Client Report - Project 2   </> See code ▾

Course DS 250

AUTHOR
Alejo Alegre Bustos

▶ Show the code

Speech

Mrs. Comeau, I want to share some observations from my analysis of the 2013 flight data. I've been looking at delays across different airports and how the weather impacts them. For example, at San Francisco International Airport, 1 in 10 flights is delayed due to weather, which is the highest average in the dataset. In contrast, at Hartsfield-Jackson Atlanta International Airport, which handles more flights, the delays due to weather are less than 1%. This analysis highlights the variability in delays and the influence of weather. Using tools like Let's Plot and pandas, we can create interactive visualizations that make it easier to understand these trends, ultimately improving our decisions and business strategies.

QUESTION 1

Fix all of the varied missing data types in the data to be consistent (all missing values should be displayed as "NaN"). In your report include one record example (one row) from your new data, in the raw JSON format. Your example should display the "NaN" for at least one missing value.

▶ Show the code

```
{
    "airport_code":{
        "0":"ATL"
    },
    "airport_name":{
        "0":"Hartsfield-Jackson Atlanta International"
    },
    "month":{
        "0":"January"
    },
    "year":{
        "0":2005.0
    },
    "num_of_flights_total":{
```

```
            "0":35048
        },
        "num_of_delays_carrier":{
            "0":1500
        },
        "num_of_delays_late_aircraft":{
            "0":NaN
        },
        "num_of_delays_nas":{
            "0":4598
        },
        "num_of_delays_security":{
            "0":10
        },
        "num_of_delays_weather":{
            "0":448
        },
        "num_of_delays_total":{
            "0":8355
        },
        "minutes_delayed_carrier":{
            "0":116423.0
        },
        "minutes_delayed_late_aircraft":{
            "0":104415
        },
        "minutes_delayed_nas":{
            "0":207467.0
        },
        "minutes_delayed_security":{
            "0":297
        },
        "minutes_delayed_weather":{
            "0":36931
        },
        "minutes_delayed_total":{
            "0":465533
        }
    }
}
```

QUESTION 2

Which airport has the worst delays? Describe the metric you chose, and why you chose it to determine the "worst" airport. Your answer should include a summary table that lists (for each airport) the total number of flights, total number of delayed flights, proportion of delayed flights, and average delay time in hours.

According to my calculations, Jackson Atlanta International Airport has the highest proportion of delayed flights. Specifically, for every 100 flights, 6 experience delays at this airport, which means 6% of flights are affected, although flights can't be split. Additionally, the average delay duration is approximately 30 minutes.

▶ Show the code

|	airport_name	airport_code	total_num_of_flights	total_num_of_delayed_flights	proportion_of_delayed_flights	ave
0	Hartsfield-Jackson Atlanta International	ATL	4430047	902443	0.06	0.2
1	Chicago O'Hare International	ORD	3597588	830825	0.05	0.3
2	San Francisco International	SFO	1630945	425604	0.03	0.1
3	Denver International	DEN	2513974	468519	0.03	0.1
4	Salt Lake City International	SLC	1403384	205160	0.01	0.0
5	Washington Dulles International	IAD	851571	168467	0.01	0.0
6	San Diego International	SAN	917862	175132	0.01	0.0

QUESTION 3

What is the best month to fly if you want to avoid delays of any length? Describe the metric you chose and why you chose it to calculate your answer. Include one chart to help support your answer, with the x-axis ordered by month. (To answer this question, you will need to remove any rows that are missing the Month variable.)

▶ Show the code

To answer these questions, I selected the following three categories of delays: carrier delays in red, delays due to late aircraft in yellow, and NAS delays in blue. These are the most common types of delays at every airport and the ones with the highest values in this dataset. I calculated the z-score for each delay type; in simple terms, z-scores represent how high or low each delay value is compared to its average. According to this chart, the best month to fly is November, as it shows the lowest average values for carrier delays, delays due to late aircraft, and NAS delays.

▶ Show the code

Z-Scores by Month

November: the best month to fly
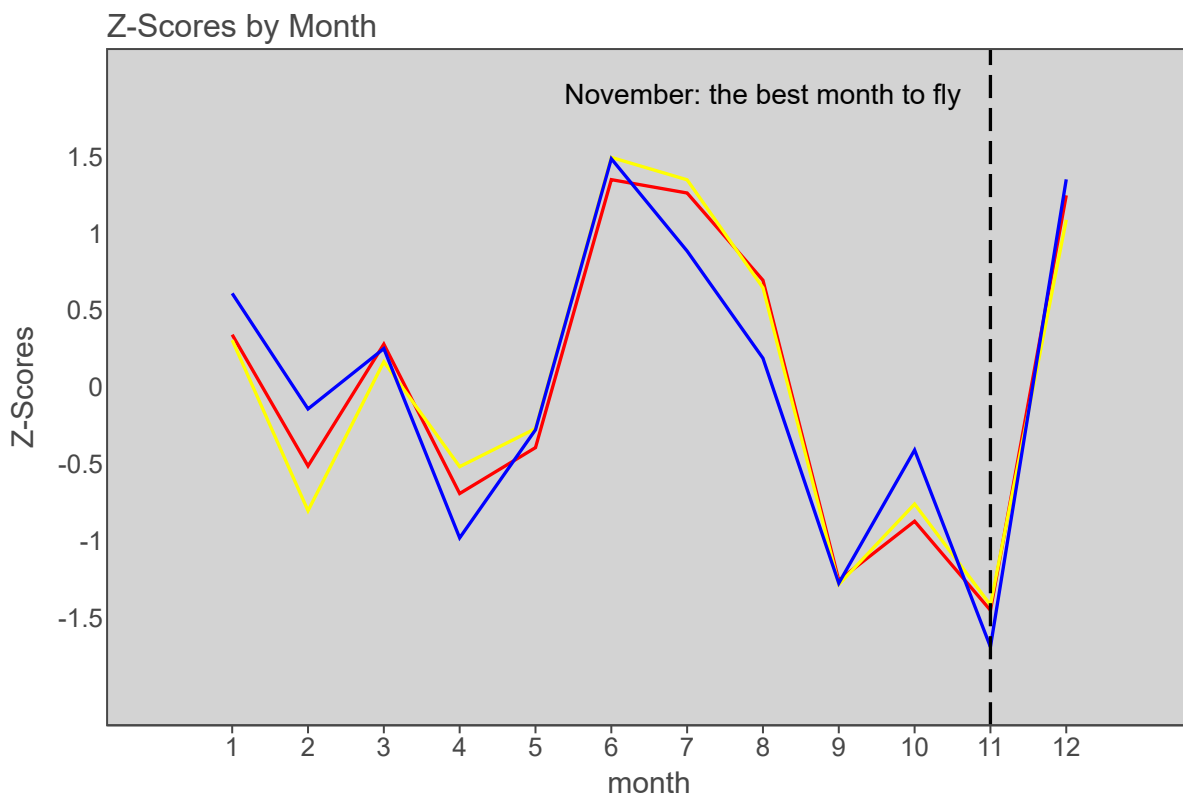
QUESTION 4

According to the BTS website, the "Weather" category only accounts for severe weather delays. Mild weather delays are not counted in the "Weather" category, but are actually included in both the "NAS" and "Late-Arriving Aircraft" categories. Your job is to create a new column that calculates the total number of flights delayed by weather (both severe and mild). You will need to replace all the missing values in the Late Aircraft variable with the mean. Show your work by printing the first 5 rows of data in a table. Use these three rules for your calculations:

100% of delayed flights in the Weather category are due to weather 30% of all delayed flights in the Late-Arriving category are due to weather From April to August, 40% of delayed flights in the NAS category are due to weather. The rest of the months, the proportion rises to 65%

▶ Show the code

	airport_name	airport_code	month	num_of_flights_total	proportion_delays_by_weather	avg_proportion_delays_b
0	San Francisco International	SFO	1	11560	0.22	0.1
1	San Francisco International	SFO	12	13827	0.20	0.1
2	San Francisco International	SFO	12	11792	0.19	0.1
3	San Francisco International	SFO	2	10016	0.19	0.1
4	San Francisco International	SFO	12	10972	0.18	0.1

◀ ▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬ ▶

QUESTION 5

Using the new weather variable calculated above, create a barplot showing the proportion of all flights that are delayed by weather at each airport. Describe what you learn from this graph.

According to this chart, 1 out of 10 flights is delayed due to weather at San Francisco International Airport, which represents the highest average delay in this data set. In addition, I've observed that delays due to weather are not as common as we might think. If we look at another airport with more flights, such as Atlanta Hartsfield-Jackson International Airport, the average of delays due to weather is less than 1%.

▶ Show the code
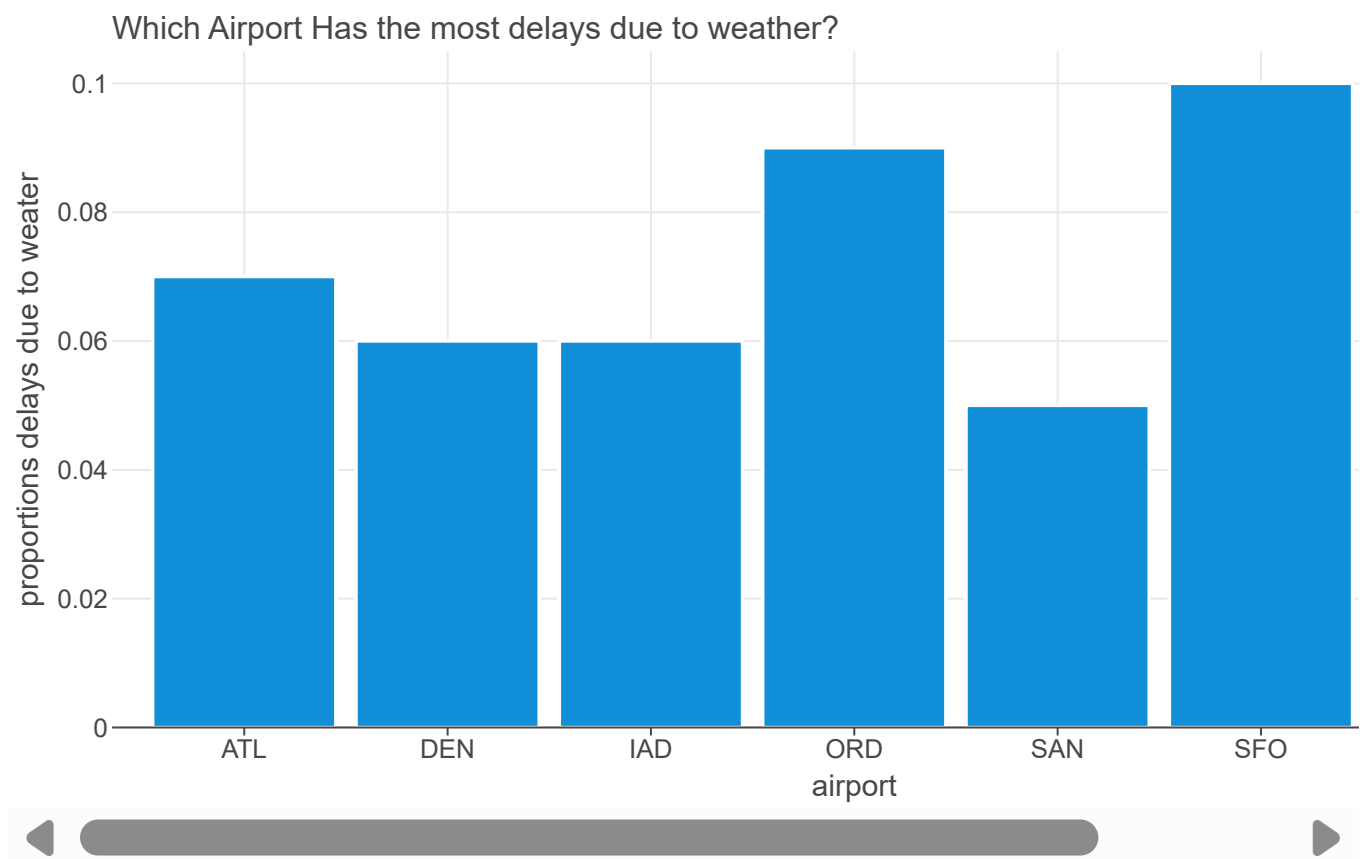
Stretch Questions

Which delay is the worst delay? Create a similar analysis as above for Weahter Delay with: Carrier Delay and Security Delay. Compare the proportion of delay for each of the three categories in a Chart and a Table. Describe your results.

▶ Show the code

	month	z_score_carrier	z_score_delays_nas	carrier_variation	nas_variation	category
0	1	0.341446	0.610199	0.000000	0.000000	high
1	2	-0.512965	-0.141708	-0.854412	-0.751907	medium
2	3	0.279910	0.251051	0.792875	0.392759	medium
3	4	-0.691965	-0.980530	-0.971875	-1.231581	low
4	5	-0.393223	-0.277186	0.298742	0.703344	medium
5	6	1.350313	1.486400	1.743536	1.763586	high

	month	z_score_carrier	z_score_delays_nas	carrier_variation	nas_variation	category
6	7	1.263355	0.885592	-0.086958	-0.600808	high
7	8	0.695764	0.188205	-0.567592	-0.697387	high
8	9	-1.257102	-1.275433	-1.952866	-1.463638	low
9	10	-0.872192	-0.408957	0.384910	0.866477	medium
10	11	-1.451354	-1.689708	-0.579163	-1.280752	low
11	12	1.248015	1.352077	2.699369	3.041785	high

This is a complex question, but we can answer it by looking at the z-scores. Z-scores represent the distance between the overall average and the monthly average for each category, helping us compare "apples to apples." In this case, as the value of each delay approaches zero, the delay worsens. With this in mind and using the table, we can compare both types of delays.

Here, we have both delays represented: the yellow line shows the values for Security delays, and the red line shows the values for Carrier delays. As you can see, both delays have their highs and lows. The yellow line seems to reach lower values, but it's still difficult to determine which of these is the worst delay.

▶ Show the code
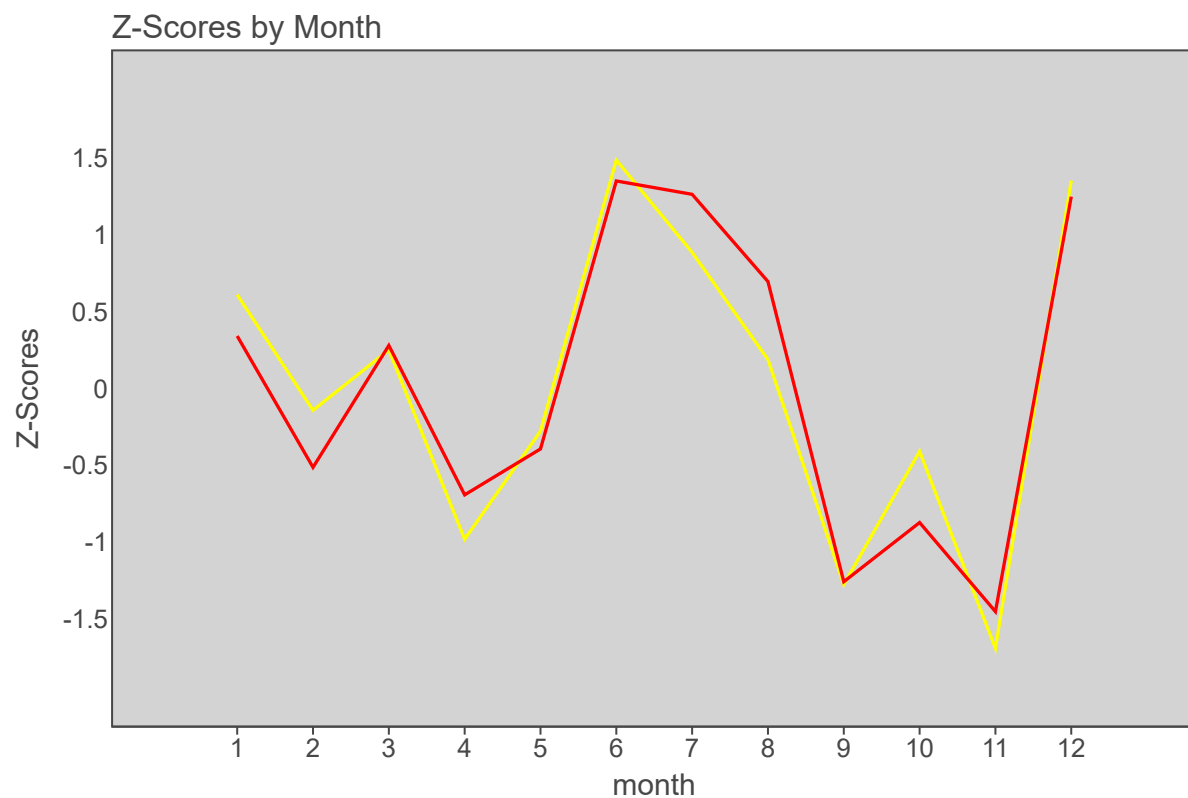
But we can use our variance column, which shows how the values for each delay type vary over the months. With that information in a graph, we can see that carrier delays exhibit more variance over time.

So, we can say that carrier delay is the worst of these delays because its values vary more across the months.

▶ Show the code

Variation by Month