Module Code: 7CCMMS61
Module Name: Statistics for Data Analysis(23~24 SEM1 000001)
Assignment: Coursework
KCL Number: 23025410
Name: Henry Cao

1. Structure of report

1.1 Table of Contents

1.2 Dataset

The dataset relevant to the report contains information on a number of residential block housing in the State of California, located on the West Coast of the United States of America (USA). The data was collected in 2001. In total there are 20,640 residential blocks. There are 10 different variables that are collected about each residential block. It seems to have been collected in order to study various trends and patterns in Californian housing, with a key focus on median house value.

The “longitude” variable provides the longitudinal coordinate of a residential block in California. Similarly, the “latitude” variable provides the latitude coordinate of a residential block in California. The variable “house_median_age” provides the median age of a house in the residential block. The variable “total_rooms” provides the count of the total number of rooms, not counting bedrooms, in all houses in the residential block. The variable “total_bedrooms” gives the count of the total number of bedrooms in all houses in the residential block.

The variable “population” returns the total number of people living in the residential block. The “households” variable returns the total number of households in the residential block. The “median_income” variable provides the median of the total household income of the houses in the

residential block. The scale for median income ranges from 0.4999 to 15.0001, so it's possible that this variable could measure median income in the tens of thousands of dollars. For interpretability, it'll be assumed that the units of median income is in the tens of thousands of USD. The variable "ocean_proximity" describes the geographic location of the block, taking on different values; "NEAR BAY" means close to the San Francisco Bay; "<1 OCEAN" means the residential block is within 1 hour travel distance of the Pacific Ocean; "INLAND" means that the residential block is greater than 1 hour away from the Pacific Ocean; "NEAR OCEAN" means that the residential block is quite close to the ocean, being even closer than the "<1H OCEAN" designation; the "ISLAND" designation means the residential block is located on an island. The last variable of note is "median_house_value", which provides the median of the household prices of all the houses in the block.

1.3 Objective

The main object of this report is to build a model that can predict the median house price of a residential housing block in California using a selection of predictors. Exploratory Data Analysis will be used to perform an initial inspection of the data and draw insight as to which model might be the most suitable for predicting median house price given the structure and patterns of the data. Extreme values will be removed to reduce the effects of outlier points with significant leverage. Partitioning of the data into subsets will be used if it leads to better results. Most of the modelling will be performed with a multilinear model without any transformation of variable or interaction terms. Various metrics and test will be used to evaluate how consistent the multilinear model is with the linear model assumptions. A model utilising the Gradient Boosting Machine method will be briefly introduced as a point of comparison. The report will analyse the findings of the multilinear and Gradient Boost Machine (GBM) models, ending with a conclusion discussing the scope of the analysis and results.

****

2. Methodology

2.1 Inspecting the data

Summary statistics will help ascertain the shape and centrality of the variables. Some of key visualisation will include, but aren't limited to: pairwise plots of quantitative variable, boxplots of quantitative variables against qualitative variables, histograms of quantitative variables, bar charts of qualitative variables. ANOVA tests will be administered on all categorical variables to see if there are significant variances in counts for specific quantitative variables. Furthermore, latitude and longitude will be shown in a pairwise plot, with the points coloured by ocean_proximity to show the geographic distribution of ocean_proximity. The normalised covariance matrix will be provided to find the level of covariance between each quantitative variable.

2.2 Reshaping the data for analysis

As the dataset may contain observations inconsistent with established patterns, the dataset will need to be reshaped. Significant values, such as outliers and leverage points, will be identified and removed accordingly. Temporary or new variables may be introduced in order to break the dataset into subsets, or provide an dimension to existing variables. The goal is to preserve the integrity of the dataset while making it more conducive for analysis.

2.3 Building the predictive models

As the chief goal of the report is to predict the median house price of residential block housing in California, United States of America, in the year 2001, two statistical models will be developed to make such predictions.

2.3.1 Multilinear Regression Model

The multilinear regression model will first involve clustering the entire dataset into subsets. This process will use k-Mean clustering and will cluster by latitude and longitude, effectively splitting the dataset into multiple regions. Each cluster will be analysed using a separate multilinear model of the following format:

$$\text{median\_house\_value}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i \quad (2.3.1.1)$$

The various β coefficients are the true population parameters for the chosen parameters to estimate the median house value. ϵ is the true random variation that isn't accounted for by the multilinear regression equation. The actual estimated responses will use the formula below:

$$\widehat{\text{median\_house\_value}} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \ldots + \widehat{\beta}_p X_p \quad (2.3.1.2)$$

The residuals are the difference between the predicted response and the actual response for median house value: $\widehat{\epsilon} = \text{median\_house\_value} - \widehat{\text{median\_house\_value}}$. The estimated beta coefficients will be derived using ordinary least squares estimation.

In order to evaluate the collective performance of separate clusters, a weighted average will be used. The R^2_{adj} of each cluster will be weighted by the number of observations in each cluster, resulting in the formula below:

$$Weighted \quad R^2_{adj} = \sum_i \frac{R^2_{adj\,i}}{n_i} \quad (2.3.1.3)$$

Where n_i is number of observations in cluster i and $R^2_{adj\,i}$ is the adjusted R^2 for cluster i. A for loop will be used to iterate through different numbers of clusters to maximise the weighted R^2_{adj}, while making sure that the minimum R^2_{adj} is checked and isn't too low of a value. Diagnostic checks will be run on each models. Checks will be done by cluster and will include: summary statistics, number of rows, variance of errors, standard deviation of errors, R^2_{adj}, skewness, kurtosis, Durbin Watson Test, qq-plot of residuals, residual plots, square root residual plot, and histogram of residuals. The final model selected will be the one that performs best given all of these criteria.

2.3.2 Gradient Boost Machine (GBM) Model

The Gradient Boosting Machines (GBM) method uses ensemble learning to generate many decision trees and combine the into a final model. The method is split randomly by a ratio of 80-20 into training and testing datasets to reduce overfitting and introduce validation into the process. Hyperparameter optimisation, mainly through grid search, will be used to test out different levels of tree depth, which determines how far down each decision tree can go, and shrinkage, which determines how much each individual tree contributes to the model, while holding number of trees and minimum number of observation in a node for splitting as constant. Although a single decision tree or regression equation can't be derived, various metrics, such as R^2, residual plots, and the Durbin-Watson Test can be used to evaluate accuracy and fit of the model. Graphing the importance of each predictor can show a predictor's relative influence within the GBM model.

2.4 Interpreting the results

The parameters of each model and their respective diagnostics will be presented and evaluated for accuracy and fit. Furthermore, each model will be interpreted in a numerical and qualitative context.

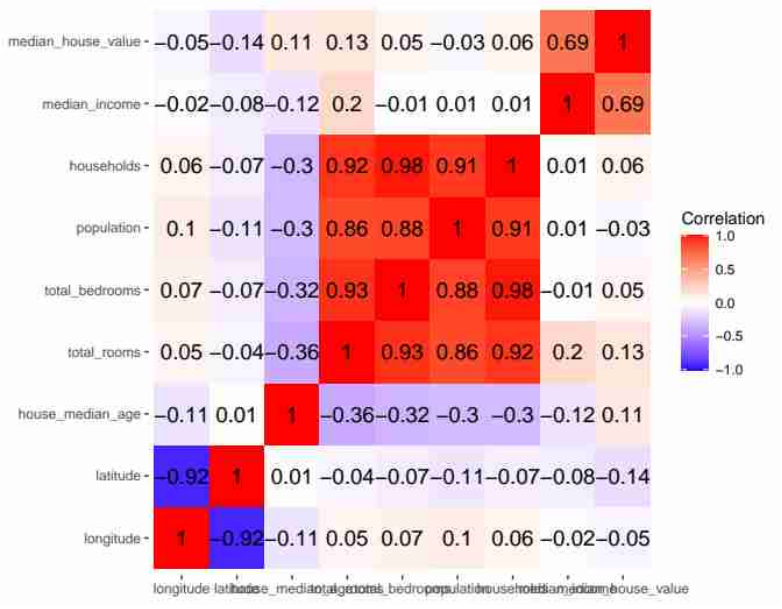
How a change in each predictor variable affects the response variable will be explained in both theoretical and practical terms. The analysis will be fit into the broader context, as key trends from the models will be laid out.

****

3. Exploratory Data Analysis

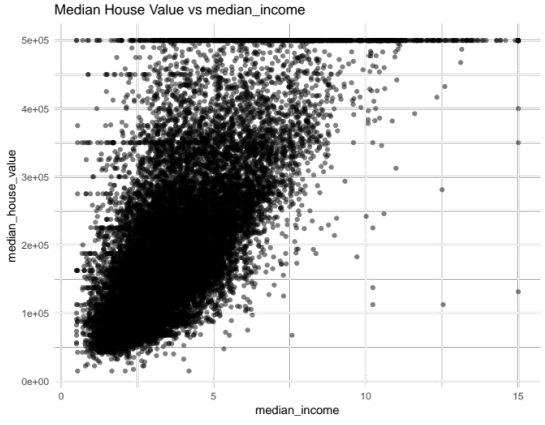
3.1 Summary Statistics and visualisations

One of the first priorities is to find any significant patterns between the variables. Below is a heat map of the correlation matrix for quantitative variables.



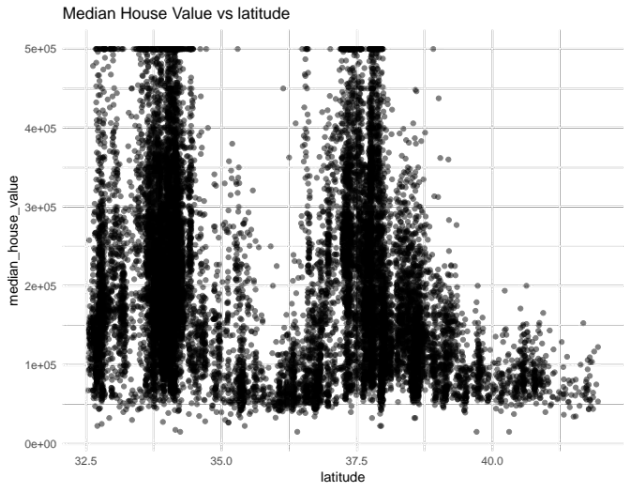
(3.1.1)

So far, only median income has any significant linear correlation with median house value (3.1.1). As such, median income must be added to any initial model. The pairwise below confirms the relatively linear relationship between median house value and median income (3.1.2).

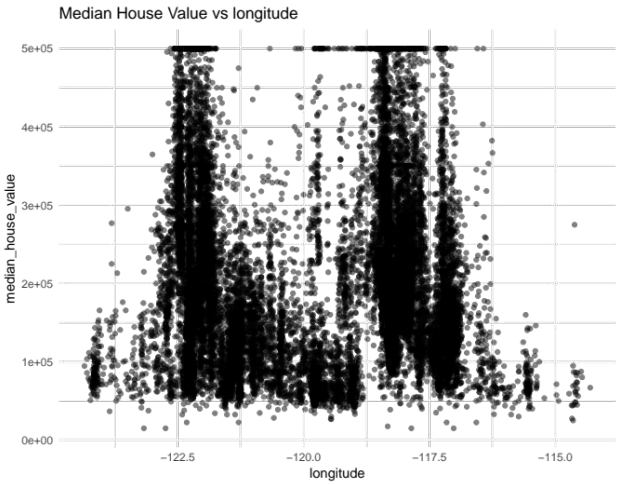


(3.1.2)

Viewing pairwise plots will help detect any non-linear patterns, which is helpful for variables with low correlation coefficients. Further analysis revealed significant non-linear relationships between median house value and the two coordinate variables (3.1.3, 3.1.4).

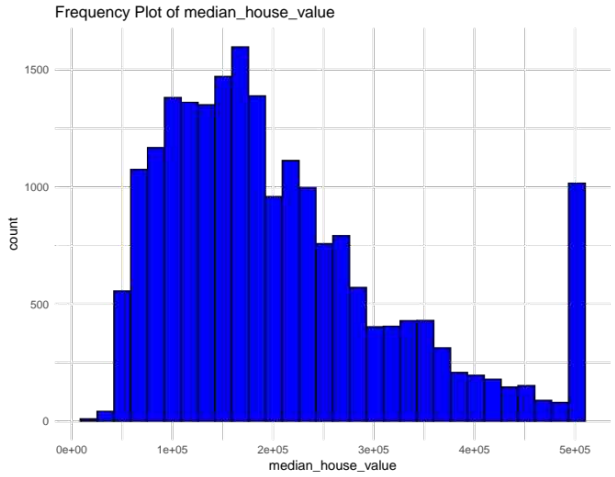


(3.1.3)

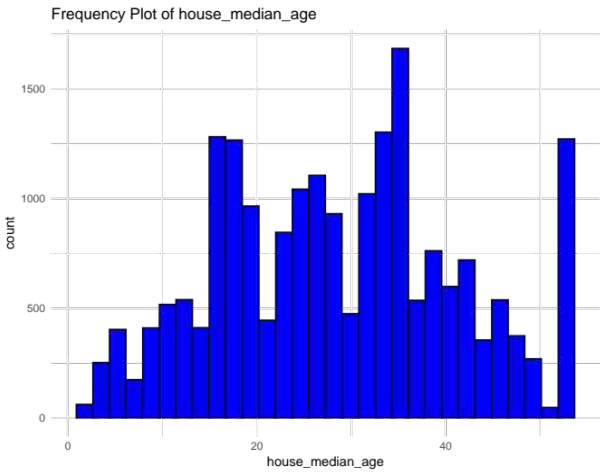


(3.1.4)

The non-linear relationships between median house value and both latitude and longitude will be modelled in the GBM method. Further inspections revealed significant number of extreme values for both median house value and house median age, which might be the result of entry errors. The next section will address how these extreme values will be handled.



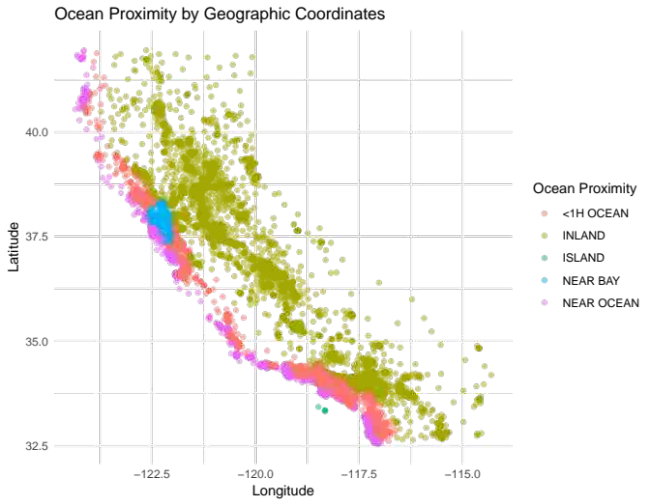
(3.1.5)



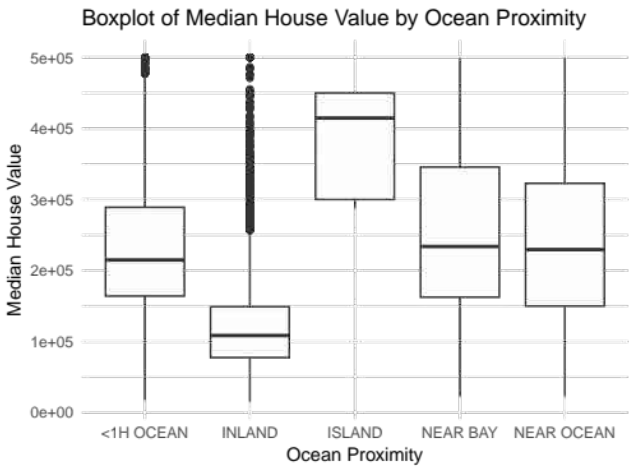
(3.1.6)

There are significant numbers of extreme values for median_house_value (3.1.5) and house_median_age (3.1.6). It's possible that the dataset has an upper limit for the values that these two variables can take on, meaning that these extreme values might actually be even higher in value without the upper limit in place. Such upper limits effectively creates multimodal distributions, which aren't ideal for linear analysis. This needs to be addressed when reshaping the data.

As ocean proximity is the only explicitly qualitative variable, additional statistics, tests, and visualisations are needed to determine its qualities and affect on median house value. Below are a series of boxplots of median house value by ocean proximity categories (3.1.8), as well as their geographic distributions (3.1.7).



(3.1.7)



(3.1.8)

It's worth noting that the counts of each categories are as follows (3.1.9):

Category	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
Total housing Blocks	9136	6551	5	2290	2658

(3.1.9)

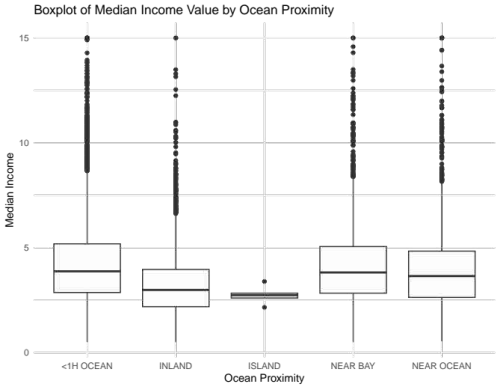
As such, an ANOVA Test is needed to determine if there's any significant differences between the ocean proximity categories regarding median house value. Here are the results (3.1.10):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ocean_proximity	4	6.544e+13	1.636e+13	1612	<2e-16
Residuals	20635	2.09e+14	1.015e+10		

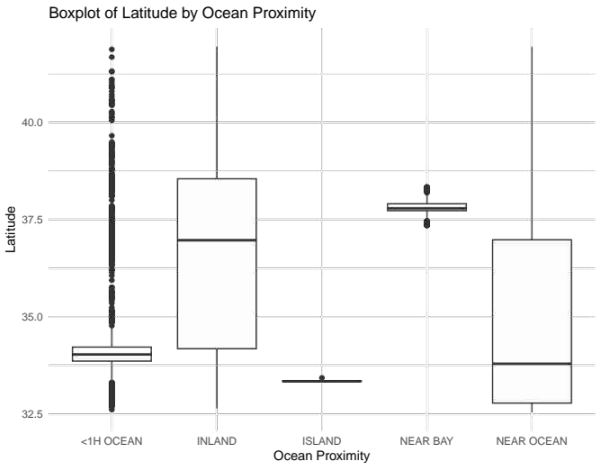
(3.1.10)

As there are 4 degrees of freedom for the numerator and 20635 degrees of freedom for the denominator, the F-value for $\alpha = 0.05$ is approximately 2.38, with the actual F value of 1632. Along with a p-value less than 2e-16, it's reasonable to conclude that ocean proximity does have a significant impact on median house value and should be included in at least one of the models.

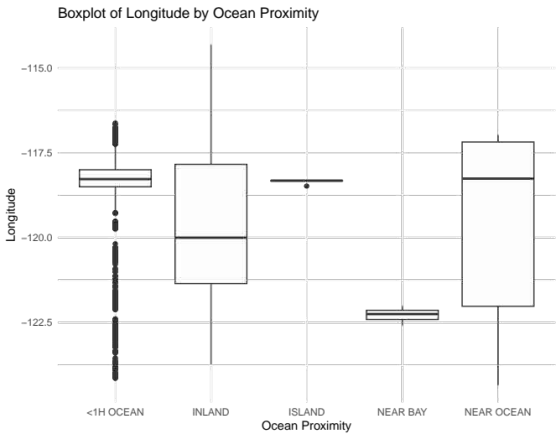
Further ANOVA tests involving ocean proximity found that the variable is significant to $\alpha = 0.05$ and less for all other quantitative variables. However, this could also be due to the large sample size, increasing the sensitivity of the ANOVA tests. Below are the boxplots for candidate predictor median income against candidate predictor ocean proximity (3.1.11), as well as the pairwise plots between median income against latitude and longitude (3.1.12, 3.1.13).



(3.1.11)

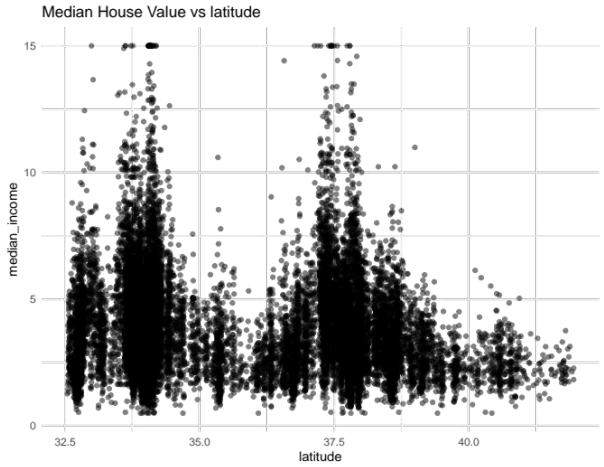


(3.1.12)

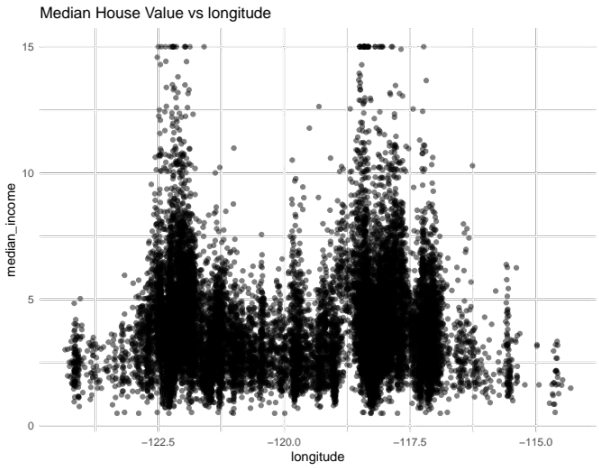


(3.1.13)

Despite the ANOVA test showing statistically significant differences, the boxplot show that the actual size of deviations of median income are quite small between categories of ocean proximity (3.1.11). However, there are significant variations of latitude and longitude against ocean proximity categories (3.1.12, 3.1.13), so the coordinate variables and ocean proximity can't be used in the same model. As GBM will be using latitude and longitude in addition to median income as predictors, pairwise plots of median income against latitude and longitude will be examined.



(3.1.14)



(3.1.15)

Although there clearly is a relation between median house value and the coordinates, the relationship is non-linear (3.1.14, 3.1.15). So far, median income, ocean proximity, latitude, and longitude are the four variables that appear to have a significant impact on median house value.

3.2 Reshaping the data

To make sure the modelling goes smoothly, all rows with null values are dropped. This step removes 207 rows. For all quantitative variables, all values beyond 3 standard deviations from the mean will be dropped (3.1.5, 3.1.6). This will ensure that the dataset is smoother and less affected by outliers. Dropping values beyond 3 standard deviations doesn't remove any rows, but choosing a narrower range might lead to too much data remove. Furthermore, this step is more of a precaution.

That said, this method fails to remove the extreme values for median house value and house median age. As such, all values greater than or equal to 500000 for median house value and equal to 52 for house median age are going to be dropped. There might be reason to remove rows where ocean

proximity equals ISLAND, as there are only 5 rows with relatively high median house values, but keeping the ISLAND rows in or removing them don't alter the results in any substantial way. This removes the abundance of extreme values for both variables. The trimming drops the number of observations from 20640 to 18362.

****

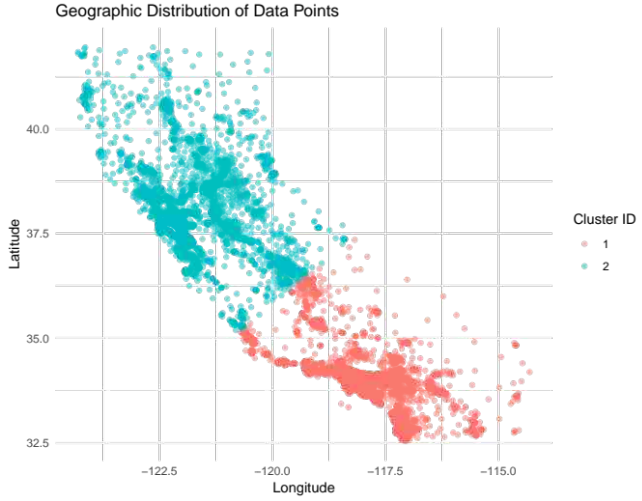
4. Inferential Statistics

4.1 Models and diagnostics

Both the multilinear model and GBM models will be presented along with statistics and visuals.

4.1.1 Multilinear Model

Latitude and longitude's significant non-linear relationship can still be account for in a multilinear model by simply clustering the dataset by latitude and longitude and running separate multilinear models for each cluster. Using k-Means clustering, the reshaped data is split into clusters. A new variable called cluster is created to facilitate the splitting of the clusters into distinct datasets. A for loop is used to find the optimal number of clusters, which was 2. Generating 6 or more clusters results in failure, as at least one of the clusters only have one category of ocean proximity. Below is the geographic distribution of each cluster (4.1.1.1).



(4.1.1.1)

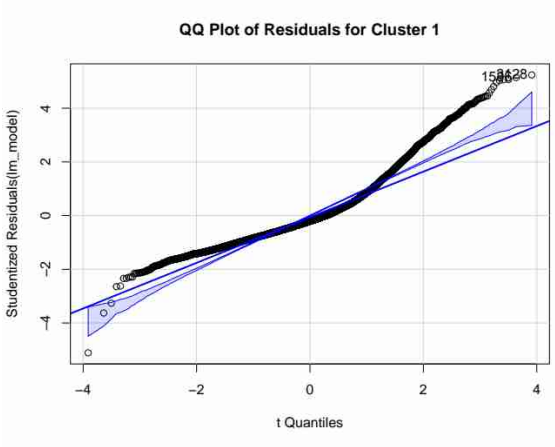
Effectively, the clusters split the dataset into Northern California, assigned to cluster 1, and Southern California, assigned to cluster 2. The multilinear model has median income, dummy variable versions of ocean proximity as the predictors and median income as the response. Note that all rows with NEAR BAY as the ocean category lie in cluster 2. Thus, only cluster 2 has a NEAR BAY dummy variable. The same can be said for all ocean proximity ISLAND rows for cluster 1. Furthermore, the ocean proximity category <1H OCEAN is modelled when all the other dummy variables equal 0. Below is the formula for cluster 1 (4.1.1.2), along with relevant diagnostic statistics and visuals (4.1.1.3, 4.1.1.4, 4.1.1.5):

$$\text{Cluster 1: } \widehat{\text{median\_house\_value}}_1 = 85230.8 + 34328.9\widehat{\text{median\_income}}_1 - 71256.6\widehat{\text{ocean\_proximity}}\text{INLAND}_1 + 201780.5\widehat{\text{ocean\_proximity}}\text{ISLAND}_1 + 1455.3\widehat{\text{ocean\_proximity}}\text{NEAR OCEAN}_1 \quad (4.1.1.2)$$

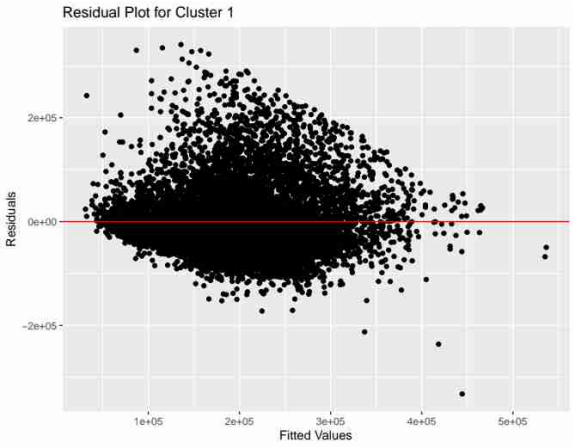
Statistics for Cluster 2	Value
Number of Rows	10844
Variance of Errors	4.242e+9

Standard Deviation of Errors	6.513e+4
R2 Adjusted	0.6683
Skewness	1.300
Kurtosis	2.425
Durbin-Watson Test	DW = 0.71423, p-value < 2.2e-16

(4.1.1.3)



(4.1.1.4)



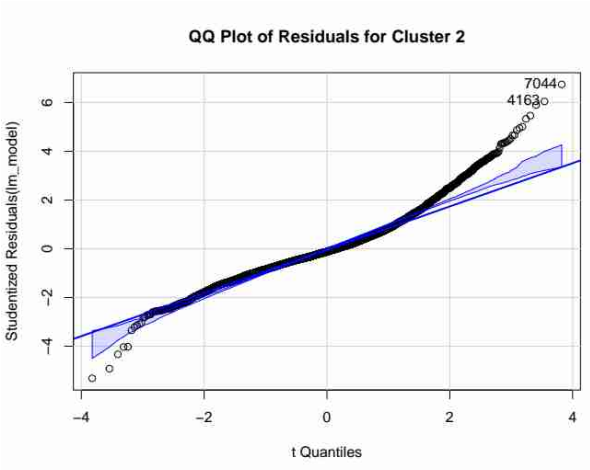
(4.1.1.5)

Below is the formula for cluster 2 (4.1.1.6), along with relevant diagnostics and visuals (4.1.1.7, 4.1.1.8, 4.1.1.9).

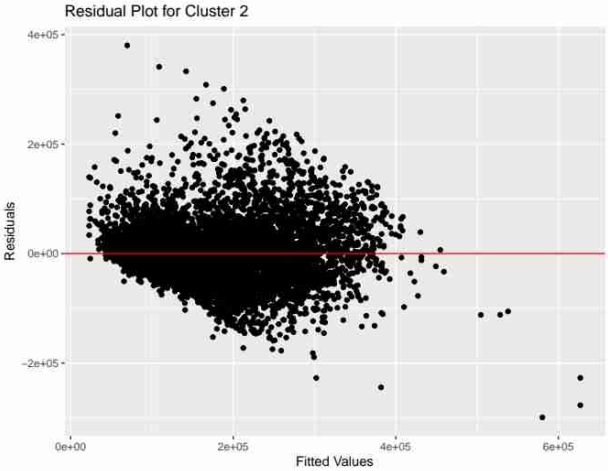
$$\text{Cluster 2: } \widehat{\text{median\_house\_value}}_2 = 73241.3 + 36920.3\widehat{\text{median\_income}}_2 \quad (4.1.1.6)$$
$$-68852.7\text{ocean\_proximity}\widehat{\text{INLAND}}_2 + 3065.4\text{ocean\_proximity}\widehat{\text{NEAR BAY}}_2$$
$$+45596.1\text{ocean\_proximity}\widehat{\text{NEAR OCEAN}}_2$$

Statistics for Cluster 2	Value
Number of Rows	7518
Variance of Errors	3.203e+9
Standard Deviation of Errors	5.660e+4
R2 Adjusted	0.5039
Skewness	0.8643
Kurtosis	2.760
Durbin-Watson Test	DW = 0.96079, p-value < 2.2e-16

(4.1.1.7)



(4.1.1.8)



(4.1.1.9)

The intercepts and beta coefficients were all statistically significant beyond $\alpha = 0.001$ (4.1.1.6), except for ocean_proximityNEAR BAY for cluster 1 with p-value of 0.14, and ocean_proximityNEAR OCEAN for cluster 2 with p-value of 0.426.

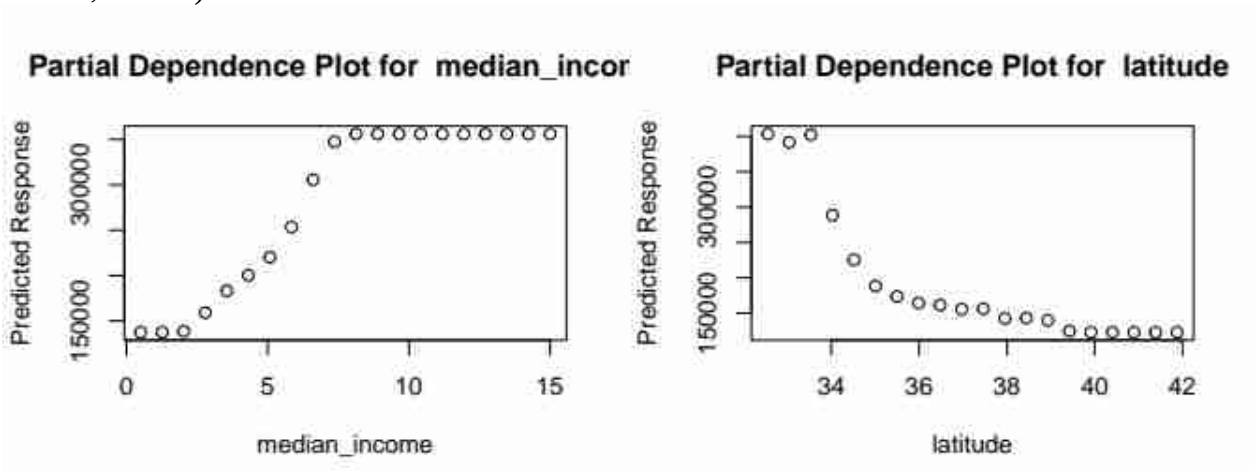
4.1.2 GBM Model

The GBM model uses the same data reshaping as the multilinear model. Predictor variables include median income, latitude, and longitude, with the response variable as median house value. Hyperparameterision is used, with shrinkage varying from 0.01 to 0.1. GBM also uses cross-validation to test the robustness of the model. As such, the dataset is configured into 5 batches for a 5-fold validation. The final GBM has the following parameters (4.1.2.1).

GBM Parameter	Value
Number of Trees	1000
Max Tree Depth	7
Shrinkage	0.1
Minimum Observations to Split Node	200

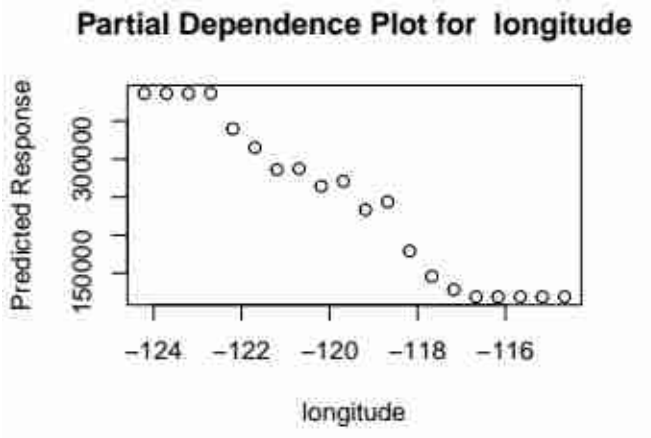
(4.1.2.1)

The following graphs show the impact of each predictor on the response variable (4.1.2.2, 4.1.2.3, 4.1.2.4, 4.1.2.5).

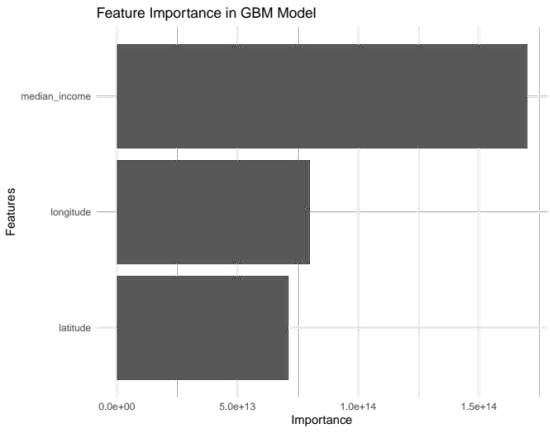


(4.1.2.2)

(4.1.2.3)



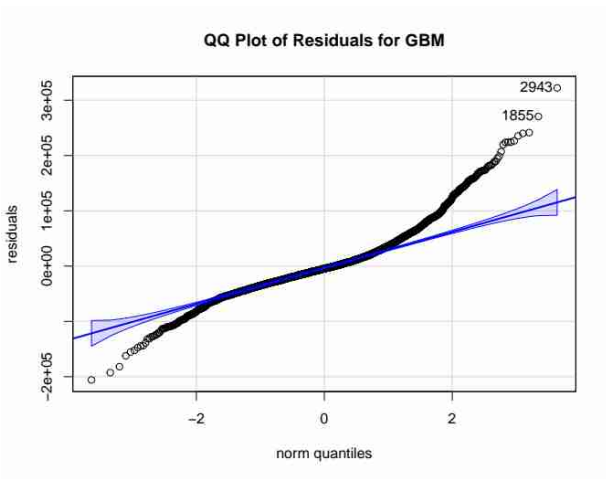
(4.1.2.4)



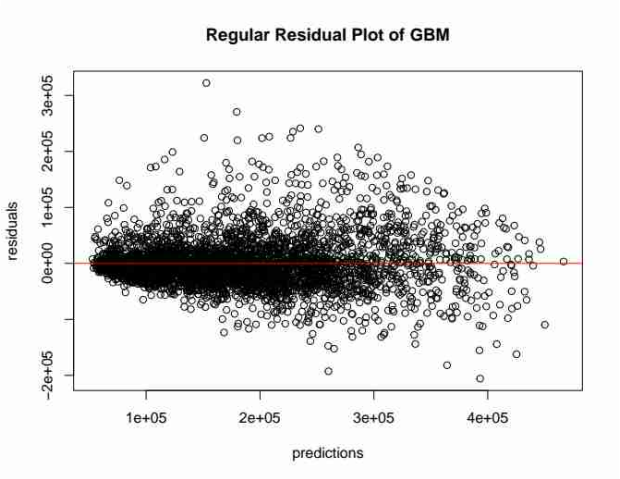
(4.1.2.5)

Below are the diagnostic variables and visuals.

Statistics for GBM	Value
R1 Score	0.8737
R2 Score	0.7633
Adjusted R2 Score	0.7634
Mean Squared Error	2.1247e+9
Variance of Errors	2.1246e+9
Standard Deviation of Errors	46093.4
Skewness	1.0836
Kurtosis	7.1523
Durbin-Watson Test	DW = 1.4996, p-value < 2.2e-16



(4.1.2.5)



(4.1.2.6)

****

5. Discussion

5.1 Fit of each model

The report will review the diagnostics of each model and assess their suitability in predicting median house value given the dataset.

5.1.1 Multilinear Regression Model

Despite attempts to cluster the dataset geographically, the multilinear model performed poorly. Despite moderate skewness values, cluster 1 is 1.300 and cluster 2 is 0.8643, and kurtosis values close to the normal distribution value of 3, cluster 1 is 2.425 and cluster 2 is 2.760, both models failed the Durbin-Watson Test, cluster 1 is 0.71423 and cluster 2 is 0.96079. This means that there is evidence of autocorrelation, in this case being greater than 0 for both clusters. Most of the variables were found to be statistically significant, with only ocean_proximityNEAR BAY of cluster 1 and ocean_proximityNEAR OCEAN of cluster 2 failing to reach 95% significance. Cluster 1 had a high R^2_{adj} of 0.6683, while Cluster 2 had a more mediocre R^2_{adj} of 0.5039. The average R^2_{adj} weighted by number of observations is 0.5712 (2.3.1.3, meaning that around 57.12% of the variance in the dataset can be explained by the multilinear regression model. Below is a brief summary of the model's adherence to the assumptions of the multilinear model.

Red = assumption not met, Green = assumption met

Assumption	Result
Linear relationship of variables	Evidence suggests that this assumption is met due to reasonably good R^2_{adj} values.
No Multicollinearity	ANOVA tests too sensitive to high sample sizes to provide strong correlation between median income and ocean proximity. Boxplots show negligible multicollinearity, so assumption met.
Independence	Durbin-Watson test failed for both clusters, assumption not met.
Homoscedasticity	Variance seems relatively constant, so assumption met.
Normal Distribution of Errors	Skewness and kurtosis values reasonable, so assumption met.

(5.1.1.1)

It's also worth noting that the inability to model latitude and longitude in a linear fashion also hurts the performance, as both of these variables have significant impact on median house value.

5.1.2 GBM Model

The GBM model has a strong fit for the dataset. Along with a high R^2_{adj} value of 0.7633, the Durbin-Watson test is passed with a value of 1.4996, which is close to the normal distribution value of 2. Although the skewness has only a moderate value of 1.0836, the kurtosis value is 7.1523, well above the normal distribution value of 3. This could be a sign of overfitting, but early tests showed that various methods for reducing overfitting, such as lowering max tree depth and number of trees, failed to significantly reduce the kurtosis value. Below is a brief summary of how the model fits the assumptions for GBM.

Assumption	Result
No Multicollinearity	There appears to be no linear relationship amongst the predictors, as latitude and longitude form the shape of California, and median income has non-linear relationships with the coordinate variables.
Balance of Response Variable	As none of the predictors are categorical in nature, assumption met.
Large Sample Size	Sample size is 18362, which is more than sufficient.
No Overfitting	Keeping the max tree depth at a low value of 5 and minimum observations for splitting nodes at a high value of 200 makes the GBM model less granular.

(5.1.2.1)

5.2 Implications of each model

Both models find predictors of significance, but how the predictors influence the response variable median house value differ.

5.2.1 Multilinear Regression Model

Assuming that the multilinear regression model is a good fit, it appears that both median income and most categories of ocean proximity have a linear relationship with median house value. Furthermore, clustering by latitude and longitude results in notable changes to the multilinear regression model. We will look at each cluster and then review the overall implication of the multilinear regression model.

Cluster 1's model makes the following claims (4.1.1.2): assuming median income is $0 and ocean proximity is <1H OCEAN, predicted median house value is $85,230.80; an increase in $10,000 of median income results in a $34,328.90 increase in median house value; having an ocean proximity of INLAND results in a $71,256.6 decrease in median house value compared to <1H OCEAN;

having an ocean proximity of ISLAND results in a $201,780.50 increase in median house value compared to <1H OCEAN; having an ocean proximity of NEAR OCEAN increases the median house value by $1,455.30 compared to <1H OCEAN. All of these individual claims are made assuming that other variables are held constant.

Cluster 2's model makes the following claims (4.1.1.6): assuming median income is $0 and ocean proximity is <1H OCEAN, predicted median house value is $73,241.30; an increase in $10,000 of median income results in a $36,920.30 in median house value; having an ocean proximity of INLAND results in a $68,852.7 decrease in median house value compared to <1H OCEAN; having an ocean proximity of NEAR BAY results in a $3,065.40 increase in median house value compared to <1H OCEAN; having ocean proximity of NEAR OCEAN increases the median house value by $45,596.10 compared to <1H OCEAN. All of these individual claims are made assuming that other variables are held constant.

Both clusters have similar beta coefficients for median income, which makes sense as median income is a strong indicator of the desirability of residential block. Furthermore, median income could also be used as a stand in for other strong indicators excluded from the dataset, such as crime rate (Metz, 2016), in which higher crime rate leads to a decrease in property value (Boggess, 2013). There seems to be similar trends in ocean proximity, with inland properties having the lowest values, and <1H OCEAN and NEAR OCEAN having some of the highest values. ISLAND has by far the highest effect on median house value. But given that there are only 2 observations in the reshaped data, it would be difficult to extrapolate this finding beyond the dataset. Although NEAR BAY has a weak level of significance, the fact that NEAR BAY properties are rather inland and still similar in median house value to <1H OCEAN means that these properties are highly desirable. This isn't a surprise, as Silicon Valley, the technology capital of the United States and to some extent the entire world, is located in the NEAR BAY area.

The fact that NEAR OCEAN properties generally have higher median house values than <1H OCEAN is surprising, given the perceived desirability of ocean front properties. It is perhaps the high desirability of finite and scarce ocean front properties could spillover and contribute to an increased demand and hence and increased median house value for NEAR OCEAN properties, which as more inland than <1H OCEAN properties. If there are only so many <1H OCEAN properties, demand will have to shift to the next best available area.

5.2.2 GBM Model

It's generally more difficult to interpret GBM models than linear models, but there are still certain takeaways that can be made. The feature importance graphs show that median income is the most impactful variable, followed by longitude and then latitude (4.1.2.5). Generally, an increase in median income increases predicted median house value, which starts off at $150,000, rises to $300,000 before plateauing for median incomes $80,000 or greater (4.1.2.2). Median house value starts off around $300,00 for lower latitudes and quickly decreases before hitting a floor at $150,000 (4.1.2.3). For longitude, going from west to east gradually decreases median house value from $300,000 to $150,000 (4.1.2.4). Partial dependencies are an approximation of what happens when a single variable is adjusted with all other predictors held constant.

However, it must be said that the GBM model result from a large number of decision trees, with each in a decision tree performing splits based on individual variables. The left branch of a tree are for variable values that meet the threshold and the right branch of a tree are for variable values that don't meet the split threshold. This can create various strata of values within one variable, allowing

for the representation of non-linear relationships. However, it's incredibly challenging to follow the exact chain of logic for GBM models in general.

****

6. Conclusion

Real estate is a complex industry, as house prices depend on a myriad of different variables, many of which aren't included in the dataset. Factors such as demographics (Khater, 2021) are relevant predictors aren't adequately captured by the dataset variables. Still, both models did have promising explanatory power, with the multilinear regression model yielding a R^2_{adj} of 0.5712 and the GBM model resulting in a R^2_{adj} of 0.7732. While the GBM model is stronger in prediction, it's much harder to extrapolate the findings due to its complex decision-making process and tendency to overfit the dataset. What is clear is that demographic variables such as median income and regional variables such as latitude, longitude, and ocean proximity have a noticeable affect on median house value of residential blocks in California in the year 2001. Going forward, it'll be a great idea to include additional variables, such as crime rate and demographics. That said, the old saying that location is the biggest factor in real estate has been more or less validated by both the multilinear regression model and the GBM model.

****

7. Bibliography

Boggess, L. N., Greenbaum, R. T., & Tita, G. E. (2013). Does crime drive housing sales? Evidence from Los Angeles. Journal of Crime and Justice, 36(3), 299–318. Taylor & Francis Online. https://doi.org/10.1080/0735648x.2013.812976

Khater, S., Kiefer, L., Yanamandra, V., & Villa, G. (2021). U.S. Population Growth: Where is housing demand strongest? Freddie Mac; Freddie Mac. https://www.freddiemac.com/fmac-resources/research/pdf/202101-Insight-12.pdf

Metz, N., & Burdina, M. (2016). How neighborhood inequality leads to higher crime rates [Review of How neighborhood inequality leads to higher crime rates]. LSE Research Online; LSE. Retrieved December 14, 2023, from https://eprints.lse.ac.uk/67733/3/blogs.lse.ac.uk-How_neighborhood_inequality.pdf