CAPSTONE PROJECT : RECOMMENDING BEST NEIGHBORHOOD IN TORONTO FOR OPENING A RESTAURANT

Applied Data Science Capstone Project : The Battle Of Neighborhoods

IBM Data Science Professional Certificate

- Avinandan Mukherjee

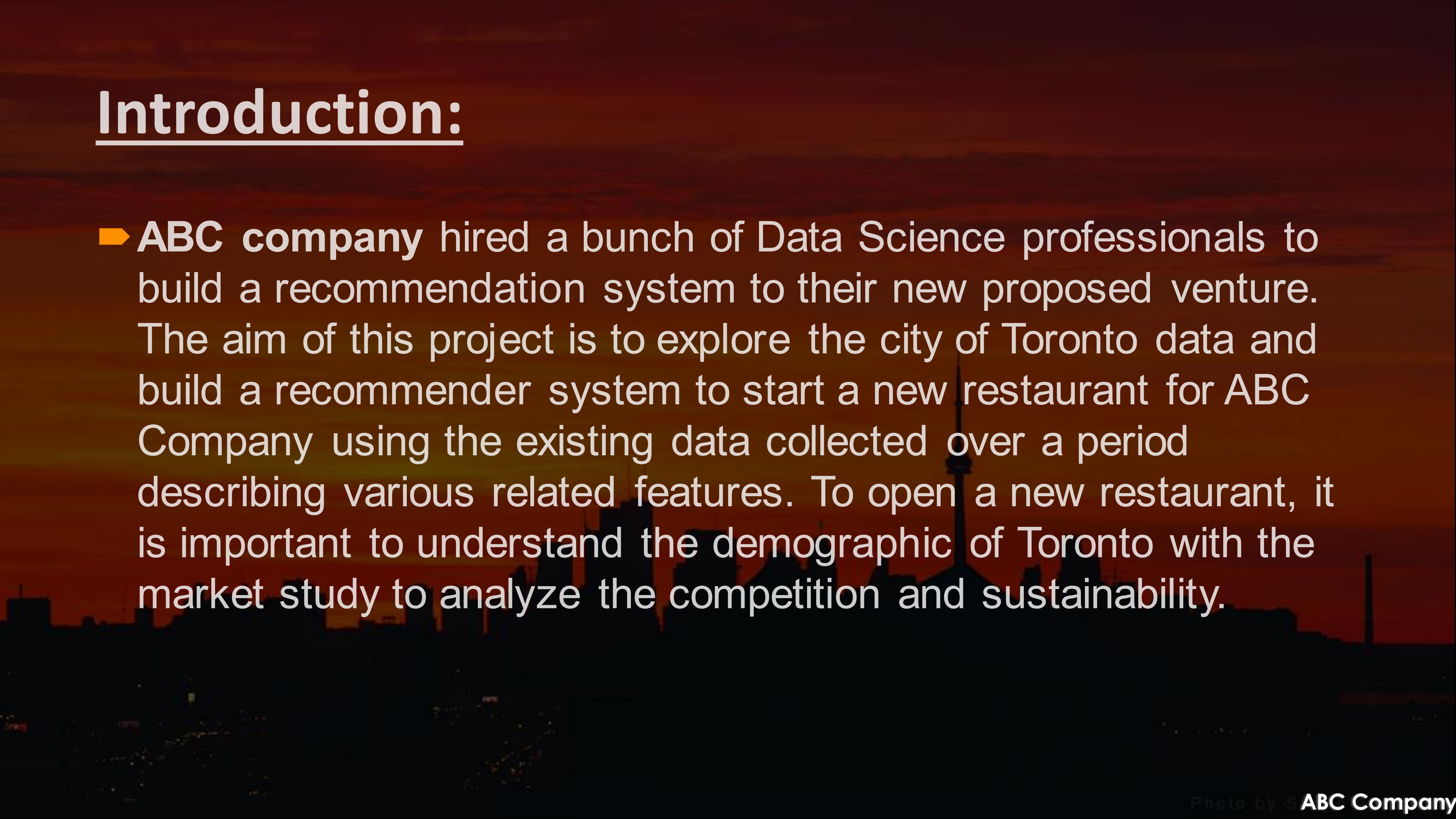
ABC Company

Agenda:

- Introduction
- Problem Description
- Data
  - Data Description
  - Data Cleaning & Data Wrangling
  - Segmenting & Clustering
- Methodology
- Result
- Discussion
- Conclusion

ABC Company

Introduction:



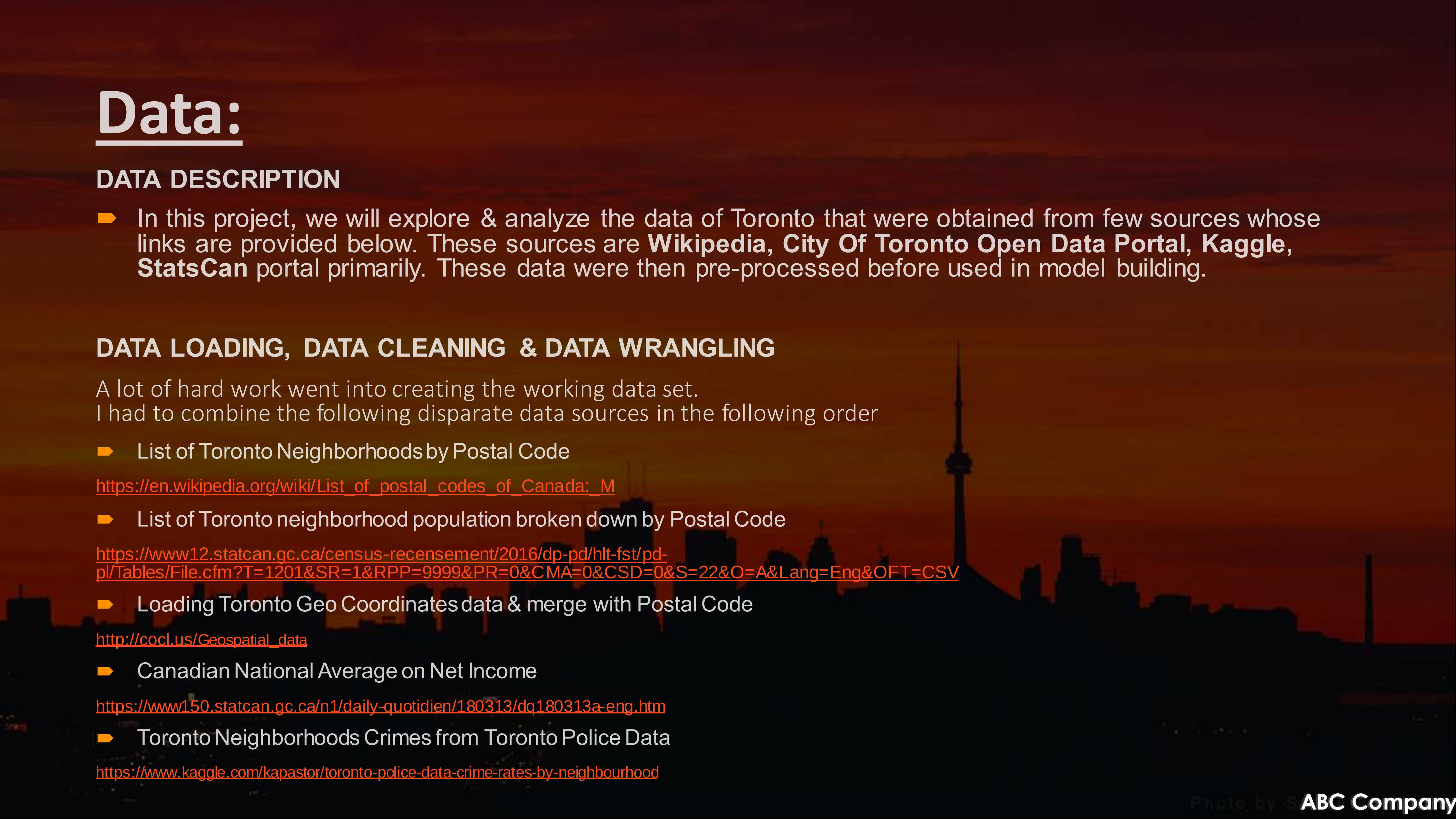
- **ABC company** hired a bunch of Data Science professionals to build a recommendation system to their new proposed venture. The aim of this project is to explore the city of Toronto data and build a recommender system to start a new restaurant for ABC Company using the existing data collected over a period describing various related features. To open a new restaurant, it is important to understand the demographic of Toronto with the market study to analyze the competition and sustainability.

Problem Description:

- Restaurants needs to be located in nicely populated area with a neighborhood.
- Restaurants needs to be located in a neighborhood where the criminal activities are lower or negligible.
- Restaurants needs to be located strategically with the existing competition in the market for a sustainable growth.
- Restaurants needs to be located in a neighborhood with above average net income of the nation.

SUCCESS CRITERIA

- Location of the restaurant with the neighborhood
- Measure existing competition by clustering common restaurants by neighborhoods
- Mean income of the neighborhood with Canadian National Average
- Population by neighborhood
- Crime in the neighborhood



Data:

DATA DESCRIPTION

- In this project, we will explore & analyze the data of Toronto that were obtained from few sources whose links are provided below. These sources are **Wikipedia, City Of Toronto Open Data Portal, Kaggle, StatsCan** portal primarily. These data were then pre-processed before used in model building.

DATA LOADING, DATA CLEANING & DATA WRANGLING

A lot of hard work went into creating the working data set.
I had to combine the following disparate data sources in the following order

- List of Toronto Neighborhoods by Postal Code

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- List of Toronto neighborhood population broken down by Postal Code

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV

- Loading Toronto Geo Coordinates data & merge with Postal Code

http://cocl.us/Geospatial_data

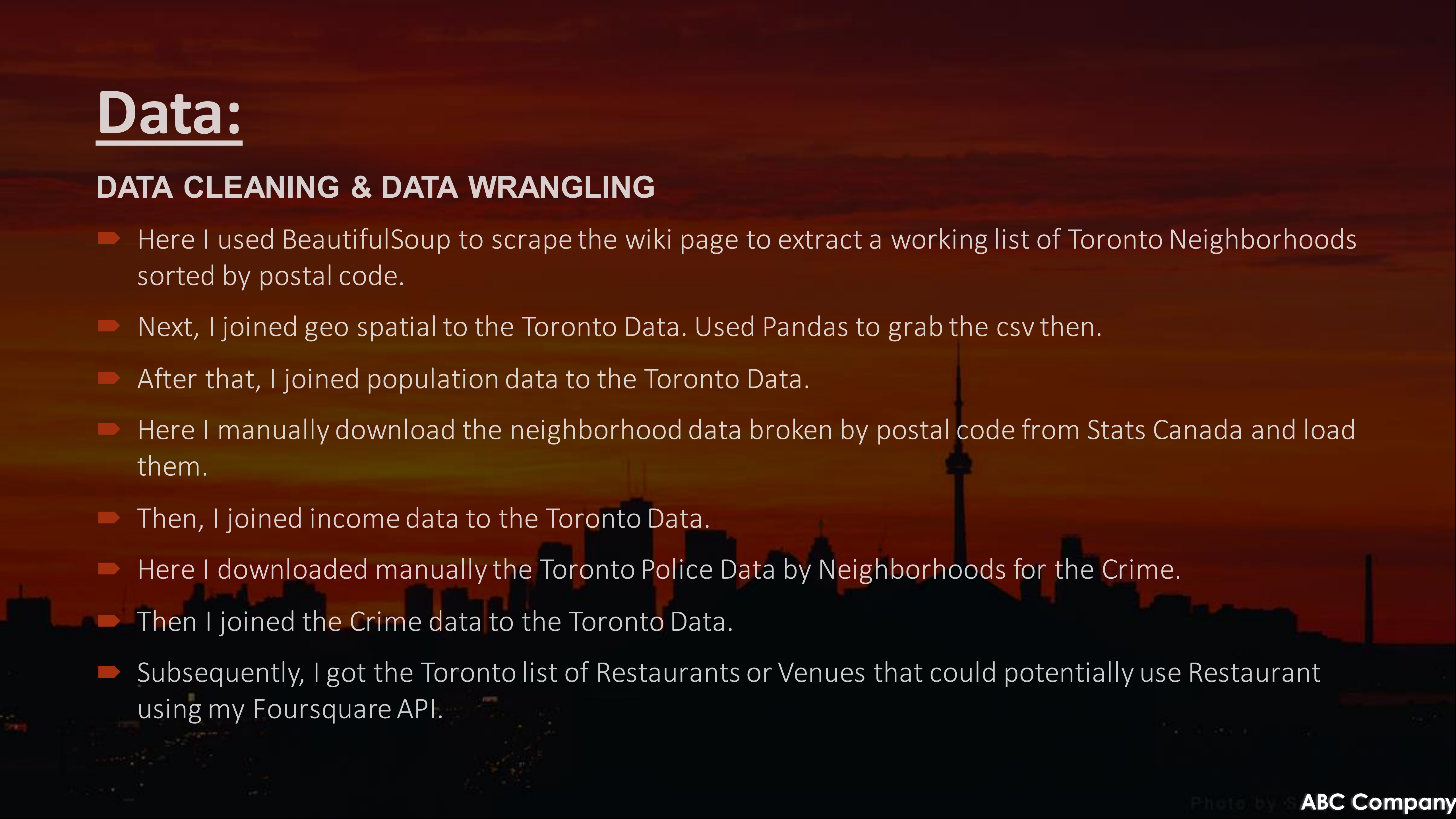
- Canadian National Average on Net Income

https://www150.statcan.gc.ca/n1/daily-quotidien/180313/dq180313a-eng.htm

- Toronto Neighborhoods Crimes from Toronto Police Data

https://www.kaggle.com/kapastor/toronto-police-data-crime-rates-by-neighbourhood

Photo by S



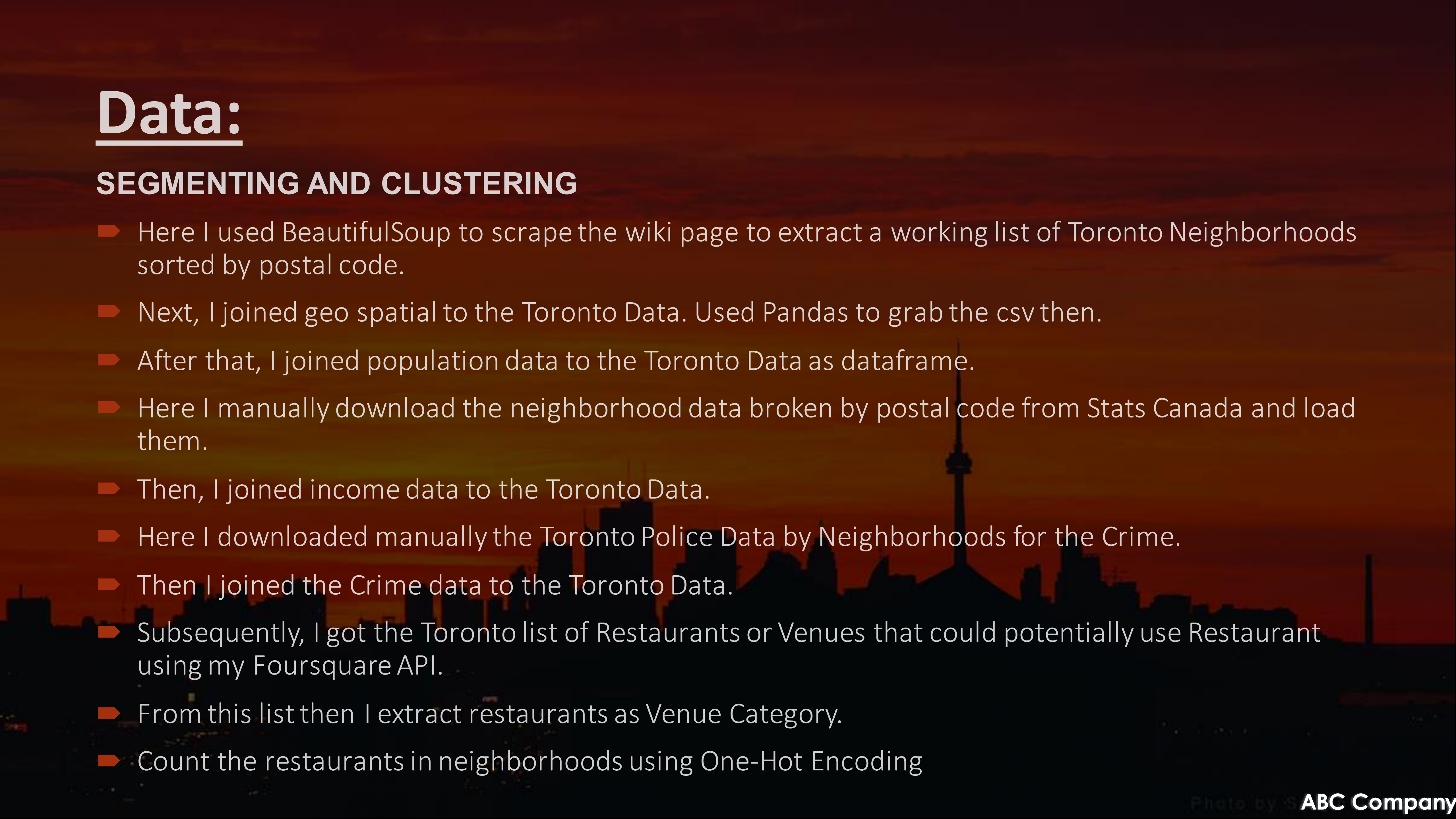
Data:

DATA CLEANING & DATA WRANGLING

- Here I used BeautifulSoup to scrape the wiki page to extract a working list of Toronto Neighborhoods sorted by postal code.
- Next, I joined geo spatial to the Toronto Data. Used Pandas to grab the csv then.
- After that, I joined population data to the Toronto Data.
- Here I manually download the neighborhood data broken by postal code from Stats Canada and load them.
- Then, I joined income data to the Toronto Data.
- Here I downloaded manually the Toronto Police Data by Neighborhoods for the Crime.
- Then I joined the Crime data to the Toronto Data.
- Subsequently, I got the Toronto list of Restaurants or Venues that could potentially use Restaurant using my Foursquare API.

Photo by S

ABC Company



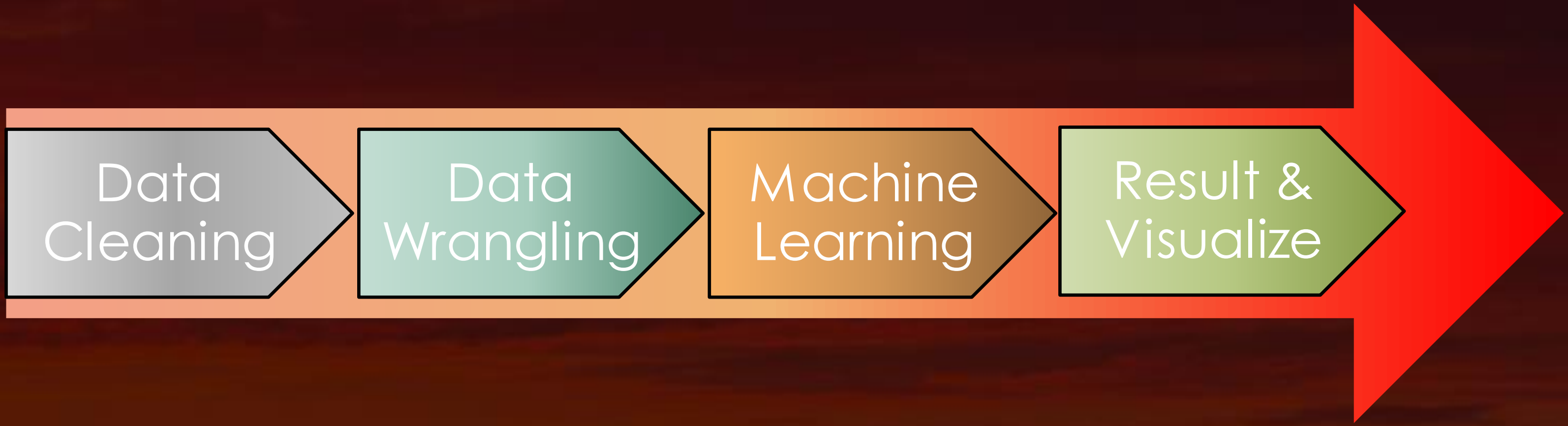
Data:

SEGMENTING AND CLUSTERING

- Here I used BeautifulSoup to scrape the wiki page to extract a working list of Toronto Neighborhoods sorted by postal code.
- Next, I joined geo spatial to the Toronto Data. Used Pandas to grab the csv then.
- After that, I joined population data to the Toronto Data as dataframe.
- Here I manually download the neighborhood data broken by postal code from Stats Canada and load them.
- Then, I joined income data to the Toronto Data.
- Here I downloaded manually the Toronto Police Data by Neighborhoods for the Crime.
- Then I joined the Crime data to the Toronto Data.
- Subsequently, I got the Toronto list of Restaurants or Venues that could potentially use Restaurant using my Foursquare API.
- From this list then I extract restaurants as Venue Category.
- Count the restaurants in neighborhoods using One-Hot Encoding

Methodology:

MACHINE LEARNING



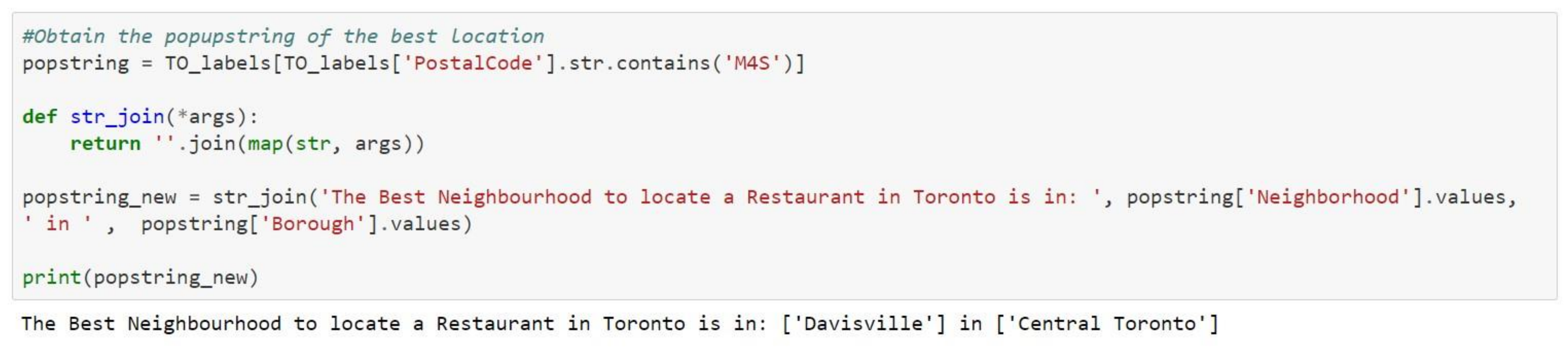
□ Algorithm Choice

- "K-Means clustering is an iterative clustering algorithm where the number of clusters K is predetermined and the algorithm iteratively assigns each data point to one of the K clusters based on the feature similarity." Here feature similarity refers to restaurant's similarity.
  Reference Link

□ Choosing the correct number of clusters

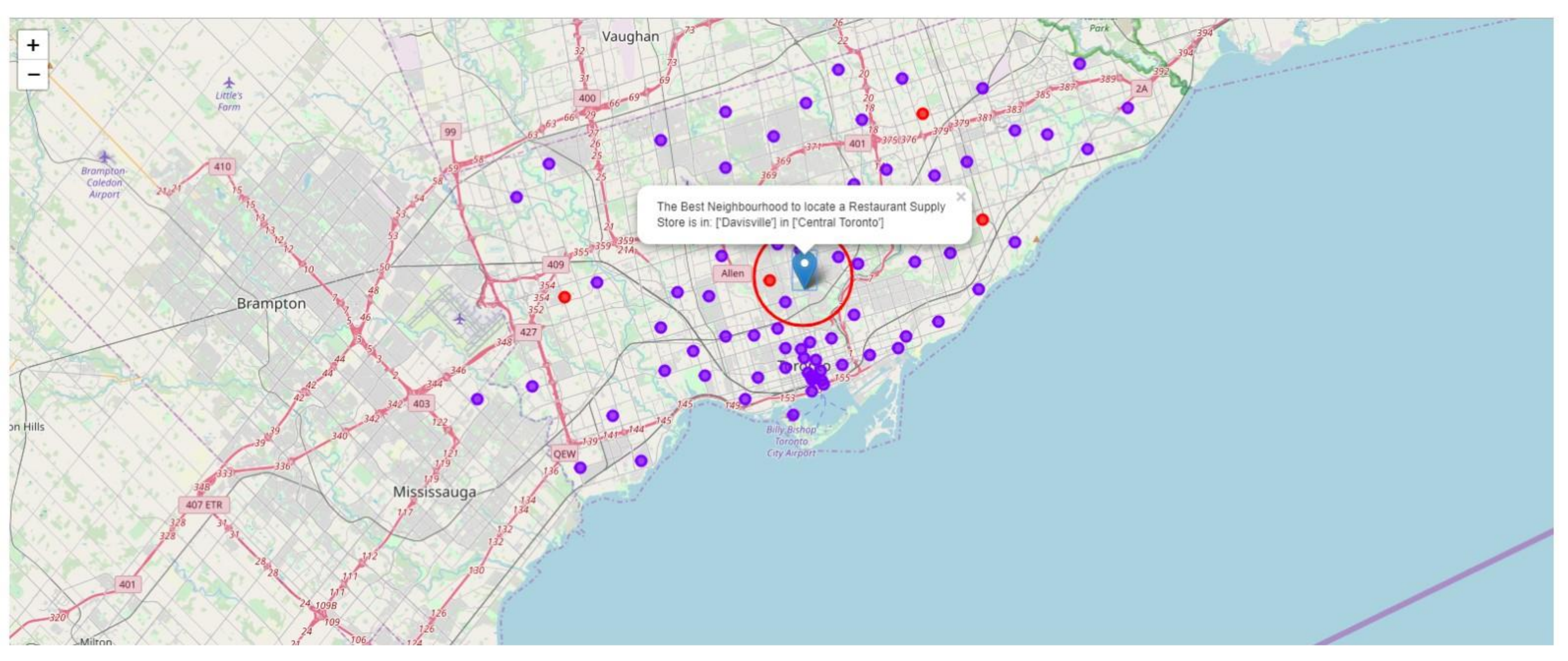
- We can use Silhouette analysis to evaluate each model. A Silhouette coefficient is calculated for observation, which is then averaged to determine the Silhouette score. The coefficient combines the average within-cluster distance with average nearest-cluster distance to assign a value between -1 and 1. A value below zero denotes that the observation is probably in the wrong cluster and a value closer to 1 denotes that the observation is a great fit for the cluster and clearly separated from other clusters. This coefficient essentially measures how close an observation is to neighboring clusters, where it is desirable to be the maximum distance possible from neighboring clusters. We can automatically determine the best number of clusters, k, by selecting the model which yields the highest Silhouette score.
  My Silhouette score was 2.
  Reference Link

Result:

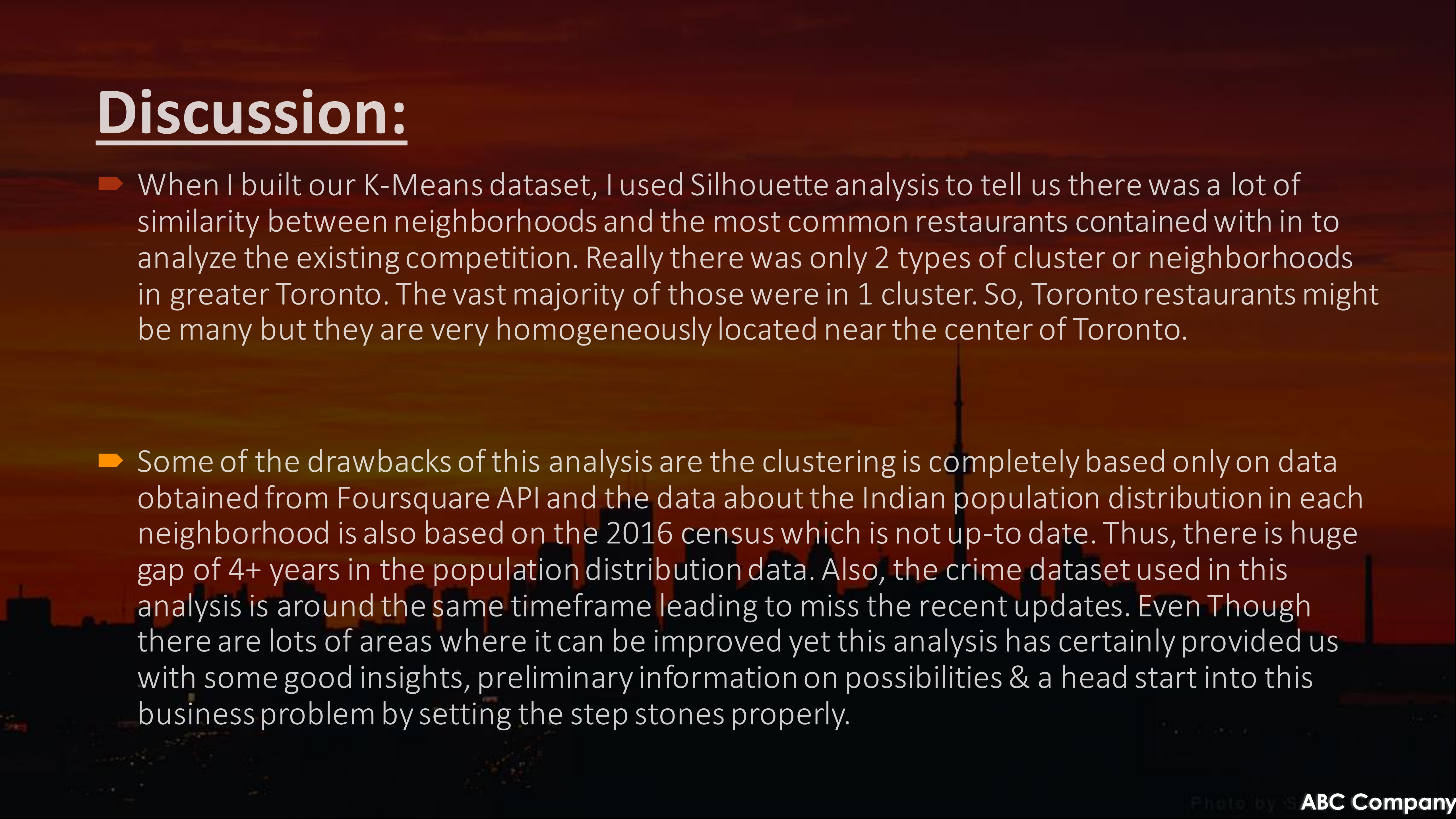


```
#Obtain the popupstring of the best location
popstring = TO_labels[TO_labels['PostalCode'].str.contains('M4S')]

def str_join(*args):
    return ''.join(map(str, args))

popstring_new = str_join('The Best Neighbourhood to locate a Restaurant in Toronto is in: ', popstring['Neighborhood'].values,
' in ' ,  popstring['Borough'].values)

print(popstring_new)
```

```
The Best Neighbourhood to locate a Restaurant in Toronto is in: ['Davisville'] in ['Central Toronto']
```



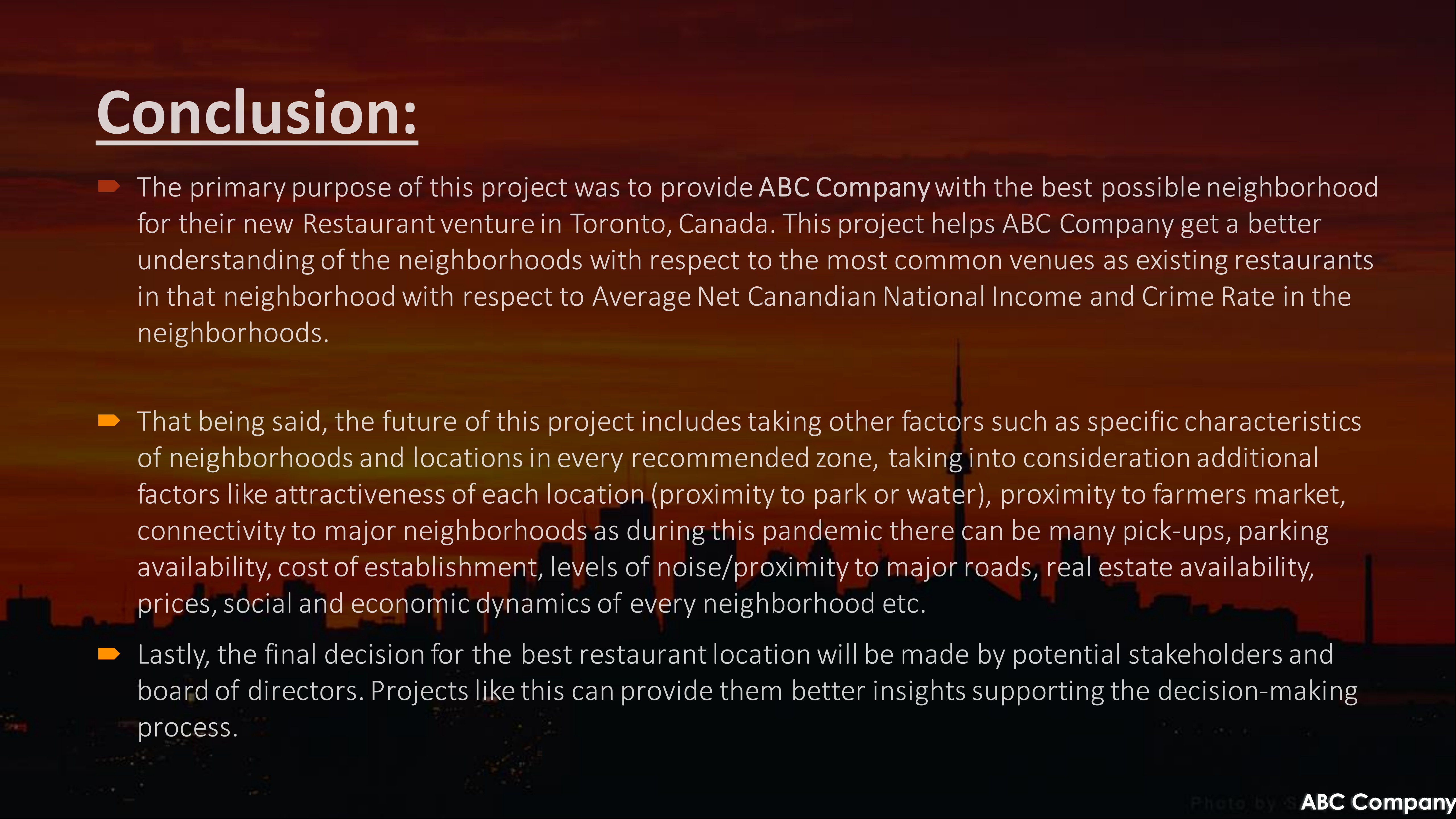
The Best Neighbourhood to locate a Restaurant Supply Store is in: ['Davisville'] in ['Central Toronto']



Discussion:

- When I built our K-Means dataset, I used Silhouette analysis to tell us there was a lot of similarity between neighborhoods and the most common restaurants contained with in to analyze the existing competition. Really there was only 2 types of cluster or neighborhoods in greater Toronto. The vast majority of those were in 1 cluster. So, Toronto restaurants might be many but they are very homogeneously located near the center of Toronto.

- Some of the drawbacks of this analysis are the clustering is completely based only on data obtained from Foursquare API and the data about the Indian population distribution in each neighborhood is also based on the 2016 census which is not up-to date. Thus, there is huge gap of 4+ years in the population distribution data. Also, the crime dataset used in this analysis is around the same timeframe leading to miss the recent updates. Even Though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.



Conclusion:

- The primary purpose of this project was to provide **ABC Company** with the best possible neighborhood for their new Restaurant venture in Toronto, Canada. This project helps ABC Company get a better understanding of the neighborhoods with respect to the most common venues as existing restaurants in that neighborhood with respect to Average Net Canandian National Income and Crime Rate in the neighborhoods.

- That being said, the future of this project includes taking other factors such as specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), proximity to farmers market, connectivity to major neighborhoods as during this pandemic there can be many pick-ups, parking availability, cost of establishment, levels of noise/proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

- Lastly, the final decision for the best restaurant location will be made by potential stakeholders and board of directors. Projects like this can provide them better insights supporting the decision-making process.

Thank you