Insights from Failed Orders

Gett, previously known as GetTaxi, is an Israeli-developed technology platform solely focused on corporate Ground Transportation Management (GTM). They have an application where clients can order taxis, and drivers can accept their rides (offers). At the moment, when the client clicks the Order button in the application, the matching system searches for the most relevant drivers and offers them the order. In this task, we would like to investigate some matching metrics for orders that did not completed successfully, i.e., the customer didn't end up getting a car.

Data Description

We have two data sets: data_orders and data_offers, both being stored in a CSV format. The data_orders data set contains the following columns:

order_datetime - time of the order

origin_longitude - longitude of the order

origin_latitude - latitude of the order

m_order_eta - time before order arrival

order_gk - order number

order_status_key - status, an enumeration consisting of the following mapping:

4 - cancelled by client,

9 - cancelled by system, i.e., a reject

is_driver_assigned_key - whether a driver has been assigned

cancellation_time_in_seconds - how many seconds passed before cancellation

The data_offers data set is a simple map with 2 columns:

order_gk - order number, associated with the same column from the orders data set

offer_id - ID of an offer

Practicalities

Make sure that the solution reflects your entire thought process including the preparation of data - it is more important how the code is structured rather than just the final result or plot.

Step 1: Data Preparation

First, let's load and inspect the datasets to understand their structure and content.

Step 2: Merge the Data Sets

Merge the orders_df and offers_df dataframes on the order_gk column.

Step 3: Distribution of Orders According to Reasons for Failure

Build the distribution of orders according to the reasons for failure: cancellations before and after driver assignment, and reasons for order rejection.

Question 1:

Build up the distribution of orders according to reasons for failure: cancellations before and after driver assignment, and reasons for order rejection. Analyze the resulting plot. Which category has the highest number of orders?

Step 4: Distribution of Failed Orders by Hour

Plot the distribution of failed orders by hours. Analyze if there is a trend.

Question 2:

Plot the distribution of failed orders by hours. Is there a trend that certain hours have an abnormally high proportion of one category or another? What hours are the biggest fails? How can this be explained?

Step 5: Average Time to Cancellation with and without Driver by Hour

Plot the average time to cancellation with and without a driver by the hour.

Question 3:

Plot the average time to cancellation with and without driver, by the hour. If there are any outliers in the data, it would be better to remove them. Can we draw any conclusions from this plot?

Step 6: Distribution of Average ETA by Hours

Plot the distribution of average ETA by hours.

Question 4:

Plot the distribution of average ETA by hours. How can this plot be explained?

Step 7: BONUS - Hexagon Visualization

Using the h3 and folium packages, calculate the hexagons containing 80% of all orders and visualize them.

Bonus Question:

Using the h3 and folium packages, calculate how many size 8 hexes contain 80% of all orders from the original data sets and visualize the hexes, coloring them by the number of fails on the map.