Uber Data Analysis

P VAISHNAVI
18MIS1086
VIT University Chennai, Tamil Nadu, India
vaishnavi.p2018@vitstudent.ac.in

Y L SAI CHARITHA
18MIS1026
VIT University Chennai, Tamil Nadu, India
ylakshmi.saicharitha2018@vitstudent.ac.in

T JAYASRI
18MIS1020
VIT University Chennai, Tamil Nadu, India
thiyyagura.jayasri2018@vitstudent.ac.in

M BHUMIKA
18MIS1024
VIT University Chennai, Tamil Nadu, India
manchikanti.bhumika2018@vitstudent.ac.in

Abstract

Over the years, data analytics has aided businesses in optimizing and improving their performance. Identifying emerging trends, investigating linkages and patterns in data, analyzing in depth, and the insights we receive from these models above are just a few of the benefits of data analytics and visualization. Is It Necessary That We Spend Time Thoroughly Examining These Concepts For All The Benefits They Provide? It is created using the 'R' programming language and utilities like ggplot2, lubridate, dplyr, and tidyr. We may learn about numerous sophisticated operations performed in data visualization through efforts like these. It will allow us to discover trends in this massive organization's data and deliver important insights into hitherto untapped data. Also, assist us in comprehending the ggplot2 library's functions.

Uber was founded eleven years ago and was already one of the fastest-growing companies in the world. UberX claims to price 30% less than taxis in Boston, which is a terrific way to catch people' attention. Machine learning and artificial intelligence are already used in practically every industry, so we tried to apply the same principles to estimating Uber cab prices. We experimented with datasets and looked at how machine learning techniques can be used to detect patterns in data in this project. We'll mostly explain how Machine Learning Algorithms are used to estimate the price of various Uber cabs. Our problem is related to regression supervised learning. We can use a variety of machine learning algorithms, including Linear Regression, Decision Tree, Random Forest Regression, and Gradient Boosting Regressor, but the one that shows to be the most accurate for price prediction is finally chosen. We must choose an algorithm that prioritizes precision.

Literature review

Past few years have seen tremendous growth in uber related data analysis using machine learning. People are coming up with various methods to analyze uber related data such as A state in which the results, k-means clustering is used to estimate the most likely collection points at a given time and to predict the best hotspots of nightlife learning trends from previous Uber pickups. The center of the taxi service decides on the space of area to be targeted for the pickup of passengers.

This can be justified by explaining that machine learning is the core of Uber and how it has impacted on tremendous growth.

- Bridging the supply demand gap

- Reduction in ETA
- Route Optimization

[1] Green Cabs vs Uber in New York City:

Driving share and share alike on a mass-scale. Poulsen, L.K, In this document applied an experiment of spatial analysis of Green cab and Uber to hotspots of New York to determine the competitive position of the NYCTLC. The resulted research showed that as demand of green cabs on the hotspots grew, the demand of Uber taxis on the hotspots also grown. This research recommends that NYCTLC creates a dashboard that analyzes and displays data in real time, as we believe this will increase its competitiveness compared to Uber. Uber is a recent taxi operator in New York and is constantly devouring the market share of the yellow and green taxis of New York Taxi and Limousine Commission (NYCTLC). The NYCTLC is an agency of the New York City Government which licenses and regulates taxis and vehicle for hire industries and also app based companies. The commission was founded on March 2 ,1971 and their headquarters are based in New York.

[2] Predicting Short-Term Uber Demand in New York City Using Spatiotemporal Modelling:

Faghih, S.S recommends a recent modelling approach in Manhattan, New York City, to capture the demand for electronic mail services, particularly the Uber application. The Uber data was added in 15-minute intervals to the Manhattan TAD level. This aggregation enables a current approach to spatio-temporal modelling to be implemented in order to acquire a geographical and temporal understanding of demand. Using Uber gathering data, two spacetime models were created, with the STAR and STAR and MSPE turns determining the models' output. According to the MSPE's findings, the Lasso-Star system should be used instead of the star design. A comparison of request for yellow and green Uber cabs in New York in 2014 and 2015 shows that demand for Uber has increased.

[3] The Impact of Uber and Lyft on Taxi Service Quality: Documentation from New York City:

Ahmed, M., has shown that by using detailed data on taxis at the travel level and on the rental vehicle and data on complaints about the level of new complaints at the level of incidents, We examine how the advent of Uber and Lyft has damaged taxi service quality in New York City. The entire impact of the scenario-based organisations, particularly the riding administrations, was massive and pervasive. One of these consequences is the intensification of competition between Uber and Lyft over taxi service quality. They measure (the absence of) service quality using a novel set of complaint data that has never been examined before. Concentrate on the quality metrics that result from the numerous complaints we present Increased competition for these connected travel services has had a considerable impact on the behaviour of taxi drivers.

[4] The competitive effects of the sharing economy: how is Uber changing axis?

Wallsten, S, stated that the results of New York and Chicago are consistent with the possibility that taxis react to the new challenge by improving quality. The emergence of Uber in New York has been connected to fewer objections to visiting the city. Using the New York City Taxi and Limousine Commission's data, they investigate the competitive impact of sharing cabs in the

taxi sector. comprehensive data set of over one billion taxi journeys in complaints and information from New York, New York, and Chicago. Google Trends on Uber's largest shared transportation service's success.

[5] Predicting Short-Term Uber Demand in New York City Using Spatiotemporal Modeling:

Faghih, S.S said that the demand for electronic mail services is growing rapidly, particularly in large cities. In the United States and New York City, Uber is the most well-known email company. A comparison of demand for yellow and Uber cabs in New York between 2014 and 2015 shows that Uber's popularity has surged. To assess the models' forecast performance, you select data for a typical day. This document's objective is to explain how these models can be used to forecast Uber demand. The Uber data includes information about the location and time of each trip's pick-ups and returns throughout the day. According to the data provided, Uber's historical data from April 2014 is available.

[6] Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR

Kumar, states that, k-means clustering is used to estimate the most likely collection points at a given time and to predict the best hotspots of nightlife learning trends from previous Uber pickups. The accuracy of travel time estimates in Delhi, India's capital city, and the National Capital Region is investigated in this study (NCR). We collected data on 610 journeys from 34 Uber users using the Uber mobile and web applications. We demonstrate the volatility of Uber cab journey time estimations empirically. We also highlight the negative effects of such unpredictability on passengers waiting for cabs, which resulted in the cancellation of 28.4 percent of requested trips. Our empirical findings deviate dramatically from the high accuracies reported in the literature on journey time estimation. These gloomy findings may prompt future research into why travel time estimations do not meet the high accuracy levels stated in the literature - (a) Is it a problem of a lack of training data in developing countries? (b) a flaw in the algorithm that fails to capture the (lack of) historical trends in developing countries.

[7] Exploring the Taxi and Uber Demand in New York City: An Empirical Analysis and Spatial Modeling:

Through quantitative assessments of Uber and taxi demand for New York City areas, this study intends to evaluate the influence of new app-based for-hire vehicles on the taxi sector (NYC). Uber and taxi rides are spatially dependent, thus demand forecasting models that account for this are being created. In their empirical investigation, the authors look at the spatio-temporal trends of Uber and taxi pick-up data. There is a strong correlation between taxi and Uber pick-ups, especially in the city's central districts. Uber journeys surged by 10 million (223.3 percent) from 2014 to 2015, whereas taxi trips (including yellow and green taxis) fell marginally by 0.8 million (1.0 percent). Manhattan has the lowest rate of Uber growth (201.2%), whereas the surrounding boroughs, such as the Bronx (597.0%) and Staten Island (597.0%), have the greatest (573.0 percent). The Moran's I tests demonstrate that both taxi and Uber demand are spatially dependent. Using socio-economic and transportation-related factors, linear models, geographical error models, and spatial lag models are created to estimate taxi and Uber demand in each neighbourhood. The spatial error models that outperform the other two by capturing

spatial dependency via a geographically lagged dependent variable are found to outperform the other two. Shorter transit access times (TAT), longer roadways, less vehicle ownership, more income, and more job prospects are all linked to higher taxi and Uber demand.

[8] Uber Data Analysis using Map Reduce:

Author states that before mapReduce, doing this type of calculation would are troublesome. Currently programmers will handle this type issues with an ease. The advanced algorithms have been coded by the knowledge scientists for frameworks so that it becomes easy to use for the programmers. They don't want the department of PhD scientists to develop a whole complex framework. mapreduce will work on the network which provides a straightforward analysis. MapReduce is gaining ground chop-chop as a result of the Apache Hadoop and Spark parallel computing systems lets programmers use mapReduce to run models and algorithms over giant distributed sets of knowledge and use advanced applied math and machine learning techniques to try to predictions, realize patterns, uncover correlations, etc.

[9] Rahul Pradhan, Praveen Kumar Mannepallli - Analysing Uber Trips using PySpark:

The majority of businesses that go online in a place where data are generated rise day by day. It is important to grow a commercial enterprise with this aggressive encircling records evaluation. This analytics project is critical in determining the utilisation of Records Analytics. Through such programs, many organizations can identify many complex activities. The Uber Data Analysis Task allows us to visualize the organization's complicated realities. It was created using the R programming language. In this article, we look into Uber pickups in New York City on a daily, monthly, and annual basis. This mission is mostly focused on Data Visualization, and it will guide you through the use of the ggplot2 package to visualize data.

[10] An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC:

"We show that the number of Uber and Lyft rides is highly related to whether it rained in New York City using all taxi, Lyft, and Uber journeys," the study's author states. When it rains, the number of Uber (Lyft) rides per hour rises by around 22 (19%), while the number of taxi rides per hour rises by only 5%, illustrating that surge pricing (prime time) encourages supply growth. We show that once Uber joined the market in May 2011, the number of taxi rides, passengers, and fare revenue all dropped significantly. Since then, taxis have not responded to increasing demand in wet hours any differently than they have in non-rainy hours. Finally, we look at whether Lyft's entry into the market has influenced Uber. Uber, according to our estimates, continued to increase when Lyft entered the market, however Uber trips during wet hours fell by roughly 9%. Our findings suggest that Lyft and Uber drivers compete for rides when demand surges unexpectedly, such as during rainy hours.

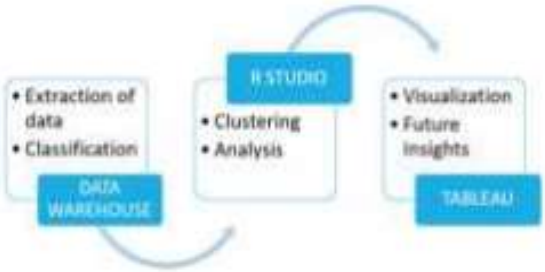
Proposed Work

We offered that we create a data visualisation project utilising R and other tools such as ggplot2 and others. Analyze numerous characteristics such as (a) trips by hour of the day (b) trips by

month of the year. Finally, construct visuals for various timeframes throughout the year. Explain how the passage of time influences customer journeys.

- Customers are frequently dissatisfied with traditional cab companies due to exorbitant rates and long wait times, and as a result, new and large markets can be exploited.
- Determine which days each basement has a higher quantity of active vehicles.
- Having the ability to tap into burgeoning markets in sub-urban locations where taxi services are not available.
- Estimated Time of Arrival can be shortened by increasing the number of Uber drivers, which will make Uber more popular with consumers, resulting in more revenue for the firm and benefits for the drivers.

Based on the data, we'll determine which destinations people visit the most and which earn the highest airline income for travel, as measured by the number of booked trips.



As shown in the diagram, the suggested system will begin with data extraction. In this scenario, we used data from April to September 2014 in New York City. For more relevant findings, data cleaning and classification are quite important. We will next go to the clustering and analysis stage, where we will use the ggplot2 library in R studio to read data in related variables and plot trips across various parameters such as week, month, and so on. The final stage entails visualising the results gathered, which we accomplished using Tableau Software. It assists us in deciphering data given in the form of graphs.

In this project, We try to find and analyze those key factors like date, month etc which helps Uber Company to enhance their business by focusing on those services and make required changes using a Kaggle dataset.

Dataset

1. Dataset URL: Dataset for visualization
   The dataset contains data on Uber pickups in New York City from April to September 2014. It has over 500k pickups (rows) and the four columns shown below:
   - Date/Time - The date and time of pickup
   - Lat - Latitude of pickup
   - Long - Longitude of pickup
   - Base
2. Dataset URL: Dataset for Prediction
   The native dataset contains 693071 rows and 57 columns which contain the data of both Uber and Lyft. But for our study, we just need the Uber data so we filter out the data according to our purpose and got a new dataset that has 322844 rows and 56 columns.

The dataset has many fields and attributes that specify us about the time, geographic location, and climatic conditions when the different Uber cabs opted.

Implementation

- R Studio: In this project, we use R to visualize the data and plot some graphical interpretation. Here, RStudio is an IDE. It includes a console, syntax, and highlighting editor, as well as graphing, history, debugging, and workspace management features for direct code execution.
- Tableau: For showing certain insights from our dataset, we used tableau. Tableau is a data visualisation application that has completely transformed the storey of using data to obtain insights. It has enabled data enthusiasts and companies to make the most of accessible analytical data.
- Google Colab: We built python models and predicted some goal values for our insights using Google Colab. Colab allows anyone to write and run python code, and it's particularly good for machine learning and data analysis.

Implemented Modules

- Data Preparation
- Data Visualization
- Data Analysis
- Feature Engineering
  - Filling NA, Drop useless column
- Modeling for prediction (Best Month and Price prediction)
  - Linear Regression
  - Decision Tress Regressor
  - Random Forest Regressor
  - Gradient Boosting
- Testing / Data Analysis
  - Mean Absolute Error
  - Mean Squared Error
  - Root Mean Squared error

Source code Files

- Code Google Colab Source code
  In this File we did the Predictions on which month the rides for a particular longitude and latitude are on much demand. We used Linear Regression, Decision Tree and Random Forest and predicted the target month. And we analyze which model fits the dataset by comparing their respective Mean Square Error, R2 scores.

- Code Google-colab - Prediction
  In this file, we see different columns of the table and try to co-relate them with others and find a relation between those two. We try to find and analyse those key factors like date, month etc which helps Uber Company to enhance their business by focusing on those services and make required changes. We use Linear reg, Decision tree, Random

forest, Gradient boosting, K-fold cross validation and had interpreted the accuracy through MSE, MAE, RMSE.

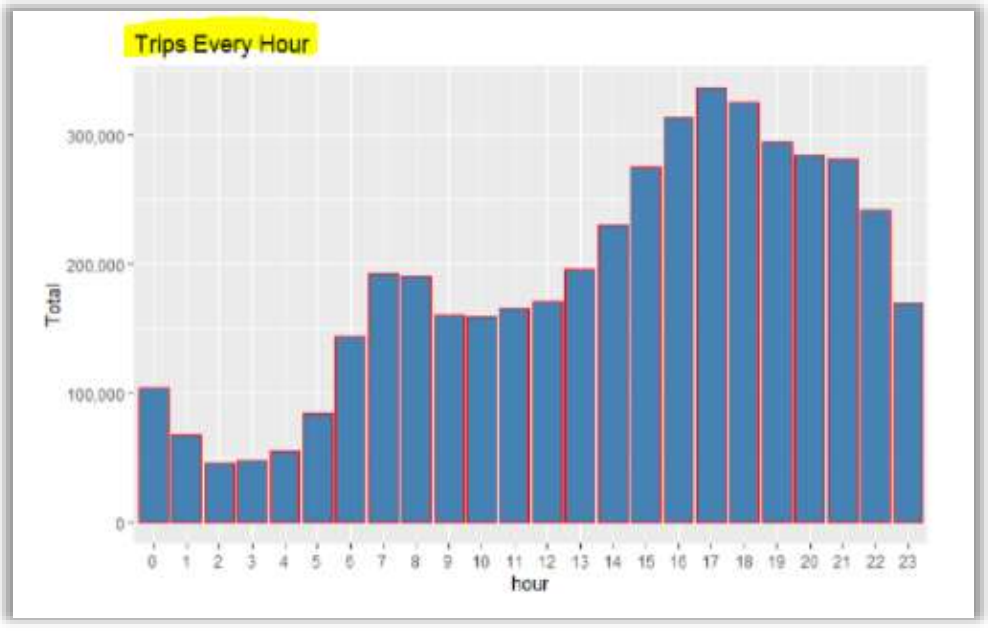
Results and Discussion

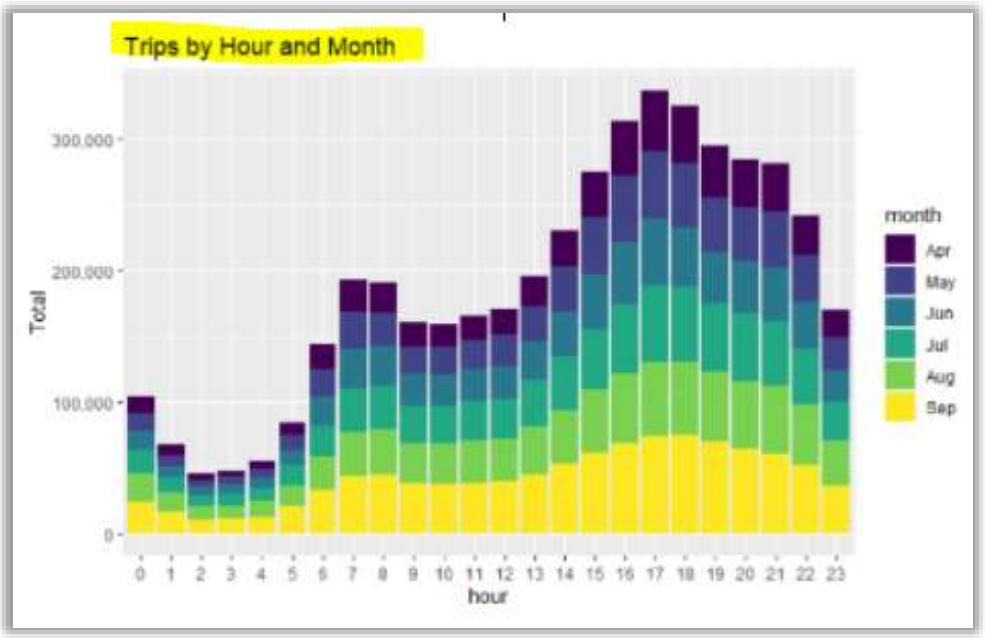
Plotting the trips by the hours in a day

Show 10 ▾ entries Search:

	hour	Total
1	0	103836
2	1	67227
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



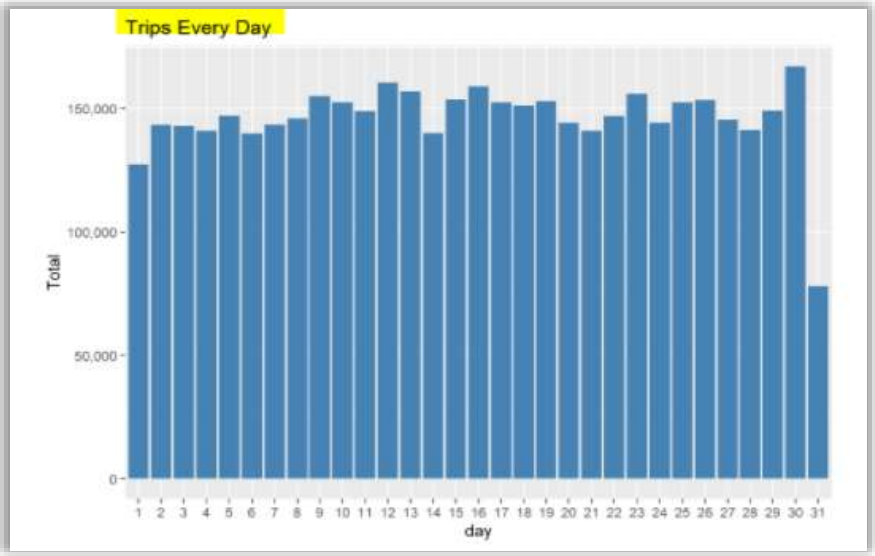


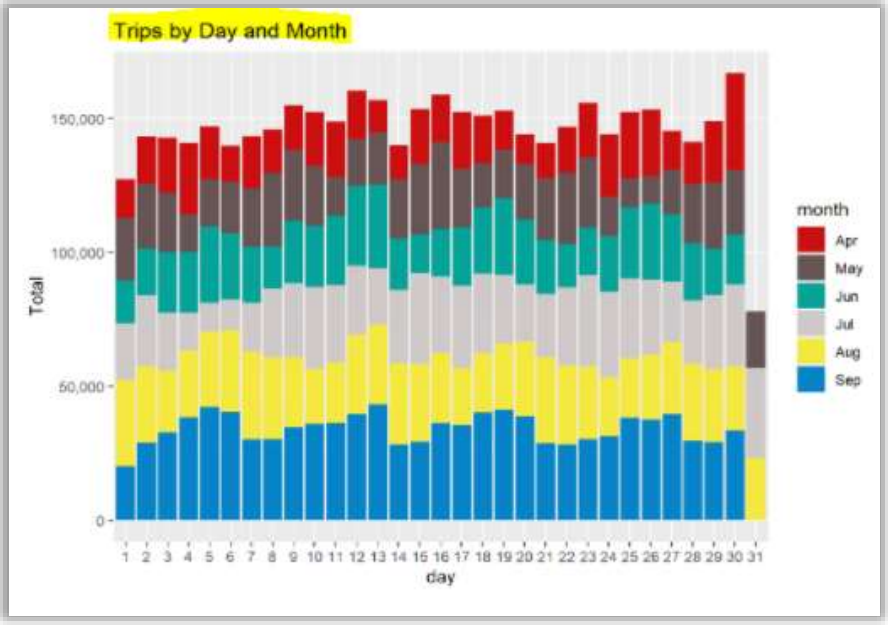
Plotting data by trips during every day of the month

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





Show 10 entries Search:

	month	day	Total
1	Apr	1	14546
2	Apr	2	17474
3	Apr	3	20701
4	Apr	4	26714
5	Apr	5	19521
6	Apr	6	13445
7	Apr	7	19550
8	Apr	8	16188
9	Apr	9	16843
10	Apr	10	20041

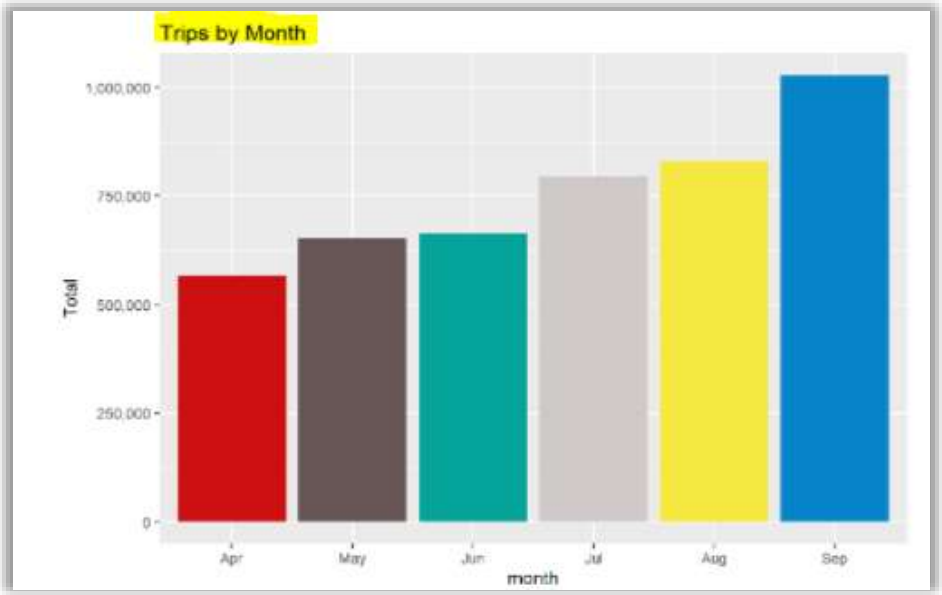
Showing 1 to 10 of 183 entries Previous 1 2 3 4 5 ... 19 Next

Number of Trips taking place during months in a year:

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

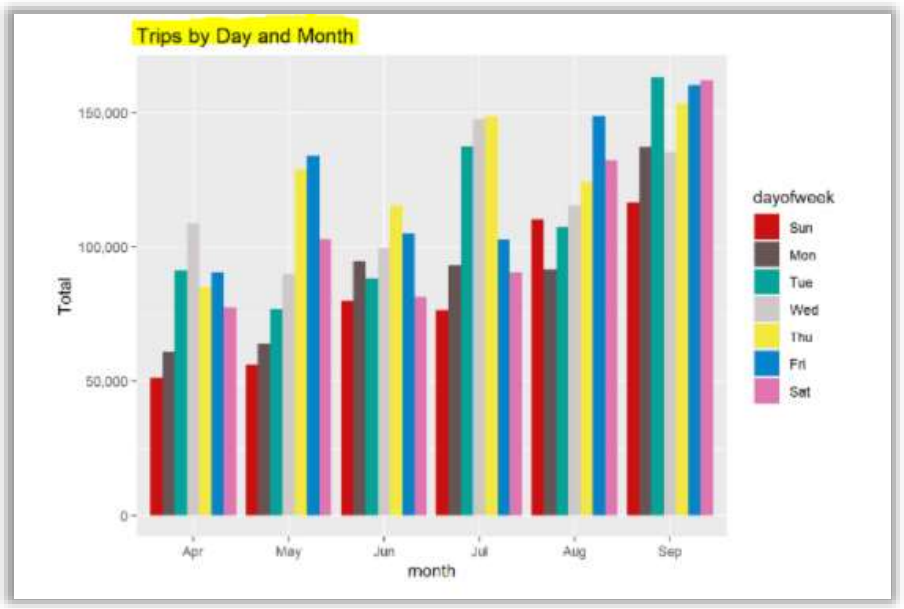


Trips by month and week_day:

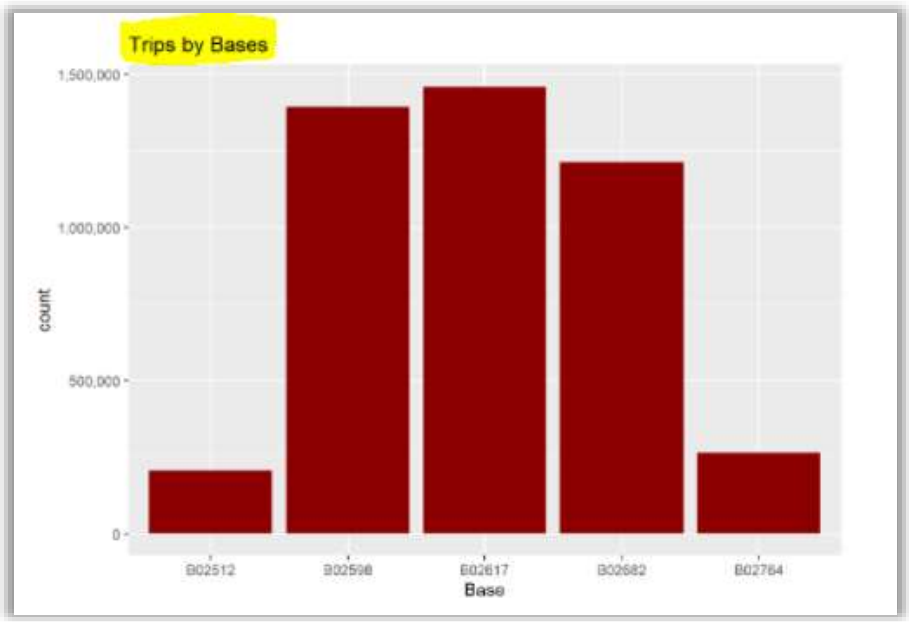
Show 10 entries Search:

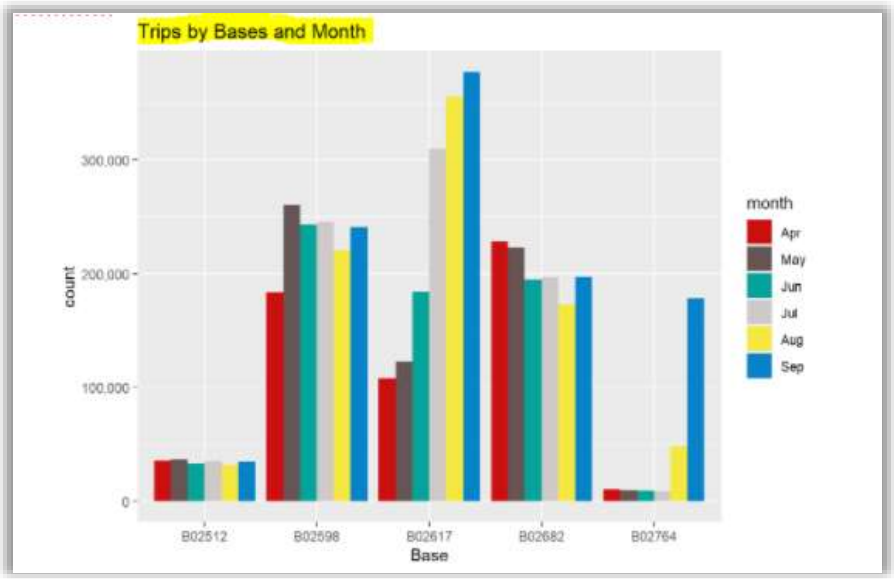
	month	dayofweek	Total
1	Apr	Sun	51251
2	Apr	Mon	60861
3	Apr	Tue	91185
4	Apr	Wed	108631
5	Apr	Thu	85067
6	Apr	Fri	90303
7	Apr	Sat	77218
8	May	Sun	58168
9	May	Mon	63846
10	May	Tue	76862

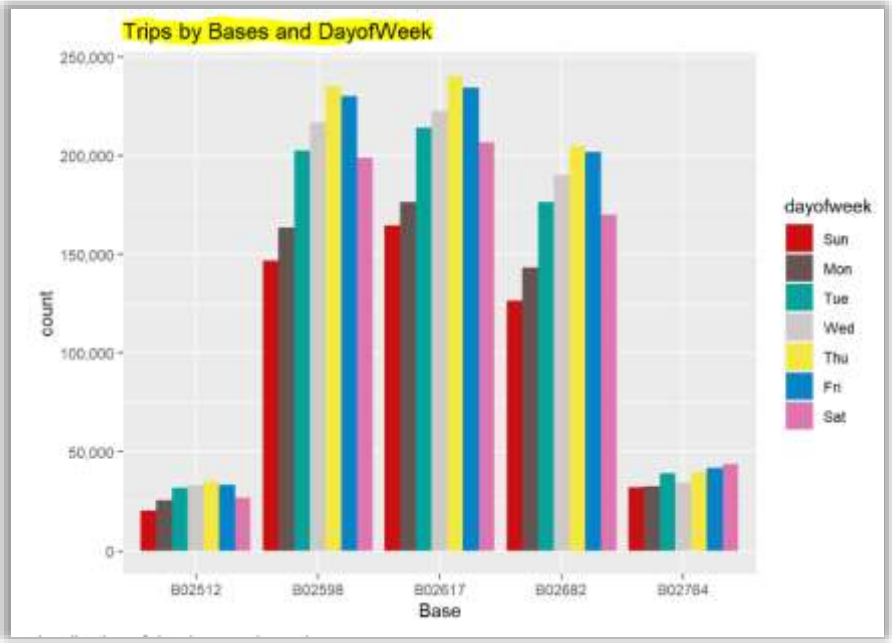
Showing 1 to 10 of 42 entries Previous 1 2 3 4 5 Next



Trips by Bases:


Trips by Bases
1,500,000
1,000,000
500,000
0
count
B02512
B02598
B02617
B02682
B02764
Base


Trips by Bases and Month
300,000
200,000
100,000
0
count
B02512
B02598
B02617
B02682
B02764
Base
month
Apr
May
Jun
Jul
Aug
Sep

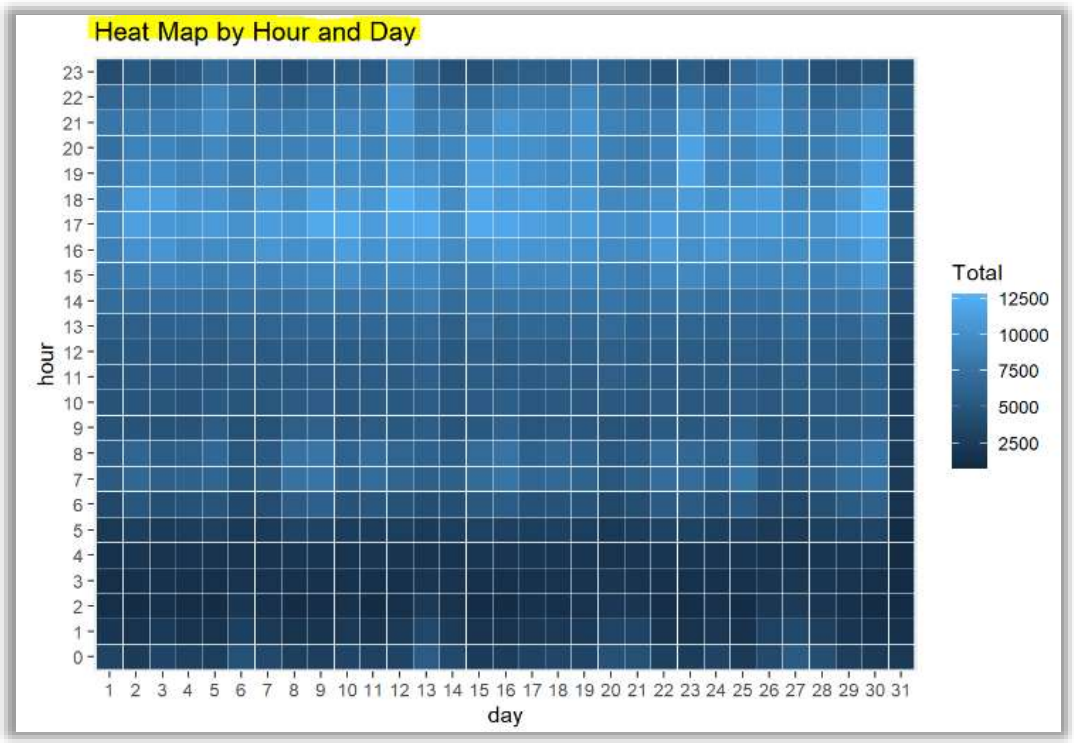


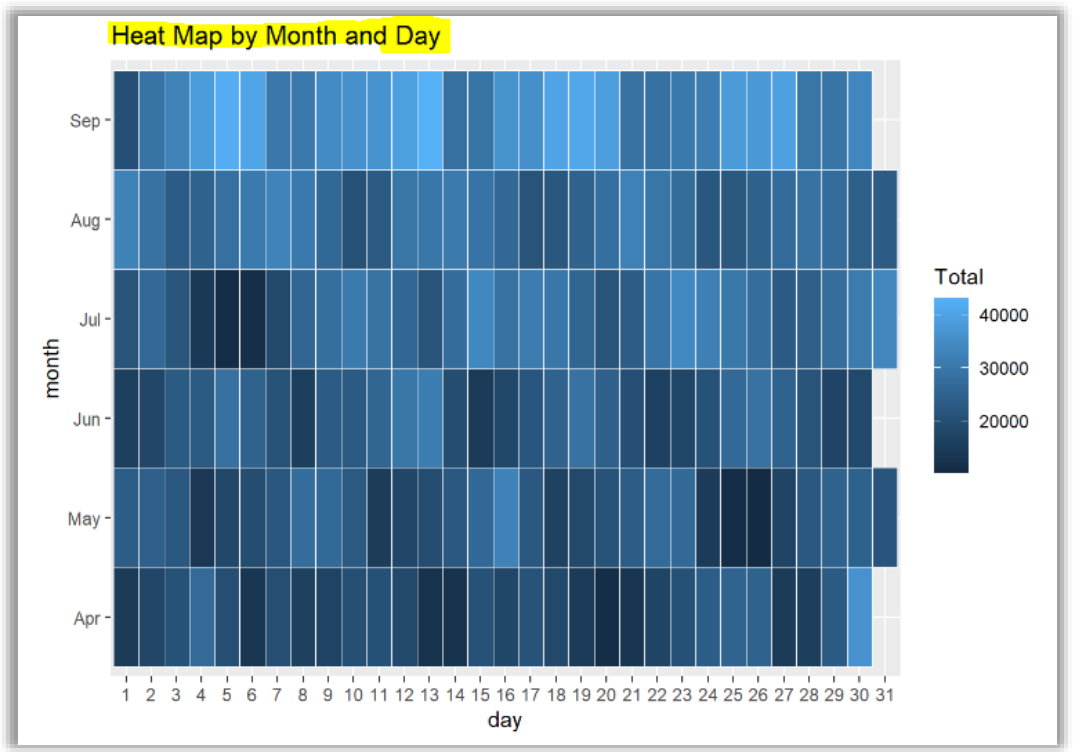
Creating a Heatmap visualization of day, hour and month:

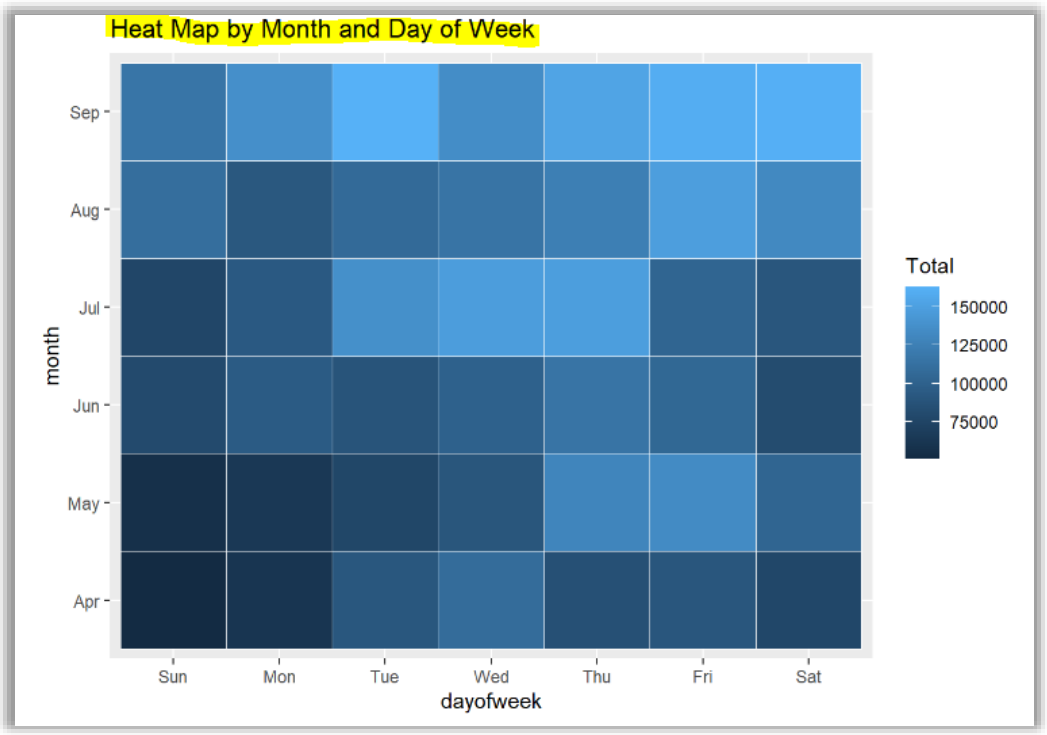
Show 10 entries Search:

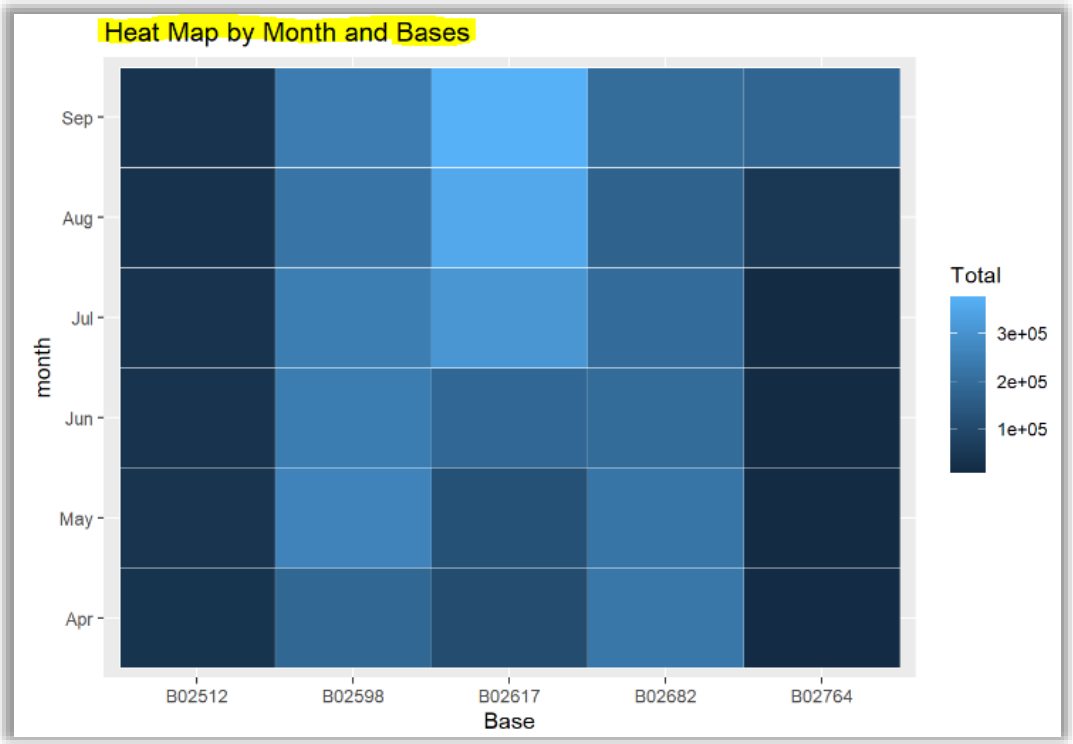
	day	hour	Total
1	1	0	3247
2	1	1	1982
3	1	2	1284
4	1	3	1331
5	1	4	1458
6	1	5	2171
7	1	6	3717
8	1	7	5470
9	1	8	5376
10	1	9	4688

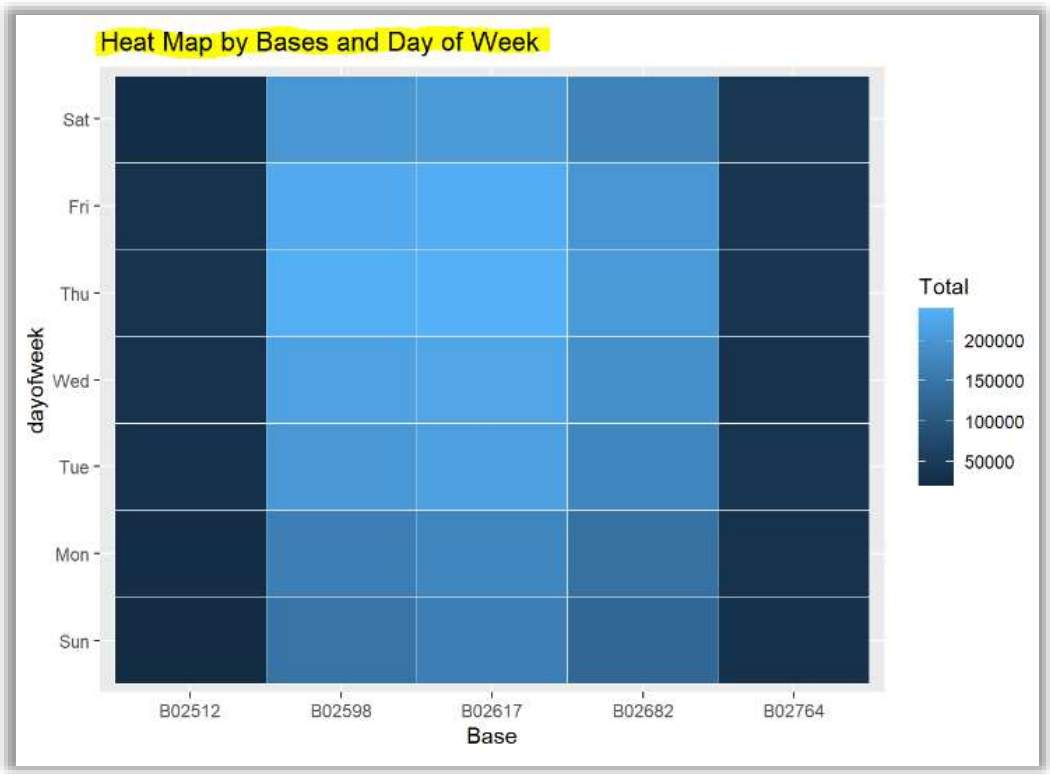
Showing 1 to 10 of 744 entries Previous 1 2 3 4 5 ... 75 Next


Heat Map by Hour and Day
hour
23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day
Total
12500
10000
7500
5000
2500

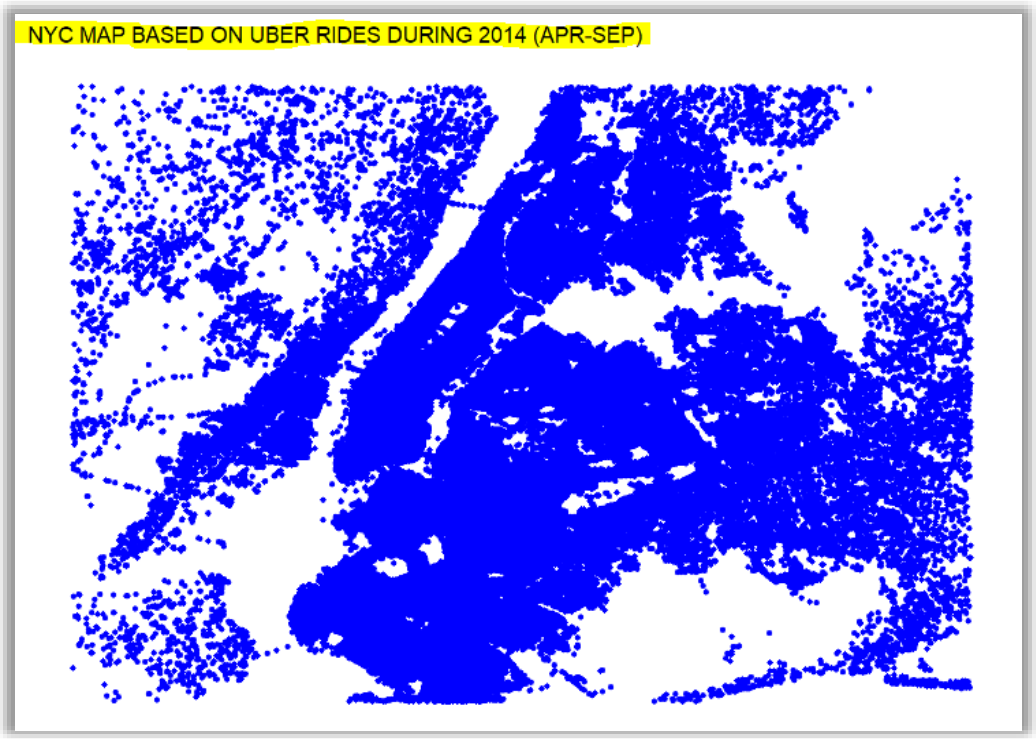

Heat Map by Month and Day
month
Sep
Aug
Jul
Jun
May
Apr
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day
Total
40000
30000
20000

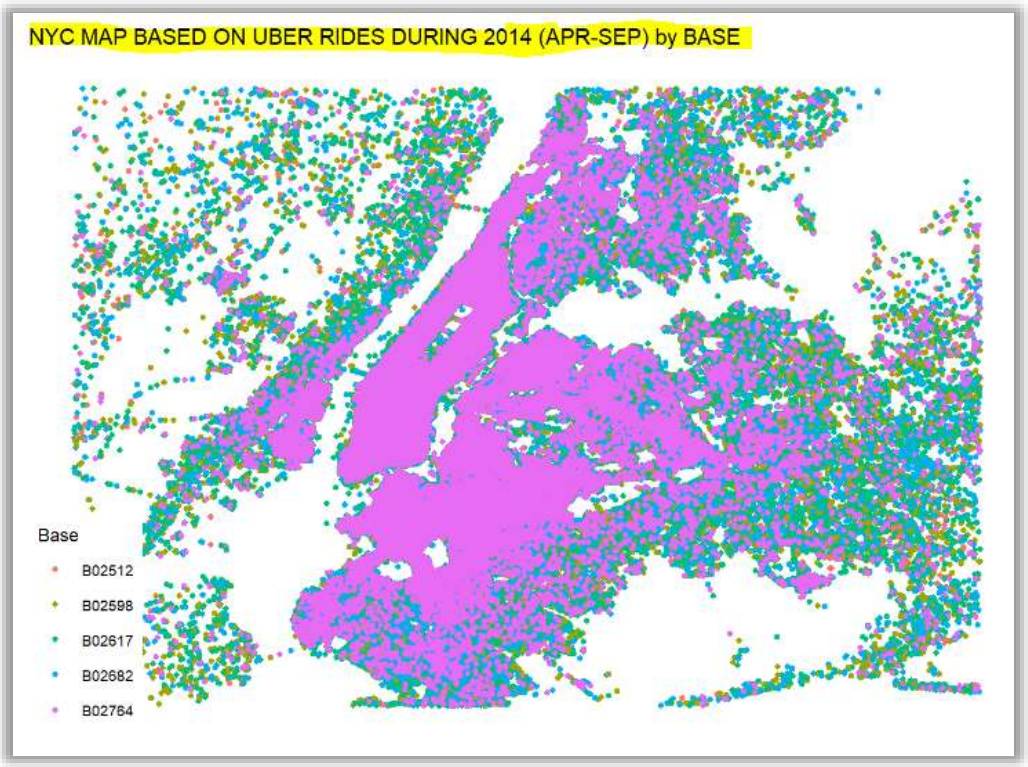

Heat Map by Month and Day of Week
Sep
Aug
Jul
Jun
May
Apr
month
Sun
Mon
Tue
Wed
Thu
Fri
Sat
dayofweek
Total
150000
125000
100000
75000


Heat Map by Month and Bases
Sep
Aug
Jul
Jun
May
Apr
month
B02512
B02598
B02617
B02682
B02764
Base
Total
3e+05
2e+05
1e+05



Creating a map visualization of rides in New York





Modelling

MSE & RMSE for predicting the month which has high demand for rides based on latitude and longitude:

```
[ ]  print("Mean Square error: ")
     print("1.Linear Regression           :",mean_squared_error(Lin,y_test))
     print("2.Decision Tree Regression    :",mean_squared_error(Dec_tree,y_test))
     print("3.Random Forest Regressor     :",mean_squared_error(Random_for,y_test))
```

```
Mean Square error:
1.Linear Regression           : 2.895833986809929
2.Decision Tree Regression    : 3.143987820684744
3.Random Forest Regressor     : 3.0101558069077683
```

```
[ ]  print("R2_Score: ")
     print("1.Linear Regression           :",r2_score(Lin,y_test))
     print("2.Decision Tree Regression    :",r2_score(Dec_tree,y_test))
     print("3.Random Forest Regressor     :",r2_score(Random_for,y_test))
```

```
R2_Score:
1.Linear Regression           : -742.8771838776281
2.Decision Tree Regression    : -3.7929186097457848
3.Random Forest Regressor     : -4.895354319399451
```

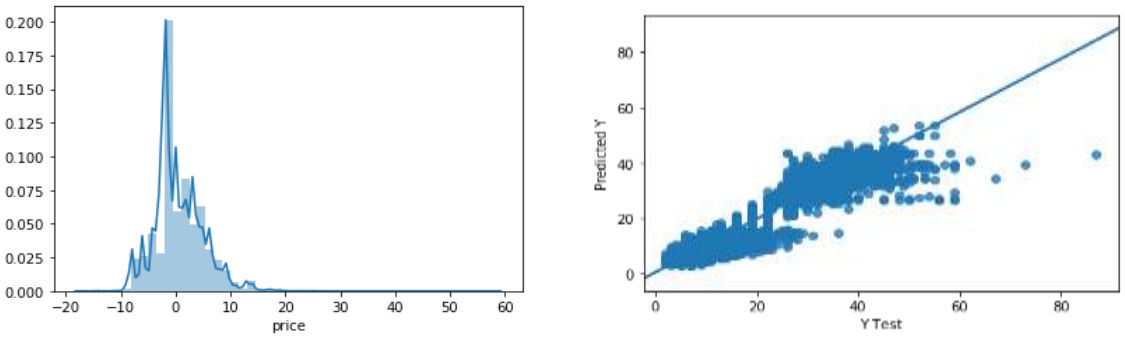
Inference: From the drafted results, we can depict that Linear Regression which has low MSE and high RMSE is the best fit model for our dataset that we have chosen.

Accuracy for Price prediction in Linear regression, Random Forest, Decision Forest, Gradient Boosting Regressor and based on MSE, MAE, RMSE:

Serial No.	Models	Accuracy
1	Linear Regression	0.957545073
2	Decision Tree	0.851791729
3	Random Forest	0.962269474
4	Gradient Boosting Regressor	0.843187213

Error table for Linear Regression:

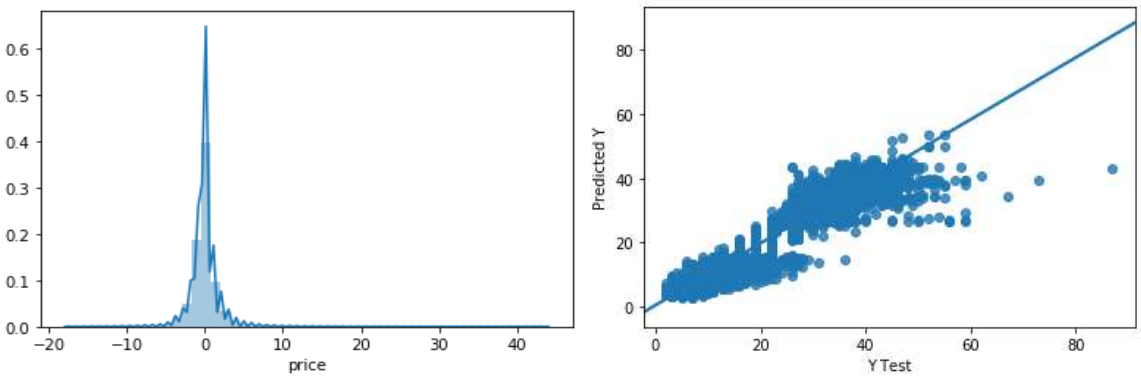
Serial No.	Models	Accuracy
1	Mean Absolute Error	3.40607721
2	Mean Squared Error	20.0334370
3	Root Mean Absolute Error	4.47587277



Scatter & Dist Plot for Linear Regression

Error table for Random Forest

Serial No.	Models	Accuracy
1	Mean Absolute Error	0.99813700
2	Mean Squared Error	2.94465361
3	Root Mean Absolute Error	10.71599930



Scatter & Dist Plot for Random Forest

Price Prediction

```
[ ]  def predict_price(name,source,surge_multiplier,icon):
         loc_index = np.where(new_uber.columns==name)[0]

         x = np.zeros(len(new_uber.columns))
         x[0] = source
         x[1] = surge_multiplier
         x[2] = icon
         if loc_index >= 0:
             x[loc_index] = 1

         return random.predict([x])[0]
```

predict_price(cab_name , source , surge_multiplier , icon)

```
[ ]  predict_price(1 , 3, 2, 0)
```

```
C:\Users\gupta\Anaconda3\lib\site-packages\ipykernel_laun

26.440184991891492
```

Inference: We show the higher in Linear regression and random forest models based on the driving outcomes. As a result, we calculated the RMSE, MSE, and MAE for both models. In the Random forest model, we found low MSE, low MAE, and high RMSE.

Thus, Price prediction has been done with dependent parameters i.e., Source(pickup address), name(vehicle type), surge_multiplier(% of rate multipler on demands) and icon(climatic conditions) for the dataset that we have taken.

References

[1] Poulsen, L.K., Dekkers, D., Wagenaar, N., Snijders, W., Lewinsky, B., Mukkamala, R.R.and Vatrapu, R., 2016, June. Green Cabs vs. Uber in New York City. In 2016 IEEEInternational Congress on Big Data (BigData Congress) (pp. 222-229). IEEE.

[2] Faghih, S.S., Safikhani, A., Moghimi, B. and Kamga, C., 2017. Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study. arXiv preprintarXiv:1712.02001.

[3] Ahmed, M., Johnson, E.B. and Kim, B.C., 2018. The Impact of Uber and Lyft on TaxiService Quality Evidence from New York City. Available at SSRN 3267082.

[4] Wallsten, S., 2015. The competitive effects of the sharing economy: how is Uber changingtaxis. Technology Policy Institute, 22, pp.1-21.

[5] Faghih, S.S., Safikhani, A., Moghimi, B. and Kamga, C., 2019. Predicting Short-TermUber Demand in New York City Using Spatiotemporal Modeling. Journal of Computing inCivil Engineering, 33(3), p.05019002.

[6] Shah, D., Kumaran, A., Sen, R. and Kumaraguru, P., 2019, May. Travel Time EstimationAccuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR∗.In Companion Proceedings of The 2019 World Wide Web Conference (pp. 130-136). ACM.

[7] Correa D., Xie K., Ozbay K. (2017). Exploring the Taxi and Uber Demands in New York City: An Empirical Analysis and Spatial Modeling. Transportation Research Board's 96th, Annual Meeting, Washington, D.C.

[8] P. Devika, Y. Prasanna , P. Swetha ,G. Akhilesh Babu(2019). Uber Data Analysis using Map Reduce.International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

[9] Rahul Pradhan, Praveen Kumar Mannepallli - Analysing Uber Trips using PySpark International Conference on Visualization. ACM, 2018, pp. 90–98.

[10] Abel Brodeura, Kerry Nieldb. An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC, Journal of Economic Behavior & Organization. Volume 152, August 2018, Pages 1-16.