**Data Collection**

The first dataset considered for this report was regarding the accidents to analyse the factors leading to the accidents in Australia and provide the business value, however, it did not hold any relevant independent variable which could used for regression.

For this reason, The final dataset considered for the analysis is based on the 1990 census data and refers to the houses found in a specific California district. It was obtained via the public access website Kaggle.com. The dataset originally comprised of 20,641 observations and included the following variables:

1. **Longitude**: A measure of how far west a home is; a larger number indicates that it is further west.

2. **Latitude**: A measure of how far north a home is; a larger value indicates that it is further north.

3. **housingMedianAge**: The median age of a home inside a block; a lower figure indicates a recent structure.

4. **totalRooms**: The total rooms in a building.

5. **totalBedrooms**: The total bedrooms in a building.

6. **population**: The total people who live in a certain block.

7. **households**: The total number of households in a block, which is a collection of individuals that live in a housing units.

8. **medianIncome**: The median income for families in a neighbourhood (measured in tens of thousands of US Dollars)

9. **medianHouseValue**: The median house value for a block of houses (measured in US Dollars)

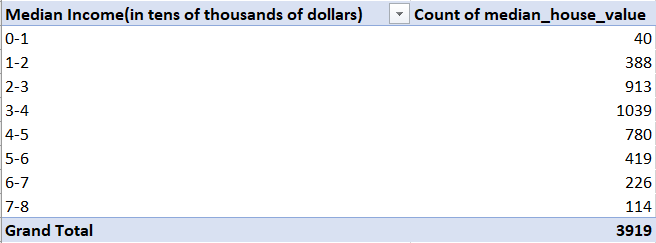
10. **oceanProximit**y: The house's proximity to the ocean/sea.

The medianHouseValue has been taken as the independent variable for the linear regression and others variable have been analysed according to its relevancy to be taken as the dependent variable.

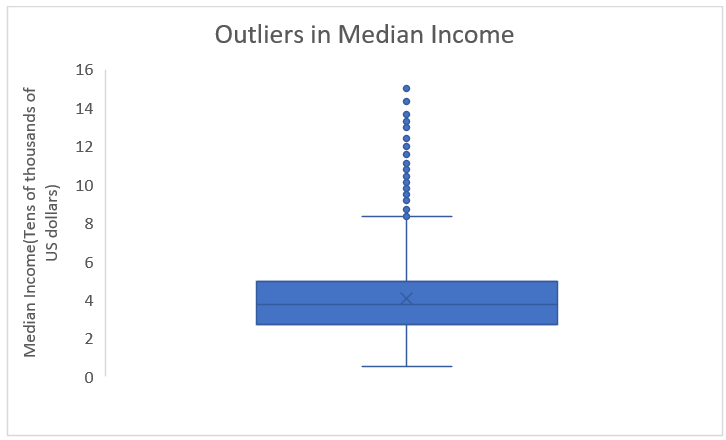
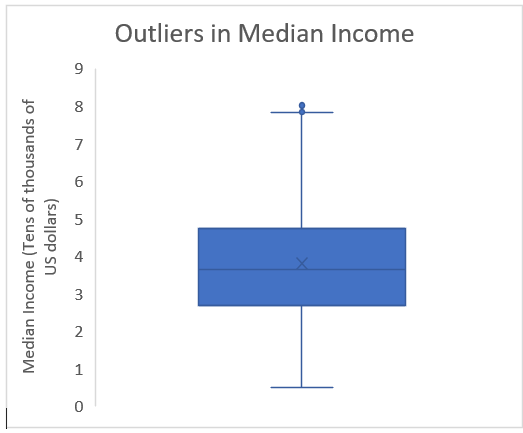
**Data cleansing**

Steps involved

1. It can be noted that latitude and longitude doesn’t hold relevancy to analyse the medianHousevalue, thus, it has been discarded.
2. The ‘oceanProximity’ has five categories, however, only two categories which are ‘NEAR BAY’ and ‘NEAR OCEAN’ were taken for this analysis. This reduced the number of record to 4949 in the dataset.
3. The significance of the variables was determined by constructing a pivot table between the Median income and the median house value. It provides us with the count of the records of the house value in the given range. Thus, there is a correlation between these variables. Similarly, other variables’ significance were also analyzed.



1. Next, The Boxplot was created for each of the variable to check the outliers. The outliers were removed by deleting the record which were above the upper limit. It was observed that the total number of records which were present above the upper limit were very less compared to the total record, thus, after removing all the outliers the total records in the dataset were 3920.

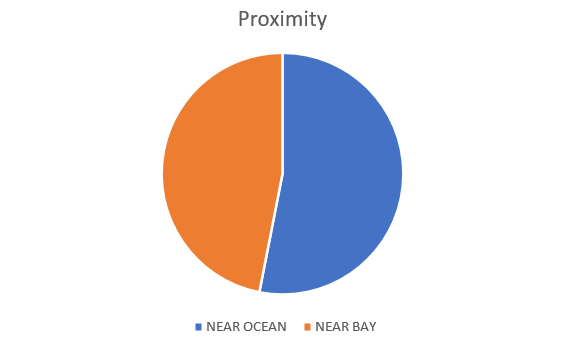
*Figure: Before Data cleansing of Median Income variable Figure: After Data cleansing of Median Income variable*

1. The variable Proximity holds two categories which are ‘NEAR BAY’ and ‘NEAR OCEAN’. It has been transformed to numerical by creating a dummy variable ‘Dproximity’ where ‘NEAR BAY’ is given value 1 and ‘NEAR OCEAN’ is given value 0.

**Analysis**

In order to analyse the given data, we are using multiple linear regression to develop a predictive model and pie chart to provide descriptive analysis. For obtaining the most accurate model, several regressions were run.

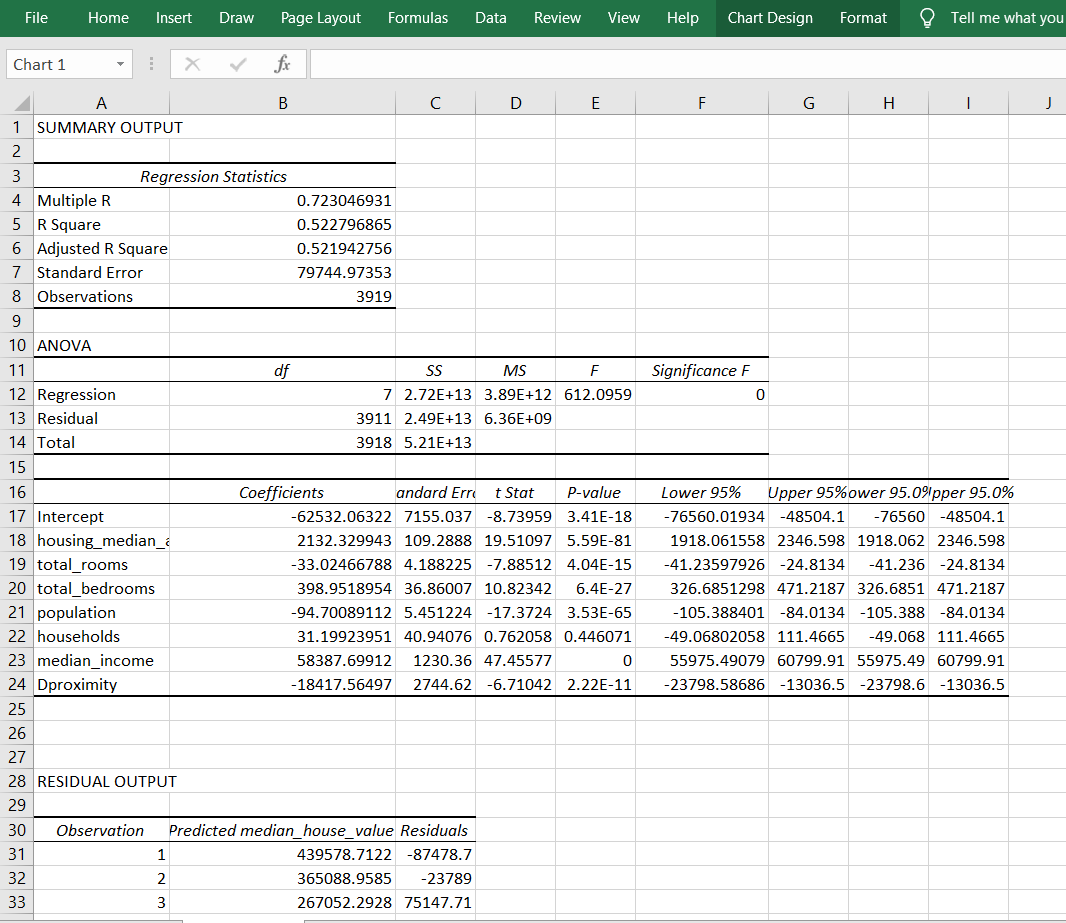
**Descriptive Analysis**



For the given dataset, the data is nearly equally divided among the proximity of ‘NEAR OCEAN’ and ‘NEAR BAY.’

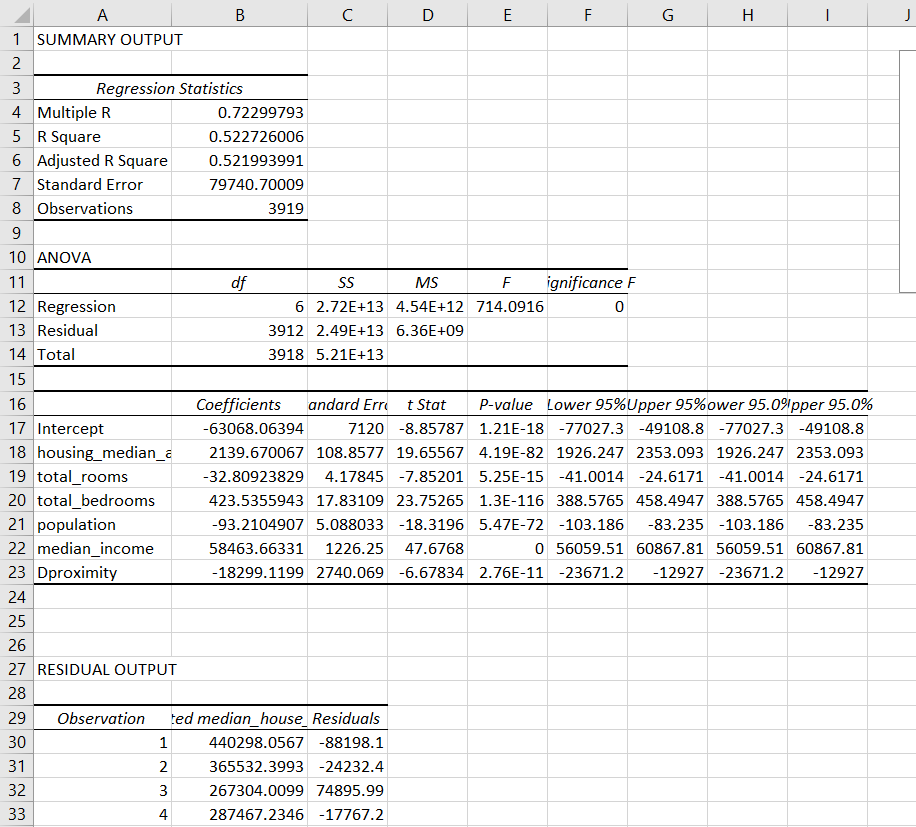
**Prescriptive analysis**

The first iteration of the predictive model generated for the multiple linear regression has the following output.



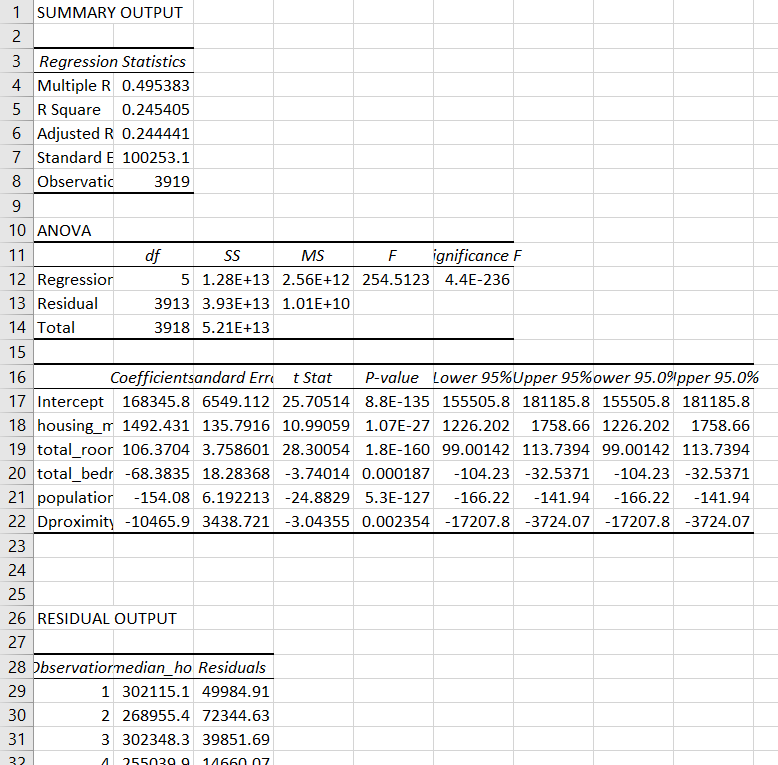
*Figure: First Iteration of the Multiple Linear Regression*

It can be observed that P-value of Households is greater than 0.05 and median income is 0. In order to increase the accuracy of the model, for the second iteration ‘households’ variable was not taken in the multiple linear regression. The output of the multiple linear regression model after second iteration can be observed below.



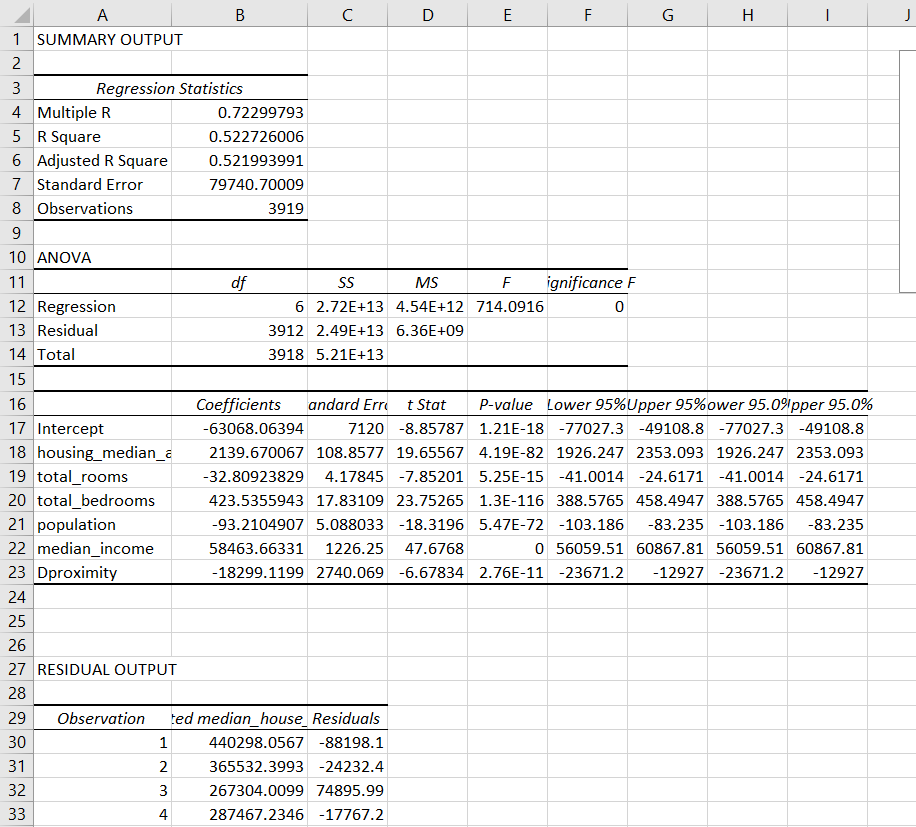
*Figure: Second Iteration of the Multiple Linear Regression*

It can be observed that the r-square value remains the same after the second iteration, however, there can changes observed in the coefficients of other variables. Also, P-values of other variables also remains less than 0.05. As the P-value of the median\_income is 0, the third iteration of the regression model was taken by removing the median\_income from the linear regression. The output of the third iteration can be observed below.



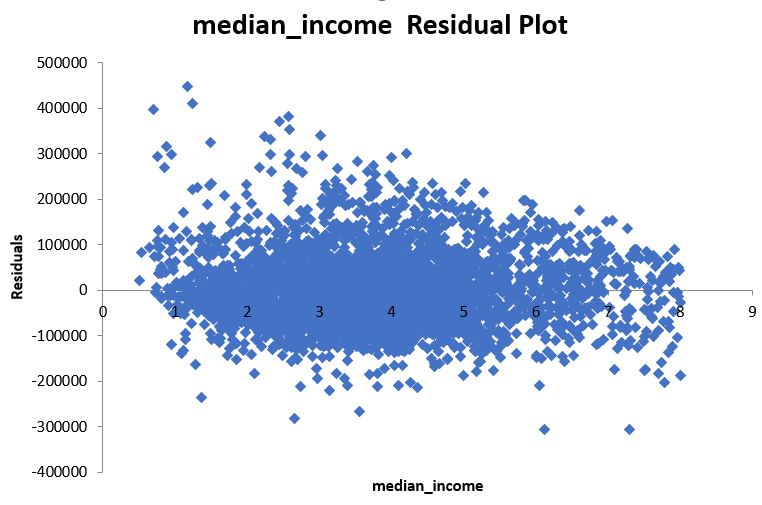
*Figure: Third Iteration of the Multiple Linear Regression*

Here, when the median\_income variable was removed from the regression, the accuracy of the model reduced by around 30 percent. This shows that the ‘median income’ has high significance in this model. Thus, the final predictive model was taken from the second iteration of the regression.

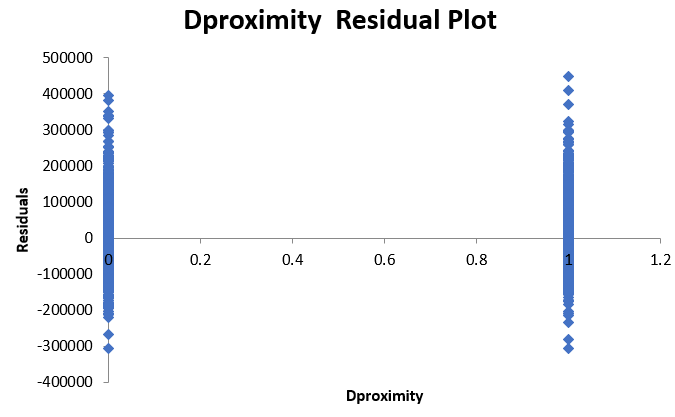


*Figure: Final Multiple Linear Regression considered for Predictive analysis*

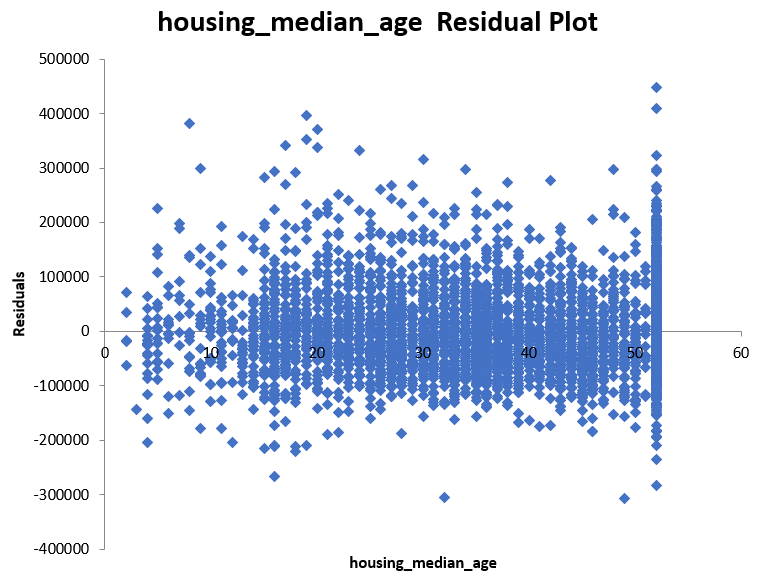
Next, when compared to other variables, the coefficients of 'Median income,' 'Dproximity,' and 'housing median age' are higher. As a result, for this research, the residual plots of these factors were taken into consideration to determine the effects of these variables on housing prices.



No pattern can be observed in the ‘median income’ residual plot as for the low median income, the residual value is present in both the negative as well as in the positive. Furthermore, the residual values decreases as the median income increases. The predicted observation shifts more towards the actual observations in the higher median income.



Similarly, there are positive and negative residuals both for the houses near the bay and the houses near the ocean.



Lastly, there are positive and negative residual values for the houses which are newly built and for the older buildings. However, it can be noted that there are high residual values for the oldest buildings. This means that some of the house prices predicted by this model were less, but in actual it was high and vice-versa.

**Discussion of the model**

Multiple linear regression was used to examine house prices in relation to various parameters such as median income, median house age, total rooms, and so on. The significance F and P-value of the dependent variable coefficients are less than 0.05, indicating a meaningful association between house values and the dependent variables.

From the model, the following equation can generated:

**Median House Value** = - 63068.06 + 2139.67 \* **Housing Median Age** - 32.81 \* **Total Rooms**

+ 423.53 \* **Total Bedrooms** – 93.21 \* **Population** + 58463.663 \* **Median Income**

* 18299.12 \* **Dproximity**

The R-square of the generated model is 0.52 which means that 48 percent of the predicted output is different from the actual output. This suggests that there are more independent variables that can explain the house values.

According to the equation, 'housing median age' has a positive coefficient, showing that every year a house gets older increases its worth by a factor of 2139.67, indicating a positive correlation in which older houses are more valuable in California than newly built ones. This could be due to older houses being built in more prominent regions such as the city, whereas newer houses being built in new areas that are expanding and becoming part of California.

'Median Income' has a positive coefficient as well, multiplying by 58463.663. This means that if the average household income rises by one, i.e. (10,000 because stated in tens of thousands of dollars), the house value rises by a factor of 58463.663. This could be attributed to increased demand for houses in California as people's purchasing power improves.

Finally, Dproximity is inversely related to housing values. The Dproximity variable is a dummy variable used to generate a numerical variable for the categorical data of proximity, in which 'NEAR BAY' is associated with value 1 and 'NEAR OCEAN' is associated with value 0. Thus, if Dproximity is 1, the house is close to the Bay and its value is reduced by a factor of 18299.12. If, on the other hand, Dproximity is 0, this indicates that the house is close to the ocean and that the house value does not fall because the coefficient is null. This means that the house built near bay is significantly less expensive than the house built near the ocean.