# Regression Analysis of Manhattan Apartment Rent

Noah Bradder
December 5,2022

## Introduction

Zillow is a real-estate marketplace company that features millions of for-sale and rental listings. Zillow provides their own valuation tool called the Zestimate which is a model built by Zillow used to produce an estimate for the value of a home. Zestimate has different accuracies depending on the location and is calculated using the properties square footage, location, number of bathrooms, market trends, as well as other details. The Zestimate is only as accurate as the data is detailed and up to date. The national median error-rate using Zestimate is 3.2% for active listings and 6.9% for off-market listings.

The focus of my analysis is to create my own model to predict property listing rent using a multiple regression model that would be simpler but similar to the model that companies like Zillow use to produce their Zestimates.

## Data

The data was not readily available from Zillow itself to do my analysis so I resorted to using a different dataset. My eyes were originally on the UCI Machine Learning "Apartments for Rent Classified" dataset but that dataset was missing from the site so instead I used the "Manhattan.csv" dataset provided by StreetEasy. StreetEasy is a real estate marketplace containing thousands of listings for both purchases and rentals in New York City.

The dataset contains the columns 'rental_id', 'rent', 'bedrooms', 'bathrooms', 'size_sqft', 'min_to_subway', 'floor', 'building_age_yrs', 'no_fee', 'has_roofdeck', 'has_washer_dryer', 'has_doorman', 'has_elevator', 'has_dishwasher', 'has_patio', 'has_gym', 'neighborhood', 'borough'.

A tibble: 6 x 18

| rental_id <dbl> | rent <dbl> | bedrooms <dbl> | bathrooms <dbl> | size_sqft <dbl> | min_to_subway <dbl> | floor <dbl> | building_age_yrs <dbl> | no_fee <dbl> |
|---|---|---|---|---|---|---|---|---|
| 1545 | 2550 | 0 | 1 | 480 | 9 | 2 | 17 | 1 |
| 2472 | 11500 | 2 | 2 | 2000 | 4 | 1 | 96 | 0 |
| 2919 | 4500 | 1 | 1 | 916 | 2 | 51 | 29 | 0 |
| 2790 | 4795 | 1 | 1 | 975 | 3 | 8 | 31 | 0 |
| 3946 | 17500 | 2 | 2 | 4800 | 3 | 4 | 136 | 0 |
| 10817 | 3800 | 3 | 2 | 1100 | 3 | 5 | 101 | 0 |

6 rows | 1-9 of 18 columns

This dataset was my primary resource for conducting my analysis but a variety of other sources were used to conduct parts of the analysis such as testing for multicollinearity.

## Research Questions

There were a couple questions I wanted to address going into this analysis. They were:

- What is the relationship between square footage and price of an apartment?
- What is the relationship between bedroom count and price of an apartment?
- What is the relationship between bathroom count and price of an apartment?
- How do the variables correlate with each other?

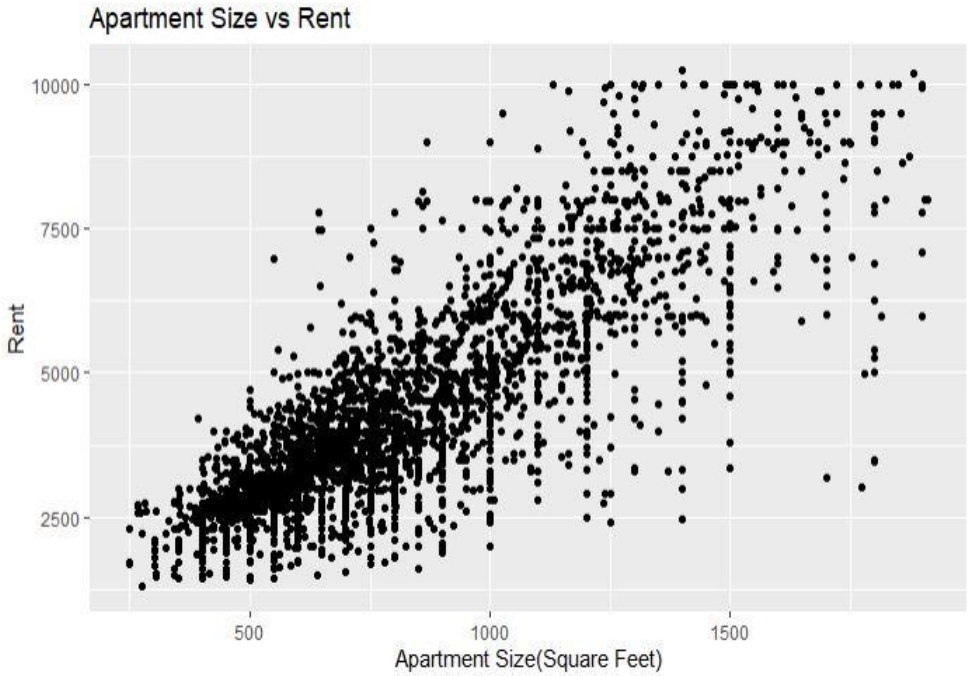
## Exploratory Analysis

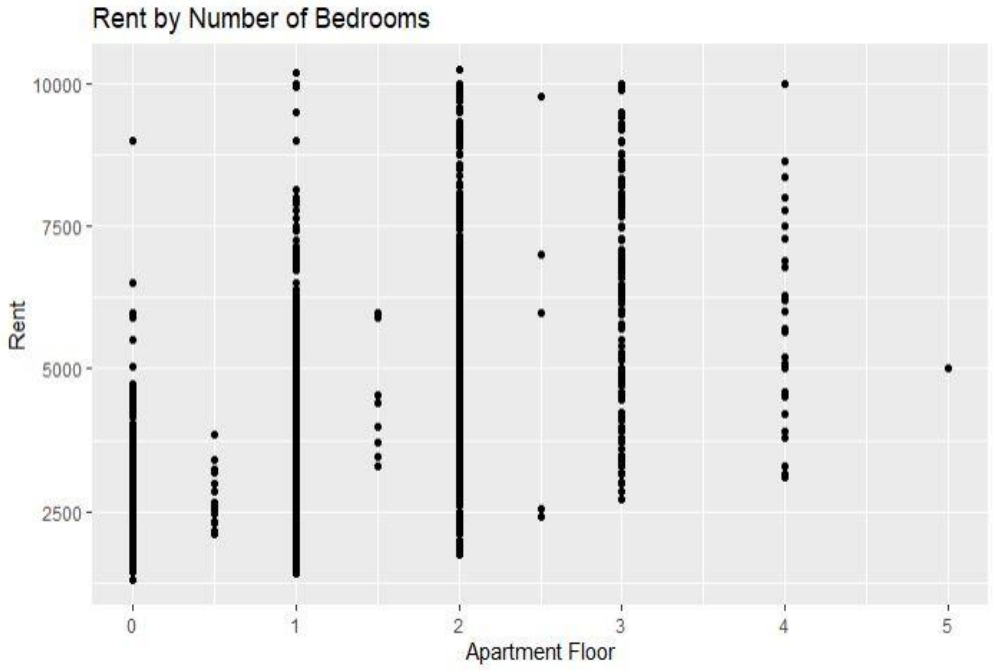
Looking at the data for the first time, my eyes were on conducting a regression analysis with either linear regression or multiple linear regression. To decide which technique to build a model around, I first had to observe and do some exploratory analysis on the data.

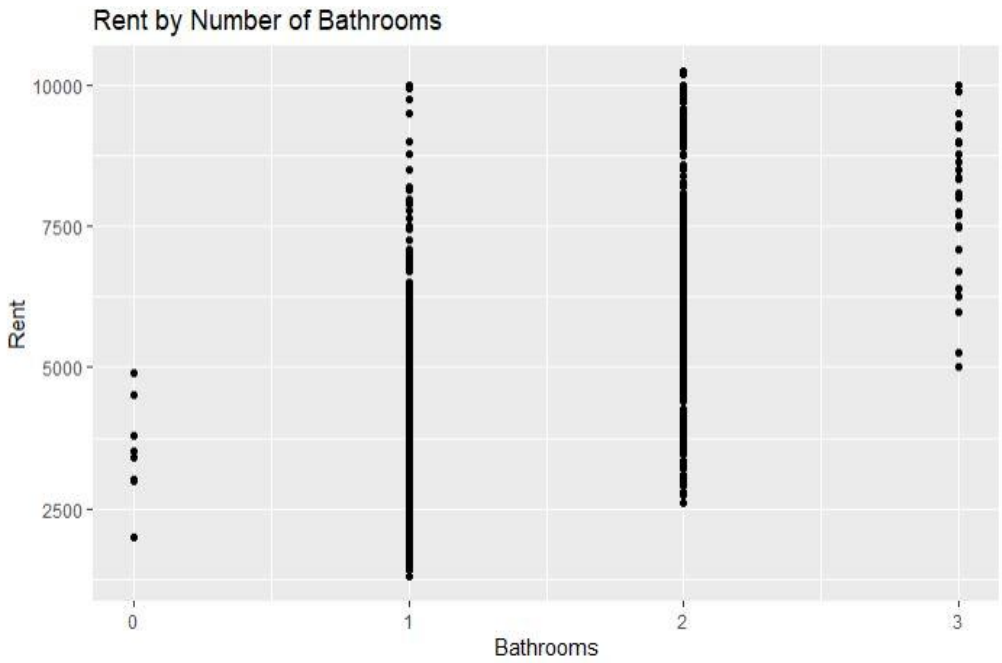
I started by looking at all the data provided in the dataframe and then summarizing each variable to look at their summary statistics. After creating boxplots to check for outliers in the dependent variable I decided to remove those outliers from the data. I decided to only remove outliers from the ‘rent’ and the ‘size_sqft’ columns and then I proceeded with the analysis.

I knew the direction I wanted to take my project and that was predicting the rent of apartments by either doing linear regression analysis or multiple linear regression. I checked the distribution of the rent prices for normality before and after removing outliers. It started skewed left and then after removing outliers was slightly skewed left. I expected this as my result due to high variation in apartment and real estate pricing especially in a city like New York City.

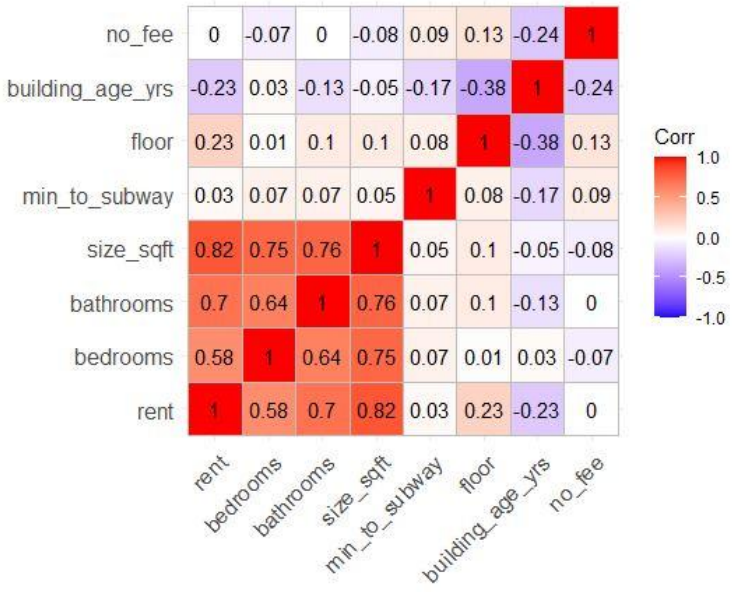
I then checked for linearity. Doing prior research on what determines the price of a unit, I decided I wanted to specifically look at the square footage, the number of bedrooms, and the number of bedrooms. I plotted all of these for linearity. Square footage had a pretty obvious positive trend when graphing, while the plots of bedrooms vs rent and bathrooms vs rent did not show any clear trend for linearity.


Apartment Size vs Rent
10000
7500
5000
2500
Rent
500
1000
1500
Apartment Size(Square Feet)


Rent by Number of Bedrooms
10000
7500
5000
2500
Rent
0
1
2
3
4
5
Apartment Floor




To look at the relationships between the variables more clearly I built a correlation matrix for all of the numeric variables. I plotted the correlation matrix in a corrplot using ggplot to easily see the relationships. As suspected the variables with the highest correlation to rent were square footage (0.82), bathrooms (0.70), bedrooms (0.58). These are the variables that I decided to use in my final model.



## Regression Modeling

The correlation matrix I had built showed me that there were three variables that correlated the most with my independent variable. Those were ‘size_sqft’, ‘bedrooms’, and ‘bathrooms’. With this information I decided to include these variables in my model as a multiple linear regression model.

```r
library(caret)

index <- createDataPartition(clean_df$rent, p=.70, list=FALSE)
train <-clean_df[index,]
test <- clean_df[-index,]
```

```r
library(car)
model <- lm(rent~size_sqft+bathrooms+bedrooms,data=train)

summary(model)
```

## Multicollinearity

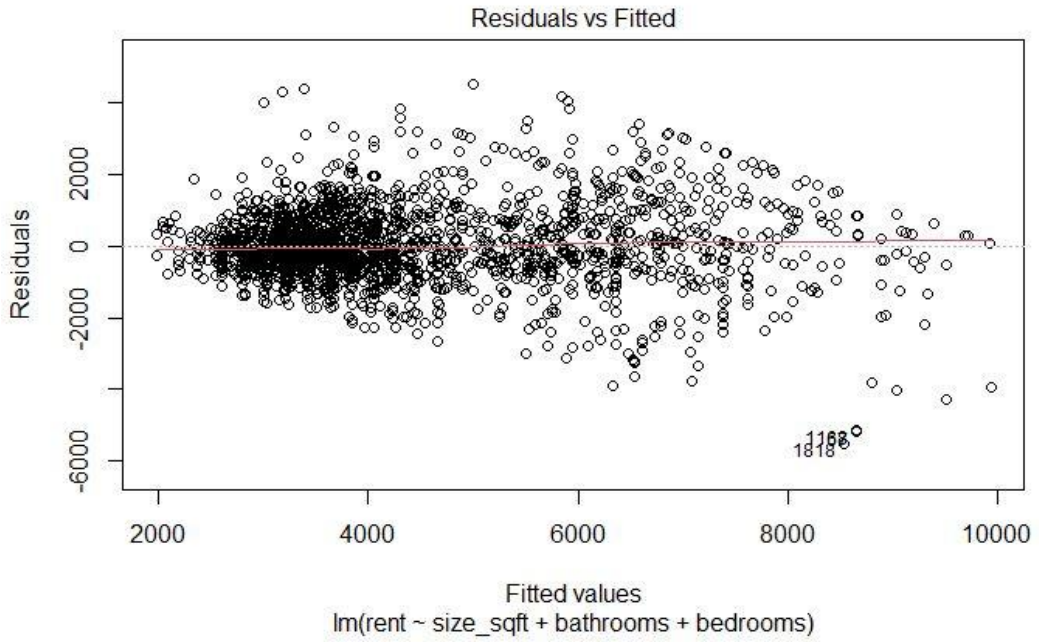
To address the possibility of multicollinearity in my model I used VIF aka Variance Inflation Factor to determine if the variables in my model have a high enough correlation to the point where it is negatively affecting the reliability of my model.

```
size_sqft bathrooms  bedrooms
 3.297335  2.413418  2.363764
```

A VIF score of over 10 indicates high multicollinearity and a VIF score of over 4 indicates possible multicollinearity. Checking the multicollinearity of the values in my model, the scores checked out as size_sqft = 3.297, bathrooms = 2.413, bedrooms = 2.363. All of the scores are lower than the acceptable levels showing that multicollinearity doesn’t exist in the model.

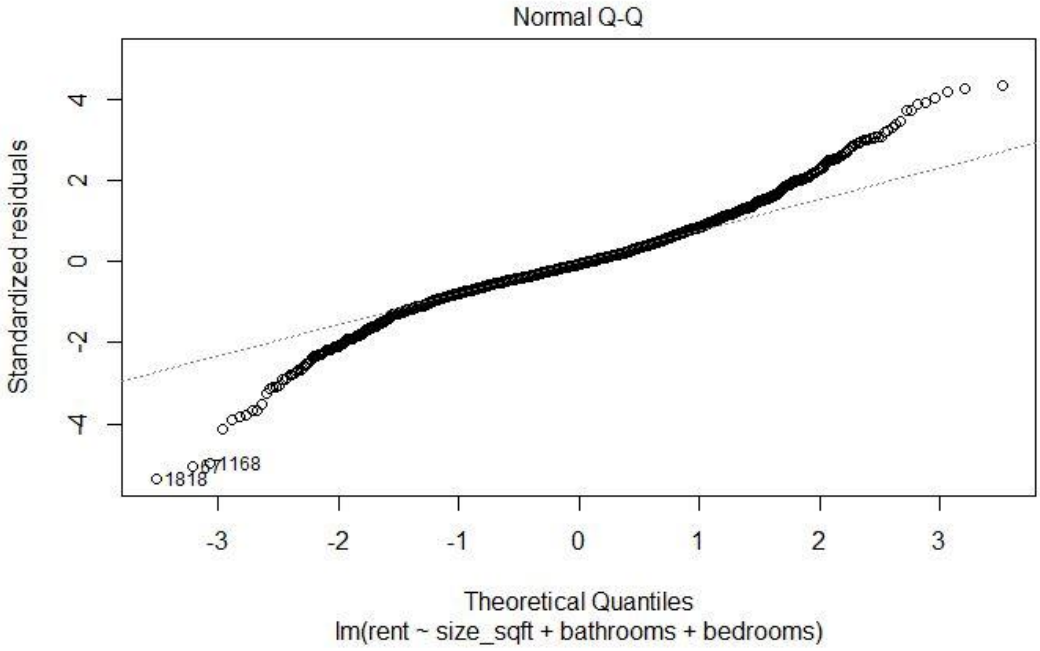
## Results

I was able to analyze the results of my model by both summarizing the model and plotting the model as well. Out of the plots the first I analyzed is a Residuals vs Fitted plot.




This plot gives multiple insights into the model performance. The red line through the plot follows a very horizontal trend showing that linearity exists in the model. Though that is the case the one issue present in this plot is that the spread of the residuals expands as the graph continues. This shows that some level of heteroscedasticity exists which breaks the rule of homoscedasticity and makes our model unreliable. At the bottom of the plot we can see two points highlighted that are possibly negatively influencing my model.

The next plot from my model is the Normal Q-Q plot:




This plot shows the normality of my model. My model follows the assumption of normality near the middle values but the model lacks normality as the data reaches really low and really high values. This negatively affects the reliability of my model. A log transformation could be added in the future to fix these issues.

Lastly, the summary of my model gave me multiple insights on the model such as my multiple linear regression equation and the accuracy of my model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.0607    69.1185   0.956    0.339
size_sqft      4.2034     0.1245  33.759  < 2e-16 ***
bathrooms    865.7271    71.9952  12.025  < 2e-16 ***
bedrooms    -239.0421    39.4651  -6.057 1.62e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1035 on 2256 degrees of freedom
Multiple R-squared:  0.6825,    Adjusted R-squared:  0.6821
F-statistic:  1616 on 3 and 2256 DF,  p-value: < 2.2e-16
```

From this summary we can derive the linear regression equation:

**Rent ≈ 66 + 4.2 x size_sqft + 865.7 x bathrooms + -239 x bedrooms**

We can also interpret the r-squared value as well as the residual standard error. The adjusted r-squared value produced by my model was 0.68. This basically means that 68% of the data can be explained by the model which is a pretty poor performance. A good performance would be an r-squared value of 0.85-1. On the other hand my residual standard error is 1035 giving me an error rate of about 23.3%. I calculated this percentage by taking the sigma of my model and dividing it by the mean rent in my cleaned dataframe.

## Conclusion

In general the model has some serious reliability issues that need to be addressed. The initial models that I had built were less reliable than the final product which is an improvement but still not great. A good way to improve upon my model is to code all the have or have not variables into binary to be able to use them in a regression classification. I think the model would be more reliable and tailored to different variables that people look for when renting apartments.

## Sources

https://www.investopedia.com/articles/personal-finance/111115/zillow-estimates-not-accurate-you-think.asp

https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/