Assignment 2 – revisit Aggregation

# Overview

This week we are going to practice aggregating (sums and counts) on a single table of flight data. Data is sampled from the bureau of transportation statistics ([BTS](https://www.transtats.bts.gov/DL_SelectFields.asp)). We are going to explore the data and answer simple questions like have the number of flights changed from January to May 2023, are flights more or less delayed on average between January and may, what is the worst day of the week to travel, what are the worst flights to take in January vs May and much more.

## What to turn in!

Please create a **word document** and submit that to me via Canvas, for homework assignment credit. Please include your name as part of your file name (e.g., Mike\_Ames\_Assignment\_2.docx). you’ll include your SQL select statement and table of results. I’ve provided example results for the first 5 questions to help guide you through this process.

## Preparation Steps

1. Create a schema, then use it.

drop schema if exists assignment\_2;

create schema assignment\_2;

use assignment\_2;

1. Load the SQL query file ***create\_table\_flights.sql*** into your workbench, run through the queries to create a table called “flights”.
2. Check the number of records in the table flights using the following code. There should be 23449 records.

SELECT count(\*) FROM assignment\_2.flights;

1. DESCRIBE the table and check fields and their respective data types.

A screenshot of a computer

Description automatically generated

# what to turn in!

*Be sure to follow the TASKs diligently! You will need to turn in a* ***word document*** *with your SQL code and Results in screenshots.*

* *Your code assignment\_2\_ <your\_name>.docx*

*See the example question/answer to understand the expected format of your submission. Unlimited resubmissions are allowed up to the deadline for the project, note the late assignment policy.*

## Results & analysis

### Result 0 – flights by airline

Example Question: Count the flights by airline, your SQL should produce something that looks like this:

|  |  |
| --- | --- |
| airline | flight\_count |
| American Airlines Inc. | 5604 |
| Delta Air Lines Inc. | 4806 |
| Southwest Airlines Co. | 4615 |
| United Air Lines Inc. | 4198 |
| Alaska Airlines Inc. | 1250 |
| JetBlue Airways | 946 |
| Spirit Air Lines | 868 |
| Frontier Airlines Inc. | 537 |
| Allegiant Air | 357 |
| Hawaiian Airlines Inc. | 268 |

Example answer:

# RES00: Query

select

airline,

count(\*) as flight\_count

from flights

group by airline

order by flight\_count desc;

# RES00: Result

A screenshot of a computer

Description automatically generated

### Result 1 – What are the worst arilines by average arr\_delay (arrival delay)

Count the flights and average the arrival delay (using the column ***arr\_delay***: The difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.) by airline, who is the worst? Your SQL should produce something that looks like this:

|  |  |  |
| --- | --- | --- |
| **airline** | **flight\_count** | **avg\_arr\_delay** |
| Frontier Airlines Inc. | 537 | 22.9441 |
| Hawaiian Airlines Inc. | 268 | 18.8507 |
| Spirit Air Lines | 868 | 12.6336 |
| JetBlue Airways | 946 | 9.9334 |
| United Air Lines Inc. | 4198 | 8.7063 |
| American Airlines Inc. | 5604 | 6.0089 |
| Allegiant Air | 357 | 4.9832 |
| Southwest Airlines Co. | 4615 | 4.7666 |
| Alaska Airlines Inc. | 1250 | 2.4112 |
| Delta Air Lines Inc. | 4806 | 0.3604 |

### Result 2 – did they get better between january and may?

Average arrival delay and flight count by airline and month, use date\_format function the %M format and fl\_date to create a month variable. Your result should look something like this.

# Hint: SELECT date\_format(fl\_date, '%M') as month\_name FROM flights;

|  |  |  |  |
| --- | --- | --- | --- |
| **month\_name** | **airline** | **flight\_count** | **avg\_arr\_delay** |
| January | Frontier Airlines Inc. | 258 | 22.2442 |
| January | Spirit Air Lines | 430 | 16.5767 |
| January | Hawaiian Airlines Inc. | 132 | 15.0909 |
| January | United Air Lines Inc. | 1984 | 10.2354 |
| January | American Airlines Inc. | 2711 | 7.6831 |
| January | Allegiant Air | 178 | 6.5955 |
| January | JetBlue Airways | 449 | 5.9421 |
| January | Delta Air Lines Inc. | 2290 | 5.4777 |
| January | Southwest Airlines Co. | 2246 | 5.2253 |
| January | Alaska Airlines Inc. | 604 | 4.3063 |
| May | Frontier Airlines Inc. | 279 | 23.5914 |
| May | Hawaiian Airlines Inc. | 136 | 22.5000 |
| May | JetBlue Airways | 497 | 13.5392 |
| May | Spirit Air Lines | 438 | 8.7626 |
| May | United Air Lines Inc. | 2214 | 7.3360 |
| May | American Airlines Inc. | 2893 | 4.4400 |
| May | Southwest Airlines Co. | 2369 | 4.3318 |
| May | Allegiant Air | 179 | 3.3799 |
| May | Alaska Airlines Inc. | 646 | 0.6393 |
| May | Delta Air Lines Inc. | 2516 | -4.2973 |

### Result 3 – can you create a January and may column?

Using similar logic to above, create a three variables january\_mean\_arr\_delay, may\_mean\_arr\_delay, january\_vs\_may\_change by airline (where january\_vs\_may\_change = may\_mean\_arr\_delay - january\_mean\_arr\_delay)

# Hint: avg(case when date\_format(fl\_date, '%M') = 'january' then arr\_delay else 0 end) as january\_mean\_arr\_delay

Your result should look something like this.

|  |  |  |  |
| --- | --- | --- | --- |
| **airline** | **january\_mean\_arr\_delay** | **may\_mean\_arr\_delay** | **january\_vs\_may\_change** |
| Delta Air Lines Inc. | 2.6101 | -2.2497 | -4.8598 |
| Spirit Air Lines | 8.2120 | 4.4217 | -3.7903 |
| Alaska Airlines Inc. | 2.0808 | 0.3304 | -1.7504 |
| Allegiant Air | 3.2885 | 1.6947 | -1.5938 |
| American Airlines Inc. | 3.7168 | 2.2921 | -1.4247 |
| United Air Lines Inc. | 4.8373 | 3.8690 | -0.9683 |
| Southwest Airlines Co. | 2.5430 | 2.2236 | -0.3194 |
| Frontier Airlines Inc. | 10.6872 | 12.2570 | 1.5698 |
| Hawaiian Airlines Inc. | 7.4328 | 11.4179 | 3.9851 |
| JetBlue Airways | 2.8203 | 7.1131 | 4.2928 |

### Result - 4 What is the worst day of the week to travel baseed on average dep\_delay?

Summarize to get the average departure delay(dep\_delay), average arrival delay and flight count by day of week. Your result should look something like this.

# Hint: SELECT DATE\_FORMAT(fl\_date, '%W') AS day\_of\_week

|  |  |  |  |
| --- | --- | --- | --- |
| day\_of\_week | avg\_dep\_delay | avg\_arr\_delay | flight\_count |
| Wednesday | 15.9487 | 10.4949 | 3314 |
| Monday | 13.1416 | 7.2341 | 3969 |
| Sunday | 11.8685 | 5.8851 | 3368 |
| Friday | 10.8262 | 5.2961 | 3164 |
| Thursday | 10.4317 | 6.2749 | 3197 |
| Tuesday | 9.7396 | 3.7385 | 3713 |
| Saturday | 7.3025 | 0.9068 | 2724 |

### Result 5 – is it better to fly on weekend or week day?

Similar to result 4, create a “week day indicator” and compare the average departure delay vs arrival delay. Note with date\_format with %w Sunday is 0 and Saturday is 6. Your result should look something like this.

|  |  |  |  |
| --- | --- | --- | --- |
| week\_day\_indicator | avg\_dep\_delay | avg\_arr\_delay | flight\_count |
| Weekend | 9.8268 | 3.6591 | 6092 |
| Weekday | 12.0286 | 6.5790 | 17357 |

Here’s a hint of how I did it, but feel free to use your own logic.

# Hint:

CASE

WHEN DATE\_FORMAT(fl\_date, '%w') = 0 OR DATE\_FORMAT(fl\_date, '%w') = 6

THEN 'Weekend'

ELSE 'Weekday' END AS week\_day\_indicator

### Result 6 - What’s the average, min, max distance flown by airline?

Using airline and ***distance*** to calculate the average, min, max distance flown by airline. Return 4 columns which are airline, mean\_distance, max\_distance, min\_distance. Sort the result by mean\_distance in ascending order.

### Result 7 - What’s the average, min, max air\_time flown by airline?

Using airline and ***air\_time*** calculate the average, min, max time by airline. Return 4 columns which are airline, mean\_time, max\_time, min\_dtime. Sort the result by mean\_time in ascending order.

### Result 8 – how many flights originate from Florida and what’s the average departure delay by airline?

On your way back from Florida, what’s the average departure delay? Using airline and ***origin\_state\_nm*** to calculate the average departure delay by airline. Return 4 columns which are airline, origin\_state\_nm (which should be Florida), flight\_count, avg\_dep\_delay. Sort the result by avg\_dep\_delay in ascending order.

### Result 9 – Top 5 longest flights by max flight time

Return 4 columns: origin\_city\_name, dest\_city\_name, max\_airtime, and max\_airtime\_hrs (i.e. divide max\_airtime by 60 to get an approximate hour). Order by max\_air\_time descending and get the top 5 using limit.

### Result 10 – TOP 5 worst origin to destination

Return 5 columns: origin\_city\_name, dest\_city\_name, mean\_airtime, and mean\_airtime\_hrs (i.e. divide mean\_airtime by 60 to get an approximate hour), flight counts. Having the average arrival delay is > 15 minutes and more than 10 flights, order by mean\_airtime descending and get the top 5 using limit.

### Result 11 – Best flights for early arrivals

Return 4 columns: origin\_city\_name, dest\_city\_name, mean of arr\_delay, and flight counts. Having average arrival delay is < -10 minutes (i.e. early arrivals) and having more than 20 flights. Order by mean\_ arr\_delay ascending and get the top 5 using limit.

### REsult 12 - What flights are likely to get a weather delay?

Return the same columns as result 11, but select flights which weather delay > 0, average arr\_delay is greater than 15 minutes, and flight count greater than 2.

### Result 13 – your analysis # 1

Think of a specific question you want to answer using the data table. Write a SQL query to obtain the relevant data that answers your question. Your question should be interesting and meaningful, and your query should include at least one aggregate function, one WHERE and HAVING statements.

### Result 14 – your analysis # 2

Think of a specific question you want to answer using the data table. Write a SQL query to obtain the relevant data that answers your question. Your question should be interesting and meaningful, and your query should include at least one CASE – WHEN structure.

### Result 15 – your analysis # 3

Think of a specific question you want to answer using the data table. Write a SQL query to obtain the relevant data that answers your question. Your question should be interesting and meaningful, and your query should include at least one subquery.

## Rubric

*Out of a total of 100 points*

* *Question 1 to 5: 6 points each*
* *Question 6 to 15: 7 points each*

*For each question, please turn in with:*

* *Question and Result #*
* *SQL query – your select statement to solve the question*
* *Result in a screenshot*