ITCS 6112

SOFTWARE SYSTEM DESIGN AND IMPLEMENTATION

A PROJECT REPORT

ON

"FLIGHT DATA ANALYSIS"

Using
PROGRAMMING LANGUAGE: ***R***

TEAM MEMBERS:

ATHREYA SURYA PRAKASH	800965127
ADITHYA KUMAR	800965120
VIKAS DESHPANDE	800964852
VENKATESH BONAGERI	800964302
NITHIN VENUGOPAL	800966213

UNDER THE GUIDANCE OF

Dr. Ali Sever

`

DEPARTMENT OF COMPUTER SCIENCE

COLLEGE OF COMPUTING AND INFORMATICS

CHARLOTTE, NORTH CAROLINA - 28262

Fall 2016

TABLE OF CONTENTS

S.NO	TOPIC
1	Abstract
2	Introduction
3	Data Description
4	Why R
5	Software & Hardware Requirements
6	Analysis Description
7	Modules
8	UML Diagrams
9	Output
10	Result Analysis
11	Test Cases
12	Conclusion
13	Future Enhancements
14	References

ABSTRACT

Flight data is huge and voluminous. Processing, analyzing and extracting useful information from such a data and projecting the results along different metrics like carrier delays, busy routes and busiest airports is the main theme of the project. The long-term objective is to assist passengers in taking better decisions with respect to selection of flights in their journey and airport authority in managing the airport in a better way by using feedbacks received from travelers and businessmen. The analysis would help the airlines in taking decisions that would better their service and businessmen in setting up their business in the most profitable way.

The system processes the flight details provided as input from an excel spreadsheet spread across different years in CSV format and displays the required metrics graphically.

- **Carrier delays:** Flight carrier delays is computed at different time granularities (Day, Week, Month, Year) and between two places
- **Busy Routes:** Busy routes are computed at different time granularities (Day, Week, Month, Year)
- **Busiest airports:** The busy airports are computed based on the number of the flights departing from the airport at different time granularities (Day, Week, Month, Year)

INTRODUCTION

The flight data analysis is one of the most important data analysis in the commercial aviation transport industry. Information such as busiest routes, busiest airports, and the delay of flights for different carriers is analyzed and presented. The stakeholders like air travelers, airports management officials, businessmen owning different business outlets at these airports and the airline companies use the results of this extensive analysis to ensure that they provide or receive services from the aviation industry in the best way possible.

Since this project deals with huge amount of data that needs to be parsed, analyzed and queried to get useful information, R programming language is used. For example, if a user is looking out for the best flight to book for a travel from Chicago to New Jersey on the Christmas day, he would want to know several details like the busiest airport, the busiest route and the

carrier delay information prior to booking his tickets. This kind of analysis is exactly what the system provides. It helps in taking the best decision with respect to travel and management. Depending on user's input, the flight data can be analyzed on a day-wise, month-wise or a year-wise basis. The user interacts with the system using the interface provided by Shiny web platform.

DATA DESCRIPTION

1. How did we obtain the data set? We got the data set from United States Department of Transportation website. It contains information about all flight departures, arrivals, flight delay details at the United States airports.
2. How the data was originally collected? The data is from the Research and Innovation Technology Administration at the Bureau of Transportation statistics and is collected by the Office of Airline Information.

WHY R?

The system has been implemented using R programming language because the project extensively involves analysis of historical data with prediction and graphical representation of the output results. For such a statistical data analysis requirement, R is one of the most preferred languages.

Another reason for using R is its compatibility with Shiny web platform. Shiny comes as an R package that makes it easy to build interactive web applications. Using R, you create a user interface and server code and Shiny compiles your code into html, CSS and JavaScript needed to display your application on the web. What makes a Shiny app particularly powerful is that it can execute R code on the backend, so the app can perform any R calculation you can run on your desktop. Perhaps you want your app to slice and dice a dataset based on user inputs. Or maybe you want your web app to run linear models, or machine learning methods on user-selected data. In either case, Shiny can help.

SOFTWARE & HARDWARE REQUIREMENTS

Software Requirements

- Windows 8/10
- R compiler version 3.3.2
- R Studio version 1.0.44 (or latest version)

- R Packages- shiny, ggplot2, dplyr

Hardware Requirements:

- RAM: 16 GB
- Processor: i5 and above
- Minimum 512 GB Hard Disk

ANALYSIS DESCRIPTION:

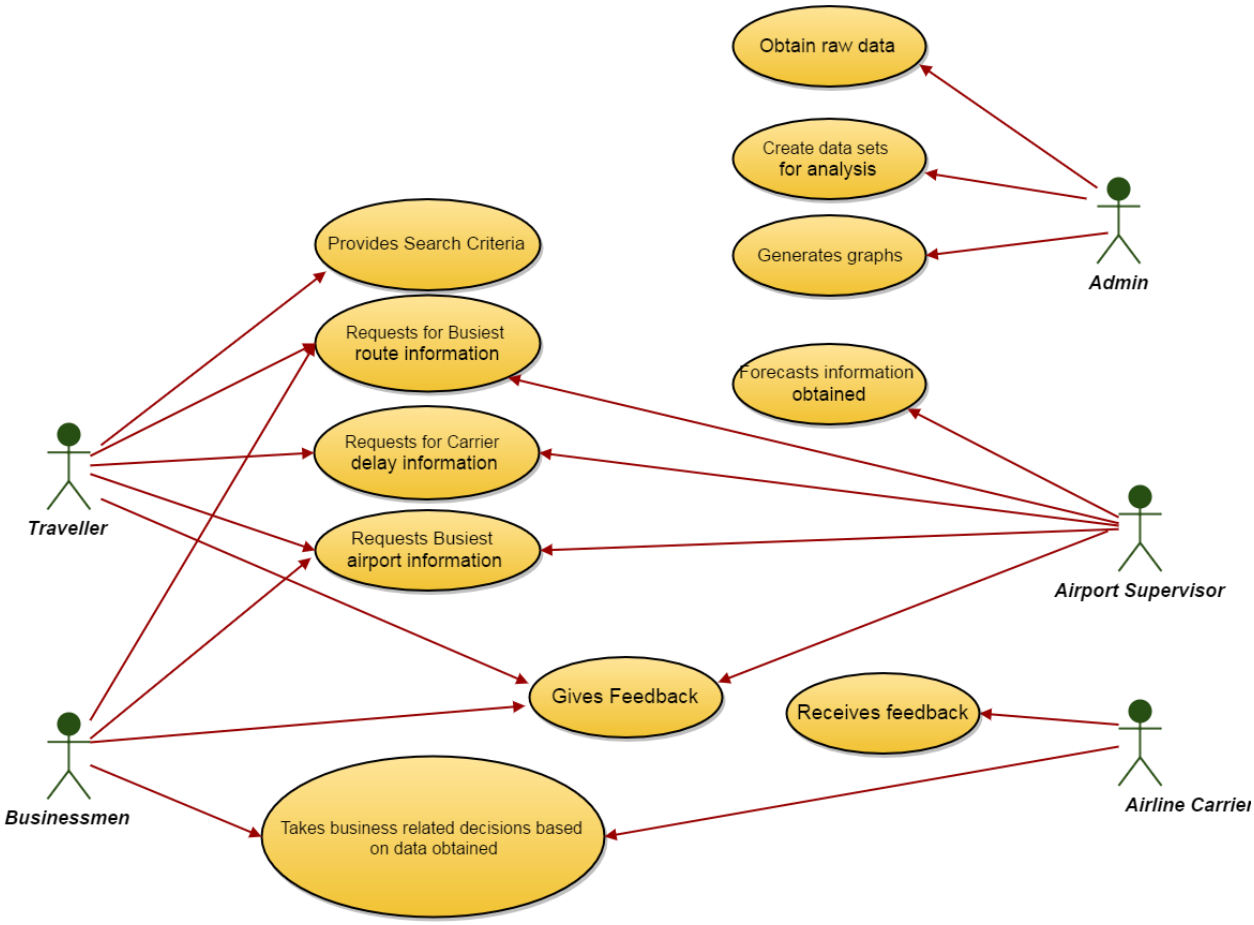
1. The process starts with acquiring the structured data sets required for flight data analysis from authentic sources. The data is taken from this link. To sample the analysis, 5 years' data from 2011 - 2015 is considered.
2. 13+ million rows of data is considered.
3. Once the system is run, the user will be prompted to select the metric (carrier delay, busiest routes, or busiest airport) he would want to analyses and to specify the time period (day-wise, month-wise or year-wise).
4. Based on the user's choice in previous step, the system will perform an extensive and meticulous analysis using the dataset considered.
5. The graphical representation of the analysis will be displayed to the user using the Shiny web interface.

MODULES:

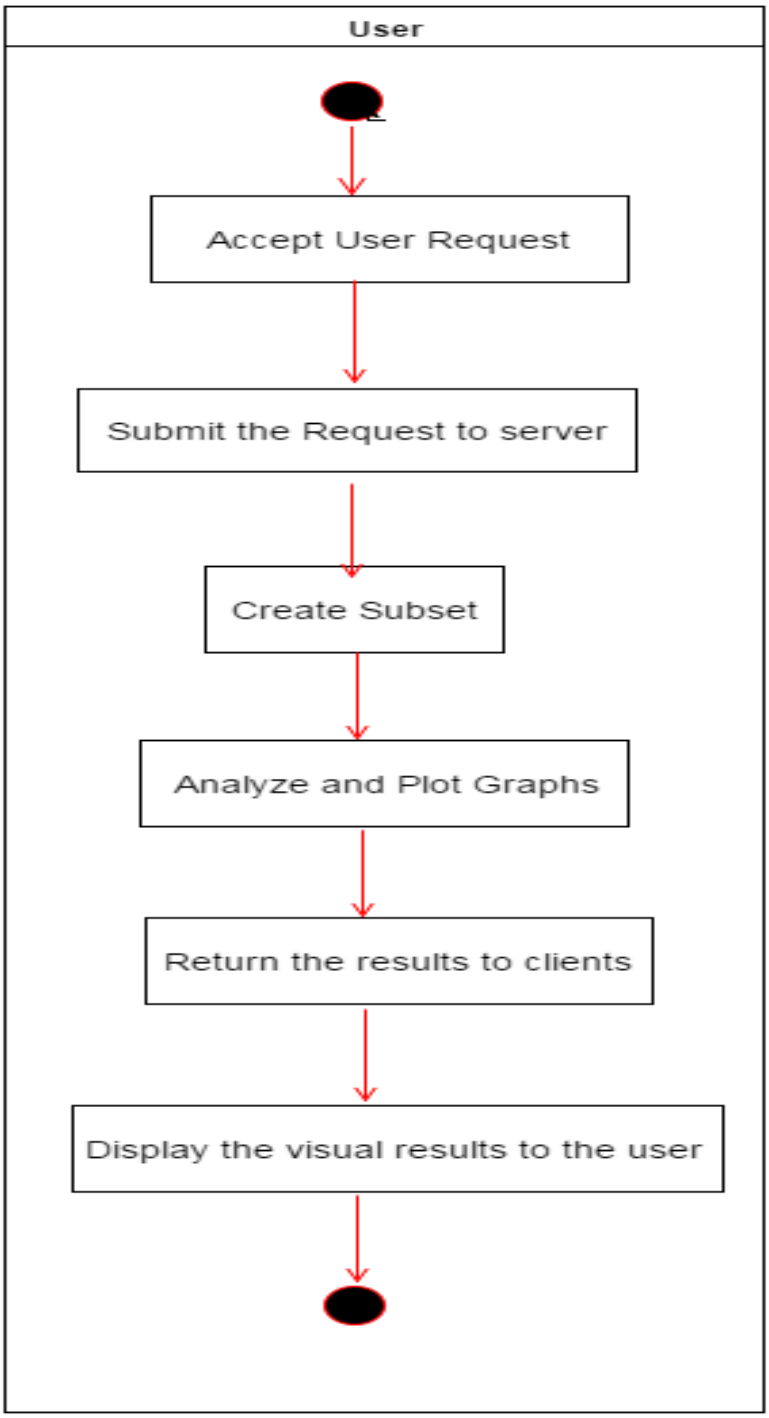
1. Input module
2. Busiest route module
3. Carrier delay module
4. Busiest airport module
5. Display module

UML DIAGRAMS

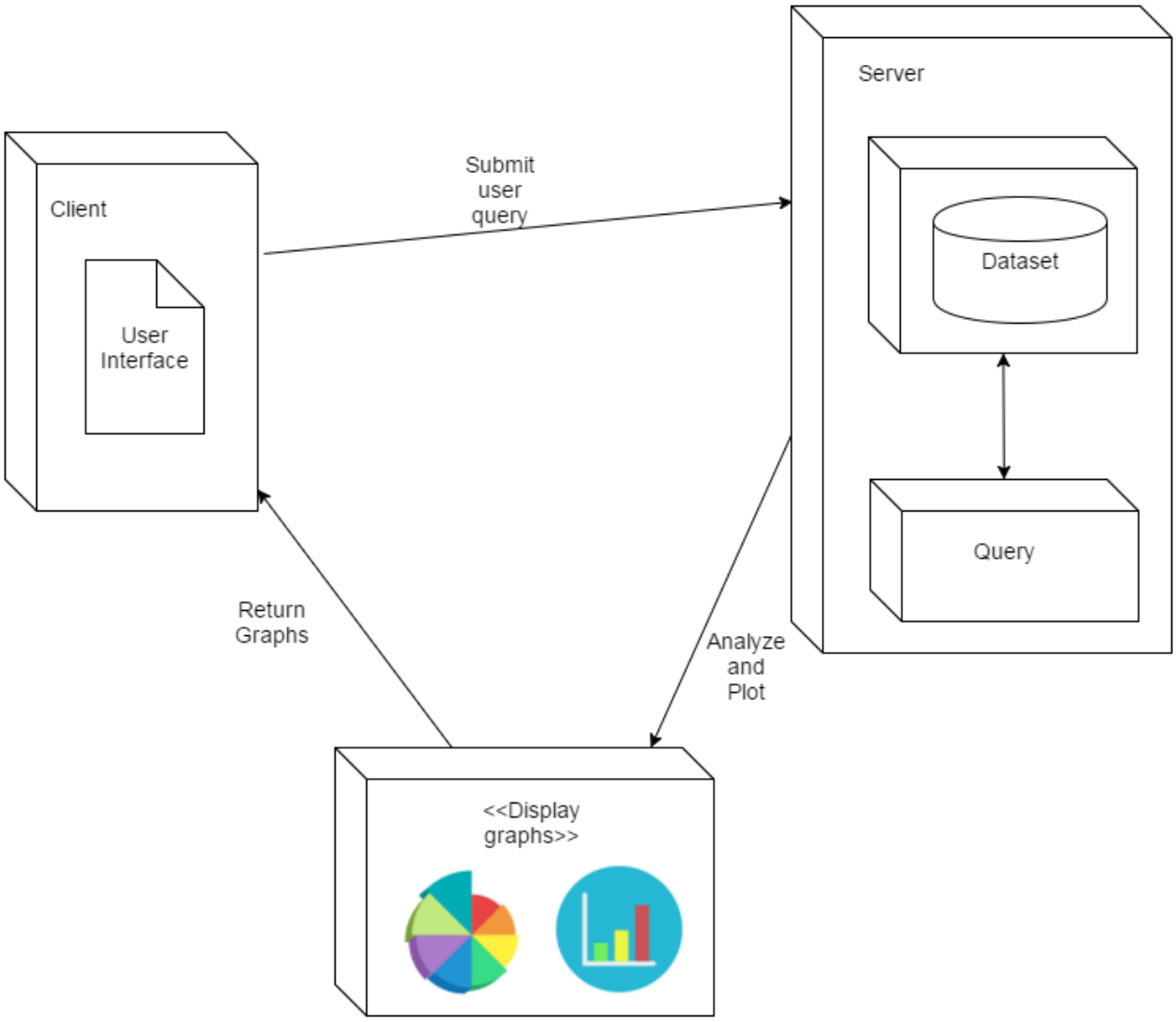
1. Use Case Diagram



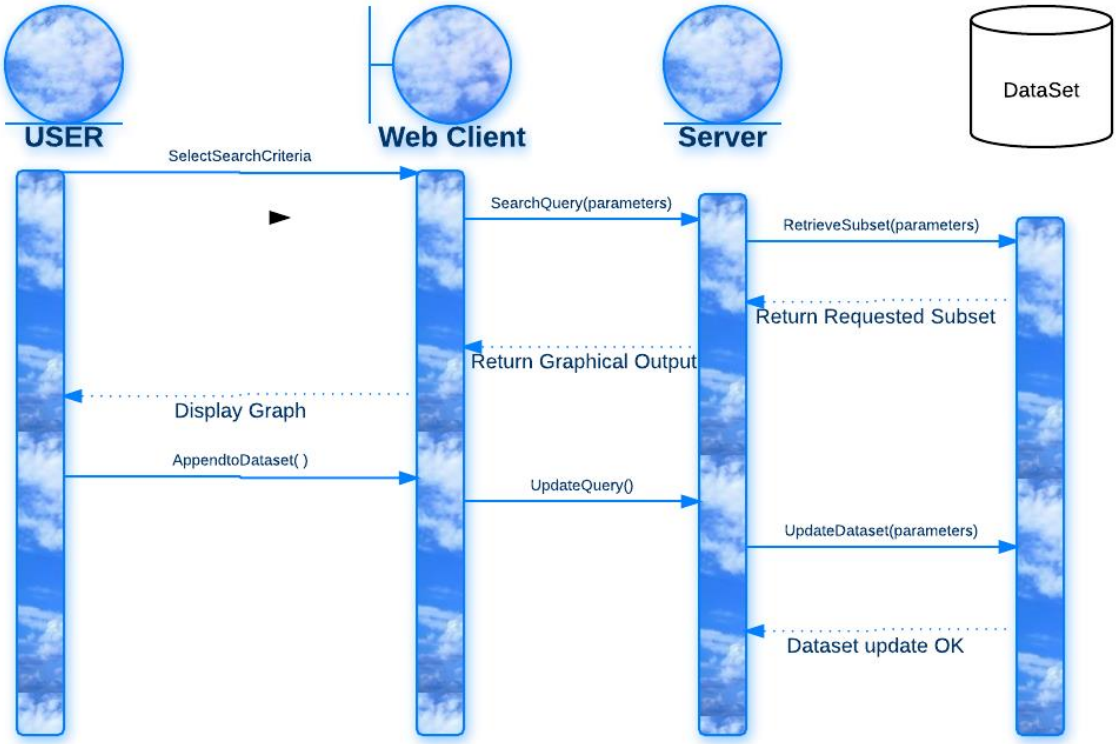
2. Activity Diagram



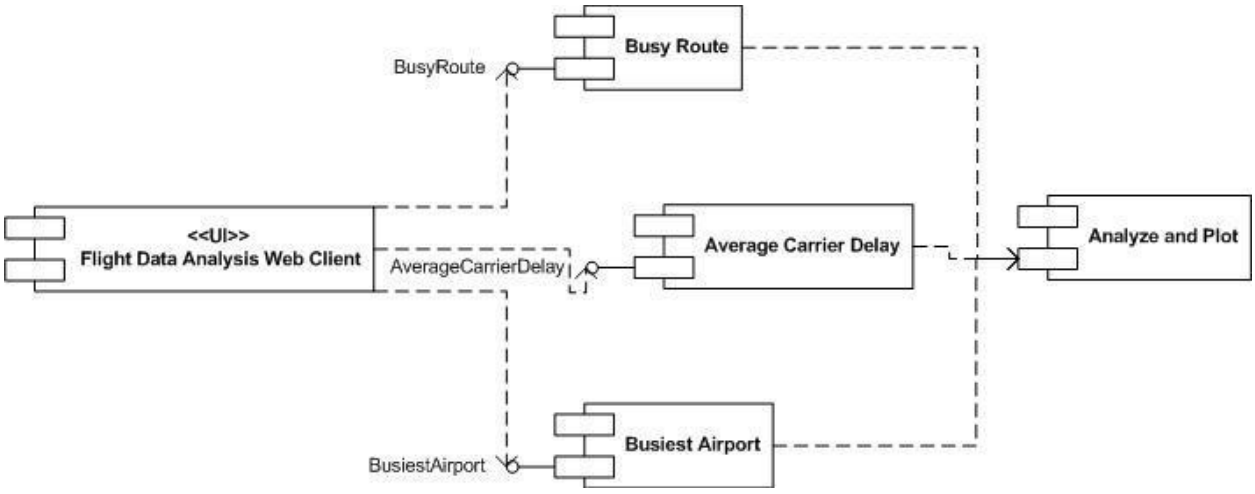
3. Deployment Diagram



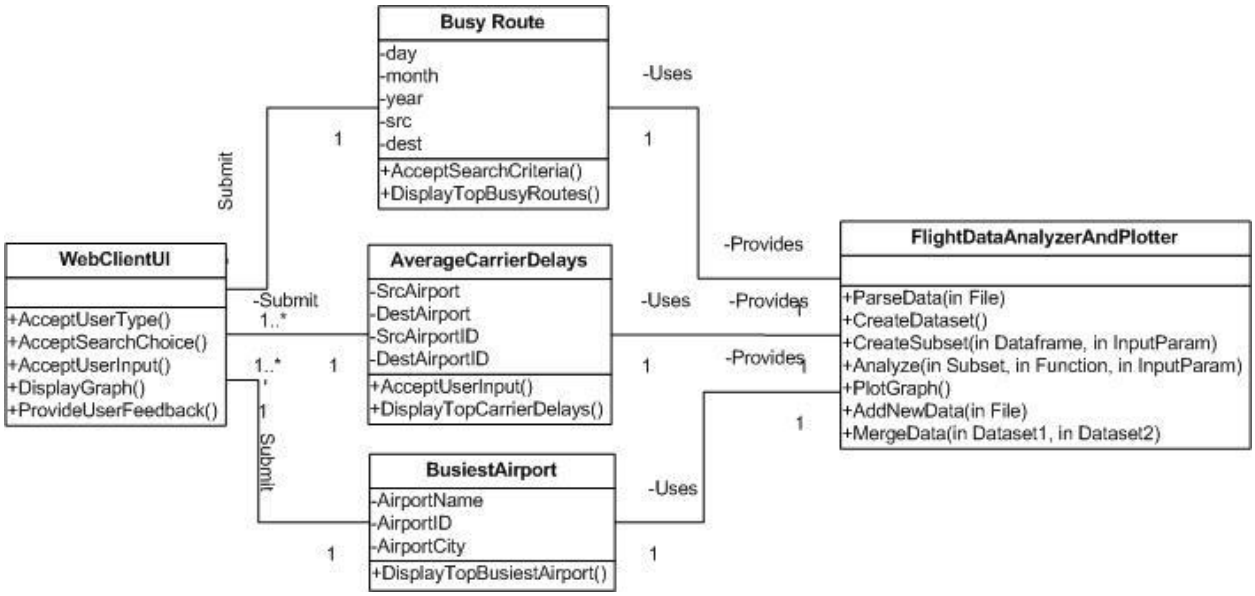
4. Sequence Diagram



5. Component Diagram



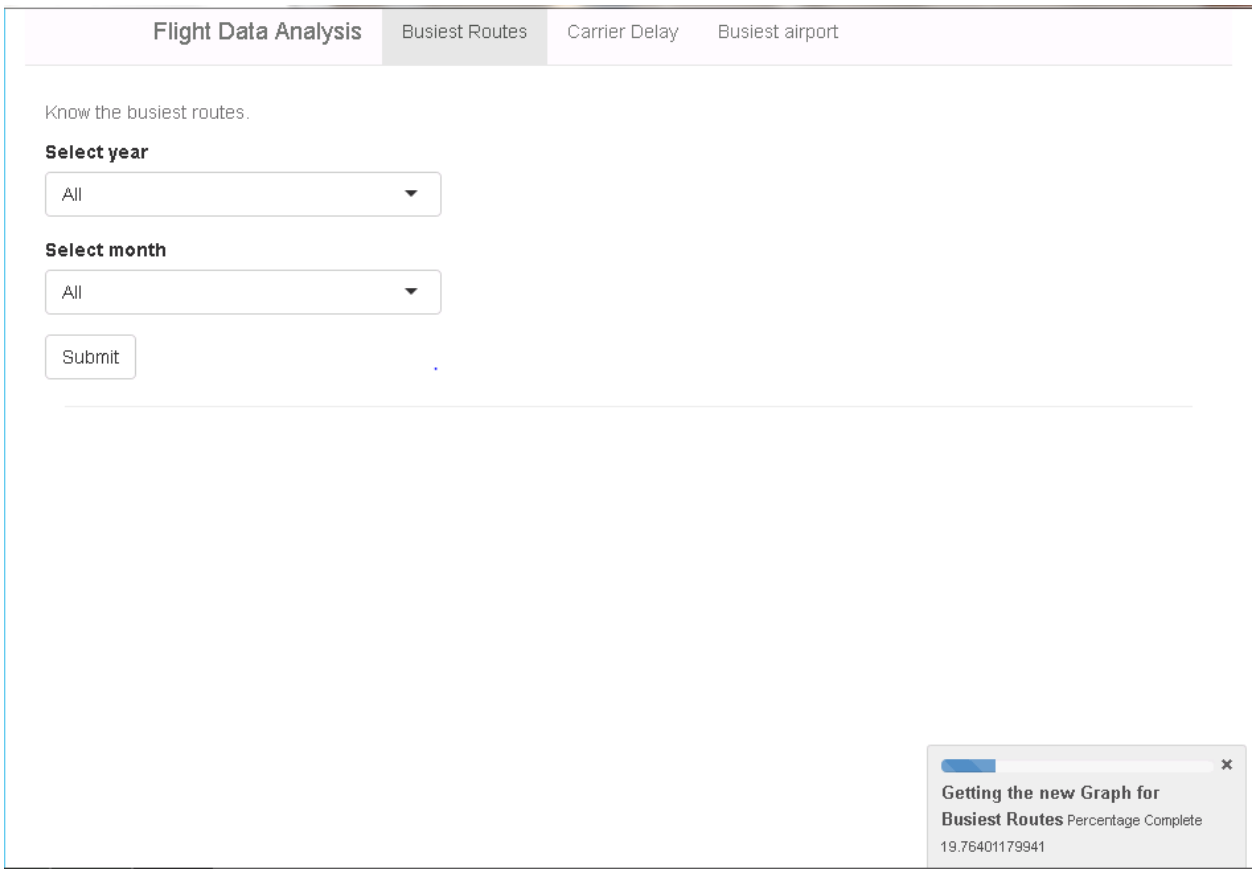
6. Class Diagram



RESULT ANALYSIS

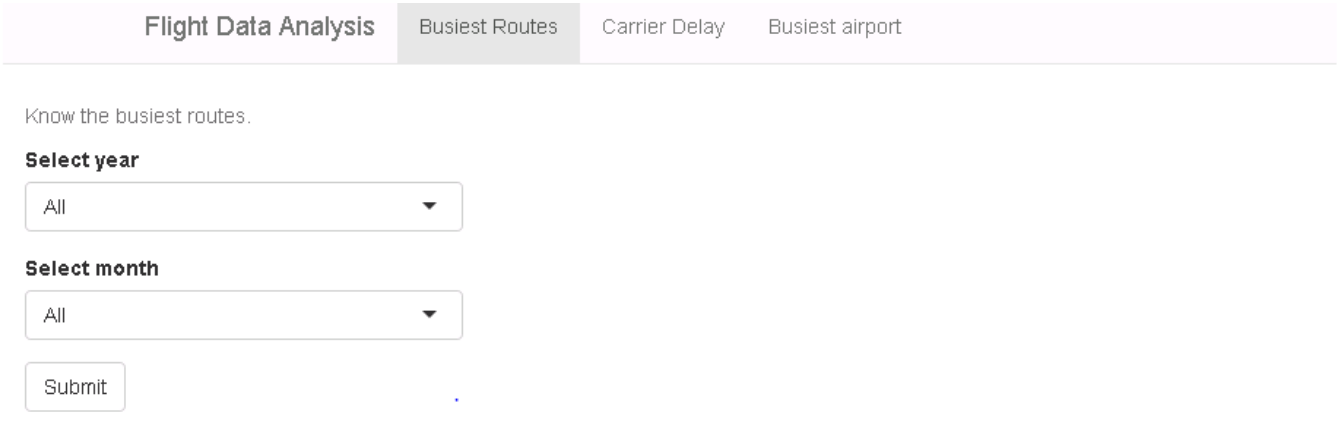
Below are some of the graphs obtained during the flight data analysis.

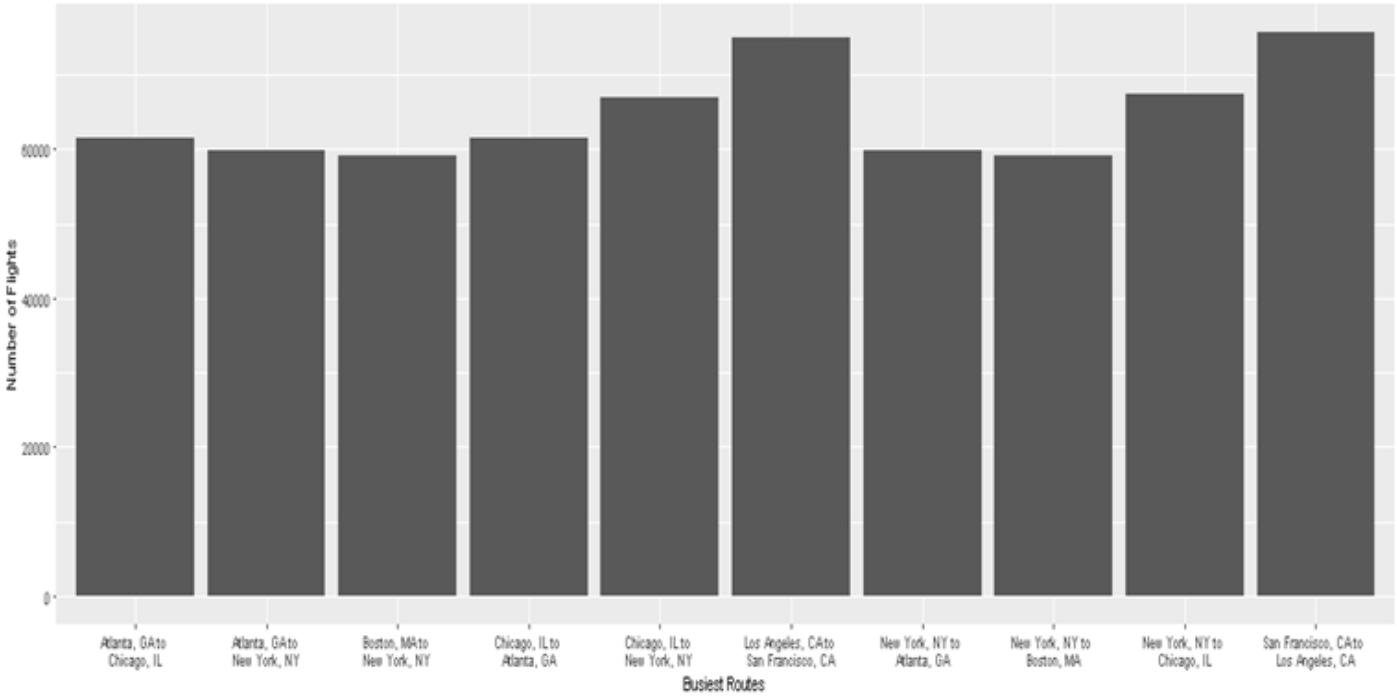
The progress bar at the right bottom corner indicates the progress of the graph generation.



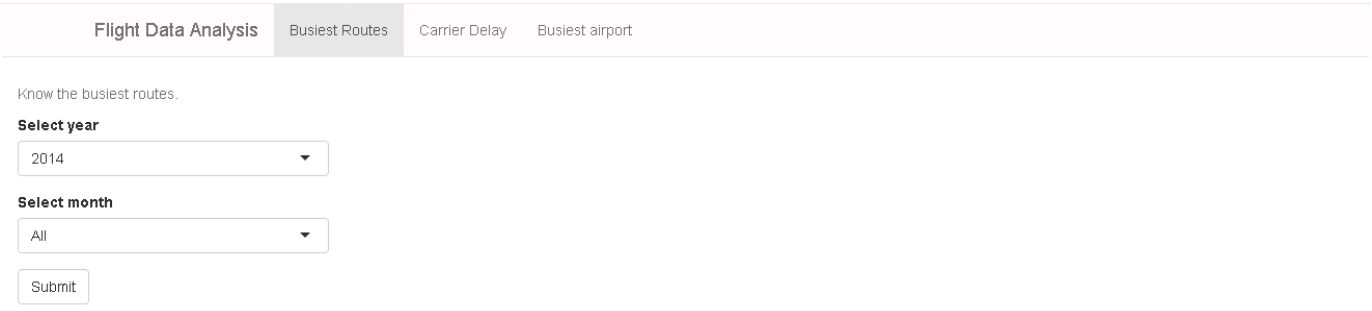
1) Busiest routes

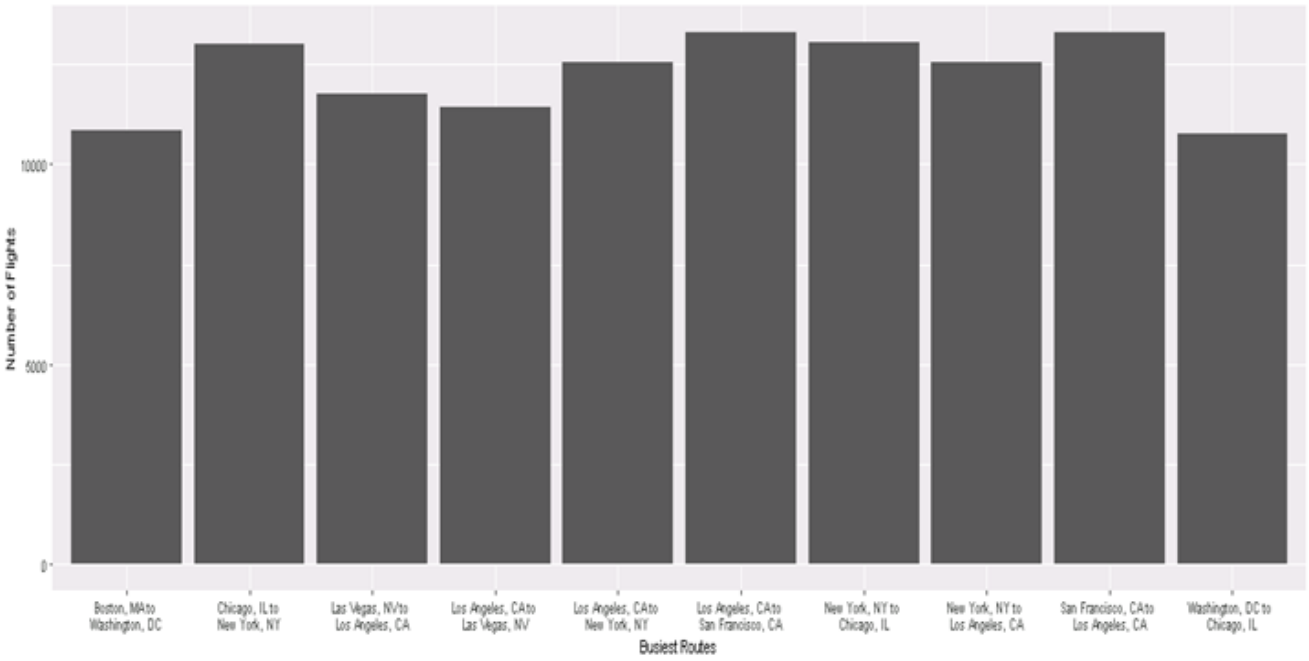
a) The below graph is obtained gives information regarding the busiest routes between the years 2011 - 2015. The busiest route in from San Francisco, CA to Los Angeles, CA with nearly 68000 flights as shown in the graph.



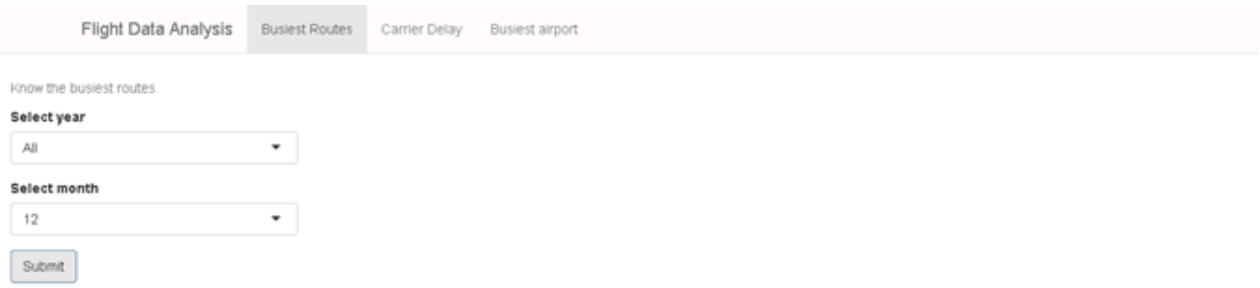


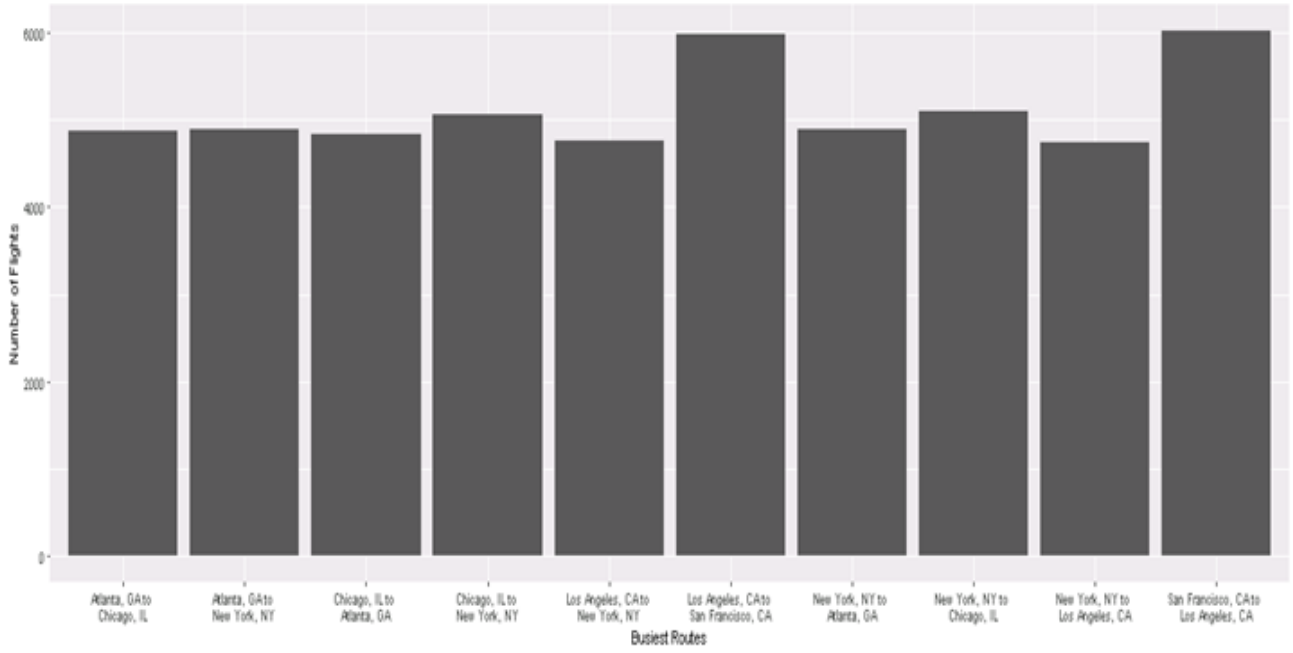
b) The below graph provides information regarding the busiest routes in the year 2014. The busiest route is from Los Angeles, CA to San Francisco, CA and approximately 13000 flights flying in this route.





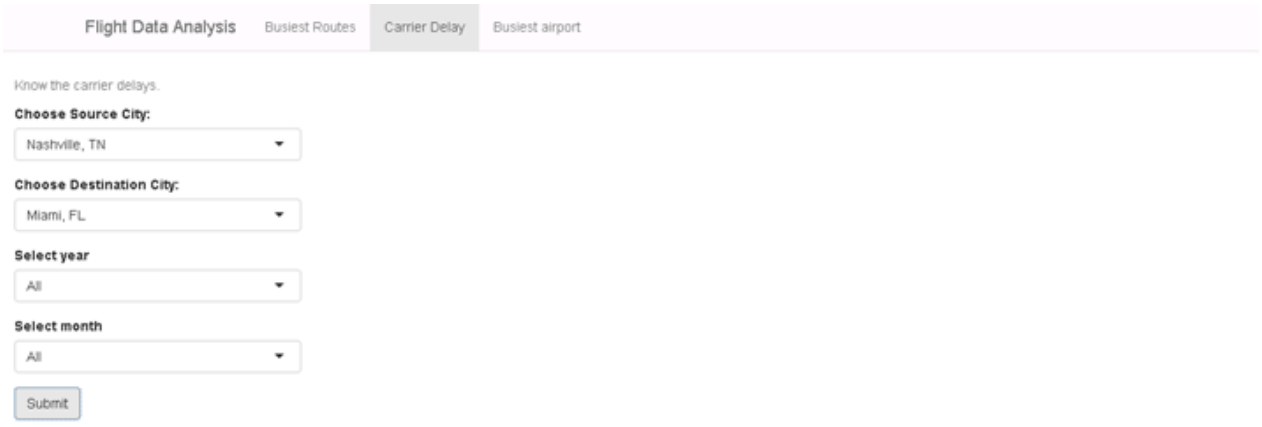
c) The below graph provides information regarding the busiest routes in the December month. The busiest route is from Los Angeles, CA to San Francisco, CA and approximately 6050 flights flying in this route in December.

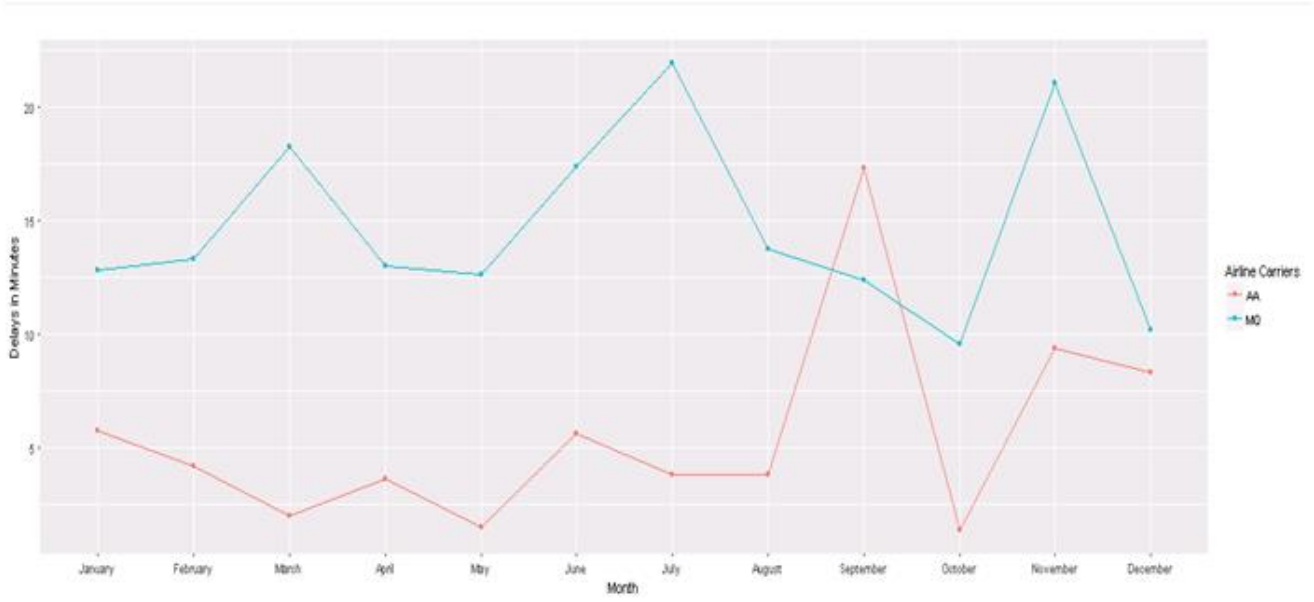




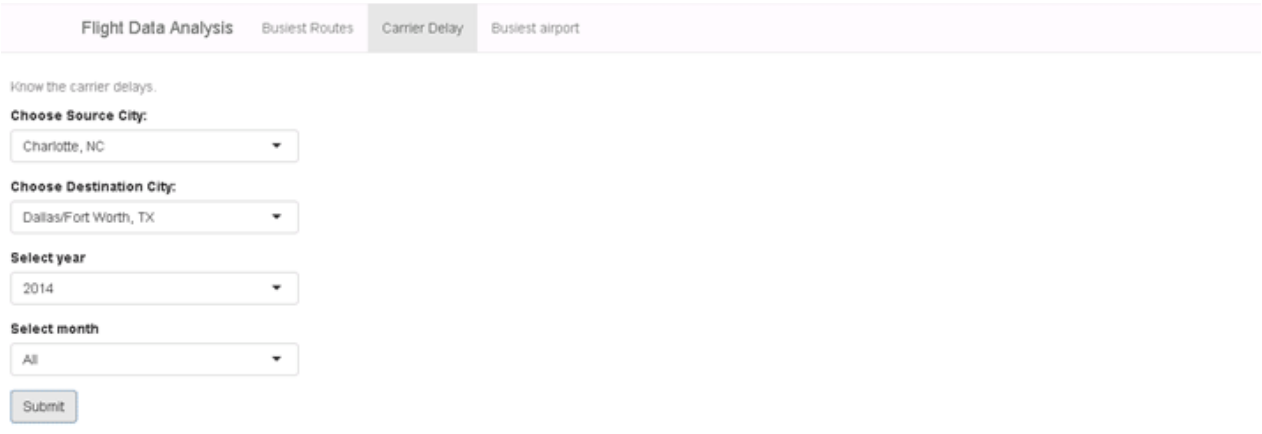
2) Carrier delay

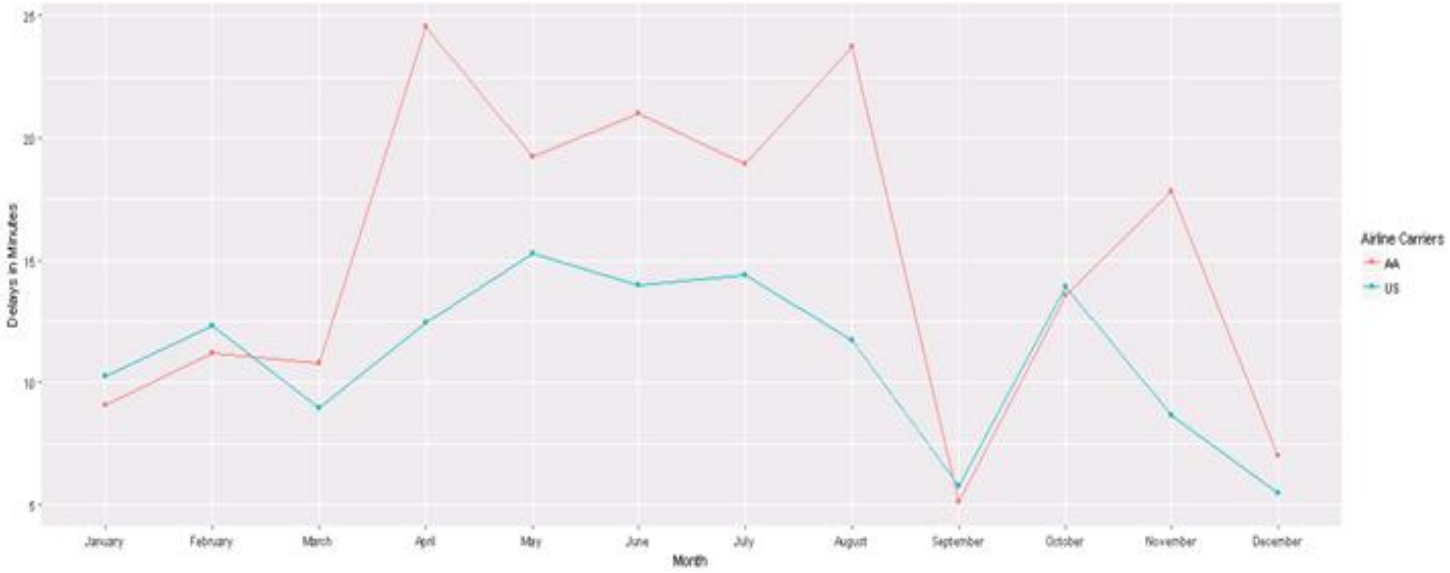
a) This graph gives you month-wise analysis of carrier delay for airline carriers flying from Nashville, TN to Miami, FL. The American Airlines, as per the analysis, is usually delayed by 18 minutes (approx) in the month of September and 2 minutes in the month of October



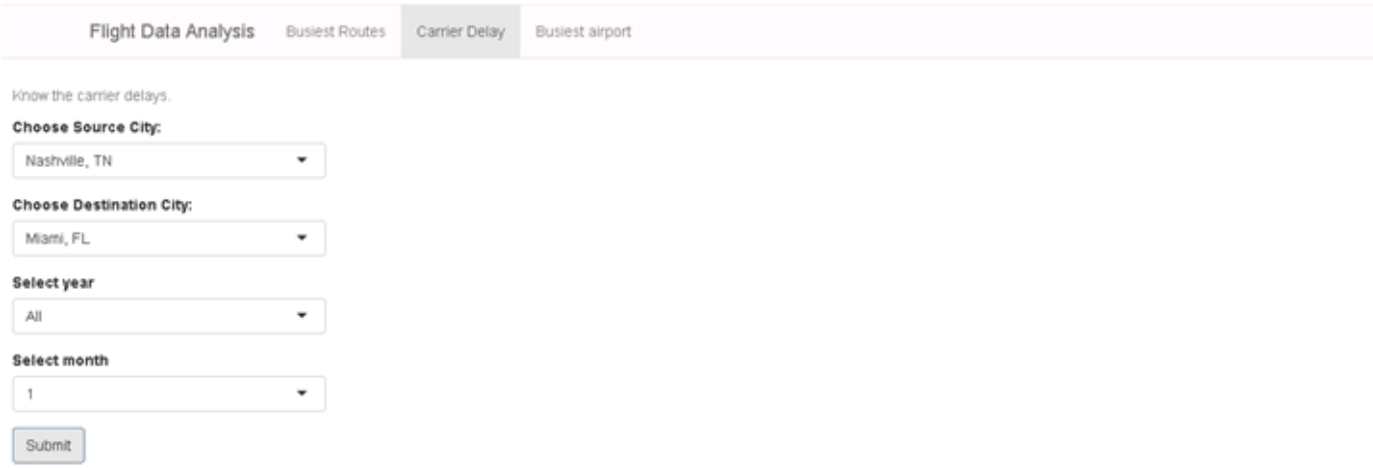


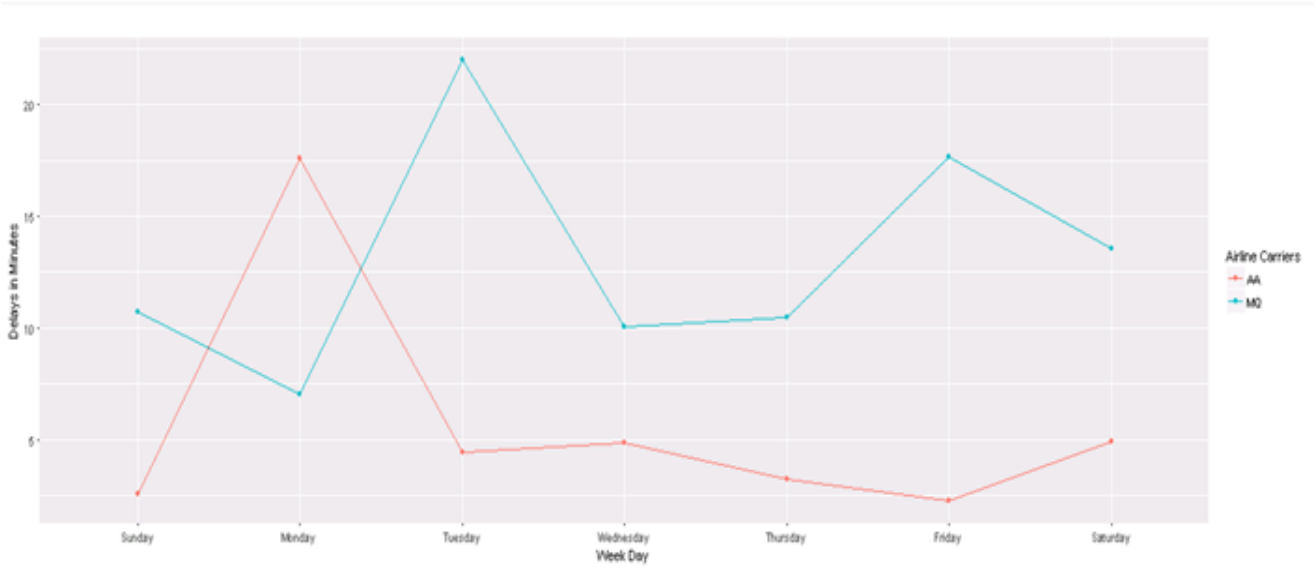
b) The below graph represents the delays for carriers flying from Charlotte, NC to Dallas, TX for each of the 12 months in the year 2014. For example, for the American Airlines, the highest delay (nearly 24 minutes) in this route has been in the month of April and the least delay (nearly 5 minutes) in September.





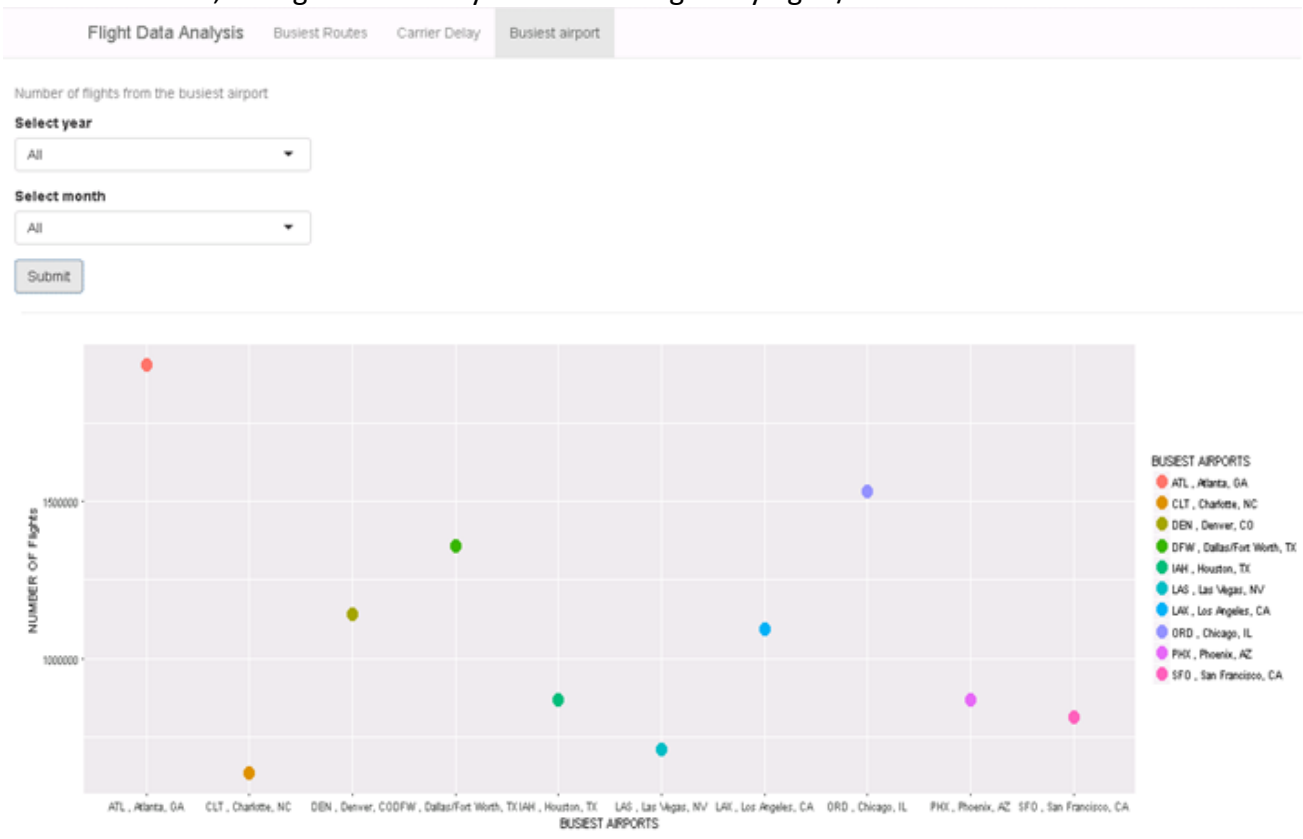
c) The below graph gives the day-wise carrier delay analysis for airlines flying from Nashville, TN to Miami, FL in the month of January. For example, the American Airlines carriers are usually delayed by 5 minutes on Fridays and 18 minutes on Mondays.





3. Busiest Airports

a) This graph gives information about the busiest airports in USA based on the flight data analysis. As shown by the graph, the busiest airport between 2011 to 2015 is the airport in Atlanta, Georgia with nearly two million flights flying to/from Atlanta.



b) The graph obtained below provides information about the top 10 busiest airports in USA in the month of June. The airport at Atlanta is the busiest with 150000 flight flying in/out of ATL Airport

c) This graph gives information about the busiest airports in the July month of 2011. ATL airport in Atlanta, Georgia is the busiest with 35000 flights flying in and out of the airport in July, 2011

TEST CASES

S. No	Test Case Description	Expected Result	Actual Result
TC01_Invalid_File_Input	An invalid file given as input (For ex: Files with different extension than .csv)	File not found	Data not found
TC02_Invalid_Selection_Of_Cities	City not present in the data set is selected	Data not found	Data not found
TC03_Carrier_Delay_Month_selection_All	Month selected as all in Carrier Delay	X axis having Month information (Jan, Feb etc)	X axis having Month information (Jan, Feb etc)
TC04_Carrier_Delay_Month_selection	Particular month selected in month input	X axis having week information(Mon, Tue, Wed etc)	X axis having week information(Mon, Tue, Wed etc)
TC05_Source_And_Destination_Are_Same	User enters the same city as source and destination	Enter proper details	Enter proper details
TC06_Invalid_Year	User select year not present in data set	Data not found	Data not found
TC07_Invalid_Month	User select month not present in data set	Data not found	Data not found
TC08_Concurrent_Request	User clicks submit button from all criteria	Request to be queued	Request is queued and displayed in first in first out manner

CONCLUSION

The following are the conclusions of our work on analysis of flight data,

- The busiest airport is O'Hare international airport, Chicago, Illinois.
- As could be expected, there is a strong correlation between departure delay and arrival delay.
- There is a difference between destination airports concerning the distribution of delayed flights. We can see from the graphs that still a lot of flights reach their destination on time or even earlier.

FUTURE ENHANCEMENTS

- A website can be created to load and analyze the data.
- More accurate results can be generated using machine learning concepts.
- Future delays can be predicted using regression model
- Enabling ticket booking functionality to the application.
- Expanding this to other means of transports also such as bus, taxi etc.

REFERENCES

1) Shiny Web Tutorial - https://shiny.rstudio.com/tutorial/
2) Lucid Charts to draw UML Diagrams - https://www.lucidchart.com.
3) R tutorial online - https://www.datacamp.com/
4) Flight information data set - http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236