Exploration and Modeling of California Housing Data

Brandon Yen

December 7, 2023

Stevens Institute of Technology

ENGR 241 (Fall 2024)

Data Science Project

Professor Jagupilla

Contents

1 Introduction

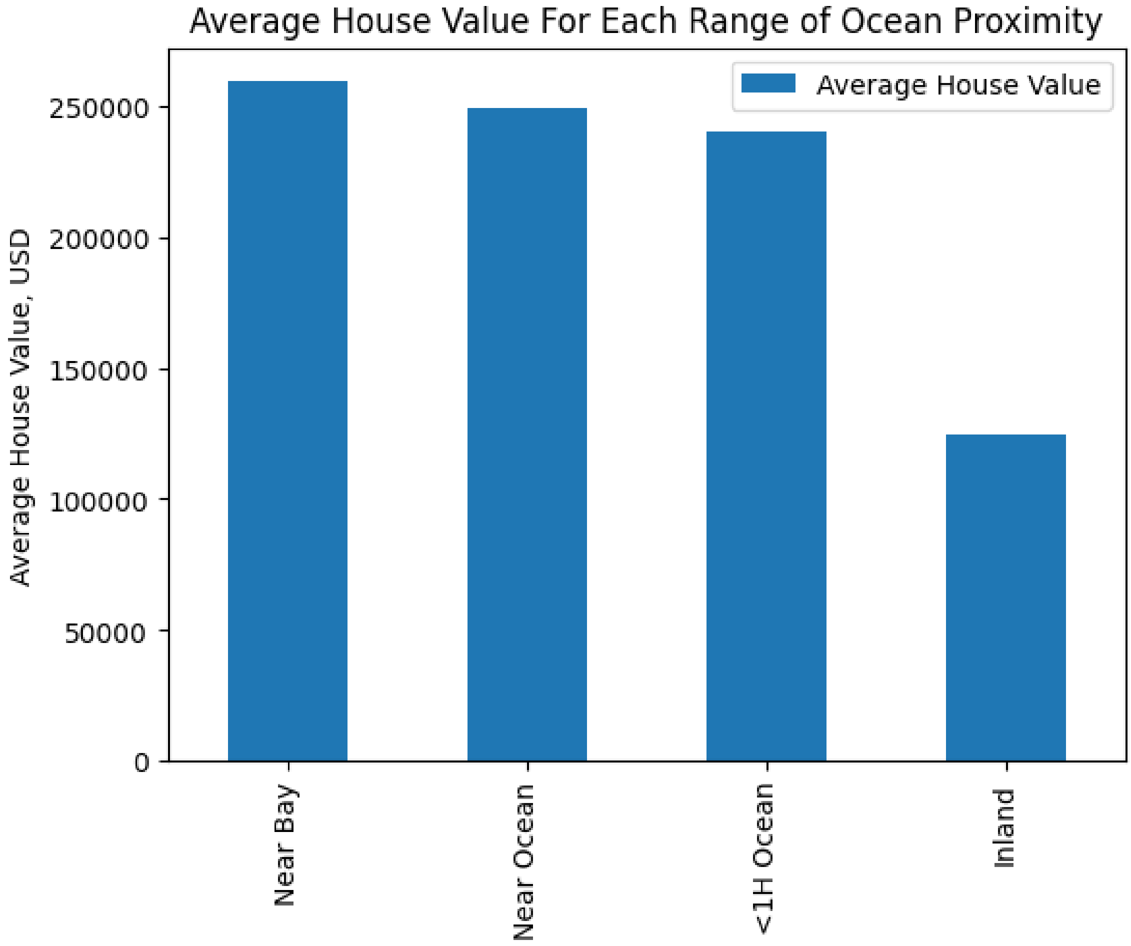
The data set provided contained information from the 1990 California census, and includes variables such as the median house age, the total rooms and bedrooms, the total population and number of households, the median income, the median house value, and the ocean proximity of a given latitude and longitude.

Initially, the data did need to be cleaned, so another variable was created: people per household. This was calculated by simply dividing the population by the number of households. Any entry with a people per household value above 20 and below 1 was removed. Any value below 1 meant that there were less residents than households, which would not be possible for a census. Any value above 20 meant that, on average, a large number of people lived in a household, which would be extremely unlikely. No thresholds were put on the population or number of households, as some locations may have their houses spread far apart (for example, a location that is primarily farmland).

2 Data Analysis

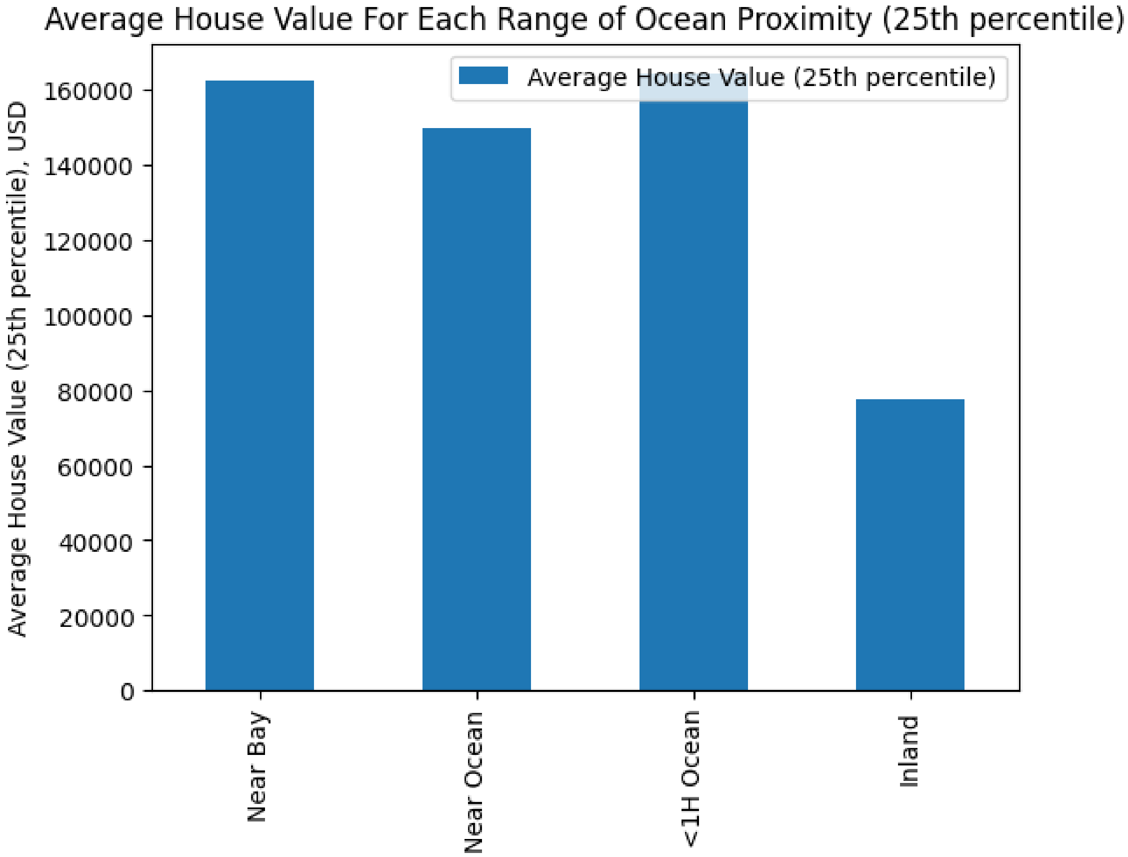
Fortunately, there were no missing or incorrect values for any of the data points. In a case where there was no value reported for a specified column, that row would be deleted to minimize errors when analyzing the data.

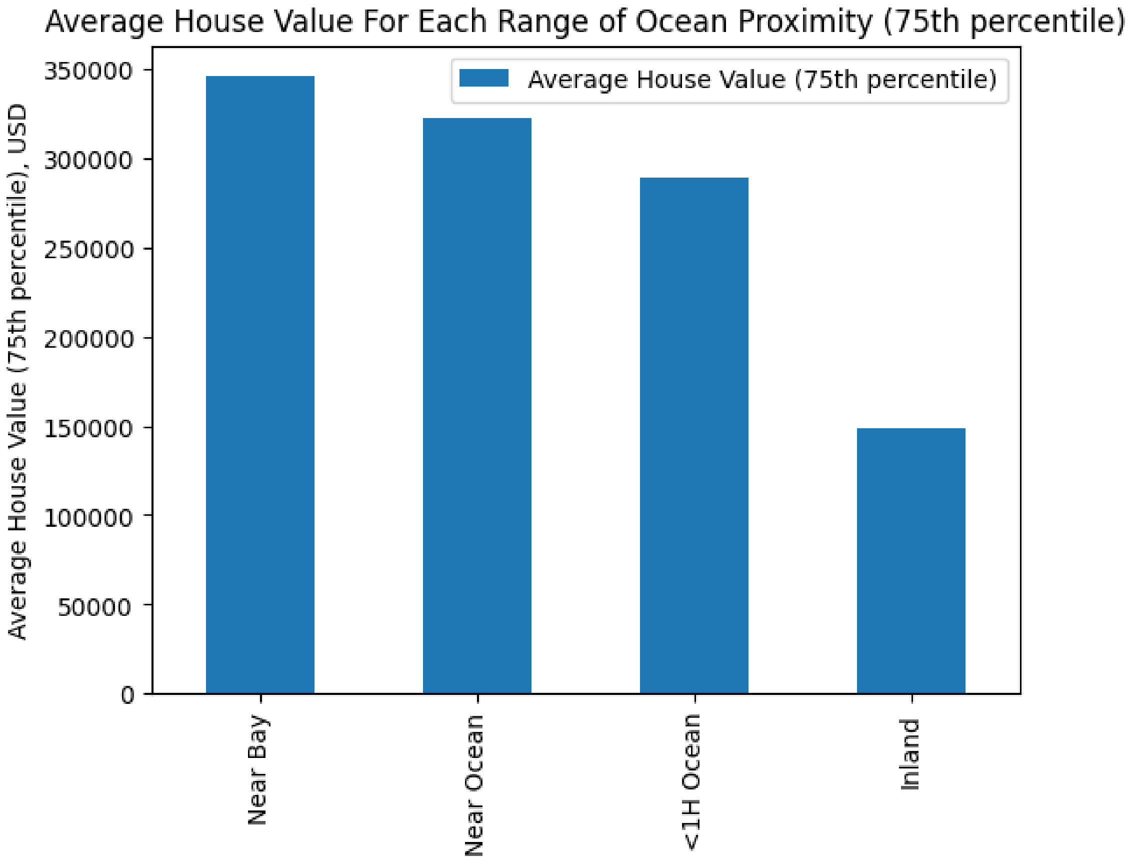
2.1 Ocean Proximity and House Value



Upon initially examining the data, there seemed to be a strong correlation between the ocean proximity of a house and its value. Looking at the average house value for each "range" of ocean proximity, we can see that houses located inland have a significantly lower average house value compared to the other three ranges of ocean proximity. Locations in California's inland tend to be less desirable to live in due to a less favorable climate and fewer job opportunities. California's inland is, on average, hotter than locations closer to the Pacific Ocean, and most of the land is sparse and dry or farmland. Large tech companies heavily dictate California's economy, and these companies tend to gravitate towards larger cities near the bay or ocean, such as the San Francisco bay area and Los Angeles. Housing would therefore become more expensive as the job opportunities in those areas become more lucrative.

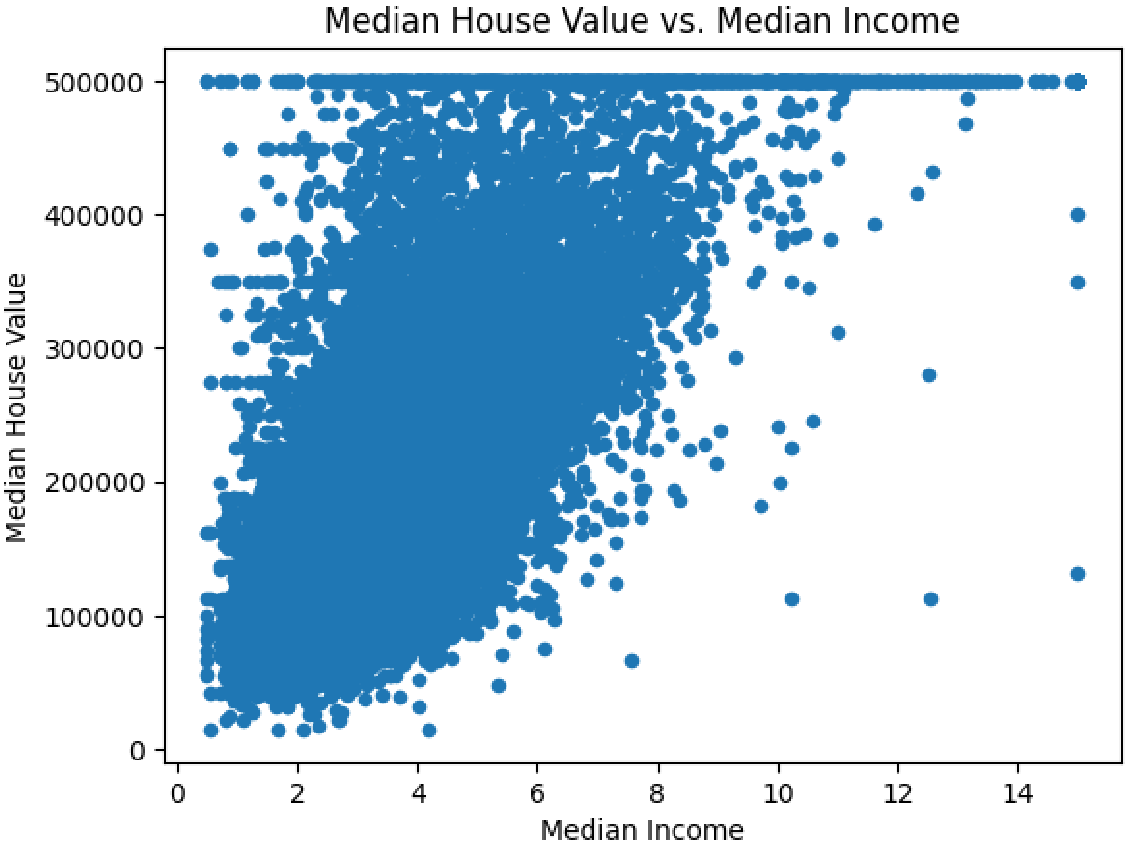
This is further confirmed by looking at more descriptive statistics for this data. Looking at the 25th and 75th percentile data, we can clearly see that in both cases, the average house price is significantly higher for non-inland houses compared to inland houses. Since the 50th percentile is susceptible to data skewing when used alone, having the 25th and 75th percentile data to corroborate these findings is helpful.


Average House Value For Each Range of Ocean Proximity (25th percentile)
Average House Value (25th percentile)
Average House Value (25th percentile), USD
0
20000
40000
60000
80000
100000
120000
140000
160000
Near Bay
Near Ocean
<1H Ocean
Inland



2.2 Median House Value and Median Income

Before trying to find a correlation between the median house value and other variables, I wanted to first look at a data point that would likely have a strong correlation with median house value: median income. This is because generally, people with higher incomes will spend more on a house.



Generally, median income had a positive correlation with median house value, confirming our findings. Running the data through a simple linear regression shows that the p-value for this test is less than 0.001, meaning that the data is very statistically significant and there is a strong correlation between the data. Additionally, the r-squared value is 0.474, so while not perfect, there is evidence that higher median income trends towards higher median house value.

3 Linear Regression Modeling

After confirming the correlation between median house value and median income, I wanted to find a correlation with other variables to a significance level of 1%. Firstly, I looked at the correlation between median house value and the people per household. Generally, those with smaller incomes tend to have bigger families (Nargund G. Declining birth rate in Developed Countries: A radical policy re-think is required. Facts Views Vis Obgyn. 2009;1(3):191-193.), which was confirmed through the data. Given that the absolute value of the t-stat, -35.549 is incredibly large, the p-value for this variable was less than 0.001. With a negative coefficient of -3.176e+04, the people per household had a strong inverse correlation with median house value.

Additionally, another strong indicator of median house value was the median age of houses. The t-stat for this variable was 15.446, and the coefficient was 952. Surprisingly, the median house value had a strong positive correlation with the median age of the house, meaning that newly built houses were actually cheaper than older, existing houses. The lack of housing in areas such as the Bay Area might have driven contractors to build apartments and other forms of cheap housing to accommodate people looking for work in lucrative industries, of which the Bay Area has plenty.

ANOVA Table for Regression

	df	SSR	MSR	F	P(>F)
People Per Household	1	1.577e+13	1.577e+13	1270.84	3.727e-270
Median Age of House	1	2.96e+12	2.96e+12	238.584	1.597e-53
Residual	20624	2.559e+14	1.24e+10		

The above ANOVA table helps us further confirm that our data is correlated. The f-stat for the people per household was extremely large, 1271, and the p-value, or probability of a data point outside this f-stat, is nearly 0, much less than our alpha value of 0.01. The same is true for the median age of the house: the f-stat was 238, and the p-value was nearly 0.

4 Interpretation and Conclusions

California's housing market has been the topic of economists for an extremely long time due to the sharp rise in technological job opportunities. Being from California, I wanted to look further into the census data to see if there were correlations between housing prices and other socioeconomic information.

Firstly, there was a strong correlation between the price of a house and its proximity to the ocean. While houses that were either near the bay, near the ocean, or within an hour of the ocean tended to have similar housing prices in the 25th percentile, 75th percentile, and mean, houses located inland were nearly half the price of all other houses.

Secondly, the positive correlation between median income and median housing price was proved by plotting the data on a scatter plot and performing a simple linear regression. The p-value was less than 0.001, and the r-squared value was 0.474, meaning that the correlation between median income and the median housing price is strong.

Thirdly, the correlation between median house value against the people per household and the median age of houses was proved at a significance level of 1% using the t-stat and ANOVA table. People per household had an inverse correlation with median house value, while the median age of houses had a positive correlation.

Appendices

A Python Code

Python code imported from Google Colab

```
from google.colab import files
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.formula.api as sm
from statsmodels.stats.anova import anova_lm
CAHousing = files.upload()
CAHousing = pd.read_excel(’California Housing Data Cleaned (2).xlsx’)
near_bay_data = CAHousing.loc[CAHousing["ocean_proximity"] == "NEAR BAY", "median_house_value"]
near_ocean_data = CAHousing.loc[CAHousing["ocean_proximity"] == "NEAR OCEAN", "median_house_value"]
ocean_data = CAHousing.loc[CAHousing["ocean_proximity"] == "<1H OCEAN", "median_house_value"]
inland_data = CAHousing.loc[CAHousing["ocean_proximity"] == "INLAND", "median_house_value"]
average_house_value = pd.DataFrame(’Ocean Proximity’:[’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’],’Average House Value’:[near_bay_data.mean(),near_ocean_data.mean(), ocean_data.mean(),inland_data.mean()])
```

```
plt = average_house_value.plot.bar()
plt.set_xticks([0, 1, 2, 3], [’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’])
plt.set_ylabel("Average House Value, USD")
plt.set_title("Average House Value For Each Range of Ocean Proximity")
print(near_bay_data.describe())
print(near_ocean_data.describe())
print(ocean_data.describe())
print(inland_data.describe())
quarter_percentile_data = pd.DataFrame(’Ocean Proximity’:[’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’],’Average House Value (25th percentile)’:[near_bay_data.quantile(0.25), near_ocean_data.quantile(0.25),ocean_data.quantile(0.25),inland_data.quantile(0.25)])
plt = quarter_percentile_data.plot.bar()
plt.set_xticks([0, 1, 2, 3], [’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’])
plt.set_ylabel("Average House Value (25th percentile), USD")
plt.set_title("Average House Value For Each Range of Ocean Proximity (25th percentile)")
three_quarter_percentile_data = pd.DataFrame(’Ocean Proximity’:[’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’],’Average House Value (75th percentile)’:[near_bay_data.quantile(0.75), near_ocean_data.quantile(0.75),ocean_data.quantile(0.75),inland_data.quantile(0.75)])
plt = three_quarter_percentile_data.plot.bar()
plt.set_xticks([0, 1, 2, 3], [’Near Bay’,’Near Ocean’,’<1H Ocean’,’Inland’])
plt.set_ylabel("Average House Value (75th percentile), USD")
plt.set_title("Average House Value For Each Range of Ocean Proximity (75th percentile)")
```

```
scatter_plot_data = pd.DataFrame('Median Income':CAHousing["median_income"],'Median House Value':CAHousing["median_house_value"])
splt = scatter_plot_data.plot.scatter(x=0,y=1)
splt.set_title("Median House Value vs. Median Income")
SLR=sm.ols(formula = 'CAHousing.median_house_value  CAHousing.median_income', data = CAHousing).fit()
SLR.summary()
MLR=sm.ols(formula = 'CAHousing.median_house_value  CAHousing.people_per_household + CAHousing.housing_median_age', data = CAHousing).fit()
MLR.summary()
anova_lm(MLR)
```