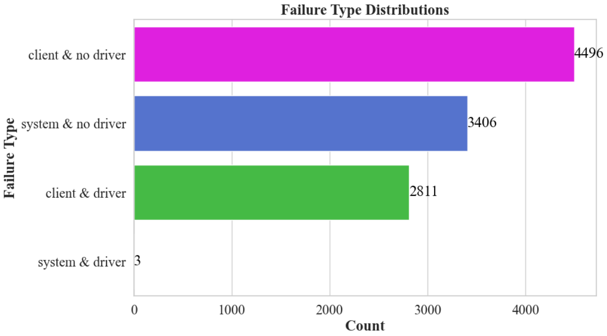
STA 326 Assignment 2 Report

Qijia He 12111211

Task 1

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

Solution:

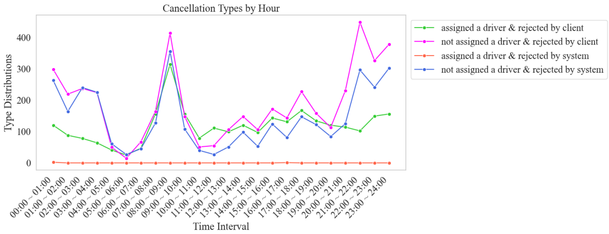


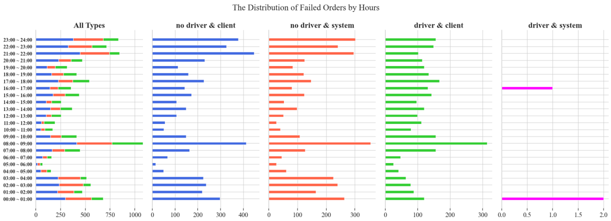
The category that has the highest failures belongs to the purple one, which is "cancelled by client & driver not assigned".

Task 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?

Solution:
For this problem, I drew 2 plots in total, the first plot is a line chart while the second is a combination of bar plot, the result of which is as follows:





As the first picture shows, usually (as the bar plot in Task 1), the normal pattern of bars should be **high - median - median - extremely low**, and some abnormal patterns happen at:

- **8:00 12:00, 19:00 20:00**: During these times, which are rushing hours, most clients need to wait longer for a driver (which can also be verified in Task 4). Some people are not willing to wait, hence the occurrence of type **4&1** (canceled by client while assigned a driver) failures is higher.

- **2:00 4:00**: During these hours, the occurrence of type **9&0** (rejected by systems and not assigned a driver) is too high, while **4&1** is low. This might be because most drivers are sleeping, resulting in only a few drivers being available in certain regions. After observing this plot, if I were a driver, I might choose to work at midnight. If I were the manager of this company, I might choose to enlarge the searching distance to find drivers for clients.

- The most abnormal time is **5:00 6:00**, where **4&0** (canceled by client and no driver assigned) is less frequent than **4&1 and 9&0**. This is the time when almost everybody goes to sleep. However, this abnormal situation may be due to the fact that there are only a few cases available at that time (as seen from the second picture), so we cannot be sure whether this situation is robust.

Task 3

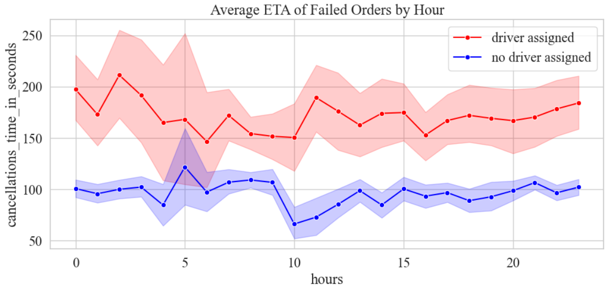
Plot the average time to cancellation with and without a driver, by the hour. If there are any outliers in the data, it would be better to remove them. Can we draw any conclusions from this plot?

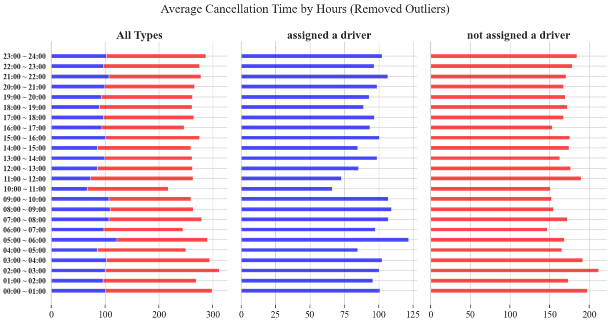
Solution:
To remove the outliers, I use the following strategy:

- Calculate the 25% and 75% percentile of the 'cancellations_time_in_seconds' column for each 'is_driver_assigned_key' type, denoted as x_1 and x_2 (for one type).

- A point is an outlier if $x < x_1 - 1.5 \times (x_2 - x_1)$ or $x > x_2 + 1.5 \times (x_2 - x_1)$. This is the method which is used for `boxplot` to detect outliers.

After removing outliers, we noticed that the average ETA Failed Orders by Hour is smaller than the original one. This is because most outliers occurred when a client decided to wait for a long time even though the driver is far away or not assigned.



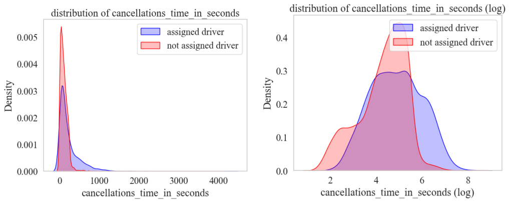


Conclusions:

1. Generally speaking, when a driver is assigned, a customer is usually willing to wait more time compared to those who did not have a driver.

2. The change of the average cancellation time when a driver is not assigned is more stable with regard to time, but when a driver is assigned, the cancellation time fluctuates more. Usually, during rush hours and evening time, the cancellation time is lower when assigned a driver.

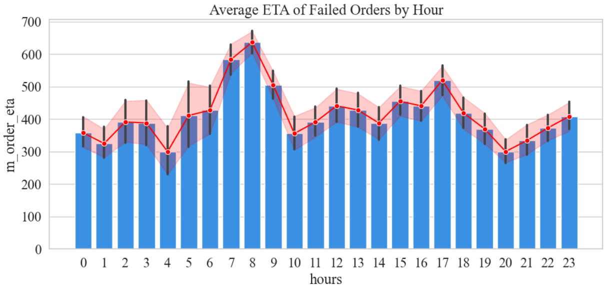
3. It can be seen from the following picture that even though the average cancellation time is longer when a driver is assigned, the most frequent cancellation time for these 2 types is almost the same. Only approximately 30% of clients are willing to wait longer, and it is those clients that make the average cancellation time become longer.



Task 4

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

Solution:



The plot shows that average ETA is higher during rush hours, indicating increased demand and potentially longer wait times for orders. There is a correlation between the number of failures and average ETA, suggesting that high failure rates contribute to longer estimated times of arrival.

Task 5

BONUS: Hexagons. Using the h3 and folium packages, calculate how many sizes 8 hexes contain 80% of all orders from the original datasets and visualize the hexes, coloring them by the number of fails on the map.

Solution:

There are 24 hexes in total, and the most frequent region on the map (the yellow one) contains a train station.