Spark SQL

Q.1>No. Of Flights Cancelled in Each Month from 2018?

Ans=

```
spark.sql("""
SELECT FL_MONTH, COUNT(CANCELLED) AS TOTAL_CANCELLED
FROM df
WHERE CANCELLED = 1
GROUP BY FL_MONTH
ORDER BY FL_MONTH
LIMIT 50
""").show()
```

```
[Stage 655:=============================================>       (6 + 1) / 7]
```

```
+--------+---------------+
|FL_MONTH|TOTAL_CANCELLED|
+--------+---------------+
|    1.0|          17169|
|    2.0|           8976|
|    3.0|          17280|
|    4.0|           6251|
|    5.0|           7155|
|    6.0|          10694|
|    7.0|          11083|
|    8.0|          12353|
|    9.0|           8132|
|   10.0|           4485|
|   11.0|           6254|
|   12.0|           6752|
+--------+---------------+
```

Observation=

The lowest number of flight cancellations is in October and the largest number in March.June,July ,August have the most number of cancelled flights.

Q.2>No. Of Flights Detoured in Each Month in 2018

Ans=

In [50]:

```
spark.sql("""
SELECT origin, COUNT(cancelled), op_carrier
FROM df where cancelled = 1
GROUP BY origin, op_carrier
""").show()
```

```
+------+---------------+----------------+
|origin|count(cancelled)|      op_carrier|
+------+---------------+----------------+
|   SLC|             16|     Mesa Airline|
|   PWM|             62|       ExpressJet|
|   TOL|             14| SkyWest Airlines|
|   HNL|             15|American Airlines|
|   ROW|              1|     Mesa Airline|
|   MIA|             39|Frontier Airlines|
|   MYR|             15|   Delta Airlines|
|   BTR|             77|       ExpressJet|
|   CMH|             20|     Mesa Airline|
|   GSP|             11|   Delta Airlines|
|   EUG|             12| SkyWest Airlines|
|   PIT|             30|     Mesa Airline|
|   AUS|             13| Republic Airways|
|   FAR|              8|       ExpressJet|
|   EVV|              7| SkyWest Airlines|
|   GSO|             66|       ExpressJet|
|   CLT|             55| SkyWest Airlines|
|   GGG|             29|        Envoy Air|
|   DEN|            199|Frontier Airlines|
|   MKG|             39| SkyWest Airlines|
+------+---------------+----------------+
```

Observation=

Mesa Airlines has the fewest detoured flights i.e.,1 while frontier Airlines has the most i.e.,199.

Q.3>Most cancelled flight

Ans=

In [70]:

```
spark.sql(""" SELECT op_carrier, COUNT(cancelled) AS Total
FROM df where cancelled=1
GROUP BY cancelled, op_carrier
ORDER BY Total desc
LIMIT 10 """).show()
```

```
[Stage 661:=============================================>       (6 + 1) / 7]

+------------------+-----+
|       op_carrier|Total|
+------------------+-----+
|Southwest Airlines|18275|
| American Airlines|14945|
|      PSA Airlines|11870|
|        Envoy Air|10655|
|  SkyWest Airlines|10610|
|  Republic Airways|10100|
|   JetBlue Airways| 6419|
|      Endeavor Air| 6355|
|       ExpressJet| 5670|
|      Mesa Airline| 5530|
+------------------+-----+
```

Observation=

Southwest Airlines had the most cancelled flights i.e, 18275.

Q.4>Flight has least amount of delay(planned departure = actual departure)

Ans=

In [76]:

```
spark.sql("""SELECT OP_CARRIER_FL_NUM, MIN(abs(CRS_DEP_TIME - DEP_DELAY))
AS min_delay
FROM df
GROUP BY OP_CARRIER_FL_NUM
```

```
        ORDER BY min_delay ASC
        LIMIT 1
        """).show()
```

```
[Stage 667:=======================================>             (5 + 2) / 7]
+----------------+----------------+
|OP_CARRIER_FL_NUM|       min_delay|
+----------------+----------------+
|          1664|0.1333333333333333|
+----------------+----------------+
```

Observation=

Flight number 1664 has minimum delay.

Q.5>No. of Flights canceled from origin city by airline flights in Each Year i.e. 2018.

In [77]:

```
spark.sql("""
SELECT origin, COUNT(cancelled), op_carrier
FROM df where cancelled=1
GROUP BY origin, op_carrier
""").show()
```

```
[Stage 670:================================================>      (6 + 1) / 7]
+------+---------------+----------------+
|origin|count(cancelled)|      op_carrier|
+------+---------------+----------------+
|   SLC|             16|     Mesa Airline|
|   PWM|              62|       ExpressJet|
|   TOL|             14| SkyWest Airlines|
|   HNL|             15|American Airlines|
|   ROW|              1|     Mesa Airline|
|   MIA|             39|Frontier Airlines|
|   MYR|             15|   Delta Airlines|
|   BTR|             77|       ExpressJet|
|   CMH|             20|     Mesa Airline|
|   GSP|             11|   Delta Airlines|
|   EUG|             12| SkyWest Airlines|
|   PIT|             30|     Mesa Airline|
|   AUS|             13| Republic Airways|
```

```
|  FAR|            8|      ExpressJet|
|  EVV|            7| SkyWest Airlines|
|  GSO|           66|      ExpressJet|
|  CLT|           55| SkyWest Airlines|
|  GGG|           29|       Envoy Air|
|  DEN|          199|Frontier Airlines|
|  MKG|           39| SkyWest Airlines|
+------+---------------+-----------------+

only showing top 20 rows
```

Observation=

Most of the cancelled flights are from frontier airlines in the Denver International Airport region i.e,199 and the lowest number is from mesa airlines from Roswell air Center region i.e,1.

Q.6>Airport has high delay(busy) most of the time

Ans=

In [79]:

```
spark.sql("""SELECT origin, AVG(dep_delay) AS avg_delay
FROM df
WHERE dep_delay > 0
GROUP BY origin
ORDER BY avg_delay DESC
LIMIT 1
""").show()
```

```
[Stage 673:=============================================>       (6 + 1) / 7]

+------+-----------------+
|origin|        avg_delay|
+------+-----------------+
|  DVL|2.3069841269841267|
+------+-----------------+
```

Observation=

Devils Lake Regional Airport frequently experiences significant delays (avg delay=2.30)

Q.7>airports have least amount of wheels off time (operations and management is good)

Ans=

In [91]:

```
spark.sql("""SELECT origin, avg(abs(wheels_off - DEP_TIME)) AS avg_wo_time
FROM df
GROUP BY origin
ORDER BY avg_wo_time ASC
LIMIT 10
""").show()
```

```
[Stage 706:=========================================>          (5 + 2) / 7]
+------+-----------------+
|origin|      avg_wo_time|
+------+-----------------+
|   OWB|11.046728971962617|
|   SMX|12.064705882352941|
|   OTZ|12.238095238095237|
|   OME|12.361794500723589|
|   HYA|12.511363636363637|
|   BKG|12.956521739130435|
|   YAK|13.530898876404494|
|   OGD|           13.744|
|   BRW|13.880952380952381|
|   LWS| 14.12290502793296|
+------+-----------------+
```

Observation=

Lewiston–Nez Perce Country airport has the least wheel-offs, indicating that its management and operations are excellent.

Q.8>Airports have high amount of wheals off time (operations and management is bad)

Ans=

In [92]:

```
spark.sql("""SELECT origin, avg(abs(wheels_off - DEP_TIME)) AS avg_wo_time
FROM df
GROUP BY origin
ORDER BY avg_wo_time desc
LIMIT 10
""").show()
```

```
[Stage 709:=============================================>       (6 + 1) / 7]

+------+-----------------+
|origin|      avg_wo_time|
+------+-----------------+
|   PPG|208.65573770491804|
|   JMS| 78.96924708377519|
|   FAI| 74.44431869624265|
|   SFO| 64.54136955030467|
|   LAX| 64.39248391459367|
|   JFK| 61.90717272842479|
|   YNG|             58.5|
|   ANC| 57.92597246503497|
|   UIN|   49.270207852194|
|   LGA| 48.57906184693933|
+------+-----------------+
```

Observation=

Pago Pago International airport has the highest wheel offs time (208.5), indicating that its management and operations are excellent.

Q.9>Appropriate time to reach at airport for travelers

Ans=

In [95]:

```
spark.sql(""" select sum(abs(CRS_DEP_TIME - DEP_TIME))/count(DEP_TIME) as
REACH_TIME from df""").show()
```

```
[Stage 715:==============================================>       (6 + 1) / 7]
+------------------+
|       REACH_TIME|
+------------------+
|34.781221267773056|
+------------------+
```

Observation=

Travelers should arrive around 35 minutes before the takeoff.