What is the framing question of your analysis, or the purpose of the model/system you plan to build?

Picture it. Hamburg, Germany. 1960. The Beatles play nightly 7 hour gigs for several months, transforming them from raw, individual musicians into the polished, cohesive group that would shoot into stardom. If musicians today could get that kind of performance and exposure experience early in their career, it would be very beneficial. Subway stations have so much throughput and "provide" a free stage for live performers. Which stations have higher foot traffic at which times of day (probably evening to play to a target audience), and are also near nightlife locations?

Who benefits from exploring this question or building this model/system?

Independent artists and small recording labels who want to expand their audience, polish their chops, and develop their people skills. The model can be changed depending on the music style and perceived audience. For example, jazz may be better received in the evening near fancier bars and restaurants while punk rock is better later at night near grittier bars and clubs.

What dataset(s) do you plan to use, and how will you obtain the data?

This will use the MTA data and another dataset with addresses and price indicators of bars/clubs/restaurants to match the stations to nearby bars and clubs based on locations and cost. I have not found a datset like that yet so that may not be possible.

What is an individual sample/unit of analysis in this project? What characteristics/features do you expect to work with?

An individual unit of analysis is the maximum number of people passing through a station near bars/clubs at specific times of day, probably 4pm-12am. I expect to work with number of entries/exits, station name, date, time, SCP, and unit or C/A.

If modeling, what will you predict as your target?

I'm not sure if I'll be modeling.

How do you intend to meet the tools requirement of the project?

I completed the quick setup. I will use a query to group the data by either UNIT or C/A and SCP and order by day and time. Then some preliminary plots by station just to see what there is to see in terms of number of people going through vs. time of day. Maybe it will be valuable to add a column delimiting weekdays/weekends. Then I'll do some more queries (or use pandas?) to find the stations with the most people. If there's time and I find an applicable dataset, I want to add the bar/club/demographic component.

All of that will use the SQL database, Python, SQLAlchemy, pandas, and matplotlib.

Are you planning in advance to need or use additional tools beyond those required?

No.

What would a minimum viable product (MVP) look like for this project?

An MVP would be the top 20 stations with the most traffic or a plot of one station showing its busiest 4-hour time period over the 3-month time period or even an analysis of one station showing the busiest time period on weekdays vs the busiest time period on weekends.