Spatial & Real Estate Data

Team Reveal Estate
Nora Barry, Laura Buchanan, Jackie Gutman

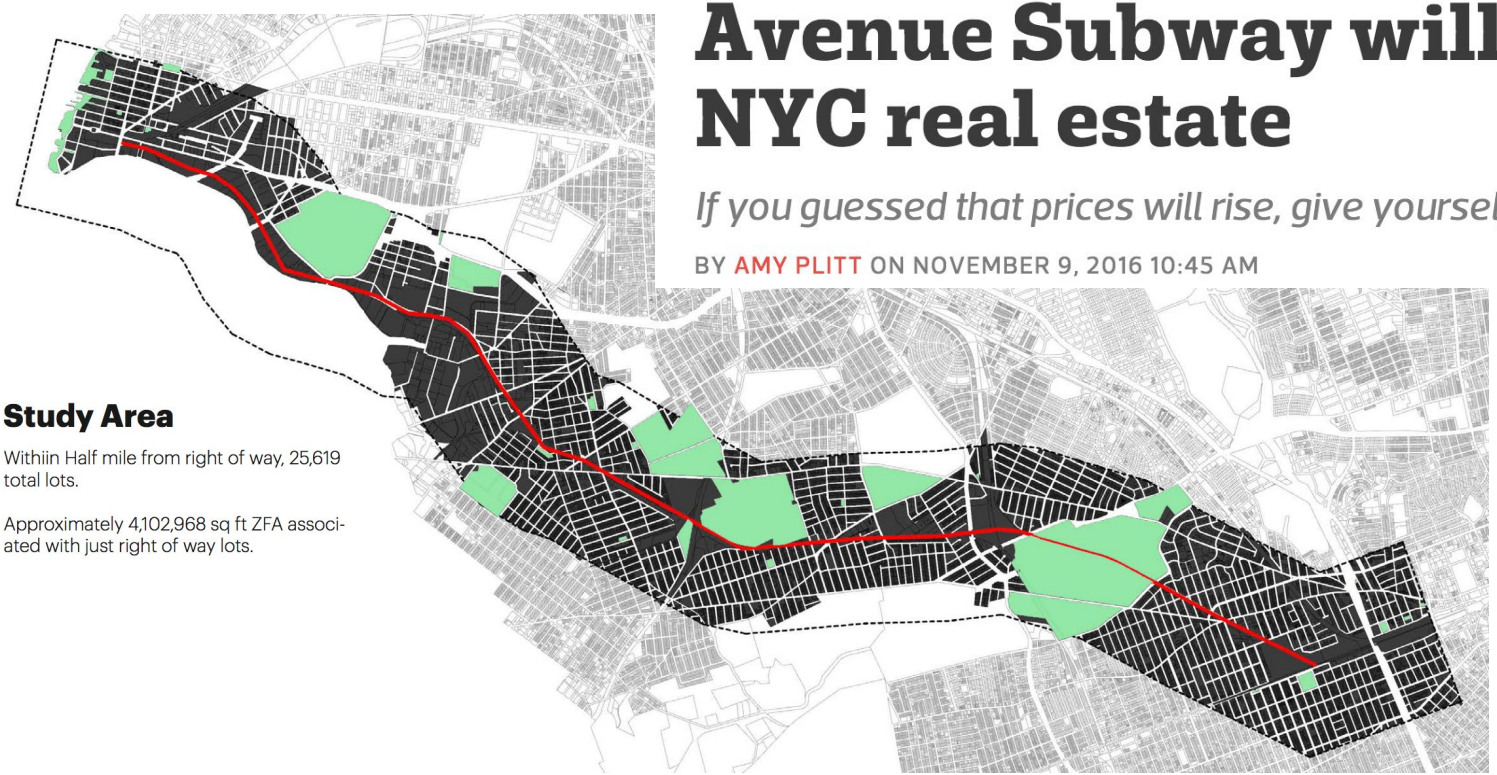
Queens Light Rail

CURBED

Here's how the Second Avenue Subway will affect NYC real estate

If you guessed that prices will rise, give yourself a gold star

BY AMY PLITT ON NOVEMBER 9, 2016 10:45 AM



Update

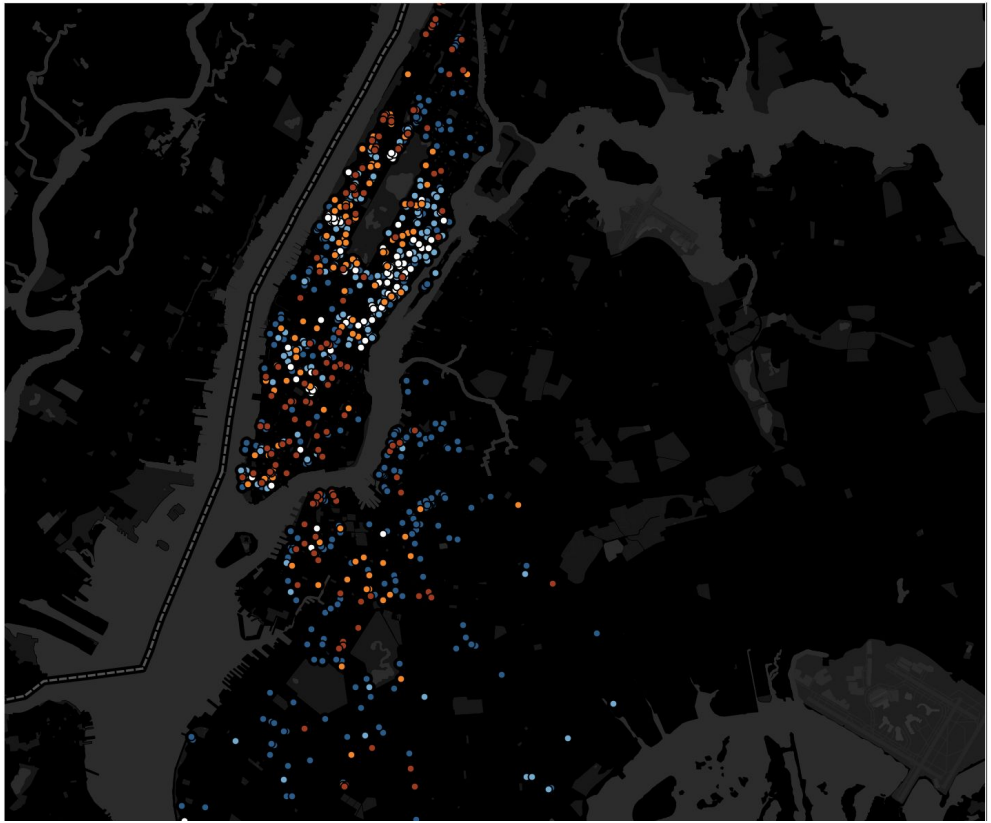
Data Clean

- Drop uninformative features in PLUTO data
- Drop PLUTO instances without latitude and longitude
- Binarize appropriate features, e.g., lots in limited **height districts**, lot in **historic districts**, lot with **landmarks** on property, lot with **basements**, etc.
- Deal with impossible values, e.g., buildings built in 2040
- Create dummy variables: **borough**, **school district**, **building class**, **lot owner type**, etc.

Data Merge

- Distance to **subway**, **PATH**, **tree census**, **botanical gardens**, **libraries**, **colleges and universities**, **outdoor cafes**, and **day cares**

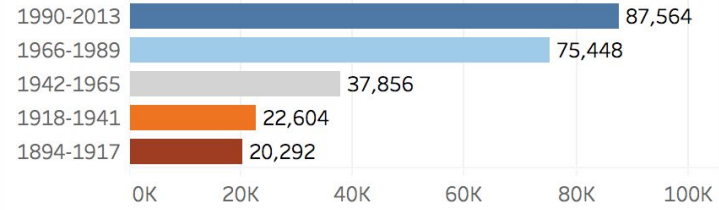
Low Price ($1 - $5 per sqft)



Year Built

1900.0 2015.0

Number of Low Prices Per Year Built



Issues & Solutions

Issues

- Many duplicate sales records for same lot in single year
- Anomalies in target variable price per sqft
- Many price per sqft < **$5**

Solutions:

- Average sales records over each year
- Filter out rows where price per sqft < **$5**
- Model MSE decreases by **> 3,000**

Preliminary Results

- Linear Regression and Random Forest models with data from:
  - PLUTO
  - Dept. of Finance & Sales
  - Subways

Model Setup		MSE	Accuracy
Linear Regression	All records	4976.12	5.43%
	Price Per Sqft > 0	3867.97	8.56%
	Price Per Sqft >= 5	3076.29	27.64%
Random Forest	All records	3835.71	13.82%
	Price Per Sqft > 0	1359.81	64.46%
	Price Per Sqft >= 5	730.32	69.66%

Next Steps

- Add more data sources (NYC Open Data)
- Determine optimal evaluation metric for comparing performance
- Iterate over various regression model families to hone in on best model
- Predict real estate values in affected lots with and without new light rail
    - Causal inference estimation methods (Athey & Imbens)