# Data Analysis Report
# Round-3 Cascade Cup 2022

Team Nerdy Duo
Rajdeep Agrawal, Nisarg Shah

February 2022

## 1 Introduction - Overview Of Data

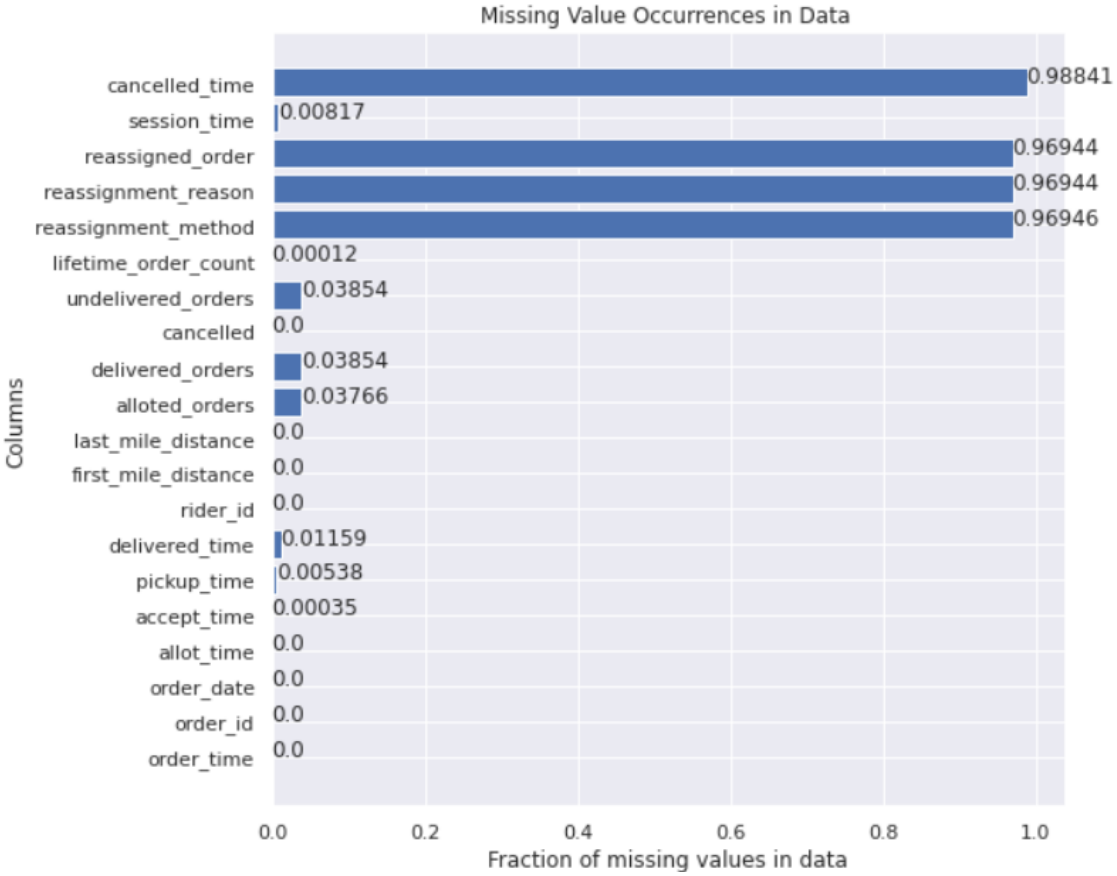
Two datasets have been provided, **train_data.csv** - which contains the target variable **cancelled** and other predictors, and **call_data.csv** as metadata of calls recorded from riders to customers or support. There are **4,50,000** observations in the former, of which one was duplicate and was dropped. Each observation has a unique order_id and other features like order_time, allot_time, accept_time, pickup_time, delivered_time, cancelled_time, etc. We won't describe the features here, however one can easily understand them as the analysis unfolds.

### 1.1 Class Label Distribution

So, this is a **classification** task to predict if a certain order has been cancelled or not based on the provided features. From below Figure, it can be clearly observed that there is heavy imbalance in class labels, with close to 99% being 0 (not cancelled).

### 1.2 Missing Value Occurrence

Now, lets have a look at missing values by column in our data.



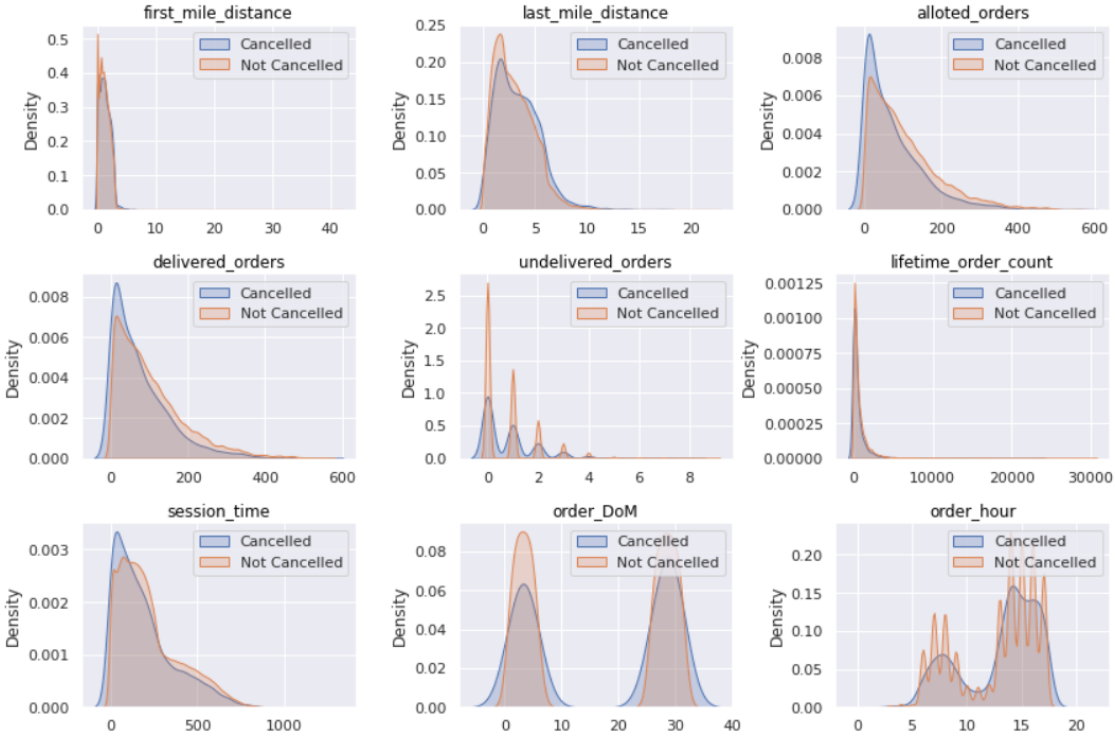

It is evident that multiple columns have missing data, some in excess of 90% (cancelled_time, reassigned_order, reassignment_reason, reassignment_method) while some as low as around 1-5% and below (delivered_orders, undelivered_orders, alloted_orders, delivered_time, pickup_time, accept_time, session_time, lifetime_order_count). While there may be a strong case for features with high percent of missing values to be dropped, we will later try to encode the NaN values to see if absence of data has anything to do with the output.

## 2 Analysis of Numerical Features

### 2.1 Distributions of numerical features by Class

Now, we analyse the different features their variation wrt the target variable. Starting with numerical features and their continuous Kernel Density plots (KDE

plots). Apart from the existing numerical features, we have also extracted the time coordinates of the order (day of week, day of month, month, hour). We use day of month and hour as numerical features.

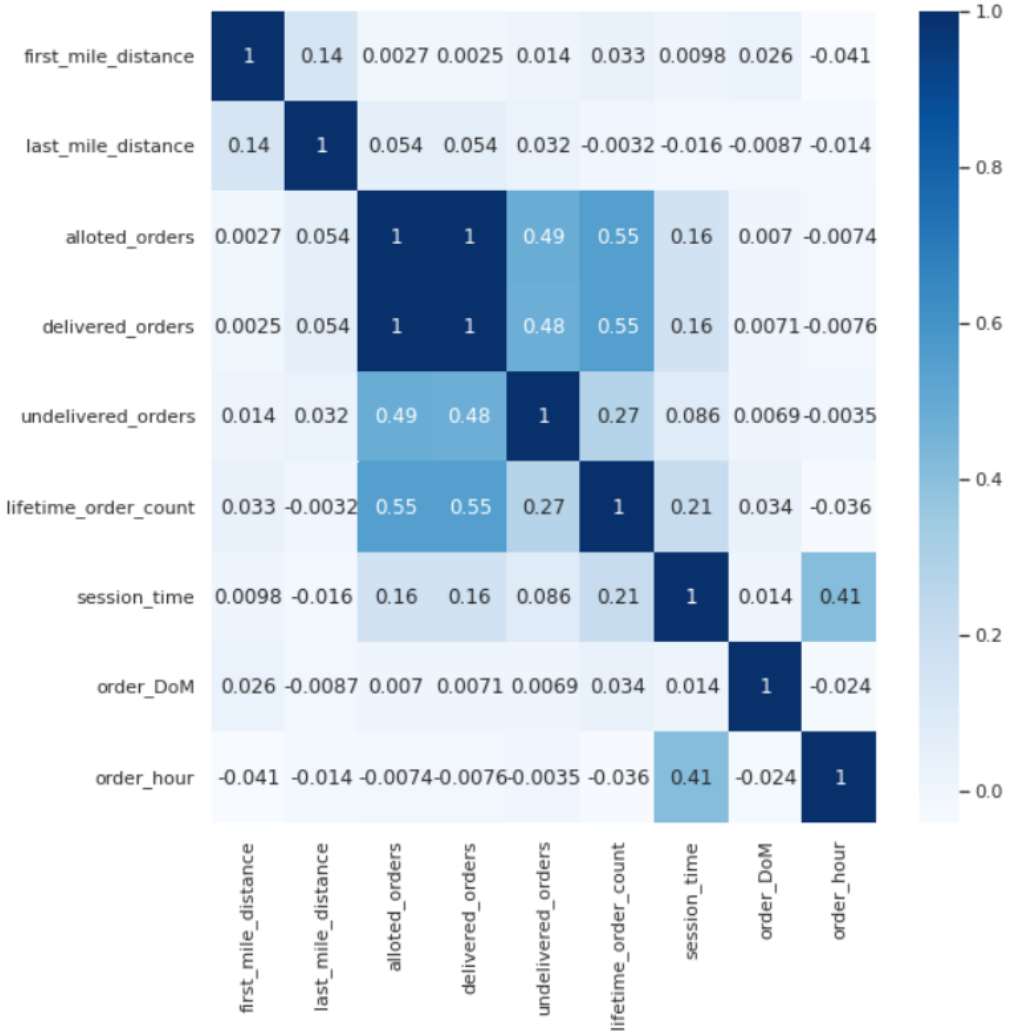



Using these plots, we get a fair idea of the range of continuous numerical features and also what kind of values are more frequent, i.e, have a higher kernel density. In addition to that, we have contrasted the same for both output classes by overlapping the curves. Some key observations include:

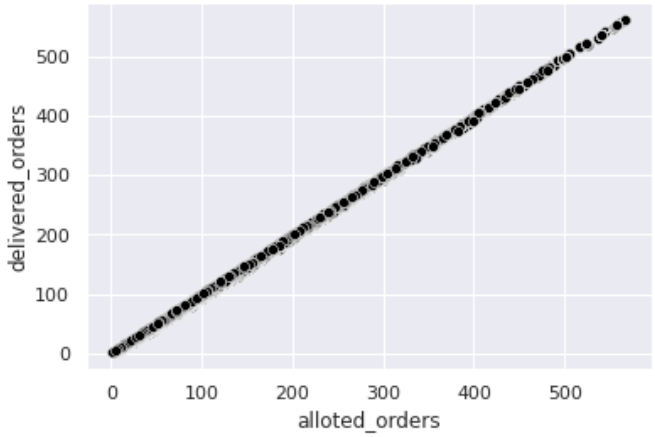
- **lifetime_order_count** is heavily right-skewed for both target labels, with a sharp peak towards the lower part of the range. Same could be said of **first_mile_distance** as well.
- **order_DoM** (day of month) represents a bimodal distribution for both target variables, however the width is larger when orders are cancelled, and peak is larger when not cancelled.
- The distributions for **alloted_orders** and **delivered_orders** are nearly identical (just keep this in mind for later!), while being fairly right-skewed too.
- One can observe a number of very sharp peaks in **undelivered_orders** and **order_hour** when orders aren't cancelled. This indicates concentration of the data to certain localised small ranges of values.

## 2.2 Correlation and Colinearity

Let us examine a correlation matrix for all the above numerical features:

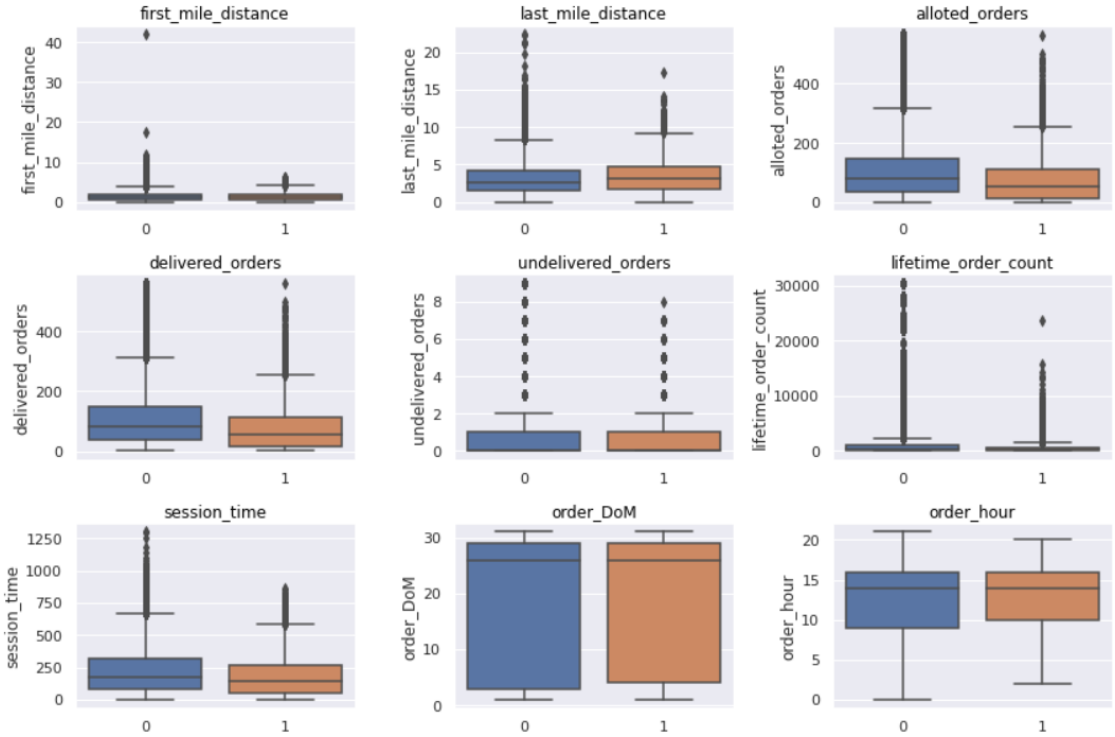



In the above symmetric correlation matrix, there seems to be only incidence of high correlation (or anti-correlation) - between **alloted_orders** and **delivered_orders**. (Remember point 3 from the previous subsection?) Hence, it might be prudent to drop one of the two, depending on the algorithm to be used. We can further validate this by plotting these 2 features together as shown in the scatterplot below:




Thus, we have now doubly confirmed the said perfect correlation.

### 2.3 Boxplots for Numerical Features




Boxplots give us an accurate representation of basic descriptive statistics of numerical features - median, IQR, maximum & minimum values and outliers. Here we plot boxplots for both target classes (0 - Not Cancelled, 1 - Cancelled) for each variable for comparison. Some key observations:

- Values in **first_mile_distance** and **lifetime_orders_count** are very closely grouped on the lower side of the range as evidenced by the compressed box-

plots, in agreement with the conclusion from the KDE plot in 2.1.

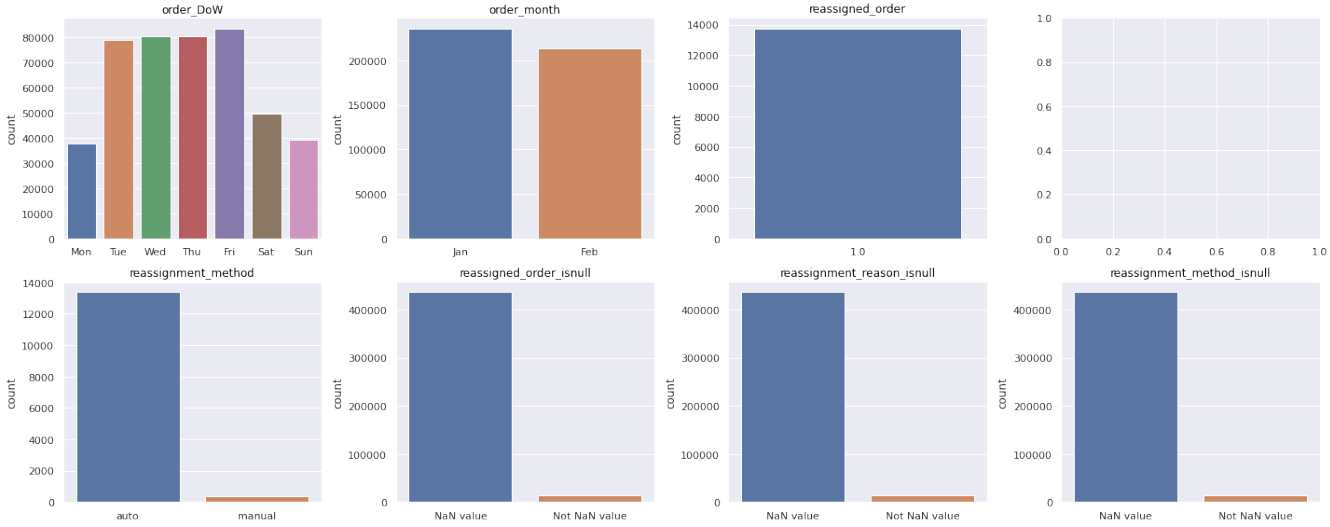
- The median day of month and order hour is roughly same for both classes. The elongated boxes for **order_DoM** imply a decent variety in values, in stark contrast to the case in point 1. The elevated position of median suggests a greater diversity of values in 2nd quartile as compared to 3rd quartile.
- The median **alloted_orders**, **delivered_orders**, **session_time** are a touch lower when orders are cancelled and the upper whisker being longer than the lower one suggests greater diversity of values in 4th quartile as compared to 1st quartile.
- In case of **undelivered_orders**, the median almost merges with the first quartile entirely for both classes.

# 3 Analysis of Categorical Features

Now lets have a look at some of the categorical features - including day of the week (order_DoW) and month (order_month). Also, for the reassignment features, we have encoded the NaN values to create an additional Boolean feature each.

## 3.1 Count of various Categories

A brief look at the various categories and their counts for each of the features:

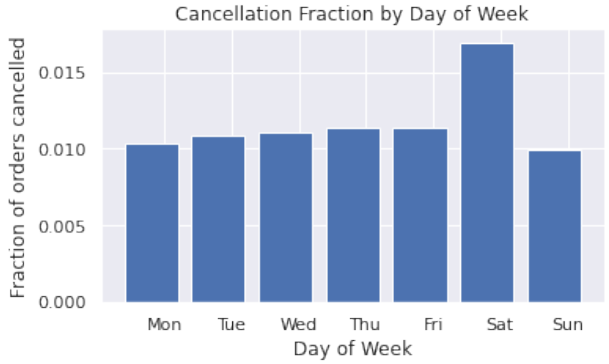


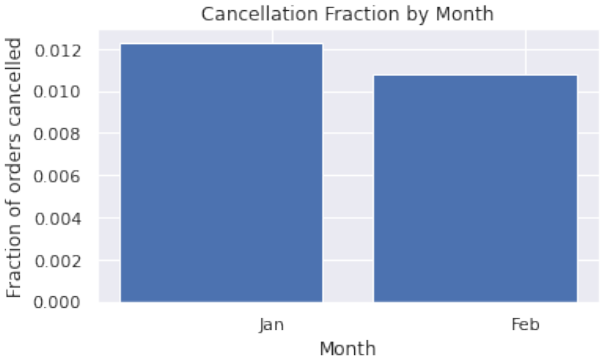

We find that all orders in the data are from months of Jan and Feb only. The reassignment features have a high fraction of NaN values and of the non-Null values, **reassigned_order** is always 1. Thus there's no record of orders not being reassigned - it's just Null. Among records where **reassignment_method** is known,

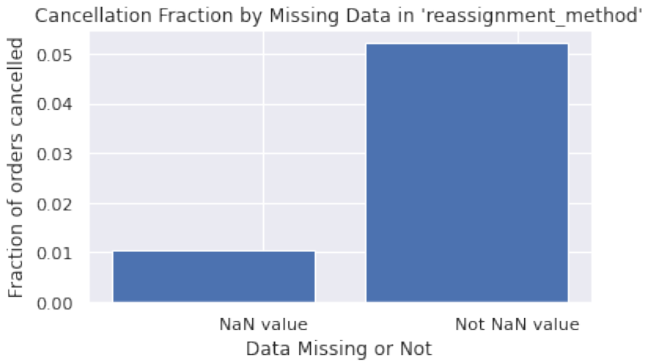
the **auto** mode is much more frequent than **manual**, indicating effective use of technology.

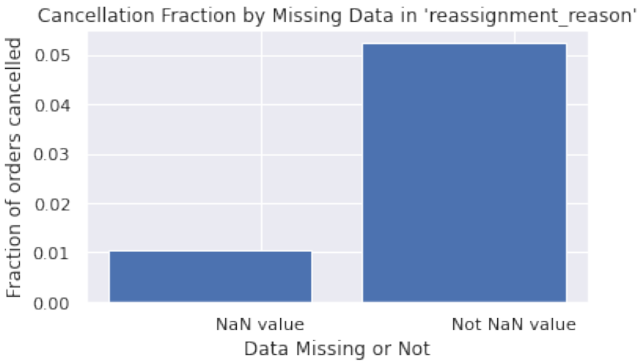
### 3.2 Order Cancellation Rate - by Category

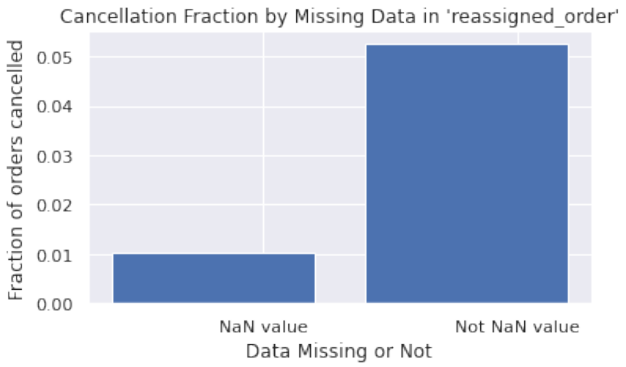
Here, we have a look at the fraction of orders cancelled per category of each categorical feature:

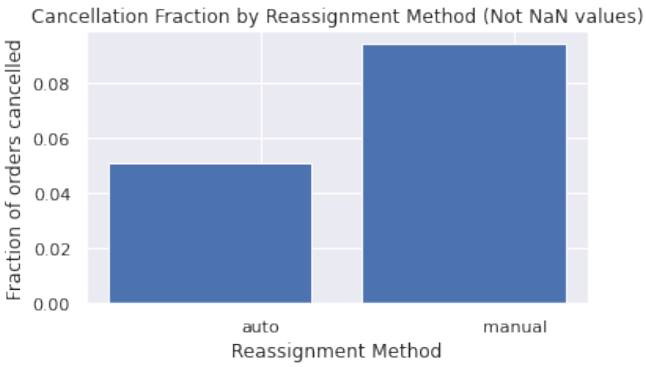


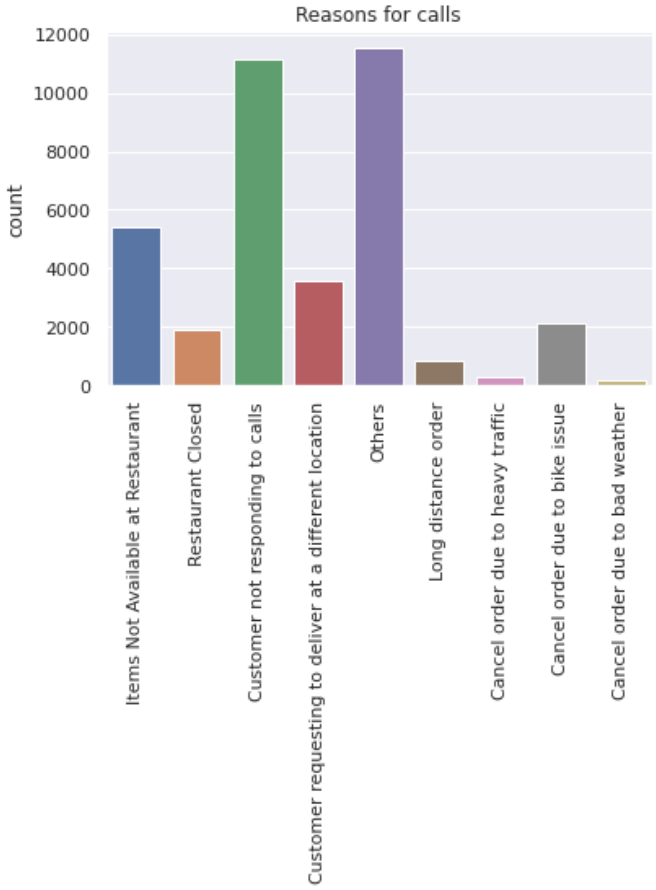




- Order Cancellation rate is slightly higher on Saturdays and slightly lower on Sundays as compared to other days of the week, which are mostly uniform.
- January had a marginally higher cancellation rate compared to February.

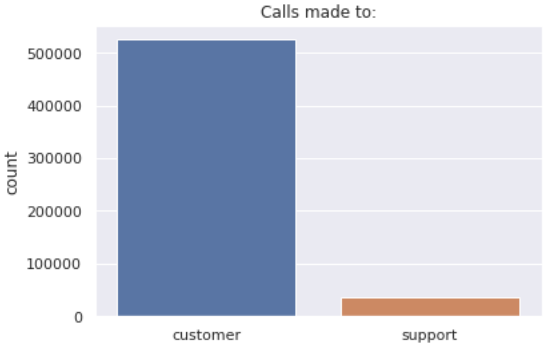
- It can be said that order cancellation is much higher when reassignment method is manual instead of auto.

## 4 Analysis with call_data.csv

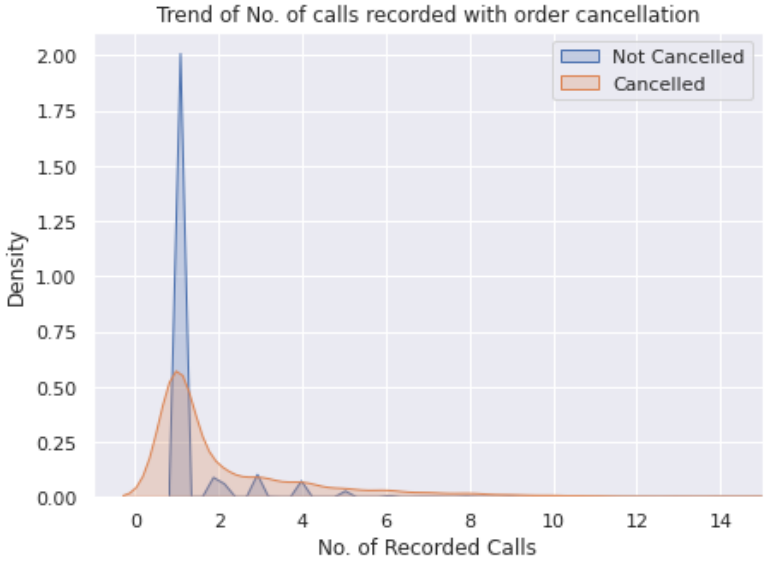
Lets have a look at the frequency of different reasons given for the recorded calls:




Clearly, Customer not responding to calls is a major reason among the ones known. Now lets look at who these calls were made to and how often:



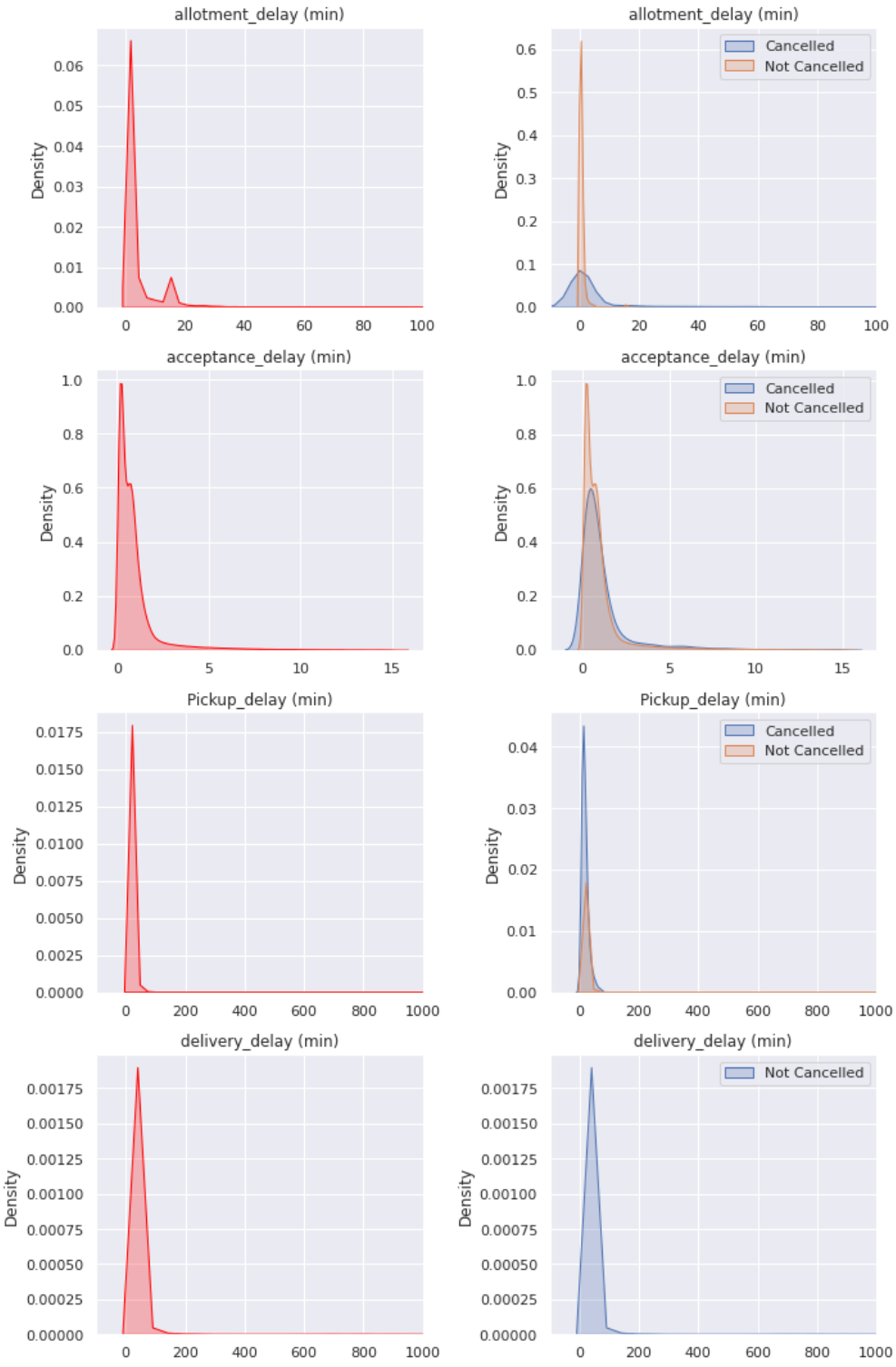
Further analysis shows that all calls whose reasons were recorded in data were made to **support** ONLY. Yet the total calls made to customers are higher and their reasons are not known.
Following is the trend of total number of calls against each target variable.




## 5 Analysis with Datetime Features

We use the provided timestamps to calculate the delay at each step and study its impact on the cancellation of orders. The order is as follows:
**order_time** $\xrightarrow[\text{delay}]{\text{allotment}}$ **allot_time** $\xrightarrow[\text{delay}]{\text{acceptance}}$ **accept_time** $\xrightarrow[\text{delay}]{\text{Pickup}}$ **pickup_time** $\xrightarrow[\text{delay}]{\text{delivery}}$ **delivered_time**

The graph below shows KDE plots of the above delays with the whole data and also by class label on the side. **allotment_delay** shows a sharp peak when orders are not cancelled and a wider curve when they are cancelled - indicating a very fine precise range of delay where orders weren't cancelled. There's barely any occurrence of allotment delay above 15-20 min. Similarly acceptance delay doesn't exceed 10 min while pickup and delivery delays have ranges on slightly higher side. Range of acceptance and pickup delays are nearly identical in both cancellation of orders and otherwise. Obviously, the delivery delay doesn't show cancelled orders as they were all delivered eventually.


allotment_delay (min)
Density
allotment_delay (min)
Cancelled
Not Cancelled
Density
acceptance_delay (min)
Density
acceptance_delay (min)
Cancelled
Not Cancelled
Density
Pickup_delay (min)
Density
Pickup_delay (min)
Cancelled
Not Cancelled
Density
delivery_delay (min)
Density
delivery_delay (min)
Not Cancelled
Density