Shahad Sulaiman Alkaltham

Project Proposal

29 Sep 2021

Background story:

In this project, our company will take over one of client's problems, as a marketing company and with amazing daily ideas, our need for great, successful and creative ideas becomes more necessary for our company growth.

For those reasons, we need to discover the rush period of a year and crowded place so that we can distribute some samples of their products in a creative way, to achieve that some data analysis will be applied, and the first place chosen is the New York subway and we will comparing between Christmas and New Year periods to know which of them is more crowded.

Questions to Answer:

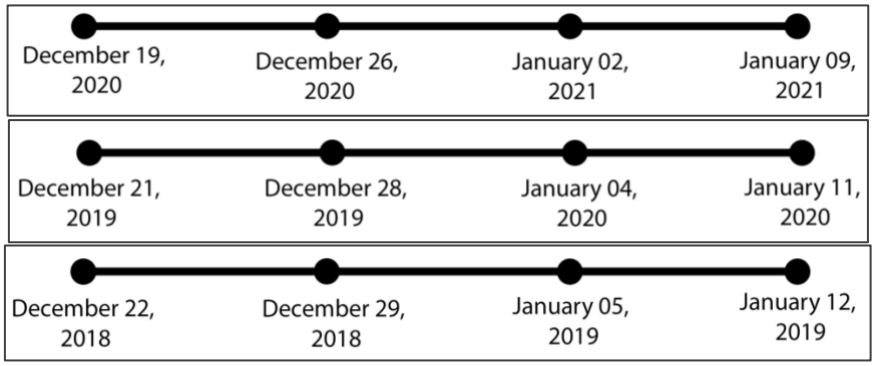
During this project some important questions will be answered, which are:

- Is the MTA subway more crowded in the Christmas or the New Year period more than other days of the year?
- then, what is the most crowded station?
- What is the most crowded day?
- What is the most crowded hour of that day?
- Is MTA subway the right place to do marketing?

All those previous questions for an easier decision-making process and based on the results or answers we will decide to be moving forward or chose another period and place, all to find a solution that satisfies our client.

Data Description:

All the datasets come from MTA website[1], the website provides weekly datasets about their subway in New York, but we will use some of them, about 2,476,441 rows, we will start by twelve weeks during four years, two of them in New Year period and other two during Christmas period, with take COVID-19 pandemic in 2020 - 2021 into consideration, since the datasets contains 2020, 2019, 2018 and 2021, we chose two weeks from each chosen period to compare it together, the figures below shows the chosen periods and each bubble represent one week.



[1] Through the link: http://web.mta.info/developers/turnstile.html

For more about our datasets, in the table in Appendix 'A' clear description of its characteristics as in MTA website[2].

To more understanding and faster processing, I will add some features shown in table below, but more features may be added in the future depends on our analysis needs.

Feature Name	**Feature Description**
ENTRY_PER_DAY	An integer value represents entries for a device in one day
EXIT_PER_DAY	An integer value represents exits for a device in one day

Tools:

Here the basic tools we will use in our data analysis

- Jupyter Notebook: for writing python codes.
- Excel: to browse the dataset.
- SQLite: to store and retrieve data.
- Some python libraries:
    - pandas: to dealing with data frames and doing some statistical operations, also read files.
    - matplotlib: to visualize the results.
    - math: doing some statistical and mathematical operations.
    - numpy: doing some statistical and mathematical operations.
    - seaborn: to visualize the results.

Conclusion:

What I expected from this project is to prove that the Christmas period or New Year Period and New York subway is a good decision for marketing the client's products.

[2] Through the link: http://web.mta.info/developers/resources/nyct/turnstile/ts_Field_Description.txt

Shahad Sulaiman Alkaltham
Project Proposal
29 Sep 2021

Appendix:

A:

Column Name	Column Description
CONTROL_AREA	It represents each control area, and our datasets contains 752 different control areas
UNIT	It represents Remote Unit for a station, and our datasets contains 450 different units
SCP	It Subunit Channel Position represents a specific address for a device, and our datasets contains 229 different SCPs
STATION	It represents the station name the device is located at, and our datasets contains 379 different stations
LINENAME	It represents all train lines that can be boarded at this station, normally lines are represented by one character, such as 456NQR represents train server for 4, 5, 6, N, Q, and R trains, and our datasets contains 114 different line names
DIVISION	It represents the Line originally the station belonged to, and our datasets contains 6 different divisions as mentioned below. **BMT**: Brooklyn–Manhattan Transit Corporation. **IND**: Independent. **IRT**: Interborough Rapid Transit. **PTH**: PATH Port Authority Trans-Hudson. **RIT**: Rochester Institute of Technology. **SRT**: Scarborough Subway.
DATE	It represents the date, and our datasets contains 84 days.
TIME	It represents the time (hh:mm:ss) for a scheduled audit event
DESC	It Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours) 1. Audits may occur more that 4 hours due to planning or troubleshooting activities. 2. Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered,
ENTRIES	It represents the cumulative entry register value for a device
EXITS	It represents the cumulative exit register value for a device