International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 1 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT19514

Airline Data Analysis

Navuluri Madhavilatha[1], Bheema Shireesha[2], Chunduru Anilkumar[3]

[1]Guest Faculty, Department of Computer Science and Engineering, Dr. APJ Abdul Kalam IIIT Ongole, Andhra Pradesh, India

[2, 3]Assistant Professor, Department of Computer Science and Engineering, Dr. APJ Abdul Kalam IIIT Ongole, Andhra Pradesh, India

ABSTRACT

In the contemporary world, Data analysis is a challenge in the era of varied inters disciplines through there is a specialization in the respective disciplines. In other words, effective data analytics helps in analyzing the data of any business system. Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. The goal of our project is to get monthly wise statistics of airline data and taking particular airport as target we are further analyzing the data to get the hourly statistics. And also we are finding out the most popular source-destination pairs and calculating the average delays at every airport. The data for this project comes from the stat-computing.org website. In particular, in the year 2008 data 70,09,728 titles recorded there which includes information on the Origin, Destination, Month, Year, DayofWeek, DayofMonth, DepDelay, ArvDelay, DepTime, ArvTime and a few other less interesting variables. Conveniently, you can export the data directly as a csv file.

Keywords : Machine Learning, Data Mining, Big Data, Statistics, Data Visualization, Data Analytics, ASCII, SRS, GNU

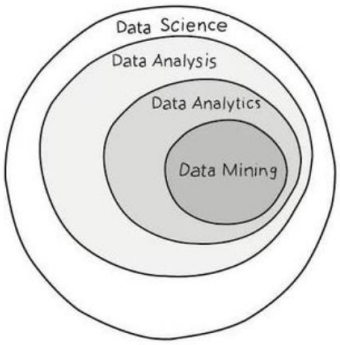
I. INTRODUCTION

The goal of our project is to get monthly wise statistics of airline data and taking particular airport as target we are further analyzing the data to get the hourly statistics. And also we are finding out the most popular source-destination pairs and calculating the average delays at every airport [1]. The data for this project comes from the statcomputing.org website in particular, in the year 2008 data 70,09,728 titles recorded there which includes information on the Origin, Destination, Month, Year, DayofWeek, DayofMonth, DepDelay, ArvDelay, DepTime, ArvTime and a few other less interesting variables. Conveniently, you can export the data directly as a csv file.

Data:

Data is a set of values of qualitative or quantitative variables. It can representation of facts as text, numbers, graphics, images, sound or video. Raw data, also known as primary data, is collected from a source. If a scientist sets up a computerized thermometer which records the temperature of a chemical mixture in a test tube every minute, the list of temperature readings for every minute, as printed out on a spreadsheet or viewed on a computer screen is "raw data" [2]. Raw data can be input to a computer program or used in manual procedures such as analyzing statistics from a survey. The term "raw data" can refer to the binary data on electronic storage devices, such as hard disk drives (also referred to as "low-level data"). All software is divided into two general categories: data and programs. Programs

are collections of instructions for manipulating data. Strictly speaking, data is the plural of datum, a single piece of information. In practice, however, people use data as both the singular and plural form of the word [3]. The term often used to distinguish binary machine-readable information from textual human-readable information. For example, some applications make a distinction



Between data files (files that contain binary data) and text files (files that contain ASCII (American Standard Code for Information Interchange) data.

In database management systems, data files are the files that store the database information, whereas other files, such as index files and data dictionaries, store administrative information, known as metadata. The seven V's sum it up pretty well – Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value [4]. Science is the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics.

Difference between data mining and data science

Data mining refers to the science of collecting all the past data and then searching for patterns in this data. You look for consistent patterns and / or relationships between variables. Once you find these insights, you validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction [5]. Data Science is an umbrella that contain many other fields like Machine learning, Data Mining, big Data, statistics, Data visualization, data analytics etc.

II. METHODS AND MATERIAL

System Requirement Specification

2.1 Software requirement specifications

It specifies the hardware and software requirements that are required in order to run the application properly. The Software Requirement Specification (SRS) is explained in detail, which includes overview of this dissertation as well as the functional and non-functional requirement of this dissertation.

Functional Requirements: R Studio

Non-Functional Requirements:Dataset (.csv file)

2.2 Hardware requirements:

Processor : Any Processor above 500 MHz
Ram : 2 GB
Hard Disk : 10 GB
Input device : Standard Keyboard and Mouse
Output device : VGA and High Resolution Monitor.

III. Technology Description

3.1 Introduction to R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc) and graphical techniques, and is highly extensible [6]. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

3.2 The R Environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities[7]. The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R is designed around a true computer language, and it allows users to add additional functionality by defining new functions. For computationally-intensive tasks, C, C++ and FORTRAN code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly. Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics. R has its own Latex-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

3.3 R Studio

R Studio is an integrated development environment (IDE) for the R programming language. Some of its features include: Customizable workbench with all of the tools required to work with R in one place (Console, source, plots, workspace, help, history, etc.). Syntax highlighting editor with code completion.

Execute code directly from the source editor (line, selection, or file). Full support for authoring Sweave and TeX documents.

Runs on all major platforms (Windows, Mac, and Linux) and can also be run as a server, enabling multiple users to access the R Studio IDE using a web browser.

IV. System Design

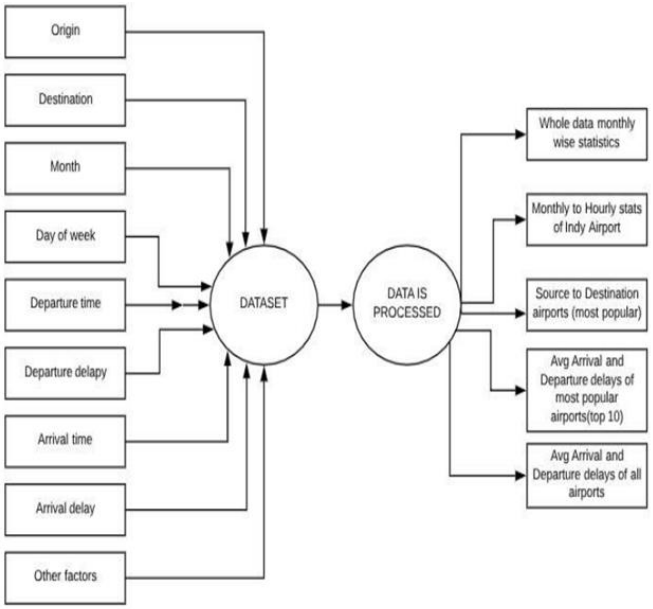
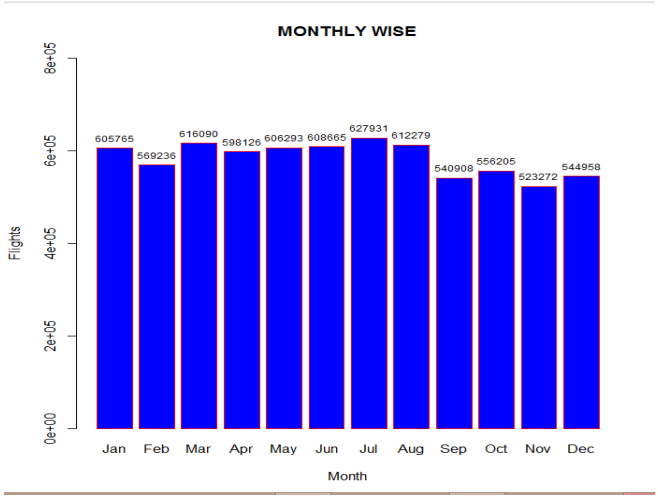


Fig. 4.1 Analysis of airline data

Data set (.csv file) contains 29 columns as "origin, destination, month, day of week departure delay, departure time, arrival time, arrival delay etc". By taking data set as an input we generate the graphs which can be easily understood by the client.

V. Experimental Results

5.1 Monthly wise statistics of entire Airline data



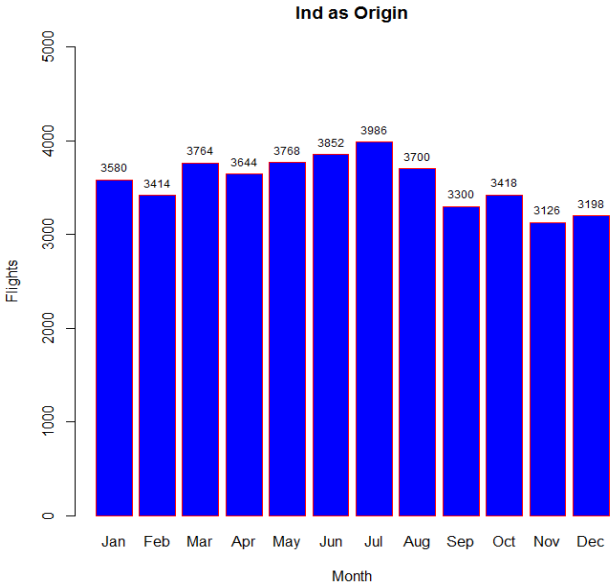
The above graph is drawn by taking Months on X-axis and Number of Flights on Y-axis. From the graph we can conclude that most number of flights (627931) are running in July month. And second most number of flights (616090) are running in March month.

5.2 Flight statistics of Indy airport

Here we are going to take Indy airport as an Origin as well as destination calculating monthly wise, weekly wise, day wise and hourly wise flight statistics.

5.2.1 Monthly wise statistics of Indy airport

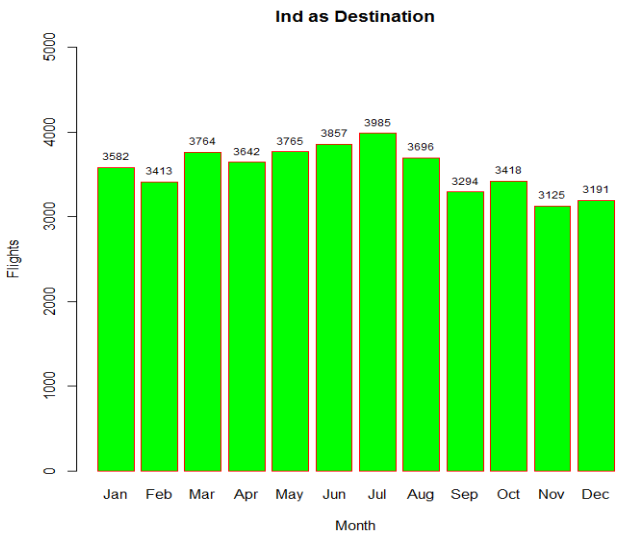
Indy as Origin:-



From the above graph we can conclude that more number of flights are leaving Indy airport in the month of July and the less number of flights in the month of November.

Indy as Destination:

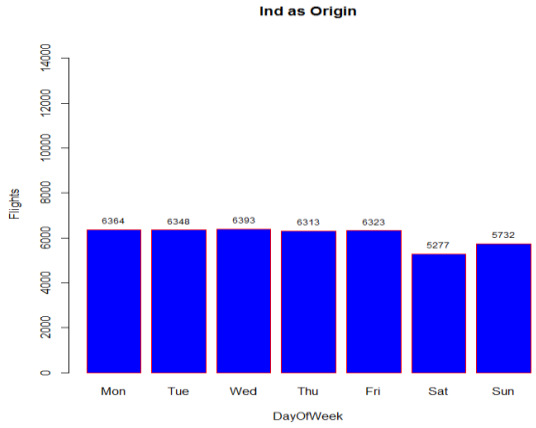
From the below graph we can conclude that more number of flights are arriving Indy airport in the month of July and the less number of flights in the month of November.



5.2.2 Weekly wise statistics of Indy airport

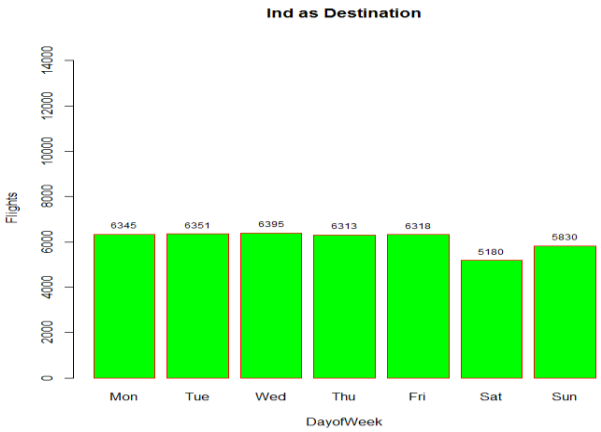
Indy as Origin:-

From the below graph we can conclude that more number of flights are leaving Indy airport on Wednesday and the less number of flights on Saturday in a week.



Indy as Destination:-

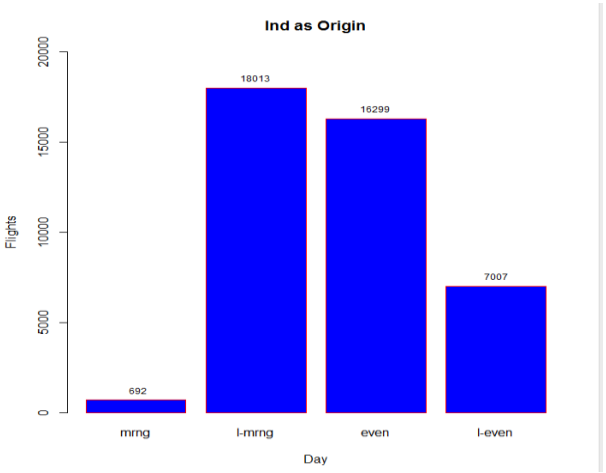
From the below graph we can conclude that more number of flights are arriving Indy airport on Wednesday and the less number of flights on Saturday in a week.



5.2.3 Day wise statistics of Indy airport

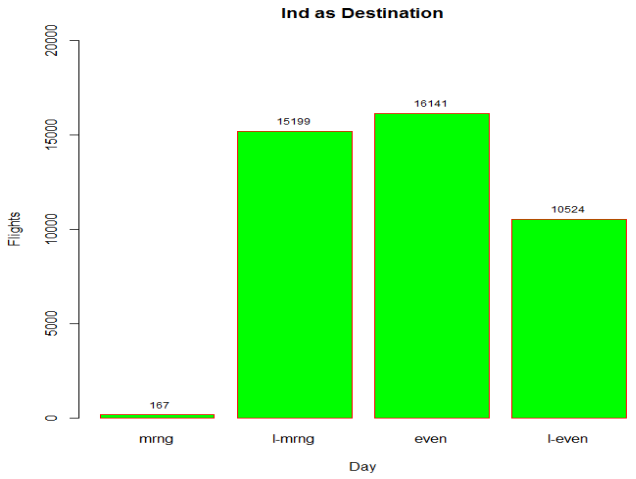
Indy as Origin:-

From the below graph we can conclude that more number of flights are leaving the Indy airport in the duration of 6am – 7am and the less number of flights in the duration of 12am – 4 am in a day.



Indy as Destination:-

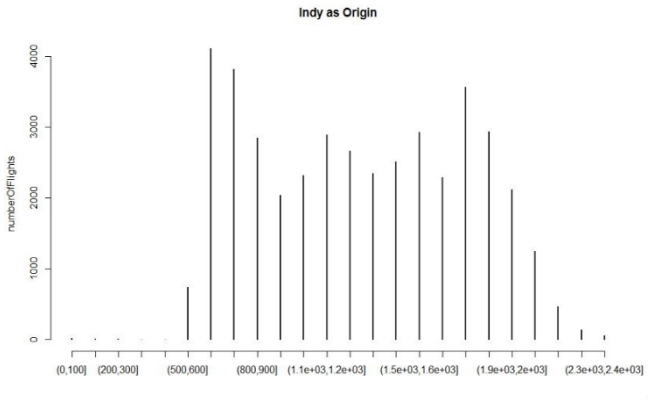
From the below graph we can conclude that more number of flights are arriving the Indy airport in the duration of 10am – 11am and the less number of flights in the duration of 2am – 6 am in a day.



5.2.4 Hourly wise statistics of Indy airport

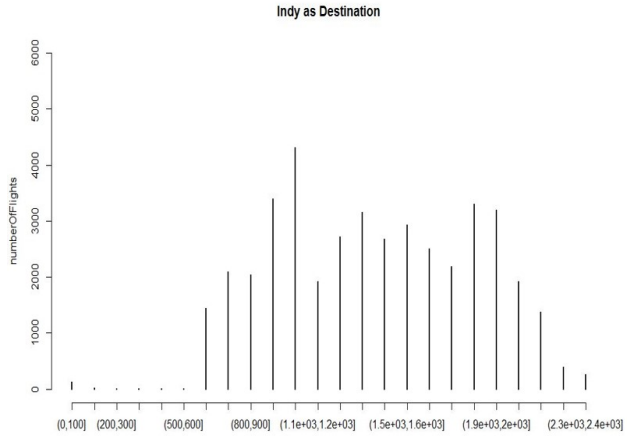
Indy as Origin:-

From the below graph we can conclude that more number of flights are leaving the Indy airport in the duration of 6am – 7am and the less number of flights in the duration of 12am – 4 am in a day.



Indy as Destination:-

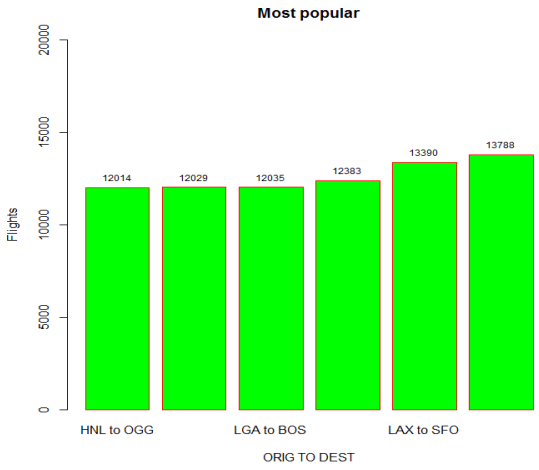
From the below graph we can conclude that more number of flights are arriving the Indy airport in the duration of 10am – 11am and the less number of flights in the duration of 2am.



5.3 Source to Destination airport pairs

Here we are going to take a look at the most and least popular airport-pairs based on the flights operating in between them (Means one as Origin and another as Destination).

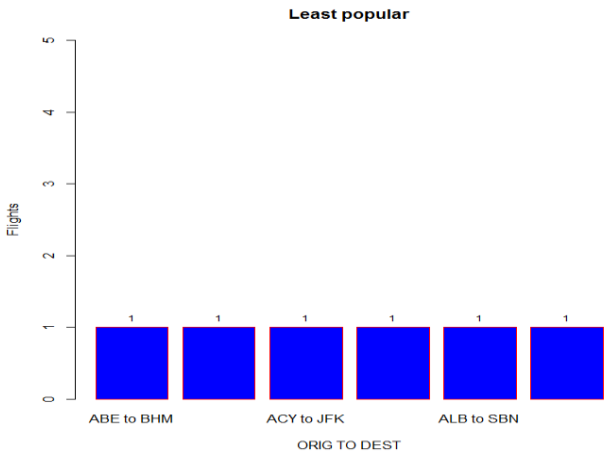
5.3.1 Most Popular



These are the top 6 airport pairs with max number of flights running between them. In these SFO to LAX stands first with 13788 running.

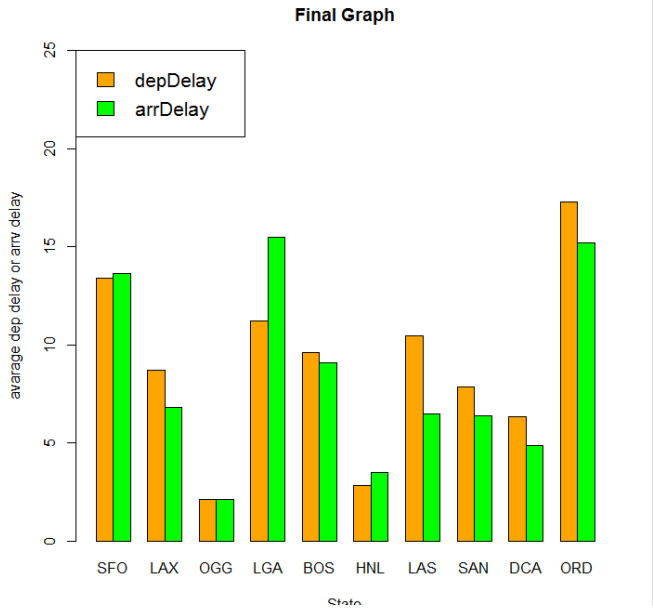
5.3.2 Least Popular

These are the least 6 airport pairs with minimum number of flights running between them.



5.4 Average delays of most popular airports

Here we are comparing the average flight delays among the top 10 most popular airports.

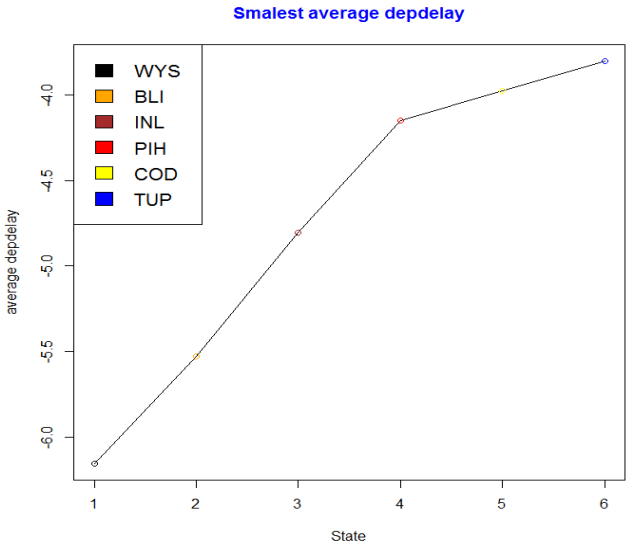


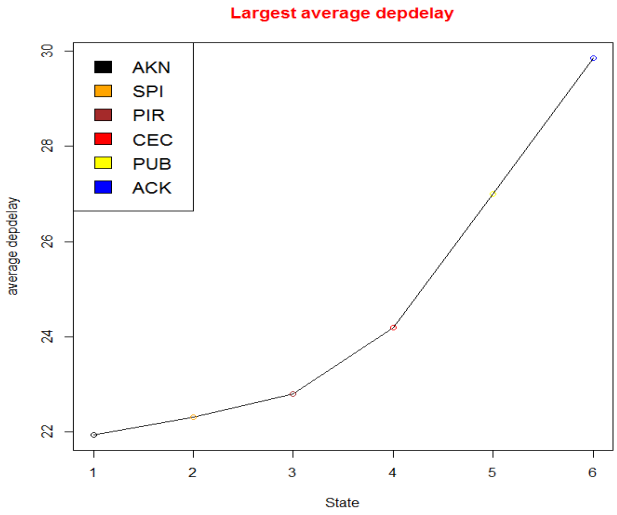
Based on the graph we can conclude that OGG airport running flights with least amount of arrival and departure delays. And ORD airport running flights with high Amount of departure delay. LGA airport running flights with high Amount of arrival delay.

5.5 Average delays for all airports

Here, we are calculating the average delays (both arrival and departure) of all the airports.

5.5.1 Departure delays for all airports

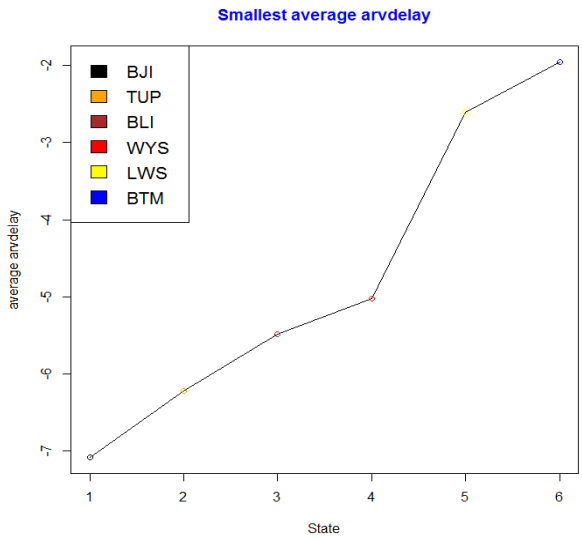


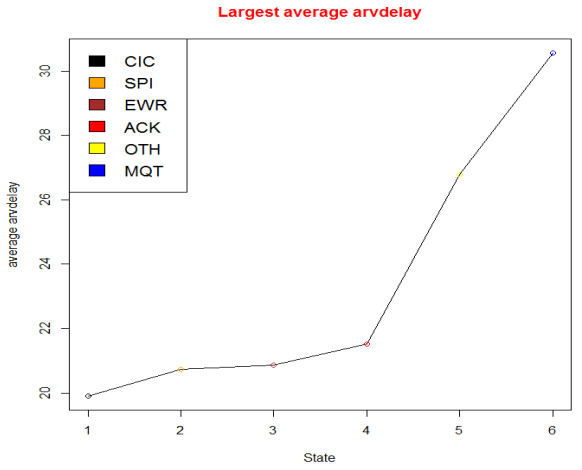


In the above graphs we are representing the largest average and smallest average departure delays of the first six (head) airports.

5.5.2 Arrival delays for all airports

In the below graphs we are representing the largest average and smallest average arrival delays of the first six (head) airports.





VI. CONCLUSION

Not all arrivals can occur when they are scheduled to, then airport congestion happens. The delay distribution of the airport can make it easier to understand the airport delay. We also found the most popular origin to destination pairs according to the number of flights running. We compared the average arrival and departure delays among them as well. And also we have calculated the average delays to all the airports.

VII. FUTURE SCOPE

Further we can implement this project by calculating the monthly wise, weekly wise, day wise and hourly wise flight statistics at popular airports we come to know the busy hours and days in which the passenger flow will be high. Those days we can arrange the alternatives in order to reduce the problems faced by the passengers as well as by the authorities like increasing the security, food and water facilities, transport facilities outside the airport ,runway clearance etc.

VIII. REFERENCES

[1]. Abdelghany, K. F., Abdelghany, A. F., and Raina S., (2004) A model for projecting flight delays during irregular operation conditions, Journal of Air Transport Management, Volume 10, Issue 6, Pages 385-394.

[2]. American Statistical Association(stat-computing.org)

[3]. Thearling, K., Data Mining and Analytic Technologies, www.thearling .com, 2004.

[4]. Tutorials point (http://www.tutorilaspoint.com/r/)

[5]. Aisling, R., and J.B. Kenneth, (1999) An assessment of the capacity and congestion levels at European airports, ERSA conference papers ersa 99, pages 241, European Regional Science Association

[6]. Bureau of Transportation Statistics, Airline On-Time Statistic. U.S. Department of Transportation. Washington, D.C. http://www.bts.gov/programs/airline information

[7]. Hansen, M., and C. Y. Hsiao (2005), Going South? An Econometric Analysis of US Airline Flight Delays from 2000 to 2004, Presented at the 84rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2005.

Cite this article as :

Navuluri Madhavilatha, Bheema Shireesha, Chunduru Anilkumar, "Airline Data Analysis", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 1, pp. 22-29, January-February 2019.

Available at doi :

https://doi.org/10.32628/CSEIT19514

Journal URL : http://ijsrcseit.com/CSEIT19514