EDA – PROJECT REPORT 2.

Name : Vikash Kumar Shrivastava.

Reg. No. : 12018607.

Roll No. : RK20RUA07.

Section : K20RU.

Submitted To : Abhijeet Dutta Sir.

DATA CLEANING.

Inspecting the dataframe

```
# Observing actual data in dataframe

uber_data
```

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp
0	619	Airport	1.0	Trip Completed	11/7/2016 11:51	11/7/2016 13:00
1	867	Airport	1.0	Trip Completed	11/7/2016 17:57	11/7/2016 18:47
2	1807	City	1.0	Trip Completed	12/7/2016 9:17	12/7/2016 9:58
3	2532	Airport	1.0	Trip Completed	12/7/2016 21:08	12/7/2016 22:03
4	3112	City	1.0	Trip Completed	13-07-2016 08:33:16	13-07-2016 09:25:47
...	...	...	...	...	...	...
6740	6745	City	NaN	No Cars Available	15-07-2016 23:49:03	NaN
6741	6752	Airport	NaN	No Cars Available	15-07-2016 23:50:05	NaN
6742	6751	City	NaN	No Cars Available	15-07-2016 23:52:06	NaN
6743	6754	City	NaN	No Cars Available	15-07-2016 23:54:39	NaN
6744	6753	Airport	NaN	No Cars Available	15-07-2016 23:55:03	NaN

```
# Check the column-wise info of the dataframe
uber_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6745 entries, 0 to 6744
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Request id         6745 non-null   int64
 1   Pickup point       6745 non-null   object
 2   Driver id          4095 non-null   float64
 3   Status             6745 non-null   object
 4   Request timestamp  6745 non-null   object
 5   Drop timestamp     2831 non-null   object
dtypes: float64(1), int64(1), object(4)
memory usage: 316.3+ KB
```

Summary of Inspecting the dataframe:

- 'Driver id' fields have decimal values however, it should be integer values
- 'Request timestamp' and 'Drop timestamp' fields have dates in different formats however, it should be in uniform format for analysis. Eg. '15-07-2016 10:00:43' vs. '11/7/2016 13:08'
- 'Driver id' fields and 'Drop timestamp' have many 'NaN' values

AFTER DATA CLEANING.

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp
0	619	Airport	1	Trip Completed	2016-07-11 11:51:00	2016-07-11 13:00:00
1	867	Airport	1	Trip Completed	2016-07-11 17:57:00	2016-07-11 18:47:00
2	1807	City	1	Trip Completed	2016-07-12 09:17:00	2016-07-12 09:58:00
3	2532	Airport	1	Trip Completed	2016-07-12 21:08:00	2016-07-12 22:03:00
4	3112	City	1	Trip Completed	2016-07-13 08:33:16	2016-07-13 09:25:47

```
uber_data.isnull().sum()
```

```
Request id           0
Pickup point         0
Driver id            0
Status               0
Request timestamp    0
Drop timestamp       0
Duration             0
Slot                 0
dtype: int64
```

ADDING SOME USEFUL FEATURES.

```
uber_data.head()
```

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp	Duration	Slot	Hour	Day
0	619	Airport	1	Trip Completed	2016-07-11 11:51:00	2016-07-11 13:00:00	69.0	Morning	11	Monday
1	867	Airport	1	Trip Completed	2016-07-11 17:57:00	2016-07-11 18:47:00	50.0	Evening	17	Monday
2	1807	City	1	Trip Completed	2016-07-12 09:17:00	2016-07-12 09:58:00	41.0	Morning	9	Tuesday
3	2532	Airport	1	Trip Completed	2016-07-12 21:08:00	2016-07-12 22:03:00	55.0	Night	21	Tuesday
4	3112	City	1	Trip Completed	2016-07-13 08:33:16	2016-07-13 09:25:47	52.5	Morning	8	Wednesday

DURATION: Shows the duration(time taken) of the trip.

SLOT : Shows the time slot in which the trip happened.

HOUR : Shows the hour in the which the request for the trip made.

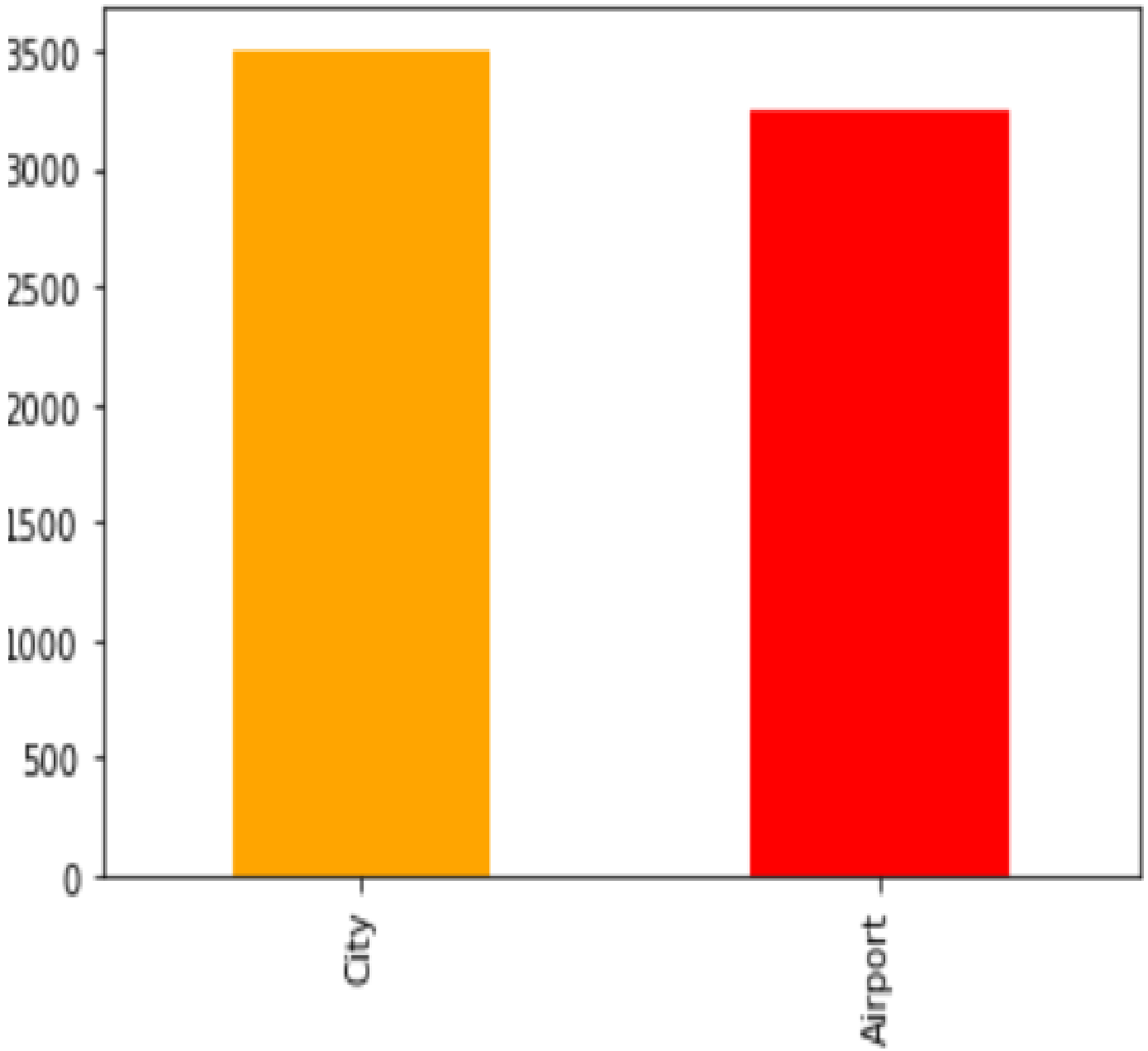
DAY : Shows the day of the trip.

UNIVARIATE ANALYSIS.

- Checking if any one of the two pick – up points is getting any uneven count of request for the trip or not.

So by plotting a graph for the same, we can see that there is not any uneven request i.e. if one pick-up point is getting a way much more request than the other.

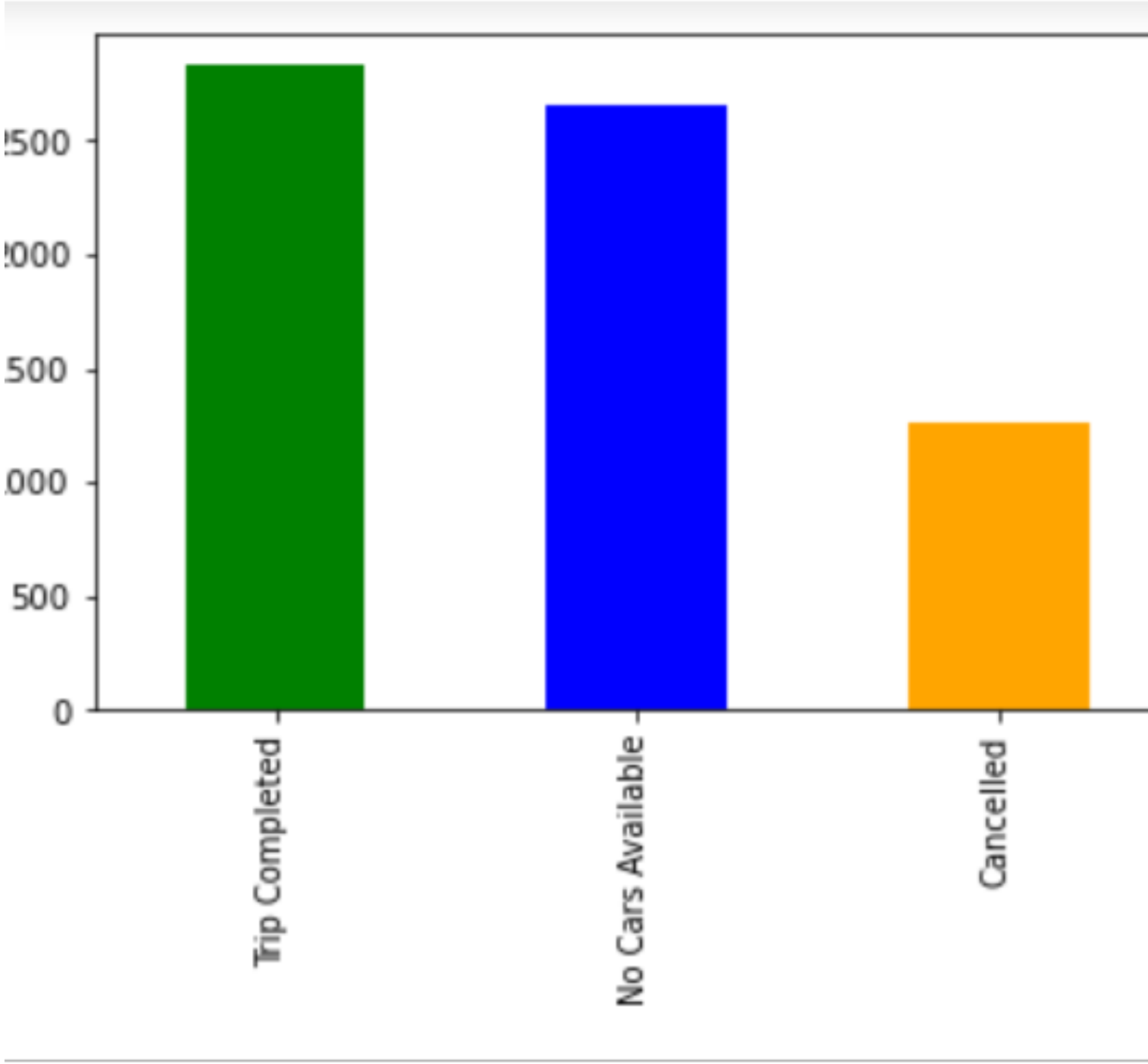
Both of the picking point are having approx equal number of requests.



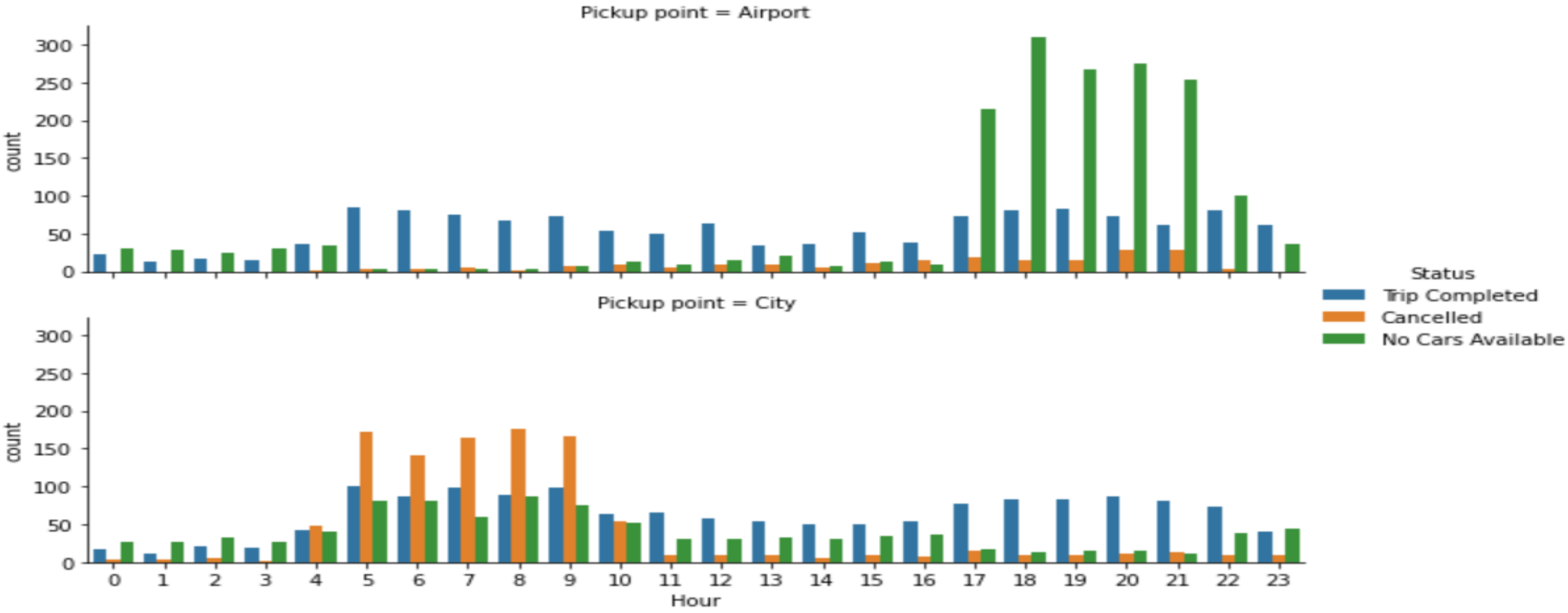
➢ Analysing the status of the trips.

Here we can see a interesting thing, we can see that more than 50% of the request is not getting fulfilled by the uber i.e.

(No car available + Cancelled) is greater than (Trip Completed).



Working On The Problem.

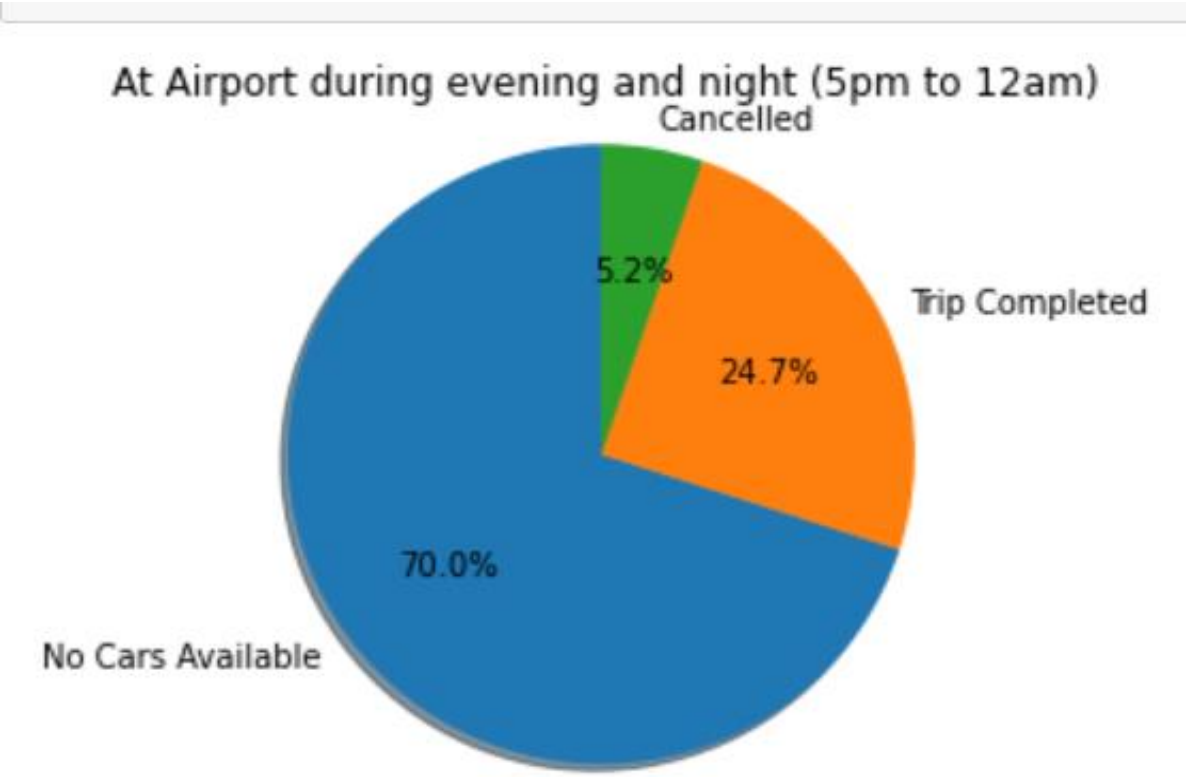


Inference 1: Uber is facing 'No Car Available' issue at Airport in evening and night (5pm to 12am).

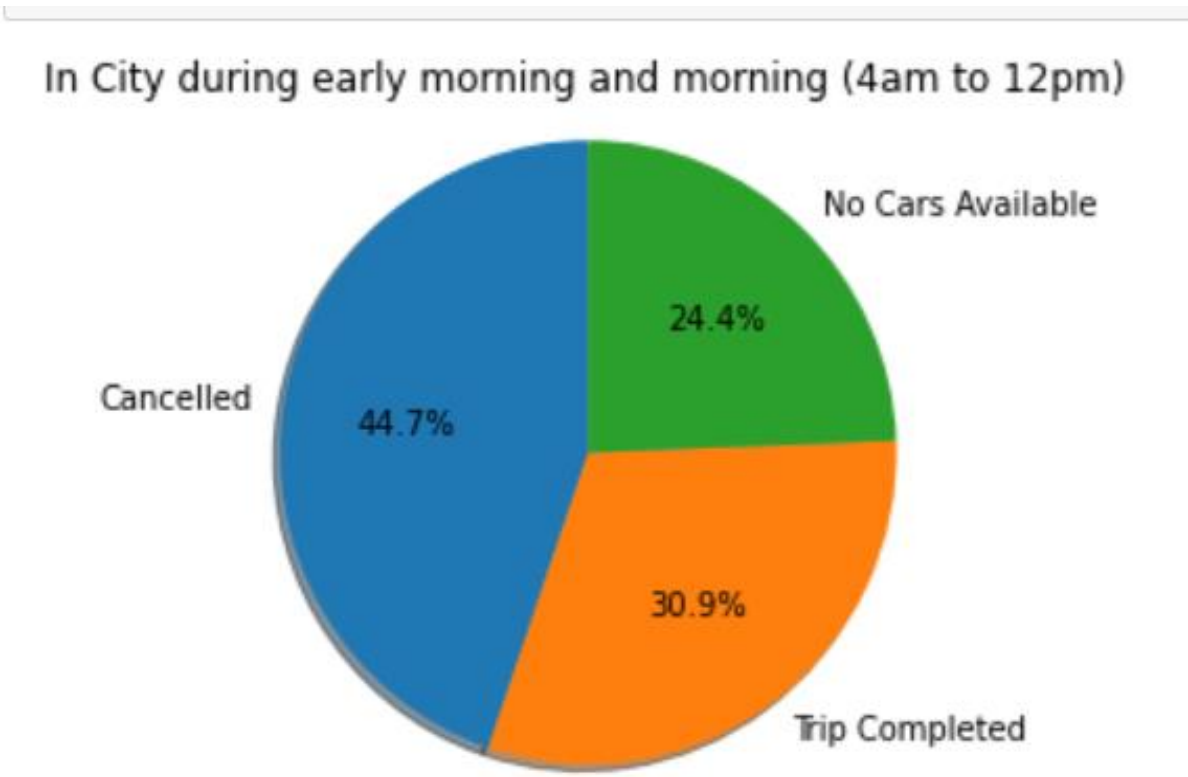
Inference 2: Uber is facing cab 'Cancelled' issue in City during early morning and morning (4am to 10am).

Inference 3: Uber is facing 'No Car Available' issue in City during early morning and morning (4am to 10am).

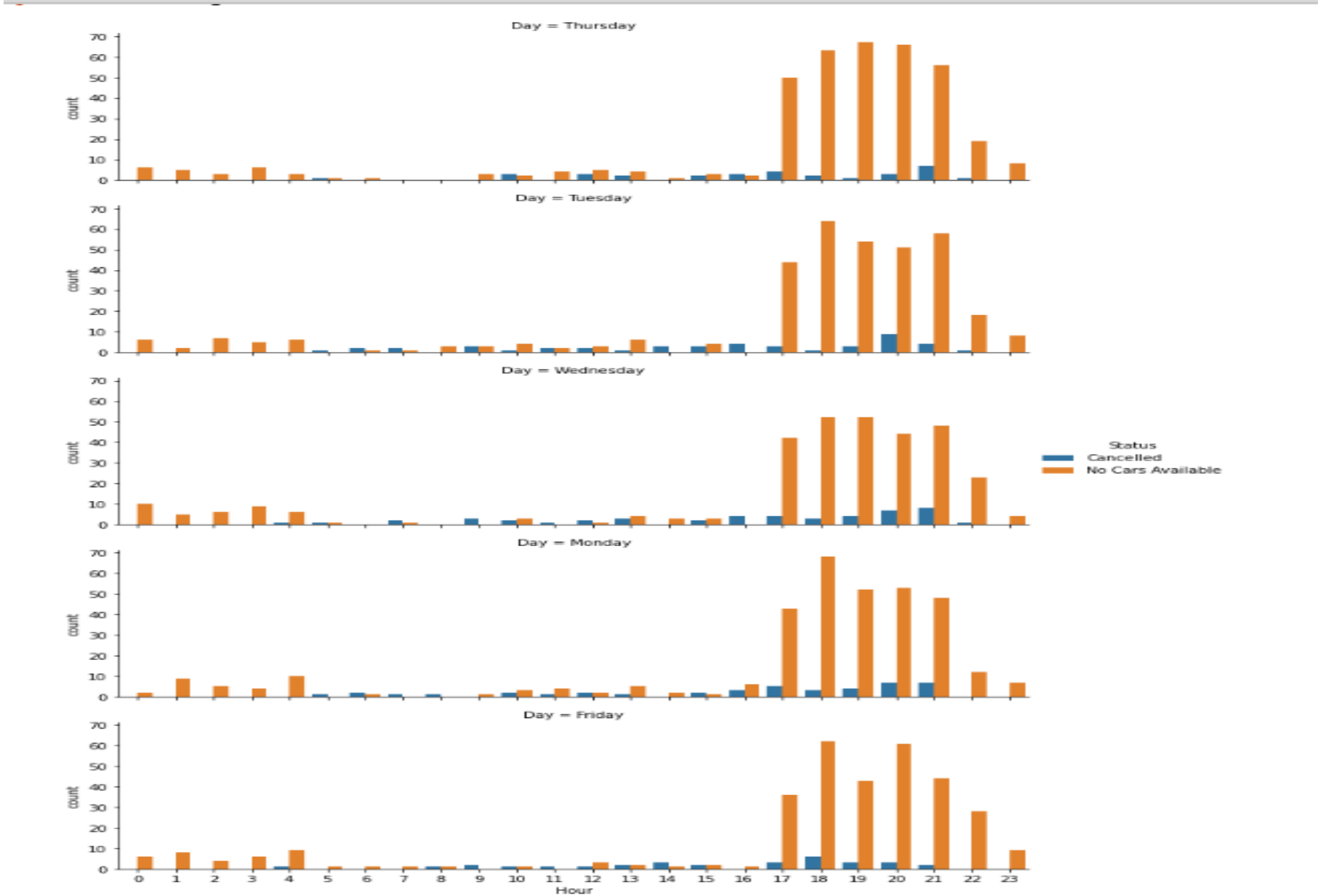
- **Only 24.7% of the demand were fulfilled during evening and night (5pm to 12am) at airport.**



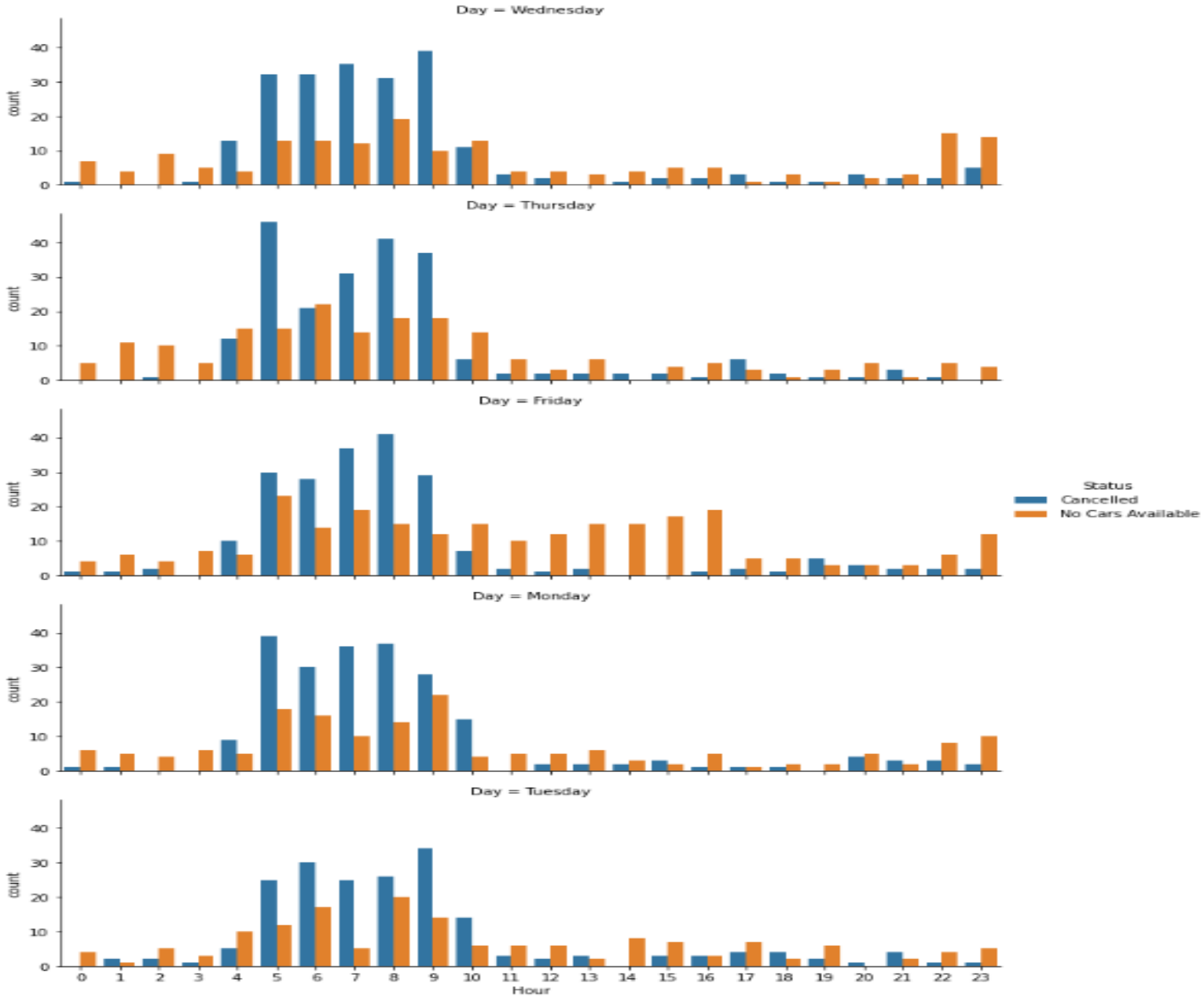
- **Only 30.9% of the demand were fulfilled during early morning and morning (4am to 12pm) at city.**



Uber is facing the same issue at airport for all the days of week.



Uber is facing the same issue at city also for all the days of the week.




Thank You!