



Realtor.com Web Scraping & A Simple Linear Model For Sale Price Prediction

Jing Wang
2017-08-02



Introduction & Motivation

1. Real estate play big role in economy
 - a. Residential real estate
 - b. Commercial real estate
2. Real estate has huge impact on employment, income and consumer spending
3. Real estate provides basic needs and tools for investment
4. A good model with high accuracy on sale price prediction has great value to guide such investment, both short term and long term.
5. Where, When and How Much?

www.realtor.com/real-estate-and-homes-detail/48U-Madison-Park-Gdns-Port-Washington-NY-11050-M37246-46222

www.rea.com/real-estate/properties/6074738 Midtown Dr. Massapequa, NY 11758 M38326-93734

Dataset & Geographical info

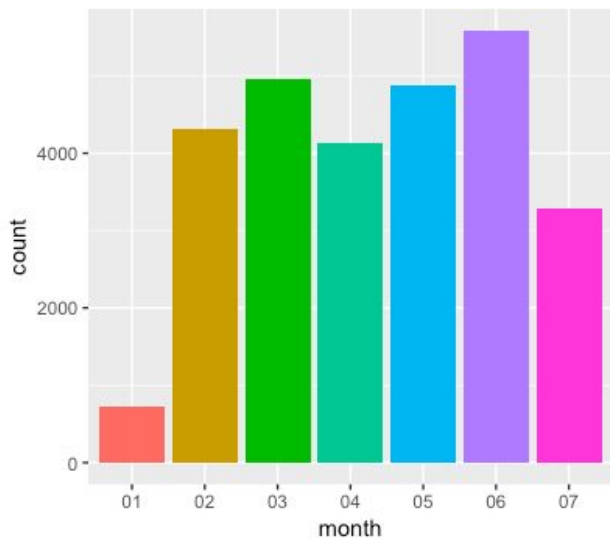
- Six counties in New York City and Long Island Area
- Key features:
 - Number of Bedroom
 - Number of Bathroom
 - Total floor size
 - Lot size
 - City
 - County
 - Zip code



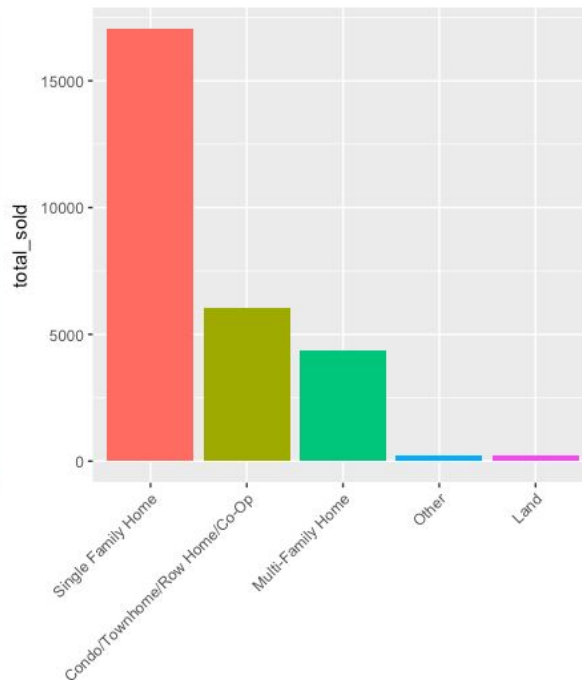


Basic EDA: Total Number of Sold Property 2017.1 ~ 2017.7

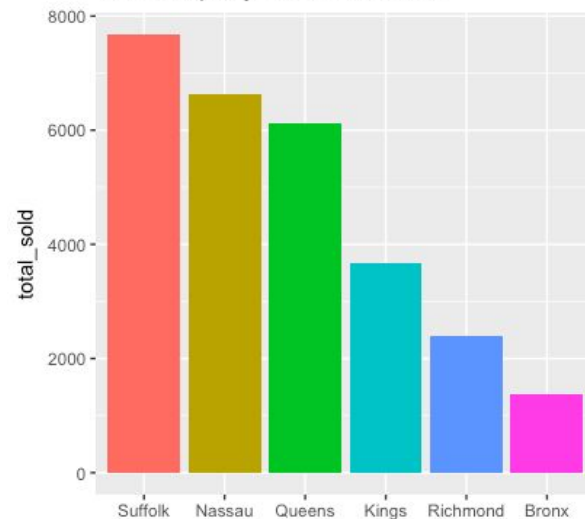
Total Sold Property Number



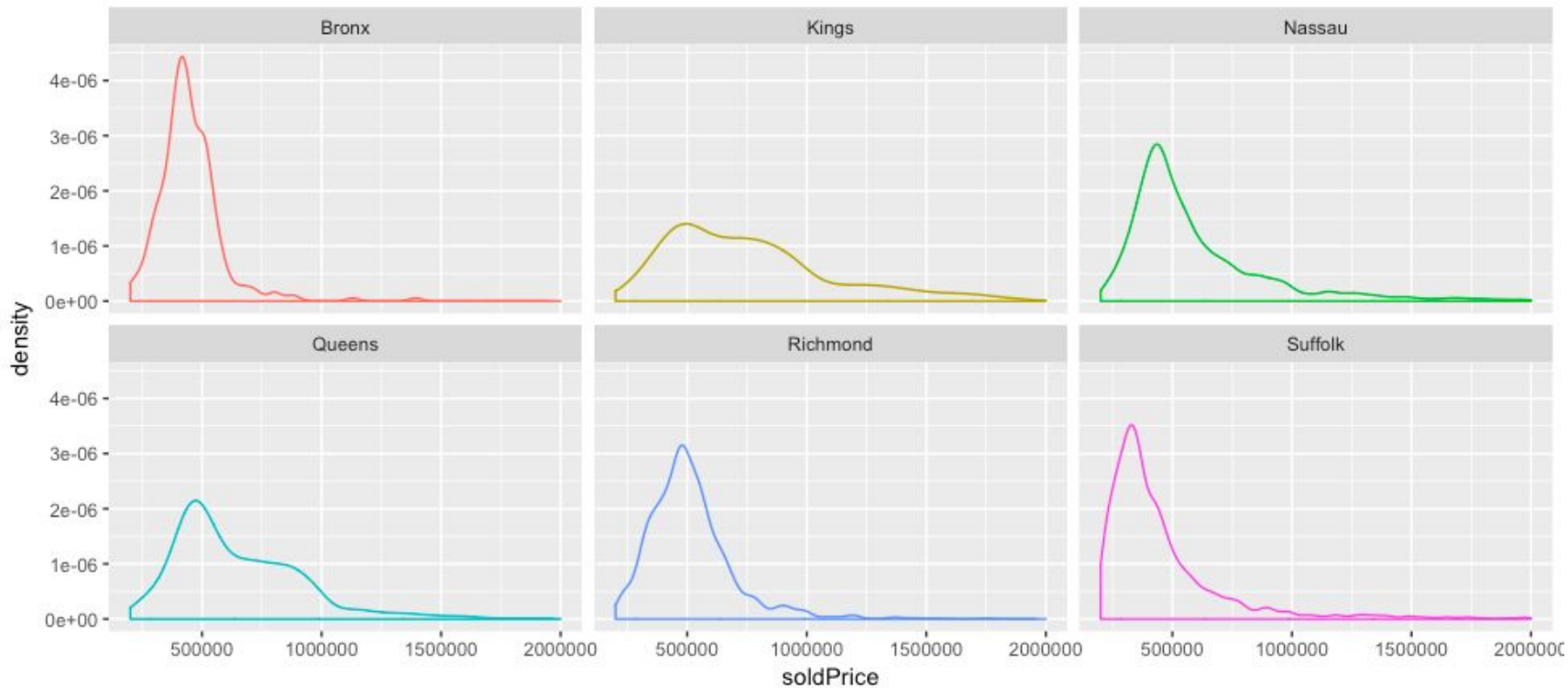
Total Sold Property Number



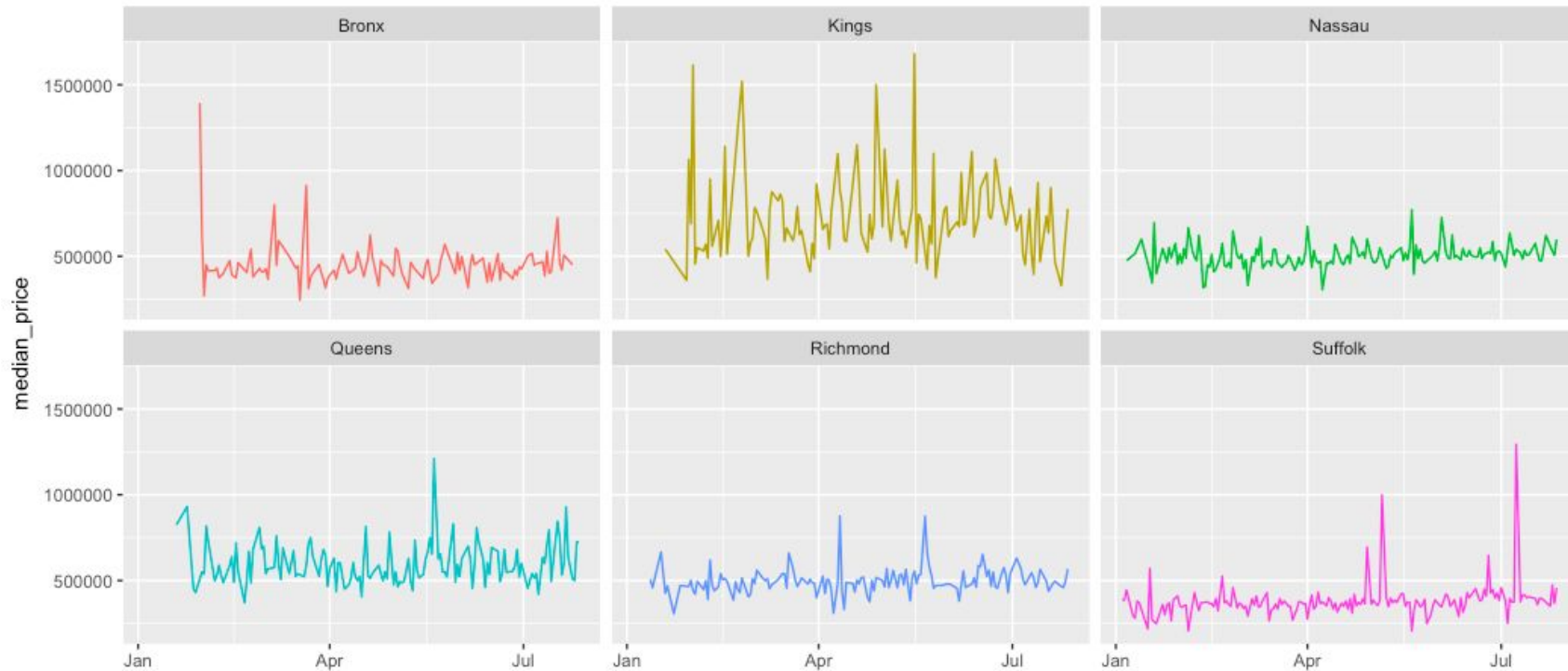
Total Property Sold in Counties



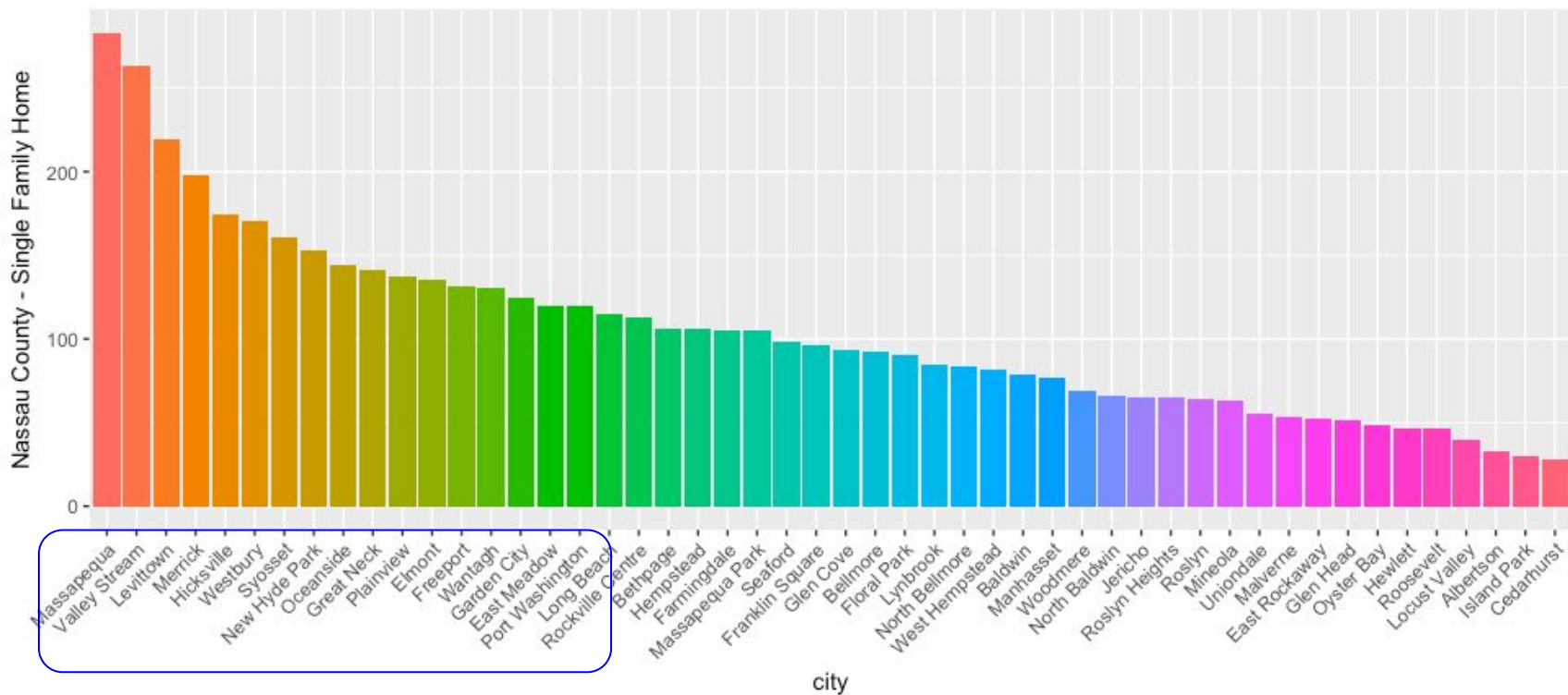
Basic EDA: Distribution of SFH Median Sale Price



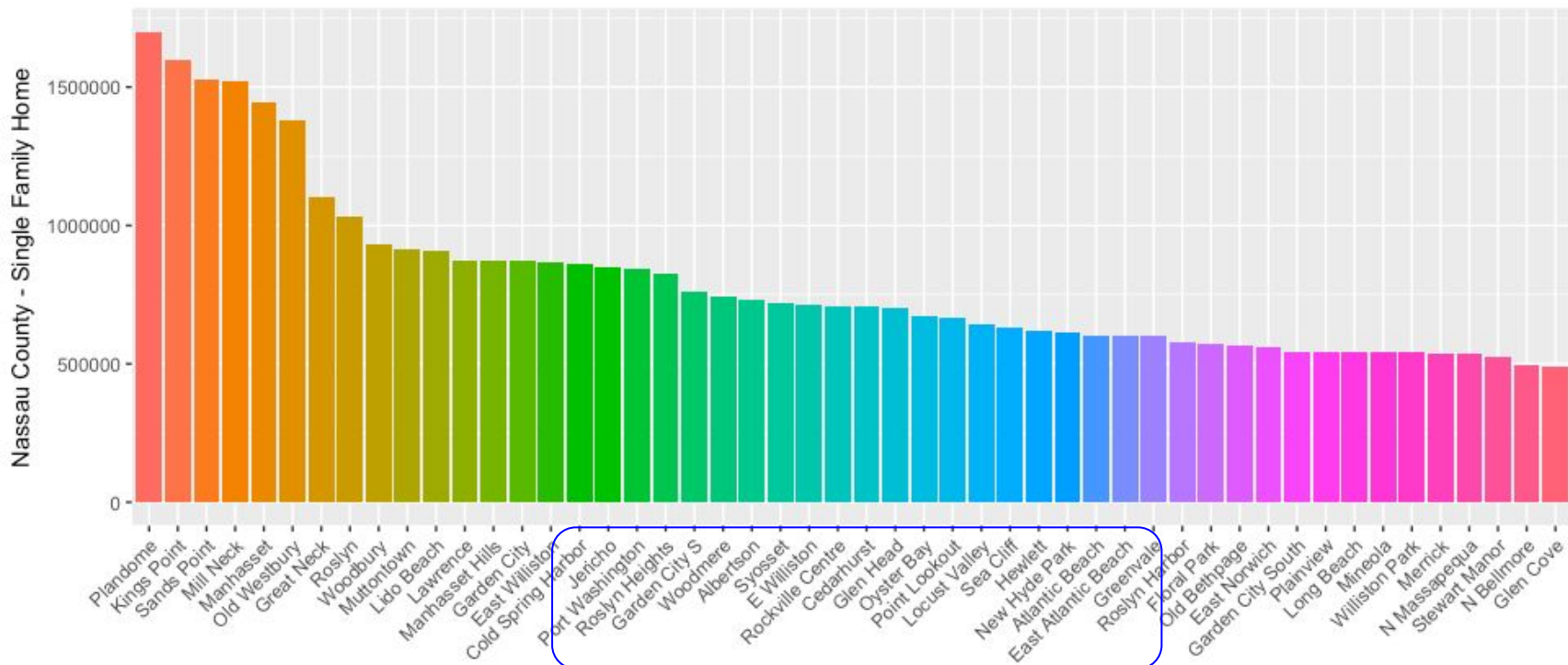
Visualization: Volatility of SFH Median Sale Price



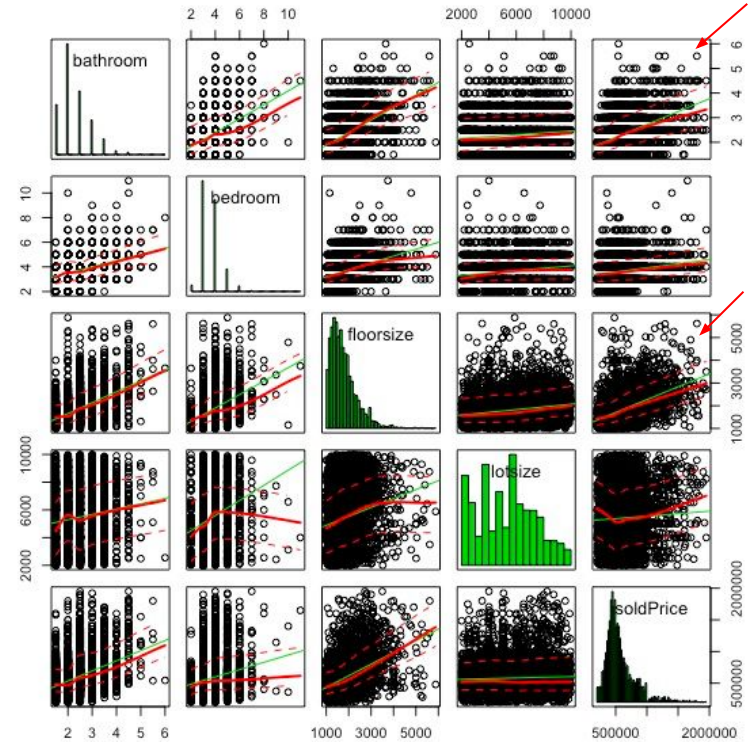
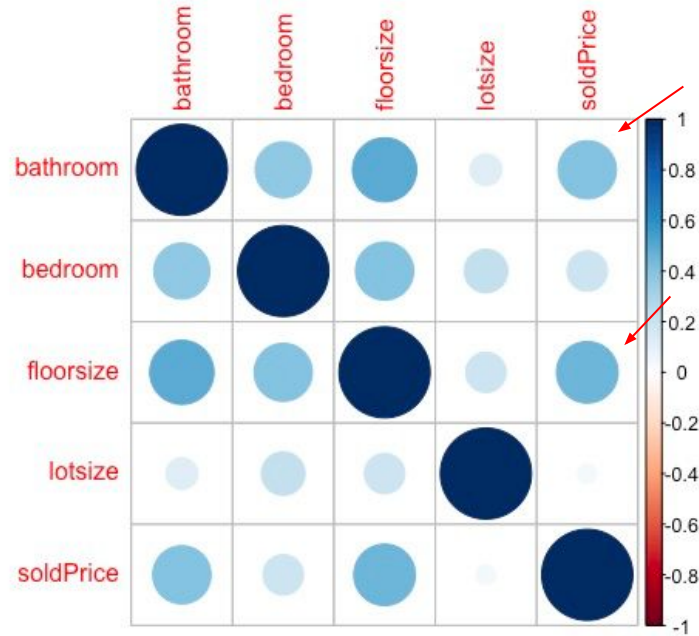
Basic EDA: Sold SFH Number Ranking in Nassau



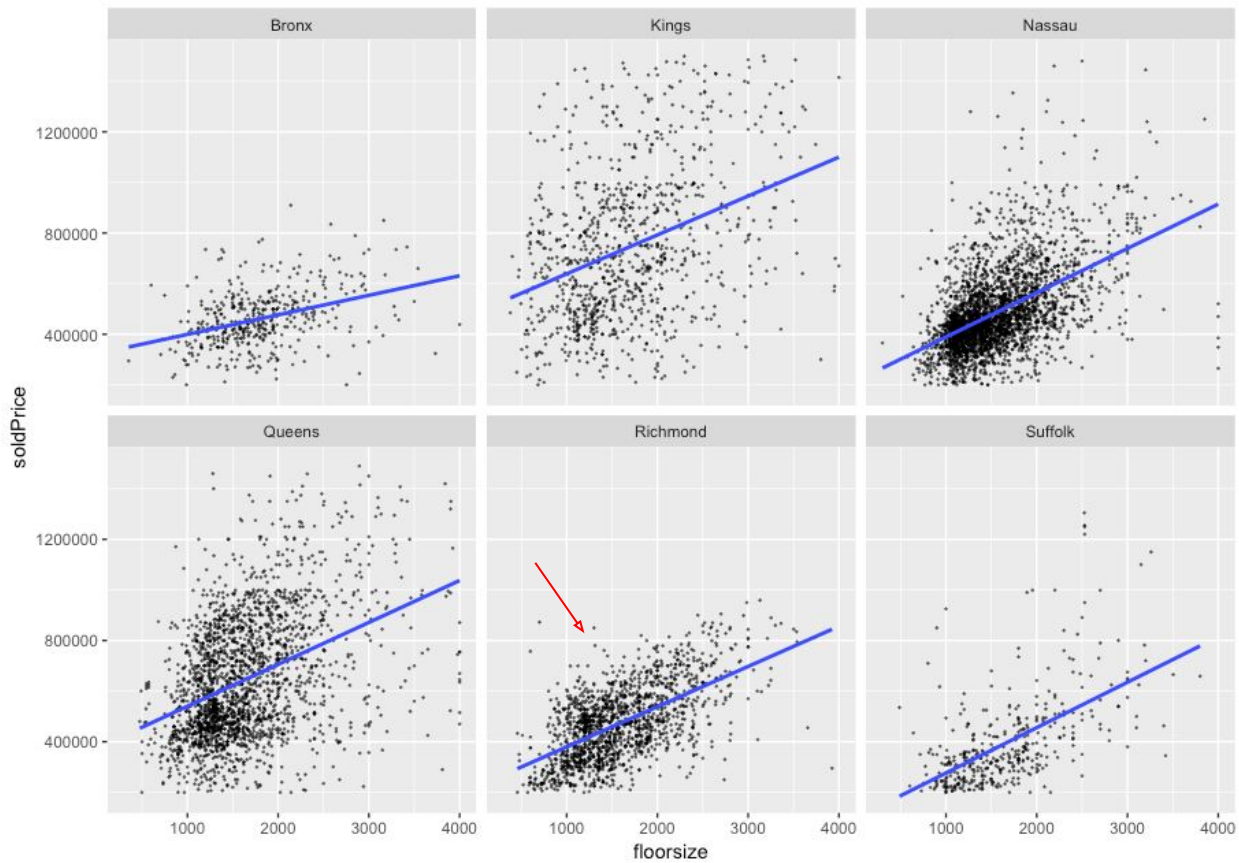
Basic EDA: SFH Median Price Ranking in Nassau



Correlation Matrix

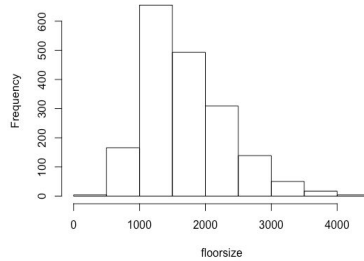


Scatter Plot: floor size ~ sold price

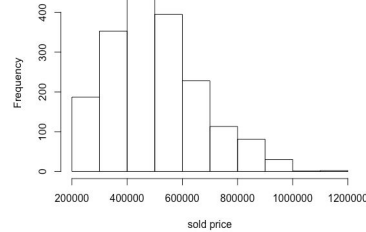


A Simple Linear Regression Model (Richmond County)

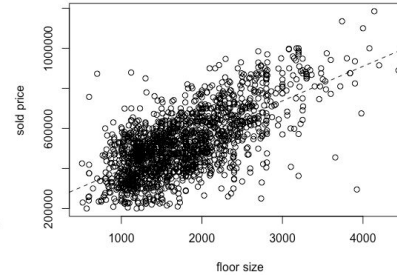
Histogram of floorsize



Histogram of sold price



Scatterplot of Realtor dataset, Richmond County



$$Y = 173 * X + 192000$$

Call:
lm(formula = soldPrice ~ floorsize, data = df2)

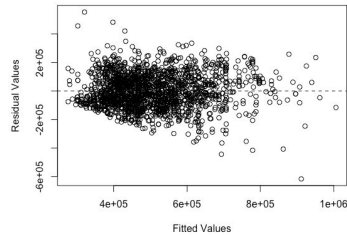
Residuals:
Min 1Q Median 3Q Max
-616611 -83303 -8800 81689 552702

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.919e+05 8.038e+03 23.88 <2e-16 ***
floorsize 1.834e+02 4.401e+00 41.67 <2e-16 ***

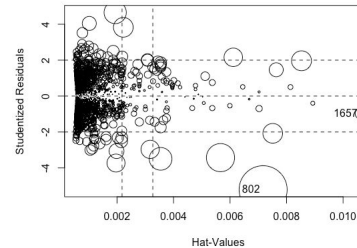
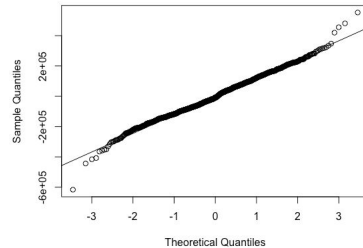
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 119600 on 1835 degrees of freedom
Multiple R-squared: 0.4862, Adjusted R-squared: 0.4859
F-statistic: 1736 on 1 and 1835 DF, p-value: < 2.2e-16

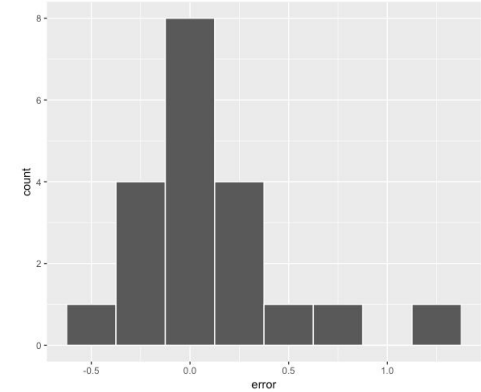
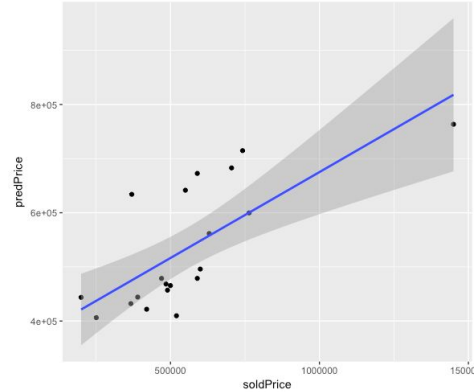
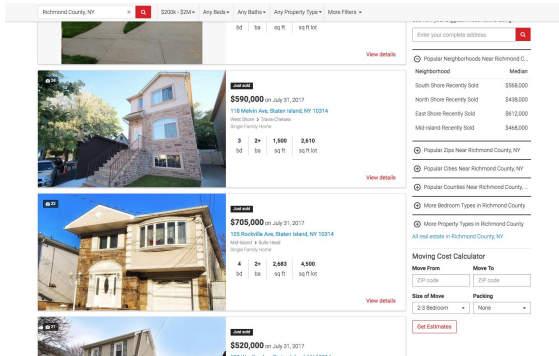
Residual Plot for Cars Dataset



Normal Q-Q Plot



Model validation & Conclusion



- Most recent property closing data from last three days (7.28 -7.31), 20 sold properties
- Floor size has significant impact to the sale price
- Need more data to build multiple linear regression model
- Need more time to search better deal

Shiny App, coming soon

Long Island Realtor Estate Market Explorer 2017

Gene | O | in | f

County

Nassau

Property Type

Single Family Home

Price Range



Bedroom



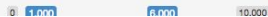
Bathroom



Floor Size



Lot Size



Sold Property Count

Median Sale Price Distribution

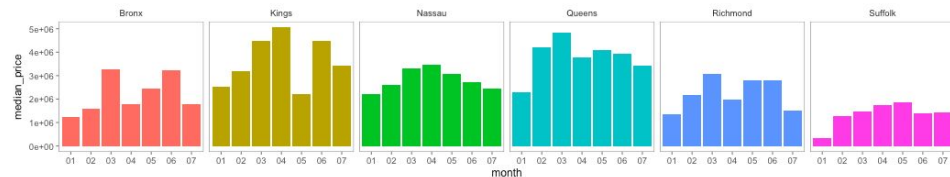
Median Sale Price Volatility

Linear Regression

Links

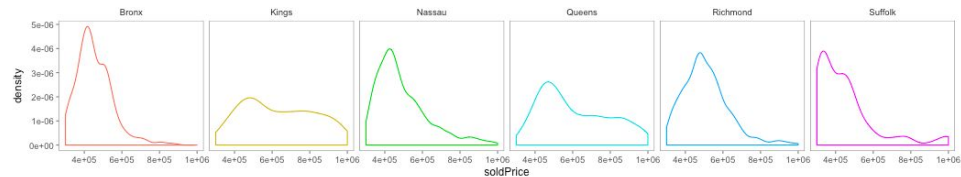
Median Sale Price Time Line

This section visualizes median sale price change from beginning of 2017.



Median Sale Price Density Plot

This section visualizes median sale price distribution by county.



Data



Tabular searchable data display similar to that found in the original source #

You can download the data with the download button above.