Group 1

Cherry Jain

Priya Harish

Ayushi Gangwal

Abhishek Singh

Kaushik Kompella

Mughundhan Chandrasekar

Sai Krishna Chaitanya Chigili

American Airlines Project

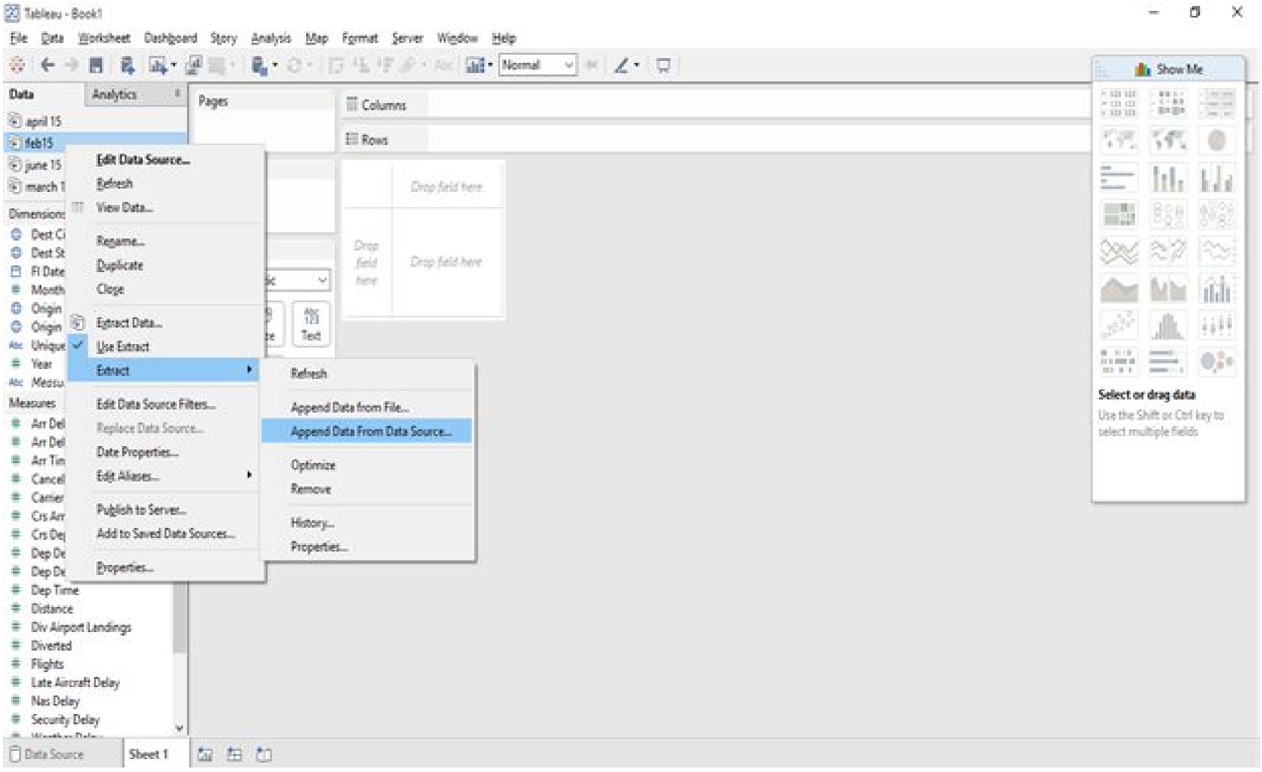
American Airlines Project October 28, 2016

1. **How you will collect and concatenate the monthly BTS datasets.**

- We have selected the timeline and collected airlines data from Bureau of Transportation Statistics (BTS) for twelve months (July 2015 to June 2016) from the link http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time and downloaded the csv file.
- We are assigned American Airlines to prepare visualization on time. Hence we have filtered out the American Airlines data from the downloaded sheet by using Filter option. We will use this to perform on-time performance of the carrier. This procedure is repeated for the span of 12 months.
- On the start page in Tableau, under Connect, we connected Our data sheet. This step created the first connection in the Tableau data source.
- After that using "New Data Source" option from Data menu in Tableau, we obtained 12 monthly reports.

Concatenation of data:

- ➢ To obtain superset of 12 months, use "Append Data from Data source" under extract in Tableau.
- ➢ After extracting all the data files, right click the data source and choose extract -> "Append data from data source" option to concatenate all the files to a single data source file.
- ➢ By this, We concatenate the data into a single database or a file to work on and we can measure our timeline for "on-time" analysis for an entire year.



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

In the group project, we are evaluating and comparing the on-time performance of the American airline flights. As per the BTS dataset provided for the project. We plan to follow the following steps to condition our BTS data set for the American Airlines

Step – 1: Variable Selection

We are classifying and selecting the important variables from the dataset that will help us in analysis of the on-time performance of flights. The variables are selected keeping the following points in mind –

- Variables that give us detailed information about time
- Variables to identify the flight and carriers
- Variables to identify origin and destination airport and city information
- Variables related to arrival, departure, and delay time information
- Variables related to Cancellation and Diversion (Optional – we can more related variables if we find the respective cancellation and diversion flags set to True)

The variables that we have selected are given below –

Variable Name	
Year	Departure Performance
Month	CRSDepTime
DayOfWeek	DepTime
Airline	DepDelay
UniqueCarrier	DepDelayMinutes
TailNum	DepTimeBlk
FlightNum	WheelsOff
Origin	Arrival Performance
OriginAirportID	WheelsOn
Origin	CRSArrTime
OriginCityName	ArrTime
Destination	ArrDelay
DestAirportID	ArrDelayMinutes
Dest	ArrTimeBlk
DestCityName	
Cancellations and Diversions	
Cancelled	
CancellationCode	
Diverted	
AirTime	
Flights	
Distance	
DistanceGroup	

Note: Couple of them might change as we proceed in the project in future

Step – 2: Handling Missing Values

In this step, we will explore the data set for the variables we have selected to see if they have missing values. Our approach would be as follows –

- Look for missing values in the variable related to Origin and Destination city names, if any then we can try to impute them using lookup tables for Origin and Destination Airport IDs.
- Missing values in the variables such as Arrival and Departure Delay Minutes can be imputed using Arrival and Departure Delay variables
- If there are missing values in any of the other variables, then we would discard those rows from the data set as these values won't give the true picture of the airline's performance even if we use the best imputation technique

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

Since the data is now collected, concatenated and conditioned well as per our project requirements, now it is possible to illustrate the performance of American Airlines using several visualization techniques. We shall perform an exploratory data analysis on the American Airlines dataset, based on the parameters such as AirTime, Distance, DepTimeBlk, DepDelayMinutes etc.

S.No	Exploration	Parameters or Methodologies Involved
1.	The busiest hour of the day shall be identified	DepTimeBlk/ArrTimeBlk
2.	The busiest airlines shall be identified	Number of flights being operated in a day
3.	Analyze distribution of flights in several regions. Identify the region with maximum traffic.	Scatter Plots, Heat maps and other visually appealing graphs
4.	The total no. Of delays or diversions (monthly) shall be analyzed	DepDelayMinutes, ArrDelayMinutes
5.	Air-Miles and the no. Of regions covered by American Airlines shall be computed. This shall be further used to identify the largest route covered by the airlines.	AirTime, Distance

We shall build an interactive dashboard which would enable the user to infer the performance of the flights depending upon several parameters (as mentioned above).

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

Implementing time and space into our visualizations:

For visualizations with time variables we can use line graphs and bar graphs. We shall be using time variables like year, month, day and Flight date. Also, we would use variables having information regarding departure time, arrival time, departure delay, arrival delay, total delay. All the above variables can be added to year/month/day/date as required respectively. Thereby various graphs like scatter plots can be plotted against various airlines and flights.

We can use geographic maps for plotting routes for different flights and airlines using origin city and destination city variables. Such graphs will help in identifying the shortest routes, monitor incoming and outgoing air traffic, monitor air traffic for certain time for a certain airport, identifying peak times of the airlines, identifying busy routes and frequently delayed flights. We can also highlight the cities the airline runs its operations and calculate the arrival/departure performance with respect to other cities. Also, we can compare one airlines with other airlines and rate the arrival/departure performance with respect to each area.