

# Statistical Analysis of Transport Frequency of Public Utility Service Vehicles

Dr. M. RameshKumar<sup>1</sup> S. Dhanakoti<sup>2</sup> S. Gayathri<sup>3</sup> T. Hemalatha<sup>4</sup>

<sup>1</sup>Assistant Professor <sup>2,3,4</sup>B.E Student

<sup>1,2,3,4</sup>Department of Computer Science and Engineering

<sup>1,2,3,4</sup>Dhirajlal Gandhi College of Technology, Tamil Nadu, India

**Abstract**— Nowadays the accumulation of the people in urban areas created the challenge for transportation. By analyzing the large amount of data, we can get the fact about real situation that happening in the current world. The idea is to provide the insights that will help us to resolve the challenges. Some of the challenge of transportation in our regular life such as traffic congestion, longer commuting, automobile dependency, inefficiency in public transport. To resolve the above-mentioned challenges, analysis is carried out based on week, day, hour and the ride frequency of the public utility service vehicles such as buses, taxis, metro trains, Flights etc. By using these insights, decision making can be performed for improving productivity and efficiency for both government and private sectors. The visualization of the data helps any one to understand the usage frequency of public utility vehicles.

**Keywords:** Data Analysis, Visualization, Public Vehicles, Ride Frequency

## I. INTRODUCTION

As the transport increases in the modern era, because of the economic development and enlarged residents, Transportation has become the requisite for everyone. The attention is required in the factors like time, traffic, number of passengers and geographical location.

More over the people are using different modes of vehicles for their mobility such as bus, car, bicycles, train etc., when comparing with all other modes of conveying, public transports are the most commonly used and cost-efficient transport. Henceforth this project deals with analysis of public transport data. Here categorization of the data based on the weekday, months, hours and frequency is performed.

The outcome of the survey is based on exploratory visualisation. This visualisation allows the user to visualize a different kind of data for the better understanding of the transport in the urban cities [9]. The examination is done by using python.

### A. Data Collection

Data are collected in the form of CSV (Comma Separated Values) format. Other than using CSV, JSON and XLS file formats can also be used. In case of unavailability it can be scrapped from data providers website. The collected data set contains pickup location, date, time, longitude, latitude etc., Datasets are collected approximately for five months along with time. Datasets are included around 5021 records of trips.

### B. Data Analysis

Data analysis is the process of collecting, cleaning, transforming and modelling the data to provide the facts about the huge amount of data which is collected. The

resultant insight of the examination is used for decision making, better understanding and allows us to take prevention measures. The basic classification of the data analysis is Descriptive analysis, Predictive analysis, Diagnostic analysis, Prescriptive analysis.

In this project, we have used the Descriptive analysis. The reason for this is it uses the past data and finds the insights of what is happened earlier.

## II. RELATED WORKS

Data Analysis has a certain impact in each and every department of the modern era. This helps various organizations or individuals in making decisions to increase efficiency of their service.

Now the below things will show about the existing system and their drawbacks of the system of them.

### A. Analysis of Performance of Local Taxi Service using R

Utilizing the travel and Booking logs available in a particular taxi service, this system performs preprocessing and analysis on the provided data to produce relationships between factors to uncover hidden facts to improve efficiency of the service provided. This system makes use of R technology for processing and analysis whereas it requires more time for processing the data. And it provides less option for on spot up gradation. It is not extensible to other systems.

## III. PROPOSED SYSTEM

The proposed system uses python as a technology for extraction, transform, load and processing of the data. The project makes use of the various open source technologies for analysing and producing the visual result. Moreover, the single interface is used for ETL (Extract Transform Load) and result is used. Instead of using the R, Python is used to increase the flexibility, efficiency and decrease the time complexity. In addition to that python can be extended as user-friendly interfaces such as web-GUI.

## IV. IMPLEMENTATION

The implementation contains the statistics and analysis of the data set.

The implementation of this project involves the process of loading dataset, mining the data, and presenting visual representation.

### A. Algorithm

The algorithm we have used in the system such as

#### 1) Linear Regression

Regression is using set of statistical models for estimating relationships between dependent variable and one or more independent variables.

We will use the simple linear regression to predict the relationship between the attribute, as we are using the two main independent variable (days, weeks, hours) and the dependent variable (number of rides).

### B. Module Description

The following system flow is the procedure used in this analysis.

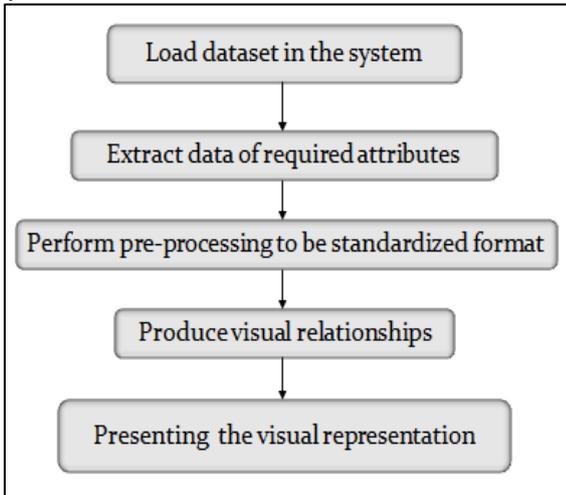


Fig. 1.1: Work Flow

#### 1) Loading the Dataset in the System

The data is collected and imported in the form of CSV format. Before analysing the data, the data transformation is carried out. The collected datasets are combined and the date/time format need to be mapped according to the respective platform. In the imported data there are eight attributes which has been shown below.

#### 2) Preprocessing

The pre-processing process involves removal of duplicates, filling missing data etc. from the extracted data. By using Python commands, we can eradicate the redundant data. After acquiring the clean dataset, the date and time are mapped by utilizing date time functions in pandas. Then Detailed comparison in performance of government and private transport have also been performed.

#### 3) Extraction of Required Attribute

Using pandas, the numerous different required attributes are extracted and the conversion of those attributes is performed.

	Date/Time	BusStopCode	RoadName	Description	Latitude	Longitude	DOM	Weekday	Hour
0	2014-09-01 00:01:00	481	Woodlands Rd	BT PANJANG TEMP BUS PK	1.383764	103.758300	1	0	0
1	2014-09-01 00:01:00	1012	Victoria St	Hotel Grand Pacific	1.296848	103.852536	1	0	0
2	2014-09-01 00:03:00	1013	Victoria St	St. Joseph's Ch	1.297710	103.853225	1	0	0
3	2014-09-01 00:06:00	1019	Victoria St	Bras Basah Cplx	1.296990	103.853022	1	0	0
4	2014-09-01 00:11:00	1029	Nth Bridge Rd	Cosmic Insurance Bldg	1.296673	103.854414	1	0	0
...	...	...	...	...	...	...	...	...	...
5016	2014-09-05 10:42:00	99139	Changi Village Rd	Blk 5	1.388195	103.987234	5	4	10
5017	2014-09-05 10:42:00	99161	Nicoll Dr	Bef Changi Beach CP 3	1.390262	103.992957	5	4	10
5018	2014-09-05 10:42:00	99171	Nicoll Dr	Changi Beach CP 2	1.391128	103.991021	5	4	10
5019	2014-09-05 10:46:00	99181	Telok Paku Rd	Bef S'pore Aviation Ac	1.387754	103.986503	5	4	10
5020	2014-09-05 10:47:00	99189	Telok Paku Rd	S'pore Aviation Ac	1.388414	103.989716	5	4	10

Table 1: Data after extraction

#### 4) Identifying Visual Relationships

By making use of plotting functions in pandas, seaborn and matplotlib, we can bring out the visual relationship. To predict the result, we are utilising exploratory visualisation as the resultant is unknown.

#### 5) Presentation of Visual Relationships

Initially hour-based analysis is performed. The frequency of the rides in the particular time is shown in the following result.

The analysis of weekday and the hour is performed to reveal the relationship between them. It provides the fact that, on which day the frequency is high and low.

In the urban city it is important to find the which place has the highest frequency number of rides. By which we can increase the transport facilities in the particular location.

## V. RESULTS

Hence the result of the project provides the visualization of the collection datasets, which can be understandable by all people. The statistical result of analysing the data shows that the weekday has the frequent usability of the vehicles. Furthermore, the location which has the highest usability of the city is provided in the result. The statistical analysis is carried out based on the week, days and hours. The implementation time and the extensibility are achieved.

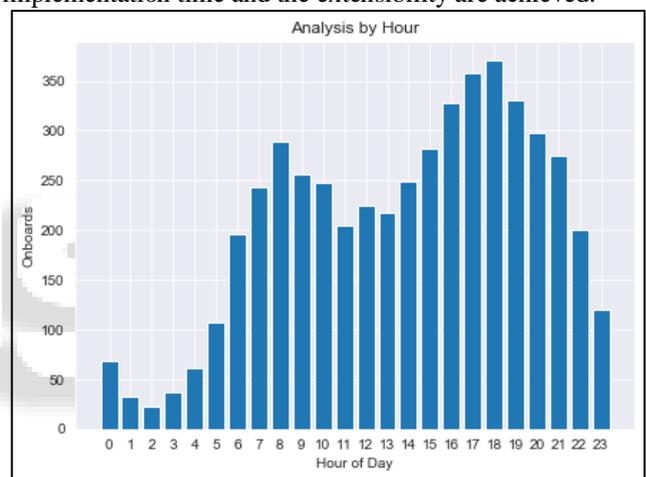


Fig. 2.1: Relationship between Rides and Hour of day

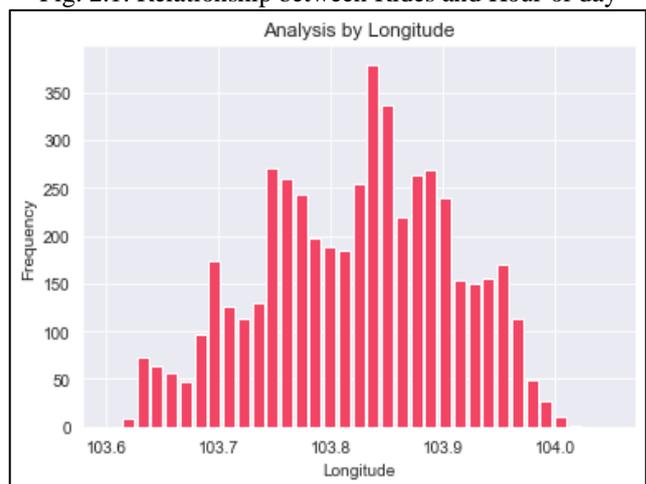


Fig. 2.2: Relationship between Location and Rides

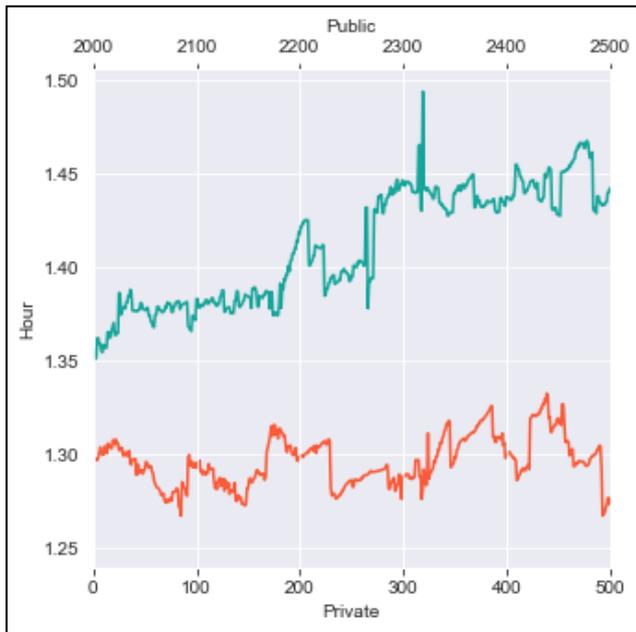


Fig. 2.3: Comparison of Public and Private Rides

## VI. FUTURE WORK

The System uses Python technology for Processing and Analysis of data. Thus, this can be easily extended to user friendly interfaces such as web and desktop application with the of various other Python Frameworks and Libraries such as Flask, Django etc.

Other than performing timeseries analysis on historical data, this can also be modified to perform analysis on real-time data by using python supported packages and libraries.

## VII. CONCLUSIONS

This massive increase in transportation using public utility vehicle services provided by both government and private sectors has led to generation of huge amount of data and this brings in the requirement for efficiency in their services provided.

This System performs analysis on the available historical data to produce relationships between the factors to uncover hidden facts which can be presented in the form of plots and charts to be used in decision making.

## ACKNOWLEDGMENT

We are thankful to Mr. S. S. Aravinth, Assistant Professor, Department of Computer Science, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India for his suggestion and encouragement in this work.

His expert suggestions and scholarly feedback had greatly enhanced the effectiveness of this work.

## REFERENCES

- [1] Jing Guo, Dr. Candace Brakewood, Dr. Cheng Liu, Dr. Kcouwai Wong, Analysis and comparison of Uber, Taxi and Uber request via Transit, August 3, 2018.
- [2] Aguinaldo Bezerra, Gisliany Alves, Ivanovitch Silva, Pierangelo Rosati, Patricia Takako Endo, Theo Lynn, A

Preliminary Exploration of Uber Data as an Indicator of Urban Liveability, International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA).

- [3] Daniel Hienert, Philipp Mayr, Visualizations in Exploratory Search – A User Study with Stock Market Information, GESIS – Leibniz Institute for the Social Sciences.
- [4] Ásdís Ólafsdóttir, Bus Service Performance Analysis, Master of Science Thesis Stockholm, Sweden 2012.
- [5] Nadia P. Kozievitch, Tatiana M. C. Gadda, Keiko V. O. Fonseca, Marcelo O. Rosa, Luiz C. Gomes-Jr, Monika Akbar, Exploratory Analysis of Public Transportation Data in Curitiba
- [6] Widdoes, Kaylene, "Case Study of Uber Data in the Central London Area" (2016). Honors Research Projects. 411.