PROJECT PROPOSAL

Analysis of MTA Turnstile Data for Vegan Restaurant Franchise

Prepared by: Sara Hawi ——— Supervised by: Dmitry Denisov

⇒ ***Backstory:***

Las Vegans Co. Ltd., a franchise company centered in the state of New York owning a chain of vegan restaurants, find their branches considerably busy on Weekends (Fridays, Saturdays, and Sundays) as compared to weekdays; therefore, they're looking to increase their revenue on those days. For that reason, they are targeting the most visited subway station in New York to open a new branch right next to and know the optimum working hours to make that happen. Also, they want to put up 3 billboards in three different stations to advertise their trade and draw in more customers.

⇒ ***Question/need:***

By performing exploratory data analysis, this project aims to answer the following questions:

1. Which are the three busiest train stations in the state New York?
2. What are the targeted peak hours on Weekends?
3. What are the targeted peak hours on Weekdays?

Answering Question (1) will help the client choose the top busiest train station to open a branch in its area along with the top three busiest stations where Las Vegans Co. Ltd. ought to put up their billboards. Questions (2-3) aim to tell the client the optimum opening hours to close the weekday-weekend revenue gap which helps increase their revenue and cuts losses of extra employee wages while keeping their revenue to a maximum.

Las Vegans Co. Ltd. primarily aims to provide culinary services for the 9.6 million US population who follow meat-free diets and as New York state listed as one of the top US states where vegans live in a study by Ispos, they want to open a new branch near a train station in NY to benefit vegans who are in a hurry while increasing their weekday revenue.

⇒ ***Data Description:***

The main source of data sought in this project will be train station turnstile machine data provided by the Metropolitan Transportation Authority (MTA) responsible for public transportation in New York. The data is posted on a weekly basis containing the station, cumulative passenger entries and exits for a daily four-hour interval, dates, and individual turnstile information, among others. Three months' worth of data from June 2021 to August 2021 is deemed enough to make an informed up to date decision as data from earlier times can be redundant and irrelevant.

For this project, trends of passenger activity are sought from exploratory data analysis using the total number of entries and exits, date and time of activity per station to locate the top three busiest NY stations and peak hours associated with the busiest station. Exploratory data analysis in this project will be performed on both categorical and numerical data e.g., station names, number of entries.

⇒ ***Tools:***

Two tools will be incorporated in the course of this project's completion: ***Python*** and ***SQL***.

SQL will be used to load the raw data and perform initial operations of data cleaning. Python will be linked to SQL via ***SQLAlchemy*** for querying, ***Pandas*** based exploratory data analysis, as well as Python's ***Matplotlib*** and ***Seaborn*** libraries for data visualization. Initially, previously mentioned tools seem to be sufficient; however, additional tools beyond those required will be decided if needed throughout the course of the project's completion such as Python's ***Bokeh*** and ***Plotly***.

⇒ ***MVP Goal:***

Choosing proper visualization tools is crucial for the comprehensive value of this project. The goal MVP will clearly communicate the top three busiest train stations in New York and the peak hours associated with the busiest station. Heat maps can be used to show peak hours of each day. The areas where the color is the darkest throughout the day will signify a peak hour and vice versa. To compare and show the extent of busyness of different New York stations, bar charts will be used.