iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI: http://doi.org/10.22214/ijraset.2020.4153

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Sentiment Analysis and Visualization on Uber Dataset

Vaibhav S. Kundu[1], Raghavendra N. Singh[2], Siddharth Porwal[3], Bindu Garg[4]
[1, 2, 3]*Student, Computer Science, Bharati Vidyapeeth (Deemed to be) University College of Engineering Pune, India.*
[4]*Professor, Dept. of Computer Engineering, Bharati Vidyapeeth (Deemed to be) University College of Engineering, Pune, India.*

***Abstract:** The main purpose of this project is to analyze the pick -up done by Uber in New York City. In this project we are giving more stress on Data visualization, which will direct you towards using 'ggplot2' library understanding the data and also the sense to understand the customer /client who avail the trips. Sentiment analysis is extracting opinions that have different views, and sentiment analysis helps us to categories data in different classes or sections and Data visualization is a method to showcase the complex data through charts, graphs and maps, with the help of data visualization, company officials can understand the complex data and take in –depth sight in data to take important decisions favorable to company`s growth.*
***Keywords:** Sentiment analysis, data visualization, ggplot2, data*

I. INTRODUCTION

Sentiment analysis is basically the contextual mining of text which mainly identifies and extracts subjective and associative information within a source material, and helping a business or any project to understand the social and computational sentiment of their product, organization, brand or service while accounting and monitoring online conversations and communication. However, analysis and the computations of social media streams and data is usually restricted to specific basic sentiment analysis and mainly count based metrics. This is quite similar to just scratching the surface and missing out on that high value insights that are yet to be discovered. It is estimated that around [1] 80% of the world's data is unstructured, or basically it is unorganized. Large quantities of text data which mainly includes emails, messages, support tickets, texts, chats, social media conversations or communications, surveys, articles, documents, files, etc. is created and developed every day but it is quite difficult to analyze, understand, and mainly sort through, and not to identify the time-consuming task it results into and can also result to be quite expensive.

II. DATA VISUALIZATION

Data visualization is mainly the graphical representation of information and data in a more constructive and organized way. By using visual elements which mainly consists of charts, graphs and different types of maps. Data visualization tools provide an efficient and accessible way to view, analyze and understand various types of trends, outliers, and patterns in present in any form of data. In the field of Big Data, [2] data visualization tools and technologies are very essential to analyze massive amounts of knowledge discovery, information and make data-driven decisions possible.

According to the information provided by the World Economic Forum, the larger part of the world produces over 2.5 quintillion bytes of data on average almost every day, and 90% of all that data has been created in the last two or three years. With such a large amount of data, it becomes extremely difficult to manage and most importantly to discover any kind of knowledge from this type of data. It would be an almost impossible task for any single person to analyze through such type of data line-by-line and come across or view any kind of distinct patterns and make observations. Data proliferation which is the advanced way of analyzing and attaining knowledge from data can be managed as part of the data science process or data science field, which includes the concept of data visualization.

III. FASTER DECISION MAKING

Data visualization often helps in making effective and faster decision making. Companies who can collect and quickly make operations on their data will always result in being more competitive in the real market place because they can easily perform knowledge-based discoveries and make informed and appropriate decisions sooner than the rest of the competition. Speed is the most essential element, and data visualization aides in the understanding and appropriate knowledge discovery of vast quantities of data by applying proper visual representations to the different forms of data. This type of visualization of data can be treated as a layer in the whole process of data mining and knowledge discovery process. In this manner the visualization layer will be typically on the top of a data warehouse or data lake which allows users to discover, analyze and explore data in a self-efficient manner.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

IV. IMPLEMENTATION

We have to download dataset of Uber Pick-up dataset. Then we need some important packages, [3] we will import some of the important library which is required for our projects and some of the important libraries are explained below: -

1) *ggplot2:* One of the popular data visualization libraries used nowadays, which is used to create aesthetic visualization plots.
2) *Lubridata:* To better understand our dataset in different time category.
3) *Tidyr:* This package keeps our data Tidy.

Now, we have to decide color for our plot, by creating vector of color which will be implemented in our projects, which will be included in our plotting functions.

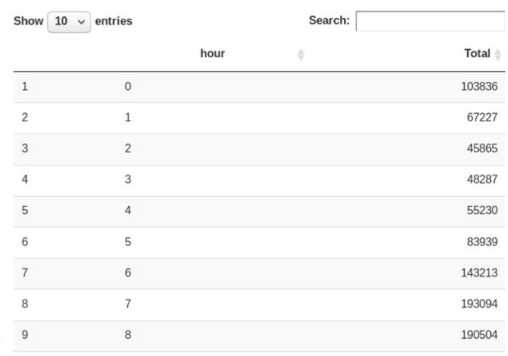
Some color codes are given below: -

a) 665555
b) CC1011
c) 05a399
d) f5e840

Next step, we have to declare variable to dataset, we go through data of Uber pick-up that contain data of 6 month of 2014, we will store them in corresponding column .After scanning of all the data , we will combine all the data in single column frame .Now, we will be formatting the Time & Date column, finishing this by creating factors of Time objects like Day , Month ,Year.

Here, we will be using ggplot function to find the trips carried by Uber drivers in New York [4] and how many times a passenger travelled through Uber. We are going to use dplyr to check if data is aggregate or not and if not aggregate the dataset. In the predicted visualization, we are going to understand the passenger fare throughout the day. We so far observed that the numbers of passenger trip are highest in time slot 5PM to 6PM in evening. Next, we are going to plot day-wise trips for month, for this we will use “groupby” clause to group by days. Then we will show the “datatable” table.

The resulting visualization is showing that the 30th of day of that month has highest trips in that month.


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

Fig 1: Screenshot

A proper graph can be visualized which helps in more effective understanding and analyzing of data. Such types of graphs make the process of knowledge discovery from any types or sets of data quite easy and is a very efficient technique. The process of decision making becomes very quick as by using proper data visualization techniques we can achieve an effective outcome very quickly.

Data visualization becomes very easy and reaches to the most optimal levels of accuracy in terms of conclusions which helps in taking right and appropriate decisions. The below graph sums up the main relation between trips performed every day in relation with number of days:

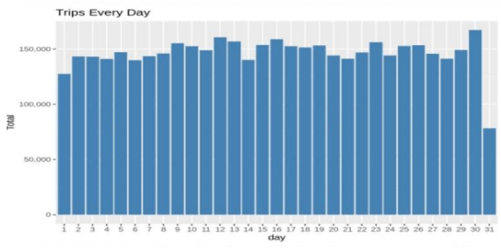


Fig 2: Graph

V. ACKNOWLEDGEMENT

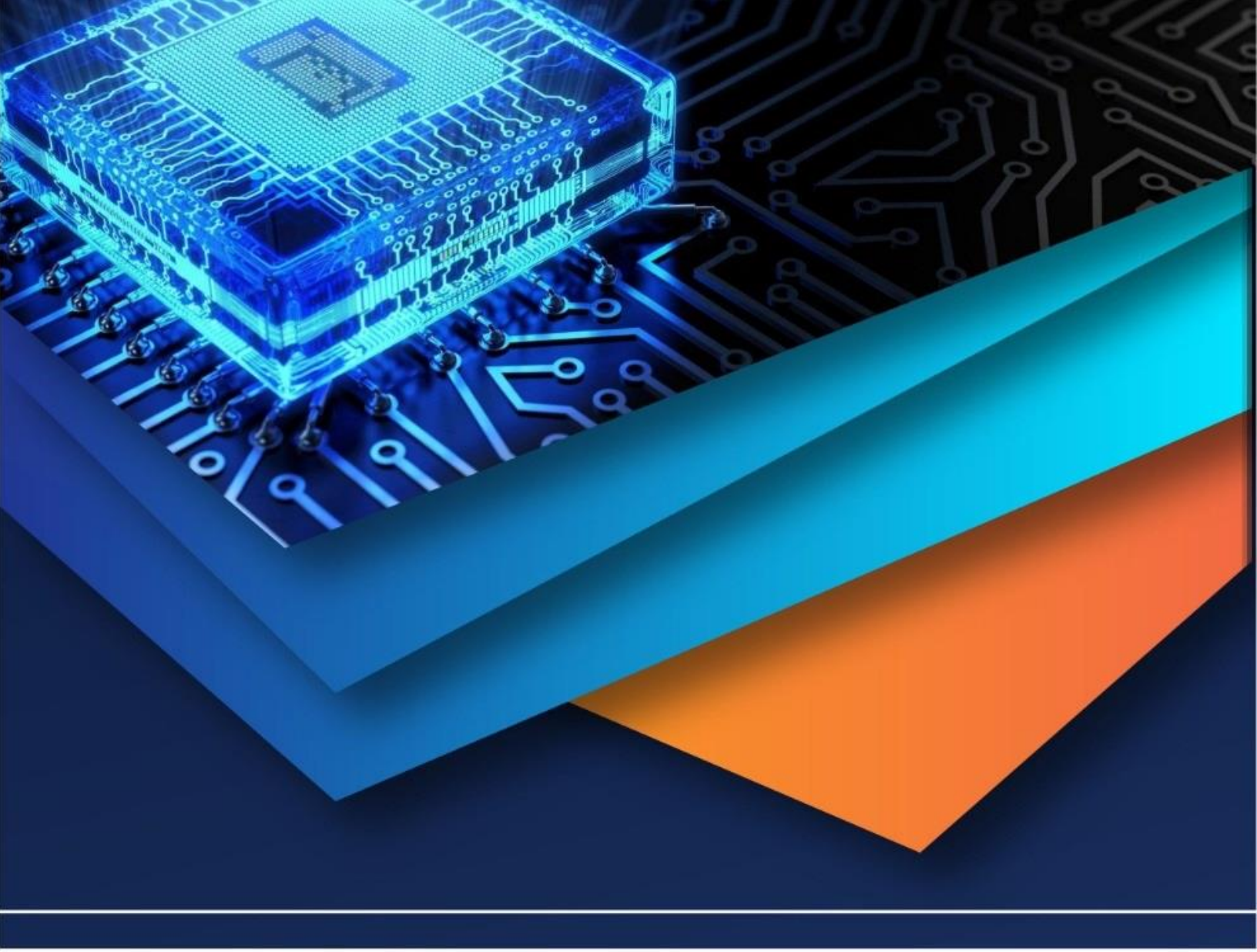
We express our deep gratitude to our project guide (Dr. Bindu Garg) for providing timely assistance wherever and whenever required. She has been a lighthouse to journey of making of this project.
We extend our gratitude to our Head of Department, Dr. Devendra Thakore who gave us this opportunity to work on this project and gave us every possible time and resource for its completion.
We would also like to thank all the faculty members of our department for their valuable insight and tips in designing the system.

VI. CONCLUSION

We performed semantic analysis on uber pickup by aggregating and collecting different datasets. The datasets mainly consist of number of pickups and time of the day. By using the function “dplyr” we aggregated the data and by using the function “groupby” we made a clause to group by days. After using proper data visualization techniques, we performed the sentiment analysis on the records or datasets available over the topic of uber pickup.

REFERENCES

[1] Tableau, articles, “*Data Visualization*”
[2] MicroStrategy, Data Visualization, “*Main Uses*”
[3] Investopedia, “*Data-mining*”
[4] Kaggle,” *Uber-pickups-in-New-York-city*”
[5] Analytics, indiamag,” *datasets for sentiment-analysis”*



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY