CITS5508 Machine Learning

Assignment 3

Mila Zhang (22756463)

THE UNIVERSITY OF WESTERN AUSTRALIA

2024 Sem 1

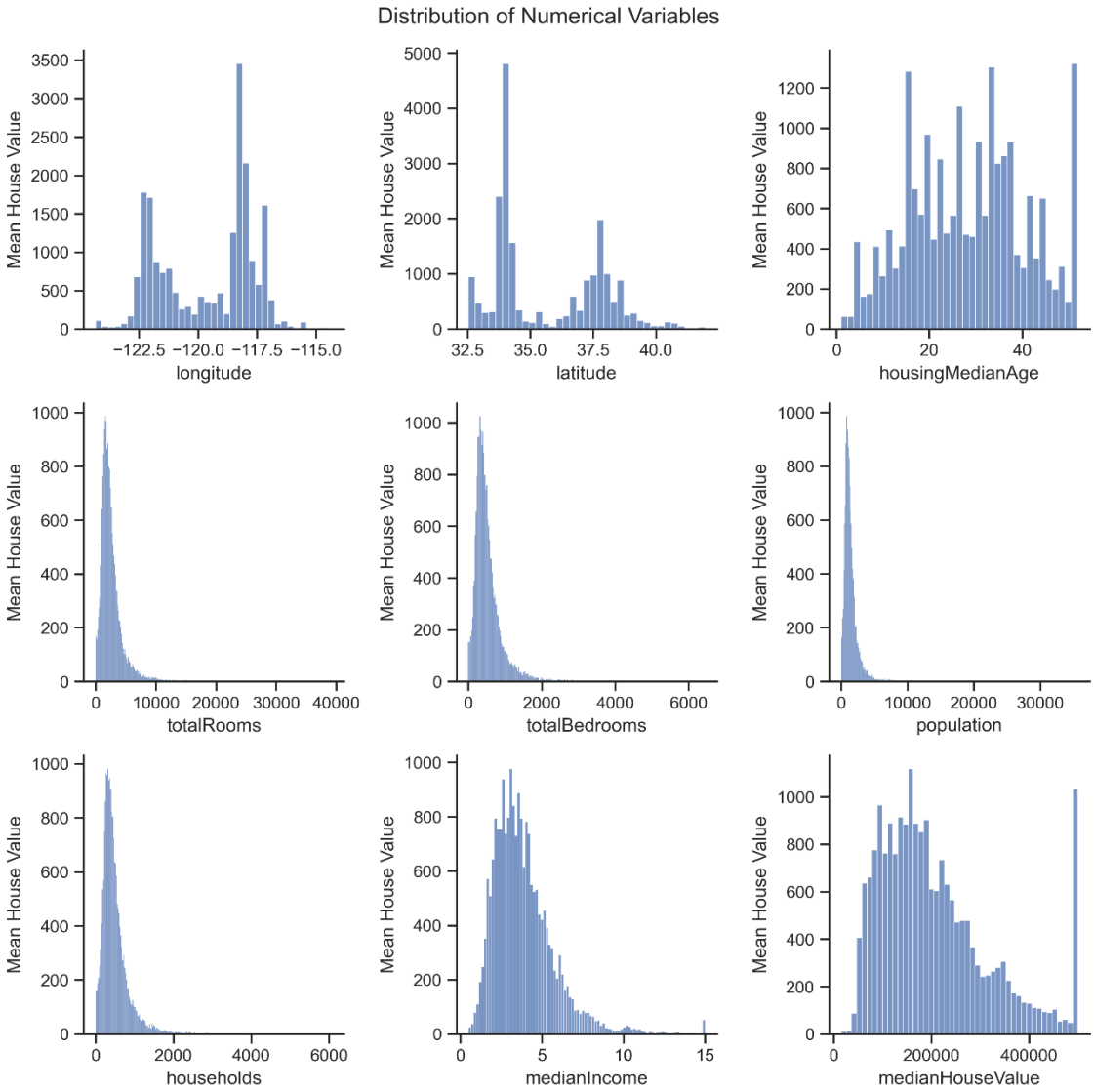
Due date: Friday, 24 May, 8pm

Table of Contents

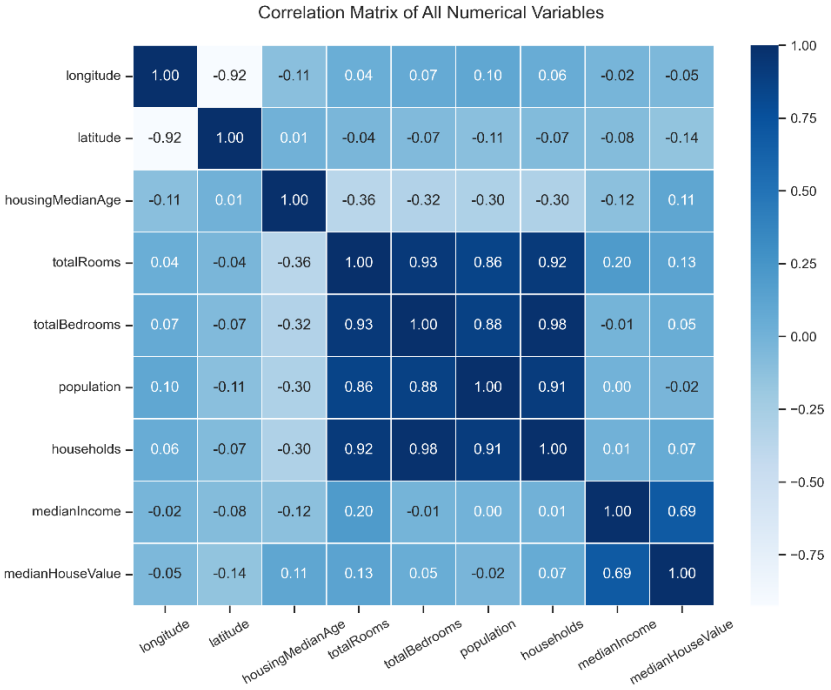
1. Reading the Dataset

1.1 Task D1

1.1 (a) Histograms

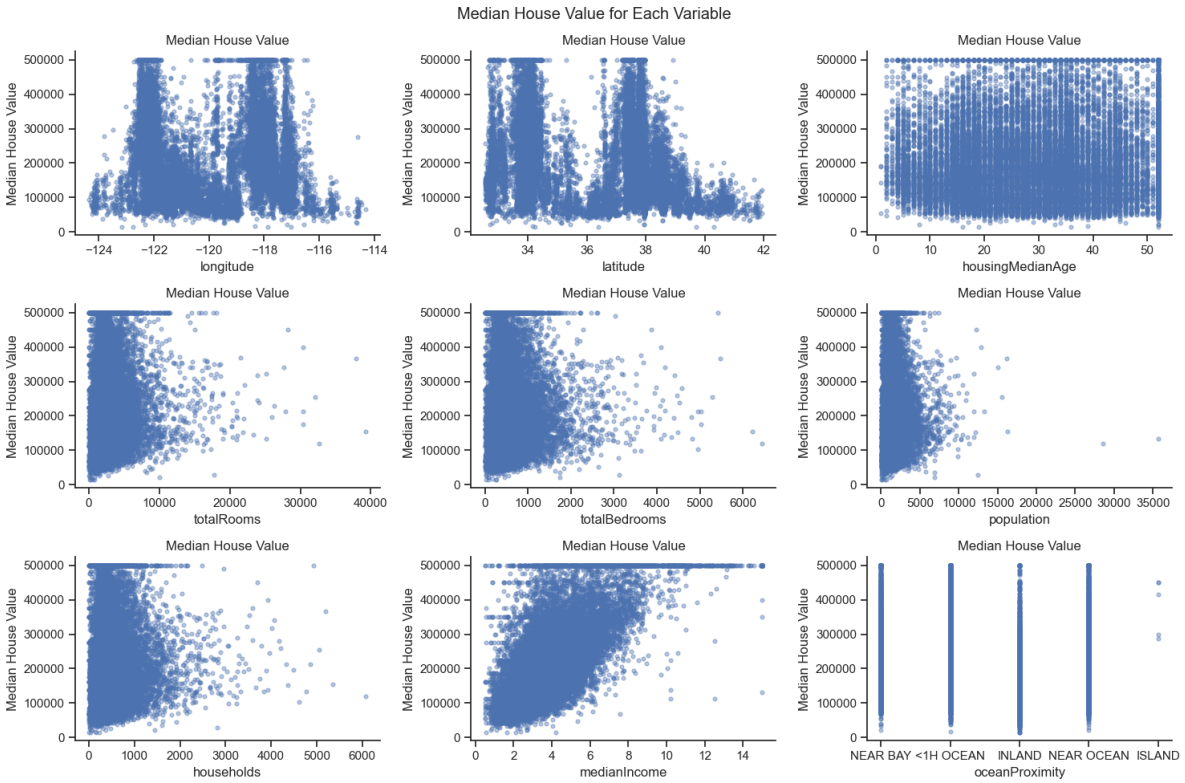


1.1 (b) Correlation Matrix



Features such as “totalRooms”, “totalBedrooms”, “population”, “households” are highly correlated with each other. Because in real life, these features tend to influence or determine one another, resulting in a strong linear relationship. For example, more households typically mean a higher population, and more rooms typically imply more bedrooms.

1.1 (c) Scatter Plot



1.2 Task D2

1.2 (a) Table

RMSE Results

	Linear Regression (Train)	Linear Regression (Test)	Lasso Regression (Train)	Lasso Regression (Test)
Data1	68607.3141	68589.3123	68660.5046	68601.8095
Data1 Standardised	68607.3141	68589.3123	68615.4411	68623.3836
Data2	0.6861	0.6859	1.1294	1.1198
Data2 Standardised	0.6861	0.6859	1.1563	1.1444

1.2 (b)

RMSE is the square root of the average squared differences between the actual and predicted values. In this context, the RMSE values are influenced by the *predictions* made by the models.

For linear regression models, the RMSE values remain *unchanged* after standardization. This is because, when features are standardized, the linear regression model's coefficients are *scaled accordingly*, maintaining the same predictions. The RMSE values for data2 and its standardized version are approximately 1/100,000 of the RMSE values for data1 and its standardized version, because the unit of the target variable was converted to hundreds of thousands of dollars.

For the Lasso regression models, the addition of the regularization term penalizes the coefficients. The L1 penalty term is *sensitive to the scale* of the features. Standardizing the features alters the coefficients, leading to different predictions and hence, different RMSE values.

1.3 Task D3

1.3 (a) Table

RMSE Results

	Linear Regression (Train)	Linear Regression (Test)	Lasso Regression (Train)	Lasso Regression (Test)
Data3	0.7095	1.1360	1.1563	1.1444
Data3 Standardised	0.7095	1.1360	1.1563	1.1444

1.3 (b)

For Linear regression, the RMSE remains *unchanged* after standardization because the model's coefficients are scaled accordingly to maintain consistent predictions.

For Lasso regression, the *identical* RMSE values indicate that standardization had minimal impact on the regularization process. This further implies that the derived three features (meanRooms, meanBedrooms, meanOccupation) effectively capture the variance in the data, rendering the model less sensitive to the scale of these features.

1.3 (c) Estimated coefficients

	Data1 LR	Data1 Lasso	Data1 Standardised LR	Data1 Standardised Lasso	Data2 LR	Data2 Lasso	Data2 Standardised LR	Data2 Standardised Lasso
longitude	-26533.2379	-26398.7585	-53194.8860	-50311.4563	-0.2653	-0.0000	-0.5319	-0.0
latitude	-25444.9108	-25420.7598	-54426.4860	-51488.4957	-0.2544	-0.0000	-0.5443	-0.0
housingMedianAge	1055.9001	1059.8418	13309.9260	13258.9162	0.0106	0.0000	0.1331	0.0
totalRooms	-6.4290	-6.4337	-14090.6494	-12015.2463	-0.0001	0.0001	-0.1409	0.0
totalBedrooms	102.9358	103.3585	43350.0643	41169.5663	0.0010	-0.0000	0.4335	0.0
population	-36.3516	-36.4043	-41771.4951	-41042.1706	-0.0004	-0.0001	-0.4177	-0.0
households	45.1305	44.8074	17290.2404	16763.7830	0.0005	-0.0000	0.1729	0.0
medianIncome	39305.2068	39291.4245	74889.2164	74413.0381	0.3931	0.0000	0.7489	0.0
oceanProximity_INLAND	-39134.8447	-38755.0381	-18231.7216	-19118.7676	-0.3913	-0.0000	-0.1823	-0.0
oceanProximity_ISLAND	153585.7019	0.0000	2672.2075	2593.7778	1.5359	0.0000	0.0267	0.0
oceanProximity_NEAR BAY	-791.4702	-0.0000	-247.4444	-0.0000	-0.0079	0.0000	-0.0025	0.0
oceanProximity_NEAR OCEAN	4935.3229	4206.6297	1648.3297	1736.2550	0.0494	0.0000	0.0165	0.0

	Data3 LR	Data3 Lasso	Data3 Standadised LR	Data3 Standadised Lasso
longitude	-0.2614	-0.0	-0.5241	-0.0
latitude	-0.2481	-0.0	-0.5306	-0.0
housingMedianAge	0.0084	0.0	0.1060	0.0
medianIncome	0.4174	0.0	0.7952	0.0
oceanProximity_INLAND	-0.3814	-0.0	-0.1777	-0.0
oceanProximity_ISLAND	1.5267	0.0	0.0266	0.0
oceanProximity_NEAR BAY	0.0587	0.0	0.0183	0.0
oceanProximity_NEAR OCEAN	0.0839	0.0	0.0280	0.0
meanRooms	-0.0801	0.0	-0.2019	0.0
meanBedrooms	0.4901	-0.0	0.2393	-0.0
meanOccupation	-0.0409	-0.0	-0.0876	-0.0

1.3 (d)

For Data1, without changing the unit of the target variable, the coefficients of the Lasso regression and linear regression are similar for both the raw and the standardized data.

For linear regression models, the coefficients for Data2 and Data3 are proportionally scaled (approximately 1/100,000) compared to Data1, reflecting the change in the target variable's unit.

Lasso regression models consistently shrink coefficients to zero across all datasets, both raw and standardized. This reduction is due to the L1 regularization term, which penalizes large

coefficients. This shows Lasso's strength in feature selection by identifying and retaining only the most significant features.

The coefficients in Data2 and Data3 are particularly small compared to those in Data1, suggesting that the change in the target variable's unit significantly affects the magnitude of the coefficients.

2. Analyzing the impact of different models

2.1 Task D4

Lasso Regression Grid Search Results

Optimal alpha	0.001
RMSE (training set)	0.7096
RMSE (test set)	1.1289

Estimated Parameter Values

	Coefficient
longitude	-0.4961
latitude	-0.5013
housingMedianAge	0.1058
medianIncome	0.7882
oceanProximity_INLAND	-0.1874
oceanProximity_ISLAND	0.0259
oceanProximity_NEAR BAY	0.0183
oceanProximity_NEAR OCEAN	0.0283
meanRooms	-0.1851
meanBedrooms	0.2225
meanOccupation	-0.0867

2.2 Task D5

2.2 (a)

Ridge Regression Grid Search Results

Optimal alpha	100
RMSE (training set)	0.7099
RMSE (test set)	1.1314

Estimated Parameter Values

	Coefficient
longitude	-0.4386
latitude	-0.4419
housingMedianAge	0.1066
medianIncome	0.7813
oceanProximity_INLAND	-0.2044
oceanProximity_ISLAND	0.0271
oceanProximity_NEAR BAY	0.0218
oceanProximity_NEAR OCEAN	0.0323
meanRooms	-0.1733
meanBedrooms	0.2094
meanOccupation	-0.0869

2.2 (b)

The α value from Ridge regression is 100,000 times larger than that of Lasso regression, indicating that Lasso regression requires a small amount of regularization to achieve optimal model performance, while Ridge regression requires a higher degree of regularization to achieve the optimal balance between bias and variance.

The coefficients values are generally similar in Lasso regression and Ridge regression, indicating that both models recognize similar strengths of impact for each independent variable, likely leading to similar optimal predictions. Both models consistently identify key predictors such as "medianIncome," "latitude," and "longitude." Specifically, both models recognize "medianIncome" as a significant positive predictor of the target variable, with coefficients of 0.7882 for Lasso and 0.7813 for Ridge. Additionally, both models identify

"latitude" and "longitude" as significant negative predictors of the target variable, with similar coefficient values.

2.3 Task D6

Decision Tree Regression Grid Search Results

Optimal 'max_depth'	9
RMSE (training set)	0.5027
RMSE (test set)	0.6015

2.4 Task D7

2.4 (a)

RMSE values: decision tree regression < Lasso regression < Ridge regression.

The decision tree regression model has the lowest RMSE for test set. The Ridge regression and Lasso regression have similar but higher RMSE.

The decision tree regression model is the best model among the three based on the RMSE on the test set. The lower RMSE indicates better predictive performance and generalization to unseen data. Decision Trees can model non-linear relationships and interactions between features more effectively than linear models like Lasso and Ridge regression.

2.4 (b)

According to EDA analysis in D1, we can observe some anomalies or outliners (e.g. in “totalRooms” and “population”) that are separated from the rest of the data point groups. Remove outliers to prevent them from disproportionately influencing the model.

Variables that are highly skewed can distort model performance and lead to biased outcomes. Applying log transformations to these variables can normalize their distributions, making them more suitable for modeling.

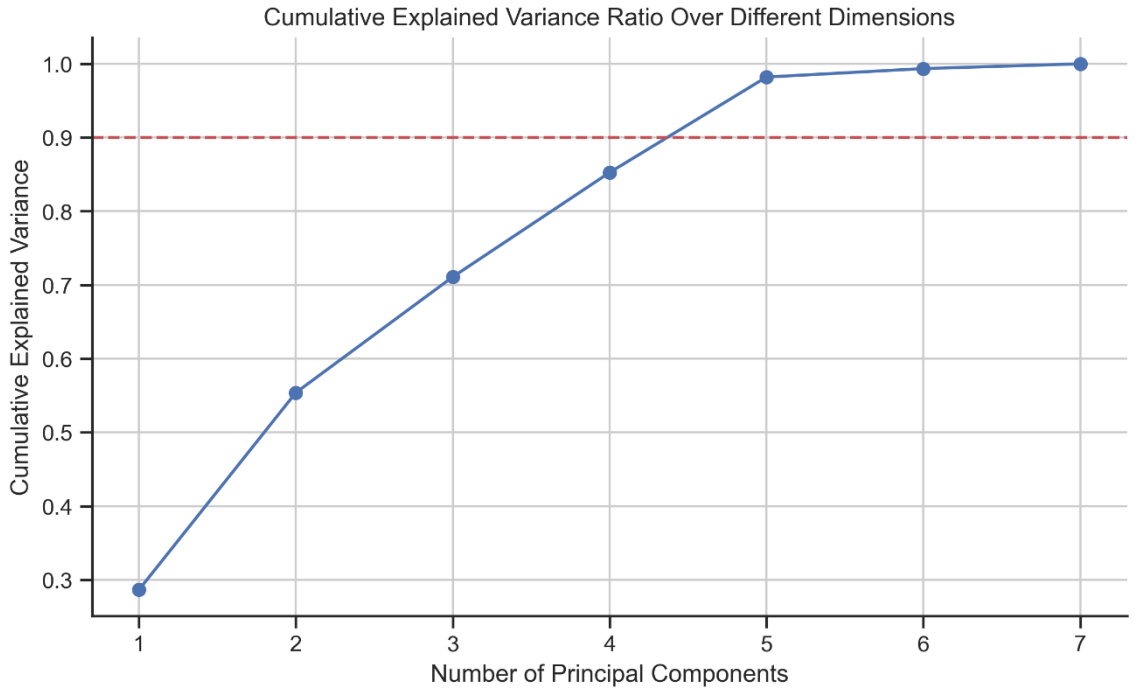
Highly correlated features can inflate the variance of coefficient estimates and lead to overfitting. To address this, features that are highly correlated should be removed. Alternatively, applying PCA can reduce dimensionality by transforming the original correlated features into a set of uncorrelated components.

Creating new derived features can significantly enhance the predictive power of a model. Derived features, such as polynomial features, can capture additional relationships within the data that were not apparent in the original features and thus improve model performance.

2.5 Task D8

Assume dummy variables created from the categorical variable are considered as categorical variables, and thus removed in the data used in D8.

2.5 (a)



2.5 (b)

The number of principal components necessary to preserve at least 90% of variance: **5**.

2.5 (c)

Linear Regression Using the Selected Number of Principal Components (>90% Variance)

RMSE (training set)	0.8059
RMSE (test set)	1.3394

2.5 (d)

Principal Components Grid Search Results

Optimal Number of Principal Components	7
RMSE (training set)	0.7189
RMSE (test set)	1.1784

2.5 (e)

Comparing (c) and (d):

The model using the optimal number of principal components, outperforms the model that retains over 90% of variance. The optimal one results in lower RMSE for both training and test sets, indicating better model performance and generalization. This highlights the importance of using a balanced number of principal components to capture essential variance while preventing overfitting, thus improving predictive capacity.

Comparing (c), (d) with D7:

RMSE values: decision tree regression < Lasso regression < Ridge regression < PCA linear regression.

Although PCA helps in reducing dimensionality, the linear regression model does not perform as well as Decision Tree Regression or the regularized linear models (Lasso and Ridge) in this case. This suggests that although PCA is useful for managing multicollinearity and reducing overfitting, it may not always capture complex relationships as effectively as other techniques.

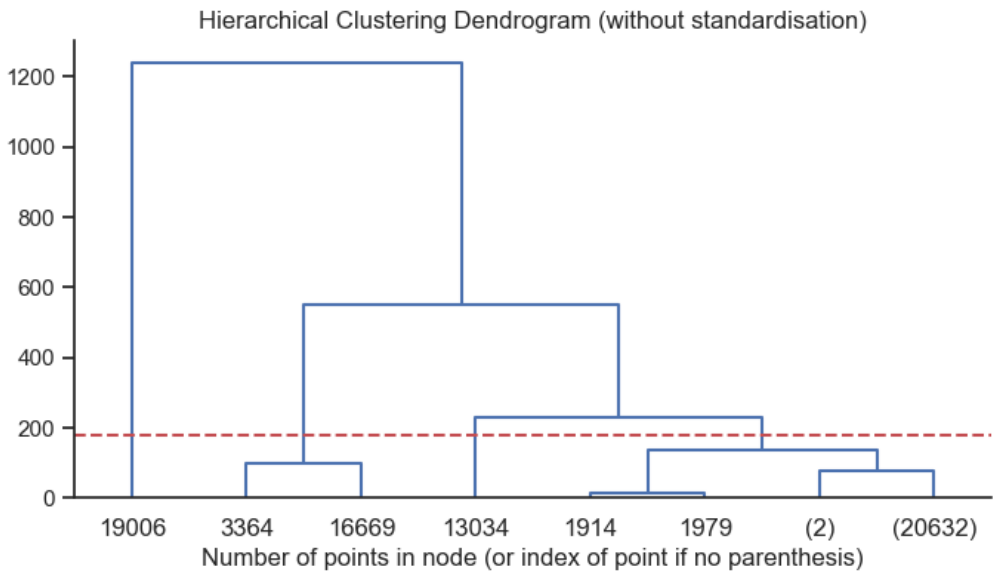
Based on RMSE values, Decision Tree Regression is the best-performing model for this dataset, with the lowest RMSE on both the training and test sets.

Lasso and Ridge regression offer balanced performance, with Lasso slightly outperforming Ridge due to its feature selection capability.

3. Clustering analysis

3.1 Task D9

3.1 (a)



Mean of Variables for Each Cluster (rounded to 2 decimal places)

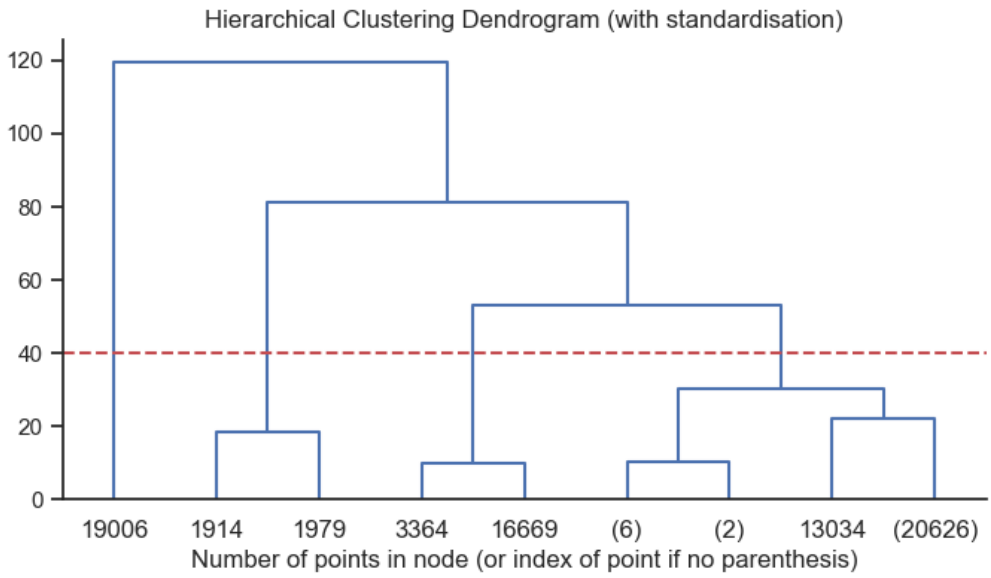
Cluster	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOccupation
0	-119.57	35.63	28.64	3.87	5.43	1.10	2.95
1	-121.98	38.32	45.00	10.23	3.17	0.83	1243.33
2	-120.60	37.86	41.00	4.89	7.11	1.23	551.09
3	-121.15	38.69	52.00	6.14	8.28	1.52	230.17

Cluster Size

Cluster	Size
0	20636
2	2
3	1
1	1

The clusters are imbalanced. Cluster 0 is the largest while other clusters are very small, with only 1 or 2 instances in each cluster. Cluster 0 has the youngest housing age, lowest median income, average home size, and smallest household size. Cluster 1 has older housing age, highest median income, smallest homes, and largest household size. Cluster 2 has moderate housing age, median income, larger homes, and larger household size. Cluster 3 has the oldest housing age, high median income, largest homes, and moderate household size.

3.1 (b)



Mean of Variables for Each Cluster

Cluster_Scaled	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOccupation
0	0.0001	-0.0003	-0.0002	-0.0002	-0.0052	-0.0059	-0.0109
1	-0.5167	1.0455	0.9821	0.5370	0.6794	0.2714	52.7660
2	-0.2597	1.5090	0.3862	-0.3267	53.2685	60.6771	-0.0488
3	-1.2031	1.2586	1.3000	3.3455	-0.9144	-0.5557	119.4191

Cluster Size

Cluster	Size
0	20635
2	2
1	2
3	1

Both methods identified one very large cluster and three very small clusters. After scaling the variables, the size of clusters has changed slightly, with cluster 0 going from 20636 to 20635 and cluster 1 going from 1 to 2. The magnitude of mean values is much smaller compared to those in D9(a). Scaling the variables ensures that each variable contributes equally to the distance calculations used in clustering, preventing variables with large numerical ranges from disproportionately influencing the clustering outcome. For example, in cluster 2, variables with smaller values like “meanRooms” and “meanBedrooms” are considered as more significant, and variables with big values like “meanOccupation” has reduced disproportionate impact, allowing for more balanced consideration across all features.

3.1 (c)

Mean of Variables for Each Cluster

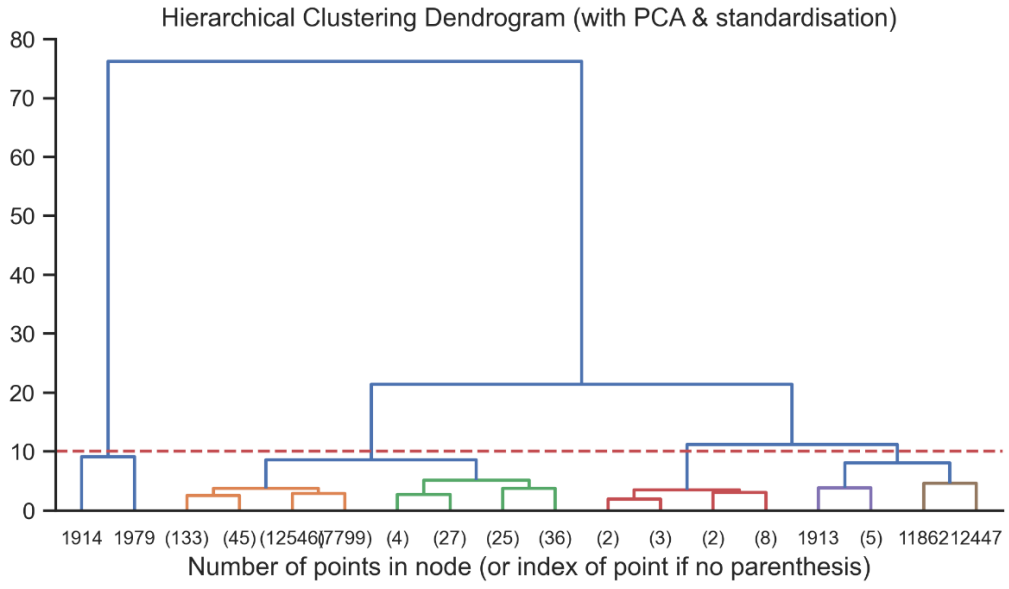
Cluster_Scaled	longitude	latitude	housingMedianAge	medianIncome	meanRooms	meanBedrooms	meanOccupation
0	0.0001	-0.0003	-0.0002	-0.0002	-0.0052	-0.0059	-0.0109
1	-0.5167	1.0455	0.9821	0.5370	0.6794	0.2714	52.7660
2	-0.2597	1.5090	0.3862	-0.3267	53.2685	60.6771	-0.0488
3	-1.2031	1.2586	1.3000	3.3455	-0.9144	-0.5557	119.4191

Cluster Size

Cluster	Size
0	20635
2	2
1	2
3	1

The results obtained from part (b) and part (c) are identical and do not indicate differences in performance. However, in this context, hierarchical clustering provides a deterministic and hierarchical structure and useful for understanding the nested relationships between clusters. K-means is largely influenced by the initialization of the centroids and is more complicated than hierarchical clustering. Therefore, in this context, hierarchical clustering might be slightly better.

3.1 (d)

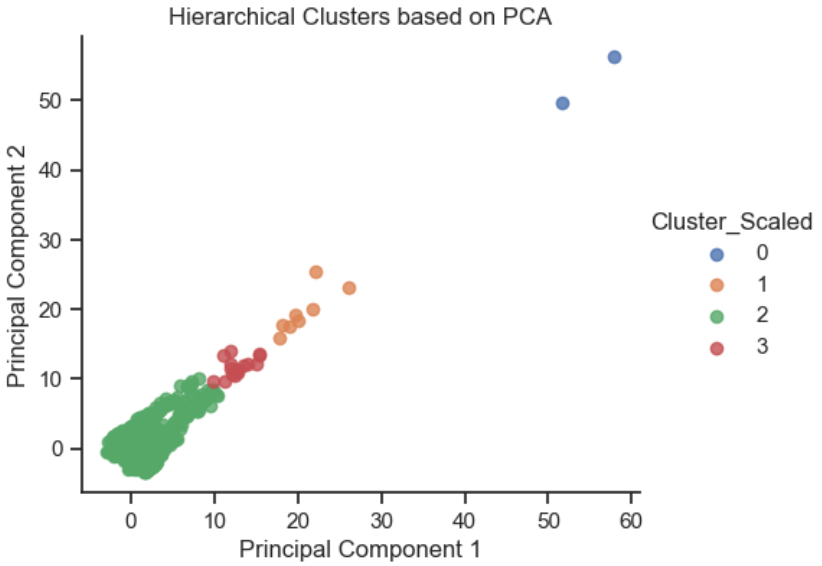


Mean Scores for Each Principal Component

Cluster_Scaled	pca0	pca1
0	54.8093	52.9323
1	20.5947	19.6582
2	-0.0226	-0.0214
3	12.7745	11.8015

Cluster Size

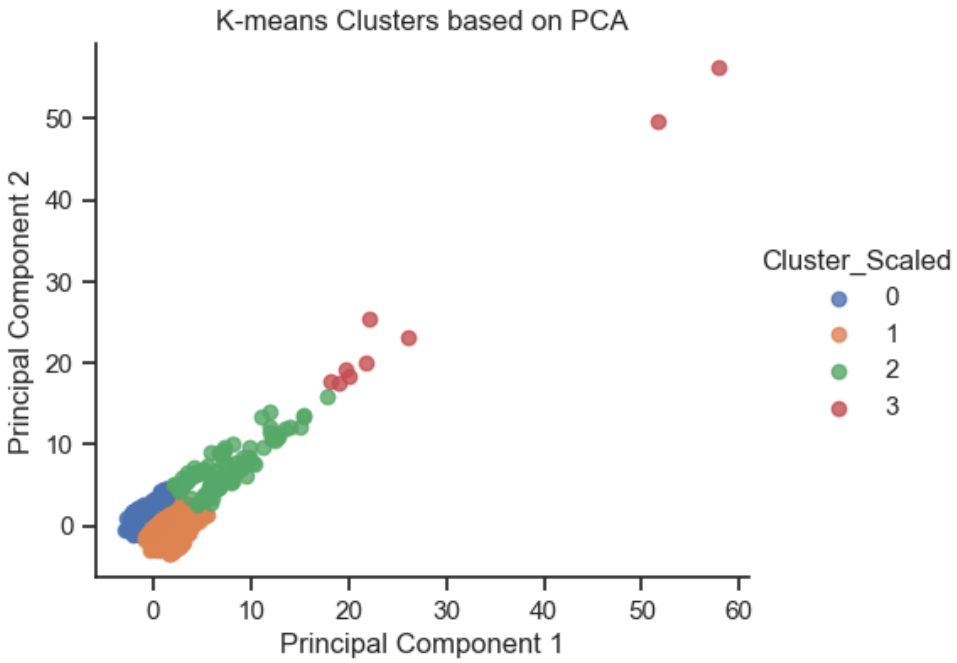
Cluster	Size
2	20615
3	15
1	8
0	2



All these methods identified one very large cluster and three very small clusters. However, the application of PCA followed by hierarchical clustering significantly improved the clustering outcomes by producing *more balanced cluster sizes*. The largest cluster shrank to 20,615 data points, while the sizes of the other clusters increased slightly. This indicates that the combination of PCA and hierarchical clustering effectively captured the underlying structure of the data, leading to a more accurate and interpretable clustering solution.

By reducing noise and redundancy, highlighting significant patterns, and forming stable groupings, PCA and hierarchical clustering together provide a robust framework for clustering complex datasets. Compared to other methods, this approach not only enhances the performance of the clustering algorithm but also ensures that the resulting clusters are meaningful and actionable for further analysis.

3.1 (e)



Mean Scores for Each Principal Component

Cluster_Scaled	pca0	pca1
0	-0.8790	0.6773
1	1.1031	-1.0571
2	7.0551	6.7874
3	28.5090	27.4786

Cluster Size

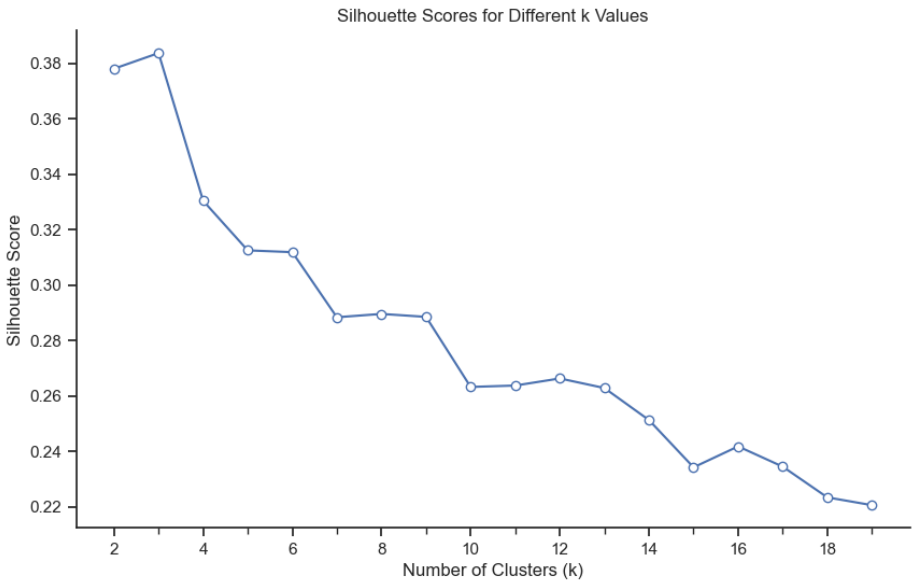
Cluster	Size
0	11939
1	8583
2	109
3	9

The application of PCA followed by k-means clustering has significantly improved the clustering outcomes, resulting in more balanced cluster sizes. The formation of two large clusters and two small clusters suggests that, compared to previous approaches, this approach has effectively distributed the data points, uncovering meaningful groupings within the dataset.

According to the scatter plot, cluster 3 includes two separated data points located in the upper right corner. These points are potentially outliners or anomalies. This approach effectively identifies clusters without designating outliers as a separate cluster, thereby reducing the impact of outliers and minimizing bias. By reducing the impact of outliers, the clustering results are more reflective of the true structure of the data, leading to more accurate and actionable insights compared to other approaches.

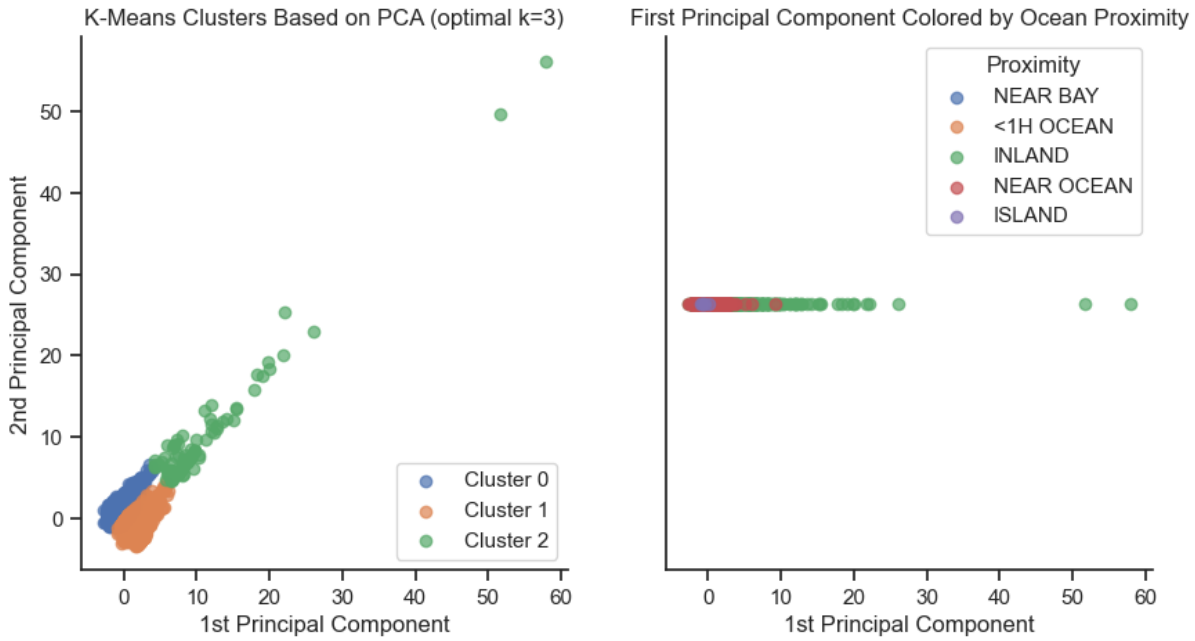
3.2 Task D10

3.2 (a)



According to this score, the optimal value of clustering is **3**, with the highest silhouette score.

3.2 (b)



From the left figure, we can see that cluster 0 and 1 are centered around the value 0 of the first principal component. Similarly, the right figure shows that the “ISLAND” and “NEAR OCEAN” categories are also located near the value 0. This suggests a possible correspondence between these clusters and these ocean proximity categories.

Additionally, the left figure indicates that cluster 2 spans a wide range, from approximately 5 to 60 on the first principal component. Correspondingly, the “INLAND” category in the right figure appears to distribute across this same range of values. This observation implies that the "INLAND" category may be associated with cluster 2.

These visual patterns suggest that certain categories of ocean proximity might correspond to different clusters, reflecting underlying relationships in the data.

3.2 (c)

Conclusions about the data that may be impacting models:

Highly skewed data can lead to biased models and inaccurate predictions. Standardizing the features helps mitigate this issue by reducing the disproportionate impact of variables, ensuring a more balanced consideration across all features.

Highly correlated features can affect the stability of model coefficients and lead to overfitting. Removing highly correlated variables or performing PCA can reduce multicollinearity. PCA transforms correlated features into a set of linearly uncorrelated components, improving model robustness and preventing overfitting.

Deriving features from existing features allows the model to learn relationships that were not apparent in the original features and thus improve the model’s ability to make accurate predictions.

Outliners and anomalies in data may distort the relationships between features, leading to misleading correlation and thus affecting the accuracy of the predictions. These extreme values can skew the model's understanding of the data, resulting in biased and unreliable predictions.