

**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) Final Report**

**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,*an, 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.

SIT Building, J12 **T** +61 2 9351 3423 ABN 15 211 513 464

The University of Sydney **F** +61 2 9351 3838 CRICOS 0

NSW 2006 Australia **E** sit.info@sydney.edu.au

**sydney.edu.au/it**

**Project Final Report ----- Group 30**

**Flight Data Visualisation**

**30.10.2017**

****

**Revision History**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Revision | Date | Status | Author | Summary of Changes |
| 0.1 | 25-Oct-2017 | Completed | Alan | Initial Draft |
| 0.2 | 29-Oct-2017 | Completed | Girish | Second Version |
| 0.3 | 29-Oct-2017 | Completed | Xinan | Third Version |
| 0.4 | 31-Oct-2017 | Completed | Girish | Incorporated Review comment |
|  |  |  |  |  |

**Table of Contents**

[1. Aim & Contributions 5](#_Toc497261716)

[1.1 Aim 5](#_Toc497261717)

[1.2 Contributions 5](#_Toc497261718)

[2. Data Set 6](#_Toc497261719)

[3. Design 8](#_Toc497261720)

[3.1 Framework of VA System – A hitchhiker’s guide to U.S.A. 8](#_Toc497261721)

[3.2 Analysis &Final Visualisation 10](#_Toc497261722)

[3.2.1 Visualisation Section1: Flight Network-View Interactive Maps 10](#_Toc497261723)

[3.2.2 Visualisation Section 1: Flight Network-View Overview 11](#_Toc497261724)

[3.2.3 Visualisation Section 1: Flight Network-View Delay Breakdowns 12](#_Toc497261725)

[3.2.4 Visualisation Section 2: Seasonal Effect 14](#_Toc497261726)

[3.2.5 Visualisation Section 3: Airline Performance 15](#_Toc497261727)

[3.2.6 Visualisation Section 4: Best/Worst Airport 16](#_Toc497261728)

[3.2.7 Visualisation Section 5: Delay Statistic Analysis 18](#_Toc497261729)

[4. Implementation 19](#_Toc497261730)

[5. Evaluation 20](#_Toc497261731)

[5.1 Results 21](#_Toc497261732)

[5.1.1 Focused group 21](#_Toc497261733)

[5.1.2 Survey (Questionnaire) Results 24](#_Toc497261734)

[5.1.3 Survey (Interview) 26](#_Toc497261735)

[5.1.4 Empirical evaluation (statistical analysis) 26](#_Toc497261736)

[5.2 Discussion 28](#_Toc497261737)

[6. Conclusion 29](#_Toc497261738)

[7. References 30](#_Toc497261739)

[8. Appendix: Group meeting minutes 31](#_Toc497261740)

[9. Appendix: Code 46](#_Toc497261741)

# Aim & Contributions

## Aim

The following objectives/questions were identified that will form the subject of this visualisation exercise and the VA system.

1. **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

1. **Visualise performance of airports and carrier year on year**

• Best performing airport

• Worst performing airport

• Best performing airline/carriers

• Worst performing airline/carriers

1. **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.**

## Contributions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project Team Members** | | | | |
| **Student name** | **Student ID** | **Contribution** | **Agree to Share** | **Signature** |
| Nagib Shah | 470360839 | 16.66% | Yes |  |
| Xinan Ma | 470489934 | 16.66% | Yes | signature paper_副本 |
| Alan Shen | 470194809 | 16.66% | Yes | sig2 |
| Young Choi | 470346338 | 16.66% | Yes | signature young |
| Qiushi Zhang | 470119101 | 16.66% | Yes | signature rachel |
| Girishkumar Dhotarkar | 470351620 | 16.66% | Yes |  |

# Data Set

The dataset we chose for this project is **Flight (delay) data set** available on [***http://stat-computing.org/dataexpo/2009/***](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.

The project team applied 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.

**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 | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 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 | [origin IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| Dest | [destination IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 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 contain Airports and Carriers information have been provided below:

**Table 1.1.2: Airports data Table 1.1.3: Carriers**

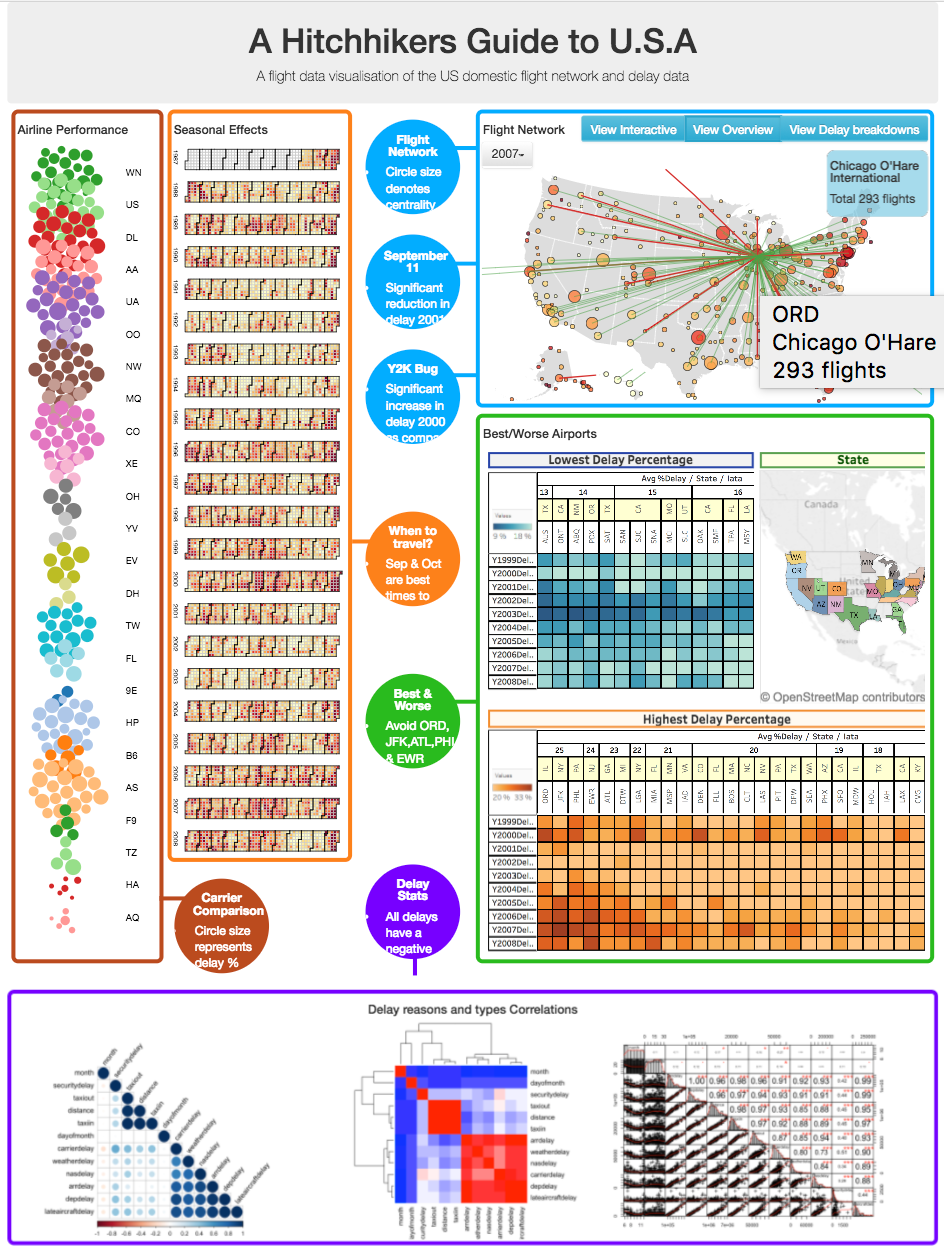
|  |  |
| --- | --- |
| **Name** | **Description** |
| iata | Iata (airport code) |
| airport | Airport Name |
| city | City Name |
| state | State Abbreviation |
| country | Country Abbreviation |
| lat | latitude |
| long | Latitude |

|  |  |
| --- | --- |
| **Name** | **Description** |
| Code | Code of Airline |
| Description | Name of Airline |

# Design

## Framework of VA System – A hitchhiker’s guide to U.S.A.

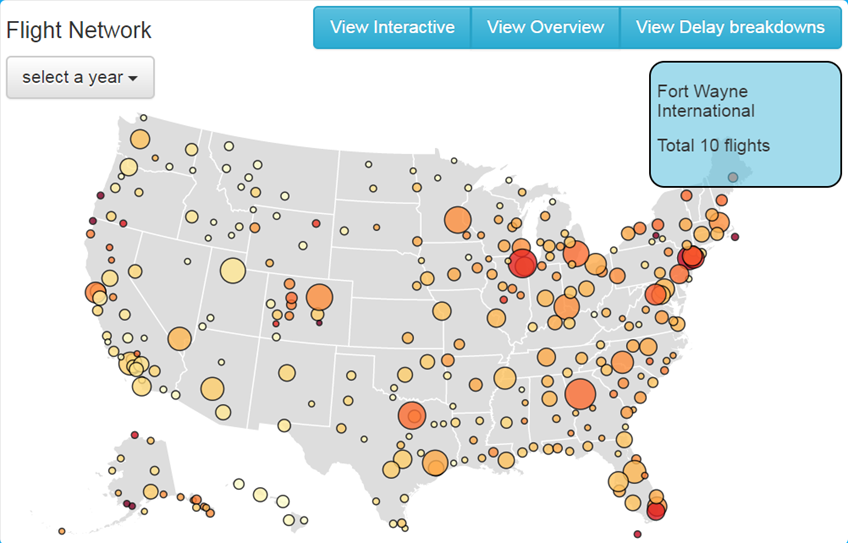
The main Flight Data Visualisation System has five visualisation sections: **Flight Network**, **Seasonal Effects**, **Airline Performance**, **Airport Performance**, and **Statistical Analysis**. There are four types of visualisation tool (D3, ArcGIS, Tableau, R) have been used to create total 10 visualisations.



**Figure 3.1.1: VA Framework**

## Analysis &Final Visualisation

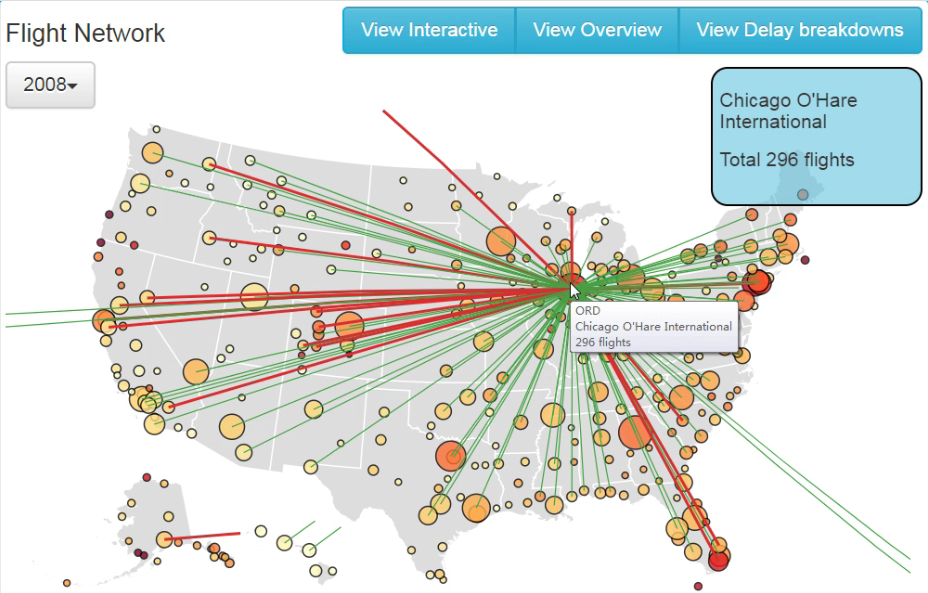
### Visualisation Section1: Flight Network-View Interactive Maps



**Figure 3.2.1: Interactive Flight Network (Nodes)**

**Purpose:** This interactive flight network designed in **D3** to show the networks of airports / airlines by flight delay in 21 years (since 1987-2008) in given data set. This visualisation is to assist to find out some events impacts and seasonal variation of airline delays. Visualisation allows user to select year, nodes (airports) to identify flight delay.

**Nodes (Circle)** size indicates airports centrality, and the colour scale of nodes indicates the delay percentage of the airports. Interaction with airport node (Hover) on graph to highlight network associated with airport alongside a brief summary.



**Edges** (Airlines route) in different weights affected by number of flights alongside custom edge attribute of delay percentage. The red colour of path for delays.

**Figure 3.2.2: Interactive Flight Network (Edges)**

**Analysis:** The effect of **9/11** can be identified by the significant reduction of airline delay nationally in 2001, and the delay performance of some particular airport and airlines gradually recovered post 911 till 2003. The effect of **Y2K Bug** shows significant delay increase in 2000 compared to years before the Y2K, and the impact has been relieved in 2001 and 2002.

**Pros:**

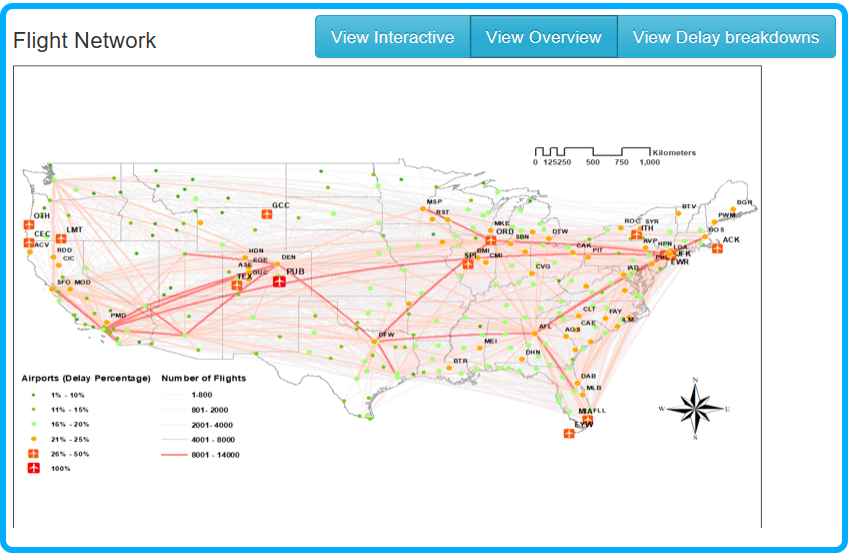
▪ User interactive feature allow users to select year and airport for visualisation

▪ Highlighting node provides additional information about airport

**Cons:**

▪ Visualisation get bit messy if try to find flight delay for states away from USA main land like Alaska, Hawaii.

### Visualisation Section 1: Flight Network-View Overview



**Figure 3.2.2: Geo-Special Layout of USA Airlines In 2008**

**Purpose:** Creating Geo-spatial layout in **ArcGIS** which will represent airports and routes which cause most delay in entire network and give overview visualization to deliver the specific airports or routes to avoid while travelling

**Nodes**: Identified in different colours and sizes (signs) by using airport delay percentage, the higher the percentage of delay, the larger nodes size. In this visualisation airport has 100% delay only has two airlines.

**Edges**: Representing 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. The colour of the airlines by using transparent 40%. The red lines show the number of flights in this airline more than 8000.

**Pros:**

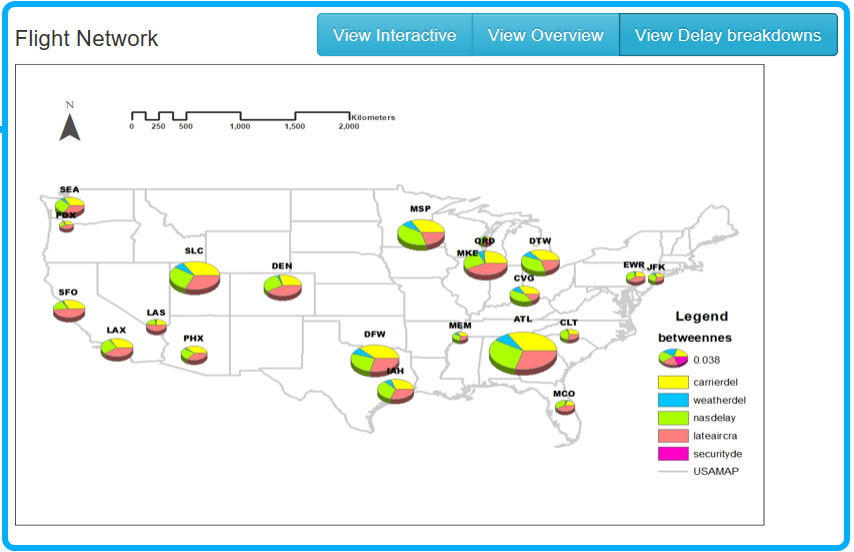
▪ Simple visualisation provide airports and routes have most delay for entire network

▪ Key information provided with distance, number of flights by size of edges

**Cons:**

▪ Static visualisation, no interactive feature for a user

### Visualisation Section 1: Flight Network-View Delay Breakdowns



**Figure 3.2.3: Delay Reasons for High Betweenness Airports In 2008**

**Purpose:** By using **ArcGIS** created Geo-spatial layout combined with pie charts which will represent airports and the percentage of different delay reasons: Carrier delay, Weather delay, NAS delay, Security delay and Late Aircraft delay. This provides traveller information about main reasons for flight delay and percentage delay mix.

**Nodes**: By using betweenness centrality analysis to choose the airports with high betweenness values, and identified these airports in different size of nodes (Pie chart).

The pie chart of each node shows the percentages of different delay reasons.

**Analysis:** It is clear to conclude from the figure that the ATL (William B Hartsfield-Atlanta International Airport) is the most critical airport to the overall flight network, and the **main reasons of delay** in **ATL** caused by carried delay, late aircraft delay and weather delay. The weather delay reason has high proportion in the centre of the continent but less in costal cites.

**Pros:**

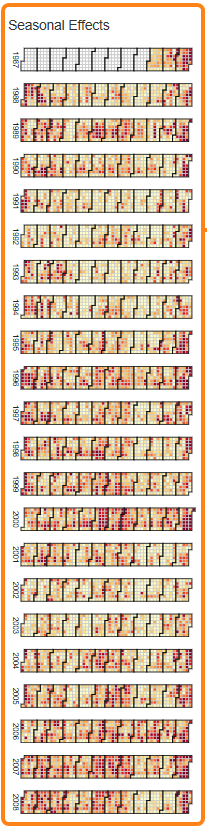
▪ Shows proportions of each delay reason at airport

**Cons:**

▪ Visualisation doesn’t show delay reasons information for all airports

▪ Not interactive visualisation with users – Visualisation shows only one year data for each map.

### Visualisation Section 2: Seasonal Effect



**Purpose:** Calendar Heat Map view of flight delay created in **D3** in order to visualise overall flight delay by day/month/year over 21 years to show which is best and worst time for travel in USA.

**Analysis:** Each day represented by squares and **heat map** technique will be used to depict flight delay percentage for the day. This visualization is showing a heat map of percentage of delays over 21 years from the year of 1987 to 2008 and developed in D3. Each day is represented by little square, each square is filled with different colour to distinguish the delays between each day. The darker the colour represents more delay; similarly the lighter colour represents less delay for that day.

It also gives some useful hints for viewers, e.g. it can be easily seen that usually the **September and the October** months have light colours meaning these two months are **best** two months **for travelling** throughout the whole year. Similarly Dec and Jan worst time to fly as have dark colours almost every day.

**Pros:**

▪ Easy to understand for user

▪ Shows entire 21 years of data and trend of flight delays

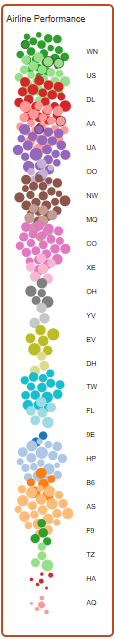
**Cons:**

▪ Not interactive visualisation

▪ Can’t get details of every day exact delay percentage and delay reasons;

**Figure 3.2.4: Seasonal Effect Calendar Heat Map**

### Visualisation Section 3: Airline Performance



**Purpose:** The interactive bubble chart (Force Layout Multiples) created in D3 aimed to visualize the flight delay performance comparison between different airlines. It helps traveller which **airline to avoid** while travelling.

**Nodes:** The nodes (circle) size influenced by the delay percentage of the airline, and each node represent one year of particular carrier, and each cluster of circles represent one airline company. Top 20 airlines performance shown from 1987 to 2008.The number of circle represents the number of delay years. Interaction with each node on graph to associated with a brief summary include

▪ Carrier

▪ Year

▪ Total Flights

▪ Number of Delays

▪ Delay Percentage

**Analysis:** Force directed algorithmis used to show good quality node size graph and simple for analysis. It is clear to show the airline performance between carriers, the suggestions to travellers is better to choose the airlines which have fewer nodes (circles) and smaller node (circle) sizes, which represent less probability to have delay compared with other carriers. From visualization EV (**Atlantic Southeast Airlines**) looks **worst airline** and HA (**Hawaiian Airlines Inc**.) looks **best airline**.

**Pros:**

▪ Easy to understand for user

▪ Shown entire 21 years of data and trend of airline delays

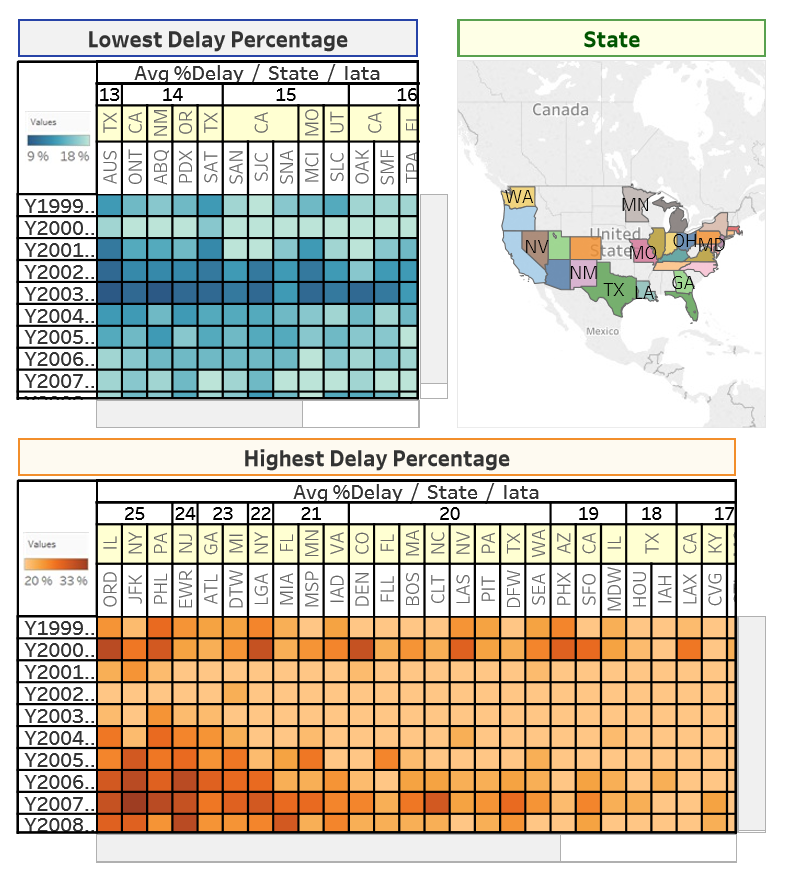
**Cons:**

▪ Can’t get detail of every day delay percentage and delay reasons

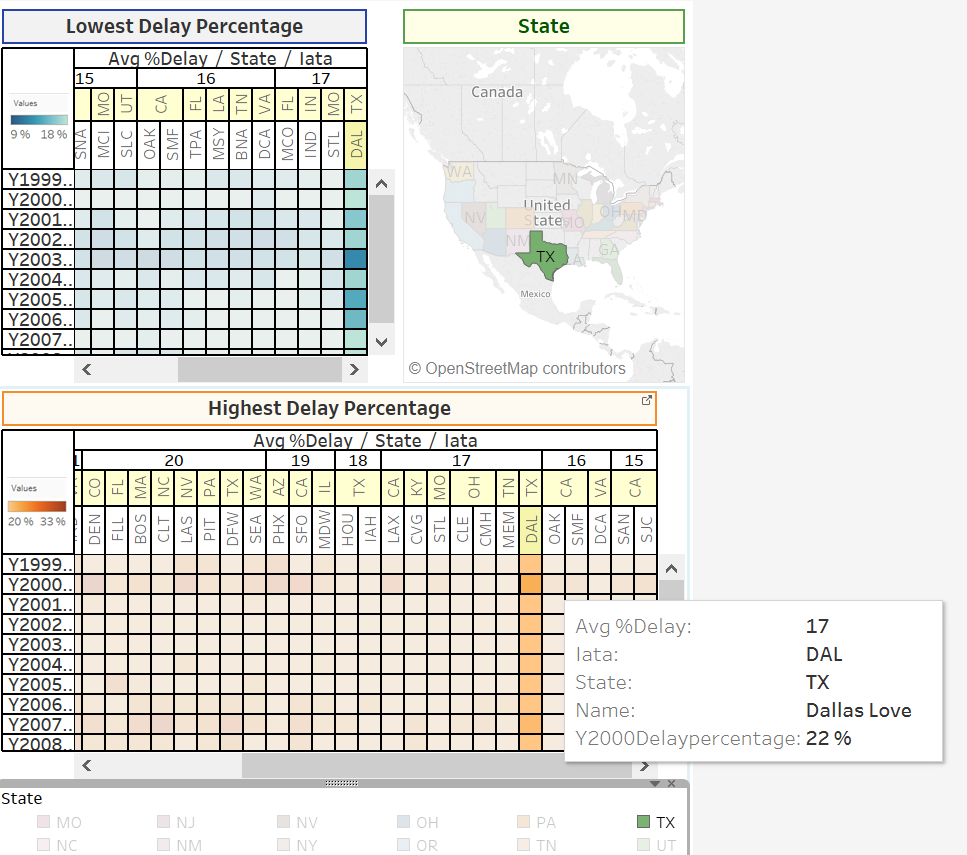
▪ Show limited number of airlines, not all

**Figure 3.2.5: Best and Worst Airline Performance Bubble Chart**

### Visualisation Section 4: Best/Worst Airport



**Figure 3.2.6: Airports Performance**



**Figure 3.2.7: Airport Detail Tooltip**

**Purpose:** A heat map visualisation representing the performance of airports from each state of the United States by its percentage of flights being delayed, showing data ranging from 1999 to 2008. This visualisation is developed in Tableau.

**Analysis:**

**Lowest Delay Percentage Heat map**: Relative performance in the number of flights delayed, between 9% to 18%, is identified as different tones of green, with a darker tone indicating a lower percentage of delay and lighter tones indicating a higher percentage of delays. Hovering the mouse over an individual box will provide detailed information regarding the airport’s home state, full name, and exact delay percentage for the selected year row. As well as showing the average delay percentage for the whole year range of 1999 to 2008 of the entire column.

**Highest Delay Percentage Heat map**: Similarly, the relative performance in percentage of delays is identified using different tones, this time representing the range between 20% and 33% and with the colour orange. Lighter tones show a lower percentage of delays and darker tones indicating higher percentage of delays. Again, same as for the Lowest Delay Percentage Heat map, hovering the mouse over an individual box will provide the airport’s home state, full name, the exact delay percentage for the year, and the average delay percentage of the entire year range.

**Airport in State of USA Map:**

The map shows every state of the US that possesses at least one airport and differentiates these states using different colours. Hovering the mouse over any one of these states will interactively highlight the same state column(s) (if found) on each heat map. Conversely hovering the mouse over any state of any of the heat maps will highlight exclusively that state on the map.

Also, highlighting a single airport of any of the heat maps will highlight the same airport column in the other heat map (if found) as well as highlight the state that airport belonged to on the map.

**Pros:**

▪ User interactive feature select any airport of top 20 to get more information

▪ Shown 10 years’ average percentage of each airport delays

▪ Shown 10 years’ average percentage of airports are grouping by state

▪ Individual box provide detailed information with airport’s home state, full name, exact delay percentage

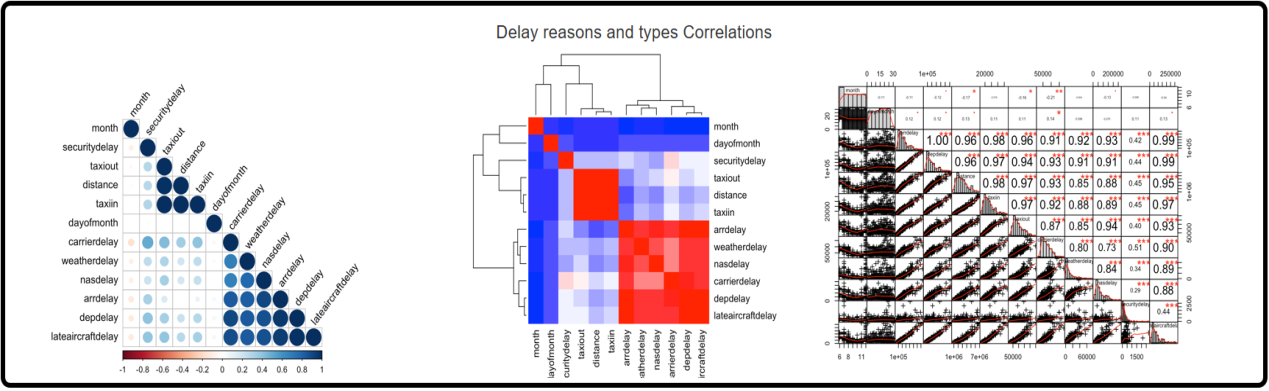
▪ Heat map of colour visualisation provide the user can easily identify for the best/worst airport

**Cons:**

▪ Can’t get detail of every day delay percentage and delay reasons

▪ Show limited number of airports

### Visualisation Section 5: Delay Statistic Analysis



**Figure 3.2.8: Delay Reasons and Correlation Matrix**

**Purpose:** By using data statistical analysis tool R (ggplot package) to visualize the correlation between each delay reasons. The purpose of this exercise is to identify and calculate the correlation figures associated with each delay reasons (% effect of reasons) and how each delay reason influence other reasons.

**Analysis:** Correlation Matrix is generated for various dimensions like whether, security, distance etc. over the predicted value of flight delay. All delays have a negative correlation with month, and late aircraft is the most correlated delay type within these reasons.

**Pros:**

* Provided specific values and colour pattern to show the correlations between reasons;

**Cons:**

* The results of the visualization have limited value for trip advising

# Implementation

The whole project implemented by the following steps:

**Step 1. Extract & Load of data**

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

**Step 2. Cleanse data & Create Indexes**

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

**Step 3. Develop & test queries for each visualisation**

1. Develop queries required for each visualisation.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

1. Get additional calculated/transformed data (e.g. % delay, performance percentage etc.)
2. Test and refine queries to get correct outputs
3. 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**

1. Carry out network analysis in NetworkX (e.g. centrality calculations) where applicable.
2. Develop code or use tool(s) to generate visualisations (e.g. JavaScript for D3).
3. Generate an overview of the visualisation of complete data
4. Reduce, refine and iterate through visualisations until satisfactory results are produced.
5. Write code or use tool options to scale colour, size, text, resolution and layout of visualisations.
6. Debug and QA code
7. QA visualisations.
8. Refine visualisations and finalise.

**Step 5. Peer review**

1. Perform peer review of each visualisation.
2. Select top 2 visualisations.

**Step 6. Evaluate and finalise visualisation**

1. Perform Evaluation for each visualisation
2. Finalize visualisations for VA system as well as final report.

# Evaluation

The table below outlines method utilised/followed for evaluation for each of the visualisation and the related tasks/aims associated.

**Table 5.1: Evaluation template of each visualisation**

|  |  |  |
| --- | --- | --- |
| **Data/ Visualization** | **Task(s)** | **Evaluation Methods** |
| **Visualization Section1:** Flight Network | ▪Identify critical routes  ▪Identify critical airports  ▪Summary of overall network | **Survey** (Questionnaire, Interview, Focused group) |
| **Visualization Section 2:** Seasonal Effect | ▪Identify month(s) for travel  ▪Identify month(s) to avoid for travel | **Survey** (Questionnaire, Interview, Focused group) |
| **Visualization Section 3**:  Airline Performance | ▪Identify Best Airline  ▪Identify Worst Airline | **Survey** (Questionnaire, Interview, Focused group)  **Empirical evaluation** (statistical analysis) |
| **Visualisation Section 4:** Best/Worst Airport | ▪Identify best Airport  ▪Identify Worst Airport | **Survey** (Questionnaire, Interview, Focused group)  **Empirical evaluation** (statistical analysis) |
| **Visualisation Section 5:**  Delay Statistic Analysis | ▪Identify main reasons for flight delay | **Survey**(Questionnaire, Interview, Focused group)  **Empirical evaluation** (statistical analysis) |

## Results

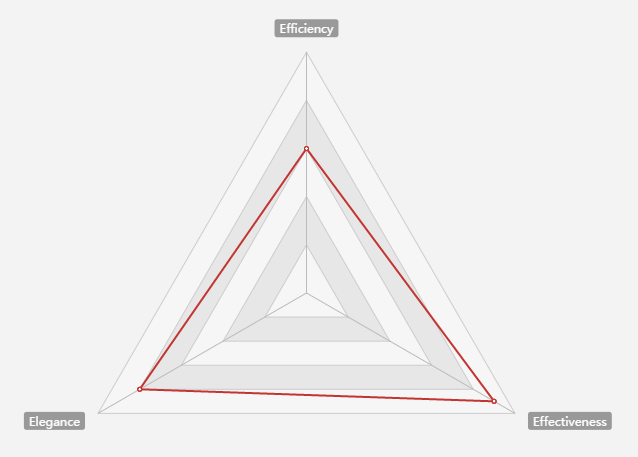
### Focused group

Team have been performed evaluation on each visualisation and following are results-

**Flight Network-View Overview**

The running time of algorithms is bit slower than other visualisation. It contains the networks of each airports / airlines by flight delay for year from 1987 to 2008 and instructiveness of visualisation makes it more effective.

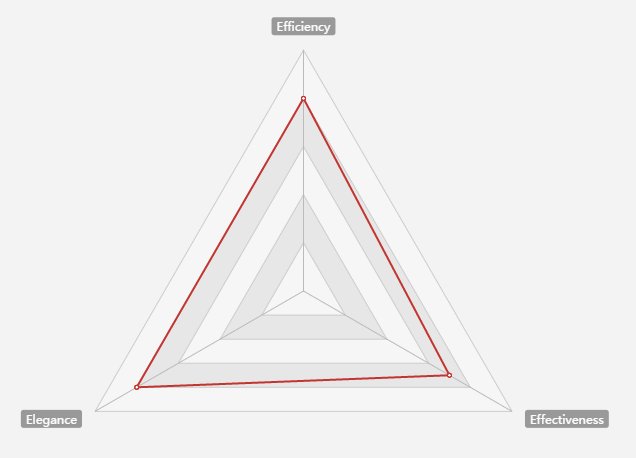
Figure.5.1.1 **Flight Network-View Interactive Evaluation**



**Flight Network-View Overview**

Visualisation efficiency is excellent.The running time of algorithms should be reasonably fast. The final drawing is beautiful. Most people firstly saw this visualization, they say the visualization is easy to understand, but in terms of effectiveness it is lacking as it gives only 1 year information and doesn’t show all airports.

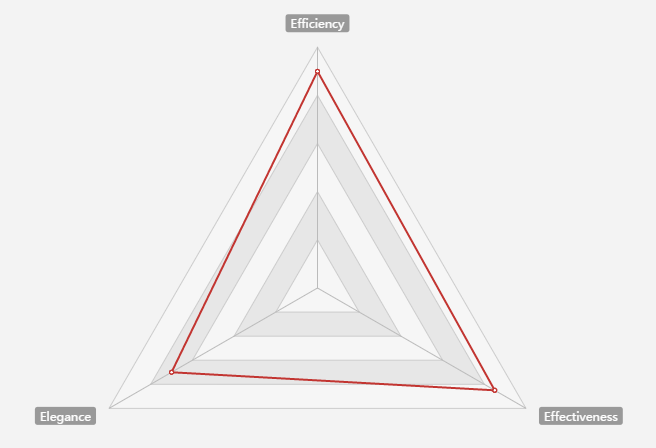
Figure.5.1.2 **Flight Network-View Overview Evaluation**



**Seasonal Effect – Heat Map**

Heat map efficiency is very good and visualisation very clearly identify flight delay pattern for 2 decades. Due to calendar view visualisation is not prettier.

Figure.5.1.3 **Seasonal Effect – Heat Map**

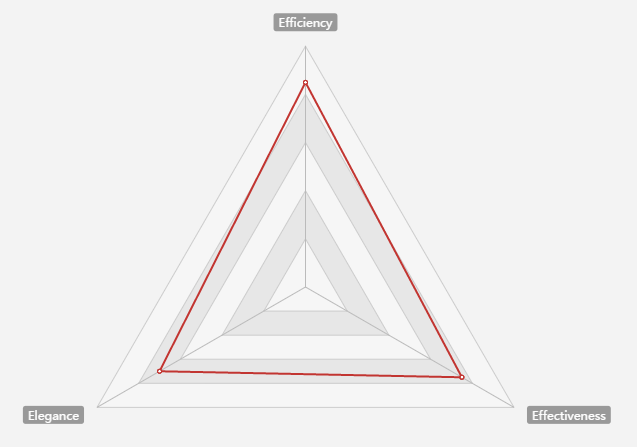


**Best and Worst Airport – Heat Map**

This Visualisation of best or worst airports performance is visualized in Tableau.

The method to use different colors is a very good way to show the difference, the running time of algorithms is reasonably fast. The final drawing is beautiful and most people can catch the point of the visualization.

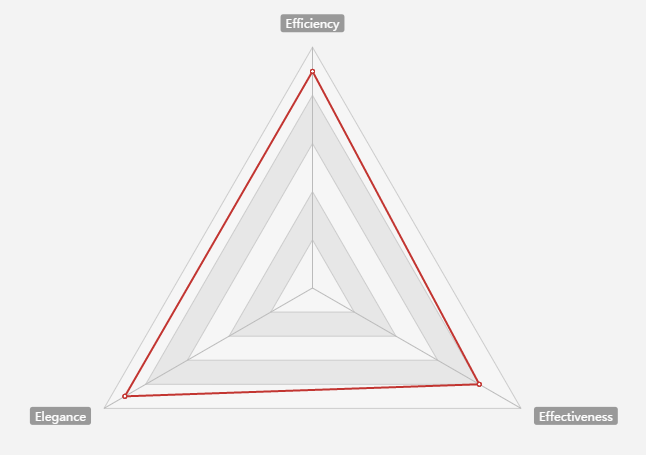
Figure.5.1.4**Best and Worst Airport – Heat Map**



**Airlines Performance – Bubble Charts**

Visualisation shows the best or worst performance airline or carriers. Due to bubble charts, the final drawing is very beautiful. Visualisation efficiency is good due to Force Directed algorithm. Audience seeing this visualisation first time raises many questions as airline code is given as label but due to interactive feature visualization is easy to understand.

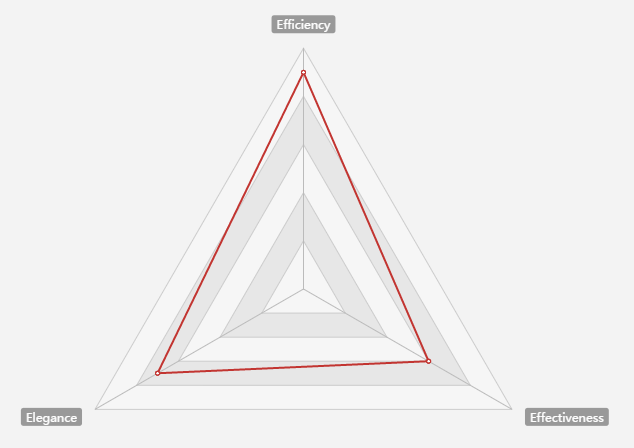
Figure.5.1.5**Airlines Performance bubble Charts**



**Delay Reasons and types of Correlations**

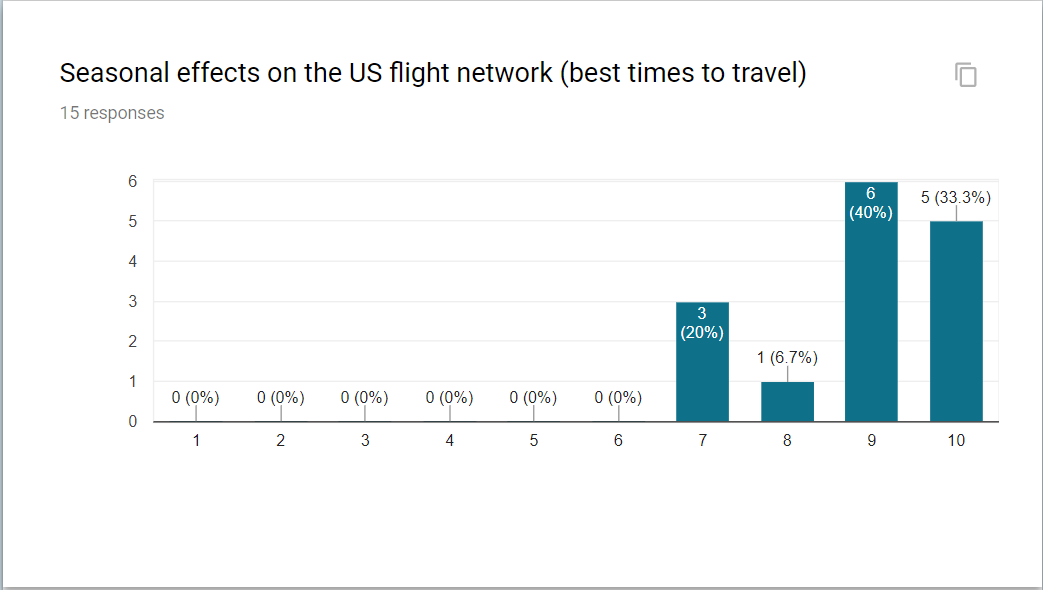
The visualization is very impressive and very efficient. However, it is not so easy to understand by most people at first glance. It shows the relationship between delay reasons and the overall flight delays.

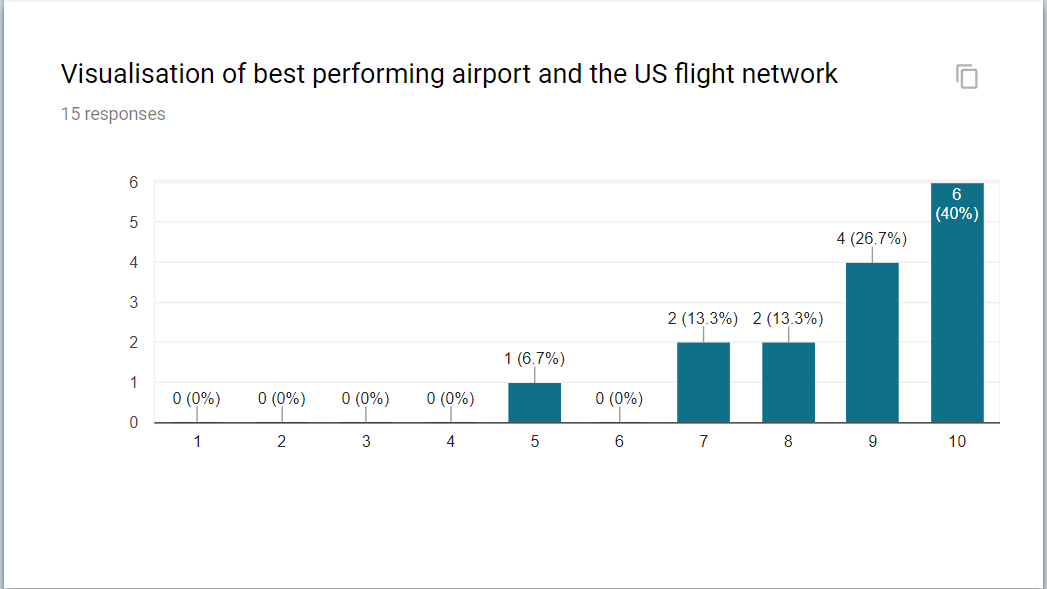
Figure.5.1.6 **Delay Reasons and types of Correlations**

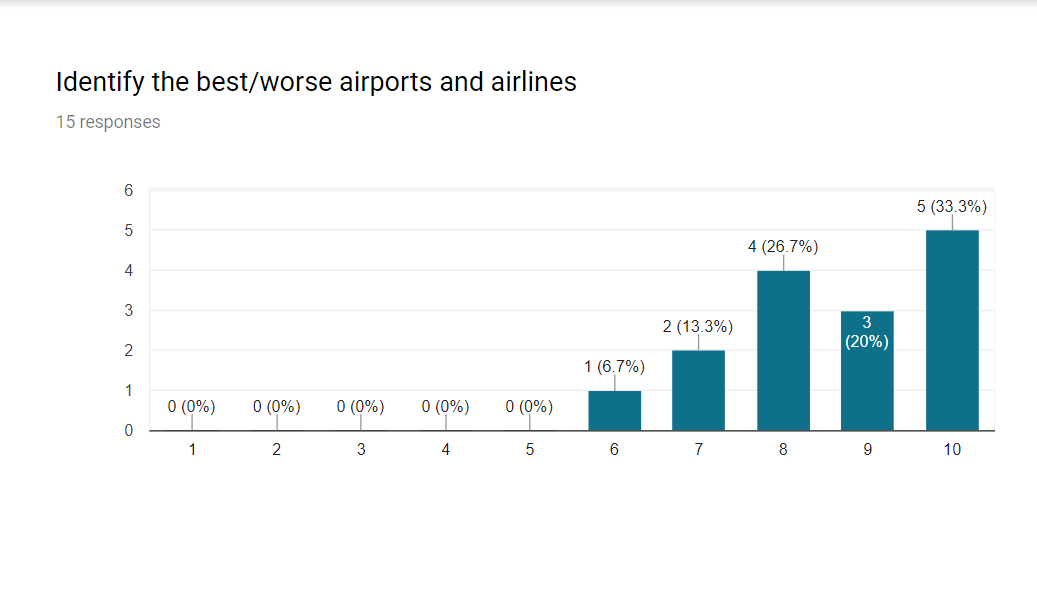


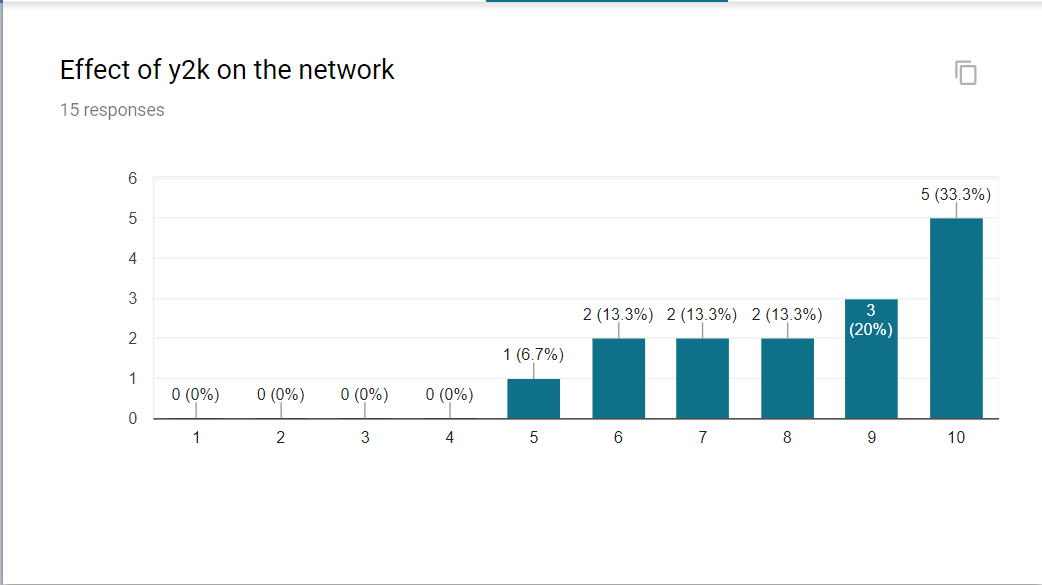
### Survey (Questionnaire) Results

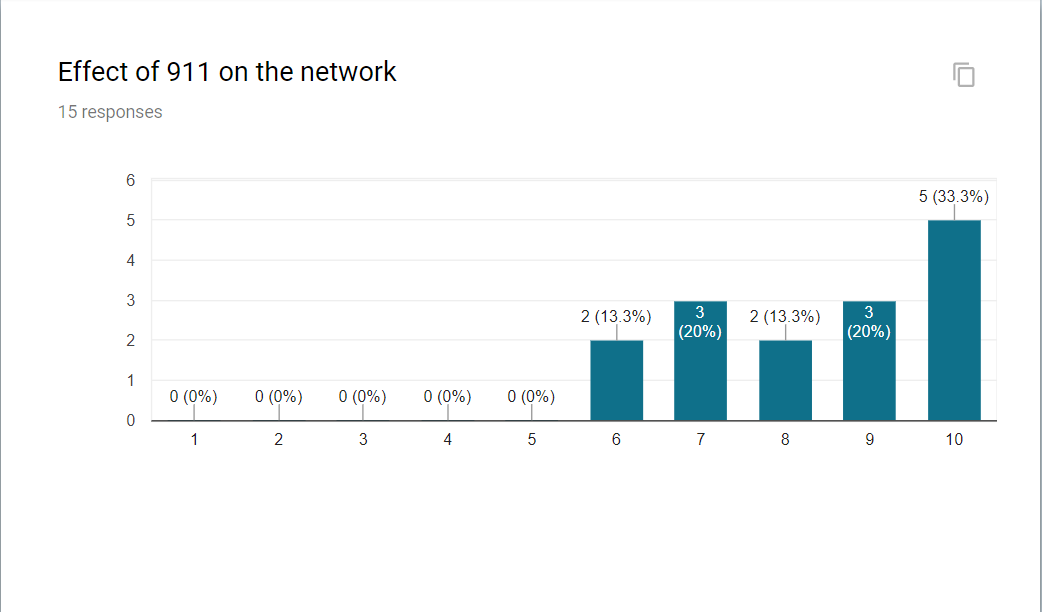
15 responses have been received on VA system specific questions/tasks, following is summary of response received, and details of survey format including questions/task are attached in appendix.

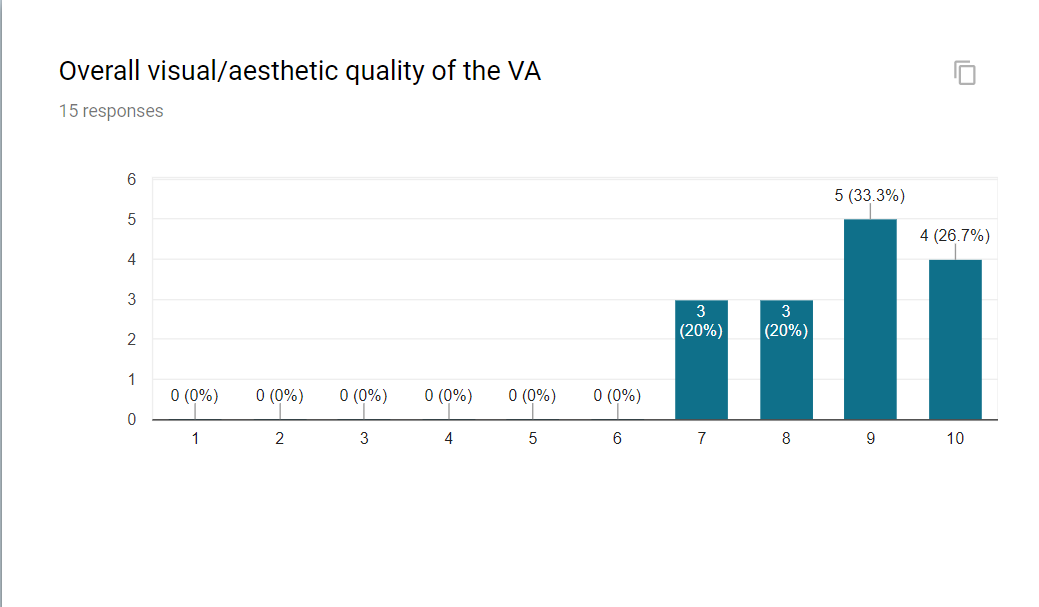
****

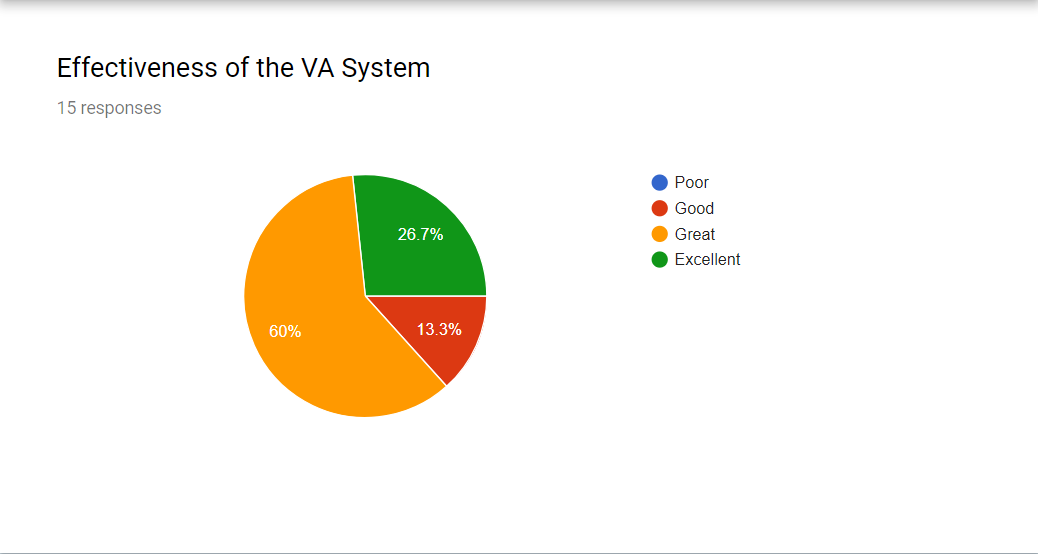
****

****

****

****

****

****

### Survey (Interview)

**Feedback on Overall VA system, details about interview is provided in appendix section.**

1. User interaction with VA system is very smooth
2. Not so clear at first glance, need explanation on visualisation from expert
3. Single page view of many visualisations
4. User can get information from VA system pertaining to airports/airlines/delay reasons

### Empirical evaluation (statistical analysis)

Statistical analysis has been performed using **Python** code to prove visualization results are accurate.

1. **Best and Worst Airlines / Carriers :**

|  |  |
| --- | --- |
| **Statistical Analysis** | Best and Worst Airlines from Visualization |
| **Null hypothesis (H0)** | The mean departure delays are the same for 2 airlines **Atlantic Southeast Airlines** and **Hawaiian Airlines** Inc. |
| **Alternative hypothesis** | The mean departure delays are different for 2 airlines **Atlantic Southeast Airlines** and **Hawaiian Airlines** Inc.(Mean departure delay for Atlantic Southeast Airlines is much higher than Hawaiian Airlines Inc.) |
| **Set α** | Significance level at 5% here. |
| **Compute p-value** | **ANOVA** : P= 4.1059403272e-39 **Kruskall-Wallis H-test**: P= 1.20170973441e-21 |
| **Result** | We, therefore, have sufficient evidence to reject the null hypothesis. Our initial guess that a statistically significant difference existed in the flight delay between airlines was backed by this statistical analysis. We have evidence to suggest that departure delay is related to carrier. |

1. **Best and Worst Airport :**

|  |  |
| --- | --- |
| **Statistical Analysis** | Best and Worst Airport from Visualization |
| **Null hypothesis (H0)** | The mean departure delays are the same for 2 airport **ORD (IL)** and **AUS (TX)** |
| **Alternative hypothesis** | The mean departure delays are different for 2 airport **ORD (IL)** and **AUS (TX)** (Mean departure delay for **ORD (IL)** is much higher than **AUS (TX)** |
| **Set α** | Significance level at 5% here. |
| **Compute p-value** | **ANOVA:** P= 2.4974769606e-07  **. . .**  **Kruskall-Wallis H-test :** P= 2.49809315473e-13 |
| **Result** | We, therefore, have sufficient evidence to reject the null hypothesis. Our initial guess that a statistically significant difference existed in the flight delay between airports was backed by this statistical analysis. We have evidence to suggest that departure delay is related to Airport. |

1. **Delay Reasons :**

|  |  |
| --- | --- |
| **Statistical Analysis** | Delay Reasons Visualisation |
| **Null hypothesis (H0)** | The departure delays have same effect for following delay reasons  ▪ Weather delay  ▪ Nas delay |
| **Alternative hypothesis** | The departure delays are different for 2 delay reasons  Mean departure delay for reason Nas delay is much higher than Weather delay |
| **Set α** | Significance level at 5% here. |
| **Compute p-value** | **ANOVA:** P= 2.27958149583e-09  **. . .**  **Kruskall-Wallis H-test:** P= 3.0630118343e-17 |
| **Result** | We, therefore, have sufficient evidence to reject the null hypothesis. Our initial guess that a statistically significant difference existed in the delay reasons was backed by this statistical analysis. We have evidence to suggest that departure delay is varies based on delay reasons. |

## Discussion

After Evaluation, following are discussion outcome -

1. Team have perform evaluation on various methods
2. Evaluation – Interview and survey taken positively by audience and most of people filled form after explanation why this survey, Received positive feedback.
3. Audience have given positive feedback about usability of VA system
4. Interview Audience asked to put more text in VA system to more clarity
5. At first glance new audience find hard to get information from VA system

Limitation of Evaluation for VA system **A Hitchhikers Guide to U.S.A.:**

1. Can’t give access to VA system to everyone, this is major Evaluation constraint.
2. This is U.S.A. flight network evaluation but performed survey and interview with Australian which might not give accurate evaluation input
3. Survey only limited to 15 people which need to increase but requires ethics approval and other approval
4. Target audience of survey may not be frequent flyer in USA
5. Evaluation may not accurate as data we have is from 1987-2008, but may not apply for today

# Conclusion

* **A Hitchhikers Guide to U.S.A. tells U.S.A. traveller –**
* Best time of year/month for travelling in USA
* Best Airports
* Best Airlines to fly
* Airport and Airline to Avoid while travelling

* **A Hitchhikers Guide to U.S.A.** also provide inside to various reasons of flight delays

* **A Hitchhikers Guide to U.S.A.** also provides information about impacts of historic events like 9/11 World trade centre, Y2K bug effect on flight delays

# References

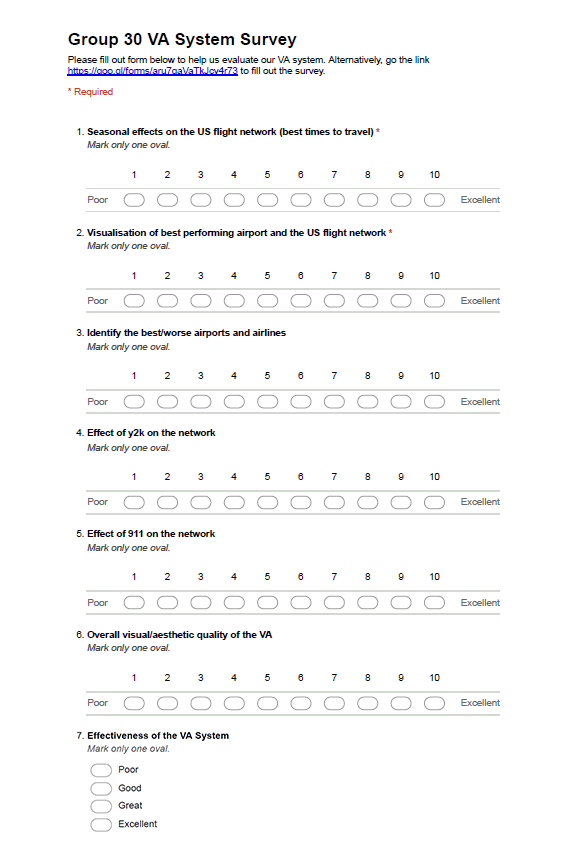
Following are website/documents referred for this assignment.

* 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 - TanujitDey • 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>

# Appendix: Group meeting minutes

**Evaluation Survey Form:**

Following are survey form created by team and distributed to get feedback from audiences. This survey form created in **Google form** to get digital response(s).



**Evaluation Interview Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name:** | Shuhao | **Interviewer:** | Alan |
| **Occupation:** | Student | **Time:** | 19/10/2017 |
| **Background:** | | Irrelevant users | |
| **Interview content:**  Hello. My name’s shuhao. I’m student.  Today, I attend this Interview for Group 30’s visualizations.  At first, when I read the overall visualization, the graph is very clear and informative. It involves plenty of information about flight line and delay reasons, something like that.  According to the first visualization, I can easily figure out when is the best month to go travelling actually. Each day is represented by squares and shows flight delay for the day.  Two maps are very attracting. Followed by that, what I can call that? Bubble chart. Actually, I don’t see that kind of chart before. The appearance is very new. | | | |
| **Suggestion:** At last, I may give some suggestions. That is, the visualization is tough and complex. If you could simply it, it could be better. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name:** | Rookit | **Interviewer:** | Alan |
| **Occupation:** | Student | **Time:** | 19/10/2017 |
| **Background:** | | Irrelevant users | |
| My name’s Rookit.  Lot of information on single page. After explaining about each graph, I can find when airline or airport could be busy. I like that graph, which show the different sizes of the airports and the plenty of connects between each airport. Different colour was used to show the different busy level.  However, I can’t understand that graph with spots clearly. Then after he told me what it is. I can understand it almost. | | | |
| **Suggestion:** The visualization could be excellent. If you guy can add more explanation, it could be better. | | | |

**Meeting Minutes:**

|  |  |
| --- | --- |
| **Assignment 2 Kick off Meeting Minutes-Week7** | |
| Date | 14September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |  |
| --- | --- | --- |
| **Agenda** | * Introduce Group Members * Nominate Roll * Select Data Sets * Discuss software, tool, method.. * Further planning and goal setting | |
| **Outcome** | * **Suggestions:**  1. Choose Flights data set (flight delay data):    * Loaded data set in SQL Data base    * Analysis of the data 2. Come up with 2 Questions from each member    * Due by 5:00 pm 16 Sep 2017 (Sat) 3. Method of communication using WhatsApp 4. Share information using github.com | |
| **Actions** | * Load data set to Data base   + Nagib * Try to ArcGIS   + Langley * Meeting Minutes   + Young * Data analyse   + everyone | |
| **Conclusion(s)** | We have very short time so we have the keep the momentum which is essential. | |
| **Assignment 2 Meeting Minutes-Week7** | | |
| Date | | 16September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Select Proposal Questions * Discuss Data set * Further planning and goal setting |
| **Outcome** | * **Suggestions:**  1. Come up with Paper Drawing for Frame from each member 2. Next meeting time    * 18 Sep 2017 (Mon) at 9:30 pm  * **Confirmed Proposal Questions:**  1. What is the best time for travelling? (Seasonal?, weekdays/weekends?) 2. Time-lapse series of graph showing performance of airports (year to year)    * Find the worse performing airport?    * Find the worse performing airline?    * City/airport performance for a single year (heat map)    * Effect of the airline/airport delay and performance during 1999 – 2000 (y2k bug) 3. Find the best/worst performing airline/carrier |
| **Actions** | * Exporting from database into CSV format   + Nagib, Girishkumar * Visualise using Python/Gephi/yEd   + Nagib * Visualise using ArcGIS   + Langley * Visualise using Calendar View   + Rachel * Visualise using ECharts   + Alan * Visualise using Tableau   + Young * Meeting Minutes   + Young * Draft initial report   + Girishkumar * Data analysis based on confirmed proposals   + everyone |
| **Conclusion(s)** | We had kept the momentum and were very productive. |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week8** | |
| Date | 18September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Discuss Drafted Design and Visualisation * Finalise Approach for Design and visualisation |
| **Outcome** | * **Discussion:**  1. Discussed Following are few design and visualisation  * Best Time to Fly - Wheel Lay-Out By Nagib * Worst Performing Airport – Heat Map by Alan * Airline On time Performance – Young  1. Next meeting time    * 21 Sep 2017 (Mon) at 9:00 pm |
| **Actions** | * Generate visualisation for each of questions along with analysis technique and visualisation lay-out   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young * finalise best design and visualisation out of all – Whole team by discussion |
| **Conclusion(s)** | We are ahead of game, Keep Momentum same. |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week8** | |
| Date | 21 September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Show and describe everyone’s own Paper Drawing for Frame * Discuss VA diagram and Visualization * Assign tasks for Design and visualization |
| **Outcome** | * **Suggestions:**  1. Assign each one’ work and show the faced question in next meeting  * **Discussion:**  1. Discussed and determined the data, tip &summary, tools and frame 2. Discussed whether following VA visualizations are proper design  * The overview map(with top 10 airports)—d3 * The map-ArcGIS * Daily heat map(Calendar)-d3 * Bubble chart(with top 10 airlines) * Coordination matrix(with security, weather and so on) * GGPlot-R * Bar chart-tableau * Line chart-tableau |
| **Actions** | * Discuss and choose the proper visualisation in the frame design * Complete best frame design and assign visualisation out of all   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young |
| **Conclusion(s)** | Everything is perfect. Fighting. |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week8** | |
| Date | 26 September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Review Group30 Initial Report * Discuss VA diagram and Visualization * Signed up Cover Sheet |
| **Outcome** | * **Suggestions:**  1. Final review Initial Report before next meeting (27/9/2017 at 9:30 pm)  * **Discussion:**  1. Review and discussion following VA visualizationsdesign  * The overview map(with top 10 airports)—d3 * The map-ArcGIS * Daily heat map(Calendar)-d3 * Bubble chart(with top 10 airlines) * Coordination matrix(with security, weather and so on) * GGPlot-R * Bar chart-tableau * Line chart-tableau |
| **Actions** | * Discuss and continue implement visualisation in the frame design * Final review Initial Report before summit   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young |
| **Conclusion(s)** | We are good progress and very productive. |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week8** | |
| Date | 27 September 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Review the Group30 Initial Report |
| **Outcome** | * **Discussion:**   + Discuss each section of initial report and finalise final draft |
| **Actions** | * Final review of Initial Report   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young * Submission of Initial Report   + Langley |
| **Conclusion(s)** | First assignment for initial report now completed |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week9** | |
| Date | 5October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Discuss the presentation structure |
| **Outcome** | * **Discussion:**   + Discuss each section of presentation and figure out the content |
| **Actions** | * **Suggestions:**  1. Assign each one’ work and show different parts in next meeting  * **Discussion:**  1. Discussed and determined the structure of slides 2. Determined the outline according to the requirement and added more detail  * Introduction * Design * Implementation * Evaluation * Progress * Planning |
| **Conclusion(s)** | The overall presentation structure was determined |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week9** | |
| Date | 6October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Detail content of slides * Discuss the different way of the presentation |
| **Outcome** | * **Discussion and Assignment:**   + Discuss detailed method to show our presentation   + Make sure the presentation is attractive |
| **Actions** | * **Discussion:**   + Discuss introduction   + To begin with a movie or question to attract audience   + Discuss the method to show our VA system   + Assign everyone work of the presentation part |
| **Conclusion(s)** | 50% of Presentation part is completed |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week9** | |
| Date | 9October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Practice and review the presentation part |
| **Outcome** | * **Practice & Review:**   + Practice the presentation   + Review the slides |
| **Actions** | * Final review of Presentation Slides * Practice the presentation   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young * Submission of Slides   + Langley |
| **Conclusion(s)** | Presentation part now completed |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week10** | |
| Date | 12October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Discuss the final report structure |
| **Outcome** | * **Discussion:**   + Discuss each section of final report and add detail |
| **Actions** | * **Suggestions:**  1. Assign everyone’s work and show different parts in next meeting  * **Discussion:**  1. Discussed and determined the structure of final report 2. Determined the outline according to the requirement and figure out the contents  * Aims and Contribution * Data set * Design * Implementation * Evaluation * Conclusion |
| **Conclusion(s)** | Start the final report |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week11** | |
| Date | 20 Oct 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Discuss and Add detail to the final report draft |
| **Outcome** | * **Discussion:**   + Discuss and Focus on the evaluation part |
| **Actions** | * **Suggestions:**  1. Assign each one’ work about the evaluation part and give everyone’s own opinion  * **Discussion:**  1. Discussed and determined the methods of evaluation 2. Discussed the result of evaluation 3. Discussed the questions about the evaluation  * How about seasonal effects on the US flight network (best times to travel) * How about visualisation of best performing airport and the US flight network * Identify the best/worst airports and airlines * How about effect of y2k on the network * How about effect of 911 on the network * How about overall visual/aesthetic quality of the VA * How about effectiveness of the VA System |
| **Conclusion(s)** | 80% of final report now completed |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week12** | |
| Date | 27October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Review the Group30 Final Report |
| **Outcome** | * **Discussion:**   + Discuss each section of final report and finalise final draft |
| **Actions** | * Final review of Final Report of Assignment 2   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young * Submission of Final Report   + Langley |
| **Conclusion(s)** | 100% of Final report now completed |

|  |  |
| --- | --- |
| **Assignment 2 Meeting Minutes-Week12** | |
| Date | 27October 2017 |

**Present**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alan Shen | X | Girishkumar Dhotarkar | X | Nagib Shah | X |
| Qiushi (Rachel) Zhang | X | Xinan (Langley) Ma | X | Young Choi | X |

**Minutes**

|  |  |
| --- | --- |
| **Agenda** | * Review the Group30 Final Report |
| **Outcome** | * **Discussion:**   + Discuss each section of final report and finalise final draft |
| **Actions** | * Final review of Final Report of Assignment 2   + Nagib   + Girishkumar   + Langley   + Rachel   + Alan   + Young * Submission of Final Report   + Langley |
| **Conclusion(s)** | 100% of Final report now completed |

# Appendix: Code

**Instruction to Get Started:**

* Download and install anaconda (this will install python 3 and relevant packages)
* Download and install postgresql (brew install postgres)
* Download and install pgadmin
* Download all necessary flight data files from <http://stat-computing.org/dataexpo/2009/>
* Ensure pg admin is up and running
* Create a new database in postgres called "FlightData"
* Clone the repository
* Load data in postgressql (FlightData DB) by either Option 1 or Option 2

Option 1: Load Data

* + In your working directory create a folder called "FlightData" and place all necessary flight data (CSVs) (intentionally excluded from this repository due to its size) - <http://stat-computing.org/dataexpo/2009/supplemental-data.html>
  + Extract the .gz/zip files for the year 2000 as well as the flight and carrier data files.
  + start the juputer notebook (if mac open a terminal and type "jupyter notebook")
  + open the Assignment2DataLoad.ipynb and run through the code (NOTE: The process might take a while to finish depending ony your hardware)
  + Once loaded close the notebook and open the 2nd notebook (FlightDataAnalaysis.ipynb)

Option 2: Load Data from Postgresql Import/export Utility (Faster)

* + Create Tables using script given on http://stat-computing.org/dataexpo/2009/sqlite.html
  + Go to table Import/Export, select import
  + Select file name, select header yes, and press ok

To Run **A Hitchhikers Guide to U.S.A.** VA System:

* Open a terminal/cmd prompt and move to "/FlightDataVisualisation/VisualisationsD3" directory
* Start the python webserver (python -m http.server 8888)
* Open a browser and go to http://localhost:8888/f[light-vaV2.html](https://github.com/nagibshah/FlightDataVisualisation/blob/master/VisualisationsD3/flight-vaV2.html)

Code:

|  |  |  |
| --- | --- | --- |
| Main code | Purpose | Language |
| sampeQuery.sql | To get appropriate data | Sql |
| FlightDataAnalaysis.ipynb | Loading data into postgressql | Python |
| [DailyFlightHeatmap.html](https://github.com/nagibshah/FlightDataVisualisation/blob/master/VisualisationsD3/DailyFlightHeatmap.html) | Seasonal Effect heat map | D3 |
| bubbleforce.html | Airline Performance code bubble chart | D3 |
| [bubbleforce2.html](https://github.com/nagibshah/FlightDataVisualisation/blob/master/VisualisationsD3/bubbleforce2.html) | Airline Performance code bubble chart | D3 |
| [flight-va.html](https://github.com/nagibshah/FlightDataVisualisation/blob/master/VisualisationsD3/flight-va.html) | Interactive Network System | D3 |
| [flight-vaV2.html](https://github.com/nagibshah/FlightDataVisualisation/blob/master/VisualisationsD3/flight-vaV2.html) | Combined VA System Code | D3 |
| [correlationMatrix.Rmd](https://github.com/nagibshah/FlightDataVisualisation/blob/master/Rworks/correlationMatrix.Rmd) | Correlation matrix | R |
| [Best & Worst Airport Statistical Evaluation.ipynb](https://github.com/nagibshah/FlightDataVisualisation/blob/master/final%20report/Evaluation%20Statistical%20Testing/Best%20%26%20Worst%20Airport%20Statistical%20Evaluation.ipynb) | Statistical Evaluation of Airport | Python |
| Best & Worst Airline Statistical Evaluation.ipynb | Statistical Evaluation of Airline | Python |
| Delay Reason Statistical Evaluation.ipynb | Statistical Evaluation of Delay reasons | Python |