# Introduction

## Rationale

* Ottawa is set to open a new light rail transit system: Confederation Line
* Metro stations along the line will now see a greater among of commuter traffic
* Will the current mix of venues at transit stations properly service this increase in commuter traffic?
  + We can compare the Ottawa mix of venues to those of a more established transit system, that of Toronto Ontario
  + Does the mix of venues at Toronto’s transit stations comparable to the mix of venues at Ottawa’s transit stations
* Foursquare consists of a large database of venues in cities across the world
  + Using the Foursquare API we can list all the venues within walking distance of each metro station along the Confederation Line and the Union Line
* Multivariate analysis can be used to compare whether the mix of venues at the metro stations are comparable between Ottawa and Toronto and identify any significant differences
* The results of this study could highlight needed venues along the Confederation Line, giving small business owners valuable insight

# Data Description

* A list of metro stations for both the Confederation Line and a comparable selection of stations from Line 1 in Toronto were found on [www.wikipedia.org](http://www.wikipedia.org)
* The Foursquare API has a detailed venue category labeling scheme. We accessed the parent categories and each of their sub-categories from (url). This was used to translate sub-categories back to their parent categories
* The Foursquare dataset was used to collect all venues within walking distance of each metro station, specifically their category

# Methodology

* Geographical coordinates for each station were assigned using the Geopy library
* Venue category labeling is quite detailed and specific in the Foursquare API it was thus necessary to use the highest level categories (or parent categories) in order for the Ottawa and Toronto venues to be comparable
  + The Foursquare venue category tree was accessed via (url)
  + I created a table listing all subcategories available in the Foursquare API and added a column showing the parent category it belonged to
* For each station, the Foursquare API was used to find all venues within 800 m of the metro station, this is the average distance someone can walk in 10 minutes.
  + The venue subcategories returned by the Foursquare API were then translated to their parent category
  + For example, ‘Sun Life Financial Centre’ was reverted to ‘Professional & Other Places’
  + The venues were grouped summed based on parent category.
  + The final table listed Stations as rows and the number of venues that belonged to each parent category as columns
* Hierarchal Clustering was performed on the the venue counts for each station
  + The counts were converted to percentages (i.e., percent of venues that belong to the food category).
    - This controlled for the greater population of Toronto
  + The data was converted to an array and normalized
  + It was than transformed into a distance matrix
  + Complete linkage hierarchical clustering was used
  + The visualised clusters were inspected for ideal cut-off and total number of clusters
* A decision tree was used to investigate the importance of different categories on the selection of clusters
  + The percentage of venues at each station were used as the feature set
  + The cluster number that the station belonged to was used as the labels for the feature set

# Results

* Table 01. Station Venue Counts
  + Station, City, Lat, Lon, Venues, Total
  + The number of venues varied 85 and 200
  + A total of 1825 venues were returned for the 12 Ottawa with a mean of 152
  + A total of 2510 venues were returned for the 14 Toronto stations with a mean of 179
  + Two categories were removed (Events, Residence) because greater than 25% of their values were NaNs.
* Figure 01. Box Plots of Venue Categories
  + Four categories have less than 10% variation
    - Outdoors & Recreation; Nightlife Spot; College & University; Arts & Entertainment
    - Apart from Nightlife Spot all the categories contain outliers; with College & University having an extreme outlier
  + Professional & Other Places – has the greatest amount of variation amongst stations
    - This is followed by Shop & Service, then by Travel & Transport
  + The box plots show several outliers, yet the majority fall within the distributions of all venues; apart from an outlier in Shop & Service and College & University
    - Those were the only stations to exceed 55% in one venue category
* Figure 01. Cluster Dendrogram
  + Cutoff was set to create seven clusters
* Table 02. Station Venue Percent (sorted by cluster)
  + Station, City, Cluster, Venues
  + Cluster 0 (n = 9; O-2; T-7)
  + Cluster 1 (n = 2; O-1; T-1)
  + Cluster 2 (n = 7; O-4; T-3)
  + Cluster 3 (n = 1; O-1; T-0)
  + Cluster 4 (n = 5; O-3; T-2)
  + Cluster 5 (n = 1; O-1; T-0)
  + Cluster 6 (n = 1; O-0; T-1)
  + More than 50% of the Toronto stations were found in one cluster (0).
    - Toronto stations were absent in two clusters (3 & 5).
  + Ottawa stations were present in all clusters apart from Cluster 6
    - Cluster 2 had the highest number of with Ottawa Stations (n = 4)
* Figure 02. Decision Tree
  + Node 0 (n = 8; Cluster 0) [Professional & Other Places > 34.5]
  + Node 1 (n = 5; Cluster 4) [Travel & Transport > 15.5]
  + Node 2 (n = 6; Cluster 2) [Travel & Transport > 6.5]
  + Node 3 (n = 3; Cluster 1, 5) [Shop & Service > 27.5]
  + Node 5 (n = 3; Cluster 0, 3, 6) [College & University > 12]

# Discussion

* Characterising Groupings
  + NODE 0 – is the first node to be resolved by the decision tree model
    - It contained 8 members of CLUSTER 0
    - It is defined by a “Professional & Other Place” cutoff of 34.5
    - One member of CLUSTER 0 (St-George) is absent from NODE 0; as its percent of “Professional & Other Place” is below the cutoff (33%).
      * In the decision tree it groups with two other stations that also have relatively high “College & University” venues: Museum (23%) and uOttawa (64%). St-George has a value of 21%
    - CLUSTER 0 contains a low number of Ottawa stations (n = 2) compared to Toronto stations (n = 7).
      * This cluster is defined by “Food” (u = 18); “Professional & Other Places” (u = 44); “Shop & Service” (u = 16);
      * **This might highlight a need for a greater number of “Professional & Other Service” type venues at Ottawa stations. In this way they would be more in line with Toronto Stations**
  + NODE 1
    - Contained all five members of CLUSTER 4
    - Defined by a “Travel & Transport” cutoff of 15.5%
    - CLUSTER 4 had an even mix of Ottawa (n = 3) and Toronto (n = 2) stations
  + NODE 2
    - Contained six members of CLUSTER 2
    - Defined by a “Travel & Transport” cut off of 6.5%
    - One member of CLUSTER 2, Dundas Station, did not make it into NODE 2
      * It forms its own NODE 4; distinguished from NODE 5 & 6 by having low “College & University”
    - CLUSTER 2 contains an even mix of Ottawa (n = 4) and Toronto (n = 3).
      * This cluster contains a even spread between “Food” (u = 22); “Professional & Other Places” (u = 23); “Shop & Service” (u = 31);
  + NODE 3
    - Contained one member of CLUSTER 5, and two members of CLUSTER 1
    - Defined by a “Shop & Service” cut off of 27.5%
    - CLUSTER 5 contains only one member: St-Laurent Station
      * This is one of the two extreme outliers observed in the box plots; with a value of 73% for shops and services
      * **The outlier status of this station might highlight the need for a greater percentage of other types of venues**
    - CLUSTER 1 contains only two members: Rideau and Queen Station
      * Like St-Laurent these stations have high percentages of “Shop & Service” 30% and 52% respectively.
      * Rideau Station is unique in having a very high percentage of “Outdoor & Recreation” venues at 28%.
        + Apart from this these two stations are identical
  + NODE 5
    - Contained one member of Cluster 0, one member of Cluster 3, and one member of Cluster 6
    - Defined by a “College & University” cut off of 12% [this could be greater]
    - CLUSTER 0 contains nine members; with eight falling within NODE 0 and being defined by high percentages of “Professional & Other Service”
      * St-George Station has relatively high “Professional & Other Service” but is unique within CLUSTER 0 in having “College & University” at 21%. The next closest is Queen’s Park at 11%
    - CLUSTER 3 contains only one member; it is one of the extreme outliers identified in the boxplot with a “College & University” at 64%
      * **Highlights the need for a greater diversity of venues needed at this station**
    - CLUSTER 6 contains only one member, Museum Station,
      * It has a “College & University” of 23%, but is also defined by an “Arts & Entertainment” of 24%.
        + The next closest station is Pimisi with a value of 15%, found in CLUSTER 2
        + **The lack of any Ottawa station in this cluster might highlight a need for a station that gives access to more of these types of venues**
  + Summary
    - Two CLUSTERS had the majority of stations, CLUSTER 0 and CLUSTER 2, these stations had a relatively balanced amount of venues fitting into “Food”, “Professional & Other Places” and “Shop & Service”.
      * Toronto stations tended to have venues with higher percentages of “Professional & Other Places”, reflected by only two out of nine stations being Ottawa stations in CLUSTER 0

# Conclusion