DM Homework 3

Matteo Emanuele, matricula 1912588

December 2020

1 Exercise 2

In this implementation we have splitted our work in two, as asked by the homework. In the first part we have worked on clusterizing the raw data.
The feature of the basic dataset are the following:

1. *longitude*: A measure of how far west a house is; a higher value is farther west

2. *latitude*: A measure of how far north a house is; a higher value is farther north

3. *housingMedianAge*: Median age of a house within a block; a lower number is a newer building

4. *totalRooms*: Total number of rooms within a block

5. *totalBedrooms*: Total number of bedrooms within a block

6. *population*: Total number of people residing within a block

7. *households*: Total number of households, a group of people residing within a home unit, for a block

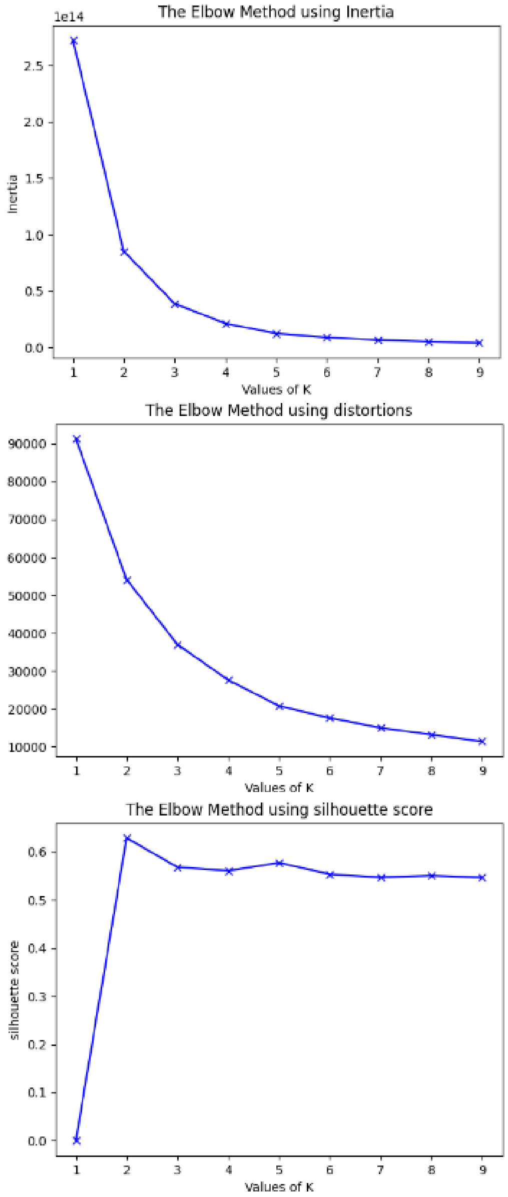
8. *medianIncome*: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

9. *medianHouseValue*: Median house value for households within a block (measured in US Dollars)

10. *oceanProximity*: Location of the house w.r.t ocean/sea

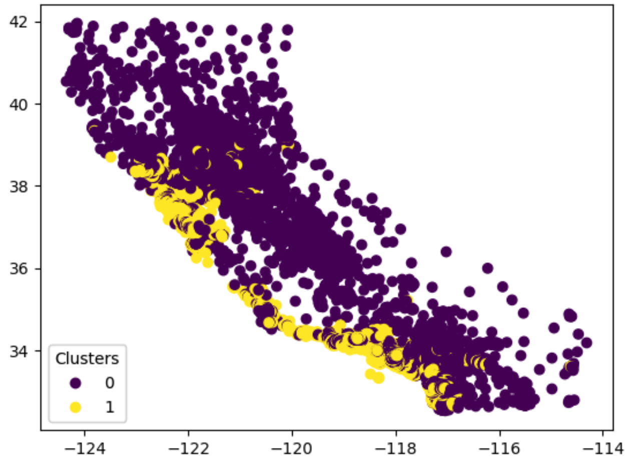
On the raw data we haven't really done anything except removing the NaN value(for technical purpose) and encoding the local_proximity categorical feature. Since this could have been considered processing of the data,it was implemented a parameter that if placed to False, would have deleted the feature,

making the clustering work on 9 features instead of 10.
The results are the following:

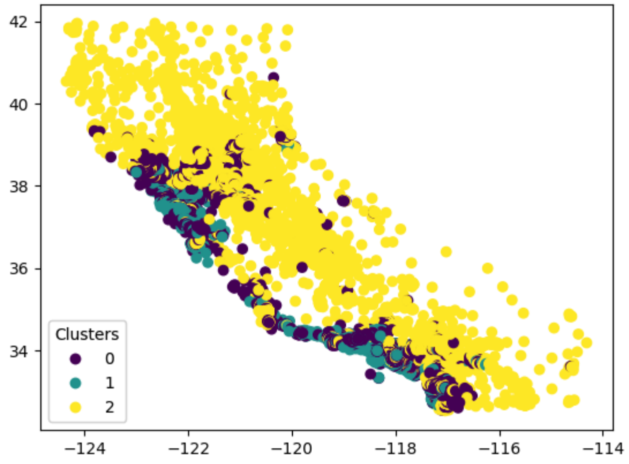


From these plots we can understand that the optimal value are either k=2(suggested by the silhouette score), or k=3(elbow tip).

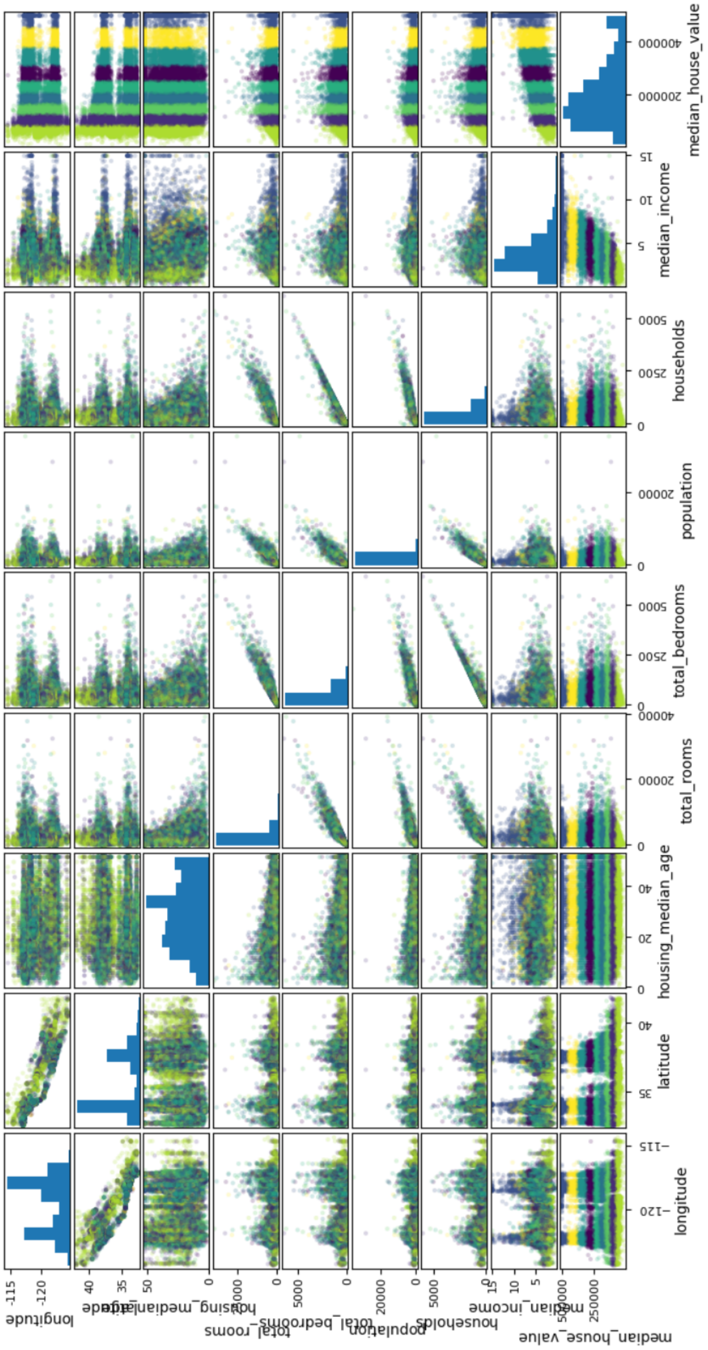
for k=2:



For k=3:



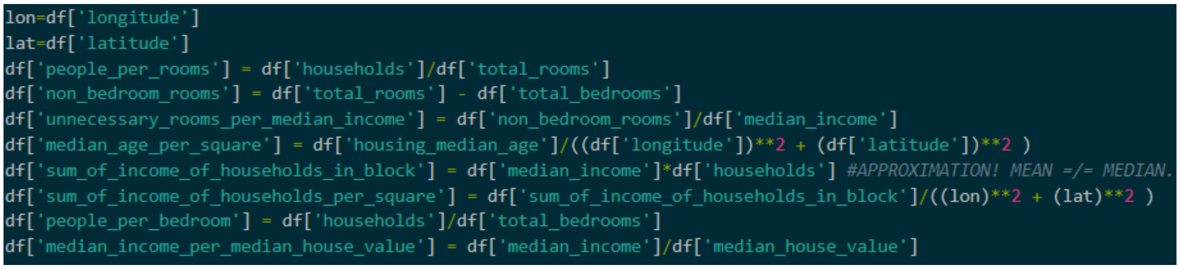
It's memorable of mention the fact that either the ocean proximity is kept, the result doesn't change.
The clusters seem to form in a way which is linked to the welfare of the region,but it is extremely dirty, since the data are not elaborated.
Here we report the scatter matrix with the respective color of plots:



The dimension of the plot were too big to be inserted horizontally.
It is possible to notice how the only feature that got clusterized significantly

are the one related to the median income, the median house value, and the geographic position.
This is not surprising! What we have discovered is that, in the California Housing Price dataset, information related to the wealth of a block are significant.
This will be the direction in which we will do feature engineering then.

Now will now be reported the work done on the **engineered data.**
The target was to polish the way the data got clusterized in the previous attempt. We want to strengthen the clusterization based on the wealfare informations. So we need to work on the feature to try to achive this.
note: *the feature built and showed now in the following lines, are not objectively the best. They were simply the one that I thought of while I was trying to engineering the data.*



```
lon=df['longitude']
lat=df['latitude']
df['people_per_rooms'] = df['households']/df['total_rooms']
df['non_bedroom_rooms'] = df['total_rooms'] - df['total_bedrooms']
df['unnecessary_rooms_per_median_income'] = df['non_bedroom_rooms']/df['median_income']
df['median_age_per_square'] = df['housing_median_age']/((df['longitude'])**2 + (df['latitude'])**2 )
df['sum_of_income_of_households_in_block'] = df['median_income']*df['households'] #APPROXIMATION! MEAN =/= MEDIAN.
df['sum_of_income_of_households_per_square'] = df['sum_of_income_of_households_in_block']/((lon)**2 + (lat)**2 )
df['people_per_bedroom'] = df['households']/df['total_bedrooms']
df['median_income_per_median_house_value'] = df['median_income']/df['median_house_value']
```

Here we will try to explain the reasoning behind the most debetable. 1. *non bedroom rooms*: is the amount of rooms that are not dedicated to sleeping. A house with high number of rooms which are not bedrooms **is indicating a luxury status: multiple bathrooms, game rooms, halls... these are all indicators for a high wealth status**.

2. *unnecessary rooms per median income*: A measure of the luxury of a household given their median income.

3. *median age per square*: a combined measure on the territory to express the age of a block. Newer blocks tends to be in the newer parts of the regions,which is also the one more devoted to economic growth.

4. *sum of income of households in a block*: another good measure for the wealth status of a household. We computed these **approximating the mean with the median. It's not optimal,but for high level analysis this can work.**

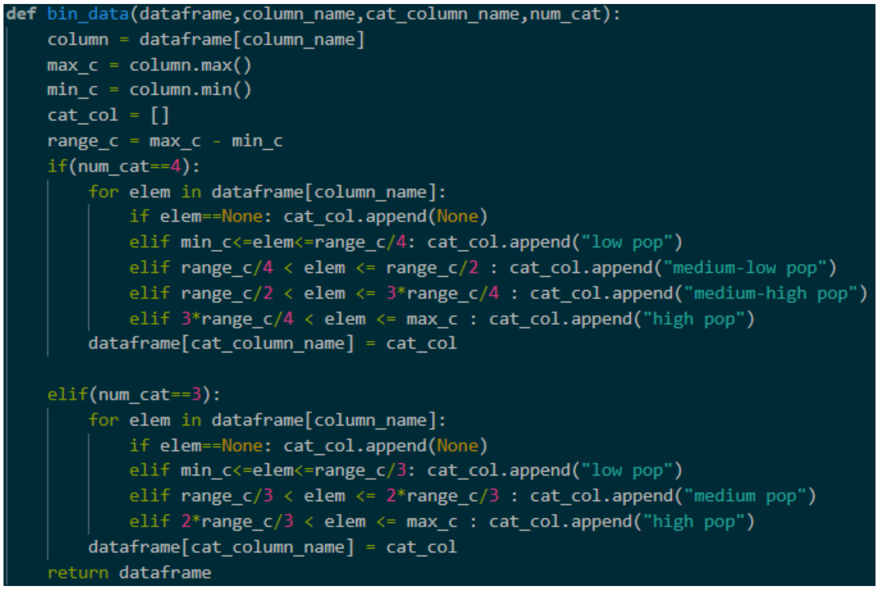
5. *sum of income of households per square*: another data created to express the welfare of households but related to the regionality, since the position on

the state is also important(west coast cities are well known richer than the others in california).

6. *median income per median house value*: this value is indicating how is the income of a median with respect to the house in which they are living. If this value is high, it means that the people there use to earn a lot in comparison with the house in which they are living.

7. *people per bedroom*: having more people sleeping inside a single bedroom, can express a wealth situation that can be interesting. For istance, poor families will be more inclined to have their sons to sleep all togheter, for space reasons.

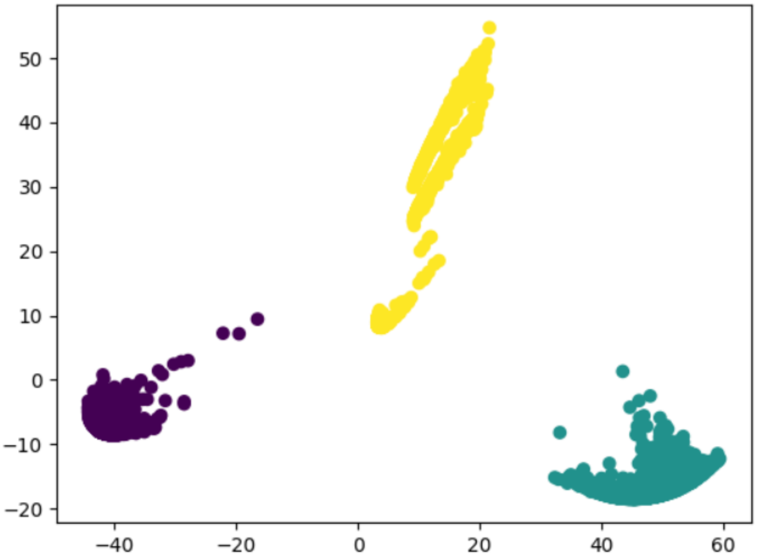
Another manipulation of the data that we did is related to the population. Instead of using the population as a number,we binned the feature to become a categorical feature. We did as such through the following function:

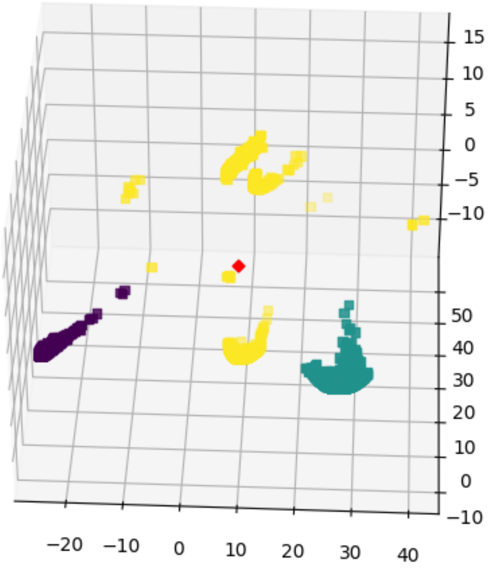


```
def bin_data(dataframe,column_name,cat_column_name,num_cat):
    column = dataframe[column_name]
    max_c = column.max()
    min_c = column.min()
    cat_col = []
    range_c = max_c - min_c
    if(num_cat==4):
        for elem in dataframe[column_name]:
            if elem==None: cat_col.append(None)
            elif min_c<=elem<=range_c/4: cat_col.append("low pop")
            elif range_c/4 < elem <= range_c/2 : cat_col.append("medium-low pop")
            elif range_c/2 < elem <= 3*range_c/4 : cat_col.append("medium-high pop")
            elif 3*range_c/4 < elem <= max_c : cat_col.append("high pop")
        dataframe[cat_column_name] = cat_col

    elif(num_cat==3):
        for elem in dataframe[column_name]:
            if elem==None: cat_col.append(None)
            elif min_c<=elem<=range_c/3: cat_col.append("low pop")
            elif range_c/3 < elem <= 2*range_c/3 : cat_col.append("medium pop")
            elif 2*range_c/3 < elem <= max_c : cat_col.append("high pop")
        dataframe[cat_column_name] = cat_col
    return dataframe
```

We turned the value of population into 4 categories.
After this,we decided to apply PCA. As needed by PCA, we first normalize the data using the sklearn.preprocessing package. During the plots of the pca clusters,they seemed to be ”little bars”, all stucked and with a high density. So to have a better rappresentation, the cosine similarity of the data was computed. Doing so, we killed the influence of the value to focus the plot only onto the relations between data. Then, the PCA was applied in 2 and 3 principal component. Here are reported the plots of the clusters. **PCA plots with cos similarity with number of clusters=3 :**





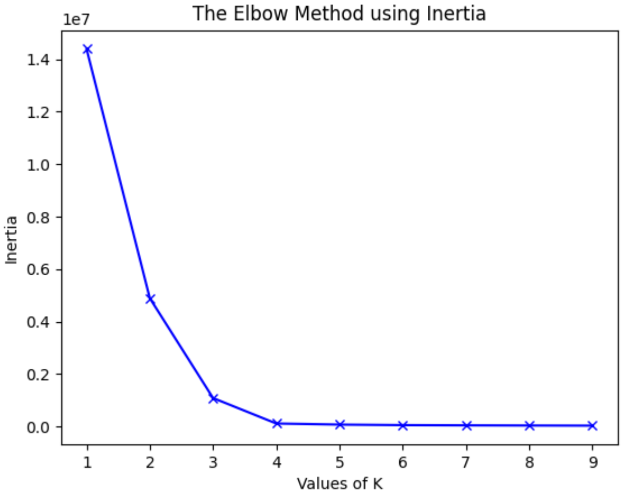
The clusters have obviously been separated.
Now the big question: **is this clusterization meaningful?**
We decided to evaluate the clusters in two ways: silhouette score and graphical analysis. The silhouette score is a meter for the goodness of a clusterization. It

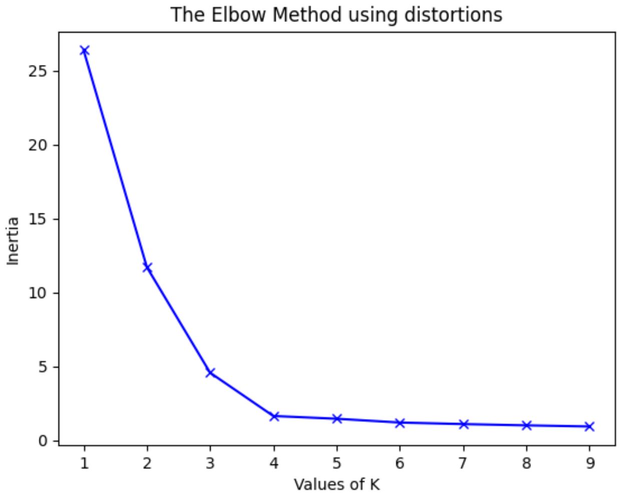
ranges from -1 to 1, where -1 means that the clusters are bad, 0 means that the clusterization is not interpretable,while values towards 1 are technically very good clusterizations.

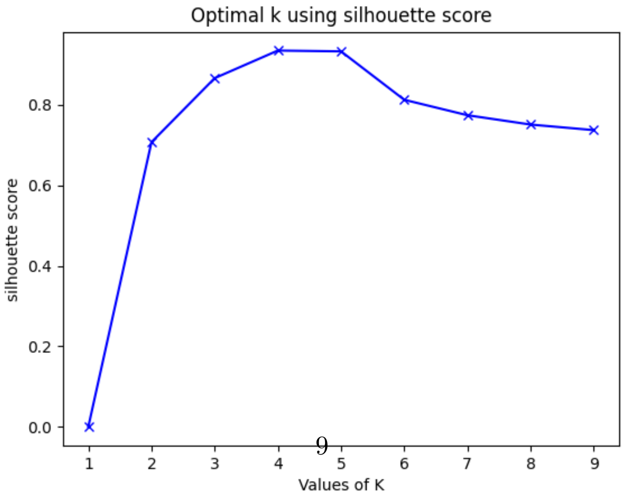
Working on the dataset we learnt that we can create features doing any kind of combinations of the starting informations, but they are not always meaningful. They can lead us to clusterize mathematically the data,but these clusters WILL NOT ALWAYS BE MEANINGFUL. This means that for those clusterization, the silhouette score will have very high values even though a different number of clusters may be more interpretable, and in fact this is exactly the case of our work.

When the clusters seems to have sense then? The features that we built are all welfare-related, so we expect to find some kind of interpretation.

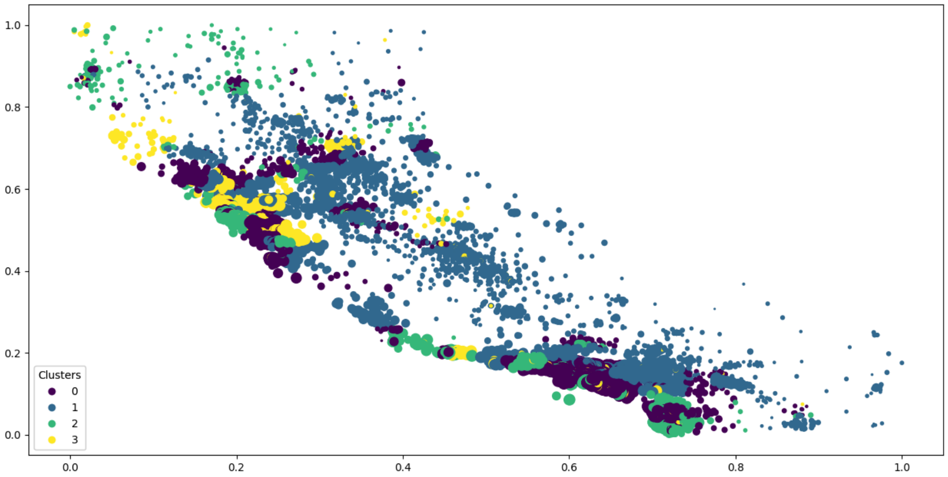
Before to move on I want to underline a reasoning about the graphical result of PCA. On the 3D representation it i almost obvious that there are actually 4 clusters. The number 4 seems to be also the highest silhouette score, and it is also suggested by the elbow method. Here are reported the graphs:


The Elbow Method using Inertia
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Inertia
1
2
3
4
5
6
7
8
9
Values of K

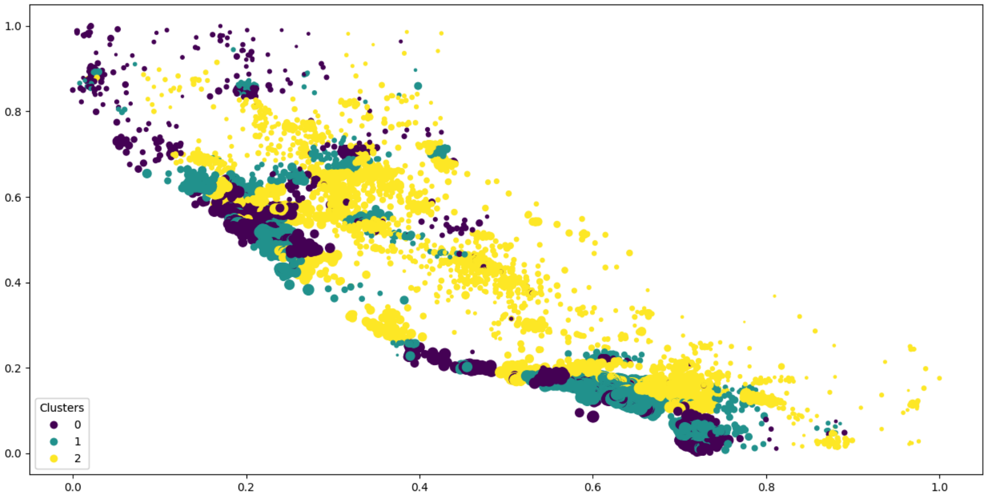

The Elbow Method using distortions
25
20
15
10
5
0
Inertia
1
2
3
4
5
6
7
8
9
Values of K


Optimal k using silhouette score
0.8
0.6
0.4
0.2
0.0
silhouette score
1
2
3
4
5
6
7
8
9
Values of K

The silhouette scores hit skyrocking values for k=4, in particular its equal to 0.951. A visual interpretation of the clusters can be done evaluated through the plotting on the california map built with the longitude-latitude data. Here I report the map for k=4:



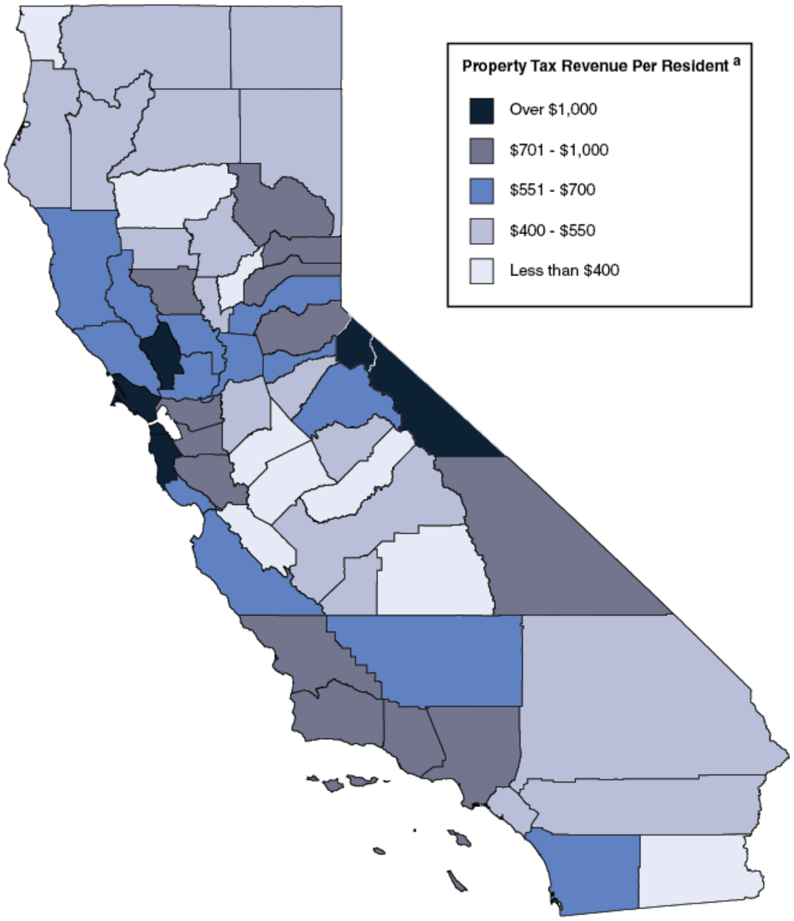
As we can see, the patterns is extremely confusing, in opposition to the result with k=3. From now on we will only focus our analysis on the 3 clusters extracted with 3 principal component.
with the following data one we obtained this clusterization,with the following feature distribution:



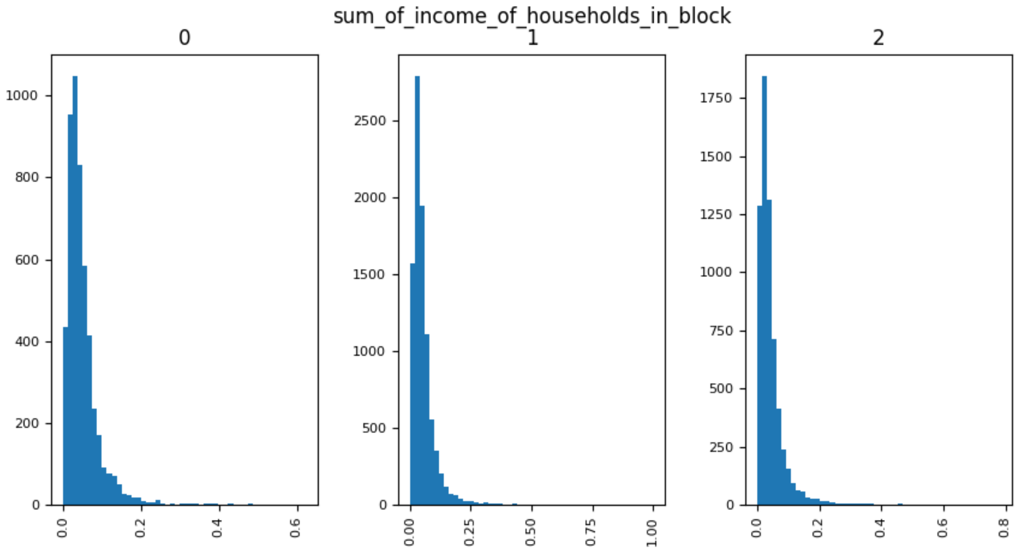
The pattern shown seems similar to the one in the previous clusterization on the raw data,but with significant differences. Two of the three clusters seems to be distributed on the richest places of the California, which are los Angeles, San Diego,and so on.
Doing some research on the internet, a map of the california about tax revenue per resident was found,and it is displayed on the next page. In the next pages are reported also the feature distributions for each cluster.

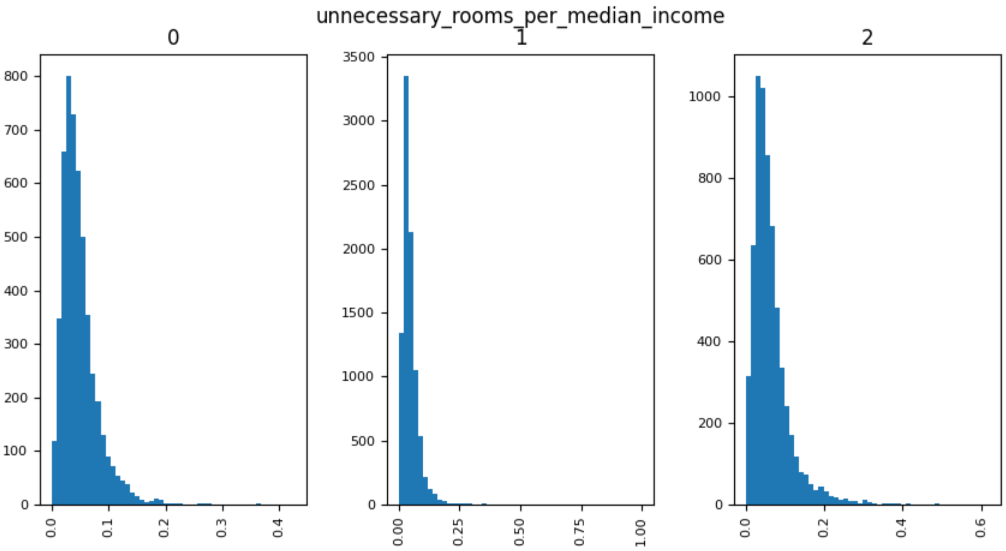
Local Governments' 1 Percent Tax Revenue Varies Widely

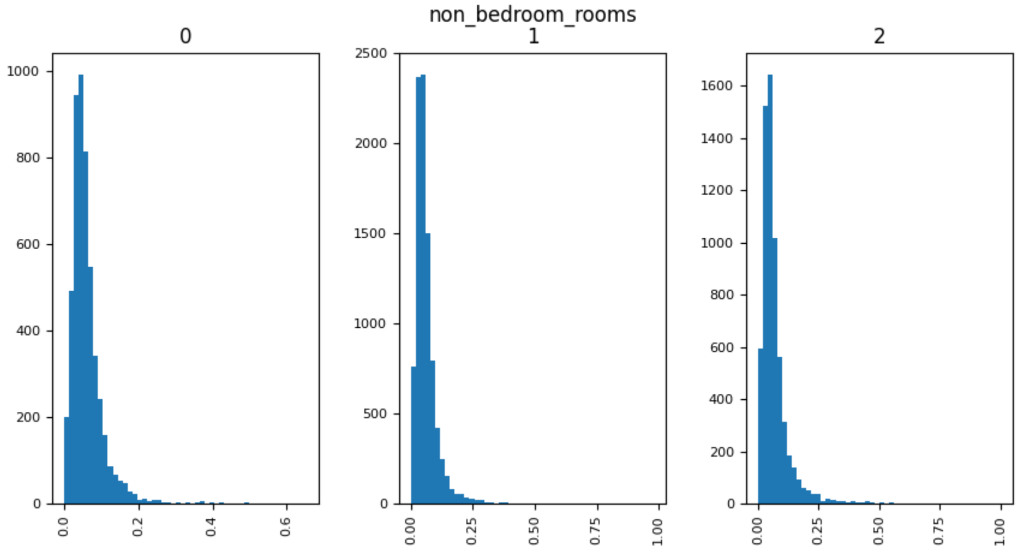
Per-Resident Property Tax Revenue For Counties, Cities, and Special Districts

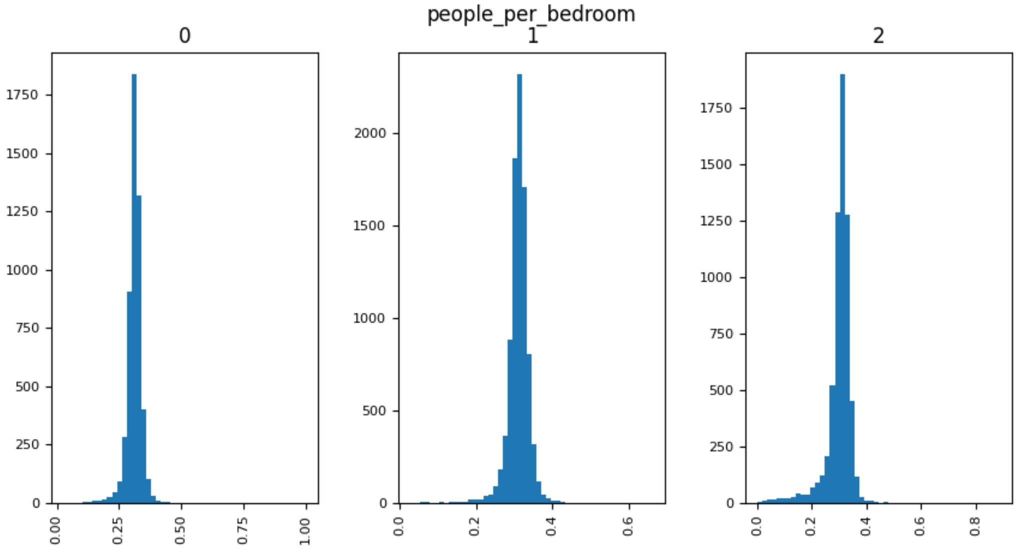


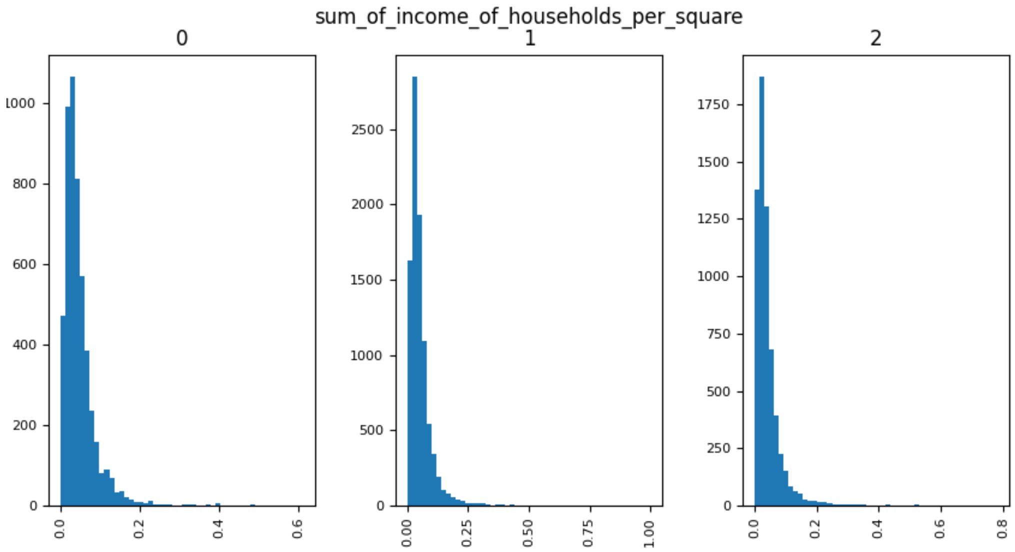
[a] Reflects LAO estimates for per-resident combined total property tax revenue for the county and the cities and special districts within the county. Estimates exclude redevelopment debt payments. Estimates assume all redevelopment debts are paid with property tax revenue from the 1 percent tax.

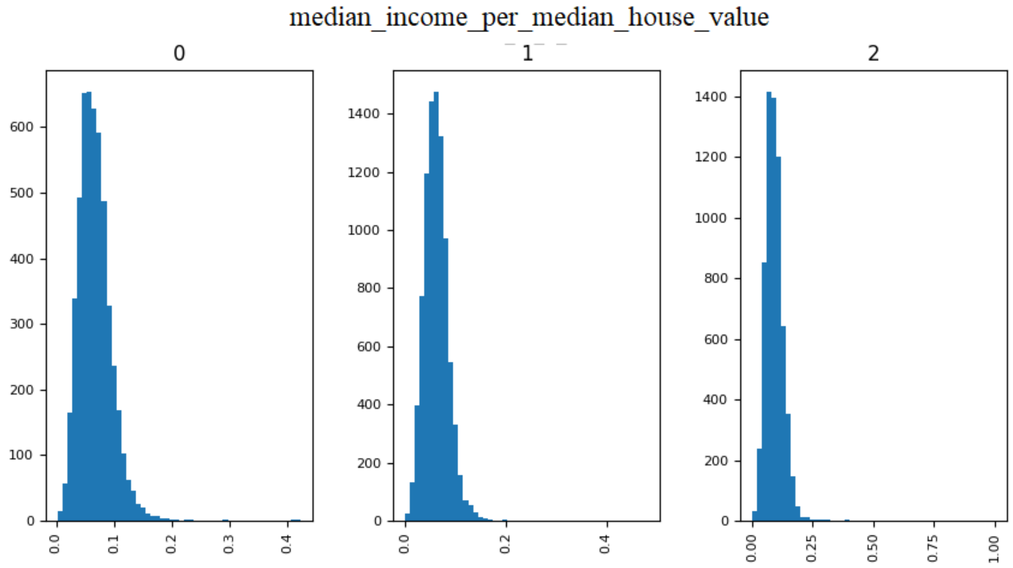

sum_of_income_of_households_in_block
0
1
2
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0
500
1000
1500
2000
2500
0.00
0.25
0.50
0.75
1.00
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8


unnecessary_rooms_per_median_income
0
1
2
0
100
200
300
400
500
600
700
800
0.0
0.1
0.2
0.3
0.4
0
500
1000
1500
2000
2500
3000
3500
0.00
0.25
0.50
0.75
1.00
0
200
400
600
800
1000
0.0
0.2
0.4
0.6


non_bedroom_rooms
0
1
2
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
2500
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
1600
1400
1200
1000
800
600
400
200
0
0.00
0.25
0.50
0.75
1.00


people_per_bedroom
0
1
2
1750
1500
1250
1000
750
500
250
0
0.00
0.25
0.50
0.75
1.00
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
1750
1500
1250
1000
750
500
250
0
0.0
0.2
0.4
0.6
0.8





Let's discuss about the final interpretation. If we look at the data, we can recognize basically that the three clusters can be identified as:
cluster 0: lower class blocks and households
cluster 1: high class blocks (very rich households)
cluster 2: middle class blocks (medium-lower to be precise)

These interpretation is, in my humble opinion, justified by the distributions. The cluster 0 presents the lowest unnecessary room per median income among all the clusters. This means that, overall, these blocks are composed of smaller

house than average if compared with the median income the people in it. On the other hand, people on cluster 1 have triple the expected value of this distribution , which is underline how these houses are very luxurious for the median income of the blocks. Also they are less distributed.
This exact reasoning can be done also for the feature "sum of income of household in block", where the reasoning applies perfectly. The unnecessary rooms("non bedroom rooms") is also reinforcing this analysis, being the people in cluster one those who lives in block with the highest numbers of rooms not dedicated to sleeping, which is a symbol of wealth as we already said.
We noticed how the map of tax revenue seems to be very similar in distribution to the clusterization that we obtained, which is promising. Although the poorest cluster and the wealthiest cluster seems to be overllapped on the coast, this could be interpreted as high social inequality(" Economic inequality in Los Angeles is driven by an expanding population at the bottom of the income and wealth distributions together with a growing share held by those at the top". Source: https://knowledge.luskin.ucla.edu/wp-content/uploads/2018/01/Haynes-Report_WideningDivide_Ong_UCLA_1.3.2017.pdf).

To conclude we managed to separate the data we have at the start into 3 clusters, with an overall satisfying result.