**CS 6301.004: Assignment 1 (Linear Regression)**

**Team Members:**

Teja Kiran Chunduri

Sneha Nandigala

Shree Charan Varma Enugandla

Prathyusha Kanmanth Reddy

**Linear regression on California housing dataset**

***Abstract***

Linear regression is a commonly used technique for predictive analysis, forecasting data and error reduction. The following report focuses on the methods used for pre-processing the data and analysing attributes of the California housing dataset.

1. **Introduction**

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. It can be used to fit a predictive model to an observed [data set](https://en.wikipedia.org/wiki/Data_set) of values of the response and explanatory variables. A fitted linear regression model can be used to identify the relationship between a single predictor variable and the response variable when all the other predictor variables in the model are "held fixed".

1. **Dataset**

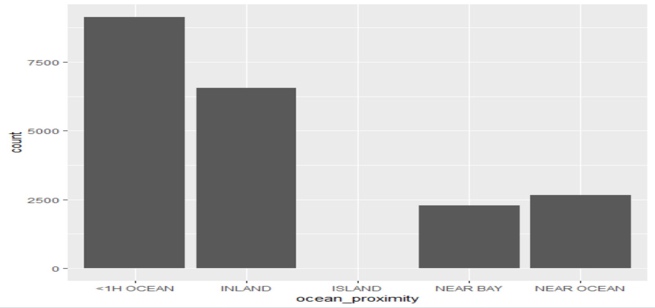
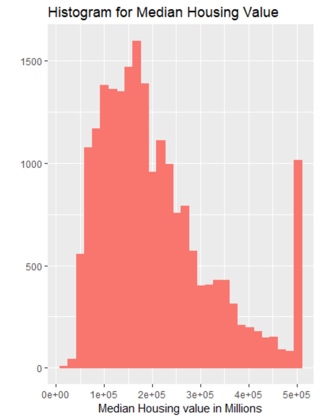
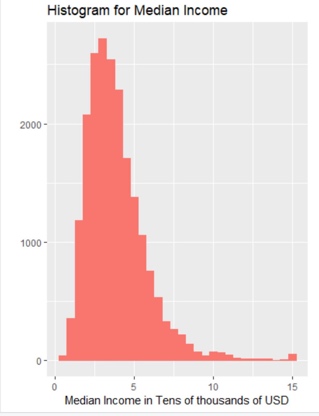
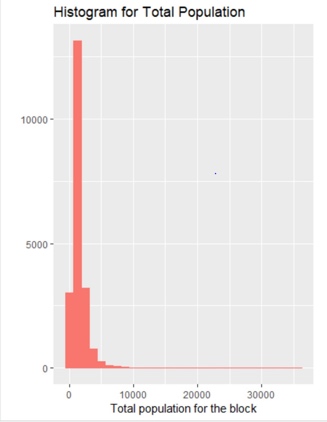
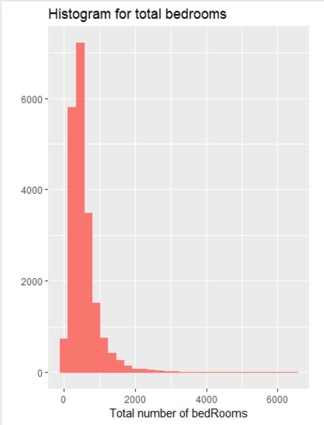
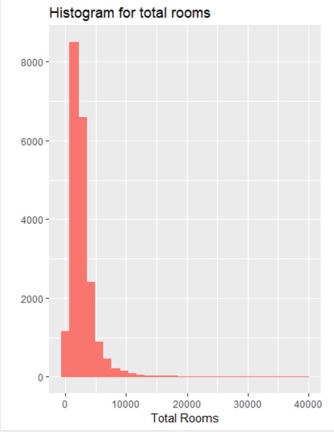
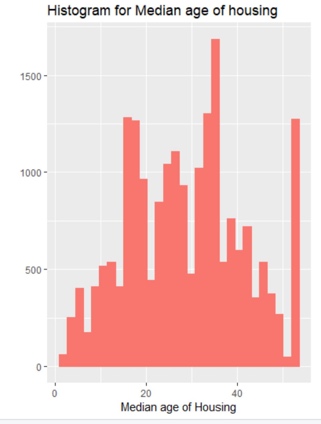
California housing dataset pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. There are 20,640 rows and 10 columns.

The column names and their data types are as follows, their names are self-explanatory:

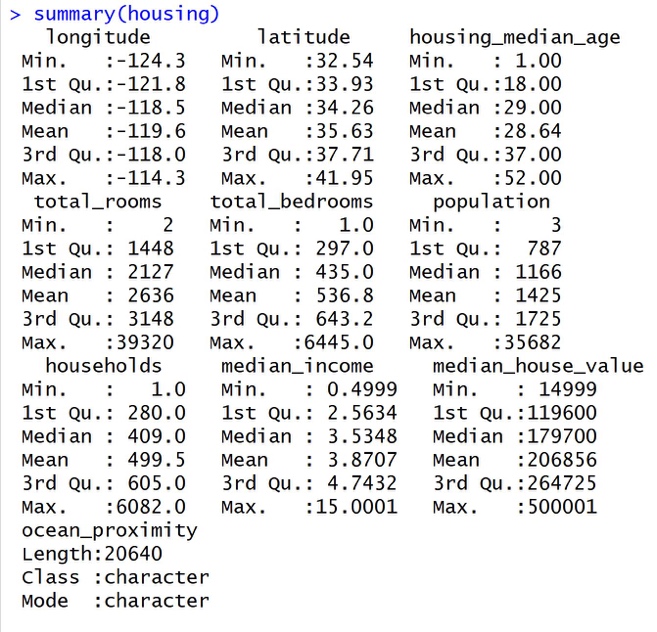
|  |  |
| --- | --- |
| *Column Name* | *Data Type* |
| Longitude | Numerical |
| Latitude | Numerical |
| Housing\_median\_age | Numerical |
| Total\_rooms | Numerical |
| Total\_bedrooms | Numerical |
| Population | Numerical |
| Households | Numerical |
| Median\_income | Numerical |
| Median\_house\_value | Numerical |
| Ocean\_proximity | Categorical |

1. **Attribute analysis**

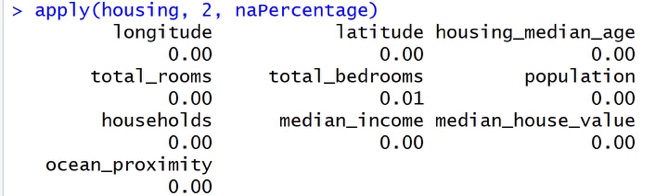
The California housing data distribution:



Summary of the dataset is as follows,



We have observed that the initial dataset contained null values. The distribution of null values is as follows:



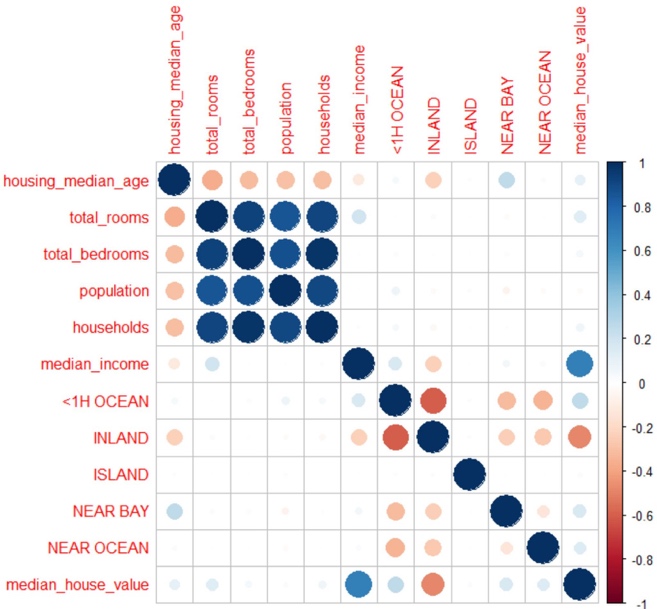
The null values in *total\_bedrooms* where replaced by the median value of that attribute.

*Ocean\_proximity* is categorical data, So the attribute values were spread and converted to numerical data. Further, the data was normalized using standard normalization.

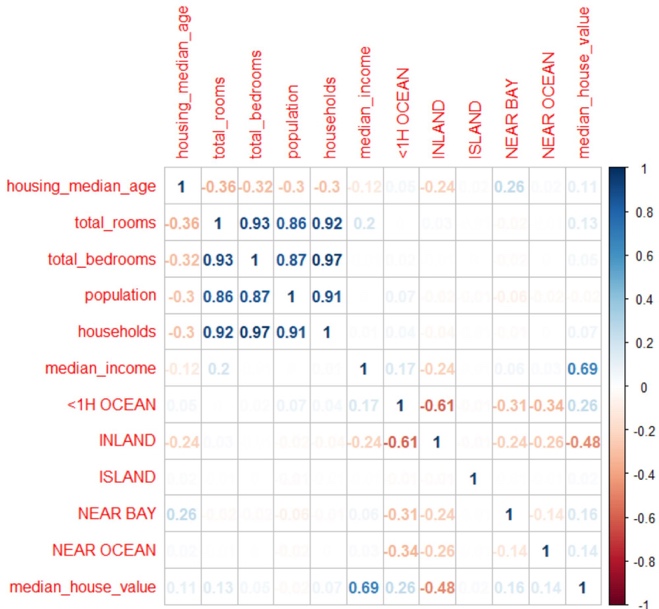
Normalized Data:



Correlation Plot - Circle:



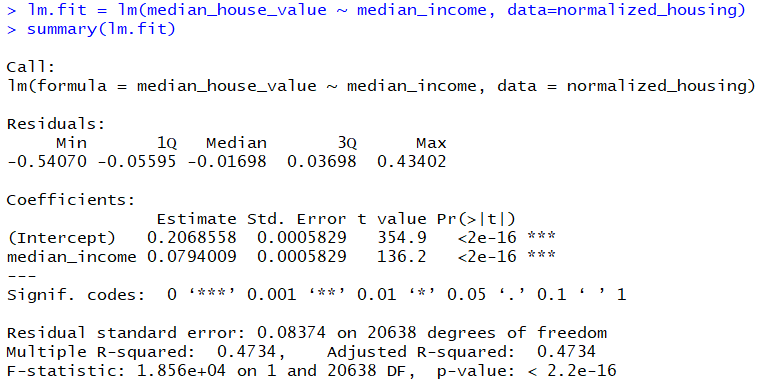
Correlation Plot - Numeric:



From the above figures, it can be observed that there is a strong correlation between median\_house\_value and median\_income. Apart from that, total\_rooms, total\_bedrooms, population and households are strongly correlated to each other.

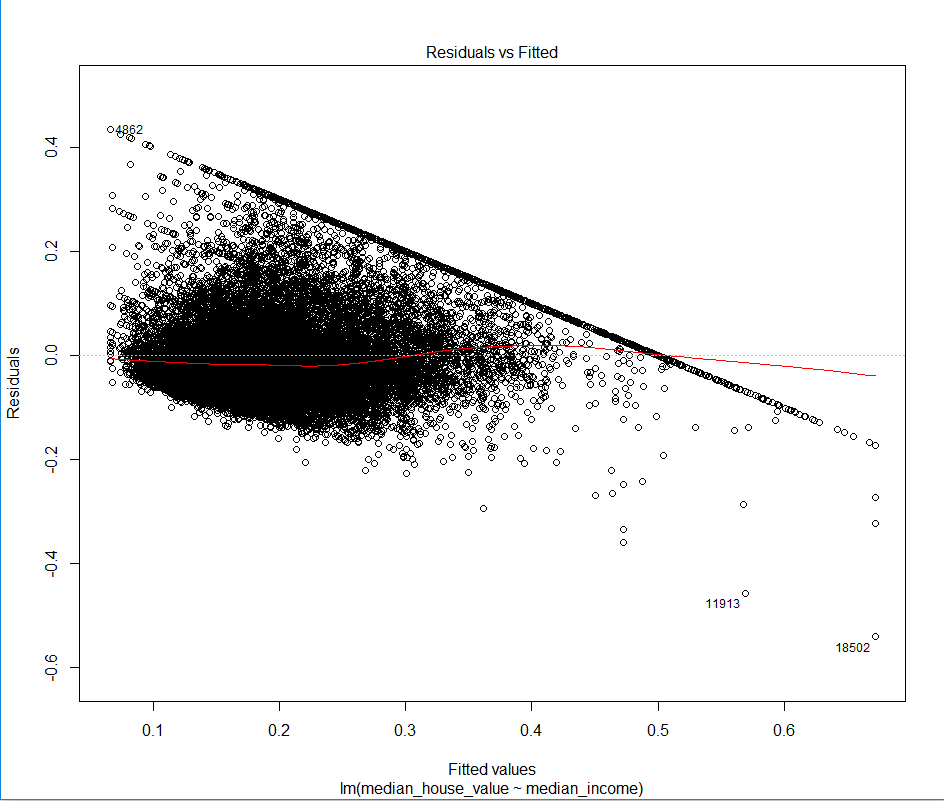
1. **Predicting Home Prices**

We tried predicting the median\_house\_value as a function of median\_income first.

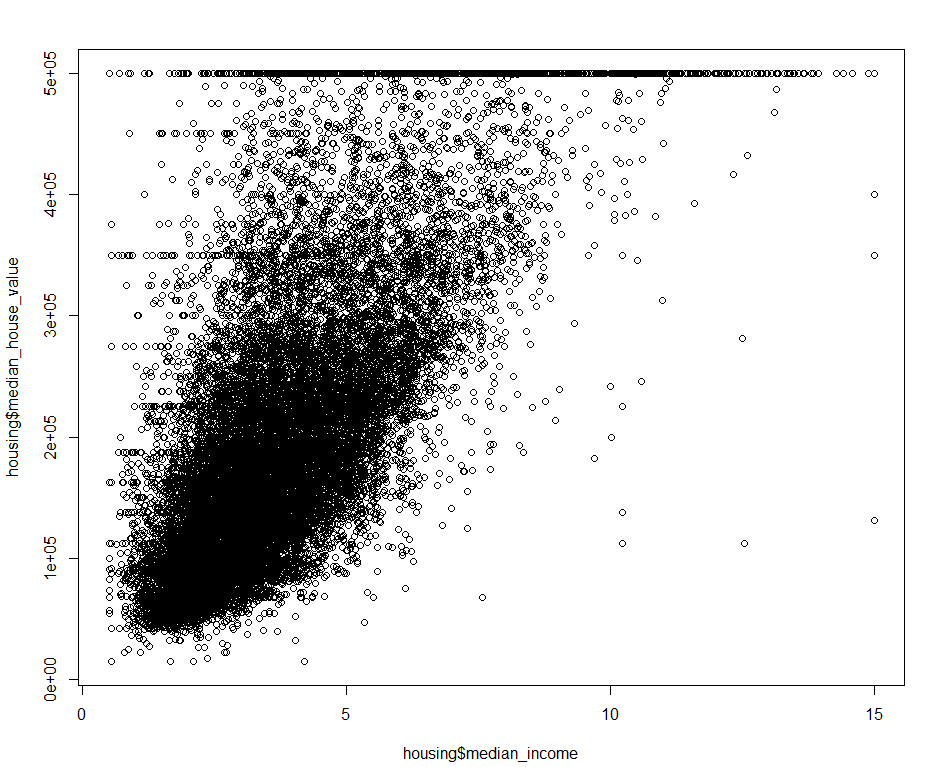


The t-value here is high and p-value is very low. This shows that the attributes are statistically significant, and we can reject the null hypothesis that the coefficient is equal to zero. F value is 101.334 and Multiple R-squared is 0.4734.

Plot of Residual vs Fitted values:

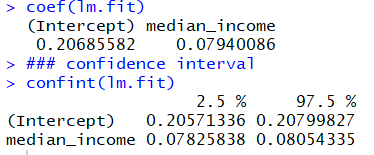


Plot of the median\_house\_value and median\_income:

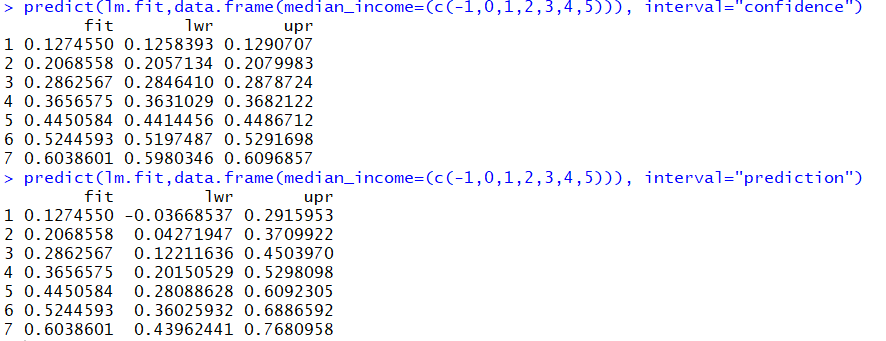


By viewing the plot and the coefficients, we can derive an equation to predict the house prices.

median\_house\_value = (0.20685582 + 0.07940086 \* median\_income)\*100000



Predicting using some test values:

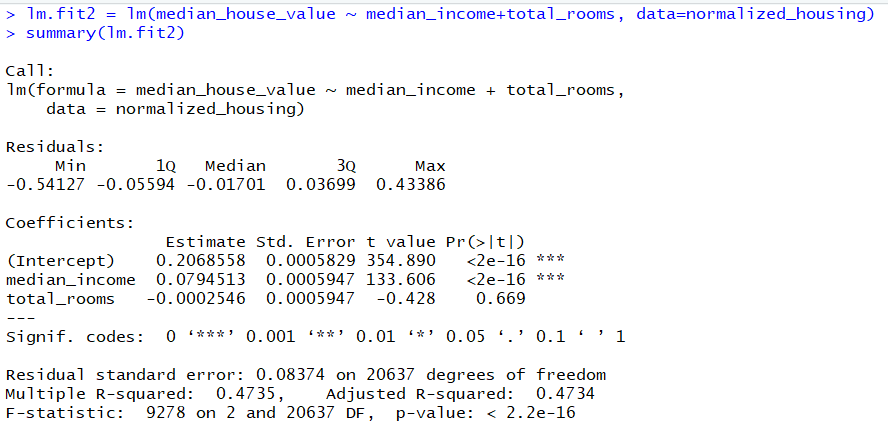


The predicted value(fit) must be multiplied with 100,000 to get the final predicted median house value. Input for median\_income is given in normalized form which ranges from -1 to 5. The predicted values are in the range of 100k – 600k which is close to what the original data depicts.

1. **Multiple Linear Regression**

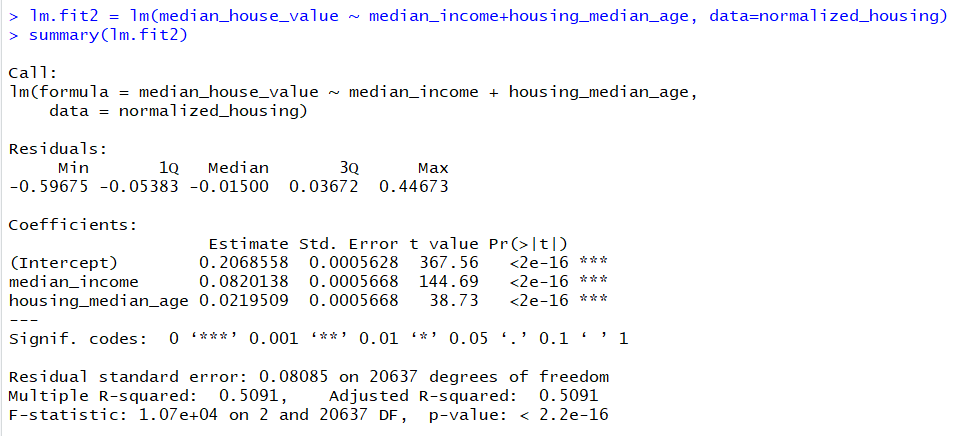
Creating the model using multiple attributes.

1. median\_house\_value against median\_income and total\_rooms:



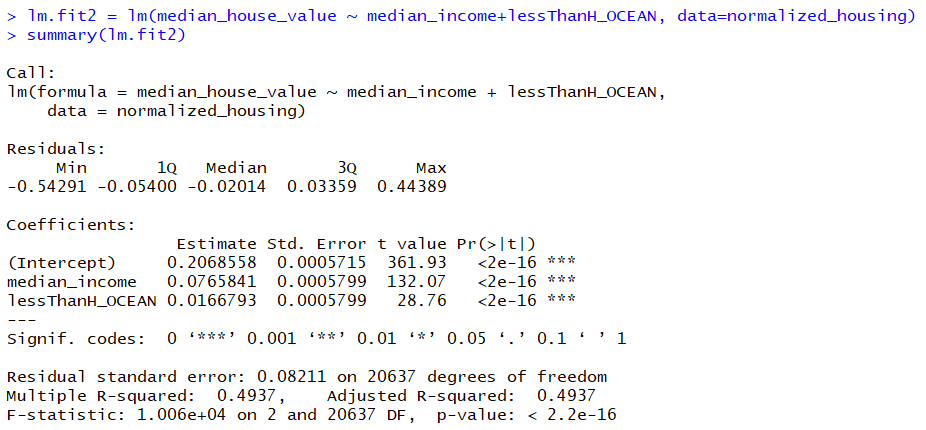
Here, we can observe that F-Statistic has increased drastically but the median\_income, p-value, t-value and residual standard error remain the same. Also, total\_rooms is not statistically significant because of its high p-value

1. median\_house\_value against median\_income and house\_median\_age:

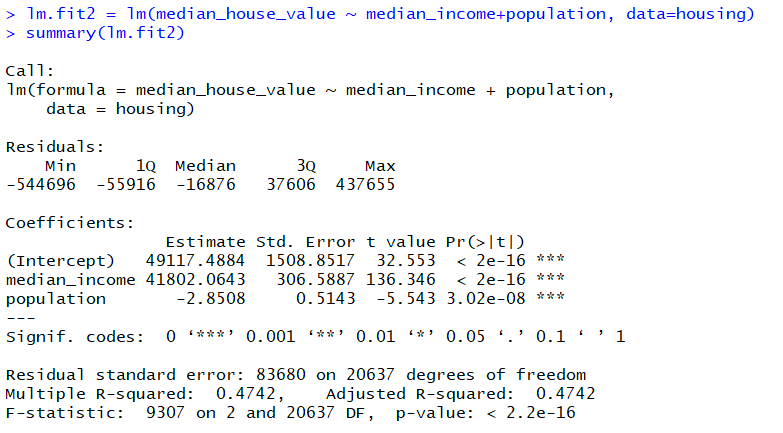


Here, we can observe that F-statistic has reduced to 58.42 and the remaining values are the same. Also, housing\_median\_age is statistically significant because of it’s low p-value.

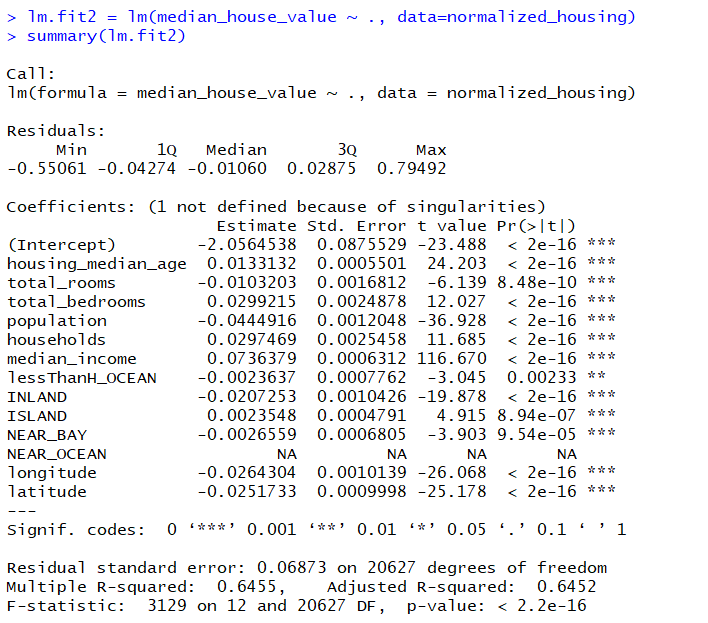
1. median\_house\_value against median\_income and lessThanH\_OCEAN:



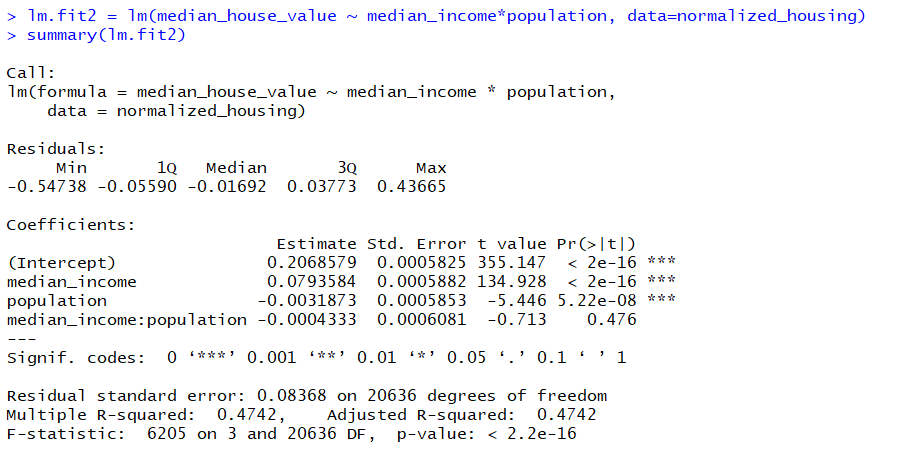
1. median\_house\_value against median\_income and population:



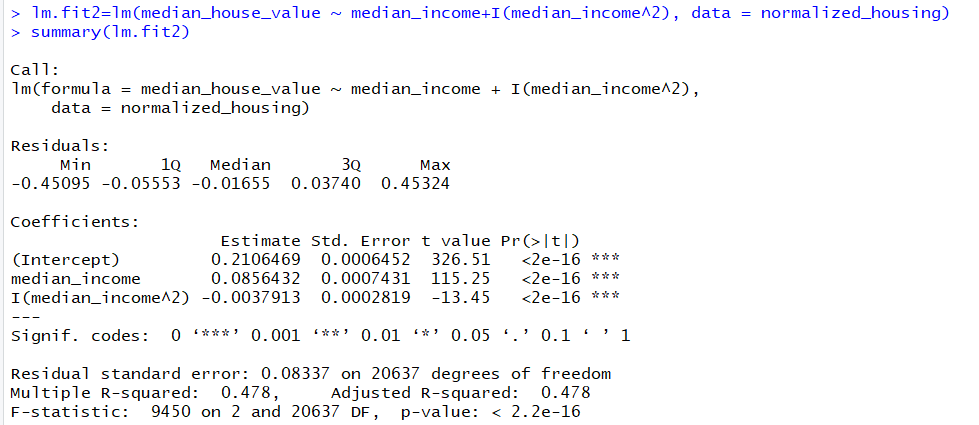
1. median\_house\_value against all the attributes:



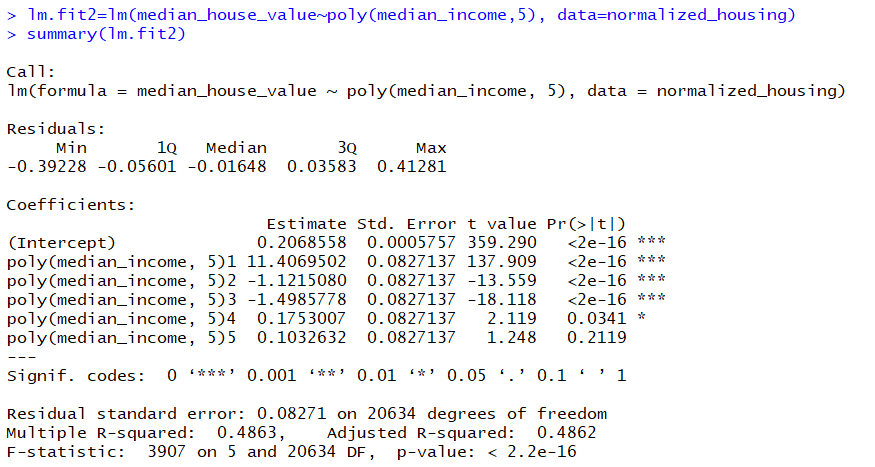
1. median\_house\_value against product of median\_income and population:



1. median\_house\_value against median\_income and median\_income squared:



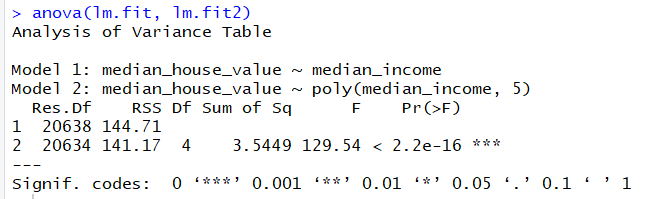
1. median\_house\_value against polynomial of median\_income:



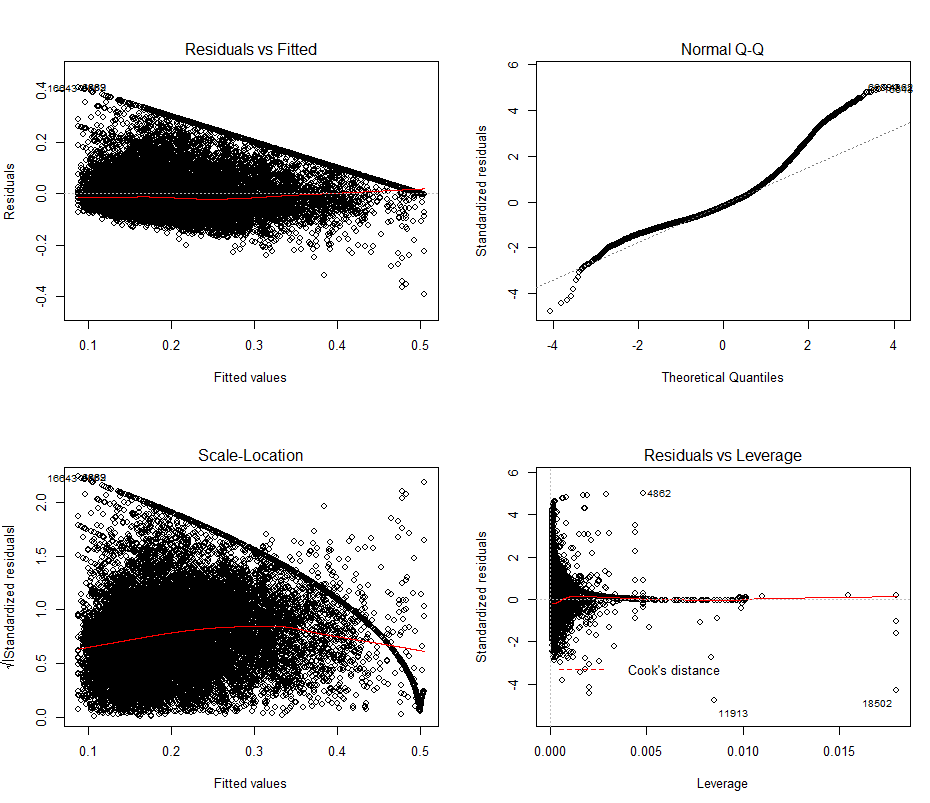
Here, we can observe that median\_income is statistically significant only upto the third degree polynomial.

Anova:

Analysis of variance table.

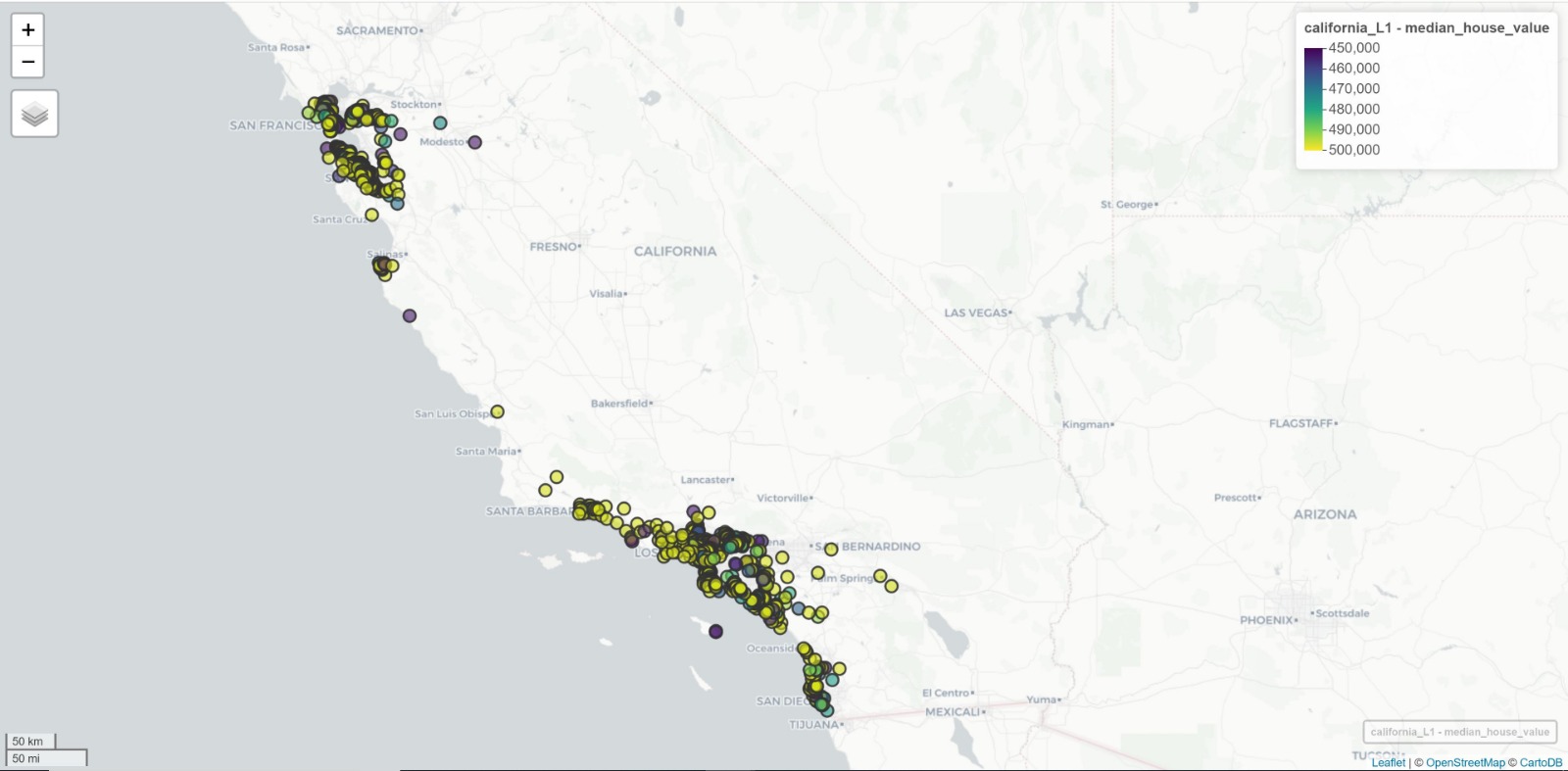


Residuals:



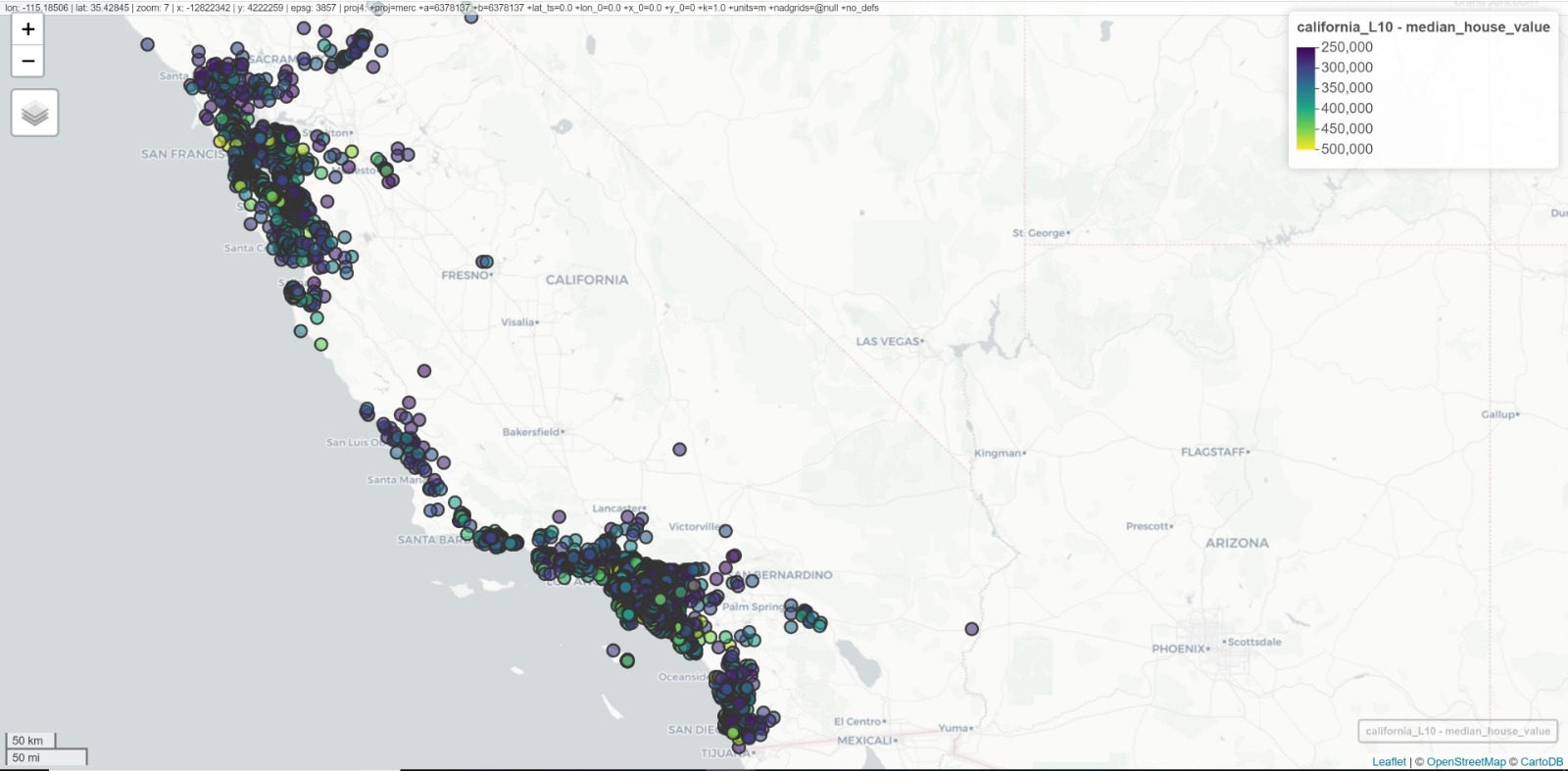
5. Observations:

We tried plotting the data based on various price ranges to see how the data is distributed. This is for all neighbourhoods with median\_house\_value ranging from 450,000 – 500,000.



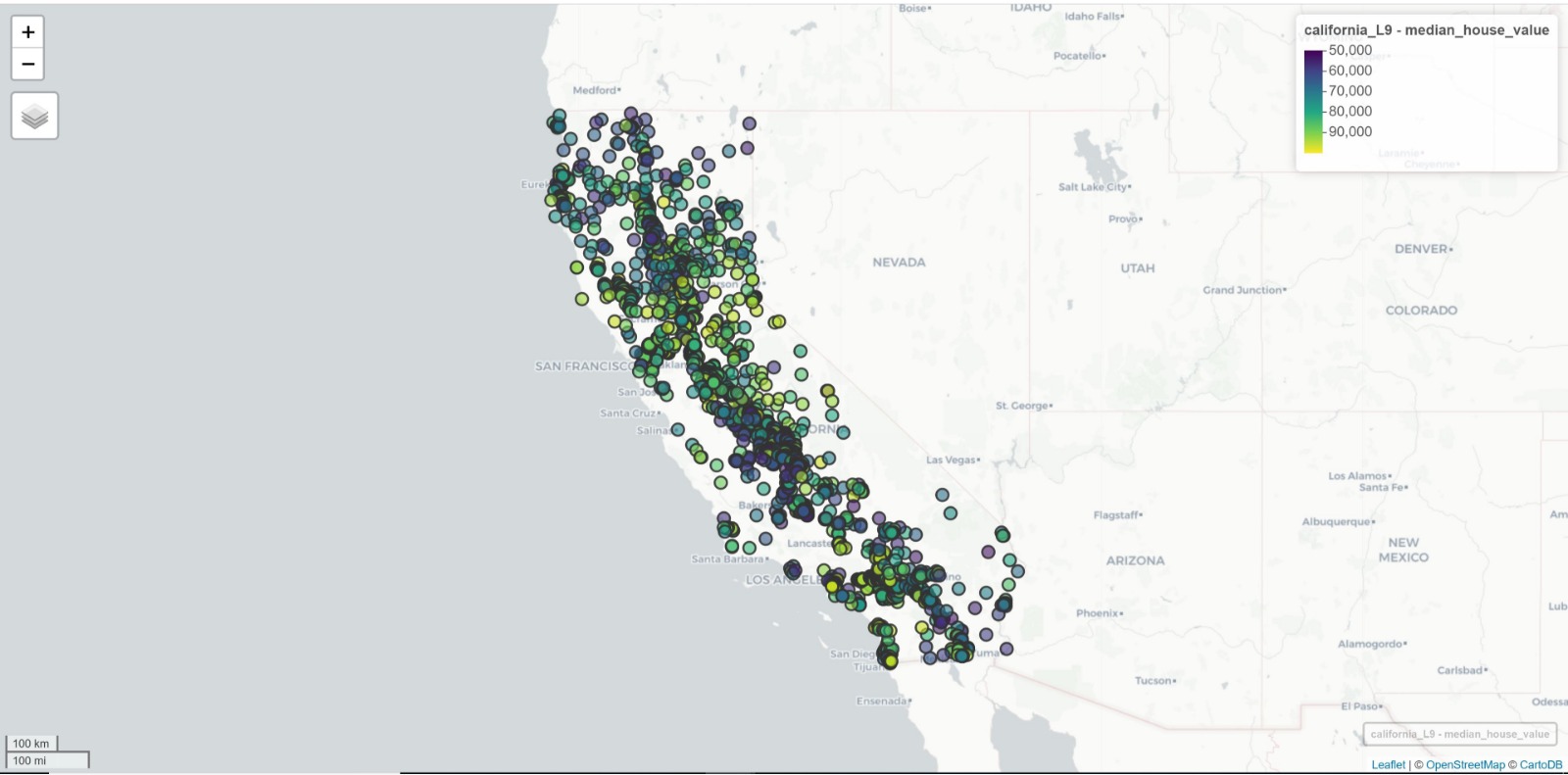
It can be observed that the areas close to the ocean are the most expensive, and there are a very few numbers of such areas.

Now, looking at 250,000 – 500,000 range:



We can see that there are high number of areas in the given range and there are quite a few areas away from the ocean.

Now, looking at 50,000 – 100,000 range:



Most of the areas tend lo lie away from the ocean (i.e. inland).