

**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

****

**29.10.2017**

**Revision History**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Revision | Date | Status | Author | Summary of Changes |
| 0.1 | 29-Oct-2017 | Completed | Girish | Initial Draft |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Table of Contents**

[1. Aim & Contributions 5](#_Toc497056825)

[1.1 Aim 5](#_Toc497056826)

[1.2 Contributions 6](#_Toc497056827)

[2. Data Set 7](#_Toc497056828)

[3. Design 9](#_Toc497056829)

[3.1 Framework of VA System - A hitchhiker’s guide to U.S.A. 9](#_Toc497056830)

[3.2 Analysis & Final Visualisation 10](#_Toc497056831)

[3.2.1 Visualisation 1: Flight Network-View Overview 10](#_Toc497056832)

[3.2.2 Visualisation 2: Flight Network-View Delay Breakdowns 11](#_Toc497056833)

[3.2.3 Visualisation 3: Flight Network-View Interactive 12](#_Toc497056834)

[3.2.4 Visualisation 4: Seasonal Effect 14](#_Toc497056835)

[3.2.5 Visualisation 5: Airline Performance 15](#_Toc497056836)

[3.2.6 Visualisation 6: Best/Worst Airport 16](#_Toc497056837)

[3.2.7 Visualisation 7: Delay reasons and types Correlations 17](#_Toc497056838)

[4. Implementation 18](#_Toc497056839)

[5. Evaluation 20](#_Toc497056840)

[5.1 Results 21](#_Toc497056841)

[5.1.1 Survey (Questionnaire) 21](#_Toc497056842)

[5.1.2 Survey (Interview) 24](#_Toc497056843)

[5.1.3 Empirical evaluation (statistical analysis) 25](#_Toc497056844)

[5.2 Discussion 27](#_Toc497056845)

[6. Conclusion 27](#_Toc497056846)

[7. References 27](#_Toc497056847)

[8. Appendix: Group meeting minutes (0.5-1 page per week: week 7-12) 28](#_Toc497056848)

[9. Appendix: Code 28](#_Toc497056849)

# 1. Aim & Contributions

## Aim

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.

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* |  |

# 2. Data Set

After analysing each data set based on criteria above, team finalized on **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.

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.

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:

Table 1.1.2: Airports data Table 1.1.3: Carriers

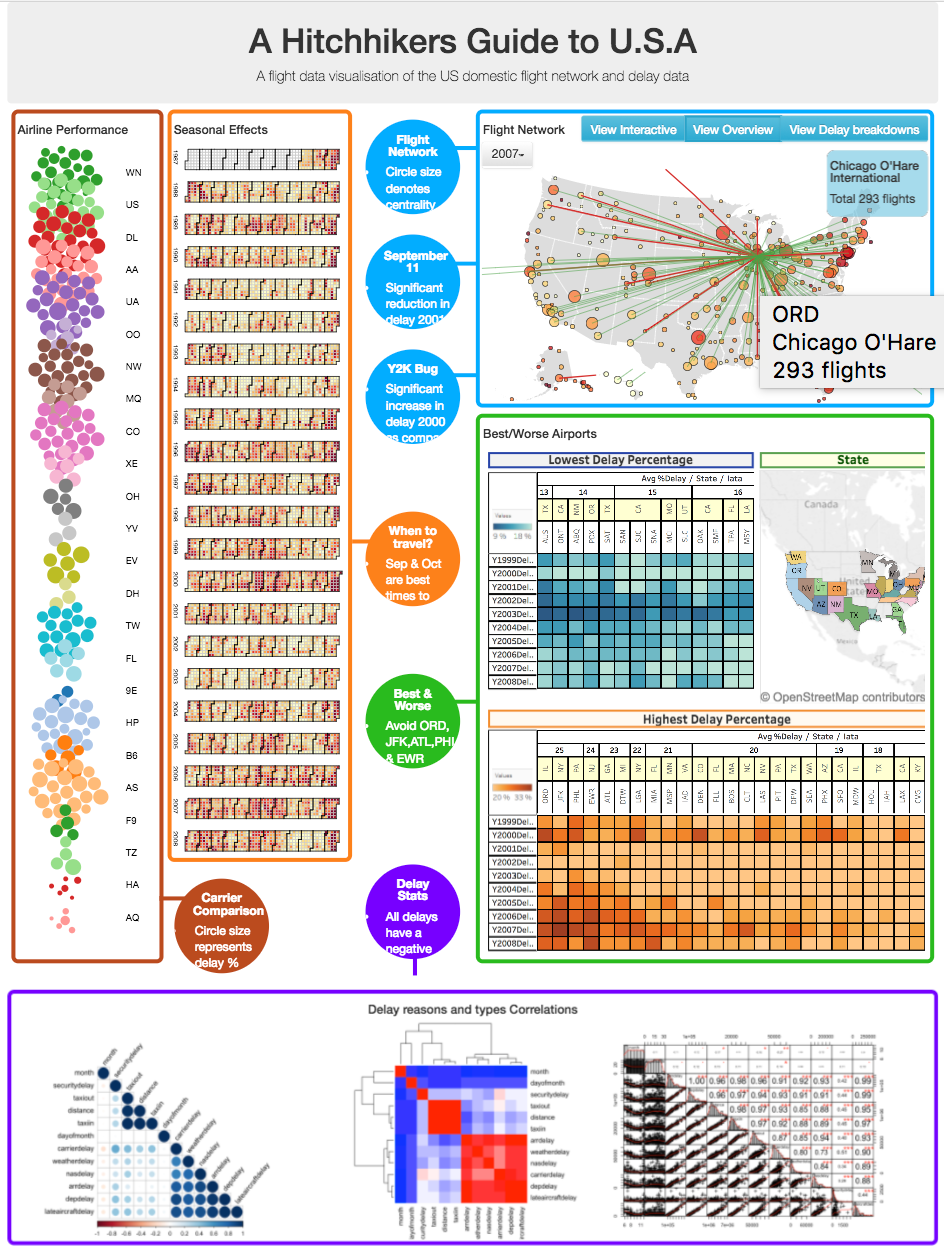
|  |  |
| --- | --- |
| **Name** | **Description** |
| iata | Iata (airport code) |
| airport | Airport Name |
| city | City Name |
| state | State Abbreviation |
| country | Country Abbreviation |
| lat | latitude |
| long | Latitude |

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

# 3. Design

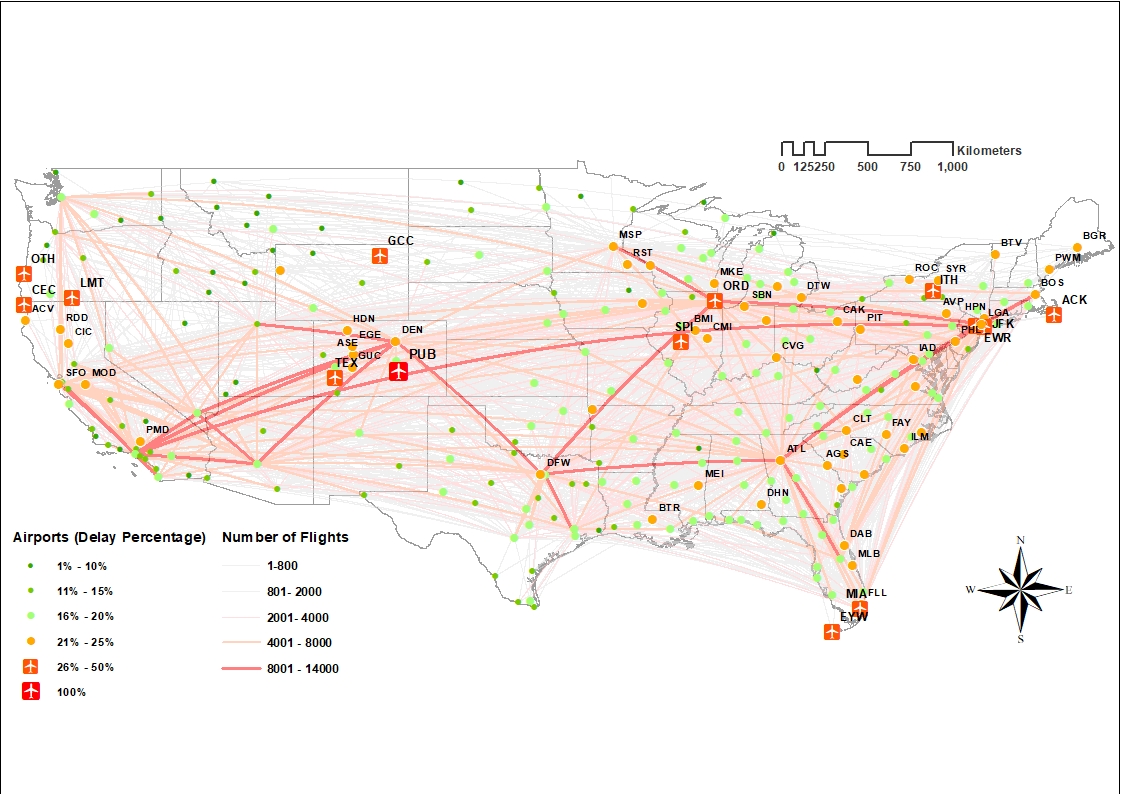
## 3.1 Framework of VA System - A hitchhiker’s guide to U.S.A.

Figure 3.1.1: VA Framework



## 3.2 Analysis & Final Visualisation

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



**Figure 3.2.1: Geo-Special Layout of USA Airlines In 2008**

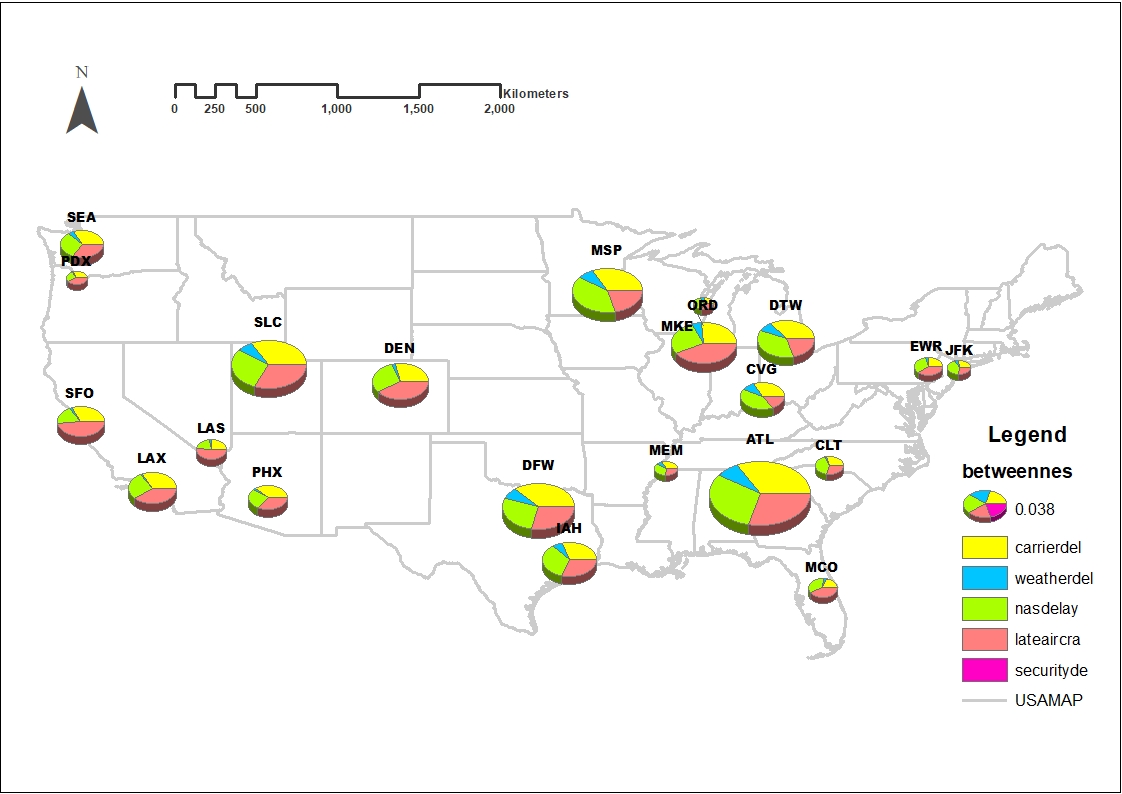
**Purpose:** Creating Geo-spatial layout in **ArcGIS** which will represent airports and airline routes which cause most delay in entire network and give overview visualization to deliver the specific airports or routes to avoid while travelling

**Node**: Identified in different colours and sizes (signs) by using airport delay percentage, the higher the percentage of delay, the larger nodes size. However, the 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 shows the number of flights in these airline more than 8000.

**Features:**

### 3.2.2 Visualisation 2: Flight Network-View Delay Breakdowns



**Figure 3.2.2: Delay Reasons for High Betweenness Airports In 2008**

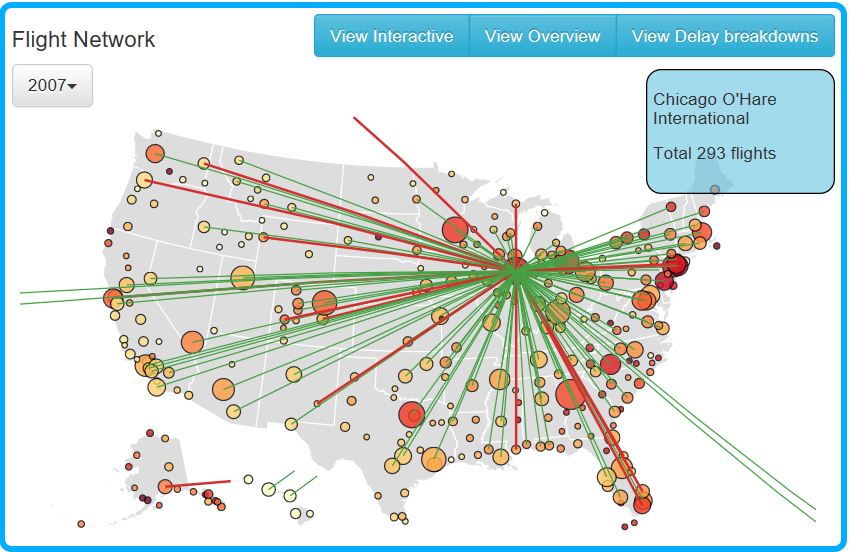
**Purpose:** By using **ArcGIS** created Geo-spatial layout combined with pie charts which will represent airports and the percentage of different delay reasons: Carrier delay, Weather delay, NAS delay, Security delay and Late Aircraft delay for higher betweenness airports.

**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 charts of each nodes 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.

### 3.2.3 Visualisation 3: Flight Network-View Interactive



**Figure 3.2.3: Interactive Visualisation to show the networks of each airports / airlines by flight delay**

**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 for higher betweenness airports.

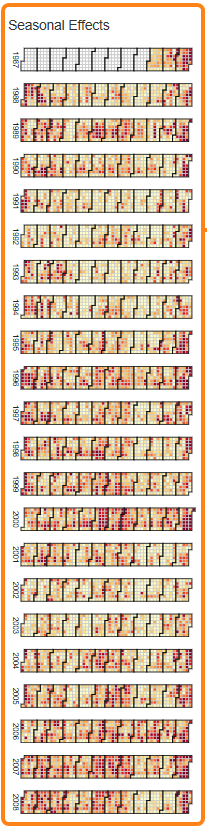
**Nodes**: By using betweenness centrality analysis to choose the airports with high betweenness values, and identified these airports in different size of nodes (Pie chart).

The pie charts of each nodes shows the percentages of different delay reasons.

**Analysis:** 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.

Software tool proposed for this visualization is D3 to cater to interactive features that are required for effective summarisation of the network.

### 3.2.4 Visualisation 4: Seasonal Effect



**Purpose:** Calendar view of flight delay to show overall flight delay by day/month/year over 21 years

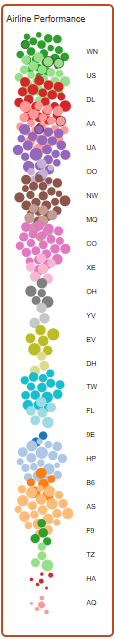
**Nodes**: Each day represented by squares and **heat map** technique will be used to depict flight delay percentage for the day. Due to the size of the dataset, heat map is considered to be the appropriate technique to represent and disseminate the delay data effectively.

Visualisation will be represented by **calendar layout**from year 1987 to 2008 and D3 is proposed as the chosen tool to perform this task.

**Analysis:** This visualization is showing a heatmap of percentage of delays（over 20 minutes）from the year of 1987 to 2008. Each day is represented by little squares, the square is filled with different colors to distinguish the delays between each day. The darker the color is, the more delays it represents, similarly the lighter the color is, less delay is represented for that day. The process to visualize the heatmap is divided into 3 parts.Firstly, I created the format of the calendar, by using SQL to extract the percentage of delay for each day and fill the format with the data, then the heatmap is completed.

The heatmap is useful for message for the time for traveling from just looking at the map, it gives an overview of the main idea we are trying to convey based on the questions we were planning to solve. It also gives some useful hints for viewers, for example,it can be easily seen that usually the September and the October are months which light colors most often display, it means that these two months are best two months for travelling throughout the whole year. Summer are bad times for flying since the dark colors are mostly displaying in this period.

### 3.2.5 Visualisation 5: Airline Performance



**Purpose:** Calendar view of flight delay to show overall flight delay by day/month/year over 21 years

**Nodes**: 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.

Visualisation will be represented by **calendar layout**from year 1987 to 2008 and D3 is proposed as the chosen tool to perform this task.

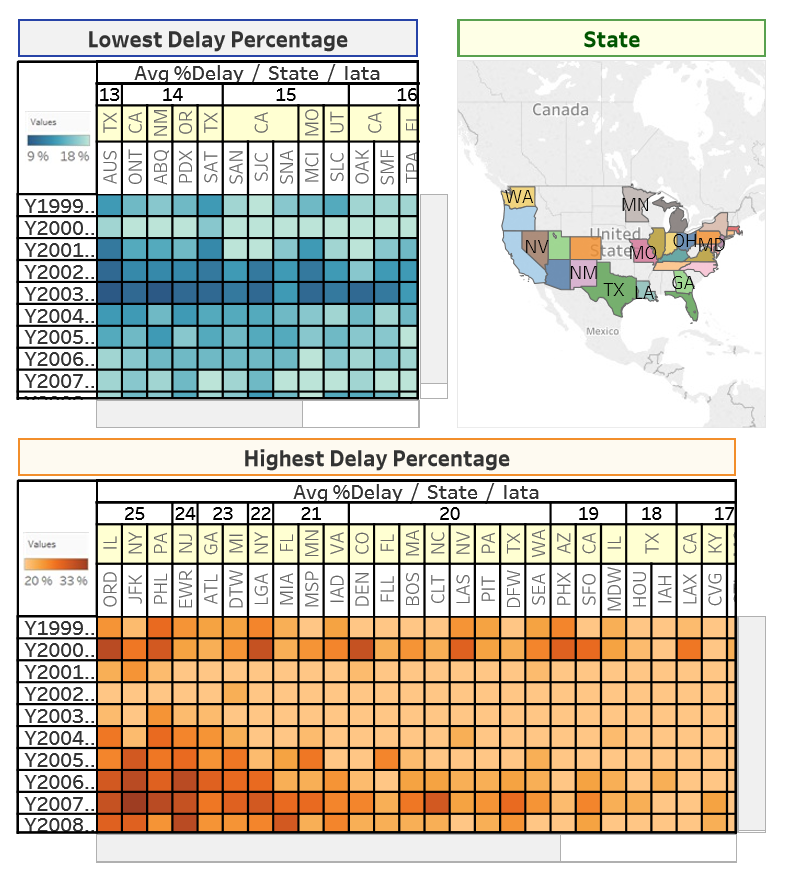
Top 10 airlines performance shown from 1987 to 2008.

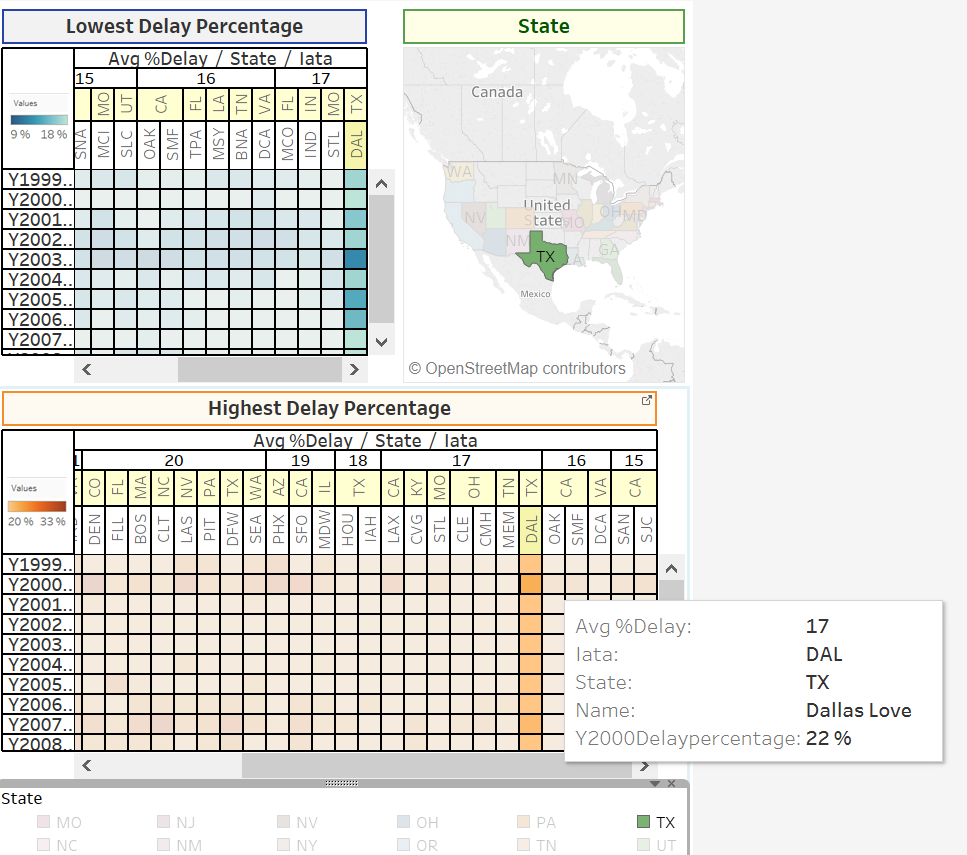
Each cluster/colour represent airline, and the circlesize to represent % delay. **Force directed algorithm**is proposed because It produces good quality result for 50–500 node size graph and simple for analysis**. D3**is proposed as the tool best for this exercise**.**

**Analysis:** Top 10 airlines performance shown from 1987 to 2008.

*Each cluster/colour represent airline, and the circlesize 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****.*** based on the questions we were planning to solve. It also gives some useful hints for viewers, for example,it can be easily seen that usually the September and the October are months which light colors most often display, it means that these two months are best two months for travelling throughout the whole year. Summer are bad times for flying since the dark colors are mostly displaying in this period.

### 3.2.6 Visualisation 6: Best/Worst Airport



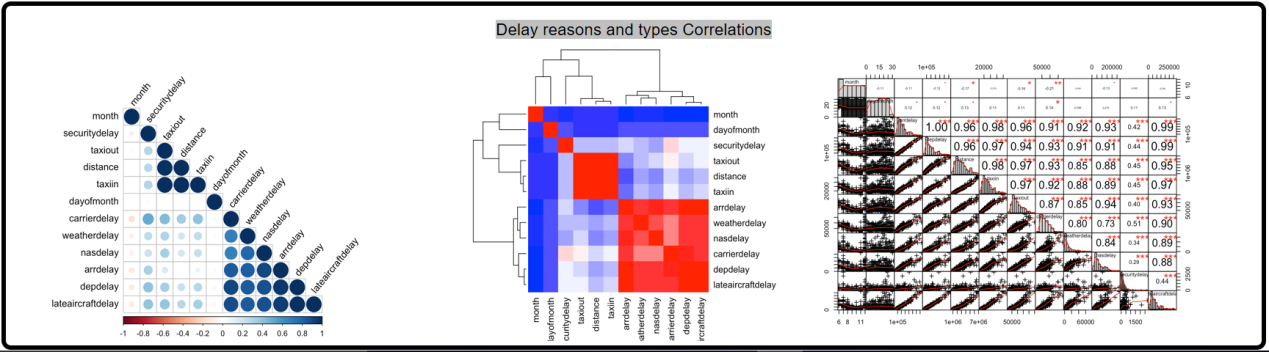
******

**Analysis:** 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**.

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.

were planning to solve. It also gives some useful hints for viewers, for example,it can be easily seen that usually the September and the October are months which light colors most often display, it means that these two months are best two months for travelling throughout the whole year. Summer are bad times for flying since the dark colors are mostly displaying in this period.

### 3.2.7 Visualisation 7: Delay reasons and types Correlations



**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 for higher betweenness airports.

**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 charts of each nodes shows the percentages of different delay reasons.

**Analysis:** 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).

**scatter plot**utilising the **ggPlot library** in **R** is proposed as the best fit tool to carry out such analysis and visualisation.

4. Implementation

Following implementation plan followed to achieve the visualisations :

**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).

|  |
| --- |
| # index samples |
|  | createindexyearon ontime(year); |
|  | createindexdateon ontime(year, month, dayofmonth); |
|  | createindexflightoriginon ontime (origin); |
|  | createindexflightdeston ontime (dest); |
|  | createindexflightarrdelayon ontime (arrdelay); |

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

1. Develop queries required for each visualisation.

|  |  |
| --- | --- |
| # e.g. delay summary all airports | |
|  | | selectf.origin, f.destas 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 |
|  | | wheref.origin=d.originandf.dest=d.destand year=2008 |
|  | | group byf.origin, f.dest, d.number\_of\_delays; |
|  |
|  |
|  |
|  |
|  |

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 scalecolour, size, text, resolution and layout of visualisations.
6. Debug and QA code
7. QA visualisations.
8. Refine visualisationsand 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.

# 5. Evaluation

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

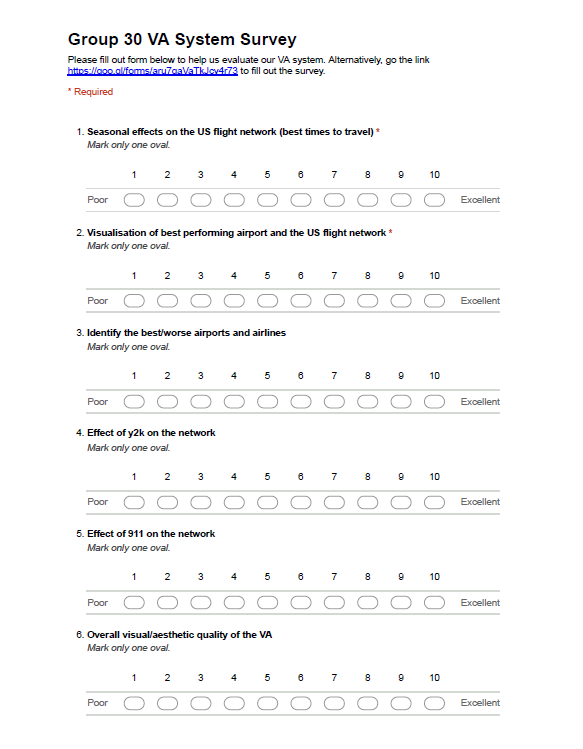
Table 5.1: Evaluation template of each visualisation

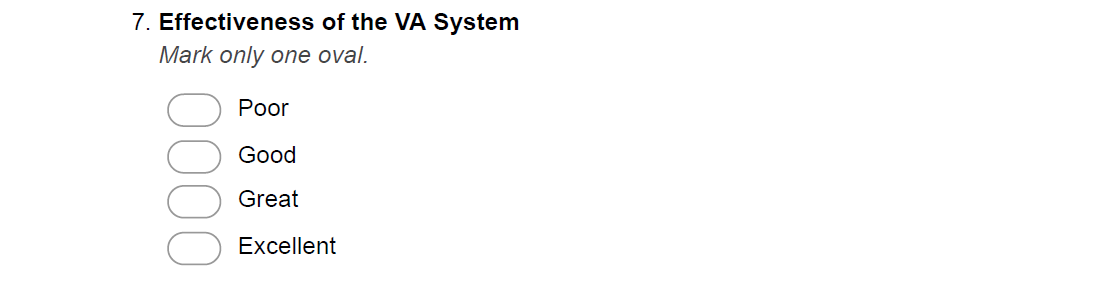
|  |  |  |
| --- | --- | --- |
| **Data / Visualization** | **Task(s)** | **Evaluation Methods** |
| **Primary Flight Network**  Visualisation 1: Flight Network-View Overview  Visualisation 2: Flight Network-View Delay Breakdowns  Visualisation 3: Flight Network-View Interactive | Identify critical routes | **Survey** (Questionnaire, Interview, Focused group) |
| Identify critical airports |
| Summary of overall network |
| Visualisation 4: Seasonal Effect | Identify month(s) for travel | **Survey** (Questionnaire, Interview, Focused group) |
| Identify month(s) to avoid for travel |
| Visualisation 5: Airline Performance | Identify Best Airline | **Survey** (Questionnaire, Interview, Focused group)  **&**  **Empirical evaluation** (statistical analysis) |
| Identify Worst Airline |
| Visualisation 6: Best/Worst Airport | Identify best Airport | **Survey** (Questionnaire, Interview, Focused group)  **&**  **Empirical evaluation** (statistical analysis) |
| Identify Worst Airport |
| Visualisation 7: Delay reasons and types Correlations | Identify main reasons for flight delay | **Survey** (Questionnaire, Interview, Focused group)  &  **Empirical evaluation** (statistical analysis) |

## 5.1 Results

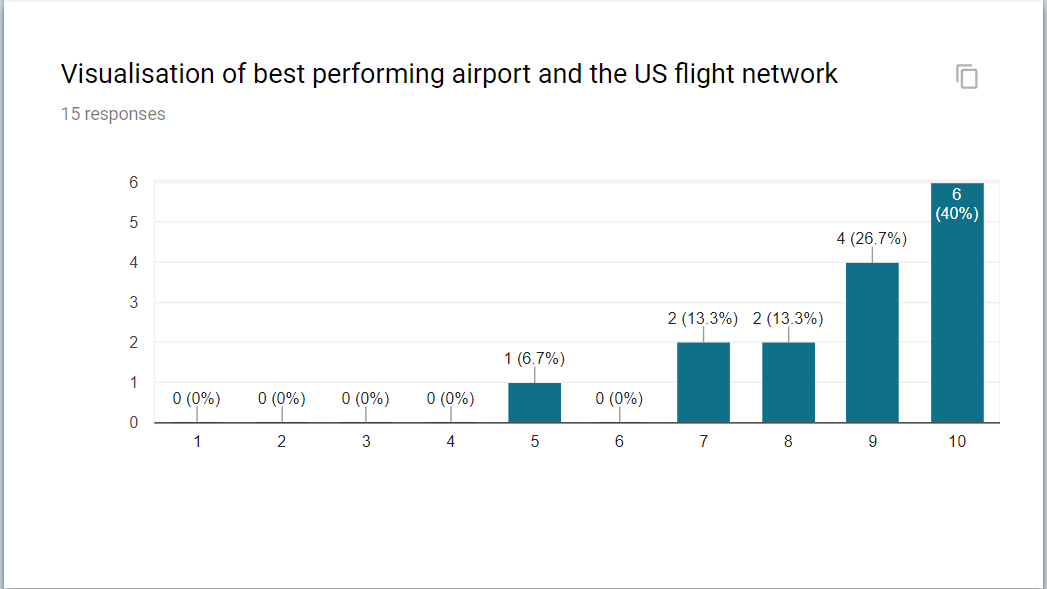
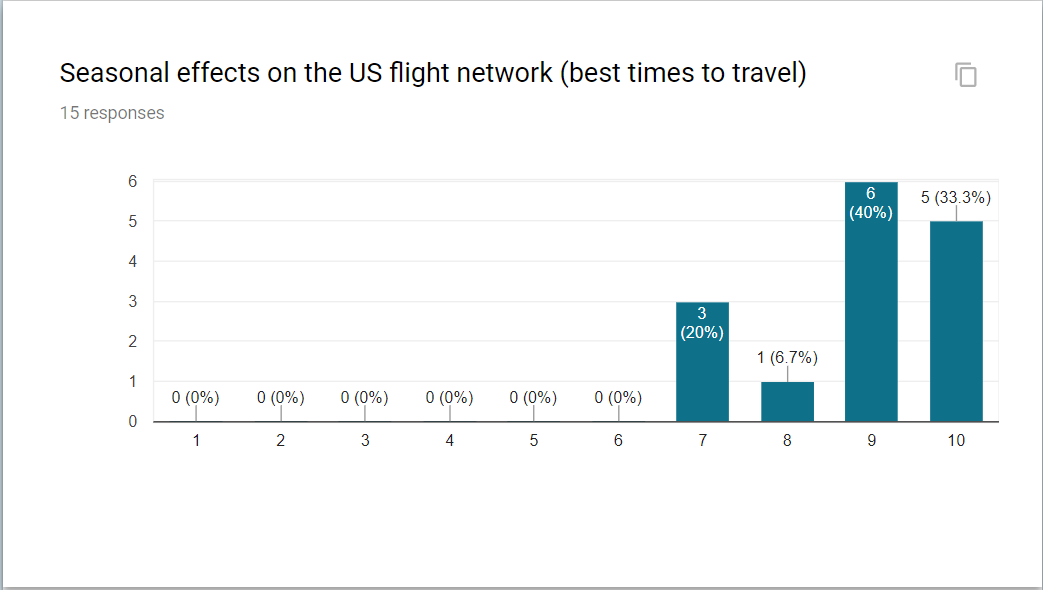
### 5.1.1 Survey (Questionnaire)

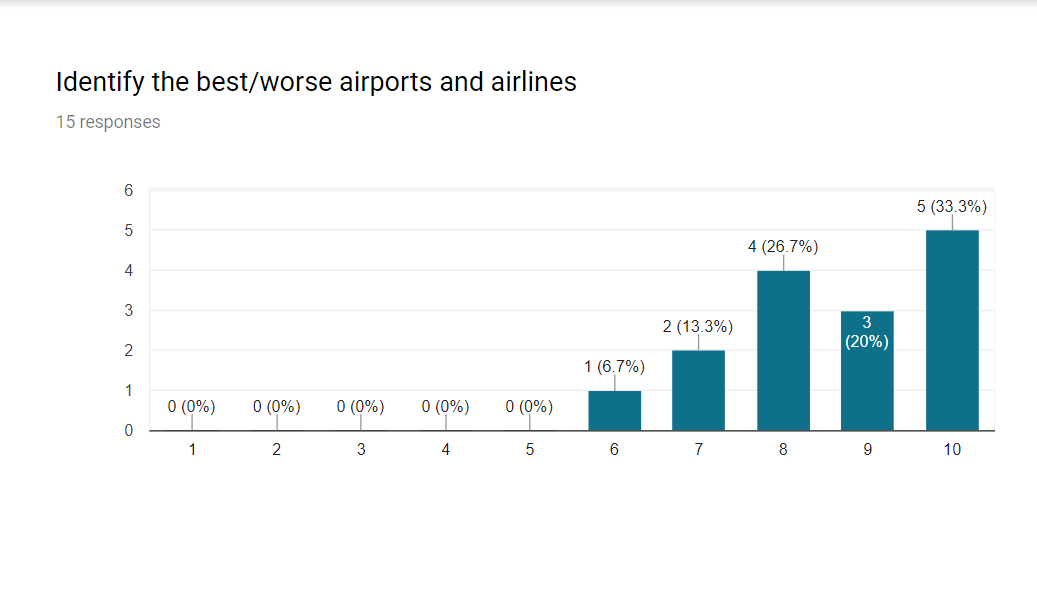
Following are survey form created by team and distributed to get feedback from classroom. This survey form created in **Google form** to get digital response(s).

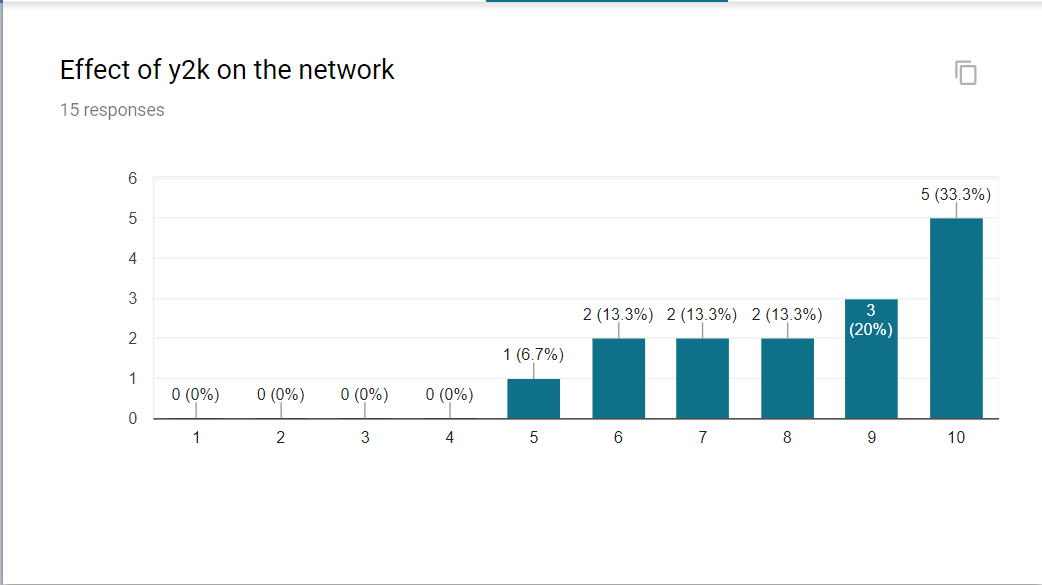


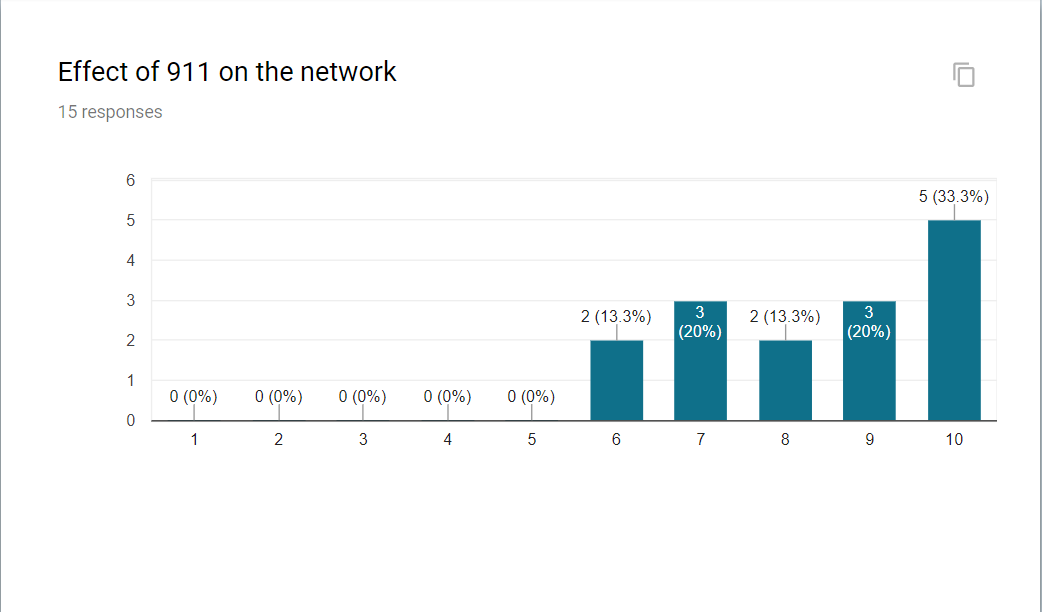


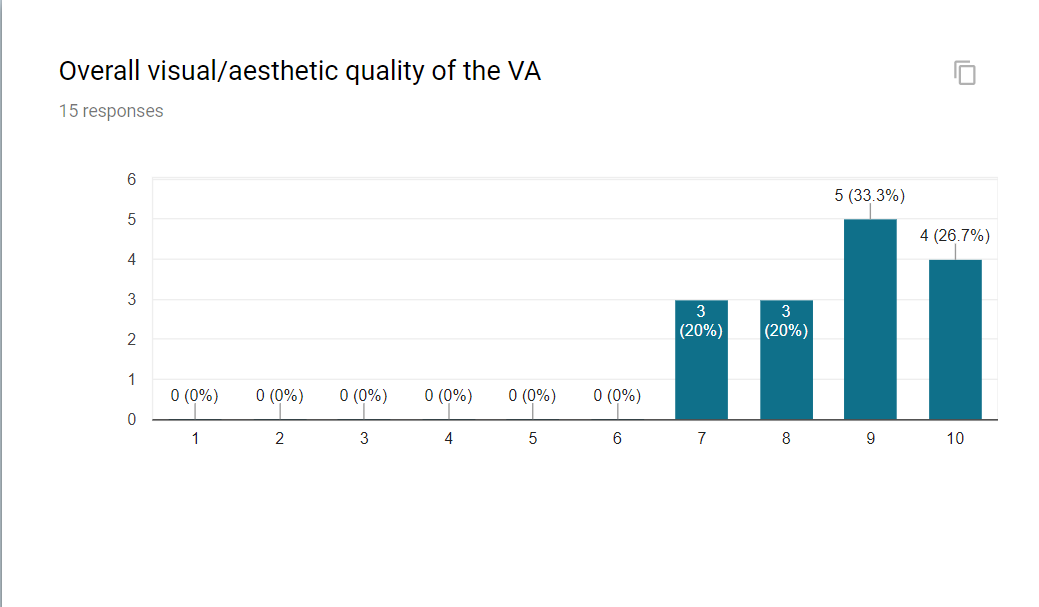
**Result (Questionnaires): 15 responses has been received, following is summary of response received, details of each response is attached in appendix.**

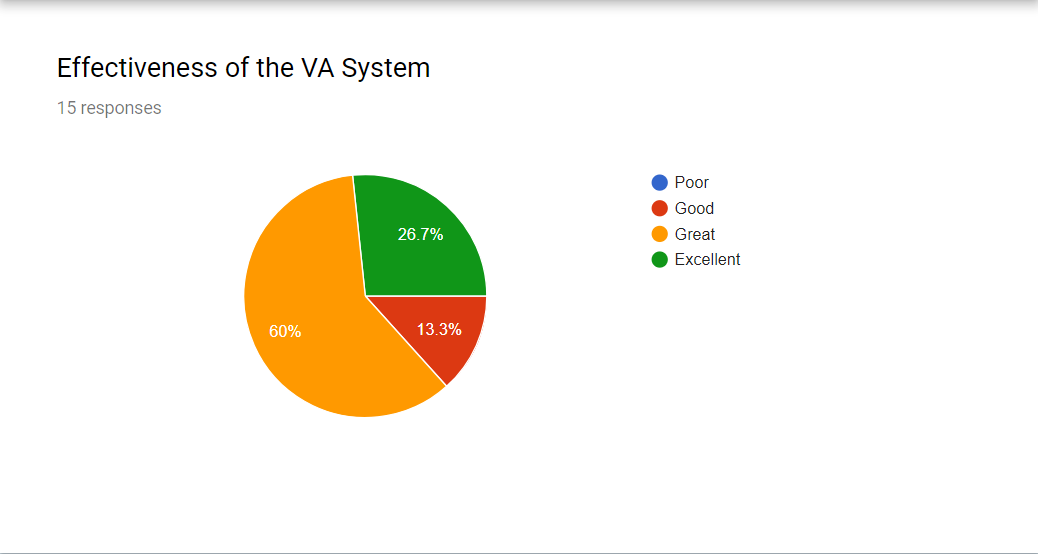
****

****

****

****

****

****

### 5.1.2 Survey (Interview)

**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 involve 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 dayis represented by squares and shows flight delay percentage for the day. According to the size of the dataset, the calendar represent the delay data effectively and directly.  Two map visualizations are very attracting. The heat map show the different airport sizes and different delay level.In the second map, colour scale indicates percentage delay. This interactive visualization contain the info from 1987 to 2008.SO great.  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 could be tough and complex. If you could simply it, it could be better.  That’s all. | | | |

### 

### 5.1.3 Empirical evaluation (statistical analysis)

Statistical analysis has been performed using python code to prove visualization is working as expected -

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

|  |  |
| --- | --- |
| **Statistical Analysis** | Best and Worst Airlines from Visualisation |
| **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 means 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 Visualisation |
| **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 means was backed by this statistical analysis. We have evidence to suggest that departure delay is related to Airport. |

## 5.2 Discussion

# 6. Conclusion

# 

# 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>

# 8. Appendix: Group meeting minutes (0.5-1 page per week: week 7-12)

# 9. Appendix: Code