

School of Information Technologies
Faculty of Engineering & IT

ASSIGNMENT/PROJECT COVERSHEET - GROUP ASSESSMENT

Unit of Study: COMP5048

Assignment name: Assignment 2-Flights data set (flight delay data)

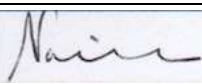
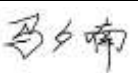
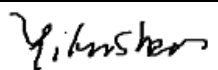



Tutorial time: 20:00 Thursday **Tutor name:** Nguyen Quan

DECLARATION

We the undersigned declare that we have read and understood the [University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy](#), and, except where specifically acknowledged, the work contained in this assignment/project is our own work, and has not been copied from other sources or been previously submitted for award or assessment.

We understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to severe penalties as outlined under Chapter 8 of the University of Sydney By-Law 1999 (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

We realise that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

Project Team Members				
Student name	Student ID	Contribution	Agree to Share	Signature
Nagib Shah	470360839	16.66%	Yes	
Xinan Ma	470489934	16.66%	Yes	
Alan Shen	470194809	16.66%	Yes	
Young Choi	470346338	16.66%	Yes	
Qiushi Zhang	470119101	16.66%	Yes	
Girishkumar Dhotarkar	470351620	16.66%	Yes	

Project Initial Report ----- Group 30

Flight Data Visualisation



27.09.2017

Revision History

Revision	Date	Status	Author	Summary of Changes
0.1	23-Sep-2017	Completed	Girish	Initial Draft
0.2	24-Sep-2017	Completed	All	Incorporated Review Comments
0.3	25-Sep-2017	Completed	All	Incorporated Review Comments
0.4	27-Sep-2017	In-Progres	Xinan	Incorporated Review Comments
0.5	27-Sep-2017	Complete	Nagib	Review and copy changes

Table of Contents

1. Introduction	5
1.1 Data Set	6
1.1.1 Selection process	6
1.1.2 Flight Data	6
1.2 Aims.....	9
2. Related Work.....	10
3. Design and Approaches	12
3.1 Framework of VA System - A hitchhiker's guide to U.S.A.....	12
3.2 Explanations of VA system framework design	13
3.3 Analysis processes and techniques of VA design	15
4. Implementation	18
5. Evaluation.....	20
5.1 Details of Evaluation Methods.....	20
5.1.1 Perform Survey	20
5.1.2 Feedback from Family and friends	20
5.1.3 Cross Validation with related work.....	21
5.1.4 Pro and Cons Analysis	21
6. Planning/Schedule & Role Assignments	22
7. References.....	24

Appendix: weekly group meeting minutes.

1. Introduction

As part of the assignment 2 of data visualisation, the project team (Group-30) has been tasked to study, analyse, identify and document the data visualisation topics and techniques, alongside the VA system that will assist in solving a series of tasks and questions as outlined in this report.

The objective of this report is to provide a detailed overview of the

- 1) Dataset selected
- 2) The problem statement(s)
- 3) High-level plans, sketches and prototypes for visualisations
- 4) Analysis tasks and techniques to be carried out for effective visualisation and summarisation of information.
- 5) VA system overview alongside the proposed tools
- 6) Project plan
- 7) Evaluation framework to address and assess if the VA system is able to cater to the problem statements/aims.

1.1 Data Set

1.1.1 Selection process

As part of assignment 2 activity, team members (Group-30) agreed on the following key criterion for selecting data set.

- Team skills profile
- Available references for Visualisation
- Visualisation achievable within timelines
- Quality of Visualisation achievable
- Complexity/Size of data

Several datasets were considered for selection by the project team and a decision was made based on the above-mentioned criterion.

1.1.2 Flight Data

After analysing each data set based on criteria above, team finalized on **Flight (delay) data set** available on <http://stat-computing.org/dataexpo/2009/>. The dataset contains largely US domestic flight network and performance details consisting of approximately 120 million flight arrival and departure information, metrics, and statistics ranging from the years 1987 till 2008. Furthermore, the dataset also contains airline/carrier, airport, and aircraft information that may provide useful insights.

After thorough study, cross reference, and consideration by the project team a decision was made to utilise the complete dataset (21 years approx.) for visualisation and analysis in order to present a complete and coherent summary of the US flight network and delay profiles. The following table outlines the structure and content of the dataset in detail.

Comp5048 Visual Analytics 2017 Group Assignment

Table 1.1.1: Base data (ONTIME)

Name	Description
Year	1987-2008
Month	Number
DayofMonth	Day
DayOfWeek	1 (Monday) - 7 (Sunday)
DepTime	actual departure time (local, hhmm)
CRSDepTime	scheduled departure time (local, hhmm)
ArrTime	actual arrival time (local, hhmm)
CRSArrTime	scheduled arrival time (local, hhmm)
UniqueCarrier	<u>unique carrier code</u>
FlightNum	flight number
TailNum	plane tail number
ActualElapsedTime	in minutes
CRSElapsedTime	in minutes
AirTime	in minutes
ArrDelay	arrival delay, in minutes
DepDelay	departure delay, in minutes
Origin	<u>origin IATA airport code</u>
Dest	<u>destination IATA airport code</u>
Distance	in miles
TaxiIn	taxi in time, in minutes
TaxiOut	taxi out time in minutes
Cancelled	was the flight cancelled
CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
Diverted	1 = yes, 0 = no
CarrierDelay	in minutes
WeatherDelay	in minutes
NASDelay	in minutes
SecurityDelay	in minutes
LateAircraftDelay	in minutes

Supplemental data Tables:

Table 1.1.2: Airports data

Name	Description
iata	Iata (airport code)
airport	Airport Name
city	City Name
state	State Abbreviation
country	Country Abbreviation
lat	latitude
long	Latitude

Table 1.1.3: Carriers

Name	Description
Code	Code of Airline
Description	Name of Airline

1.2 Aims

Team have gone through multiple iterations to identify aims of project exercise. The following objectives/questions were identified that will form the subject of this visualisation exercise and the VA system.

- a.** Visualise best time of year for travelling with minimal flight delays.
 - Any specific seasons/months is good or bad for travelling;
 - Weekends cause more delay than weekdays;
 - Any specific airport or air route to avoid while travel booking.

- b.** Visualise performance of airports and carrier year on year
 - Best performing airport
 - Worst performing airport
 - Best performing airline/carriers
 - Worst performing airline/carriers

- c.** Visualise negative effect of specific incident or event on flight delay
 - 1999-2000 Y2K bug effect on flight delay
 - 9/11 World trade centre terrorist attack effect on flight delay

- d.** Visualise best performing airport(s) performance for a single year.

2. Related Work

Winner 1: Congestion in the sky

<http://stat-computing.org/dataexpo/2009/posters/wicklin-allison.pdf>

Brief: Effective visualisation of the US flight network that indicates delays and overall performance of airports and airlines. The work/submission includes a comprehensive set of visualisations ranging from network overview, to temporal and carrier analysis as well as statistical inference of the dataset. This submission has the most amount of synergy with our project aims and tasks.

Table 2.1: Strengths and weakness analysis for related work (I)

Strengths	Weaknesses
<ul style="list-style-type: none"> The interactive graph for percentage of flight delays in the network and provides an easy to use and clean visualisation. Effective visualisation of the seasonal effects of the US domestic flight network Historical comparison of each airport in the form of Box-Plots 	<ul style="list-style-type: none"> The carrier effect visualisation is not very effective at conveying the overall carrier performance.

Winner 3: A tale of two Airports

<http://stat-computing.org/dataexpo/2009/posters/wickham.pdf>

Brief: Visualisations and charting aimed at providing insights related to specific events and incidents only in the US flight network. The work focuses mostly on two specific airports (SFO and OAK) and the effects of major world events (e.g. 911)

Table 2.2: Strengths and weakness analysis for related work (II)

Strengths	Weaknesses
<ul style="list-style-type: none"> Clean VA framework focusing on a small subset (2 airports) and therefore deals with limited data. Focuses on specific events and targeted visualisation to show effect. Good use of colour schemes on line chart to show carrier and airport effects. 	<ul style="list-style-type: none"> Only focuses on two airports as opposed to the flight network Does not utilise any network graphs and relies mostly on charting and plotting to infer information and patterns. Excessive utilisation of line charts with minimal variety.

Independent work: Flight data visualisation using google fusion tables -

<http://xliberation.com/googlecharts/d3flights.html>

Brief: This reference is very useful as it attempts to build/augment rich interactive features with visualisation of the US flight network. The work includes custom airport/node centrality calculations, custom edge weights, as well as incorporation of colours scales and schemes to effectively convey the overall status and the local network of a specific airport(s).

Table 2.3: Strengths and weakness analysis for related work (III)

Strengths	Weaknesses
<ul style="list-style-type: none">• Interactive features that allows selection of specific airports and its immediate neighbours.• Effective use of colour schemes to indicate performance• Easy identification of key airports by node size.• Summary of airport and key statistics shown with airport selection	<ul style="list-style-type: none">• Does not provide a grand overview of the flight network. User must interact and select airport first before any visualisations are presented.• Does not indicate US states outside of the mainland (e.g. Alaska, Hawaii)• Does not utilise edge bundling to limit the total number of edges for overviews for a busy network.

3. Design and Approaches

3.1 Framework of VA System - A hitchhiker's guide to U.S.A

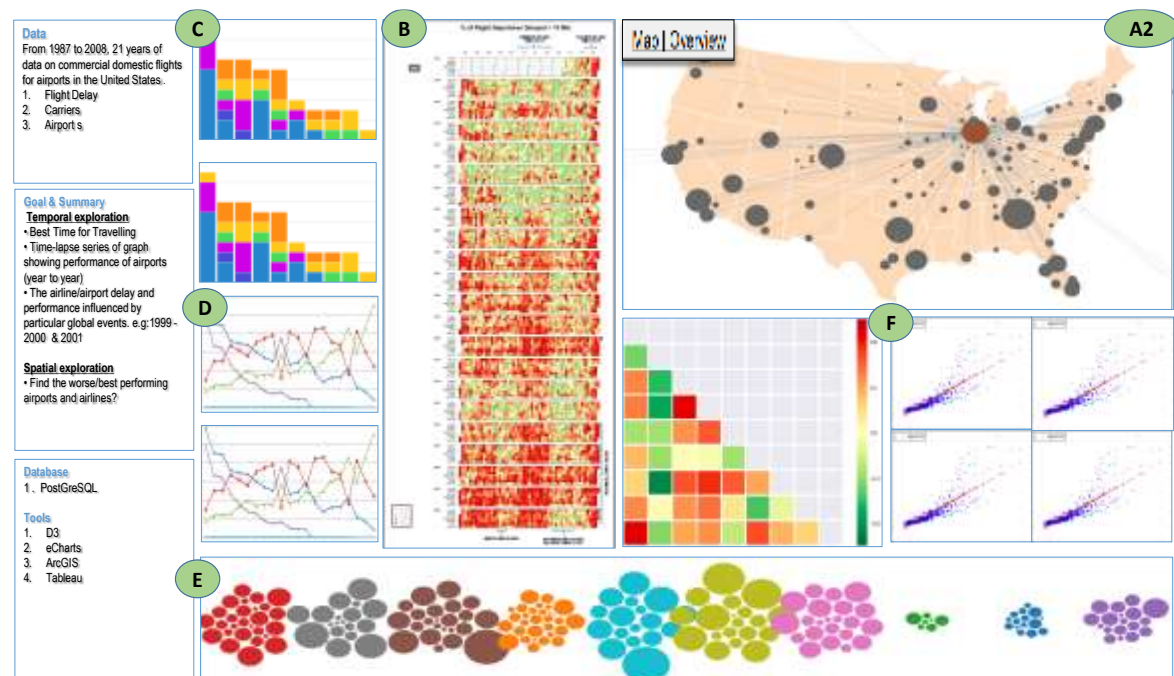
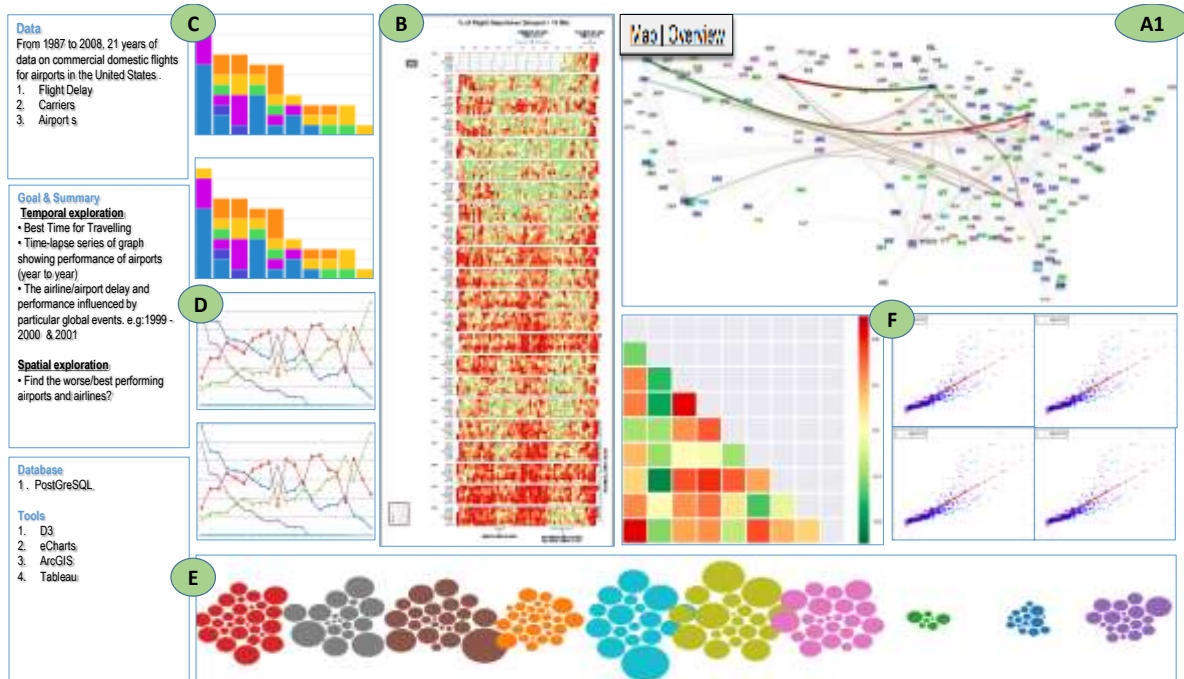


Figure 3.1.2: VA Framework with interactive network visualisation

3.2 Explanations of VA system framework design

Table 3.2.1: Explanation of each visualization design

Label	Purpose	Explanation & Justification
A1	Primary Network Graph - Summarize and visualise 21 years of US domestic flight network data for flight delays	<p>Nodes identified in different colours by using k-core analysis in order to visualise sub-graph which represents only airports with minimum number of flights (done to reduce size of the flight network).</p> <p>Edges (Airlines route) in different weights affected by number of flights alongside custom edge attribute of delay percentage. This is done to reduce the overall edges within the network.</p> <p>Betweenness centrality analysis will be used to identify airports by their shortest path and thereby critical to the overall flight network.</p> <p>For visualisation software tool ArcGIS is proposed for creating Geo-spatial layout which will represent airport and airline routes which cause most delay in entire network.</p>
A2	Interactive Visualisation to show the networks of each airports / airlines by flight delay	<p>Airport circle size represent out-degree centrality, colour scale will be applied to indicate percentage delay. This interactive visualisation will have drop down for year from 1987 to 2008. Interaction with airport node (Hover) on graph to highlight network associated with airport alongside a brief summary.</p> <p>Software tool proposed for this visualization is D3 to cater to interactive features that are required for effective summarisation of the network.</p> <p>There will be switch button [Map Overview] created to toggle between A1 and A2 visualisation.</p>
B	Calendar view of flight delay to show overall flight delay by day/month/year over 21 years	<p>Each day represented by squares and heat map technique will be used to depict flight delay percentage for the day. Due to the size of the dataset, heat map is considered to be the appropriate technique to represent and disseminate the delay data effectively.</p> <p>Visualisation will be represented by calendar layout</p>

Comp5048 Visual Analytics 2017 Group Assignment

		from year 1987 to 2008 and D3 is proposed as the chosen tool to perform this task.
C	Bar Charts for representing top performing airports	<p>Performance of airport can be measured by percentage delay. Statistical analysis will be used on airport performance by year from 1987 to 2008 to identify top 12 airports.</p> <p>Visualisation will be represented by Bar charts as it is considered the simplest form of information summarisation from an end user perspective. Tableau is proposed for this visualisation piece.</p>
D	Line Charts representing performance of best airport for a year	<p>Summary of the best performing airport(s) analysed by statistical analysis method. Identify months of the year which cause more flight delay for top performing airport. Summarised monthly flight delay data can be represented effectively in line chart. Tableau is the proposed tool for this exercise.</p>
E	Bubble Charts for airlines comparison analysis for top performing airlines	<p>Top 10 airlines performance shown from 1987 to 2008.</p> <p>Each cluster/colour represent airline, and the circle size to represent % delay. Force directed algorithm is proposed because It produces good quality result for 50–500 node size graph and simple for analysis. D3 is proposed as the tool best for this exercise.</p>
F	Correlation Matrix showing the relationship between delay reasons and the overall arrival delays to identify primary effect on flight delay	<p>To show effect of various delay reasons Statistical Analysis will be performed. Correlation Matrix to be generated for various dimensions like whether, security, distance etc. over the predicted value of flight delay. The purpose of this exercise is to identify and calculate the correlation figures associated with each delay reasons (% effect of reasons).</p> <p>scatter plot utilising the ggPlot library in R is proposed as the best fit tool to carry out such analysis and visualisation.</p>

3.3 Analysis processes and techniques of VA design

Table 3.3.1: Analysis techniques illustration of each visualization design

Label	Data Processing	Analysis Technique / Algorithm / Methods
A1	Primary Network Graph - Summarize and visualise 21 years of US domestic flight network data for flight delays Tables/Entities used: 1. Ontime 2. airports	K-core Betweenness centrality
A2	Interactive Visualisation to shows the networks of each airports / airlines by flight delay Tables/Entities used: 1. Ontime 2. Airports 3. Carriers	Out-degree Centrality
B	Calendar view of flight delay to show overall flight delay by day/month/year over 21 years Tables/Entities used: Ontime	Heat map
C	Bar Charts for representing top performing airports Tables/Entities used: 1. Ontime 2. Airports	Statistical analysis
D	Line Charts representing performance of best airport for a year Tables/Entities used: 1. Ontime 2. Airports	Statistical analysis
E	Force Bubble Charts for airlines comparison analysis for top performing airlines Tables/Entities used: 1. Ontime 2. Carriers	Force Directed algorithm
F	Correlation Matrix and scatter plots Table/Entities used: Ontime	Statistical Analysis /Correlation Calculation

3.4 Summary of Visualisation

Table 3.4.1: Summary of visualization system design

Label	Visualisation	Addressing Aims	Visualisation Layout	Tool Proposed
A1	Primary Network Graph - Summarize and visualise 21 years of US domestic flight network data for flight delays	Any specific airport or routes to avoid while travelling	Geo-spacial Layout	PostgreSQL NetworkX ARCGIS
A2	Interactive Visualisation to show the networks of each airports / airlines by flight delay	<ul style="list-style-type: none"> Any specific airport or routes to avoid while travelling 9/11 World trade centre terrorist attack effect on flight network Identify effect of Y2K bug during 1999-2000 period 	Geo-spacial Layout	PostgreSQL D3, python
B	Calendar view of flight delay to show overall flight delay by day/month/year over 21 years	<ul style="list-style-type: none"> Identify seasonal patterns in flight network that may assist with avoiding specific months/weekdays/weekends while travelling. 1999-2000 Y2K bug effect on flight delay 9/11 World trade centre terrorist attack effect on flight network 	Calendar Layout	PostgreSQL , D3
C	Bar Charts for representing top performing airports	Best/Worst performing airport which should be avoided while booking for travel	Bar Charts	PostgreSQL Tableau

Comp5048 Visual Analytics 2017 Group Assignment

D	Line Charts representing performance of best airport for a year	Spotlight and visualise best airport(s) performance over a period	Line Charts	PostgreSQL Tableau
E	Bubble Charts for airlines comparison analysis for top performing airlines	Best/Worst performing airline/carriers	Bubble Chart Layout	PostgreSQL D3
F	Correlation Matrix depicting the relationship between delay reasons/factors to identify and quantify the overall flight delay	Quantify and calculate the % effect of various factors to the flight delay (e.g. 15% delays likely caused by bad weather)	Correlation Matrix Scatter Plot	PostgreSQL R ggPlot

4. Implementation

Initial brief of full implementation is given in following steps:

Step 1. Extract & Load of data

- Download all data files from websites and put it in local drive.
- Use PostgreSQL utility to import data into database.
- Load all 21 years of flight delay data into Ontime table.
- Load airport data into Airport table and carrier data into carrier table.
- Check errors and remediate error while loading data.

Step 2. Cleanse data & Create Indexes

- Verify data has been completely loaded and no errors.
- Correct data type for all columns.
- Create indexes for all important columns and combinations to allow fast retrieval of information (query performance).

index samples

```
create index year on ontime(year);
create index date on ontime(year, month, dayofmonth);
create index flightorigin on ontime (origin);
create index flightdest on ontime (dest);
create index flightarrdelay on ontime (arrdelay);
```

Step 3. Develop & test queries for each visualisation

- Develop queries required for each visualisation.

e.g. delay summary all airports

```
select f.origin, f.dest as destination, count(*) as count,
sum(arrdelay) as total_arrival_delay_minutes,
d.number_of_delays,
(d.number_of_delays * 100)/count(*) as delayPercentage
from ontime f,
(select origin,
dest,
count(arrdelay) as number_of_delays
from ontime
where arrdelay > 15
and year=2008
group by origin, dest) d
where f.origin = d.origin and f.dest = d.dest and year=2008
group by f.origin, f.dest, d.number_of_delays;
```

Comp5048 Visual Analytics 2017 Group Assignment

- b.** Get additional calculated/transformed data (e.g. % delay, performance percentage etc.)
- c.** Test and refine queries to get correct outputs
- d.** Download query results in csv, excel, graphml formats for easy port and load in visualisation tools.

Step 4. Develop code or use software tools to visualise & perform testing

- a.** Carry out network analysis in NetworkX (e.g. centrality calculations) where applicable.
- b.** Develop code or use tool(s) to generate visualisations (e.g. JavaScript for D3).
- c.** Generate an overview of the visualisation of complete data
- d.** Reduce, refine and iterate through visualisations until satisfactory results are produced.
- e.** Write code or use tool options to scale colour, size, text, resolution and layout of visualisations.
- f.** Debug and QA code
- g.** QA visualisations.
- h.** Refine visualisations and finalise.

Step 5. Peer review

- a.** Perform peer review of each visualisation.
- b.** Select top 2 visualisations.

Step 6. Evaluate and finalise visualisation

- a.** Perform Evaluation for each visualisation
- b.** Finalize visualisations for VA system as well as final report.

5. Evaluation

The table below outlines the template to be utilised/followed for evaluation for each of the graphs and the related tasks/aims associated.

Table 5.1: Evaluation template of each visualisation

Data/Visualisation	Visualisation name (e.g. A1 – Primary Network Graph)
Task(s)	<ul style="list-style-type: none"> Identify critical routes Identify critical airports/nodes Summary of overall network
Evaluation framework/Methods	<ul style="list-style-type: none"> Peer Review Feedback from focus group (questionnaire) Cross validation utilising past references
Pros/Cons Analysis	Identify and document pros/cons and limitations

5.1 Details of Evaluation Methods

5.1.1 Perform Survey

Various visualisation for same data set will be taken to focus group for visualisation quality and provide feedback via simple questionnaire/survey. Visualisation which rated higher will be selected for final report.

Table 5.1.1: Evaluation survey table

Visualisation Versions	Probably Worst	-----			Probably Best	Selected Option
Version 1	1	2	3	4	5	
<div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>						
Version 2	1	2	3	4	5	
<div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>						

5.1.2 Feedback from Family and friends

Every team member to show and explain Visualisation to their family / friends and provide feedback to the group. Selection of Visualisation for final report based on feedback provided.

5.1.3 Cross Validation with related work

Perform comparison with flight delay data Visualisation outlined in related work section.

5.1.4 Pro and Cons Analysis

Analyse and document the pros and cons of each visualization alongside any limitations.

6. Planning/Schedule & Role Assignments

Table 6.1: Assignment time schedule

Tasks	Start Date	End Date	Actual Start Date	Actual End Date	Member Involved	Status
Analyse data sets and finalize data set	14/09	15/09	14/09	15/09	All	Completed
Decide on visualisation aims	16/09	18/09	16/09	18/09	All	Completed
Visualisation draft preparation and finalize	19/09	21/09	19/09	21/09	All	Completed
Initial report 1 documentation	21/09	28/09	21/09	28/09	All	Completed
Initial Draft	21/09	23/09	21/09	23/09	Xinan, Girish	Completed
Review revise	24/09	26/09	24/09	26/09	All	Completed
Finalize initial report	26/09	28/09	26/09	28/09	All	Completed
Design	21/09	28/09				In Progress
Flight delay visualisation (Heat Map)	21/09	28/09	21/09		Qiushi, Alan	In Progress
Airport causing flight delay (K-Core)	21/09	28/09	21/09		Nagib, Xinan	In Progress
Airport causing flight delay interactive visualisation (Out-Degree Centrality)	21/09	28/09	21/09		Nagib, Girish	In Progress
Weather Impact on Flight Delay (Co-relation Matrix)	21/09	28/09	21/09		Xinan, Nagib, Young	In Progress
Flight Delay due to various parameters e.g. Security, Weather etc. (ggPlot scatter plot)	21/09	28/09	21/09		Xinan, Nagib, Young	In Progress
Airline Performance Comparison (Bubble Chart)	21/09	28/09	21/09		Xinan, Alan	In Progress
Best airport visualisation (Bar charts and line chart)	21/09	28/09	21/09		Young, Girish	In Progress
Implementation	14/09	12/10	15/09			Not Started
Extract and load of data	15/09	22/09	16/09	22/09	All	Completed
Cleanse data and create indexes	22/09	25/09	22/09	26/09	All	Completed
Develop & test queries for each visualisation	26/09	30/09	26/09		All	In Progress
Coding / Use software						
Flight delay visualisation (Heat Map)	29/09	10/10	25/09		Qiushi, Alan	In-Progress
Airport causing flight delay (K-Core)	29/09	10/10			Nagib, Xinan	Not Started
Airport causing flight delay interactive visualisation (Out-Degree Centrality)	29/09	10/10	25/09		Nagib, Girish	In-Progress
Weather impact on flight delay (Co-relation Matrix)	29/09	10/10	25/09		Xinan, Young	Not Started

Comp5048 Visual Analytics 2017 Group Assignment

Flight delay due to various parameters e.g. Security, Whether etc. (ggPlot scatter plot)	29/09	10/10			Xinan, Young	Not Started
Airline performance comparison (Bubble Chart)	29/09	10/10			Xinan, Alan	Not Started
Best Airport Visualisation (Bar Charts and Line Chart)	29/09	10/10			Young, Girish	Not Started
Peer Review	10/10	12/10			All	
Evaluation	13/10	19/10			All	Not Started
Presentation	29/10	5/10			TBD	Not Started
Initial Draft	29/10	5/10			TBD	Not Started
Movie Creation	29/10	5/10			TBD	Not Started
Review	29/10	5/10			TBD	Not Started
Finalize Presentation	29/10	5/10			TBD	Not Started
Final Submission Documentation	5/10	19/10			TBD	Not Started
Initial Draft	5/10	19/10			TBD	Not Started
Review	5/10	19/10			TBD	Not Started
Finalize Initial Report	5/10	19/10			TBD	Not Started

Table 6.2: Assignment roles schedule

Roles	Team Members
Designer	Nagib, Xinan, Young
Developer	Nagib, Xinan, Qiushi, Alan, Young, Girish
Technical Writer	Qiushi, Girish, Young
Reviewer	Nagib, Xinan, Qiushi, Alan, Young, Girish
Evaluation Team Members	Nagib, Xinan, Qiushi, Alan, Young, Girish

7. References

Following are website/documents referred for this assignment.

- Force Layout Multiples
<https://bl.ocks.org/mbostock/1804889>
- Congestion in the sky – Data expo winner
<http://stat-computing.org/dataexpo/2009/posters/wicklin-allison.pdf>
- Data Expo runner up – flight data
<http://stat-computing.org/dataexpo/2009/posters/hofmann-cook.pdf>
- Data Expo runner up 2 – flight data
<http://stat-computing.org/dataexpo/2009/posters/wickham.pdf>
- Minimizing Flight Delay - Tanujit Dey • David Phillips • Patrick Steele
<http://stat-computing.org/dataexpo/2009/posters/dey-phillips-steele.pdf>
- Circular Heat map - <https://github.com/prcweb/d3-circularheat>
- Interactive flight network visualisation
<http://mbostock.github.io/d3/talk/20111116/airports.html>
- Interactive flight network visualisation using google fusion tables
<http://xliberation.com/googlecharts/d3flights.html>
- Circular Visualisation D3
<https://github.com/nicgirault/circosjs>