Are flights coming from the North delayed more?

Databases Group Project WS 23/24

Group U: *Wilson Liu, Cajus Marvin Schneider*

Objective 1: Wilson Liu

Objective 2: Cajus Schneider

Objective 3: Wilson Liu

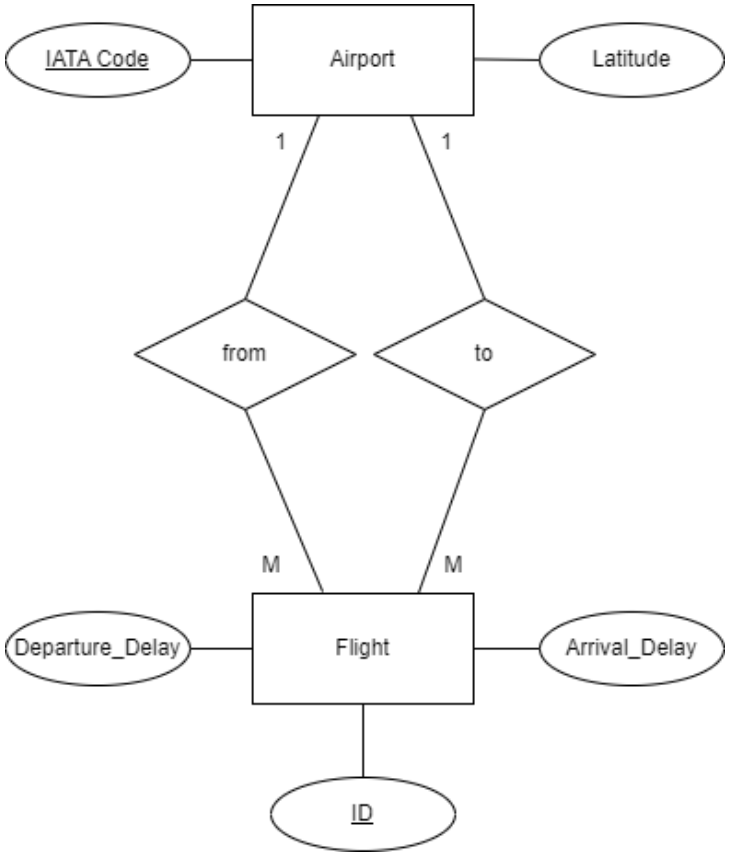
Objective 4: Cajus Schneider

Concept

Dataset: 2015 Flight Delays and Cancellations (link)

- **Research problem:** We would like to investigate whether flights coming from/going to more northern airports typically have longer delays and to what extent? Our current plan to answer this is by ingesting the data from .csv files to a SQL-based database, making the relevant queries on airport latitude, delay time, etc., and graphing the relationship using libraries like Pandas or Seaborn.
- **Motivation:** In Canada there is a common rumor, concerning this research problem. We are curious to see whether this common rumor is based in fact and statistics.
- **Method:** A relational database is ideal for this project as we will need to search for and connect information about two different entities – airports and flights. Also, you have multiple entities with more complex relationships (e.g., 1:M), which would not naturally lend itself to being represented in a spreadsheet or just one .csv file. The airlines and airports can be saved as foreign keys in the *flights* relation, which naturally comes from the dataset.

Objective 2 –
data modeling and data ingestion




```
(1) /*Downloaded Datasets*/
(2) /*Create Entities*/
CREATE TABLE Airport(
            IATA_CODE VARCHAR(8) PRIMARY KEY,
            LATITUDE FLOAT);
CREATE TABLE Flight(
            ID SERIAL PRIMARY KEY,
            DEPARTURE_DELAY FLOAT,
            ARRIVAL_DELAY FLOAT,
            ORIGIN_AIRPORT VARCHAR(8),
            DESTINATION_AIRPORT VARCHAR(8),
            FOREIGN KEY (ORIGIN_AIRPORT) REFERENCES Airport(IATA_CODE) ON
            DELETE CASCADE,
            FOREIGN KEY (DESTINATION_AIRPORT) REFERENCES Airport(IATA_CODE)
            ON DELETE CASCADE);
(3) /*Ingest Airport Data
Using temporary table to load .csv file*/
(4) /*Create temporary table*/
(5) /*Load .csv file in Temporary Table*/
(6) /*ingest data from temporary table into Airport table*/
(7) /*Ingest Flight Data
Using temporary table to load .csv file*/
(8) /*Create temporary table*/
(9) /*Load data from temporary table*/
(10)/*ingest data from temporary table into Flight  table*/
(11)/*Not importing tuples where FK Values are missing (caused an error before)*/
(12)/*Delete temporary tables*/
```

Objective 3 – database queries

To quantify 'delay', for the purposes of this study we defined it as departure delay.

Early flight departures (with a negative value for Departure_Delay) were set to a value of 0, so as to not skew the average delay time.

To get an accurate picture of whether latitude is correlated with flight delay, we will calculate the average delay in blocks of 5 degrees latitude.

```
UPDATE Flight
      SET DEPARTURE_DELAY = 0.0
      WHERE DEPARTURE_DELAY < 0.0;

SELECT A.Latitude, F.Departure_Delay
    FROM Airport AS A
INNER JOIN Flight AS F ON F.ORIGIN_AIRPORT = A.IATA_CODE;

SELECT AVG(Departure_Delay)
      FROM (Airport AS A
      INNER JOIN Flight as F
      ON F.ORIGIN_AIRPORT = A.IATA_CODE)
WHERE A.Latitude BETWEEN %s AND %s;
```

Objective 4

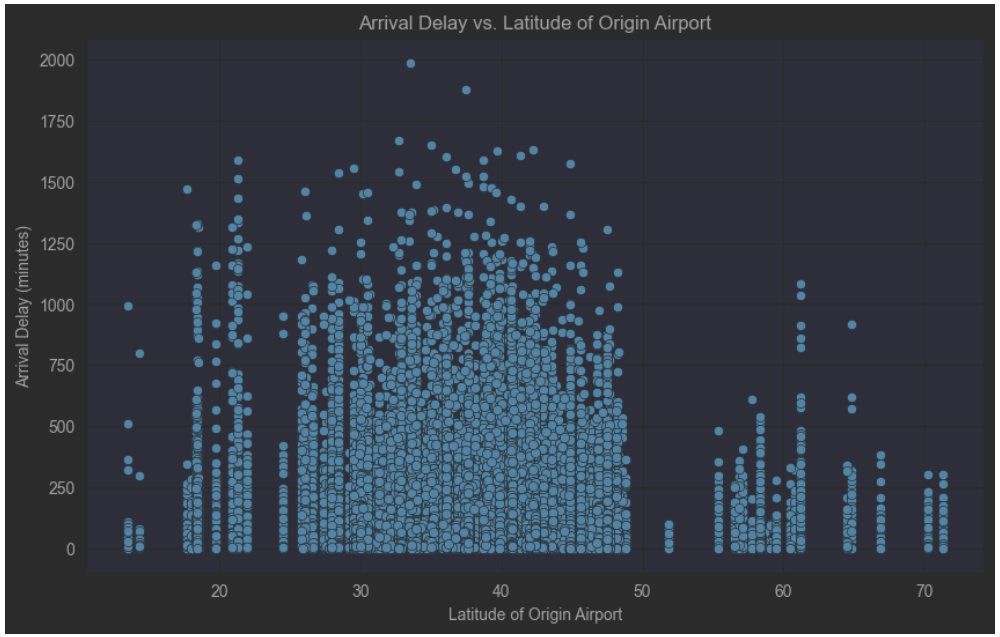
Using a mix of JDBC Prepared Statements (Statement 1,2) and JDBC Callable Statements (Statement 3) we processed the data using pandas, matplotlib and seaborn libraries.

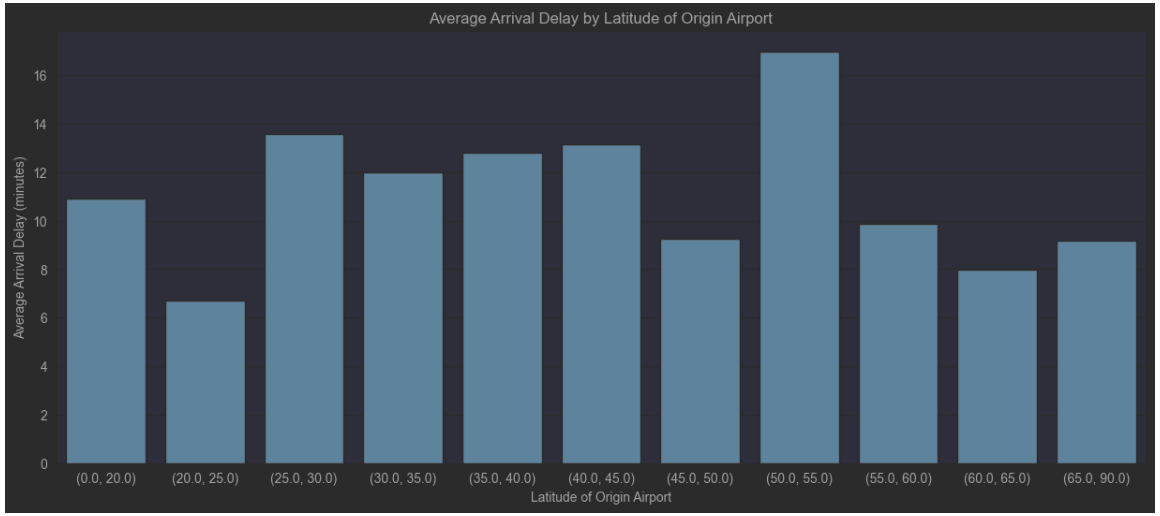
After executing the first Statement, we plot all arrival delays vs. the latitudes of the Origin Airport. We cannot see any correlation (corr = 0.00).

We then call Statement 3 by passing a list of parameters for the „%s“ placeholder. These parameters group airports by latitude and calculate an average. Again, we cannot see any correlation (corr = -0.06).

We fail to reject the H0: there is no relationship between the latitude of the airport and the departure delay.

Further investigation could focus on a) the relative number of flights coming from the North that have a general delay, b) control for the winter months only or c) test whether airplanes going to the North are delayed more





Reproducibility

This project on Github