Influential Factors for NYC Housing Prices in 2015

RAN JIN, XIAO WANG, SHIBO YU, ZHUOWEI HAN (GROUP 13)

1 Introduction:

Housing prices in NYC vary in a wide range in different five boroughs. So except the obvious factors such as the different functions that these boroughs are assigned, does there exist some other incredible factors that could potentially influence the housing prices, even though they seem to be unrelated to the real estate market at all? More specifically, for consumers, what factors are essential to consider while buying a house in NYC? We choose the numbers of death and injury caused by car collisions, the total GHG emissions and the total number of trees as factors to represent safety, pollution level, and vegetation coverages, repsectively in these five boroughs. By analyzing thoroughly these three datasets (including Total Number of Trees in 2015[1], Numbers of Deaths And Injuries Caused By Car Collision in 2015[2] adn Total Greenhouse Gas (GHG) Emission in 2013[3]), we are trying to find out if there exists some correlations between the housing price in every borough and these factors. And finally, we will construct 5 linear regression models using these three datasets and housing price in each borough in 2015[4] to complete our analysis.

Our questions are:

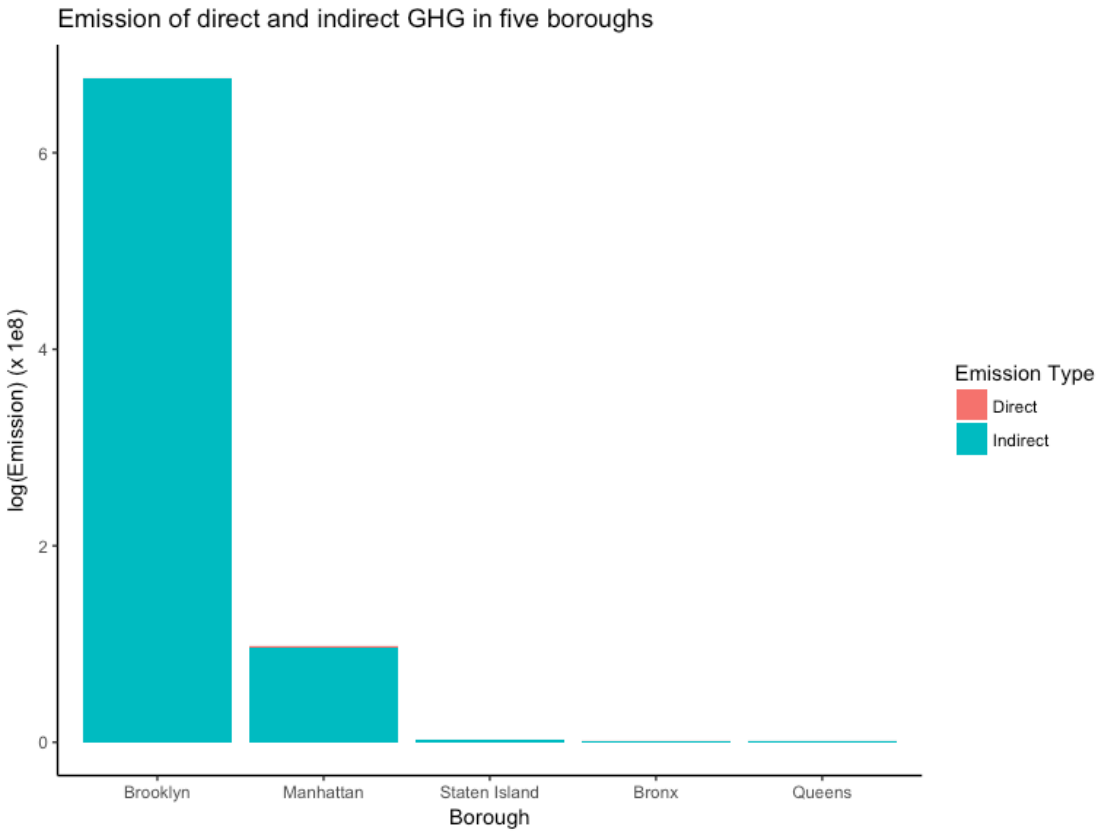
- What is the hypothesis we can derive from direct/indirect GHH emission in each borough?
- What is the hypothesis we can derive from pedetrians killed, cyclists killed, and motorists killed from car collision in each borough?
- What is the hypothesis we can derive from pedetrians injured, cyclists injured, and motorists injured from car collision in each borough?
- Whether or not that the housing price can be affected by the trees distribution of 5 boroughs in New York City. If the answer is YES, how would the relationship be? Is it linear relationship or can be transform into linear relationship?
- Can we fit good linear models of NYC housing prices in these 5 boroughs regressing on some or all of the variables within the datasets we display later on in the report? If so, how are the housing price in each borough linked to the factors? (Positive correlation or negative.)

And the followings are our analysis and interpretations.

2 The Database Table:

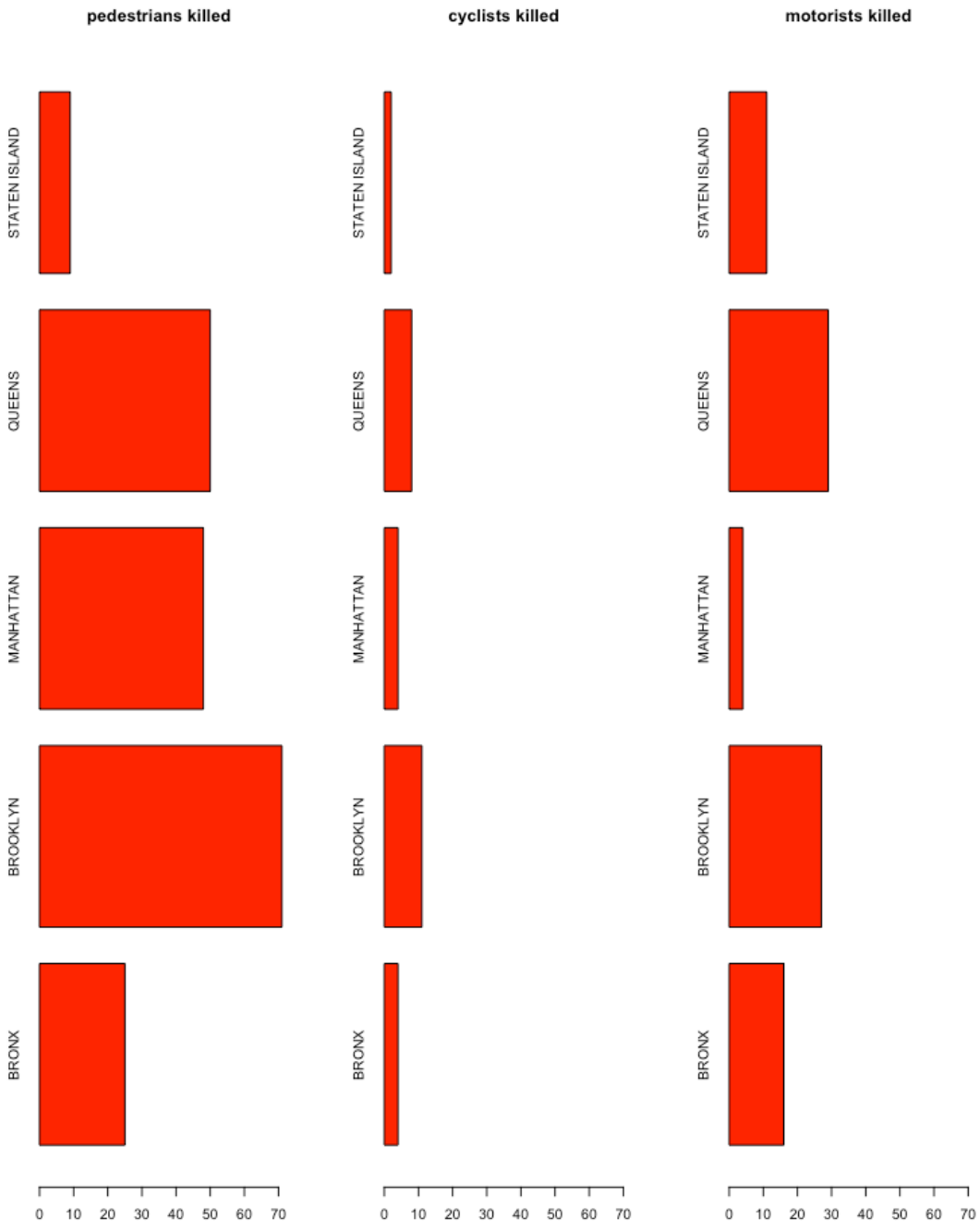
table	Description
nychousing	The information of houses that were sold in 2015 in New York
tree	the information of all the trees in New York in 2015
database	The information of car collision in New York in 2015
energy	The information about inspected GHG in New York in 2013

3 Analysis:

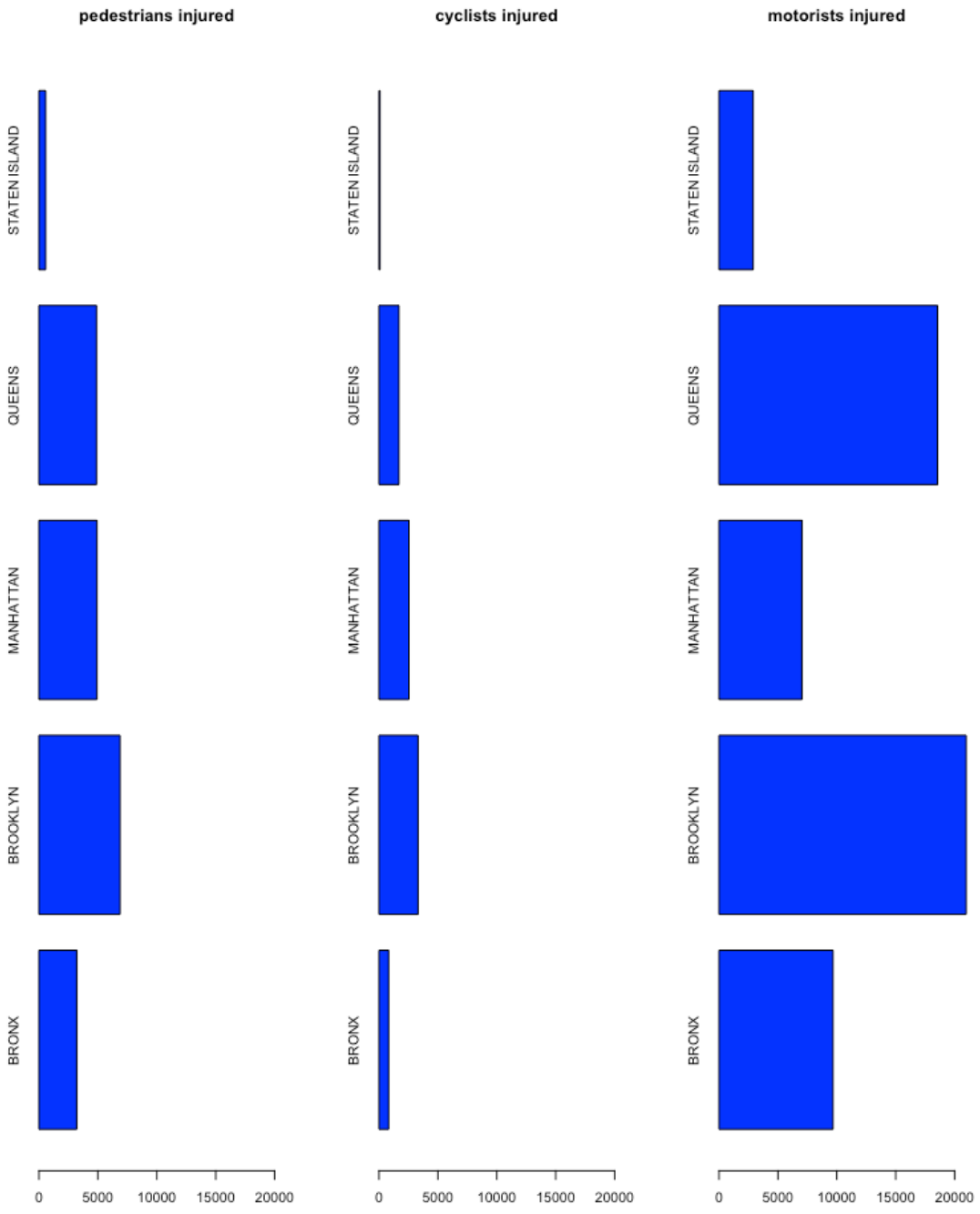


- The number of indirect emission in Broolyn and Manhattan seems way much bigger than that of direct emssion.

- From the graph above, we may guess the housing price in Broolyn and Manhattan may be lower than other places due to the GHG.

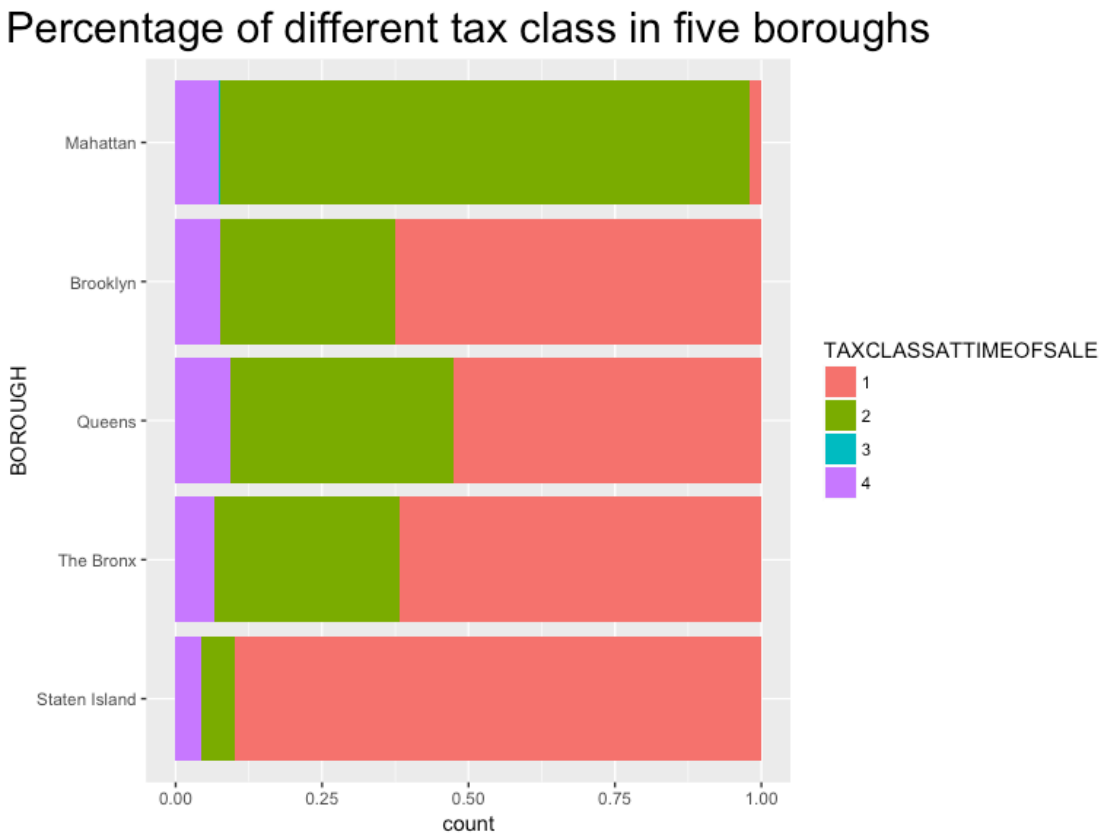


- The 3 graphs above represent the numbers of pedestrians killed, cyclists killed, motorists killed respectively from car collisions for each borough.
- Pedestrians killed have a significantly larger number than the other two.
- Brooklyn and Queens appear to have the relatively higher numbers of incidents which might drive the housing prices in those two areas down.

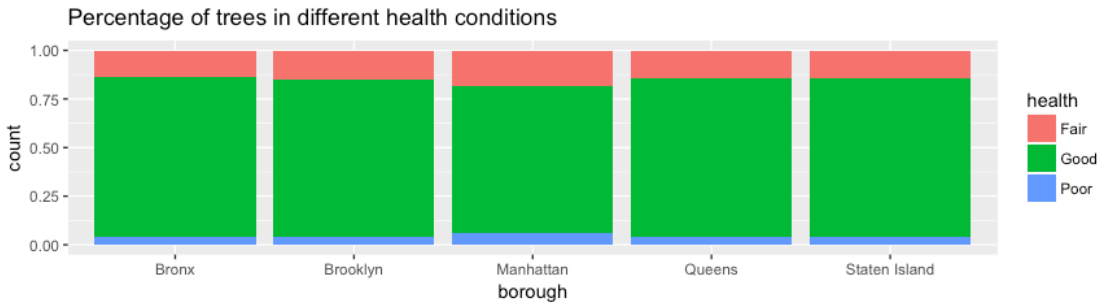


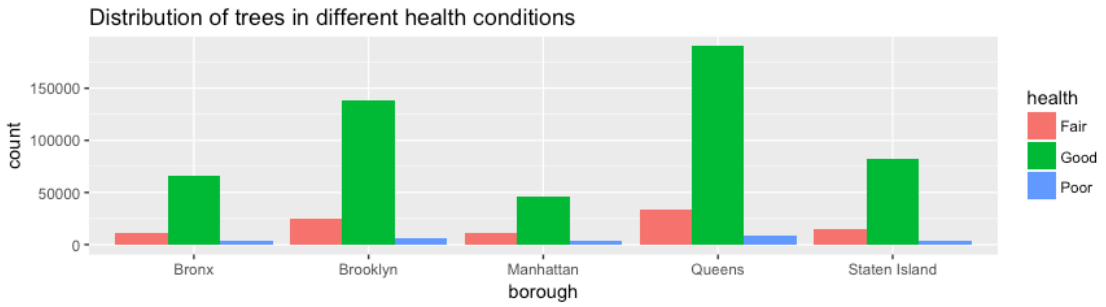
- The 3 graphs above represent the numbers of pedestrians injured, cyclists injured, motorists injured respectively from car collisions for each borough.
- motorists injured have a significantly larger number than the other two.
- Brooklyn and Queens still appear to have the relatively higher numbers of incidents which might drive the housing prices in those two areas down.

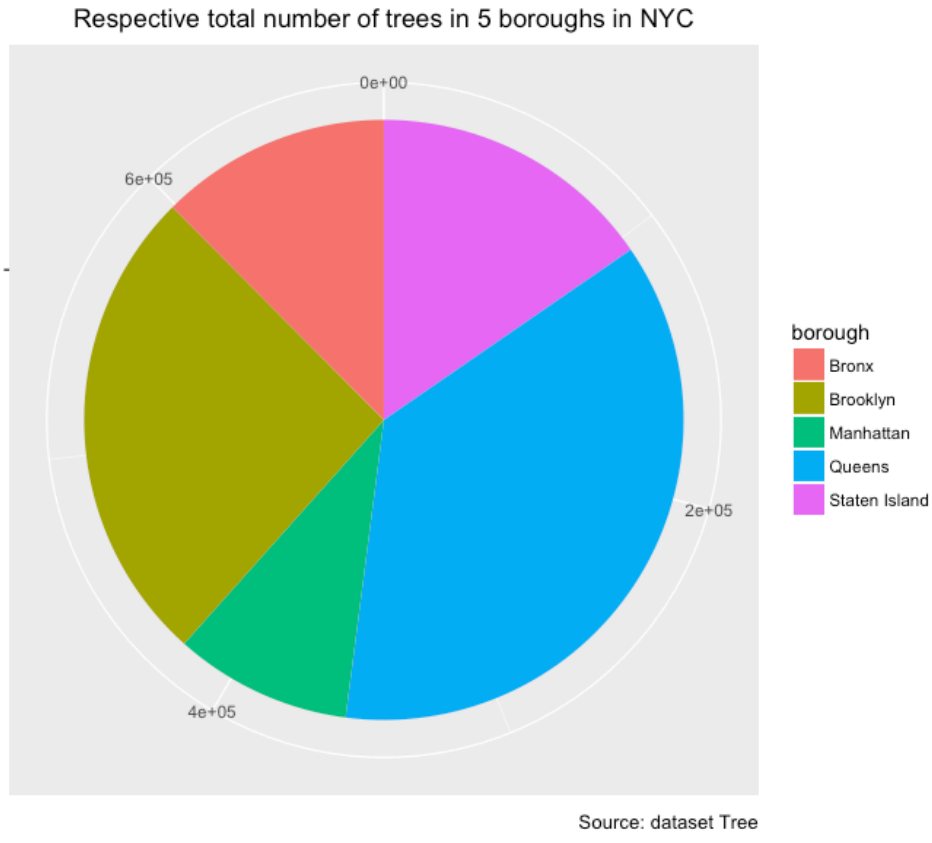
Variable	Description
borough	county-level administrative divisions
saleprice	transaction price
tax class of time of sale	the housing tax level at the time of sale
building class category	the building category at the time of construction



- The numbers 1,2,3,4 in the above graph represent the different tax level 19.991%, 12.892%, 10.934% and 10.574. There is no doubt that the different level of tax could show the different level of houses.
- The percentage of the houses that should be paid the lowest tax in the five boroughs are similar.
- The highest tax is paid most in the Staten Island.
- In Mahattan, the ratio of highest tax is incredible low.
- The graph gives us some hint that the house size in Mahattan may be smaller and in Staten Island, people are wealthy enough to buy some huge houses.
- So the most busiest place may not have the highest housing price.







- The above three plots can help us find out and easily compare the distributions of trees in the five bouroughs of New York City, which may give us some clues on how the trees distribution would affect the housing price here.

- As seen from the 1st plot, the percentages of trees in different health conditions are almost the same for the five bouroughs.

- The 2nd plot conveys very different information from the 1st plot because the number distribution of trees in different health conditions are quite distinct, except for "Poor" trees. Surprisingly but understandably, the most developed Manhanttan has least "Fair" trees and "Good" trees, which may be substituted with countless skyscapers.

- The 3rd pie-chart further shows the number of trees without differentiating the health conditions. The colorful areaes shows the proportion of respect total number of trees in the five bouroughs, with the specific numbers accumulated in a clockwise direction.

Variable	Description
zipcode	Five-digit zipcode in New York City
sp	Average housing saleprice in each zipcode
pk	Number of people killed by accidents in each zipcode
pj	Number of people injured by accidents in each zipcode
numtree	Number of trees located in each zipcode
borough[1]	Name of borough in New York City
borocode[2]	Code for borough in which tree point is located
emission	Total emission of GHG[3]

Explanation to the table

- borough: Manhattan, Bronx, Brooklyn, Queens, Staten Island.

- borocode: 1(Manhattan), 2(Bronx), 3(Brooklyn), 4(Queens), 5(Staten Island).

- GHG: greenhouse gases including CO_2, CH_4, and N_2O. For further information, please visit https://www.archibus.com/ai/abizfiles/v21.1_help/archibus_help/Subsystems/webc/Content/gloss/carbon_footprint/carbon_diox_equiv_def.htm.

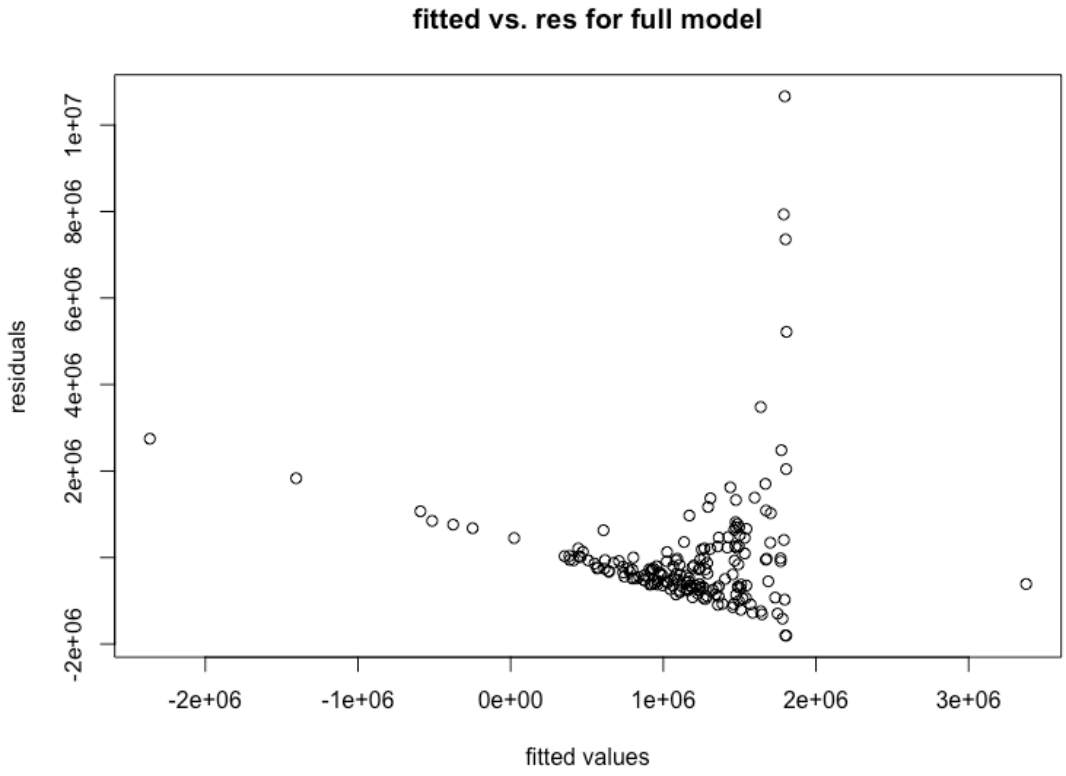
Interpretation of the tables

- The above table shows the variables extracted from the original four raw dataset by "INNER JOIN" all of them via their common variable "zipcode".

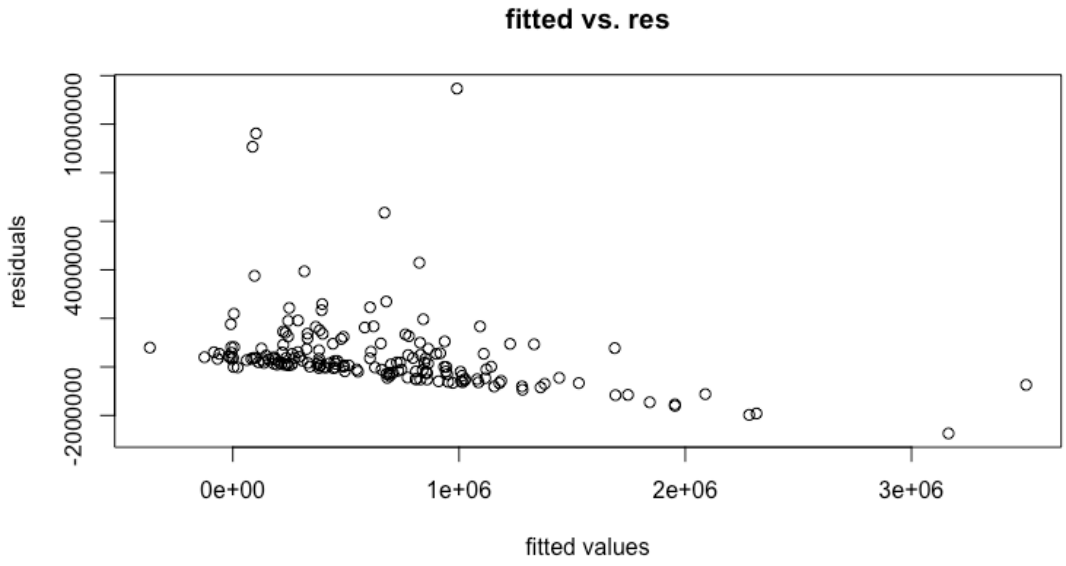
- These variables would be used for our FINAL analysis on how the housing price can be influenced by other variables.

- Given that the same variable may have varying degrees of impact on housing price in different boroughs, the above table is further divided into five sub-tables according to the five boroughs in New York.

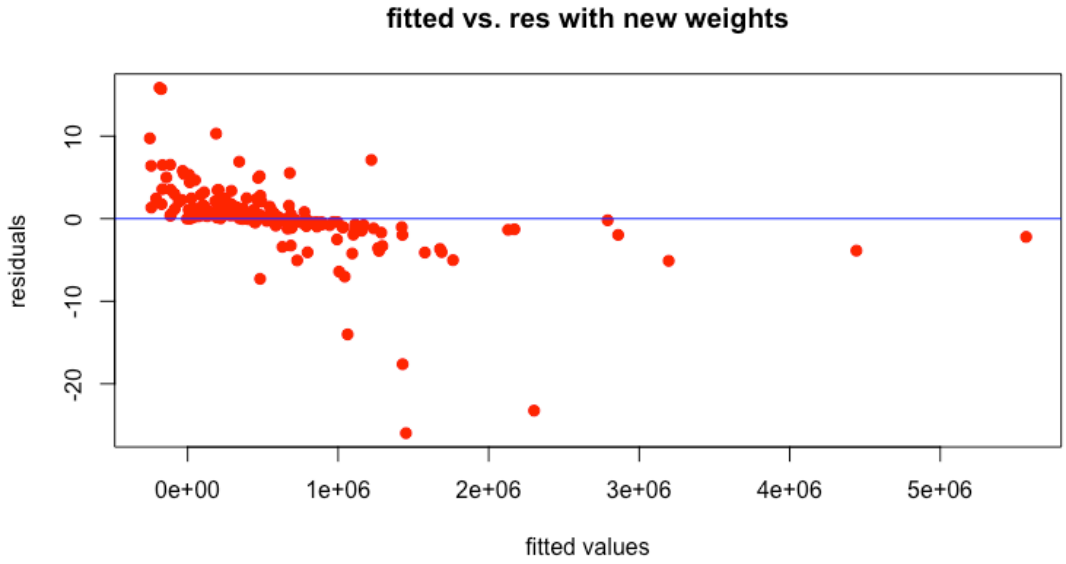
- Then, five independent and systematic regression analyses are performed with regard to five boroughs.



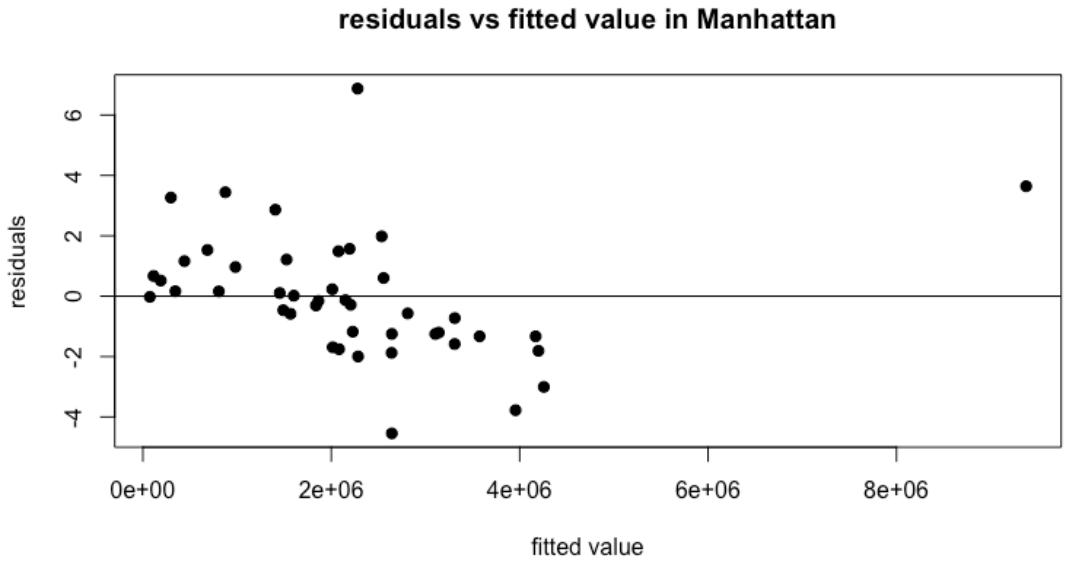
- we first fitted the full model using all four variables including total people killed, total people injured, number of trees, and the total emissions of GHG as the regressors and we found that R squared is relatively low which is 0.1367186.
- the full model also has a graph of fitted values vs. residuals of a non-constant looking variances. (heteroskedasticity)
- Since the housing price is never 0, the beta 0 intercept would be pointless.
- Thus, we fitted the new full model using all four variables without the intercept and got a lower adjusted R squared 0.1818077 than the previous model with the intercept 0.1168731.
- 5.767977710^{4}, 1520.0768756, -38.7918515, 0.0035028 are coefficients for pk,pj,numtree, and emission.
- 0.1818077, 0.1779964, 0.1787586, 0.1296181 are adjusted R squared from full model, model with three variables, model with two variables, and model with one variable.



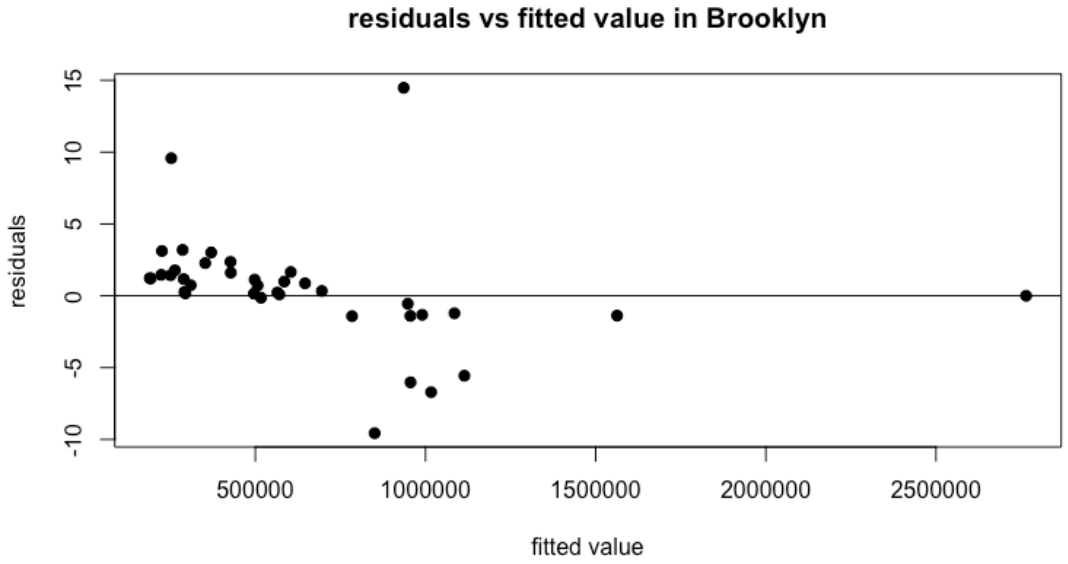
- Since the fitted values vs. residuals plot does not look spreaded here. We will use some different weights on the regressors.



- Now the fitted values vs. residuals plot looks much better than the previous one with some different weights we have applied since the points are much more spreaded. Thus, we will use the same weighting method and non-intercept linear model for the each of the regression analysis of the boroughs.

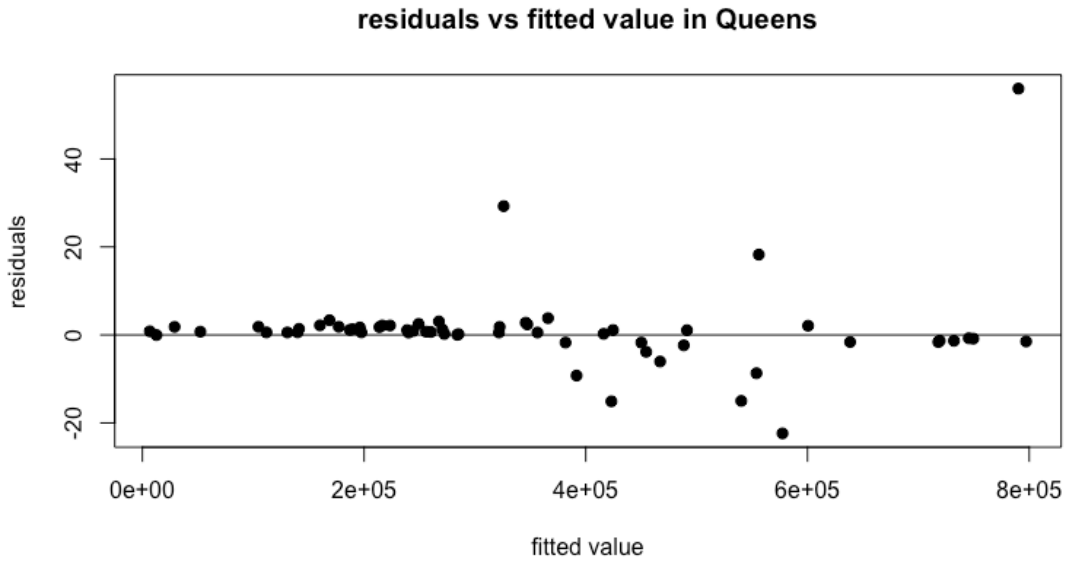


- -1.12838910^{5}, 2327.2723975, 763.9135198, 0.2452886 are the coefficients for pk, pj, numtree, and emission respectively.
- 0.9499079 is the R squared for the new full model in Manhattan.

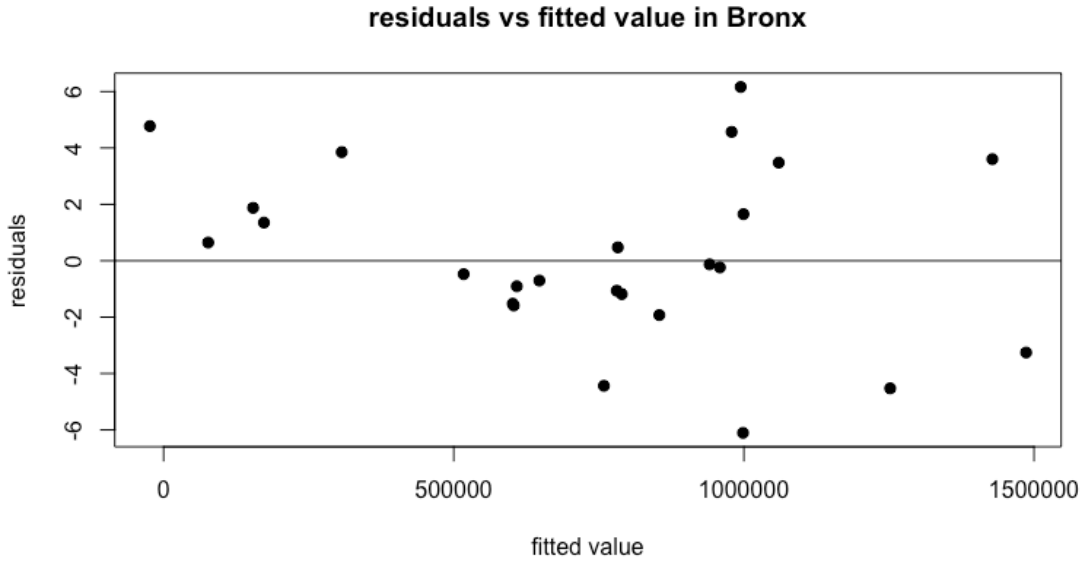


- 4.511129510^{4}, -535.5723458, 184.3339288, 0.0032634 are the coefficients for pk, pj, numtree, and emission respectively.

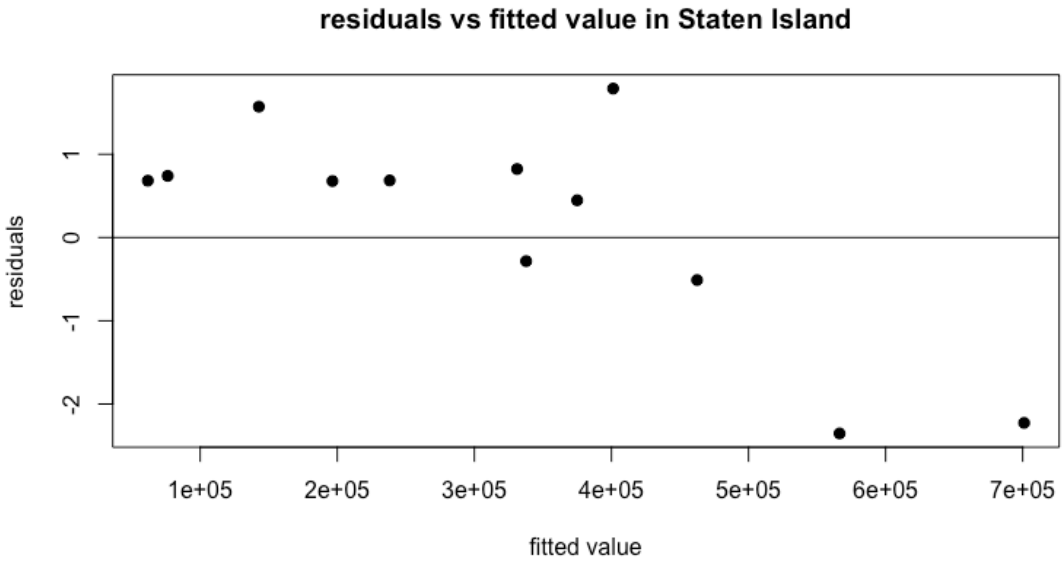
- 0.9999925 is the R squared for the new full model in Brooklyn.



- -2.642873710^{4}, 642.0372387, 20.7993276, 2.4679097 are the coefficients for pk, pj, numtree, and emission respectively.
- 0.8016953 is the R squared for the new full model in Queens.

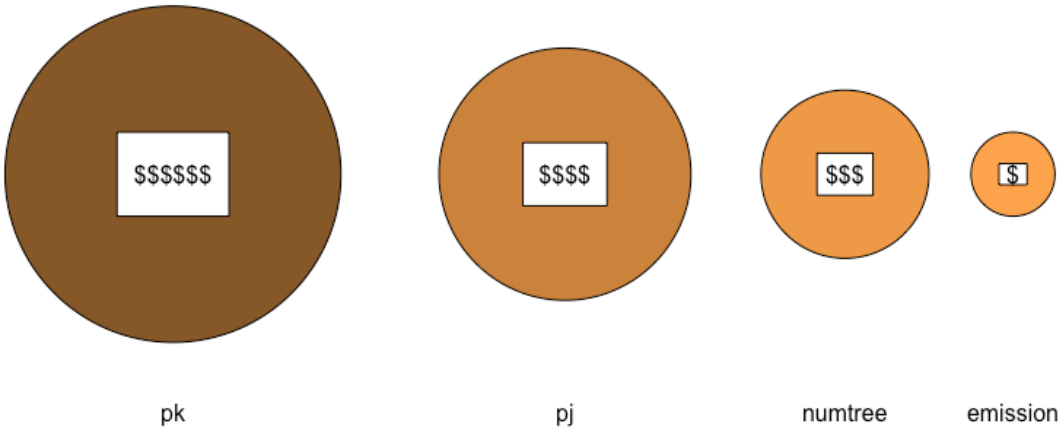


- 8.862730110^{4}, 1909.4003313, -150.5882879, 2.4159853 are the coefficients for pk, pj, numtree, and emission respectively.
- 0.9463711 is the R squared for the new full model in Bronx.



- -2092.577046, 455.8168327, 22.7365022, 0.3930806 are the coefficients for pk, pj, numtree, and emission respectively.

- 0.973902 is the R squared for the new full model in Staten Island.



- The four coins of different sizes and colors represent the different influence of the four variables for housing prices.

4 Conclusion:

Housing prices in NYC vary in a wide range. We choose the numbers of death and injury casued by car collisons, the total GMG emissions and the condition of trees to represent safety, pollution condition and vegetation coverages separately in these five boroughs. By analyzing thoroughly these datasets (including Total Number of Trees in 2015, Numbers of Deaths And Injuries Caused By Car Collision in 2015 and Total Greehhouse Gas (GHG) Emission in 2013). After constructing linear regression models using these three datasets and Housing Price In 2015 to give quantative interpretations, we finally found some models after numerous modifications.However, even though the R squared value is high, the constancy of errors is still a violation of the assumption of Linear Regression Model. In addition, the coefficients we found differ tremendously which is somewhat abnormal. Thus, we will conclude that the the linear regression model might not be a good model to use for our problem and these four datasets have no significant influence on the housing price of NYC compared with the other outstanding factors.

5 References:

[1] Department of Parks and Recreation (DPR). (2017). 2015 Street Tree Census - Tree Data -- .[Data file]. Available from NYC OpenData Web site https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh

[2] NYPD. (2017). Vehicle Collisions in NYC, 2015-Present [Data file]. Available from Kaggle Web site https://www.kaggle.com/nypd/vehicle-collisions

[3] Mayor's Office of Long Term Planning and Sustainability (OLTPS). (2017). Energy and Water Data Disclosure for Local Law 84 (2013)[Data file]. Available from NYC OpenData Web site https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/rgfe-8y2z

[4] Department of Finance (DOF).(2017). Annualized Rolling Sales Update. [Data file]. Available from NYC OpenData Web site https://data.cityofnewyork.us/Housing-Development/Annualized-Rolling-Sales-Update/uzf5-f8n2