Tableau Flight Data Project

William McKee **September 2017**

Data Set Links:

The first version of the flight data visualization can be found here.

The final version can be found at this location.

Summary:

This investigation looks at a data set for flight delay data for United States domestic flight arrivals from the Bureau of Transportation Statistics website. This report covers the period from January 2007 through December 2016. The report examines flight delays by breaking them down into type. Special attention is given to the most frequently travelled airports and the most common carriers, especially those with below average on-time record.

A flight is considered delayed when the flight arrives at the destination gate at least 15 minutes later than the scheduled time.

There are five flight delay causes defined in the data set:

- **Carrier**: Delays caused by a problem under the airline's control, including maintenance issues, airplane cleaning, baggage loading, and fueling
- **Weather**: Delays are due to significant weather events such as thunderstorms, snow storms, hurricanes, and/or tornadoes
- **National Aviation System (NAS)**: Delays attributed to overall system conditions including heavy airline traffic, airport operations, and air traffic control
- **Late-arriving aircraft**: Delay occurs because a previous flight using the same aircraft arrived late
- **Security**: Delays due to terminal breaching, security incident on the plane, or security lines above 29 minutes

Each row in the data set contains the following information:

- **Year**: 2007 to 2016 in this data set
- **Month**: 1 through 12
- **Carrier**: Airline two-letter or number code (e.g., AA)
- **Carrier Name**: Airline full name (e.g., American Airlines)
- **Airport**: Airport's three letter code (e.g., PHL)
- **Airport Name**: Full Airport name (e.g., Philadelphia International Airport)

- **Arr_flights**: Total arrivals at the airport for a carrier
- **Arr_del15**: The number of flights delayed at least 15 minutes
- **Carrier_ct**: The number of flights delayed due to carrier issues
- **Weather_ct**: The number of flights delayed due to weather issues
- **Nas_ct**: The number of flights delayed due to national air system issues
- **Security_ct**: The number of flights delayed due to security incidents
- **Late_aircraft_ct**: The number of flights delayed due to late arriving aircraft
- **Arr_cancelled**: The number of cancelled flights
- **Arr_diverted**: The number of flights diverted to airports other than the destination airport
- **Arr_delay**: The number of minutes for delayed flights
- **Carrier_delay**: The number of minutes for carrier delayed flights
- **Weather_delay**: The number of minutes for weather delayed flights
- **Nas_delay**: The number of minutes for national air system delayed flights
- **Security_delay**: The number of minutes for security delayed flights
- **Late_aircraft_delay**: The number of minutes for late aircraft delayed flights

A flight delay may be given multiple causes and is counted fractionally under each cause. For example, a delay blamed equally on the carrier and the weather is counted as 0.5 carrier and 0.5 weather. The flight delay minutes are then divided in half between these two delay types.

Each row of data represents one month of flight delay data for one carrier and one airport (e.g., January 2013 for Delta Airlines at Chicago Midway Airport).

Initial Design:

I started the data exploration by looking at the yearly flight delay counts and flight delay minutes. I decided to enhance my understanding of the yearly flight delay data by adding the following fields:

- I calculated on-time, delayed, cancelled, and diverted flights as a percentage of all flights
- I added calculations of carrier, weather, NAS, security, and late arriving flight delay counts as a percentage of all flight delay counts
- I also added calculations of each delay type's minutes as a percentage of all flight delay minutes
- Finally, I computed the average flight delay times for each type of delay

I have the following Dashboards in the initial visualization. The chart types are listed in each Dashboard's description. The initial visualization can be found here.

- **Flight Status Percentages by Year**: This dashboard displays information via text tables and stacked bar charts. The information shows the percentage of on-time, delayed, cancelled, and diverted flights each year. A user can click on an entry on the chart and highlight the corresponding entry in the table (and vice versa). The stacked bar chart shows the on-time percentage for each year very well.
- **Delay Type Percentages by Year**: This dashboard also displays percentages – this time breaking down the delays by the five delay types. Text boxes and stacked bar charts are used here, like the first dashboard.
- **Average Flight Delay by Year (in Minutes)**: This dashboard displays the average number of minutes per delay type each year. Since this is a measurement over time, a line graph is shown. One can see which delay type averages the longest number of minutes each year.
- **Delay Type Minutes Percentage by Year**: This dashboard displays percentages of the flight delay minutes attributed to each delay type. One again, a stacked bar chart is used to show each year's breakdown.
- **Delays for Airports with at least One Million Flights**: This dashboard compares the flight delay counts against the total number of flight arrivals over the ten-year period for frequently travelled airports. Since two variables are compared, a scatterplot is used. There is a color coding comparing airports with on-time records above and below average.
- **Delays for Top 10 Carriers**: This dashboard is like the previous one, measuring the same two variables for the top carriers. Same chart type is used.
- **Please note**: On the last three dashboards, only carriers and airports with below average on-time records are examined.
- **Average Flight Delay Comparison (in Minutes)**: This dashboard shows the audience the average flight delays for both airports and carriers with below average on-time records. Highlight tables are used to differentiate between the lowest and highest average flight delays.
- **Carriers Flight Delay Percentages by Type**: On this dashboard, the flight delays for the carriers are broken down by type – with a percentage breakdown. A stacked bar chart is used to compare the percentage breakdown every year.
- **Airports Flight Delay Percentages by Type**: This dashboard, like the previous one, breaks down the delay types by percentage. Airports are measured this time. Once again, a stacked bar chart is used.

Feedback:

My wife is a research scientist at the University of Pennsylvania in Philadelphia. She presents data every two weeks and attends presentations at least once per week at her job. She was not familiar with the data set before looking at my first presentation. She provided lots of valuable feedback. She made the following points:

1. Each description, at the top of the slide, only needs one or two sentences.
2. Charts plus text boxes are redundant, especially since one can hover over a bar and get the data values.
3. X and Y axis labels must be updated. Tableau's "Value" does not adequately describe the data. Some axes can have better titles (e.g., "Arr Del15")
4. Some legends have redundant information in title and color labels. Remove titles or change field names.
5. Dashboard #4 confused my wife. I need to distinguish flight delay counts versus flight delay minutes better. If this slide does not provide more valuable information, I will eliminate the slide.
6. Keep the table for dashboards #5 and #6 – with the scatterplot. Update the title for slide #5 and include color legend for above and below on-time record airports and carriers. Consider larger scatter points for better on-time records.
7. The seventh slide ("Average Flight Delay Comparison (in Minutes)") could benefit from a heat map in lieu of a highlighted table. Explore other color schemes such as blue-white-orange.
8. "Grand Total" should be labelled "Average" since average is depicted in the tables.
9. In some cases, information from a later slide refers to information in an earlier slide. Could I provide a link or refer to previous slide effectively?
10. Add a slide explaining the delay categories.

Final Design:

Following feedback, I made several updates to the slides, improving the data to ink ratio. The titles of each story point have been updated. The improved version of the visualization can be found at this location.

- **Flight Status Percentages**: This dashboard displays information on a stacked bar chart. The information shows the percentage of on-time, delayed, cancelled, and diverted flights each year. A user can click on an entry on the chart to see the exact percentage value. The table was removed.
- **Terminology**: The meaning of flight delay and the delay types are explained to the audience on this slide. This is the additional slide my wife suggested.
- **Delay Type Count Percentages**: This dashboard also displays percentages – this time breaking down the delays by the five delay types. A stacked bar chart is presented, like the first dashboard. The table was removed.
- **Average Flight Delays (in Minutes)**: This dashboard displays the average number of minutes per delay type each year. Since this is a measurement over time, a line graph is shown. One can see which delay type averages the longest number of minutes each year. The table was removed.

- **Delay Type Minutes Percentages**: This dashboard displays percentages of the flight delay minutes attributed to each delay type. One again, a stacked bar chart is used to show each year's breakdown. The table was removed.
- **Delays for Top 20 Airports**: This dashboard compares the flight delay counts against the total number of flight arrivals over the ten-year period for frequently travelled airports. Since two variables are compared, a scatterplot is used. There is a color coding comparing airports with on-time records above and below average, explained by the legend.
- **Delays for Top 10 Carriers**: This dashboard is like the previous one, measuring the same two variables for the top carriers. Same chart type is used.
- **Please note**: On the last three dashboards, only carriers and airports with below average on-time records are examined.
- **Average Flight Delays (in Minutes)**: This dashboard shows the audience the average flight delays for both airports and carriers with below average on-time records. Heat maps are used to differentiate between the lowest and highest average flight delays. A legend aids the audience's understanding of the heat map.
- **Carriers Flight Delay Percentages by Type**: On this dashboard, the flight delays for the carriers are broken down by type – with a percentage breakdown. A stacked bar chart is used to compare the percentage breakdown every year. The table was removed.
- **Airports Flight Delay Percentages by Type**: This dashboard, like the previous one, breaks down the delay types by percentage. Airports are measured this time. Once again, a stacked bar chart is used. Once again, the table was removed.

Resources: I did not consult with any Web sites, books, forums, blog posts, nor GitHub repositories in the submission of this work