ABSTRACT
The goal of this project was to perform Exploratory Data Analysis (EDA) on New York City subway data to identify high-volume traffic locations to propose potential, targeted, new business locations. Using turnstile data from the Metropolitan Transit Authority (MTA), I performed data cleaning and preparation, data transformation and manipulation, and data visualization to determine which MTA stations experience the most traffic, by day and time. The resulting analysis allowed me to propose data-driven recommendations based on NYC MTA traffic patterns.

DESIGN
Bagels & Brew, a café featuring breakfast offerings, with considerations to open a new location in New York City. Planned for spring of 2022, they are hoping for increased MTA ridership as NYC traffic commute returns to near pre-pandemic levels. More importantly, Bagels & Brew is seeking to make an informed decision about potential locations to open their café, with particular emphasis on weekday morning hours as it closely aligns with the specialty of their business.

To address the needs of Bagels & Brew and the scope of the project, I used pre-pandemic MTA turnstile data to identify high-volume ridership by station, day, and time. Leveraging data analysis tools to better visualize MTA turnstile traffic and draw actionable insight on behalf of the client.

DATA
The New York MTA publishes weekly turnstile data. The dataset examined contained 13 weeks of turnstile data (JAN-MAR, 2019), with over 2.5 million rows of multiple field names (features).

Note that each row is an observation at a moment in time for a *specific* turnstile (defined by combining the C/A, Unit, and SCP). These turnstile readings occur every 4 hours, recording both the entry and exit counts for that particular MTA station the turnstile is located. Most importantly, the entry and exit recordings are *not* counts of people for a given timeframe, but instead act more like an odometer where the counts are recorded *cumulatively*.

Below you can find detailed descriptions of each MTA column (feature) found in this dataset:

FIELD NAME	***DESCRIPTION***
C/A	Control Area
UNIT	Remote Unit for a station
SCP	Subunit Channel Position represents a specific address for a device
STATION	Represents the station name the device is located
LINENAME	Represents all train lines that can be boarded at this station
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	Represents the type of audit event (“REGULAR” or “RECOVR_AUD”). Scheduled events occur every 4 hours
ENTRIES	The cumulative entry register value for a device
EXITS	The cumulative exit register value for a device

METHODS

Performed EDA using the following processes:

Data Querying and Importing (with SQL and SQLAlchemy)

- Accessed a local MTA database file, performed a SQL query, and imported the dataset into my environment (Jupyter Notebook)

Data Preparation and Cleaning (with Pandas and Python)

- Duplicate turnstile readings having multiple recorded events for a given observation (day, time, station) were removed
- Turnstiles readings with negative turnstile counts (entries minus - entries) due to reversed counters were changed into positive values while calculations resulting in null values were dropped

- Turnstile readings with extreme outliers caused by missing or incorrect data were removed or fixed accordingly

Data Formatting (with Pandas)

- Transformed column data types and added additional columns, including day of the week (by name), previous entries, previous exits etc
- Filtered out turnstile recordings outside the scope of the EDA (afternoons/evenings, and weekends)
- Sorted the data by station, turnstile, time, and day - resulting in a more user-friendly dataset to calculate total entry and exit counts, organized by location and timeframe

Data Manipulation, Exploration, and Analysis (with Pandas)

- Explored (total) entry and exit readings by location
- Explored (total) entry and exit readings by day and time
- Explored exit readings only, by location, day, and time

Data Visualization (with Matplotlib and Seaborn)

- Visualized total traffic (entry + exit counts) of the top MTA stations
- Visualized busiest and slowest MTA stations by exit traffic count only
- Visualized total traffic by weekday

TOOLS

Languages used in this project included Python and SQL. The following libraries were additionally used:

- SQLAlchemy for database creation and data querying/importing
- Pandas for data cleaning, transformation, manipulation for analysis
- Matplotlib and Seaborn for data visualization

COMMUNICATION

More details of the findings from this project can be viewed here. This project along with relevant files, code, and images can also be found on my personal Github repo.