**A Data Visualization Project**

**On**

**Flights Cancellation and Delay depending upon various circumstances**

**Jasmeet Singh**

**X14121336**

**MSCDA (2015-16)**

1. **Abstract**

This project is about the flight cancellation and delay information of all the domestic airports in United States of America (USA). It contains the details of only the month of January, 2008. Data is cleaned to make suitable for analysis and generating reports or graphs. Graphical representations are the best way to understand the actual data in the correct manner. This project deals with data related to airlines such as which airline is preferred by the passengers, how many flights got delayed and for what reasons.

1. **Introduction**

This data is based on the flights details which contains fields related to cancellation and diversion of flights. It states various reasons of delay in flights. The data also contains information about the producer of planes. This project covers all the possible reasons for flight cancellation and delay issues based on that data is cleaned to implement the successful generation of delay reports and the main focus of this project is to calculate the flow amount of flights of a particular airport to measure the air traffic situations which may cause huge amount of loss for airport authorities and results in flight delay issues. This project helps in predicting the reasons for late arrivals and late departures. Airport authorities can use the visualized reports to find out the problems and can fix them within the timeframe, unless they have to bear huge losses.

1. **Dataset**

Dataset contains some data which needs to be cleaned such as time format which is recorded in the four-digit format (0000) and later transformed to actual time format (HH:MM). Some fields in the dataset are irrelevant, so they are removed. After a lot of cleansing process, the data turned into meaningful sense and is ready to use.

A flight is considered delayed when it arrived 15 or more minutes than the scheduled time. Delayed time is calculated for delayed flights only.

When multiple causes are assigned to one delayed flight, each cause is computed based on delayed minutes, it is responsible for. The displayed values are rounded and may not add up to the total.  
SOURCE: Bureau of Transportation Statistics, Airline Service Quality Performance 234.

Download Link: <http://stat-computing.org/dataexpo/2009/the-data.html>

**Variable descriptions**

|  |  |  |
| --- | --- | --- |
|  | Name | Description |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) (Expected Departure Time) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) (Expected Arrival Time) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes (converted to hh:mm format) |
| 13 | CRSElapsedTime | in minutes (converted to hh:mm format) |
| 14 | AirTime | in minutes (converted to hh:mm format) |
| 15 | ArrDelay | arrival delay, in minutes (converted to HH:MM format) |
| 16 | DepDelay | departure delay, in minutes (converted to HH:MM format) |
| 17 | Origin | origin IATA airport code (international airport abbreviation code) |
| 18 | Dest | destination IATA airport code (international airport abbreviation code) |
| 19 | Distance | in miles |
| 20 | Cancelled | Was the flight cancelled? |
| 21 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 22 | Diverted | 1 = yes, 0 = no |
| 23 | CarrierDelay | in minutes |
| 24 | WeatherDelay | in minutes |
| 25 | NASDelay | in minutes (National Aviation System) |
| 26 | SecurityDelay | in minutes |
| 27 | LateAircraftDelay | in minutes |

Further related information is also available on the web about the aircrafts, weather and carriers (<http://stat-computing.org/dataexpo/2009/supplemental-data.html>).

## Airports

Describes the locations of all the airports in United States of America, with the following fields:

* IATA: the international airport abbreviation code
* Name of the airport
* City and Country: specifies where airport is located.
* Lat and Long: latitude and longitude represents the actual geographical coordinates of the airport.

This majority of this data comes from the Federal Aviation Association (FAA).

## Carrier codes

List of all carrier codes with full description.

## Planes

This file contains the information about the airplane’s tail numbers, manufacturers, year of manufacture, engine type and many more.

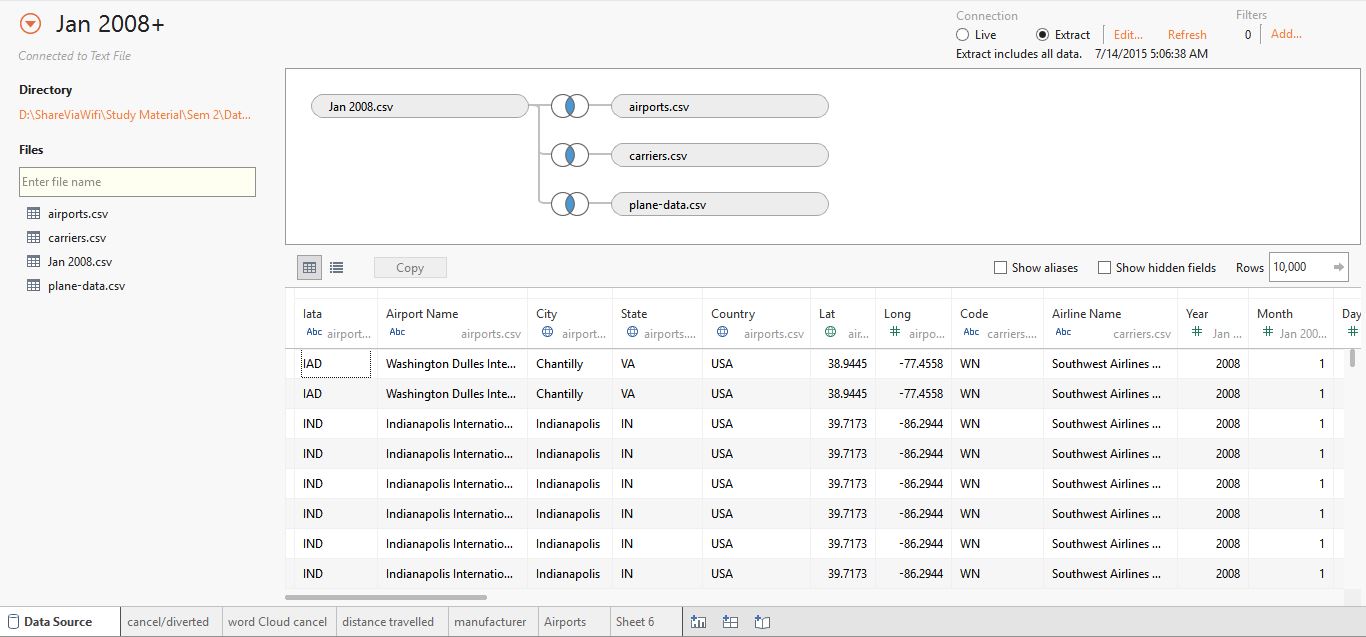
1. **Data Visualization Tool**

The tool which is used to find out the meaning from the dataset is Tableau v9.0. It is the best tool to represent large amount of data in the form of pictorial representations in a user friendly manner. Tableau is easy to use software, by using simple drag and drop feature so that the data could be converted into meaningful information.

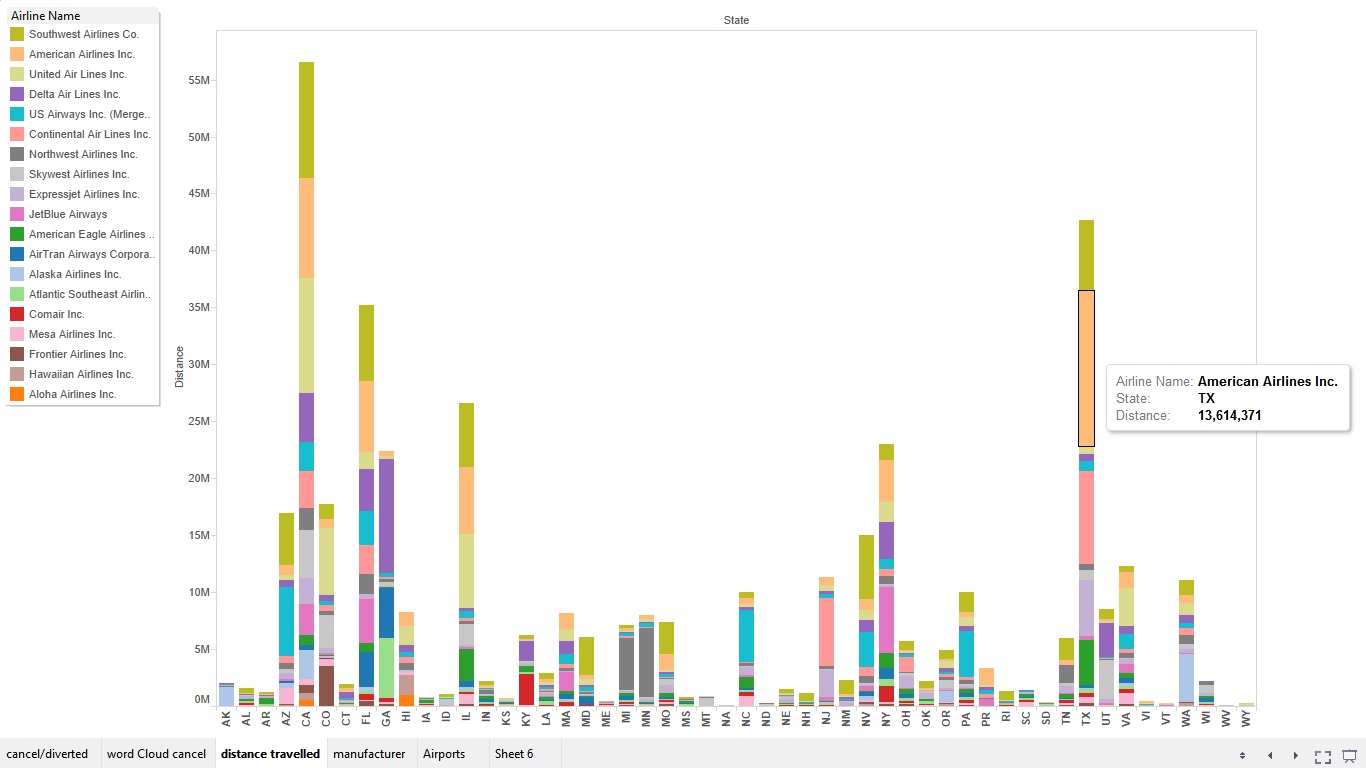
Google Refine is used to clean the data using the text facet filter to group all the airline names as they are stored in database with different spellings. In order to combine them into a group this tool is used for cleaning the data and make it suitable for visualization.

**4.1. Working with Tableau**

Linking all the datasets together to perform some actions on the combined data. Different datasets are joined on the basis of some common field between them. For example, on the basis of Iata (international airport abbreviation code) dataset “airport.csv“ is joined with “Jan2008.csv”, similarly “carriers.csv” is connected on the basis of “unique carrier” and “plane-data.csv” is joined with respect to “tail number” respectively. The below shown figure represents the combined fields of datasets and their respective values.

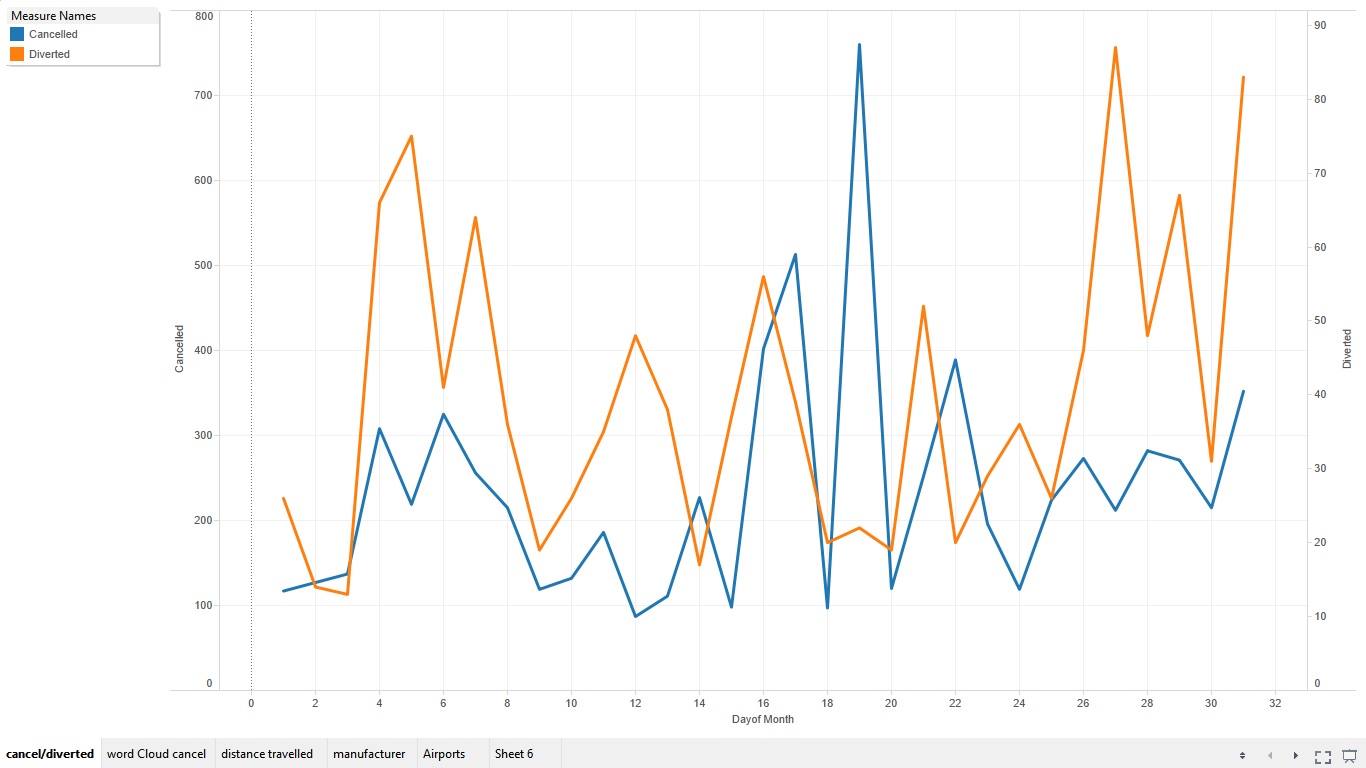


1. **Business Queries**
   1. Is there any impact of people travelling frequently in the state and the distance travelled by the airline?



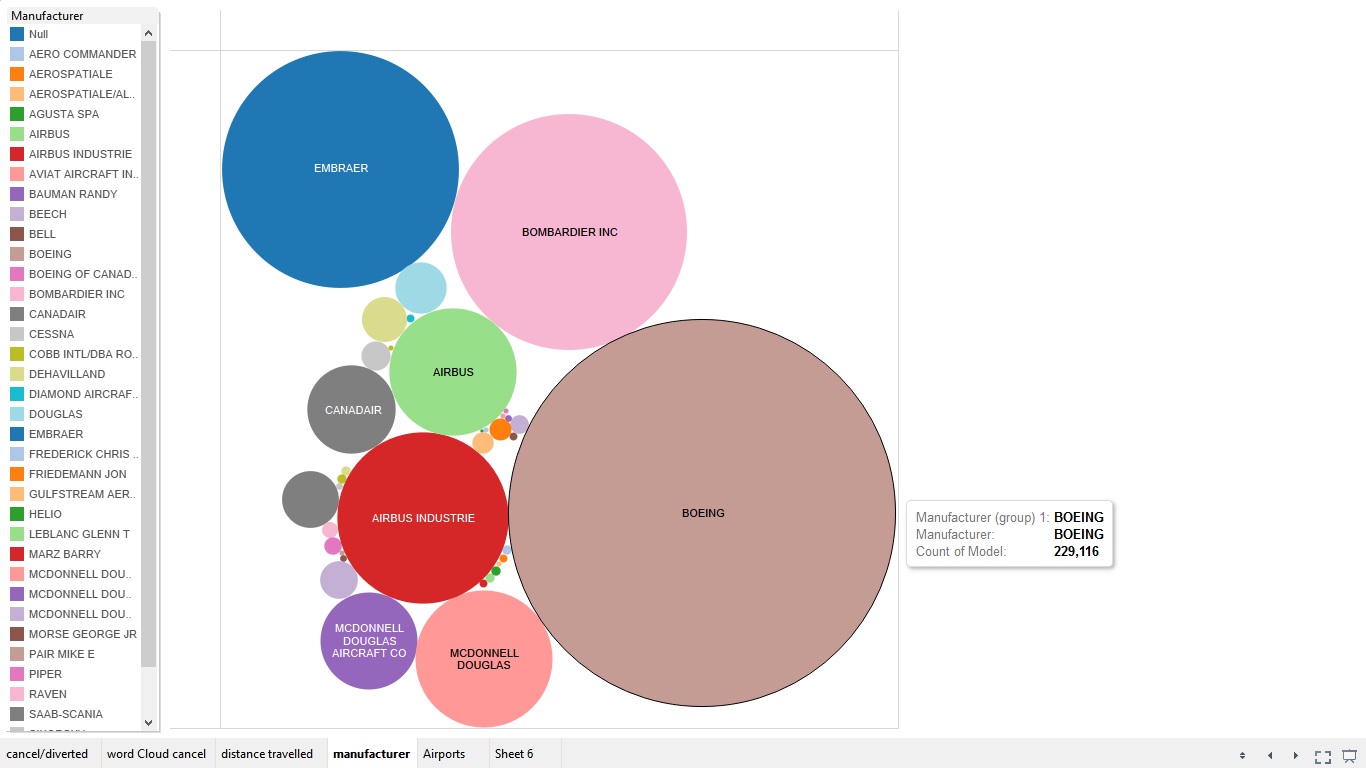
List of all the airlines used in every state in USA and which airline is most preferred by the citizens of a particular state. This information is collected by the distance travelled by that airline. It indicates that which airline is more popular in that state. For example, “United Air Lines Inc.” is mostly used by the citizens of “California”. Similarly, “American Airlines Inc.” is widely used in “Texas”.

* 1. Which day in the January record more diversions and cancellations in the flight schedule?



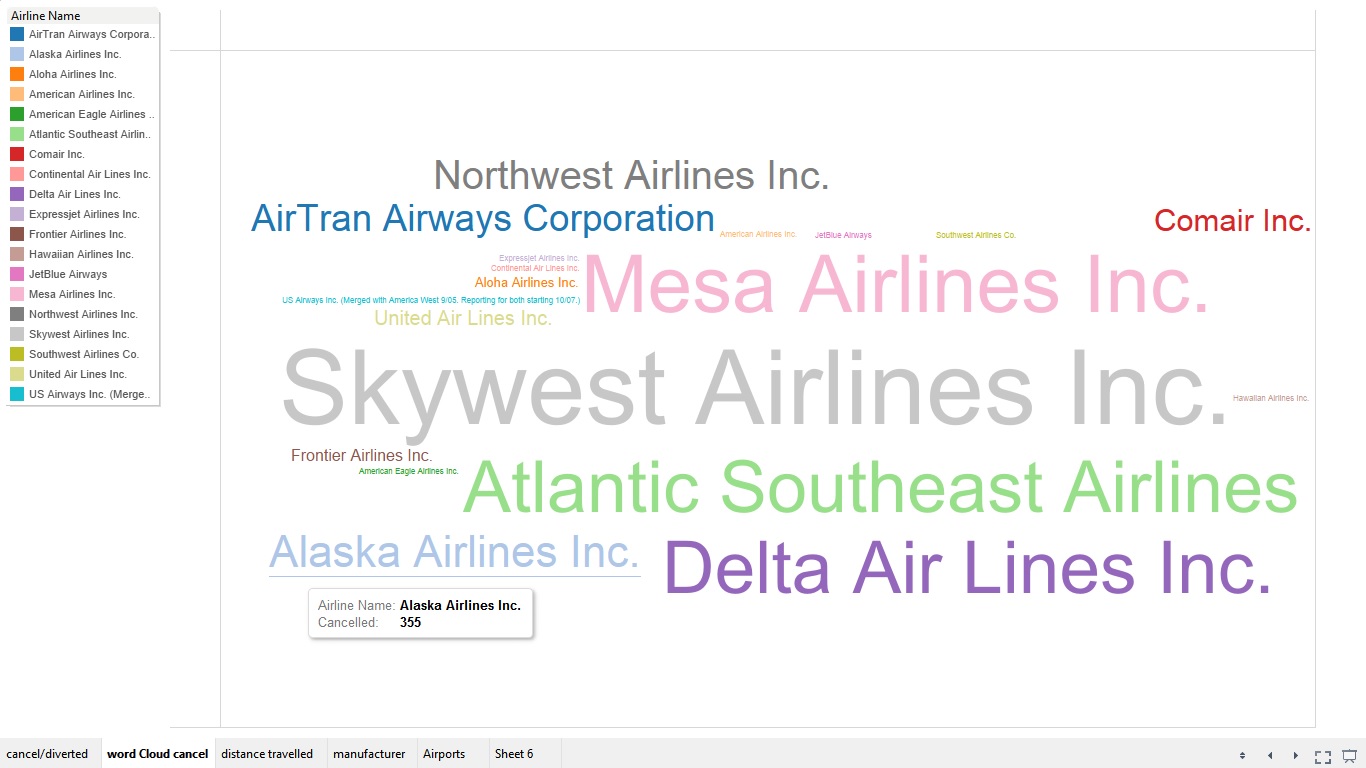
The trends of cancelled and diverted for “Day of Month”, color shows details about cancelled and diverted flights. This figure represents the cancellation and delay flights, such as, 19th of January records highest number of cancellation flights (760) and day 12 records the minimum number of cancellation which is 87.on the other hand, maximum number of delayed flights were recorded as 87 on 27th day of January, while 3rd January holds least amount of delayed flights i.e. 13.

* 1. Which manufacturer is highly preferred by the airlines?



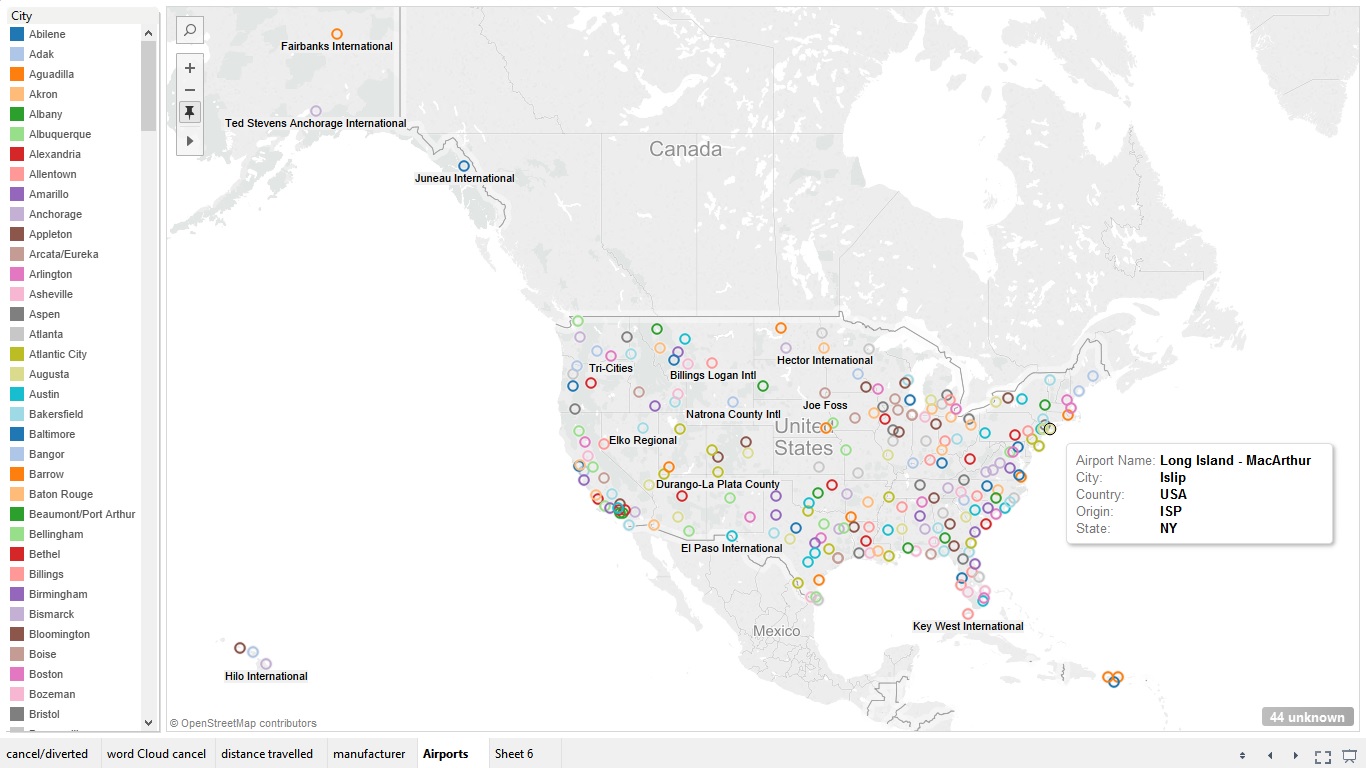
This data shows that which aircraft is produced more and which is less. As the results shows that in January 2008, Boeing planes were recorded as 229,116 which is the highest number of manufactured planes followed by Embraer, Bombardier and Airbus such as 85797, 85144 and 44849 respectively. Whereas, Agusta Spa ended at 18 which the least manufactured aircraft.

* 1. Word Cloud of airlines based upon the cancellation of flights.

****

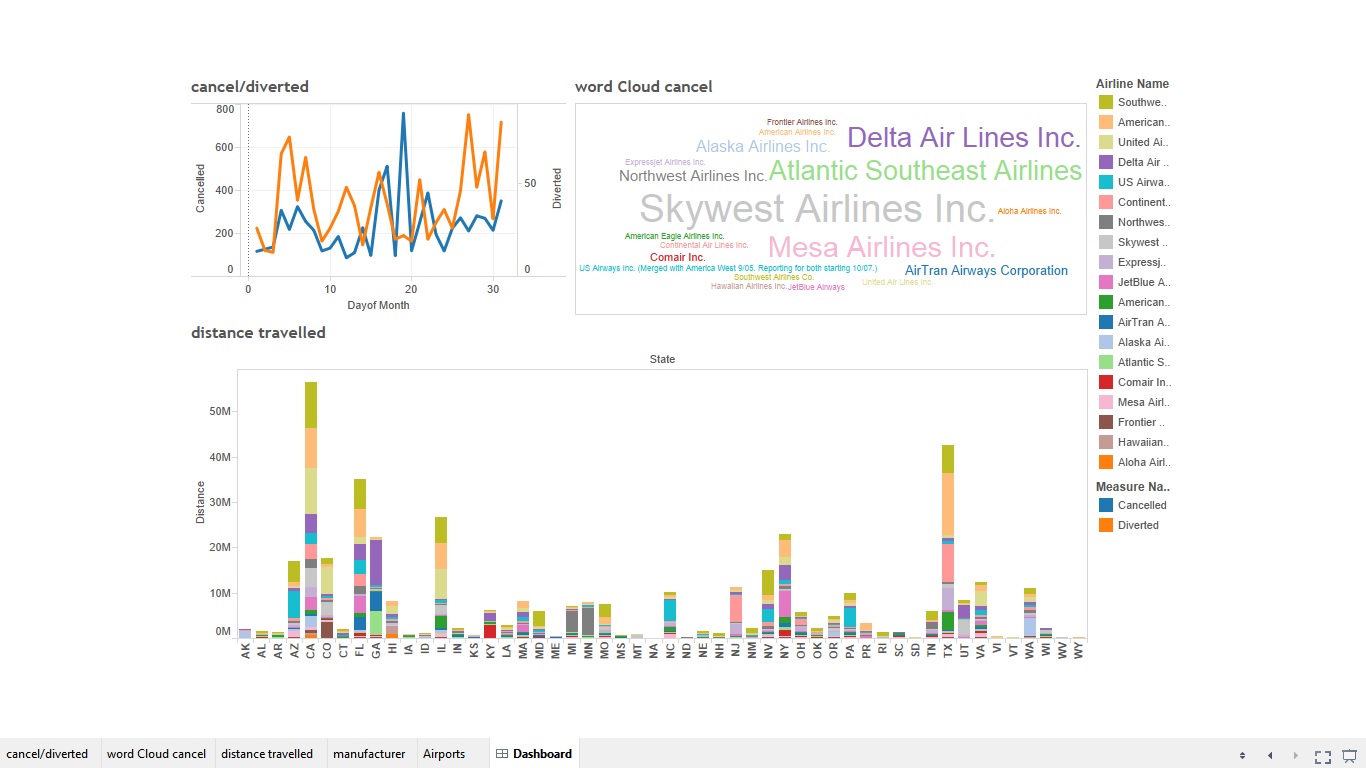
This figure shows that which airline was cancelled most and the results are represented by word cloud. Bigger size of airline names shows the maximum number of cancellations whereas smaller size of airline names represents least number of cancellations of flights. Flights of Skywest Airlines was cancelled 2,511 times whereas flights of American Airlines Inc., American Eagle Airlines Inc., Continental Air Lines Inc., Expressjet Airlines Inc., Hawaiian Airlines Inc., JetBlue Airways, Southwest Airlines Co. and US Airways Inc. were never cancelled.

* 1. List of all the domestic airports in USA.

****

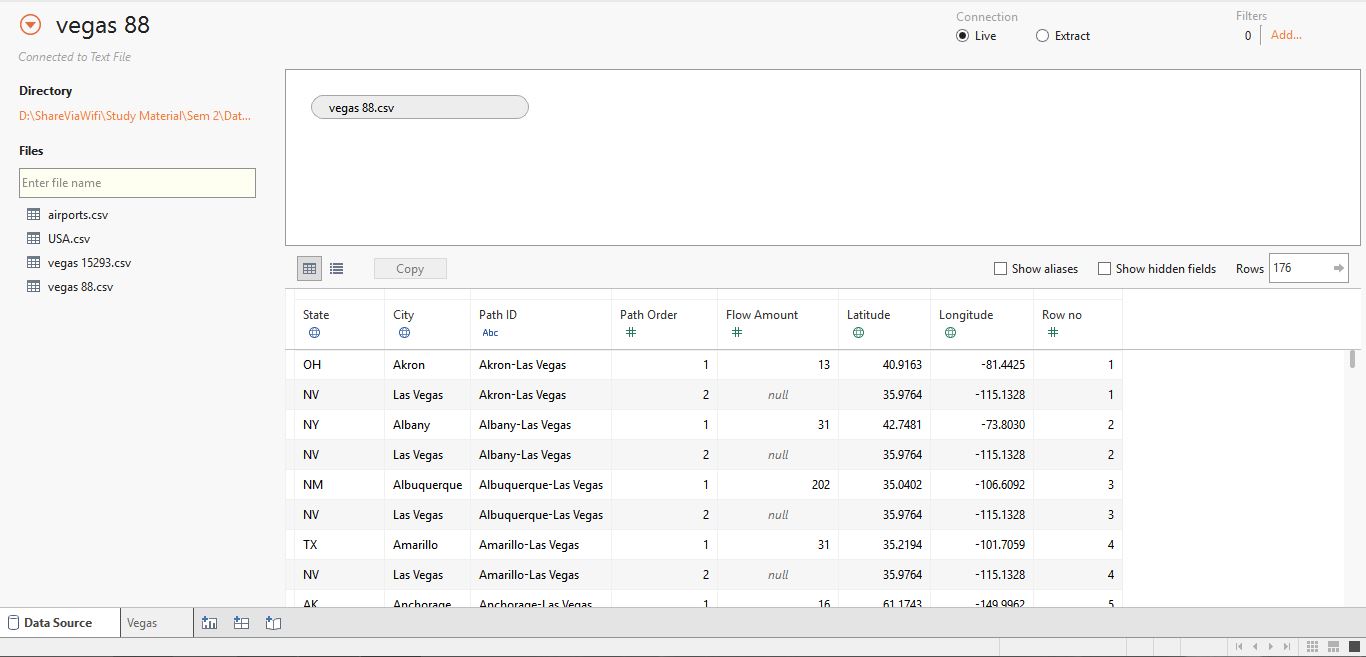
Map based on Longitude (generated) and Latitude (generated). Color shows details about City. The marks are labeled by Airport Name. Details are shown for State, Country and Origin city. The view is filtered on Origin and Destination of flight, which keeps track of 278 different airports.

* 1. Dashboard



Here, we can see the results in single window. It keep tracks of all the recorded data. On the basis of airlines names displayed at right side of the dashboard, the respective charts get affected and shows the data of only the selected flights. Dynamic results can be seen at the same time in a single view.

* 1. Flights to Vegas.

****

A new dataset is created manually on the basis of Jan2008.csv contains data only 88 source cities, which contains following information of flights flying towards Los Vegas in the month of January 2008.

State: Name of state.

City: Name of City.

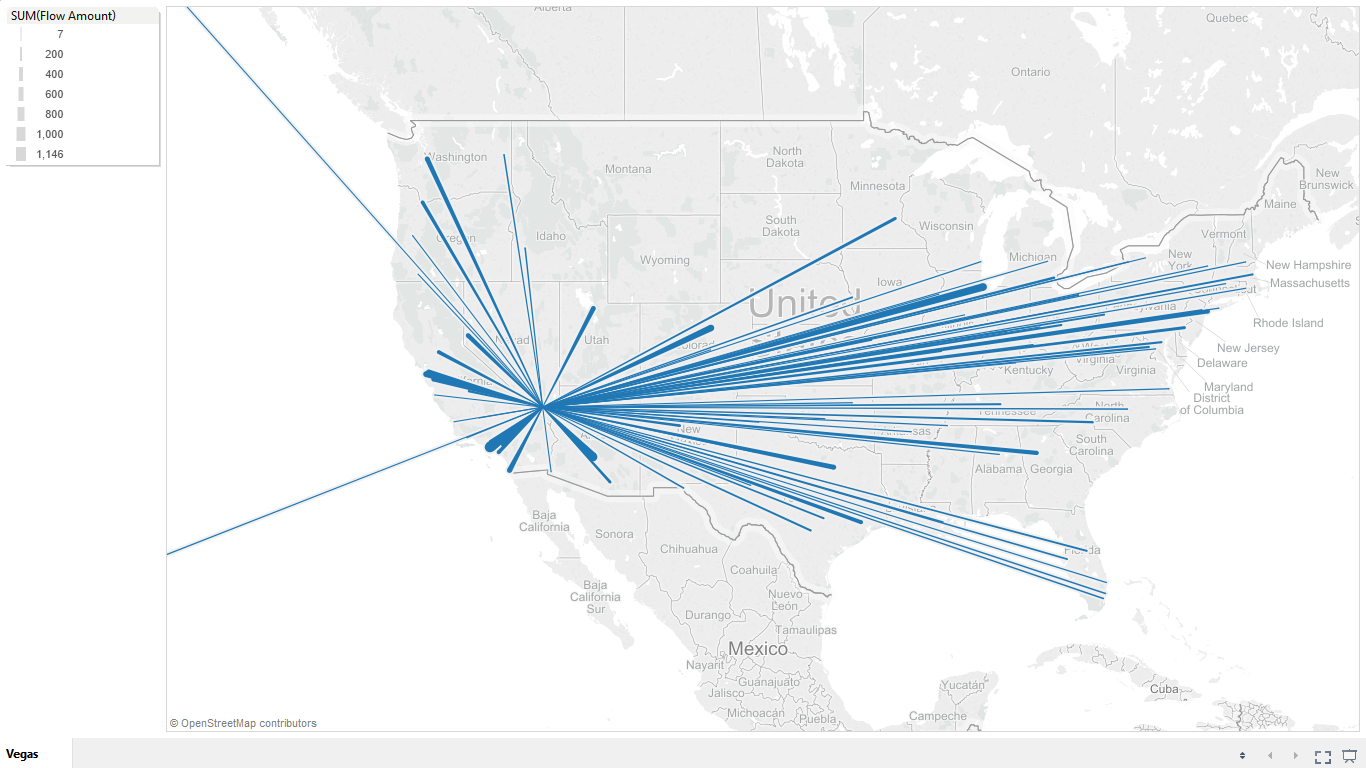
Path ID: Calculated by concatenating the source and destination of flights separated by hyphen.

Path Order: 1 = source city and 2 = destination city.

Flow Amount: Represents the number of flights towards that city.

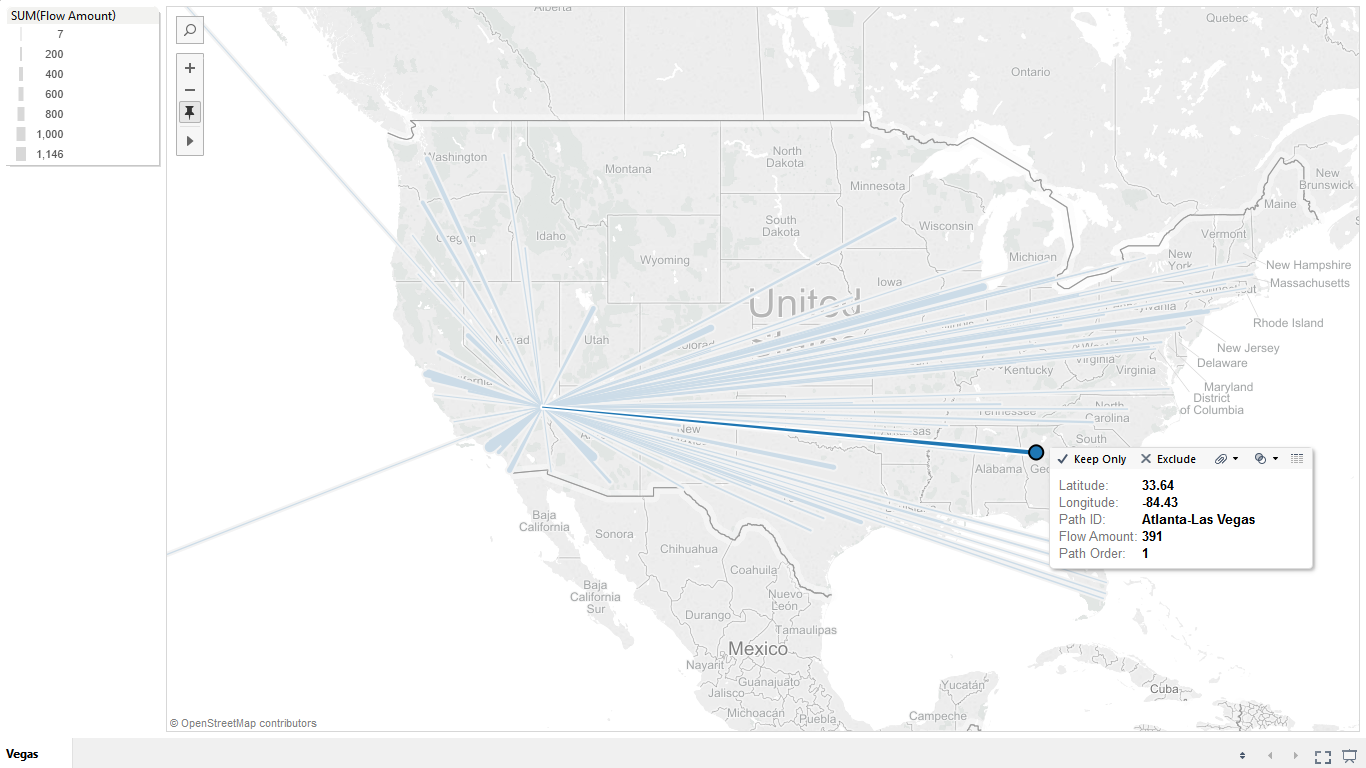
Latitude & Longitude: Represents the actual geographical coordinates of the airport.

* 1. Visual representation of all the flighs having Las Vegas as their destination.



Above figure displays all the flights in the month of January 2008 having Las Vegas as destination.

Path ID is a mandatory field to create lines from source to destination city. Thickness of every line is defined by Flow Amount which is calculated on the basis of number of flights between two cities in a month.

****

As we can see there are various lines between Las Vegas and rest of the cities in USA. Some lines or paths are thin and some are thick. This thickness is measured using Flow Amount field between two cities. For example, Path ID shows that the flight departed from Atlanta towards Las Vegas and in the month of January 2008 only 391 flights departed towards Las Vegas from Atlanta. If the number of flights were more, the displayed line would be thick otherwise it would be represented as a thin. Therefore, thickness of the line represents the flow of flights.

1. **Conclusion**

Data is analyzed visually to learn the patterns generated in the form of graphs and can be used for future predictions. This report helps us to analyze the cancellation and delay studies based on weather conditions, carrier delay, National Aviation System (NAS) Delay, Security Delay and Late Aircraft Delay. Results shows the exact issues which should be considered to avoid the cancellation of flights. These visualizations will help the aircraft industries to know the actual reasons of flight cancellation and delay problems as well as traffic flow for every airport and flight. We can also measure the flow amount of air traffic of a particular area. This will help us to know which is the busiest airport at a given point of time. Future work can be done on this project to avoid cancellation and delay of flights by determining the actual traffic of the destination airport, whether the runway is free to land the aircraft, otherwise the airport authority has to pay for the delay in landings.