ITESO

Exploración Datos Ponpare

PAP Modelación Matemática para el desarrollo de planes y proyectos de negocio

Sara Eugenia Rodríguez Reyes

10/10/2015

Introducción

Ponpare es el sitio líder de cupones en Japón, ofrece grandes descuentos en diversos productos.

Lo que esta compañía busca es predecir los cupones que un cliente va a comprar en un periodo de tiempo determinado.

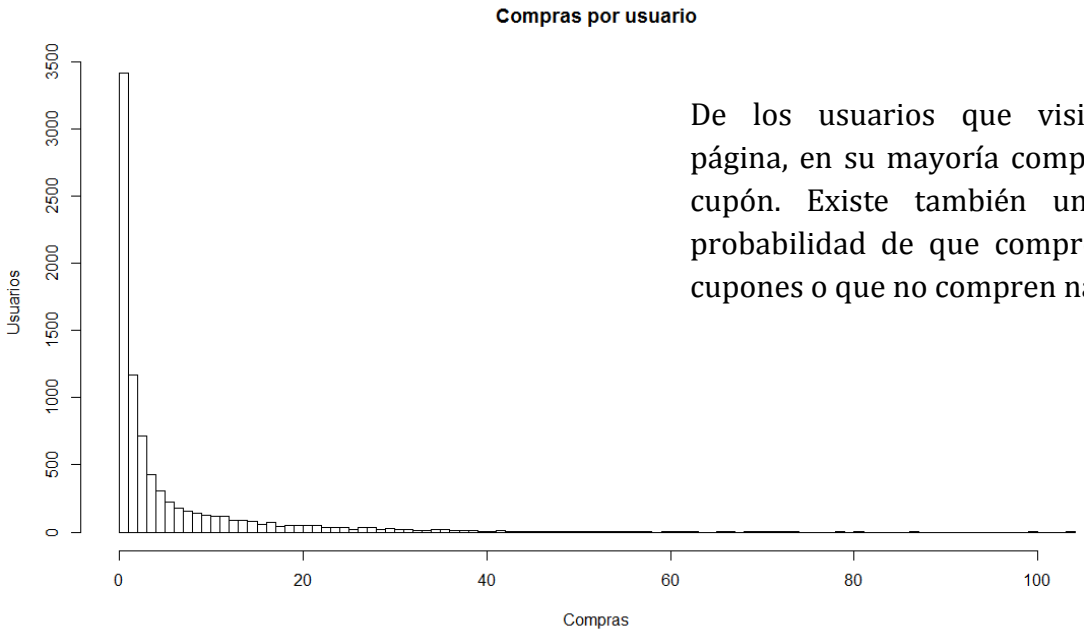
Para poder predecir esto, primero es necesario conocer los datos y explorarlos. Que es lo que se hará en esta primera etapa.

Se tienen cinco archivos:

1. Coupon_list_train.csv: lista de cupones que son considerados como conjunto de entrenamiento
2. User_list.csv: lista de todos los usuarios en el conjunto de datos
3. Coupon_visit_train.csv: visitas de usuarios y que buscaron cupones durante el periodo de entrenamiento
4. Coupon_detail_train.csv: detalles de las compras de los usuarios que compraron

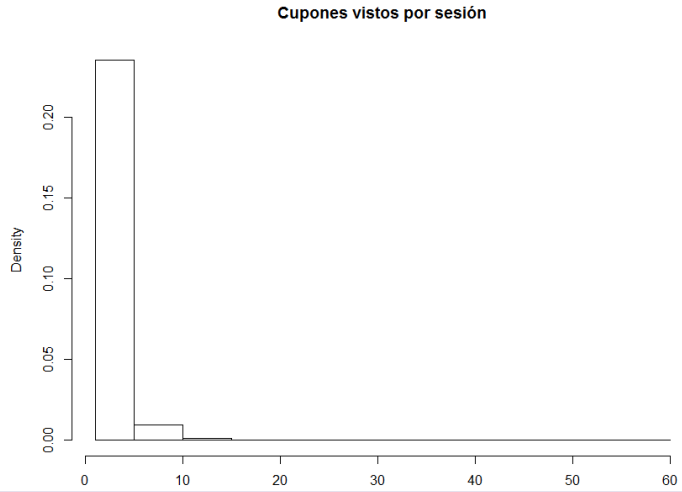
A continuación se presenta el código con el que se hizo la exploración de datos; este mismo es comentado. Se presentan gráficas con explicaciones de patrones encontrados en los datos.

```
##Fijar el directorio de trabajo
setwd("~/ITESO/PAP2/CouponPurchase")

#Cargar librerías
library(plyr)
library(dplyr)
library(ggplot2)
library (randomForest)
##Cargar archivos

##Listado del área cupones
CouponAreaTrain <- read.csv("coupon_area_train.csv",header=TRUE)
##Registro de compra de los usuarios que compran cupones
CouponDetailTrain <- read.csv("coupon_detail_train.csv",header=TRUE)
##Listado de cupones
CouponListTrain <- read.csv("coupon_list_train.csv",header=TRUE)
##Registro de búsqueda de cupones de los usuarios
CouponVisitTrain <- read.csv("coupon_visit_train.csv",header=TRUE)
##Lista de usuarios
UserList <- read.csv("user_list.csv",header=TRUE)

##EXPLORACIÓN DE DATOS

#Lista de usuarios que visitaron y compraron (varias veces)
Compradores   <- filter(CouponVisitTrain, PURCHASE_FLG == 1)


#Función para contar datos
cuentasIdsDif <- function(ids.vec){
  return(length(unique(ids.vec))) #Tamaño de datos únicos
}

usuario<- unique(CouponVisitTrain$USER_ID_hash)  #Vector de los usuarios
numUsuarios<-cuentasIdsDif(CouponVisitTrain$USER_ID_hash)#Cantidad de
usuarios

#Cuantas veces compró cada usuario
ComprasUsuario <- aggregate(PURCHASE_FLG ~ USER_ID_hash, data =
CouponVisitTrain, sum)
names(ComprasUsuario) <- c("Usuario", "Compras")
table(ComprasUsuario[,2])
hist(ComprasUsuario[,2],breaks=105, main="Compras por usuario",
xlab="Compras", ylab="Usuarios")
```



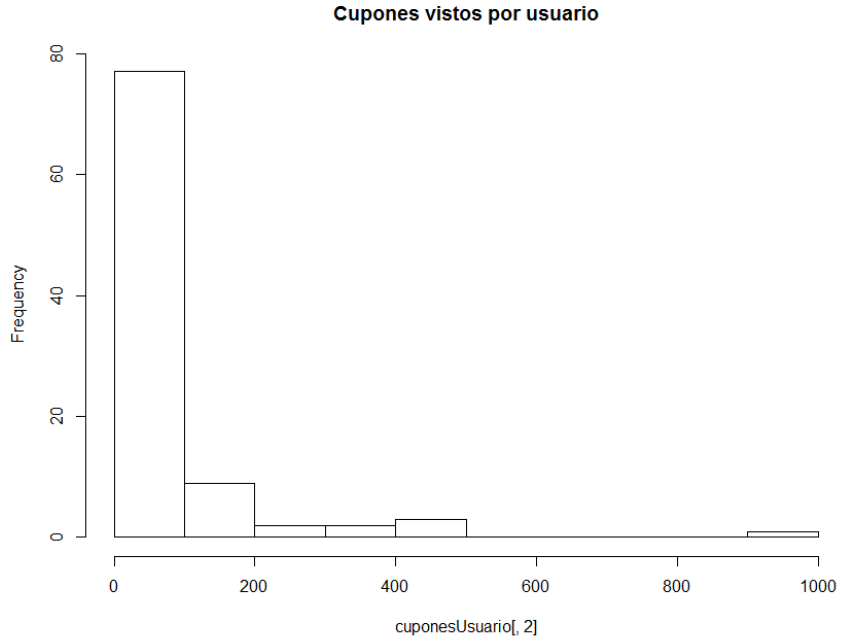
De los usuarios que visitan la página, en su mayoría compran un cupón. Existe también una alta probabilidad de que compren dos cupones o que no compren nada.

```
#Cuantos cupones se vieron por sesión
cuponesSesion <- aggregate(CouponVisitTrain$COUPON_ID_hash,
by=list(CouponVisitTrain$SESSION_ID_hash) , FUN=cuentasIdsDif)
names(cuponesSesion) <- c("Sesion", "Cupones Vistos")
breaks <- c(1,5,10,15,20,60)
hist(cuponesSesion[,2], breaks=breaks, freq=FALSE)
table(cuponesSesion[,2])
```



Cuando un usuario entra en la página, lo más probable es que visite de 1 a 5 cupones por sesión.

```
#Cuantos cupones vio cada usuario
cuponesUsuario <- aggregate(CouponVisitTrain$COUPON_ID_hash,
by=list(CouponVisitTrain$USER_ID_hash), FUN=cuentasIdsDif)
names(cuponesUsuario) <- c("Usuario", "Cupones Vistos")
hist(cuponesUsuario[,2], main="Cupones vistos por usuario")
```



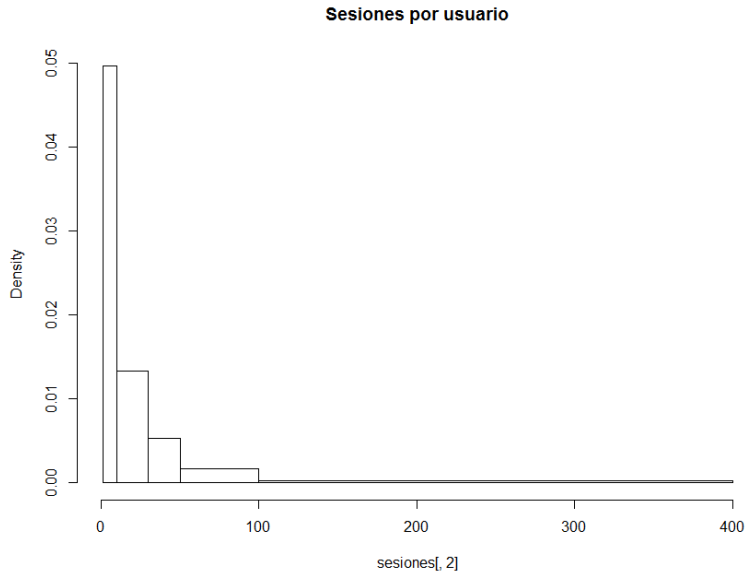
La mayoría de los usuarios ven de 1 a 20 cupones.

```
#Cuantas compras se hicieron por sesión
comprasSesion <- aggregate(CouponVisitTrain$PURCHASEID_hash,
by=list(CouponVisitTrain$SESSION_ID_hash) , FUN=cuentasIdsDif)
names(comprasSesion) <- c("Sesion", "Compras")
table(comprasSesion[,2])
```

Compras	1	2	3	4	5	6	11	21
Sesión	3191	310	42	10	1	3	1	1

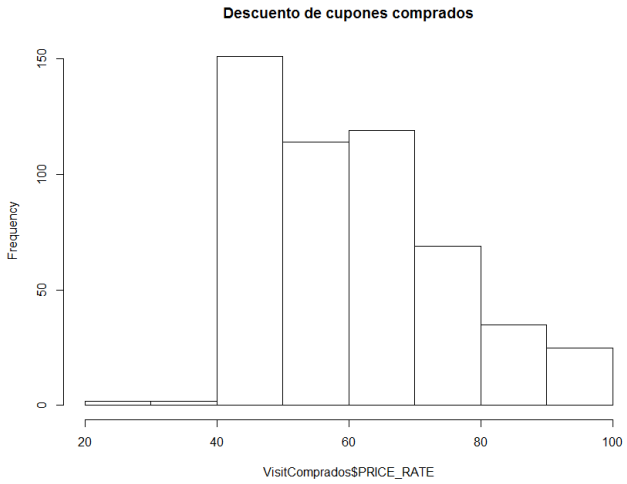
Cuando un usuario entra a una sesión, existe una mayor probabilidad de que solo realice una compra; en ocasiones puede hacer 2 o 3.

```
#Cuantas sesiones hizo cada usuario
sesiones <- aggregate(CouponVisitTrain$SESSION_ID_hash,
by=list(CouponVisitTrain$USER_ID_hash), FUN=cuentasIdsDif)
names(sesiones) <- c("Usuario", "Sesiones")
table(sesiones[,2])
breaks <- c(1,10,30,50,100,400)
hist(sesiones[,2], breaks=breaks)
```



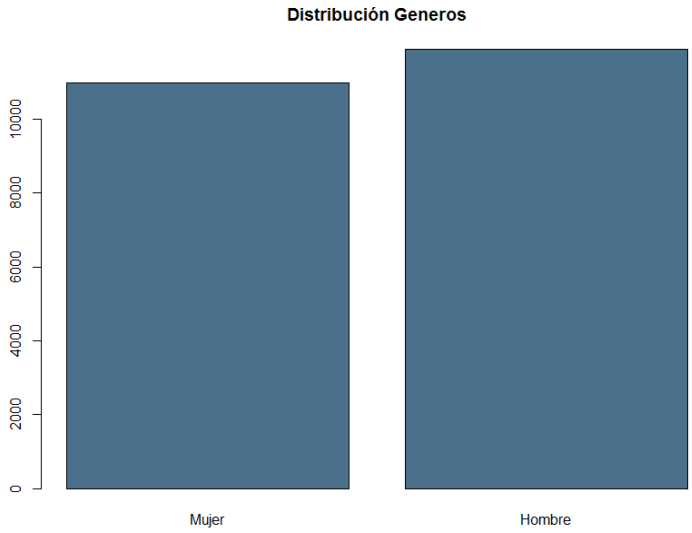
Existe una mayor probabilidad de que cada usuario realice entre 1 y 10 sesiones.

```
##Ver cómo influyen las promociones en las ventas
##Seleccionar cupones que se compraron
cuponesComprados <- CouponVisitTrain[CouponVisitTrain$PURCHASEID_hash!="",]
VisitComprados <- merge(CouponListTrain, cuponesComprados,
by="COUPON_ID_hash") #de CouponListTrain solo los cupones que se compraron
table(VisitComprados$PRICE_RATE)
hist(VisitComprados$PRICE_RATE, main="Descuento de cupones comprados")
```



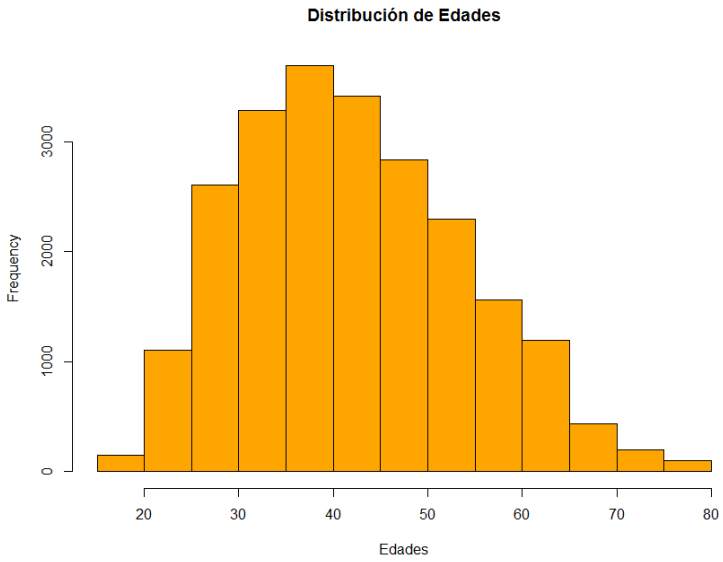
Si el descuento se encuentra entre el 40% y 50%, es más probable que se realice una compra de un cupón.

```
##Distribución de géneros
table(UserList$SEX_ID)
barplot(table(UserList$SEX_ID), names.arg = c("Mujer", "Hombre"), main =
"Distribución Generos", col = "skyblue4")
```



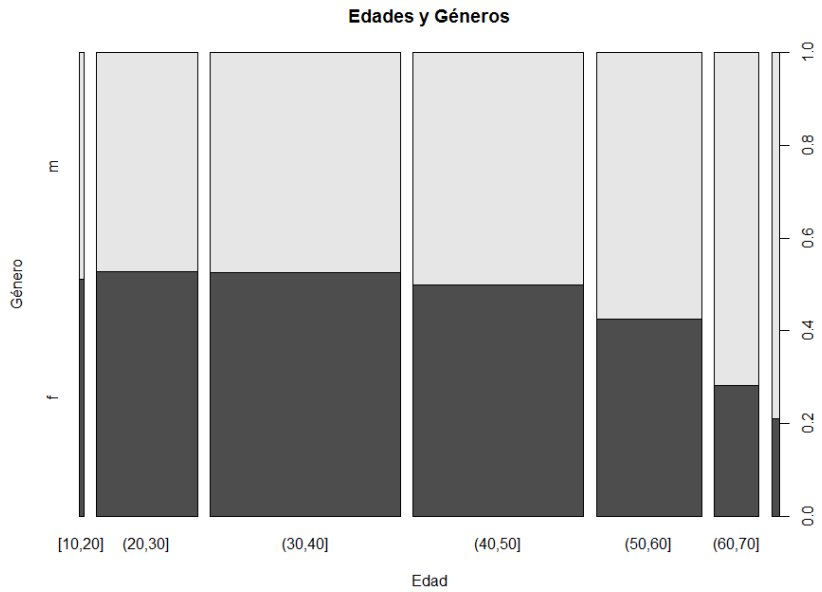
De los usuarios, 10983 son mujeres, mientras que 11890 son hombres.

```
## Histograma de edades
hist(UserList$AGE, main = "Distribución de Edades", xlab = "Edades", col =
"Orange")
```



Los usuarios que más utilizan el servicio están entre los 35 y 40 años. En Japón, el 38.1% de la población se encuentra en este rango de edad. En estas edades, las personas tienen un trabajo estable, y en su mayoría, tienen una familia que mantener por lo que el uso de cupones es mayor con la finalidad de ahorrar dinero.

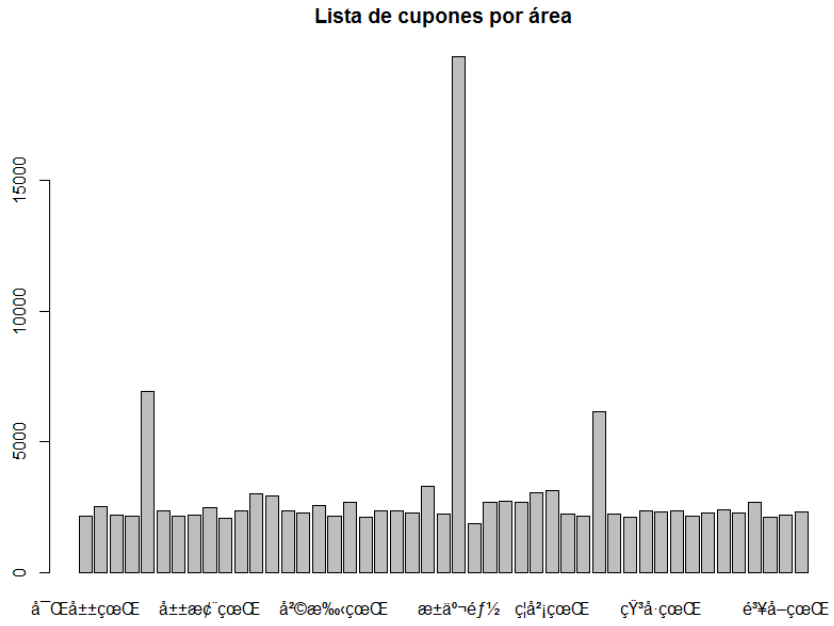
```
##Diferencia de usuarios por género y edades
edades <- c(10, 20, 30, 40, 50, 60, 70, 80)
#La Función cut sirve para cortar una variable continua en categórica (edad)
UserList$AGE_GROUP <- cut(UserList$AGE, edades, include.lowest = TRUE)
plot(UserList$AGE_GROUP,UserList$SEX_ID, main = "Edades y Géneros", xlab =
"Edad", ylab = "Género")
```




```
ggplot(UserList, aes(x=AGE, fill=SEX_ID)) + geom_histogram(binwidth=5,
position="dodge")
```

En las gráficas se puede observar, que mientras más avanza la edad de la gente, aumenta el número de hombres que utilizan el servicio. Esto se puede explicar ya que en Japón los hombres gozan de salarios mayores a los de las mujeres; mientras que muchas mujeres se dedican al hogar. Es por esto que los hombres tienen mayor posibilidad económica de comprar cupones.

```
##Distribución de cupones por área
barplot(table(CouponAreaTrain$PREF_NAME),main = "Lista de cupones por área")
```



En la gráfica anterior se puede observar las áreas en las que hay más cupones.

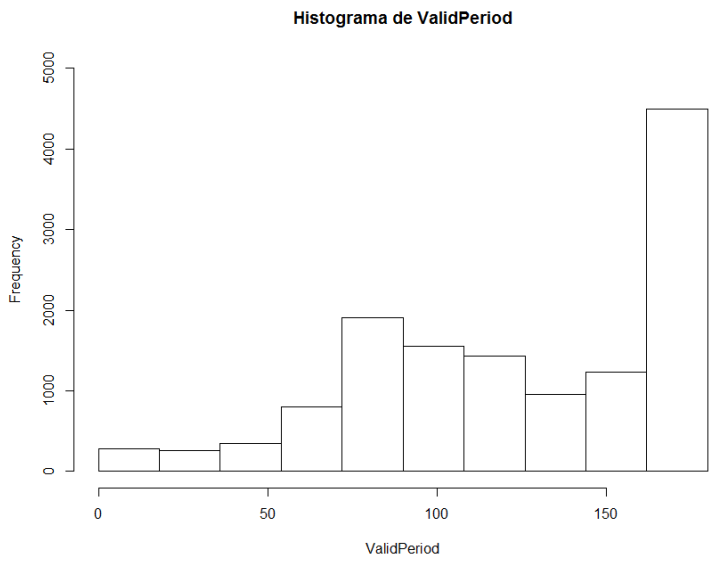
De la misma manera, la siguiente tabla muestra la frecuencia de cupones comprados por tienda.

```
##Frecuencia de cupones por tienda
couponTiendas <- as.data.frame(table(CouponAreaTrain$PREF_NAME))
colnames(couponTiendas) <- c("Tienda", "Frecuencia")
arrange(couponTiendas, desc(Frecuencia))
```

```
         Tienda Frecuencia
1     æ□±äº¬éƒ½      19720
2     å¤§é˜ªåºœ       6942
3   ç¥žå¥ˆå·□çœŒ       6152
4     æ„›çŸ¥çœŒ       3275
5     ç¦□å²¡çœŒ       3135
6     åŸ¼çŽ‰çœŒ       3035
7     å…µåº«çœŒ       3027
8     å□ƒè‘‰çœŒ       2922
9     åŒ—æµ·é□“       2715
10    äº¬éƒ½åºœ       2697
11    é□™å²¡çœŒ       2692
12    å²□é˜çœŒ       2682
13    åºƒå³¶çœŒ       2664
14    å²¡å±±çœŒ       2558
15    ä¸‰é‡□çœŒ       2527
16    å±±æ¢¨çœŒ       2466
```

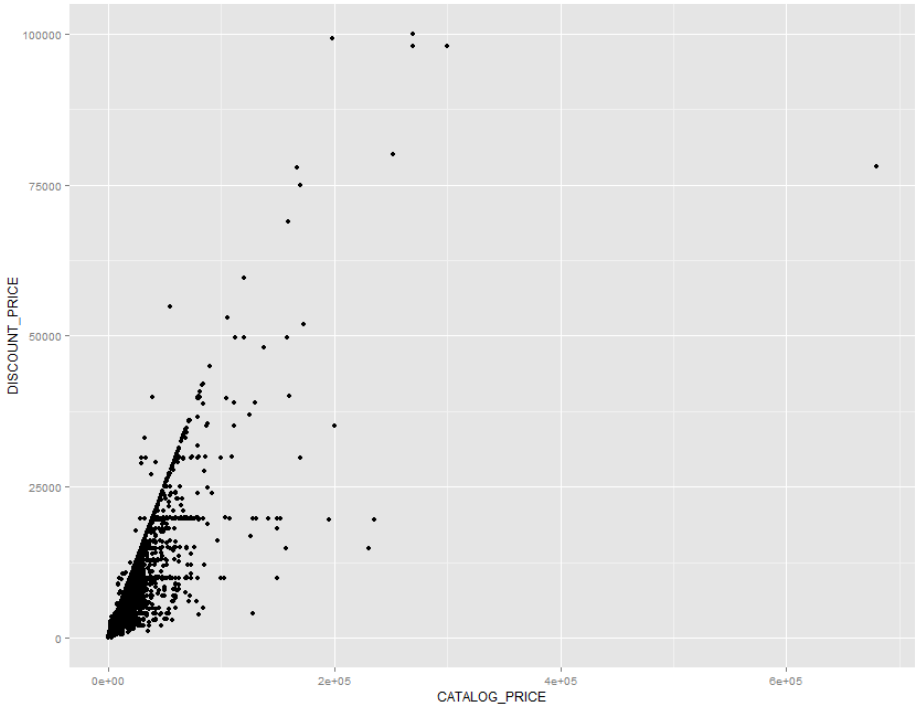
```
17    é•·é‡ŽçœŒ       2388
18    å®®åŸŽçœŒ       2374
19    æ ƒæœ¨çœŒ       2368
20    æ–°æ½ŸçœŒ       2364
21    é¦™å·□çœŒ       2357
22    å¥ˆè‰¯çœŒ       2347
23    ä½□è³€çœŒ       2335
24    ç¾¤é¦¬çœŒ       2335
25    çŸ³å·□çœŒ       2329
26    èŒ¨åŸŽçœŒ       2316
27    å¾³å³¶çœŒ       2282
28    é•·å´ŽçœŒ       2270
29    é□'æ£®çœŒ       2264
30    æ„›åª›çœŒ       2254
31    ç§‹ç”°çœŒ       2248
32    ç¦□å³¶çœŒ       2215
33    æ»‹è³€çœŒ       2213
34    é³¥å•–çœŒ       2206
35 å'ŒæŒå±±çœŒ      2196
36    å±±å½¢çœŒ       2194
37    å¯Œå±±çœŒ       2169
38    å²©æ‰‹çœŒ       2162
39    å±±å□£çœŒ       2159
40    å¤§å^†çœŒ       2144
41    ç¦□äº•çœŒ       2141
42    é«˜çŸ¥çœŒ       2135
43    ç†Šæœ¬çœŒ       2097
44    å³¶æ ¹çœŒ       2092
45 é¹¿å…□å³¶çœŒ       2092
46    å®®å´ŽçœŒ       2078
47    æ²–ç¸„çœŒ       1852
```

```
##Histograma del Periodo de Validación de los cupones
dfValidPeriod <- CouponListTrain
##Remplazamos los valores NA por el promedio
dfValidPeriod$VALIDPERIOD[is.na(dfValidPeriod$VALIDPERIOD)] <-
mean(dfValidPeriod$VALIDPERIOD)
breaksValidPeriod <- c(0,18,36,54,72,90,108,126,144,162,180)
hist(CouponListTrain$VALIDPERIOD,main="Histograma de
ValidPeriod",breaks=breaksValidPeriod, xlim=c(0,180),
ylim=c(0,5000),xlab="ValidPeriod")
```



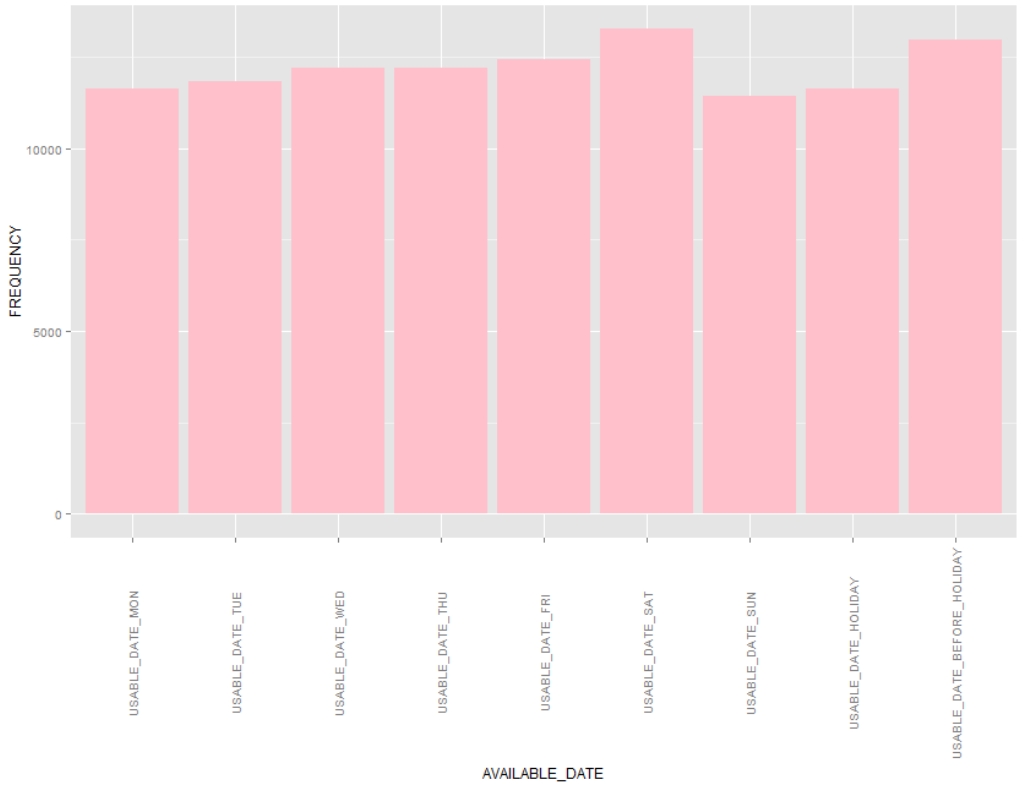
En el histograma anterior, se puede observar que la mayoría de los cupones tienen un periodo de validación entre 162 y 180 días (6 meses). Se infiere que a las personas les conviene más tener un cupón que sea válido durante un mayor rango de tiempo que si se vence entre una y dos semanas.

```
##Tranformar datos como fechas
CouponListTrain$VALIDFROM <- as.Date(CouponListTrain$VALIDFROM)
CouponListTrain$VALIDEND <- as.Date(CouponListTrain$VALIDEND)
CouponListTrain$DISPFROM <- as.Date(CouponListTrain$DISPFROM)
CouponListTrain$DISPEND <- as.Date(CouponListTrain$DISPEND)
ggplot(CouponListTrain)+geom_point(aes(x=CATALOG_PRICE, y = DISCOUNT_PRICE))
```



Se observa una relación lineal entre el precio catálogo y el precio de descuento. Mientras más aumenta el precio señalado en el catálogo, más aumenta el precio de descuento.

```
##Ver qué días se usan más los cupones
usable_dates <-
c("USABLE_DATE_MON","USABLE_DATE_TUE","USABLE_DATE_WED","USABLE_DATE_THU","U
SABLE_DATE_FRI","USABLE_DATE_SAT","USABLE_DATE_SUN","USABLE_DATE_HOLIDAY","U
SABLE_DATE_BEFORE_HOLIDAY")
sum <- data.frame(AVAILABLE_DATE = character(), FREQUENCY = integer())
for (i in usable_dates){
  frequency = sum(CouponListTrain[!is.na(CouponListTrain[,i]),][,i])
  sum = rbind(sum, data.frame(AVAILABLE_DATE = i, FREQUENCY = frequency))
}
ggplot(sum, aes(AVAILABLE_DATE,FREQUENCY)) + geom_bar(fill="pink",
stat="identity") + theme(axis.text.x  = element_text(angle=90, vjust=0.5))
```



Los días en que se usan los cupones con mayor frecuencia son los sábados y los días antes de vacaciones. Esto puede ser ya que los sábados son días en que no se trabajan y las personas tienen más tiempo para ver y comprar los cupones. De la misma forma, los días antes de vacaciones, la gente busca promociones para ahorrar dinero durante las vacaciones; ya sea paquetes vacacionales, transporte, comidas, etc.

```
##Análisis por edad
CouponDetailTrain$age <-
UserList[match(CouponDetailTrain$USER_ID_hash,UserList$USER_ID_hash),"AGE"]
subset_age_10 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 10 &
as.numeric(CouponDetailTrain$age) <= 20) ,]
mean(subset_age_10$age)
table(subset_age_10$age)
```

Entre 10 y 20 años, el promedio de edad en el que se compran cupones es de 19 años. Mientras que el mayor número de jóvenes dentro de este rango que compran cupones tienen 20 años.

Edad	15	16	17	18	19	20
#Personas	30	40	62	33	87	218

```
subset_age_20 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 20 &
as.numeric(CouponDetailTrain$age) <= 30) ,]
mean(subset_age_20$age)
table(subset_age_20$age)
```

Entre 20 y 30 años, el promedio de edad en el que se compran cupones es de 27 años. Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 29 años.

Edad	21	22	23	24	25	26	27	28	29	30
#Pers	323	811	922	1451	1587	2359	2762	3228	3278	3247

```
subset_age_30 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 30 &
as.numeric(CouponDetailTrain$age) <= 40) ,]
mean(subset_age_30$age)
table(subset_age_30$age)
```

Entre 30 y 40 años, el promedio de edad en el que se compran cupones es de 36 años. Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 39 años.

Edad	31	32	33	34	35	36	37	38	39	40
#Pers	3356	4357	3728	3654	3837	4626	4636	5099	5330	5103

```
subset_age_40 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 40 &
as.numeric(CouponDetailTrain$age) <= 50) ,]
mean(subset_age_40$age)
table(subset_age_40$age)
```

Entre 40 y 50 años, el promedio de edad en el que se compran cupones es de 45 años. Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 42 años.

Edad	41	42	43	44	45	46	47	48	49	50
#Pers	4998	6101	4817	4711	4813	4237	5025	5183	4778	5521

```
subset_age_50 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 50 &
as.numeric(CouponDetailTrain$age) <= 60) ,]
mean(subset_age_50$age)
table(subset_age_50$age)
```

Entre 50 y 60 años, el promedio de edad en el que se compran cupones es de 55 años. Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 51 años.

Edad	51	52	53	54	55	56	57	58	59	60
#Pers	4997	4931	4900	4353	3355	3365	3121	2837	2775	2534

```
subset_age_60 <- CouponDetailTrain[(as.numeric(CouponDetailTrain$age) > 60 &
as.numeric(CouponDetailTrain$age) <= 70) ,]
mean(subset_age_60$age)
table(subset_age_60$age)
```

Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 62 años.

Edad	61	62	63	64	65	66	67	68	69	70
#Pers	2512	2594	2140	2003	1940	1227	634	958	625	485

```
subset_age_70 <-CouponDetailTrain[(as.numeric(CouponDetailTrain$age)>70) ,]
mean(subset_age_70$age)
table(subset_age_70$age)
```

Entre 70 y 80 años, el promedio de edad en el que se compran cupones es de 74 años. Mientras que el mayor número de personas dentro de este rango que compran cupones tienen 71 años.

Edad	71	72	73	74	75	76	77	78	79	80
#Pers	593	490	290	162	150	189	129	129	196	34

Predicción

Para hacer la predicción se utilizará un modelo de Boosted Trees. La variable a predecir será purchase flag. La cual es 1 si hubo compra o 0 si no hubo.

Hacemos un data set nuevo uniendo CouponListTrain y CouponVisitTrain.

```
names(CouponListTrain)[length(CouponListTrain)] <- "COUPON_ID_hash"
data <- merge(CouponListTrain, CouponVisitTrain, by ="COUPON_ID_hash")
month <- month(data$DISPFROM)  #separar el mes de la fecha
data$month <- month
years <- year(data$DISPFROM)   #separar el año de la fecha
data$years <- years
weeks<- week(data$DISPFROM)    #separar la semana de la fecha
data$weeks <- weeks
```

Las variables a utilizar para la predicción son: Capsule Text, genre name, price rate, catalog price, discount price, validperiod, small area name, large area name, month, years y weeks.

```
data <- select(data,CAPSULE_TEXT,GENRE_NAME,PRICE_RATE,
                CATALOG_PRICE, DISCOUNT_PRICE, VALIDPERIOD,PURCHASE_FLG,
               small_area_name,large_area_name, month, years, weeks)

data$small_area_name <- as.integer(data$small_area_name)
data$CAPSULE_TEXT <- as.integer(data$CAPSULE_TEXT)
data$large_area_name <- as.integer(data$large_area_name)
data$GENRE_NAME <- as.integer(data$GENRE_NAME)
```

Se separa un set de entrenamiento con el 70% de los datos y otro set de prueba.

```
#Conjunto de entrenamiento y prueba
train <- sample(1:nrow(data),nrow(data)*0.7)
test <- data[-train,"PURCHASE_FLG"]
```

Para la creación del boosted tree, se prueba con distintos valores de número de árboles y de parámetros; para proseguir calculando el mínimo error y ver cuáles son los valores óptimos.

```
set.seed(1)
ntree <- c(25,100,250,500)
interact <- seq(1,4,1)
lambda <- seq(0.25,1,0.25)
error <- array(dim = c(4,4,4))
for (i in 1:length(ntree)){
  for(j in 1:length(interact)){
    for (k in 1:length(lambda)){
      Boost <- gbm(PURCHASE_FLG ~. , data =data[train,], distribution =
"gaussian",
                        n.trees = ntree[i], interaction.depth = interact[j],
shrinkage = lambda[k])
      yhatBoost <- predict(Boost, newdata=data[-train,], n.trees=ntree[i])
      mse <- mean((yhatBoost-test)^2)
      error[i,j,k] <- mse
    }
  }
}
```

El mínimo error es de 3.252093 % y se da con ntree=25, interaction depth=2 y lambda=0.25. Por lo que se vuelve a calcular el boosted tree con estos valores óptimos.

```
Boost <- gbm(PURCHASE_FLG ~. , data =data[train,], distribution =
"gaussian",
             n.trees = 25, interaction.depth = 2, shrinkage = 0.25)
yhatBoost <- predict(Boost, newdata=data[-train,], n.trees=25)
mse <- mean((yhatBoost-test)^2)
```

El error con este modelo y estos parámetros es del 3.255993%

Conclusión

La gente es diferente en lo que se refiere a su voluntad de comprar más o menos de un bien a diversos precios.

Los distintos consumidores tienen distintas curvas de demanda o planes para diversos productos y servicios. Mientras que los propietarios de empresas inicialmente no conocen esta información acerca de las costumbres de compra de sus potenciales clientes.

Es por esto, que la exploración de datos ofrece un mejor acercamiento para las empresas que ofertan cupones con la finalidad de que predecir quiénes, dónde y cómo se hace la compra de cupones. Una vez teniendo una idea de cómo se hace este proceso de compra; las compañías pueden mejorar sus procesos de oferta de cupones, ahorrar dinero y aumentar sus ganancias.

De manera cualitativa, lo que ayuda a predecir si se comprará un cupón son varias cosas:

- Un usuario realiza de 1 a 10 sesiones; de cada una de esas sesiones, ve de 1 a 5 cupones y realiza por lo general una compra.
- Si el descuento del cupón está entre el 40% y 50% es probable que se realice una compra.
- Los usuarios entre 35 y 40 años son los que más compran
- Entre más jóvenes, las mujeres compran más. Mientras va aumentando la edad, los hombres compran en mayor cantidad. Por lo que si el usuario tiene entre 20 y 40 años, es más probable que sea una mujer, por lo que se recomienda promover productos para el mercado femenino. De la misma forma, si el usuario es mayor a 40 años, es más probable que sea un hombre, por lo que se recomienda promover productos y servicios para el mercado masculino.
- Si el periodo de validación es de aproximadamente 6 meses, se compran más cupones.
- Los sábados y los días antes de vacaciones son los días en los que se compran más cupones.