Tarea 4 SEM Javiera San Martín

June 14, 2023

Section 5: Structural Equation Modelling

0.1 Housekeeping and Data

[1]:
```
pip install semopy
```

```
Requirement already satisfied: semopy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (2.3.9)
Requirement already satisfied: sklearn in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy) (0.0)
Requirement already satisfied: pandas in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy)
(1.2.4)
Requirement already satisfied: scipy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy)
(1.6.2)
Requirement already satisfied: statsmodels in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy)
(0.12.2)
Requirement already satisfied: sympy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy) (1.8)
Requirement already satisfied: numpy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from semopy)
(1.20.1)
Requirement already satisfied: python-dateutil>=2.7.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
pandas->semopy) (2.8.1)
Requirement already satisfied: pytz>=2017.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
pandas->semopy) (2021.1)
Requirement already satisfied: six>=1.5 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from python-
dateutil>=2.7.3->pandas->semopy) (1.15.0)
Requirement already satisfied: scikit-learn in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
sklearn->semopy) (1.2.2)
Requirement already satisfied: joblib>=1.1.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
```

```
learn->sklearn->semopy) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
learn->sklearn->semopy) (2.1.0)
Requirement already satisfied: patsy>=0.5 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
statsmodels->semopy) (0.5.1)
Requirement already satisfied: mpmath>=0.19 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from sympy->semopy)
(1.2.1)
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
pip install factor_analyzer
```

```
Requirement already satisfied: factor_analyzer in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (0.4.1)
Requirement already satisfied: scikit-learn in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
factor_analyzer) (1.2.2)
Requirement already satisfied: pandas in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
factor_analyzer) (1.2.4)
Requirement already satisfied: scipy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
factor_analyzer) (1.6.2)
Requirement already satisfied: pre-commit in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
factor_analyzer) (3.3.2)
Requirement already satisfied: numpy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
factor_analyzer) (1.20.1)
Requirement already satisfied: python-dateutil>=2.7.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
pandas->factor_analyzer) (2.8.1)
Requirement already satisfied: pytz>=2017.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
pandas->factor_analyzer) (2021.1)
Requirement already satisfied: six>=1.5 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from python-
dateutil>=2.7.3->pandas->factor_analyzer) (1.15.0)
Requirement already satisfied: nodeenv>=0.11.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from pre-
commit->factor_analyzer) (1.8.0)
Requirement already satisfied: virtualenv>=20.10.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from pre-
commit->factor_analyzer) (20.23.0)
Requirement already satisfied: identify>=1.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from pre-
```

```
commit->factor_analyzer) (2.5.24)
Requirement already satisfied: pyyaml>=5.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from pre-
commit->factor_analyzer) (5.4.1)
Requirement already satisfied: cfgv>=2.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from pre-
commit->factor_analyzer) (3.3.1)
Requirement already satisfied: setuptools in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
nodeenv>=0.11.1->pre-commit->factor_analyzer) (52.0.0.post20210125)
Requirement already satisfied: filelock<4,>=3.11 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
virtualenv>=20.10.0->pre-commit->factor_analyzer) (3.12.2)
Requirement already satisfied: platformdirs<4,>=3.2 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
virtualenv>=20.10.0->pre-commit->factor_analyzer) (3.5.3)
Requirement already satisfied: distlib<1,>=0.3.6 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
virtualenv>=20.10.0->pre-commit->factor_analyzer) (0.3.6)
Requirement already satisfied: joblib>=1.1.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
learn->factor_analyzer) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
learn->factor_analyzer) (2.1.0)
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
pip install stepmix
```

```
Requirement already satisfied: stepmix in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (1.1.1)
Requirement already satisfied: pandas in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
(1.2.4)
Requirement already satisfied: numpy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
(1.20.1)
Requirement already satisfied: scikit-learn>=1.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
(1.2.2)
Requirement already satisfied: scipy in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
(1.6.2)
Requirement already satisfied: tqdm in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
(4.59.0)
Requirement already satisfied: matplotlib in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from stepmix)
```

```
(3.3.4)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
learn>=1.0.0->stepmix) (2.1.0)
Requirement already satisfied: joblib>=1.1.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from scikit-
learn>=1.0.0->stepmix) (1.2.0)
Requirement already satisfied: pillow>=6.2.0 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->stepmix) (8.2.0)
Requirement already satisfied: cycler>=0.10 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->stepmix) (0.10.0)
Requirement already satisfied: python-dateutil>=2.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->stepmix) (2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->stepmix) (2.4.7)
Requirement already satisfied: kiwisolver>=1.0.1 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
matplotlib->stepmix) (1.3.1)
Requirement already satisfied: six in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
cycler>=0.10->matplotlib->stepmix) (1.15.0)
Requirement already satisfied: pytz>=2017.3 in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (from
pandas->stepmix) (2021.1)
Note: you may need to restart the kernel to use updated packages.
```

[4]:
```
pip install graphviz
```

```
Requirement already satisfied: graphviz in
/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages (0.20.1)
Note: you may need to restart the kernel to use updated packages.
```

[68]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn
import scipy
from scipy.linalg import eigh, cholesky
from scipy.stats import norm
import linearmodels.panel as lmp
from pylab import plot, show, axis, subplot, xlabel, ylabel, grid
```

```
import semopy
import seaborn as sns
from factor_analyzer import FactorAnalyzer
from sklearn.decomposition import PCA
from IPython.display import Image


%matplotlib inline
```

0.2 Pregunta 1

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.

- El análisis de los datos se realizó en el anexo 1.

Descripciones:

- AcceptedCmp1 = 1 si el cliente aceptó la oferta en la 1ra campaña, 0 de lo contrario
- AcceptedCmp2 = 1 si el cliente aceptó la oferta en la 2da campaña, 0 de lo contrario
- AcceptedCmp3 = 1 si el cliente aceptó la oferta en la 3ra campaña, 0 de lo contrario
- AcceptedCmp4 = 1 si el cliente aceptó la oferta en la 4ta campaña, 0 de lo contrario
- AcceptedCmp5 = 1 si el cliente aceptó la oferta en la 5ta campaña, 0 de lo contrario
- Response (target) = 1 si el cliente aceptó la oferta en la última campaña, 0 de lo contrario
- Complain = 1 si el cliente se quejó en los últimos 2 años
- DtCustomer = fecha de inscripción del cliente en la empresa
- Education = nivel de educación del cliente
- Marital = estado civil
- Kidhome = número de niños pequeños en el hogar
- Teenhome = número de adolescentes en el hogar
- Income = ingresos familiares anuales
- MntFishProducts = cantidad gastada en pescados (productos del mar) en los últimos 2 años
- MntMeatProducts = cantidad gastada en carne en los últimos 2 años
- MntFruits = cantidad gastada en frutas en los últimos 2 años
- MntSweetProducts = cantidad gastada en productos dulces en los últimos 2 años
- MntWines = cantidad gastada en vino en los últimos 2 años
- MntGoldProds = cantidad gastada en productos de "lujo" en los últimos 2 años
- NumDealsPurchases = número de compras realizadas con descuento
- NumCatalogPurchases = número de compras realizadas utilizando el catálogo
- NumStorePurchases = número de compras realizadas directamente en las tiendas
- NumWebPurchases = número de compras realizadas a través del sitio web de la empresa
- NumWebVisitsMonth = número de visitas al sitio web de la empresa en el último mes
- Recency = número de días desde la última compra

[69]:
```
df_food=pd.read_csv('../data/ifood_df.csv')
#data description
df_food.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2205 entries, 0 to 2204
Data columns (total 39 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Income                2205 non-null   float64
 1   Kidhome               2205 non-null   int64  
 2   Teenhome              2205 non-null   int64  
 3   Recency               2205 non-null   int64  
 4   MntWines              2205 non-null   int64  
 5   MntFruits             2205 non-null   int64  
 6   MntMeatProducts       2205 non-null   int64  
 7   MntFishProducts       2205 non-null   int64  
 8   MntSweetProducts      2205 non-null   int64  
 9   MntGoldProds          2205 non-null   int64  
 10  NumDealsPurchases     2205 non-null   int64  
 11  NumWebPurchases       2205 non-null   int64  
 12  NumCatalogPurchases   2205 non-null   int64  
 13  NumStorePurchases     2205 non-null   int64  
 14  NumWebVisitsMonth     2205 non-null   int64  
 15  AcceptedCmp3          2205 non-null   int64  
 16  AcceptedCmp4          2205 non-null   int64  
 17  AcceptedCmp5          2205 non-null   int64  
 18  AcceptedCmp1          2205 non-null   int64  
 19  AcceptedCmp2          2205 non-null   int64  
 20  Complain              2205 non-null   int64  
 21  Z_CostContact         2205 non-null   int64  
 22  Z_Revenue             2205 non-null   int64  
 23  Response              2205 non-null   int64  
 24  Age                   2205 non-null   int64  
 25  Customer_Days         2205 non-null   int64  
 26  marital_Divorced      2205 non-null   int64  
 27  marital_Married       2205 non-null   int64  
 28  marital_Single        2205 non-null   int64  
 29  marital_Together      2205 non-null   int64  
 30  marital_Widow         2205 non-null   int64  
 31  education_2n Cycle    2205 non-null   int64  
 32  education_Basic       2205 non-null   int64  
 33  education_Graduation  2205 non-null   int64  
 34  education_Master      2205 non-null   int64  
 35  education_PhD         2205 non-null   int64  
 36  MntTotal              2205 non-null   int64  
 37  MntRegularProds       2205 non-null   int64  
 38  AcceptedCmpOverall    2205 non-null   int64  
dtypes: float64(1), int64(38)
memory usage: 672.0 KB
```

[70]:
```
df_food.describe()
```

[70]:
```
              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]
```

Correr anexo 1 luego el resto del código.

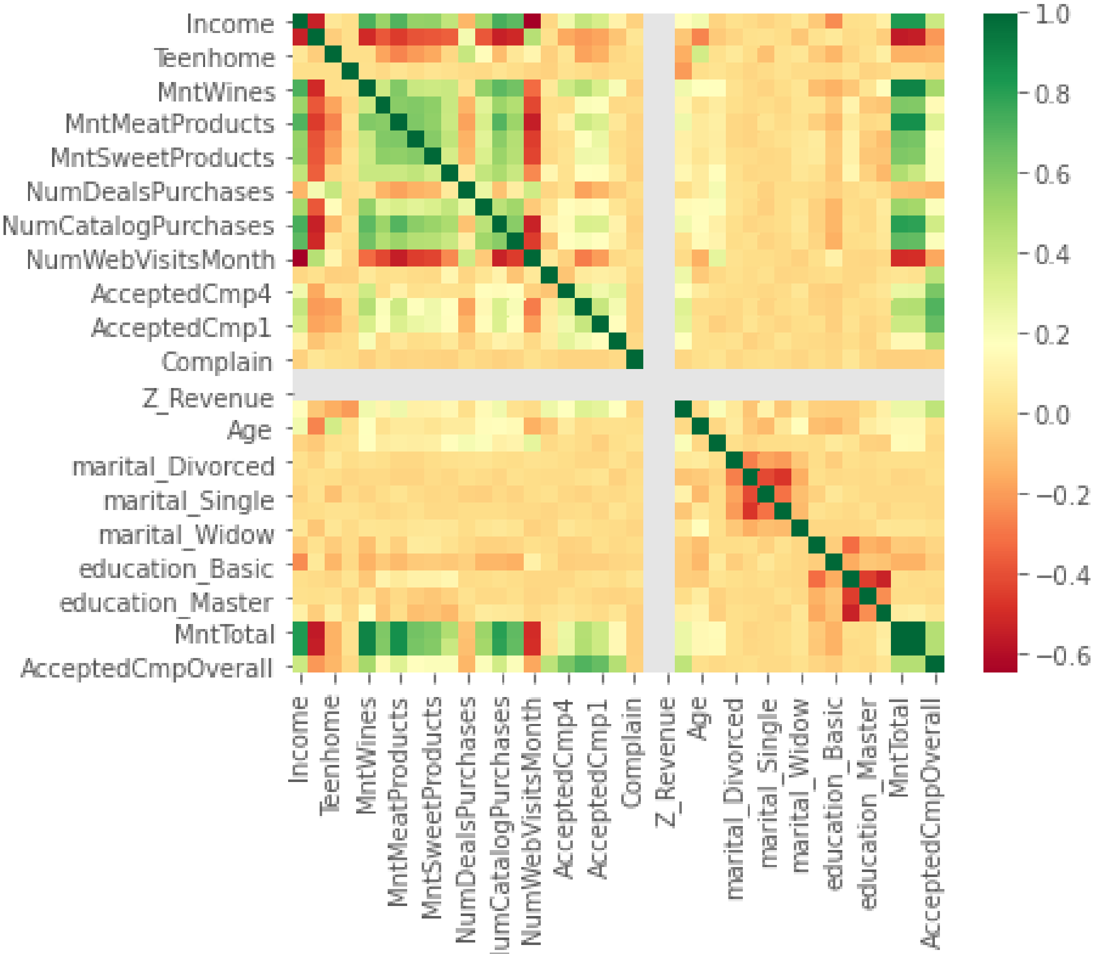
[83]:
```python
df_food2= df_food[['MntWines',"MntFruits", "MntMeatProducts",␣
↪"MntFishProducts", "MntSweetProducts", "MntGoldProds", "NumDealsPurchases",␣
↪"NumWebPurchases", "NumCatalogPurchases",␣
↪"NumStorePurchases","NumWebVisitsMonth"]]
df_food.describe()
```

[83]:
```
              Income      Kidhome     Teenhome      Recency     MntWines  \
count    2181.000000  2181.000000  2181.000000  2181.000000  2181.000000
mean    51479.366804     0.446584     0.512150    49.025676   303.548372
std     20552.087114     0.538017     0.544753    28.987854   334.635002
min      1730.000000     0.000000     0.000000     0.000000     0.000000
25%     35196.000000     0.000000     0.000000    24.000000    24.000000
50%     51124.000000     0.000000     0.000000    49.000000   174.000000
75%     67893.000000     1.000000     1.000000    74.000000   505.000000
max    113734.000000     2.000000     2.000000    99.000000  1493.000000

         MntFruits  MntMeatProducts  MntFishProducts  MntSweetProducts  \
count  2181.000000      2181.000000      2181.000000       2181.000000
mean     26.146263       161.785420        37.286566         26.814764
std      39.519130       212.083811        54.409878         40.893861
min       0.000000         1.000000         0.000000          0.000000
25%       2.000000        16.000000         3.000000          1.000000
50%       8.000000        67.000000        12.000000          8.000000
75%      33.000000       224.000000        49.000000         33.000000
max     199.000000       984.000000       259.000000        262.000000

       MntGoldProds  …  marital_Together  marital_Widow  education_2n Cycle  \
count   2181.000000  …       2181.000000    2181.000000         2181.000000
mean      43.841357  …          0.259055       0.034846            0.088950
std       51.544891  …          0.438217       0.183433            0.284737
min        0.000000  …          0.000000       0.000000            0.000000
25%        9.000000  …          0.000000       0.000000            0.000000
50%       24.000000  …          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      2181.000000           2181.000000       2181.000000    2181.000000
```

```
mean          0.024301                0.503439            0.165979        0.217331
std           0.154017                0.500103            0.372147        0.412525
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  2181.000000      2181.000000         2181.000000
mean    555.581385       511.740028            0.292526
std     568.109267       545.940078            0.668892
min       4.000000      -283.000000            0.000000
25%      55.000000        42.000000            0.000000
50%     341.000000       283.000000            0.000000
75%     956.000000       873.000000            0.000000
max    2262.000000      2145.000000            4.000000

[8 rows x 39 columns]
```

[86]:
```
fig, ax = plt.subplots(1,1, figsize = (6,5))
sns.heatmap(df_food.corr(), cmap='RdYlGn')
```

[86]:
```
<AxesSubplot:>
```



0.3 PCA

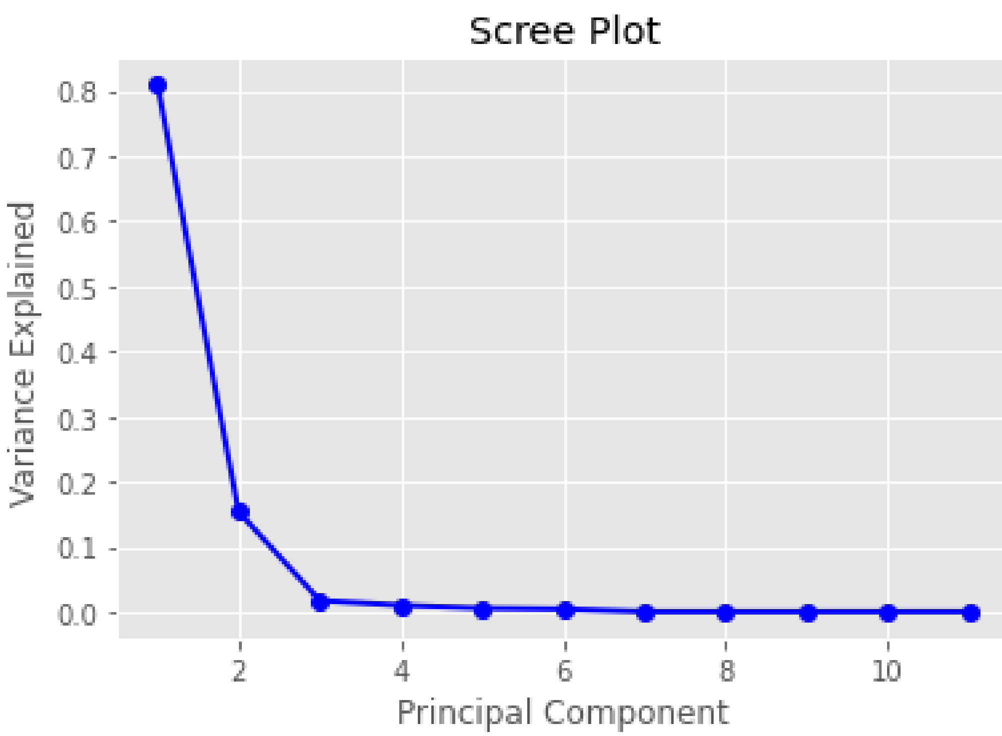
0.4 Pregunta 2

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?

[87]:
```
pca = PCA(n_components=11)
pca_food = pca.fit_transform(df_food2)
print(pca.explained_variance_ratio_)
```

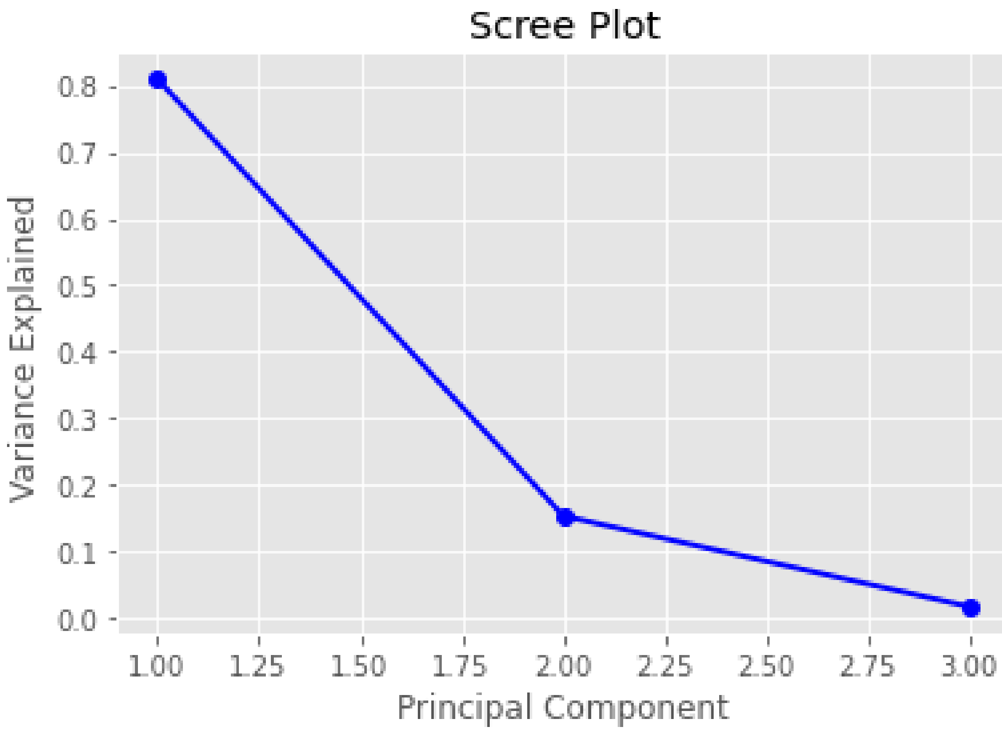
```
[8.09613601e-01 1.53633686e-01 1.72435003e-02 9.80944075e-03
 5.41209224e-03 4.16871385e-03 4.04401296e-05 3.49445475e-05
 1.82800045e-05 1.47332062e-05 1.05674635e-05]
```

[88]:
```python
#scree plot using explained variance proportion

PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2,␣
 ↪color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
plt.show()
```



[89]:
```python
pca = PCA(n_components=3)
pca_features = pca.fit_transform(df_food2)
print(pca.explained_variance_ratio_)

PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2,␣
 ↪color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
plt.show()
```

```
[0.8096136  0.15363369 0.0172435 ]
```



Se realizó un un scree plot con el total de variables utilizadas. En este se observa que hay 3 variables significativas que explican gran parte de la varianza. Se realiza un nuevo PCA con las 3 variables, se grafica y se obtienen los valores propios de 0.8096136, 0.15363369 y 0.0172435

[93]:
```
pca_vectors = pd.DataFrame(data = pca.components_)
pca_vectors.head()
```

[93]:
```
          0         1         2         3         4         5         6  \
0  0.891857  0.051131  0.435924  0.073443  0.052437  0.060932 -0.000156
1 -0.450746  0.093793  0.870967  0.139092  0.091635  0.040584 -0.002899
2 -0.028419  0.321124 -0.206453  0.585234  0.322424  0.637666  0.001508

          7         8         9        10
0  0.004092  0.005627  0.005892 -0.002718
1 -0.001684  0.003838  0.001108 -0.005467
2  0.010642  0.009924  0.011377 -0.005649
```

[94]:
```
pca_df = pd.DataFrame(data=pca_features,columns=["PC1", "PC2", "PC3"])
pca_df.describe().apply(lambda s: s.apply('{0:.3f}'.format))
```

[94]:
```
            PC1       PC2       PC3
count  2181.000  2181.000  2181.000
mean      0.000    -0.000     0.000
```

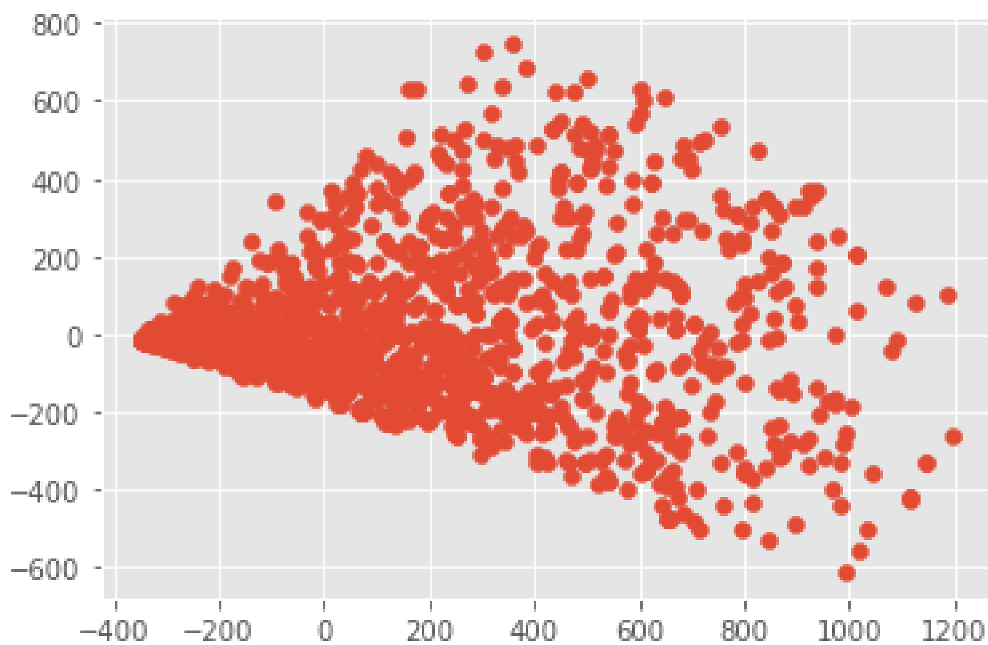
```
std     366.431   159.623    53.477
min    -348.152  -610.364  -192.779
25%    -319.762   -50.157   -24.796
50%    -141.615   -11.246   -14.260
75%     240.733    19.298    15.020
max    1197.561   745.618   261.020
```

[95]:
```
pca_df.corr().apply(lambda s: s.apply('{0:.3f}'.format))
```

[95]:
```
        PC1     PC2     PC3
PC1   1.000  -0.000   0.000
PC2  -0.000   1.000  -0.000
PC3   0.000  -0.000   1.000
```

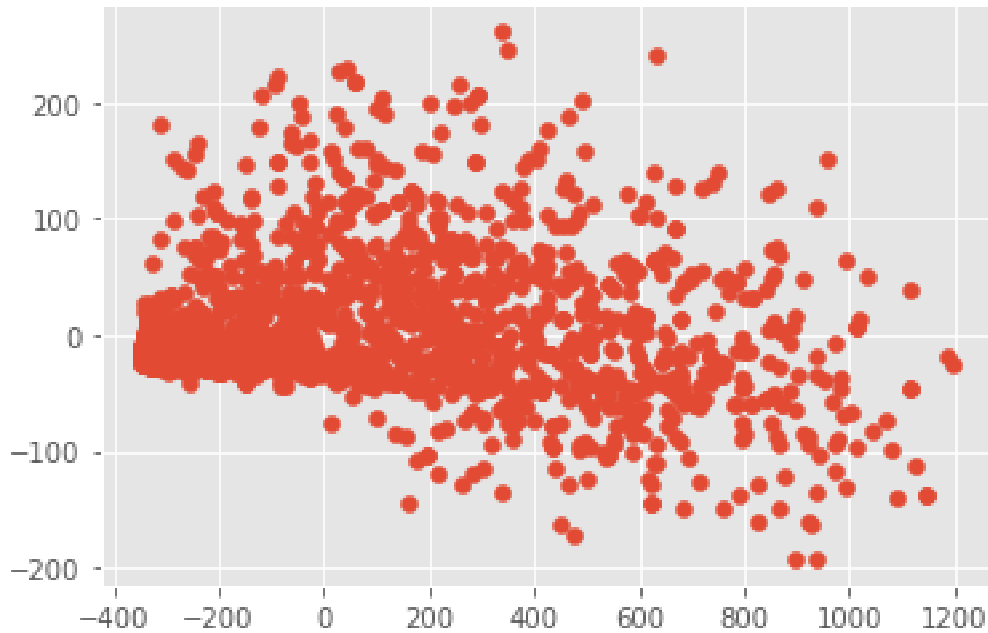
[97]:
```
plt.scatter(pca_df["PC1"],pca_df['PC2'])
```

[97]:
```
<matplotlib.collections.PathCollection at 0x7fb8ed56e7f0>
```



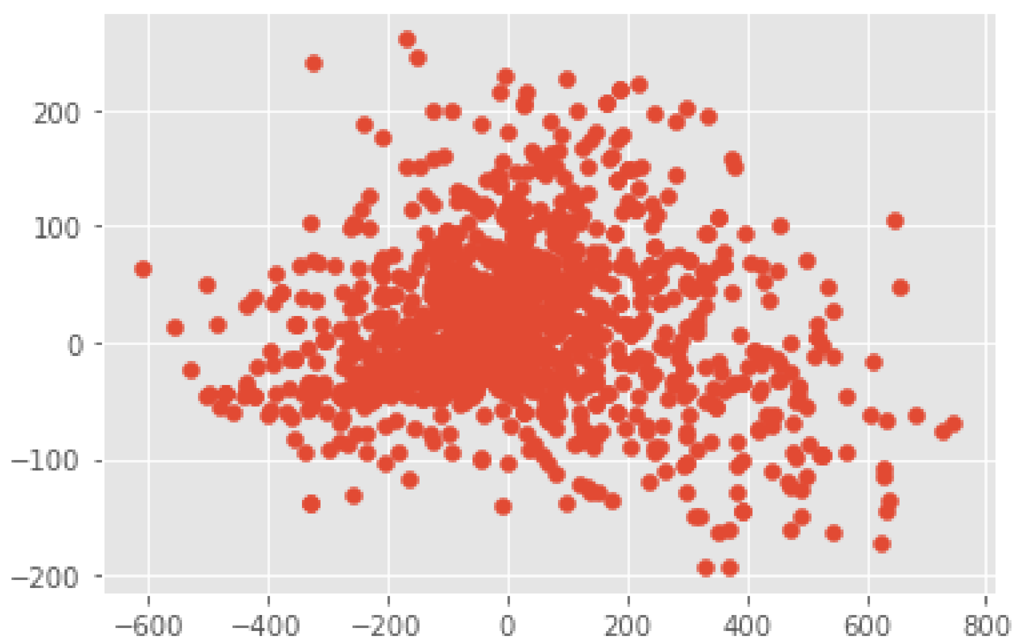
[98]:
```
plt.scatter(pca_df["PC1"],pca_df["PC3"])
```

[98]:
```
<matplotlib.collections.PathCollection at 0x7fb8ed2faa90>
```



[99]:
```
plt.scatter(pca_df['PC2'],pca_df["PC3"])
```

[99]: <matplotlib.collections.PathCollection at 0x7fb907e11550>



Se observa en el gráfico PC1 vs PC2 cierta tendencia horizontal con pendiente negativa. En el gráfico PC1 vs PC3 se observa una tendencia horizontal, mientras que en el gráfico PC2 vs PC3 se observa una dispersión de los datos por lo que no se puede realizar una conclución con certeza.

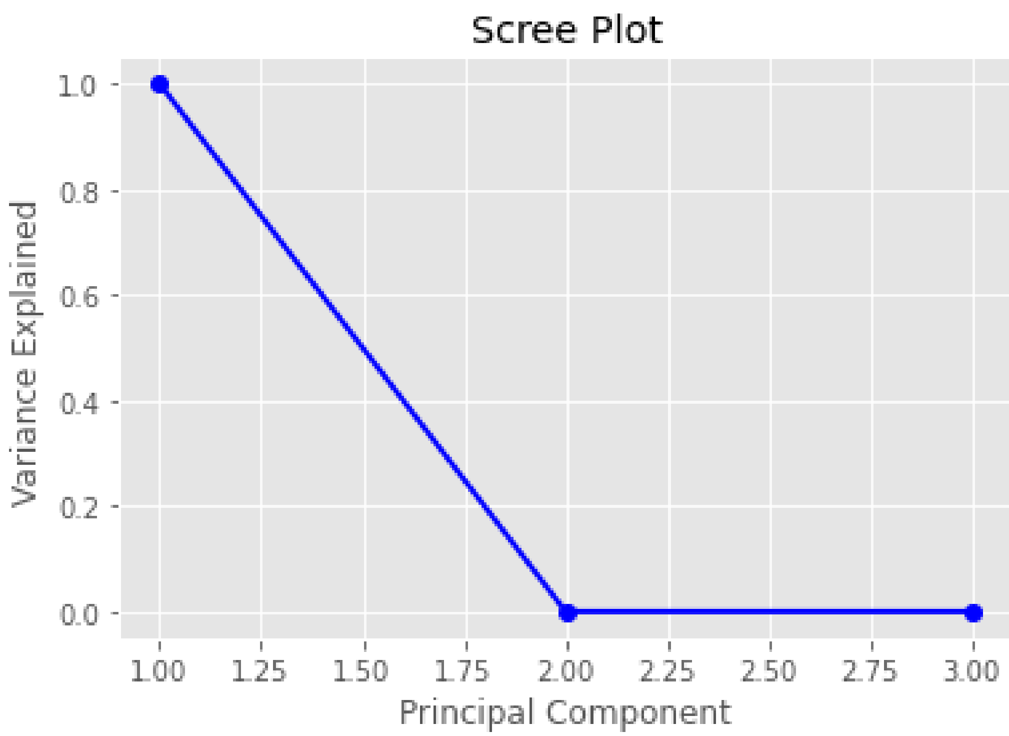
0.5 Pregunta 3

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?

[117]:
```
df_food_2 = df_food[[ "Income", "Kidhome", "education_2n Cycle",␣
↪"education_Basic", "education_Graduation", "education_Master",␣
↪"education_PhD", "Recency"]]
```

[118]:
```
pca = PCA(n_components=3)
pca_features = pca.fit_transform(df_food_2)
print(pca.explained_variance_ratio_)

PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2,␣
↪color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
plt.show()
```

```
[9.99998009e-01 1.98925361e-06 8.08638668e-10]
```



[119]:
```
pca_vectors = pd.DataFrame(data = pca.components_)
pca_vectors.head()
```

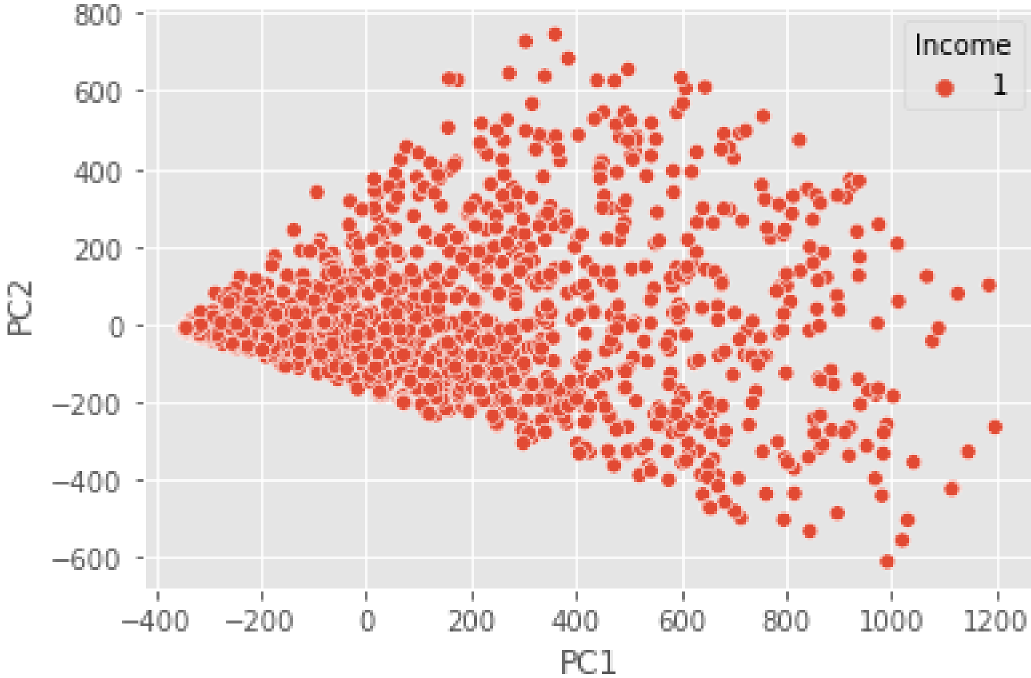
[119]:
```
              0         1             2         3             4             5  \
0  1.000000e+00 -0.000014 -8.036468e-07 -0.000002  3.326685e-07  3.696327e-07
1 -1.161512e-05  0.000298 -3.298870e-05 -0.000024  5.599436e-04 -3.141458e-04
2 -5.898070e-07  0.000043  9.586595e-02  0.022041 -8.404247e-01  2.540342e-01

          6         7
0  0.000002  0.000012
1 -0.000189  1.000000
2  0.468483  0.000642
```

Existe una alta correlación entre PC1 e Income; PC2 y Recency; PC3 y education_PhD

[120]:
```
V1 = "Income";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

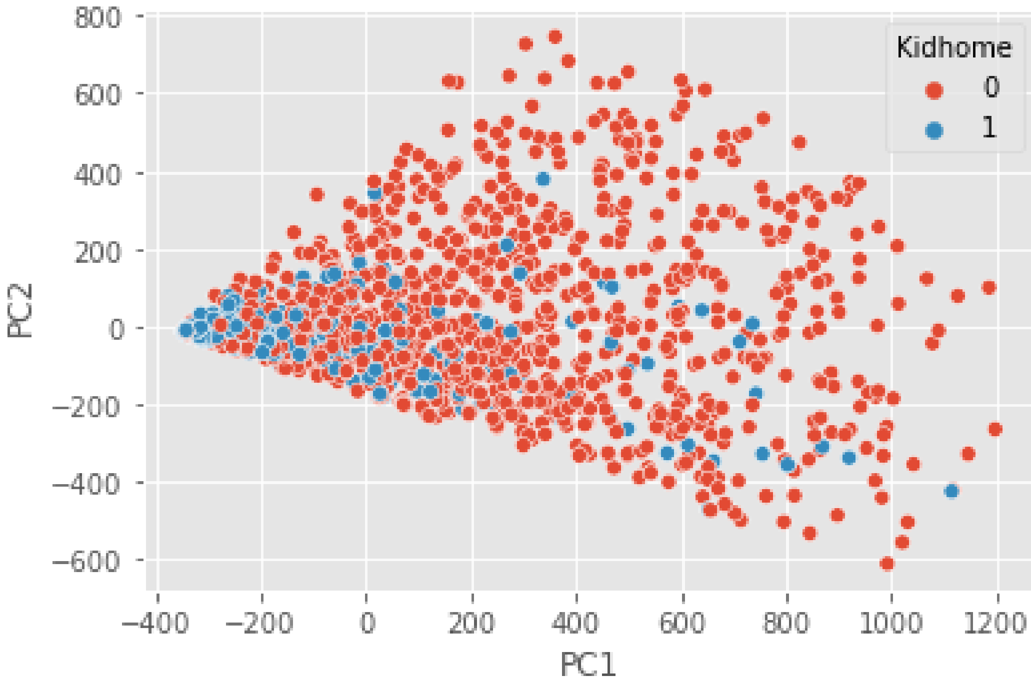
```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
```

```
result in an error or misinterpretation.
  warnings.warn(
```



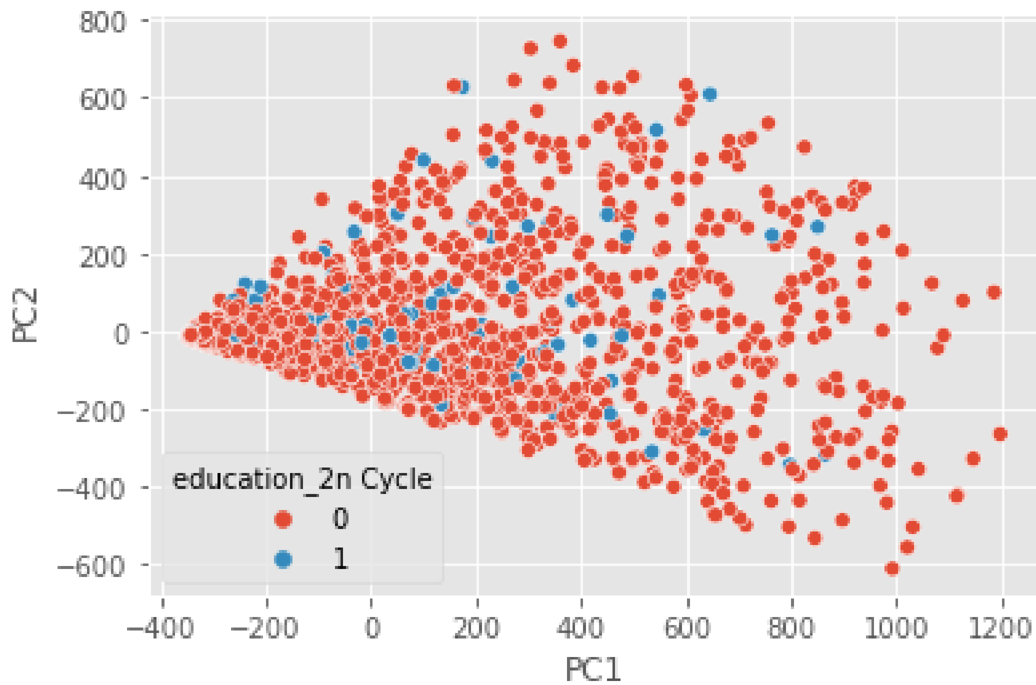
[121]:
```
V1 = "Kidhome";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



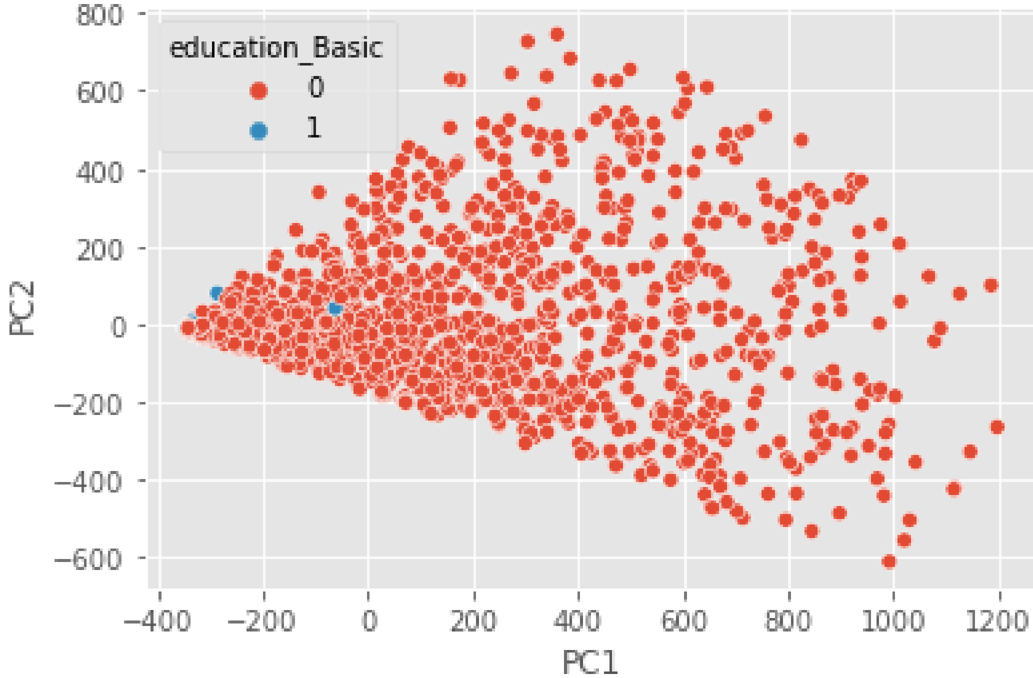
[122]:
```
V1 = "education_2n Cycle";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



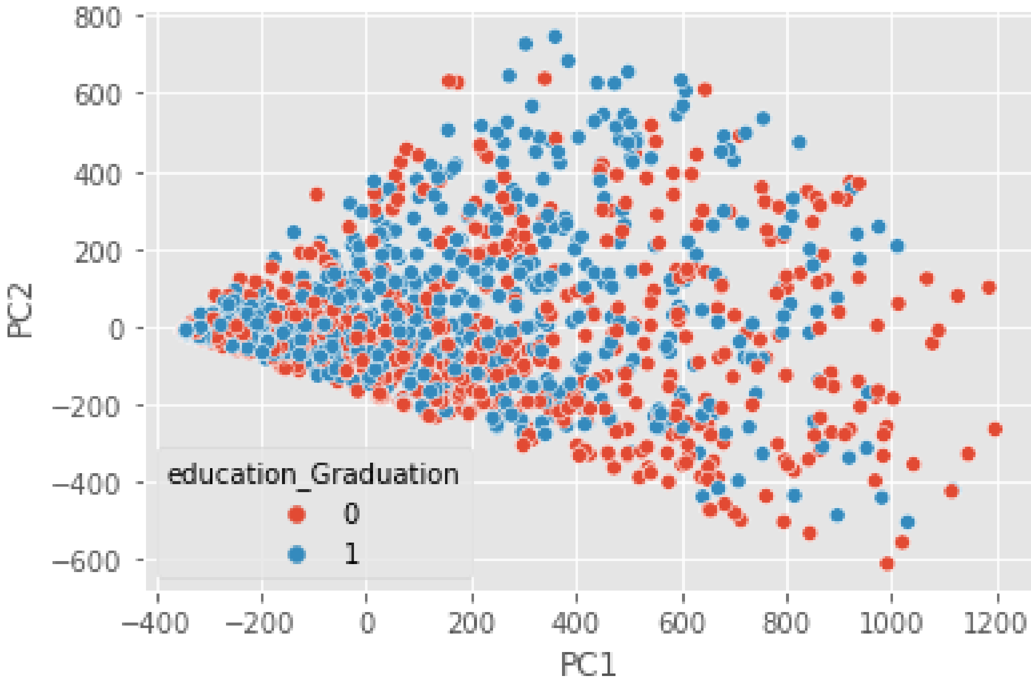

```
[123]: V1 = "education_Basic";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



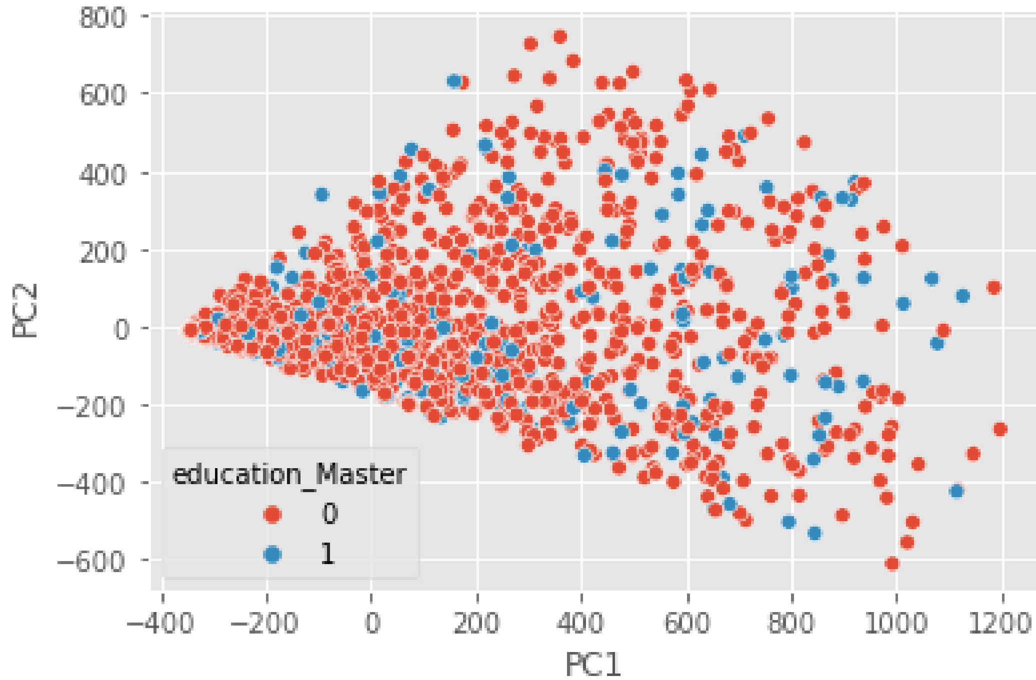
[124]:
```
V1 ="education_Graduation";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



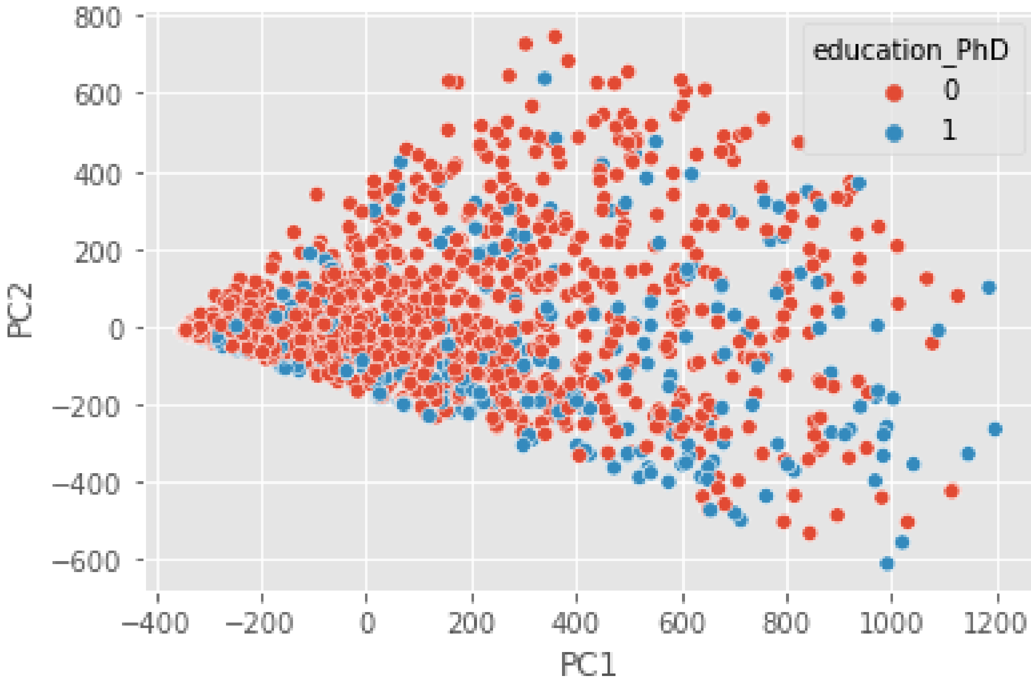
[125]:
```
V1 = "education_Master";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



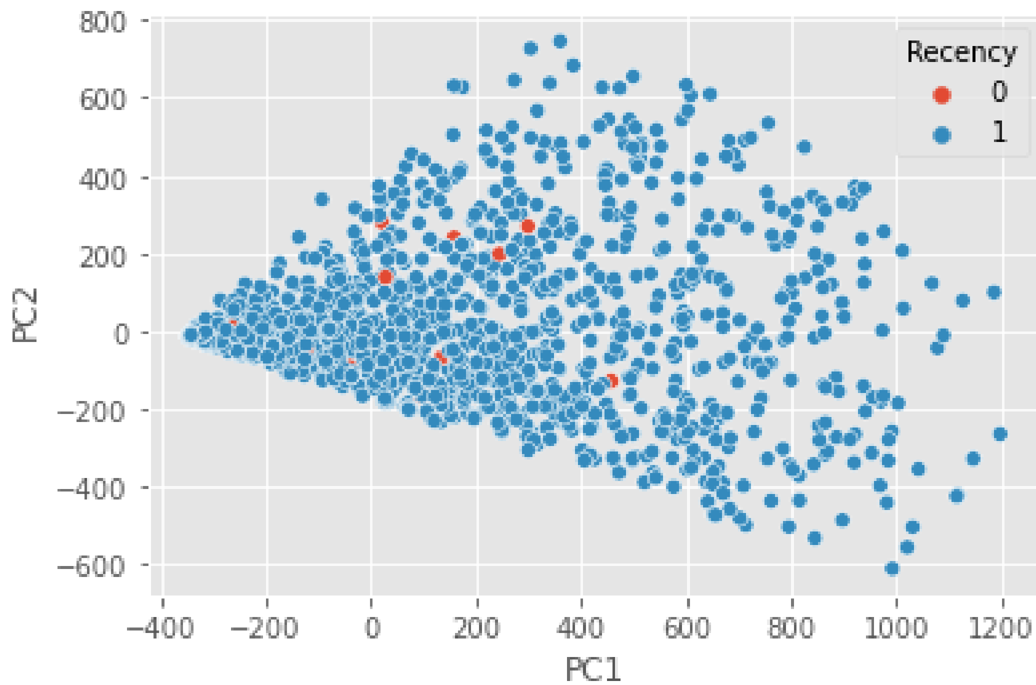
[126]:
```
V1 = "education_PhD";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



[127]:
```
V1 = "Recency";
pca_df[V1] = 0;
pca_df[V1] = np.where(df_food[V1] > 0, 1, pca_df[V1]);
sns.scatterplot("PC1", "PC2", data=pca_df, hue=V1);
```

```
/Users/macbookair/opt/anaconda3/lib/python3.8/site-
packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables
as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```



Para las variables Income, education_2n Cycle, education_Basic no se muestrna diferencias significativas o visbles, por Lo que no se puede asumir que alguno de los dos factores tengan mayor influencia. Para las variables Kidhome, education_Graduation, education_Master, education_PhD se observa una diferencia significativa en el eje horizontal por lo que se puede asumir que el factor PC2 tiene mayor relación.

0.6 EFA

0.7 Pregunta 4

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.

[135]:
```
# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer(rotation='promax')
fa.fit(df_food2)
```

[135]:
```
FactorAnalyzer(rotation_kwargs={})
```

[129]:
```
fa.loadings_
```

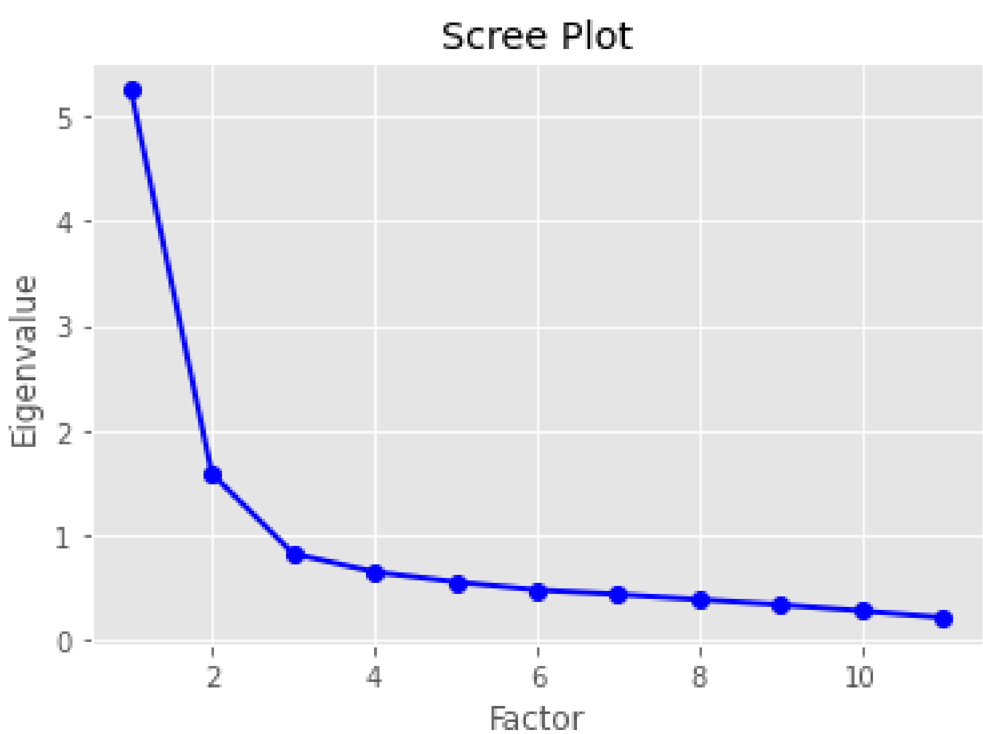
[129]:
```
array([[ 1.07516614, -0.27808647, -0.02330356],
       [-0.08833869,  0.81505812, -0.03128251],
       [ 0.42440732,  0.36619971, -0.24866408],
```

```
       [-0.0941634 ,  0.84901382, -0.04932061],
       [-0.04215036,  0.76001832, -0.02803789],
       [ 0.20249502,  0.42644171,  0.16994652],
       [ 0.09657193, -0.02952742,  0.62094574],
       [ 0.54320024,  0.16808687,  0.46217645],
       [ 0.64679782,  0.20954901, -0.15335873],
       [ 0.64296387,  0.14461263,  0.04811019],
       [-0.29679   , -0.1899502 ,  0.54607509]])
```

[130]:
```
fa.get_eigenvalues()
```

[130]:
```
(array([5.24761553, 1.58768356, 0.82500567, 0.65101447, 0.5547266 ,
        0.47883697, 0.43628435, 0.38710866, 0.33764836, 0.27889466,
        0.21518119]),
 array([ 4.84504462e+00,  1.14420518e+00,  6.14439920e-01,  1.40485497e-01,
         1.08696311e-01,  5.11181298e-02, -2.86317525e-03, -6.33838298e-02,
        -1.53794563e-01, -2.54448799e-01, -3.04960458e-01]))
```

[140]:
```
values = np.arange(1,12)
eigenvalues = pd.DataFrame(data=fa.get_eigenvalues())
plt.plot(values, eigenvalues.loc[0], 'o-', linewidth=2, color='blue')
plt.title('Scree Plot')
plt.xlabel('Factor')
plt.ylabel('Eigenvalue')
plt.show()
```



El número óptimo de valores propios es de dos, ya que se observa en el gráfico solamente dos factores son relevantes.

[141]:
```
fa.get_factor_variance()
```

[141]:
```
(array([2.61978206, 2.48608865, 1.01866813]),
 array([0.23816201, 0.22600806, 0.09260619]),
 array([0.23816201, 0.46417006, 0.55677626]))
```

EFA example using semopy

[142]:
```
print(semopy.efa.explore_cfa_model(df_food2, pval=0.05))
```

```
eta1 =~ MntMeatProducts + MntFishProducts + MntFruits + MntSweetProducts +
MntWines + MntGoldProds
```

Utilizando semopy se obtiene solo un valor propio el cual se compone de distintos factores.

0.8 Latent classes

R package adapted for Python can be used, called stepmix (install with pip)

[2]:
```
from stepmix.stepmix import StepMix

# Continuous StepMix Model with 3 latent classes
model = StepMix(n_components=3, measurement="categorical", verbose=1,␣
 ↪random_state=123)

# Fit model and predict clusters
model.fit(df_food2)
df_food2['pred']=model.predict(df_food2)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-e83a51807554> in <module>
      5
      6 # Fit model and predict clusters
----> 7 model.fit(df_food)
      8 df_food2['pred']=model.predict(df_food2)

NameError: name 'df_food' is not defined
```

0.9 Latent growth

Latent growth modelling is not available on Python at this time. Example available in R for the lavaan library at https://lavaan.ugent.be/tutorial/growth.html

Latent trajectory class (growth curves and class membership) is not available on Python at this time. Example available in R using the LCTMtools library at https://rstudio-pubs-static.s3.amazonaws.com/522393_3aa7f65898f8426e9c0a92d7971b619d.html.

0.10 General CFA

0.11 PREGUNTA 5

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.

[168]:
```
mod = """
# measurement model
Mnt1 =~ MntMeatProducts + MntFishProducts + MntFruits + MntSweetProducts +␣
 ↪MntWines + MntGoldProds
Num1 =~ NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +␣
 ↪NumStorePurchases + NumWebVisitsMonth
    """

model = semopy.Model(mod)
out=model.fit(df_food2)
print(out)
```

```
Name of objective: MLW
Optimization method: SLSQP
Optimization successful.
Optimization terminated successfully
Objective value: 1.666
Number of iterations: 234
Params: 0.448 0.306 0.318 3.155 0.347 -802.935 -1231.049 -1268.667 718.095
1388.794 3.944 5.026 1630.062 22619.162 2.136 55967.191 5.315 2249.190 1335.237
3.499 0.000 -0.123 4972.862
```

[169]:
```
model.inspect(mode='list', what="names", std_est=True)
```

[169]:
```
                 lval  op                  rval     Estimate  Est. Std  \
0     MntMeatProducts   ~                  Mnt1     1.000000  0.424533
1     MntFishProducts   ~                  Mnt1     0.448084  0.616322
2           MntFruits   ~                  Mnt1     0.305613  0.508013
3    MntSweetProducts   ~                  Mnt1     0.317731  0.515273
4            MntWines   ~                  Mnt1     3.155213  0.685109
5         MntGoldProds  ~                  Mnt1     0.347267  0.458806
```

6	NumDealsPurchases	~	Num1	1.000000	0.000864
7	NumWebPurchases	~	Num1	-802.934639	-0.490585
8	NumCatalogPurchases	~	Num1	-1231.048853	-0.806019
9	NumStorePurchases	~	Num1	-1268.666541	-0.674958
10	NumWebVisitsMonth	~	Num1	718.094637	0.504620
11	Num1	~~	Num1	0.000003	1.000000
12	Num1	~~	Mnt1	-0.122759	-1.076875
13	Mnt1	~~	Mnt1	4972.861827	1.000000
14	MntSweetProducts	~~	MntSweetProducts	1388.793530	0.734494
15	NumWebVisitsMonth	~~	NumWebVisitsMonth	3.944282	0.745358
16	NumStorePurchases	~~	NumStorePurchases	5.026366	0.544432
17	MntFishProducts	~~	MntFishProducts	1630.061796	0.620147
18	MntMeatProducts	~~	MntMeatProducts	22619.162300	0.819772
19	NumCatalogPurchases	~~	NumCatalogPurchases	2.135553	0.350333
20	MntWines	~~	MntWines	55967.190822	0.530626
21	NumWebPurchases	~~	NumWebPurchases	5.315315	0.759326
22	MntGoldProds	~~	MntGoldProds	2249.189594	0.789497
23	MntFruits	~~	MntFruits	1335.236849	0.741923
24	NumDealsPurchases	~~	NumDealsPurchases	3.499456	0.999999

	Std. Err	z-value	p-value
0	-	-	-
1	0.024977	17.939968	0.0
2	0.018659	16.378717	0.0
3	0.019257	16.499077	0.0
4	0.168644	18.709316	0.0
5	0.022413	15.494214	0.0
6	-	-	-
7	20992.793033	-0.038248	0.96949
8	32185.843661	-0.038248	0.96949
9	33169.368209	-0.038248	0.96949
10	18774.642528	0.038248	0.96949
11	0.000137	0.019124	0.984742
12	3.209545	-0.038248	0.96949
13	500.407802	9.937618	0.0
14	43.748192	31.745164	0.0
15	0.12455	31.668377	0.0
16	0.170915	29.408488	0.0
17	53.233495	30.62098	0.0
18	700.430928	32.293209	0.0
19	0.091074	23.448504	0.0
20	1917.583421	29.186313	0.0
21	0.167298	31.771515	0.0
22	70.026645	32.119054	0.0
23	41.987765	31.800617	0.0
24	0.105971	33.022717	0.0

[170]:
```
semopy.calc_stats(model)
```

[170]:
```
       DoF  DoF Baseline         chi2  chi2 p-value  chi2 Baseline       CFI  \
Value   43            55  3633.739501           0.0   13645.644069  0.735793

            GFI      AGFI       NFI       TLI     RMSEA        AIC  \
Value  0.733707  0.659393  0.733707  0.662061  0.195717  42.667823

              BIC    LogLik
Value  173.481214  1.666089
```

[47]:
```
semopy.semplot(model, "model.png")
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/execute.py in␣
↪run_check(cmd, input_lines, encoding, quiet, **kwargs)
     80         else:
---> 81             proc = subprocess.run(cmd, **kwargs)
     82     except OSError as e:

~/opt/anaconda3/lib/python3.8/subprocess.py in run(input, capture_output,␣
↪timeout, check, *popenargs, **kwargs)
    492
--> 493     with Popen(*popenargs, **kwargs) as process:
    494         try:

~/opt/anaconda3/lib/python3.8/subprocess.py in __init__(self, args, bufsize,␣
↪executable, stdin, stdout, stderr, preexec_fn, close_fds, shell, cwd, env,␣
↪universal_newlines, startupinfo, creationflags, restore_signals,␣
↪start_new_session, pass_fds, encoding, errors, text)
    857
--> 858             self._execute_child(args, executable, preexec_fn, close_fds

    859                                 pass_fds, cwd, env,

~/opt/anaconda3/lib/python3.8/subprocess.py in _execute_child(self, args,␣
↪executable, preexec_fn, close_fds, pass_fds, cwd, env, startupinfo,␣
↪creationflags, shell, p2cread, p2cwrite, c2pread, c2pwrite, errread, errwrite ␣
↪restore_signals, start_new_session)
   1705                         err_msg = os.strerror(errno_num)
-> 1706                     raise child_exception_type(errno_num, err_msg,␣
↪err_filename)
   1707                 raise child_exception_type(err_msg)

FileNotFoundError: [Errno 2] No such file or directory: PosixPath('dot')

The above exception was the direct cause of the following exception:
```

```
ExecutableNotFound                          Traceback (most recent call last)
<ipython-input-47-65db599301fb> in <module>
----> 1 semopy.semplot(model, "model.png")

~/opt/anaconda3/lib/python3.8/site-packages/semopy/plot.py in semplot(mod,␣
 ↪filename, inspection, plot_covs, plot_exos, images, engine, latshape,␣
 ↪plot_ests, std_ests, show)
    122                     label = str()
    123                 g.edge(rval, lval, label=label, dir='both', style='dashed')
--> 124     g.render(filename, view=show)
    125     return g

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/_tools.py in wrapper(*args,␣
 ↪**kwargs)
    169                               category=category)
    170
--> 171             return func(*args, **kwargs)
    172
    173         return wrapper

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/rendering.py in␣
 ↪render(self, filename, directory, view, cleanup, format, renderer, formatter,␣
 ↪neato_no_op, quiet, quiet_view, outfile, engine, raise_if_result_exists,␣
 ↪overwrite_source)
    120         args.append(filepath)
    121
--> 122         rendered = self._render(*args, **kwargs)
    123
    124         if cleanup:

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/_tools.py in wrapper(*args,␣
 ↪**kwargs)
    169                               category=category)
    170
--> 171             return func(*args, **kwargs)
    172
    173         return wrapper

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/rendering.py in␣
 ↪render(engine, format, filepath, renderer, formatter, neato_no_op, quiet,␣
 ↪outfile, raise_if_result_exists, overwrite_filepath)
    322     cmd += args
    323
--> 324     execute.run_check(cmd,
    325                       cwd=filepath.parent if filepath.parent.parts else␣
 ↪None,
    326                       quiet=quiet,
```

```
~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/execute.py in␣
 ↪run_check(cmd, input_lines, encoding, quiet, **kwargs)
     82     except OSError as e:
     83         if e.errno == errno.ENOENT:
---> 84             raise ExecutableNotFound(cmd) from e
     85         raise
     86 

ExecutableNotFound: failed to execute PosixPath('dot'), make sure the Graphviz␣
 ↪executables are on your systems' PATH
```

No logré que se pudiera ver el modelo en mi computador.

0.12 Complete SEM example

0.13 Pregunta 5

Finalmente, implemente un SEM completo usando la estructura propuesta en la Pregunta 5. En particular, estime un modelo donde los factores explican la variable Response, junto con otras variables demograficas que existen en la base de datos. Ademas utilice dichas variables relevantespara explicar los factores latentes si lo considera apropiado. Las variables a incluir en el modelo final deben tener sustento teorico y el modelo final debe optimizar el ajuste a los datos, en base a los criterios vistos en clase. Que puede concluir en base a sus resultados?

[174]:
```python
import semopy
import pandas as pd
desc = semopy.examples.political_democracy.get_model()
print(desc)
```

```
# measurement model
ind60 =~ x1 + x2 + x3
dem60 =~ y1 + y2 + y3 + y4
dem65 =~ y5 + y6 + y7 + y8
# regressions
dem60 ~ ind60
dem65 ~ ind60 + dem60
# residual correlations
y1 ~~ y5
y2 ~~ y4 + y6
y3 ~~ y7
y4 ~~ y8
y6 ~~ y8
```

[173]:
```python
mod = """
# measurement model
Mnt1 =~ MntMeatProducts + MntFishProducts + MntFruits + MntSweetProducts +␣
 ↪MntWines + MntGoldProds
```

```
Num1 =~ NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +␣
↪NumStorePurchases + NumWebVisitsMonth

# regressions
Response ~ Mnt1 + Num1 + Income + Kidhome + education_2n Cycle +␣
↪education_Basic + education_Graduation + education_Master + education_PhD +␣
↪Recency

# residual correlations

    """

model = semopy.Model(mod)
out=model.fit(df_food)
print(out)
```

```
Traceback (most recent call last):

  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/semopy/
↪parser.py", line 189, in parse_desc
    kind, items = separate_token(line)

  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/semopy/
↪parser.py", line 57, in separate_token
    raise SyntaxError(f'Invalind syntax for line:\n{token}')

  File "<string>", line unknown
SyntaxError: Invalind syntax for line:
Response ~ Mnt1 + Num1 + Kidhome + education_2n Cycle + education_Basic +␣
↪education_Graduation + education_Master + education_PhD + Recency


During handling of the above exception, another exception occurred:

Traceback (most recent call last):

  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/IPython/core/
↪interactiveshell.py", line 3437, in run_code
    exec(code_obj, self.user_global_ns, self.user_ns)

  File "<ipython-input-173-847abba67dd2>", line 12, in <module>
    model = semopy.Model(mod)

  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/semopy/model.
↪py", line 105, in __init__
    super().__init__(description)
```

```
  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/semopy/
↪model_base.py", line 53, in __init__
    effects, operations = parse_desc(description)

  File "/Users/macbookair/opt/anaconda3/lib/python3.8/site-packages/semopy/
↪parser.py", line 203, in parse_desc
    raise SyntaxError(f"Syntax error for line:\n{line}")

  File "<string>", line unknown
SyntaxError: Syntax error for line:
Response ~ Mnt1 + Num1 + Kidhome + education_2n Cycle + education_Basic +␣
↪education_Graduation + education_Master + education_PhD + Recency
```

[175]:
```
data = semopy.examples.political_democracy.get_data()
mod = semopy.Model(desc)
res = mod.fit(data)
```

[178]:
```
print(mod.inspect())
```

```
     lval  op   rval  Estimate  Std. Err   z-value   p-value
0   dem60   ~  ind60  1.482379  0.399024  3.715017  0.000203
1   dem65   ~  ind60  0.571912  0.221383  2.583364  0.009784
2   dem65   ~  dem60  0.837574  0.098446  8.507992       0.0
3      x1   ~  ind60  1.000000         -         -         -
4      x2   ~  ind60  2.180494  0.138565 15.736254       0.0
5      x3   ~  ind60  1.818546  0.151993  11.96465       0.0
6      y1   ~  dem60  1.000000         -         -         -
7      y2   ~  dem60  1.256819  0.182687  6.879647       0.0
8      y3   ~  dem60  1.058174  0.151521  6.983699       0.0
9      y4   ~  dem60  1.265186  0.145151  8.716344       0.0
10     y5   ~  dem65  1.000000         -         -         -
11     y6   ~  dem65  1.185743  0.168908  7.020032       0.0
12     y7   ~  dem65  1.279717  0.159996   7.99841       0.0
13     y8   ~  dem65  1.266084  0.158238  8.001141       0.0
14  dem60  ~~  dem60  3.950849  0.920451  4.292296  0.000018
15  dem65  ~~  dem65  0.172210  0.214861  0.801494  0.422846
16  ind60  ~~  ind60  0.448321  0.086677  5.172345       0.0
17     y1  ~~     y5  0.624423  0.358435  1.742083  0.081494
18     y1  ~~     y1  1.892743   0.44456  4.257565  0.000021
19     y2  ~~     y4  1.319589   0.70268  1.877937   0.06039
20     y2  ~~     y6  2.156164  0.734155  2.936934  0.003315
21     y2  ~~     y2  7.385292  1.375671  5.368501       0.0
22     y3  ~~     y7  0.793329  0.607642  1.305585  0.191694
23     y3  ~~     y3  5.066628  0.951722  5.323646       0.0
24     y4  ~~     y8  0.347222  0.442234  0.785154  0.432363
25     y4  ~~     y4  3.147911  0.738841  4.260605   0.00002
```

26	y6	~~	y8	1.357037	0.5685	2.387047	0.016984
27	y6	~~	y6	4.954364	0.914284	5.418843	0.0
28	x2	~~	x2	0.119894	0.069747	1.718973	0.085619
29	y5	~~	y5	2.351910	0.480369	4.896044	0.000001
30	x1	~~	x1	0.081573	0.019495	4.184317	0.000029
31	x3	~~	x3	0.466732	0.090168	5.176276	0.0
32	y8	~~	y8	3.256389	0.69504	4.685182	0.000003
33	y7	~~	y7	3.430032	0.712732	4.812512	0.000001

[51]:
```
semopy.semplot(mod, "semmodel.png")
```

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/execute.py in␣
 ↪run_check(cmd, input_lines, encoding, quiet, **kwargs)
     80         else:
---> 81             proc = subprocess.run(cmd, **kwargs)
     82     except OSError as e:

~/opt/anaconda3/lib/python3.8/subprocess.py in run(input, capture_output,␣
 ↪timeout, check, *popenargs, **kwargs)
    492
--> 493     with Popen(*popenargs, **kwargs) as process:
    494         try:

~/opt/anaconda3/lib/python3.8/subprocess.py in __init__(self, args, bufsize,␣
 ↪executable, stdin, stdout, stderr, preexec_fn, close_fds, shell, cwd, env,␣
 ↪universal_newlines, startupinfo, creationflags, restore_signals,␣
 ↪start_new_session, pass_fds, encoding, errors, text)
    857
--> 858             self._execute_child(args, executable, preexec_fn, close_fds

    859                                 pass_fds, cwd, env,

~/opt/anaconda3/lib/python3.8/subprocess.py in _execute_child(self, args,␣
 ↪executable, preexec_fn, close_fds, pass_fds, cwd, env, startupinfo,␣
 ↪creationflags, shell, p2cread, p2cwrite, c2pread, c2pwrite, errread, errwrite ␣
 ↪restore_signals, start_new_session)
   1705                         err_msg = os.strerror(errno_num)
-> 1706                     raise child_exception_type(errno_num, err_msg,␣
 ↪err_filename)
   1707                 raise child_exception_type(err_msg)

FileNotFoundError: [Errno 2] No such file or directory: PosixPath('dot')

The above exception was the direct cause of the following exception:

ExecutableNotFound                        Traceback (most recent call last)
<ipython-input-51-b1f1c8bae510> in <module>
```

```
----> 1 semopy.semplot(mod, "semmodel.png")

~/opt/anaconda3/lib/python3.8/site-packages/semopy/plot.py in semplot(mod,␣
↪filename, inspection, plot_covs, plot_exos, images, engine, latshape,␣
↪plot_ests, std_ests, show)
    122                 label = str()
    123             g.edge(rval, lval, label=label, dir='both', style='dashed')
--> 124     g.render(filename, view=show)
    125     return g

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/_tools.py in wrapper(*args␣
↪**kwargs)
    169                                   category=category)
    170
--> 171             return func(*args, **kwargs)
    172
    173         return wrapper

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/rendering.py in␣
↪render(self, filename, directory, view, cleanup, format, renderer, formatter,␣
↪neato_no_op, quiet, quiet_view, outfile, engine, raise_if_result_exists,␣
↪overwrite_source)
    120         args.append(filepath)
    121
--> 122         rendered = self._render(*args, **kwargs)
    123
    124         if cleanup:

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/_tools.py in wrapper(*args␣
↪**kwargs)
    169                                   category=category)
    170
--> 171             return func(*args, **kwargs)
    172
    173         return wrapper

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/rendering.py in␣
↪render(engine, format, filepath, renderer, formatter, neato_no_op, quiet,␣
↪outfile, raise_if_result_exists, overwrite_filepath)
    322     cmd += args
    323
--> 324     execute.run_check(cmd,
    325                       cwd=filepath.parent if filepath.parent.parts else␣
↪None,
    326                       quiet=quiet,

~/opt/anaconda3/lib/python3.8/site-packages/graphviz/backend/execute.py in␣
↪run_check(cmd, input_lines, encoding, quiet, **kwargs)
```

```
     82     except OSError as e:
     83         if e.errno == errno.ENOENT:
---> 84             raise ExecutableNotFound(cmd) from e
     85         raise
     86 

ExecutableNotFound: failed to execute PosixPath('dot'), make sure the Graphviz␣
  ↪executables are on your systems' PATH
```

Tarea 3

Instrucciones

Los resultados de los ejericicios propuestos se deben entregar como un notebook por correo electronico a *juancaros@udec.cl* el dia 9/6 hasta las 21:00. Es importante considerar que el código debe poder ejecutarse en cualquier computadora con la data original del repositorio. Recordar la convencion para el nombre de archivo ademas de incluir en su documento titulos y encabezados por seccion. Utilizar la base de datos *ifood_df.csv*.

Como se indica en la Tabla 1, las variables describen el comportamiento de un set de consumidores en una tienda de retail. Las variables categoricas (e.g. educacion, estado civil) ya han sido convertidas a variables binarias (una por cada categoria).

[8]:
```
Image(filename='../data/dictionary.png', width=600)
```

[8]:

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment with the company
Education	customer's level of education
Marital	customer's marital status
Kidhome	number of small children in customer's household
Teenhome	number of teenagers in customer's household
Income	customer's yearly household income
MntFishProducts	amount spent on fish products in the last 2 years
MntMeatProducts	amount spent on meat products in the last 2 years
MntFruits	amount spent on fruits in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines in the last 2 years
MntGoldProds	amount spent on *gold* products in the last 2 years
NumDealsPurchases	number of purchases made with discount
NumCatalogPurchases	number of purchases made using catalogue
NumStorePurchases	number of purchases made directly in stores
NumWebPurchases	number of purchases made through company's web site
NumWebVisitsMonth	number of visits to company's web site in the last month
Recency	number of days since the last purchase

Table 1: Meta-data table

1 Anexo 1: analisis de la información

- AcceptedCmp1 = 1 si el cliente aceptó la oferta en la 1ra campaña, 0 de lo contrario
- AcceptedCmp2 = 1 si el cliente aceptó la oferta en la 2da campaña, 0 de lo contrario
- AcceptedCmp3 = 1 si el cliente aceptó la oferta en la 3ra campaña, 0 de lo contrario
- AcceptedCmp4 = 1 si el cliente aceptó la oferta en la 4ta campaña, 0 de lo contrario
- AcceptedCmp5 = 1 si el cliente aceptó la oferta en la 5ta campaña, 0 de lo contrario
- Response (target) = 1 si el cliente aceptó la oferta en la última campaña, 0 de lo contrario
- Complain = 1 si el cliente se quejó en los últimos 2 años
- DtCustomer = fecha de inscripción del cliente en la empresa
- Education = nivel de educación del cliente
- Marital = estado civil
- Kidhome = número de niños pequeños en el hogar
- Teenhome = número de adolescentes en el hogar
- Income = ingresos familiares anuales
- MntFishProducts = cantidad gastada en pescados (productos del mar) en los últimos 2 años
- MntMeatProducts = cantidad gastada en carne en los últimos 2 años
- MntFruits = cantidad gastada en frutas en los últimos 2 años

- MntSweetProducts = cantidad gastada en productos dulces en los últimos 2 años
- MntWines = cantidad gastada en vino en los últimos 2 años
- MntGoldProds = cantidad gastada en productos de “lujo” en los últimos 2 años
- NumDealsPurchases = número de compras realizadas con descuento
- NumCatalogPurchases = número de compras realizadas utilizando el catálogo
- NumStorePurchases = número de compras realizadas directamente en las tiendas
- NumWebPurchases = número de compras realizadas a través del sitio web de la empresa
- NumWebVisitsMonth = número de visitas al sitio web de la empresa en el último mes
- Recency = número de días desde la última compra

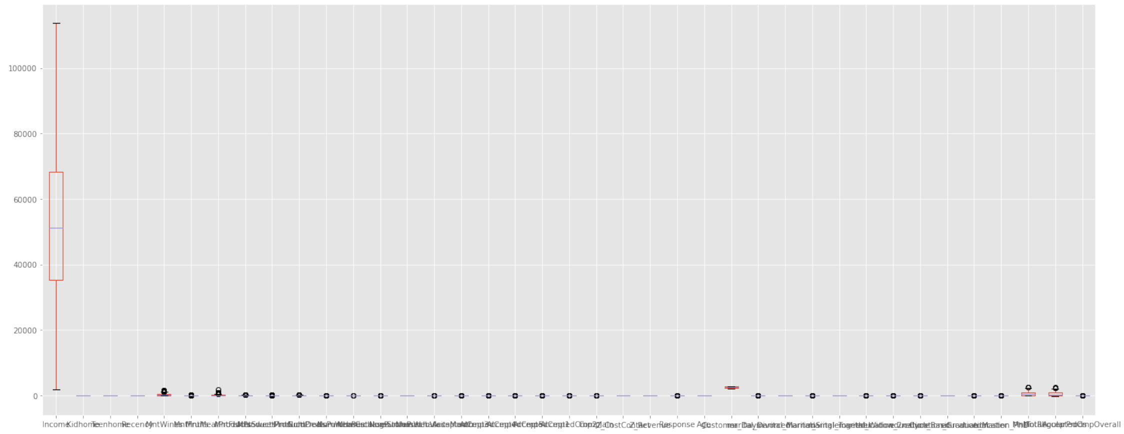
[71]:
```python
# Check for missing data
print("Valores de Null en el Dataframe:")
print(df_food.isnull().sum())
```

```
Valores de Null en el Dataframe:
Income                  0
Kidhome                 0
Teenhome                0
Recency                 0
MntWines                0
MntFruits               0
MntMeatProducts         0
MntFishProducts         0
MntSweetProducts        0
MntGoldProds            0
NumDealsPurchases       0
NumWebPurchases         0
NumCatalogPurchases     0
NumStorePurchases       0
NumWebVisitsMonth       0
AcceptedCmp3            0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Z_CostContact           0
Z_Revenue               0
Response                0
Age                     0
Customer_Days           0
marital_Divorced        0
marital_Married         0
marital_Single          0
marital_Together        0
marital_Widow           0
education_2n Cycle      0
education_Basic         0
```

```
education_Graduation    0
education_Master        0
education_PhD           0
MntTotal                0
MntRegularProds         0
AcceptedCmpOverall      0
dtype: int64
```

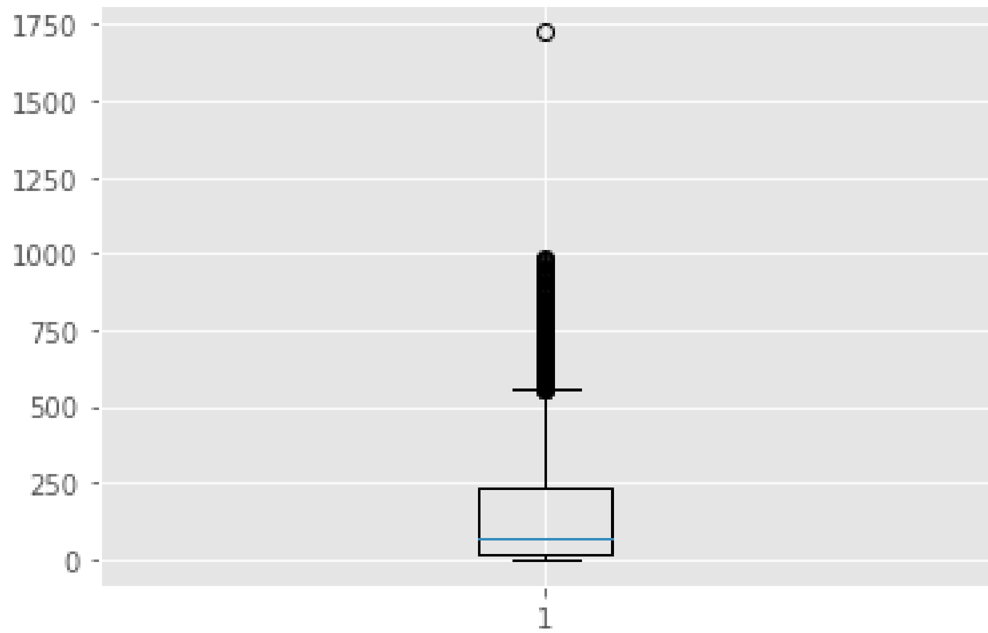
[72]:

```python
# Create a figure and axis
fig, ax = plt.subplots(figsize=(25, 10))

# Plot the box plots for all columns
df_food.boxplot(ax=ax)

# Show the plot
plt.show()
```



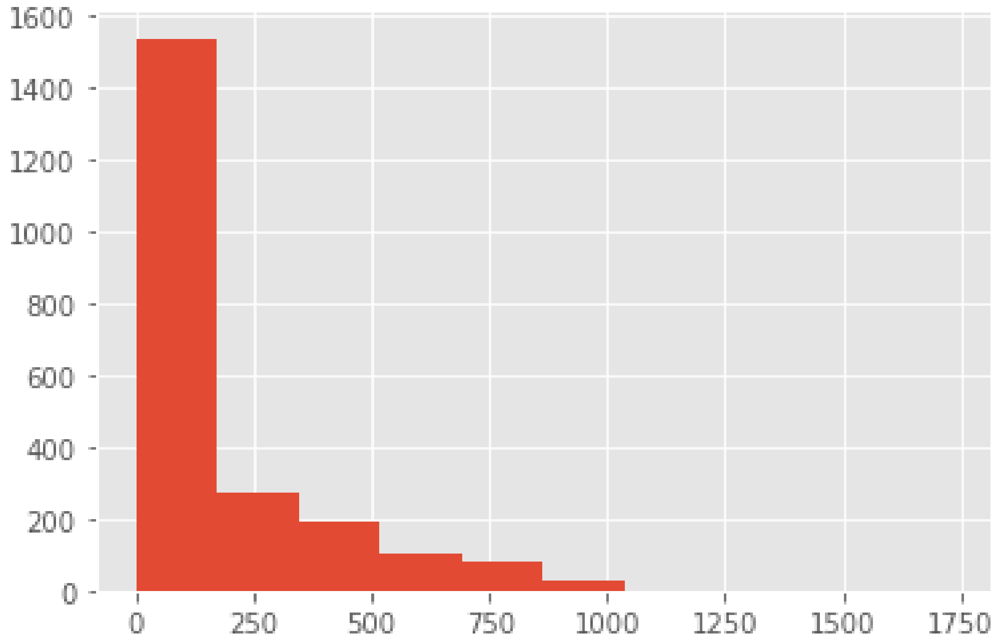
[73]:

```python
plt.boxplot(df_food['MntMeatProducts'])
plt.show()
```



[74]:
```
A="MntMeatProducts"
plt.style.use('ggplot')
df_food['MntMeatProducts'].hist();
df_food.value_counts(A)
```

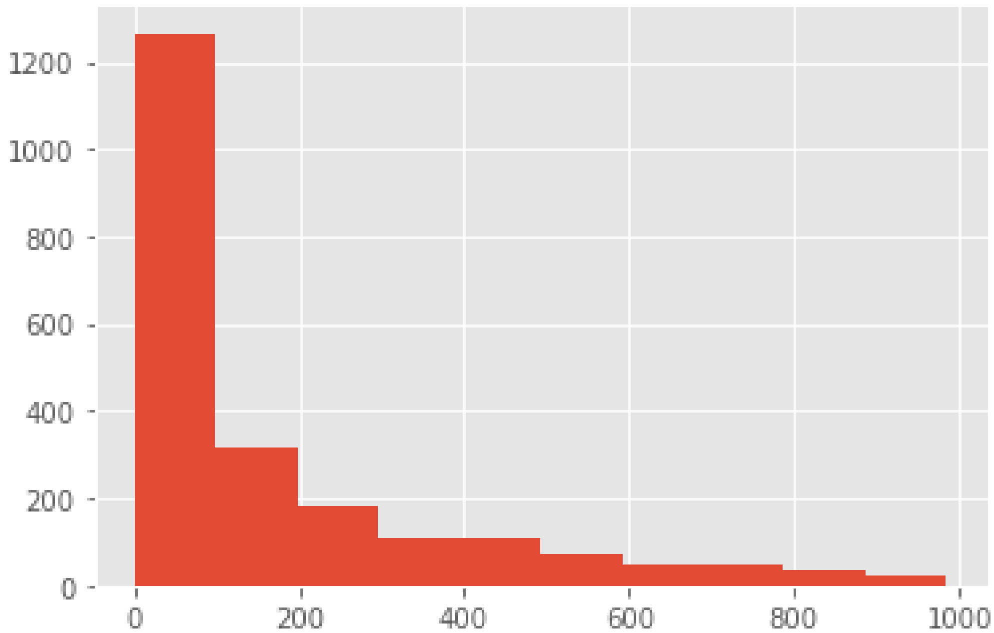
[74]:
```
MntMeatProducts
7       53
5       49
11      49
8       44
6       42
        ..
444      1
450      1
452      1
454      1
1725     1
Length: 551, dtype: int64
```



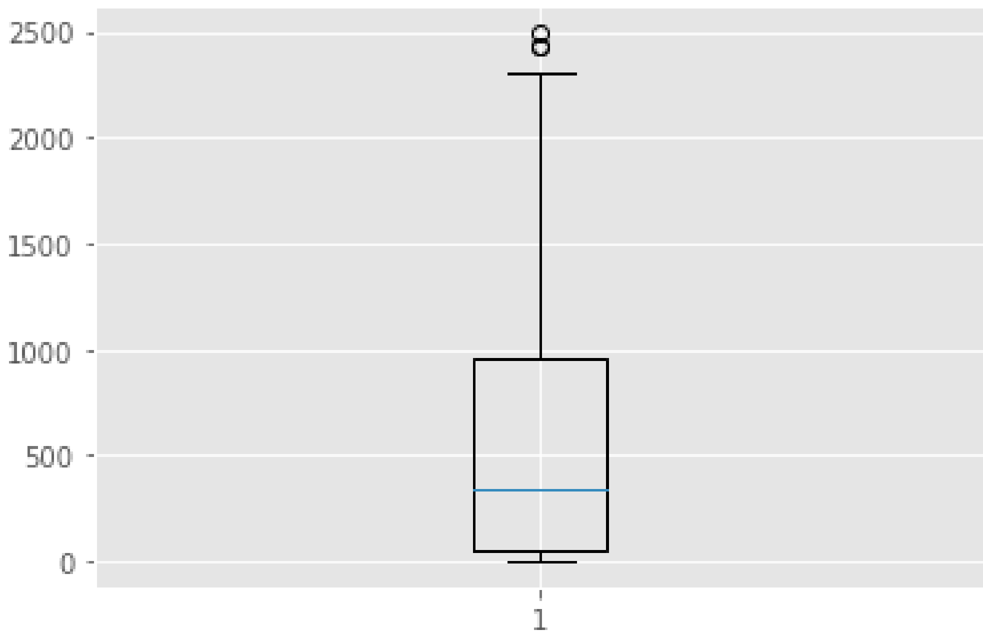
[75]:
```
df_food = df_food[(df_food["MntMeatProducts"] <1100) &␣
↪(df_food["MntMeatProducts"] > 0)];
print(len(df_food));
A="MntMeatProducts"
plt.style.use('ggplot')
df_food['MntMeatProducts'].hist();
df_food.value_counts(A)
```

```
2203
```

[75]:
```
MntMeatProducts
7      53
5      49
11     49
8      44
6      42
       ..
444     1
450     1
452     1
454     1
984     1
Length: 549, dtype: int64
```



[77]:
```
plt.boxplot(df_food['MntTotal'])
plt.show()
```

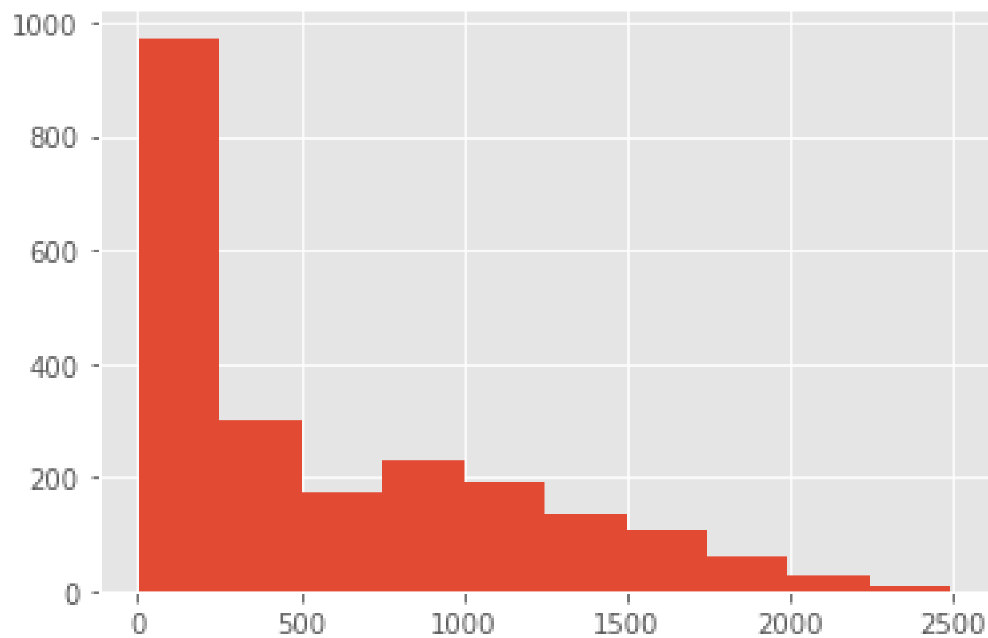


[78]:

```
A="MntTotal"
plt.style.use('ggplot')
df_food['MntTotal'].hist();
df_food.value_counts(A)
```

[78]:

```
MntTotal
39      30
41      25
19      24
40      24
16      24
        ..
841      1
839      1
836      1
826      1
2491     1
Length: 896, dtype: int64
```



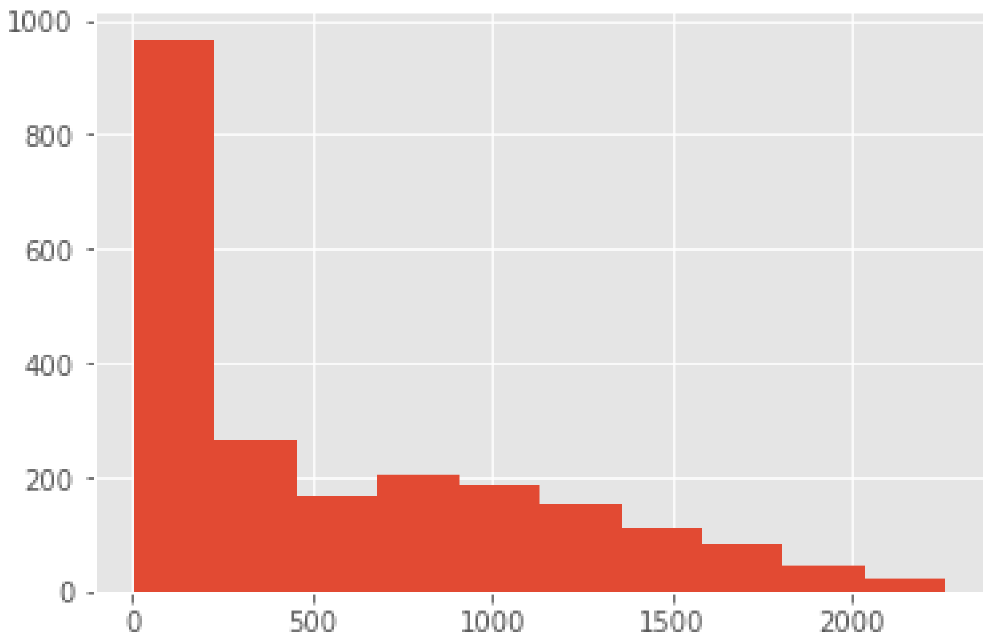
[79]:

```
df_food = df_food[(df_food["MntTotal"] <2300) & (df_food["MntTotal"] > 0)];
print(len(df_food));
A="MntTotal"
plt.style.use('ggplot')
df_food['MntTotal'].hist();
df_food.value_counts(A)
```

```
2198
```

[79]:

```
MntTotal
39      30
41      25
40      24
19      24
16      24
        ..
841      1
839      1
836      1
826      1
2262     1
Length: 893, dtype: int64
```



[80]:

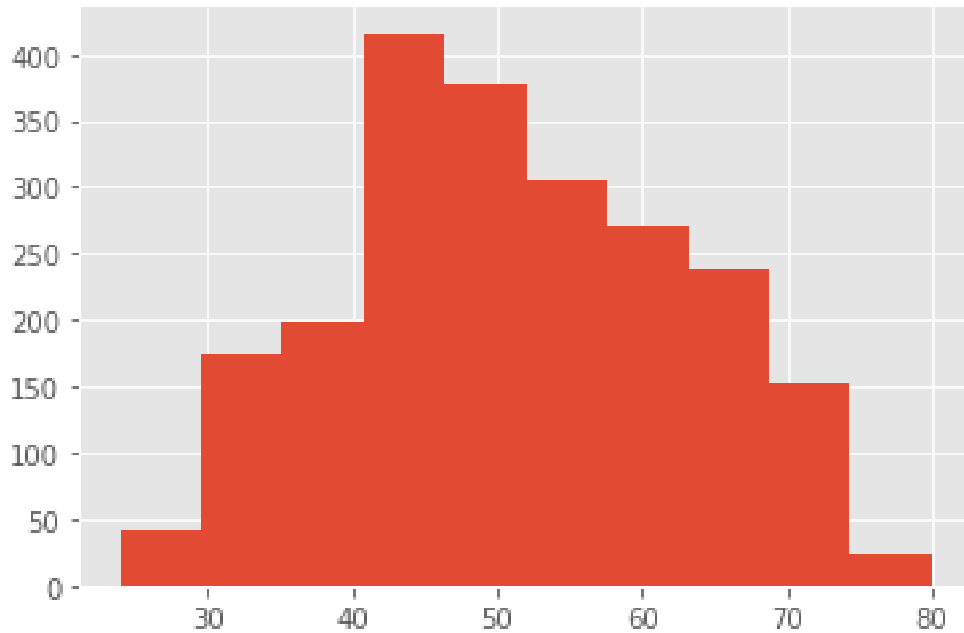
```
A="Age"
plt.style.use('ggplot')
df_food['Age'].hist();
df_food.value_counts(A)
```

[80]:

```
Age
44    88
49    85
45    81
48    78
```

42	76
50	74
55	74
47	71
51	69
46	69
64	55
41	52
68	52
62	52
52	51
43	50
61	50
54	50
66	49
60	49
65	48
58	44
57	44
53	44
38	43
69	42
56	41
37	41
34	41
63	41
40	39
39	38
36	38
59	35
67	35
35	32
31	29
71	29
70	29
32	28
33	27
72	21
30	18
74	16
73	16
29	13
28	13
75	8
76	7
77	6
25	5

```
27      5
26      3
24      2
79      1
80      1
dtype: int64
```

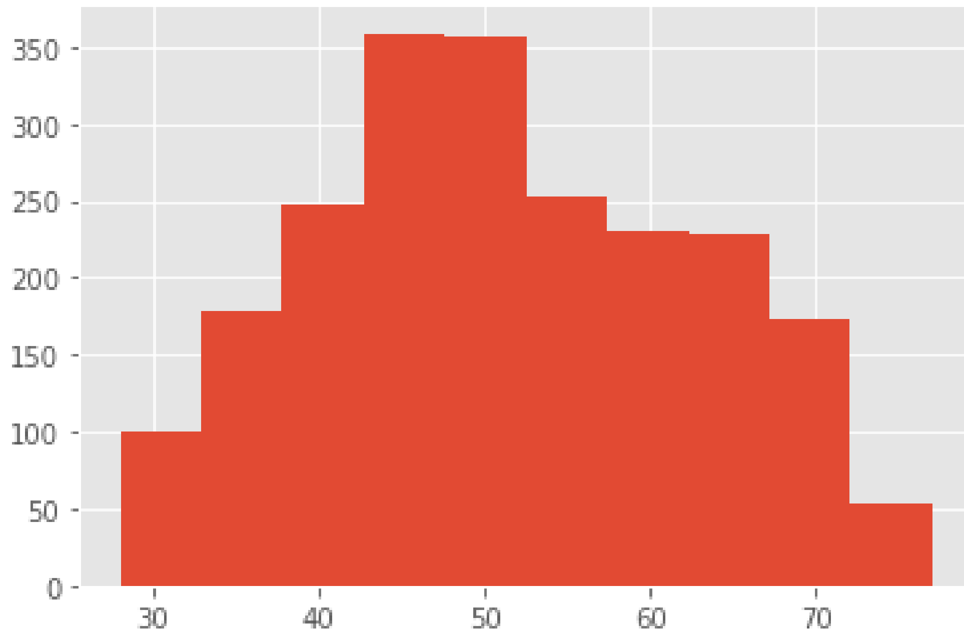


[81]:
```
df_food = df_food[(df_food["Age"] < 78) & (df_food["Age"] > 27)];
print(len(df_food));
A="Age"
plt.style.use('ggplot')
df_food['Age'].hist();
df_food.value_counts(A)
```

```
2181
```

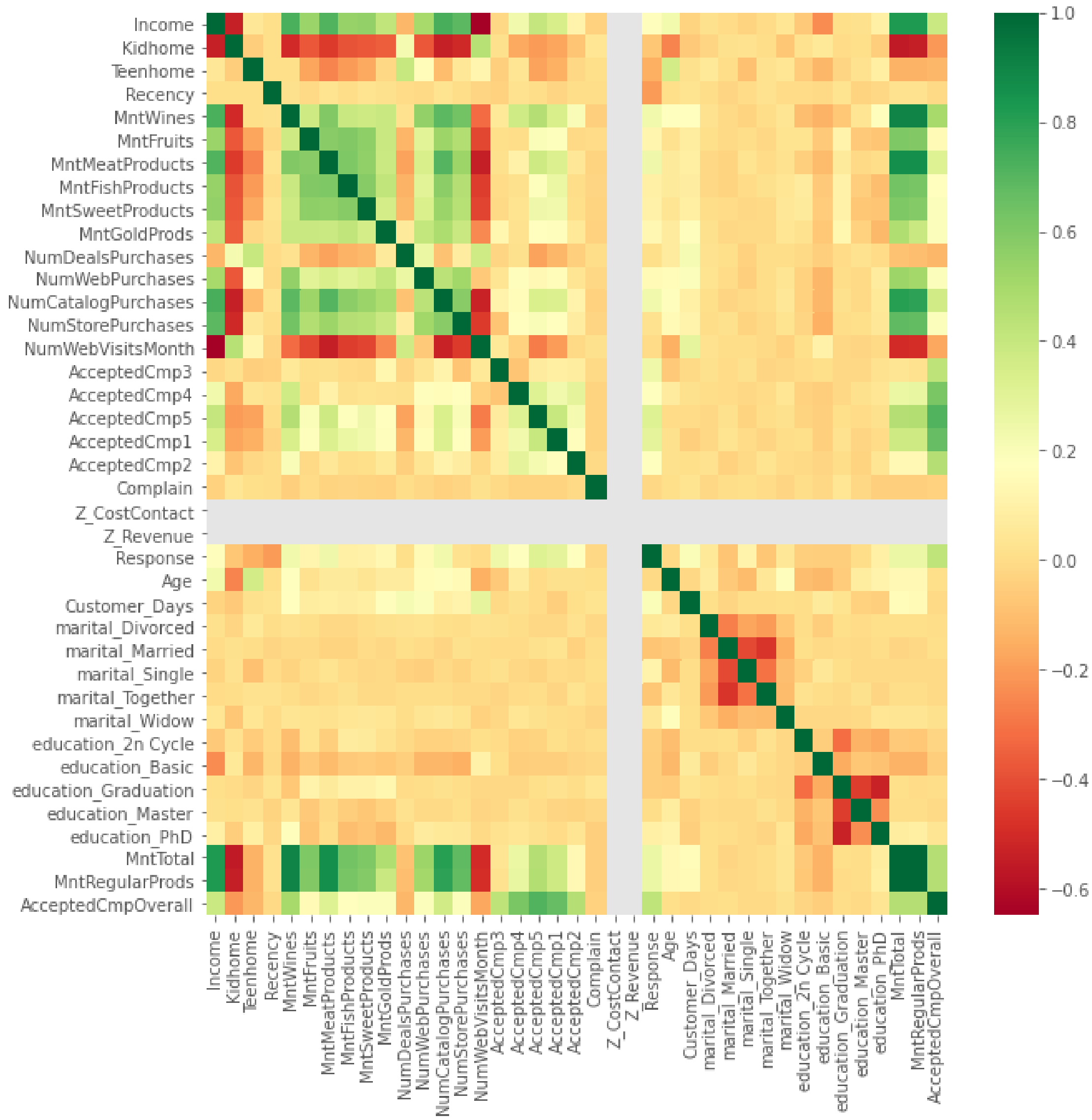
[81]:
```
Age
44     88
49     85
45     81
48     78
42     76
50     74
55     74
47     71
51     69
```

```
46    69
64    55
41    52
62    52
68    52
52    51
61    50
54    50
43    50
60    49
66    49
65    48
58    44
57    44
53    44
38    43
69    42
34    41
63    41
56    41
37    41
40    39
36    38
39    38
59    35
67    35
35    32
71    29
31    29
70    29
32    28
33    27
72    21
30    18
74    16
73    16
28    13
29    13
75     8
76     7
77     6
dtype: int64
```



[82]:
```
fig, ax = plt.subplots(1,1, figsize = (10,10))
sns.heatmap(df_food.corr(), cmap='RdYlGn')
```

[82]: <AxesSubplot:>



Income
Kidhome
Teenhome
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
AcceptedCmp3
AcceptedCmp4
AcceptedCmp5
AcceptedCmp1
AcceptedCmp2
Complain
Z_CostContact
Z_Revenue
Response
Age
Customer_Days
marital_Divorced
marital_Married
marital_Single
marital_Together
marital_Widow
education_2n Cycle
education_Basic
education_Graduation
education_Master
education_PhD
MntTotal
MntRegularProds
AcceptedCmpOverall
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


```
[ ]:
```