BUSINESS DATA VISUALIZATION

Project Plan

Group 3

JANUARY 11, 2017

SHIKHA SHAH 650358700
KOUSALYA D 669348510
KARAN PATEL 668208243

1. How will you collect and concatenate the monthly BTS Data?

The data is collected from the below-mentioned link:
http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
The timeframe selected is from September 2015 to October 2016. We have used the Filter Year and Filter period option from the BTS website, to download the file. The data is saved in .csv format. Important attributes are selected through the checkboxes (For example, Unique Carrier, Airline ID, Flight Number, Origin, Origin City Name, etc.) from the look-up table. This procedure is repeated for the span of twelve months till September 2016. The downloaded file is in .csv format and contains detailed information regarding all the airlines.

As per the airline assigned, data from BTS shall be first filtered only for that concerned airline (using the lookup information from BTS Website in the *Airline ID* lookup field)
Using the lookup key to be '*airline code*' for the assigned airline, we will filter out the data successfully and saved the file for a month. The same procedure is repeated for twelve months.
[Besides the airline assigned, for analysis or competitors and comparison, we will look for further information on other airlines using the same steps]

Once all the files are downloaded, they are saved on the desktop and using the tool Tableau we can extract the files.
Once all the files extracted using the above steps, Concatenation is done by the following:
Extract > Append Data from Data Source

2. How you will condition the BTS datasets for use in the group project

For conditioning the data, we need to consider various attributes/variables as well as remove all redundant values like Null Values, Duplicate Values, etc.

Variable Selection:
The variable selection process will depend on the following parameters:

- Attributes associated with State of Origin and Destination (For example New York to Illinois)
- Attributes associated with Time (Departure and Arrival time)
- Attributes pertaining to traffic information for that state
- Attributes pertaining to delays and types of delay (Security Delay, Weather Delay, etc.)
- Attributes associated with cancellations and diversion
- Attributes used for comparison regarding cancellation and diversion (Cancellation Code, Diversion Code, Weather, etc.)

Note: These are tentative attributes. These would vary as we proceed with the project.

Cleaning the Missing Values/Null Values:
Based on the above parameters our approach would be to clean the data using various tools like Excel, Tableau Data Interpreter ETL tools (Clover ETL, TalenD).

Sometimes the observation is that null values are unintended/by mistake. In such cases, we would try to have specific lookups from the related fields to find values for the same. For example, a null value in *Airline Name* can be looked-up using '*Flight ID*' field.
In other cases, where there is no way left to lookup/find the null values, we discard those variables and go ahead with the information available. For figures such as the flight duration, we will use the average of that month. Another tool our team will be using for cleaning the headers/subheader abnormalities and also missing values is the *Tableau Data Interpreter.*

It is also very important to spot outliers in the data. This can be done by using Box Plot in Tableau. This gives us an awareness of how many values contain significant outliers. However, care will be taken in identifying actual outliers.

3. How will you explore your airline's data for meaningful patterns and trends?

For performing exploratory analysis, we use the attribute mentioned above.
Some typical examples of these would be:

Sr. No	Visualization	Purpose
1	Bar Plot	To identify number of flights from origin state to various destination states
2	Scatter Plot	To identify on-time schedule performance of various airlines in a state
3	Geographic Plot/Symbol Map	To analyze which states have maximum delays
4	Geographic Plot/Symbol Map	To identify frequency of visits in different states
5	Heat Map	To identify states with maximum traffic
6	Line Chart	To analyze average delays through various dates

Note: These are tentative examples. These would vary as we proceed with the project.

4. How will you incorporate time and space into your visualizations?

Incorporation of time variables can be done by using bar graphs, line graphs. It is useful to include attributes like date, month, quarter, year to perform the analysis. Also, YTD (Year till date) data can be used to compare various delays/cancellations between different flights for a state. Also, there can comparisons on month-on-month, year-on-year basis for different flights.
For space dimension, we may use Symbol plots/Geographical plots to depict the highest amount of traffic or highest amount of delays or cancellations various states. Air traffics can be plotted by using Heat maps to give insights on how which state possess the maximum traffic.

Team and Work Distribution:

As per the team discussion, the majority of the work will be done by meeting together at least twice every week. Together after discussion and collaboration, the majority of the concepts can be implemented there and then. Since our team consists of only 3 members as of now, it is easy to coordinate the meetings and also distribute work efficiently.

[Besides Tableau and Tableau Data Interpreter, we will use tools like Excel and other ETL tools if need be; especially for cleaning data]