Business Intelligence Project

New York Property Insights Using Tableau

Master’s in Data Science & Busines Analytics
ESSEC & Centrale Supélec

Victor Hong
B00802726
victor.hong@essec.edu

1. Introduction

The goal of this project is to employ Tableau on publicly available data relevant to assessing the attractiveness of properties in New York. The objective is help an investor prospect for a property to stay and rent. The hypothetical client in question possesses the following characteristics and preferences - the criteria was chosen to reflect the likely preferences of contemporary home-owners and investors:

- Budget: $2 million USD
- Location: Manhattan, New York
- Type of Property: Residential
- Accessibility: Close proximity to subway stations
- Leisure: Close proximity to tourist attractions
- Rental Preferences: Prefer to rent during winter months (client goes somewhere warm during winter season)
- Other Concerns: Prefer area to be safe from violent crimes or home break-ins

The goal of the study is to use visualisation and other preparatory tools to provide an investor as much useful information as possible in a concise manner to not overload the investor but also facilitate the investor to make an informed decision at the end of the day. These include factors that likely play a role for future home-owners or travellers, such as space, safety, accessibility, attractions and price.

Guided by the above, I needed authoritative sources of information to avoid erroneous conclusions. Datasets were taken from official and authoritative sources as much as possible (NY Open Data mostly, or other official NY government sources were used) to ensure completeness, security and legitimacy of the information. Only the tourism dataset was taken from an outline source as this was not the most material consideration in our study.

Given the richness of dimensions and size of the datasets, conventional tools such as Excel would be hard-pressed to process, much less, present the data. Hence Python was used to clean and process the data, as well as engineer new features where necessary.

Tableau was used as a visualisation tool owing to its efficiency in handling large amounts of data, as well as its diversity of tools to present data in an elegant, intuitive and dynamic manner for users. Having a presentation tool that effectively presents the processed data is as important as the data processing and gathering in itself; at the end of the day, the goal is after all, to educate a consumer in simple terms without conveying too much complexity or causing confusion.

2. Data Modeling and Preparation

In preparation for this study, 5 sources of data were obtained to cover dimensions that reflect preferences for ownership. These concerned rent potential, nearby tourist attractions, proximity or availability of subway stations, crime data in surrounding area and history of property transactions. The data sources are reflected below:

2.1 Data Sources:

	Item	Source
1.	Airbnb rental data	Inside AirBnb: http://insideairbnb.com/get-the-data/ *Note: Data (I used the full dataset as it was more rich in variables) was downloaded from this link.
2.	Tourist attraction geolocation	MyGeoData: https://mygeodata.cloud/data/download/osm/tourist-attractions/united-states-of-america--new-york/new-york-county *Note: Data was exported as csv
3.	Subway geolocation	NYC Open Data: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49 *Note: Data was exported as csv from the website
4.	Crime data in surrounding area	NYC Open Data: https://data.cityofnewyork.us/Public-Safety/Crime-Map-/5jvd-shfj *Note: Data was exported as csv from the website.
5.	Historical transaction data	NYC Department of Finance (*detailed annual sales report by Borough)*: https://www.nyc.gov/site/finance/taxes/property-annualized-sales-update.page *Note: Data was downloaded from the year 2013 to 2021.

2.2 Data Pre-processing

Once we had all datasets, the columns that were pertinent for the analysis were chosen and/or pre-processed. Granular step-by-step details of this process are commented in the ipynb file attached alongside this submission. For the purpose of this report, I will describe the thinking process behind the pre-processing steps and leave the granularity in the ipynb file:

1. **Reducing dimensionality:** First, identify columns that would be interesting for this study;
2. **Combining Datasets:** This was specifically needed for the historical transaction data as the datasets were split by borough and by year. For e.g., for the year 2021, the datasets were available as - this meant I had to download 5 files for each year ranging from 2013 to 2021, a total of 45 files:
    a. 2021_Bronx.csv
    b. 2021_Queens.csv
    c. 2021_Manhattan.csv
    d. 2021_Staten_Island.csv
    e. 2021_Brooklyn.csv
3. **Pre-processing:** For columns that were engineered, they were pre-processed to ensure they were in suitable formats.
4. **Feature Engineering**: Since the study required extensive use of geolocation, I employed geopy and geocoder to complete the datasets (where necessary) to obtain features such as addresses, latitudes and longitudes and zipcodes. While possible, geopy and geocoder could encode ‘neighbourhood’ features as well, but they were not accurate or often missing.

2.3 Data Modelling

Once the data were obtained, cleaned, pre-processed and feature engineered where necessary, the data was modelled to represent how the structure of the data would be retrieved, processed and visualised once visualised on Tableau. The conceptual entity relationship schema and logical snowflake model are presented in the following figures. The data models also summarise which variables/features were ultimately used in the visualisation of the data for the user.

I had previously used JMerise to graph the following, but it has since crashed (i have followed the steps given to rectify JMerise but am still encountering issues - in any case I have taken screenshots of the encountered errors, and graphed the schemas using powerpoint directly on my own - i apologise for this inconvenience).

Figure 1: Entity-Relationship Schema

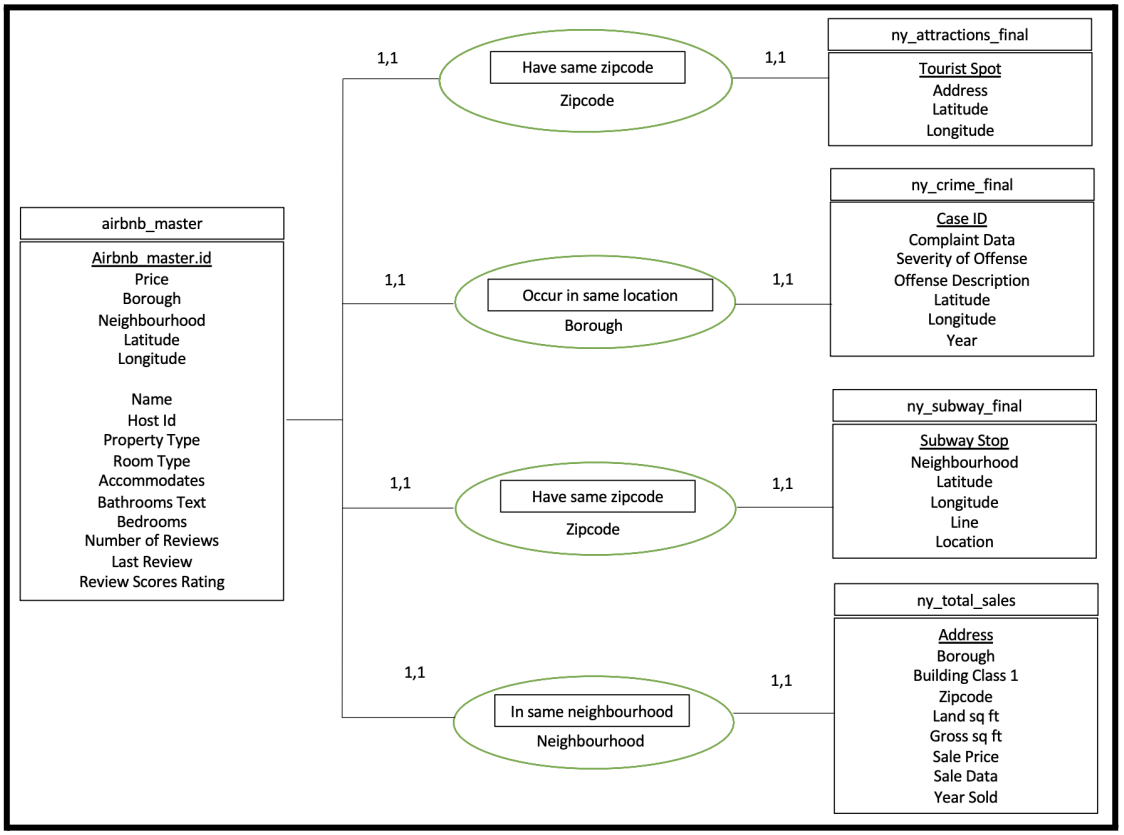
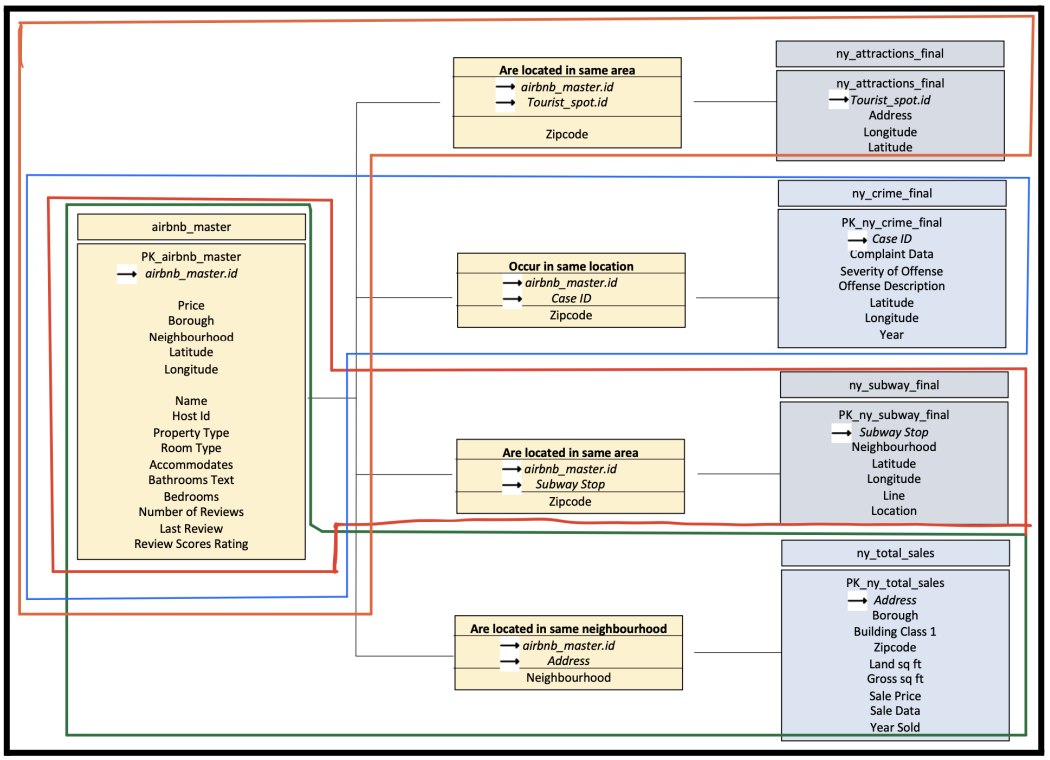


Figure 2: Logical Snowflake Schema



The key features were usually trivial or self-evident, with the exception of the following for clarity - features already explained in previous rows were not repeated, but should they have a different representation, they are explained accordingly:

Fact Table	Features	Description
Airbnb_master (data set containing rental information)	Price	Rental price of Airbnb
	Borough	One of the five boroughs of New York (Queens, Brooklyn, Manhattan, Staten Island, Bronx)
	Neighbourhood	Neighbourhoods within boroughs
	Latitude, Longitude	Geo-coordinates
Ny_attractions_ final (data set containing information on attractions)	Address	Address of attraction location
ny_crime_final (data set containing crime information)	Complaint_date	Data complaint was lodged
	Severity of Offense	Classified by misdemeanour, violation and felony, in increasing order of severity by penalty.
ny_total_sales (data set compiling sales data of property sold from 2013 to 2021)	Land sq ft	Land area of property listed in square feet
	Gross sq ft	The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building, including the land area and space within any building or structure on the property
	Building Class Category	This is a field that so that users of the Rolling Sales Files can easily identify similar properties by broad usage (e.g. One Family Homes) without looking up individual Building Classes.

***** IMPORTANT : The source files downloaded above exceed the zip capacity of 100 mb. Hence in my zip submission, I have only included my ipynb file, along with the processed data in excel (airbnb_master, ny_attractions, ny_crime_final, ny_total_sales, ny_subway_final - Please reach reach me at my email victor.hong@essec.edu / victorhong2017@gmail.com) if you have queries *****

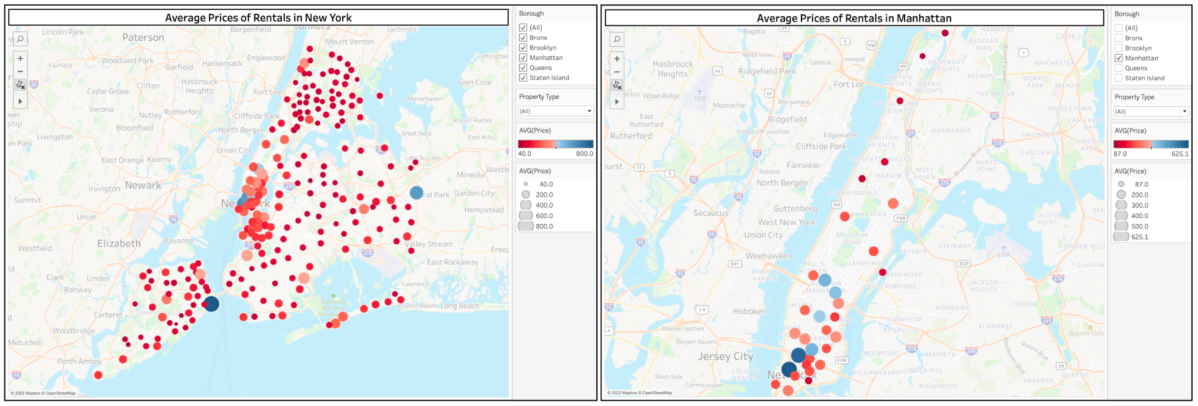
3. Application

Once the dataset was prepared, the visualisations were produced on Tableau. The Tableau file in this submission will also contain dashboards and the story feature (included in appendix) - they are appended here for completeness, along with the explanations I had similarly presented physically in class.

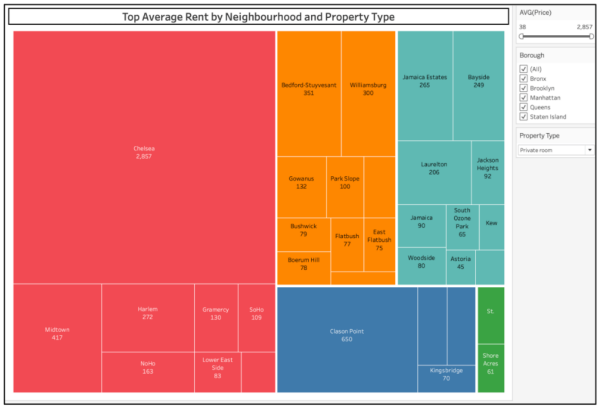
Goal: To determine if Manhattan meets the criteria set in the introduction:

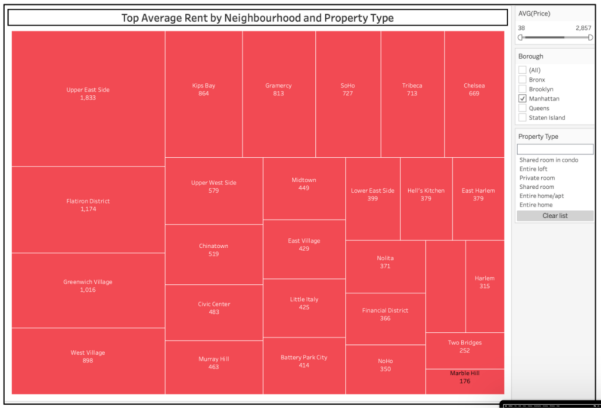
- Budget: $2 million USD
- Location: Manhattan, New York
- Type of Property: Residential
- Accessibility: Close proximity to subway stations
- Leisure: Close proximity to tourist attractions
- Rental Preferences: Prefer to rent during winter months (client goes somewhere warm during winter season)
- Other Concerns: Prefer area to be safe from violent crimes or home break-ins

Step 1: First, we use geolocation to identify hotspots that prioritise areas that yield the highest rent on average. As the client is interested in Manhattan, we take a deeper look and find Manhattan to indeed be promising in terms of rental yield.

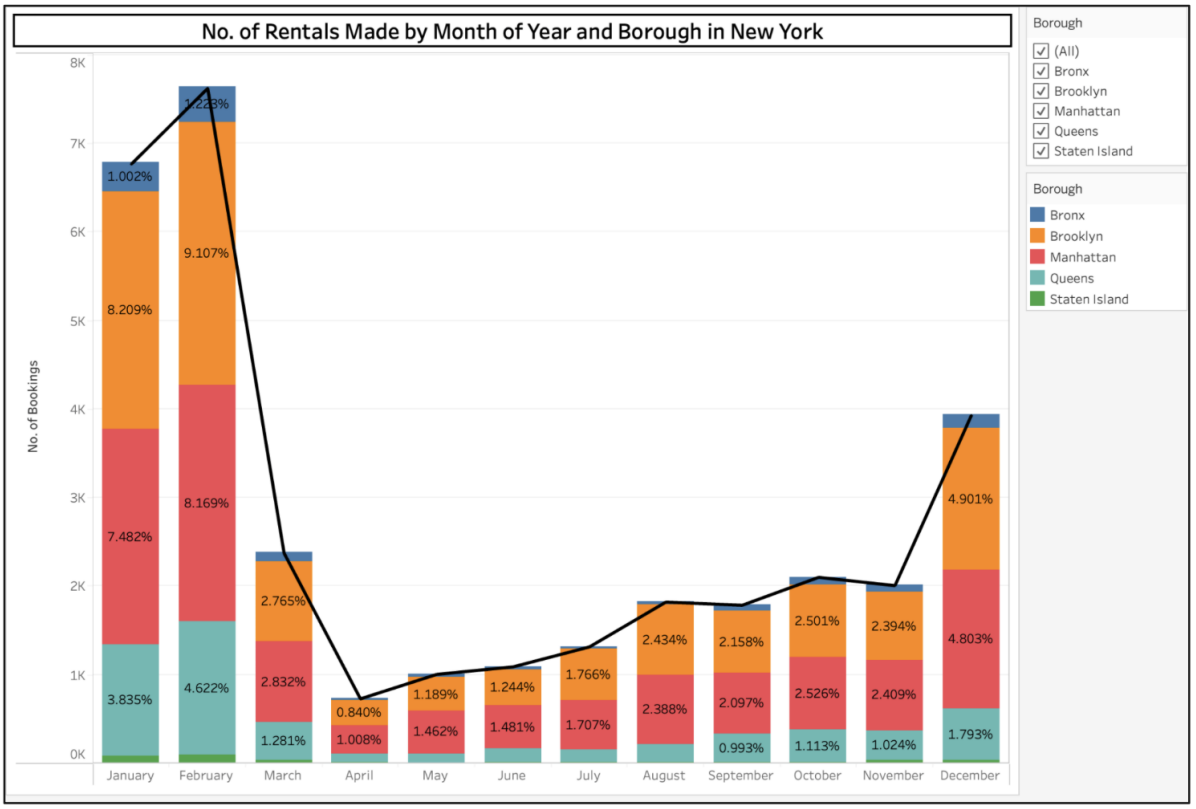


Step 2: Once we establish there is a market for rentals in Manhattan, we want to see if this is generally the case for residential properties. The next visualisation shows Manhattan has respectable rental potential among different types of residential units.

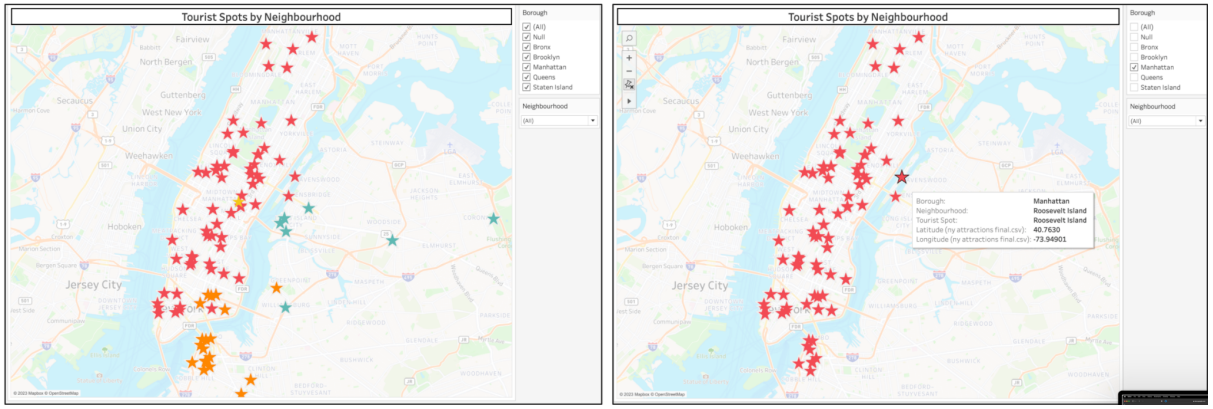




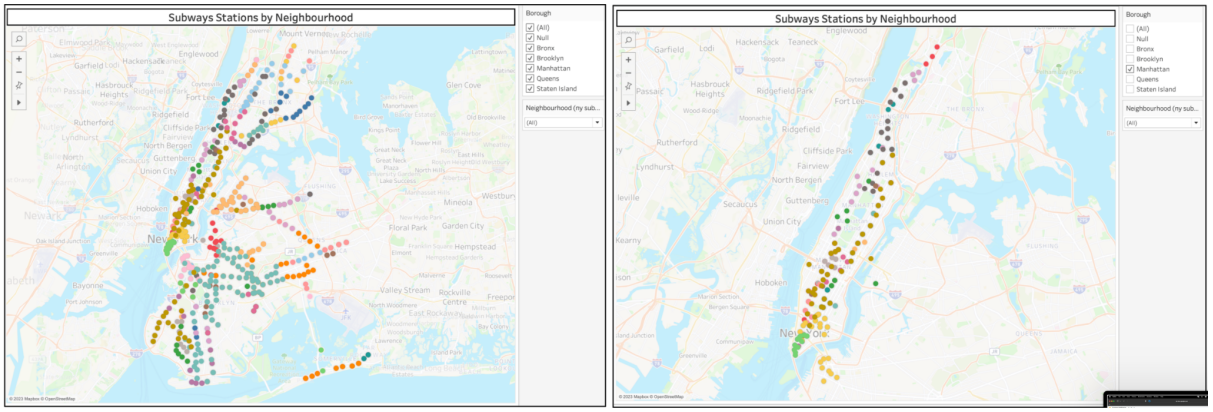
Step 3: Next, we look at temporal data to see if rentals are indeed seasonal in New York. A quick aggregate of monthly data shows that the number of rents made are the highest between the months of December, January and February - which suits the needs of the client.



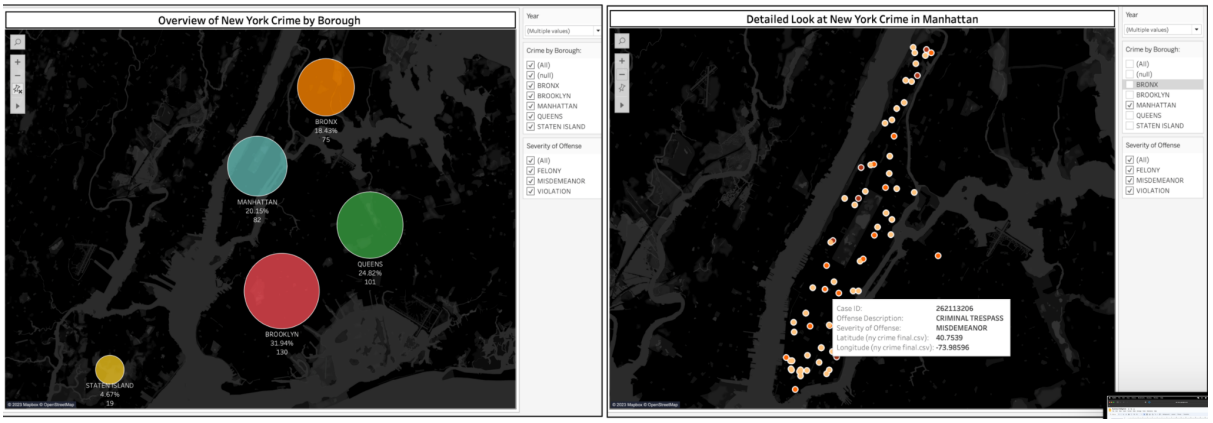
Step 4: Once we have met the criteria of rent potential, we start to look at the other preferences of the client concerning leisurely activities. Here we see that locations in Manhattan are generally accessible with tourist attractions.



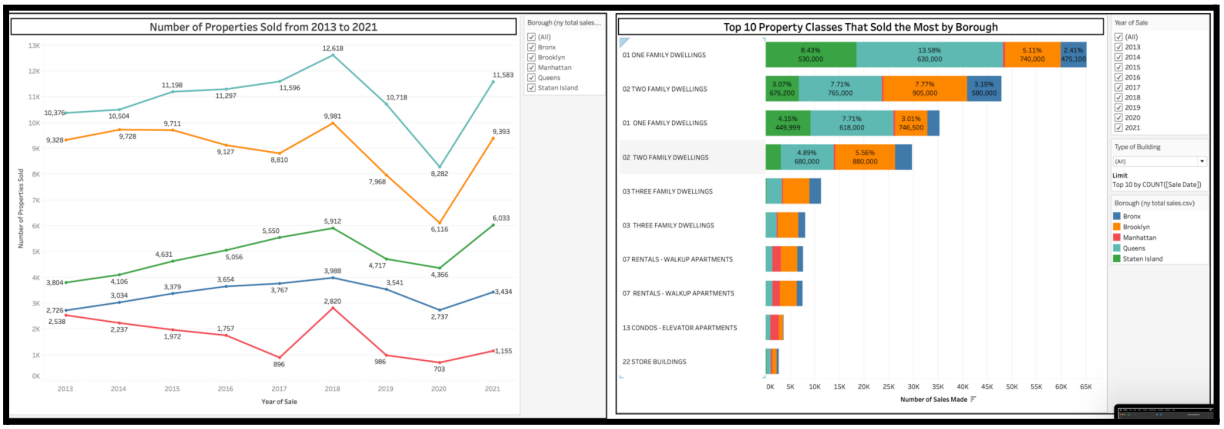
Step 5: Similarly, Manhattan is generally accessible by metro locations from the north, all the way to the south.



Step 6: Now we move on to safety of neighbourhoods. On aggregate, Manhattan ranks fourth relative to the other boroughs. Closer examination visually shows the spread of reported crime is sporadic and not focused in a particular neighbourhood. One improvement that could included here is to show the date of the complaint as well as further description on the nature of the report to provide more information (and assurance) for users.

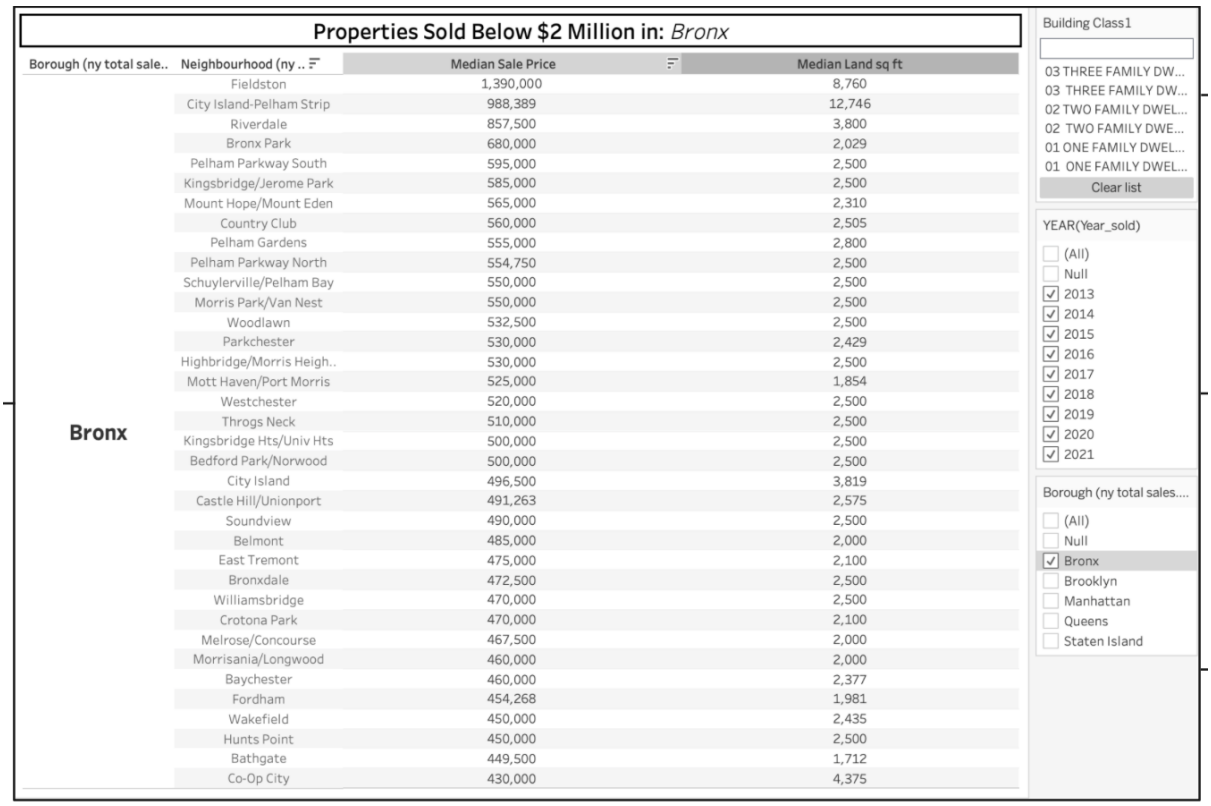


Step 7: Now that we are done accessing rent potential, suitability of location arising from convenient locations to subways and tourist attractions, as well as safety, we can start looking into rental volumes and trends from past years. Here we observe that relative to other boroughs, Manhattan seems to have lower sales volumes. Additionally, for the dwellings that our client desires, these dwellings are in shorter supply relative to other boroughs. We can briefly conclude that housing in Manhattan may be in short supply and high demand:



Step 8: Looking at a bubble chart characterised by gross square feet of the property, land square feet of the property and price sold for the property, we observe that at a price of $2 million dollars, there were not many properties that were sold in that price range in Manhattan. In fact, more options existed in other boroughs for that price range at a larger area (for either land square feet or gross square feet). Hence the preliminary conclusion is that its possible the investor might need to consider other alternative properties in other boroughs in addition to the preference for Manhattan.

Step 9: Finally, we take a look at prospects in other boroughs, split by neighbourhoods, to observe the price comparisons of properties in those locations and their respective areas. In this figure, we look at Bronx and find possible alternatives.However, this has to be re-weighed against the initial preferences of the investor, and if the investor is willing to accept trade-offs of his/her initial preferences in lieu of meeting the budget constraint.



Properties Sold Below $2 Million in: *Bronx*

Borough (ny total sale..	Neighbourhood (ny ..	Median Sale Price	Median Land sq ft
Bronx	Fieldston	1,390,000	8,760
	City Island-Pelham Strip	988,389	12,746
	Riverdale	857,500	3,800
	Bronx Park	680,000	2,029
	Pelham Parkway South	595,000	2,500
	Kingsbridge/Jerome Park	585,000	2,500
	Mount Hope/Mount Eden	565,000	2,310
	Country Club	560,000	2,505
	Pelham Gardens	555,000	2,800
	Pelham Parkway North	554,750	2,500
	Schuylerville/Pelham Bay	550,000	2,500
	Morris Park/Van Nest	550,000	2,500
	Woodlawn	532,500	2,500
	Parkchester	530,000	2,429
	Highbridge/Morris Heigh..	530,000	2,500
	Mott Haven/Port Morris	525,000	1,854
	Westchester	520,000	2,500
	Throgs Neck	510,000	2,500
	Kingsbridge Hts/Univ Hts	500,000	2,500
	Bedford Park/Norwood	500,000	2,500
	City Island	496,500	3,819
	Castle Hill/Unionport	491,263	2,575
	Soundview	490,000	2,500
	Belmont	485,000	2,000
	East Tremont	475,000	2,100
	Bronxdale	472,500	2,500
	Williamsbridge	470,000	2,500
	Crotona Park	470,000	2,100
	Melrose/Concourse	467,500	2,000
	Morrisania/Longwood	460,000	2,000
	Baychester	460,000	2,377
	Fordham	454,268	1,981
	Wakefield	450,000	2,435
	Hunts Point	450,000	2,500
	Bathgate	449,500	1,712
	Co-Op City	430,000	4,375

4. Conclusion

4.1 Recommendations to Investor Client

The final recommendations to the client can be summarised in below, resulting from the visualisations above, can be used to conclude for the client's preferences:

- Budget: $2 million USD for residential property in Manhattan, New York
  - Possible, but the market might be short in supply for Manhattan.
- Accessibility and Leisure:
  - Not an issue in the Manhattan area
- Rental Preferences: Prefer to rent during winter months
  - Rental seasonality fits client's needs
- Other Concerns: Prefer area to be safe from violent crimes or home break-ins
  - Reasonable assurance that crime is not a deterrent for purchase

4.2 My Observations and Lessons Learnt

In summary, I was able to combine the use of Python (for data preparation, pre-processing and feature engineering) for the data preparation phase. Tableau was then used to create a pitch for an investor, by presenting a story of how a typical property hunt could seem like, while using a variety of data sources and data visualisations to build a case and present insights useful for an investor to make a decision. Compared to other tools I have used for

visualization (Altair, Panel, matplotlib), I am definitely more fond of using Tableau as it is much simpler to use, and I can easily generate sophisticated graphs quickly.

There were three lessons that I learned from this project. As in most data science cases, data preparation is the most time consuming aspect of the project. Finding good data sources was a huge exercise in itself. Once data was obtained, data cleaning was also extremely time-consuming. The two processing tasks that took up most of my time were firstly, using geopy to synthesize geolocation data (latitude, longitude, postcode, etc.), and secondly, combining massive datasets using pandas (in the case of the properties sold dataset, there were 5 data sets per yea from 2013 to 2021 that I had to individually download, then collectively clean and combine and create new features). This resulted in a dataset of over 700,000 entries - painstaking work!

The second lesson I learned was using entity-relationship schemas to be able to fully benefit from Tableau. With smaller tables, Tableau is already a very versatile tool - but combining multiple datasets was where I discovered Tableau's true potential, which I have definitely not fully mastered yet and will continue to experiment to be more familiar at entity-relationship schemas and using Tableau to visualise multiple data sources across large datasets. Tableau can perform way more in-depth analysis beyond the superficial drag-and-drop methods I had learned on DataCamp.

The final lesson I learned are the array of functionalities within Tableau itself. I have barely scratched the surface of what Tableau is capable of doing; beyond the bread-and-butter bar charts, histograms and line charts, my research on Tableau online has shown me so many other visualisation techniques I have yet to learn. One area i would like to go into now is building user friendly dashboards and publishing them online.

Sources

1. "Inside Airbnb - Get the Data." Inside Airbnb. Retrieved from http://insideairbnb.com/get-the-data/

2. "Tourist Attractions in New York County, United States of America (New York) - OSM Data." mygeodata.cloud. Retrieved from https://mygeodata.cloud/data/download/osm/tourist-attractions/united-states-of-america--new-york/new-york-county

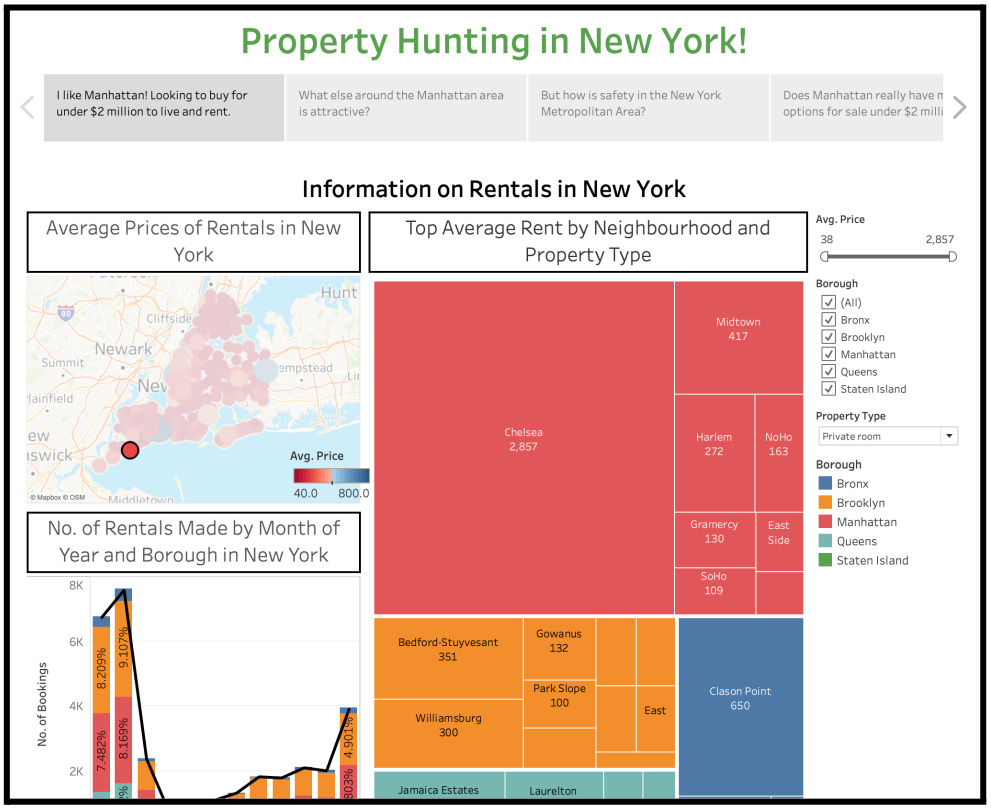
3. "Subway Stations." NYC Open Data. Retrieved from https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

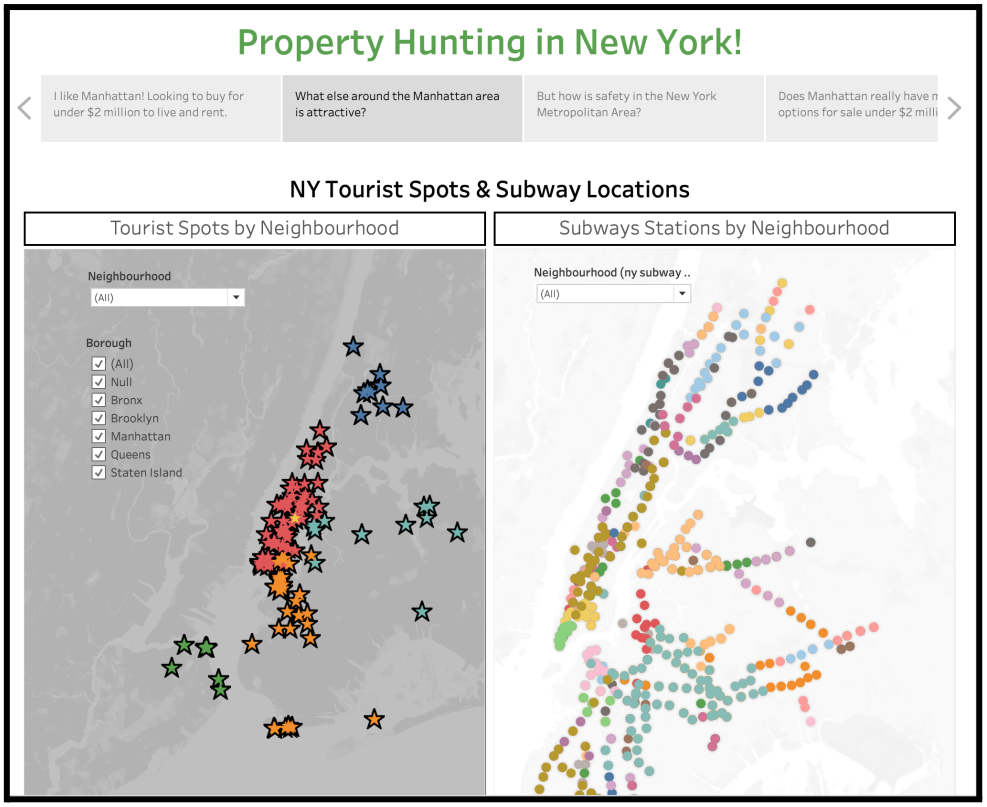
4. "Crime Map." NYC Open Data. Retrieved from https://data.cityofnewyork.us/Public-Safety/Crime-Map-/5jvd-shfj

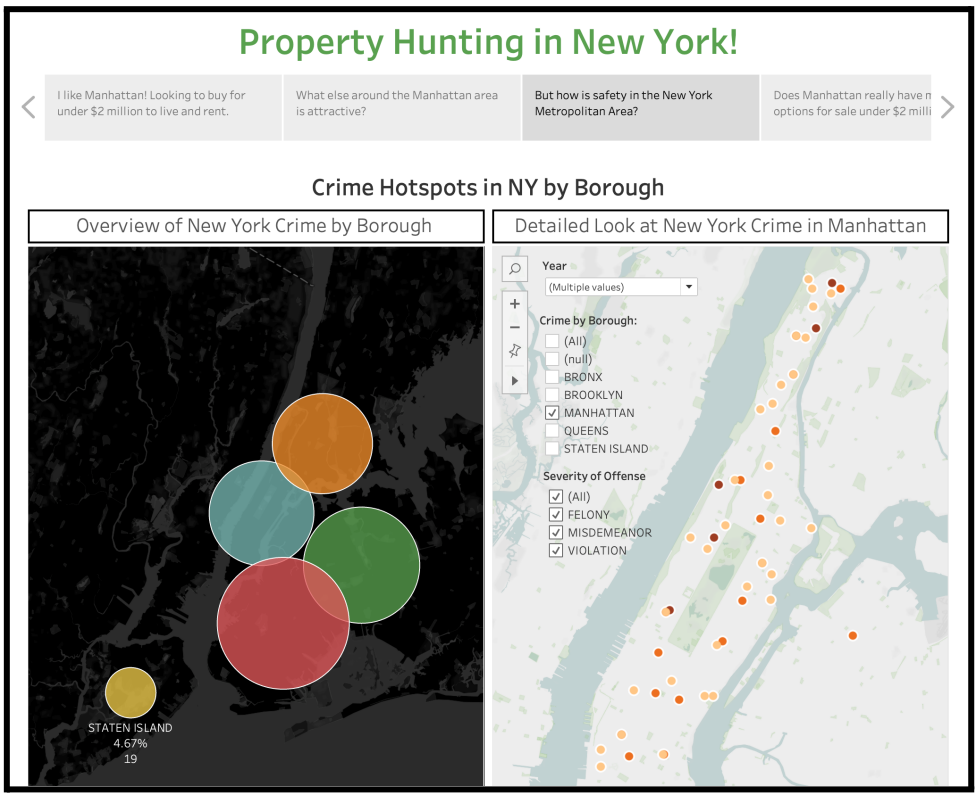
5. "Property Annualized Sales Update." New York City Department of Finance. Retrieved from https://www.nyc.gov/site/finance/taxes/property-annualized-sales-update.page

Appendices:

Tableau Story Screenshots:




Property Hunting in New York!
I like Manhattan! Looking to buy for under $2 million to live and rent.
What else around the Manhattan area is attractive?
But how is safety in the New York Metropolitan Area?
Does Manhattan really have n options for sale under $2 milli
NY Tourist Spots & Subway Locations
Tourist Spots by Neighbourhood
Subways Stations by Neighbourhood
Neighbourhood
(All)
Borough
(All)
Null
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Neighbourhood (ny subway ..
(All)

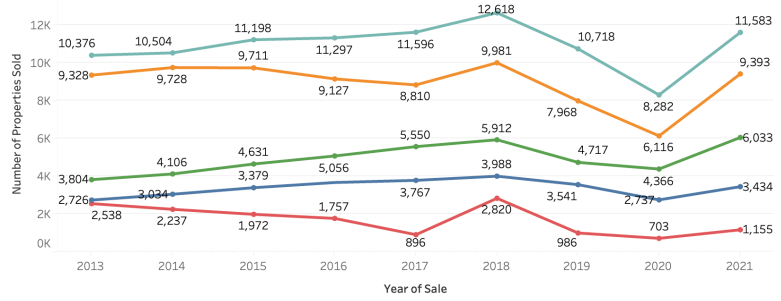

Property Hunting in New York!
I like Manhattan! Looking to buy for under $2 million to live and rent.
What else around the Manhattan area is attractive?
But how is safety in the New York Metropolitan Area?
Does Manhattan really have n options for sale under $2 milli
Crime Hotspots in NY by Borough
Overview of New York Crime by Borough
Detailed Look at New York Crime in Manhattan
STATEN ISLAND
4.67%
19
Year
(Multiple values)
Crime by Borough:
(All)
(null)
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
Severity of Offense
(All)
FELONY
MISDEMEANOR
VIOLATION

Property Hunting in New York!

looking to buy for live and rent. | What else around the Manhattan area is attractive? | But how is safety in the New York Metropolitan Area? | **Does Manhattan really have many options for sale under $2 million** | Closin at alte less th

Trends of NY Properties Sold

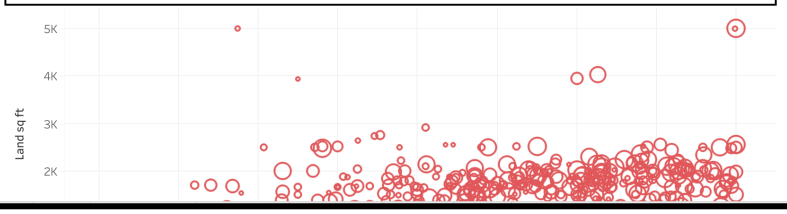
Number of Properties Sold from 2013 to 2021



Borough (ny total sales.c..
- [x] (All)
- [x] Bronx
- [x] Brooklyn
- [x] Manhattan
- [x] Queens
- [x] Staten Island

Year of Year_sold
- [ ] (All)
- [ ] 2013
- [ ] 2014
- [ ] 2015
- [ ] 2016
- [ ] 2017
- [ ] 2018
- [ ] 2019
- [ ] 2020
- [x] 2021

Year: *2021* and Borough: *Manhattan* of Purchase



Borough (ny total sales.c..
- [ ] (All)
- [ ] Bronx
- [ ] Brooklyn
- [x] Manhattan
- [ ] Queens
- [ ] Staten Island

Property Hunting in New York!

around the Manhattan area ve? | But how is safety in the New York Metropolitan Area? | Does Manhattan really have many options for sale under $2 million | **Closing Thoughts: Maybe we could look at alternatives to Manhattan that sell less than $2 million?**

Borough (ny total sale..	Neighbourhood (ny ..	Median Sale Price	Median Land sq ft
Bronx	Fieldston	1,390,000	8,760
	City Island-Pelham Strip	988,389	12,746
	Riverdale	857,500	3,800
	Bronx Park	680,000	2,029
	Pelham Parkway South	595,000	2,500
	Kingsbridge/Jerome Park	585,000	2,500
	Mount Hope/Mount Eden	565,000	2,310
	Country Club	560,000	2,505
	Pelham Gardens	555,000	2,800
	Pelham Parkway North	554,750	2,500
	Schuylerville/Pelham Bay	550,000	2,500
	Morris Park/Van Nest	550,000	2,500
	Woodlawn	532,500	2,500
	Parkchester	530,000	2,429
	Highbridge/Morris Heigh..	530,000	2,500
	Mott Haven/Port Morris	525,000	1,854
	Westchester	520,000	2,500
	Throgs Neck	510,000	2,500
	Kingsbridge Hts/Univ Hts	500,000	2,500
	Bedford Park/Norwood	500,000	2,500
	City Island	496,500	3,819
	Castle Hill/Unionport	491,263	2,575
	Soundview	490,000	2,500
	Belmont	485,000	2,000
	East Tremont	475,000	2,100
	Bronxdale	472,500	2,500
	Williamsbridge	470,000	2,500
	Crotona Park	470,000	2,100
	Melrose/Concourse	467,500	2,000
	Morrisania/Longwood	460,000	2,000

Building Class1

03 THREE FAMILY DW...
03 THREE FAMILY DW...
02 TWO FAMILY DWEL...
02 TWO FAMILY DWE...
01 ONE FAMILY DWEL...
01 ONE FAMILY DWEL...
Clear list

Year of Year_sold
- [ ] (All)
- [ ] Null
- [x] 2013
- [x] 2014
- [x] 2015
- [x] 2016
- [x] 2017
- [x] 2018
- [x] 2019
- [x] 2020
- [x] 2021

Borough (ny total sales.c..
- [ ] (All)
- [ ] Null
- [x] Bronx
- [ ] Brooklyn
- [ ] Manhattan
- [ ] Queens
- [ ] Staten Island