Analysis of Boston Rent Prices

William Dean

Introduction

Boston is known for its high housing prices. However, these costs vary greatly upon the property location within the city. Throughout Boston, different places provide different sets of amenities as well as the proximity to various places of interest that make certain areas more desirable to live than others. Also, desirability of certain housing styles throughout the city can impact the price of rentals. Overall, when looking at rental prices in Boston, there is a delicate balance between the access the location provides and its affordability. The aim of this project is to investigate the the direct impact amenities have on the price to rent a room in Boston.

Data

Data Collection

The website PadMapper provides information about rent through the city. Along with this rent information, the number of rooms being rented was provided. This gives users the ability to look at the price of rent per rooms. Google's Places API provides information regarding shopping, dinner, grocery, entertainment, transportation options, nightlife, and what other amenities that are available in an area of a city. Information regarding parks and outdoor resources can be found on Boston's city website.

Data Description

After aggregating the data about the listings from the sources and extracting characteristics, each room listings had the following attributes:

Variables to capture the proximity of local amenities. Namely, the closest distance in miles to all of the following: T stop, bar, grocery store, coffee shop, restaurant, historical site, landmark, bike path, dog park, and park. These variables describe the local transportation, food, leisure, and entertainment and give insight to how accessible they are for that area.

Some variables reflecting where the listing is in the city. These include an indicator if the listing is within a half mile of a sports arena, another indicator if downtown, and one if the listing is within a half mile of the Boston Common. Similarly, the neighborhood of the city is attached to each listing and the distance in miles to the Boston Common.

There are two variables to capture the characteristics of the listings and other listings within the area. They are the average number of rooms of the closest 10 other apartments which tell what living situations in the area is like as well as an indicator if the listing is a studio apartment.

With each listings, the average traffic count of the closest 3 intersections was recorded in order to describe how busy the location is.

The variable of interest to estimate is the rent, however the log(Rent) for the listing provides a better comparison with the right skewness of rent prices. Figure 1 shows the location of the data points throughout the city of Boston and the log(Rent) price of that data point.

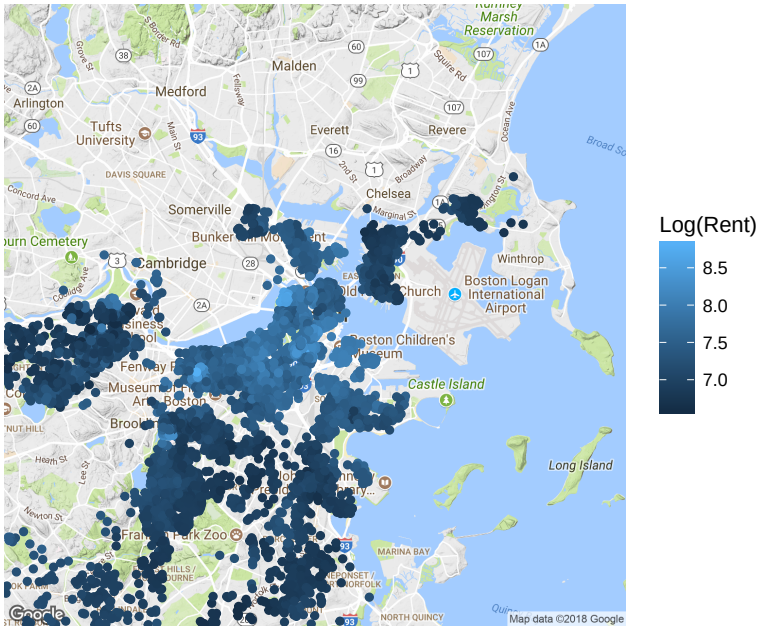

Figure 1: Price of rent per room throughout Boston (Log Scale)

A Need for GLM

Even if you are not familiar with Boston, it is probably apparent that the rent prices are going vary, not only throughout the city, but throughout the neighborhoods in the city. It's also likely to believe how much they vary will depend upon where they are and what amenities the area provides.

For instance, Figure 2 shows that not only does the average rent change between neighborhoods but how much that rent varies also depends upon the neighborhood. Areas like Allston have many types of living situations ranging from family homes to new luxury apartments which increase the variability of a room's price throughout the neighborhood. On the other hand, some areas provide fairly uniform living arrangements. For example, a downtown neighborhood like Chinatown has very similar apartment/loft style living or a very residential area like Roslindale may only have family homes which would affect the variability of the rent prices. Although a crude example, the dispersion of the data is likely to also depend upon the covariates of interest.

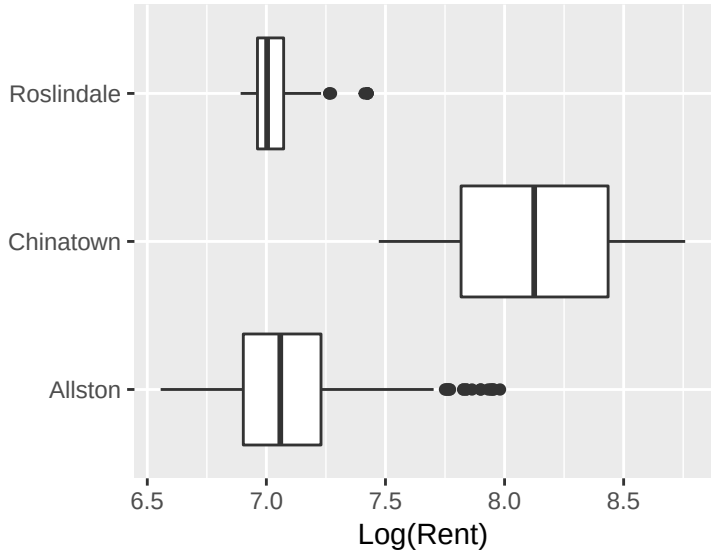

Figure 2: An early indication that the dispersion of the data may depend upon the covariates

That being said, instead of just modeling the rent on the covariates, modeling both the mean and dispersion will likely capture the variation of the data. Using a GLM to jointly model the mean and dispersion will allow for the model such a scenario. That is, after modeling the mean of the log(Rent) with a linear model and the dispersion of data with a gamma model, a new model will take in account both the covariate's effect on the mean and dispersion and will better reflect the data.

EDA

Investigating the relationship the covariates have with the outcome, many appear to have the same relationship with the rent regardless of the location of the listing. For instance, Figure 6 in the appendix shows how as the average number of rooms of closest listings increases on the log scale, the rent per room also decreases on the log scale. Not only does it decreases, but it generally decreases for all neighborhoods. This appears to be the case for these two variables as well as other relationships. Neighborhood appears to have no effect on the trend of the relationship. For many other relationships, the location does not change the effect of the covariate.

In comparison, the effect that some of the covariates appear to vary depending on the neighborhood. Consider Figure 3 where the relationship the log of distance to closest grocery store has with the rent changes. The rent on the log scale in Charlestown appears to decrease linearly as the closest to a grocery store increases, however in Jamaica Plain the relationship between these two variables appears a little different and moves in the opposite direction. The appendix shows a similar neighborhood effect with the variable for closeness to bicycle paths. There appears to exist interactions between some of the variables and the neighborhood of the listings.

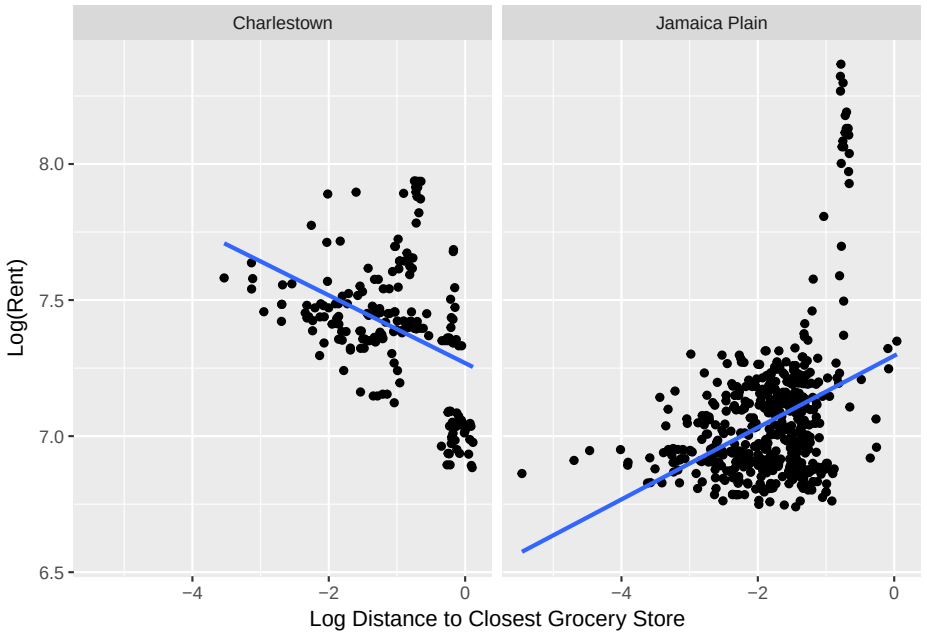

Figure 3: Neighborhood Effect Among Covariates

Many of the variables appear to be more linearly related with the response variable when taking the log of the variable. That is, the marginally the data is very right skewed, having many amenities close, however a few listings that have a sparsity of amenities. From here, all distance variables are used on the log scale to account for this characteristic.

With all of this into consideration, it seems apparent to include an interaction term between the log distance to grocery store and neighborhood as well as the log distance to a bike path and the neighborhood. Many of the other variables suggest less of an impact of interaction with neighborhood so they'll be modeled without an interaction term with the neighborhood.

Model

Mean

With the response of log(Rent) being a continuous variable, we are able to model the mean of the data with a linear model.

This linear model has variables with no interaction term with the neighborhood of the listings. Namely, the log of the distance to the Boston Common, the log distance to the closest: T stop, luxurious apartment building, dog park, park, historical site, landmark, coffee shoe, restaurant, and bar all do not have an interaction term with neighborhood. Also the log of the number of rooms does not have an interaction.

In order to capture the effect of the neighborhood, an interaction term is included between the neighborhood with log distance to grocery store as well as the log distance to closest bike path.

Two other indicators are present in this model. They are indicators for if the listing is a studio and if it is within a mile of a sports arena.

After fitting the this model, we are able to then model the residuals and capture the missing variability from this model.

Dispersion

After fitting just the linear model for the mean, it is apparent that the model does not capture all the variability of the data. Consider Figure 8 in the appendix where the squared residuals of the first model have a some sort of relationship with the average number of rooms of the surround listings.

Modeling the squared residuals from the mean model with a gamma GLM give insight to the dispersion's dependency on the covariates. A log link is used to relate the mean of the dispersion to the parameters.

The predicted dispersion from this model can be used weights to refit the linear regression.

Final Model

Refitting the first linear model with new weights, this model now takes into account the varying dispersion because of our covariates. The linear models use the same covariates but only differ by the weighting of the variances.

Model Comparison

The first model fit does not model the dispersion even though it is likely to have varying dispersion based off our covariates. Without modeling this dispersion, our AIC is -4057. In comparison, weighting our linear regression with a modeled dispersion, the model has an AIC of -5919.

Using a lower AIC to choose between models, it suggest that modeling both the mean and dispersion is provides a better fit for the data.

Similarly, in Figure 4, the residuals of the first model in comparison to the second model where the dispersion is also modeled, we see that there no longer the a drastic increase in variance of the residuals as the fitted values increases which is promising. That is, the first model does not explain the variation of the data as well as the second model as the fitted values are larger.

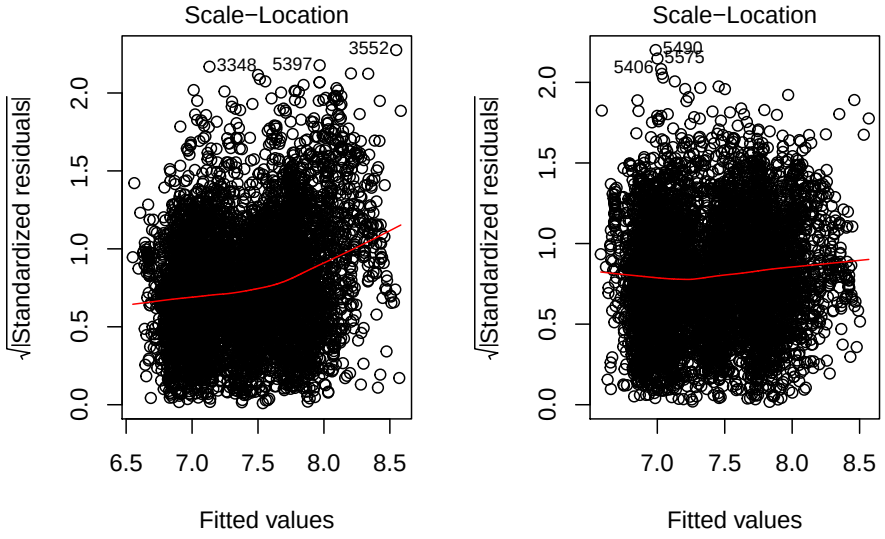

Figure 4: Residuals of First Model (Left) vs Second Model (Right)

The second model appears to capture the variation of the data better than the first model. The regression with the modeled dispersion has an R squared of 0.894 which is an improvement from the first regression's R squared of 0.85.

These few signs suggest that there has been an improvement in the fit of the data after modeling the dispersion, which better reflects our data.

Discussion

Effects of Covariates

Many of the covariates effects appear to follow intuition. For instance, the final model suggests a significantly negative effect of increasing log distance from closest T stop with a 95% confidence interval for the parameter is between -0.022 and -0.011. This suggests that as you get further away from the closest T stop, the price of a room will generally decrease. This makes sense with the T transport being used many people through the city and is clearly favorable to be near city transportation. Similarly, the model suggests that as the average number of rooms in the area increases on the log scale, the expected rent will decrease. This suggests that areas where there are living situations offer rooms will have generally have lower rent per bedroom prices.

The model suggests also some insights that may not be so apparent as well. For instance, as a listing gets farther away from its closest restaurant, there is a statistically significant decrease in its rent compared to other listings. Being closer to a restaurant will, in general, increase the value of the room being rented. On the contrary, the closer a listing is to a bar, the lower on average the rent will be per room. The data suggests that living closely to a bar will generally have a lower rent than a comparable listing further away from a bar.

Consider the effect proximity a coffee shops has on the rent, our model suggest that there is no significant effect. The 95% confidence interval for the log distance to the closest coffee shop has a lower bound of -0.006 and upper bound 0.006. This model suggest that effects of a close coffee shop is insignificant after controlling for all other variables.

It is to note that the effect of the log distance to the closest coffee shop in the first model has a 95% confidence interval lower bound of 0.009 and upper bound 0.025. This is contrary to what the second model says; however, our second model reflects the data better so that is used for inference.

One effect that seems contrary to intuition is that the model suggests having a studio room will decrease the rent on average. The model's 95% confidence interval for this parameter is statistically negative, which seems rather odd that this effect would be the case. However, looking at the data in Figure 5, we can see that non-studio rooms with comparable average number of rooms as a studio do appear to have lower rents, making some sense of this negative parameter estimate.

Figure 5: Justification of Studio Effect

A table of 95% confidence intervals for all the non-interaction effects of both models are provided in the appendix.

Challenges

While modeling this problem, there is a balance between capturing the relationships from the data while keeping a model that does not over simplify the situation. Most notably, being able to capture the effects of each listings attribute without having the added complexity that comes along with the interaction with the neighborhood of the listing. As we saw in the EDA, there potentially exists an interaction effect covariates. In order to keep a model that is rather interpretable and parsimonious many of these may be overlooked to provide that simplicity.

The data had no attributes which described the interior of home which could lead to further explanation of the variability of a room's rental price. For example, year of home or condition may greatly reflect in the rental price and could improve the fit of our model.

Conclusion

Figuring out what factors influence the price of a room rental throughout Boston could be explained with a linear model. However, it also seemed apparent that variability in the price also was influenced because of these factors as well. Because of that, it seemed appropriate to use a gamma GLM to model the dispersion of the data as well. After doing so, we saw improvements in our linear model ability to explain variability of the log price of a room rental through a few measures of goodness of fit.

After improving our model fit, we turned to inference from the model where we saw many conclusion that meet intuition as well as some insights about the effects of surrounding amenities and home features.

Jointly modeling the mean and dispersion of our data allows for a better fit model, leading to better inference about our parameters, and provides a understanding behind the amenities that drive room prices in Boston.

Appendix

EDA

Many covariates are linear with the rent on a log scale. Also, regardless of the neighborhood, the relationship is fairly similar. In the Figure 6 example, increasing the log of the average number of rooms in the area generally decreases the price of the room in the same manner regardless of location in the city.

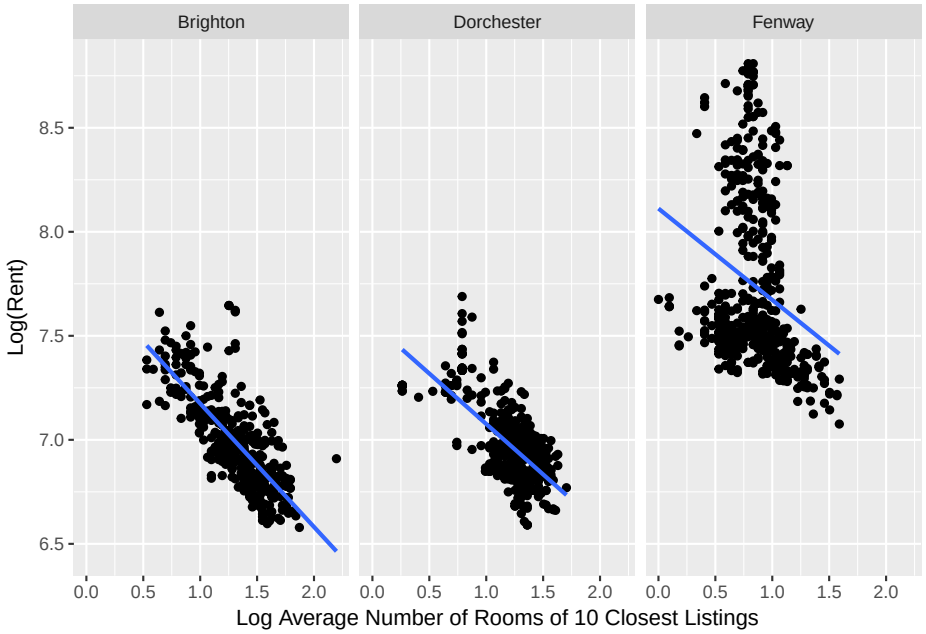

Figure 6: Example of No Effect of Neighborhood

Another case in Figure 7 where the relationship that a response has with a covariate depends upon the neighborhood. That is, in Brighton, there does not seem to be much of a relationship between the closest bike path and the rent. However, Chinatown shows a negative relationship, and Fenway appears to be some other relationship that differs from the both of them.

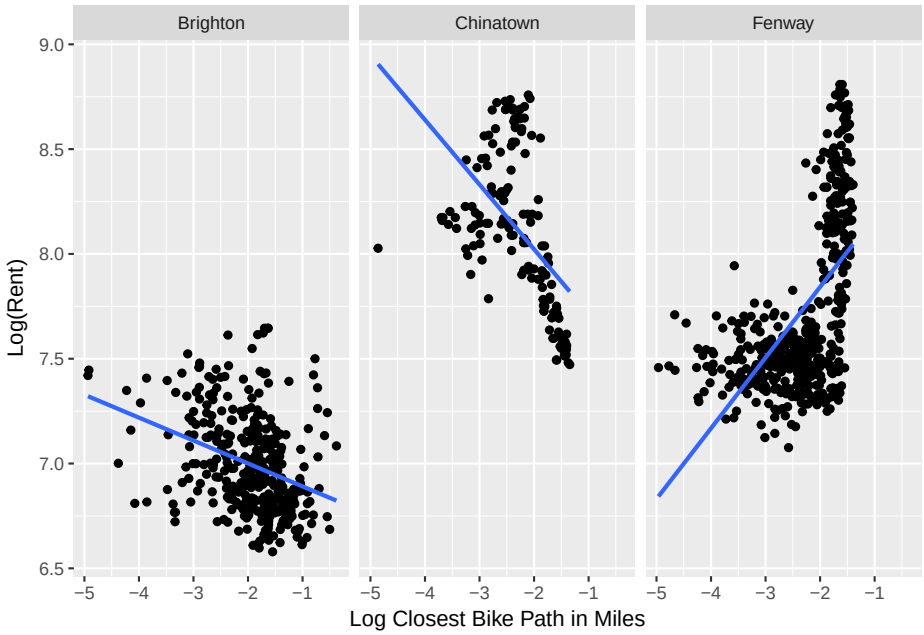

Figure 7: Example of Neighborhood Effect with the Covariate

Dispersion

After fitting the model on just the mean of the data, it is apparent that the dispersion depend upon the covariates as well. Figure 8 suggests that the dispersion is also dependent upon the number of rooms.

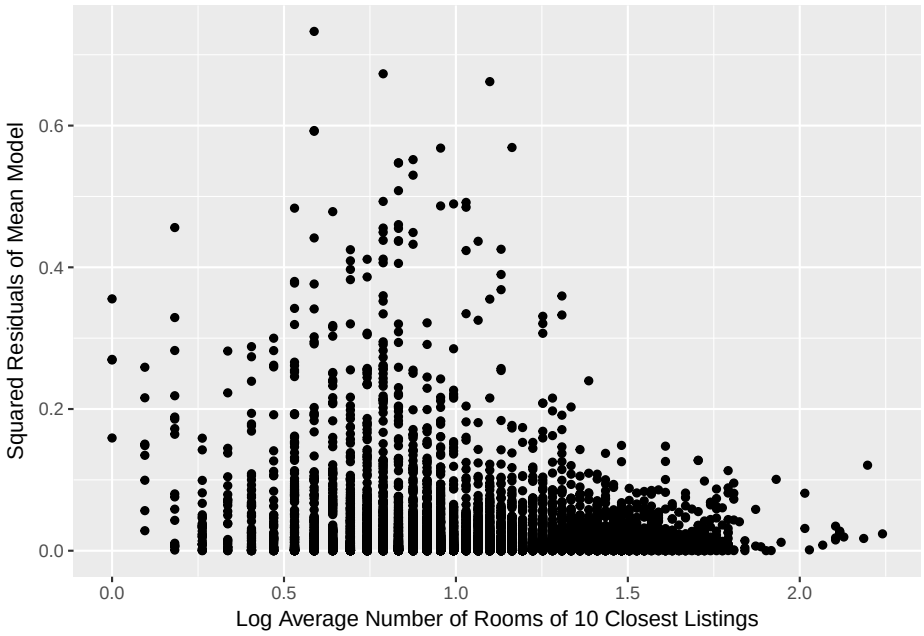

Figure 8: Modeling the Dispersion after First Model

Effect of Covariates

Below are the 95% confidence intervals for the final model's parameter estimates that are do not include an interaction.

The intercept, distance to closest park, closest historical location, closest Bar, and is within half a mile of an arena are all statistically positive. Similarly, the distance to the Boston Common, closest T stop, luxurious apartment complex, restaurant, log number of rooms, and the studio indicator are all statistically negative.

```
##                      2.5 %  97.5 %
## (Intercept)         7.8957  8.0639
## log(closest_bc)    -0.2140 -0.1774
## log(closest_T)     -0.0224 -0.0112
## log(closest_lux)   -0.0304 -0.0193
## log(closest_dog)   -0.0085  0.0040
## log(closest_park)   0.0042  0.0154
## log(closest_hist)   0.0120  0.0228
## log(closest_land) -0.0048  0.0061
## log(closest_Cof)   -0.0062  0.0056
## log(closest_food) -0.0133 -0.0004
## log(closest_Bar)    0.0012  0.0138
## log(rooms2 + 1)    -0.5111 -0.4762
## studioyes          -0.1320 -0.0709
## num_pSport1         0.0848  0.1761
```

Similarly, the 95% confidence intervals from the first model that did not model dispersion for comparison. Because the fit of this model is not as good as the final model, we do not use it for inference.

```
##                      2.5 %  97.5 %
## (Intercept)         7.9221  8.0886
## log(closest_bc)    -0.1886 -0.1436
## log(closest_T)     -0.0180 -0.0036
## log(closest_lux)   -0.0520 -0.0389
## log(closest_dog)   -0.0103  0.0063
## log(closest_park)   0.0077  0.0221
## log(closest_hist)   0.0173  0.0316
## log(closest_land) -0.0149 -0.0001
## log(closest_Cof)    0.0094  0.0247
## log(closest_food) -0.0134  0.0030
## log(closest_Bar)   -0.0095  0.0072
## log(rooms2 + 1)    -0.5343 -0.4928
## studioyes          -0.1698 -0.1156
## num_pSport1         0.1208  0.1862
```