Shannon Dulaney

ECON-494-01

Dr. Levkoff

October 28, 2020

Descriptive Analytics Project: CA Housing Price

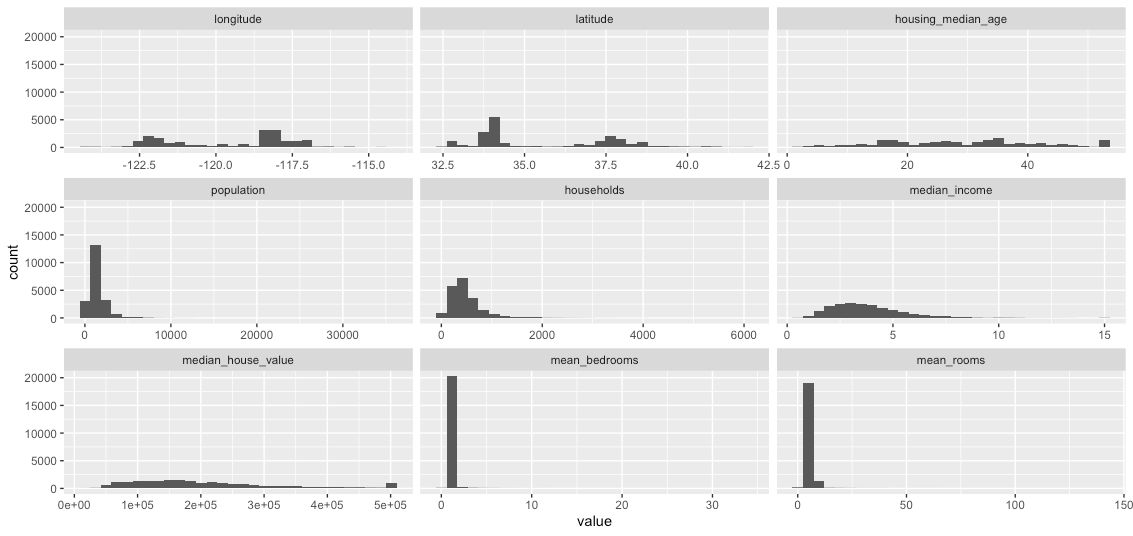
Executive summary

The objective of this report is to describe how various housing factors affect housing prices in California. The data was collected from the 1990 California census report and contains nine numerical variables and one categorical variable. The numerical variables include longitude and latitude values which measure the distance west and north, respectively. The remaining variables are measured as within a block of the longitude and latitude coordinates and those are as follows: housing median age, total rooms, total bedrooms, population, households (defined as a group of people residing within a home unit), median income (measured in tens of thousands of USD), and median house value (measured in USD). The categorical variable of “Ocean Proximity” contains five categories: <1H Ocean, Inland, Near Ocean, Near Bay, Island. The overall structure of the dataset includes 20,640 observations and 10 variables.

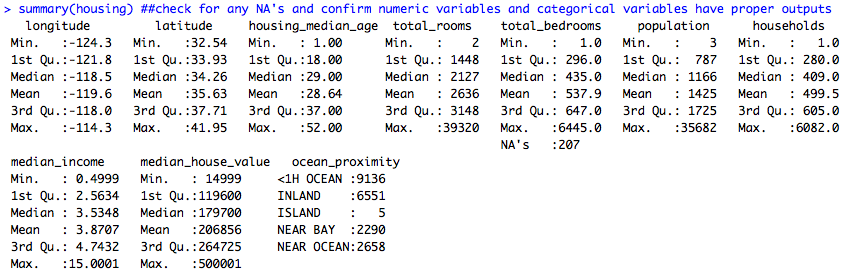
Preprocessing and Data Cleaning

The first step in my preprocessing was to conduct an initial exploratory analysis of the dataset utilizing the ggplot application (Figure 1). I utilized the head, tail, and summary commands to view the overall structure of the dataset. I also used the levels function to view the classes of the categorical variable “Ocean Proximity”. Once I got a high-level overview of my dataset, I analyzed the summary output to check for any NA’s and verify the numeric and categorical variables have the proper outputs. While the variable types had proper outputs, there were 207 NA’s identified in the “total\_bedroom” variable (Figure 1). There is also a large variance between the Max values for each variable, which is to be expected based on the difference between types of variables described. For example, the housing\_median\_age and median\_housing\_value variables. I also utilized the summary data to investigate further the structure of the dataset, specifically the average of key variables of interest, to determine if the population reflects expected housing statistics of 1990. The average house value of this dataset is $206,856, the average house age is 28.64 years, and the average household income is $38,707. These averages are within the realm of possibilities for the housing market of 1990 in California and we can conclude that the population is acceptable for further analysis. Lastly, I generated a ggplot of the ocean proximity categorical variables to determine where the population is mainly distributed (Figure 2). From this visual we can see that the majority of distributions are located <1H Ocean, followed by Inland, Near Ocean, Near Bay, and Island. After conducting a Jarque-Bera test for normality, we can conclude that the median house value is not normally distributed. This is also reflected in the histogram of the median house value as seen below in Figure 3.

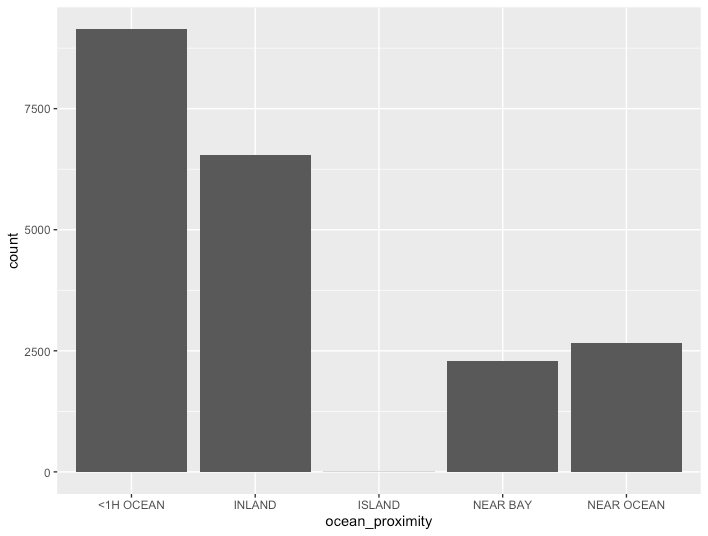
**Figure 1**



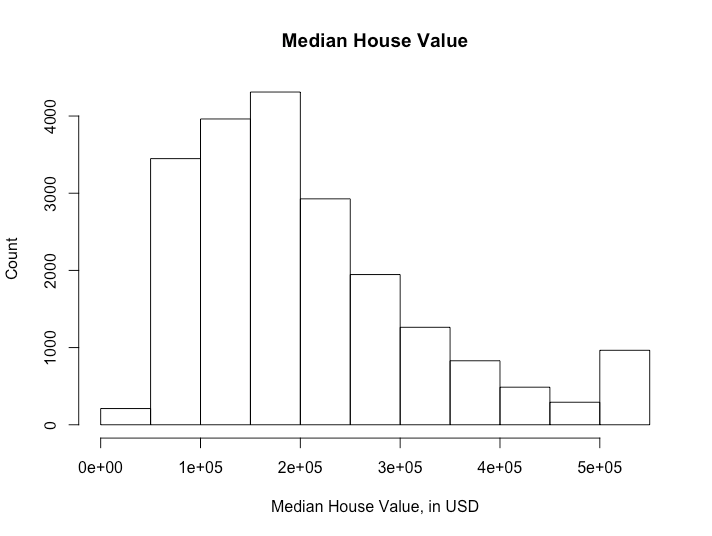
**Figure 2**



**Figure 3**

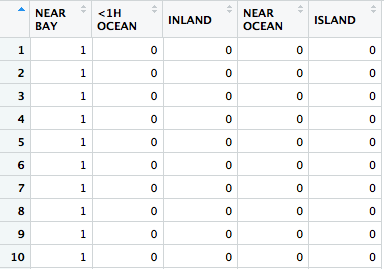
****

**Figure 4**



After the initial analysis, I then started cleaning the raw dataset based on my findings to prepare for the exploratory analysis. I first recoded the total\_bedrooms variable which contained NA’s to compute the median of the variable using an is.na command. Next, I transformed the mean\_bedrooms and mean\_rooms by dividing the total\_bedrooms by households and total\_rooms by households. The average number of bedrooms and rooms is a better descriptive method for this analysis as the data for each input is based on a housing block and there is a chance of a wide variety of houses in a given block, for example, an apartment building on a block of single houses. I then renamed the column headers of total rooms to mean rooms and total bedrooms to mean bedrooms by creating a vector that would drop the old header names to reflect this change. The next step in the cleaning process is to turn the categorical variables into booleans using the for command. To do this, I separated the levels of the variable into “categories” and populated a separate data frame matrix with just the categorical variables with each level as its own column, and zero as their value initially. I then implemented a for loop command to this matrix that would populate the data with “1” to their corresponding category. After the matrix had been populated, I ran a drop command to remove the original column from the data frame so only the names at the top of the column remain (Figure 4). At this stage, I now address the issue of different scaling in the numerical variables and proceed to rescale using the scale command. After rescaling is completed, I then used the cbind command to combine the categorical matrix generated prior, to the rescaled numerical variables to create my final cleaned data set. This data set is labeled “housing\_cleaned” and now contains rescaled numerical variables where the coefficients are given equal weight and a binary output for the categorical variables.

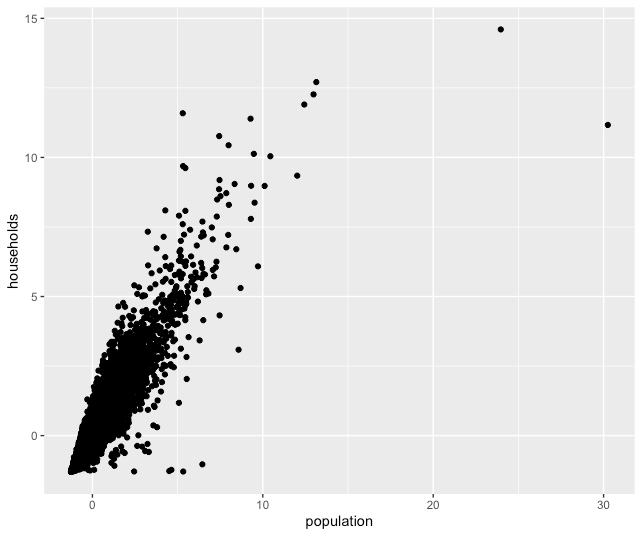
**Figure 5**



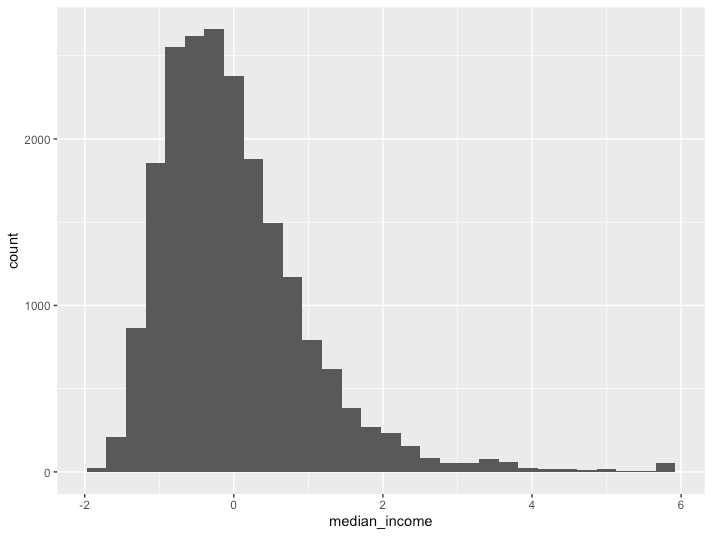
Exploratory Analysis

For my analysis, I first wanted to generate a plot to check the correlation between population and number of households with population on the X-axis and households on the Y-axis which is shown in figure 5. This plot correctly shows that as population increases, the number of households increases. Secondly, I wanted to check the distribution of median income and it appears that the distribution is slightly skewed with a few outliers (Figure 6). There is also a correlation between the age of the home and a subsequent decrease in value, which is also expected.

**Figure 6**



**Figure 7**

****

Conclusion

Overall, this dataset appears to accurately reflect what we would expect housing data to be in California from the 1990’s census. The raw dataset was fairly descriptive to begin with and we were able to preform several descriptive analyses leveraging the ggplot program and built in R commands such as summary statistics. A few cleaning processes were undertaken, the major change being converting the categorical variable into a binary matrix and rescaling the numerical variables. This allowed us to use a clean data set to preform an exploratory descriptive analysis. From this analysis, we determined that the variables are correlated appropriately. Ggplot is a highly complex system that has a several useful applications, however I found it to be very challenging to use. Future analysis could benefit from a different cleaning approach to ease in the use of the ggplots. There is also room for improvement in optimizing the full range of ggplot to create more dynamic visualizations.