Dylan Kasanders

MATH 342

Project Draft

**Abstract**

We constructed a linear regression model for predicting the median house value for a particular block in California in 1990. With this, we aim to find a linear relationship between characteristics of a housing block and the median house value within that block. We additionally aim to seek variables that are attributing the most to housing prices within California.

**Introduction**

We are using California housing data pulled from “Sparse spatial autoregressions” by Pace, R. Kelly and Ronald Barry. The housing data they used was retrieved through US Census data. The data consists of characteristics of housing blocks in California in 1990 as stated in the US Census. We used this data in order to create a linear model that predicts median house value for a block. Using the data, we also constructed two variables that gave the distance between the block’s location and two key cities in California: Los Angeles and San Francisco. With this we are aiming to see the extent the distance from these cities can be used to predict the median housing price.

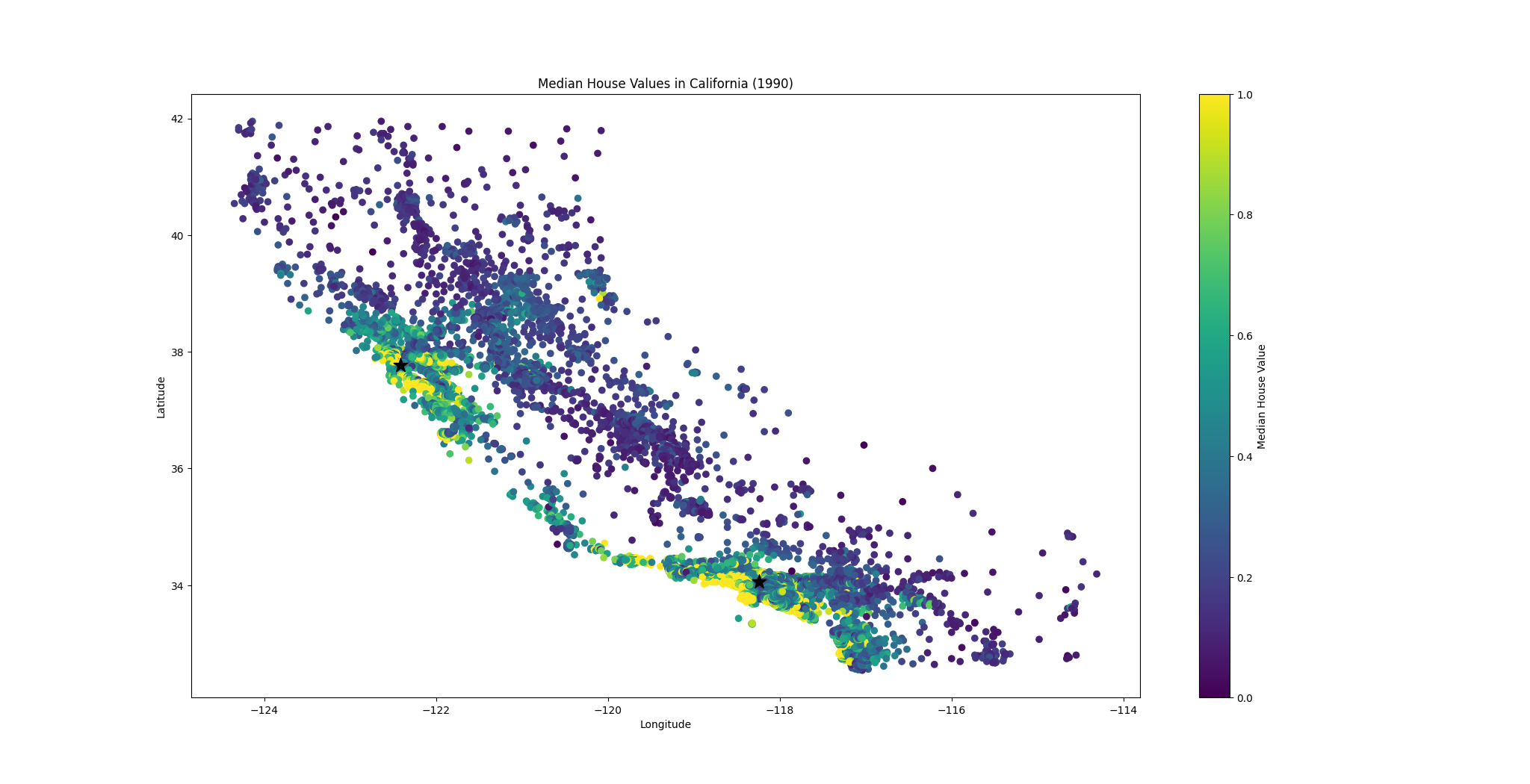
**Methodology**

The linear regression model we have created utilizes the following variables, either taken straight from the dataset or extracted.

* Median Income – Average income of the block
* Distance to Los Angeles – Distance the block has to Los Angeles
* Distance to San Francisco – Distance the block has to San Francisco
* Median Housing Age – Median Age of the houses on the block
* Total Rooms – Total number of rooms in a block
* Total Bedrooms – Total number of bedrooms in a block
* Households – Number of households in a block
* Median Income – Median income of residents within a block

The model we ran on the dataset using these features is as follows:

Looking at a heatmap of housing values in California, we can notice an increase in house values that are near Los Angeles and San Francisco. This increase calls for a parameter that can measure the closeness to these two respective cities.



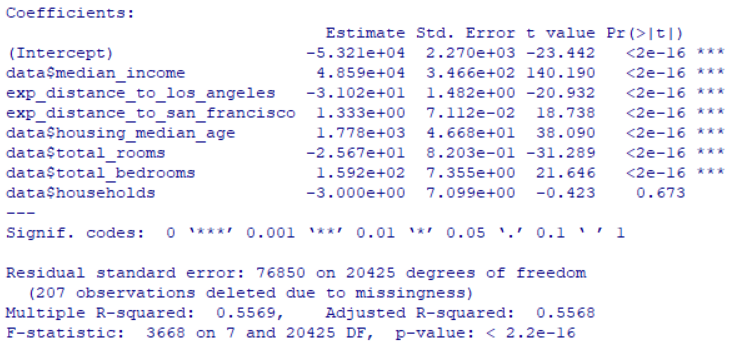
Los Angeles and San Francisco are highlighted.

We found that when initially constructing models only utilizing the distances from Los Angeles and San Francisco, the model that had a exponential relationship between the median house value and the distances to have a greater value compared to a linear relationship (0.27 vs 0.21). This suggests that the impact of house values is better captured with an exponential function rather than a linear one.

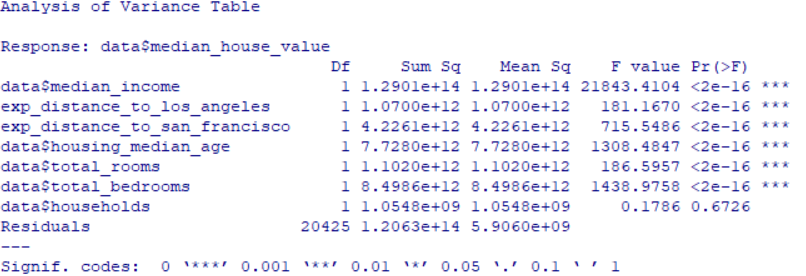
Our model therefore assumes an exponential relationship between the median house value and distances to key cities Los Angeles and San Francisco. Additionally, using variables median income, median housing age, and people per household as linear variables to determine median house value.

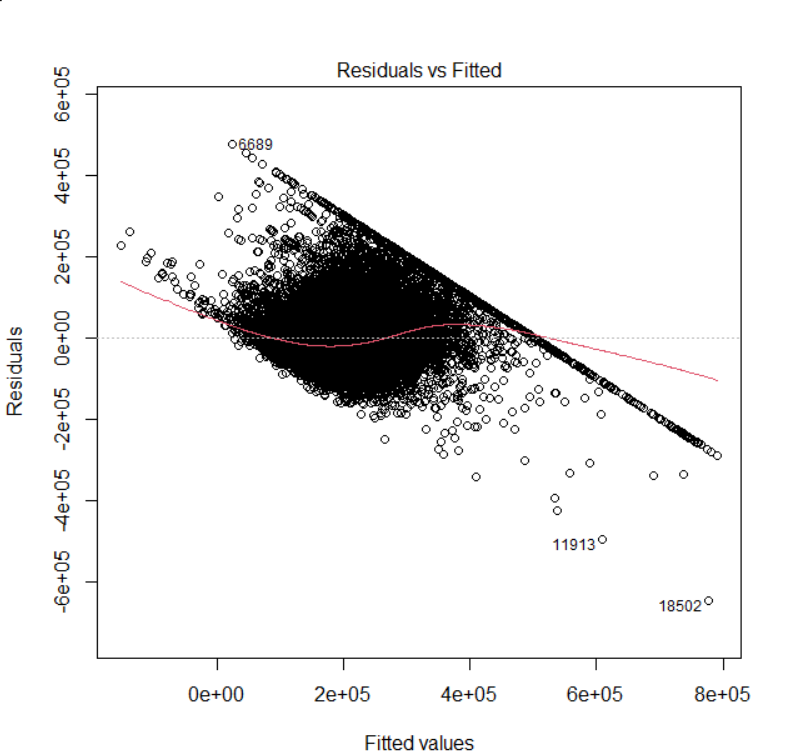
**Results**

With the following parameters, we have achieved the following results:



The model obtained a of 0.5569, having increased from the model that only utilizes the distances from Los Angeles and San Francisco. Additionally, this model has a residual standard error of 76850, while the standard deviation of the median house values in the dataset is 115395.6. This notes room for improvement in adjusting the model, as for practical applications a lower residual standard error may be desired especially in scenarios where confidence intervals of predicted values are desired.





**Conclusion**

We have created a linear model that explains 55.69% of the variability of the median house value in a block with the available data. Utilizing the distance from key cities Los Angeles and San Francisco assisted in further predicting the median house value in a given block. This model has room to improve with future feature extraction with the dataset provided and feature transformation of the chosen variables.