the basic description of input data cannot be omitted.

• Brief Description about Citibike Data

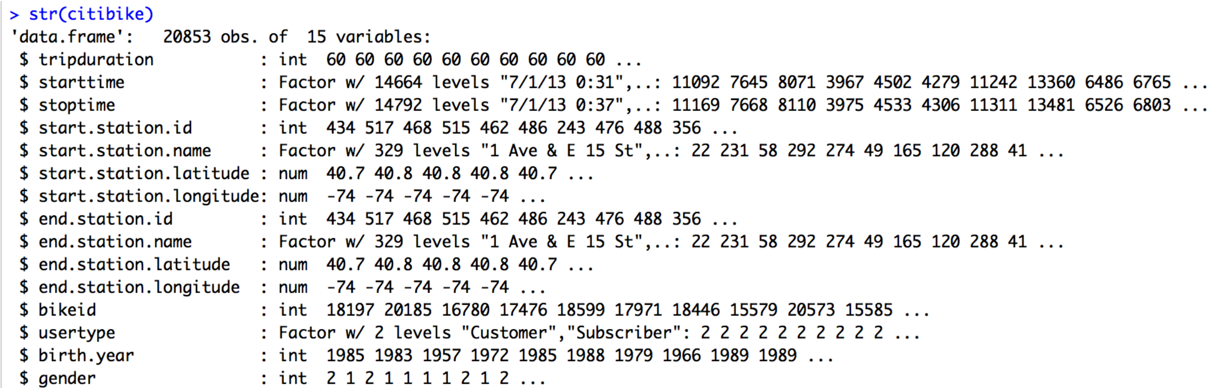
Identified the path of “CitiBike Data.csv” file, and got a big picture of it.



```
getwd()
citibike<-read.csv("~/Desktop/CitiBike Data.csv",header=TRUE)
#explore data
names(citibike)
dim(citibike)
class(citibike)
head(citibike)
str(citibike)
```

So, based on the result, we could grab some features about the citibike data,

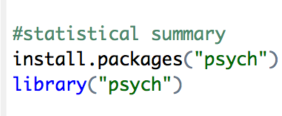
- The data frame dimension is 20853*15, which means the table includes 20853 records(observations) and 15 attributes(variables).
- For the whole data structure such as variable name, variable type, we could get a brief picture which I represented below,



```
> str(citibike)
'data.frame':	20853 obs. of  15 variables:
 $ tripduration           : int  60 60 60 60 60 60 60 60 60 60 ...
 $ starttime              : Factor w/ 14664 levels "7/1/13 0:31",..: 11092 7645 8071 3967 4502 4279 11242 13360 6486 6765 ...
 $ stoptime               : Factor w/ 14792 levels "7/1/13 0:37",..: 11169 7668 8110 3975 4533 4306 11311 13481 6526 6803 ...
 $ start.station.id       : int  434 517 468 515 462 486 243 476 488 356 ...
 $ start.station.name     : Factor w/ 329 levels "1 Ave & E 15 St",..: 22 231 58 292 274 49 165 120 288 41 ...
 $ start.station.latitude : num  40.7 40.8 40.8 40.8 40.7 ...
 $ start.station.longitude: num  -74 -74 -74 -74 -74 ...
 $ end.station.id         : int  434 517 468 515 462 486 243 476 488 356 ...
 $ end.station.name       : Factor w/ 329 levels "1 Ave & E 15 St",..: 22 231 58 292 274 49 165 120 288 41 ...
 $ end.station.latitude   : num  40.7 40.8 40.8 40.8 40.7 ...
 $ end.station.longitude  : num  -74 -74 -74 -74 -74 ...
 $ bikeid                 : int  18197 20185 16780 17476 18599 17971 18446 15579 20573 15585 ...
 $ usertype               : Factor w/ 2 levels "Customer","Subscriber": 2 2 2 2 2 2 2 2 2 2 ...
 $ birth.year             : int  1985 1983 1957 1972 1985 1988 1979 1966 1989 1989 ...
 $ gender                 : int  2 1 2 1 1 1 1 2 1 2 ...
```

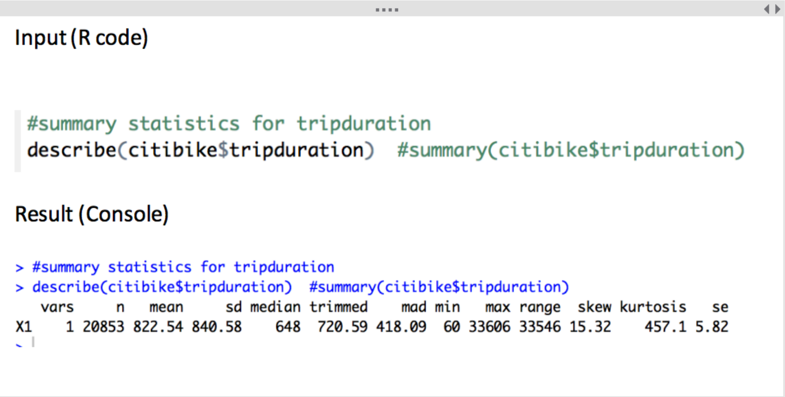
• Analytics Questions

I adopted the **describe function** from **“psych”** package for the statistical summary.



```
#statistical summary
install.packages("psych")
library("psych")
```

1.Compute summary statistics for tripduration.



Input (R code)

```
#summary statistics for tripduration
describe(citibike$tripduration)  #summary(citibike$tripduration)
```

Result (Console)

```
> #summary statistics for tripduration
> describe(citibike$tripduration)  #summary(citibike$tripduration)
   vars     n   mean     sd median trimmed    mad min   max range  skew kurtosis   se
X1    1 20853 822.54 840.58    648  720.59 418.09  60 33606 33546 15.32    457.1 5.82
```

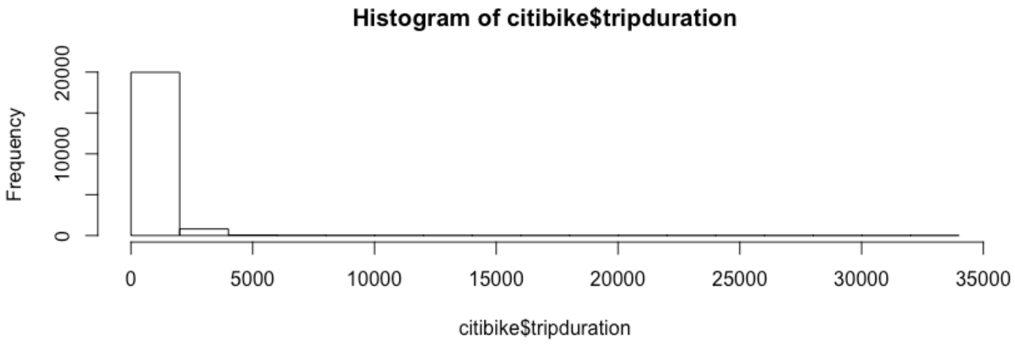
Summary

As we could see on trip duration,

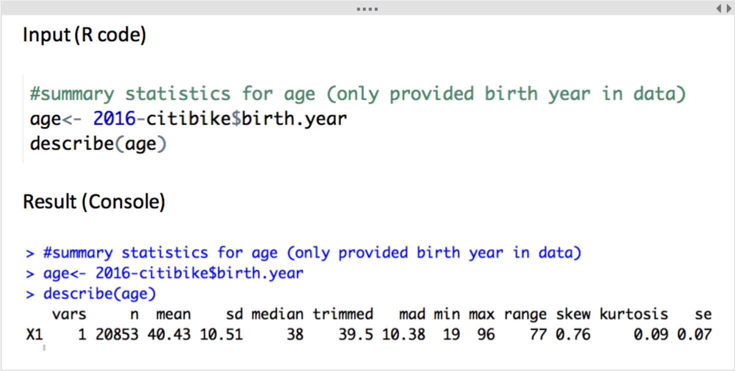
[V_{sd} is very large]: The duration time are widely spread out and far from mean value.

[V_{range} is very large]: The range of duration time is huge.

[$V_{kurtosis}$ and V_{skew} are very large]: Most people spent the similar tripduration time, which is much smaller than the median of tripduration.



2.Compute summary statistics for age



Input (R code)

```
#summary statistics for age (only provided birth year in data)
age<- 2016-citibike$birth.year
describe(age)
```

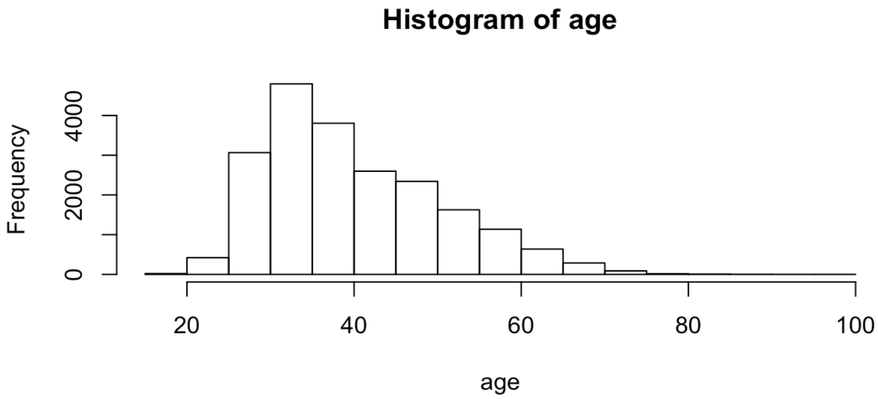
Result (Console)

```
> #summary statistics for age (only provided birth year in data)
> age<- 2016-citibike$birth.year
> describe(age)
   vars     n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 20853 40.43 10.51     38    39.5 10.38  19  96    77 0.76     0.09 0.07
```

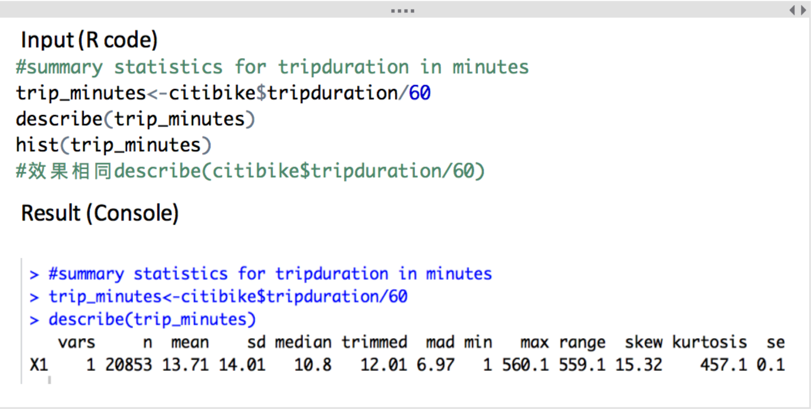
Summary

As we could see on age,

[$V_{kurtosis}$ and V_{skew} are small]: the age of riders is spread out, which includes different generations. In addition, comparative frequent riders dropped in the younger generation.



3.Compute summary statistics for tripduration in minutes (Need to transform tripduration from seconds to minutes)



Input (R code)

```
#summary statistics for tripduration in minutes
trip_minutes<-citibike$tripduration/60
describe(trip_minutes)
hist(trip_minutes)
#效果相同describe(citibike$tripduration/60)
```

Result (Console)

```
> #summary statistics for tripduration in minutes
> trip_minutes<-citibike$tripduration/60
> describe(trip_minutes)
   vars     n  mean    sd median trimmed  mad min   max range  skew kurtosis  se
X1    1 20853 13.71 14.01   10.8   12.01 6.97   1 560.1 559.1 15.32    457.1 0.1
```

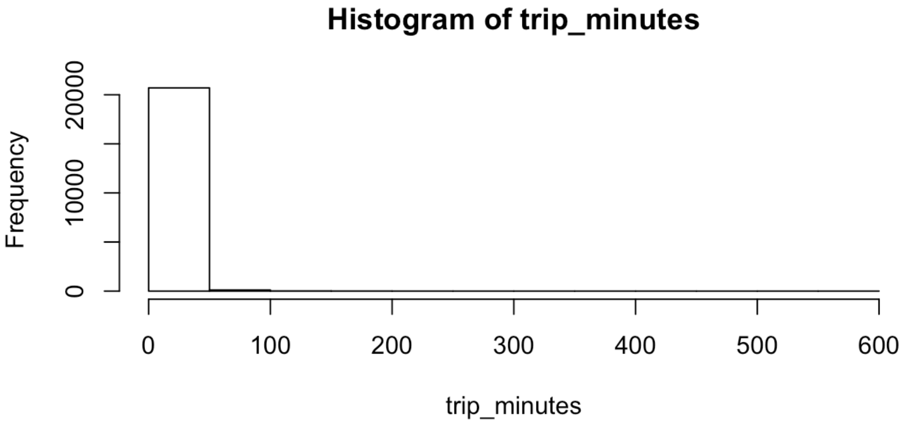
Summary

The same as tripduration in seconds,

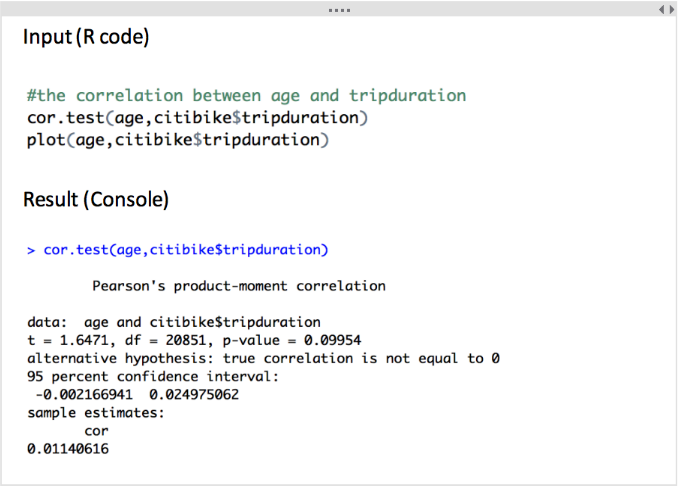
[V_{sd} is very large]: The duration time are widely spread out and far from mean value.

[V_{range} is very large]: The range of duration time is huge.

[$V_{kurtosis}$ and V_{skew} are very large]: Most people spent the similar tripduration time, which is much smaller than the median of tripduration.



4. Compute the correlation between age and tripduration



Input (R code)

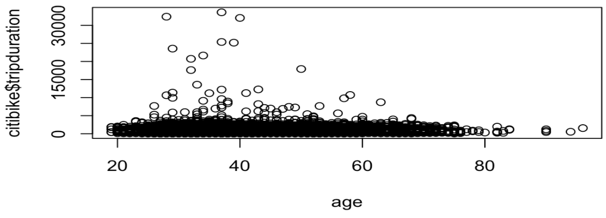
```
#the correlation between age and tripduration
cor.test(age,citibike$tripduration)
plot(age,citibike$tripduration)
```

Result (Console)

```
> cor.test(age,citibike$tripduration)

        Pearson's product-moment correlation

data:  age and citibike$tripduration
t = 1.6471, df = 20851, p-value = 0.09954
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.002166941  0.024975062
sample estimates:
       cor
0.01140616
```

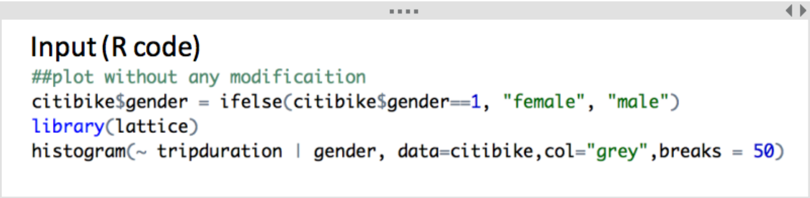
Summary

As we see the p-values is larger than 0.05, the variable age and tripduration are not strongly correlated. Below is the scatter plot (age~tripduration),which we could also see that correlation directly,



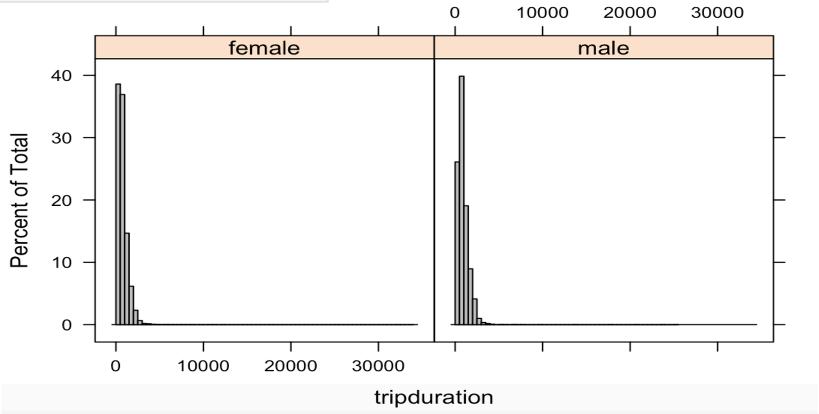
5. Plot the histograms and box plots for tripduration by gender

As we did the statistical summary above, the tripduration value is extremely huge, varied, and densely distributed with extreme value, which means direct plot cannot represent clearly. As we could see the histogram with original tripduration value below.

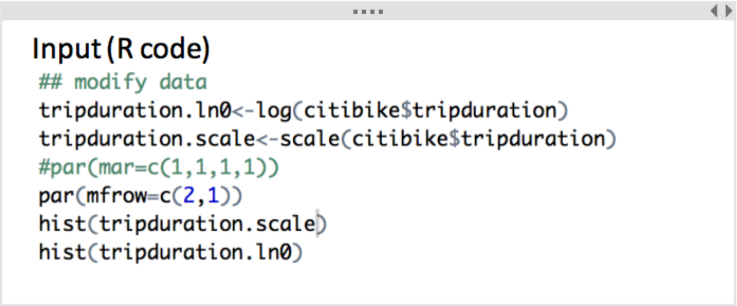


Input (R code)

```
##plot without any modificaition
citibike$gender = ifelse(citibike$gender==1, "female", "male")
library(lattice)
histogram(~ tripduration | gender, data=citibike,col="grey",breaks = 50)
```

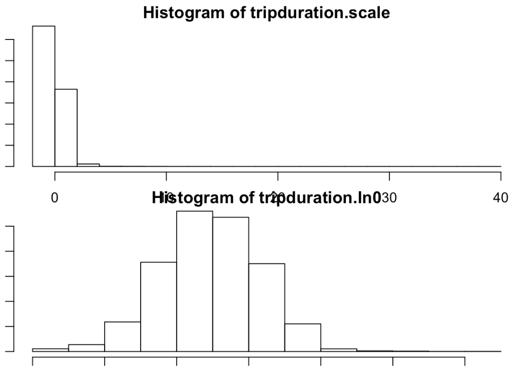


So, before plot any graphs, first we need to modify the data, and I used log()method not scale()method, because the former one could make the histogram closer to the normal distribution.



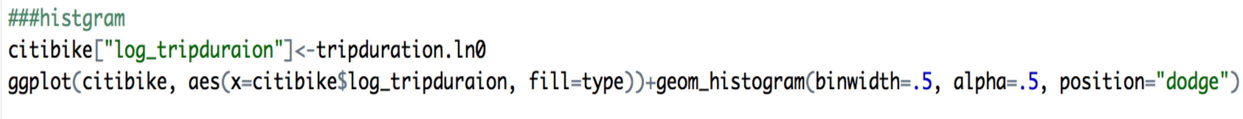
Input (R code)

```
## modify data
tripduration.ln0<-log(citibike$tripduration)
tripduration.scale<-scale(citibike$tripduration)
#par(mar=c(1,1,1,1))
par(mfrow=c(2,1))
hist(tripduration.scale)
hist(tripduration.ln0)
```

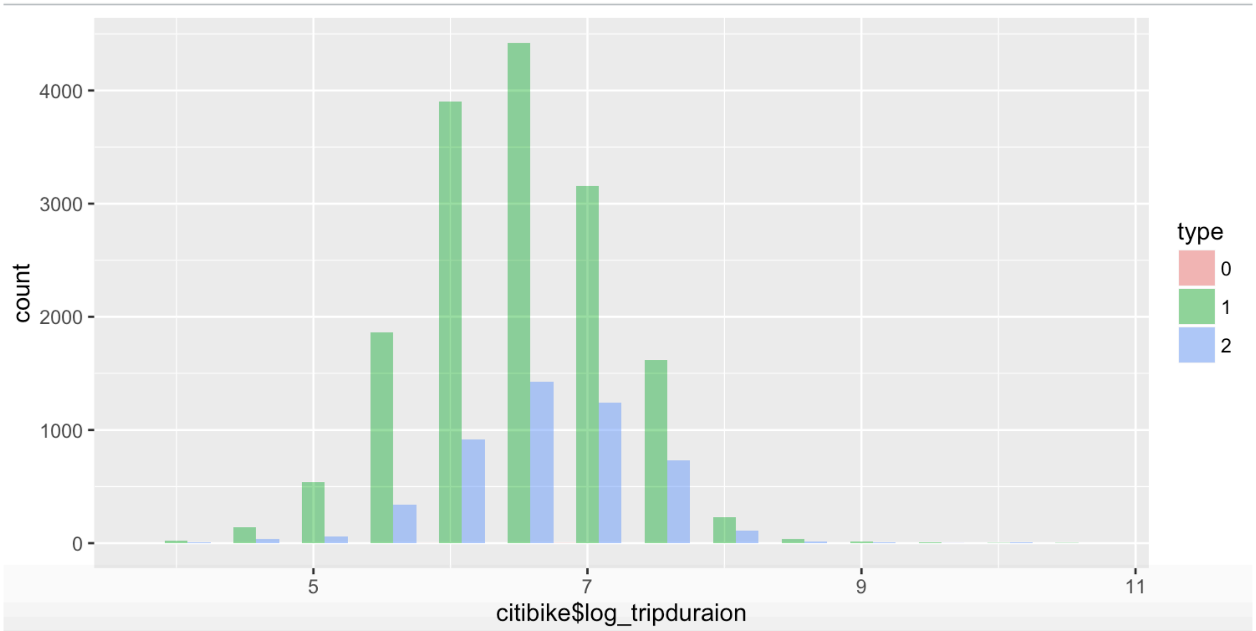


Based on the modified tripduration data, I plot the graphs based on the ggplot package and split method,

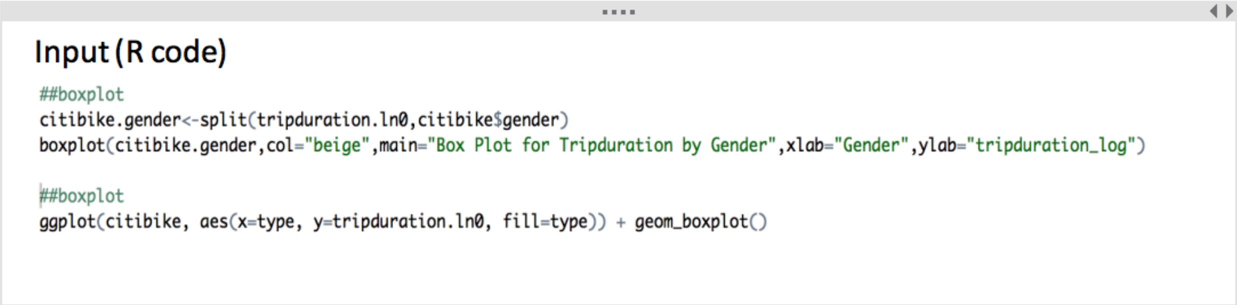
For the histogram, I used the ggplot package. As I used the modified tripduration data based on log method, I need to add this variable to the citibike dataframe. So all the code is represented below,



```
###histgram
citibike["log_tripduraion"]<-tripduration.ln0
ggplot(citibike, aes(x=citibike$log_tripduraion, fill=type))+geom_histogram(binwidth=.5, alpha=.5, position="dodge")
```

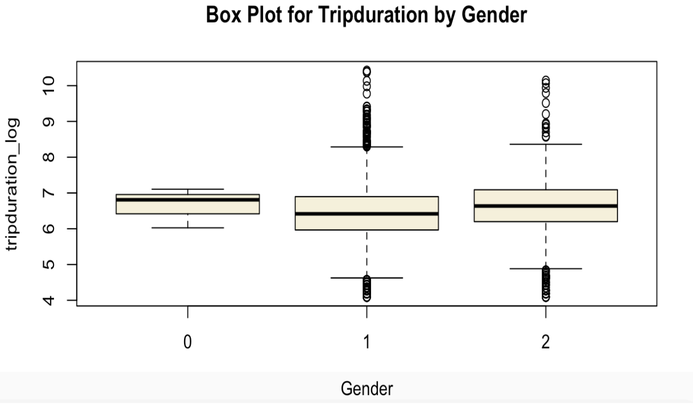


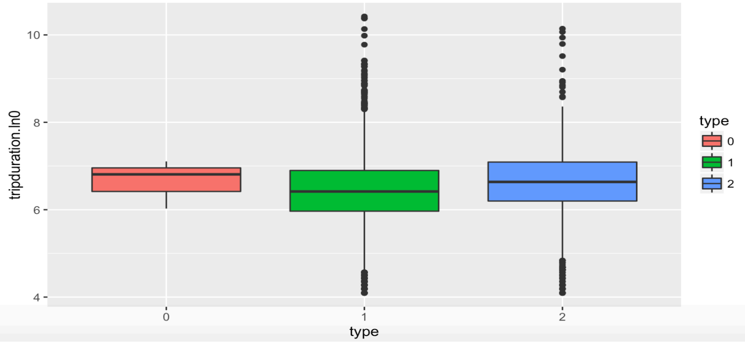
For the boxplot, I used two ways to represent: one is based on the split function, another is based on the ggplot package.



Input (R code)

```
##boxplot
citibike.gender<-split(tripduration.ln0,citibike$gender)
boxplot(citibike.gender,col="beige",main="Box Plot for Tripduration by Gender",xlab="Gender",ylab="tripduration_log")

##boxplot
ggplot(citibike, aes(x=type, y=tripduration.ln0, fill=type)) + geom_boxplot()
```





- ## Business Questions

1. What is the total revenue assuming all users riding bikes from 0 to 45 minutes pay $3 per ride and user exceeding 45 minutes pay an additional $2 per ride?

Method

In order to get the total revenue, we need to know each trips price and duration time:

$$Total\ Revenue = \sum Price(trip\ i)$$

For the Price, we could build a function on its trip duration time,

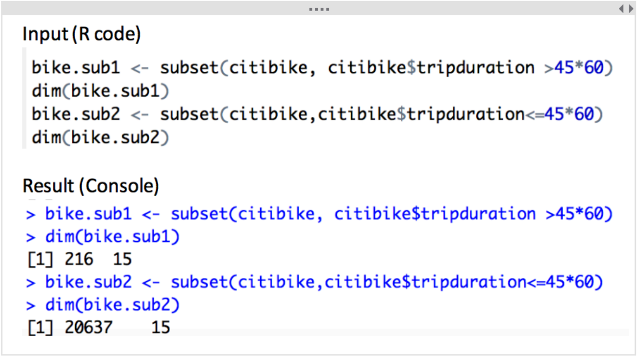
$$Price(i) = \begin{cases} 3 & (tripduration \leq 45min) \\ 5 & (tripduration > 45min) \end{cases}$$

For the tripduration time, I use the subset() function to separate the data into two groups, and use the dim()function to get how much trips in the group which duration time is larger than 45 minutes (45*60 seconds) or smaller. So,

$$Total\ Revenue = \sum Price(if\ tripDurantion\ i \leq 45 * 60) + \sum Price\ (if\ tripDuration\ i > 45 * 60\ )$$

Implementation

Use the R code to get the two subsets of data and its corresponding total number of trips,



Input (R code)

```
bike.sub1 <- subset(citibike, citibike$tripduration >45*60)
dim(bike.sub1)
bike.sub2 <- subset(citibike,citibike$tripduration<=45*60)
dim(bike.sub2)
```

Result (Console)

```
> bike.sub1 <- subset(citibike, citibike$tripduration >45*60)
> dim(bike.sub1)
[1] 216  15
> bike.sub2 <- subset(citibike,citibike$tripduration<=45*60)
> dim(bike.sub2)
[1] 20637    15
```

As we see,
For the group(tripduration<=45min),
Number of trips=20637
For the group(tripduration>45min),
Number of trips=216
So, Total Price
= 20637*3+216*5=$62991

Result

As calculated above, the total revenue is $62991.

2.Looking at tripduration in minutes, what can you say about the variance in the data.

a.What does this mean for the pricing strategy?

b.What does this mean for inventory availability?

As we summarized above, the variance of tripduration in minutes is very large, which means the time people used citibike are largely varied and spread out from the average duration time, and there exits some extreme value, which means only a minimum people take the citibike for an extremely long time.

Based on that,
a. For the pricing itself, as a significant factor that reflect the balance between product supply and demand, there are several factors we need to consider when relates the price strategy:

- Fixed and variable costs
- Competition
- Company objectives
- Proposed positioning strategies
- Target group and willingness to pay

And all the price strategy should serve the company objectives. Based on what I researched, the citibike program is aimed to encourage more people to use cycling as a new transport to reduce emissions, road wear, collisions, and road and transit congestion. At the same time, considering cycling is more used for a short trip, so I think there are several points related with pricing strategy,

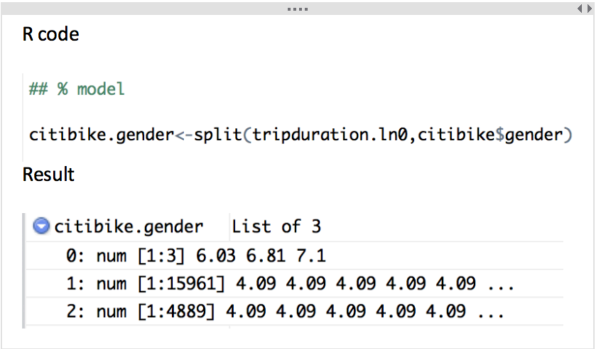
- The starting price is more acceptable compared with other public transport. Method adopted: as most people used it as a short trip transport, in order to encourage more people,
    - The usage time for the starting price need include the majority trip duration
    - Adding membership or more discount encourage people take it as a mainly transport.
- Towards some extreme tripduarion, in order to decrease the maintenance fee for all the bikes and stations,
    - Except the starting price, set one or two levels to control some extremes duration time.

b. For the for inventory availability, as most people use the bike for a short time, the inventory basically could satisfy the public demand. But for some tourist sites, it still depends on the station density and public demand. In addition, in order to make sure the availability, the citibike adopted a method to set certain usage time for unlimited pass or membership, which could encourage riders return the bike to some certain stations in a certain time period.

3. A business manager wants to reallocate the $5M marketing budget using a gender segmentation strategy. Specifically, the manager is asking you to create two models:
a. A model that use % of male vs females in the dataset
b. A model based on average trip duration by gender

a. % of males and females Model
As I searched from the citibike dataset explanation, based on Gender (Zero=unknown; 1=male; 2=female), the % of males and females model could represented below,

Using the split() function or subset() function to divide the whole dataset into 3 groups(gender=0,gender=1,gender=2), we could get the number of each group.



R code

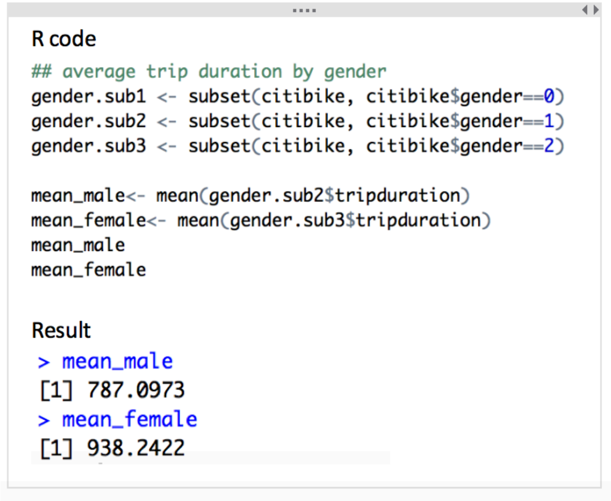
```
## % model

citibike.gender<-split(tripduration.ln0,citibike$gender)
```

Result

```
citibike.gender    List of 3
  0: num [1:3] 6.03 6.81 7.1
  1: num [1:15961] 4.09 4.09 4.09 4.09 4.09 ...
  2: num [1:4889] 4.09 4.09 4.09 4.09 4.09 ...
```

Omitting the unknown value,
N(male)=15961
N(female)=4889
%(male)=76.55%
%(female)=23.45%
B(male)=3,827,500
B(female)=1,172,500
So based on the % model, the budget used on male is about $3,827,500, and for female is about $1,172,500, which means their target customer is male.

b. A model based on average trip duration by gender
In order to get the average trip duration by gender, I use the subset function to get the average trip duration for male and female.



R code

```
## average trip duration by gender
gender.sub1 <- subset(citibike, citibike$gender==0)
gender.sub2 <- subset(citibike, citibike$gender==1)
gender.sub3 <- subset(citibike, citibike$gender==2)

mean_male<- mean(gender.sub2$tripduration)
mean_female<- mean(gender.sub3$tripduration)
mean_male
mean_female
```

Result

```
> mean_male
[1] 787.0973
> mean_female
[1] 938.2422
```

As we calculated on the left,
Dura_Avg(male) =787.0973 sec
Dura_Avg(female) =938.2422 sec
%(male)=45.62%   1725.3395
%(female)=54.38 %
B(male)=2,281,000
B(female)=2,719,000
So based on the average trip duration model by gender, the budget used on male is about $2,281,000 and for female is about $2,719,000.

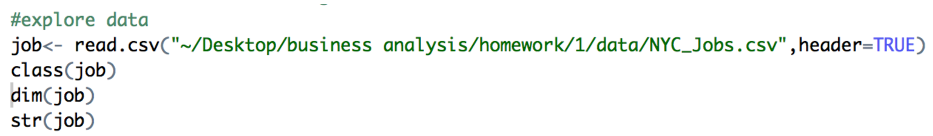
Part 2: Teach me something

This part of the assignment is fairly simple and open-ended. Your first task is to get yourself a data set that you like and teach me something about it. Anything. It doesn’t have to be profound, it doesn’t have to be earth changing, it should just use your skills from this lesson. Some thoughts on choosing your dataset:

Target Dataset: NYC_job (nyc opendata)

This dataset contains current job postings available on the City of New York’s official jobs site (http://www.nyc.gov/html/careers/html/search/search.shtml). Internal postings available to city employees and external postings available to the general public are included.

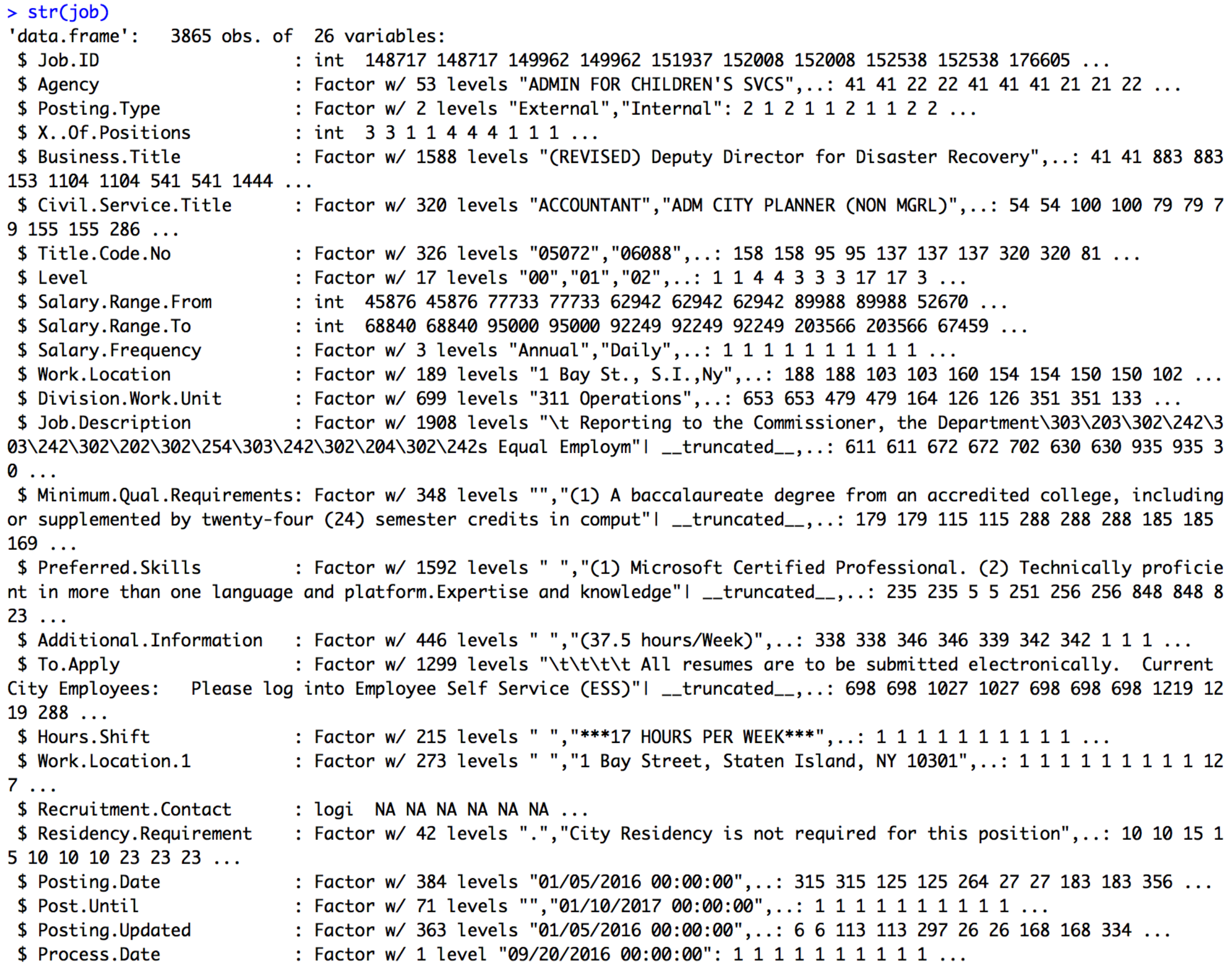
Dataset explanation



```
#explore data
job<- read.csv("~/Desktop/business analysis/homework/1/data/NYC_Jobs.csv",header=TRUE)
class(job)
dim(job)
str(job)
```

Based on the code above, the dataset is explained below,

- The dataset imported as a data frame has 3865 observations and 26 variables(3865rows*26colums).
- For each variable's characteristic,



```
> str(job)
'data.frame':	3865 obs. of  26 variables:
 $ Job.ID                       : int  148717 148717 149962 149962 151937 152008 152008 152538 152538 176605 ...
 $ Agency                       : Factor w/ 53 levels "ADMIN FOR CHILDREN'S SVCS",..: 41 41 22 22 41 41 41 21 21 22 ...
 $ Posting.Type                 : Factor w/ 2 levels "External","Internal": 2 1 2 1 1 2 1 1 2 2 ...
 $ X..Of.Positions              : int  3 3 1 1 4 4 4 1 1 1 ...
 $ Business.Title               : Factor w/ 1588 levels "(REVISED) Deputy Director for Disaster Recovery",..: 41 41 883 883
153 1104 1104 541 541 1444 ...
 $ Civil.Service.Title          : Factor w/ 320 levels "ACCOUNTANT","ADM CITY PLANNER (NON MGRL)",..: 54 54 100 100 79 79 7
9 155 155 286 ...
 $ Title.Code.No                : Factor w/ 326 levels "05072","06088",..: 158 158 95 95 137 137 137 320 320 81 ...
 $ Level                        : Factor w/ 17 levels "00","01","02",..: 1 1 4 4 3 3 3 17 17 3 ...
 $ Salary.Range.From            : int  45876 45876 77733 77733 62942 62942 62942 89988 89988 52670 ...
 $ Salary.Range.To              : int  68840 68840 95000 95000 92249 92249 92249 203566 203566 67459 ...
 $ Salary.Frequency             : Factor w/ 3 levels "Annual","Daily",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Work.Location                : Factor w/ 189 levels "1 Bay St., S.I.,Ny",..: 188 188 103 103 160 154 154 150 150 102 ...
 $ Division.Work.Unit           : Factor w/ 699 levels "311 Operations",..: 653 653 479 479 164 126 126 351 351 133 ...
 $ Job.Description              : Factor w/ 1908 levels "\t Reporting to the Commissioner, the Department\303\203\302\242\3
03\242\302\202\302\254\303\242\302\204\302\242s Equal Employm"| __truncated__,..: 611 611 672 672 702 630 630 935 935 3
0 ...
 $ Minimum.Qual.Requirements: Factor w/ 348 levels "","(1) A baccalaureate degree from an accredited college, including
or supplemented by twenty-four (24) semester credits in comput"| __truncated__,..: 179 179 115 115 288 288 288 185 185
169 ...
 $ Preferred.Skills             : Factor w/ 1592 levels " ","(1) Microsoft Certified Professional. (2) Technically proficie
nt in more than one language and platform.Expertise and knowledge"| __truncated__,..: 235 235 5 5 251 256 256 848 848 8
23 ...
 $ Additional.Information       : Factor w/ 446 levels " ","(37.5 hours/Week)",..: 338 338 346 346 339 342 342 1 1 1 ...
 $ To.Apply                     : Factor w/ 1299 levels "\t\t\t\t All resumes are to be submitted electronically.  Current
City Employees:   Please log into Employee Self Service (ESS)"| __truncated__,..: 698 698 1027 1027 698 698 698 1219 12
19 288 ...
 $ Hours.Shift                  : Factor w/ 215 levels " ","***17 HOURS PER WEEK***",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Work.Location.1              : Factor w/ 273 levels " ","1 Bay Street, Staten Island, NY 10301",..: 1 1 1 1 1 1 1 1 1 12
7 ...
 $ Recruitment.Contact          : logi  NA NA NA NA NA NA ...
 $ Residency.Requirement        : Factor w/ 42 levels ".","City Residency is not required for this position",..: 10 10 15 1
5 10 10 10 23 23 23 ...
 $ Posting.Date                 : Factor w/ 384 levels "01/05/2016 00:00:00",..: 315 315 125 125 264 27 27 183 183 356 ...
 $ Post.Until                   : Factor w/ 71 levels "","01/10/2017 00:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Posting.Updated              : Factor w/ 363 levels "01/05/2016 00:00:00",..: 6 6 113 113 297 26 26 168 168 334 ...
 $ Process.Date                 : Factor w/ 1 level "09/20/2016 00:00:00": 1 1 1 1 1 1 1 1 1 1 ...
```

Reason

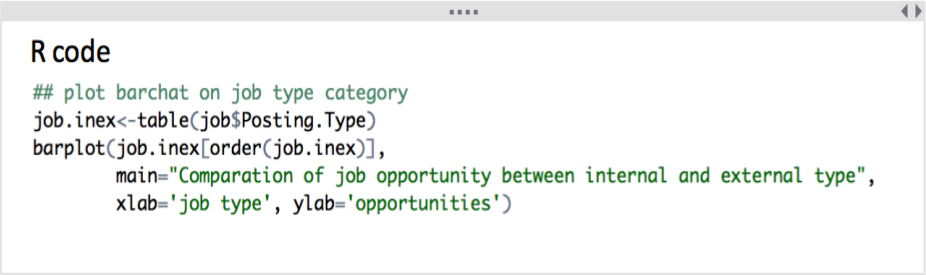
- Get some ideas about job information, which is essential for full-time students. And based on the dataset, I plan to get some ideas about position type, No of position available, and salary.
- The size of this dataset is reasonable to deal with. Originally I plan to analyze the taxi data, but 2GB is too big to download and analyze.
- Still, the dataset has one disadvantage, all the job listed in the dataset is provided by government agency, which means it cannot be the representative for

the whole job market. But on the other hand, it could minimize some other impact factor among different sections (like financing agency, commercial agency, or IT).

- So, I would clarify all the analysis below is only focusing on the government job market.

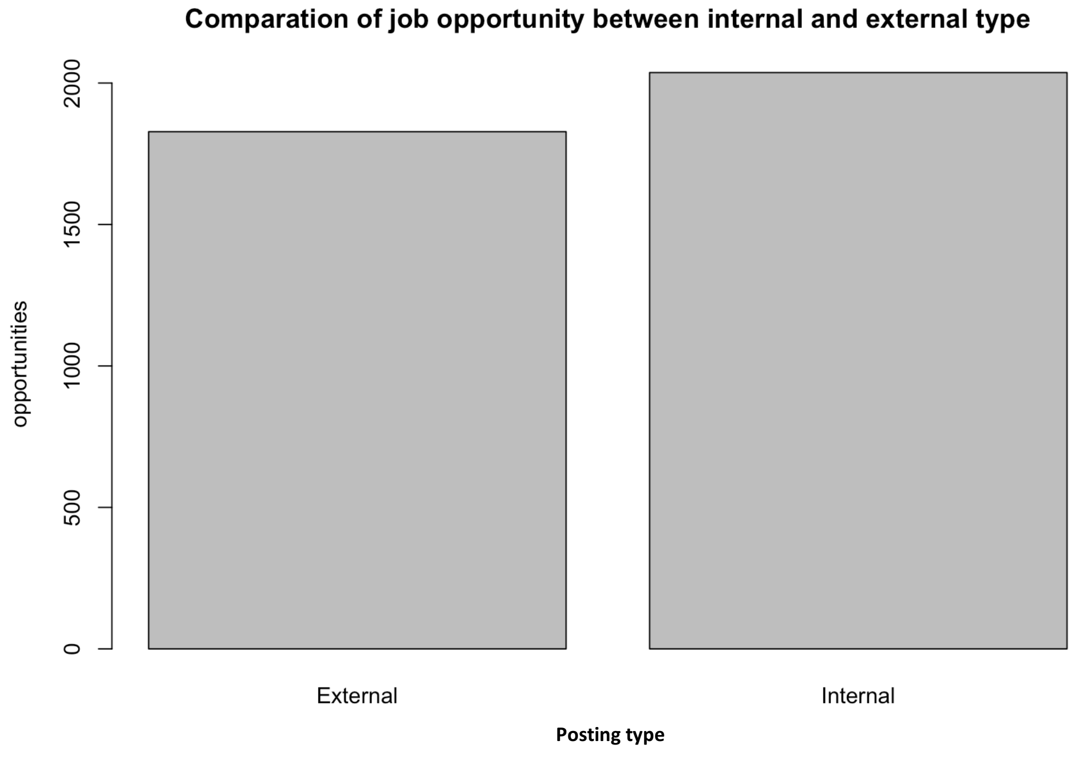
Dataset exploration

✓ The internal positions are much more than the external positions in the job dataset.



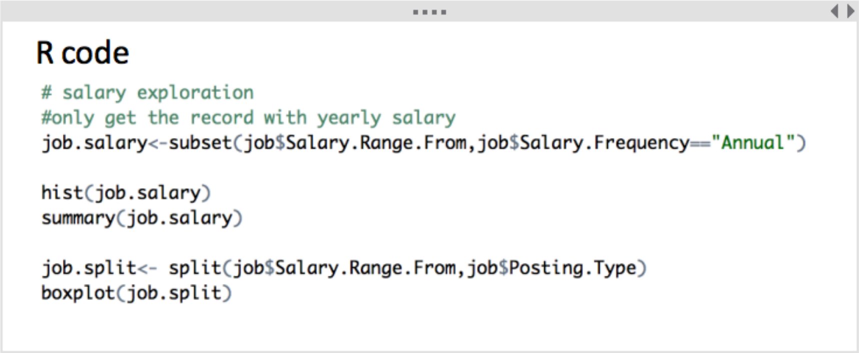
R code

```
## plot barchat on job type category
job.inex<-table(job$Posting.Type)
barplot(job.inex[order(job.inex)],
        main="Comparation of job opportunity between internal and external type",
        xlab='job type', ylab='opportunities')
```



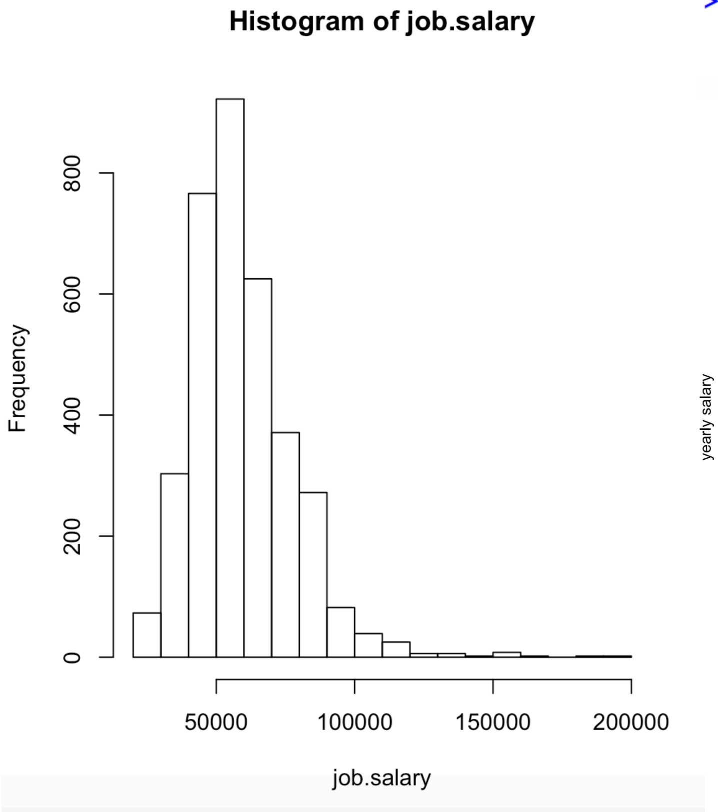
To some extent, it explained why networking is so important, especially networking with the people in the targeted company. Because there are much more opportunities for people already inside, and if they could help you, you will get more chances to have a job.

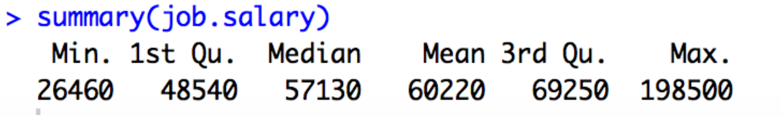
- ✓ The average salary per year is at least $60220. And the Min. of salary of the government position is in the range of (26460, 198500). In addition, there is no big difference between the salary of internal one and external one.



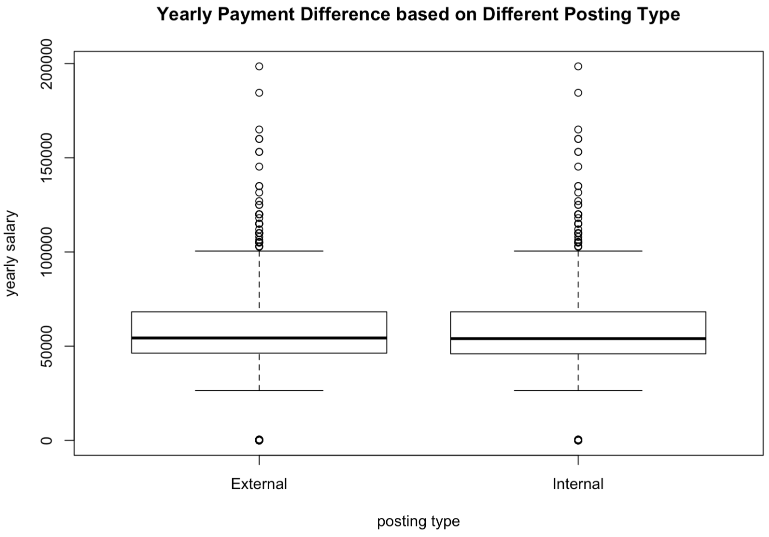
R code

```
# salary exploration
#only get the record with yearly salary
job.salary<-subset(job$Salary.Range.From,job$Salary.Frequency=="Annual")

hist(job.salary)
summary(job.salary)

job.split<- split(job$Salary.Range.From,job$Posting.Type)
boxplot(job.split)
```





```
> summary(job.salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26460   48540   57130   60220   69250  198500
```



Business application and possibility

Possibility 1: Creating a New Professional Job Searching and Analysis Website

This website is not only designed to provide the employment information, but also some specific job analysis reports related with all the job or employment data.

As I analyzed in the data exploration part, we could get more information if we could get all the data related the whole job market or some specific industry. It could help us, especially the students with no working experience. For one thing, the specific industry job report could help the student to get some direct perspectives towards this field. And of course it will also help the student to prepare the interview. In addition, it could guide the student find his own job niche, which is a good way to think about the future career path.

Possibility 2: Networking Platform Focused on Job Market – (LinkedIn)

From the government career market, we could see there are more opportunities that are not open to the public, which means the importance of networking in the job market. So, if there is a platform mainly focused on the professional career networking, it will get a bunch of supporters and account a big part in the job market. As we see now, LinkedIn is the one designed to do something in this field.

Appendix (R code)

- **First part Citibike data analysis**

```
getwd()
citibike<-read.csv("~/Desktop/CitiBike Data.csv",header=TRUE)
#explore data
names(citibike)
dim(citibike)
class(citibike)
head(citibike)
str(citibike)

#statistical summary
install.packages("psych")
library("psych")

#summary statistics for tripduration
describe(citibike$tripduration)  #summary(citibike$tripduration)
hist(citibike$tripduration)

#summary statistics for age (only provided birth year in data)
age<- 2016-citibike$birth.year
describe(age)
hist(age)

#summary statistics for tripduration in minutes
trip_minutes<-citibike$tripduration/60
describe(trip_minutes)
hist(trip_minutes)
#效果相同 describe(citibike$tripduration/60)

#the correlation between age and tripduration
cor.test(age,citibike$tripduration)
plot(age,citibike$tripduration)

#plot graph
#         Plot the histograms and box plots for tripduration by gender
install.packages("ggplot2")
library("ggplot2")
##plot without any modificaition
citibike$gender = ifelse(citibike$gender==1, "female", "male")
library(lattice)
```

```
histogram(~ tripduration | gender, data=citibike,col="grey",breaks = 50)


## modify data
tripduration.ln0<-log(citibike$tripduration)
tripduration.scale<-scale(citibike$tripduration)
#par(mar=c(1,1,1,1))
par(mfrow=c(2,1))
hist(tripduration.scale)
hist(tripduration.ln0)

##histgram
citibike["log_tripduraion"]<-tripduration.ln0
ggplot(citibike, aes(x=citibike$log_tripduraion, fill=type))+geom_histogram(binwidth=.5,
alpha=.5, position="dodge")

#histgram
library(plyr)
cdat <- ddply(citibike, "type", summarise, tripduration.mean=mean(tripduration.ln0))
cdat
ggplot(citibike, aes(tripduration.ln0, fill=type))+geom_histogram(binwidth=.5, alpha=.5,
position="dodge")+geom_vline(data=cdat,
aes(xintercept=tripduration.mean,colour=gender),linetype="dashed", size=1)

##boxplot
citibike.gender<-split(tripduration.ln0,citibike$gender)
boxplot(citibike.gender,col="beige",main="Box Plot for Tripduration by
Gender",xlab="Gender",ylab="tripduration_log")

##boxplot
ggplot(citibike, aes(x=type, y=tripduration.ln0, fill=type)) + geom_boxplot()
```

- **Second Part: Teaching me something**

```
# homework1 teach me something
#explore data
job<- read.csv("~/Desktop/business analysis/homework/1/data/NYC_Jobs.csv",header=TRUE)
class(job)
dim(job)
str(job)

## plot barchat on job type category
job.inex<-table(job$Posting.Type)
```

```
barplot(job.inex[order(job.inex)],
        main="Comparation of job opportunity between internal and external type",
        xlab='posting type', ylab='opportunities')

# salary exploration
#only get the record with yearly salary
job.salary<-subset(job$Salary.Range.From,job$Salary.Frequency=="Annual")

hist(job.salary)
summary(job.salary)

job.split<- split(job$Salary.Range.From,job$Posting.Type)
boxplot(job.split,main="Yearly Payment Difference based on Different Posting
Type",xlab='posting type',ylab='yearly salary')
```