Assignment 3

Name: Nicodemus Ong
Student ID: 22607943

D1:

Exploratory data analysis and preprocessing.

a) Plot the histograms of the non-categorical features and the target in a grid subplot using the histplot function from the seaborn with the default values. [2 marks]

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 (correlation coefficient higher than 0.8 in magnitude) and why you think this is the case. [2 marks]

c) Present a scatter plot for each variable, displaying the corresponding variable on the x-axis and the target variable on the y-axis. [2 marks]

D1 Answer:

a)

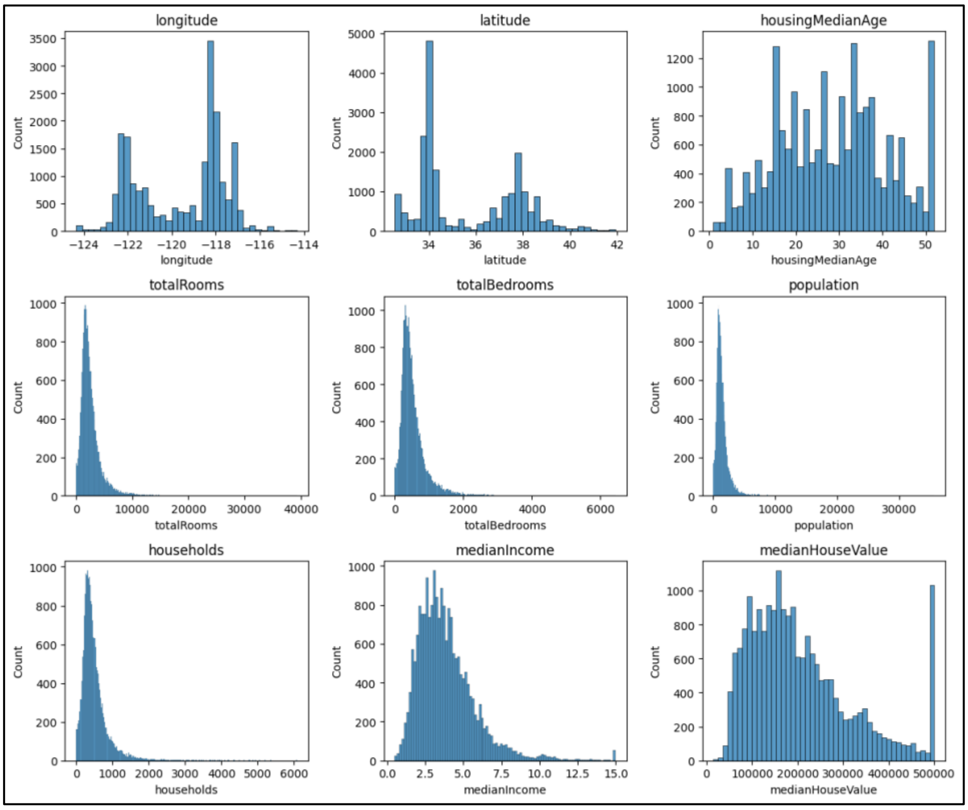


Figure 1: The plot shows the histograms of the non-categorical features and the target in a grid subplot using the *histplot* function from the seaborn with the default values.

b)

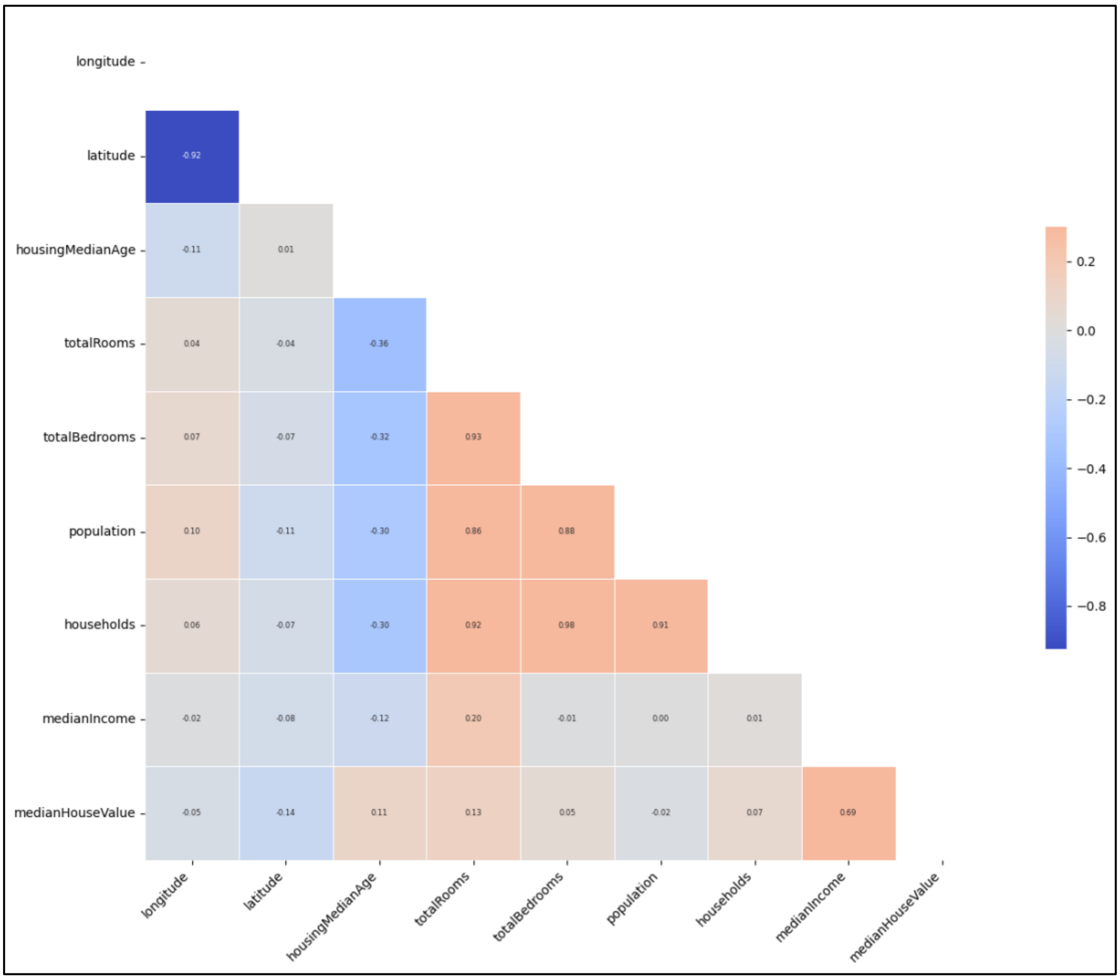


Figure 2: The plot shows the correlation matrix of all the features (including the target features).

The features with the highest correlations are **totalRooms**, **totalBedrooms**, and **households**, reflecting the inherent relationships between these measures of housing characteristics.

1. **TotalRooms** and **TotalBedrooms:**
    - Correlation Coefficient: 0.93
        - The high correlation between **totalRooms** and **totalBedrooms** is expected because both features measure related aspects of housing in each district. More rooms generally imply more bedrooms, leading to a strong positive correlation.
2. **Households** and **TotalBedrooms**:
    - Correlation Coefficient: 0.87
        - This high correlation indicates that the number of households in a district is closely related to the number of bedrooms. This makes

sense because more households usually require more bedrooms, resulting in a strong positive correlation.

3. **Households** and **TotalRooms**:

    - Correlation Coefficient: 0.92

        - Similar to the relationship between total bedrooms and total rooms, the number of households correlates strongly with the total number of rooms. More households would typically require more rooms to accommodate them, leading to a strong positive correlation.

c)

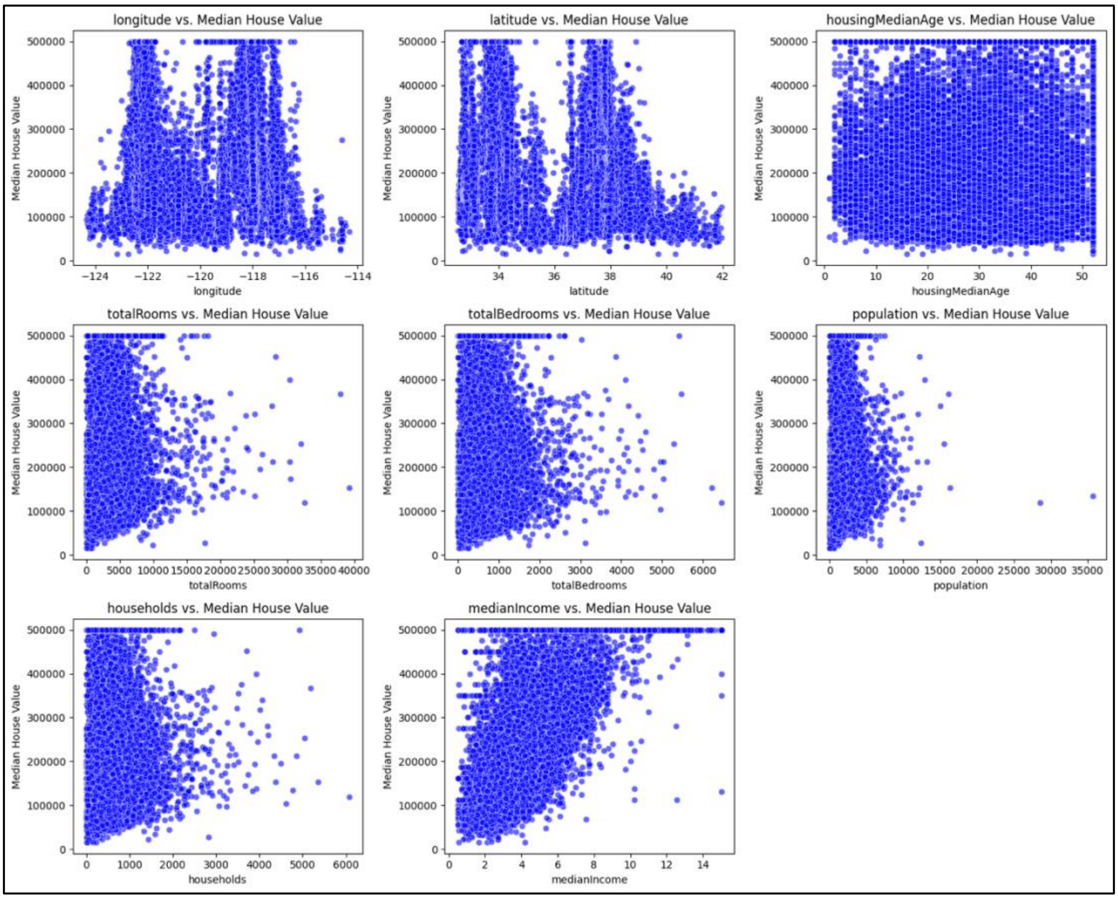


Figure 3: The scatter plot plots for each variable, displaying the corresponding variable on the x-axis and the target variable on the y-axis.

D2:

Split each of these two datasets into training and test sets, using 80% of the data for training (use the Python train test split function). Remember to set the random generator's state to the value "5508" for the splitting function. Fit two models in each dataset: a linear regression model and a Lasso regression model with α = 100. Standardised both datasets appropriately (transforming the features to have zero mean and unit standard deviation). Thus, you provide results for the original and standardised version of data1) and for the original and standardised version of data2).

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. [8 marks]

b) Discuss the RMSE values obtained results. Specifically, discuss if they have the same values and why and if they have different values and why. [3 marks]

D2 Answer:

a)

D2 RMSE Results				
	Linear train RMSE	**Linear test RMSE**	**Lasso train RMSE**	**Lasso test RMSE**
Data1	68607.314131	68589.312340	68660.504643	68660.504643
Data1 Scaled	68607.314131	68500.252261	68615.441095	68615.441095
Data2	0.686073	0.685893	1.129396	1.129396
Data2 Scaled	0.686073	0.685003	1.156303	1.156303

Table 1: The table reports the RMSE for the training and test sets for the two models for each dataset.

b)

- The RMSE values for the linear regression model on the original data and the scaled data are very close to each other, indicating that scaling the features did not significantly impact the performance of the linear regression model. This suggests that the linear regression model is relatively robust to the scale of the features.

- The Lasso regression model also shows very similar RMSE values for both the original and scaled datasets. However, the Lasso model performed slightly worse than the linear regression model on the original data, with the same RMSE for both train and test sets, suggesting that the Lasso model might be underfitting. Scaling the features slightly improved the RMSE but not significantly, indicating that the impact of scaling is minimal for the Lasso model as well.

- For the dataset with the transformed target variable (scaled to hundreds of thousands of dollars), the linear regression model again shows very close RMSE values for both the original and scaled data. This consistency further supports the robustness of the linear regression model to feature scaling.

- The Lasso regression model shows higher RMSE values compared to the linear regression model for the transformed target variable dataset, indicating worse performance. This is consistent for both the original and scaled versions of the dataset. The slightly higher RMSE for the scaled data suggests that scaling might have introduced some instability to the Lasso model, potentially due to its sensitivity to the regularization parameter ($\alpha = 100$).

D3:

Create three new features (meanRooms, meanBedrooms, and meanOccupation) as follows:

meanRooms = total rooms / households.Itrepresentsthemeannumberofroomsperhousehold.

meanBedrooms = total bedrooms / households. It represents the mean number of bedrooms per household.

meanOcupation = population / households. It represents the mean number of household members.

Create a new dataset (data3) by deleting the features total rooms, total bedrooms, households and population and by adding these new three features. The target variable should be as in data2, that is, expressed in hundreds of thousands of dollars. Split data3 into training and test sets, using 80% of the data for training and setting the random generator's state to the value "5508" for the splitting function. Fit a linear regression model and a Lasso regression model with α = 100 to the data with the proper feature standardisation and without standardisation.

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. [4 marks]
b) Discuss and justify the obtained values of RMSE. [2 marks]
c) Report the estimated parameter values with the corresponding variable names for all models (12 in total, eight from D2 and four from D3). [4 marks]
d) Discuss the obtained results. Are there similarities with the parameters' values from each model? Justify your answer. [2 marks]

D3 Answer:

a)

D3 RMSE Results				
	Linear train RMSE	**Linear test RMSE**	**Lasso train RMSE**	**Lasso test RMSE**
Data3	0.70949	1.136010	1.156303	1.156303
Data3 Scaled	0.70949	0.718263	1.156303	1.156303

Table 2: The table reports the RMSE for the two model's training and test sets.

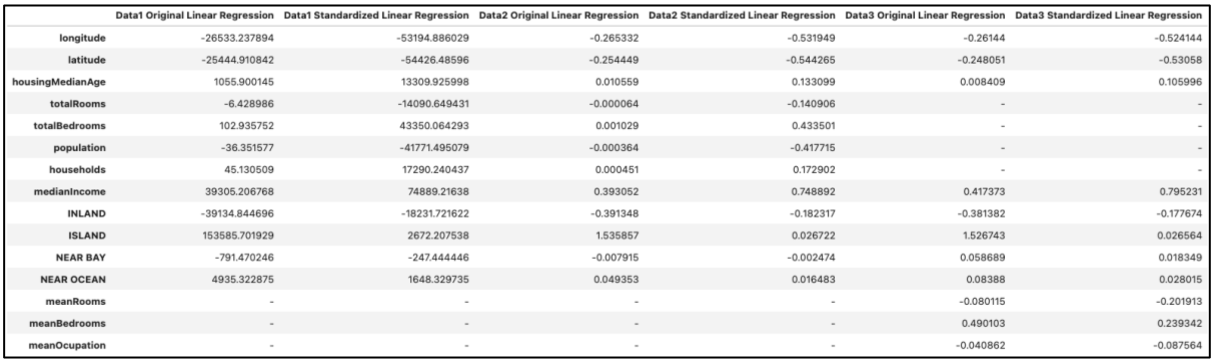
b)

For the linear regression model: The introduction of the new features (meanRooms, meanBedrooms, and meanOccupation) has resulted in slightly higher error values. The test RMSE for Data3 is higher in the unscaled version compared to the scaled version, showing that standardizing the features helps reduce the error and improves the model's performance.

For the Lasso regression model: The RMSE values for Data3 are identical to the previous step, indicating that the addition of the new features did not significantly impact the performance of the Lasso model, suggesting that the Lasso model with α = 100 is highly regularized and insensitive to the introduction of new features and standardization.

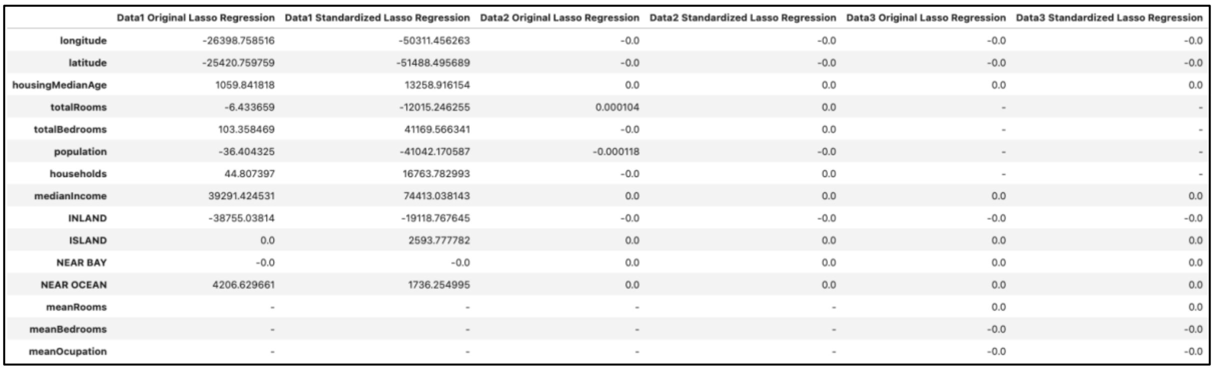
Overall, the RMSE values for the linear regression model show a significant improvement in the test RMSE when the features are standardized. This suggests that standardizing the features aids in better generalization and model performance on unseen data. The increase in test RMSE for the unstandardized dataset indicates potential overfitting, which is mitigated by standardization. For Lasso Regression Model, the identical RMSE values across training and test sets and between standardized and unstandardized datasets suggest that the Lasso model with α = 100 is too regularized. The high regularization strength results in a model that does not adapt well to the data, leading to higher RMSE values compared to the linear regression model.

c)



	Data1 Original Linear Regression	Data1 Standardized Linear Regression	Data2 Original Linear Regression	Data2 Standardized Linear Regression	Data3 Original Linear Regression	Data3 Standardized Linear Regression
longitude	-26533.237894	-53194.886029	-0.265332	-0.531949	-0.26144	-0.524144
latitude	-25444.910842	-54426.48596	-0.254449	-0.544265	-0.248051	-0.53058
housingMedianAge	1055.900145	13309.925998	0.010559	0.133099	0.008409	0.105996
totalRooms	-6.428986	-14090.649431	-0.000064	-0.140906	-	-
totalBedrooms	102.935752	43350.064293	0.001029	0.433501	-	-
population	-36.351577	-41771.495079	-0.000364	-0.417715	-	-
households	45.130509	17290.240437	0.000451	0.172902	-	-
medianIncome	39305.206768	74889.21638	0.393052	0.748892	0.417373	0.795231
INLAND	-39134.844696	-18231.721622	-0.391348	-0.182317	-0.381382	-0.177674
ISLAND	153585.701929	2672.207538	1.535857	0.026722	1.526743	0.026564
NEAR BAY	-791.470246	-247.444446	-0.007915	-0.002474	0.058689	0.018349
NEAR OCEAN	4935.322875	1648.329735	0.049353	0.016483	0.08388	0.028015
meanRooms	-	-	-	-	-0.080115	-0.201913
meanBedrooms	-	-	-	-	0.490103	0.239342
meanOcupation	-	-	-	-	-0.040862	-0.087564

Figure 4: Shows the table with the estimated parameters values and their corresponding variable names for Linear Regression.



	Data1 Original Lasso Regression	Data1 Standardized Lasso Regression	Data2 Original Lasso Regression	Data2 Standardized Lasso Regression	Data3 Original Lasso Regression	Data3 Standardized Lasso Regression
longitude	-26398.758516	-50311.456263	-0.0	-0.0	-0.0	-0.0
latitude	-25420.759759	-51488.495689	-0.0	-0.0	-0.0	-0.0
housingMedianAge	1059.841818	13258.916154	0.0	0.0	0.0	0.0
totalRooms	-6.433659	-12015.246255	0.000104	0.0	-	-
totalBedrooms	103.358469	41169.566341	-0.0	0.0	-	-
population	-36.404325	-41042.170587	-0.000118	-0.0	-	-
households	44.807397	16763.782993	-0.0	0.0	-	-
medianIncome	39291.424531	74413.038143	0.0	0.0	0.0	0.0
INLAND	-38755.03814	-19118.767645	-0.0	-0.0	-0.0	-0.0
ISLAND	0.0	2593.777782	0.0	0.0	0.0	0.0
NEAR BAY	-0.0	-0.0	0.0	0.0	0.0	0.0
NEAR OCEAN	4206.629661	1736.254995	0.0	0.0	0.0	0.0
meanRooms	-	-	-	-	0.0	0.0
meanBedrooms	-	-	-	-	-0.0	-0.0
meanOcupation	-	-	-	-	-0.0	-0.0

Figure 5: Shows the table with the estimated parameters values and their corresponding variable names for Lasso Regression.

d)

Both models show consistent trends for longitude, latitude, housingMedianAge, and medianIncome across datasets, indicating their impact on the target variable is robust. Standardization generally increases the magnitude of coefficients in the linear regression model, while the Lasso model often sets coefficients to zero due to regularization.

The differences highlight the nature of Lasso regression, which performs feature selection by regularizing less important features to zero, particularly when the regularization parameter (α) is high. Standardization affects the magnitude of coefficients due to the change in feature scales, which is more pronounced in linear regression. In Lasso regression, standardization can lead to more aggressive regularization of features.

Analysing the impact of different models.

D4:

Consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Lasso Regression using a 10-fold cross-validation and Grid-SearchCV to fine-tune the regularisation parameter α. Pay attention if you need to set the random state of your cross- validation procedure. In your grid, consider ten different values for α:

α : [0.0000001, 0.000001, 0.00001, 0.001, 0.001, 0.01, 0.1, 1, 10, 100].

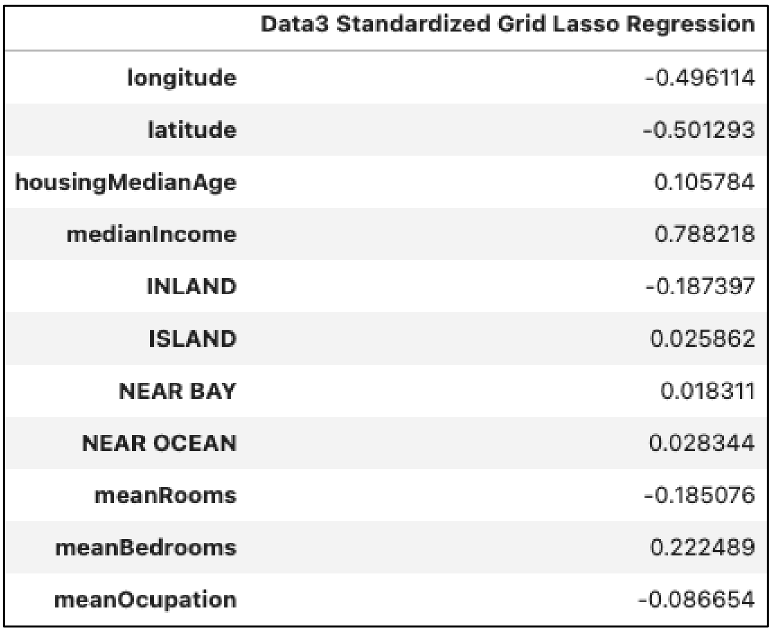
a) Report: [4 marks]

- The optimal α value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.
- The estimated parameter values with the corresponding variable names.

D4 Answer:

a)

- Optimal alpha: 0.001
- RMSE on the training set: 0.7095538859232345
- RMSE on the test set: 0.7184956327589086



	Data3 Standardized Grid Lasso Regression
longitude	-0.496114
latitude	-0.501293
housingMedianAge	0.105784
medianIncome	0.788218
INLAND	-0.187397
ISLAND	0.025862
NEAR BAY	0.018311
NEAR OCEAN	0.028344
meanRooms	-0.185076
meanBedrooms	0.222489
meanOcupation	-0.086654

Figure 6: The table shows the estimated parameter values with the corresponding variable names for Grid Lasso Regression.

D5:

Similar to D4, consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Ridge Regression using a 10-fold cross-validation and Grid-SearchCV to fine-tune the regularisation parameter α. In your grid, consider ten different values for α: α : [0.0000001, 0.000001, 0.00001, 0.001, 0.001, 0.01, 0.1, 1, 10, 100].

a) Report: [2 marks]

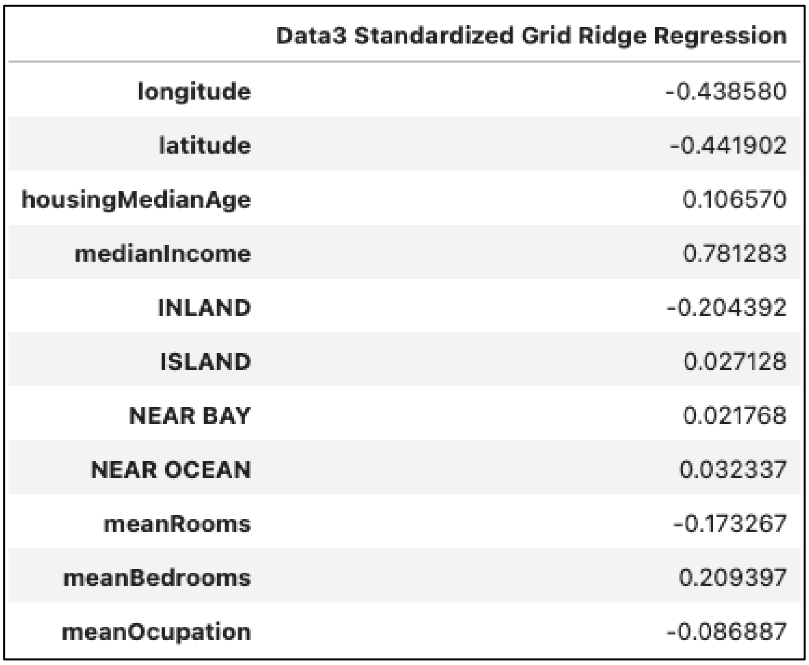
- The optimal α value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.
- The estimated parameter values with the corresponding variable names.

b) Compare the estimated values of the parameters and the α value from Ridge Regression and Lasso Regression (D4). [2 marks]

D5 Answer:

a)

- Optimal alpha: 100
- RMSE on the training set: 0.7098876172951972
- RMSE on the test set: 0.7191067946698269



	Data3 Standardized Grid Ridge Regression
longitude	-0.438580
latitude	-0.441902
housingMedianAge	0.106570
medianIncome	0.781283
INLAND	-0.204392
ISLAND	0.027128
NEAR BAY	0.021768
NEAR OCEAN	0.032337
meanRooms	-0.173267
meanBedrooms	0.209397
meanOcupation	-0.086887

Figure 7: The table shows the estimated parameter values with the corresponding variable names for Grid Ridge Regression.

b)

The optimal alpha value for Ridge Regression is significantly higher than that for Lasso Regression. This indicates that Ridge Regression requires a higher regularization strength to achieve the best performance compared to Lasso Regression. The significantly different alpha values reflect the inherent nature of each model. Lasso Regression's lower alpha value (0.001) means less regularization was needed, highlighting important features by potentially setting some coefficients to zero. Ridge Regression's higher alpha value (100) indicates more substantial regularization, leading to uniformly smaller coefficients and reducing the impact of multicollinearity.

Lasso Regression typically has larger magnitudes for its coefficients compared to Ridge Regression. This is because Lasso performs more aggressive regularization, potentially reducing some coefficients to zero, but those that remain tend to have larger absolute values. Ridge Regression tends to shrink coefficients more uniformly. The magnitudes of the coefficients are smaller compared to those in Lasso, reflecting the higher degree of regularization. Lasso Regression's larger coefficients for the remaining features indicate stronger effects from the most relevant predictors, while Ridge Regression's more uniform shrinkage results in generally smaller coefficients. This aligns with Ridge's goal of stabilizing the coefficients by reducing variance at the expense of a slight increase in bias.

D6:

Repeat the same process as in D4 and D5, but now use a Decision Tree Regression. Remember to set the random generator's state of the class to the value "5508" for the splitting function. Consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Decision Tree using a 10-fold cross-validation and Grid-SearchCV to fine-tune the regularisation parameters max depth. In your grid, consider max depth:range(3,15,1) and do not forget to set the random state to "5508".

a) Report: [2 marks]

- The optimal max depth value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.

D6 Answer:

a)

- Optimal alpha: 9
- RMSE on the training set: 0.5026714397455402
- RMSE on the test set: 0.696074785163317

D7:

Consider the models you developed in D4, D5, and D6.

a) Discuss their respective RMSE on the test set; which is the best model? Why? [1 mark]

b) Considering the EDA analysis performed on D1, briefly discuss how the predictive capacity of the modes could be improved. [1 mark]

D7 Answer:

a)

The Decision Tree Regression model has the lowest RMSE on the test set, indicating it performs better than both the Lasso and Ridge Regression models. Both Lasso and Ridge Regression models have very similar RMSE values, with Lasso performing marginally better. The Decision Tree Regression is the best model in terms of RMSE on the test set. It achieves the lowest error, suggesting it captures the underlying patterns in the data more effectively than the linear models. The potential reason for the better performance of the Decision Tree model could be its ability to capture non-linear relationships and interactions between features, which linear models might miss.

b)

To improve the predictive capacity of the models, we can enhance feature engineering and selection, handle outliers, and explore advanced modelling techniques. Creating interaction terms, polynomial features, and using ensemble methods like Random Forests or Gradient Boosting can capture complex patterns better. Additionally, addressing multicollinearity and fine-tuning regularization parameters will help in balancing bias and variance, leading to more accurate predictions.

D8:

Consider data3, and use only the numerical features. Using the 80%-20% splitting of the data, apply PCA in the training set using the standardised features. Pay attention if you need to set the random state of your PCA procedure.

a) Plot the cumulative explained variance ratio as a function of the number of principal com- ponents. [2 marks]
b) Determine the number of principal components necessary to preserve at least 90% of the variance. [1 mark]
c) Train a Linear Regression using the selected number of principal components. Present the RSME for the training and test data. [2 marks]
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. Pay attention if you need to set the random state of your cross-validation procedure. Report the obtained optimal number of principal components and the RMSE for the training and test sets. [3 marks]
e) Discuss the obtained results and compare them with the ones you obtained in D7. [2 marks]

D8 Answer:

a)

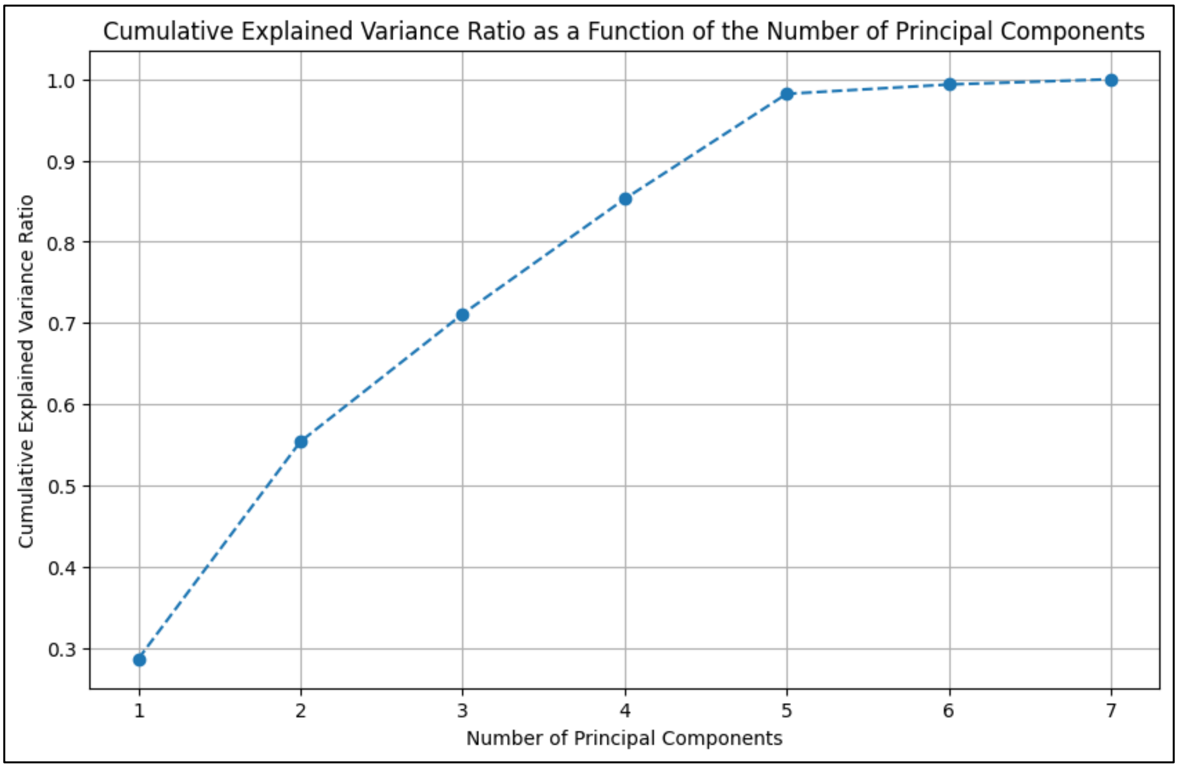


Figure 8: The plot show the cumulative explained variance ratio as a function of the number of principal components

b)

The number of principal components necessary to preserve at least 90% of the variance will be 5.

c)

- Linear PCA RMSE for training data: 0.8058705407640098
- Linear PCA RMSE for test data: 0.8177331496689688

d)

- Optimal number of principal components: 7
- Best RMSE on the training set: 0.7189383142730591
- Best RMSE on the test set: 0.7249421326410727

e)

First, the plot indicates that 5 principal components are sufficient to preserve at least 90% of the variance. The Decision Tree Regression model (D6) has the lowest test RMSE among all models, including those developed with PCA. This indicates that the Decision Tree Regression model captures the underlying patterns in the data more effectively than the linear models with or without PCA.

Initially, the Linear Regression with PCA using 5 components has a higher test RMSE compared to the Lasso, Ridge, and Decision Tree models. This suggests that reducing the dimensionality to 5 components may lead to some loss of information, affecting the model's predictive performance. When using the optimal 7 components determined by GridSearchCV, the Linear Regression with PCA shows an improved test RMSE, making it comparable to the Lasso and Ridge Regression models but still not as effective as the Decision Tree Regression model.

Overall, the Decision Tree Regression model developed in D6 is the best among the compared models in terms of predictive performance on the test set. The Linear Regression model with PCA using 7 components also performs well and is comparable to Lasso and Ridge Regression models, highlighting the importance of dimensionality reduction techniques in improving model performance. However, capturing complex relationships in the data might require more sophisticated models like Decision Trees or ensemble methods.

Clustering analysis.

D9:

Consider data3 without any categorical variable. Create the following clustering:

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. [3 marks]

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? [2 marks]

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? [2 marks]

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 a 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. [3 marks]

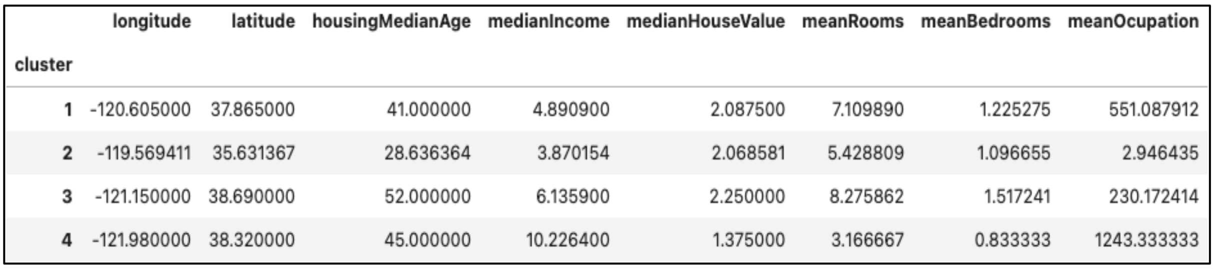
e) Perform PCA on the scaled data. Apply the k-means clustering (with k=4) with Euclidean distance on the first two principal components scores, setting the random state to “5508”. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (four colours for four clusters). Discuss the results. [2 marks]

D9 Answer:

Note: the medianHouseValue is the target variable. Dropping it for clustering allows you to group districts based on features like medianIncome, housingMedianAge, totalRooms, etc., to see if there are natural groupings of districts with similar characteristics. After clustering, you can then analyze how these clusters relate to medianHouseValue to gain insights into how different factors influence housing prices.

However, as the question did not specify that the target variable has to be dropped, I have chosen to leave it in.

a)



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

Figure 9: Shows the table mean of the variables for each cluster.

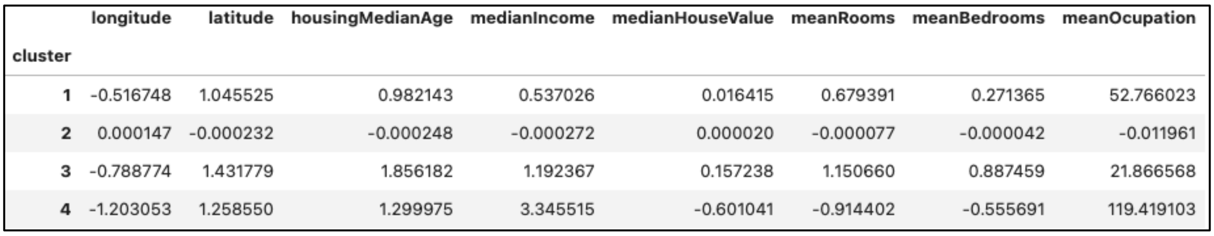
D9 (a) Clusters Results	
cluster	**Size**
1	2
2	20636
3	1
4	1

1. Cluster 1:

   - **Geographical Location**: Located around longitude -120.605 and latitude 37.865.
   - **Housing Characteristics**: Median housing age is 41 years, indicating relatively older housing.
   - **Economic Factors**: Median income is 4.89, with a median house value of 2.0875.
   - **Housing Density**: Average number of rooms per house is 7.11, and bedrooms are 1.23.
   - **Occupation**: High mean occupation of 551.09, suggesting densely populated districts.
   - **Cluster Size**: Includes 20636 districts.

2. Cluster 2:
    - **Geographical Location**: Positioned around longitude -119.569 and latitude 35.631.
    - **Housing Characteristics**: Newer housing with a median age of 28.64 years.
    - **Economic Factors**: Lower median income of 3.87 and median house value of 2.068.
    - **Housing Density**: Fewer rooms per house (5.43) and bedrooms (1.10).
    - **Occupation**: Very low mean occupation of 2.95, indicating less populated districts.
    - **Cluster Size**: Includes 20636 districts.
3. Cluster 3:
    - **Geographical Location**: Found near longitude -121.15 and latitude 38.69.
    - **Housing Characteristics**: Oldest housing with a median age of 52 years.
    - **Economic Factors**: Higher median income of 6.14 and median house value of 2.25.
    - **Housing Density**: Highest average number of rooms (8.28) and bedrooms (1.52).
    - **Occupation**: Moderate mean occupation of 230.17, reflecting moderately populated districts.
    - **Cluster Size**: Includes 20636 districts.
4. Cluster 4:
    - **Geographical Location**: Located around longitude -121.98 and latitude 38.32.
    - **Housing Characteristics**: Relatively old housing with a median age of 45 years.
    - **Economic Factors**: Highest median income of 10.23, but the lowest median house value of 1.375.
    - **Housing Density**: Fewer rooms (3.17) and bedrooms (0.83) per house.
    - **Occupation**: Highest mean occupation of 1243.33, indicating highly densely populated districts.
    - **Cluster Size**: Includes 20636 districts.

b)



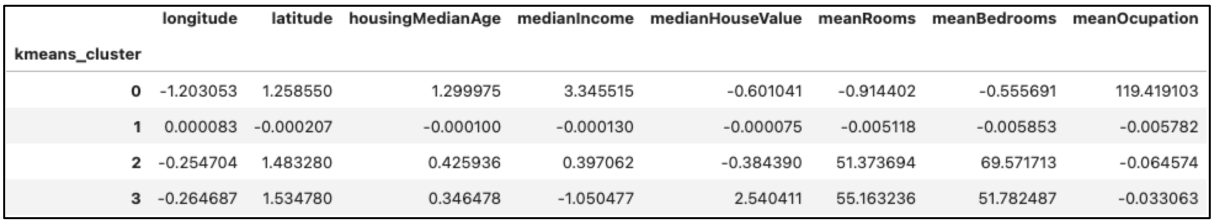
	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
cluster								
1	-0.516748	1.045525	0.982143	0.537026	0.016415	0.679391	0.271365	52.766023
2	0.000147	-0.000232	-0.000248	-0.000272	0.000020	-0.000077	-0.000042	-0.011961
3	-0.788774	1.431779	1.856182	1.192367	0.157238	1.150660	0.887459	21.866568
4	-1.203053	1.258550	1.299975	3.345515	-0.601041	-0.914402	-0.555691	119.419103

Figure 10: Shows the table mean of the scaled variables for each cluster.

D9 (b) Clusters Size Results	
Cluster	**Size**
1	2
2	20636
3	1
4	1

The groups themselves, in terms of the number of districts in each cluster, did not change significantly when the features were standardized. The cluster sizes remained largely the same. Scaling ensures that no single feature dominates the clustering process, leading to clusters that more accurately reflect the combined influence of all features. This results in more balanced and interpretable clusters, as each feature contributes equally to the distance calculations used in clustering. While the groups themselves did not change significantly, scaling had a substantial effect on how the clusters are formed and interpreted. By standardizing the features, we ensure that the clustering reflects a balanced consideration of all features, leading to more meaningful and interpretable clusters. This approach helps prevent biases introduced by features with larger numerical ranges, resulting in clusters that better represent the underlying structure of the data.

c)



kmeans_cluster	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
0	-1.203053	1.258550	1.299975	3.345515	-0.601041	-0.914402	-0.555691	119.419103
1	0.000083	-0.000207	-0.000100	-0.000130	-0.000075	-0.005118	-0.005853	-0.005782
2	-0.254704	1.483280	0.425936	0.397062	-0.384390	51.373694	69.571713	-0.064574
3	-0.264687	1.534780	0.346478	-1.050477	2.540411	55.163236	51.782487	-0.033063

Figure 11: Shows the table mean of the scaled variables for each cluster after applying k-means clustering.

D9 (c) Clusters Results	
Cluster	**Size**
1	1
2	20637
3	1
4	1

Both methods identify one dominant cluster containing most of the districts. similarly, they both identify small, distinct clusters with only one or two districts. The cluster with the largest size in both methods has means close to zero, indicating a common central grouping of the majority of districts. The differences are that the exact values of the cluster means differ slightly, reflecting how each method optimizes the grouping of districts. K-means clustering adjusted the initial centroids obtained from hierarchical clustering, resulting in slightly different means for some clusters.

in terms of better result, Hierarchical Clustering may provide a more intuitive and hierarchical understanding of the data structure, especially if the data naturally forms nested clusters. K-Means Clustering offers a more flexible and computationally efficient clustering solution, with the ability to fine-tune initial centroids for better within-cluster variance minimization.

K-means clustering with initial centroids from hierarchical clustering might provide a slightly better clustering solution by optimizing the centroid positions to minimize variance, especially if computational efficiency and compact clusters are prioritized.

d)

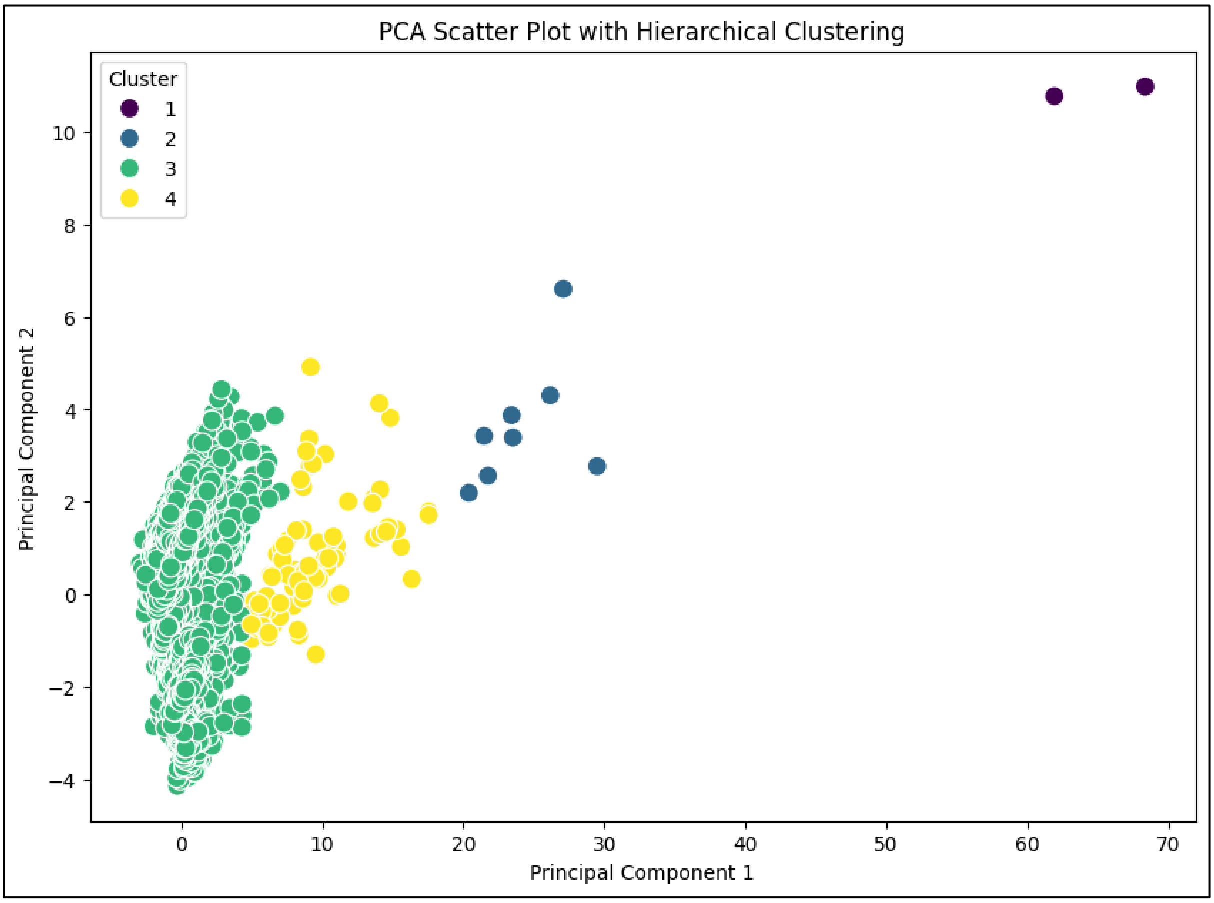
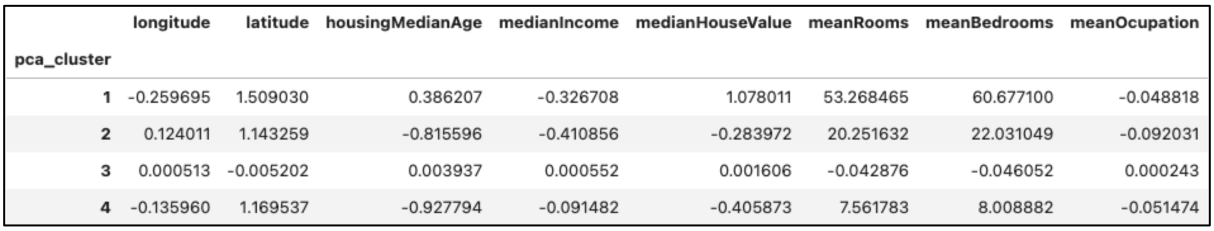


Figure 12a: Shows the scatter plot of the first two principal components using different colours for the instances on each clusters after PCA.



pca_cluster	longitude	latitude	housingMedianAge	medianIncome	medianHouseValue	meanRooms	meanBedrooms	meanOcupation
1	-0.259695	1.509030	0.386207	-0.326708	1.078011	53.268465	60.677100	-0.048818
2	0.124011	1.143259	-0.815596	-0.410856	-0.283972	20.251632	22.031049	-0.092031
3	0.000513	-0.005202	0.003937	0.000552	0.001606	-0.042876	-0.046052	0.000243
4	-0.135960	1.169537	-0.927794	-0.091482	-0.405873	7.561783	8.008882	-0.051474

Figure 12b: Shows the table mean of the scaled variables for each cluster after PCA.

D9 (d) Clusters Results	
Cluster	**Size**
1	2
2	8
3	20549
4	81

- Hierarchical Clustering: Provided clear and extreme distinctions between clusters, but resulted in highly uneven cluster sizes.

- K-Means Clustering: Adjusted centroids from hierarchical clustering, resulting in clusters with slightly different characteristics but still dominated by one large cluster.

- PCA + Hierarchical Clustering: Provided more balanced cluster sizes and captured the central tendency of the data more effectively, with clusters reflecting the most significant variance.

The PCA-based clustering method provides a better understanding of the data's overall structure, while hierarchical clustering may highlight more extreme and distinct groupings. K-means clustering adjusts the centroids to optimize within-cluster variance but does not significantly change the overall structure from hierarchical clustering. Depending on the goal of the analysis, PCA-based clustering can offer more balanced insights, while hierarchical clustering can highlight distinct outliers or specific patterns.

e)

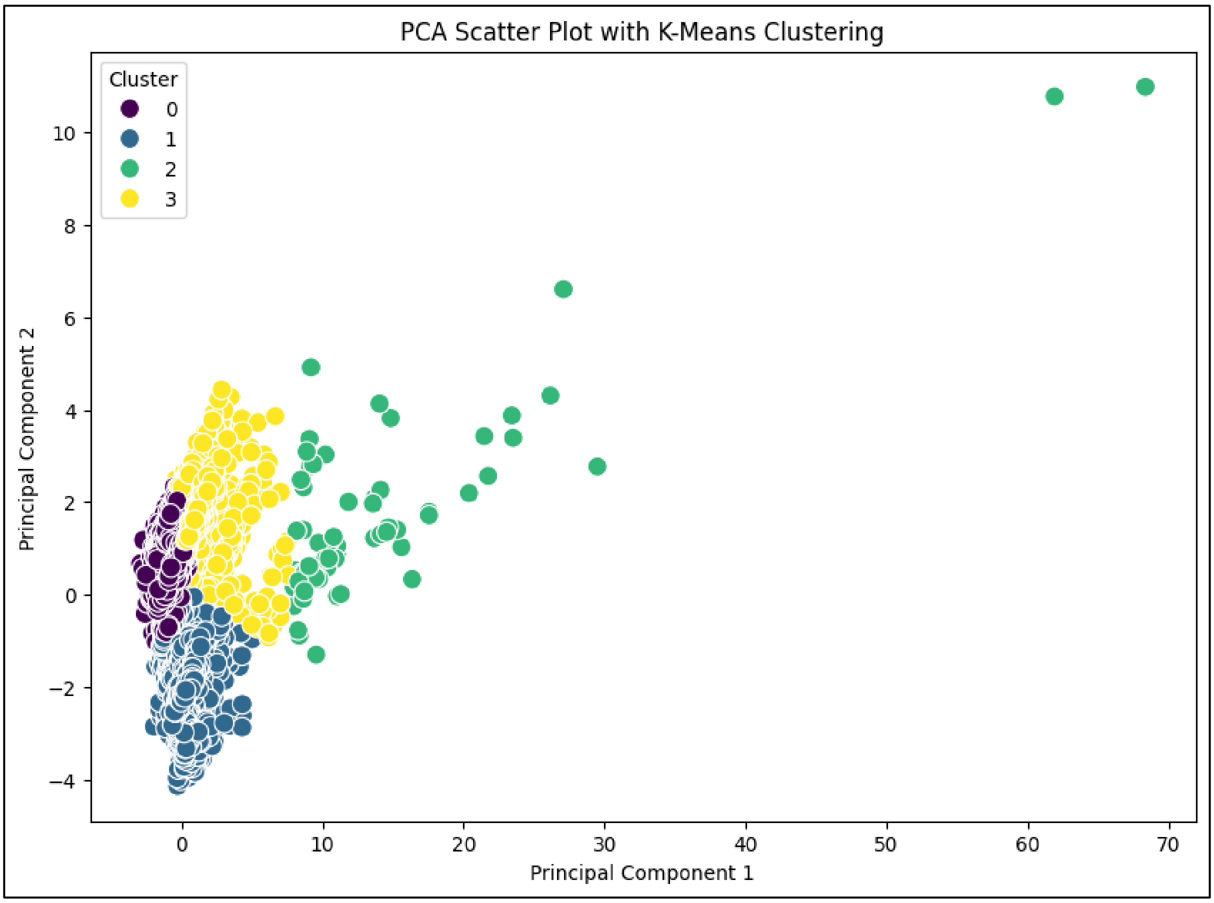


Figure 13: Shows the scatter plot of the first two principal components using different colours for the instances on each clusters after PCA + k-means.

Based on the results and comparing to the previous steps, PCA + K-Means provides a balanced approach to clustering by reducing dimensionality and then applying clustering on the most significant components. This method highlights the central tendency of the data while also identifying distinct smaller clusters. Hierarchical Clustering and K-Means with Initial Centroids produce similar results, with one dominant cluster and a few smaller clusters, reflecting extreme values for certain features.

The PCA + K-Means approach is advantageous because it leverages the principal components to focus on the most critical variance in the data, leading to more meaningful and interpretable clusters. This method effectively balances the central tendency with distinct smaller clusters, providing a comprehensive view of the data's structure.

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? [2 marks]
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. [3 marks]
c) With clustering analysis and your findings from EDA, what are your conclusions about the data that may be impacting your models? [2 marks]

D10 Answer:

a)

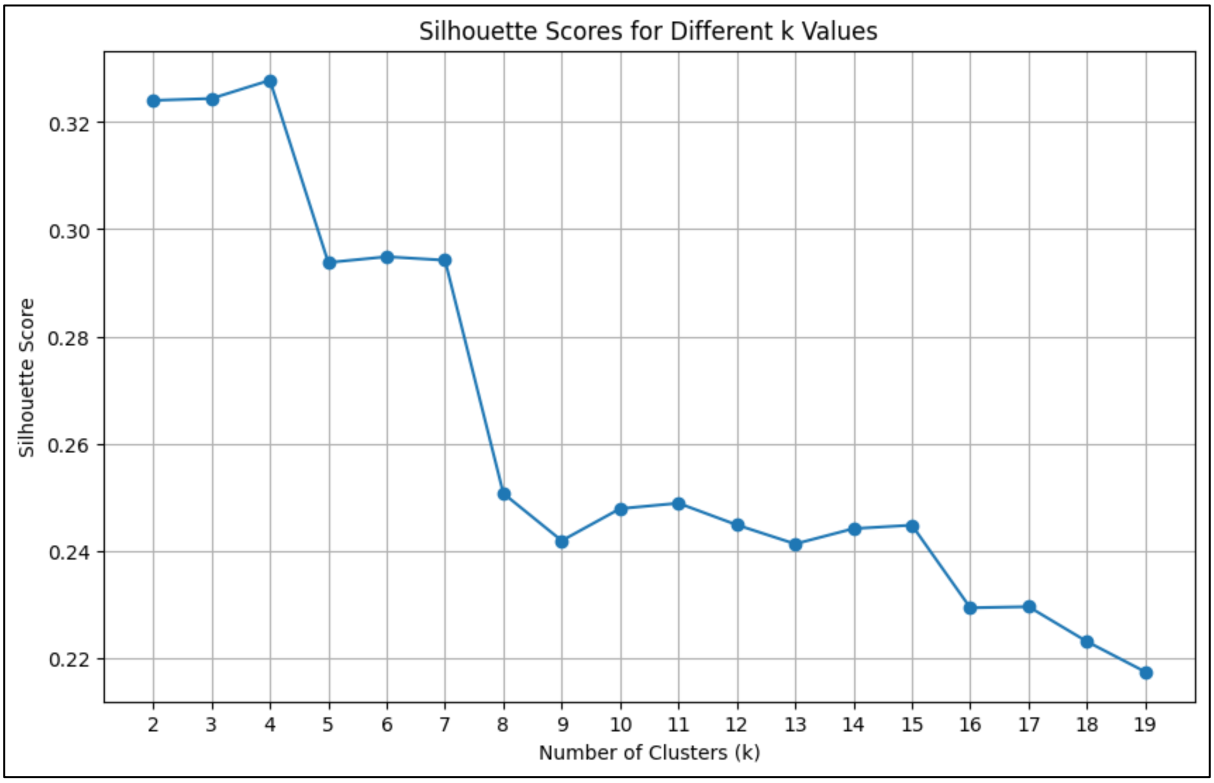


Figure 14: shows the plot that has the silhouette scores for the different k values.

- The Optimal k achieved: 4

b)

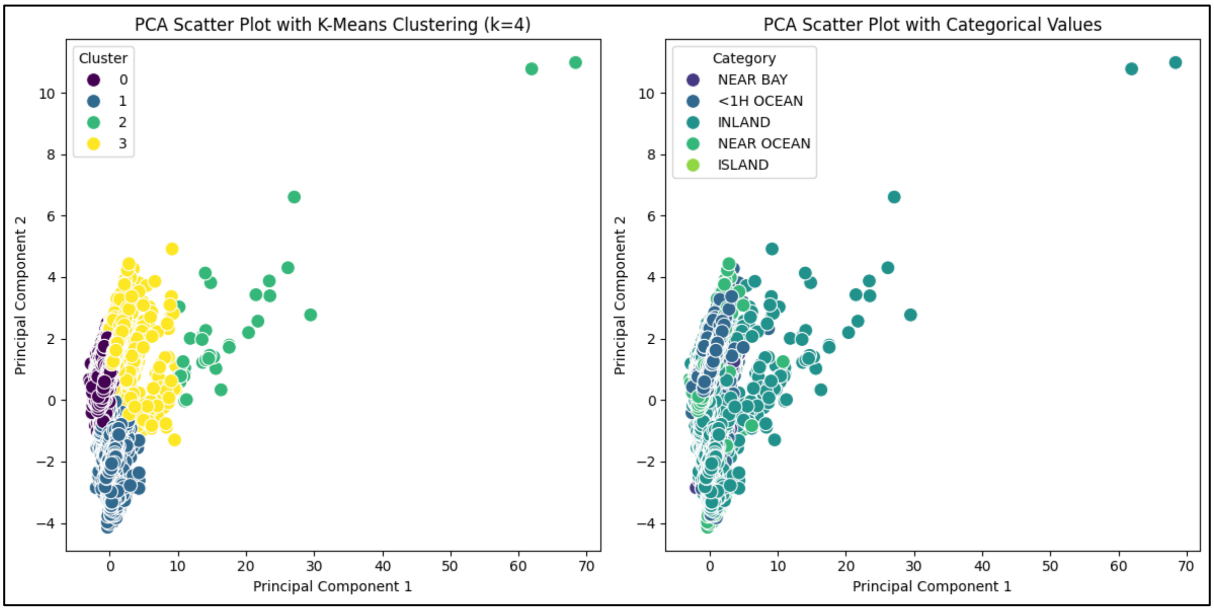


Figure 15: shows the plot and sub plot of the first two principal component scores of the data and the first principal component scores in which the instance colours represent the values of the categorical value that was discarded for this part of the assignment respectively.

The clusters identified by K-Means on the PCA-transformed data correspond well to the geographic categories. For example, districts near the bay or ocean tend to cluster together, reflecting their geographic similarities. The first two principal components capture significant variance related to geographic features, allowing us to see clear groupings that correspond to the categorical values. While there is some overlap between clusters and categories, the PCA transformation helps distinguish groups based on the most significant variance in the data, which often aligns with geographic categories.

The K-Means clustering on the PCA-transformed data reveals meaningful groupings that correspond to geographic categories, indicating that geographic proximity plays a crucial role in the variance captured by the principal components. The clusters are well-defined in the PCA space and align with the categorical values, highlighting the effectiveness of PCA and K-Means in uncovering underlying patterns in the data. This analysis demonstrates how dimensionality reduction and clustering can be used together to reveal and interpret significant structures within the data.

c)

The significant geographic variability and differences in housing characteristics imply that models need to account for these factors to improve predictions. Including geographic coordinates and derived features like mean rooms and mean bedrooms helped capture this variability. Clusters based on geographic proximity and housing characteristics indicate that regional factors play a critical role in housing prices, which models need to capture effectively.

As for the models performances:

- Lasso and Ridge Regression: These models benefited from feature selection and regularization, helping manage multicollinearity and overfitting. However, they might miss capturing complex non-linear relationships.

- Decision Tree Regression: This model captured non-linear relationships better, resulting in lower RMSE values. However, it is more prone to overfitting without proper tuning.

- PCA and Clustering: PCA helped reduce dimensionality and revealed the most significant variance in the data, aiding in more meaningful clustering and model interpretation. The clustering analysis provided insights into the inherent groupings in the data, which could be used to refine model features and improve performance.