CITS5508 Machine Learning - Assignment 3

Franco Meng 23370209

Outline:

This report is aimed at exploring the California Housing dataset using various machine learning techniques, including Principal Component Analysis (PCA), clustering, and supervised learning algorithms. The goal is to analyse and model housing prices in California based on different features:

The dataset has 20,640 rows and 10 columns (eight numeric variable, one categorical variable and one target variable).

The variables of the dataset are:

- longitude: district group longitude.
- latitude: district group latitude.
- housingMedianAge: median house age in the district.
- totalRooms: the total number of rooms in the district.
- totalBedrooms: the total number of bedrooms in the district
- population: district population.
- households: the total number of households in the district.
- medianIncome: median income in the district.
- oceanProximity: whether each district is near the ocean, near the Bay area, inland or on an island (categorical).
- medianHouseValue: the median house value (target variable).

Deliverables:

D1 (a) Plot the histograms of the non-categorical features and the target in a grid subplot.

- Answer: fig 1.

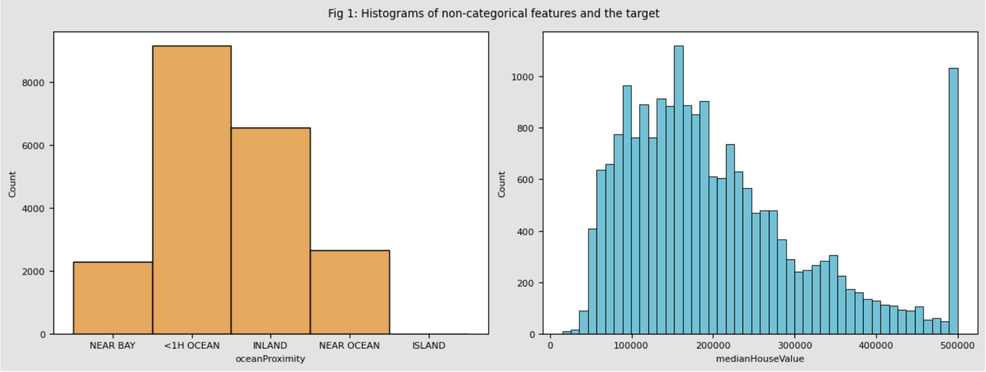


Fig 1: Histograms of non-categorical features and the target

D1 (b): Compute the correlation matrix of all features (including the target features). Do not use the categorical variable (ocean proximity). Describe which features are more correlated.

- Answer: fig 2.

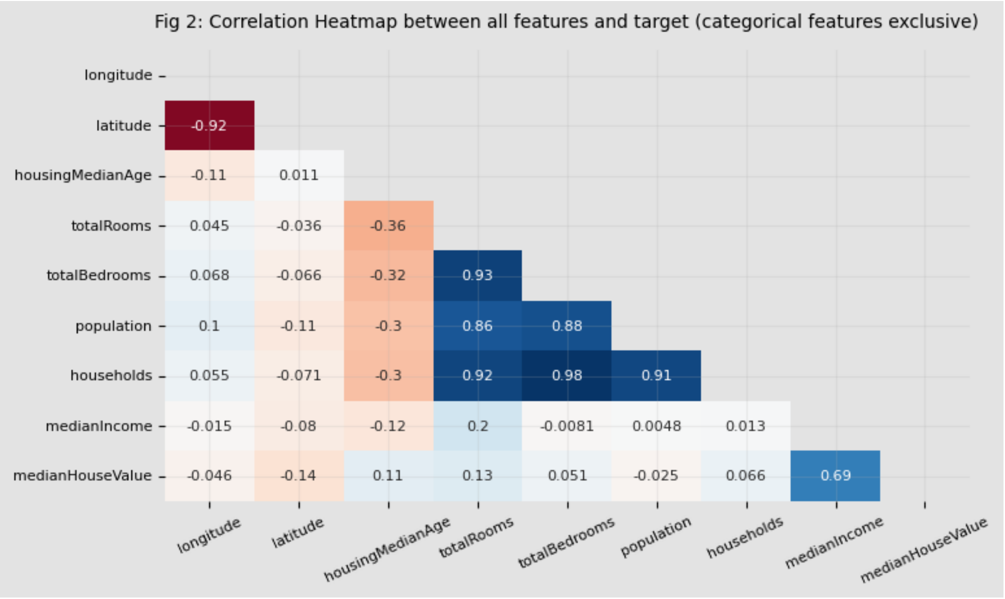


Fig 2: Correlation Heatmap between all features and target (categorical features exclusive)

- Comment:

  The highly correlated features :

  All 6 pairwise combinations between **: TotalRooms, TotalBedrooms, Populatation, Households**

  It makes lots of sense where these above features are highly correlated, as the total rooms, total bedrooms would naturally be strongly correlated with households, and population (which is the district population).

  More people mean more households, then with more demand for number of total rooms.

D1 (C): Present a scatter plot for each variable, displaying the corresponding variable on the x-axis and the target variable on the y-axis.

- Answer: fig 3

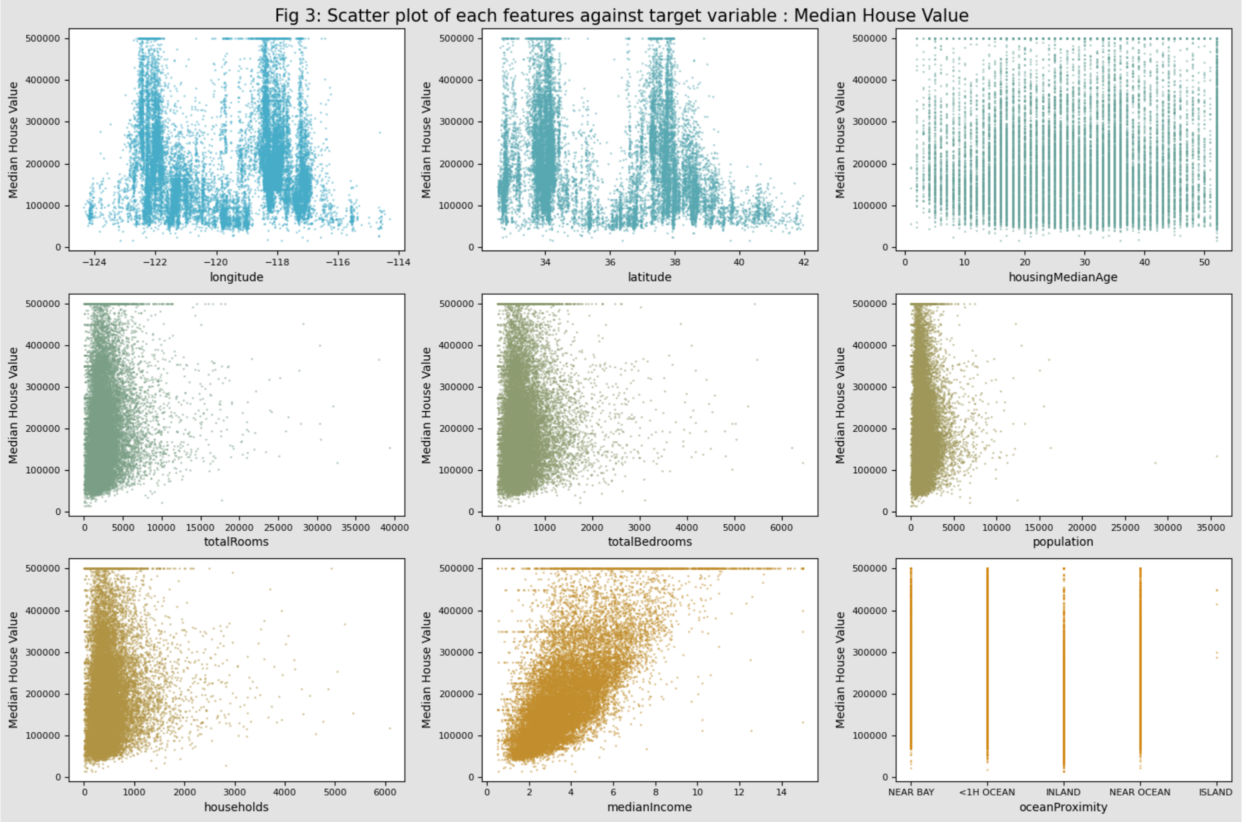


Fig 3: Scatter plot of each features against target variable : Median House Value

D2(a): In a table, report the RMSE for the training and test sets for the two models for each dataset. That is, your table should contain four rows with four values each.

- Answer: Table 1.

Table 1: RMSE for different models/data set

	Linear Reg (Training)	**Linear Reg (Testing)**	**Lasso Reg* (Training)**	**Lasso Reg* (Testing)**
Data 1	68607.31	68589.31	68660.50	68601.81
Data 2	0.69	0.69	1.13	1.12
Data 1 (Scaled)	68607.31	68589.31	68666.24	68622.01
Data 2 (Scaled)	0.69	0.69	1.16	1.14

*Alpha in Lasso Models are set to 100.

D2(b): Discussion

- Answer:
  As we can see from the table above, the scaling of the data didn’t affect the RMSE for linear models, regardless the unit differences in target variables in Data 1 and Data 2.
  The linear regression models are scale-invariant, when the input features are scaled, it will change the parameters/weights of each feature, but the RMSE will remain the same.

However, there are slightly differences in Lasso Regression models. As the effect of hyperparameter alpha, which is the strength of penalty, are directly related to the scale of the features. After the standardisation, alpha would have the same effects across all features.
The differences shown in the table may not be obvious as the Alpha = 100 is just a random constant, without fine tuning to find the best effective alpha.

D3 (a): Report the RMSE for the two models' training and test sets in a table. Your table should contain two rows with four values each.

- Answer: Table 2.

Table 2. RMSE for different model and dataset

	Linear Reg (Training)	**Linear Reg (Testing)**	**Lasso Reg* (Training)**	**Lasso Reg* (Testing)**
Data 3	0.71	1.14	1.16	1.14
Data 3 (Scaled)	0.71	1.14	1.16	1.14

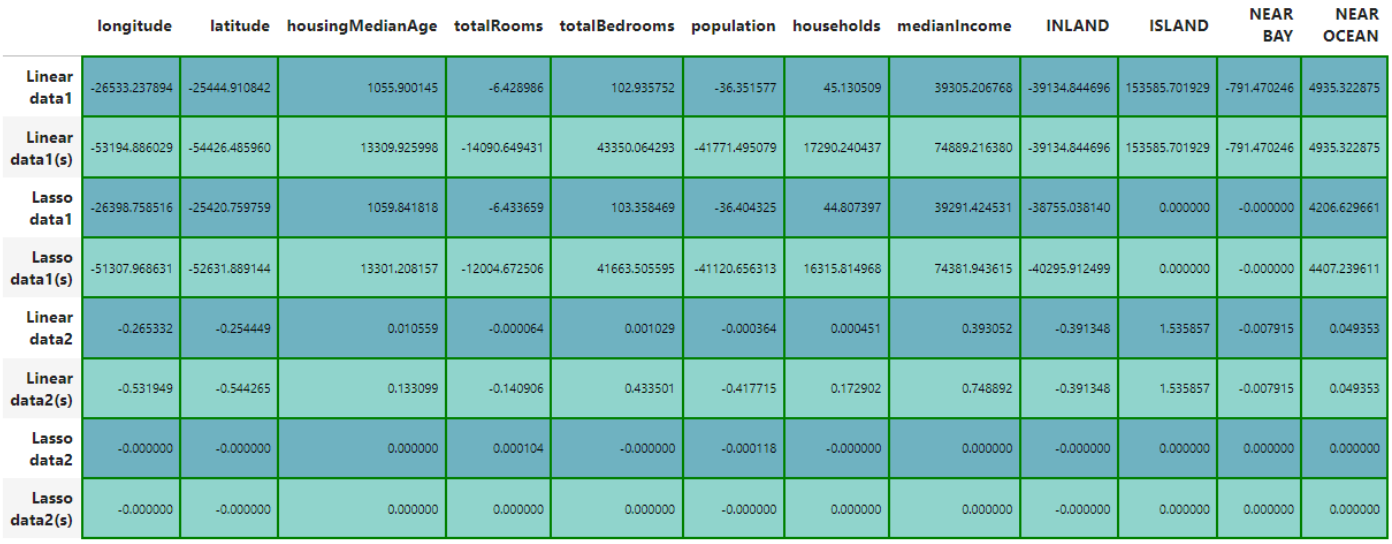
D3 (b): Discussion.

- Answer: as the table shows, all RMSE remains the same between Data 3 and Data 3(scaled). By transforming those four highly correlated columns into three new columns, it has improved the data by standardizing these columns in different way (all divided by households). The features are now may already have similar scales, therefore the standardisation may have minimum effect.
Also the alpha = 100 may be too large, where all the weights are penalised to zero, apart from bias term. It is interesting / coincidence to see the RMSE on testing set remains, regardless of the regularisation.

D3 (c): Report the estimated parameter values with the corresponding variable names for all models (12 in total, eight from D2 and four from D3)

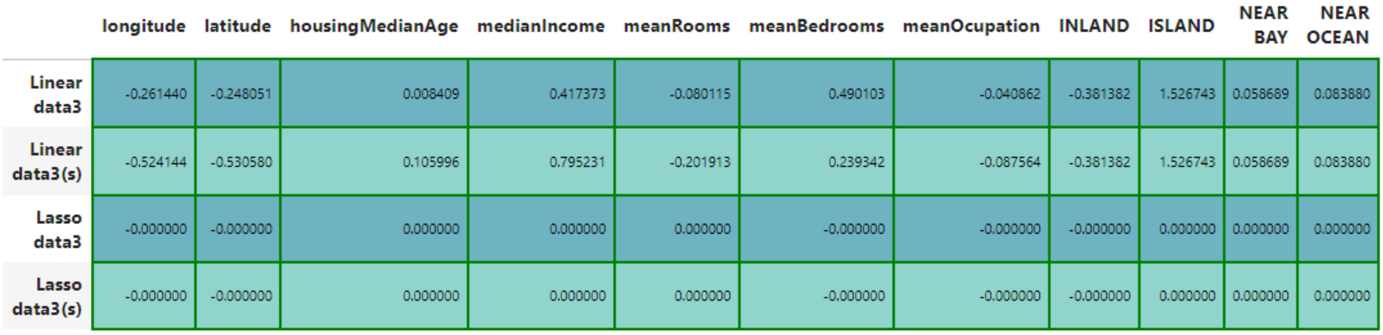
- Answer: Table 3 & 4.

Table 3 Reported Weights for all Features / Models in D2



	longitude	latitude	housingMedianAge	totalRooms	totalBedrooms	population	households	medianIncome	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Linear data1	-26533.237894	-25444.910842	1055.900145	-6.428986	102.935752	-36.351577	45.130509	39305.206768	-39134.844696	153585.701929	-791.470246	4935.322875
Linear data1(s)	-53194.886029	-54426.485960	13309.925998	-14090.649431	43350.064293	-41771.495079	17290.240437	74889.216380	-39134.844696	153585.701929	-791.470246	4935.322875
Lasso data1	-26398.758516	-25420.759759	1059.841818	-6.433659	103.358469	-36.404325	44.807397	39291.424531	-38755.038140	0.000000	-0.000000	4206.629661
Lasso data1(s)	-51307.968631	-52631.889144	13301.208157	-12004.672506	41663.505595	-41120.656313	16315.814968	74381.943615	-40295.912499	0.000000	-0.000000	4407.239611
Linear data2	-0.265332	-0.254449	0.010559	-0.000064	0.001029	-0.000364	0.000451	0.393052	-0.391348	1.535857	-0.007915	0.049353
Linear data2(s)	-0.531949	-0.544265	0.133099	-0.140906	0.433501	-0.417715	0.172902	0.748892	-0.391348	1.535857	-0.007915	0.049353
Lasso data2	-0.000000	-0.000000	0.000000	0.000104	-0.000000	-0.000118	-0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
Lasso data2(s)	-0.000000	-0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000

Table 4 Reported Weights for all Features / Models in D2



	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOcupation	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Linear data3	-0.261440	-0.248051	0.008409	0.417373	-0.080115	0.490103	-0.040862	-0.381382	1.526743	0.058689	0.083880
Linear data3(s)	-0.524144	-0.530580	0.105996	0.795231	-0.201913	0.239342	-0.087564	-0.381382	1.526743	0.058689	0.083880
Lasso data3	-0.000000	-0.000000	0.000000	0.000000	0.000000	-0.000000	-0.000000	-0.000000	0.000000	0.000000	0.000000
Lasso data3(s)	-0.000000	-0.000000	0.000000	0.000000	0.000000	-0.000000	-0.000000	-0.000000	0.000000	0.000000	0.000000

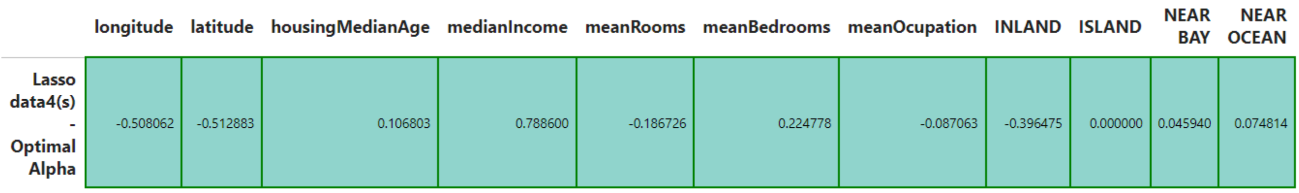
D3 (d): Discussion.

- Answer: as above table shown, the binary/dummy variable weights remain the same between original and standardised version of each model. Due to the standardisation process didn't involve these hot encoded transform the variable, where the standardisation only applied on the original numerical values.
- As assumed, the α = 100 for lasso regression for data 3, has heavily panelised all the weights to 0 in table 4, it also panelised ISLAND and NEAR BAY in table 3, regardless of target data differences between data 1 and data 2.

D4. Lasso regression fine tuning.

- The optimal α value according to the Grid-Search: **0.01**
- The RMSE on the training set: **0.71**
- The RMSE on the test set: **1.13**
- The estimated parameter values with the corresponding variable names: **table 5**

Table 5 . parameter values for each feature (lasso regression)



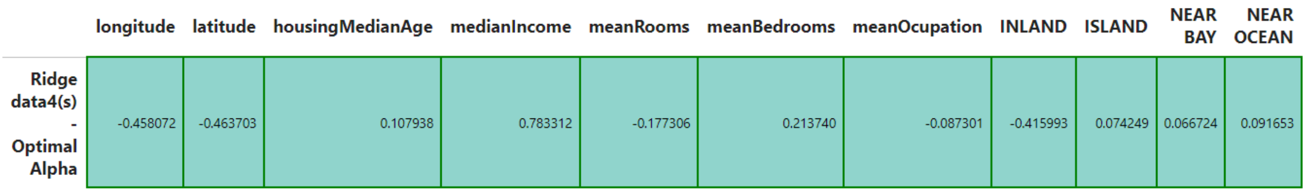
	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOcupation	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Lasso data4(s) - Optimal Alpha	-0.508062	-0.512883	0.106803	0.788600	-0.186726	0.224778	-0.087063	-0.396475	0.000000	0.045940	0.074814

D5. Ridge regression fine tuning.

(a)

- The optimal α value according to the Grid-Search: **100**
- The RMSE on the training set: **0.71**
- The RMSE on the test set: **1.13**
- The estimated parameter values with the corresponding variable names: **table 6**

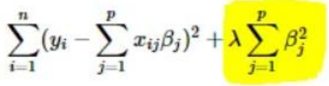
Table 6 . parameter values for each feature (Ridge regression)



	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOcupation	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Ridge data4(s) - Optimal Alpha	-0.458072	-0.463703	0.107938	0.783312	-0.177306	0.213740	-0.087301	-0.415993	0.074249	0.066724	0.091653

(b) Discussion:

- The ridge optimal α is 10,000 times larger than the Lasso regression, however the it has achieved the same RMSE on the training and testing set.
- The Lasso regression completed penalised the weights if ISLAND to 0, therefore this feature can be removed. However the Ridge won't be able to do so. Overall, there are minor differences in terms of the absolute values of weights between two models.
- The reason where α is much larger in Ridge: Due to the L2 norm penalty term in ridge regression, the penalty grows quadratically with the size of coefficient, here our coefficients are all between -1 to 1, therefore the square term of coefficients in penalty made it even smaller, hence the bigger α need in order to imply appropriate penalty effect. (below equation λ = α), where Lasso use absolute of coefficient value in penalty term, therefore the Alpha would be smaller.



$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

D6. Decision Tree Regression.

- The optimal max_depth according to the Grid-Search: **10**
- The RMSE on the training set: **0.46**
- The RMSE on the test set: **0.60**

D7. Discussion.

- According to RMSE on the test set, the decision tree has the best performance and is the best model for this dataset. In our data, there are lots of outlier issues, these may limit the regression model performance, however by using decision tree these issues maybe handled better in decision-based algorithm, rather than trying to fit the best line.

- Consider the EDA in D1, polynomial transformation of the features may improve the model.

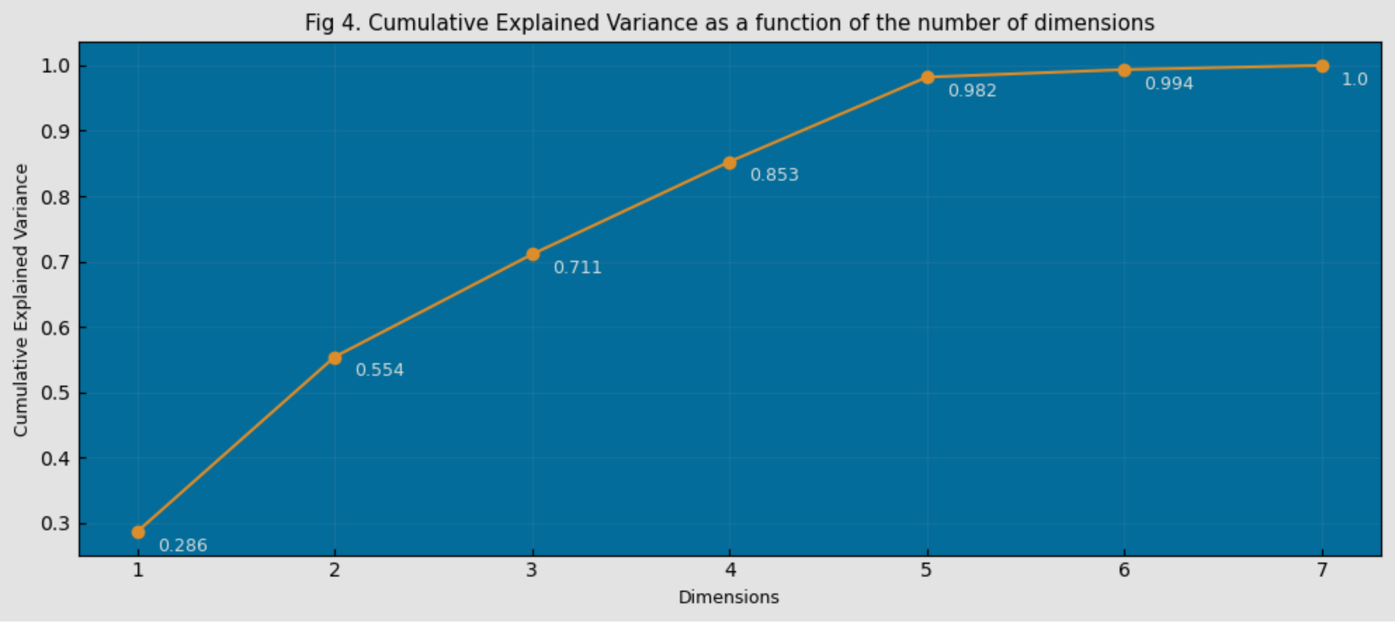
D8.

Note: For below D8, D9, D10, the target variable 'MedianHouseValue' has been added back in, as for the unsupervised learning, there is no target variables to be isolated.

Also, as required to use only numerical features. Here I've dropped all dummy variables : ('INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN'). However, It is a bit confusing here as I don't think these should be called 'categorical' variables, as they are now binary/numeric variabls. If I do not drop these columns, then the 'use only numerical features / without any categorical variables' requirements become redundant.

(a): Plot the cumulative explained variance ratio as a function of the number of principal components.

- Answer . Fig 4



(b): Determine the number of principal components necessary to preserve at least 90% of the variance.

- Answer: **5**

(c): Train a Linear Regression using the selected number of principal components. Present the RSME for the training and test data.

- Answer:
  RSME for training: **0.81**
  RSME for testing: **1.34**

(d): Use GridSearchCV to find the optimal number of principal components according to a 10 fold cross-validation and use a Linear Regression as the base model. Report the obtained optimal number of principal components and the RMSE for the training and test sets.

- Answer:
  Optimal number of Principal Components : **6**
  RSME for training: **0.73**
  RSME for testing: **1.17**

(e): Discuss the obtained results and compare them with the ones you obtained in D7:

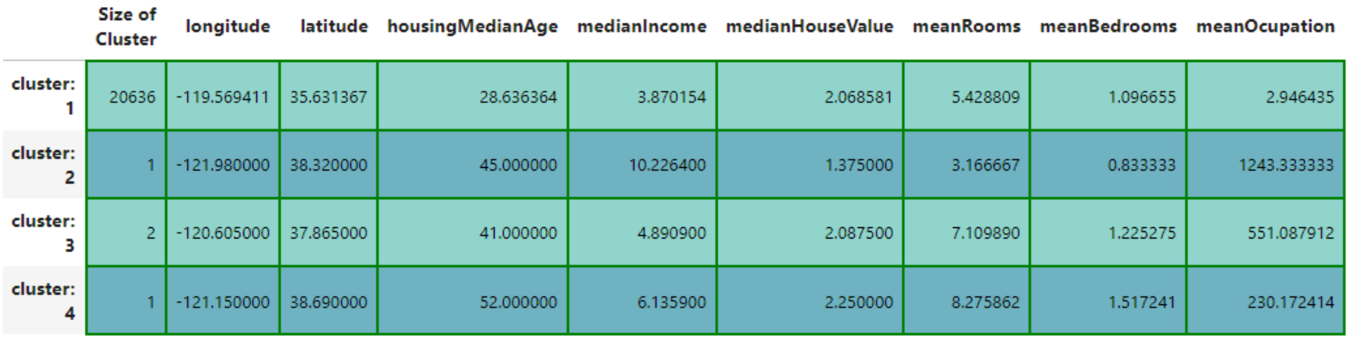
- Answer: The PCA with reduced number of features to be used in the regression models, can still achieve a reasonable performance, however the RSME for testing set is still double of the one achieved in the decision tree model in D7. Further proved the effectiveness of decision tree regressor for our data.

D9.

(a) Using this data, perform a hierarchical clustering with average linkage and Euclidean distance to cluster the districts. Cut the dendrogram at a height that results in four distinct clusters. Present the mean of the variables for each cluster and briefly summarise the characteristics of the districts in the four groups, including the size of each cluster.

- Answer: Table 7

Table 7. Size of each cluster and mean of each feature.



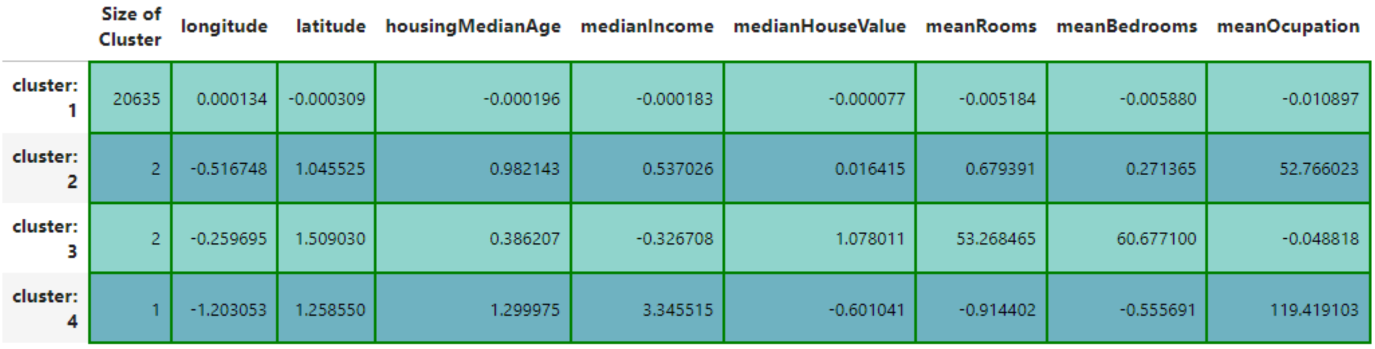
	Size of Cluster	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
cluster: 1	20636	-119.569411	35.631367	28.636364	3.870154	2.068581	5.428809	1.096655	2.946435
cluster: 2	1	-121.980000	38.320000	45.000000	10.226400	1.375000	3.166667	0.833333	1243.333333
cluster: 3	2	-120.605000	37.865000	41.000000	4.890900	2.087500	7.109890	1.225275	551.087912
cluster: 4	1	-121.150000	38.690000	52.000000	6.135900	2.250000	8.275862	1.517241	230.172414

- As we can see from the size of clusters, due to the nature of hierarchical clustering based on the Average distance between points of the clusters here, the result is not idea, nearly all instances got grouped into cluster 1, where only 1 or 2 in other clusters.
- From the mean of The MeanOcupation (last column) amongst all the clusters, we can see that there are huge differences between cluster 1 with other clusters, these may be the issue with extreme outliers, where only 1 or 2 outliers become its own cluster, but all the rest of the data instances become 1 giant cluster.

(b) Using standardised features, perform a hierarchical clustering with average linkage and Euclidean distance to cluster the districts. Cut the dendrogram at a height that results in four distinct clusters. Did the groups change? What effect does scaling the variables have on the hierarchical clustering obtained?

- Answer: Table 8

Table 8. Size of each cluster and mean of each feature (standardised).



	Size of Cluster	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
cluster: 1	20635	0.000134	-0.000309	-0.000196	-0.000183	-0.000077	-0.005184	-0.005880	-0.010897
cluster: 2	2	-0.516748	1.045525	0.982143	0.537026	0.016415	0.679391	0.271365	52.766023
cluster: 3	2	-0.259695	1.509030	0.386207	-0.326708	1.078011	53.268465	60.677100	-0.048818
cluster: 4	1	-1.203053	1.258550	1.299975	3.345515	-0.601041	-0.914402	-0.555691	119.419103

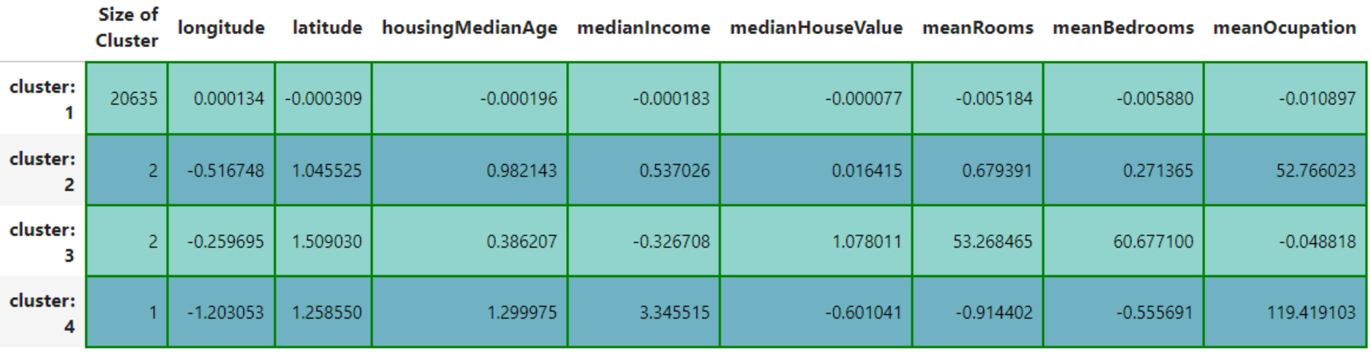
- From the above table, we can see the cluster barely changed, there are a few instances got assigned different clusters, but overall, the standardisation still maintained the characteristics of the outliers. Here we can see why those very few outliers become it is own groups, from MeanBedrooms and MeanOcupation columns.
- More specifically, below shows all the instances where the assigned cluster differ, compared to (a)

Instance Index	**1914**	**1979**	**3364**	**13034**	**16669**	**19006**
Cluster Assignment in (a)	0	0	2	3	2	1
Cluster Assignment in (b)	2	2	1	0	1	3

(c) Using standardised features, apply the k-means clustering (with k=4) with Euclidean distance. Set the initial centroids of the k-means as the group means obtained from the hierarchical clustering in part (b). Compare the results with the hierarchical clustering from part(b). Which one do you think provides a better result?

- Answer: Table 9

Table 9. Size of each cluster and mean of each feature (standardised).



	Size of Cluster	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
cluster: 1	20635	0.000134	-0.000309	-0.000196	-0.000183	-0.000077	-0.005184	-0.005880	-0.010897
cluster: 2	2	-0.516748	1.045525	0.982143	0.537026	0.016415	0.679391	0.271365	52.766023
cluster: 3	2	-0.259695	1.509030	0.386207	-0.326708	1.078011	53.268465	60.677100	-0.048818
cluster: 4	1	-1.203053	1.258550	1.299975	3.345515	-0.601041	-0.914402	-0.555691	119.419103

- As we can see from above table, with initialising the centroids, and change hierarchical clustering to K-means. The result barely changed.
- Both clustering techniques providing poor results. As by initialising the centroids from (b), the K-mean clustering won't adjust to a better result. As the centroids were manually initialised to fall on, or very close to the outliers.

(d) Perform PCA on the scaled data. Perform hierarchical clustering with average linkage and Euclidean distance on the first two principal component scores. Cut the dendrogram at height that results in four distinct clusters. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (four colours for four clusters). Compare the group characteristics to the group characteristics obtained in the previous tasks.

- Answer: Fig 5

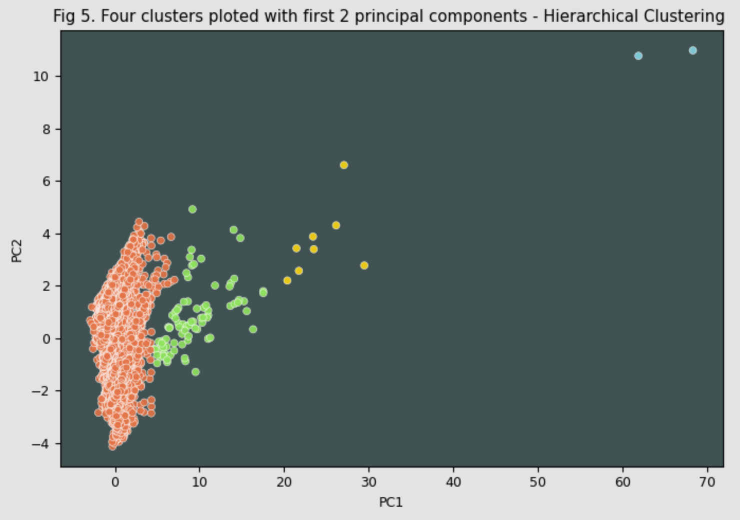
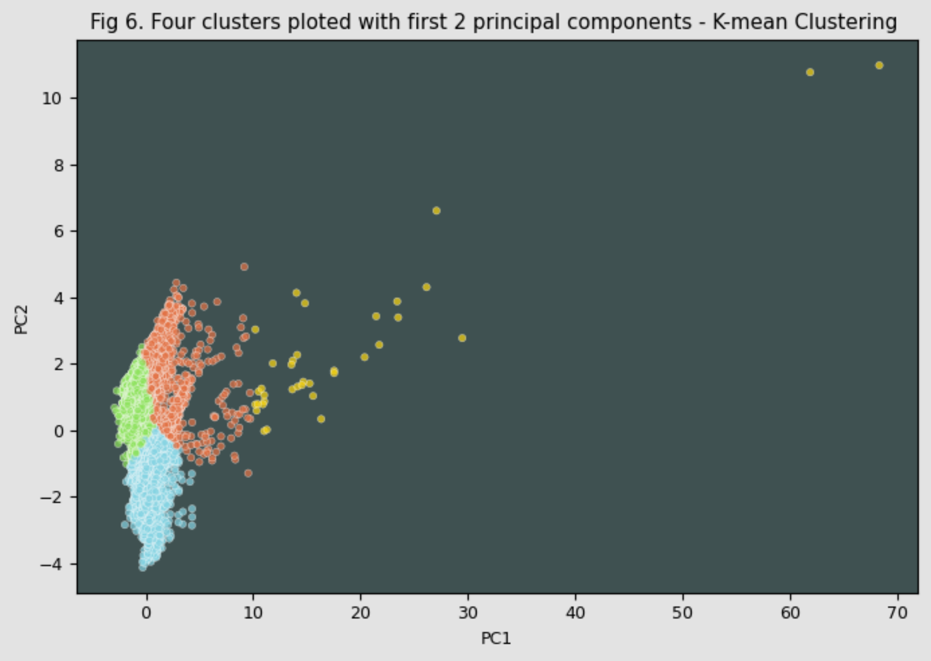


Fig 5. Four clusters ploted with first 2 principal components - Hierarchical Clustering

- After applying the PCA, the first cluster (red) still contains majority of the data instances, however the other clusters have improved by incorporating more instances.
- From the plot clearly, we can see there are 2 extreme outliers in our data. Which proves again PCA and Clustering are very useful tool to inspect the data in EDA process.

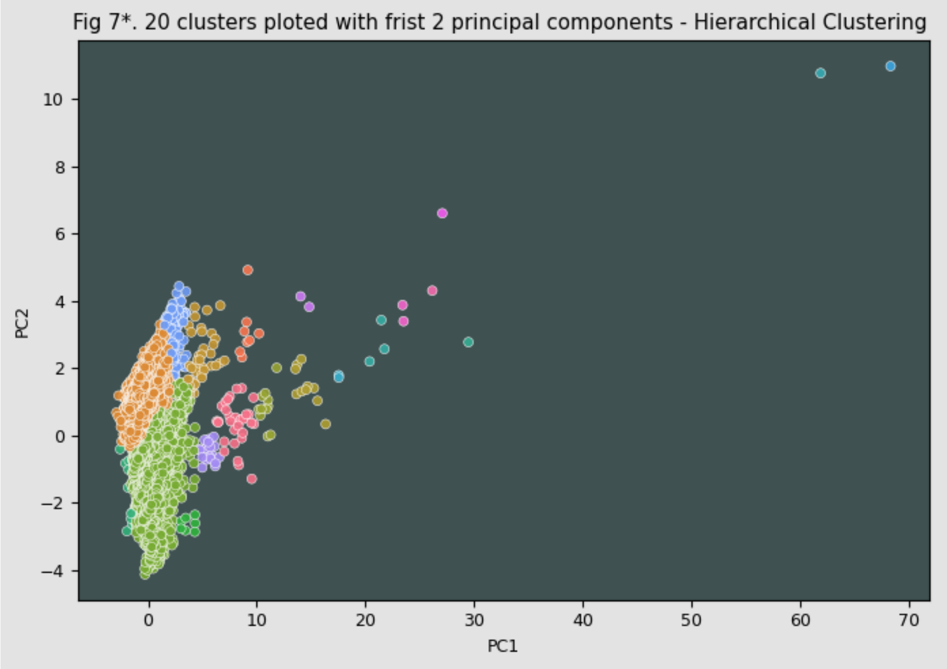
(e) Perform PCA on the scaled data. Apply the k-means clustering (with k=4) with Euclidean distance on the first two principal components scores, setting the random state to "5508".Present the scatterplot of the first two principal components using different colours for the instances on each cluster (four colours for four clusters). Discuss the results.

- Answer: Fig 6



- It is quite clear; the performance of the K-Mean clustering is a lot better then hierarchical clustering.
  Mainly because we have limited cut the dendrogram at height that results in four distinct clusters. Where the outliers have become it is own clusters.
  If we were to increase the height, the hierarchical clustering would still be able to yield a good resolution.
  But overall, when there are outliers, or part of data are far from most of the data points, hierarchical clustering may not be the best solution, where K-mean as in Fig 6 yellow cluster shows, can be a faster solution to group those points into single cluster.

  Below Fig 7* is an extra plot showing, when use hierarchical clustering but increase the K to 20, the result can still be meaningful, but overall, not so effective in our case.



D10

Consider data3 without any categorical variable. Compute the silhouette score by applying k□means on this dataset after scaling the features to have zero mean and unit standard deviation. Use values for k in range(2,20,1). Remember to set random state=5508 for the KMeans class.

(a) Plot the silhouette scores for the different k values. According to this score, what was the optimal value of clustering?

- Answer: based on Fig 8, K = 3 would be the optimal K value.

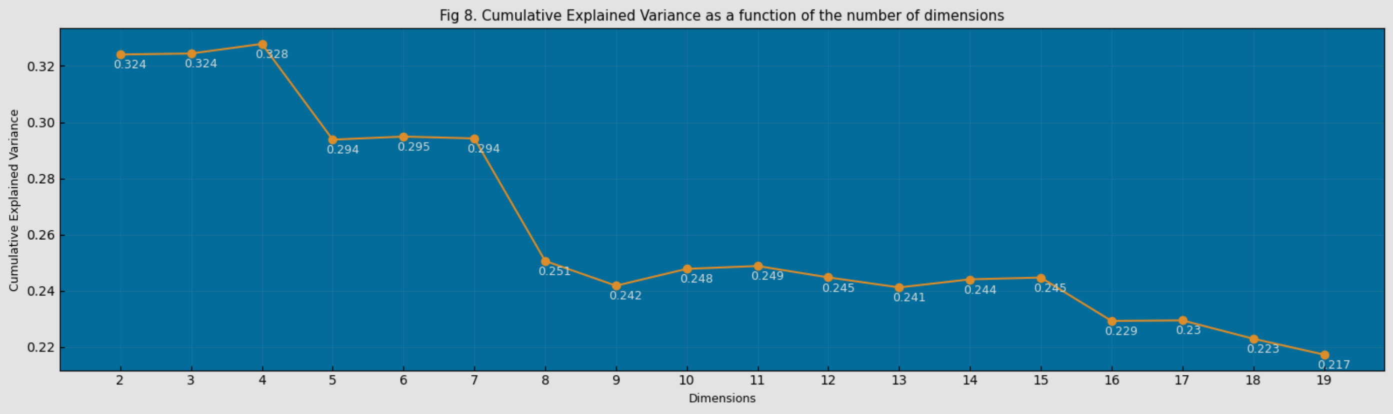


Fig 8. Cumulative Explained Variance as a function of the number of dimensions

(b) Considering the optimal k value obtained in the previous item, plot the k groups (using different colours for the instances in each group) on the first two principal component scores of the same data. In a side plot, plot the first principal component scores in which the instance colours represent the values of the categorical value you discarded for this part of the assignment. Comment on the relationship between the groups in these two plots.

- Answer: Fig 9 and Fig 10

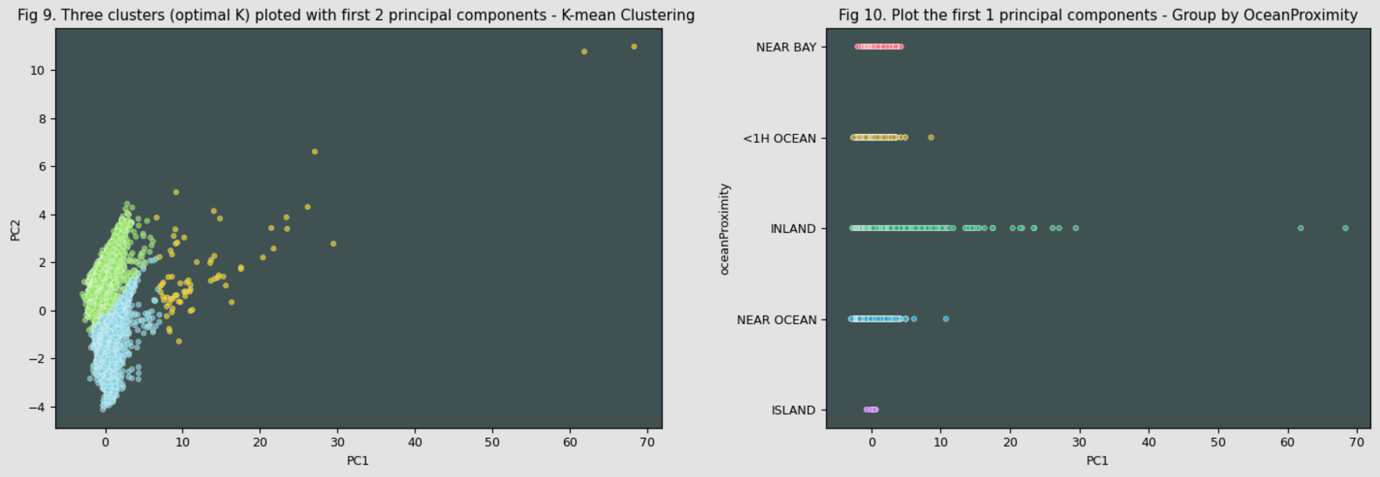


Fig 9. Three clusters (optimal K) ploted with first 2 principal components - K-mean Clustering

Fig 10. Plot the first 1 principal components - Group by OceanProximity

- As we can see, the optimal K to be set to 3 in K-mean clustering, seems like a good clustering result.
- The first PCA seems like having a relationship with Ocean Proximity, where the INLAND group tend to own the large variances. But the Near Ocean, <1H Ocean, and Near Bay are similar.
- The ISLAND group owns the minimum variance in the first principal component.

(c) With clustering analysis and your findings from EDA, what are your conclusions about the data that may be impacting your models?

- The data present to have issues with outliers, which were not removed due to the strictly following the instructions, however in real project, these outliers may need special attention before modelling.
  The outliers can affect the clustering result, especially the hierarchical in our case, however overall, as clustering is the distance-based algorithm, it is more volatile to outliers. Same with PCA.

- It is also worth noting that, when we transform the categorical Ocean Proximity into dummy binary variables, should we include these variables when we scale the dataset?

  I have experimented with both scenarios, for the regression model, it didn't affect to much of the model performance, however when all variables were scaled, the RMSE on the testing set would be improved slightly.

  Furthermore, based on my further experiment, If we do PCA on all features including numerical and dummy categorical. These dummy binary categorical variables must be scaled as well. As the testing error is so much larger if we don't scale these binary variables.   Which makes sense, as PCA and Clustering are distance based algorithms, If we don't scale these binary variables, the mean of these features are between 0 – 1, where all the numeric features was scaled to mean of 0 with 1 unit of standard deviation.