Academic Year 2020/2021

Data Mining Project

Bacchiocchi Francesco: f.bacchiocchi@campus.fct.unl.pt
Castelli Daniele: d.castelli@campus.fct.unl.pt

November 21, 2021

Part I - Defintion of the project

The data contains general information about houses blocks found in a given California district and some summary stats about them based on the 1990 census data. In particular each block is described by 10 features. Since the "ocean proximity" field is not numerical in the original data we encoded it in a discrete categorical representation (0-4). Moreover we dropped the NaN values getting a number of observations equal to 20433. Plotting the histograms we can immediately observe that two features, namely "housing median age" and "median house value" have been cut off (meaning that all values larger than an upperbound are reported as the upperbound itself). For this reason we excluded the data concentrated in the upperbound since we would not be able to infer any information about the real distribution of this large misreported values.

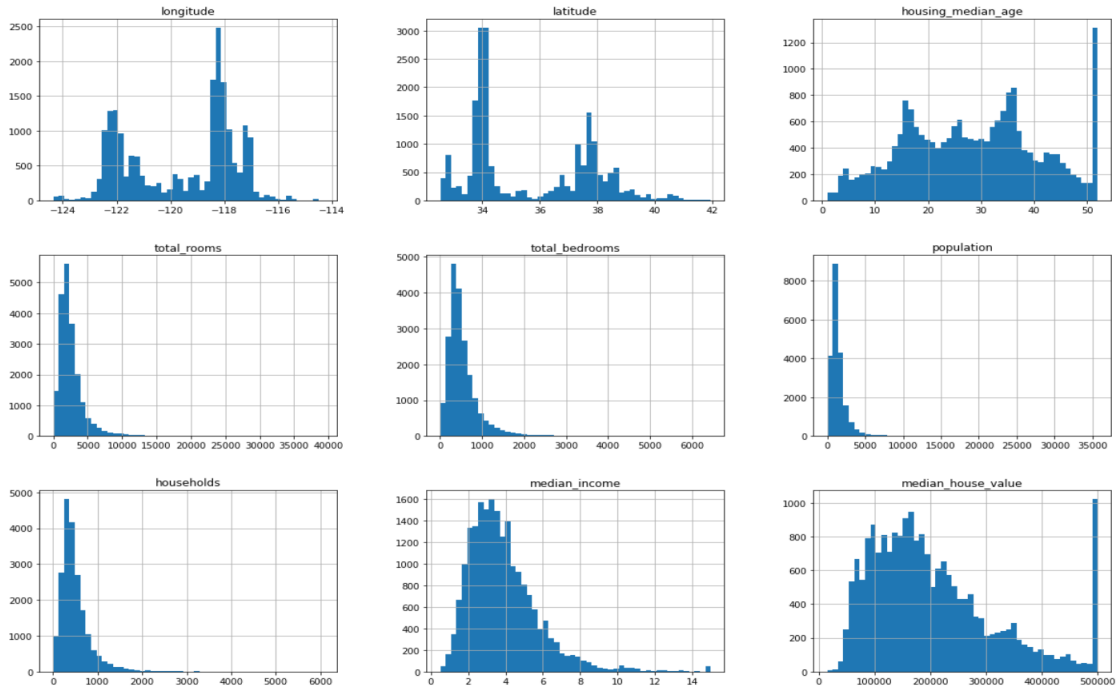


Figure 1: Histograms of Data

After removing the aforementioned observations we were left with 18336 data points. We made a scatterplot of the latitude and the longitude, synthetically showing the geographical distribution of median house values using a color scale. As expected, we observe that houses closer to the coast are generally more expensive. Finally we underline the fact that the scatter plot of the geographical coordinates has the shape of the California state as previously mentioned.

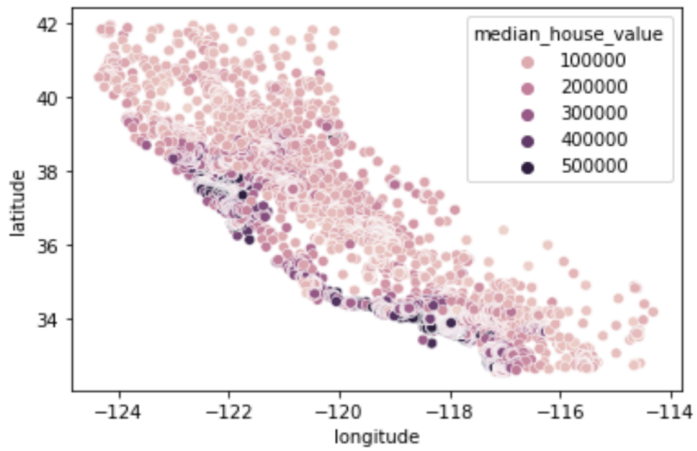


Figure 2: Geographical distribution of House Values

To look at possible correlation between features we have plotted the correlation matrix, which is reported below. It is fairly evident that features related to the size of the households are much correlated.

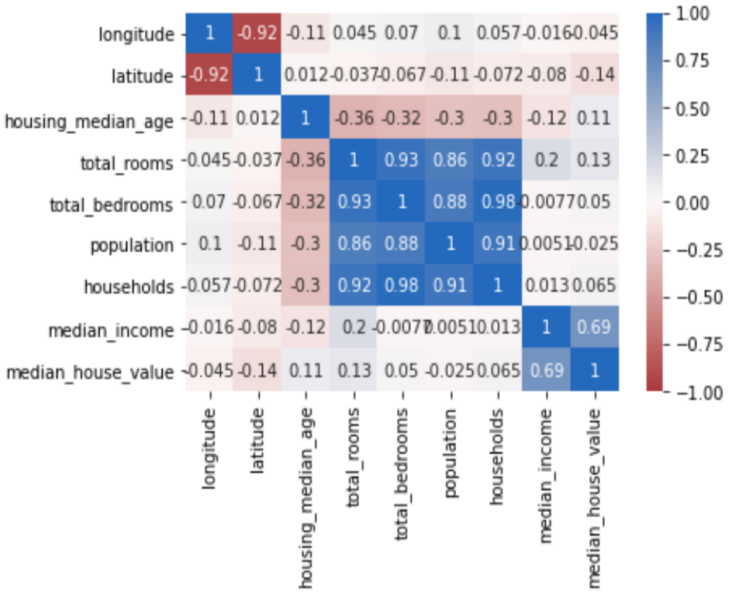


Figure 3: Correlation heatmap

Part II - Linear Regression

In this section we have performed a linear regression between two covariates of the chosen dataset: "total bedrooms" and "total rooms", taking as:

- Independent Variable (X): "total bedrooms"
- Response Variable (Y): "total rooms"

Point a)

First of all we have performed a scatter plot of the total bedrooms and the total number of rooms. We can clearly observe a linear trend between the two variables and so it is reasonable to fit a linear regression model on this data.

Point b)

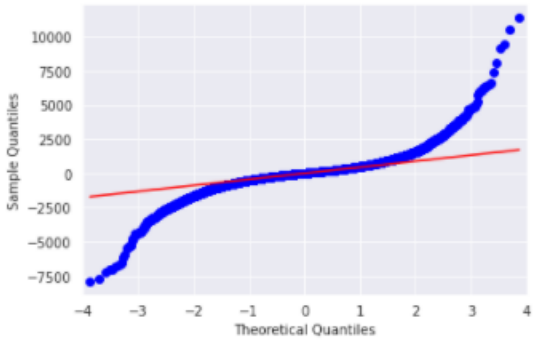
We can now fit a linear regression model, given by:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \;\; i.i.d. \;\; \forall i = 1...n$$

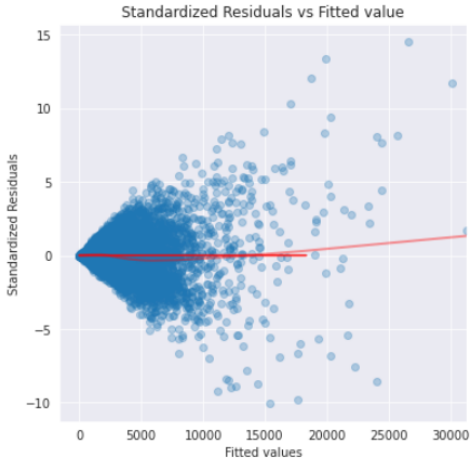
Plotting both the data and the regression line we have:

To check the validity of the following model we had to verify four different hypothesis:

- **Normality of residuals**. To check if residuals are normally distributed we can provide a normal qq-plot, to compare the sample quantiles of the standardized residuals with the one of a standard normal distribution. In this case we can observe two skews which indicate that residuals are not Gaussian. This intuition is confirmed by the p-value of the Shapiro test which is almost zero.



- **Homoschedasticity of residuals**. To check this hypothesis we have to verify that the variance of the errors is constant along the fitted values. In this case we can observe many heteroschedastic effects indeed a smaller variance of the residuals is associated to smaller fitted values while an higher variance is associated to bigger ones. This is probably imputable to the presence of many outliers in the two data distributions.



- **Linearity of the model**. This assumption is confirmed by the scatter-plot of the data which clearly indicate a linear response in the number of rooms with respect the number of bedrooms. Moreover, since we didn't observe any curvature in the trend of the residuals, we can state that this assumption is completely fulfilled.
- **Independence of residuals** Since all the measurements are taken in different places and are related to different houses we can assume the observations are independent.

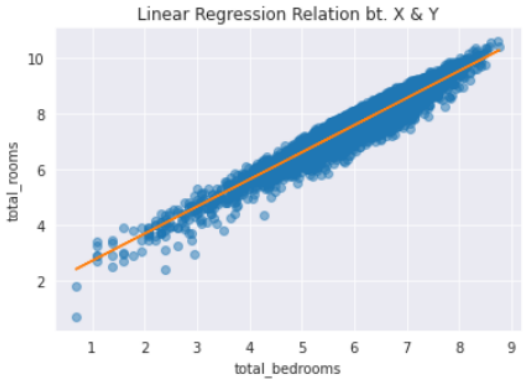
In conclusion, despite the r^2=0.87, we have not verified the all assumptions required by a linear regression model. This could be a consequence of the long tailed distribution of both response and the dependent variables.

Point c)

To overcome the presence of outliers in both the dependent and independent variables we have fitted a linear regression on log-transformed variables. In particular we have fitted the following model:

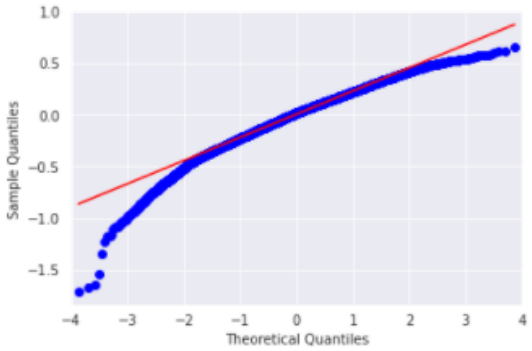
$$log(Y_i) = \beta_0 + \beta_1 log(X_i) + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i = 1...n$$

In this case we get the following scatter plot of the data and regression line:

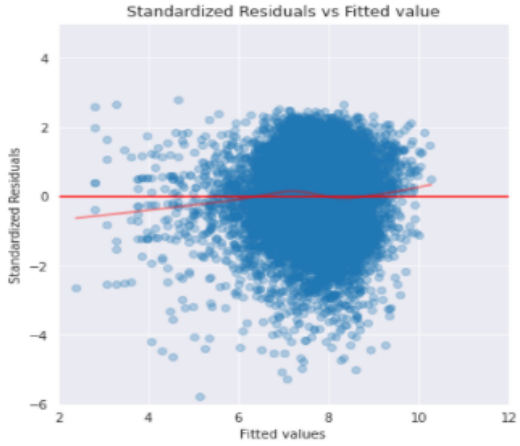


Also in this case we had to verify the same hypothesis presented before.

- **Normality of residuals**. To check if residuals are normal distributed we can provide a normal qq-plot as previously done. Also in this case we can observe two skews standing for a non gaussianity of residuals. Indeed, providing a Shapiro test, we get also in this case a p-value of almost zero.



- **Homoschedasticity of residuals**. To check this hypothesis we have to verify that the variance of the errors is constant along the fitted values. In this case we observe a homogeneous residuals' trend along the fitted values and so we can assume that this assumption is fulfilled. This is clearly a consequence of the log-transformation we applied to the data.



- **Linearity of the model**. This assumption is confirmed also in this case by the scatter-plot of the data which clearly shows a linear response in the number of rooms with respect the number of bedrooms.

- **Independence of residuals** As observed before, since all the measurements are taken in different places and are related to different houses, we can assume that observations are independent.

Remark: From now on we write with capital letters the original variables and with small letters the log transformed ones, i.e. $y_i = log(Y_i)$ and $x_i = log(X_i)$

Point d)

The linear regression model on log-trasformed data is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i = 1...n$$

where we have estimate the following parameters:

- $\hat{\beta}_0$= 1.72721 which represents the intercept of the ERE. $\hat{\beta}_0$ is the estimate of β_0 which represents the expected value of the response variable (that is the natural logarithm of the total number of rooms) when the natural logarithm of the number of bedrooms is equal to zero.

- $\hat{\beta}_1$= 0.97422 which represents the slope of the ERE. $\hat{\beta}_1$ is the estimate of β_1 which represents the increment in the response variable (which is the natural logarithm of the total number of rooms) for an unitary increment of the dependent variable (which is the natural logarithm of the number of bedrooms).

Point e)
The determinacy coefficient r^2 is defined as:

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

or equivalently, since $SSR = SST - SSE$ we get:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

r^2 represents the portion of variability of the response variable explained by the ERE and thus it ranges from 0 to 1. As we can see from the second formulation, its maximum occurs when $\hat{y}_i = y_i \ \forall i$ (perfect fit). In our case the computation gives $r^2 = 0.9012$ which clearly shows that the fitted regression model is able to capture the data variability well.
The sample correlation coefficient $\hat{\rho}$ is defined as:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sqrt{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n}\sqrt{\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n}}$$

In our case the computation gives $\hat{\rho} = 0.9493$. (Note that in this case, since we are performing a simple linear regression, we have that $\hat{\rho} = +\sqrt{r^2}$)
Point f)
Thanks to the fact that the statistics $t = \hat{\beta}_1 / Se(\hat{\beta}_1)$ is distributed as a t-student with $n-(m+1)$ degrees of freedom we can perform a t-test to check if there exists a linear relationship between the two variables. In particular we have performed the following t-test:

$$\begin{cases} H0 : \beta_1 = 0 \\ H1 : \beta_1 \neq 0 \end{cases}$$

where we have defined:

$$Se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2/n}} = 0.00238$$

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-m-1}} = 0.23326$$

$m = 1$, number of regressors.
$n = 18336$, total number of observations
Since we obtain a $p - value = 0$ we have statistical evidence to affirm that there exists a linear relationship between the two variables.

Point g)
The confidence interval at level 95% for the unknown true slope of the regression line is given by:

$$CI_{0.95}(\beta_1) = [\hat{\beta}_1 \pm t_{n-2,0.975} Se(\hat{\beta}_1)] = [0.96954, 0.97890]$$

Point h)
Assuming that both x_i and y_i come from a Gaussian distribution for all $i = 1...n$, the confidence interval at level 95% for the correlation coefficient is given by:

$$CI_{0.95}(\rho) = [\hat{\rho} \pm t_{n-2,0.975} \sqrt{\frac{1-\hat{\rho}^2}{n-2}}] = [0.94455, 0.95367]$$

Point i)
Choosing an house with $X_p = 110$ bedrooms which approximately corresponds to $x_p = 4.7$, the confidence interval at level 95% for the mean of the response variable is equal to:

$$CI_{0.95}(E[y|x_p]) = [\hat{y}_p \pm t_{n-2,0.975} \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}] = [6.30438, 6.30777]$$

Point j)
For the same choice of $x_p = 4.7$ the prediction interval at level 95% for the real value of the response variable is given by:

$$CI_{0.95}(y|x_p) = [\hat{y}_p \pm t_{n-2,0.975} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}] = [6.19941, 6.41274]$$

As the intuition suggests the prediction interval is wider than the confidence interval, this holds in general as we can see from their definitions.

Part III - Principal Component Analysis

Points a)

In this section we have analyzed six features of our data set, leaving out: "longitude", "latitude", "median house value" and the categorical one "ocean proximity". To do so we have performed a PCA to detect the presence of particular data pattern or hidden structure in a meaningful and lower dimensional space. We have firstly considered two different way to standardize the data and then projected them in the 2D PC plane. In particular we have that:

- Standardization by standard deviations (Z-scoring):

$$z_{ij} = \frac{x_{ij} - \hat{\mu_j}}{\hat{\sigma}_j} \quad i = 1...n, \quad j = 1...6$$

where:
$\hat{\mu_j}$: is the sample mean of the j-th column of the dataset
$\hat{\sigma}_j$: is the sample standard deviation of the j-th column
performing a standardization by standard deviations and projecting the data along the 1st and 2nd PC components we have obtained the following:

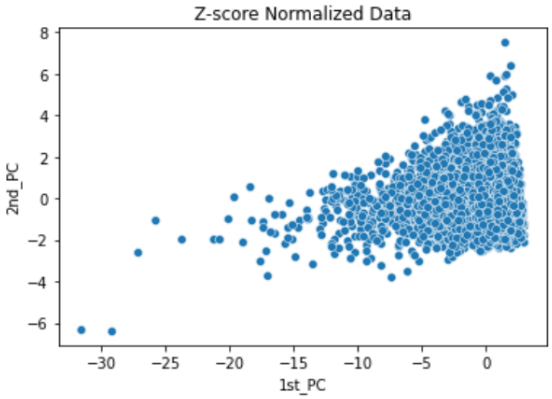


Figure 4: 2D PC Plane - Standardization by standard deviations

- Standardization by range:

$$z_{ij} = \frac{x_{ij} - \hat{\mu_j}}{r_j} \quad i = 1...n, \quad j = 1...6$$

where:
$\hat{\mu_j}$: is the sample mean of the j-th column of the dataset
r_j: is the difference between the maximum and the minimum of the j-th column
performing a standardization by range and projecting the data along the 1st and 2nd PC components we have obtained the following:

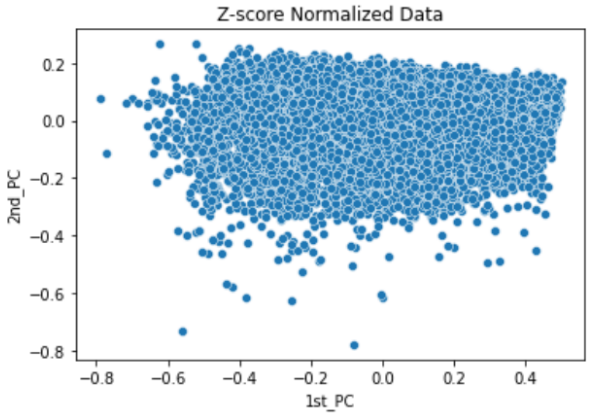


Figure 5: 2D PC Plane - Standardization by range

Points b)

In order to perform dimensionality reduction we decided to use SVD on the data matrix, to ensure the numerical stability of the procedure. Our data matrix D has features on columns and observations on rows, thus the Covariance matrix is given by $C = (D^T D)/(n-1)$ where n is the number of observations. Standard PCA performs a diagonalization of the covariance matrix, namely $C = W\Lambda W^T$. The singular value decomposition applied to the data matrix leads to the representation $D = U\Sigma V^T$. Thus we get

$$C = \frac{(U\Sigma V^T)^T(U\Sigma V^T)}{n-1} = \frac{V\Sigma^T U^T U\Sigma V^T}{n-1} = \frac{V\Sigma^2 V^T}{n-1}$$

We can see that the orthonormal basis to which we earlier referred as W is nothing more than than the matrix of right singular vectors V, and there is a simple relation between the "principal values" λ_i and the singular values σ_i, namely: $\lambda_i = \frac{\sigma_i^2}{n-1}$. Looking at the boxplots of the selected variables after the range normalization we can clearly observe that the two features, namely "housing median age" and "median income" present a bigger boxplot than the others. This badly affect the PCA algorithm since all the variability of the data will be captured by the latter features ("housing median age" and "median income") so that the direction of the eigenvalues associated to the first and second principal components will almost coincide with those variables. For this reason we have decided to proceed with the standardization by standard deviations which does not presents this kind of issue.

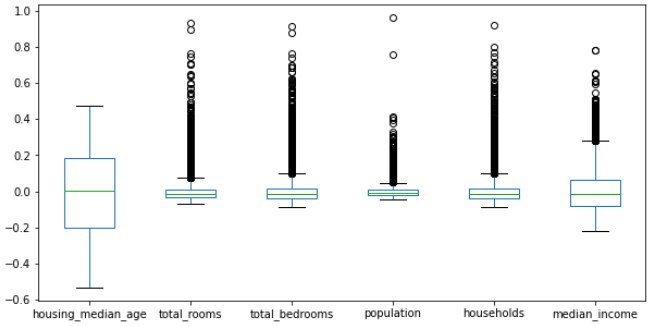


Figure 6: Boxplot of the data after the range standardization

Point c)
Looking at the data projected along the first two principal components and assigning a darker colour to those houses having an higher price we can immediately recognize a significant data structure. Observations having a higher PC2 are associated to more expensive houses while we cannot recognize a clear separation in terms of price along PC1. To further understand the results we can look at the matrix V^T whose rows are the vector's basis of the new reference system provided by the PCA.

$$V^T = \begin{pmatrix} 0.22193098 & -0.48852597 & -0.49202143 & -0.47078436 & -0.49335729 & -0.07045993 \\ -0.50883161 & 0.05305942 & -0.13517319 & -0.14044988 & -0.13064837 & 0.82656405 \\ 0.82974318 & 0.12133116 & 0.02554927 & 0.08735314 & 0.06817802 & 0.5327977 \\ & & \dots & \dots & & \end{pmatrix}$$

We can observe that the first row of V^T corresponds to an "average" between the features: "total rooms", "total bedroom", "population" and "house holds". On the other hand the second row has much smaller values associated to the above mentioned features, giving instead particular importance to the first and the last one, which are: "housing median age" and "median income". The observed separation of data with respect to "median house value" along the 2nd PC let us deduce that this quantity is very much influenced by the two features enhanced in the 2nd PC. We show below the scatter plots of the data both in the 2D and 3D PC spaces, highlighting the aforementioned separation with respect to the "median house value" feature.

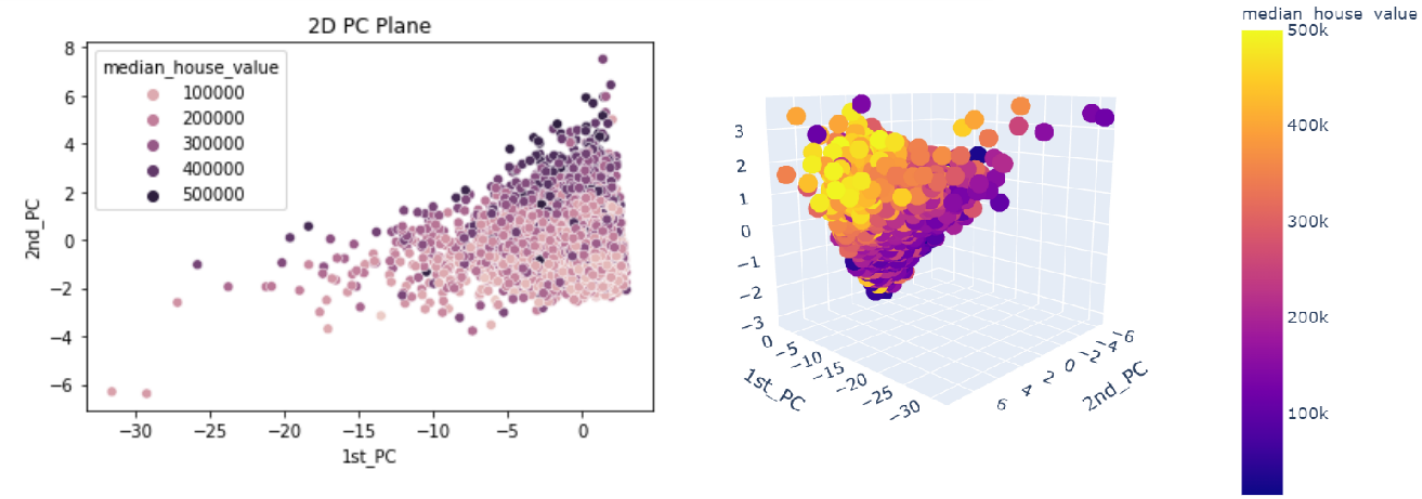


Figure 7: Projection of the data in 2D and 3D PC plane

Point d)

In order to evaluate the quality of the PC projection we show the plots of both the variance explained by the individual principal components and the cumulative explained variance.Since we performed PCA through the SVD of the Data matrix the variance explained by the $i-th$ principal component is given by $\frac{\sigma_i^2}{\sum_{j=1}^{r}\sigma_j^2}$ and accordingly the cumulative variance explained by the first k components is $\frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{j=1}^{r}\sigma_j^2}$. We can see that the first three principal components are able to capture more than 95% of the total variance while the first two almost the 85%. For this reason we think that two PC are enough to capture well the variability of the data.

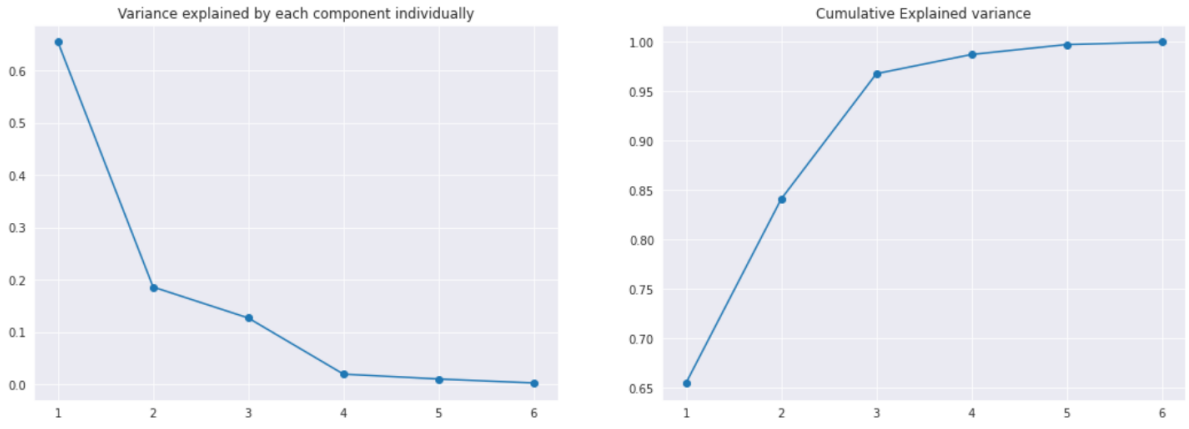


Figure 8: Quality of the Principal Components

Part IV - Fuzzy Clustering with Anomalous Patterns

Remark Due to the large number of observations in our dataset all the algorithms involved in this part of the project would have required a big amount of time to converge. For this reason we have decided to reduce the size of our dataset to 1000 observations.

Point a)

We applied the FCM algorithm with the parameter c ranging form 2 to 9. For each value we run the algorithm for multiple random initialization (actually three); we then stored the cost functions associated to the different values of c. In particular, for each c, we took the minimum of the cost function among the different random initializations. The plot shows that this cost function is monotonic decreasing with c and it does not present a remarkable elbow. Therefore it does not represent a good criterion to choose the optimal value of c. Moreover for each value of c we analyzed the variability of the solutions (final prototypes and membership matrix) corresponding to different seeds. This was achieved introducing a variance metric contained in the function "my_norm". It turns out that for every fixed c the algorithm converged to very close solutions for each seed, thus also this analysis was not enough to establish an optimal value of c.

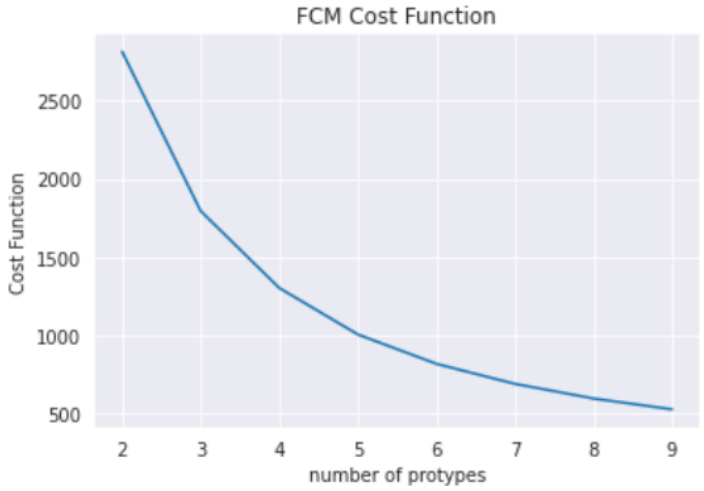


Figure 9: Cost Function for different values of c

Point b)

The algorithm of Anomalus Pattern can be a very effective tool to find a good initialization for clustering algorithms. The procedure is the following:

- 1) Standardized dat by centering (mandatory) and possibly normalizing.
- 2) Find Anomalous cluster and store its cardinality and centroid.
- 3) Remove the cluster just found from the dataset. If the dataset is empty the algorithm has ended otherwise go back to step 2.

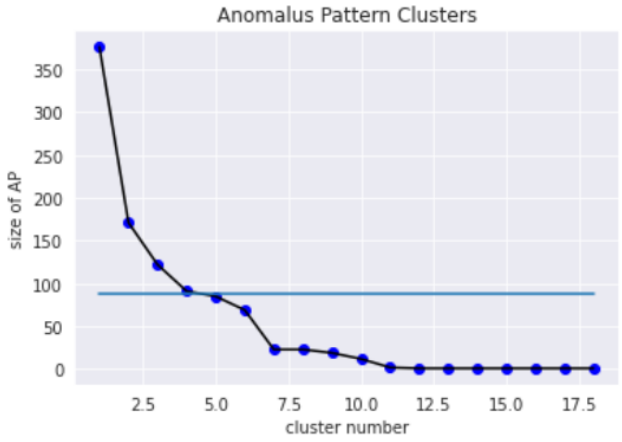
In particular step 2 performs an "almost" 2-Means with initial centers being respectively 0 (i.e the grand mean since the data is centered) and the observation furthest away from 0. The only difference with respect to a 2-Means is that the 0 centroid stays fixed the whole time. We have finally applied this algorithm to our data in order to better initialize the FCM. We will present the obtained result in next section.

Points c-d)

We have compared the FCM with a random initialization with the FCM initialized with the protypes provided by the AP algorithm getting the following:

Algorithm	number of iterations
FCM with random inizialization (c=3)	35
FCM with AP inizialization (c=3)	28
FCM with random inizialization (c=4)	88
FCM with AP inizialization (c=4)	40
FCM with random inizialization (c=5)	86
FCM with AP inizialization (c=5)	38

As we can see in the above table, the number of iterations required to the AP-FCM to converge is slightly smaller than those necessary to the randomly initialized version of the FCM. This holds for many values of the parameter c, clearly showing that the AP-FCM helps in improving the medium rate of convergence. For both of the applications of the fuzzy c-means algorithm the Python Toolbox uses the norm of the difference of the membership matrix within two consecutive iterations: the algorithm stops when this quantity is smaller than a tolerance ϵ, that we set equal to 0.01. Also a maximum number of iterations can be tuned to stop the algorithm, we set it to 1000. Applying the AP algorithm to better initialize the FCM and taking as threshold $\tau = 90$ we get 4 different prototypes, meaning that, a reasonable value of c could be 4. Moreover if we look at the following graph we can observe an elbow in the size of anomalous clusters from 4 to 5 confirming that taking c=4 could be a good initialization.



In order to get advantage of the PCA representation we recomputed all matrices and quantities of interest for the reduced dataset used throughout this section. The crisp representation of the fuzzy clusters obtained with the AP-FCM setting c=4 and finally projected in the 2D PC plane is given by:

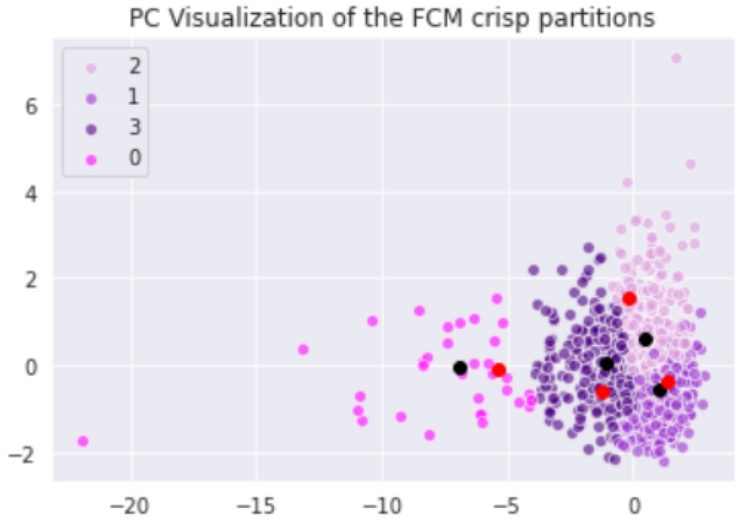


Figure 10: AP-FCM Crisp clusters

In addition to the clustered points we also plotted the initial prototypes in red (given by the AP) and the final ones in black. As the plot displays, the initial guess was quite near to the final solution, improving the speed of convergence. Moreover in order to better visualize the fuzzy partitions we plotted in a single figure the membership values associated to each partition.

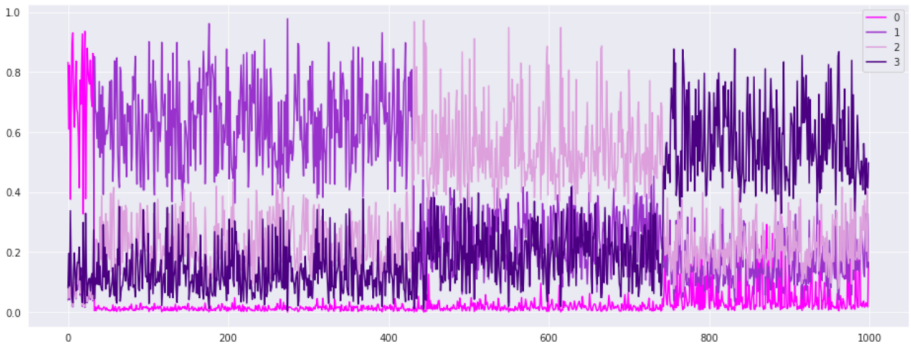


Figure 11: Fuzzy partitions with c=4

Point e)
We implemented and applied to the all the fuzzy c-partitions obtained in point a) the two following validation indexes:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^{m}||v_i - v_j||^2}{N min_{l\neq s}||v_l - v_s||^2}$$

Considering the membership degree and the structure of datasets, Xie and Beni proposed the XB index to measure the overall average compactness and separateness, and the smaller its value, the better the partition result.

$$PCAES = \sum_{i=1}^{c}\sum_{j=1}^{N}\frac{u_{ij}^2}{u_M} - \sum_{i=1}^{c} exp(-\frac{min_{k\neq i}||v_l - v_s||^2}{\beta}) \quad \text{where:}$$

$$u_M = min_{1\leq i\leq c}\sum_{j=1}^{N} u_{ij}^2 \ , \quad \beta = \frac{\sum_{i=1}^{c}||v_i - \overline{v}||^2}{c} \ , \quad \overline{v} = \frac{\sum_{j=1}^{N} x_j}{N}$$

Wu and Yang proposed the index PCAES by combining the normalized partition coefficient with the exponential separateness degree of each cluster, and the larger its value, the better the partition result. We now show the plot of both indexes when we let c range from 2 to 9.

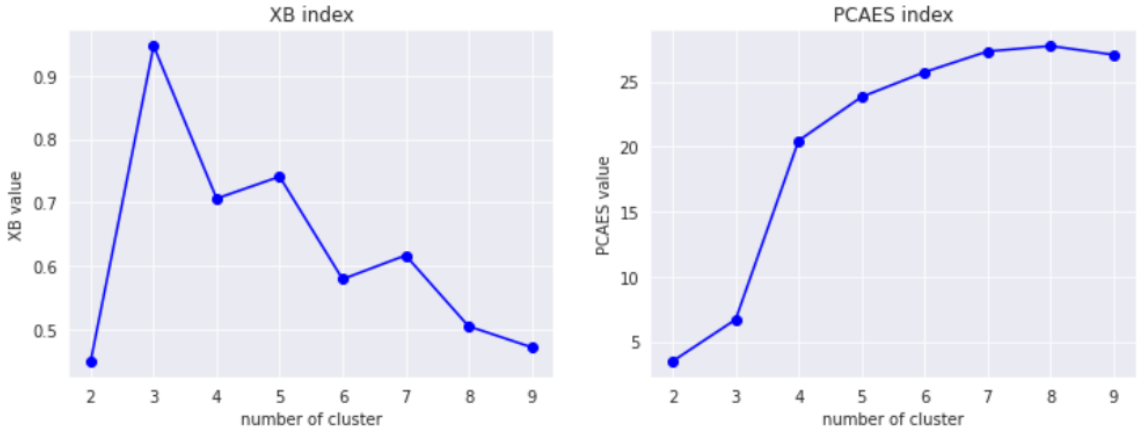


Figure 12: Validation indexes

From the two plots together we may conclude that the optimal number of cluster based on these indexes is around 8. Anyhow, keeping in mind the results obtained through the AP we notice that in correspondence of 4 clusters we have both a local minimum of the XB and a remarkable increasing elbow in the PCAES. This facts eventually led us to choose 4 as the most reasonable number of cluster for our problem confirming what we have previously observed with the AP-FCM. Finally we computed the value of these indexes for both the AP-FCM and the randomly initialize FCM with c=4. We summarize the results in this table:

Algorithm	Index value
XB_AP-FCM	0.7031
XB_random	0.7092
PCAES_AP-FCM	20.4158
PCAES_random	20.4579

Point f)
Taking advantage of the previous considerations made about the PC plot, we can affirm that the cluster 2 in figure (10) represent the observations that have an higher PC2 and thus are associated to more expensive houses, while cluster one will be associated to less expansive houses. For what concerns cluster 0 we have that its values will have a lower PC1, therefore it will be probably associated to bigger houses as observed before.

References

[1] Larose T., Larose C.. *Data Mining and Predictive Analytics*. Wiley Series on Methods and Applications in Data Mining, Wiley (2 nd edition), 2015.

[2] Zaki, M., and Meira, Jr, W. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press (2nd Edition), 2020.

[3] Mirkin, B. *Core Data Analysis: Summarization, Correlation, Visualization*. Springer Nature (2nd Edition), 2019.

[4] S. Nascimento. *Fuzzy Clustering via Proportional Membership Model*. Vol 119 of Frontiers of Artificial Intelligence and Applications, 2005.