Team Noobs
Varun Nagpal and Sachin Gautam

Introduction to the Data

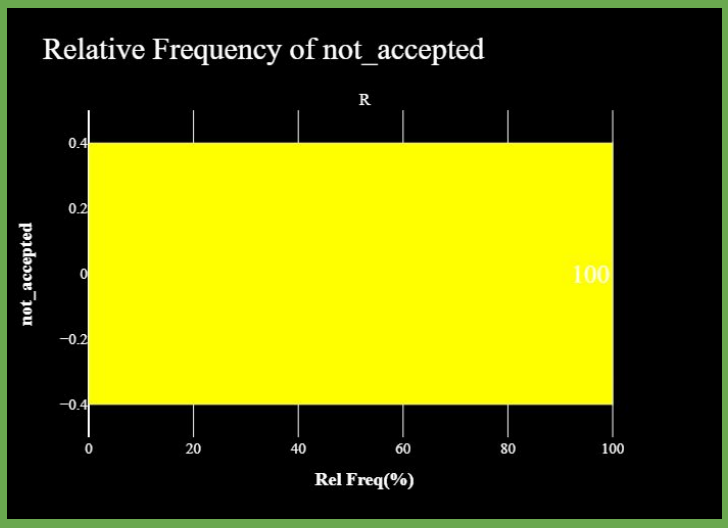
In this data analysis report we are going to be looking at the trends with a **major focus on the cancelled order data** and also compare it with the non cancelled data. **We will only be looking at the features where we saw some possibly significant trend.**

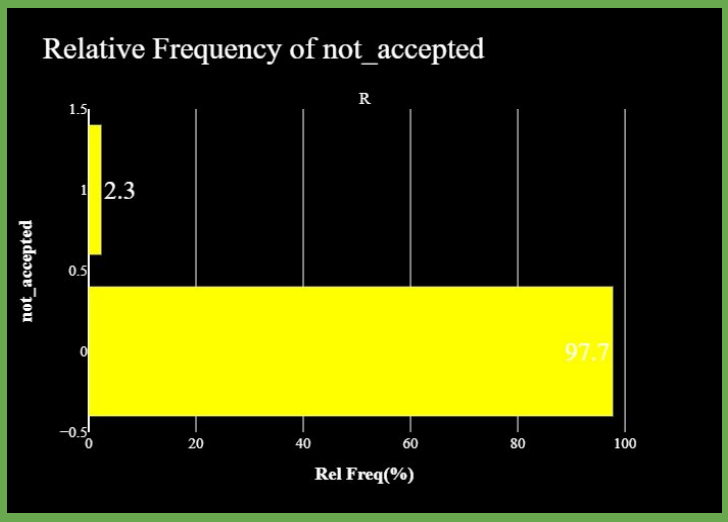
The target variable in this data ('cancelled') has 2 classes (0 and 1 i.e. not cancelled and cancelled) and the data is **highly imbalanced**(majority class is 0 and the ratio is around 85:1).

There were some 'data leakage' features like 'cancelled_time' which were dropped for this analysis.

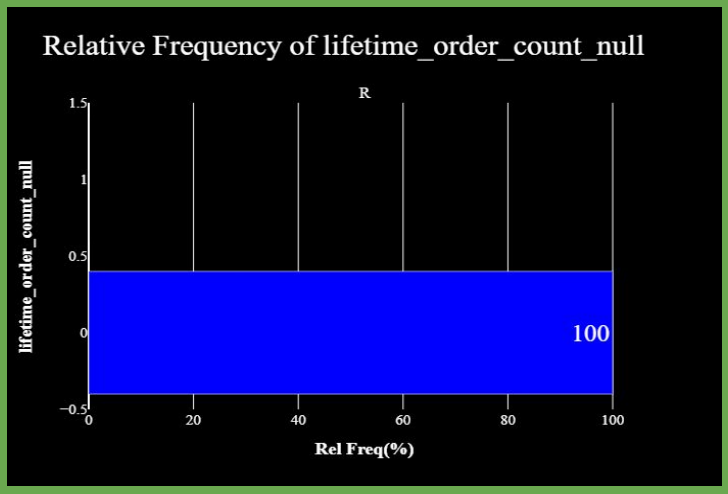
When it comes to the **null values seen in our data**, the null values in the **Reassignment Features** are only there where there has been no reassignment. Then, there are small amount of null values seen in **lifetime_order_count, session_time** and relatively significant amount in **delivered_orders, alloted_orders and undelivered_orders**. Since, undelivered_orders = alloted_orders - delivered_orders and so null values can be handled accordingly. There are a small amount of null values in **accept_time/ accept_date** which seem to be there if the order was not accepted. On the next page we further look at certain null values and possible errors

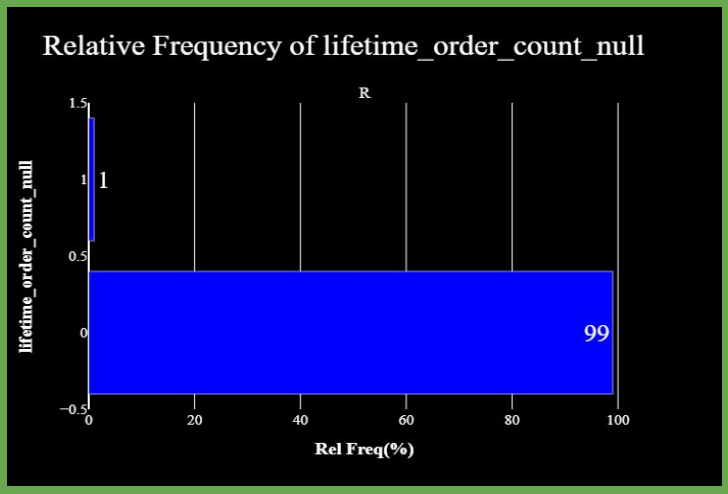
Here we have 3 engineered features which mark where there was a null value or seemingly wrong entry for certain features. We can see the distribution for all 3 is similar, 100% of the 0 class (non-null class/ right entry class) for non-cancelled data but a small percentage of the null class/wrong entry class in cancelled data.



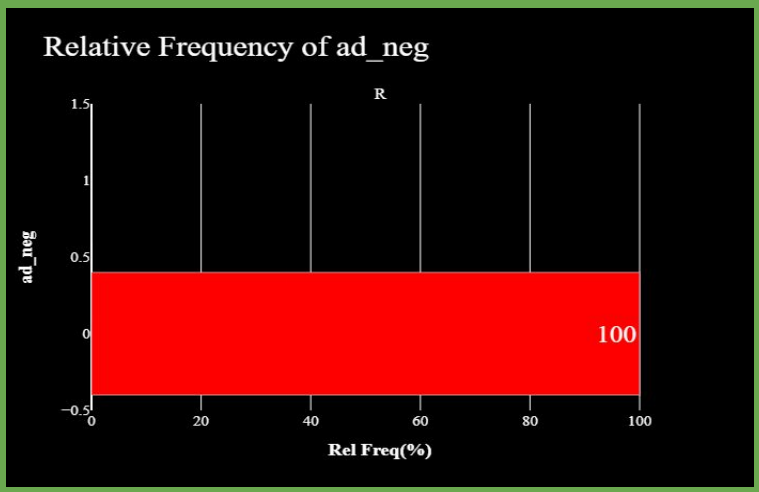


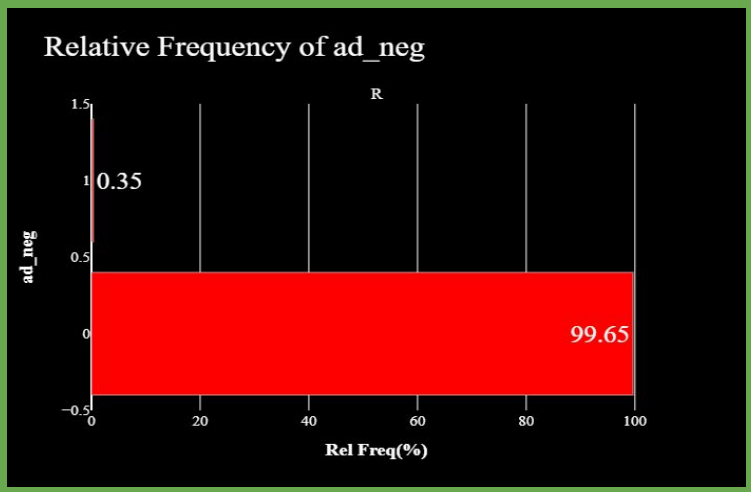
Not_accepted (accept_time)





Lifetime_order_count_null (lifetime_order_count)



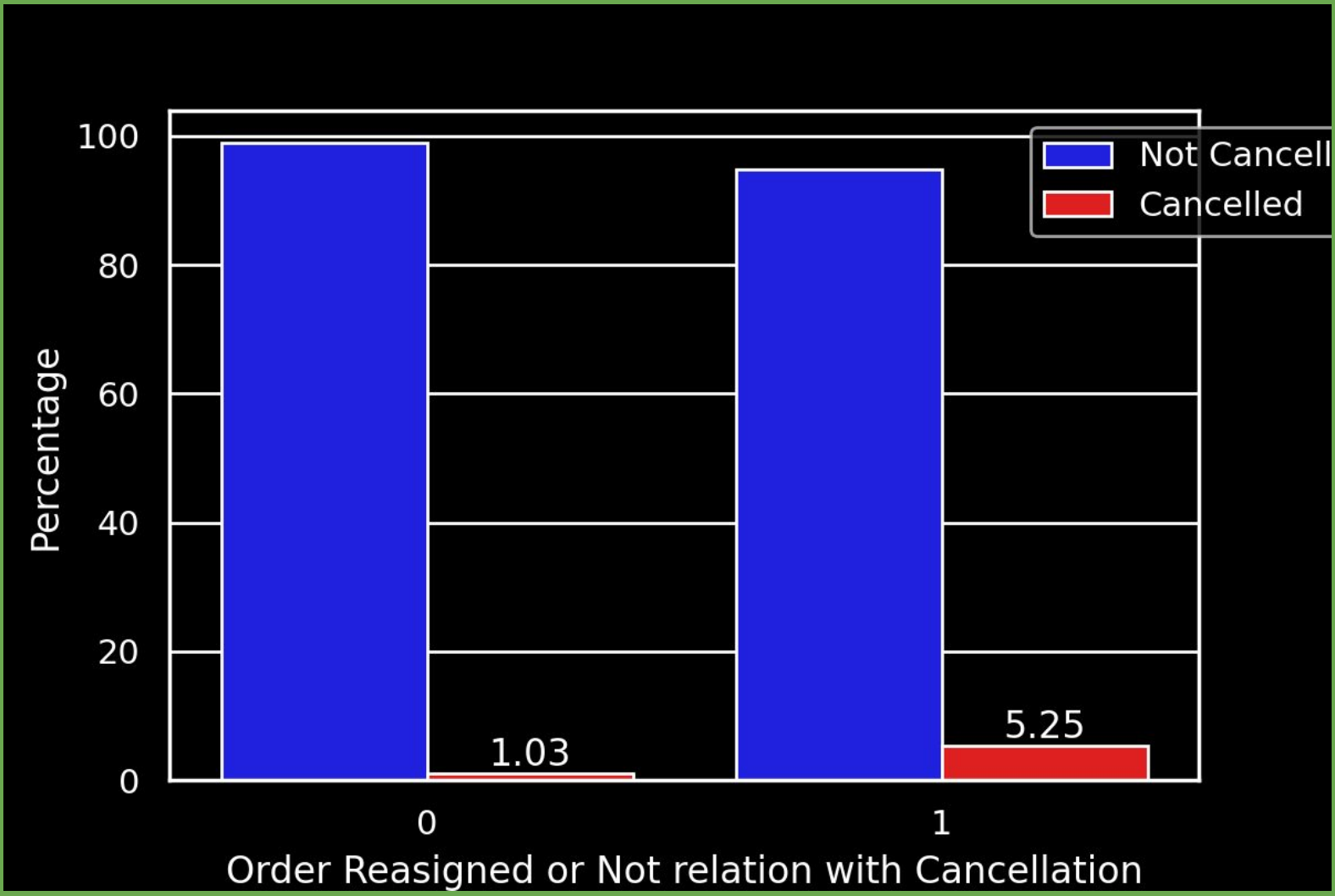


Ad_neg (acceptance_delay/ad)
(Error/Wrong entry where ad is -ve)

Inferences based on Univariate Analysis

1. Reassignment Features

   The **reassignment features were very imbalanced** indicating most of the orders in the dataset are not ones that were reassigned (for both cancelled and non-cancelled data).

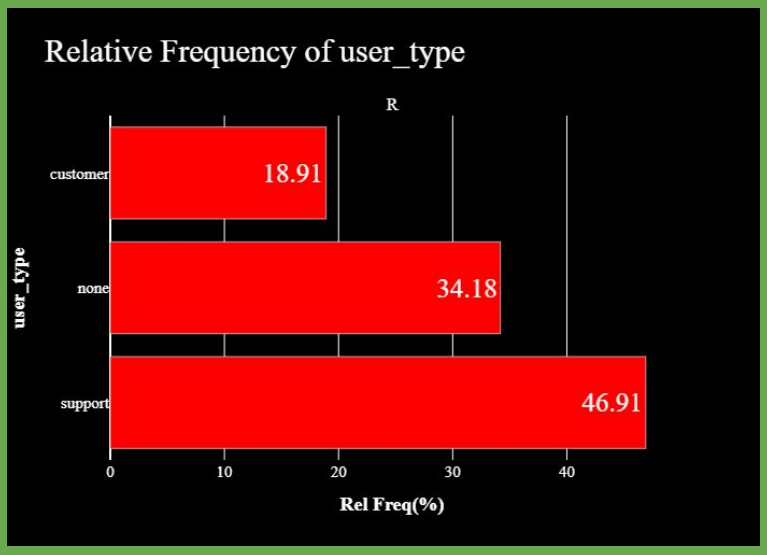


Here we can see that when order was reassigned, the **percentage of cancellation was much higher then when the order was not reassigned.**

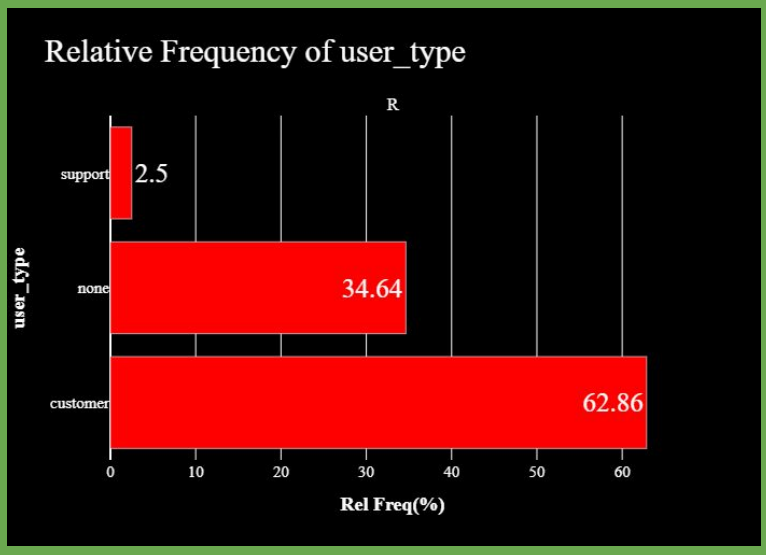
* 0 = Not Reassigned and 1 = Reassigned

2. User_type in Call_Data

For the **non-cancelled data**, we see from the call data that most of the orders involved communication between the **rider and the end customer**. The percentage of calls made to the support staff was very low. But for the **cancelled data** we can see that the majority of the orders involved a **call to the support staff**.



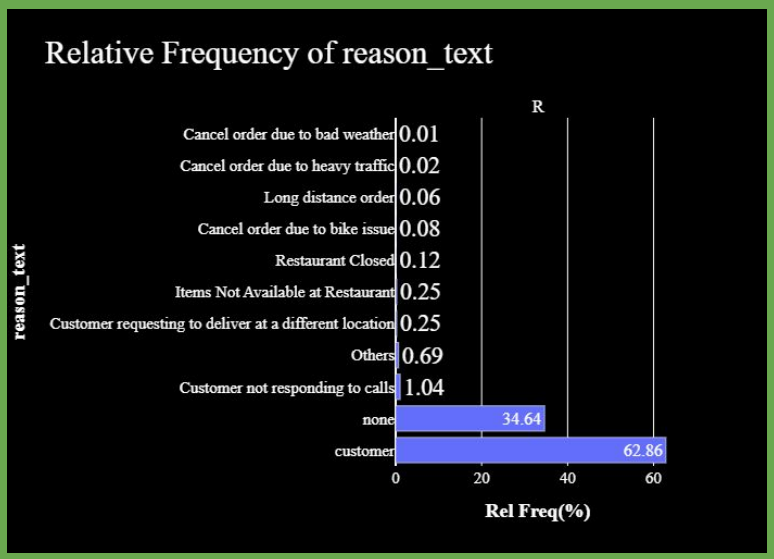
Cancelled Data



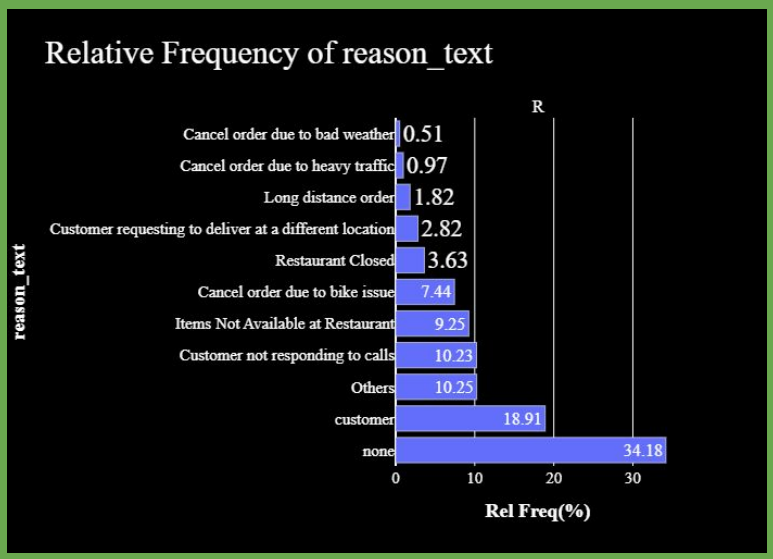
Non-Cancelled Data

3. Reason_text in Call_Data

For the **non-cancelled data**, most of the orders have a call b**etween the rider and customer only or there is no call.** But for the **cancelled data**, we can see that most of the orders have either **no call** or there is a good percentage of **calls with various reason texts which are very rarely seen for non-cancelled data.**



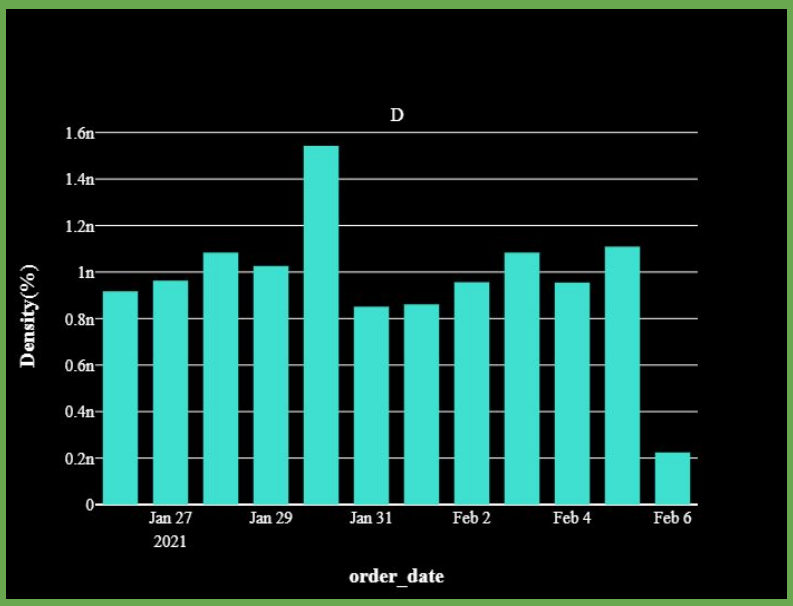
Non- Cancelled Data



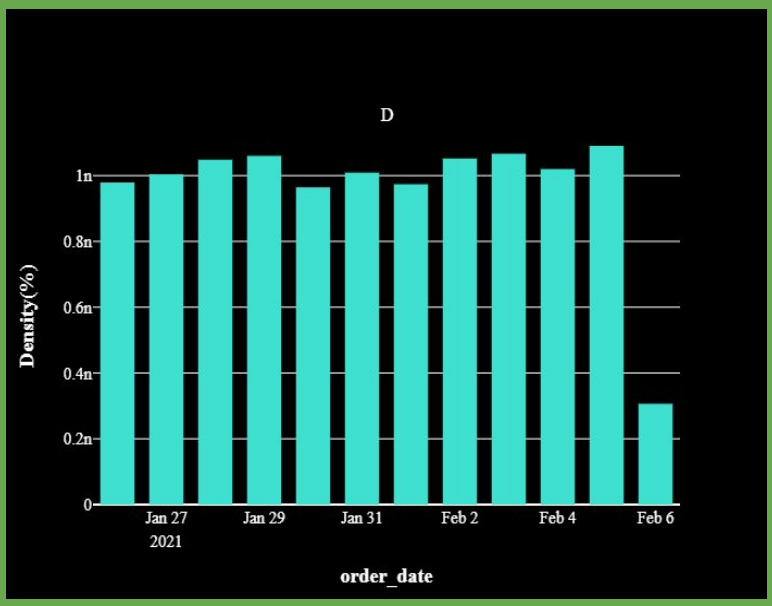
Cancelled Data

4. Order_date

For the cancelled data we can see that the distribution of orders is similar for all dates except 30th Jan and 6th Feb. We see that the orders were low on 6th Feb but this is explained by the fact that the data was not available for the whole day. On 30th Jan we can see that the number of orders cancelled had a higher percentage/ density while for non-cancelled data there was no such spike, in fact we see a minor dip. **So, we infer that there is a need to look into the events of 30th Jan more closely.**



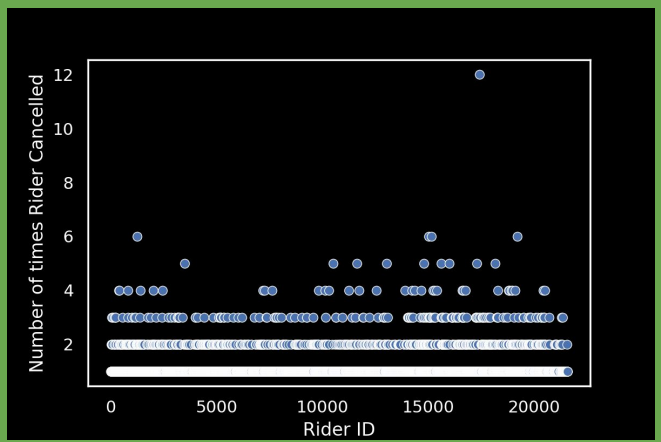
Cancelled Data



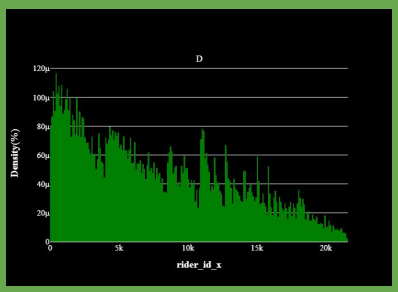
Non-Cancelled Data

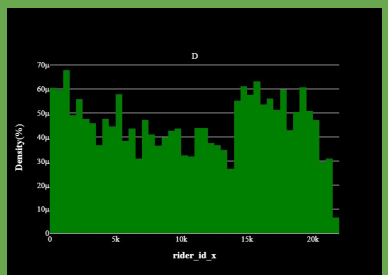
5. Rider_id

Most of the riders that cancelled, did it only once but there were riders that did it multiple times. In particular, **rider 17416 cancelled 12 times**. Also, there were **some that cancelled 5-6 time**s. Also, for rider id we can see that **density** of rider_id **increases** for **cancelled orders in the range 15k-20k** while it is continuously decreasing for non-cancelled data which s**hows higher cancellation in the mentioned range relative to non-cancellation**.



Non-Cancelled

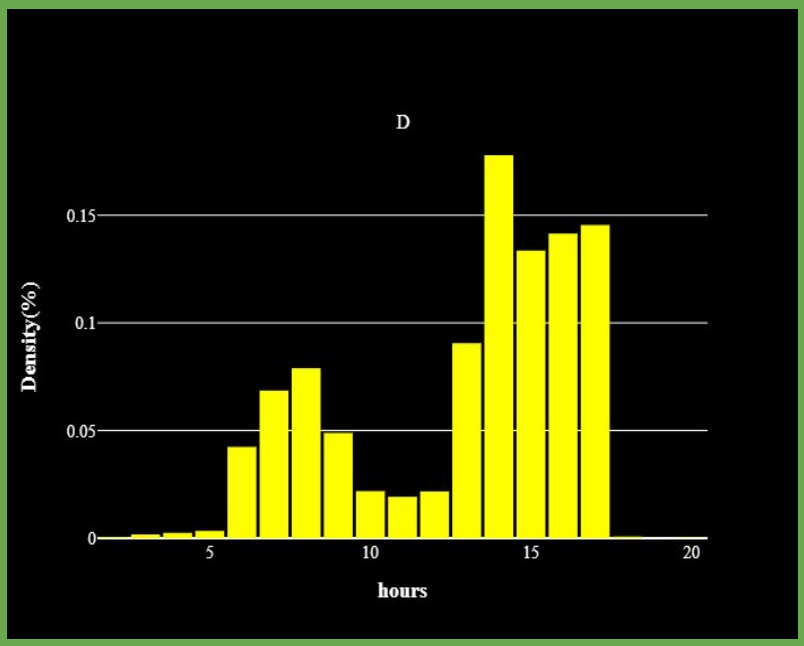


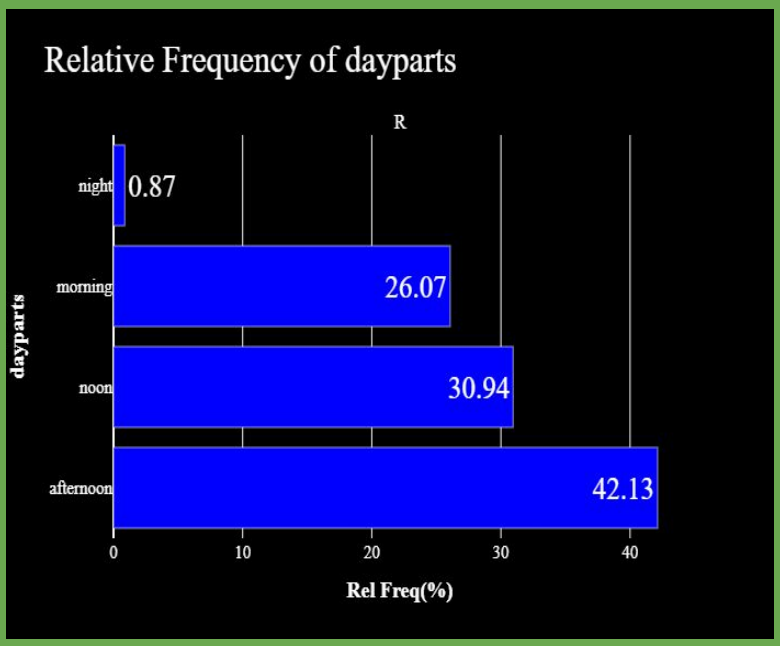


Cancelled

6. Dayparts and Hours of Day

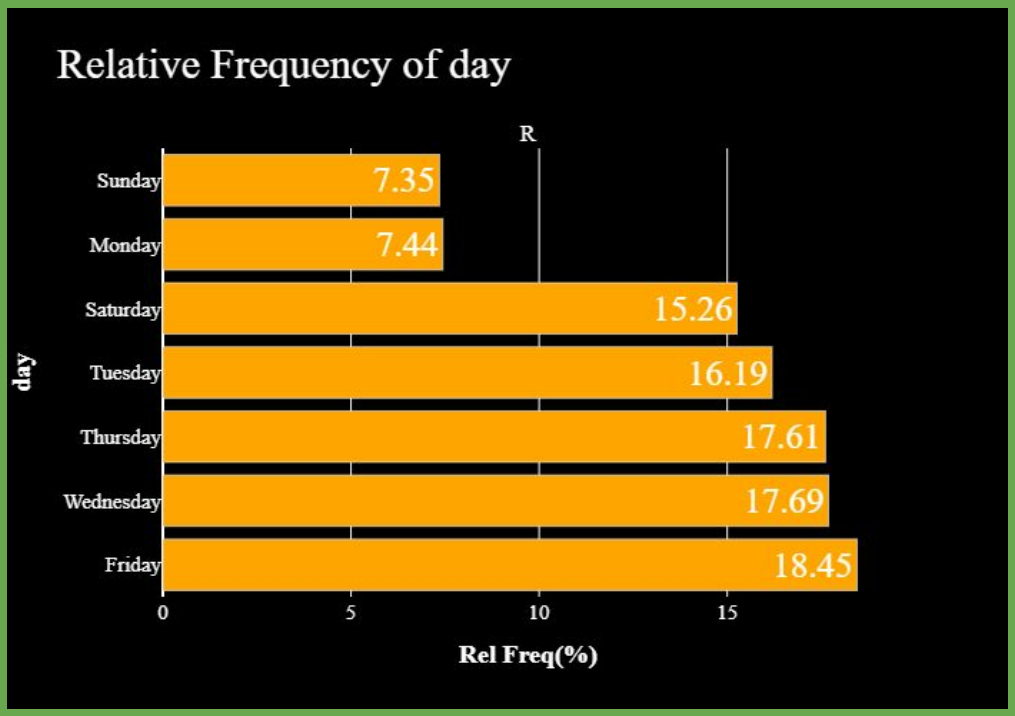
Most of the orders are seen during the afternoon hours and the minimum during the night hours. Majority of the orders are seen in the 1 pm to 5 pm timeframe and then in the 6 am to 9 am time frame. Similar trend is seen for both cancelled and non-cancelled data



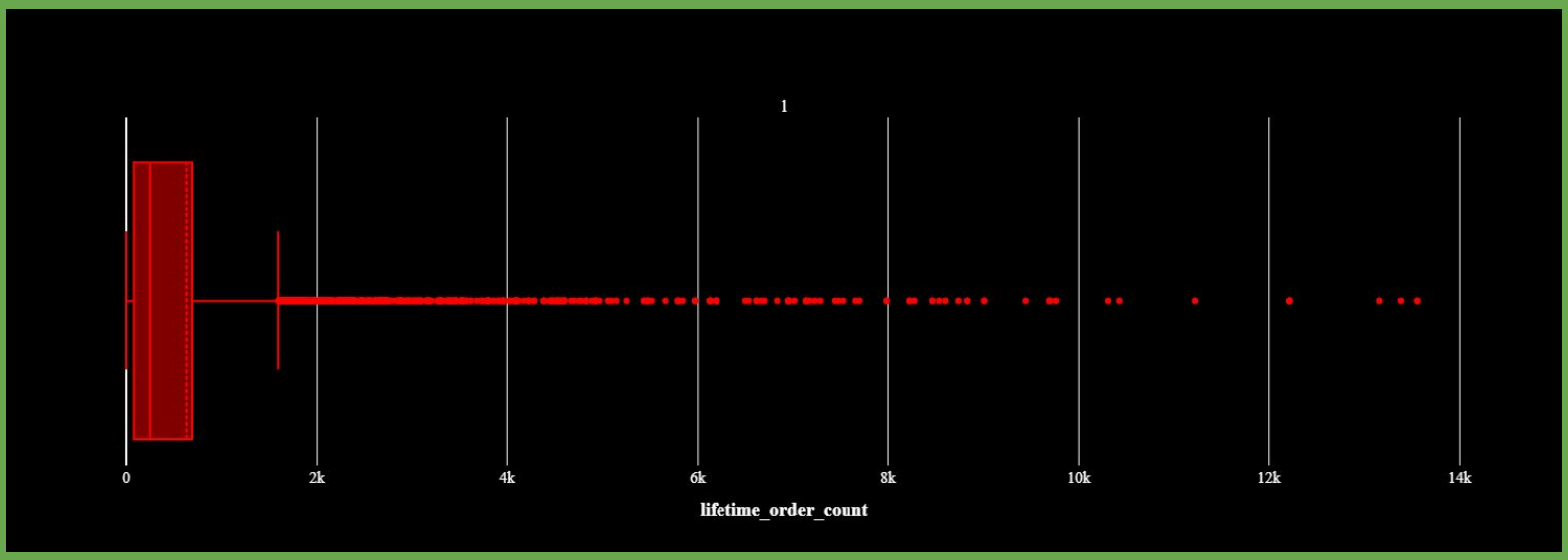


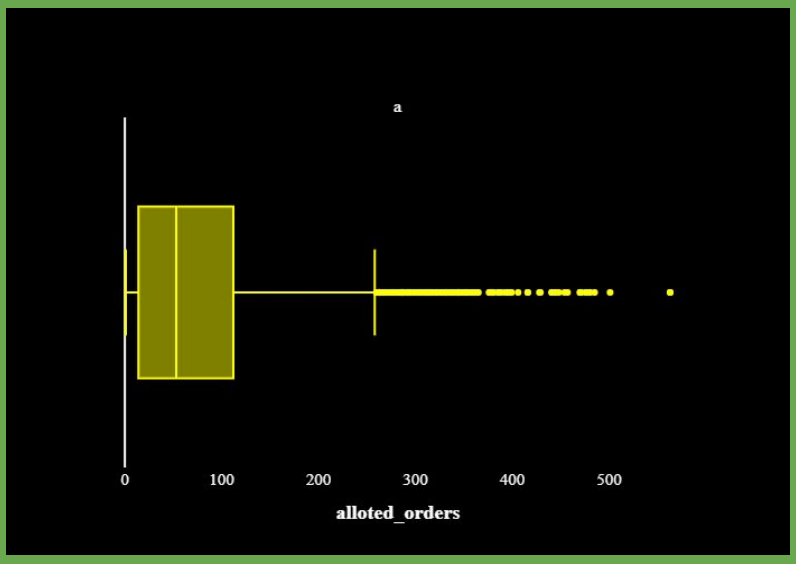
7. Day of Orders

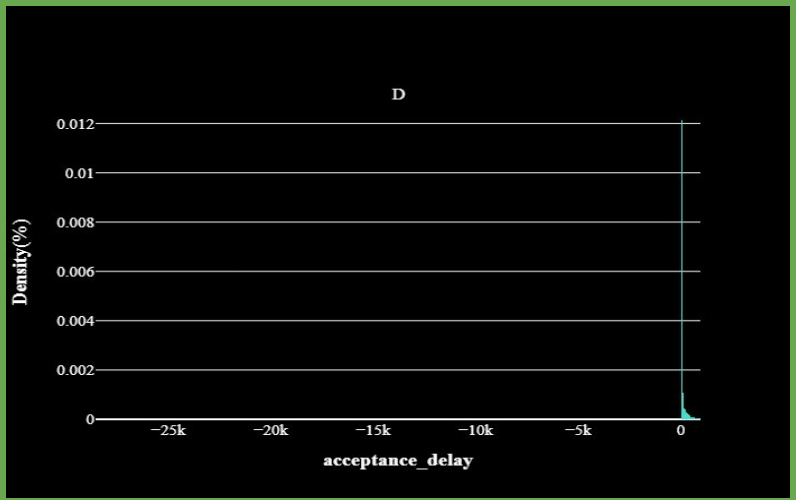
For the day feature we can see that **majority of the orders were seen on Friday and the lower end was seen on Sunday/Monday.** Similar trend was seen for both cancelled and non-cancelled data

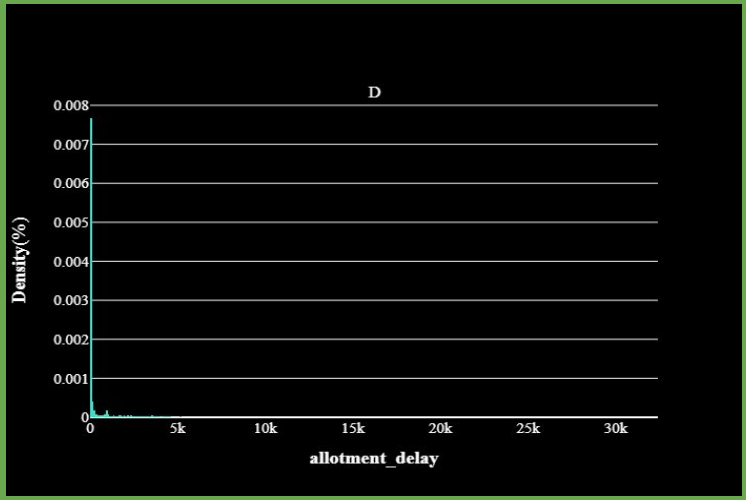


Outliers + Skewness in Continuous Features like alloted_orders, delivered_orders, lifetime_order_count,acceptance_delay,allotment_delay









Feature Engineering

Some of the engineered features include:

- Total_distance = first_mile_distance + last_mile_distance
- Day,hours,month, dayparts,is_weekend from datetime features
- Acceptance_delay and allotment_delay from differences in accept time and allot time , allot time and order time
- Bin_last , binning the last_mile_distance feature
- As mentioned earlier in our report, some features were also created to mark where there were null values in certain features (not_accepted, lifetime_order_count_null and ad_neg).

The Special feature engineering trick which gives some features highly correlated with the target variable:

- Aggregations of numeric features

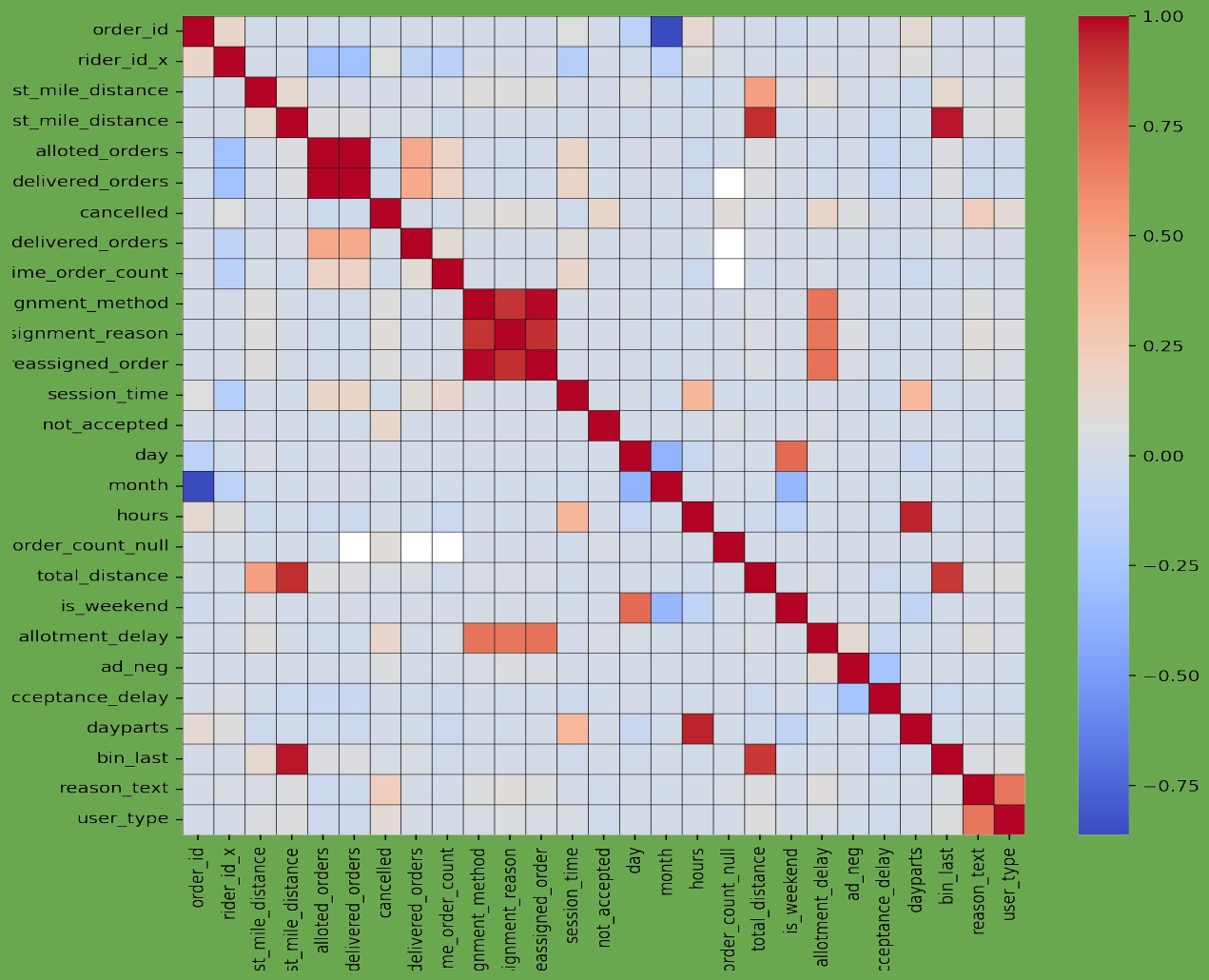
Computed statistics for all the numeric columns. To do so, we groupby the rider_id and merge the result back into the training data. The agg function calculated based on mean,min,max,sum and count operations.

- Categorical features

Like numeric features, operations are performed here on categorical features. The sum columns represent the count of that category for the associated client and the mean represents the normalized count.

Inferences based on Multivariate Analysis

Correlation Plot



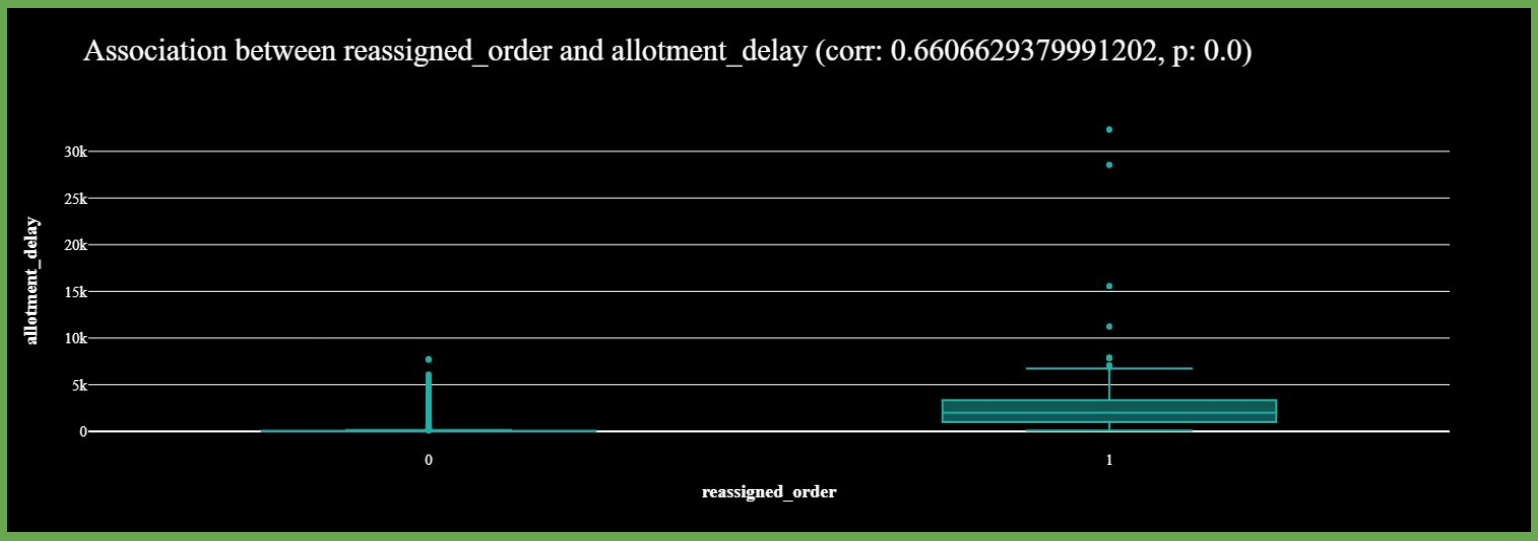
Some of the obvious strong correlations seen are between : (first_mile_distance,last_mile_distance, bin_last and total_distance),(alloted_orders, delivered_orders and undelivered_orders), (user_type and reason_text).

Other Strong correlations seen are:

- Order_id and month (Order_id is negatively correlated with time in general, which is unexpected because one would expect a positive correlation)
- Reassignment features and allotment_delay
- Rider_id and alloted_orders and delivered_orders
- Month and is_weekend (Since we only have a limited number of months and days in dataset)
- Session_time, dayparts and hours (as the day progresses, session time can only increase and hence we see a positive correlation here)
- No strong correlation was seen with the target variable. Allotment_delay, undelivered_orders and not_accepted had moderate correlation. Also, from call_data , reason_text had decent correlation.

Allotment_delay and reassigned_order and relation with Cancellation

Allotment delay tends to be higher if the order was reassigned for both cancelled and non-cancelled data . We perform **point biserial correlation** test to confirm this. A p-value of < 0.05 rejects the null hypothesis that there is no difference between the reassigned and non-reassigned orders for cancelled data.
Also, the median and mean were higher for cancelled data than non-cancelled data for allotment delay. We perform **t-test** here and see p value of <0.05 which tells us that there is significant difference in allotment delay between cancelled and non-cancelled orders .



Thank You!