The Magic of a Borough

Victoria Michalska vm2@williams.edu
Computer Science
3045790

1. Introduction

In 2022, the value of the NYC housing market was estimated at $3.51 trillion, and it appears that the market is not expected to stop growing anytime soon. The idea that real estate is a wise investment is a heavily debated topic, especially as young people grow more and more hesitant towards making such a purchase; but if one is to make it, it should be a wise one that functions more like an investment.

New York City has been publishing large sets of property data on the internet since 2003. Online, the Annualized Sales Updates feature all sales since 2003, broken down by detail and by borough; furthermore, their website also features Rolling Sales Data, which features sales from the past 12 months. Property Valuation and Assessment Data can provide data on properties that haven't been sold in the past 20 years.

While it is widely known that the highest price per square foot can be found in Manhattan, I'd like to discover the value added exclusively by nature of a property being featured in a given borough, accounting for the most notable luxuries and pitfalls of living in a given area. Those qualities include but are not limited to: the quality of the transportation in area, the square footage of apartments in the borough, and the median income level of the surrounding area.

The population data by zip code was gathered from data.betaNYC data.BetaNYC, income and demographic data was gathered from The U.S. Census Bureau Bureau, and the sales data is from NYC OpenData for the year 2021. Now the question is regarding where to buy: which location provides the most bang for your buck? In which of the boroughs would one be paying the most for exclusively the right to admit to living in a given borough? That will be my treatment variable. I hypothesize that Manhattan is the most overrated, and will result in the price of having a place in Manhattan (and only the value of being able to claim a location in Manhattan) being the highest.

2. Preliminaries

A directed acyclic graph is a series of nodes connected by directed edges, which can be articulated as a touple consisting of the vertices and their connecting edges with independent errors equipped with the do-operator (Pearl, 2009). Such a graph cannot contain a cycle. A causal model of a such a graph where each variable is considered a function of its parents and the independent error term, so the probability of a given variable V which can factorize via the following:

$$p(V) = \prod_{V_i \in V} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i)),$$

where $\text{pa}_{\mathcal{G}}(V_i)$ is the parents of V_i in $\mathcal{G}$. Within the context of the causal model, $\text{pa}_{\mathcal{G}}(V_i) \to V_i$ can be understood as $\text{pa}_{\mathcal{G}}(V_i)$ is the direct cause of V_i.

Conditional Independence's in $p(V)$ can be read off from the DAG via d-separation, i.e., $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. Backdoor adjustment requires a valid set that blocks any paths from the treatment to the variable, ignoring the directionality of the edges. This set allows for the following formula to be true and provides the average causal effect according to backdoor adjustment criterion:

$$\sum_{Z} p(Z) \times \mathbb{E}[Y \mid A = a, Z]$$

This backdoor adjustment can be used to find the average causal affect if it is run multiple times and then averaged, especially if bootstrapping is used. Bootstrapping is a technique where a dataset is sampled in order to manufacture more data, having calculated the corresponding probabilities.

The aim of this paper will be to find the average causal effect (ACE) of a variable a on an outcome Y; ACE is most efficiently defined as the difference between the expected value of the outcome when the variable is present versus when it is not.

Given that my data is not from a randomized control set, Greedy Equivalence Search (GES) will be run on my data in order to find possible causal structures in my data. GES is a score-based method for learning DAGs (Chickering, 2002), assuming that there are no unmeasured confounders in my data.

3. Methods

Because this data is not collected from a randomized control trial, backdoor adjustment is a viable option to be used to adjust for confounding variables. My data, having been collected from the U.S. Census Bureau, the NYC Open Data portal, and the BetaNYC's Community Data Portal, was cleaned and organized in such a way so that for each given property collected from the 2021 New York City Sales Data, the listed variables are included:

1. **SALE_PRICE**: The price a given property was sold at.
2. **ELEVATOR**: A binary variable marking presence of an elevator in the property.
3. **RESIDENTIAL_UNITS**: The number of residential units in the property.
4. **COMMERCIAL_UNITS**: The number of commercial units in the property.
5. **LAND_SQUARE_FEET**: The number of square feet on the property.
6. **SERVICES_COUNT**: The number of subway lines that run through the subway stops in the neighborhood.

7. **COOP**: Binary variable describing whether or not the property is in a co-op.

8. **CONDO**: Binary variable describing whether or not the property is a condo.

9. **FAMILY_DWELLING**: Binary variable describing whether or not the property is a house, rather than an apartment.

10. **MEDIAN_INCOME**: The estimated median income in the property's zip code.

11. **POPULATION_DENSITY**: The population density of the property's zip code.

12. **PERCENT_WHITE**: The percent of white people in the property's zip code.

13. **STATEN_ISLAND**: Binary variable indicating whether or not the location is in Staten Island.

14. **BRONX**: Binary variable indicating whether or not the location is located in the Bronx.

15. **MANHATTAN**: Binary variable indicating whether or not the location is located in Manhattan.

16. **QUEENS**: Binary variable indicating whether or not the location is located in Queens.

17. **BROOKLYN**: Binary variable indicating whether or not the location is located in Brooklyn.

In order to create connections between all of these variables, I choose to run a Greedy Equivalence Search in Tetrad to generate statistically likely structural relationships between the confounding variables.

Considering the binary nature of my treatment variables, I will be able to run backdoor adjustment on my DAG (which has been generated from Tetrad) in order to find the ACE of the property being in a given borough on the cost of the property.

I will be featuring 5 treatment variables, which all affect one another directly: **BRONX**, **BROOKLYN**, **MANHATTAN**, **STATEN_ISLAND**, and **QUEENS**. Only one could be true at any one given time because a property can only be in one given borough at any one time, and therefore should be a direct causal relationship between all of these variables: by being in Brooklyn, you are therefore in no other boroughs.

4. Results

Having run a Greedy Equivalence Search, the graph featured in Figure 1 was found. This graph, while complex, can be used to find the backdoor adjustment set by grabbing all of the parents of a given treatment variable.

Not accounting for the backdoor set, I calculated the average causal inference was calculated using the backdoor mean to be the following in the table:

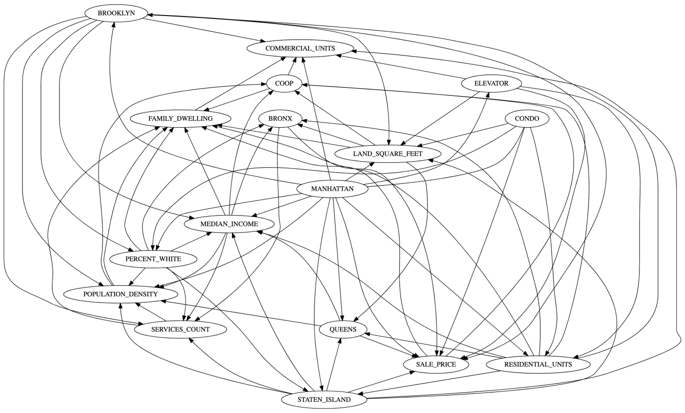

Figure 1: Graph found from Greedy Equivalence Search

Treatment	Average Causal Effect	Confidence Interval
BROOKLYN	1855955.2275742907	(1731246.543128, 1983469.167729)
STATEN_ISLAND	2148301.997333407	(2030414.113547, 2271603.073753)
BRONX	1938562.7800815117	(1819083.726922, 2050635.983169)
QUEENS	2316297.045536933	(2149038.90658, 2493400.412503)
MANHATTAN	1314578.4015892013	(1267715.87901, 1353817.567006)

Once accounting for the backdoor adjustment sets for each treatment variable, the results are shown in the following chart:

Treatment	Average Causal Effect	Confidence Interval
BROOKLYN	1672985.6098442036	(1578968.988044, 1782279.858110)
STATEN_ISLAND	2060435.0927335	(1912357.127318, 2213646.601449)
BRONX	1848148.208520	(1733844.171711, 1975548.650634)
QUEENS	2078241.34739	(1955108.002055, 2219801.21975)
MANHATTAN	1350921.892919	(1298765.64200, 1404818.780730)

5. Discussion and Conclusion

The graph found from the Greedy Equivalence Search did not feature edges between all of the treatment variables, which was expected; all of the treatment variables are connected by definition of only one of the binary variables ever being set to the value 1. This could be alleviated in future trials by including a required set of edges in the Greedy Equivalence Search.

Regarding the average causal effect, accounting for the backdoor set resulted in slightly shifted confidence intervals, but not smaller. The ordering of the treatments' average causal effects also did not change as a result of accounting for the backdoor set, only changing the differences between the different treatments. However, what is of note is the remarkably small margin by which Queens surpasses Staten Island; this small margin suggests that there may be unmeasured confounding variables involved in the graph.

In my introduction, I mention that I assume linear relationships between all of the variables: this may be, in reality, inaccurate. The most notable example of this may be the connection between residential units: while a single apartment may be in high demand, a property that contains two residential units may be less desirable, resulting in a lower price. However, purchasing a whole building is efficient and more valuable, meaning that then purchasing more residential units is more valuable. There is no variable that accounts for that in my data.

So, given all of these causal effects, what does it mean? Queens having the highest average causal effect means what? It means that purchasing a piece of property in Queens involves paying the most for the mere title of being in Queens, without the attached benefits of being in the borough. I suppose that means that according to my calculations, Manhattan, even though it tends to have the most expensive apartments, is the most worth it, considering related aspects.

References

U.S. Census Bureau. New york city neighborhoods populations and density. URL https://data.census.gov/.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

data.BetaNYC. New york city neighborhoods populations and density. URL https://data.beta.nyc/en/dataset/pediacities-nyc-neighborhoods/resource/7caac650-d082-4aea-9f9b-3681d568e8a5.

Judea Pearl. *Causality*. Cambridge University Press, 2009.