NYC Apartment Sales Bayesian Regression Modeling Report

Joo Kim

11/24/2018

1. Introduction & Data

For the research challenge, I chose the New York City Property Sales data, which lists properties that sold between October 2017 and September 2018 in the city's five boroughs: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. With the data, I aimed to explore how the property sale prices differ by borough. Property sale prices can be seen as a marker of economic strength and development potential, but they also betray economic disparities. Studying the differences in property sale prices can shed light on the gaps in development and help locate areas of inequality, so city officials and urban planners may allocate more funding to boost economic activity as well as provide support such as subsidized housing to low-income residents living in boroughs with higher levels of inequality.

The datasets are publicly available through the New York City Department of Finance and include the following variables (those added to the model are bolded):

- **Borough**
- Neighborhood
- Building class category
- Tax class at present
- Block
- Lot
- Ease-ment
- Building class at present
- Address
- Apartment number
- Residential units
- Commercial units
- **Total units**
- **Land square feet**
- **Gross square feet**
- **Year built**
- Tax class at time of sale
- Building class at time of sale
- **Sale price**
- Sale date

The full dataset has 81947 observations and includes individual apartment units as well as entire buildings used for various purposes (e.g. civic, industrial, residential). I expect building type to affect sale price, but given time constraints and for ease of interpretation, I will be limiting my scope and analyzing only entire residential buildings. This precludes generalization of my findings to all properties in New York but will increase the precision of my estimates, which may be applied to understand price differences of residential buildings between boroughs in New York City.

2. Hypothesis

My hypothesis is that property sale prices differ by borough. In particular, I predict that as the land square footage of properties increase, the rate at which the property sale prices increase differs by borough.

3. Methodology

I will be fitting Bayesian models for this analysis. Bayesian methods quantify uncertainy in the model using probability and simulation; they combine data with prior information, which usually comes from prior research on the subject, to make the coefficient estimates more stable. For this research challenge, I used default priors that are weakly informative (mean 0 and standard deviation 2.5 for all coefficients except the intercept) that come with the 'rstanarm' Bayesian modeling package. Please note that the Bayesian method summaries do not provide p-values. Instead, I will be looking to standard errors to assess if an estimate is muddled by noise.

4. Assumptions

The first assumption I make is that the variables in the data sufficiently explain variation in the housing market. Second, I assume that the data is accurate and recorded mostly without errors; I removed observations that seemed egregiously inaccurate from my analysis. Finally, I make a rather strong assumption that the missing values in the data are "missing at random." Missing values can be quite tricky to work with because they can bias the results. There are several mechanisms for why an observation might be missing, but the assumption that I make here is that the probability that a variable is missing depends only on available information. In other words, the probability of a missing value in sale price depends only on other, fully-recorded variables in the data set. This assumption allows me to model and predict outcomes for missing values using other predictors in the data.

Please note the results do not imply causality. Thus, my analysis won't tell us whether a building in one borough vs another causes an increase or decrease in price. What the results do tell us is whether or not there is a relationship between property price and borough and the magnitude of that relationship.

5. Exploratory Data Analysis and Descriptive Statistics:

I began by exploring the data to understand its structure and the distributions of the variables I am interested in including in my analysis.

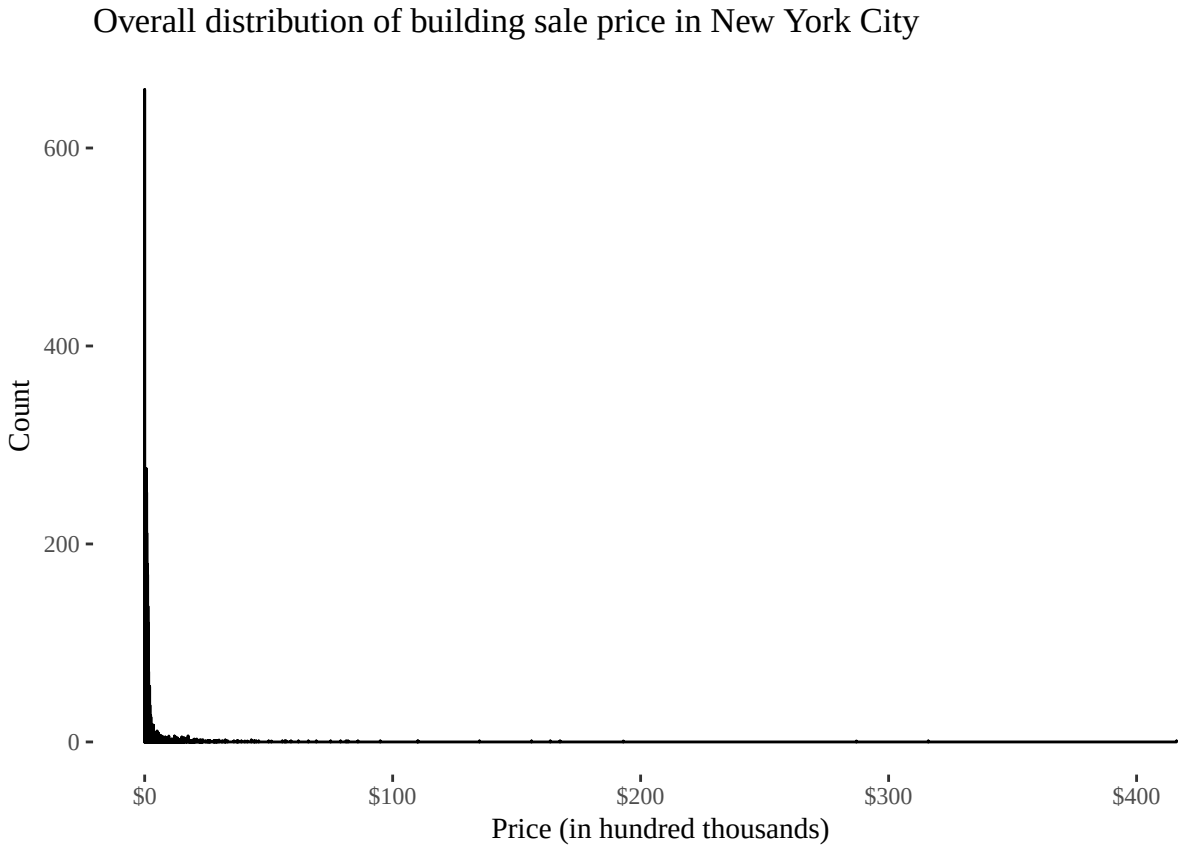
Borough

The filtered dataset, which only includes entire buildings for residential use, has 40606 observations and is divided by borough as follows:

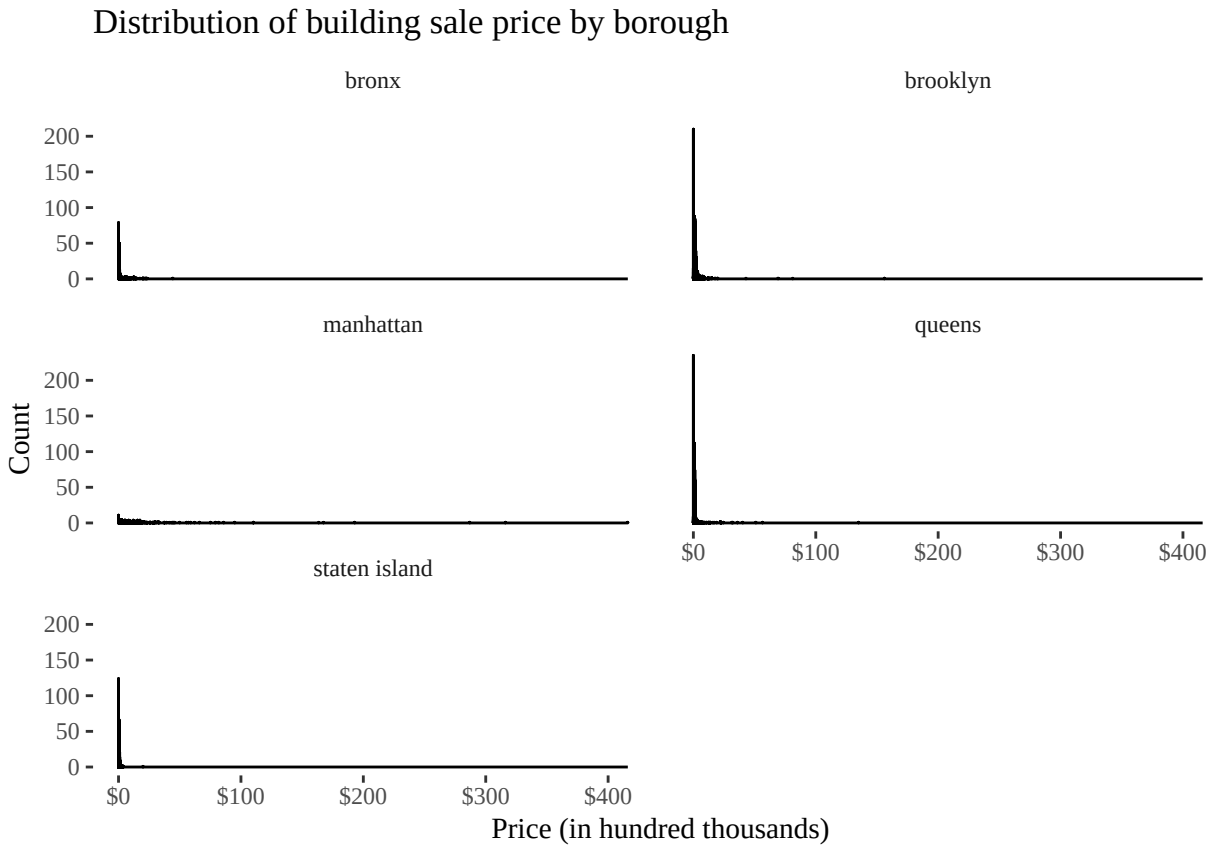
```
## # A tibble: 5 x 2
##   borough       count
##   <chr>         <int>
## 1 queens        15842
## 2 brooklyn      11900
## 3 staten island  6845
## 4 bronx          4869
## 5 manhattan      1150
```

Property sale prices

Property prices range from 1 to over 416 million dollars. The middle 50% of the values fall between 450,000 and 920,000 dollars. The below histogram illustrates the distribution of sale prices across all five boroughs.

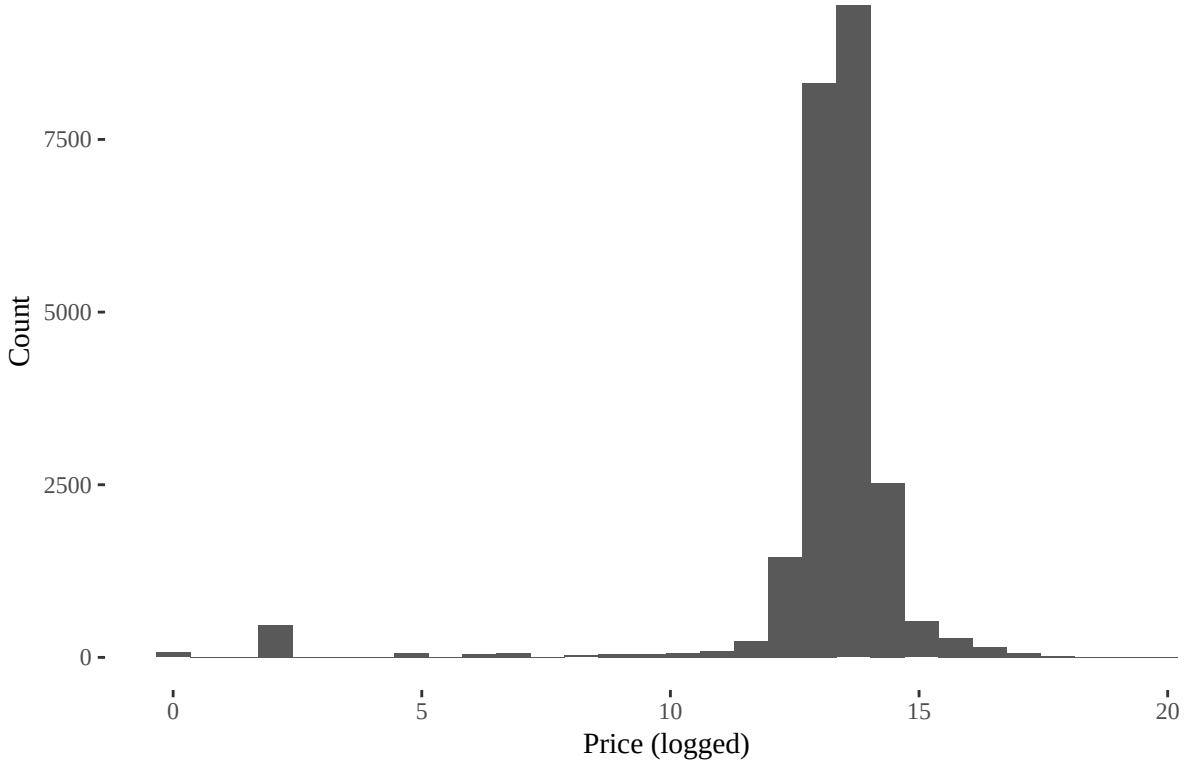


The graph on original scale shows that the distribution of the property prices is expectedly right tailed. By borough, the trend is similar, but Manhattan's property prices are relatively evenly distributed when compared to prices of buildings in other boroughs.

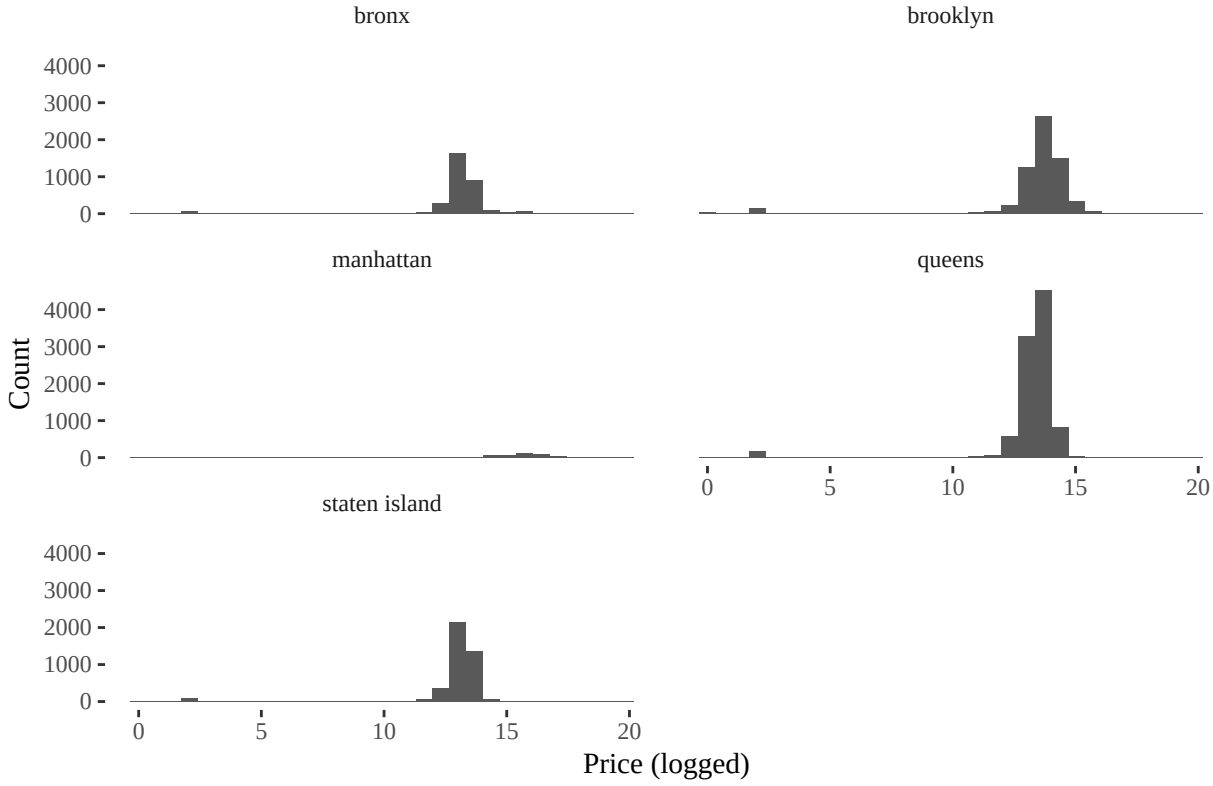


For modeling, I logged the sale prices. Logarithmic transformations are appropriate when dealing with outcomes like sale prices, which are necessarily all positive. Furthermore, they are often used when additivity and linearity are not reasonable assumptions. Below are the distributions of the logged property prices in the five boroughs, for reference.

Overall distribution of building sale price (logged) in New York City



Distribution of building sale price (logged) by borough



Finally, I've found that there are values of sale prices that are infeasibly low. I don't know a lot about real estate, but I know that $1 for a property in New York is not reasonable! I am going to keep them as is, but erroneous records like these can severely bias the model and the coefficient estimates. In addition, there are 13,836 missing sale prices (about 34% of the observations) in the data. As mentioned above, missing values also bias the model. There are several ways of handling missing values, but being mindful of time, I applied a simple deterministic imputation, in which I regressed observed sale prices on the predictors to impute sale prices of the missing records. The advantages and disadvantages of this approach are outlined at the end of the report.

Other variables of interest

- Land square feet ranges from 1 to 3217, while gross square feet ranges from 1 to 3053. I suspect these are also erroneously recorded. For simplicity's sake, I removed observations that have less than 100 square feet for land square and gross square footage, which amounted to about 9% of all observations. Another way to remedy the issue would be to impute these observations. I've also standardized the two variables by subtracting their respective means and dividing by their respective standard deviations to get more interpretable models. Standard deviations can be seen as a measure of practical significance; measuring the effect of increasing the square footage by 1 is not as meaningful as the effect of increasing it by the variable's standard deviation. The means of land square feet and gross square feet are 1186.0 and 959.8, while the standard deviations are 659.5 and 698.1, respectively.
- The "units" variable tells us the total number of units in the property. I centered the units by subtracting its mean (19.2) for ease of interpretation of the coefficients.
- In addition, the data contain buildings that were built between 1800 and 2018. This variable, "buildage," likely affects sale prices because newer buildings tend to command higher prices. There are 90 observations (less than 0.3% of the data) that have 0 for the year built value, likely due to a recording error. These were also removed from the dataset.

The final clean dataset ready for modeling has 36800 observations in total, with a tally for each borough as shown below:

```
## # A tibble: 5 x 2
##   borough       count
##   <fct>         <int>
## 1 bronx          4487
## 2 brooklyn      11065
## 3 manhattan       919
## 4 queens        14611
## 5 staten island  5718
```

6. The Models

First Model

I first start by fitting a simple model, without any interactions.

$$log(price_i) = \alpha + \beta_1 borough_i + \beta_2 landsqft_i + \beta_3 grosssqft_i + \beta_4 units_i + \beta_5 buildage_i + \epsilon_i$$

The model coefficient summaries:

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      price_log ~ .
##  observations: 36800
```

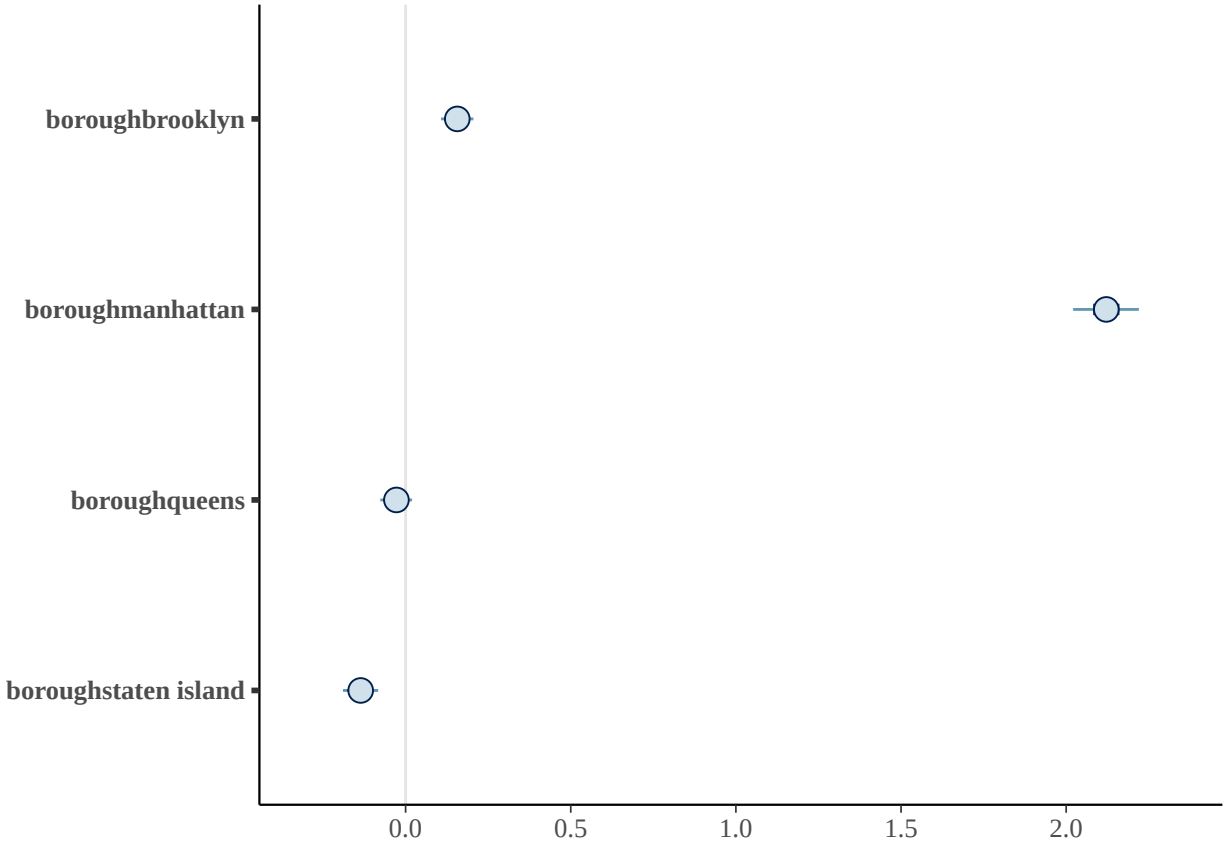
```
##  predictors:   9
## ------
##                       Median MAD_SD
## (Intercept)           13.2    0.0
## boroughbrooklyn        0.2    0.0
## boroughmanhattan       2.1    0.1
## boroughqueens          0.0    0.0
## boroughstaten island -0.1    0.0
## land_z                 0.1    0.0
## gross_z                0.1    0.0
## units_c                0.0    0.0
## years_since            0.0    0.0
## sigma                  1.6    0.0
##
## Sample avg. posterior predictive distribution of y:
##          Median MAD_SD
## mean_PPD 13.2    0.0
##
## ------
## For info on the priors used see help('prior_summary.stanreg').
```

The below table gives a closer look at the coefficient estimates and their standard errors:

```
##                       simple_model_coef simple_model_se
## (Intercept)              13.1648321133    0.0349468758
## boroughbrooklyn           0.1564385250    0.0308195450
## boroughmanhattan          2.1215152154    0.0590633726
## boroughqueens            -0.0278385884    0.0288085855
## boroughstaten island     -0.1360450859    0.0322716363
## land_z                    0.0849465865    0.0087766523
## gross_z                   0.0614101628    0.0092768851
## units_c                   0.0064933119    0.0004700875
## years_since              -0.0007666432    0.0003040254
```

The coefficient of the intercept, 13.16, is the predicted log price of a building built in 2018 with 19.2 units, 1186.0 land square footage, and 959.8 gross square footage in the Bronx. This corresponds to exp(13.16) = 519176.9 dollars.

The coefficient of boroughmanhattan is the predicted difference in log price between the Bronx and Manhattan, holding everything else constant. As expected, the difference between Bronx and Manhattan prices is quite large. The predicted price of a buiding in Manhattan built in 2018 with 19.2 units, 1186.0 land square footage and 959.8 gross square footage is exp(2.12) = 834.38% higher than a building with the same characteristics but in the Bronx. In Brooklyn, the price of a similar building is expected to be exp(0.16) = ~17.0% higher, while exp(-0.03) = ~3% lower in Queens and exp(-0.14) = ~12.7% lower in Staten Island. The coefficient for Queens is rather small with a large standard error, which tells us that the point estimate is uncertain and could change in sign or magnitude if given additional information. The following plots display the point estimates of the variables of interest along with their confidence intervals.



The regression model has a residual standard deviation σ of 1.6, implying that approximately 68% of log prices will be within 1.6 of the predicted value. On the original scale, this means that 68% of price will be within a factor of exp(1.6) = 4.95 of the prediction. For example, a brand new building in the Bronx with 19.2 units, 1186.0 land square footage, and 959.8173 gross square footage is predicted to be between [exp(13.16 - 1.6), exp(13.16 + 1.6)] = [104820 dollars, 2571500 dollars]. It's a wide range and demonstrates that this model does not predict price well.

Improved Model

I then include an interaction of borough and land square footage in the model in an attempt to get more precise estimates.

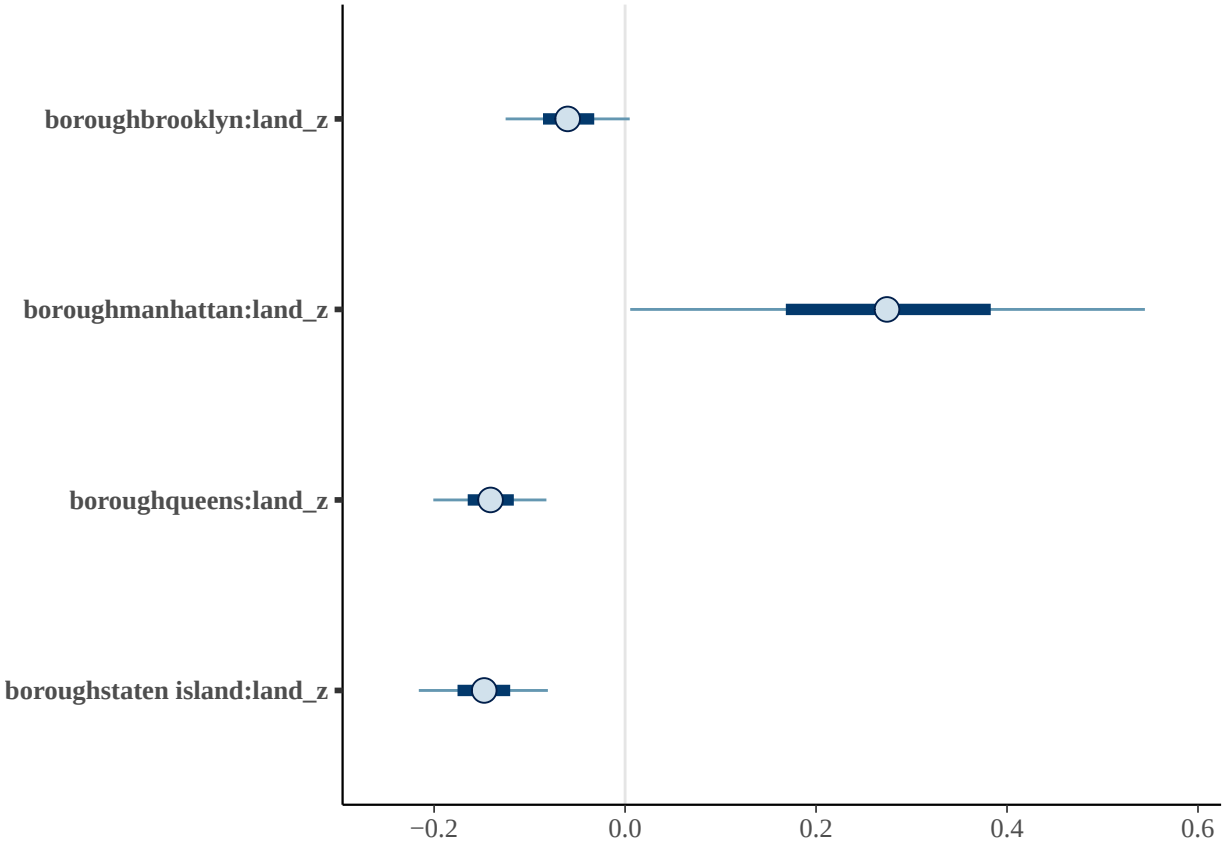
$log(price_i) = \alpha + \beta_1 borough_i + \beta_2 landsqft_i + \beta_3 grosssqft_i + \beta_4 units_i + \beta_5 buildage_i + \beta_6 borough_i * landsqft_i + \epsilon_i$

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      price_log ~ . + land_z:borough
##  observations: 36800
##  predictors:   13
## ------
##                            Median MAD_SD
## (Intercept)                13.2    0.0
## boroughbrooklyn             0.1    0.0
## boroughmanhattan            2.5    0.2
## boroughqueens              -0.1    0.0
## boroughstaten island       -0.2    0.0
## land_z                      0.2    0.0
```

```
## gross_z                      0.1    0.0
## units_c                      0.0    0.0
## years_since                  0.0    0.0
## boroughbrooklyn:land_z      -0.1    0.0
## boroughmanhattan:land_z      0.3    0.2
## boroughqueens:land_z        -0.1    0.0
## boroughstaten island:land_z -0.1    0.0
## sigma                        1.6    0.0
##
## Sample avg. posterior predictive distribution of y:
##          Median MAD_SD
## mean_PPD 13.2    0.0
##
## ------
## For info on the priors used see help('prior_summary.stanreg').
```

```
##                              interactive_model_coef interactive_model_se
## (Intercept)                            13.2120293396         0.0380646843
## boroughbrooklyn                         0.1267418872         0.0328532907
## boroughmanhattan                        2.5364692942         0.1899391625
## boroughqueens                          -0.0647958917         0.0314457914
## boroughstaten island                   -0.1841439134         0.0360826942
## land_z                                  0.1997736644         0.0336435269
## gross_z                                 0.0601673655         0.0089488432
## units_c                                 0.0064446482         0.0004752101
## years_since                            -0.0007648535         0.0003195881
## boroughbrooklyn:land_z                 -0.0601004340         0.0395972802
## boroughmanhattan:land_z                 0.2742771886         0.1583008135
## boroughqueens:land_z                   -0.1408872529         0.0357684405
## boroughstaten island:land_z            -0.1475274137         0.0408173707
```

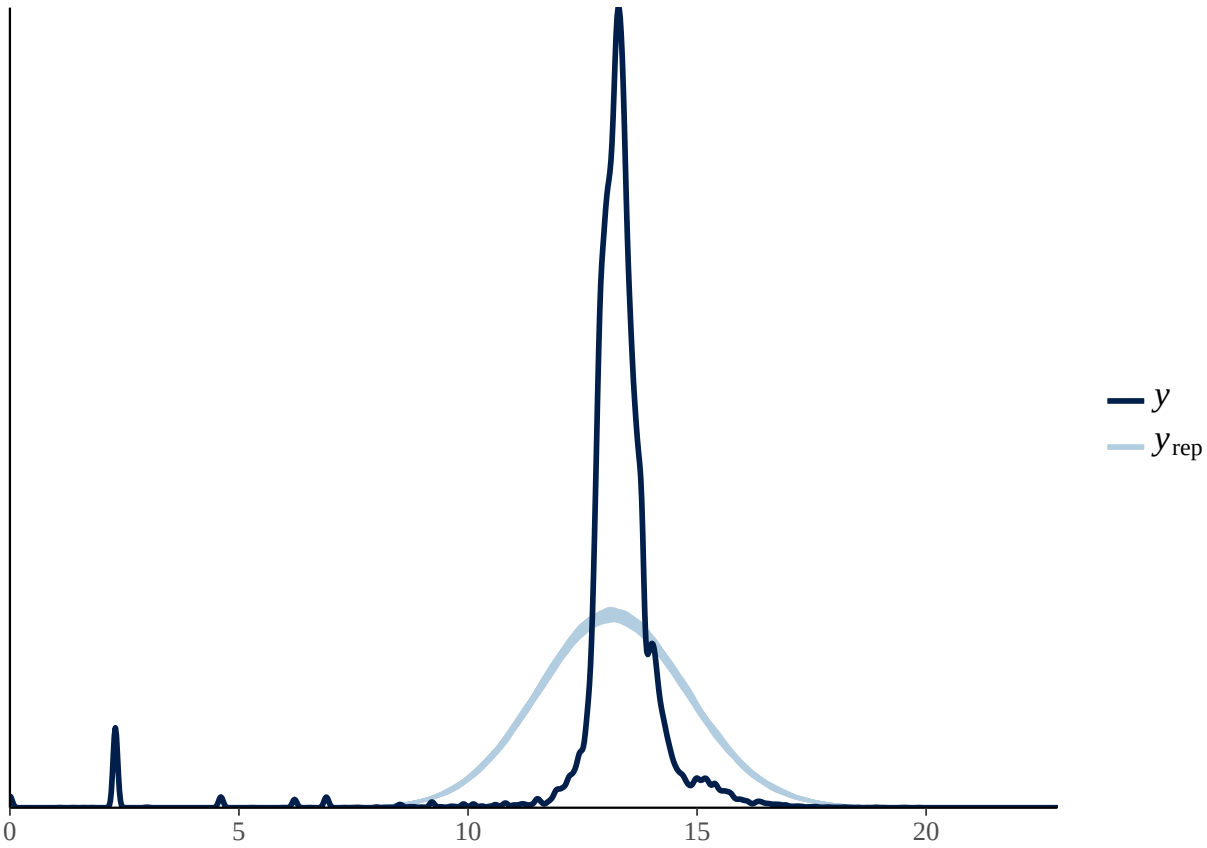
The coefficients for the interacted variables convey the differences in slopes between the predictive differences of sale price across the five boroughs. For example, the coefficient for boroughmanhattan:land_z, 0.27, when exponetiated, 1.31, suggests that a 659.5 square foot increase in land size corresponds to 31% more of an increase in price in Manhattan than in the Bronx. Interestingly, the rate of price increase with increasing land size is slower in all other boroughs; a 659.5 square feet greater land size corresponds to a 5.8% slower increase in Brooklyn, 13.14% slower increase in Queens, and 13.76% slower increase in Staten Island.



The graphs show the point estimates and standard errors of the interaction variables. The Manhattan:land interaction variable has a large standard error, which suggests a large degree of uncertainty in the price change rate between the two boroughs. This is also likely attributable to the fact that there are only 919 data points for Manhattan. The Brooklyn:land interaction variable, which sits across 0, suggests that the difference in the rate of price change by land square footage could be 0 or even positive, given additional information.

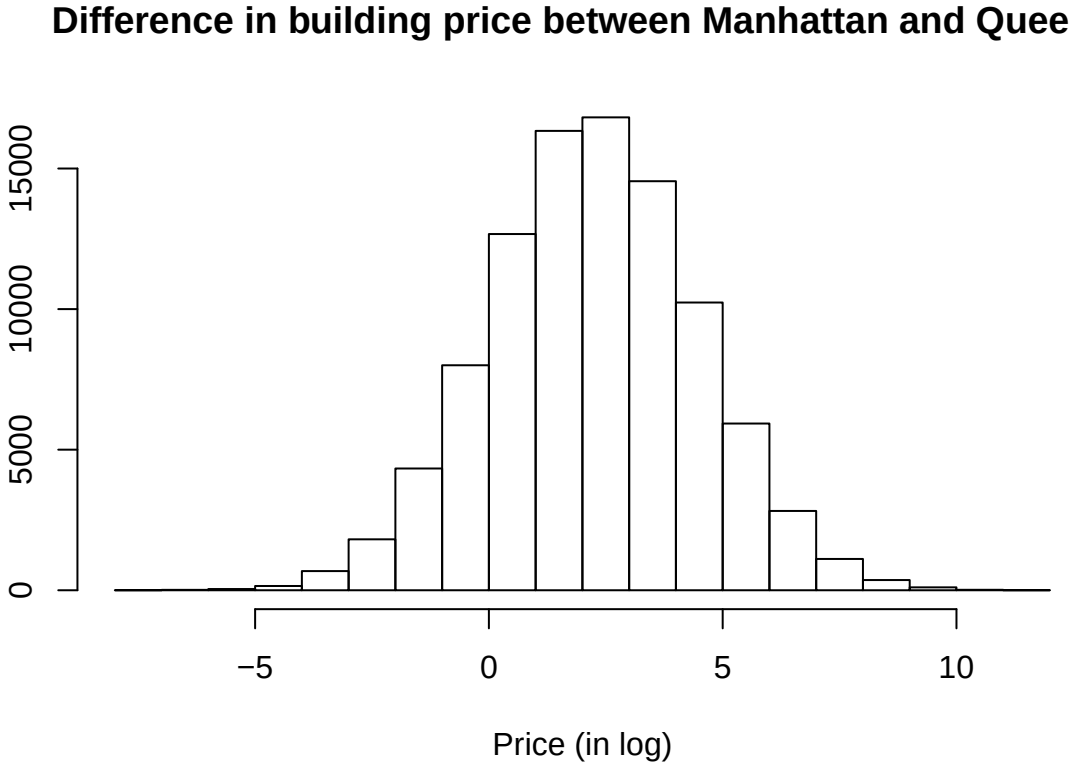
7. Simple Predictive Checks

After fitting the data, a good next step is to compare real data against simulated data to gauge whether or not the model is a good fit. I first made predictions with the model, then got 500 random draws of the distribution of the replicated data and plotted them against real data.



The darker blue line (labeled y) is the density plot of real price data (logged), while the lighter blue lines (labeled y_rep) represent 500 predictive replications of simulated data from the interaction model. Unfortunately, the plot shows that the model is a poor fit. It does a relatively good job of predicting sale prices at the lower and higher ends of the spectrum, but is a poor fit for those in the middle of the range. The model is severely overestimating log prices between 9 and 13, while underestimating log prices between 13 and 14. I suspect there are several reasons for this: 1) For interpretability, I did not include all the variables from the original dataset in the model; the selected predictors alone fail to sufficiently and accurately explain variability in the housing market. Adding some of the removed predictors into the model might induce a better fit. 2) The data contain (a lot of) noise! About a third of the outcome values were missing, and erroneous records were scattered across predictors such as land square feet, gross square feet, and years built. As mentioned above, noise can severely bias the model.

As a final step, one can simulate data to see how a change in a predictor value might affect the outcome. Although the model is overall a poor fit of the data, it predicts with some precision sale prices that are above 17.5 in log prices (approximately 39.8 million dollars). For this example, I simulated data to measure how keeping everything else the same but changing the borough from Queens to Manhattan would affect pricing for the most expensive buildings.



The plot shows a distribution of the differences in pricing with simulated data. The plot illustrates a large variance in the prediction of the price difference between Queens and Manhattan for buildings that are 39 million dollars and up. Furthermore, although it is centered slightly to the right of 0, the distribution still contains 0, which indicates no difference in pricing between Manhattan and Queens for the most expensive properties in the data.

8. Recommended Next Steps

Modeling is an iterative process that one can continuously tweak and improve. If I had more time, I would have loved to...

- Conduct external research to gain some domain knowledge in real estate pricing and understand what other factors influence property prices.
- Use multiple imputation for missing values. My deterministic approach for imputing missing data is flawed because it ignores predictive uncertainty by using point estimates to make those predictions. Furthermore, central tendency is an artifact of a deterministic imputation procedure; we fail to correctly predict for those that might have unusually higher or lower values, which can happen with real data. One approach to remedy this issue is to use multiple imputation, by creating several imputed values, each of which comes from a slightly different model. One can then run a model with the different completed datasets and combine the inferences to account for the uncertainty of the imputed values.
- Incorporate informative priors. By default, the stan_glm() function uses weakly informative normal priors, but with more domain knowledge, one could use informative prior information! For example, we can set the log price to be centered around 14 with longer tails for a more accurate estimate. Using informative priors is especially useful when the data are noisy, and so priors play a bigger role in keeping the estimates more reasonable.
- Try out different models. A multilevel model seems particularly appropriate for answering this question. Multilevel models are similar to the interaction model I've used for the analysis, but allow the varying slopes (and intercepts when appropriate) to share information with partial pooling.