INFSCI 510 Final Project: How Housing Features Affect Price

Drew Whiteside
School of Computing and Information
University Of Pittsburgh
Pittsburgh, PA, USA
daw196@pitt.edu

Hayden Burget
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
hdb17@pitt.edu

Cullen McDonald
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
cqm5@pitt.edu

***Abstract*–Research question: What features of a house will affect its listing price online, and to what extent? This question is interesting to us because answering it would provide insight into what aspects of a house may cause its price to be valued more than its counterparts. The results of this study may show prospective buyers or renters of houses with a sense of how certain listings should be priced, given the features they have. Our model can be used by anyone looking to rent a property or list one themselves.**

Plan for data analysis: In order to train our model, we will use a dataset titled "USA Housing Listings" on Kaggle, which displays data from over 300,000 Craigslist listings on housing throughout the country up until January 2020 [1]. It includes features of the listings along with the price they were listed at. We will select 4-6 pertinent features of these listings and train a model in an attempt to show the extent to which each feature affects the listing price of an apartment.

I. INTRODUCTION

In the housing market, properties can vary based on a wide array of factors. Some of these factors include location, size, and the different amenities they offer. Because of the variability in the different elements of each property, it becomes unclear which factors truly affect how much a property is listed for. Therefore, the aim of the project is to form a machine learning model that can predict the relative cost of an apartment based on its factors. As a source of data to train our model on, we will use a dataset from properties sold on Craigslist, which includes thousands of listings that have been put up from the United States. Through our model, we will attempt to determine whether region, size, beds, baths, furnish status, or parking availability affects the monthly rent price of an apartment.

II. METHODOLOGY

To begin analyzing our data, we uploaded the data into a Jupyter Notebook and loaded the data into a Pandas dataframe. According to the "type" column of the CSV file, there were a few types of listings available, such as apartment, condo, house, and duplex. Therefore, because the dataset is extremely large, and in an effort to reduce as many external factors as possible, we chose to focus solely on apartments as these are most relevant to other people of our age group. Choosing apartments also came with the benefit of having the most data points available compared to the other list types. Then, we decided on a set of 6 features that were originally in the file:

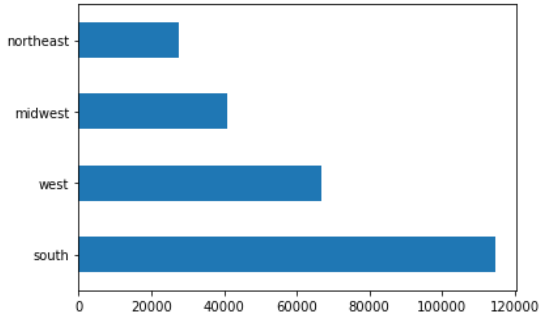
- state
- sqfeet
- beds
- baths
- comes_furnished
- parking_options

Since 50 states would result in too much noise for a machine learning model, we modified the column by mapping the "state" value of every listing into a set of four regions: south, west, northeast, and midwest. Below is a list of all of our features for the model, including the justification for using them and descriptive information about them:

Feature 1: Region

Justification: An apartment's price could vary largely based on where it is situated in the country. Most areas of the country change drastically in median income, population density, weather, and overall desires of the people for specific types of housing. With the United States being incredibly geographically diverse, it is important to include region as a variable that will affect the cost of an apartment.

FIGURE I. BAR GRAPH OF MAPPED APARTMENT REGIONS



South: 114,653
West: 66,855
Midwest: 40,829
Northeast: 27,489

Examining the value counts of the mapped regions showed that there was at least a somewhat even distribution for the bins that were chosen for each state.

Feature 2: Square Footage

Justification: The bigger the apartment, the more expensive it is if all other variables are equal. Our justification for removing outliers was to keep square footage between 300 - 3000 square feet to accommodate very small studios and very large but possible apartments. We wanted to prevent too many outliers in square footage from skewing our distributions too much to where the models would be less capable of predicting the price of average apartments. From eliminating the extreme outliers, we have an average square footage of 961 square feet which is slightly above the reported average in the United States but within range when compensating for standard deviation of 253 square feet. We lastly applied a square root transformation so it would be easier to analyze the distribution in the data by having a more normalized scale to our density graph.

FIGURE II. DESCRIPTIVE STATISTICS FOR SQUARE FOOTAGE COLUMN

count	249826.000000
mean	961.546220
std	253.777384
min	300.000000
25%	780.000000
50%	950.000000
75%	1100.000000
max	3000.000000

FIGURE III. BOXPLOT OF SQUARE FOOTAGE BEFORE REMOVING OUTLIERS

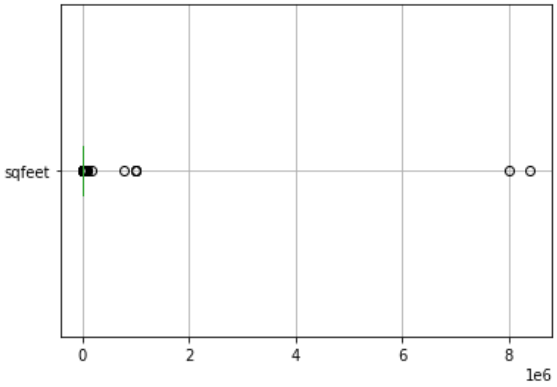


FIGURE IV. BOXPLOT OF SQUARE FOOTAGE AFTER REMOVING OUTLIERS.

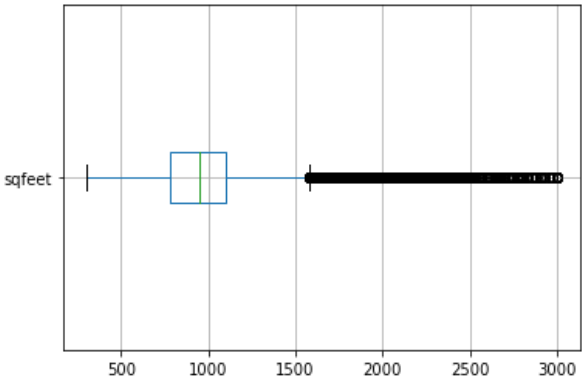


Fig IV. Maximum square footage is now 3000 square feet.

FIGURE V. DENSITY PLOT OF SQUARE FOOTAGE BEFORE SQUARE ROOT TRANSFORMATION

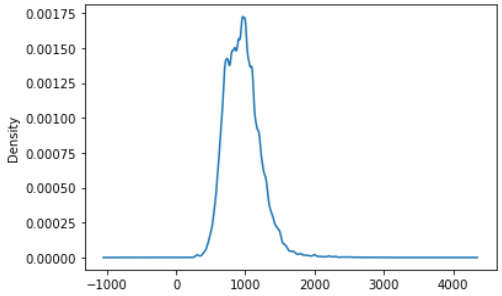
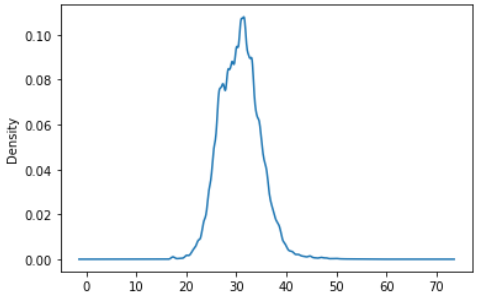


FIGURE VI. SQUARE FOOTAGE DENSITY PLOT AFTER SQUARE ROOT TRANSFORMATION



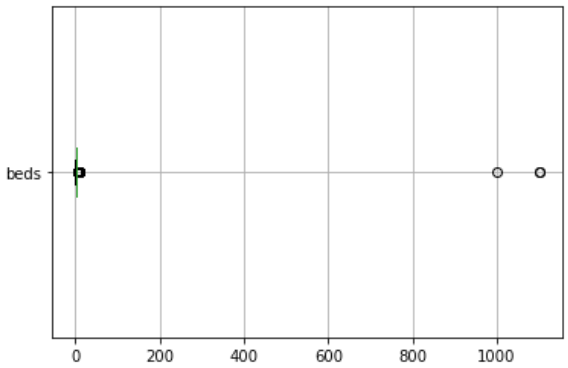
Feature 3: Beds

Justification: having more or less beds can affect the price of a house in relation to how much space is remaining for living areas, kitchens, storage, etc. For outliers we decided to keep the apartments that had one to five bedrooms. This way we can relatively compare the apartments without having extremes or just flat out mistakes in the bed room count. As shown by our density plot, it makes perfect sense to have mostly two bed rooms followed by one bedroom and tailed by three bedrooms.

FIGURE VII. DESCRIPTIVE STATISTICS FOR BEDS COLUMN

count	249826.000000
mean	1.824698
std	0.692842
min	1.000000
25%	1.000000
50%	2.000000
75%	2.000000
max	5.000000

FIGURE VIII. BOXPLOT OF BEDS BEFORE OUTLIER REMOVAL



Boxplot for number of beds shows extreme outliers that indicate an error in the data entry as we would not expect bedroom counts as high as 1000.

FIGURE IX. BOXPLOT OF BEDS AFTER OUTLIER REMOVAL

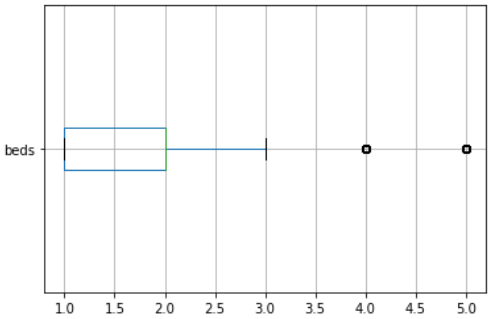


Fig IX. After removing the outliers, we have a maximum of apartments with bedroom counts between and including 1 to 5.

FIGURE X. DENSITY PLOT OF BED COUNTS

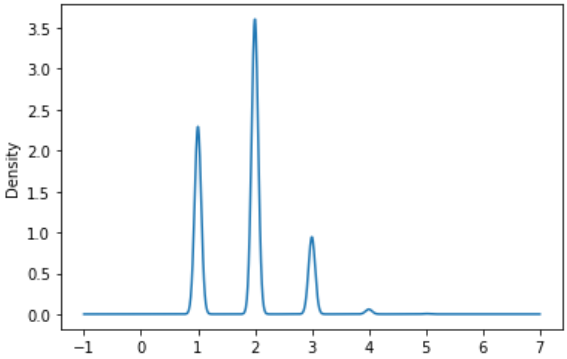


Fig. X. Density plot showing the frequency of the number of bedrooms in each apartment – two bedrooms being the most frequent.

Feature 4: Baths

Justification: Similarly to beds, people may be willing to pay more or less depending on how many bathrooms are available. We decided to only keep the data that contained one to four bathrooms because this will help mitigate noisy data that will skew the data too much and create inaccuracies in our model;sp redicitions and remove impossibilities from errors in entering the data such as is present with an apartment with over 70 bathrooms. Most of the apartments did come with 1 bathroom as is most common in 1 and 2 bedrooms with 2 bedrooms being the second most common. It is rarer but not impossible to have 2 bedrooms that have 1.5 bathrooms, so we were not surprised to see that as more common than apartments with 3 or more bathrooms as only the biggest apartments will have that.

FIGURE XI. DESCRIPTIVE STATISTICS FOR BATHS COLUMN

```
count    249826.000000
mean          1.480921
std           0.517148
min           1.000000
25%           1.000000
50%           1.000000
75%           2.000000
max           4.000000
```

FIGURE XII. BOXPLOT OF BATHS BEFORE OUTLIER REMOVAL

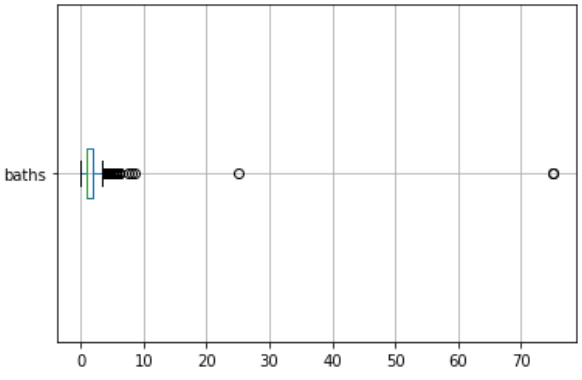


FIGURE XIII. BOXPLOT OF BATHS AFTER OUTLIER REMOVAL

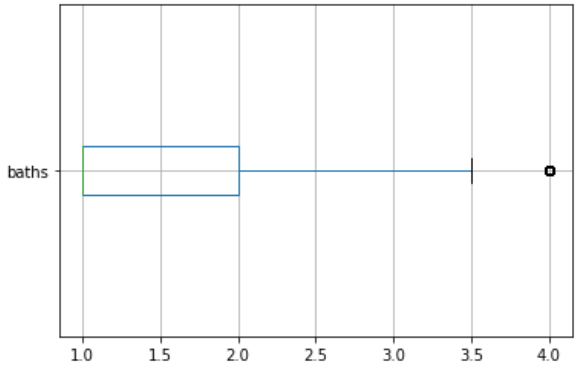
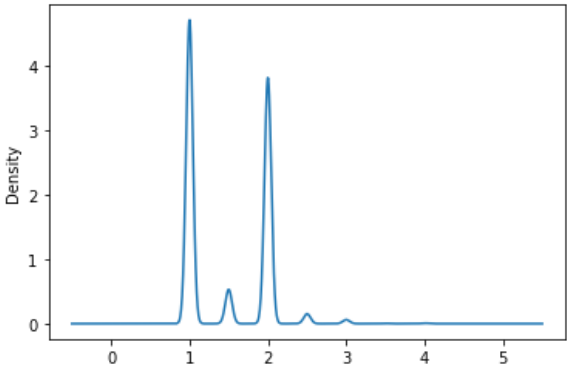


FIGURE XIV. DENSITY PLOT OF BATHROOM COUNTS



Feature 5: Furnishing Status

Justification: An apartment that comes fully furnished will cost a lot more than the exact same apartment that does not come furnished. Typically the option to have an apartment that comes furnished will imply an area that is more expensive to live in which may help to isolate the luxury apartments away from non luxury apartments even if their square footage, bedroom count, and bathroom counts are identical.

FIGURE XV. VALUE COUNTS FOR COLUMN COMES_FURNISHED

```
0    242582
1      7244
Name: comes_furnished
```

Value counts for furnishing status. Less than 3% of apartments come furnished which represents a minority of our data and makes this predictor very minor.

Feature 6: Available Parking

Justification: Different options for parking a car can add a lot of value to a home. For example, the same home with street parking compared to if it had a driveway and a garage will change the price significantly. For pieces of data that were null, we assumed it was not available since places that had garage or covered parking would typically correctly list this since it positively affects the overall price of the unit. This can hopefully isolate apartments in lesser developed areas or cheaper areas from apartments that may have underground garage parking, or a garage space that may help to imply a more expensive apartment.

FIGURE XVI. VALUE COUNTS FOR COLUMN PARKING_ATTACHED

```
0    231519
1     18307
Name: parking_attached
```

Value counts for available parking at each apartment. About 7.3% of apartments were advertised with parking options attached to the apartment. This represents a very small minority of the data that will hopefully lead to a small subset being more expensive when compared to similar apartments without parking attached.

III. RESULTS

From this dataset, we attempted to find if we could accurately predict the monthly price of an apartment that we had never seen before just based on a few core variables. In order to test if this was possible, we used 3 machine learning models on our data: Naive-Bayes, K Nearest Neighbor, and Decision Tree.

We found the best accuracy was using the K Nearest Neighbor model which yielded an accuracy score of about 66.6%. Decision Tree gave us an accuracy score of 63.9% accuracy followed up with Naive-Bayes that was 63.28% accurate.

K Nearest Neighbors was most likely the best model to most accurately predict our intended response variable due to the fact that apartment rent prices can most easily be predicted if they are split into groups based on square footage, number of bedrooms, and number of bathrooms primarily. Without seeing the apartments in person to judge them based on the neighborhoods they are in, or how nice they physically appear, you are forced to group apartments to attempt to predict which price range it will fit into. To improve the accuracy of these models, you could expand the price buckets so that the model has more flexibility and higher chances to be correct. However, this might ruin the meaningfulness of the overall model since the price ranges will be too large to give someone a good idea of what they will be paying each month.

This analysis was more accurate than we had first predicted considering the data that we were given which can be identical but can represent very different apartments. Although the models cannot offer a perfect basis on what an apartment should cost, they might be able to offer a decent starting point to expand on if given more variables in the future that can help to specify a more accurate rent price.

IV. DISCUSSION

The overall goal of this project was to create a model that someone could easily use to give them confidence in the price range of an apartment in an area based on a few factors. Although we reached about a 67% accuracy score, it is still not enough for someone to fully trust this model repeatedly in estimating rent prices. In this dataset we were limited by not having certain variables that we know can vastly increase an apartments monthly rent cost such as the date of its last renovation, if it considered a 'luxury' apartment, or if it is at the top floor on an apartment complex vs the ground floor just to name a few. The exact same apartment can change prices drastically if just moved a few streets closer to a major attraction such as a city park or a better view of the ocean. This information would need to be included in order to create a model that someone could confidently use in order to consistently predict the price of almost any apartment.

Contributions to Overall Project

Drew Whiteside: Worked on initial project abstract, helped find Kaggle dataset used for the project, did initial data cleaning process for the "data story," wrote abstract/introduction for the final paper

Hayden Burget: Worked on initial project abstract, contributed to data story, primarily worked

on methodology section, assisted with data manipulation in Jupyter notebook.

Cullen McDonald: Worked on initial project abstract, assisted in creating Jupyter notebook for data story, updated Jupyter notebook for final report and worked primarily on Results + Discussion section of final report.

V. REFERENCES

[1] Reese, Austin. "USA Housing Listings." Kaggle. https://www.kaggle.com/datasets/austinreese/usa-housing-listings.

[2] "Average Rent by State 2022." *Average Rent by State 2022*, https://worldpopulationreview.com/state-rankings/average-rent-by-state.