Analysis of Airline Datasets

A Project report submitted in partial fulfilment of the requirements of the award of the degree of

**Bachelor of Technology**

# in Computer Engineering

by

# Rishabh Jain, Reg No: PCE20CS155 Sharma Chetan, Reg No: PCE20CS208 Tushar Sharma, Reg No: PCE20CS189 Sarthak Bharadwaj, Reg No: PCE20CS167

under the guidance of

**Mrs. Neha Shrotriya, Assistant Professor**



(Session 2021-22)

**Department of Computer Engineering**

**Poornima College of Engineering**

ISI-6, RIICO Institutional Area, Sitapura, Jaipur – 302022

**December, 2021**

**Department Certificate**

This is to certify that **Mr. Rishabh Jain, Mr. Sharma Chetan, Mr. Tushar Sharma & Mr. Sarthak Bharadwaj** Registration no. **PCE20CS155**, **PCE20CS208, PCE20CS189, PCE20CS167** of the **2nd Year** Department of Computer Engineering, has submitted this project report entitled **“Analysis of Airline Datasets”** under the supervision of Dr. / Prof. **Sonam Gour**, working as **Assistant Professor** in department of Computer Engineering as per the requirements of the Bachelor of Technology program of Poornima College of Engineering, Jaipur.

Dr. Surendra Kumar Yadav Ms. Sonam Gour Head, Dept. of Computer Engineering Coordinator-Project

# CANDIDATE’S DECLARATION

I hereby declare that the work which is being presented in this project report entitled **“Analysis of Airline Datasets**” in the partial fulfillment for the award of the Degree of Bachelor of Technology in (Computer Engineering), submitted in the Department of Computer Engineering, Poornima College of Engineering, Jaipur, is an authentic record of my own work done during the period from July 2021 to December 2021 under the supervision and guidance of **Sonam Gaur, Assistant Professor.**

We have not submitted the matter embodied in this project report for the award of any other degree.

Dated:

Place: Jaipur

## SUPERVISOR’S CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Signature)

Dated: Mrs. Sonam Gour

Place: Jaipur Assistant Professor

# ACKNOWLEDGEMENT

I would like to convey my profound sense of reverence and admiration to my supervisor **Neha Shrotriya, Assistant Professor Department of Computer Engineering, Poornima College of Engineering,** for his intense concern, attention, priceless direction, guidance and encouragement throughout this research work.

I am grateful to **Dr. Mahesh Bundele**, Director of Poornima College of Engineering for his helping attitude with a keen interest in completing this dissertation in time.

I extend my heartiest gratitude to all the teachers, who extended their cooperation to steer the topic towards its successful completion. I am also thankful to non-teaching staff of the department to support in preparation of this dissertation work.

My special heartfelt gratitude goes to **Dr. Surendra Kumar Yadav, HOD,Department of Computer Engineering, Ms. Sonam Gour, Project Coordinator, Department of Computer Engineering, Poornima College of Engineering**, for unvarying support, guidance and motivation during the course of this research.

I would like to express my deep sense of gratitude towards management of Poornima College of Engineering including **Dr. S. M. Seth**, Chairman Emeritus, Poornima Group and former Director NIH, Roorkee, **Shri Shashikant Singhi**, Chairman, Poornima Group, **Mr. M. K. M. Shah**, Director Admin & Finance, Poornima Group and **Ar. Rahul Singhi**, Director Poornima Group for establishment of institute and providing facilties my studies.

I would like to take the opportunity of expressing my thanks to all faculty members of the Department, for their kind support, technical guidance, and inspiration throughout the course.

At last but not least I would like to give special thanks to God who enabled me to complete my dissertation on time.

## Rishabh Jain , Department of Computer Engineering, PCE20CS155 Sharma Chetan , Department of Computer Engineering, PCE20CS208 Tushar Sharma , Department of Computer Engineering, PCE20CS189 Sarthak Bharadwaj , Department of Computer Engineering, PCE20CS167

**LIST OF FIGURES**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Fig. No.** | **Description** | **Page**  **No.** |
| 1. | 1.1 | Cancelled flight by month |  |
| 2 | 1.2 | Cancelled flight by year |  |

**ABSTRACT –**

In this paper, the analysis of the airline data set is performed using Microsoft Azure HDInsight which runs Hadoop in the cloud. Hive and Hive QL statements have been used for querying the data. Data visualization has been done by extracting the output of the HIVE query in excel and plotting the data using line and scatter plot charts. The visualization of the data shows some patterns that exist between flight diversions and flight distance, flight cancellation and flight distance and so forth. Keywords: Hadoop, HDInsight, Big Data, Hive, Data Visualization, Data Analysis

**1. INTRODUCTION** -

There is no doubt that a lot of excitement exists with the term Big Data. Big Data in simple words can be large-scale data which does not have a well-defined structure. The size of the data itself is so huge that it is not practically easy for a single computer to store and process all the data by itself. Traditional computing approached the problem in a different way, the focus was always to increase the processing speed and power of the computer. As the data grows exponentially, the processing power of the single computer becomes a bottleneck and thus a new approach was needed to address the issue at hand. A new way was developed where many non-expensive commodity computers all worked together in harmony with each other, in order to store and process this big data in parallel that allows us to extract meaningful information from a large data set. Moreover, current technologies using the cloud infrastructure allows us to easily create clusters of computers by renting them for as much time as required and then releasing the computing resources when no longer needed. Thus with cloud technologies we get the computing power of the clusters of computers with minimal investment. The airline data has been taken from the United States Department of Transportation, Bureau of Transportation Statistics [1]. The data consists of the arrival and departure records of all US domestic flights from the period 2012 to 2014. Section 2 gives a brief introduction of the Airline Data set, Hive and Hindsight. Section 3 describes the mechanism by which the data set is analyzed. Section 4 describes the experimental observations from the data set. Section 5 is the conclusion of this paper.

Problem statement :-

1) Big amount of data generated on hourly basis.

2) There are many active fights

3) Many flights delay daily

The data that is maintained is big in size and it is increasing. Processing the data multiple times is a time taking process. Visualization tools need to fetch the data in real time and the graphs or charts made on top of data need to be updated quickly.

Internal data containing real-time information about an airline’s resources and their scheduled utilization over time, and external data for factors such as current and future weather forecasts, competitor activities, and air traffic control, are necessary for efficient operations (Nathans, 2015). These bits of information and data must be readily available and accessible to represent drivers

and constraints for scenarios induced by irregular operations, so as to facilitate the development of effective self-governing platforms for airline disruption management. In addition, whenever a new airline system is replaced or upgraded, new data sources are typically integrated into the existing framework (Gershkoff, 2016). The new data must be maintained for both existing and new applications, and thus present cost-intensive challenges for mitigating disruption because many facets of the airline infrastructure are impacted.

**1.1. The Problem**

While there have been consistent improvements to the existing decision support systems used by human controllers in the Airline Operations Control Center (AOCC), two factors have continued to limit the performance of the disruption resolutions that are applied. First, the decision support systems do not explicitly proffer solutions to specific schedule disruptions. As such, human controllers in the AOCC are required to be reactive in addressing disruptions by using their best judgement based upon their prior experience from resolving the same (or similar) disruption. Second, majority of the decision support (computer) systems used by multiple departments (including the AOCC) in an airline and other air transportation stakeholders (e.g. airports) are not designed or developed at the same time nor by the same vendor (Nathans, 2015). As such, information and data are required to be entered into multiple computer systems thereby exposing human controllers in the AOCC to data input problems and errors. Therefore, information entered into a decision support system for disruption management may be out of sync with other systems and yield incorrect decisions due to lack of data integrity.

Furthermore, extant literature on the adoption of machine learning for decision support systems have only focused on addressing irregular operations for system-wide disruption management in the national airspace system as a whole.

In other words, existing literature address irregular operations for system-wide disruption management by exploiting airport data from multiple airlines in the air transportation system. For instance, Liu et al. (2019) used machine learning to analyze air traffic management actions for incidences of ground delay programs across major airports in the United States. The authors employed a two-stage framework, such that the first stage used support vector machines to correlate incidences of ground delay programs with regional convective (weather) activity while the second stage trained logistic regression and random forest models by using a weather score obtained from the first stage. In addition, presented an approach for estimating aggregate flight departure delay at airports in China through supervised learning models. They applied four separate types of airport-related aggregate characteristics (including weather) to predict expected departure delays at a major airport in China, by employing linear regression, support vector machines, randomized trees, and Light GBM.

While the models presented in both research studies provided acceptable predictive capacity for disruption management at the airport level in the respective air transportation networks, they (i.e. the models) do not describe the disruption management proclivities of a decision support system commissioned by a specific airline in the air transportation network.

In the bid to improve data integrity and fidelity for existing decision support systems, airlines have significantly invested in creating better localized data collection platforms within their respective organizations, which can amass information from different sources within and outside the organization that is easily accessible through a centralized data server (Amadeus IT Group, 2016).

As such, there is a need to fully leverage the ubiquity and accessibility of information (data) collected by existing platforms in the AOCC to enhance agile decision-making capabilities of the AOCC during airline disruption management. To that effect, this paper provides a comprehensive discussion on exploratory analysis administered on historical scheduling and operations recovery

data supplied by a major airline in the United States. This exploratory analysis serves as the basis for the development of credible predictive and prescriptive models for airline disruption management.

**1.2. Contributions**

To the best of our knowledge, we provide the first literature that strictly employs the statistics of big data and machine learning to demonstrate the characterization and evaluation of a functional role (i.e. domain) in the AOCC of a major U.S. airline for disruption management. Thus, in contrast and complement to existing literature, our work provides research on irregular operations

and disruption management accustomed to a specific U.S. airline based upon data provided by the airline. To achieve the aforementioned objectives, this paper enhances prior research and literature on irregular operations and airline disruption management through the following contributions:

1. We introduce and explore several abstraction methods for applying, enhancing, and sequestering raw features and labels in a historical data set on airline scheduling and operations recovery from a major U.S. airline, to readily identify relevant cognitive patterns for key drivers during airline

disruption management.

2. We investigate the application of appropriate machine learning techniques for revealing patterns, pertinence, and properties of abstracted data features, which provide necessary a priori information (i.e. beliefs) for Bayesian and pseudo-Bayesian methods. These methods are subsequently employed in future work for developing functional models in an intelligent multiagent system for airline disruption management.

**1.3. Paper Organization**

The next section in this paper provides an overview of the elements of historical scheduling and operations recovery data retrieved from a major U.S. Airline, followed by a section that expansively discuses several relevant and interrelated processes for exploratory data analysis. We conclude with a summary of pertinent findings and areas for further research in Section 4.

**1.4 Data Overview**

The raw data utilized for demonstrating the exploratory analysis discussed in this paper was provided by Southwest Airlines. Like many major U.S. airlines, Southwest Airlines employs an integrated AOCC organization wherein all functional roles share the same physical space (at the airline’s headquarters in Dallas) and are hierarchically dependent on AOCC supervisors for multiple problem dimensions in airline operations recovery (Hagel et al., 2017). As the largest carrier in the United States in terms of originating domestic passengers boarded with more than 4,100 flight schedule operations daily to over 100 destinations, the supervisors (and controllers) at Southwest Airlines Network Operations Control (SWA-NOC) seek to use technology to see the impact of their decisions to make better ones for improved disruption management. For many years, the controllers at SWA-NOC relied on gut instincts to track and understand how their disruption resolution actions cascaded throughout the airline’s network, but could not inform their instincts with data. To address this issue, upper management at SWA-NOC created the Baker workgroup; an integrated team of supervisors and software developers dedicated to improving decision-making during disruption management by developing and enhancing a suite of computerized decision support systems called the Baker tool. In order to better support the Baker tool, the workgroup created an autonomous data collection platform to record flight schedules that are subject and not subject to different disruption incidents in the Southwest Airlines route network.

A brief description of the functional disruption resolution domains (or roles) in SWA-NOC highlighted in Table 1 is as follows:

**1. Customer Hold:**

This functional domain addresses disruptions related to holding aircraft for passengers on inbound flight connections and holding aircraft to accommodate passengers off cancelled and delayed flights. As such, the customer hold functional domain resolves the aircraft and passenger problem dimensions in airline disruption management. Disruption instances for the customer hold domain accounted for about 11% of delayed flight schedules in the Southwest Airlines route network over the one-year period (i.e. September 2016 to September 2017).

**2. Dispatch CSC:**

This functional domain manages disruptions related to flight dispatch activities by the airline that also includes holding flights to accommodate international flight schedule slot times. To that effect, the Dispatch CSC functional domain addresses the aircraft and crew problem dimensions during disruption management, and disruption instances related to Dispatch CSC represented 4% of delayed flight schedules in the airline operations between September 2016 and September 2017.

**3. Flight Operations:**

This functional domain resolves disruptions defined by Pilot (cockpit crew) scheduling activities as they relate to Pilot tardiness and normal aircraft readiness, and addresses the crew problem dimension of airline disruption management. Between September 2016 and September 2017, disruption instances related to Flight Operations represented about 8.5% of delayed flight schedules, 5.7% of cancelled flight schedules, and 13.5% of diverted flight schedules in Southwest Airlines operations.

**4. Fuel Management:**

This functional role in SWA-NOC manages disruptions related to aircraft fueling and other energy administration activities, and addresses the aircraft problem dimension during disruption management. Disruption instances related to Fuel Management between September 2016 and September 2017 represented 1.1% of delayed flight schedules in Southwest Airlines operations.

**5. Ground Operations:**

This functional domain in SWA-NOC manages disruptions defined by several activities ranging from passenger boarding and aircraft provisioning to ramp services and aircraft towing, and as such, resolves the aircraft and passenger problem dimensions in airline disruption management. Over the one year period of airline operations, disruptions related to Ground Operations accounted for the largest percentage of total flight schedule delays of 39%, and the third highest percentage of total flight schedule cancellations of 13%.

**6. Inflight:**

Similar to Flight Operations, Inflight resolves disruptions defined by Flight Attendant (cabin crew) scheduling activities as they relate to Flight attendant tardiness and normal aircraft preparedness, and thus addresses the crew problem dimension of airline disruption management. Between September 2016 and September 2017, disruption instances related to Flight Operations represented about 18.5% of delayed flight schedules, 5.1% of cancelled flight schedules, and about 1% of diverted flight schedules in Southwest Airlines operations.

**7. Maintenance:**

This functional domain resolves disruptions defined by aircraft maintenance and inspection activities, and as such, addresses the aircraft problem dimension of airline disruption management. Disruption instances related to Maintenance represented 7.8% of delayed flight schedules and about 0.3% of cancelled flight schedules during Southwest Airlines operations from September 2016 to September 2017.

**8. NAS:**

This adopted functional role in SWA-NOC manages disruptions defined by air traffic control activities related to gate hold for congestion at departure and arrival airport stations. As such, the NAS functional domain addresses uncontrollable disruptions representing all problem dimensions during airline disruption management. Disruption instances associated with NAS represented 5.3% of delayed flight schedules, 13.5% of cancelled flight schedules and 6.4% of diverted flight schedules during Southwest Airlines operations from September 2016 to September 2017.

**9. Security:**

This functional domain addresses disruptions defined by security measures enforced to ensure the safety and convenience of passengers at airports prior to aircraft boarding. Its responsibilities includes managing disruptions due to baggage screening by TSA (Transportation Security Administration) at the skycap or ticket counter. As such, the Security functional domain resolves the passenger problem dimension during airline disruption management. Between September 2016 and September 2017, disruption instances related to Security represented the least percentages of total delayed, cancelled, and diverted flight schedules of 0.7%, 0.03%, and 0.16%, respectively, in Southwest Airlines operations.

**10. Technology:**

This functional role manages all disruption activities defined by system-wide technology outages, and thus aims to resolve all problem dimensions during airline disruption management. Disruption instances related to Technology accounted for 2.1% of all delayed flight schedules in Southwest Airlines operations between September 2016 and September 2017.

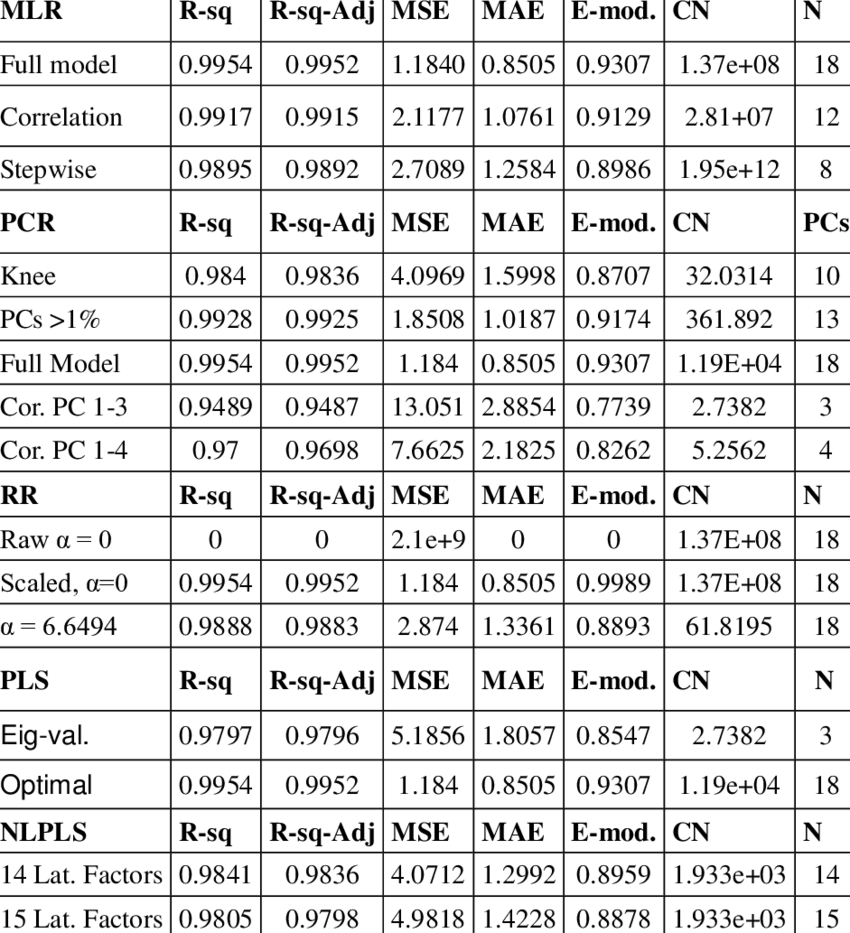
**11. Weather:**

Similar to NAS, this adopted functional domain in SWANOC manages all kinds of uncontrollable disruption defined by inclement weather activities. To that effect, the Weather functional role aims to resolve the aircraft, crew and passenger problem dimensions during disruption management. Disruption instances associated with the Weather functional domain accounted for the highest percentage of cancelled and diverted flight schedules (62.6% and 72.8% respectively) among all functional domains in SWA-NOC between September 2016 and September 2017. In addition, delayed flight instances related to Weather represented 2.9% of the total delayed flight instances addressed by all functional domains in SWA-NOC over the one year data collation period.

**1.5. Data Analysis**

The previous section provided a macroscopic overview of disruption activities for different functional roles in SWA-NOC and revealed that about 42% of all flight schedules for Southwest Airlines route network operations from September 2016 to September 2017 were disrupted. As a result, the functional disruption resolution domains in SWA-NOC were most likely to address irregular operations due to delayed flight schedules, which represent approximately 94% of all disrupted flight schedules recorded from September 2016 to September 2017.

Furthermore, there are two separate chunks of data which are defined by the occurrence of disruption during flight schedule execution. The first chunk, which is known as the non-disrupted data set, represents the larger chunk that contains instances of flight schedules that executed without any disruption.



**2. Proposed Approach:-**

The development of prediction models and decid implies a ﬁeld of tension between prediction accuracy and interpretability. The prediction accuracy describes the relation between the model and the real data. High

prediction accuracy means that there is a strong correlation between the pre-diction and the real value. In our context the usage of empirical distributions for delay predictions would have high prediction accuracy for short-term forecasts. However, this can lead to erroneous interpretations of the underlying mechanisms and result in wrong decision making. An alternative approach is to focus on the delay genereating mechanisms. This leads to the aspect of model interpretability. The beneﬁt in this approach is the understanding of a substantial relationship between cause and effect. On

the other hand, the prediction accuracy might be lower as only the most important patterns are captured. Both targets of prediction accuracy interpretability are addressed in the analysis. An exploratory analysis is pre-pended to statistical modeling. The derivation of decision rules and the generation of predictive models are closely related to the field of data mining. Data Mining is often deﬁne das the extraction of unexpected patterns in large data sets (Hand et al. 2001;Hastie Et Al.2009). It uses statistical and algorithmic methods for descriptive and predictive problems. Large data sets with thousands or millions of variables and observations pose challenges to formal statistical reasoning. For example, performing a large number of signiﬁcance tests will reject by design a certain percentage of Null Hypotheses (e.g., Efron 2010). Moreover, with large sample sizes, standard errors of estimators tend to become so small, that even unimportant’ differences between measured and true value share reported as signiﬁcant. In predictive modeling, Big Data risks to favor complex models that ‘mimic’ the sample and its statistical ﬂuctuation, but do not necessarily extract its underlying mechanisms (Hand et al. 2001, Chapter. 4.6.2 While for the purpose of short-term prediction some of these issues are resolved (for example by assessment of the bias-variance tradeoff), the data mining methodology does currently not provide a sound basis for the automated extraction of interpretable patterns (Breiman 2001a;Cox 2006; Cox and Wermuth 1996). As mentioned above, our strategy to avoid these pitfalls is to rely on descriptive meth- ods, complemented by formal inferences, whenever possible.

## 3. Description of the Data

The following analysis is performed on a data set consisting of 2.5 million ﬂight delay records provided by a major European airline for the time from March 2003 to

February 2007. Only continental passage line ﬂights are considered. Due to night ﬂying restrictions there are only occasional ﬂights between 10 p.m. and 6 a.m. which are excluded. Eventually, 2.2 million ﬂight records are used for the analysis. Besides the number of passengers per ﬂight, the available attributes can be differentiated into time, location and delay reason. The time aspect is characterized by scheduled and actual departure and arrival times in Central European Time (CET) stamps. For the determination of the local departure and arrival times we integrate information on time zones and daylight-saving time per airport, provided by openﬂight.org.

The location is represented by the departure and arrival airport. The route attribute can be derived from origin-and- destination pairs. The network is based on a hub-and- spoke structure with two major hubs, where 38.7 % of all ﬂights depart. 24.45 % of all ﬂights are spoke-to-spoke connections. In addition to airports and routes, we take the net-work structure into consideration by distinguishing between the following directions: hub-to-spoke, spoke-to- hub and spoke-to-spoke. A delay is deﬁned as the nonnegative deviation between the scheduled and actual departure time. The departure time is deﬁned as the time the aircraft leaves the gate. For every ﬂight, up to four different departure delay reasons and their durations are recorded, based on standardized IATA Delay Codes. They deﬁne primary delays as exogenous effects with codes from 1 to 89, containing airline internal reasons, disruptions of the turnaround process, technical damages, or airport and airspace congestion, just to mention the main categories. The group of reactionary delays include the codes from 90 to 96. These include waiting for passenger or load connections, for the ate arrival of a resource such as aircraft or crew, and for decisions from operations control. Of course, the transition between endogenous and exogenous effects is ﬂuent. The usage of the standardized IATA Delay Codes ensures the general adaptability of the approach. frequencies per number of departure delay records. Note that in this study we concentrate on positive delay values, early departures are declared to be on-time and thus set to a delay of 0 min. This is because negative delays do not propagate. In case of multiple records, secondary delay is mostly recorded ﬁrst. Further- more, only the delay is recorded that lead to late departures. Delay reasons that overlap in time are not entirely

* recorded, see Fig. 2 for illustration. Both of these effects
* lead to an underestimation of delay. In detail, 47.6 % of
* ﬂights are primarily and 19.75 % secondarily delayed.
* 8.63 % of the secondarily delayed ﬂights also contain
* primary delays.

**Conclusion :-**

From the above experimental results, we can see that interesting sets of trends and patterns exists in large data sets which help us to get a better understanding of the data. Recent advancement in cloud technologies helps us to harness the power of parallel processing of a cluster of computers with little investment and almost no maintenance of the underlying computer hardware. From the experimental results we also see the following observations: a. Average flight departure delay is at peak during the months of June and July every year and there is a sharp increase in the average delay from November to December. b. Average flight departure delay is increasing continuously over the period 2012 to 2014 in spite of the fact that the total number of flights have decreased from 2013 to 2014. c. Highest average departure delay for flights has been observed for flight distance of less than 500 miles. d. Highest numbers of flights which are cancelled have a flight distance of less than 1000 miles. e. There is a trend of increasing flights which are cancelled every year from 2012 to 2014. f. There is a sharp rise in the number of flights which are cancelled from the month of November to January every year for 2012 to 2014.

**Future Scope** :-

Mobility and its pillars of transport (air, inland and maritime) are at the very center of our socio-economic fabric. They underpin social connections and facilitate access to goods and services, including trade, jobs, health care and education. In today’s world, mobility by air, road and water is all about efficiencies, speed, interconnectivity and accessibility by all. However, this raises the issue about sustainability. The UN predicts that by 2050 two thirds of the world population will live in cities1. How can we adapt and enhance today’s already-stretched mobility system for it to respond to our expectations and increased demands? How can mobility be reinvigorated for it to be sustainable and support the 2030 Agenda of Sustainable Development and its 17 Sustainable Development Goals (SDGs)?

For a start, mobility actors should come together in a shared vision. This is where the World Bank-led Sustainable Mobility for All (SuM4All) steps in. For the first time ever, the SuM4All provides the transport sector and its modes of transport with the opportunity to speak with one voice and jointly unpack a Roadmap of Actions that is tailored to countries and cities to implement on a voluntary basis. The SuM4All includes all modes of transport, including aviation. Aviation facilitates access to countries and cities, increases multi layered efficiencies in travel and makes safety and security in travel top priorities. The aviation sector is rapidly taking gender equality at heart.

**REFERENCES:-**

1. Airline Data Set, United States Department of Transportation, Office of the Assistant Secretary for Research and Technology, Bureau of Transportation Statistics,<http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>
2. What is Hive? <http://www-01.ibm.com/software/data/infosphere/hadoop/hive/>.
3. Introduction to Hadoop in HDInsight: Big-data analysis and processing in the cloud, https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
4. An Introduction to Windows Azure BLOB Storage, https:/[/www.si](http://www.simpletalk.com/cloud/cloud-data/an-introduction-to-windows-azure-blob-)m[pletalk.com/cloud/cloud-data/an-introduction-to-windows-azure-blob-](http://www.simpletalk.com/cloud/cloud-data/an-introduction-to-windows-azure-blob-) storage-/
5. Explorer for Microsoft Azure Storage: Freeware Client,<http://www.cloudberrylab.com/free-microsoft-azure-explorer.aspx>
6. Upload data for Hadoop jobs in HDInsight,<https://azure.microsoft.com/enus/documentation/articles/hdinsight-upload-data/>
7. “Market Basket Analysis Algorithms with MapReduce”, Jongwook Woo, DMKD- 00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-4795