SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Project Proposal

Prepared by:

Fatimah Abdullah AlShammari

1. Summary

XY is a business incubators company. XY would like to post an advertisement at stations with the highest number of EXISTS in the morning from 7-9 AM on working days from Monday to Fridays. Because XY wants to attract customers who want to create a startup company, or startups that need support.

2. Dataset Description

In this project I will use MTA turnstile dataset, which is containing 11 columns and the number of rows is unknown because the dataset updated weekly. So, I will study data for the last 3 months.

The table below shows the description for each column in the dataset.

Feature	Description
C/A	Control Area (A002)
UNIT	Remote Unit for a station (R051)
SCP	Subunit Channel Position represents a specific address for a device (02-00-00)
STATION	Represents the station name the device is located at
LINENAME	Represents all train lines that can be boarded at this station Normally lines are represented by one character. LINENAME 456NQR represents train server for 4, 5, 6, N, Q, and R trains.
DIVISION	Represents the Line originally the station belonged to BMT, IRT, or IND
DATE	Represents the date (MM-DD-YY)
TIME	Represents the time (hh:mm:ss) for a scheduled audit event
DESc	= 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
ENTRIES	The cumulative entry register value for a device
EXIST	The cumulative exit register value for a device

The Table below shows the descriptions for each new column that will be added in the dataset.

Feature	Description
Date_time	To combine date with time and convert it to datetime type
weekday	To find every date corresponding to any day of the week
Num_of_EXISTS	The exact number of people who left the station in the last 4 hours
TIME_24_HOUR	Contains time per hour par day
NUM_OF_EXISTS_PER_HOUR	Contains number of people who left the station per hour

3. Tools

I will use pandas and numpy libraries to analysis the data, and then I will display the data results by using the matplotlib or seaborn library.

4. Conclusion

After studying the data, I will identify the stations that people frequently leave between 7-9 AM, so I can conclude that these stations are close to some companies. Therefore, we have determined the appropriate location to publish the advertisement.