## **Airlines dataset**

## Data Set

Airlines dataset comes from the Research and Innovative Technology Administration (RITA), from the United States Department of Transportation. Data is from 1987 to 2008 and consists of figures about the arrival and departure details for the commercial flights throughout the 3,376 airports in the USA. The data is split down into 22 yearly chunk files, year by year. The dataset is 12GB with nearly about 120 million records.

## Additional data requirement

The dataset needed to be combined with the supplementary datasets provided to complete the analysis and visualization.

* **Airport data** - Required to get airport names, city, longitude, and latitude of the airports
* **Carrier data** - Required to get Airline's name

## Data Cleansing

Though the data set was already formatted, it still required additional data cleansing activities such as:

* Date is split into 3 columns (year month, day of month), it had to be combined
* Some columns like Arrival delay and departure delay contained value as 'NA' wherever not applicable. This needed to be converted to zero before performing any arithmetic operation
* Some of the airport names contain ',' in their names which was replaced with spaces

## Aims and Contribution

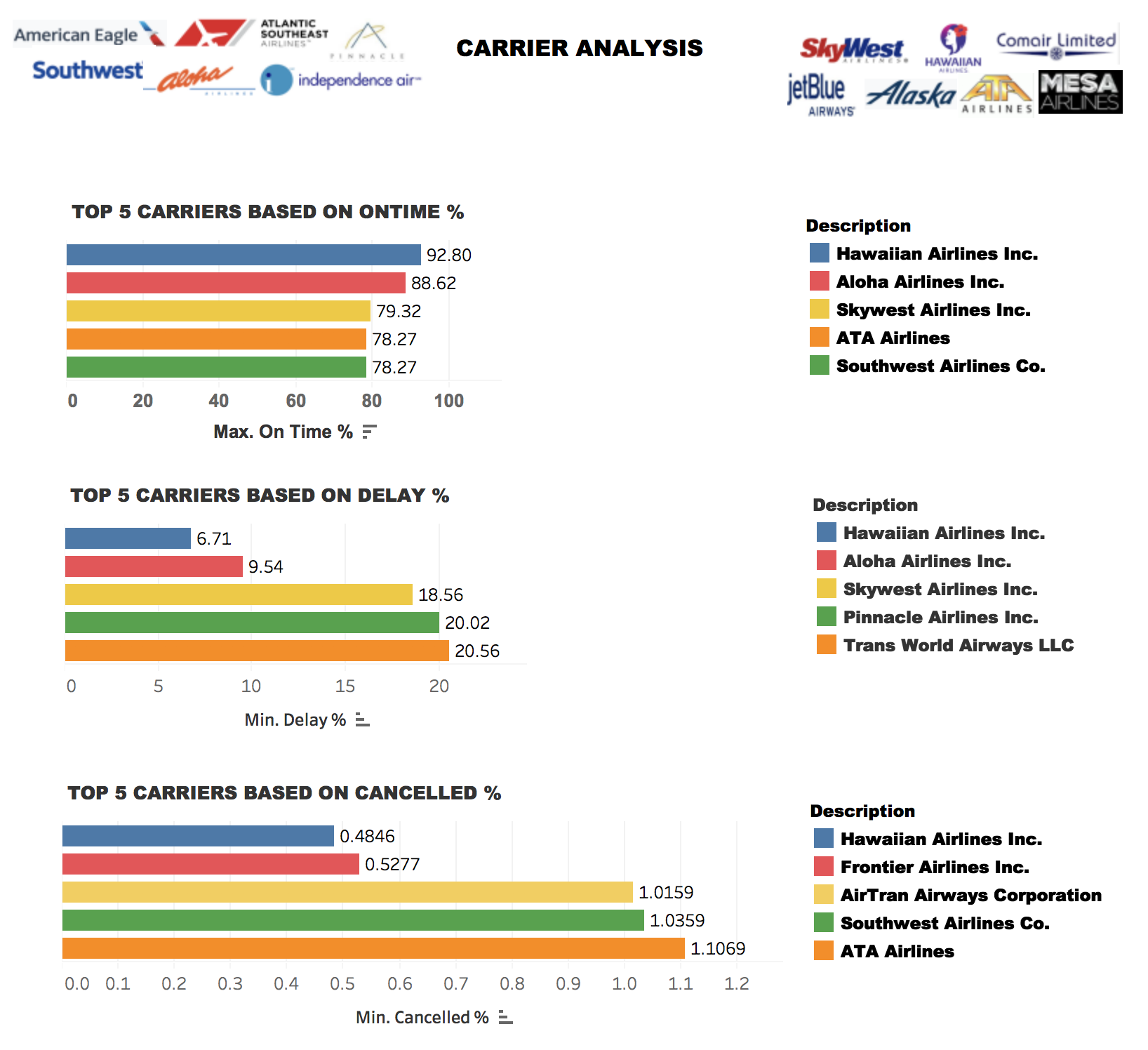
We have performed analysis and produced visualizations on the Airlines data in order to identify:

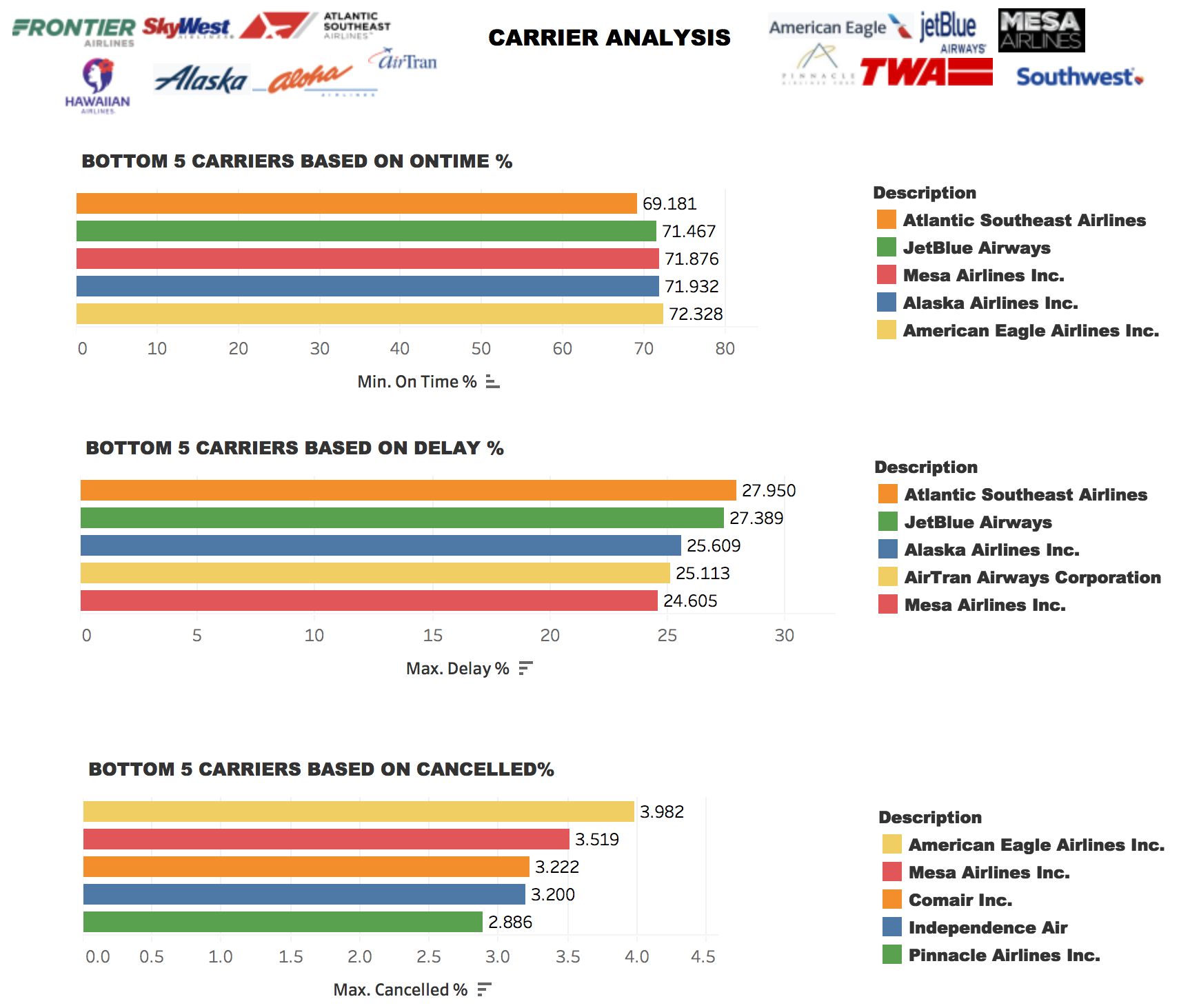
* Usual and unusual patterns in the data
* Best Month/day of the travel to avoid delay/cancellation
* Best and Worst Airlines
* Best and Worst Airports
* Identify reasons for Cancellations
* Analysis of Airport Routes based on the best and Worst Airports
* Factors that contributed to the airline’s delay/cancellation such as 9/11 attack,Global financial crisis, weather and public holidays

# Visualisation

### Visualisation – Carrier Analysis

A ***bar*** visualisation for the carrier analysis has been created to identify Top and Bottom Carriers based on percentage of On Time,Delay and Cancelled from the Visualisation 2 a and Visualisation 2 b





**Patterns Identified:**

**Top Carriers:**

* Ontime -Hawaiian Airlines
* Delay Percentage:Hawaiian Airlines
* Cancellation Percentage:Hawaiian Airlines

**Bottom Carriers**

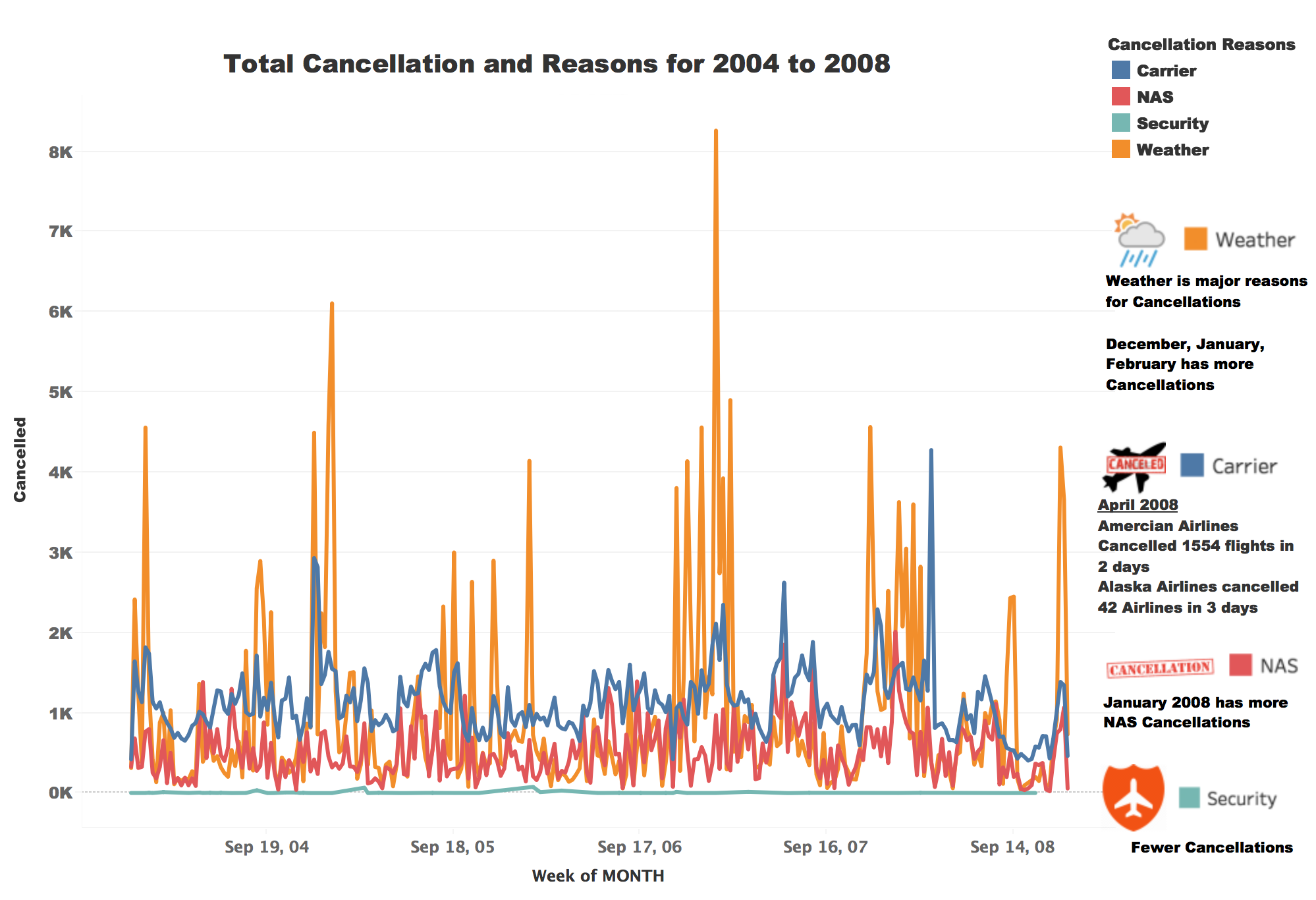
* Ontime -Atlantic Southeast Airlines
* Delay Percentage:Atlantic Southeast Airlines
* Cancellation Percentage:American Eagle Airlines

### 

### Visualisation - Total Cancellations and Reasons during a 5-year period

A ***lines (continuous)*** visualisation have been created for the total cancellations VS the reason of flight cancellation. The information depicted by this visualisation will assist us in identifying the reason which led to the cancellation of a particular flight at a specific period of time. The X-axis has been plotted as the month and the Y-axis as the cancelled flights. The colour of the lines in the graph represents the reason for cancellation of flight. These reasons are categorized as below:

* A → cancellation of flight due to carrier
* B → cancellation of flight due to the weather
* C → cancellation of flight due to NAS
* D → cancellation of flight due to security reasons



**Pattern Identified:**

* From the visualisation 5, We can observe that weather is a major reasons for Cancellations from 2004 to 2008
* December,January and February has more number of Cancellations due to weather.
* The Second highest reason for cancellation is due to Carrier.In April 2008,American Airlines cancelled 1554 flights in two days.Alaska Airlines cancelled 42 Airlines in three days[3]
* January 2008 has more number of NAS Cancellations
* There are fewer Cancellations due to Security

### Visualisation - Total number of flights and cancellations per week during a 5-year period

A ***heatmap*** visualisation for the total number of cancellations/delays per week /month (Visualisation 6 a to 6 f) have been created, which will assist us in identifying the weeks/months with the maximum and the minimum number of cancellations/delays.

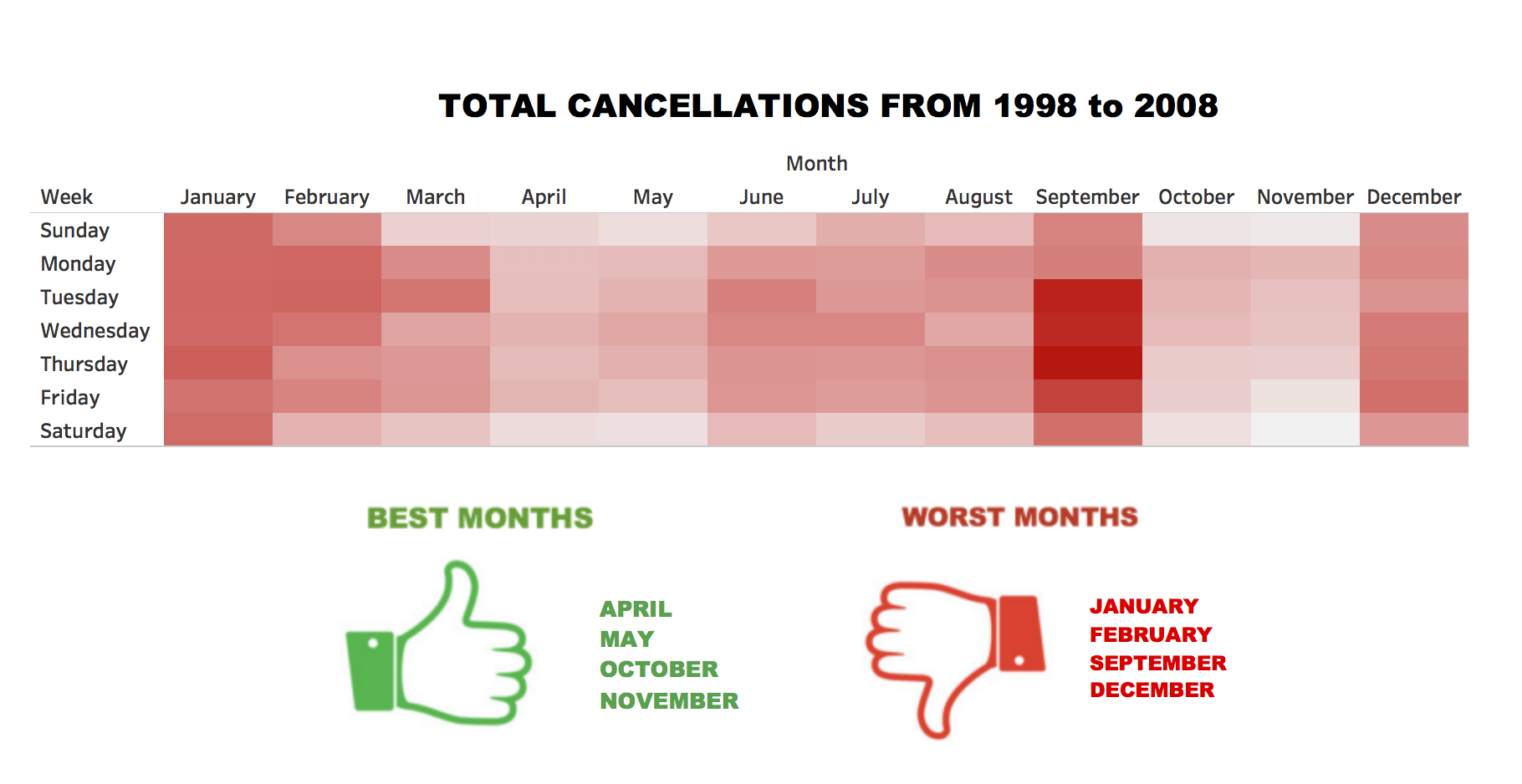
**Pattern Identified:**

**Based on Cancellations**

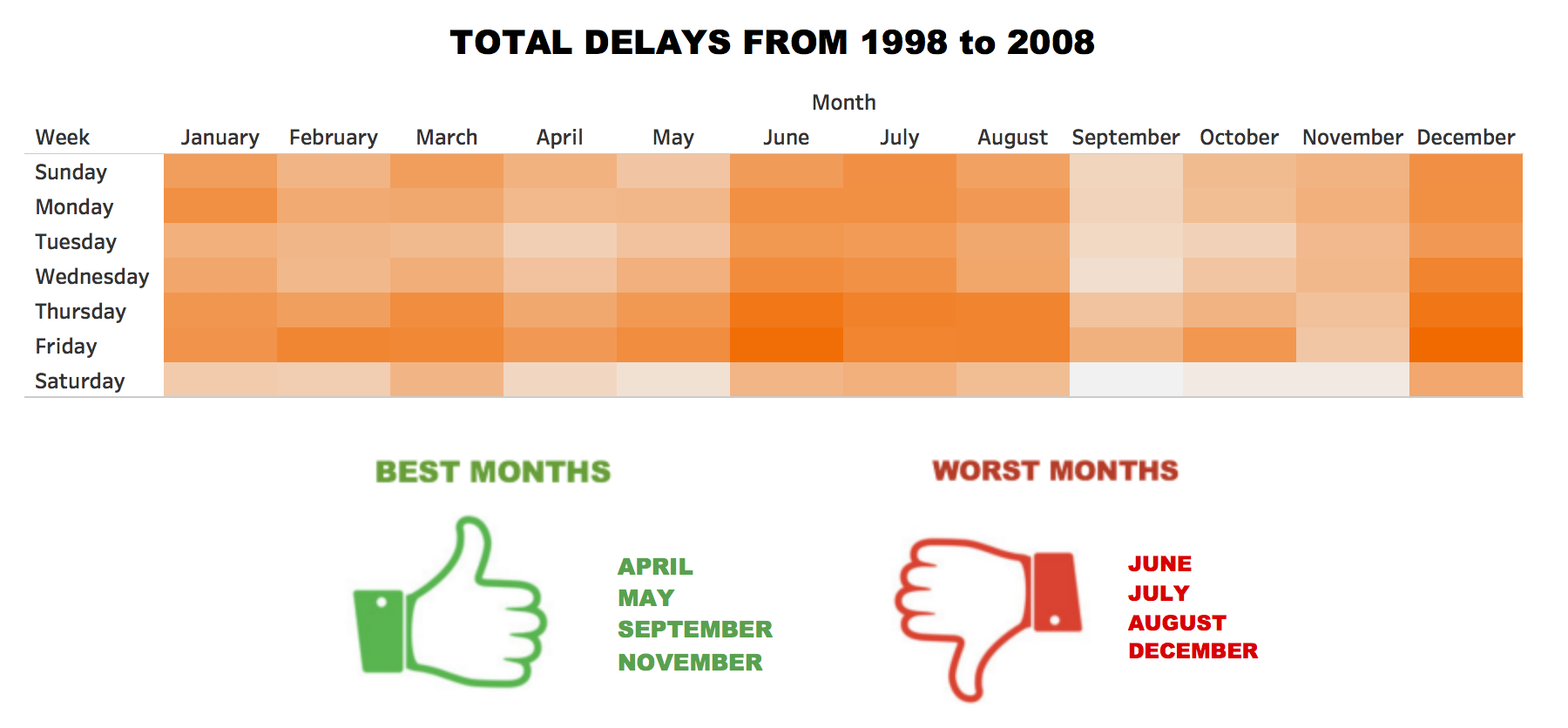
* Best Month: April,May,October,November
* Worst Month:January,February,September,December
* Best Days:Saturday,Sunday
* Worst Days:Monday,Tuesday

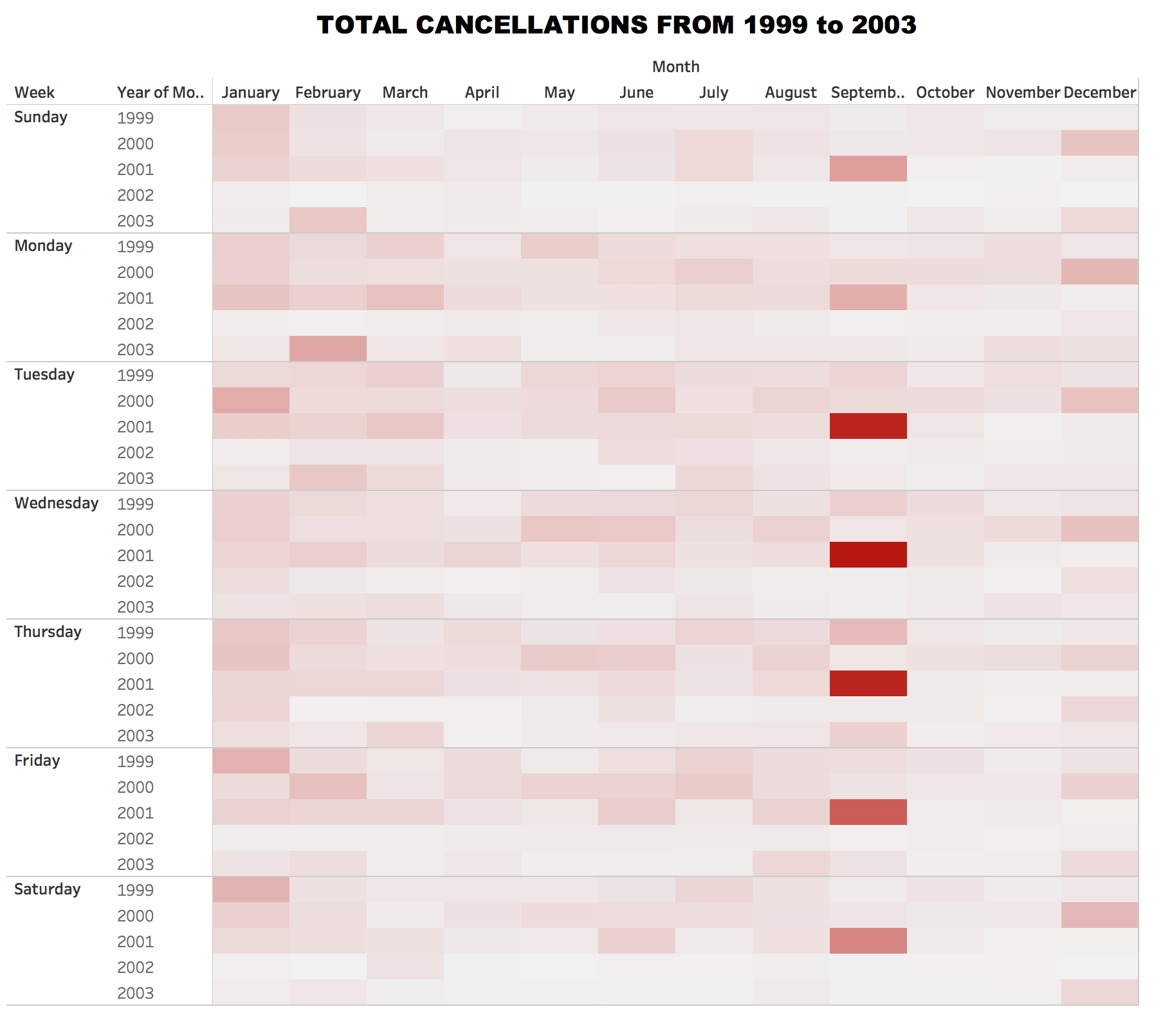
**Based on Delays:**

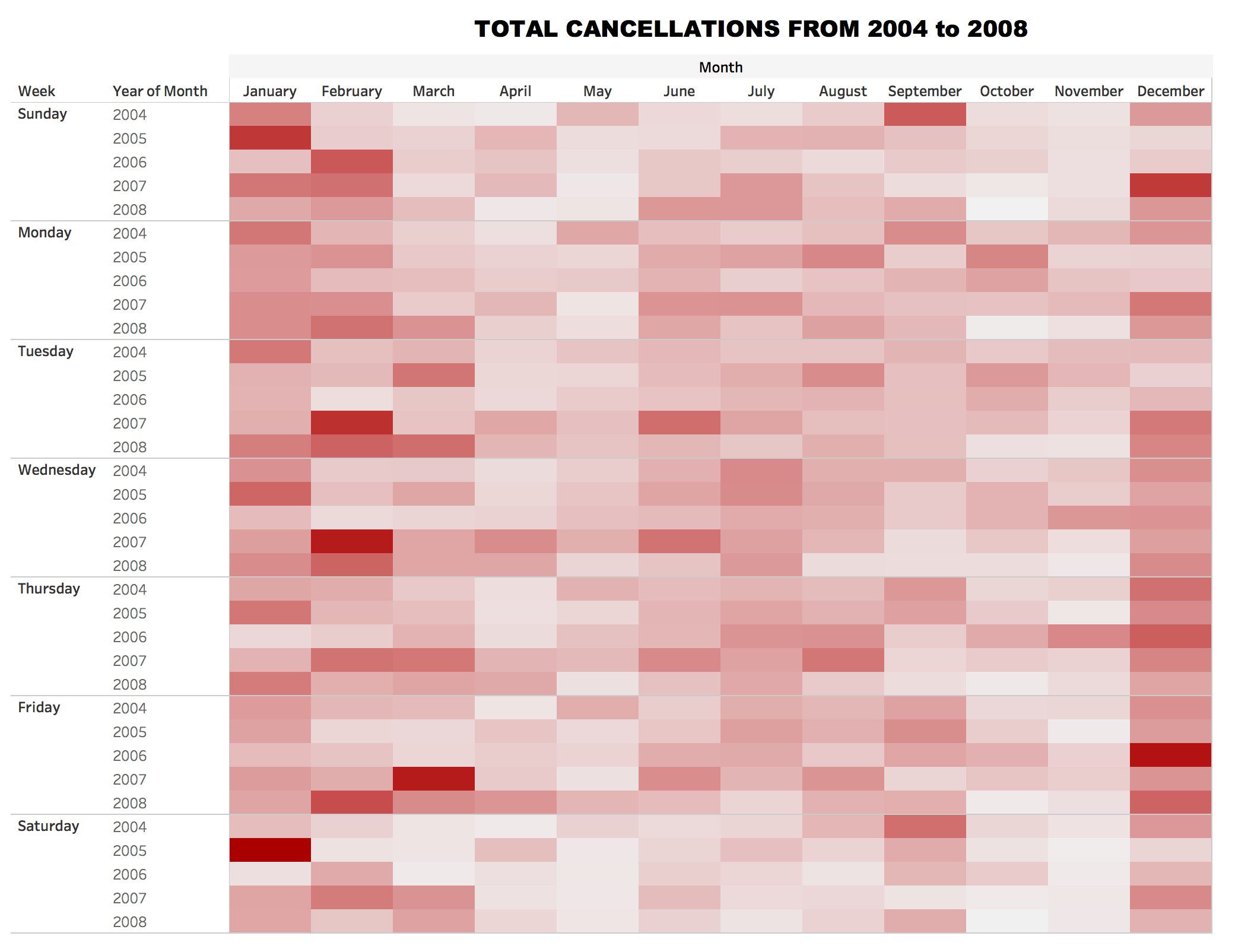
* Best Month:April,May,September,November
* Worst Month:June,July,August,December
* Best Days:Tuesday,Saturday
* Worst Days:Thursday,Friday

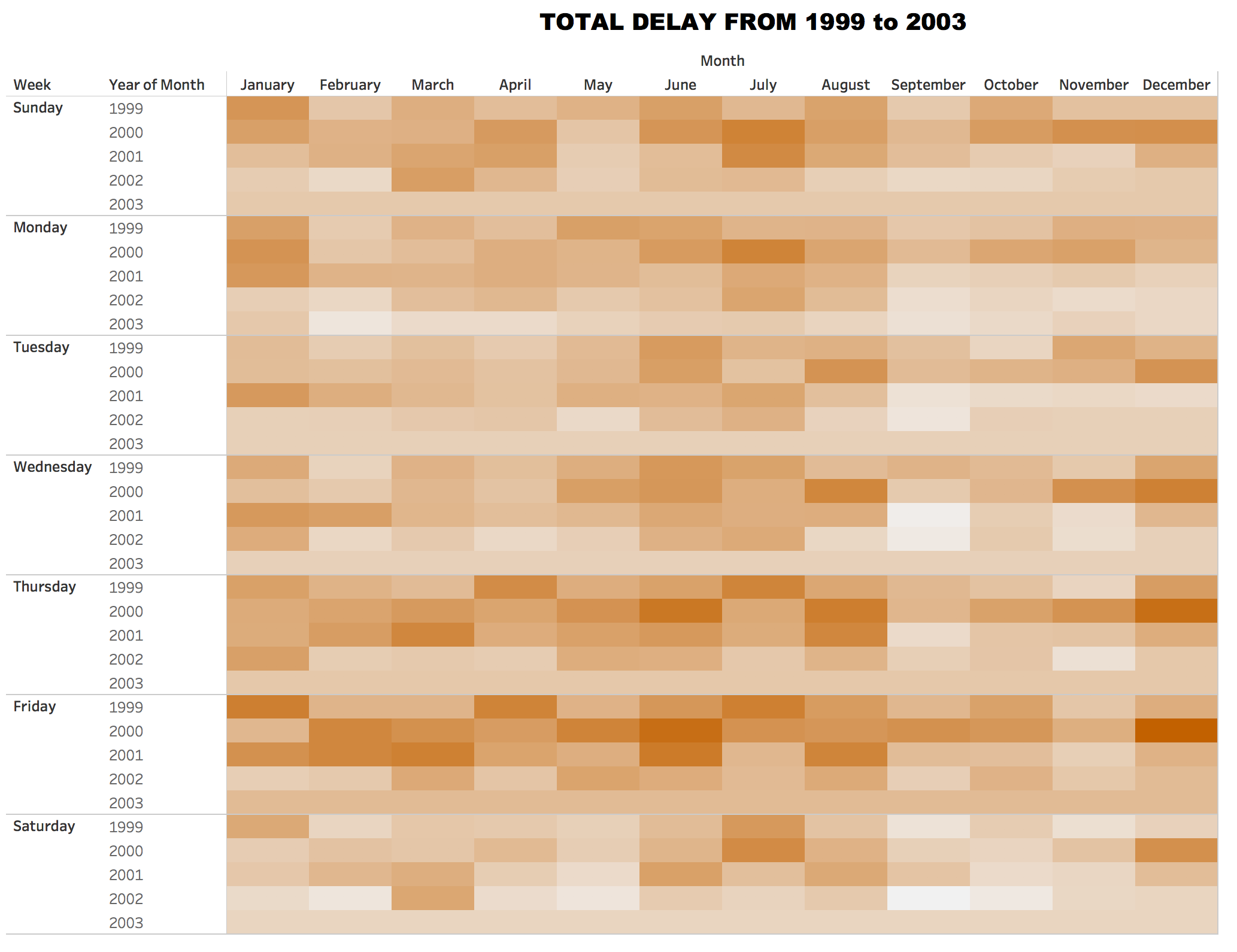


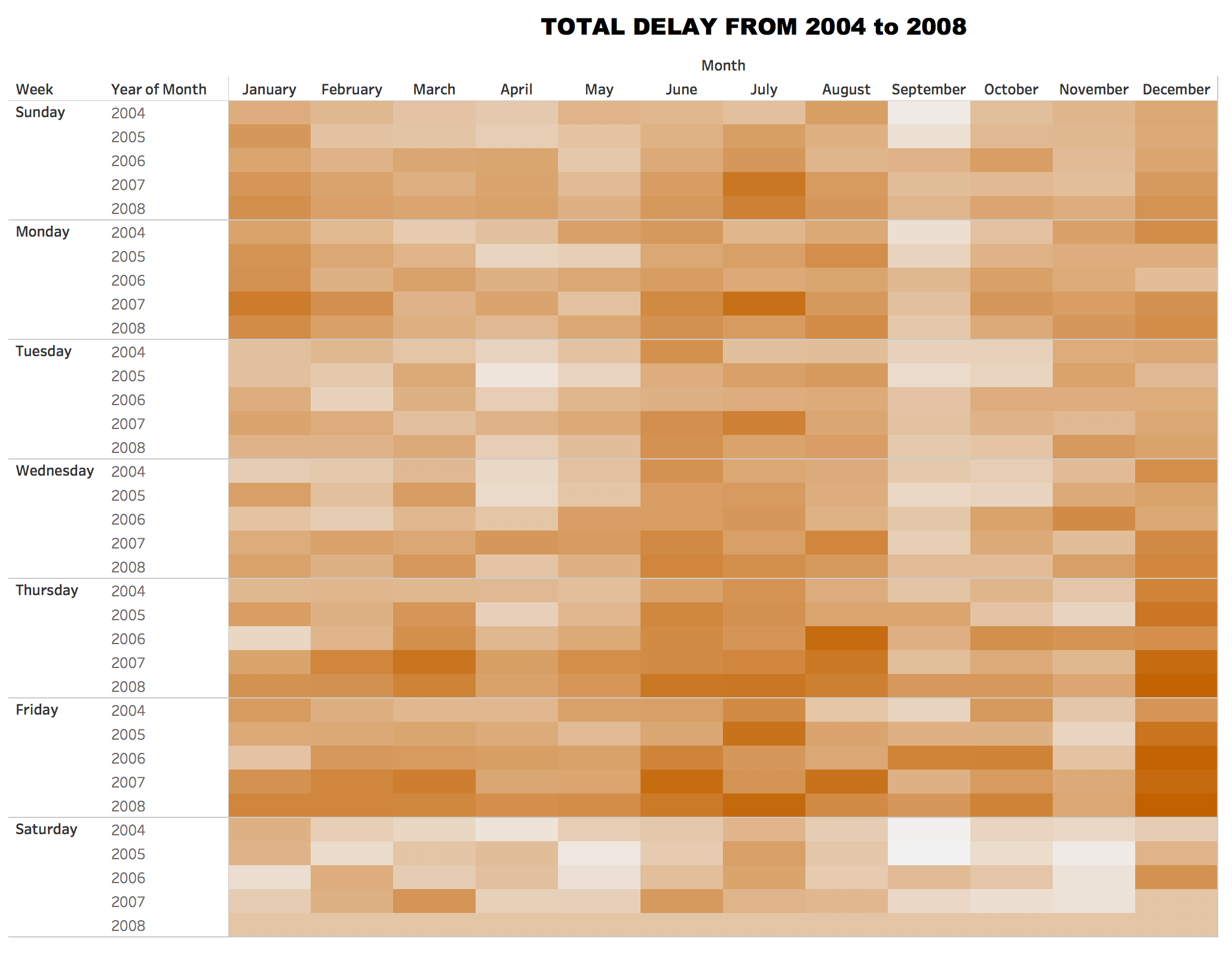




**



**

**

### Visualisation - Trend Analysis

A ***lines (continuous)*** visualisation have been created for the based on the Top Five Airlines and Airports with respect to Cancellation and Delays to analyse the Trends across years.

# Screen Shot 2017-11-01 at 11.06.00 pm.png

From the Visualisation 10 a we can observe the trend For Cancellation and Delays for Carriers.We have Identified the Top 5 Carriers from Visualisation 2 a and 2b-Carrier Analysis.We can observe that SouthWest Airlines experienced higher Cancellation percentage and Aloha Airlines had higher delay percentage in 2001 due september 11 terrorist attack.On 26th Nov and 17th December of 2006,there was 3805 and 4140 cancellations respectively due to weather(From Visualisation 5).Frontier,Southwest and ATA Airlines has higher Cancellations percentage on 2006.After 2006,the cancellation percentage of Frontier had gradually decreased.

From the Visualisation 10 b we can observe the trend For Cancellation and Delays for Airports.All the five airports(Chicago O’Hare International,Gen Edw L.Logan International,LaGuardia,Newark International and Ronald Reagan Washington National) had higher cancellation Percentage i*n 2001 due september 11 terrorist attack.In 2002,all the five airports experienced a sudden dip in the cancellation Percentage.In 2007,all the 5 airports had highest delay Percentage and comparatively higher Cancellation percentage.According to USA Today Article published in 2007, the main reasons for the cause of delays are due to shortages of pilot,longer time to refuel and mechanical breakdowns.[4]*