Consultas en PySpark con Joins

En este proyecto buscaremos analizar tres datasets relacionales aplicaremos filtrados especificos para analizar y poder responder las preguntas clave en nuestro proyecto, para así denotar los valores de las distintas variables que lo conforman.

En este proyecto trabajamos con tres tablas en PySpark: • flights_spark: contiene información de vuelos (origen, destino, fechas, retrasos, avión utilizado, etc.). • airports_spark: contiene información de aeropuertos (código FAA, nombre, ciudad, estado, latitud, longitud). • planes_spark: contiene información de aviones (número de cola tailnum, fabricante, modelo, año de fabricación, etc.).

1. Aeropuertos destino más frecuentes
2. Mostrar vuelos junto con fabricante del avión
3. Vuelos agrupados por fabricante del avión
4. Contar vuelos por aeropuerto (origen y destino juntos)
5. Promedio de distancia por aeropuerto de origen
6. Vuelos de más de 2000 km con modelo del avión
7. Número de vuelos por año de fabricación del avión
8. Vuelos con origen y destino mostrando ciudad de ambos aeropuertos
9. Promedio de retraso de salida por aeropuerto
10. Promedio de retraso de llegada por modelo de avión
11. Contar vuelos por ciudad origen y destino
12. Vuelos con aviones Boeing
13. Contar vuelos por fabricante
14. Vuelos con más de 2 horas de retraso en salida y llegada
15. Promedio de distancia recorrida por fabricante de avión

```
pip install findspark
```

```
Collecting findspark
  Downloading findspark-2.0.1-py2.py3-none-any.whl.metadata (352 bytes)
Downloading findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1
```

```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.12/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.12/dist-packages (from pyspark) (0.10.9.7)
```

```
import findspark
findspark.init()
import pyspark
from pyspark.sql import SparkSession
```

```
#Create Spark Session for builder
from pyspark.sql import SparkSession
spark=SparkSession.builder.master("local[1]")\
.appName("SparkByExamples.com")\
.getOrCreate()
print(spark.sparkContext)
print("Spark App Name :"+spark.sparkContext.appName)
```

```
<SparkContext master=local[1] appName=SparkByExamples.com>
Spark App Name :SparkByExamples.com
```

```
from google.colab import drive
# Montar Google Drive
drive.mount('/content/drive', force_remount=True)
```

```
import pandas as pd
import numpy as np
flightsdf = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/flights_small.csv')
airportsdf = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/airports.csv')
planesdf = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/planes.csv')
```

```
flights_spark = spark.read.csv('/content/drive/MyDrive/Colab Notebooks/flights_small.csv', header= True, inferSchema = True)
airports_spark = spark.read.csv('/content/drive/MyDrive/Colab Notebooks/airports.csv', header= True, inferSchema = True)
planes_spark = spark.read.csv('/content/drive/MyDrive/Colab Notebooks/planes.csv', header= True, inferSchema = True)
```

```
flights_spark.show()
```

```
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
|year|month|day|dep_time|dep_delay|arr_time|arr_delay|carrier|tailnum|flight|origin|dest|air_time|distance|hour|minute|
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
|2014|   12|  8|     658|       -7|     935|       -5|     VX| N846VA|  1780|   SEA| LAX|     132|     954|   6|    58|
|2014|    1| 22|    1040|        5|    1505|        5|     AS| N559AS|   851|   SEA| HNL|     360|    2677|  10|    40|
|2014|    3|  9|    1443|       -2|    1652|        2|     VX| N847VA|   755|   SEA| SFO|     111|     679|  14|    43|
|2014|    4|  9|    1705|       45|    1839|       34|     WN| N360SW|   344|   PDX| SJC|      83|     569|  17|     5|
|2014|    3|  9|     754|       -1|    1015|        1|     AS| N612AS|   522|   SEA| BUR|     127|     937|   7|    54|
|2014|    1| 15|    1037|        7|    1352|        2|     WN| N646SW|    48|   PDX| DEN|     121|     991|  10|    37|
|2014|    7|  2|     847|       42|    1041|       51|     WN| N422WN|  1520|   PDX| OAK|      90|     543|   8|    47|
|2014|    5| 12|    1655|       -5|    1842|      -18|     VX| N361VA|   755|   SEA| SFO|      98|     679|  16|    55|
|2014|    4| 19|    1236|       -4|    1508|       -7|     AS| N309AS|   490|   SEA| SAN|     135|    1050|  12|    36|
|2014|   11| 19|    1812|       -3|    2352|       -4|     AS| N564AS|    26|   SEA| ORD|     198|    1721|  18|    12|
|2014|   11|  8|    1653|       -2|    1924|       -1|     AS| N323AS|   448|   SEA| LAX|     130|     954|  16|    53|
|2014|    8|  3|    1120|        0|    1415|        2|     AS| N305AS|   656|   SEA| PHX|     154|    1107|  11|    20|
|2014|   10| 30|     811|       21|    1038|       29|     AS| N433AS|   608|   SEA| LAS|     127|     867|   8|    11|
|2014|   11| 12|    2346|       -4|     217|      -28|     AS| N765AS|   121|   SEA| ANC|     183|    1448|  23|    46|
|2014|   10| 31|    1314|       89|    1544|      111|     AS| N713AS|   306|   SEA| SFO|     129|     679|  13|    14|
|2014|    1| 29|    2009|        3|    2159|        9|     UA| N27205|  1458|   PDX| SFO|      90|     550|  20|     9|
|2014|   12| 17|    2015|       50|    2150|       41|     AS| N626AS|   368|   SEA| SMF|      76|     605|  20|    15|
|2014|    8| 11|    1017|       -3|    1613|       -7|     WN| N8634A|   827|   SEA| MDW|     216|    1733|  10|    17|
|2014|    1| 13|    2156|       -9|     607|      -15|     AS| N597AS|    24|   SEA| BOS|     290|    2496|  21|    56|
|2014|    6|  5|    1733|      -12|    1945|      -10|     OO| N215AG|  3488|   PDX| BUR|     111|     817|  17|    33|
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
only showing top 20 rows
```

```
airports_spark.show()
```

```
+---+--------------------+---------------+-----------------+----+---+---+
|faa|                name|            lat|              lon| alt| tz|dst|
+---+--------------------+---------------+-----------------+----+---+---+
|04G|   Lansdowne Airport|     41.1304722|      -80.6195833|1044| -5|  A|
|06A|Moton Field Munic...|     32.4605722|      -85.6800278| 264| -5|  A|
|06C| Schaumburg Regional|     41.9893408|      -88.1012428| 801| -6|  A|
|06N|     Randall Airport|      41.431912|      -74.3915611| 523| -5|  A|
|09J|Jekyll Island Air...|     31.0744722|      -81.4277778|  11| -4|  A|
|0A9|Elizabethton Muni...|     36.3712222|      -82.1734167|1593| -4|  A|
|0G6|Williams County A...|     41.4673056|      -84.5067778| 730| -5|  A|
|0G7|Finger Lakes Regi...|     42.8835647|      -76.7812318| 492| -5|  A|
|0P2|Shoestring Aviati...|     39.7948244|      -76.6471914|1000| -5|  U|
|0S9|Jefferson County ...|     48.0538086|     -122.8106436| 108| -8|  A|
|0W3|Harford County Ai...|     39.5668378|      -76.2024028| 409| -5|  A|
|10C|  Galt Field Airport|     42.4028889|      -88.3751111| 875| -6|  U|
|17G|Port Bucyrus-Craw...|     40.7815556|      -82.9748056|1003| -5|  A|
|19A|Jackson County Ai...|     34.1758638|      -83.5615972| 951| -4|  U|
|1A3|Martin Campbell F...|     35.0158056|      -84.3468333|1789| -4|  A|
|1B9| Mansfield Municipal|     42.0001331|      -71.1967714| 122| -5|  A|
|1C9|Frazier Lake Airpark|54.0133333333333|-124.768333333333| 152| -8|  A|
|1CS|Clow Internationa...|     41.6959744|      -88.1292306| 670| -6|  U|
|1G3|  Kent State Airport|     41.1513889|      -81.4151111|1134| -4|  A|
|1OH|     Fortman Airport|     40.5553253|      -84.3866186| 885| -5|  U|
+---+--------------------+---------------+-----------------+----+---+---+
only showing top 20 rows
```

```
planes_spark.show()
```

```
+-------+----+--------------------+----------------+--------+-------+-----+-----+---------+
|tailnum|year|                type|    manufacturer|   model|engines|seats|speed|   engine|
+-------+----+--------------------+----------------+--------+-------+-----+-----+---------+
| N102UW|1998|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N103US|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N104UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N105UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
```

```
| N107US|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N108UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N109UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N110UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N111US|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N11206|2000|Fixed wing multi ...|          BOEING| 737-824|      2|  149|   NA|Turbo-fan|
| N112US|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N113UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N114UW|1999|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N117UW|2000|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N118US|2000|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N119US|2000|Fixed wing multi ...|AIRBUS INDUSTRIE|A320-214|      2|  182|   NA|Turbo-fan|
| N1200K|1998|Fixed wing multi ...|          BOEING| 767-332|      2|  330|   NA|Turbo-fan|
| N1201P|1998|Fixed wing multi ...|          BOEING| 767-332|      2|  330|   NA|Turbo-fan|
| N12114|1995|Fixed wing multi ...|          BOEING| 757-224|      2|  178|   NA|Turbo-jet|
| N121DE|1987|Fixed wing multi ...|          BOEING| 767-332|      2|  330|   NA|Turbo-fan|
+-------+----+--------------------+----------------+--------+-------+-----+-----+---------+
only showing top 20 rows
```

```
flights_spark.createOrReplaceTempView("flights")
airports_spark.createOrReplaceTempView("airports")
planes_spark.createOrReplaceTempView("planes")
```

```
spark.catalog.listTables()
```

```
[Table(name='airports', catalog=None, namespace=[], description=None, tableType='TEMPORARY', isTemporary=True),
 Table(name='flights', catalog=None, namespace=[], description=None, tableType='TEMPORARY', isTemporary=True),
 Table(name='planes', catalog=None, namespace=[], description=None, tableType='TEMPORARY', isTemporary=True)]
```

```
sqlDF = spark.sql("SELECT * FROM flights")
sqlDF.show()
```

```
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
|year|month|day|dep_time|dep_delay|arr_time|arr_delay|carrier|tailnum|flight|origin|dest|air_time|distance|hour|minute|
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
|2014|   12|  8|     658|       -7|     935|       -5|     VX| N846VA|  1780|   SEA| LAX|     132|     954|   6|    58|
|2014|    1| 22|    1040|        5|    1505|        5|     AS| N559AS|   851|   SEA| HNL|     360|    2677|  10|    40|
|2014|    3|  9|    1443|       -2|    1652|        2|     VX| N847VA|   755|   SEA| SFO|     111|     679|  14|    43|
|2014|    4|  9|    1705|       45|    1839|       34|     WN| N360SW|   344|   PDX| SJC|      83|     569|  17|     5|
|2014|    3|  9|     754|       -1|    1015|        1|     AS| N612AS|   522|   SEA| BUR|     127|     937|   7|    54|
|2014|    1| 15|    1037|        7|    1352|        2|     WN| N646SW|    48|   PDX| DEN|     121|     991|  10|    37|
|2014|    7|  2|     847|       42|    1041|       51|     WN| N422WN|  1520|   PDX| OAK|      90|     543|   8|    47|
|2014|    5| 12|    1655|       -5|    1842|      -18|     VX| N361VA|   755|   SEA| SFO|      98|     679|  16|    55|
|2014|    4| 19|    1236|       -4|    1508|       -7|     AS| N309AS|   490|   SEA| SAN|     135|    1050|  12|    36|
|2014|   11| 19|    1812|       -3|    2352|       -4|     AS| N564AS|    26|   SEA| ORD|     198|    1721|  18|    12|
|2014|   11|  8|    1653|       -2|    1924|       -1|     AS| N323AS|   448|   SEA| LAX|     130|     954|  16|    53|
|2014|    8|  3|    1120|        0|    1415|        2|     AS| N305AS|   656|   SEA| PHX|     154|    1107|  11|    20|
|2014|   10| 30|     811|       21|    1038|       29|     AS| N433AS|   608|   SEA| LAS|     127|     867|   8|    11|
|2014|   11| 12|    2346|       -4|     217|      -28|     AS| N765AS|   121|   SEA| ANC|     183|    1448|  23|    46|
|2014|   10| 31|    1314|       89|    1544|      111|     AS| N713AS|   306|   SEA| SFO|     129|     679|  13|    14|
|2014|    1| 29|    2009|        3|    2159|        9|     UA| N27205|  1458|   PDX| SFO|      90|     550|  20|     9|
|2014|   12| 17|    2015|       50|    2150|       41|     AS| N626AS|   368|   SEA| SMF|      76|     605|  20|    15|
|2014|    8| 11|    1017|       -3|    1613|       -7|     WN| N8634A|   827|   SEA| MDW|     216|    1733|  10|    17|
|2014|    1| 13|    2156|       -9|     607|      -15|     AS| N597AS|    24|   SEA| BOS|     290|    2496|  21|    56|
|2014|    6|  5|    1733|      -12|    1945|      -10|     OO| N215AG|  3488|   PDX| BUR|     111|     817|  17|    33|
+----+-----+---+--------+---------+--------+---------+-------+-------+------+------+----+--------+--------+----+------+
only showing top 20 rows
```

```
query = "SELECT origin, dest, COUNT(*) as N FROM flights GROUP BY origin, dest"
# Run the query
flights_counts = spark.sql(query)
# Convert the results to a pandas DataFrame
pd_counts = flights_counts.toPandas()
# Print the head of pd_counts
print(pd_counts.head())
```

```
  origin dest    N
0    SEA  RNO    8
1    SEA  DTW   98
2    SEA  CLE    2
3    SEA  LAX  450
4    PDX  SEA  144
```

```
flights_spark.join(airports_spark, flights_spark.origin == airports_spark.faa)\
.select("year", "month", "day", "origin", "dest", "name")\
.show(10)
```

```
+----+-----+---+------+----+------------------+
|year|month|day|origin|dest|              name|
+----+-----+---+------+----+------------------+
|2014|   12|  8|   SEA| LAX|Seattle Tacoma Intl|
|2014|    1| 22|   SEA| HNL|Seattle Tacoma Intl|
|2014|    3|  9|   SEA| SFO|Seattle Tacoma Intl|
|2014|    4|  9|   PDX| SJC|      Portland Intl|
|2014|    3|  9|   SEA| BUR|Seattle Tacoma Intl|
|2014|    1| 15|   PDX| DEN|      Portland Intl|
|2014|    7|  2|   PDX| OAK|      Portland Intl|
|2014|    5| 12|   SEA| SFO|Seattle Tacoma Intl|
|2014|    4| 19|   SEA| SAN|Seattle Tacoma Intl|
|2014|   11| 19|   SEA| ORD|Seattle Tacoma Intl|
+----+-----+---+------+----+------------------+
only showing top 10 rows
```

```
from pyspark.sql import functions as F
#Agrupar por Origen y destino , y contar
flights_counts = flights_spark.groupBy("origin", "dest").agg(F.count("*").alias("N"))
#Convertir a pandas
pd_counts = flights_counts.toPandas()
#Imprimir las 10 primeras filas
print(pd_counts.head())
```

```
  origin dest    N
0    SEA  RNO    8
1    SEA  DTW   98
2    SEA  CLE    2
3    SEA  LAX  450
4    PDX  SEA  144
```

Aeropuertos destino más frecuentes

```
flights_with_dest_airport_names = flights_spark.join(airports_spark, flights_spark.dest == airports_spark.faa)
most_frequent_dest_airports = flights_with_dest_airport_names.groupBy(airports_spark.name).agg(F.count("*").alias("count"))
most_frequent_dest_airports\
.orderBy("count", ascending=False)\
.show(10)
```

```
+--------------------+-----+
|                name|count|
+--------------------+-----+
|  San Francisco Intl|  787|
|    Los Angeles Intl|  615|
|         Denver Intl|  586|
|Phoenix Sky Harbo...|  530|
|      Mc Carran Intl|  520|
|Ted Stevens Ancho...|  449|
|  Chicago Ohare Intl|  439|
| Salt Lake City Intl|  396|
|Dallas Fort Worth...|  371|
|Norman Y Mineta S...|  369|
+--------------------+-----+
only showing top 10 rows
```

Mostrar vuelos junto con fabricante del avión

```
flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)\
.select("flight","origin", "dest", "manufacturer")\
.show(10)
```

```
+------+------+----+------------+
|flight|origin|dest|manufacturer|
+------+------+----+------------+
|  1780|   SEA| LAX|      AIRBUS|
|   851|   SEA| HNL|      BOEING|
|   755|   SEA| SFO|      AIRBUS|
|   344|   PDX| SJC|      BOEING|
|   522|   SEA| BUR|      BOEING|
|    48|   PDX| DEN|      BOEING|
```

```
|  1520|   PDX| OAK|      BOEING|
|   755|   SEA| SFO|      AIRBUS|
|   490|   SEA| SAN|      BOEING|
|    26|   SEA| ORD|      BOEING|
+------+------+----+------------+
only showing top 10 rows
```

Vuelos agrupados por fabricante del avión

```
from pyspark.sql import functions as F


flights_with_manufacturer = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)


flights_by_manufacturer = flights_with_manufacturer.groupBy("manufacturer").agg(F.count("*").alias("count"))

flights_by_manufacturer.orderBy("count", ascending=False).show()
```

```
+--------------------+-----+
|        manufacturer|count|
+--------------------+-----+
|              BOEING| 6660|
|      BOMBARDIER INC|  895|
|              AIRBUS|  811|
|    AIRBUS INDUSTRIE|  743|
|             EMBRAER|  273|
|   MCDONNELL DOUGLAS|   47|
|            CANADAIR|    8|
|              CESSNA|    4|
|ROBINSON HELICOPT...|    3|
|  CIRRUS DESIGN CORP|    2|
|       BARKER JACK L|    1|
|                BELL|    1|
+--------------------+-----+
```

Conteo de vuelos por aeropuerto (origen y destino)

```
all_airports_df = flights_spark.selectExpr("origin as airport").unionAll(flights_spark.selectExpr("dest as airport"))
airport_counts_df_api = all_airports_df.groupBy("airport").agg(F.count("*").alias("flight_count"))
airport_counts_df_api.orderBy("flight_count", ascending=False).show()
```

```
+-------+------------+
|airport|flight_count|
+-------+------------+
|    SEA|        6898|
|    PDX|        3403|
|    SFO|         787|
|    LAX|         615|
|    DEN|         586|
|    PHX|         530|
|    LAS|         520|
|    ANC|         449|
|    ORD|         439|
|    SLC|         396|
|    DFW|         371|
|    SJC|         369|
|    OAK|         334|
|    SMF|         283|
|    SAN|         271|
|    ATL|         258|
|    MSP|         238|
|    IAH|         226|
|    SNA|         198|
|    LGB|         175|
+-------+------------+
only showing top 20 rows
```

Promedio de distancia por aeropuerto de origen

```
avg_distance_by_origin = flights_spark.groupBy("origin").agg(F.avg("distance").alias("average_distance"))
avg_distance_by_origin.orderBy("average_distance", ascending=False).show(10)
```

```
+------+------------------+
|origin|  average_distance|
+------+------------------+
|   SEA|1276.5170269469943|
|   PDX|1065.9026494146642|
+------+------------------+
```

Vuelos de más de 2000 km con modelo del avión

```
long_distance_flights = flights_spark.filter(flights_spark.distance > 2000)
flights_with_model = long_distance_flights.join(planes_spark, long_distance_flights.tailnum == planes_spark.tailnum)
flights_with_model.select("flight","distance", "model").show()
```

```
+------+--------+---------+
|flight|distance|    model|
+------+--------+---------+
|   851|    2677|  737-890|
|    24|    2496|  737-890|
|   616|    2378| A320-214|
|    29|    2640| A330-243|
|   815|    2701|  737-890|
|    18|    2554|737-990ER|
|  1598|    2182|737-932ER|
|  1275|    2402|737-924ER|
|    12|    2496|  737-890|
|    25|    2603| A330-243|
|  1358|    2454|  737-832|
|    32|    2378|737-990ER|
|  2497|    2172|737-932ER|
|  1473|    2422|  757-2Q8|
|   875|    2701|  737-890|
|  1929|    2182|  767-332|
|   774|    2520|  737-890|
|   264|    2422| A320-232|
|  1857|    2279| A321-231|
|    38|    2717|737-990ER|
+------+--------+---------+
only showing top 20 rows
```

Número de vuelos por año de fabricación del avión

```
flights_with_plane_year = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)
flights_by_plane_year = flights_with_plane_year.groupBy(planes_spark.year).agg(F.count("*").alias("flight_count"))
flights_by_plane_year.orderBy(planes_spark.year).show()
```

```
+----+------------+
|year|flight_count|
+----+------------+
|   0|           8|
|1959|           1|
|1975|           1|
|1980|           3|
|1984|          10|
|1985|          35|
|1986|          13|
|1987|          33|
|1988|          52|
|1989|          53|
|1990|         105|
|1991|          74|
|1992|         269|
|1993|          97|
|1994|         157|
|1995|         126|
|1996|         194|
|1997|         222|
|1998|         622|
|1999|         650|
```

```
+----+------------+
only showing top 20 rows
```

Vuelos con origen y destino mostrando ciudad de ambos aeropuertos

```
sql_query = """
SELECT
  f.flight,
  f.origin,
  origin_airport.name AS origin_airport_name,
  f.dest,
  dest_airport.name AS dest_airport_name
FROM flights f
JOIN airports origin_airport ON f.origin = origin_airport.faa
JOIN airports dest_airport ON f.dest = dest_airport.faa
"""


spark.sql(sql_query).show()
```

```
+------+------+------------------+----+--------------------+
|flight|origin|origin_airport_name|dest|   dest_airport_name|
+------+------+------------------+----+--------------------+
|  1780|   SEA|Seattle Tacoma Intl| LAX|    Los Angeles Intl|
|   851|   SEA|Seattle Tacoma Intl| HNL|       Honolulu Intl|
|   755|   SEA|Seattle Tacoma Intl| SFO|  San Francisco Intl|
|   344|   PDX|      Portland Intl| SJC|Norman Y Mineta S...|
|   522|   SEA|Seattle Tacoma Intl| BUR|            Bob Hope|
|    48|   PDX|      Portland Intl| DEN|         Denver Intl|
|  1520|   PDX|      Portland Intl| OAK|Metropolitan Oakl...|
|   755|   SEA|Seattle Tacoma Intl| SFO|  San Francisco Intl|
|   490|   SEA|Seattle Tacoma Intl| SAN|      San Diego Intl|
|    26|   SEA|Seattle Tacoma Intl| ORD|  Chicago Ohare Intl|
|   448|   SEA|Seattle Tacoma Intl| LAX|    Los Angeles Intl|
|   656|   SEA|Seattle Tacoma Intl| PHX|Phoenix Sky Harbo...|
|   608|   SEA|Seattle Tacoma Intl| LAS|      Mc Carran Intl|
|   121|   SEA|Seattle Tacoma Intl| ANC|Ted Stevens Ancho...|
|   306|   SEA|Seattle Tacoma Intl| SFO|  San Francisco Intl|
|  1458|   PDX|      Portland Intl| SFO|  San Francisco Intl|
|   368|   SEA|Seattle Tacoma Intl| SMF|     Sacramento Intl|
|   827|   SEA|Seattle Tacoma Intl| MDW| Chicago Midway Intl|
|    24|   SEA|Seattle Tacoma Intl| BOS|General Edward La...|
|  3488|   PDX|      Portland Intl| BUR|            Bob Hope|
+------+------+------------------+----+--------------------+
only showing top 20 rows
```

Promedio de retraso de salida por aeropuerto.

```
from pyspark.sql import functions as F
avg_dep_delay_by_origin = flights_spark.groupBy("origin").agg(F.avg("dep_delay").alias("average_departure_delay"))
avg_dep_delay_by_origin.orderBy("average_departure_delay", ascending=False).show()
```

```
+------+-----------------------+
|origin|average_departure_delay|
+------+-----------------------+
|   SEA|      6.334918983202022|
|   PDX|      5.513178294573644|
+------+-----------------------+
```

Promedio de retraso de llegada por modelo de avión

```
flights_with_model = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)
avg_arr_delay_by_model = flights_with_model.groupBy(planes_spark.model).agg(F.avg("arr_delay").alias("average_arrival_delay"))
avg_arr_delay_by_model.orderBy("average_arrival_delay", ascending=False).show()
```

```
+-------------+--------------------+
|        model|average_arrival_delay|
+-------------+--------------------+
|      777-224|                72.0|
```

```
|      A319-114|                44.5|
|      A330-223|                36.0|
|       737-705|                33.0|
|       737-3K2|                26.0|
|     767-432ER|                20.5|
|DC-9-83(MD-83)|    18.75862068965517|
|       767-33A|                17.0|
|       737-301|   15.285714285714286|
|       757-231|   14.966666666666667|
|       737-7BD|   14.807017543859649|
|       737-3Y0|                14.0|
|       737-3Q8|                12.0|
|   CL-600-2B19|    9.373983739837398|
|      A319-112|    7.614457831325301|
|       737-3H4|    7.442446043165468|
|      A319-111|    7.392857142857143|
|       737-76N|    6.857142857142857|
|       767-332|    6.230769230769231|
|       737-490|    6.040339702760085|
+--------------+---------------------+
only showing top 20 rows
```

Contar vuelos por ciudad origen y destino

```
from pyspark.sql import functions as F
from pyspark.sql.functions import col # Import col here


flight_counts_by_cities = flights_spark.alias("f").join(
    airports_spark.alias("origin_airport"),
    col("f.origin") == col("origin_airport.faa")
).join(
    airports_spark.alias("dest_airport"),
    col("f.dest") == col("dest_airport.faa")
).groupBy(
    col("origin_airport.name").alias("origin_city"),
    col("dest_airport.name").alias("dest_city")
).agg(F.count("*").alias("flight_count"))


flight_counts_by_cities.orderBy("flight_count", ascending=False).show()
```

```
+------------------+--------------------+------------+
|       origin_city|           dest_city|flight_count|
+------------------+--------------------+------------+
|Seattle Tacoma Intl|  San Francisco Intl|         482|
|Seattle Tacoma Intl|    Los Angeles Intl|         450|
|Seattle Tacoma Intl|Ted Stevens Ancho...|         380|
|Seattle Tacoma Intl|      Mc Carran Intl|         364|
|Seattle Tacoma Intl|         Denver Intl|         351|
|Seattle Tacoma Intl|Phoenix Sky Harbo...|         321|
|      Portland Intl|  San Francisco Intl|         305|
|Seattle Tacoma Intl|  Chicago Ohare Intl|         282|
|Seattle Tacoma Intl|Dallas Fort Worth...|         245|
|      Portland Intl|         Denver Intl|         235|
|Seattle Tacoma Intl| Salt Lake City Intl|         225|
|Seattle Tacoma Intl|Norman Y Mineta S...|         213|
|Seattle Tacoma Intl|Metropolitan Oakl...|         213|
|      Portland Intl|Phoenix Sky Harbo...|         209|
|Seattle Tacoma Intl|     Sacramento Intl|         190|
|Seattle Tacoma Intl|      San Diego Intl|         180|
|      Portland Intl| Salt Lake City Intl|         171|
|Seattle Tacoma Intl|George Bush Inter...|         169|
|Seattle Tacoma Intl|Minneapolis St Pa...|         166|
|      Portland Intl|    Los Angeles Intl|         165|
+------------------+--------------------+------------+
only showing top 20 rows
```

Vuelos con aviones Boeing

```
boeing_flights = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)\
.filter(planes_spark.manufacturer == "BOEING")\
```

```
.select(flights_spark.tailnum, planes_spark.manufacturer, flights_spark.flight)


boeing_flights.show()
```

```
+-------+------------+------+
|tailnum|manufacturer|flight|
+-------+------------+------+
| N559AS|      BOEING|   851|
| N360SW|      BOEING|   344|
| N612AS|      BOEING|   522|
| N646SW|      BOEING|    48|
| N422WN|      BOEING|  1520|
| N309AS|      BOEING|   490|
| N564AS|      BOEING|    26|
| N323AS|      BOEING|   448|
| N305AS|      BOEING|   656|
| N433AS|      BOEING|   608|
| N765AS|      BOEING|   121|
| N713AS|      BOEING|   306|
| N27205|      BOEING|  1458|
| N626AS|      BOEING|   368|
| N8634A|      BOEING|   827|
| N597AS|      BOEING|    24|
| N519AS|      BOEING|   300|
| N793SA|      BOEING|  1617|
| N520AS|      BOEING|   306|
| N549AS|      BOEING|   604|
+-------+------------+------+
only showing top 20 rows
```

Contar vuelos por fabricante

```
flights_with_manufacturer = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)
flights_by_manufacturer = flights_with_manufacturer.groupBy("manufacturer").agg(F.count("*").alias("count"))
flights_by_manufacturer.orderBy("count", ascending=False).show()
```

```
+--------------------+-----+
|        manufacturer|count|
+--------------------+-----+
|              BOEING| 6660|
|      BOMBARDIER INC|  895|
|              AIRBUS|  811|
|    AIRBUS INDUSTRIE|  743|
|             EMBRAER|  273|
|   MCDONNELL DOUGLAS|   47|
|            CANADAIR|    8|
|              CESSNA|    4|
|ROBINSON HELICOPT...|    3|
|  CIRRUS DESIGN CORP|    2|
|       BARKER JACK L|    1|
|                BELL|    1|
+--------------------+-----+
```

Vuelos con más de 2 horas de retraso en salida y llegada

```
delayed_flights = flights_spark.filter((flights_spark.dep_delay > 120) & (flights_spark.arr_delay > 120))\
.select(flights_spark.dep_delay, flights_spark.arr_delay, flights_spark.flight)

delayed_flights.show()
```

```
+---------+---------+------+
|dep_delay|arr_delay|flight|
+---------+---------+------+
|      155|      170|  1598|
|      223|      212|  1596|
|      144|      142|  4612|
|      164|      149|   794|
|      133|      145|  5438|
|      202|      198|   538|
|      132|      135|    18|
|      274|      279|  5325|
|      328|      318|   157|
|      274|      268|   511|
```

```
|      179|      187|  4634|
|      150|      133|   508|
|      266|      284|  1397|
|      211|      200|  1556|
|      302|      314|  1121|
|      273|      261|   419|
|      135|      125|  3458|
|      370|      371|  1702|
|      202|      191|   970|
|      123|      126|  1054|
+---------+---------+------+
only showing top 20 rows
```

Promedio de distancia recorrida por fabricante de avión

```
flights_with_manufacturer = flights_spark.join(planes_spark, flights_spark.tailnum == planes_spark.tailnum)
avg_distance_by_manufacturer = flights_with_manufacturer.groupBy(planes_spark.manufacturer).agg(F.avg("distance").alias("average_
avg_distance_by_manufacturer.orderBy("average_distance", ascending=False).show()
```

```
+--------------------+------------------+
|        manufacturer|  average_distance|
+--------------------+------------------+
|  CIRRUS DESIGN CORP|            1693.5|
|              CESSNA|            1616.0|
|                BELL|            1616.0|
|   MCDONNELL DOUGLAS| 1554.787234042553|
|              AIRBUS|1375.0628853267572|
|    AIRBUS INDUSTRIE|1318.1736204576043|
|              BOEING| 1245.056006006006|
|       BARKER JACK L|             965.0|
|ROBINSON HELICOPT...| 925.3333333333334|
|      BOMBARDIER INC| 749.3497206703911|
|             CANADAIR|            731.25|
|              EMBRAER|127.29304029304029|
+--------------------+------------------+
```