CASCADE CUP 2022
ROUND 3 REPORT

TEAM : BRUTE_FORCE

TEAM MEMBERS

1)KHUSHAL VINOD RATHI

2)SIDDARTH NILOL K S

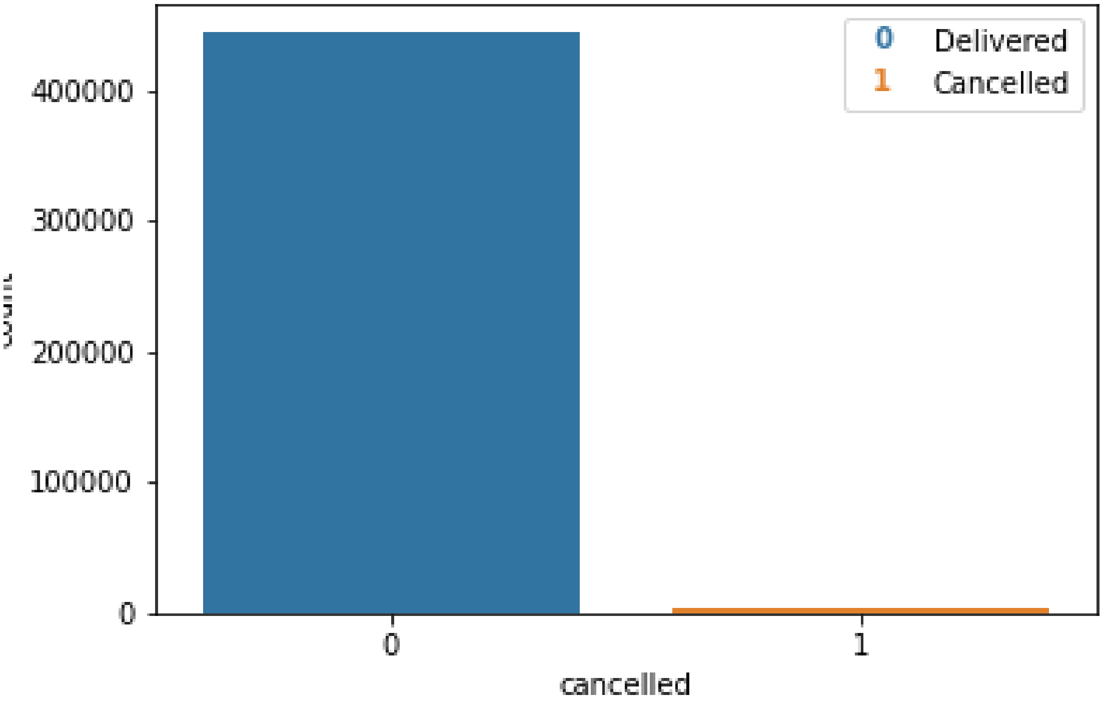
COLLEGE: IIT MADRAS

INTRODUCTION:

The dataset provided is related to a typical order made on **Zomato** and **Swiggy**. Once the order is placed, the timestamps of different actions taken like order time, allot time, accept time, etc., are recorded for a particular order. The rider details are also attached while information regarding whether the allotted rider cancels the order, rider characteristics including lifetime count of orders, delivered orders etc.

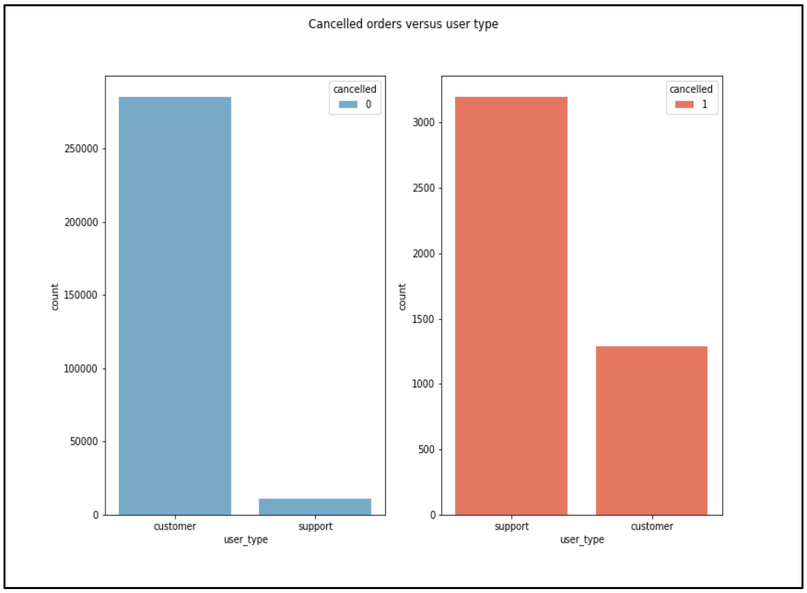
The rider also has the option to get the order cancelled before delivery by calling the support centre. So, the call data is also available.

The on-hand task is to compile a detailed data analysis report based on the dataset.



The graph above shows the number of cancelled and delivered orders in the dataset provided. The number of delivered orders are **444782**, whereas the number of cancelled orders is **5218**. There is a high class imbalance between the two classes. In the upcoming observations, we will study the impact of different features on order cancellation in depth.

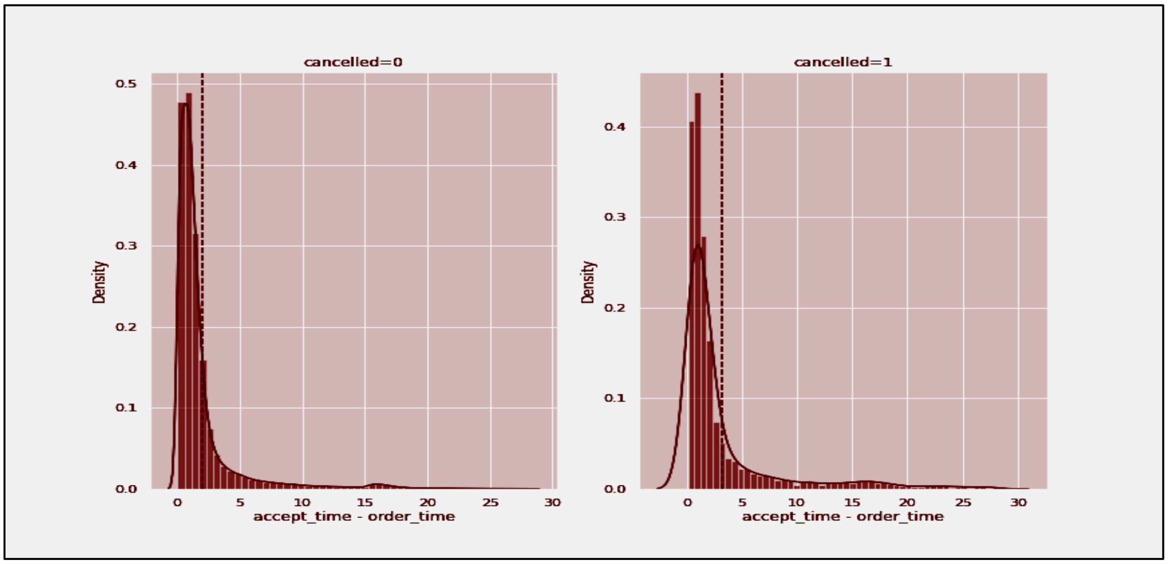
Observation 1:
Does a rider calling a support call centre have a higher chance of order cancellation than calling a customer?



- After merging the call_data with the train_data on the basis of order_id, we get the call data of each order.
- We can observe from the 2 figures that when the orders were not canceled (canceled=0), the riders called the customers most of the time.
- On the contrary, when the orders were canceled (canceled=1), the riders called the support call centre most of the time.

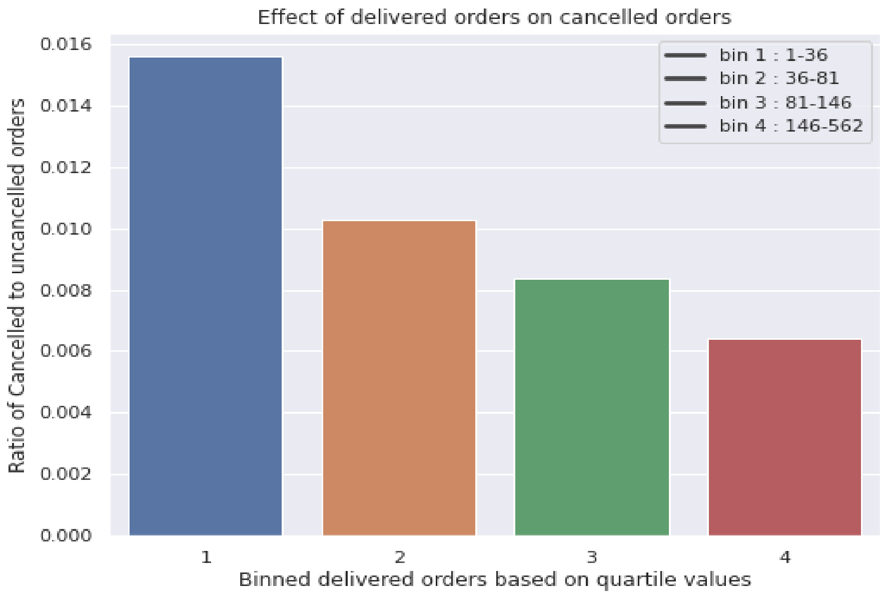
User_type	Cancelled = 0	Cancelled = 1
Customer	285372	1288
Support	11352	3196
Ratio (Support/Customer)	0.04	2.48

Observation 2:
What is the effect of time taken to accept the order by a rider from the time of order on the chances of order cancellation?



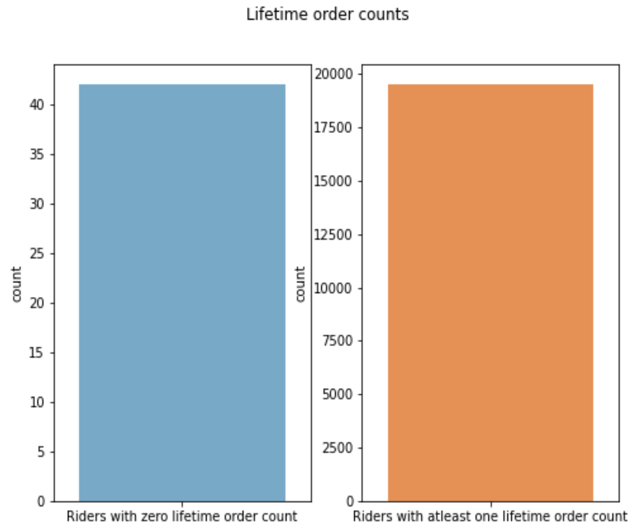
- We found that as the time difference between the accept time and order time increases, the chances of order cancellation increase.
- On doing hypothesis testing of mean comparison with having null hypothesis: (µ1=µ2) and alternative hypothesis: (µ1<µ2), found that **p-value** equal to **0.0374** (less than 0.05). So rejecting null hypothesis confirms that both means are significantly different.
- As we can observe, the mean time difference when orders are not cancelled (left fig) is significantly **less than** when the orders were cancelled (right fig). In conclusion, the time difference and order cancellation are directly related.

Observation 3:
What impact does delivered orders of riders have on new order delivery?



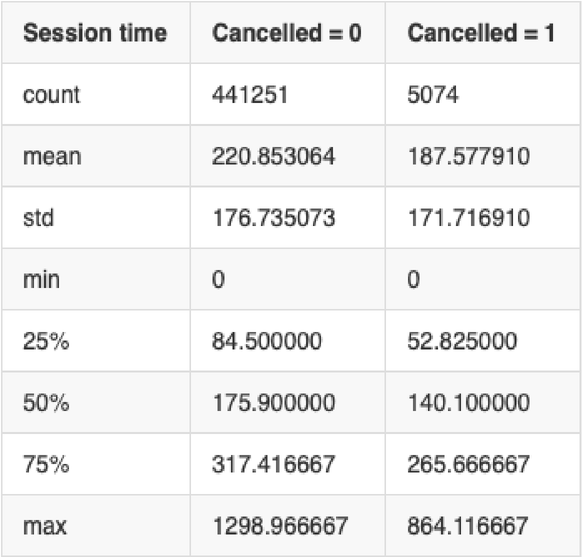
- The delivered orders were binned into **4 categories** based on their quartile values. In different categories, we found the ratio of cancelled to uncancelled orders.
- From the graph, it can be seen that category 1 has the highest cancellation ratio whereas category 4 has the least cancellation ratio. This clearly implies that, if orders delivered by a rider are high, chances of a new order getting canceled are low.
- Thus, delivered orders and cancellations have an **inverse relationship**.

Observation 4:
How many new riders were present in the dataset?



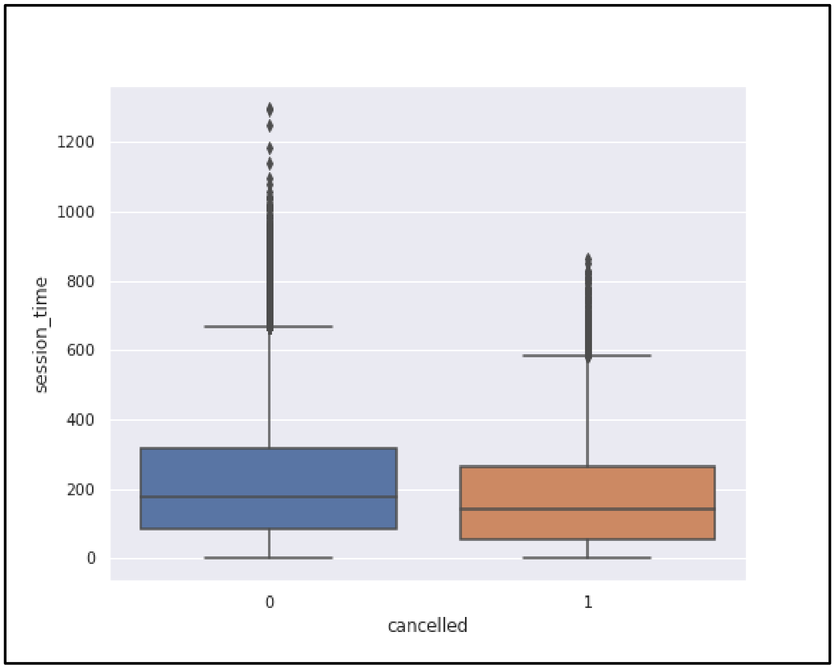
- From the graph, we could observe that there were **42** new riders in the dataset.
- We assumed that if the lifetime_order_count = 'NaN' or '0' value, then the rider has freshly joined as there is no record of previous data.

Observation 5 :
How does session time affect the chances of an order getting cancelled?

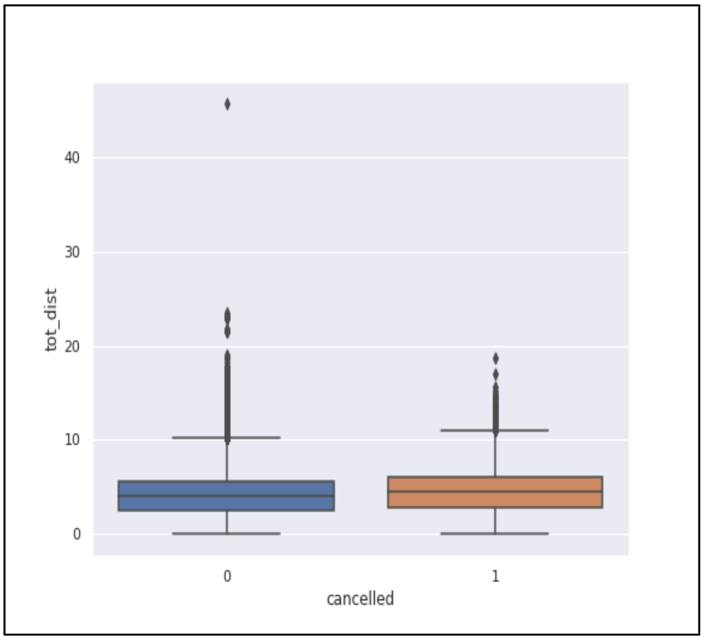


Session time	Cancelled = 0	Cancelled = 1
count	441251	5074
mean	220.853064	187.577910
std	176.735073	171.716910
min	0	0
25%	84.500000	52.825000
50%	175.900000	140.100000
75%	317.416667	265.666667
max	1298.966667	864.116667

- As the session time increases, the probability of an order getting cancelled decreases.
- The mean session time of riders when an order is delivered is **220.853064** seconds (left).
- The mean session time of riders when an order is not delivered is **187.577910** seconds (right).
- From the image and statistics, it is evident that riders with **higher session time** will **most likely** deliver the order.



Observation 6:
Does the total distance a rider has to cover to complete the order delivery affect the order cancellation?

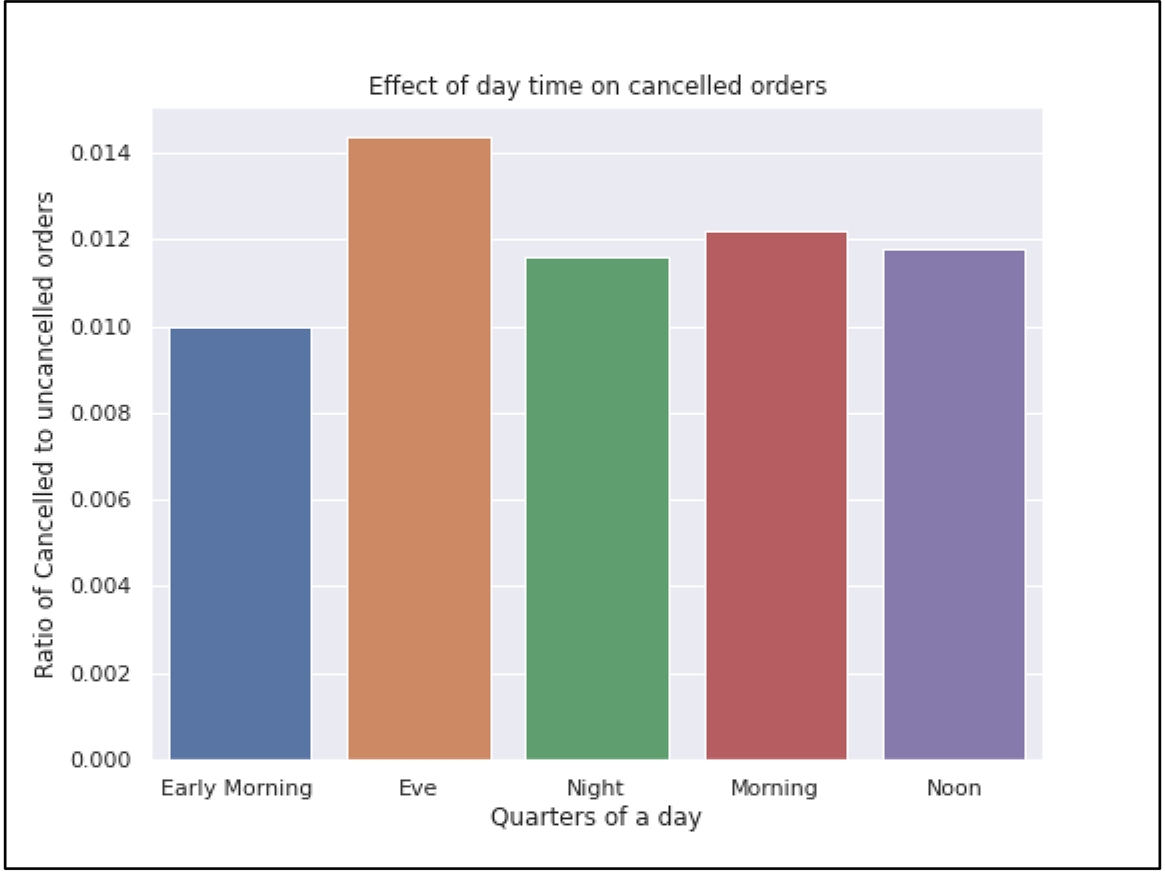


- The first_mile_distance represents the distance between the rider and the restaurant
- The last_mile_difference represents the distance between the restaurant and the customer location.
- The sum of first_mile_distance and last_mile_distance gives the total distance the rider has to travel to complete the order delivery.

We removed outliers to make a clear observation and we found that as the total distance increases the chances of order cancellation increases because the mean total distance when orders were not canceled **(μ1=4.194)** is less than the mean total distance when orders were canceled **(μ2=4.599)**.

Observation 7:

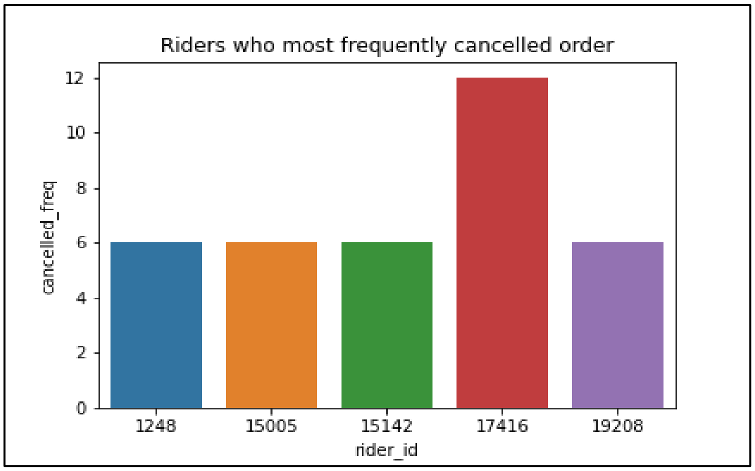
Does the quarter of a day in which order has been placed matter for the rider?

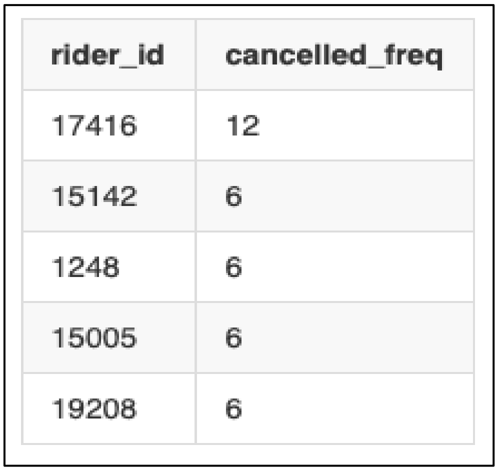


- We have divided the day into **five quarters** depending on the time.
- In the below graph, the y-axis represents the ratio of cancelled to uncancelled orders in a particular quarter of a day.
- From the graph, we can see that the ratio is close to **0.011** in different quarters of the day, and there is no significant difference in ratios throughout the day.
- Therefore, we can conclude that the cancelled feature is **independent** of order time and date.

Observation 8:
Which rider has most frequently cancelled the orders?

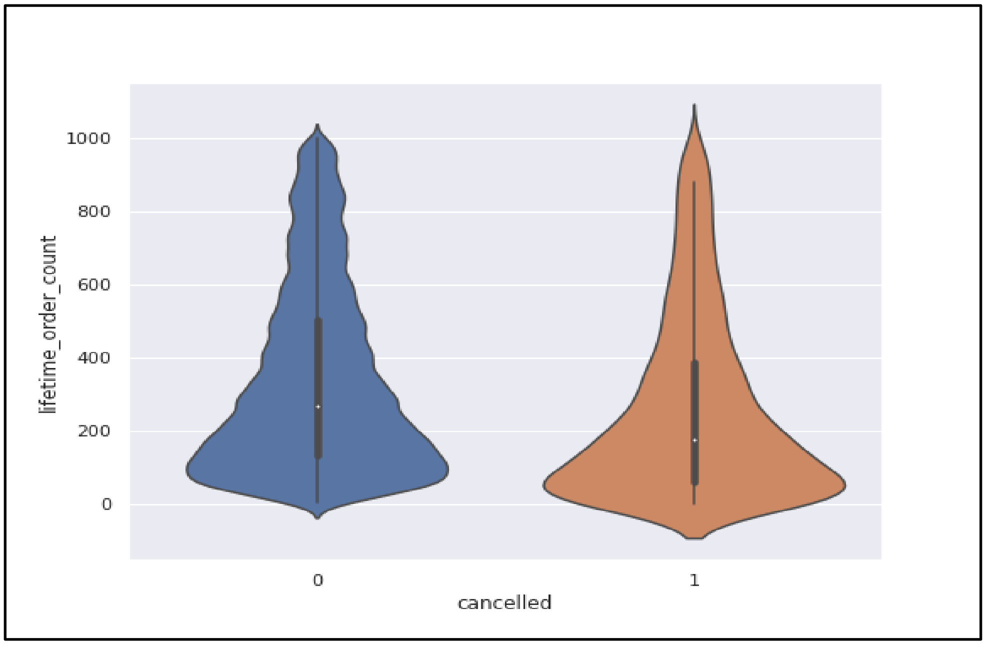
The graph shows that the rider with **ID:17416** has most frequently cancelled the orders **(12 times)**.





rider_id	cancelled_freq
17416	12
15142	6
1248	6
15005	6
19208	6

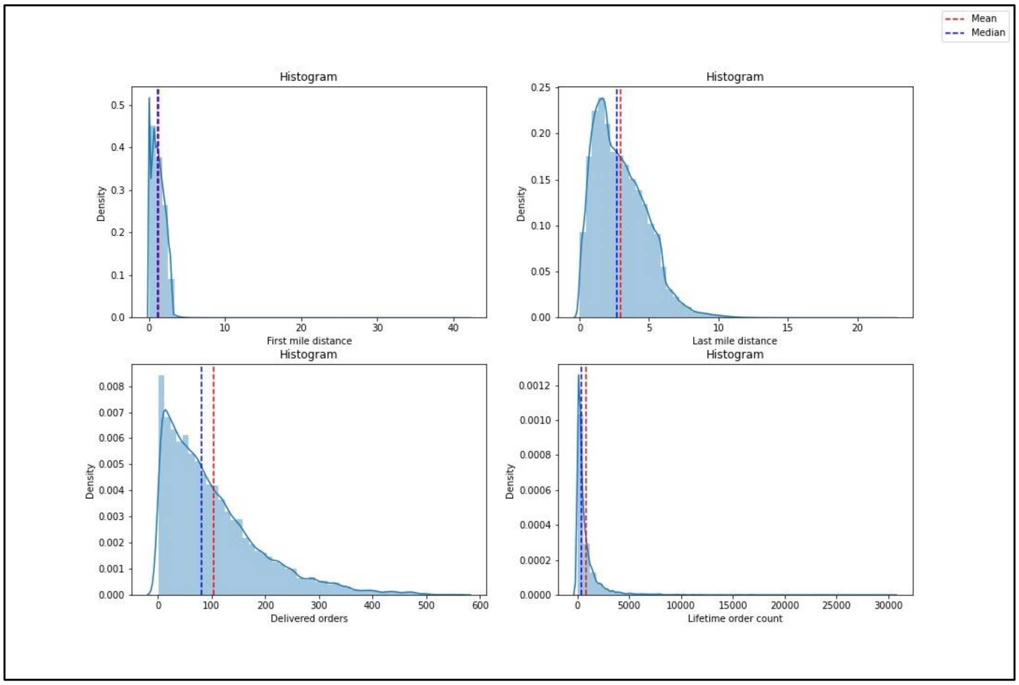
Observation 9:
Does the lifetime order count of a rider affect the chances of the order being cancelled?



- The lifetime order counts of riders were grouped on the basis of cancelled feature.
- The mean lifetime order count of riders when order is delivered is **338.502160** (left).
- The mean lifetime order count of riders when order is cancelled is **259.238816** (right).
- From the violin plot, it is evident that as lifetime order counts increase, the chances of order getting cancelled decrease.

Observation 10:
Are the numerical features in the dataset skewed?

- We observed that the numerical columns like **first_mile_distance, last_mile_distance, lifetime_order_count, and delivered_orders** are highly right-skewed.
- Skewness coeffcients:
  - First_mile_distance: **0.7588891596379476**
  - Last_mile_distance: **0.8267678468439186**
  - Lifetime_order_count: **6.757141606225104**
  - Delivered_orders: **1.3878097334046695**
- As we know that skewness coefficient **greater than 0.5** indicates that the data is highly skewed, it clearly seen that the above features are **highly right-skewed**.



CONCLUSION:

- The data is **highly imbalanced** based on the “cancelled” category.
- The rider calling a support call centre has a **higher chance** of order cancellation than calling a customer.
- The time difference between accept time and order time and order cancellation are **directly related**.
- The rider history of the number of delivered orders and order cancellations have an **inverse relationship**.
- There were **42** new (freshly recruited) riders in the dataset.
- The riders with **higher session time** will most likely deliver the order.
- As the total travelling distance of the rider increases, the chance of order cancellation increases.
- The order cancelled feature is **independent** of order time and date.
- The rider with **ID:17416** has most frequently cancelled the orders.
- As lifetime order counts increase, the chances of order getting cancelled decrease.
- The numerical columns are highly right-skewed.

THANK YOU