Jeremiah Philipps & Max Nowak

5/11/2022

CS 305

**Final Project Report**

*Introduction*

For this project, we examined a collection of data concerning the Boston housing market in the late 2000’s – early 2010’s. The data consisted of a csv file from Kaggle.com The file contained 1460 rows of data, each row describing the sale of a home in the greater Boston area. The file consisted of 81 columns of information about each house sold (Sale price, square footage, overall quality, garage/lot information, date of sale, etc.) Our goal was to identify trends between different columns with the sale price. We essentially came at this dataset with the question “What are important characteristics that affect the sale price of homes in Boston?” The goal behind this would be to inform hypothetical clients to aid in the home buying process.

*Work Breakdown*

Due to schedule differences, all work on this work on this project was completed asynchronously. The primary method of contact was email for questions, suggestions, and reporting while GitHub was used for filesharing. While both partners had a hand in the project from start to finish, it was decided to break the majority of the work in to preprocessing (assigned to Jeremiah) and visualization (assigned to Max). This separation of work felt optimal in terms of workload balance and subject interest. With that being said, despite this split both partners did assist each other on their respective parts to some degree.

**Project Initiation**

Both partners spent time searching Kaggle and other online resources for datasets to work with, emailing possible candidates back and forth before the Boston housing dataset was agreed upon

**Preprocessing**

The bulk of the preprocessing was completed by Jeremiah, with input and assistance from Max. The preprocessing was done using Python (Pandas, Numpy, Matplotlib.pyplot, and Seaborn libraries were used specifically) in a Jupyter Notebook environment. Besides importing the data and libraries the preprocessing consisted of cleaning up null values, removing unnecessary columns, and removing outliers. Some rows had an extremely high percentage of null values (one column had 99% nulls). After examining these columns, we determined they were all rather niche characteristics (Pool quality, alley [Y/N], etc.) that were very unlikely to affect the sale price. We ultimately decided to drop any columns that had contained over 5% null values as they were both unnecessary information and too sparsely populated to accurately estimate/fill the nulls. After removing these columns, there were eight more columns that had null values. The highest amount of nulls was 38 and many of the null rows were shared across these columns. Because 38 was such a small subset of our overall dataset, we decided to simply remove any rows containing null values.

Once our dataset was free of null values, we did some preliminary visualization in Python to further guide the data cleaning and guide later visualization. The most useful tool for this was the correlation matrix shown in Figure 1.

Graphical user interface, chart

Description automatically generated

Figure 1

Chart, scatter chart

Description automatically generated By examining the right most column (relating to the sale price) we found many light-colored rows that mean those features are highly correlated with the sale price. We plotted the square footage and sale prices to get an idea for what this correlation looks like, as shown in figure 2.

Figure 2

Chart, scatter chart

Description automatically generated It was very evident, visually that there were outliers, the two dots in the lower right hand. We removed these which resulted in a much cleaner looking correlation as shown in figure 3.

Figure 3

Going into this, we expected square footage to be one of the main factors affecting sale prices which is why we thought it important to remove any outliers, but after discussion we agreed it would be impractical to do so for the remaining 74 columns. At this point, the preprocessing was completed, the data was saved to a new csv to be used for visualization in Tableau.

**Visualization**

The visualization was completed by Max, with input and assistance from Jeremiah. The visualization was created in Tableau. The process for creating the four required graphs and final dashboard was essentially just trial and error. After combining countless combinations of columns with the sale price we determined that some of the major factors that affected sale price were: Square footage, building type, neighborhood, and the time of year sale occurs. While analyzing these relationships we continued to think of our hypothetical client and what other information/relationships would be useful to them. These other useful relationships we discovered where price/sq. foot as it relates to building type and neighborhood, as well as neighborhood popularity and popular times for selling. The overall visualization and summary of our findings is shown in figure 4. Chart

Description automatically generated

*Figure 3*

*Conclusion*

In conclusion, we believe we achieved our initial goal of understanding the influences of sale prices in the Boston housing market. We adequately cleaned our data and created very useful visualizations that could help a hypothetical homebuyer understand if they are getting a good or fair price when analyzing a given home. In the process we even uncovered more correlations than anticipated in relation to price/sq. foot and the seasonality of sales and pricing.