STARBUCKS LOCATION ANALYSIS

Coursera Capstone Project

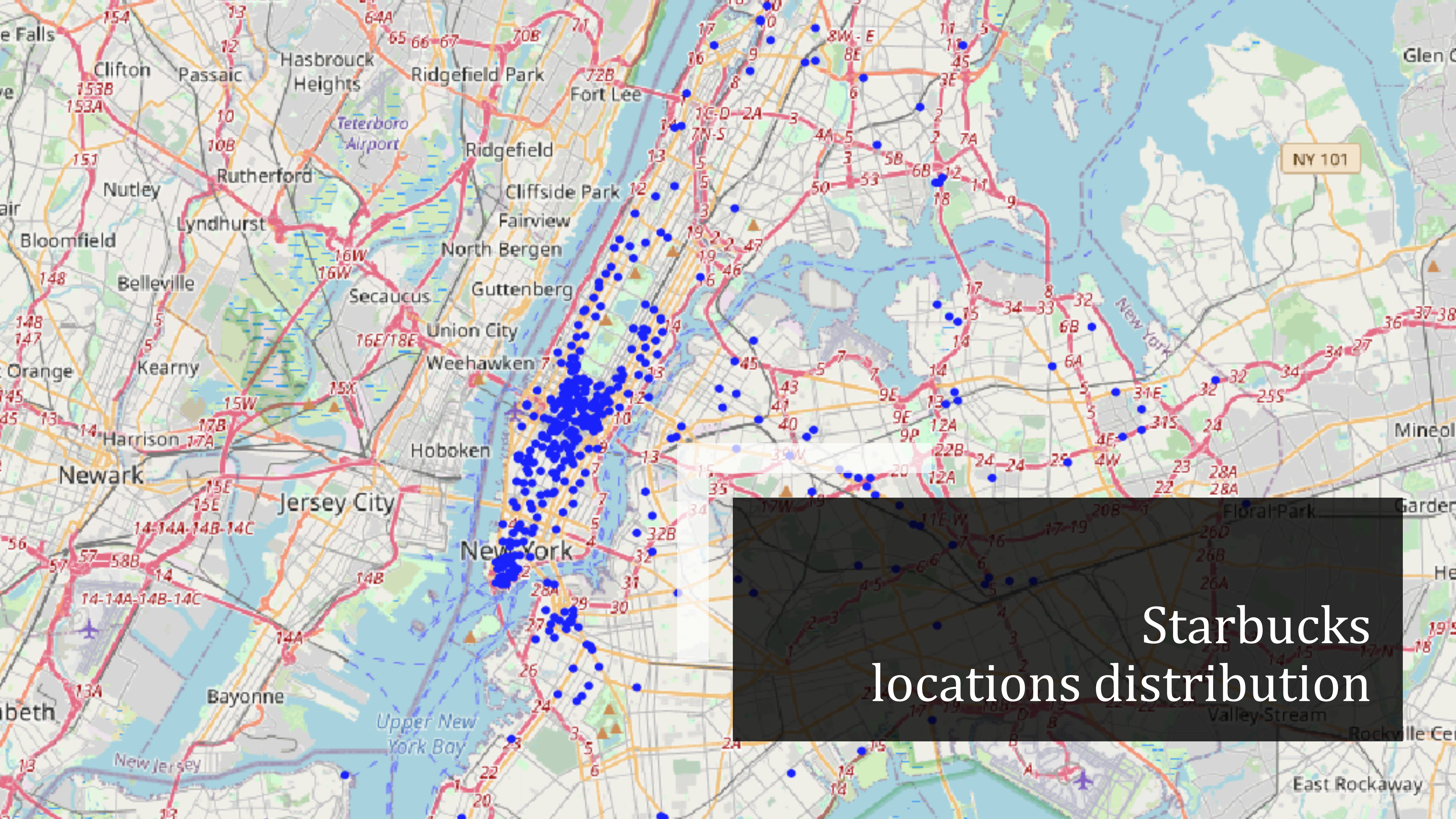
Leping Xu

1. Introduction

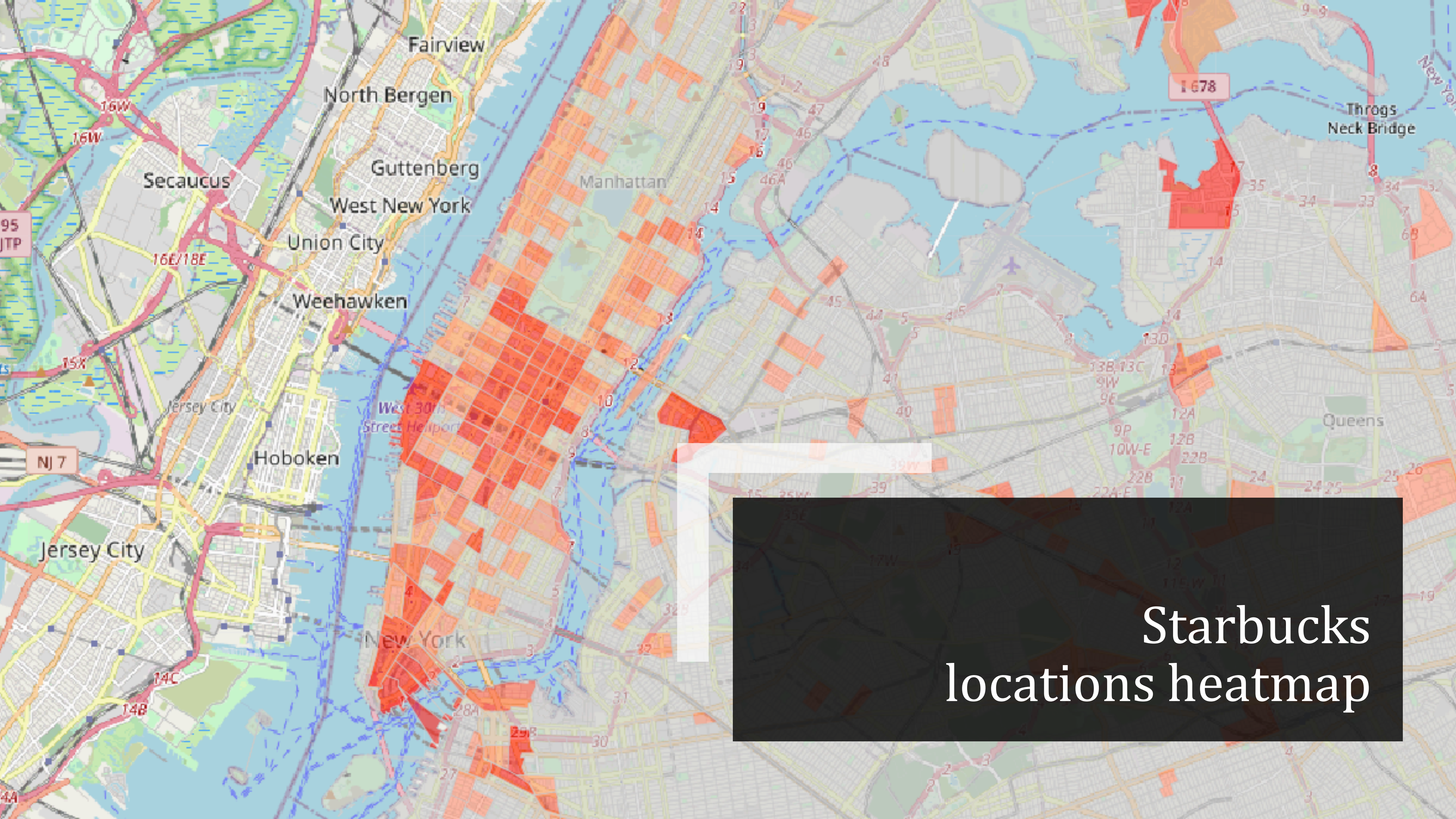
- “How does Starbucks always choose such a popular location?”
- Wstudy all Starbucks locations in five boroughs of New York and try to figure out the similarities among these locations based on the venues within those locations' proximity.
- We will use Foursquare API to retrieve location data, define a "Starbucks proximity score" for a given location, and implement machine learning techniques to score any given location.
- The score will provide guidance for new Starbucks store location selection, but more importantly, for aspiring individual investors to open similar stores, for example, a fast food store or a tea house.

2. Data Preparation and Pre-Processing

- We work with New York City Census Data to divide the city into small fractions
- Use Foursquare API to extract location data. Then we use explore function to get all venues within a certain distance of each tract.
- Search through *Loopnet.com*, the go-to website for commercial property search, to find available spaces.



Starbucks locations distribution



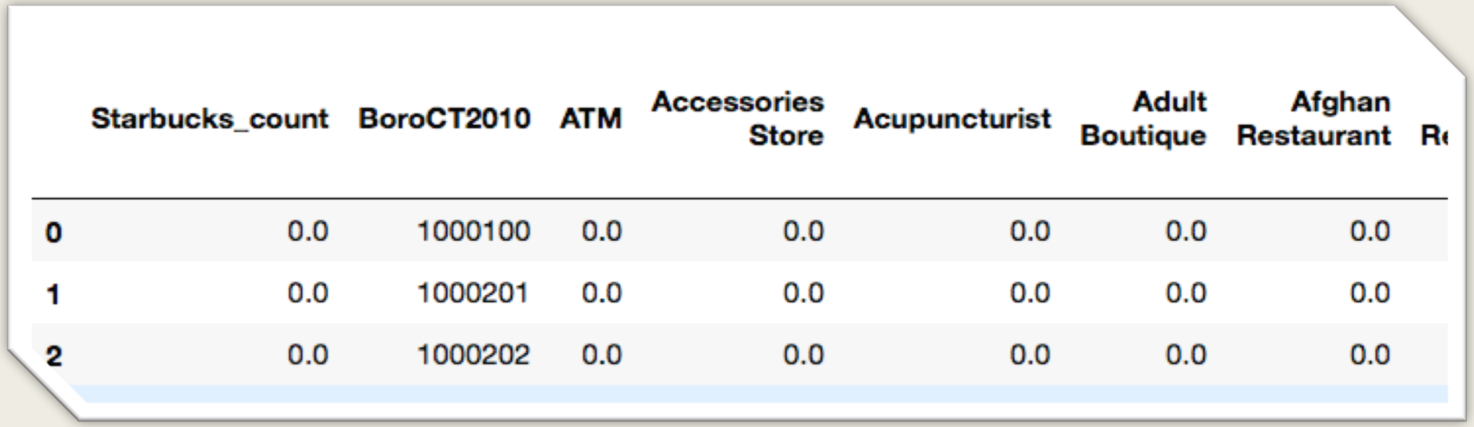
Starbucks locations heatmap

2. Data Preparation and Pre-Processing (contd.)

- Get venues data for each tract to construct features variables (X’s)
- Transform the data using one-hot encoding
- Group by each tract to get each tract’s venues distribution
- “Starbucks_count” column is y label as and columns 2 through end as features (X’s):

2. Data Preparation and Pre-Processing (contd.)

- Split data into training-testing sets for cross validation
- Solve the class imbalance problem by repeating data with label greater than 0 multiple times
- Final training set has 6,668 observations and 538 features.



	Starbucks_count	BoroCT2010	ATM	Accessories Store	Acupuncturist	Adult Boutique	Afghan Restaurant	Re
0	0.0	1000100	0.0	0.0	0.0	0.0	0.0	
1	0.0	1000201	0.0	0.0	0.0	0.0	0.0	
2	0.0	1000202	0.0	0.0	0.0	0.0	0.0	

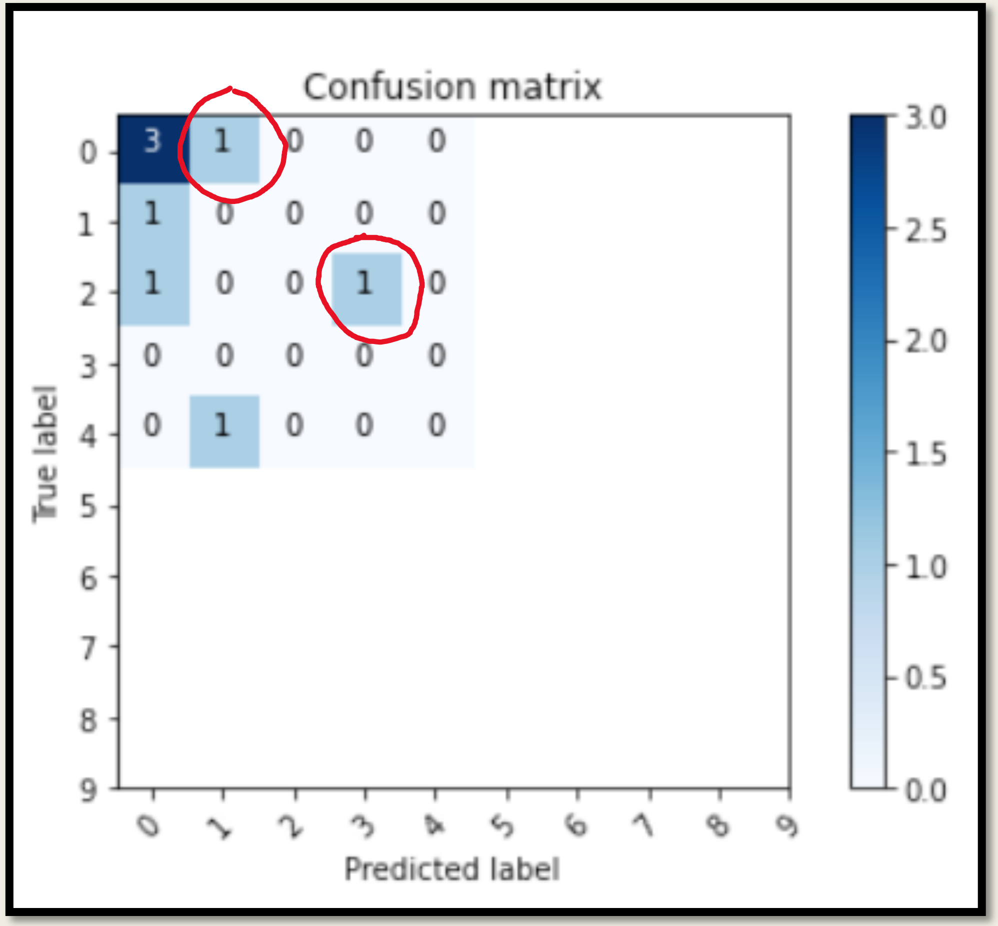
3. Modeling / Methodology

- After trying different classification models, we landed on a single layer neural network due to performance.
- Other models compared includes K nearest neighbours, random forest, adaboost, etc.
- Final model:

  -
    ```
    clf = MLPClassifier(solver='adam', alpha=1e-3,
    learning_rate = 'adaptive', max_iter= 5000,
    hidden_layer_sizes=(16), random_state=0)
    ```
- For hyperparameters tuning, used cross-validation to split the data into 5 batches and compute validation score for a full cross-validation.

3. Modeling / Methodology (Contd.)

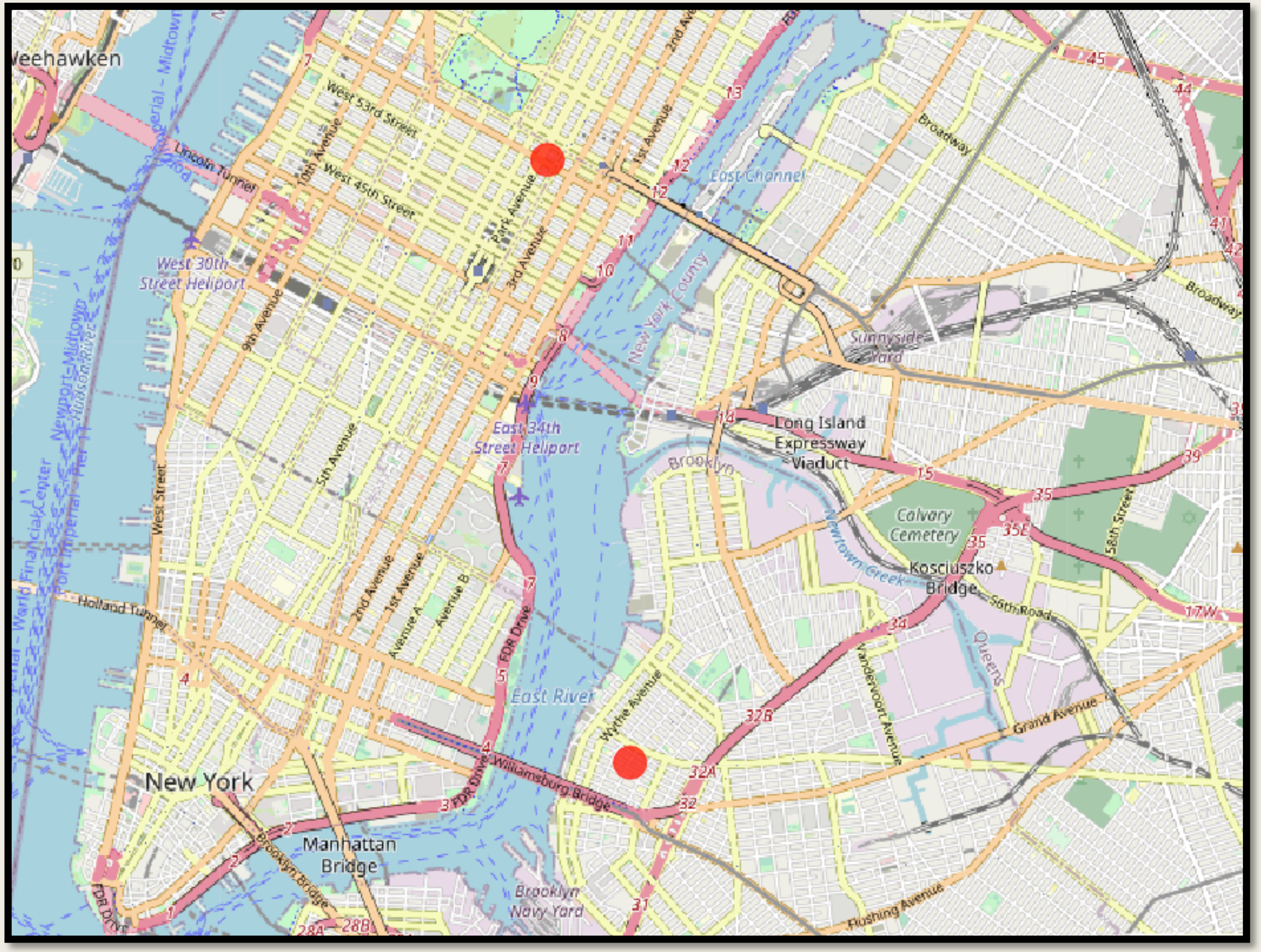
- Fit the testing dataset, i.e. the locations found on *Loopnet.com* and compute confusion matrix as shown on the right:
- Best locations are those with predicted label greater than True label, indicating growth potential
- Both red circles meet the requirement.



	0	1	2	3	4
0	3	1	0	0	0
1	1	0	0	0	0
2	1	0	0	1	0
3	0	0	0	0	0
4	0	1	0	0	0

4. Result

- 2 Red circles shown on the right are selected best locations for opening a new Starbucks or similar store
- Both stores are in dense areas, namely Manhattan and Williamsburg, meaning the market is yet to saturate.



5. Discussion

- The project achieved its goal using simple machine learning techniques and data science tools.
- Noted points and future improvements:
  - 1. The problem we seek solution is not a standard classification problem, in the sense that even the data available is not considered absolutely right, because we can actually select any location to open more stores to change the data. So classification accuracy in the testing data is not as important as in the training data, since we need the inaccuracy to provide insights.
  - 2. The data is very sparse, so we could further explore other options in data processing and modeling to better fit such dataset. Potentially a model designed specifically for sparse data would do a better job.

5. Discussion (Contd.)

- 3. Sample size is limited compared to number of features available.
  - *This work only focus on extracting venues data that is close to a certain area while there are huge amounts of other types of data that can be easily incorporated such as demographic;*
  - *Even then, we have 538 features compared to only ~2,000 data points. This could potentially lead to overfitting, so we tried using Principal Component Analysis(PCA)/Sparse PCA to reduce dimensionality, but the result is not ideal, so it was dropped in the end.*
  - *Nonetheless, future work could be done on expanding the data size by potentially include data from more cities with similar characteristics as New York such as Chicago, London, etc.*
- 4. The data is very sparse, so we could further explore other options in data processing and modeling to better fit such dataset. Potentially a model designed specifically for sparse data would do a better job.
- 5. Again, there are other excellent sources of data available for improvement. From a more realistic perspective, only looking at Starbucks would not be enough. A collection of similar stores would be a good starting point for next steps.

6. Conclusion

- Our analysis procedure :
  - *Extract Starbucks locations throughout New York City, and combine it with New York City census data to allocate Starbucks locations into tracts.*
  - *Pulled venues data from Foursquares API to characterize each tract by venues within them.*
  - *Process the data we have, combined with Loopnet data to construct our training-validation-testing datasets.*
  - *Deploy a neural network model to predict each location's Starbucks store count, contrasting it with the actual store count to get new store potentials.*

6. Conclusion (Contd.)

- What we've learnt from the above analysis:
  - *Starbucks locations clusters towards dense areas such as Midtown and Downtown Manhattan, Downtown Brooklyn, etc. Not only that, potential new stores also tend to fall into these areas, indicating that the market is still yet to be saturated;*
  - *Even though the data size is not very large, a good model still provides excellent insight in predicting and evaluating business location selection.*