Flight Delay Dataset - Summary

Author: Oliver Kröning

Date: 18/11/2018

Tableau Story URL:

https://public.tableau.com/profile/oliver.kr.ning#!/vizhome/Udacity_DAND_Tableau_Story/DAND_Fli
publish=yes

This report documents the analysis and visualization process of a flight delay dataset within the framework of Udacity's Data Analyst Nanodegree Project 08: "Create a Tableau Story".

At first, I want to present the dataset and the questions which should be answered in this project using data visualizations via Tableau. Afterwards, the approach to analyze the research subject is described. On the basis of this approach, defined steps are performed and expored results are visualized in the following section. In the last section, the received users' feedback is presented and the next steps to remedy defiencies are described.

Dataset

Airlines that have 0.5 percent of total domestic scheduled-service passenger revenue report on-time data together with the causes of delays and cancellations to the Bureau of Transportation Statistics since 2003. Reported causes of delay are available from June 2003 to the most recent month. In general, there is a delay, when the aircraft arrives later than 15 minutes than the official time of arrival.

The data are given for each month and combination of carrier and airport. Therefor, the dataset only considers arrival delays. We also obtained geographical data to show spatial characteristics. Furthermore, the causes of delay are reported in broad categories that were created by the Air Carrier On-Time Reporting Advisory Committee. The categories are Air Carrier, National Aviation System, Weather, Late-Arriving Aircraft and Security. The causes of cancellation are the same, with the exception of the late-arriving aircraft category:

- Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.
- Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas. (Source: www.bts.gov)

The causes of delays are given as a float to describe the proportion of the absolute number of delays. Furthermore, the overall delay as well as the separate causes are presented in minutes.

Questions

To start with the analysis, we want to focus on the following questions:

1. What are the different causes of flight delays/cancellations/diversions and which causes are more frequent in general?
2. How are the causes of delays/cancellations/diversions temporally distributed? Are there differences for certain months or seasons?
3. Which carriers cause most of the delays? Are there any (temporal) correlations between the amount of flights and the proportion of delays?
4. Which airports cause most of the delays? Are there any relationships with certain carriers? What are the reasons for these delays?
5. Which are the top 10 largest airports as measured by flight count? Does the size of the airports affect the proportion of delays and does the proportion of delay causes differ from the general overview?

Methodology

Within the visualization, we start with a global illustration of the average proportions of flight delays, cancellations and diversions. We also add a date filter to explore changes between different years. In case of a delay, another sheet is created to explore the average proportions of delay causes. For both sheets, calculated fields containing the proportions have to be created. We decided to use pie charts to visualize the general proportions.

All sheets apply the same color code within the visualization:

- Blue = on-Time
- Yellow = delayed
- Orange = diverted
- Violet = cancelled
- Yellow Green = Carrier delay
- Pink = Security delay
- Light Grey = NAS delay
- Light Brown = Late Aircraft delay
- Dark Grey = Weather delay

The next sheets handle temporal relationships of delays, cancellations and diversions as well as the cause of delay. First, we take a look at the absolute number of all flights over time in a line plot. The date is presented continuously from 2010 to 2018 to show temporal changes developments in number of flights. Furthermore, an aggregation of the count of flights per month is plotted over discrete months to clarify seasonal dependencies.

Carriers and their likelihood of being late are handled in the next sheets. Here, we compare all carriers and sum up the minutes of delay. This statistic is not very meaningful, since the number of flights differ a lot between the carriers and, hence, carriers with a large number of flight also have a large number of delay minutes or cancellations. Thus, we look at the average proportion of delay flights compared to the total number of flights. This, is also the measure for the following sheets to analyze carriers with a high probability of delays. Another visualization shows the development of the delay proportion over time for each carrier in discrete line plot with different carriers ordered

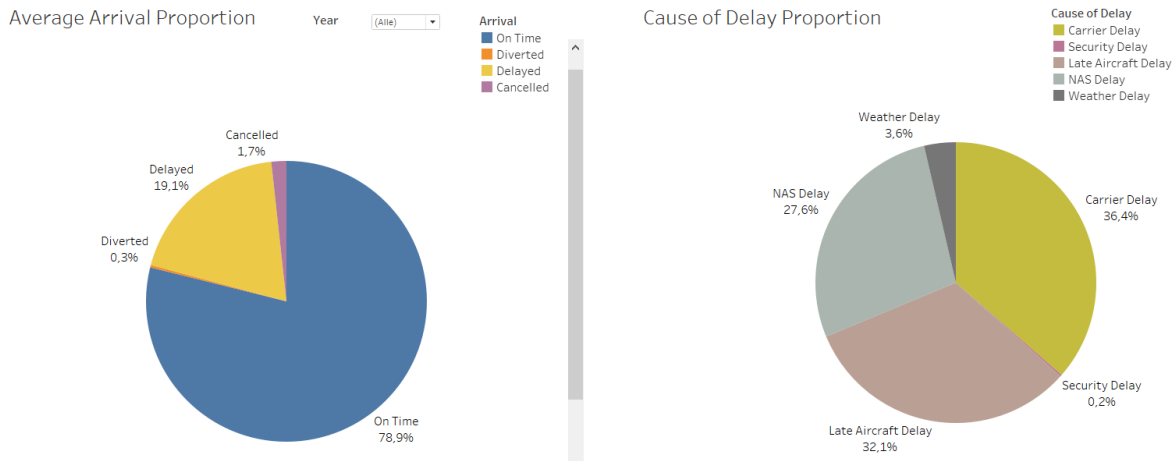
vertically in several diagrams. Here, the plots show the data yearly and monthly to visualize the relationships continuously within the range of time and seasonally for the different months of the year.

This delay proportion can also be spatially analyzed by viewing the airports in a map plot. The information of the number of flight for each airport are described by the size of the bubbles. The color represents the proportion of flights that are on time. A color spectrum from red (low percentage of on time flights) and blue (high percentage of on time flights) is chosen to emphasize the differences within the data. Here, we have chosen a range of 50% to 100% to have a good contrast between the values. Furthermore, a filter is implemented to filter for carriers and years. Thus, we can detect bad combinations of certain carriers and airports and their temporal developments. Another diagram connected to the bubble map shows additionally the causes for not being on time. Here, the proportions of the delay causes as well as the proportions of cancellations and diversions are presented in a bar chart.

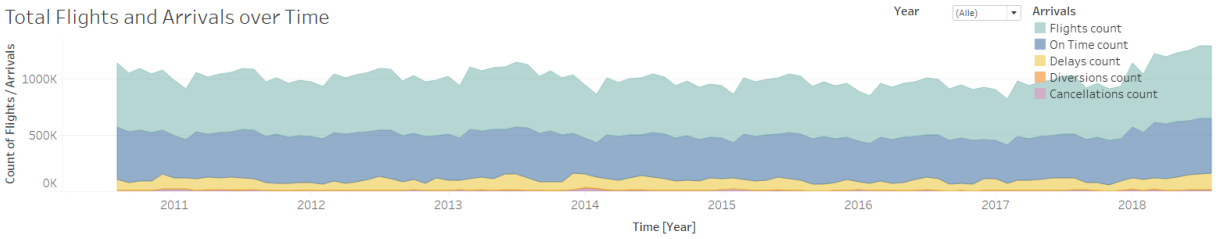
The final sheets visualize the top 10 airports and top 10 carriers measured by flight count. At first, we have to create two sets in dependence of the number of flights and apply them to our sheets as a filter. To find reliable airports and carriers, we create a boxplot for the on-time proportion of the airports. To consider the number of flights, we want to visualize how many flights are operated with a certain carrier. So we can find good/bad combinations of airports and airlines. Furthermore, we want to analyze the causes of the delays, that occured at certain airports.

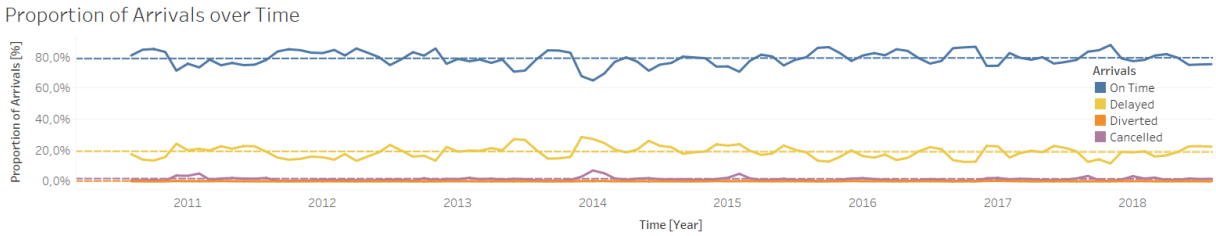
Results

The general overview shows that around 78% of all flights are on time. The remaining flights are delayed (around 19%), cancelled (around 1.7%) or diverted (0.3%). The proportion of on time flights fluctuates between 75% (2014) and 81.5 (2016). Most of the delays are due to circumstances within the airline's control (36.4%) and the National Aviation System (27.6%). Another large proportion of flights are delayed because of a late arrival of a previous flight causing a late departure of the airplane (32.1%). Only a few delays can be explained by meteorological or security circumstances. These results are our reference for the further detailled analysis, so differences and variations can be identified.

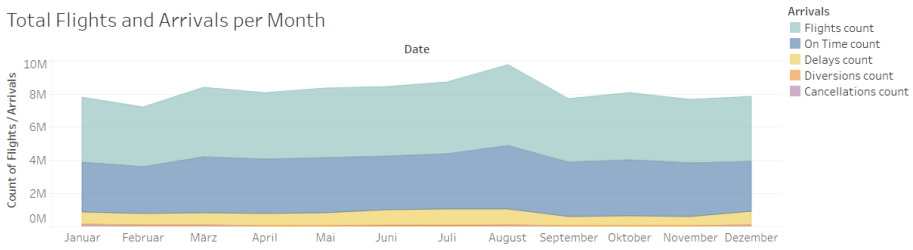


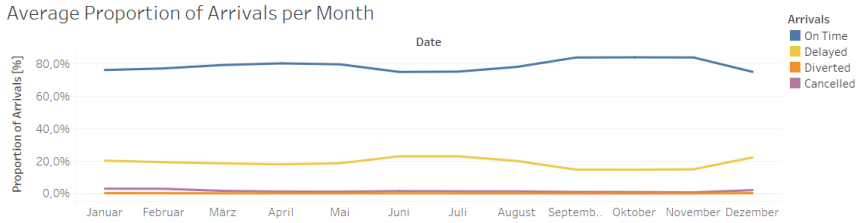
In the area chart, the total amount of flights as well as the arrivals are visualized over time. We might suggest that the number of flights periodically changes since there are minima for each February and maxima in summer months. The total number of flights has increased since 2017. The arrivals follow this trend. The proportions of arrivals visualized in the line plot below. are nearly fixed and not affected by this increase. However, seasonal flutuations can be identified.



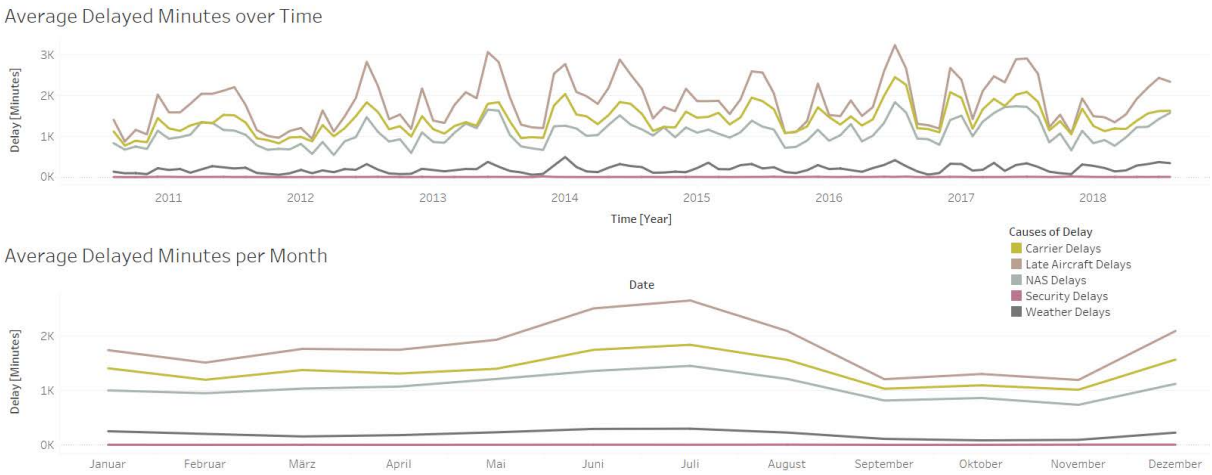


Seasonal changes can be visualized by using discrete time plots. In this case, an area and time plot is applied to count all flights and arrivals and calculate the average of the arrival proportions for each month of the year. We can see that there is an increase of flights during the summer months. This might be due to the holiday season, when more people use aircrafts to go on vacation. This causes also an increase of the delay rate within summer. Additionally, the on time rate drops at the end of the year in conjunction with a rise of delayed and cancelled flights. However, the total number of flights is nearly constant in November, December and January. This might be a result of a higher volume of passengers due to Christmas time.

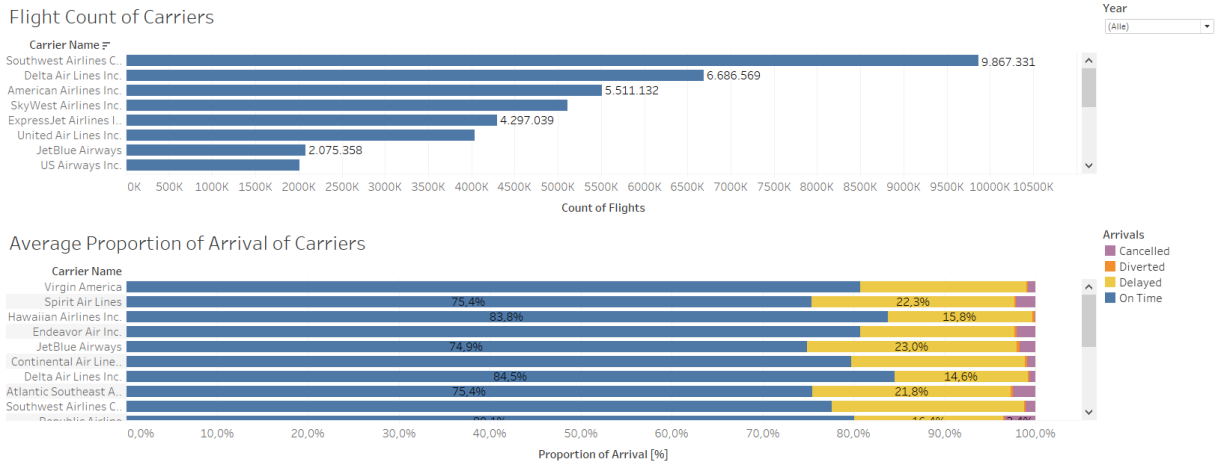




These line charts show us the distribution of occurrences of the several delay causes measured in delayed minutes over time. It is obvious that every delay cause follows the same overall trend. There is a seasonal relationship, that was also described on the previous slides. Here, the increase of each delay cause in the summer months and winter months is clarified. The peak in the summer months might be a result of the large amount of flights maybe because of the holiday season, which can also be seen in the previous slide. Since the number of total flights is nealy constant during the winter months, the increase in delay minutes in december might be a result of the bad weather conditions (which also could affect a delayed aircraft departure) and a higher volume of passengers during the Christmas season. An increase of carrier delays in this period of time might also a result of an additional burden within baggage loading.

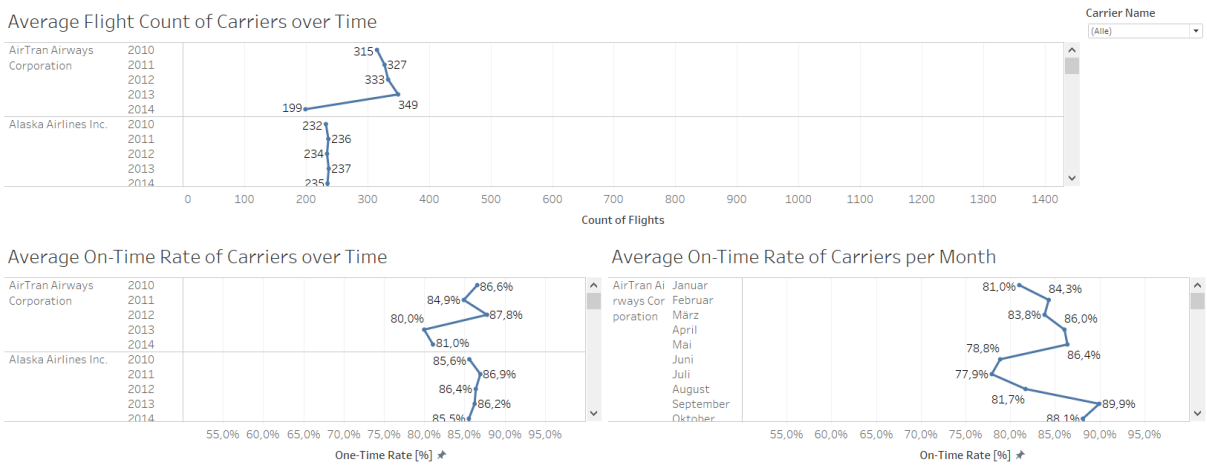


Let's take a look at the carriers o find relationships between delays / delay causes and certain carriers. Each carrier operates a different number of flights each month. There are carriers with over 1,000,000 arrivals per year like Southwest Airlines Co. in 2013 and small carriers with an overall flight count of around 65,000 like Allegiant Air or Comair Inc.. Thus, an analysis of the overall delay minutes or cancellation count is not significant. A visualization of the proportion of on-time arrivals in combination with delays, cancellations and diversions gives us a better understanding of the punctuality and reliability of carriers. In the proportional bar chart below, the highest overall rate of on-time arrivals is achieved by Alaska Airlines Inc. and Delta Air Lines Inc. with around 85%. The latter carrier has also the second highest number of flights. Other carriers with a high number of flights have a moderate delay rate of around 77% like Southwest Airlines Co. The lowest overall on-time percentage is achieved by PSA Airlines Inc., a carrier with only 185,000 flight in the recent 8 years. Furthermore, 4,5% of flights of this carrier were cancelled, which is the highest cancellation rate in this dataset.

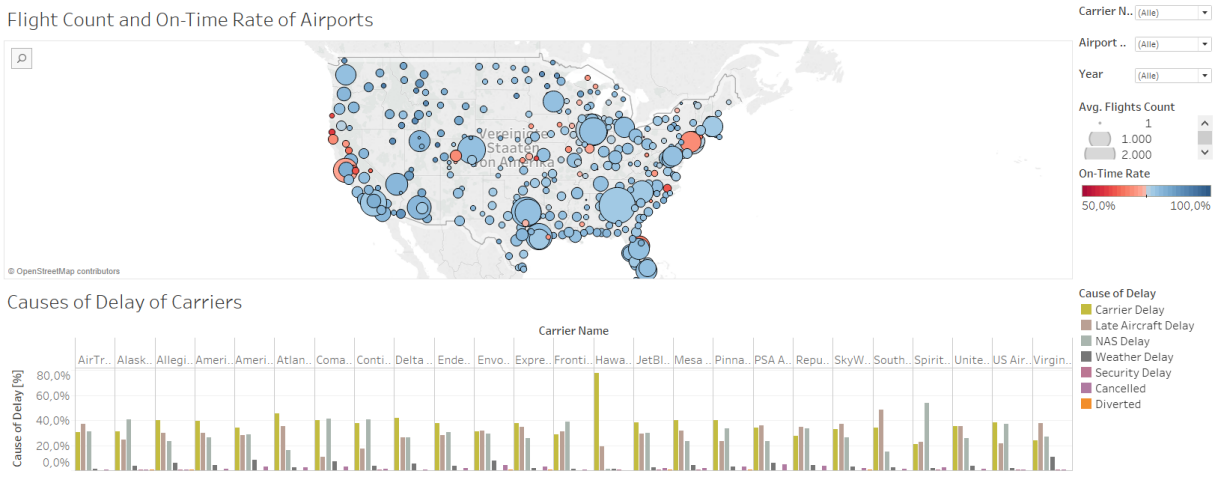


To visualize a temporal development of each carrier, we plot the monthly average flight count over time as well as the proportion of on-time arrivals per discrete year and month in line charts. The filter function within this dashboard helps us to compare carriers and their development within the recent 8 years. The flight count chart gives information about the success of the carrier. Airlines like JetBlue Airwys or Hawaiian Airlines Inc. operate a constant number of flights. However, other carrier like ExpressJet Airlines Inc. have suffered a massive drop of flights. There are also carriers that have significantly increased their number of flights like American Airlines Inc. in 2015 and 2016. The on-time proportion shows us whether this development can be connected to the reliability of the carrier. An interesting correlation can be detected for Southwest Airlines Co., which suffered a major drop in flights count in 2013 and 2014. During these years, this carrier had also the lowest on-time proportion within the recent 8 years. In contrast, the positive development of the number of flights of United Air Lines Inc. in 2011 resulted in a massive drop of around 15% of on-time arrivals within the following years. In the monthly overview of on-time arrival

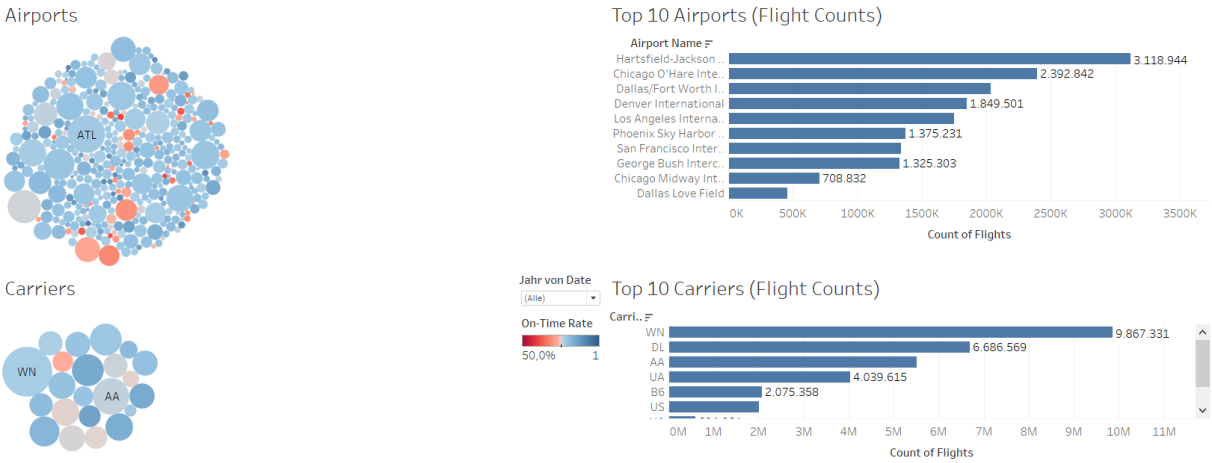
percentages, the described seasonal trend within summer and winter months can be observed for almost every carrier. Note that there are data missing for some years or month for some carriers.



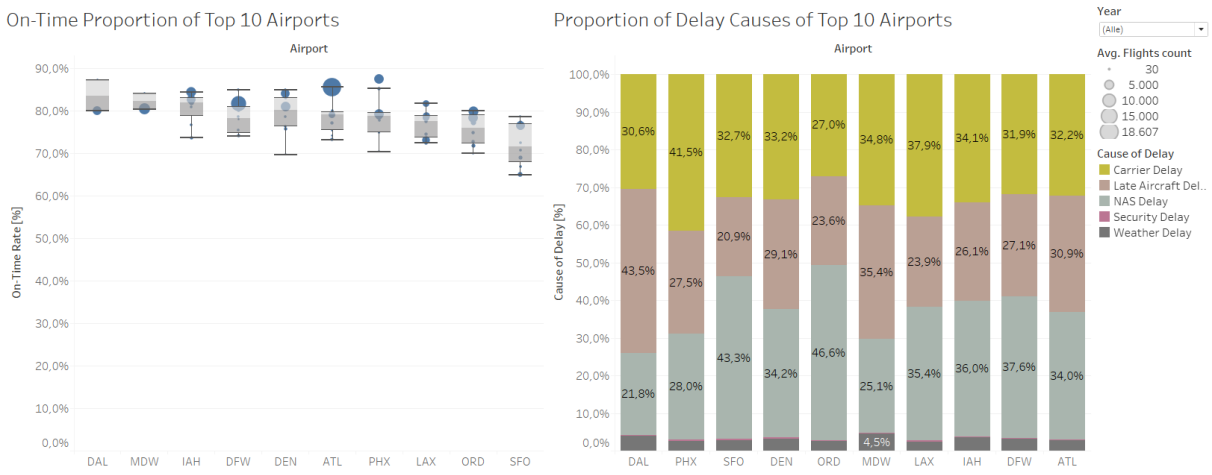
A spatial overview of the the airport's flight count in combination with it's on-time arrival proportion gives us information about the correlation of the airport's size and the amount of delayed or cancelled flights. Surprisingly, most of the larger airports achieved a mediocre or even very good rate of on-time arrivals. Exceptions are Newark Liberty International (69.1%) and Orlando Sanford International (68%), which are far below the overall average of 78.9%. It is striking, that especially small airports at the east and west coast have a high rate of delays and/or cancellations like Southwest Oregon Regional (61.4%), Jach McNamara Field (59,9%) or Pitt Greenville (62.7%). Another surprising aspect is the distribution of delay causes for each airport also in combination with certain carriers. E.g., Newark Liberty International shows a very high rate of NAS delays (> 60%) for almost all carriers. However, Houston's William P Hobby airport has a higher rate of carrier delays. Additionally, the cause of delay depends on the carrier. E.g. 78.1% of the delayed flights by Hawaiian Airlines Inc. are caused by the carrier. Thus, airports operating flight's of this carrier also suffered from this kind of delay.



Let's take a look at the top 10 largest airports and carriers, which can be identified by the total number of flights. The airports with the highest number of flight within the recent 8 years are Hartsfield-Jackson Atlanta International and Chicago O'Hare International. We can see that most of the large airports have an average rate of delays respectively mediocre proportion of on-time arrivals. An Exceptions is San Francisco International (71.8%). A similar trend can be observed for the carriers. Airlines with a large number of flights have a high proportion of on-time arrivals like Delta Air Lines Inc.. The top 10 airlines and carriers can be found on the right. We will inspect them in detail in the following slide.



The box plot on the left shows the top 10 airports and their on-time arrival proportion, when we consider the top 10 carriers. Thereby the size of the bubble indicates the number of flights operated by a certain carrier. The highest median in on-time rate is achieved by Dallas Dallas Love Field (83.6%) followed by Chicago Midway International (82.2%). The lowest rate of the top 10 airports can be detected at the San Francisco International (71.6%). It is noticable that the carriers that operate most of the flights for certain airports seem to achieve a higher on-time rate. The causes of delay differ for each airport, but as shown before the top 3 are late aircraft delays, carrier delays and NAS delays.



Feedback

Feedback for version 1 of the Tableau story given by Amelia P.:

1. Be careful about the information quantity for each dashboard

- I splitted one sheet into two separate sheets to distribute the information a little bit.

2. All charts must be labels and axis descriptions and them cannot be cut

- I labelled all charts and added a axis description.

3. The chart descriptions can be more summarized. Use the report to put all text.

- I shortened the chart description for all sheets and put the findings in the report above.

4. Put titles in all charts. In the third dashboard, for example, is not clear which information we have in the charts

- I put titles in all charts of the story to clarify the presented information.

5. In dashboard 6 we have green and red. We should avoid using the colors green and red together because colorblind people will have difficulty analyzing the graph. See this article on this: https://www.tableau.com/about/blog/2016/4/examining-data-viz-rules-dont-use-red-green-together-53463"

- I used new colors as described in the color code above.

Feedback for version 1 of the Tableau story given by Manar J.:

"you have done hard work on your story, it is great work, I have two points to improve the design, the first point is that I think better to put the legends to be clearer by show them without the list, and the second point is that you put labels and axis ticks in the bar charts and one of the line charts, I think you can hide the ticks of the axis or remove the labels. I hope that will be useful for your story, with best of luck in your project"

- I relocated the legends and tidied up the sheets.
- I labelled all charts and axis. I removed the labels for the bar chart in the map plot slide. Here, the label was not useful since it was only partially visible and the color legend on the right gives the same information.