What is the worst airline company to travel with (powered by data)?

By Thomas Ricci - Sit-2a
22.01.2018

Introduction

First things first, we got to have the names of every company in the world. Well that isn't really possible since they all move around airports, and are bought, sold constantly this isn't feasible for the small amount of time we have.
So from that point I decided to care more about finding the data and see which companies are the most frequent.

Stage 1 - determine what criteria defines "the worst airline"

There are numerous things we can factor in. But we have to narrow it down to the data we can find online, and not just rumors, and news feeds. So now defining the worst airline would include:

- Customer satisfaction reports
- Ticket prices raise for each flight
- Lost luggage
- Delays, and frequencies of these
- Total flight counts

Now that's alot to go through if we were to get every airline in every single country. The best found on the internet, with enough data is in the United States.

Since most of the data found for this experiment was for US travel, the question now becomes "What is the worst airline to travel with in the US?"

That still doesn't narrow down enough data. That is because every airport most likely has a customer report, this means going through roughly 5000 airport survey sites, and finding by hand the data for each one, which isn't feasible in the time frame allocated to this project.

Next up: Ticket prices. We can't use that neither because the price of a ticket varies through time, depending on the time of registration. This requires monitoring ticket prices for every flight over the globe, the data most likely exists, but won't be free.

The only thing we need to plan now is how we are going to record and be able to determine which is the worst airline. We will need to score every criteria we have and leverage that to determine “which is the worst airline to travel with in the US?”.
There will be a problem with just raw counting, this doesn’t account for smaller airlines. This means that a Phoenix Air might not have the same amount of list luggage than American Airlines, and more delays. We need to scale all of the aggregated data on each airline based on the number of total flights made by the company.

So now we have flight delays, and lost luggage that will determine the worst airline in the US.

Stage 2 - Exploring and finding the data

Let’s start the digging.

Flight delay data

There is data available on https://www.flightstats.com/v2. The problem is that the realtime tracker is free, but not the historical data, which is what we need now. The historical data is to be paid for, and would cost around 5,000$ just for the export of 2017.
Instead there is a special bureau in the US for transportation : The “Department of Transportation”. This bureau has put out in 2017 the dataset on flights across the world.

Here is a link to the latest dataset https://www.transtats.bts.gov/DL_SelectFields.asp

TSA claims

Then comes the lost luggage. There is a statistic put out by the TSA (Transportation Security Administration) every year with all lost luggage and if it was claimed. This gives us insight on how much bagages has been misplaced, and by which company.

here is a link to the dataset https://www.dhs.gov/tsa-claims-data

Stage 3 - Normalizing & sanitizing the data

Our objective is to find the worst airline company, this means filtering through each document and reducing the number of airlines available. This means normalizing airline names. Airline names can contain the company type (e.g Amazon Co), we don’t need that piece of data.

Next up you need to sanitize the headers in all of your CSV files. Make sure that all the headers don't contain any spaces. The airline flight delay csv file headers contains spaces, please be sure to trim them.

Stage - 4 Graphs

The final graphs concern only a subset of the available data. All graphs were plotted using mattplot lib in python. All graph plotting can be found in the code/plot.py

Graphs are available in the **figures** folder.

Data and where you can find it

All data was exported from web portals.

The TSA claims was exported from `https://www.dhs.gov/tsa-claims-data`
and imported on 12.01.2018

The flight delays was exported from https://www.transtats.bts.gov/Fields.asp?Table_ID=236
and imported on the 12.01.2018

Conclusion

We can't truly determine which airline is the worst. We haven't determined if claims or delay is more important. Than there is a problem with the data.

If we had to draw a brief conclusion on which airline is the worst:

- "Delta Air Lines" has lost the most luggage
- "Southwest Airlines" has had the most delays

Both datasets might not be complete. Secondly not every airline is reportorated in the TSA claims report, that is if the airline screen their luggage privatly and so aren't reported in the report.

The transportation bureau must have alot of the data, but there might not be everything. Considering the fact that flightstats.com, when exporting data, shows a total number of rows of over 50'000 only for 2017. And the files provided by the Bureau Of Transport provides only 10'000 rows for 2017. This biases alot the analysis

Plus we are missing a crucial piece of data: The number of flights per airline. This changes completly this analysis, and allows for a much better conclusion. Having the amount of flights

per airline allows the data we found for delays to be scaled. This means that big companies with alot of flights, therefor alot of delays won't be penalised.

Sources

TSA claims
https://www.dhs.gov/tsa-claims-data

Delays
http://ourairports.com/data/

Flights
https://www.transtats.bts.gov/DL_SelectFields.asp
https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time
https://www.transtats.bts.gov/Fields.asp?Table_ID=236