Timila Kulkarni

STAT 311

Group 31

Final Project Part 1

Introduction

Our dataset consists of 384976 observations and was built by scraping data from Craigslist. It contains data on all relevant information provided by Craigslist on privately sold housing options. The data in our version of this dataset include 4 quantitative and 6 categorical variables. The categorical variables include different ways of grouping the data, like the type of housing unit or if dogs/cats are allowed. The quantitative variables include different measurements and different types of measurements, from price to the number of beds to square footage. These will be used as predictor and response variables in our analysis to see the relationship, if any, between the variables.

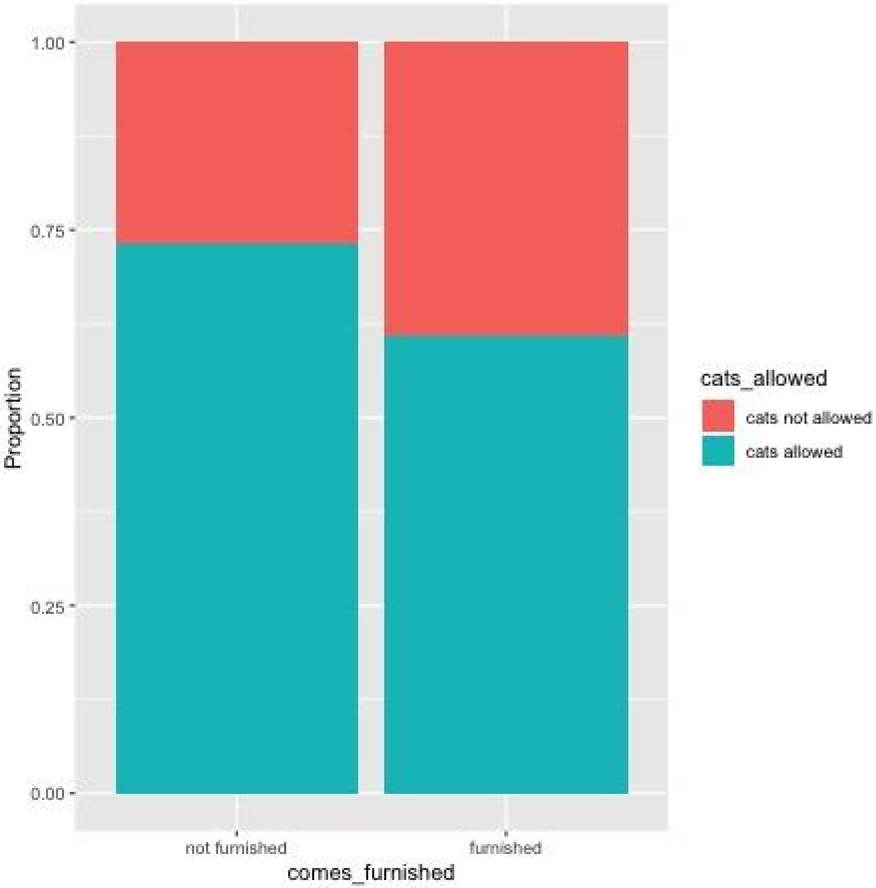
Exploratory Data Analysis (Timila)

I began my Exploratory Data Analysis by creating a two-way contingency table to explore the relationship between my two categorical variables, comes_furnished and cats_ allowed. The table was as follows:

	Cats not allowed	Cats allowed
Furnished	26.7%	73.3%
Not furnished	39%	61%

From this, we can see that there is a possibility of dependence between these two variables since it seems like there is a higher percentage of homes that are not furnished that allow cats, as opposed to furnished homes.

To assess this possibility, I created a bar chart:



Visually, we can see that there is some dependence between the two variables since the proportion of cats allowed changes as furnishing status does.

Now, for the quantitative variables. I first created 7-number summaries for each of the variables, price and beds. They are as follows:

Minimum	Q1	Median	Q3	Maximum	Mean	Standard Deviation
0	805	1036	1395	2768307249	8825.722	4462200

Minimum	Q1	Median	Q3	Maximum	Mean	Standard Deviation
0	1	2	2	1100	1.905345	3.494572

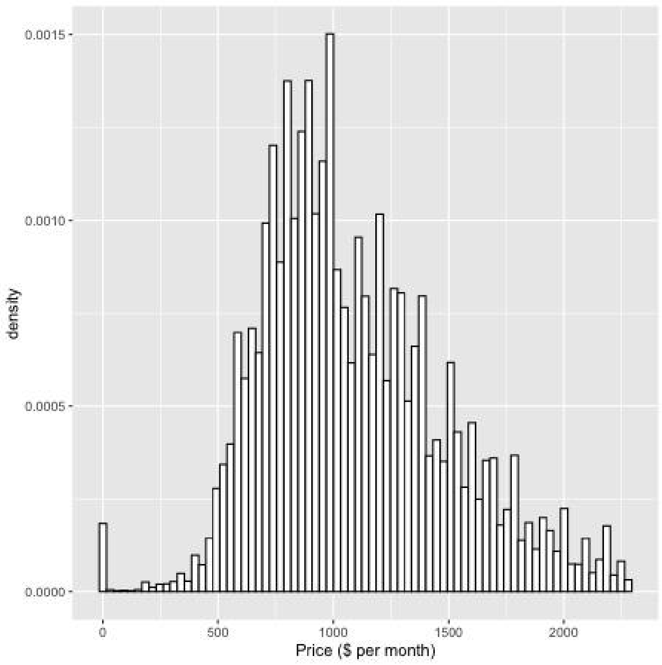
From these summary statistics, we can see that the mean rent per month (price) is over $8800, with a standard deviation of $4,462,200. This shows that there is a high variance in our data. However, the median is $1036, suggesting that there are a lot of outliers with high leverage.

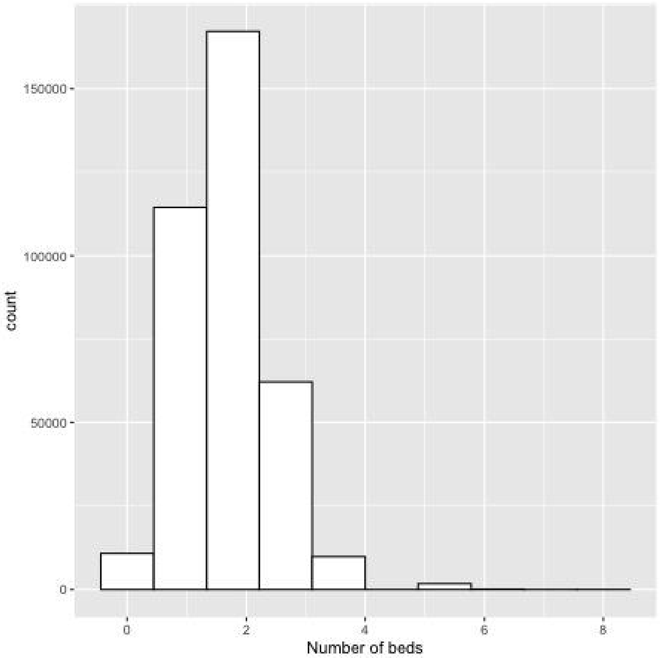
The first and third quartiles are $805 and $1395 respectively, again backing up that there seem to be a lot of outliers. Also, it is important to note that the minimum is $0, so this may influence our data a lot as well.

For the beds variable, we see that the mean number of beds is almost 2, with a standard deviation of about 3.5. However, we have at least one outlier since our maximum is 1100; unlike with the price variable, there don't seem to be too many such observations since the mean and standard deviation look pretty reasonable.

Since there seemed to be so many outliers and variance with the price variable, I decided to clean up the data. So, in the new, cleaned-up dataset, I removed observations that were outliers using the 25th or 75th quartile – 1.5*IQR rule. I also removed the outliers of 1100 and 1000 beds for the beds variable, and all price observations of 0, since they were unnecessarily skewing the data.

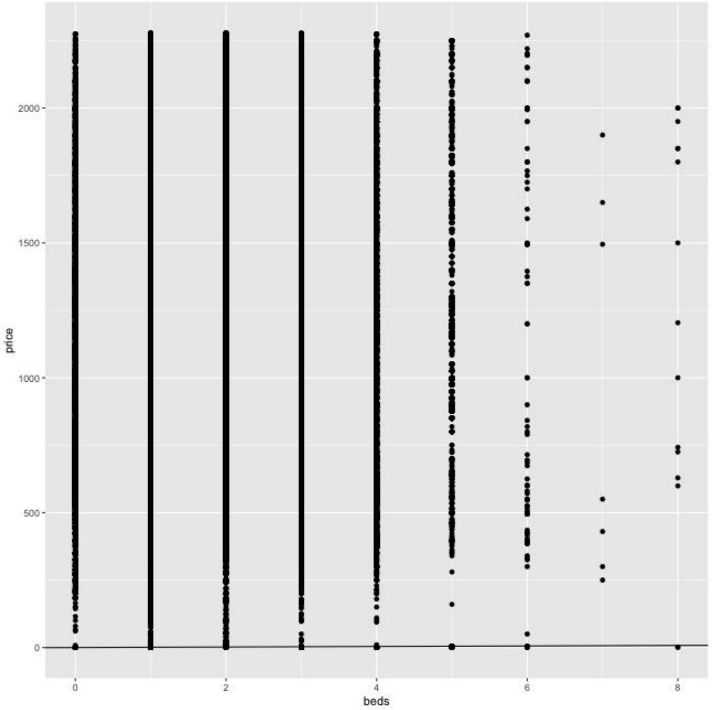
Then, I created histograms for both variables:





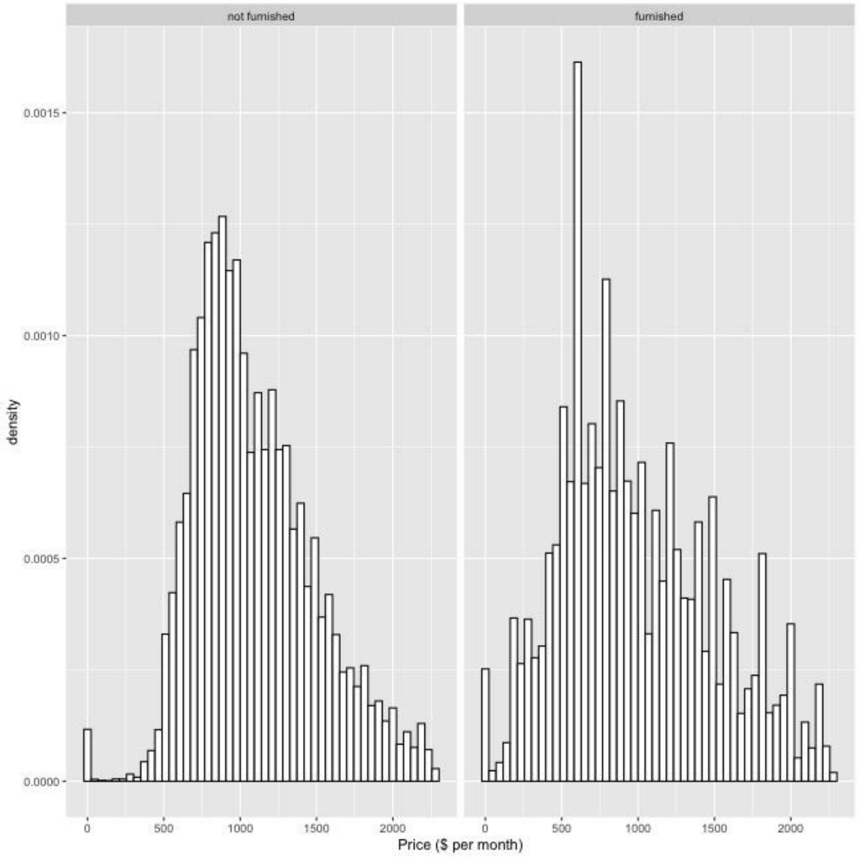
The histogram for price is unimodal with a right skew, with the mode at about $990. It is very wide and spread out. The beds histogram is also unimodal, with a slight right skew, but much taller and skinnier. The mode is at 2 beds.

Now, we are ready to look at the relationship between the two variables. So, I plotted a scatter plot:



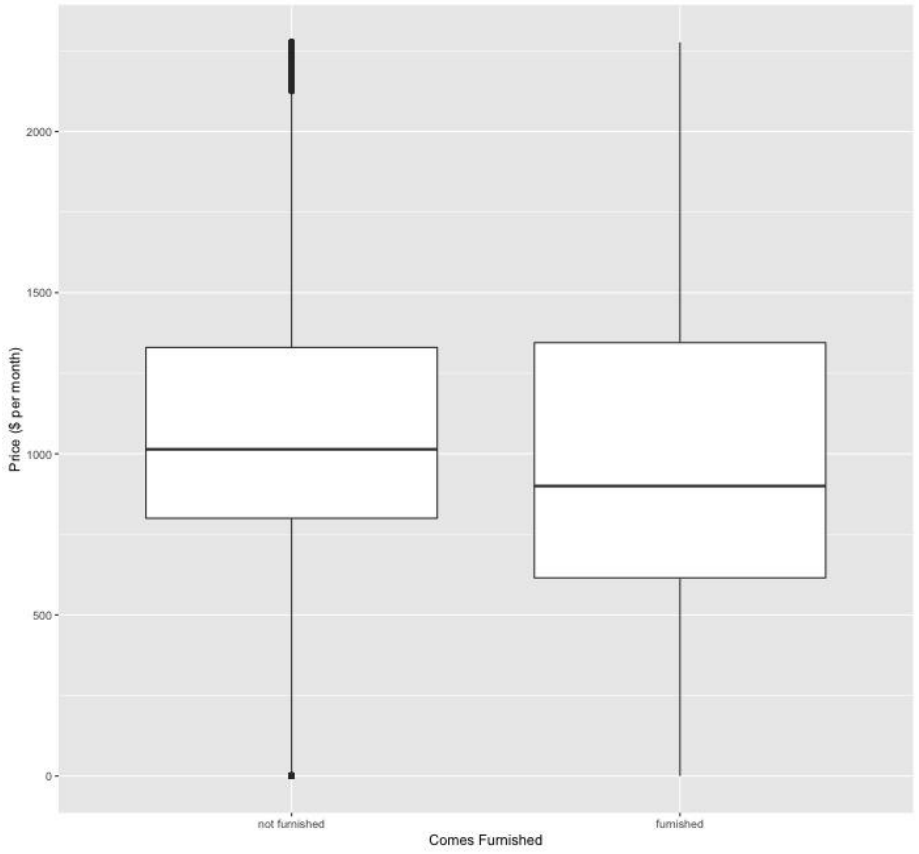
Unfortunately, it doesn't look like there is any kind of linear relationship between the two variables. I also tried using a log transformation on the price variable, but the results, although cleaner, were similar to the above plot and didn't yield a linear relationship of any kind.

Finally, to explore the relationship between the quantitative and categorical variables, I plotted faceted histograms and comparative boxplots. Of all the histograms plotted, the relationship between furnishing and price looks promising, with the following histogram:



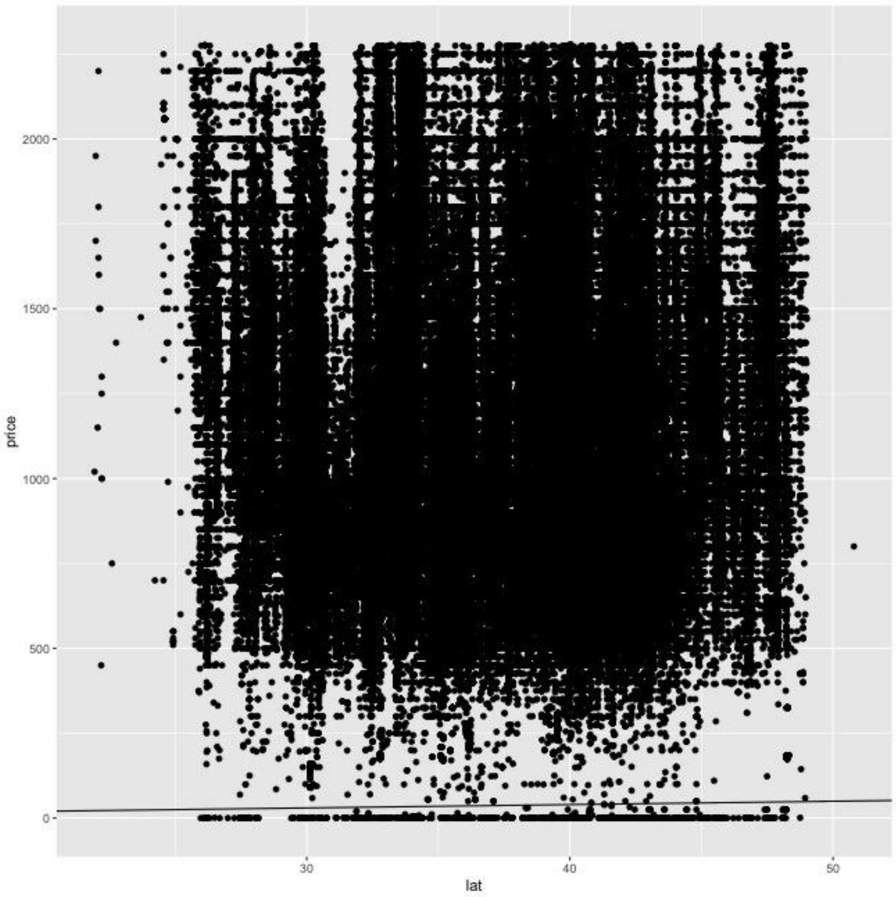
Both histograms are unimodal and right-skewed. The histogram for furnished units has a much higher mode but is also more spread out other than that. The histogram for not furnished units has more skew and is skinnier.

To further explore this relationship, I plotted a comparative boxplot for this relationship:



Unexpectedly, the median price for furnished units is actually lower. However, there is more variability and spread in the price of furnished units than unfurnished, as can be seen from the Inter Quartile Range. There are also quite a few outliers for the not furnished box plot, which probably caused the higher median that we did not expect.

Finally, since the regression for Part 3 would not work out with the *beds* variable, I went back and plotted a quick scatter plot for the *lat* variable and price:



There is a *very* roughly linear relationship between these variables, but I should be able to carry out a regression analysis on this.

Hypotheses and Identified Analysis

After discussing the results of all the individual Exploratory Data Analyses, we developed the following questions that we would like to explore:

<u>Timila</u>

- Regression: Latitude can be used to predict price (rent per month).
- Hypothesis test: Furnished units have lower prices than unfurnished units, on average (two-sample z-test for difference between means).

<u>Matthew</u>

- Regression: Square footage can be used to predict price, based on whether or not dogs are allowed
- Hypothesis test: Bigger housing units per square footage are more likely to allow dogs, on average (t-test for difference between means)

Mikayla

- Regression: Square Feet can be used to predict Price (rent per month)
- Hypothesis test: flat and loft housing are more likely to have wheelchair accessibility than other housing types, on average. (two-sample t-test for difference in means)