PRAC2 – Limpieza y Validación de los Datos

Autor: Camilo Octavio Baez Ramos
Junio 2019

Descripción del dataset

El dataset utilizado es Black Friday el cual se encuentra en el repositorio de datasets del sitio Kaggle, contiene 12 columnas y 538K registros que hacen referencia a las compras de productos realizados por clientes durante el Black Friday. 11 de las 12 columnas del dataset son categóricas, solamente 1 columna (Purchase) es continua. El dataset permite aplicar actividades de pre-procesamiento y limpieza, por lo que permite aplicar los conocimientos adquiridos en la materia. El dataset permitirá hacer preguntas de negocio referentes a preferencias de productos con base a las características de los clientes, así como poder definir estrategias para captar ciertas poblaciones de clientes que no están estimuladas a comprar en estas fechas.

A	B	C	D	E	F	G	H	I	J	K	L
User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	10	A	2	0	12			1422
1000001	P00085442	F	0-17	10	A	2	0	12	14		1057
1000002	P00285442	M	55+	16	C	4+	0	8			7969
1000003	P00193542	M	26-35	15	A	3	0	1	2		15227
1000004	P00184942	M	46-50	7	B	2	1	1	8	17	19215
1000004	P00346142	M	46-50	7	B	2	1	1	15		15854
1000004	P0097242	M	46-50	7	B	2	1	1	16		15686
1000005	P00274942	M	26-35	20	A	1	1	8			7871
1000005	P00251242	M	26-35	20	A	1	1	5	11		5254
1000005	P00014542	M	26-35	20	A	1	1	8			3957
1000005	P00031342	M	26-35	20	A	1	1	8			6073
1000005	P00145042	M	26-35	20	A	1	1	1	2	5	15665
1000006	P00231342	F	51-55	9	A	1	0	5	8	14	5378
1000006	P00190242	F	51-55	9	A	1	0	4	5		2079

Limpieza y acondicionado de datos

Definición de librerías a utilizar dentro del script.

```
library(ggplot2)
library(dplyr)
library(corrgram)
library(FactoMineR)
library(factoextra)
library(gridExtra)
```

Cargue de la fuente de datos (archivo csv) y análisis primario de las variables.

```
data <- read.csv('01/BlackFriday.csv')
str(data)
summary(data)
head(data)
```

```
'data.frame':   537577 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1000004 1000005 ...
 $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",..: 671 2375 851 827 2733 1830 1744 3319 3597 2630 ...
 $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1 7 3 5 5 5 3 ...
 $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2 2 2 1 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3 5 4 3 3 3 2 ...
 $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
 $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
 $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
   User_ID           Product_ID     Gender        Age          Occupation     City_Category Stay_In_Current_City_Years
Min.   :1000001   P00265242:  1858   F:132197   0-17 : 14707   Min.   : 0.000   A:144638      0 : 72725
1st Qu.:1001495   P00110742:  1591   M:405380   18-25: 97634   1st Qu.: 2.000   B:226493      1 :189192
Median :1003031   P00025442:  1586              26-35:214690   Median : 7.000   C:166446      2 : 99459
Mean   :1002992   P00112142:  1539              36-45:107499   Mean   : 8.083                 3 : 93312
3rd Qu.:1004417   P00057642:  1430              46-50: 44526   3rd Qu.:14.000                 4+: 82889
Max.   :1006040   P00184942:  1424              51-55: 37618   Max.   :20.000
                  (Other)   :528149             55+  : 20903
Marital_Status   Product_Category_1 Product_Category_2 Product_Category_3    Purchase
Min.   :0.0000   Min.   : 1.000     Min.   : 2.00      Min.   : 3.0       Min.   :  185
1st Qu.:0.0000   1st Qu.: 1.000     1st Qu.: 5.00      1st Qu.: 9.0       1st Qu.: 5866
Median :0.0000   Median : 5.000     Median : 9.00      Median :14.0       Median : 8062
Mean   :0.4088   Mean   : 5.296     Mean   : 9.84      Mean   :12.7       Mean   : 9334
3rd Qu.:1.0000   3rd Qu.: 8.000     3rd Qu.:15.00      3rd Qu.:16.0       3rd Qu.:12073
Max.   :1.0000   Max.   :18.000     Max.   :18.00      Max.   :18.0       Max.   :23961
                                    NA's   :166986     NA's   :373299
```

Se identifican las siguientes variables:

- User_ID: Identificador del comprador.
- Product_ID: Identificador del producto
- Gender: Sexo del comprador.
- Age: Rango de edad del comprador.
- Occupation: Ocupación del comprador.
- City_Category: Ciudad del comprador.
- Stay_In_Current_City_Years: Número de años de residencia en la ciudad del comprador.
- Marital_Status: Estado civil.
- Product_Category_1: Categoría del producto comprado.
- Product_Category_2: Otra categoría del producto comprado.
- Product_Category_3: Otra categoría del producto comprado.
- Purchase: Valor de la compra.

A continuación, se verifican los posibles datos nulos del dataset:

```
colSums(is.na(data))
```

```
          User_ID        Product_ID            Gender               Age
                0                 0                 0                 0
       Occupation     City_Category Stay_In_Current_City_Years    Marital_Status
                0                 0                 0                 0
Product_Category_1 Product_Category_2 Product_Category_3          Purchase
                0            166986            373299                 0
```

En las columnas Product_Category_2 y Product_Category_3 se encuentran una gran cantidad de valores nulos, esto se debe a que los productos no tienen asignadas otras categorías.

Los valores nulos de las columnas mencionadas serán reemplazados por ceros:

```
data[is.na(data)]<-0
colSums(is.na(data))
```

```
          User_ID         Product_ID             Gender                Age 
                0                  0                  0                  0 
       Occupation      City_Category Stay_In_Current_City_Years     Marital_Status 
                0                  0                  0                  0 
Product_Category_1 Product_Category_2 Product_Category_3           Purchase 
                0                  0                  0                  0 
```

Y datos vacíos, no se encuentran datos vacíos para las columnas.

```
colSums(data=="")
```

```
          User_ID         Product_ID             Gender                Age 
                0                  0                  0                  0 
       Occupation      City_Category Stay_In_Current_City_Years     Marital_Status 
                0                  0                  0                  0 
Product_Category_1 Product_Category_2 Product_Category_3           Purchase 
                0                  0                  0                  0 
```

Ahora, realizamos una vista general de las variables encontrando que en su gran mayoría todas son categóricas excepto la variable Purchase

```
# Â¿Para quÃ© variables tendrÃa sentido un proceso de discretizaciÃ³n?
apply(data,2, function(x) length(unique(x)))
```

```
          User_ID         Product_ID             Gender                Age 
             5891               3623                  2                  7 
       Occupation      City_Category Stay_In_Current_City_Years     Marital_Status 
               21                  3                  5                  2 
Product_Category_1 Product_Category_2 Product_Category_3           Purchase 
               18                 18                 16              17959 
```

A continuación, se procede a discretizar las variables con pocas clases convirtiéndolas en factores:

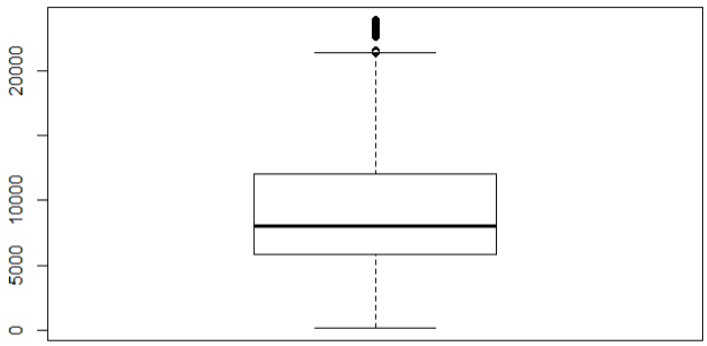
```
# Discretizamos las variables con pocas clases
cols<-c("Gender","Age","Occupation","City_Category","Stay_In_Current_City_Years","Ma
rital_Status","Product_Category_1","Product_Category_2","Product_Category_3")
for (i in cols){
  data[,i] <- as.factor(data[,i])
}
summary(data)
```

```
   User_ID           Product_ID     Gender        Age           Occupation     City_Category Stay_In_Current_City_Years
Min.   :1000001   P00265242:  1858   F:132197   0-17 : 14707   4      : 70862   A:144638      0 : 72725              
1st Qu.:1001495   P00110742:  1591   M:405380   18-25: 97634   0      : 68120   B:226493      1 :189192              
Median :1003031   P00025442:  1586              26-35:214690   7      : 57806   C:166446      2 : 99459              
Mean   :1002992   P00112142:  1539              36-45:107499   1      : 45971                 3 : 93312              
3rd Qu.:1004417   P00057642:  1430              46-50: 44526   17     : 39090                 4+: 82889              
Max.   :1006040   P00184942:  1424              51-55: 37618   20     : 32910                                        
                  (Other)   :528149              55+  : 20903   (Other):222818                                        
Marital_Status Product_Category_1 Product_Category_2 Product_Category_3    Purchase     
0:317817       5      :148592     0      :166986     0      :373299     Min.   :  185  
1:219760       1      :138353     8      : 63058     16     : 32148     1st Qu.: 5866  
               8      :112232     14     : 54158     15     : 27611     Median : 8062  
               11     : 23960     2      : 48481     14     : 18121     Mean   : 9334  
               2      : 23499     16     : 42602     17     : 16449     3rd Qu.:12073  
               6      : 20164     15     : 37317     5      : 16380     Max.   :23961  
               (Other): 70877     (Other):124975     (Other): 53569                    
```

Finalmente, se identifican los outliers de la única variable continua del conjunto de datos “purchase”:

```
boxplot.stats(data$Purchase)$out
boxplot(data$Purchase)
```

```
 [766] 23129 23209 23425 21500 21421 23300 23169 23072 23187 21435 23594 23258 23314
21419 23083 23479 23419
 [783] 23215 23475 23663 23611 21423 23462 23633 23196 23359 23313 23585 23547 23594
23723 23637 21405 23395
 [800] 23353 23280 23798 23333 23389 23472 23488 23792 23349 21555 21462 21512 21568
23610 23634 23322 23848
 [817] 23861 23913 23691 23254 23739 23301 23314 23445 21451 23145 23047 23638 23443
23076 23678 23631 23361
 [834] 21506 23715 23412 23906 23807 21518 23741 23389 23124 23261 23883 23614 23314
23396 23080 23323 23487
 [851] 23735 23600 23519 23763 23241 23417 23270 23936 23949 23659 23143 23704 23180
23784 23759 23698 23087
 [868] 23914 23181 23523 23915 23101 23836 23193 21436 21418 23738 23585 23167 23562
23372 23861 21462 23192
 [885] 21522 23363 23671 23496 23913 23125 23699 23091 23087 23546 23676 23837 23105
23760 23222 23423 23528
 [902] 23146 23353 23081 21475 23665 23043 21416 23580 23703 23944 23543 23128 21563
21408 23431 23041 23246
 [919] 23518 22710 21428 23306 23378 23475 23575 23125 23323 23714 23281 23215 21481
23092 23847 23155 21491
 [936] 23620 23320 23631 23643 23783 23409 23766 23292 23531 23933 21450 21468 21562
23853 23371 23726 21401
 [953] 23677 23328 23433 21477 23523 23725 21391 21487 23405 23318 23360 23685 23486
21564 23052 23435 21418
 [970] 23341 21423 21442 23835 23046 23425 23714 23611 23847 23258 23455 23178 23774
23564 23649 23435 23057
 [987] 23550 23255 23830 23387 23739 23307 23830 23463 23838 23729 23572 22651 23442
23877
 [ reached getOption("max.print") -- omitted 1665 entries ]
```



De lo visualizado se puede concluir que los valores considerados por el algoritmo como outliers en realidad corresponden a valores de artículos, dado que existen más de 2500 valores con valores con valores por encima del rango de los 21.000 lo cual no aparenta ser un outlier respecto a la población de valores de productos. El manejo que se dará a los outliers es mantenerlos como están dentro del dataset.

Finalmente, se exportan el dataset limpio:

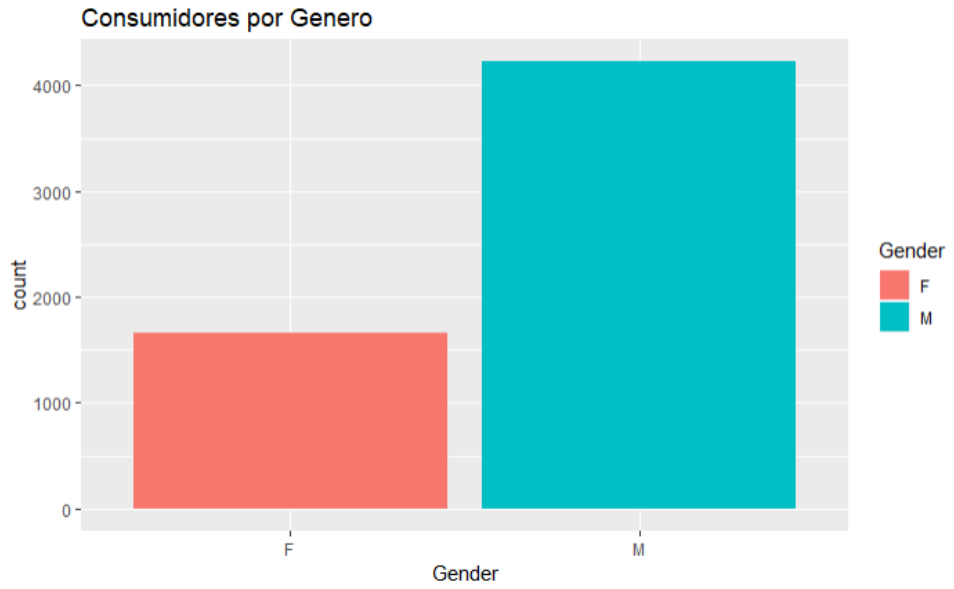
```
write.csv(data, "BlackFriday_clean.csv")
```

Análisis de los datos

Ahora analizaremos como están conformadas las diferentes dimensiones del dataset, empezaremos graficando la cantidad de personas por género que realizaron compras

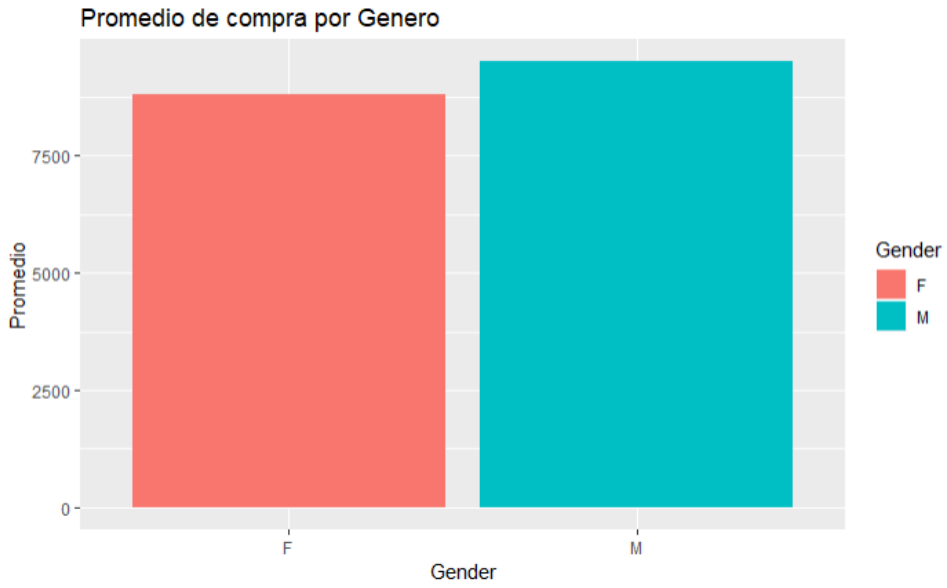
en el almacén. En la gráfica se evidencia que existe una mayoría notable entre la distribución de compradores de sexo masculino a las de sexo femenino.

```
data_genero = data %>%
                  dplyr::select(User_ID, Gender) %>%
                  group_by(User_ID) %>%
                  distinct()

ggplot(data = data_genero) +
  geom_bar(mapping = aes(x = Gender, y = ..count.., fill = Gender)) +
  labs(title = 'Consumidores por Genero')
```



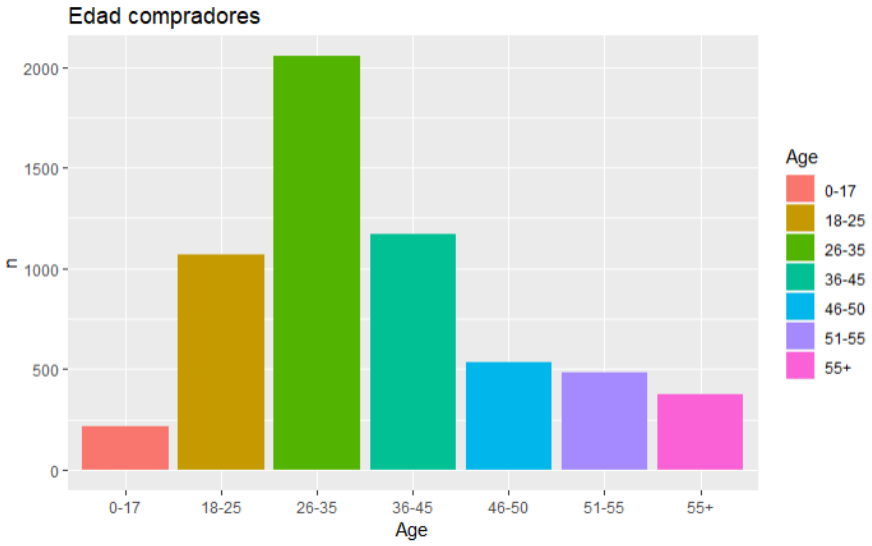
Dada la diferencia de compradores, sería interesante verificar la medida de tendencia central en relación al promedio de compra por sexo. **Como se puede ver en la gráfica, la diferencia entre promedios de compra por género no es tan alta como la diferencia entre las cantidades de compradores por sexo, lo que indica que el evento llama más la atención a hombres que a mujeres, esta información vendría bien para definir una buena estrategia de marketing orientada al público femenino.**

```
compras_x_usuario = data %>%
                      dplyr::select(User_ID, Gender, Purchase) %>%
                      group_by(User_ID,Gender) %>%
                      summarise(Total_compra = sum(Purchase),
                        Cantidad=n())

promedio_x_sexo = compras_x_usuario %>%
                    group_by(Gender) %>%
                    summarise(Promedio=sum(as.numeric(Total_compra))/sum(as.numeric(
Cantidad)))

ggplot(data = promedio_x_sexo) +
  geom_bar(mapping = aes(x = Gender, y = Promedio, fill = Gender), stat = 'identity'
) +
  labs(title = 'Promedio de compra por Genero')
```



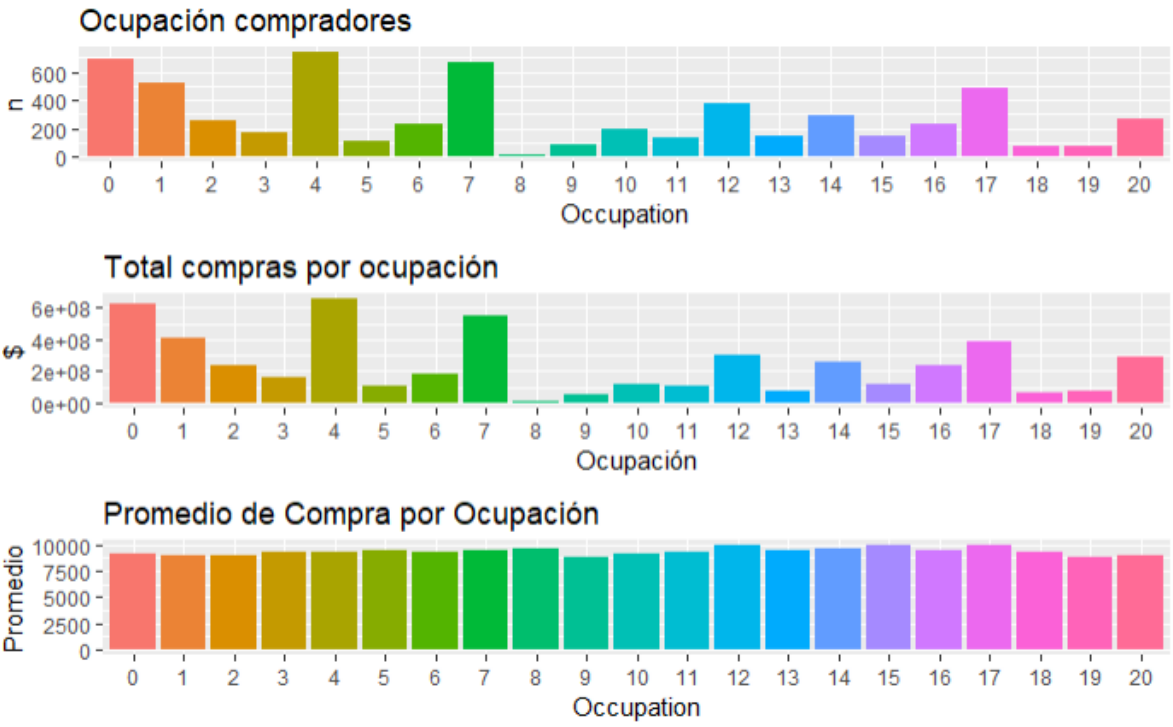
Continuamos el análisis por la columna de edad, vamos a realizar conteo por cada categoría de edad:

```
edad_compradores = data %>%
                   dplyr::select(User_ID, Age) %>%
                   distinct() %>%
                   count(Age)

ggplot(data = edad_compradores) +
  geom_bar(stat = 'identity', mapping = aes(x = Age, y = n, fill = Age)) +
  labs(title = 'Edad compradores')
```



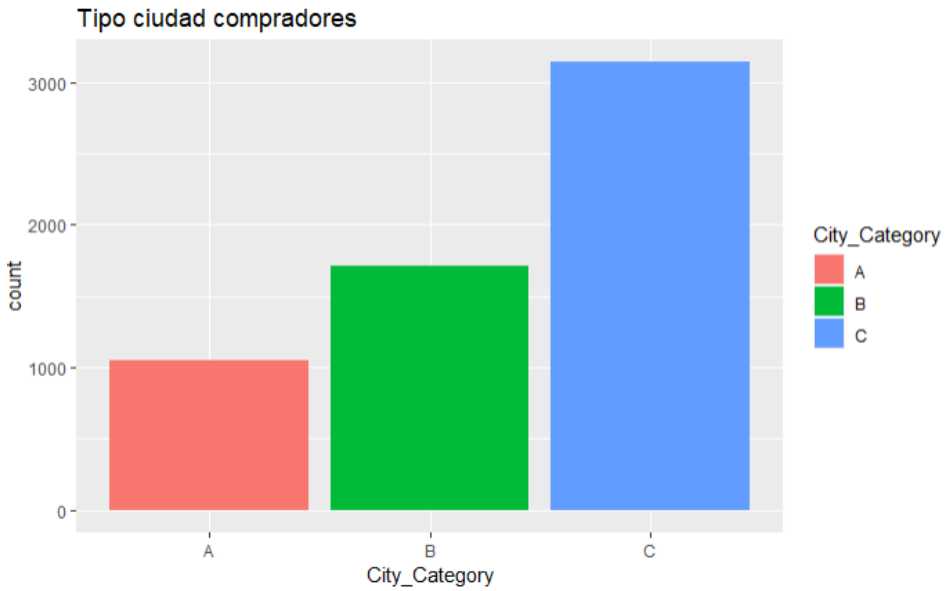
Se observa que la mayor población de compradores se encuentra ubicada entre los 18 y los 45 años de edad. Ahora analizamos la columna ocupaciones de los clientes, encontrando que en promedio, ningún cargo tiene a gastar más por compra.

```
ocupacion_compradores = data %>%
                    dplyr::select(User_ID, Occupation) %>%
                    distinct() %>%
                    count(Occupation)

g1<-ggplot(data = ocupacion_compradores) +
  geom_bar(stat = 'identity', mapping = aes(x = Occupation, y = n, fill = Occupation)) +
  labs(title = 'OcupaciÃ³n compradores') + theme(legend.position = "none")

compras_x_ocupacion = data %>%
                  group_by(Occupation) %>%
                  summarise(Compras = sum(Purchase))

g2<-ggplot(data = compras_x_ocupacion, aes(x = Occupation, y = Compras, fill = Occupation)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Total compras por ocupaciÃ³n', y = '$', x = 'OcupaciÃ³n') + theme(legend.position = "none")

compras_x_ocupacion = data %>%
                      dplyr::select(User_ID, Occupation, Purchase) %>%
                      group_by(User_ID,Occupation) %>%
                      summarise(Total_compra = sum(Purchase),
                        Cantidad=n())

promedio_x_ocupacion = compras_x_ocupacion %>%
                    group_by(Occupation) %>%
                    summarise(Promedio=sum(as.numeric(Total_compra))/sum(as.numeric(Cantidad)))

g3<-ggplot(data = promedio_x_ocupacion) +
  geom_bar(mapping = aes(x = Occupation, y = Promedio, fill = Occupation), stat = 'identity') +
  labs(title = 'Promedio de Compra por OcupaciÃ³n') + theme(legend.position = "none")

grid.arrange(g1,g2,g3,ncol=1)
```



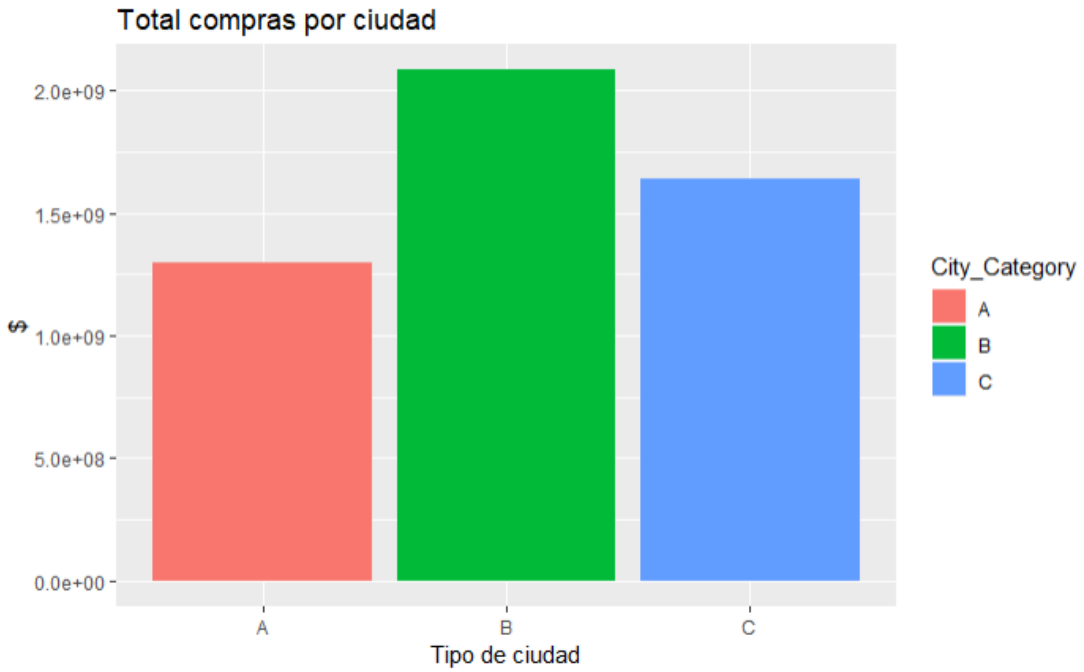
Ahora observemos de donde vienen los compradores, vemos que principalmente vienen de las ciudades de categoría C.

```
ciudad_compradores =  data %>%
                        dplyr::select(User_ID, City_Category) %>%
                        distinct()

ggplot(data = ciudad_compradores) +
  geom_bar(mapping = aes(x = City_Category, y = ..count.., fill = City_Category)) +
  labs(title = 'Tipo ciudad compradores')
```



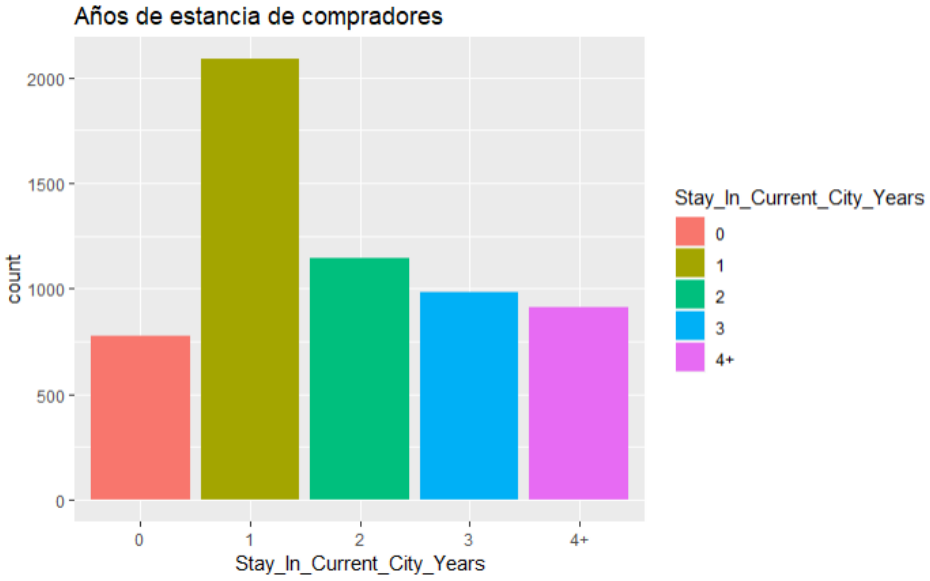
Ahora veamos las compras por ciudad, sorprendentemente encontramos que los mayores compradores (en cantidad de dinero) no son de la ciudad con mayor volumen de compradores (C).

```
compras_x_ciudad = data %>%
                   group_by(City_Category) %>%
                   summarise(Compras = sum(Purchase))

ggplot(data = compras_x_ciudad, aes(x = City_Category, y = Compras, fill = City_Categor
y)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Total compras por ciudad', y = '$', x = 'Tipo de ciudad')
```



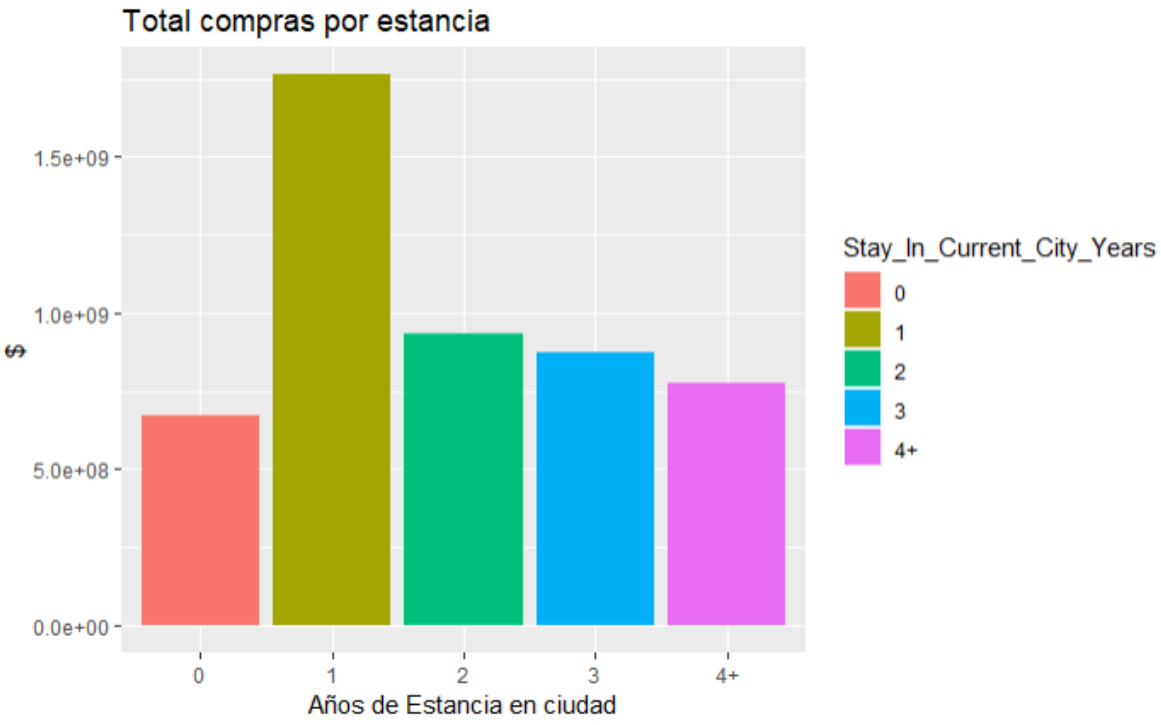
Ahora examinaremos los compradores por tiempo de residencia:

```
estancia_compradores =  data %>%
                        dplyr::select(User_ID, Stay_In_Current_City_Years) %>%
                        distinct()

ggplot(data = estancia_compradores) +
  geom_bar(mapping = aes(x = Stay_In_Current_City_Years, y = ..count.., fill = Stay_In_
Current_City_Years)) +
  labs(title = 'AÃ±os de estancia de compradores')
```



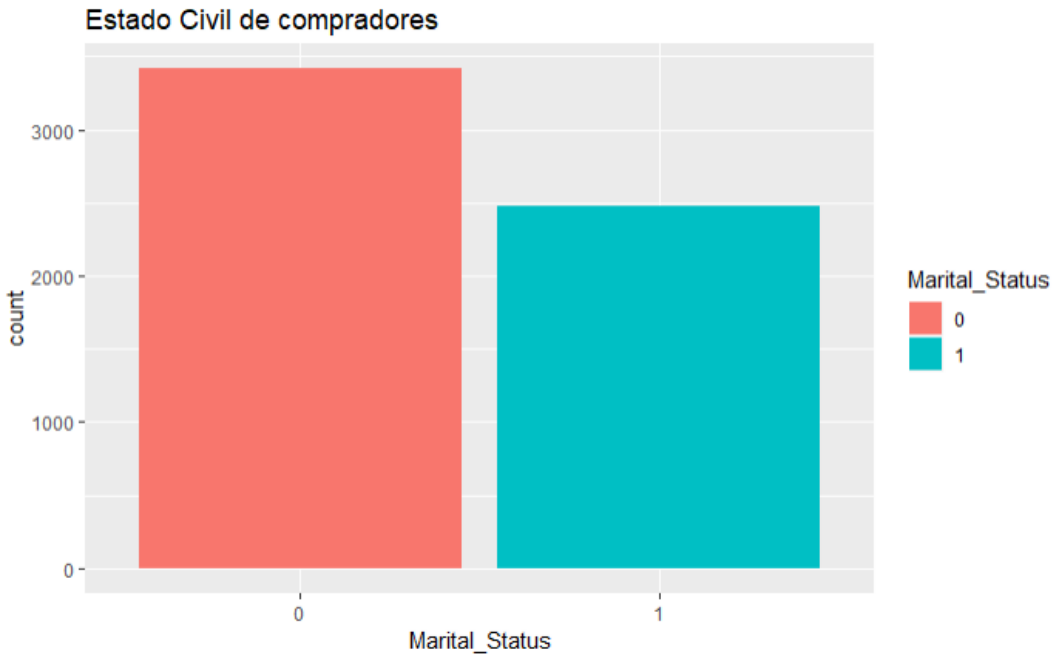
Ahora veamos las compras por años de estancia.

```
compras_x_estancia = data %>%
                  group_by(Stay_In_Current_City_Years) %>%
                  summarise(Compras = sum(Purchase))

ggplot(data = compras_x_estancia, aes(x = Stay_In_Current_City_Years, y = Compras, fill
 = Stay_In_Current_City_Years)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Total compras por estancia', y = '$', x = 'AÃ±os de Estancia en ciudad'
)
```



Ahora examinaremos los compradores por estado civil:

```
est_civil_compradores =  data %>%
                       dplyr::select(User_ID, Marital_Status) %>%
                       distinct()

ggplot(data = est_civil_compradores) +
  geom_bar(mapping = aes(x = Marital_Status, y = ..count.., fill = Marital_Status)) +
  labs(title = 'Estado Civil de compradores')
```



Ahora veamos las compras por estado civil:

```
compras_x_est_civ = data %>%
                    group_by(Marital_Status) %>%
                    summarise(Compras = sum(as.numeric(Purchase)))

ggplot(data = compras_x_est_civ, aes(x = Marital_Status, y = Compras, fill = Marital_St
atus)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Total compras por estado civil', y = '$', x = 'Estado Civil')
```

Selección de grupos de datos a analizar

Se seleccionan los grupos de variables que pueden ser interesantes dentro del análisis:

```
data.hombre <- data[data$Gender == "M",]
data.mujer <- data[data$Gender == "F",]
data.marital_0 <- data[data$Marital_Status == "0",]
data.marital_1 <- data[data$Marital_Status == "1",]
data.age_17 <- data[data$Age == "0-17",]
data.age_25 <- data[data$Age == "18-25",]
data.age_35 <- data[data$Age == "26-35",]
data.age_45 <- data[data$Age == "36-45",]
data.age_50 <- data[data$Age == "46-50",]
data.age_55 <- data[data$Age == "51-55",]
data.age_55plus <- data[data$Age == "55+",]
```

Comprobación de normalidad de variable continua

Se procede a verificar la normalidad de la variable continua Purchase:

```
library(nortest)
alpha = 0.05
p_val = ad.test(data$Purchase)$p.value
if (p_val < alpha) {
  cat("La variable Purchase no sigue una distribución normal. \n")
}else{
  cat("La variable Purchase sigue una distribución normal. \n")
}
cat("P-value: ")
cat(p_val)
```

```
La variable Purchase no sigue una distribución normal.
P-value: 3.7e-24
```

Encontrando que dicha variable no está normalizada, esto dado a que el p-value de la prueba es inferior a 0,05.

Evaluación de la varianza de los grupos

A continuación, se realizará test de Fligner-Killeen para verificar la homogeneidad de la varianza entre algunos grupos de compradores y sus compras.

```r
fligner.test(Purchase ~ Marital_Status, data = data)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Marital_Status
Fligner-Killeen:med chi-squared = 6.6163, df = 1, p-value = 0.01011
```

```r
fligner.test(Purchase ~ Age, data = data)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Age
Fligner-Killeen:med chi-squared = 78.885, df = 6, p-value = 6.072e-15
```

```r
fligner.test(Purchase ~ Gender, data = data)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  Purchase by Gender
Fligner-Killeen:med chi-squared = 1608.5, df = 1, p-value < 2.2e-16
```

Como se visualiza, todos los p-value obtenidos son <0,05 por lo que las varianzas de los grupos no son consideradas homogéneas.

Evaluación de hipótesis a través de prueba estadística

Se realiza prueba de contraste de hipótesis sobre las muestras identificadas para hombres y mujeres, la hipótesis a evaluar es si los hombres compran más que las mujeres.

```r
data.hombre.compras <- data.hombre$Purchase
data.mujer.compras <- data.mujer$Purchase
t.test(data.hombre.compras, data.mujer.compras,alternative = "less")
```

```
        Welch Two Sample t-test

data:  data.hombre.compras and data.mujer.compras
t = 45.673, df = 238460, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 720.0405
sample estimates:
mean of x mean of y
 9504.772  8809.761
```

Con un p-value=1, se valida estadísticamente la hipótesis planteada.

Conclusiones

- Sobre los datos se aplicó labor de preprocesamiento para manejar los casos de ceros o elementos nulos y valores extremos (outliers). Para el caso de ceros no

se encontraron dentro del dataset, para el caso de nulos se realizó imputación por valor 0 a fin de no eliminar los registros donde se presentaran estos casos. Para el caso de los outliers, estos se dejaron dentro del dataset al no tratarse de valores del todo atípicos, el gran número de casos y por corresponder a posibles valores reales de compras.

- De acuerdo al análisis descriptivo realizado se puede identificar que la diferencia entre promedios de compra por género no es tan alta como la diferencia entre las cantidades de compradores por sexo, lo que indica que el evento llama más la atención a hombres que a mujeres, esta información vendría bien para definir una buena estrategia de marketing orientada al público femenino.

- La prueba estadística permitió validar que la población masculina tiene un promedio de compras superior a la femenina. Hecho que también pudo verificarse en el análisis descriptivo.

- Se encuentra que los mayores compradores (en cantidad de dinero) no son de la ciudad con mayor volumen de compradores (C).

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	**COBR**
Redacción de las respuestas	**COBR**
Desarrollo código	**COBR**