Project Proposal- Linear Regression/Web Scraping

Question/need:

- I plan to build a model that aims to predict house prices in Staten Island based on metrics like square footage, number of bedrooms/bathrooms, proximity to schools/transit and additional features (backyard, basement, attic). This model aims to help home-buyers in selecting a home that would provide the most value for its cost, based on the above features.

Data Description:

- I plan on using data from recently sold houses on Trulia with similar features to serve as the observed value of the dependent variable (house price), with which to compare the predictions from my model.
- An individual sample/unit of analysis in this project would be a house listing. Each house listing would be evaluated by features such as square footage, number of bedrooms/bathrooms, proximity to schools/transit and additional features (backyard, basement, attic).

Tools:

- The tools I plan on using are BeautifulSoup, Selenium, Sklearn, Pandas, Matplotlib, Seaborn, Numpy
- Additional tools may be needed in order to calculate proximity to transit/schools if I cannot find that metric on Trulia or elsewhere- I am not sure what those tools may be yet.

MVP Goal:

- An MVP for this project would be a jupyter notebook showcasing a working regression model that predicts the value of at least one recently sold home (based on the above features) and comparing that value to the observed value when it was sold.