Python project for Data Science

Prepared By :- Karan Singal

Under the guidance of :- Tareq Jaber PhD.

INTRODUCTION

Data Source : The U.S. DOT's Bureau of Transportation Statistics

This data is from US DOT showing all the Delayed, Diverted, Cancelled flights operated by Major Air carriers in USA in 2015

Ref:- https://www.kaggle.com/usdot/flight-delays

Prediction of flight delay in USA

Steps to follow for Data science project

- Identifying the Dataset
- Data Cleaning
- Exploratory data analysis (EDA)
- Visualization
- Conclusion

Preparing the Data

First step is to Import all the necessary Libraries

```
# Import Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Load the Dataset in Python Console i.e. CSV files

```
flights= pd.read_csv('flights.csv', low_memory=False)
airlines = pd.read_csv('airlines.csv')
airports = pd.read_csv('airports.csv')
```

DATAFRAME

Airlines

airlines - DataFrame

Index	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways
5	OO	Skywest Airlines Inc.
6	AS	Alaska Airlines Inc.
7	NK	Spirit Air Lines
8	WN	Southwest Airlines Co.
9	DL	Delta Air Lines Inc.
10	EV	Atlantic Southeast Air…
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

Airport

airports - DataFrame

Index	IATA_CODE	AIRPORT	CITY
0	ABE	Lehigh Valley International…	Allentown
1	ABI	Abilene Regional Airp…	Abilene
2	ABQ	Albuquerque International…	Albuquerque
3	ABR	Aberdeen Regional Airp…	Aberdeen
4	ABY	Southwest Georgia Regio…	Albany
5	ACK	Nantucket Memorial Airp…	Nantucket
6	ACT	Waco Regional Airport	Waco
7	ACV	Arcata Airport	Arcata/Eurek
8	ACY	Atlantic City International…	Atlantic Cit
9	ADK	Adak Airport	Adak
10	ADQ	Kodiak Airport	Kodiak
11	AEX	Alexandria International…	Alexandria
12	AGS	Augusta Regional Airp…	Augusta

Flights

flights - DataFrame

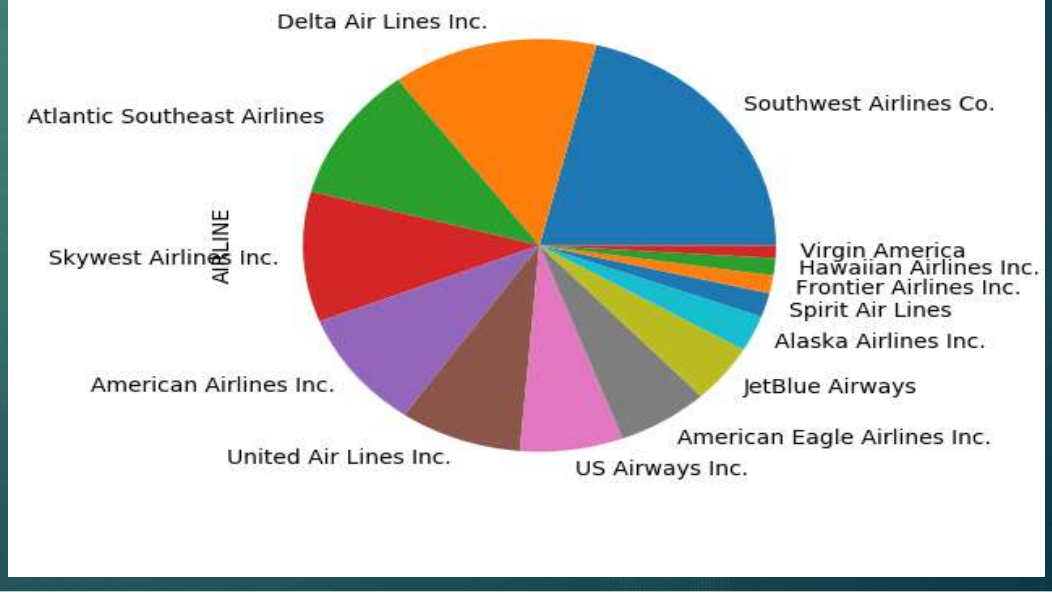
Index	YEAR	MONTH	DAY	'_OF_W	AIRLINE_x	FLIGHT_NUMBER	TAIL_NUMBER
0	2015	January	1	4	AS	98	N407AS
1	2015	January	1	4	AS	135	N527AS
2	2015	January	1	4	AS	108	N309AS
3	2015	January	1	4	AS	122	N413AS
4	2015	January	1	4	AS	130	N457AS
5	2015	January	1	4	AS	136	N431AS
6	2015	January	1	4	AS	134	N464AS
7	2015	January	1	4	AS	144	N514AS
8	2015	January	1	4	AS	114	N303AS
9	2015	January	1	4	AS	695	N607AS
10	2015	January	1	4	AS	730	N423AS
11	2015	January	1	4	AS	81	N577AS
12	2015	January	1	4	AS	162	N792AS

Q1:- Which is the largest airline in terms of number of flights

```
flights_max = flights['AIRLINE'].value_counts()
print('The maximum flights are from ' + (flights_max.nlargest(1).to_string()))
flights_max.plot.pie()
```

The maximum flights are from Southwest Airlines

```
In [41]: flights_max
Out[41]:
Southwest Airlines Co.          1261855
Delta Air Lines Inc.             875881
American Airlines Inc.           725984
Skywest Airlines Inc.            588353
Atlantic Southeast Airlines      571977
United Air Lines Inc.            515723
American Eagle Airlines Inc.     294632
JetBlue Airways                  267048
US Airways Inc.                  198715
Alaska Airlines Inc.             172521
Spirit Air Lines                 117379
Frontier Airlines Inc.            90836
Hawaiian Airlines Inc.            76272
Virgin America                    61903
Name: AIRLINE, dtype: int64
```



Q2:- Which month of the year has the most number of cancelled flights?

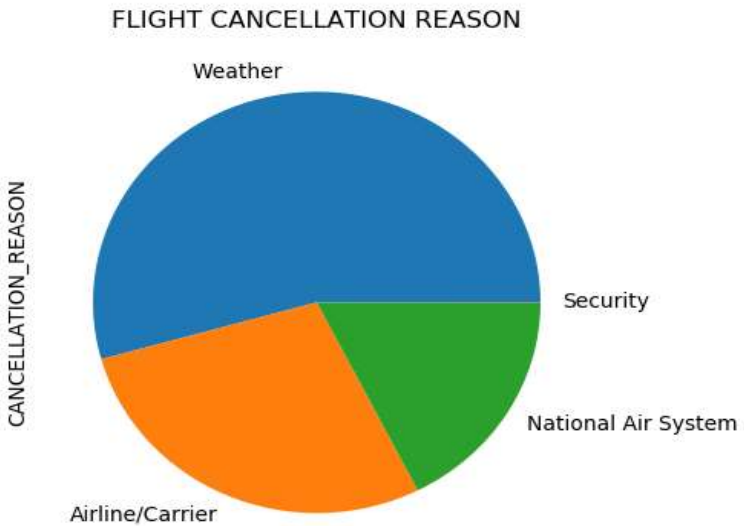
```
flights_cancelled_month = flights.groupby('AIRLINE', as_index= False)['ARRIVAL_DELAY']
                                 .agg('sum').rename(columns={"ARRIVAL_DELAY":"ARRIVAL_TOTAL"})
flights_cancelled_month = flights.groupby(by='MONTH')['CANCELLED'].agg('count')
flights_cancelled_month.plot.bar(title='CANCELLED FLIGHTS BY MONTH')
```

The maximum flights cancelled by Month

Q3:- What is the most common reason for flight cancellation ?

```
flights_cancel_reason = flights['CANCELLATION_REASON'].value_counts()
print('"The reason for cancellation on most flights is',flights_cancel_reason.nlargest(1)
                                                    .to_string(),'in total numbers'' "')
flights_cancel_reason.plot(kind='pie',title='FLIGHT CANCELLATION REASON')
```

The most common reason for flight cancellations is due to weather



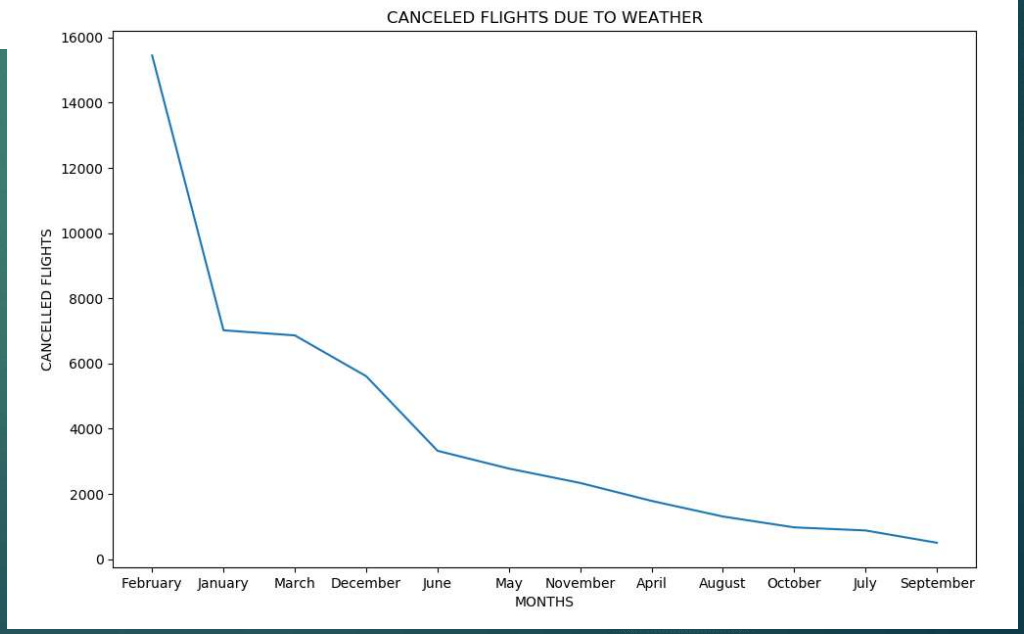
Q4:- How many flights are cancelled due to weather in each month ?

```
cancelled_due_to_weather = flights[flights['CANCELLATION_REASON'] == 'Weather'].groupby('MONTH',
   as_index= False)['CANCELLATION_REASON'].agg('count').sort_values('CANCELLATION_REASON',ascending=False)
cancelled_due_to_weather
```

In February month most flights are cancelled due to weather

Out[74]:

	MONTH	CANCELLATION_REASON
3	February	15447
4	January	7020
7	March	6864
2	December	5613
6	June	3325
8	May	2780
9	November	2339
0	April	1789
1	August	1310
10	October	977
5	July	882
11	September	505



Q5:- Which airline company has the most delayed flight?

```
flights_most_delay = flights.groupby('AIRLINE', as_index= False)['ARRIVAL_DELAY']
                .agg('count').rename(columns={"ARRIVAL_DELAY":"ARRIVAL_DELAY_CNT"})
flights_most_delay_sort = flights_most_delay.sort_values('ARRIVAL_DELAY_CNT',ascending=False)
print('"The carrier/airline with the most number of delayed flights is',flights_most_delay_sort
                                    ['AIRLINE'].head(1).to_string(index=False).upper(),'"')
```

The most delayed flights is from SOUTHWEST AIRLINES CO.

```
In [118]: print('"The carrier/airline with the most number of delayed flights
is',flights_grouped_cnt_sort['AIRLINE'].head(1).to_string(index=False).upper()
,'"')
"The carrier/airline with the most number of delayed flights is  SOUTHWEST
AIRLINES CO. "
```

Q6:- Which airline company has the least delayed flight?

```
flights_grouped_sum = flights.groupby('AIRLINE', as_index= False)['ARRIVAL_DELAY']
                  .agg('sum').rename(columns={"ARRIVAL_DELAY":"ARRIVAL_DELAY_SUM"})

flights_delayed_avg = flights_grouped_cnt.merge(flights_grouped_sum, left_on=
                                'AIRLINE', right_on='AIRLINE', how='inner')
flights_delayed_avg.loc[:,'AVG_DELAY_AIRLINE'] = flights_delayed_avg
                  ['ARRIVAL_DELAY_SUM'] / flights_delayed_avg['ARRIVAL_DELAY_CNT']

flights_delayed_avg_sort = flights_delayed_avg.sort_values('AVG_DELAY_AIRLINE', ascending=True)
print('" The carrier/airline with the least average time of delayed flights is',
      flights_delayed_avg_sort['AIRLINE'].head(1).to_string(index=False).upper(),' "')
```

The least delayed flights is from SOUTHWEST AIRLINES CO.

```
In [42]: print('" The carrier/airline with the least average time of delayed flights
is',flights_delayed_avg_sort['AIRLINE'].head(1).to_string(index=False).upper(),' "')
" The carrier/airline with the least average time of delayed flights is  ALASKA AIRLINES INC.  "
```

Conclusion

- Which is the biggest airline- Southwest Airlines Co.
- You can predict the best airline in terms of least delay ALASKA AIRLINES INC.
- How weather could affects your schedule
- You can avoid flying in the airlines having most delayed flights

Ref:- https://www.kaggle.com/usdot/flight-delays

Thank You