PROJECT UBER

Foundation Of Data Science

Group 8

S Abhishek	AM.EN.U4CSE19147
Arvind Kumar K	AM.EN.U4CSE19109
Bikash Chandra	AM.EN.U4CSE19114
Harsha Sathish	AM.EN.U4CSE19123
Navneet Kumar Singh	AM.EN.U4CSE19138

· Table Of Contents

Abstract

- All of us at some point in our life have used uber, so here we will be doing the exploratory analysis of uber pickups. Which will help us better analyze the business. Data analysis helps travel organizations to provide better recommendations for investing in their future trips based on its business and personal trips.
- Data Analysis is being used in many fields such as health care, manufacturing, information technology and so on. With increasing amounts of urban data being collected and made available, new possibilities of data-driven research emerge lead to changes in people living through scientific proof-based decision-making and strategies. Here we concentrate on an urban data set of particular significance which is taxi rides.
- Taxis are essential, and taxi-related information can provide ground-breaking insight into different facets of city life, from economic activity and human behavior to mobility trends. Taxi data involves geographical components as well as several variables associated with each ride.

- Data insight is gain through data pre-processing, feature engineering and data exploratory analysis. This transformation of raw data will enable us to have the meaningful insight into the data and understand the mobility pattern of New York City.
- The primary methodology behind this study is to analyze and find the accuracy of the most frequent category of trip among all trips taken by a customer in a region using data analysis.
- Uber Data Analysis task permits us to recognize the complicated factual visualization of this large organization.
- We will be using 'Python' programming language. Here we analyze the Daily, Monthly and Yearly Uber Pickups in New York City. This mission is primarily based on Data Visualization that will give us information towards use of ggplot2 library for perception of the data. There are many questions that can be answered but here we will be focusing on,
    - Uber Pickups and distribution in NYC
    - Time when Uber pickups happen regularly.
    - Days when pickup happens regularly.
    - Pickup distribution in the zones.
    - Finding out the hotspot areas

Introduction

- Taxi is one of the urban public transports in many busy countries. Unlike other public transports, taxi rides provide accessibility, convenience, yet privacy to passengers.
- A competitive and reasonable taxi pricing is worth the ride for private car users to switch to a taxi service.
- Millions of taxi trips data are generated on monthly basis, which this data can be useful to gain the insight of the traffic patterns and obtain a clear view of urban city life.
- Nowadays, a Real-time prediction for experience provider demand (always mirrored by way of the wide variety of pickups) is increasingly more critical for the motive of enhancing the effectively and sustainability of the city transportation system.
- Newly aroused utility matters like experience sharing and independent mobility dispatching are primarily based on strong demand predictions.
- The growth of the science of geographic information provides new possibilities for urban understanding and planning. With automated data collection of taxi movement, a city's operation can be extracted from geospatial data in both spatial and temporal point of view.

- It provides a more accurate depiction of the nature of a region, considering that daily movement and activities found in geospatial data indicate the social-economic properties of urban functions.
- We need to advocate a deep gaining knowledge of primarily based strategy to make dynamic predictions for uber pickups the use of historical data.
- Other attributes like taxi Id, total fare amount, and number of passengers are also recorded which allows researchers to study the traffic congestion, economics of fare pricing, and optimal fleet size.

Motivation & Objectives

- According to the United Nations in 2008, for the first time in history, half of the world's population was living in urban areas (United Nations, 2008).
- In 1950 there were only two metropolises with at least 10 million inhabitants. In 1975 only three metropolises broke that barrier.
- Today there are 21 megacities with more than 10 million inhabitants, and in 2025, the United Nations estimates that there will be 27 cities (United Nations, 2012).

- This is clear evidence of the fast growth of urbanization in terms of population and size.
- The demand for better services (e.g. public transportation, energy, communications) and urban planning (e.g. infrastructures, environments, policies) increases with the rapid growth of urban areas.
- In order to maintain a constant flow of people and vehicles, we need to reduce the use of individual means of transport (e.g. car) and stimulate the use of public transportation (e.g. bus, metro, train).
- Traffic is one of the major sources of toxic compounds present in combustion gases that negatively impact the health of urban inhabitants (EEA, 2011), (Borrego, et al., 2000), (Zavala, et al., 2006), (Ndoke & Jimoh, 2005), (Becker, et al., 2000).
- There is a need to address this issue today while low-carbon transport systems are still being developed.
- However, we need to improve the public transportation system in order to meet citizens' needs.
- A more efficient public transportation system can lead to a reduction in traffic congestions and consequent reduction of energy consumption and pollution.

- Nevertheless, to optimize the public transportation network it is essential to understand what drives the common citizen and what their needs are.
- We need a better understanding of city dynamics.
- Gathering data from the traditional public transportation (e.g. bus, train, metro) can provide us with a relevant database and information on general passengers' movement.
- However, it does not provide the exact origin and destination for each passenger, since these transportation modes rely on pre-designated stops and paths.
- The taxi can be a way to retrieve a large data set of information with higher precision when we focus on the origin and destination of each trip.
- Taxis can pick-up the passengers right where they are standing, and drop them off precisely at their desirable destination, without being bound to a pre-determined path.
- The process of data collecting is transparent and non-intrusive to the passenger. Additionally, taxis can be used as a probe for traffic conditions (Castro, et al., 2012), (Yuan, et al., 2011a), (Gühnemann, et al., 2004), (Liu, et al., 2009a).

- At the same time, we are experiencing new developments in pervasive and ubiquitous computing technologies, such as a global system for mobile communications (GSM) and a global positioning system (GPS), which provide useful tools for sensing social and traffic activities in cities.
- Nowadays we are able to access a wider variety of devices, with a growing number of features and computational capabilities. This technological diversity provides us with the tools to sense urban spaces. It allows us to either take a collective snapshot of all urban activity or simply follow the pattern of a single vehicle or individual.
- Analyzing GPS-enabled vehicle traces and mobile phone activity thus provides, to some extent, an overview of how the city functions.
- Taxis are currently equipped with GPS devices for better monitoring and dispatching.
- Their traces have been used to study various aspects of the traffic network as they provide fine-grained data that reflects the state of traffic flow in a city.
- These traces typically carry occupancy information on pick-up and drop-off location.

- The ubiquity of taxis has attracted considerable attention for a while, in order to extract information and develop prediction systems, which led to a significant amount of research work being performed around the exploration of taxi-GPS traces.
- Facing the challenges of growing cities and by taking the opportunistic sensing approach, a main question is posed: to what extent can GPS traces of taxis be used to infer the city's dynamics, namely the inhabitants' patterns?
- Furthermore, what is the role of taxis in the complex relations amid the diversity of urban data sources? Although previous research on this topic led to important findings, there are still challenges yet to explore that we aim to analyze in this thesis.
- Our work deepens the spatiotemporal analysis and the study of predictability of taxi trips by using complementary data (e.g. Points Of Interest, weather conditions); explores the underlining relationship between taxi volume and mobile phone activity, two important urban data sources; and further extends the study of the relation among urban data sources by examining the relationship between taxi mobility

patterns, weather conditions and the level of concentration of exhaust gases, to estimate the concentrations of gas in urban areas.

Study System

What is Data Analysis?

- Data science is often thought to consist of advanced statistical and machine learning techniques.
- The process of cleaning, transforming, manipulating data into useful information is known as Data analysis.
- When we take a particular decision based on previous data that is data analysis. We can make future decisions using data analysis.

What is Exploratory Data Analysis?

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- It is a good practice to understand the data first and try to gather as many insights from it.
- Exploratory Data Analysis is all about making sense of data in hand.

Why we use Data Analysis?

- Exploratory data analysis (EDA) is a classical and under-utilized approach that helps you quickly build a relationship with the new data.
- It is always better to explore each data set using multiple exploratory techniques and compare the results.
- This step aims to understand the dataset, identify the missing values & outliers if any using visual and quantitative methods to get a sense of the story it tells.
- It suggests the next logical steps, questions, or areas of research for your project.
- All the business has lots of data. To grow business, sometimes data analysis required.
- By analysing data, we get important topics on which work out and make our plan for the future through which made perfect future decisions.

- Most of the businesses going online where the data generate increases day by day.
- To grow business with this competitive environment data analysis is necessary.

Materials & Methods

- In this project, data is obtained through Kaggle repository titled “Uber TLC FOI Response”
- This directory contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015.
- Trip-level data on 10 other for-hire vehicle (FHV) companies, as well as aggregated data for 329 FHV companies, is also included. All the files are as they were received on August 3, Sept. 15 and Sept. 22, 2015.
- This data was used for four FiveThirtyEight stories: Uber Is Serving New York’s Outer Boroughs More Than Taxis Are, Public Transit Should Be Uber’s New Best Friend, Uber Is Taking Millions Of Manhattan Rides Away From Taxis, and Is Uber Making NYC Rush-Hour Traffic Worse?.

The Data

- The dataset contains, roughly, four groups of files:
- Uber trip data from 2014 (April - September), separated by month, with detailed location information
- Uber trip data from 2015 (January - June), with less fine-grained location information
- Non-Uber FHV (For-Hire Vehicle) trips. The trip information varies by company, but can include day of trip, time of trip, pickup location, driver's for-hire license number, and vehicle's for-hire license number.
- Aggregate ride and vehicle statistics for all FHV companies (and, occasionally, for taxi companies)

Uber trip data from 2014

- There are six files of raw data on Uber pickups in New York City from April to September 2014.
- For coarse-grained location information from these pickups, the file taxi-zone-lookup.csv shows the taxi Zone (essentially, neighborhood) and Borough for each locationID.
- The files are separated by month and each has the following columns,

- Date/Time : The date and time of the Uber pickup
- Lat : The latitude of the Uber pickup
- Lon : The longitude of the Uber pickup
- Base : The TLC base company code affiliated with the Uber pickup

- The file uber-raw-data-janjune-15.csv has the following columns,
  - Dispatching_base_num : The TLC base company code of the base that dispatched the Uber
  - Pickup_date : The date and time of the Uber pickup
  - Affiliated_base_num : The TLC base company code affiliated with the Uber pickup
  - locationID : The pickup location ID affiliated with the Uber pickup

Non-Uber FLV trips

- The dataset also contains 10 files of raw data on pickups from 10 for-hire vehicle (FHV) companies.
- The trip information varies by company, but can include day of trip, time of trip, pickup location, driver's for-hire license number, and vehicle's for-hire license number.

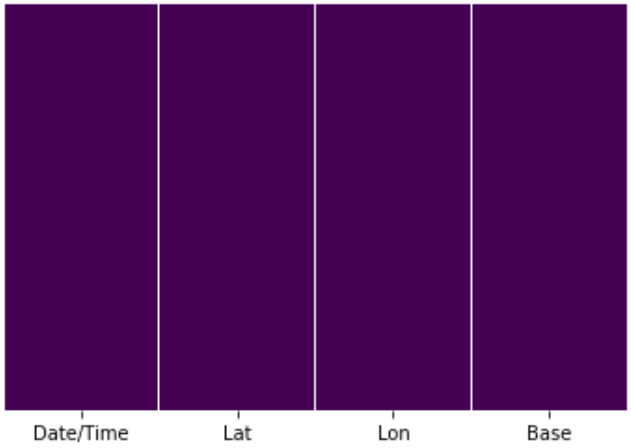
Aggregate Statistics

- There is also a file other-FHV-data-jan-aug-2015.csv containing daily pickup data for 329 FHV companies from January 2015 through August 2015.
- The file Uber-Jan-Feb-FOIL.csv contains aggregated daily Uber trip statistics in January and February 2015.

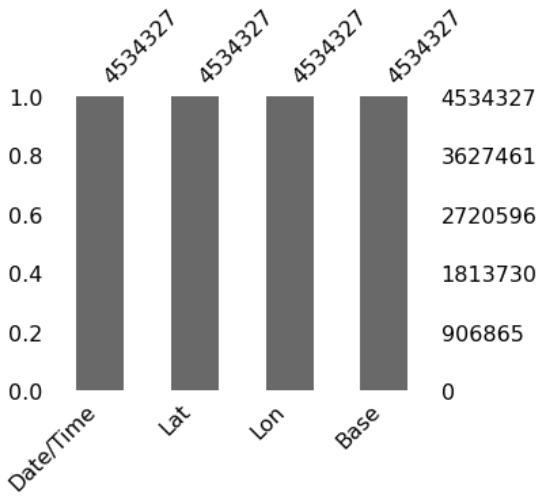
Data Cleaning

- Descriptive Statistics
    - Describe the basic features of dataset and obtain a brief summary of the data
    - Describe() in Pandas library helps us to have a brief summary of the dataset
    - It automatically calculates basic statistics for all numerical variables including NaN values
- Check for columns with Null Values

- **Heatmaps**
    - Visualize the data in a 2-d format in the form of colored maps.
    - Color variation gives visual cues to the readers about the magnitude of numeric values.
- It can describe the density or intensity if variables, visualize patterns, variance and even anomalies.

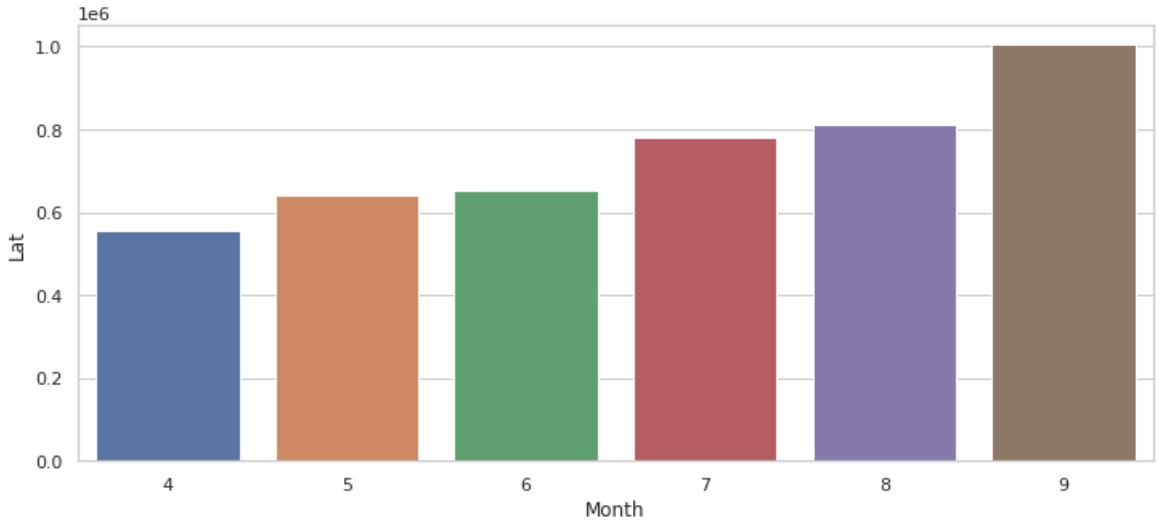


- Missingno Library
    - Missingno library provides the distribution of missing values in the dataset using the plots which helps to locate the missing values present in each column and check if there is a correlation between missing values of different columns.



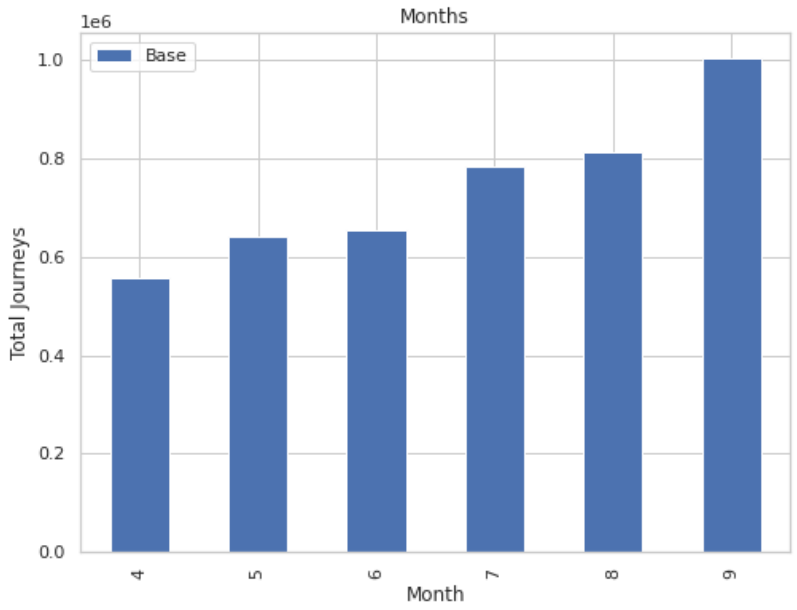
Exploratory Data Analysis

Trip by Month



- Using Seaborn, we have plotted the bar plot of Month (x - axis) vs Latitude(y-axis)

- From the above graph, we can see that there is a steady increase in the number of pickups every month
- A slightly larger increase is noted from August to September (Total Journey Ratio = 1 * 10^6)
- Next, we plot the graph showing the number of pickups every month

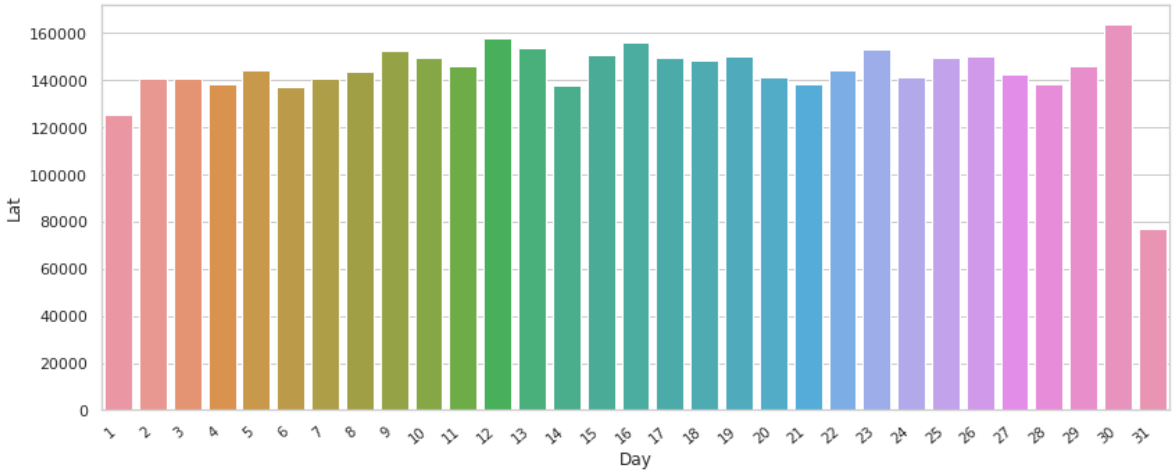


- From the above graph, it is observed that there is a significant increase in the number of pickups every month.

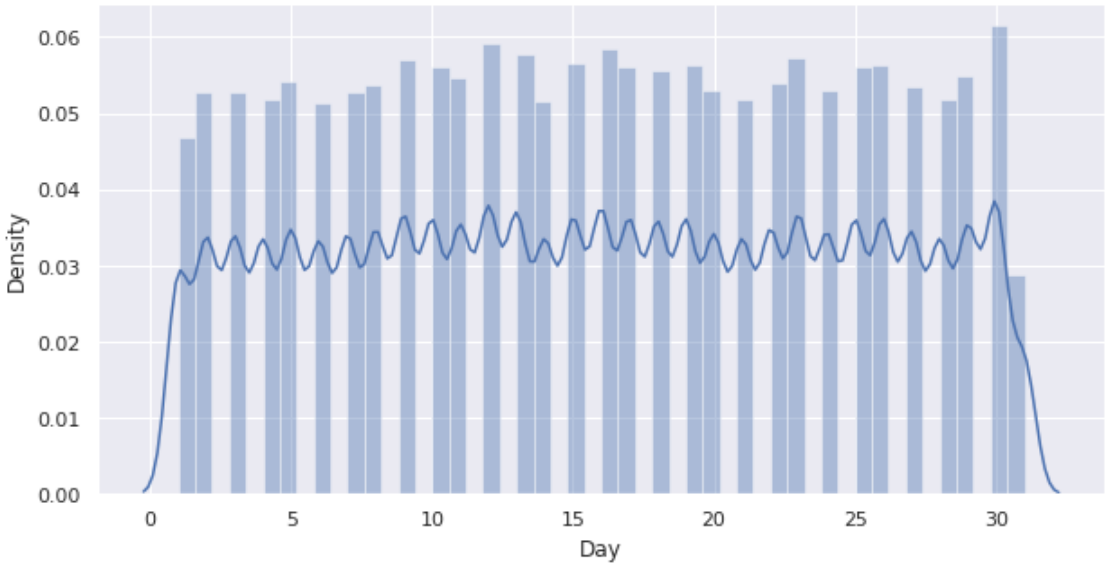
Trip by Week Day

- Find out which day had more journeys than other days over 6 months
- Using Seaborn, we have plotted the bar plot of Day (x - axis) vs Latitude(y-axis)

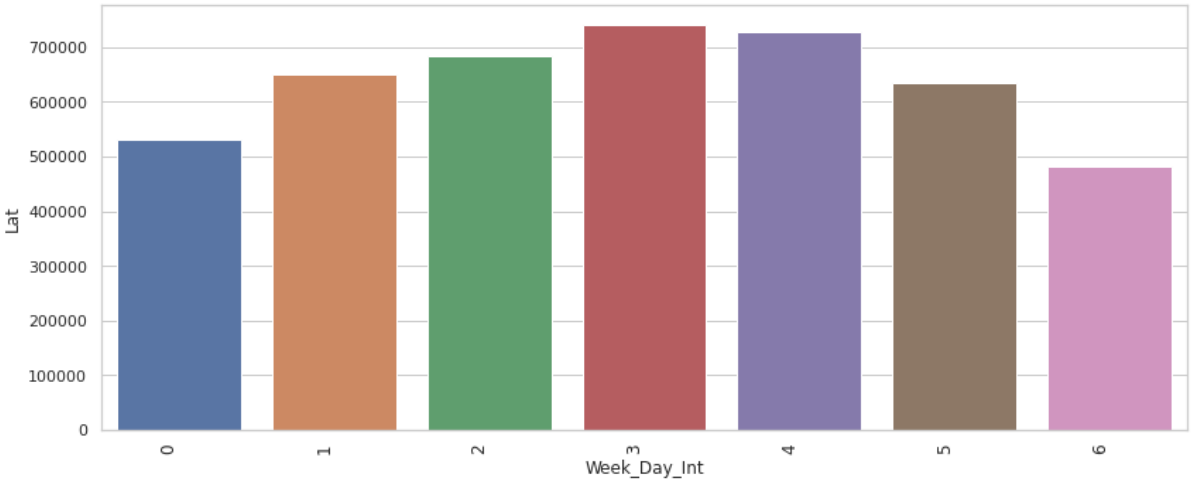
- The y-axis value w.r.t to the corresponding day in the x-axis shows the sum total of fares on that particular date from each month, that's why 31 has the least as all the months don't have 31 days.



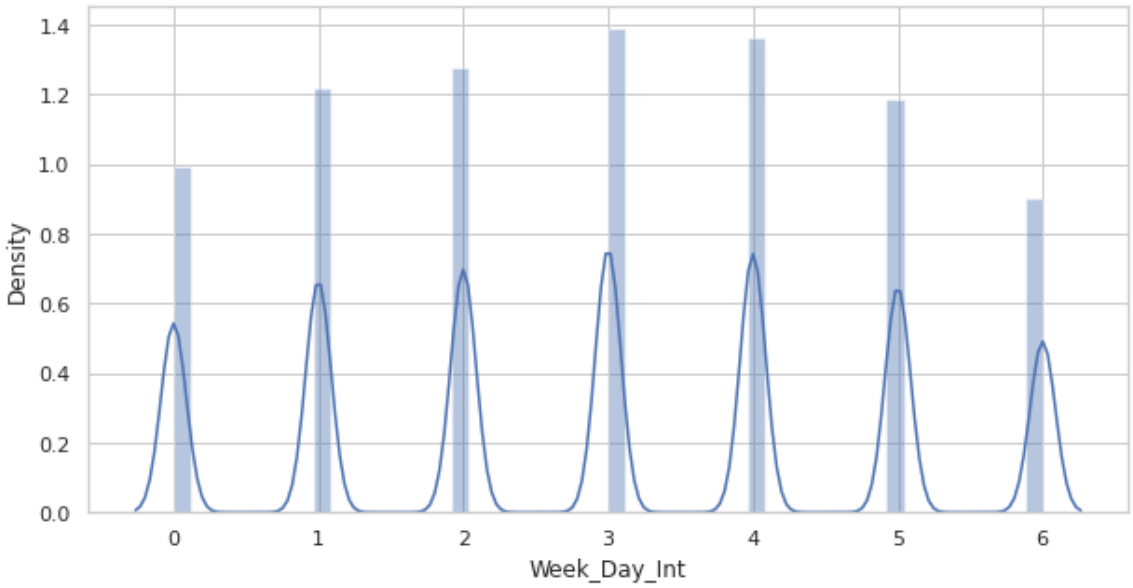
- From this graph we realize that the 1st day of the month has least fares (31 is exception) and 30 has the most fares, while the other dates are almost the same.
- Next is a distplot to depict the data in more continuous form, by showing in the range of 5 days each in the x-axis rather than the discrete specific dates as shown in the bar plot above, and the y-axis shows destiny.



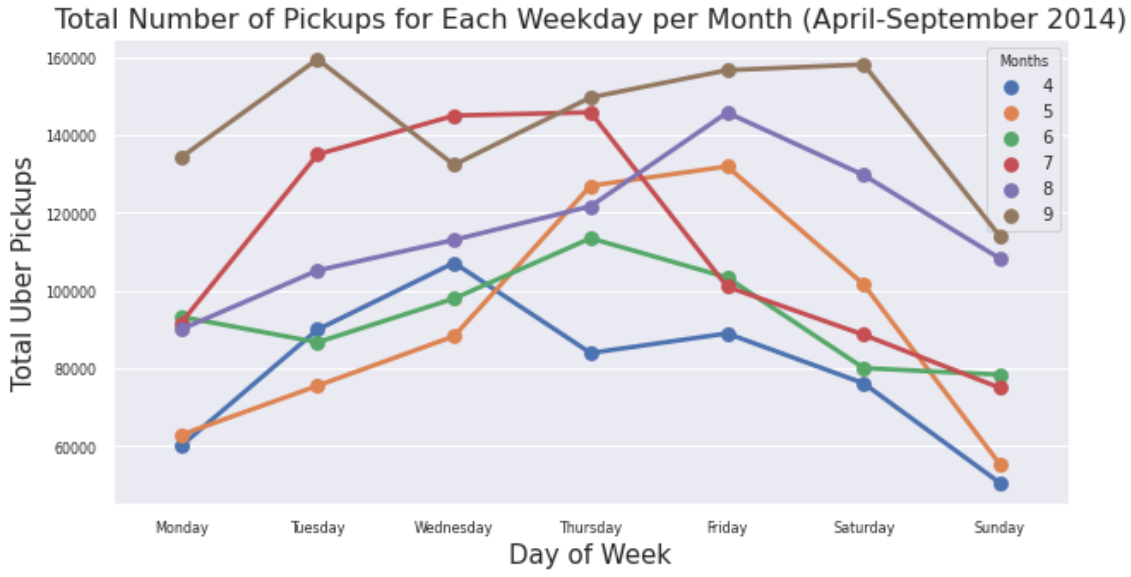
- From this graph we get the same trends as in the bar plot, we did before this as the data is same.
- Next is a bar plot to depict the data of a particular day of the week rather than dates as done before.
- In x-axis 0 refers to monday,1–tuesday,2-wednesday,3-thursday,4-friday,5-saturday,6-sunday.
- The y-axis depicts the total number of pickups on that specific day throughout the six months (April - September).



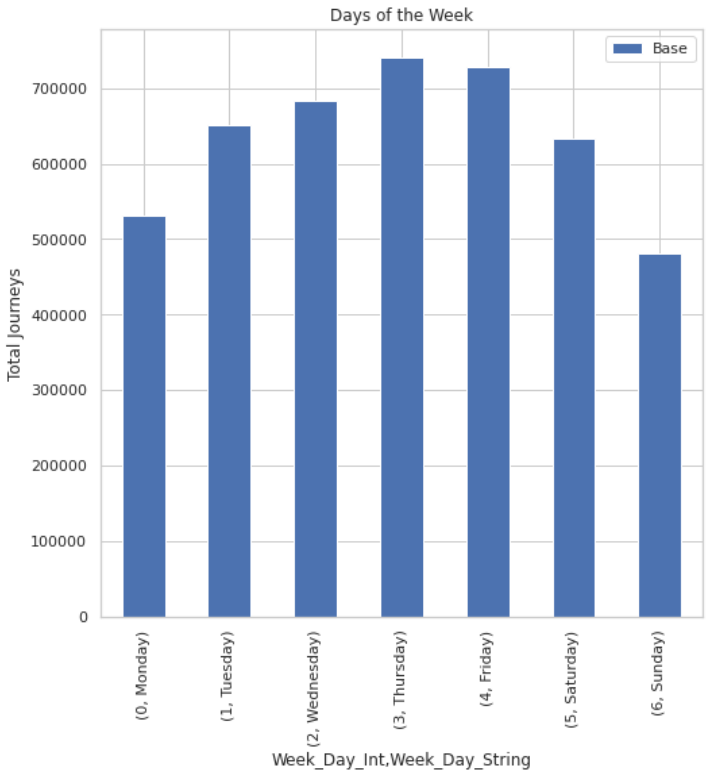
- From this graph we notice a strange trend where the closer we are to the middle of the week the greater number of pickups and as we move to the start and end of the week (Mon and Sun) it reduces.
- Next is a distplot to depict the data in more continuous form, by depicting destiny rather than Latitude on y-axis, just to notice the increase and decrease in fares in a more pictorially detailed way.



- Next is a point plot to depict the data of the number of journeys on each day of the week, by depicting number of journeys on y-axis, and the week days on x –axis.
- We did this because Point plots can be more useful than bar plots for focusing comparisons between different levels of one or more categorical variables.
- They are particularly adept at showing interactions: how the relationship between levels of one categorical variable change across levels of a second categorical variable.
- The lines that join each point from the same hue level allow interactions to be judged by differences in slope, which is easier for the eyes than comparing the heights of several groups of points or bars.



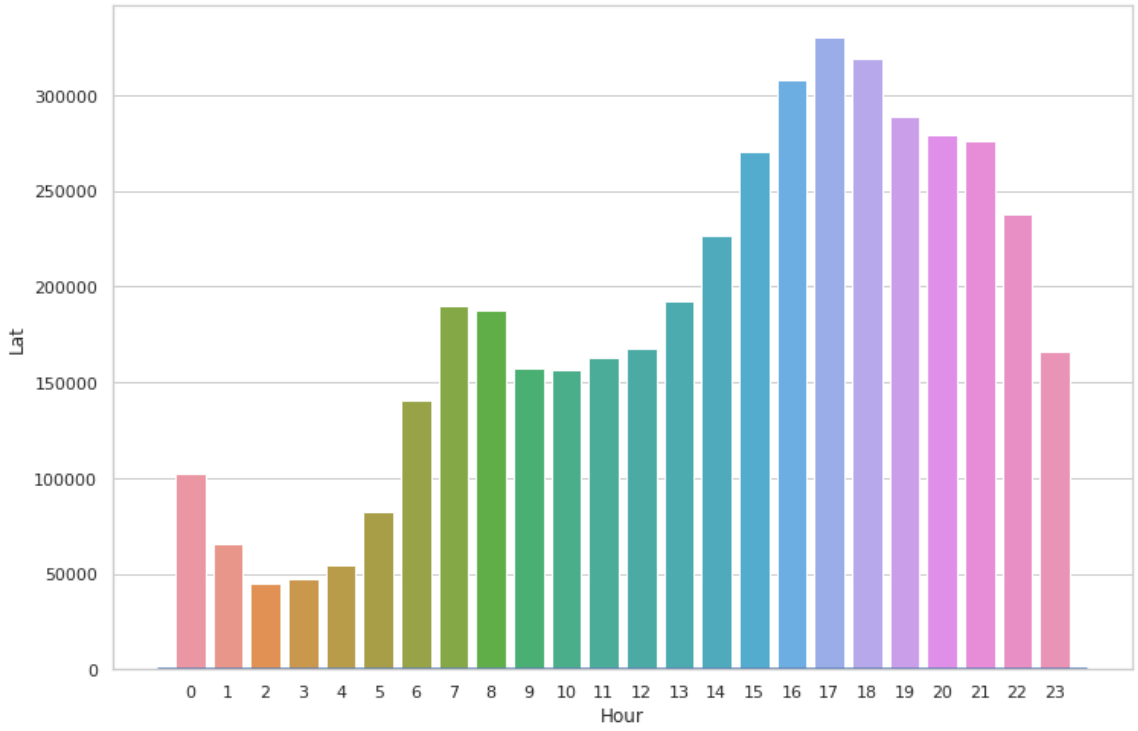
- Here, for the first time we are not just plotting the data of each weekday as the sum total of the number of fares in the whole of 6 months, but also, we have given a separate line for each month which are formed from their respective dots.
- We realize that each week day has varying trends of number of journeys each month, April and July have Wednesday as the busiest, September has Tuesday as the busiest, august and may have Friday at the peak, while June has Thursday at the top.
- But Sunday has the least number of journeys in al the months as it's a holiday.
- Next, we made a final bar plot to summarize the number of journeys on each day of the week in a more uniform way.



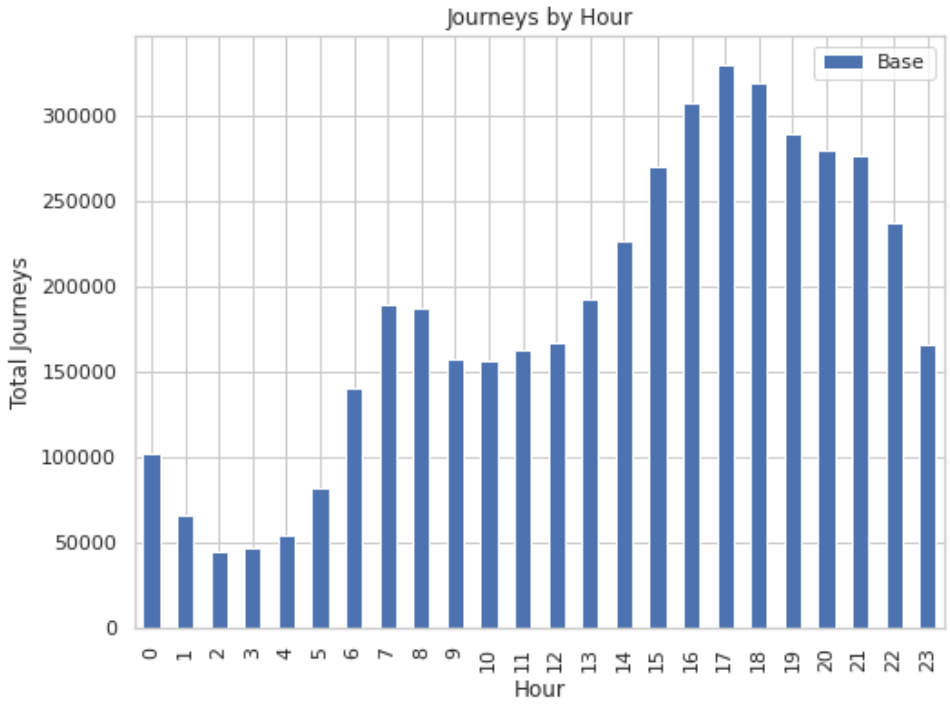
- Interestingly, there were more rides on Thursday and Friday is closely followed.
- It's also interesting that there were more rides on even Wednesday and Tuesday than on Saturdays

Trip by Hour

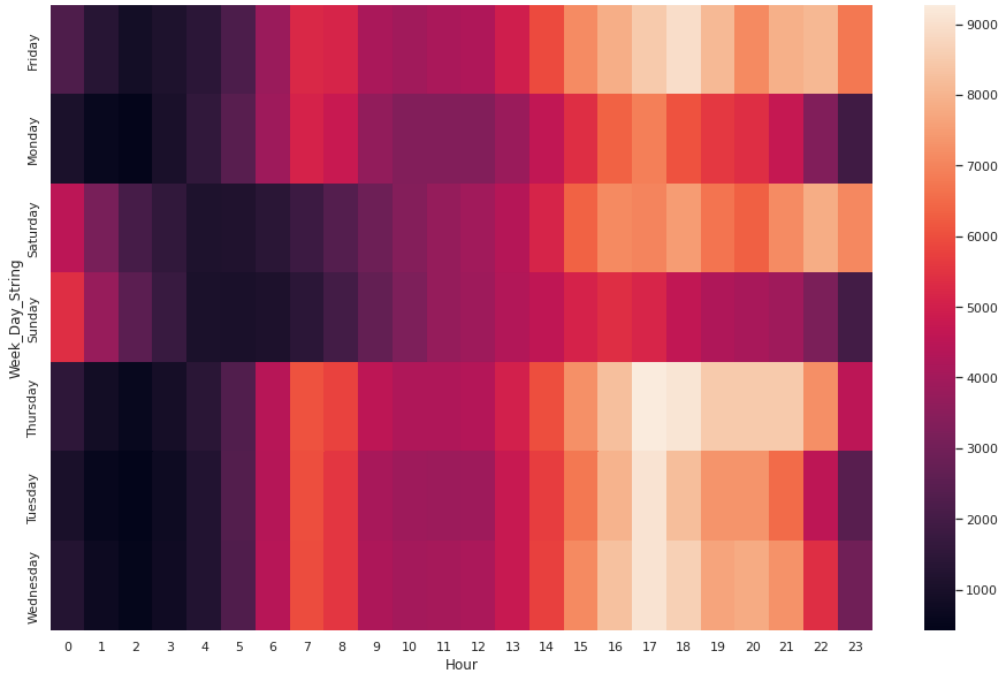
- Using Seaborn, we have plotted two bar plots of Hour (x - axis) vs attitude(y-axis) below,



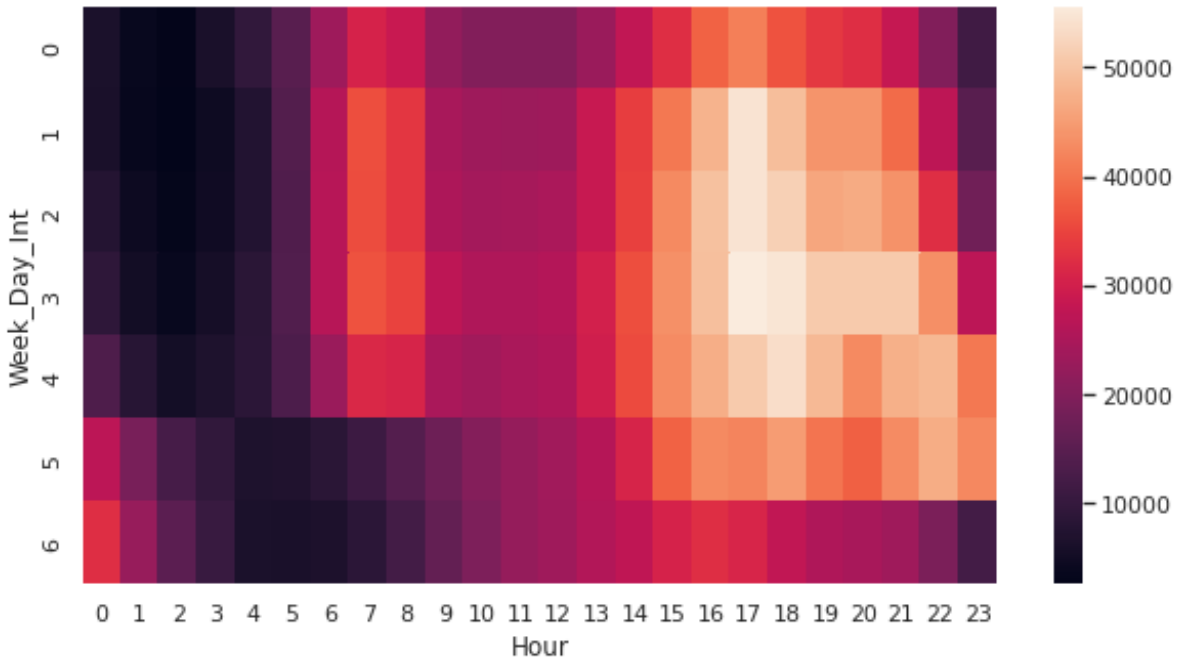
- In this second bar plot we have given a common base, just for uniform representation.



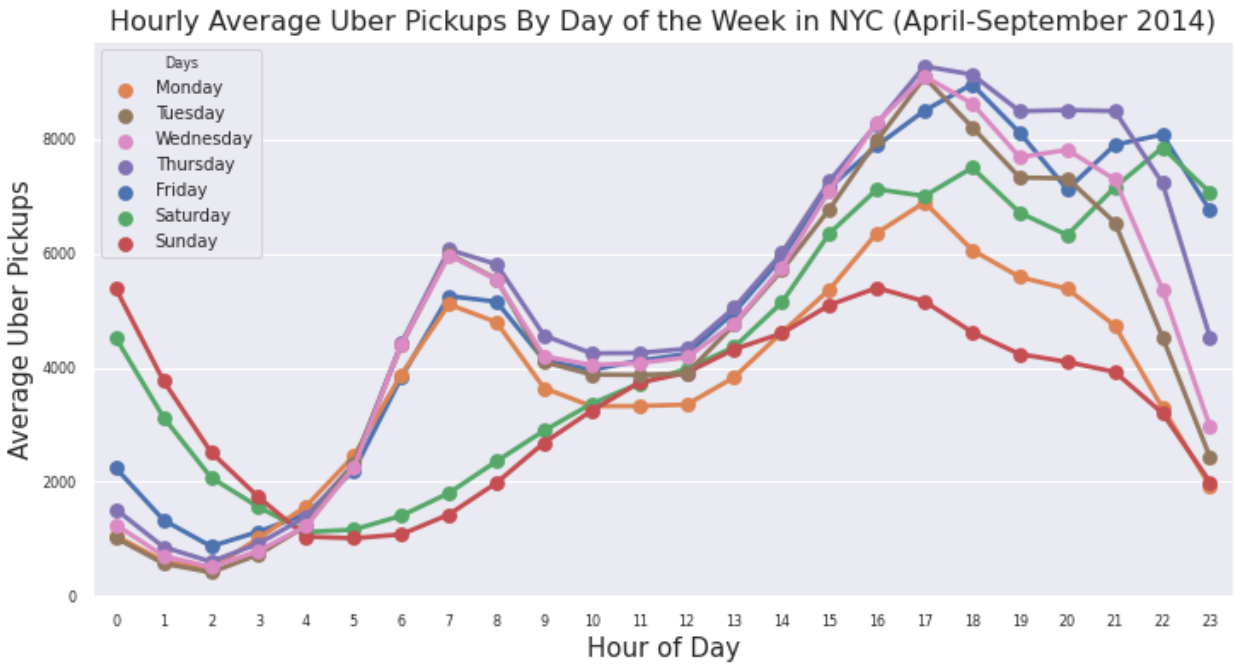
- There are some interesting trends which we get from the above two bagplots.
- If we look at the extremes of the graph 17th hour of the day(5pm) has the greatest number of rides while 2am has the least number of rides.
- There are three local maxima in the graph, the first one being at the start, I.e., at midnight, from which it keeps on decreasing till 2 am, then again starts increasing till 7am which is the 2nd maxima, after which it keeps on dropping till 10am and again increase till 5pm the last maxima after which it keeps on decreasing till the day ends.
- Next, we have two heat maps, where x-axis representing the hours of the day, while y-axis is taken by the week days, to show how the number of rides changes with the time in every particular week day side by side.
- If we look at the extremes of the graph 17th hour of the day(5pm) has the greatest number of rides while 2am has the least number of rides.
- A heatmap contains values representing various shades of the same color for each value to be plotted. Here the darker shades of the chart represent lower values than the lighter shade. For a very different value a completely different color can also be used.



- This heat map has days in y-axis (written in words, with not the order of days in a week)

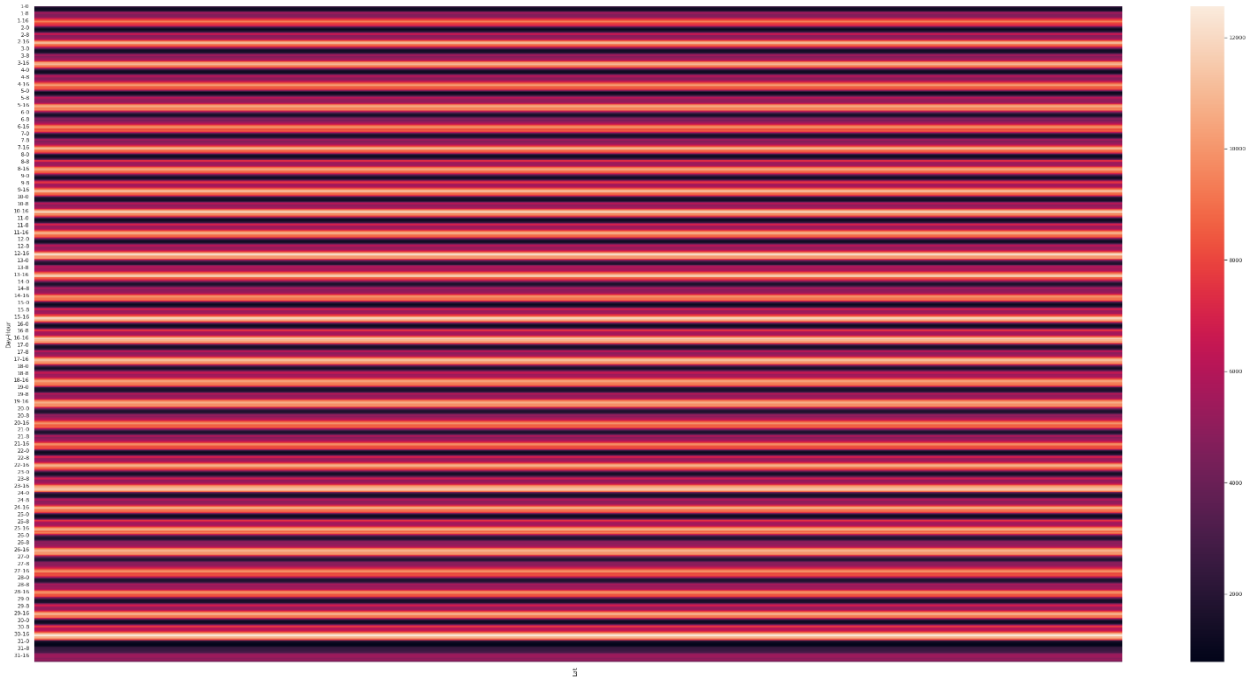


- This heat map has order from 6 being Sunday to 1 being Monday.
- This heat map shows that only the weekends, I.e., Saturday and Sunday have different trends when compared to the working days.
- In the working days (monday to Friday) it starts with pitch dark indicating the least number of rides at midnight to increasing in a linear rate till 21st hour(9pm) and then again decreasing towards the end of the day.
- This shows that the majority rides are taken by people having jobs or some activity at the morning which increases till night as people start returning to their homes and then decreases towards the end.
- On the other hand, in weekends, it is dark most the times with having some lighter shades between 10 am to 6 pm.
- This shows that people don't go out much in the weekends except for a few hours for recreational activities.
- Here we have a dot plot next to show a more detailed depiction of data so that we can differentiate the trends of each on respective hours of the day with respect to the respective week days separately as each individual line is representing different day of the week.



- As we can see in the graph the green and red line are having different slopes when compared to the other four lines, this is because they are the weekends while other are working days.
- As discussed, earlier weekends have rides mostly between 10am to 6pm, so green and red have a slight bump in that time period.
- On the lines of the working days there are two bumps, the first one on 7-8 am and the other being 4-6pm ,which represents the arrival and departure of the people with jobs from their office.
- Next, we have another heat map based on day and hour, but in this one in the y-axis, the date is provided along with 1st eight hours and the next

eight hours respectively for each date of the month. The light shade represents higher number of rides.

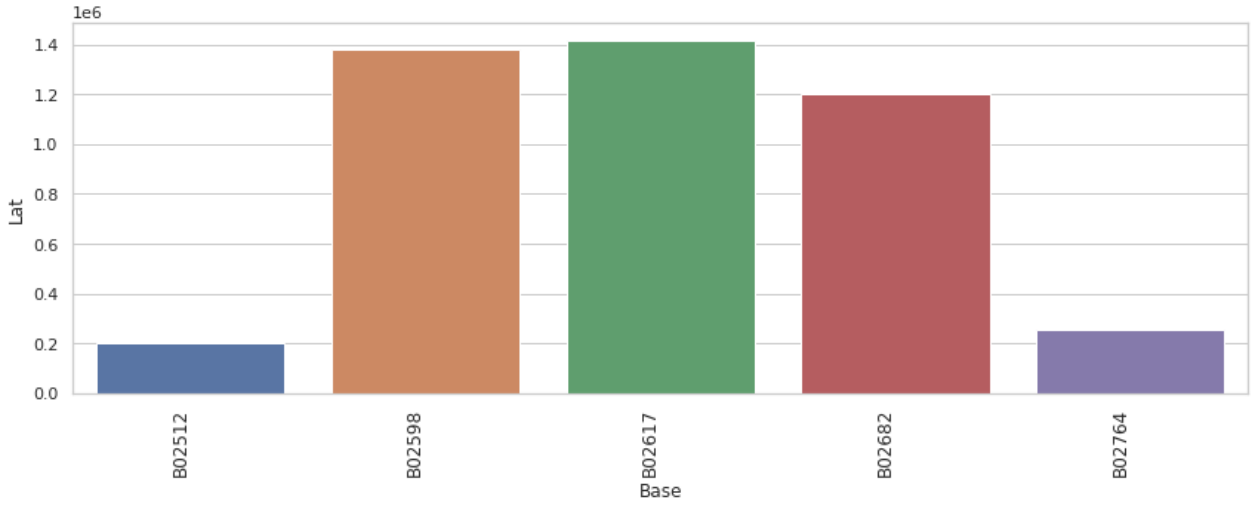


- As we are showing data of 30 days of a month with each day being divided into two parts this graph is very much compact.
- From looking at the graph we see that there's a pattern in which a bright band is adjacent to dark one and so on.
- This is because we are showing 0-8 hours of the particular date first and then 8-16 next, for 0-8 hours have least activity we get a dark band while 8-16 is the most active hour so we get a bright band, and hence we see this uniform pattern.

- Also, as we go down the graph the dark bands get more darker while the bright bands are fading a bit.
- This shows that when approaching end of a month the overall activity tends to go down a bit.

Trip by Base

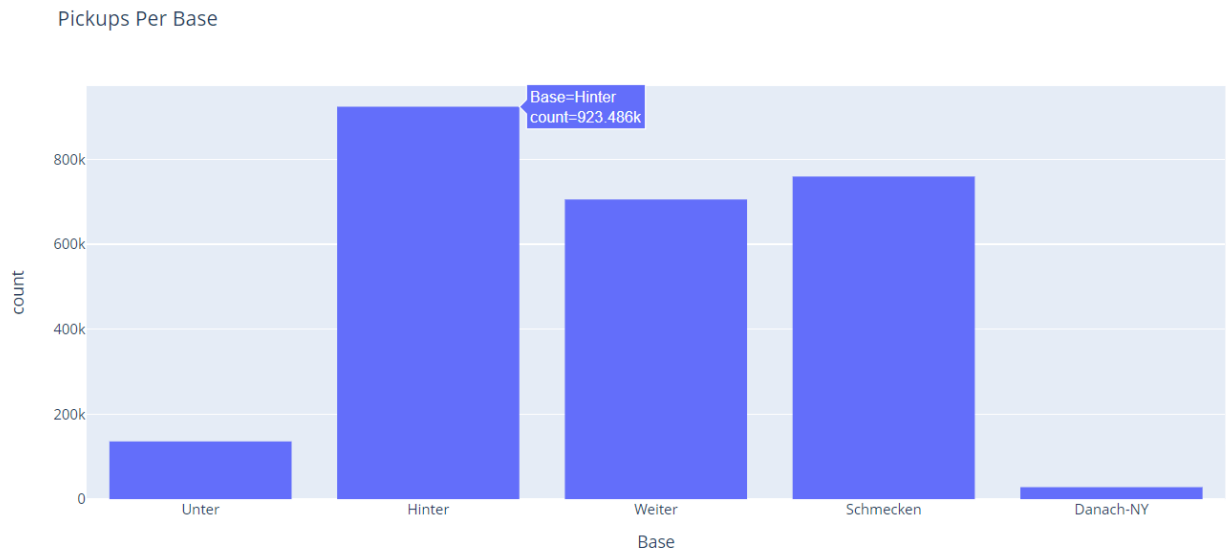
- This is a bar plot generated for Uber pick-up statistics from the six Uber bases of New York City.



- To understand the graph better we first generated the number of rides per base in tabular form.

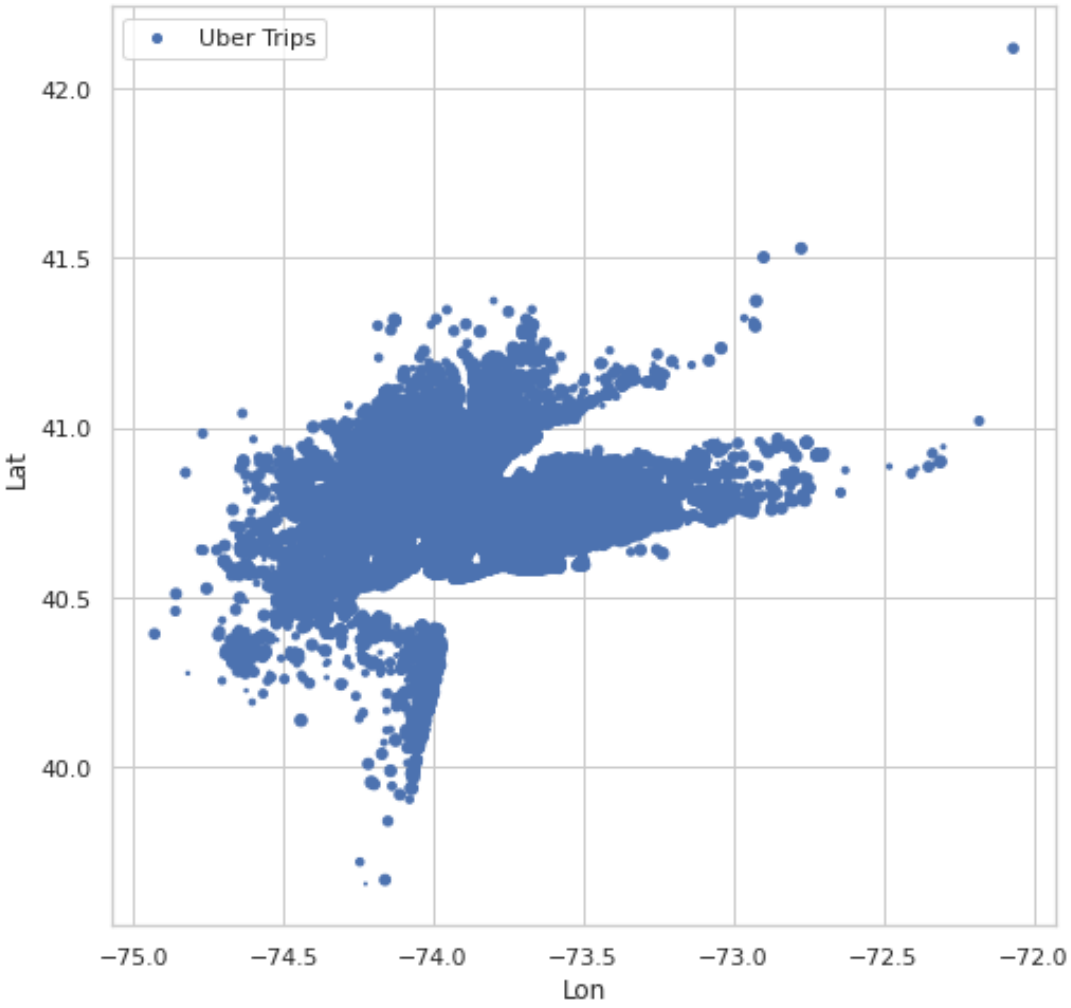
```
B02617    1417983
B02598    1379578
B02682    1198901
B02764     254931
B02512     200353
```

- Next, we have a histogram, in which the names of the six bases is provided on x-axis and the count of rides is scaled on y-axis.



- The highest pickups have been from the base Hinter (923.486k) and the lowest pickups have been from the base Danach-NY.

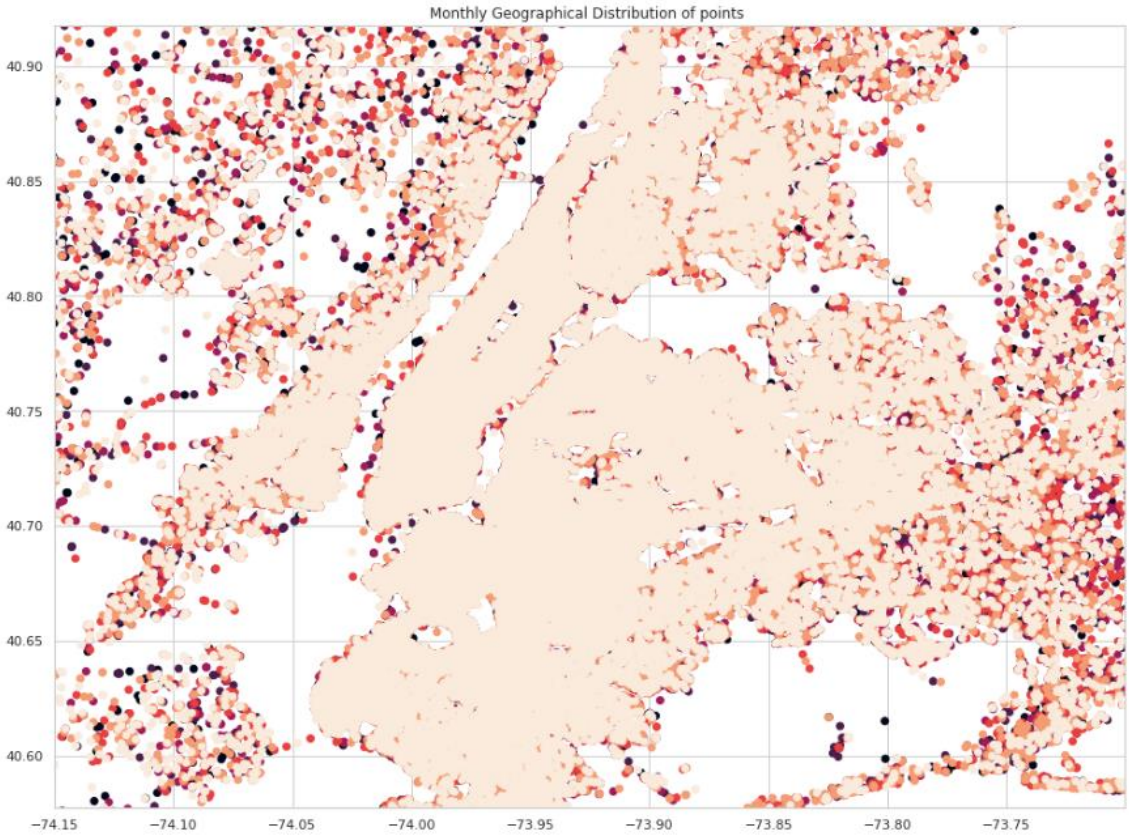
Density of Uber Trips



- Here on the x-axis, we are having the longitude and, on the y-axis, we are having the latitude and we are showing the density of uber trips.
- As we can see that there are dots spread towards the exterior region while at the central part it is super dense, which means that the majority of the rides are taken in the central city.

Monthly Geographical Distribution

- The best way to depict Monthly Geographical Distribution is via scatterplot. Here y represents latitude while x gives the longitudes.
- The darker shades represent most activity while the lighter region shows less activity.



- This graph shows us the monthly Geographical Distribution of points.
- Also, we can observe that its denser at the left and the right part.

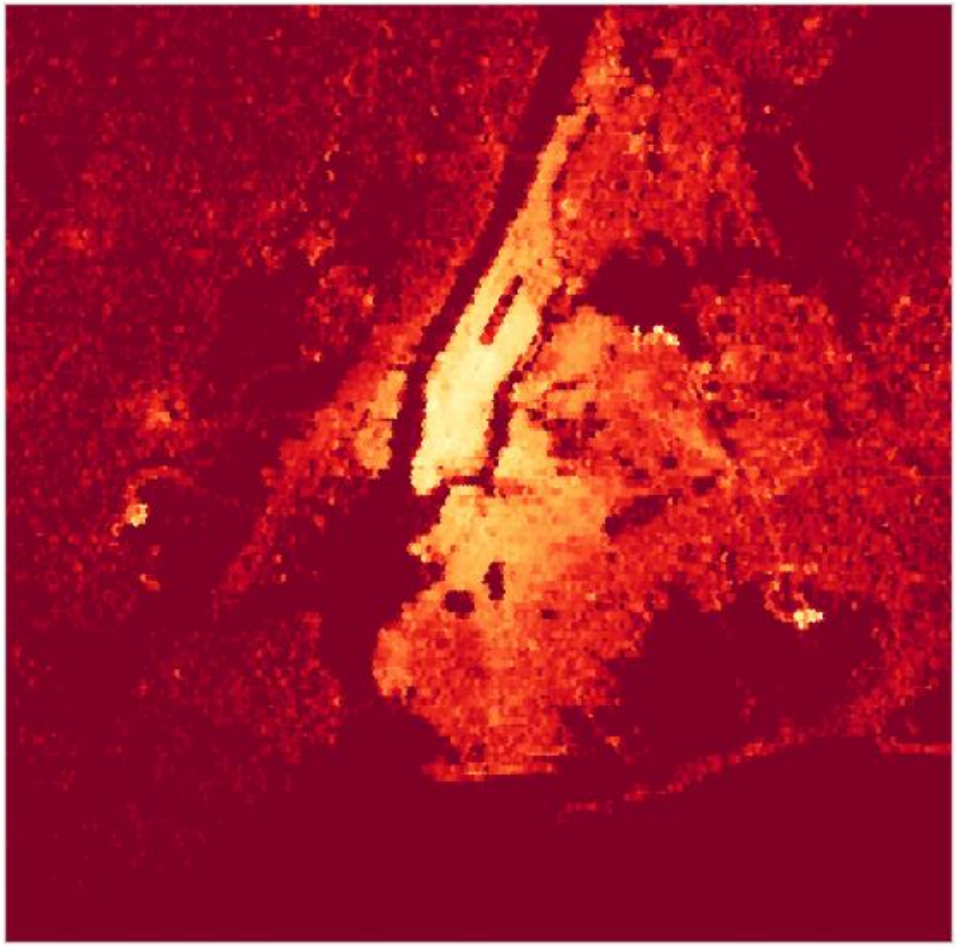
Daily Geographical Distribution

- Here we have another scatterplot where we are giving the plot of a day rather than a month.

- In this graph we are showing the daily geographical distributions of points.
- We can observe that its highest in the middle region whereas density decreases as we move away from the middle.
- This is due to the fact that there are more places in the core of the city that people intend to visit.

Heatmap of Total Pickup

- In this graph we are showing the Heatmap of Total pickups and we can see the locations where the pickups have been the most in the region and the least as well.

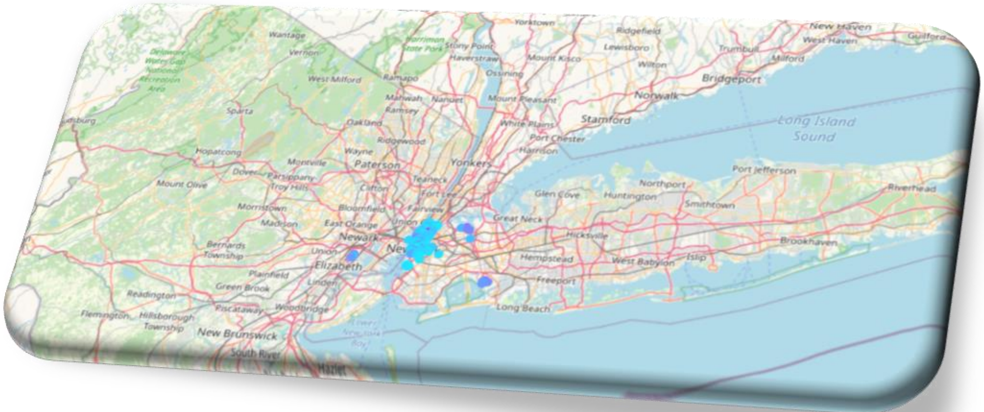


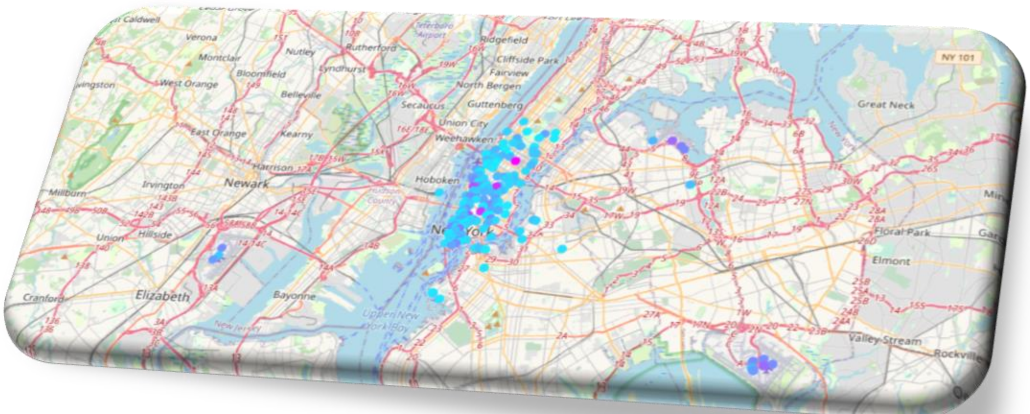
- We can see that in middle region there has been the greatest number of pickups.

Highly Clustered/Hotspot Areas

- While it can be useful to look at the raw data set of our heatmap and gain intuition, looking at snap shots doesn't give us enough information to make good conclusions, and plotting hours of individual pick-ups just makes a mess.
- We want actionable conclusions from this data set, so we'll need to be able to make quantitative comparisons between different regions in the city and we may not use the Bases column since they cover huge areas.
- We'll implement the DBSCAN clustering method from scikit-learn instead of structured spatial binning.
- The K-Means algorithm is likely the most common clustering algorithm.
- But for spatial data, the DBSCAN algorithm is far superior in this kind of scenario where you are also using geographical data.
- The DBSCAN algorithm will group points together that meet a specified density metric.
- Basically, we'll define a maximum distance to make two individual points count as neighbors, as well as a minimum number of neighbors for a group of points to qualify as a cluster.

- The algorithm will sort the points into groups which meet the criteria and discard all of the outliers.
- Once DBSCAN has identified all applicable clusters, we can easily calculate the centroid using the MultiPoint class from Shapely and plot the results.
- This allows us to precisely identify locations which experience a high volume of pick-ups during a specified time frame.
- By using the total number of pickups in an individual cluster as a metric for coloring the hot spot locations, we can visualize the intensity of a given hotspot in addition to its centroid.
- First, we'll write a function which runs the clustering algorithm and returns the "hot spots." We'll get the coordinates of the centroid and the number of pickups in each cluster.


Wantage
Vernon
West Milford
Ramapo
Stony Point
Haverstraw
Ossining
Yorktown
Mount Kisco
Ridgefield
Lewisboro
Wilton
Trumbull
New Haven
West Haven
Milford
Guilford
Bridgeport
Norwalk
Stamford
Long Island Sound
Sparta
Oakland
Mahwah
Ramsey
Nanuet
Mount Pleasant
White Plains
Port Chester
Harrison
Ridgewood
Wayne
Paterson
Yonkers
Hopatcong
Montville
Dover
Parsippany-Troy Hills
Mount Olive
Morristown
Madison
East Orange
Bloomfield
Clifton
Teaneck
Fort Lee
Fairview
Union City
Great Neck
Glen Cove
Huntington
Northport
Port Jefferson
Smithtown
Riverhead
Hicksville
Brookhaven
Newark
Bernards Township
Union
Elizabeth
Plainfield
Linden
Hempstead
West Babylon
Islip
Freeport
Green Brook
Readington
Piscataway
Woodbridge
Long Beach
Flemington
Hillsborough Township
New Brunswick
South River

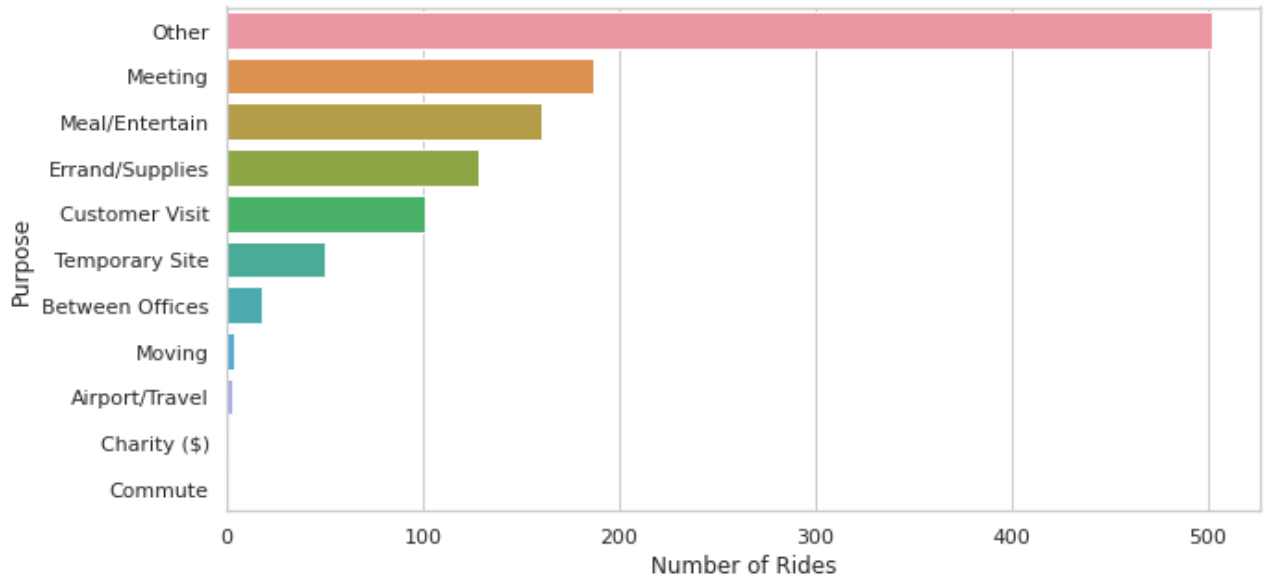

Caldwell
Verona
Montclair
Nutley
Rutherford
Ridgefield
Cliffside Park
Fairview
North Bergen
Guttenberg
Secaucus
Union City
Weehawken
Hoboken
Bloomfield
Lyndhurst
Belleville
Livingston
West Orange
East Orange
Kearny
Harrison
Newark
Irvington
Millburn
Hillside
Union
Cranford
Elizabeth
Bayonne
Upper New York Bay
New York
Great Neck
NY 101
Floral Park
Elmont
Valley Stream
Rockville


CR 678
CR 670
CR 653
US 1
US 9
Hoboken
I 78
NJTP
Holland Tunnel
I 78
Jersey City
New York
Manhattan Bridge
Brooklyn
Calvary Cemetery

- The maps above show the areas experiencing more than 25 pickups that occur within 50 meters of each other after 4:00 PM on August 14, 2014.

Network Analysis

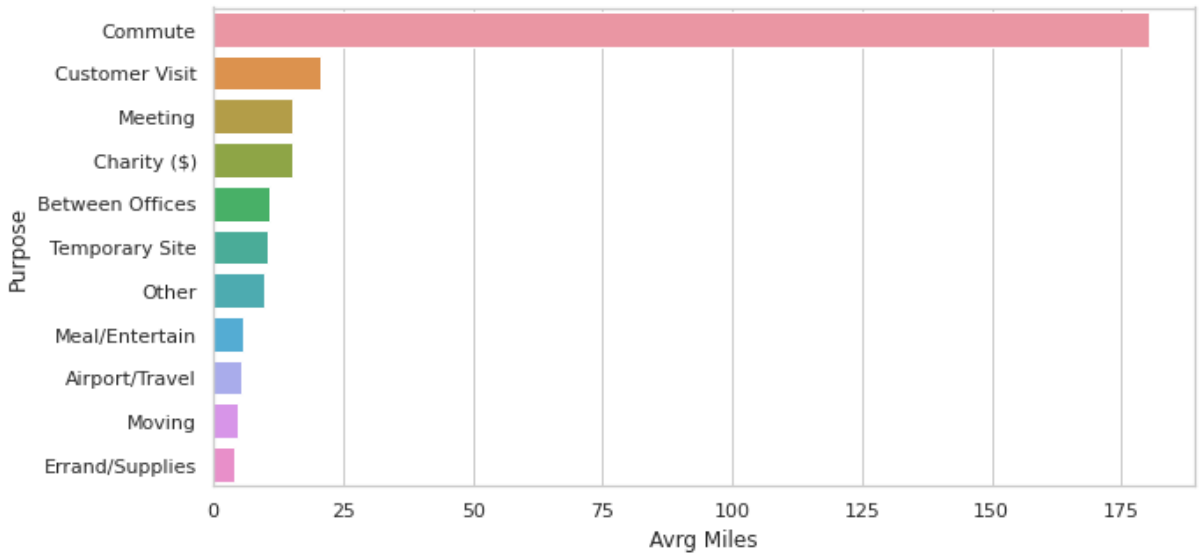
Purpose vs Rides



- The above graph depicts the number of uber rides availed for the given purposes.
- The x-axis shows the number of rides and the y-axis shows purpose of availing the uber ride.

- It is observed that most of the rides are for miscellaneous purposes, while some are related to work

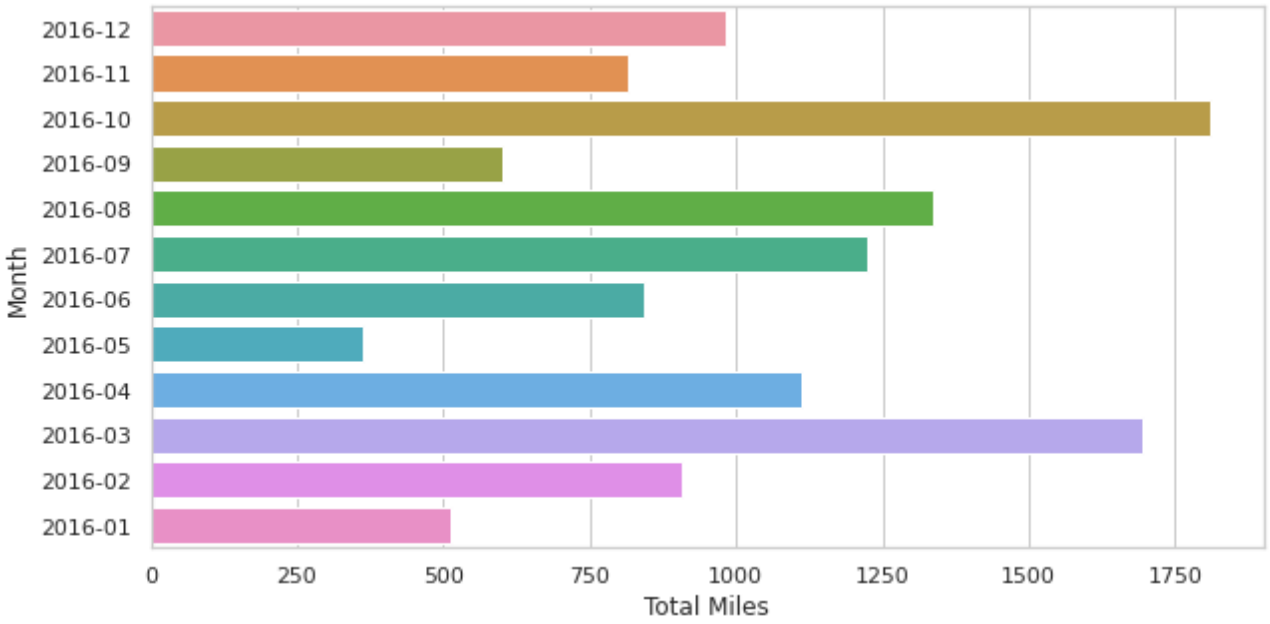
Purpose vs Miles



- The graph shows the average miles travelled by an uber for each purpose for travel
- The x-axis shows the average miles travelled and the y-axis shows the purpose for travel.
- It is observed that distance travelled due to work is relatively lower than normal commute distance.
- The average distance travelled by customers of each category:

PURPOSE*	MILES*
Airport/Travel	5.500000
Between Offices	10.944444
Charity ($)	15.100000
Commute	180.200000
Customer Visit	20.688119
Errand/Supplies	3.968750
Meal/Entertain	5.698125
Meeting	15.247594
Moving	4.550000
Other	9.748008
Temporary Site	10.474000

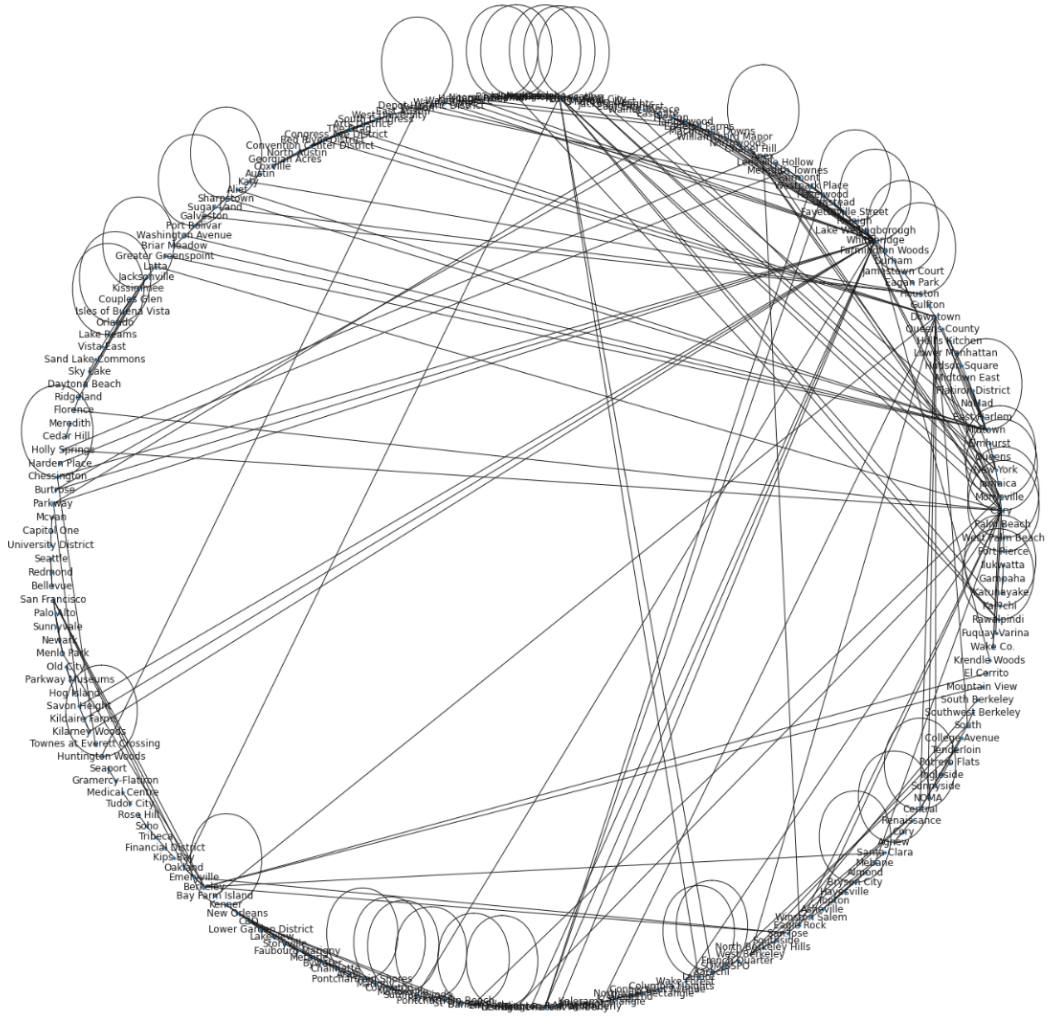
Total Miles per Month



- The graph depicts the total miles travelled through uber for the year 2016

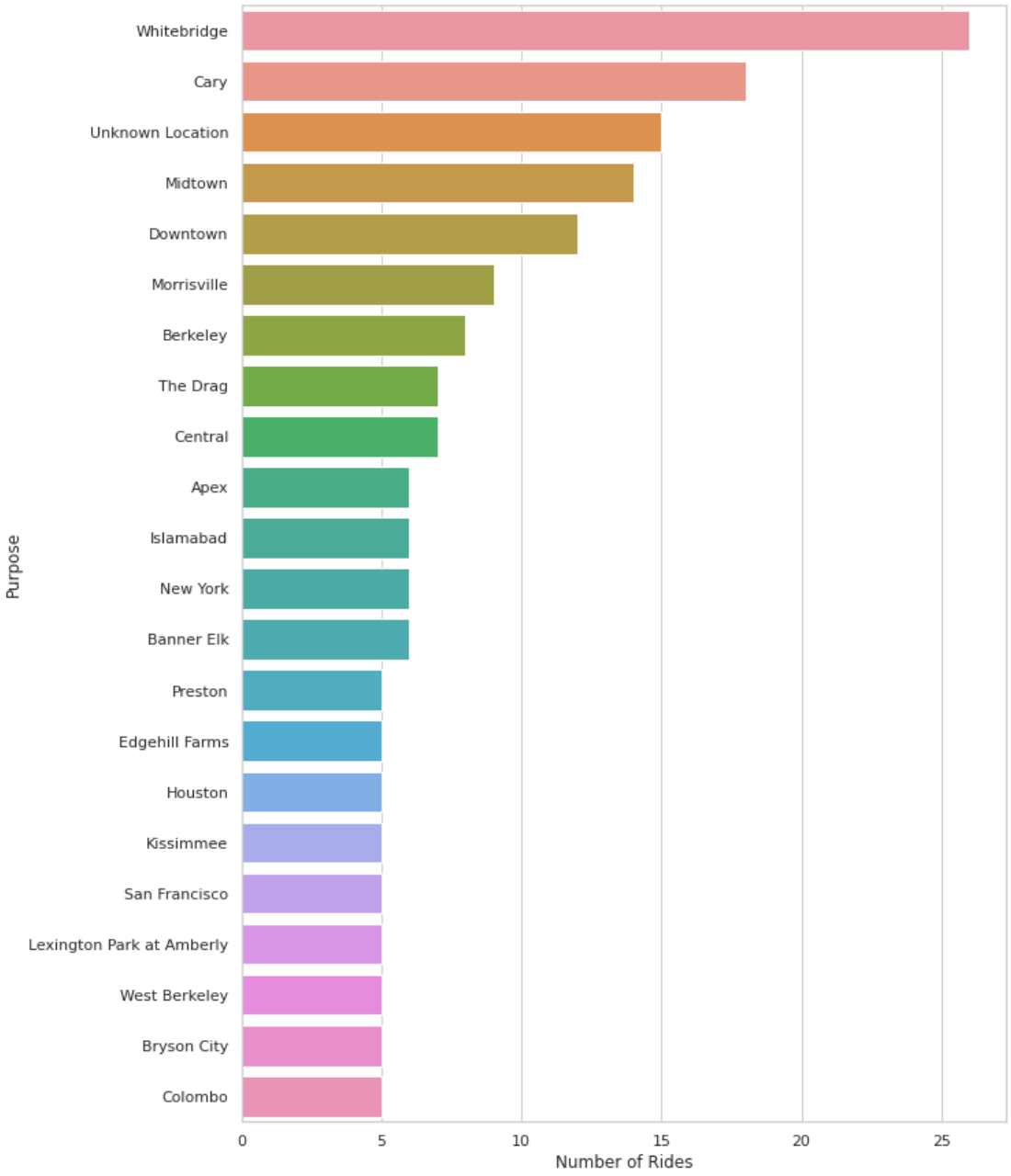
- It is observed that there was more demand for uber in the months of October and March.
- Halloween and Thanksgiving, two of the major festivals celebrated in the United States, is in the month of October. Thus, the miles travelled is highest for this month

Location Visited



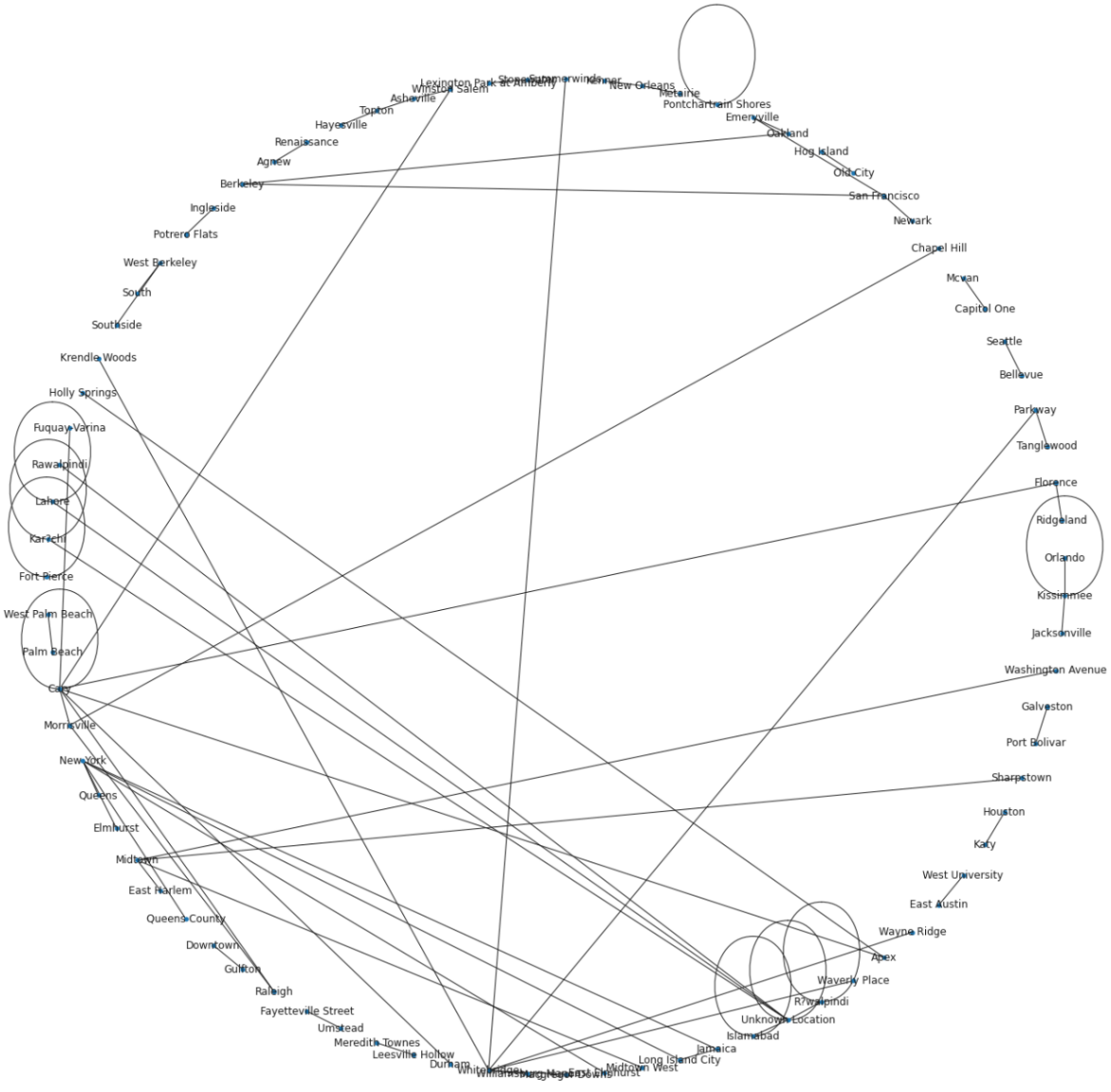
- The above graph shows the starting and ending locations that have been recorded in the dataset
- The graph is plotted using NetworkX package

Locations visited at least 5 times



- The graph displays the locations that have been visited more than 5 times are displayed in the graph.
- It is observed that Whitebridge and Cary have the most pick-up and drop-off compared to other locations.

Frequent Network



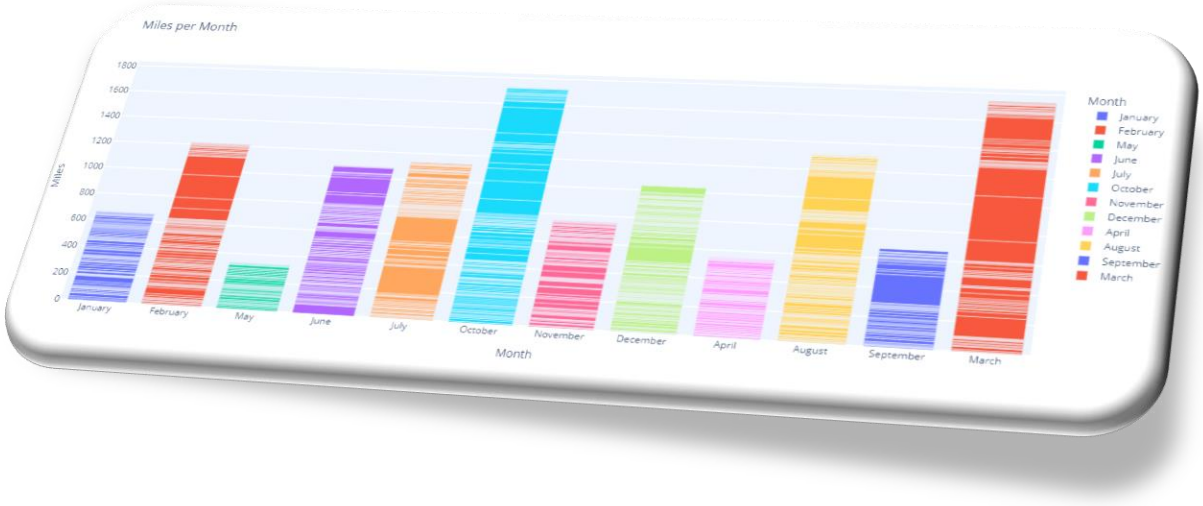
- In the previous graphs, we had notices that the highest number of uber rides were for meeting purposes.

Visualize Miles and Months:

	Week_Day	sum
0	Friday	2542.2
1	Monday	1386.2
2	Saturday	1674.6
3	Sunday	1767.8
4	Thursday	2315.5
5	Tuesday	1321.1
6	Wednesday	1197.3

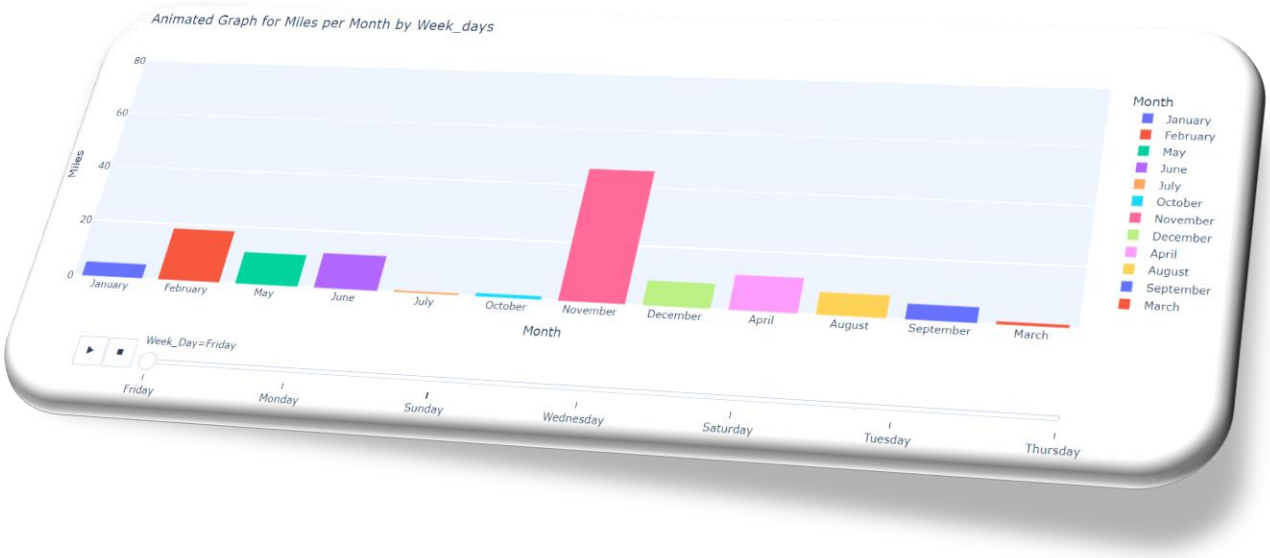
- A new Data frame is created, with the cumulative sum of the miles travelled in each day of the week.

Miles per Month



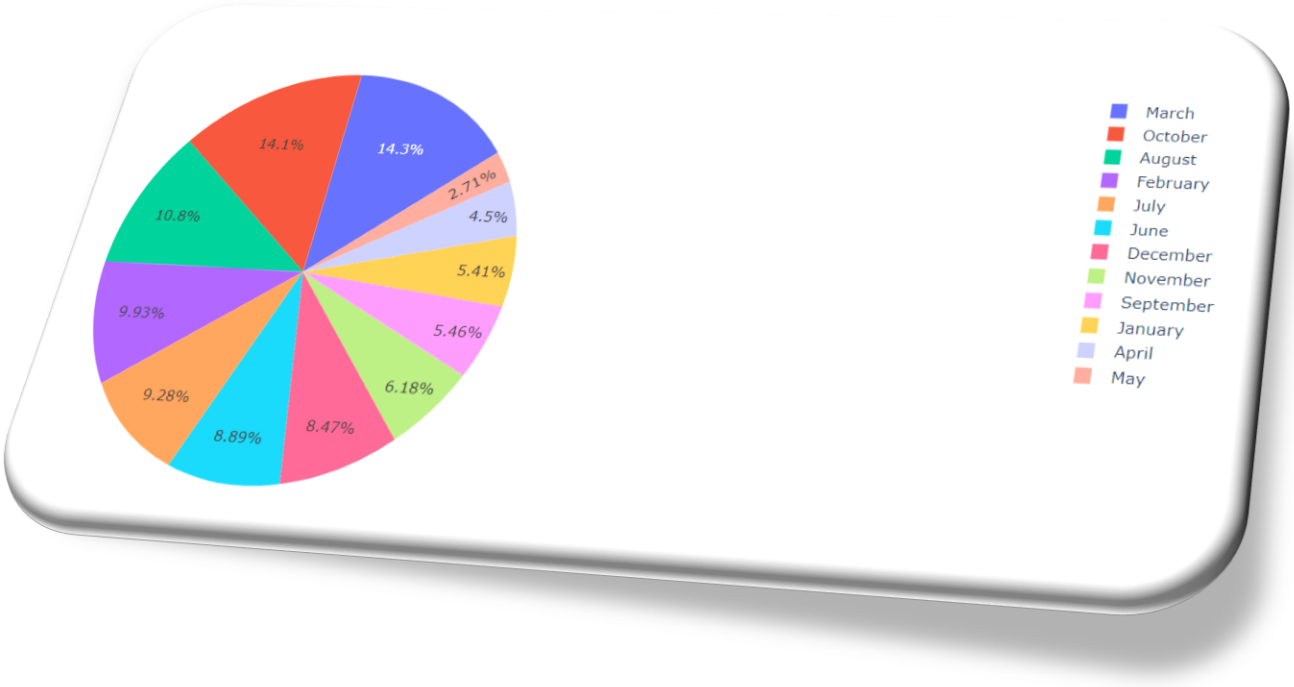
- The above graph represents the cumulative sum of the total miles travelled for each day of the week

Animated Graph for Miles per Month by Weekdays:



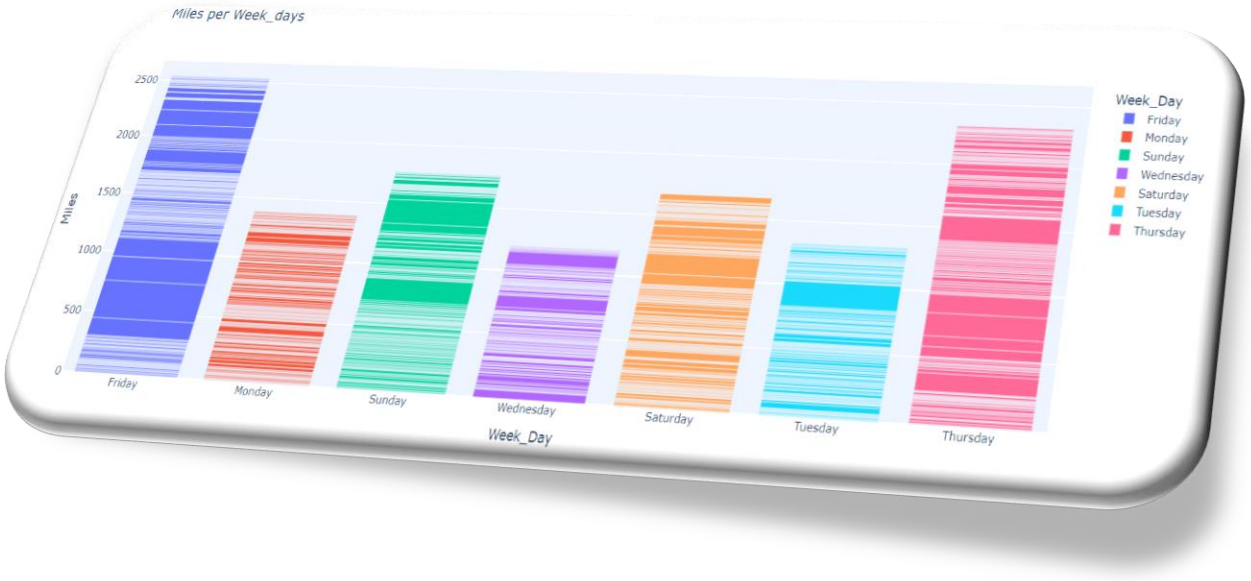
- The animated graph shows the miles travelled for each day of the week, for the entire year.
- The x-axis contains the months of the year and the y-axis represents the total miles travelled.
- A bar is created below the graph, representing the days of the week. The cursor can be dragged to select the particular day.
- The animated graph is plotted with the help of ***plotly.express.***

- Pi Graph For Miles Per Month:

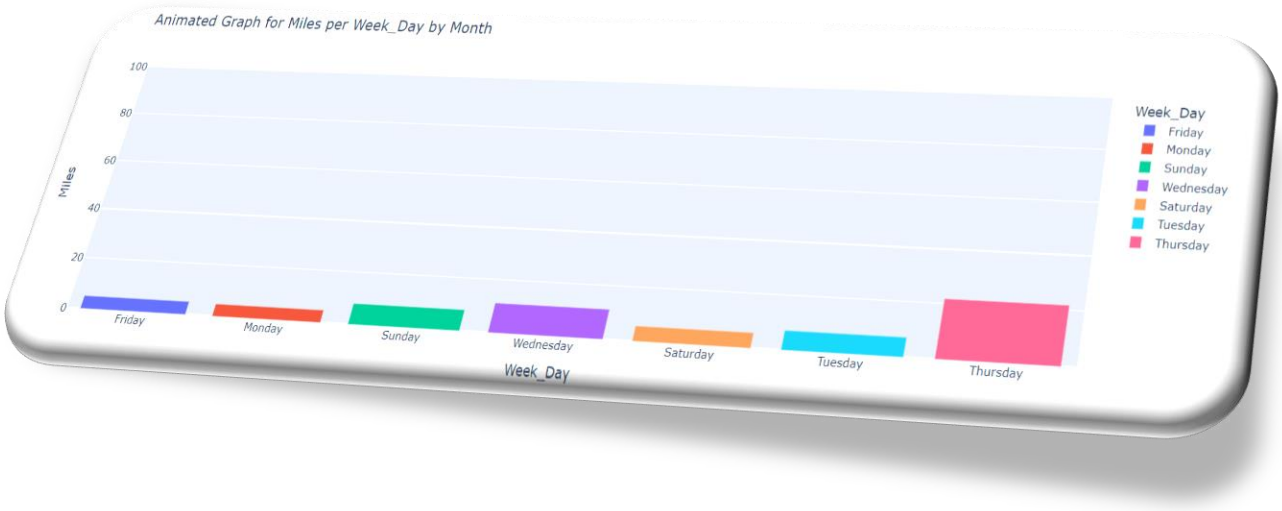


- The above pie graph depicts the percentage share of each month in the total miles travelled.
- It is observed that March has the highest percentage of 14.3%, closely followed by October having 14.1%.

Visualize "Miles" and "Week Day"

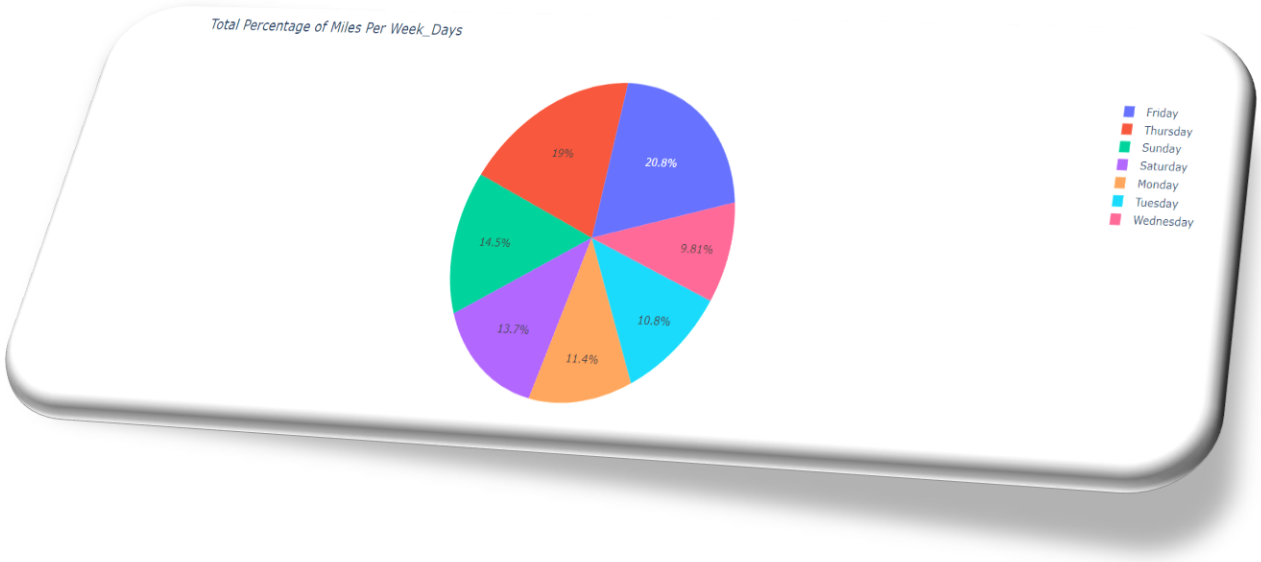


- The graph denotes the miles travelled through uber per week day.

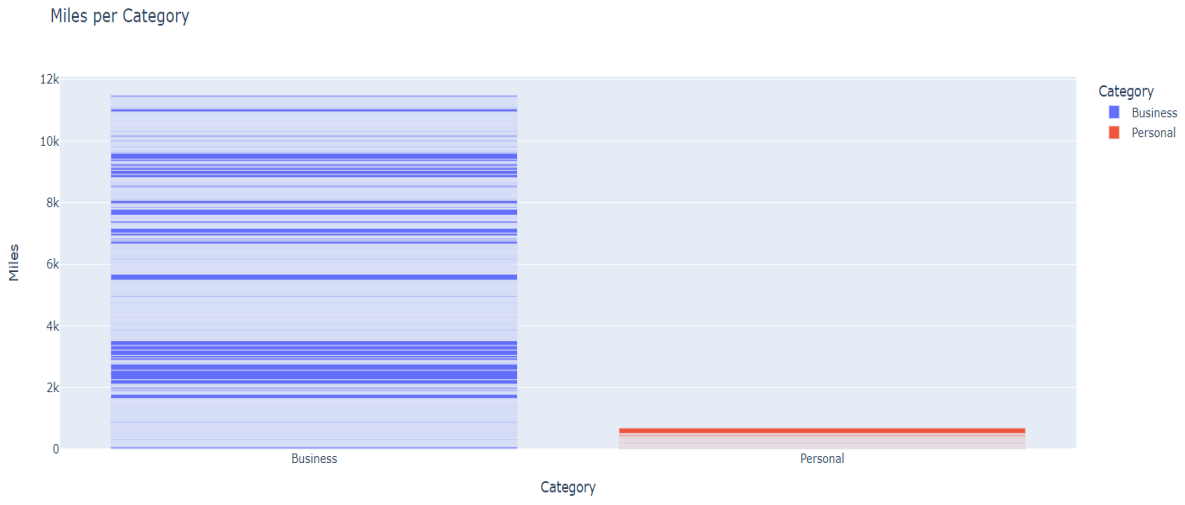


- The graph depicts the miles per week day, by each month.

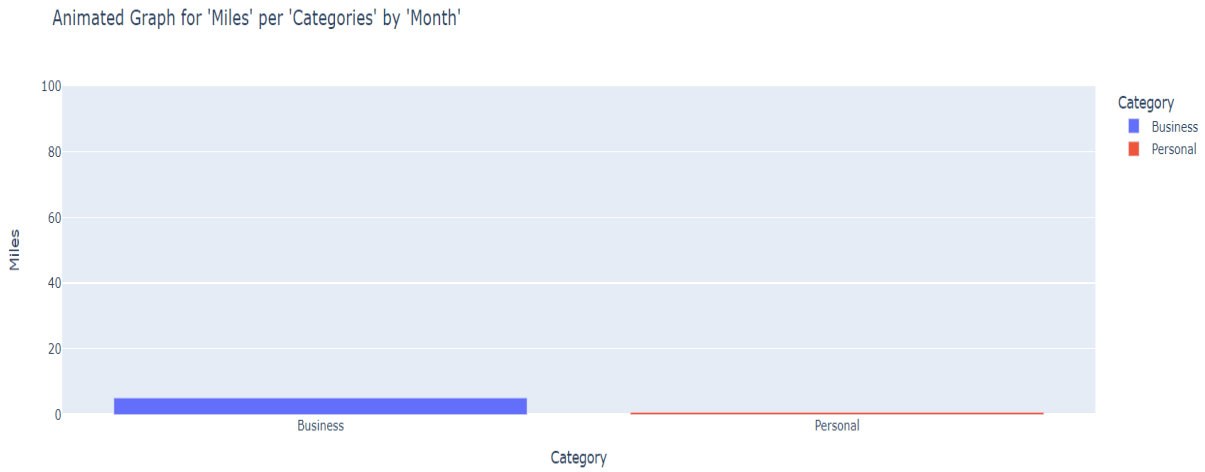
Pi Graph of "Miles" Per "Week Day"



- The pie graph shows the distribution of total miles travelled by their week day.
- It is observed that maximum number of uber rides is availed on Friday's, and the least number of rides is availed on Wednesday's.

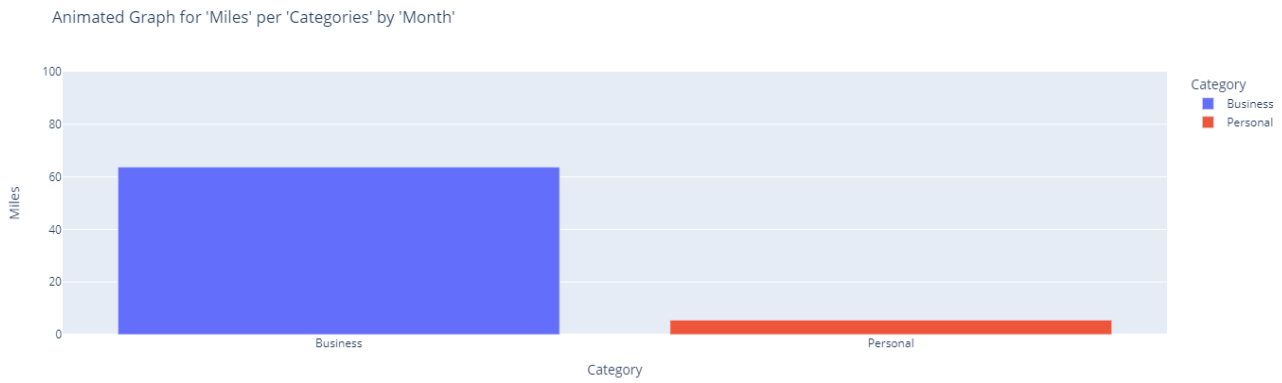


- The above graph shows the miles travelled per uber ride for Business and Personal Travel
- It is observed that the distance travelled for a Business Trip is much more when compared to that of a Personal Trip

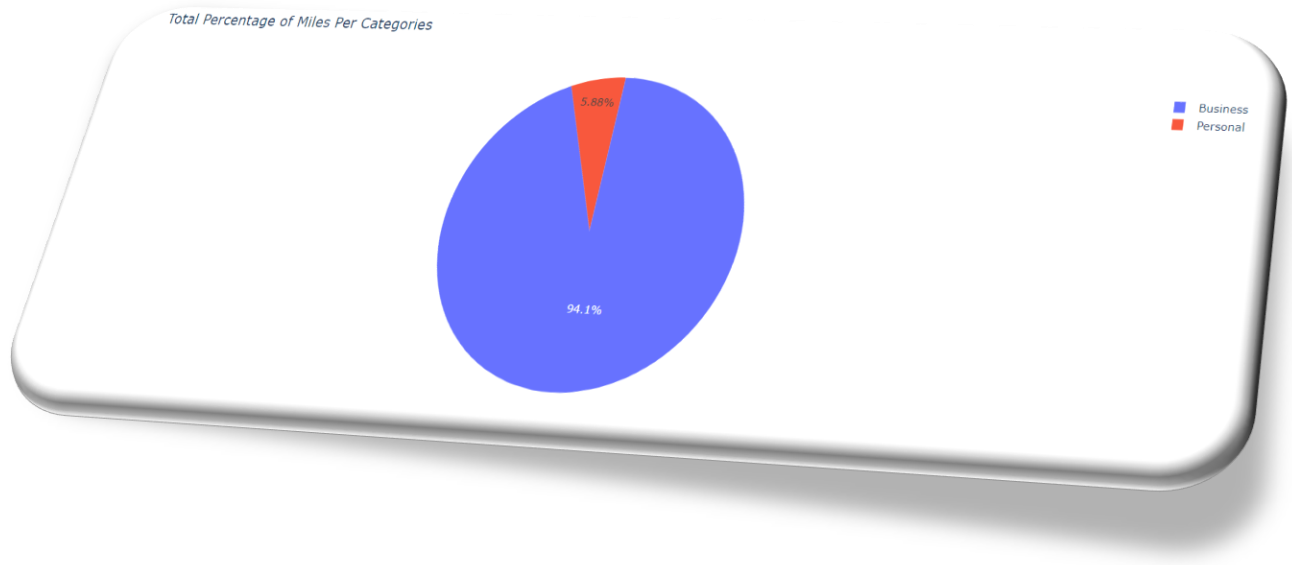


- The graph is an animated graph, which depicts the miles travelled per month for either Business or Personal Trips.
- It is observed that for the month of June, there is a significant hike In the total miles travelled for a Business purpose, whereas in the month of July, similarly there is a hike in the total miles travelled for Personal reasons.

Relation of "Miles" and "Category" per "Week_Day"

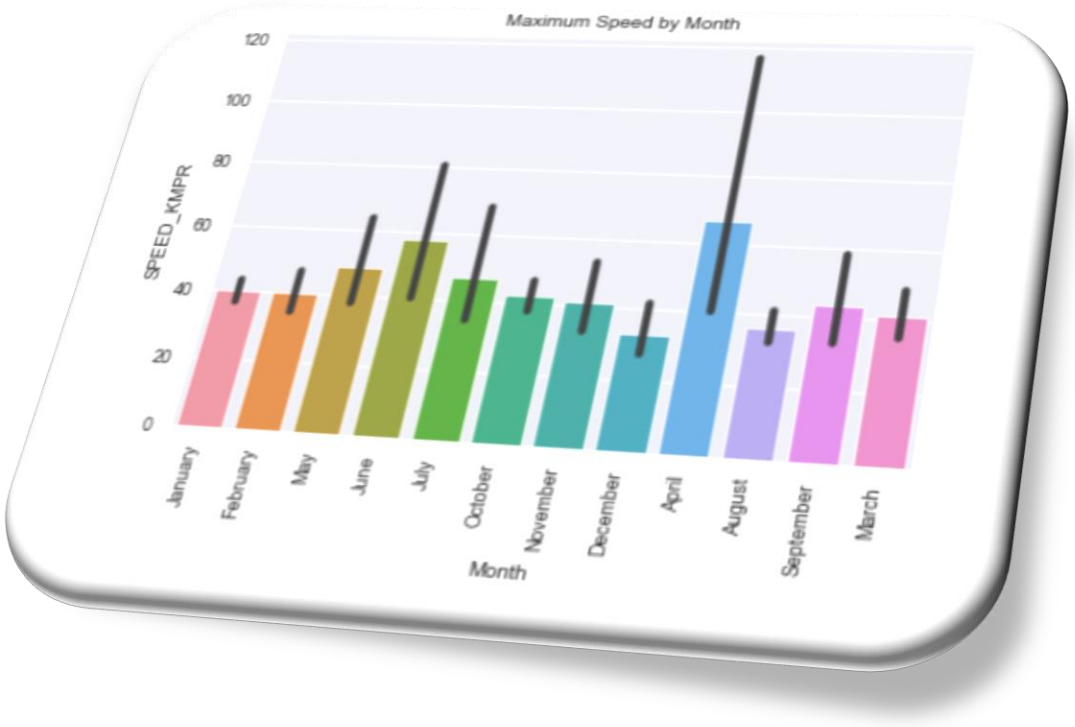


- The animated graph depicts the cumulative miles travelled per week day.

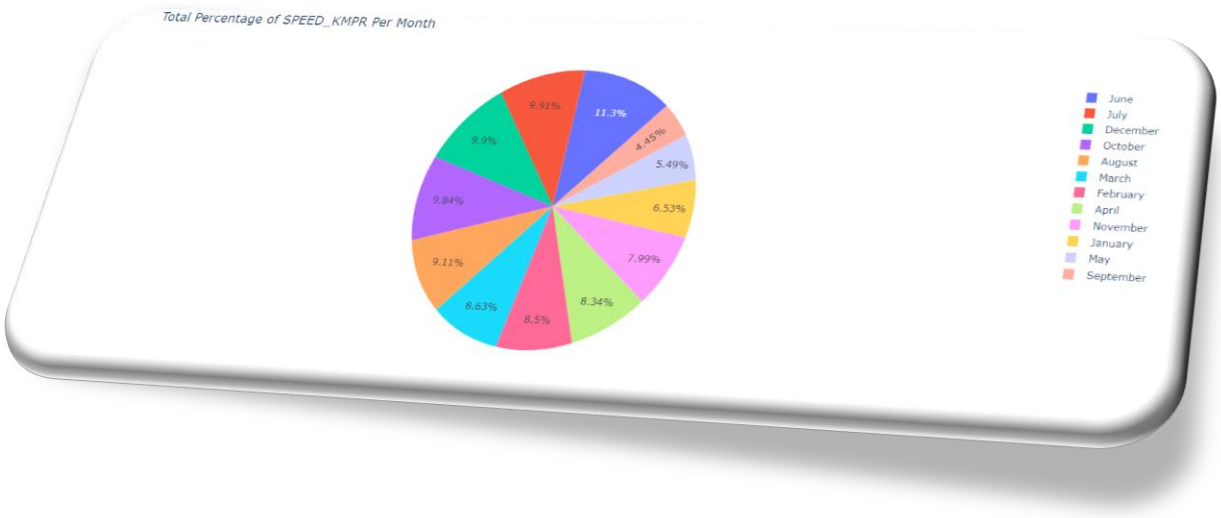


- The pie chart displays the percentage per categories in the total miles travelled in an Uber.

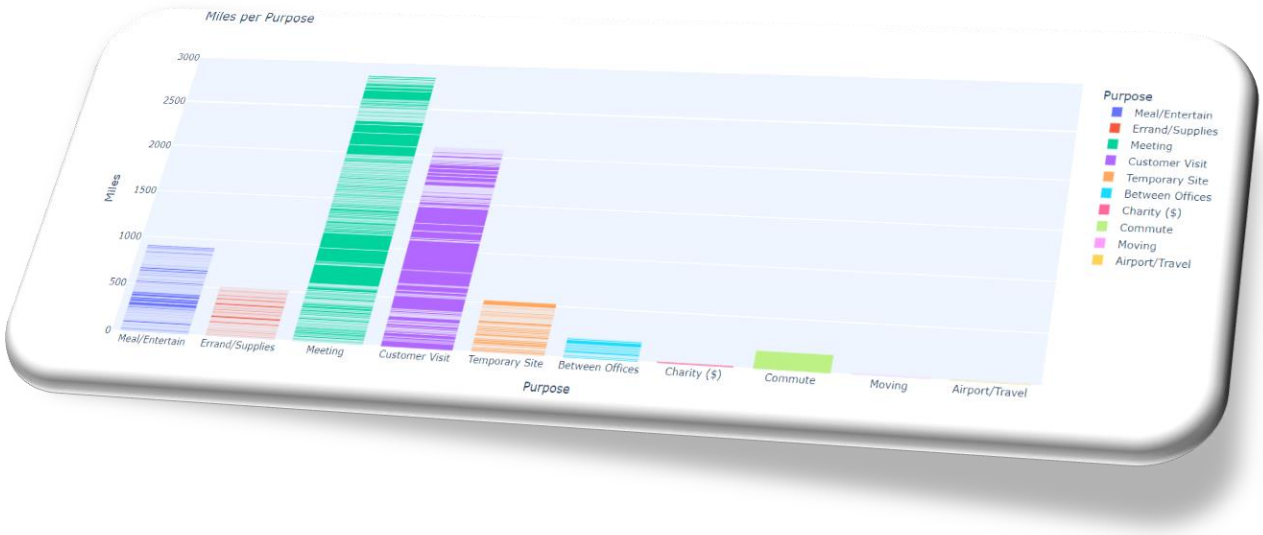
- A big chunk of the total miles travelled is for Business purposes (94.1%). Personal travel forms just 5.8% of the total miles travelled.



- The above graph displays the maximum speed of an uber taxi recorded in the months of 2016.
- Max Speed of 65 km/hr was recorded in the month of April, and the Least Speed of 35 km/hr was recorded in the month of December.

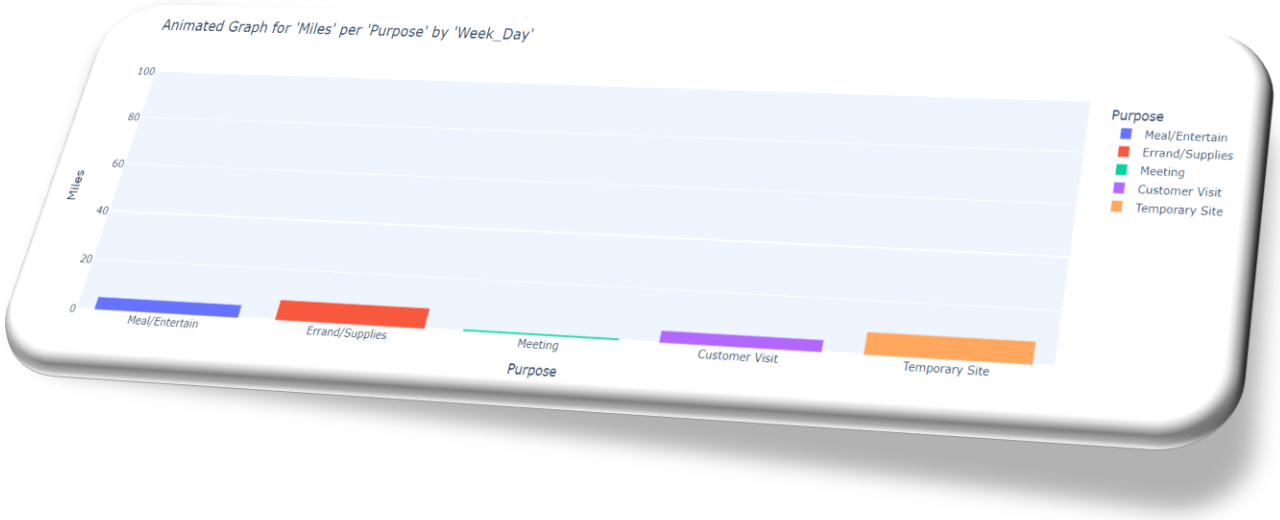


- The pie graph denotes the Speed in kilometer/hour per month.



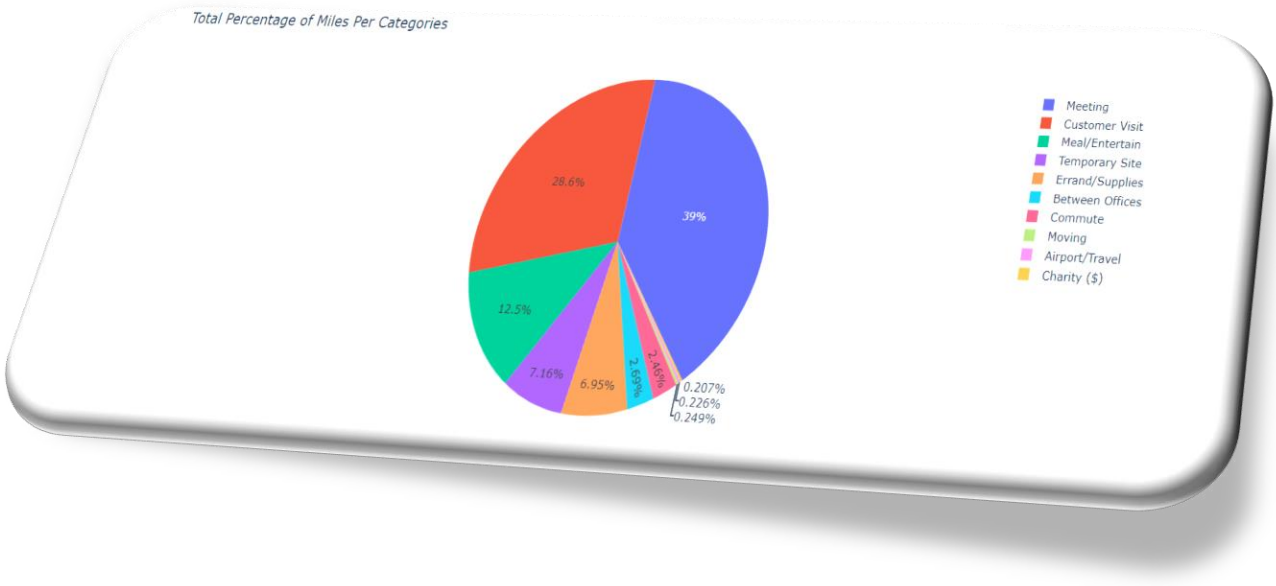
- The above Bar Plot shows the miles travelled for a particular purpose
- The x-axis denotes the Purpose for Travel and the y-axis denotes the miles travelled.

- It is observed that travelling for meetings and customer visits are the reasons for travelling the most distance.
- Airport travel, Regular Commute and Moving only form a small portion of the total miles travelled, as the distance from source to destination is relatively small.



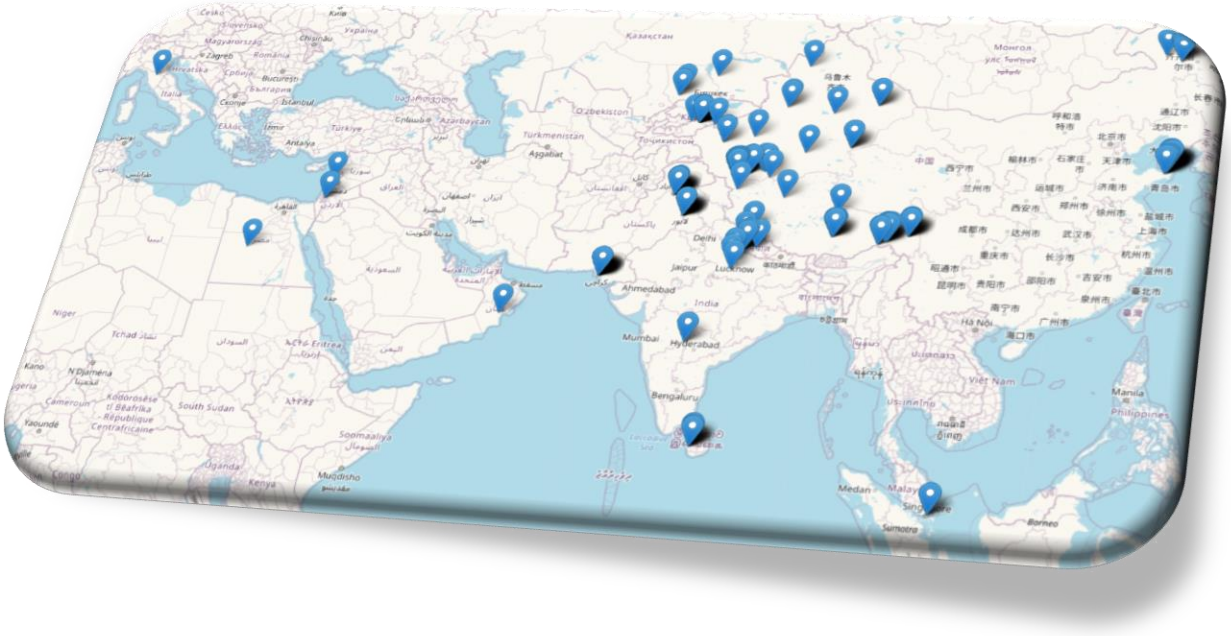
- The animated graph represents the miles travelled for some purpose, considering each day of the week.

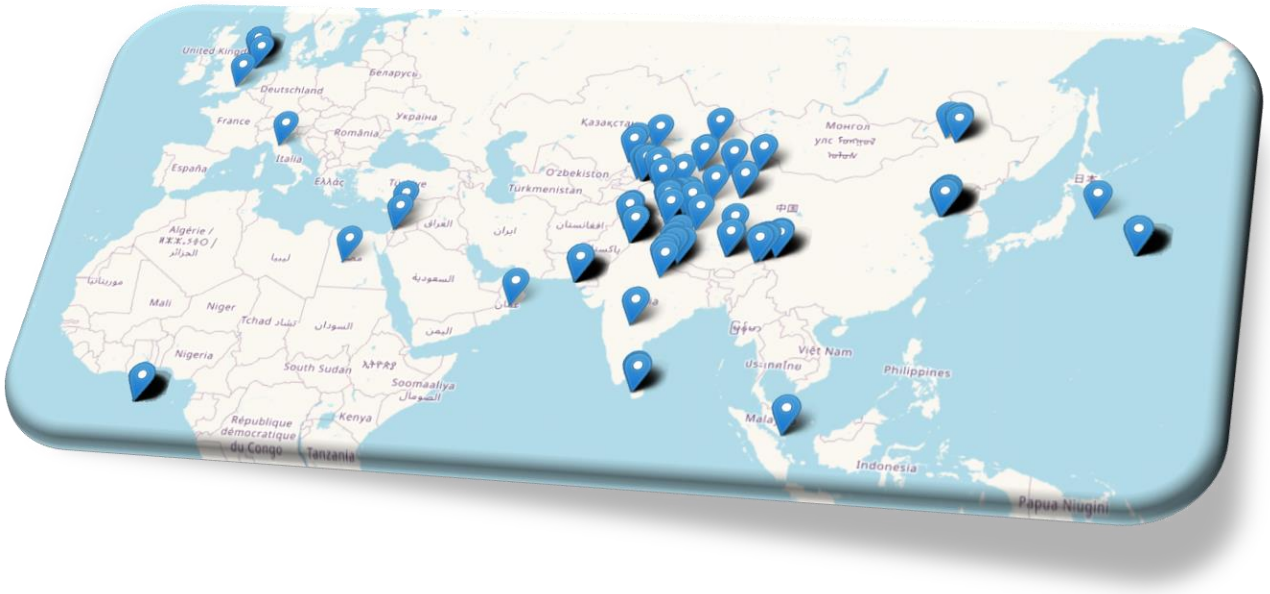
- The animated graph represents the miles travelled for some purpose, considering each day of the week.



- The Pie Graph represents the percentage share of each category in the total miles travelled.

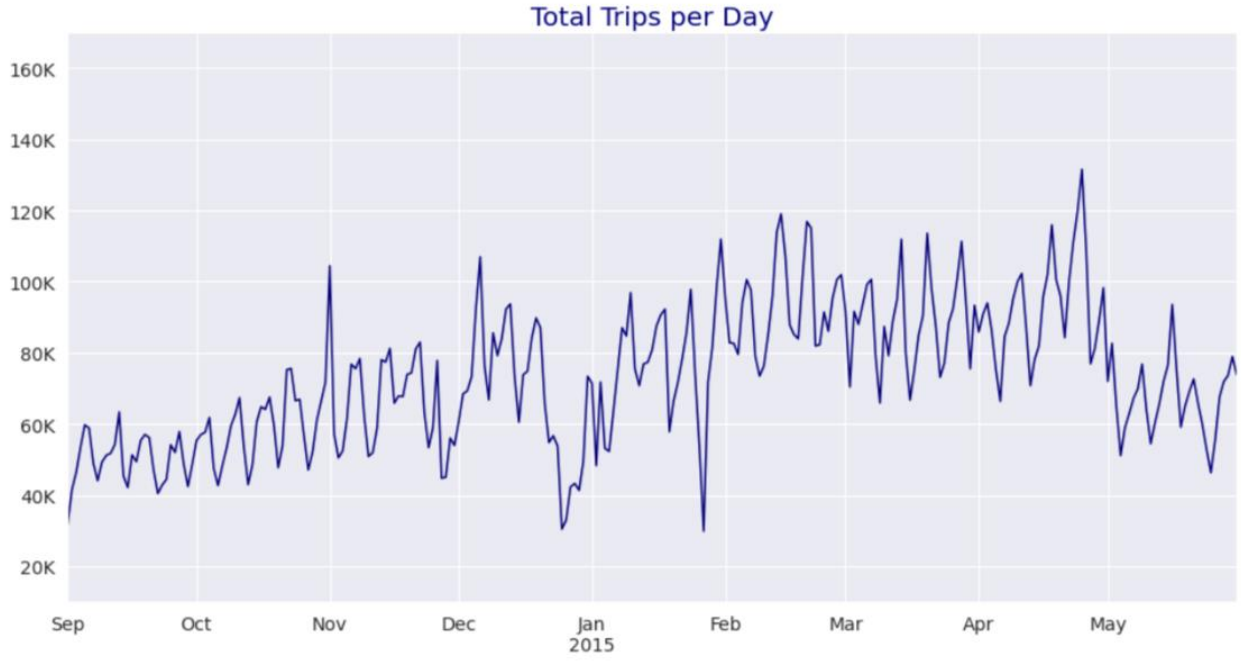
Location Mapping





Visualization

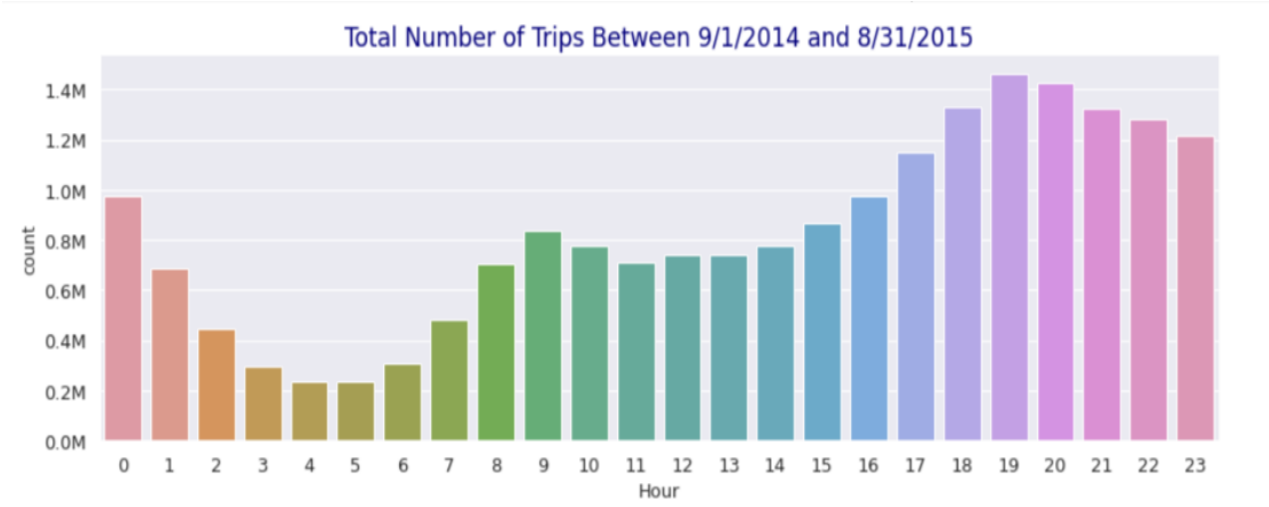
Total Trips Per Day



- In this plot we can see which month there are highest and lowest number of trips.
- It is evident that in the month of May there has been highest number of trips whereas in the month of Feb there has been lowest number of trips.
- It highlights some changepoints associated with major holidays and other weather and touristic/cultural events.

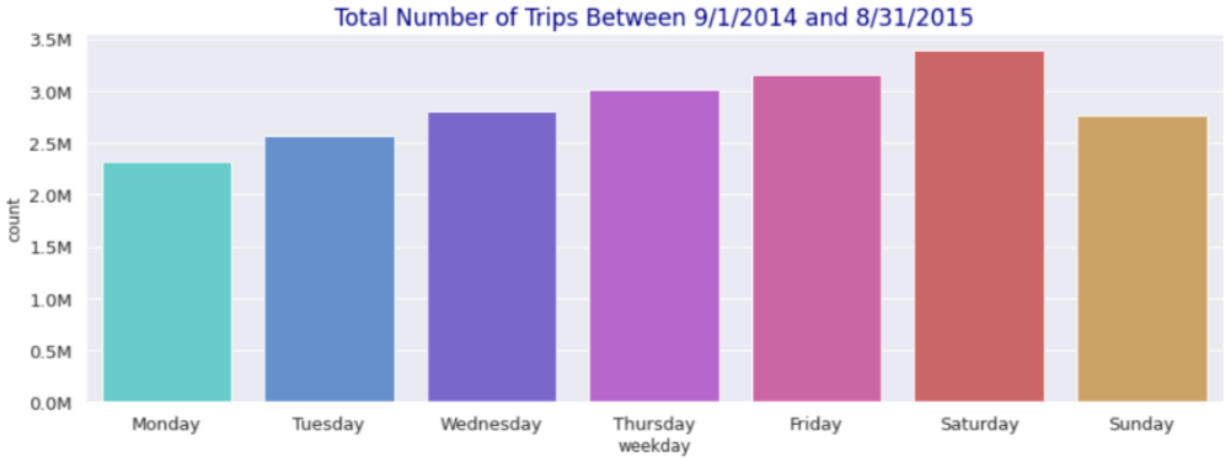
🜸 Total Number of Trips in Hours

- The effect of time on demand for Uber rides distribution per hour, weekday, and month.



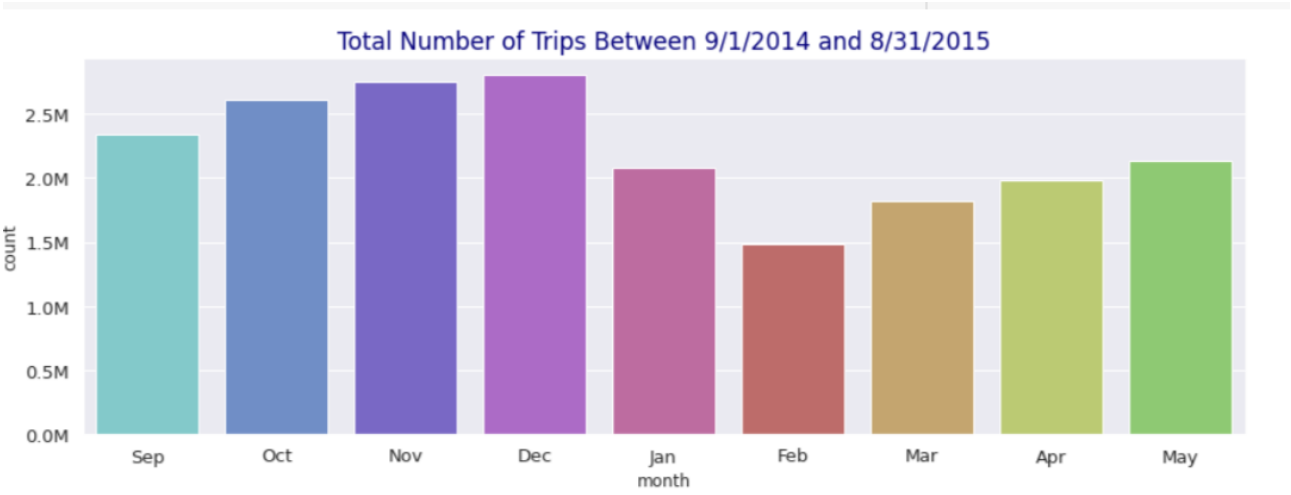
- We can see that in the plot of hour and count the count has been on a higher end in the later part of the day whereas in the morning it is low.

Total Number of Trips in Weekdays



- Highest number of trips are happening on Sunday and the lowest is happening of Monday.

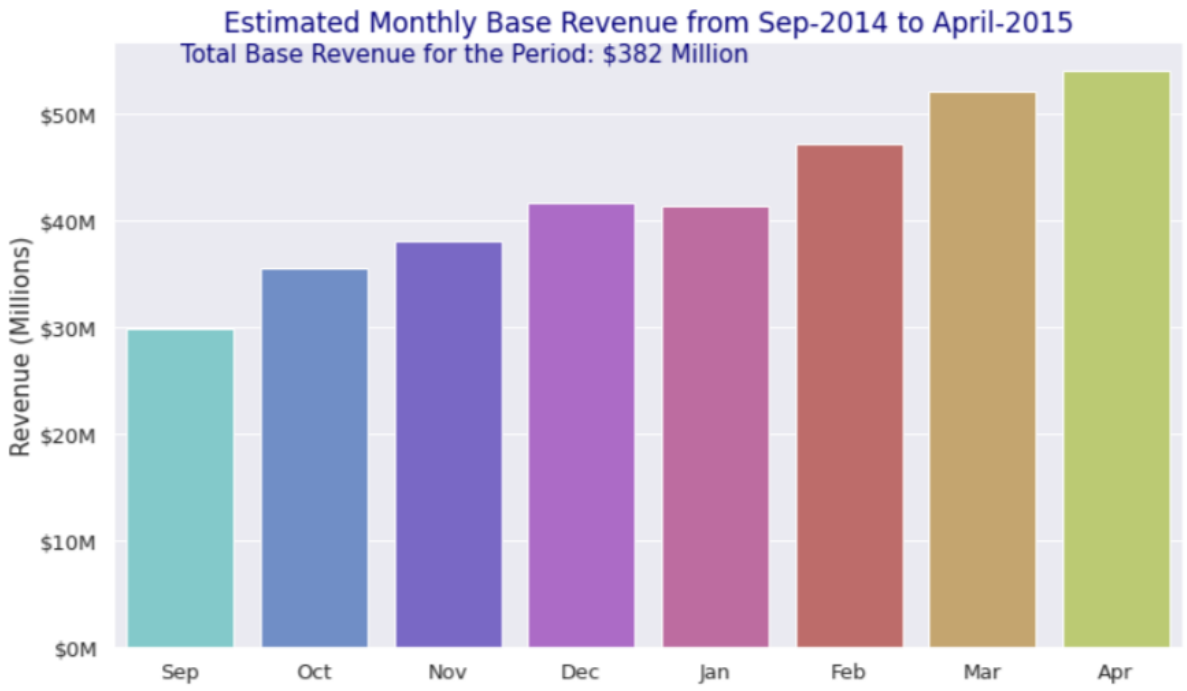
Total Number of Trips in Month



- We can see over here that during the month of December the trips have been the highest.

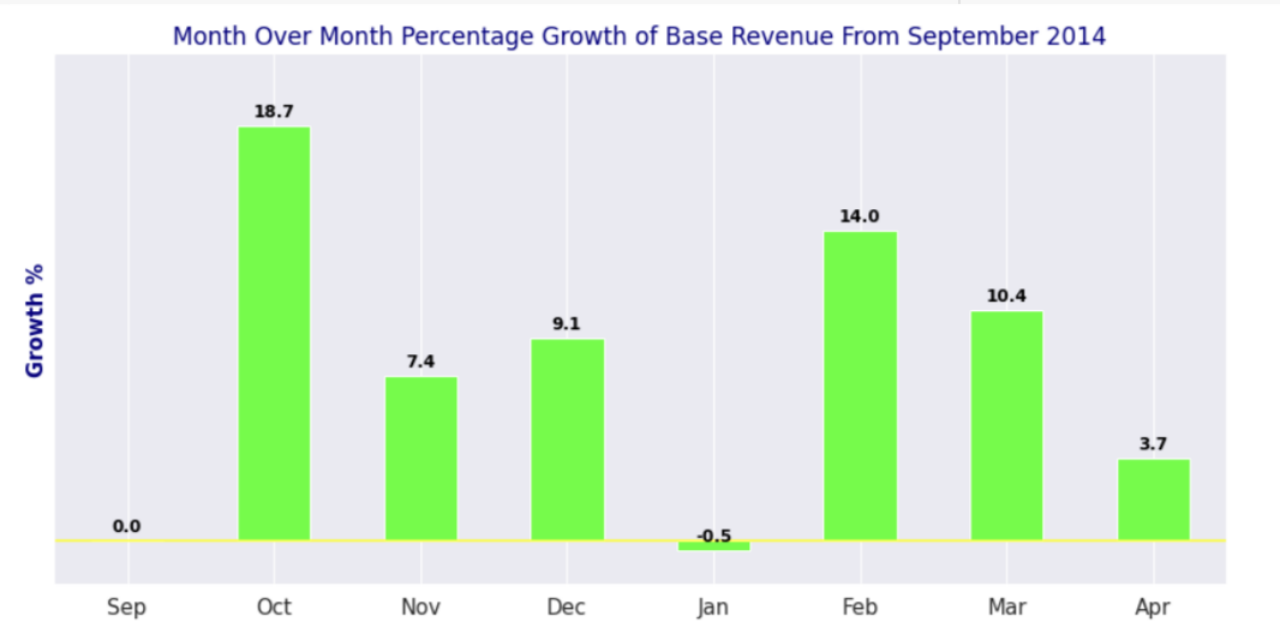
- During the month of Feb number of trips have been the lowest.

- **Total Base Revenue Per Month**



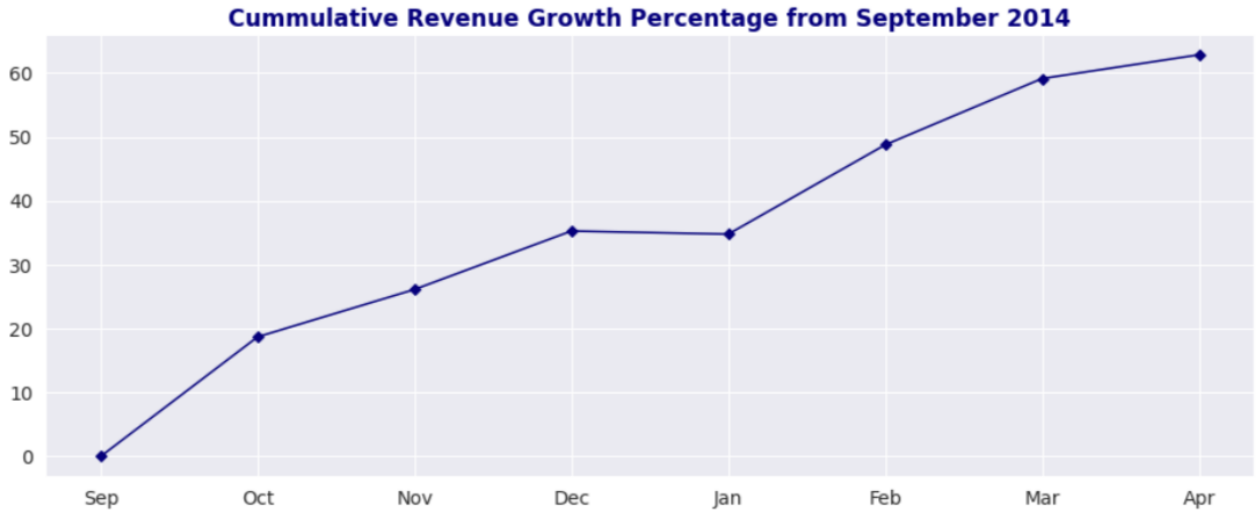
- The highest total revenue per month is during the month of April.
- And the lowest total revenue per month is during the month of Sep.

Total Cumulative Growth



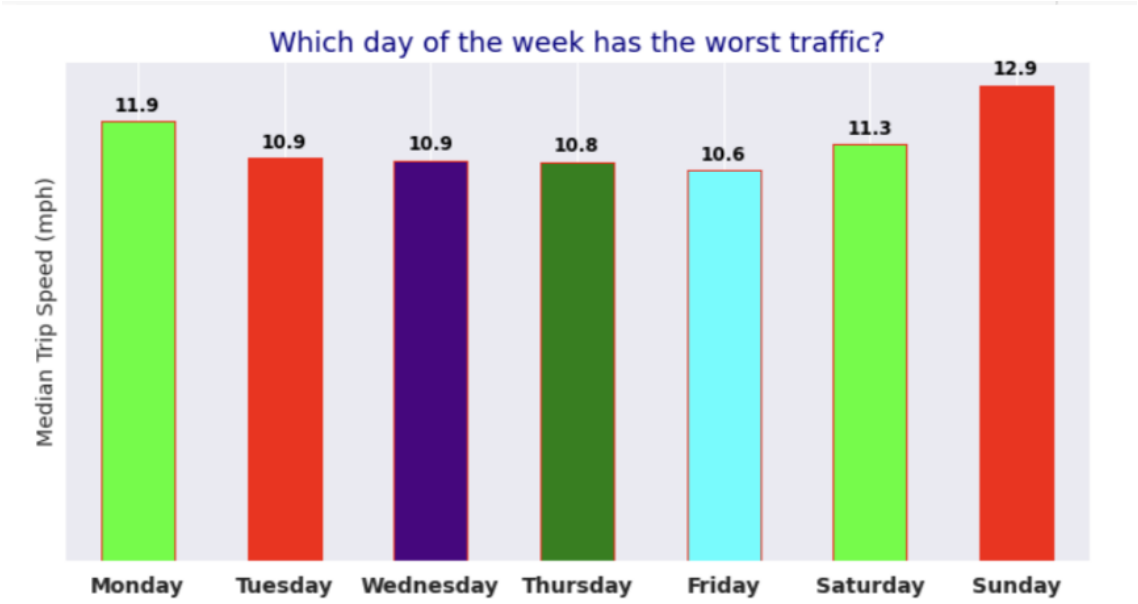
- The highest Growth % is seen in the month of October and lowest has been seen In the month of Sep as well as Jan.

Cumulative Base Revenue Growth



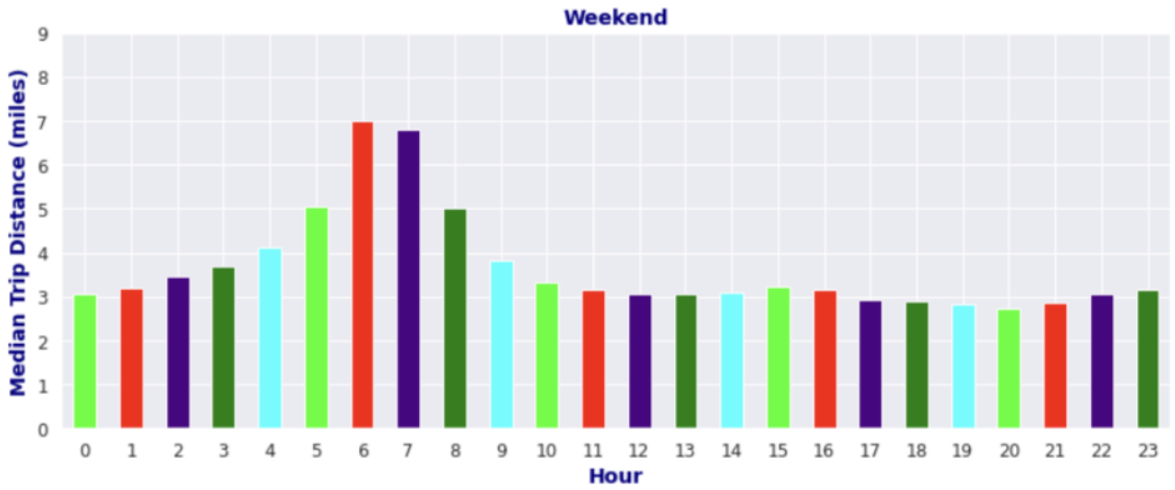
- This is a plot between Months and Revenue Growth.
- Its visible that the revenue has increased during dec and January there was a flat curve apart from that it has shown continuous growth.

Traffic Analysis



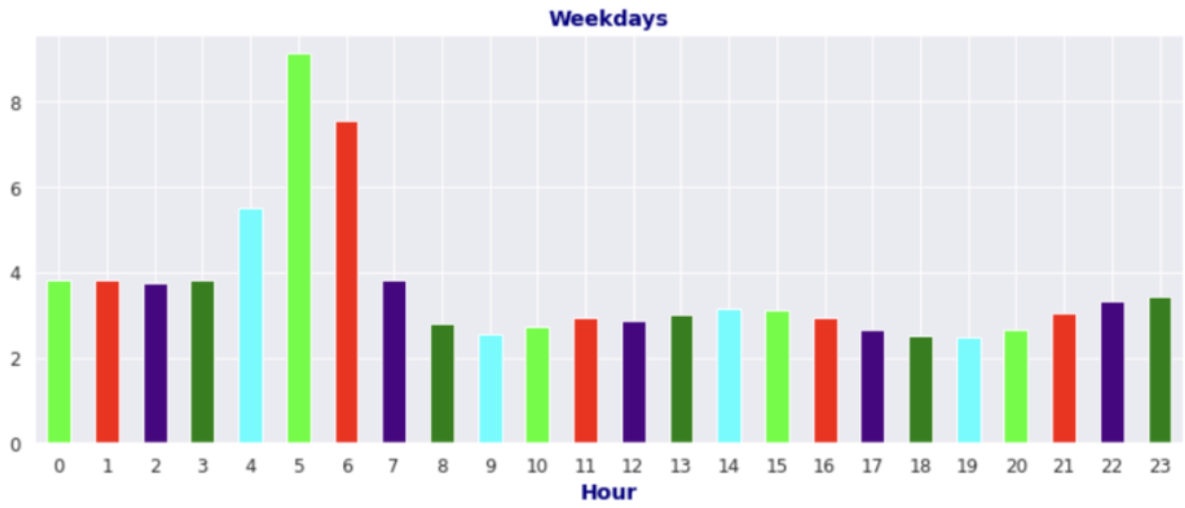
- Here we are analyzing the traffic with median trip speed.
- The highest being on Sunday.
- The lowest being on Friday.

Median Distance Traveled Per Trip



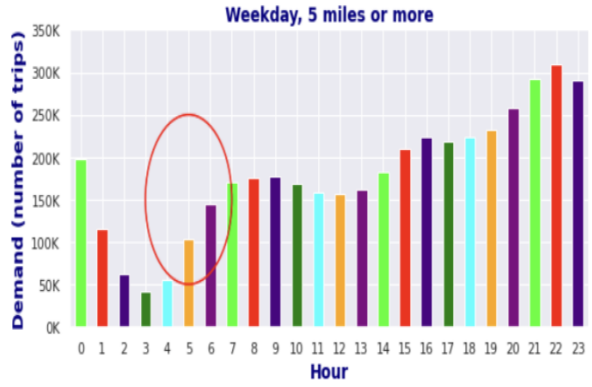
- Here we have plotted hour with Median Trip distance.
- After careful observation we can see that maximum is during the morning while at other hours its nearly the same.

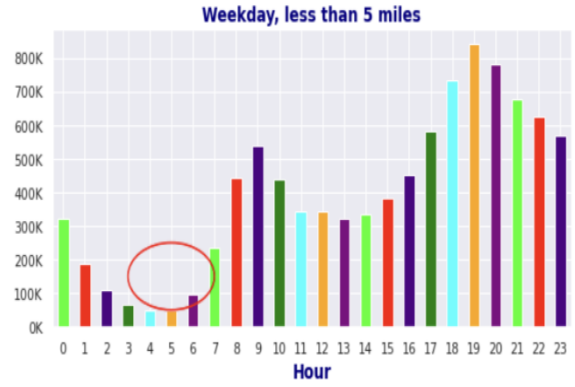
Week Days

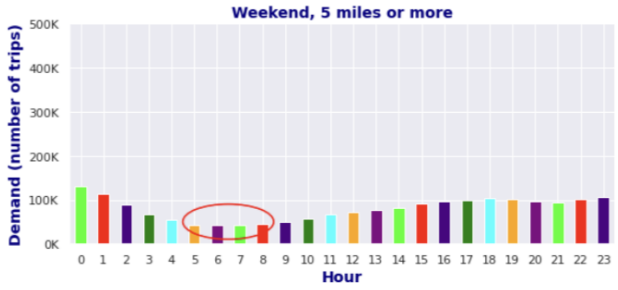


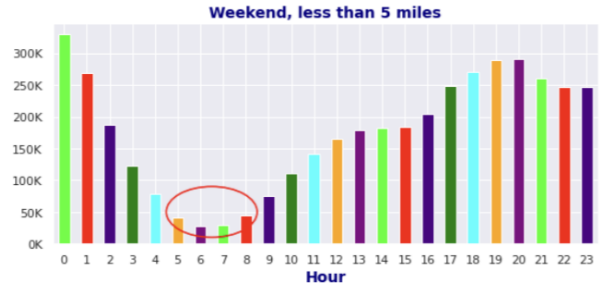
- On week days the highest median trip distance is at 5Am and the lowest being at 7pm in the night.

Count of Trips Per Hour



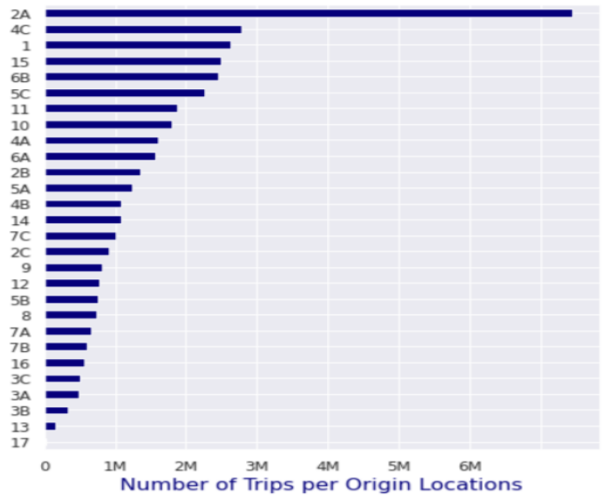


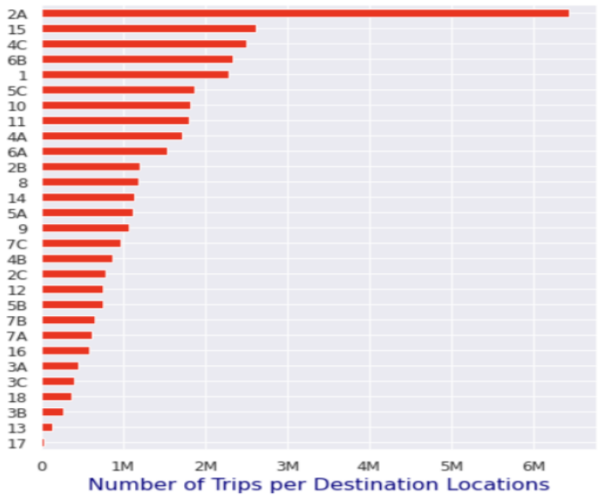




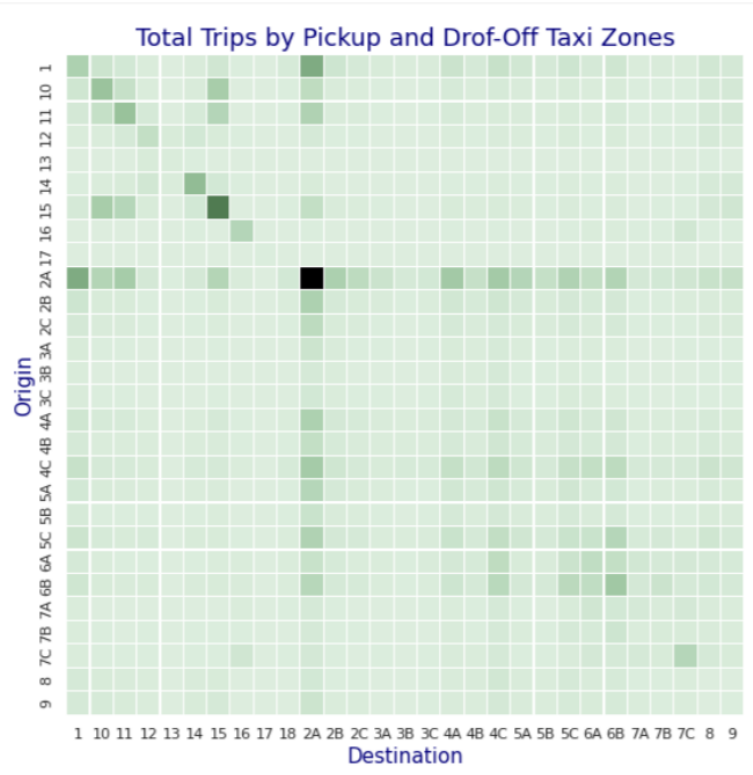
- It is quite evident from the graph that the lowest demand is during the morning hours.
- Whereas the highest demand being the evening.

Pickup and Drop - Off Zones





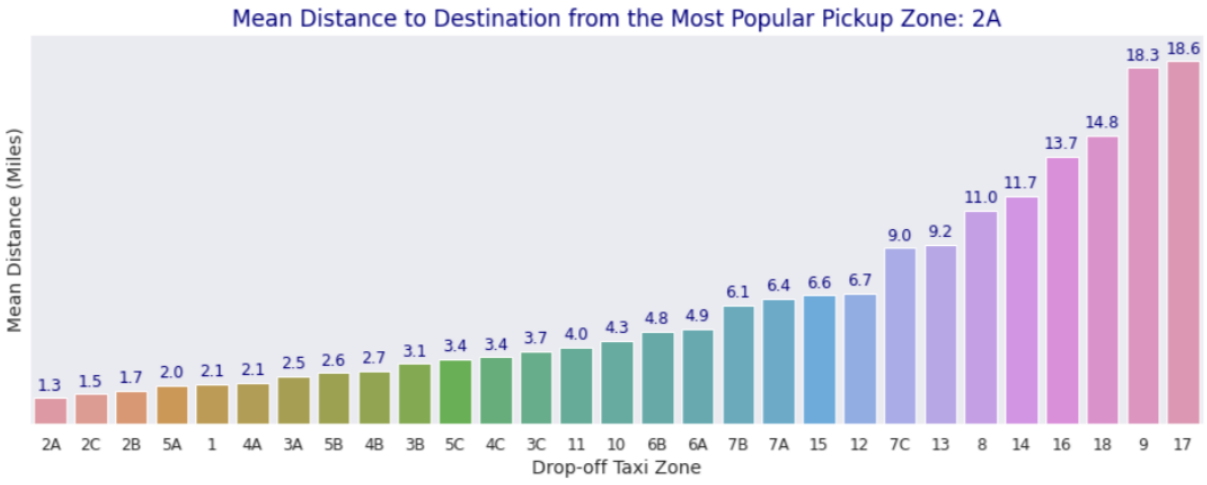
- Here we are plotting number of trips with origin locations.
- And also, number of trips with destination locations.



- We can see that the location where maximum number of picks ups and drop offs have taken place is 2A and the second highest seems to be 15 and others are having nearly same.
- The darker the small square box means the greater number of trips while the lighter it is means the least number of trips have either originated or haven been the final destination of the trip.

Mean Distance to Destination

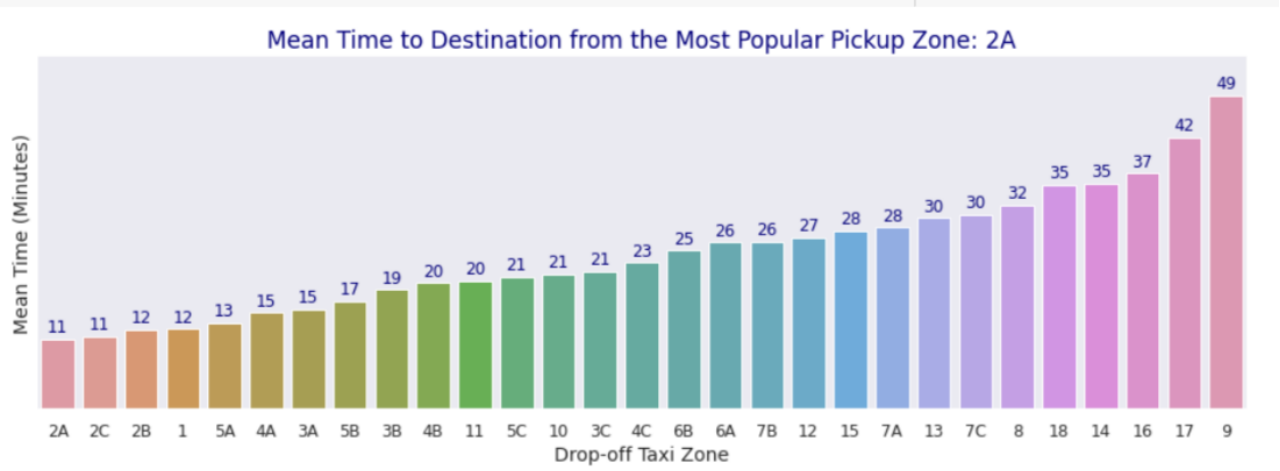
- Mean distance to destination from the most popular pick-up location



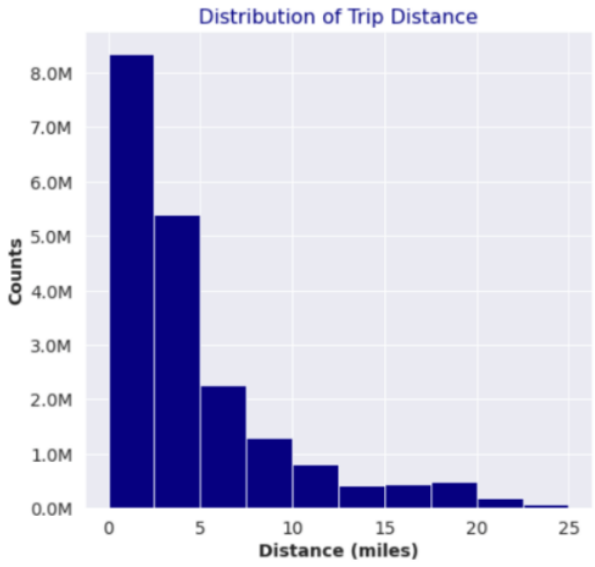
- The lowest mean distance was from drop off taxi zone 2A.
- The highest mean distance was from drop off taxi zone 17.

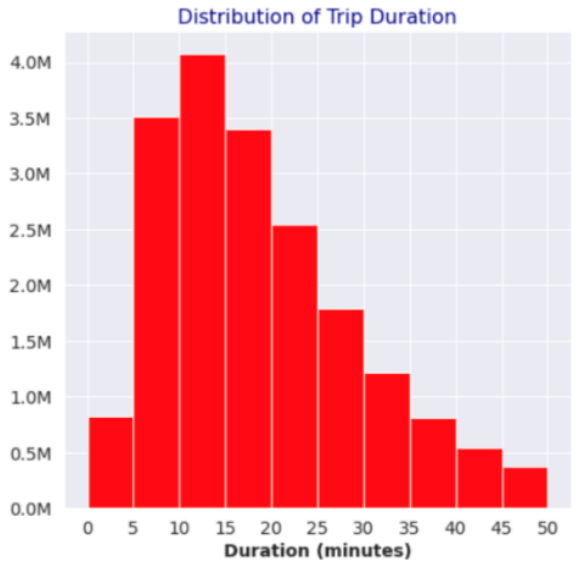
Mean Time to Destination

- Mean Time to destination from the most popular pick-up location



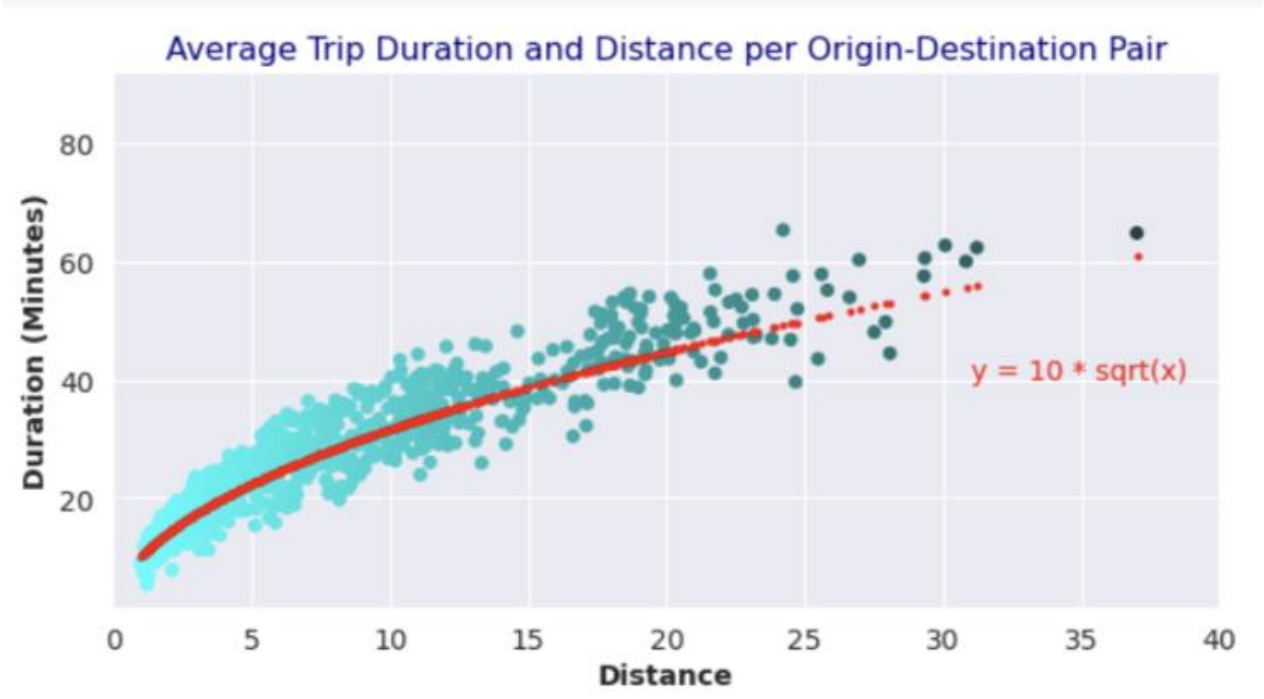
- The highest mean time was observed from taxi zone 9.
- The lowest mean time was observed from the taxi zone 2A.





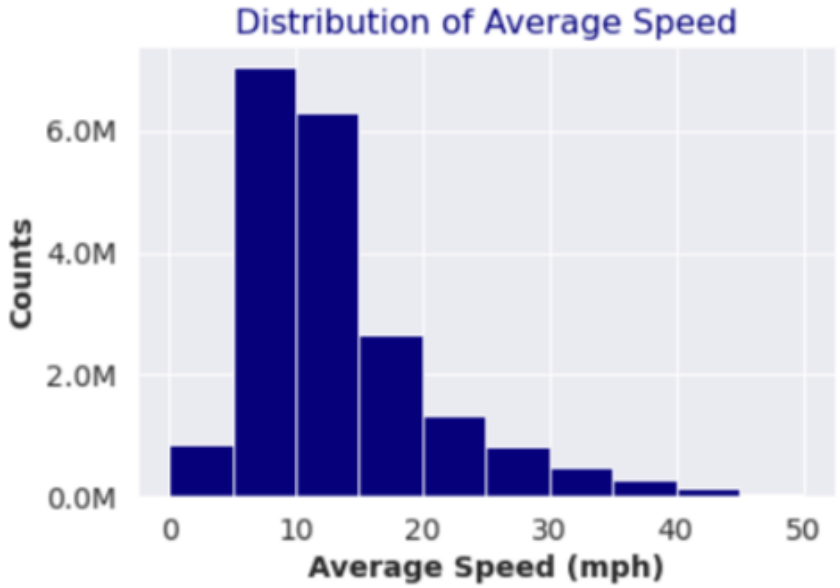
- We can see that the maximum number of trips are happening for 0-5 miles.
- And as the number of miles increase the number of trips also decreases.
- As far as minutes vs count is concerned, we can observe that the maximum number of trips are happening when the trip duration is between 10 and 15 minutes.
- And the minimum is when the trip duration is either less than 5minutes and more than 45 minutes.

Trip Distance vs Trip Duration



- Here on the x-axis, we are plotting the distances and, on the y-axis, we are plotting the Duration in (Minutes).
- This graph shows us the average trip duration and distance per Origin-Destination pair.

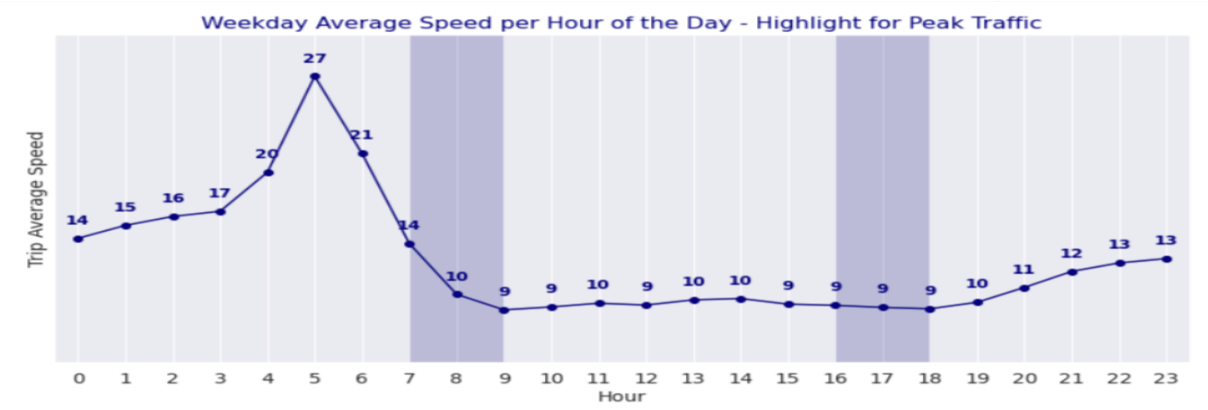
Average Speed Per Trip



- The highest average speed was found to be at 10 mph and the lowest average speed was found to be between 40 and 50 mph.

- Average Speed Per Hour

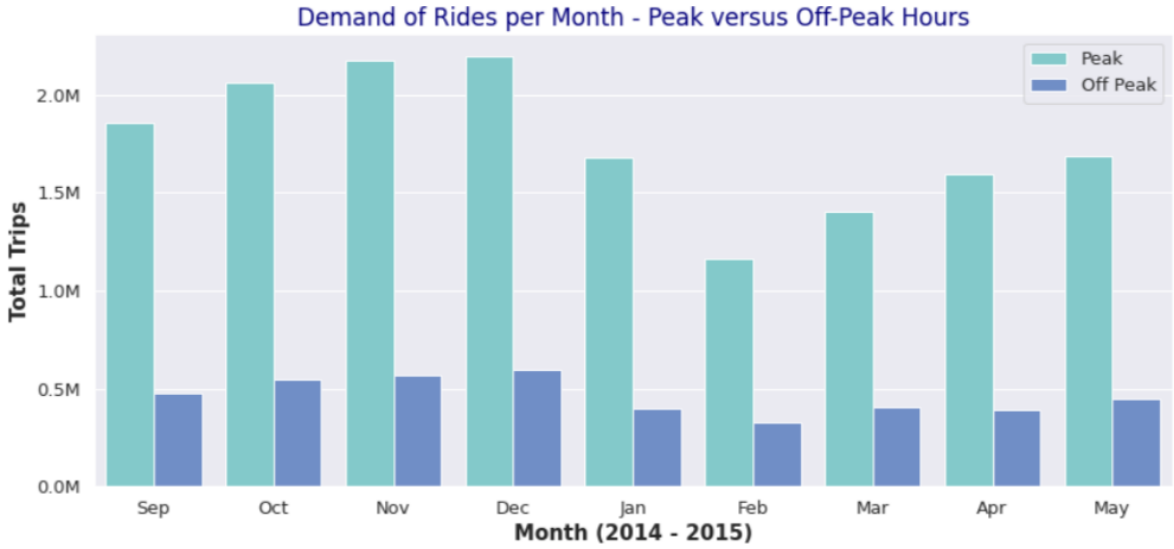
- Average Speed Per Hour during Weekdays



- The highest speed was observed to be at 5Am and the lowest at 9 AM and 6 PM.

- **Average Trips Per Month**

- Trips per month during peak hours and off-peak hours



- As we can see from the graph that the highest trips are in the month of December in peak hours.
- The lowest trips in peak hours are in the month of February.
- The highest trips in off peak hours are during the month of December.
- The lowest trips during off peak hours are in the month of February.
- Both in peak and off-peak hours lowest number of trips is in the month of February.

Results

- Our findings were generated in four phases, one loading the data of uber as whole and cleaning the data, I.e., removing duplicate, redundant data and also removing data with null values.
- Then exploring the data at the basic levels to get tables and graphs of uber in new york,2014 in terms of months, days, dates, bases and hours.
- Then the final step was to use the observations and already explored data to find further relationship between the data, and finding useful results which can help in further understanding the case study.

- Through this project we learned how to use data, load it, and then analyze it.
- Also, how to represent it through tabular form, histogram, bar graph, dot plot, scatterplot, geographic representation, distplot, heat graphs and many more features.
- Through this project we learnt the importance of data Science brings together the domain expertise from programming, mathematics, and statistics to create insights and make sense of data.
- At the end of the Uber data analysis R project, we observed how to create data visualizations.
- We made use of several packages that allowed us to plot various types of visualizations that pertained to several time-frames of the year.
- With this, we could conclude how time affected customer trips. Finally, we made a geo plot of New York that provided us with the details of how various users made trips from different bases.

Future Scope

- We can use this data for training a model using ML and building a smart AI based predictive system.
- Model can automatically send the insights to the authorities or drivers related to areas having most trips and passenger count in certain areas.
- This big data can be used to study passenger's behavior.

Literatures Cited

[1] https://ggplot2.tidyverse.org/

[2] https://rpubs.com/Unsa/582359

[3] https://www.skyfilabs.com/project-ideas/uberdata-analysis

[4] https://www.uber.com/us/en/careers/teams/datascience/

[5] https://iedu.us/tag/project-in-r-uber-dataanalysis-project/

[6] https://growvation.com/paritoshsankhla/project/uber-data-analysis/5e95ee80-9455- 4473-acaf-b670fe2abc8b

[7] https://acadgild.com/blog/mapreduce-use-case-uber-data-analysis

hadoop.apache.org/

[8] https://www.owler.com/reports/acadgild/acadgild-blog-mapreduce-us e-case---uber-data-analy/1470309362454

[9] https://ieeexplore.ieee.org/document/8389665

[10] https://github.com/vickyg12/Uber-MapReduce-Data-analysis

Conclusion

- Before working on features first we need to know about the data insights which we get to know by EDA.
- Apart from that, we visualize the data by drawing various plots, due to which we understand the relationship between values in data set.
- Uber uses a mixture of internal and external data to estimate fares.
- Uber calculates fares automatically using street traffic data, GPS data and its own algorithms that make alterations based on the time of the journey.

Thankyou!!