Data Analytics Project

2015 Flight Delays and Cancellations

Introduction

The project objectives were:

- Experience contact with a **large dataset**
- Deal with data with **no previous contact**
- Use **data cleansing** and **data manipulation** techniques
- Use the clean data to **create visualizations**
- **Find insights** to solve questions that arise

The chosen database was published by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics and contains information about the performance of US domestic flights, including both on-time, delayed, canceled and diverted flights, for the year of 2015. It contains almost six million rows, and working with this considerable amount of data was one of the reasons to choose this dataset.

The project will use Google BigQuery, for data cleanse and manipulation, integrated to Google Data Studio for Dashboard visualizations.

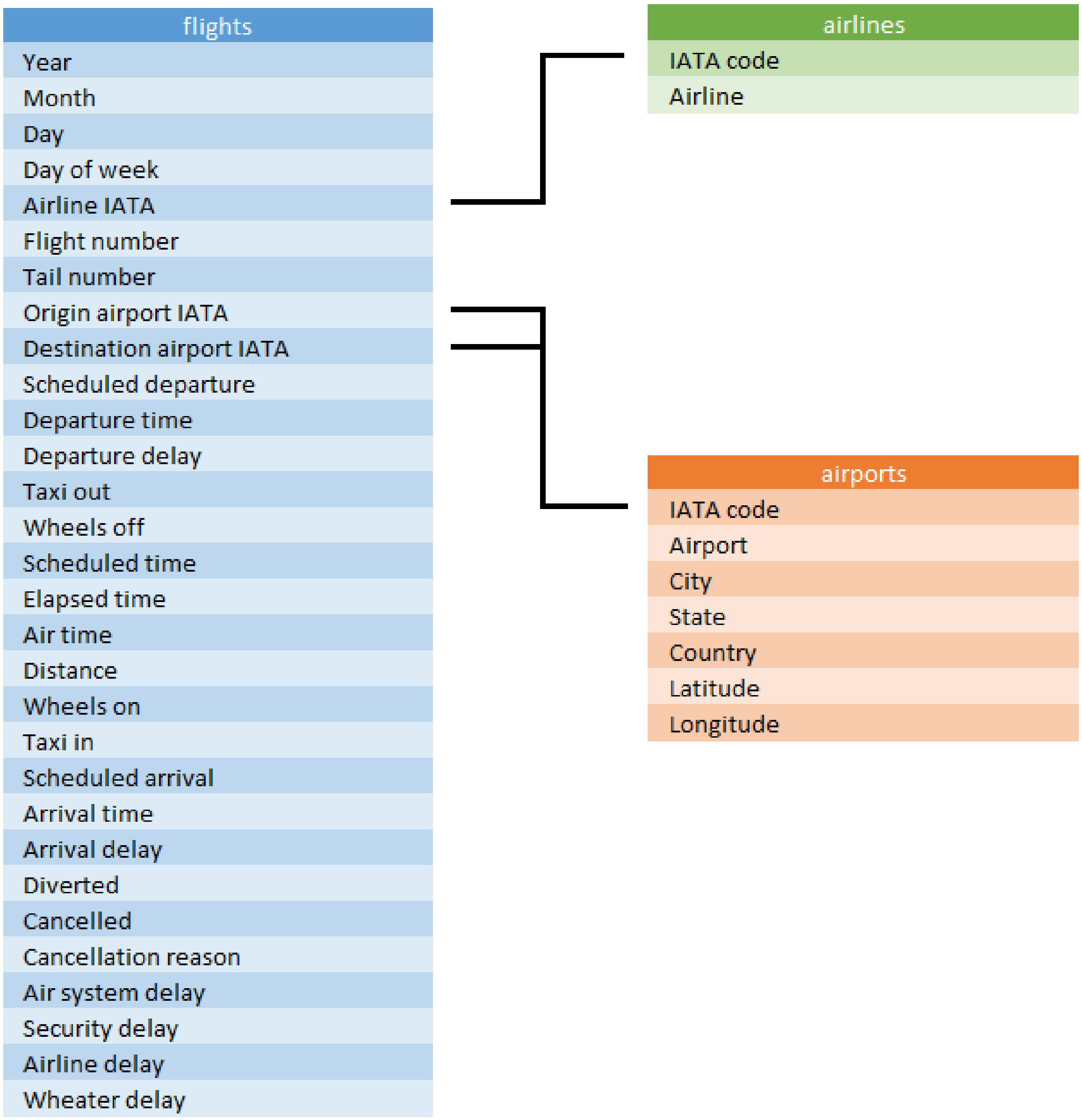
Importing Data

The database consists in three tables: **flights**, **airlines** and **airports**. The main one is the flights table, consisting in more than 5,800,000 rows, in a CSV archive of almost 600 MB in size. The other two tables are auxiliary to the main one, related to the main one by the IATA code of the airports (3-digit code) and the airlines (2-digit code).

This large table was divided, in order to be possible to upload CSVs direct to BigQuery, as it has a limitation for this manual upload method of 100 MB archive size. Also, uploading by this method would result in exercising UNION functions via SQL. So it was divided into **6 tables**, where each table contains information about 2 months of the year.

The flight table also has many columns, with information about dates, flight info, scheduled and actual flight times, and specific delay information. The airport table contains localization information about it, besides the name, while the airline table contains only the airline name.

Here there is the tables scheme:



flights
Year
Month
Day
Day of week
Airline IATA
Flight number
Tail number
Origin airport IATA
Destination airport IATA
Scheduled departure
Departure time
Departure delay
Taxi out
Wheels off
Scheduled time
Elapsed time
Air time
Distance
Wheels on
Taxi in
Scheduled arrival
Arrival time
Arrival delay
Diverted
Cancelled
Cancellation reason
Air system delay
Security delay
Airline delay
Wheater delay

airlines
IATA code
Airline

airports
IATA code
Airport
City
State
Country
Latitude
Longitude

After importing the data, the attributed data types to the columns will be checked. Most of the numerical columns were automatically attributed as a float.

Data Manipulation And Cleansing

Complete SQL code for the following steps will be annexed in a SQL file to this project.

The first step worked here was to merge the year, month, and day columns, in order to have date information stored in a date column type. This process involved the use of the CONCAT function, where we would join the date parts and hyphens between them. Also, it was needed to add zeros at the left of the 1-digit numbers. The RIGHT function was then used after concatenating zeros on the left, to select only the last 2 digits of this inner concatenation. One example of this step can be seen here:

```
CONCAT(
    YEAR,
    '-',
    RIGHT( CONCAT('00', MONTH), 2),
    '-',
    RIGHT( CONCAT('00', DAY), 2)
  ) AS EVENT_DATE
```

It was also needed to join the six flight tables. As mentioned before, the flights table was divided into 6 parts before importing to BigQuery, where each part represents 2 months of data. So a simple UNION ALL command would be enough to join the data from these tables, if repeated sequentially for each table.

The second step was to correct attributed data types, using CAST functions. The concatenated date from the previous step was changed to DATE data type.

Columns that were attributed as float, like for example the Departure Delay, Taxi Out, Air Time, and many others, were changed to INT64 data type.

The columns that represented time (hours and minutes), like the Scheduled Departure column, for example, were also wrongly attributed as a float. So it was needed to concatenate back the zero digit on left, for hours that should have it, just like in the previous step related to the date concatenation. One example of this step can be seen here:

```
CAST(
  RIGHT(
    CONCAT(
      '000',
      SCHEDULED_DEPARTURE)
    , 4)
  AS STRING
) AS SCHEDULED_DEPARTURE
```

Then it was added colons between the hour and minutes digits. Also, some hours values had a 24h value instead of 00h, which also needed to be replaced to achieve a proper conversion to DateTime. One example of this step can be seen here:

```
CONCAT(
  IF (
    LEFT(SCHEDULED_DEPARTURE, 2) = '24',
    '00',
    LEFT(SCHEDULED_DEPARTURE, 2)
  ),
  ':',
  RIGHT(SCHEDULED_DEPARTURE, 2)
) AS SCHEDULED_DEPARTURE_HM
```

At last, all the hour and minutes fields had been transformed to timestamps, adding the date at the beginning, and 00 seconds at the end. This was done in order to be able to check delayed flights, as we don't have delay calculations for all the flights. Comparing the actual arrival time timestamps with the scheduled timestamps when can then check if the flight was delayed or not.

One observation on this last step is that for some flights at the very begging of the day, for example, a flight scheduled to depart at 00:30, if it departs ahead of time, and prior to 00:00, its event date for event timestamp concatenation should be the day before. The same applies to flights departing at the end of the day, and also for arrival, wheels off, and wheels on time. To generalize these cases, it was considered that a flight scheduled to depart prior to 06:00 in the morning, but actually departed between 18:00 and 00:00, would have departed on the day prior. Similar to it, flights scheduled to depart after 18:00, but with actual departure between 00:00 and 06:00 would be considered to have the following date. This approximation might be wrong for very specific cases, of flights with very long delays, but it was considered that these exceptional cases would not interfere with the general analysis.

After looking at the results of the date approximation for the flights with long delays registered, some tweaking was done on the intervals mentioned. The maximum and minimum delays of the database were also checked, resulting in a range that goes from -82 to 1988 minutes. Arrivals delay has a similar range. So there is no need to consider flights past 02:00 in the morning as early flights leaving on the day prior, which would reduce the mistakes made in the high delay flights. The final adjustments were to consider flights scheduled prior to 02:00 with actual departure after 22:00 as departed on the previous day. Flights not included in this situation, that had departure time 2 hours prior to the scheduled time, were considered as departed on the following day, covering even huge delays. Besides these cases, all other flights were considered to depart on the same day as scheduled. The same logic applies to arrivals. One example of this step can be seen here:

```
CAST(
  CONCAT(
    CASE
      WHEN CAST( LEFT(SCHEDULED_DEPARTURE_HM,2) AS INT64) < 2
        AND CAST( LEFT(DEPARTURE_TIME_HM,2) AS INT64) >= 22
        AND IFNULL(DEPARTURE_DELAY,0) <= 0
      THEN DATE_SUB( EVENT_DATE, INTERVAL 1 DAY)
      WHEN CAST( LEFT(DEPARTURE_TIME_HM,2) AS INT64) <
          CAST( LEFT(SCHEDULED_DEPARTURE_HM,2) AS INT64)
        AND CAST( LEFT(SCHEDULED_DEPARTURE_HM,2) AS INT64) -
          CAST( LEFT(DEPARTURE_TIME_HM,2) AS INT64) >= 2
        AND IFNULL(DEPARTURE_DELAY,0) > 0
      THEN DATE_ADD(EVENT_DATE, INTERVAL 1 DAY)
      ELSE EVENT_DATE
    END,
    ' ',
    DEPARTURE_TIME_HM,
    ':00'
  ) AS DATETIME
) AS DEPARTURE_TIME
```

All the steps made were saved as a **View**, which will then be used in the following steps.

To start, overall data was checked, without month or localization aggregations. Data calculated in this step includes the numbers of flights scheduled for the year, flights that actually departed, on-time flights, delayed flights, diverted flights, and canceled flights. There is also the number of canceled flights per cancellation reason.

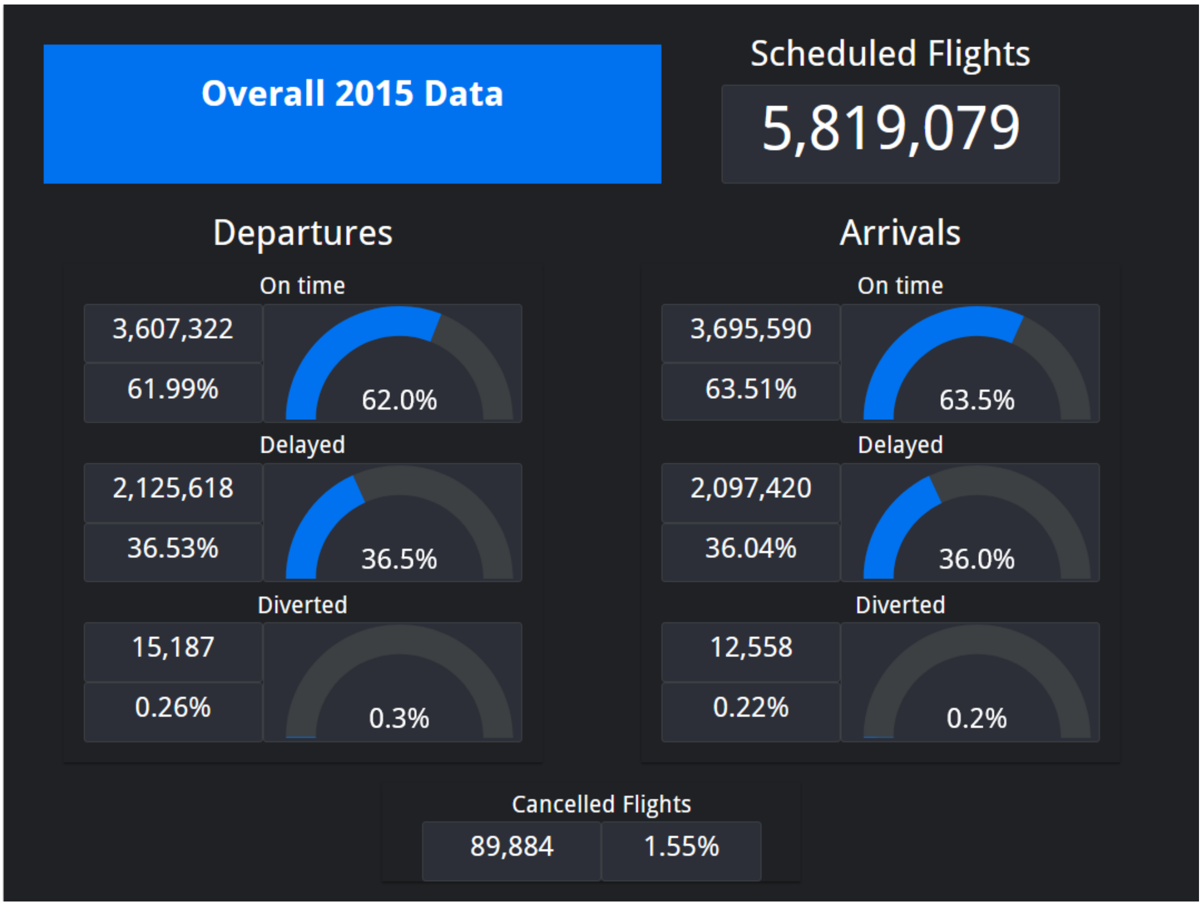
Data segmented by month was also checked, and segmented by day also. When looking at this data in the dashboard, we will be to see seasonal effects, and also outlier days. Also, it was checked some specific airport data. After checking which were the 5 busiest airports in the year, overall views specific to them were created.

Result Dashboard

The result Dashboard consists of 3 sections:

- Overall Data
- Breakdowns
- Airports

The first page in the Overall Data section shows us an overview of all departures and arrivals on time, delayed, diverted, and canceled flights. The first conclusion here is that we have a lot of delayed flights, reaching more than one third of all scheduled flights, as we can see below:



Overall 2015 Data

Scheduled Flights

5,819,079

Departures

On time

3,607,322

61.99%

Delayed

2,125,618

36.53%

Diverted

15,187

0.26%

Arrivals

On time

3,695,590

63.51%

Delayed

2,097,420

36.04%

Diverted

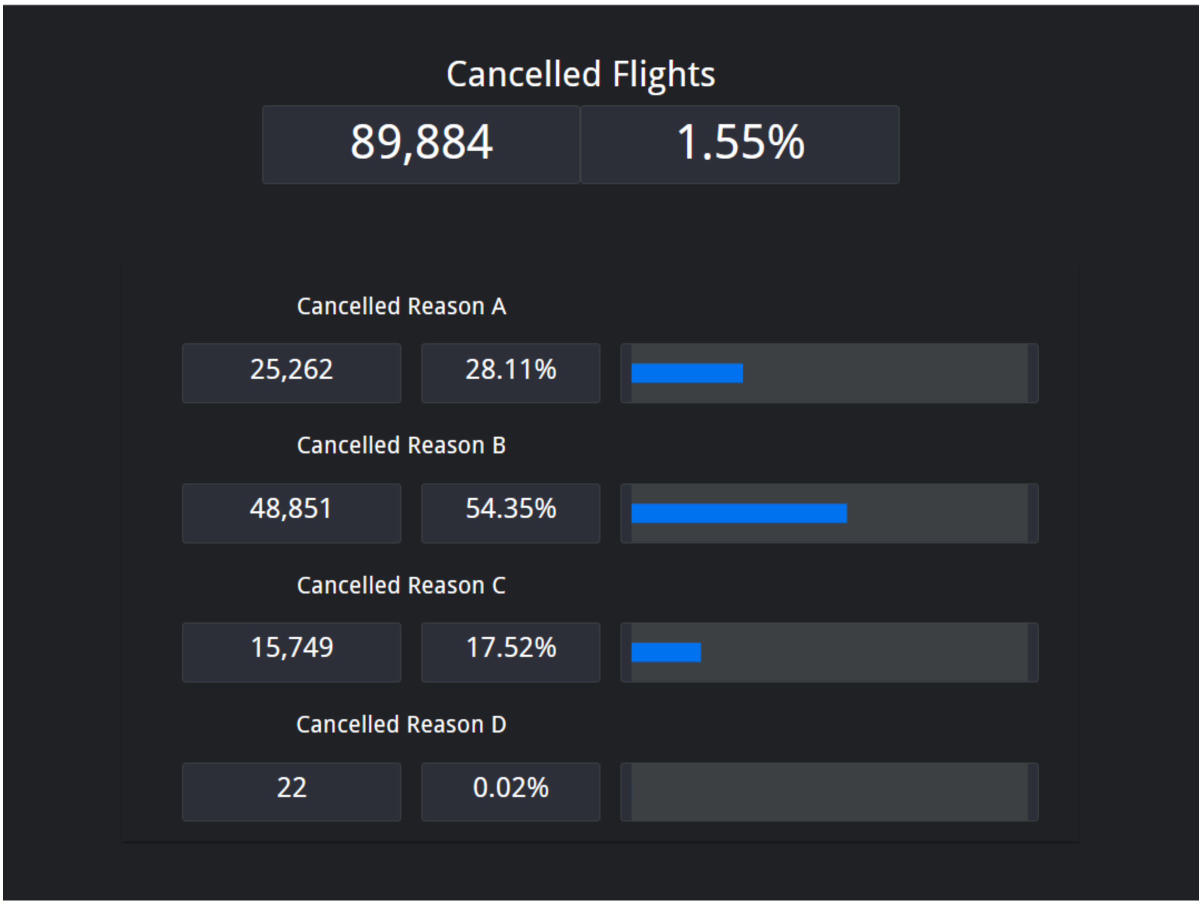
12,558

0.22%

Cancelled Flights

89,884 1.55%

The second page shows detail about cancelations and the proportion of each cancellation reason. Reason B occurs 54% of the time, while on the other side Reason D occured very few times, as we can see below:



Cancelled Flights

89,884 1.55%

Cancelled Reason A

25,262 28.11%

Cancelled Reason B

48,851 54.35%

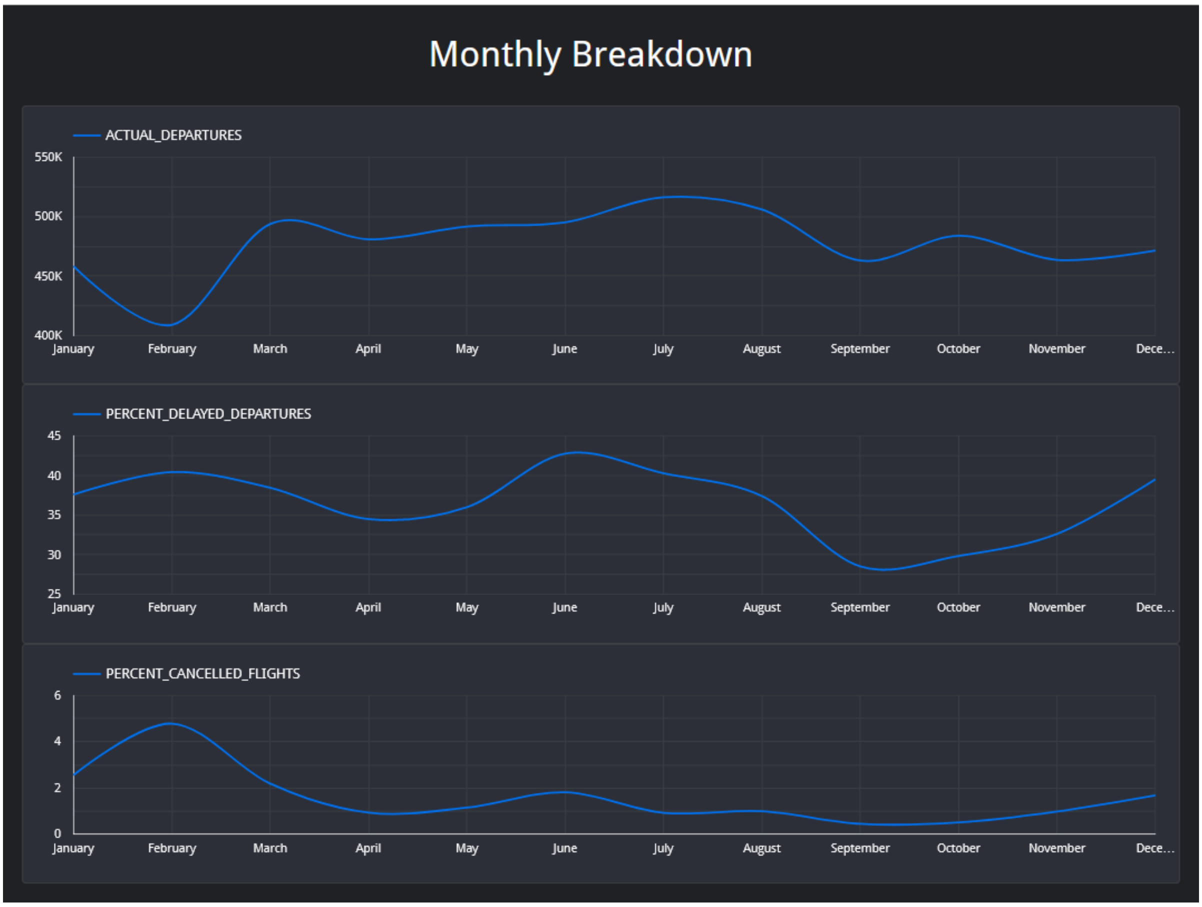
Cancelled Reason C

15,749 17.52%

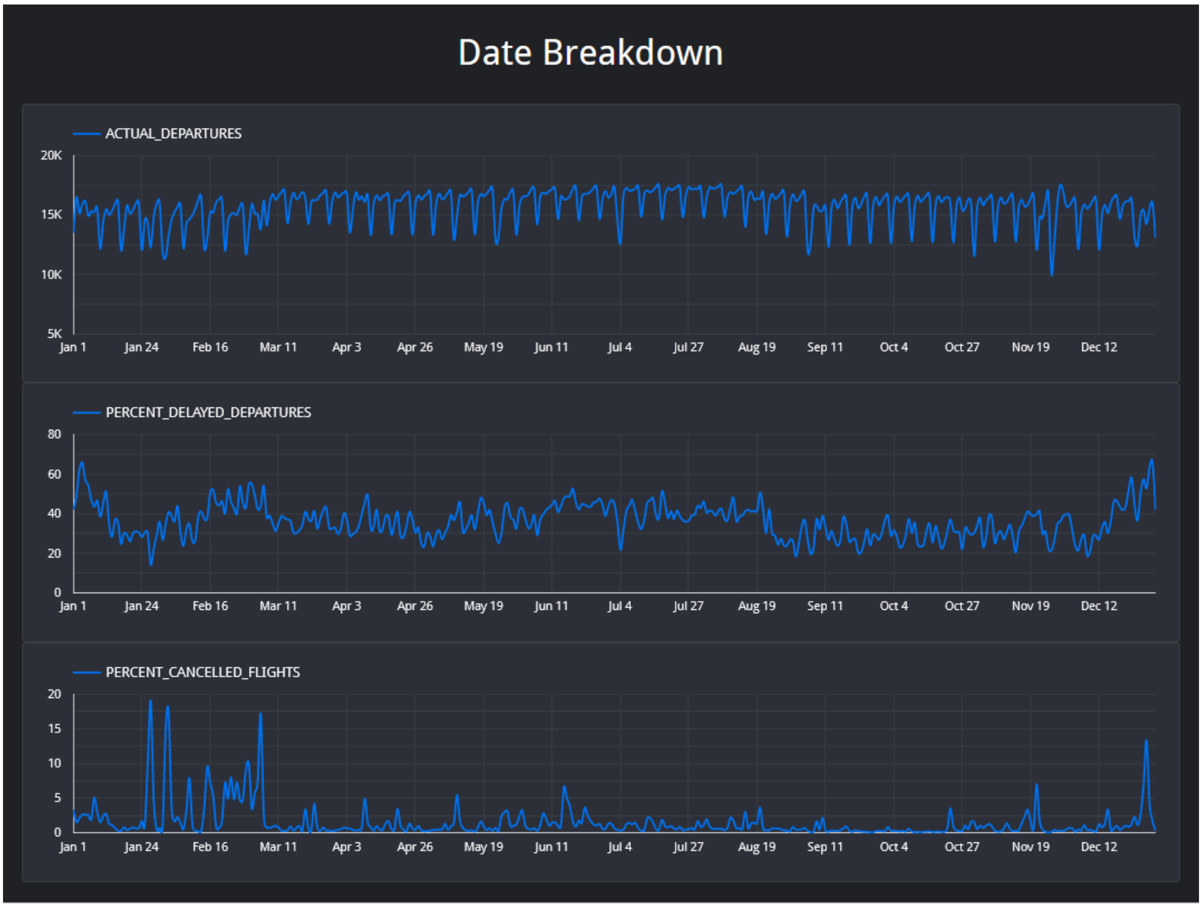
Cancelled Reason D

22 0.02%

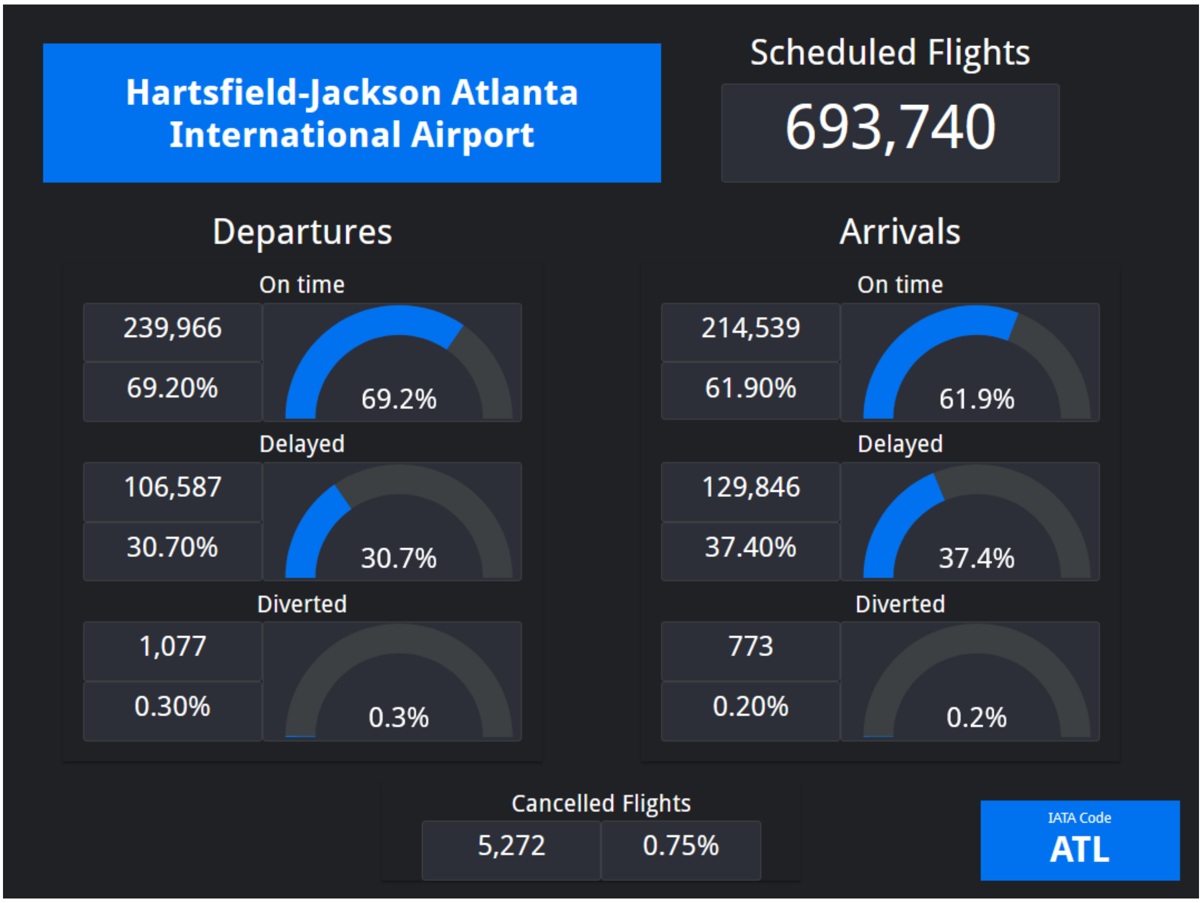
In the Breakdown section, first, we can see the monthly breakdown. It shows a higher cancellation rate in February, besides being the month with fewer flights scheduled. September and October showed fewer delayed departures.



On the Date Breakdown page, we can see clearly the weekly pattern, where weekends have fewer flights per day. Delayed departures seem to reach higher values closer to the turns of the years. Canceled flights seem to reach higher value on specific days, which probably had some cause of greater nature involved.



At last, we have the Airports section, showing us the overall performance of each of the five busiest airports. Atlanta Airport showed considerably better on-time departures percentage compared to the USA overall, even while receiving slightly more delayed flights (arrivals). Also lower cancelld flights here.



Hartsfield-Jackson Atlanta International Airport

Scheduled Flights

693,740

Departures

On time

239,966

69.20%

Delayed

106,587

30.70%

Diverted

1,077

0.30%

Arrivals

On time

214,539

61.90%

Delayed

129,846

37.40%

Diverted

773

0.20%

Cancelled Flights

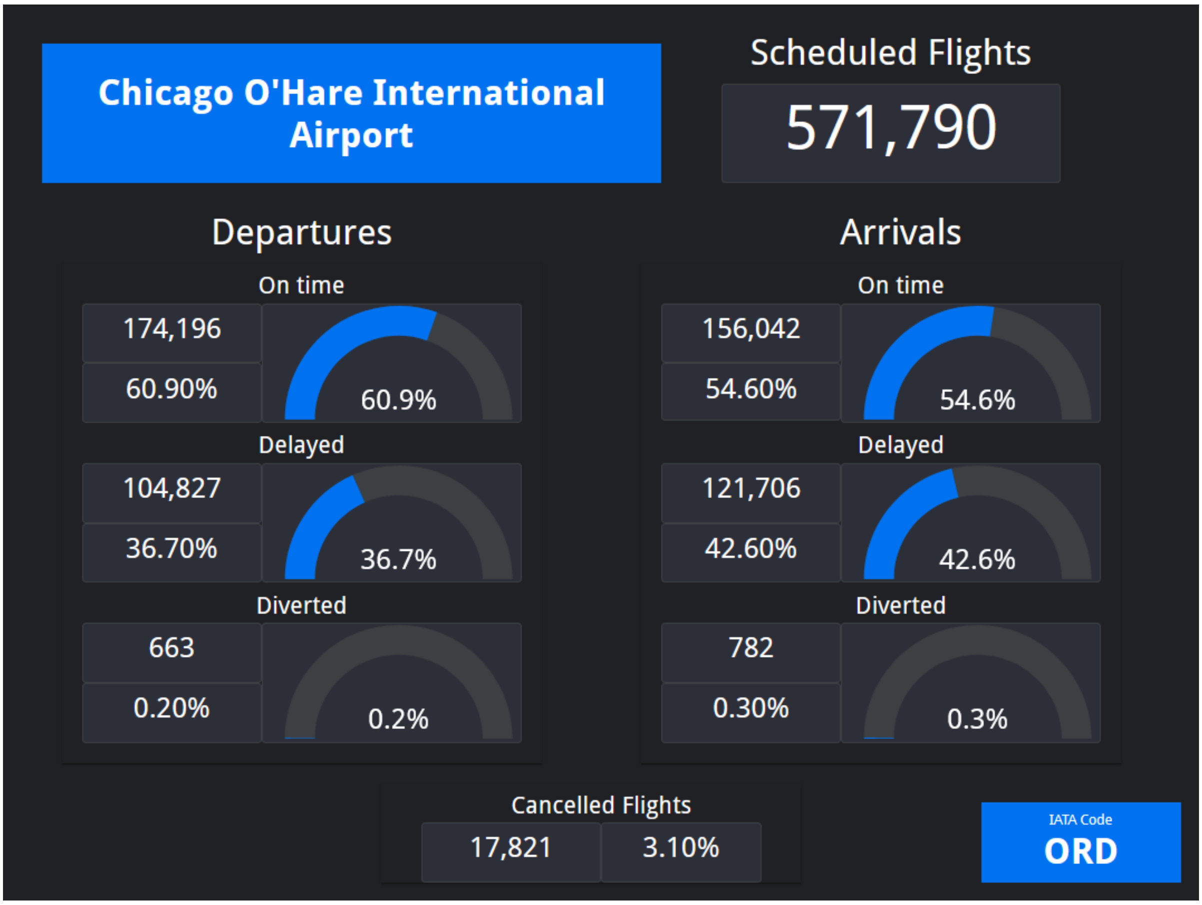
5,272

0.75%

IATA Code

ATL

Chicago Airport, on the other hand, shows slightly lower on-time departures, but it received a much higher delayed flight amount. The number of canceled flights was higher here.



Chicago O'Hare International Airport

Scheduled Flights

571,790

Departures			Arrivals		
On time			On time		
174,196	60.90%	60.9%	156,042	54.60%	54.6%
Delayed			Delayed		
104,827	36.70%	36.7%	121,706	42.60%	42.6%
Diverted			Diverted		
663	0.20%	0.2%	782	0.30%	0.3%

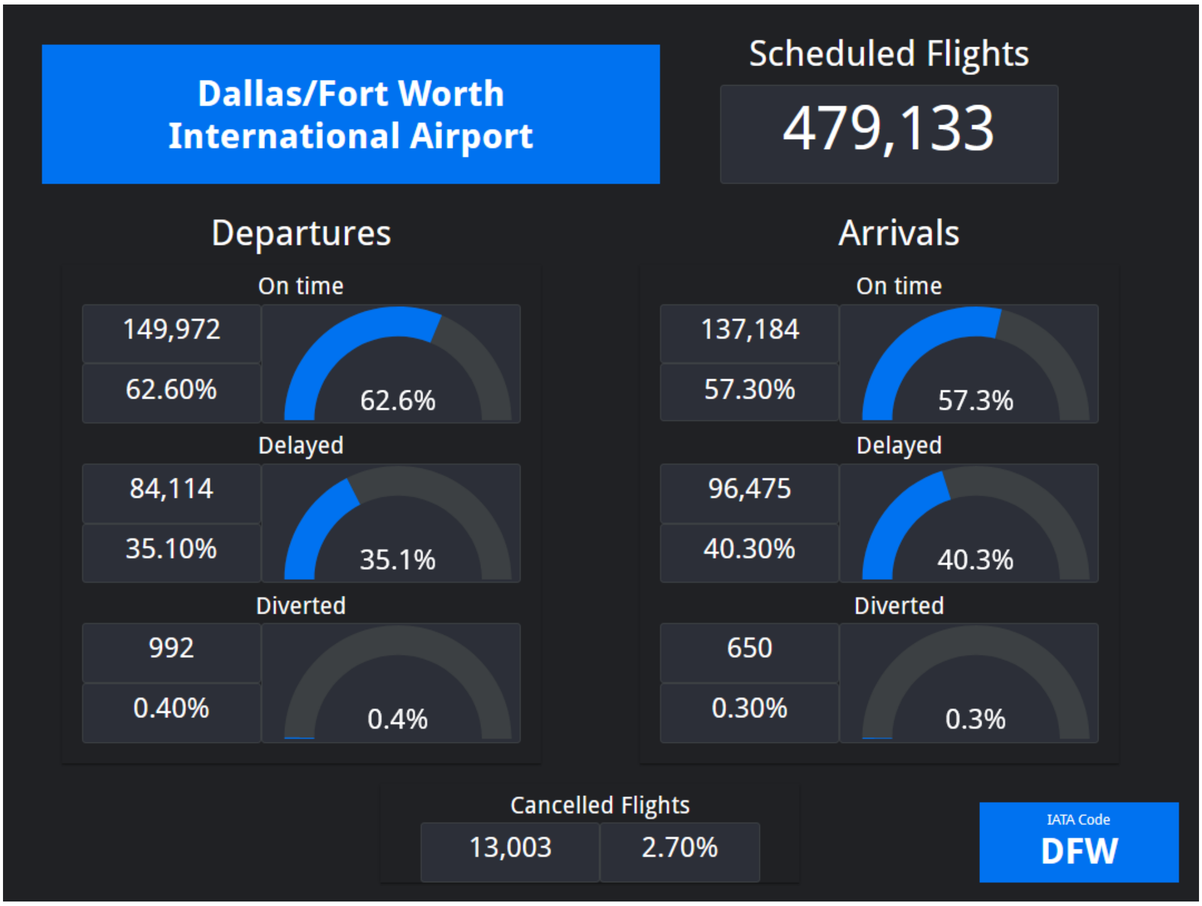
Cancelled Flights

17,821 3.10%

IATA Code

ORD

Dallas/Fort Worth Airport had a close to the average percentage of on-time departures, and a lower on-time arrivals percentage. Also had a higher amount of canceled flights.



Dallas/Fort Worth International Airport

Scheduled Flights

479,133

Departures

On time

149,972

62.60%

Delayed

84,114

35.10%

Diverted

992

0.40%

Arrivals

On time

137,184

57.30%

Delayed

96,475

40.30%

Diverted

650

0.30%

Cancelled Flights

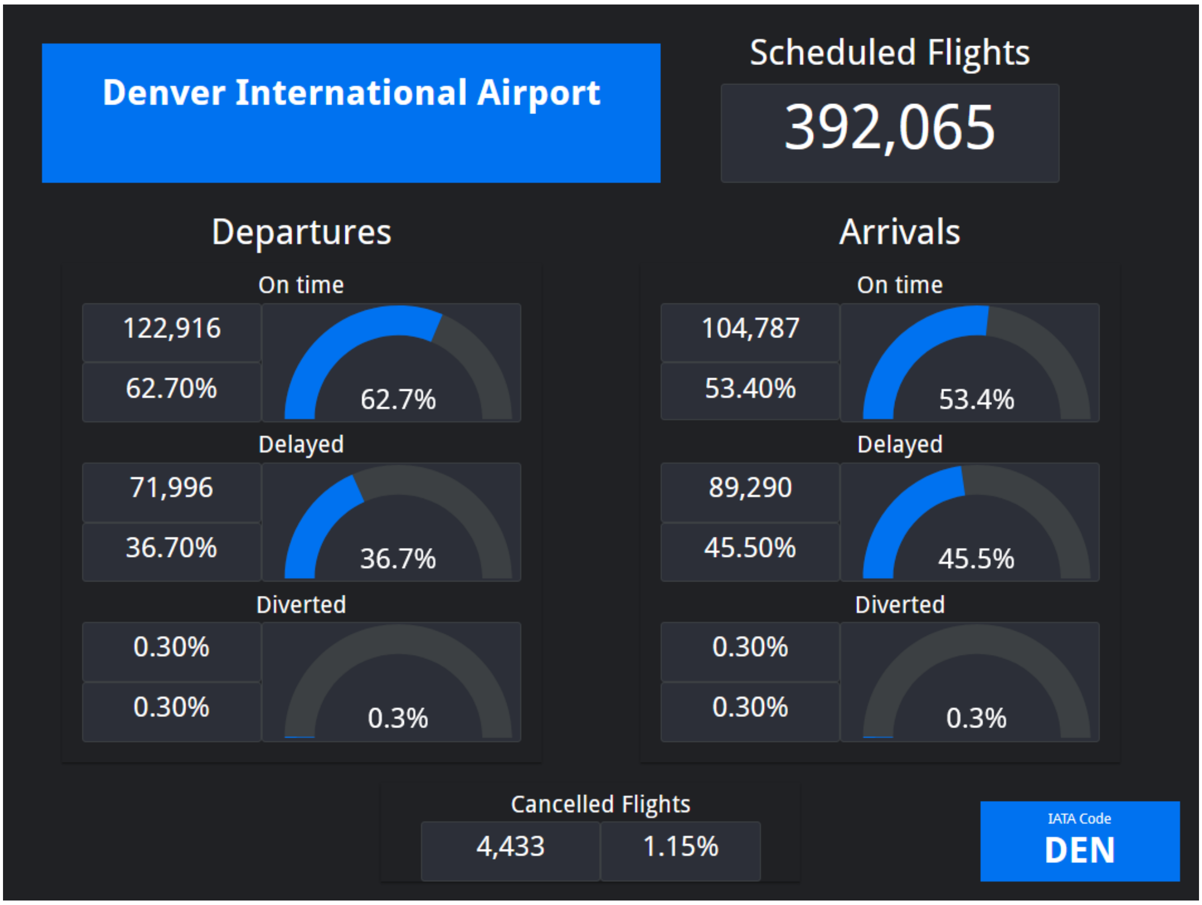
13,003

2.70%

IATA Code

DFW

Denver also had a close to the average on-time departure percentage, while having the lower on-time arrivals percentage of the five analyzed airports. Slightly lower canceled flights percentage.



Denver International Airport

Scheduled Flights

392,065

Departures

On time

122,916		
62.70%	62.7%	

Delayed

71,996	
36.70%	36.7%

Diverted

0.30%	
0.30%	0.3%

Arrivals

On time

104,787	
53.40%	53.4%

Delayed

89,290	
45.50%	45.5%

Diverted

0.30%	
0.30%	0.3%

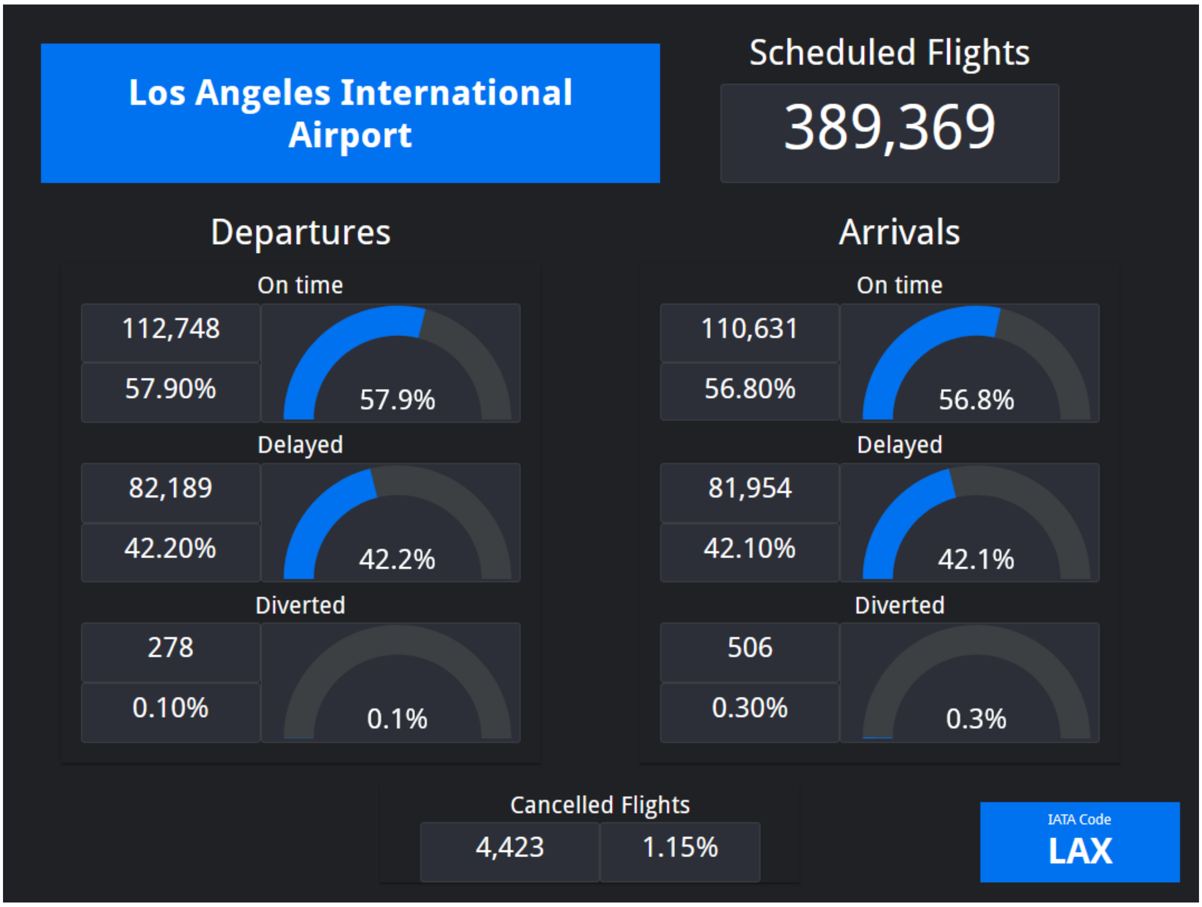
Cancelled Flights

4,433	1.15%

IATA Code

DEN

Los Angeles Airport had both lower on-time percentages of departures and arrivals, with a slightly lower cancelled percentage.



Los Angeles International Airport

Scheduled Flights
389,369

Departures

On time
112,748
57.90%
57.9%

Delayed
82,189
42.20%
42.2%

Diverted
278
0.10%
0.1%

Arrivals

On time
110,631
56.80%
56.8%

Delayed
81,954
42.10%
42.1%

Diverted
506
0.30%
0.3%

Cancelled Flights
4,423
1.15%

IATA Code
LAX

Conclusion

A more thorough analysis could follow from here, but for the purpose of demonstrating the ability to collect data, manipulate it, and create visualizations, the work on this project will end here. All the work was done starting only from the data files in CSV, using the tools BigQuery and Google Data Studio.