Understanding the Internal and External Factors that Effect Housing Sale Price Change in New York Borough: Staten Island

Eren Con
Center for Urban Science+Progress, New York University
e-mail: ec2986@nyu.edu

Introduction

Housing prices changes depending on various factors. These factors may be internal factors that are properties/situation of building itself such as floor area; external factors that are features effective on prices such as distance to landfill areas. This study aims to determine the effect of the internal and external factors on housing prices in Staten Island based on the time interval between 2003 and 2008.

For this study, Staten Island is selected as a borough that is used for analysis. In addition to the internal factors: floor area and gross area; Staten Island is selected for understanding the relation between housing prices and distance to solid waste landfill area.

Data Description, Cleaning and Processing

The housing sale data is obtained from the Enigma data portal which provides and combine government official data from the different open data portals. The dataset includes 21 different attributes. The floor area, gross area and sale year columns are used for the determination of the internal factor. For the external factors, distance is calculated from the address column in the housing sale data. For doing this, The Open Street Map API is the used for the reverse geocoding to obtain latitude and longitude information of the each housing units.

In the data cleaning process, unnecessary fields and inappropriate data values which are '0' sale price values are removed from the dataset. 2003 and 2008 time interval is used since data is available from 2003 and lifetime of the landfill area ended in 2008. In addition to that, after the distance calculations, maximum distance is defined as 10 km which is the maximum distance from the landfill area in Staten Island, and data is also cleaned depending on this value.

In the analysis processes, python programming language and Jupyter Notebook is selected as development environment. numpy, pandas, pygeocoder, statsmodels python packages/modules are used for data management and data analysis.

Methodology, Data Analysis and Interpretation of the Results

In this study; linear regression analysis (OLS Method) is used for understanding the relations between both internal and external factors that create two model on housing prices. First model examine the relation between the distance and sale price change. Our prior belief for this model is accordance with the negative effect of the landfill area on house prices and accepted as a null hypothesis. Similarly, second model examine the relation between sale prices change and the total and building area of the houses. Our prior belief for the second model is accordance with the positive effect of area increase in sales prices.

As a result of the first model, R^2 value is calculated as a value of 0.007. This value is very low and it indicates that changes in distance can explain the 0.7% of the changes in the sale prices. Coefficient value reflects that 0.01 mile = 16 meters change in distance effects the sale prices at a very low value of 3.94 dollars. In a 95% confidence interval this change between the 1.526 and 786.60 dollars. T-value in the analysis is very close to the 1.96 critical value in %95 confidence interval. This means that there is not a strong evidence to explain the change in price for this model. In addition to that, p value is 0.049 and this means that null hypothesis (relation between the distance and price) can be rejected.

Figure 1. First Model Analysis Results and Distance-Sale Price Graphs

In the second model, R^2 value is calculated as a value of 0.926. This means that changes in land and gross are explain the changes in the sale prices as 92.6%. Coefficient values are calculated as 24.7072 and 86.3838 respectively. This means that 1 unit (square meter) change in increase the housing price 24.7072 dollars for land area and 86.3838 for gross area. T-values are bigger than

remember that the null hypothesis is the opposite of your theory (because you want to try and falsify your theory! so here the null hypothesis is that the landfill has NO EFFECT on the price, which is falsified.

you really really need to clean your data more here: the slope of your regression line is entirely driven by the datapoint at y = 4000000. i bet you $$ to peanuts that if you remove it you get a very different result and the importance of distance to landfield will decrease.

below you have a whole bunch of points at 0 sq feet with assigned prices. if you remove them your result will change (they have influence because they are at the edge of your distribution. particularly, i bet you $$ to peanuts that this will steepen your relationship a lot

the critical value of 1.96 and the p values are 0. This means that null hypothesis (relation between the land, gross areas and price) can be accepted.


```
                          OLS Regression Results
==============================================================================
Dep. Variable:              saleprice   R-squared:                       0.929
Model:                            OLS   Adj. R-squared:                  0.929
Method:                 Least Squares   F-statistic:                     3535.
Date:                Sat, 12 Dec 2015   Prob (F-statistic):           1.06e-311
Time:                        17:46:05   Log-Likelihood:                -7373.3
No. Observations:                 545   AIC:                         1.475e+04
Df Residuals:                     542   BIC:                         1.477e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     1.777e+05   1.04e+04     17.133      0.000    1.57e+05   1.98e+05
land_sq_feet    24.7072      2.131     11.594      0.000      20.521     28.893
gross_sq_feet   86.3838      1.513     57.077      0.000      83.411     89.357
==============================================================================
Omnibus:                      779.579   Durbin-Watson:                   1.886
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           324147.612
Skew:                           7.237   Prob(JB):                         0.00
Kurtosis:                     121.595   Cond. No.                     1.12e+04
==============================================================================
```
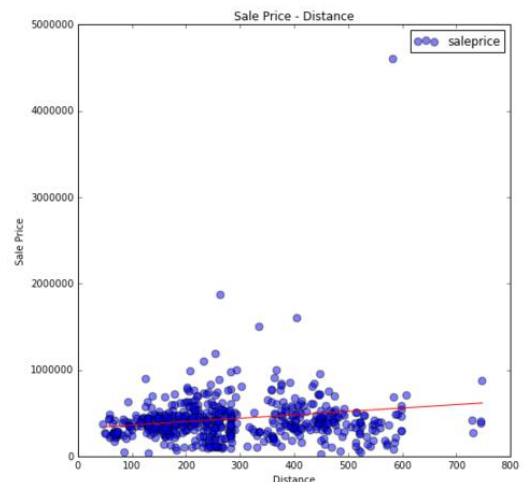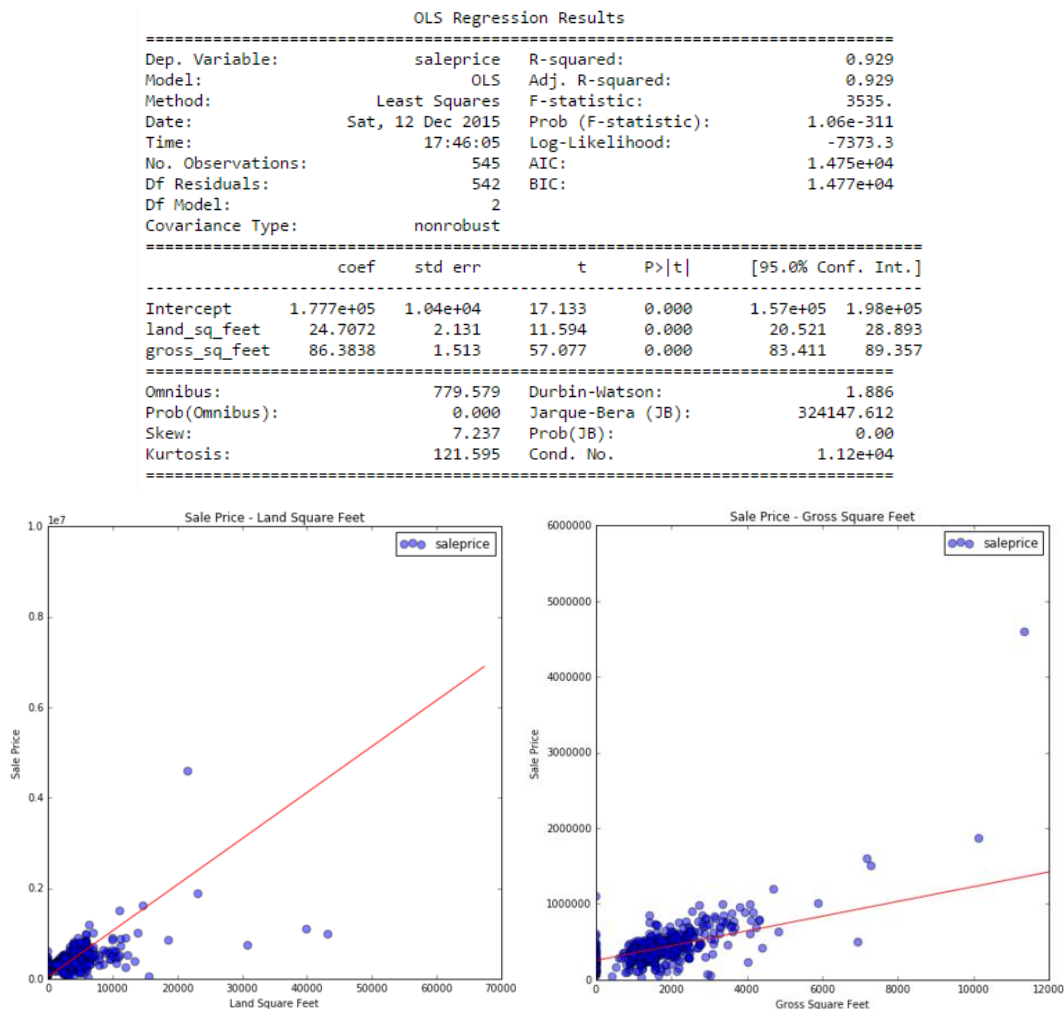

Figure 2. Second Model Analysis Results and Land & Gross Area-Sale Price Graphs

To sum up, this study demonstrates the relation between the internal and external factors and their effects on the housing sales prices in a given context. It is understood that there is not a strong evidence that the proximity the landfill area in the Staten Island effect the housing prices in negatively. On the other hand, it is concluded that increase in housing land and gross area effect the housing prices positively.

Future Work

Future study will cover the before and after use of the landfill area and their effect on the housing sell prices. In addition to that, housing prices will be analyzed depending on effect of the each land use on price. These factors will be community parks, bus stops, subway stations or central business districts and this approach will be implemented other New York Borough's.

<u>Bibliography</u>

Housing Sale Prices Data for Staten Island, last accessed 2015. https://app.enigma.io/table/us. states.ny.cities.nyc.property.annualized-sales?row=0&col=0&page=1&documentView=false

Reverse Geocoding Code Resources, last accessed December 2015. http://geocoder. readthedocs.org/

Software Codes for this analysis, https://github.com/fedhere/PUI2015_EC