**1. Introduction: Business Problem**

The battel of the neighborhood is an important data science project targeting the clusters of similar types to provide a user to exploit this as a tool for an ample number of business choices. This project is based on the application of different data science techniques learned in the entire course.

**1.1 Objective**

With the given set of objectives, the battle of neighborhoods has been implemented for the exploration of data using Segmenting and Clustering techniques applied to the neighborhoods in Toronto. The objective of Applied Data Science Capstone is given as follows:

1. To learn about clustering and k-means clustering in particular.
2. To showcase this project in the form of the public repository using the GitHub platform.
3. To learn how to use the Foursquare API and clustering to segment and cluster the neighborhoods in New York City.
4. To learn how to use the Beautifulsoup Python package to scrape websites and parse HTML code.
5. To apply the skills acquired so far in this course to segment and cluster neighborhoods in the city of Toronto.

Aiming to achieve above objectives the data has been collected from Wikipedia page, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> . For this purpose knowledge of Python, Data Science, Visualization, Machine Learning Algorithms, Segmentation, HTML parsing, Repository site GitHub, and Foursquare is required. How do I utilize the complete program to build the “The Battle of Neighborhoods” project have been shown in the background section. [Top](http://localhost:8888/notebooks/Desktop/Coursera/Coursera_Capstone/PGA%20Segmenting%20and%20Clustering%20Neighborhoods%20in%20Toronto.ipynb#toc)

**2. Data**

Based on the problem statement:

* Following data sources will be needed to extract/generate the required information:, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Geocoder**
* number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
* coordinate of Toronto center will be obtained using **Geocoder** of well known Toronto location
* The data frame will consist of three columns: **PostalCode, Borough, and Neighborhood**
* Needs to process only those cells that have an assigned Borough. Ignore cells with a Borough that is Not assigned.
* More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, one can see that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
* If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the Borough. So for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighborhood columns will be Queen's Park.
* Needs to clean the data using appropriate methods.

**3. Background**

The battel of the neighborhood was started from day one of the IBM Data Science Professional Certificate course designed uniquely to become data scientist (Santarcangelo, Ahuja, et al. 2018) . The introduction of each of the faculty members involved into delivering this course have been given in the cited web source as shown in the reference section of this report. Following set of courses described how it has been utilized to develop the Capstone Project.

1. The course started with learning “What is data Science”, which was taught by Alex Aklson and Polong Lin. This module motivated us to lead towards the path of data science (Aklson and Lin, What is Data Science 2018). This module gives the insight about the statistics involved in the field of data science, the demand for data scientists, and some of the qualities of excelling data scientists. According to this course module data science is considered the sexiest job in the 21st century. The term "cloud" was explained by the scientists. Different tools, algorithms, and technologies they use on a daily basis were elaborated and also motivated by the data science professionals as they give advice to anyone who is passionate about data science.
2. Open Source tools such as Jupyter Notebook, Apache Zeppelin Notebooks, and IBM Data Science Experience and Cognitive Class Labs (IBM Watson) for Data Science were discussed by Polong Lin (Lin 2018). The entire lab practice sessions for all other courses were completed using IBM Watson.
3. The course Data Science Methodology presented by Alex and Lin, explained the characteristics of the modeling process, the Modeling and the Model Evaluation stages pertaining to any data science problem. This module helped us to understand data, ways in which data is prepared, the purpose of data modeling, means to evaluate a model and ways in which a model is evaluated, means to prepare or clean data and finally to complete the Data Understanding and the Data Preparation stages pertaining to any data science problem (Aklson and Lin, Data Science Methodology 2018).
4. The course Python for Data Science and AI started a practical journey, which was delivered by Joseph Santarcangelo. This module covered Python Basics, first program, Types, Expressions and Variables, String Operations, Python Data Structures such as Lists and Tuples, Sets, Dictionaries. Programming Fundamentals such as Conditions and Branching, Loops, Functions, and Objects and Classes were discussed. This course also covered Reading files with open, writing files with open, loading data with Pandas, working with and Saving data with Pandas and Numpy (Santarcangelo, Python for data science 2018).
5. The Databases and SQL for Data Science course taught by Rav Ahuja, helped us to understand SQL and Relational Databases, to write basic SQL statements, Practice basic SQL statements hands-on on a live database, and to create a database instance on the IBM Watson Cloud (Ahuja 2018).
6. The course Data Analysis with Python taught by Joseph Santarcangelo, covered data analysis from scratch. This is the biggest and very important course in this program. The PGA also found helpful to develop our concept of Data Analytics with Python (Santarcangelo, Data Analysis with Python 2018).
7. Data Visualization with Python is the 7th course in this series. Alex Aklson taught in this course about reading CSV files into a pandas data frame and process and manipulate the data in the data frame followed by data visualization and some of the best practices to keep in mind when creating plots and visuals. He has also explained different kinds of plots using Matplotlib (Aklson, Introduction to Data Visualization Tools 2018).
8. Machine Learning with Python is the final derivable of this certification course. The course has taught by Saeed Aghabozorgi. He has demonstrated the Python libraries for Machine Learning. The course covered Supervised vs. Unsupervised algorithms with examples. Simple and Multiple Regression, Linear and Non-Linear Regression have been on a dataset for estimation. Different Classification methods and its application have been shown in various datasets to solve real-world problems. Recommender systems have been discussed, in this connection, we will get introduced with the main idea behind recommendation engines, then two main types of recommendation engines, namely, content-based and collaborative filtering (Aghabozorgi 2018).
9. The GitHub repository is a way to showcase our implemented ideas in front of the world. Github has been setup using steps given in a blog post entitled “How to Properly Setup Your Github Repository — Windows Version” by our faculty Alex Aklson (Aklson 2019).

All the above topics have been utilized to find the solution to problem effectively.

**1. Introduction: Business Problem**

The battel of the neighborhood is an important data science project targeting the clusters of similar types to provide a user to exploit this as a tool for an ample number of business choices. This project is based on the application of different data science techniques learned in the entire course.

**1.1 Objective**

With the given set of objectives, the battle of neighborhoods has been implemented for the exploration of data using Segmenting and Clustering techniques applied to the neighborhoods in Toronto. The objective of Applied Data Science Capstone is given as follows:

1. To learn about clustering and k-means clustering in particular.
2. To showcase this project in the form of the public repository using the GitHub platform.
3. To learn how to use the Foursquare API and clustering to segment and cluster the neighborhoods in New York City.
4. To learn how to use the Beautifulsoup Python package to scrape websites and parse HTML code.
5. To apply the skills acquired so far in this course to segment and cluster neighborhoods in the city of Toronto.

Aiming to achieve above objectives the data has been collected from Wikipedia page, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> . For this purpose knowledge of Python, Data Science, Visualization, Machine Learning Algorithms, Segmentation, HTML parsing, Repository site GitHub, and Foursquare is required. How do I utilize the complete program to build the “The Battle of Neighborhoods” project have been shown in the background section. [Top](http://localhost:8888/notebooks/Desktop/Coursera/Coursera_Capstone/PGA%20Segmenting%20and%20Clustering%20Neighborhoods%20in%20Toronto.ipynb#toc)

**2. Data**

Based on the problem statement:

* Following data sources will be needed to extract/generate the required information:, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using **Geocoder**
* number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
* coordinate of Toronto center will be obtained using **Geocoder** of well known Toronto location
* The data frame will consist of three columns: **PostalCode, Borough, and Neighborhood**
* Needs to process only those cells that have an assigned Borough. Ignore cells with a Borough that is Not assigned.
* More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, one can see that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
* If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the Borough. So for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighborhood columns will be Queen's Park.
* Needs to clean the data using appropriate methods.

**3. Background**

The battel of the neighborhood was started from day one of the IBM Data Science Professional Certificate course designed uniquely to become data scientist (Santarcangelo, Ahuja, et al. 2018) . The introduction of each of the faculty members involved into delivering this course have been given in the cited web source as shown in the reference section of this report. Following set of courses described how it has been utilized to develop the Capstone Project.

1. The course started with learning “What is data Science”, which was taught by Alex Aklson and Polong Lin. This module motivated us to lead towards the path of data science (Aklson and Lin, What is Data Science 2018). This module gives the insight about the statistics involved in the field of data science, the demand for data scientists, and some of the qualities of excelling data scientists. According to this course module data science is considered the sexiest job in the 21st century. The term "cloud" was explained by the scientists. Different tools, algorithms, and technologies they use on a daily basis were elaborated and also motivated by the data science professionals as they give advice to anyone who is passionate about data science.
2. Open Source tools such as Jupyter Notebook, Apache Zeppelin Notebooks, and IBM Data Science Experience and Cognitive Class Labs (IBM Watson) for Data Science were discussed by Polong Lin (Lin 2018). The entire lab practice sessions for all other courses were completed using IBM Watson.
3. The course Data Science Methodology presented by Alex and Lin, explained the characteristics of the modeling process, the Modeling and the Model Evaluation stages pertaining to any data science problem. This module helped us to understand data, ways in which data is prepared, the purpose of data modeling, means to evaluate a model and ways in which a model is evaluated, means to prepare or clean data and finally to complete the Data Understanding and the Data Preparation stages pertaining to any data science problem (Aklson and Lin, Data Science Methodology 2018).
4. The course Python for Data Science and AI started a practical journey, which was delivered by Joseph Santarcangelo. This module covered Python Basics, first program, Types, Expressions and Variables, String Operations, Python Data Structures such as Lists and Tuples, Sets, Dictionaries. Programming Fundamentals such as Conditions and Branching, Loops, Functions, and Objects and Classes were discussed. This course also covered Reading files with open, writing files with open, loading data with Pandas, working with and Saving data with Pandas and Numpy (Santarcangelo, Python for data science 2018).
5. The Databases and SQL for Data Science course taught by Rav Ahuja, helped us to understand SQL and Relational Databases, to write basic SQL statements, Practice basic SQL statements hands-on on a live database, and to create a database instance on the IBM Watson Cloud (Ahuja 2018).
6. The course Data Analysis with Python taught by Joseph Santarcangelo, covered data analysis from scratch. This is the biggest and very important course in this program. The PGA also found helpful to develop our concept of Data Analytics with Python (Santarcangelo, Data Analysis with Python 2018).
7. Data Visualization with Python is the 7th course in this series. Alex Aklson taught in this course about reading CSV files into a pandas data frame and process and manipulate the data in the data frame followed by data visualization and some of the best practices to keep in mind when creating plots and visuals. He has also explained different kinds of plots using Matplotlib (Aklson, Introduction to Data Visualization Tools 2018).
8. Machine Learning with Python is the final derivable of this certification course. The course has taught by Saeed Aghabozorgi. He has demonstrated the Python libraries for Machine Learning. The course covered Supervised vs. Unsupervised algorithms with examples. Simple and Multiple Regression, Linear and Non-Linear Regression have been on a dataset for estimation. Different Classification methods and its application have been shown in various datasets to solve real-world problems. Recommender systems have been discussed, in this connection, we will get introduced with the main idea behind recommendation engines, then two main types of recommendation engines, namely, content-based and collaborative filtering (Aghabozorgi 2018).
9. The GitHub repository is a way to showcase our implemented ideas in front of the world. Github has been setup using steps given in a blog post entitled “How to Properly Setup Your Github Repository — Windows Version” by our faculty Alex Aklson (Aklson 2019).

All the above topics have been utilized to find the solution to problem effectively.