My statistical questions from the dataset on housing were: What variables drive the price of a house in the housing market? What features of a house drive its price? Do certain locations of houses attract a certain type of buyer? (i.e. income) and Do certain home features attract a certain type of buyer?

While doing my descriptive statistics, I thought that it was best not to include the longitude, latitude, and ocean proximity variables. This is because the data seemed like it was captured using coordinates, i.e. longitude and latitude, so there are no repeats. The mean and variance of these two variables would not be significant because they are treated more like identifiers than values. Ocean proximity is a qualitative variable, so the mean, variance and mode would not be applicable. The mode of all of the variables can be depicted in the histograms as the tallest bar.

I found it very difficult to conduct a PMF in my analysis because all of the variables are numeric and measured within a block, not individual houses or households. That being said, it was very hard to find a scenario to compare within a variable, so I chose to find the PMF of the median age of houses within a block.

I chose to find the CDF of the median housing age and the exponential CDF of the median house value, which was my dependent variable. The CDF of the median housing age seemed to have a normal distribution, meaning that the CDF was a straight line with a consistent positive slope. This would indicate that there are the same number of new houses on the market (new to 25 years old) as older houses (26 to 50+ years old) on the market. It may also mean that it does not matter to Californians how old the house is when buying real estate; people will buy whatever is available. The exponential CDF showed a relatively normal distribution with a standard exponential curve. There were many houses that increased in price at lower end of the price range, and after around $350,000, not many houses varied in price. The $500,000 mark is an outlier, especially in this graph. Because there was not much explanation when I originally obtained the dataset, this may be accounting for all the houses in the dataset that were more expensive than $500,000.

The scatterplots of the total number of rooms within a block and value of houses in USD both were compared to, what I declared to be, the dependent variable of the median income of households. I had originally thought that a higher income would mean a higher number of rooms, therefore, a bigger house, but the plot shows that even with a large income, neighborhoods still like to have less than about 17,000 rooms per block. I was expecting there to be clumps on the left lower corner and right upper corner but there was somewhat of a uniform distribution, with a slight increase after $200,000. Again, the $500,000 mark is an outlier that may be accounting for all the houses in the dataset more expensive than that.

Finally, I thought that a multiple linear regression was going to be the best fit for my dataset. All of the p-values of my variables resulted in being significant, and all were measured at 0.000. This is actually not a good thing because the regression calculated that each of the variables impacted the dependent variable equally, which I know would not be the case in real life. So,

this regression is inconclusive. This might be because I retrieved this file off of the Internet and did not collect it myself. One of the other challenges that I faced was with errors in the code, but I hope that with time and more practice, I'll see less errors. I think that if all of the variables were measured by house or household, then the analysis would have gone more smoothly; because I had to think and analyze in terms of blocks, it was much more difficult.

Sources:

Nugent, C. (2017). California Housing Prices. Kaggle. https://www.kaggle.com/camnugent/california-housing-prices.