

# Applied Data Science Project

Nades Nadeswaran

15<sup>th</sup> Oct 2018

# Executive Summary

- Project focused on solving the problem of  
“ Enabling Vegans on AI & Data Science Training in NY & Toronto with valuable insights of Boroughs & Neighborhoods that are popular for them ”
- 3 Trusted Data Sources (2 Publicly available plus Foursquare DB) were leveraged for data understanding, mapping, modelling, mining, grouping, extracting, and concluding etc.
- Results & Findings
  - NY has 3 very popular neighborhoods in same borough (= Manhattan) for Vegans
  - Toronto has one very popular neighbourhood (=Parkdale Village) in the borough of West Toronto
- The approach and methodology are limited in scope given the project timeline and can be further expanded as appropriate

# “Problem Statement” Description

IT companies world-wide continue to face the challenge of shortages in skilled and productive Data Scientists to be deployed for AI, Machine Learning, & Deep Learning related projects that can truly deliver technology driven business value to enterprises and institutions globally.

New-Hire Data Scientists (DS) from India need to travel to NA (NY / Toronto) for 6 month in-house training as part of their DS Orientation.

Majority of the New DSs are Vegans and their expectations and requirements need to be respected and addressed by mgmt.

**" How can IT companies help their overseas Vegan staff to choose between which Neighborhoods in NY / Toronto to live during their 6-months training ? "**

# Trusted Data Sources and how they will be used ....

The following trusted data sources will be used to conduct Data exploration and other forms of data analysis as part of the methodology for the project.

- 1) NY Neighborhood Data - <https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json> & [https://geo.nyu.edu/catalog/nyu\\_2451\\_34572](https://geo.nyu.edu/catalog/nyu_2451_34572)
- 2) Toronto Neighborhood Data - '[https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)'
- 3) Foursquare Location Data (Venues [Restaurants, Community Centers], TOP tips, Favorites, User Experience, etc) will be used to cluster, segment, target, and position to craft recommendations for the vegans end-user community.

Using the data available in the above 3 trusted sources, I will be conducting clustering and neighborhood based analysis leveraging primarily Foursquare APIs and tools such as KNN and relevant unsupervised learning methods to deliver recommendations options to the target user community primarily comprising of the newly hired vegans.

# Methodology

The methodology employed is aligned with the Data Science 10 steps program that was discussed during an earlier module of the course.

Let me outline below the core areas of focus and tasks performed as part of the methodology to address the problem of "helping the Vegans" for making the right choices regarding their training centers in NY and Toronto.

- a) Obtain relevant vegan places data from Foursquare and clean it for data understanding and grouping etc.
- b) Explore the data for clusters & patterns of Neighborhoods in Boroughs
- c) Group the places of high interest into relevant neighborhood and Borough pairings
- d) Generate the markers on both NY & Toronto maps to highlight the neighborhoods that are in focus.
- e) Use Foursquare venue, Categories etc to enable the profiling the primary & focus neighborhoods.
- f) Generate the short-listed neighborhoods for both NY & Toronto

# Results

- **Here is the Final short-list of neighborhoods for Vegans in NY**
  - **neighborhood**
    - Chelsea and Clinton 6
    - Gramercy Park and Murray Hill 1
    - Greenpoint 2
    - Greenwich Village and Soho 5
    - Lower East Side 2
    - Northwest Brooklyn 3
    - Upper East Side 1
- **It was clear that Greenwich, Chelsea, and Northwest were highly favored for Vegans in NY**
- **Get the TOP 3 Vegan Places in Toronto**
  - **neighborhood**
    - Bloorcourt 1
    - East Toronto 1
    - Kensington Market 1
    - Parkdale Village 3
- **It was clear that Parkdale Village is highly favored for Vegans in Toronto**[1](#)

# Discussion..

While this project was able to derive specific recommendations in terms of places to live in NY & Toronto for Vegans, it has limitations given that the assumptions for the model are simplistic in nature.

The main variable of focus was only Food & Leisure where as Vegans may be interested in a number of other amenities and interests that will make it a multi-variable focused data science analysis.

From my perspective the project has afforded me a tremendous opportunity to explore and apply most of the knowledge and insights gained during the previous 8 modules of this course.

# Summary & Conclusion

- It is very evident that a Data Scientist will need to establish clarity and crisply identify the "Business Problem" or "Idea of Research" that he/she is undertaking and be able to articulate clearly the value being delivered to the beneficiaries / stakeholders / end-users etc.
- Once the Business Problem / Idea has been confirmed, its also equally important to explore / research for the trusted data sources that can be leveraged to undertake the Data Science framework based problem solving.
- Next comes the critical step of choosing the methods (Unsupervised or Supervised and within each of them the specific methods for machine learning / model dev't and training etc ) for data analysis/mining/patterns recognition etc...
- It is also vital that relevant presentation and visualization techniques are used to share and present the findings to the project sponsor / stakeholders / end-users etc. It may be also appropriate to comment any future work that may be valuable to augment the solution and enrich it.