Análisis de ventas a través de los comportamientos del consumidor

Motivación

El conjunto de datos es una muestra de transacciones realizadas en una tienda. Nos gustaría conocer mejor el comportamiento de compra del cliente frente a diferentes productos. Un primer problema es, entonces, predecir el monto de la compra de un usuario con la ayuda de la información contenida en las otras variables.

El problema de clasificación también se puede resolver en este conjunto de datos, ya que varias variables son categóricas y algunos otros enfoques podrían ser por ejemplo “Predecir la edad del consumidor” o incluso “Predecir la categoría de los bienes comprados”.

Este conjunto de datos también es particularmente conveniente para agrupar y quizás encontrar diferentes grupos de consumidores dentro de él como así también ser capaces de generar recomendaciones de productos.

Breve exploración del dataset

El dataset *retail_sales.zip* se encuentra en el siguiente classroom: https://classroom.google.com/u/0/c/MzE2MDE2MTY4NzBa (https://classroom.google.com/u/0/c/MzE2MDE2MTY4NzBa) descargarlo y descomprimirlo.

Cargamos el dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                     537577 non-null int64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 non-null object
Occupation                  537577 non-null int64
City_Category               537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status              537577 non-null int64
Product_Category_1          537577 non-null int64
Product_Category_2          370591 non-null float64
Product_Category_3          164278 non-null float64
Purchase                    537577 non-null float64
dtypes: float64(3), int64(4), object(5)
memory usage: 49.2+ MB
```

Out[4]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_
403939	1002095	P00295842	M	55+	13	C	
91200	1002010	P00163342	M	18-25	4	B	
470713	1000533	P00118542	M	26-35	12	A	
437164	1001290	P00021742	M	36-45	0	C	
425316	1005493	P00358442	M	36-45	12	B	

Descripción de las columnas:

- $User\_ID$: identificador unívoco de cada usuario.
- $Product\_ID$: identificador unívoco de cada producto.
- $Gender$: género del usuario, F $\rightarrow$ Femenino, M $\rightarrow$ Masculino.
- Age: edad del usuario representada por rangos, es decir, no se conoce la edad exacta del usuario sino el rango de edad al cual pertenece.
- $Occupation$: ocupación del usuario, existen 21 ocupaciones distintas, cada una de ellas está identificada con un número del 0 al 20.
- $City\_Category$: categoría de ciudad en la que vive el usuario, existen 3 categorías: A - B - C.
- $Stay\_In\_Current\_City\_Years$: tiempo de permanecia del usuario en la ciudad actual expresada en años. Nota: el valor 0 significa que el tiempo de permanencia del usuario en esa ciudad ha sido menor a un año, no es un valor inválido.
- $Marital\_Status$: estado civil del usuario, valor 0 $\rightarrow$ Soltero, valor 1 $\rightarrow$ Casado
- $Product\_Category\_1$: categoria 1 del producto.
- $Product\_Category\_2$: categoria 2 del producto, probablemente sea una subcategoria de 1.
- $Product\_Category\_3$: categoria 3 del producto, probablemente sea una subcategoria de 2.
- $Purchase$: precio que el usuario pagó por el producto expresado en dólares.

En Primer lugar se extraen las columnas *Product_Category_2* y *Product_Category_3*

Out[6]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_
29632	1004497	P00240142	M	51-55	0	B	
525578	1003007	P00053742	M	36-45	0	C	
338822	1004116	P00235642	M	36-45	7	A	
346781	1005412	P00192542	M	26-35	12	A	
220076	1003931	P00133742	M	26-35	5	C	

Out[7]:

	User_ID	Occupation	Marital_Status	Product_Category_1	Purchase
count	5.375770e+05	537577.00000	537577.000000	537577.000000	537577.000000
mean	1.002992e+06	8.08271	0.408797	5.295546	93.338599
std	1.714393e+03	6.52412	0.491612	3.750701	49.810221
min	1.000001e+06	0.00000	0.000000	1.000000	1.850000
25%	1.001495e+06	2.00000	0.000000	1.000000	58.660000
50%	1.003031e+06	7.00000	0.000000	5.000000	80.620000
75%	1.004417e+06	14.00000	1.000000	8.000000	120.730000
max	1.006040e+06	20.00000	1.000000	18.000000	239.610000

1. Cantidad total de órdenes, usuarios y productos.

```
El número total de órdenes asciende a  537577
```

Out[9]:

	0
User_ID	5891
Product_ID	3623

2. Calcular estadísticos: media, mediana, moda, desviación estándar, valor mínimo, valor máximo de la cantidad de compras por usuario. Gráficar. A qué distribución conocida corresponde? Qué se puede concluir?

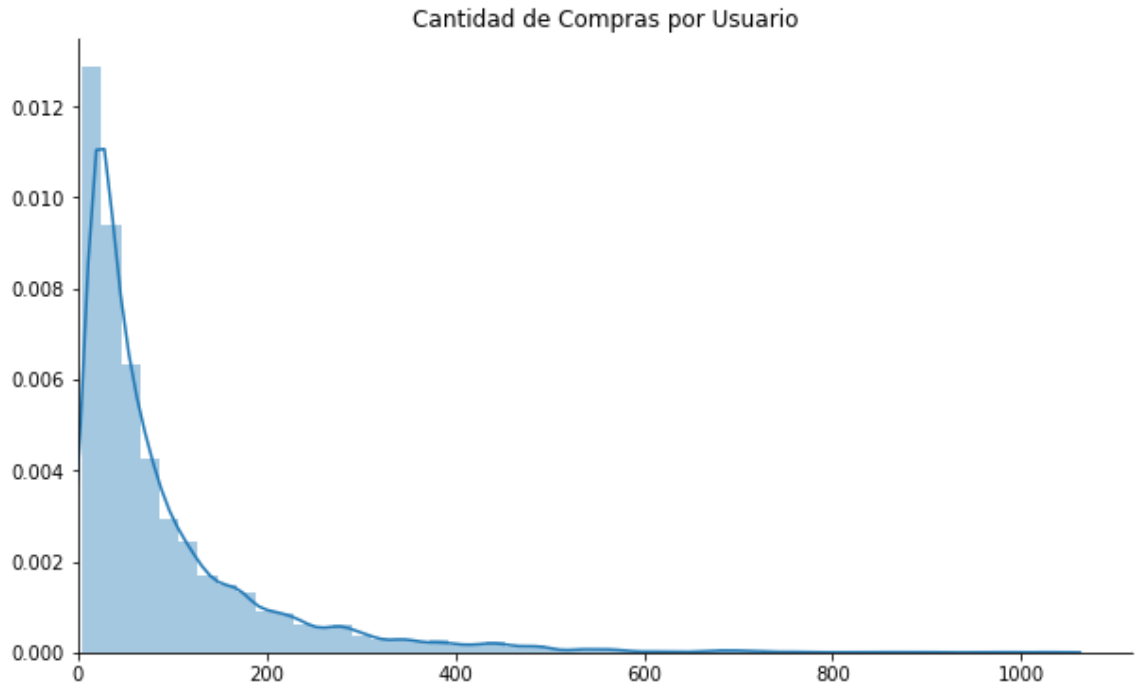
Out[10]:

	0
count	5891.000000
mean	91.253947
std	105.929800
min	5.000000
25%	25.000000
50%	53.000000
75%	114.000000
max	1025.000000

```
La mediana de compra por cada usuario se ubicó en  0    53.0
dtype: float64
 y la moda de las compras por usuario fue igual a      0
0  17
```

Out[13]:

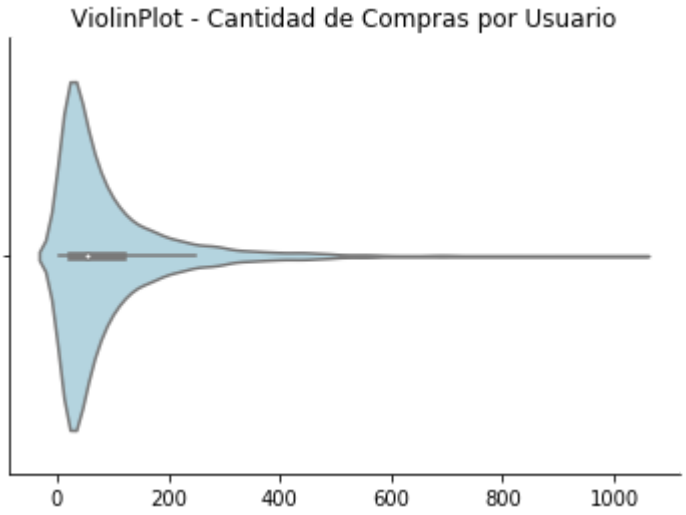
```
(0, 1116.6209571460904)
```



La cantidad de compras por usuario es marcadamente asimétrica derecha, lo que es señal de que esta variable sigue una distribución exponencial. Esto significa que el comportamiento más repetido (la moda) es una pequeña cantidad de compras (17), y que las compras de mayores cantidades son cada vez más extrañas entre los usuarios. Se puede observar además en base a la mediana, que la mitad de los usuarios compró como máximo 53 veces, aunque el promedio de compras sea de 91.25. Este último valor es afectado por la clara asimetría de la distribución, por lo que no es una buena medida representativa de los datos.

Out[14]:

```
Text(0.5, 1.0, 'ViolinPlot - Cantidad de Compras por Usuario')
```



El resultado obtenido en el ViolinPlot coincide con la gráfica observada previamente: valores concentrados en números altos de la serie, lo que responde a la asimetría de los datos. En estos casos, tanto el ViolinPlot como el BoxPlot no resultan de utilidad para la detección de **valores atípicos**

3. Calcular media, mediana, desviación estándar, valor mínimo, valor máximo de los valores de compras. Determinar valores atípicos (outliers) y graficar. El porcentaje de valores atípicos es muy alto? Analizar por género, edad, ocupación de los consumidores.

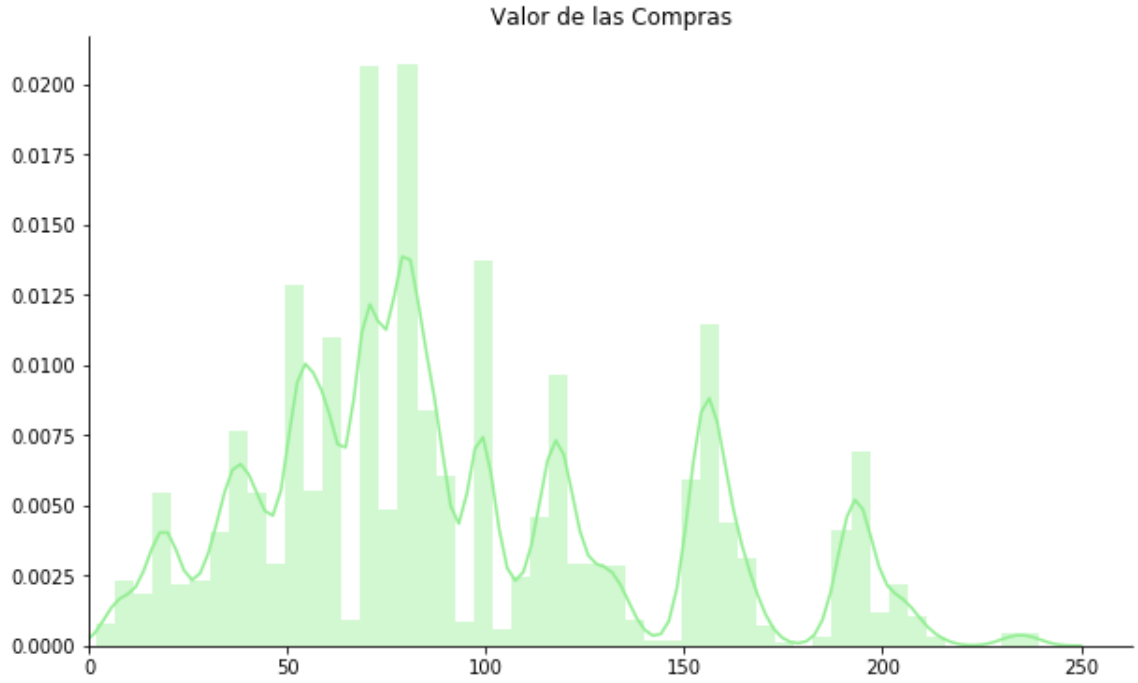
Out[15]:

	Purchase
count	537577.000000
mean	93.338599
std	49.810221
min	1.850000
25%	58.660000
50%	80.620000
75%	120.730000
max	239.610000

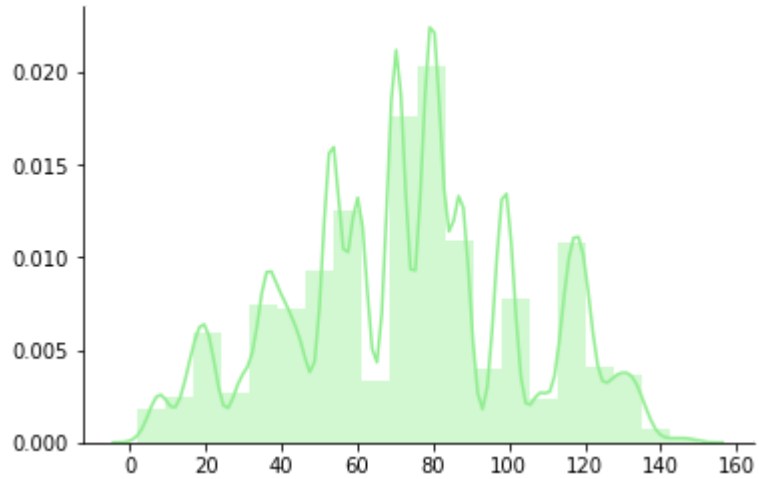
```
La mediana de cada compra se ubicó en  Purchase    80.62
dtype: float64
 y la moda de las compras fue igual a     Purchase
0     68.55
```

Out[17]:

```
(0, 262.98460164028415)
```

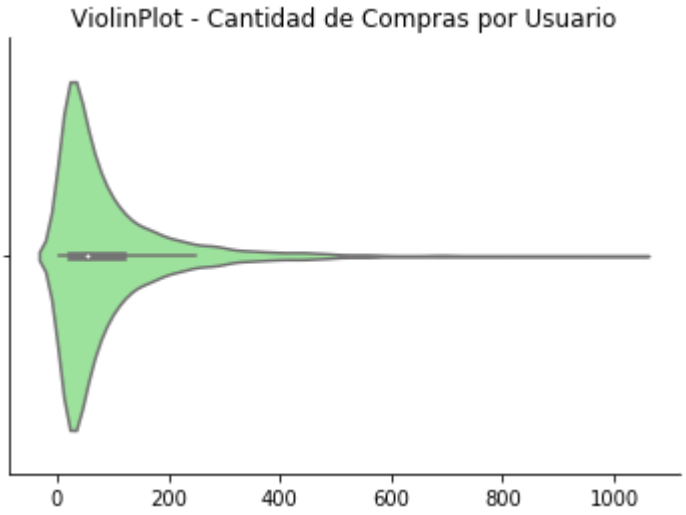


Como se ve en el gráfico, los datos no siguen una distribución normal en todo su rango. Probablemente entre el rango 0-150 podría asimilarse a una función normal, pero viendo la serie en su totalidad, esta presenta una asimetría derecha, dado que en el extremo derecho de la serie aún persiste una frecuencia elevada de montos comprados.



Out[19]:

```
Text(0.5, 1.0, 'ViolinPlot - Cantidad de Compras por Usuario')
```



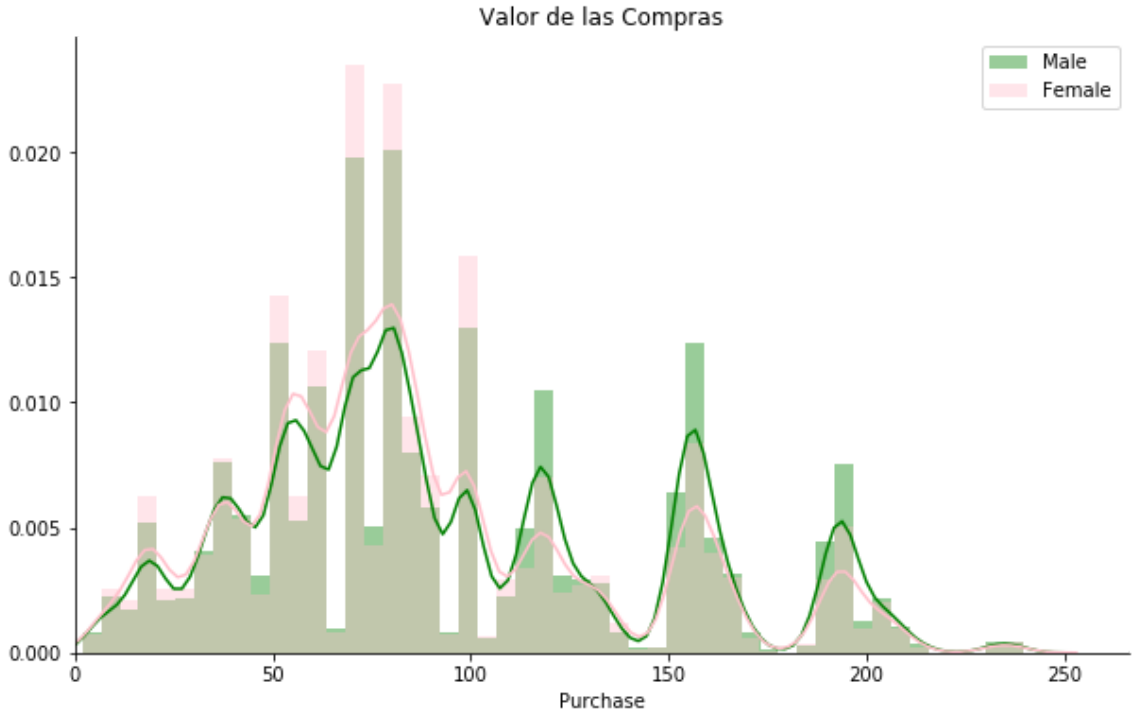
Identificación de los valores atípicos

```
El primer cuartil es igual a  58.66

El tercer cuartil es igual a  120.73

El rango intercuartil es igual a  62.07000000000001

El último cuartil es igual a  213.83500000000004

Cantidad de valores mayores al limite superior : 2665

Proporción de valores mayores al limite superior : 0.004957429354306453
```

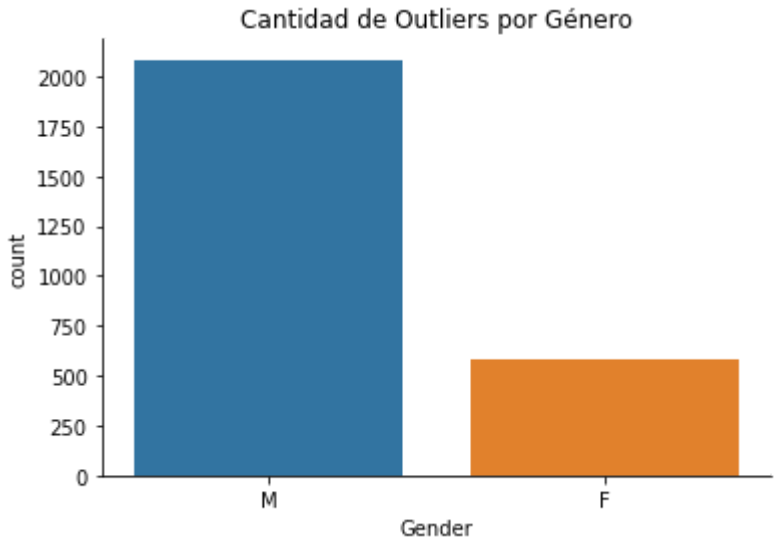
Outliers por Género

Out[27]:

```
<matplotlib.legend.Legend at 0x1600ad406a0>
```

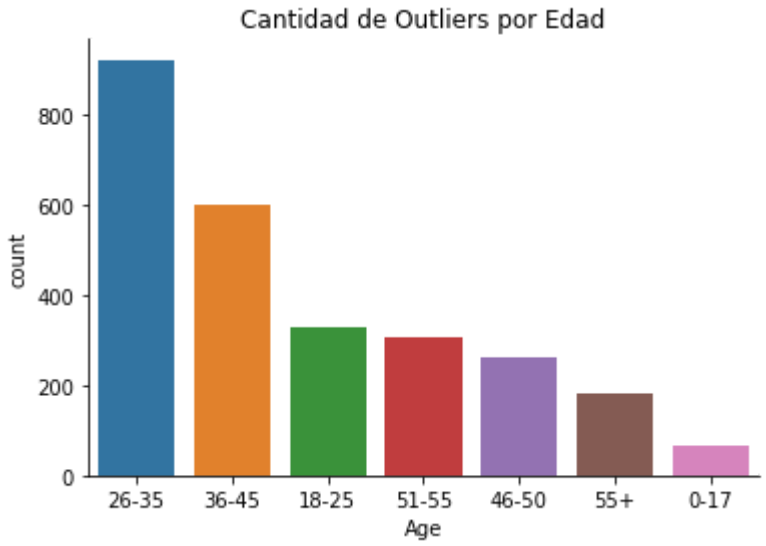


En el gráfico precedente se puede obervar como las mujeres presentan una mayor proporci+on de outliers en realción a los hombres en compras menos costosas, mientras que se presenta una mayor cantidad de outliers masculinos en compras de mayor valor; por encima de los U$S 150.



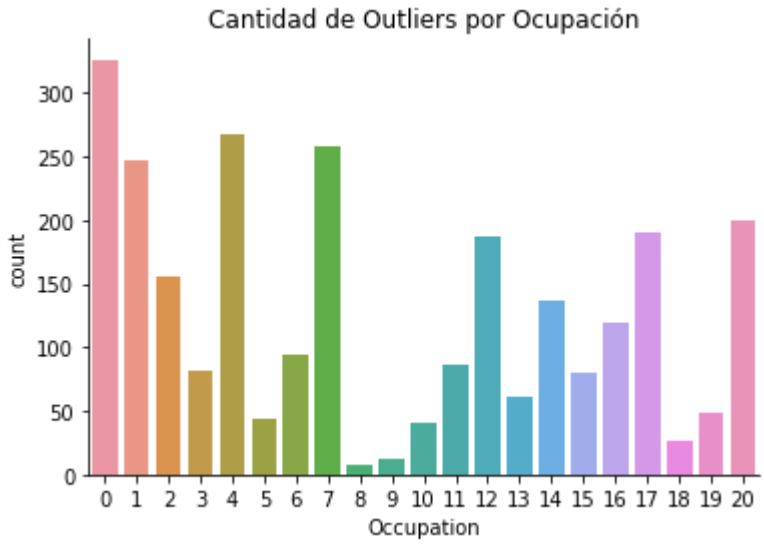
Los outliers masculinos superan los 2000 compradores mientras que los outliers de género femenino superan apenas los 500.

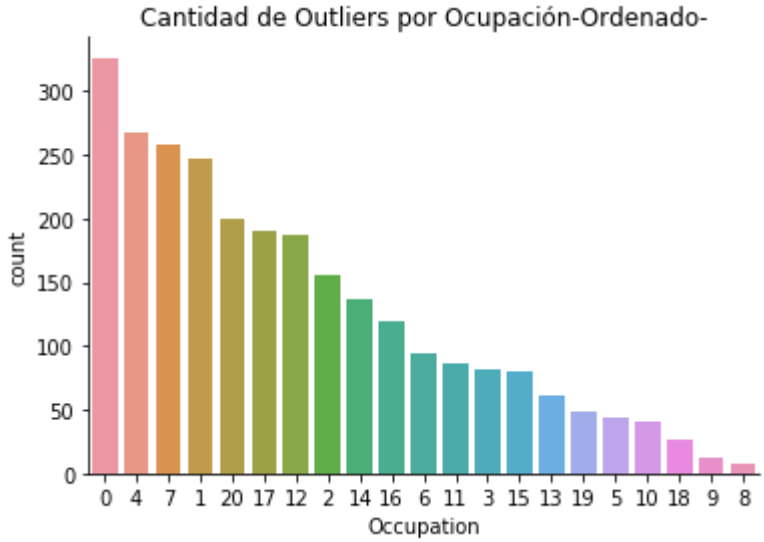
Outliers por Edad



La mayor cantidad de outliers se ubican en el rango de edad que va desde los 26 a los 35 años; en dicho rango los outliers superan los 800 compradores. Le siguen en orden de importancia los usuarios que tienen entre 36 y 45 años. Aquellos compradores entre 0 y 17 años solo presentan alrededor de 100 outliers.

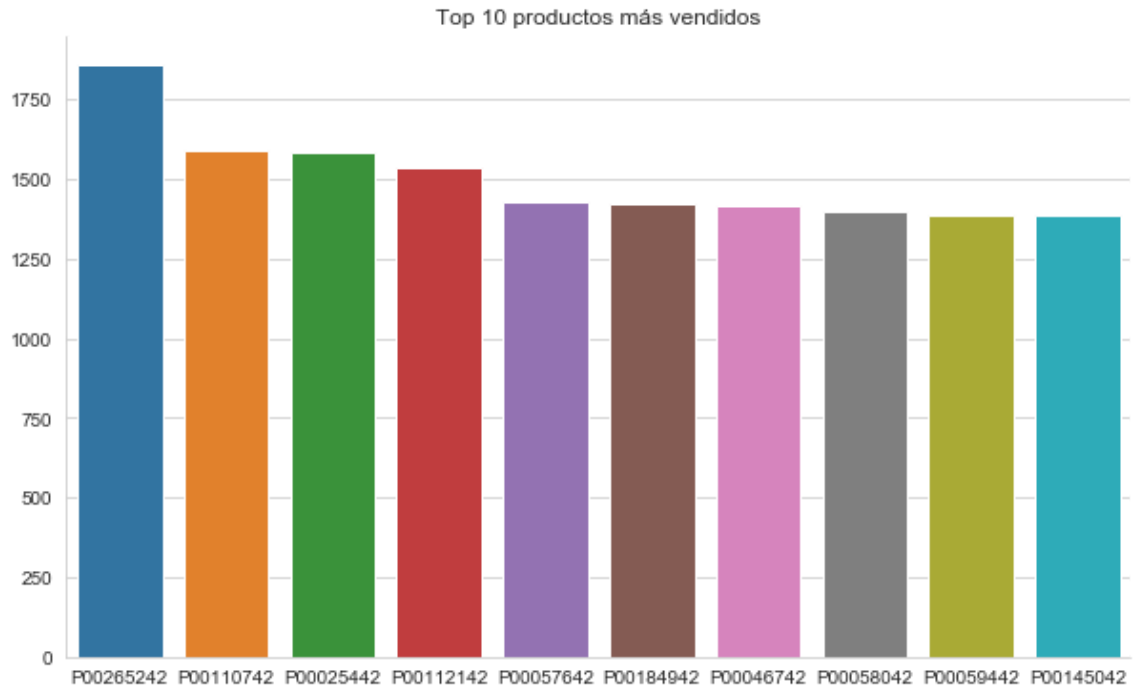
Outliers por Ocupación





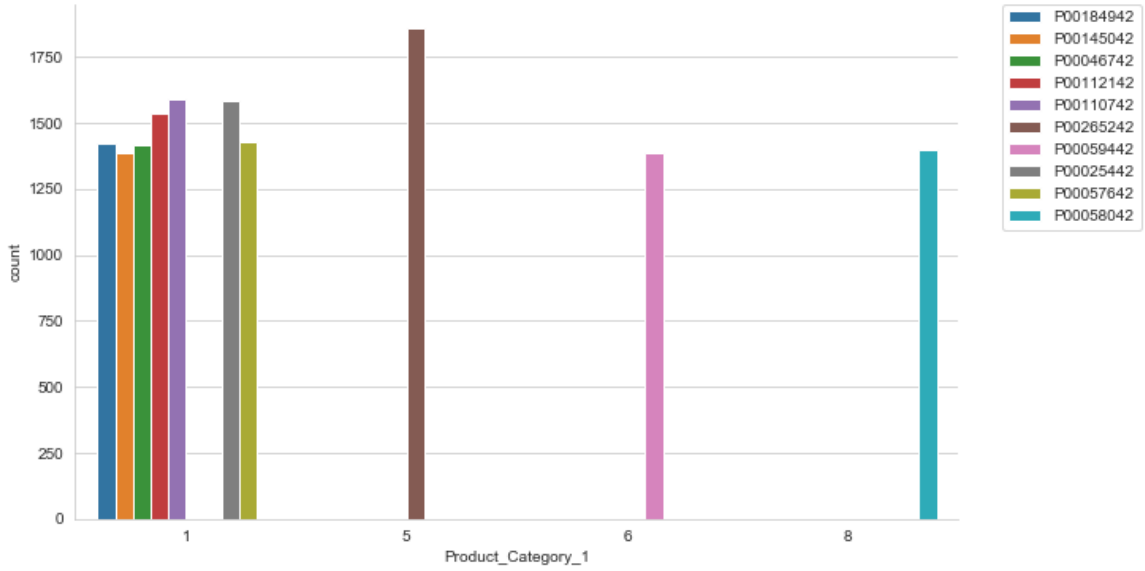
La profesión 0 es la que mas outliers registra, superando los 300 compradores. Le siguen las profesiones 4 y 7 que contienen 250 outliers, respectivamente, en promedio. La ocupación que menos compradores fuera de serie que presenta es la 8, con menos de 25 outliers.

4. Top productos más vendidos, a qué categorias pertenecen? (tener en cuenta solo la columna Product_Category_1). Estos productos son consumidos por usuarios de todas las edades o algunos rangos en particular? Graficar productos más vendidos por edad y por género

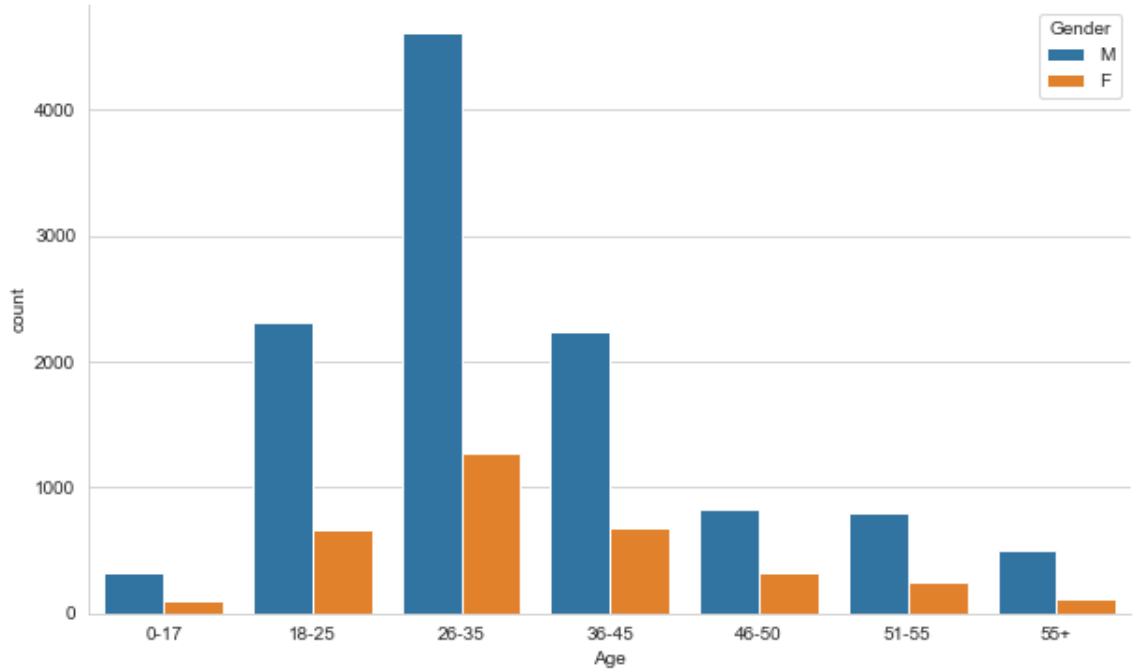


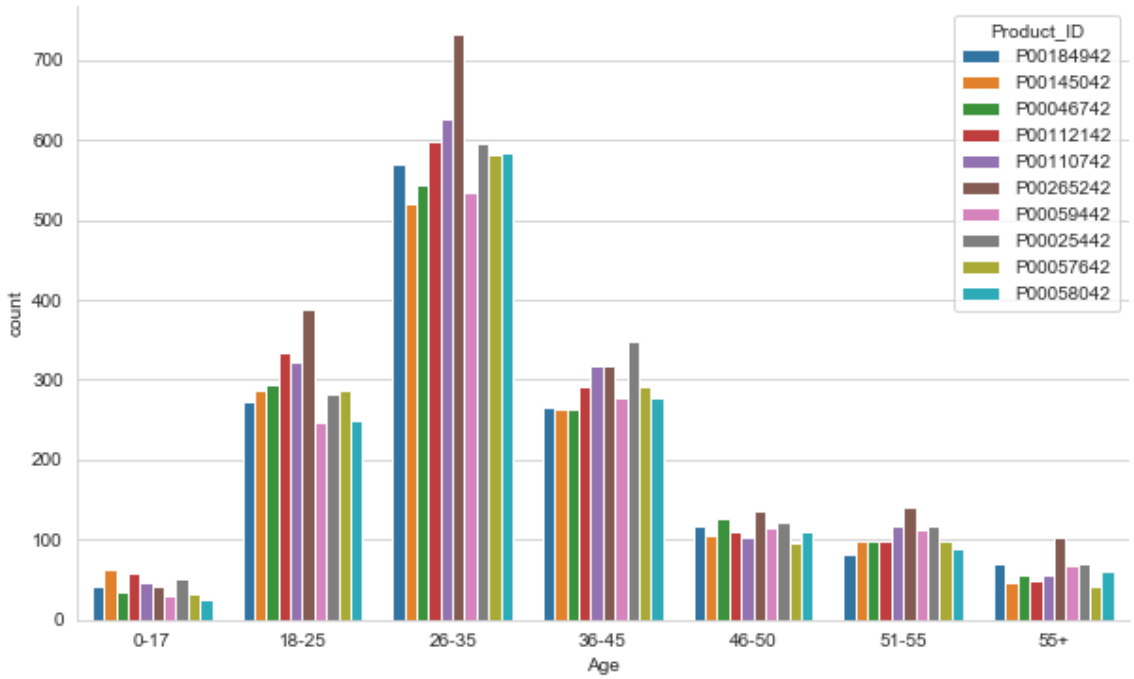
Out[33]:

Product_ID	0
P00265242	1858
P00110742	1591
P00025442	1586
P00112142	1539
P00057642	1430
P00184942	1424
P00046742	1417
P00058042	1396
P00145042	1384
P00059442	1384

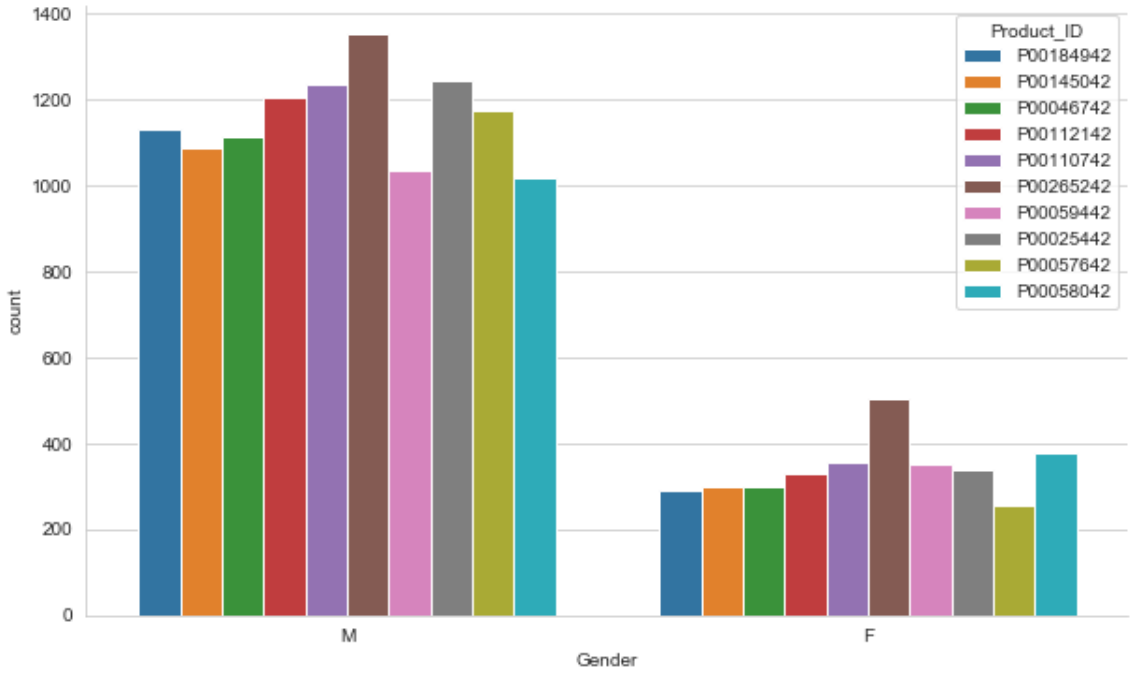


El 0.6909854087547471 de los productos más vendidos pertence a la categoría 1





El Producto PO0265242 es el más consumido entre todos los rangos etarios, a excepción de 2 rangos: 0-17 y 36-45. En el primer caso, el producto más consumido consiste en el producto PO0145042, que por lidera el ránking de productos consumidos entre las edades 0-17 años. En el segundo caso, el producto más adquirido entre las personas que tienen 36-45 años es el bien PO0025442.



Al analizar los productos más vendidos, distinguiendo el género, se observa que coincide tanto para hombre como para mujeres que el producto más adquirido es el PO0265242. Seguidamente, se observa que el producto que se posiciona en segundo lugar en el género femenino, es el menos adquirido por los masculinos; PO0058042.

Out[43]:

	Gender							Product_Category_1						
Age	**0-17**	**18-25**	**26-35**	**36-45**	**46-50**	**51-55**	**55+**	**0-17**	**18-25**	**26-35**	**36-45**	**46-50**	**51-55**	**55+**
Product_ID														
P00025442	52	282	595	348	122	117	70	52	282	595	348	122	117	70
P00046742	36	293	544	264	126	98	56	36	293	544	264	126	98	56
P00057642	33	288	581	291	96	99	42	33	288	581	291	96	99	42
P00058042	26	250	584	277	111	88	60	26	250	584	277	111	88	60
P00059442	30	247	535	277	115	112	68	30	247	535	277	115	112	68
P00110742	46	323	627	317	104	117	57	46	323	627	317	104	117	57
P00112142	58	335	597	292	111	98	48	58	335	597	292	111	98	48
P00145042	63	286	520	264	106	99	46	63	286	520	264	106	99	46
P00184942	43	274	570	266	118	83	70	43	274	570	266	118	83	70
P00265242	41	388	732	317	136	140	104	41	388	732	317	136	140	104

5. La categoría de productos menos vendidos corresponde a productos muy costosos? Cuáles categorías de productos son más consumidas por edad? y por género?

Los productos menos vendidos son los siguientes:

Out[45]:

	0
Product_ID	
P00056542	1
P00013442	1
P00013542	1
P00013842	1
P00206542	1

Out[46]:

```
Index(['P00056542', 'P00013442', 'P00013542', 'P00013842', 'P00206542',
       'P00062442', 'P00062342', 'P00275042', 'P00314742', 'P00142542',
       ...
       'P00204042', 'P00260742', 'P00308042', 'P00135942', 'P00306542',
       'P00073342', 'P00126742', 'P00306942', 'P00338242', 'P00203942'],
      dtype='object', name='Product_ID', length=141)
```

Out[49]:

	Product_ID	Age	Gender	Purchase	Product_Category_1
173592	P00074542	26-35	M	203.23	7
424219	P00341542	55+	M	202.91	6
137562	P00308042	51-55	F	192.06	10
514546	P00315142	26-35	M	184.68	10
511641	P00075042	36-45	M	184.56	9
97045	P00135942	26-35	M	169.54	7
518979	P00038842	51-55	F	168.95	7
402677	P00292142	36-45	F	166.09	7
528719	P00295642	26-35	M	159.66	2
303737	P00166442	36-45	M	159.66	6

Inicialmente consideramos los productos cuyo valor se ubica por encima del 75% del total de productos vendidos. Es decir, solo se consideran aquellos bienes que superan el precio que contiene al 75% del total de observaciones.

```
Del total de productos menos consumidos, solo  17 se encuentran fuera del
75% de las obervaciones
```

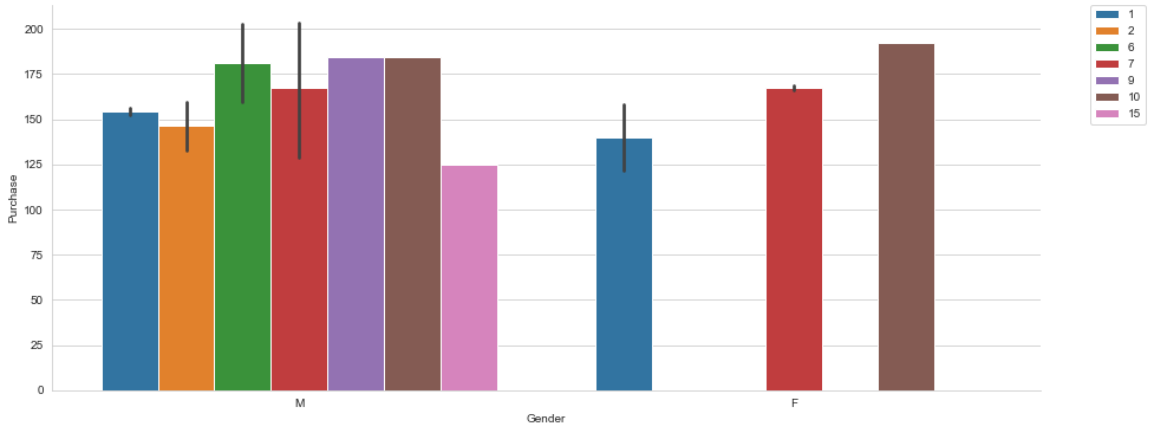
Es decir, de los 141 productos menos vendidos, tan solo 17 superan el tercer rango intercuartil que representa a productos que superan los U$S 120 valor de compra. En otras palabras, solo 17 productos superan el precio promedio del 75% de los productos en la muestra.

Al considerar la categoría de los productos menos vendidos, se observa que la categoría 8 aglutina a mas de 60 de los 141 productos menos vendidos. Le sigue en orden de importancia la categoría 5 que contiene a mas de 30 de los 141 productos menos vendidos.

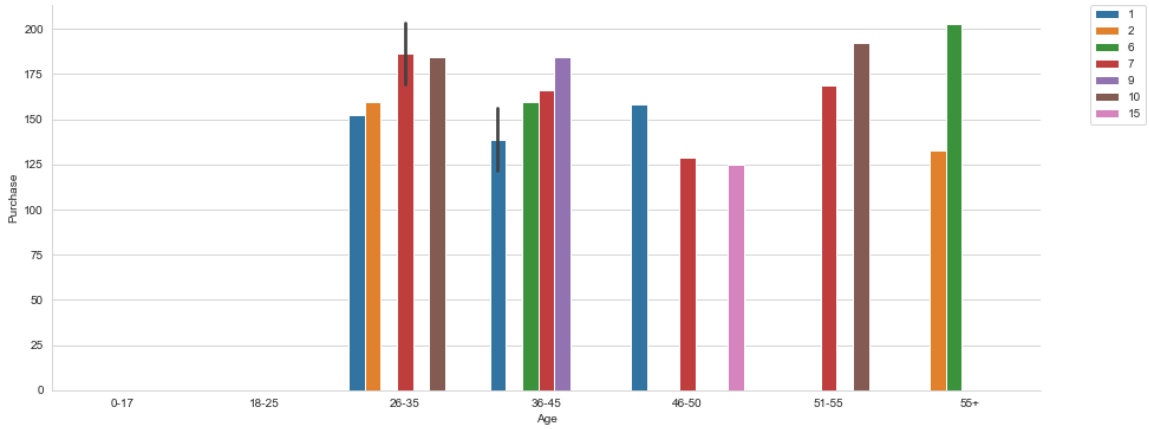
Out[52]:

```
0.7092198581560284
```

El 71% de los productos menos vendidos corresponden a productos de la categoría 8 y 5.



Del total de productos menos vendidos y cuyo precio supera al precio que contiene al 75% del total de la muestra, se observa que ninguno de los 17 pertenece a la categoría 8 o 5. Más aún, se observa que de los productos más caros y menos vendidos, solo la categoría 1, 7 y 15 es compartida por personas de ambos géneros; mientras que el resto de las categorías representan a productos menos adquiridos por el género masculino.



Al analizar el rango etario de los productos menos adquiridos y más costosos, se observa una concentración entre los rangos 26-35 y 36-45, que a su vez consisten en los rangos etarios que consumen productos más costosos; tal como se señaló en el punto 3. A su vez, se observa que mismos rangos etarios comparten la categoría de productos 1 y 7 como las menos consumidas; consitente con lo observado cuando se analizo la distribución por género de la categoría de productos menos consumidos.

Adicionalmente, se observa que la categoría de productos 7 atraviesa a gran parte de los consumidores de variada edad; con la excepción de aquellos consumidores mayores a los 55 años. A su vez, la categoría 15 al igual que la categoría 9 solo figuran una sola vez en los siguientes rangos etarios: 46-50 y 36-45, respectivamente.

A continuación, se analiza el caso de los productos menos adquiridos y cuyos precios se ubican por debajo del 25% de los precios del total de las observaciones. En otras palabras, se estudian los productos menos adquiridos y más baratos.

```
Del total de productos menos consumidos, solo  67 se encuentran fuera del
25% inferior de las obervaciones
```

Out[70]:

	Product_Category_1						Product_ID						Purchase		
Age	**18-25**	**26-35**	**36-45**	**46-50**	**51-55**	**55+**	**18-25**	**26-35**	**36-45**	**46-50**	**51-55**	**55+**	**18-25**	**26-35**	**36-45**
Gender															
F	5.0	1.0	2.0	NaN	NaN	2.0	5.0	1.0	2.0	NaN	NaN	2.0	5.0	1.0	2.0
M	23.0	14.0	7.0	6.0	2.0	5.0	23.0	14.0	7.0	6.0	2.0	5.0	23.0	14.0	7.0

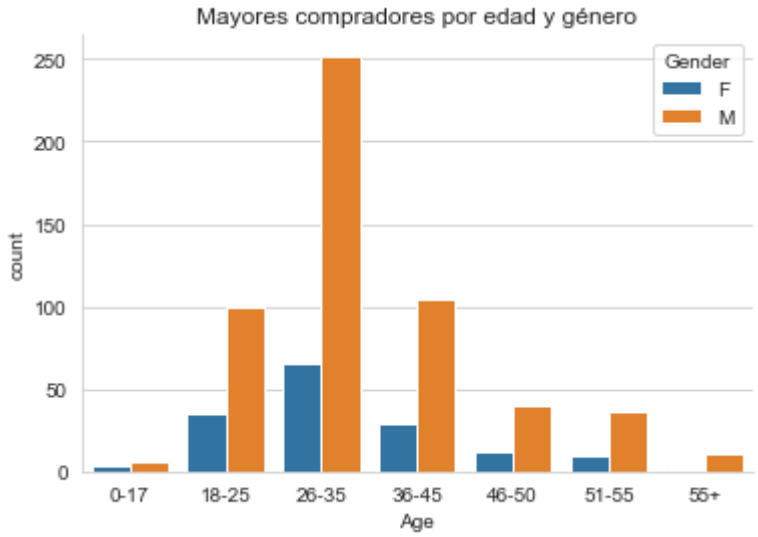
En línea con lo observado previamente, los productos menos consumidos y más económicos corresponden a consumidores entre 18 y 25 años y masculinos.

Out[75]:

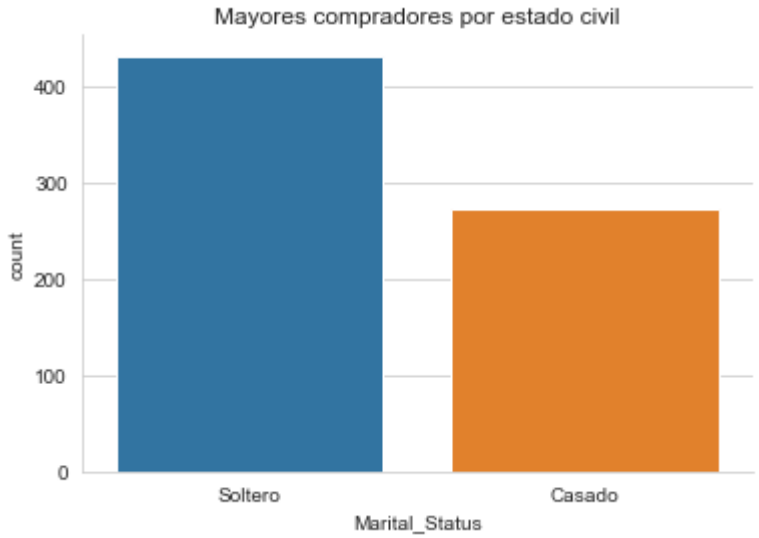
```
0.5957446808510638
```

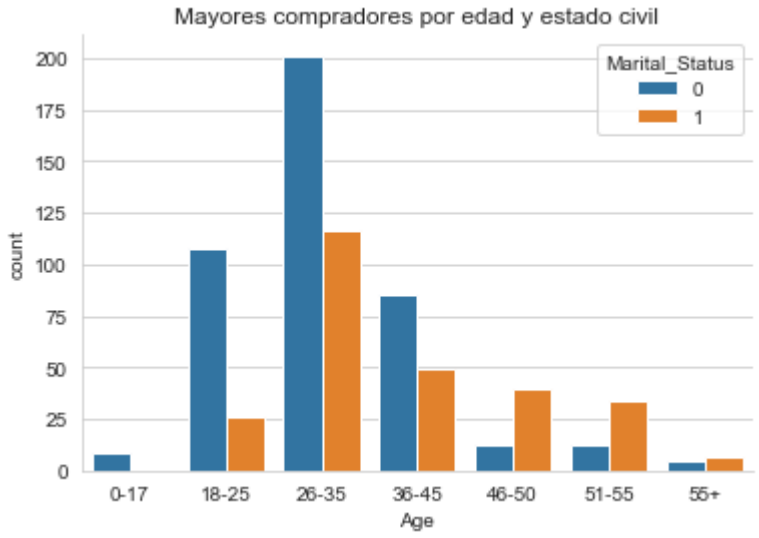
El precio del 60% de los productos menos consumidos se encuentran fuera del 50% de las observaciones que están alrededor de la mediana.

6. Los usuarios que más compran (usuarios con más de 200 órdenes) poseen características en común? Cuáles?

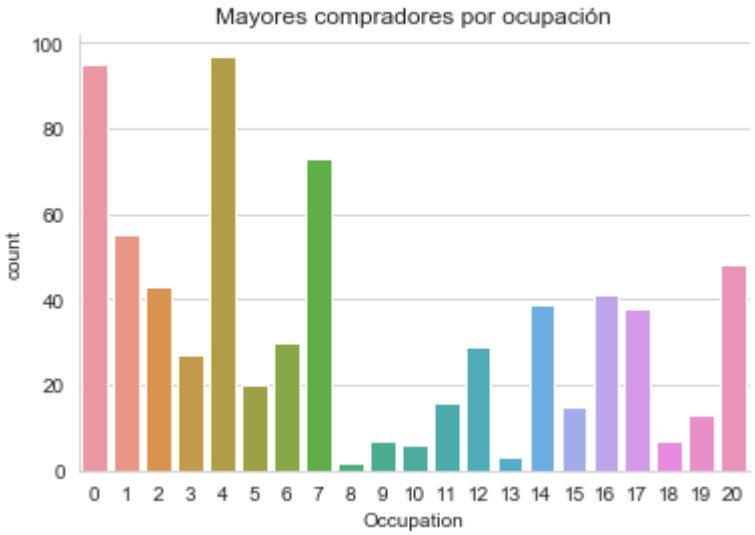


La gran mayoría de los usuarios que realizaron más de 200 compras son hombres, casi triplicando el número de mujeres. A su vez, más del 40% de los grandes usuarios se encuentran entre los 26 y 35 años de edad, seguido por aquellos usuarios entre 18 y 25 años y 36 y 45 años, cada uno con un 20% del total. Es decir que entre los 18 y 45 años se encuentra el 80% de todos los usuarios que hicieron más de 200 compras.Viendo la cantidad de estos usuarios por edad y género, se observa que las proporciones entre hombres y mujeres se mantienen estables a lo largo de los distintos rangos etarios.





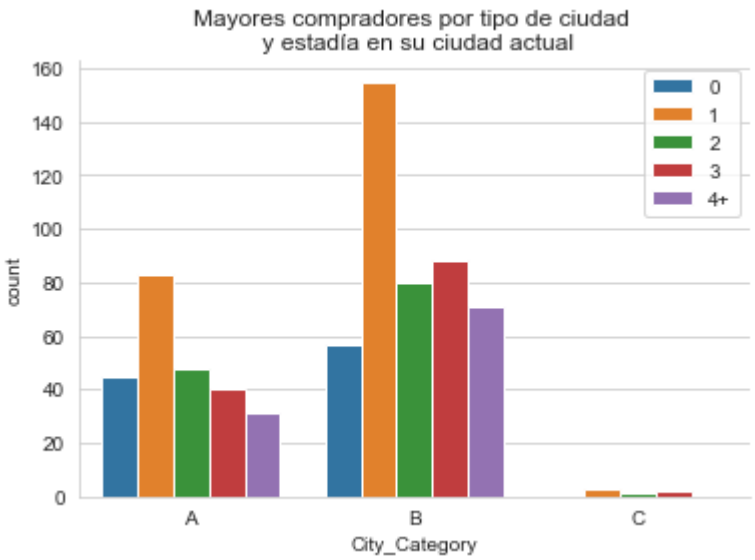
Analizando el estado marital de estos usuarios, se observa que aproximadamente el 60% de estos son solteros mientras que los restantes se encuentran casados, aunque esta relación varía notablemente a lo largo de los años. Hasta los 45 años la proporción de solteros es mayor a los casados, mientras que pasada esa edad la proporción de estos últimos supera a la de los primeros con amplitud. De esta gráfico, se puede observar también que el grupo que más usuarios concentra es aquél que contiene solteros entre 26 y 35 años, y que acumula aproximadamente el 28% del total.



En cuanto a la ocupación de los grandes usuarios, también se distinguen unas pocas actividades que concentran una gran proporción del total, estas son la ocupación 0, 4 y 7, que en total contemplan más del 35% del total de usuarios.

Out[95]:

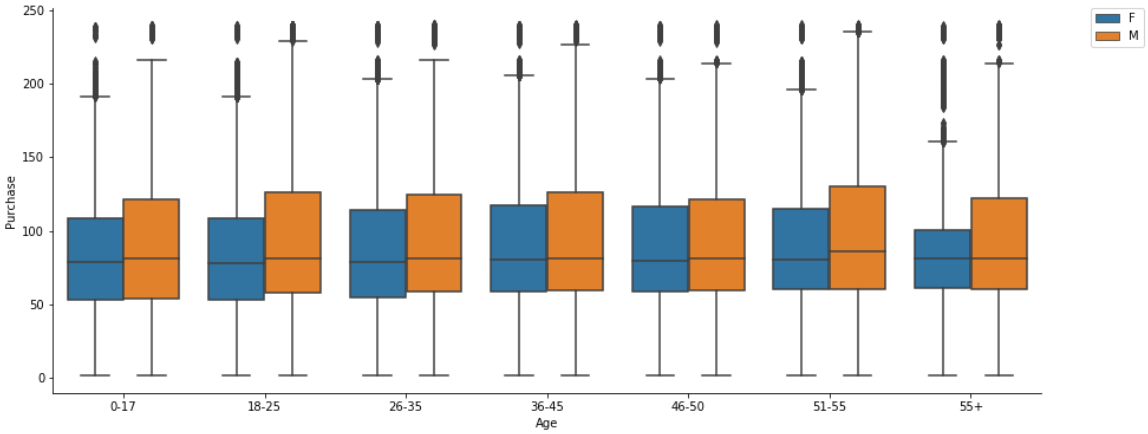
```
<matplotlib.legend.Legend at 0x1600df153c8>
```



En cuanto a la ciudad y tiempo de estadía en esta, el recuento muestra que la gran mayoría de los compradores provienen de las ciudades de tipo B, y dentro de esta sobresale el grupo de personas que ha estado solamente un año en esta ciudad y representa más del 20% del total de grandes usuarios. Este mismo grupo de personas es el más numeroso también entre los grandes usuarios de las ciudades de tipo A. Por último, la proporción de usuarios provenientes de ciudades tipo C es casi nula.

7. Analizar la distribución de ventas por edad, género y estado civil. Graficar.

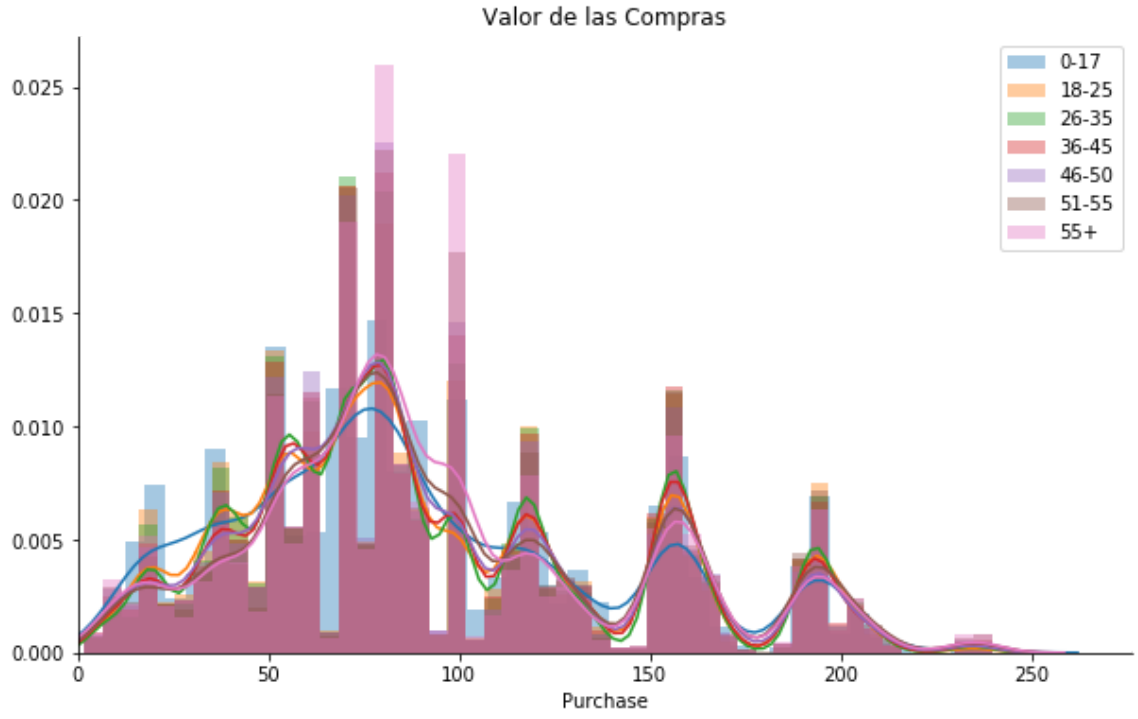
Distribución de las ventas por edad



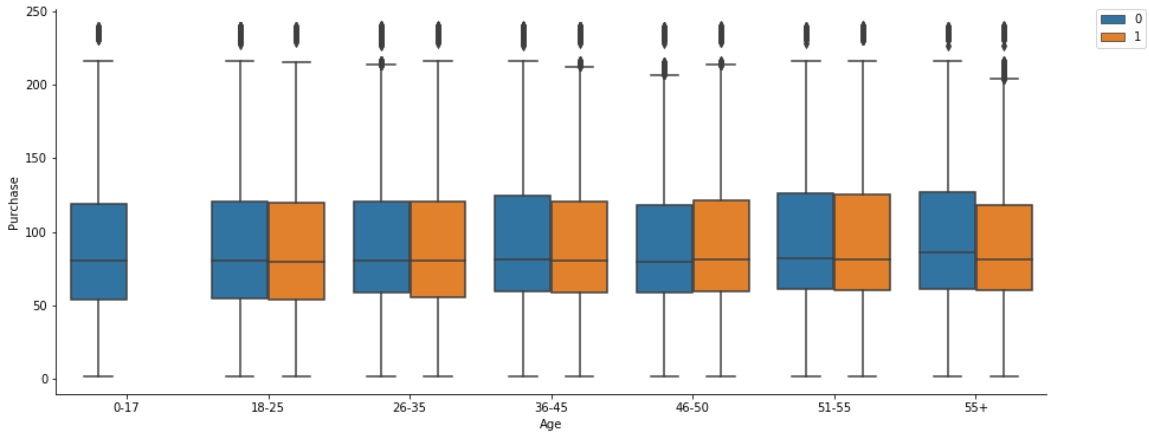
Del análisis de la distribución de las ventas por edad y por sexo, surge que la mediana de compra de los hombres siempre se ubica por encima (o en igual posición) que las mujeres, independientemente el rango etario considerado. A su vez, el 3° rango intercuartílico de los hombres (que acumula el 75% de las observaciones) siempre se ubica por encima de las mujeres, mientras que el primer rango intercuartílico de las mujeres siempres se ubica por debajo en relación al de los hombres. Lo señalado anteriormente, permite sugerir que los hombres presentan una mayor prediposición en adquirir productos más costosos frente a las mujeres.

Out[323]:

```
<matplotlib.legend.Legend at 0x16461f56eb8>
```



Distribución de las ventas por Estado Civil



La distribución de las compras por edad y estado civil deja en evidencia que las personas solteras adquieren productos más costosos cuando superan los 55 años y cuando se encuentran entre los 36-45 años, considerando aún el rango etario 0-17, que por generalidad son personas aún solteras. En el resto de los rangos etarios definidos, las personas casadas adquieren productos más costosos.

8. Qué categoría de ciudad posee el mayor porcentaje de compras dada la proporción de usuarios que contiene? Graficar.

Out[368]:

	City_Category	Cantidad Usuarios
0	A	1045
1	B	1707
2	C	3139

Out[369]:

	City_Category	Cantidad Ordenes
0	A	144638
1	B	226493
2	C	166446

Out[371]:

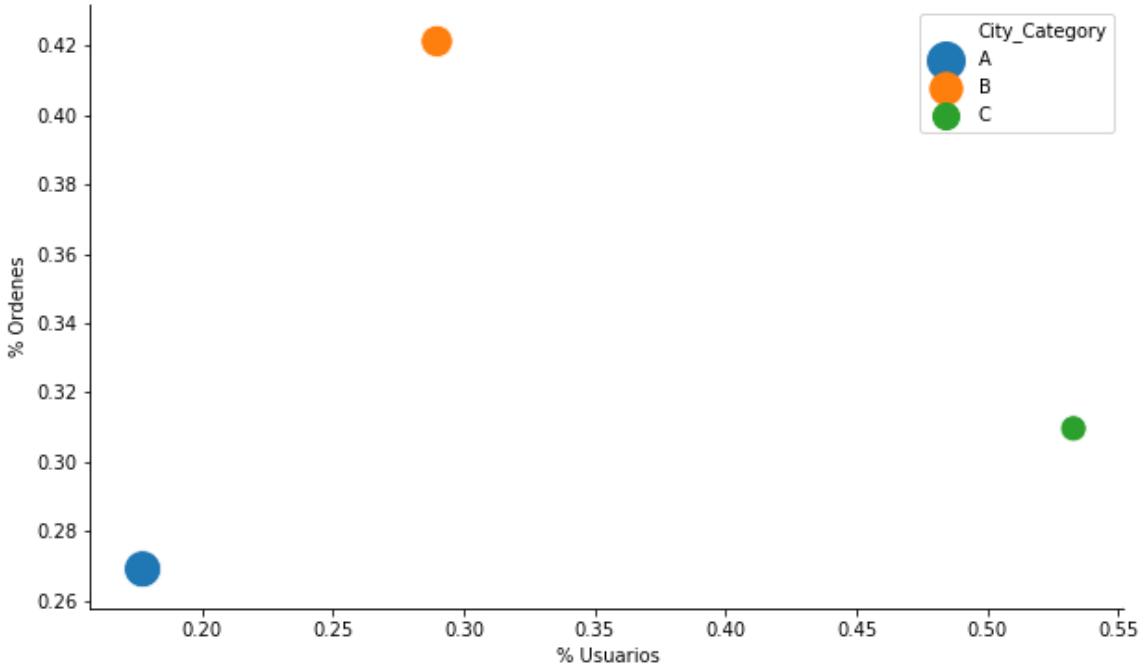
	City_Category	Cantidad Usuarios	Cantidad Ordenes	Compra Promedio por usuario	% Usuarios	% Ordenes	Ciudad más demandante
0	A	1045	144638	138.409569	0.177389	0.269055	0.091666
1	B	1707	226493	132.684827	0.289764	0.421322	0.131558
2	C	3139	166446	53.025167	0.532847	0.309623	-0.223224

Del análisis sobre la cantidad de órdenes por ciudad, surge que la ciudad "B" lidera el ránking al concentrar el 42% del total de ordenes analizadas. En segundo lugar, se ubica la ciudad "C" que concentra el el 31% del total de órdenes y la ciudad "A" completa el 27% restante. No obstante, al considerar la cantidad de usuarios que concentra cada ciudad, surge que "C" contiene el 53% del total de consumidores en la muestra, mientras que las ciudades "A" y "B" contran el 18% y 29%, respectivamente. Por lo tanto, si bien la ciudad "C" se ubica en segundo lugar al considerar la cantidad de órdenes ejecutadas, concentra a más de la mitad de los compradores, por lo que debería esperarse un mayor nivel de órdenes.

Lo comentado previamente se refuerza al observar que la cantidad de ordenes promedio por usuario en la ciudad "C" representan menos de la mitad de las órdenes promedio por usuario en las ciudades "A" y "B".

Out[334]:

	City_Category	% Usuarios	% Ordenes
0	A	0.177389	0.269055
1	B	0.289764	0.421322
2	C	0.532847	0.309623

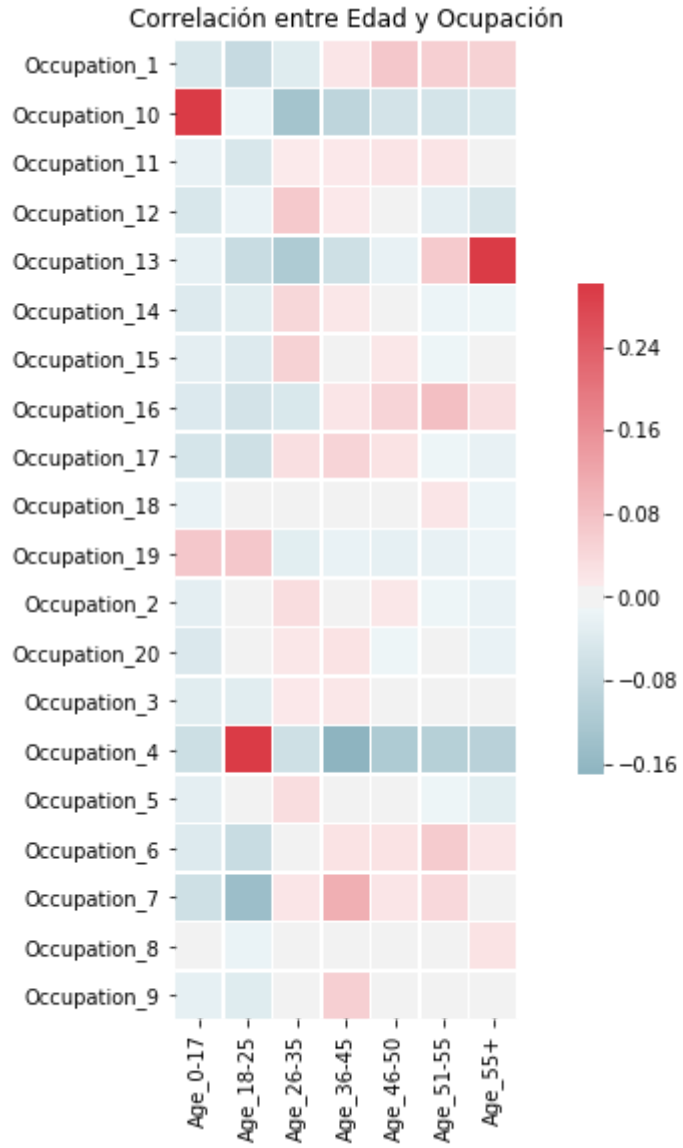


En el Scatter-Plot se logra apreciar que la ciudad "B" se ubica por encima de la línea de equidistribución mientras que la ciudad "C" se ubica por debajo de la línea de equidistribución. La línea de equidistribución es una recta que imaginaria de correspondencia 1-1 entre el % de órdenes y el % de usuarios. En otras palabras, si el % de ordenes coincidiera con el % de usuarios, todos los puntos se ubicarían sobre una recta de 45° que parte del origen. Esto implica, que en la ciudad "B" la proporción de órdenes supera ampliamente la proporción de usuarios radicados en la ciudad "B". Por otro lado, en la ciudad "C" se presenta un caso opuesto, dado que el porcentaje de usuarios supera ampliamente el % de órdenes. En cuanto a la ciudad "A" se observa que el % de ordenes supera levemente al % de usuarios.

9. Analizar la correlación entre edad y ocupación de los consumidores

Out[44]:

```
Text(0.5,1,'Correlación entre Edad y Ocupación')
```



El mapa de calor muestra que en las edades mas bajas (entre 0 y 17), las ocupaciones se encuentran menos dispersos, ubicandose la mayoría en la categoría 10, y unos pocos en la categoría 0. La gran concentración en la categoría 10 podría ser un indicativo de que esta representa a los estudiantes. Luego, en el grupo 18-25, la ocupación que concentra la gran mayoría es la 4, aunque aparecen varias categorías con usuarios que no se encontraban en el grupo anterior. A medida que observamos los grupos de edad mas grandes los puestos de trabajo se distribuyen más homogéneamente a lo largo de las distintas ocupaciones, lo que no significa que todas ocupen las misma cantidad de usuarios. Esta dispersión es mayor en el grupo 26-35, pero progresivamente decae y por último, en el rango etario de 55 años o más, se da una gran concentración en la ocupación 13, por lo que esta categoría podría pertenecer a los jubilados.