```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.metrics import mean_absolute_percentage_error
import statsmodels.api as sm
import scipy.stats as stats
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import StratifiedKFold,
RandomizedSearchCV, train_test_split
from statsmodels.formula.api import glm
import statsmodels.api as sm
from sklearn.preprocessing import OrdinalEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report, accuracy_score
from sklearn.tree import DecisionTreeClassifier
```

Data import and aggregation steps:

```
I.   Collate the information specific to flights, airports (like type
of airport, elevation etc) and runway(length_ft, width_ft, surface
etc.). Get all those fields in single dataset which you believe may
impact the delay.

airlines = pd.read_excel('Airlines.xlsx')
airports = pd.read_excel('airports.xlsx')
runways = pd.read_excel('runways.xlsx')

airlines.head()

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay
0   1      CO     269         SFO       IAH          3    15     205
1
1   2      US    1558         PHX       CLT          3    15     222
1
2   3      AA    2400         LAX       DFW          3    20     165
1
3   4      AA    2466         SFO       DFW          3    20     195
1
4   5      AS     108         ANC       SEA          3    30     202
0

airports.head(2)

       id ident           type                   name  latitude_deg  \
0    6523   00A       heliport     Total Rf Heliport     40.070801
```

```
1  323361  00AA  small_airport  Aero B Ranch Airport     38.704022

   longitude_deg  elevation_ft continent iso_country iso_region
municipality  \
0     -74.933601          11.0       NaN          US      US-PA
Bensalem
1    -101.473911        3435.0       NaN          US      US-KS
Leoti

  scheduled_service gps_code iata_code local_code home_link
wikipedia_link  \
0                no      00A       NaN         00A       NaN
NaN
1                no     00AA       NaN        00AA       NaN
NaN

  keywords
0      NaN
1      NaN

runways.head()

       id  airport_ref airport_ident  length_ft  width_ft surface
lighted  \
0  269408         6523           00A       80.0      80.0  ASPH-G
1
1  255155         6524          00AK     2500.0      70.0    GRVL
0
2  254165         6525          00AL     2300.0     200.0    TURF
0
3  270932         6526          00AR       40.0      40.0   GRASS
0
4  322128       322127          00AS     1450.0      60.0    Turf
0

   closed le_ident  le_latitude_deg  le_longitude_deg  le_elevation_ft
\
0       0       H1              NaN               NaN              NaN

1       0        N              NaN               NaN              NaN

2       0        1              NaN               NaN              NaN

3       0       H1              NaN               NaN              NaN

4       0        1              NaN               NaN              NaN


   le_heading_degT  le_displaced_threshold_ft he_ident
he_latitude_deg  \
0              NaN                        NaN      NaN
```

```
NaN
1               NaN                           NaN         S
NaN
2               NaN                           NaN        19
NaN
3               NaN                           NaN        H1
NaN
4               NaN                           NaN        19
NaN

   he_longitude_deg  he_elevation_ft  he_heading_degT  \
0               NaN              NaN              NaN
1               NaN              NaN              NaN
2               NaN              NaN              NaN
3               NaN              NaN              NaN
4               NaN              NaN              NaN

   he_displaced_threshold_ft
0                        NaN
1                        NaN
2                        NaN
3                        NaN
4                        NaN

airports.head(2)

       id ident           type                   name  latitude_deg  \
0    6523   00A       heliport     Total Rf Heliport     40.070801
1  323361  00AA  small_airport  Aero B Ranch Airport     38.704022

   longitude_deg  elevation_ft continent iso_country iso_region
municipality  \
0     -74.933601          11.0       NaN          US      US-PA
Bensalem
1    -101.473911        3435.0       NaN          US      US-KS
Leoti

  scheduled_service gps_code iata_code local_code home_link
wikipedia_link  \
0                no      00A       NaN        00A       NaN
NaN
1                no     00AA       NaN       00AA       NaN
NaN

  keywords
0      NaN
1      NaN
```

```
airport_run = pd.merge(airports, runways, left_on = 'ident', right_on
= 'airport_ident', how = "left")
airport_run.head(2)
```

```
     id_x ident           type                   name  latitude_deg  \
0    6523   00A       heliport     Total Rf Heliport     40.070801
1  323361  00AA  small_airport  Aero B Ranch Airport     38.704022

   longitude_deg  elevation_ft continent iso_country
iso_region  ...  \
0     -74.933601          11.0       NaN          US      US-PA  ...

1    -101.473911        3435.0       NaN          US      US-KS  ...


  le_longitude_deg le_elevation_ft le_heading_degT
le_displaced_threshold_ft  \
0              NaN             NaN             NaN
NaN
1              NaN             NaN             NaN
NaN

  he_ident he_latitude_deg he_longitude_deg he_elevation_ft
he_heading_degT  \
0      NaN             NaN               NaN             NaN
NaN
1      NaN             NaN               NaN             NaN
NaN

   he_displaced_threshold_ft
0                        NaN
1                        NaN

[2 rows x 38 columns]
```

```
count_runway = airport_run.groupby('airport_ident')
[['id_y']].count().sort_values(by = 'id_y', ascending =
False).reset_index()
count_runway.head(2)
```

```
  airport_ident  id_y
0          KORD    11
1          KNHU    10
```

```
air_run = pd.merge(airports, count_runway, how = 'left', left_on =
'ident', right_on = 'airport_ident')[['iata_code', 'type',
'elevation_ft','id_y']]
air_run.rename(columns = {'id_y': 'runway_count'}, inplace = True)
air_run.head(2)
```

```
  iata_code           type  elevation_ft  runway_count
0       NaN       heliport          11.0           1.0
1       NaN  small_airport        3435.0           NaN

air_run.dropna().to_csv('run_2.csv', index = False)

airlines.head(2)

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay
0   1      CO     269         SFO       IAH          3    15     205
1
1   2      US    1558         PHX       CLT          3    15     222
1

combined_data = pd.merge(airlines, air_run, how = 'left', left_on =
'AirportFrom', right_on = 'iata_code')

new_names = list(combined_data[air_run.columns].columns +
'_source_airport')
old_names = list(combined_data[air_run.columns].columns)

combined_data.rename(columns = {old:new for old,new in zip(old_names,
new_names)}, inplace = True)
combined_data.head(2)

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   2      US    1558         PHX       CLT          3    15     222
1

  iata_code_source_airport type_source_airport
elevation_ft_source_airport  \
0                      SFO       large_airport
13.0
1                      PHX       large_airport
1135.0

   runway_count_source_airport
0                          4.0
1                          3.0

combined_data.columns

Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo',
'DayOfWeek',
       'Time', 'Length', 'Delay', 'iata_code_source_airport',
       'type_source_airport', 'elevation_ft_source_airport',
```

```
       'runway_count_source_airport'],
      dtype='object')

combined_data = pd.merge(combined_data, air_run, how = 'left', left_on
= 'AirportTo', right_on = 'iata_code')

new_names = list(combined_data[air_run.columns].columns +
'_dest_airport')
old_names = list(combined_data[air_run.columns].columns)
combined_data.rename(columns = {old:new for old,new in zip(old_names,
new_names)}, inplace = True)
combined_data.head(2)

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   2      US    1558         PHX       CLT          3    15     222
1

  iata_code_source_airport type_source_airport
elevation_ft_source_airport  \
0                      SFO       large_airport
13.0
1                      PHX       large_airport
1135.0

   runway_count_source_airport iata_code_dest_airport
type_dest_airport  \
0                          4.0                    IAH
large_airport
1                          3.0                    CLT
large_airport

   elevation_ft_dest_airport  runway_count_dest_airport
0                       97.0                       5.0
1                      748.0                       4.0

# drop iata_code columns
combined_data.drop(columns =
list(combined_data.columns[combined_data.columns.str.startswith('iata_
code')]), inplace = True)

combined_data.head()

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   2      US    1558         PHX       CLT          3    15     222
1
```

```
2   3      AA    2400          LAX       DFW          3    20      165
1
3   4      AA    2466          SFO       DFW          3    20      195
1
4   5      AS     108          ANC       SEA          3    30      202
0

  type_source_airport  elevation_ft_source_airport  \
0       large_airport                         13.0
1       large_airport                       1135.0
2       large_airport                        125.0
3       large_airport                         13.0
4       large_airport                        152.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
97.0
1                          3.0     large_airport
748.0
2                          4.0     large_airport
607.0
3                          4.0     large_airport
607.0
4                          3.0     large_airport
433.0

   runway_count_dest_airport
0                        5.0
1                        4.0
2                        7.0
3                        7.0
4                        4.0

test =
pd.read_html("https://en.wikipedia.org/wiki/List_of_airlines_of_the_Un
ited_States")

len(test)

21

test[0]

                 Airline  Image IATA ICAO         Callsign  \
0        Alaska Airlines    NaN   AS  ASA           ALASKA
1          Allegiant Air    NaN   G4  AAY        ALLEGIANT
2      American Airlines    NaN   AA  AAL         AMERICAN
3         Avelo Airlines    NaN   XP  VXP            AVELO
4         Breeze Airways    NaN   MX  MXY             MOXY
5        Delta Air Lines    NaN   DL  DAL            DELTA
```

6	Eastern Airlines	NaN	2D	EAL	EASTERN	
7	Frontier Airlines	NaN	F9	FFT	FRONTIER FLIGHT	
8	Hawaiian Airlines	NaN	HA	HAL	HAWAIIAN	
9	JetBlue	NaN	B6	JBU	JETBLUE	
10	Ravn Alaska	NaN	7H	RVF	RAVN FLIGHT	
11	Southwest Airlines	NaN	WN	SWA	SOUTHWEST	
12	Spirit Airlines	NaN	NK	NKS	SPIRIT WINGS	
13	Sun Country Airlines	NaN	SY	SCX	SUN COUNTRY	
14	United Airlines	NaN	UA	UAL	UNITED	

	Primary hubs, secondary hubs	Founded	\
0	Seattle/Tacoma Anchorage Portland (OR) San Fra...	1932	
1	Las Vegas Cincinnati Destin/Ft. Walton Beach I...	1997	
2	Dallas/Fort Worth Charlotte Chicago–O'Hare Mia...	1926	
3	Burbank New Haven Orlando Raleigh/Durham Wilmi...	1987	
4	Charleston (SC) Hartford New Orleans Norfolk P...	2018	
5	Atlanta Detroit Minneapolis/St. Paul New York–...	1924	
6	Miami	2010	
7	Denver Atlanta Chicago–O'Hare Cincinnati Cleve...	1994	
8	Honolulu Kahului	1929	
9	New York–JFK Boston Los Angeles Fort Lauderdal...	1998	
10	Ontario	2021	
11	Dallas–Love Atlanta Baltimore Chicago–Midway D...	1967	
12	Fort Lauderdale Atlantic City Atlanta Detroit ...	1980	
13	Minneapolis/St. Paul Dallas/Fort Worth Las Vegas	1982	
14	Chicago–O'Hare Denver Houston–Intercontinental...	1926	

	Notes
0	Founded as McGee Airways and commenced operati...
1	Founded as WestJet Express and began operation...
2	Founded as American Airways and commenced oper...
3	First did business as Casino Express Airlines ...
4	Founded as Moxy Airways but was renamed due to...
5	Founded as Huff Daland Dusters and commenced o...
6	NaN
7	NaN
8	Founded as Inter-Island Airways in early 1929 ...
9	Founded as New Air and commenced operations in...
10	Founded as Northern Pacific Airways.
11	Founded as Air Southwest and commenced operati...
12	Founded as Charter One.
13	Commenced operations in 1983. Operates some Am...
14	Founded as Varney Air Lines and commenced oper...

II. Different airline companies may perform differently in terms of on time arrival. The performance may depend on the experience of the airline company. Pull the information specific to different airlines from the Wikipedia page https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States. Use web scaping to fetch the information about how long the airlines has been in the business.

```
website_url =
requests.get('https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States').text
soup = BeautifulSoup(website_url,'lxml')
My_table = soup.findAll("table",{"class":"wikitable"})

len(My_table)
```

```
7
```

```
airlines_wiki_list = []
for tab in My_table:
    temp = pd.read_html(str(tab))
    temp = pd.DataFrame(temp[0])
    airlines_wiki_list.append(temp)

airlines_wiki = pd.concat(airlines_wiki_list)

airlines_wiki.head(2)
```

```
           Airline  Image IATA ICAO   Callsign  \
0  Alaska Airlines    NaN   AS  ASA     ALASKA
1    Allegiant Air    NaN   G4  AAY  ALLEGIANT

                        Primary hubs, secondary hubs  Founded  \
0  Seattle/Tacoma Anchorage Portland (OR) San Fra...   1932.0
1  Las Vegas Cincinnati Destin/Ft. Walton Beach I...   1997.0

                                                Notes
0  Founded as McGee Airways and commenced operati...
1  Founded as WestJet Express and began operation...
```

III. Get all the information pulled so far in one table.

```
combined_data.head(2)
```

```
   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
```

```
1
1   2      US    1558         PHX       CLT          3    15     222
1

  type_source_airport  elevation_ft_source_airport  \
0       large_airport                         13.0
1       large_airport                       1135.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
97.0
1                          3.0     large_airport
748.0

   runway_count_dest_airport
0                        5.0
1                        4.0
```

finding the year founded of airlines

```
airlines_founded =
pd.merge(combined_data[['Airline']].drop_duplicates(),airlines_wiki[['
IATA', 'Founded']].drop_duplicates(),
         how = 'left', left_on = 'Airline', right_on = 'IATA')

airlines_founded
```

```
   Airline IATA  Founded
0       CO  NaN      NaN
1       US  NaN      NaN
2       AA   AA   1926.0
3       AS   AS   1932.0
4       DL   DL   1924.0
5       B6   B6   1998.0
6       HA   HA   1929.0
7       OO   OO   1972.0
8       9E   9E   1985.0
9       OH   OH   1979.0
10      EV  NaN      NaN
11      XE   XE   2016.0
12      YV   YV   1980.0
13      UA   UA   1926.0
14      MQ   MQ   1984.0
15      F9   F9   1994.0
16      WN   WN   1967.0
```

```
# will fill in missing values later
```

IV. Look into Wikipedia page: https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States Total passenger traffic may also contribute to the delay of flights. The term hub is used to identify busy commercial airports. Large hubs are the airports that each account for at least one percent of total U.S. passenger enplanements. Medium hubs are defined as airports that each account for between 0.25 percent and 1 percent of the total passenger enplanements.

Pull passenger traffic data using web scraping and collate in a table.

```
website_url =
requests.get('https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States').text
soup = BeautifulSoup(website_url,'lxml')
My_table = soup.findAll("table",{"class":"wikitable"})

hub_data = {}
i = 0
for tab in My_table:
    hub_data[i] = pd.read_html(str(tab))
    hub_data[i] = pd.DataFrame(hub_data[i][0])
    i +=1
```

We need only hub data hence first two table

```
hub_data
```

```
{0:     Rank (2022)                                    Airports (large)
IATA Code  \
 0             1  Hartsfield–Jackson Atlanta International Airport
ATL
 1             2           Dallas/Fort Worth International Airport
DFW
 2             3                      Denver International Airport
DEN
 3             4                      O'Hare International Airport
ORD
 4             5                 Los Angeles International Airport
LAX
 5             6             John F. Kennedy International Airport
JFK
 6             7                  Harry Reid International Airport
LAS
 7             8                     Orlando International Airport
MCO
 8             9                       Miami International Airport
MIA
 9            10           Charlotte Douglas International Airport
CLT
 10           11              Seattle–Tacoma International Airport
SEA
```

11	12	Phoenix Sky Harbor International Airport	PHX		
12	13	Newark Liberty International Airport	EWR		
13	14	San Francisco International Airport	SFO		
14	15	George Bush Intercontinental Airport	IAH		
15	16	Logan International Airport	BOS		
16	17	Fort Lauderdale–Hollywood International Airport	FLL		
17	18	Minneapolis–Saint Paul International Airport	MSP		
18	19	LaGuardia Airport	LGA		
19	20	Detroit Metropolitan Airport	DTW		
20	21	Philadelphia International Airport	PHL		
21	22	Salt Lake City International Airport	SLC		
22	23	Ronald Reagan Washington National Airport	DCA		
23	24	San Diego International Airport	SAN		
24	25	Baltimore/Washington International Airport	BWI		
25	26	Tampa International Airport	TPA		
26	27	Austin–Bergstrom International Airport	AUS		
27	28	Dulles International Airport	IAD		
28	29	Nashville International Airport	BNA		
29	30	Midway International Airport	MDW		
30	31	Daniel K. Inouye International Airport	HNL		

	Major cities served	Metro area	State	2022[2]	\
0	Atlanta	Atlanta	GA	45396001	
1	Dallas and Fort Worth	Dallas–Fort Worth	TX	35345138	
2	Denver	Denver	CO	33773832	

3	Chicago	Chicagoland	IL	33120474
4	Los Angeles	Greater Los Angeles	CA	32326616
5	New York City	New York Metro	NY	26919982
6	Las Vegas	Las Vegas	NV	25480500
7	Orlando	Orlando	FL	24469733
8	Miami	Miami Metro	FL	23949892
9	Charlotte	Charlotte	NC	23100300
10	Seattle and Tacoma	Seattle Metro	WA	22157862
11	Phoenix	Phoenix	AZ	21852586
12	Newark and New York City	New York Metro	NJ	21572147
13	San Francisco	San Francisco Bay Area	CA	20411420
14	Houston	Houston	TX	19814052
15	Boston	Boston	MA	17443775
16	Fort Lauderdale and Hollywood	Miami Metro	FL	15370165
17	Minneapolis and Saint Paul	Minneapolis–Saint Paul	MN	15242089
18	New York City	New York Metro	NY	14367463
19	Detroit	Detroit	MI	13751197
20	Philadelphia	Philadelphia	PA	12421168
21	Salt Lake City	Salt Lake City	UT	12383843
22	Washington, D.C.	Washington Metro	VA	11553850
23	San Diego	San Diego	CA	11162224
24	Baltimore and Washington, D.C.	Baltimore	MD	11151169
25	Tampa	Tampa	FL	10539459
26	Austin	Austin	TX	10382573
27	Washington, D.C.	Washington Metro	VA	

```
10266324
 28                        Nashville                Nashville    TN
9829062
 29                          Chicago              Chicagoland    IL
9650281
 30                         Honolulu                 Honolulu    HI
8828395

      2021[3]   2020[4]   2019[5]   2018[6]   2017[7]   2016[8]
2015[9]  \
 0   36676010  20559866  53505795  51865797  50251964  50501858
49340732
 1   30005266  18593421  35778573  32821799  31816933  31283579
31589839
 2   28645527  16243216  33592945  31362941  29809097  28267394
26280043
 3   26350976  14606034  40871223  39873927  38593028  37589899
36305668
 4   23663410  14055777  42939104  42624050  41232432  39636042
36351272
 5   15273342   8269819  31036655  30620769  29533154  29239151
27782369
 6   19160342  10584059  24728361  23795012  23364393  22833267
21857693
 7   19618838  10467728  24562271  23202480  21565448  20283541
18759938
 8   17500096   8786007  21421031  21021640  20709225  20875813
20986349
 9   20900875  12952869  24199688  22281949  22011251  21511880
21913166
 10  17430195   9462411  25001762  24024908  22639124  21887110
20148980
 11  18940287  10531436  22433552  21622580  21185458  20896265
21351504
 12  14514049   7985474  23160763  22797602  21571198  19923009
18684818
 13  11725347   7745057  27779230  27790717  26900048  25707101
24190560
 14  16242821   8682558  21905309  21157398  19603731  20062072
20595881
 15  10909817   6035452  20699377  20006521  18759742  17759044
16290362
 16  13598994   8015744  17950989  17612331  15817043  14263270
13061632
 17  12211409   7069720  19192917  18361942  18409704  18123844
17634273
 18   7827307   4147116  15393601  15058501  14614802  14762593
14319924
 19  11517696   6822324  18143040  17436837  17036092  16847135
```

16255520
 20   9820222   5753239 16006389 15292670 14271243 14564419
15101349
 21 10795906   5753239 12840841 12226730 11615954 11143738
10634538
 22   6731737   3573489 11595454 11367176 11506310 11470854
11242375
 23   7836360   4637856 12648692 12174224 11139933 10340164
9985763
 24   9253561   5451355 13284687 13371816 12976554 12340972
11738845
 25   8847197   4966775 10978756 10368514   9548580   9194994
9150458
 26   6666215   3141505   8683711   7921797   6973115   6095545
5797547
 27   7227875   3862658 11884117 11621623 11024306 10596942
10363974
 28   7594049   4013995   8935654   8017347   6902771   6338517
5715205
 29   7680617   4236603 10081781 10678018 10912074 11044387
10830850
 30   5830928   3126391   9988678   9578505   9743989   9656340
9656340

     2014[10] 2013[11]
 0   46604273 45308407
 1   30804567 29038128
 2   26000591 25496885
 3   33843426 32317835
 4   34314197 32425892
 5   26244928 25036358
 6   20620248 19946179
 7   17278608 16884524
 8   19471466 19420089
 9   21537725 21346601
 10 17888080 16690295
 11 20344867 19525109
 12 17773405 17546506
 13 22770783 21704626
 14 19772087 18952840
 15 15507561 14810153
 16 12031860 11538140
 17 16972678 16280835
 18 13535372 13372269
 19 15775941 15683523
 20 14792339 14727945
 21 10139065   9668048
 22 10057794   9838034
 23   9333152   8878772

```
 24  11022200  11132731
 25   8531561   8267752
 26   5219982   4900959
 27  10415948  10570993
 28   5396958   5050989
 29  10311996   9915646
 30   9463000   9466995  ,
 1:     Rank (2021)                              Airports (medium hubs)
IATA Code  \
 0            32                                  Dallas Love Field
DAL
 1            33                     Portland International Airport
PDX
 2            34            St. Louis Lambert International Airport
STL
 3            35                           William P. Hobby Airport
HOU
 4            36                  Sacramento International Airport
SMF
 5            37  Louis Armstrong New Orleans International Airport
MSY
 6            38               Raleigh–Durham International Airport
RDU
 7            39    Norman Y. Mineta San José International Airport
SJC
 8            40                                  John Wayne Airport
SNA
 9            41    San Francisco Bay Oakland International Airport
OAK
 10           42            Southwest Florida International Airport
RSW
 11           43             Luis Muñoz Marín International Airport
SJU
 12           44                  Kansas City International Airport
MCI
 13           45                  San Antonio International Airport
SAT
 14           46            Cleveland Hopkins International Airport
CLE
 15           47                Indianapolis International Airport
IND
 16           48                                       Kahului Airport
OGG
 17           49                  Pittsburgh International Airport
PIT
 18           50  Cincinnati/Northern Kentucky International Air...
CVG
 19           51          John Glenn Columbus International Airport
CMH
```

```
 20           52                    Palm Beach International Airport
PBI
 21           53                 Jacksonville International Airport
JAX
 22           54                         Hollywood Burbank Airport
BUR
 23           55                      Bradley International Airport
BDL
 24           56                      Ontario International Airport
ONT
 25           57            Milwaukee Mitchell International Airport
MKE
 26           58                    Charleston International Airport
CHS
 27           59         Ted Stevens Anchorage International Airport
ANC
 28           60                  Albuquerque International Sunport
ABQ
 29           61                                    Boise Airport
BOI
 30           62                                 Eppley Airfield
OMA
 31           63                      Memphis International Airport
MEM
 32           64                    Richmond International Airport
RIC
 33           65                    Reno–Tahoe International Airport
RNO

         City served              Metro Area State  2022[2]  2021[3]
2020[4]  \
 0            Dallas       Dallas–Fort Worth    TX  7819129  6487563
3669930
 1          Portland                Portland    OR  7241882  5759879
3455877
 2         St. Louis                St. Louis    MO  6709080  5070471
3041765
 3           Houston                  Houston    TX  6462948  5560780
3127178
 4        Sacramento                Sacramento    CA  6040824  4760275
2710342
 5       New Orleans              New Orleans    LA  5931899  4017147
2632606
 6           Raleigh                  Raleigh    NC  5849665  4311049
2337496
 7          San Jose  San Francisco Bay Area    CA  5590137  3619690
2283186
 8         Santa Ana     Greater Los Angeles    CA  5536313  3807205
1824836
```

9	Oakland	San Francisco Bay Area	CA	5506232	4011953	2271294
10	Fort Myers	Southwest Florida	FL	5132694	5080805	2947139
11	San Juan	San Juan	PR	5039771	4738725	2362851
12	Kansas City	Kansas City	MO	4796476	3795290	2167616
13	San Antonio	San Antonio	TX	4751610	3677643	1919958
14	Cleveland	Cleveland	OH	4237795	3552402	1990156
15	Indianapolis	Indianapolis	IN	4209416	3487100	1989126
16	Kahului	Maui	HI	4125311	2933315	1135141
17	Pittsburgh	Pittsburgh	PA	3918968	3069259	1742406
18	Cincinnati	Cincinnati	KY	3702997	3050597	1729395
19	Columbus	Columbus	OH	3618555	2825259	1577596
20	West Palm Beach	Miami Metro	FL	3257730	2567897	1518732
21	Jacksonville	Jacksonville	FL	3177393	2425685	1367501
22	Burbank	Greater Los Angeles	CA	3054729	1942417	1056838
23	Hartford	Hartford	CT	2844713	2273259	1150033
24	Ontario	Greater Los Angeles	CA	2840758	2201528	1237946
25	Milwaukee	Milwaukee	WI	2660187	2231010	1263385
26	Charleston	Charleston	SC	2608497	2015277	944660
27	Anchorage	Anchorage	AK	2604308	2184959	1157301
28	Albuquerque	Albuquerque	NM	2317836	1688646	868922
29	Boise	Boise	ID	2230467	1809000	991241
30	Omaha	Omaha	NE	2204395	1829912	1036245
31	Memphis	Memphis	TN	2163692	1793073	1015981
32	Richmond	Richmond	VA	4068689	2033816	1604459
33	Reno	Reno	NV	2132856	1781785	

976937

	2019[5]	2018[6]	2017[7]	2016[8]	2015[9]	2014[10]	2013[11]
0	8408457	8134848	7876769	7554596	7040921.0	4522341.0	4023779.0
1	9797408	9940866	9435473	9071154	8340234.0	7878760.0	7452603.0
2	7946986	7822274	7372805	6793076	6239231.0	6108758.0	6216104.0
3	7069614	6937061	6741870	6285181	5937944.0	5800726.0	5377050.0
4	6454413	6031630	5460526	4969366	4816440.0	4384616.0	4255145.0
5	6717105	6565482	6005527	5569705	5329696.0	4870569.0	4576539.0
6	6919429	6416822	5851004	5401714	4954717.0	4673869.0	4482016.0
7	7828885	7140616	6225148	5321603	4885690.0	4621003.0	4315839.0
8	5153276	5317149	5195047	5217242	4945175.0	4584147.0	4540628.0
9	6560230	6798321	6530308	5934639	5506672.0	5069257.0	4770716.0
10	5144467	4719568	4461304	4350650	4231134.0	4025959.0	3788870.0
11	4590117	4033412	4163587	4343354	4218785.0	4150828.0	4103197.0
12	5759419	5935131	5744918	5391557	5135127.0	4982722.0	4836221.0
13	5022980	4844427	4521611	4179994	4091389.0	4046856.0	4005874.0
14	4894541	4836580	4562740	4205739	4083476.0	3686315.0	4375448.0
15	4709183	4695040	4376432	4216766	3889567.0	3605908.0	3535015.0
16	3791807	3572133	3442189	3352813	3220753.0	3019338.0	2955304.0
17	4715947	4670033	4327431	3986114	3890677.0	3827860.0	3812460.0
18	4413457	4269258	3926158	3269979	3036697.0	2874684.0	2776377.0
19	4172067	4054572	3765007	3567864	3312496.0	3115501.0	3063822.0
20	3460429	3263042	3166532	3100624	3113485.0	2926242.0	2844507.0
21	3479923	3118540	2759067	2799587	2716465.0	2589198.0	2549070.0
22	2988720	2680240	2402106	2077892	1973897.0	1928491.0	

```
1918011.0
 23  3323614  3330734  3214976  2982194  2926047.0  2913380.0
2681181.0
 24  2723002  2499171  2247645  2127387  2089801.0  2037346.0
1970538.0
 25  3374073  3548817  3452544  3383271  3229876.0  3228607.0
3214811.0
 26  2375868  2192893  1945699  1811695  1669988.0  1539326.0
1441415.0
 27  2713843  2642901  2556188  2563524  2525876.0  2381826.0
2325030.0
 28  2641450  2647269  2412328  2341719  2323883.0  2354184.0
2477783.0
 29  2057750  1943181  1777642  1633507  1487777.0  1378352.0
1313741.0
 30  2455274  2457087  2303223  2127387  2046155.0  2020354.0
1975339.0
 31  2318442  2213083  2102739  2016089  1873716.0  1800268.0
2301003.0
 32  4038000  4077763  3657479  3421034        NaN        NaN
NaN
 33  2162250  2048916  1953028  1771864  1669876.0  1611572.0
1671926.0  ,
 2:     Rank                                          Airport name  \
 0       1              John F. Kennedy International Airport
 1       2                       Miami International Airport
 2       3                  Los Angeles International Airport
 3       4               George Bush Intercontinental Airport
 4       5               Newark Liberty International Airport
 5       6            Dallas/Fort Worth International Airport
 6       7   Hartsfield–Jackson Atlanta International Airport
 7       8                         O'Hare International Airport
 8       9    Fort Lauderdale–Hollywood International Airport
 9      10            Washington Dulles International Airport
 10     11                San Francisco International Airport
 11     12  General Edward Lawrence Logan International Ai...
 12     13            Charlotte Douglas International Airport
 13     14                        Denver International Airport
 14     15                      Orlando International Airport
 15     16               Seattle–Tacoma International Airport
 16     17           Phoenix Sky Harbor International Airport
 17     18                 Philadelphia International Airport
 18     19          Detroit Metropolitan Wayne County Airport
 19     20                   Harry Reid International Airport

                            Location IATA Code  2021[12]  2020[13]
2019[14]
 0                  Queens, New York       JFK  12466165   8219317
33432159
```

```
 1                    Miami, Florida       MIA  11592445   6565834
20735658
 2           Los Angeles, California       LAX   7862532   6246602
25210140
 3                    Houston, Texas       IAH   6458473   3491935
10764589
 4                Newark, New Jersey       EWR   6250880   3688541
14087622
 5                     Irving, Texas       DFW   5852397   3268822
9103438
 6             College Park, Georgia       ATL   5474264   3347184
12268779
 7                 Chicago, Illinois       ORD   5148494   3481860
13412885
 8          Fort Lauderdale, Florida       FLL   4016553   2839383
8524251
 9                   Dulles, Virginia       IAD   3230027   1917510
7990292
 10  South San Francisco, California       SFO   3139041   3210024
14357960
 11            Boston, Massachusetts       BOS   2046561   1574712
7534504
 12        Charlotte, North Carolina       CLT   1989704   1069001
3405907
 13                Denver, Colorado       DEN   1856124    934563
3037012
 14                Orlando, Florida       MCO   1837706   1525177
6957048
 15               SeaTac, Washington       SEA   1393603   1273179
5392147
 16                Phoenix, Arizona       PHX   1223856    750138
1958468
 17       Philadelphia, Pennsylvania       PHL    988733    682030
3847253
 18                Romulus, Michigan       DTW    966375    873744
3717775
 19                 Paradise, Nevada       LAS    738257    711614
3462627  ,
 3:   Rank                                      Airport name  \
   Rank                                    Airport name
 0    1                     Memphis International Airport
 1    2        Ted Stevens Anchorage International Airport
 2    3      Louisville Muhammad Ali International Airport
 3    4                      O'Hare International Airport
 4    5                       Miami International Airport
 5    6                  Los Angeles International Airport
 6    7  Cincinnati/Northern Kentucky International Air...
 7    8                 Indianapolis International Airport
 8    9            Dallas/Fort Worth International Airport
 9   10                      Ontario International Airport
```

```
                  Location IATA code        Cargo
                  Location IATA code         Ibs. % chg. 2017/16
 0       Memphis, Tennessee       MEM  23949525780         00.35%
 1        Anchorage, Alaska       ANC  17337337377         02.79%
 2     Louisville, Kentucky       SDF  13403682652         04.68%
 3        Chicago, Illinois       ORD  10373559593        010.84%
 4           Miami, Florida       MIA   7963988407         00.82%
 5  Los Angeles, California       LAX   7197930264         03.85%
 6         Hebron, Kentucky       CVG   5700282994        033.32%
 7    Indianapolis, Indiana       IND   5138500318        0-3.58%
 8            Irving, Texas       DFW   4155362297         07.65%
 9      Ontario, California       ONT   3522510318        015.81%  }
```

```
large_hub = hub_data[0].copy()
med_hub = hub_data[1].copy()

large_hub.insert(loc =1, column= 'Hub_type', value = 'large')
med_hub.insert(loc =1, column= 'Hub_type', value = 'medium')

# before combinig lets work with column names

# remove any special characters or things in bracket
large_hub.columns
```

```
Index(['Rank (2022)', 'Hub_type', 'Airports (large)', 'IATA Code',
       'Major cities served', 'Metro area', 'State', '2022[2]',
'2021[3]',
       '2020[4]', '2019[5]', '2018[6]', '2017[7]', '2016[8]',
'2015[9]',
       '2014[10]', '2013[11]'],
      dtype='object')
```

```
# remove refrences from brackets
column_temp =
large_hub.columns.str.split('[([]').str[0].str.strip().str.lower().str
.replace(' ','_').values
column_temp[list(map( lambda x : x.isnumeric(), column_temp))]  =
'data_' + column_temp[list(map( lambda x : x.isnumeric(),
column_temp))]
large_hub.columns = column_temp
large_hub.columns
```

```
Index(['rank', 'hub_type', 'airports', 'iata_code',
'major_cities_served',
       'metro_area', 'state', 'data_2022', 'data_2021', 'data_2020',
       'data_2019', 'data_2018', 'data_2017', 'data_2016',
'data_2015',
       'data_2014', 'data_2013'],
      dtype='object')
```

```
# remove refrences from brackets
column_temp =
med_hub.columns.str.split('[([]').str[0].str.strip().str.lower().str.r
eplace(' ','_').values
column_temp[list(map( lambda x : x.isnumeric(), column_temp))]  =
'data_' + column_temp[list(map( lambda x : x.isnumeric(),
column_temp))]
med_hub.columns = column_temp
med_hub.columns
```

```
Index(['rank', 'hub_type', 'airports', 'iata_code', 'city_served',
       'metro_area', 'state', 'data_2022', 'data_2021', 'data_2020',
       'data_2019', 'data_2018', 'data_2017', 'data_2016',
'data_2015',
       'data_2014', 'data_2013'],
      dtype='object')
```

```
large_hub.rename(columns = {'major_cities_served':'city_served'},
inplace = True)

final_hub_data = pd.concat([large_hub, med_hub])

final_hub_data.head(2)
```

```
   rank hub_type                                          airports
iata_code  \
0     1    large  Hartsfield–Jackson Atlanta International Airport
ATL
1     2    large           Dallas/Fort Worth International Airport
DFW

             city_served         metro_area state  data_2022
data_2021  \
0                Atlanta            Atlanta    GA   45396001
36676010
1  Dallas and Fort Worth  Dallas–Fort Worth    TX   35345138
30005266

   data_2020  data_2019  data_2018  data_2017  data_2016
data_2015  \
0   20559866   53505795   51865797   50251964   50501858  49340732.0

1   18593421   35778573   32821799   31816933   31283579  31589839.0


    data_2014   data_2013
0  46604273.0  45308407.0
1  30804567.0  29038128.0
```

```
final_hub_data.data_2019.isnull().sum()
```

```
0
```

```
final_hub_data.isnull().sum()
```

```
rank           0
hub_type       0
airports       0
iata_code      0
city_served    0
metro_area     0
state          0
data_2022      0
data_2021      0
data_2020      0
data_2019      0
data_2018      0
data_2017      0
data_2016      0
data_2015      1
data_2014      1
data_2013      1
dtype: int64
```

```
final_hub_data.columns
```

```
Index(['rank', 'hub_type', 'airports', 'iata_code', 'city_served',
       'metro_area', 'state', 'data_2022', 'data_2021', 'data_2020',
       'data_2019', 'data_2018', 'data_2017', 'data_2016',
'data_2015',
       'data_2014', 'data_2013'],
      dtype='object')
```

```
combined_data_pax = pd.merge(combined_data,
final_hub_data[['iata_code', 'data_2019']],how = 'left' , left_on =
'AirportFrom', right_on = 'iata_code')
```

```
combined_data_pax.rename(columns = {'iatacode':
'iatacode_source' ,'data_2019': 'data_2019_source_airport'}, inplace =
True)
```

```
combined_data_pax = pd.merge(combined_data_pax,
final_hub_data[['iata_code', 'data_2019']],how = 'left' , left_on =
'AirportTo', right_on = 'iata_code')
```

```
combined_data_pax.rename(columns = {'iata_code':
'iatacode_dest' ,'data_2019': 'data_2019_dest_airport'}, inplace =
True)
```

```
combined_data_pax =
combined_data_pax.loc[:,~combined_data_pax.columns.str.startswith('iat
acode')].copy()
```

```
combined_data_pax

            id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time
Length  \
0            1      CO     269         SFO       IAH          3    15
205
1            2      US    1558         PHX       CLT          3    15
222
2            3      AA    2400         LAX       DFW          3    20
165
3            4      AA    2466         SFO       DFW          3    20
195
4            5      AS     108         ANC       SEA          3    30
202
...        ...     ...     ...         ...       ...        ...   ...
...
518551  539377      B6     717         JFK       SJU          5  1439
220
518552  539378      B6     739         JFK       PSE          5  1439
223
518553  539379      CO     178         OGG       SNA          5  1439
326
518554  539382      UA      78         HNL       SFO          5  1439
313
518555  539383      US    1442         LAX       PHL          5  1439
301

        Delay type_source_airport  elevation_ft_source_airport  \
0           1       large_airport                         13.0
1           1       large_airport                       1135.0
2           1       large_airport                        125.0
3           1       large_airport                         13.0
4           0       large_airport                        152.0
...       ...                 ...                          ...
518551      1       large_airport                         13.0
518552      1       large_airport                         13.0
518553      0      medium_airport                         54.0
518554      1       large_airport                         13.0
518555      1       large_airport                        125.0

        runway_count_source_airport type_dest_airport  \
0                               4.0     large_airport
1                               3.0     large_airport
2                               4.0     large_airport
3                               4.0     large_airport
4                               3.0     large_airport
...                             ...               ...
518551                          4.0     large_airport
518552                          4.0    medium_airport
518553                          2.0     large_airport
```

```
518554                            6.0     large_airport
518555                            4.0     large_airport

        elevation_ft_dest_airport  runway_count_dest_airport
iata_code_x  \
0                           97.0                         5.0
SFO
1                          748.0                         4.0
PHX
2                          607.0                         7.0
LAX
3                          607.0                         7.0
SFO
4                          433.0                         4.0
ANC
...                          ...                         ...          .
..
518551                       9.0                         2.0
JFK
518552                      29.0                         1.0
JFK
518553                      56.0                         2.0
OGG
518554                      13.0                         4.0
HNL
518555                      36.0                         4.0
LAX

        data_2019_source_airport iata_code_y  data_2019_dest_airport
0                     27779230.0         IAH              21905309.0
1                     22433552.0         CLT              24199688.0
2                     42939104.0         DFW              35778573.0
3                     27779230.0         DFW              35778573.0
4                      2713843.0         SEA              25001762.0
...                          ...         ...                     ...
518551                31036655.0         SJU               4590117.0
518552                31036655.0         NaN                     NaN
518553                 3791807.0         SNA               5153276.0
518554                 9988678.0         SFO              27779230.0
518555                42939104.0         PHL              16006389.0

[518556 rows x 19 columns]
```

addd founded column

```
airlines_founded

   Airline IATA  Founded
0       CO  NaN      NaN
1       US  NaN      NaN
```

```
2       AA   AA   1926.0
3       AS   AS   1932.0
4       DL   DL   1924.0
5       B6   B6   1998.0
6       HA   HA   1929.0
7       OO   OO   1972.0
8       9E   9E   1985.0
9       OH   OH   1979.0
10      EV  NaN      NaN
11      XE   XE   2016.0
12      YV   YV   1980.0
13      UA   UA   1926.0
14      MQ   MQ   1984.0
15      F9   F9   1994.0
16      WN   WN   1967.0
```

```
combined_data_pax = pd.merge(combined_data_pax,
airlines_founded[['Airline', 'Founded']], on = 'Airline')

combined_data_pax.head(2)
```

```
   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   6      CO    1094         LAX       IAH          3    30     181
1

  type_source_airport  elevation_ft_source_airport  \
0       large_airport                         13.0
1       large_airport                        125.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
97.0
1                          4.0     large_airport
97.0

   runway_count_dest_airport iata_code_x  data_2019_source_airport  \
0                        5.0         SFO                27779230.0
1                        5.0         LAX                42939104.0

  iata_code_y  data_2019_dest_airport  Founded
0         IAH              21905309.0      NaN
1         IAH              21905309.0      NaN
```

1. Check the missing values in each field. Perform missing value treatment. Justify your actions

combined pax

```
combined_data_pax.shape
```

```
(518556, 20)
```

```
combined_data_pax.isna().sum().sort_values(ascending = False)
```

```
Founded                          83601
data_2019_source_airport         83586
iata_code_x                      83586
data_2019_dest_airport           83536
iata_code_y                      83536
runway_count_source_airport         31
runway_count_dest_airport           31
elevation_ft_dest_airport           31
type_dest_airport                   31
elevation_ft_source_airport         31
type_source_airport                 31
Airline                              0
Delay                                0
Length                               0
Time                                 0
DayOfWeek                            0
AirportTo                            0
AirportFrom                          0
Flight                               0
id                                   0
dtype: int64
```

for type runway count and elevation lets get the airports for which information is missing

```
combined_data_pax[combined_data_pax.type_source_airport.isna()].Airport
tFrom.unique()
```

```
array(['CYS'], dtype=object)
```

```
combined_data_pax[combined_data_pax.type_dest_airport.isna()].AirportT
o.unique()
```

```
array(['CYS'], dtype=object)
```

As we see information for only CYS is missing Lets check for this information using data dictionary and match the description and name of the airport to fetch information

```
airport_dict = pd.read_excel('Data Dictionary.xlsx', sheet_name =
'airlines',header = 29)
airport_dict.head(2)
```

```
  Aiport ID    Description  Unnamed: 2
0       ABE   RAF Calveley         NaN
1       ABE  Bisho Airport         NaN
```

```
airport_dict = pd.read_excel('Data Dictionary.xlsx', sheet_name =
'airlines',header = 29, usecols = [0,1])
airport_dict.head(2)
```

```
  Aiport ID    Description
0       ABE   RAF Calveley
1       ABE  Bisho Airport
```

```
airport_dict[airport_dict['Aiport ID'] == 'CYS'].Description
```

```
194    Cheyenne Regional Jerry Olson Field
Name: Description, dtype: object
```

```
name = airport_dict[airport_dict['Aiport ID'] ==
'CYS'].Description.values[0]
name.lower()
```

```
'cheyenne regional jerry olson field'
```

```
air_miss = airports.loc[name.lower() == airports.name.str.lower(),
['ident', 'name', 'iata_code', 'type', 'elevation_ft']]
```

```
air_miss.head(2)
```

```
      ident                                     name iata_code
type  \
34675  KCYS  Cheyenne Regional Jerry Olson Field       NaN
medium_airport

       elevation_ft
34675        6159.0
```

```
air_miss_comb = pd.merge(air_miss, runways[['airport_ident', 'id']],
how = 'left', left_on = 'ident', right_on = 'airport_ident')
runway_count_miss = air_miss_comb.groupby('ident')
[['id']].count().sort_values(by = 'id', ascending =
False).reset_index()
runway_count_miss
```

```
  ident  id
0  KCYS   2
```

```
air_miss_data = pd.merge(air_miss,runway_count_miss ).rename(columns =
{'id' : 'runway_count'})[['iata_code', 'type', 'elevation_ft',
'runway_count']]

combined_data_pax.loc[combined_data_pax.AirportFrom == 'CYS',
'type_source_airport'] = air_miss_data.type.values[0]
combined_data_pax.loc[combined_data_pax.AirportFrom == 'CYS',
'elevation_ft_source_airport'] = air_miss_data.elevation_ft.values[0]
combined_data_pax.loc[combined_data_pax.AirportFrom == 'CYS',
'runway_count_source_airport'] = air_miss_data.runway_count.values[0]

combined_data_pax.loc[combined_data_pax.AirportTo == 'CYS',
'type_dest_airport'] = air_miss_data.type.values[0]
combined_data_pax.loc[combined_data_pax.AirportTo == 'CYS',
'elevation_ft_dest_airport'] = air_miss_data.elevation_ft.values[0]
combined_data_pax.loc[combined_data_pax.AirportTo == 'CYS',
'runway_count_dest_airport'] = air_miss_data.runway_count.values[0]

combined_data_pax.isna().sum().sort_values(ascending = False)
```

```
Founded                         83601
data_2019_source_airport        83586
iata_code_x                     83586
data_2019_dest_airport          83536
iata_code_y                     83536
Airline                             0
runway_count_dest_airport           0
elevation_ft_dest_airport           0
type_dest_airport                   0
runway_count_source_airport         0
id                                  0
type_source_airport                 0
Delay                               0
Length                              0
Time                                0
DayOfWeek                           0
AirportTo                           0
AirportFrom                         0
Flight                              0
elevation_ft_source_airport         0
dtype: int64
```

```
airline_dict = pd.read_excel('Data Dictionary.xlsx', sheet_name =
'airlines',header = 10, usecols = [0,1])
airline_dict.head(2)
```

```
  Airlines ID Description
0          WN   Southwest
1          DL       Delta
```

```
miss_founded =
combined_data_pax[combined_data_pax.Founded.isna()].Airline.unique()
print(airline_dict[airline_dict['Airlines ID'].isin( ['EV', 'CO',
'US'])])
```

```
  Airlines ID                          Description
5          US  PSA (initially US Airway Express)
7          EV                          ExpressJet
9          CO     United Airlines (initially CO)
```

```
miss_val = {'US' : 1967, 'CO' : 1934, 'EV' : 1986}
for aline in miss_founded:
    combined_data_pax.loc[(combined_data_pax.Founded.isna()) &
                          (combined_data_pax.Airline == aline), 'Founded']
= miss_val[aline]
```

```
(combined_data_pax.isna().sum().sort_values(ascending =
False)/combined_data_pax.shape[0])*100
```

```
data_2019_source_airport       16.118992
iata_code_x                    16.118992
data_2019_dest_airport         16.109350
iata_code_y                    16.109350
id                              0.000000
Airline                         0.000000
runway_count_dest_airport       0.000000
elevation_ft_dest_airport       0.000000
type_dest_airport               0.000000
runway_count_source_airport     0.000000
elevation_ft_source_airport     0.000000
type_source_airport             0.000000
Delay                           0.000000
Length                          0.000000
Time                            0.000000
DayOfWeek                       0.000000
AirportTo                       0.000000
AirportFrom                     0.000000
Flight                          0.000000
Founded                         0.000000
dtype: float64
```

For missing pax data use median value based on 'type' of airport

```
combined_data_pax.groupby('type_source_airport')
[['data_2019_source_airport']].median()
```

```
                      data_2019_source_airport
type_source_airport
large_airport                      21905309.0
```

```
medium_airport                          3323614.0
small_airport                                 NaN
```

```
med_val = combined_data_pax.groupby('type_source_airport')
[['data_2019_source_airport']].median()
med_val
```

```
                      data_2019_source_airport
type_source_airport
large_airport                      21905309.0
medium_airport                      3323614.0
small_airport                             NaN
```

```
for typ in combined_data_pax.type_source_airport.unique():
      combined_data_pax.loc[(combined_data_pax.type_source_airport ==
typ)& (combined_data_pax.data_2019_source_airport.isna()),
                       'data_2019_source_airport'] =
med_val.loc[typ].values[0]
```

```
combined_data_pax.columns
```

```
Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo',
'DayOfWeek',
       'Time', 'Length', 'Delay', 'type_source_airport',
       'elevation_ft_source_airport', 'runway_count_source_airport',
       'type_dest_airport', 'elevation_ft_dest_airport',
       'runway_count_dest_airport', 'iata_code_x',
'data_2019_source_airport',
       'iata_code_y', 'data_2019_dest_airport', 'Founded'],
      dtype='object')
```

```
# med_val_dest = combined_data_pax.groupby('type_dest_airport')
[['data_2019_dest_airport']].median()
# med_val_dest
```

```
for typ in combined_data_pax.type_source_airport.unique():
      combined_data_pax.loc[(combined_data_pax.type_dest_airport ==
typ)& (combined_data_pax.data_2019_dest_airport.isna()),
                       'data_2019_dest_airport'] =
med_val.loc[typ].values[0]
```

```
combined_data_pax.head(2)
```

```
   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   6      CO    1094         LAX       IAH          3    30     181
1

  type_source_airport  elevation_ft_source_airport  \
```

```
0       large_airport                          13.0
1       large_airport                         125.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
97.0
1                          4.0     large_airport
97.0

   runway_count_dest_airport iata_code_x  data_2019_source_airport  \
0                        5.0         SFO                27779230.0
1                        5.0         LAX                42939104.0

  iata_code_y  data_2019_dest_airport  Founded
0         IAH              21905309.0   1934.0
1         IAH              21905309.0   1934.0
```

```
(combined_data_pax.isna().sum().sort_values(ascending =
False)/combined_data_pax.shape[0])*100
```

```
iata_code_x                   16.118992
iata_code_y                   16.109350
data_2019_source_airport       0.226205
data_2019_dest_airport         0.224855
id                             0.000000
Airline                        0.000000
runway_count_dest_airport      0.000000
elevation_ft_dest_airport      0.000000
type_dest_airport              0.000000
runway_count_source_airport    0.000000
elevation_ft_source_airport    0.000000
type_source_airport            0.000000
Delay                          0.000000
Length                         0.000000
Time                           0.000000
DayOfWeek                      0.000000
AirportTo                      0.000000
AirportFrom                    0.000000
Flight                         0.000000
Founded                        0.000000
dtype: float64
```

Since % of values missing is 0.2% we can simply eliminate these rows

2. Perform data visualization and share your insights related to following aspects:

I. According to the data provided, around 70% of the flights are delayed for Southwest airlines. Visualize to compare the same for other airlines.
II. No delayed flights on different weekdays. Which days of the week are safest to travel.
III. Which airlines to recommend for short, medium and long length of travel.

Do you observe any pattern in the time of departure of flights of long duration

```
# get id for "southwest Airlines"
id_airline =
airline_dict.loc[airline_dict['Description'].str.strip().str.lower()
== 'southwest', 'Airlines ID'].values[0]

round(combined_data_pax[combined_data_pax.Airline ==
id_airline].Delay.sum()/
      combined_data_pax[combined_data_pax.Airline ==
id_airline].Delay.size*100)
```

```
70
```

```
def percent_Delay(x):
    return round(x.sum()/x.size * 100,2)

delay_perc = combined_data_pax.groupby('Airline')
['Delay'].agg(percent_Delay)

delay_perc
```

```
Airline
9E    39.77
AA    38.85
AS    33.93
B6    46.70
CO    56.62
DL    45.05
EV    40.22
F9    44.90
HA    32.02
MQ    34.81
OH    27.73
OO    45.29
UA    32.39
US    33.60
WN    69.78
XE    37.89
```

```
YV    24.29
Name: Delay, dtype: float64
```

```
delay_perc = delay_perc.reset_index()

plot_data = pd.merge(delay_perc, airline_dict, left_on = 'Airline',
                     right_on = 'Airlines ID', how = 'left')
[['Airline', 'Description', 'Delay']]
plot_data
```

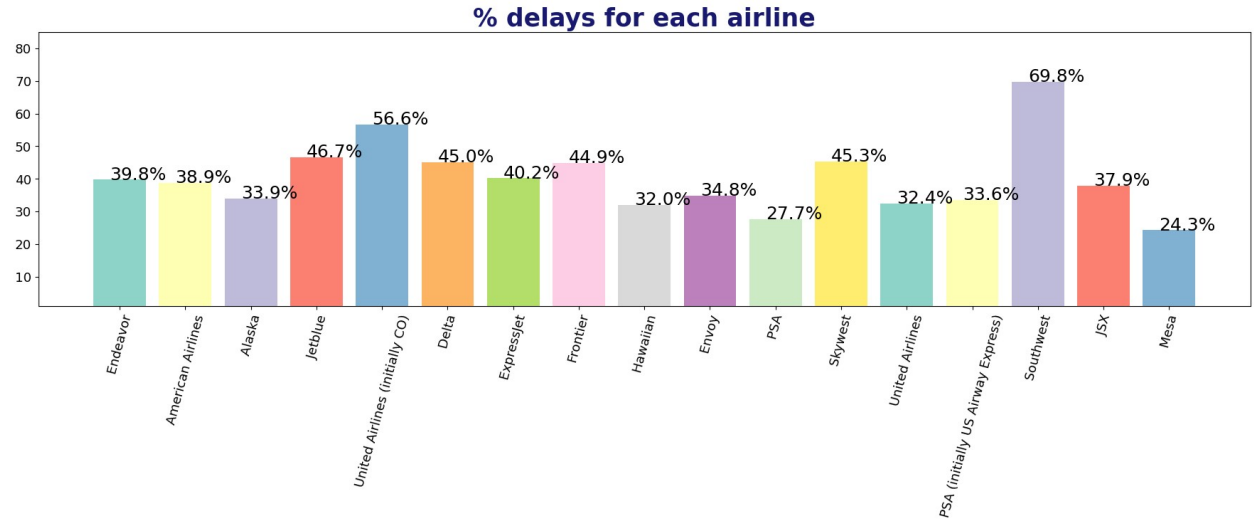
	Airline	Description	Delay
0	9E	Endeavor	39.77
1	AA	American Airlines	38.85
2	AS	Alaska	33.93
3	B6	Jetblue	46.70
4	CO	United Airlines (initially CO)	56.62
5	DL	Delta	45.05
6	EV	ExpressJet	40.22
7	F9	Frontier	44.90
8	HA	Hawaiian	32.02
9	MQ	Envoy	34.81
10	OH	PSA	27.73
11	OO	Skywest	45.29
12	UA	United Airlines	32.39
13	US	PSA (initially US Airway Express)	33.60
14	WN	Southwest	69.78
15	XE	JSX	37.89
16	YV	Mesa	24.29

```
plt.figure(figsize = (22,5))
plt.bar(plot_data.Description, height =  plot_data.Delay, color =
plt.get_cmap('Set3').colors)
for v, idx in zip(plot_data.Delay.values,plot_data.index ):
    plt.annotate('{:.1f}%'.format(v), xy = (idx-0.15, v), size = 18,
family = 'times')
plt.ylim(1,85)
plt.xticks(size = 13, rotation = 75)
plt.yticks(size = 13)
plt.title('% delays for each airline', size = 25, color =
'midnightblue', weight = 'heavy', family = 'times')
plt.show()
```

```
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
```

```
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
```

```
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
```




```
II.  No delayed flights on different weekdays. Which days of the week
are safest to travel.

combined_data_pax.head()

   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
0   1      CO     269         SFO       IAH          3    15     205
1
1   6      CO    1094         LAX       IAH          3    30     181
1
2  11      CO     223         ANC       SEA          3    49     201
1
3  18      CO    1496         LAS       IAH          3    60     162
0
4  20      CO     507         ONT       IAH          3    75     167
0

  type_source_airport  elevation_ft_source_airport  \
0       large_airport                         13.0
1       large_airport                        125.0
2       large_airport                        152.0
3       large_airport                       2181.0
4       large_airport                        944.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
```

```
97.0
1                            4.0     large_airport
97.0
2                            3.0     large_airport
433.0
3                            4.0     large_airport
97.0
4                            2.0     large_airport
97.0

   runway_count_dest_airport iata_code_x  data_2019_source_airport  \
0                        5.0         SFO                27779230.0
1                        5.0         LAX                42939104.0
2                        4.0         ANC                 2713843.0
3                        5.0         LAS                24728361.0
4                        5.0         ONT                 2723002.0

  iata_code_y  data_2019_dest_airport  Founded
0         IAH              21905309.0   1934.0
1         IAH              21905309.0   1934.0
2         SEA              25001762.0   1934.0
3         IAH              21905309.0   1934.0
4         IAH              21905309.0   1934.0
```

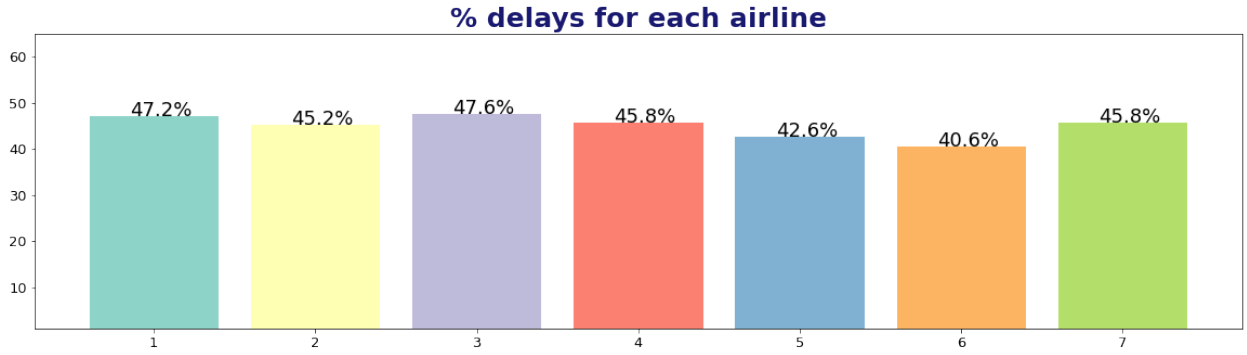
```
delay_perc_weekday = combined_data_pax.groupby('DayOfWeek')
['Delay'].agg(percent_Delay)
delay_perc_weekday
```

```
DayOfWeek
1    47.22
2    45.21
3    47.58
4    45.78
5    42.56
6    40.56
7    45.77
Name: Delay, dtype: float64
```

```
plt.figure(figsize = (20,5))
plt.bar(delay_perc_weekday.index, height =  delay_perc_weekday.values,
color = plt.get_cmap('Set3').colors)
for v, idx in zip(delay_perc_weekday.values, range(1,
len(delay_perc_weekday.index)+1)):
    # print(v, idx)
    plt.annotate('{:.1f}%'.format(v), xy = (idx-0.15, v), size = 18,
family = 'times')
plt.ylim(1,65)
plt.xticks(size = 13)
plt.yticks(size = 13)
plt.title('% delays for each airline', size = 25, color =
```

```
'midnightblue', weight = 'heavy', family = 'times')
plt.show()
```

```
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
findfont: Font family 'times' not found.
```



```
III.  Which airlines to recommend for short, medium and long length of
travel.

duration_data = combined_data_pax[['Airline', 'Length',
'Delay']].copy()
```

```
duration_data.head()
```

	Airline	Length	Delay
0	CO	205	1
1	CO	181	1
2	CO	201	1
3	CO	162	0
4	CO	167	0

```
duration_data['duration'] = pd.cut(duration_data.Length, 3, labels =
['short', 'medium', 'long'])

duration_data.head()
```

	Airline	Length	Delay	duration
0	CO	205	1	short
1	CO	181	1	short
2	CO	201	1	short
3	CO	162	0	short
4	CO	167	0	short

```
duration_data_grp = duration_data.groupby(['Airline','duration'])
['Delay'].agg(
    percent_Delay).reset_index().pivot(index = 'Airline',
                                       columns = 'duration').fillna(0)
['Delay']
duration_data_grp.columns = duration_data_grp.columns.astype(str)
duration_data_grp.reset_index()
```

duration	Airline	short	medium	long
0	9E	39.77	0.00	0.00
1	AA	37.62	43.25	60.40
2	AS	32.58	38.17	0.00
3	B6	45.70	51.05	0.00
4	CO	52.88	64.96	66.87
5	DL	43.88	50.24	48.62
6	EV	40.22	50.00	0.00
7	F9	45.03	43.56	0.00
8	HA	30.16	40.48	0.00
9	MQ	34.82	27.42	0.00
10	OH	27.61	39.20	0.00
11	OO	45.25	53.03	0.00
12	UA	29.92	37.10	39.26
13	US	31.96	40.72	0.00
14	WN	69.12	77.61	0.00
15	XE	37.87	53.70	0.00
16	YV	24.28	25.86	0.00

```
duration_data.index
```

```
RangeIndex(start=0, stop=518556, step=1)
```

```
# get names of airlines also

airline_dict
```

```
    Airlines ID                Description
0            WN                  Southwest
1            DL                      Delta
2            OO                    Skywest
3            AA          American Airlines
4            MQ                      Envoy
..          ...                        ...
683         XNA  Nambour Hospital Helipad
684         YAK        Aussenkehr Airport
685         YAK        Congo Town Airport
686         YAK          Yalkulka Airport
687         YUM          Yuinmery Airport

[688 rows x 2 columns]
```

```
airline_dict.Description = airline_dict.Description.str.strip()
duration_data_grp = pd.merge(duration_data_grp,airline_dict[['Airlines
ID', 'Description']],
         left_on = 'Airline', right_on = 'Airlines ID',
         how = 'left')
duration_data_grp
```

```
    short  medium   long Airlines ID
Description
0   39.77    0.00   0.00          9E
Endeavor
1   37.62   43.25  60.40          AA                    American
Airlines
2   32.58   38.17   0.00          AS
Alaska
3   45.70   51.05   0.00          B6
Jetblue
4   52.88   64.96  66.87          CO     United Airlines (initially
CO)
5   43.88   50.24  48.62          DL
Delta
6   40.22   50.00   0.00          EV
ExpressJet
7   45.03   43.56   0.00          F9
Frontier
8   30.16   40.48   0.00          HA
Hawaiian
9   34.82   27.42   0.00          MQ
Envoy
10  27.61   39.20   0.00          OH
PSA
```

```
11  45.25   53.03   0.00          OO
Skywest
12  29.92   37.10  39.26          UA                      United
Airlines
13  31.96   40.72   0.00          US  PSA (initially US Airway
Express)
14  69.12   77.61   0.00          WN
Southwest
15  37.87   53.70   0.00          XE
JSX
16  24.28   25.86   0.00          YV
Mesa
```

```
combined_data_pax.Airline.nunique()
```

```
17
```

```
long = duration_data_grp[duration_data_grp.long ==
duration_data_grp.long.min()].Description.values.tolist()
print('Airlines with no delays for long flights :\n',', '.join(long))
medium = duration_data_grp[duration_data_grp.medium ==
duration_data_grp.medium.min()].Description.values.tolist()
print('\nAirlines with no delays for medium flights :\n', ',
'.join(medium))
short = duration_data_grp[duration_data_grp.short ==
duration_data_grp.short.min()].Description.values.tolist()
print('\nAirlines with no delays for short flights :\n',  ',
'.join(short)     )
```

```
Airlines with no delays for long flights :
 Endeavor, Alaska, Jetblue, ExpressJet, Frontier, Hawaiian, Envoy,
PSA, Skywest, PSA (initially US Airway Express), Southwest, JSX, Mesa

Airlines with no delays for medium flights :
 Endeavor

Airlines with no delays for short flights :
 Mesa
```

```
IV.  Do you observe any pattern in the time of departure of flights of
long duration
```

```
combined_data_pax['duration'] = pd.cut(combined_data_pax.Length, 3,
labels = ['short', 'medium', 'long'])
```

```
combined_data_pax.head(2)
```

```
   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
```

```
0   1      CO     269          SFO       IAH          3    15     205
1
1   6      CO    1094          LAX       IAH          3    30     181
1

  type_source_airport  ...  type_dest_airport
elevation_ft_dest_airport  \
0       large_airport  ...      large_airport
97.0
1       large_airport  ...      large_airport
97.0

  runway_count_dest_airport  iata_code_x  data_2019_source_airport  \
0                       5.0          SFO                27779230.0
1                       5.0          LAX                42939104.0

  iata_code_y  data_2019_x data_2019_y  Founded  duration
0         IAH   21905309.0  21905309.0   1934.0     short
1         IAH   21905309.0  21905309.0   1934.0     short

[2 rows x 22 columns]
```

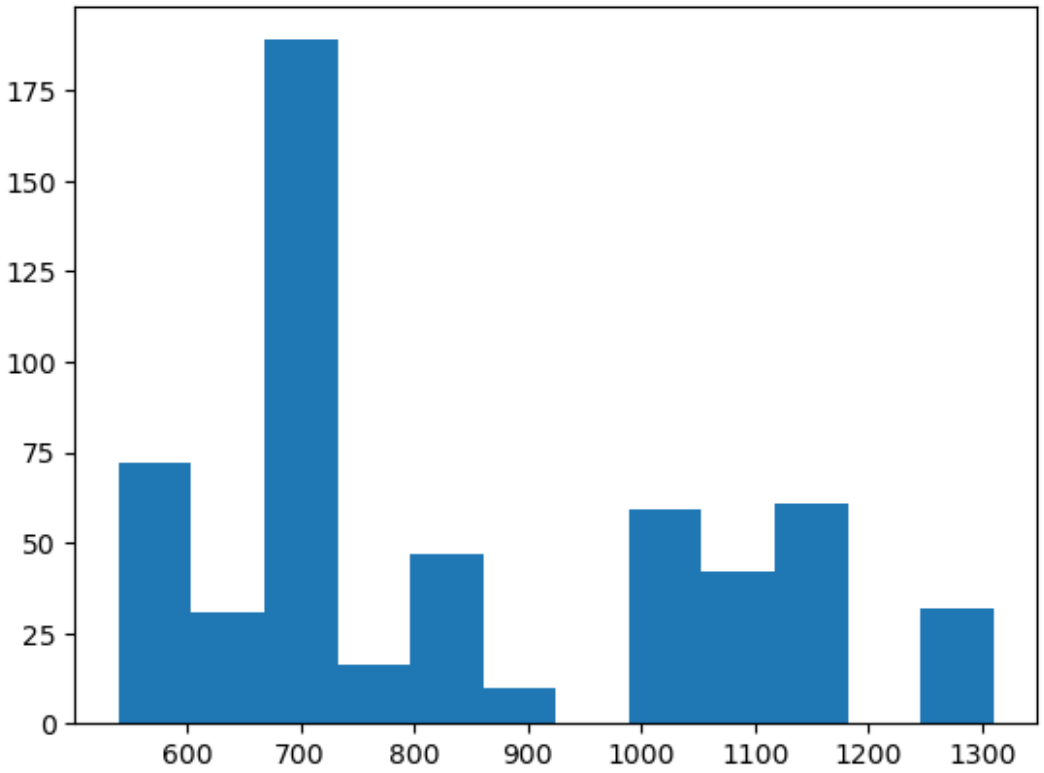
```
pd.crosstab(combined_data_pax.Time, combined_data_pax.duration)
['long']
```

```
Time
10      0
15      0
20      0
21      0
25      0
       ..
1428    0
1430    0
1431    0
1435    0
1439    0
Name: long, Length: 1131, dtype: int64
```

```
y = pd.crosstab(combined_data_pax.Time, combined_data_pax.duration)
['long'].index
x = pd.crosstab(combined_data_pax.Time, combined_data_pax.duration)
['long'].values

filter_data = combined_data_pax.loc[combined_data_pax.duration ==
'long', ['Time', 'duration']]

filter_data.Time.describe()
```

```
count     559.000000
mean      840.635063
```

```
std       221.020092
min       540.000000
25%       670.000000
50%       717.000000
75%      1045.000000
max      1310.000000
Name: Time, dtype: float64

plt.hist(filter_data.Time, bins = 12)
plt.show()
```



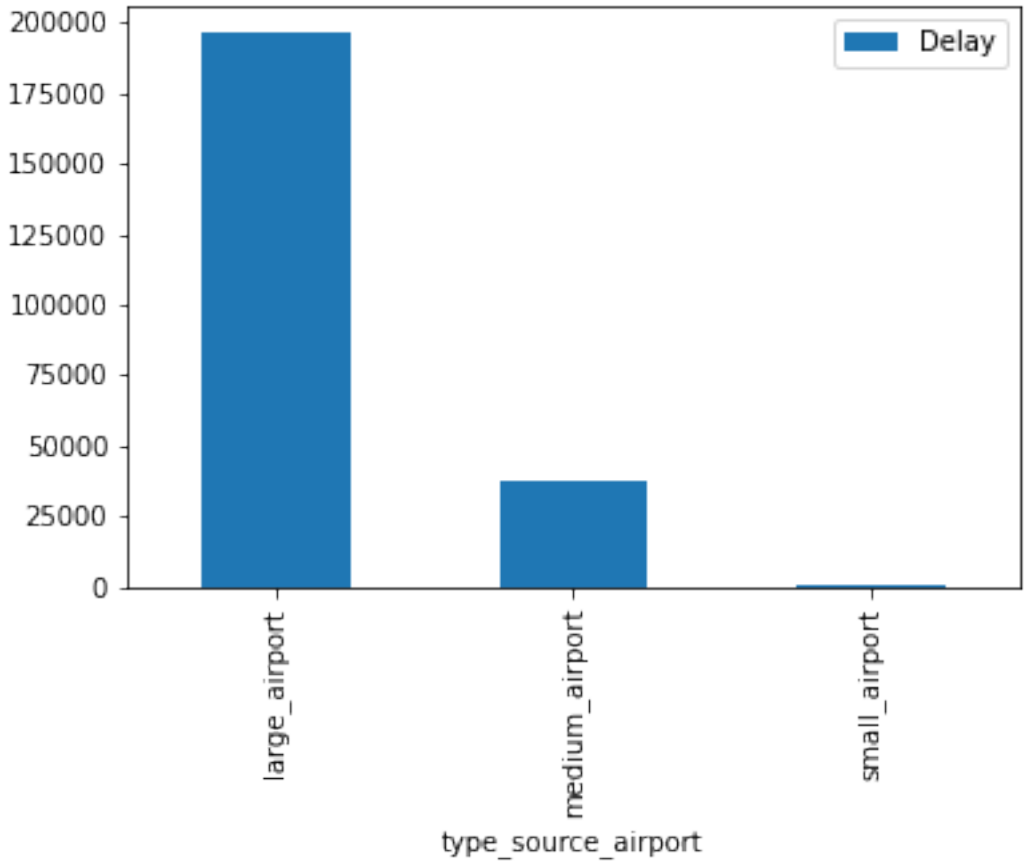
3. How Large Hubs compare to Medium hubs in terms of count of delayed flights. Use appropriate visualization to represent your findings.

```
combined_data_pax.head()
```

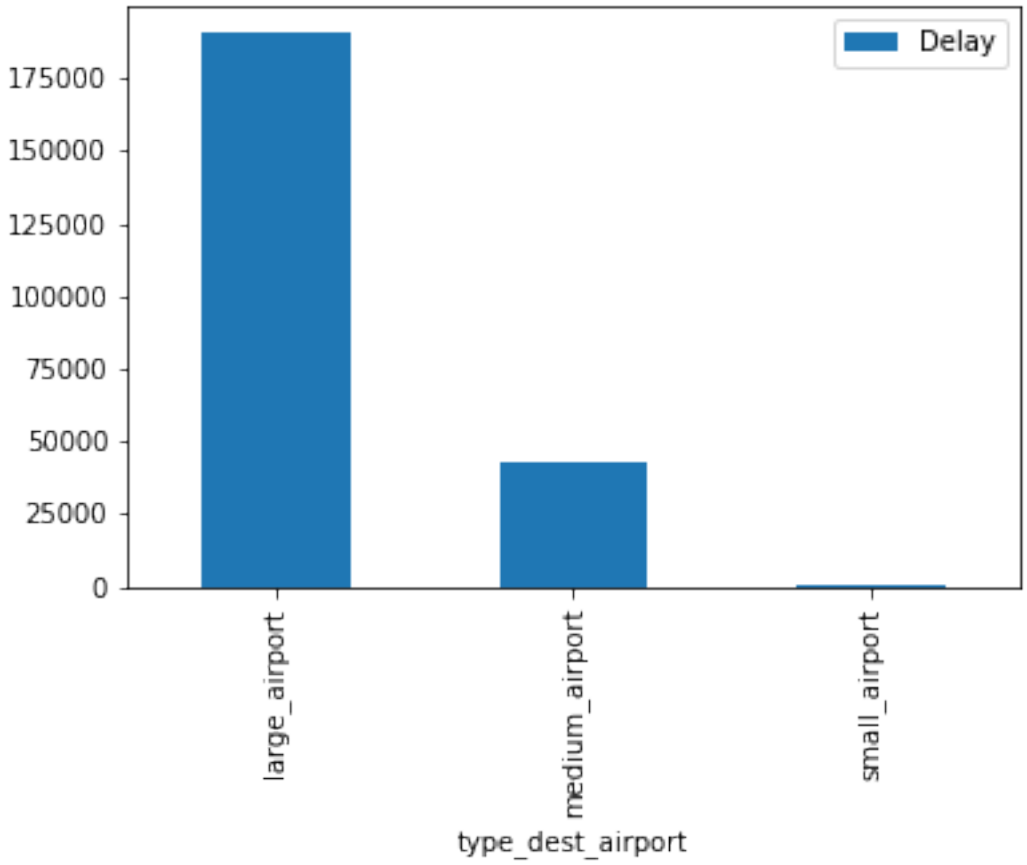
	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
0	1	CO	269	SFO	IAH	3	15	205	1
1	6	CO	1094	LAX	IAH	3	30	181	1
2	11	CO	223	ANC	SEA	3	49	201	

```
1
3  18      CO    1496          LAS       IAH          3    60     162
0
4  20      CO     507          ONT       IAH          3    75     167
0

  type_source_airport  ...  runway_count_source_airport
type_dest_airport  \
0       large_airport  ...                         4.0
large_airport
1       large_airport  ...                         4.0
large_airport
2       large_airport  ...                         3.0
large_airport
3       large_airport  ...                         4.0
large_airport
4       large_airport  ...                         2.0
large_airport

  elevation_ft_dest_airport  runway_count_dest_airport  iata_code_x  \
0                     97.0                         5.0          SFO
1                     97.0                         5.0          LAX
2                    433.0                         4.0          ANC
3                     97.0                         5.0          LAS
4                     97.0                         5.0          ONT

  data_2019_source_airport  iata_code_y data_2019_dest_airport
Founded  \
0               27779230.0          IAH             21905309.0
1934.0
1               42939104.0          IAH             21905309.0
1934.0
2                2713843.0          SEA             25001762.0
1934.0
3               24728361.0          IAH             21905309.0
1934.0
4                2723002.0          IAH             21905309.0
1934.0

   duration
0     short
1     short
2     short
3     short
4     short

[5 rows x 21 columns]

combined_data_pax.groupby('type_source_airport')
[['Delay']].agg('sum').plot.bar()
```

```
<AxesSubplot: xlabel='type_source_airport'>
```




```
combined_data_pax.groupby('type_dest_airport')
[['Delay']].agg('sum').plot.bar()

<AxesSubplot: xlabel='type_dest_airport'>
```



1. Perform hypothesis testing techniques to learn:

   I. Has the altitude of the airport anything to do with flight delays. Check for incoming and outgoing flights

   II. Has surface-type of runways of airports anything to do with flight delays

   III. Has length, duration of flight, anything to do with flight delays

I. Has the altitude of the airport anything to do with flight delays. Check for incoming and outgoing flights

2 sample t test

for outgoing

```
sample1 = combined_data_pax[combined_data_pax.Delay ==
1].elevation_ft_source_airport
sample2 = combined_data_pax[combined_data_pax.Delay ==
0].elevation_ft_source_airport

t, p = stats.ttest_ind(sample1, sample2)

if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'

result
```

```
'reject null'
```

for incoming flights

```
sample1 = combined_data_pax[combined_data_pax.Delay ==
1].elevation_ft_dest_airport
sample2 = combined_data_pax[combined_data_pax.Delay ==
0].elevation_ft_dest_airport

t, p = stats.ttest_ind(sample1, sample2)

if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'

result
```

```
'reject null'
```

Conclusion : Significant difference in avg elevation wrt flight delay for both incoming and outgoing flights

is no. of runway at airport for delayed < for non delayed

```
combined_data_pax
```

```
             id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time
Length  \
0             1      CO     269         SFO       IAH          3    15
205
1             6      CO    1094         LAX       IAH          3    30
181
2            11      CO     223         ANC       SEA          3    49
201
3            18      CO    1496         LAS       IAH          3    60
162
4            20      CO     507         ONT       IAH          3    75
167
...         ...     ...     ...         ...       ...        ...   ...
...
518551  538750      WN    2601         LAS       SMF          5  1230
85
518552  538783      WN    1936         SMF       SAN          5  1235
85
518553  538810      WN    2629         LAS       RNO          5  1240
75
518554  538833      WN    1226         SFO       LAX          5  1245
75
518555  538834      WN    2370         LAX       SFO          5  1245
75

        Delay type_source_airport  ...  runway_count_source_airport  \
0           1       large_airport  ...                          4.0
1           1       large_airport  ...                          4.0
2           1       large_airport  ...                          3.0
3           0       large_airport  ...                          4.0
4           0       large_airport  ...                          2.0
...       ...                 ...  ...                          ...
518551      1       large_airport  ...                          4.0
518552      1       large_airport  ...                          2.0
518553      1       large_airport  ...                          4.0
518554      1       large_airport  ...                          4.0
518555      1       large_airport  ...                          4.0

        type_dest_airport elevation_ft_dest_airport  \
0           large_airport                      97.0
1           large_airport                      97.0
2           large_airport                     433.0
```

```
3           large_airport                       97.0
4           large_airport                       97.0
...                   ...                        ...
518551      large_airport                       27.0
518552      large_airport                       17.0
518553      large_airport                     4415.0
518554      large_airport                      125.0
518555      large_airport                       13.0

        runway_count_dest_airport  iata_code_x
data_2019_source_airport  \
0                             5.0          SFO
27779230.0
1                             5.0          LAX
42939104.0
2                             4.0          ANC
2713843.0
3                             5.0          LAS
24728361.0
4                             5.0          ONT
2723002.0
...                           ...          ...                          ..
.
518551                        2.0          LAS
24728361.0
518552                        1.0          SMF
6454413.0
518553                        3.0          LAS
24728361.0
518554                        4.0          SFO
27779230.0
518555                        4.0          LAX
42939104.0

        iata_code_y data_2019_dest_airport  Founded  duration
0               IAH             21905309.0   1934.0     short
1               IAH             21905309.0   1934.0     short
2               SEA             25001762.0   1934.0     short
3               IAH             21905309.0   1934.0     short
4               IAH             21905309.0   1934.0     short
...             ...                    ...      ...       ...
518551          SMF              6454413.0   1967.0     short
518552          SAN             12648692.0   1967.0     short
518553          RNO              2162250.0   1967.0     short
518554          LAX             42939104.0   1967.0     short
518555          SFO             27779230.0   1967.0     short

[518556 rows x 21 columns]
```

```
s1 = combined_data_pax[combined_data_pax.Delay ==
1].runway_count_source_airport
s2 = combined_data_pax[combined_data_pax.Delay ==
0].runway_count_source_airport

t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

```
s1 = combined_data_pax[combined_data_pax.Delay ==
1].runway_count_dest_airport
s2 = combined_data_pax[combined_data_pax.Delay ==
0].runway_count_dest_airport

t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

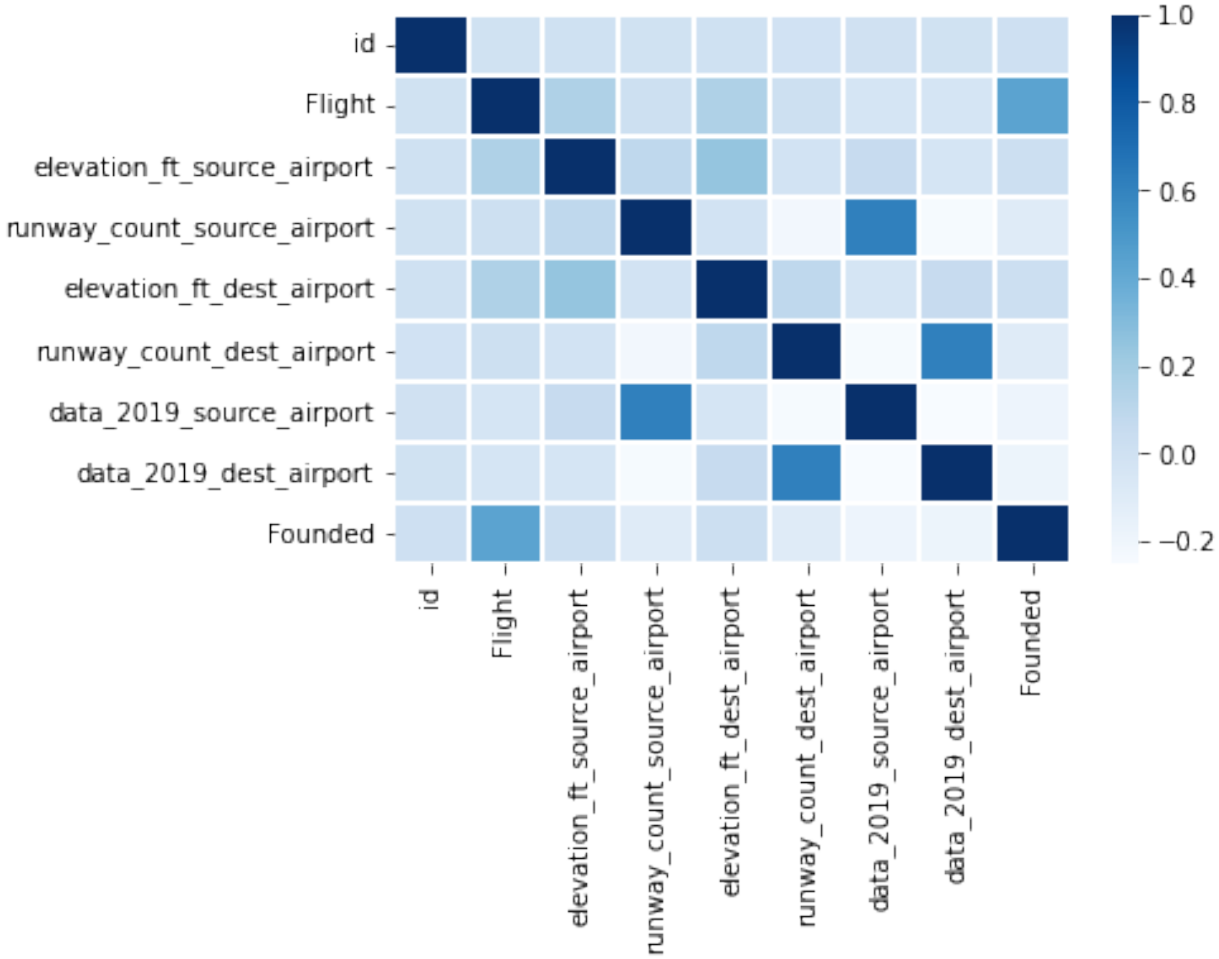
```
reject null
```

```
combined_data_pax.columns
```

```
Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo',
'DayOfWeek',
       'Time', 'Length', 'Delay', 'type_source_airport',
       'elevation_ft_source_airport', 'runway_count_source_airport',
       'type_dest_airport', 'elevation_ft_dest_airport',
       'runway_count_dest_airport', 'iata_code_x',
'data_2019_source_airport',
       'iata_code_y', 'data_2019_dest_airport', 'Founded',
'duration'],
      dtype='object')
```

```
# Find correlation matrix amongst predictors of flight delay. Create a
heatmap to visualize. Share your findings.

correlation_matix = combined_data_pax.drop(columns = ['DayOfWeek',
'Time', 'Length',
'Delay','type_source_airport','type_dest_airport']).corr()

sns.heatmap(correlation_matix, cmap='Blues',linecolor='white',
linewidths=2)
plt.show()
```

```
/tmp/ipykernel_71/2729550883.py:3: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  correlation_matix = combined_data_pax.drop(columns = ['DayOfWeek',
'Time', 'Length',
'Delay','type_source_airport','type_dest_airport']).corr()
```



Conclusion : avg runway count at destination airport for delayed filghts < avg runway count at destination airport for delayed filghts for Incoming flights

Has length, duration of flight, anything to do with flight delays!

```
s1 = combined_data_pax[combined_data_pax.Delay == 1].Length
s2 = combined_data_pax[combined_data_pax.Delay == 0].Length
```

```
t, p = stats.ttest_ind(s1, s2)

if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

```
# there is isgnificant difference

cs = pd.crosstab(combined_data_pax.duration, combined_data_pax.Delay)
cs
```

```
Delay           0       1
duration
short      255324  204474
medium      28991   29208
long          252     307
```

```
chi, p, df, ex = stats.chi2_contingency(cs)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

```
t, p = stats.ttest_ind(s1, s2)
if p < 0.05:
    result = 'reject null'
else :
    result = 'fail to reject null'
print(result)
```

```
reject null
```

Conclusion : avg duration for delayed filghts and non Delayed flights are significantly different.

```
    - avg duration of flights is less for non delayed flights
    - short duration flights get delayed less.
```

check info of dat

```
combined_data_pax.head(2)
```

```
   id Airline  Flight AirportFrom AirportTo  DayOfWeek  Time  Length
Delay  \
```

```
0   1      CO     269          SFO       IAH          3    15     205
1
1   6      CO    1094          LAX       IAH          3    30     181
1

  type_source_airport  ...  runway_count_source_airport
type_dest_airport  \
0       large_airport  ...                          4.0
large_airport
1       large_airport  ...                          4.0
large_airport

  elevation_ft_dest_airport  runway_count_dest_airport  iata_code_x  \
0                     97.0                         5.0          SFO
1                     97.0                         5.0          LAX

  data_2019_source_airport  iata_code_y data_2019_dest_airport
Founded  \
0               27779230.0          IAH             21905309.0
1934.0
1               42939104.0          IAH             21905309.0
1934.0

   duration
0     short
1     short

[2 rows x 21 columns]



combined_data_pax.columns

Index(['id', 'Airline', 'Flight', 'AirportFrom', 'AirportTo',
'DayOfWeek',
       'Time', 'Length', 'Delay', 'type_source_airport',
       'elevation_ft_source_airport', 'runway_count_source_airport',
       'type_dest_airport', 'elevation_ft_dest_airport',
       'runway_count_dest_airport', 'iata_code_x',
'data_2019_source_airport',
       'iata_code_y', 'data_2019_dest_airport', 'Founded',
'duration'],
      dtype='object')

combined_data_pax.to_csv('combined_data_pax.csv', index=False)
```

6. Use Onehotencoder and Ordinalencoder to deal with categorical variables.

```
combined_data_pax.isna().sum()
```

```
id                                0
Airline                           0
Flight                            0
AirportFrom                       0
AirportTo                         0
DayOfWeek                         0
Time                              0
Length                            0
Delay                             0
type_source_airport               0
elevation_ft_source_airport       0
runway_count_source_airport       0
type_dest_airport                 0
elevation_ft_dest_airport         0
runway_count_dest_airport         0
iata_code_x                   85001
data_2019_source_airport       1173
iata_code_y                   84949
data_2019_dest_airport         1166
Founded                           0
duration                          0
dtype: int64
```

```
combined_data_pax.dropna(inplace = True)

combined_data_pax.drop(columns = ['id', 'Flight',  'duration'],
inplace = True)

combined_data_pax.head(2)
```

```
  Airline AirportFrom AirportTo  DayOfWeek  Time  Length  Delay  \
0      CO         SFO       IAH          3    15     205      1
1      CO         LAX       IAH          3    30     181      1

  type_source_airport  elevation_ft_source_airport  \
0       large_airport                         13.0
1       large_airport                        125.0

   runway_count_source_airport type_dest_airport
elevation_ft_dest_airport  \
0                          4.0     large_airport
97.0
1                          4.0     large_airport
97.0

   runway_count_dest_airport iata_code_x  data_2019_source_airport  \
0                        5.0         SFO                27779230.0
1                        5.0         LAX                42939104.0

  iata_code_y  data_2019_dest_airport  Founded
```

```
0          IAH              21905309.0   1934.0
1          IAH              21905309.0   1934.0
```

```
combined_data_pax.type_dest_airport.unique()
```

```
array(['large_airport', 'medium_airport'], dtype=object)
```

```
ordinal = OrdinalEncoder(categories=[['medium_airport',
'large_airport'],['medium_airport', 'large_airport']])
ordinal.fit(combined_data_pax[['type_source_airport',
'type_dest_airport']])
```

```
OrdinalEncoder(categories=[['medium_airport', 'large_airport'],
                           ['medium_airport', 'large_airport']])
```

```
combined_data_pax[['type_source_airport', 'type_dest_airport']] =
ordinal.transform(combined_data_pax[['type_source_airport',
'type_dest_airport']])
```

```
model_data = combined_data_pax.drop(columns = ['Airline',
'AirportFrom', 'AirportTo'])
```

```
model_data.shape
```

```
(349772, 15)
```

```
dummy = pd.get_dummies(model_data)
dummy.shape
```

```
(349772, 141)
```

```
airlines.shape
```

```
(518556, 9)
```

```
dummy.Founded = 2022 - dummy.Founded
```

```
dummy.head(2)
```

```
   DayOfWeek  Time  Length  Delay  type_source_airport  \
0          3    15     205      1                  1.0
1          3    30     181      1                  1.0

   elevation_ft_source_airport  runway_count_source_airport  \
0                         13.0                          4.0
1                        125.0                          4.0

   type_dest_airport  elevation_ft_dest_airport
runway_count_dest_airport  \
0                1.0                       97.0
5.0
1                1.0                       97.0
5.0
```

```
   ...  iata_code_y_SAT  iata_code_y_SEA  iata_code_y_SFO
iata_code_y_SJC  \
0  ...                0                0                0
0
1  ...                0                0                0
0

   iata_code_y_SJU  iata_code_y_SLC  iata_code_y_SMF  iata_code_y_SNA
\
0                0                0                0                0

1                0                0                0                0


   iata_code_y_STL  iata_code_y_TPA
0                0                0
1                0                0

[2 rows x 141 columns]
```

```
model_data.reset_index(drop = True, inplace = True)

np.random.seed(12)
deploy_idx = np.random.choice(model_data.index, replace = False, size
= 5000)

deploy = model_data.loc[deploy_idx]

X_deploy = deploy.drop(columns = 'Delay')

model_dev = model_data.loc[~model_data.index.isin(deploy.index)]

deploy.reset_index(drop = True, inplace = True)
model_dev.reset_index(drop = True, inplace = True)

dummy.dropna(inplace=True)

X = dummy.drop(columns = 'Delay')
y = dummy.Delay
```

Split data into train and test

Standardise data

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

x_train,x_test,y_train,y_test =
train_test_split(X,y,stratify=y,random_state=0)
```

```
st = StandardScaler()
x_train_std = st.fit_transform(x_train)
x_test_std = st.transform(x_test)
```

Apply logistic regression (use stochastic gradient descent optimizer) and decision tree models

```
from sklearn.linear_model import SGDClassifier

sgdcModel = SGDClassifier()
sgdcModel.fit(x_train_std,y_train)
```

```
SGDClassifier()
```

```
# train score
sgdcModel.score(x_train_std,y_train)
```

```
0.5979476153989837
```

```
# train score
sgdcModel.score(x_test_std,y_test)
```

```
0.5984241162814633
```

Accuracy report

```
from sklearn.metrics import classification_report

y_train_pred_sgd = sgdcModel.predict(x_train_std)
y_test_pred_sgd = sgdcModel.predict(x_test_std)

print(classification_report(y_train,y_train_pred_sgd))

print(classification_report(y_test,y_test_pred_sgd))
```

```
              precision    recall  f1-score   support

           0       0.60      0.67      0.64    137227
           1       0.59      0.52      0.55    125102

    accuracy                           0.60    262329
   macro avg       0.60      0.59      0.59    262329
weighted avg       0.60      0.60      0.60    262329

              precision    recall  f1-score   support

           0       0.60      0.67      0.64     45743
           1       0.59      0.52      0.55     41700

    accuracy                           0.60     87443
   macro avg       0.60      0.59      0.59     87443
```

```
weighted avg       0.60      0.60      0.60     87443
```

Decision Tree Model

```
from sklearn.tree import DecisionTreeClassifier

dtModel = DecisionTreeClassifier()

dtModel.fit(x_train,y_train)

# train Score
dtModel.score(x_train,y_train)

#test score
dtModel.score(x_test,y_test)

y_train_pred_dt = dtModel.predict(x_train)
y_test_pred_dt = dtModel.predict(x_test)

print(classification_report(y_train,y_train_pred_dt))

print(classification_report(y_test,y_test_pred_dt))
```

```
              precision    recall  f1-score   support

           0       0.78      0.93      0.85    137227
           1       0.90      0.71      0.79    125102

    accuracy                           0.82    262329
   macro avg       0.84      0.82      0.82    262329
weighted avg       0.84      0.82      0.82    262329

              precision    recall  f1-score   support

           0       0.61      0.70      0.65     45743
           1       0.60      0.50      0.55     41700

    accuracy                           0.60     87443
   macro avg       0.60      0.60      0.60     87443
weighted avg       0.60      0.60      0.60     87443
```

Decision tree is overfitted

```
dtModel =
DecisionTreeClassifier(min_samples_split=12,min_samples_leaf=12) # try
different values

dtModel.fit(x_train,y_train)
```

```
# train Score
dtModel.score(x_train,y_train)

#test score
dtModel.score(x_test,y_test)

y_train_pred_dt = dtModel.predict(x_train)
y_test_pred_dt = dtModel.predict(x_test)

print(classification_report(y_train,y_train_pred_dt))

print(classification_report(y_test,y_test_pred_dt))
```

```
              precision    recall  f1-score   support

           0       0.72      0.79      0.75    137227
           1       0.74      0.67      0.70    125102

    accuracy                           0.73    262329
   macro avg       0.73      0.73      0.73    262329
weighted avg       0.73      0.73      0.73    262329

              precision    recall  f1-score   support

           0       0.65      0.70      0.67     45743
           1       0.64      0.58      0.61     41700

    accuracy                           0.64     87443
   macro avg       0.64      0.64      0.64     87443
weighted avg       0.64      0.64      0.64     87443
```

```
from sklearn.ensemble import GradientBoostingRegressor

gbmodel = GradientBoostingRegressor()
gbmodel.fit(x_train,y_train)
```

```
GradientBoostingRegressor()
```

```
print(gbmodel.feature_importances_)
```

```
[1.56380565e-02 2.99486898e-01 4.01268392e-02 0.00000000e+00
 1.84954034e-02 2.41557682e-02 0.00000000e+00 8.72237069e-03
 3.77336530e-03 1.85768628e-02 5.03400046e-03 4.02016537e-01
 0.00000000e+00 0.00000000e+00 1.15823037e-04 0.00000000e+00
 0.00000000e+00 0.00000000e+00 0.00000000e+00 6.67197363e-04
 0.00000000e+00 1.99087674e-03 0.00000000e+00 7.67567565e-05
 1.23645493e-02 0.00000000e+00 0.00000000e+00 1.51649888e-03
 9.69921881e-03 0.00000000e+00 0.00000000e+00 0.00000000e+00
 0.00000000e+00 0.00000000e+00 8.81862351e-04 3.52066492e-04
 0.00000000e+00 1.67387780e-04 0.00000000e+00 0.00000000e+00
 0.00000000e+00 1.90580081e-03 0.00000000e+00 2.95469193e-03
```

```
 0.00000000e+00 0.00000000e+00 2.26044321e-02 0.00000000e+00
 0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00
 9.13229642e-03 1.45307761e-03 0.00000000e+00 0.00000000e+00
 3.53111876e-03 0.00000000e+00 0.00000000e+00 7.89199788e-03
 7.77901365e-03 0.00000000e+00 0.00000000e+00 0.00000000e+00
 3.56256481e-04 0.00000000e+00 0.00000000e+00 0.00000000e+00
 2.19017613e-03 0.00000000e+00 0.00000000e+00 3.90277224e-04
 2.28432622e-03 0.00000000e+00 2.01841733e-04 6.27928865e-04
 0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00
 0.00000000e+00 0.00000000e+00 0.00000000e+00 2.71731452e-03
 1.69927741e-04 0.00000000e+00 0.00000000e+00 0.00000000e+00
 2.43248248e-02 0.00000000e+00 0.00000000e+00 0.00000000e+00
 1.41493662e-02 0.00000000e+00 3.34470459e-04 0.00000000e+00
 2.52772723e-04 0.00000000e+00 0.00000000e+00 3.49299440e-04
 0.00000000e+00 1.00038825e-03 0.00000000e+00 0.00000000e+00
 8.28737601e-04 1.26496360e-03 0.00000000e+00 1.07612387e-03
 0.00000000e+00 3.71755337e-04 2.43086007e-03 0.00000000e+00
 2.13814967e-04 0.00000000e+00 0.00000000e+00 0.00000000e+00
 0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00
 0.00000000e+00 0.00000000e+00 3.98802386e-04 1.13619433e-02
 4.67436456e-03 1.55931768e-04 0.00000000e+00 0.00000000e+00
 0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00
 4.03021317e-03 0.00000000e+00 7.79917181e-04 1.72524762e-03
 0.00000000e+00 0.00000000e+00 2.27385338e-04 0.00000000e+00]
```

```
pd.DataFrame({'Features':gbmodel.feature_names_in_,'Importance':gbmode
l.feature_importances_}).sort_values("Importance",ascending=False)
```

```
                        Features  Importance
11                       Founded    0.402017
1                           Time    0.299487
2                         Length    0.040127
88               iata_code_y_CLT    0.024325
5    runway_count_source_airport    0.024156
..                           ...         ...
55               iata_code_x_ONT    0.000000
54               iata_code_x_OMA    0.000000
51               iata_code_x_MSY    0.000000
50               iata_code_x_MSP    0.000000
139              iata_code_y_TPA    0.000000

[140 rows x 2 columns]
```

```
gbmodel.score(x_train,y_train)
```

```
0.10945666158862555
```

```
gbmodel.score(x_test,y_test)
```

```
0.10417735013457807
```