Simon Mettler

Data Science

Final project milestone #2: Project Write-up

**Project Problem Statement:**

Real estate is a big driver of the New York economy, and real estate prices in the city are among the highest in the world. It seems that people are willing to pay a significant premium in order to live in the city, and especially in certain neighborhoods. Some of the reasons for this are well understood- New York is a hub of ideas, innovation, and creativity that attracts people from around the world. It is also rich in restaurants, cultural institutions, and other amenities that make life in the city great. However, not every part of the city is equally valued by current or potential residents, and real estate prices vary widely by neighborhood. There are, of course, many potential reasons that can make some neighborhoods more desirable than others. One major reason likely is the proximity to, and accessibility of, public transit. New York as a neighborhood is known for its public transit system (especially its subways), and proximity to a subway is often advertised as a desirable living trait. This project will aim to more closely explore the relationship between proximity to subways and the real estate market.

**Hypothesis:**

Even when controlling for other variables that could affect real estate values , proximity to subways will have a positive effect on real estate prices throughout the city.

**Approach to problem:**

From a statistical standpoint, a regression would likely form the basis of a predictor of real estate values. The model could be structured to predict a continuous number (property value), based on various inputs, including proximity to a subway line. While proximity to subway will likely have a significant impact on values, one of the purposes of this project will be to attempt to quantify just how much of an impact it has. At the same time, the model could also identify which other variables (e.g. size, crime, school quality) have a significant effect on property prices.

More immediately, a particularly beneficial impact of the model intersects with several ongoing policy discussions in the city of New York. The L Train is now slated to shut down for a full 18 months in 2019. Before then, the first phase of the new Second Avenue Subway is slated to open in the Upper East side. And the mayor has proposed a streetcar to connect Astoria to Sunset Park. In all three of those cases, the impact on real estate prices is a big topic of conversation. In the latter, the city is even proposing to finance the streetcar’s construction through bonds backed by the property taxes likely to be collected as a result of increased property values. Therefore, the model will have several useful applications.

**Datasets:**

Several publicly available or derived datasets exist that could be helpful in this analysis. The most important is the DOF dataset, which lists estimated market values for every parcel in New York City. (The creation of that dataset was its own project, detailed extensively [here](http://chriswhong.com/open-data/liberating-data-from-nyc-property-tax-bills/)). In addition to that, NYC’s MapPluto ([described here](http://www1.nyc.gov/assets/planning/download/pdf/data-maps/open-data/pluto_datadictionary.pdf)) has a trove of information available about each parcel, and can be joined to the previous dataset for more richness. To get at the various potential control (independent variables), values in MapPluto, or based on lookups in other city datasets, will be essential. This includes data on school quality ([here](http://schools.nyc.gov/Accountability/tools/report/default.htm)), crime rates ([here](http://www.nyc.gov/html/nypd/html/crime_prevention/crime_statistics.shtml)), and demographics ([here](http://maps.nyc.gov/census/), among other places). And of course, the data on subway accessibility itself. Several routes are possible here, but the easiest first step is to use this [analysis](https://cwhong.carto.com/tables/combined/public/map), which has already done some of the geospatial work involved. (The analysis created a “walkshed” around each subway station in the city, analyzing what areas fell within a 5, 7.5, or 10 minute walk of a station). The next step then will be to match that map to a list of all parcels in the city, which can be done using a tool like CartoDB. Once complete, a binary variable can then be created for whether a given parcel is within 5, 7.5, or 10 minutes of a subway stop.

**Domain Knowledge:**

As an employee of the city, I frequently use the various datasets of city data that exist, and have a good understanding of which data are helpful to use where. I also frequently conduct geospatial analyses, which will be helpful in this project.

Various research efforts have attempted to analyze this problem, and while the general conclusion has been that proximity to a subway has a positive correlation with real estate prices, there is disagreement over the magnitude of the effect and of the other variables to analyze that also account for differences in prices.

**Project Concerns:**

* While a regression will likely be able to identify some sort of relationship, it is unclear whether

1. A simple regression will suffice as the correct approach (and whether some of the underlying independent variables are distributed in a way that enables regression)
2. Whether it will ever be possible to capture all possible independent variables that could affect real estate values in a regression, and whether data about the independent variables can be collected at a unit of analysis that is sufficiently small

* Assumptions and caveats
  + The DOF property value data is not perfect- it uses a three year rolling average (and so doesn’t capture rapid changes in appreciation/depreciation), and it is only an estimate; an actual dataset of all real estate transactions over the past 5 years, or of all new rental contracts, would be more helpful
  + The method for determining proximity to a subway, while effective, is somewhat crude by being binary; a more nuanced approach could look at exact walk time or distance from a subway as a continuous variable, for example; however, this would require significantly more computing power and would be challenging to achieve within the scope of this project
  + For the purposes of this project, all subways are considered equally and weighted the same- in reality, of course, certain subway stations and lines are much more “valuable” to be near, and would have a different effect on prices. Future versions of this analysis could consider this, but for this project the assumption will be made that all subways are “equal”
  + One workaround to the above issue, which this project will consider, is looking at a few “all else equal” scenarios; e.g. identifying neighborhoods that, for all intents and purposes, should be identical in their effect on property values (e.g. similar demographic and economic info), except that some properties within the neighborhoods are closer than others to the subway
  + While the overall DOF data/mappluto is comprehensive (e.g. includes every parcel in the city), the final analysis will likely have to drop several values where data is missing or incomplete; however, since the initial dataset is ~800k+ rows, even a dataset with only half of those rows would still be significant.
* Risks
  + The model is unlikely to be wrong directionally, but of course could be inaccurately in its quantifying of magnitude. In that, the risks of using it in actual financial or policy-making decisions could be problematic. However, if interpreted as a heuristic and as a guide, rather than as a precise predictor, it can be helpful to both real estate professionals and city policymakers
  + While none of the data is known to be incorrect, it is certainly possible that there are missing values or otherwise flawed assumptions in some of the datasets; there is also a risk that in joining different datasets created at different units of analysis (or different geographies), error is introduced into the model.

**Outcomes**

The outcome of the project will be a model that predicts real estate values based on various factors, and a set of coefficients that can be used, heuristically, to approximate the benefit of proximity to a subway. The target audience would be real estate professionals, policymakers, academics, and others with an interest in cities. The initial version of the model will likely not have to be that complicated, but will be more useful to start a conversation than to serve as the final word on anything. This is the type of project that can undergo many iterations, tweaking of assumptions, and scenario adjustments before being “deployed”; rather than being deployed directly, some of the underlying findings are likely to be incorporated into deal decisions and policies. Several near-term applications include decisions related to the L train shutdown, the Second Avenue Subway, and the BQX.