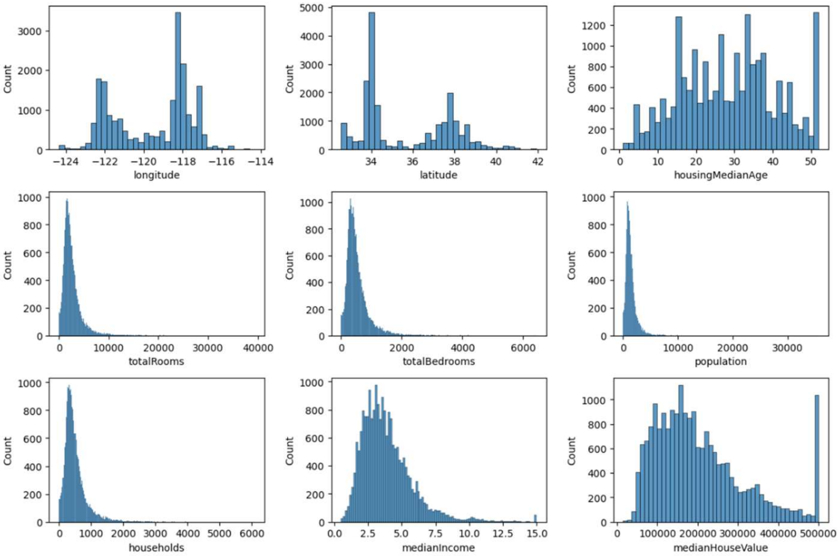
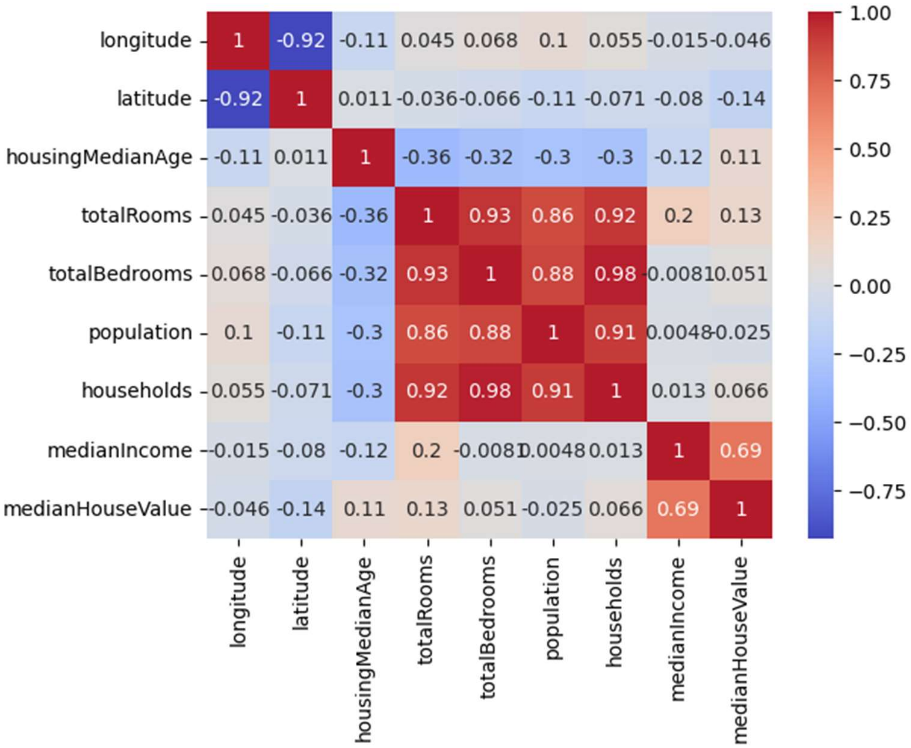
D1a


Count
longitude
−124 −122 −120 −118 −116 −114
0 1000 2000 3000
latitude
34 36 38 40 42
0 1000 2000 3000 4000 5000
housingMedianAge
0 10 20 30 40 50
0 200 400 600 800 1000 1200
totalRooms
0 10000 20000 30000 40000
totalBedrooms
0 2000 4000 6000
population
0 10000 20000 30000
0 200 400 600 800 1000
households
0 1000 2000 3000 4000 5000 6000
medianIncome
0.0 2.5 5.0 7.5 10.0 12.5 15.0
medianHouseValue
0 100000 200000 300000 400000 500000

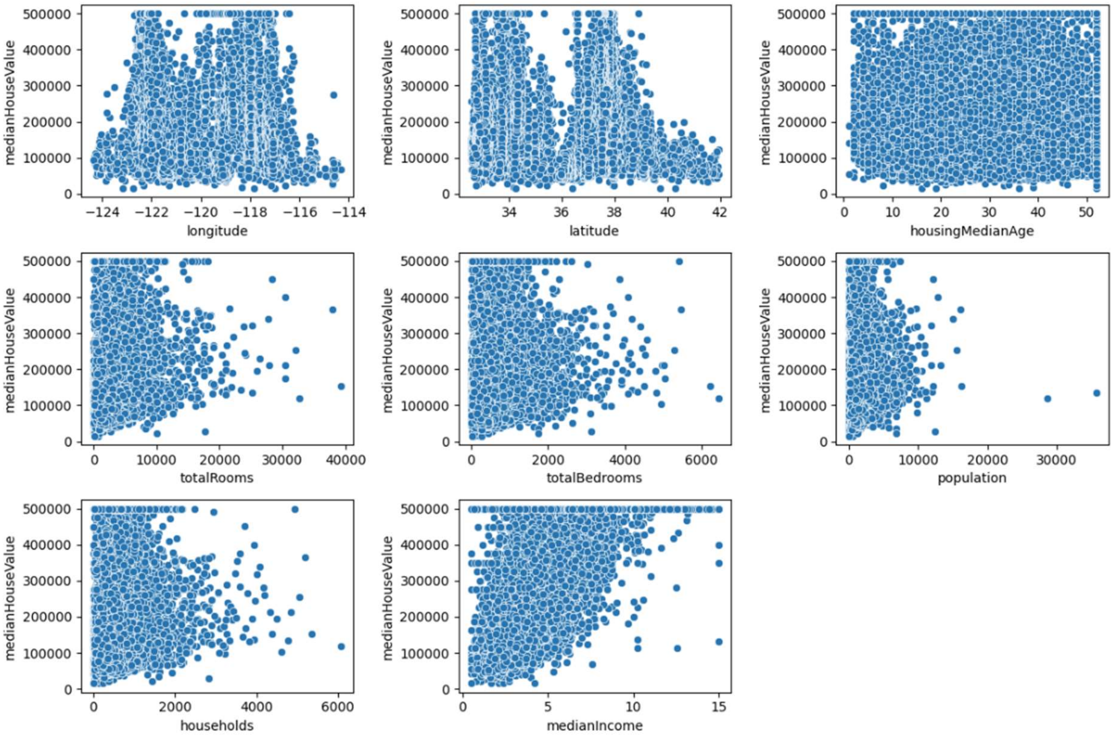
D1B


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
longitude latitude housingMedianAge totalRooms totalBedrooms population households medianIncome medianHouseValue
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

Features with correlation value more than 0.8 are – 1)totalBedrooms and totalRooms, 2)Population and totalRooms, 3)Population and totalBedrooms, 4)households and totalBedrooms, 5)household and totalRooms, 6)household and population.

For 1) It is quite obvious that more bedrooms would mean total number of rooms itself is higher. Similary for 4) and 5) More household means that each house will have atleast a room, so it is correlating. For 6) More population would mean more houses need for that population to live, and given that houses have rooms, even 2) and 3) are justified.

D1C



D2a

	Linear Original	Linear Scaled	Lasso Original	Lasso Scaled
Data1 Train RMSE	68607.314131	68607.314131	68660.504643	68615.441095
Data1 Test RMSE	68589.31234	68589.31234	68601.8095	68623.383563
Data2 Train RMSE	0.686073	0.686073	1.129396	1.156303
Data2 Test RMSE	0.685893	0.685893	1.119761	1.144382

D2b

Given that Linear Regression is invariant to the scale of features, the algorithm naturally adjusts the coefficients to compensate for any change in scale, we see that RMSE for both data set doesn’t change for original or scaled features.
Meanwhile for Lasso we see that that there is a slight change after scaling the data. This is because of the regularization term present in Lasso regression, which isn’t the case in Linear Regression.

D3A

Model/Dataset	Linear Original	Linear Scaled	Lasso Original	Lasso Scaled
Train RMSE	0.7094901591048489	0.7094901591048489	1.156302732313301	1.156302732313301
Test RMSE	1.1360103992432073	1.136010399243208	1.1443821141210113	1.1443821141210113

D3B

Linear Regression- As expected, the respective train and test RMSE doesn’t change with scaling, as it is invariant to scaling, given that the algorithm naturally adjusts the coefficients to compensate for any change in scaling.

Lasso Regression- In this case, even respective RMSE for Lasso doesn’t change, which is different from Q2. This suggests that the optimal coefficients found by Lasso under regularization are proportionally scaling with the feature scales, which leads to equivalent prediction error irrespective of scaling. This change is brought feature engineering.

D3C

Data1

Coefficient	Linear Original	Lasso Original	Linear Scaled	Lasso Scaled
longitude	-26533.24	-26398.76	-53194.89	-50311.46
latitude	-25444.91	-25420.76	-54426.49	-51488.50
housingMedianAge	1055.90	1059.84	13309.93	13258.92
totalRooms	-6.43	-6.43	-14090.65	-12015.25
totalBedrooms	102.94	103.36	43350.06	41169.57
population	-36.35	-36.40	-41771.50	-41042.17
households	45.13	44.81	17290.24	16763.78
medianIncome	39305.21	39291.42	74889.22	74413.04
INLAND	-39134.84	-38755.04	-18231.72	-19118.77
ISLAND	153585.70	0.00	2672.21	2593.78
NEAR BAY	-791.47	-0.00	-247.44	-0.00
NEAR OCEAN	4935.32	4206.63	1648.33	1736.25

Data2

Coefficient	Linear Original	Lasso Original	Linear Scaled	Lasso Scaled
longitude	-0.2653	-0.00	-0.5319	-0.00
latitude	-0.2544	-0.00	-0.5443	-0.00
housingMedianAge	0.01056	0.00	0.1331	0.00
totalRooms	-0.000064	0.000104	-0.1409	0.00
totalBedrooms	0.001029	-0.00	0.4335	0.00
population	-0.0003635	-0.0001176	-0.4177	-0.00
households	0.0004513	-0.00	0.1729	0.00
medianIncome	0.3931	0.00	0.7489	0.00
INLAND	-0.3913	-0.00	-0.1823	-0.00
ISLAND	1.536	0.00	0.02672	0.00
NEAR BAY	-0.007915	0.00	-0.002474	0.00
NEAR OCEAN	0.04935	0.00	0.01648	0.00

Data3

Coefficient	Linear Original	Lasso Original	Linear Scaled	Lasso Scaled
longitude	-0.2614	-0.00	-0.5241	-0.00
latitude	-0.2481	-0.00	-0.5306	-0.00
housingMedianAge	0.008409	0.00	0.1060	0.00
medianIncome	0.4174	0.00	0.7952	0.00
INLAND	-0.3814	-0.00	-0.1777	-0.00
ISLAND	1.527	0.00	0.02656	0.00
NEAR BAY	0.05869	0.00	0.01835	0.00
NEAR OCEAN	0.08388	0.00	0.02801	0.00
meanRooms	-0.08011	0.00	-0.2019	0.00
meanBedrooms	0.4901	-0.00	0.2393	-0.00
meanOccupation	-0.04086	-0.00	-0.08756	-0.00

D3D

Linear Regression- Coefficients are non-zero across all 3 datasets, showing that linear regression utilizes all available features for predictions.

Lasso Regression- Most of the coefficients in data 2 and data 3 are either close to 0 or actually 0. This just shows the overall reduction usage of features for predictions. In data 1 Near Bay is the only one with 0 coefficient for both scaled and original data.

The scaled versions for both linear and Lasso regressions often show different magnitudes in coefficients compared to the original data.

D4.

Optimal Alpha (Best Alpha):

0.001

RMSE on the Training Set:

0.7096

RMSE on the Test Set:

1.1289

Estimated Parameter Values with Corresponding Variable Names:

Variable	**Coefficient**
longitude	-0.4961137036010406
latitude	-0.501292825211876
housingMedianAge	0.10578435482923555
medianIncome	0.788218063149755
INLAND	-0.18739733533884162
ISLAND	0.025861656990015972
NEAR BAY	0.01831088665682883
NEAR OCEAN	0.0283435889696055
meanRooms	-0.18507638596852444
meanBedrooms	0.22248876801980447
meanOccupation	-0.08665361600750339

D5a

Best Alpha for Ridge:

100

Training RMSE for Ridge:

0.7099

Test RMSE for Ridge:

1.1314

Ridge Regression Parameter Estimates:

Variable	**Coefficient**
longitude	-0.43857994909306636
latitude	-0.4419022957627164
housingMedianAge	0.1065697010827994
medianIncome	0.7812833302024725
INLAND	-0.20439210126365845
ISLAND	0.027127850440601408
NEAR BAY	0.021768328527110456
NEAR OCEAN	0.03233665330687603
meanRooms	-0.17326736212785135
meanBedrooms	0.2093972628926058
meanOccupation	-0.08688698503018567

D5B

Ridge, has a higher α=100, because it penalizes the square of the coefficients, leading to a more distributed and uniform shrinkage across all coefficients. In contrast, Lasso, which has a much lower α=0.001, uses a linear penalty that reduces less important coefficients to zero, thus performing feature selection. This leads to Ridge requiring a stronger penalty to achieve a similar regularizing effect as Lasso because it does not zero out coefficients but rather minimizes their impact uniformly.

D6

- Best max_depth for Decision Tree according to grid search: 9
- Training RMSE for Decision Tree: 0.5027
- Test RMSE for Decision Tree: 0.6015

D7a

Model D4 – Test RMSE is 1.1289

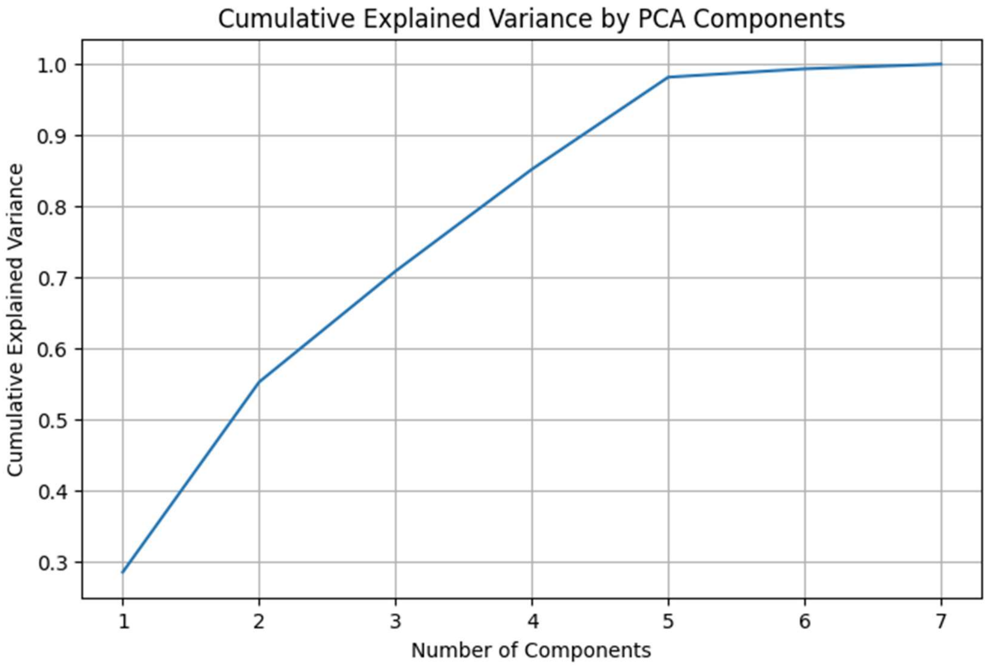
Model D5 – Test RMSE is 1.1314, worst amongst the 3

Model D6 – Test RMSE is 0.6015, is by far the best. This is because decision tree inherently handles mixture of numeric and categorical data very well. It also performs well even if the data is inherently non-linear, which seems to be the case here

D7b

Given that all of the plots, seem to be skewed, using log transformation (given that its right skewed) will be beneficial. Also removing outliers using z-score or any other technique would be beneficial

D8a



D8b

Number of components needed to explain at least 90% variance is 5

D8c

Training RMSE: 0.8065
Test RMSE: 1.3424

D8d

Optimal number of principal components: 7

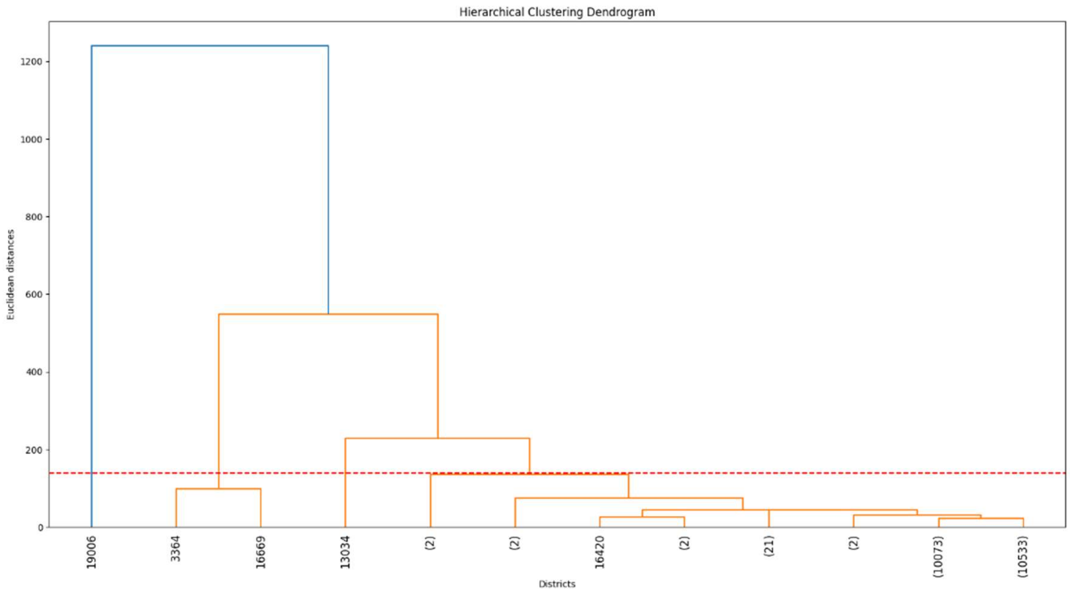
Best Test RMSE: 0.7417689765138596

Best Train RMSE: 0.7232480434778568

D8e

Model D5, still remains the best on basis of Test RMSE. Mode D8d, with a test RMSE of 0.7418, does better than most models, and is the second best after D5. Model D8c has the worst Test RMSE of 1.3424, and is massively overfitting, with training RMSE of 1.3424.

D9a



Cluster	Longitude	Latitude	Housing Median Age	Median Income	Median House Value	INLAND	ISLAND	NEAR BAY	NEAR OCEAN	Mean Rooms	Mean Bedrooms	Mean Occupation	Size
1	-120.605	37.865	41	4.8909	2.0875	0.5	0.0	0.0	0.5	7.109890	1.225275	551.087912	2
2	-119.569411	35.631367	28.636364	3.870154	2.068581	0.31731	0.000242	0.110971	0.128756	5.428809	1.096655	2.946435	20636
3	-121.150	38.690	52	6.1359	2.25	1.0	0.0	0.0	0.0	8.275862	1.517241	230.172414	1
4	-121.980	38.320	45	10.2264	1.375	1.0	0.0	0.0	0.0	3.166667	0.833333	1243.333333	1

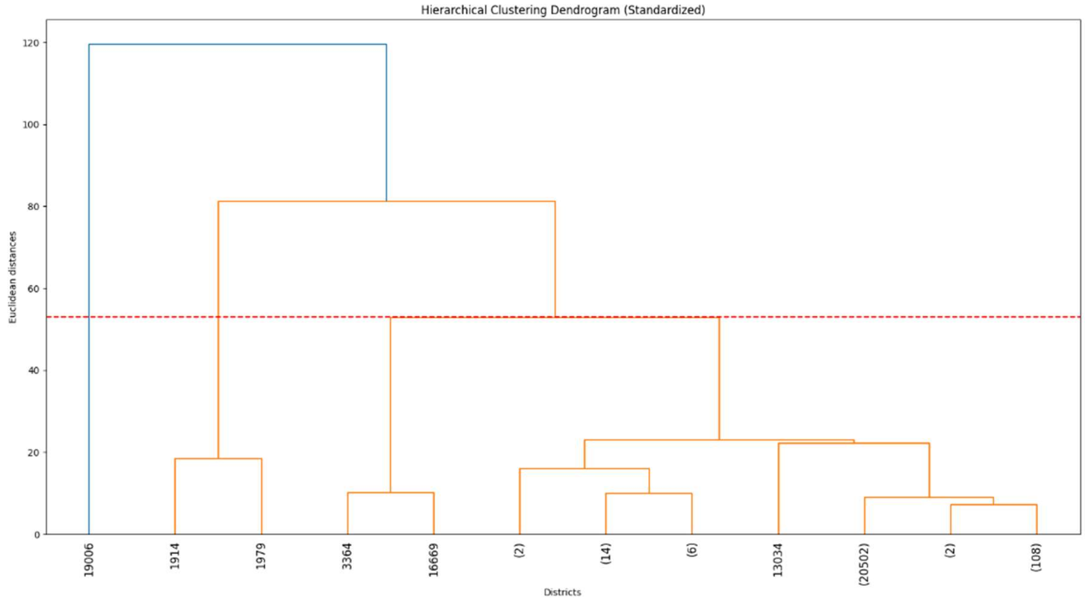
Cluster 1 has just 2 districts with median housing average age of 41 years and a high median income of 4.89, 1 inland and 1 near-ocean location.

Cluster 2 is the largest group with 20,636 districts, features younger housing with a median age of 28.6 years and a lower median income of 3.87. This captures almost all districts in data, hence can be very representative of overall data.

Cluster 3 includes just one district with the housing age of 52 years, the highest median income (6.14), and is in inland.

Cluster 4 also consists of a single district but with relatively new housing (45 years), median income is 10.23, extremely high occupation density (about 1243 people per household), suggesting that it could be an anomaly/outlier hence detected as a different cluster.

9b



```
Original Cluster Sizes
1       2
2   20636
3       1
4       1
Scaled Cluster Sizes
1       2
2       2
3   20635
4       1
```

Due to scaling, groups changed, and also the total distance of each groups, according to dendogram, changed.

To add to this, cluster 2, which had 20636 districts before scaling, now has 2. And cluster 3 which had 1 district, now has 20635.

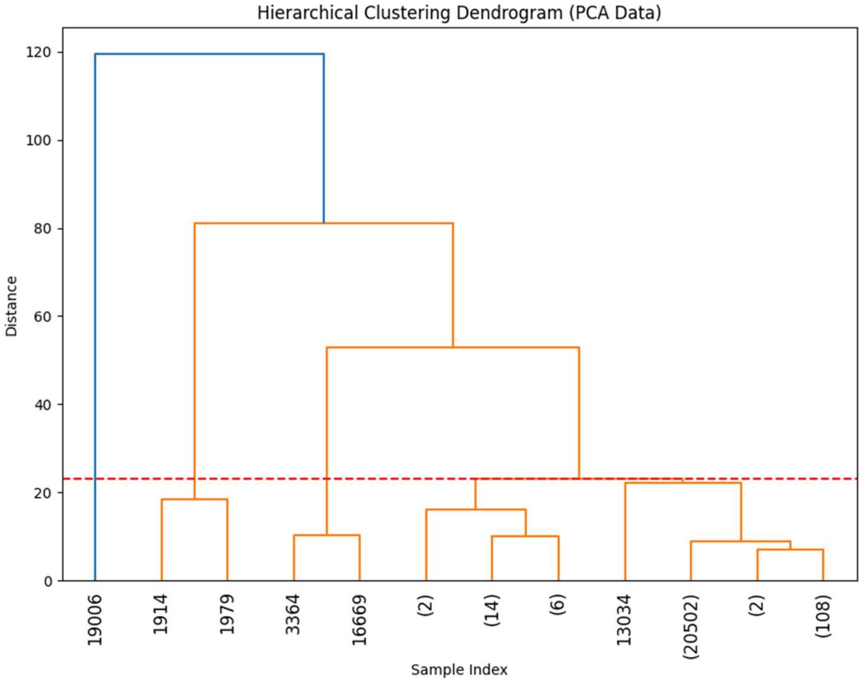
Scaling did not result in a major change for Hierarchical Clustering.

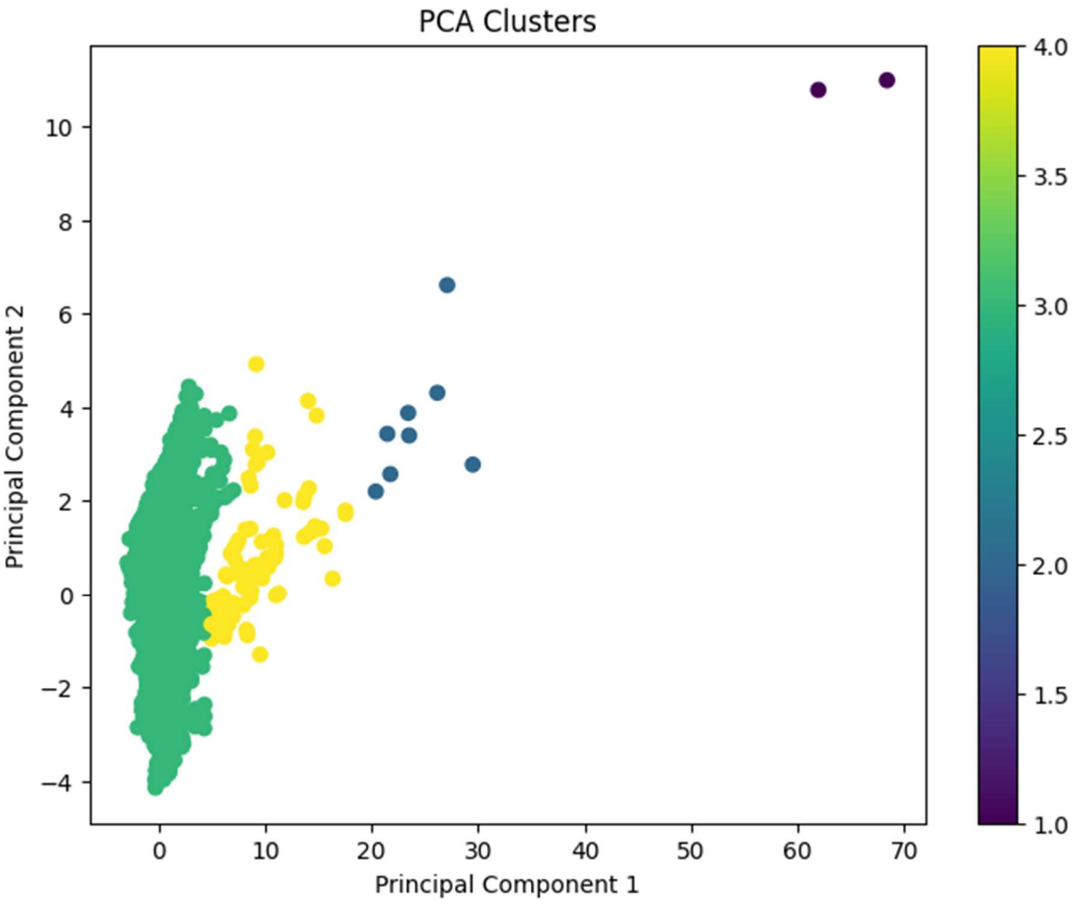
D9c

```
K-Means Cluster Sizes
1       2
2       2
3   20635
4       1
Hierarchical Cluster Sizes (Scaled)
1       2
2       2
3   20635
4       1
```

As we can see, K-Means and Hierarchical Clustering, both result in exactly same clusters. This is due to the fact that initial centroids have been taken on basis of already done Hierarchical Clustering. Both of them provide same results, but I would choose K-Means as its result could have been different without initial centroid.

D9d



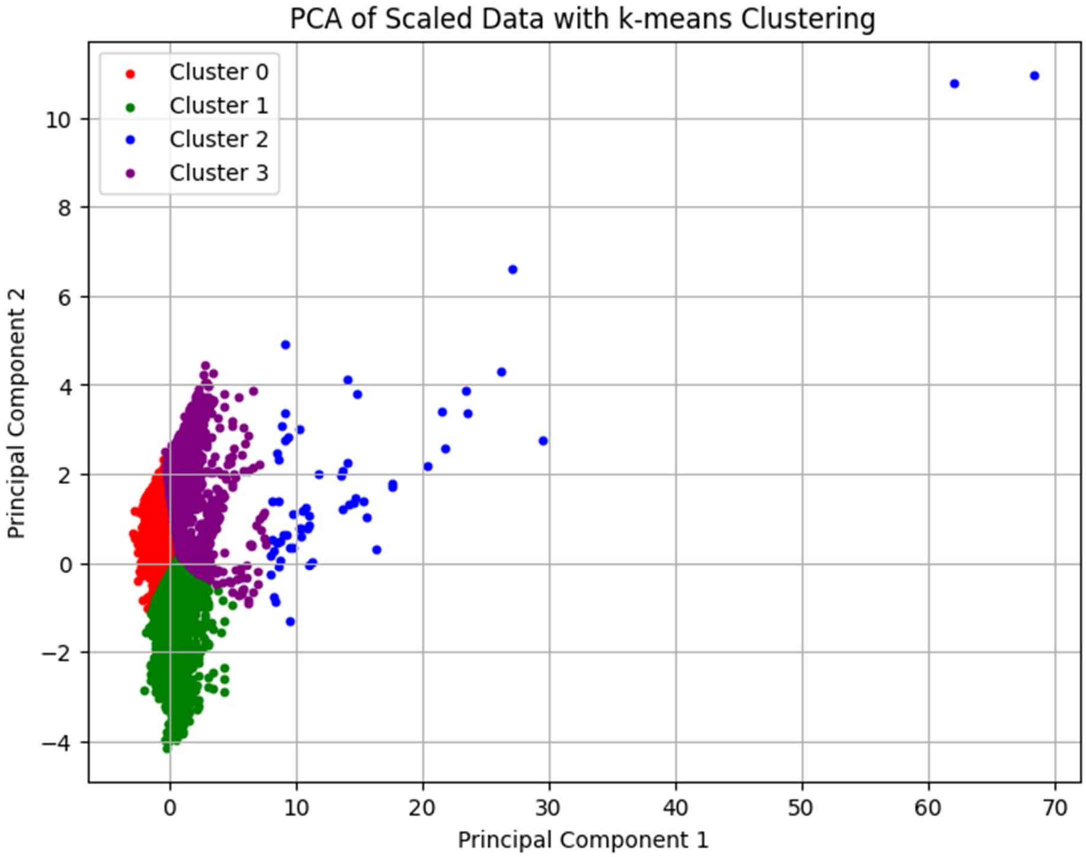


Clusters

```
3   20549
4      81
2       8
1       2
```

We see some improvement in terms of cluster size, as compared to previous tasks. Now we have a cluster with 81, and another cluster with 8 districts. Yet, we continue to see a pattern of 1 cluster having majority of districts. This could be partly because of outliers, and algorithms just mainly dividing between normal data and outliers

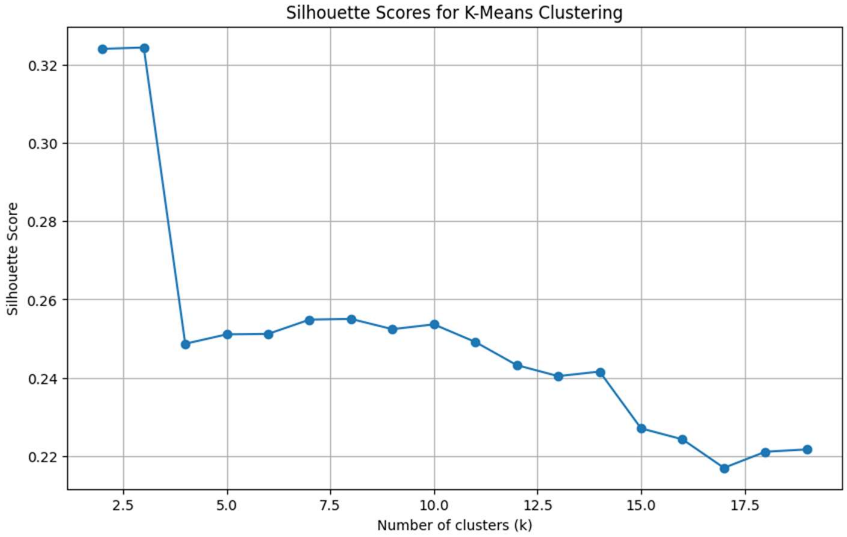
D9 e



```
1    9244
2    8373
4    2962
3      61
```

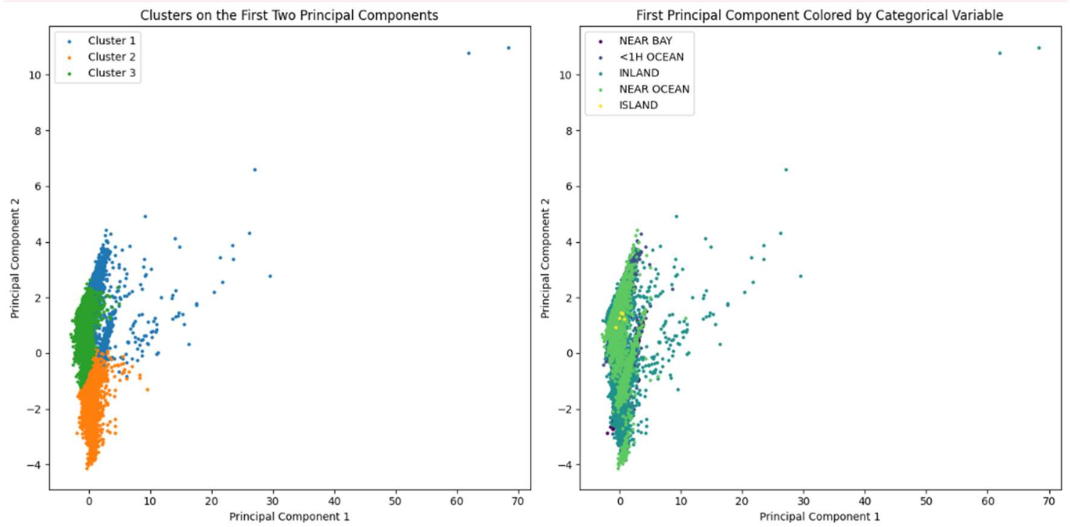
We see a major improvement in terms of cluster sizes, with Cluster 1 having 9244, Cluster 2 having 8373, Cluster 3 having 61, Cluster 4 having 2962 districts. From the plot, we can see that using PCA and K-means, we were able to divide the 'major blob' of data in 3 parts, while in all other methods we did earlier, that entire area of districts were considered the same. This could be partly because of outliers, and algorithms just mainly dividing between normal data and outliers. Outliers also tend to be further away from each other itself, hence resulting in algorithms categorizing them as different than each other.

D10 a



According to the plot, K=3 is the best value to pick, it has the highest Silhoutte Score.

10 b



As we can see, out categorical variable, ‘oceanProximity’ is completely aligns with both Principal Component, it is a very dominant feature in the dataset. This could also means that first 2 principal component represents geographical information, so any further analysis in this would help us understand geographical distances.

10 C

There is high correlation between many features and data seems to have lots of anomalies and outliers. Apart from that, the data itself is skewed, which is expected in such type of data which deals with income, housing etc. Using log transformation would be very beneficial in this case. Given that we know now that 'oceanProximity' is one the most influential variable, more feature engineering with respect to that, and in general analysing its importance can be beneficial