QUANTUM
-ANALYTICS-

NYC Taxi Project

About the dataset

- This dataset contains 6 tables in csv format, along with a geospatial map in TopoJSON and Shapefile formats
- The 4 Taxi Trips tables contain a total of 28 million Green Taxi trips in New York City from 2017 to 2020. Each record represents one trip, with fields containing details about the pick-up/drop-off times and locations, distances, fares, passengers, and more
- The 454 Calendar table contains a fiscal calendar (2017-2020) used by the Taxi & Limousine Commission, with fields containing the date and fiscal year, quarter, month, and week
- The Taxi Zones table contains information about 265 zone locations in New York City, including the location id, borough, and service zone
- The Taxi Zones Map files contain a map of New York City with divisions for the 265 locations that can be used to create custom map visuals in Power BI (TopoJSON) or Tableau (Shapefile)

For this project, you'll be playing the role of a new Data Analyst for the New York City Taxi & Limousine Commission. It's your first week on the job, and you just received the following email from the Lead Dispatcher:

Welcome to the team!

We've been collecting trip data for ~4 years now, but without a proper analyst we haven't been able to put it to good use. That's where you come in!

The raw data has some issues, so we'll need to make the following adjustments and assumptions to clean and prep the data:

- Let's stick to trips that were NOT sent via "store and forward"
- I'm only interested in street-hailed trips paid by card or cash, with a standard rate
- We can remove any trips with dates before 2017 or after 2020, along with any trips with pickups or drop-offs into unknown zones
- Let's assume any trips with no recorded passengers had 1 passenger
- If a pickup date/time is AFTER the drop-off date/time, let's swap them

- We can remove trips lasting longer than a day, and any trips which show both a distance and fare amount of 0
- If you notice any records where the fare, taxes, and surcharges are ALL negative, please make them positive
- For any trips that have a fare amount but have a trip distance of 0, calculate the distance this way: (Fare amount - 2.5) / 2.5
- For any trips that have a trip distance but have a fare amount of 0, calculate the fare amount this way: 2.5 + (trip distance x 2.5)

Once the data is cleaned up, I'm hoping you can build me a dashboard to help with weekly planning and logistics. For any given fiscal week, I'd like to be able to use historical data to answer the following questions:

1. What's the average number of trips we can expect this week?
2. What's the average fare per trip we expect to collect?
3. What's the average distance traveled per trip?
4. How do we expect trip volume to change, relative to last week?
5. Which days of the week and times of the day will be busiest?
6. What will likely be the most popular pick-up and drop-off locations?

I realize this is a lot to ask for, but this type of analysis will have a huge impact on our business!

Thanks in advance,

Marie Jones (Lead Dispatcher, NYC Green Taxis)

For this challenge, your task is to build a dashboard that meets Marie's requirements.

NB: Generate your own insights to add to this.

Samuel Josiah
Program Lead
Samuel@quantumanalyticsco.org

Please follow our LinkedIn **HERE**