SDAIA T5 Boot Camp Project Proposal

Question/Need:

The Riyadh metro organization is the governing body for the metro project in Riyadh, with such a new project some adjustment, but wouldn't it be nice to have a map that displays the prediction for high traffic stations in advance to prepare in advance and avoid complications down the road?

By using data from the MTA which is the governing body for the subway system in NY, high traffic stations in NY could identified then extracting features of their locations (e.g.: near working area, school zone, housing area, tourist attraction, etc.…).

Data Description:

Data is from the MTA open dataset, the MTA was chosen for the following reasons:

- It is one of the 10 biggest subway systems worldwide.
- Has a lot of easily accessible data from the organization's website.
- Riyadh is witnessing a growth in population due to job opportunities and visitors which resample NY.

From the collected data it is expected to need the following information:

- Station name.
- Date and time.
- Number of people who entered/exited the station.

NOTE: the data of the collected data is from 2020/1/1 until 2020/12/31.

Dataset features/columns:

Feature Name	Feature Type
Control Area	Text
Remote Unit	Text
Device Address	Text
Station Name	Text
Lines Served by The Station	Text
Date	Text
Time	Text
Audit Scheduled	Text
Station's Division	Text
Sum of Entries	Number
Sum of Exits	Number

Number of entries: ~ 10,715,796

Tools:

Tools: SQL, Python, and Jupyter Notebook.

Libraries: SQLAlchemy, Pandas, and Plotly.

MVP Goal:

Return which stations are predicted to have high traffic levels, based on their location.