Northeastern University

PPUA 5262 – Final Project

Analysis of Craigslist Housing Postings in Boston, Massachusetts

Evaluation of long-term rental listing price trends against demographics and places of interest (2021-2022)

Jessica Tanumihardja
4-24-2023
Seattle, WA

Executive Summary

Analysis of Boston's housing through urban informatics can reveal community patterns and the housing market. The "Craigslist Postings in Massachusetts" dataset (Castro, Amiri, et al., Craigslist Postings in Massachusetts 2020) provided by the Boston Area Research Initiative is explored and dissected to gain a general understanding of Boston's neighborhoods that reveal whether the community has certain housing preferences and characteristics. The focus of the analysis is the impact of those characteristics on the long-term rental listing price per month. This analysis can be beneficial for the government and rule-makers to determine which area should be the focus of attention/ need improvement or for private developers in determining areas to be developed or types of development in a certain area. This study can also inform renters to decide on their preferred rental location and affordable housing area.

R studio is utilized to process, parse, analyze, model, and visualize the dataset. Internal analysis of the Craigslist postings dataset reveals that the median long-term rental price for the City of Boston ($2,500) is higher than the median of the whole state of Massachusetts ($1,895). There is also a greater supply of rental property (more posting observed) during the summer months. From the text mining of the post's description, most of the postings in the City of Boston are for apartment-related property (61%). Pet owners are also accommodated well as ~40% of the property allows dogs and ~55% of the property allows cats.

To reveal the underlying features impacting the listing price, Boston's places of interest data from Google Places and demographic data from recent census records are extracted to provide additional measures. These data combinations broaden the analysis from a housing focus to a community focus study. The median listing price per square foot of property is utilized for comparison to standardize the price data.

Mapping these additional measures shows that higher listing prices are found in the same place as the cluster of the city development, which is located around the downtown area. This area is also where the food and entertainment services are concentrated in. However, more postings are found in Allston/Brighton area than in the central/downtown area. This might reflect the impact of the university area on the rental housing market, especially for those posting on the website.

Pairwise correlation between posting counts and median listing price per square foot and the places of interest and demographics measures per census tracts ID. The number of postings does not have a strong correlation with any of the variables measured. The listing price has a relatively strong correlation (coefficient magnitude > 0.5) with, from highest to lowest, the percentage of commute methods (by walking, by automotive, and commute length between 10-30 minutes), age range (under 18 and 18-34 years old), race (Black and White races), food services count, and entertainment places counts. This quantifies and proves the hypothesis made previously on what affects the listing price. However, the impact of places of interest is less than what was expected. The importance of commute length and methods in determining people's willingness to spend on a rental property is also reflected in the ANOVA analysis. The downtown and industrial/institutional neighborhood type has a higher median price per square foot than the residential area.

The result of the correlation analysis is strengthened by the findings of regression analysis of listing price against the independent variables. The final linear model is built with the top 5 variables with the highest standardized beta: commute length between 10-30 minutes, commuting by walking, age under 18 years old, Asian, and area of census tracts (acres). This model can explain ~65% of the variance in listing price. Although the model can predict the trend well, the visualization shows that there is quite a bias observed. A more complex regression model such as the bagging and boosting model might increase the performance metrics and reduce the bias.

Introduction

Housing trends are an important aspect of a community's economic development. Although the majority of housing units are owner-occupied, approximately 35% of occupied housing is a rental property (Jeffrey Stupak 2023). Therefore, community housing insights can be discovered through an analysis of rental house trends and their distribution.

To narrow the exploration into practical tasks, this document investigates the rental housing trends from the "Craigslist Postings in Massachusetts" dataset (Castro, Amiri, et al., Craigslist Postings in Massachusetts 2020) provided by the Boston Area Research Initiative (BARI). The dataset contains Craigslist property rental listings in various regions of Massachusetts but is narrowed down to the City of Boston only for a more targeted analysis. This is an intriguing dataset as one can find the characteristics of Boston, the pulse of the city, or any of its corresponding neighborhood's recent rental property trends and supply-demand markets. It should be noted that this dataset mainly covers the COVID-19 pandemic time so further data scrapping is needed to get the "normal" (before 2020) and "new normal" (after 2022) trend.

Through the dataset variables, one can draw property desirability insights as the conceptual big picture of Boston's housing characteristics. The analysis is concentrated on what contributes to the desirability of a property based on the listing price. Does Bostonian prefer convenience over space? (Ryan, Ward and Pederson 2022). This question can be answered by analyzing the price per area of property and mapping listing prices over different neighborhoods. This preference is especially true for properties located in downtown areas where the population density is higher and available space is limited. Proximity to places of interest (restaurants, universities, entertainment places, etc.), which is a measure of convenience, and the type of neighborhood a property is located are also analyzed to answer this question.

To further understand Boston's housing market characteristics, the dataset is analyzed against the demographic characteristics related to housing. Does median household income correlate with the median housing price in Boston's neighborhood? Do generational differences reflect in the housing trend of different neighborhoods? Does gender or race determine the distribution of housing properties in the neighborhood? Statistical analysis of these socioeconomic factors in the dataset can filter the significant factors to the price and desirability of the properties. The common hypothesis is that housing prices have a positive relationship with median household income. Is this hypothesis true? These questions are answered through this data exploration.

Data and Methods

This document is a combination of all the city exploration discussions (Tanumihardja 2023), with detailed R Studio code provided in the Appendix. The main source of data is provided by the "Craigslist Postings in Massachusetts" dataset (Castro, Amiri, et al., Craigslist Postings in Massachusetts 2020). R Studio is used to pre-process, clean, analyze, model, and visualize the dataset. Four other administrative datasets are utilized to expand the analysis:

- Tract Census Data provided by BARI (Castro, Amiri, et al., Harvard Dataverse 2020)
- Google Place Places of Interest (Gibbons, et al. 2020)
- Census Indicator estimate, ACS_1418_TRACT.tab (O'Brien, Ciomek and Tucker 2019)
- Boston neighborhood shapefile (Boston Neighborhoods 2020)

Exploratory Data Analysis

Exploratory data analysis (EDA) was the first step to gaining insight from the raw Craigslist dataset. It is looking at the data from different angles and getting preliminary visualization or analysis to better understand the data. The dataset contains 205,450 unique Craigslist postings (identified by unique ID) gathered between March 2020 to December 2021. Each observation has 13 other variables such as time of posting, price, and location.

The analysis will focus on the City of Boston, which has 180 census tracts and 16 neighborhoods according to the 2010 Census records (Boston Planning & Development Agency 2023). After the data cleaning, outlier removal, and filtering of the census tracts to include only the City of Boston, the sample size is reduced to 17,649 postings. Sampling some of the postings, most of the listing seems to be professionally posted by property management. It appears that Craigslist is still a popular site to advertise places to rent in Boston even though there is a newer web posting option.

Among the 13 original variables, only the listing price is a true numerical variable. Hence, the focus of the analysis will be based on this listing price variable. A new variable, price per square foot, is added to generalize the price data and compare similar properties (based on location, size, style, etc.). It is not the best valuation metric but one of the simplest before further diving into other complicated variables. It also can give an overview of the demand for a particular location since a higher-demand location is expected to have a higher price per square foot value. Two more new variables are added and related as both indicate the type of property: apartment (multi-family) or house (single-family). This category is often an important distinction for renters as they might prefer one over another. The presence and amount of apartment or multi-family property listings might be a measure of population density. This modified dataset is also merged with other econometrics and demographic datasets as well as the Google Place dataset to increase the understanding of housing trends in the City of Boston.

Most of the useful text information is in the `BODY` variable so text analysis and mining were conducted on this 1 feature to be able to extract useful information from it. Some of the descriptions listed in the BODY variable are similar among the postings such as the amount of deposit, laundry location, number of beds, and baths. Three new variables (binary class) are generated to identify whether the postings are apartment, house, or student related by doing a keyword search in the `BODY` variable. The apartment type distribution is shown in Figure 1.

One important note is these listings are for rental properties. Though it may explain the housing market in the Boston area, it is only explaining the rental property market, not for house owners. It is good to know the limit of the dataset. In addition, some neighborhood or census tracts ID has more postings than others, which leads to inherent bias when analyzing the characteristics of each census tract.

New Measures: Aggregation

The main objective of the Craigslist housing analysis is to understand the variation in listing prices based on location, demographics, and amenities. The listing price ($/month) is the only numerical variable in the dataset. Thus, the only numerical measure from the original dataset is the median listing price differences among the census tracts by aggregating posts by tract IDs and calculating the median/mean/minimum/maximum listing price of each tract. The median price is chosen for further analysis since it is insensitive to outliers. This aggregated dataset is merged with other datasets listed below.

New Measures: Places of Interest

Since the variables in the original dataset are limited, the dataset is merged with the places of interest count in each census tract provided by Google Places. The Google Places dataset contains places of interest recorded in Google records. There are 98 unique tags to identify those places and not all of them are beneficial for the housing analysis. Therefore, only 6 types of categories are gathered to be compared with the Craigslist dataset. The categories are:

- Food-related services (restaurant, bakery, bar, cafe, meal takeaway, and food)
- Grocery services (grocery and convenience stores)
- Transit stations (bus, train, subway, and light rail stations) (Help 2023)
- School-related (schools and universities)
- Entertainment-related (zoo, stadium, bowling alley, movie theater, art gallery, nightclub, aquarium, museum, and shopping mall)
- Parks

This dataset is aggregated by census tract by totaling the places of interest count according to the six categories above. The merged dataset allows the evaluation of “property desirability” in each census tract. The focus of the discussion is to check if the median listing price is impacted by the number of places of interest in a certain census tract.

New Measures: Demographics

The aggregated Craigslist dataset is also merged with the census indicator data. The dataset contains the percentage of demographic features in each census tract. Therefore, the impact of econometrics (population density, sex ratio, age, and household income), race, and commute methods on the listing price can be quantified (Boston Area Research Map 2021). Along with the places of interest data, the dataset has enough measures to find the underlying trend of long-term rental property pricing in the City of Boston. Correlation and regression analyses are conducted with these merged datasets.

Visualization

The comparison of all these measures is achieved by plotting, analyzing groups by t-test & ANOVA, and conducting correlation & regression analysis. Visualizations also accompany the analysis result to convey the summary efficiently. The visualization includes bivariate plots (dot, bar, and line plots), choropleth maps, and correlation matrices. The Ggplot2 library is utilized to produce the plots along with necessary extensions (GGally, patchwork, and ggcorrplot). For mapping, the Boston Neighborhood shapefile is merged with the dataset to produce the census tract boundaries via sf library. Insights are drawn based on the comparison results.

Result and Discussion

Statistical Analysis & Comparison

Data distribution analysis is an important preliminary step before building a more complicated analysis. The pricing data distribution for the City of Boston follows a normal distribution with a median listing price of $2,500. This number is higher than the statewide average of $1,895, which indicates a higher listing price is expected in the city. The highest posting counts were found in the month of June, followed by May and July. Higher summer months' availability might indicate the importance of schools and universities in Boston's housing characteristics. It might also indicate that people tend to move during better weather. There is no clear distinction for a day of the month although the posting counts decrease as the end of the month is approaching.

From the allowance of pets variables, more listings allow cats (more than half of the listings, 56%) than dogs (40% of the listings). The median listing price of the property that allows pets is higher than the one that doesn't as shown in Table 1 below.

Table 1 Allowance of Pets distribution in Craigslist Housing Dataset

	Allows Dogs	Allows Cats
Percentage of postings	40%	56%
Median Price (allowed)	$2,654	$2,500
Median Price (not allowed)	$2,400	$2,450

Figure 1 (left) shows the distribution of Craigslist posting counts per neighborhood and whether the posting is identified as an apartment property (61% of postings) through text mining. The highest number of postings are found in Allston/Brighton neighborhood and followed by Jamaica Plain. The lack of postings in the Central neighborhood (North End, West End, South End, and Downtown) was not expected, but it might be explained by lower supply (less availability) and higher demand in that area. The data imbalance for some neighborhoods, as well as tracts ID, may cause bias in the analysis.

On the other hand, the highest median listing price is found in the South Boston neighborhood ($2900/month), followed by South End ($2762) and Central ($2700) as shown in Figure 1 (right). The higher availability in Allston/Brighton neighborhood with relatively affordable pricing makes it a great neighborhood option to live in Boston (Bungalow 2022).

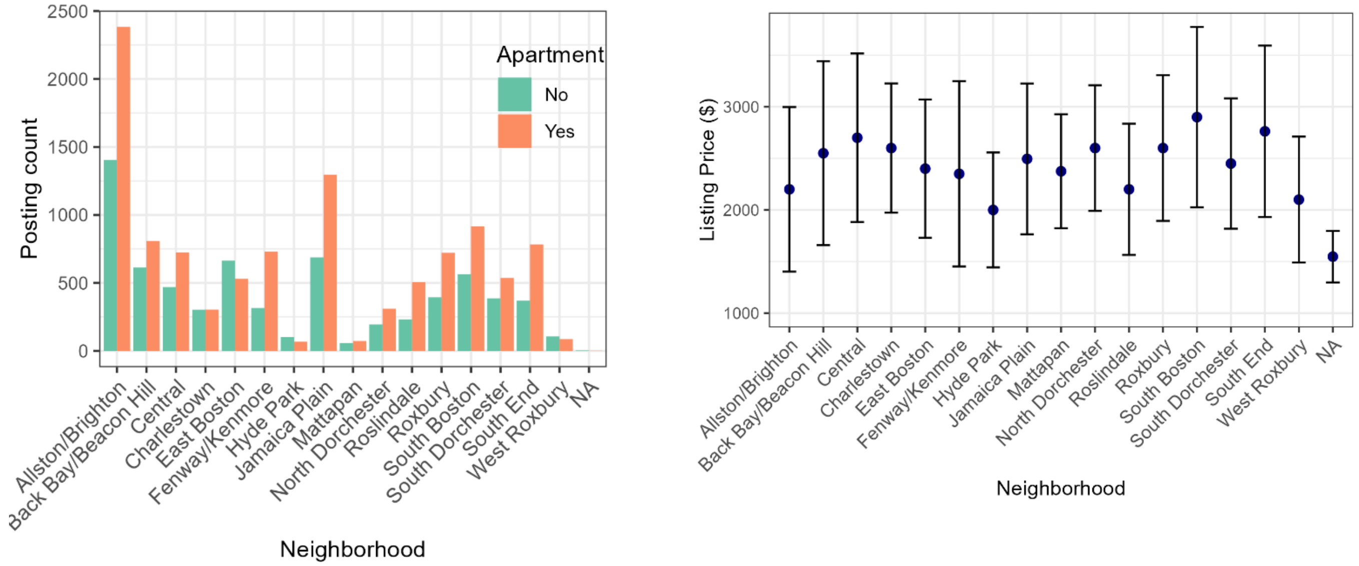
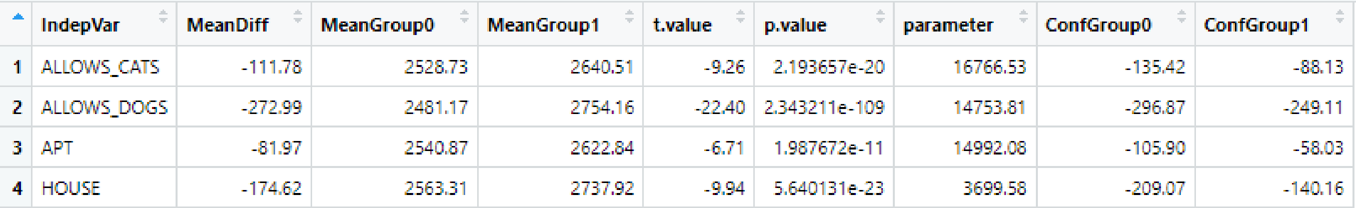


Figure 1 Craigslist Rental Postings (left) and Listing Price (right, with 1 standard deviation error bars) Distribution According to Boston's Neighborhood.

To aid with the comparison, a t-test is conducted for these binary variables (ALLOWS_CATS, ALLOWS_DOGS, APT, and HOUSE) to prove the hypothesis created beforehand, whether listing prices differ between these two types of property, apartment, or house, respectively. The result is summarized in a table shown below.

Table 2 Summary of t-test analysis with Craigslist housing binary variables in the City of Boston



	IndepVar	MeanDiff	MeanGroup0	MeanGroup1	t.value	p.value	parameter	ConfGroup0	ConfGroup1
1	ALLOWS_CATS	-111.78	2528.73	2640.51	-9.26	2.193657e-20	16766.53	-135.42	-88.13
2	ALLOWS_DOGS	-272.99	2481.17	2754.16	-22.40	2.343211e-109	14753.81	-296.87	-249.11
3	APT	-81.97	2540.87	2622.84	-6.71	1.987672e-11	14992.08	-105.90	-58.03
4	HOUSE	-174.62	2563.31	2737.92	-9.94	5.640131e-23	3699.58	-209.07	-140.16

Welch Two Sample t-test is utilized to calculate the analysis above. Group 0 indicates a FALSE value (don't allow pets or is not an APT/HOUSE) and Group 1 indicates a TRUE value (allow pets/ an APT/ a HOUSE). The negative values mean that Group 0 has a lower listing price than Group 1. For the ALLOWS_CATS variable, the mean price difference is approximately $112 with a very small p-value (a very small chance that it is due to chance). The t-statistics => 3, which means it is significant. The mean difference in the ALLOWS_DOGS variable is greater ($273), which means the two groups have more obvious price differences compared to the cat groups. This supported the result found in Assignment 3 as the property with an allowance for dogs has a higher listing price. The p-value is almost zero and t-statistic is >3.

The mean differences for the APT variable are about $82, which is the lowest among all four independent variables analyzed. The mean difference for the HOUSE variable is much higher with approximately a $175 difference between Groups 0 and 1. It is expected as single-family homes usually have higher square footage and are bigger than apartments. For both variables, the p-value is also very small, and the t-statistic value is >3 indicating that the higher price is not by chance and is significant. The mean differences for all four categorical variables are outside of the confidence of interval range for Group 1 (TRUE). Therefore, the difference is significant if the p-value < 0.05.

Mapping New Measures

The comparison among different neighborhood areas can be expressed clearly with the choropleth maps. The boundary division indicates the 2010 census tracts boundary. The color fill indicates the range of each category. To standardize the price, the variable median listing price per square foot is used. Figure 2 shows the median household income, posting count, and median price per square foot for each census tract. The neighborhood names are provided in the first map (left). The grey area means the amount exceeds the range shown (high number).

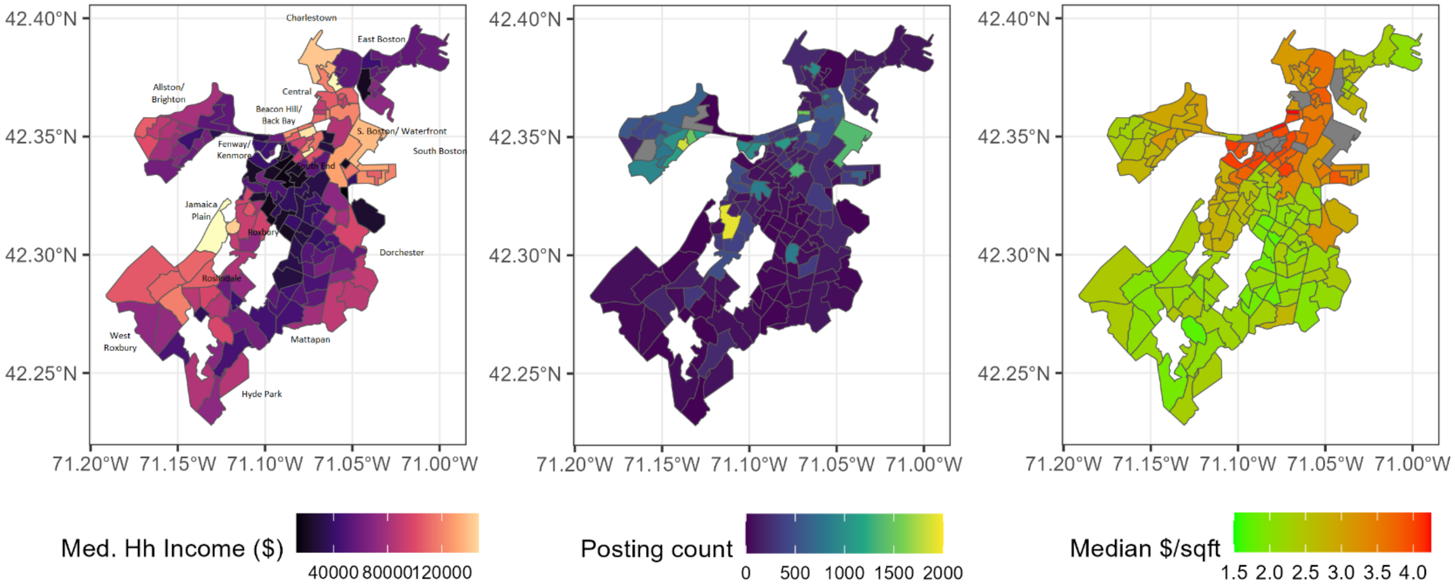


Figure 2 Map of Median Household Income, Craigslist Posting Count, and Median Listing Price/ sq.ft. for the City of Boston (2020-2021 data) per census tract ID.

As shown in the bar graph in Figure 1, a higher posting count is observed in Allston/Brighton and South Boston neighborhoods. Higher prices are also found around Downtown/Central, Back Bay/Beacon Hill, and South Boston. It should be noted that the demographic data provided "median gross rent" across different CT IDs. Although most of the trends are similar, there are some areas where the result is the opposite of what the Craigslist data provided. For example, South Hyde Park and Charlestown have higher median gross rent values than the median listing price. Caution shall be exercised when using the map since Craigslist data only a fraction of all available rental properties.

The median household income somewhat reflects the median price map except for some areas like West Roxbury and Jamaica Plain. The relationship is explained further in the correlation section below.

For places of interest analysis, there is no strong correlation between the 6 categories analyzed and the listing price. A slightly positive relationship with the number of postings (and price to a lesser extent) is found in the food-related services and entertainment places. More postings are found in the neighborhood with more food-related services and the median price of the postings is also slightly higher as shown in Figure 3. The left map shows the median listing price ($) per square foot of property (green is the lowest). The center map shows the number of food-related services located in each CT (dark blue is the lowest). The right map shows the number of entertainment-related places of interest in each CT (blue is the lowest). The grey color for each map means that the number for that respective CT is higher than the limit of the range (high value).

The grey area of the food-related places of interest map on the right, where nearby (same census tract) food-related establishment is greater than 50 stores, corresponds to the red color in the median listing price map on the left. This proves the positive relationship between the two variables. Another strong relationship is found in the entertainment services as more postings and higher median prices are found in neighborhoods with more entertainment places, especially in the South Boston neighborhood. This positive relationship is quantified with a correlation matrix in the next section.

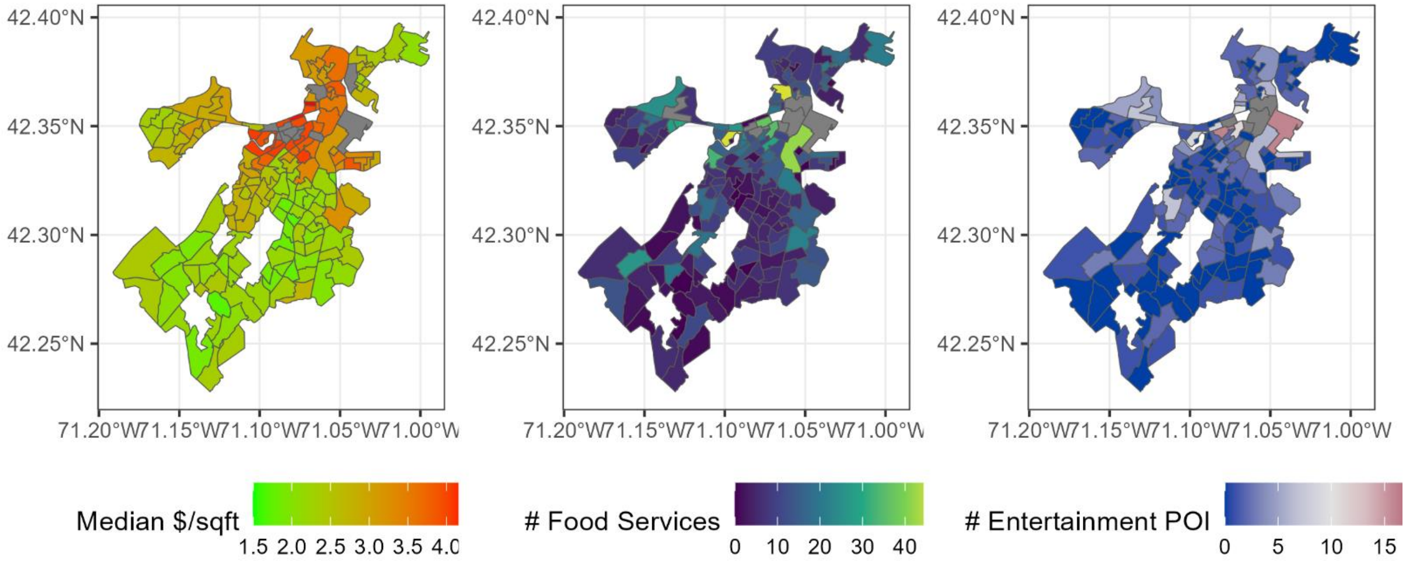


Figure 3 Map of Craigslist Median Listing Price/ sq.ft., Food-related services, and Entertainment services for the City of Boston per census tract ID.

There is no clear correlation found between proximity to grocery/supermarket or transit with the amount of posting or the median price. This is unexpected as they are considered a necessity for most people. Furthermore, data imbalance is found in the park and university-school categories. Therefore, generalizations should be made with caution as they will contain a higher bias. In addition, since the aggregation of the property is conducted per census tract instead of direct distance to places of interest, properties close to the borders of the tracts might have disadvantages. It might be close to the place of interest, but it is excluded because they are in different tracts IDs.

Correlation Analysis

To quantify the relationship found above, correlation matrices were utilized. It is a bit challenging to analyze each property since there is no parcel ID associated with the property (for security reasons). Therefore, analysis was done based on the census tracts ID or CT ID (neighborhood). The focus of the variables will be the number of postings and median listing price (per sq. ft.) as utilized before. All other new variables will be analyzed with these two variables. Pearson correlation between variables can be calculated using multiple methods in R as shown in the Appendix. Preliminary correlation analysis with only the Craigslist data shows no correlation among the variables except for a weak correlation between CT_ID_10 and the listing price.

There are several important correlations from the original dataset. Within the original variables in aggregated Craigslist dataset, the n_postings variable has a strong positive correlation with n_allowsCat (0.973) and n_allowsDog (0.755) with p-value = 0 (highly significant). However, it has a weak negative correlation with median_price with a higher p-value (less significant). This indicates that more postings listed pet allowance, but the allowance of pets does not impact the median listing price extensively.

The min_price has a medium negative correlation with n_allowsCat and n_allowsDog (p-value very small, it is a significant variable). Similarly, the max_price has a medium positive correlation with n_allowsCat and n_allowsDog (p-value very small, it is significant). Therefore, the allowance of pets impacts the minimum/maximum pricing more than places of interest variables.

The correlation analysis with the new measures is divided into four different categories to increase the matrix clarity.

Places of Interest

The pairwise correlation matrix produced by the GGally library for the Places of Interest variables is shown in Figure 4. The diagonal line shows the distribution of each variable. The lower left triangle shows the dot plots between two variables. The upper right triangle shows the correlation value (r-value) of the pairwise correlation. The “n” in front of the variables stands for “number of”.

The median_price (and mean_price to some degree) has a weak positive correlation with n_food, n_school, and n_entertainment (besides other price-related variables). The p-values are <0.01 (n_entertainment is lowest). The additional variable created to standardize the median listing price with property area (median_priceSqft) is also utilized here. It has a weak-med positive correlation with n_allowsCat, n_allowsDog, n_food, and n_entertainment (0.2-0.3).

The highest correlation value magnitude with the original Craigslist variable is a weak correlation between “median price per sq.ft.” & “entertainment” (0.322) or “food” (0.316) with a p-value < 0.001 (significant). This supports the finding in the mapping section above. The median price per sq. ft. has a higher general correlation compared to the median price, so this variable is chosen as the representative of the listing price from this point onwards.

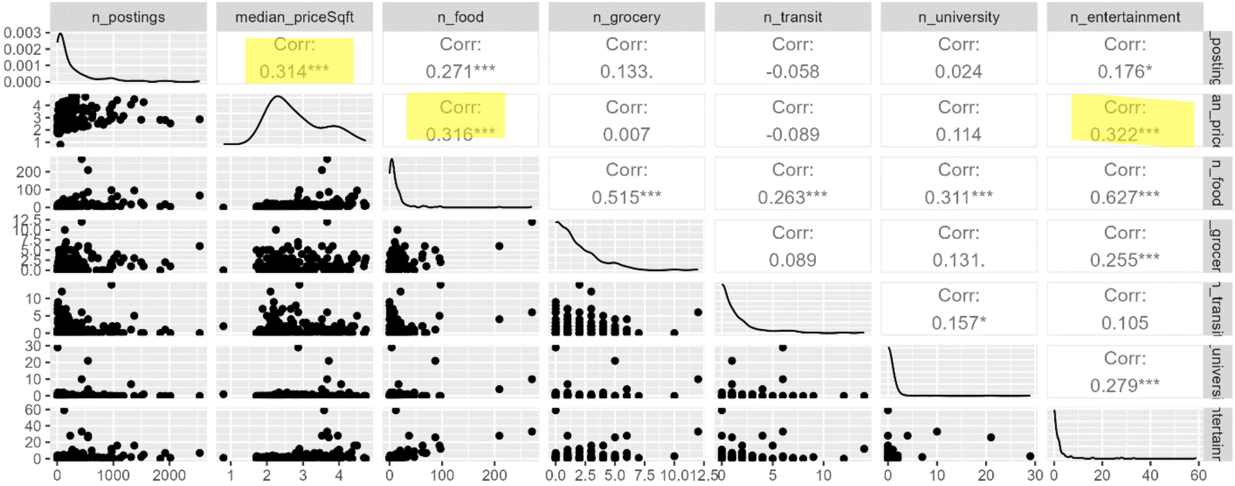


Figure 4 Correlation Matrix between Number of Craigslist Postings, Median Price/sq.ft. and Places of Interest count per census tract ID. The top 3 correlation coefficients are highlighted.

Among the places of interest variables, it is found that there is a strong correlation of n_food with n_convenience and n_entertainment (>0.62). In addition, there is a medium correlation of n_food with n_grocery or n_school (0.40 – 0.52). This indicates that commercial areas tend to cluster together like malls and plazas. Both have p-values < 0.001 (the actual number is 0 or very close to 0).

Econometrics

The correlation matrix with econometrics variables from Census Indicators is shown in Figure 5. For the n_posting variable, it has a medium correlation with age under 18 years old (negative) and age between 18 – 34 years old (positive). This indicates that there are more postings (supply) when the area has a higher adult population than other age ranges. It makes sense since internet postings are targeted at millennials and Gen Z. Similar trend is observed between median price per sq. ft. with ages under 18 & age 34-64 might indicate young adults and your family (with no children) tend to live in higher median price areas. A family with kids might opt to live in the suburb for a quieter and/or larger property.

PopDen (population density) has a medium positive correlation with median_priceSqft (very small p-value, significant). This indicates that urban areas tend to have higher rental properties which are expected. There is no correlation with the sex ratio variable, which is a great thing! In addition, there is little to no correlation between any variables with median household income, which is unexpected.

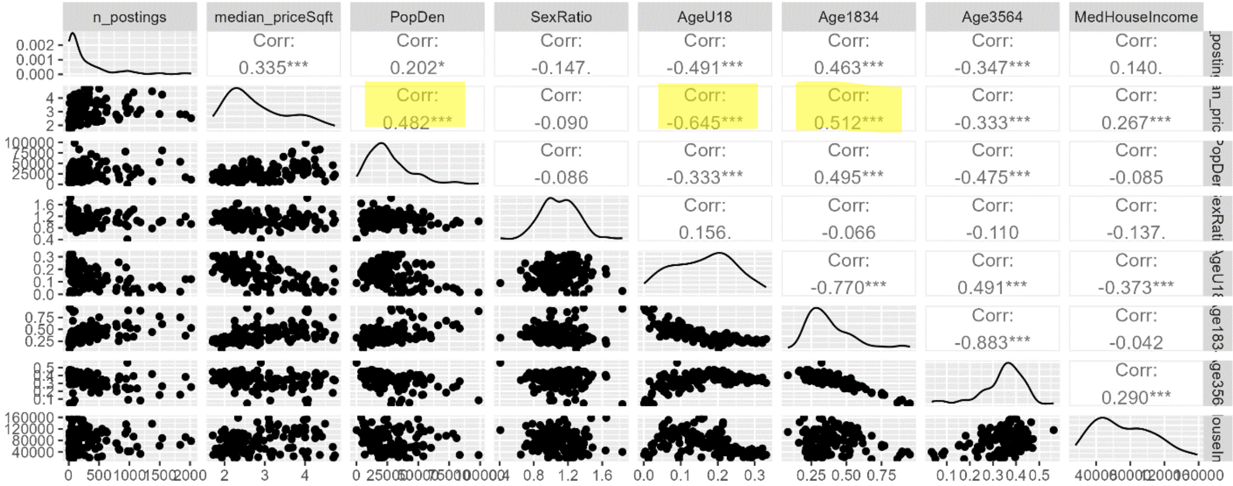


Figure 5 Correlation Matrix between Number of Craigslist Postings, Median Price/sq.ft. and Econometrics variables per census tract ID. The top 3 correlation coefficients are highlighted.

Race

For the race matrix shown in Figure 6, the number of postings variable has a weak correlation with all the races. The most prominent r value is with the white and black communities. It is not a strong correlation though so it will be interesting to know whether this value decreases over time as racial equity increases.

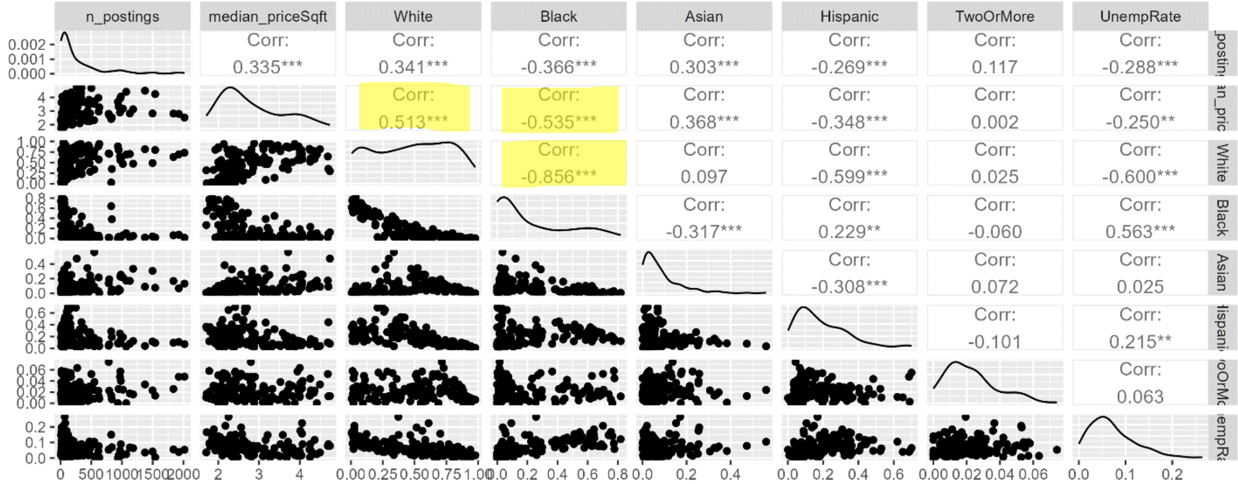


Figure 6 Correlation Matrix between Number of Craigslist Postings, Median Price/sq.ft. and Race variables per census tract ID. The top 3 correlation coefficients are highlighted.

The median price/sq.ft. has a medium negative correlation to the black race with a low p-value (significant). This is not a strong enough value as the plots are more exponential than normal correlation plots. The r value is an almost similar but opposite sign with the white race. This also coincides with the unemployment rate column as it is higher for the black race. Since the unemployment rate correlation with median price/sqft is weak and negative, it seems like the impact of race is more prominent than the unemployment rate for n_posting and median price/sq.ft. variables.

Commute Method

Different commute methods from the census indicators are listed in Figure 7. The highest correlation value is found between median price/sqft with ByWalk (0.752) and Commute1030 (0.697) with a p-value < 0.001. This supports the hypothesis of residential property closer to high amenities (POI, work area, university area) and, likely located in an urban area, has higher rental value.

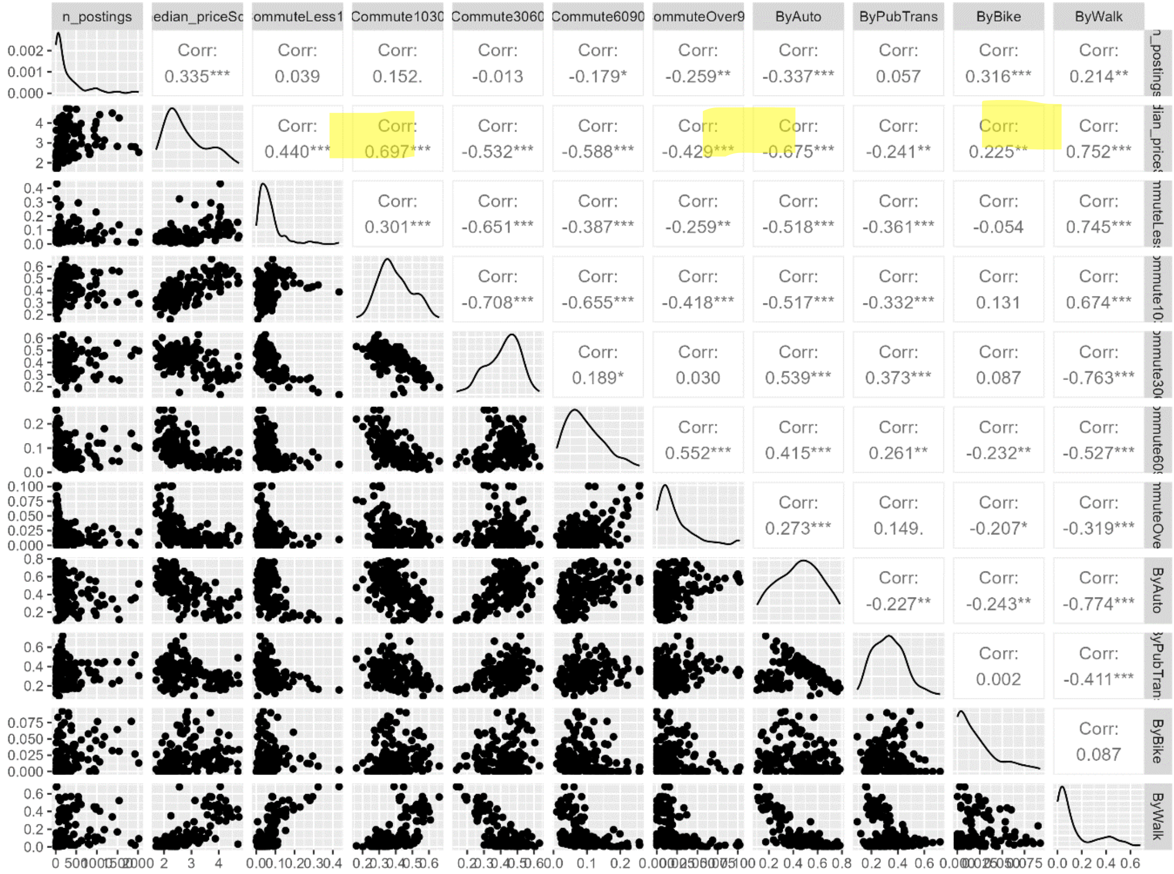


Figure 7 Correlation Matrix between Number of Craigslist Postings, Median Price/sq.ft. and Commute Method variables per census tract ID. The top 3 correlation coefficients are highlighted.

The next highest correlation value is median price/sq.ft. with Commute3060 and Commute6090 (negative). This is also expected since commuting longer than 30 minutes tends to be achieved by car or public transit. The further the distance from the destination, the lower the price is.

It should be noted that there is not much correlation with n_postings except a weak negative correlation with commuting ByAuto (car) and a weak positive correlation with ByWalk.

ANOVA Analysis

To assess the impact of neighborhood type on rental pricing, ANOVA (Analysis of Variance) is utilized for comparing groups similar to the t-test calculation above. Reports came out in the form of F-statistic (F-value), which compares between-group variation and within-group variation. If all variation is accounted for, R-squared = 1. If all groups have identical distribution, R-squared = 0 (no variation). In between values are evaluated with sums of squared values (SS or Sum Sq). In the summary(aov_heat), “grade” indicates the categories’ degree of freedom (DOF), and “residuals” indicates the rest of the variation DOF. Post-hoc reporting allows further assessment of the result of ANOVA (“a series of t-tests comparing the means of pairs of categories”).

From the ANOVA test, the sum square within groups (type = 27.97) is smaller than between groups (residuals = 78.84). With a high F-value and low probability, it reveals that there is a significant difference among the groups. The mean for each type of neighborhood is shown in Figure 8.

Further analysis using TukeyHSD() discovered that the highest difference is found between Park and Downtown neighborhood types (-1.4 mean price difference). This is reasonable since downtown property tends to have a lower price and more parks are likely to be found in suburban areas.

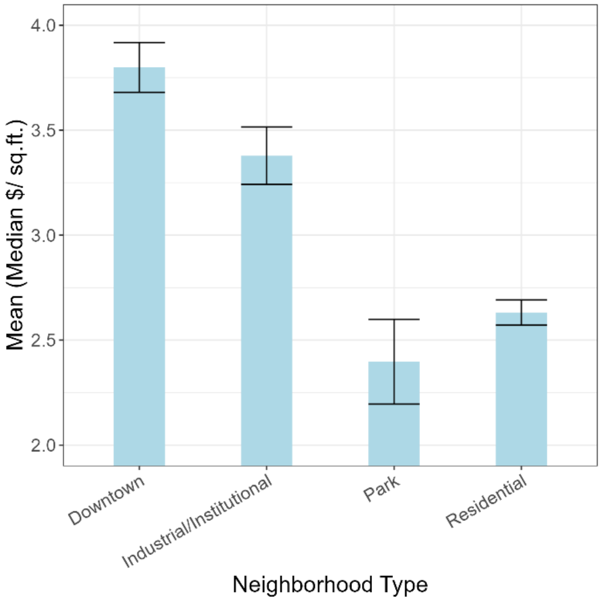


Figure 8 ANOVA of Neighborhood Types with Median Price/ sq.ft.

Interestingly, the residential-downtown type has the second-highest mean price difference. One would expect that a residential area is a highly sought place to live for a peaceful and safe neighborhood. But as proven in the correlation analysis above, those busy areas such as downtown and industrial/institutional area have a higher correlation with the median price per square feet variable than any other parameters. This also supports the importance of business and commercial areas in affecting rental prices in Boston.

Regression Analysis

The result of the regression analysis is shown in the Appendix. Four models were tested, and a final model was generated to include the features with the highest explained variance. The three models tested are the top 5 features with the highest correlation values, places of interest variables, race variables, and other demographic variables.

From the correlation section above, the highest correlation factor found between median price per square feet is commute-related demographics. Thus, those demographic variables that have correlation coefficients magnitude of > 0.5 are further explored with linear regression analysis. The main independent variables are AgeU18 (age under 18), Commute1030 (between 10-30 minutes), (commute) ByAuto, and (commute) ByWalk. Other independent variables are also analyzed for completion.

For the first model, the multiple R-squared (0.64) and adjusted R-squared (0.63) are >0.5. This means that the model explains ~60% of the variance in the median price per square feet variable, which is relatively good considering there are so many factors contributing to property pricing! The residual standard errors are relatively small (0.5 from a range of 0 to 100). The coefficient standard error is quite large considering all values are <1.

The p-value for this model is < 0.05, indicating that the independent variable is quite significant (a low likelihood that the relationship happens by chance). The highest significance is the Commute1030 variable, followed by the AgeU18 variable. ByAuto variable is the lowest (least significant) but still <0.05. The linear model analyzed unstandardized beta (coefficient or slope), hence, the QuantPsyc library is used below for standardized beta value.

Standardized betas value is equal to R (D. T. O'Brien 2022). The R^2 values indicate the percentage of each variable explaining the variance in the dependent variable. The highest is Commute1030, but it only explained 12% (0.35^2) of the variance in the dataset. Even with all those 4 variable variances combined, they only explained 26% of the price variance, indicating there are many other factors contributing to the price values.

The second model is the regression analysis of median_priceSqft with POI variables. However, the model has a low R-squared value, so it is not a good model. However, the standardized beta value for n_food is the highest so far and the p-value is also infinitesimal. It indicates that this variable is a good feature to be included in the final model.

The third model is the regression of median_priceSqft with Race variables. This model resulted in an R-squared value of 0.35 (low fitting). The correlation coefficient with White and Black variables was >0.5. The standardized beta value for the White variable is the highest so far, but it is greater than one. This might happen if it is multi-collinearity. The p-value is > 0.05 though, so it is not significant. This might not be a good feature for the final model. Based on the standardized beta value, the Black or Asian variable should be included in the final model.

Regression of median_priceSqft with other demographic variables (fourth model) is also conducted to check the importance of the feature to the price value. The R^2 of this model is not relatively low (<0.5). However, the White and Population Density variables have significant beta values. These might be important features for the final model.

Considering the linear model results from the four previous models, the variables with the highest beta values are used in the final model. The independent variable is limited to 5 variables: Asian, Commute1030, AgeU18, ByWalk, and AREA_ACRES. The chosen variables also indicate different areas of the demographic and socioeconomic measures to avoid similarity. This ensures the model is not overcrowded and reduces the chance of multi-collinearity. The Asian race is chosen among all the races as it gave the highest R-squared and beta coefficient in the final model. The AREA_ACRES can also represent the population density of each census tract.

The regression plot for each variable is shown below (except AREA_ACRES). All the variables have a positive relationship with the median listing price except the Age under 18 years old variable. The plot shows high variance (scattered dots) of all four variables.

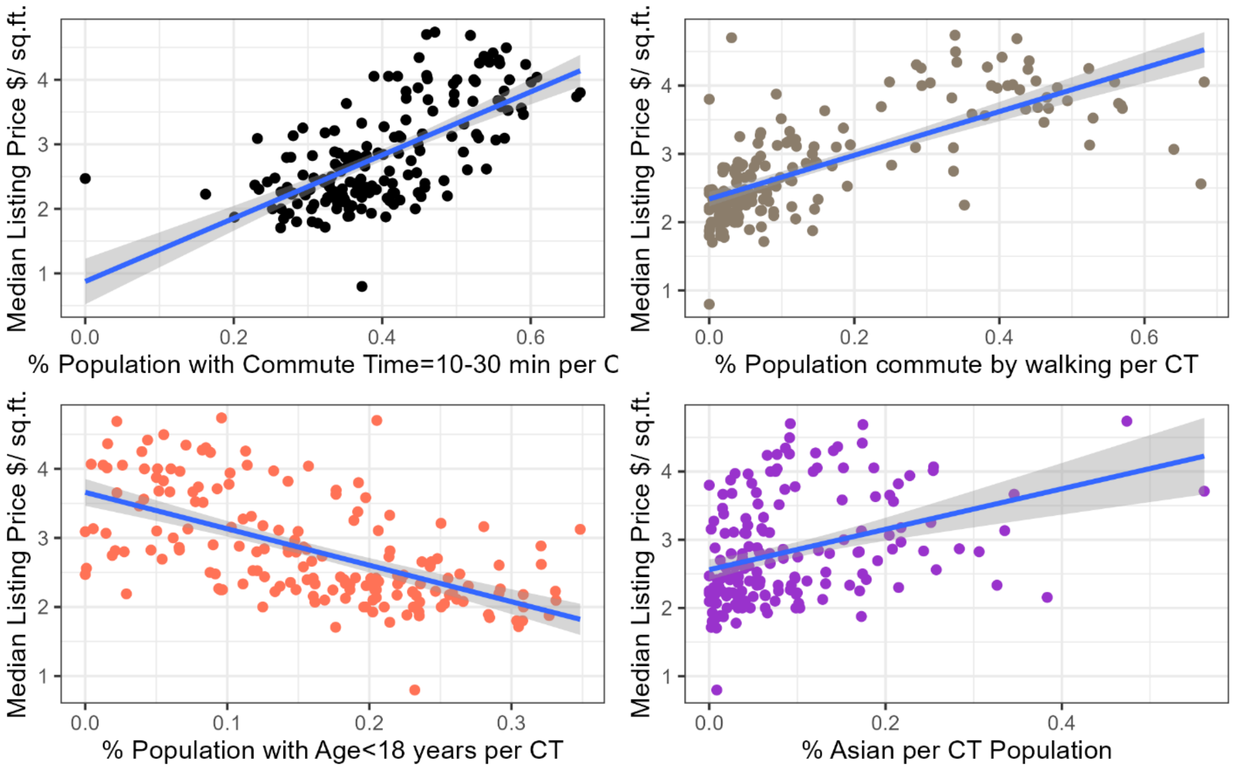


Figure 9 Linear Regression Plot of the final independent variables to explain the variance in Craigslist median listing price per sq. ft.

The final model explained 65% of the variance in the median listing price, which is 2% higher than the first model. It should be noted that a simple linear model was utilized in this analysis. It should be noted that there is no single independent variable that can explain the variation in listing price for more than 14%, which shows there are numerous factors that impact the property listing price.

Analysis of feature importance with boosting or bagging regression model will show a better feature importance rank. The neural network might better explain the complicated relationship between demographic and housing prices. We can arrange from best to worst beta and take only the top features. Another way is to conduct PCA. However, the contribution of each feature is unknown.

Conclusion

Community housing insights were discovered through an analysis of Craigslist rental house trends and distribution in the City of Boston. The median listing price for Boston is $2,500, which is significantly higher than the Massachusetts median of $1,895. The Craigslist data shows more rental property available (posted) during summer months. More than 50% of the property is cat-friendly although fewer options are for dog owners (40%). One should expect slightly higher rental prices when pets are allowed on the property.

From the t-test conducted for allowance of pets in the rental property, it is found that the property that allows dogs has a $273 higher mean listing price than the one that does not allow dogs. The p-value is very low, and the t-value is high, so these numbers are significant. This supported the findings in earlier analysis with an exact value. This means the difference is lower in the property that allows cats. This is interesting since it implies that there are properties that allow cats only and not dogs.

When comparing whether the property is an apartment or a house, the mean price difference from the t-test is found to be lower. It is expected that the mean difference will be higher when the property is a house, and it is found to be $175 higher with high significance. Therefore, tenants should expect an approximately $175 higher rental price for a house compared to an apartment in the Boston area.

Merging the Craigslist dataset with the Places of Interest and Census Indicators facilitates more insights into the listing price trend, especially the property desirability and whether some demographic explained higher/lower listing prices. Since Craigslist is only a subset of the rental properties available in the neighborhood, it can only illustrate part of the pulse of the neighborhood. The Census Indicators, articles, and BARI maps complete the picture for our neighborhood analysis. Therefore, caution should be exercised when analyzing a specific dataset obtained from one source. Another example is the Google Places of Interest dataset has no record of transit stations in the neighborhood while there are a lot of bus and streetcar stations.

The correlation analysis proves some of the hypotheses about places of interest (POI), especially the relationship between median listing price and food & entertainment amenities, but to a smaller extent. Both POI categories have p-values < 0.001 (significant variables). A medium correlation between the places of interest variable especially food-related, entertainment, and grocery supports the hypothesis that amenities (commercial areas) tend to cluster together like in malls and plazas.

The highest correlation value with the median price is found in a census tract that has the highest percentage of people community commuting by walking (0.752). This is also supported by the negative correlation between commuting by car and with median listing price per square foot. Compared to the highest correlation with food-related POI (0.316), the correlation analysis indicates that commuting places a higher importance in determining the listing price of a property (demand) than amenities (places of interest). This is also supported by a medium positive correlation with age between 18 – 34 years old (university students and young professionals).

Another important correlation, to a lesser extent, is the impact of population density on the listing price. The medium correlation between the White and Black races with the listing price might show that one race is found living in a certain range of listing price more than the others.
A multi-group analysis using ANOVA was conducted for the aggregated dataset between median listing price per square feet (dependent variable) and types of the neighborhood (independent variable). The types of neighborhoods are grouped into four categories: “Downtown”, “Industrial/Institutional”, “Park”, and “Residential”.

It was expected that Downtown would have the highest mean price difference with other groups followed by the residential type. However, it is found that although downtown has the highest price, the next highest price is in the industrial/institutional area. This is intriguing since house-type listings with higher mean listing prices should be in residential areas. However, not many house-type listings are found in the Boston area so the lack of samples might not impact the mean price much. These results also show the impact of the university area (institutional) on the City of Boston, with higher prices found in this neighborhood.

In addition, the highest mean price difference is found between Park and Downtown neighborhood types. This is reasonable since downtown property tends to have a lower price and more parks are likely to be found in suburban areas. These findings supported the correlation analysis before that busy area such as downtown and industrial/institutional area have a higher correlation with a median price per square feet than any other parameters. This also supports the importance of business and university areas in affecting rental prices in the City of Boston.

Regression analysis using a linear model for base R is conducted for the Craigslist Housing Dataset. Since price is the only numerical value in the dataset, it is used as the dependent variable. The regression analysis discovered that the Commute1030, ByWalk, and AgeU18 variables are making a significant contribution to the variance of the median listing price.

A linear model was also analyzed using other independent variables to ensure no important variable was left behind. It is found that one of the race variables (esp. Asian) and Area of CT (AREA_ACRES) also plays a role in determining the price, with less degree for the latter. The final model is built based on these 5 independent variables. The R-squared value of the linear model is 0.65 (both multiple and adjusted). Since the value is > 0.5, this model fits the listing price dataset quite well. The squared error is also relatively small. The p-value is close to 0 (infinitely small value).

Based on the standardized beta values, however, there is no single independent variable that can explain the variation in listing price for more than 14%, which shows there are numerous factors that impact the property listing price. The demographics and econometrics findings discussed here are only one facet of the social phenomenon.

In conclusion, clusters of places of interest and certain demographic indicators are affecting the rental listing price in the City of Boston. The main indicators are commuting method by walking, younger age range (<34 years old), lower commute length (<30 minutes), and proximity to food-related and entertainment services. This is reflected in higher-priced properties when it is closer to downtown. The importance of lower commute length and walking method are reflected in the renters' willingness to spend on a rental property; convenience is preferred when picking a place to live in. However, these indicators only explain less than 50% of the variance in price, which shows the complexity of long-term rental pricing and many factors affecting it.

References

2021. *Boston Area Research Map.* Northeastern University. Accessed 02 08, 2023. https://boston-area-research-map-nu.hub.arcgis.com/ .

2020. *Boston Neighborhoods.* 12 07. Accessed 2023. https://data.boston.gov/dataset/boston-neighborhoods.

2023. *Boston Planning & Development Agency.* Accessed 2023. https://www.bostonplans.org/getattachment/e35cecef-cf91-41ef-b28a-f3f227a059d8/.

Bungalow. 2022. *The 9 Best Neighborhoods to Live in Boston in 2022.* Bungalow. 03 06. Accessed 02 07, 2023. https://bungalow.com/articles/best-neighborhoods-in-boston-massachusetts.

Castro, Edgar, Mehrnaz Amiri, Riley Tucker, and Daniel O'Brien. 2020. *Craigslist Postings in Massachusetts.* 08 24. Accessed 2023. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/52WSPT.

Castro, Edgar, Mehrnaz Amiri, Riley Tucker, and Daniel T. O'Brien. 2020. *Harvard Dataverse.* Boston Area Research Initiative's Boston Data Portal. 08 24. Accessed 01 23, 2023. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/52WSPT.

Gibbons, Sage, Edgar Castro, Petros Papadopoulos, Riley Tucker, and Dan O'Brien. 2020. *Places of Interest.* 08 25. Accessed 2023. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/V2UTWH.

Help, Transit Partners. 2023. *Create your data feed: Stops & Stations.* Accessed 2023. https://support.google.com/transitpartners/answer/6377423?hl=en.

Jeffrey Stupak, Lida R. Weinstock. 2023. *Introduction to U.S. Economy: Housing Market.* 01 03. Accessed 04 22, 2023. https://sgp.fas.org/crs/misc/IF11327.pdf.

Kollar, Jessica Semega and Melissa. 2022. *Income in the United States: 2021.* U.S. Census Bureau. 09 13. Accessed 02 08, 2023. https://www.census.gov/library/publications/2022/demo/p60-276.html.

O'Brien, Dan, Alexandra Ciomek, and Riley Tucker. 2019. *Massachusetts Census Indicators.* Boston Area Research Initiative, Northeastern University. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XZXAUP&version=3.0.

O'Brien, Daniel T. 2022. "12. Unpacking Mechanisms Driving Inequities: Multivariate Regression." In *Urban Informatics: Using Big Data to Understand and Serve Communities*. Boston: Boca Raton, FL, Chapman Hall / CRC Press. https://ui.danourban.com/mapping.html#mapping.

Ryan, Bill, Karina Ward, and Kellie Pederson. 2022. *Housing (A part of the Downtown Market Analysis Toolbox program).* Accessed 04 22, 2023. https://economicdevelopment.extension.wisc.edu/articles/evaluating-housing-opportunities/.

Slant. 2023. *Foursquare vs. Google Maps.* Accessed 2023. https://www.slant.co/versus/2576/6934/~foursquare_vs_google-maps.

Tanumihardja, Jessica. 2023. *SPPUA5262 Big Data for Cities.* WordPress. 01. https://sppua5262.wordpress.com/2023/01/30/pulse-of-the-city-craigslist-housing/.

Appendix – Methodology

R Studio is utilized to process the dataset. The snippet of the codes necessary for the analysis is shown below.

Exploratory Data Analysis

```
#load required packages
library(tidyverse)
library(tinytex) #for knitting
library(stats) #for splicing and processing, if needed
library(ggplot2) #for plotting
```

```
#load the dataset and view
df_cl <- read.csv("/dataverse_files/CRAIGSLIST.Listings.csv")
```

```
dim(df_cl) #returns both numbers of rows and columns

## [1] 205450     14

summary(df_cl) #get the basic statistical description for numeric variables and the number of missing values

#set LISTING_MONTH as a factor instead of a character

df_cl$LISTING_MONTH<-as.factor(df_cl$LISTING_MONTH)

sapply(df_cl, class) #after: returns data type per columns

##    LISTING_ID  LISTING_YEAR LISTING_MONTH   LISTING_DAY  LISTING_TIME
##     "numeric"     "integer"      "factor"     "integer"   "character"
##  RETRIEVED_ON          BODY         PRICE     AREA_SQFT   ALLOWS_CATS
##   "character"   "character"     "integer"     "numeric"     "integer"
##   ALLOWS_DOGS       ADDRESS      LOCATION      CT_ID_10
##     "integer"   "character"   "character"     "numeric"
```

From the CRAIGSLIST.DOCUMENTATION.2021, the variables are:

- LISTING_ID is a unique identifier for each listing that is generated from the listing's URL address.
- LISTING_YEAR indicates the year the listing was posted online
- LISTING_MONTH indicates the month the listing was posted online
- LISTING_DATE indicates the day of the month the listing was posted online
- LISTING_TIME indicates the time the listing was posted online
- RETRIEVED_ON indicates the date and time the BARI scraper collected the listing information
- BODY is a string variable that includes the text from the main body section of the listing.
- PRICE is the price of the listed property in dollars.

- AREA_SQFT is the area of the listed property in square feet.
- ALLOWS_CATS is a binary variable where “1” indicates that the listed property allows cats, and “0” indicates the property does not allow cats.
- ALLOWS_DOGS is a binary variable where “1” indicates that the listed property allows dogs, and “0” indicates the property does not allow dogs.
- ADDRESS is the address posted for the listing.
- LOCATION is a string variable containing the location posted with the listing.
- CT_ID_10 is the 2010 Census Tract ID number.
  - This is found by extracting longitude and latitude from listed locations and spatially overlaying them onto BARI’s Census Tracts shapefile.

Missing Values

Handling missing values, the reference here: missing values.

```
# remove the whole row NA first
df_cl2 <- filter(df_cl, rowSums(is.na(df_cl)) != ncol(df_cl))
dim(df_cl2) #only removed the 5 rows containing all NAs
```

```
## [1] 205445     14
```

```
summary(df_cl2) #get the basic statistical description for numeric variables
and the number of missing values
df_cl2$PRICE_PER_SQFT <- df_cl2$PRICE / df_cl2$AREA_SQFT
```

Filter only City of Boston postings & Outliers

Use the `Tract Census Data` provided in Module 10 to get types of neighborhoods per CT ID

```
#from Module 10
tracts_census <- read.csv("//Modules/Tract Census Data.csv")

#get the neighborhood name from the tracts_geo file
tracts_census_sub <- tracts_census %>%
  select(CT_ID_10, BRA_PD_ID, BRA_PD, ISD_Nbhd) %>%
  distinct()
```

```
# filter only in Boston's CT
CT_boston <- unique(tracts_census_sub$CT_ID_10)

# filter Craigslist dataset to the Boston area only per CT ID
df_cl_boston <- df_cl2 %>%
  filter(CT_ID_10 %in% CT_boston) %>%
  filter(PRICE > 800 & PRICE <5500) %>% #filter unreasonable price
  filter(AREA_SQFT >200 & AREA_SQFT<3000) #filter unreasonable area

#filter unreasonable price and area by eliminating 1% of min and max samples
  #refer to assignment 3
```

Add the additional variable: Apartment or house from the description

```
df_cl_boston$APT <- as.numeric(str_detect(
  df_cl_boston$BODY,
  paste(c("Apartment", "apartment", "apt", "multi-family", "multi-family"),
        collapse = '|'))) #to include all possibility
table(df_cl_boston$APT)

#use the same df above
df_cl_boston$HOUSE <- as.numeric(str_detect(
  df_cl_boston$BODY,
  paste(c("House", "house", "single-family", "single-family"),
        collapse = '|'))) #include all possibility of house
table(df_cl_boston$HOUSE)
```

Merge with the tract census dataset to add neighborhood names:

```
#merge with the dfs
df_cl_boston_nb <- merge(df_cl_boston, tracts_census_sub,
                 by='CT_ID_10',
                 all.x=TRUE) #want all rows from df_cl_boston

#if want to write the df in folder
#write.csv(df_cl_boston_nb, "C:/Users/jtanu/Documents/Northeastern/2023SPRING
/PPUA5262/Assignment/df_cl_boston_nb.csv", row.names=FALSE)
```

Merge Dataset: Google Places

```
#load the dataset and view
GooglePlaces <- read.csv("Assignment/dataverse_files/GooglePlaces.POI.csv")

## Data cleaning
#convert CT as numeric, format it as numeric
GooglePlaces$CT_ID_10 <- as.numeric(GooglePlaces$CT_ID_10)

## Warning: NAs introduced by coercion

#remove NA for GP
GooglePlaces <- GooglePlaces %>% filter(!is.na(CT_ID_10))
#check unique CT_ID after NA removal
length(as.list(unique(GooglePlaces$CT_ID_10)))

## [1] 181
```

Get the designated POI count:

```
#create dataframe to calculate number of desired places of interest within a
given CT
  #choose to use group_by instead of merge() so we can create multiple column
at the same time
```

```r
GP_POI <- GooglePlaces %>%
  group_by(CT_ID_10) %>%
  summarise(n_food = sum(Tag_1 %in% c("restaurant", "bakery", "bar",
                                      "cafe", "meal_takeaway", "food")),
            n_grocery = sum(Tag_1 == "grocery_or_supermarket"),
            n_transit = sum(Tag_1 %in% c("transit_station", "bus_station",
                            "train_station", "subway_station",
"light_rail_station")),
              #transit supposed includes bus, train, subway, and lightrail
              #can also include OR for Tag_2 if desired
            n_convenience = sum(Tag_1 == "convenience_store"),
            n_school = sum(Tag_1 == "school"),
            n_university = sum(Tag_1 == "university"),
            n_entertainment = sum(Tag_1 %in% c("zoo", "stadium",
"bowling_alley","movie_theater", "art_gallery", "night_club", "aquarium",
"museum", "shopping_mall")),
            n_park = sum(Tag_1 == "park"))
```

Aggregate and merge POI (Google Places) dataset

```r
#aggregate df_cl2
df_cl_agg <- df_cl2 %>%
  group_by(CT_ID_10) %>%
  summarise(n_postings = n(),
            median_price = median(PRICE),
            mean_price = mean(PRICE),
            min_price = min(PRICE), #maybe use quantile 10% to exclude
outliers?
            max_price = max(PRICE), #maybe use quantile 90% to exclude
outliers?
            #note: the aggregate below removes NA so we still get values
            #does not mean it represents all postings in the area
            median_priceSqft = median(PRICE_PER_SQFT, na.rm = TRUE),
            n_allowsCat = sum(ALLOWS_CATS, na.rm=T),
            n_allowsDog = sum(ALLOWS_DOGS, na.rm=T))

#inner join two df to get CT ID in boston only
df_cl_boston_POI <- df_cl_agg %>%
  inner_join(GP_POI, by = "CT_ID_10")

#add total column
df_cl_boston_POI <- df_cl_boston_POI %>%
  mutate(n_allPOI = rowSums(across(n_food:n_park)))
```

Aggregate and merge Census Indicator (ACS) dataset

```r
#load census indicator
  #to get more variables, we will use "ACS_1418_TRACT.tab" data (2014-2018 AC
S Tracts indicator)
```

```
census_indicator <-read.csv("//Assignment/dataverse_files/ACS_1418_TRACT.csv
")

df_clBos_POI_CI <- merge(df_cl_boston_POI,census_indicator,
                  by='CT_ID_10',all.x=TRUE) #want all rows
```

Boston Neighborhood shapefile (for Mapping)

```
library(sf)

## Linking to GEOS 3.9.3, GDAL 3.5.2, PROJ 8.2.1; sf_use_s2() is TRUE

##Import a shapefile##
tracts_geo<-st_read("/Assignment/dataverse_files/Tracts_Boston_2010_BARI/Trac
ts_Boston BARI.shp")

boston_nb<-st_read("/Assignment/dataverse_files/Boston_Neighborhoods/Boston_N
eighborhoods.shp")

#load census indicator
  #to get more variables, we will use "ACS_1418_TRACT.tab" data (2014-2018 AC
S Tracts indicator)

census_indicator <-read.csv("Assignment/dataverse_files/ACS_1418_TRACT.csv")
tracts_geo_CI <- merge(tracts_geo,census_indicator,
                  by='CT_ID_10',all.x=TRUE) #want all rows

tracts_geo_CI_cl <- merge(tracts_geo_CI, df_cl_boston_POI,
                  by='CT_ID_10',all.x=TRUE) #want all rows
```

Visualization: Mapping

All the mapping and visualization are produced using ggplot2 or its extension. The sample codes are shown below (some codes are eliminated due to similarity). Only the tracts with total population of greater than 500 is plotted to filter outlier.

```
#mapping using data from census indicators

#map median listing price

map_medIncome <- ggplot() +
  geom_sf(data=tracts_geo_CI_cl[tracts_geo_CI_cl$TotalPop>500,], #filter outl
ier
          aes(fill=MedHouseIncome),
          inherit.aes = FALSE) +
  scale_fill_viridis_c(option = "magma") + #limits=c(20,3000)
  labs(fill="Med. Hh Income ($)")+
  theme_bw() +
  theme(legend.position = "bottom",
        legend.text = element_text(size=8))
```

```
map_medIncome

#map number of postings

map_nPost_cl <- ggplot() +
  geom_sf(data=tracts_geo_CI_cl[tracts_geo_CI_cl$TotalPop>500,], #filter outl
ier
          aes(fill=n_postings), inherit.aes = FALSE) +
  #scale_fill_gradient(high = "red", low = "green",
  #                     limits=c(400,3500))+
  scale_fill_viridis_c(limits=c(20,3000)) +
  labs(fill="Posting count")+
  theme_bw() +
  theme(legend.position = "bottom",
        legend.text = element_text(size=8))

map_nPost_cl

#map median price per square feet

map_medPriceSqft_cl <- ggplot() +
  geom_sf(data=tracts_geo_CI_cl[tracts_geo_CI_cl$TotalPop>500,], #filter outl
ier
          aes(fill=median_priceSqft),
          inherit.aes = FALSE) +
  scale_fill_gradient(high = "red", low = "green",
                      limit = c(1.5,4.25))+
  #scale_fill_viridis_c() +
  labs(fill="Median $/sqft")+
  theme_bw() +
  theme(legend.position = "bottom")
        #legend.text = element_text(size=8))

map_medPriceSqft_cl

#put the plots side by side (can also use gridExtra shown next chunk):
library(patchwork)
map_medIncome + map_nPost_cl + map_medPriceSqft_cl
```

Correlation

Places of interest correlations:

```
#Follow Chap 10.3 Urban Informatics textbook
# need to identify is numeric
df_cl_boston_POI %>%
  select(where(is.numeric)) %>%
  na.omit() %>%
  cor()

# check significance by measuring p-values
```

```r
library(Hmisc)

rcorr_matrix_clPOI <- df_cl_boston_POI %>%
  select(where(is.numeric)) %>% #if error, rerun library(tidyverse)
  na.omit() %>%
  as.matrix() %>%
  rcorr()
rcorr_matrix_clPOI[3]

#visualize correlation matrix

library(GGally)

df_cl_boston_POI %>%
  select_if(is.numeric) %>%
  na.omit() %>%
  ggpairs(column = c(2,7, 10:12, 15:16)) #pick important columns to reduce figure size
```

Census indicator correlations:

```r
#filter only is numeric
library(Hmisc)
rcorr_matrix_CI <- df_clBos_POI_CI %>%
  select(is.numeric) %>%
  na.omit() %>% #omit rows with NA so that we can get values for all variables
  as.matrix %>%
  rcorr()
#call the first matrix (r-value)
rcorr_matrix_CI[1]

#plot per category: Econometrics

library(GGally)

df_clBos_POI_CI %>%
  select_if(is.numeric) %>%
  na.omit() %>%
  ggpairs(column = c(2,7, 19:23, 32)) #pick columns per category

#plot per category: Race
library(GGally)

df_clBos_POI_CI %>%
  select_if(is.numeric) %>%
  na.omit() %>%
  ggpairs(column = c(2,7, 26:30, 36)) #pick important columns to reduce figure size

#plot per category: Commute method
library(GGally)
```

```
df_clBos_POI_CI %>%
  select_if(is.numeric) %>%
  na.omit() %>%
  ggpairs(column = c(2,7, 49:57)) #pick important columns to reduce figure si
ze
```

ANOVA analysis

```
#get the correct column for anova
anova_type <- aov(median_priceSqft~Type, data=df_clBos_POI_CI_nb)
summary(anova_type)

##               Df Sum Sq Mean Sq F value   Pr(>F)    
## Type           3  27.97   9.325   20.58 1.84e-11 ***
## Residuals    174  78.84   0.453                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness

#post hoc analysis to get pairwise comparison between groups
TukeyHSD(anova_type)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = median_priceSqft ~ Type, data = df_clBos_POI_CI_nb)
## 
## $Type
##           diff        lwr        upr     p adj
## I-D -0.4208353 -1.0144803  0.1728098 0.2587233
## P-D -1.4018074 -2.0887115 -0.7149033 0.0000021
## R-D -1.1677924 -1.6962457 -0.6393391 0.0000003
## P-I -0.9809721 -1.5432170 -0.4187273 0.0000653
## R-I -0.7469571 -1.0984466 -0.3954676 0.0000008
## R-P  0.2340150 -0.2589025  0.7269326 0.6076971
```

ANOVA visualization:

```
#need to melt the data to plot based on mean
library(reshape2)

#index for median_priceSqft = 7
#index for Type = 75
melted_type <- melt(df_clBos_POI_CI_nb[c(7,75)],
               id.vars=c("Type"))
means_type <- aggregate(value~Type, data=melted_type, mean)
#levels(means_type$Type) <- c()

#rename column
names(means_type)[2] <- "mean"
```

```
#add standard error
ses_type <-aggregate(value~Type,
               data=melted_type,
               function(x) sd(x, na.rm=TRUE)/sqrt(length(!is.na(x))))
names(ses_type)[2]<-'se'

# add SES to the previous means_type
means_type <- merge(means_type, ses_type,
                    by='Type')
#add lower and upper bars for the standard errors
means_type <- transform(means_type, lower=mean-se, upper=mean+se)

#rename the types for clarity
levels(means_type$Type)<- c("Downtown",
                           "Industrial/Institutional",
                           "Park",
                           "Residential")

#update graph with errorbars
ggplot(data=means_type, aes(x=Type, y=mean)) +
  geom_bar(stat="identity",position="dodge",
           fill="lightblue", width = 0.5) +
  geom_errorbar(aes(ymax=upper, ymin=lower),
                position=position_dodge(.9), width = 0.5)+
  theme_bw()+
  coord_cartesian(ylim=c(2,4))+ #adjust to show differences
  labs(y = "Mean")
```

Regression Analysis

Since the code for different variables are similar, only the highest correlation value variables and the final regression variables are listed here.

```
#Regression in R
regression <-lm(median_priceSqft~ByWalk+Commute1030+ByAuto+AgeU18, data=df_cl
Bos_POI_CI)
summary(regression)

##
## Call:
## lm(formula = median_priceSqft ~ ByWalk + Commute1030 + ByAuto +
##     AgeU18, data = df_clBos_POI_CI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56120 -0.25440 -0.02081  0.25967  1.91975
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2100     0.2294   9.634  < 2e-16 ***
## ByWalk        1.0539     0.3954   2.665 0.008443 **
## Commute1030   2.6901     0.4558   5.902 1.93e-08 ***
## ByAuto       -0.7142     0.3109  -2.297 0.022842 *
## AgeU18       -1.9014     0.5363  -3.545 0.000508 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4749 on 168 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.6353, Adjusted R-squared:  0.6266
## F-statistic: 73.15 on 4 and 168 DF,  p-value: < 2.2e-16
```

Calculate standardize beta:

```
##Observe standardized betas##
require(QuantPsyc)

lm.beta(regression)

##      ByWalk Commute1030      ByAuto      AgeU18
##   0.2374101   0.3589061  -0.1685796  -0.2145562
```

Final regression variables (kept below 5 variables):

```
#Regression with median_price
regression_final <-lm(median_priceSqft~Asian+Commute1030+AgeU18+ByWalk+AREA_A
CRES, data=df_clBos_POI_CI)
class(regression_final)

## [1] "lm"

summary(regression_final)

##
## Call:
## lm(formula = median_priceSqft ~ Asian + Commute1030 + AgeU18 +
##     ByWalk + AREA_ACRES, data = df_clBos_POI_CI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51119 -0.28959 -0.02119  0.29227  1.88053
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7614219  0.2457279   7.168 2.42e-11 ***
## Asian        0.8468899  0.4479347   1.891 0.060423 .
## Commute1030  3.0250504  0.4998257   6.052 9.30e-09 ***
## AgeU18      -1.9216451  0.5714862  -3.363 0.000960 ***
## ByWalk       1.1527520  0.3364342   3.426 0.000772 ***
## AREA_ACRES  -0.0005720  0.0001992  -2.871 0.004623 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4679 on 165 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.6516, Adjusted R-squared:  0.641
## F-statistic: 61.71 on 5 and 165 DF,  p-value: < 2.2e-16
```

```
#calculate standardized betas
lm.beta(regression_final)
```

```
##       Asian Commute1030      AgeU18      ByWalk  AREA_ACRES
##   0.1021222   0.3828314  -0.2128783   0.2585127  -0.1390901
```

```
View(data.frame(regression_final$residuals))
```

```
#save residuals
#View(data.frame(regression_final$residuals))
df_clBos_POI_CI <-
merge(df_clBos_POI_CI,data.frame(regression_final$residuals),
                      by='row.names', all.x=TRUE)
```

Visualization for regression:

###Commute1030

```
plot_commute<-ggplot(data=df_clBos_POI_CI, aes(x=Commute1030, y=median_priceS
qft))+
  geom_point() +
  xlab("% Population with Commute Time=10-30 min per CT") + ylab("Median List
ing Price $/ sq.ft.")+
  geom_smooth(method=lm)+
  theme_bw()

plot_commute
```

Similarly, plot other variables and combine into one figure.

```
#save plot
plot_final <- arrangeGrob(plot_commute, plot_walk, plot_age, plot_asian, nrow
=2)

ggsave(file="Assignment9_regressionPlotFinal.png", plot_final)
```