Coursera Capstone Final Assignment Wee2 Final Report

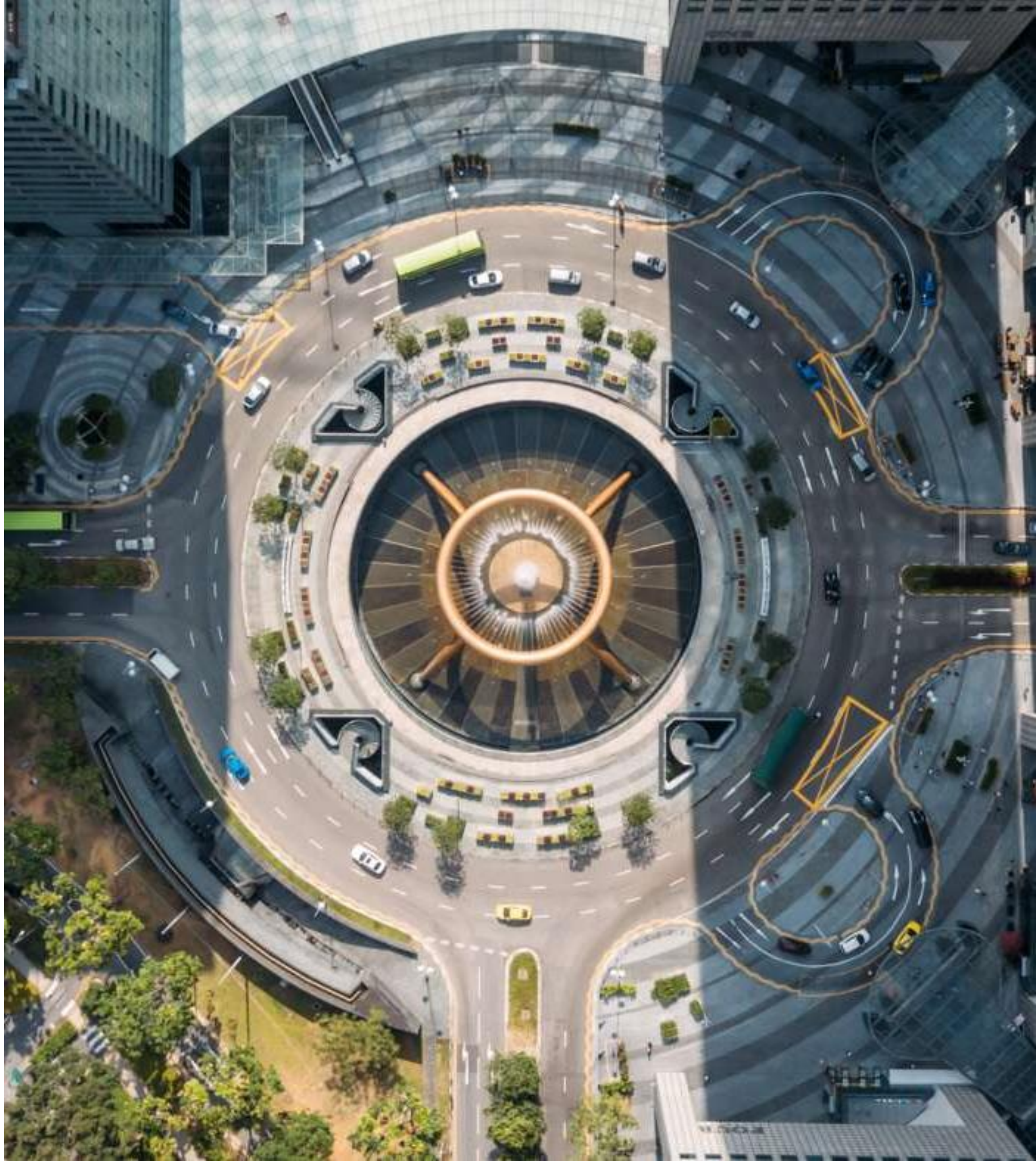
Applied Data Science
Capstone by IBM Coursera

Liu Xudong

2020-11-30

Content

- Background: Business Problem
- Datasets
- Methodology and Analysis
- Conclusion



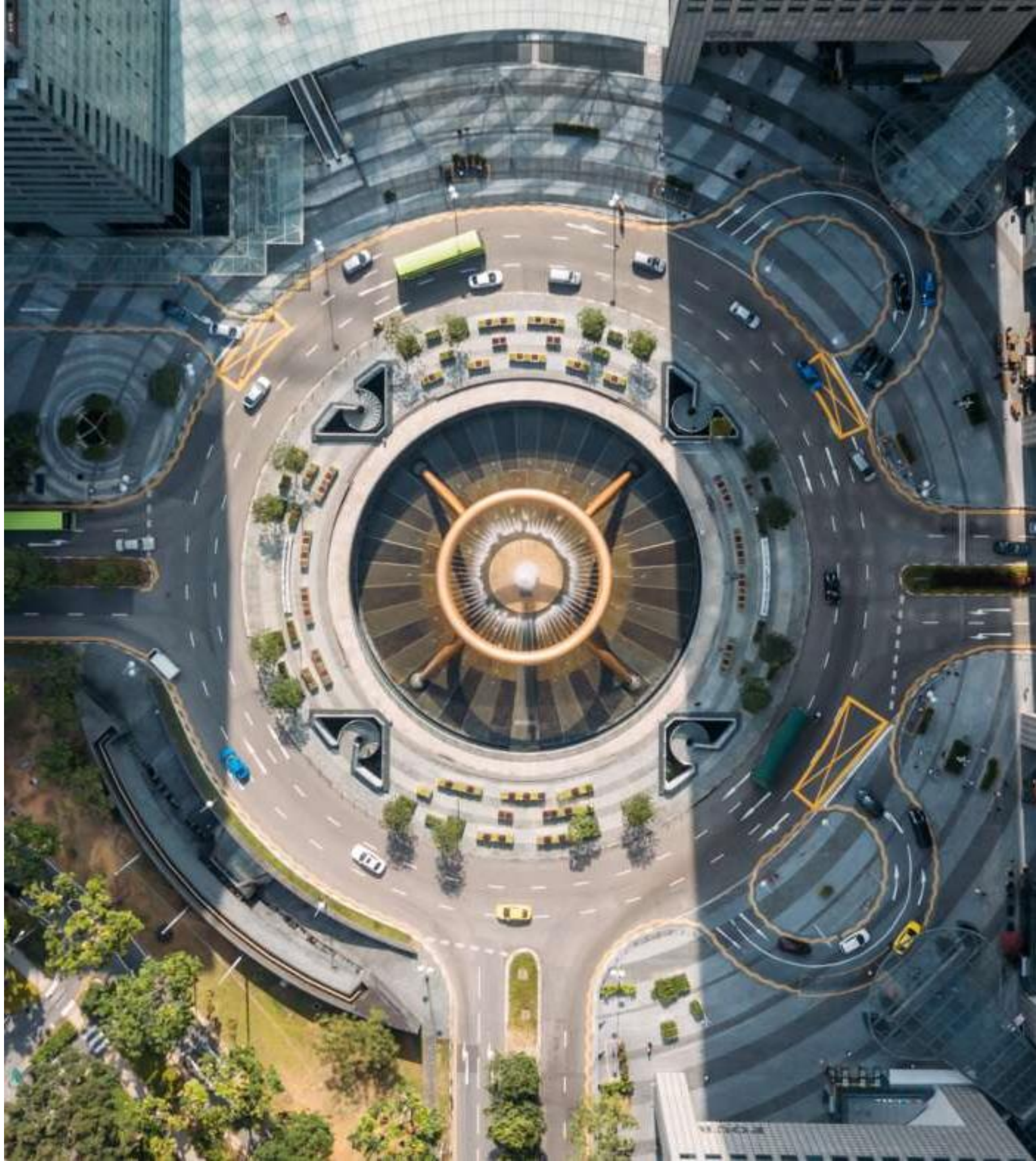
Background: Business Problem

" Is this place good enough to open a restaurant? "

- Existing Restaurants nearby?
- Population / population chagne in recent year?
- Number of shopping malls / bars / parks / etc ?
- Other food venue (bars / fast food / etc) ?

Content

- Background: Business Problem
- Datasets
- Methodology and Analysis
- Conclusion



Data Acquisition

- The neighborhood name, population and total income information is extracted from [Toronto Neighborhood Profile].
- The latitude and longitude of the neighborhood is given by google geocoding.
- Number of restaurants/stations/parks/etc and their location in every neighborhood will be obtained using Foursquare API.

df_toronto_neighbors

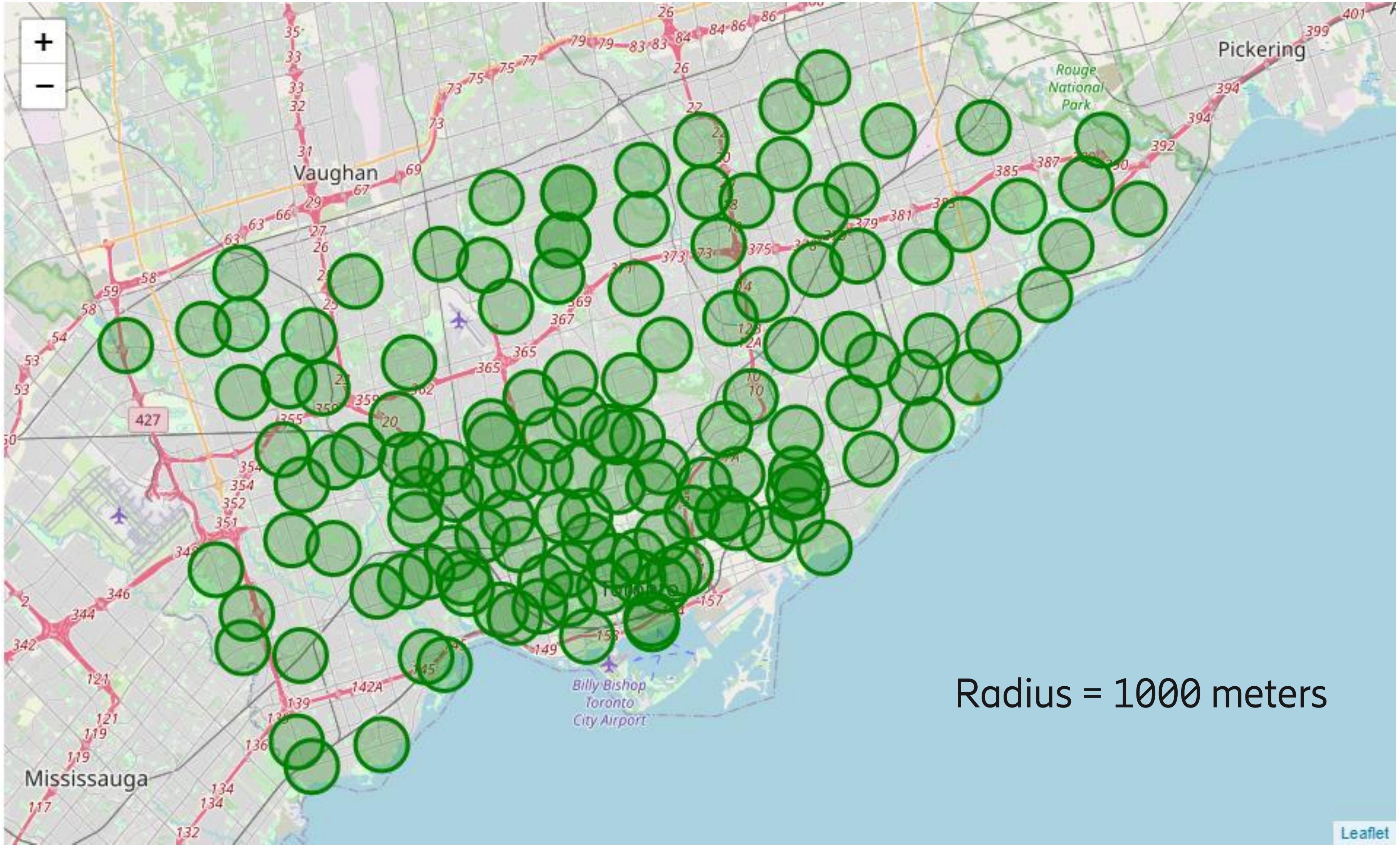
	Neighborhood	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Latitude	Longitude
0	Agincourt North	29,113	-3.90%	25,005	43.808053	-79.266502
1	Agincourt South-Malvern West	23,757	8.00%	20,400	43.788009	-79.283882
2	Alderwood	12,054	1.30%	10,265	43.601710	-79.545238
3	Annex	30,526	4.60%	26,295	43.669833	-79.407585
4	Banbury-Don Mills	27,695	2.90%	23,410	43.744847	-79.340923
...	...	...	...	...	...	...
135	Wychwood	14,349	2.60%	11,345	43.677910	-79.420102
136	Yonge-Eglinton	11,817	11.70%	9,995	43.706431	-79.398642
137	Yonge-St.Clair	12,528	7.50%	11,170	43.688098	-79.394117
138	York University Heights	27,593	-0.40%	23,530	43.766449	-79.477446
139	Yorkdale-Glen Park	14,804	0.80%	12,065	43.708236	-79.453975

140 rows × 6 columns

df_toronto_neighbors_merged

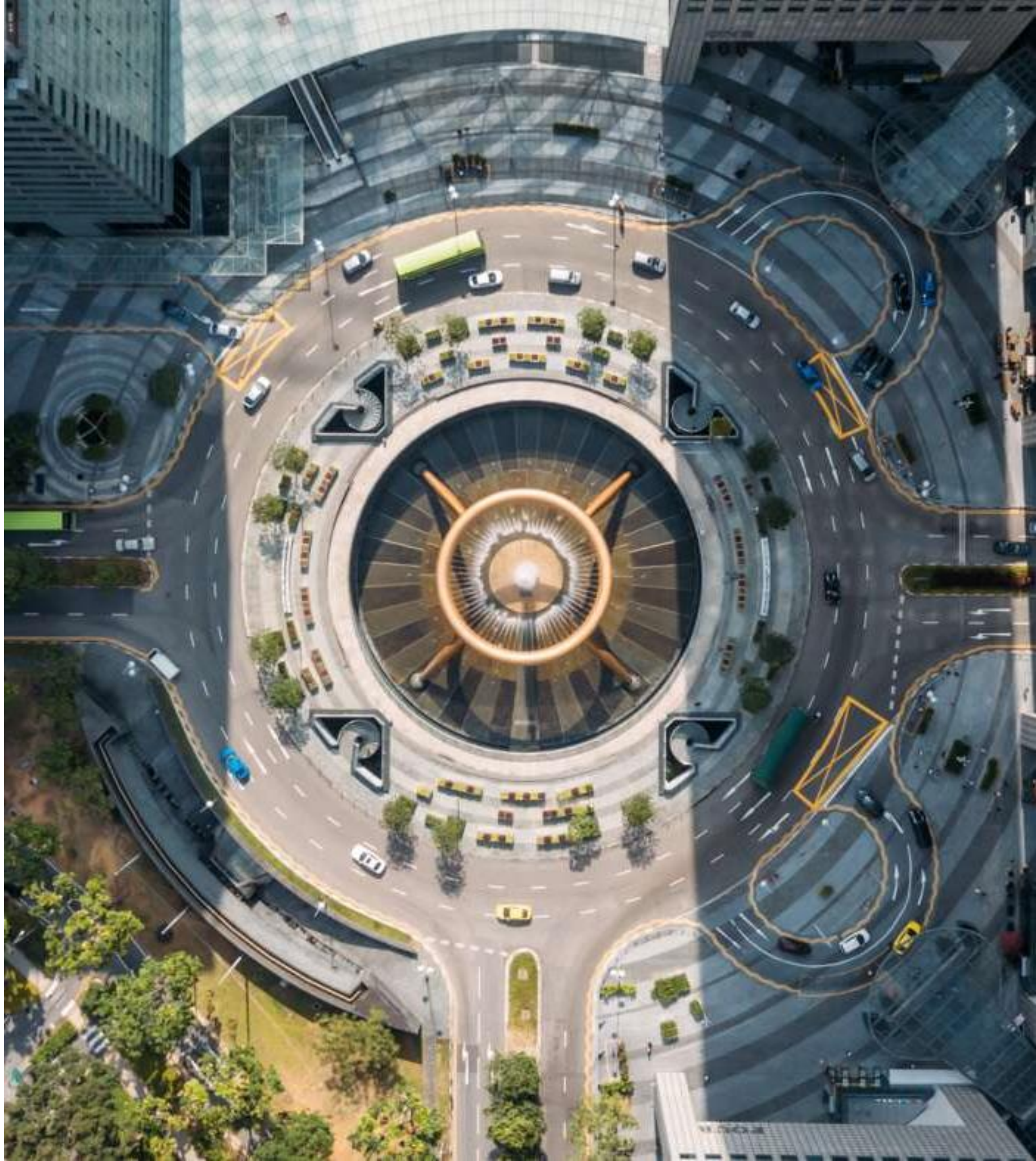
	Neighborhood	Latitude	Longitude	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Restaurant	Mall/Shops	Station/Subway	Bar/Drinks	Park	Hotel	School
0	Agincourt North	43.808053	-79.266502	29113.0	-0.039	25005.0	9	10	0	4	2	0	0
1	Agincourt South-Malvern West	43.788009	-79.283882	23757.0	0.080	20400.0	16	8	0	3	1	0	0
2	Alderwood	43.601710	-79.545238	12054.0	0.013	10265.0	3	9	0	2	2	0	0
3	Annex	43.669833	-79.407585	30526.0	0.046	26295.0	36	24	0	12	1	1	2
4	Banbury-Don Mills	43.744847	-79.340923	27695.0	0.029	23410.0	5	2	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
135	Wychwood	43.677910	-79.420102	14349.0	0.026	11345.0	34	25	0	13	2	0	1
136	Yonge-Eglinton	43.706431	-79.398642	11817.0	0.117	9995.0	38	26	0	16	1	0	0
137	Yonge-St.Clair	43.688098	-79.394117	12528.0	0.075	11170.0	17	10	1	4	2	1	0
138	York University Heights	43.766449	-79.477446	27593.0	-0.004	23530.0	8	11	0	6	0	0	0
139	Yorkdale-Glen Park	43.708236	-79.453975	14804.0	0.008	12065.0	14	10	1	3	0	0	0

Neighborhood Exploration

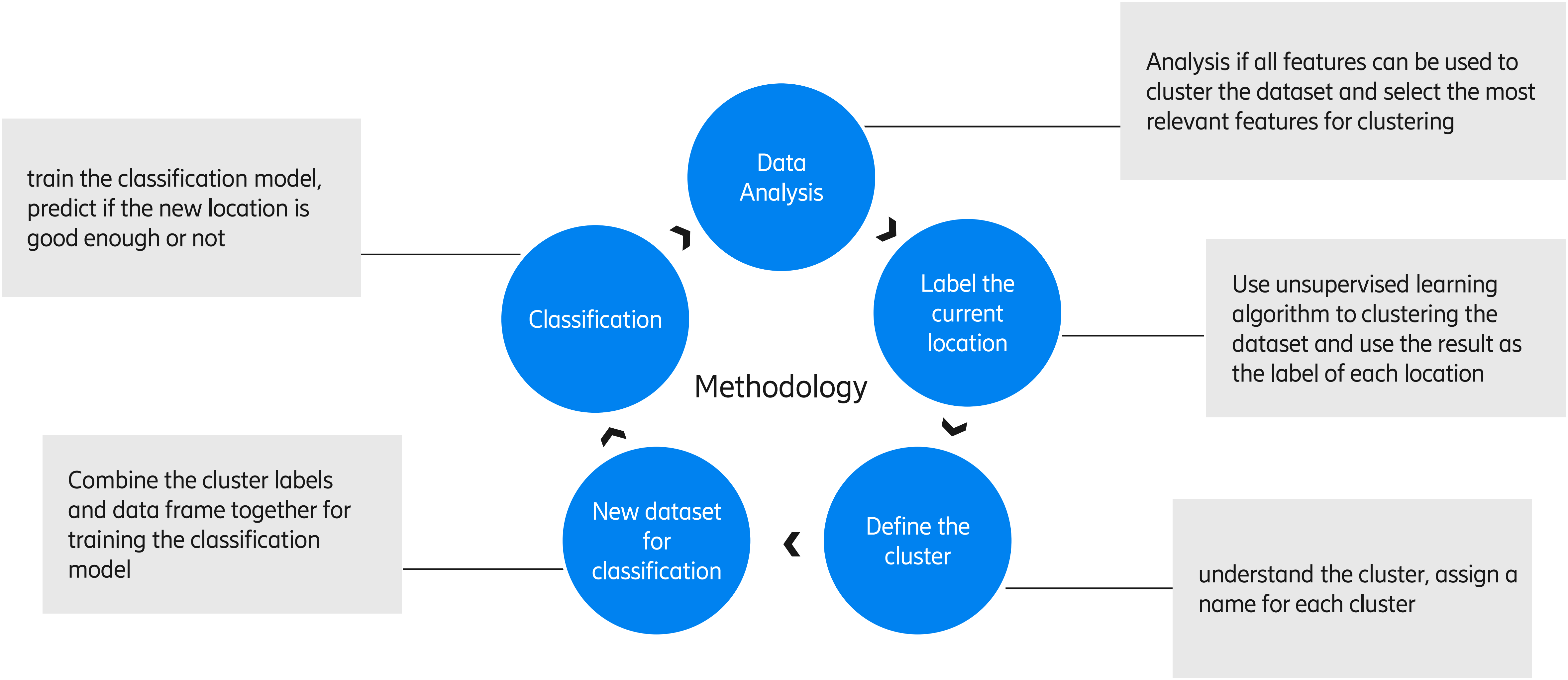


Content

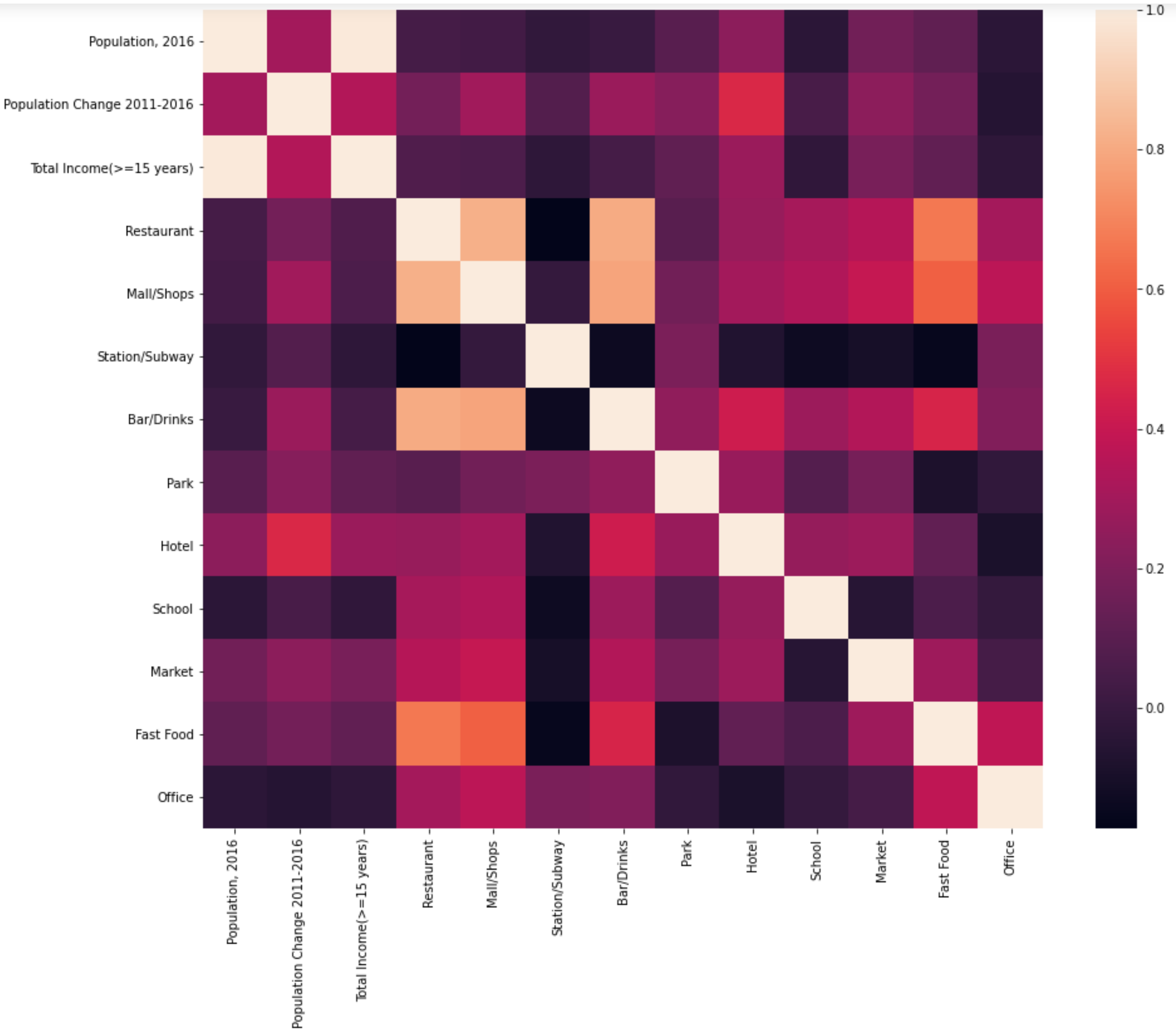
- Background: Business Problem
- Datasets
- Methodology and Analysis
- Conclusion

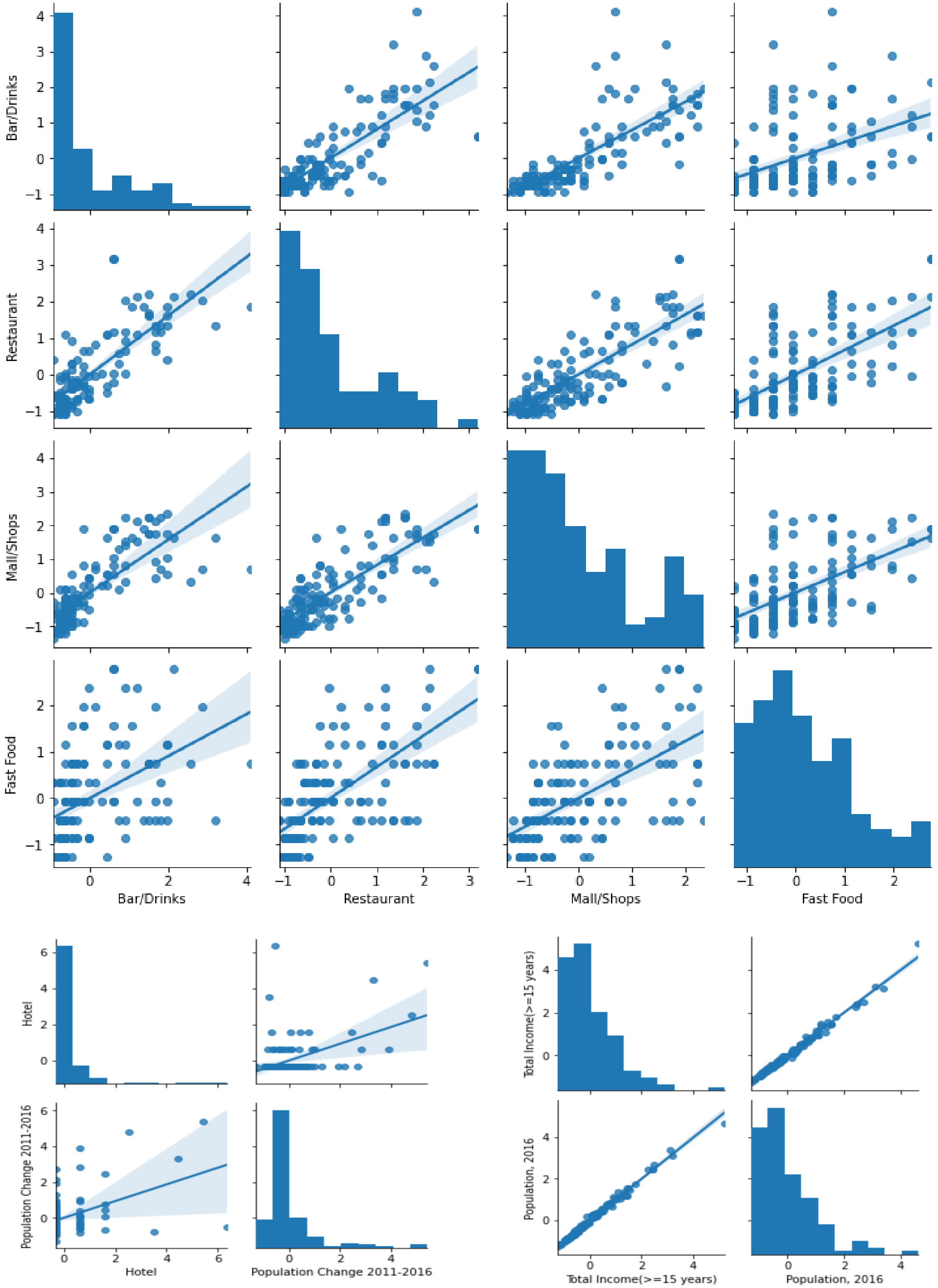


Methodology

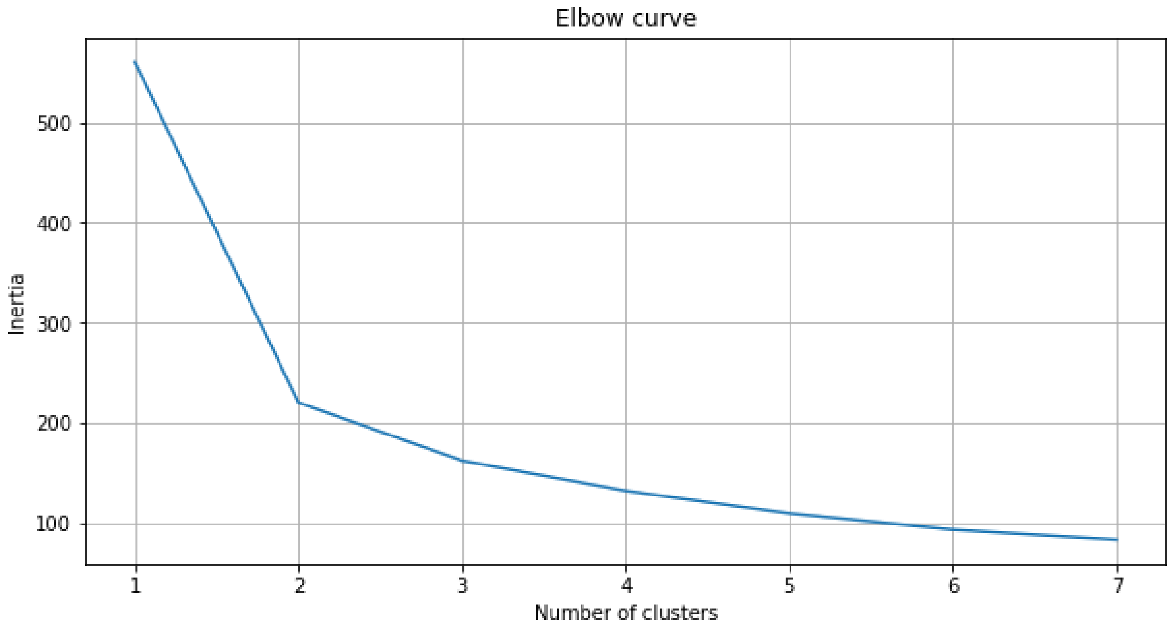


Data Analysis





Label the current location

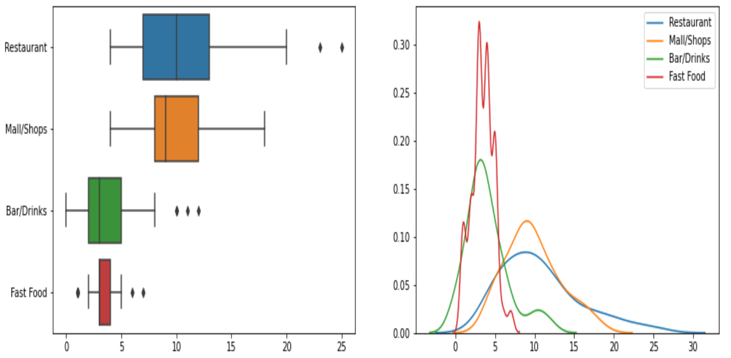


- Use K=5 in K-Means algorithm to do clustering
- Merge the cluster labels into data frame together with all other features for each location

Longitude	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Restaurant	Mall/Shops	Station/Subway	Bar/Drinks	Park	Hotel	School	Market	Fast Food	Office	Cluster Label
-79.266502	29113.0	-0.039	25005.0	9	10	0	4	2	0	0	0	4	2	0
-79.283882	23757.0	0.080	20400.0	16	8	0	3	1	0	0	1	4	1	0
-79.545238	12054.0	0.013	10265.0	3	9	0	2	2	0	0	0	1	0	2
-79.407585	30526.0	0.046	26295.0	36	24	0	12	1	1	2	0	5	1	1
-79.340923	27695.0	0.029	23410.0	5	2	0	1	0	0	0	0	0	2	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
-79.420102	14349.0	0.026	11345.0	34	25	0	13	2	0	1	1	7	1	4
-79.398642	11817.0	0.117	9995.0	38	26	0	16	1	0	0	1	5	1	1
-79.394117	12528.0	0.075	11170.0	17	10	1	4	2	1	0	1	1	1	0
-79.477446	27593.0	-0.004	23530.0	8	11	0	6	0	0	0	0	1	0	0
-79.453975	14804.0	0.008	12065.0	14	10	1	3	0	0	0	0	5	3	0

Define the cluster

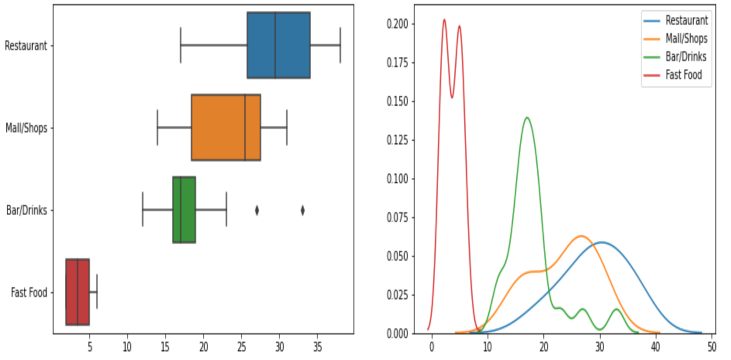
Cluster_0



- Invest restaurant near shopping mall

The distribution of Restaurant is very consistent with the shopping Mall and the total number is not large. So considering to locate the Restaurant near the shopping mall.

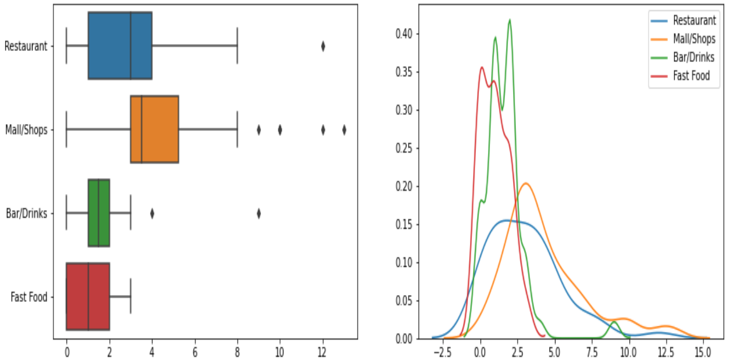
Cluster_1



- Invest bar/drinks near existing restaurant

The distribution of Restaurant is quite similarly with the shopping Mall (not as good as cluster_0), but where the Restaurant is concentrated the Bar/Drinks is few. So considering to invest Bar/Drinks near a restaurant.

Cluster_2

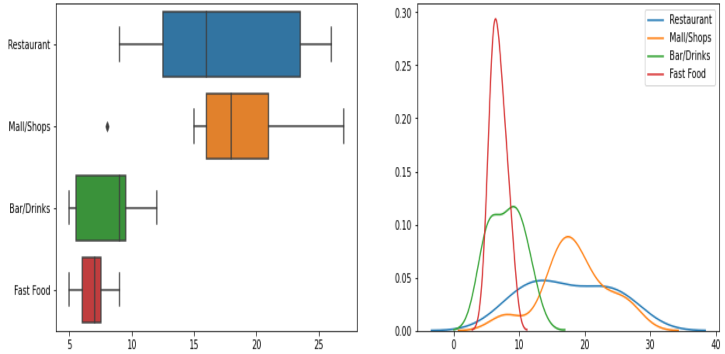


- Good place to open a restaurant

All of the venues are distributed similar, and number of each is not many. So this is a very good choice to invest Restaurant.

Define the cluster

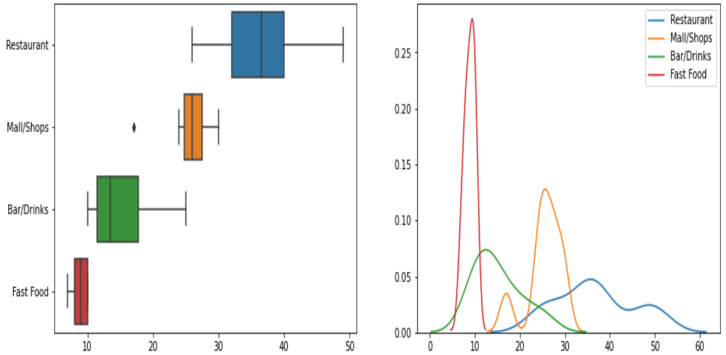
Cluster_3



- Invest restaurant near bar/drinks

The distribution range of Restaurant is large, and number is not small. But there are not enough restaurant at the large number of Bar/Drinks (high correlation), so a restaurant near a bar/drinks can be considered.

Cluster_4



- Not a good choice

The distribution of all the venues is very scattered, and number of Restaurant is large. So this is not a good choice to open a restaurant.

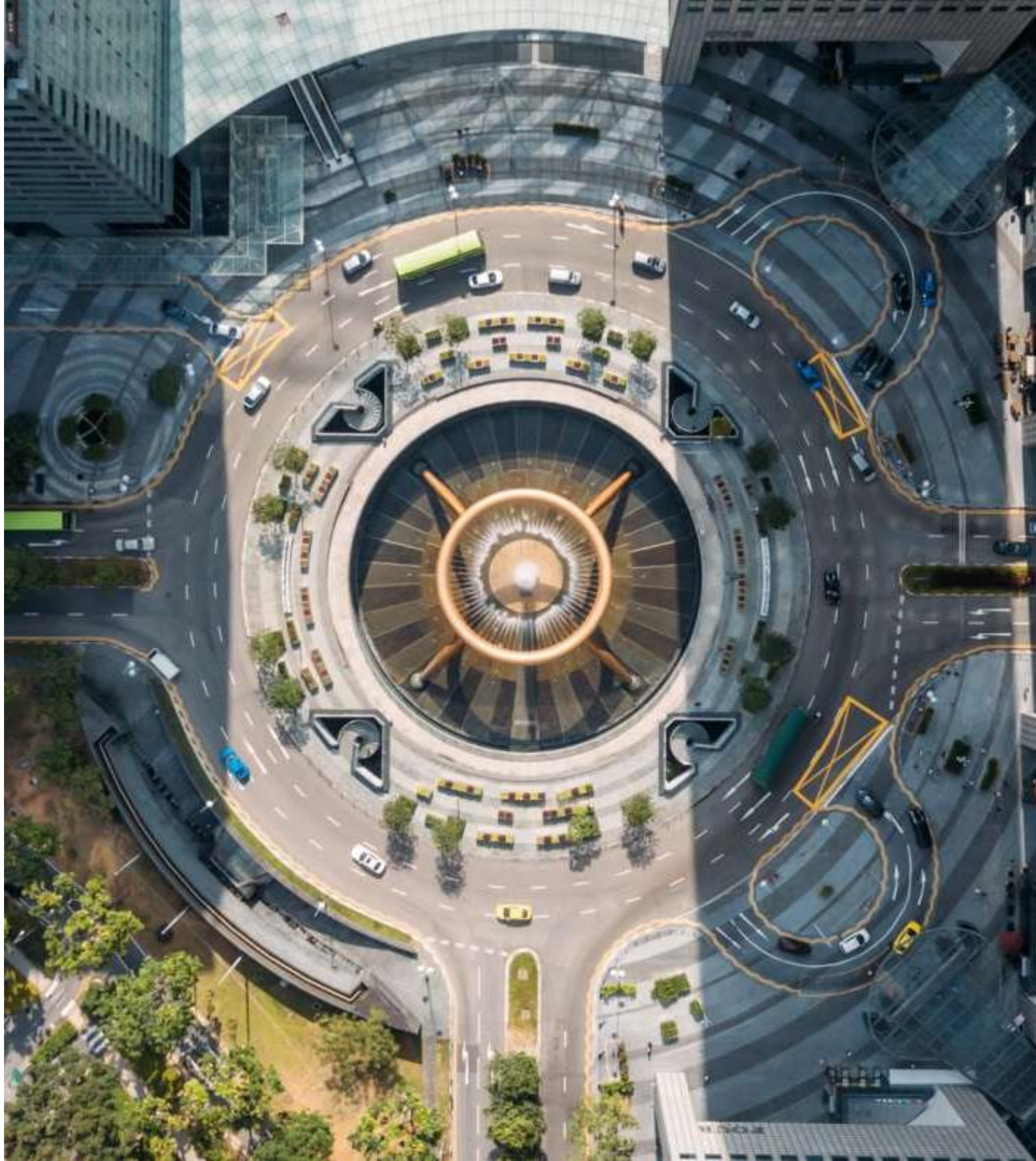
New dataset for classification & Classification

- Dataset
  - X = df_classify[['Restaurant','Mall/Shops','Bar/Drinks','Fast Food']]
  - y = df_classify['Cluster Label']
  - X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1)
- Classification Result
  - Overall F1-Scroe = 93%
  - Cluster_3 F1-Score = 80%
  - Cluster_4 F1-Score =80%
  - This may because the samples don't include as many samples as other clusters, so the model for this two cluster is not trained as good as other.
  - This can be improved by collecting more data.

```
______ Training Report ______
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        38
           1       1.00      1.00      1.00        16
           2       1.00      1.00      1.00        44
           3       1.00      1.00      1.00         9
           4       1.00      1.00      1.00         5

    accuracy                           1.00       112
   macro avg       1.00      1.00      1.00       112
weighted avg       1.00      1.00      1.00       112

______ Test Report ______
              precision    recall  f1-score   support

           0       1.00      0.91      0.95        11
           1       0.80      1.00      0.89         4
           2       1.00      1.00      1.00         8
           3       0.67      1.00      0.80         2
           4       1.00      0.67      0.80         3

    accuracy                           0.93        28
   macro avg       0.89      0.92      0.89        28
weighted avg       0.95      0.93      0.93        28
```

Content

- Background: Business Problem
- Datasets
- Methodology and Analysis
- Conclusion



Conclusion

- Obtain needed data from different sources
  - Radius used to explore the location can be more dynamic
- Use correlation analysis to analyse data and select features
  - More detailed categories can be defined
- Cluster the location by K-Means algorithm
- Predict which cluster the location belongs to
  - More data can be collected to help improve prediction accuracy