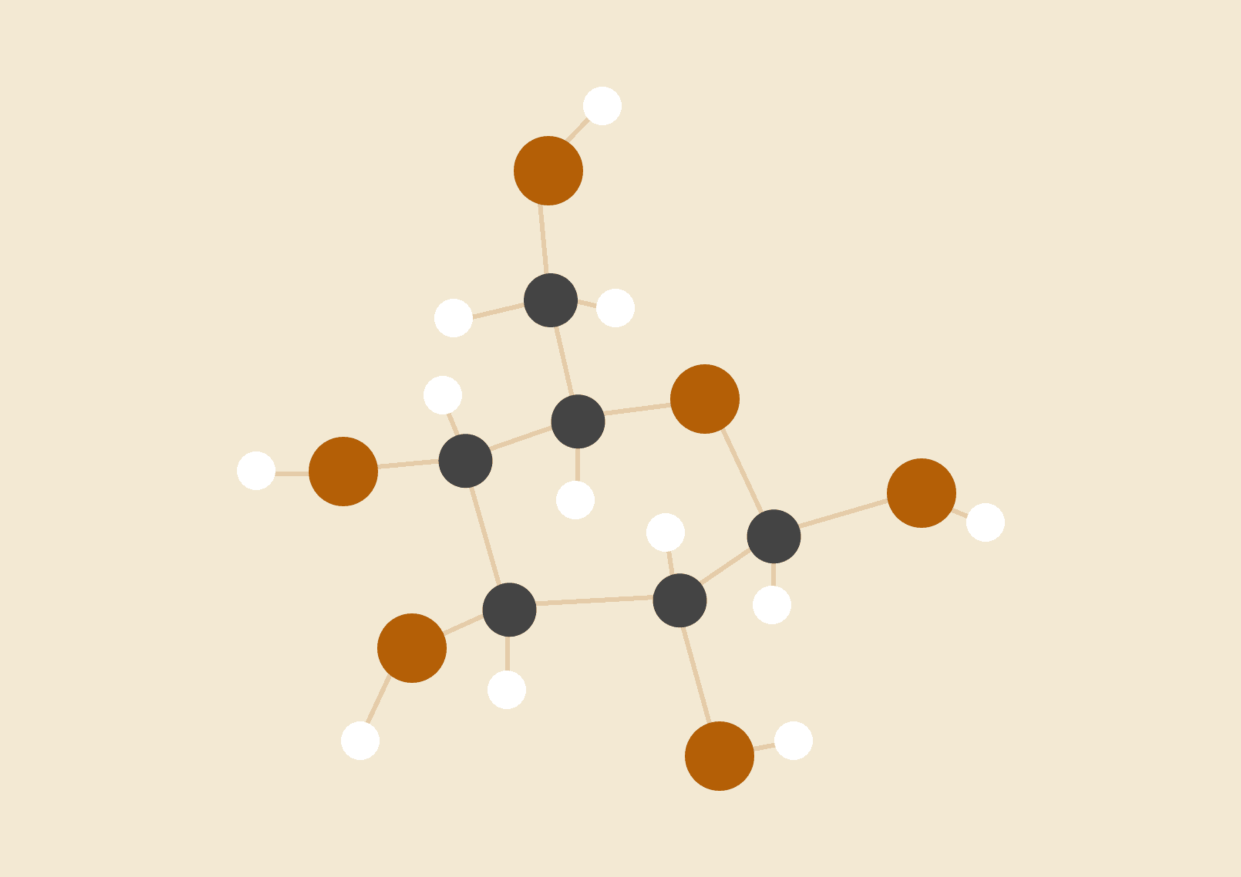
Cascade Cup 2022 Round 3

Rider-Driven Cancellation Data Analysis



Data Wizards

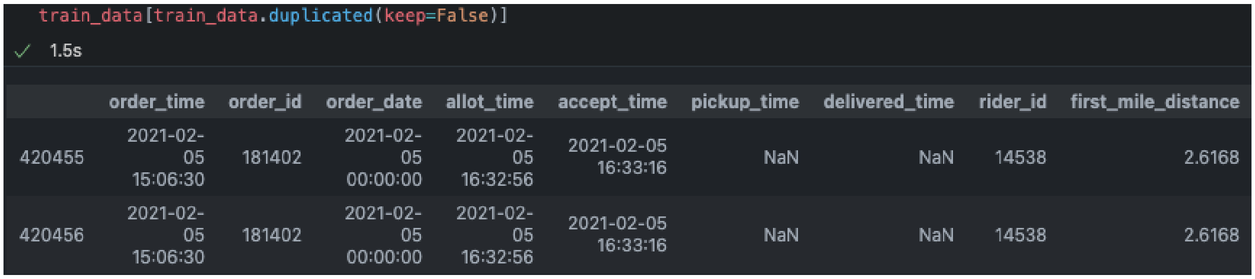
Adarsh Subramanian & Abhiram S

Indian Institute of Technology, Madras

INSIGHTS AND OBSERVATIONS

1. Duplicate Records and Invalid Values in Columns

Order ID **181402** has a duplicate record in `train_data.csv`. This order happens to be a cancelled order.

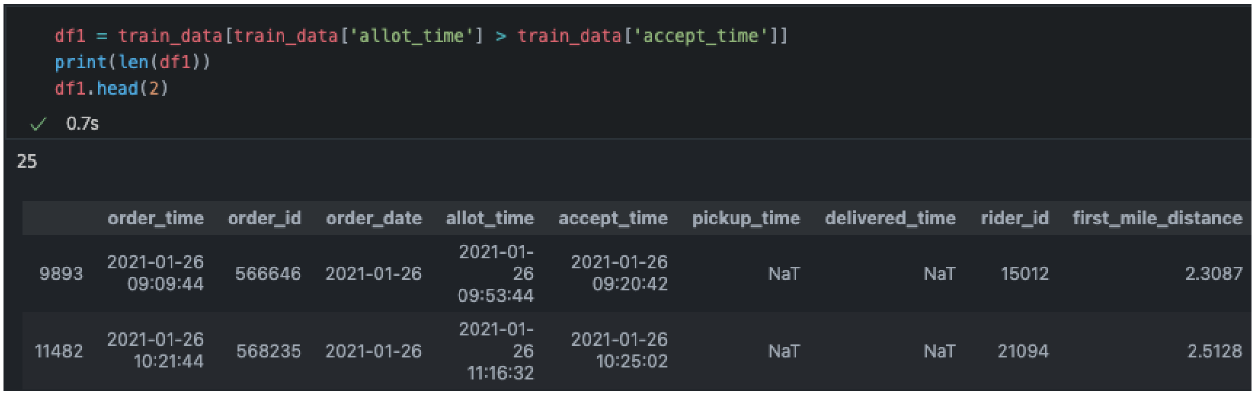


```
train_data[train_data.duplicated(keep=False)]
```

✓ 1.5s

	order_time	order_id	order_date	allot_time	accept_time	pickup_time	delivered_time	rider_id	first_mile_distance
420455	2021-02-05 15:06:30	181402	2021-02-05 00:00:00	2021-02-05 16:32:56	2021-02-05 16:33:16	NaN	NaN	14538	2.6168
420456	2021-02-05 15:06:30	181402	2021-02-05 00:00:00	2021-02-05 16:32:56	2021-02-05 16:33:16	NaN	NaN	14538	2.6168

There are 25 orders where `allot_time` is after `accept_time` and 10 orders where `accept_time` is after `pickup_time`. This is physically impossible.

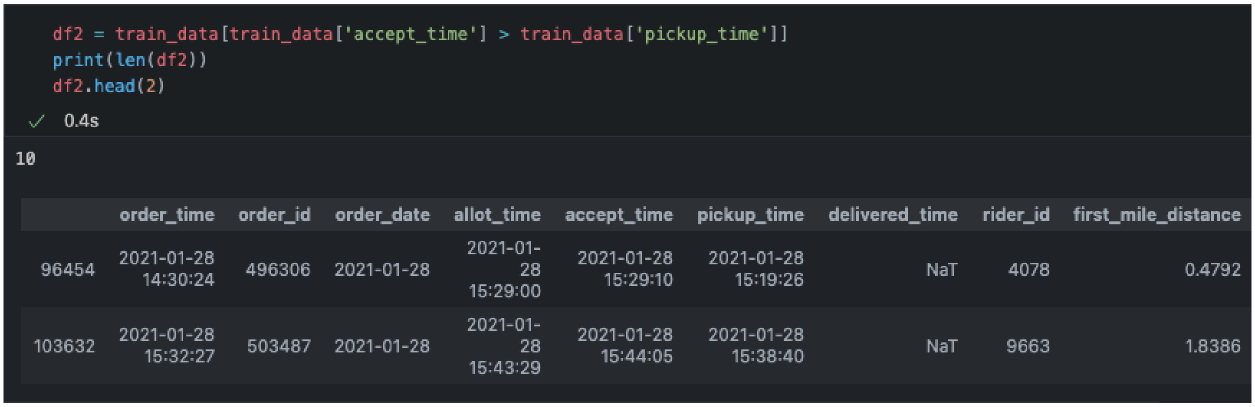


```
df1 = train_data[train_data['allot_time'] > train_data['accept_time']]
print(len(df1))
df1.head(2)
```

✓ 0.7s

```
25
```

	order_time	order_id	order_date	allot_time	accept_time	pickup_time	delivered_time	rider_id	first_mile_distance
9893	2021-01-26 09:09:44	566646	2021-01-26	2021-01-26 09:53:44	2021-01-26 09:20:42	NaT	NaT	15012	2.3087
11482	2021-01-26 10:21:44	568235	2021-01-26	2021-01-26 11:16:32	2021-01-26 10:25:02	NaT	NaT	21094	2.5128



```
df2 = train_data[train_data['accept_time'] > train_data['pickup_time']]
print(len(df2))
df2.head(2)
```

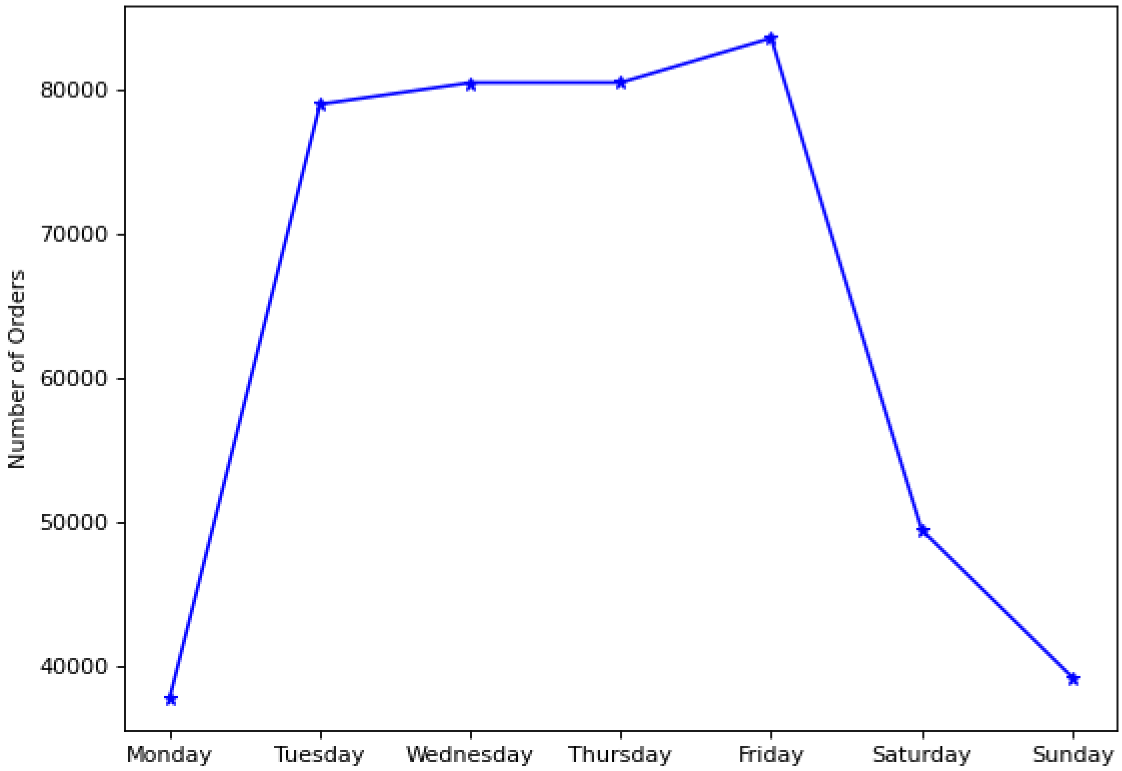
✓ 0.4s

```
10
```

	order_time	order_id	order_date	allot_time	accept_time	pickup_time	delivered_time	rider_id	first_mile_distance
96454	2021-01-28 14:30:24	496306	2021-01-28	2021-01-28 15:29:00	2021-01-28 15:29:10	2021-01-28 15:19:26	NaT	4078	0.4792
103632	2021-01-28 15:32:27	503487	2021-01-28	2021-01-28 15:43:29	2021-01-28 15:44:05	2021-01-28 15:38:40	NaT	9663	1.8386

These erroneous records might have been caused by some internal system error.

2. Number of Orders on Each Day of the Week

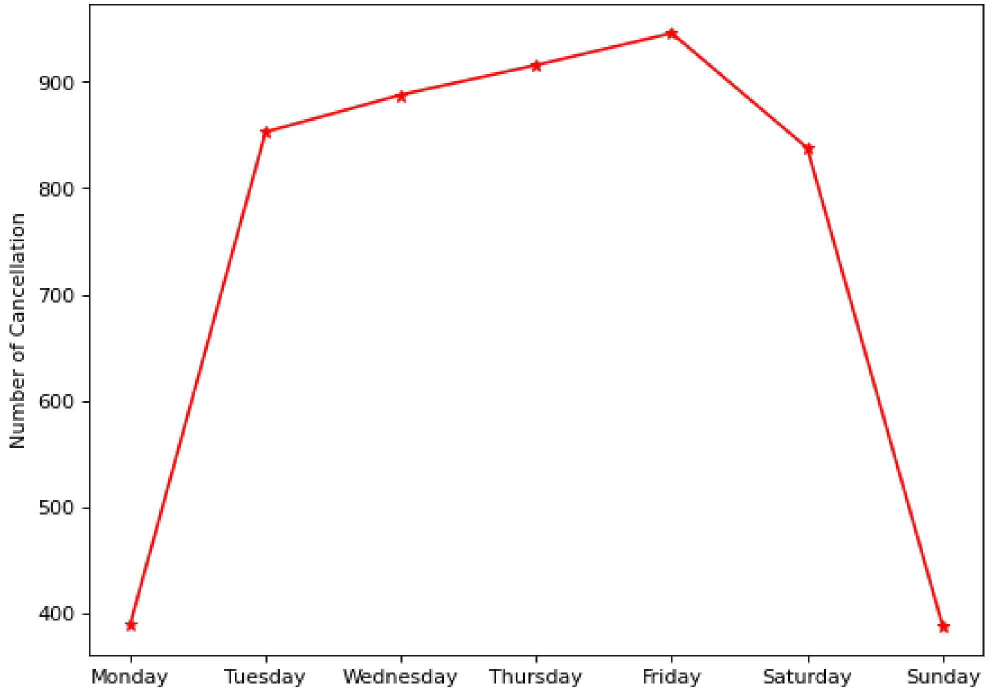


From the data, we can see that **the number of orders is highest during the weekdays, with the highest order on Fridays**. This can be logically deduced from the fact that people would be more busy during weekdays and won't get time to cook food at home. On weekends however, people prefer home food.

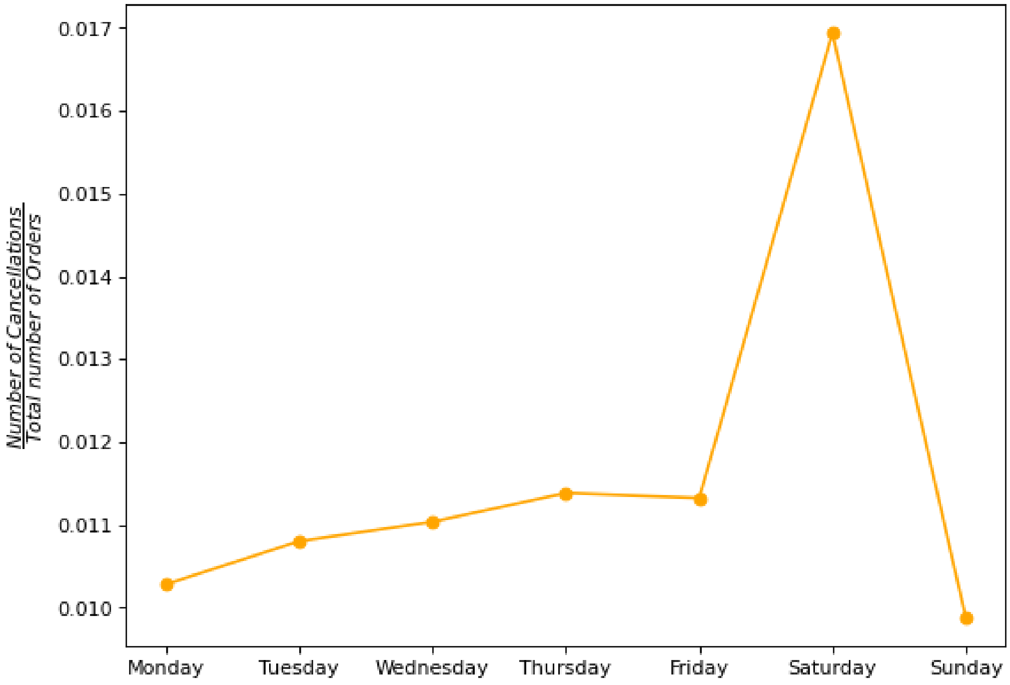
One specific point to note here is that **the number of orders is the least on Mondays**. This is probably because people are able to cook food/make preparations for cooking food on Sunday itself making it easier for them to cook on monday. This is not possible for the subsequent days.

3. Number of Cancellations on Each Day of the Week

Using the given data, we see that **the number of cancellations of orders is proportional to the number of orders — higher the number of orders, higher the number of cancellations** (which is expected). The trend has been plotted below:

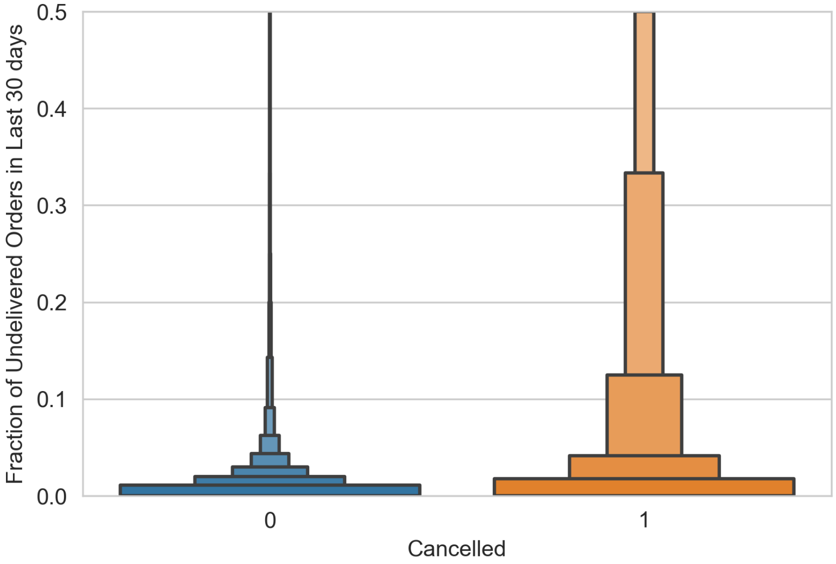


Herein, the proportionality trend is followed for all days but one. **On Saturdays, the number of orders is not very high (around 50000), yet, the number of cancellations is high**. On plotting the curve of $\frac{Number\ of\ cancellations}{Total\ Number\ of\ Orders}$ vs Day of Week,



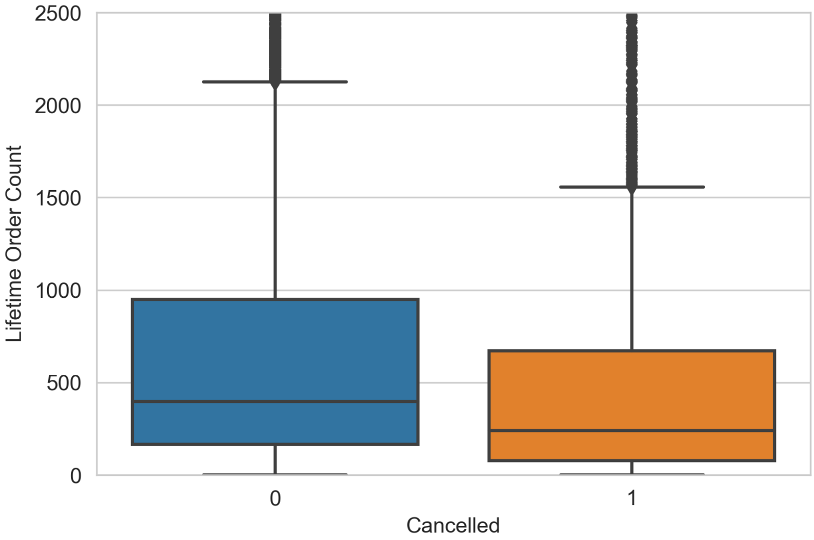
We see that this ratio is very high on Saturdays, indicating that the probability of an order being cancelled by riders on Saturdays is the highest. This information will be crucial as the operators can be well prepared for cancellations and re-allocations.

4. Influence of Rider History



The letter-value plot above shows the distribution of the fraction of orders undelivered by riders in the month prior to an accepted order.

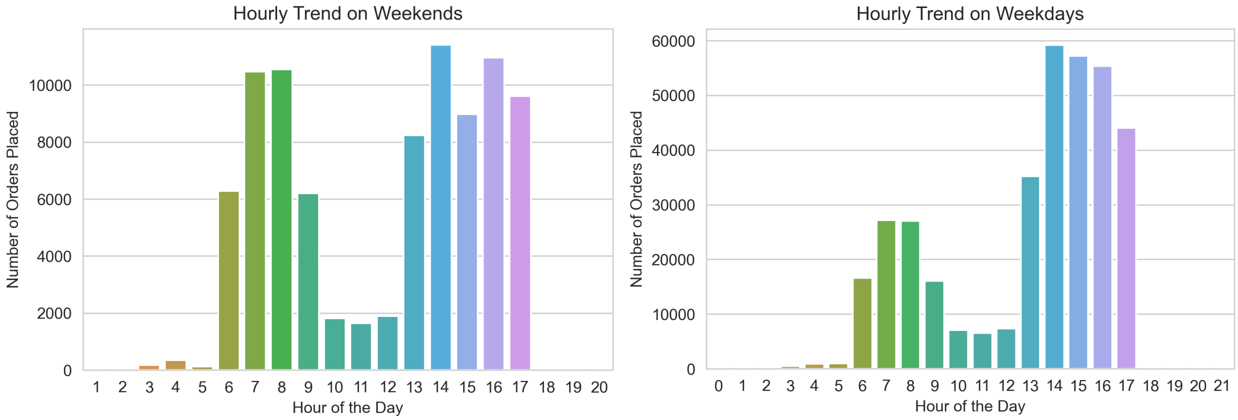
From the plot above, we can see that **riders who have already cancelled orders in the past month are more likely to cancel the current order**. This information is potentially useful in predicting rider-driven cancellations before they actually happen.



The box plot above shows the distribution of the number of orders delivered by a rider in their lifetime prior to an accepted order.

From the plot above, we can see that **riders who have delivered more orders in the past are more likely to successfully deliver the current order**. This information is potentially useful while reassigning orders after a rider requests to do so. Reassigning the order to a more experienced rider might ensure successful delivery.

5. Hourly Trend of Orders on Weekdays and Weekends

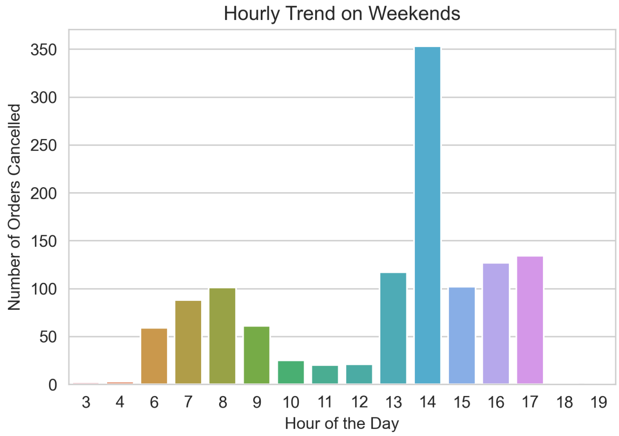


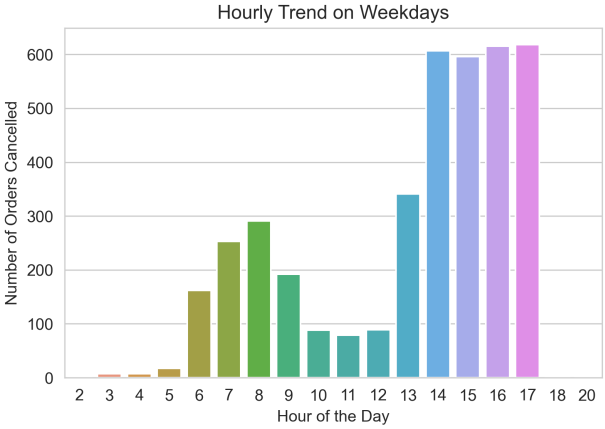
The bar graphs above show the number of orders placed during each hour of the day on both weekdays and weekends.

One trend is common to both the graphs: **the demand peaks during breakfast (6 AM to 9 AM) and lunch (1 PM to 5 PM)**. This could be because the people would be getting ready for work in the morning and as a result, they wouldn’t get time to prepare their breakfast or lunch.

One way the two graphs differ is that **during weekends, the demand is roughly the same for both breakfast and lunch** while **during weekdays, the demand is much higher for lunch than for breakfast**.

6. Hourly Trend of Cancellations on Weekdays and Weekends





The bar graphs above show the number of cancellations during each hour of the day on both weekdays and weekends.

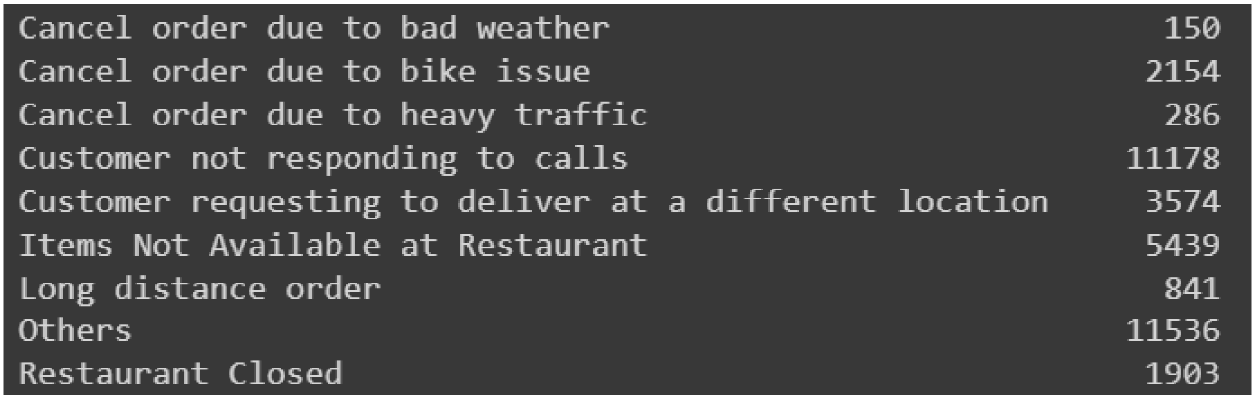
On weekdays, the trend is similar to that of the orders placed, i.e. **peaks during breakfast and lunch time, and the peak at lunchtime is higher than the peak at breakfast time**. This is what is expected, since cancellations cannot happen without orders being placed in the first place.

On weekends however, the trend is slightly different. **The cancellations still peak at breakfast time and lunchtime, however, the peaks themselves are at a similar height**. An anomaly here is that **the number of cancellations at 2 PM is significantly higher than during other hours**. This is likely due to a **one off event** happening on a particular Saturday and doesn't reflect the general trend.

7. Call Data Analysis:

From the call data pool, we see that

The following are the number of cancellations attributed to each reason:



```
Cancel order due to bad weather                            150
Cancel order due to bike issue                            2154
Cancel order due to heavy traffic                          286
Customer not responding to calls                         11178
Customer requesting to deliver at a different location    3574
Items Not Available at Restaurant                         5439
Long distance order                                        841
Others                                                   11536
Restaurant Closed                                         1903
```

We see that the most common reason for cancellation is of the “Others” category (which is presumably personal reasons).

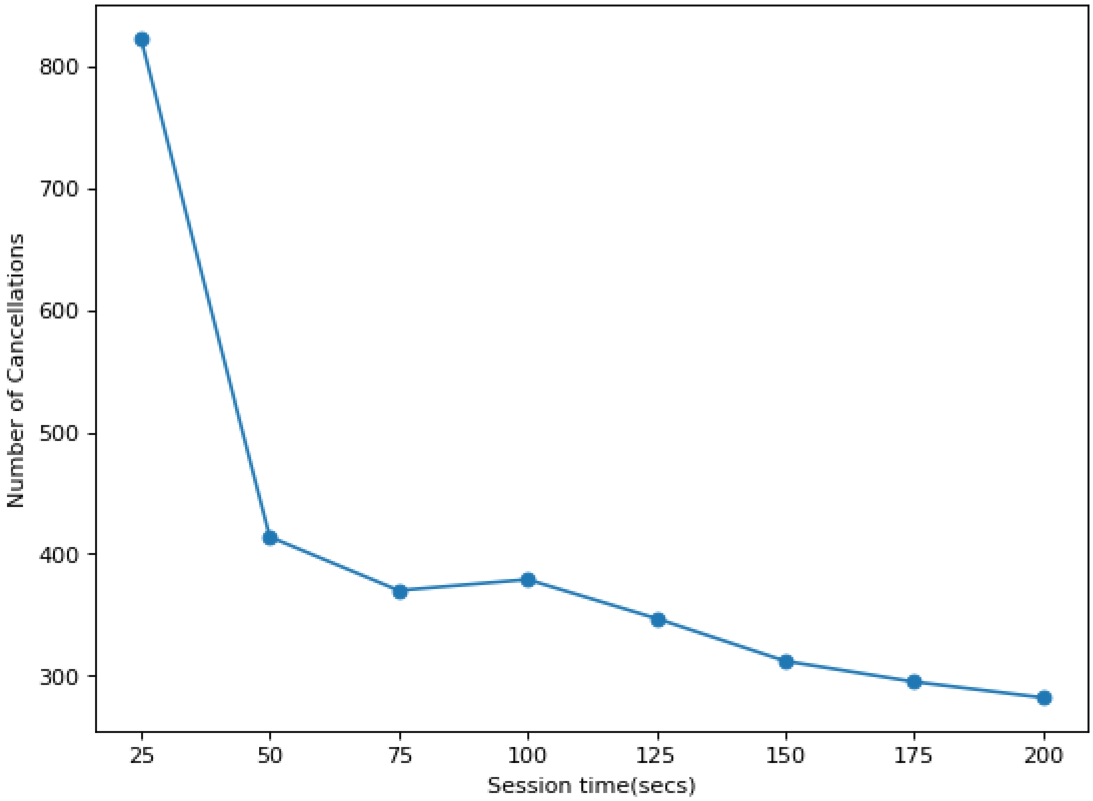
The 2nd most frequent **reason is that of the customer not responding to calls, and the frequency of this reason is pretty high.**We can also see that mitigating this issue can reduce a lot of cancellations. A few measures that can be taken include

- ➔ Asking the customer to fill in a back-up phone number
- ➔ Giving a chat option as well, for communicating with the customer

Another point to note is that, among the cancellations made under this reason, **more than 5000 cancellations were made just after 1 call not being picked up**. **A mandate should be enforced on the delivery agents to make at least 5 calls** to the non-responding customer before cancelling the order(from the data, around 7300 out of the 11178 cancellations made under this reason, were done with fewer than 5 call attempts).

8. Session time correlation with cancellation:

On analysing the data, we see that the riders tend to cancel rides initially; 823 cancellations have been made by riders who are in their 1st 25 seconds of session time. As the session time increases, the riders are more likely to deliver the order(as they would have fallen in the groove of the work). The trend is as follows:



Hence, the operators should be more prepared to handle cancellations from riders in their initial deliveries for the day.

Conclusion:

The data is very insightful and a lot of operational changes can be made using the trends in the data. We were able to see many more mild and semi-mild trends, but we have only included the most important ones in this report.

We had a great experience taking part in this competition. Our learning curve was steep and we got hands-on experience with challenging problems. All in all, it was a great, and enjoyable experience!

Thank You!