Data Story

Cab booking cancellation is an upcoming problem that often leads to business loss and customer dissatisfaction. If this continues it can lead to bigger problems including business competitions. This data set will look into different variables that cause cab booking cancellations including different package types. The clients will be towards cab companies to see the trends of cab bookings and the cases in which cancelations occur. Based on the analysis of the data the cab companies will be able to see data regarding cases of cancellations. It is also geared towards customers to show them which times are the best to book a cab.

The data set contains multiple fields that are helpful for the user to understand. The common primary fields are booking ID which is unique to each customer and the user id which is based on mobile number. The vehicle model id is used to differentiate between every vehicle to see if there is correlation which the type of vehicle causing booking cancellations. The package id focuses on the type of package which depends on the distance that will be traveled. A value of 1 - 5 is a categorical value that represents the distance which ranges from 40 km to 120 km. The travel type id represents the type of travel in which 1 - 3 are categorical values that relate to the distance. The from area id and to area id represents the location to which the passengers start and go to this is specified for point-to-point travel. Another important field is booking type that represents whether a booking was done on desktop or mobile which can be used to determine if there is a correlation between the number of cab

cancellations. Another field that could be of use the latitude and longitude of the from address and to address to show if there is trend as to which locations tend to cause cab cancellations.

Some fields that should probably be included is the difference in time from which a cab booking is made to the time in which the cab comes. This difference in time can be used to show if there is an increase in cancellation for shorter times. My predictions are for shorter differences the increase is greater as there is an increased chance of not having enough cabs and a lack of planning for backups. Another field that could be included is the number of times a customer has booked from the agency. This field could be used to see if there is difference for a new customer compared to a long term customer. I would predict that for a new customer that chances a cab being canceled is higher as an agency would prefer to service their older customers. A final field that would be helpful is the delay time that maps the time a customer should be picked to when they were actually picked. This would be helpful to show that different delays times have an impact towards the cancellation rate. Realistically a longer delay time would lead towards a higher cancellation rate.

For this data there will be cleaning in terms of modifying the NULL values to NA to symbolize data that is not available. There will also be work to add new columns that relate to the existing columns this will take a while to understand and create. The hardest part of this assignment will be to apply the skills learnt.

One exploration I learned is that if the car cancelation is 1 the cost of error is 100. The vehicle id of 12 and 28 don’t have data for the package id. In general there are a lot of NULL values that need to be explored in depth to make sure there isn’t any missing details. This will take the longest to understand as the data and fields are long.

Based on the finding I am going to work the most on cleaning the data as this will be the hardest part. Afterwards I will focus on grouping the data based on the different fields in my case the booking id type and whether a booking is done online or through a mobile device.