Yoga Today EDA Project Write-up

Using EDA to find ideal locations for a new yoga studio

By: Linda Ngo

Abstract

The goal of the project was to use exploratory data analysis to help determine which areas the stakeholders of Yoga Studio should consider when scouting for a new location for their studio. The MTA Turnstile Dataset was used in this exercise in order to aid in determining areas of high foot traffic that would be convenient for yoga practitioners. Yoga Today was a popular and hot yoga studio before the pandemic but due to the restrictions and drop in attendance, the studio was forced to close its doors. Sensing the resurgence of normal life, the owners of Yoga Today have reached out to us to request our help in narrowing down areas of interest in New York City that would be ideal for the new studio.

Design and data

For a yoga studio to have high sign ups, the location and time needs to be convenient for yoga practitioners to incorporate into their schedule. The location needs to be near convenient transportation and given that one of the most popular forms of transportation in NYC is the subway, we choose to use the MTA Turnstile dataset as a main basis for our analysis. The dataset contains over 3 million rows and includes 379 stations. There are about 3-4 million riders a day so we believe that this is an adequate dataset for the exercise requested of us.

Algorithms

SQL and Pandas were used to clean up the data and aggregate it so that it would be usable for data analysis. Effort was taken to clean the column names, join them together or aggregating as to calculate the totals by station, turnstile, month, day, week etc. We also used pandas to drop certain columns that we didn't need or erroneous data that was skewing the dataset. After examining the data, we were able to deduce that the data was taken as a snapshot every 4 hours and the entries and exits were cumulative values. While the number of entries did not match the number of exits (maybe people took an Uber home from work?), we decided the best approach would be to drop the exits and only use the daily entries for our analysis. After we were done cleaning the data, we concatenated it so that we could calculate totals to see which stations had the most traffic. Following that, we calculated which days of the week had the most ridership and it turned out to be the weekdays. Some stations had a close amount of weekend riders as weekdays and we suspect those may be in locations that are more touristy. We recommend that the owners of Yoga Today consider opening an office near one of the top 10 busiest stations to take advantage of the foot traffic of people that could join the yoga studio before or after work.

Further Ideas to Pursue

Due to the constraints of time, we were not able to do analysis with other ideas that we thought of. Further consideration should be taken in terms of targeting the customer base for the studio. The dataset from the Census should be taken into observation to see the gender and age demographics as we would like to aim for an area that has a high amount of women between the ages of 18 and 45. Another topic that should be taken into consideration is how much the rent is in these areas as some areas are probably much more expensive than others and rent expenses would be a major driver when picking a location.

Tools

- NumPy and Pandas for data manipulation
- Matplotlib, Altair and Folium for plotting, maps and graphs
- SQL (DB Browser and SQLAlchemy) for data manipulation and loading into Python