SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

MTA Turnstile Data Analysis

Final Phase

Prepared by:

Fatimah Abdullah AlShammari

Abstract

The goal of the project was to get the most station or turnstile that contains the largest number of EXITS in the two hours 7-9 AM . I worked with MTA turnstile dataset. and I did some steps in the preprocessing stage, I showed the results that I got by using a bar chart in the matplotlib library.

1. Problem Statement

A business incubators company 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. Data

2.1. Features

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

2.2. Derived Features

Feature	Description
datetime	Combine date with time to convert it to datetime type
day_of_week	Contains every date corresponding to any day of the week
weekday_OR_weekend	Contains every which type of day, weekend or weekday
PREV_EXITS	To subtract it from EXITS to find exact number of people who left the station for each 4 hours
EXISTS2	The exact number of people who left the station for each 4 hours
Num_of_EXITS_in_1_hour	Contains number of people who left the station per hour

3. Tools

I used pandas and numpy libraries to analysis the data, and then I displayed the data results by using the matplotlib library.

4. Result

	C/A	UNIT	SCP	STATION	DATE	Num_of_EXITS_in_1_hour
84035	N510	R163	02-00-02	14 ST	08/12/2021	191840.50
139400	R604	R108	03-00-05	BOROUGH HALL	06/03/2021	174626.25
82219	N506	R022	00-05-04	34 ST-HERALD SQ	06/03/2021	137251.25
44854	N068	R012	03-00-02	34 ST-PENN STA	07/15/2021	122592.00
98374	R137	R031	02-03-00	34 ST-PENN STA	08/23/2021	105794.75
66635	N325A	R218	00-00-00	ELMHURST AV	07/27/2021	104007.25
132474	R526	R096	00-05-03	82 ST-JACKSON H	06/30/2021	98967.25
57704	N137	R354	00-00-00	104 ST	06/29/2021	79800.50
103591	R177	R273	01-00-01	145 ST	07/16/2021	48943.50
42589	N063	R011	02-00-00	42 ST-PORT AUTH	08/20/2021	38088.00

Figure 1: subset of the dataset that contains our goal

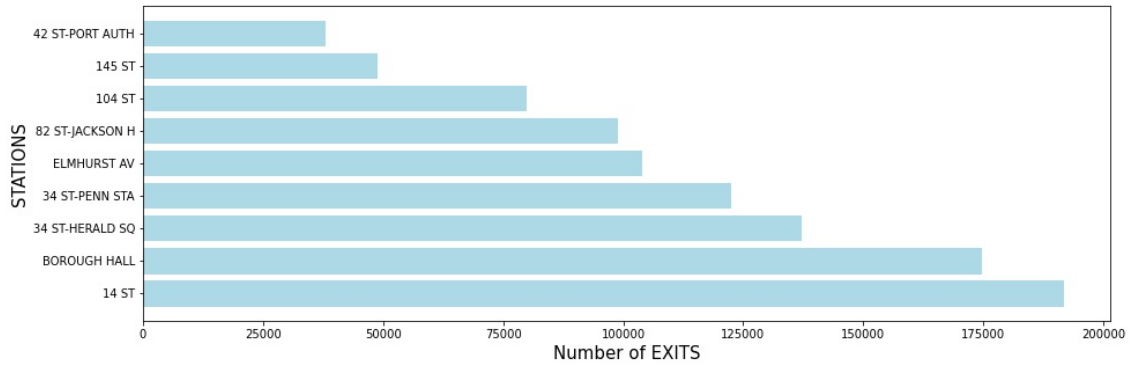

Figure 2: chart represent top 10 station with the number of EXITS for each hour

5. Conclusion

After working on the dataset, I conclude that the 14 ST station has the largest number of exits that's mean the most workplaces or companies are stationed next to 14 ST station. So, we can place the advertisement in 14 ST to attract the targets.