UBER SUPPLY-DEMAND GAP

VISUALISATION AND ANALYSIS

PROBLEM STATEMENT

We're considering trips made on Uber to-and-from the Airport for five days. Our goals is to understand whether there is a supply-demand gap, try to find out what factors influence the supply-demand gap, if it is present. Then, find quantitatively the extent of such a gap, if it exists.

Our assumptions include the following:

- Supply in the context of Uber is a completed trip. If a trip is cancelled or if no cars were available, the supply is non-existent for that trip.
- For our summarizations such as Mean and Median, we'll only consider trips that have both a request and a drop time for our quantitative analysis. In other words, NA in any one of those makes a trip irrelevant because it wasn't completed.

DATA CLEANING

The data provided had various consistency issues and discrepancies. Some of these are enumerated below:

- The date and time were inconsistent, and also aren't of standard date-time format.
- There are NA values which were treated as a part of our assumptions.
- There were no duplicates found in the dataset.

DERIVED METRICS AND VARIABLES

We have created some derived metrics and variables. These are enumerated below:

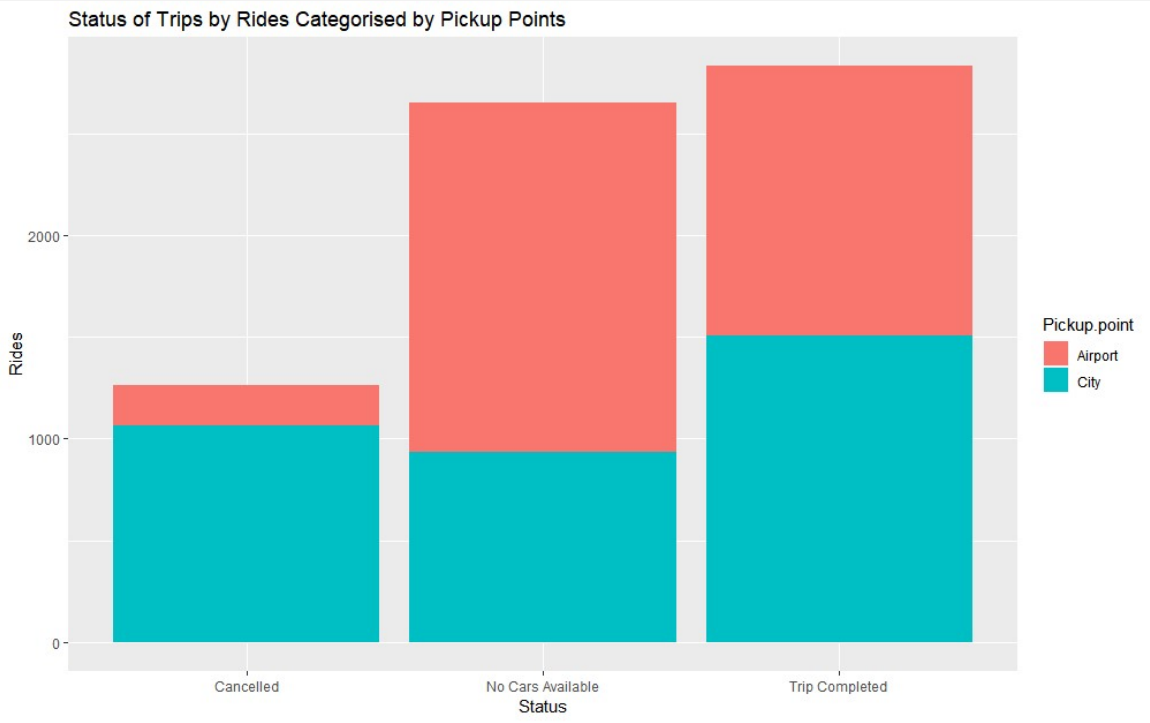
- **Status.bool** – This variable is a simpler factorized version of the Status column. This tells whether a trip was completed (1) or not (0). This would help us understand and visualize better when we aren't interested in the reason for why the trip didn't happen.
- **Trip.duration** – This is the trip duration for each trip which is simply a subtraction of the drop and request times.
- **Day.of.week** – This tells us which day of the week it is so that we could test our hypothesis that the days of the week impact the supply and demand.
- **Driver.gap** – This variable calculates each driver's gap between rides. However, this couldn't provide us with a lot of information as there are times where the gaps are over six hours which could mean the driver has just changed shifts.
- **is.Peak** – This variable tells us whether the ride was made during a Peak/Rush hour or not. The peak and rush hours were calculated on the basis of a time plot further in the presentation.

STATUS BY RIDES

The following plot shows us a histogram categorized by Pickup Points.

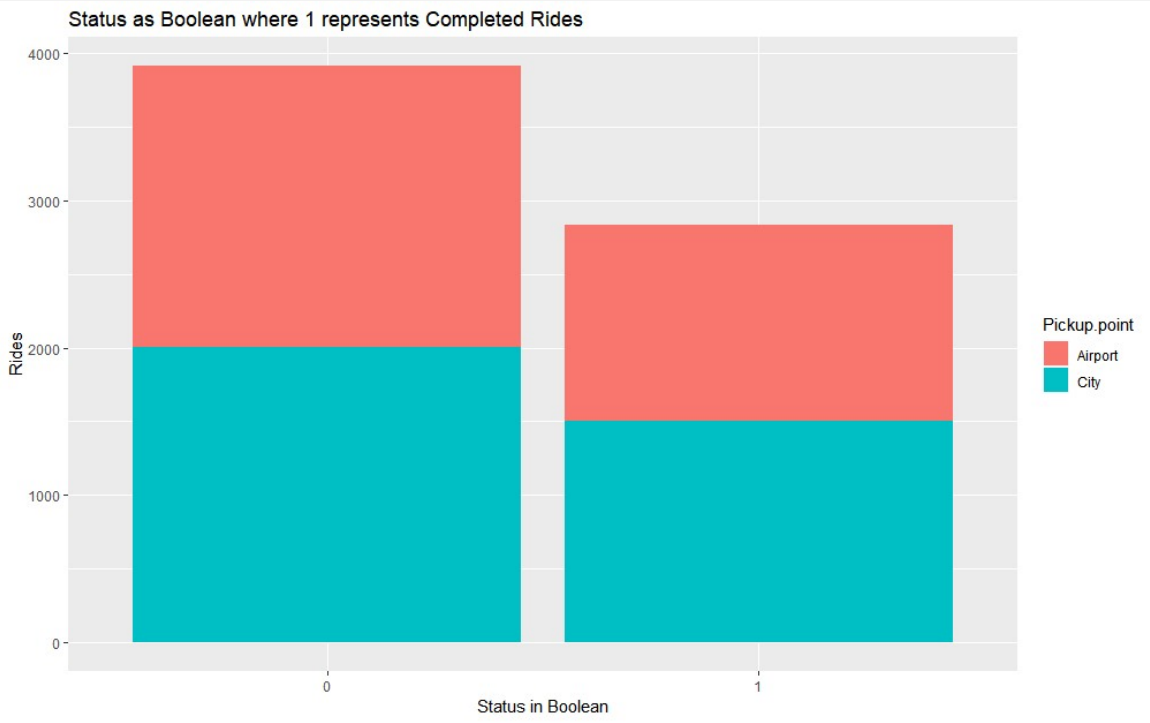
We can clearly see that the rides with No Cars Available is comparable to Trips Completed.

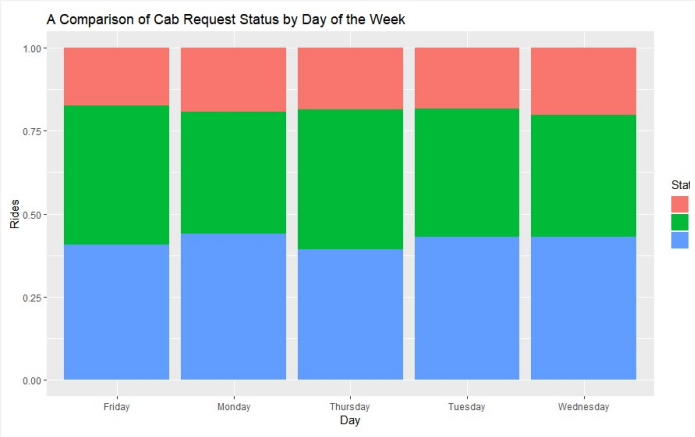
So, we shall create another plot with a Boolean Factor of Trip Completed or Not instead of three Status variables.

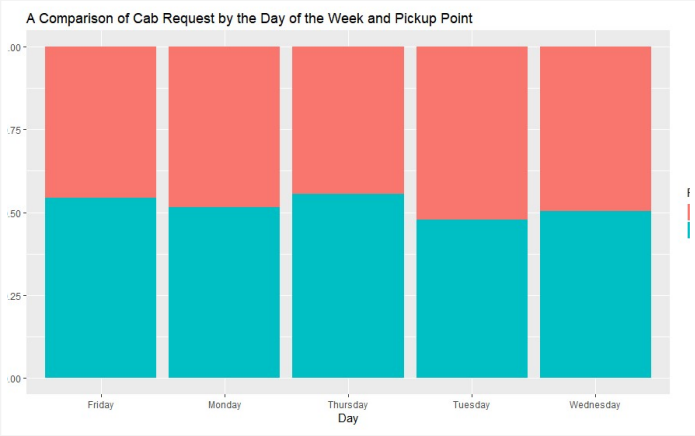


STATUS BY RIDES

This plot shows with clarity that the number of rides being demanded are greater than the number of rides being supplied as more rides are incomplete (Cancelled/Unavailable) than the rides which are complete.







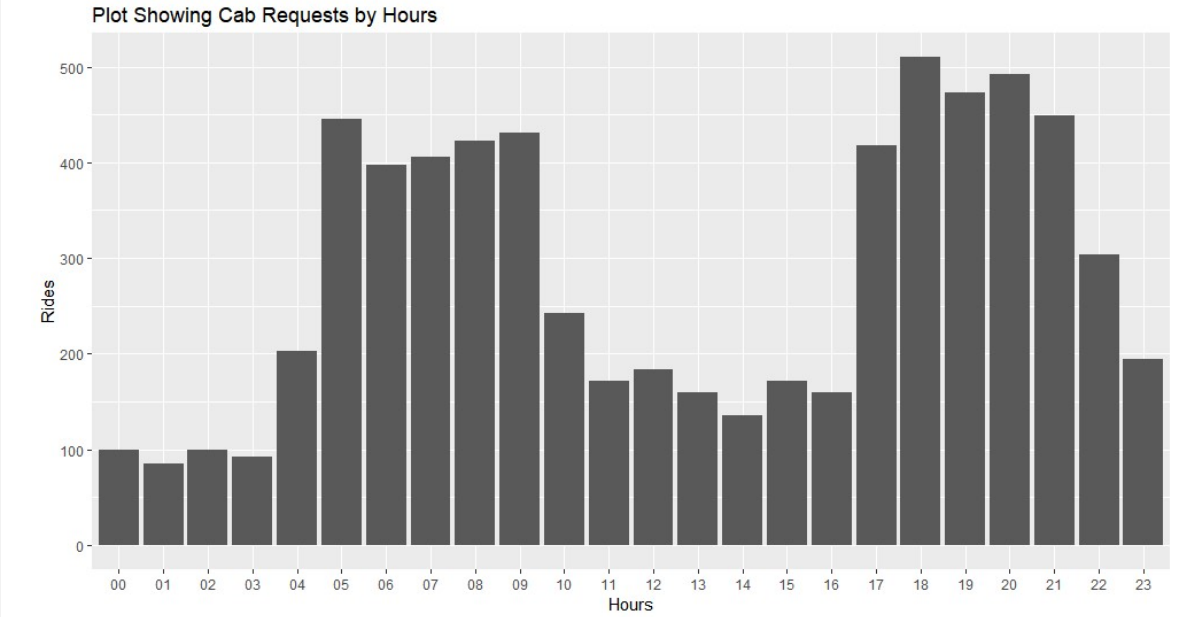
RIDES BY THE DAYS OF THE WEEK

An initial hypothesis was that rides will increase on weekends. However, upon plotting the graphs for the days of the week, two inferences were clear.

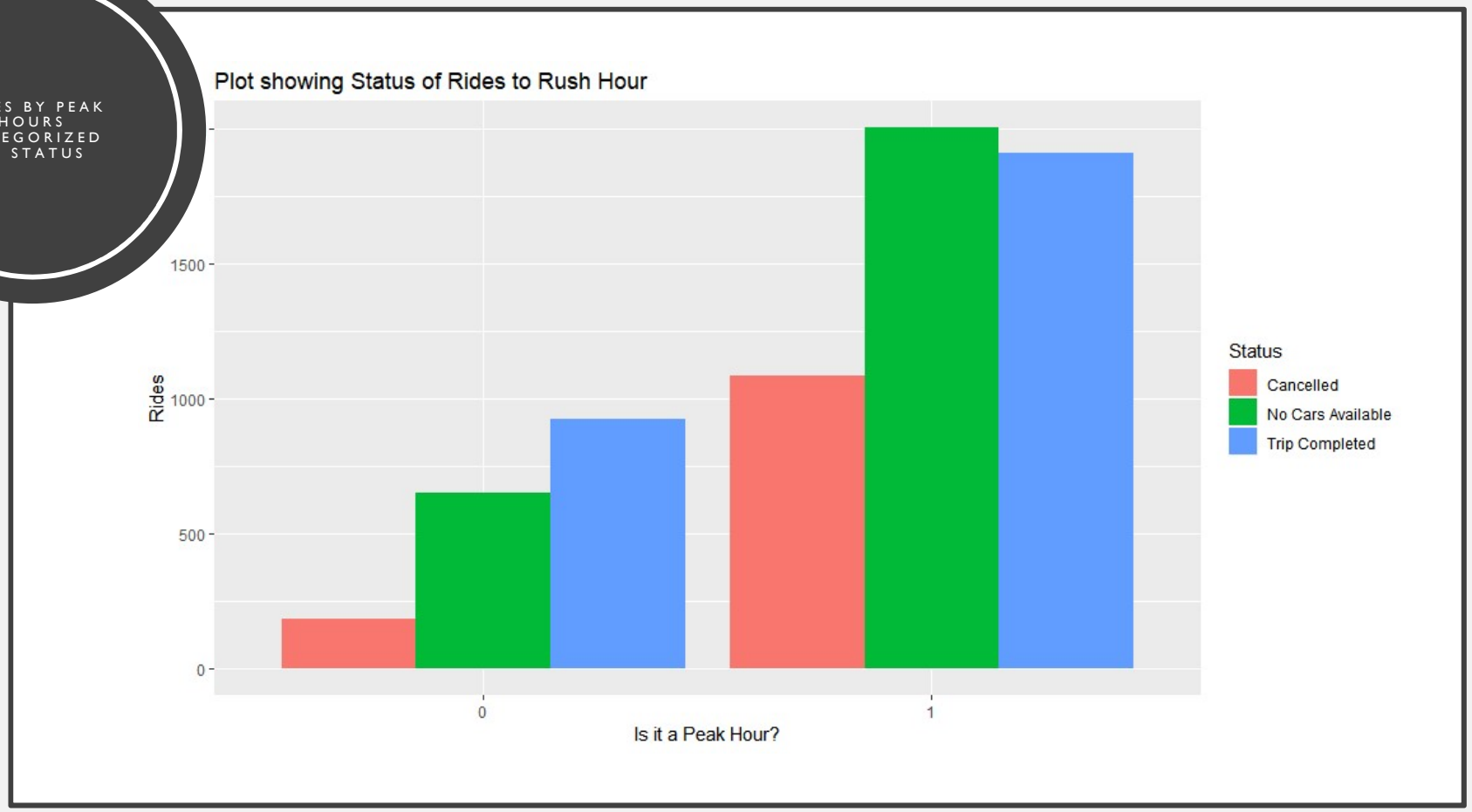
- The data did not include Saturdays and Sundays. So, our hypothesis could not be tested.
- There wasn't much of a difference between the Status of the rides when viewed opposed to which day of the week it was. Therefore, we can say that the Day of the Week **does not** impact the supply-demand gap which clearly exists based on the previous plots.

FINDING PEAK HOURS

- It is rather clear that the slots from 5AM to 10AM and 5PM to 10PM can be considered as peak hours given our data set.
- We'll use this understanding to analyze the data further.

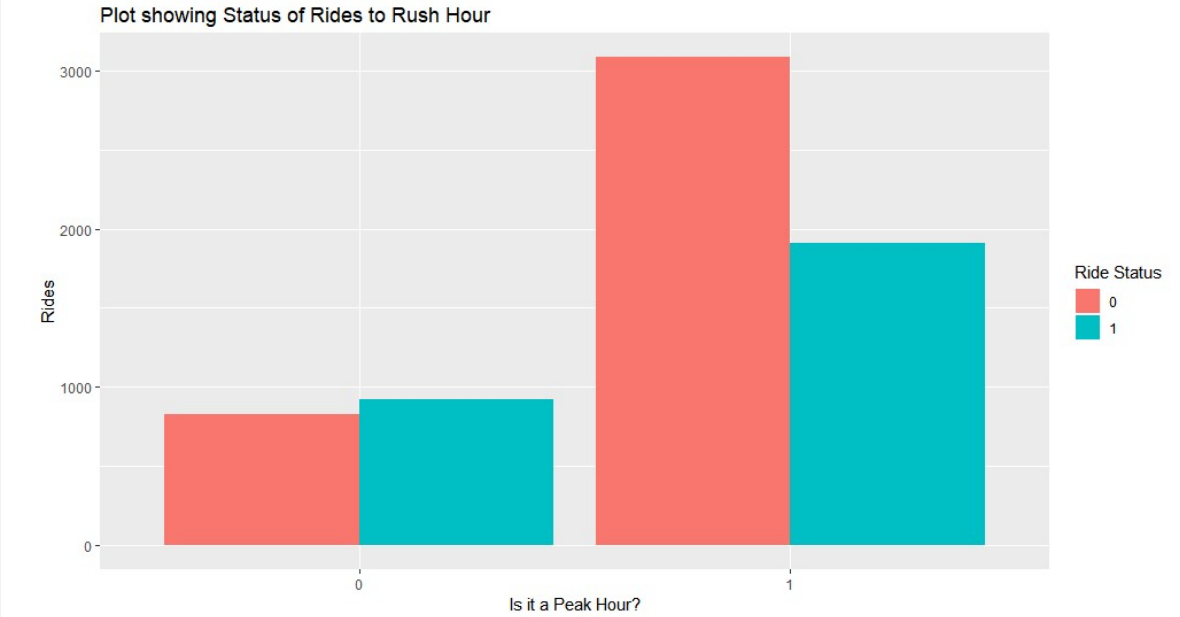


RIDES BY PEAK HOURS CATEGORIZED BY STATUS



RIDES BY PEAK HOURS CATEGORIZED BY SIMPLIFIED STATUS

- Following our previous approach, we took the simplified Boolean Factor of Status derived by us to understand that there were more rides being Cancelled or Not Available during Peak Hours than the rides being Completed.
- We can also clearly see that the number of rides Completed exceeds the number of rides Incomplete if the hours aren't peak.



QUANTITATIVE ANALYSIS

- We're considering the demand to be the sum of the rides requested which is the size of our dataset.
- Supply is the number of rides completed based on our assumptions.
- Lag is the number of rides which aren't completed.
- We can see that the lag increases further during peak hours.
- Also, there is a huge gap in demand and supply.

	Overall	Peak	Not Peak
Demand	6745	4992	1753
Supply	2831 (42%)	1908 (38%)	923 (48%)
Lag	3914 (58%)	3084 (62%)	830 (52%)

CONCLUSION

There is a huge gap in supply and demand. In general, the supply and demand have a gap. There are a lot of rides which go incomplete (Unavailable or Cancelled) in comparison to being Completed.

Irrespective of how we consider the data, the Lag is greater than the Supply. In fact, the lag goes further down during peak hours but is still less than the Completed/Supplied trips.

In other words, the supply-demand gap is clear. It can be fixed perhaps, by focusing more on the peak hours first as they contribute in extra to the already existing gap. That should be the place to start.