While doing EDA on my datasets, I found that the prices between 2015 bedroom prices and 2019 prices were highly correlated, using the OLS function to generate both the coefficients and the R2 value, so I used them to fill missing values in the data.

To find out whether or not the data was normally distributed, I graphed the histograms of each and took the chi-square of each variable. When using the chi-square test, I determined the data was not normally distributed, but would become normal when taking the log of the data. Using scipy.stats.chisquare, I was able to get the p-value without doing it manually. To evaluate the p-values, I used the following criteria:

Null hypothesis: The dataset is normally distributed.

Alternate hypothesis: The dataset is not normally distributed.

Since the p\_values were 0.0 for each dataset, we were able to reject the null hypothesis.

The datasets used did not have many other features besides median income and population, though, by using the pearsonr method from scipy.stats we were able to determine that bedroom prices and income were significantly correlated, but only moderately, with an observed correlation score of 0.6. This was also true with the Population variable, but with a lower correlation score of 0.38. The criterion used for this one is as follows:

The null hypothesis: 1 bedroom rent prices and income are not correlated with one another

The alternate hypothesis: 1 bedroom rent prices and income are correlated with one another

The null hypothesis: 1 bedroom rent prices and population are not correlated with one another

The alternate hypothesis: 1 bedroom rent prices and population are correlated with one another