Uncovering how and why the Average Listing Price varies across Massachusetts Craigslist Data

Executive Summary

The average pricing of a listing in Massachusetts differs across census tracts due to various intrinsic factors such as seasonality, the square footage of a listing, its pet policy and so forth.

The relationship between the size of a listing (square footage) and its price did not have a straightforward and substantial positive correlation thus suggesting that more analysis into the extrinsic factors was needed and would be more informative.

I chose to look at extrinsic factors such as the median income and the education rates. The analysis gave proof of the fact that higher the education rates, the higher is the median income and the higher the median income of a census tract, the higher was the average price of a listing.

Various regression models, that are explained in a detailed manner in latter sections, have been written and their performances compared by assessing the model report and summaries. In this manner, in addition to understanding how such driving factors shape valuation, a more comprehensive approach can be leveraged to optimize pricing strategies.

Introduction

During the course, I came across an article that spoke about the topic of affordable housing and what it does to the nearby property values. The article spoke of the various results of research that spoke about how it either increased(more often than not) the property valuation of places by small margins or decreased the valuation by very small margins in regions that had a very high median income.
I thought that this research was interesting and wanted to delve more into how the median income of a census tract affects the average price of a listing in that census tract and to what extent.

I came across another article which spoke about how education and training could affect the economy. The key takeaway was that people with more advanced degrees got paid more by industries and more generally speaking, the countries who had a greater number of people with advanced degrees were seen to be more developed. However, achieving such high forms of education are not a one-step solution/access to a large number of people.

Incorporating equity in data analysis is of utmost importance because I think that it has the power to create imbalances that are more visible when we ask questions such as who this data is about? How and who the results of this analysis will affect? What perspective of a certain demographic does this create?

Thereby, in order to account for equity in my analysis, I added information about the education rates in a neighborhood/census tract. To be specific- I included data of how many people in a census tract had a Graduate/Professional degree and also how many people didn't complete a high school degree. This is one such metric I believe that has an impact on a person's income and what consequently the price bracket that they stay in. Showing the difference in education rates across census tracts will lay out how not everyone has the same opportunities towards attaining high incomes.

This will further help strengthen the correlation and causation between median income and the average listing price in a neighborhood. I think that this analysis will uncover the importance of accessible education to people from all walks of life so as to provide equal income opportunities. And as previously stated, one of the direct consequences of higher incomes is better purchasing power and higher valued homes.

Data and Methods

i) Craigslist and tidycensus data

The Craigslist dataset includes information about housing listings on Craigslist for the state of Massachusetts. This data has been processed and scraped by BARI (Boston Area Research Initiative). The dataset covers all 5 regions of Boston, Cape Cod, South Coast, Worcester and Western Mass and has been aggregated across census tracts. It consists of 205450 observations of 14 variables/features. These are both numerical and non-numerical features such as Listing Year, Month, Day, Body (which consists of textual description of the listing), Allows dogs, cats, Square footage, Price of listing and so forth.

In addition, I've used the tidycensus package available in R to read in external data of median income across census tracts and also data that reports the number of people in each census tract who have either a graduate/professional degree and the number of people who have not completed high school. The easy-to-use get_acs() function, which is used to access the U.S. Census Bureau 2017-2021 American Community Survey(ACS) 5-Year Estimates data, is available under the tidycensus package and is used to read in this data. The 'variables' parameter in this function is set to the data (income or degree attainment information) that we want to read.

ii) Adding new variables

For the sake of preliminary analysis and visualizations, I've chosen to create three new variables to the original dataset namely - ALLOWS_BOTH, Price_Range and Area_Range. I believe that these variables make the data's content more interpretable. The Price range explains what category of price the listings fall under. I've used 3 groups, namely- Cheap, Mid-range and Pricy. I've chosen the median value and 3rd quartile values as limits for the categories. Anything below the median value falls in the first category (cheap), the values in between fall in the second category (mid-range) and the others fall in the last category- Pricy.
Similarly, the Area_Range variable denotes the category of area in sq ft of the listings. According to the bucket it falls in, the listings are either classified as 'Small', 'Mid-sized' and 'Large-Area' spaces. Using the knowledge of what size the listing is and the price category that it falls in, the user can make their pick quicker and in an easier manner. For instance, if a user sees a listing that falls in the 'Cheap' price category and is also a 'Mid-sized' house as opposed to a listing that is 'Pricy' and 'Small' in space, they'll probably pick/ want to explore the former option more quicker. Lastly, I thought it would be beneficial for users to see if the listings allowed both dogs and cats or either pet. It makes for more direct interpretation.

For the sake of modeling, I chose to keep only certain selected columns from the original dataset. In order to combine the above two new datasets to the original modified one, they were merged on the Census Tract ID.
The group by and mean functions were used to find the average sq.ft and average listing price in each census tract. A new variable was created called Grad_or_Prof rate which grouped the census tracts into three separate groups having namely: Highest, Average or lowest Graduate/Professional degree rate.
Similarly another variable called LessThan_HS_rate was created whose categories denoted either Highest or Fewer number of people who have incomplete high school degrees.

In order to choose numerical limits to group the census tracts into each of these categories, I chose to set a record that had a value higher than the mean/average value (number of graduates/people who didn't complete high school) into the highest category. Further, those that lay below the average value got categorized into the lesser category.
The process and logic for the various categorizations and data manipulations are available in the Methodology and code excerpts appendix.

Statistical Analysis

i) Preliminary Analysis and discussion

It was interesting to note as in **Figure 1** how the prices varied across different months. The months of May, June and July saw the highest prices most probably because these months coincide with school/university vacations and most students /families are choosing this particular time frame to rent out a home. In addition, I thought it was noteworthy to realize how the pet policy affected pricing across the state. As shown in **Figure 2**, the listings that allowed both dogs and cats had a higher median price for a listing as opposed to those that allowed either one or none. I attribute this to the fact that listings that allowed pets on their premises would require the owners/hosts to invest more in the upkeep and general cleanliness of their property.

Price vs Listing Month

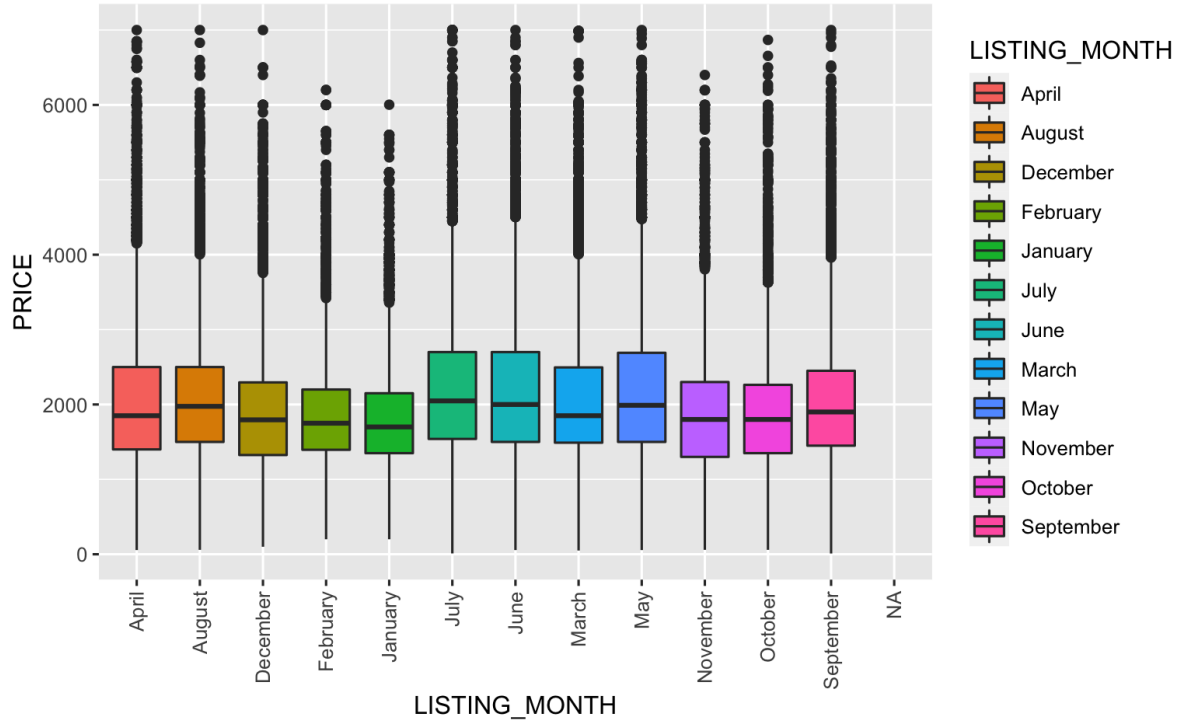

Figure 1.

Price vs Pet Policy

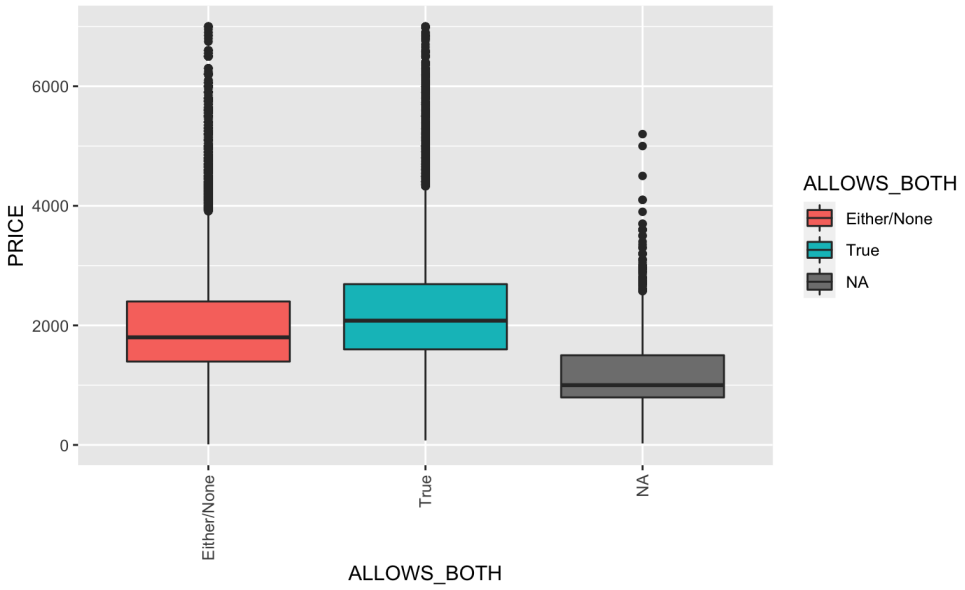

Figure 2.

ii) Initial Regression model, Median Income visualization and discussion

The first regression model I chose to build was a linear regression model to evaluate the impact that predictor variables Area range and Allows_both (indicating pet policy) had on my target variable-Price. The summary of the regression was as below:

```
Call:
lm(formula = PRICE ~ ALLOWS_BOTH + Area_Range, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2451.9  -471.9   -66.2   400.4  4533.1

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2466.929      7.530  327.62   <2e-16 ***
ALLOWS_BOTHTrue      133.416      5.606   23.80   <2e-16 ***
Area_RangeMid-sized -351.094      8.602  -40.82   <2e-16 ***
Area_RangeSmall     -700.716      8.218  -85.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 805.2 on 85084 degrees of freedom
  (120362 observations deleted due to missingness)
Multiple R-squared:  0.09474,   Adjusted R-squared:  0.09471
F-statistic:  2968 on 3 and 85084 DF,  p-value: < 2.2e-16
```

As we can see the P-values for the respective T- tests are much lesser than 0.5 which gave me enough proof that these predictors were statistically significant.

Before moving onto the main section of my analysis, I chose to spatially visualize the distribution of median income across Massachusetts and the result as seen below in **Figure 3**.

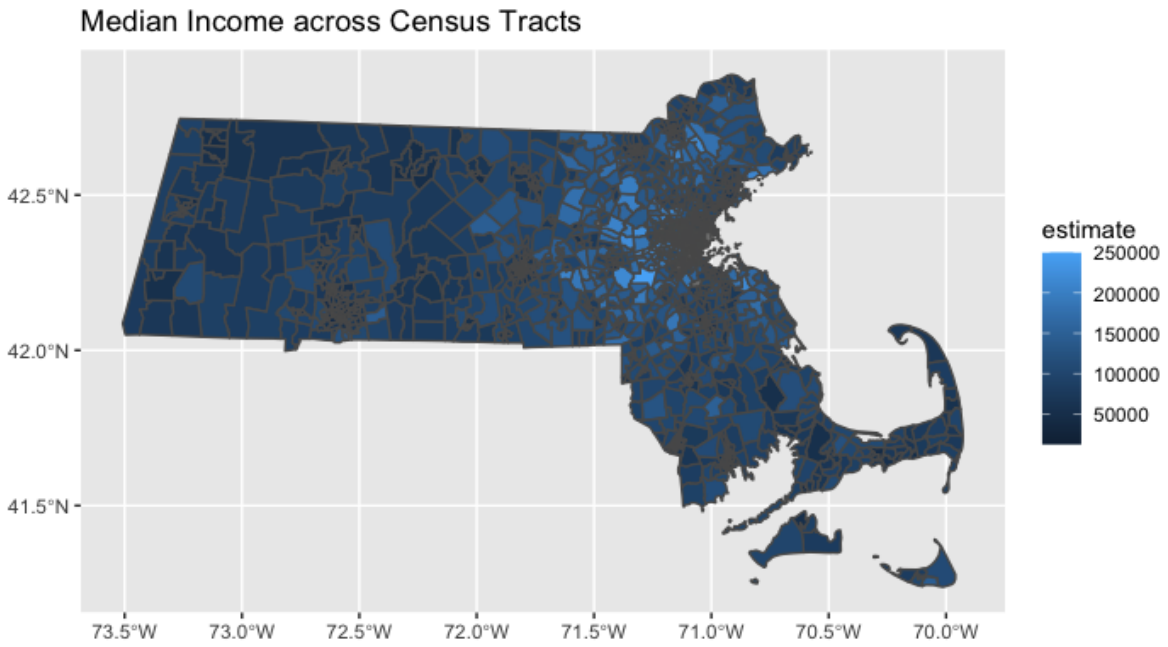

Figure 3.

The distribution of wealth across the state is seemingly obvious. Boston is seen to have the highest median income that's in the range of $200,000 - $250,000. The surrounding areas are also seen to lie in relatively high price brackets. Moving away from Boston in both east and west directions the income brackets are seen to decline.

iii) Hypothesis, Validation and discussion

The initial hypothesis I had was the fact that if a census tract had a higher number of graduates/professionals the average price of listing would be higher. Reason being- advanced degrees give people higher purchasing powers and incomes (which I also hypothesize to be highly correlated to the average listing price) and in turn have more authority in valuing their listings better.

The below correlation plot was generated to test this hypothesis:

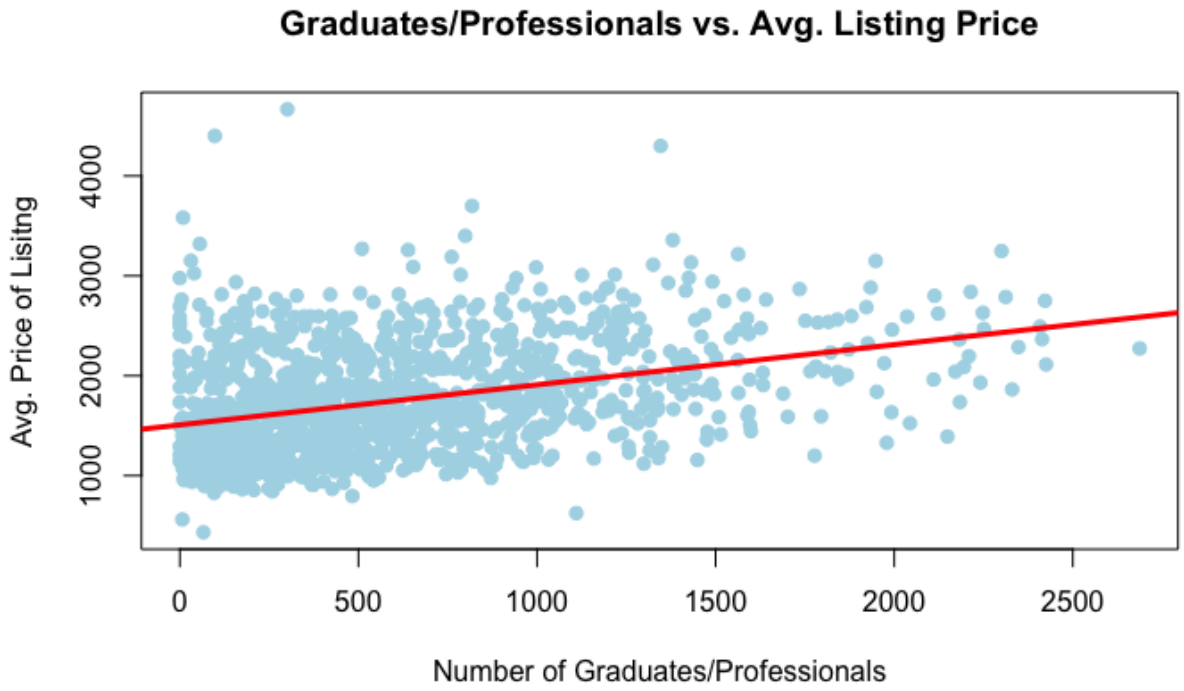

Figure 4.

Of course, as **Figure 4** suggests with its positively correlated trend line, our initial hypothesis is confirmed. Of course, this correlation does not suggest causality. The consequent figures- **Figure 5 & 6** indicate the correlation behind the fact that a greater number of advanced professionals dictate higher median incomes and lastly that tracts with higher median incomes see higher average listing prices. These relationships are enough to draw the conclusion that census tracts that see higher incomes are seen to push out lower income residents as a result of which creates a higher cost of living which brings along with it higher priced listings.

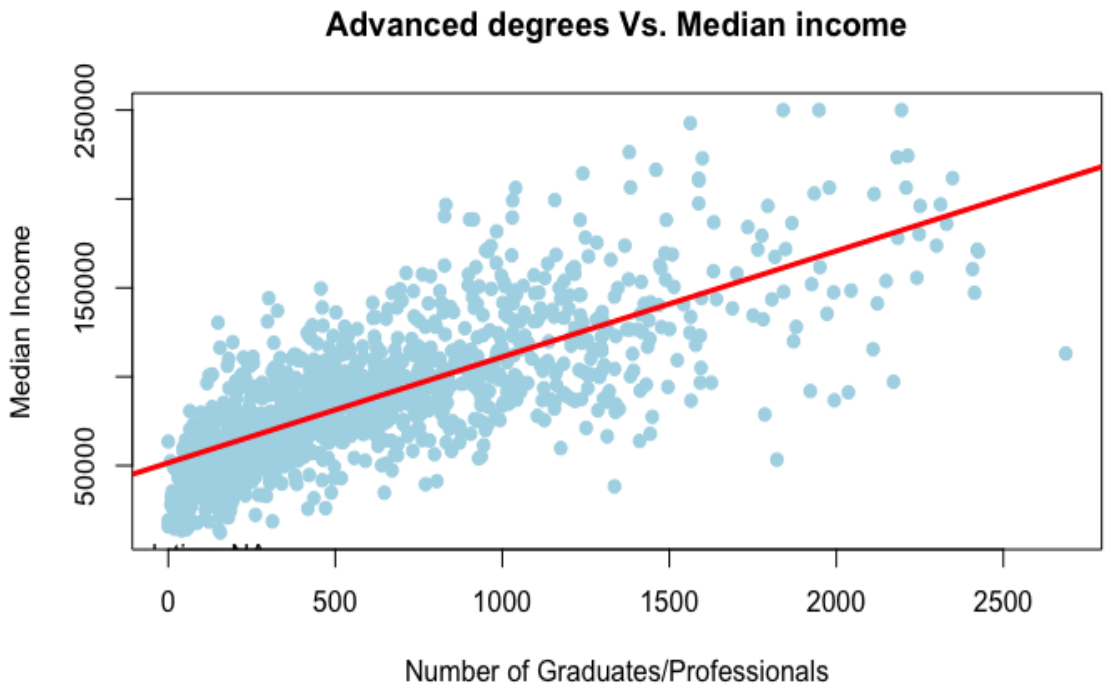

Figure 5.

Figure 6.

The consequent plot only supports this statement further by showing how the median income is negatively correlated with the number of people who have not completed high school. That is, the greater is the number of people with incomplete high school degrees- the lesser is their annual median income.

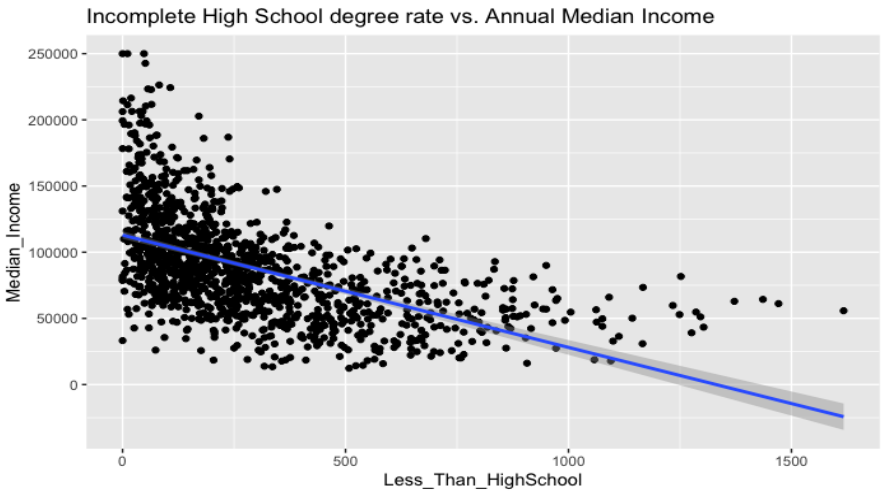

Figure 7.

iv) ANOVA Analysis and discussion

Next, I chose to run ANOVA tests on a one way, two way and three way model. The one way model has been constructed using an independent variable grad_prof_rate that's been constructed using thresholds as mentioned above. The logic behind the thresholding is the fact is as follows: The maximum number of graduate/professional degree holders in a census tract in Massachusetts is 2687, whereas the mean/average is 608. So any census tract that has less than 608 number of professionals is classified as having a "Lesser-Grads/Profs" and one that has >608 and <=1000 are classified as "AverageRate_Grads/Profs". Lastly, if the number is >1000 then this census tract is classified as having "Highest_Grads/Profs".

The below was the summary of my one way model:

```
                Df    Sum Sq  Mean Sq F value Pr(>F)
grad_prof_rate   2  48506072 24253036   91.18 <2e-16 ***
Residuals     1257 334342324   265984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The residual sum of squares we have is quite high as we can see. The p-value of the model is less than 0.001. Hence, the variable Grad_or_Prof definitely has an impact on the target variable- Average Price of a listing.

The similar logic of creating thresholds was used for the creation of one more variable namely: LessThan_HS_rate. This variable signifies the number of people in a census tract that have not completed a high school certification. Next, an ANOVA test was done for a two.way model that incorporates both of these two categorical variables. The summary of this model was as below:

```
                  Df    Sum Sq  Mean Sq F value Pr(>F)
grad_prof_rate     2  43491648 21745824  81.661 <2e-16 ***
LessThan_HS_rate   1     27756    27756   0.104  0.747
Residuals       1010 268956021   266293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The sum of squared residuals of the LessThan_HS_rate variable is lesser than the grad_prof_rate variable. To then select the best model, the AIC based selection was done:

```
Model selection based on AICc:

         K     AICc Delta_AICc AICcWt Cum.Wt       LL
two.way  5 15550.90       0.00      1      1 -7770.42
one.way  4 19319.66    3768.75      0      1 -9655.81
```

The two.way model performed way better than the one.way model.
The two.way model has a lower AIC value and explains 100% of the total variation in the dependent variable (Price) that can be explained by these full sets of models.

It made sense to include the average sqft. of the house as well because there was a positive relationship between the avg. price of a listing and this variable as seen in the gally plot(**Figure 8**). Adding the average sqft. variable to the model makes for our three.way model. AIC based model selection was performed again to assess whether the addition of this variable was a good idea.

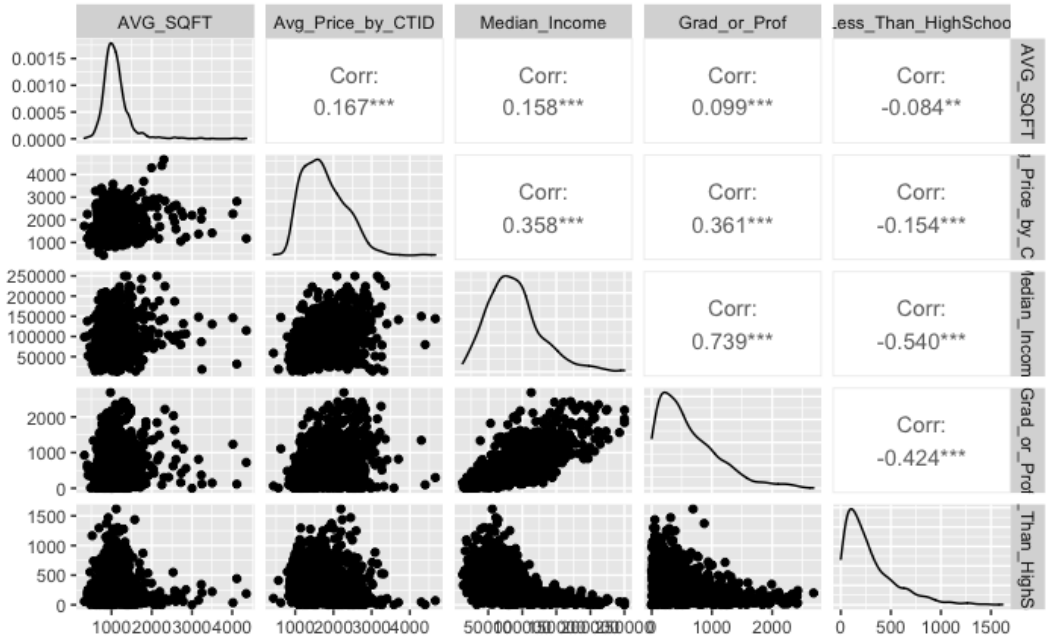

Figure 8.

To finally select the best among all three model, the AIC based selection was done and the result was as below:

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
three.way	6	15541.55	0.00	0.99	0.99	-7764.73
two.way	5	15550.90	9.36	0.01	1.00	-7770.42
one.way	4	19319.66	3778.11	0.00	1.00	-9655.81

The three.way model has the lowest AIC value and explains close to a 100% (99%) of the total variation in the dependent variable (Average Price of listing) that can be explained by these full sets of models.

I finally ran the following regression model after the confidence from my three.way model. The result of the regression was as follows:

```
Call:
lm(formula = Avg_Price_by_CTID ~ grad_prof_rate + LessThan_HS_rate +
    AVG_SQFT, data = df_grad)

Residuals:
     Min       1Q   Median       3Q      Max
-1442.57  -378.10   -59.73   315.88  2869.49

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1663.0091    60.7579  27.371  < 2e-16 ***
grad_prof_rateHighest_Grads/Profs  299.2732    48.2482   6.203 8.08e-10 ***
grad_prof_rateLesser-Grads/Profs  -211.9513    41.9468  -5.053 5.16e-07 ***
LessThan_HS_rateHighest_HS_incomplete  22.9705    42.6236   0.539 0.590063
AVG_SQFT                            0.1505     0.0446   3.374 0.000768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 513.4 on 1009 degrees of freedom
Multiple R-squared:  0.1489,       Adjusted R-squared:  0.1455
F-statistic: 44.12 on 4 and 1009 DF,  p-value: < 2.2e-16
```

Given the size of the dataset. The regression model performs well as indicated by the residual standard error of 513.4 on 1009 degrees of freedom. Moreover, the F-statistic is 44.12 which is way greater than 1. And so we can reject the null hypothesis that there is no relationship between the dependent and target variables. The distribution of the residuals is quite normally distributed which indicates that the predictions are not far off from their actual values.

Conclusion

Education has long since been a key driver of the wealth of a nation, offering individuals to improve their lives and that of their families. It is thus ever more important to recognize in our technical analysis that not everyone has the same access to quality education and training and as evidenced by the data analysis, access to basic forms of education. This inadvertently creates inequities across census tracts and generally speaking across regions.

The lack of quality education can have massive negative impacts, especially on marginalized communities and underrepresented groups who have additional barriers such as poverty, discrimination and so forth. This serves to undermine communities and staggers their tendency for growth. As a result, it creates a vicious and systemic cycle of exclusion.

Quality education as asserted and evidenced through the analysis is a crucial factor towards attaining higher incomes. When the income in a region is high it indicates a higher standard of living. Of course there are other factors that we took into consideration in the analysis while determining what affects the valuation of listings such as the pet policy, the square footage. In addition crime rates, local amenities, school district etc are other factors that could help ascertain their valuation. While all of these and more factors affect property valuation, median income is a key factor and cannot be ignored as it provides insight into the economic prosperity of an area.

The efforts of the overall analysis was to indicate the importance of providing access to quality education so as to give people equal opportunities towards higher purchasing powers and as a direct consequence - property valuations. Using the census tract IDs, one can ascertain which counties/tracts require more attention towards their upliftment. This can be in the forms of scholarship opportunities, increased funding towards public education systems, better integration efforts and so forth.

Bibliography

1. "ANOVA in R," Statsandr, accessed April 24, 2023, https://statsandr.com/blog/anova-in-r/.

2. "Education Affects Income," Investopedia, accessed April 24, 2023, https://www.investopedia.com/articles/economics/09/education-training-advantages.asp.

3. "Impact of Affordable Housing to Nearby Property Values," Bloomberg, May 2, 2022, https://www.bloomberg.com/news/articles/2022-05-02/does-affordable-housing-lower-property-values.

4. Kajal, Neha. "Education Leads to Higher Priced Homes," Bachelor thesis, KTH Royal Institute of Technology, 2019, http://www.diva-portal.org/smash/get/diva2:1346009/FULLTEXT01.pdf.

5. Walker, Kyle. "Basic Usage of the tidycensus Package." Accessed April 24, 2023. https://walker-data.com/tidycensus/articles/basic-usage.html.

Appendix

Methodology and Code Excerpts

The following code section depicts the key steps in coding to add and manipulate the three new variables: Area Range, Price Range and Allows Both.

I've chosen the median value and 3rd quartile values as limits for the categories. Anything below the median value falls in the first category (cheap), the values in between fall in the second category (mid-range) and the others fall in the last category- Pricy. Similarly, the Area_Range variable denotes the category of area in sq ft of the listings. According to the bucket it falls in, the listings are either classified as 'Small', 'Mid-sized' and 'Large-Area' spaces.

The variables ALLOWS_CATS and ALLOWS_DOGS are both denoted with values 1 if they are allowed. Hence, for the new variable ALLOWS_BOTH: if the values under both these attributes add up to 2 (1 present under each column) then ALLOWS_BOTH is set to True. If not, they are set to saying “Either/None” to denote either one pet allowed or none.

```{r}
df <- read.csv('/Users/riddhinarayan/Downloads/CRAIGSLIST.Listings.csv')
#Displaying the first few rows of the data
head(df)
#add 'Price_Range' column based on values in 'PRICE' columns
df$Price_Range <- with(df, ifelse(PRICE> 2500, 'Pricy',
                          ifelse(PRICE > 1895, 'Mid-range', 'Cheap')))
```

```{r}
#add 'Price_Range' column based on values in 'PRICE' columns
df$Area_Range <- with(df, ifelse(AREA_SQFT> 1304, 'Large_Area',
                          ifelse(AREA_SQFT > 925, 'Mid-sized', 'Small')))
```

```{r}
library(dplyr)
df <- transform(
  df, ALLOWS_BOTH= ifelse(ALLOWS_CATS+ALLOWS_DOGS ==2, 'True', 'Either/None'))
```

The following code section depicts the key steps in coding to compute the average price and average square footage of a listing, as well as the addition of three new variables: Number of graduates/professionals, lesser than high school equivalents and median

income across census tracts.

```r
#Load required libraries
library(dplyr)
library(tidycensus)
library(tidyverse)
options(tigris_use_cache = TRUE)

#Read the data
df <- read.csv('/Users/riddhinarayan/Downloads/CRAIGSLIST.Listings.csv')

#Compute average Sq. ft per census tract ID
df_footage<-filter(df, !is.na(AREA_SQFT))
df_footage<-df_footage %>%
  group_by(CT_ID_10) %>%
  summarise(AVG_SQFT=mean(AREA_SQFT, na.rm = TRUE) )
df_footage

df_footage <- setNames(df_footage,
                       c("GEOID", "AVG_SQFT"))

#Compute Avg. Price of Listing per census tract ID
df_agg <- aggregate(df[ , 8], by = list(df$CT_ID_10), FUN = mean)
df_AvgPrice <- setNames(df_agg,
                       c("GEOID", "Avg_Price_by_CTID"))
head(df_AvgPrice)

#Merge by Census Tract ID
df1 <-merge(df_footage, df_AvgPrice, by = 'GEOID')
df1_mod <- filter(df1,df_footage$AVG_SQFT<5000)
df1_mod
```

```r


#Get Annual Median Income data
ma_income <- get_acs(
  geography = "tract",
  variables = "B19013_001",
  state = "MA",
  year = 2020,
  geometry = TRUE
)

head(ma_income)

#Merge median income data
ma_inc_SelectCols <- ma_income[,c("GEOID","estimate")]

#merge on GEOID
df_w_income <-merge(df1_mod,ma_inc_SelectCols, by = 'GEOID')

#Rename columns
df_w_income <- df_w_income %>%
       rename('Census_Tract_ID' = 'GEOID',
              'Median_Income' = 'estimate')
head(df_w_income)

#Get number of people who didn't complete High School per Census Tract
ma_lt_hs <- get_acs(
  geography = "tract",
  variables = "B06009_002",
  state = "MA",
  year = 2020,
  geometry = TRUE
)
```

```r

head(ma_lt_hs)

#Rename columns
ma_lt_hs <- ma_lt_hs %>%
       rename('Census_Tract_ID' = 'GEOID',
              'Less_Than_HighSchool' = 'estimate')

#Merge
ma_hs_SelectCols <- ma_lt_hs[,c("Census_Tract_ID","Less_Than_HighSchool'
df_lt_hs<-merge(df_w_income,ma_hs_SelectCols, by = 'Census_Tract_ID')
head(df_lt_hs)

# Get number of people who completed either a Graduate or some other Pro
ma_grad <- get_acs(
  geography = "tract",
  variables = "B06009_006",
  state = "MA",
  year = 2020,
  geometry = TRUE
)

head(ma_grad)

#Rename
ma_grad <- ma_grad %>%
       rename('Census_Tract_ID' = 'GEOID',
              'Grad_or_Prof' = 'estimate')

#Merge to get final dataframe
ma_grad_SelectCols <- ma_grad[,c("Census_Tract_ID","Grad_or_Prof")]
df_grad<-merge(df_lt_hs,ma_grad_SelectCols, by = 'Census_Tract_ID')
head(df_grad)
```