2. Data acquisition and cleaning

2.1 Data sources

The data utilized in this project comes from three main sources,

- Borough data taken from wikipedia - link
- Crime dataset taken from nydata - link
- NYC sales data taken from kaggle - link
- Foursquare API - Venues for locations

2.1A Borough data

The borough data provides details in regards to the neighbourhood with corresponding boroughs, although the population data that's in the table isn't accurate to current population numbers we dont require those metrics hence it is ignored.

2.1B Crime data

The details of all crime reports in the city allows us to determine the locations where crimes are more frequent, the boroughs that report a high number of crimes are identified with this data and the data fields that are irrelevant are dropped and renamed.

- The crime data is over the time period of 11 months from nov-18' to sept-19'
- The unwanted columns are removed from the dataframe as they add no value to the objective at hand as they mainly comprise of codes of reference
- The columns are renamed to make them more readable to the viewer and descriptions are given below

2.1C NYC housing sales

Here we have a dataset taken from kaggle that gives us property sales from the years 2016 and 2017, we will utilize this to determine the neighbourhoods that support moderate housing costs, the data is considerably large and there are multiple missing values. After removing the rows with multiple Nan’s we also get a filter on the housing cost and the square feet area. We will utilize this to determine the neighbourhoods that support moderate housing costs, and help us identify neighbourhoods that are skyhigh when it comes to rentals and property values.

2.2 Data Cleaning

The common practices for data cleaning are utilized and nan values from important fields are addressed, along with the outliers that are identified and removed.

Community Boards in New York City

Community Board (CB)	Area km^2	Pop. Census 2010	Pop./ km^2	Neighborhoods
Bronx CB 1	7.17	91,497	12,761	Melrose, Mott Haven, Port Morris
Bronx CB 2	5.54	52,246	9,792	Hunts Point, Longwood
Bronx CB 3	4.07	79,762	19,598	Claremont, Concourse Village, Crotona Park, Morrisania
Bronx CB 4	5.28	146,441	27,735	Concourse, Highbridge
Bronx CB 5	3.55	128,200	36,145	Fordham, Morris Heights, Mount Hope, University Heights
Bronx CB 6	4.01	83,268	20,765	Bathgate, Belmont, East Tremont, West Farms
Bronx CB 7	4.84	139,286	28,778	Bedford Park, Norwood, University Heights
Bronx CB 8	8.83	101,731	11,521	Fieldston, Kingsbridge, Kingsbridge Heights, Marble Hill, Riverdale, Spuyten Duyvil, Van Cortlandt Village
Bronx CB 9	12.41	172,298	13,884	Bronx River, Bruckner, Castle Hill, Clason Point, Harding Park, Parkchester, Soundview, Unionport
Bronx CB 10	16.76	120,392	7,183	City Island, Co-op City, Locust Point, Pelham Bay, Silver Beach, Throgs Neck, Westchester Square
Bronx CB 11	9.32	113,232	12,149	Allerton, Bronxdale, Indian Village, Laconia, Morris Park, Pelham Gardens, Pelham Parkway, Van Nest
Bronx CB 12	14.56	152,344	10,463	Baychester, Edenwald, Eastchester, Fish Bay, Olinville, Wakefield, Williamsbridge, Woodlawn
Brooklyn CB 1	12.82	160,338	12,507	Greenpoint, Williamsburg, Williamsburg Houses
Brooklyn CB 2	7.72	98,620	12,775	Boerum Hill, Brooklyn Heights, Brooklyn Navy Yard, Clinton Hill, Dumbo, Fort Greene, Fulton Ferry, Fulton Mall, Vinegar Hill
Brooklyn CB 3	7.67	143,867	18,757	Bedford-Stuyvesant, Ocean Hill, Stuyvesant Heights
Brooklyn CB 4	5.31	104,358	19,653	Bushwick
Brooklyn CB 5	14.61	173,198	11,855	City Line, Cypress Hills, East New York, Highland Park, New Lots, Starrett City
Brooklyn CB 6	9.01	104,054	11,549	Carroll Gardens, Cobble Hill, Gowanus, Park Slope, Red Hook
Brooklyn CB 7	10.96	120,063	10,955	Greenwood Heights, Sunset Park, Windsor Terrace
Brooklyn CB 8	4.25	96,076	22,606	Crown Heights, Prospect Heights, Weeksville
Brooklyn CB 9	4.07	104,014	25,556	Crown Heights, Prospect Lefferts Gardens, Wingate
Brooklyn CB 10	10.57	122,542	11,593	Bay Ridge, Dyker Heights, Fort Hamilton

The data taken from wikipedia that contains a table containing information regarding the neighbourhoods and boroughs in new york city and the the other details aren't accurate and are removed

The data is formatted into a pandas dataframe and the community board column is transformed into another field only containing the borough names.The neighbourhoods are separated into separate values and iterated to each borough, The other columns are not required and are dropped.

The table is then parsed to get relevant details with the geopy library to get location data and add that into the dataframe for each corresponding neighbourhood and borough.

	search_name	lat	lon	importance	type	Neighbourhood	Borough	boundingbox	class
0	Melrose Bronx	40.8256703	-73.9152416	0.536855	station	Melrose	Bronx	[40.8206703, 40.8306703, -73.9202416, -73.9102...	railway
1	Mott Haven Bronx	40.8089897	-73.9229147	0.623710	neighbourhood	Mott Haven	Bronx	[40.8089397, 40.8090397, -73.9229647, -73.9228...	place
2	Port Morris Bronx	40.8015147	-73.9095811	0.606728	neighbourhood	Port Morris	Bronx	[40.8014647, 40.8015647, -73.9096311, -73.9095...	place
3	Hunts Point Bronx	40.8126008	-73.8840247	0.673005	neighbourhood	Hunts Point	Bronx	[40.8125508, 40.8126508, -73.8840747, -73.8839...	place
4	Longwood Bronx	40.8162916	-73.8962205	0.532702	station	Longwood	Bronx	[40.8112916, 40.8212916, -73.9012205, -73.8912...	railway
5	Claremont Bronx	40.8341667	-73.9102778	0.350000	park	Claremont	Bronx	[40.8341167, 40.8342167, -73.9103278, -73.9102...	leisure
6	Concourse Village Bronx	40.823868149999996	-73.92118091218828	0.500000	residential	Concourse Village	Bronx	[40.8227742, 40.8249622, -73.9229217, -73.9194...	landuse
7	Crotona Park Bronx	40.838901899999996	-73.89386451155042	0.644356	park	Crotona Park	Bronx	[40.8344186, 40.8434525, -73.9011324, -73.886757]	leisure
8	Morrisania Bronx	40.8292672	-73.9065253	0.540718	neighbourhood	Morrisania	Bronx	[40.8292172, 40.8293172, -73.9065753, -73.9064...	place
9	Concourse Bronx	40.8185618	-73.927303	0.589675	station	Concourse	Bronx	[40.8135618, 40.8235618, -73.932303, -73.922303]	railway
10	Highbridge Bronx	40.83653245	-73.92959563928645	0.400000	residential	Highbridge	Bronx	[40.8351109, 40.8387844, -73.9306484, -73.9280...	landuse
11	Fordham Bronx	40.8614754	-73.8905439	0.545961	station	Fordham	Bronx	[40.8564754, 40.8664754, -73.8955439, -73.8855...	railway
12	Morris Heights Bronx	40.8498223	-73.919859	0.631291	neighbourhood	Morris Heights	Bronx	[40.8497723, 40.8498723, -73.919909, -73.919809]	place

Crime data

The cleaning is done by removing a lot of data points that aren't required, the data also only stores information regarding the boroughs that correspond to each crime that was committed and the total data points are more than eighty thousand, however there are multiple Nan values and the location data although does contain latitudinal and longitudinal fields lacks over more than 80 percent of its data, hence it cannot be utilized to obtain the neighbourhoods that correspond to the crime incidents.

We will for the sake of interpretation be removing multiple fields that are deemed unnecessary for the scope of our project as they aren't relevant in determining our ultimate goal which is to determine the safety of locations in the city of new york.

CMPLNT_NUM - removed
CMPLNT_FR_DT - complaintdate
CMPLNT_FR_TM - complainttime
CMPLNT_TO_DT - complaintlastdate
CMPLNT_TO_TM - complaintlasttime
ADDR_PCT_CD - removed
RPT_DT - reporteddate
KY_CD - removed
OFNS_DESC - offencedescription
PD_CD - removed
PD_DESC - removed
CRM_ATPT_CPTD_CD - status
LAW_CAT_CD - crimecategory
BORO_NM - borough
LOC_OF_OCCUR_DESC - crimelocation
PREM_TYP_DESC - premisestype
JURIS_DESC - removed
JURISDICTION_CODE - removed
PARKS_NM - removed

HADEVELOPT - removed
HOUSING_PSA - removed
X_COORD_CD - removed
Y_COORD_CD - removed
SUSP_AGE_GROUP - suspectage
SUSP_RACE - suspectrace
SUSP_SEX - suspectsex
TRANSIT_DISTRICT - removed
Latitude - removed
Longitude - removed
Lat_Lon - location
PATROL_BORO - removed
STATION_NAME - removed
VIC_AGE_GROUP - victimeage
VIC_RACE - victimerace
VIC_SEX - victimesex

The data frame is cleaned and the significant outliers are removed.

	complaintdate	complainttime	complaintlastdate	complaintlasttime	reporteddate	offencedescription	status	crimecategory	borough	crimelocation	premisestype	suspectage	suspectrace	suspectsex	
0	11/30/2018	22:00:00	01/07/2019	17:00:00	01/09/2019	PETIT LARCENY	COMPLETED	MISDEMEANOR	BROOKLYN	INSIDE	RESIDENCE - APT. HOUSE	UNKNOWN	UNKNOWN	F	(40.65775788 -73.95177405
1	11/30/2018	06:50:00	NaN	NaN	01/06/2019	HARRASSMENT 2	COMPLETED	VIOLATION	QUEENS	INSIDE	RESIDENCE-HOUSE	25-44	BLACK	M	(40.70821893 -73.73603388
2	11/30/2018	13:15:00	01/10/2019	16:00:00	01/10/2019	PETIT LARCENY	COMPLETED	MISDEMEANOR	QUEENS	INSIDE	STREET	25-44	BLACK HISPANIC	M	(40.69959352 -73.89406972
3	11/30/2018	11:00:00	12/05/2018	09:00:00	01/06/2019	THEFT-FRAUD	COMPLETED	FELONY	BROOKLYN	INSIDE	RESIDENCE - APT. HOUSE	UNKNOWN	UNKNOWN	U	(40.60864137 -73.99048623
4	11/30/2018	00:01:00	11/30/2018	23:59:00	01/08/2019	HARRASSMENT 2	COMPLETED	VIOLATION	STATEN ISLAND	FRONT OF	RESIDENCE-HOUSE	25-44	BLACK	F	(40.64135410 -74.09069605

.
The data does contain a significant number of missing values but they are not in the important fields that we need to use for our objective.

Sales data
The boroughs are marked as binned values, using the information provided they are renamed to their names.The sale prices have a few outliers that need to be removed to have the sales price ranges more practical and within the scope of what we need to be searching for hence its set from 10000 to 1000000 dollars.The same is applied to the land square feet field that sets the range filter to a realistic housing area from 100 to10000 square feet.

	Borough	Neighborhood	Building Class Category	Tax Class At Present	Block	Lot	Building Class At Present	Address	Apartment Number	Zip Code	Residential Units	Commercial Units	Total Units	Land Square Feet	Gross Square Feet	Year Built	Tax Class At Time Of Sale	Building Class At Time Of Sale	Sale Price	Sale Date
0	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	392	6	C2	153 AVENUE B		10009	5	0	5	1633	6440	1900	2	C2	6625000	2017-07-19 00:00:00
1	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	26	C7	234 EAST 4TH STREET		10009	28	3	31	4616	18690	1900	2	C7	-	2016-12-14 00:00:00
2	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	39	C7	197 EAST 3RD STREET		10009	16	1	17	2212	7803	1900	2	C7	-	2016-12-09 00:00:00
3	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	402	21	C4	154 EAST 7TH STREET		10009	10	0	10	2272	6794	1913	2	C4	3936272	2016-09-23 00:00:00
4	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	404	55	C2	301 EAST 10TH STREET		10009	6	0	6	2369	4615	1900	2	C2	8000000	2016-11-17 00:00:00