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

BARDIYA B.

1. PROBLEM DESCRIPTION:

We want to have a way to point out neighborhoods with potentially unused business opportunities and suggestions for promising venue types in those neighborhoods, so that a human investor can look into the suggestions and make his decisions. We will try come up with such suggestions by using the information about the neighborhood, in particular information about the number of venues of other types.

The idea is that a large number of venues of some types should support and create business opportunities for other types of venues too. For example in an area with a lot of bars it is very likely that there is potential for food places being profitable too.

Now a simple idea is to make use of the face that in reality the potential is often already exploited and the occurrence of venues types that are supportive of each other should correlate relatively strongly. Therefore we could use some machine learning model (e.g. decision tree or neural network) to predict the number of actual venues of one type by the number of venues of other types in the same area. By then making the reasonable assumption that in the most popular/busy areas the actual number of venues is relatively close to the optimal number (if much larger some businesses would fail and the number should reduce, if much lower the potential is usually exploited until the number grows), we could then interpret a large deviation between predicted values and real values as either an oversaturation of the market, where it would be difficult to open a new venue or a not saturated markets with potential business opportunities.

However, such models would predict for highly correlated venue types A and B, whose occurrence is correlated because they are competing with each other, instead of supporting each other, in a certain area with a large number of venues of type A also a large number of venues of type B. This would most likely be a correct prediction, as business advice not helpful. Opening a large burger joint in an area with a lot of other food places, just because the machine learning algorithm suggests a large number of burger joints in that area too, might be not very wise. In fact we might want to place our burger joint into an area with fewer burger joints than expected and with a large number of venues that are usually highly correlated with burger joints, indicating that they might be supportive of our business and avoid areas with a lot of venues of other types, whose occurrence is usually also highly correlated with burger joints, due to its competing nature.

Furthermore, a full-fledged machine learning model might be also a bit of an overkill, since we are not in building a model that makes business decisions on its own, but only a system that points out possible business opportunities for a human investor to look into more closely. Therefore a simpler model, which allows easy interpretation of the data might be more suitable.

The idea we are going to present it to cluster the correlation matrix into different groups of venues, who are similarly correlated in occurrence with each other and also with venues types of other groups and then let a human decide which venues in a cluster seem to be supportive and which competitive of each other. If we then find an area with a lower number of a certain type of venue, than one would expect from the number of supporting venues we could suggest a business opportunity, especially if the number of competing venue types is low. In a way the clustering the correlation matrix and let a human sub-cluster the groups further would be just a method of feature-engineering for a model suggesting business opportunites

2. DATA:

The data we are going to use is the same as before.

1. *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
2. *We are going to use the CSV file provided in the previous assignments to get the coordinates of the neighborhoods of Toronto*
3. *We will use the Foursquare API to get information about venues in the different neighborhoods.*

3. METHODOLOGY

In this section we will describe how to use clustering to find venues whose occurrence in a region is correlated, and how to use that information to suggest how many types of venue type A should be in a certain region, based on the occurrence of venues of type B. The reasoning for this is, that if they occur in a strongly correlated fashion, chances are that they do because they support each other by providing business opportunities for each other, just like bars and food places, since an area where people go out to drink they will eat a lot too.

In order to find the correlation between areas we will look at areas with a lot of venues. We could divide areas into discrete lattice and count the occurrence of all

types of venues in a given radius. For simplicity we will look here at neighborhoods and count the occurrence of different types of venues in the whole area. We will end up with a table of the following form:

Neighborhood	Plaza	Steakhouse	Museum	Farmers Market	Diner	Theater	Thai Restaurant	Mediterranean Restaurant	Concert Hall	Hotel	Comic Shop	Japanese Restaurant
Adelaide, King, Richmond	1.364651	1.433881	0.948831	1.170009	0.626260	1.453397	1.167228	1.216060	1.367451	1.236531	0.785607	1.001911
Berczy Park	1.388224	1.443571	0.981254	1.157728	0.622091	1.405578	1.171968	1.202566	1.386947	1.280386	0.777446	1.015558
Brockton, Exhibition Place, Parkdale Village	0.446877	0.467526	0.293527	0.361853	0.194796	0.405841	0.306989	0.359663	0.420309	0.229557	0.226256	0.230862
Business Reply Mail Processing Centre 969 Eastern	0.196595	0.200476	0.107360	0.141951	0.080425	0.245889	0.223040	0.208049	0.199769	0.138205	0.130483	0.161857
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara	0.616739	0.722034	0.459055	0.584584	0.287271	0.740262	0.585896	0.529804	0.717789	0.475847	0.369592	0.473916

We will then focus on the top 25% types of venues with the highest total number of occurences in other to have enough statistics to for our correlation calculation.

The Correlation C_{xy} between two types of venues x, y is given by the formular

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

Where $n_{x,i}$ is the number of venues of type x in neighborhood i, μ_{n_x} is the average number of venues of types x over the whole set of neighborhoods and σ_{n_x} is the corresponding standard deviation. The resulting correlation matrix C gives the correlation between all different types of venues.

Taking the rows (or alternatively the columns, which is equivalent due to the symmetry $C_{xy} = C_{yx}$) we then obtain vectors, with dimension equal to the number of different venues taken into consideration. Those vectors can be taken as feature vectors for a clustering algorithm to cluster the venues into groups of venues which are similarly correlated between each other, and similarly correlated to the venues of other clusters. We will use the simplest algorithm, the K-Mean algorithm here.

In a linearized model we can then write for a neighborhood the expected number of venues of type x, in terms of all other venues y $(y \neq x)$ in the same cluster in the form

$$n_{x,i}^{pred} = \sum_{y,y\neq x} A_{xy} n_{y,i}.$$

Where A_{xy} is for fixed x the regression vector obtained by linear regression over all neighborhoods and therefore independent of i.

The business opportunity $opp_{x,i}$ for a venue of type x in a neighborhood i can then be defined by taking the difference between predicted number of venue and real number of venues for x, i.e.

$$\begin{aligned} opp_{x,i} &= n_{x,i}^{pred} - n_x \\ &= \sum_{y,y\neq x} A_{xy} n_{y,i} - n_x. \end{aligned}$$

A positive value for $opp_{x,i}$ then indicates that one would expect in neighborhood i more venues of type x then currently exist, based on the other venue types in that area. We therefore suggest a business opportunity. A negative value on the other hand suggests an oversaturated market.

However, as we said, since venues can be correlated due to the competitive nature, not the supportive, we will look at a chosen cluster K and picking a certain venue type, e.g. Japanese Restaurants. Then we will subdivide Cluster K further into clusters K_1 and K_2, by reason, where we will put into K_1 the Japanese restaurant and all other types of venues in K, which are likely to compete with Japanese restaurants, e.g. all other types of restaurants. Into K_2 we will put the remaining types of venues.

We will then make predictions for all venues in K_1 based on the venues of K_2, i.e.

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

We can then define

$$\begin{aligned} opp_{x,i} &= n_{x,i}^{pred} - n_x \\ &= \sum_{y\in K_2} A_{xy} n_{y,i} - n_x, \end{aligned}$$

as the business opportunity for that special type of venue based on non-competitive venues in the same area and

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

As the overall business opportunity for the whole competing subcluster K_1, i.e. in the example with the Japanese restaurant we then have information about the business opportunity of restaurants in general and Japanese restaurants in particular. This way we could conclude that, even though a certain area has too few Japanese restaurants, the total number of restaurants in over-saturated and opening a restaurant might be unwise due to high number of competitive restaurants of other type.

4. RESULTS

In this section we go step by step through our results. The full results can be viewed in the Jupyter notebook.

We first start scraping information about all meighborhoods in Toronto from the Wikipedia page (see section 2). Scraping reults in a table of the form

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
...	...	...	...

From the previous assignments we have a csv file which provided information about coordinates for each postal code. Merging the tables results in a table of the form

```
toronto_df
```

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
5	M4P	Central Toronto	Davisville North	43.712751	-79.390197
6	M4R	Central Toronto	North Toronto West	43.715383	-79.405678
7	M4S	Central Toronto	Davisville	43.704324	-79.388790
8	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160
9	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049
...	...	...	...	...	...

We then use the Fourequare to get information about all venues in the neighborhoods:

```
toronto_venues = getNearbyVenues(names=toronto_df['Neighbourhood'],
                                 latitudes=toronto_df['Latitude'],
                                 longitudes=toronto_df['Longitude']
                                )
```

```
The Beaches
The Danforth West, Riverdale
The Beaches West, India Bazaar
Studio District
Lawrence Park
Davisville North
North Toronto West
Davisville
Moore Park, Summerhill East
Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West
```

In order to obtain information about the number of venues per neighborhood we first transform the table using one hot encoding

```
# one hot encoding
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep=""

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = toronto_onehot.columns.tolist()
fixed_columns.remove('Neighborhood')
fixed_columns = ['Neighborhood'] + fixed_columns
toronto_onehot = toronto_onehot[fixed_columns]
toronto_onehot
```

	Neighborhood	Airport	American Restaurant	Amphitheater	Antique Shop	Aquarium	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	I J
0	The Beaches	0	0	0	0	0	0	0	0	0	0	
1	The Beaches	0	0	0	0	0	0	0	0	0	0	
2	The Beaches	0	0	0	0	0	0	0	0	0	0	
3	The Beaches	0	0	0	0	0	0	0	0	0	0	
4	The Beaches	0	0	0	0	0	0	0	0	0	0	
5	The Beaches	0	0	0	0	0	0	0	0	0	0	
6	The Beaches	0	0	0	0	0	0	0	0	0	0	
7	The Beaches	0	0	0	0	0	0	0	0	0	0	
8	The Beaches	0	0	0	0	0	0	0	0	0	0	
9	The Beaches	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	

We then filter for the top 25% venues which occur the most to have enough statistics available for further calculations.

```
venue_count=toronto_onehot.drop('Neighborhood', axis=1).T.sum(axis=1).to_frame()
venue_count=venue_count.rename(columns={0:'count'})
venue_count=venue_count.sort_values('count', ascending=False)
top_venue_types=venue_count.iloc[0:int(venue_count.shape[0]/4)].index.values.tolist()
venue_count
```

	count
Coffee Shop	257
Café	236
Park	155
Italian Restaurant	153
Bakery	93
Pizza Place	86
Hotel	79
Japanese Restaurant	77
Restaurant	75
Bar	68
...	...

Grouping over neighborhoods and summing over venues of the same type we then obtain a table of the form above, with the number of venues of a type in a certain neighborhood.

A heat matrix of the correlation between the values then looks as follows:

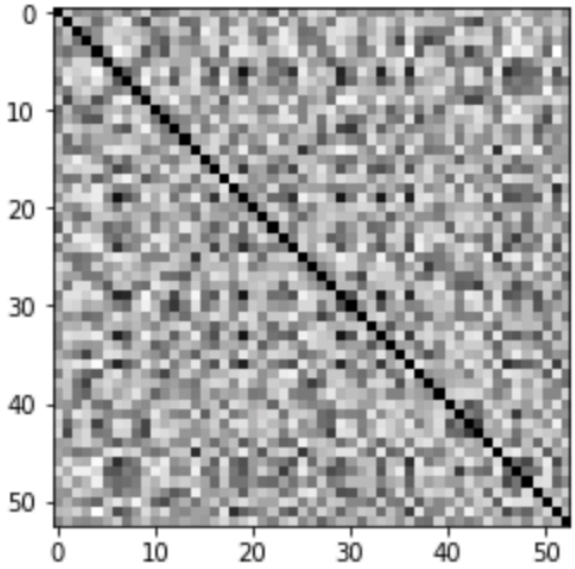


Fig 1: Heat Map for the correlation between the top 0.25 venue types

Running k-mean clustering algorithm over the rows of the matrix with different number of clusters results in a accuracy diagram

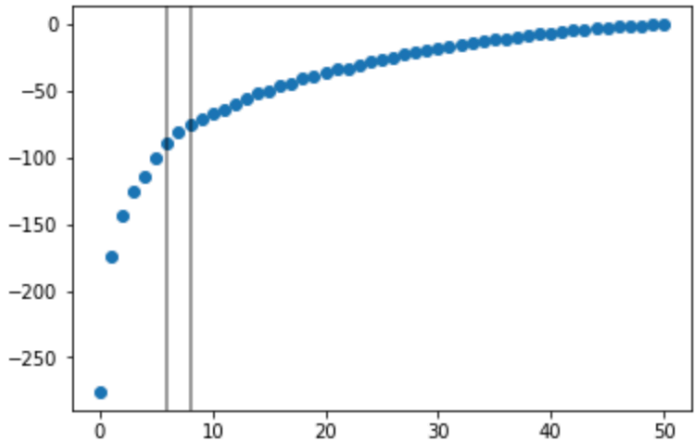


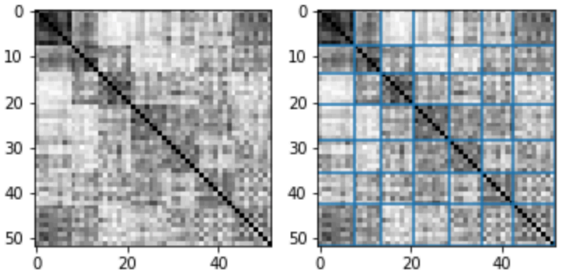
Fig. 2: Accuracy for k-means clustering of the correlation matrix

The elbow point seems to be at k=7.

Sorting the heat-max into the corresponding clusters and sorting the clusters by mean correlation strength results for k=7 in:

```
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



We see a relatively strongly correlated cluster of size 8 (Cluster with new label 1). Further analysis shows that this cluster remains stable when varying the number of clusters around k=7. So this might be a good candidate cluster to investigate business opportunities.

The Cluster is given by the following venues:

```
Cluster1=[

        'Farmers Market', 'Plaza', 'Steakhouse', 'Hotel',
        'Concert Hall', 'Mediterranean Restaurant'
        'Thai Restaurant', 'Theater'
]
```

We see that there are some types of food places in that cluster. We plan to make predictions for restaurants and we divide Cluster 2 into a subcluster K1 of restaurants and a subcluster K2 of remaining venues.

```
K1      =       ['Steakhouse', 'Mediterranean Restaurant',
                'Thai Restaurant']


K2      =       ['Farmers Market', 'Plaza', 'Theater',
                'Concert Hall', 'Hotel']
```

Using linear regression as follows, we will then make predictions for the restaurant types in K1 based on the venues in K2:

```
for i in subcluster1:
    X=np.array(toronto_venues[subcluster2])
    y=np.array(toronto_venues[[i]])
    reg=LinearRegression(fit_intercept=False).fit(X, y)
    subcluster1_predictions[[i]]=np.dot(np.array(toronto_venues[subcluster2]), np.array(reg.coef_[0]))
```

This will give us the prediction table

```
subcluster1_predictions.head()
```

Neighborhood	Steakhouse	Mediterranean Restaurant	Thai Restaurant
Adelaide, King, Richmond	1.563131	1.353034	2.254969
Berczy Park	1.745775	1.139392	2.362033
Brockton, Exhibition Place, Parkdale Village	-0.230257	0.578186	0.635286
Business Reply Mail Processing Centre 969 Eastern	0.027390	0.284257	0.550327
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara	0.886487	0.482941	0.946370

and subtracting the table with real venue numbers

```
subcluster1_venues.head()
```

Neighborhood	Steakhouse	Mediterranean Restaurant	Thai Restaurant
Adelaide, King, Richmond	1	1	2
Berczy Park	2	1	2
Brockton, Exhibition Place, Parkdale Village	0	0	1
Business Reply Mail Processing Centre 969 Eastern	0	0	0
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara	1	1	1

We get the table of business opportunities

```
subcluster1_business_opps.head()
```

Neighborhood	Steakhouse	Mediterranean Restaurant	Thai Restaurant	overall opp
Adelaide, King, Richmond	0.563131	0.353034	0.254969	1.171134
Berczy Park	-0.254225	0.139392	0.362033	0.247200
Brockton, Exhibition Place, Parkdale Village	-0.230257	0.578186	-0.364714	-0.016786
Business Reply Mail Processing Centre 969 Eastern	0.027390	0.284257	0.550327	0.861974
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara	-0.113513	-0.517059	-0.053630	-0.684202

which we have enriched by the overall opportunity number.

Sorting the table by the overall business opportunity number, which can see the neighborhoods with best business opportunity and which time of restaurant has the most potential:

```
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

Sorting the table by the business opportunity number for the different types of venues, we can on the other hand see which neighborhood has the most potential for each venue, and the corresponding overall opportunity in that neighborhood:

```
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

```
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

```
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

5. DISCUSSION & CONCLUSION

As we have seen we have built some type of systems to recognize outliers in the expected number of venues, based on the number of other venues with which they correlate, in order to suggest over- and under-saturated markets and find business opportunities.

We have further split a cluster of correlated venues by reason into subclusters of competing similar venues, and other non-competing venues which might then me supportive. With that we have suggested business opportunities for restaurants based on non-restaurant venues they correlate with.

However we should keep in my a few problems.

Our correlations are based on a very small number of venues and neighborhoods.

We might want to get more neighborhoods and data into account, so that we can also calculate correlation of occurences based on a smaller radius and make the whole correlation calculation more likely to reflect a connection between different venues.

Furthermore we need more statistic to make significance tests. To be able to confidently say that we have found an business opportunity, we would need to make sure that the deviation between predicted value and real value is big enough so that it is a statistical outlier and could be interpreted as a special event, rather than usual statistical error.

Also one might want to use more complex models such a polynomial regression, random forests and neural networks, instead of just using a linear. But again for such complex models we have to extend our approach to use much more data.