MGT 6203 Team #21 Final Report

D. Han, A. Karwoski, S. Kwon, J. Place, R. Weber

[Overview of Project 2](#_Toc12255419)

[Background 3](#_Toc1717667306)

[Approach 3](#_Toc2015005060)

[Initial hypotheses 3](#_Toc1963822166)

[Overview of Data 3](#_Toc1457380444)

[Cleaning Process 4](#_Toc1428137191)

[Exploratory Data Analysis 5](#_Toc200148865)

[Overview of Modeling 6](#_Toc892175716)

[Final Notes 10](#_Toc1026483121)

[Conclusion 11](#_Toc2130184345)

[Works Cited 11](#_Toc1504681027)

[Appendix 12](#_Toc2001092523)

[Data Sources 12](#_Toc1552171277)

[Realtor.com Housing Price dataset 12](#_Toc1869202166)

[US County Demographics dataset 12](#_Toc974530420)

[US County Crime dataset 12](#_Toc930364427)

# Overview of Project

## Background

One of the largest purchases one can make is to buy a home, and houses are more expensive than ever. In order to make sound real-estate decisions, one must be able to quickly and accurately estimate the price of a home, either to maximize their return on a sale or to minimize overspending when purchasing. To minimize the influence of human factors, we wanted to generate a model which would estimate the purchase price of a home, based on construction, location, and locality information. This model will allow us to determine the value of a home as compared to other homes sold in a relatively close geographic area.

## Approach

The base problem is to generate a model which will predict housing prices. Since we desired a continuous response and expected to have consistent inputs, we decided to pursue generating linear models for our predictions.

Our approach was to combine housing data with statistics based on crime and demographic information, and then filter that data down to a single US state (New York) to ensure a manageable scale. Next, we evaluated parameters and determined which would be best modelled as categorical or numerical, as well as what type of scaling/normalization should be performed. We fit several linear models against the data, evaluate the level of impact from each parameter, and evaluated what (if any) permutation of those parameters should be performed. Our final linear model provides a reasonable estimate of the housing market without providing any location information.

## Initial hypotheses

Our initial hypothesis was that housing models abstract away many environmental factors on prices by generalizing on postal code. We expected that what people describe as a ‘desirable location’ is in fact influenced by multiple factors, such as the age makeup of the locals, income and poverty levels, crime rates, and others. We expected to find some or all of these factors correlated with housing prices.

# Overview of Data

## Cleaning Process

Cleaning our data required significant effort on our behalf. Firstly, we started with three distinct data sets, which did not include a common identifier. Additionally, one of our data sets (demographic information) was several gigabytes in size and included several thousand columns. In order to successfully merge the data, we split the process into three stages: restrict our demographic data to a single state, grab only the latest version of the information provided in the data set, and then finally join our ancillary data sets on US county into a single set which could be joined to the realtor data based postal code.

Once we combined our data into a single data set, we needed to perform additional cleaning. We found that there were significant amounts of data on housing missing, both intentionally (there were empty lots in the data set) and unintentionally (the data was mis-scraped), there were outliers in the data which would throw off analysis, and there was an issue where was not a clean separation between postal codes and counties, leading to some data duplication.

In order to alleviate these concerns, we eliminated duplicate data rows by selecting the county which was more populated and dropping the less populated one. It was assumed that the more populated counties would dominate regarding crime and demographic data. Outliers were eliminated based on having pricing data +/- three σ from the mean. Rows with excessive missing data were deleted, and the remaining missing values were imputed using the median value of the column. After cleaning, it was determined that our cleaned data set was large enough for initial modelling without having to find additional data.

From our cleaned data set, we performed stepwise regression and then linear regression, and then examined the summary of our linear models to determine which variables were most important. Of all our variables, the key variables found were: number of baths (bath), number of beds (bed), lot size (acre\_lot), and our generated boolean feature for listings that were plain lots (lot\_flagTrue).

Our data was provided via kaggle.com, with parent data being scraped from realtor.com (for housing data), ICPSR (crime data), and USDA/US Census data (demographics). Links to the kaggle.com data sets are in the appendix, with links to the parent data found on the kaggle.com pages.

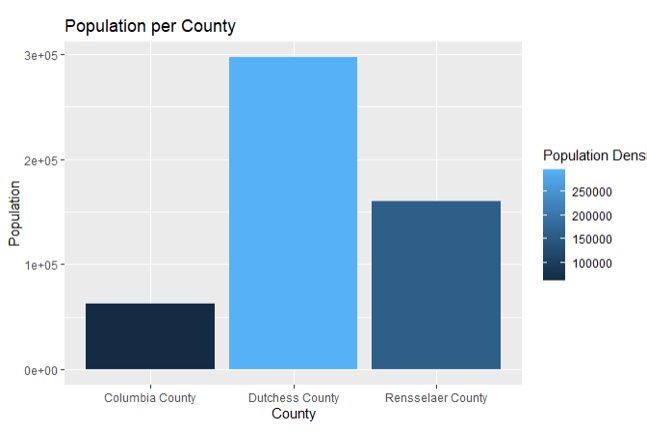
After evaluation and feature removal, our final cleaned data consisted of the following features:

|  |  |
| --- | --- |
| Feature | Description |
| Price | Our dependent variable, which we were attempting to find analytically |
| Bed | # of bedrooms |
| Bath | # of bathrooms |
| Acre\_lot | Size of home lot, in acres |
| Zip\_code | Property postal code (as a factor) |
| House\_size | Home size, in sq. ft. |
| Lot\_flag | Boolean, does property have only land (is lot) or no (is lot + home) |

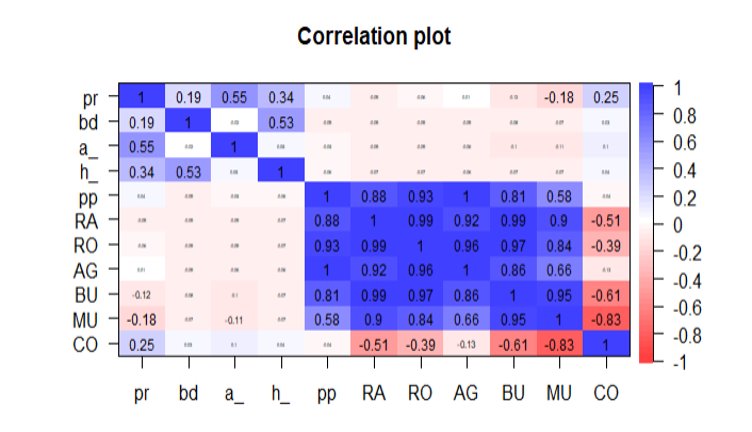
The final cleaned data set consists of approximately 600 data points.

## Exploratory Data Analysis

Our EDA revealed that in the process of cleaning the data, most New York counties had their listings entirely removed. We ended up having 3 counties: Columbia, Dutchess, and Rensselaer. This confirmed our previous assumption that the realtor data had not been scraped properly and impacted our results. Additional results showed that the counties with listings were relatively high in population, which explained why they had many postings which were properly formatted.

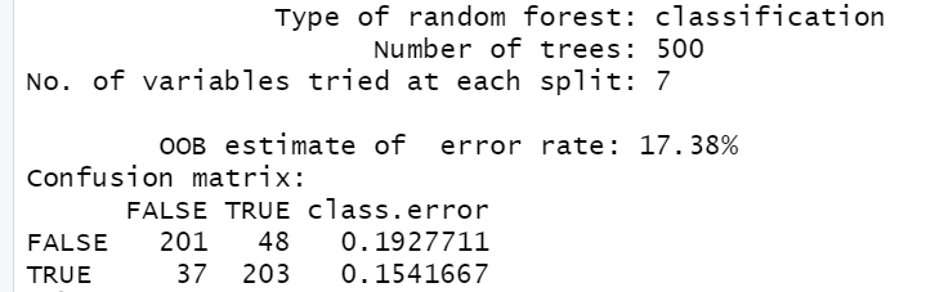


Another feature found during EDA was that even though our data came from multiple sources, many features had high correlation between them. This led to many features being dropped in order to improve model performance.



# Overview of Modeling

We used Random Forest models to split our homes into expensive and inexpensive groups.

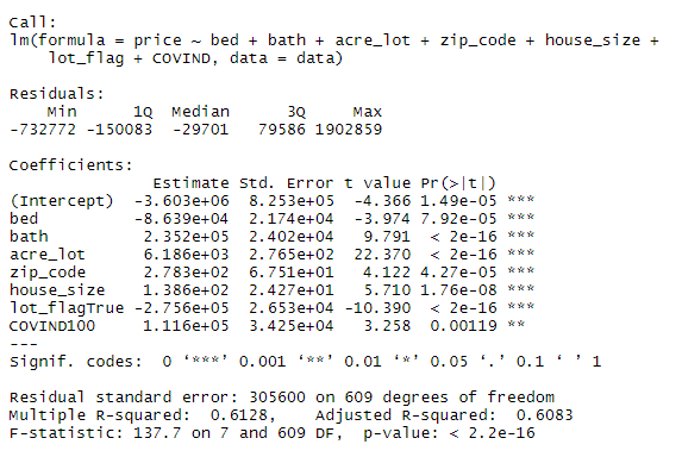


We split the dataset into two groups, based on if the price was greater or less than the median home price. Fortunately, the datasets split into almost exactly even sizes, and we used a train-test split of 80/20 for the training and testing of our random forest model.

This was performed because of an intuition that buyers in certain price ranges may be looking for features that may not be important to buyers in the other. Our RF model was capable of splitting the homes into these categories with an accuracy of ~82.63%, indicating that there was a definite split in features between the two. The model was also better at classifying inexpensive houses more accurately than expensive houses. The inclusion of 500 trees helps reduce the bias introduced by the training dataset.

We used these results to generate two linear regression models and then compared results to see which was more effective at predicting house prices overall.

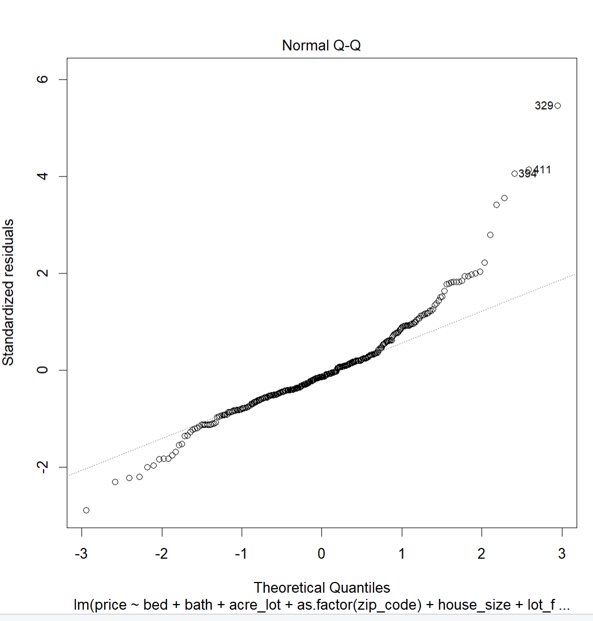
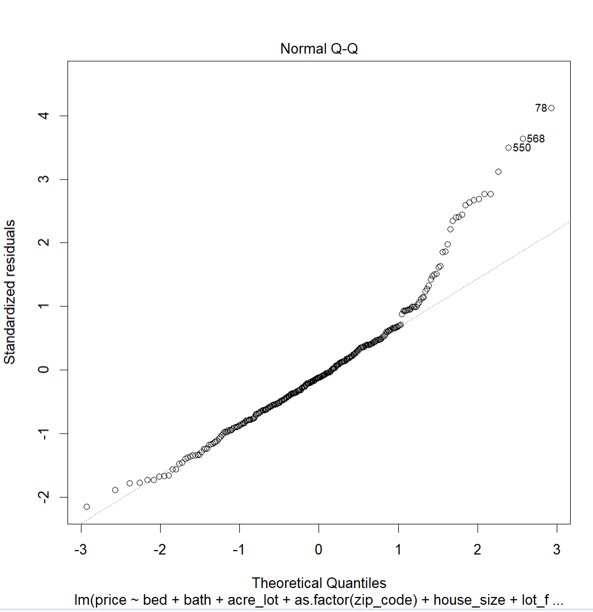
We additionally performed stepwise regression on our dataset, in order to determine which features were useful to both models, so that we could compare the significance of these features between the final two models.



This was one of the more interesting points in our experiment, because although we had included many crime and demographic features, they were all removed via this selection process as they all correlated strongly with zip codes. Even more interesting was that when zip codes were factorized and tested in our regressive models, they were largely unimportant, suggesting that our initial hypotheses was entirely incorrect. However, a selection of the zip codes is statistically significant, meaning some of the zip codes have much different crime and demographic features than most of the others.

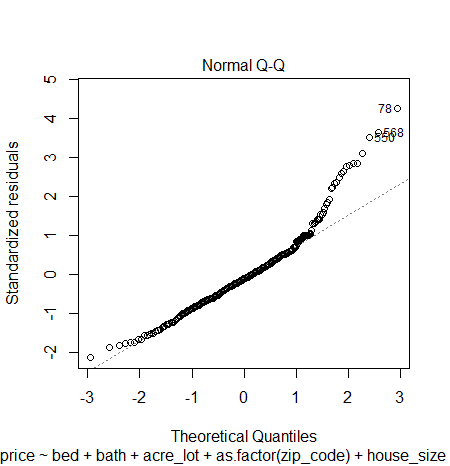
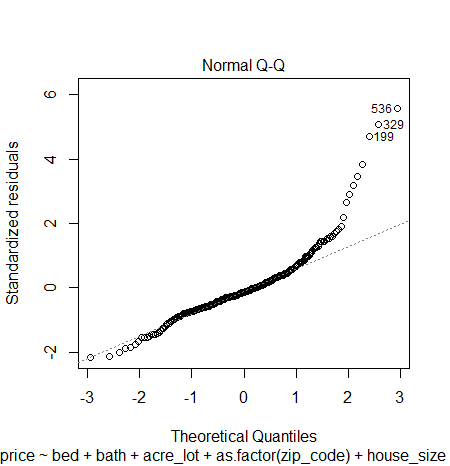
For our main model, we generated two standard linear regression models. We generated two of these, based on the results of our splitting the data set into Expensive home purchases and Inexpensive home purchases. The data was split using a Random Forest model, as described above. Using the predicted classes, we split the dataset into the Expensive and Inexpensive classes. Fortunately, both predicted classes had almost the exact same amount of datapoints, one having 308 datapoints, and the other having 309 datapoints.

The two models performed roughly similar in terms of R-squared values; however, when we examine the Q-Q plots of each:



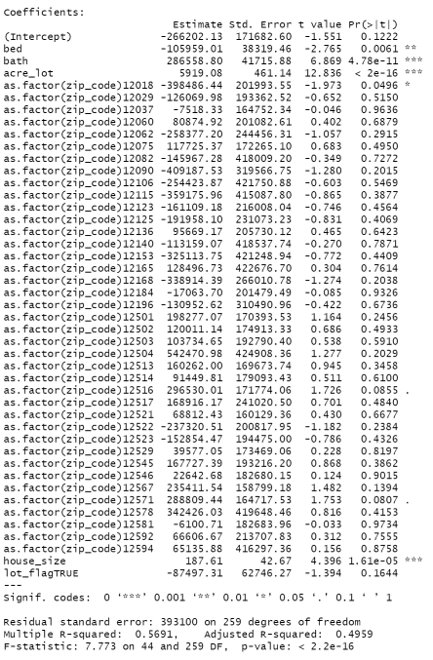
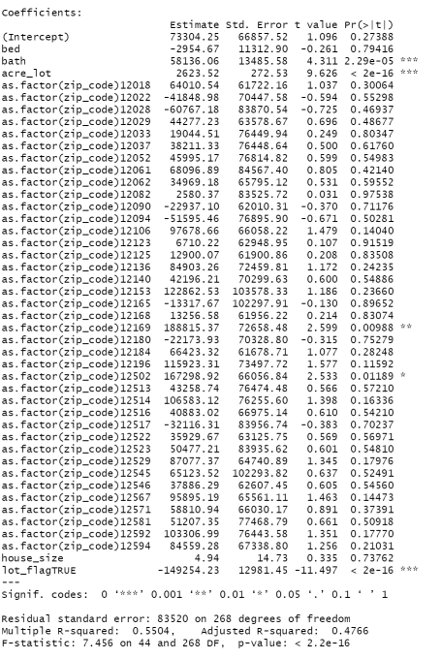
As we can see the residuals do not appear to follow the normality constraint and constant variance assumption. Therefore, for our final models we performed a log transformation using the box-cox method to find the optimal lambda. The optimal lambda value was 0 meaning log(price) was the transformation we used for the final two models.

Below are the QQ plots for the inexpensive and expensive models with the transformed response. The Inexpensive model is on the left.



The transformations helped satisfy the normality and constant variance constraints making the data follow the distribution more closely. Accepting these is our final models we created two new linear regression models using the variables selected in the stepwise regression algorithm.

Below are the outputs of our linear regression models with inexpensive on the left.



The R squared of these models is slightly lower than the R squared of the original stepwise model. However, the inexpensive model has a substantially lower RSE of 83520 compared to 305600 seen in the stepwise model. It is also shown that the RSE of the expensive model is higher in the final model than in the original stepwise model using the full dataset. This is due to the high variation in housing prices at the right tail of expensive home prices. These extreme values cause the model to underperform at higher values.

Overall, both models have an overall p-value of ~0, meaning at a model level, the predictors are significant in determining the response variable, price. It would be interesting for future observation to run an Anova test with a less complex model to see if any of the predictors are significantly 0.

These observations show that the models we produced are much better at predicting lower housing prices. This is evidenced both by the higher accuracy in the Random Forest Model and in the Linear Regression Model at classifying and predicting inexpensive houses respectively. We accept these models, regardless of the R square, because this decrease is likely due to the higher influence of tail values as a result of the lower number of datapoints.

We also use these models to analyze the variables that are important in each model. As we can see, for inexpensive houses, the number of beds and house size are not significant, but for the expensive models these variables are significant with 95% and 99% confidence respectively.

# Final Notes

We originally wanted to generate pricing estimates based on data which was not focused on postal codes, as our original hypotheses was that demographic / locality features would determine which postal codes were desirable. However, we were not able to determine which demographic features were highly correlated with home value. Because of this, we were forced to pivot and change how we developed our model.

One area we would have liked to examine had we more time and computational resources would have been to develop our own method of scraping, instead of being reliant on the Kaggle dataset. We found numerous issues with the Kaggle realtor data and believe that we could have achieved a much higher level of accuracy if the amount of data were higher.

Another area of possible improvement would have been to examine more closely the demographics data, and generation of features from that data. Unfortunately, due to the way that the data was structured, and our limited computing resources, we were not able to get as in-detail as we would have liked with this dataset.

# Conclusion

Home price prediction is very difficult to achieve, especially when considering homes with wide ranging price values. Although we overcame these difficulties and found some measure of success, we would have liked to have seen higher accuracy in our models.

We added additional data to our set to help us find insights but ended up removing most of it when our feature selection process showed that it wasn’t useful. Our key takeaway from this point was that more data does not equal better data, or more useful. If we had realized this sooner, we could have focused on generating features from our parent data set and achieved more accurate results.

Web-scraped data is prone to having missing values and bugs. Although there was enough data to perform analysis, if the data had been better scraped, we could have avoided many cleaning steps. Even better than scraping would have been to find an offered API for home prices, such as going directly to the MLS listings, and pulling the data straight from the source. This would have eased cleaning requirements and enabled us to make better models.

# Works Cited

Ceccato, V., & Wilhelmsson, M. (2019, September 12). *Do crime hot spots affect housing prices?* Taylor & Francis. Retrieved July 4, 2022, from <https://www.tandfonline.com/doi/full/10.1080/2578983X.2019.1662595>

Rockoff, J. E., & Linden, L. (2008). *Estimates of the impact of crime risk on property values from Megan's laws*. American Economic Review. Retrieved July 4, 2022, from <https://www0.gsb.columbia.edu/faculty/jrockoff/aer.98.3.pdf>

# Appendix

## Data Sources

### Realtor.com Housing Price dataset

<https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset>

### US County Demographics dataset

<https://www.kaggle.com/datasets/bitrook/us-county-historical-demographics?select=us_county_demographics.csv>

### US County Crime dataset

<https://www.kaggle.com/datasets/mikejohnsonjr/united-states-crime-rates-by-county/code>