Coursera

IBM Data Science Capstone Project

Problem Definition & Data

Background

The Coursera IBM Data Science Professional Certification course consists of 9 online courses that covers topics including open-source tools and libraries, Python, databases, SQL, data visualization, data analysis, statistical analysis, predictive modeling, and machine learning algorithms. The course finishes with a Capstone project.

The purpose of this Capstone project is to demonstrate the use of the data science toolsets, methodologies, and skills that have been acquired during this course to help solve a business problem.

Problem

In this project I am a hypothetical bike GPS location device and services vendor looking to introduce the new GPS location device that is hidden within the seat tube and a subscription-based service that monitors the location of the bike in real time.

Although bike theft has always been a common issue, especially in urban areas, it has increased during the pandemic in a number of cities. The pandemic led to an unprecedented boom in bikes sales. The rising demand, increase in ridership, and shortage of bikes nationwide among other factors, has likely contributed to a rise in bike theft.

With the introduction of this new product, I will be using the theft data and known bike shop locations to target my approach to market with this new product.

Audience

This is hypothetical, but the target audience for the outcome of this project could be a real manufacturer of bike products that are promoted to reduce bike thefts.

This could be used to determine potential local businesses so that marketing campaigns can targeted to promote their products/services.

Data

For this project, the data that will be used:

- List of districts on Toronto
    - Data will be acquired through web scraping the Canadian Postal Code's Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Geospatial coordinates of the neighborhoods and boroughs of Toronto
    - Data will be captured using the Geocoder package and stored into a csv file for easy consumption – the use of the Geocoder package is no longer free
- Bicycle theft data will be used to ascertain high theft areas using the coordinates of the neighborhoods and boroughs of Toronto
    - Data will be captured from the Toronto Police Services website and stored into a csv file for easy consumption (https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts/about)
- Bike Shop venues within the neighborhoods and boroughs of Toronto
    - Data will be acquired through the use of the Foursquare API

Methodology

This project will compare suburbs and will determine similarities based on clustering techniques using location data services.

This project uses web scraping techniques to retrieve data from the Canadian Postal Code's Wikipedia page.

The data is then acquired and cleansed in preparation for clustering.

The geospatial locations data import will be merged with the post code data which will enable the data to be visualised over a map of the area.

The bicycle theft locations data import will also be merged with the post code data which will enable the data to be visualised over a map of the area.

The data will be clustered and plotted over the map.

The clustering is carried out by K Means and the clusters are plotted using the Folium Library.

The data will be mapped across Toronto and then focused/clustered in on boroughs.