DATA SECTION REPORT

- **Clustering the Districts:**

Finally, we try to cluster these 5 districts based on the venue categories and use dbscan clustering. So our expectation would be based on the similarities of venue categories, these districts will be clustered.

DBSCAN offers several advantages over k-means clustering when it comes to categorising geographical data. Firstly, k-means minimises variance rather than geographical distance. Further away from the equator (at high-latitudes), there is substantial distortion amongst groupings. The DBSCAN algorithm works better with arbitrary distances and can be implemented into scikit-learn's version of the DBSCAN algorithm. DBSCAN clusters a spatial data set based on two parameters: a

physical distance from each point, and a minimum cluster size. This method works much better for spatial latitude-longitude data. Additionally, DBSCAN generates the number of cluster for us, rather than requiring us to specify the number as with the k-means approach.

The DBSCAN method first calculates the centroid's coordinates. Then I use Python's built-in min function to find the smallest member of the cluster in terms of distance to that centroid. The key argument does this with a lambda function that calculates each point's distance to the centroid in meters, via geopy's great circle function. Finally, the coordinates of the point that was the least distance from the centroid are returned to us. To use this function, we map it to series of clusters from out pandas dataframe. In other words, for each element (i.e., cluster) in the series, it gets the center-most point and then assembles all these center-most points into a new series called *centermost_points*. Then I turn these center-most points into a pandas dataframe of points which are spatially representative of my clusters (and in turn, the full data set.

- **Results and Discussion:**

We reached at the end of the analysis, where we got a sneak peak of the 5 major wards of Toronto and, as the business problem started with benefits and drawbacks of opening a lunch restaurant in one of the busiest districts, the data exploration was mostly concentrated on the restaurants. I have used data from web resources like Wikipedia, python libraries like Geopy, and Foursquare API, to set up a very realistic data-analysis scenario.

The result of this analysis means that each venue in Toronto is assigned to a specific DBSCAN cluster, which can then be provided to our client so that an optimal delivery service can be devised. Whilst the final implementation of this will be up to the client, we can recommend that they assign delivery riders/drivers to a specific cluster. A full presentation of all venues is not really feasible however as there are too many to list visually. Instead, these will have to be maintained in some form of database this information can be updated. Any new venues, or changes to existing venues can then be re- categorised in the DBSCAN model and added to the client's venue roster. Finally and of note is that the central region of Toronto (or more specifically, the central cluster) is

somewhat isolated from the rest of the city. This is due to this area being the most densely populated.

•Conclusion

Outcome of this job will be the function wich use Forsquare API data, geo data for particular city and category of veneu Customer want to analyze and find best place. I can try this function to find best places in Toronto.