Capstone Final

J. Hiba A. Kader

Wednesday, December 9, 2015

# Ponpare Coupons

Ponpare is Japanese coupon site that offers discount coupons on many activities, goods and services. The goal of the following analysis and model is to improve Ponpare recommendation system

We have few datasets provided by the company that give us information about the users and the coupons purchased.

## Descriptive Aspect

**USERS**

***Who are our Users? and what do we know about them?***

let us have a look at the *users list* to see the distribution of the **AGE** of the users.

str(user\_list\_en)

## 'data.frame': 22873 obs. of 7 variables:  
## $ PREF\_NAME : Factor w/ 48 levels "","å¯<U+008C>å±±ç<U+009C><U+008C>",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ REG\_DATE : Factor w/ 22842 levels "2010-07-21 13:44:02",..: 20897 20623 16670 19295 13750 19510 17293 15907 18913 18113 ...  
## $ SEX\_ID : Factor w/ 2 levels "f","m": 1 1 1 2 2 2 2 1 2 2 ...  
## $ AGE : int 25 28 35 25 30 36 22 24 44 31 ...  
## $ WITHDRAW\_DATE: Factor w/ 919 levels "2011-07-07 19:51:30",..: NA NA NA NA NA NA NA NA NA NA ...  
## $ USER\_ID\_hash : Factor w/ 22873 levels "0000b53e182165208887ba65c079fc21",..: 19542 15730 5693 5991 20014 6788 10199 5477 18081 10240 ...  
## $ en\_pref\_name : Factor w/ 47 levels "Aichi Prefecture",..: NA NA NA NA NA NA NA NA NA NA ...

We have 22,875 observations and 7 variables including AGE, SEX ID(Gender), USER ID hash, PREF NAME (which is the name of the user's area in Japanese), en pref name (the English translation of the PREF NAME), Registered date and Withdrawal date (dates of the user registration and withdrawal from the website).

for my analysis, I will only be using Age, Sex\_id, prefecture name and the User ID. Now, let us explore Age.

hist(user\_list\_en$AGE, main="Age Distribution", sub="Figure 1",cex.sub=0.7, col="blue", xlab= "USERS AGE")

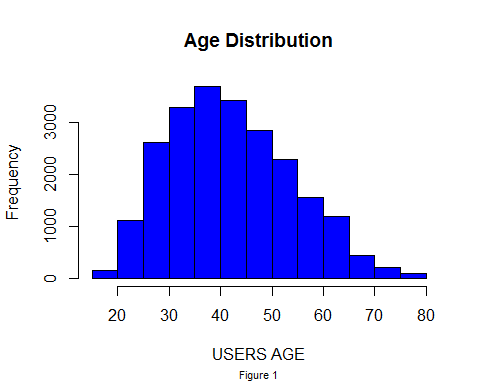


Figure (1) shows that the *AGE* variable has a normal distribution

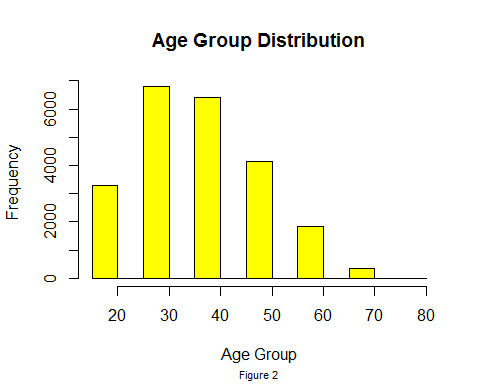
describe(user\_list\_en$AGE)

## user\_list\_en$AGE   
## n missing unique Info Mean .05 .10 .25 .50   
## 22873 0 66 1 42.5 25 28 33 41   
## .75 .90 .95   
## 51 59 64   
##   
## lowest : 15 16 17 18 19, highest: 76 77 78 79 80

We have 66 different ages category, our users' age varies between 15 to 80 with a mean of 42.5 years and standard deviation of 11.84.

I categorized age based on 10 years interval and called the new variable *age group*

attach(user\_list\_en)  
user\_list\_en$age\_group[AGE<=19 & AGE>=15] = "18"  
user\_list\_en$age\_group[AGE<=29 & AGE>=20] = "20"  
user\_list\_en$age\_group[AGE<=39 & AGE>=30] = "30"  
user\_list\_en$age\_group[AGE<=49 & AGE>=40] = "40"  
user\_list\_en$age\_group[AGE<=59 & AGE>=50] = "50"  
user\_list\_en$age\_group[AGE<=69 & AGE>=60] = "60"  
user\_list\_en$age\_group[AGE<=79 & AGE>=70] = "70"  
user\_list\_en$age\_group[AGE<=89 & AGE>=80] = "80"  
user\_list\_en$age\_group=as.numeric(user\_list\_en$age\_group)  
hist(user\_list\_en$age\_group, main="Age Group Distribution", col="yellow",sub="Figure 2", cex.sub=0.7, xlab="Age Group")



By creating age groups to categorize users, I was able to determine that most users are in their 30's and 40's; \* 30% of users are in their 30's \* 28% of users in their 40's and \* 18% of users in their 50's

This could be explained by the fact that young USERS in these age groups are more net-savvy and use on-line shopping more frequently than the older less tech-savvy age groups.

Now that we know the age, Let us explore our USERS's **GENDER**.

describe(user\_list\_en$SEX\_ID)

## user\_list\_en$SEX\_ID   
## n missing unique   
## 22873 0 2   
##   
## f (10983, 48%), m (11890, 52%)

The results show that we slightly have more MALE users that make up 52% of our entire users.

***Does the gender of users change in differnt age groups? Are there more males than females in some age groups or vise versa?***

Let us calculate the probabilities of **Gender** per **Age** category.

t= table(user\_list\_en$SEX\_ID,user\_list\_en$age\_group)  
barplot(prop.table(t,2), legend=paste(unique(user\_list\_en$SEX\_ID)), ylab="Probability", xlab="Group Age", main="Gender & Age Group", sub="Figure 3", cex.sub=0.7, col=c("pink","skyblue"))

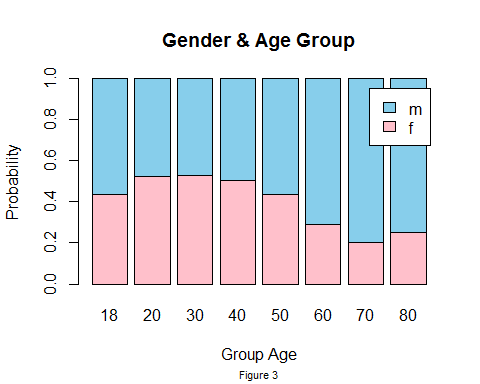


Figure (3) shows that in the (20, 30 and 40) age groups we almost have equal gender presence, with slightly more females than males.

Interestingly, the genders of users start shifting slightly in the 50's age group, with a drastic shift in the 60's age group where males make up 70% of users, and in the 70's age group where males increase to 80% of users in this age group, this could be due to the traditional role of men in Japan.

Let us look at the **AREA** were our users come from and determine the distribution of our users based on the prefecture name.

unique(user\_list\_en$en\_pref\_name)

## [1] <NA> Toyama Prefecture Mie   
## [4] Wakayama Prefecture Oita Prefecture Osaka prefecture   
## [7] Nara Prefecture Yamaguchi Prefecture Yamagata Prefecture   
## [10] Yamanashi Prefecture Miyazaki Prefecture Miyagi Prefecture   
## [13] Hyogo Prefecture Chiba Prefecture Saga Prefecture   
## [16] Tokushima Prefecture Okayama Prefecture Iwate Prefecture   
## [19] Gifu Prefecture Shimane Prefecture Niigata Prefecture   
## [22] Tochigi Prefecture Ehime Prefecture Aichi Prefecture   
## [25] Shiga Prefecture Tokyo Okinawa   
## [28] Kyoto Hokkaido Hiroshima Prefecture  
## [31] Saitama Prefecture Fukuoka Prefecture Fukushima Prefecture  
## [34] Fukui Prefecture Kanagawa Prefecture Akita   
## [37] Kumamoto Prefecture Gunma Prefecture Ishikawa Prefecture   
## [40] Kagawa Prefecture Kochi Prefecture Nagasaki Prefecture   
## [43] Nagano Prefecture Aomori Prefecture Shizuoka Prefecture   
## [46] Kagoshima prefecture Tottori Prefecture Ibaraki Prefecture   
## 47 Levels: Aichi Prefecture Akita Aomori Prefecture ... Yamanashi Prefecture

Our users are distributed over 47 prefectures.

***Are our users clustered in some prefecture more than other prefecture?***

barplot(table(user\_list\_en$en\_pref\_name), name.arg=row.names(user\_list\_en),main ="Prefecture Distribution\n Figure 4", las=2, cex.names=0.6, space=NULL, col="blue")

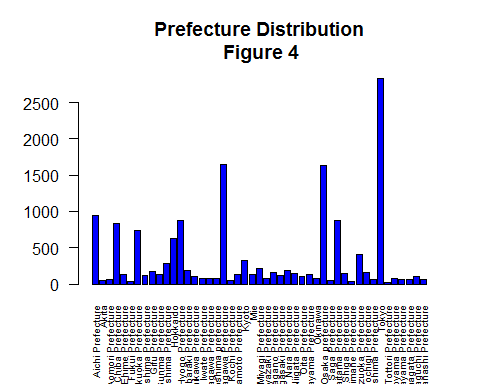


Figure 4, shows the distribution of users in the 47 Prefectures where Tokyo (with 2830 = 12% of users), Kanagawa and Osaka (with 7% of users) have the most users and Tottori Prefecture where we have the least users with only 25 users. However, I have 7256 missing values, which could change this finding.

**COUPONS**

Now let us look at the attributes of the product that Ponpare offers our users **The Coupons**.

Ponpare provided us with three datasets with many variables about the coupons, the first dataset is called "*coupon\_detail\_train\_en*" which includes "Item Count" (count of coupons at each purchase), "small area" (geographical/location), "Coupon ID", "User ID", and "Purchased ID" (log of the purchase transactions).

str(coupon\_detail\_train\_en)

## 'data.frame': 168996 obs. of 7 variables:  
## $ SMALL\_AREA\_NAME: Factor w/ 55 levels "å¯<U+008C>å±±","ä¸<U+0089>é<U+0087>""| \_\_truncated\_\_,..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ ITEM\_COUNT : int 2 1 1 1 1 2 1 2 1 1 ...  
## $ I\_DATE : Factor w/ 166422 levels "2011-07-01 00:10:42",..: 86299 28099 76944 38683 165619 146357 78824 56528 17787 144933 ...  
## $ PURCHASEID\_hash: Factor w/ 168996 levels "0000655d8c1e67679c3c1a8887a97d10",..: 94168 158381 94801 158790 15988 18847 147581 162910 41018 69770 ...  
## $ USER\_ID\_hash : Factor w/ 22782 levels "0000b53e182165208887ba65c079fc21",..: 3484 17541 14913 13914 16181 14977 13313 671 21959 5176 ...  
## $ COUPON\_ID\_hash : Factor w/ 19368 levels "000eba9b783cec10658308b5836349f6",..: 15099 1925 19310 6692 11725 2196 16387 18467 6610 6148 ...  
## $ en\_small\_area : Factor w/ 55 levels "Aichi","Akasaka, Roppongi, Azabu",..: 49 49 49 49 49 49 49 49 49 49 ...

We have 168,996 observations in this dataset based on "Purchase ID", we have 22,782 unique "Coupon ID" and 19,368 unique "User ID".

The second dataset is called *Coupon\_list\_train\_en* which also includes many details about the coupons

str(coupon\_list\_train\_en)

## 'data.frame': 19413 obs. of 29 variables:  
## $ large\_area\_name : Factor w/ 9 levels "ä¸­å<U+009B>½","å<U+009B><U+009B>å<U+009B>½",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ ken\_name : Factor w/ 47 levels "å¯<U+008C>å±±ç<U+009C><U+008C>","ä¸<U+0089>é<U+0087>ç<U+009C><U+008C>""| \_\_truncated\_\_,..: 16 16 16 29 16 16 29 16 16 29 ...  
## $ small\_area\_name : Factor w/ 55 levels "å¯<U+008C>å±±","ä¸<U+0089>é<U+0087>""| \_\_truncated\_\_,..: 17 17 17 35 17 17 35 17 17 35 ...  
## $ GENRE\_NAME : Factor w/ 13 levels "ã<U+0082>¨ã<U+0082>¹ã<U+0083><U+0086>","ã<U+0082>®ã<U+0083><U+0095>ã<U+0083><U+0088>ã<U+0082>«ã<U+0083>¼ã<U+0083><U+0089>",..: 4 4 9 3 1 9 9 9 4 9 ...  
## $ CAPSULE\_TEXT : Factor w/ 25 levels "ã<U+0082>¨ã<U+0082>¹ã<U+0083><U+0086>","ã<U+0082>¤ã<U+0083><U+0099>ã<U+0083>³ã<U+0083><U+0088>",..: 6 6 10 4 1 15 15 10 6 15 ...  
## $ PRICE\_RATE : int 65 50 50 50 91 53 50 50 57 50 ...  
## $ CATALOG\_PRICE : int 6038 7600 7800 4179 72000 10800 25000 39600 3654 9000 ...  
## $ DISCOUNT\_PRICE : int 2100 3800 3900 2089 6000 5000 12500 19800 1560 4450 ...  
## $ DISPFROM : Factor w/ 378 levels "2011-06-27 12:00:00",..: 17 277 222 94 174 69 136 243 105 299 ...  
## $ DISPEND : Factor w/ 373 levels "2011-07-01 12:00:00",..: 13 273 217 91 173 65 133 238 101 295 ...  
## $ DISPPERIOD : int 2 3 2 2 5 1 3 3 2 4 ...  
## $ VALIDFROM : Factor w/ 376 levels "2011-07-02","2011-07-03",..: NA NA 215 90 171 64 131 236 NA 289 ...  
## $ VALIDEND : Factor w/ 514 levels "2011-07-15","2011-07-17",..: NA NA 271 131 305 214 230 371 NA 438 ...  
## $ VALIDPERIOD : int NA NA 85 70 163 179 128 164 NA 178 ...  
## $ USABLE\_DATE\_MON : int NA NA 1 1 1 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_TUE : int NA NA 1 1 1 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_WED : int NA NA 1 1 1 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_THU : int NA NA 1 1 1 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_FRI : int NA NA 0 0 1 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_SAT : int NA NA 0 0 1 1 2 2 NA 2 ...  
## $ USABLE\_DATE\_SUN : int NA NA 0 0 0 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_HOLIDAY : int NA NA 0 0 0 1 1 1 NA 1 ...  
## $ USABLE\_DATE\_BEFORE\_HOLIDAY: int NA NA 0 0 1 1 2 2 NA 2 ...  
## $ COUPON\_ID\_hash : Factor w/ 19413 levels "000eba9b783cec10658308b5836349f6",..: 16946 7720 9936 2261 12523 6037 134 14782 1959 13330 ...  
## $ en\_capsule : Factor w/ 24 levels "Beauty","Class",..: 4 4 12 6 22 11 11 12 4 11 ...  
## $ en\_genre : Factor w/ 13 levels "Beauty","Delivery service",..: 2 2 7 3 13 7 7 7 2 7 ...  
## $ en\_small\_area : Factor w/ 55 levels "Aichi","Akasaka, Roppongi, Azabu",..: 36 36 36 14 36 36 14 36 36 14 ...  
## $ en\_ken : Factor w/ 47 levels "Aichi Prefecture",..: 31 31 31 11 31 31 11 31 31 11 ...  
## $ en\_large\_area : Factor w/ 9 levels "China","East Sea",..: 1 1 1 1 1 1 1 1 1 1 ...

This dataset has 19,413 observations based on unique "coupon ID" with 29 variables including "large area" (geographical/location), "ken name" (another name for prefecture), "genre", "capsule name" (geographical/location),"price rate" (discount rate),"catalogue price", "discount price" (price after discount), "desp from", "disp end" ,"disp period" (dates of the dispensed coupon), "valid from", "valid end", "valid period" (coupons validity dates), "usable dates" (some coupons are usable on certain days),

To create a comprehensive dataset that includes all variables for Coupons, we need to merge the two datasets

cpdtr = coupon\_detail\_train\_en  
cpltr= coupon\_list\_train\_en  
train <- merge(cpdtr,cpltr)

names(train)

## [1] "COUPON\_ID\_hash" "en\_small\_area"   
## [3] "SMALL\_AREA\_NAME" "ITEM\_COUNT"   
## [5] "I\_DATE" "PURCHASEID\_hash"   
## [7] "USER\_ID\_hash" "large\_area\_name"   
## [9] "ken\_name" "small\_area\_name"   
## [11] "GENRE\_NAME" "CAPSULE\_TEXT"   
## [13] "PRICE\_RATE" "CATALOG\_PRICE"   
## [15] "DISCOUNT\_PRICE" "DISPFROM"   
## [17] "DISPEND" "DISPPERIOD"   
## [19] "VALIDFROM" "VALIDEND"   
## [21] "VALIDPERIOD" "USABLE\_DATE\_MON"   
## [23] "USABLE\_DATE\_TUE" "USABLE\_DATE\_WED"   
## [25] "USABLE\_DATE\_THU" "USABLE\_DATE\_FRI"   
## [27] "USABLE\_DATE\_SAT" "USABLE\_DATE\_SUN"   
## [29] "USABLE\_DATE\_HOLIDAY" "USABLE\_DATE\_BEFORE\_HOLIDAY"  
## [31] "en\_capsule" "en\_genre"   
## [33] "en\_ken" "en\_large\_area"

I need to clean up some of the variables that I will exclude from our analysis. I will only include the following variables:

train <- train[,c("COUPON\_ID\_hash","USER\_ID\_hash","en\_genre","DISCOUNT\_PRICE","PRICE\_RATE","USABLE\_DATE\_MON","USABLE\_DATE\_TUE","USABLE\_DATE\_WED","USABLE\_DATE\_THU","USABLE\_DATE\_FRI","USABLE\_DATE\_SAT","USABLE\_DATE\_SUN","USABLE\_DATE\_HOLIDAY","USABLE\_DATE\_BEFORE\_HOLIDAY","en\_large\_area", "en\_ken","en\_small\_area","ITEM\_COUNT")]

str(train)

## 'data.frame': 90916 obs. of 18 variables:  
## $ COUPON\_ID\_hash : Factor w/ 19368 levels "000eba9b783cec10658308b5836349f6",..: 1 2 2 2 2 2 2 2 2 2 ...  
## $ USER\_ID\_hash : Factor w/ 22782 levels "0000b53e182165208887ba65c079fc21",..: 3407 14136 2146 20673 6675 462 22395 10397 4410 13889 ...  
## $ en\_genre : Factor w/ 13 levels "Beauty","Delivery service",..: 7 2 2 2 2 2 2 2 2 2 ...  
## $ DISCOUNT\_PRICE : int 3500 1575 1575 1575 1575 1575 1575 1575 1575 1575 ...  
## $ PRICE\_RATE : int 51 78 78 78 78 78 78 78 78 78 ...  
## $ USABLE\_DATE\_MON : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_TUE : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_WED : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_THU : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_FRI : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_SAT : int 2 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_SUN : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_HOLIDAY : int 1 NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_BEFORE\_HOLIDAY: int 2 NA NA NA NA NA NA NA NA NA ...  
## $ en\_large\_area : Factor w/ 9 levels "China","East Sea",..: 2 6 6 6 6 6 6 6 6 6 ...  
## $ en\_ken : Factor w/ 47 levels "Aichi Prefecture",..: 23 41 41 41 41 41 41 41 41 41 ...  
## $ en\_small\_area : Factor w/ 55 levels "Aichi","Akasaka, Roppongi, Azabu",..: 50 43 43 43 43 43 43 43 43 43 ...  
## $ ITEM\_COUNT : int 2 1 1 1 1 1 1 1 1 1 ...

Now let us merge the resulted dataset that contains all coupons variables with our *"users"* dataset to create a dataset that contains all users and coupons. This dataset will help us understand relations between users and the type of coupons.

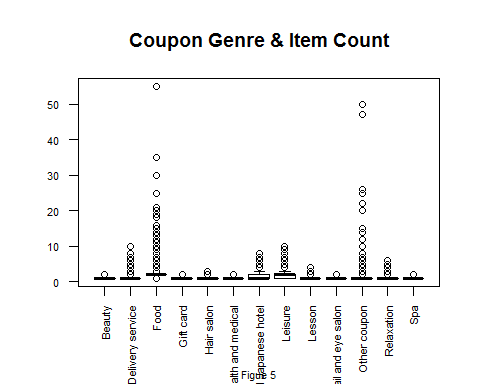
users <-user\_list\_en[,c("SEX\_ID","AGE","USER\_ID\_hash","age\_group")]  
train.2 <- merge(train, users)

The result of merging the previous datasets is a new dataset that has 19,368 "unique coupon ID" and 22,782 "unique users" with 21 variables.

Let us explore some of them.

***Is there a coupon genre that is more popular than another? let us see which coupons our users purchased the most?***

plot(train.2$en\_genre, train.2$ITEM\_COUNT, las=2,cex.names=0.2, cex.axis=0.7, main="Coupon Genre & Item Count", sub="Figue 5",cex.sub=0.7 )



We can determine from (Figure 5) that among the 13 genres we have, users purchased coupons in the "Food" genre in big counts, where some users purchased 10 or more coupons at a time, followed by "other coupons" (not clear about what other coupons genre is), "Delivery Services", "Hotels" and "Leisure".

***But if users buy more than 10 or 20 coupons at a time in a specific genre, does that mean that this genre the the most popular?***

describe(train.2$en\_genre)

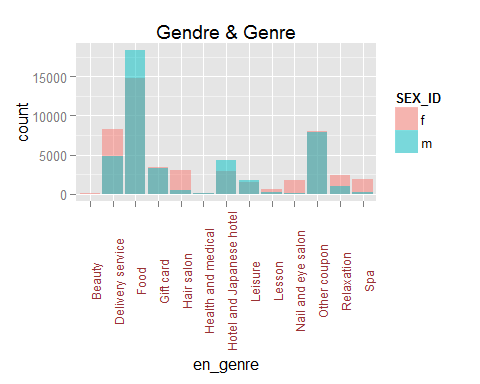
## train.2$en\_genre   
## n missing unique   
## 90916 0 13   
##   
## Beauty (143, 0%), Delivery service (13004, 14%)   
## Food (33040, 36%), Gift card (6724, 7%)   
## Hair salon (3507, 4%), Health and medical (160, 0%)   
## Hotel and Japanese hotel (7152, 8%)   
## Leisure (3199, 4%), Lesson (780, 1%)   
## Nail and eye salon (1862, 2%)   
## Other coupon (15950, 18%), Relaxation (3349, 4%)   
## Spa (2046, 2%)

Looking at the results above we conclude that the "Food" genre is the most popular which makes up 36% of the purchased coupons, followed by "other coupons" and then "Delivery Service".

Therefore, this confirms that the "Food" genre is the most popular.

***Do female users and male users share the same preference in genres?***

GG=ggplot(train.2, aes(x=en\_genre, fill=SEX\_ID)) + geom\_histogram(binwidth=.5, alpha=.5, position="identity") +theme(axis.text.x = element\_text(face="plain", color="#993333", size=9, angle=90))  
GG+ggtitle("Gendre & Genre")



From the figure above, we find that some genres are more popular with females than males, such as beauty salons, spas and delivery service. However, Food and hotel coupons are more popular with males.

***Does the Discount Rate of the coupons matter?***

Let us plot the price rate (discount rate), but before that I will recode our (Price Rate) to intervals of 5's. i will focus on price rate 30 and above, since there were very few coupons that were sold under this rate.

***Does the different "usable day" of the coupon makes some coupons more popular?***

variables with the (usable date) consist of three values (0,1 and 2), if we assume number(2) means certain kind of coupons with certain conditions, number (1) means usable that day and number(0) means not usable that day, we come up with the following findings after looking at each "usable day":

Coupons usability coded with number (2) were not purchased widely during the week (Mon, Tue, Wed, Thu and on holidays), but the purchase of these coupons increased on Friday and Sunday and very notably increased on Saturday and before the holiday.

After looking at the frequency of Coupons usability coded with number (1), i found that the usability on different days of the week does not make a significant difference on the purchase. All coupons with different usable dates were purchased similarly, except for the slight decrease on coupons "usable Saturday" which could be explained with the spike in purchases of coupons type (2) usable Saturdays

Monday

##   
## 0 1 2   
## 3399 51199 902

Tuesday

##   
## 0 1 2   
## 2548 51840 1112

Friday

##   
## 0 1 2   
## 2887 50025 2588

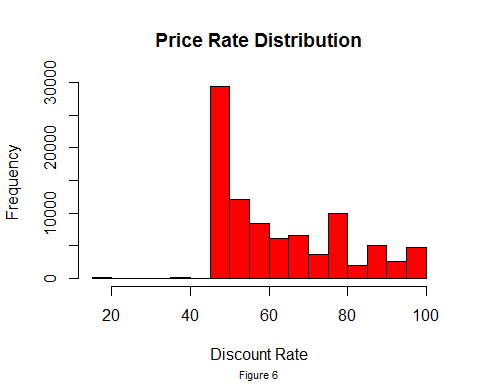
Sat

##   
## 0 1 2   
## 5540 43598 6362

attach(train.2)

train.2$Discount\_Rate[PRICE\_RATE<=19 & PRICE\_RATE>=0]= "15"  
train.2$Discount\_Rate[PRICE\_RATE<=29 & PRICE\_RATE>=20]= "20"  
train.2$Discount\_Rate[PRICE\_RATE=30]= "30"  
train.2$Discount\_Rate[PRICE\_RATE<=35 & PRICE\_RATE>30]= "35"  
train.2$Discount\_Rate[PRICE\_RATE<=40 & PRICE\_RATE>=36]= "40"  
train.2$Discount\_Rate[PRICE\_RATE<=45 & PRICE\_RATE>=41]= "45"  
train.2$Discount\_Rate[PRICE\_RATE<=50 & PRICE\_RATE>=46]= "50"  
train.2$Discount\_Rate[PRICE\_RATE<=55 & PRICE\_RATE>=51]= "55"  
train.2$Discount\_Rate[PRICE\_RATE<=60 & PRICE\_RATE>=56]= "60"  
train.2$Discount\_Rate[PRICE\_RATE<=65 & PRICE\_RATE>=61]= "65"  
train.2$Discount\_Rate[PRICE\_RATE<=70 & PRICE\_RATE>=66]= "70"  
train.2$Discount\_Rate[PRICE\_RATE<=75 & PRICE\_RATE>=71]= "75"  
train.2$Discount\_Rate[PRICE\_RATE<=80 & PRICE\_RATE>=76]= "80"  
train.2$Discount\_Rate[PRICE\_RATE<=85 & PRICE\_RATE>=81]= "85"  
train.2$Discount\_Rate[PRICE\_RATE<=90 & PRICE\_RATE>=86]= "90"  
train.2$Discount\_Rate[PRICE\_RATE<=95 & PRICE\_RATE>=91]= "95"  
train.2$Discount\_Rate[PRICE\_RATE<=100 & PRICE\_RATE>=96]= "100"

train.2$Discount\_Rate=as.numeric(train.2$Discount\_Rate)  
hist(train.2$Discount\_Rate, main="Price Rate Distribution", col="Red", sub="Figure 6", cex.sub=0.7, xlab="Discount Rate")



From Figure 6, we can clearly see the popularity of the (50%) discount rate with the highest frequency, followed by discount rate (55%) and (80%)

## Inferential Aspect

In the next step, I did some inferential analysis to find the effect of explanatory variables on Price Rate, for this purpose I used a linear regression model.

***Does the AGE or GENDER of users have an influence on their choice of the discount rate?***

* Influence of AGE on Discount Rate

discount.age = lm(Discount\_Rate~AGE)  
summary(discount.age)

##   
## Call:  
## lm(formula = Discount\_Rate ~ AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -54.156 -13.448 -5.351 12.574 41.049   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.787973 0.207012 351.61 <2e-16 \*\*\*  
## AGE -0.172956 0.004535 -38.14 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.87 on 90843 degrees of freedom  
## (71 observations deleted due to missingness)  
## Multiple R-squared: 0.01576, Adjusted R-squared: 0.01575   
## F-statistic: 1455 on 1 and 90843 DF, p-value: < 2.2e-16

From the univariate regression model (discount.age) to test the effect of age on discount rate, I realize that each year increase in age causes a decrease in discount rate by 1.17, so young users seek higher discount rate.

discount.age.gender = lm(Discount\_Rate~AGE+SEX\_ID)  
summary(discount.age.gender)

##   
## Call:  
## lm(formula = Discount\_Rate ~ AGE + SEX\_ID)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -54.414 -13.431 -5.363 12.580 41.285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.904015 0.207402 351.510 <2e-16 \*\*\*  
## AGE -0.166194 0.004605 -36.088 <2e-16 \*\*\*  
## SEX\_IDm -0.893182 0.107202 -8.332 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.86 on 90842 degrees of freedom  
## (71 observations deleted due to missingness)  
## Multiple R-squared: 0.01651, Adjusted R-squared: 0.01649   
## F-statistic: 762.6 on 2 and 90842 DF, p-value: < 2.2e-16

In this multivariate regression model we find that the effects of both sex and age are significant on discount rate (P < 0.05), which means that after controlling AGE, GENDER has negative change , so our female users (reference group) seek higher discount rates than our male users.

***Does that mean that female users seek cheaper items, or do they just go after a good discount rate?***

If we run another linear regression with the "PRICE" as the dependent variable like the following model.

price.age.gender = lm(DISCOUNT\_PRICE~AGE+SEX\_ID)  
summary(price.age.gender)

##   
## Call:  
## lm(formula = DISCOUNT\_PRICE ~ AGE + SEX\_ID)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2864 -1777 -769 530 98078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1753.5656 43.2673 40.529 < 2e-16 \*\*\*  
## AGE 14.0569 0.9607 14.631 < 2e-16 \*\*\*  
## SEX\_IDm -183.2086 22.3628 -8.193 2.59e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3310 on 90913 degrees of freedom  
## Multiple R-squared: 0.002705, Adjusted R-squared: 0.002683   
## F-statistic: 123.3 on 2 and 90913 DF, p-value: < 2.2e-16

we can infer from the results above that for each year older in age the price increases . The mean increase of the price is 14.0569 (14.1) for each year increase in age (adjusted for the sex).

And interestingly, males seek cheaper items (on average 183) comparing with the price our female users seek (adjusted for the age).

I will include the Discount Rate in the model, after i test the correlation between Price Rate and Discount Rate

cor(DISCOUNT\_PRICE,PRICE\_RATE)

## [1] -0.2550081

The correlation is not high, therefore I can include the discount rate in the regression.

price.age.gender.rate = lm(DISCOUNT\_PRICE~AGE+SEX\_ID+Discount\_Rate)  
summary(price.age.gender.rate)

##   
## Call:  
## lm(formula = DISCOUNT\_PRICE ~ AGE + SEX\_ID + Discount\_Rate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3151 -1433 -775 290 97935   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5446.4204 64.4216 84.543 < 2e-16 \*\*\*  
## AGE 5.6521 0.9377 6.027 1.67e-09 \*\*\*  
## SEX\_IDm -226.8563 21.6828 -10.462 < 2e-16 \*\*\*  
## Discount\_Rate -50.7041 0.6708 -75.586 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3207 on 90841 degrees of freedom  
## (71 observations deleted due to missingness)  
## Multiple R-squared: 0.06172, Adjusted R-squared: 0.06169   
## F-statistic: 1992 on 3 and 90841 DF, p-value: < 2.2e-16

The effect of sex on price rate highly related to the discount rate. When I put the discount rate in the model, the regression coefficient for sex changes dramatically.

But the Adjusted R-squared= 0.06169 which means that I have explained only 6 % of the variability in DISCOUNT\_PRICE by these three variable.

***But do other variables have an effect on the purchase transaction (YES purchase = 1, NO purchase = 0) of the coupon?***

Ponpare provided us with a huge dataset *coupon visit train* that contains the viewing logs for every visit a User made during a period of time that spans from 2011-07-01 to 2012-06-23. This dataset has 2,833,180 observations with 8 variables, which makes it a very large dataset to work with.

The variable of interest in this dataset is (Purchase Flag) which records whether the user purchased a coupon (value "1") or did not purchase the coupon (value "0") for every visit.

The other variables describe more details about each user's visit to the website "I DATE" (date of the visit), "PAGE SERIAL", "REFERRER hash", "COUPON ID hash","USER ID hash", "SESSION ID hash","PURCHASEID hash" (record of the purchase transaction).I will not use these variables in my analysis.

To create a more manageable dataset, I subtracted a smaller dataset from the *coupon visit train*, I subset it based on a period of a week (May 13,2013 to 19,2013), then I merged the week dataset with the other dataset (which contains all variables of coupons and users), I did the merge based on the unique (coupon ID hash).

str(train.W1)

## 'data.frame': 51968 obs. of 28 variables:  
## $ SEX\_ID : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AGE : int 18 18 18 19 19 19 19 19 19 19 ...  
## $ COUPON\_ID\_hash : Factor w/ 2033 levels "000fb2d3790cd8860e337290db189abe",..: 1654 1680 1680 67 964 964 964 1057 1791 1883 ...  
## $ USER\_ID\_hash : Factor w/ 6599 levels "000cc06982785a19e2a2fdb40b1c9d59",..: 6330 6330 6330 2249 2249 2249 2249 2249 2249 2249 ...  
## $ PURCHASE\_FLG : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ I\_DATE : Factor w/ 48732 levels "2012-05-13 00:00:05",..: 11499 19825 23833 5939 14761 14773 22165 22138 5934 29826 ...  
## $ DATE : Factor w/ 7 levels "2012-05-13","2012-05-14",..: 3 3 4 2 3 3 4 4 2 5 ...  
## $ en\_pref\_name : Factor w/ 47 levels "Aichi Prefecture",..: 33 33 33 9 9 9 9 9 9 9 ...  
## $ ITEM\_COUNT : int NA NA NA NA NA NA NA NA NA NA ...  
## $ PRICE\_RATE : int NA NA NA NA NA NA NA NA NA NA ...  
## $ CATALOG\_PRICE : int NA NA NA NA NA NA NA NA NA NA ...  
## $ DISCOUNT\_PRICE : int NA NA NA NA NA NA NA NA NA NA ...  
## $ DISPPERIOD : int NA NA NA NA NA NA NA NA NA NA ...  
## $ VALIDPERIOD : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_MON : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_TUE : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_WED : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_THU : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_FRI : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_SAT : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_SUN : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_HOLIDAY : int NA NA NA NA NA NA NA NA NA NA ...  
## $ USABLE\_DATE\_BEFORE\_HOLIDAY: int NA NA NA NA NA NA NA NA NA NA ...  
## $ en\_capsule : Factor w/ 18 levels "Delivery service",..: NA NA NA NA NA NA NA NA NA NA ...  
## $ en\_genre : Factor w/ 12 levels "Delivery service",..: NA NA NA NA NA NA NA NA NA NA ...  
## $ en\_small\_area : Factor w/ 54 levels "Aichi","Akasaka, Roppongi, Azabu",..: NA NA NA NA NA NA NA NA NA NA ...  
## $ en\_ken : Factor w/ 46 levels "Aichi Prefecture",..: NA NA NA NA NA NA NA NA NA NA ...  
## $ en\_large\_area : Factor w/ 9 levels "China","East Sea",..: NA NA NA NA NA NA NA NA NA NA ...

Now that we have a variable with a binary value, let us run a logistic regression to see if our variables have an effect on this response variable.

We will start with *GENDER*.

log.gender=glm(PURCHASE\_FLG~SEX\_ID, train.W1, family="binomial")  
summary(log.gender)

##   
## Call:  
## glm(formula = PURCHASE\_FLG ~ SEX\_ID, family = "binomial", data = train.W1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.2964 -0.2964 -0.2964 -0.2791 2.5553   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.10320 0.02915 -106.469 < 2e-16 \*\*\*  
## SEX\_IDm -0.12260 0.04489 -2.731 0.00631 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 17730 on 51967 degrees of freedom  
## Residual deviance: 17722 on 51966 degrees of freedom  
## AIC: 17726  
##   
## Number of Fisher Scoring iterations: 6

We can see that the slope for sex is negative (female is the reference group), meaning that the probability of event (PURCHASE\_FLG) is higher for females.

Now let us test for *AGE*

log.age=glm(PURCHASE\_FLG~AGE, train.W1, family="binomial")  
summary(log.age)

##   
## Call:  
## glm(formula = PURCHASE\_FLG ~ AGE, family = "binomial", data = train.W1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.3024 -0.2920 -0.2886 -0.2848 2.5610   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.300741 0.099298 -33.241 <2e-16 \*\*\*  
## AGE 0.002983 0.001996 1.495 0.135   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 17730 on 51967 degrees of freedom  
## Residual deviance: 17727 on 51966 degrees of freedom  
## AIC: 17731  
##   
## Number of Fisher Scoring iterations: 6

The p value is not significant so the probability of PURCHASE FLG does not relate to the age of users.

## BUILDING A MODEL FOR CLUSTERING

### Creating a sample

To prepare a smaller sample that is representative of the users, I opted to base the sample on the top 100 users.

I created a new variable based on the "frequency" of the user and picked the highest 100 users

y = count(train.2, "USER\_ID\_hash")

train.3 = merge(train.2, y)  
train.4 = train.3[order(train.3$freq),]  
train.5 = subset(train.4, !duplicated(train.4$USER\_ID\_hash))

sample.1=tail(train.5,100)

Now that we have the frequency of the users as a variable, let us test if 'Gender' and 'Age' have an effect on determining the frequency of a user

reg.freq=lm(freq~AGE + SEX\_ID, sample.1)  
summary(reg.freq)

##   
## Call:  
## lm(formula = freq ~ AGE + SEX\_ID, data = sample.1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.537 -5.429 -2.449 5.460 22.242   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.95044 3.21519 14.292 <2e-16 \*\*\*  
## AGE -0.03058 0.06902 -0.443 0.659   
## SEX\_IDm 1.35157 1.44864 0.933 0.353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.191 on 97 degrees of freedom  
## Multiple R-squared: 0.01006, Adjusted R-squared: -0.01035   
## F-statistic: 0.4929 on 2 and 97 DF, p-value: 0.6124

From the result of the regression we can see that male users are more frequent users than females with 1.35 males to 1 females after controlling the age.

sample.3 = sample.2[,-c(1,2,5,17,21)]

y.2=y[order(y$freq),]  
freq.user=tail(y.2,100)  
sample.6=merge(freq.user, train.2)

To prepare the sample for clustering, I excluded the User Id, Coupon Id. I also used only one variable "en ken" that contains geographical location information, which I find sufficient to provide the location of our users. Therefore, I excluded the rest of the geographical location variables "large area" and "small area" to avoid increasing the dimensional of the dataset

sample.7=sample.6[,c(4,5,6,7,8,9,10,11,12,13,14,15,17,20,22)]

We have NA's in the "Usable Days" variable especially in the "Service Delivery Genre", so it is safe to assume that the value of the "usable days" could be "1" since it is a delivery service and it should be usable any day of the week.

sample.8 = sample.7  
sample.8[is.na(sample.8)]<-1

We have 13 levels in the "genre" variables. Looking at them closely we can recode some of them to minimize the levels for easier analysis.

attach(sample.8)

en\_genre\_recode = recode(en\_genre, "c('Beauty', 'Nail and eye salon') = 'Beauty';c('Delivery service')='Delivery service';c('Food')='Food';c('Lesson','Health and medical','Other coupon')='Other coupon';c('Spa','Relaxation')='Spa Relaxation';c('Hotel and Japenese hotel')='Hotel and Japenese hotel';c('Leisure')='Leisure';c('Hair salon')='Hair salon'")  
sample.9=cbind(sample.8, en\_genre\_recode)

sample.10=sample.9  
sample.10$en\_genre = NULL

I included "Beauty" and "Nail and eye salon" under the same genre "Beauty". Considering the low number of coupons in the "lesson", "health and medical" I included both genres under the genre "other coupon", I also included "spa" and "relaxation" under one genre "spa relaxation".

Now we have only 9 genres.

unique(sample.10$en\_genre\_recode)

## [1] Food Other coupon   
## [3] Gift card Delivery service   
## [5] Hotel and Japanese hotel Spa Relaxation   
## [7] Hair salon Leisure   
## [9] Beauty   
## 9 Levels: Beauty Delivery service Food Gift card ... Spa Relaxation

I decided to exclude the 9 variables that indicate the "usable date" for two reasons:

1. The results of my descriptive analysis showed that there is no significant variance in the amount of purchased coupons among the different usable dates. There was a very slight difference and I consequently determined that the "usable dates" were not big factors that influence the purchase of the coupons
2. These variables contain values of "0, 1 and 2". There was no explanation provided by the company as to what value "2" means in this context, I assumed it means a coupon with special terms. Taking into consideration this lack of information, the results of any model will be hard to interpret properly.

sample.11 = sample.10[,-c(3:11)]

### Clustering

We will create clusters based on the following variables:

names(sample.11)

## [1] "DISCOUNT\_PRICE" "PRICE\_RATE" "en\_ken" "SEX\_ID"   
## [5] "age\_group" "en\_genre\_recode"

Since our dataset includes categorical and numeric variables we need to assign dummy variables to the factor

sample.11[,"dummy"]=1  
sample.12=as.data.frame(model.matrix(dummy~.,sample.11))  
sample.12[,1]=NULL

Now I will split the dataset to 70% for "training" and the remaining for "test".

rn\_train <- sample(nrow(sample.12),floor(nrow(sample.12)\*0.7))  
sample.12.train<-sample.12[rn\_train,]  
sample.12.test<-sample.12[-rn\_train,]

### Hierarchical Clustering

Now that our dataset is ready, let us compute the distance between all data points using the "euclidean" distance

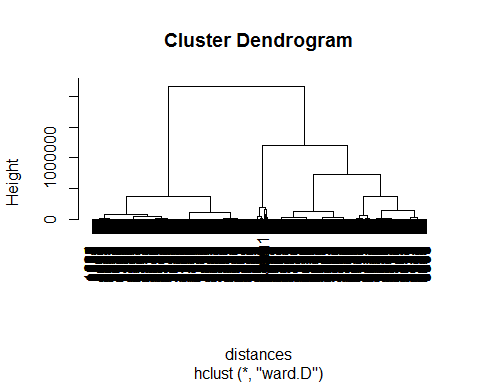
distances = dist(sample.12.train, method="euclidean")

Now we will use the hierarchical clustering using method "ward.D" which cares about he distances between clusters using centroid distance and also the variance in each of the clusters.

clustersample.12 = hclust(distances, method = "ward.D")

let us have a look at the dendrogram to help us pick the number of clusters.

plot(clustersample.12)



I will use 10 clusters, considering i have 100 unique users (10^2 = 100)

hc.group = cutree(clustersample.12, k=10)

now let us explore our clusters.

cluster.1 = subset(sample.12.train, hc.group==1)

1. Price mean = 878.12
2. Price range = 550 to 1150
3. Age group = mean = 41 (mostly in 50's)
4. Gender = female and male are balanced in this cluster
5. Area= most user are from Tokyo =49%, Osaka = 20% , Fukuoka 5% Genre = Food (72%) , gift card 7%, delivery service 8%, leisure 4%

cluster.2 = subset(sample.12.train, hc.group==2)

1. Price mean = 4000
2. Price range = 2680 to 4000
3. Age Group = mostly in their 30's and 50's
4. Gender = predominantly female (302f to 170m)
5. Area = Tokoy 65%, Osaka = 14%, Kanagawa 4%, Hokkaido 3%, Hyogo 2%,Fukuoka 2%, Genre = Food 48%, Spa 16%, Delivery Service 8%, Hair = 7%

cluster.3 = subset(sample.12.train, hc.group==3)

1. Price mean = 330
2. Age Group = mostly in their 40's
3. Gender = slightly more male
4. Area = Tokyo 80%, Osaka 7%, Fukuoka 4%, Kanagawa 1%, Koyoto 1%
5. Genre = Food 70%, other 20%, Gift cards 4%

cluster.5 = subset(sample.12.train, hc.group==5)

1. Price mean = 2377
2. Price range = 2050 to 2625
3. Age group = mostly in 50's
4. Gender = more males
5. Area = Tokyo 45%, Osaka 20%, Kyoto 5%, Genre = Food 80%, Delivery Service5%,

cluster.7 = subset(sample.12.train, hc.group==7)

1. Price mean = 5286
2. Price range = 4200 to 7500
3. Age Group = mostly in their 50's
4. Gender = more female
5. Area = Tokyo 50% , Osaka 12% , Kyoto 7%, Hokkaido 5%, Fukuoka 4%, Kanagawa 4% , Hyogo 3%
6. Genre = Food 31%, Hotel 27%, Spa 13%, Hair 12%,

### K-Mean Clustering

Using the Kmean, I created 7 cluster

KMC = kmeans(sample.12.train, centers=7, iter.max=1000)

In the ("felxclust") package we have the object class KCCA (K-Centroids Cluster Analysis). We need to convert the information from the clustering algorithim to an object of the class KCCA, this is needed before using the "predict" function on the test set

KMC.kcca = as.kcca(KMC,sample.12.train)  
user.clustors = predict(KMC.kcca, newdata=sample.12.test)

Now We will add the clusters as a column to our train and test.

train.validation = cbind(sample.12.train, KMC$cluster)  
test.validation = cbind(sample.12.test, user.clustors)

Let us explore our K-MEAN clusters

kmcluster.1 = train.validation[KMC$cluster==1,]

**Kmean Cluster 1**

1. Price mean= 430
2. Age = mostly in the 40 and 50
3. Gender = slightly more males
4. Area = Tokyo 69% , Osaka 14%, Fukuoka 4%, Hokkiado 4%, Kanagawa 3%,
5. Genre = Food 74%, other 14%, gift card 3%,

kmcluster.3 = train.validation[KMC$cluster==3,]

**Kmean Cluster 3**

1. Price mean = 4763
2. Price range = 3675 to 7000
3. Age = 50s Gender = more females
4. Area = Tokyo 57% ,Osaka 10%,Kyoto 6%,Hokkiado 4%,Fukuoka 4%, Kanagawa 4%, Saga 2%
5. Genre = Food 32%, Hotel 20%, Spa 13%, Hair 12%, Delivery service 7%

kmcluster.6 = train.validation[KMC$cluster==6,]

**Kmean Cluster 6**

1. Price range = 1890 to 3625
2. Age = 50s Gender = more females
3. Area = Tokyo 55%, Osaka 21%, Kanagawa 4%, Fukuoka 4%, Hokkiado 3%, Hyogo 2% Genre = Food 68%, other 12%, Spa 10%, Delivery service 7%, Hair 2%

#### Testing

Let us see if we have similar results when we look at the test clusters

kmcluster.test.3 = test.validation[user.clustors==3,]

1. Price mean = 4,737
2. Gender = more females
3. Area = Tokyo 60% , Osaka 17%, Hokkaido 3%, Kanagawa 2%
4. Genre= Food 36%, Hotel 20%, Hair 13%, Spa 11%

kmcluster.test.6 = test.validation[user.clustors==6,]

1. Price mean= 2,493
2. Gender = slightly more females
3. Area = Tokyo 49%, Osaka 18%, Hokkaido 9%, Kanagawa 5%, Fukuoka 5%
4. Genre = Food 65%, Spa 13%, Delivery service 7%, Hair 3%, Hotel 3%

kmcluster.test.7 = test.validation[user.clustors==7,]

1. Price mean = 1,282
2. Gender = more males
3. Area = Tokyo 49%, Osaka 18%, Hokkaido 9%,Kanagwa 5%,
4. Genre = Food 70%, Delivery service 14%, Gift cards 5%, Spa 3%