Homework 3

Matini Gabriele, 1934803

1 Exploratory Data Analysis of the dataset

The dataset is composed by 10 features of the housing prices, being latitude, longitude positions, housing median age, total rooms, total bedrooms, population, households (families), median income, and median house value, for a total of 20640 rows. Of these, 207 have NULL total bedrooms value, and they will not be considered in the study, since their contributions would not change the final results (they constitute 1% of the dataset).

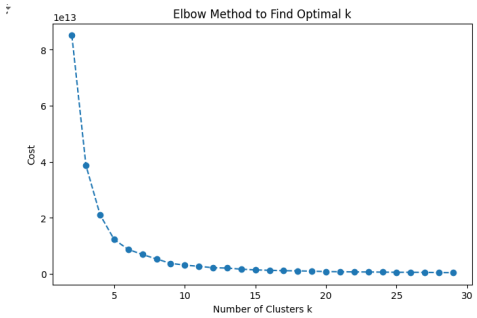
Plotting the distributions reveals that most of them are normal distributions, except for some anomalies, such as median housing value, which has a high number of points at the end of the distribution, and latitude and longitude that look bimodal. We also study the correlation of all variables, finding out that latitude and longitude seem to have a very high anti-correlation (-0.9), while total rooms, total bedrooms, population and households are all highly correlated between each other, as we could expect (all above 0.8). Finally, housing median income and housing median value are somewhat correlated (0.7 circa).

2 Clustering on Raw Data

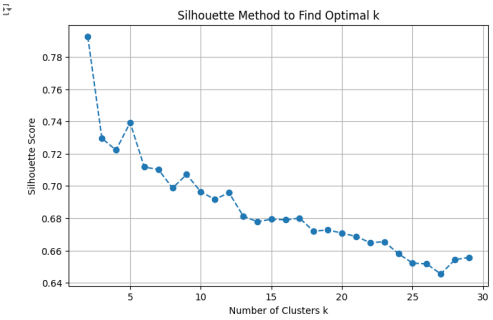
K-Means++ was used as an algorithm. We use the elbow method, that plots a cost calculated as within-cluster squared sums (WCSS) for values of k from 2 to 30, which is a measure of how compact the clusters are. We also use the silhouette method, which for each point i with its cluster C_k, and for C_m nearest cluster to i such that $k \neq m$, calculates this quantity:
$S(i) = \frac{Cohesion(i) - Separation(i)}{max\{Cohesion(i), Separation(i)\}}$, with $Cohesion(i) = \frac{1}{|C_k|-1} \sum_{j=1 j \neq i}^{|C_k|} d(i,j)$ $Separation(i) = \frac{1}{|C_m|} \sum_{j=1}^{|C_m|} d(i,j)$
. In our evaluation, we take the average of all Silhouette values of all points. The Silhouette method provides a measure of how well defined the clusters are, for a value that can go from -1 to 1. Both of these metrics will provide a useful framework throughout the study, and the raw data plotting will serve as a baseline for improvement with feature eng techniques as well as the registered time to run K-Means++.
Other than the silhouette score and the elbow method, which are the primary ways we evaluate clustering quality, we also use the visualization of clustered points projected onto a 2D space. This visualization is achieved via an additional application of PCA for dimensionality reduction. It is important to note that this 2D representation does not directly reflect the original clusters as computed but serves as an approximation. While the transformations applied to the data are evaluated through the silhouette score and the elbow method, the plotted clusters are derived from a separate PCA step and should be interpreted with this limitation in mind.
From the analysis of raw data graphs, it's clear that the WCSS is too high, in the order of billions and trillions, while the silhouette score seems ok. The time to cluster was $\sim$ 13 seconds, although it can go up to $\sim$ 16. We will use feature engineering to improve from here.

3 Feature Engineering

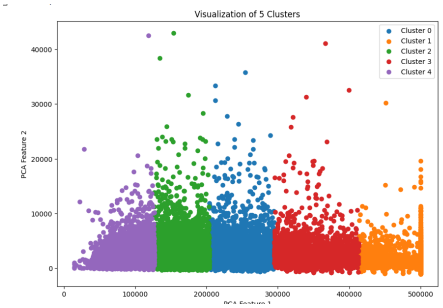
To discover feature engineering techniques, we adopt an incremental experimentation approach, where techniques are systematically tested and evaluated. If a technique demonstrates improvements in clustering performance, as measured by either of the two metrics, it is incorporated into subsequent experiments. Before listing the experiments, it should be noted that the ocean_proximity field was the only categorical field and was thus one-hot encoded. Please note also that all performances talked about here are documented in the graphs plotted and saved in the elbow and silhouette folders.



(a) As we can see, the cost for values is in the order of trillions, going in the order of billions for k ¿ 5.



(b) We can see a certain complementarity with the WCSS cost, as the greater k is, the less is the cost, however the clusters become redundant and less well defined and the silhouette goes to 0.



(c) Showing clustering for k=5. The clustering is done by applying PCA in order to project the points onto the spaces with the two most contributing dimensions in terms of variance, which define the overall geometry. In this way we are able to plot an n-dimensional point into a 2D space.

Figure 1: Analysis of cost and silhouette score for clustering on raw data.

Standardization We try to emphasize the scales of some of the less important variables through standardization. The reasoning for this stems from the fact that the median_house_value field has an extremely high variance, and thus dominates in the Euclidean space when computing K-Means++. We can see this, by conducting a study on the raw data, where we remove each time one feature and see how the PCA clustering plot changes: it changing significantly is a sign that we found the important feature. What our standardization does, in particular, is centering the data around the mean and dividing by the standard deviation, performing a stdev based normalization. Results: we have the expected effect of lowering the WCSS, from trillions all the way down to 140000 (indicating a much better compactness), however the silhouette score degraded a bit, because one dimensions is not dictating the whole clustering anymore. Since standardization worked well on the dataset, the following experiments will be tried using it.

Spatial features removal and Decorrelation We remove latitude and longitude or decorrelate them in two separate sub experiments. Latitude and longitude may not be so important to cluster the points, as such we might want to remove them to improve computation time without impacting performances. At the same time, latitude and longitude have a very high anti-correlation, so we also see what happens if we decorrelate them by coupling the two into a single feature. Decorrelation was achieved by applying PCA to project the two features onto a single principal component that captures the majority of the variance between them. Note that the dataset in general does not have very important spatial patterns, since only ocean proximity and these two features say anything about the space dimension. Results: decorrelation does no noticeable improvement (which suggests again that the spatial features are not really withing the groupings and patterns of the dataset), while removing latitude and longitude improves slightly WCSS, indicating that probably latitude and longitude don't offer much in the way of meaningful information for clustering. We will thus make a case for removing the spatial features from the study: latitude, longitude and ocean_proximity. As a sidenote, a third

subexperiment was tried by decorrelating the four features that are highly correlated (population, household, total_rooms and total_bedrooms): the various combinations tried revealed just a slow but progressive worsening of the silhouette score.

Feature removal We find out what other features we can remove to sharpen k-means results. It turns out many of them are quite repetitive, others are just useless and they damage the clustering by adding dimensionality. In particular, latitude, longitude, ocean_proximity and housing median age just do not provide enough information compared to the dimensionality problems they add, while others such as total bedrooms, population and households are almost redundant due to their strong correlation between each other and the total rooms field, so they end up adding very little useful information to the overall dataset, as total rooms in some way represents them. In the end, we end up with 3 dimensions, which improves massively clustering: from the elbow curve is clear that, for the same values of k the cost decreases (from 200000 for k=2 we get to less than 40000, for example), and the silhouette curve slightly improves too.

Dimensionality reduction with PCA we use PCA in order to identify top k most important dimensions to get 90% of the cumulative variance. It turns out that k=3, since the contributions are around 56% for the first one, second is 24%, third is 12%. The results are not as good as the raw features removal method, however we are using a less arbitrary measure and trying to coalesce every dimension, that would otherwise get entirely discarded, into a linear combination. In the next experiment, we will thus try to use PCA too along with whatever experiment engineering.

Integrate new features Integrate new features by having the old ones add information instead of being redundant. We start by integrating difference between rooms and bedrooms and various ration between population, households, total_bedrooms and total_rooms. Their integration worsens the results.

Taking the square of some features We square certain features to amplify their variance, making them potentially more influential in clustering, and see the results both when we replace and we don't replace the old features. This approach yields improvements in clustering metrics when we replace the old features with the squared ones: both the WCSS decreases (indicating tighter clusters), and the silhouette score increases, reflecting better-defined clusters. These improvements can be attributed to two factors: first, squaring emphasizes larger values, pushing outliers further away from other points, making them more likely to form distinct clusters. Second, squaring increases the variance of these features, giving them greater weight in clustering computations. However, there are tradeoffs. Examining the correlation matrix shows that squaring weakens correlations between features, which may suggest less interpretability. Additionally, squaring alters the relationships between features, potentially introducing non-linear biases that could affect downstream analyses or the generalizability of the clusters.

Multiplying features with the dominant feature This approach is conceptually similar to squaring the features, but it was explored to determine if it could provide better results. Unfortunately, the improvement was not as significant as anticipated. While there was a modest enhancement in terms of WCSS, the results were less impactful compared to squaring the features. Consistent with the previous experiment, removing the original features after transformation yielded better results than retaining them.

3.1 Conclusions

From the experiments, we can take that standardization of data is key, along with PCA that helps to eliminate noise and irrelevant features, enabling a clearer understanding of the underlying patterns within the dataset that clustering aims to uncover. Squaring the features can enhance clustering results, although it may introduce nonlinearity into the data. Removing the original features after transformation generally leads to better results, possibly by preventing redundancy and reducing noise. With respect to time, only removing features (either with PCA or not) seems to change time noticeably, likely due to having less dimensions (we go from around 13 seconds to around 3 or 4).