EDA Kings County, Seattle blog post
LaShanni Butler
1/24/19

I found this exploratory data analysis (EDA) of the Kings County, Seattle dataset quite challenging. However, I'm glad that we had this exercise and it tied together all the many months of lessons in Module 1. I must admit, I took several days just staring at a blank Jupyter Notebook for this assignment, trying to figure out how the heck am I going to do this. It definitely was stressful because I've never done anything like this before. Luckily there were some great resource provided by the instructors for this project.

For the first step of my EDA, I loaded the necessary libraries (e.g. – NumPy, Pandas, Matplotlib, etc.) along with the Kings County dataset from the repo and began my analyses. After looking at the columns and rows using .info(), .head() and .tail(), I initially thought several columns would really drive the price of homes in Kings County. At a quick glance, columns like waterfront, square feet, view, and zip code (i.e. – location, location, location) would really be the most important to the buyer and predict housing prices in my dataset. I mean categories like this just make sense (at least to me) and are important for many home buyers. Interestingly as I explored some columns, like waterfront, I noticed the values in this column contained NaNs. So, waterfront appeared to be a Boolean value and was scored as 0 or 1 (0 = No waterfront, 1 = waterfront). Thus, I determined this column did not have much value to me after all (go figure). A few more columns in this dataset either contained NaNs or all zeros, so I thought it would either best to drop these columns or fix the NaNs. This was a great exercise, because coming from a research background, I have never come across a clean and perfect dataset. So, I knew learning how to clean this data would work to my advantage in future jobs and projects.

My second step in this EDA was to explore my dataset with descriptive statistics and visualizations. Running .describe() allowed me to get the 5-point statistics for my predictor, price. Looking at the 5-point statistics is very powerful because it generates a summary of the following:

- Min: minimum value in the data set
- 1st Quartile: 25% of the data falls below this value.
- Median (2nd Quartile): 50% of the data falls below this value
- 3rd Quartile: 75% of the data falls below this value
- Max: largest value in the dataset

Once I had an idea of the lowest, mean (or average), max, and everything in between for the Kings County home prices, I then proceeded to generate some histograms and distribution plots. My graphs showed that home prices in Kings County showed a normal distribution. However, there was a positive skewness to both graphs. I believe the positive skewness is due to the upper 1% buying the most expensive homes in the area. It makes sense that this trend would affect the data and skew it. Also, I wanted to look at a histogram and distribution plot of square feet, since I figured it greatly influenced home prices. Like price, these graphs were also positively skewed and showed a normal distribution.

Next was the heat map, which is a really cool way to look on a colored grid of all the columns and rows in your dataset. The heat map compares each column and row against each other and generates a scale of what is correlated which each other. So, for instance, my heat map showed price and sqft_living

were highly correlated. Based on the heat map correlated values, I intended to look at bathrooms, sqft_living, sqft_above, grade, yr_built, and sqft_living15 for my regression. However, before doing that a multicollinearity check would be needed.

This multicollinearity check brought me to the final steps in this EDA process. As I had stated earlier in this blog post, I thought bathrooms, sqft_living, sqft_lot, grade, sqft_above, yr_built, and sqft_living15 would be the best predictors of homes in Kings County. I ran a predictor test of the dataset using the .corr() function. For this exploration, I considered any value with a correlation of 0.50 to be a high absolute value (and thus it would be omitted from my analyses). Lo and behold, all the categories I was going to analyze were over that 0.50 mark! So, this was truly interesting to me because an initial correlation check of the dataset essentially skewed my notions of what I thought drives home prices in this area. So, with this new multicollinearity information, I had a new set of categories to analyze which were:

- bedrooms
- floors
- condition
- yr_built
- zip code
- sqft_lot15

Both waterfront and yr_reno were also included in this new group. However, since these categories had either missing values or issues with how consistent the data was recorded, I did not include them in my regression analysis.

From here I created my X arrays (see bulleted list above) and Y array (price). Then using scikit-learn, I ran a train_test_split. This function allowed me to split my dataset into a training and test dataset, as well as, help fit my model to check the score of my data predictions. Then I compared each X predictor value against the Y predictor, generated scatter plots and least squares data tables that contained the R-squared and p values. What I found was interesting and really made me question if I choose the right predictors. All my R-squared values were low; however, my p values were not greater than 0.5. So, did I choose the right categories for this analysis? In running the multiple regression, I looked at all the X predictors against the Y predictor in one table and generated another R-squared. Guess what, the same trend was there too! I now had an overall R-squared of 0.16 and all my p values were less than 0.5 in my multiple regression. So, this really begged the question if this was a good representation of Kings County housing data. I did a quick Google search and saw that a low R-squared is not necessarily a bad thing. But naturally, I wanted a nice R-squared value that was close to at least .85. Maybe its back to the drawing board to see what I can change to make the model stronger.

EDA Kings County, Seattle blog post
LaShanni Butler
1/24/19

Even though I wasn't truly satisfied with my low R-squared, I did ascertain some interesting facts about Kings County housing data:

1) The more bedrooms a home has, influences the price of the home

2) Homes with 2 floors had the widest range in price in this dataset. So, I assumed many buyers are attracted to home with 2 floors

3) Homes with a condition of 3 also have the widest price range. So, this might be attractive to buyers in Kings County

4) Year built did not seem to influence the price too much. I gather buyers have no preference if homes were built in 1900 or in 2019

5) The most desirable zip codes for buyers were in the 98000, 98050 and 98100 as there are many price clusters in these areas

6) Land lots to the nearest 15 neighbors were clustered at 0 to approx. 75,000

Lastly what I really learned was factors that I personally thought would drive home sales (e.g. - waterfront, view, etc.) really had no bearing on how a home price is predicted in Kings County. In closing, I found this EDA very challenging, somewhat fun (only if you take away the stress level, lol). I'm really glad that I made it this far and now have my first data science project in my portfolio!