DRAFT

GENERAL ASSEMBLY: DATA SCIENCE

Final Project: Predicting Real Estate Values by proximity to transit

By Simon Mettler

August 2016

Context: Real estate market in NYC is crazy

Context: Transit has an effect on real estate- right?

BROKERS WEEKLY

New subway line already impacting home prices along Second Avenue

BY HOLLY DUTTON • MAY 13, 2016

The Second Avenue Subway line, long derided as fantasy by New Yorkers who had seen plans for construction stall and linger for decades, finally has an end date in sight — and it's poised to shake things up for the Upper East Side.

Areas of the neighborhood became affordable again in the past few years, as more and more people

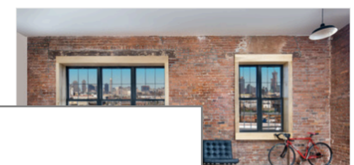
$300M Nov

WILLIAMSBURG, GREENPOINT & BUSHWICK Real Estate Transportation

How an L Train Closure Might Hurt Brooklyn Real Estate

By Amy Zimmer and Gwynne Hogan | January 28, 2016 7:31am

BROOKLYN — If the L train's tunnel connecting Manhattan and Brooklyn shuts down for a prolonged period of time — as is



Get our daily Williamsburg, Greenpoint & Bushwick news and alerts!

Email Address | Zip | Subscribe

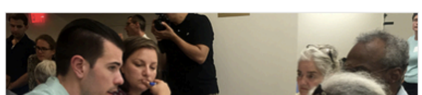
By clicking subscribe, I agree to be bound by the

COBBLE HILL, CARROLL GARDENS & RED HOOK Transportation

Brooklyn Heights Residents Like BQX, but Fear Property Tax Hike Along Route

By Alexandra Leon | June 21, 2016 3:44pm
@alexandraaleon

BROOKLYN HEIGHTS — The city's streetcar proposal would be backed by its neighbors in Brooklyn



Get our daily Cobble Hill, Carroll Gardens & Red Hook news and alerts!

Goal: Build a model that predicts real estate values based on proximity to transit

- Multivariate linear regression
- Geo-spatial variables
- Focus on transit accessibility as primary independent variable

Approach: Identifying the dependent variable

- City Department of Finance maintains property tax records for every lot in the city
- As part of these records, they calculate an estimated market value for each lot
- Methodology isn't perfect, but good proxy for real estate values

NYC Finance

Statement Details

June 5, 2015
Nyc Dcas
1 Centre St.
1-00121-0001
Page 3

Overpayments/Credits	Activity Date	Due Date	Amount
Finance-Property Tax		01/01/2013	$-48,036.68
Finance-Property Tax		07/01/2013	$-39.64
Finance-Property Tax		04/01/2014	$-37,188.32
Finance-Property Tax		01/01/2015	$0.00
Credit Transfer Received	04/17/2015		$-394.81
Payment Transferred to	04/03/2015		$-2,053.96
Fire-Prevention Inspection 83826		10/01/2013	$-280.00
Total overpayments/credits remaining on account			**$-486,588.5**

Annual Property Tax Detail

Tax class 4 - Commercial Property
Current tax rate — **Tax rate** 10.6840%

Estimated market value $180,900,000

	Billable assessed value		Tax rate		Taxes
Tax before exemptions and abatements	**$69,839,910**	**X**	**10.6840%**	**=**	**$7,461,696**
Dept Real Est (Dcas)	$-69,839,910				$-7,461,696
Annual property tax					**$0**

While the data is public, it's in this format- luckily someone has already scraped it and created a single, ~1M row CSV for every parcel in the city

Approach: Join BBL-level DOF data to MapPluto dataset

- MapPLUTO can be thought of as the dataset of record for all landuse related data about the city
  - Unit of analysis is the BBL- Borough Block Lot
  - ~800k rows
  - Includes hundreds of columns, including info on
    - Zoning designations
    - Districts
    - Building age/size
    - Special rules/regulations
    - ...and much more- a treasure trove to identify potential control variables!!
- Every BBL in the city is geo-coded, and therefore can be used for spatial analysis
- All of the DOF data is also available at BBL level, facilitating the join

Approach: Join BBL-level DOF data to MapPluto dataset

- Joined on BBL
- Scrubbed for rows with missing or incomplete values
- Eliminated non residential real estate or unbuilt properties
- Resulted in dataframe of ~600k rows, ready for further analysis

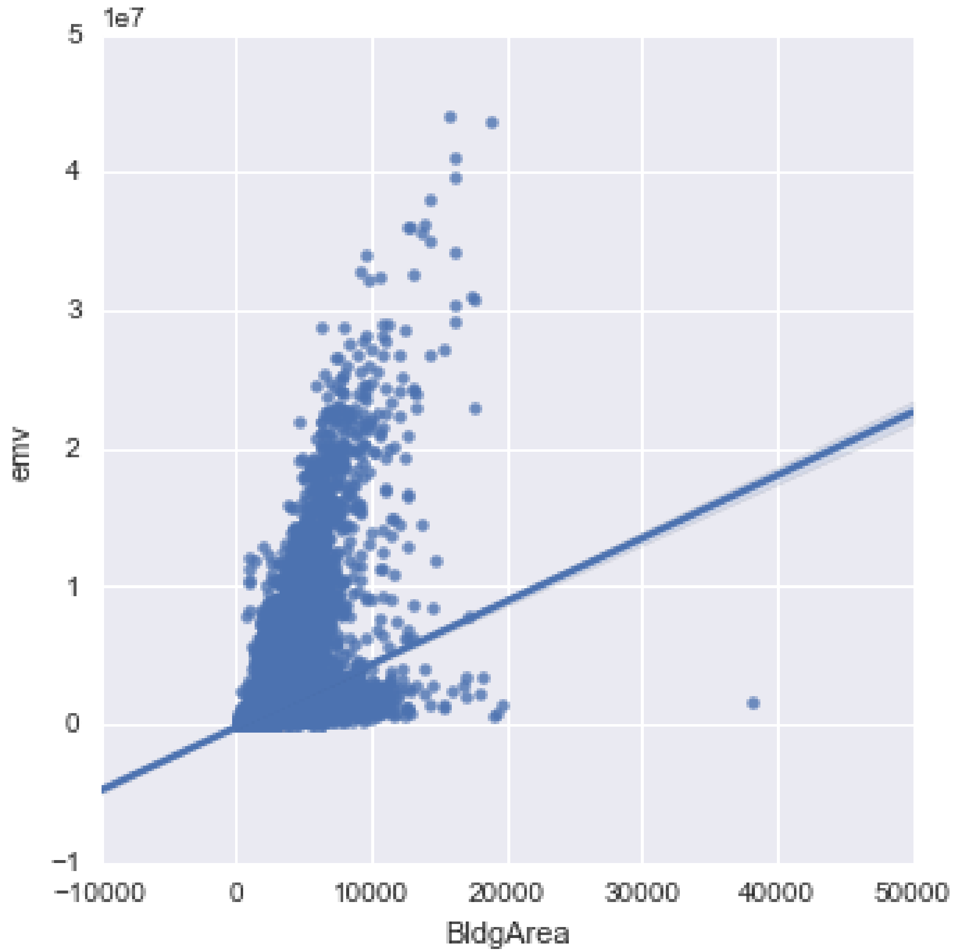
	BBL	Borough	SchoolDist	PolicePrct	BldgArea	emv
1	1000970045	MN	2.0	1.0	1845	2424000.0
2	1000970055	MN	2.0	1.0	13015	8644000.0
3	1000970144	MN	2.0	1.0	1880	1955000.0
4	1001350011	MN	2.0	1.0	11515	9104000.0
5	1001400024	MN	2.0	1.0	11913	7400000.0

Example of df.head()- many more columns…

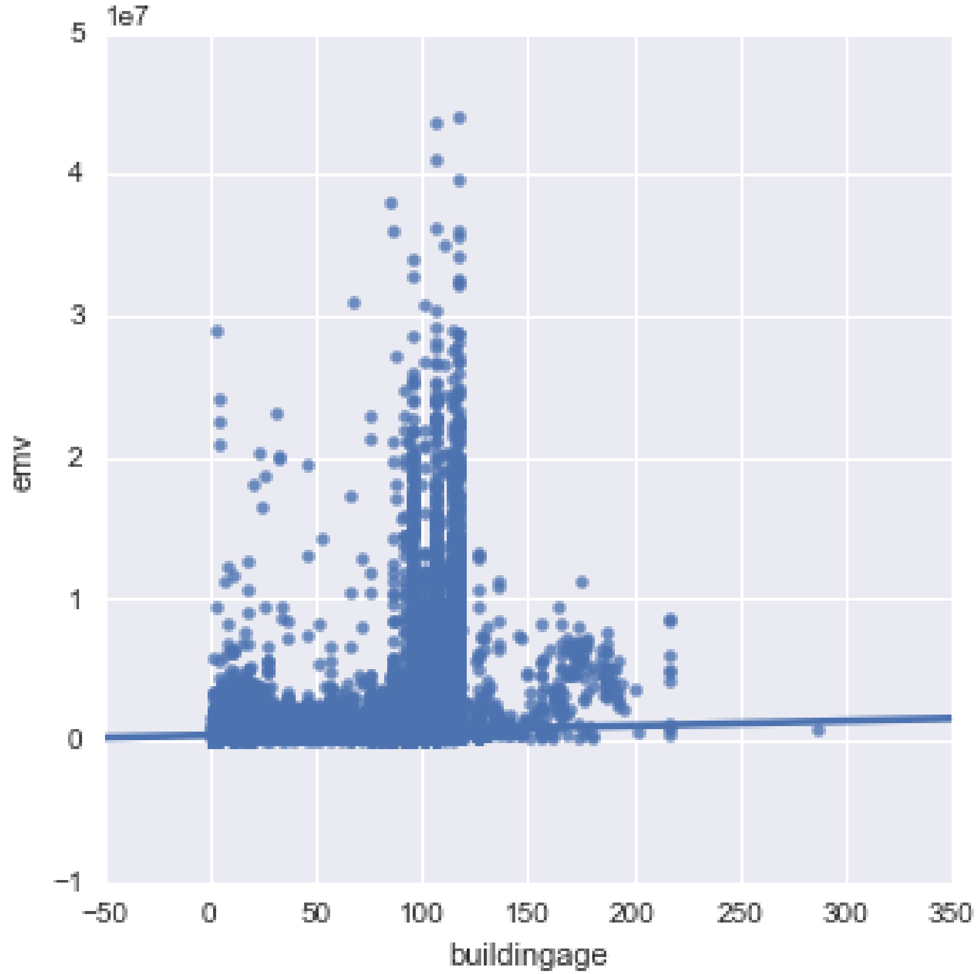
Approach: Identify potential control variables and identify source for them

Source	Potential control variable	How to encode	Variable Details
MapPLUTO	Building Area	Use value in MapPluto	Total sq. feet of building floor area
	Building Age	Use year built to calculate age	Years, drop any value >200 (likely data error)
	Historic District	Encode dummy based on whether in historic district	True/False
	Landmarked Building	Encode dummy based on whether landmarked	True/False
	Borough	Run getdummies for all boroughs	T/F for all boroughs except 1
Police Dep.	Crime Rate	Join police crime rate data by precinct on precint field on MapPLUTO	# of crimes per 1000 residents
DOE	School Quality	Join school grad data by school district on schooldist field on MapPLUTO	4 year grad rate in district
Census data	Race, Income	Join Census demographic data on census block/NTA	% white, median income, in NTA

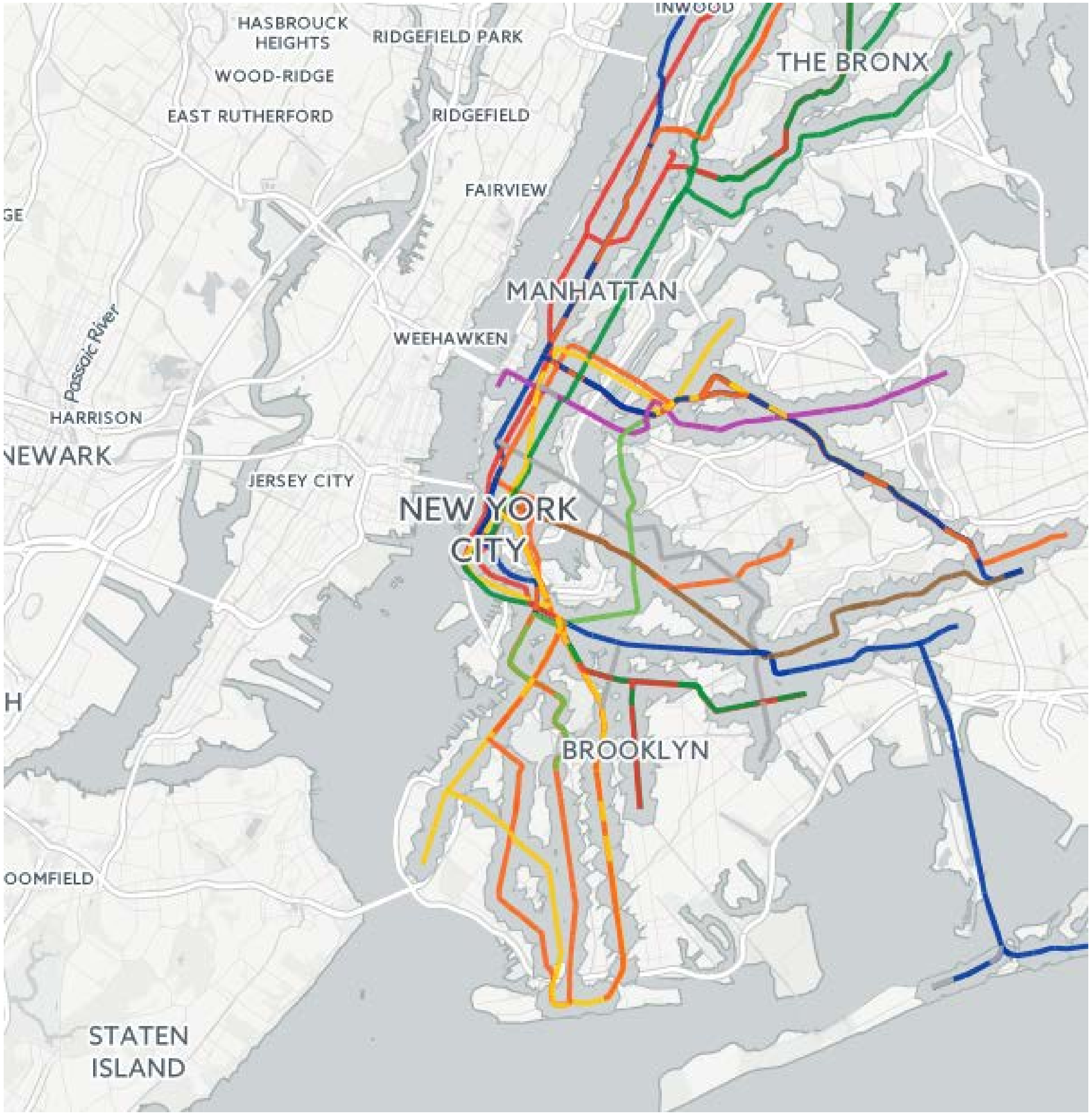
Some variables are clearly correlated with EMV
Correlation: Building Area & EMV, R^2=.48



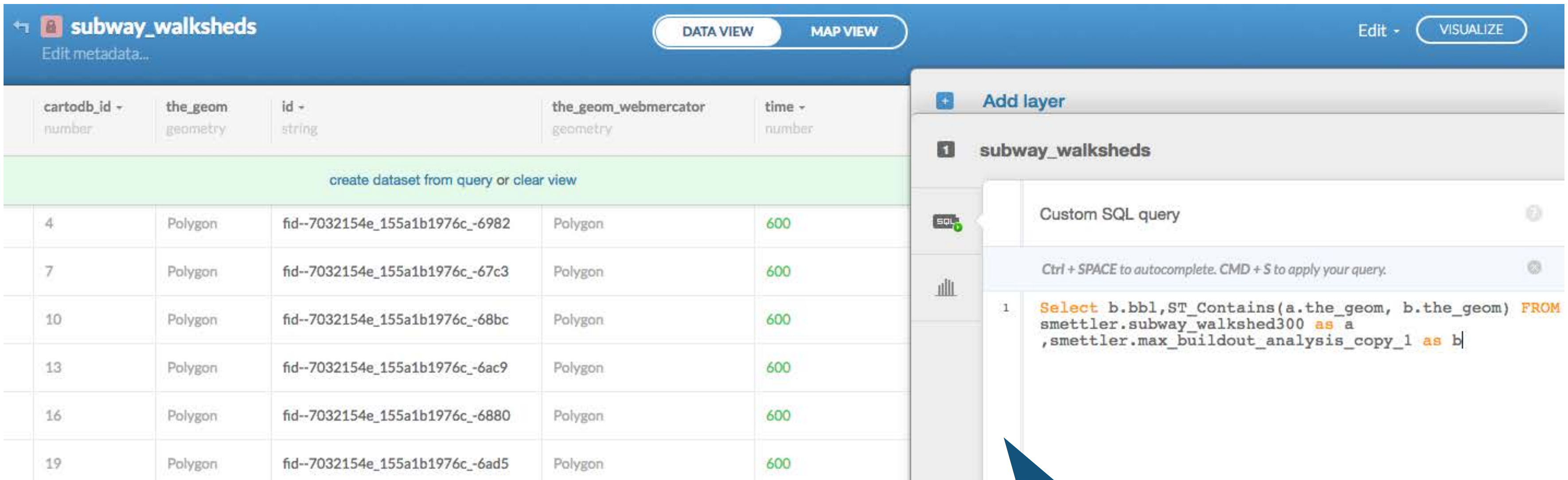
Others are less clearly correlated with EMV
Correlation: Building Age & EMV, R^2=.12



Approach: How to evaluate access to transit



Approach: Using subway walkshed map and cartodb, create a table of bbls within 5, 7.5, and 10 min of a subway stop



Custom SQL query to create CSV files of BBLs- those files are then loaded into Python with pd.Load_csv and joined into the overall DF

Preliminary analysis: Being in a subway walkshed appears to be a significant variable

ANOVA table with 3 different subway dummy variables

	df	sum_sq	mean_sq	F	PR(>F)
Borough	4.0	1.177774e+17	2.944435e+16	85944.569301	0.000000e+00
within300	1.0	3.793012e+12	3.793012e+12	11.071355	8.767758e-04
within450excl	1.0	7.604393e+13	7.604393e+13	221.963246	3.454351e-50
within600excl	1.0	1.178184e+14	1.178184e+14	343.898020	9.558487e-77
Residual	528562.0	1.810837e+17	3.425969e+11	NaN	NaN

ANOVA table with singular subway dummy variables

	df	sum_sq	mean_sq	F	PR(>F)
Borough	4.0	1.177774e+17	2.944435e+16	85950.044407	0.000000e+00
within600	1.0	2.085055e+14	2.085055e+14	608.641828	2.632935e-134
Residual	528564.0	1.810729e+17	3.425751e+11	NaN	NaN

Initial Regression: All variables significant

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    emv   R-squared:                       0.588
Model:                            OLS   Adj. R-squared:                  0.588
Method:                 Least Squares   F-statistic:                 5.023e+04
Date:                Tue, 09 Aug 2016   Prob (F-statistic):               0.00
Time:                        23:22:33   Log-Likelihood:            -7.6678e+06
No. Observations:              528570   AIC:                         1.534e+07
Df Residuals:                  528554   BIC:                         1.534e+07
Df Model:                          15
Covariance Type:            nonrobust
======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------------
Intercept           -1.339e+06   9229.567   -145.042      0.000     -1.36e+06 -1.32e+06
ishist[T.True]       7.268e+05   5077.913    143.136      0.000      7.17e+05  7.37e+05
islandmark[T.True]   7.785e+05    3.64e+04    21.379      0.000      7.07e+05   8.5e+05
BldgArea              302.8127      0.924    327.697      0.000       301.002   304.624
school_grad_rate     2.915e+05   9386.951     31.052      0.000      2.73e+05   3.1e+05
crime_rate          -1745.0011    263.551     -6.621      0.000     -2261.552 -1228.450
is_BK                4.726e+05   2957.690    159.795      0.000      4.67e+05  4.78e+05
is_BX                3.484e+05   3614.031     96.415      0.000      3.41e+05  3.56e+05
is_MN                4.312e+06   9041.076    476.935      0.000      4.29e+06  4.33e+06
is_QN                3.591e+05   2494.038    143.973      0.000      3.54e+05  3.64e+05
buildingage          1414.5918     27.376     51.674      0.000      1360.937  1468.247
Med_Income             10.0170      0.051    195.273      0.000         9.916    10.118
perc_white           2.553e+05   3104.883     82.215      0.000      2.49e+05  2.61e+05
within300           -4.462e+04     1.5e+04    -2.967      0.003     -7.41e+04 -1.51e+04
within450excl        1.215e+04   4416.617      2.752      0.006      3496.929  2.08e+04
within600excl        1.225e+04   3302.256      3.711      0.000      5781.382  1.87e+04
==============================================================================
Omnibus:                  1028137.063   Durbin-Watson:                   0.470
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       8468613974.126
Skew:                          14.916   Prob(JB):                         0.00
Kurtosis:                     622.380   Cond. No.                     3.47e+06
==============================================================================
```

Mostly as expected, some surprises:

- Within300 is negative but the other transit ones are positive
- Landmarks and historic districts are very strong- may be capturing something else
- Building age has a negligible effect- likely because of a non-linear distribution

Regression # 2: Remove non-helpful independent variables

With some adjustments, looks slightly better:
- R squared value about the same @ .59
- Changing dummy variable for transit to simply be within 10 minutes of subway or not results in stronger coefficient
- Eliminating building age has negligible effect on overall model

OLS Regression Results

Dep. Variable:	emv	R-squared:	0.586
Model:	OLS	Adj. R-squared:	0.586
Method:	Least Squares	F-statistic:	6.225e+04
Date:	Tue, 09 Aug 2016	Prob (F-statistic):	0.00
Time:	23:41:44	Log-Likelihood:	-7.6691e+06
No. Observations:	528570	AIC:	1.534e+07
Df Residuals:	528557	BIC:	1.534e+07
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.253e+06	9106.161	-137.625	0.000	-1.27e+06	-1.24e+06
ishist[T.True]	7.636e+05	5034.673	151.670	0.000	7.54e+05	7.73e+05
islandmark[T.True]	8.232e+05	3.65e+04	22.556	0.000	7.52e+05	8.95e+05
BldgArea	295.5614	0.916	322.688	0.000	293.766	297.357
school_grad_rate	3.004e+05	9407.813	31.936	0.000	2.82e+05	3.19e+05
crime_rate	-1787.6201	264.367	-6.762	0.000	-2305.771	-1269.470
is_BK	5.203e+05	2818.728	184.591	0.000	5.15e+05	5.26e+05
is_BX	3.812e+05	3567.135	106.867	0.000	3.74e+05	3.88e+05
is_MN	4.388e+06	8945.866	490.467	0.000	4.37e+06	4.41e+06
is_QN	3.965e+05	2393.308	165.689	0.000	3.92e+05	4.01e+05
Med_Income	9.8308	0.051	191.816	0.000	9.730	9.931
perc_white	2.616e+05	3109.240	84.130	0.000	2.55e+05	2.68e+05
within600	2.686e+04	2860.545	9.388	0.000	2.12e+04	3.25e+04

Omnibus:	1027302.856	Durbin-Watson:	0.466
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8429637978.700
Skew:	14.889	Prob(JB):	0.00
Kurtosis:	620.953	Cond. No.	3.47e+06

Further analyses: Adjusting and fine-tuning regression models

	Model	Training R^2	Test R^2	Coefficient on transit dummy	Comments
1	**Baseline**	0.5848969376	0.5866645106	2.95E+04	Baseline model previously described
2	**Baseline, w/ intercept of 0**	0.5700045197	0.5719172058	3.27E+04	Makes school quality a negative indicator
3	**Baseline, normalized**	0.5848969376	0.5866645106	2.95E+04	Doesn't appear to have any effect over non-normalized equivalent
4	**Baseline, w/ intercept 0, normalized**	0.5700045197	0.5719172058	3.27E+04	Doesn't appear to have any effect over non-normalized equivalent
5	**Baseline, Lasso**	~ .57	~ .57	~3E04	Higher alpha results in slightly lower transit coefficient
6	**Baseline, Ridge**	~ .5-.55	~ .5-.55	Up to ~9E04	Higher alpha increases transit coefficient, but decreases R squared

Even with some adjustments and fine-tuning, R^2 steady ~0.6, and roughly consistent performance across test/train; Ridge suggests transit coefficient could more significant than others

However, a simple dummy model performs better

Created a simple dummy model that only considers NTA, neighborhood, building size, and the binary transit variable

- Considers all ~180 NTAs (corresponding roughly to neighborhoods) as their own dummy variable and regresses on them
- Assumes that within that neighborhood “dummy”, all things that could affect value- e.g. attractiveness of neighborhood, crime, school quality, etc.- are already captured
- Coefficient for binary transit variable: 3.22220691e+04
  - Consistent with other regressions/models
  - Suggests ~$32K of value can be attributed *only to* being within 10 min of a subway
- R squared on test: 0.79,
- R squared on train: 0.80

Question for further analysis: What other “quantifiable” metrics might exist that could explain what makes a neighborhood valuable in real estate

Summary

Conclusion

- Regardless of model used or way parameters were cut, the binary "within 10 min of subway" variable is always significant and contributes ~$25-$40K to the value of a property
- Most attributes that were analyzed perform as expected, but there are certainly other predictive variables are out there that can be analyzed/quantified
- It seems safe to extrapolate this model as predictive- in an "all things equal" scenario, e.g. the opening of a new subway will add value to properties in East Harlem- BUT of course the opening of the new subway could also change the neighborhood more fundamentally in other ways that increase value (e.g. independent variables are related)

Shortcomings

- Data at different levels/units, so not always granular enough
- EMV methodology is imperfect doesn't always reflect up to date reality- Opportunity to analyze other data, e.g. zillow transaction data, rents
- Could associate an exact time to subway for each BBL instead of a simple binary variable
- Count areas with more subways differently, weigh the "quality" of a certain subway stop
- Consider buses and ferries and their effects