In [1]:
```
from PIL import Image
airplane = Image.open('/Users/saileshkumarm/Downloads/US_Airways-Logo.jpg')
airplane
```

Out[1]:

U·S AIRWAYS

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
```

In [3]:
```
%matplotlib inline
warnings.filterwarnings('ignore')
```

In [4]:
```
airline = pd.read_excel('/Users/saileshkumarm/Downloads/United_States_Airlines_Analysis_Capstone_2/Airlin
```

In [5]:
```
airline.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 518556 entries, 0 to 518555
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id           518556 non-null  int64
 1   Airline      518556 non-null  object
 2   Flight       518556 non-null  int64
 3   AirportFrom  518556 non-null  object
 4   AirportTo    518556 non-null  object
 5   DayOfWeek    518556 non-null  int64
 6   Time         518556 non-null  int64
 7   Length       518556 non-null  int64
 8   Delay        518556 non-null  int64
dtypes: int64(6), object(3)
memory usage: 35.6+ MB
```

In [6]:
```
airports = pd.read_excel('/Users/saileshkumarm/Downloads/United_States_Airlines_Analysis_Capstone_2/airpo
```

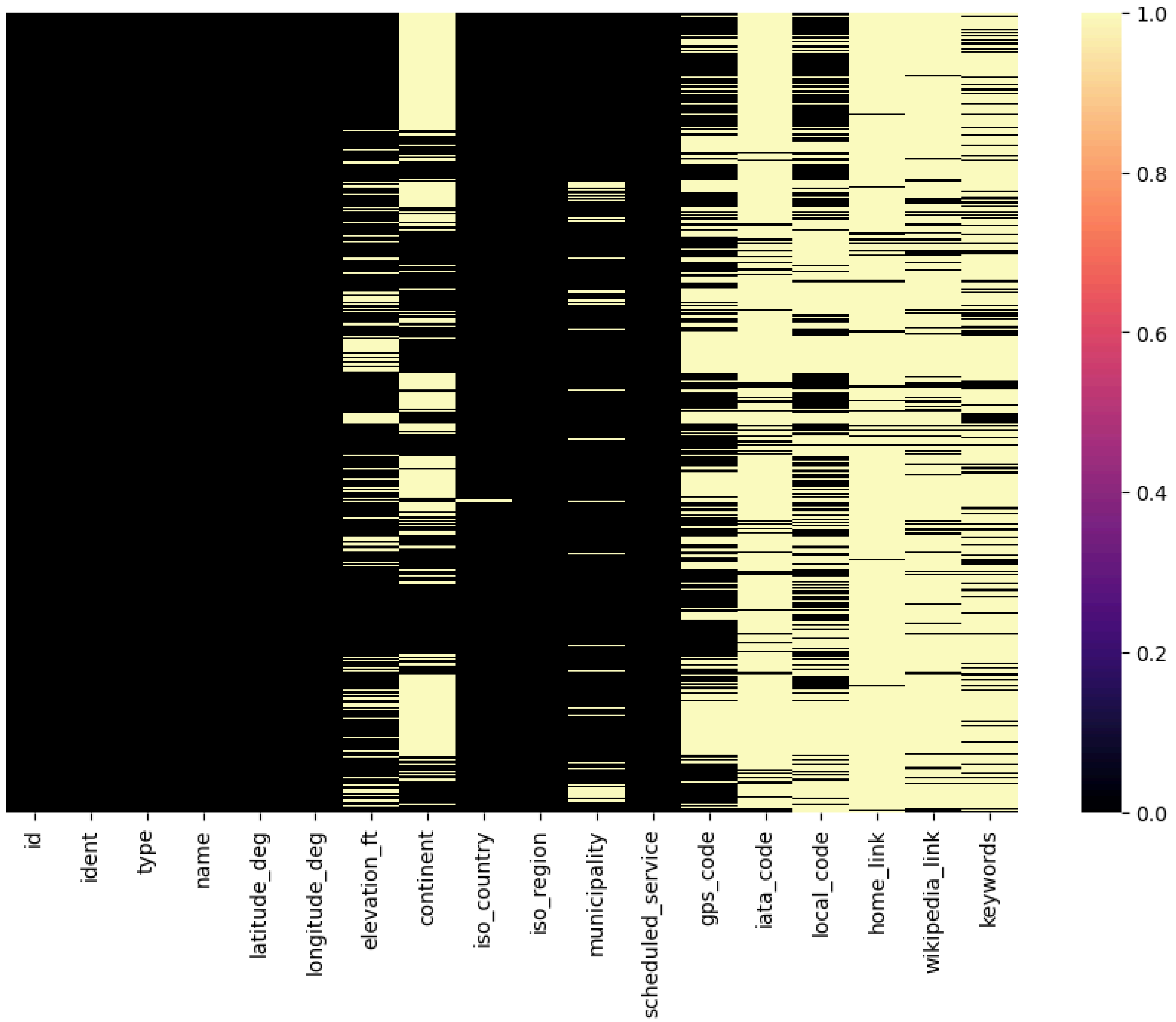
In [7]:
```
#fig, ax = plt.subplots(figsize=(11, 7))
#sns.heatmap(df_airline.isnull(),yticklabels=False,cbar=False,cmap = 'viridis')
```

In [7]:
```
airports.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73805 entries, 0 to 73804
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 73805 non-null  int64  
 1   ident              73805 non-null  object 
 2   type               73805 non-null  object 
 3   name               73805 non-null  object 
 4   latitude_deg       73805 non-null  float64
 5   longitude_deg      73805 non-null  float64
 6   elevation_ft       59683 non-null  float64
 7   continent          38086 non-null  object 
 8   iso_country        73546 non-null  object 
 9   iso_region         73805 non-null  object 
 10  municipality       68739 non-null  object 
 11  scheduled_service  73805 non-null  object 
 12  gps_code           42996 non-null  object 
 13  iata_code          9160 non-null   object 
 14  local_code         32975 non-null  object 
 15  home_link          3492 non-null   object 
 16  wikipedia_link     10705 non-null  object 
 17  keywords           13951 non-null  object 
dtypes: float64(3), int64(1), object(14)
memory usage: 10.1+ MB
```

In [8]:
```
fig, ax = plt.subplots(figsize=(11, 7))
sns.heatmap(airports.isnull(),yticklabels=False,cbar=True,cmap = 'magma')
```

Out[8]:
```
<Axes: >
```



In [9]:
```
runways = pd.read_excel('/Users/saileshkumarm/Downloads/United_States_Airlines_Analysis_Capstone_2/runway
```

In [10]:
```
runways.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43977 entries, 0 to 43976
Data columns (total 20 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   id                         43977 non-null  int64  
 1   airport_ref                43977 non-null  int64  
 2   airport_ident              43977 non-null  object 
 3   length_ft                  43753 non-null  float64
 4   width_ft                   41088 non-null  float64
 5   surface                    43518 non-null  object 
 6   lighted                    43977 non-null  int64  
 7   closed                     43977 non-null  int64  
 8   le_ident                   43793 non-null  object 
 9   le_latitude_deg            15016 non-null  float64
 10  le_longitude_deg           15000 non-null  float64
 11  le_elevation_ft            12781 non-null  float64
 12  le_heading_degT            14624 non-null  float64
 13  le_displaced_threshold_ft  2883 non-null   float64
 14  he_ident                   37332 non-null  object 
 15  he_latitude_deg            14971 non-null  float64
 16  he_longitude_deg           14973 non-null  float64
 17  he_elevation_ft            12620 non-null  float64
 18  he_heading_degT            16428 non-null  float64
 19  he_displaced_threshold_ft  3176 non-null   float64
dtypes: float64(12), int64(4), object(4)
memory usage: 6.7+ MB
```

In [11]:
```
##Drop the columns that will not play an important role in the model building
```

In [12]:
```
#Remove the feature from the airpot data that is not useful
```

In [13]:
```
airports.columns
```

Out[13]:
```
Index(['id', 'ident', 'type', 'name', 'latitude_deg', 'longitude_deg',
       'elevation_ft', 'continent', 'iso_country', 'iso_region',
       'municipality', 'scheduled_service', 'gps_code', 'iata_code',
       'local_code', 'home_link', 'wikipedia_link', 'keywords'],
      dtype='object')
```

In [14]:
```
airports.drop(['continent', 'iso_country', 'iso_region','municipality','gps_code','local_code', 'home_lin
'wikipedia_link', 'keywords'], axis=1, inplace=True)
airports
```

Out[14]:

	id	ident	type	name	latitude_deg	longitude_deg	elevation_ft	scheduled_service	iata_code
0	6523	00A	heliport	Total Rf Heliport	40.070801	-74.933601	11.0	no	NaN
1	323361	00AA	small_airport	Aero B Ranch Airport	38.704022	-101.473911	3435.0	no	NaN
2	6524	00AK	small_airport	Lowell Field	59.947733	-151.692524	450.0	no	NaN
3	6525	00AL	small_airport	Epps Airpark	34.864799	-86.770302	820.0	no	NaN
4	6526	00AR	closed	Newport Hospital & Clinic Heliport	35.608700	-91.254898	237.0	no	NaN
...	...	...	...	...	...	...	...	...	...
73800	46378	ZZ-0001	heliport	Sealand Helipad	51.894444	1.482500	40.0	no	NaN
73801	307326	ZZ-0002	small_airport	Glorioso Islands Airstrip	-11.584278	47.296389	11.0	no	NaN
73802	346788	ZZ-0003	small_airport	Fainting Goat Airport	32.110587	-97.356312	690.0	no	NaN
73803	342102	ZZZW	closed	Scandium City Heliport	69.355287	-138.939310	4.0	no	ZYW
73804	313629	ZZZZ	small_airport	Satsuma I≈çjima Airport	30.784722	130.270556	338.0	no	NaN

73805 rows × 9 columns

In [15]:
```
runways.drop(['le_ident', 'le_latitude_deg','le_longitude_deg','le_elevation_ft', 'le_heading_degT',
'le_displaced_threshold_ft', 'he_ident','he_latitude_deg','he_longitude_deg', 'he_elevation_ft', 'he_head
'he_displaced_threshold_ft'], axis = 1,inplace=True)
runways
```

Out[15]:

	id	airport_ref	airport_ident	length_ft	width_ft	surface	lighted	closed
0	269408	6523	00A	80.0	80.0	ASPH-G	1	0
1	255155	6524	00AK	2500.0	70.0	GRVL	0	0
2	254165	6525	00AL	2300.0	200.0	TURF	0	0
3	270932	6526	00AR	40.0	40.0	GRASS	0	0
4	322128	322127	00AS	1450.0	60.0	Turf	0	0
...	...	...	...	...	...	...	...	...
43972	235186	27243	ZYTX	10499.0	148.0	CON	1	0
43973	235169	27244	ZYYJ	8530.0	148.0	CON	1	0
43974	354997	317861	ZYYK	8202.0	NaN	NaN	0	0
43975	346789	346788	ZZ-0003	1800.0	15.0	Turf	0	0
43976	313663	313629	ZZZZ	1713.0	82.0	concrete	0	0

43977 rows × 8 columns

In [16]:
```
#Merge the runways and airport data
```

In [17]:
```
merged_df = pd.merge(runways,airports, left_on='airport_ident', right_on='ident')
merged_df
```

Out[17]:

	id_x	airport_ref	airport_ident	length_ft	width_ft	surface	lighted	closed	id_y	ident	type	name	latitude_d
0	269408	6523	00A	80.0	80.0	ASPH-G	1	0	6523	00A	heliport	Total Rf Heliport	40.0708
1	255155	6524	00AK	2500.0	70.0	GRVL	0	0	6524	00AK	small_airport	Lowell Field	59.9477
2	254165	6525	00AL	2300.0	200.0	TURF	0	0	6525	00AL	small_airport	Epps Airpark	34.8647
3	270932	6526	00AR	40.0	40.0	GRASS	0	0	6526	00AR	closed	Newport Hospital & Clinic Heliport	35.6087
4	322128	322127	00AS	1450.0	60.0	Turf	0	0	322127	00AS	small_airport	Fulton Airport	34.9428
...	...	...	...	...	...	...	...	...	...	...	...	...	
43972	235186	27243	ZYTX	10499.0	148.0	CON	1	0	27243	ZYTX	large_airport	Shenyang Taoxian International Airport	41.6398
43973	235169	27244	ZYYJ	8530.0	148.0	CON	1	0	27244	ZYYJ	medium_airport	Yanji Chaoyangchuan Airport	42.8828
43974	354997	317861	ZYYK	8202.0	NaN	NaN	0	0	317861	ZYYK	medium_airport	Yingkou Lanqi Airport	40.5425
43975	346789	346788	ZZ-0003	1800.0	15.0	Turf	0	0	346788	ZZ-0003	small_airport	Fainting Goat Airport	32.1105
43976	313663	313629	ZZZZ	1713.0	82.0	concrete	0	0	313629	ZZZZ	small_airport	Satsuma I≈çjima Airport	30.7847

43977 rows × 17 columns

In [18]:
```
merged_df.drop(['id_x','id_y'],axis=1,inplace=True)
```

In [19]:
```
merged_df
```

Out[19]:

	airport_ref	airport_ident	length_ft	width_ft	surface	lighted	closed	ident	type	name	latitude_deg	longitude_deg
0	6523	00A	80.0	80.0	ASPH-G	1	0	00A	heliport	Total Rf Heliport	40.070801	-74.93360
1	6524	00AK	2500.0	70.0	GRVL	0	0	00AK	small_airport	Lowell Field	59.947733	-151.69252
2	6525	00AL	2300.0	200.0	TURF	0	0	00AL	small_airport	Epps Airpark	34.864799	-86.77030
3	6526	00AR	40.0	40.0	GRASS	0	0	00AR	closed	Newport Hospital & Clinic Heliport	35.608700	-91.25489
4	322127	00AS	1450.0	60.0	Turf	0	0	00AS	small_airport	Fulton Airport	34.942803	-97.81801
...	...	...	...	...	...	...	...	...	...	...	...	..
43972	27243	ZYTX	10499.0	148.0	CON	1	0	ZYTX	large_airport	Shenyang Taoxian International Airport	41.639801	123.48300
43973	27244	ZYYJ	8530.0	148.0	CON	1	0	ZYYJ	medium_airport	Yanji Chaoyangchuan Airport	42.882801	129.45100
43974	317861	ZYYK	8202.0	NaN	NaN	0	0	ZYYK	medium_airport	Yingkou Lanqi Airport	40.542524	122.35860
43975	346788	ZZ-0003	1800.0	15.0	Turf	0	0	ZZ-0003	small_airport	Fainting Goat Airport	32.110587	-97.35631
43976	313629	ZZZZ	1713.0	82.0	concrete	0	0	ZZZZ	small_airport	Satsuma I≈çjima Airport	30.784722	130.27055

43977 rows × 15 columns

In [20]:
```
final_df = pd.merge(airline, merged_df, left_on='AirportFrom', right_on='iata_code',how='inner')
final_df
```

Out[20]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	airport_ref	...	lighted	closed	ident	type
0	1	CO	269	SFO	IAH	3	15	205	1	3878	...	1	0	KSFO	large_airport
1	1	CO	269	SFO	IAH	3	15	205	1	3878	...	1	0	KSFO	large_airport
2	1	CO	269	SFO	IAH	3	15	205	1	3878	...	1	0	KSFO	large_airport
3	1	CO	269	SFO	IAH	3	15	205	1	3878	...	1	0	KSFO	large_airport
4	4	AA	2466	SFO	DFW	3	20	195	1	3878	...	1	0	KSFO	large_airport
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2160271	488365	CO	2	GUM	HNL	3	400	430	0	5433	...	1	0	PGUM	large_airport
2160272	506855	CO	2	GUM	HNL	4	400	430	1	5433	...	1	0	PGUM	large_airport
2160273	506855	CO	2	GUM	HNL	4	400	430	1	5433	...	1	0	PGUM	large_airport
2160274	525138	CO	2	GUM	HNL	5	400	430	1	5433	...	1	0	PGUM	large_airport
2160275	525138	CO	2	GUM	HNL	5	400	430	1	5433	...	1	0	PGUM	large_airport

2160276 rows × 24 columns

In [21]:
```
final_df.drop_duplicates(subset=['id'], keep='first', inplace=True)
final_df
```

Out[21]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	airport_ref	...	lighted	closed	ident	type
0	1	CO	269	SFO	IAH	3	15	205	1	3878	...	1	0	KSFO	large_airport
4	4	AA	2466	SFO	DFW	3	20	195	1	3878	...	1	0	KSFO	large_airport
8	9	DL	2606	SFO	MSP	3	35	216	1	3878	...	1	0	KSFO	large_airport
12	129	DL	1580	SFO	DTW	3	345	270	0	3878	...	1	0	KSFO	large_airport
16	150	UA	756	SFO	DEN	3	348	158	0	3878	...	1	0	KSFO	large_airport
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2160266	451344	CO	2	GUM	HNL	1	400	430	1	5433	...	1	0	PGUM	large_airport
2160268	469866	CO	2	GUM	HNL	2	400	430	1	5433	...	1	0	PGUM	large_airport
2160270	488365	CO	2	GUM	HNL	3	400	430	0	5433	...	1	0	PGUM	large_airport
2160272	506855	CO	2	GUM	HNL	4	400	430	1	5433	...	1	0	PGUM	large_airport
2160274	525138	CO	2	GUM	HNL	5	400	430	1	5433	...	1	0	PGUM	large_airport

518525 rows × 24 columns

In [22]:
```
# When it comes to on-time arrivals, different airlines perform differently based on the amount of experi
# they have. The major airlines inthis field include US Airways Express (founded in 1967) Continental Air
# (founded in 1934), and Express Jet (founded in 19860. Pull such information specific to various airline
# the Wikipedia page link given below.https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States
```

In [23]:
```
## Now lets use the web scrapping to import the data frome the wikipedia.
```

In [24]:
```
url = "https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States"
tables = pd.read_html(url)
```

In [25]:
```
print(tables)
```

```
[               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   Southwest Airlines    NaN   WN  SWA        SOUTHWEST
11      Spirit Airlines    NaN   NK  NKS     SPIRIT WINGS
12  Sun Country Airlines   NaN   SY  SCX      SUN COUNTRY
13      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
```

In [26]:
```
tables[0]
```

Out[26]:

	Airline	Image	IATA	ICAO	Callsign	Primary hubs, secondary hubs	Founded	Notes
0	Alaska Airlines	NaN	AS	ASA	ALASKA	Seattle/Tacoma Anchorage Portland (OR) San Fra...	1932	Founded as McGee Airways and commenced operati...
1	Allegiant Air	NaN	G4	AAY	ALLEGIANT	Las Vegas Cincinnati Destin/Ft. Walton Beach I...	1997	Founded as WestJet Express and began operation...
2	American Airlines	NaN	AA	AAL	AMERICAN	Dallas/Fort Worth Charlotte Chicago–O'Hare Mia...	1926	Founded as American Airways and commenced oper...
3	Avelo Airlines	NaN	XP	VXP	AVELO	Burbank New Haven Orlando Hartford Lakeland Ra...	1987	First did business as Casino Express Airlines ...
4	Breeze Airways	NaN	MX	MXY	MOXY	Charleston (SC) Hartford New Orleans Norfolk P...	2018	Founded as Moxy Airways but was renamed due to...
5	Delta Air Lines	NaN	DL	DAL	DELTA	Atlanta Detroit Minneapolis/St. Paul New York–...	1924	Founded as Huff Daland Dusters and commenced o...
6	Eastern Airlines	NaN	2D	EAL	EASTERN	Miami	2010	NaN
7	Frontier Airlines	NaN	F9	FFT	FRONTIER FLIGHT	Denver Atlanta Chicago–O'Hare Cincinnati Cleve...	1994	NaN
8	Hawaiian Airlines	NaN	HA	HAL	HAWAIIAN	Honolulu Kahului	1929	Founded as Inter-Island Airways in early 1929 ...
9	JetBlue	NaN	B6	JBU	JETBLUE	New York–JFK Boston Los Angeles Fort Lauderdal...	1998	Founded as New Air and commenced operations in...
10	Southwest Airlines	NaN	WN	SWA	SOUTHWEST	Dallas–Love Atlanta Baltimore Chicago–Midway D...	1967	Founded as Air Southwest and commenced operati...
11	Spirit Airlines	NaN	NK	NKS	SPIRIT WINGS	Fort Lauderdale Atlantic City Atlanta Detroit ...	1980	Founded as Charter One.
12	Sun Country Airlines	NaN	SY	SCX	SUN COUNTRY	Minneapolis/St. Paul Dallas/Fort Worth Las Vegas	1982	Commenced operations in 1983. Operates some Am...
13	United Airlines	NaN	UA	UAL	UNITED	Chicago–O'Hare Denver Houston–Intercontinental...	1926	Founded as Varney Air Lines and commenced oper...

In [27]:
```
tables[4]
```

Out[27]:

	Airline	Image	IATA	ICAO	Callsign	Primary hubs, secondary hubs	Founded	Notes
0	21 Air	NaN	2I	CSB	CARGO SOUTH	Greensboro	2014.0	NaN
1	ABX Air	NaN	GB	ABX	ABEX	Wilmington (OH) Cincinnati Miami	1980.0	Founded as Airborne Express. Operates some Ama...
2	Air Cargo Carriers	NaN	2Q	SNC	NIGHT CARGO	Milwaukee Cincinnati	1986.0	Commenced operations in 1980.
3	AirNet Express	NaN	NaN	USC	STAR CHECK	Columbus–Rickenbacker	1974.0	Founded as Financial Air Express.
4	Air Transport International	NaN	8C	ATN	AIR TRANSPORT	Wilmington (OH) Cincinnati	1978.0	Founded as US Airways and commenced operations...
5	Alaska Central Express	NaN	KO	AER	ACE AIR	Anchorage	1996.0	NaN
6	Aloha Air Cargo	NaN	KH	AAH	ALOHA	Honolulu	1946.0	Founded as Trans-Pacific Airlines and separate...
7	Alpine Air Express	NaN	5A	AIP	ALPINE AIR	Provo Billings Sioux Falls	1971.0	NaN
8	Amazon Air	NaN	AFW	KAFW	AMAZON AIR	Ft. Worth–Alliance Cincinnati Leipzig/Halle Sa...	2015.0	Formerly Amazon Prime Air
9	Ameriflight	NaN	A8	AMF	AMFLIGHT	Dallas/Ft. Worth Burbank	1968.0	Founded as California Air Charter.
10	Amerijet International	NaN	M6	AJT	AMERIJET	Miami Port of Spain	1974.0	NaN
11	Ameristar Jet Charter	NaN	7Z	AJI	AMERISTAR	Dallas–Addison El Paso Willow Run	2000.0	NaN
12	Asia Pacific Airlines	NaN	P9	MGE	MAGELLAN	Guam Honolulu	1998.0	NaN
13	Atlas Air	NaN	5Y	GTI	GIANT	New York–JFK Anchorage Cincinnati Houston–Inte...	1992.0	Commenced operations in 1993. Operates some Am...
14	Bemidji Airlines	NaN	CH	BMJ	BEMIDJI	Bemidji Minneapolis/St. Paul	1946.0	Commenced operations in 1947.
15	Castle Aviation	NaN	NaN	CSJ	CASTLE	Akron/Canton	1986.0	NaN
16	Corporate Air	NaN	NaN	CPT	AIRSPUR	Billings	1981.0	NaN
17	CSA Air	NaN	NaN	IRO	IRON AIR	Iron Mountain	1998.0	NaN
18	Empire Airlines	NaN	EM	CFS	EMPIRE	Coeur d'Alene Spokane	1977.0	NaN
19	Everts Air Cargo	NaN	5V	VTS	EVERTS	Fairbanks Anchorage	1995.0	NaN
20	FedEx Express	NaN	FX	FDX	FEDEX	Memphis Anchorage Cologne/Bonn Dubai Ft. Worth...	1971.0	Founded as Federal Express and commenced opera...
21	Freight Runners Express	NaN	NaN	FRG	FREIGHT RUNNERS	Milwaukee	1985.0	NaN
22	IFL Group	NaN	IF	IFL	EIFFEL	Waterford Miami	1983.0	Founded as Air Contract Cargo.
23	Kalitta Air	NaN	K4	CKS	CONNIE	Ypsilanti Anchorage Bahrain Cincinnati Hong Ko...	1967.0	Founded as American International Airways.
24	Kalitta Charters	NaN	CB	KFS	KALITTA	Ypsilanti	NaN	NaN
25	Lynden Air Cargo	NaN	L2	LYC	LYNDEN	Anchorage	1995.0	NaN
26	Martinaire	NaN	NaN	MRA	MARTEX	Addison	1978.0	NaN
27	Merlin Airways	NaN	NaN	MEI	AVALON	Billings Miami San Juan	1983.0	NaN
28	Mountain Air Cargo	NaN	C2	MTN	MOUNTAIN	Kinston	1974.0	NaN
29	National Airlines	NaN	N8	NCR	NATIONAL CARGO	Orlando/Sanford	1985.0	Commenced operations in 1986.
30	Northern Air Cargo	NaN	NC	NAC	YUKON	Anchorage Miami	1956.0	NaN
31	Polar Air Cargo	NaN	PO	PAC	POLAR	Anchorage Cincinnati Hong Kong Honolulu Los An...	1993.0	NaN
32	Royal Air Freight	NaN	NaN	RAX	AIR ROYAL	Waterford	1961.0	NaN
33	Ryan Air Services	NaN	7S	RYA	RYAN AIR	Anchorage Aniak Bethel Emmonak Kotzebue Nome S...	1953.0	Founded as Unalakleet Air Taxi.
34	Sky Lease Cargo	NaN	GG	KYE	SKY CUBE	Miami	1969.0	Founded as Wrangler Aviation and commenced ope...
35	Skyway Enterprises	NaN	KI	SKZ	SKYWAY-INC	NaN	1981.0	Commenced operations in 1983.
36	StratAir	NaN	NaN	NaN	NaN	Miami	2018.0	Virtual airline on behalf of Northern Air Cargo.
37	Trans Executive Airlines	NaN	KH	MUI	RHOADES EXPRESS	Honolulu	1982.0	NaN
38	UPS Airlines	NaN	5X	UPS	UPS	Louisville Chicago/Rockford Cologne/Bonn Colum...	1988.0	NaN
39	USA Jet Airlines	NaN	UJ	JUS	JET USA	Ypsilanti Laredo	1994.0	NaN
40	West Air	NaN	NaN	PCM	PAC VALLEY	Las Vegas Oakland Ontario Sacramento San Diego	1988.0	NaN
41	Western Global Airlines	NaN	KD	WGN	WESTERN GLOBAL	Fort Myers Anchorage Hong Kong Liege Los Angel...	2013.0	NaN
42	Wiggins Airways	NaN	WG	WIG	WIGGINS AIRWAYS	Manchester	1929.0	NaN

In [28]:
```
tables[6]
```

Out[28]:

	Airline	Image	IATA	ICAO	Callsign	Primary hubs, secondary hubs	Founded	Notes
0	Comco	NaN	NaN	NaN	NaN	NaN	2002	NaN
1	Janet	NaN	NaN	WWW	JANET	Las Vegas	1972	NaN
2	Justice Prisoner and Alien Transportation System	NaN	NaN	JUD	JUSTICE	Oklahoma City	1980	Commenced operations in 1995.

In [29]:
```
# First merge all wikipedia table.
tables = [tables[0],tables[1],tables[2],tables[3],tables[4],tables[5],tables[6]]
```

In [30]:
```
wiki_tables = pd.concat(tables, ignore_index=True)
```

In [31]:
```
wiki_tables
```

Out[31]:

	Airline	Image	IATA	ICAO	Callsign	Primary hubs, secondary hubs	Founded	Notes
0	Alaska Airlines	NaN	AS	ASA	ALASKA	Seattle/Tacoma Anchorage Portland (OR) San Fra...	1932.0	Founded as McGee Airways and commenced operati...
1	Allegiant Air	NaN	G4	AAY	ALLEGIANT	Las Vegas Cincinnati Destin/Ft. Walton Beach I...	1997.0	Founded as WestJet Express and began operation...
2	American Airlines	NaN	AA	AAL	AMERICAN	Dallas/Fort Worth Charlotte Chicago–O'Hare Mia...	1926.0	Founded as American Airways and commenced oper...
3	Avelo Airlines	NaN	XP	VXP	AVELO	Burbank New Haven Orlando Hartford Lakeland Ra...	1987.0	First did business as Casino Express Airlines ...
4	Breeze Airways	NaN	MX	MXY	MOXY	Charleston (SC) Hartford New Orleans Norfolk P...	2018.0	Founded as Moxy Airways but was renamed due to...
...	...	...	...	...	...	...	...	...
135	Lifestar	NaN	NaN	NaN	NaN	NaN	NaN	NaN
136	Life Lion	NaN	NaN	NaN	NaN	NaN	NaN	NaN
137	Comco	NaN	NaN	NaN	NaN	NaN	2002.0	NaN
138	Janet	NaN	NaN	WWW	JANET	Las Vegas	1972.0	NaN
139	Justice Prisoner and Alien Transportation System	NaN	NaN	JUD	JUSTICE	Oklahoma City	1980.0	Commenced operations in 1995.

140 rows × 8 columns

In [32]:
```
#  Extract columns from wikipedia table that we need to merge
```

In [33]:
```
df_wiki = wiki_tables[['IATA', "Founded"]]
```

In [34]:
```
df_wiki
```

Out[34]:

	IATA	Founded
0	AS	1932.0
1	G4	1997.0
2	AA	1926.0
3	XP	1987.0
4	MX	2018.0
...	...	...
135	NaN	NaN
136	NaN	NaN
137	NaN	2002.0
138	NaN	1972.0
139	NaN	1980.0

140 rows × 2 columns

In [35]:
```
# Now we gather all the information that we got from wiki pedia link and the data that we have.
new_df = final_df.merge(df_wiki, left_on ='Airline', right_on = "IATA")
new_df
```

Out[35]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	airport_ref	...	ident	type	name	la
0	4	AA	2466	SFO	DFW	3	20	195	1	3878	...	KSFO	large_airport	San Francisco International Airport	
1	231	AA	526	SFO	DFW	3	360	215	0	3878	...	KSFO	large_airport	San Francisco International Airport	
2	234	AA	552	SFO	MIA	3	360	315	1	3878	...	KSFO	large_airport	San Francisco International Airport	
3	905	AA	810	SFO	ORD	3	385	255	0	3878	...	KSFO	large_airport	San Francisco International Airport	
4	1739	AA	24	SFO	JFK	3	425	325	1	3878	...	KSFO	large_airport	San Francisco International Airport	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
434919	497838	9E	4292	LWB	JFK	3	890	110	1	20390	...	KLWB	medium_airport	Greenbrier Valley Airport	
434920	516333	9E	4292	LWB	JFK	4	890	110	0	20390	...	KLWB	medium_airport	Greenbrier Valley Airport	
434921	534123	9E	4292	LWB	JFK	5	890	110	0	20390	...	KLWB	medium_airport	Greenbrier Valley Airport	
434922	69058	9E	3752	ABR	MSP	7	410	76	1	3358	...	KABR	medium_airport	Aberdeen Regional Airport	
434923	189396	9E	3752	ABR	MSP	7	410	76	0	3358	...	KABR	medium_airport	Aberdeen Regional Airport	

434924 rows × 26 columns

In [36]:
```
# The total passenger traffic may also contribute to flight delays. The term hub
# refers to busy commercial airports. Large hubs are airports that account for at
# least 1 percent of the total passenger enplanements in the United States. Airports that account for
#0.25 percent to 1 percent of total passenger enplanements
# are considered medium hubs. Pull passenger traffic data from the Wikipedia
# page given below using web scraping and collate it in a table.
```

In [37]:
```
url2 = "https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States"
table2 = pd.read_html(url2)
table2
```

Out[37]:
```
[    0                                                 1
 0 NaN  Graphs are unavailable due to technical issues...,
     Rank (2023)                                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                Los Angeles International Airport       LAX
 4             5                     O'Hare International Airport       ORD
 5             6             John F. Kennedy International Airport       JFK
 6             7                    Orlando International Airport       MCO
 7             8                 Harry Reid International Airport       LAS
 8             9         Charlotte Douglas International Airport       CLT
 9            10                      Miami International Airport       MIA
 10           11              Seattle–Tacoma International Airport       SEA
 11           12              Newark Liberty International Airport       EWR
 12           13              San Francisco International Airport       SFO
 13           14         Phoenix Sky Harbor International Airport       PHX
 14           15              George Bush Intercontinental Airport       IAH
 15           16                      Logan International Airport       BOS
```

In [38]:

```
table2[1]
```

Out[38]:

	Rank (2023)	Airports (large)	IATA Code	Major cities served	Metro area	State	2023[2]	2022[3]	2021[4]	2020[5]	2019[6]	2018[7]	2017
0	1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	Atlanta	GA	50950023	45396001	36676010	20559866	53505795	51865797	502519
1	2	Dallas/Fort Worth International Airport	DFW	Dallas and Fort Worth	Dallas–Fort Worth	TX	39246196	35345138	30005266	18593421	35778573	32821799	318169
2	3	Denver International Airport	DEN	Denver	Denver	CO	37863966	33773832	28645527	16243216	33592945	31362941	298090
3	4	Los Angeles International Airport	LAX	Los Angeles	Greater Los Angeles	CA	36676975	32326616	23663410	14055777	42939104	42624050	412324
4	5	O'Hare International Airport	ORD	Chicago	Chicagoland	IL	35843081	33120474	26350976	14606034	40871223	39873927	385930
5	6	John F. Kennedy International Airport	JFK	New York City	New York Metro	NY	30493867	27154885	15273342	8269819	31036655	30620769	295331
6	7	Orlando International Airport	MCO	Orlando	Orlando	FL	28033177	24469733	19618838	10467728	24562271	23202480	215654
7	8	Harry Reid International Airport	LAS	Las Vegas	Las Vegas	NV	27896019	25480500	19160342	10584059	24728361	23795012	233643
8	9	Charlotte Douglas International Airport	CLT	Charlotte	Charlotte	NC	25896193	23100300	20900875	12952869	24199688	22281949	220112
9	10	Miami International Airport	MIA	Miami	Miami Metro	FL	24716890	23949892	17500096	8786007	21421031	21021640	207092
10	11	Seattle–Tacoma International Airport	SEA	Seattle and Tacoma	Seattle Metro	WA	24594202	22157862	17430195	9462411	25001762	24024908	226391
11	12	Newark Liberty International Airport	EWR	Newark	New York Metro	NJ	24505862	21774690	14514049	7985474	23160763	22797602	215711
12	13	San Francisco International Airport	SFO	San Francisco	San Francisco Bay Area	CA	24191117	20411420	11725347	7745057	27779230	27790717	269000
13	14	Phoenix Sky Harbor International Airport	PHX	Phoenix	Phoenix	AZ	23880446	21852586	18940287	10531436	22433552	21622580	211854
14	15	George Bush Intercontinental Airport	IAH	Houston	Houston	TX	22228829	19814052	16242821	8682558	21905309	21157398	196037
15	16	Logan International Airport	BOS	Boston	Boston	MA	19962577	17443775	10909817	6035452	20699377	20006521	187597
16	17	Fort Lauderdale–Hollywood International Airport	FLL	Fort Lauderdale and Hollywood	Miami Metro	FL	17042632	15370165	13598994	8015744	17950989	17612331	158170
17	18	Minneapolis–Saint Paul International Airport	MSP	Minneapolis and Saint Paul	Minneapolis–Saint Paul	MN	17019086	15242089	12211409	7069720	19192917	18361942	184097
18	19	LaGuardia Airport	LGA	New York City	New York Metro	NY	16173072	14367463	7827307	4147116	15393601	15058501	146148
19	20	Detroit Metropolitan Airport	DTW	Detroit	Detroit Metro	MI	15378558	13751197	11517696	6822324	18143040	17436837	170360
20	21	Philadelphia International Airport	PHL	Philadelphia	Philadelphia Metro	PA	13656020	12421168	9820222	5753239	16006389	15292670	142712
21	22	Salt Lake City International Airport	SLC	Salt Lake City	Salt Lake City	UT	12905239	12383843	10795906	5753239	12840841	12226730	116159
22	23	Baltimore/Washington International Airport	BWI	Baltimore and Washington, D.C.	Baltimore	MD	12849636	11151169	9253561	5451355	13284687	13371816	129765
23	24	Ronald Reagan Washington National Airport	DCA	Washington, D.C.	Washington Metro	VA	12365011	11553850	6731737	3573489	11595454	11367176	115063
24	25	San Diego International Airport	SAN	San Diego	San Diego	CA	12190159	11162224	7836360	4637856	12648692	12174224	111399
25	26	Dulles International Airport	IAD	Washington, D.C.	Washington Metro	VA	12073231	10266324	7227875	3862658	11884117	11621623	110243
26	27	Tampa International Airport	TPA	Tampa	Tampa	FL	11677560	10539459	8847197	4966775	10978756	10368514	95485
27	28	Nashville International Airport	BNA	Nashville	Nashville	TN	11227159	9829062	7594049	4013995	8935654	8017347	69027
28	29	Austin–Bergstrom International Airport	AUS	Austin	Austin	TX	10833394	10382573	6666215	3141505	8683711	7921797	69731
29	30	Midway International Airport	MDW	Chicago	Chicagoland	IL	10659401	9650281	7680617	4236603	10081781	10678018	109120
30	31	Daniel K. Inouye International Airport	HNL	Honolulu	Honolulu	HI	10149761	8828395	5830928	3126391	9988678	9578505	97439

In [39]:
```
table2[1] = table2[1].drop(['2021[4]'], axis=1)
```

In [40]:
```
table2[1].head()
```

Out[40]:

	Rank (2023)	Airports (large)	IATA Code	Major cities served	Metro area	State	2023[2]	2022[3]	2020[5]	2019[6]	2018[7]	2017[8]	2016[9]	2015[10]
0	1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	Atlanta	GA	50950023	45396001	20559866	53505795	51865797	50251964	50501858	49340732
1	2	Dallas/Fort Worth International Airport	DFW	Dallas and Fort Worth	Dallas–Fort Worth	TX	39246196	35345138	18593421	35778573	32821799	31816933	31283579	31589839
2	3	Denver International Airport	DEN	Denver	Denver	CO	37863966	33773832	16243216	33592945	31362941	29809097	28267394	26280043
3	4	Los Angeles International Airport	LAX	Los Angeles	Greater Los Angeles	CA	36676975	32326616	14055777	42939104	42624050	41232432	39636042	36351272
4	5	O'Hare International Airport	ORD	Chicago	Chicagoland	IL	35843081	33120474	14606034	40871223	39873927	38593028	37589899	36305668

In [41]:
```
table2[1]['traffic_Chg19_20'] = table2[1]['2020[5]'] - table2[1]['2019[6]']
```

In [42]:
```
table2[1]['traffic_Chg18_19'] = table2[1]['2019[6]'] - table2[1]['2018[7]']
table2[1]['hubs'] = str('large_hub')
```

In [43]:
```
table2[1]
```

Out[43]:

	Rank (2023)	Airports (large)	IATA Code	Major cities served	Metro area	State	2023[2]	2022[3]	2020[5]	2019[6]	2018[7]	2017[8]	2016
0	1	Hartsfield–Jackson Atlanta International Airport	ATL	Atlanta	Atlanta	GA	50950023	45396001	20559866	53505795	51865797	50251964	505018
1	2	Dallas/Fort Worth International Airport	DFW	Dallas and Fort Worth	Dallas–Fort Worth	TX	39246196	35345138	18593421	35778573	32821799	31816933	312835
2	3	Denver International Airport	DEN	Denver	Denver	CO	37863966	33773832	16243216	33592945	31362941	29809097	282673
3	4	Los Angeles International Airport	LAX	Los Angeles	Greater Los Angeles	CA	36676975	32326616	14055777	42939104	42624050	41232432	396360
4	5	O'Hare International Airport	ORD	Chicago	Chicagoland	IL	35843081	33120474	14606034	40871223	39873927	38593028	375898
5	6	John F. Kennedy International Airport	JFK	New York City	New York Metro	NY	30493867	27154885	8269819	31036655	30620769	29533154	292391
6	7	Orlando International Airport	MCO	Orlando	Orlando	FL	28033177	24469733	10467728	24562271	23202480	21565448	202835
7	8	Harry Reid International Airport	LAS	Las Vegas	Las Vegas	NV	27896019	25480500	10584059	24728361	23795012	23364393	228332
8	9	Charlotte Douglas International Airport	CLT	Charlotte	Charlotte	NC	25896193	23100300	12952869	24199688	22281949	22011251	215118
9	10	Miami International Airport	MIA	Miami	Miami Metro	FL	24716890	23949892	8786007	21421031	21021640	20709225	208758
10	11	Seattle–Tacoma International Airport	SEA	Seattle and Tacoma	Seattle Metro	WA	24594202	22157862	9462411	25001762	24024908	22639124	218871
11	12	Newark Liberty International Airport	EWR	Newark	New York Metro	NJ	24505862	21774690	7985474	23160763	22797602	21571198	199230
12	13	San Francisco International Airport	SFO	San Francisco	San Francisco Bay Area	CA	24191117	20411420	7745057	27779230	27790717	26900048	257071
13	14	Phoenix Sky Harbor International Airport	PHX	Phoenix	Phoenix	AZ	23880446	21852586	10531436	22433552	21622580	21185458	208962
14	15	George Bush Intercontinental Airport	IAH	Houston	Houston	TX	22228829	19814052	8682558	21905309	21157398	19603731	200620
15	16	Logan International Airport	BOS	Boston	Boston	MA	19962577	17443775	6035452	20699377	20006521	18759742	177590
16	17	Fort Lauderdale–Hollywood International Airport	FLL	Fort Lauderdale and Hollywood	Miami Metro	FL	17042632	15370165	8015744	17950989	17612331	15817043	142632
17	18	Minneapolis–Saint Paul International Airport	MSP	Minneapolis and Saint Paul	Minneapolis–Saint Paul	MN	17019086	15242089	7069720	19192917	18361942	18409704	181238
18	19	LaGuardia Airport	LGA	New York City	New York Metro	NY	16173072	14367463	4147116	15393601	15058501	14614802	147625
19	20	Detroit Metropolitan Airport	DTW	Detroit	Detroit Metro	MI	15378558	13751197	6822324	18143040	17436837	17036092	168471
20	21	Philadelphia International Airport	PHL	Philadelphia	Philadelphia Metro	PA	13656020	12421168	5753239	16006389	15292670	14271243	145644
21	22	Salt Lake City International Airport	SLC	Salt Lake City	Salt Lake City	UT	12905239	12383843	5753239	12840841	12226730	11615954	111437
22	23	Baltimore/Washington International Airport	BWI	Baltimore and Washington, D.C.	Baltimore	MD	12849636	11151169	5451355	13284687	13371816	12976554	123409
23	24	Ronald Reagan Washington National Airport	DCA	Washington, D.C.	Washington Metro	VA	12365011	11553850	3573489	11595454	11367176	11506310	114708
24	25	San Diego International Airport	SAN	San Diego	San Diego	CA	12190159	11162224	4637856	12648692	12174224	11139933	103401
25	26	Dulles International Airport	IAD	Washington, D.C.	Washington Metro	VA	12073231	10266324	3862658	11884117	11621623	11024306	105969
26	27	Tampa International Airport	TPA	Tampa	Tampa	FL	11677560	10539459	4966775	10978756	10368514	9548580	91949
27	28	Nashville International Airport	BNA	Nashville	Nashville	TN	11227159	9829062	4013995	8935654	8017347	6902771	63385
28	29	Austin–Bergstrom International Airport	AUS	Austin	Austin	TX	10833394	10382573	3141505	8683711	7921797	6973115	60955
29	30	Midway International Airport	MDW	Chicago	Chicagoland	IL	10659401	9650281	4236603	10081781	10678018	10912074	110443
30	31	Daniel K. Inouye International Airport	HNL	Honolulu	Honolulu	HI	10149761	8828395	3126391	9988678	9578505	9743989	96563

In [44]:
```
table2[1] = table2[1][['IATA Code', 'traffic_Chg19_20', 'traffic_Chg18_19','hubs']]
table2[1]
```

Out[44]:

	IATA Code	traffic_Chg19_20	traffic_Chg18_19	hubs
0	ATL	-32945929	1639998	large_hub
1	DFW	-17185152	2956774	large_hub
2	DEN	-17349729	2230004	large_hub
3	LAX	-28883327	315054	large_hub
4	ORD	-26265189	997296	large_hub
5	JFK	-22766836	415886	large_hub
6	MCO	-14094543	1359791	large_hub
7	LAS	-14144302	933349	large_hub
8	CLT	-11246819	1917739	large_hub
9	MIA	-12635024	399391	large_hub
10	SEA	-15539351	976854	large_hub
11	EWR	-15175289	363161	large_hub
12	SFO	-20034173	-11487	large_hub
13	PHX	-11902116	810972	large_hub
14	IAH	-13222751	747911	large_hub
15	BOS	-14663925	692856	large_hub
16	FLL	-9935245	338658	large_hub
17	MSP	-12123197	830975	large_hub
18	LGA	-11246485	335100	large_hub
19	DTW	-11320716	706203	large_hub
20	PHL	-10253150	713719	large_hub
21	SLC	-7087602	614111	large_hub
22	BWI	-7833332	-87129	large_hub
23	DCA	-8021965	228278	large_hub
24	SAN	-8010836	474468	large_hub
25	IAD	-8021459	262494	large_hub
26	TPA	-6011981	610242	large_hub
27	BNA	-4921659	918307	large_hub
28	AUS	-5542206	761914	large_hub
29	MDW	-5845178	-596237	large_hub
30	HNL	-6862287	410173	large_hub

In [45]:
```
table2[2].head()
```

Out[45]:

	Rank (2023)	Airports (medium hubs)	IATA Code	City served	Metro Area	State	2023[2]	2022[3]	2021[4]	2020[5]	2019[6]	2018[7]	2017[8]	2016[9]	2015[10]
0	32	Dallas Love Field	DAL	Dallas	Dallas–Fort Worth	TX	8559009	7819129	6487563	3669930	8408457	8134848	7876769	7554596	7040921
1	33	Portland International Airport	PDX	Portland	Portland	OR	8123024	7241882	5759879	3455877	9797408	9940866	9435473	9071154	8340234
2	34	St. Louis Lambert International Airport	STL	St. Louis	St. Louis	MO	7307544	6709080	5070471	3041765	7946986	7822274	7372805	6793076	6239231
3	35	Raleigh–Durham International Airport	RDU	Raleigh	Research Triangle	NC	7118953	5849665	4311049	2337496	6919429	6416822	5851004	5401714	4954717
4	36	William P. Hobby Airport	HOU	Houston	Houston	TX	6800214	6462948	5560780	3127178	7069614	6937061	6741870	6285181	5937944

In [46]:
```
table2[2]['traffic_Chg19_20'] = table2[2]['2020[5]'] - table2[2]['2019[6]']
table2[2]['traffic_Chg18_19'] = table2[2]['2019[6]'] - table2[2]['2018[7]']
table2[2]['hubs'] = str('Medium_hub')
```

In [47]:
```
table2[2] = table2[2][['IATA Code', 'traffic_Chg19_20', 'traffic_Chg18_19','hubs']]
table2[2]
```

Out[47]:

	IATA Code	traffic_Chg19_20	traffic_Chg18_19	hubs
0	DAL	-4738527	273609	Medium_hub
1	PDX	-6341531	-143458	Medium_hub
2	STL	-4905221	124712	Medium_hub
3	RDU	-4581933	502607	Medium_hub
4	HOU	-3942436	132553	Medium_hub
5	SMF	-3744071	422783	Medium_hub
6	MSY	-4084499	151623	Medium_hub
7	SJC	-5545699	688269	Medium_hub
8	SJU	-2227266	556705	Medium_hub
9	SNA	-3328440	-163873	Medium_hub
10	MCI	-3591803	-175712	Medium_hub
11	OAK	-4288936	-238091	Medium_hub
12	SAT	-3103022	178553	Medium_hub
13	RSW	-2197328	424899	Medium_hub
14	CLE	-2904385	57961	Medium_hub
15	IND	-2720057	14143	Medium_hub
16	PIT	-2973541	45914	Medium_hub
17	CVG	-2684062	144199	Medium_hub
18	CMH	-2594471	117495	Medium_hub
19	PBI	-1941697	197387	Medium_hub
20	OGG	-2656666	219674	Medium_hub
21	JAX	-2112422	361383	Medium_hub
22	ONT	-1485056	223831	Medium_hub
23	BUR	-1931882	308480	Medium_hub
24	BDL	-2173581	-7120	Medium_hub
25	CHS	-1431208	182975	Medium_hub
26	MKE	-2110688	-174744	Medium_hub
27	ANC	-1556542	70942	Medium_hub
28	ABQ	-1772528	-5819	Medium_hub
29	OMA	-1419029	-1813	Medium_hub
30	MEM	-1302461	105359	Medium_hub
31	RIC	-1343096	142216	Medium_hub
32	BOI	-1066509	114569	Medium_hub

In [48]:
```
# Merge all Wikipedia tables
```

In [49]:
```
final_wiki = [table2[1],table2[2]]
final_wiki = pd.concat(final_wiki, ignore_index=True)
final_wiki
```

Out[49]:

	IATA Code	traffic_Chg19_20	traffic_Chg18_19	hubs
0	ATL	-32945929	1639998	large_hub
1	DFW	-17185152	2956774	large_hub
2	DEN	-17349729	2230004	large_hub
3	LAX	-28883327	315054	large_hub
4	ORD	-26265189	997296	large_hub
...	...	...	...	...
59	ABQ	-1772528	-5819	Medium_hub
60	OMA	-1419029	-1813	Medium_hub
61	MEM	-1302461	105359	Medium_hub
62	RIC	-1343096	142216	Medium_hub
63	BOI	-1066509	114569	Medium_hub

64 rows × 4 columns

In [50]:
```
final_data = new_df.merge(final_wiki, left_on ='iata_code', right_on = "IATA Code")
```

In [51]:
```
final_data
```

Out[51]:

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay	airport_ref	...	longitude_deg	elevation_ft	scheduled
0	4	AA	2466	SFO	DFW	3	20	195	1	3878	...	-122.375000	13.0	
1	231	AA	526	SFO	DFW	3	360	215	0	3878	...	-122.375000	13.0	
2	234	AA	552	SFO	MIA	3	360	315	1	3878	...	-122.375000	13.0	
3	905	AA	810	SFO	ORD	3	385	255	0	3878	...	-122.375000	13.0	
4	1739	AA	24	SFO	JFK	3	425	325	1	3878	...	-122.375000	13.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
362690	506267	9E	4052	DAL	MEM	4	370	90	0	3479	...	-96.851799	487.0	
362691	512858	9E	3704	DAL	MEM	4	705	92	1	3479	...	-96.851799	487.0	
362692	518247	9E	4060	DAL	MEM	4	990	90	0	3479	...	-96.851799	487.0	
362693	524678	9E	4052	DAL	MEM	5	370	90	1	3479	...	-96.851799	487.0	
362694	530841	9E	3704	DAL	MEM	5	705	92	0	3479	...	-96.851799	487.0	

362695 rows × 30 columns

In [52]:
```
final_data.drop_duplicates(subset=['id'],keep='first',inplace=True)
```

In [53]:
```
final_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 362695 entries, 0 to 362694
Data columns (total 30 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   id                 362695 non-null  int64  
 1   Airline            362695 non-null  object 
 2   Flight             362695 non-null  int64  
 3   AirportFrom        362695 non-null  object 
 4   AirportTo          362695 non-null  object 
 5   DayOfWeek          362695 non-null  int64  
 6   Time               362695 non-null  int64  
 7   Length             362695 non-null  int64  
 8   Delay              362695 non-null  int64  
 9   airport_ref        362695 non-null  int64  
 10  airport_ident      362695 non-null  object 
 11  length_ft          362695 non-null  float64
 12  width_ft           362695 non-null  float64
 13  surface            362695 non-null  object 
 14  lighted            362695 non-null  int64  
 15  closed             362695 non-null  int64  
 16  ident              362695 non-null  object 
 17  type               362695 non-null  object 
 18  name               362695 non-null  object 
 19  latitude_deg       362695 non-null  float64
 20  longitude_deg      362695 non-null  float64
 21  elevation_ft       362695 non-null  float64
 22  scheduled_service  362695 non-null  object 
 23  iata_code          362695 non-null  object 
 24  IATA               362695 non-null  object 
 25  Founded            362695 non-null  float64
 26  IATA Code          362695 non-null  object 
 27  traffic_Chg19_20   362695 non-null  int64  
 28  traffic_Chg18_19   362695 non-null  int64  
 29  hubs               362695 non-null  object 
dtypes: float64(6), int64(11), object(13)
memory usage: 83.0+ MB
```

In [54]:
```
# Remove columns that are not usable

final_data.drop(['id','AirportFrom','airport_ident','iata_code','AirportTo','surface','ident',
'IATA', 'IATA Code','name'], axis=1,inplace=True)
```

In [55]:
```
final_data.info()
```

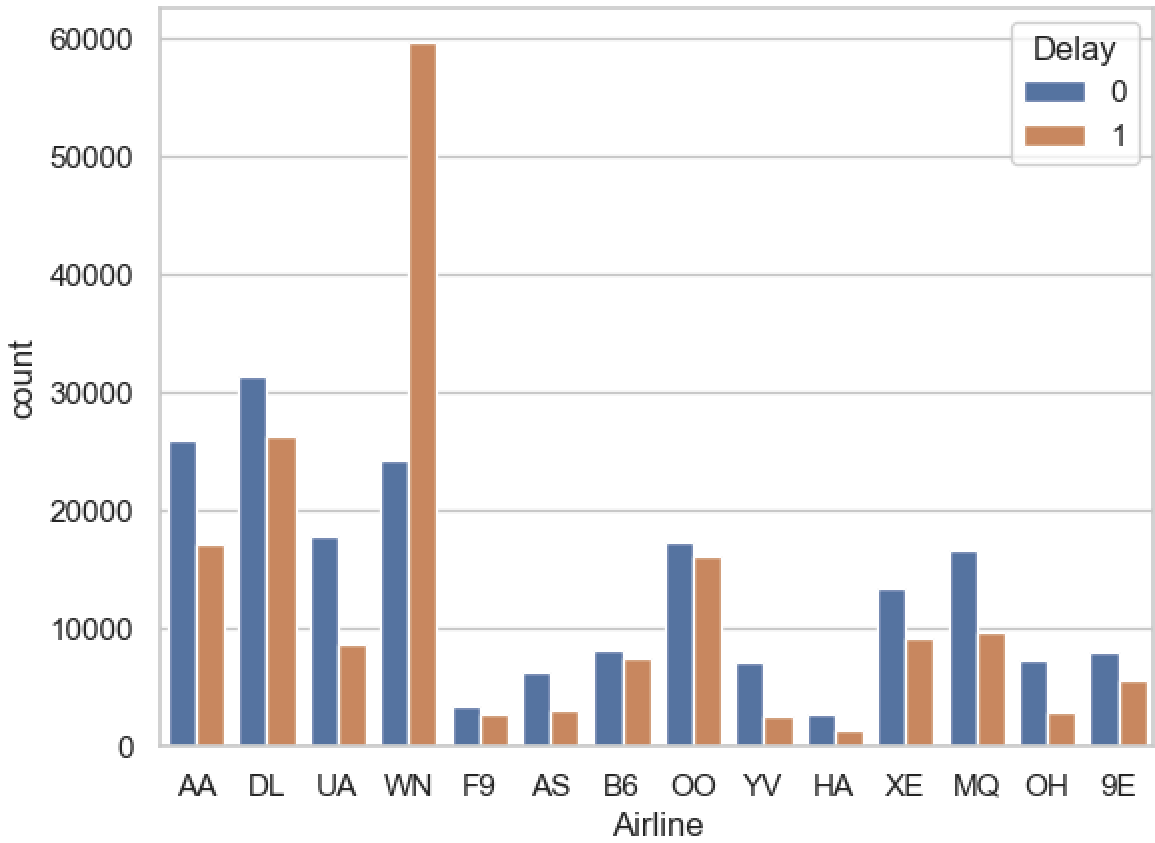
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 362695 entries, 0 to 362694
Data columns (total 20 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Airline            362695 non-null  object 
 1   Flight             362695 non-null  int64  
 2   DayOfWeek          362695 non-null  int64  
 3   Time               362695 non-null  int64  
 4   Length             362695 non-null  int64  
 5   Delay              362695 non-null  int64  
 6   airport_ref        362695 non-null  int64  
 7   length_ft          362695 non-null  float64
 8   width_ft           362695 non-null  float64
 9   lighted            362695 non-null  int64  
 10  closed             362695 non-null  int64  
 11  type               362695 non-null  object 
 12  latitude_deg       362695 non-null  float64
 13  longitude_deg      362695 non-null  float64
 14  elevation_ft       362695 non-null  float64
 15  scheduled_service  362695 non-null  object 
 16  Founded            362695 non-null  float64
 17  traffic_Chg19_20   362695 non-null  int64  
 18  traffic_Chg18_19   362695 non-null  int64  
 19  hubs               362695 non-null  object 
dtypes: float64(6), int64(10), object(4)
memory usage: 55.3+ MB
```

In [56]:
```
# Check for any NULL values and treat them accordingly

final_data.isnull().sum()
```

Out[56]:
```
Airline              0
Flight               0
DayOfWeek            0
Time                 0
Length               0
Delay                0
airport_ref          0
length_ft            0
width_ft             0
lighted              0
closed               0
type                 0
latitude_deg         0
longitude_deg        0
elevation_ft         0
scheduled_service    0
Founded              0
traffic_Chg19_20     0
traffic_Chg18_19     0
hubs                 0
dtype: int64
```

In [57]:
```
sns.set_theme(style='whitegrid')

sns.countplot(x=final_data['Airline'],hue= final_data['Delay'])
```

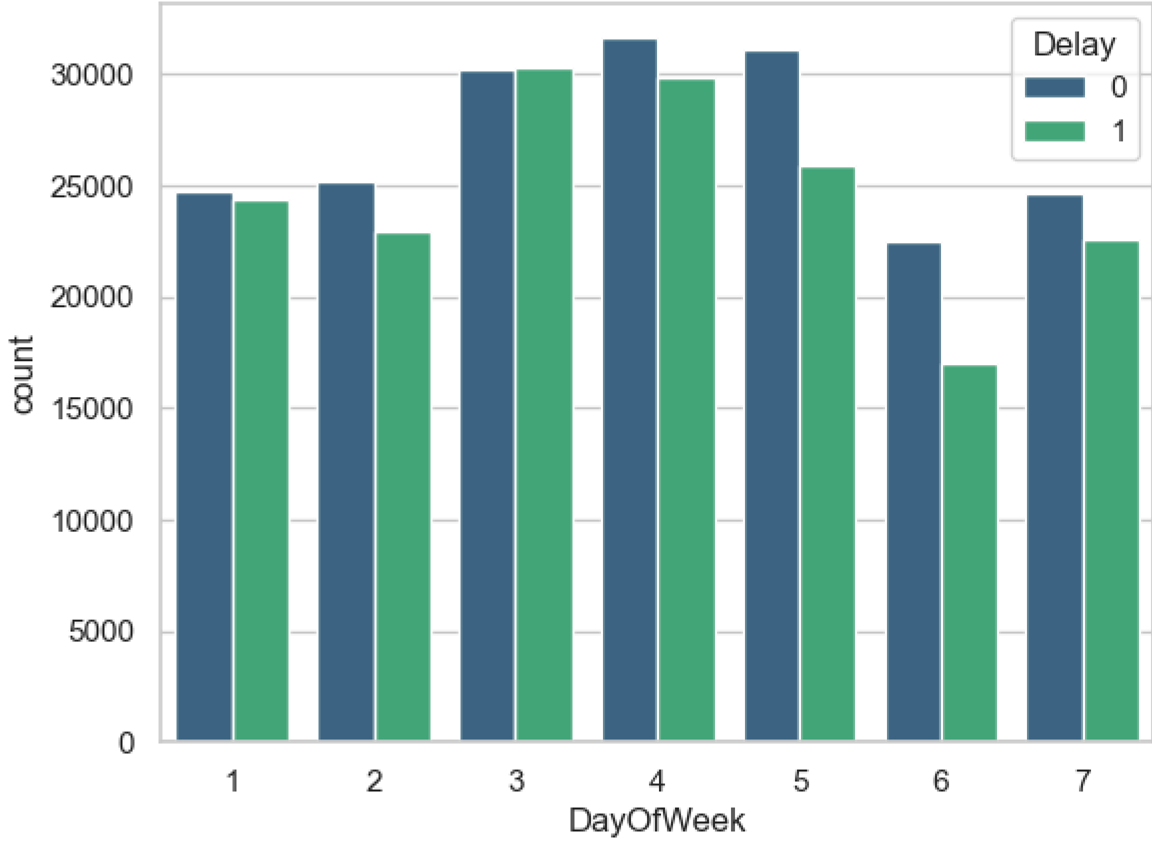
Out[57]:
```
<Axes: xlabel='Airline', ylabel='count'>
```



The above graph shows that more than 70% of delayed flights belong to Southwest Airlines

In [58]:
```
sns.countplot(x=final_data['DayOfWeek'],hue= final_data['Delay'],palette = 'viridis')
```

Out[58]: <Axes: xlabel='DayOfWeek', ylabel='count'>



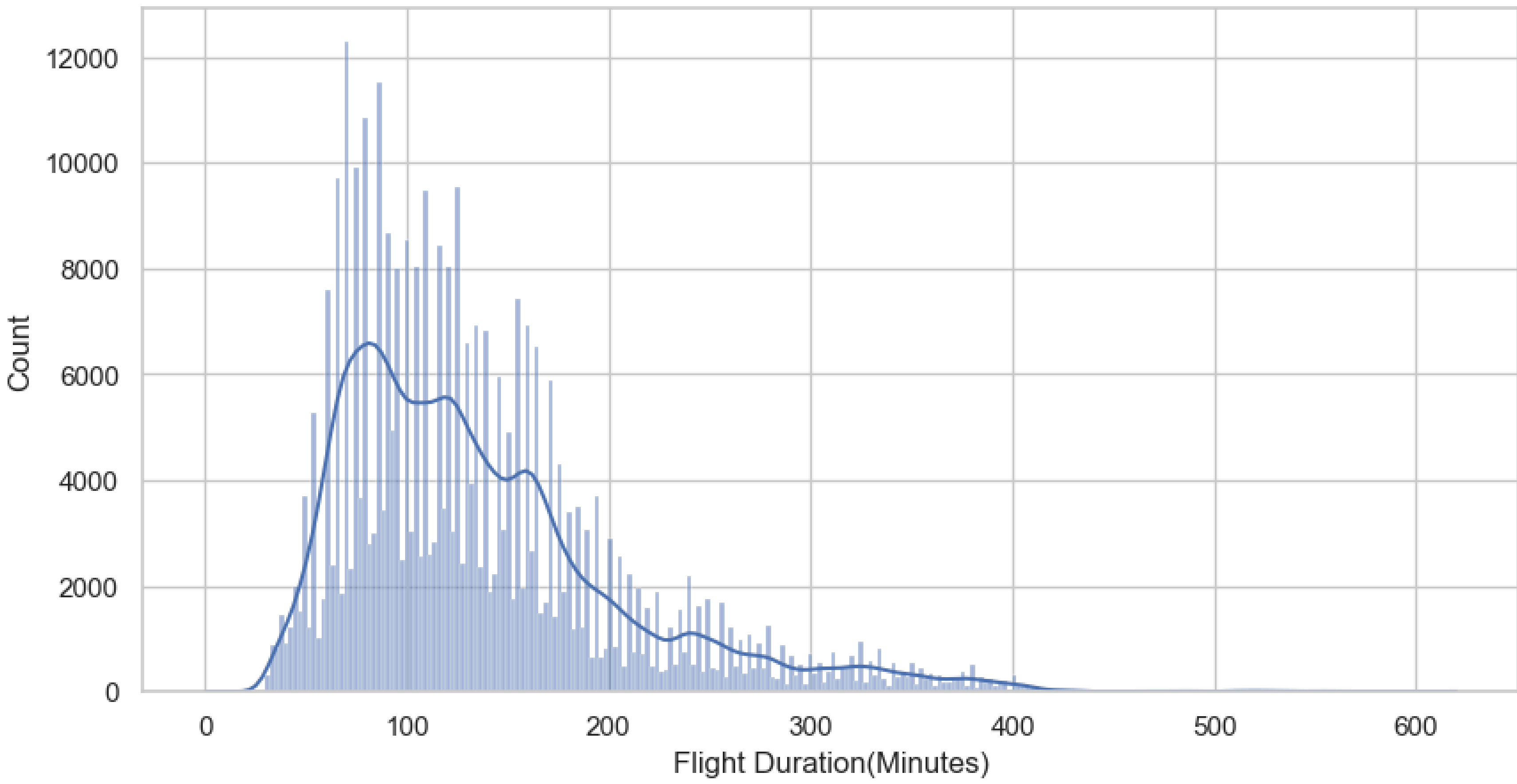
The above graph shows that on 6th day of the week we have least delayed flights

In [59]:
```
final_data['Length'].max()
```

Out[59]: 620

In [60]:
```
fig, ax = plt.subplots(figsize=(10, 5))

sns.histplot(data = final_data,x='Length',bins = 'auto' ,kde=True)
plt.xlabel('Flight Duration(Minutes)')
```

Out[60]: Text(0.5, 0, 'Flight Duration(Minutes)')



In [61]:
```
final_data['Airline'][final_data['Length']<180].value_counts()
```

Out[61]:
```
Airline
WN    71327
DL    40148
OO    31764
MQ    25488
AA    24637
XE    21670
UA    15777
9E    13507
B6     9888
OH     9798
YV     9610
AS     5996
F9     5105
HA     3034
Name: count, dtype: int64
```

In [62]:
```
## The above airlines are recommended for Short distance travel as flight duration lasts between 30 minut
```

In [63]:
```
final_data['Airline'][(final_data['Length']>180 ) & (final_data['Length'] <360)].value_counts()
```

Out[63]:
```
Airline
DL    16504
AA    16051
WN    11395
UA     9385
B6     4847
AS     2822
OO     1576
F9     1075
XE      994
HA      751
MQ      604
OH      433
YV      246
9E       47
Name: count, dtype: int64
```

In [64]:
```
## The above airlines are recommended for Medium distance travel as flight duration lasts between 3 to 6
```

In [65]:
```
final_data['Airline'][final_data['Length'] > 360].value_counts()
```

Out[65]:
```
Airline
UA    1304
AA    1081
DL     842
B6     822
AS     540
HA     252
WN      52
Name: count, dtype: int64
```

In [66]:
```
## The above airlines are recommended for Long distance travel as flight duration lasts more than 6 hours
```

In [67]:
```
final_data[final_data['Length'] > 360].describe().T
```

Out[67]:

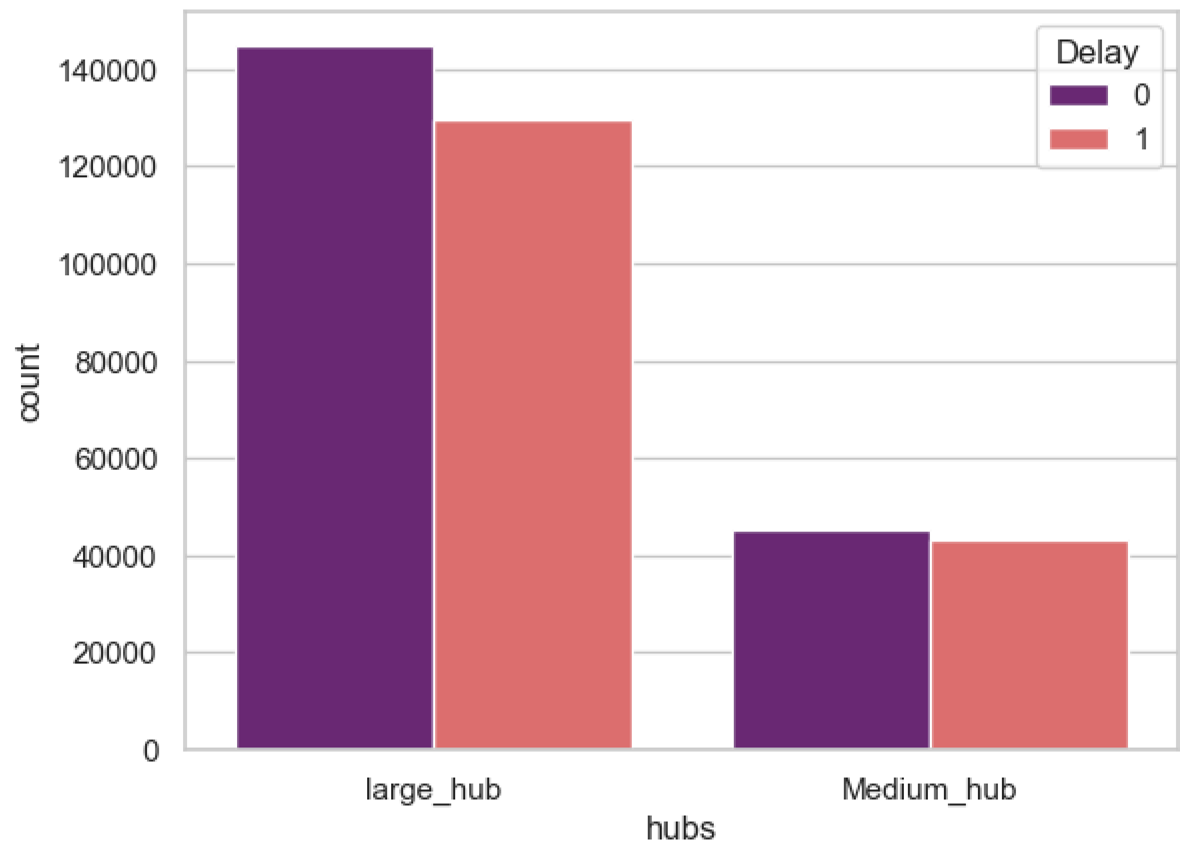
	count	mean	std	min	25%	50%	75%	max
Flight	4893.0	5.592187e+02	7.509137e+02	1.000000e+00	5.900000e+01	2.090000e+02	8.490000e+02	3.760000e+03
DayOfWeek	4893.0	4.003270e+00	1.925851e+00	1.000000e+00	2.000000e+00	4.000000e+00	6.000000e+00	7.000000e+00
Time	4893.0	8.287658e+02	2.857339e+02	1.000000e+02	5.500000e+02	8.850000e+02	1.080000e+03	1.435000e+03
Length	4893.0	3.957466e+02	4.073634e+01	3.610000e+02	3.740000e+02	3.850000e+02	4.000000e+02	6.200000e+02
Delay	4893.0	4.161046e-01	4.929617e-01	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
airport_ref	4893.0	3.741812e+03	4.890171e+02	3.384000e+03	3.602000e+03	3.622000e+03	3.670000e+03	6.384000e+03
length_ft	4893.0	1.005470e+04	2.102188e+03	4.892000e+03	7.861000e+03	1.100000e+04	1.207900e+04	1.207900e+04
width_ft	4893.0	1.716636e+02	2.506597e+01	1.000000e+02	1.500000e+02	1.500000e+02	2.000000e+02	2.000000e+02
lighted	4893.0	9.513591e-01	2.151382e-01	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
closed	4893.0	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
latitude_deg	4893.0	3.893579e+01	6.497714e+00	1.843940e+01	3.894450e+01	4.063945e+01	4.193851e+01	6.117440e+01
longitude_deg	4893.0	-8.828553e+01	2.515070e+01	-1.579242e+02	-9.703800e+01	-7.377932e+01	-7.377932e+01	-6.600180e+01
elevation_ft	4893.0	2.385510e+02	7.470823e+02	8.000000e+00	1.300000e+01	1.800000e+01	1.250000e+02	5.431000e+03
Founded	4893.0	1.939004e+03	2.692440e+01	1.924000e+03	1.926000e+03	1.926000e+03	1.932000e+03	1.998000e+03
traffic_Chg19_20	4893.0	-1.763080e+07	6.656975e+06	-3.294593e+07	-2.276684e+07	-1.718515e+07	-1.414430e+07	-1.556542e+06
traffic_Chg18_19	4893.0	5.823171e+05	4.670072e+05	-1.434580e+05	4.101730e+05	4.158860e+05	6.928560e+05	2.956774e+06

The majority of long-duration flights seem to depart between late morning and early evening, with a significant number of flights departing in the early afternoon.

Also there is a wide range of departure times, indicating that long-duration flights are spread throughout the day, from early morning to late night.

In [68]:
```
sns.countplot(data = final_data, x='hubs',hue='Delay',palette='magma')
```

Out[68]: <Axes: xlabel='hubs', ylabel='count'>



In [69]:
```
# From the large hubs its clear approx. 1,20,000 filghts are delayed but from the medium hubs aprrox 40,0
```

Using hypothesis testing to determine if the airport's altitude has anything to do with flight delays for incoming and departing flights

In [70]:
```
from scipy.stats import chi2_contingency
table = [final_data['latitude_deg'],final_data['Delay']]
stat, p, dof, expected = chi2_contingency(table)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
 print('Probably independent')
else:
 print('Probably dependent')
```

```
stat=194163.649, p=1.000
Probably independent
```

In [71]:
```
#So its clear from the above hypothesis testing that altitude is not a contributing factor for flight del
```

Check if the number of runways at an airport affects flight delays

In [72]:
```
 from scipy.stats import chi2_contingency
table = [final_data['airport_ref'],final_data['Delay']]
stat, p, dof, expected = chi2_contingency(table)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
  print('Probably independent')
else:
  print('Probably dependent')
```

```
stat=199667.595, p=1.000
Probably independent
```

In [73]:
```
#So its clear from the above hypothesis testing that number of runways at an airport is not a
#contributing factor for flight delay
```

Check if the duration of a flight (length) affects flight delays

In [74]:
```
from scipy.stats import spearmanr
data1 = final_data['Length']
data2 = final_data['Delay']
stat, p = spearmanr(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
  print('Probably independent')
else:
  print('Probably dependent')
```

```
stat=-0.001, p=0.561
Probably independent
```

In [75]:
```
#Since both the variables are independent, hence length of the flight is not contributing to the flight d
```

In [76]:
```
final_data
```

Out[76]:

	Airline	Flight	DayOfWeek	Time	Length	Delay	airport_ref	length_ft	width_ft	lighted	closed	type	latitude_deg	longitude
0	AA	2466	3	20	195	1	3878	7500.0	200.0	1	0	large_airport	37.618999	-122.3
1	AA	526	3	360	215	0	3878	7500.0	200.0	1	0	large_airport	37.618999	-122.3
2	AA	552	3	360	315	1	3878	7500.0	200.0	1	0	large_airport	37.618999	-122.3
3	AA	810	3	385	255	0	3878	7500.0	200.0	1	0	large_airport	37.618999	-122.3
4	AA	24	3	425	325	1	3878	7500.0	200.0	1	0	large_airport	37.618999	-122.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
362690	9E	4052	4	370	90	0	3479	7752.0	150.0	1	0	medium_airport	32.847099	-96.8
362691	9E	3704	4	705	92	1	3479	7752.0	150.0	1	0	medium_airport	32.847099	-96.8
362692	9E	4060	4	990	90	0	3479	7752.0	150.0	1	0	medium_airport	32.847099	-96.8
362693	9E	4052	5	370	90	1	3479	7752.0	150.0	1	0	medium_airport	32.847099	-96.8
362694	9E	3704	5	705	92	0	3479	7752.0	150.0	1	0	medium_airport	32.847099	-96.8

362695 rows × 20 columns

Checking for correlation between flight delay predictors

In [77]:
```
predictors = final_data.drop(['Delay'], axis=1)

numeric_predictors = predictors.select_dtypes(include=[np.number])

numeric_predictors
```

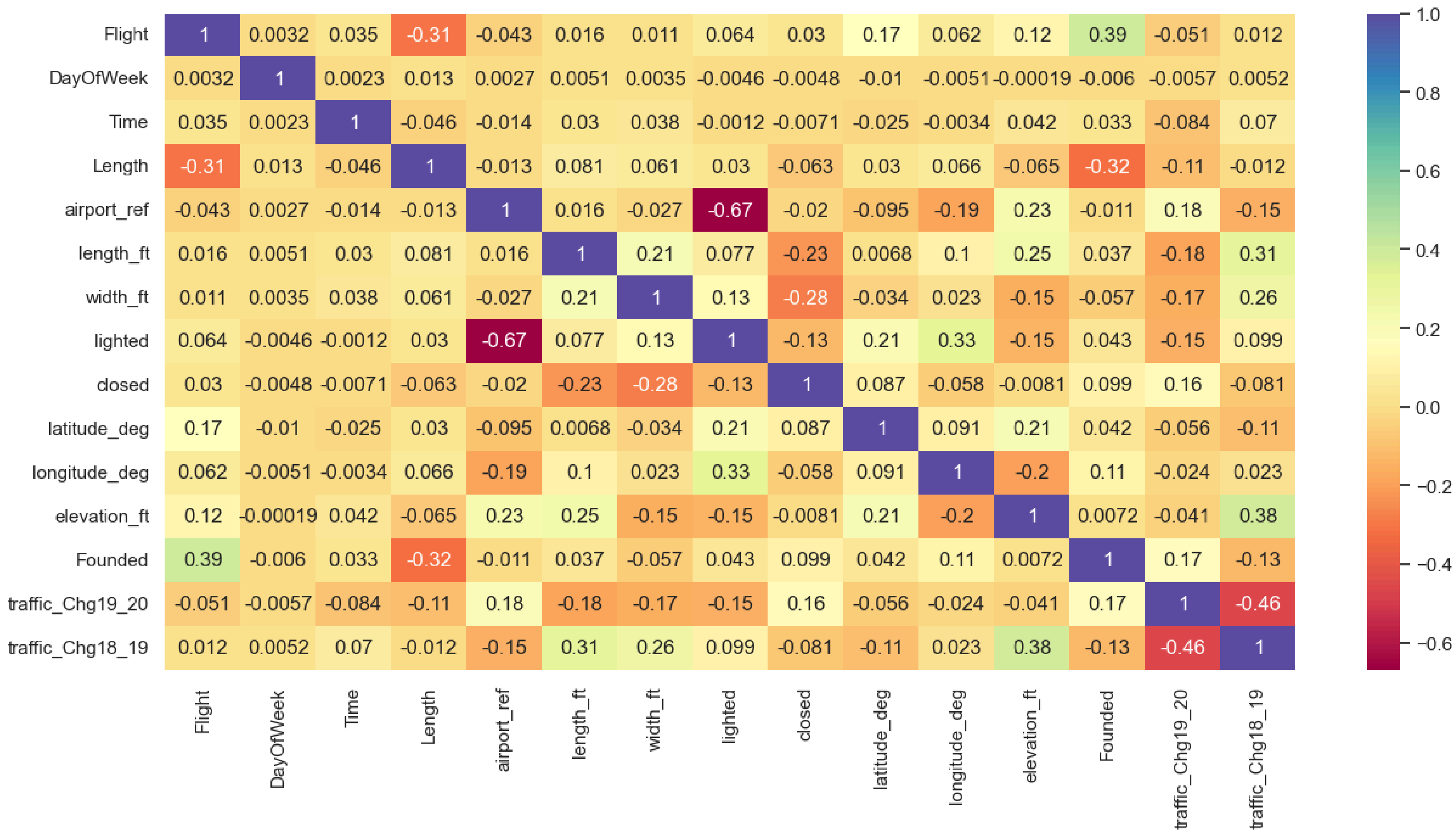
Out[77]:

	Flight	DayOfWeek	Time	Length	airport_ref	length_ft	width_ft	lighted	closed	latitude_deg	longitude_deg	elevation_ft	Founded	t
0	2466	3	20	195	3878	7500.0	200.0	1	0	37.618999	-122.375000	13.0	1926.0	
1	526	3	360	215	3878	7500.0	200.0	1	0	37.618999	-122.375000	13.0	1926.0	
2	552	3	360	315	3878	7500.0	200.0	1	0	37.618999	-122.375000	13.0	1926.0	
3	810	3	385	255	3878	7500.0	200.0	1	0	37.618999	-122.375000	13.0	1926.0	
4	24	3	425	325	3878	7500.0	200.0	1	0	37.618999	-122.375000	13.0	1926.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
362690	4052	4	370	90	3479	7752.0	150.0	1	0	32.847099	-96.851799	487.0	1985.0	
362691	3704	4	705	92	3479	7752.0	150.0	1	0	32.847099	-96.851799	487.0	1985.0	
362692	4060	4	990	90	3479	7752.0	150.0	1	0	32.847099	-96.851799	487.0	1985.0	
362693	4052	5	370	90	3479	7752.0	150.0	1	0	32.847099	-96.851799	487.0	1985.0	
362694	3704	5	705	92	3479	7752.0	150.0	1	0	32.847099	-96.851799	487.0	1985.0	

362695 rows × 15 columns

In [78]:
```
corr_matrix = numeric_predictors.corr()
```

In [79]:
```
fig, ax = plt.subplots(figsize=(15, 7))
sns.heatmap(corr_matrix, yticklabels = True,annot=True, cmap='Spectral')
plt.show()
```



Using OneHotEncoder and OrdinalEncoder to deal with categorical variables

In [80]:
```
final_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 362695 entries, 0 to 362694
Data columns (total 20 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Airline            362695 non-null  object 
 1   Flight             362695 non-null  int64  
 2   DayOfWeek          362695 non-null  int64  
 3   Time               362695 non-null  int64  
 4   Length             362695 non-null  int64  
 5   Delay              362695 non-null  int64  
 6   airport_ref        362695 non-null  int64  
 7   length_ft          362695 non-null  float64
 8   width_ft           362695 non-null  float64
 9   lighted            362695 non-null  int64  
 10  closed             362695 non-null  int64  
 11  type               362695 non-null  object 
 12  latitude_deg       362695 non-null  float64
 13  longitude_deg      362695 non-null  float64
 14  elevation_ft       362695 non-null  float64
 15  scheduled_service  362695 non-null  object 
 16  Founded            362695 non-null  float64
 17  traffic_Chg19_20   362695 non-null  int64  
 18  traffic_Chg18_19   362695 non-null  int64  
 19  hubs               362695 non-null  object 
dtypes: float64(6), int64(10), object(4)
memory usage: 55.3+ MB
```

In [81]:
```
final_data['Airline'].value_counts()
```

Out[81]:
```
Airline
WN    83831
DL    57713
AA    43200
OO    33415
UA    26551
MQ    26324
XE    22679
B6    15619
9E    13554
OH    10259
YV     9856
AS     9477
F9     6180
HA     4037
Name: count, dtype: int64
```

In [82]:
```
final_data['type'].value_counts()
```

Out[82]:
```
type
large_airport     342440
medium_airport     20255
Name: count, dtype: int64
```

In [83]:
```
final_data['scheduled_service'].value_counts()
```

Out[83]:
```
scheduled_service
yes    362695
Name: count, dtype: int64
```

In [84]:
```
final_data['hubs'].value_counts()
```

Out[84]:
```
hubs
large_hub     274167
Medium_hub     88528
Name: count, dtype: int64
```

In [85]:
```
#The scheduled_service column has same value,so it will not help in prediction. It will be removed and ot
```

In [86]:
```
final_data = final_data.drop(['scheduled_service'], axis=1)
```

In [87]:
```
# Using the ordinal encoder
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [88]:
```
final_data['Airline'] = le.fit_transform(final_data['Airline'])
final_data['type'] = le.fit_transform(final_data['type'])
final_data['hubs'] = le.fit_transform(final_data['hubs'])
```

In [89]:
```
final_data.head(50)
```

Out[89]:

	Airline	Flight	DayOfWeek	Time	Length	Delay	airport_ref	length_ft	width_ft	lighted	closed	type	latitude_deg	longitude_deg	elevatio
0	1	2466	3	20	195	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
1	1	526	3	360	215	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
2	1	552	3	360	315	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
3	1	810	3	385	255	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
4	1	24	3	425	325	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
5	1	600	3	440	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
6	1	1929	3	455	90	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
7	1	39	3	495	345	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
8	1	12	3	565	334	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
9	1	2222	3	565	210	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
10	1	1894	3	585	90	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
11	1	1972	3	610	255	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
12	1	1964	3	625	215	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
13	1	1512	3	670	255	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
14	1	1492	3	730	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
15	1	16	3	750	335	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
16	1	1923	3	755	85	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
17	1	442	3	765	320	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
18	1	2282	3	795	210	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
19	1	554	3	855	260	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
20	1	1931	3	875	80	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
21	1	20	3	885	325	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
22	1	564	3	905	205	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
23	1	1528	3	990	205	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
24	1	2578	3	1015	80	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
25	1	618	3	1025	245	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
26	1	1943	3	1150	80	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
27	1	272	3	1255	315	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
28	1	18	3	1380	329	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
29	1	1522	3	1435	240	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
30	1	2466	4	20	195	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
31	1	526	4	360	215	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
32	1	552	4	360	315	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
33	1	810	4	385	255	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
34	1	24	4	425	325	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
35	1	600	4	440	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
36	1	1929	4	455	90	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
37	1	39	4	495	345	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
38	1	12	4	565	334	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
39	1	2222	4	565	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
40	1	1894	4	585	90	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
41	1	1972	4	610	255	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
42	1	1964	4	625	215	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
43	1	1512	4	670	255	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
44	1	1492	4	730	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
45	1	16	4	750	335	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
46	1	1923	4	755	85	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
47	1	442	4	765	320	1	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
48	1	2282	4	795	210	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	
49	1	554	4	855	260	0	3878	7500.0	200.0	1	0	0	37.618999	-122.375	

Applying logistic regression (use stochastic gradient descent optimizer) and decision tree models. Using the stratified five-fold method to build and validate the models

In [90]:
```
# Assigning the features to x and y variables
X = final_data.drop(['Delay'], axis= 1)
y = final_data['Delay']
```

In [91]:
```
from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
X = scaler.fit_transform(X)
```

In [96]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.9,random_state=101)
```

In [97]:
```
# Applying the logistic regression with the randomsearchcv hypermeter tunning

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import RandomizedSearchCV

lr = LogisticRegression()
```

In [98]:
```
params = {"penalty": ["l1","l2"],
'solver': ['newton-cg', 'liblinear']}
# Cross Validation
folds = 5
rscv = RandomizedSearchCV(estimator = lr,param_distributions = params,scoring = "accuracy",verbose = 1,cv
```

In [99]:
```
rscv.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 4 candidates, totalling 20 fits
```

Out[99]:
```
▸ RandomizedSearchCV
▸ estimator: LogisticRegression
  ▸ LogisticRegression
```

In [100]:
```
print(rscv.best_params_)
print(rscv.best_score_)
```

```
{'solver': 'newton-cg', 'penalty': 'l2'}
0.5945776211993566
```

In [101]:
```
lr = LogisticRegression(penalty= 'l2', solver= 'newton-cg')
lr.fit(X_train,y_train).score(X_train,y_train)
```

Out[101]: 0.5944734625105308

In [102]:
```
lr.score(X_test, y_test)
```

Out[102]: 0.5962227736421285

In [103]:
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier()
params = {
'criterion': ["gini", "entropy"],
'min_samples_leaf' : [2,3,4,5,6,7,8,9],
"max_depth": [2,3,4,5,6,7,8,9]
}
rscv = RandomizedSearchCV(estimator = dt,param_distributions= params,scoring = "accuracy",cv= 5,verbose=1
rscv.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
```

Out[103]:
```
▸ RandomizedSearchCV
▸ estimator: DecisionTreeClassifier
  ▸ DecisionTreeClassifier
```

In [104]:
```
print(rscv.best_params_)
print(rscv.best_score_)
```

```
{'min_samples_leaf': 4, 'max_depth': 9, 'criterion': 'entropy'}
0.6484736156850731
```

In [105]:
```
dtc = DecisionTreeClassifier(max_depth= 9, criterion='entropy',min_samples_leaf= 6)
dtc.fit(X_train, y_train).score(X_train, y_train)
```

Out[105]:
```
0.6543953434939113
```

In [106]:
```
dtc.score(X_test, y_test)
```

Out[106]:
```
0.6497932175351531
```

In [107]:
```
#Based on the result it's evident that decision tree has good accuracy
```

Using the stratified five-fold method to build and validate the models through XGB classifier

In [108]:
```
from xgboost import XGBClassifier
from sklearn.model_selection import RandomizedSearchCV

# Create the parameter grid: gbm_param_grid
gbm_param_grid = {
    'n_estimators': range(8, 20),
    'max_depth': range(6, 10),
    'learning_rate': [.4, .45, .5, .55, .6],
    'colsample_bytree': [.6, .7, .8, .9, 1]
}

# Instantiate the classifier: gbm
gbm = XGBClassifier()

# Perform random search: xgb_random
xgb_random = RandomizedSearchCV(param_distributions=gbm_param_grid,estimator=gbm, scoring="accuracy",verb


xgb_random.fit(X_train, y_train)


print("Best parameters found: ", xgb_random.best_params_)

print("Best accuracy found: ", xgb_random.best_score_)
```

```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
Best parameters found:  {'n_estimators': 16, 'max_depth': 9, 'learning_rate': 0.45, 'colsample_bytree':
0.6}
Best accuracy found:  0.6626912754083568
```

In [109]:
```
xgb = XGBClassifier(n_estimators=14, max_depth=9, learning_rate=0.45,colsample_bytree=0.9)
xgb.fit(X_train,y_train).score(X_train,y_train)
```

Out[109]:
```
0.6845799188174926
```

In [110]:
```
# Compare all the methods.
print(lr.score(X_test, y_test))
print(dtc.score(X_test, y_test))
print(xgb.score(X_test, y_test))
```

```
0.5962227736421285
0.6497932175351531
0.6655638268541494
```

In [111]:
```
#From the above values of accuracy of diffrent models, we get the best result when XGBclassifier utlized.
```