ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"»

ANALYSIS OF THE HOUSEHOLDS INCOME/ EXPENDITURE SURVEY IN QUITO

Homework Project "2018/2019"

"Lottery" team:
Freire Rubén

MSc Program "Data Science"
1st Year

Moscow 2018

Table of Contents

1. INTRODUCTION

1.1. Choice of the Dataset

The main data set corresponds to the National Survey of Income and Expenditures of Urban and Rural Households 2011-2012 of Ecuador (ENIGHUR). This survey provides data on the amount, distribution and structure of household income and expenditure, based on the demographic and socioeconomic characteristics of its members. The analysis of this information will serve to identify the characteristics of households according to their income and expenditures, find groups of vulnerable households as well as the main relationships among the selected variables. For the proposed analysis, 200 households from the city of Quito will be selected.

1.2. Dataset: Variables

In this data set, each row corresponds to a household. We find socio-economic information about the head of the household, which is the member of the household that provides the highest percentage of income. There is information regarding the composition of the household: the size of the household. In addition to household income, the total expenses and sources of expenses are presented: food, clothing, education, transportation and communication, housing, health, recreation and others. In the following tables, a summary of the variables used is presented.

	Variable	**Type**	**Categories**
Head of Household	Gender_HH	Qualitative	1. Male 2. Female
	Age_HH	Quantitative	
	Age_g	Qualitative	1. 17-25 2. 26-35 3. 36-45 4. 46-55 5. 56-65 6. >66
	Civil_status_HH	Qualitative	1. Married 2. Divorced 3. Separated 4. Single 5. Free union 6. Widower
	Education_level_HH	Qualitative	1. No education 2. Primary 3. Secondary 4. University
	Employmen_group_HH	Qualitative	1. Self-employed 2. Private employee 3. Public employee 4. Employer 5. Domestic Services 6. Unpaid Family Workers

Table 1. Head of household variables.

		Variable	Type
Household	Income	Household_size	Quantitative
		Earners	Quantitative
		Total_domestic_income	Quantitative
		Per_capita_income	Quantitative
	Outcome	Feeding	Quantitative
		Clothing	Quantitative
		Housing	Quantitative
		Health	Quantitative
		Transportation and Communication	Quantitative
		Recreational	Quantitative
		Education	Quantitative
		Others	Quantitative
		Total_domestic_outcome	Quantitative
		Per_capita_outcome	Quantitative

Table 2. Household variables

Once the variables were defined, the households that had completed information were selected. One of the problems found is that when the value of the expenses is null, in the database "NaN" is assigned instead of the value zero, this problem was corrected by simply replacing the correct value. In this information, 200 households were selected by simple random sampling for our analysis.

1.3. Dataset: Overview

For a better understanding of the behavior of households in relation to their income and expenses, descriptive statistics will be obtained. Python 3 was used for the analysis, and several tables of descriptive information were made using SPSS 21.

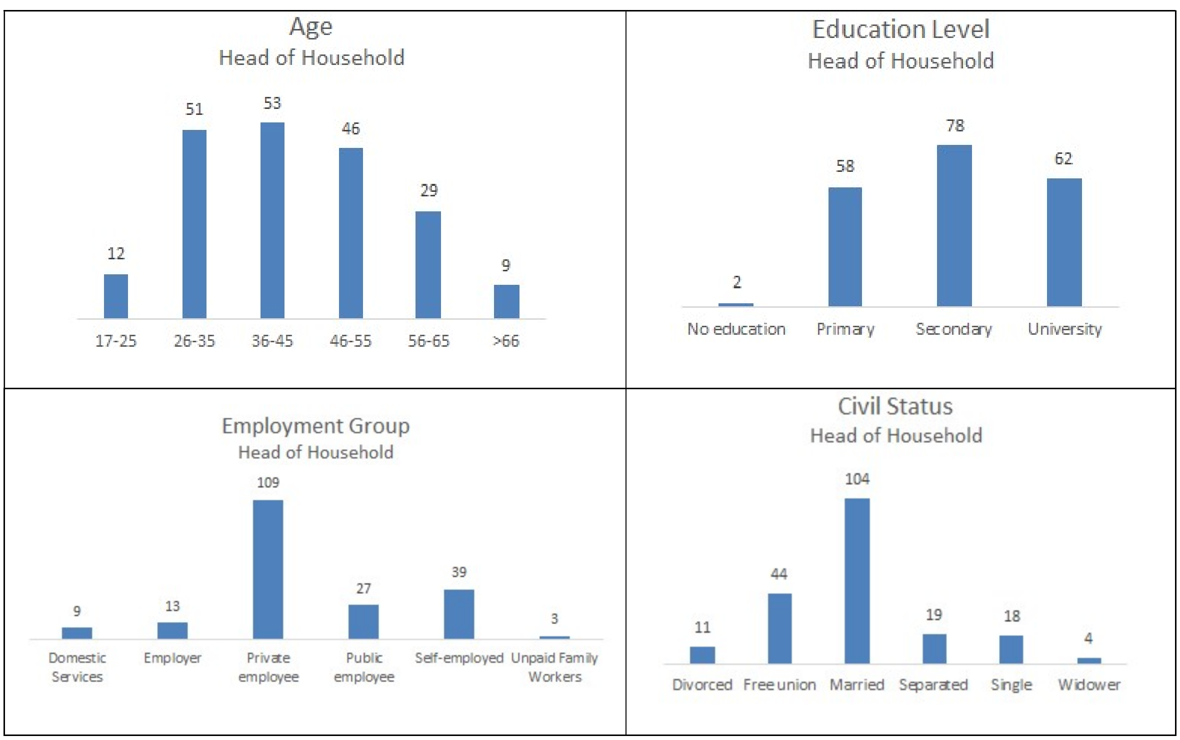


Figure 1. Head of Household: Frequency tables.

The heads of households, mostly do not have university studies, are private employees and are married.

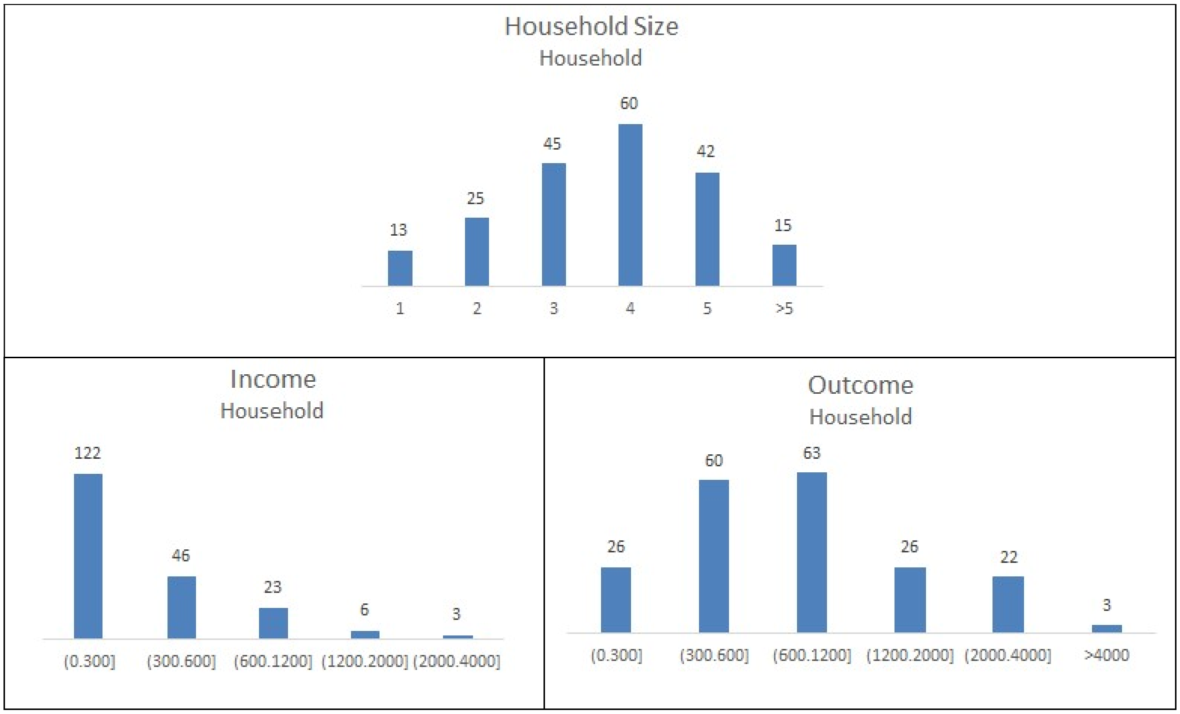


Figure 2. Household: Frequency tables.

About households, they are composed mostly of up to four members, their income is not higher than six hundred dollars. Household expenditures have a different distribution for the same ranges.

		Income(mean)	Outcome(mean)
Civil_ Status	Divorced	1585,81	1002,10
	Free union	1207,73	870,48
	Married	1759,23	1261,52
	Separated	637,11	504,67
	Single	683,46	478,63
	Widower	1130,65	920,52
Education level	No education	621,68	676,24
	Primary	786,28	552,05
	Secondary	1203,31	907,56
	University	2286,59	1584,63
Employment group	Domestic Services	483,66	319,87
	Employer	2530,95	1751,80
	Private employee	1281,61	939,34
	Public employee	2560,08	1673,81
	Self-employed	805,47	660,43
	Unpaid Family Workers	1662,55	1139,33

Table 3. Households Income/Outcome by household head characteristics.

The characteristics of the household head determine the income. Where we find a higher level of education the income is higher on average. Households where the head of household is a public employee receive higher income, salaries in the public sector are better. And in the case of the "Married", this difference is due to the fact that in households, 2 members contribute to the income.

		Food	Clothing	Housing	Health	T&C	Recreational	Education	Others
Education level	No education	166,28	38,01	64,28	2,1	338,87	32,05	0	34,66
	Primary	188,12	42,1	88,36	44,97	103,04	19,77	18,48	47,21
	Secondary	233,21	68,37	147,54	49,13	215,59	51,19	60,24	82,29
	University	281,96	103,77	223	91,06	394,05	113,16	212,57	165,05
Employment group	Domestic Services	110,55	21,45	79,93	22,36	44	11,12	0,05	30,42
	Employer	300,44	70,63	258,62	153,49	471,61	109,26	204,61	183,15
	Private employee	228,8	72,7	145,38	50,79	226,46	53,72	68,97	92,52
	Public employee	316,38	121,94	217,14	109,18	385,87	125,87	256,79	140,63
	Self-employed	196,52	46,85	116,98	23,59	133,86	33,28	41,42	67,93
	Unpaid Family Workers	289,75	42,86	78,4	163	350,52	49,12	74,31	91,39

Table 4. Households Outcome by household head characteristics.

As expenses are related to income, we observe that groups with greater acquisitive power spend on average more in education, health, clothing and recreation. These variables can help us establish groups.

2. Data Analysis

2.1. K-means Algorithm (Homework 2)

The features is based on Table 4, for the features education, health, clothing and recreation we see a greater difference in the means with respect to the groups of education level and employment group for the head of household. On these variables we will apply the K-Mean algorithm to divide the households according to their expenses.

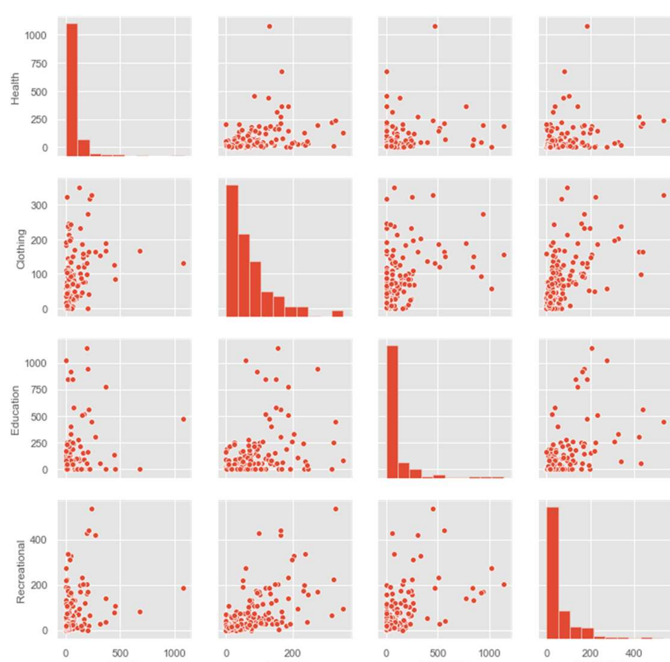


Figure 3. Pair Plot for the selected variables.

For this part I will use the K-mean algorithm of the Python library “sklearn” and I will also show the code of an own implementation (see appendix).

K=5		K=9	
Random initialization	Inertia	Random initialization	Inertia
1	3.129.943	1	2.145.600
2	3.062.034	2	2.142.497
3	3.137.906	3	2.062.551
4	3.129.943	4	2.040.858
5	3.129.943	5	1.916.273
6	3.472.598	6	1.916.209
7	3.200.292	7	2.062.551
8	2.954.701	8	2.031.042
9	3.134.880	9	2.268.306
10	3.066.523	10	2.055.717

Figure 4. Inertia

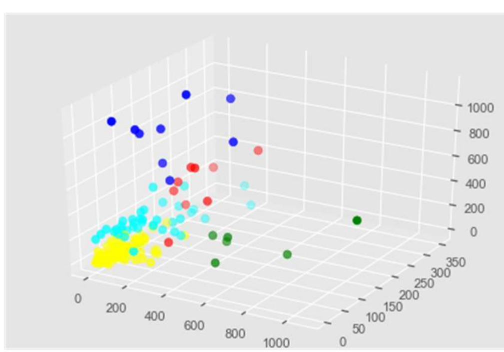
The value Inertia is the within-cluster sum-of-squares, so we will choose the smallest value. For these values let’s build a table with the average values of each feature on each cluster.

K=5, Means

Cluster	Count	Education	Recreational	Health	Clothing
0	9	311,08	360,93	130,37	211,27
1	6	110,00	85,88	555,96	137,94
2	9	841,51	148,56	124,67	145,10
3	39	154,07	91,60	64,55	118,45
4	137	13,93	25,88	28,77	41,09
Total	**200**	**94,75**	**61,10**	**60,45**	**71,42**

K=9, Means

Cluster	Count	Education	Recreational	Health	Clothing
0	11	785,54	182,49	134,84	150,49
1	21	212,05	87,07	36,66	75,26
2	5	38,33	65,89	451,51	139,53
3	45	43,76	52,15	37,36	69,36
4	9	236,81	328,10	110,68	214,64
5	12	33,97	38,90	160,03	80,38
6	8	36,82	91,23	83,19	245,72
7	1	468,33	185,86	1078,17	129,98
8	88	4,47	15,59	14,07	25,43
Total	**200**	**94,75**	**61,10**	**60,45**	**71,42**



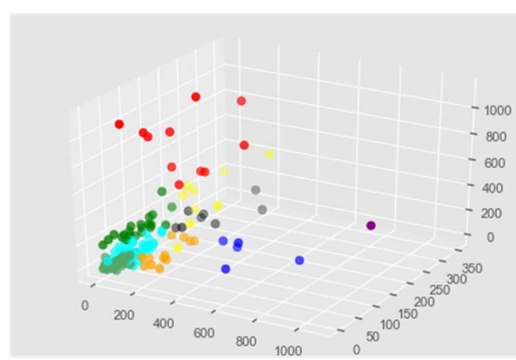


Figure 5. Clusters

Although the inertia in K = 5 is greater than in K = 9, we can see that for K = 5 a better interpretation is easier, for example, about the last clusters (3, 4) we can say that these groups represent low income households that survive with the basics. Cluster 0 is interesting because it represents the households with the highest average income ($ 5770) but that allocates their expenses to recreation and clothing. The other clusters (1, 2) divide their expenses between education and health, which would represent average households. I would select the division in K=5 because the averages of the features are further away than in the K=9, a better interpretation is easier. And by the Elbow Method we can see that after K=5 the sum-of-squares does not decrease too much.

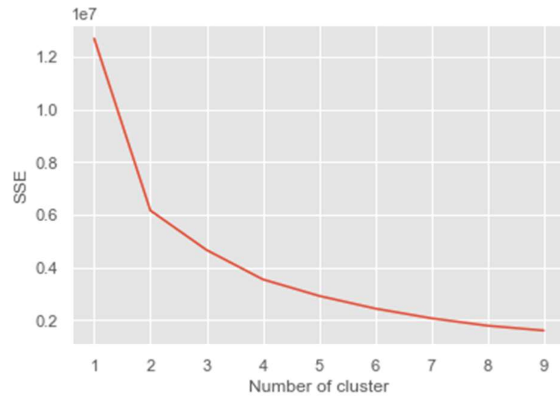


Figure 6. Elbow Method

2.2.1. Bootstrapping

I select the feature= Recreational and the cluster number 4 and 3, then calculate the 95% confidence interval for the feature.

For the feature

mean_feature= 61.1

Bootstrap_pivotal=[49.35 , 72.81]

Bootstrap:non-pivotal=[50.03 , 73.47]

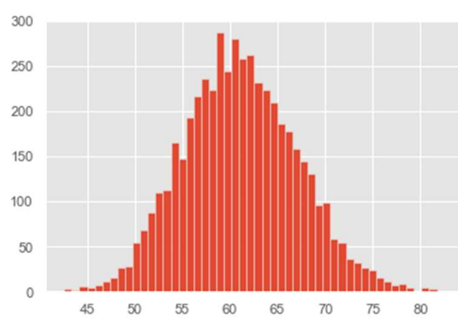


Figure 7. Bootstrapping

And did bootstrapping for the differences:

Feature vs Cluster 4	Feature vs Cluster 3 mean_cluster3= 91.60	Cluster 3 vs Cluster 4 mean_cluster4= 25.88
Bootstrap pivotal= [22.44 , 47.85] Bootstrap non-pivotal= [22.88, 48.76]	Bootstrap pivotal= [-52.31 , -8.55] Bootstrap non-pivotal= [-51.99 , -8.15]	Bootstrap pivotal= [-84.56 , -46.6] Bootstrap non-pivotal= [-84.64 , -46.43]

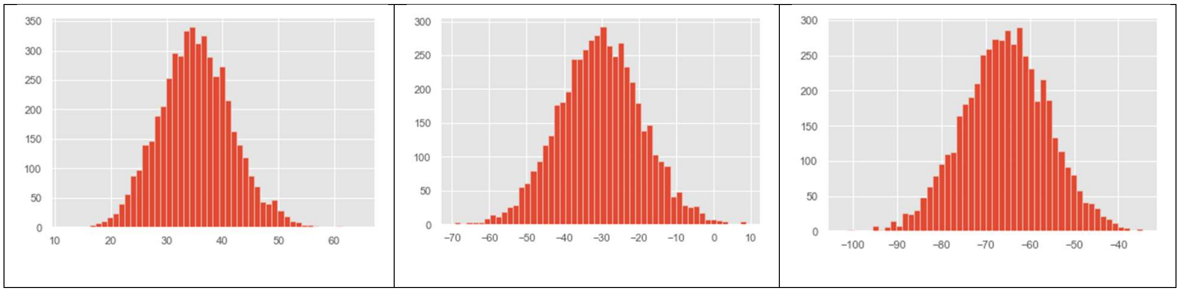


Figure 8. Comparison with Bootstrap

The 0 value it is not in the 95% confidence intervals, so we can conclude that there are differences in the means.

2.3. Contingency Table (Homework 3)

For this task I took the features: Gender, Civil Status and Level of Education. Let's start by building the contingency tables for these features.

Gender - Civil Status

Conditional Frequency Table

Gender	Divorced	Free union	Married	Separated	Single	Widower	Total
	Civil_status						
Female	20%	9%	5%	39%	23%	5%	100%
Male	1%	26%	65%	1%	5%	1%	100%
Total	6%	22%	52%	10%	9%	2%	100%

Relative Frequencies

Gender	Divorced	Free union	Married	Separated	Single	Widower	Total
	Civil_status						
Female	5%	2%	1%	9%	5%	1%	22%
Male	1%	20%	51%	1%	4%	1%	78%
Total	6%	22%	52%	10%	9%	2%	100%

Quetelet relative index table

Gender	Divorced	Free union	Married	Separated	Single	Widower
	Civil_status					
Female	271,9	-58,7	-91,3	306,7	152,5	127,3
Male	-76,7	16,6	25,7	-86,5	-43,0	-35,9

Chi-square: 115.69

Summary Quetelet index: 57.84

Degrees of freedom: 5

Figure 9. Gender-Civil Status

Let's analyze the households, when the head of household is male, the civil status of the head is more likely to be: married or free union. When a woman is the head of household, her civil status has a greater possibility of being: separated or divorced.

With the Quetelet index we confirm that, for example, the gender females raises the frequency for the civil status category divorced by 271.9%. The gender female provides for a strong increase in the probabilities. The average knowledge of the civil status “adds” 57.84% to frequency of gender.

Gender – Level of Education

	Conditional Frequency Table Education_level				
Gender	No education	Primary	Secondary	University	Total
Female	2%	34%	41%	23%	100%
Male	1%	28%	38%	33%	100%
Total	1%	29%	39%	31%	

	Relative Frequencies Education_level				
Gender	No education	Primary	Secondary	University	Total
Female	1%	8%	9%	5%	22%
Male	1%	22%	30%	26%	78%
Total	1%	29%	39%	31%	100%

	Quetelet relative index table Education_level			
Gender	No education	Primary	Secondary	University
Female	127,3	17,6	4,9	-26,7
Male	-35,9	-5,0	-1,4	7,5

Chi-square: 2.72

Summary Quetelet index: 1.36

Degrees of freedom: 3

Figure 10. Gender – Education Level

For these features we cannot say much, regardless of the sex of the head of household, their education is secondary. It is true that there are more masculine heads of households with higher education but this is because there are more male heads in the households. By the Quetelet Qndex we only see that being a female head of household raises the frequency of the level of education: not education by 127.3%. The average knowledge of the education level “adds” only a 1.36% to frequency of gender.

Analyzing the Summary of the Quetelet index, we can see that they are more related to the features “Gender-Civil Status” than “Gender – Educational Level”. Civil status “adds” 57.84% to the Gender Frequency. Let’s look at the chi-square statistic.

	Chi-square	Degrees of freedom	Probability	Critical value
Gender – Civil Status	115,69	5	0,95%	11,1
			0,99%	15,1
Gender – Education Level	2,71	3	0,95%	7,82
			0,99%	11,35

Table 5. Hypothesis test of independency.

For the "Gender - Civil status" features we reject the null hypothesis (features are independent). We conclude that they are dependent, that there is an association between the two variables at 95% and 99% of confidence.

For the "Gender – Education level" features we can´t reject the null hypothesis (features are independent). We conclude that they are independent, that there is not an association between the two variables at 95% and 99% of confidence. For this pair of features, we can reject the null hypothesis if N=577 for 95% of confidence and N=839 for 99% of confidence.

2.4. Principal Component Analysis (Homework 4)

For starting database I will take the features used for the K-means algorithm: Education, Recreation and Health. These features refer to household expenses according to these features. I choose these features because when finding clusters it was possible to identify groups that characterized the households.

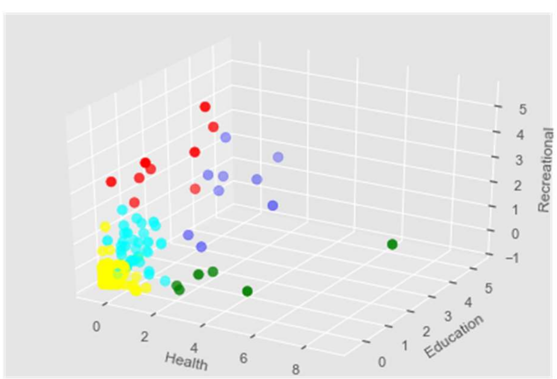
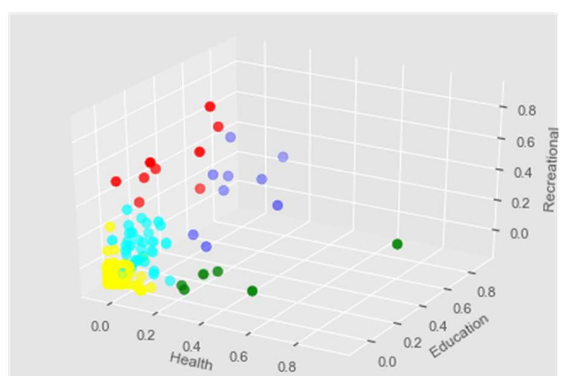
Task 1. Standardize the selected subset; compute its data scatter and determine contributions of all the principal components to the data scatter, naturally and per cent.

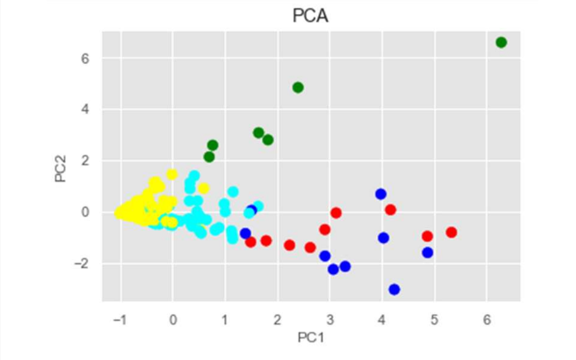
Singular values: [18.7406 12.5313 9.5789]

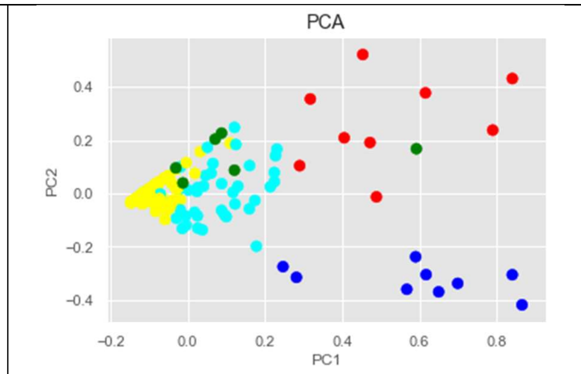
Data_scatter: 600

Contributions of principal components	
Natural	**Percent**
18,7406	59%
12,5313	26%
9,5789	15%

Task 2,3,4. Visualize the data with these features using standardization with two versions of normalization: (a) over ranges and (b) over standard deviations. And apply PCA. Compute and interpret a hidden factor behind the selected features.

Over standard deviations	**Over ranges**
	





"Hidden factor"

PC	Health	Education	Recreation
1	47%	61%	63%
2	88%	-40%	-26%

"Hidden factor"

PC	Health	Education	Recreation
1	21%	73%	65%
2	25%	-69%	68%

Figure 11. PCA

About the visualization of the data with respect to the two types of standardization, I do not see significant differences, it may be because I considered the groups obtained by the K-means method of the first part of the work.

When performing the PCA for two components, the graphics do vary. The first component always separates the groups of households with the best income (right) from those with the worst income (left), but the second component changes:

- For standardization with standard deviation this second axis separates the clusters where households spends more on health and less on education (above) and those that spend more on education and less on health (below).
- For the standardization with the range, this second axis separates the groups by their expenses in recreation(above) and education (below)

2.5. Linear regression (Homework 5)

For this part I will take the features: "Total_domestic_income" and "Total_domestic_outcome". The expenses are related to the income, a household cannot spend more of the money it receives.

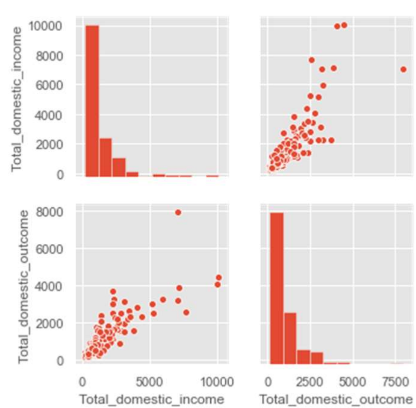
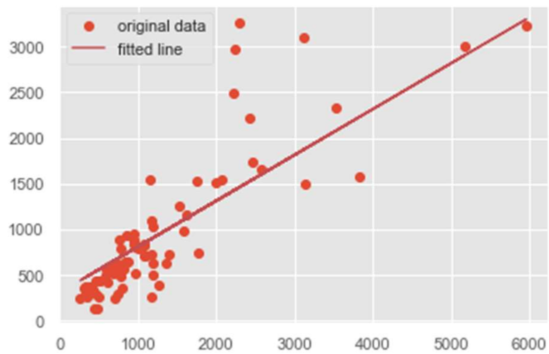


Figure 12. Scatter plot for linear regression.

For the linear regression model, the independent variable is: Total_domestic_income and the dependent variable is: Total_domestic_outcome. As a result of the linear regression model we have:



Slope= [0.50]

Intercept=310.73

R^2= 0.626

correlation=0.79

Figure 13. Linear regression

The slope=0.5873 and means that as the **Total_domestic_income** variable increases by 1, the predicted value of Total_domestic_outcome decreases by 0.5873. For this case the intercept is 310.73, we have households that spends more than they earn. The determinacy coefficient= 0.63, so in our model, 63% of the variability in Total_domestic_outcome can be explained using Total_domestic_income. The correlation of r = 0.79 suggests a strong, positive association between the two variables.

Values	Prediction
500	561.48
1000	812.23
2000	1313.74

About predictions we can say that low-income households have problems when they spend, they spend more than they earn. When the income is higher this relationship improves.

The MAE=262.29, measures the average magnitude of the errors in our set of predictions. This value, in comparison with the scale of our data is small, and shows to some extent the degree of explanation of the variability reached.

3. References

- Lebart Ludovic, Piron Marie, Morineau Alain, 2006, Statistique exploratoire multidimensionnelle, 4ta. Edición Dunod, París.
- Peña Daniel, 2002, Análisis de datos multivariantes, McGraw-Hill, España.
- Anderson T. W, 2003, An introduction to multivariate statistical analysis, Wiley Interscience.
- Mirkin B, 2011, Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer.
- https://machinelearningmastery.com
- https://www.wikipedia.org/

A. Appendix

Annex, the principal parts of code used to obtain the data. The code is on Python 3.

A1. Homework 2

```
#Kmeans method with Python
datas = pd.read_csv("dataSP1.csv",sep = ';')
X = datas[["Health","Clothing","Education","Recreational"]]

# Some variables
RANDOM_STATE = 42
NUM_CLUSTERS =9
NUM_ITER = 20
NUM_ATTEMPTS = 10
data_sample = X

from sklearn.cluster import KMeans

km = KMeans(n_clusters=NUM_CLUSTERS, init='random', max_iter=500, n_init=1)#, verbose=1)
km.fit(data_sample)

final_cents = []
final_inert = []
label=[]

for sample in range(NUM_ATTEMPTS):
    print('\nCentroid attempt: ', sample)
    km = KMeans(n_clusters=NUM_CLUSTERS, init='random', max_iter=500, n_init=1)#, verbose=1)
    km.fit(data_sample)
    inertia_start = km.inertia_
    intertia_end = 0
    cents = km.cluster_centers_

    for iter in range(NUM_ITER):
        km = KMeans(n_clusters=NUM_CLUSTERS, init=cents, max_iter=500, n_init=1)
        km.fit(data_sample)
        print('Iteration: ', iter)
        print('Inertia:', km.inertia_)
        print('Centroids:', km.cluster_centers_)
        inertia_end = km.inertia_
        cents = km.cluster_centers_
        scores= km.labels_

    final_cents.append(cents)
    final_inert.append(inertia_end)
    label.append(scores)


#Elbow curve
sse = {}
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k, max_iter=1000).fit(X)
    sse[k] = kmeans.inertia_ # Inertia: Sum of distances of samples to their closest cluster center
plt.figure()
plt.plot(list(sse.keys()), list(sse.values()))
plt.xlabel("Number of cluster")
plt.ylabel("SSE")
plt.show()



My kmeans


```python
#Kmeans Algorithm
#X dataset
#k number of clusters
def kmean(X,k):
    #Chose  random centroids
    id = np.random.randint(0, X.shape[0], size=k)
    C=X[id,:]

    # Store centrois
    C_old = np.zeros(C.shape)
    clusters = np.zeros(X.shape[0])
    #Inicialize loop
    error = np.linalg.norm(C- C_old)

    while error != 0:

        for i in range(len(X)):
            distances = np.linalg.norm(X[i]-C,axis=1)
            cluster = np.argmin(distances)
            clusters[i] = cluster
```
```

```
        C_old = copy.deepcopy(C)

        for i in range(k):
            points = [X[j] for j in range(len(X)) if clusters[j] == i]
            C[i] = np.mean(points, axis=0)

        error = np.linalg.norm(C- C_old)

    return clusters, C


#For Kmeans criterion
def crit(X,C,cluster,k):
    criterio=0
    for i in range(len(X)):
        labels=cluster.astype(np.int64)
        j=labels[i]
        criterio=criterio+np.linalg.norm(X[i]-C[j,:])**2

    return criterio
```

A2. Homework 3

```
#Create contingency table, chi-squared value
def conting(X):

    n=X.shape[0]
    T=X/X.sum(axis=0)
    C=np.zeros((X.shape[0],X.shape[1]))
    D=np.zeros((X.shape[0],X.shape[1]))
    for i in range(n):
        C[i,:]=X.sum(axis=0)*(X[i,:].sum()/X.sum())

    D=(X-C)**2/C
    chi=D.sum()
    return chi,T


#Create quetelet index
def quetelet(X):
    n=X.shape[0]

    Q=np.zeros((X.shape[0],X.shape[1]))
    D=np.zeros((X.shape[0],X.shape[1]))
    C=X/X.sum()
    for i in range(n):
        D[i,:]=(C[i,:])/C[i].sum()
        Q[i,:]=100*(D[i,:]-C.sum(axis=0))/C.sum(axis=0)

    quetelet=np.sum(C*Q)
    return C,Q,quetelet

#Bootstrap
plt.hist(Mean_cluster, bins=50)
plt.show()

meanc=np.mean(Mean_cluster)
stdc=np.std(Mean_cluster)

p1c=np.percentile(Mean_cluster, 2.5)
p2c=np.percentile(Mean_cluster, 97.5)

print("mean_cluster=",np.mean(boot2))
print("Bootstrap pivotal=","[",meanc-1.96*stdc,",",meanc+1.96*stdc,"]")
print("Bootstrap non-pivotal=","[",p1c,",",p2c,"]")
```

A3. Homework 4

```
XP= datas[["Health","Education","Recreational"]]

#Standarized data
XP_std=(XP-XP.mean(axis=0))/XP.std(axis=0)
XP_range=(XP-XP.mean(axis=0))/(XP.max(axis=0)-XP.min(axis=0))


#Data scatter

from numpy.linalg import svd
```

```
U, s, V = np.linalg.svd(XP_std)
data_sct=0
for i in range(len(s)):
    data_sct=data_sct+s[i]**2
```

#PCA

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(XP_std)
principalDf = pd.DataFrame(data = principalComponents
             , columns = ['principal component 1', 'principal component 2'])

finalDf = pd.concat([principalDf, datas[['K5']]], axis = 1)
pd.DataFrame(pca.components_)
```

A4. Homework 5

```
from sklearn import linear_model
from sklearn.metrics import r2_score
from scipy import stats
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn.model_selection import train_test_split

X=datas[["Total_domestic_income"]]
y=datas[["Total_domestic_outcome"]]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=1)

regression_model = linear_model.LinearRegression(fit_intercept=False)
regression_model.fit(X_train, y_train)
y_pred = regression_model.predict(X_test)



m=regression_model.coef_[0]
b=regression_model.intercept_
r=regression_model.score(y_test, y_pred)
print("slope=",m, "intercept=",b,"r^2=",r,"correlation",r**(1/2))
regression_model.predict([[500]])

mean_absolute_error(y_pred, y_test)
```