Assignment 3

Adharsh Sundaram Soudakar (23796349)

D1: *Exploratory data analysis and preprocessing*.

(a) **Histograms:**

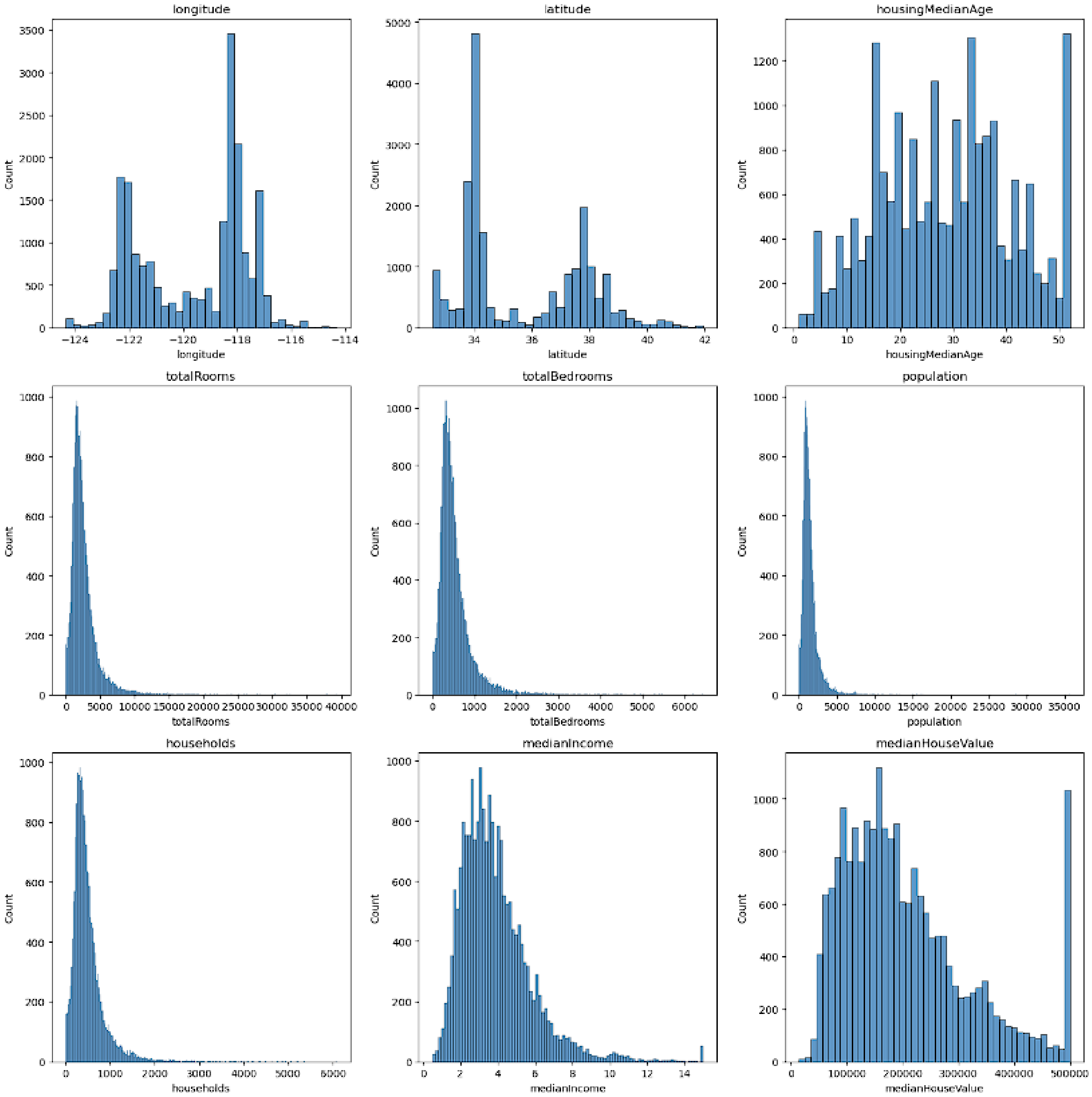


Fig. Histogram of the non-categorical features against target variable.

(b) **Highly Correlated features*:**

totalBedrooms
Population
Households
totalRooms

***against each other**

Comments:

- This because all these features are dependant on each other when it comes to anything housing related. For example, the higher the population the higher the households and more households would lead to higher numbers of rooms including bedrooms.
- To summarise this correlation reflects how housing and population metrics are inherently related.

(c) **Scatter plots:**

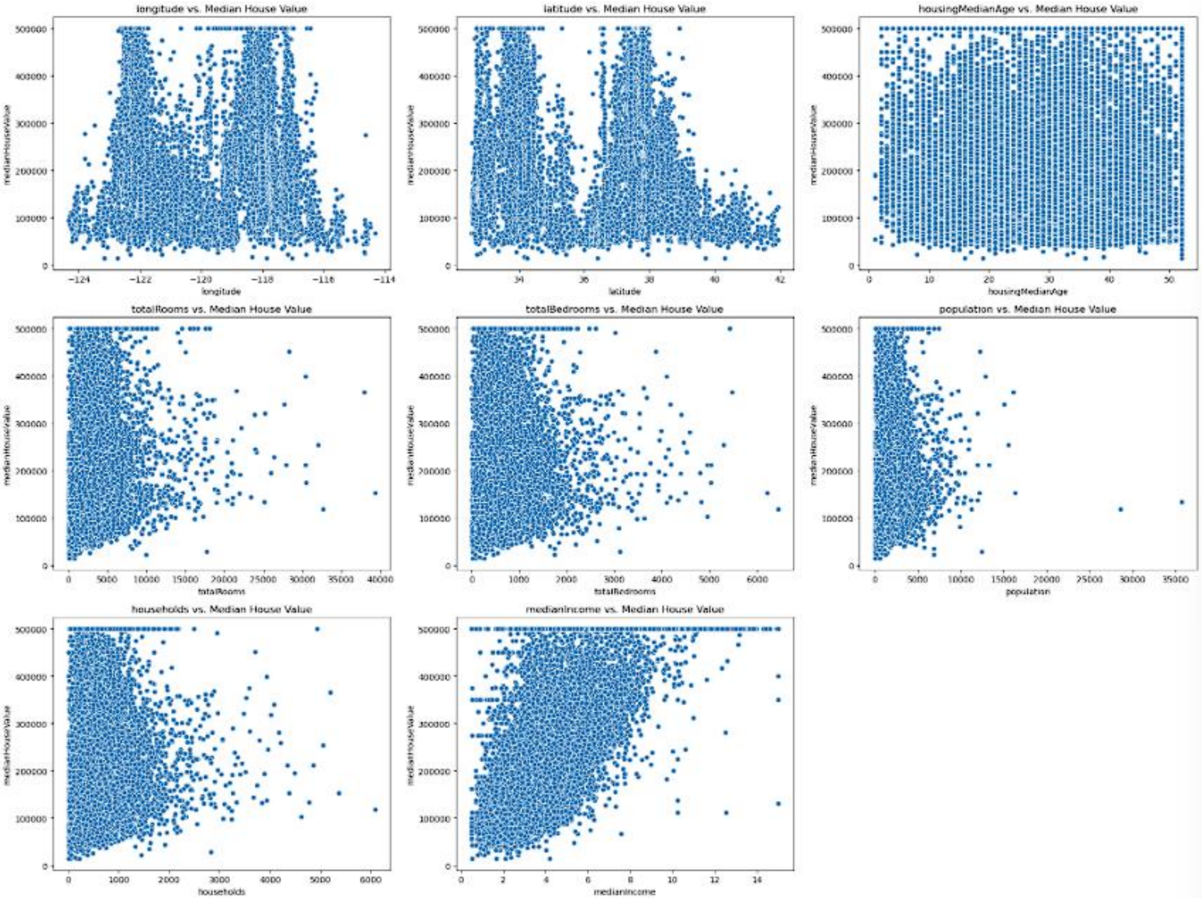


Fig. Scatter plots of each variable against the target variable.

Analysing the impact of different data transformations

D2:

(a) **Table reporting the RMSE for training and test sets for the two models for each dataset:**

Dataset/Model	LR RMSE train set	LR RMSE test set	Lasso RMSE train set	Lasso RMSE test set
Data1 original	68607.31	68589.31	68660.5	68601.81
Data1 standardised	68607.31	68589.31	68615.44	68623.38
Data2 original	0.686073	0.685893	1.129396	1.119761
Data2 standardised	0.686073	0.685893	1.156303	1.44382

(b) **Comments on results:**

- With Linear Regression for data1 and data2 (both original and standardised), there's no difference in RMSE between them. This could be because Linear regression is robust to changes in feature scaling (feature scaling didn't affect the predictive performance).
- But, with Lasso Regression, feature scaling did affect the predictive performance. RMSE on standardised data is higher than unstandardised data. Lasso might have over-penalised certain coefficients, hence the poor performance.
- But overall, the RMSE on test set is slightly lower than train set in all cases (except on standardised data1/Lasso model), indicating good generalization performance of the models.

D3:

(a) **Table reporting the RMSE for training and test sets for the two models for data3:**

Dataset/Model	LR RMSE train set	LR RMSE test set	Lasso RMSE train set	Lasso RMSE test set
Data3 original	0.70949	1.13601	1.156303	1.144382
Data3 standardised	0.70949	1.13601	1.156303	1.144382

(b) **Comments on results:**

- The results reflect the trade-off between fitting the training data well and generalizing to unseen data. With LR, lower training error but higher test error, overfitting. LassoR on the other hand, generalises better but at the cost of higher training error.
- Standardisation (feature scaling) has no effect overall; this could be because the exclusion of certain features and addition of new features resulted in a mean and a standard deviation closer to zero and one respectively or standardisation did not have any effect.

(c) **Reporting the estimated parameter values with the corresponding variable names for all models:**

NOTE: Fields with '0' as value are either '0' or were approximated to 0 as they were too small (10e-7). Fields with '-' are either fields that were removed in that dataset or don't exist.

For data1:

Features	Original		Standardised	
	LR	LassoR	LR	LassoR
Longitude	-26533.24	-26398.75	-53194.88	-50311.45
Latitude	-25444.91	-25420.76	-54426.48	-51488.49
Housing Median Age	1055.90	1059.84	13309.92	13258.91
Total Rooms	-6.428986	-6.433659	-14090.64	-12015.24
Total Bedrooms	102.9357	103.3584	43350.06	41169.56
Population	-36.35157	-36.40432	-41771.49	-41042.17
Households	45.13051	44.80739	17290.24	16763.78
Median Income	39305.21	39291.42	74889.21	74413.03
Ocean Proximity Near Bay	-791.4702	0	-247.4444	0
Ocean Proximity Island	153585.7	0	2672.207	2593.777
Ocean Proximity Near Ocean	4935.322	4206.629	1648.329	1736.254
Ocean Proximity Inland	-39134.84	-38755.03	-18231.72	-19118.76
Mean Occupation	-	-	-	-
Mean Bedrooms	-	-	-	-
Mean Rooms	-	-	-	-

For data2:

Features	Original		Standardised	
	LR	LassoR	LR	LassoR
Longitude	-0.265332	0	-0.531949	0
Latitude	-0.254449	0	-0.544265	0
Housing Median Age	0.010559	0	0.133099	0
Total Rooms	-0.000064	0.000104	-0.140906	0
Total Bedrooms	0.001029	0	0.433501	0
Population	-0.000364	-0.000118	-0.417715	0
Households	0.000451	0	0.172902	0
Median Income	0.393052	0	0.748892	0
Ocean Proximity Near Bay	-0.007915	0	-0.002474	0
Ocean Proximity Island	1.535857	0	0.026722	0
Ocean Proximity Near Ocean	0.049353	0	0.016483	0
Ocean Proximity Inland	-0.391348	0	-0.182317	0
Mean Occupation	-	-	-	-
Mean Bedrooms	-	-	-	-
Mean Rooms	-	-	-	-

For data3:

Features	Original		Standardised	
	LR	LassoR	LR	LassoR
Longitude	-0.261440	0	-0.524144	0
Latitude	-0.248051	0	-0.530580	0
Housing Median Age	0.008409	0	0.105996	0
Total Rooms	-	-	-	-
Total Bedrooms	-	-	-	-
Population	-	-	-	-
Households	-	-	-	-
Median Income	0.417373	0	0.795231	0
Ocean Proximity Near Bay	0.058689	0	0.018349	0
Ocean Proximity Island	1.526743	0	0.026564	0
Ocean Proximity Near Ocean	0.083880	0	0.028015	0
Ocean Proximity Inland	-0.381382	0	-0.177674	0
Mean Occupation	-0.040862	0	-0.087564	0
Mean Bedrooms	0.490103	0	0.239342	0
Mean Rooms	-0.080115	0	-0.201913	0

(d) **Comments on results:**

- With regards to LR, there are similarities to parameter values of the variables between data1 and data2 but in data2 they are changed in hundreds of thousands. This might be due to the manual alteration that we performed on the data. There's a slight difference with the parameter values with regards to dataset3. Again, due to manual alteration, deletion of certain variables and addition of new ones.
- With regards to LassoR, as discussed earlier, over penalisation of certain coefficients results in many values to be either too small (10e-7) or 0. This shows that LassoR is too aggressive when it comes to feature selection. For dataset2 and dataset3, all the parameter values are zero or too small, this indicates none of the features impact the target variable, but this is not right.
- Standardisation resulted in massive changes. This could be due to the difference in scales of each field in the datasets.

D4:

(a) **Tables reporting LassoR model results:**

- The optimal alpha value according to Grid-Search:

Optimal alpha	0.001

- The RMSE on the training set:

RMSE on train set	0.734956

- The RMSE on the test set:

RMSE on test set	1.128937

- The estimated parameter values with the corresponding variable names:

Features	Parameter vales
Longitude	-0.496116
Latitude	-0.501295
Housing Median Age	0.105784
Median Income	0.788216
Mean Rooms	-0.185073
Mean Bedrooms	0.222486
Mean Occupation	-0.086653
Ocean Proximity Inland	-0.187397
Ocean Proximity Island	0.025861
Ocean Proximity Near Bay	0.018311
Ocean Proximity Near Ocean	0.028343

D5:

(a) **Tables reporting RidgeR model results:**

- The optimal alpha value according to Grid-Search:

Optimal alpha	100

- The RMSE on the training set:

RMSE on train set	0.709887

- The RMSE on the test set:

RMSE on test set	1.131426

- The estimated parameter values with the corresponding variable names:

Features	Parameter vales
Longitude	-0.438579
Latitude	-0.441902
Housing Median Age	0.106569
Median Income	0.781283
Mean Rooms	-0.173267
Mean Bedrooms	0.209397
Mean Occupation	-0.086886
Ocean Proximity Inland	-0.204392
Ocean Proximity Island	0.027127
Ocean Proximity Near Bay	0.021768
Ocean Proximity Near Ocean	0.032336

(b) **Comparison of alpha and parameter values between D4 and D5:**

- In Lasso Regression, the optimal alpha value is relatively low (0.001), indicating that the model penalizes the coefficients more aggressively whereas in Ridge Regression, the optimal alpha value is higher (100), indicating that the model applies a milder regularization.
- Looking at the estimated parameter values, we can see that they are generally similar between the two models, albeit with slight differences.

D6:

(a) **Tables reporting Decision TreeR model results:**

- The optimal max depth according to Grid-Search:

Optimal max depth	9

- The RMSE on the training set:

RMSE on train set	0.502671

- The RMSE on the test set:

RMSE on test set	0.603567

D7: Comparison between models from D4, D5 and D6

(a) **Comments on RMSE on the test set:**

- Decision Tree Regression model has the lowest error on test set of all models. Although all the models have poor generalisation capacity, Decision Tree Regression model has the lowest error (high predictive accuracy) (most of the relationships are non-linear), hence it's the best model.

(b) **Comments on improving model's predictive capacity:**

- From EDA (scatter plots) we could understand that there exist non-linear relationships, these could be transformed (exponentially or logarithmically) to help models from LassoR and RidgeR predict better as these models do not capture non-linear relationships well. Also from histograms, we could see certain features are right skewed, we could normalise these which would help improve the models' performance. These are some of the ways to improve the models' performance.

D8:

(a) **Plot:**

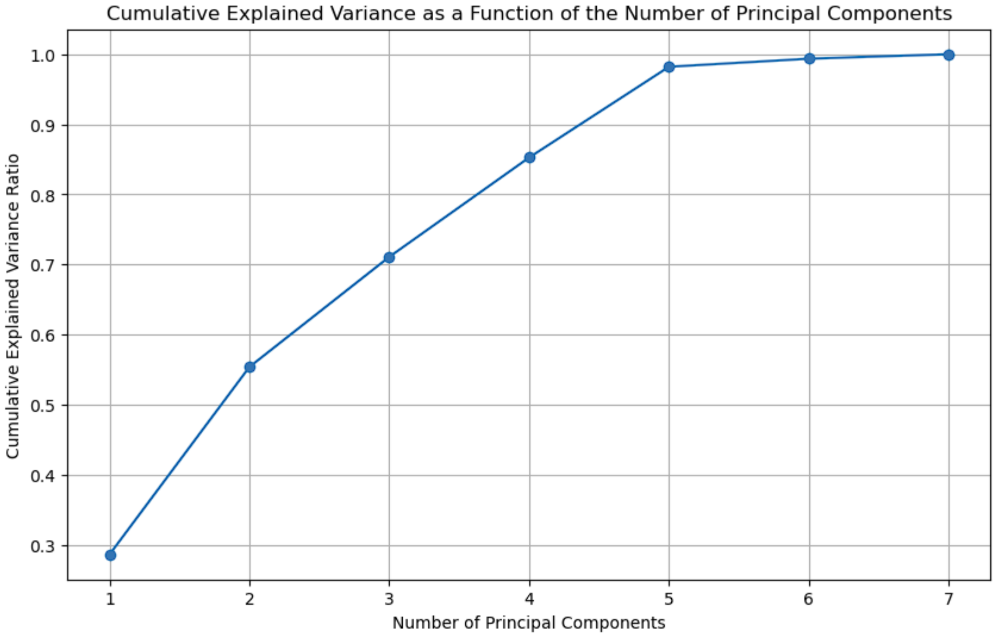


Fig. Cumulative explained variance as a function of the number of principal components.

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

Necessary principal components	5

(c) **RMSE on Train and Test Set:**

RMSE on training set	0.805871
RMSE on test set	1.339414

(d) **Report:**

Optimal number of principal components	7
RMSE on training set	0.718938
RMSE on test set	1.178386

(e) **Discussion and comparison of results:**

- Using the optimal number of principal components has resulted in a lower RMSE on both train and test sets, improving the predictive accuracy of the model.
- But the accuracy is not better than the 3 previous models. In case of LR and LassoR, this could be because of regularization which could have prevented overfitting. With Decision TreeR, it would have captured non-linear relationships better.

D9:

(a) **Hierarchical Clustering without Standardisation:**

Mean of the variable for each cluster:

Feature/Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Longitude	-120.605	-119.569411	-121.15	-121.98
Latitude	37.865	35.631367	38.69	38.32
HousingMedianAge	41.0	28.6364	52.0	45.0
MedianIncome	4.8909	3.8702	6.1359	10.2264
MeanRooms	7.1099	5.4288	8.2759	3.1667
MeanBedrooms	1.2253	1.0967	1.5172	0.8333
MeanOccupation	551.0879	2.9464	230.1724	1243.333

Summary:

Cluster 1:

- Size: 2 districts
- Characteristics:
  - Located at coordinates around -120.605000 longitude and 37.865000 latitude.
  - High housing median age (41 years).
  - Median income is moderately high (4.8909).
  - Houses have a high number of rooms (7.1099).
  - Moderate number of bedrooms (1.2253).
  - High mean occupation (551.0879), indicating potentially high density or number of occupants per housing unit.

Cluster 2:

- Size: 20636 districts
- Characteristics:
  - Located at coordinates around -119.569411 longitude and 35.631367 latitude.
  - Moderate housing median age (28.636 years).
  - Lower median income (3.8702).
  - Houses have fewer rooms (5.4288).
  - Slightly fewer bedrooms (1.0967).
  - Very low mean occupation (2.9464), indicating lower density or fewer occupants per housing unit.

Cluster 3:

- Size: 1 district
- Characteristics:
  - Located at coordinates around -121.150000 longitude and 38.690000 latitude.
  - Very high housing median age (52 years).
  - High median income (6.1359).
  - Houses have a very high number of rooms (8.2759).

        - High number of bedrooms (1.5172).
        - Moderate mean occupation (230.1724), indicating moderate density.

Cluster 4:

- Size: 1 district
- Characteristics:
    - Located at coordinates around -121.980000 longitude and 38.320000 latitude.
    - High housing median age (45 years).
    - Very high median income (10.2264).
    - Houses have the fewest rooms (3.1667).
    - Fewest number of bedrooms (0.8333).
    - Very high mean occupation (1243.3333), indicating extremely high density or number of occupants per housing unit.

(b) **Hierarchical Clustering with Standardisation:**

Mean of the variable for each cluster:

Feature/Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Longitude	-120.09	-120.605	-119.56943	-121.98
Latitude	38.355	37.865	35.631202	38.32
HousingMedianAge	33.5	41.0	28.637	45.0
MedianIncome	3.25	4.89	3.87	10.226
MeanRooms	137.221	7.109	5.416	3.166
MeanBedrooms	29.851	1.225	1.093	0.8333
MeanOccupation	2.5636	551.087	2.8574	1243.33

Size of clusters:

Cluster 1	2
Cluster 2	2
Cluster 3	20635
Cluster 4	1

Interpretation:

- Yes, the groups changed. The cluster distribution has changed, cluster 3 now has most districts meaning this is the group where features are very close to their respective means.

(c) K-means Clustering with Initial Centroids from Hierarchical Clustering (Standardised features):

Mean of the variable for each cluster:

Feature/Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Longitude	-121.98	-120.08	-119.5695	-120.1
Latitude	38.32	38.8	35.631	38.91
HousingMedianAge	45.0	34.0	28.6382	33.0
MedianIncome	10.2264	4.625	3.8704	1.875
MeanRooms	3.1666	132.5333	5.4163	141.9091
MeanBedrooms	0.8333	34.0666	1.0939	25.6363
MeanOccupation	1243.33	2.4	3.0106	2.7273

Size of clusters:

Cluster 1	1
Cluster 2	1
Cluster 3	20637
Cluster 4	1

Interpretation:

- Hierarchical clustering appears to be the better model since it provides a clear hierarchical structure and identifies distinct outliers and the clusters are interpretable and the distribution, although imbalanced, highlights the natural grouping in the data.
- K-means Clustering, starting with hierarchical centroids, produces similar results but with more pronounced outliers. This might not be as useful for understanding the natural structure of the data unless the focus is specifically on outlier detection.

(d) **Plot:**

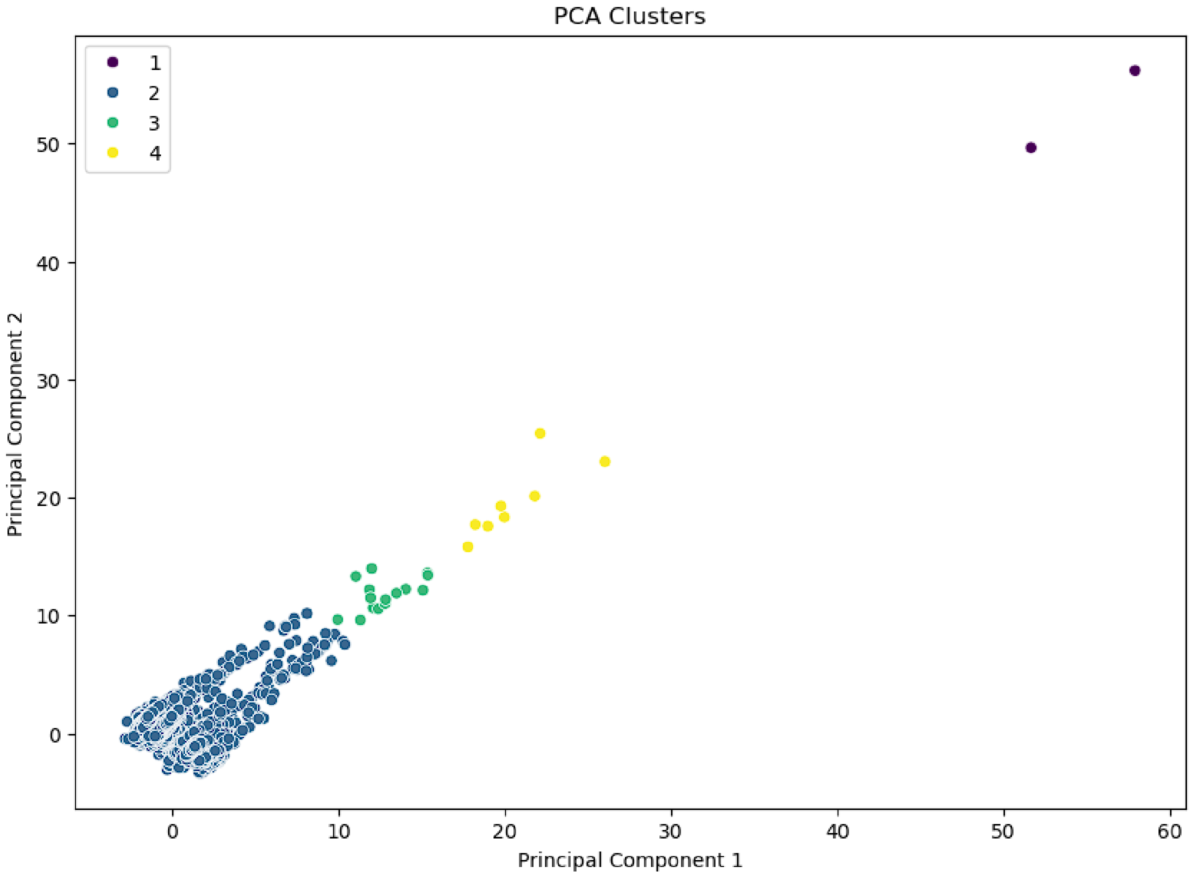


Fig. Scatterplot of the first two principal components for the instances on each cluster.

Comparison:

- The cluster distribution has changed. The size has changed as a result with lesser number of outliers. There are no more cluster with just one district.
- Cluster 1 has the least number of districts (2 to be precise).

(e) **Plot:**

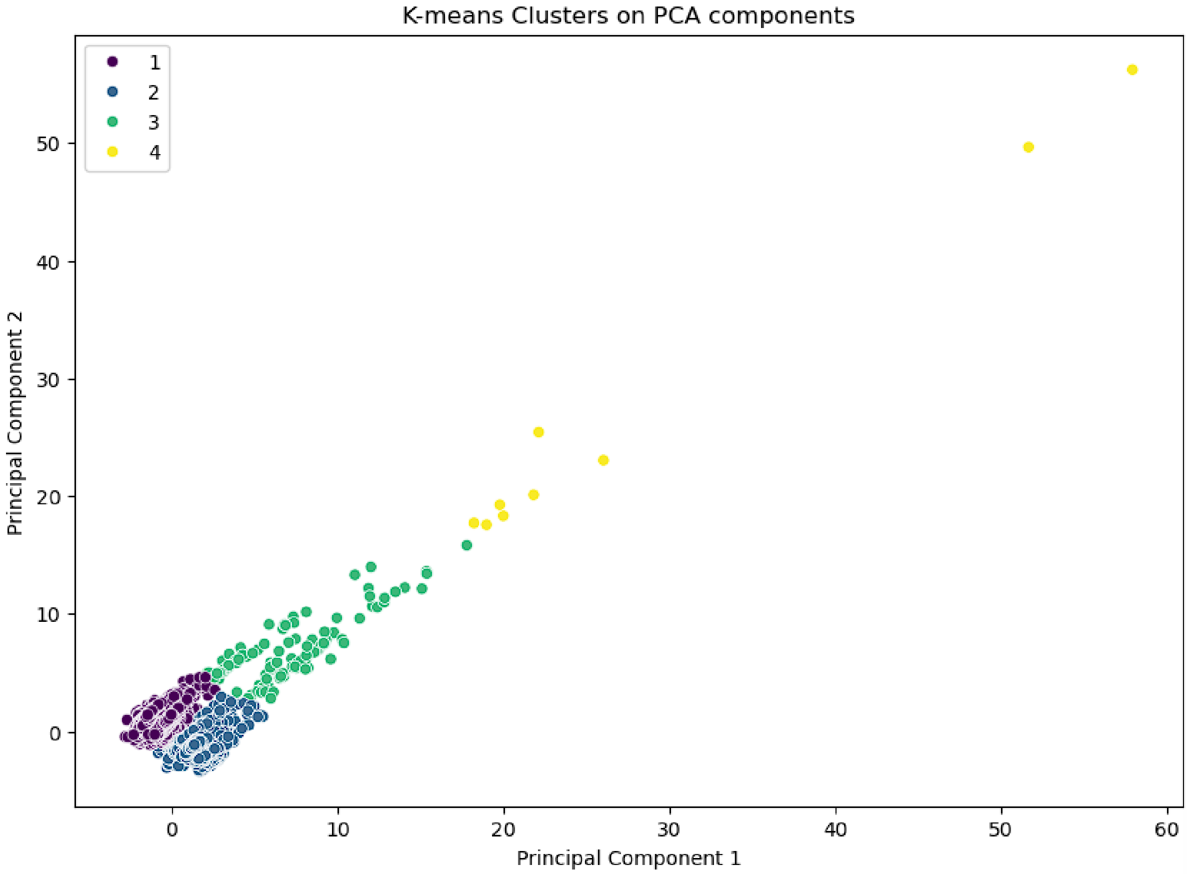


Fig. Scatterplot of the first two principal components for the instances on each cluster.

Discussion on results:

- The number of outliers is the lowest with this model.
- The cluster distribution is almost balanced (at least between cluster 1 and cluster 2). This could mean that biasing has reduced.
- Cluster 4 has the least number of districts (9 to be precise).

D10:

(a) **Plot:**

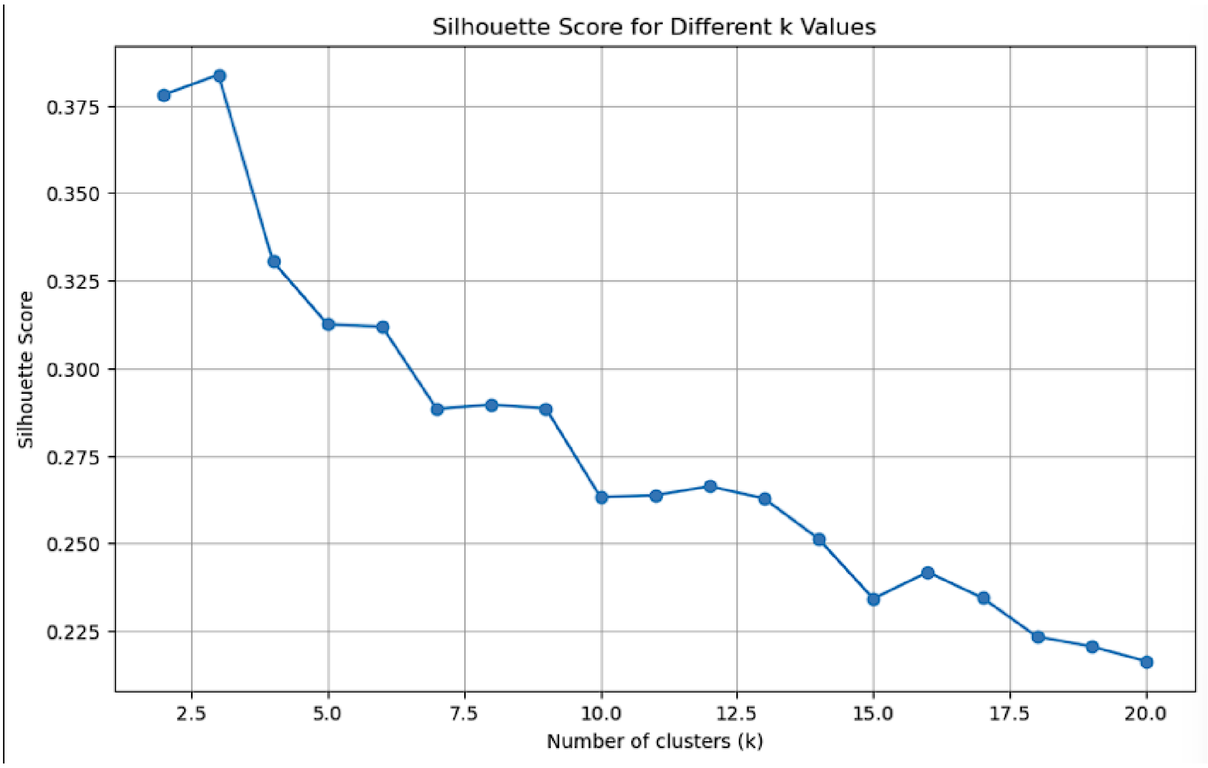
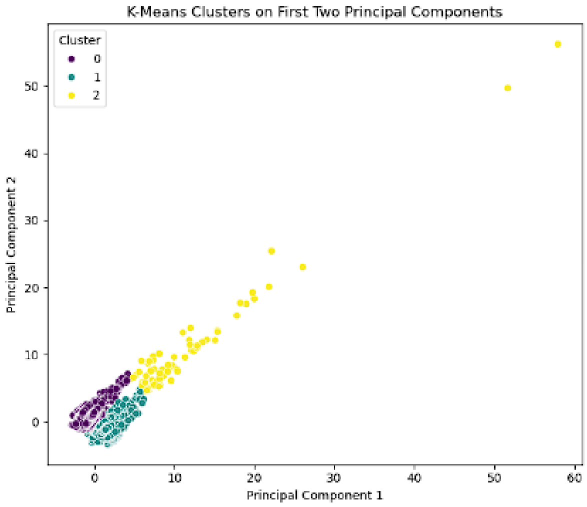
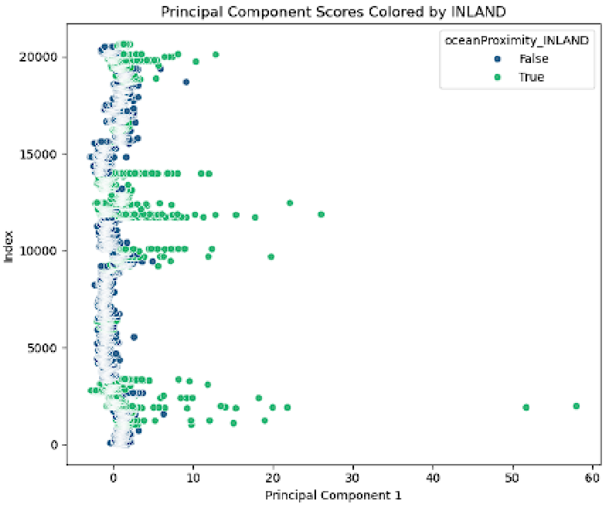


Fig. Silhouette scores for the different K values.

Optimal number of cluster(k)	3

(b) **Plots:**





Fig(left). Scatterplot of the K groups on first two principal components.

Fig(right). Scatterplot of the first principal component scores for categorical field OceanProximity(INLAND).

Relationship:

- The clusters identified by K-Means (left plot) show distinct groupings with cluster 2 having a wider spread along the first principal component.
- There is a noticeable overlap between the inland classification (right plot, green points) and the spread of cluster 2, suggesting that the inland proximity contributes significantly to the variance captured by the first principal component.
- The more compact clusters (0 and 1) might represent non-inland points, possibly closer to the ocean, reinforcing the influence of geographical features on the PCA and clustering results.

(c) **Conclusion on impact of data on Models:**

There might be several reasons as to how the data impacts model training. Some of the reasons I have identified are:

- *Skewed features*: In skewed data the tail region may act as an outlier for the models. This could be the reason why we had distinct outliers in all models.
- *Trivial features*: Certain features could have had no importance/significance with the target feature. These features during training could contribute to overfitting and hence lesser generalization capacity.
- *Non-linear relationships*: Certain models do not perform well when there exist non-linear relationships between features.