tarea_3_nicolas_netz

June 14, 2023

1 Laboratorio de métodos aplicados: Tarea 3

Estudiante: Nicolás Netz

N°Mat: 2018458791

Correo: nnetz2018@udec.cl

Prof. Juan Carlos Caro

2 Importar los módulos necesarios y la data

En el presente notebook se utilizan distintas librerías para apoyar los cálculos y las visualizaciones. Además, se incluye una serie de funciones extra, para reutilizar código en distintas celdas.

3 Pregunta 1: Cargar base de datos

Cargue la base de datos y realice los ajustes necesarios para su uso (missing values, recodificar variables, etcetera). Identifique los tipos de datos que se encuentran en la base, realice estadisticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

	income	kidhome	teenhome	recency	mntwines	mntfruits	mntmeatproducts	\
0	58138.0	0	0	58	635	88	546	
1	46344.0	1	1	38	11	1	6	
2	71613.0	0	0	26	426	49	127	
3	26646.0	1	0	26	11	4	20	
4	58293.0	1	0	94	173	43	118	

	mntfishproducts	mntsweetproducts	mntgoldprods	…	marital_together	\
0	172	88	88	…	0	
1	2	1	6	…	0	
2	111	21	42	…	1	
3	10	3	5	…	1	
4	46	27	15	…	0	

	marital_widow	education_2n cycle	education_basic	education_graduation	\
0	0	0	0	1	
1	0	0	0	1	

```
2              0                    0                0                     1
3              0                    0                0                     1
4              0                    0                0                     0

   education_master  education_phd  mnttotal  mntregularprods  \
0                 0              0      1529             1441
1                 0              0        21               15
2                 0              0       734              692
3                 0              0        48               43
4                 0              1       407              392

   acceptedcmpoverall
0                   0
1                   0
2                   0
3                   0
4                   0

[5 rows x 39 columns]
```

3.1 Notas al leer la data

Se tienen 2205 entradas en 39 columnas, las cuales corresponden a lo indicado en la tabla:

- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- Response
- Complain
- DtCustomer
- Education
- Marital
- Kidhome
- Teenhome
- Income
- MntFishProducts
- MntMeatProducts
- MntFruits
- MntSweetProducts
- MntWines
- MntGoldProds
- NumDealsPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebPurchases
- NumWebVisitsMonth

- Recency
- Age
- Customer_Days
- marital_divorced
- marital_married
- marital_single
- marital_together
- marital_widow
- education_basic
- education_2n
- education_graduation
- education_master
- education_phd
- MntTotal
- z_costcontact
- z_revenue
- mntregularprods
- acceptedcmpoverall

No se identificaron valores NaN en ninguna de las columnas. A continuación se presentan los estadísticos descriptivos de las variables, para analizar rangos (mínimos y máximos), promedio y count. Lo unico extraño que se identifica es que en la variable mntregularprods hay valores negativos, indicando que una persona recibió dinero al comprar mntregularprods ó algo similar. Debido a esto, se filtran esos valores.

A continuación se presentan los estadísticos descriptivos de las variables.

3.2 Notas sobre estadísticos descriptivos

Se revisó que los valores mínimos y máximos de las distintas variables estén en sus rangos.

Algunas cosas que notar: * mntregularprods tiene valores negativos, por lo que se filtran.

	income	kidhome	teenhome	recency	mntwines	\
count	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	
mean	51622.094785	0.442177	0.506576	49.009070	306.164626	
std	20713.063826	0.537132	0.544380	28.932111	337.493839	
min	1730.000000	0.000000	0.000000	0.000000	0.000000	
25%	35196.000000	0.000000	0.000000	24.000000	24.000000	
50%	51287.000000	0.000000	0.000000	49.000000	178.000000	
75%	68281.000000	1.000000	1.000000	74.000000	507.000000	
max	113734.000000	2.000000	2.000000	99.000000	1493.000000	

	mntfruits	mntmeatproducts	mntfishproducts	mntsweetproducts	\
count	2205.000000	2205.000000	2205.000000	2205.000000	
mean	26.403175	165.312018	37.756463	27.128345	
std	39.784484	217.784507	54.824635	41.130468	
min	0.000000	0.000000	0.000000	0.000000	
25%	2.000000	16.000000	3.000000	1.000000	

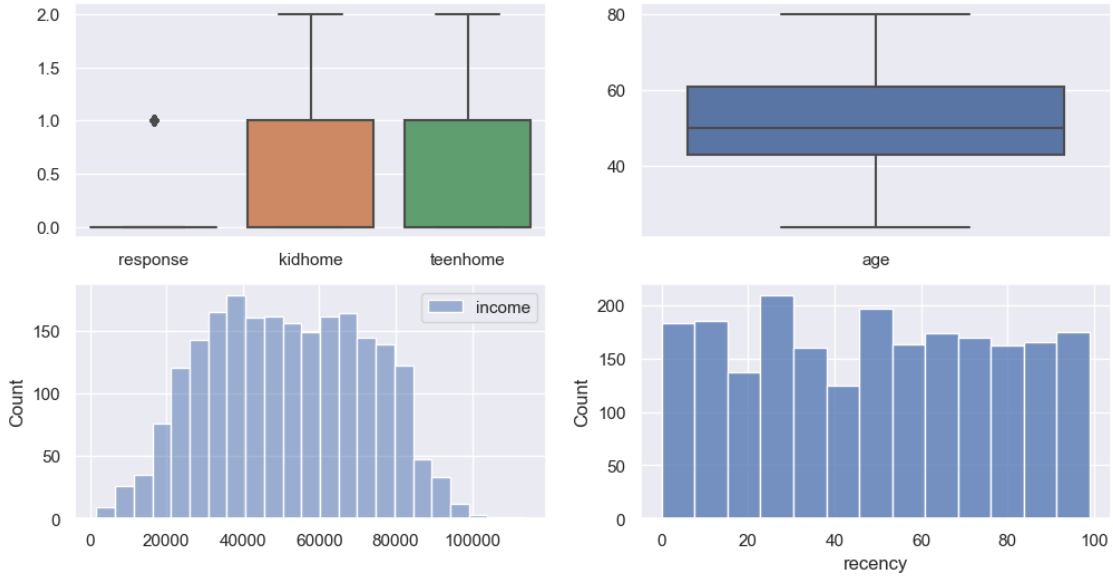
```
50%       8.000000        68.000000        12.000000          8.000000
75%      33.000000       232.000000        50.000000         34.000000
max     199.000000      1725.000000       259.000000        262.000000

       mntgoldprods  …  marital_together  marital_widow  education_2n cycle  \
count   2205.000000  …       2205.000000    2205.000000         2205.000000
mean      44.057143  …          0.257596       0.034467            0.089796
std       51.736211  …          0.437410       0.182467            0.285954
min        0.000000  …          0.000000       0.000000            0.000000
25%        9.000000  …          0.000000       0.000000            0.000000
50%       25.000000  …          0.000000       0.000000            0.000000
75%       56.000000  …          1.000000       0.000000            0.000000
max      321.000000  …          1.000000       1.000000            1.000000

       education_basic  education_graduation  education_master  education_phd  \
count      2205.000000           2205.000000       2205.000000    2205.000000
mean          0.024490              0.504762          0.165079       0.215873
std           0.154599              0.500091          0.371336       0.411520
min           0.000000              0.000000          0.000000       0.000000
25%           0.000000              0.000000          0.000000       0.000000
50%           0.000000              1.000000          0.000000       0.000000
75%           0.000000              1.000000          0.000000       0.000000
max           1.000000              1.000000          1.000000       1.000000

          mnttotal  mntregularprods  acceptedcmpoverall
count  2205.000000      2205.000000          2205.00000
mean    562.764626       518.707483             0.29932
std     575.936911       553.847248             0.68044
min       4.000000      -283.000000             0.00000
25%      56.000000        42.000000             0.00000
50%     343.000000       288.000000             0.00000
75%     964.000000       884.000000             0.00000
max    2491.000000      2458.000000             4.00000

[8 rows x 39 columns]
```

3.3 Notas sobre la matriz de correlación y la distribución de las variables

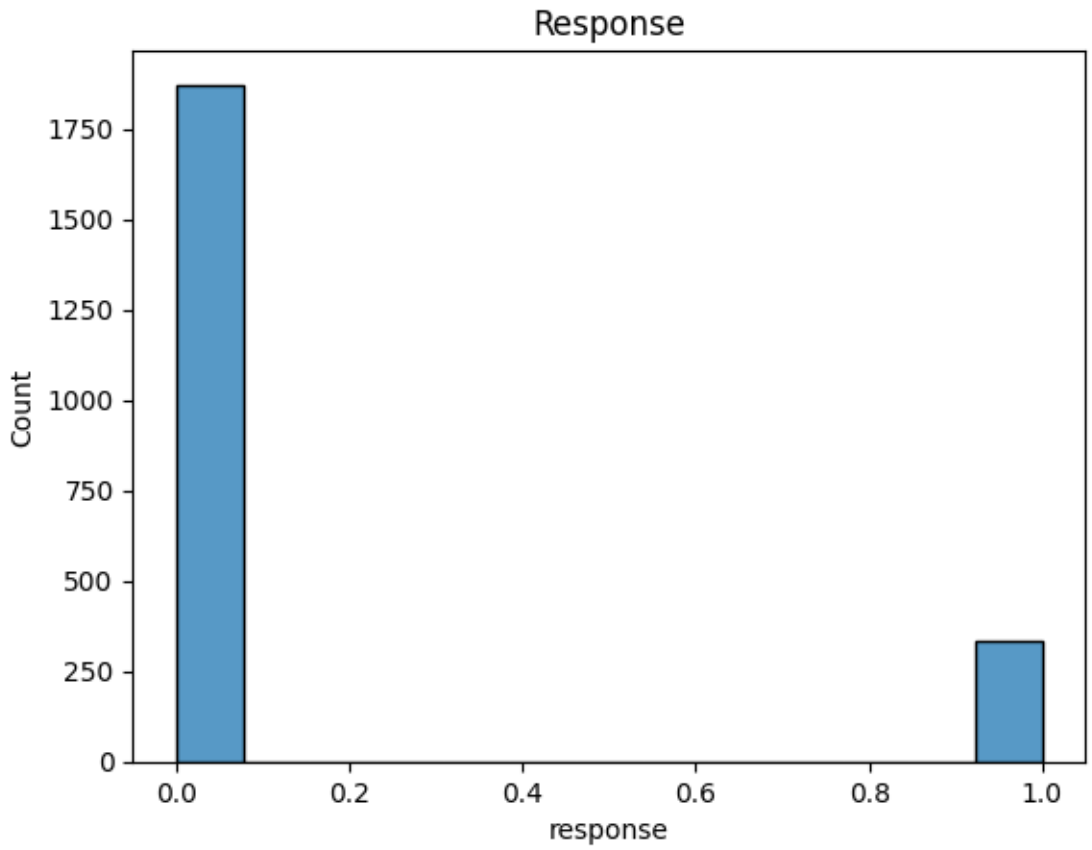
En los gráficos de esta sección se revisaron algunas distribuciones de las variables. Notamos que las personas con response = 1 son pocas. Se identifica que la mayoría de las personas no tienen un hijo en la casa, ni un adolescente, o a lo más tienen uno. En términos de rango etario, la mayoría se concentra entre los 40 y los 60 años. La mayoría de las personas tiene un income de entre 40000 y 75000. Notamos que la variable Recency tiene una distribución predominantemente uniforme, sin embargo, la de income tiene más forma de camapana. Es importante notar esto, pues en un análisis de más adelante, se convierte la variable income en tramos. (low, medium y high income).

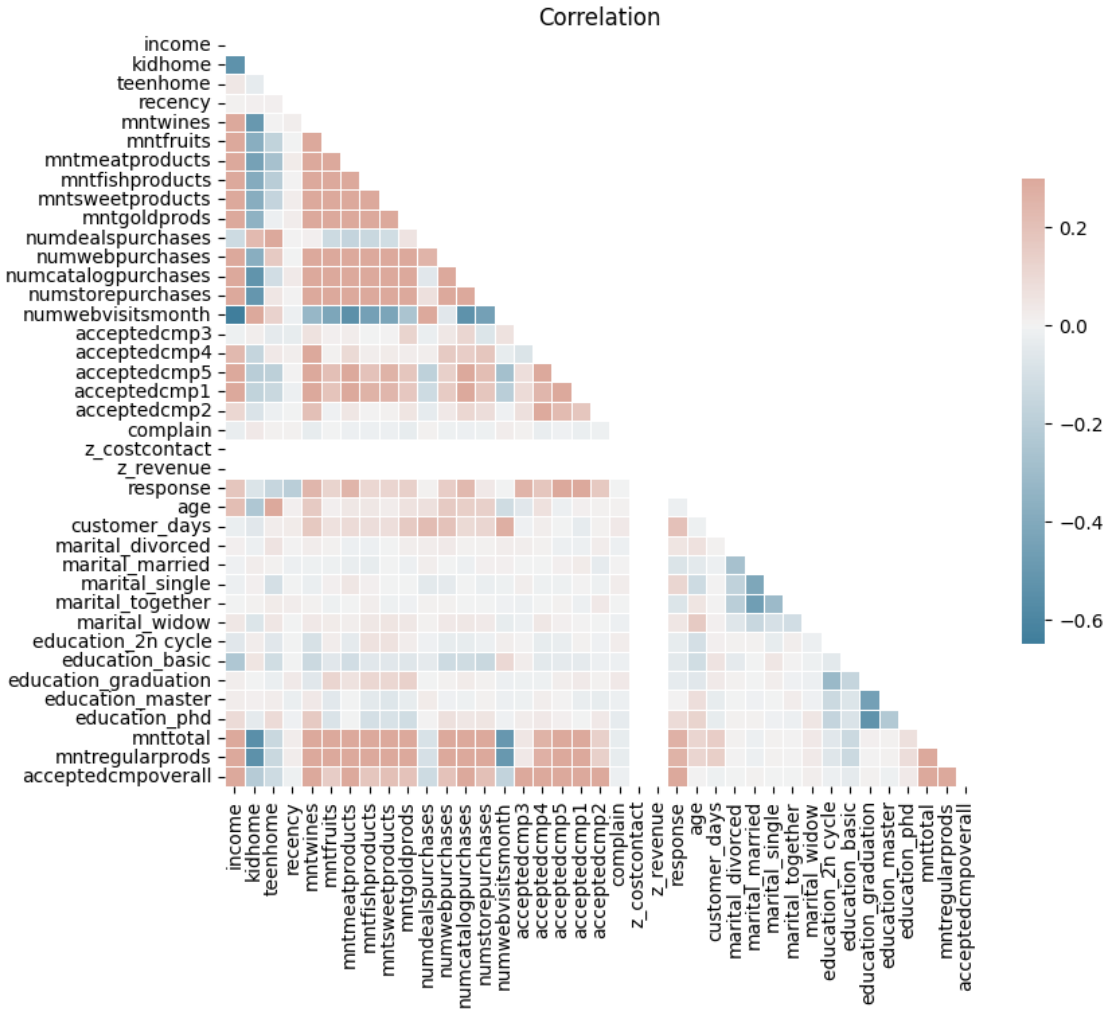
Al ver la matriz de correlación, se identifican valores con alta correlación (negativa). Las variables con correlación más fuerte son las de tipo num% y mnt%. No hay correlación perfecta.

```
<Axes: xlabel='recency', ylabel='Count'>
```




```
<bound method Text.set_x of Text(0.5, 1.0, 'Response')>
```





4 Pregunta 2: PCA

Realice un PCA usando las variables de numero de compras y cantidad gastada en los diversos items. En particular, identifique los valores propios y determine el numero optimo de componentes. Luego estime y grafique la distribucion de los componentes. Ademas discuta la importancia relativa de las variables sobre cada uno de los componentes estimados. Que se puede concluir de este analisis?

4.1 Respuesta 2: PCA

Se realiza un PCA usando las variables (11 variables): * 'numdealspurchases','numwebpurchases', 'numcatalogpurchases', 'numstorepurchases','numwebvisitsmonth','mntwines', 'mnt-fruits','mntmeatproducts', 'mntfishproducts', 'mntsweetproducts','mntgoldprods'

Para utilizar estas variables, que tienen distintas unidades de medida y escalas, se usa la clase

StandardScaler de scikit-learn para aplicar una transformación de estandarización a los datos. (Se le quita la media y se escala a la varianza unitaria).

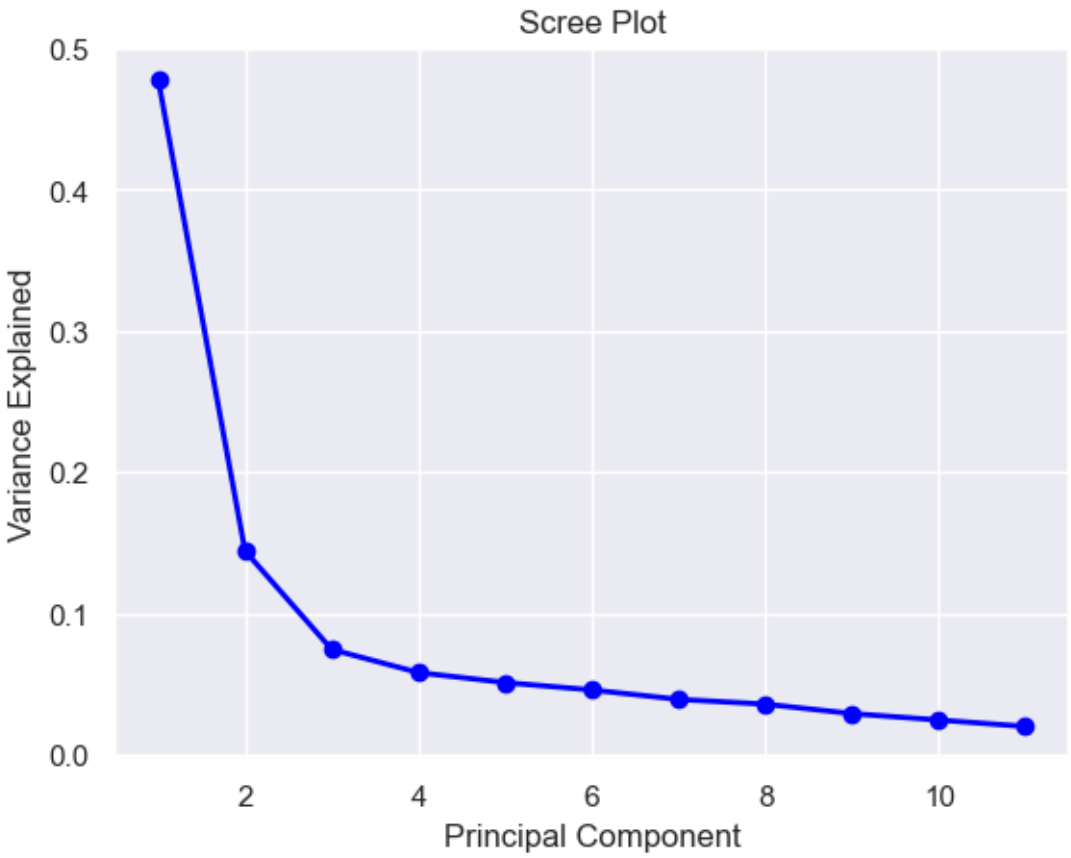
Notar que no se utilizó la variable mnttotal, pues no es más que la suma de las otras "mnt%" variables anteriores, lo que generaría problemas de especificación.

A partir de este PCA, se obtienen los siguientes valores propios:

variable	valor propio
PCA1	0.4776
PCA2	0.1443
PCA3	0.0745
PCA4	0.0581
PCA5	0.0511
PCA6	0.0458
PCA7	0.0392
PCA8	0.0358
PCA9	0.0291
PCA10	0.0245
PCA11	0.0200

Notamos que, a partir del Scree Plot siguiente, se genera un punto de inflexión con 2 Componentes principales, por lo que se decide usar ese número de componentes (explica el 0.47 + 0.14 = 0.62 de la varianza , mientras que si se agrega el tercero se llegaría solo a 0.69 y sólo complicaría el análisis.)

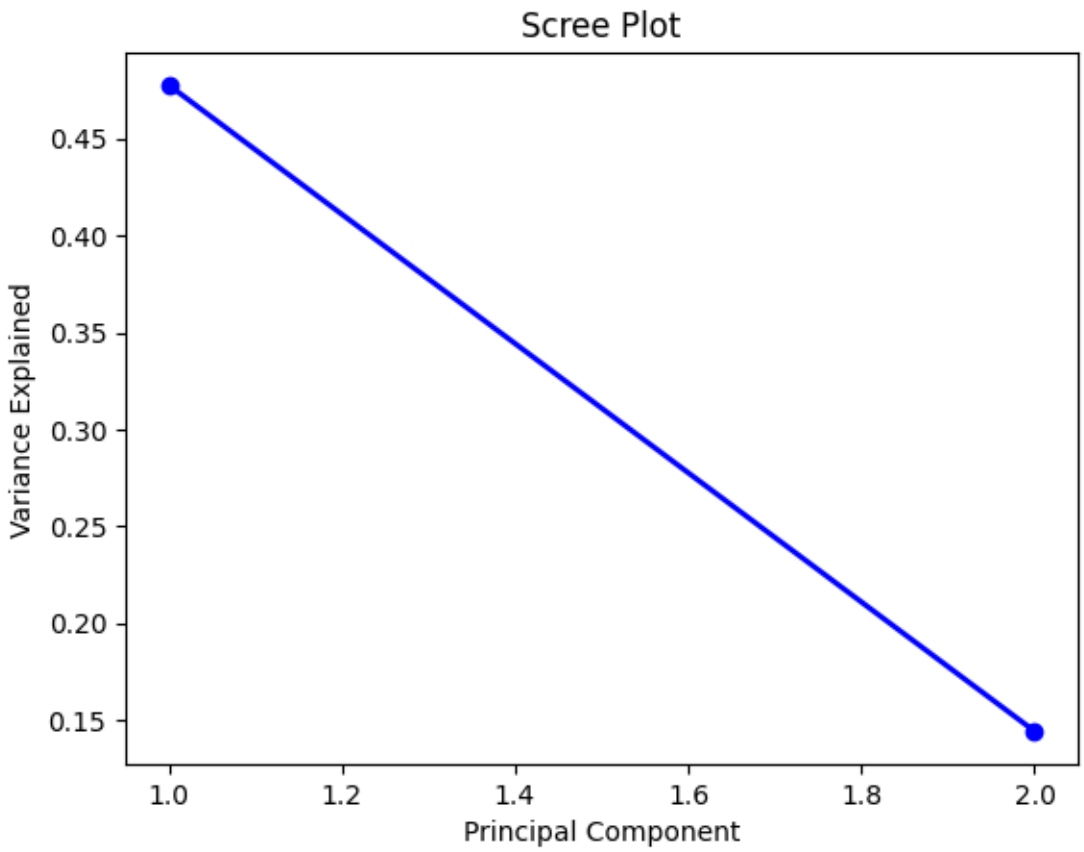
```
[0.4776 0.1443 0.0745 0.0581 0.0511 0.0458 0.0392 0.0358 0.0291 0.0245
 0.0200]
```



Así, según lo discutido anteriormente, se realiza un nuevo PCA con sólo 2 componentes. Se obtienen los siguientes valores propios:

variable	valor propio
PCA1	0.4776
PCA2	0.1443

Donde la figura siguiente muestra el scree plot.



4.1.1 Sobre la importancia relativa de las variables sobre cada componente estimado.

En la siguiente tabla se presentan los vectores propios. Al analizar estos, podemos ver que las distintas variables (columnas) tienen diferentes importancias relativas sobre los componentes 1 y 2 (filas). Notamos que algunas variables tienen mayor importancia sobre un componente que sobre otro. Esto ayuda a indicar que tal vez un grupo de variables que tiene similar importancia sobre los componentes pertenencen a un grupo. Todo esto se justificaría por los efectos de la varianza a la hora de calcular los componentes.

Tabla con vectores propios

	numdealspurchases	numwebpurchases	numcatalogpurchases	\
PC1	-0.048176	0.249046	0.367340	
PC2	0.643732	0.493824	0.014259	

	numstorepurchases	numwebvisitsmonth	mntwines	mntfruits	\
PC1	0.329704	-0.275900	0.325823	0.318278	
PC2	0.187767	0.402126	0.224310	-0.128922	

	mntmeatproducts	mntfishproducts	mntsweetproducts	mntgoldprods

PC1	0.358649	0.328884	0.317219	0.266472
PC2	-0.123291	-0.141465	-0.112040	0.172859

5 Pregunta 3: PCA 3 componentes

Con los resultados de la Pregunta 2, mantenga los primeros 3 componentes principales y repita el analisis. Graficamente y estadisticamente indique si existen diferencias o relaciones significativas entre los valores de los PCA y las siguientes variables: Income, Kidhome, Education y Recency. Que puede concluir de los resultados?

5.1 Respuesta 3: PCA 3

Se ejecuta el modelo con los 3 componentes principales. Al revisar el scree plot a continuación, se identifica que los componentes principales alcanzan a explicar cerca del 70% de la varianza. La importancia relativa de las variables sobre los componentes se puede ver en la tabla con vectores propios más abajo. Con esa tabla, se puede de identificar grupos de variables que tienen efectos similares sobre los componentes.

variable	valor
PC1	0.47762042
PC2	0.14426904
PC3	0.07451216

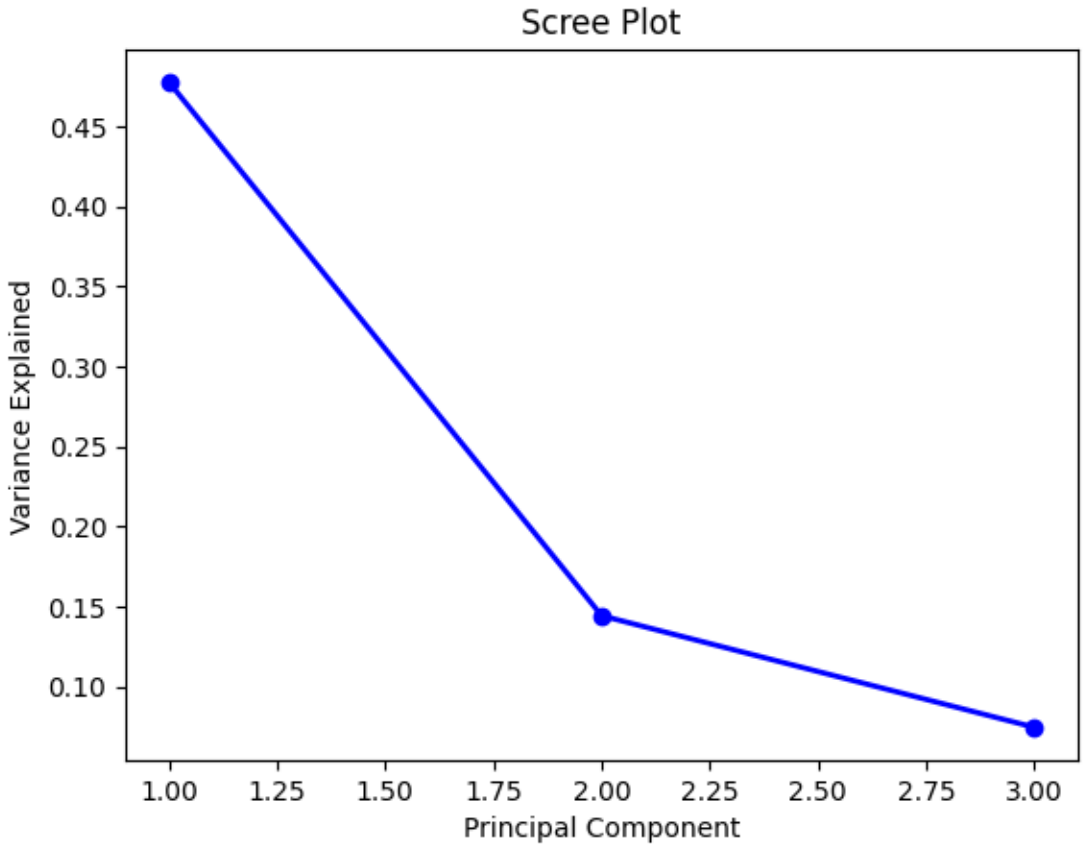


Tabla con vectores propios

	numdealspurchases	numwebpurchases	numcatalogpurchases	\
PC1	-0.048176	0.249046	0.367340	
PC2	0.643732	0.493824	0.014259	
PC3	0.234977	-0.064529	-0.190959	

	numstorepurchases	numwebvisitsmonth	mntwines	mntfruits	\
PC1	0.329704	-0.275900	0.325823	0.318278	
PC2	0.187767	0.402126	0.224310	-0.128922	
PC3	-0.264638	0.256848	-0.465169	0.393674	

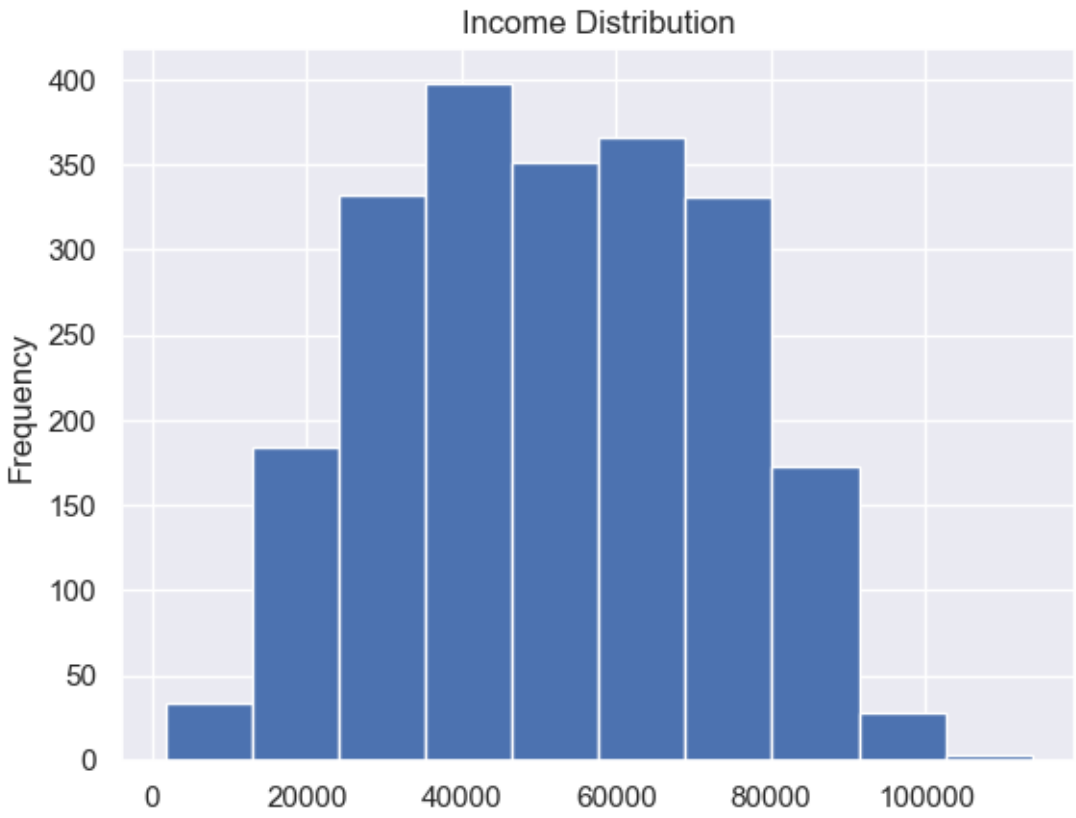
	mntmeatproducts	mntfishproducts	mntsweetproducts	mntgoldprods
PC1	0.358649	0.328884	0.317219	0.266472
PC2	-0.123291	-0.141465	-0.112040	0.172859
PC3	-0.122385	0.368499	0.348843	0.352605

5.1.1 Valores PCA y Income

Para analizar la relación entre el PCA e Income, se revisa gráficamente la distribución de esta variable para generar 3 grupos. low income [0,25000], medium income [25000, 80000] y high income [80000,max(income)].

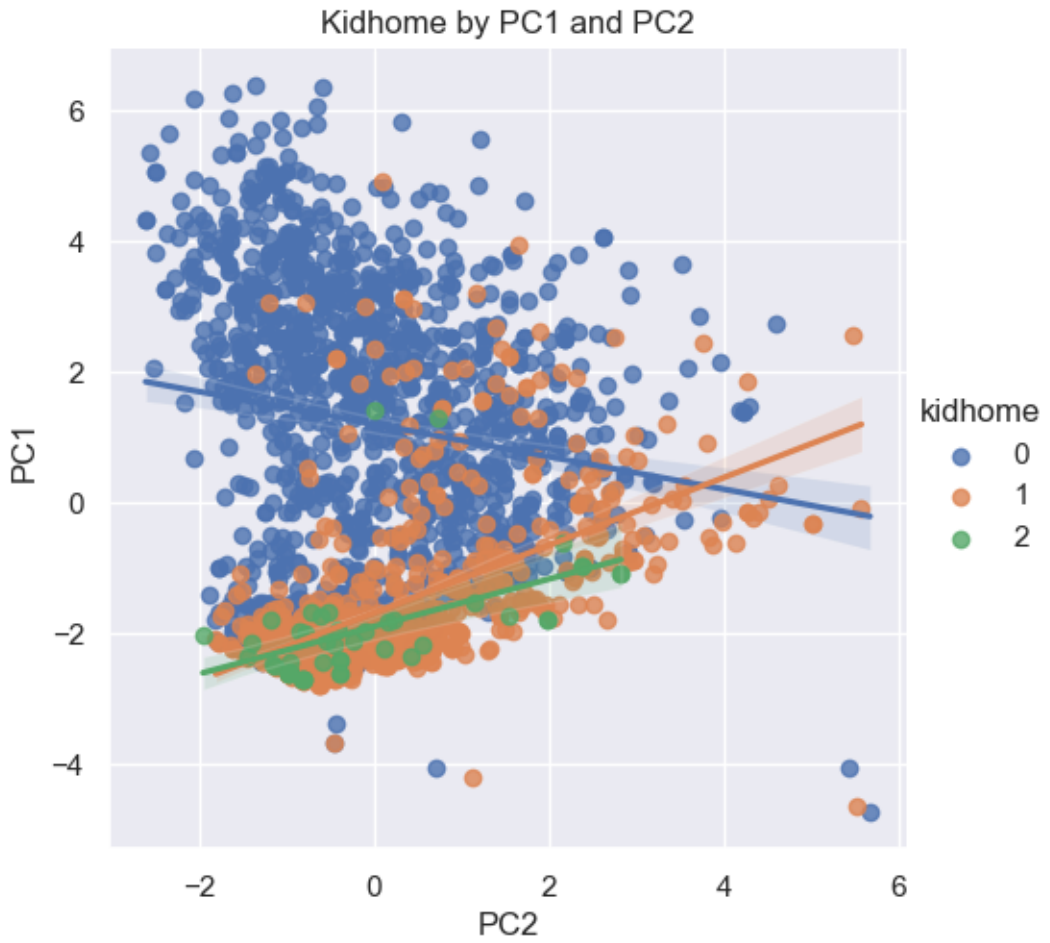
En la figura Income by PC1, PC2 se ajustó una regresión a cada grupo sobre los componentes principales 1 y 2. En cada figura se puede ver la recta con la que se ajustó y sus intervalos de confianza. Se puede identificar que estos 3 grupos son significativamente distintos pues ninguna de las rectas ni sus intervalos de confianza del error se intersectan. Se podría hacer el mismo análisis numericamente.

```
<Axes: title={'center': 'Income Distribution'}, ylabel='Frequency'>
```


Income Distribution
Frequency
400
350
300
250
200
150
100
50
0
0
20000
40000
60000
80000
100000

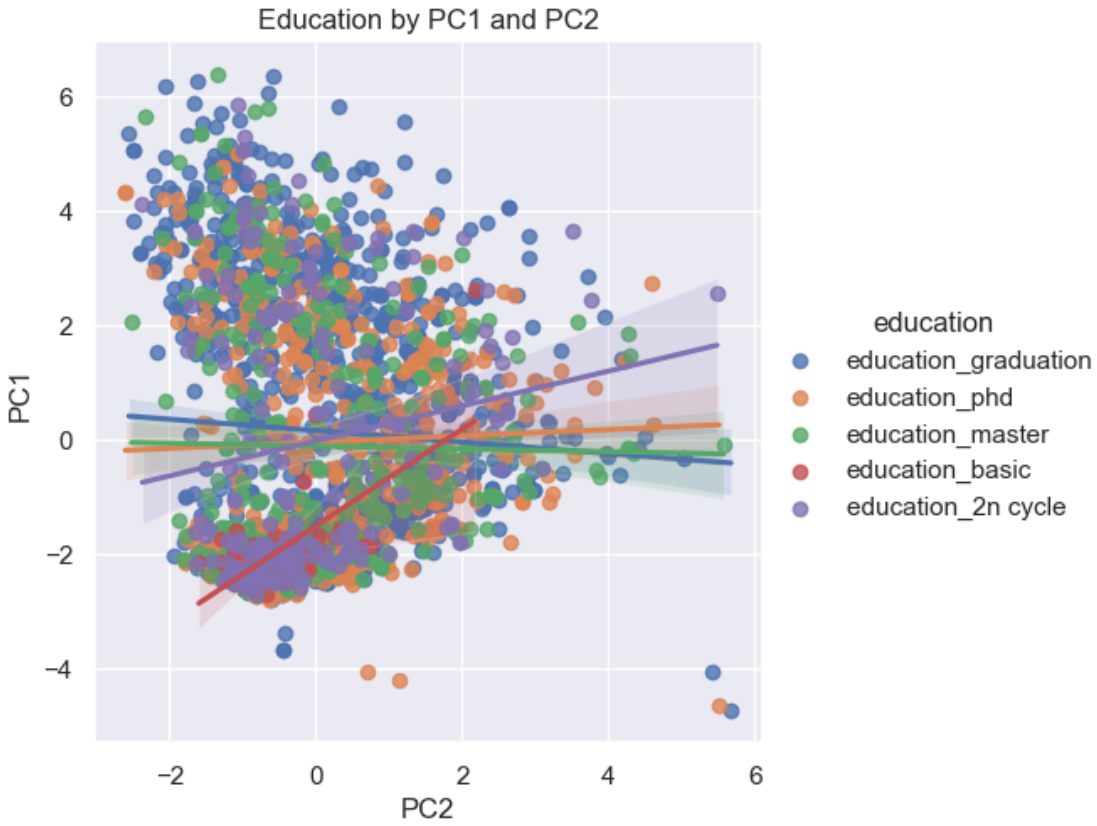
5.1.2 Valores PCA y Kidhome

Se sigue la misma metodología anterior, y se identifica que el grupo de kidhome que no tiene un hijo en casa tiene un comportamiento distinto a los que tienen 1 hijo o más en casa, pues las rectas de regresión son distintas.



5.1.3 Valores PCA y Education

Se identifica que sobre la variable education, el grupo education_basic es significativamente distinto de los demás, sin embargo entre los otros no es posible identificar una diferencia significativa.

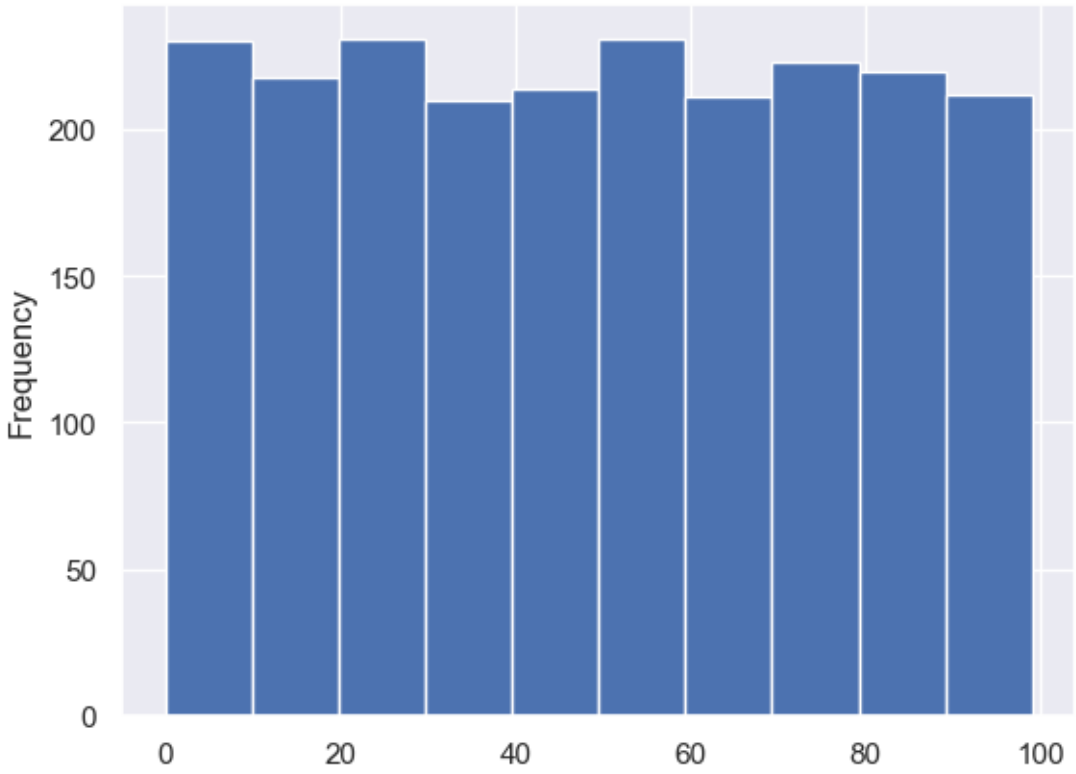


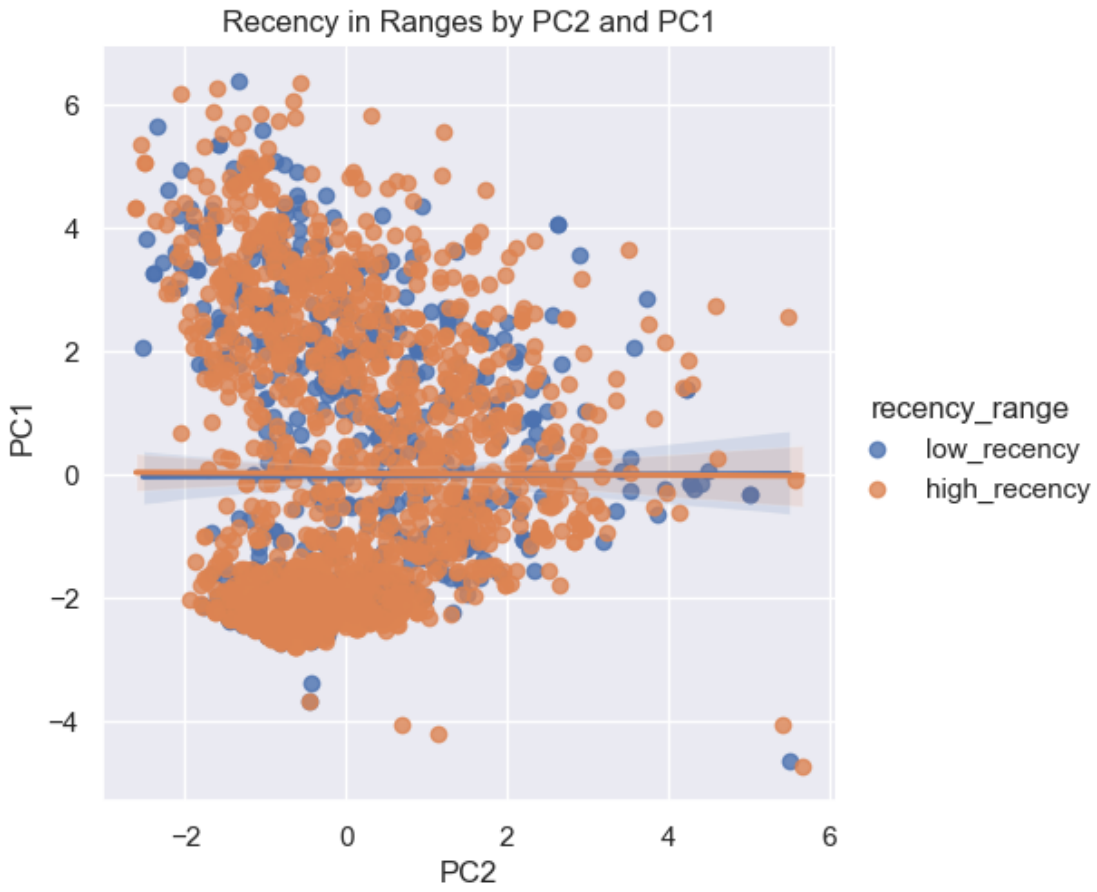
5.1.4 Valores PCA y Recency

Se identifica que la distribución de la variable es más bien uniforme, sin embargo, se agrupa en dos partes, quienes tienen bajos valores de recency y quienes tienen altos (0,30) y (30, max(recency)).

Al ver la figura Recency in Ranges, no se identifica una diferencia significativa de los valores del PCA sobre los grupos.

```
<Axes: ylabel='Frequency'>
```


Frequency
200
150
100
50
0
0
20
40
60
80
100



6 Pregunta 4: EFA

A partir del mismo set de variables de la pregunta 2 realice un EFA. En particular determine el numero optimo de factores y las variables que se asocian a cada factor. Tambien discuta si existen variables que no son informativas

En esta sección se usan las variables de la pregunta 2:

- numdealspurchases
- numwebpurchases
- numcatalogpurchases
- numstorepurchases
- numwebvisitsmonthmntwines
- mntfruits
- mntmeatproducts
- mntfishproducts
- mntsweetproducts
- mntgoldprods

Se aplica un Análisis Factorial Exploratorio, donde todas las medidas se relacionan con las variables

latentes. Se utiliza el análisis factorial mediante 'promax'. Notar que los datos no se normalizan como en las secciones anteriores, pues los efectos de escala no influyen en el resultado.

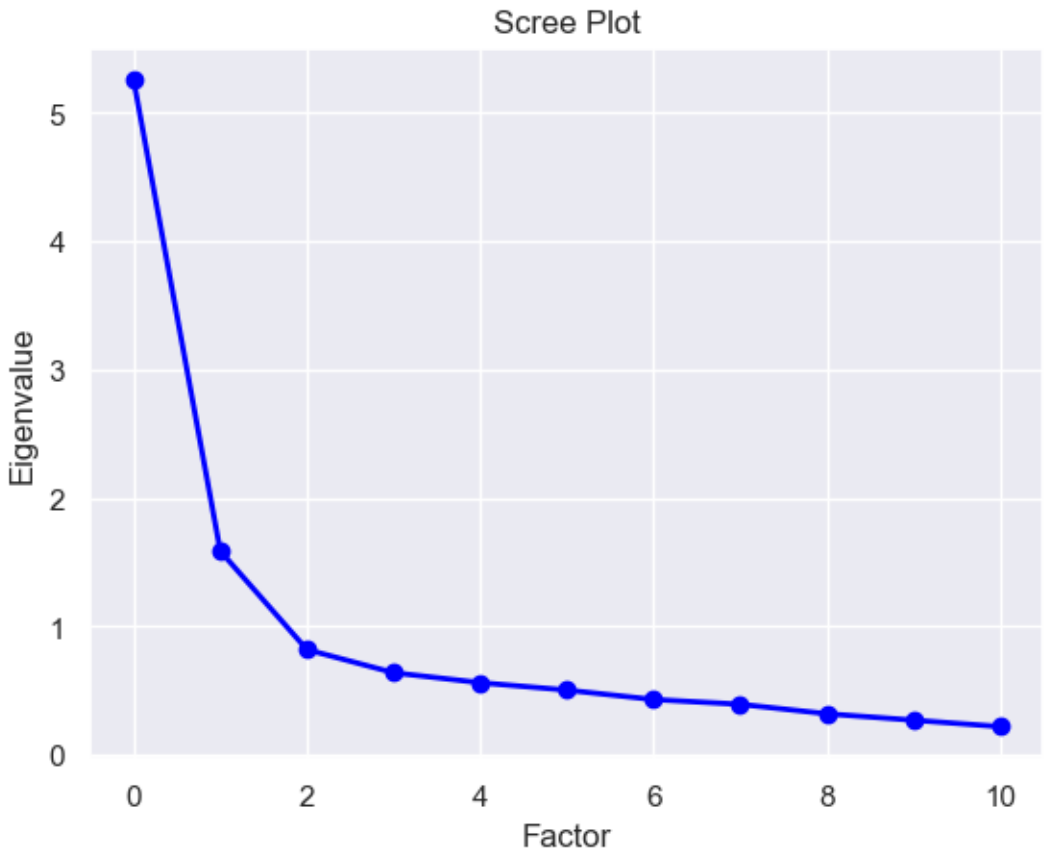
```
FactorAnalyzer(rotation_kwargs={})
```

```
Resultados analisis factorial exploratorio
```

	Factor1	Factor2	Factor3
numdealspurchases	0.100497	-0.036368	0.588414
numwebpurchases	0.557527	0.178358	0.495534
numcatalogpurchases	0.653613	0.185869	-0.149049
numstorepurchases	0.651740	0.135909	0.053503
numwebvisitsmonth	-0.330337	-0.153133	0.577575
mntwines	1.049109	-0.254270	-0.006035
mntfruits	-0.080855	0.808227	-0.035443
mntmeatproducts	0.455455	0.327126	-0.248123
mntfishproducts	-0.082705	0.837038	-0.056700
mntsweetproducts	-0.028239	0.750323	-0.032244
mntgoldprods	0.202965	0.433575	0.155904

A partir de la tabla anterior "Resultados analisis factorial exploratorio", se identifica la magnitud (en desviaciones estándar) con la que una variable influye en cada factor. Así, se identifica por ejemplo, que el factor 1 se relaciona fuertemente con mntwines, mntmeatproducts, numcatalogpurchases y numstorepurchases.

No se identifica una variable que sea poco informativa, pues todas tienen una influencia de al menos 0.5 en algún factor.



6.0.1 Número optimo de factores

Al ver el Scree Plot anterior, se identifica que el punto de inflexión ocurre con 3 factores. Por lo tanto, el número optimo de factores sería 3.

```
eta1 =~ mntmeatproducts + mntfishproducts + mntfruits + mntsweetproducts +
mntwines + mntgoldprods
```

7 Pregunta 5: CFA

Con los resultados obtenidos en la Pregunta 4, proponga un CFA donde cada variable solo se asocia con un factor. Entregue un nombre a cada factor que representa el concepto comun entre todas las variables. Reporte la importancia de cada medida (variable) a cada factor e indique la correlacion entre factores.

```
Name of objective: MLW
Optimization method: SLSQP
Optimization successful.
Optimization terminated successfully
Objective value: 0.578
```

```
Number of iterations: 134
Params: 1.272 9.880 0.062 1.392 0.037 4.766 923.267 56968.387 3.555 969.835
7.713 1513.135 1326.893 600.494 646.679 57.123 860.163 59.278 0.000
```

	lval	op	rval	Estimate	Est. Std	\
0	mntsweetproducts	~	luxury	1.000000e+00	6.183852e-01	
1	mntgoldprods	~	luxury	1.271834e+00	6.501121e-01	
2	mntwines	~	luxury	9.879684e+00	7.121202e-01	
3	numwebpurchases	~	luxury	6.227114e-02	5.729126e-01	
4	mntfruits	~	standard	1.000000e+00	6.944842e-01	
5	mntfishproducts	~	standard	1.391954e+00	7.239692e-01	
6	numcatalogpurchases	~	discount	1.000000e+00	6.951380e-09	
7	numdealspurchases	~	discount	3.727229e-02	3.816421e-10	
8	luxury	~~	luxury	6.004945e+02	1.000000e+00	
9	luxury	~~	standard	6.466791e+02	8.997958e-01	
10	luxury	~~	discount	5.712259e+01	1.207480e+08	
11	standard	~~	standard	8.601630e+02	1.000000e+00	
12	standard	~~	discount	5.927778e+01	1.046955e+08	
13	discount	~~	discount	3.726893e-16	1.000000e+00	
14	numwebpurchases	~~	numwebpurchases	4.765706e+00	6.717712e-01	
15	mntfruits	~~	mntfruits	9.232667e+02	5.176917e-01	
16	mntwines	~~	mntwines	5.696839e+04	4.928848e-01	
17	numdealspurchases	~~	numdealspurchases	3.554730e+00	1.000000e+00	
18	mntsweetproducts	~~	mntsweetproducts	9.698352e+02	6.175998e-01	
19	numcatalogpurchases	~~	numcatalogpurchases	7.712672e+00	1.000000e+00	
20	mntfishproducts	~~	mntfishproducts	1.513135e+03	4.758687e-01	
21	mntgoldprods	~~	mntgoldprods	1.326893e+03	5.773543e-01	

	Std. Err	z-value	p-value
0	-	-	-
1	0.051011	24.932414	0.0
2	0.370492	26.666406	0.0
3	0.002758	22.575609	0.0
4	-	-	-
5	0.051432	27.064099	0.0
6	-	-	-
7	0.019476	1.913779	0.055648
8	39.874883	15.059466	0.0
9	33.403132	19.359834	0.0
10	2.559552	22.317418	0.0
11	52.802344	16.290242	0.0
12	2.738448	21.646491	0.0
13	2.995079	0.0	1.0
14	0.155897	30.56955	0.0
15	38.032494	24.275734	0.0
16	2092.9327	27.219407	0.0
17	0.10726	33.141284	0.0

```
18  32.532953  29.810857       0.0
19   3.004093   2.567388  0.010247
20  67.841271  22.304054       0.0
21  45.550396  29.130224       0.0

       DoF  DoF Baseline         chi2  chi2 p-value  chi2 Baseline       CFI  \
Value   17            28  1271.270989           0.0    7722.513269  0.836992

            GFI      AGFI       NFI       TLI     RMSEA        AIC  \
Value  0.835381  0.728863  0.835381  0.731515  0.183172  36.844299

              BIC   LogLik
Value  145.072339  0.57785
```