Cascade Cup

Rider-Driver Cancellation

Given the order and rider details as described, predicting rider-driven cancellation in advance via analysis and machine learning models can be very useful for Shadowfax.

Section I - Exploratory Data Analysis

We start the analysis by visualising the successfully delivered orders against the cancelled orders on an everyday basis utilising a dual line chart.

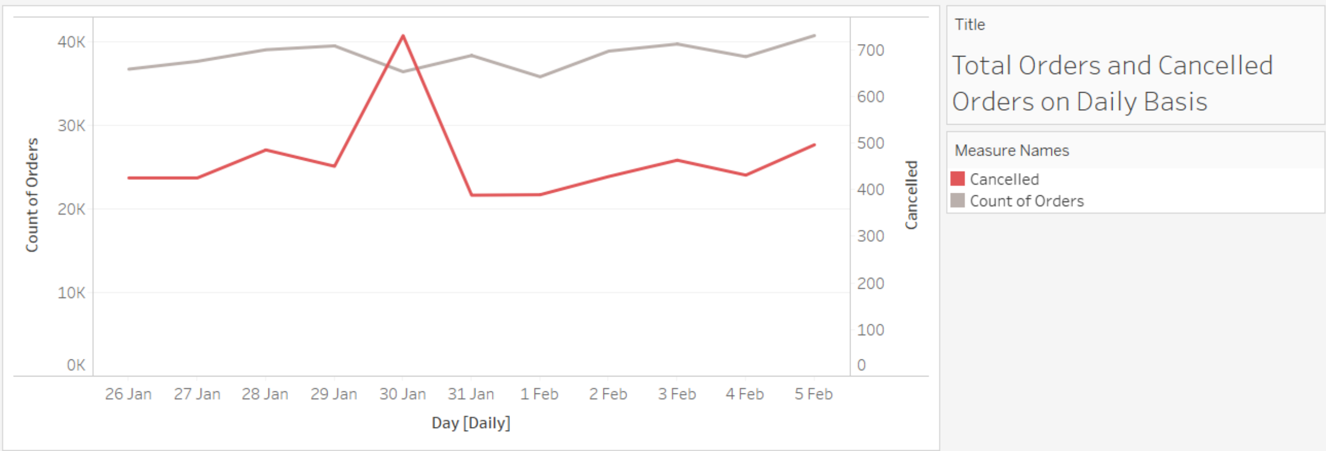


Figure 1. Delivered Orders, Cancelled Orders vs Day [Daily]

From the above graph, we noticed that there is a similar trend between the delivered orders and cancelled orders with the exception of one day - January 30, 2021 which happens to be a Saturday, where we notice a sharp rise in the number of cancelled orders.

Next, instead of using a daily approach for the visualisation, we use an hourly approach for additional inferences.

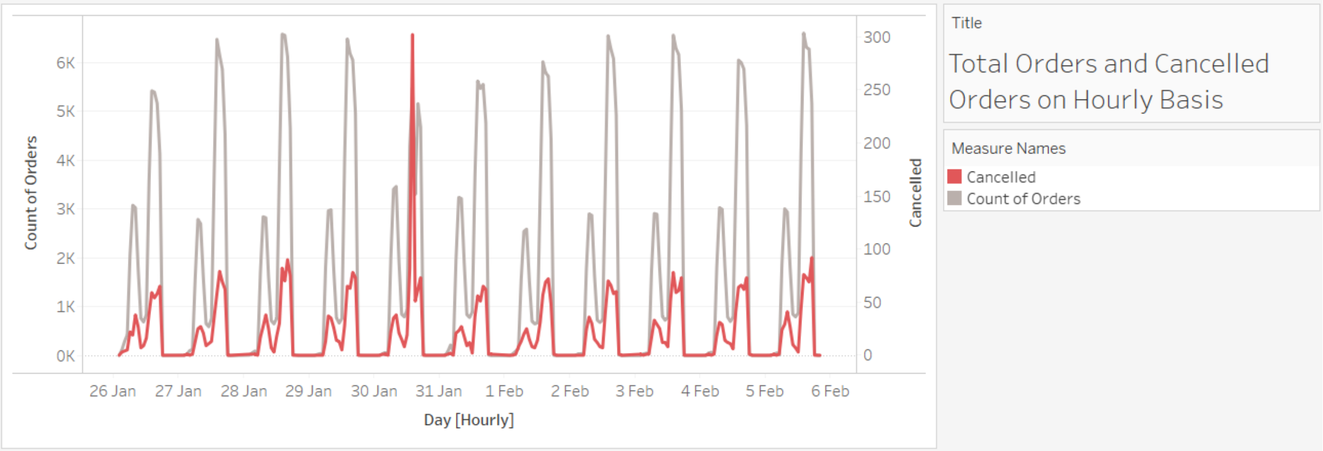


Figure 2. Delivered Orders, Cancelled Orders vs Day [Hourly]

From the above graph, we noticed that there is a common trend observed everyday which consists of a small spike followed by a large spike, which drops to 0 for around 8 hours. During January 30, 2021 (Saturday) we observed a much larger second spike compared to other days. To clearly understand the various sessions involved in this comparison, we use continuous area charts as given below.

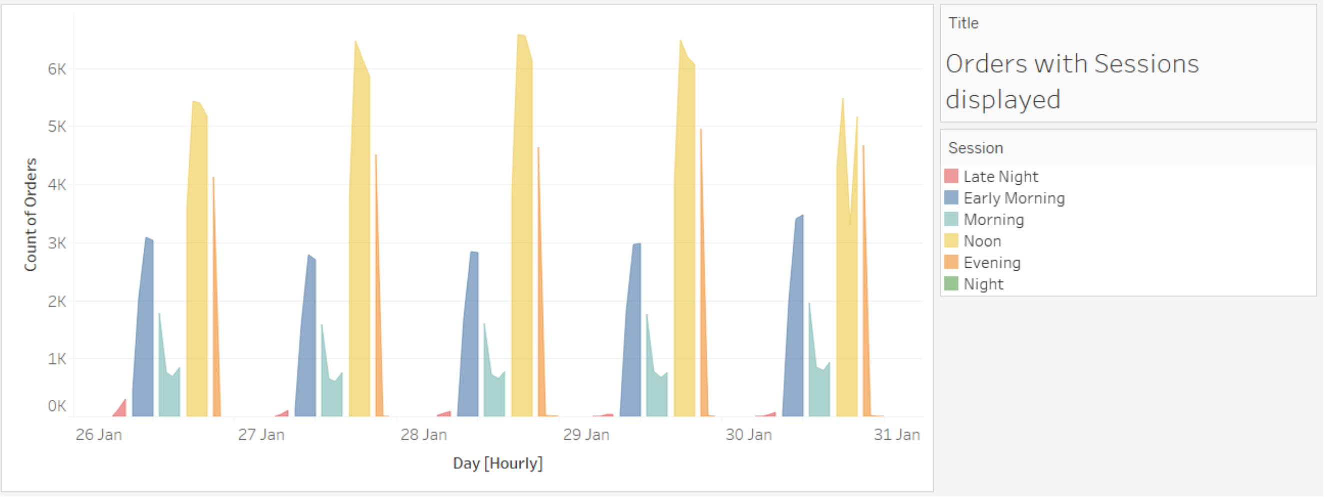


Figure 3. Total Orders vs Day [Hourly] with Sessions

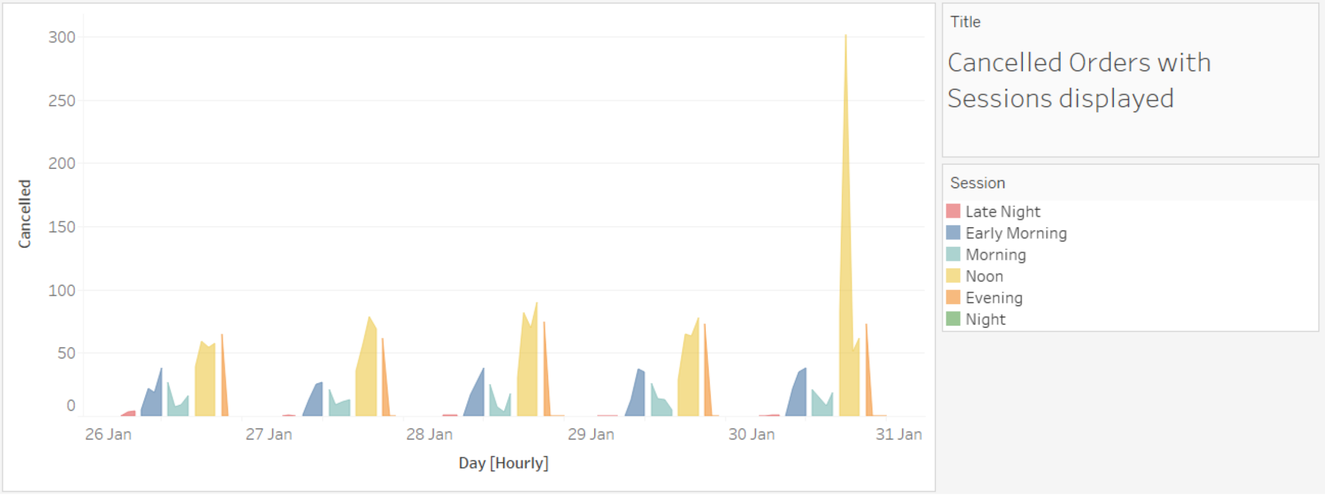


Figure 4. Cancelled Orders vs Day [Hourly] with Sessions

From the above graphs, we can see that most of the orders are placed/cancelled during Noon and the lowest number of orders placed/cancelled is during Late Night. We also notice that there are no orders placed/cancelled during the Night, which signifies that the companies that Shadowfox works with for the given dataset close their services by 1800 hours (6 P.M.) and resume them again by 0200 hours (2 A.M.). So, for the entirety of the report we assume the same working conditions.

Next, we try to visualise the trend based on the days of the week using Area charts. From the graphs given below (Figure 5., and Figure 6.) we can conclude that almost all the days follow a similar trend for orders placed/cancelled and hence are almost indistinguishable - with the exception of Saturday. However, we are unable to confirm this assumption since we do not have the data available for 6 February, 2021 (2nd Saturday in the Dataset).

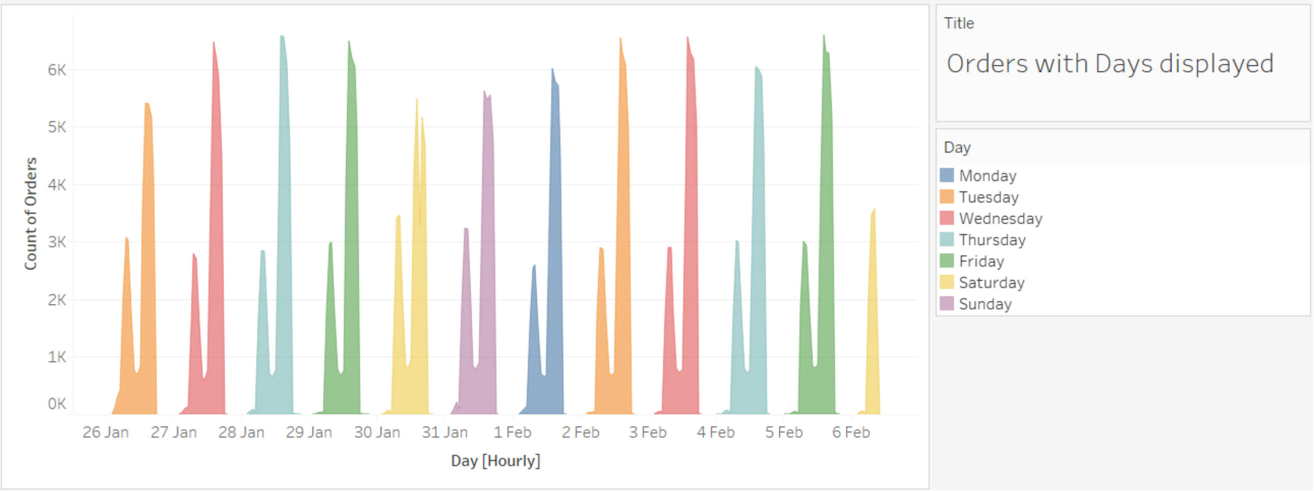


Figure 5. Total Orders vs Day [Hourly] with Days

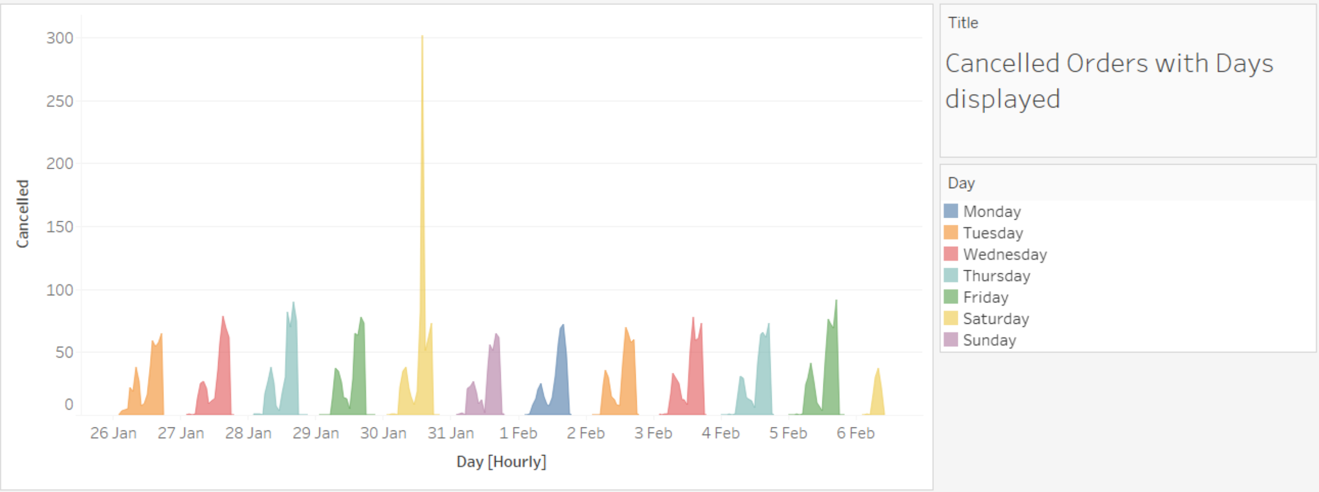


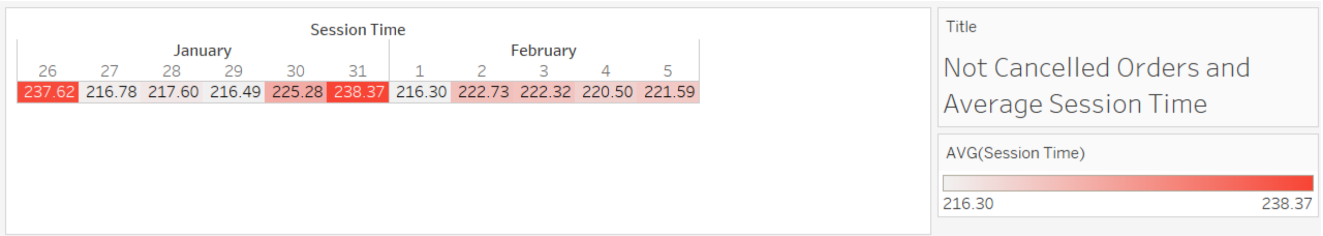
Figure 6. Cancelled Orders vs Day [Hourly] with Days

The key points discussed in the first section are,

- Similar trend between placed orders and cancelled orders (except Saturday)
- Very large number of orders are cancelled during Saturday compared to other days
- During a day, most of the orders are placed/cancelled during Noon
- During a day, least number of orders are placed/cancelled during Late Night
- During a day, no orders are placed during Night
- All days follow a similar pattern for orders placed/cancelled, i.e., Monday is similar to Tuesday, which is also similar to the 2nd Tuesday (except Saturday) - this requires more data to be confirmed

Section II - Specific Attribute Analysis

In this section, we consider the remaining attributes that impact the cancellation of orders, and try to understand their relation. We start by visualising the average session time a user takes when an order gets/does not get cancelled using a highlight table.

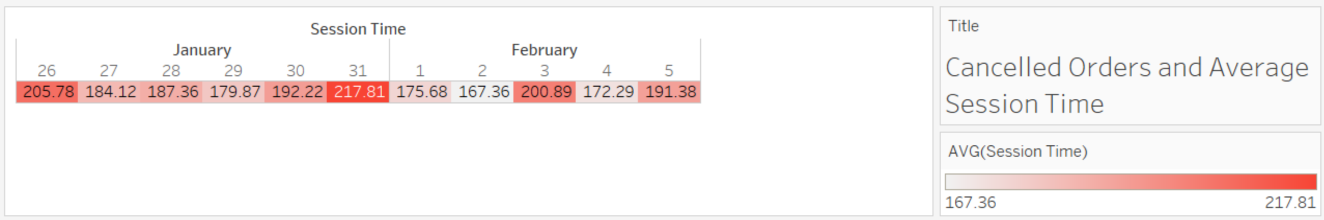


Session Time

January						February				
26	27	28	29	30	31	1	2	3	4	5
237.62	216.78	217.60	216.49	225.28	238.37	216.30	222.73	222.32	220.50	221.59

Title
Not Cancelled Orders and Average Session Time
AVG(Session Time)
216.30
238.37

Figure 7. Not Cancelled and Avg. Session Time



Session Time

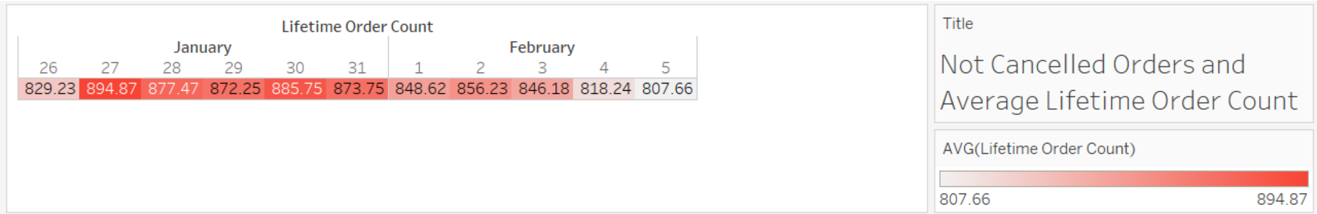
January						February				
26	27	28	29	30	31	1	2	3	4	5
205.78	184.12	187.36	179.87	192.22	217.81	175.68	167.36	200.89	172.29	191.38

Title
Cancelled Orders and Average Session Time
AVG(Session Time)
167.36
217.81

Figure 8. Cancelled and Avg. Session Time

From the above highlight graphs we notice that the average session time is slightly lower for riders who cancel orders. This could be due to them spending less time on the app since they have cancelled their orders, so it could be used as a metric for the modelling. However, the difference observed is not too big for it to be considered a major factor.

Next, we perform the same visualisation with lifetime order count.

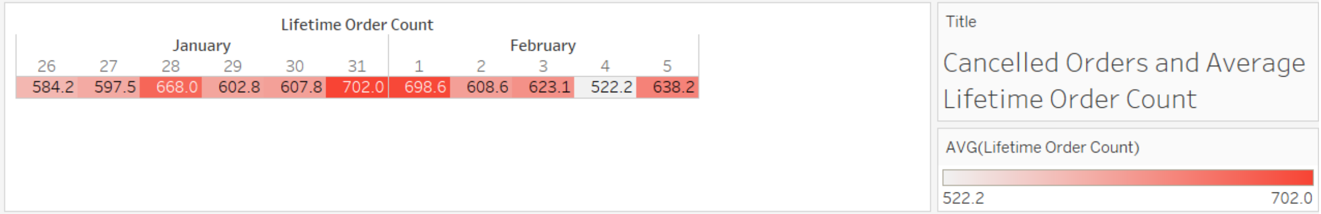


Lifetime Order Count

January						February				
26	27	28	29	30	31	1	2	3	4	5
829.23	894.87	877.47	872.25	885.75	873.75	848.62	856.23	846.18	818.24	807.66

Title
Not Cancelled Orders and Average Lifetime Order Count
AVG(Lifetime Order Count)
807.66
894.87

Figure 9. Not Cancelled and Avg. Lifetime Order Count



Lifetime Order Count

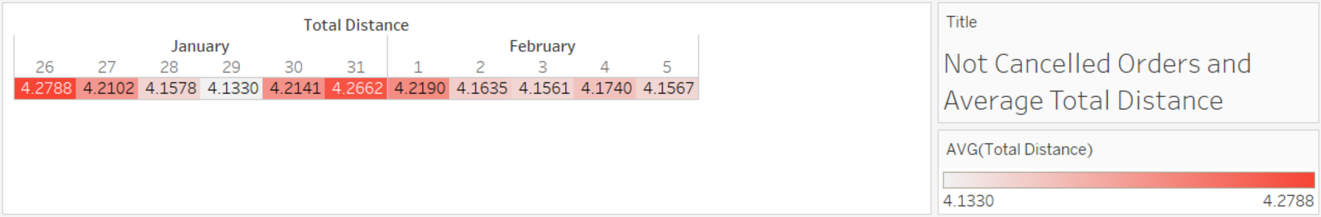
January						February				
26	27	28	29	30	31	1	2	3	4	5
584.2	597.5	668.0	602.8	607.8	702.0	698.6	608.6	623.1	522.2	638.2

Title
Cancelled Orders and Average Lifetime Order Count
AVG(Lifetime Order Count)
522.2
702.0

Figure 10. Cancelled and Avg. Lifetime Order Count

From the above highlight graphs, we observe that the average lifetime order count of riders who cancelled their orders is much lower than that of those who do not frequently cancel orders. This could be due to the fact that a high lifetime order count signifies experience, reliance and hence a lesser chance of cancelling an order. Therefore, this is also considered as an important factor for modelling.

Finally, we perform the same visualisation with total distance (td = first mile distance + last mile distance).



Total Distance

January						February				
26	27	28	29	30	31	1	2	3	4	5
4.2788	4.2102	4.1578	4.1330	4.2141	4.2662	4.2190	4.1635	4.1561	4.1740	4.1567

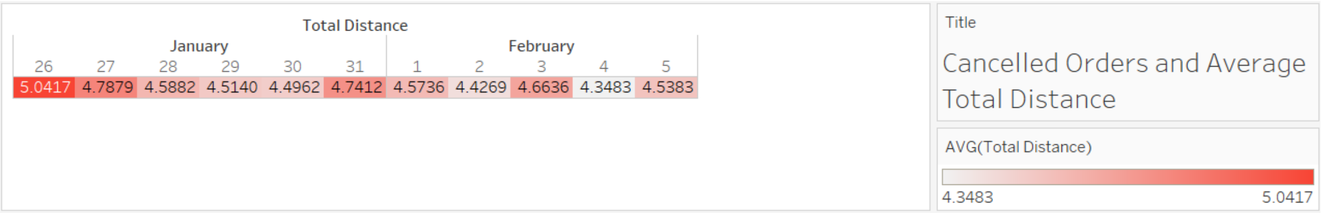
Title

Not Cancelled Orders and Average Total Distance

AVG(Total Distance)

4.1330 4.2788

Figure 11. Not Cancelled and Avg. Total Distance



Total Distance

January						February				
26	27	28	29	30	31	1	2	3	4	5
5.0417	4.7879	4.5882	4.5140	4.4962	4.7412	4.5736	4.4269	4.6636	4.3483	4.5383

Title

Cancelled Orders and Average Total Distance

AVG(Total Distance)

4.3483 5.0417

Figure 12. Cancelled and Avg. Total Distance

From the above highlight graphs, we observe that the average total distance of riders who cancelled their orders is much higher than that of those who do not frequently cancel orders. The total distance could hence be a major reason for which riders cancel their orders, and is discussed in detail in the next section. So, the total distance is also considered as an important factor for modelling.

The key points discussed in the second section are,

- Impact of Session Time, Lifetime Order Count and Total Distance for Cancelled orders
- Nature of impact - positive (or) negative (i.e., smaller lifetime order count signifies higher chance to cancel an order, whereas smaller total distance indicates lower chance to cancel an order)

Hence, for the model built to predict rider cancellation, we combine the general trends observed in Section I with the selected attributes that directly impact cancellation, as given in Section II.

In the final section of the report, we look deeper into the call data and suggest few changes to prevent unnecessary cancellation and methods to reallocate the orders better.

Section III - Call Data Analysis

In this section, we analyse the call data while combining it with the given training data to draw various insights. We start by visualising the reasons given by riders with their total count present in the dataset using horizontal bars.

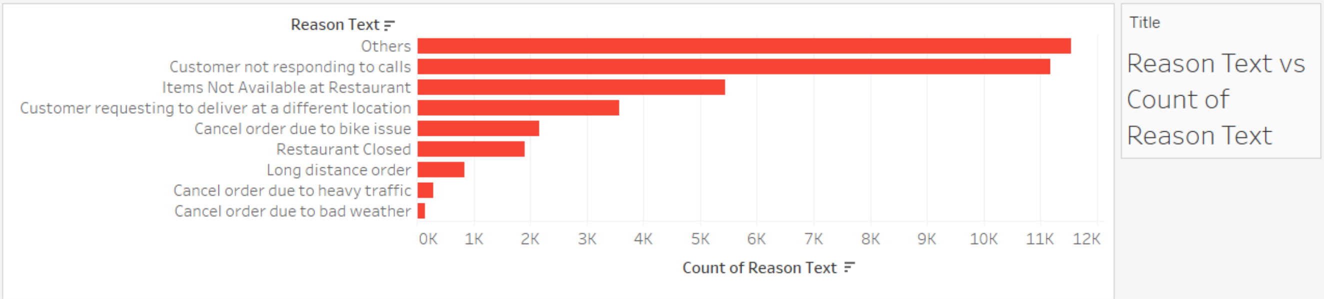
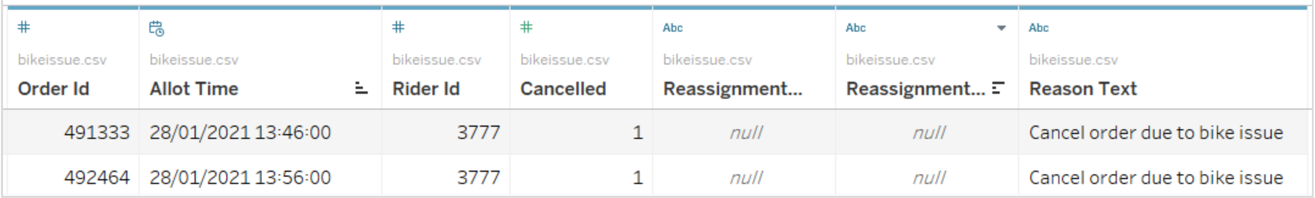


Figure 13. Reason Text vs Count(Reason Text)

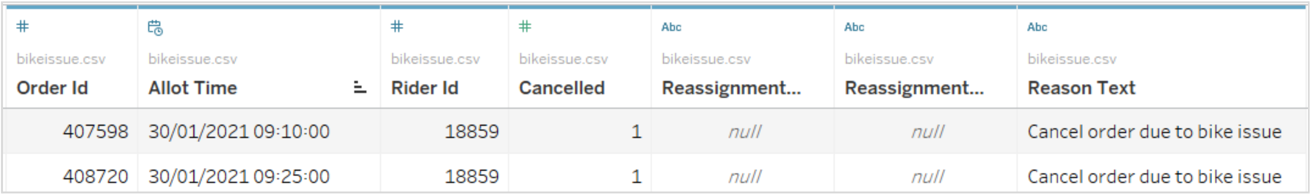
From the above graph, we see that there are 9 reasons a rider gives for cancellation, or can also decide to leave it empty. Out of the 9 reasons, we see that 'Others' is used most frequently compared to the rest. This means that the other 8 default reasons provided are insufficient to categorise the cancellation reasons and hence the 'Others' reason could be looked into deeper to derive few more default reasons for cancellation.

When each reason was analysed individually, the 'Cancel order due to bike issue' stood out among the rest. This was because when a rider selected the 'Cancel order due to bike issue' as the reason for cancellation, at times he would get allotted another order subsequently. Since the bike issue might still exist, the rider ends up cancelling the order again. As a result, the rider is unable to deliver orders and even though he quoted a genuine reason, the system increases his undelivered count. The above claim is seen in an extract from the combination of the train and call data taken below,



bikeissue.csv Order Id	bikeissue.csv Allot Time	bikeissue.csv Rider Id	bikeissue.csv Cancelled	bikeissue.csv Reassignment...	bikeissue.csv Reassignment...	bikeissue.csv Reason Text
491333	28/01/2021 13:46:00	3777	1	null	null	Cancel order due to bike issue
492464	28/01/2021 13:56:00	3777	1	null	null	Cancel order due to bike issue

Figure 14. An extract of details of Rider ID 3777



bikeissue.csv Order Id	bikeissue.csv Allot Time	bikeissue.csv Rider Id	bikeissue.csv Cancelled	bikeissue.csv Reassignment...	bikeissue.csv Reassignment...	bikeissue.csv Reason Text
407598	30/01/2021 09:10:00	18859	1	null	null	Cancel order due to bike issue
408720	30/01/2021 09:25:00	18859	1	null	null	Cancel order due to bike issue

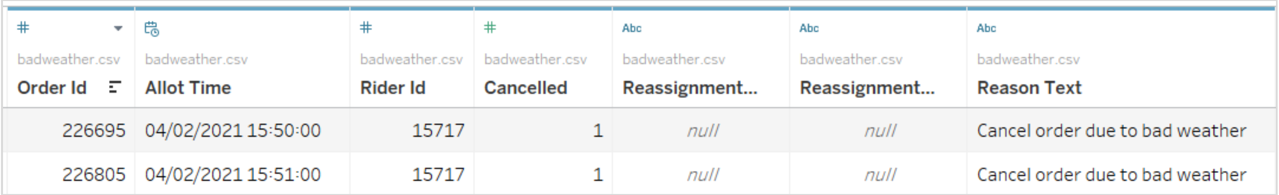
Figure 15. An extract of details of Rider ID 18859

From the above extracts, we see that a rider gets allotted multiple orders even after cancelling one due to bike issues. In Figure 14., Rider ID 3777 has an order allotted to him at 1346 hours (1:46 P.M.) which he cancels due to bike issue. Then, he gets another order allotted to him at 1356 hours (1:46 P.M.) which he has to cancel again.

This trend was observed across multiple riders such as Rider ID 3777, 18859, 18883, 21354, etc. A simple solution to fix this issue would be the implementation of a cooldown feature. If the rider is experiencing a bike issue, he generally has a good idea on how long it would take him to resolve that issue.

So, in the above example (Figure 14.), the rider 3777 cancels the order at 1346 hours due to a bike issue. If he knew that the issue would be resolved by 1700 hours, he could use the cooldown feature to indicate this. So, during this time period - no orders would be allotted to the rider 3777. Hence, he would not have to make unnecessary cancellations and ruin his record.

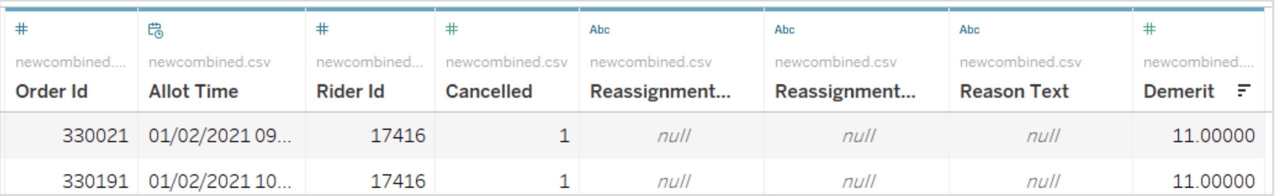
To a very small extent this issue was also observed in 'Cancel order due to bad weather' - only 3 riders were present with this issue compared to the 61 in the previous reason. Hence, the same cooldown feature could also be extended to this reason.



badweather.csv	badweather.csv	badweather.csv	badweather.csv	badweather.csv	badweather.csv	badweather.csv
Order Id	Allot Time	Rider Id	Cancelled	Reassignment...	Reassignment...	Reason Text
226695	04/02/2021 15:50:00	15717	1	null	null	Cancel order due to bad weather
226805	04/02/2021 15:51:00	15717	1	null	null	Cancel order due to bad weather

Figure 16. An extract of details of Rider ID 15717

There are also many riders that cancel orders without providing any reason. To identify such riders, we introduce a new attribute called Demerit that keeps track of cancellations without any reasons. On performing this in the dataset, we are able to filter out those who generally do not give reasons for their cancellations. This is given in the below figure,

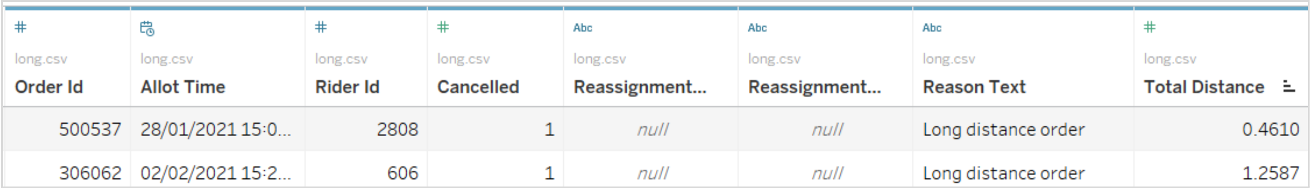


newcombined....	newcombined.csv	newcombined...	newcombined.csv	newcombined.csv	newcombined.csv	newcombined.csv	newcombined....
Order Id	Allot Time	Rider Id	Cancelled	Reassignment...	Reassignment...	Reason Text	Demerit
330021	01/02/2021 09...	17416	1	null	null	null	11.00000
330191	01/02/2021 10...	17416	1	null	null	null	11.00000

Figure 17. An extract of details of Rider ID 17416

From the above extract we identify that rider 17416 has the highest demerits in the given dataset with a staggering total of 11! This means that he cancels orders frequently without providing any reason, and this should be reviewed by his employer.

We can also extend the demerits feature to identify riders who select incorrect reasons for namesake. This is especially visible in the 'Long distance order' reason.

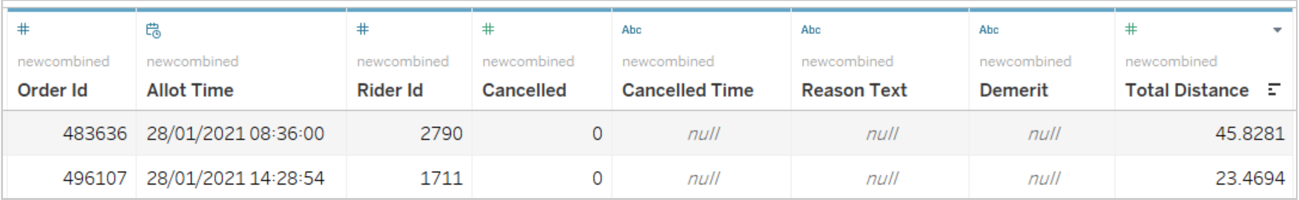


#	📅	#	#	Abc	Abc	Abc	#
long.csv	long.csv	long.csv	long.csv	long.csv	long.csv	long.csv	long.csv
Order Id	Allot Time	Rider Id	Cancelled	Reassignment...	Reassignment...	Reason Text	Total Distance
500537	28/01/2021 15:0...	2808	1	*null*	*null*	Long distance order	0.4610
306062	02/02/2021 15:2...	606	1	*null*	*null*	Long distance order	1.2587

Figure 18. An extract of details of Rider ID 2808 and Rider ID 606

In the above example we see that even though the total distance is 0.46 miles, 1.25 miles, the riders report it as a 'Long distance order' and cancel the order. This is a false claim especially when we consider that the mean total distance of the entire dataset is 4.19 miles. So, we identify such false reasons and change their Reason Text to Null. As a result, for every Null entry, their demerits increase. In this way, we can identify poorly performing riders.

There also exist riders that cover incredibly long distances to deliver their allotted orders. Few of them are given below,



#	📅	#	#	Abc	Abc	Abc	#
newcombined	newcombined	newcombined	newcombined	newcombined	newcombined	newcombined	newcombined
Order Id	Allot Time	Rider Id	Cancelled	Cancelled Time	Reason Text	Demerit	Total Distance
483636	28/01/2021 08:36:00	2790	0	*null*	*null*	*null*	45.8281
496107	28/01/2021 14:28:54	1711	0	*null*	*null*	*null*	23.4694

Figure 19. An extract of details of Rider ID 2790 and Rider ID 1711

So, we can create a new attribute that keeps track of the average distances a rider covers. The next time an order has to be reassigned due to long distance, it can automatically be reassigned to those riders who have a higher average distance.

The key points discussed in the third section are,

- Need to create more default reasons for cancellation by generalising the already existing ones present in 'Others'
- Add a cooldown feature for riders who cancel orders due to a bike issue
- Extend this feature for smaller reasons such as bad weather, heavy traffic that could be used in the future
- Add a demerits attribute to mark riders who frequently cancel orders without providing any reason
- Extend this feature to detect false reasons provided for cancellation, such 'Long distance order' when the actual distance is under a mile
- Add an average distance attribute, to keep track of riders who prefer covering long distances
- Automatically reallocate orders to riders with higher average distance when the reason given is 'Long distance order'

With this, we conclude the analysis of the given rider-cancellation prediction problem, with a few suggestions to reduce the number of cancellations in the future.

Thank You

Tools Used,
Python 3.10.1
Microsoft Excel
Weka 3.8.5
Tableau Public

Find the documentation at,
Google Colab

Guna M.
mg2kofl@gmail.com

Sairam Vaidya M.
sairamvaidya2661@gmail.com