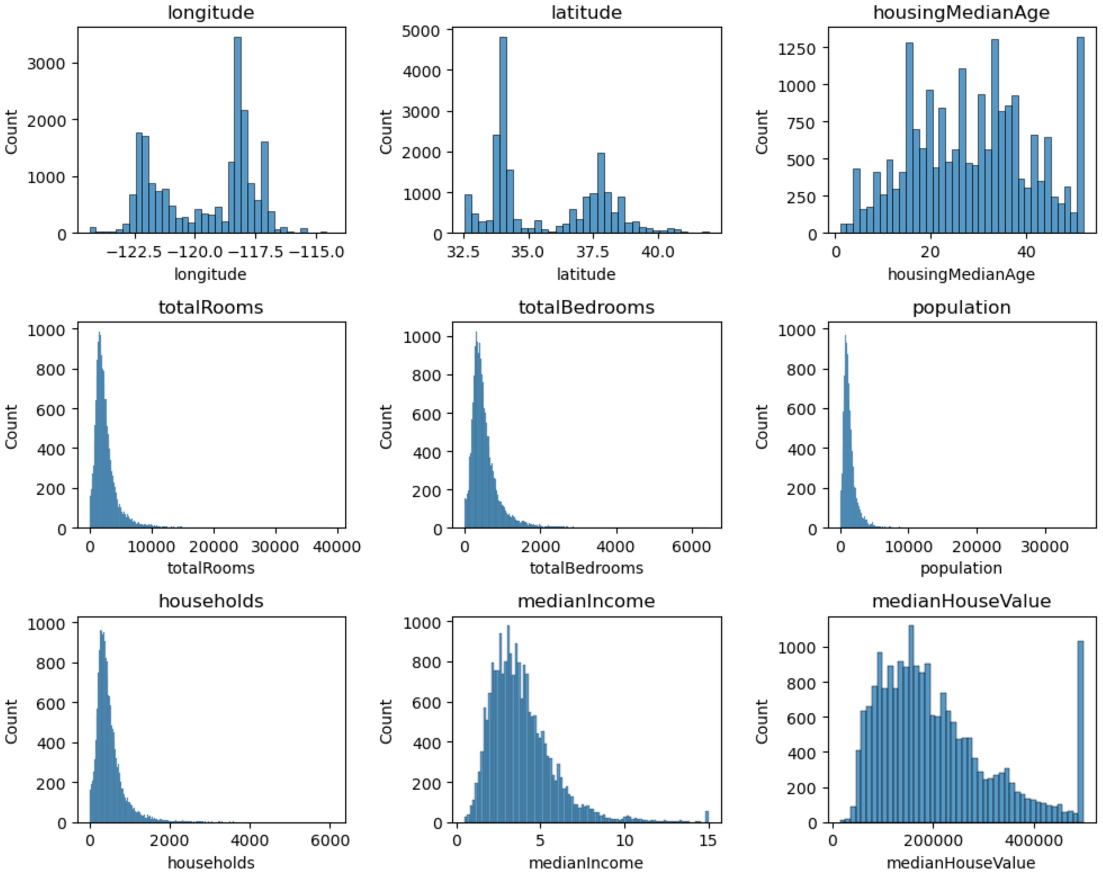
Assignment 3 – Allison Lau (23123849)

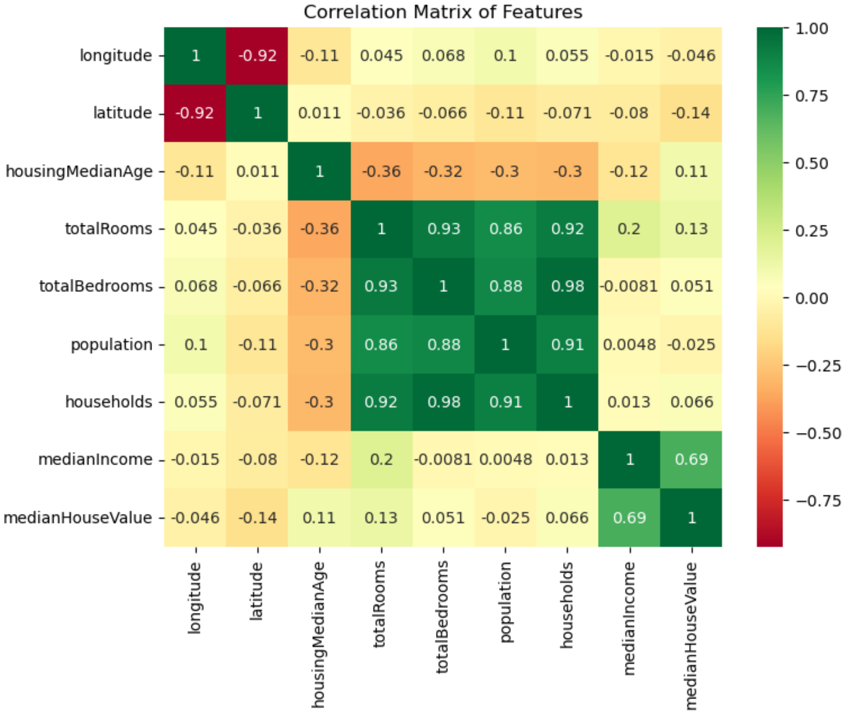
1 Reading the dataset

D1 Exploratory data analysis and preprocessing

(a) Histograms



(b) Correlation Matrix

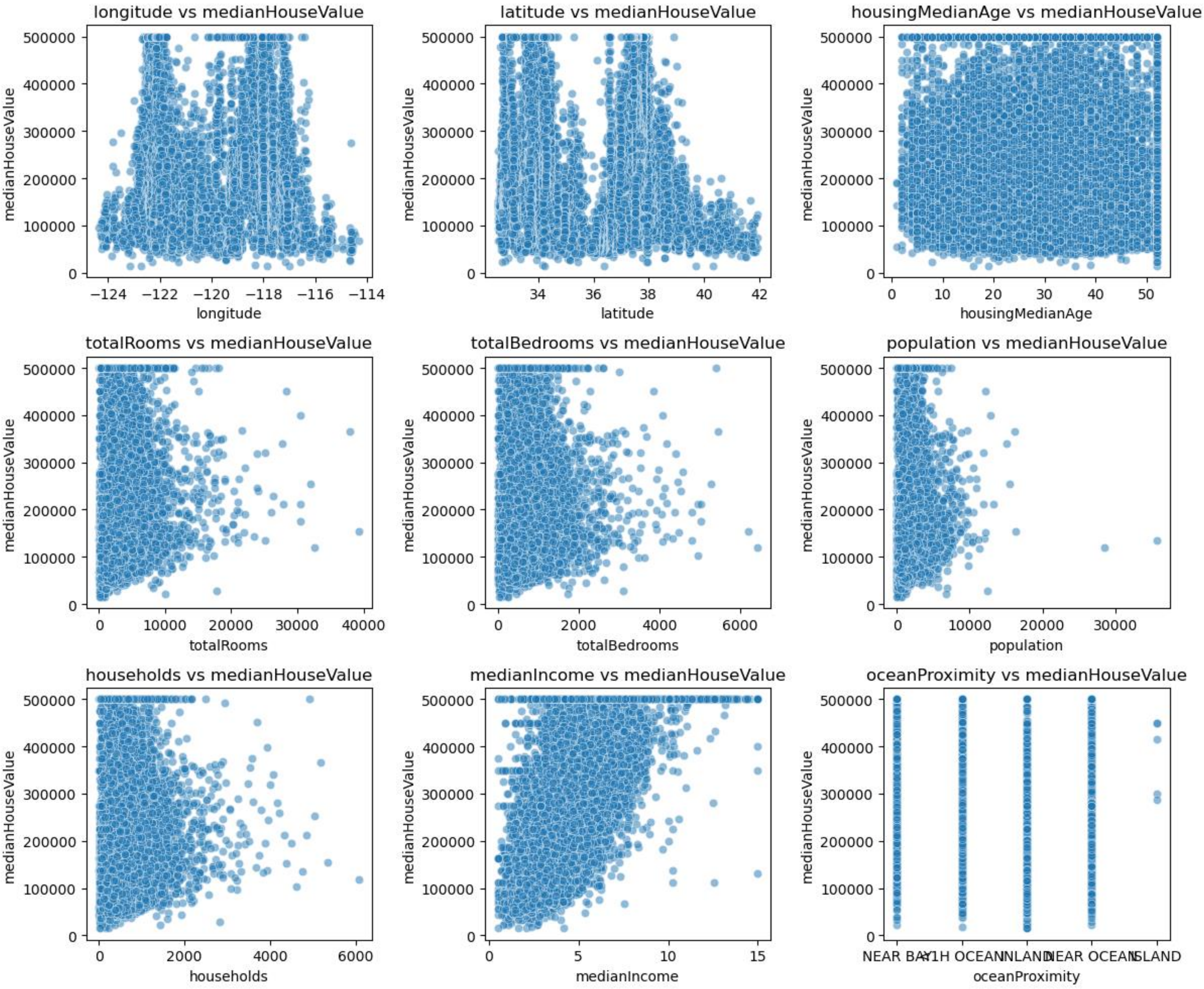

Correlation Matrix of Features

	longitude	latitude	housingMedianAge	totalRooms	totalBedrooms	population	households	medianIncome	medianHouseValue
longitude	1	-0.92	-0.11	0.045	0.068	0.1	0.055	-0.015	-0.046
latitude	-0.92	1	0.011	-0.036	-0.066	-0.11	-0.071	-0.08	-0.14
housingMedianAge	-0.11	0.011	1	-0.36	-0.32	-0.3	-0.3	-0.12	0.11
totalRooms	0.045	-0.036	-0.36	1	0.93	0.86	0.92	0.2	0.13
totalBedrooms	0.068	-0.066	-0.32	0.93	1	0.88	0.98	-0.0081	0.051
population	0.1	-0.11	-0.3	0.86	0.88	1	0.91	0.0048	-0.025
households	0.055	-0.071	-0.3	0.92	0.98	0.91	1	0.013	0.066
medianIncome	-0.015	-0.08	-0.12	0.2	-0.0081	0.0048	0.013	1	0.69
medianHouseValue	-0.046	-0.14	0.11	0.13	0.051	-0.025	0.066	0.69	1

totalRooms, **totalBedrooms**, **population** and **households** have high positive correlation (above 0.8) with each other. **totalRooms** and **totalBedrooms** have high positive correlation because houses with more bedrooms typically have more total number of rooms. **households** have high positive correlation with **totalRooms** and **totalBedrooms** because the number of rooms/bedrooms in a house is typically proportional to the number of households. **population** and **household** have high positive correlation because a higher population in an area typically means more households.

longitude and **latitude** have high negative correlation (-0.92), suggesting that the region is aligned in a diagonal direction which creates a pattern where increases in one feature correspond to decreases in the other feature.

(c) Scatter plot



2 Analysing the impact of different data transformations

D2 Original and standardised data1 and data2 for Linear and Lasso Regression

(a) Report table of RMSE values for training and test sets

Type of Dataset	Linear Training RMSE	Linear Test RMSE	Lasso Training RMSE	Lasso Test RMSE
Original data1	68607.314131	68589.312340	68660.504643	68601.809500
Standardised data1	68607.314131	68589.312340	68615.441095	68623.383563
Original data2	0.686073	0.685893	1.129396	1.119761
Standardised data2	0.686073	0.685893	1.156303	1.144382

(b) Discussion of RMSE values obtained results

Linear regression models produced the same RMSE values for both original and standardised data. This is because it is scale invariant as it relies on the ratios between the features and not their magnitudes. Hence, the scaling of the input features does not affect the RMSE values.

Lasso regression models produced different RMSE values for original and standardised data. This is because it includes an L1 regularisation term that places a penalty on the magnitude of the coefficients. The scale of variables will affect how much of a penalty will be applied on their coefficients. Without standardising the data, variables with larger scales can dominate the penalty, leading to less effective regularisation. Standardisation ensures that all variables are on the same scale, allowing for balanced penalisation and different RMSE values.

The RMSE values for data2 are in decimals because the target variable has been transformed to hundreds of thousands of dollars. This causes the target variable to be smaller in magnitude which also reduces the magnitude of the RMSE values.

D3 Original and standardised data3 for Linear and Lasso Regression

(a) Report table of RMSE values for training and test sets

Type of Dataset	Linear Training RMSE	Linear Test RMSE	Lasso Training RMSE	Lasso Test RMSE
Original data3	0.70949	1.13601	1.156303	1.144382
Standardised data3	0.70949	1.13601	1.156303	1.144382

(b) Discussion and justification of obtained RMSE values

Lasso regression now produces the same RMSE values for both original and standardised data. This is due to the replacement of highly correlated features with new features (meanRooms, meanBedrooms, meanOccupation) which are likely less correlated with each other and more uniformly scaled. This means that they tend to have similar variances, allowing the Lasso regularisation term to impact them uniformly and resulting in the same RMSE values for both the original and standardised datasets.

(c) Report tables of estimated parameter values with corresponding variable names

Linear Regression Models on Original Data

Data1

Feature	Coefficient
longitude	-26533.237894
latitude	-25444.910842
housingMedianAge	1055.900145
totalRooms	-6.428986
totalBedrooms	102.935752
population	-36.351577
households	45.130509
medianIncome	39305.206768
INLAND	-39134.844696
ISLAND	153585.701929
NEAR BAY	-791.470246
NEAR OCEAN	4935.322875

Data2

Feature	Coefficient
longitude	-0.265332
latitude	-0.254449
housingMedianAge	0.010559
totalRooms	-0.000064
totalBedrooms	0.001029
population	-0.000364
households	0.000451
medianIncome	0.393052
INLAND	-0.391348
ISLAND	1.535857
NEAR BAY	-0.007915
NEAR OCEAN	0.049353

Data3

Feature	Coefficient
longitude	-0.261440
latitude	-0.248051
housingMedianAge	0.008409
medianIncome	0.417373
INLAND	-0.381382
ISLAND	1.526743
NEAR BAY	0.058689
NEAR OCEAN	0.083880
meanRooms	-0.080115
meanBedrooms	0.490103
meanOccupation	-0.040862

Linear Regression Models on Standardised Data

Data1

Feature	Coefficient
longitude	-53194.886029
latitude	-54426.485960
housingMedianAge	13309.925998
totalRooms	-14090.649431
totalBedrooms	43350.064293
population	-41771.495079
households	17290.240437
medianIncome	74889.216380
INLAND	-18231.721622
ISLAND	2672.207538
NEAR BAY	-247.444446
NEAR OCEAN	1648.329735

Data2

Feature	Coefficient
longitude	-0.531949
latitude	-0.544265
housingMedianAge	0.133099
totalRooms	-0.140906
totalBedrooms	0.433501
population	-0.417715
households	0.172902
medianIncome	0.748892
INLAND	-0.182317
ISLAND	0.026722
NEAR BAY	-0.002474
NEAR OCEAN	0.016483

Data3

Feature	Coefficient
longitude	-0.524144
latitude	-0.530580
housingMedianAge	0.105996
medianIncome	0.795231
INLAND	-0.177674
ISLAND	0.026564
NEAR BAY	0.018349
NEAR OCEAN	0.028015
meanRooms	-0.201913
meanBedrooms	0.239342
meanOccupation	-0.087564

Lasso Regression Models on Original Data

Data1

Feature	Coefficient
longitude	-26398.758516
latitude	-25420.759759
housingMedianAge	1059.841818
totalRooms	-6.433659
totalBedrooms	103.358469
population	-36.404325
households	44.807397
medianIncome	39291.424531
INLAND	-38755.038140
ISLAND	0.000000
NEAR BAY	0.000000
NEAR OCEAN	4206.629661

Data2

Feature	Coefficient
longitude	-0.000000
latitude	-0.000000
housingMedianAge	0.000000
totalRooms	0.000104
totalBedrooms	-0.000000
population	-0.000118
households	-0.000000
medianIncome	0.000000
INLAND	-0.000000
ISLAND	0.000000
NEAR BAY	0.000000
NEAR OCEAN	0.000000

Data3

Feature	Coefficient
longitude	-0.0
latitude	-0.0
housingMedianAge	0.0
medianIncome	0.0
INLAND	-0.0
ISLAND	0.0
NEAR BAY	0.0
NEAR OCEAN	0.0
meanRooms	0.0
meanBedrooms	-0.0
meanOccupation	-0.0

Lasso Regression Models on Standardised Data

Data1

Feature	Coefficient
longitude	-50311.456263
latitude	-51488.495689
housingMedianAge	13258.916154
totalRooms	-12015.246255
totalBedrooms	41169.566341
population	-41042.170587
households	16763.782993
medianIncome	74413.038143
INLAND	-19118.767645
ISLAND	2593.777782
NEAR BAY	0.000000
NEAR OCEAN	1736.254995

Data2

Feature	Coefficient
longitude	-0.0
latitude	-0.0
housingMedianAge	0.0
totalRooms	0.0
totalBedrooms	0.0
population	-0.0
households	0.0
medianIncome	0.0
INLAND	-0.0
ISLAND	0.0
NEAR BAY	0.0
NEAR OCEAN	0.0

Data3

Feature	Coefficient
longitude	-0.0
latitude	-0.0
housingMedianAge	0.0
medianIncome	0.0
INLAND	-0.0
ISLAND	0.0
NEAR BAY	0.0
NEAR OCEAN	0.0
meanRooms	0.0
meanBedrooms	-0.0
meanOccupation	-0.0

(d) Discussion of obstained results from above tables
The signs of coefficients for the same features tend to be consistent across all models. For example, ‘longitude’ and ‘latitude’ coefficients are consistently negative, whereas ‘housingMedianAge’ and ‘medianIncome’ are consistently positive for all models. The consistency in the signs of the coefficients for these features suggests a robust relationship between these features and the target variable, regardless of different data transformations.

Lasso model on original and standardised data1 did not shrink most of the coefficients to zero, meaning all features were considered important. However, all coefficients in data2 and data3 (both original and standardised) were shrunk to zero by the Lasso model. This suggests that the Lasso model found these features to be less important for predicting the target variable. This may be due to the data transformation applied to the target variable in data2 and data3 to be expressed in hundreds of thousands of dollars.

3 Analysing the impact of different models

D4 Lasso Regression Model with optimised alpha value

Optimal α value : 0.01
Training set RMSE : 0.7158
Test set RMSE : 1.0706

Feature (Variable name)	Coefficient (Estimated parameter values)
longitude	-0.246387
latitude	-0.240308
housingMedianAge	0.103832
medianIncome	0.724969
INLAND	-0.274275
ISLAND	0.019529
NEAR BAY	0.017847
NEAR OCEAN	0.031162
meanRooms	-0.033403
meanBedrooms	0.070794
meanOccupation	-0.078468

D5 Ridge Regression Model with optimised alpha value

(a) Report results

Optimal α value : 100
Training set RMSE : 0.7099
Test set RMSE : 1.1314

Feature (Variable name)	Coefficient (Estimated parameter values)
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
meanOccupation	-0.086887

(b) Compare Lasso Regression and Ridge Regression

Lasso regression is used for feature selection by shrinking some coefficients to exactly zero. The optimal α value for Lasso is 0.01, indicating a strong penalty being applied to the coefficients. Ridge regression is used for feature selection by shrinking the coefficients towards zero but not exactly zero. The optimal α value for Ridge is 100, indicating a light penalty being applied to the coefficients. This can be observed by the generally smaller coefficient magnitudes by the Lasso model, compared to Ridge model.

Lasso Regression model resulted in a slightly higher training RMSE but a lower test RMSE compared to the Ridge Regression model. This suggests that Lasso's feature selection helped to generalise better to the unseen test data, while Ridge might be overfitting slightly more to the training data.

D6 Decision Tree Regression Model with optimised max_depth value

Optimal max_depth : 9
Training set RMSE : 0.5027
Test set RMSE : 0.6015

D7 Compare models from D4, D5, and D6

(a) Discussion on the best model and their RMSE values

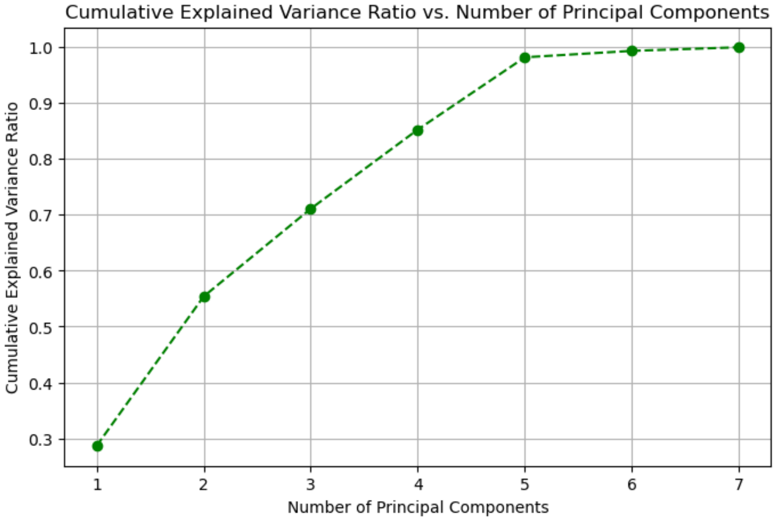
Decision Tree Regression model performed the best among the models because it yielded the lowest RMSE on the test set with 0.6015, compared to Lasso regression model with 1.0706 and Ridge regression model with 1.1314. This means that its predictions are closer to the actual values in the test set compared to the other models, meaning it has the best generalisation and predictive capacity on the test set. This may be because of its ability to capture non-linear relationships in the data more effectively than linear models like Lasso and Ridge regression.

(b) Discuss how the predictive capacity of the models could be improved from EDA

Observing the correlation matrix in D1 suggests that many features have non-linear relationships with each other, as indicated by their correlation coefficients not being close to 1 or -1. Non-linear or complex models can be used instead of linear models to better capture these non-linear relationships and improve predictive capacity. Additionally, dimensionality reduction techniques such as Principal Component Analysis (PCA) can help reduce the number of features while preserving important information, to reduce noise and multicollinearity which can improve the performance of the models. Also, anomaly detection techniques can be used to identify and handle outliers that may skew the predictions to improve the model's robustness.

D8 Principal Component Analysis

(a) Plot cumulative explained variance ratio against number of principal components



(b) Number of principal components necessary to preserve 90% of variance = 5

(c) Linear Regression Model trained using 5 principal components

Training set RMSE : 0.8059

Test set RMSE : 1.3394

(d) GridSearchCV with 10-fold cross-validation

Optimal number of principal components : 7

Training set RMSE : 0.7189

Test set RMSE : 1.1784

(e) Discussion of the obtained results and compare with models in D7

The Linear Regression Model trained using 5 principal components has higher RMSE values for both training and test sets compared to using 7 principal components. This highlights the importance of selecting an appropriate number of principal components when using PCA for dimensionality reduction. Using too few may not be sufficient to capture the variability in the data. This shows the effectiveness of using grid search with cross-validation for selecting the optimal number of principal components. This method can effectively identify the number of components required to still capture the patterns in the data without including noise. The RMSE values for training set and test set for both models in D8 were higher than the RMSE values obtained by all models in D7.

4 Clustering analysis

D9 Different Clustering Techniques

(a) Hierarchical Clustering with average linkage and Euclidean distance (original data)

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

Assumption: Dendrogram plot was not asked to be presented but can be found in the code. **Cluster 1** has the largest size (20636), it has lower *medianIncome* compared to the other clusters, which indicate a mix of middle to lower-middle income households. **Cluster 2** only has a size of 1, it has a very high *medianIncome* which indicates an affluent area. It has a significantly high *meanOccupation* which suggests that it is a commercial or non-residential area. **Cluster 3** has a size of 2, it has a moderate *medianIncome* and *medianHouseValue*, indicating middle-income. It has a high *meanRooms* and *meanOccupation* count which suggests larger homes, may be a dormitory or communal living. **Cluster 4** also only has a size of 1, it has a higher *housingMedianAge* than the rest of the clusters, with a moderately high *medianIncome* and *meanOccupation*. Clusters 2, 3, and 4 might be outliers or errors in the data.

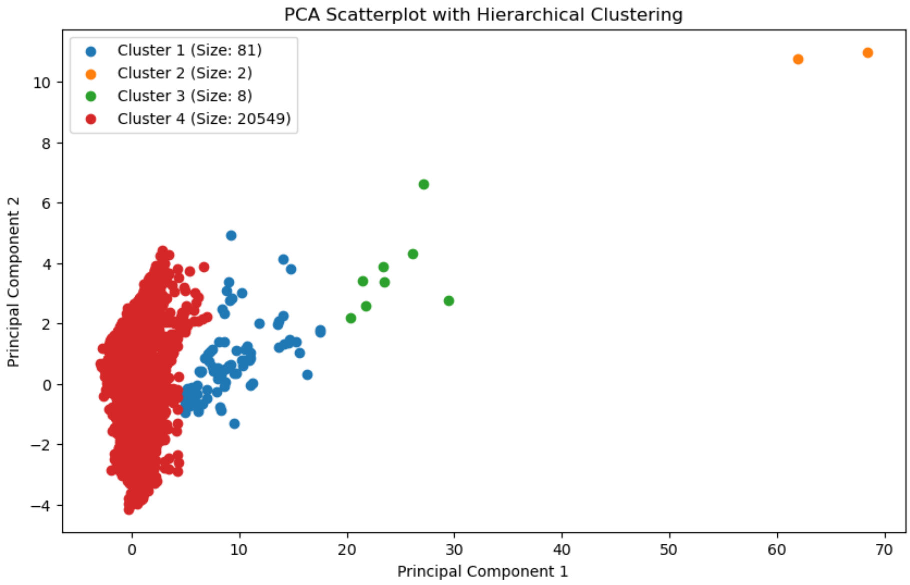
(b) Hierarchical Clustering with average linkage and Euclidean distance (standardised)

Assumption: Results and dendrogram plot were not asked to be presented but can be found in the code. The size of the clusters remains mostly consistent with the largest cluster (Cluster 1) containing the majority of the data points, whereas clusters 2, 3, and 4 each still have a very small number of data points, indicating potential outliers or errors in the data. Cluster 1 has a size of 20635, cluster 2 and 3 have a size of 2, and cluster 4 have a size of only 1. Standardising the dataset prevents features with larger magnitudes from dominating the clustering process, allowing each feature to be considered equally by the distance metrics.

(c) K-means Clustering (k=4) with Euclidean distance (standardised)

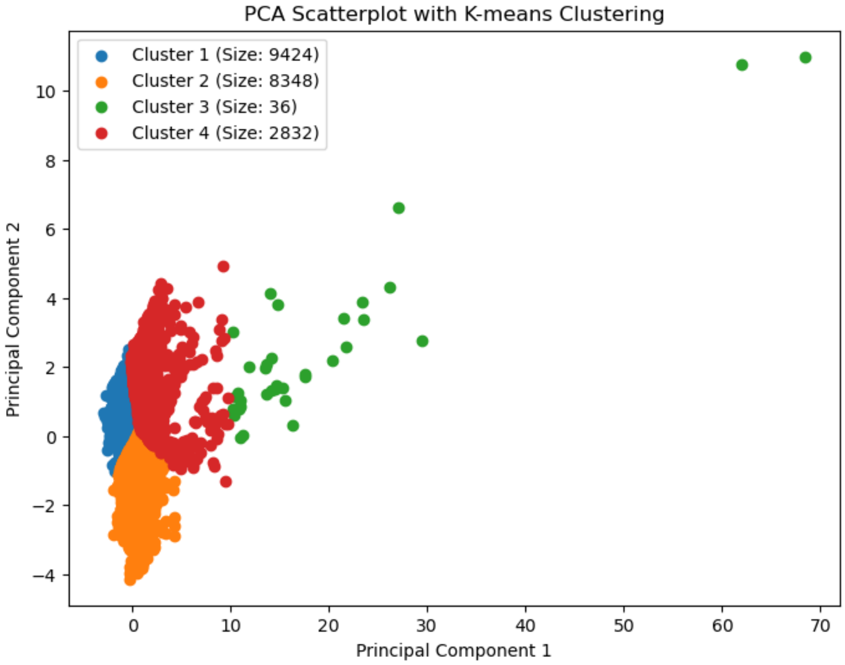
Assumption: Results did not need to be presented but can be found in the code. Both hierarchical clustering from part (b) and k-means clustering from part (c) resulted in the same clusters, same mean of the variables for each cluster, and same sizes of each cluster. Both methods provided the same clustering outcome. The only difference is that K-means clustering is faster and more efficient. However, hierarchical clustering does not require initial centroids to start the clustering process.

(d) Scatterplot of the first two principal components (Hierarchical clustering)



Assumption: Dendrogram plot was not asked to be presented but can be found in the code. The cluster sizes have changed, meaning that PCA on the standardised data affected the clustering outcome due to the dimensionality reduction. Cluster 4 (red) is the largest cluster, containing 20549 data points that are densely packed together, indicating similarity in terms of the principal components. Cluster 1 (blue) consists of 81 data points that are more spaced apart, suggesting that they are more dissimilar to each other. Cluster 3 with 8 data points also very sparsely distributed, indicating greater dissimilarity among the data points. Cluster 2 has 2 outlier data points, that are significantly far away from the other clusters. Comparing with previous group characteristics, there is still a dominant cluster with a majority of the data points, while the other three clusters have less data points.

(e) Scatterplot of the first two principal components (K-means clustering)

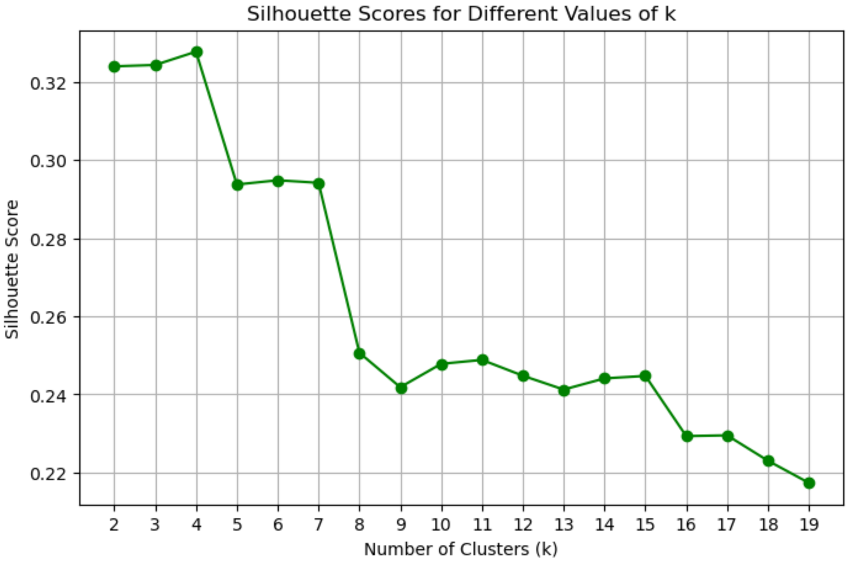


Previous clustering has unbalanced clustering sizes with one very large cluster and other smaller clusters. K-means clustering on PCA-transformed standardised data resulted in more balanced cluster sizes. There are two large clusters, cluster 1 and cluster 2 have sizes of 9424 and 8348 respectively and are densely packed together, showing that the data points within the clusters share a lot of common characteristics. Cluster 4 is moderately sized with 2832 data points and cluster 3 is a small cluster with 36 data points, representing outliers.

Hierarchical clustering builds a cluster hierarchy by merging the closest clusters, creating a dendrogram. It does not require the number of clusters to be specified but is computationally slower. K-means clustering aims to partition data into K clusters by iteratively assigning points to the nearest centroid and updating centroids. It is computationally faster but requires predefining K.

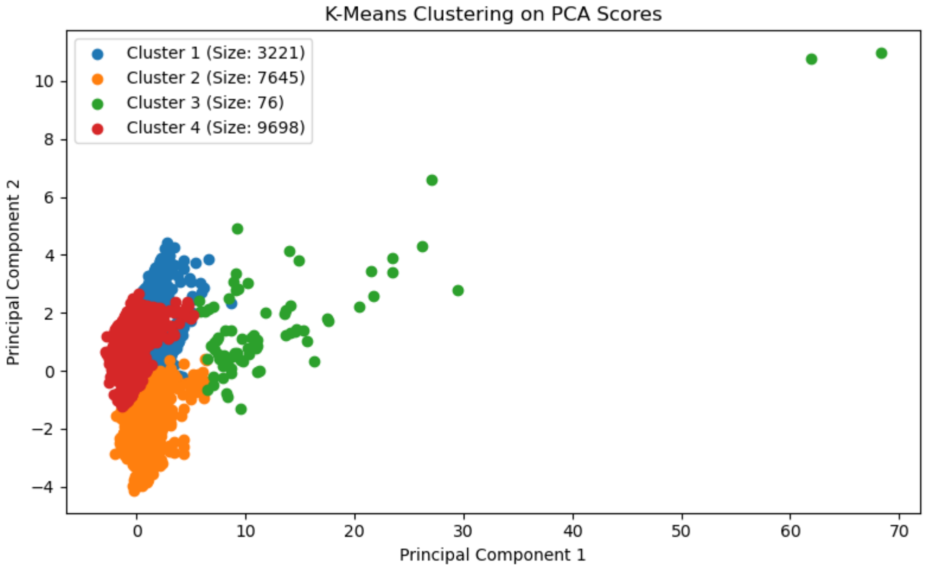
D10 Silhouette scores for k-means clustering with k=2 to 20

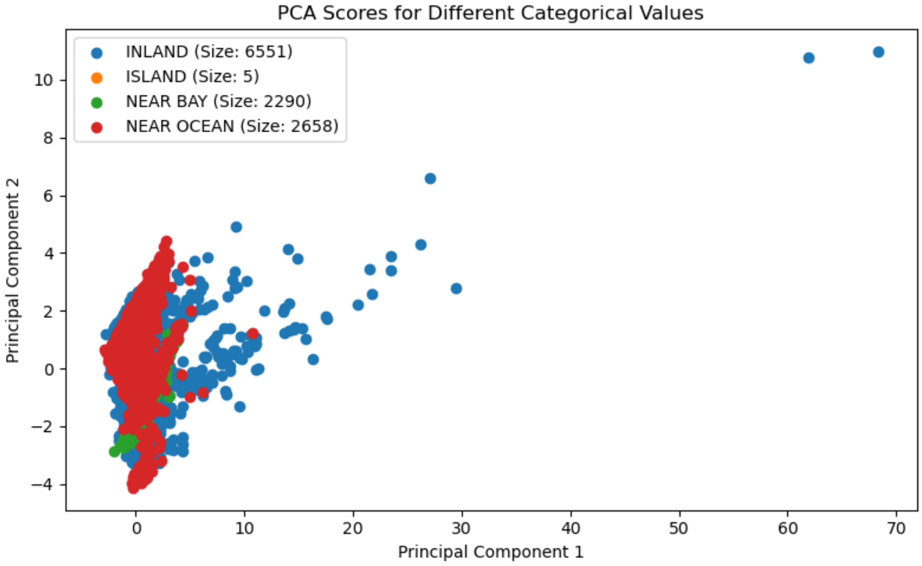
(a) Plot silhouette scores for different k values



Optimal value of k for clustering = 4

(b) Scatterplot of the first two principal component scores and the side plot





In the plot showing the k groups, you can observe more distinct clusters of data points with each colour representing a different group. These clusters are more well-separated with little overlapping. However, in the side plot, you can observe more overlapping patterns. Cluster 3 (green in first plot) is associated with the categorical variable ‘INLAND’ (blue in second plot).

(c) Conclusions about the data based on EDA and clustering analysis

From Exploratory Data Analysis, it can be observed that many features are not highly correlated, resulting in non-linear and complex relationships that can impact linear models such as Linear Regression. It might be more suitable to use a complex model such as Decision Trees or Random Forests to capture these complex relationships and improve model performance. Other than that, feature selection and data transformation can help to reduce the number of features and simplify the model, which can improve its interpretability and generalisation ability.

From Clustering Analysis, it is evident that the dataset contains outliers (potentially errors in the data). These outliers can significantly affect the model by disproportionately influencing the model's parameters and predictions, leading to decreased performance and generalisation to new data. Furthermore, categorical variables in the dataset such as *oceanProximity* is crucial in determining housing prices. Hence, it is important to encode categorical variables properly so that models can account for the impact of these variables on the target variable for their predictions.