Data Practice Homework 2

唐麟杰 12111204

2024 年 4 月 25 日

1 Task 1

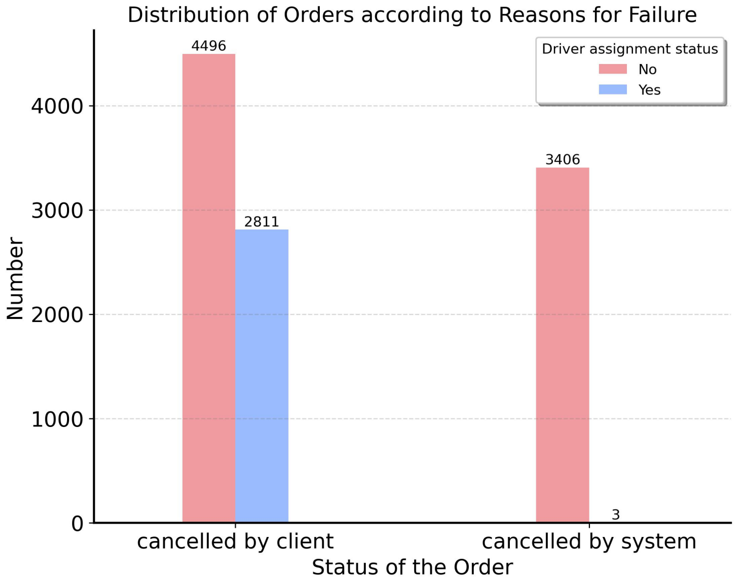

图 1: Distribution of Orders according to Reasons for Failure

As shown in the chart, we divide into two main categories: orders cancelled by clients and orders cancelled by the system. In these two categories, we can further distinguish whether a driver had been assigned at the time the order was cancelled.

- Among the orders cancelled by clients, the most were cancelled before a driver was assigned, amounting to 4496 orders, while 2811 orders were cancelled after a driver had been assigned.
- Among the orders cancelled by the system, again the majority were cancelled before a driver was assigned, with 3406 orders. Only 3 orders were cancelled by the system after a driver had been assigned.
- Therefore, the category with the highest number of orders is orders cancelled by clients before a driver has been assigned. This may be due to changes in customer needs or traffic conditions at the time. Secondly, the number of orders cancelled by the system is also significant, indicating that the taxi dispatch system needs further improvement.

2 Task 2

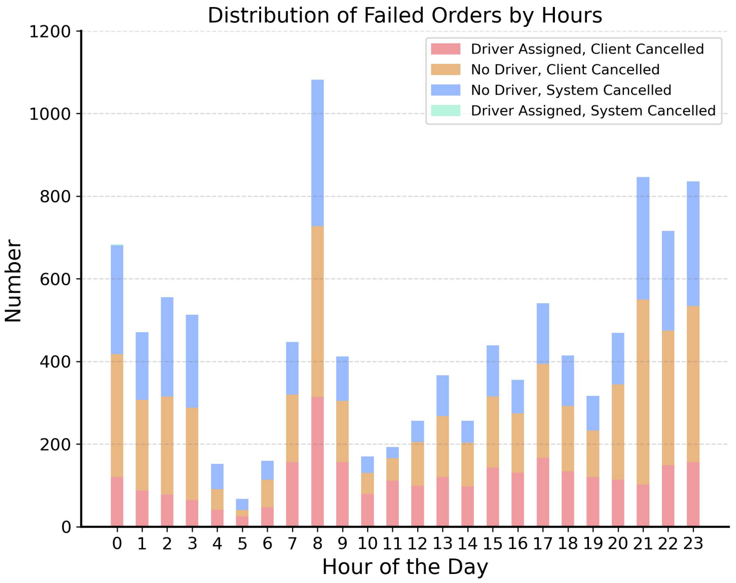

图 2: Distribution of Failed Orders by Hours

The chart above shows that:

- There are certain hours that have an abnormally high proportion of one category: Clear peaks in the number of failed orders are observed at 8 AM and between 9 PM to 1 AM. In terms of categories, there seem to be two particularly prominent types of failed orders: those with no driver and cancelled by the system (light blue bars), and those with no driver and cancelled by the client (yellow bars).
- The peak period of failures occurs at 8 AM.
- Possible explanations: The peak at 8 AM might be due to the rush hour, where passenger demand for rides increases sharply. The number of available drivers may not meet this requirement. The high rate of failures in the evening and early morning hours might be due to a reduction in the number of drivers at night or system maintenance.

3 Task 3

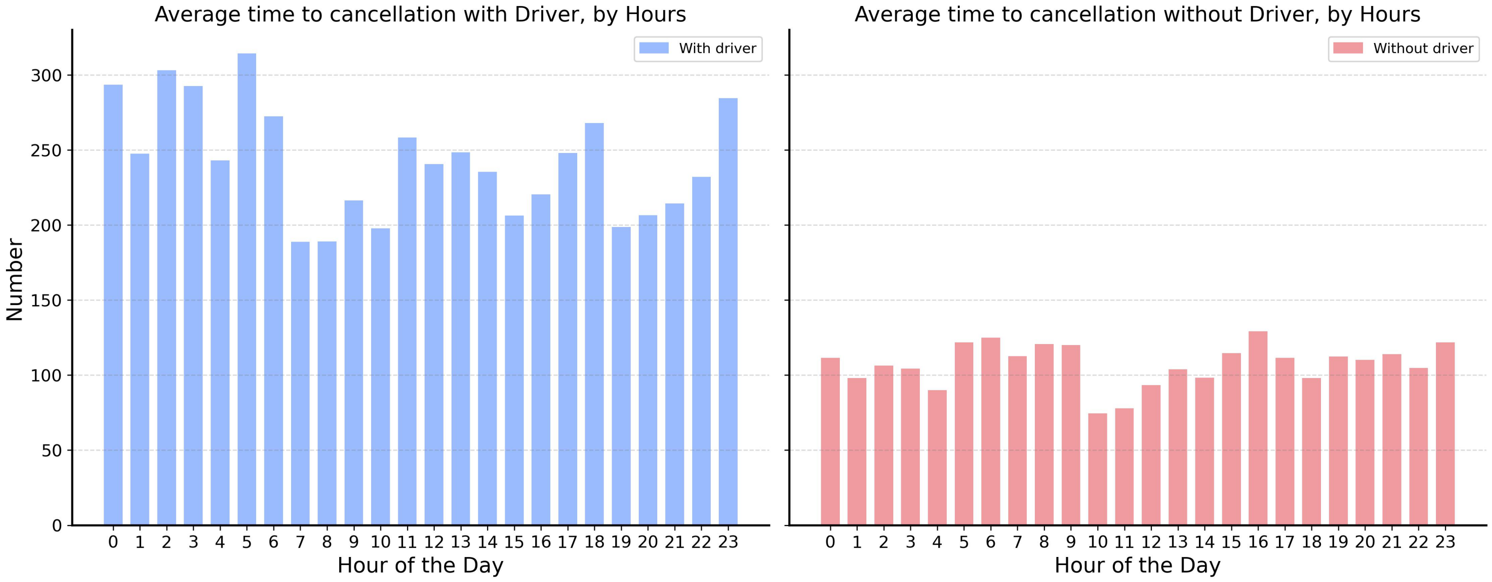

图 3: Average time to cancellation with or without Driver, by Hours (before data cleaning)

Based on the chart, we can draw the following preliminary conclusions:

- When a driver is assigned to an order, the average cancellation time is longer. This may suggest that users are more willing to wait because they know a driver is on the way to pick them up.
- The cancellation times are relatively shorter when no driver is assigned. This could be due to users choosing to cancel their orders after a long waiting time without a driver being allocated.
- Overall, there is not much fluctuation in cancellation times, regardless of whether a driver is assigned or not.

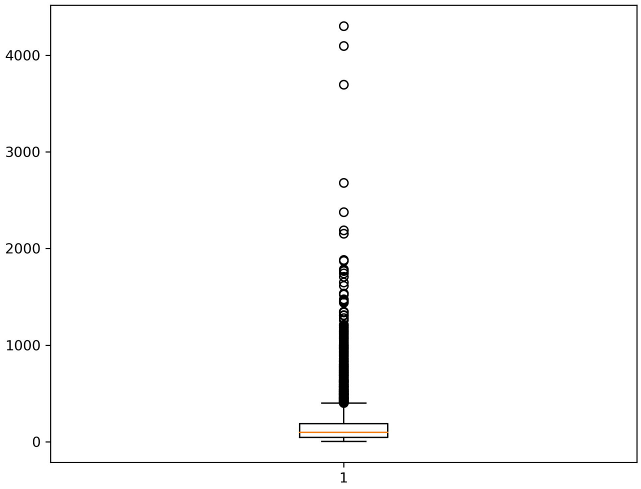

图 4: Boxplot of the data

Then we draw the boxplot to see whether there are outliers in the data. In this case, an outlier is defined as any data point that lies below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR, where Q1 is the first quartile, Q3 is the third quartile, and the IQR is the interquartile range.

According to this criteria, we can find that there is a big number of very large extreme values in the dataset. We need to drop them out.

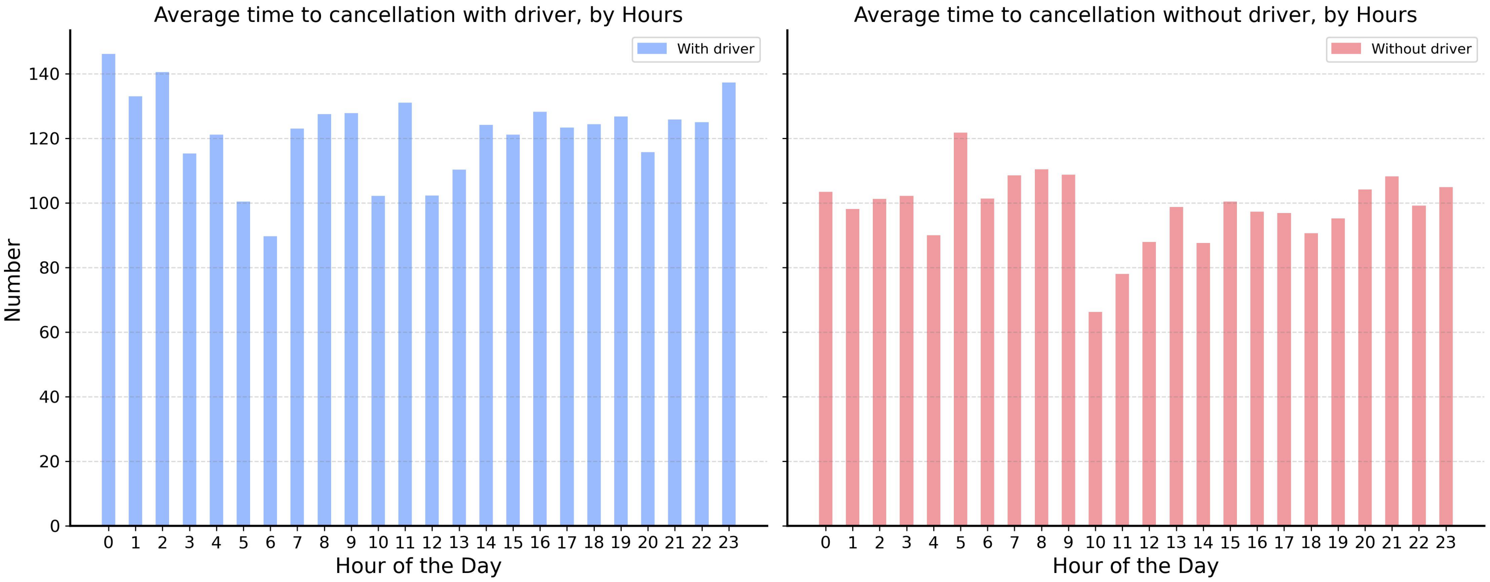

图 5: Average time to cancellation with or without Driver, by Hours (after data cleaning)

Here is the plot after deleting the outliers. By comparing these two charts, we can draw the following conclusions:

After removing the outliers, it can be observed that the average cancellation time for orders with a driver assigned has decreased significantly, yet it is still evident that the orders with a driver assigned have a longer average cancellation time.

4 Task 4

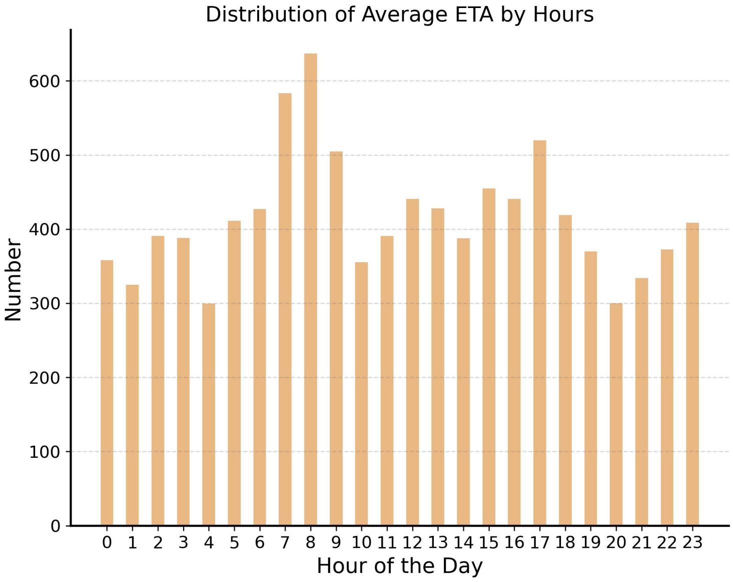

图 6: Distribution of Average ETA by Hours

From the above chart, we can observe:

- The average ETA does not show a linear relationship with the average number of failed orders in Task 2.
- During morning and evening rush hours (around 8 AM and 5 PM), the average ETA is higher. This suggests that during peak times, either a traffic jam or a larger number of orders causes drivers to take longer to reach their destinations.
- During late night to early morning, the average ETA is lower. It may be due to less traffic on the roads. Drivers can arrive at their destinations more quickly.
- Based on these findings, taxi companies could deploy more drivers during times with longer ETAs to improve service efficiency and customer satisfaction. Similarly, they could consider reducing the rate of empty rides during periods with shorter ETAs to save on costs.

5 Task 5

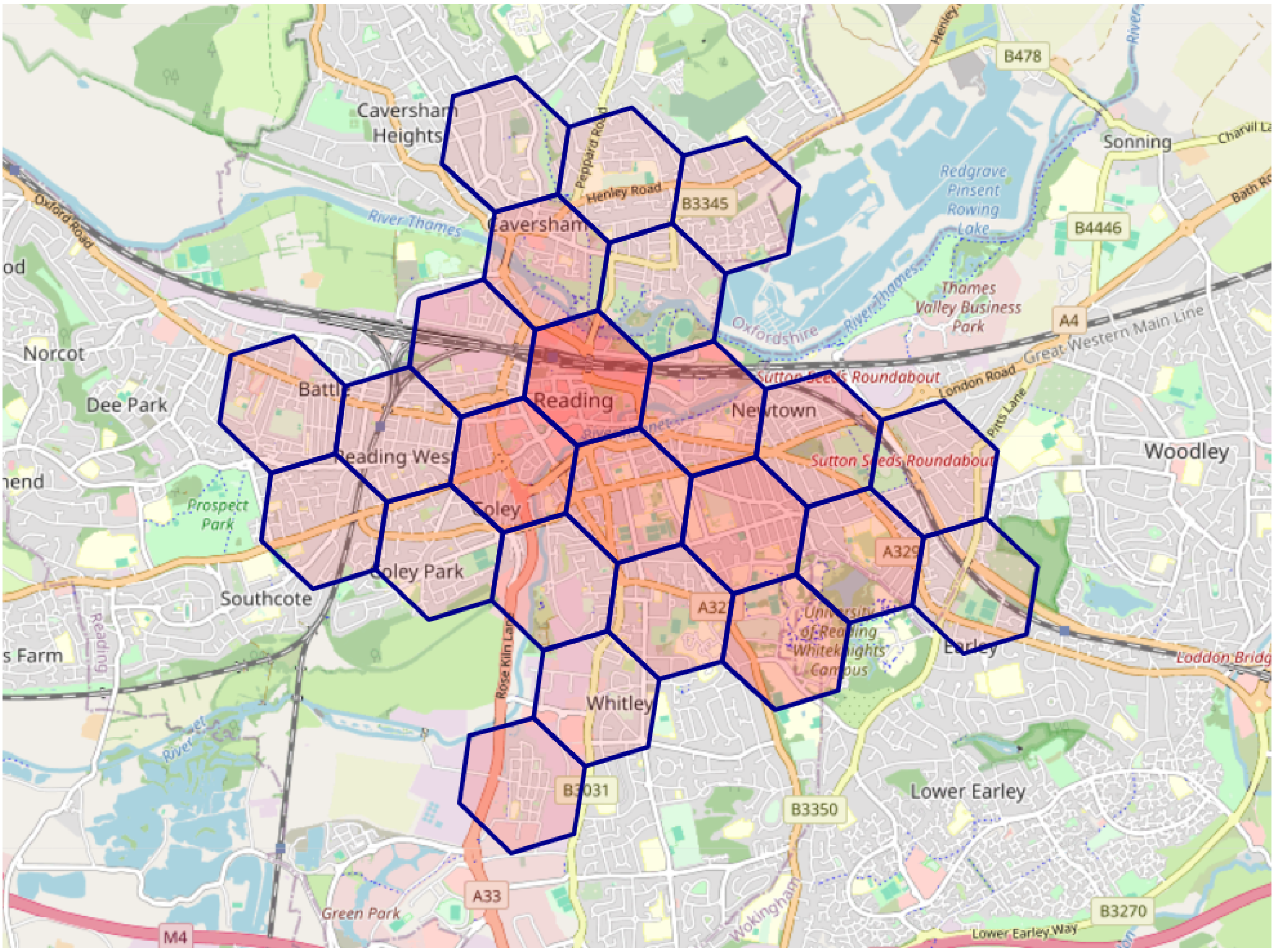

图 7: Sizes 8 hexes with 80% of all orders

We visualise the map and colour hexes by the number of fails. We can see that there are 24 sizes 8 hexes contain 80% of all orders. A large number of order cancellations occur near Reading.