Airlines Delay

NINJAZ

Agenda

1. Dataset Description
2. Data Cleaning
3. Dashboard
4. PIG Latin Questions

Airlines Delay Dataset

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers.

This version of the dataset was compiled from the Statistical Computing Statistical Graphics 2009 Data Expo

contains 30 columns and 1,936,758 entries

#	Column	Dtype
0	Unnamed: 0	int64
1	Year	int64
2	Month	int64
3	DayofMonth	int64
4	DayOfWeek	int64
5	DepTime	float64
6	CRSDepTime	int64
7	ArrTime	float64
8	CRSArrTime	int64
9	UniqueCarrier	object
10	FlightNum	int64
11	TailNum	object
12	ActualElapsedTime	float64
13	CRSElapsedTime	float64
14	AirTime	float64
15	ArrDelay	float64
16	DepDelay	float64
17	Origin	object
18	Dest	object
19	Distance	int64
20	TaxiIn	float64
21	TaxiOut	float64
22	Cancelled	int64
23	CancellationCode	object
24	Diverted	int64
25	CarrierDelay	float64
26	WeatherDelay	float64
27	NASDelay	float64
28	SecurityDelay	float64
29	LateAircraftDelay	float64

Features Description

no.	feature	description
1	Year	2008
2	Month	1-Dec
3	DayofMonth	Jan-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	1. unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number: aircraft registration, unique aircraft identifier
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes: A flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS).
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	*was the flight cancelled
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.
26	WeatherDelay	in minutes: Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.
27	NASDelay	in minutes: Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
28	SecurityDelay	in minutes: Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.
29	LateAircraftDelay	in minutes: Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as

Data Cleaning

Dropping useless columns:

```
list = ['DayofMonth' , 'DayOfWeek' ,'Diverted' ,'CancellationCode' , 'Cancelled' ,'WeatherDelay', 'NASDelay' ,'SecurityDelay' ,'LateAircraftDelay']

# Drop unuseful column
for column in list:
    df.drop(columns= column,axis=1,inplace=True)
```

Dropping nulls:

```
#drop null values
df.dropna(axis=0,inplace=True)
```

Dashboard

1st Question

How many minutes the arrival delay for each flight with distance greater than 500 miles?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',') AS
( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines GENERATE FlightNum,ArrDelay,Distance;
distance= FILTER airlines_info BY Distance > 500;
DUMP distance;
```

Result

```
(3920,34.0,515)
(509,57.0,1591)
(1333,80.0,828)
(675,15.0,1489)
(4,16.0,838)
(188,47.0,1093)
(362,64.0,972)
(422,72.0,765)
(1056,29.0,1052)
(3244,21.0,888)
(1035,49.0,587)
(205,38.0,666)
(54,31.0,816)
(23,36.0,742)
(302,78.0,2237)
(1079,38.0,758)
(2021,54.0,1090)
(1018,49.0,2298)
(419,31.0,1588)
(2032,20.0,1588)
(3940,60.0,1588)
(538,24.0,520)
(2383,52.0,1987)
(1774,18.0,2106)
(2632,32.0,2106)
```

2nd Question

What is the average minute delay for each carrier?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',') AS
( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines GENERATE UniqueCarrier,CarrierDelay;
carrier_delay = GROUP airlines BY (UniqueCarrier);
Avg_arr = FOREACH carrier_delay GENERATE AVG(airlines.CarrierDelay);
DUMP Avg_arr;
```

Result

```
(28.973984771573605)
(18.841187558906693)
(21.926728971962618)
(37.47969802081208)
(17.052489905787347)
(10.093)
(35.73943661971831)
(16.380253310305125)
(24.752113193678795)
(27.363258026159333)
(22.197816150330116)
(18.913820090231102)
(16.561383928571427)
(10.445792190867627)
(15.377529057253552)
(39.73010380622837)
```

3rd Question

What is the difference between the arrival and departure time of the flights and the actual elapsed time for the flight?

Pig Script

```
airlines = LOAD '/user/maria_dev/practice/Airlines.csv' USING PigStorage(',') AS
( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines GENERATE FlightNum,ArrTime-DepTime,ActualElapsedTime;
DUMP airlines_info;
```

Result

```
(3920,130.0,90.0)
(509,100.0,240.0)
(1333,201.0,121.0)
(675,188.0,228.0)
(4,203.0,123.0)
(54,96.0,56.0)
(623,97.0,57.0)
(188,235.0,155.0)
(362,227.0,147.0)
(422,115.0,135.0)
(1056,233.0,153.0)
(3244,356.0,136.0)
(186,109.0,69.0)
(3154,110.0,70.0)
(1035,276.0,96.0)
(205,223.0,143.0)
(3430,11.0,71.0)
(1580,79.0,99.0)
(2195,26.0,86.0)
(54,123.0,143.0)
(23,207.0,127.0)
(1574,94.0,54.0)
```

4th Question

How many flights are departing from Indiana airport to Baltimore airport?

Pig Script

```
join_airlines= JOIN airlines1 by (FlightNum,Origin), airlines2 by (FlightNum,Dest);
origin_filter= FILTER join_airlines by airlines1::Origin matches 'IND';
dest_filter= FILTER join_airlines by airlines2::Dest matches 'BWI';
cogroup_data = COGROUP origin_filter by airlines1::Origin, dest_filter by airlines2::Dest;
DUMP cogroup_data;
```

Result

```
(BWI,{},{(2008,1,1650.0,1625,1815.0,1745,WN,3326,N686SW,85.0,80.0,70.0,30.0,25.0,BWI,MHT,377,6,9,0.0,2008,
(IND,{(2008,1,1323.0,1255,1526.0,1510,WN,4,N674AA,123.0,135.0,110.0,16.0,28.0,IND,TPA,838,4,9,0.0,2008,1,1
```

5th Question

Flights coming from Indiana Airport?

Pig Script

```
airlines = LOAD '/user/maria_dev/practice/Airlines.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines GENERATE FlightNum,Origin,Dest;
origin_filter = FILTER airlines_info by Origin matches 'IND';
DUMP origin_filter;
```

Result

```
(3920,IND,BWI)
(509,IND,LAS)
(1333,IND,MCO)
(675,IND,PHX)
(4,IND,TPA)
(3920,IND,BWI)
(509,IND,LAS)
(829,IND,MDW)
(1144,IND,PHX)
(1746,IND,BWI)
(378,IND,JAX)
(1827,IND,MDW)
(1746,IND,BWI)
(675,IND,PHX)
(1333,IND,MCO)
(675,IND,PHX)
(810,IND,MCI)
(815,IND,BWI)
(2885,IND,JAX)
(619,IND,MCI)
(1405,IND,MDW)
(1623,IND,MDW)
(540,IND,BWI)
```

6th Question

What is the maximum delay ?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier_delay = GROUP airlines BY (UniqueCarrier);
max_delay = FOREACH carrier_delay GENERATE MAX(airlines.CarrierDelay);
limited_result = limit max_delay 1;
DUMP limited_result;
```

Result

```
(1094.0)
```

7th Question

What is the minimum delay ?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier_delay = GROUP airlines BY (UniqueCarrier);
min_delay = FOREACH carrier_delay GENERATE MIN(airlines.CarrierDelay);
limited_result = limit min_delay 1;
DUMP limited_result;
```

Result

```
(0.0)
```

8th Question

Count the number of flights of each unique carrier.

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

carrier = GROUP airlines BY (UniqueCarrier);

count_carrier = FOREACH carrier GENERATE FLATTEN(group) as (UniqueCarrier), COUNT(airlines.UniqueCarrier);
DUMP count_carrier;
```

Result

```
(9E,4728)
(AA,4244)
(DL,2675)
(EV,4901)
(F9,1486)
(FL,3000)
(HA,142)
(MQ,10422)
(NW,5442)
(OH,3364)
(OO,11814)
(UA,10861)
(US,5376)
(WN,19951)
(XE,6969)
(YV,4624)
```

9th Question

What is the Actual elapsed time for flights in ascending order?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines Generate FlightNum,ActualElapsedTime;
ActualElapsedTime_ordered = ORDER airlines_info BY ActualElapsedTime ASC;
STORE ActualElapsedTime_ordered INTO 'ActualElapsedTime_ASC_ordered';
```

Result

```
2475 25.0
1012 25.0
6659 25.0
5775 26.0
5747 26.0
555  26.0
5775 27.0
5775 27.0
5775 27.0
240  27.0
5781 27.0
5707 27.0
5340 27.0
6659 27.0
5740 28.0
5718 28.0
5740 28.0
```

10th Question

What is the CRSelapsed time for flights in ascending order?

Pig Script

```
airlines = LOAD '/user/maria_dev/Airlines/Airlines_delay_processed.csv' USING PigStorage(',')
AS ( Year:int, Month:int,DepTime:float,CRSDepTime:int,ArrTime:float,
CRSArrTime:int,UniqueCarrier:chararray, FlightNum:int,TailNum:chararray,ActualElapsedTime:float,
CRSElapsedTime:float,AirTime:float,ArrDelay:float,DepDelay:float,Origin:chararray,Dest:chararray,
Distance:int,TaxiIn:int,TaxiOut:int,CarrierDelay:float);

airlines_info = FOREACH airlines Generate FlightNum,CRSElapsedTime;
CRSElapsedTime_ordered = ORDER airlines_info BY CRSElapsedTime ASC;
STORE CRSElapsedTime_ordered INTO 'CRSElapsedTime_ASC_ordered';
```

Result

```
405  19.0
6790 20.0
5707 26.0
5707 26.0
5707 26.0
179  29.0
192  29.0
172  29.0
144  29.0
240  29.0
190  29.0
6797 30.0
5781 31.0
180  31.0
5778 31.0
5427 31.0
5775 31.0
5427 31.0
5778 31.0
```

THANK YOU!