An approach to predict taxi-passenger demand using quantitative histogram on Uber data

A Bharathi, S Surya Prakash
Bannari Amman Institute of Technology
Sathyamangalam, India
nishabharathi93@gmail.com
suryaprakash.it16@bitsathy.ac.in

Abstract— **The precise prediction of the day to day and monthly transactions is of great value for companies. This information can be beneficial for the companies in analyzing their ups and downs and draw other plans. Moreover, a precise prediction method can optimize the performance of a company. The branch of analytics that deals with prediction is known as predictive analytics. This paper presents the use of data analytics in analyzing the transaction dataset provided by Uber to predict the possible outcomes and the changes to be made. The histograms and heat maps drawn provide us a clear visualization of the dataset and we must predict the rest out of it.**

Keywords— Uber trip data;Data analytics; histograms; Data prediction; heat map

I. Introduction

In this generation many technologies are created to support our day to day needs. In 2014, technology based applications on transportation were at peak. The scopes of technology in the field of transportation were booming. This gave birth to one such company named Uber, which later on became one of the pioneers of the world in the field of technology based transportation.

Many new companies started to turn their attention towards technology based transportation, which they believed to be the future of transportation industry. The key players in this system are the people who are the end users. The companies started focusing on the existing customers instead gaining new customers. The companies used social media for this purpose. This was a great strategy. The company started to get responses from their customers which later were the source of knowledge about customer needs and behaviors which played an important role in maintaining the relationship between the customer and the management. This information were of great help to the Uber in overpowering Yellow cab rides in the centre and Green cab rides in the outskirts of New York city.

Taxi drivers need to decide which place is the most suitable place for picking up the customers. Passengers also prefer to find a taxi no sooner than they are ready for pickup.

Drivers don't have adequate data about where the travelers and the different cabs are and to plan to go accordingly. Along these lines, a taxi focus can sort out the taxi armada and proficiently convey them as per the request from the whole city. To construct such a taxi center, a wise framework that can anticipate the future request all through the city is required.

II. Related Work

From the taxi informational series we can measure and count on the met taxi request, that is, the quantity of the taxi administrations developed what's more, will develop at various areas. In any case, the neglected taxi request, e.g., the ratio between the number of individuals who require a taxi at a particular time and the number of taxis readily available at that time may not be the same. To fix this problem, late papers try and collect the disregarded taxi request from the taxi informational collection. In [7] the creators are a part of flight landing with taxi request and foresee the tourist request at various air terminals in Singapore make use of queuing theory.

Taxi request expectation issue has pulled in more considerations as of late due to the reachable of taxi informational index [3]. Mukai forecast the taxi request from earlier taxi information with a neural community device.

Kai Zhao[1] targeted on predicting taxi call for through reading maximum predictability (Π max) of taxi call for most predictability is defined by using entropy of taxi demand together with both randomness and temporal correlation.

Moreira-Matias [10] introduced a methodology for predicting spatial distribution of taxi-customers using streaming data. Histogram time series is used to get the frequency of the taxi demand.

978-1-7281-3250-1/19/$31.00 ©2019 IEEE

K. Zhao [9] finding the city region based on the maximum taxi visits. In this paper Association rules are used to infer the functional regions in the city.

K. Harish [6] finds the number of trips occurring in a month and classifies according to the frequency of number of transactions taking place in a day using map reduce technique on Hadoop framework. In this paper the missing values have not been taken into consideration.

Lasse Korsholm Poulsen [2] finds the popularity between Green cabs and Uber and presents the growing rates and shows us comparison between the two.

III. Implementation

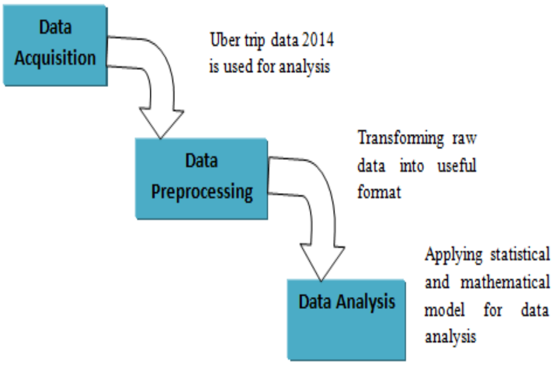
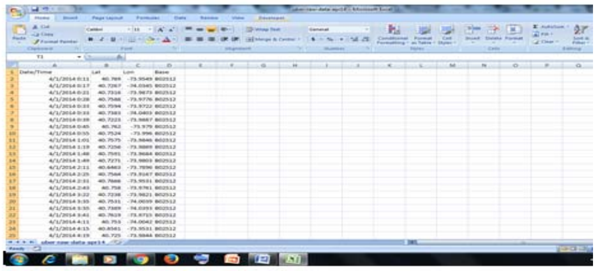

Data Acquisition
Uber trip data 2014 is used for analysis
Data Preprocessing
Transforming raw data into useful format
Data Analysis
Applying statistical and mathematical model for data analysis

Fig. 1. Flow Diagram of the Uber data analysis

A. Dataset acquisition & Dataset Description Uber Dataset

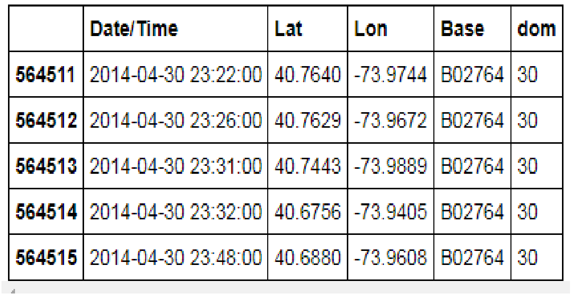
The dataset consists of trip record data of the Uber. It is made up of following fields.
Date/Time: The date and time of the Uber pickup data
Lat: latitude of the Uber pickup data
Lon: longitude of the Uber pickup data
Base: The TLC () base company code affiliated with the Uber pickup



1	Date/Time	Lat	Lon	Base
2	4/1/2014 0:11	40.769	-73.9549	B02512
3	4/1/2014 0:17	40.7267	-74.0345	B02512
4	4/1/2014 0:21	40.7316	-73.9873	B02512
5	4/1/2014 0:28	40.7588	-73.9776	B02512
6	4/1/2014 0:33	40.7594	-73.9722	B02512
7	4/1/2014 0:33	40.7383	-74.0403	B02512
8	4/1/2014 0:39	40.7223	-73.9887	B02512
9	4/1/2014 0:45	40.762	-73.979	B02512
10	4/1/2014 0:55	40.7524	-73.996	B02512
11	4/1/2014 1:01	40.7575	-73.9846	B02512
12	4/1/2014 1:19	40.7256	-73.9869	B02512
13	4/1/2014 1:48	40.7591	-73.9684	B02512
14	4/1/2014 1:49	40.7271	-73.9803	B02512
15	4/1/2014 2:11	40.6463	-73.7896	B02512
16	4/1/2014 2:25	40.7564	-73.9167	B02512
17	4/1/2014 2:31	40.7666	-73.9531	B02512
18	4/1/2014 2:43	40.758	-73.9761	B02512
19	4/1/2014 3:22	40.7238	-73.9821	B02512
20	4/1/2014 3:35	40.7531	-74.0039	B02512
21	4/1/2014 3:35	40.7389	-74.0393	B02512
22	4/1/2014 3:41	40.7619	-73.9715	B02512
23	4/1/2014 4:11	40.755	-74.0042	B02512
24	4/1/2014 4:15	40.8541	-73.9531	B02512
25	4/1/2014 4:19	40.725	-73.9844	B02512

Fig. 2.Uber dataset 2014

Pandas is an open-source, fast and flexible python library which is used to import the data. In data preprocessing stage, the raw data is transformed into desirable format as shown in figure 3.



	Date/Time	Lat	Lon	Base	dom
564511	2014-04-30 23:22:00	40.7640	-73.9744	B02764	30
564512	2014-04-30 23:26:00	40.7629	-73.9672	B02764	30
564513	2014-04-30 23:31:00	40.7443	-73.9889	B02764	30
564514	2014-04-30 23:32:00	40.6756	-73.9405	B02764	30
564515	2014-04-30 23:48:00	40.6880	-73.9608	B02764	30

Fig. 3. Data Frame format

2. Data Analysis

In this model DOM (Date of the month) is calculated to find the day in which the most number of trips had taken place.

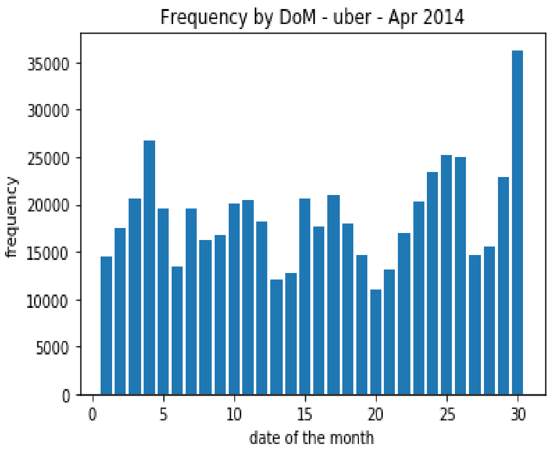


Fig. 4. Frequency by Date of the month

Calculating the frequency by hour is useful for finding the time in which the most number of trips had taken place

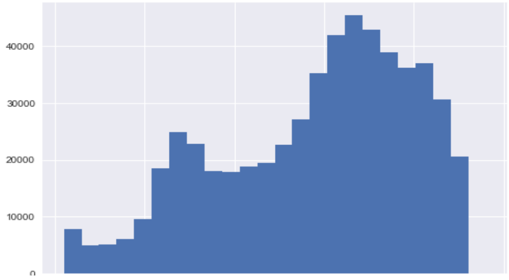


Fig. 5. Analyze the hour

Next we are analyzing the weekday in which the most number of trips had taken place.

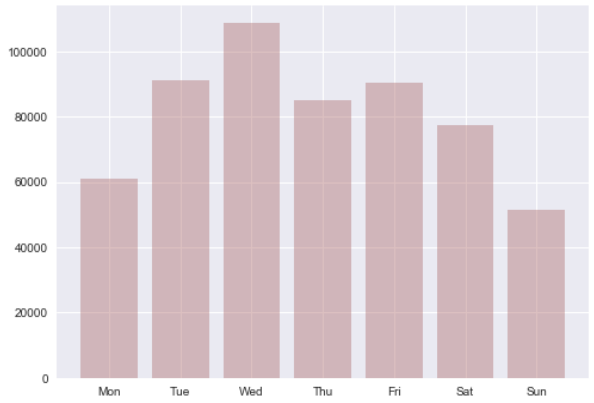


Fig. 6. Analyze the Weekday

Cross analysis is performed over hours of a day and days of a week with the help of a heatmap. Seaborn is a python package that is used to create a heatmap.

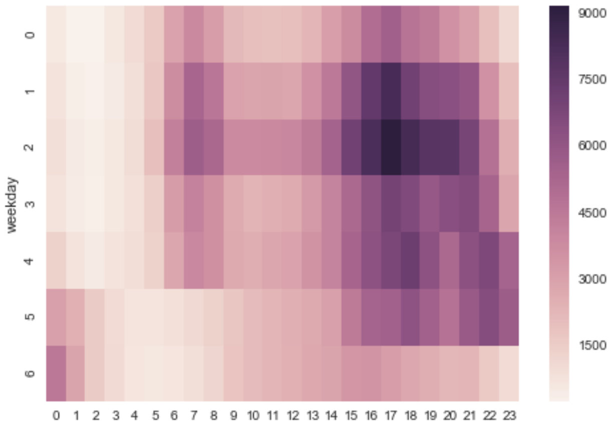


Fig. 7. Cross analysis (hour, Dow)

Figure 8 shows the number of transactions that had taken place in the corresponding latitude.

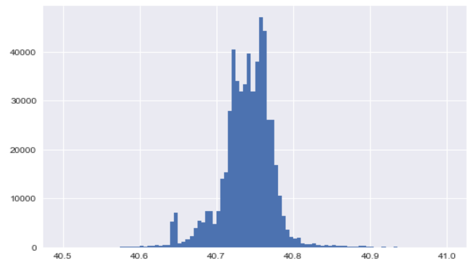


Fig. 8. Frequency by latitude

Figure 9 shows the number of transactions that had taken place in the corresponding longitude.

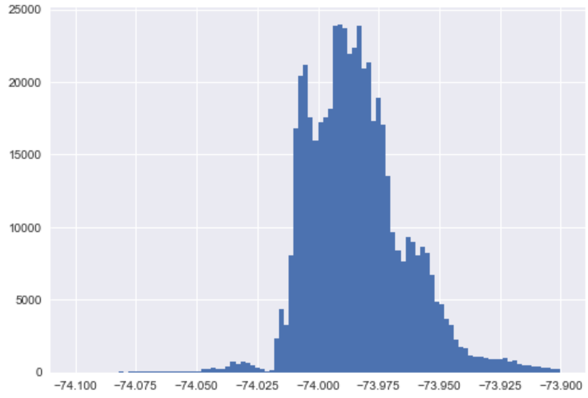


Fig. 9. Frequency by longitude

Fig 10 shows the number of transactions that had taken place over both the latitude and the longitude.

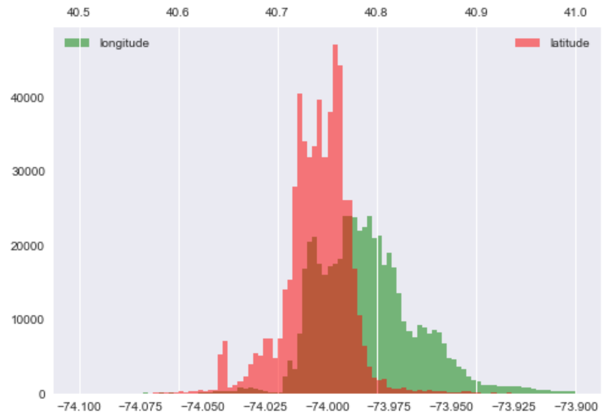


Fig. 10. Frequency by considering both latitude and longitude

IV. CONCLUSION

In this paper we have made an attempt to use histogram based approach to analyze the Uber dataset. From this we come to know the busy zones in the interior of the city where more number of transactions take place. This information will be useful in knowing the taxi necessities of each zone within the city limits and also in the outskirts. The peak hours calculated is helpful in predicting the taxi necessities at each hour of a day which is useful in planning the time of the taxi pickup accordingly. The latitude and the longitude coordinates are helpful in allocating the number of taxis per region based on the location and the business of that zone. The above information is useful in providing a better customer service.

REFERENCES

[1] Kai Zhao, Denis Khryashchev, Juliana Freire,” Predicting Taxi Demand at High Spatial Resolution: Approaching the Limit of Predictability”, 2016 IEEE International Conference on Big Data

[2] Lasse Korsholm Poulsen1, 2016,” Green cabs vs. Uber in New York City” IEEE International Congress on Big Data”, DOI 10.1109/BigDataCongress.2016.35

[3] N. Mukai and N. Yoden, Taxi Demand Forecasting Based on Taxi Probe Data by Neural Network. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 589–597.

[4] S. Silverstein, “These animated charts tell you everything about Uber prices in 21 cities,” http://www.businessinsider.com/uber-vs-taxi-pricing-by-city-2014-10?IR=T, 10 2014. 1

[5] A. Tangel, “Green taxis gaining ground in New York city,” http://www.wsj.com/articles/ green-taxis-gaining-ground-in-new-york-city-1403145481, 06 2014. 1

[6] Saravana M K, Harish K,” A Case Study on Analyzing Uber Datasets using Hadoop Framework”, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

[7] A. Anwar, M. Volkov, and D. Rus, “Changinow: A mobile application for efficient taxi allocation at airports,” in ITSC 2013, Oct 2013, pp. 694–701.

[8] Y. Huang and J. W. Powell, “Detecting regions of disequilibrium in taxi services under uncertainty,” in SIGSPATIAL’12, Redondo Beach, CA, USA, November 7-9, 2012, 2012, pp. 139–148.

[9] K. Zhao, M. P. Chinnasamy, and S. Tarkoma, “Automatic city region analysis for urban routing,” in IEEE ICDMW 2015, Atlantic City, NJ, USA, November 14-17, 2015, 2015, pp. 1136–1142.

[10] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, “Predicting taxi-passenger demand using streaming data,” IEEE Trans. Intelligent Transportation Systems, vol. 14, no. 3, pp. 1393–1402, 2013

[11] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, “Prediction of urban human mobility using large-scale taxi traces and its applications,” Frontiers of Computer Science in China, vol. 6, no. 1, pp. 111–121, 2012.

[12] A. Hern, “New York taxi details can be ex- tracted from anonymised data,researcherssay,”http://www.theguardian.com/technology/2014/jun/7/new-york-taxi-details-anonymised-data-researchers-warn, 06 2014. 8

[13] Berti-Equille Laure ; Bonifati Angela ; Milo Tova,” Machine Learning to Data Management: A Round Trip,” 2018 IEEE 34th International Conference on Data Engineering (ICDE), 10.1109/ BigData Congress. 2016.35