**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT Post, Bangalore-54

A Dissertation Report on

Airfare Aggression Analysis (2K16)

Submitted by

**Abhishek Raja**  1MS13CS008

**Adithya**  1MS13CS010

**Akshay P Sarashetti** 1MS13CS019

**Raghavendra** 1MS14CS417

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*

Aug – Dec 2016

****

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S. RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu)

# CERTIFICATE

I hereby certify that the work which is being presented in the B.E. report for partial fulfillment of the requirements for the award of the **Bachelor of Engineering in Computer Science & Engineering** and submitted to the **Department of Computer Science & Engineering** of M S Ramaiah Institute of Technology is an authentic record of our own work carried out during the period **August 2016 to December 2016** (**7th semester**) under the supervision of Prof Dr. T. N. R. Kumar

**Signature of Student(s)**

**Abhishek Raja Adithya**

**Akshay P Sarashetti Raghavendra**

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

**Signature of Supervisor**

Date: 14-12-2016

**ACKNOWLEDGEMENTS**

I express my sincere gratitude to Prof Dr. T. N. R. Kumar, Dept. of Computer Science, MSRIT, for his stimulating guidance, continuous encouragement and supervision throughout the course of present work.

**Signature of student(s)**

**Abhishek Raja Adithya**

**Akshay P Sarashetti Raghavendra**

**ABSTRACT**

The World Wide Web has made access to the Internet part of the structure of everyday life. Millions of people all over the world search the Web every day. But the commercial technology of searching large collections has remained largely unchanged since the 1960s.

Online Ticket fare estimation is one major advantage of the WWW. Earlier only tickets were booked online but this new estimation system gives the fare of the ticket based on the Booking Date, the Date of Journey, and other factors. Hence, this system is advantageous to choose the best offer available for the same booking from among the various booking rates of different companies. This new system not only gives offer for specific airlines but also gives deals offered by various airlines along with their details. The user thus has the ultimate power to choose the airline required based on his/her own requirements such as food, class of ticket, etc.

**CONTENTS**

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

***List of Figures***

**1** **INTRODUCTION**

1.1 Problem Statement

1.2 Objectives

1.3 Project Deliverables

1.4 Current and Future Scope

**2 LITERATURE SURVEY**

**3 PROJECT ORGANIZATION**

**4 SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.2.5 Storage Interfaces

4.3 Functional Requirements

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.4.6 Performance

4.5 Performance Requirements

4.6 Database Requirement

**5 DESIGN**

5.1 Introduction

5.2 Architecture Design

5.3 Graphical User Interface

5.4 Data Flow Diagram

5.5 Sequence Diagram

5.6 Class diagram

**6 IMPLEMENTATION**

6.1 Tools Introduction

6.2 Technology Introduction

6.3 Overall view of the project in terms of implementation

6.4 Explanation of Algorithm and how it is been implemented

6.5 Information about the implementation of Modules

**7 TESTING**

**8 CONCLUSION & SCOPE FOR FUTURE WORK**

**9 REFERENCES**

**1. INTRODUCTION**

Online booking is a convenient way of booking your travel over the internet. Using the regular online services you can:

* Book a flight and pay online for your travel.
* Request for a particular seat, special meal or service.
* Redeem your frequent flyer miles online.
* Check Flight status.
* View the Schedule/timetable.
* Web Check in.
* Make an online booking using the NGPAY mobile app.

The already available online booking system does not provide the best offer for the same booking by comparing many other prices of various companies. Hence this new addition can give a whole easy and economical planning of the travel.

* 1. **PROBLEM STATEMENT**

Here we try to analyse the airline fare aggression based on various derivatives like:

* How early tickets are booked.
* The time of flight.
* Which day of the week the flight is booked.

Basically we try to analyse the fares considering:

* Various airline companies.
* Different source-destination, and
* Different time,

So that we could advice the customer before-hand to avoid booking ticket from a particular company by ranking the companies based on how frequently and aggressively they change their fares compared to others and give the customer a better advice if he has options.

**1.2 OBJECTIVES**

The main objective of this project is to provide an easy and convenient mode to book an air ticket such that the user is happy by paying the least fare possible for the desired starting point and destination. Apart from suggesting the airline based on price alone, it can give suggestions based on requirements such as food, travel comfort, availability, upgradability, etc. Another objective of the project is to develop it into a full-fledged android application so that it becomes easy to use and available for all. The main advantage of the project is that it not only compares the air fares of on particular airline but it also compares the fare of all available airline from the given source point to the destination. The project also includes different payment modes as well. Security protocols and standardizations are used to keep the data private and secured.

**1.3 PROJECT DELIVERABLES**

The main deliverable of this project is a working prototype of the Online Ticket Fare Analyser which successfully identifies lowest available ticket fare for the travel from a given source to a desired destination. The project is very helpful to determine the best prices based on the required facilities as well.

**1.4 CURRENT AND FUTURE SCOPE**

In this project we have successfully built the AIRFARE AGGRESSION ANALYSIS where the best fare for the given travel specification is obtained from the available data of the airlines. The future work aimed for this project is to enable the AIRFARE AGGRESSION ANALYSIS to advice the customers to choose the best airline that provides least charges for the fly and also provide best catering to the customers flying. Apart from this, we intend to make it a full-fledged mobile application so that it reaches the targeted audience.

As new users are entering the network and new services are evolving, the behavior of both user and service providers are going to dynamically change. So a system which adapts to the new types of attacks and responds must be formulated. This makes the project more robust and fault tolerant.

**2. LITERATURE SURVEY**

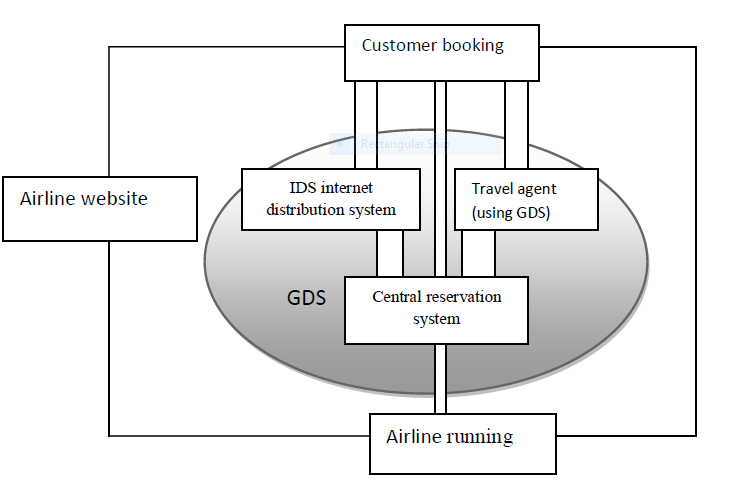
In the arena of global competition, organizations all over the world are competing through the use of the most comprehensive and advanced technological features. The most common example of innovation is in the area of information technology and communication. Various industries are using technology and the advancements of software and internet to maintain and monitor their business transactions. In the application of the informative systems, the airline industry is the most commonly used system.

**2.1 History of Airline Reservation Information System**

American Airlines was the first to establish an automated booking system in 1946. Using a system to track information and improve efficiency was a highly appealing aim in the industry, and drew the attention of other airlines worldwide. The system endured years of development and alterations. Later, other airlines invested more in research and development to launch improved systems, and through the late 1960s and early 1970s, airlines established their own systems.

United Airlines developed the Apollo Reservation System, and shortly after allowed travel agents access. The Apollo system was the foundation for many further developments, which spread from just US airlines to European airlines as well. The research and development of Airline Reservation System became a significant aspect of the industry and all its air carrier companies, and partnerships between airlines and technology gurus emerged.

Airline deregulation occurred in 1978, magnifying the importance of computerized airline reservation systems and their accessibility. During the early 1970s, as travel agents pushed for access to reservation systems, and certain airline executives made investments for the sake of accessing the systems of other airlines, antitrust laws came into focus. The purpose of the 1978 Airline Deregulation Act in the United States was to eliminate government control over commercial aviation, and ensure competitive behavior and fair business practices in the airline industry. Passengers could gain knowledge of market forces and new market entry in the industry. Information on specific airlines and the industry as a whole became more widely and readily accessible, evolving the airline reservation systems from "standalone" operations toward GDS. Today, airline reservation systems have developed into computerized reservation systems which are of mission critical to the airline industry, about six major airline reservations systems are used by international airlines.



These were the main reasons for an online air ticket reservation. But to enhance this facility even more, we built the AIRFARE AGGRESSION ANALYSIS which is more beneficial to the users as it provides better savings to the users.

**3. PROJECT ORGANIZATION**

**3.1 SOFTWARE PROCESS MODEL**

Development of this software included splitting of development work into distinct phases or stages containing activities to ensure better planning and management. The methodology used includes the pre-definition of specific deliverables and artefacts that are created and completed by project team to develop or maintain this application.

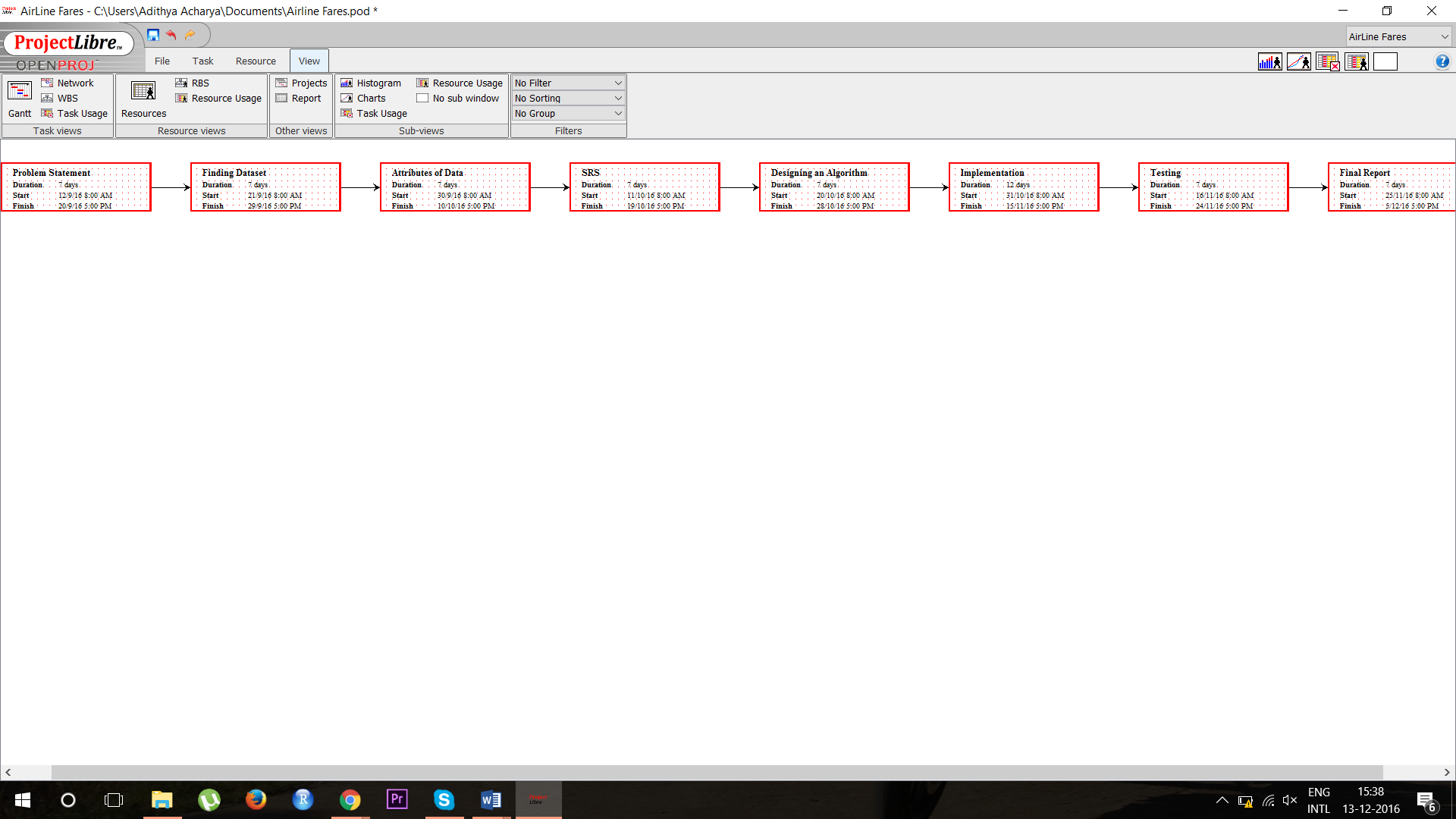
The steps included in software development life-cycle were as given below:

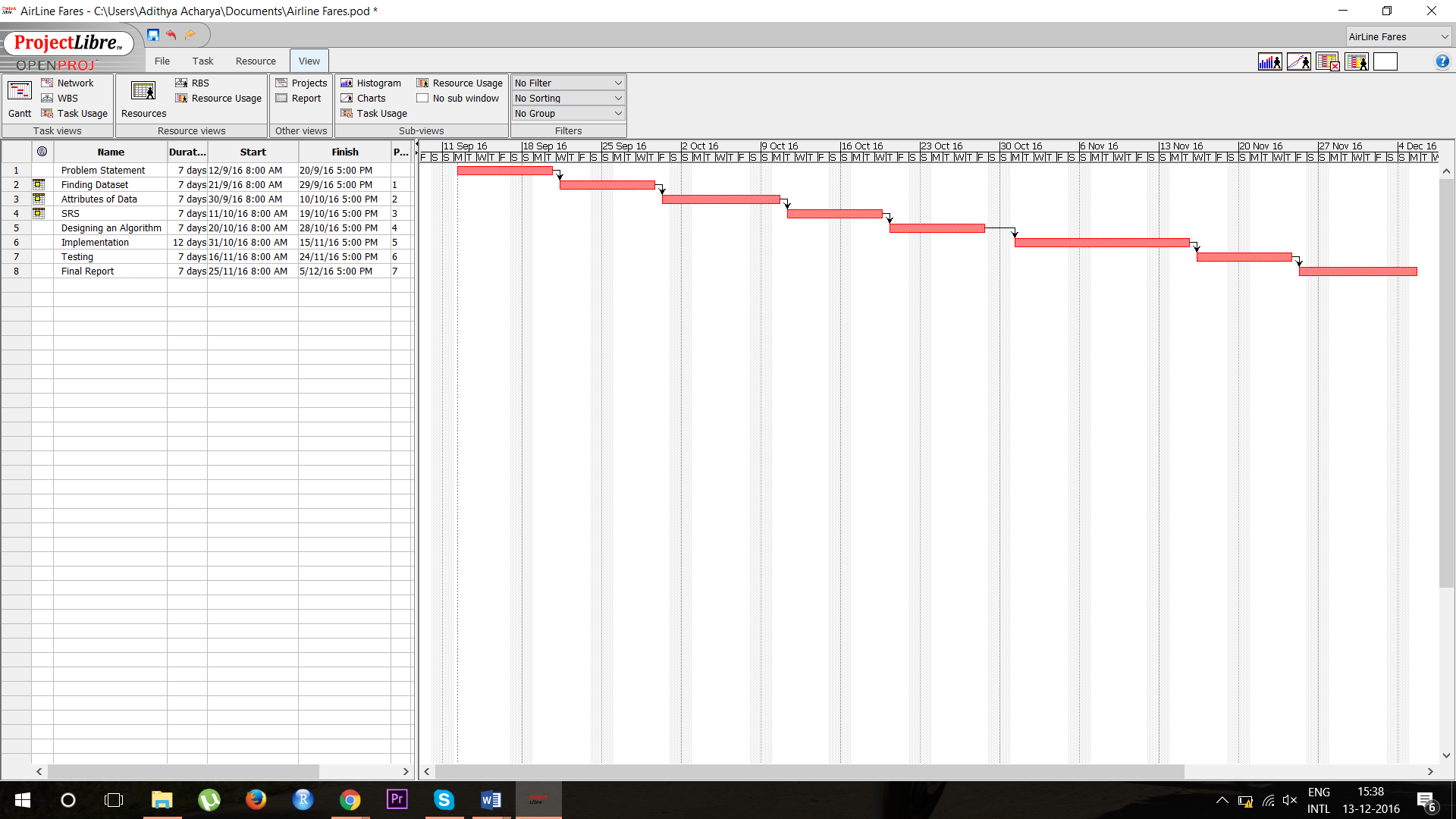
|  |  |
| --- | --- |
| i) System  Planning: | Suitable problem definition is chosen and is checked for system feasibility. If the system is feasible, the project is chosen. Else, problem definition is discarded and another problem can be focused. |
| ii) Requirement  Analysis and Specification | This step is a critical and crucial step which decides the success of the project. Requirements are analysed and specifications are described in this step. Requirements may be hardware, software, functional, non-functional, structural or behavioural. Use cases are designed to make sure that all the functions are specified without any ambiguity. |
| iii) Architecture  and  Detailed Design: | High level design is first developed with the help of requirements. A detailed design is then built using architecture design which includes functionality of each requirement module. |
| iv) Implementation  and Execution: | Suitable code is written in a suitable language or platform which is specific, reliable and portable for each module. After all modules are programmed atomically, they are integrated and executed as a whole. |
| v) Testing: | Bugs or errors in function modules may be present in the system. Testing is done to remove such bugs and errors from the system. Testing may be unit testing, system testing, integration testing. System is also tested for its reliability, portability, functionality, security and performance. |
| vi) Deployment  and maintenance: | After proper testing, project is deployed and released. There may arise some bugs which were unable to be identified during testing. Such bugs are taken care by proper maintenance and management of the project. |

The methodology followed in this project is an agile (Extreme) methodology. This method mainly involves evolving learning since it has to produce the lowest price based on the current booking time and the date of travel. And other facilities like online payment makes it a much more beneficial project for the users.

**3.2 ROLES AND RESPONSIBILITIES**

Each team member had an equally important role to play in this project. Each team member was assigned with specific tasks and responsibilities to perform by the team leader during the project life-cycle. The task assigned to each member at different times was equally important irrelevant to the complexity of it since each and every minute details was important to build the project. The tasks started from constructing the database and every step from there was a key step to building the project.





**4. SOFTWARE REQUIREMENT SPECIFICATIONS**

The introduction of the Software Requirements Specification (SRS) provides an overview of the entire SRS with purpose, scope, definitions, acronyms, abbreviations, references and overview of the SRS. The aim of this document is to gather and analyze the available database provided by the various airlines and suggest the flight that provides the least cost to travel from a desired source point to a desired destination.

**4.1 PRODUCT OVERVIEW**

The AIRFARE AGGRESSION ANALYSIS is basically an extension of the online ticket reservation system. The main advantage of this new system is that the user is benefitted with getting the lowest available price that is available among the list of airlines available from desired source point to the desired destination and also at the required date and time.

**4.2 EXTERNAL INTERFACE REQUIREMENTS**

* **Software Interface**: Here we are using RStudio as a software interface for our project and also Microsoft excel to read the training data. RStudio is the IDE that is used to run the r code implemented in the project. **RStudio** is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. **R** is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.
* **Hardware Interface**: The main interface that has been developed as of now is a website. But the final goal of the project is to make it a mobile application so that it can be used by each and everyone. Minimum Requirements: RStudio runs on a laptop with min 4GB RAM, 64 bit configuration

Requirements: 4GB RAM/40GB Hard Disk/ I3 processor. (Recommended)

* **Communication Interface**: As mentioned earlier the website platform is going to be the main communication interface between the airlines and the customer.
* **Storage Interface**: The key storage interface planned for the project is cloud since it can easily accessed by the airways to make bookings.

**4.3 FUNCTIONAL REQUIREMENTS**

The hypothesis must be able clearly distinguish the airlines based on their behaviour towards the increase or decrease in airline fares for a certain day. The project also to determine the best offers based on certain criteria required by users as well.

**4.4 SOFTWARE SYSTEM ATTRIBUTES**

**4.4.1 Reliability**

The software system meets all the necessary functional requirements without any failure and displays the appropriate data.

**4.4.2 Availability**

The program is available 24x7 provided we have proper internet connection to interact with cloud database.

**4.4.3 Security**

The code developed relies on the Internet, and uses cloud as database. Hence it may be vulnerable to attacks. Security mechanisms are standardized to provide privacy and to keep data secured by avoiding vulnerable threats.

**4.4.4 Portability**

This system is easily portable and can be run anywhere. It includes different devices such as computers, mobiles, tablets, etc.

**4.4.5 Maintainability**

Hardware used is user friendly and easy to use. The code is written clearly and well documented.

**4.4.6 Performance**

The performance of the project mainly depends on the internet speed available to the users and also other facilities for online payments.

**4.5 PERFORMANCE REQUIREMENTS**

In this project, we deal with a cloud database. Hence it is crucial for us to give importance to Internet connectivity. To get the best output from the system the user must also be flexible with the time of flight on the same date.

**4.6 DATABASE REQUIREMENT**

The device requires an extensive database. The database is mainly required in order to identify the availability of the tickets in any airlines and also maintain the cost of the tickets. This database has to be updated real-time to maintain proper availability of the tickets. We have used the cloud as our database storage medium as it is fault tolerant, durable and consistent. Cloud enables efficient querying to fetch the predetermined values, and perform the appropriate action.

**5. DESIGN**

**5.1 ARCHITECTURE DESIGN**

The architectural design is the design of the entire software system; it gives a high-level overview of the software system, it provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.

C:\Users\Adithya Acharya\Downloads\2.png

**Figure 1 – System Architecture.**

**5.2 GRAPHICAL USER INTERFACE:**

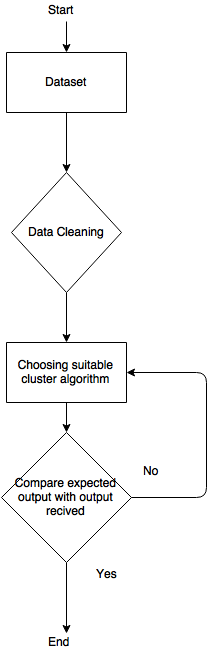
Plots are used to represent data in Graphical format. Plots of different kinds namely scatter plots, histograms are used.

**5.3 SEQUENCE DIAGRAM:**

C:\Users\Adithya Acharya\Downloads\3.png

**Figure 2 – Sequence Diagram.**

**5.4 FLOW DIAGRAM:**



**Figure 2 – Flow Diagram**

**6. IMPLEMENTATION**

**6.1 TOOLS INTRODUCTION**

**Rstudio:**

Rstudio is a free and open-source integrated development for R, a programming language for statistical computing and graphics.

RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop version and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server.

RStudio is written in the C++ programming language and uses the QT framework for its Graphical User Interface.

R studio consists of a console, a workspace, a files tab, a History Tab, plot tab. Console is where the code is written and executed. Workspace is where all the codes can be saved for future purposes. A files tab shows all the files and folders on your pc. Plots tab is where all the plots are displayed. History tab keeps a record of all previous commands, which is used in testing purposes. The Rscript is where you keep a record of your work.

New packages can be installed using install.packages command,that lets you choose from a variety of libraries to be installed.

Rstudio is known for its graphical representation of data. There are different plots available for different kinds of data. They are:

1. Strip Charts

2. Histograms

3. Boxplots

4. Scatter Plots

R also allows users to represent their data in 3D models.

**6.2 TECHNOLOGY INTRODUCTION**

**R(programming language)**

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and FORTRAN code can be linked and called at run time. Advanced users can write C, C++, Java or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages.

Like other similar languages such as APL and MATLAB, R supports matrix arithmetic. R's data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists. R's extensible object system includes objects for (among others): regression models, time-series and geo-spatial coordinates. The scalar data type was never a data structure of R. Instead, a scalar is represented as a vector with length one.

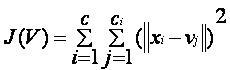
**6.3 OVERVIEW OF THE PROJECT IN TERMS OF IMPLEMENTATION**

Raw data was obtained from the company TravelSpends. Data was of 2 types GDS and LLC. The data provided to us had to be cleaned to remove redundant tuples. After cleaning the data, the dataset had about 60k data points. We calculated the difference in booking date and date of journey and was added to the data set. The data obtained was now normalized, to see through that it is not left skewed or right skewed. Data analysis was done, by applying Clustering technique. The Clustering algorithm k-means, is an algorithm that groups’ data into different groups based on clusters.

**6.4 ALGORITHMS**

**K-means Clustering Algorithm**

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:



Where,

‘||xi - vj||’ is the Euclidean distance between xi and vj.

‘Ci’ is the number of data points in ith cluster.

‘C’ is the number of cluster centers.

Algorithm:

Given an initial set of *k* means, the algorithm proceeds by alternating between two steps:

**Assignment step**: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.

**Update step**: Calculate the new means to be the centroids of the observations in the new clusters.

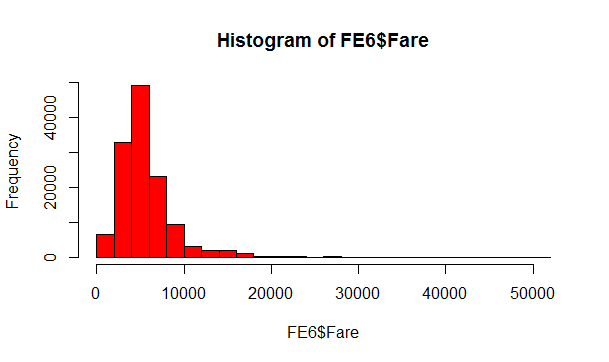
Regarding computational complexity, finding the optimal solution to the k-means clustering problem for observations in d dimensions is:

NP-hard in general Euclidean space d even for 2 clusters

NP-hard for a general number of clusters k even in the plane

If k and d (the dimension) are fixed, the problem can be exactly solved in time O(n^{dk+1}\log {n})} O(n^{dk+1}\log {n}), where n is the number of entities to be clustered

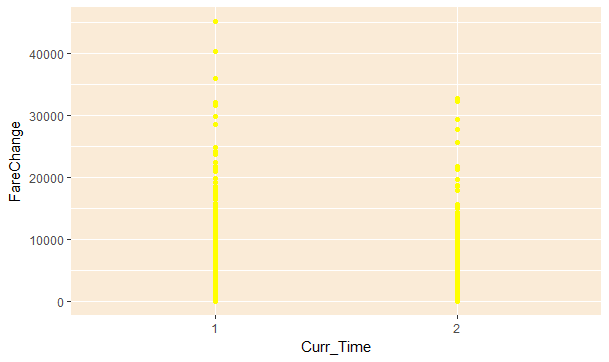
**6.5 INFORMATION ABOUT THE IMPLEMENTATION OF MODULES**



**Fig: Distribution of fare**

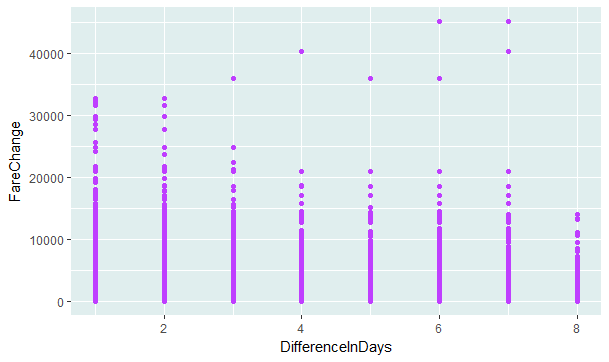
**Module 1: Fare with respect to current time**

Airfare for a particular company changes with respect to the flight journey timings. The data obtained has fares based on, whether the departure is morning or evening. To use these values mathematically, we use transform function to transform these categorical values to mathematical values.



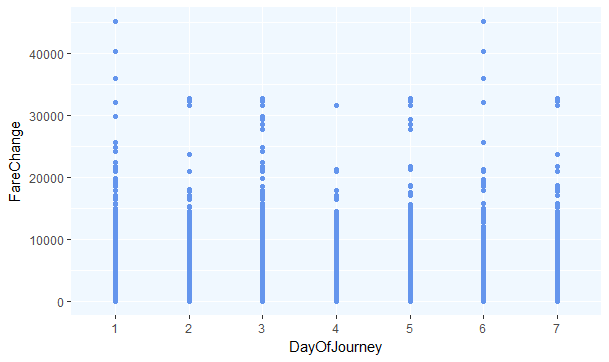
**Module 2: Fare with respect to difference in days**

Airfare is usually calculated based on how early one books a flight ticket. Thus difference in booking date and date of journey was calculated. As the difference is less the fare is higher and for a larger value of difference in days, the fare was usually lower.



**Module 3: Fare with respect to day of the week**

General mentality of public is that fares are usually high on Friday evenings and Sunday nights. But to the contrary, fares are usually high on Monday mornings. This leads to our analysis statement, to derive which company aggressively increases the price on a Monday morning.

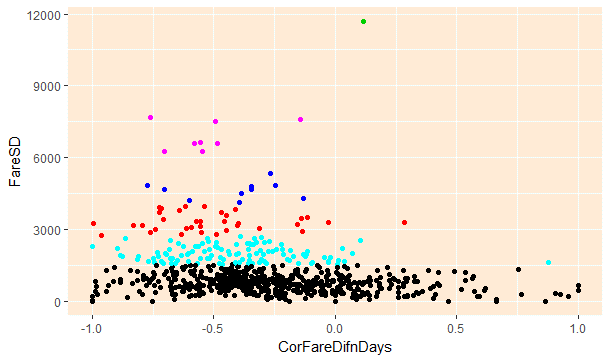


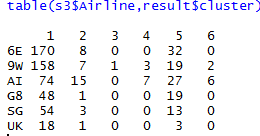
**7. TESTING, RESULTS AND SNAPSHOTS**

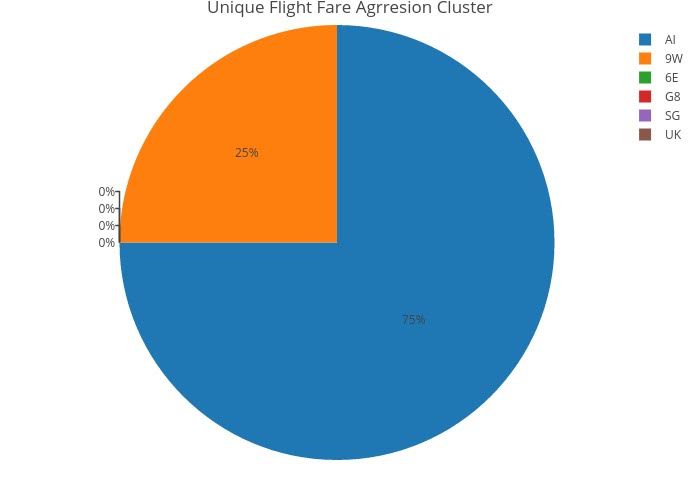
**Analysis 1:**

Unique flights were determined based on the day of the journey, difference in days and current time. Standard deviance and variation of fares are calculated. Correlation, covariance for all the unique flights with respect to their fare and difference in days were also calculated. The data generated were now stored in a new dataset where the dataset had 732 unique tuples.

Attributes correlation, standard deviation and covariance were taken into consideration to apply k-means clustering algorithm where the data was clustered into 6 different clusters, based on the airline company. The results obtained are shown below:



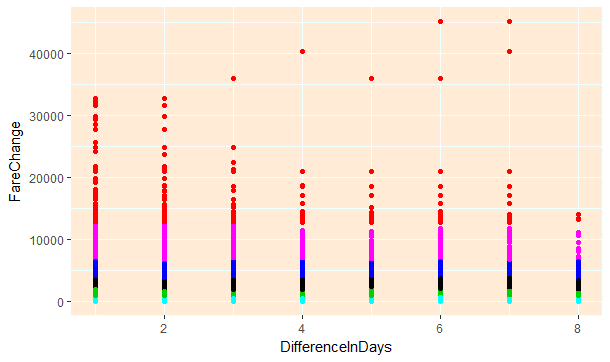


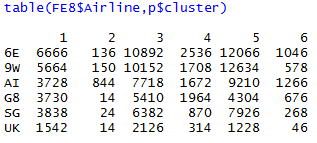


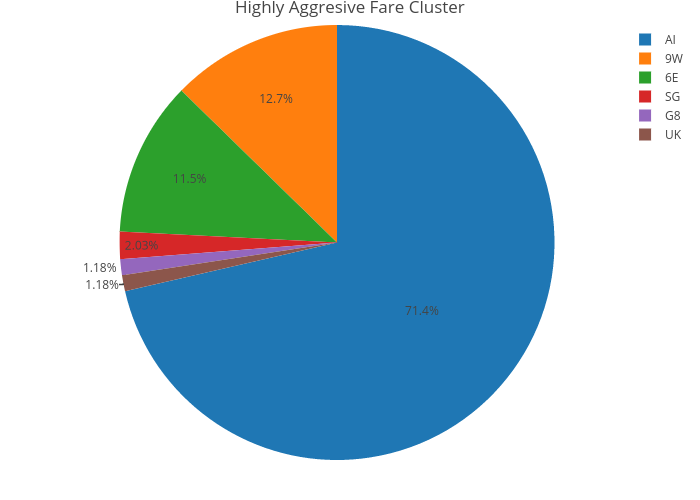
From the above plots it is clearly significant that airline company AI is the company that increases its prices more aggressively.

**Analysis 2:**

Fare minimum value is calculated for each unique flight of given source, destination and Airline Company. For every value in the population, fare change is calculated with respect to the minimum value. The data generated is now applied with k-means clustering algorithm based on the current time, day of journey, difference in days and fare change, with k-value 6. The result obtained was stored into a table and plot of cluster values against airline and fare was done to obtain the final plot. The results obtained are shown below:







Just as analysis 1, analysis 2 provided the same results with much more accuracy also stating that both the analysis are true and valid.

**8. CONCLUSION**

In this project we have successfully built the AIRFARE AGGRESSION ANALYSIS where the best fare for the given travel specification is obtained from the available data of the airlines. The future work aimed for this project is to enable the AIRFARE AGGRESSION ANALYSIS to advice the customers to choose the best airline that provides least charges for the fly and also provide best catering to the customers flying.

As new users are entering the network and new services are evolving, the behavior of both user and service providers are going to dynamically change. So a system which adapts to the new types of attacks and responds must be formulated. This makes the project more robust and fault tolerant.

Designing and implementing novel devices for the benefit of the society is definitely in progress. However, focusing on the technology aspect alone is not enough. Conforming to established laws and regulations, upholding privacy of users while providing security to their device, ensuring that the product is sustainable and has a positive social impact adds to its economic value and renders it more feasible. Although the usage of cloud as a storage medium has potential advantages such as easy maintainability, pay-per-use, lesser energy consumption etc., the vulnerabilities to surface attacks, supplier stability and accessibility also need to be considered. It is important to find a balance between the feasibility of the device, its economic value and most importantly customer satisfaction.

**9. REFERENCES**

1. <https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&uact=8&ved=0ahUKEwi4y7v_4_DQAhVLLY8KHc38CPgQFggwMAQ&url=http%3A%2F%2Fwww.iaeng.org%2Fpublication%2FWCE2010%2FWCE2010_pp465-468.pdf&usg=AFQjCNG_AYqylJbpQe0XfnWsfeAzxy8WDg&sig2=AEahB8NrvmUYmKXorHp94g>
2. <https://en.wikipedia.org/wiki/Internet_booking_engine>
3. <http://www.travelpd.com/travel-portal-agile-methodologies>
4. <http://www.airfrance.fr/FR/en/common/transverse/footer/edito_psc.htm>
5. <https://www.philippineairlines.com/AboutUs/LegalNotices/OnlineBooking>
6. <https://kr.flyasiana.com/C/en/homepage.do?menuId=003002002000000&menuType=CMS>
7. <http://www.criminaldefenselawyer.com/crime-penalties/federal/computer-crimes.htm>l
8. <https://www.researchgate.net/profile/Bruce_Schatz/publication/2955040_Digital_libraries_Technological_advances_and_social_impact/links/543be39a0cf2d6698be344ef.pdf>
9. <http://www.otrams.com/blog/trends/the-challenges-and-opportunities-in-online-travel-and-hotel-booking>
10. Aksoy, S., Atilgan, E., &amp; Akinci, S. (2003). Airline services marketing by domestic and foreign firms: differences from the customers’ viewpoint. Journal.
11. “Constrained K-means Clustering”, Kiri Wagstaff and Claire Cardie Department of Computer Science, Cornell University, Ithaca, NY
12. <https://en.wikipedia.org/wiki/RStudio>
13. The Changing Low-Cost Airline Model: An Analysis of Spirit Airlines David E.Rosenstein.
14. O’Connell, J.F., &amp; Williams, G. (2005). Passengers’ perceptions of low cost airlines and full service carriers: A case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines. Journal of Air Transport Management, 11, 259-272.
15. <http://www.statmethods.net/advstats/cluster.html>
16. <https://www.r-bloggers.com/k-means-clustering-in-r/>
17. <https://en.wikipedia.org/wiki/K-means_clustering>
18. <http://www.rdatamining.com/examples/kmeans-clustering>
19. “Exploring Airline Fare Pricing” Keith A. Willoughby Department of Finance and Management Science Edwards School of Business, University of Saskatchewan.