Analyzing Uber's Performance in New York City: Assessing Seasonal Variations in Ride-for-Hire Services in New York City

Payel Ghosal

2024-04-06

Introduction:

The transportation landscape in New York City has undergone significant changes with the rise of ride-for-hire services like Uber. This report aims to analyze Uber's performance in the city during the months of January and February 2015, focusing on key metrics such as trip volume, active vehicle utilization rate, and daily trip distribution. By leveraging data from the "Uber-Jan-Feb-FOIL" dataset, this report provides insights into Uber's operational dynamics and its impact on mobility patterns in New York City.

Data:

The analysis is based on data collected from the Uber-Jan-Feb-FOIL dataset, which includes information on dispatching base numbers, dates, active vehicles, and trips. The data was processed and visualized using the R programming language, with the ggplot2 and dplyr libraries utilized for data manipulation and visualization.

```
uber <- read.csv("D:\\MS-2\\Sem-4\\STAT-6220_Consulting\\Uber-Jan-Feb-FOIL.csv",
                 sep=",",header = T)
head(uber)
```

```
##   dispatching_base_number       date active_vehicles trips
## 1                  B02512 01-01-2015             190  1132
## 2                  B02765 01-01-2015             225  1765
## 3                  B02764 01-01-2015            3427 29421
## 4                  B02682 01-01-2015             945  7679
## 5                  B02617 01-01-2015            1228  9537
## 6                  B02598 01-01-2015             870  6903
```

```
str(uber)
```

```
## 'data.frame':    354 obs. of  4 variables:
##  $ dispatching_base_number: chr  "B02512" "B02765" "B02764" "B02682" ...
##  $ date                   : chr  "01-01-2015" "01-01-2015" "01-01-2015" "01-01-2015" ...
##  $ active_vehicles        : int  190 225 3427 945 1228 870 785 1137 175 890 ...
##  $ trips                  : int  1132 1765 29421 7679 9537 6903 4768 7065 875 5506 ...
```

Data Cleaning:

Necessary data cleaning is done to ensure the correct form of `dates` in the data.

```
uber$date <- rep(seq(from = as.Date("2015-01-01"),
                     to = as.Date("2015-02-28"), by = 'day'),each=6)
```

```r
# Convert 'date' column to date format
uber$date <- as.Date(uber$date)
```

Storyboarding:

Seasonal fluctuations in transportation demand can have significant implications for ride-for-hire services in urban areas like New York City. This story aims to analyze the data for January and February 2015 to assess seasonal variations in ride-for-hire services, focusing on dispatching base numbers, dates, active vehicles, and trips. By examining trip volumes, vehicle utilization rates, and trends over the two-month period, the story will uncover insights into how seasonal factors impact the operation and utilization of ride-for-hire services in the city.

Visualization:

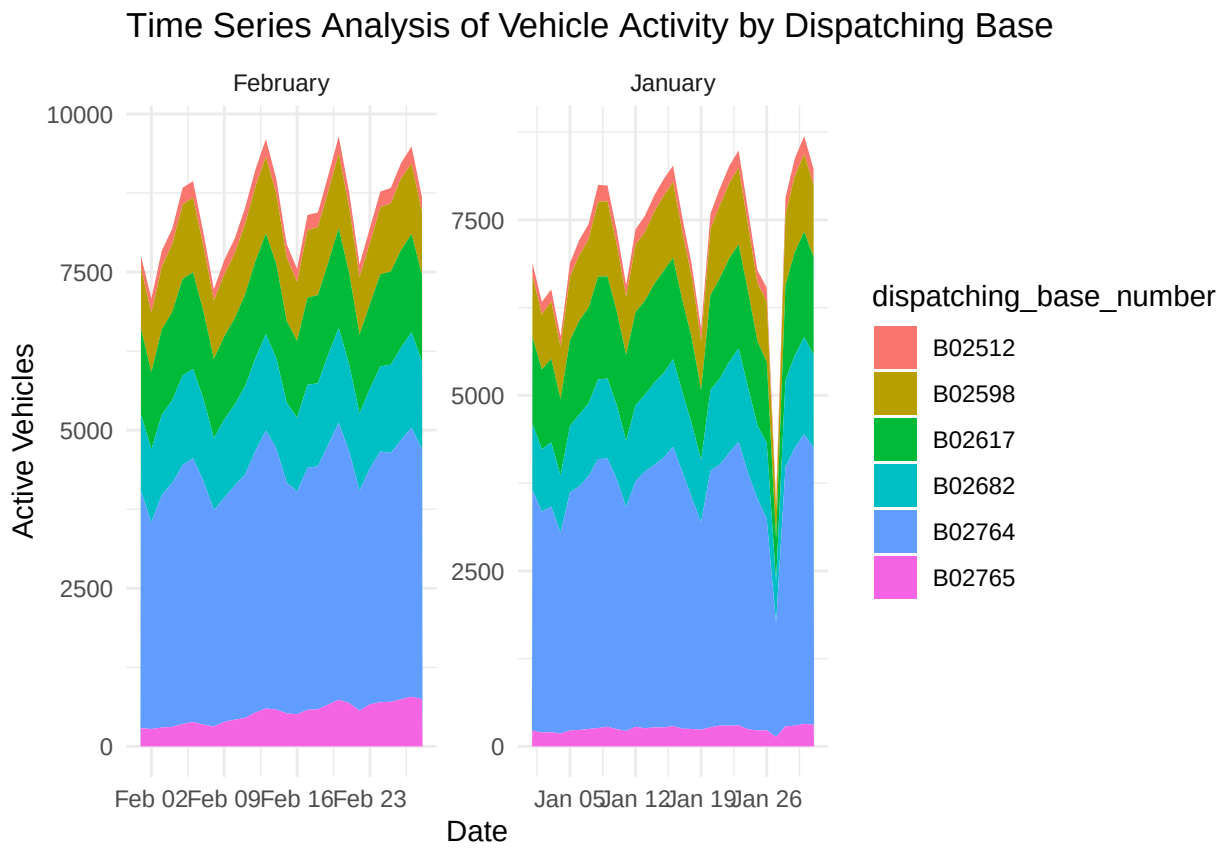
1. Time Series Analysis of Vehicle Activity:

Separate stacked area charts or line graphs showing the distribution of active vehicles by dispatching base number over time for January and February. These visualizations will reveal any trends or patterns in vehicle availability and fleet size for each month.

```r
# Time Series Analysis of Vehicle Activity
time_series_vehicle_activity <- uber %>%
  mutate(month = format(date, "%B")) %>%
  group_by(month, date, dispatching_base_number) %>%
  summarise(active_vehicles = sum(active_vehicles)) %>%
  ggplot(aes(x = date, y = active_vehicles, fill = dispatching_base_number)) +
  geom_area() +
  facet_wrap(~month, scales = "free") +
  labs(title = "Time Series Analysis of Vehicle Activity by Dispatching Base",
       x = "Date",
       y = "Active Vehicles") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'month', 'date'. You can override using the
## `.groups` argument.
```

```r
time_series_vehicle_activity
```



The time series analysis illustrated trends in vehicle activity by dispatching base number over the two-month period. Variations in the distribution of active vehicles over time were observed, indicating fluctuations in fleet size and availability.

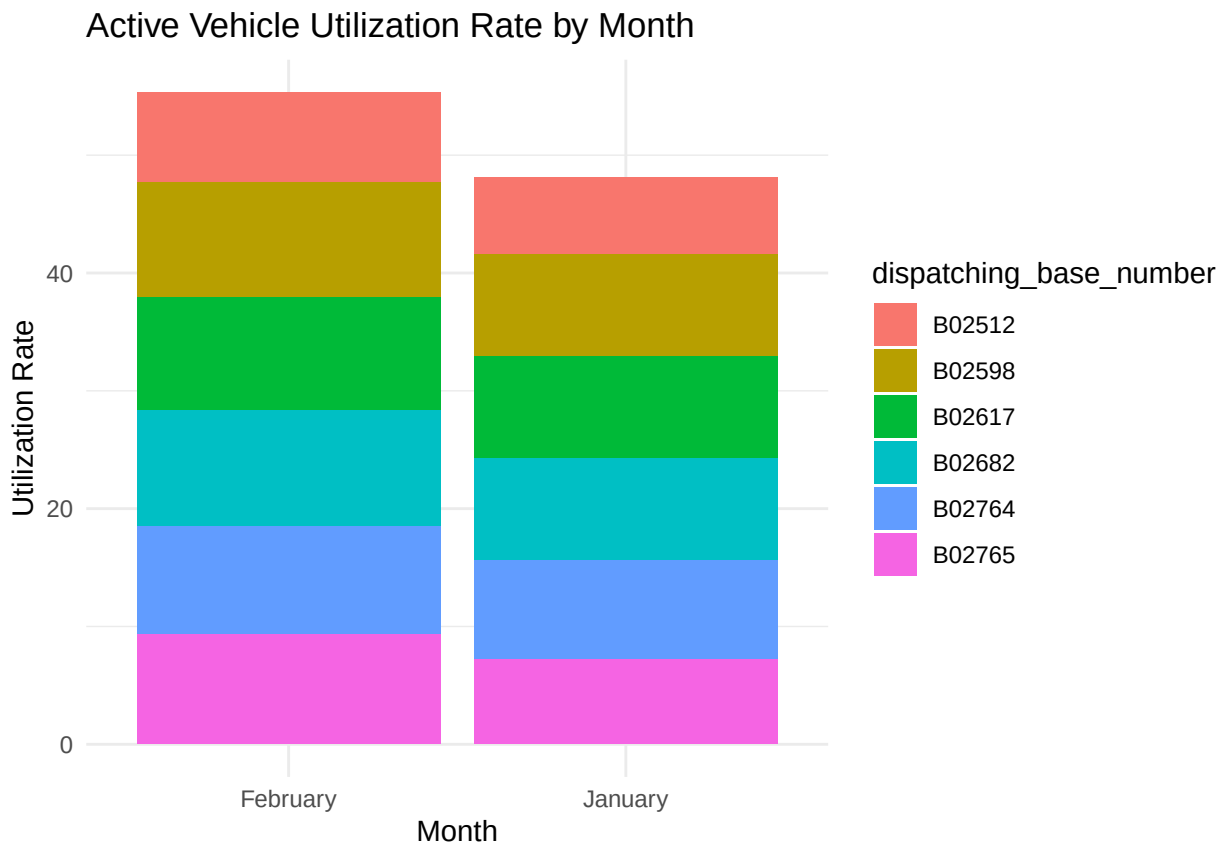
2. Active Vehicle Utilization Rate by Month:

A bar chart displaying the utilization rate of active vehicles for each dispatching base number, separately for January and February. This visualization will illustrate variations in vehicle utilization efficiency between the two months.

```
# 2. Active Vehicle Utilization Rate by Month
active_vehicle_utilization <- uber %>%
  mutate(month = format(date, "%B")) %>%
  group_by(month, dispatching_base_number) %>%
  summarise(utilization_rate = sum(trips) / sum(active_vehicles)) %>%
  ggplot(aes(x = month, y = utilization_rate, fill = dispatching_base_number)) +
  geom_bar(stat = "identity") +
  labs(title = "Active Vehicle Utilization Rate by Month",
       x = "Month",
       y = "Utilization Rate") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

```
active_vehicle_utilization
```



- The analysis of active vehicle utilization rates indicated differences in efficiency among dispatching bases.
- B02512, B02765 demonstrated higher utilization rates in February, suggesting more effective deployment of vehicles to meet demand.

3. Daily Trip Distribution Heatmap for January and February:

Two separate heatmaps overlaying trip distribution by dispatching base number and date for January and February. These visualizations will identify any differences in peak days and times of trip activity between the two months.

```
# 3. Daily Trip Distribution Heatmap for January and February
daily_trip_distribution_Jan <- uber %>%
  mutate(day = format(date, "%d")) %>%
  filter(format(date, "%m") %in% c("01")) %>%
  group_by(day, dispatching_base_number) %>%
  summarise(trip_count = sum(trips)) %>%
  ggplot(aes(x = dispatching_base_number, y = day, fill = trip_count)) +
  geom_tile() +
  labs(title = "Daily Trip Distribution Heatmap for January",
       x = "Dispatching Base Number",
       y = "Day of Month") +
  theme_minimal() +
  scale_fill_viridis_c()
```

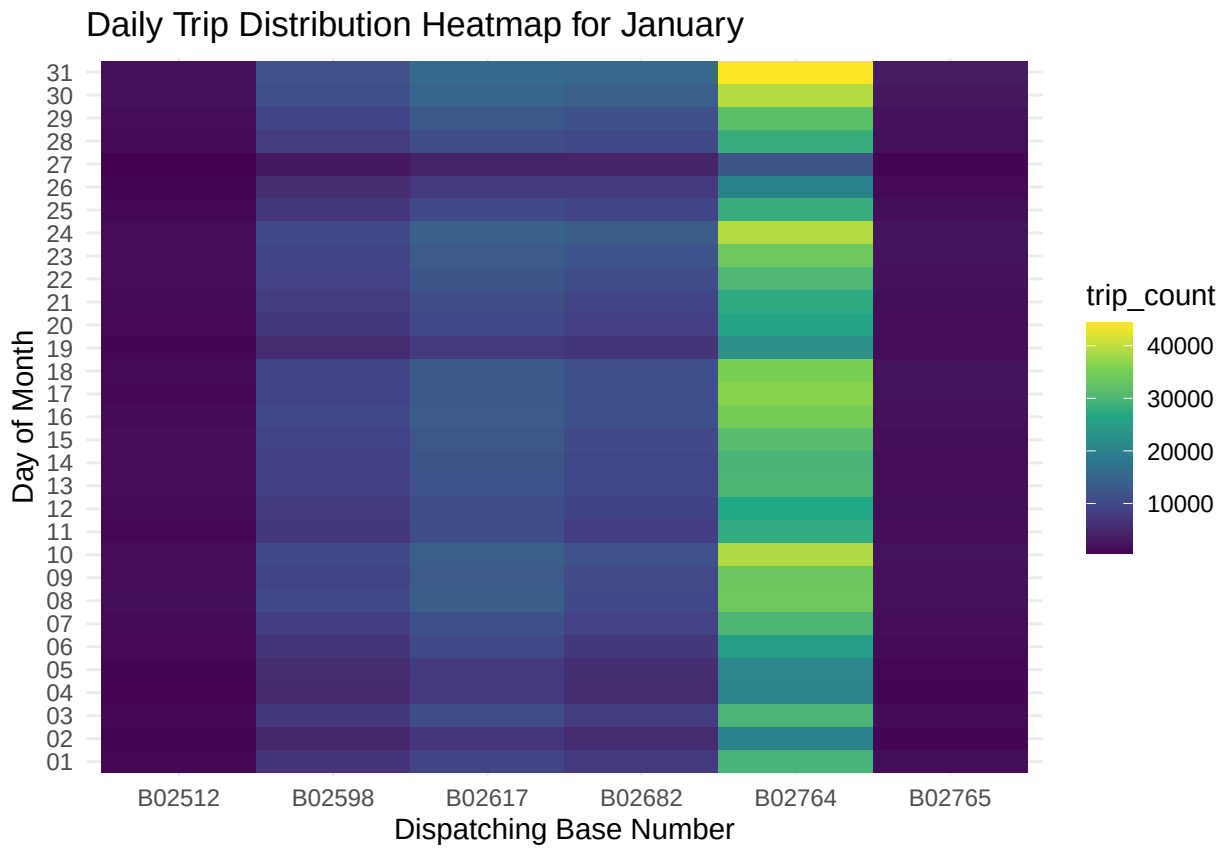
```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
```

```
## argument.
```

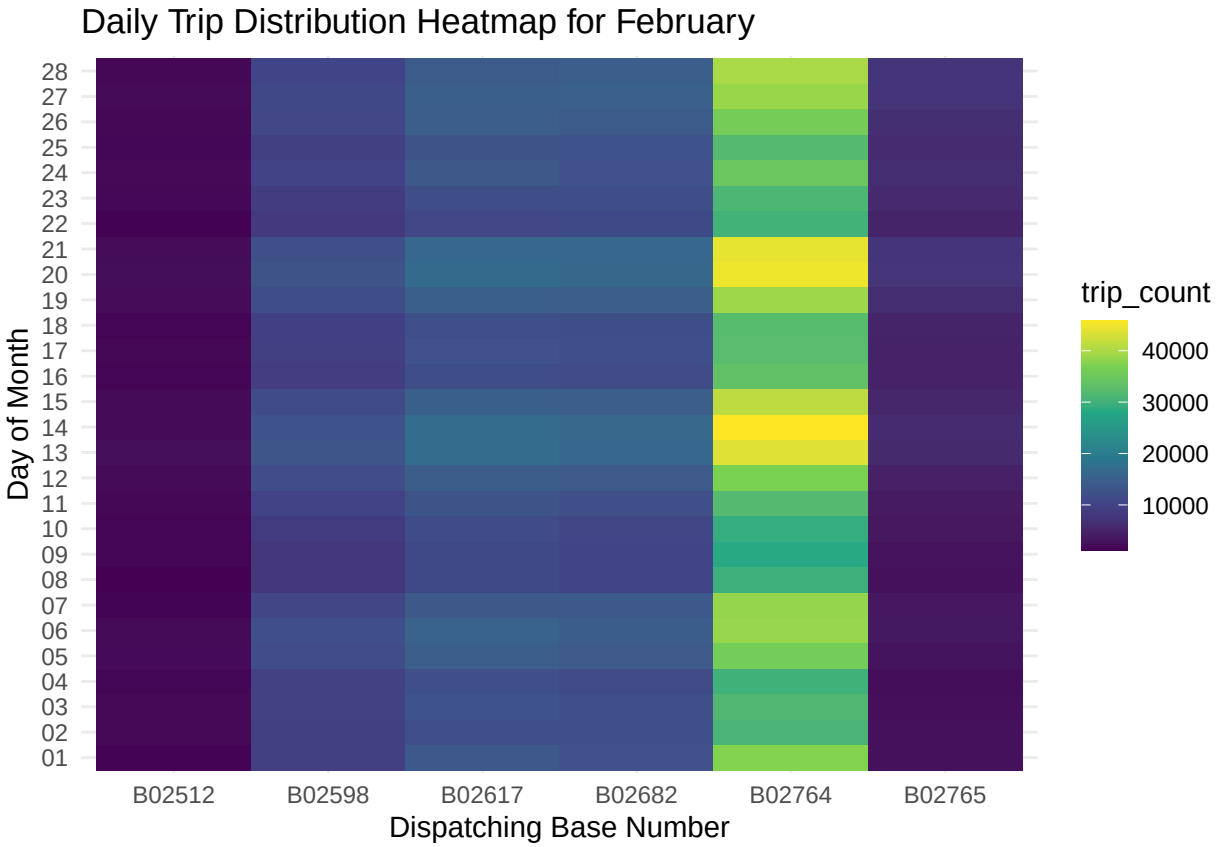
```
daily_trip_distribution_Feb <- uber %>%
  mutate(day = format(date, "%d")) %>%
  filter(format(date, "%m") %in% c("02")) %>%
  group_by(day, dispatching_base_number) %>%
  summarise(trip_count = sum(trips)) %>%
  ggplot(aes(x = dispatching_base_number, y = day, fill = trip_count)) +
  geom_tile() +
  labs(title = "Daily Trip Distribution Heatmap for February",
       x = "Dispatching Base Number",
       y = "Day of Month") +
  theme_minimal() +
  scale_fill_viridis_c()
```

```
## `summarise()` has grouped output by 'day'. You can override using the `.groups`
## argument.
```

```
daily_trip_distribution_Jan
```




```
daily_trip_distribution_Feb
```



- The heatmap visualization illustrated the distribution of daily trips across dispatching bases during the two months.
- Certain days with lighter color and dispatching base B02764 showed higher trip counts, indicating fluctuations in demand over time.

4. Comparison of Top Performing Dispatching Bases:

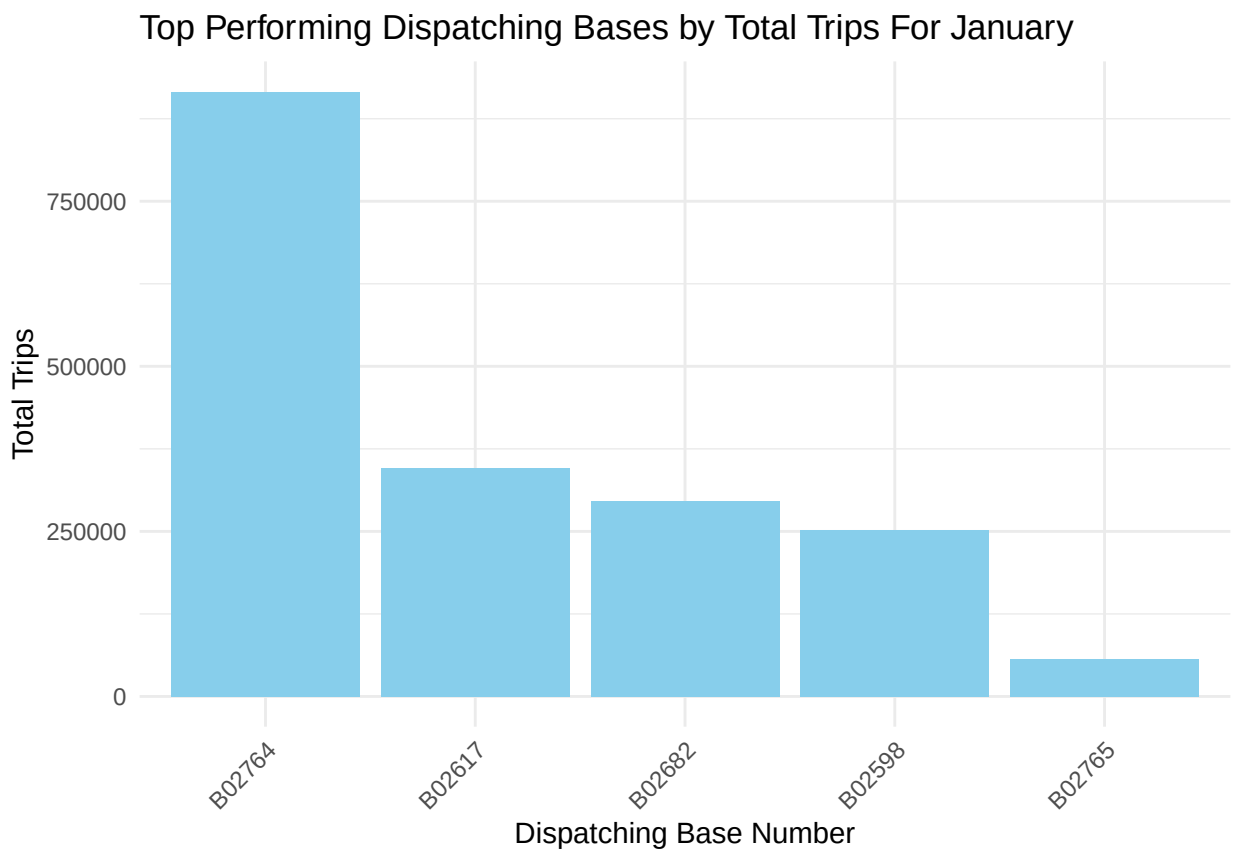
A bar chart or pie chart comparing dispatching base numbers based on total trips or average trip volume per active vehicle for January and February. This visualization will highlight any shifts in top-performing bases between the two months.

```
# Comparison of Top Performing Dispatching Bases
top_dispatching_bases_Jan <- uber %>%
  mutate(day = format(date, "%d")) %>%
  filter(format(date, "%m") %in% c("01")) %>%
  group_by(dispatching_base_number) %>%
  summarise(total_trips = sum(trips),
            avg_trip_per_vehicle = sum(trips) / sum(active_vehicles))

# Bar chart for total trips
top_dispatching_bases_total_Jan <- top_dispatching_bases_Jan %>%
  arrange(desc(total_trips)) %>%
  top_n(5) # Top 5 performing bases based on total trips
```

```
## Selecting by avg_trip_per_vehicle
```

```
ggplot(top_dispatching_bases_total_Jan, aes(x = reorder(dispatching_base_number,
                                                        -total_trips), y = total_trips)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top Performing Dispatching Bases by Total Trips For January",
       x = "Dispatching Base Number",
       y = "Total Trips") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



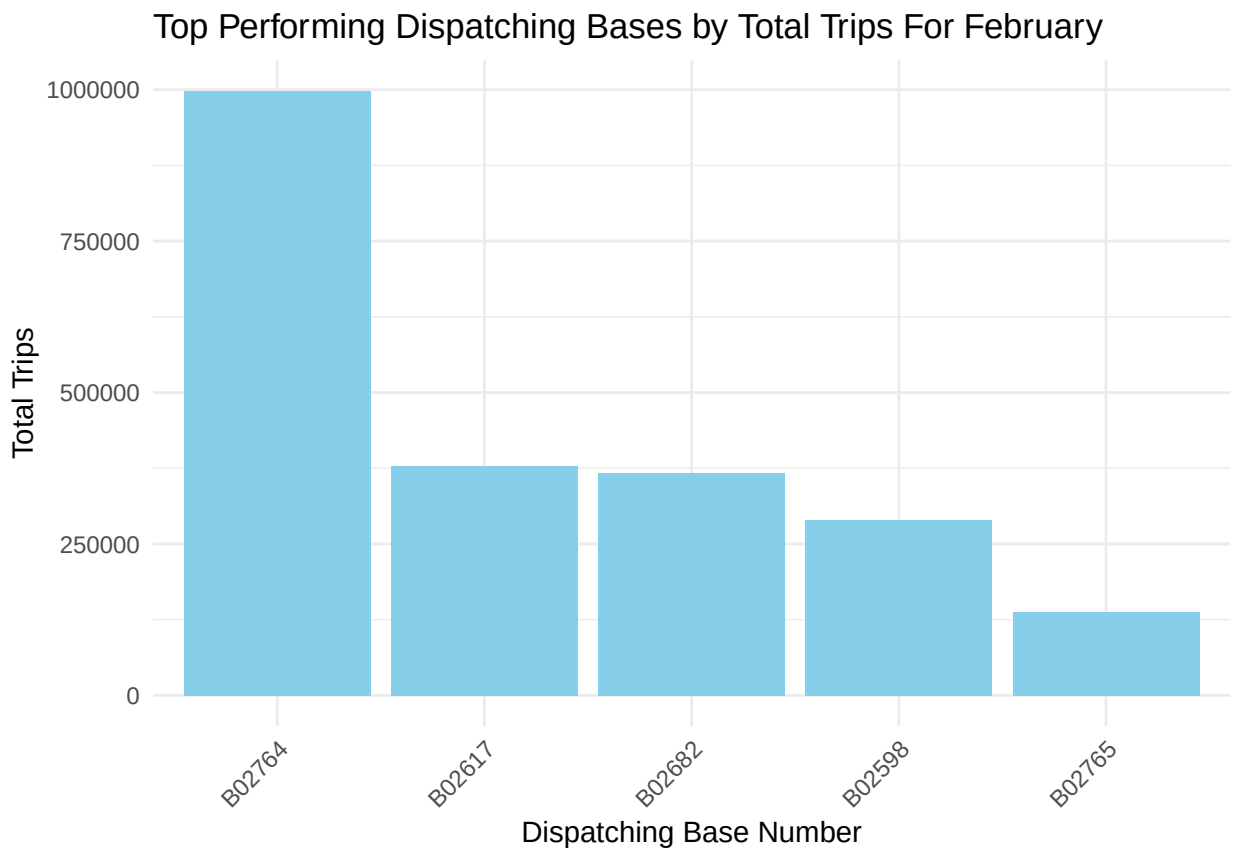

```
top_dispatching_bases_Feb <- uber %>%
  mutate(day = format(date, "%d")) %>%
  filter(format(date, "%m") %in% c("02")) %>%
  group_by(dispatching_base_number) %>%
  summarise(total_trips = sum(trips),
            avg_trip_per_vehicle = sum(trips) / sum(active_vehicles))

# Bar chart for total trips
top_dispatching_bases_total_Feb <- top_dispatching_bases_Feb %>%
  arrange(desc(total_trips)) %>%
  top_n(5) # Top 5 performing bases based on total trips
```

```
## Selecting by avg_trip_per_vehicle
```

```
ggplot(top_dispatching_bases_total_Feb, aes(x = reorder(dispatching_base_number,
                                                        -total_trips), y = total_trips)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top Performing Dispatching Bases by Total Trips For February",
```

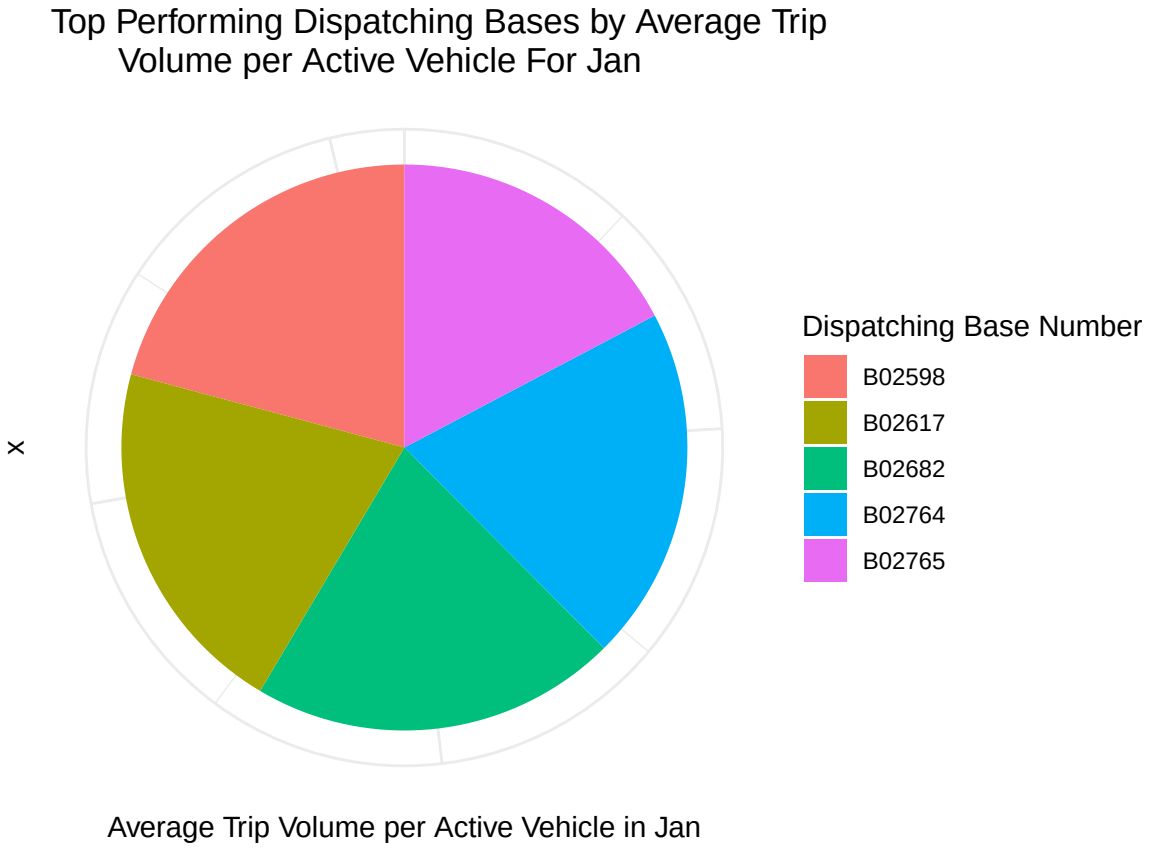
```
      x = "Dispatching Base Number",
      y = "Total Trips") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Pie chart for average trip volume per active vehicle
top_dispatching_bases_avg_Jan <- top_dispatching_bases_Jan %>%
  arrange(desc(avg_trip_per_vehicle)) %>%
  top_n(5) # Top 5 performing bases based on average trip volume per active vehicle
```

```
## Selecting by avg_trip_per_vehicle
```

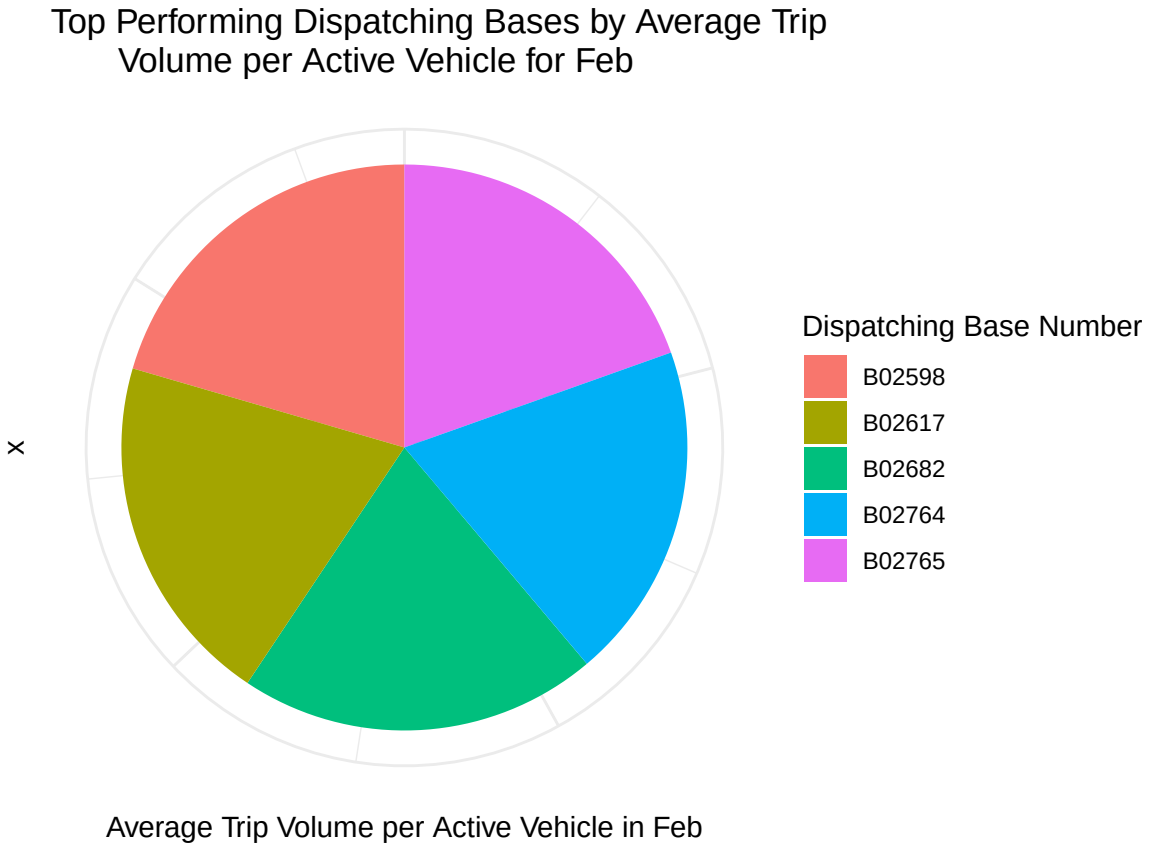
```
ggplot(top_dispatching_bases_avg_Jan, aes(x = "", y = avg_trip_per_vehicle,
                                          fill = dispatching_base_number)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Top Performing Dispatching Bases by Average Trip
       Volume per Active Vehicle For Jan",
       fill = "Dispatching Base Number",
       y = "Average Trip Volume per Active Vehicle in Jan") +
  theme_minimal() +
  theme(axis.text.x = element_blank())
```




```
# Pie chart for average trip volume per active vehicle
top_dispatching_bases_avg_Feb <- top_dispatching_bases_Feb %>%
  arrange(desc(avg_trip_per_vehicle)) %>%
  top_n(5) # Top 5 performing bases based on average trip volume per active vehicle
```

```
## Selecting by avg_trip_per_vehicle
```

```
ggplot(top_dispatching_bases_avg_Feb, aes(x = "", y = avg_trip_per_vehicle,
                                          fill = dispatching_base_number)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Top Performing Dispatching Bases by Average Trip
       Volume per Active Vehicle for Feb",
       fill = "Dispatching Base Number",
       y = "Average Trip Volume per Active Vehicle in Feb") +
  theme_minimal() +
  theme(axis.text.x = element_blank())
```




```
top_dispatching_bases_avg_Jan
```

```
## # A tibble: 5 x 3
##   dispatching_base_number total_trips avg_trip_per_vehicle
##   <chr>                         <int>                <dbl>
## 1 B02682                       295941                 8.72
## 2 B02598                       251658                 8.66
## 3 B02617                       345988                 8.59
## 4 B02764                       915976                 8.42
## 5 B02765                        56287                 7.18
```

```
top_dispatching_bases_avg_Feb
```

```
## # A tibble: 5 x 3
##   dispatching_base_number total_trips avg_trip_per_vehicle
##   <chr>                         <int>                <dbl>
## 1 B02682                       366568                 9.78
## 2 B02598                       289133                 9.77
## 3 B02617                       379037                 9.61
## 4 B02765                       137383                 9.32
## 5 B02764                       998473                 9.20
```

Clearly February has higher total and average trip volumes in compared to January, specially for B02682 and B02764.

Conclusion:

Through these visualizations, the story will provide insights into how seasonal variations impact the operation and utilization of ride-for-hire services in New York City, informing discussions on resource allocation, service planning, and adaptation strategies to seasonal demand fluctuations. February has a higher rate of deployment with respect to January. In conclusion, this report provides a comprehensive analysis of Uber's performance in New York City based on the data from January and February 2015.

Future Plan:

We further plan to dig deeper to track the evolution of Uber's performance over time and assess its impact on the transportation landscape of New York City. More data on other months might be helpful in our study. If we can get the hourly data we can also work on the visualization of the peak hours.

References:

- Uber-Jan-Feb-FOIL dataset
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.