Modeling of pedestrian traffic volume in Toronto based on venue categories in close proximity

Noel Weber
Capstone Project, Mai 2021

Predicting pedestrian traffic volume: motivation

- Predicting pedestrian volume is important for the city of Toronto
    - To manage traffic
    - To install traffic lights / other management tools
    - To ensure safety around the city

- Pedestrian volumes may change with local offers; as neighborhoods change and new venues are introduced, pedestrian traffic may change as well

- Thus, predicting the volume is crucial for continuous traffic management

Data acquisition

- Two sources
  - Database of the city of Toronto: https://open.toronto.ca/dataset/traffic-signal-vehicle-and-pedestrian-volumes/
    - 2018 Data collection on traffic volume at crossings in Toronto
  - FourSquare Data, set to similar timeframe

- FourSquare data was acquired separately for each crossing under investigation, with 100 venues limit per crossing

Data preparation

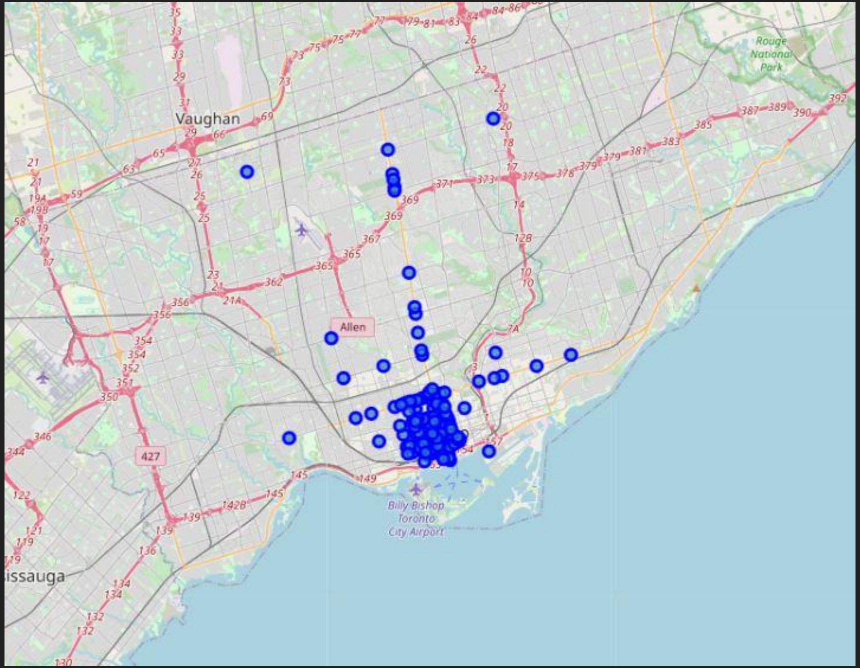
- Vehicle volume and date of data collection was ignored

- 2280 crossings were included in the dataset (rows) with 11 columns each
  - The 150 most popular crossings were considered

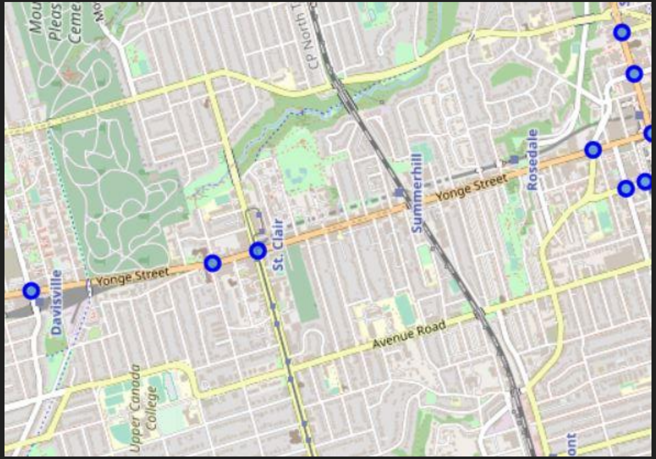
- The 25 closest venues to each crossing and their categories were considered

- 233 unique venue categories were found, making for 233 features in the cleaned dataset

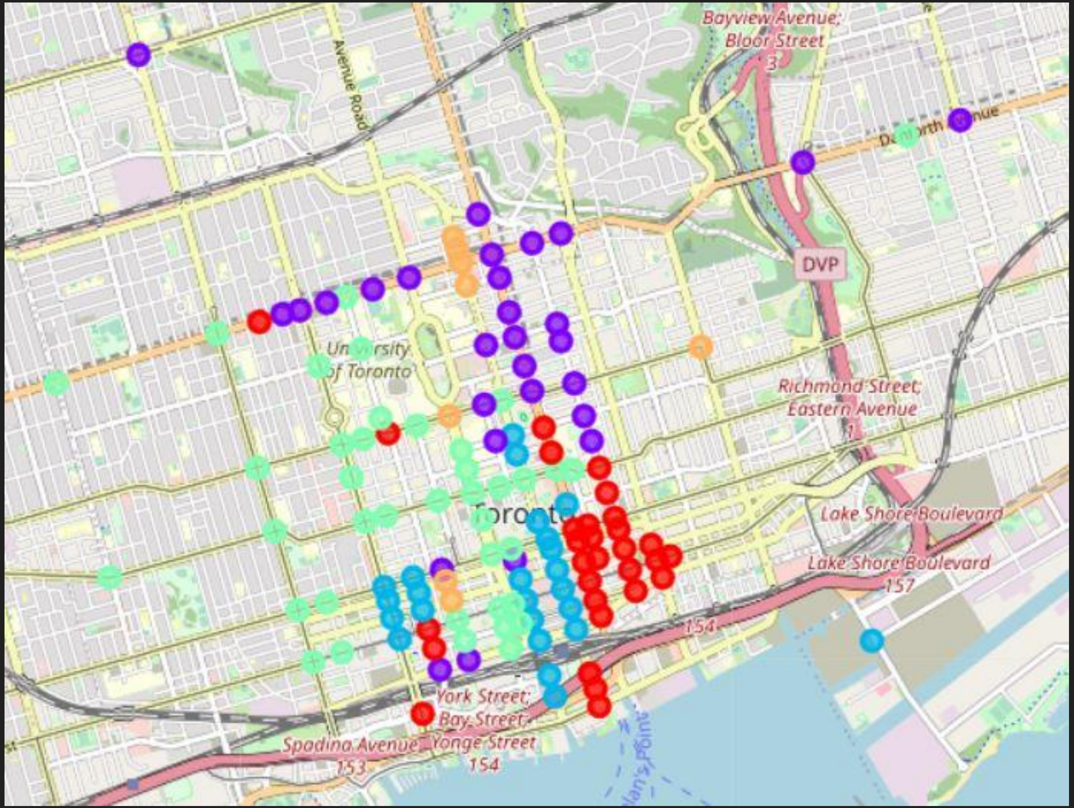
The crossings with the highest pedestrian traffic can be found in downtown Toronto, only few outliers



Most popular crossings can be found along few main streets



Clustering of crossings based on venues



- Crossings are clustered into 5 possible clusters based on the occurrence of venue categories in close proximity
- Most clusters centred around certain streets
- Does not impede accuracy of results

Machine learning approach

- Two approaches:
  - KNN model
  - Logistic regression

- The target variable (pedestrian volume) was subdivided into 5 brackets
  - Possible outliers may cause distortion

Machine learning insights

- Poor model accuracy

- Jaccard score:
  - KNN: 0.277
  - Logistic regression: 0.132
- F1 score:
  - KNN: 0.43
  - Logistic regression: 0.23

Explanation:

- Small data sets (crossing)

- Division of target variable too affected by outliers

Improvement opportunities

1. Excluding outliers from the results dataset to ensure a more evenly trained model

2. Increasing the dataset to reduce the effect of outliers