Harkeerat Sidhu

California Housing Prices

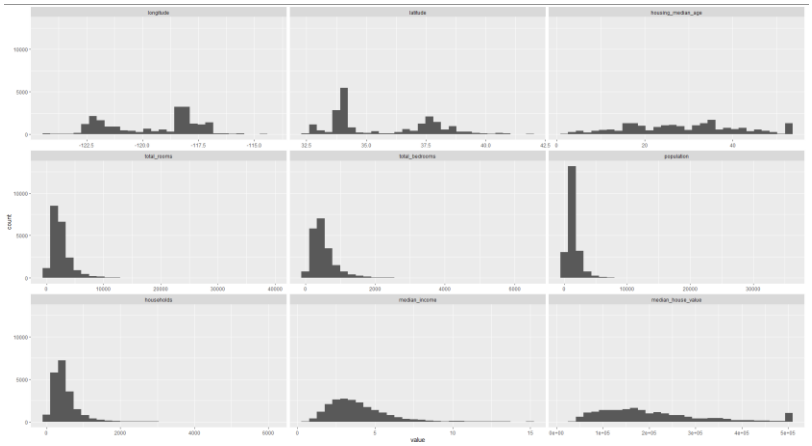
Introduction

The dataset that I used is the California Housing Prices dataset which contains the median house prices for California districts based on the 1990 Census. While this dataset is older and in most cases using an outdated dataset to do data science or analysis can unnecessarily increase your error because these old data sets don't have data that is relevant to today's problems. For example, in my data, the median housing values were capped at $500,001, however, if you were to look at California today the average house costs at least one million dollars. So this data isn't relevant for today, however for my needs, it does the job.
It does contain a good starting point for predicting housing prices and classifying the location of a house based on other factors in the data set. The dataset does contain a good starting point for predicting housing prices and classifying the location of a house based on other factors in the data set.

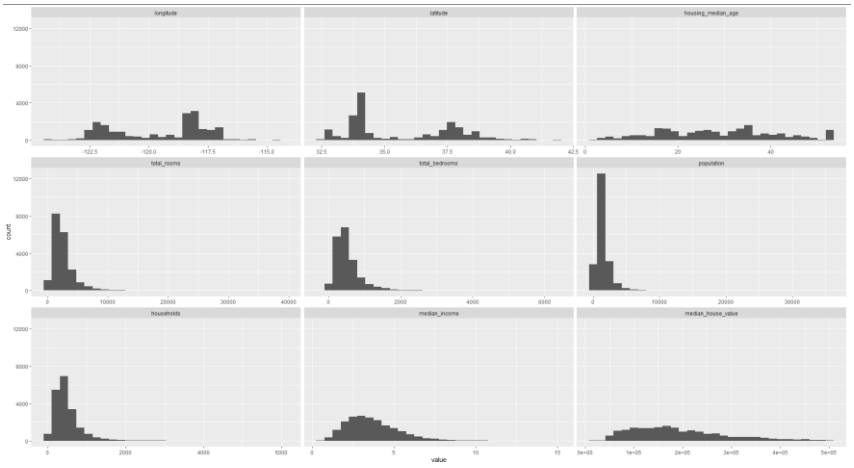
Data Visualization and Data Cleaning

Taking a cursory glance at the dataset, I see that there are ten columns. These columns are longitude, latitude, housing median age, the total rooms, the total bedrooms, population, households, median income, median housing value, and ocean proximity. The observations for everything excluding ocean proximity, latitude, and longitude are the mean values per city block. I noticed that there is lots of variation in these columns of the dataset. I first looked at a summary and a visualization of the dataset to watch for any discrepancies that could affect my analysis.

In the figure below, I saw that there were many entries in the tail end of the histogram for the housing median value and the housing median income. When I looked at those data points, I discovered that the housing values of the dataset were capped off at 500,001. I believe that this meant that values of 500,001 indicated housing values greater than that amount. Looking at the histograms of entries with values of 500,001, I saw that there was a big deal of variation in this outlier data. The range is varied and large. It is almost identical to the range of the original dataset. It is possible that the high median house values include apartment buildings as well as expensive homes. While these entries accounted for approximately 5% of my dataset, I will commit to keeping those entries removed from the final cleaned dataset as that value accounts for housing values that could be more than 500,001.



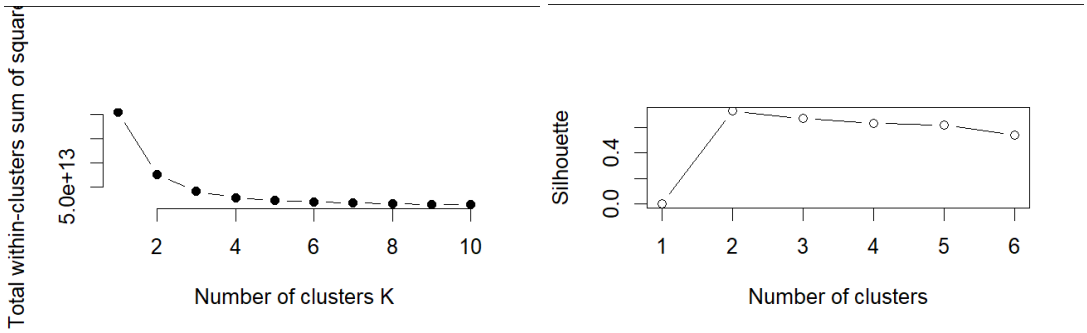
Another discrepancy that I noticed was that there were many missing values in the data frame column that contained the total number of bedrooms. I wanted to clean the dataset by imputing values into the missing values. The first approach was to use mean imputation. However, I did not feel confident in this method as it can disturb the distribution. I considered logistic regression as a way to impute those values. However, the class slides recommend that this should be avoided whenever possible as it can increase correlations and underestimate the variance. With that being said, I settled on stochastic regression. It preserves the distribution of the observed response variables and preserves the correlation between the response variable and the other variables. Using stochastic regression, I created 3 data frames that I then combined into one final cleaned data frame. A histogram of the final cleaned data frame is shown below.



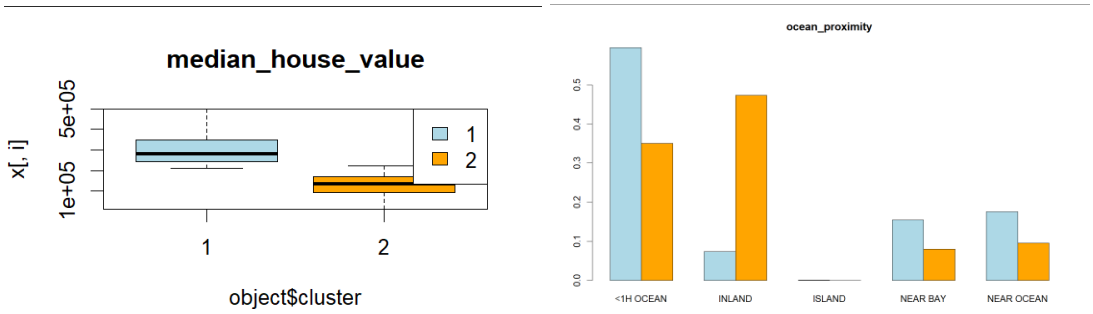
Classification and Cluster Analysis

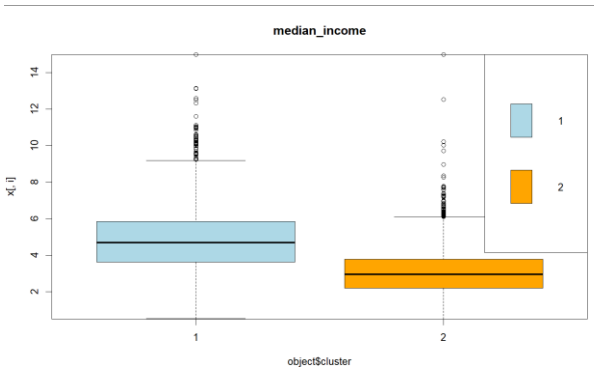
I wanted to originally see if it is possible for me to use the data set to differentiate between different groups. Since I am using a mixed type dataset, k-prototype was the desired algorithm for me to use. Before I could do the cluster analysis, I needed to pick an

appropriate k. By plotting the within-clusters sum of the squares for multiple clusters, I used the elbow rule to decide that the optimal number of clusters would be at three. To double check, I also compared the silhouette widths to ensure that the maximum width was at k=2. The figures below are the results of this analysis. I could have gone with more clusters, but I did not want to risk overfitting this model.

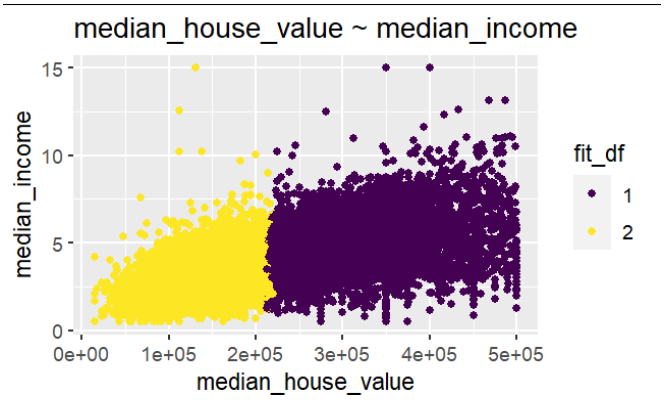


By using the cwe profiles function in R, I was able to interpret the results to see which types of variables belong to which cluster. From my output, I am able to see which clusters belong to which levels of the variables. Two variables that I was interested in were the ocean proximity, median housing value, and median income. Below are figures of the breakdown of the clustering of these two variables. It is important to note that while there are plots for ten variables, I am only including these two for the sake of brevity.





It seems that the first cluster mostly contains points that are closer to a body of water. The median house value for the first cluster seems to be high. Additionally, the median income of the first cluster is also high. The second cluster contains more data points of the houses that are more than an hour inland. The second cluster has lower values for income and house values compared to the first cluster. An additional relationship that I wanted to look at is how plotting the median house value against the median income would look like when I label the points according to the points' cluster. I can see that there is a relationship between income and housing value. This goes hand in hand with the clustering results that I have gotten.



Multinomial Model

I wanted to fit a multinomial model to the data set to predict the ocean proximity variable based on the other variables in the dataset. I used a multinomial model because the ocean proximity variable has five response variables. One important step in predicting the ocean proximity variable was to remove the longitude and latitude variables from the dataset. These would not be very helpful for predicting the ocean proximity when I am given specific coordinates. The final multinomial model that I used was able to correctly predict ocean proximity 80% of the time. Below is a table that shows how each variable was classified.

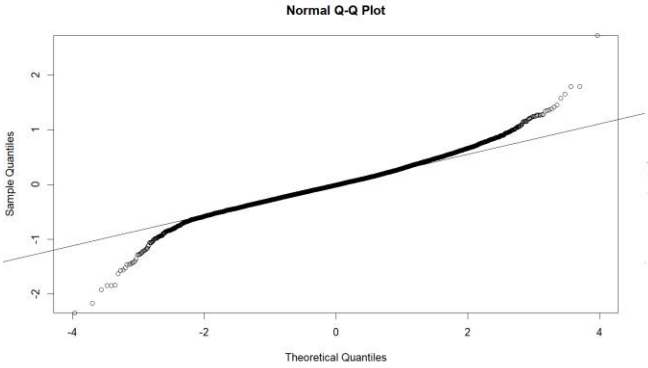
	testL				
LogRes	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
<1H OCEAN	2260	120	0	122	468
INLAND	33	1853	0	5	30
ISLAND	0	0	0	0	0
NEAR BAY	144	21	0	494	102
NEAR OCEAN	104	2	1	0	144

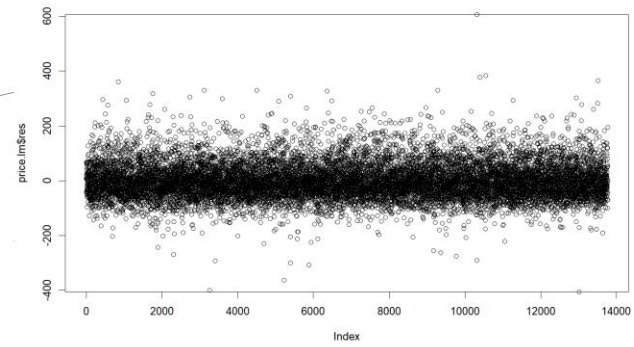
I can see that the houses that are inland are the most accurately predicted at a rate of 96%. It was also able to correctly predict houses that are less than one hour from the ocean 76% of the time. The model was not able to correctly predict the houses on islands. However, there are only five data points for this variable so I will not worry about this too much. The model did have

the most difficulty distinguishing between houses near the bay and near the ocean. However, one could make the argument that they are essentially the same thing.

Linear Model

I wanted to model the relationship between median housing values against all the other variables in the dataset. I created a linear regression model to look at this relationship. Using a log transform of the response variable, I was able to increase the adjusted R-squared from .6148 to .6556. Checking the summary of the model showed me that all the variables are significant. As such, I decided to not drop any of the variables. However, looking at the normal qq-plot, I can see that the normality assumption is being violated. Additionally, in both models, the variance assumptions have also been violated. I decided that the linear model was not the best model to predict housing prices. Therefore, I will not consider this the most useful model. Below is evidence of these assumptions being violated after the variable transformation.





Random Forest Model

Rather than use a linear regression model, I decided to use a random forest model. Iset the response variable as the median house value and the rest of the data as the predictors. To train my model, I used eighty percent of the dataset and set the remaining twenty percent as to test the model. Iran the algorithm by creating 750 trees. The variable that I was most interested in was the RMSE. I created a vector of predicted values using the test data. Using this vector, I then used it to create my RMSE. I found that the random forest model was able to predict the median house value within approximately $42,000. I found that this was the better of the two models. One important note about this observation is that the data is from a time when housing values were much lower. While this error is small by today's standards, it is much larger by the standards of 1990.

Conclusion

After some data cleaning and imputation, I was able to conduct a k-prototype analysis on the dataset. I found out that partitioning the two groups tended to separate the inland homes from the more coastal homes. The coastal homes had a higher housing value than the inland homes. I also attempted to create a multinomial regression model to classify the ocean

proximity. While the inland homes and homes that are within one hour of the ocean were accurately classified, the model struggled with the island homes, bay homes, and ocean homes. The best model that I had was the random forest model. It was able to predict a home price within $44985.20. While that was quite a bit when compared to the home prices of 1990, it was one of my better models. Overall, I see that the trends in California have not changed since the 1990 Census.

Bibliography

https://rpubs.com/areyhan02/c3segm

https://stackoverflow.com/questions/66931997/silhouette-value-of-each-cluster-using-clustmixtype

https://stats.stackexchange.com/questions/293877/optimal-number-of-clusters-using-k-prototypes-method-in-r