Housing Prices in California

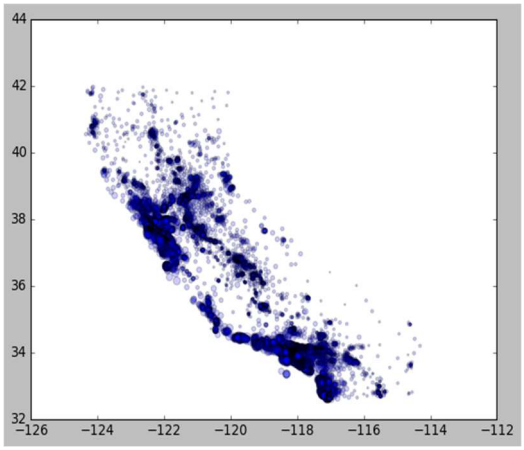
By: Scott Johnson, Pranav Kutty, Shanrui Huang, Zhaoyi Yang, Xiaoyao Li

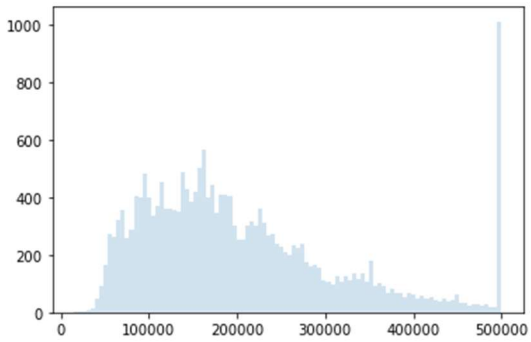
Description of Project:

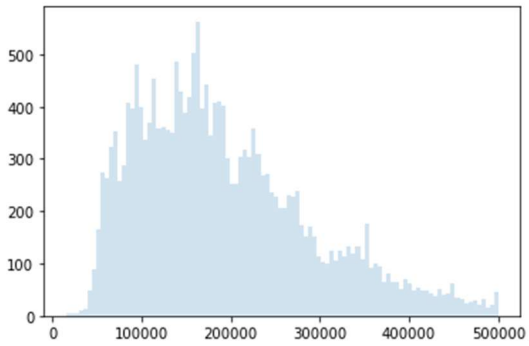
Processing and Manipulation of Data in Python:

Before beginning the manipulation and processing of the data, first it had to be imported into python. This was done using the pd.read_csv() function to open the file of California housing data, allowing us to use the data in a table in python.

One of the questions we had was whether location had an impact on the median house price. We explored this first by graphing it geographically. This was done using the plt.scatter(data,x,y,s,alpha) function. Using the longitudinal and latitudinal data supplied in the housing prices data set. A scale factor was also implemented to scale the size of the circle off each location relative to the median house price multiplied by 0.0001. This allowed us to visually see where the median house price was low and high geographically.

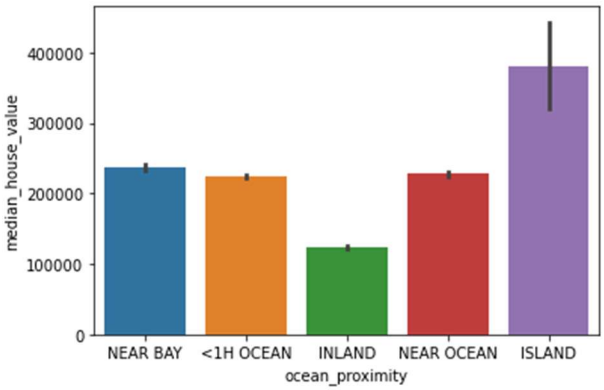




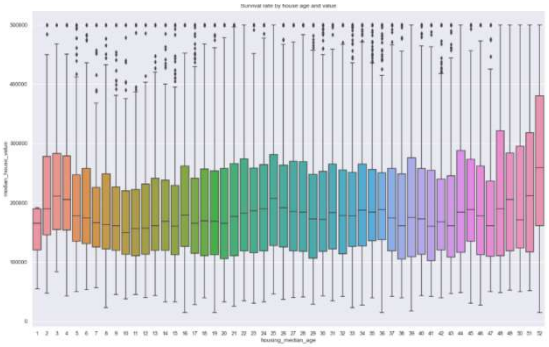


We also graphed the amount of median house prices in each price range with the range being every $5000. This was graphed using the plt.hist(data, histtype,bins,alpha) function to graph it in the form of a histogram to make it easy to analysis visually. It was found to have a great deal of houses in the 496000-500001 price range. After investigating the data, it was found that there were a little over 1000 locations with a median house price of 500001. We assumed from this figure that any median house price over the price of 500001 was rounded down to this value, thus we treated them as an outlier, removing them from the data by making an altered data frame removing any median house prices above the value of 500000. This removed to the major discrepancy in the data.

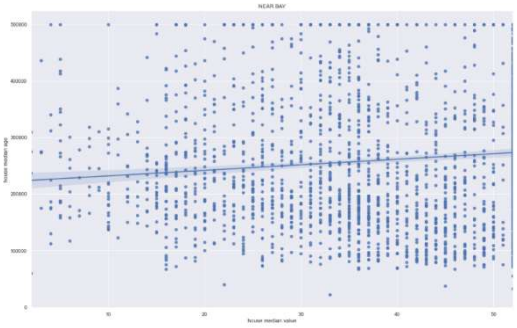
The next bit of data manipulation we performed was using a bar plot to plot the average median house price of each location's proximity to the ocean. This was done using the sns.barplot(data,x,y) function. This represents the average median house price of each ocean proximity.

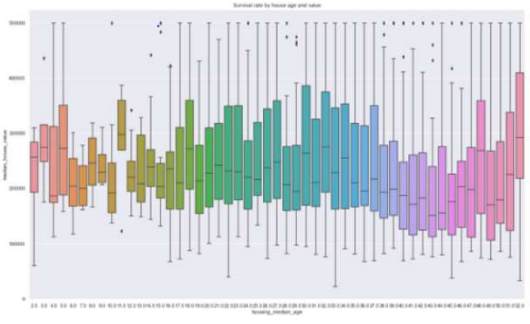


Another question we had was whether the age of the house had an effect on the price We want to make a box plot of all the houses in California with their houses ages and value data. So, we use sns.boxplot(data,x,y) function, at this time, the data will be housing data which has been definite above and x and y will be the columns with housing_median_age data and median_house_value data. plt. title is for adding a title to the graph we make. As there are lots of value and the graph is too small to show it clearly, we use sns.set() function to set the graph in a better size.



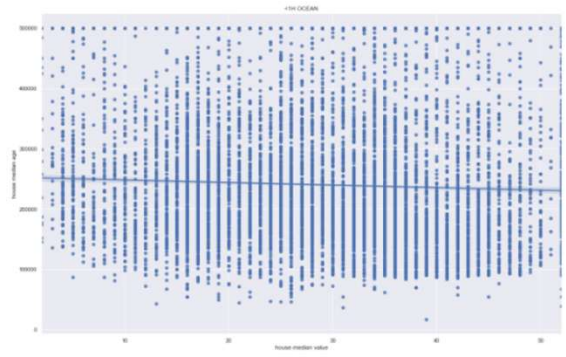
We have already had a look at the overall relationship between all the house value and house age, now we will analysis the data separately to see in different area which is different ocean proximity, what is the relationship between house ages and value. To find out if different area will affect the house value. Because we need look at these data separately, so for each part we need drop the rows we don't need, which is not the areas we analysis on at this moment. We need clear the data in housing_median_age and median_house_value first to make sure only the data of area we analysing is showed, so. mask() function is be used to drop value which is not satisfied with the condition at the moment. We need a graph of regplot first, to see the slope of gradient first, it will tell use in this area, as the house become older, the value of the houses will trend up or down. After that we need put the clear value as an array so it can be graph by sns.regplot function. For each point on the graph is a house with corresponding house age and value. plt.xlabel and plt.ylabel for adding x and y label to the graph. We also representd it in the form of a

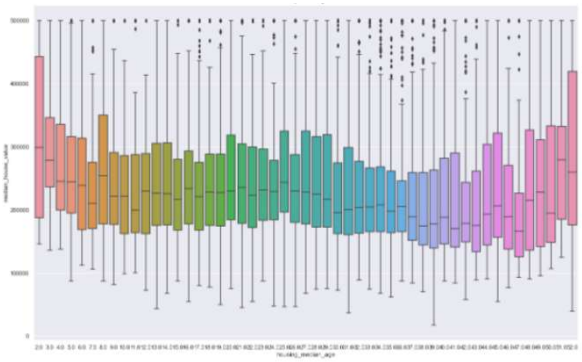


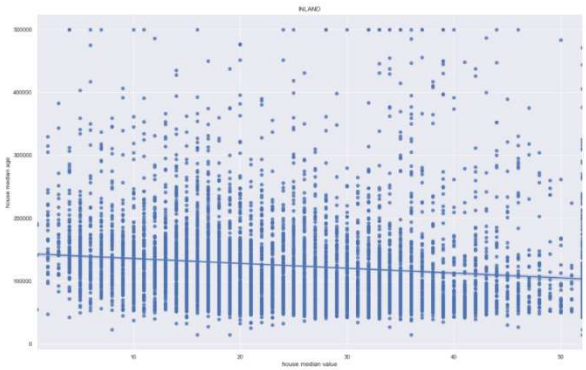


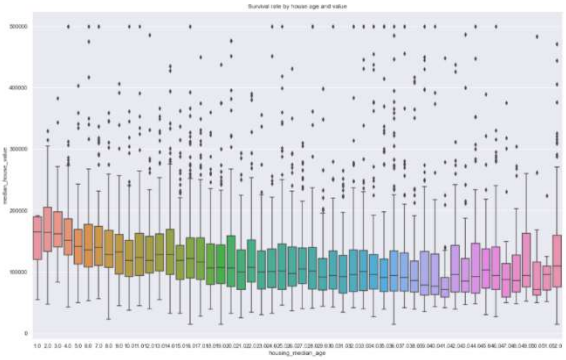
boxplot, the same way as the other one was done. The box graph will show the distribution proportion of houses values in different ages. Therefore, we can determine the percentage of houses ages with corresponding values in the total houses amount.

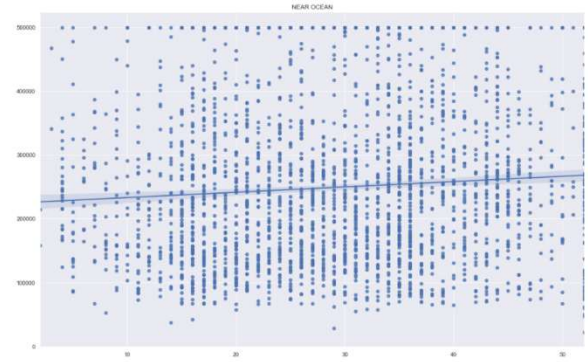
We also graphed the rest of the ocean proximities relationship graph and boxplot.

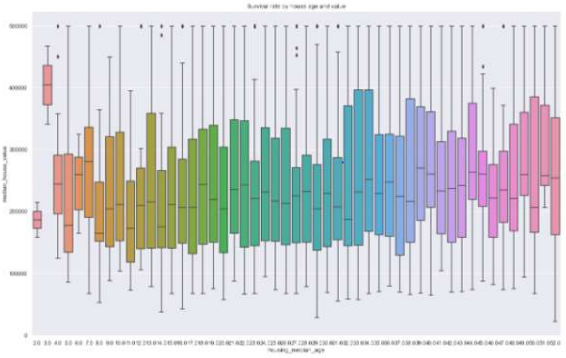


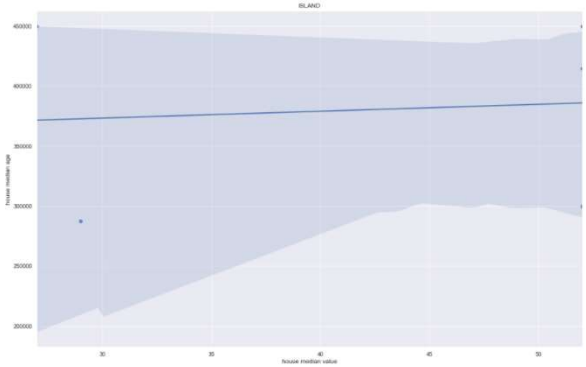


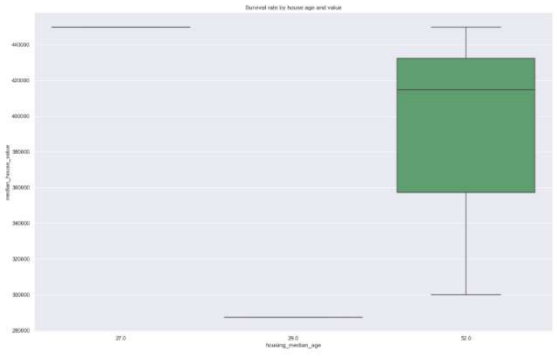




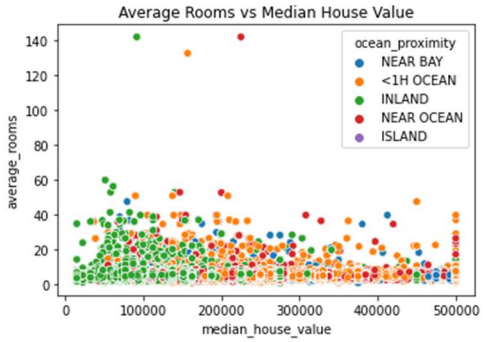


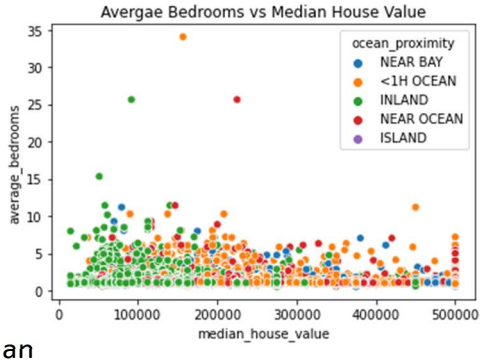




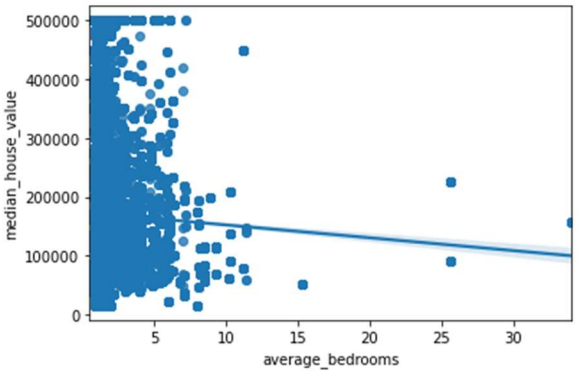


The next question we had was, does the average amount of bedrooms and rooms per house have an effect on the price. The first thing we need to do is to process the raw data because total_bedrooms represents the sum of the bedrooms of a certain number of households. So, we created an average bedrooms value by dividing the total bedrooms by the amount of houses. This was then put into a data frame and merged with the original data. The same was done with total rooms, creating a average rooms value. After obtaining the data, it was then put into a scatter plot using the sns.scatterplot(data,x,y,hue) function, graphing the average bedrooms vs median house price and average rooms vs median house price. The graph was also coloured into categories of ocean proximity to further explore whether location has an effect on the number of rooms, perhaps having an effect on price.

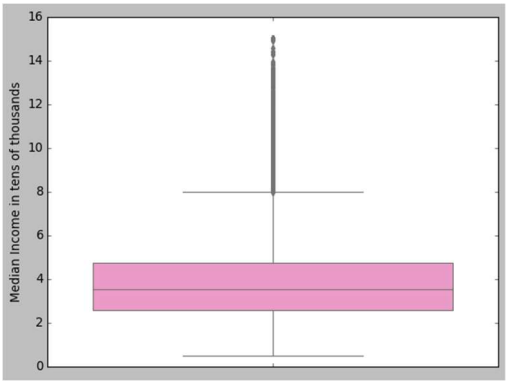




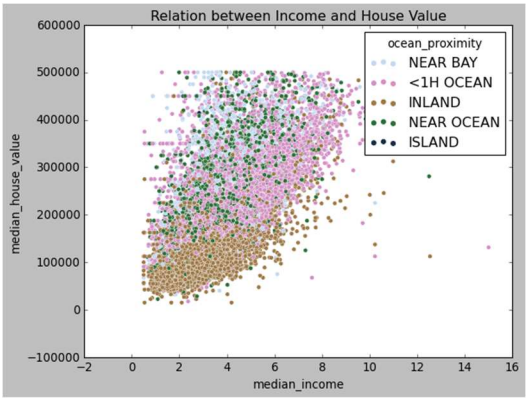
We also graphed the average amount of bedrooms relationship to median house price to see whether more bedrooms increases or decreases price.



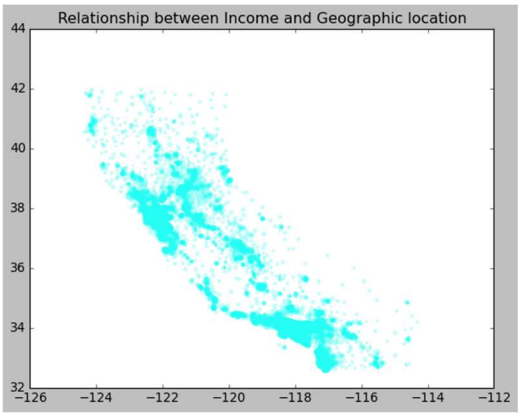
The final question that we wanted to look into was how median income in each area affected median house value, as well as other variables such as ocean proximity and geographic location. The initial step we took when analysing this variable was to create a simple box plot using sns.boxplot() to have a visual understanding of where the median and outliers are approximately. The median lies between $30,000 and $40,000 and the outliers are found approximately above $80,000. We decided not to remove the outliers for this variable, as it gives us an illustration of some of the highest earners in California and how they spend their money on houses.



After looking at the box plot and understanding the distribution of the median income variable, we decided we'd use a scatter plot to observe the correlation in order to come to some conclusions for this question. The scatter plot was from seaborn (sns.scatterplot()), and featured income and house value on the axes, with color coding based on ocean proximity. Just from looking at the graph we can see some correlations. For example, the positive correlation between the axes indicate that as the median income of an area increases, the median house value also increases. This suggests that those who earn less would be less willing to spend more money on houses and those who are more afluent in terms of earnings seem to be more likely to buy more expensive homes. From looking at the ocean proximity variable, we can also come to a few conclusions. We can see that for both lower house value and lower income, the houses tend to be inland, whereas areas with higher median incomes tend to be near the bay, near the ocean or within an hour of the ocean. Interestingly, the previously mentioned outliers for the median income seem to have varying house values, but they all seem to be near the bay or ocean.



For the next graph we wanted to look at the median income variable and look at how it changes based on geographic location. To do this we used Matplotlib's scatterplot feature, plt.scatter, putting the longitude and latitude variables on the axes and differentiating income by the scale of each point. From this graph we can reach conclusions on how geographic location affects median income. One of the conclusions we can clearly see is that most of the larger median incomes are located along the western border. This is likely due to the western coast of California being along the Pacific Ocean, meaning houses will be located near beaches and get coastal views. This will likely attract buyers with larger median incomes. Another conclusion we can come to is that the largest points are located in and around major cities, such as Los Angeles, San Jose and San Diego.



Conclusions:

At the beginning of our project, we set out to learn about what effect different variables had on housing prices in the state of California in the year 1990. In particular, we wanted to know what impact location, number of rooms or bedrooms, age and income had on the price of houses in this particular Census. Through data manipulation using Python and its libraries, we were able to come to conclusions for the questions we posed.

Firstly, we have found that there is a correlation between location and median house price. As seen in the data, more expensive houses tend to be around major cities and along the coast, whereas houses further inland tend to be cheaper. Californians would rather spend more money to live near the shoreline and bustling cities and would live rather live inland if they wished to save their money.

In contrast, the correlation between the number of bedrooms or rooms and the house value is much weaker. Based on the data we analysed, the number of rooms and bedrooms in an area doesn't seem to have had an impact on the median house value of the given area. A possible reason for this may be that most people would rather spend more money on larger plots with bigger rooms rather than many smaller rooms.

With the relationship between age and house value, there are a few interesting correlations to look at. For example, houses near the bay and near the ocean tend to get more expensive as they get older, whereas houses within 1 hour of the ocean and houses inland tend to be more expensive if they're newer. A conclusion we can come to about this is that individuals that want to stay further inland would want to spend more on newer houses rather than old run-down houses. In contrast, individuals that want to stay near the bay or ocean may want to spend premium on more vintage style homes rather than more modern homes.

Finally, we can see a strong correlation between income and house value, namely, areas with higher median incomes tend to have higher median house values. Areas with higher house values within an hour of the ocean also tend to have higher median incomes than areas with high median house values near the bay. A simple conclusion we can come to for this is that those with more money would likely be more willing to spend more money on their place of residence, and those with lower rate of earning would be more likely to be more conservative with their money in terms of house purchase.

Overall, the California Housing dataset has been enlightening to work on, and we've been able to use it to find the answers to the questions we proposed to ourselves at the beginning of the project.