Anthony Medina

Final Report:

San Francisco California Housing Prices

Problem Statement

Where can we afford to live? San Francisco is one of the most expensive places to live now. What about back in the 90s? This project explores the 1990 census data as an introduction to regression modeling.

Data Wrangling

The initial data set had 20640 rows and 9 columns with various features. Even though it was a good sized data set, it needed some love to be usable. Most of the columns were of the wrong data type, even though there was a lot of data, there were quite a few duplicates, currency in the wrong denomination and missing values.

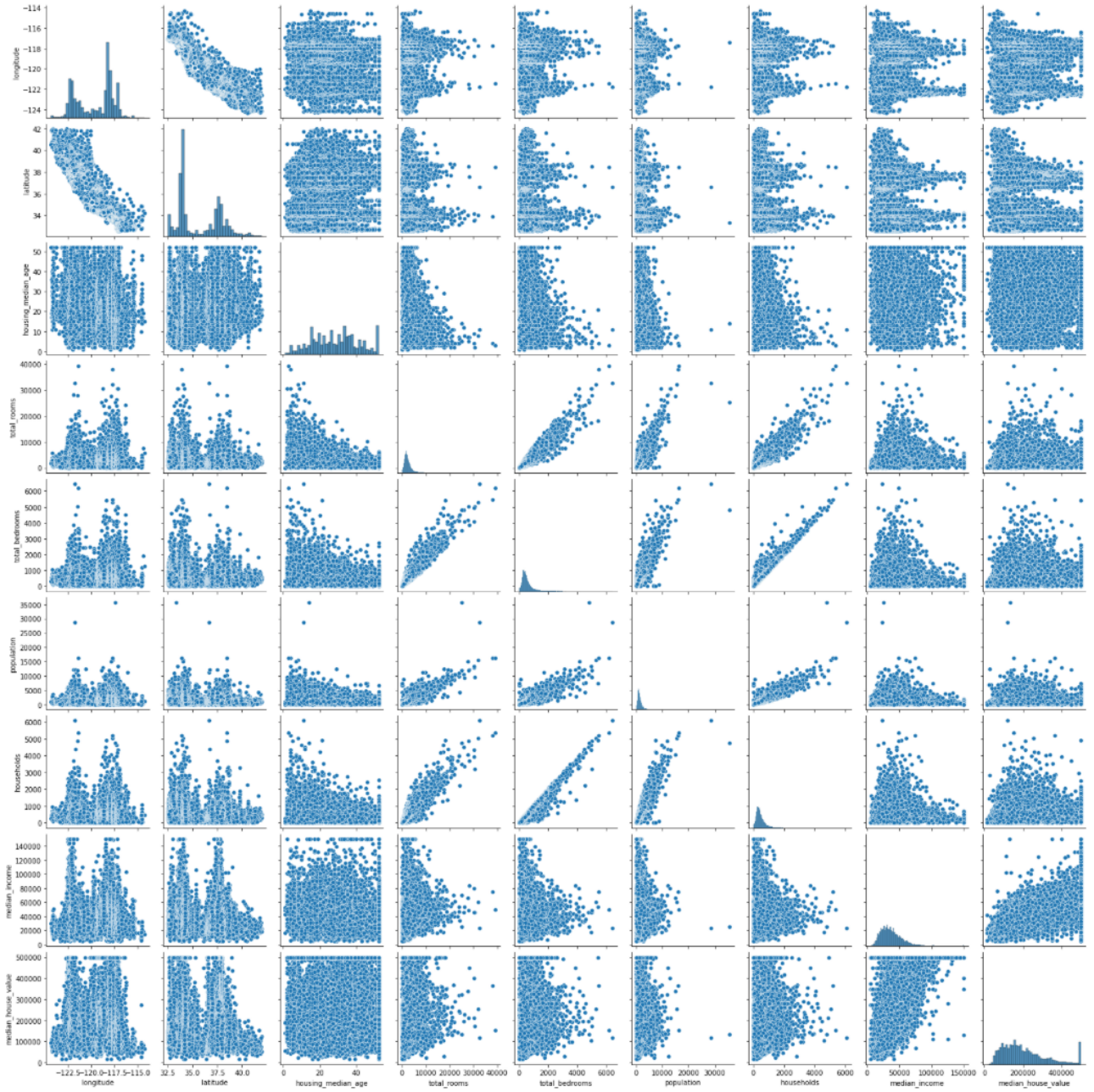
After looking through the data to get a feel for the data, I made note of the data types, and the summary statistics of each variable. Then it was time to deal with missing data. Since the missing data was only found in a single column, I looked to see if there was a reason for the missing data. I explored replacing the data with the mean, median, zero and dropping the data altogether. In the end, I decided to drop the data since it represented a fraction of a percent.

Exploratory Data Analysis

EDA began with deep diving into each feature individually. I collected summary statistics and looked at the range of values. This is also where I looked at the distribution of each variable individually. Some of the variables were skewed. This is

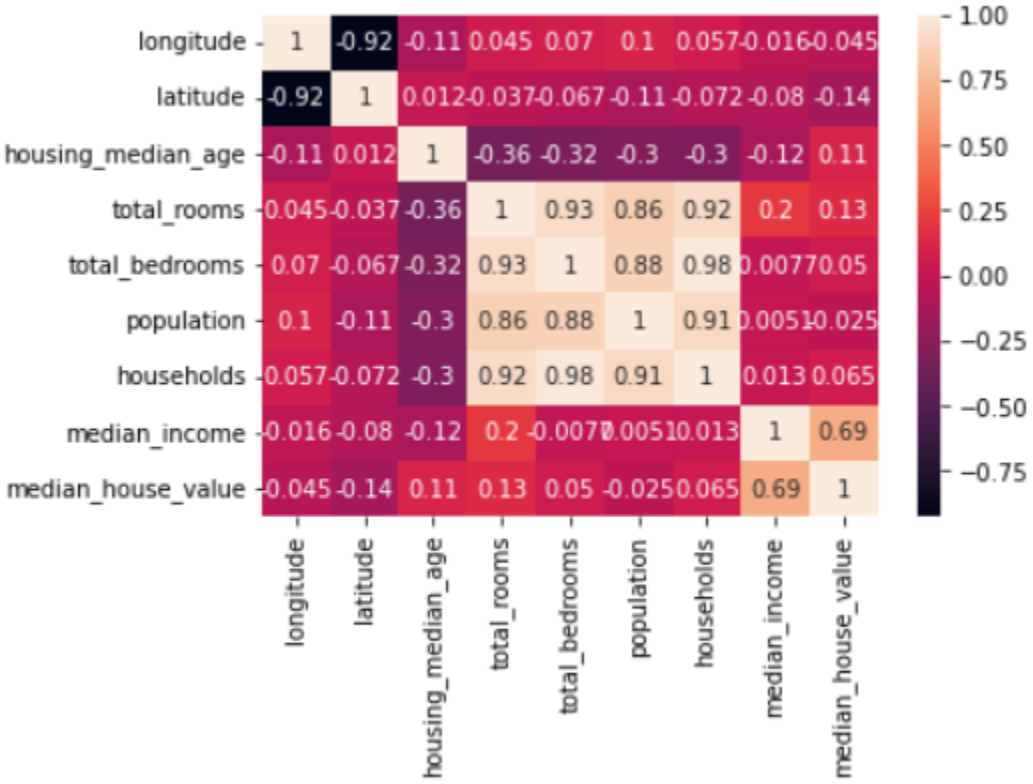
where I discovered that there were only 5 data points that had Island as their location. This will come into play later since island points seem to be really off.

After discovery of variables individually, I started looking at their relationship to the target variable. It started with a scatter plot of all of the variables.

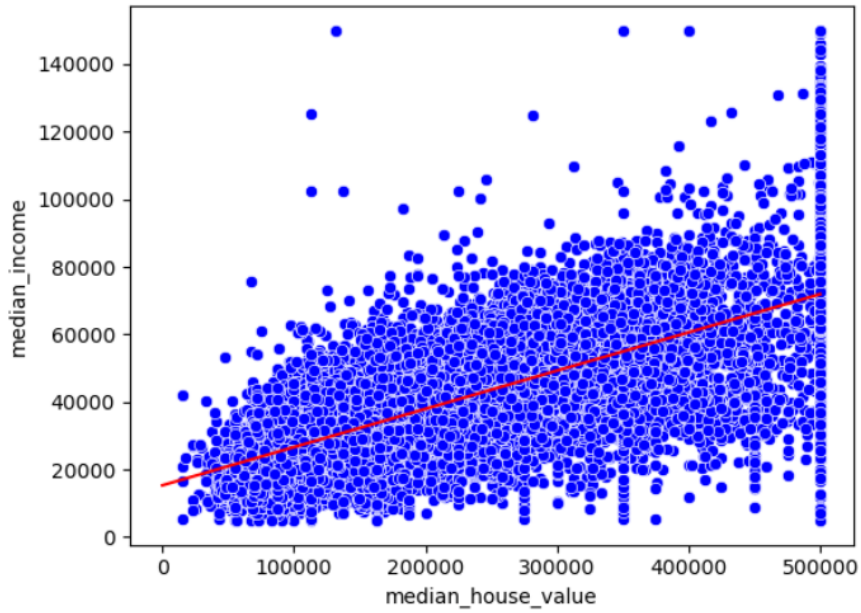


When latitude and longitude variables are put together, it's very clear that we're looking at a data set in california.

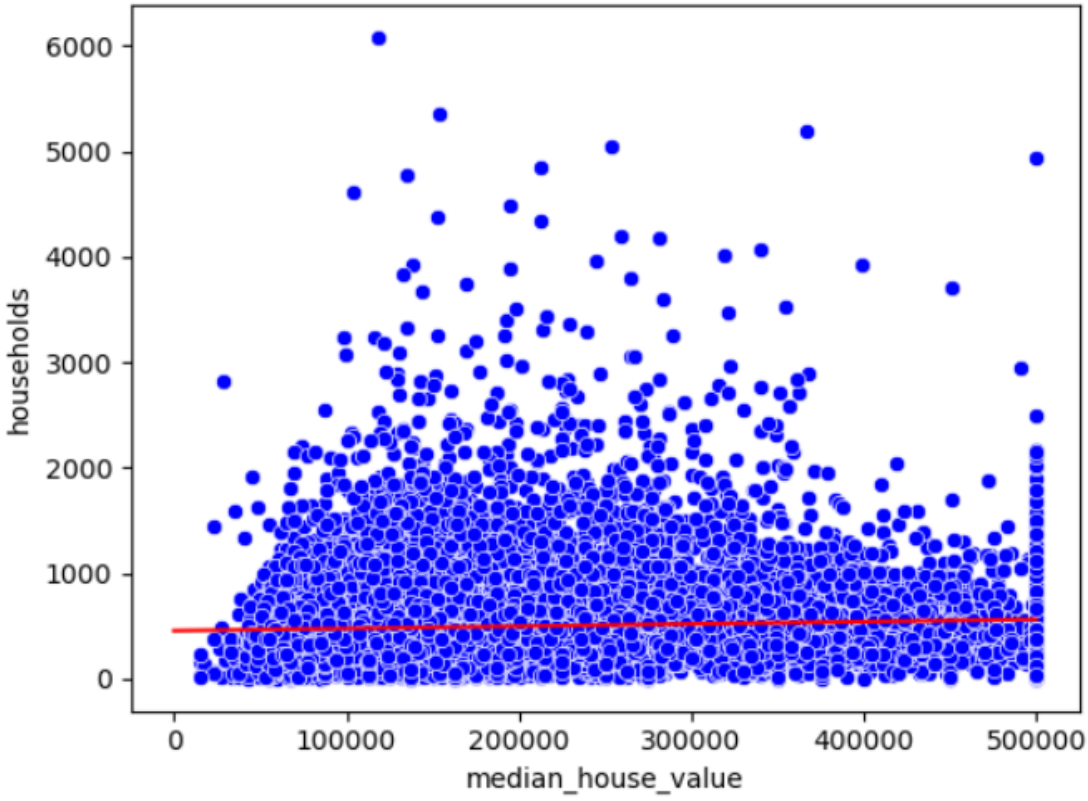
After looking at scatter plots of the and making some notes of variables that might be correlated, I printed out a heatmap of all of the variables.



Confirmation from the heatmap told me to further investigate Median house value with median income and the number of households.



It's kind of difficult, with the number of data points, to see the correlation, but it does slightly slant up.



It's difficult to see the line, but math doesn't lie.

Model Selection

Since this is a regression problem, I started with Linear Regression and Polynomial Regression. After I met with my advisor, he recommended I try a couple more models to see if I could bring the R-squared higher, and the RMSE smaller. This is when I learned that you should only use one type of scaler for the data.

After using the MinMaxScaler on all of the numeric columns and creating dummy variables for the categorical variables, I tried 4 models. I tried Linear Regression, Polynomial Regression, Ridge Regression, and Lasso Regression.

Linear Regression Model:
R-squared: 0.6352576876332579
MSE: 4907868535.531821
RMSE: 70056.18127996859

Lasso Regression Model:
R-squared: 0.6352516232554682
Mean Squared Error (MSE): 4907950136.069877
RMSE: 70056.76367111085

The Lasso and Linear regression models were very similar. I went with the Linear Regression model since it is a less complicated model to use.

Future Research

To test the model, I used the median value of each of the features except for the location. I created an array for each of the locations. I would be curious to do a deeper dive into the Ocean Front property. I would imagine that should be a lot more.

Here are the values that the model predicted for each location.

Inland $417,603.01

Island $426,431.03

Near_Bay $385,457.96

Near_Ocean $326,171.32

Ocean_Front $257,276.63

Since these values are from 1990, I used an inflation calculator to see their value in 2023. The results are not surprising.

Inland $967,883.48

Island $989,773.50

Near_Bay $894,672.40

Near_Ocean $757,064.35

Ocean_Front $597,155.40