King County House Prices

Description of Data Set

King County is the most populous center in the state of Washington with a population of 1,931,249 people and 789,232 households according to the 2010 US Census. The data set is a sample of 21,597 homes in King County with the following attributes:

- Identification notation for a house
- Date house was sold
- Price of house
- Number of bedrooms
- Number of bathrooms
- Square footage of house
- Square footage of lot
- Total floors of house
- If house has view to a waterfront
- If a house has been viewed
- Condition of house
- Grade of house
- Square footage of house apart from basement
- Square footage of the basement
- Year built
- Year renovated
- Zip code
- Latitude
- Longitude
- Living area in 2015
- Lot area in 2015

Initial Plan for Data Exploration

The objective of the project is to find how various attributes will affect the overall prices of the homes sampled. After running summary statistics, the median price for homes is $450,000, but the maximum was $7.7 million, so to examine the spread of the data, a density curve was plotted which indicated that the distribution was highly skewed to the left. The next step was to clean the data and then create visualizations to get a better idea of the data, as well as see any trends or patterns which might indicate

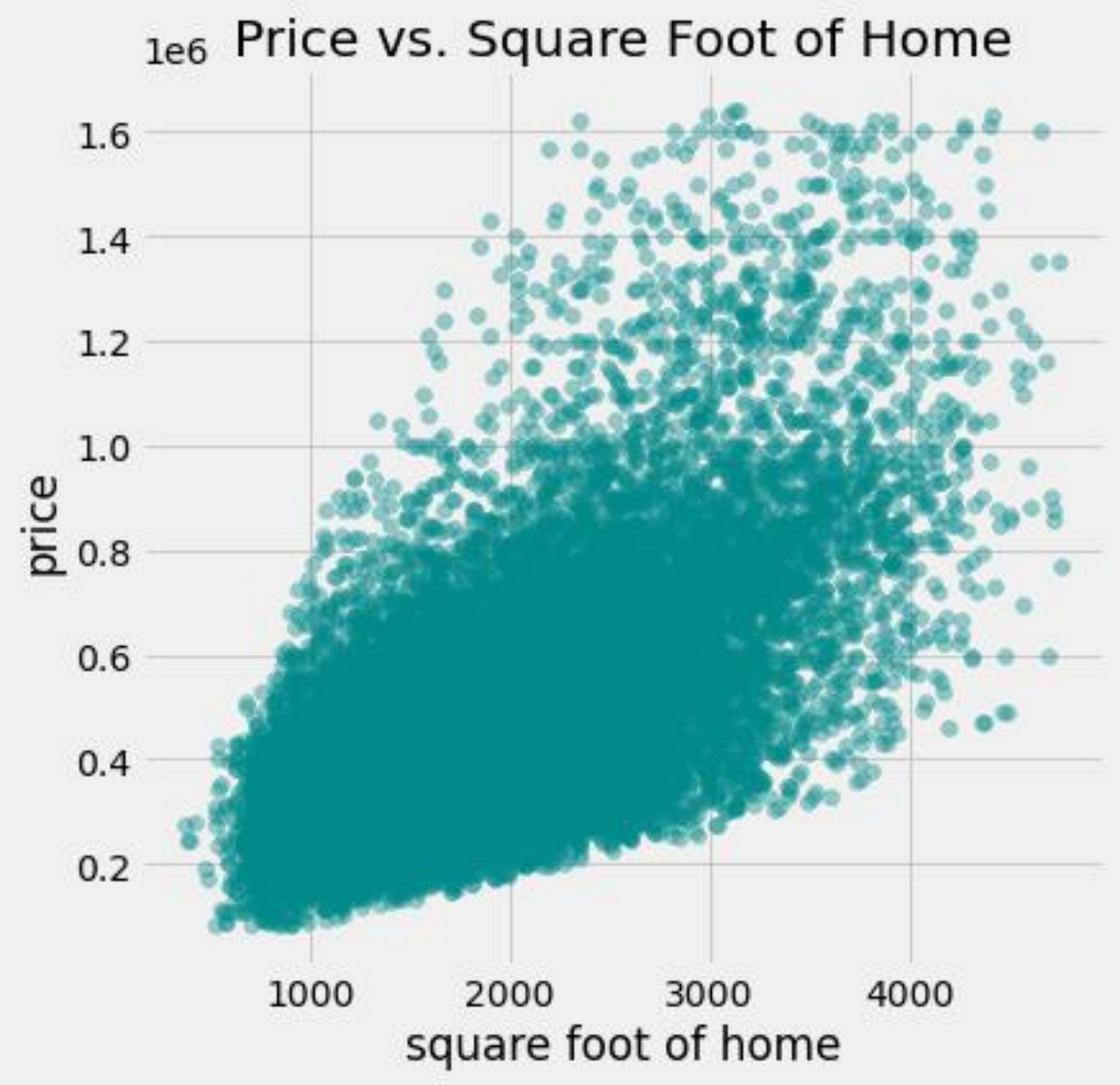
a significant relationship which would affect the prices of the homes. Finally, if an interesting and seemingly meaningful finding arises, it would be tested for significance.

Actions Taken for Data Cleaning and Feature Engineering

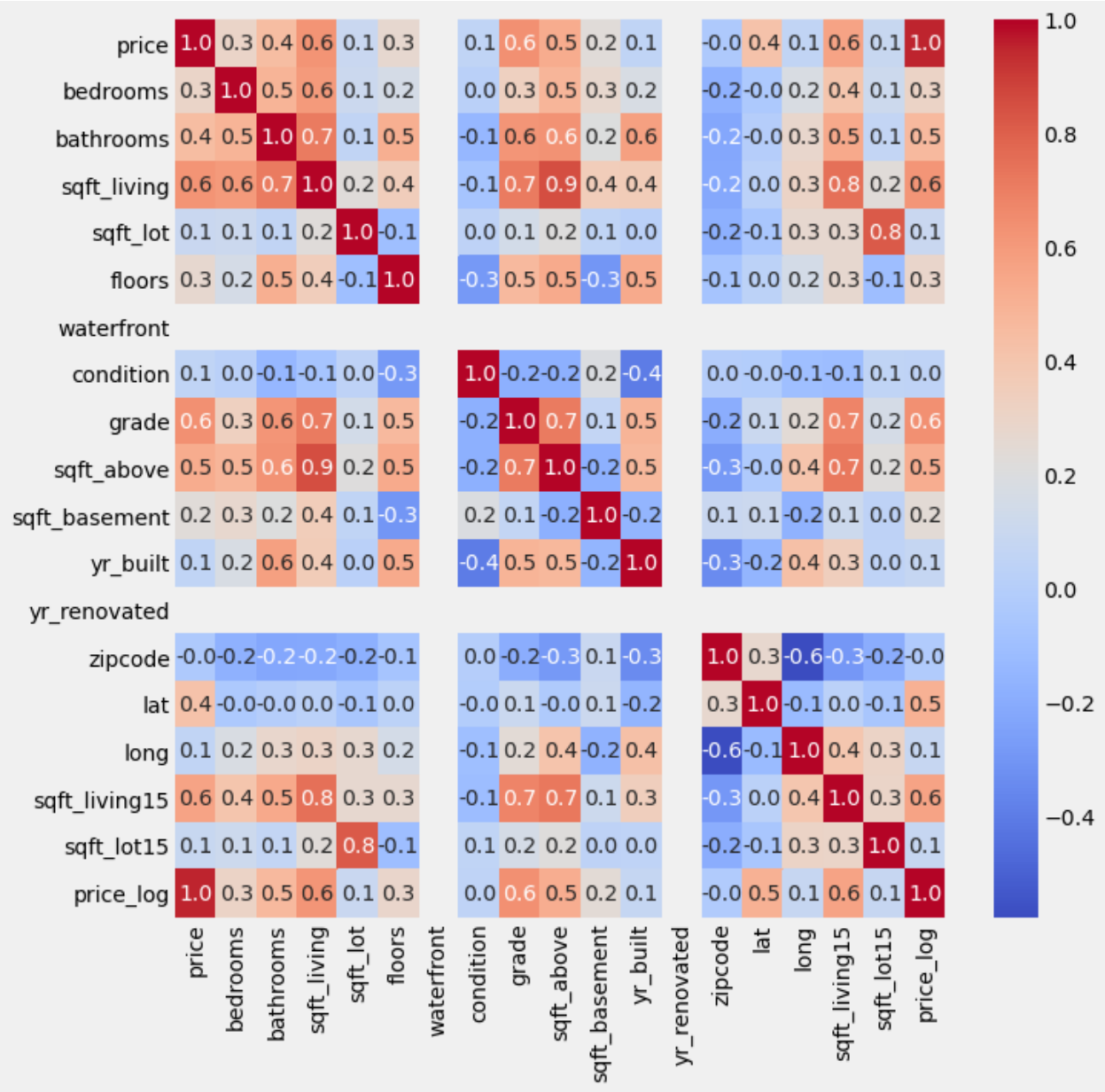
After examining the data set for null values, none were present. Three attributes were removed, as they had no bearing in influencing prices: 1. Identification number, 2. Date of purchase, 3. Number of views. Because all of the remaining values were numerical and in manageable scales, feature engineering was determined to be unnecessary. The summary statistics, density curve, and boxplot for price indicated the presence of outliers that would significantly affect the analysis, so observations with z-scores larger than 3 were removed and 19,121 observations remained.

Key Findings and Insights

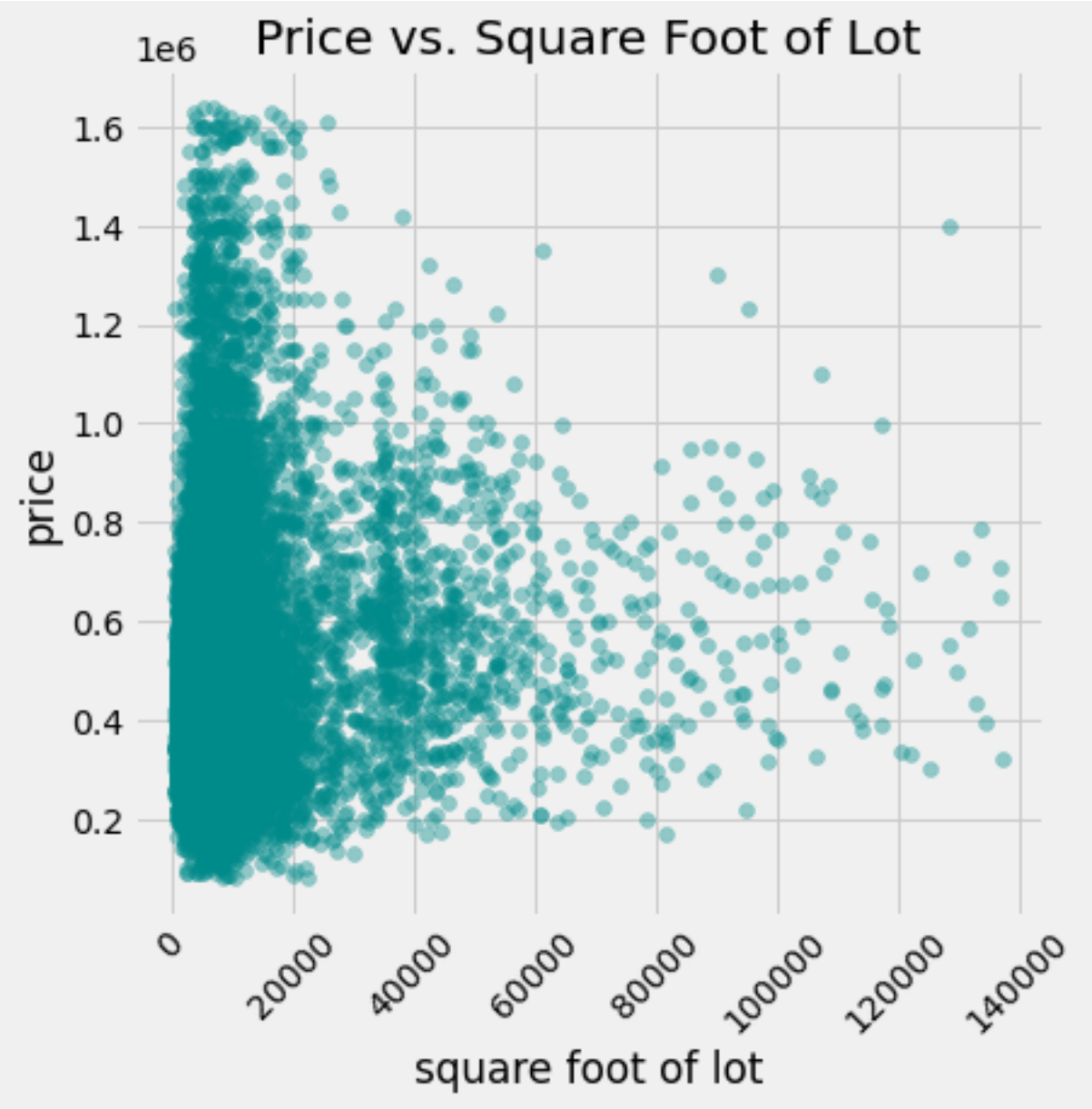
The prices appear to have a linear relationship with the square footage of the house. Perhaps this makes practical sense, that bigger homes will fetch bigger prices.



The following correlation heatmap indicates that square footage of living area has a relatively high positive correlation to the price of homes.

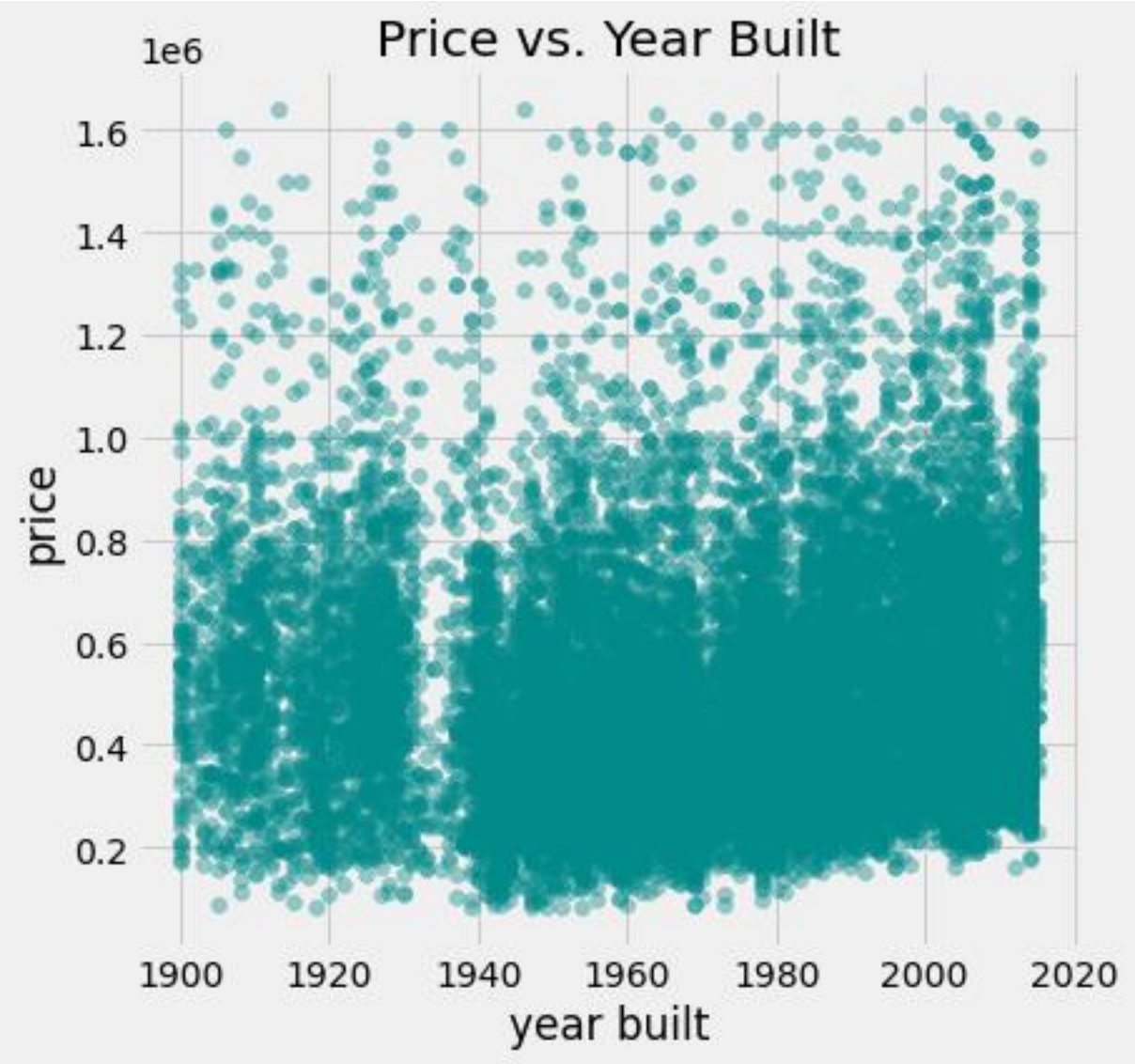


However, the size of the lot does not bear any sort of linear relationship with price. Most homes, including some of the most expensive have lot sizes less than 20,000 square feet. Perhaps this is due to homes in more rural parts of the county with big lot sizes, but smaller home sizes and less desirable zip codes.

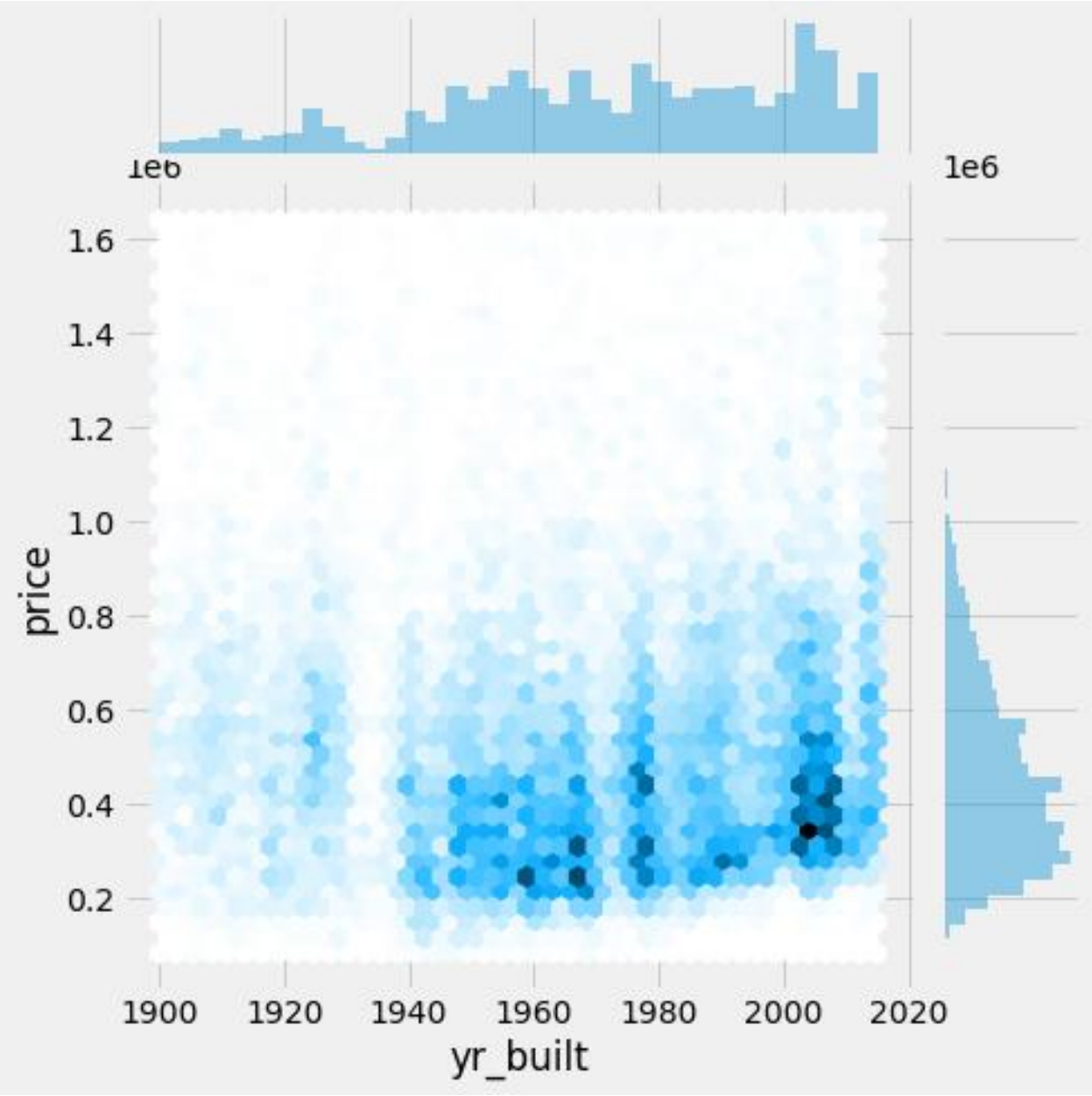


The number of bedrooms seem to have relationship with house price, although it doesn't follow that with more rooms, the higher the price, although one bedroom homes cap at the lowest. Some of the most expensive homes are those with 3 to 5 bedrooms.

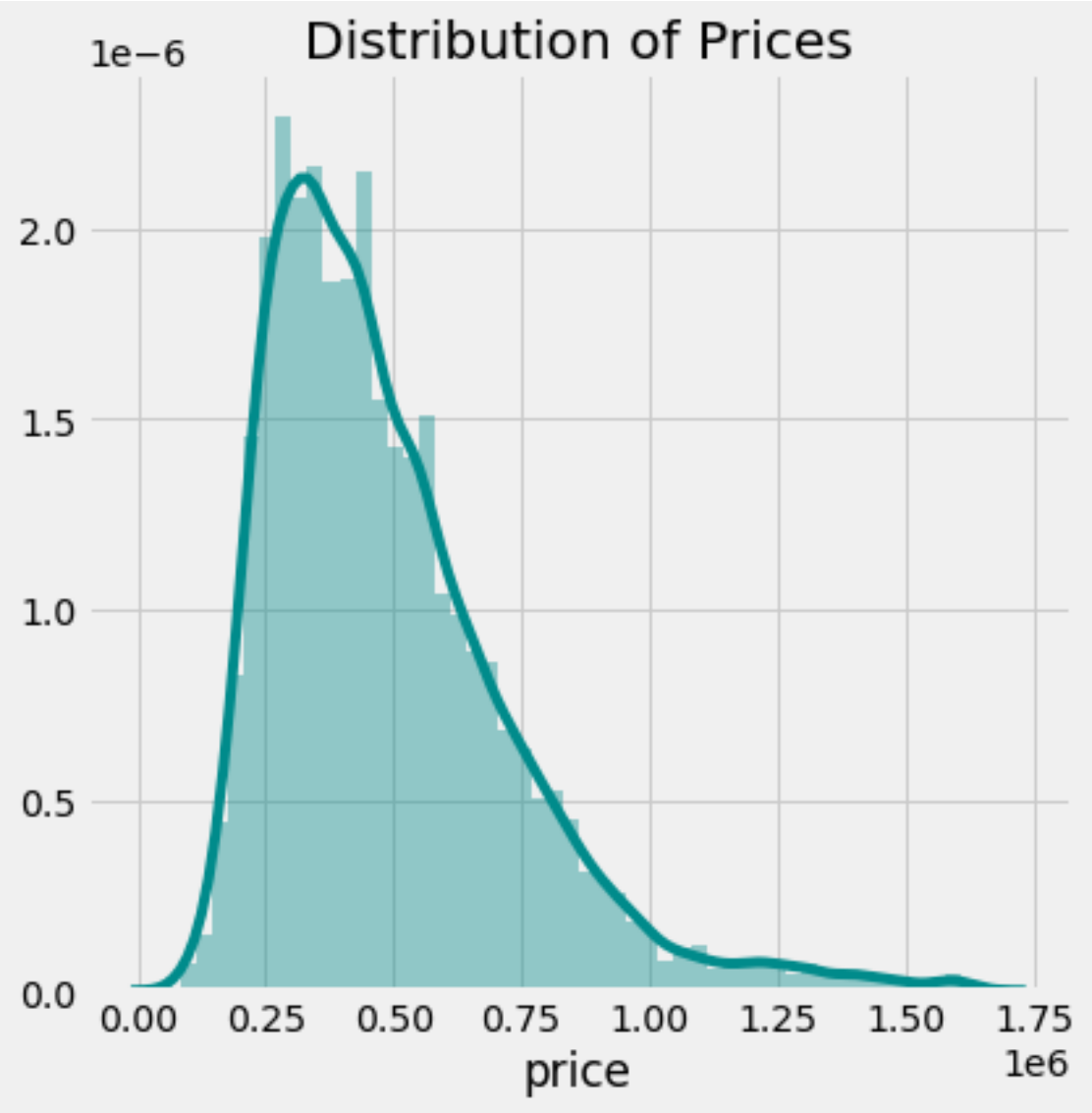
There is no clear discernable pattern with house prices in King County in relation to the year they were built.



However, upon examining the joint plot, there appears to be a high density of homes around the median price range that were built between 2000 and 2010.



The distribution of house prices in King County is left skewed. This indicates that the more expensive homes are significantly more costly and in fewer numbers than a more typical home close the median price.



Hypotheses About Data

1. Larger square footage of living area means a tendency for higher house prices.
2. More expensive homes tend to be in more densely populated urban areas where the square footage of the lot is not as big.
3. The average price of homes in King County is approximately $500,000.

Significance Test on Hypothesis 3

After running a one tailed t-test on the distribution of house prices in King County with the null hypothesis being the average price of homes is not around $500,000 and the alternative hypothesis being the average price of homes is around $500,000, the p-value turned out to be about 8.63e-12, which is far below the 0.05 threshold for alpha, so we can reject the null hypothesis and accept the alternative hypothesis that the average price of homes in King County is approximately $500,000. The true mean of house prices in this data set is about $488,192

Next Steps in Analyzing Data

The eventual goal of this project would be to create linear regression models to fit a line in order to predict house prices in King County. However, the distribution of prices is left skewed and in order to facilitate linear regression, the prices need to be log scaled to achieve a more or less normal curve. A pair plot of the dataset to get a better idea of overall trends should also be pursued, as well as other visualizations such as bar graphs to get a better grasp of the data. There are many features in this data set and those will need to be explored as well aside from what was done so in this preliminary analysis to see how they affect the target variable of house price. Finally, using the appropriate library, visualize the locations of each data point on a geographical map of King County as the data set provides longitude and latitude coordinates.

Quality of Data set

The data set for King County house prices was clean, well-formatted, and overall easy to work with. The few categorical variables were already transformed into numerical values and the data set contained no null values, so cleaning the data set took minimal effort. The data set also contained a sufficient number or observations, although additional ones can be helpful for more analysis, as well as training machine learning algorithms.