DBMS Final Project Presentation

Kate Halushka & Josie Libbon

Our Objective

Translate semi-structured JSON data on flight operations into a PostgreSQL structured format. Host this database with AWS and perform relevant analysis.

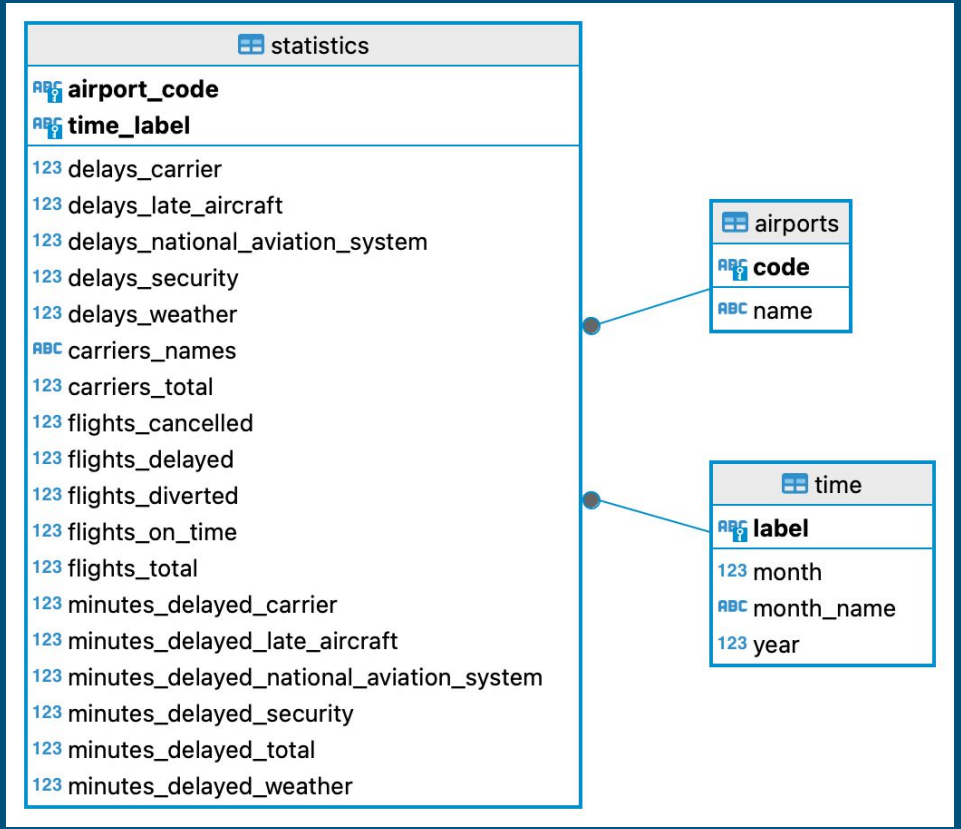
Our Dataset

- JSON Airport dataset
  - https://think.cs.vt.edu/corgis/datasets/json/airlines/airlines.json


```
{
  "Airport": {
    "Code": "ATL",
    "Name": "Atlanta, GA: Hartsfield-Jackson Atlanta International"
  },
  "Time": {
    "Label": "2003/06",
    "Month": 6,
    "Month Name": "June",
    "Year": 2003
  },
  "Statistics": {
    "# of Delays": {
      "Carrier": 1009,
      "Late Aircraft": 1275,
      "National Aviation System": 3217,
      "Security": 17,
      "Weather": 328
    },
    "Carriers": {
      "Names": "American Airlines Inc.,JetBlue Airways,Continental Air Lines Inc.,Delta Air Lines Inc.,Atlantic Southeast Airlines,AirTran Airways Corporation,America West Airlines Inc.,Northwest Airlines Inc.,ExpressJet Airlines Inc.,United Air Lines Inc.,US Airways Inc.",
      "Total": 11
    },
    "Flights": {
      "Cancelled": 216,
      "Delayed": 5843,
      "Diverted": 27,
      "On Time": 23974,
      "Total": 30060
    },
    "Minutes Delayed": {
      "Carrier": 61606,
      "Late Aircraft": 68335,
      "National Aviation System": 118831,
      "Security": 518,
      "Total": 268764,
      "Weather": 19474
    }
  },
},
```

Database Translation

- Created 3 tables
  - Airports
  - Time
  - Statistics
- Populated table using a python script
  - Encountered data cleaning issue with O'Hare (due to apostrophe)
  - Luckily all of the entries had the same keys



Populated Database

- Used DBeaver to connect to RDS Postgres instance

	code	name
1	ATL	Atlanta, GA: Hartsfield-Jackson Atlanta International
2	BOS	Boston, MA: Logan International
3	BWI	Baltimore, MD: Baltimore/Washington International Thurgood Marshall
4	CLT	Charlotte, NC: Charlotte Douglas International
5	DCA	Washington, DC: Ronald Reagan Washington National
6	DEN	Denver, CO: Denver International
7	DFW	Dallas/Fort Worth, TX: Dallas/Fort Worth International
8	DTW	Detroit, MI: Detroit Metro Wayne County
9	EWR	Newark, NJ: Newark Liberty International
10	FLL	Fort Lauderdale, FL: Fort Lauderdale-Hollywood International
11	IAD	Washington, DC: Washington Dulles International
12	IAH	Houston, TX: George Bush Intercontinental/Houston
13	JFK	New York, NY: John F. Kennedy International
14	LAS	Las Vegas, NV: McCarran International
15	LAX	Los Angeles, CA: Los Angeles International
16	LGA	New York, NY: LaGuardia

	label	month	month_name	year
1	2003/06	6	June	2,003
2	2003/07	7	July	2,003
3	2003/08	8	August	2,003
4	2003/09	9	September	2,003
5	2003/10	10	October	2,003
6	2003/11	11	November	2,003
7	2003/12	12	December	2,003
8	2004/01	1	January	2,004
9	2004/02	2	Febuary	2,004
10	2004/03	3	March	2,004
11	2004/04	4	April	2,004
12	2004/05	5	May	2,004
13	2004/06	6	June	2,004
14	2004/07	7	July	2,004
15	2004/08	8	August	2,004
16	2004/09	9	September	2,004
17	2004/10	10	October	2,004

	airport_code	time_label	delays_carrier	delays_late_aircraft	delays_national_avi
1	ATL	2003/06	1,009	1,275	
2	BOS	2003/06	374	495	
3	BWI	2003/06	296	477	
4	CLT	2003/06	300	472	
5	DCA	2003/06	283	268	
6	DEN	2003/06	516	323	
7	DFW	2003/06	986	1,390	
8	DTW	2003/06	376	371	
9	EWR	2003/06	322	519	
10	FLL	2003/06	247	256	
11	IAD	2003/06	320	295	
12	IAH	2003/06	329	730	
13	JFK	2003/06	376	226	
14	LAS	2003/06	511	678	

Analytics

- 7 functions
- get_extreme_rates()
- Prints the best and worst delay and cancellation rates among all months/airports
- Output:

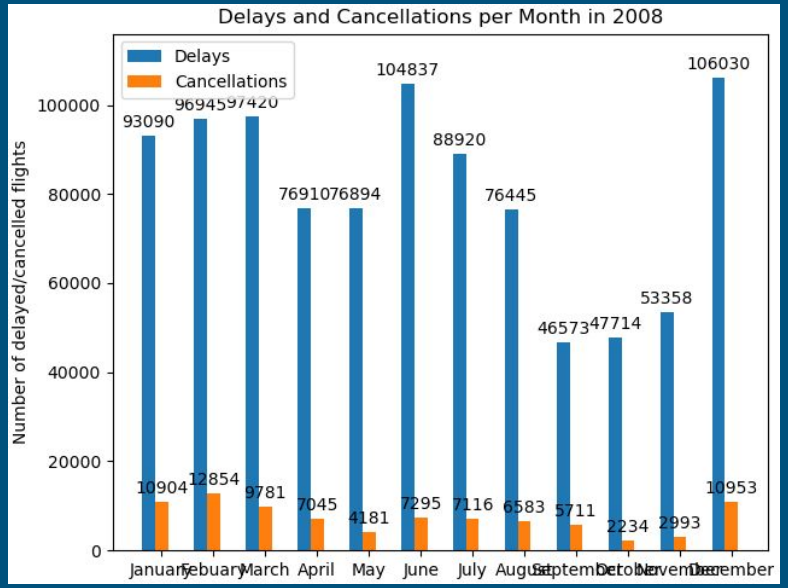
```
Best and worst flight delay and cancellation rates by airport and month:
The month and airport with the highest flight cancellation rate was DCA in Febuary 2010 with 22.58% of flights being cancelled.
The month and airport with the highest flight delay rate was SFO in January 2008 with 44.39% of flights being delayed.
The month and airport with the lowest flight cancellation rate was TPA in October 2013 with 0.06% of flights being cancelled.
The month and airport with the lowest flight delay rate was SLC in November 2009 with 5.98% of flights being delayed.
```

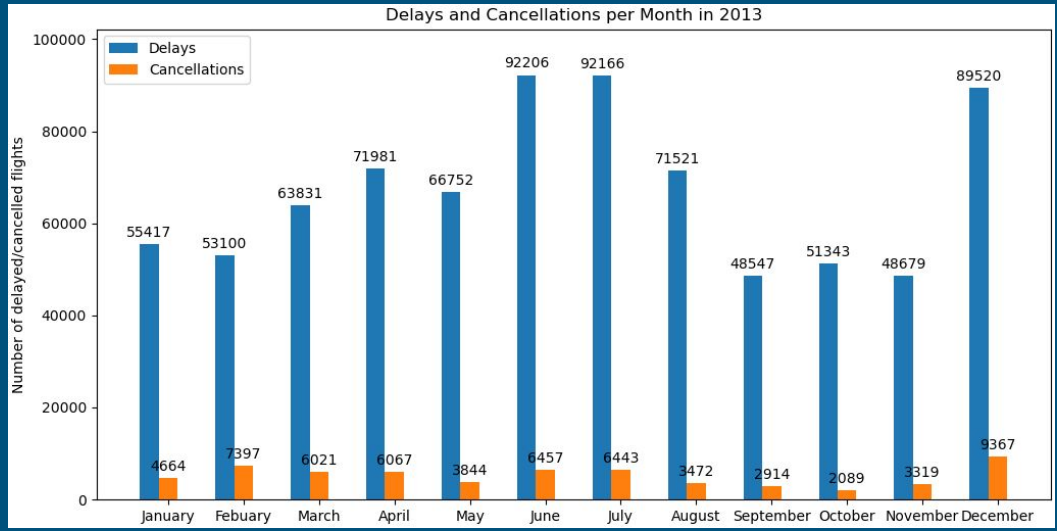
Analytics

- plot_flight_data_by_year(year)
- Outputs a plot with the total number of delays and cancellations per month across all airports for a given year

```
cursor.execute(f"""SELECT airport_code, time.month_name, time_label, flights_cancelled, flights_delayed
FROM statistics JOIN time on statistics.time_label = time.label WHERE time.year = {year} ORDER BY
time.month""")
```

Flight Data by Year



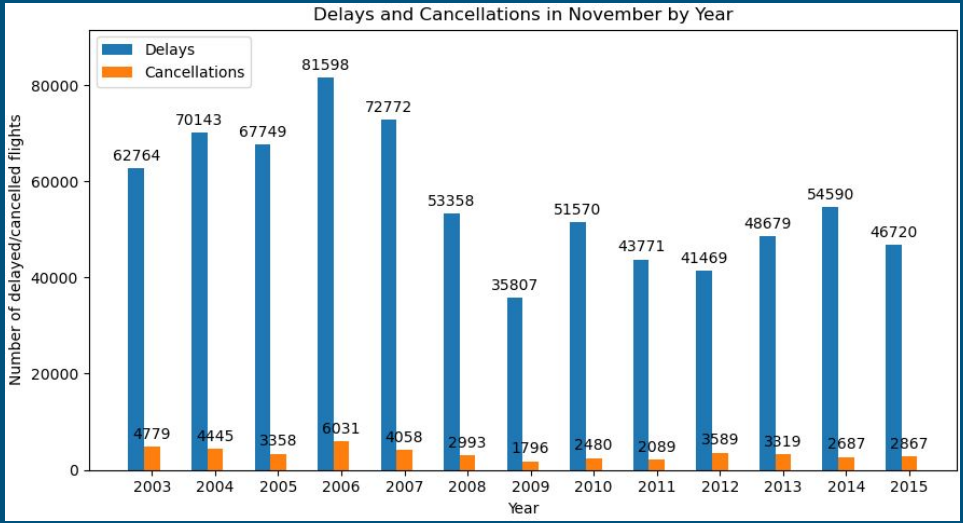


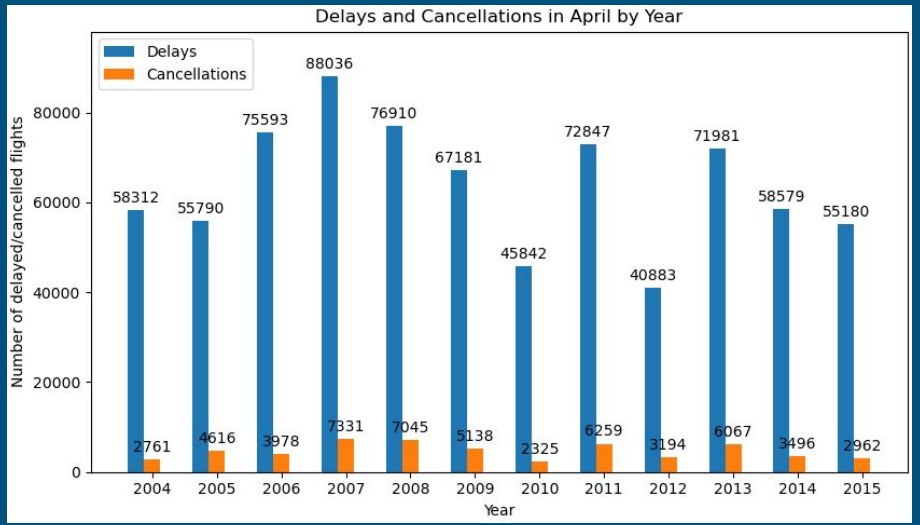
Analytics

- plot_flight_data_by_month(month)
- Outputs a plot with the total number of delays and cancellations per year across all airports for a given month

```
cursor.execute(f"""SELECT airport_code, time.month_name, time_label, flights_cancelled,
flights_delayed, time.year FROM statistics JOIN time on statistics.time_label = time.label WHERE
time.month = {month} ORDER BY time.year""")
```

Flight Data by Month





Analytics

- best_and_worst_airports()
- Outputs the best and worst airports in terms of on-time flight rates across all months of the dataset, and the average on-time rate across the entire dataset

```
Best and worst airports with on-time flights all-time:
The airport with the lowest rate of on-time flights is: EWR with a rate of 67.59% of flights on time.
The airport with the highest rate of on-time flights is: SLC with a rate of 84.36% of flights on time.
The average on-time rate across all airports is 78.10%
```

Analytics

- average_delay_times()
- Outputs the average amount of time that a flight is delayed for by reason (5 reasons in the dataset) across all months/airports

```
Average delay times by reason for delay:
The average time a flight is late for carrier delays is 60.95 minutes.
The average time a flight is late for aircraft delays is 62.62 minutes.
The average time a flight is late for national aviation system delays is 47.22 minutes.
The average time a flight is late for security delays is 37.98 minutes.
The average time a flight is late for weather delays is 80.25 minutes.
```

Analytics

- most_frequent_delay_type_by_airport()
- Outputs the most frequent delay type at each airport, as well as the overall most frequent delay type
- Only reasons cited were Late Aircraft and National Aviation System
- National Aviation System was also the most frequent reason overall

```
Most frequent delay type by airport:
The most frequent reason for delays at ATL is National Aviation System.
The most frequent reason for delays at BOS is National Aviation System.
The most frequent reason for delays at BWI is Late Aircraft.
The most frequent reason for delays at CLT is National Aviation System.
The most frequent reason for delays at DCA is National Aviation System.
The most frequent reason for delays at DEN is Late Aircraft.
The most frequent reason for delays at DFW is Late Aircraft.
The most frequent reason for delays at DTW is National Aviation System.
The most frequent reason for delays at EWR is National Aviation System.
The most frequent reason for delays at FLL is National Aviation System.
The most frequent reason for delays at IAD is Late Aircraft.
The most frequent reason for delays at IAH is National Aviation System.
The most frequent reason for delays at JFK is National Aviation System.
The most frequent reason for delays at LAS is Late Aircraft.
The most frequent reason for delays at LAX is Late Aircraft.
The most frequent reason for delays at LGA is National Aviation System.
The most frequent reason for delays at MCO is Late Aircraft.
The most frequent reason for delays at MDW is Late Aircraft.
The most frequent reason for delays at MIA is National Aviation System.
The most frequent reason for delays at MSP is National Aviation System.
The most frequent reason for delays at ORD is National Aviation System.
The most frequent reason for delays at PDX is Late Aircraft.
The most frequent reason for delays at PHL is National Aviation System.
The most frequent reason for delays at PHX is Late Aircraft.
The most frequent reason for delays at SAN is Late Aircraft.
The most frequent reason for delays at SEA is Late Aircraft.
The most frequent reason for delays at SFO is National Aviation System.
The most frequent reason for delays at SLC is Late Aircraft.
The most frequent reason for delays at TPA is Late Aircraft.
The most frequent reason for delays across all airports is National Aviation System.
```

Analytics

- compare_delta()
- Doesn’t Ever Leave The Airport?

Analytics

- Performs 2 statistical t-tests:
  - Comparing number of delayed flights where Delta was a possible carrier to flights where Delta was not a possible carrier
    - Not statistically significant
  - Comparing number of minutes that a flight was delayed where Delta was a possible carrier to flights where Delta was not a possible carrier
    - Statistically significant

```
Does Delta actually delay more than other airlines?
The p-value is: 0.22718375867068558
The result is not significantly significant. The p-value is > 0.05, therefore the increased number of delayed flights where Delta was a carrier is likely by chance.
The p-value is: 0.015981673642602837
The result is significantly significant. The p-value is < 0.05, therefore the increased number of minutes that flights were delayed by where Delta was a carrier is likely not by chance.
```

Thank you!

Any questions?