Uber Pickup Data Analysis

Debalina Chakraborty & Samantha Lobo

4/18/2022

Motivation & Overview

With the increasing amount of car share services, we have seen that there have been a lot of customer dissatisfaction leading to poor company reviews and customer experiences. Our objective is to draw insights from the historical Uber Pickup data available and suggest solutions which will further enhance customer experience. This is a predominantly data visualization project using the ggplot2 library for understanding the data and for developing a story to further understand the customers who avail the trips in New York.

Related work

Previous work has been done to predict customer behaviour based on Uber pickups & determining the rush hours. We have used the base idea of this project and have performed exploratory data analysis based on visualizations only. a(Https://www.analyticsvidhya.com/blog/2021/10/end-to-end-predictive-analysis-on-ubers-data/ (Https://www.analyticsvidhya.com/blog/2021/10/end-to-end-predictive-analysis-on-ubers-data/))

** Initial questions**

The questions we try to answer in this project are: 1. How do passengers fare throughout the day? 2. Which part of the day do we have more number of trips? 3. Which day, month contributed to the highest trips in the year? 4. How passengers made trips from different bases?

> Through the course of this project, we evolve into questions such as: How has time affected the customer trips?

R Markdown

```
###### Loading the libraries
library(ggplot2)
```

```
## Warning in register(): Can't find generic `scale_type` in package ggplot2 to
## register S3 method.
```

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggmap':
##
##     wind
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.3
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(DT)
```

```
## Warning: package 'DT' was built under R version 4.1.3
```

```
library(scales)
library(leaflet)
```

#**Data** ###### Collecting the data : The data was collected from Kaggle competition a(‘https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city (https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city)’). We are loading the data sets which are individual files of pickup data from april 2014 to September 2014 and combine them to create a data set for the whole year of 2014.

```
###Create a vector of our colors that will be included in our plotting functions
colors = c("#E6E6FA", "#D8BFD8", "#DDA0DD", "#EE82EE", "#DA70D6", "#BA55D3", "#8A2BE2")
getwd()
```

```
## [1] "C:/Users/saman/OneDrive/Desktop/Adv R/archive (2)"
```

```
###### Load the data & data wrangling, storing them in dataframes to be used later.
apr_data <- read.csv("uber-raw-data-apr14.csv")
may_data <- read.csv("uber-raw-data-may14.csv")
jun_data <- read.csv("uber-raw-data-jun14.csv")
jul_data <- read.csv("uber-raw-data-jul14.csv")
aug_data <- read.csv("uber-raw-data-aug14.csv")
sep_data <- read.csv("uber-raw-data-sep14.csv")
data_2015 <- read.csv("uber-raw-data-janjune-15.csv")
data_2014 <- rbind(apr_data,may_data, jun_data, jul_data, aug_data, sep_data)

data_2014$Date.Time <- as.POSIXct(data_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S")

data_2014$Time <- format(as.POSIXct(data_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format= "%H:%M:%S")

data_2014$Date.Time <- ymd_hms(data_2014$Date.Time)

data_2014$day <- factor(day(data_2014$Date.Time))
data_2014$month <- factor(month(data_2014$Date.Time, label = TRUE))
data_2014$year <- factor(year(data_2014$Date.Time))
data_2014$dayofweek <- factor(wday(data_2014$Date.Time, label = TRUE))

data_2014$hour <- factor(hour(hms(data_2014$Time)))
data_2014$minute <- factor(minute(hms(data_2014$Time)))
data_2014$second <- factor(second(hms(data_2014$Time)))
data_2014$Base_names <- if_else(data_2014$Base == "B02512", "Unter",
                                if_else(data_2014$Base == "B02598", "Hinter",
                                if_else(data_2014$Base == "B02682", "Schmecken",
                                if_else(data_2014$Base == "B02764", "Danach-NY",
                                if_else(data_2014$Base == "B02617", "Weiter",
                                if_else(data_2014$Base == "B02765", "Grun",
                                if_else(data_2014$Base == "B02835", "Dreist",
                                if_else(data_2014$Base == "B02836", "Drinnen",data_2014$Base))))))))
```

#Exploratory Data Analysis

Below graph plotting highlights the trips by the hours in a day in the year 2014

```
 hour_data <- data_2014 %>%
  group_by(hour) %>%
  dplyr::summarize(Total = n())
datatable(hour_data)
```

Show 10 ⌄ **entries** **Search:** ▭

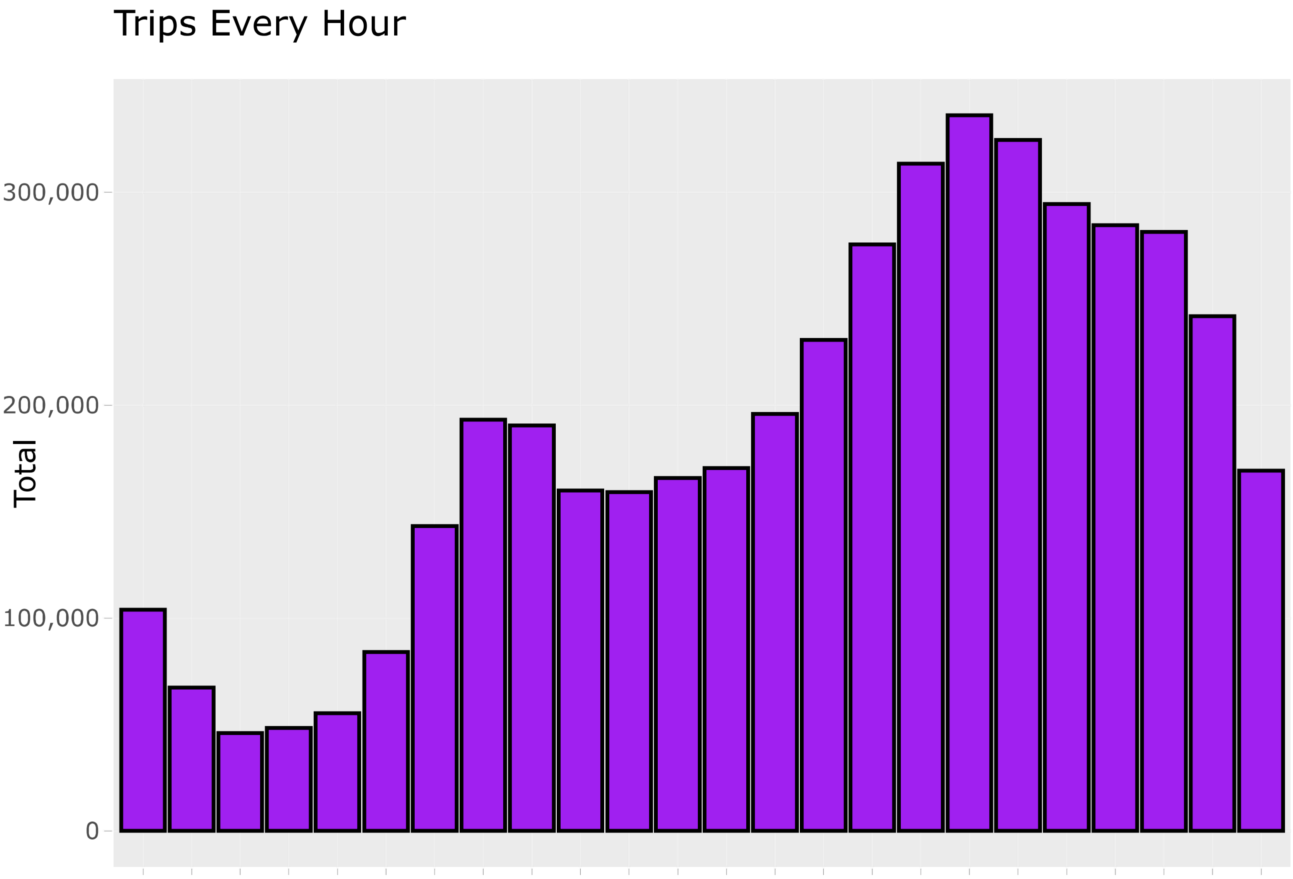
	hour	Total
1	0	103836
2	1	67227

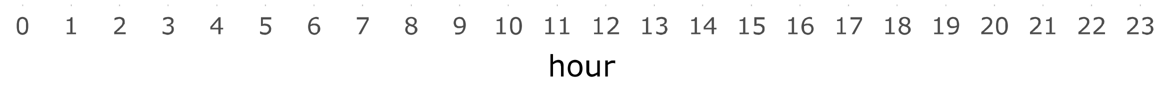
	hour	Total
3	2	45865
4	3	48287
5	4	55230
6	5	83939
7	6	143213
8	7	193094
9	8	190504
10	9	159967

Showing 1 to 10 of 24 entries

Previous 1 2 3 Next

```
Trip_p_hr<- ggplot(hour_data, aes(hour, Total)) +
  geom_bar( stat = "identity", fill = "purple", color = "black") +
  ggtitle("Trips Every Hour") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = comma)
ggplotly(Trip_p_hr)
```



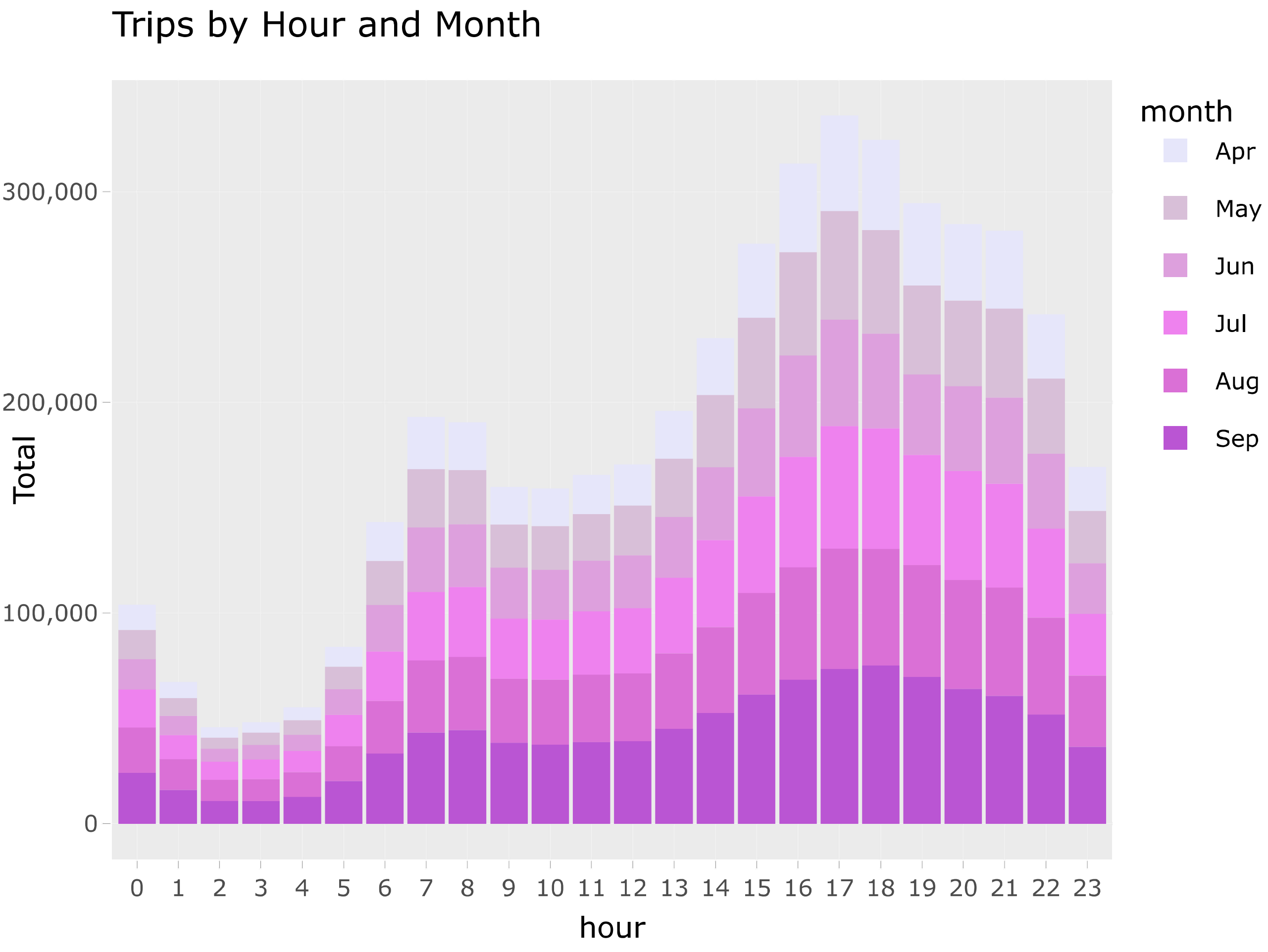


Below graph plotting highlights the trips by the hours for every month in the year 2014

```
month_hour <- data_2014 %>%
  group_by(month, hour) %>%
  dplyr::summarize(Total = n())
```

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

```
Trip_p_hr_month <- ggplot(month_hour, aes(hour, Total, fill = month)) +
  geom_bar( stat = "identity") +
  ggtitle("Trips by Hour and Month") +
  scale_y_continuous(labels = comma)+
  scale_fill_manual(values = colors)
ggplotly(Trip_p_hr_month)
```



This shows the trip every day in a month. We see a general trend that over 15,000 trips were covered on an average between 11-16th of a month & again a small spike towards the end of the month.

```
day_group <- data_2014 %>%
  group_by(day) %>%
  dplyr::summarize(Total = n())
datatable(day_group)
```

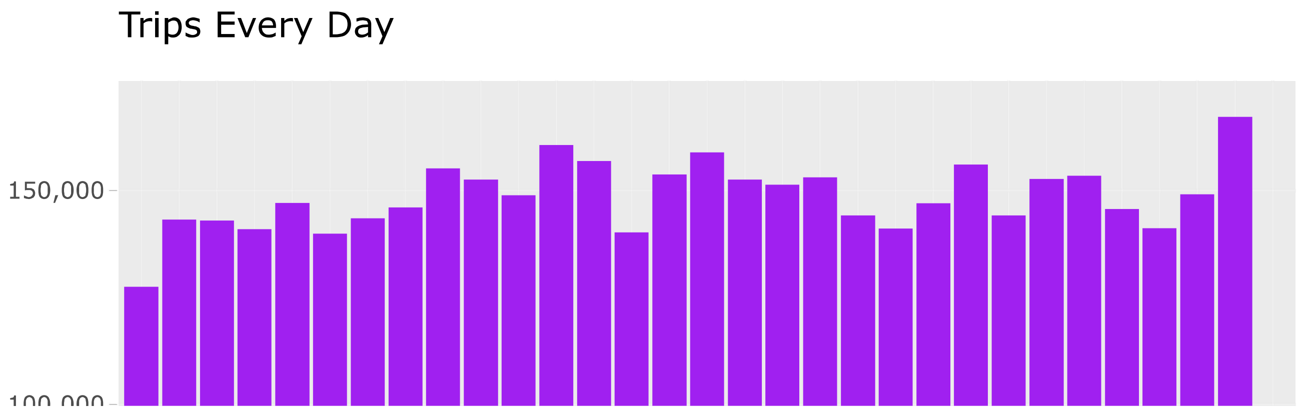
Show 10 entries Search:

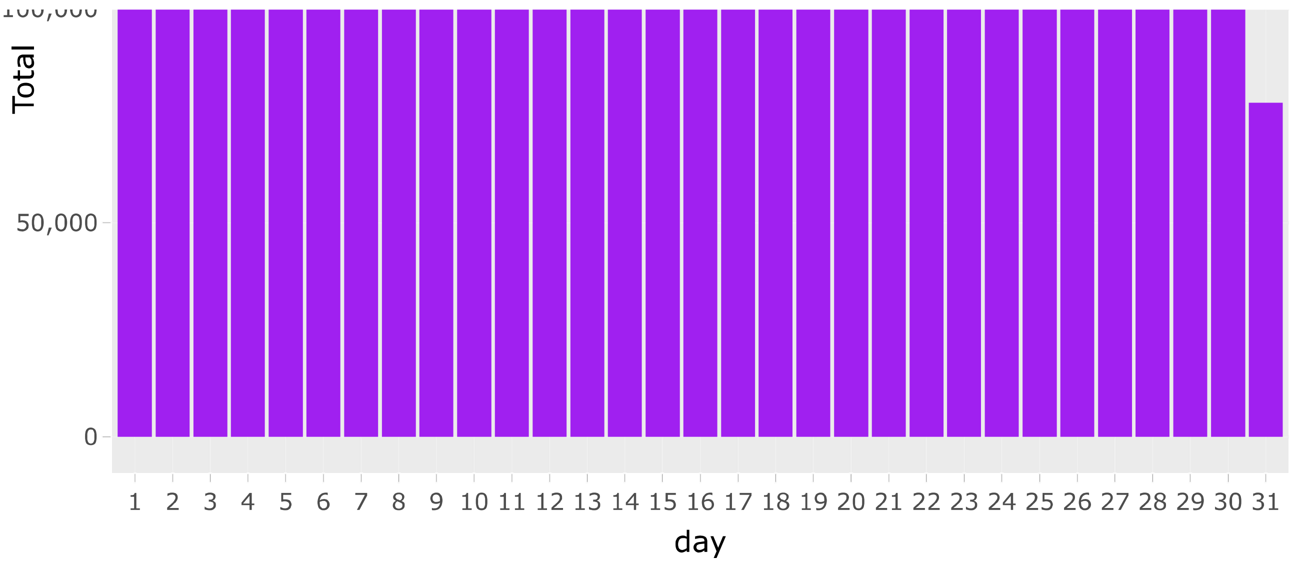
	day	Total
1	1	127430
2	2	143201
3	3	142983
4	4	140923
5	5	147054
6	6	139886
7	7	143503
8	8	145984
9	9	155135
10	10	152500

Showing 1 to 10 of 31 entries

Previous 1 2 3 4 Next

```
Trip_p_day <- ggplot(day_group, aes(day, Total)) +
  geom_bar( stat = "identity", fill = "purple") +
  ggtitle("Trips Every Day") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = comma)+
  scale_fill_manual(values = colors)
ggplotly(Trip_p_day)
```



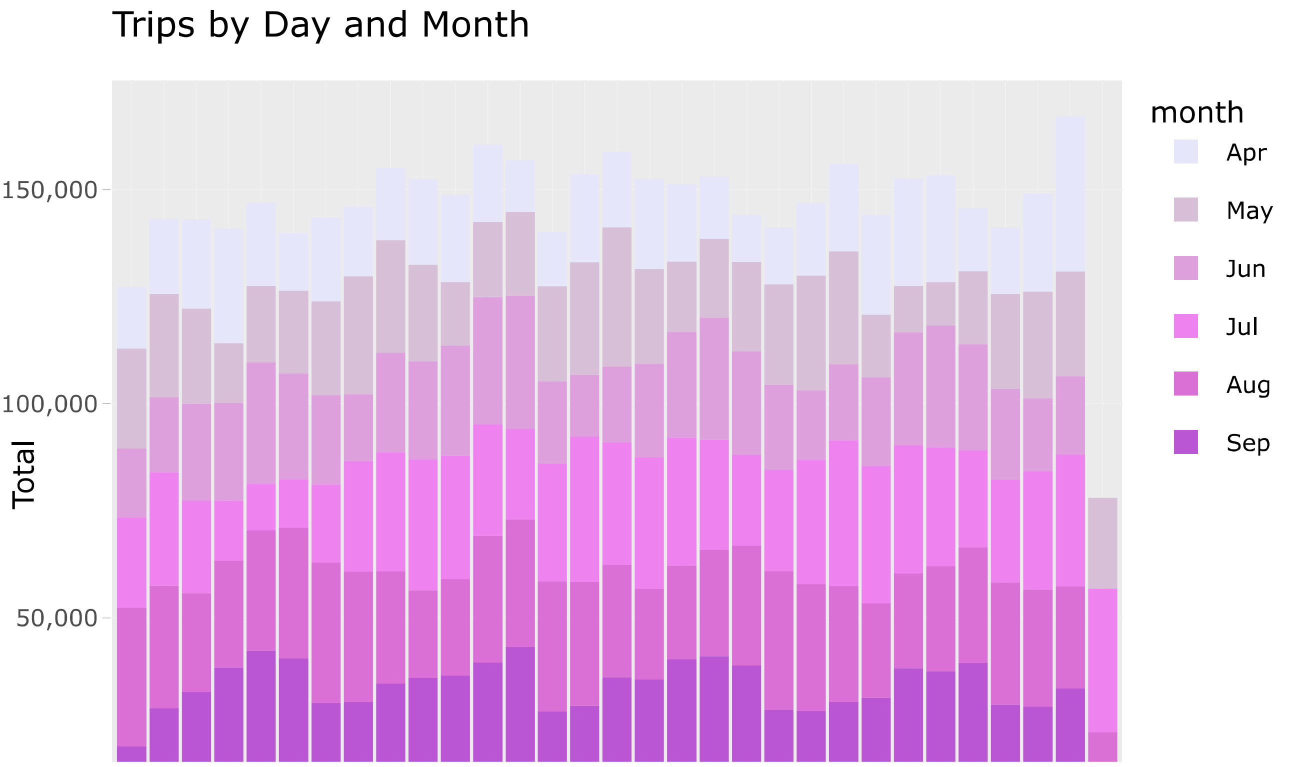


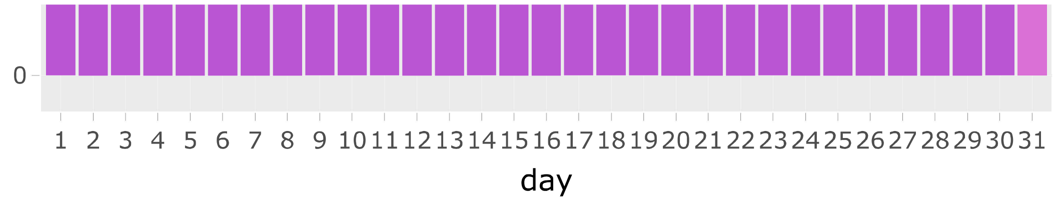
Let’s see how the above daily trips spread across the months, to get a better idea of when we see a peak in Uber pickups.

```
day_month_group <- data_2014 %>%
  group_by(month, day) %>%
  dplyr::summarize(Total = n())
```

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

```
Trip_p_day_month <- ggplot(day_month_group, aes(day, Total, fill = month)) +
  geom_bar( stat = "identity") +
  ggtitle("Trips by Day and Month") +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = colors)
ggplotly(Trip_p_day_month)
```





Here we can see that the highest uber pickups were between July-Sept. Considering that being the tourist season in NYC the numbers are justified.

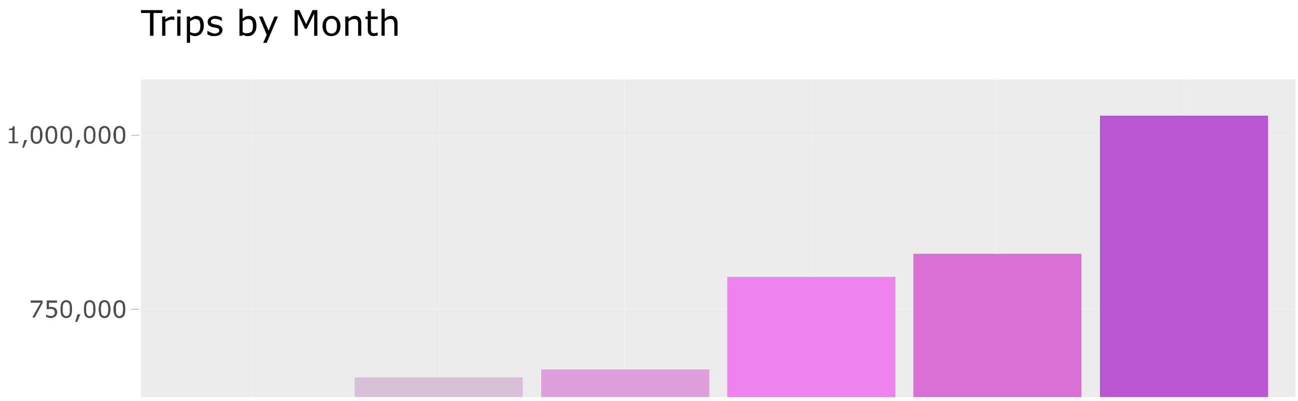
```
month_group <- data_2014 %>%
  group_by(month) %>%
  dplyr::summarize(Total = n())
datatable(month_group)
```

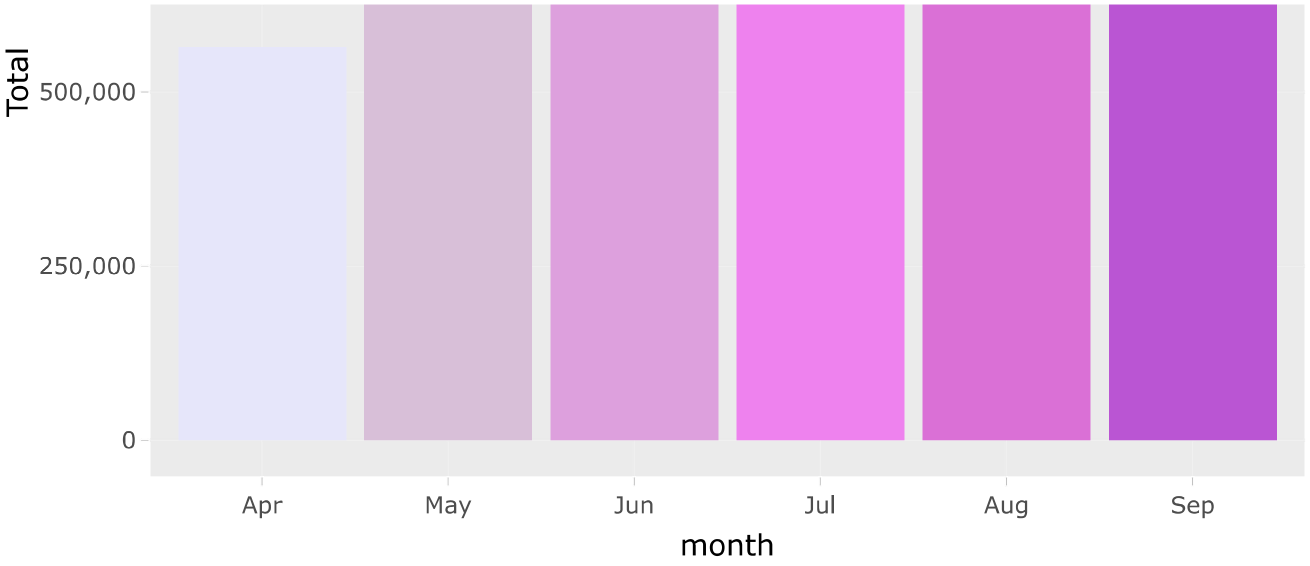
Show 10 **entries** **Search:**

	month	Total
1	Apr	564516
2	May	652435
3	Jun	663844
4	Jul	796121
5	Aug	829275
6	Sep	1028136

Showing 1 to 6 of 6 entries Previous 1 Next

```
Trips_p_month <- ggplot( month_group, aes(month, Total, fill = month)) +
  geom_bar( stat = "identity") +
  ggtitle("Trips by Month") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = colors)
ggplotly(Trips_p_month)
```



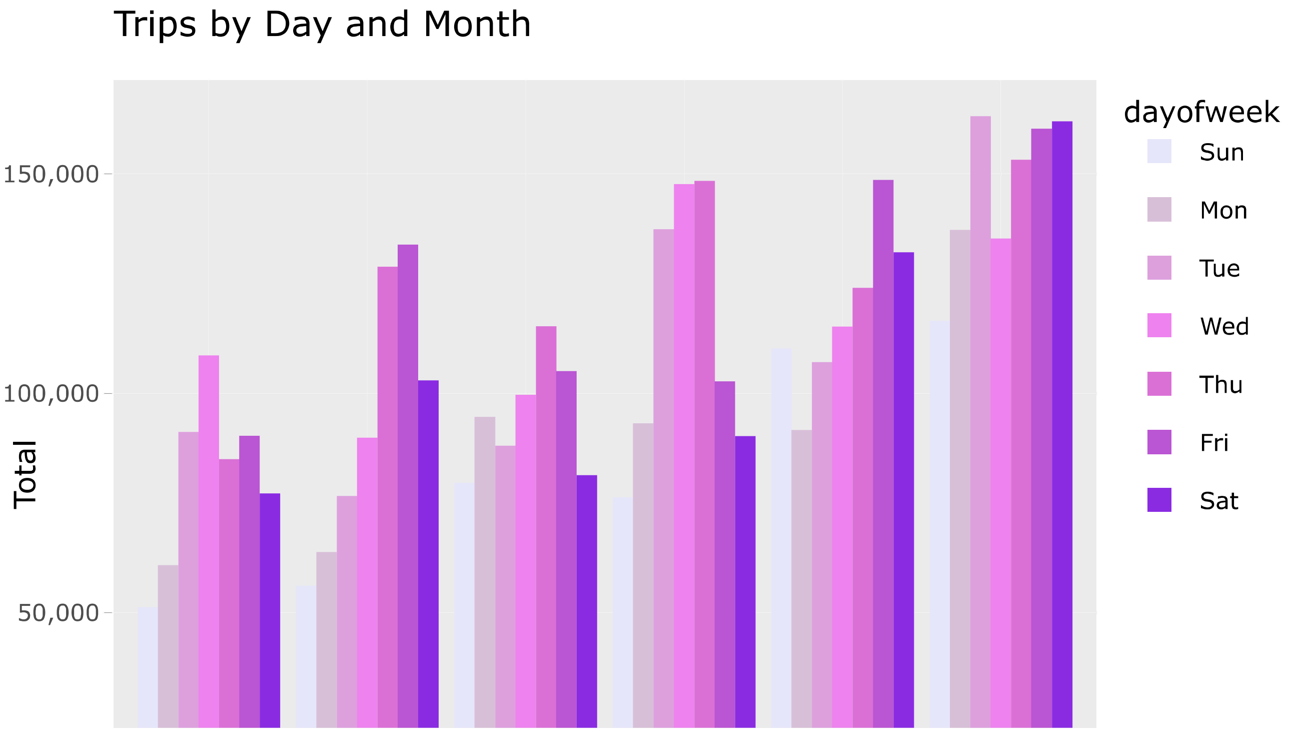


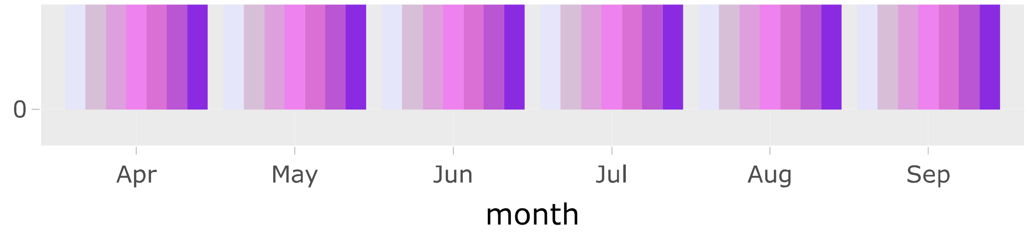
As we can see, from the below graph, the peak pickup between April - June is during the weekdays, but as we move towards July-September we can see that the highest pickups were during the weekends (Friday-Sat).

```
month_weekday <- data_2014 %>%
  group_by(month, dayofweek) %>%
  dplyr::summarize(Total = n())
```

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

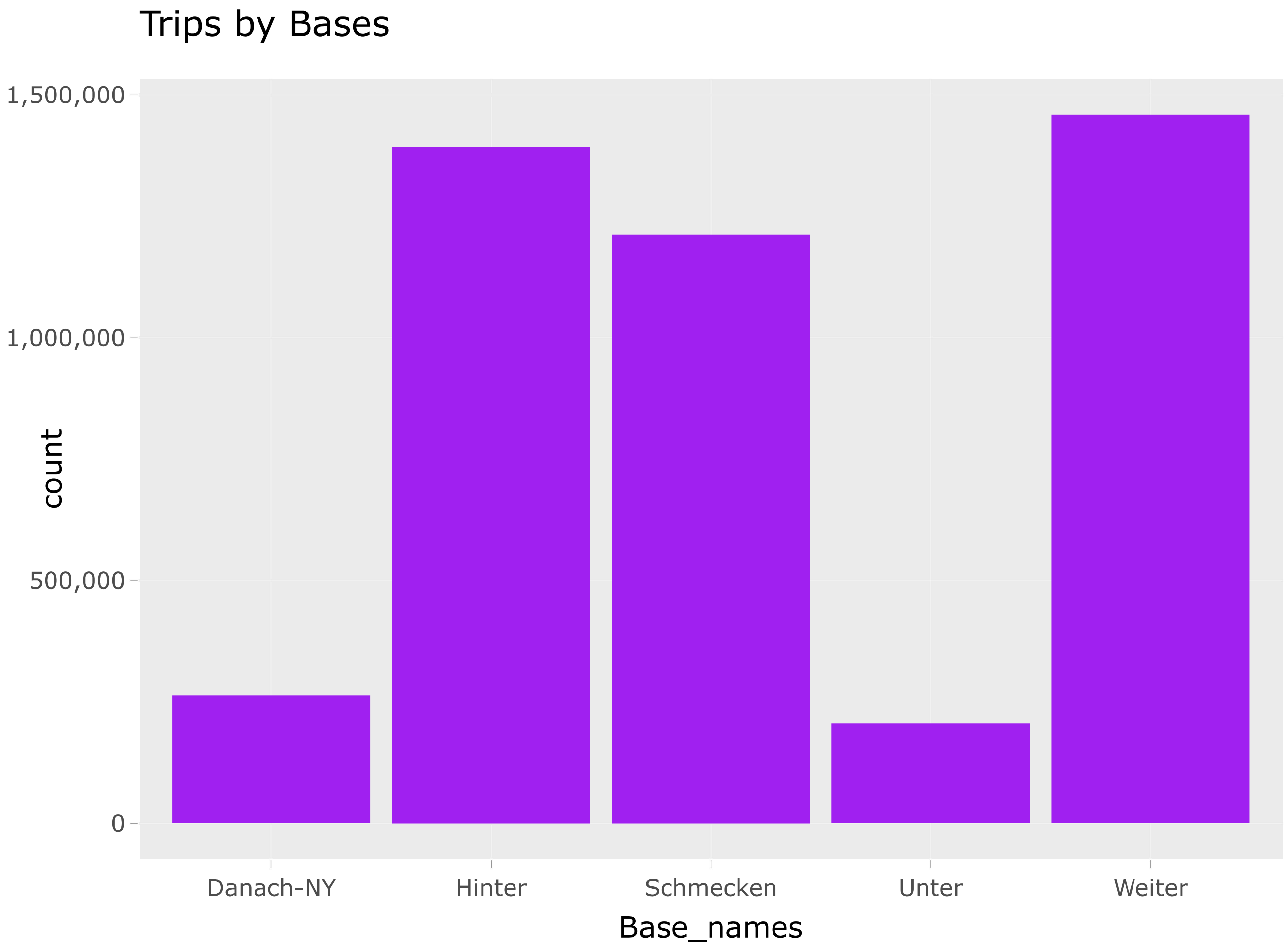
```
Trips_p_day_month <- ggplot(month_weekday, aes(month, Total, fill = dayofweek)) +
  geom_bar( stat = "identity", position = "dodge") +
  ggtitle("Trips by Day and Month") +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = colors)
ggplotly(Trips_p_day_month)
```





We are now analysing Uber pickup by Base locations, most of the pickups happen from NYC bases like Hinter, Weiter & Schmecken

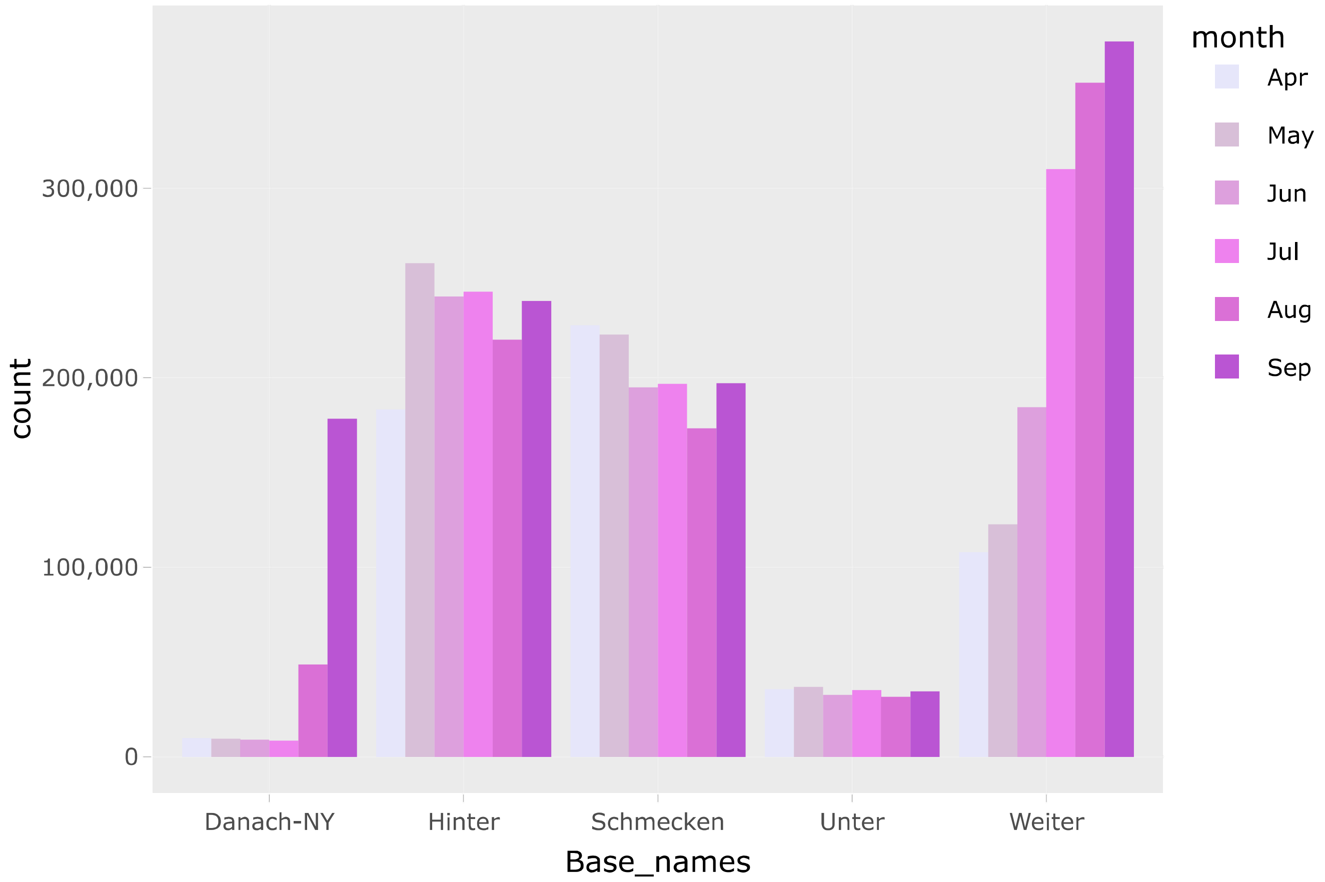
```
Trip_By_Base <- ggplot(data_2014, aes(Base_names)) +
  geom_bar(fill = "purple") +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by Bases")
ggplotly(Trip_By_Base)
```



Below graph highlights the picups made across these Uber bases in every month. All across the bases the number of trips increased between July & September. Except in Hinter & Schmecken where the count of trips were higher in Aril & May.

```
Trip_By_Base_Month <- ggplot(data_2014, aes(Base_names, fill = month)) +
  geom_bar(position = "dodge") +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by Bases and Month") +
  scale_fill_manual(values = colors)
ggplotly(Trip_By_Base_Month)
```

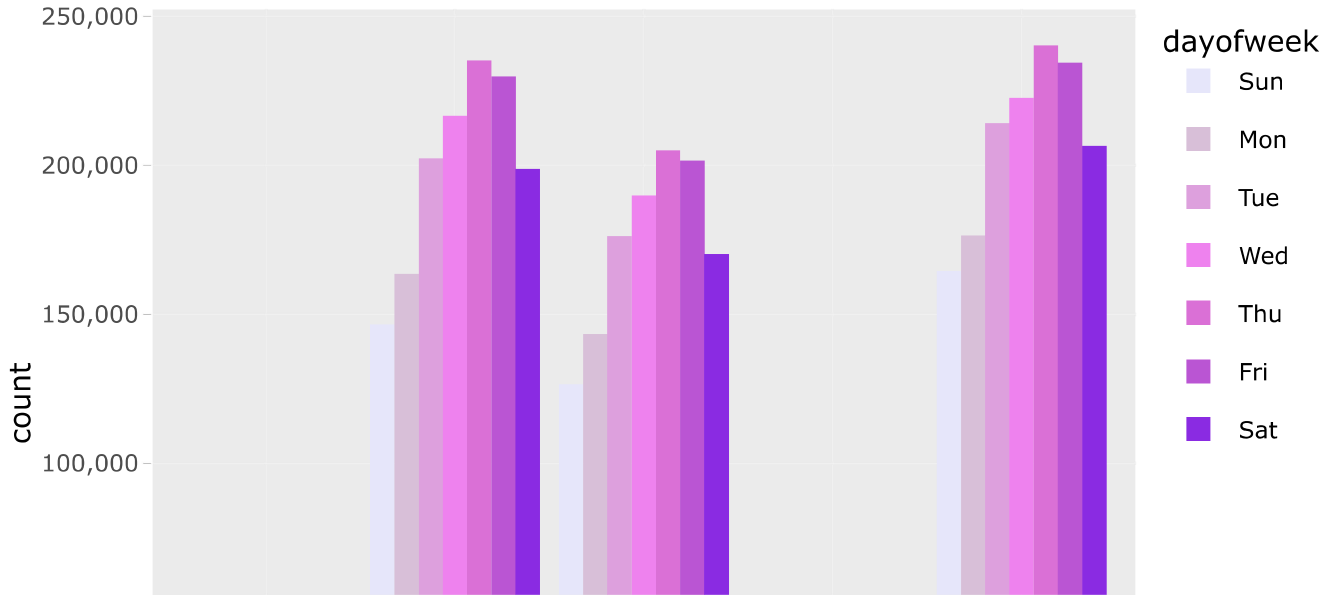
Trips by Bases and Month

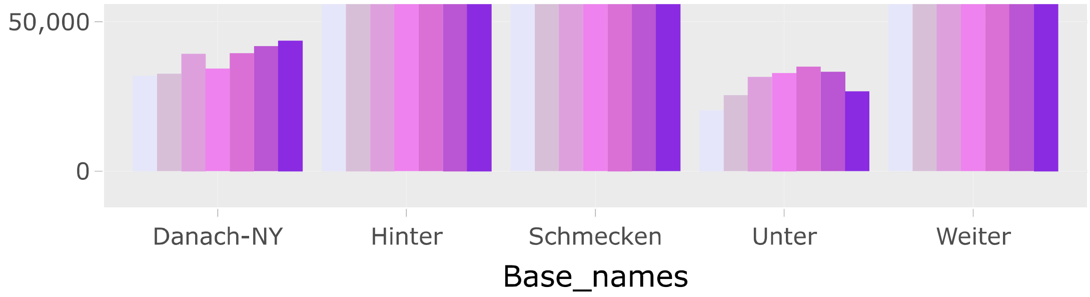


We are now analysing Uber pickup by Base locations by week, most of the trips are during weekdays.

```
Trip_ByWeek_Base <- ggplot(data_2014, aes(Base_names, fill = dayofweek)) +
  geom_bar(position = "dodge") +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by Bases and DayofWeek") +
  scale_fill_manual(values = colors)
ggplotly(Trip_ByWeek_Base)
```

Trips by Bases and DayofWeek





Creating a Heatmap visualization of day, hour and month.

a) The first graph helps us understand that almost everyday 6-8 am or the first 6-8th hour is the peak pickup time for uber rides & similarly the next rush hour is between 4-8 PM that is between 16th & 20th hour of the day.

```
day_and_hour <- data_2014 %>%
  group_by(day, hour) %>%
  dplyr::summarize(Total = n())
```

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

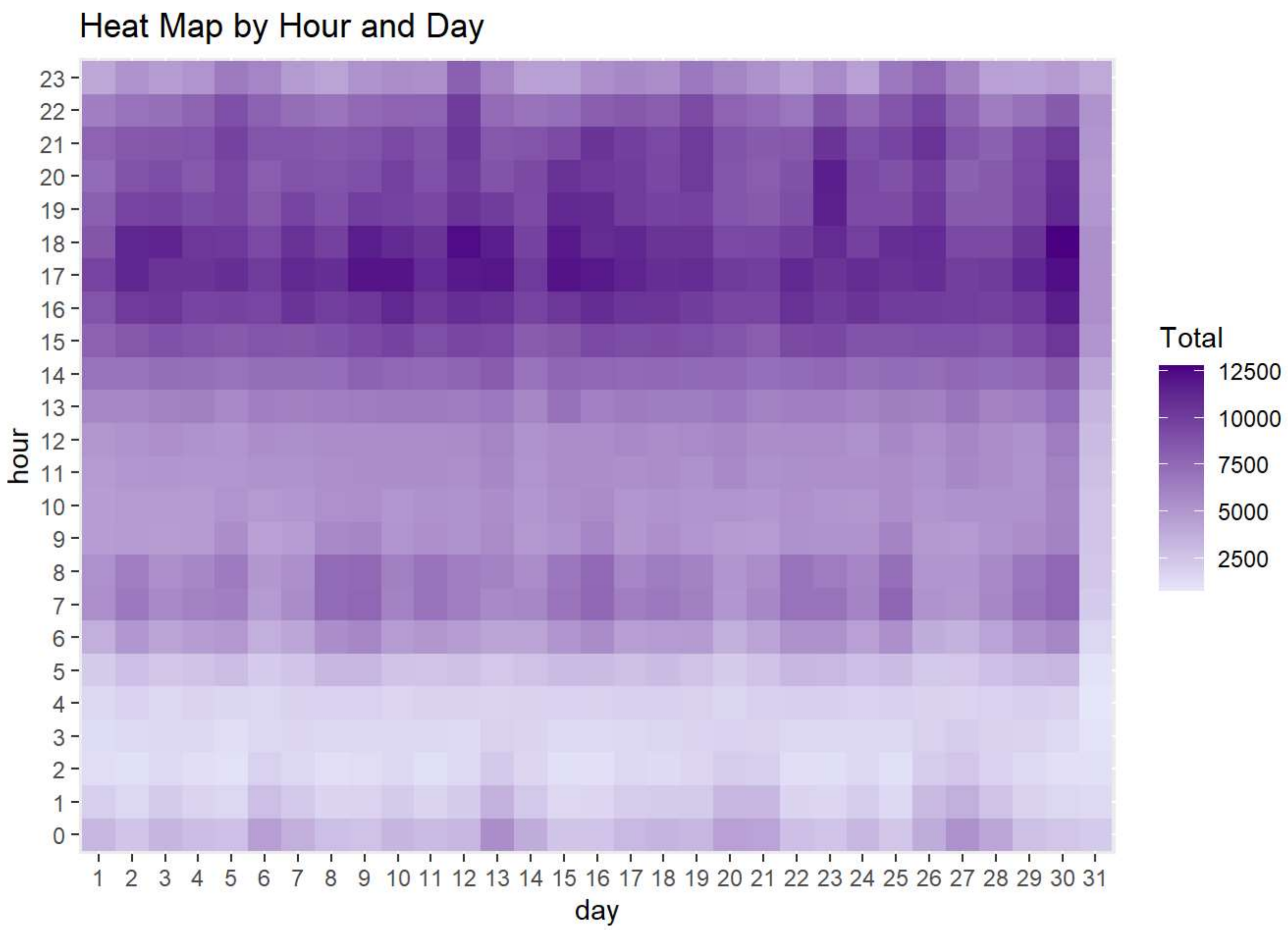
```
datatable(day_and_hour)
```

Show 10 **entries** **Search:**

	day	hour	Total
1	1	0	3247
2	1	1	1982
3	1	2	1284
4	1	3	1331
5	1	4	1458
6	1	5	2171
7	1	6	3717
8	1	7	5470
9	1	8	5376
10	1	9	4688

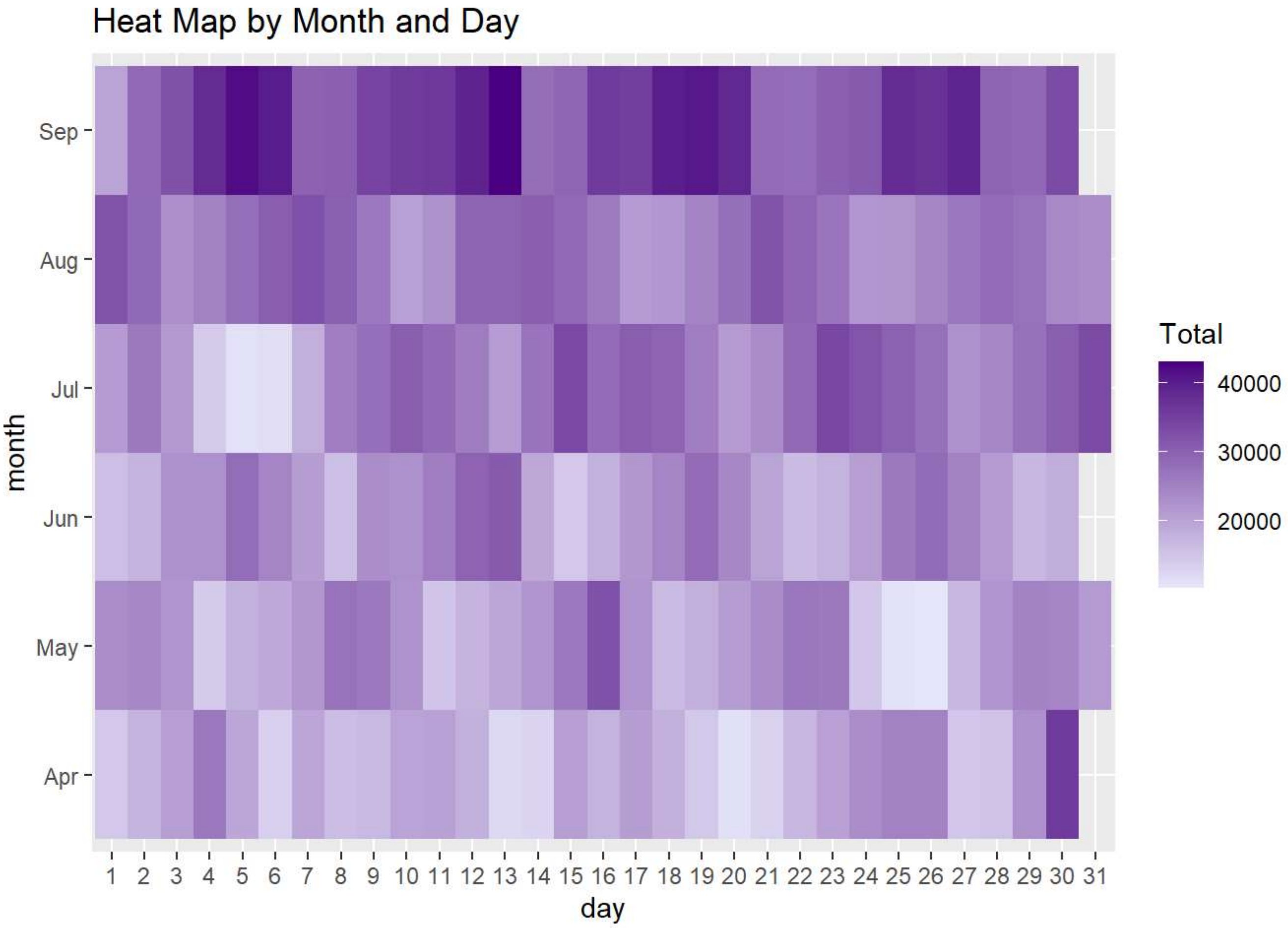
Showing 1 to 10 of 744 entries Previous 1 2 3 4 5 … 75 Next

```
ggplot(day_and_hour, aes(day, hour, fill = Total)) +
  geom_tile() +
  scale_fill_gradient(low = "#E6E6FA", high= "#4B0082", guide = "colourbar")+
  ggtitle("Heat Map by Hour and Day")
```



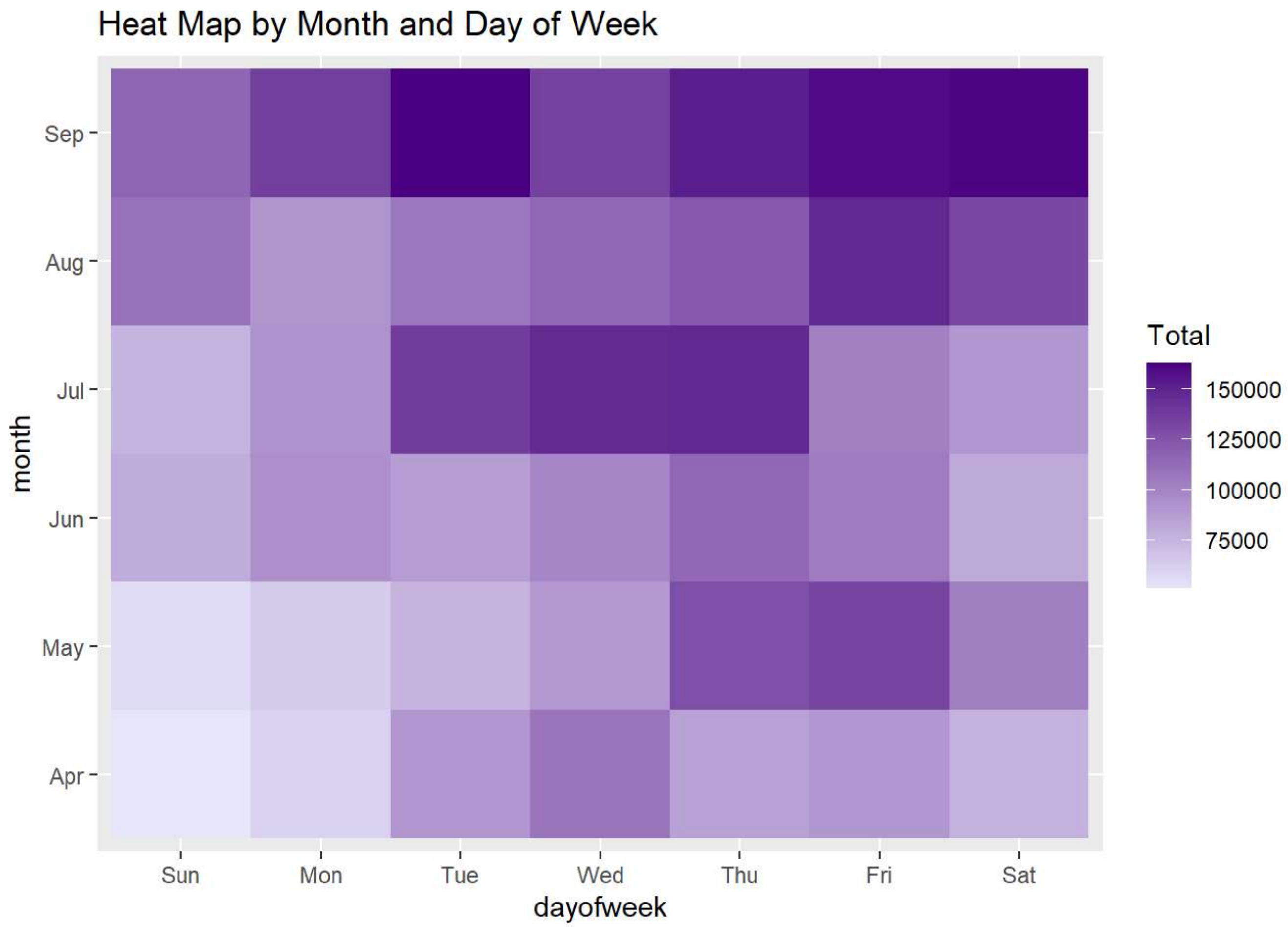
b) The second graphs shows the peak uber rides happen in the Month of Aug-Sep.

```
ggplot(day_month_group, aes(day, month, fill = Total)) +
  geom_tile() +
  scale_fill_gradient(low = "#E6E6FA", high= "#4B0082", guide = "colourbar")+
  ggtitle("Heat Map by Month and Day")
```



c) The third graphs shows the pickup ride distribution by month & day of the week, where we see a clear distinguishment between the trend till July where max pickups are in Weekdays & then during Septemeber it shifts towards the weekend.

```
ggplot(month_weekday, aes(dayofweek, month, fill = Total)) +
  geom_tile() +
  scale_fill_gradient(low = "#E6E6FA", high= "#4B0082", guide = "colourbar")+
  ggtitle("Heat Map by Month and Day of Week")
```



Further diving deeper we see the trend between the pickups in a month by Base locations

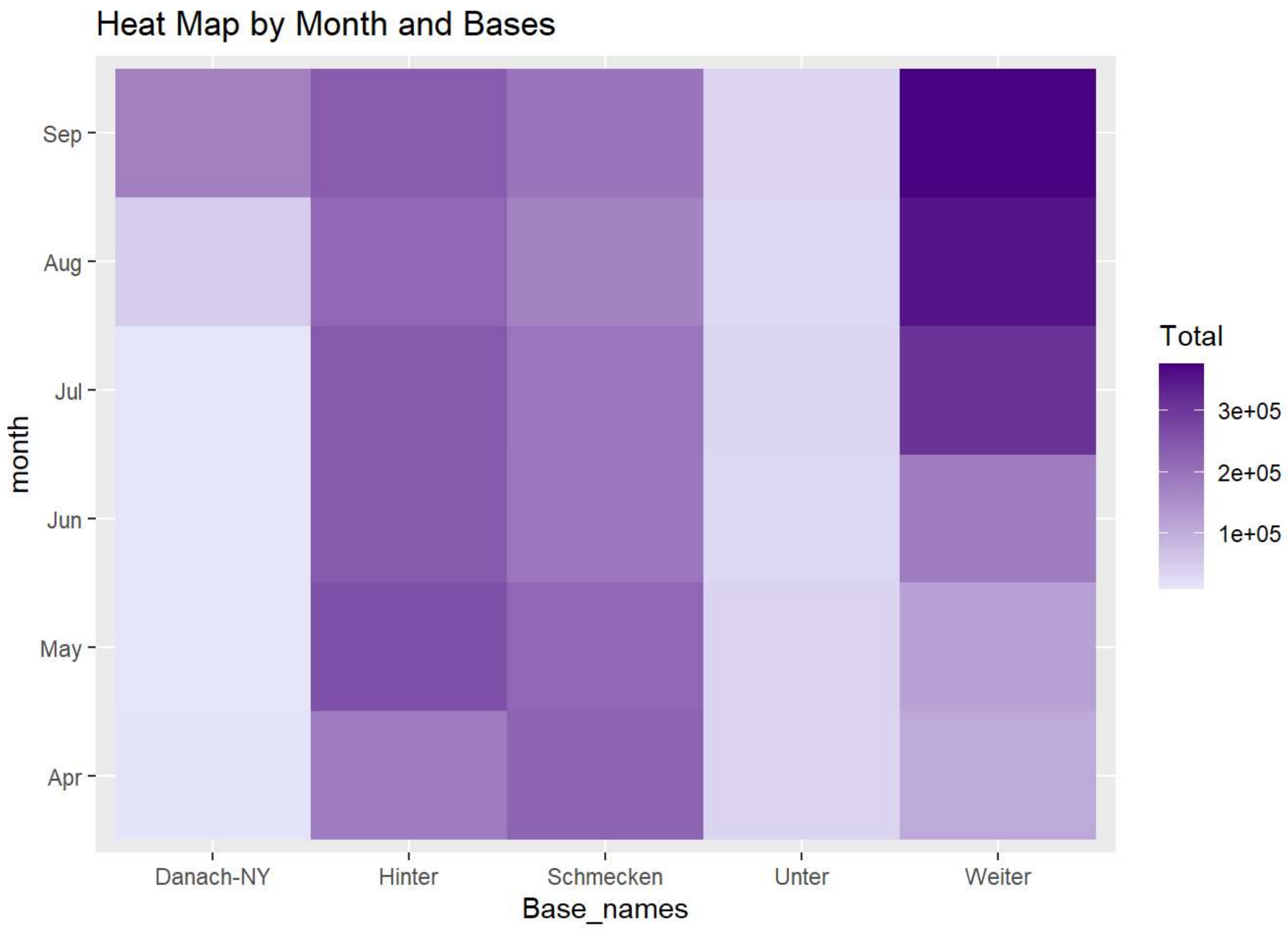
```
month_base <-  data_2014 %>%
  group_by(Base_names, month) %>%
  dplyr::summarize(Total = n())
```

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

```
day0fweek_bases <-  data_2014 %>%
  group_by(Base_names, dayofweek) %>%
  dplyr::summarize(Total = n())
```

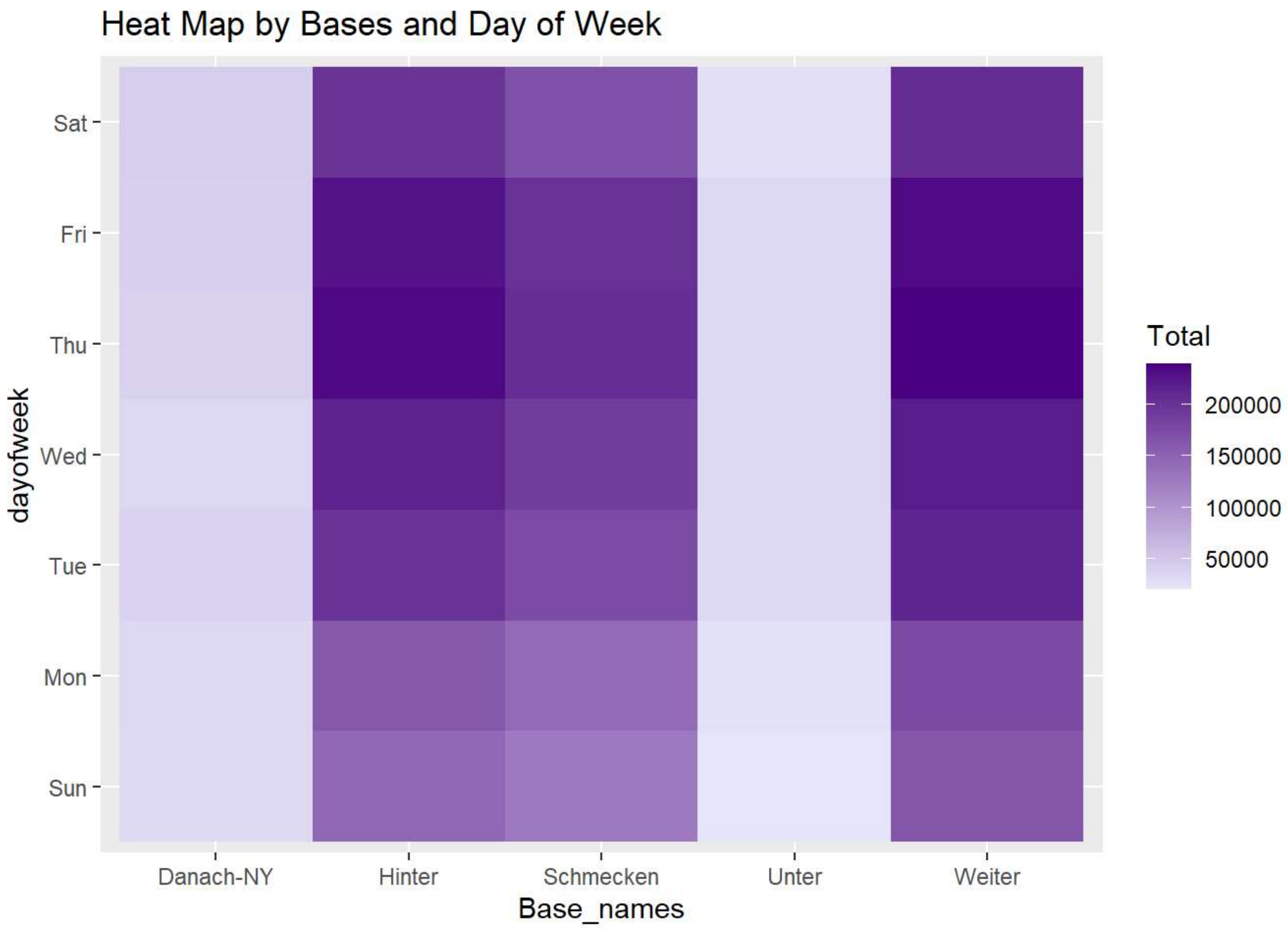
```
## `summarise()` has grouped output by 'Base_names'. You can override using the `.groups` argume
nt.
```

```
ggplot(month_base, aes(Base_names, month, fill = Total)) +
 geom_tile() +
  scale_fill_gradient(low = "#E6E6FA", high= "#4B0082", guide = "colourbar")+
   ggtitle("Heat Map by Month and Bases")
```



And Pickup from a base in every day of a week.

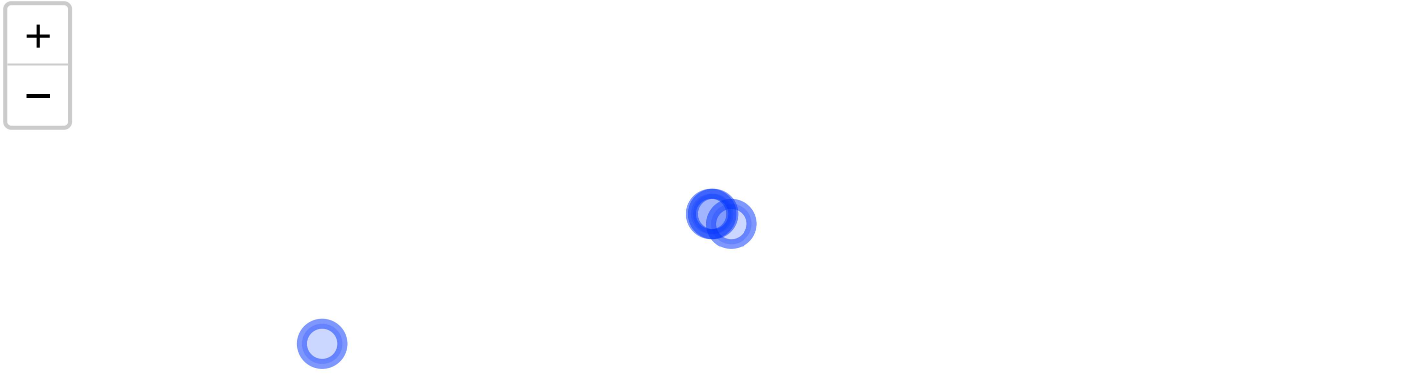
```
ggplot(day0fweek_bases, aes(Base_names, dayofweek, fill = Total)) +
  geom_tile() +
  scale_fill_gradient(low = "#E6E6FA", high= "#4B0082", guide = "colourbar")+
  ggtitle("Heat Map by Bases and Day of Week")
```



Analysing the Top Pickup locations in NYC - We can see from the graphh that the top pickup location are JFK International airport, LaGuardia Airport, TWA Hotels near JFK, LaGuardia delta Airlines Terminal, 9/11 Museum & LaGuardia Southwest airline terminal.

```
uber_2014_top <- data_2014 %>%
  mutate(Lat_3 = round(Lat, 3),Lon_3 = round(Lon, 3)) %>%
  count(Lat_3, Lon_3, sort = TRUE) %>%
  head()

uber_2014_top_map <- leaflet(uber_2014_top)
uber_2014_top_map <- addTiles(uber_2014_top_map)
uber_2014_top_map <- addCircleMarkers(uber_2014_top_map, lng = ~Lon_3,
                                        lat = ~Lat_3)
uber_2014_top_map
```



Leaflet (https://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

#Narrative & summary

1. Maximum number of trips occur between 16th to 19th hour of the day which is approximately between 4PM to 7 Pm.
2. The total pickups are fewer in the months of April to Jun but increases significantly from July to September, which is the preferred months for Toursist to vist NYC.
3. We see a general trend that over 15,000 trips were covered on an average between 11-16th of a month & again a small spike towards the end of the month.
4. When we compare the data between the days in a week we see that between April - June most pickups are during the weekdays, but as we move towards July-September we can see that the highest pickups were during the weekends (Friday-Sat).
5. Further analysing Uber pickup by Base locations, most of the pickups happen from NYC bases like Hinter, Weiter & Schmecken.
6. Below are the top pickup locations NYC which contribute to maximum rides.
    - JFK International airport
    - LaGuardia Airport
    - TWA Hotels near JFK
    - LaGuardia delta Airlines Terminal
    - 9/11 Museum
    - LaGuardia Southwest airline terminal.

We can conclude that 50% of the rides are available during 4-8 PM within the city & July-September being the preferred months of travel for tourists, if we can make Uber rides available near the Top 5 locations at max capacity Uber might be able to reduce their customer dissatisfaction scores.