Assignment #4: Advanced Spark SQL on Flight Data
Max Levine

My repository is at https://github.com/maxlevinestuff/assignment-4-advanced-spark-sql-on-flight-data-maxlevinestuff

Setup/Process

For task 1, I determined the actual and scheduled flight times and got the discrepancies between them, ranking for each airline the flights with greatest discrepancies. First, I took the input and added two new columns, ScheduledTravelTime and ActualTravelTime, which were gotten from taking the difference of ScheduledArrival and ScheduledDeparture, and ActualArrival and ActualDeparture, respectively, and casting them as longs. I then calculated a Discrepancy column as the absolute difference between these. I then use a Window to partition by CarrierCode and order by Discrepancy. I then use row_number() to compute a rank number within each CarrierCode which was sorted by Discrepancy, then filter to leave only those with rank 1. I then join the resulting data frame with carriers_df to include flight information, aliasing the two to avoid name conflicts with CarrierCode, and selecting relevant columns to be sent to the output csv.

For task 2, I ranked airlines by their on-time consistency, gotten from the standard deviation of the differences between their ActualDeparture and ScheduledDeparture. First, I took the input and added a DepartureDelay column, set to the difference between actual and scheduled departure. Grouping by CarrierCode, I then calculated the standard deviation of departure delay within that carrier, also counting the number of flights to filter out carriers with less than 100 flights. I then join with carriers_df to get the carrier names, select relevant columns, then, ordering by the standard deviation, output to csv.

For task 3, I calculated the ratio of canceled flights (where ActualDeparture is null) to total flights for each origin-destination pair. First, I created an IsCanceled column, which is 1 if ActualDeparture is null, and 0 otherwise. I then grouped by origin and destination; I summed up the IsCanceled column as NumCanceled, counted the total number of flights as FlightNum, then divided the former by the latter to get a CancellationRate column. I stored the result in cancellation_stats. I then do two joins on cancellation_stats with the airports_df to get the airport name and city for both the origin and destination airport. Each join uses Origin/Destination from the cancellation_stats and corresponds it to the airport codes in airports_df. I order by CancellationRate and output to csv.

For task 4, I ranked each carrier and time of day (morning, afternoon, evening, and night) by average departure delay. First, I got a ScheduledHour column by extracting the hour from the scheduled departure from the flights.csv. Then, I created a TimeOfDay column, setting this to either “Morning”, “Afternoon”, “Evening”, or “Night”, depending on if, respectively, the ScheduledHour was between 6 and 12, 12 and 18, 18 and 24, and 24 and 6. I then calculated the DepartureDelay as the difference between departure and actual departure. Then, grouping by carrier and time of day, I calculated their AvgDepartureDelay by averaging their DepartureDelay. I then join with the carriers_df to get the carrier name, select the relevant columns ordered by time of day and average delay, and output to csv.

Results

Here are some of the results from task 1:
FlightNum,CarrierName,Origin,Destination,ScheduledTravelTime,ActualTravelTime,Discrepancy,CarrierCode

1003,American Airlines,LAX,DEN,900.0,754.0,146.0,AA
2147,Air France,DFW,MIA,780.0,930.0,150.0,AF
7080,British Airways,SFO,ATL,840.0,989.0,149.0,BA

This shows that American Airlines had a discrepancy of 146, Air France had 150, and British Airways had 149.

Here are the top 3 results from task 2:

CarrierName,NumFlights,StdDevDepartureDelay
British Airways,1059,42.86507919688348
American Airlines,957,42.94906780833314
Lufthansa,986,42.97764237708067

British Airways had the most consistent and lowest standard deviation in departure delays at 42.87, followed by American Airlines at 42.95, followed by Lufthansa at 42.98.

Here are the top 3 results for task 3:

OriginAirport,OriginCity,DestinationAirport,DestinationCity,CancellationRate
Dallas/Fort Worth International,Dallas,Los Angeles International,Los Angeles,21.904761904761905
Miami International,Miami,Dallas/Fort Worth International,Dallas,18.095238095238095
Logan International,Boston,San Francisco International,San Francisco,17.857142857142858

The flight type with the highest cancellation rate had a rate of 21.90%, followed by 18.10%, followed by 17.86%.

Here are the top 3 results from task 4:

CarrierName,TimeOfDay,AvgDepartureDelay
Lufthansa,Afternoon,2251.2
Qantas,Afternoon,2309.6137339055795
Delta Airlines,Afternoon,2601.9718309859154

This shows that for the time period of afternoon, Lufthansa had the lowest average departure delay, followed by Qantas, then Delta Airlines.