Assessment of relation between pedestrian traffic and popular venues, Toronto

Author: Noel Weber
Date: 14-5-2021

1. Business Problem

For a city, in order to control traffic flows and manage availability of public transport options and traffic regulation, it is helpful to be able to analyze and predict the flow of pedestrian traffic throughout a city. If there are hotspots where many people gather, certain precautions may be required in order to ensure public safety and / or manage traffic in other forms.

However, it is also good to know how pedestrian traffic changes with the availability of new venues. If, for example, new restaurants or cafés open in an area, how will the amount of foot-traffic change? It would be beneficial for the municipalities to know this upfront in order to account for increases.

That is the business problem this project is concerned with. For the city of Toronto, the pedestrian foot-traffic at peak hours at popular intersections is assessed, combined with information on close-by venues. A model will be extracted to predict foot-traffic based on the type of venues that are closest to the intersections.

The audience for this report is the department of traffic management of the city of Toronto. The insights aim to make better predictions of pedestrian traffic, thus helping with management issues.

2. Data

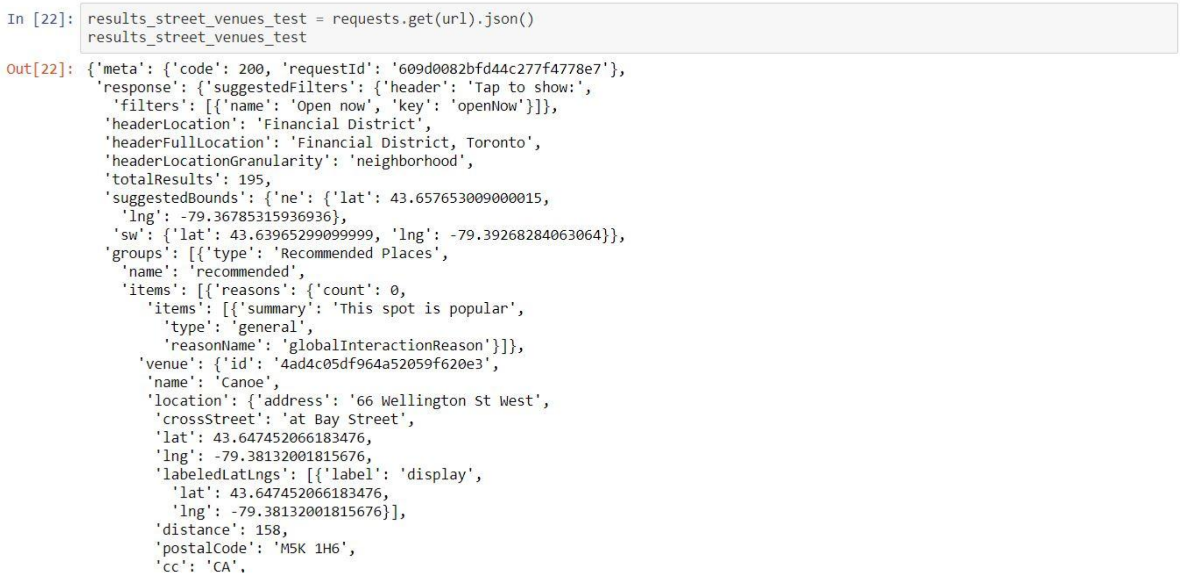
2.1. Data sources

Two data sources are used to conduct this project. They are explained in the following.

FourSquare Data
FourSquare is a global location data repository that yields data on locations and venues. It is the foundation for many popular applications, such as uber. Contributions to FourSquare can be made by the general public, ensuring that the database keeps on growing and improving.

The FourSquare data that is used for this project is venue and location data, meaning that for a certain location (provided in latitude and longitude), venues are found that are within a certain radius around this location.
The parameters that need to be set are the radius around the location, the limit of number of venues to be returned, as well as the user specific FourSquare access tokens. All these are converted to a user and problem specific url, which is used to extract the venue data.

The venue data consists of the name of the venue, the location (latitude, longitude), the postal code, the category, the exact address, and information such as whether it is a popular venue. The data is returned as an object and is unpacked into a dataframe to be able to work with the data. Hereby, the name, venue category and the latitude and longitude are the only parameters considered important. An extract of the returned object is shown.



```
In [22]: results_street_venues_test = requests.get(url).json()
         results_street_venues_test
```

```
Out[22]: {'meta': {'code': 200, 'requestId': '609d0082bfd44c277f4778e7'},
          'response': {'suggestedFilters': {'header': 'Tap to show:',
            'filters': [{'name': 'Open now', 'key': 'openNow'}]},
           'headerLocation': 'Financial District',
           'headerFullLocation': 'Financial District, Toronto',
           'headerLocationGranularity': 'neighborhood',
           'totalResults': 195,
           'suggestedBounds': {'ne': {'lat': 43.657653009000015,
             'lng': -79.36785315936936},
            'sw': {'lat': 43.63965299099999, 'lng': -79.39268284063064}},
           'groups': [{'type': 'Recommended Places',
             'name': 'recommended',
             'items': [{'reasons': {'count': 0,
                'items': [{'summary': 'This spot is popular',
                  'type': 'general',
                  'reasonName': 'globalInteractionReason'}]},
               'venue': {'id': '4ad4c05df964a52059f620e3',
                'name': 'Canoe',
                'location': {'address': '66 Wellington St West',
                 'crossStreet': 'at Bay Street',
                 'lat': 43.647452066183476,
                 'lng': -79.38132001815676,
                 'labeledLatLngs': [{'label': 'display',
                   'lat': 43.647452066183476,
                   'lng': -79.38132001815676}],
                 'distance': 158,
                 'postalCode': 'M5K 1H6',
                 'cc': 'CA',
```

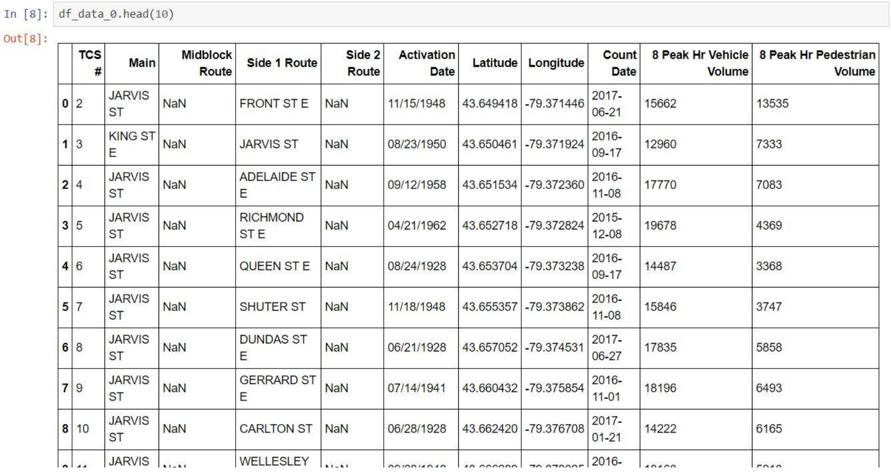
Image 1: snippet of FourSquare data export

Pedestrian volume data

The data on volume of pedestrians in the city of Toronto can be extracted from an Excel File that can be downloaded from the website of the city of Toronto (https://open.toronto.ca/dataset/traffic-signal-vehicle-and-pedestrian-volumes/). The data contains information on traffic volume at traffic lights from 2018.

The datafile is read into the project notebook and transformed into a dataframe for easier manipulation. The information contained in the data is the main street of the intersection, the first and (if applicable) second side street, the date when the traffic light was activated, the Latitude, the Longitude, the count date of traffic, the peak volume within 8 hours for vehicles and pedestrian separately.

An extract from the data is shown.



```
In [8]: df_data_0.head(10)
```

Out[8]:

	TCS #	Main	Midblock Route	Side 1 Route	Side 2 Route	Activation Date	Latitude	Longitude	Count Date	8 Peak Hr Vehicle Volume	8 Peak Hr Pedestrian Volume
0	2	JARVIS ST	NaN	FRONT ST E	NaN	11/15/1948	43.649418	-79.371446	2017-06-21	15662	13535
1	3	KING ST E	NaN	JARVIS ST	NaN	08/23/1950	43.650461	-79.371924	2016-09-17	12960	7333
2	4	JARVIS ST	NaN	ADELAIDE ST E	NaN	09/12/1958	43.651534	-79.372360	2016-11-08	17770	7083
3	5	JARVIS ST	NaN	RICHMOND ST E	NaN	04/21/1962	43.652718	-79.372824	2015-12-08	19678	4369
4	6	JARVIS ST	NaN	QUEEN ST E	NaN	08/24/1928	43.653704	-79.373238	2016-09-17	14487	3368
5	7	JARVIS ST	NaN	SHUTER ST	NaN	11/18/1948	43.655357	-79.373862	2016-11-08	15846	3747
6	8	JARVIS ST	NaN	DUNDAS ST E	NaN	06/21/1928	43.657052	-79.374531	2017-06-27	17835	5858
7	9	JARVIS ST	NaN	GERRARD ST E	NaN	07/14/1941	43.660432	-79.375854	2016-11-01	18196	6493
8	10	JARVIS ST	NaN	CARLTON ST	NaN	06/28/1928	43.662420	-79.376708	2017-01-21	14222	6165
[illegible]	[illegible]	JARVIS	[illegible]	WELLESLEY	[illegible]	[illegible]	[illegible]	[illegible]	2016-	[illegible]	[illegible]

Image 2: Pedestrian volume as extracted from Toronto data file

2.2 Data preparation

Firstly, the pedestrian volume data is imported. Since the data contains multiple columns of irrelevant information, as discussed above, it is sorted by the column ‘8 Peak Hr Pedestrian Volume’ to gain an overview of the data. The shape of the dataset is also included. From the shape, it is observed that data for 2280 pedestrian crossing locations is available. This is considered excessive, since the main interest is for downtown Toronto (where it is expected that the highest number of foot-traffic occurs that is relevant for traffic management). Thus, the 150 crossings with the highest pedestrian volume are selected.

The data is prepared by dropping the columns that are not needed for further assessment, “TCS #”, “Midblock Route”, and “8 Peak Hr Vehicle Volume”.

To confirm that the data set is centered around central Toronto, the crossings are plotted using the folium library. The plotted results are seen below:

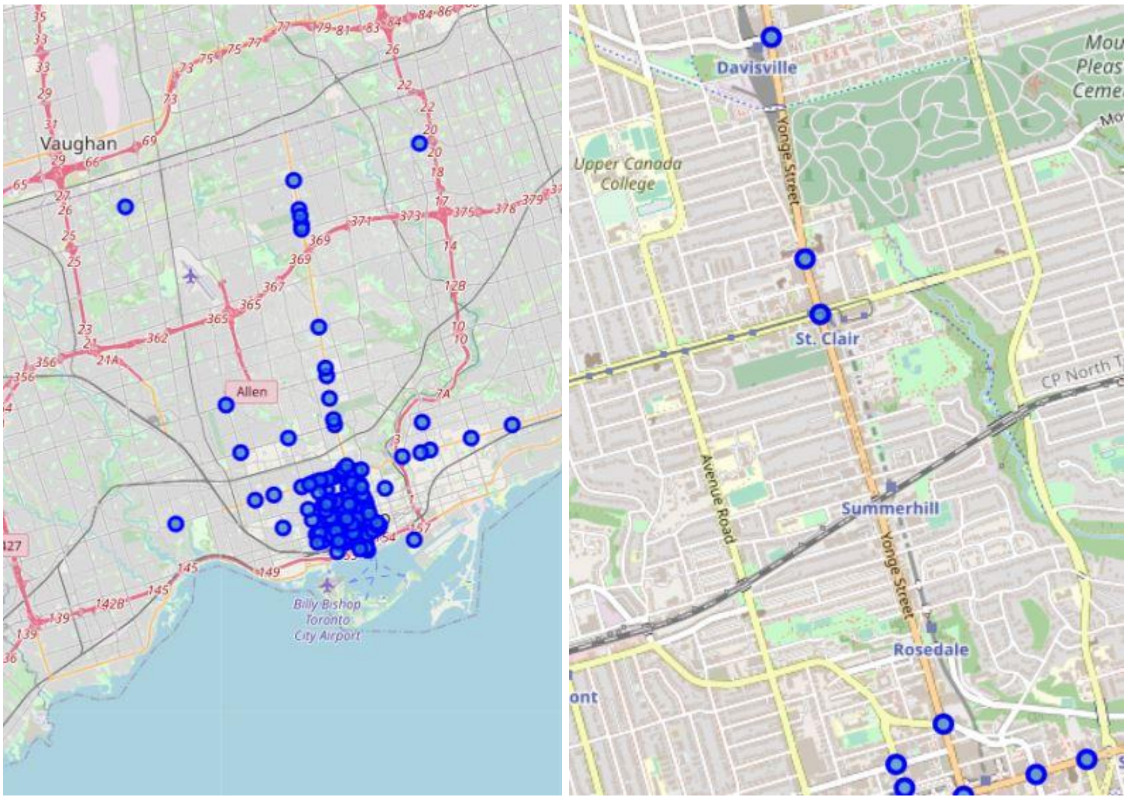
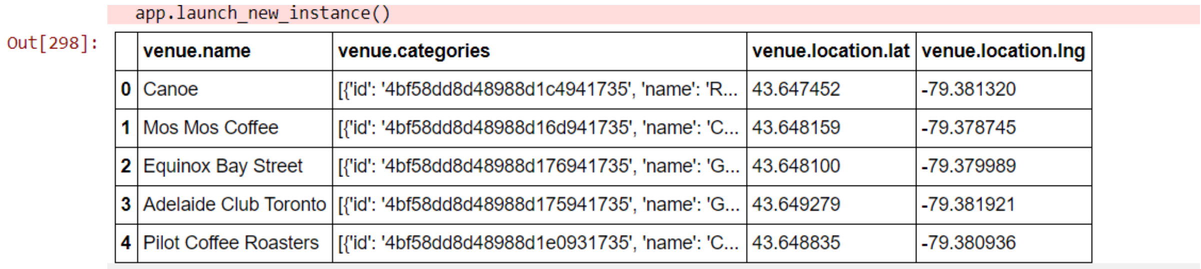


Image 3: Top 150 pedestrian crossings by pedestrian volume in Toronto

As can be seen in image 3, most of the crossings of top pedestrian volume are located around central Toronto. Furthermore, it is obvious that certain streets have high volume extending out of the downtown area, for example Yonge street, which is featured prominently in the top crossings by volume.

To prepare the corresponding FourSquare data, several steps are taken.
Firstly, a formula is written that extracts the name of the category of a venue from the FourSquare object that is returned for a point of interest. The returned FourSquare object is shown in image 4, the formula in image 5, the resulting table in image 6:

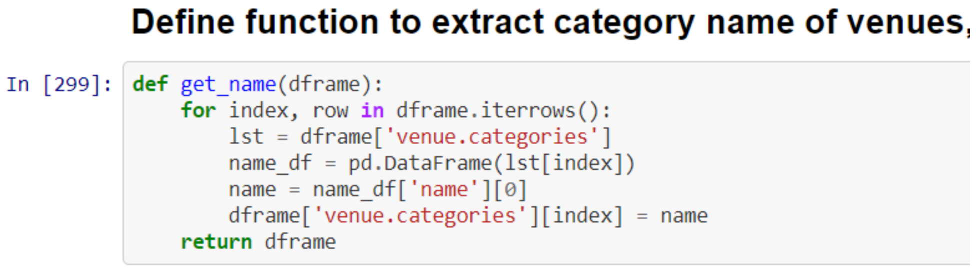


```
app.launch_new_instance()
```

Out[298]:

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Canoe	[{'id': '4bf58dd8d48988d1c4941735', 'name': 'R...	43.647452	-79.381320
1	Mos Mos Coffee	[{'id': '4bf58dd8d48988d16d941735', 'name': 'C...	43.648159	-79.378745
2	Equinox Bay Street	[{'id': '4bf58dd8d48988d176941735', 'name': 'G...	43.648100	-79.379989
3	Adelaide Club Toronto	[{'id': '4bf58dd8d48988d175941735', 'name': 'G...	43.649279	-79.381921
4	Pilot Coffee Roasters	[{'id': '4bf58dd8d48988d1e0931735', 'name': 'C...	43.648835	-79.380936

Image 4: FourSquare output in data table prior to data preparation

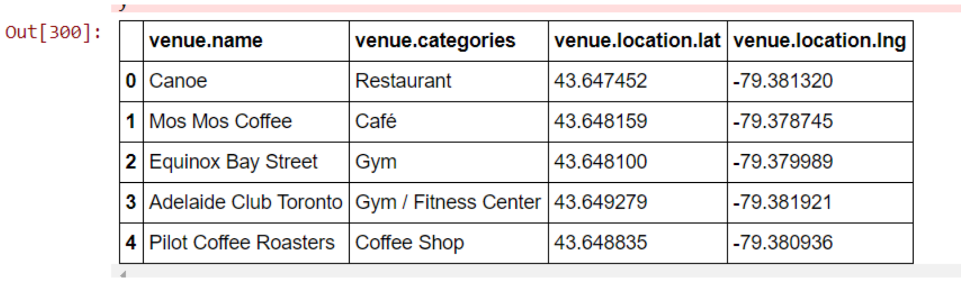


Define function to extract category name of venues,

In [299]:

```
def get_name(dframe):
    for index, row in dframe.iterrows():
        lst = dframe['venue.categories']
        name_df = pd.DataFrame(lst[index])
        name = name_df['name'][0]
        dframe['venue.categories'][index] = name
    return dframe
```

Image 5: snippet of code to extract category name



Out[300]:

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Canoe	Restaurant	43.647452	-79.381320
1	Mos Mos Coffee	Café	43.648159	-79.378745
2	Equinox Bay Street	Gym	43.648100	-79.379989
3	Adelaide Club Toronto	Gym / Fitness Center	43.649279	-79.381921
4	Pilot Coffee Roasters	Coffee Shop	43.648835	-79.380936

Image 6: FourSquare output of venues after category extraction from object

Then, the dataframe including the crossings is extended by as many columns as venues that should be attributed to the crossing. These venues are the closest venues to the crossing. 25 venues were chosen to be attributed to each crossing. Next, the code iterates through the dataframe with the 100 crossings and extracts 100 venues centered around the location of the crossing (in latitude, longitude), for each crossing. From these 100 venues, the 25 venues that are closest to the crossing are chosen and added to the crossing dataframe. The distance is computed using the following equation, where D = distance, ∆lat is difference in latitude, ∆lon is difference in longitude:

D^2 = ∆lat^2 + ∆lon^2

Now that the closest venues to each crossing are found, the data preparation is concluded.

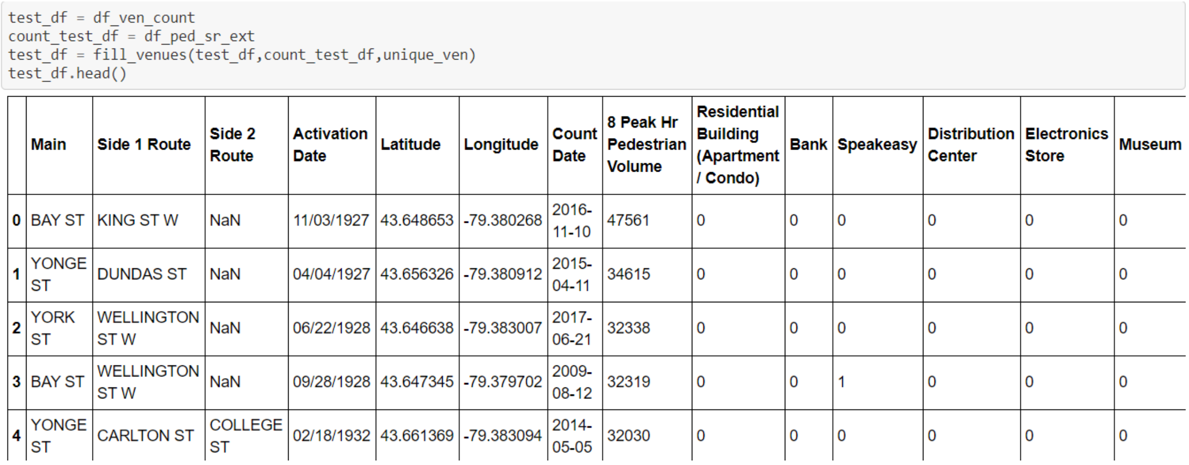
3. Methodology

The following section describes the methodology for data assessment.

As was described earlier, 150 crossing are assessed, with the closest 25 venues to each crossing being part of the assessment for correlating pedestrian volume to venue proximity. Thus, the first task is to identify what type of venue is present in the dataset and how many unique venues there are. This is crucial, since there would be very little possible correlation between the crossings and their venues if too many distinct categories were present.

Thus, first, the unique venues are extracted from the dataframe. 233 unique venue categories are found, for a possible 3750 possible unique venue categories (150*25). While this may not seem like much, there can still be a great variety across the 150 crossings in the locations that they are surrounded by. This may reduce the accuracy of any model that is built and needs to be kept in mind.

Next, the 233 unique venue categories, that are extracted and stored as a list, are added as columns to a dataframe that contains each crossing, the location, and the corresponding pedestrian traffic number. For each unique venue category column, it is counted how often the venue category occurs in the top 25 closest venues to this specific crossing. An extract of the resulting table is shown in image 7.



```
test_df = df_ven_count
count_test_df = df_ped_sr_ext
test_df = fill_venues(test_df,count_test_df,unique_ven)
test_df.head()
```

	Main	Side 1 Route	Side 2 Route	Activation Date	Latitude	Longitude	Count Date	8 Peak Hr Pedestrian Volume	Residential Building (Apartment / Condo)	Bank	Speakeasy	Distribution Center	Electronics Store	Museum
0	BAY ST	KING ST W	NaN	11/03/1927	43.648653	-79.380268	2016-11-10	47561	0	0	0	0	0	0
1	YONGE ST	DUNDAS ST	NaN	04/04/1927	43.656326	-79.380912	2015-04-11	34615	0	0	0	0	0	0
2	YORK ST	WELLINGTON ST W	NaN	06/22/1928	43.646638	-79.383007	2017-06-21	32338	0	0	0	0	0	0
3	BAY ST	WELLINGTON ST W	NaN	09/28/1928	43.647345	-79.379702	2009-08-12	32319	0	0	1	0	0	0
4	YONGE ST	CARLTON ST	COLLEGE ST	02/18/1932	43.661369	-79.383094	2014-05-05	32030	0	0	0	0	0	0

Image 7: count of venue categories around a each crossing

3.1 Clustering

Having determined which venue categories occur most frequently around a crossing, the crossings can be clustered based on these venue categories. For this, KMeans clustering is used since it is a good cluster approach for categorizing data based on the occurrence of individual numerical pointers.

To perform the clustering, all non-numerical columns are dropped from the dataframe. Next, the pedestrian volume column is dropped as well, since it is expected that it would offset the results too much, and since it is not relevant for the purely venue based clustering.
The remaining dataframe is normalized.

The actual clustering is performed by selecting a number of clusters. 5 are selected for the project at hand. The clustering is performed with random_state=0. The resulting cluster labels are re-inserted into the original dataframe with non-numeric entries.

It could be observed that most clusters at the top of the dataframe, corresponding to crossings with high pedestrian volume, are in cluster 1. This is not surprising, since it is expected that downtown venues with high pedestrian volume in close proximity are similar in type.

Nevertheless, the clusters are visualized for confirmation as before. In image 8 it can be seen that most clusters are either central or extend around the city. This confirms the earlier suspicion.

3.2 Model training

To predict pedestrian volume from venue categories in close proximity, machine learning models need to be trained.

To do so, first the features for the machine learning algorithm need to be selected. Since only the occurrence of a venue and the pedestrian volume is important for the model, all non-numeric values are dropped again.

Next, the target values, the pedestrian volume, are excluded from the dataframe to form the predictor dataframe, X. The target pedestrian volume is collected in the target dataframe y. However, y still needs to be categorized in order to work for the selected machine learning approach since continuous values can be a challenge. In order to do so, brackets are formed, into which the pedestrian volumes are sorted. 5 brackets are formed, in even distances from 0 pedestrian traffic to the next highest 10,000 over the highest pedestrian volume in the dataset (50,000 in the case at hand).

Now that y is categorized, X is prepared by preprocessing using preprocessing from the sklearn library:

```
X= preprocessing.StandardScaler().fit(X).transform(X)
```

The data is split into a test and a train partition, with a test-train split of 0.2, meaning the train set takes 80% of the data.

To increase comparability and potentially accuracy, two models are chosen instead of one:

- K nearest neighbors classifier
- Logistic regression

Both these models are suitable for categorization and modeling of numerical data. More models, such as a decision tree based model, would have been possible as well.

3.2.1 KNN model

The KNN model is started by selecting the best k for the task at hand. For this, the model is built for each k-value from 1 to 50, and the standard accuracy of the results y_hat is found when modeling for X_test is compared to y_test. K = 2 is found to yield the highest accuracy, and the model is rebuilt with k=2.

3.2.2 logistic regression

For logistic regression, the solver needs to be chosen. A Newton-cg solver is selected since it was found that it copes well with multi-categorical data while not being restricted to large datasets only. Since the results of y are not binary but subdivided into 5 brackets, this was deemed most appropriate. The model is trained with c=0.01.

3.3 Assessment of accuracy

The accuracy of the models is tested next. To not only rely on a single model evaluation, 3 approaches are taken:

- Jaccard score
- F1 score

Both jaccard and F1 score should be close to 1 ideally.

4. Results

4.1 Clustering

As was mentioned earlier, the clusters are well centred in either the central areas or around the city. This could be seen numerically, from the cluster labels as discussed before, and can also be observed from the mapping of the clusters across Toronto, as shown below.

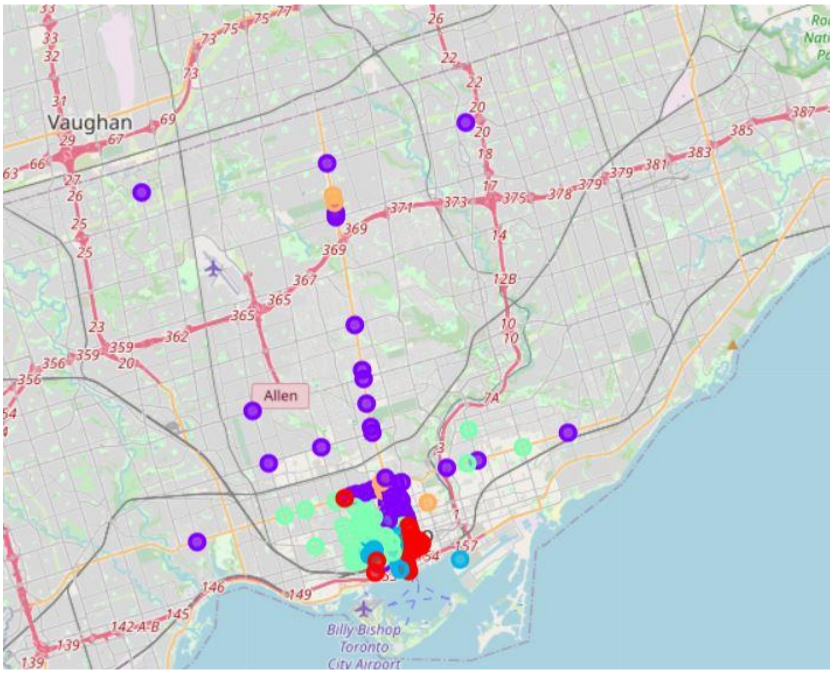


Image 8: Clusters are either centred in the city center or spread around it

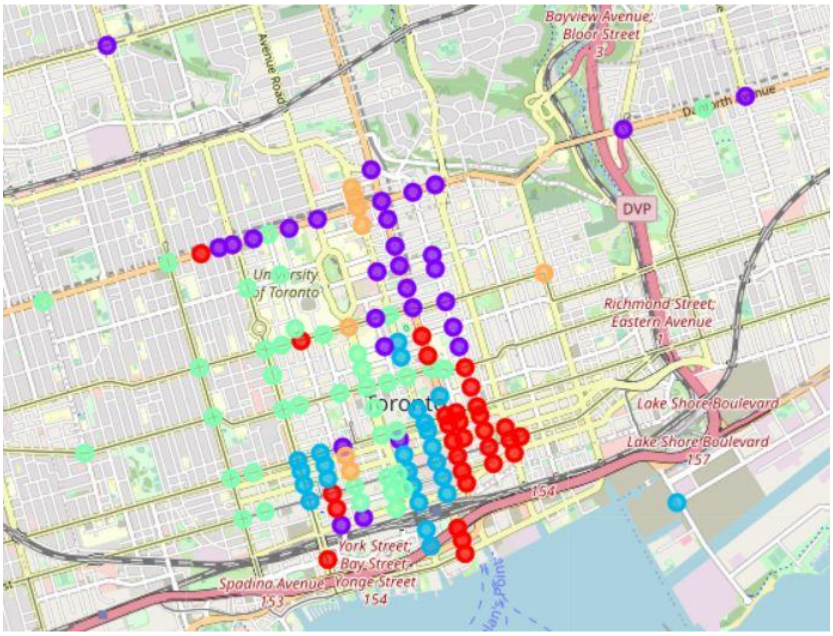


Image 9: Clusters seem to be centred around certain streets in Toronto

The centering of clusters around certain streets seen in image 9 is particularly interesting.

4.2 Machine Learning

The results from the machine learning approach are, unfortunately, not particularly promising, but fall in line with the clustering results. The following accuracies were achieved:

Jaccard score:

- KNN: 0.277
- Logistic regression: 0.132

F1 score:

- KNN: 0.43
- Logistic regression: 0.23

Since they should ideally be close to 1, this result can be considered as poor. In the next chapter it will be discussed how this result could have been affected by the data and the selection process, and what could be done better in future iterations of the models.

5. Discussion of results

As was seen in the previous section, the results are not very promising for making a strong case for using machine learning approaches on venue-category data to predict pedestrian volumes. However, there may be some possible iterations and improvements, and major factors that could have affected the outcome.

As was seen, the clustering is mainly focused around certain streets for certain clusters. This could have been the fact due to the way that the dataset was built. By taking the top 25 venues around a crossing, many crossings in the same streets will have the same locations listed as closest to them, thus naturally making them more likely to be clustered together. Furthermore, certain areas may have a higher occurrence of certain venue types. This is not necessarily a bad thing, since these certain types and crossings with certain venues close to them may attract more or fewer pedestrians. However, it is an explanation for the cluster distribution.

However, the selection of only 150 crossings and only 25 locations may have skewed the machine learning algorithm, as the dataset was rather small and because the categorization was limited. Nevertheless, it was tested whether selecting 100 crossing would have made things worse, but the results were similar. Thus, the number of crossings is not necessarily a factor for a better model. However, the increased dataset could have been an advantage.

The far bigger expected factor is the categorization of results in y, where the pedestrian volume was grouped into 5 groups. These were particularly unevenly distributed, heavily skewing the model towards low pedestrian volumes because of few outliers with many pedestrians.

2 possible improvements are suggested:

- Excluding outliers from the results dataset to ensure a more evenly trained model
- Increasing the dataset to reduce the effect of outliers

6. Conclusion

This report details the efforts to develop a machine learning model to predict pedestrian volumes in Toronto from the categories of venues close to popular pedestrian crossings. Data to do so was acquired from the website of the city of Toronto as well as FourSquare.

Venue categories were attributed to crossings by identifying the closest venues to a crossing. The resulting correlated data was clustered and two machine learning models were trained to identify pedestrian volume from the venue categories.

From the clustering it was seen that most clusters are centred around certain streets. It was determined that this is the logical conclusion from the way the data was prepared, but may well reflect reality as many similar venues are centred around certain streets as well.

The machine learning models unfortunately did not perform in a satisfying way. The highest achieved F1 score for a KNN model was at 0.43. The two key recommendations for improvements are the exclusion of pedestrian volume outliers as well as the increase of number of crossings and venues per crossing to even out outlier-effects.

Nevertheless, it can be concluded that there is some legitimacy to modeling pedestrian volume by venue categories around crossings. However, the models need further improvement.