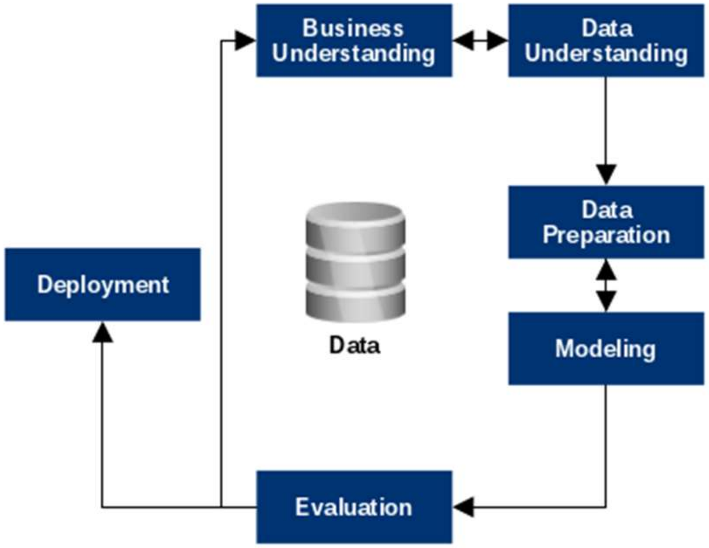
Data Mining Project: Insight of NYC districts regarding the tobacco licensing system for soon to be tobacco business owner, Josh.

Alvin Munar
Katherine Gallego

Data Mining Process

- With the help of CRISP-DM, we can easily implement a process of steps for data mining and analysis.
- CRISP-DM (Cross-Industry Standard Process for Data Mining) is a detailed model consisting of 6 steps which include:
- 1. Business Understanding
- 2. Data understanding
- 3. Data Preparation
- 4. Modeling
- 5. Evaluation
- 6. Deployment



Business Problem

➡ Josh is a prospective business owner and wants to open shop. He is uncertain that opening up a shop in a district where there is a high volume of tobacco retailers will have more competition and become less profitable. Should Josh decide to open a shop in a district where there is less volume of smoke shops?

➡ Solution: Josh will rely on a dataset to see which area has the less amount of smoke shops so he has less competition.

Business Understanding

- We need a dataset to understand the objectives for the project and understand the intended goals. From there, we can determine what the problem is going to be.
- How will data mining answer the business problem?
  - Data mining solution will give us a better perspective of how to approach a business problem using a data set.
- The purpose of this project is to provide an overview/insight of a tobacco licensing system that includes the quantity, location and type of tobacco product while complying to government regulation.
- We want to dissect whether or not it will be beneficial to open a smoke shop in a district with a higher concentration of tobacco retail shops vs. a lower concentration of tobacco retail shops.

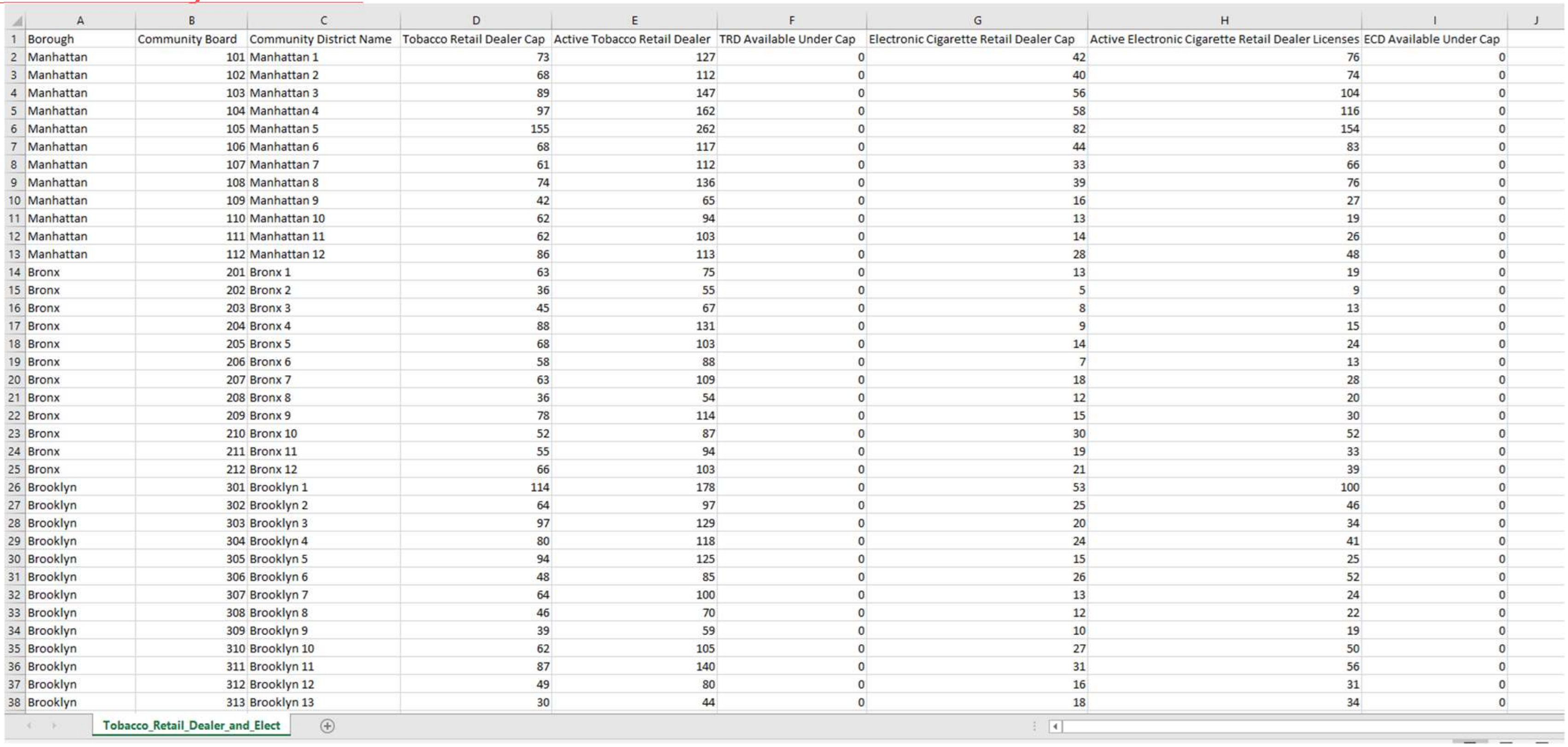
Business Understanding cont.

- We will utilize data from NYC open data from 2020 in order to construct Machine Learning Models (Supervised data mining). The models will denote which district will best predict where Josh can open his smoke shop.
- How (precisely) will a data mining solution address the business problem?
    - Among the districts, where is it likely that Josh will open up his smoke shop?
    - Will using regression to estimate or predict whether Josh will be more susceptible in opening his smoke shop business in a district where there is more or less of a presence in smoke shop businesses?
- Similarly, we believe that by implementing clustering methods might help denote concentrations of tobacco localities and can help Josh to make a suitable choice on where to put his smoke shop.

Data Analysis

- We found a reliable data from *NYC Open Data 2020.* (NYCOpenData)
- We will use the data aggregated from the NYC Open Data to find a solution on the business problem and see if data mining will help prove the answer to the solution.
- Similarly, we can use this community district map to locate each tobacco licensed retail shop. (Craig Newark Graduate School of Journalism) (NYC Planning | Community Profiles)
- We obtained a total of number of Tobacco Retail Dealer licenses allowed throughout the districts of NYC (n=3879) known as supervised data.

Tobacco Retail Dealer and Electronic Cigarette Retail Dealer Caps by Community District



	A	B	C	D	E	F	G	H	I	J
1	Borough	Community Board	Community District Name	Tobacco Retail Dealer Cap	Active Tobacco Retail Dealer	TRD Available Under Cap	Electronic Cigarette Retail Dealer Cap	Active Electronic Cigarette Retail Dealer Licenses	ECD Available Under Cap	
2	Manhattan	101	Manhattan 1	73	127	0	42	76	0	
3	Manhattan	102	Manhattan 2	68	112	0	40	74	0	
4	Manhattan	103	Manhattan 3	89	147	0	56	104	0	
5	Manhattan	104	Manhattan 4	97	162	0	58	116	0	
6	Manhattan	105	Manhattan 5	155	262	0	82	154	0	
7	Manhattan	106	Manhattan 6	68	117	0	44	83	0	
8	Manhattan	107	Manhattan 7	61	112	0	33	66	0	
9	Manhattan	108	Manhattan 8	74	136	0	39	76	0	
10	Manhattan	109	Manhattan 9	42	65	0	16	27	0	
11	Manhattan	110	Manhattan 10	62	94	0	13	19	0	
12	Manhattan	111	Manhattan 11	62	103	0	14	26	0	
13	Manhattan	112	Manhattan 12	86	113	0	28	48	0	
14	Bronx	201	Bronx 1	63	75	0	13	19	0	
15	Bronx	202	Bronx 2	36	55	0	5	9	0	
16	Bronx	203	Bronx 3	45	67	0	8	13	0	
17	Bronx	204	Bronx 4	88	131	0	9	15	0	
18	Bronx	205	Bronx 5	68	103	0	14	24	0	
19	Bronx	206	Bronx 6	58	88	0	7	13	0	
20	Bronx	207	Bronx 7	63	109	0	18	28	0	
21	Bronx	208	Bronx 8	36	54	0	12	20	0	
22	Bronx	209	Bronx 9	78	114	0	15	30	0	
23	Bronx	210	Bronx 10	52	87	0	30	52	0	
24	Bronx	211	Bronx 11	55	94	0	19	33	0	
25	Bronx	212	Bronx 12	66	103	0	21	39	0	
26	Brooklyn	301	Brooklyn 1	114	178	0	53	100	0	
27	Brooklyn	302	Brooklyn 2	64	97	0	25	46	0	
28	Brooklyn	303	Brooklyn 3	97	129	0	20	34	0	
29	Brooklyn	304	Brooklyn 4	80	118	0	24	41	0	
30	Brooklyn	305	Brooklyn 5	94	125	0	15	25	0	
31	Brooklyn	306	Brooklyn 6	48	85	0	26	52	0	
32	Brooklyn	307	Brooklyn 7	64	100	0	13	24	0	
33	Brooklyn	308	Brooklyn 8	46	70	0	12	22	0	
34	Brooklyn	309	Brooklyn 9	39	59	0	10	19	0	
35	Brooklyn	310	Brooklyn 10	62	105	0	27	50	0	
36	Brooklyn	311	Brooklyn 11	87	140	0	31	56	0	
37	Brooklyn	312	Brooklyn 12	49	80	0	16	31	0	
38	Brooklyn	313	Brooklyn 13	30	44	0	18	34	0	

Data Understanding *cont.*

- In the data, it shows the attributes such as:
  - Borough: Where the business is located.
  - Community Board: Community District Number where the Business is located.
  - Community District Name
  - Tobacco Retail Dealer Cap: Maximum number of Tobacco Retail Dealer licenses allowed in the Community District. Calculated as 50 percent of the number of active non-pharmacy Tobacco Retail Dealer licenses as of the cutoff date (February 24, 2018).
  - Active Tobacco Retail Dealer Licenses: Number of active Tobacco Retail Dealer licenses in the Community District as of the last run date of the report, minus all Tobacco Retail Dealer licenses held by pharmacies.

Data Understanding *cont.*

- TRD (Tobacco Retail Dealer) Under Cap
- Electronic Cigarette Retail Dealer Cap: Maximum number of Electronic Cigarette Retail Dealer licenses assigned to the Community District. Calculated as 50 percent of the number of active Electronic Cigarette Retail Dealer licenses as of the cutoff date (August 23, 2018).
- Active Electronic Cigarette Retail Dealer Licenses: Number of active Electronic Cigarette Retail Dealer licenses in the Community District as of the last run date of the report.
- ECD Available Under Cap

Data Preparation

- Data will be used from NYC Open Data:
  - Tobacco Retail Dealer and Electronic Cigarette Retail Dealer Caps by Community District
  - We will concentrate on the Sum of Active Tobacco Retail Dealer Licenses and the Sum of Tobacco Retail Dealer Cap.
- Based on the dataset, we used a pivot table to show which borough, then district has the less densely populated smoke shop.
- The pivot table provided better insight of the aggregated data.
- The Bronx consists of districts 1-10, the following districts with neighborhoods are listed:
  - Bronx District 1: Melrose, Mott Haven, Port Morris
  - Bronx District 2: Hunts Point, Longwood
  - Bronx District 3: Claremont, Crotona Park East, Melrose, Morrisania
  - Bronx District 4: Concourse, Concourse Village, East Concourse, HIghbridge, Mount Eden
  - Bronx District 5: Fordham, Morris Heights, Mount Hope, University Heights

Data Preparation cont.

- The Bronx consists of districts 1-12, the following districts with neighborhoods are listed:
  - Bronx District 6: Bethgate, Belmont, Bronx Park South, East Tremont, West Farms
  - Bronx District 7: Bedford Park, Fordham, Kingsbridge Heights, Norwood, University Heights
  - Bronx District 8: Fieldston, Kingsbridge, Marble Hill, North Riverdale, Riverdale, Spuyten Duyvil
  - Bronx District 9: Bronx River, Castle Hill, Clason Point, Harding Park, Parkchester, Soundview, Soundview-Bruckner, Unionport
  - Bronx District 10: City Island, Co-op City, Country Club, Edgewater Park, Pelham Bay, Schuylerville, Throgs Neck, Westchester Square
  - Bronx District 11: Allerton, Bronxdale, Indian Village, Morris Park, Pelham Gardens, Pelham Parkway, Van Nest
  - Bronx District 12: Baychester, Eastchester, Edenwald, Olinville, Wakefield, Williamsbridge, Woodlawn.

Data Preparation

- Manhattan consists of districts 1-12, the following districts with neighborhoods are listed below:
    - Manhattan District 1: Battery Park City, Civic Center, Ellis Island, Governors Island, Liberty Island, South Street Seaport, Tribeca, Wall Street, World Trade Center
    - Manhattan District 2: Greenwich Village, Hudson Square, Little Italy, NoHo, SoHo, South Village, West VIllage
    - Manhattan District 3: Claremont, Crotona Park East, Melrose, Morrisania
    - Manhattan District 4: Chelsea, Clinton, Hudson Yards
    - Manhattan District 5: Flatiron, Gramercy Park, Herald Square, Midtown, Midtown South, Murray Hills, Times Square, Union Square
    - Manhattan District 6: Beekman Place, Gramercy Park, Murray Hill, Peter Cooper Village, Stuyvesant Town, Sutton Place, Tudor City, Turtle Bay
    - Manhattan District 7: Lincoln Square, Manhattan Valley, Upper West Side
    - Manhattan District 8: Carnegie Hill, Lenox Hill, Roosevelt Island, Upper East Side, Yorkville
    - Manhattan District 9: Hamilton Heights, Manhattanville, Morningside Heights, West Harlem
    - Manhattan District 10: Central Harlem
    - Manhattan District 11: East Harlem, Harlem, Randall’s Island Park, Wards Island Park
    - Manhattan District 12: Inwood, Washington Heights

Data Preparation

- Brooklyn consists of districts 1-18, the following districts with neighborhoods are listed below:
  - Brooklyn District 1: East Williamsburg, Greenpoint, Northside, Southside, Williamsburg
  - Brooklyn District 2: Boerum Hill, Brooklyn Heights, Clinton Hill, Downtown Brooklyn, DUMBO, Fort Greene, Fulton Ferry, Navy Yard, Vinegar Hill
  - Brooklyn District 3: Bedford-Stuyvesant, Stuyvesant Heights, Tompkins Park North
  - Brooklyn District 4: Bushwick
  - Brooklyn District 5: Broadway Junction, City Line, Cypress Hills, East New York, Highland Park, New Lots, Spring Creek, Starrett City
  - Brooklyn District 6: Carroll Gardens, Cobble Hill, Columbia St, Gowanus, Park Slope, Red Hook
  - Brooklyn District 7: Sunset Park, Windsor Terrace
  - Brooklyn District 8: Crown Heights, Prospect Heights, Weeksville
  - Brooklyn District 9: Crown Heights South, Prospect Lefferts Garden, Wingate
  - Brooklyn District 10: Bay Ridge, Dyker Heights, Fort Hamilton
  - Brooklyn District 11: Bath Beach, Bensonhurst, Gravesend, Mapleton
  - Brooklyn District 12: Borough Park, Kensington, Ocean Parkway
  - Brooklyn District 13: Brighton Beach, Coney Island, Gravesend, Homecrest, Sea Gate, West Brighton
  - Brooklyn District 14: Ditmas Park, Flatbush, Manhattan Terrace, Midwood, Ocean Parkway, Prospect Park South
  - Brooklyn District 15: Gerritsen Beach, Gravesend, Homecrest, Kings Highway, Manhattan Beach, Plumb Beach, Sheepshead Bay
  - Brooklyn District 16: Broadway Junction, Brownsville, Ocean Hill
  - Brooklyn District 17: East Flatbush, Farragut, Flatbush, Northeast Flatbush, Remsen Village, Rugby, Erasmus
  - Brooklyn District 18: Bergen Beach, Canarsie, Flatlands, Georgetown, Marine Park, Mill Basin, Mill Island, Paerdegat Basin

Data Preparation

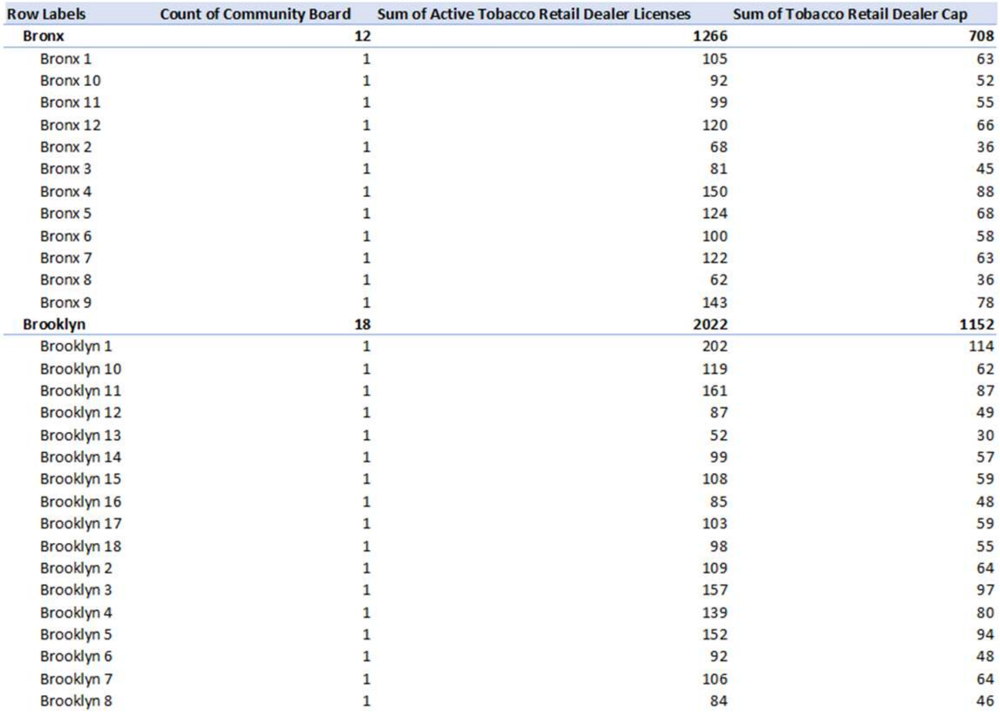
- Queens consists of districts 1-14, the following districts with neighborhoods are listed below:
    - Queens District 1: Astoria, Astoria Heights, Queensbridge, Dutch Kills, Long Island City, Ravenswood, Rikers Island (BX), Steinway
    - Queens District 2: Blissville, Hunters Point, Long Island City, Sunnyside, Sunnyside Gardens, Woodside
    - Queens District 3: East Elmhurst, Jackson Heights, North Corona
    - Queens District 4: Corona, Corona Heights, Elmhurst, Lefrak City
    - Queens District 5: Glendale, Maspeth, Middle Village, Ridgewood
    - Queens District 6:Forest Hills, Forest Hills Gardens, Rego Park
    - Queens District 7: Auburndale, Bay Terrace, Beechhurst, Clearview, College Point, Downtown Flushing, East Flushing, Flushing, Malba, Murray Hill, Queensboro Hill, Waldheim, Whitestone
    - Queens District 8: Briarwood, Fresh Meadows, Hillcrest, Holliswood, Jamaica, Jamaica Estates, Jamaica Hills, Kew Gardens Hills, Pomonok, Utopia
    - Queens District 9: Kew Gardens, Ozone Park, Richmond Hill, Woodhaven
    - Queens District 10: Howard Beach, Lindenwood, Old Howard Beach, Ozone Park, South Ozone Park
    - Queens District 11: Auburndale, Bayside, Douglaston, Hollis Hills, Little Neck, Oakland Gardens
    - Queens District 12: Hollis, Jamaica, Jamaica Center, North Springfield Gardens, Rochdale, South Jamaica, St. Albans
    - Queens District 13: Bellaire, Bellerose, Brookville, Cambria Heights, Floral Park, Glen Oaks, Laurelton, New Hyde Park, Queens Village, Rosedale, Springfield Gardens
    - Queens District 14: Arverne, Bayswater, Belle Harbor, Breezy Point, Broad Channel, Edgemere, Far Rockaway, Hammels, Neponsit, Rockaway Park, The Rockaways, Roxbury, Seaside, Somerville

Data Preparation

- Staten Island consists of districts 1-3, the following districts with neighborhoods are listed below:
  - Staten Island District 1: Arlington, Castleton Corners, Clifton, Elm Park, Fox Hills, Graniteville, Grymes Hill, Howland Hook, Livingston, Mariners Harbor, New Brighton, Old Place, Park Hill, Port Ivory, Port Richmond, Randall Manor, Rosebank, Shore Acres, Silver Lake, St. George, Stapleton, Sunnyside, Tompkinsville, Ward Hill, West Brighton, West New Brighton, Westerleigh, Willowbrook
  - Staten Island District 2: Arrochar, Bloomfield, Bulls Head, Chelsea, Concord, Dongan Hills, Egbertville, Emerson Hill, Grant City, Grasmere, Heartland Village, Lighthouse Hill, Manor Heights, Midland Beach, New Dorp, New Dorp Beach, New Springville, Old Town, South Beach, Todt Hill, Travis, Willowbrook
  - Staten Island District 3: Annadale, Arden Heights, Bay Terrace, Butler Manor, Charleston, Eltingville, Fresh Kills, Great Kills, Greenridge, Huguenot, Oakwood, Oakwood Beach, Oakwood Heights, Pleasant Plains, Prince's Bay, Richmond Town, Richmond Valley, Rossville, Sandy Ground, Tottenville, Woodrow

Data Preparation *cont.*

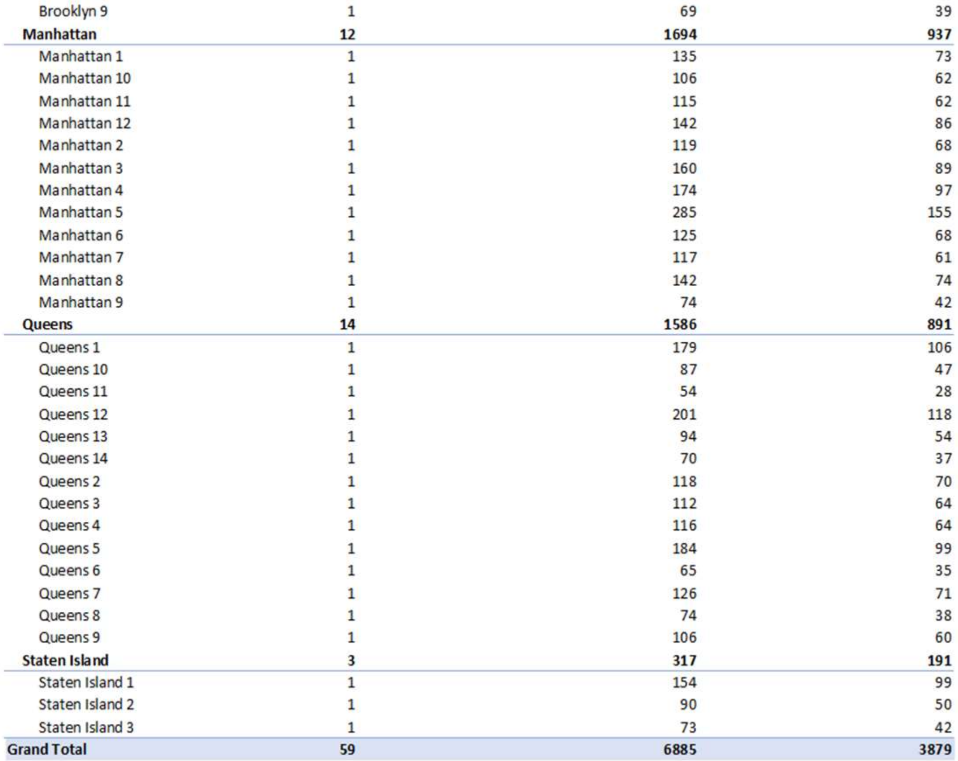
- For this Pivot table model, we focused on the sum of Tobacco Retail Dealer, Cap in relation to the districts for each borough.
- Bronx shows that it has 1266 active tobacco retail dealer cap across its 12 Community Board.
- Brooklyn has 2022 active retail dealer cap across 18 board.



Row Labels	Count of Community Board	Sum of Active Tobacco Retail Dealer Licenses	Sum of Tobacco Retail Dealer Cap
Bronx	**12**	**1266**	**708**
Bronx 1	1	105	63
Bronx 10	1	92	52
Bronx 11	1	99	55
Bronx 12	1	120	66
Bronx 2	1	68	36
Bronx 3	1	81	45
Bronx 4	1	150	88
Bronx 5	1	124	68
Bronx 6	1	100	58
Bronx 7	1	122	63
Bronx 8	1	62	36
Bronx 9	1	143	78
Brooklyn	**18**	**2022**	**1152**
Brooklyn 1	1	202	114
Brooklyn 10	1	119	62
Brooklyn 11	1	161	87
Brooklyn 12	1	87	49
Brooklyn 13	1	52	30
Brooklyn 14	1	99	57
Brooklyn 15	1	108	59
Brooklyn 16	1	85	48
Brooklyn 17	1	103	59
Brooklyn 18	1	98	55
Brooklyn 2	1	109	64
Brooklyn 3	1	157	97
Brooklyn 4	1	139	80
Brooklyn 5	1	152	94
Brooklyn 6	1	92	48
Brooklyn 7	1	106	64
Brooklyn 8	1	84	46

Data Preparation *cont.*

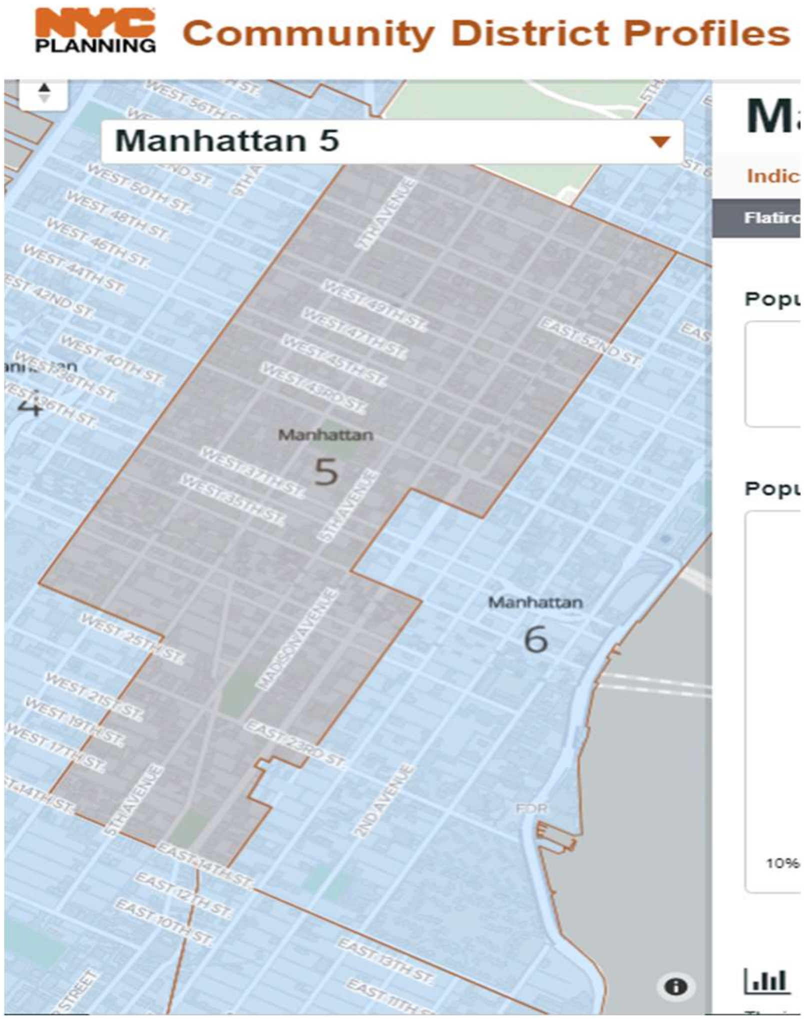
- Manhattan has 12 Community Board and 1694 active tobacco dealer.
- Queens has 14 Community Board and 1586 active dealers.
- Staten Island has 317 active tobacco retail dealer and 3 community board.



Brooklyn 9	1	69	39
Manhattan	**12**	**1694**	**937**
Manhattan 1	1	135	73
Manhattan 10	1	106	62
Manhattan 11	1	115	62
Manhattan 12	1	142	86
Manhattan 2	1	119	68
Manhattan 3	1	160	89
Manhattan 4	1	174	97
Manhattan 5	1	285	155
Manhattan 6	1	125	68
Manhattan 7	1	117	61
Manhattan 8	1	142	74
Manhattan 9	1	74	42
Queens	**14**	**1586**	**891**
Queens 1	1	179	106
Queens 10	1	87	47
Queens 11	1	54	28
Queens 12	1	201	118
Queens 13	1	94	54
Queens 14	1	70	37
Queens 2	1	118	70
Queens 3	1	112	64
Queens 4	1	116	64
Queens 5	1	184	99
Queens 6	1	65	35
Queens 7	1	126	71
Queens 8	1	74	38
Queens 9	1	106	60
Staten Island	**3**	**317**	**191**
Staten Island 1	1	154	99
Staten Island 2	1	90	50
Staten Island 3	1	73	42
Grand Total	**59**	**6885**	**3879**

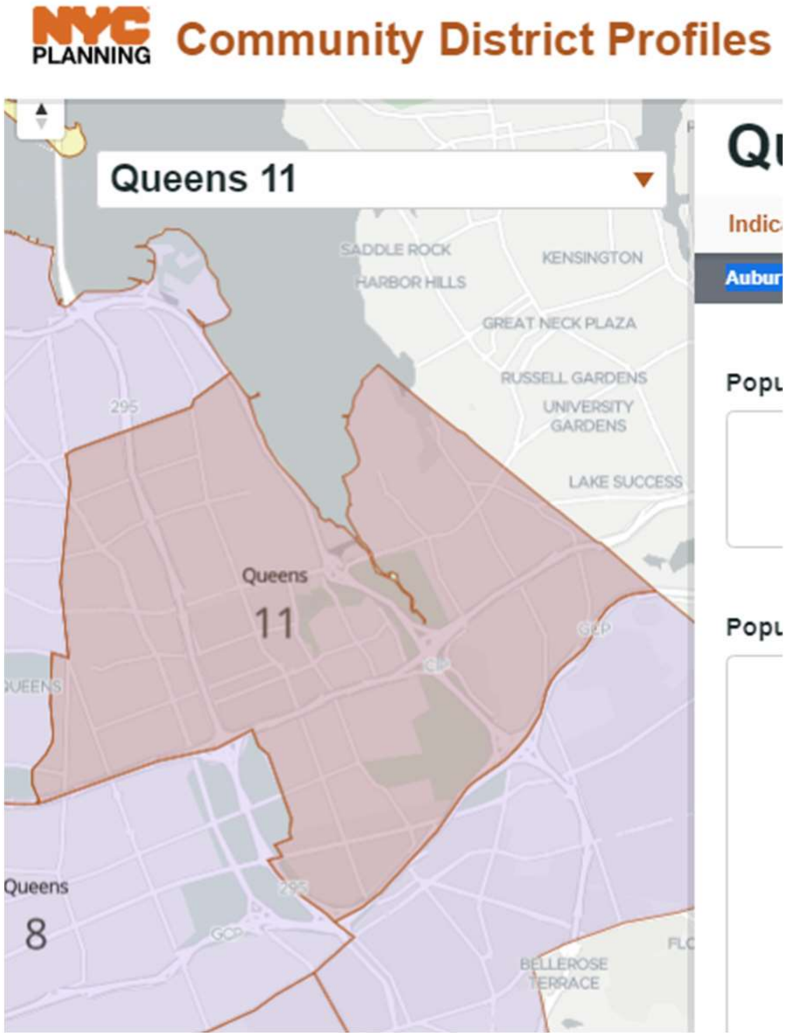
Data Preparation *cont.*

- District 5 of Manhattan, which is all of Flatiron, Gramercy Park, Herald Square, Midtown, Midtown South, Murray Hill, Times Square, Union Square, has 155 retailer tobacco shops.
- There are 285 active active licenses but don't have a shop set up yet will eventually set up a shop.
- District 5 has the highest concentration number of smoke shops out of all the districts and boroughs.
- With the highest concentration number of smoke shops, we conclude that competition is highly visible creating an environment inducing more smoke usage.



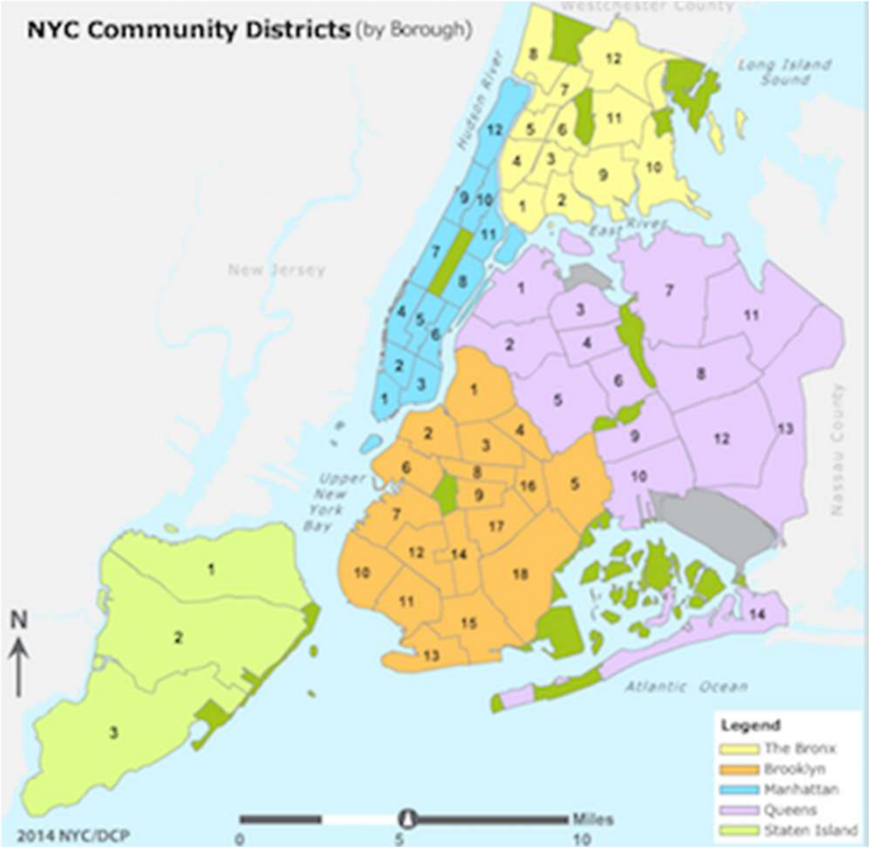
Data Preparation *cont.*

- District 11 of Queens Auburndale, Bayside, Douglaston, Hollis Hills, Little Neck, Oakland Gardens has 28 retailer tobacco shops.
- There are 54 active licenses but don't have a shop set up yet but will eventually set up a shop.
- Similarly, we analyzed that District 11 in Queens had the lowest concentration number of smoke shops.
- With the lowest concentration number of smoke shops, we conclude that competition is low visible creating an environment inducing less smoke usage.



Data Preparation cont.

- These findings may reflect consumer demand.
- We concluded that convenience and proximity play a factor in District 11 of Queens.
- Queens District 11 is considered to be more rural in comparison to District 5 of Manhattan.
- It would be more convenient to set up a store at walking distance for consumers.



Modelling

- Based on the data we gathered using pivot tables to visualize the amount of smoke shop per district.
- We looked at how many district there are in each borough and by using a pivot table we can see how many smoke shops are in each district.
- We can use a supervised data mining technique. With regression to determine the best possible outcome for where Josh can set up shop.
- Clustering algorithms such as K-means clustering, are noted as a distance-based method that create clusters so that the intra-cluster variation is minimized. In each cluster for k-means, it is represented by a centroid which represents a mean of points assigned to the cluster.
- This should solve the business problem because the algorithm will indicate where the best possible outcome should be. It helps determine where the least amount of smokes shop will be.

Modelling *cont.*

- Pros:
  - Regression attempts to estimate the numerical value of the smoke shops per district and cross reference that with the maps of each district to determine the density of the district. (NYC Community District Profiles)
  - The algorithm has data inputs and specific targets and will use the data to predict the outcome.
  - Use Link Prediction to predict the connections between data items such as the smoke shops in the district with the borough.
  - Linear regression will optimize the model to fit the data.
- Cons:
  - An unsupervised data mining technique can use profiling as a data mining task to profile the district to see which one has the most and least smokers. (NYC Community District Profiles)
  - Clustering can cluster all of the smoke shops in each district to a centroid which is the district. Then cluster the district to a centroid which is the borough to determine which borough has the least densely populated borough and district by smoke shops.
  - With clustering it will give the k-means of the value which is the density of the district/borough.
  - Co-occurence the amount of smoke shops with the amount of smokers in the district/borough.

Evaluation

- We used linear regression to estimate or predict whether Josh will be more susceptible in opening his smoke shop business in a district where there is more or less of a presence in smoke shop businesses.
- Similarly, classification technique can also be used to predict which borough and district in a given data to determine the most efficient location for a smoke shop.
- Linear regression algorithm is mostly likely going to be the supervised data mining task because of the the algorithm will optimize to fit the model to the data.
- The regression analysis results can also indicate that districts with a high number of distribution in active tobacco retail dealer licenses may have a correlation to the number of active smokers for each district.
- Our goal for this study is to determine whether the distribution of tobacco licenses and the amount of active retail tobacco dealer will denote where the more profitable district will be in terms of how many active tobacco retailers are in each district?

Deployment

- The results will indicate which borough then district has the least density of smoke shops.
- Josh will then use that results to help make his decision on where he should set up his smoke shop.
- There is no cyber ethical issues. No data is being mined from individuals or businesses regarding personal or important information.
- Aside from cyber moral issues. There are health issues. This data is useful in terms of limiting the amount of tobacco retailers products throughout NYC districts. NYC is implementing strategies to reduce where and by whom tobacco products are sold.

References

- https://communityprofiles.planning.nyc.gov/
- https://researchguides.journalism.cuny.edu/NYCResearch
- https://data.cityofnewyork.us/Business/Tobacco-Retail-Dealer-and-Electronic-Cigarette-Ret/ymyu-3dbp/data