CAPSTONE PROJECT (WEEK 3 & 4) THE BATTLE OF THE NEIGHBORHOODS

BARDIYA B.

PROBLEM DESCRIPTION:

- Investigation of the types and numbers of venues in neighborhoods
- Identify needed venue types with business potential
- Identify neighborhoods with need for some venue types and business potential

IDEA:

- Investigate correlation between different venue types
- Make an educated guess about competetive and supporting venues in a strongly correlated cluster
- Subcluster into K1: competitive venues, K2: supproting venues for K1
- I.e. K1: Different types of restaurant, K2: venues which don‘t compete but correlate strongly with K1

IDEA PART 2:

- For each neighborhood:
  - Train a model to predict the number of occurencies of venues in K1 based on K2
  - Interpret the difference between predicted/expected number and real number as business opportunity or over-saturation of the market

DATA :

- The data we are going to use is the same as before.
  - ***We are going to scrape the following Wikipedia site for information about Toronto:*** ***https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M***
  - ***We are going to use the CSV file provided in the previous assignments to get the coordinates of the neighborhoods of Toronto***
  - ***We will use the Foursquare API to get information about venues in the different neighborhoods.***

METHODOLOGY:

Correlation C_{xy} between two types of venues x, y :

$$C_{xy} = \sum_i \frac{(n_{x,i} - \mu_{n_x})(n_{y,i} - \mu_{n_x})}{\sigma_{n_x}\sigma_{n_y}}$$

$n_{x,i}$: number of venues of type x in neighborhood i,

μ_{n_x}: average number of venues of type x over all neighborhoods

σ_{n_x}: corresponding standard deviation

METHODOLOGY PART 2:

- Restrict to venues that occur often enough, for proper statistics
- Find cluster of stronly correlated venues
- Split cluster:
  - K1: Competitive venues to make predictions for
  - K2: Other venues to base prediction on

METHODOLOGY PART 3:

- Use linear regression to make predictions:

$$n_{x,i}^{pred} = \sum_{y \in K_2} A_{xy} n_{y,i} \ , \qquad x \in K_1$$

- Define Business opportunity for venue type x and neighborhood i

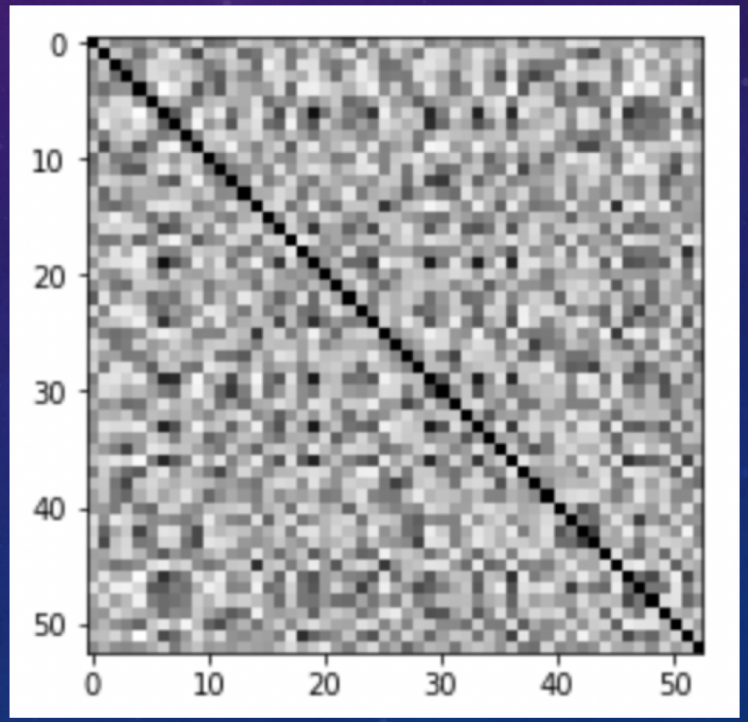
$$opp_{x,i} = n_{x,i}^{pred} - n_x$$

- Define overall business opportunity in cluster K1 by

$$overallopp_{K_1,i} = \sum_{x \in K_1} opp_{x,i}$$

RESULTS:

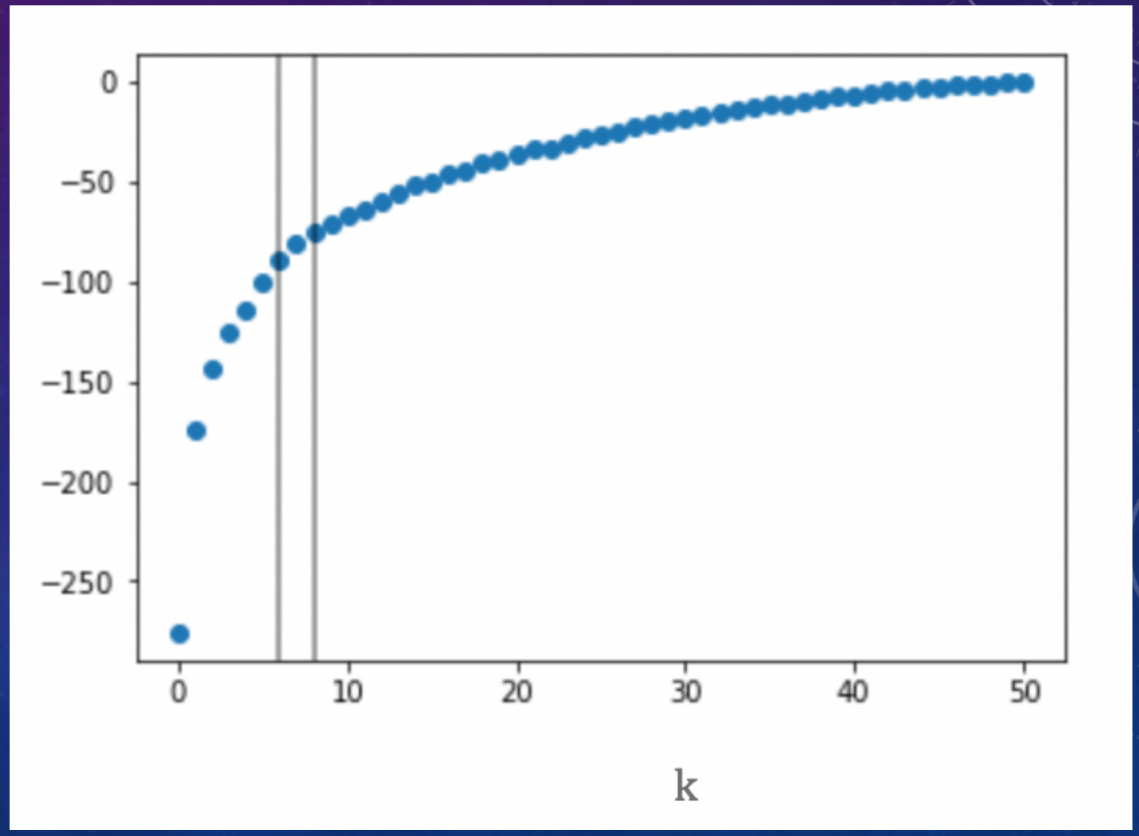
- Taking the top 25% venues in the neighborhoods of torronto we find the correlation heat map:



RESULTS:

- Running K-Mean on the rows of the correlation matrix

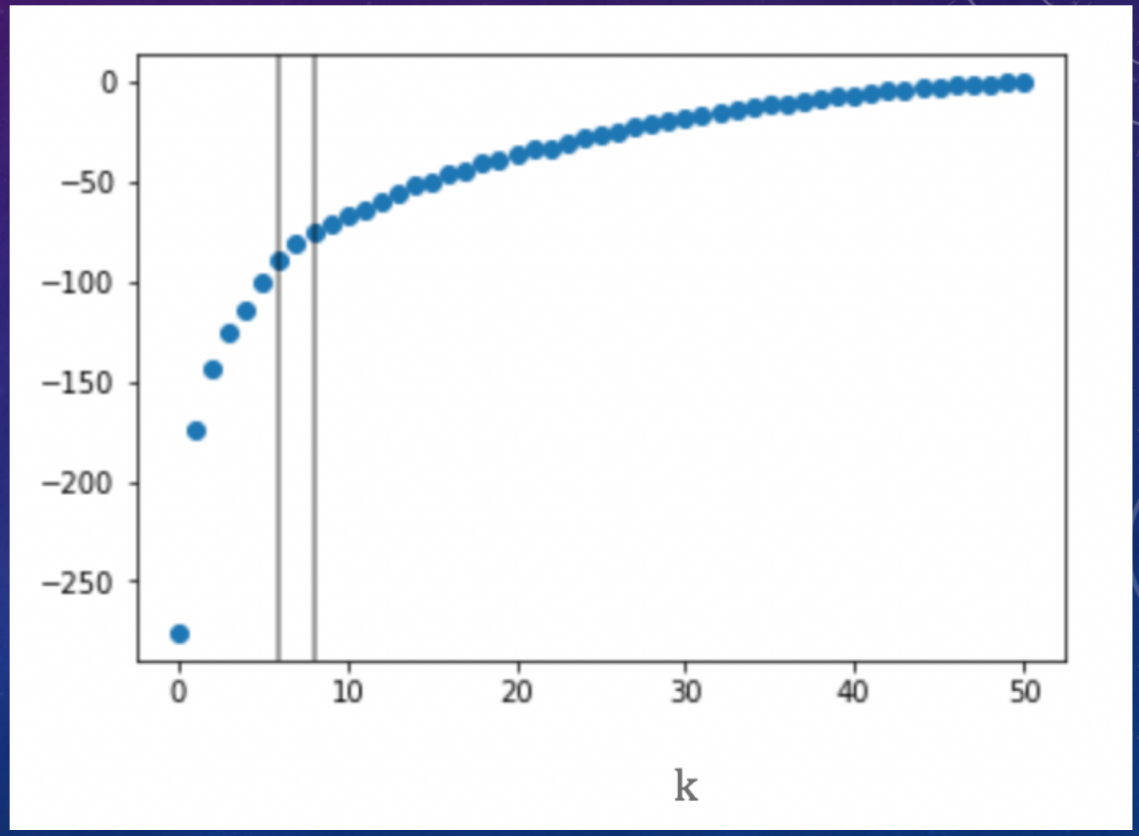
- Elbow-point at k=7



RESULTS:

- Running K-Mean on the rows of the correlation matrix

- Elbow-point at k=7

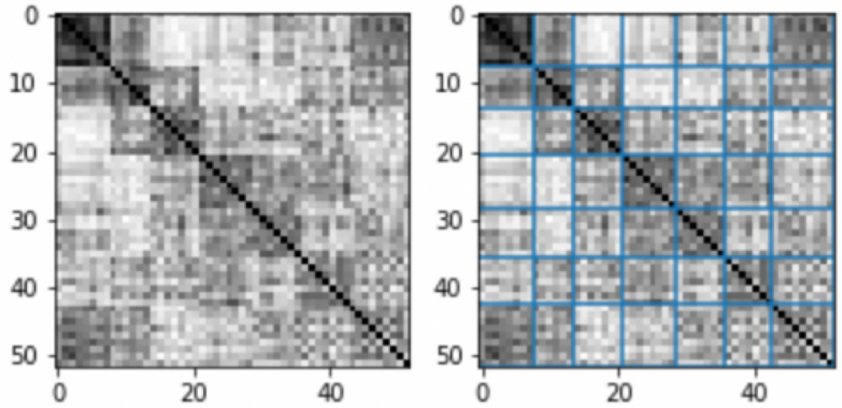


RESULTS:

- Strongly correlated cluster of size 8
- Cluster stays stable when varrying k around k=7

Number of clusters 7

	new label	corr_strength	cluster_size
2	1	0.684999	8
5	2	0.533677	6
3	3	0.494104	7
1	4	0.446302	8
6	5	0.391797	7
4	6	0.362063	7
0	7	0.235594	9



['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']

RESULTS:

- Cluster given by:

  ['Farmers Market', 'Plaza', 'Steakhouse', 'Hotel‘, 'Concert Hall', 'Mediterranean Restaurant’ ,

  'Thai Restaurant', 'Theater’]

- Split into

  K1: 'Steakhouse', 'Hotel‘, 'Concert Hall', 'Mediterranean Restaurant’ , 'Thai Restaurant’, K2: Rest

- Make predictions for the number of venues in K1 based on K2 from linear regression

['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']

RESULTS:

- Business opportunities for venues in K1, from difference between prediction/expectation vs. Real number of values

- Best overall opportunities

```
subcluster1_business_opps.sort_values('overall opp', ascending=False).head()
```

Neighborhood	Steakhouse	Mediterranean Restaurant	Thai Restaurant	overall opp
First Canadian Place, Underground city	1.160317	0.180937	0.299696	1.640949
Adelaide, King, Richmond	0.563131	0.353034	0.254969	1.171134
Business Reply Mail Processing Centre 969 Eastern	0.027390	0.284257	0.550327	0.861974
St. James Town	0.342961	-0.032706	0.406760	0.717015
Design Exchange, Toronto Dominion Centre	0.160317	0.180937	0.299696	0.640949

['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']

RESULTS:

- Best opportunities for Steakhouses

```
cols=['Steakhouse', 'overall opp']
subcluster1_business_opps[cols].sort_values(cols, ascending=False).head()
```

Neighborhood	Steakhouse	overall opp
First Canadian Place, Underground city	1.160317	1.640949
Adelaide, King, Richmond	0.563131	1.171134
The Danforth West, Riverdale	0.412011	0.328186
St. James Town	0.342961	0.717015
The Annex, North Midtown, Yorkville	0.256756	0.114095

['Steakhouse', 'Mediterranean Restaurant',
 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant',
 'Thai Restaurant']

RESULTS:

- Best opportunities for Medditerranean Restaurants

```
cols=['Mediterranean Restaurant', 'overall opp']
subcluster1_business_opps[cols].sort_values(cols, ascending=False).head()
```

Neighborhood	Mediterranean Restaurant	overall opp
Brockton, Exhibition Place, Parkdale Village	0.578186	-0.016786
Harbourfront East, Toronto Islands, Union Station	0.441413	0.169477
The Danforth West, Riverdale	0.439093	0.328186
Adelaide, King, Richmond	0.353034	1.171134
Business Reply Mail Processing Centre 969 Eastern	0.284257	0.861974

['Steakhouse', 'Mediterranean Restaurant',
 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant',
 'Thai Restaurant']

RESULTS:

- Best opportunities for Thai Restaurants

```
cols=['Thai Restaurant', 'overall opp']
subcluster1_business_opps[cols].sort_values(cols, ascending=False).head()
```

Neighborhood	Thai Restaurant	overall opp
Forest Hill North, Forest Hill West	0.595054	0.331789
Business Reply Mail Processing Centre 969 Eastern	0.550327	0.861974
St. James Town	0.406760	0.717015
Berczy Park	0.362033	0.247200
Ryerson, Garden District	0.362033	0.247200

['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']

DISCUSSION & CONCLUSIONS:

- We have suggested a method to identify neighborhood with business opportunities & identification of venue types with potential
- Problems:
  - Need more data for better statistics
  - Need to take into account more neighborhoods, cities, countries
  - Therefore need generalized model, mixed data split into train and test set
  - Try out other models, polynomial regression, random forrests, neural networks

['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']
['Steakhouse', 'Mediterranean Restaurant', 'Thai Restaurant']

THANK YOU