Uber Supply-Demand Gap Analysis

Submitted by Aswathy Gopalakrishnan

1 INTRODUCTION

The project analyses Uber trip data to understand when there are more ride requests than available drivers. By examining trips statuses, driver availabilities, and time segments throughout the day, the project aims to identify times and areas with supply and demand gaps. The goal is to find patterns and help provide suggestion to improve driver availability during busy times. The initial data preprocessing is already done in SQL. Basic data analysis was then done in Excel and the dashboards are created to understand the data, and identify the supply-demand gap. Here, in the notebook file, detailed temporal analysis will be done to identify the peak time for requests and the availability of drivers during the same time frame.

The analysis is done in 3 phases:

SQL – Data Preprocessing

Excel – Dashboard Creation

Python – EDA

2 SQL ANALYSIS FOR DATA PREPROCESSING

The dataset *'Uber Request Data.csv'* was uploaded using table import feature on MySQL. A new database *'uber_analysis'* was created to perform the preprocessing tasks.

```
create database uber_analysis;
use uber_analysis;
select * from uber_request_data;
```

The timestamp columns were in text formats. These were converted to date columns.

```
-- Update the columns to datetime formats
Alter table uber_request_data
add column  request_time datetime,
add column drop_time datetime,
add column response_time int;

--  Update 'request_time' by parsing 'Request timestamp' string
Update uber_request_data
set response_time = case
    when drop_time is null then null
    when timestampdiff(minute, request_time, drop_time) < 0 then null
    else timestampdiff(minute, request_time, drop_time)
end;
```

```sql
-- Similarly,  Update 'drop_time' with converted date, handling different formats
Update uber_request_data
set request_time = case
    when `Request timestamp` like '%/%' then
        str_to_date(`Request timestamp`, '%m/%d/%Y %H:%i')
    when `Request timestamp` like '%-%' then -- checking if format is day-month-year
        str_to_date(`Request timestamp`, '%d-%m-%Y %H:%i:%s')
    else
        null -- handling unexpected formats
end
where `Request timestamp` is not null;

-- Set 'drop_time' to NULL where 'Status' indicates trip was canceled or no driver found
Update uber_request_data
set drop_time = case
    when `Drop timestamp` like '%NA%' or `Drop timestamp` is null then null
    when `Drop timestamp` like '%-%' then
        str_to_date(`Drop timestamp`, '%d-%m-%Y %H:%i:%s')
    when `Drop timestamp` like '%/%' then
        str_to_date(`Drop timestamp`, '%m/%d/%Y %H:%i')
    else null
end;
```

An additional column named *'response_time'* was created by taking the difference between request and drop times. This helped in understanding any ambiguities in the date columns.

```sql
-- Calculate response time in minutes
Update uber_request_data
set response_time = case
    when drop_time is null then null
    when timestampdiff(minute, request_time, drop_time) = 0 then null
    else timestampdiff(minute, request_time, drop_time)
end;
```

It was observed that some columns had negative values as the month and date columns in the *'response_time'* were switched for these rows. This was updated to correct this discrepancy.

```sql
-- There are drop times less than request time
Select * from uber_request_data where drop_time < request_time;

-- this happened due to the problems while entering the data, updating the dates to correct this
Update uber_request_data
set request_time = STR_TO_DATE(
  CONCAT(
    '2016-07-12 ',
    date_format(request_time, ' %H:%i:%s')
  ),
  '%Y-%m-%d %H:%i:%s'
)
where request_time > drop_time;
```

Then, there were some response times that were huge (>4000). This also happened due to errors that happened during the entry of the dates. In the *'drop_time'*, instead of '2016-11-08', the user updated it as '2016-12-07' which extended the time by almost one month. This was updated to the correct date to solve the discrepancy.

```
-- Some response times were greater than 4000
select * from uber_request_data where response_time >1000;
-- Correcting this, by adjusting the dates
Update uber_request_data
set drop_time = str_to_date(
  concat(
    '2016-11-08',
    time_format(drop_time, '%H:%i:%s')
  ),
  '%Y-%m-%d %H:%i:%s'
)
where response_time > 1000;
```

Additional columns were created by splitting datetime columns to date and time of the day or hour columns. This was done for easy analysis in the future phases.

```
-- Creating additional columns by splitting date and hour of request and drop times
Alter table uber_request_data
add column request_date date,
add column request_time_hhmm varchar(5),
add column drop_date date,
add column drop_time_hhmm varchar(5);

Update uber_request_data
set
  request_date = date(request_time),
  request_time_hhmm = date_format(request_time, '%H:%i'),
  drop_date = date(drop_time),
  drop_time_hhmm = date_format(drop_time, '%H:%i');
```

Finally, the *'Driver_id'* column was formatted by replacing *'NA'* values by *'0'*, and then changing the datatype to *'int'*.

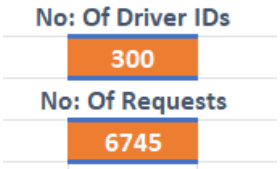
After this, the file was exported for further analysis in Excel.

3 Excel Analysis and Dashboard Creation

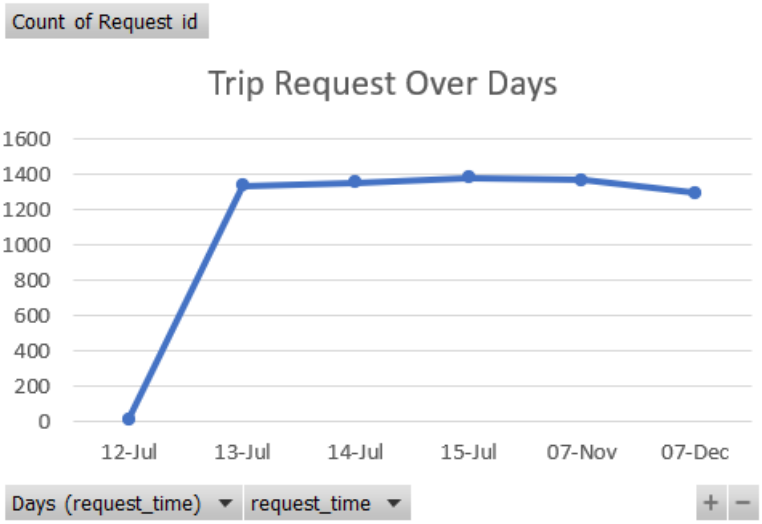
In the Excel phase, a basic understanding on the supply-demand gap was analysed to understand the given problem statement better. Pivot tables were made and corresponding graphs were plotted to prepare the dashboard.

The following visualizations were made and analysed:

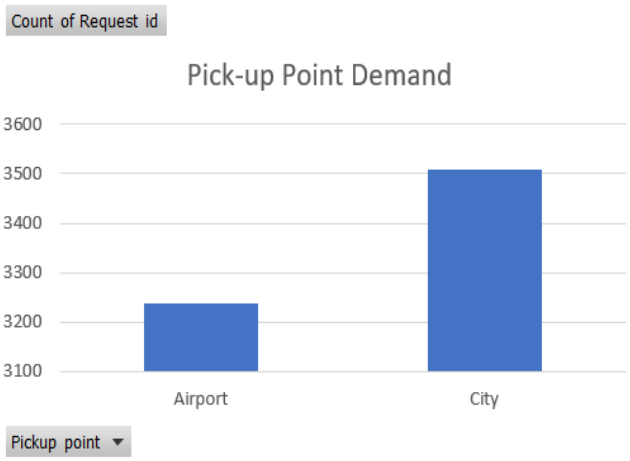
1. **Total Count** - The total number of requests(demand) and drivers(supply) were calculated. There are 300 drivers and 6745 distinct requests.



2. **Trip Requests Over Days** – This showed the numbers of requests that were made on distinct days in the dataset. This gives an idea on the demand on each day, and how consistent it is over time. It was observed that it is fairly consistent. The initial dip is due to inconsistent dataset which does not have enough values for 12th June.

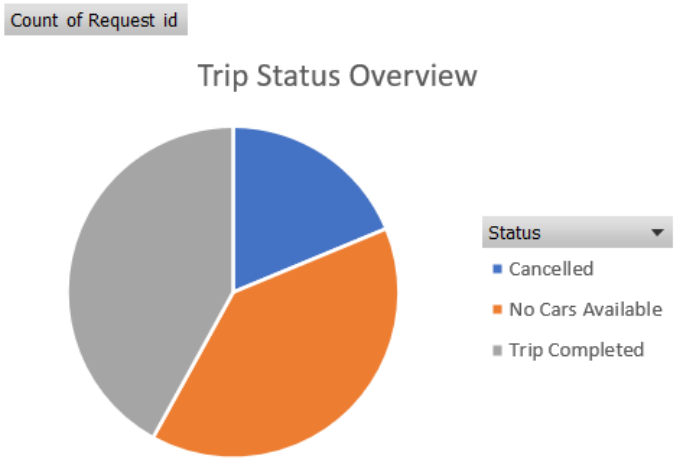


3. **Pick-up Location Demand** – There are two pick-up locations – *'Airport'* and *'City'*. This visualisation gives an idea on which location gets more request and hence has more demand. Visibly, city has more demand than airport.

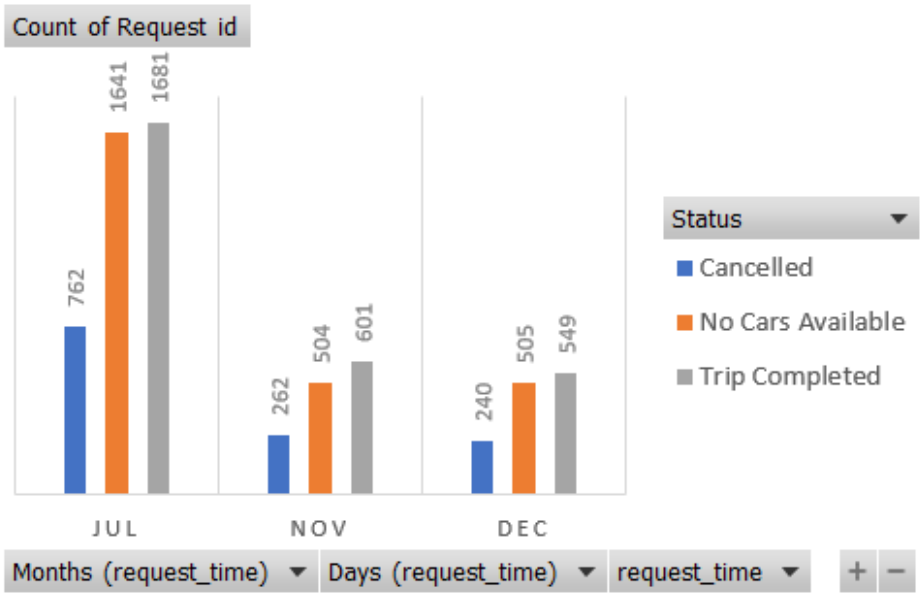


4. **Trip Status Overview** – There are three statuses – *'Cancelled'*, *'No cars Available'*, and *'Trip Completed'*. The percentage of each of these were analysed to understand how much demand is met. Most of the trips that were not completed were due to no cars being

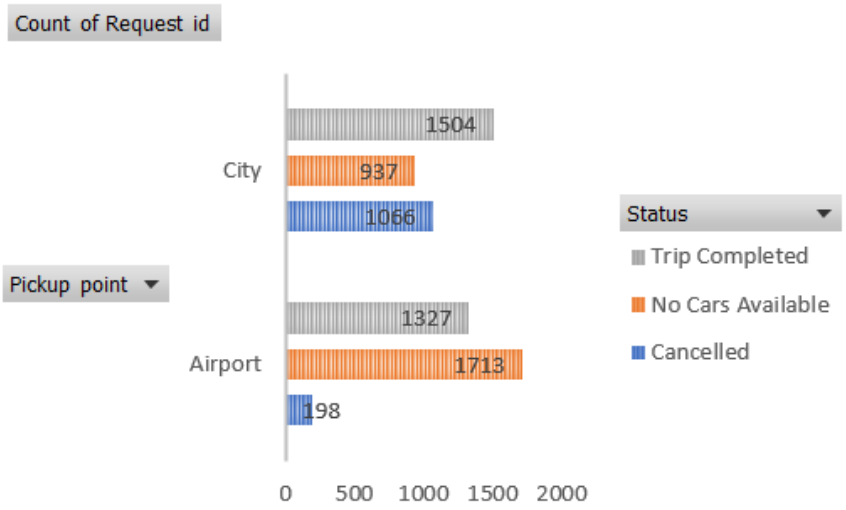
available. The cancellations were comparatively low. The reasons for this needs to be analysed using timeframes in advanced analysis.



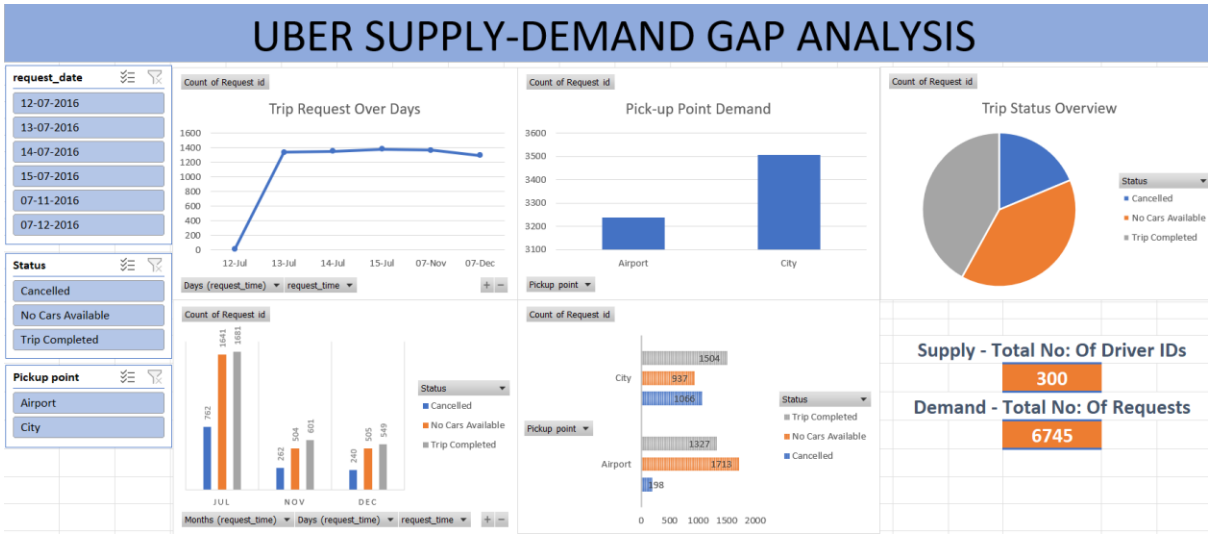
5. **Status Over Months** – The analysis was done to see if the pattern changes over time. July showed high number of requests. However, the dataset isn't of high quality or has balanced dataset. This could be the reason for the bias.



6. **Demand Based on Location and Status** – The bar chart helps to understand how the distribution of various status is occurring in the two pick-up locations. In cities, both reasons for trips not being completed are fairly similar (937 and 1066). However, in airport, the major reason is the cars not being available.



A dashboard has been created with the visualizations:



The major insights from excel analysis are as follows:

- There is a clear need to optimize driver allocation, especially in high-demand area, the city.
- Managing supply proactively during peak periods could reduce unavailability and thereby helps in improving the customer satisfaction.
- Monitoring-specific statuses across locations can help identify operational bottlenecks.

Clearly, there is a gap between the demand and supply. However, we need to analyse further to understand the main reasons for this to happen. For this, EDA using python in Jupyter Notebook was performed.

4 PYTHON EDA FOR TEMPORAL ANALYSIS

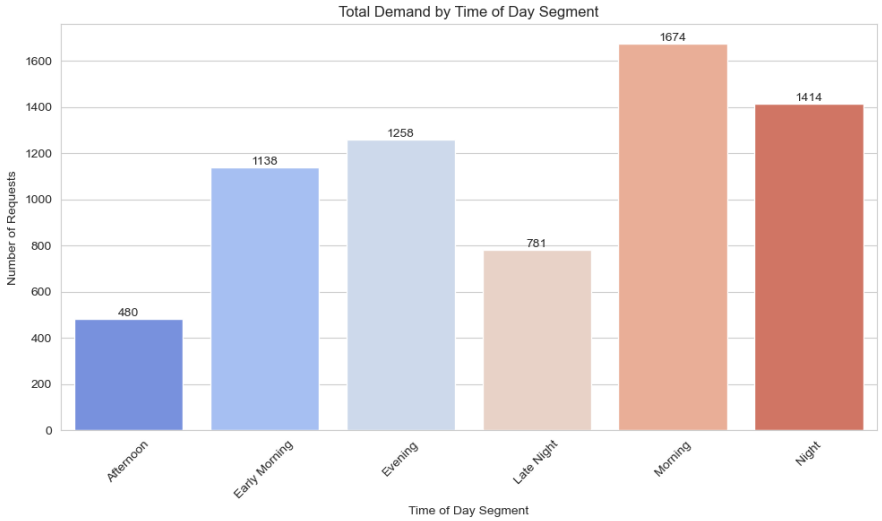
The excel analysis revealed that there is in fact a pattern in the trips not being completed. Further analysis needs to be time to identify trends over different timeframes of the day. For this, an additional feature *'time_of_day_segment'* was created. In this, the durations of the day are split into different segment as shown below in the code snippet.

```
if 0 <= hour < 3:
    return 'Late Night'
elif 3 <= hour < 7:
    return 'Early Morning'
elif 7 <= hour < 12:
    return 'Morning'
elif 12 <= hour < 15:
    return 'Afternoon'
elif 15 <= hour < 19:
    return 'Evening'
elif 19 <= hour < 22:
    return 'Night'
elif 22 <= hour < 24:
    return 'Late Night'
else:
    return 'Unknown'
```

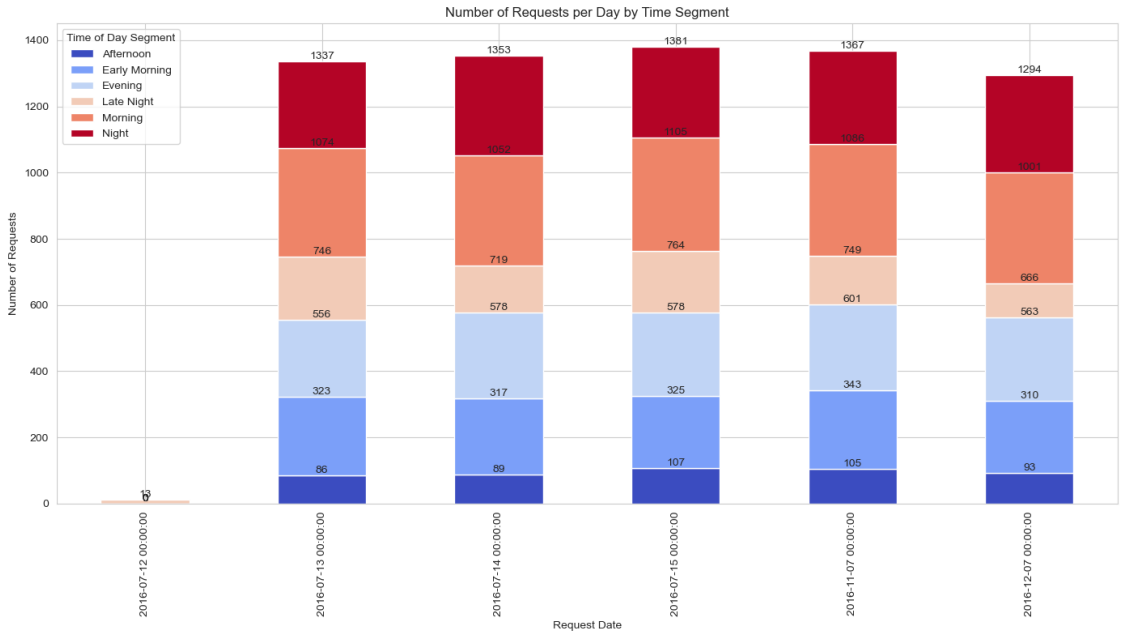
After this, the temporal analysis was done using visualizations of demand, supply, and gap over various factors.

The following visualizations were made:

1. **Total Demand by Time-of-Day Segment**
   The visualization shows how the requests are spread across different time segments along a day. It will help to understand the segments where demand is really high as well as low. The highest demands are during 'Morning' and 'Night' segments with 1674 and 1414 total requests respectively. The lowest demand is in the 'Afternoon' time segment. More resources can be shifted to morning and night rides, and some resources can be spared during the afternoon. There may be reasons why the drivers choose specific timeslots. Driver's availability and location preference must be considered before making the choices.



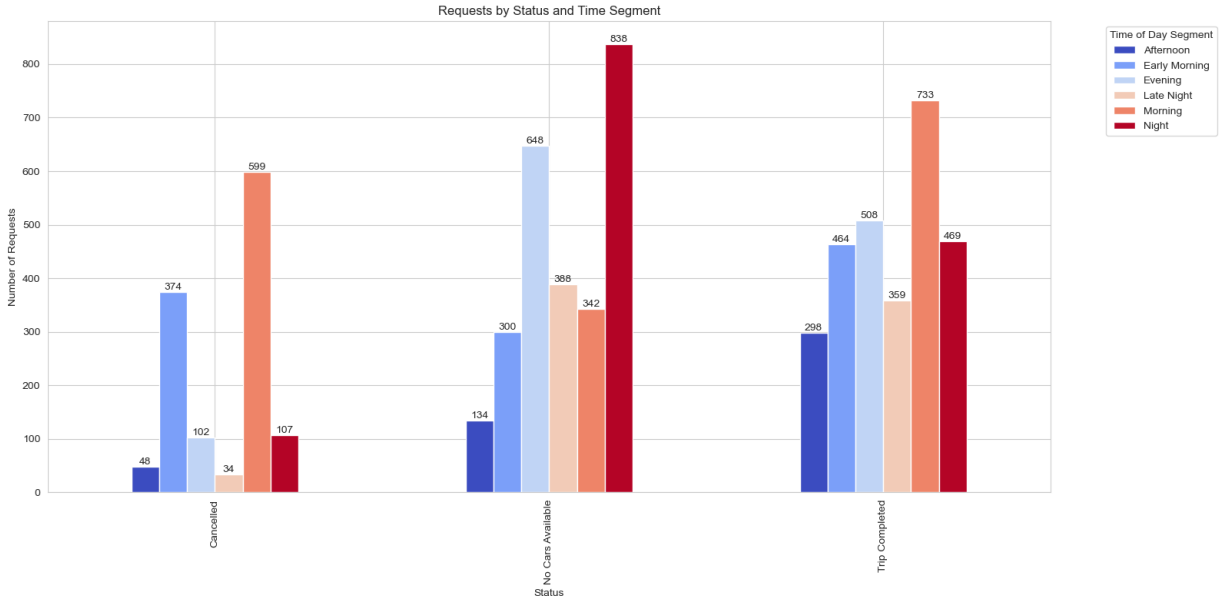
2. **Number of Requests per Day by Time Segment**
   The total number of requests on distinct request dates were analysed based on the time segments. This helps to identify if the trends we identified earlier are consistent over the days. The chart shows that the pattern is fairly similar with more requests in the morning and night segments and the lowest in the afternoon on all days, reaffirming whatever we have observed earlier. This confirms that the proper allocation of resources must be done on a daily basis to reduce the demand-supply gap.



3. Requests by Status and Time Segment

Now that the time segments with high and low demands have been identified, the analysis is done to understand the primary reasons for demand-supply gap occurring in these segments. This visualization helps to understand that. It is observed that, the gap occurs during night mostly because no cars are available during the segment. However, morning rides are mostly cancelled.

Driver's preferences play a major role in this as we had discussed earlier. Not many drivers are interested in taking night rides. In the morning, due to the high demand, the drivers may have more options to choose from based on the location and the estimated charge of the trip. We had observed earlier that cancellations are less in airport. This could be cos of the higher fare. When given an option, drivers would choose the higher one. Some system must be employed to reduce cancellations. However, to limit this won't be a fair change considering the driver's preferences though.

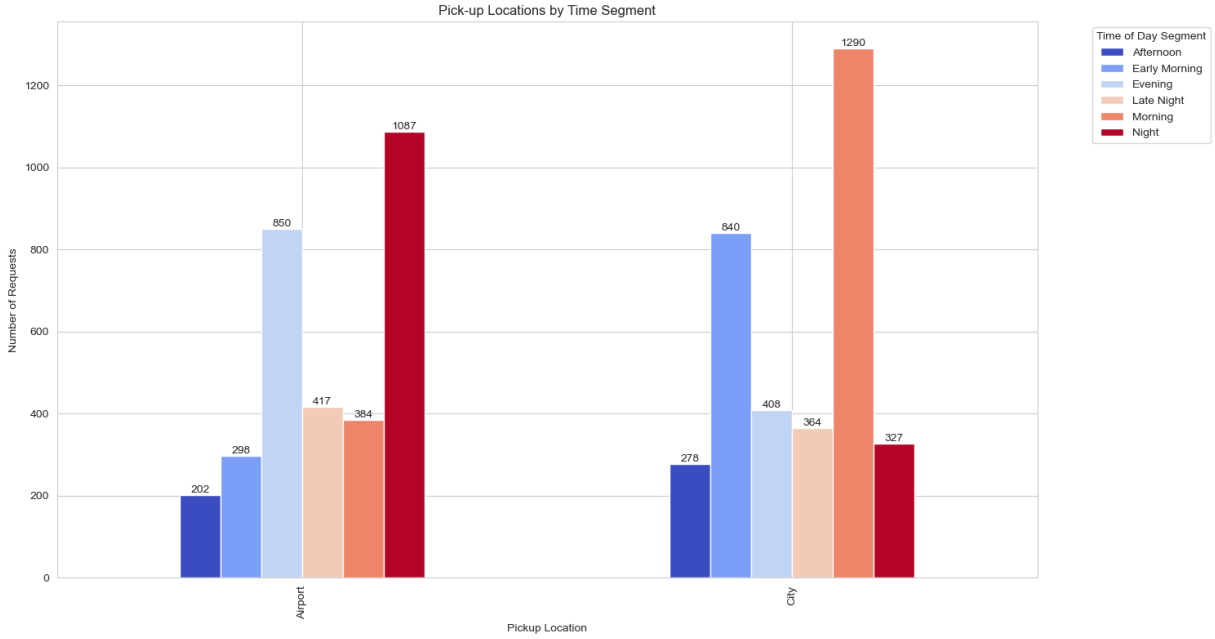


4. **Pick-up Locations by Time Segment**

It was already discussed that cancellations happen more in the airport. Further analysis was done here to see how the distribution of demand is in the locations in different time segments.

It is observed that there is a high demand for airport trips at night compared to city. In previous analysis, we have seen that the supply is less at night. Connecting these two observations, it can be safely said that the less supply is causing more cancellations for airport trips at night. On the other side, city trips have high demand in the morning. There are enough drivers active in the morning. However, they may be choosing the trips that can pay them extra, leading to more cancellations.

This indicates that the supply must be increased in the morning, and better allocation practices need to be set up for the morning rides.

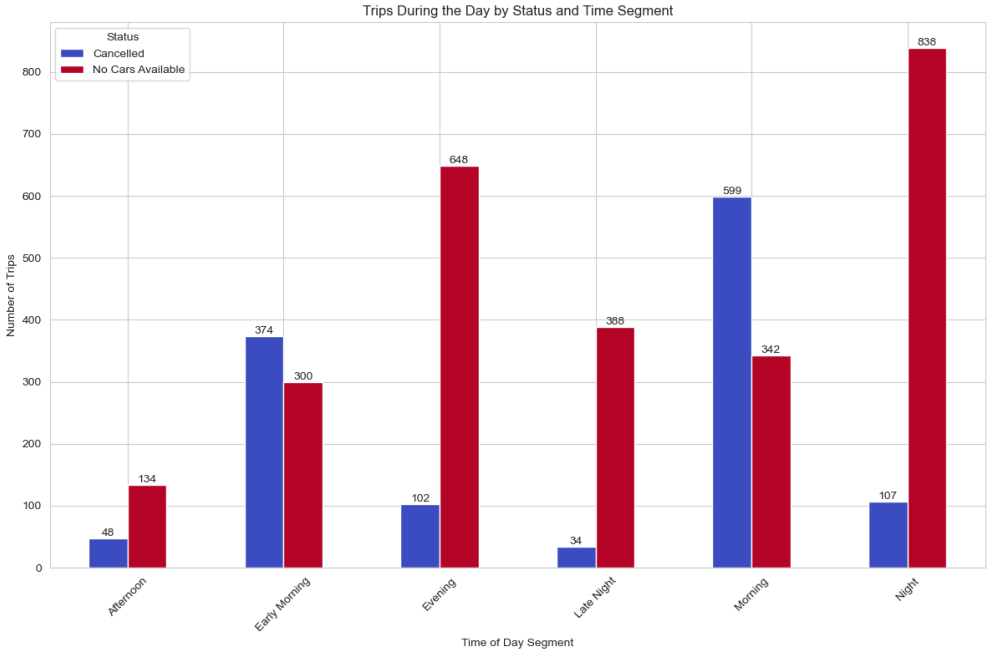


5. **Trips During the Day by Status and Time Segment**

The visualization helps to identify the reason for gap in specific time segments. This was done to reaffirm the observations to help with better allocation of resources.

Night has a really high bar for no cars being available. It's the same for late night and evening travels as well. For early morning and morning rides, cancellations are higher. Afternoons seem to have less demand and less supply too. So, we can safely say that, from late afternoon, the demand is much more than supply.

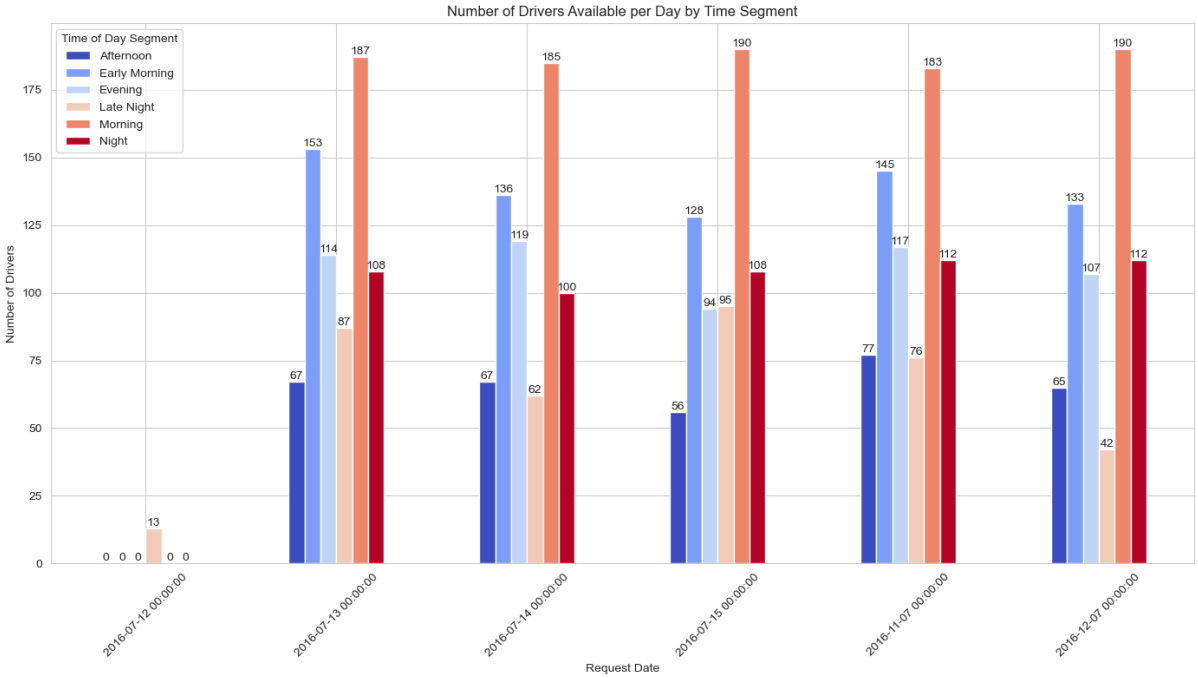
There must be some tactics included for travels after evening to meet the demand. Bonuses, points, or badges that can help the driver improve their profile can be some of the options to this. This way, more drivers will be motivated to work during these time slots.



6. **Number of Drivers Available per Day by Time Segment**

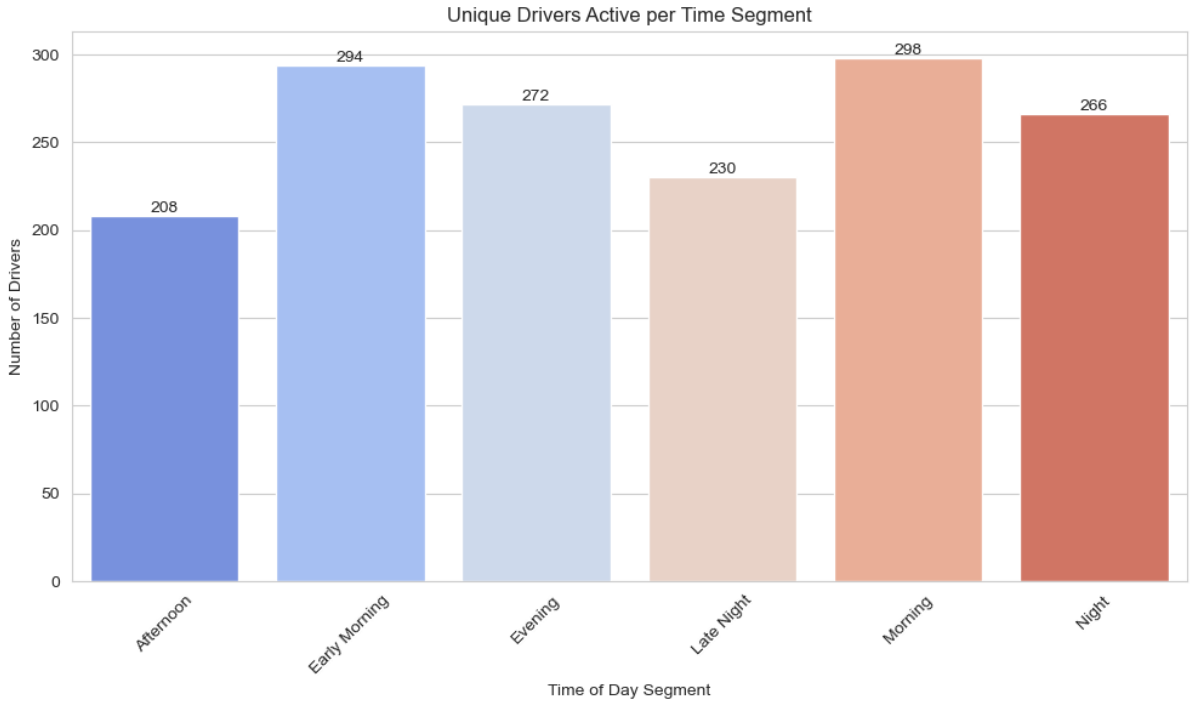
The previous analysis clearly showed how the demand is affected based on several features. The following analysis is part of identifying the problems in the supply, how the drivers are distributed across different segments.

More drivers are available in the morning and early morning segments. There are very few available during night and late-night segments. This clearly defines the gap. More drivers must be encouraged to work during night time.



7. **Unique Drivers Active per Time Segment**

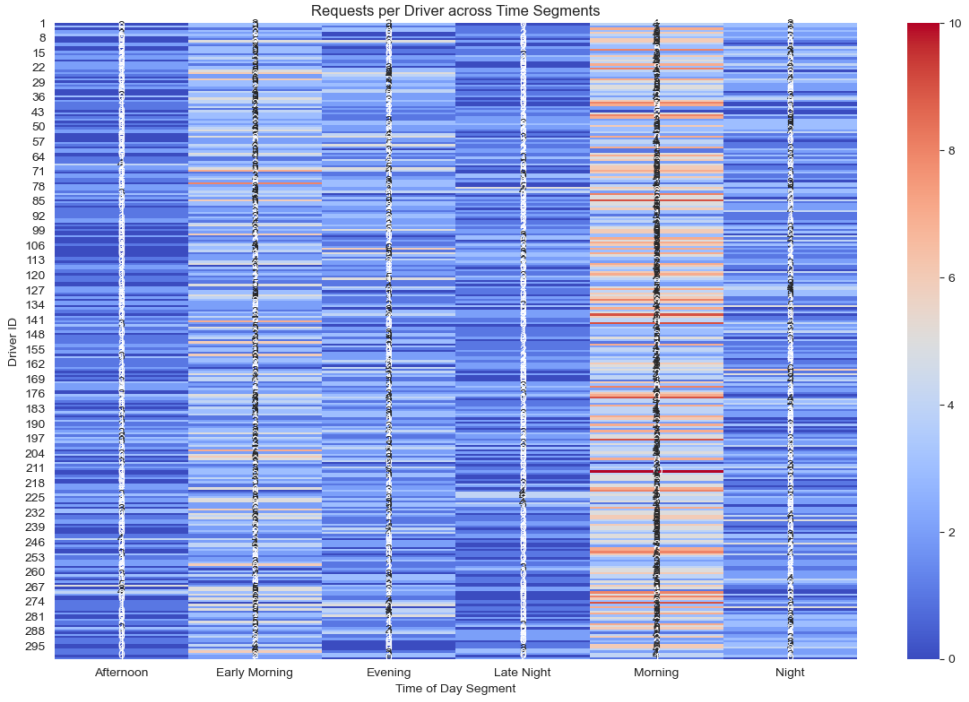
By identifying unique drivers in each segment, it is possible to identify the distribution of the 300 drivers across the day. Almost all the drivers are active in the morning (298 out of 300). However, at night, it reduces to 266 and 230 during late night. Drivers prefer to work more in the morning due to high demands and better locations.



8. Requests per Driver across Time Segments

A heatmap has been visualised to understand the number of requests taken by individual drivers over the day. This gives a clear as in why drivers choose to work more in the morning times. Clearly, morning times provide more rides for each driver. Some has up to 10 rides in the morning. Most of the drivers have below 5 rides in the night segments. This clearly shows the reason for the drivers choosing to work more in the morning and not in the late evenings and night times.

Without enough rides, it won't be fair to push the drivers more to the night rides. Maybe, the allocation needs to be done better. Most of the drivers work throughout the day. A fair and beneficial allocation must be done to meet the demand.



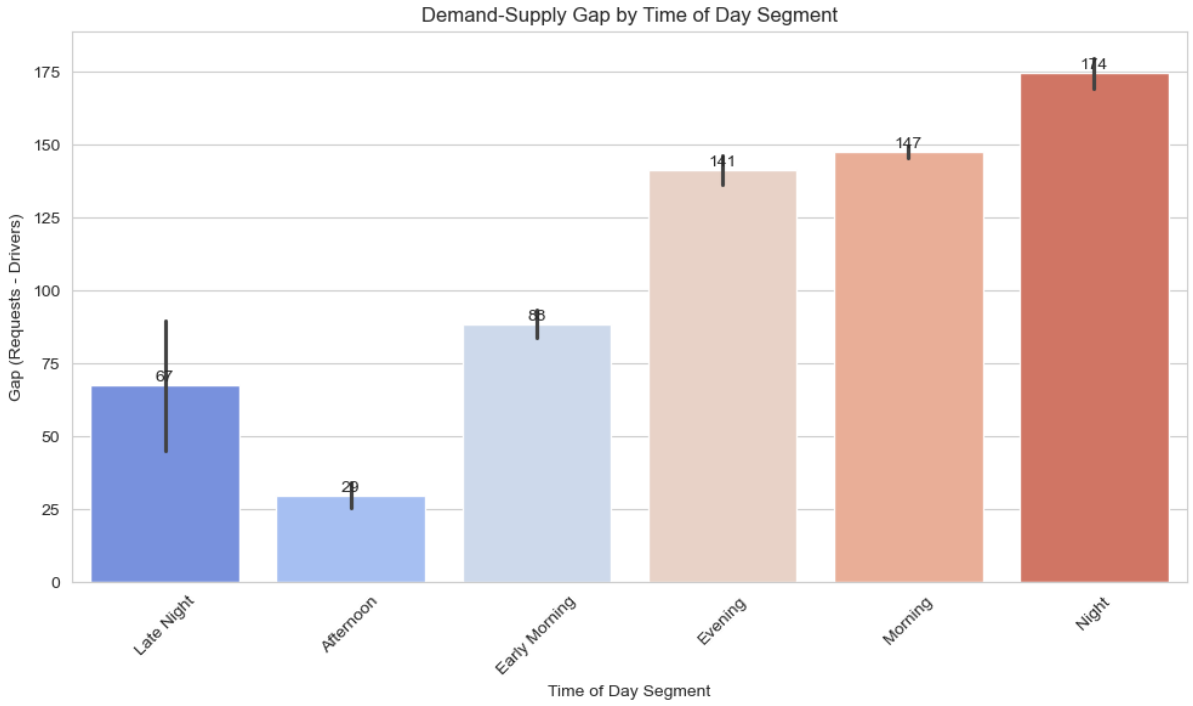
9. Demand-Supply Gap by Time-of-Day Segment

The analysis on demand and supply has been completed. The following visualization helps to summarise the gap for the time segment. The gap was calculated by taking the difference of supply and demand. The lines above each bar in the chart are error bars, representing variability or uncertainty. In this context, they could indicate the range of variation in the gap data or the confidence interval.

Main insights are:

- Morning and night segments have the highest gaps, indicating significant unmet demand.
- Early Morning and late-night show lower gaps, suggesting demand and supply are more balanced.
- Afternoon also has a relatively low gap.

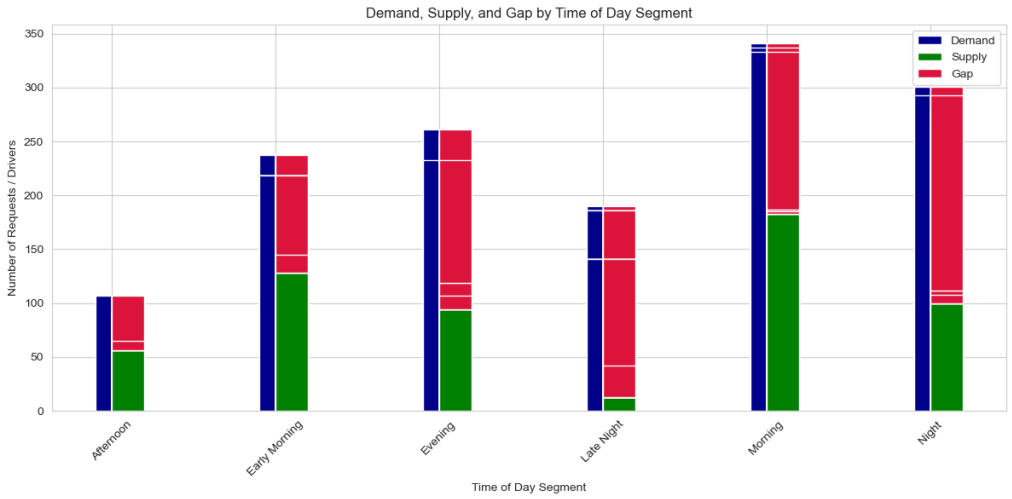
Focus must be morning and night segments as these show highest unmet demand. Driver shifts must be adjusted to cover peak times, taking into consideration the location demands as well.



10. Demand, Supply, and Gap by Time-of-Day Segment

A final graph was plotted to see the demand, supply and gap in one place for better comparison. The demand (blue) is consistently higher than supply (green), especially during Morning and Night. The gap (red) indicates unmet demand, which highlights periods where driver supply may need to be improved. The Morning and Night segments have the highest gaps, suggesting these are critical periods for driver deployment.

More focus should be on increasing driver availability during the times of High demand and low supply. Low gap periods may not need immediate actions. However, proper monitoring must be implemented to identify changes in future.



5 Major Insights and Solutions from Analysis

Based on your detailed analysis and visualizations, the following are the business solutions that directly address the identified supply-demand gaps:

- Increase driver shifts during peak times (morning and night) to meet high demand, especially in high-gap locations.
- Offer incentives, bonuses, or badges for drivers who work in low supply but high demand periods such as late-night and early mornings.
- Focus on increasing driver availability in high-demand areas, such as airports and city hotspots. Use geo-analytics to identify congestion areas and deploy drivers strategically in real-time.
- Offer flexible working hours, especially for drivers preferring nighttime shifts.
- Introduce shift reviews periodically to match supply with demand effectively.
- Implement predictive analytics to forecast demand trends hourly and daily to optimize driver allocation.
- Encourage riders to use services during low-demand periods with discounts.
- Use driver and user feedbacks to continually improve scheduling practices and identify bottlenecks.

By implementing these practices, a better supply for the demands, thereby reducing the gap can be achieved.

6 Conclusion

The analysis provides a comprehensive analysis of Uber's demand and supply dynamics, highlighting critical time segments that experience the highest unmet demand. Through detailed data preprocessing, exploratory data analysis, and visualizations, the key areas were identified where driver supply needs to be optimized to reduce gaps. The insights gained emphasize the importance of strategic driver scheduling, location-specific deployment, and incentivization during high-demand periods. Implementing these targeted interventions can enhance operational efficiency, improve customer satisfaction, and ensure a balanced supply-demand ecosystem.