BA Assignment II

Kunyang Que, N12065518, kq395

Problem I: Citibike

Import library and load data

In [2]:
```
library(lubridate)
library(dplyr)
library(ggplot2)
```

In [3]:
```
file.path<-"https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/JC-201709-citibike-tripdata.csv"
data<-read.csv(file.path,header=TRUE,stringsAsFactors=FALSE)
head(na.omit(data))
```

A data.frame: 6 × 15

	tripduration	starttime	stoptime	start.station.id	start.station.name	start.station.latitude	start.s
	<int>	<chr>	<chr>	<int>	<chr>	<dbl>	
1	364	2017-09-01 00:02:01	2017-09-01 00:08:05	3183	Exchange Place	40.71625	
2	357	2017-09-01 00:08:12	2017-09-01 00:14:09	3187	Warren St	40.72112	
3	432	2017-09-01 00:10:12	2017-09-01 00:17:24	3195	Sip Ave	40.73074	
4	934	2017-09-01 00:10:11	2017-09-01 00:25:46	3272	Jersey & 3rd	40.72333	
5	932	2017-09-01 00:10:16	2017-09-01 00:25:48	3272	Jersey & 3rd	40.72333	
6	414	2017-09-01 00:15:32	2017-09-01 00:22:26	3186	Grove St PATH	40.71959	

1.1 Compute summary statistics for tripduration

In [4]:
```
summary(data$tripduration)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   61.0   238.0   355.0   756.9   610.0 2181628.0 
```

1.2 Compute summary statistics for age

Get the current year

In [5]:
```
dates<-as.Date(data$starttime, "%Y-%m-%d", tz = "UTC")
get.year<-year(dates)
head(get.year)
```

2017 · 2017 · 2017 · 2017 · 2017 · 2017

Calculate the age

In [6]:
```
a<-as.numeric(data$birth.year)
age<-get.year-a
head(age)
```

```
Warning message in eval(expr, envir, enclos):
"NAs introduced by coercion"
```

28 · 37 · 29 · 26 · 24 · <NA>

Summary statistics

In [7]:
```
summary(age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  16.00   30.00   34.00   36.88   42.00  130.00    2384 
```

1.3 Compute summary statistics for tripduration in minutes

Transfer seconds to minutes

In [8]:
```
tripd.min<-data$tripduration/60
head(tripd.min)
```

6.06666666666667 · 5.95 · 7.2 · 15.5666666666667 · 15.5333333333333 · 6.9

Summary statistics

In [9]:
```
summary(tripd.min)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.02    3.97    5.92   12.62   10.17 36360.47 
```

1.4 Compute the correlation between age and tripduration

In [10]:
```
cor(age, tripd.min, use="na.or.complete")
```

0.00705514845205564

1.5 Business Questions

- 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.

In [11]:
```
data<-mutate(data,revenue=if_else(data$tripduration/60<=45,3,5))

paste("The total revenue is: ", sum(data$revenue))
```

'The total revenue is: 100651'

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

In [12]:
```
var(tripd.min)
```

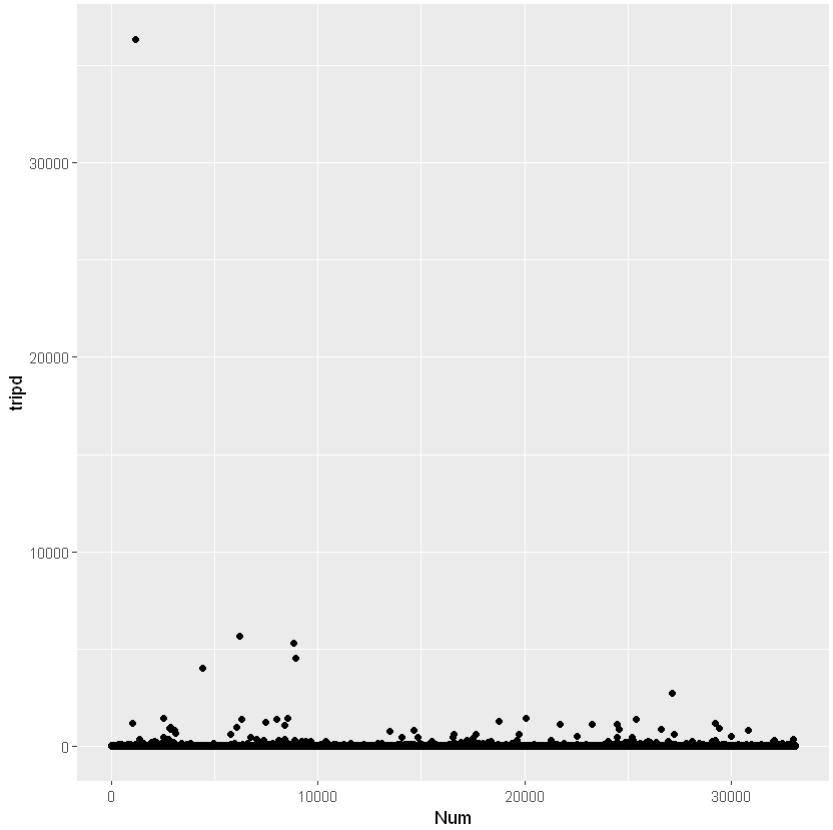
44300.2432734721

Such a large variance indicates that the user's riding time ranges from short to very long, which means that users could be segmented to several groups according to their riding time and a corresponding pricing strategy should be made for each category.

- What does this mean for the pricing strategy?

Let's view the scatter plot at first.

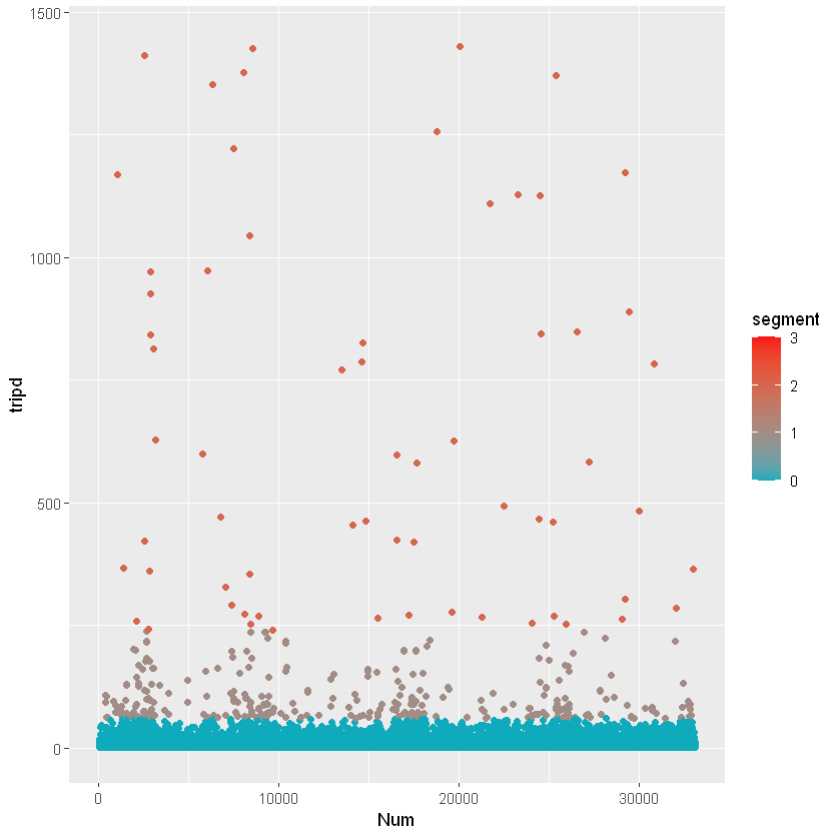
In [13]:
```
count<-1: length(tripd.min)
plot.data<-data.frame(Num=count, tripd=tripd.min)
ggplot(plot.data, aes(x=Num,y=tripd))+geom_point()
```



As we can see from the scatter plot, almost all data are concentrated between 0 to 1440(1 day) minutes.We enlarge the scatter plot, analyze the data in the 0-1440 field, and did a segmentation.

In [14]:
```
plot.data<-mutate(plot.data,segment=if_else(tripd<=60,0,
                                            if_else(tripd<=4*60,1,
                                                    if_else(tripd<=24*60,2,3))))
ggplot(plot.data, aes(x=Num,y=tripd,group=segment,color=segment))+
    ylim(0,1440)+geom_point()+scale_colour_gradient(low="#0fabbc", high="#fa1616")
```

```
Warning message:
"Removed 6 rows containing missing values (geom_point)."
```



1. We set recodes within one hour of riding time(blue points) as **"Short Distance"**. For this group, renting a bicycle may just meet the short-distance riding demand from the subway station to the destination. A very large number of records stay in this range. Therefore, the company can give corresponding preferential plans for this length of rental, such as a "Daily Pass" that allows one hour of cycling per day.
2. The one-hour to four-hour of riding(brown points) is set as the **"Medium Distance"**. Cycling during this time period may be exercise or sightseeing trips. Therefore, we suggest that the cycling discount for this time period can be placed on weekends.
3. The one-hour to four-hour of riding(orange points) is set as the **"Long Distance"**. We recommended that a segmented pricing method could be took. For example, 4 hours to 6 hours are charged 10 dollars, 6 hours to 12 hours for the part exceeding 4 hours is charged 1 dollar per hour, and the part exceeding 12 hours is charged 1.5 dollor per hour.
4. Only a very small part of the record is more than 1 day(doesn't show on the second scatter plot), we set it as **"Lost"**. Penalty measures can be formulated for this part of users.

- What does this mean for inventory availability?

1. Schedule regular maintenance plan for bicycles with high frequency of use.
2. Increase the amount of bicycle in areas with greater bike demand.
3. Count the bicycles that cannot be returned for a long time and replenish new bicycles in time.

Problem II: Zagat

Load data

In [15]:
```
url<-"https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/zagat.CSV"
data_zagat<-read.csv(url, header=TRUE,stringsAsFactors=FALSE)
head(data_zagat)
```

A data.frame: 6 × 5

	Name	Food	Decor	Service	Price
	<chr>	<int>	<int>	<int>	<int>
1	107 West	16	13	16	26
2	2nd Street cafe	14	13	15	21
3	44 & Hell's kitchen	22	19	19	42
4	55 wall	21	22	21	54
5	55 wall street	21	22	21	54
6	92	15	15	15	43

Calculate the statistical index

In [16]:
```
food.sd=sd(data_zagat$Food)
decor.sd=sd(data_zagat$Decor)
service.sd=sd(data_zagat$Service)
price.sd=sd(data_zagat$Price)
total.sd<-food.sd+decor.sd+service.sd+price.sd
total.sd
```

27.0859240306785

In [17]:
```
food.weight=food.sd/total.sd
food.weight
```

0.136332348818796

In [18]:
```
decor.weight=decor.sd/total.sd
decor.weight
```

0.18274911975718

In [19]:
```
service.weight=service.sd/total.sd
service.weight
```

0.131682904331738

In [20]:
```
price.weight=price.sd/total.sd
price.weight
```

0.549235627092286

Calculate the score

Based on the result, we calculate the score with the following formula: **food.weight *food* + *price.weight* price + service.weight *service* + *decor.weight* decor**.

In [21]:
```
score=food.weight*data_zagat$Food+price.weight*data_zagat$Price+service.weight*data_zag
at$Service+decor.weight*data_zagat$Decor
summary(score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.48   20.57   26.78   27.82   34.54   55.06 
```

In [22]:
```
data_zagat$Score=score
head(data_zagat)
```

A data.frame: 6 × 6

	Name	Food	Decor	Service	Price	Score
	<chr>	<int>	<int>	<int>	<int>	<dbl>
1	107 West	16	13	16	26	20.94411
2	2nd Street cafe	14	13	15	21	17.79358
3	44 & Hell's kitchen	22	19	19	42	32.04142
4	55 wall	21	22	21	54	39.30752
5	55 wall street	21	22	21	54	39.30752
6	92	15	15	15	43	30.37860

2.1 Describe in your index in English (marketing lingo)

score = food.weight *food* + *price.weight* price + service.weight *service* + *decor.weight* decor

- According to the formula, to evaluate the quality of a restaurant, price is the most important factor accounting for 54.92%. Followed by decoration, accounting for 18.27%. Then there are food and service, which accounted for 13.63% and 13.17%.

2.2 Using your index, how can a restaurant become more profitable?

To make a restaurant more profitable, the manager should try to obtain more income while controlling expenses.Therefore, there are several alternative methods.

1. On the premise of keeping the quality of the dishes unchanged, appropriately lower the prices of the dishes with large amounts of orders to attract more customers.
2. Improve the quality of high-end dishes while increasing prices to make customers who are not sensitive to prices more willing to pay for these dishes. Make these dishes produce a premium effect.
3. Appropriately reduce the cost of decoration and services.

2.3 Describe the characteristics of restaurant operating at the bottom 25% and top 25% of your index

TOP 25%

In [23]:
```
top25<-data_zagat[order(data_zagat$Score, decreasing=T), ][0:as.integer(length(data_zagat$Score)/4),]
head(top25)
```

A data.frame: 6 × 6

	Name	Food	Decor	Service	Price	Score
	<chr>	<int>	<int>	<int>	<int>	<dbl>
272	Veritas	27	22	26	80	55.06406
103	Four Season	26	27	26	78	54.74300
205	Nobu	28	23	24	74	51.82436
212	Oceana	27	25	26	72	51.21842
213	Oceana	27	25	26	72	51.21842
265	Union Pacific	26	26	25	72	51.13316

In [24]:
```
summary(top25[,2:6])
```

```
      Food           Decor          Service          Price      
 Min.   :17.00   Min.   :13.00   Min.   :15.00   Min.   :45.00  
 1st Qu.:21.00   1st Qu.:19.00   1st Qu.:20.00   1st Qu.:52.00  
 Median :23.00   Median :21.00   Median :21.00   Median :55.00  
 Mean   :22.95   Mean   :20.93   Mean   :21.31   Mean   :57.01  
 3rd Qu.:25.00   3rd Qu.:23.00   3rd Qu.:23.00   3rd Qu.:61.00  
 Max.   :28.00   Max.   :27.00   Max.   :26.00   Max.   :80.00  
     Score      
 Min.   :34.60  
 1st Qu.:37.63  
 Median :39.84  
 Mean   :41.07  
 3rd Qu.:43.21  
 Max.   :55.06  
```

In [25]:
```
bottom25<-data_zagat[order(data_zagat$Score), ][0:as.integer(length(data_zagat$Score)/4
),]
head(bottom25)
```

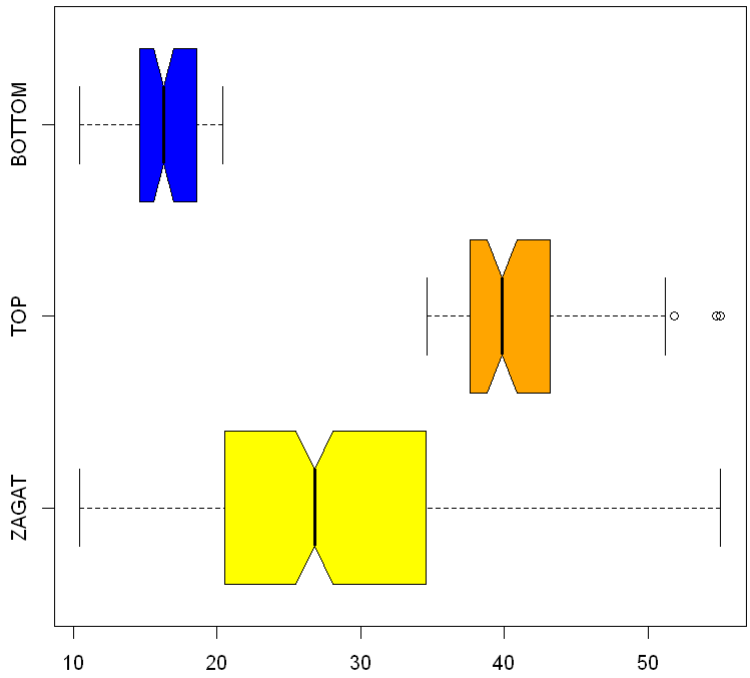
A data.frame: 6 × 6

	Name	Food	Decor	Service	Price	Score
	<chr>	<int>	<int>	<int>	<int>	<dbl>
281	vinnie's pizza	20	3	13	10	10.47913
29	Bo-ky	17	4	8	12	10.69294
92	Ess a bajel	23	5	13	9	10.70439
159	Magnolia Bakery	25	10	13	8	11.34156
26	big wong	22	3	11	12	11.58690
232	Pump Energy food, the	19	4	14	12	11.75570

In [26]:
```
summary(bottom25[,2:6])
```

```
      Food           Decor          Service          Price      
 Min.   : 9.00   Min.   : 3.00   Min.   : 8.00   Min.   : 8.00  
 1st Qu.:15.00   1st Qu.: 8.00   1st Qu.:12.00   1st Qu.:16.00  
 Median :17.00   Median :10.00   Median :13.00   Median :19.00  
 Mean   :17.49   Mean   :10.24   Mean   :13.23   Mean   :18.91  
 3rd Qu.:20.00   3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:22.00  
 Max.   :25.00   Max.   :20.00   Max.   :17.00   Max.   :25.00  
     Score      
 Min.   :10.48  
 1st Qu.:14.64  
 Median :16.29  
 Mean   :16.38  
 3rd Qu.:18.60  
 Max.   :20.39  
```

In [27]:
```
boxplot(data_zagat$Score,top25$Score,bottom25$Score,names=c("ZAGAT","TOP","BOTTOM"),
        col=c("yellow","orange","blue"),horizontal=TRUE,notch=TRUE)
```



1. From the point of view of scores, the average score of the Top group is 41.07, and the average score of the Bottom group is 16.38. Compared with the top group, the score distribution of the bottom group is more concentrated.
2. From the aspect of food and price, the average score of bottom in price is only about 5 points lower than top, but the price is about 38 points lower.

2.4 If you were hired to advise a new restaurant operator, what would you recommend in terms of the balance & trade-offs between food, decor, service, and price?

1. Achieving the industry average level of service and decoration, and focusing on improving the cost-effectiveness of dishes.
2. Introducing dishes with lower prices than competing products of the same quality to attract more customers.
3. Prioritize price control in improving the quality of dishes and controlling prices.