**Assignment 2**

1. Spark Data Analysis using Spark SQL –

**(Participants are open to apply EDA and Analysis using queries using Spark on Pyspark)**

**Dataset:** Flights\_Delay.csv

Dataset Description:

ID: Rows ID

YEAR: 2015

MONTH: 1-12

DAY: 1-31

DAY\_OF\_WEEK: 1 (Monday) - 7 (Sunday)

AIRLINE: Airline CODE

FLIGHT\_NUMBER: Flight Number

TAIL\_NUMBER: Flight’s tail number

ORIGIN\_AIRPORT: Origin IATA airport code

DESTINATION\_AIRPORT: Destination IATA airport code

SCHEDULED\_DEPARTURE: Actual departure time (local, hhmm)

DEPARTURE\_TIME: Scheduled departure time (local, hhmm)

DEPARTURE\_DELAY: Departure delay, in minutes

TAXI\_OUT: Taxi out time in minutes

WHEELS\_OFF:

SCHEDULED\_TIME: Scheduled arrival time (local, hhmm)

ELAPSED\_TIME: in Minutes

AIR\_TIME: in Minutes

DISTANCE: in Miles

WHEELS\_ON:

TAXI\_IN: Taxi in time, in minutes

SCHEDULED\_ARRIVAL: Scheduled arrival time (local, hhmm)

ARRIVAL\_TIME: Actual arrival time (local, hhmm)

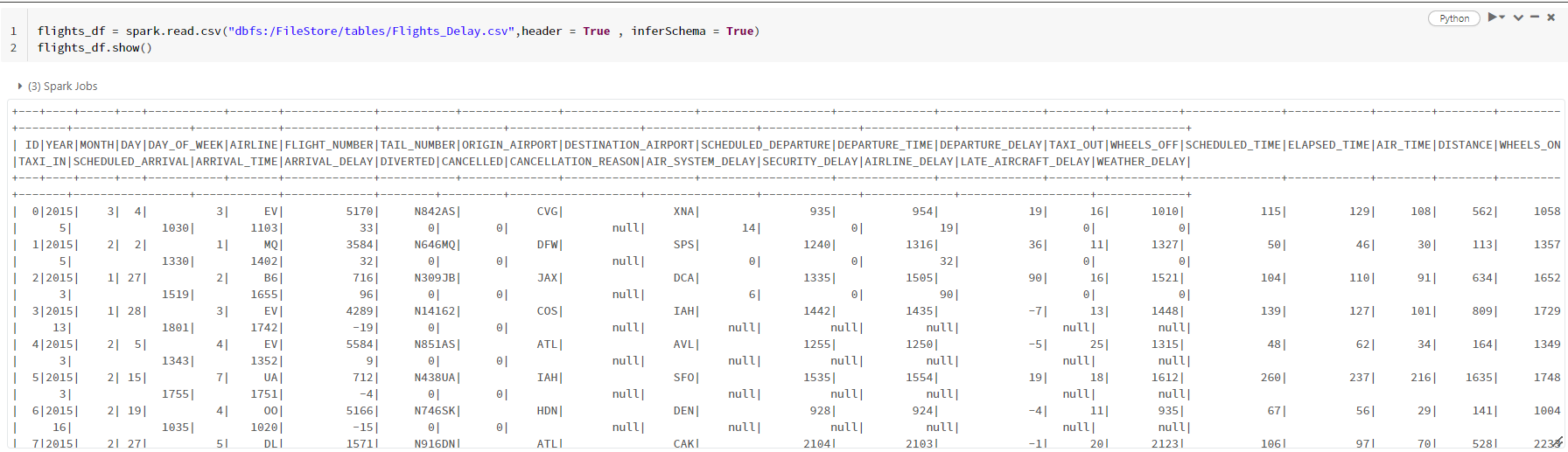
ARRIVAL\_DELAY: Arrival delay, in minutes

DIVERTED: 1 = yes, 0 = no

CANCELLED: 1 = yes, 0 = no

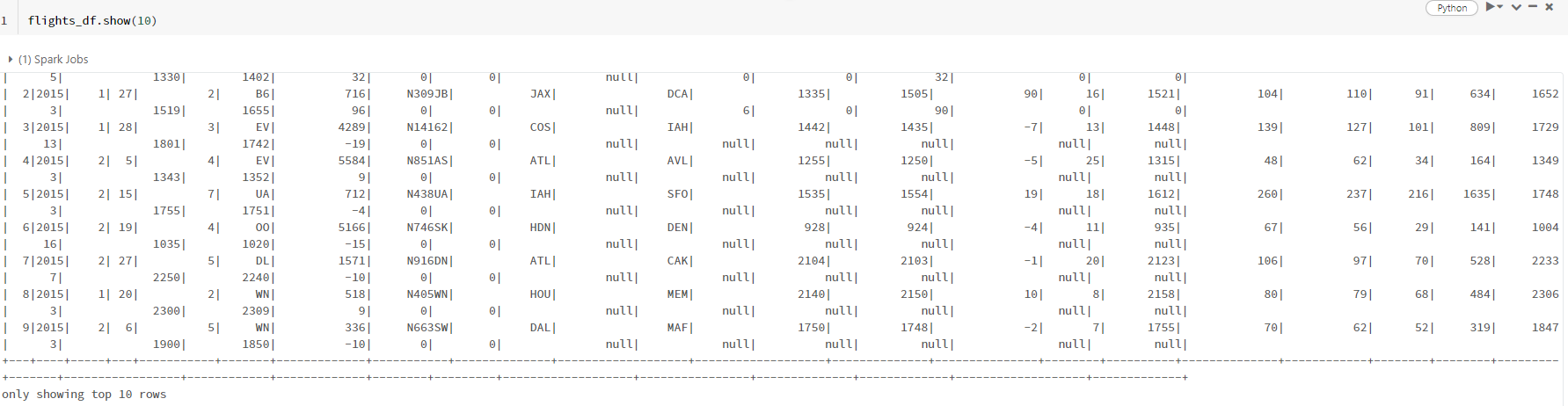
1. Create Airline Pysaprk sql dataFrame.

Answer:



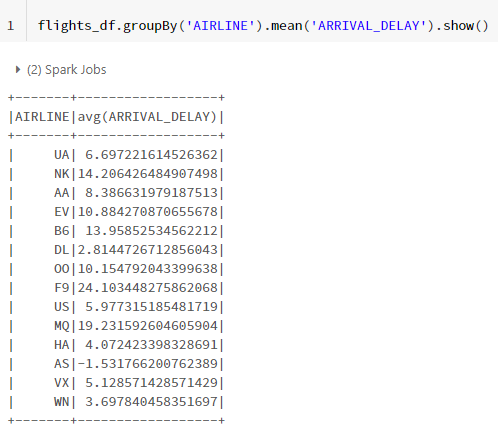
1. Describe the table schema & show 10 rows of Dataset

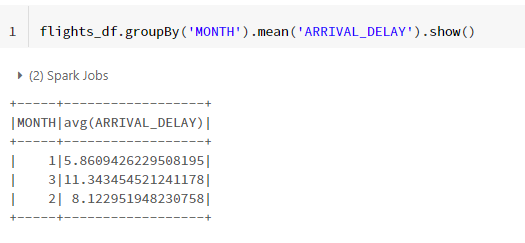
Answer:



Perform Data Analysis & EDA using Pyspark on this Dataset showing following into considerations:

1. Average arrival delay caused by airlines.

Answer:

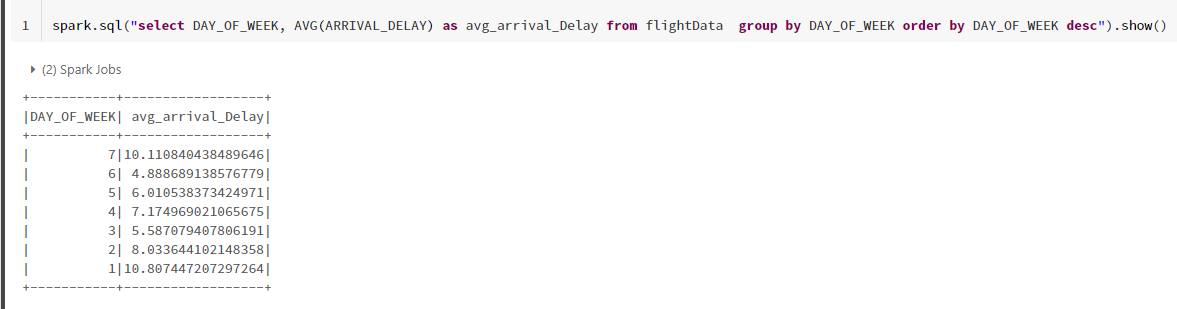
1. Days of months with average Arrival delays.

Answer:

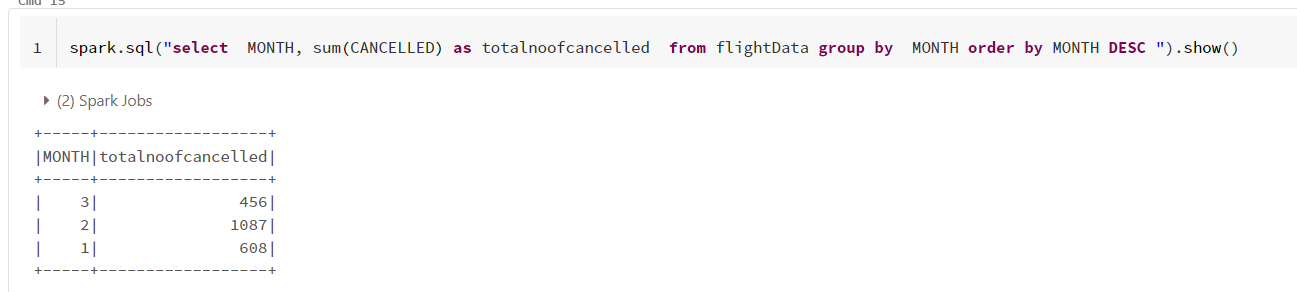
1. Days of months with median Departure delays.

Answer: 

1. Arrange weekdays with respect to the average arrival delays.

Answer: 

1. Show Analysis for each month with total number of cancellations. (Descending).

Answer: 

1. Show the top 10 busiest airport. [Hint: Sum Total of Arrival and Departures from an Airport].

Answer:

1. Show the airlines that make the maximum number of cancellations.

Answer:

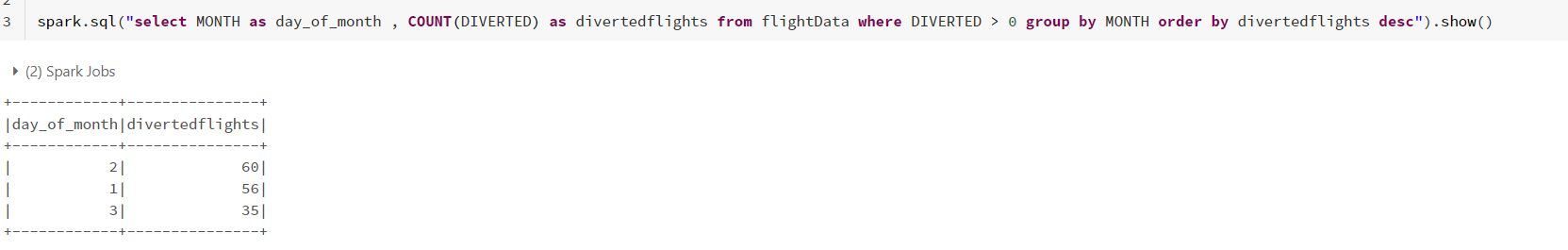


1. Find and order airlines in descending that make the most number of diversions.

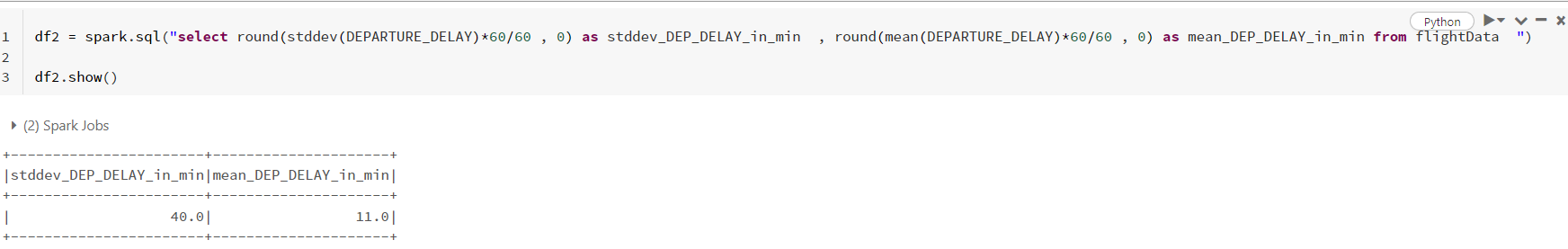
Answer:



1. Show days of month that seen the most number of diversion.

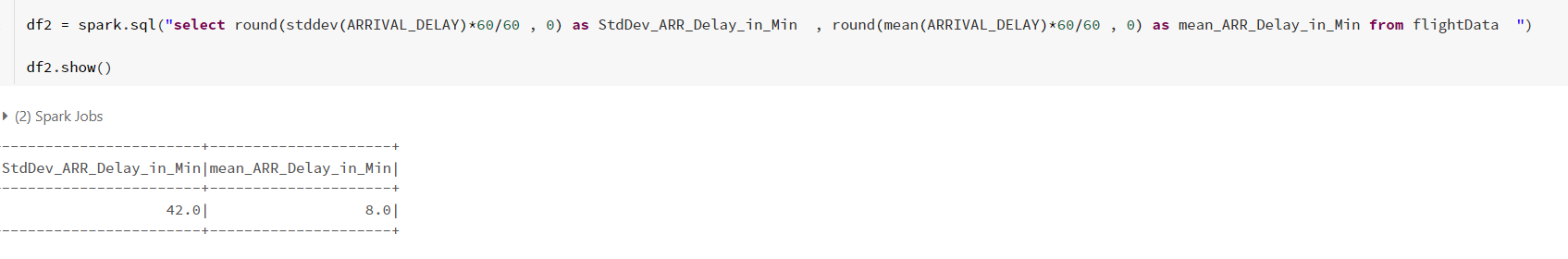
Answer: 

1. Calculate mean and standard deviation of departure delay for all flights in minutes.

Answer: 

1. Calculate mean and standard deviation of arrival delay for all flights in minutes.

Answer:



1. Find all diverted Route from a source to destination Airport & which route is the most diverted.

Answer:

1. Finding AIRLINES with its total flight count, total number of flights arrival delayed by more than 30 Minutes, % of such flights delayed by more than 30 minutes when it is not Weekends with minimum count of flights from Airlines by more than 10. Also Exclude some of Airlines 'AK', 'HI', 'PR', 'VI' and arrange output in descending order by % of such count of flights.