Term Project

Airlines On-Time Performance, Delays, Cancellations and Diversions

MILESTONE 1 - Data selection and EDA

Introduction: Airline cancellations or delays are one of the major causes of passenger inconvenience. With publicly available dataset, using data science, I am hoping to gain meaningful insights into the best-performing airlines and understand the causes of delays, diversions and cancellations across different airline carriers.

For the final project, I would like to analyze airline data to identify different factors and their effects on a carrier's performance. Using the available performance measures I would like to be able to predict the chances of a flight being on-time/delayed/cancelled.

Data Source: Excel files from BTS. The Excel data has airline performance factors such as cancelled, diverted, delayed and on-time data. The downloaded raw data has up to 34 columns. https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E (Download Raw Data link for data).

Problem statement addressed:

This study will benefit Customers as it will help predict a flights performance. Customers can lookup the chances of their flight reaching on-time during their booking or even before heading to the airport. Airlines can also benefit by comparing airline performances and predicting possibilities of delay based on aircraft/origin/destination and apply corrective measures to reduce cancellations and delays and improve on-time performance.

Data Transformation

In the data transformation step, I will be modifying the following:

1. Cancellation reason in the flight dataset is represented as A, B, C and D. I will be updating the cancellation code as follows:

   A Carrier

   B Weather

   C National Air System

   D Security

1. I will be adding a new column 'Status' with the status of a flight such as, On-Time, Delayed, Cancelled, Diverted.

2. Diverted column is of binary value which can be modified to a Yes/No

In [1]:
```
#Load necessary libraries
```

```python
import pandas as pd
import numpy as np
import plotly.express as px
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```python
#Read flight data from "https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E"

flight_data_df = pd.read_csv('T_ONTIME_MARKETING_May.csv')
flight_data_df.head()
```

Out[2]:

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	MKT_UNIQUE_CARRIER	OP_UNIQUE_CAI
0	2022	2	5	1	7	5/1/2022 12:00:00 AM	AA	
1	2022	2	5	1	7	5/1/2022 12:00:00 AM	AA	
2	2022	2	5	1	7	5/1/2022 12:00:00 AM	AA	
3	2022	2	5	1	7	5/1/2022 12:00:00 AM	AA	
4	2022	2	5	1	7	5/1/2022 12:00:00 AM	AA	

5 rows × 39 columns

1. DROP DUPLICATES

Duplicates cause inconsistent results when dealing with statistics. Hence dropping duplicate rows.

In [3]:

```python
print('Dataframe before dropping duplicates :', flight_data_df.shape)
flight_data_df = flight_data_df.drop_duplicates() # 1,389 rows dropped
print('Dataframe after dropping duplicates :',flight_data_df.shape)
```

```
Dataframe before dropping duplicates : (602950, 39)
Dataframe after dropping duplicates : (601561, 39)
```

2. Update Null values and Drop null rows, if any

Drop null rows, if any and update null values to 0 for delays

In [4]:

```python
#Drop null values
print('Dataframe before dropping null rows :', flight_data_df.shape)
flight_data_df.dropna()
print('Dataframe after dropping null rows :', flight_data_df.shape)
#Update null values to 0
flight_data_df.DISTANCE = flight_data_df.DISTANCE.fillna(0)
flight_data_df.DEP_DELAY = flight_data_df.DEP_DELAY.fillna(0)
flight_data_df.ARR_DELAY = flight_data_df.ARR_DELAY.fillna(0)
flight_data_df.CARRIER_DELAY = flight_data_df.CARRIER_DELAY.fillna(0)
flight_data_df.WEATHER_DELAY = flight_data_df.WEATHER_DELAY.fillna(0)
flight_data_df.NAS_DELAY = flight_data_df.NAS_DELAY.fillna(0)
flight_data_df.SECURITY_DELAY = flight_data_df.SECURITY_DELAY.fillna(0)
flight_data_df.LATE_AIRCRAFT_DELAY = flight_data_df.LATE_AIRCRAFT_DELAY.fillna(0)
```

```
Dataframe before dropping null rows : (601561, 39)
Dataframe after dropping null rows : (601561, 39)
```

In [5]:
```
flight_data_df.loc[pd.isna(flight_data_df.CANCELLATION_CODE), 'CANCELLATION_CODE']='Z'
```

3. Add new features

Cancellation code is represented as A, B, C and D, which is not very informative. The BTS website provided details on this code as follows:

In [6]:
```
flight_data_df['CANCELLATION_REASON'] = ''
flight_data_df.CANCELLATION_REASON = np.where(flight_data_df.CANCELLATION_CODE=='A', 'Ca
                                 np.where(flight_data_df.CANCELLATION_CODE=='B', 'Weathe
                                          np.where(flight_data_df.CANCELLATION_CODE=='C'
                                                   np.where(flight_data_df.CANCELLATION_
                                                             np.where(flight_data_df.CAN

flight_data_df.groupby(['CANCELLATION_REASON'])['CANCELLATION_REASON'].count().sort_inde
```

Out[6]:
```
CANCELLATION_REASON
Carrier                  4902
National Air System      1394
Not Cancelled          590957
Security                    1
Weather                  4307
Name: CANCELLATION_REASON, dtype: int64
```

Adding a new column 'STATUS' that tells the status of a flight

In [7]:
```
flight_data_df['STATUS'] = ''

flight_data_df.STATUS = np.where(flight_data_df.CANCELLED==1, 'Cancelled',
                                 np.where(flight_data_df.DIVERTED==1, 'Diverted',
                                          np.where(flight_data_df.ARR_DELAY<=15, 'On-Tim
                                                   np.where(flight_data_df.ARR_DELAY>15,
flight_data_df.groupby(['STATUS'])['STATUS'].count().sort_index()
```

Out[7]:
```
STATUS
Cancelled     10604
Delayed      119624
Diverted       1581
On-Time      469752
Name: STATUS, dtype: int64
```

Creating a new column 'ARR_DELAYED'. A flag that represents if a flight was delayed. Similar to CANCELLED and DIVERTED As a step to data reduction, I will be considering flights arriving 15 minutes or later as delayed

In [8]:
```
flight_data_df.loc[(flight_data_df['ARR_DELAY']>15), 'ARR_DELAYED'] = True
flight_data_df.loc[(flight_data_df['ARR_DELAY']<=15), 'ARR_DELAYED'] = False

flight_data_df.groupby(['ARR_DELAYED'])['ARR_DELAYED'].count().sort_index()
```

Out[8]:
```
ARR_DELAYED
False    481937
True     119624
Name: ARR_DELAYED, dtype: int64
```

Adding a new column 'DELAY_REASON' that tells the reason for a flight getting delayed

In [9]:
```
flight_data_df['DELAY_REASON'] = np.where(flight_data_df.CARRIER_DELAY != 0, 'Carrier',
                                          np.where(flight_data_df.LATE_AIRCRAFT_DELAY !=
                                                   np.where(flight_data_df.WEATHER_DELAY
```

```
                                            np.where(flight_data_df.NAS_
                                                 np.where(flight_dat

flight_data_df.groupby(['DELAY_REASON'])['DELAY_REASON'].count().sort_index()
```

Out[9]:

```
DELAY_REASON
                 477611
Carrier           74794
LateAircraft      26097
NAS               18695
Security            142
Weather            4222
Name: DELAY_REASON, dtype: int64
```

Data Visualization:

Implementing arithmetic functions for statistical analysis

Creating a new dataframe with total number of flights per operating carrier to calculate the %

In [10]:

```
flight_totals = flight_data_df.value_counts(subset=['OP_UNIQUE_CARRIER']).reset_index()
flight_totals_df = pd.DataFrame(flight_totals)
flight_totals_df.columns = ['OP_UNIQUE_CARRIER','TOTAL']
flight_totals_df['PERCENTAGE'] = round(flight_totals_df.TOTAL/flight_totals_df.TOTAL.sum

flight_totals_df = flight_totals_df.sort_values('PERCENTAGE',ascending=False)
flight_totals_df.head(5)
```

Out[10]:

	OP_UNIQUE_CARRIER	TOTAL	PERCENTAGE
0	WN	107950	17.94
1	DL	76021	12.64
2	AA	71471	11.88
3	OO	66615	11.07
4	UA	53535	8.90

Calculate percentage by carrier and flight status

In [11]:

```
flight_status = flight_data_df.value_counts(subset=['OP_UNIQUE_CARRIER','STATUS']).reset
flight_status_df = pd.DataFrame(flight_status) #create a dataframe
flight_status_df.columns = ['OP_UNIQUE_CARRIER','STATUS', 'COUNT'] #Add column names
flight_status_df = flight_status_df.sort_values('OP_UNIQUE_CARRIER') #Sort by operating

flight_status_df['PERCENTAGE'] = ''

for index, row in flight_status_df.iterrows():
    tot = flight_totals.loc[flight_totals.OP_UNIQUE_CARRIER==row.OP_UNIQUE_CARRIER].TOTA
    val = (row.COUNT/tot * 100)
    flight_status_df.at[index,'PERCENTAGE'] = round(val[0].astype(float),2) #Calculate t

flight_status_df.head(10)
```

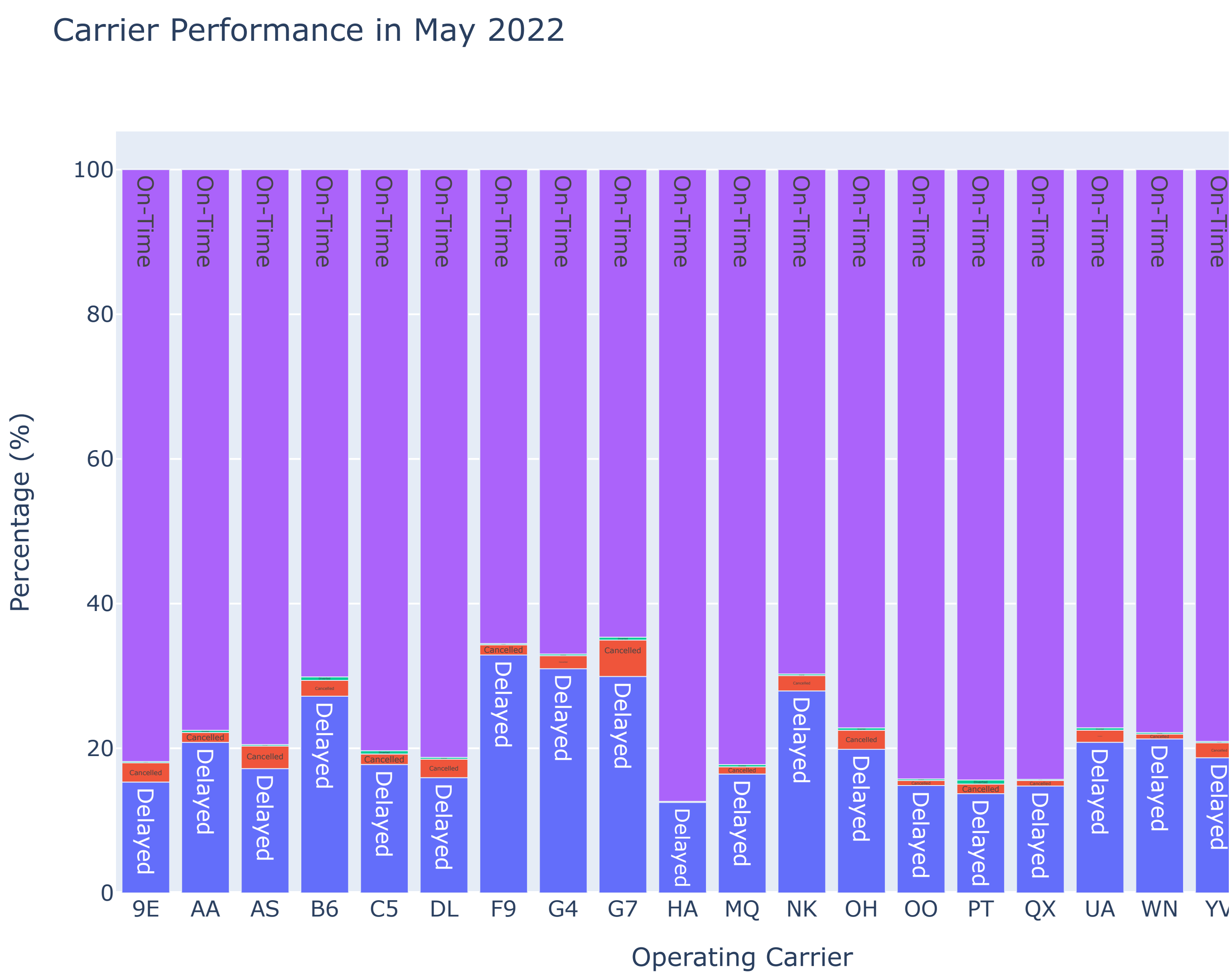
Out[11]:

	OP_UNIQUE_CARRIER	STATUS	COUNT	PERCENTAGE
33	9E	Delayed	3113	15.33
48	9E	Cancelled	542	2.67
74	9E	Diverted	35	0.17

8	9E	On-Time	16613	81.83
41	AA	Cancelled	973	1.36
56	AA	Diverted	215	0.3
3	AA	On-Time	55403	77.52
11	AA	Delayed	14880	20.82
47	AS	Cancelled	608	3.12
10	AS	On-Time	15502	79.49

Bar chart for carier performance in May 2022

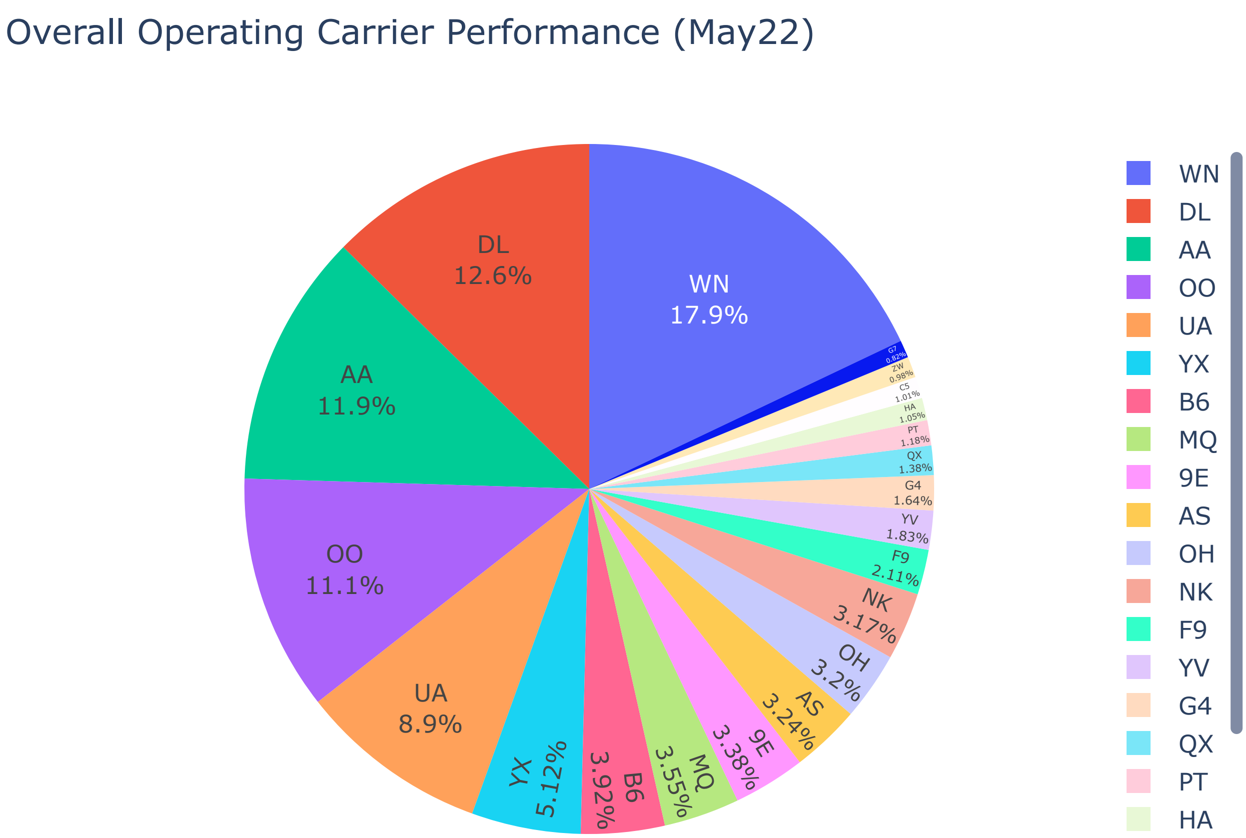
In [12]:
```
fig = px.bar(flight_status_df, x="OP_UNIQUE_CARRIER", y="PERCENTAGE", title="Carrier Per
             color="STATUS", text="STATUS",
             labels={"OP_UNIQUE_CARRIER": "Operating Carrier",
                     "PERCENTAGE": "Percentage (%)"})
fig.update_layout(autosize=False,width=900, height=600)
fig.show()
```



Hawaian airlines had the best on-time performance in May'22 followed by Air Wisconsin(ZW). Frontier airlines(F9) had the most number of delays at 32.9% GoJet had the most cancellations at 7%

Pie chart for Overall Carrier performance in May'22

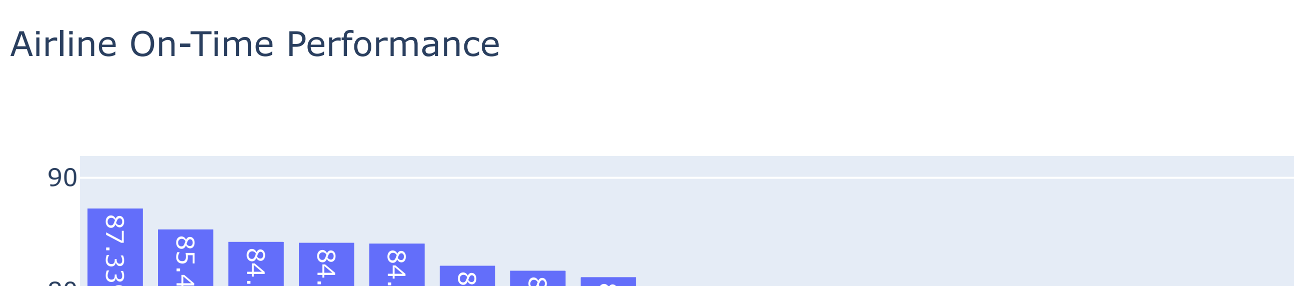
In [13]:
```
fig = px.pie(flight_totals_df, values='PERCENTAGE', names='OP_UNIQUE_CARRIER',
             title='Overall Operating Carrier Performance (May''22)')
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.show()
```

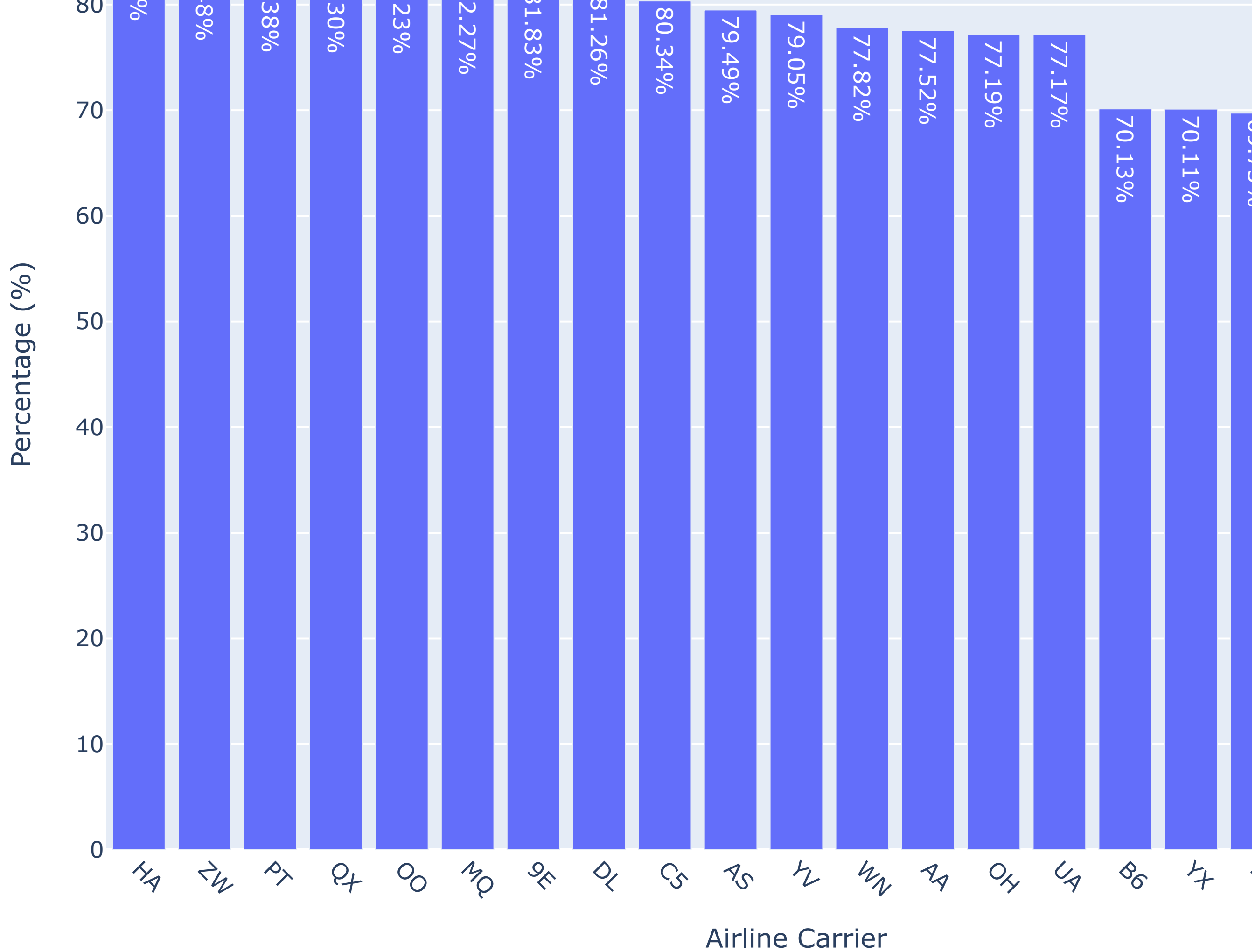


We can see southwest carrier (WN) had the most number of flights in May 2022.

Bar plot for Airline with best on-time performance

In [14]:
```
airline_on_time_performance = flight_status_df[flight_status_df.STATUS == 'On-Time'].sor

fig=px.bar(airline_on_time_performance,
           x=airline_on_time_performance.OP_UNIQUE_CARRIER,
           y=airline_on_time_performance.PERCENTAGE, title="Airline On-Time Performance"
           text=airline_on_time_performance.PERCENTAGE.apply(lambda x: '{0:1.2f}%'.forma
           labels=dict(OP_UNIQUE_CARRIER="Airline Carrier", PERCENTAGE="Percentage (%)")
fig.update_xaxes(tickangle=45)
fig.update_layout(autosize=False,width=900, height=700)
```





Hawaiian airline was the best performing airline in May'22 with 87.33% on time performance and Go-Jet is the least performing airline with 64.6% on-time performance.

In [15]:
```
#Load csv file with airport names for origin and destination
airport_data_df = pd.read_csv('L_AIRPORT.csv')
airport_data_df.head()
```

Out[15]:

	Code	Description
0	01A	Afognak Lake, AK: Afognak Lake Airport
1	03A	Granite Mountain, AK: Bear Creek Mining Strip
2	04A	Lik, AK: Lik Mining Camp
3	05A	Little Squaw, AK: Little Squaw Airport
4	06A	Kizhuyak, AK: Kizhuyak Bay

In [16]:
```
#Create a new dataframe with the percentage by origin airport and status
flight_origin_totals = flight_data_df.value_counts(subset=['ORIGIN']).reset_index()   #ge
flight_origin_totals_df = pd.DataFrame(flight_origin_totals)#create a dataframe
flight_origin_totals_df.columns = ['ORIGIN','TOTAL']#Add column names
#Calculate the percentage by origin airport
flight_origin_totals_df['PERCENTAGE'] = round(flight_origin_totals_df.TOTAL/flight_origi


origin_airport_delays = flight_data_df.value_counts(subset=['ORIGIN','STATUS']).reset_in
origin_airport_df = pd.DataFrame(origin_airport_delays) #create a dataframe
origin_airport_df.columns = ['ORIGIN','STATUS', 'COUNT'] #add column names
```

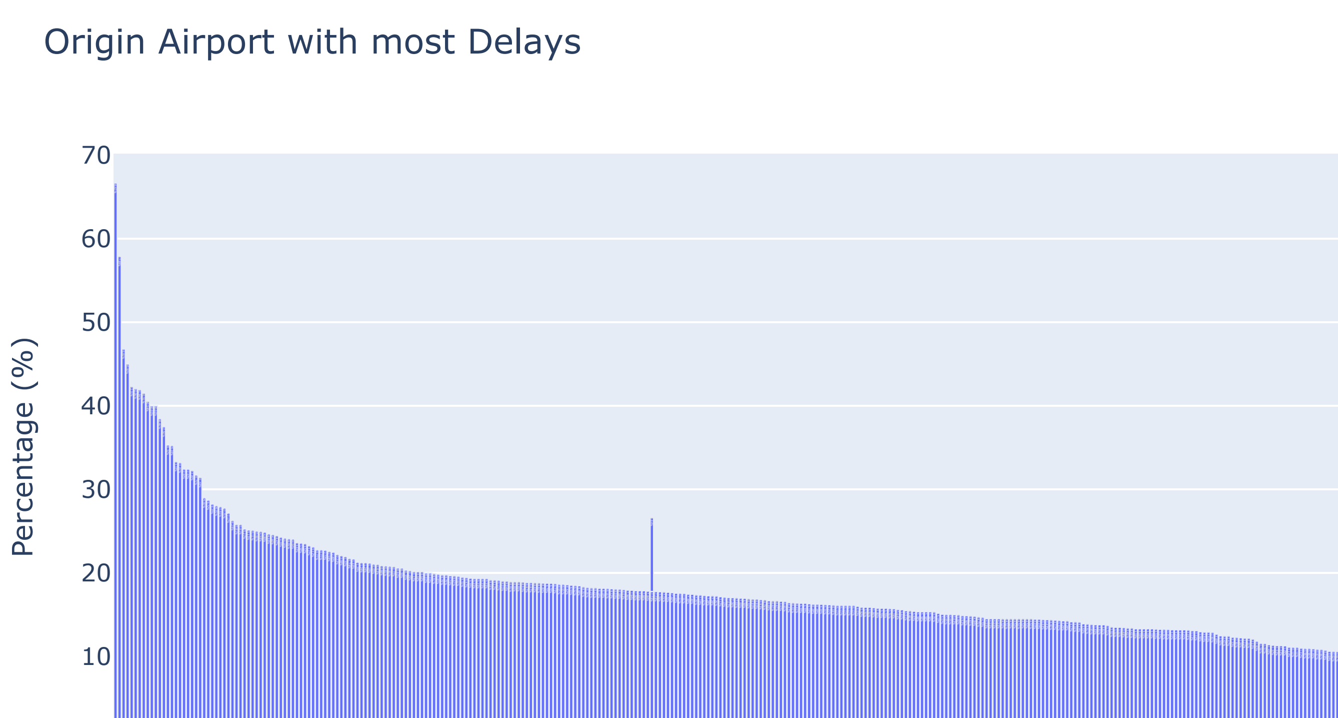
```
origin_airport_df = origin_airport_df.sort_values('ORIGIN') #sort by origin
origin_airport_df['PERCENTAGE'] = ''

for index, row in origin_airport_df.iterrows():
    tot = flight_origin_totals.loc[flight_origin_totals.ORIGIN==row.ORIGIN].TOTAL.values
    val = (row.COUNT/tot * 100)
    origin_airport_df.at[index,'PERCENTAGE'] = round(val[0].astype(float),2)  #calulate

origin_airport_df.head(10)
origin_airport_df = origin_airport_df.sort_values('PERCENTAGE',ascending=False) #sort by

#Add the airport name from the airport_data_df and add as a new column to the origin_air
origin_airport_df=pd.merge(origin_airport_df, airport_data_df, how='left', left_on='ORIG
origin_airport_df.rename(columns={'Description':'ORIGIN_AIRPORT_NAME'}, inplace=True)
del origin_airport_df['Code']

new = origin_airport_df.ORIGIN_AIRPORT_NAME.str.split(":", n = 1, expand = True)
origin_airport_df["ORIGIN_AIRPORT_NAME"] =  new[1]
origin_airport_df.head()
```

Out[16]:

	ORIGIN	STATUS	COUNT	PERCENTAGE	ORIGIN_AIRPORT_NAME
0	GST	On-Time	12	100.0	Gustavus Airport
1	STC	On-Time	1	100.0	St. Cloud Regional
2	LWS	On-Time	95	96.94	Lewiston Nez Perce County
3	BGM	On-Time	30	96.77	Greater Binghamton/Edwin A. Link Field
4	DRT	On-Time	60	96.77	Del Rio International

Bar chart for Origin airport with most delays

In [17]:
```
fig = px.bar(origin_airport_df[origin_airport_df.STATUS=="Delayed"], x="ORIGIN_AIRPORT_N
             title="Origin Airport with most Delays",
             text=origin_airport_df[origin_airport_df.STATUS=="Delayed"].PERCENTAGE.appl
             labels=dict(ORIGIN_AIRPORT_NAME="Origin Airport", PERCENTAGE="Percentage (%
fig.update_xaxes(tickangle=80)
fig.update_layout(autosize=False,width=900, height=700)
fig.show()
```



It appears Tri Cities has multiple entries for different origin airports. Identify and update the airport name.

In [18]:
```
origin_airport_df[origin_airport_df.ORIGIN_AIRPORT_NAME.str.contains('Tri Cities')]
```

Out[18]:

	ORIGIN	STATUS	COUNT	PERCENTAGE	ORIGIN_AIRPORT_NAME
29	PSC	On-Time	451	90.56	Tri Cities
207	TRI	On-Time	302	81.4	Tri Cities
506	TRI	Delayed	66	17.79	Tri Cities
708	PSC	Delayed	44	8.84	Tri Cities
1018	TRI	Cancelled	3	0.81	Tri Cities
1093	PSC	Diverted	2	0.4	Tri Cities
1178	PSC	Cancelled	1	0.2	Tri Cities

Updating the airport name for PSC

In [19]:
```
origin_airport_df.loc[origin_airport_df["ORIGIN"] == "PSC", "ORIGIN_AIRPORT_NAME"] = 'Tr
```

In [20]:
```
origin_airport_df[origin_airport_df.ORIGIN_AIRPORT_NAME.str.contains('Tri Cities')]
```

Out[20]:

	ORIGIN	STATUS	COUNT	PERCENTAGE	ORIGIN_AIRPORT_NAME
29	PSC	On-Time	451	90.56	Tri Cities(PSC)
207	TRI	On-Time	302	81.4	Tri Cities
506	TRI	Delayed	66	17.79	Tri Cities
708	PSC	Delayed	44	8.84	Tri Cities(PSC)
1018	TRI	Cancelled	3	0.81	Tri Cities
1093	PSC	Diverted	2	0.4	Tri Cities(PSC)
1178	PSC	Cancelled	1	0.2	Tri Cities(PSC)

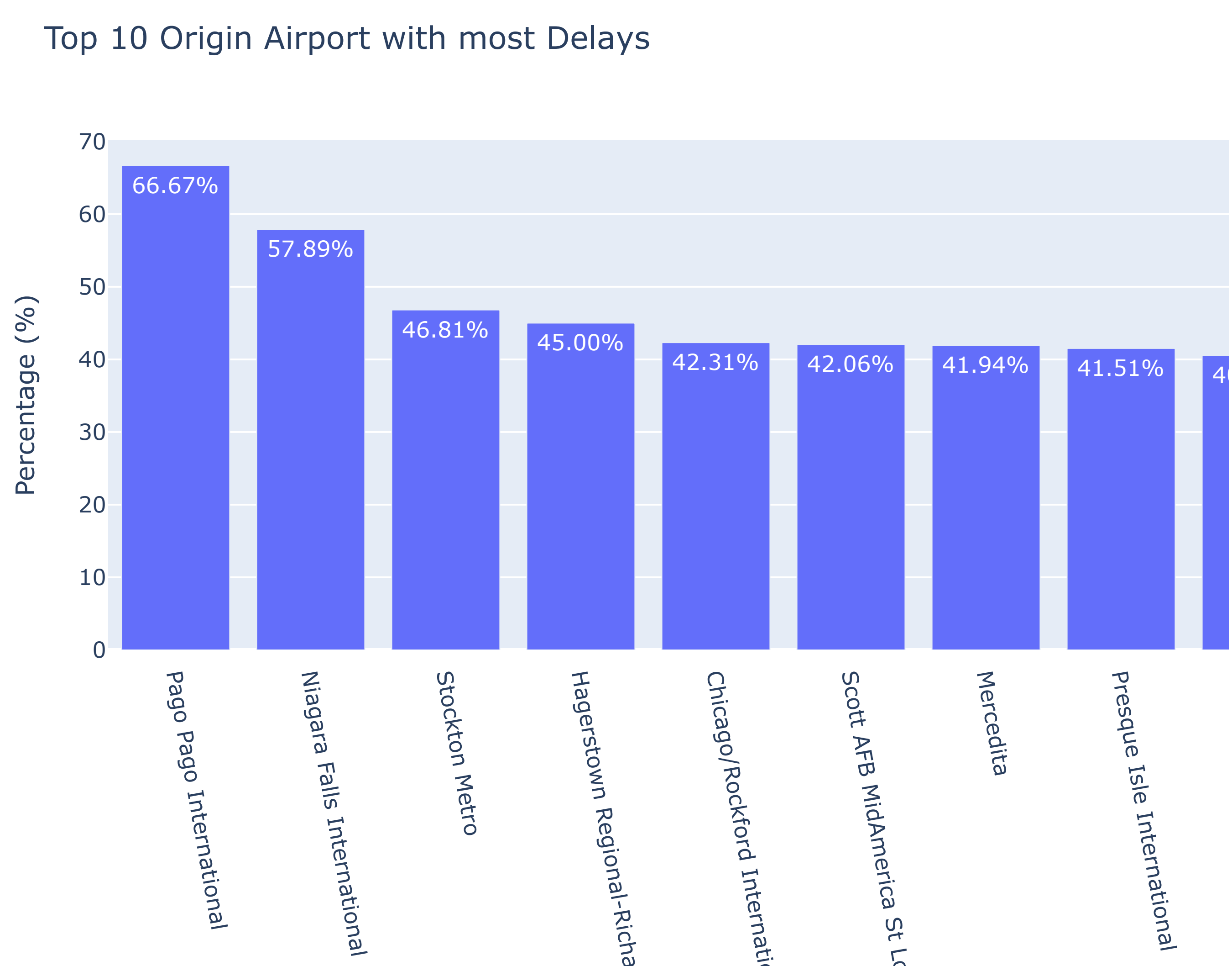
Since the chart has many airports to fit, filtering the list to get the top 10 origin airports with most delays

In [21]:
```
top_10_origin_delay_airports = origin_airport_df[origin_airport_df.STATUS=='Delayed'].he
top_10_origin_delay_airports
```

Out[21]:

	ORIGIN	STATUS	COUNT	PERCENTAGE	ORIGIN_AIRPORT_NAME
344	PPG	Delayed	2	66.67	Pago Pago International
361	IAG	Delayed	22	57.89	Niagara Falls International
368	SCK	Delayed	22	46.81	Stockton Metro
370	HGR	Delayed	9	45.0	Hagerstown Regional-Richard A. Henson Field
371	RFD	Delayed	22	42.31	Chicago/Rockford International
373	BLV	Delayed	45	42.06	Scott AFB MidAmerica St Louis
374	PSE	Delayed	39	41.94	Mercedita
375	PQI	Delayed	22	41.51	Presque Isle International
376	USA	Delayed	30	40.54	Concord Padgett Regional
378	RIW	Delayed	14	40.0	Central Wyoming Regional

In [22]:
```
fig = px.bar(top_10_origin_delay_airports[top_10_origin_delay_airports.STATUS=="Delayed"
             title="Top 10 Origin Airport with most Delays",
             text=top_10_origin_delay_airports[top_10_origin_delay_airports.STATUS=="Del
             labels=dict(ORIGIN_AIRPORT_NAME="Origin Airport", PERCENTAGE="Percentage (%
fig.update_xaxes(tickangle=80)
fig.update_layout(autosize=False,width=900, height=700)
fig.show()
```



Flights originating from Pago Pago International are delayed 66.67%

DESTINATION

In [23]:
```
#Create a new dataframe with the percentage by origin airport and status
flight_dest_totals = flight_data_df.value_counts(subset=['DEST']).reset_index() #get the
flight_dest_totals_df = pd.DataFrame(flight_dest_totals)#create a dataframe
flight_dest_totals_df.columns = ['DEST','TOTAL']#Add column names
#Calculate the percentage by destination airport
flight_dest_totals_df['PERCENTAGE'] = round(flight_dest_totals_df.TOTAL/flight_dest_tota


dest_airport_delays = flight_data_df.value_counts(subset=['DEST','STATUS']).reset_index(
dest_airport_df = pd.DataFrame(dest_airport_delays) #create a dataframe
dest_airport_df.columns = ['DEST','STATUS', 'COUNT']#add column names
dest_airport_df = dest_airport_df.sort_values('DEST')#sort by destination
dest_airport_df['PERCENTAGE'] = ''

for index, row in dest_airport_df.iterrows():
    tot = flight_dest_totals.loc[flight_dest_totals.DEST==row.DEST].TOTAL.values  #get t
    val = (row.COUNT/tot * 100)
    dest_airport_df.at[index,'PERCENTAGE'] = round(val[0].astype(float),2) #calulate the

dest_airport_df.head(10)
dest_airport_df = dest_airport_df.sort_values('PERCENTAGE',ascending=False)#sort by perc

#Add the airport name from the airport_data_df and add as a new column to the dest_airpo
dest_airport_df=pd.merge(dest_airport_df, airport_data_df, how='left', left_on='DEST', r
dest_airport_df.rename(columns={'Description':'DEST_AIRPORT_NAME'}, inplace=True)
del dest_airport_df['Code']

new = dest_airport_df.DEST_AIRPORT_NAME.str.split(":", n = 1, expand = True)
dest_airport_df["DEST_AIRPORT_NAME"] =  new[1]
dest_airport_df.head()
```

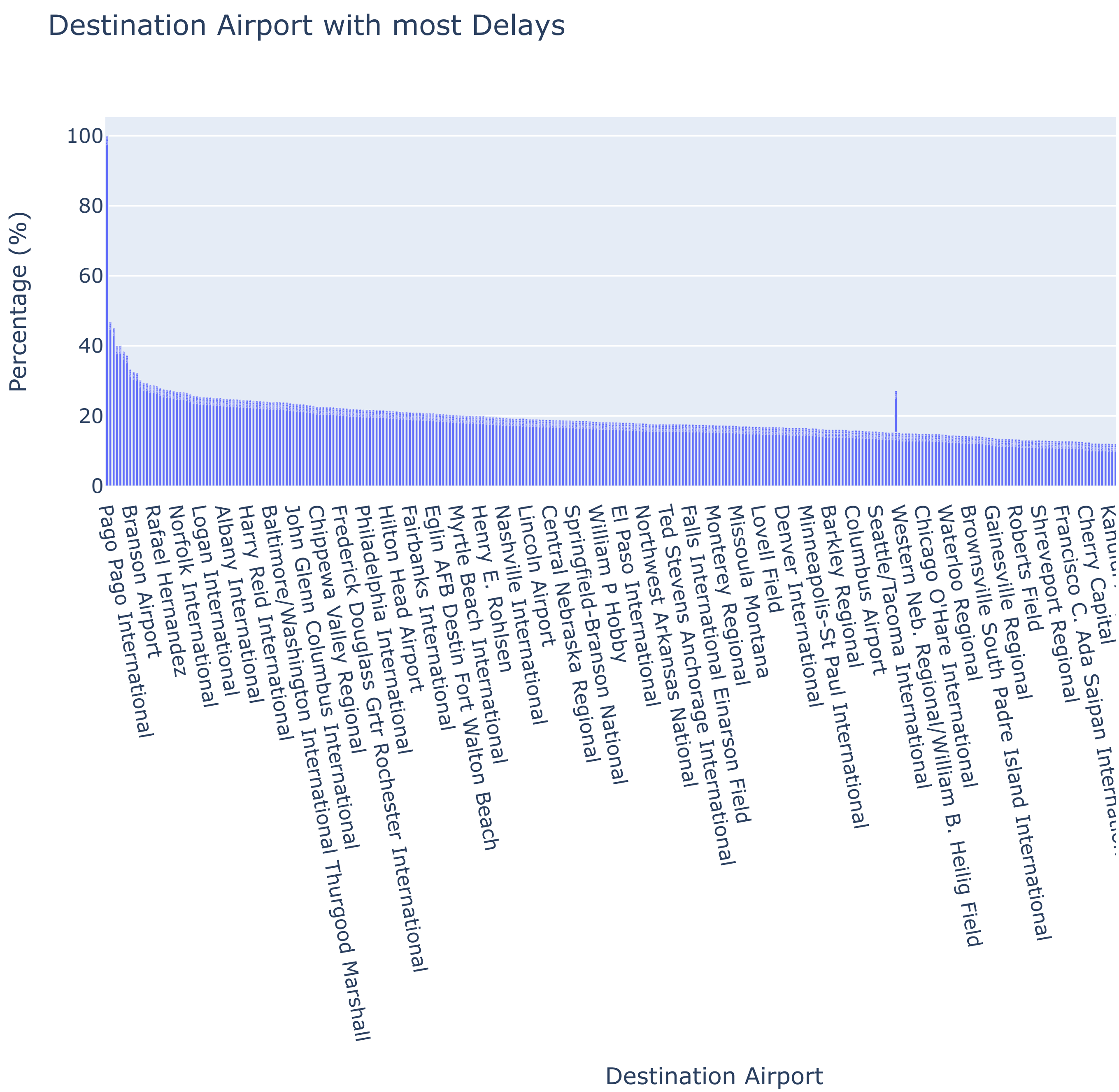
Out[23]:

	DEST	STATUS	COUNT	PERCENTAGE	DEST_AIRPORT_NAME
0	GST	On-Time	12	100.0	Gustavus Airport
1	STC	On-Time	1	100.0	St. Cloud Regional
2	PPG	Delayed	3	100.0	Pago Pago International
3	TWF	On-Time	31	96.88	Joslin Field - Magic Valley Regional
4	PIH	On-Time	30	96.77	Pocatello Regional

Bar chart for Destination Airports with most delays

In [24]:
```
fig = px.bar(dest_airport_df[dest_airport_df.STATUS=="Delayed"], x="DEST_AIRPORT_NAME",
             title="Destination Airport with most Delays",
             text=dest_airport_df[dest_airport_df.STATUS=="Delayed"].PERCENTAGE.apply(la
             labels=dict(DEST_AIRPORT_NAME="Destination Airport", PERCENTAGE="Percentage
fig.update_xaxes(tickangle=80)
```

```
fig.update_layout(autosize=False,width=900, height=700)
fig.show()
```



Updating Destination name for PSC

In [25]:
```
dest_airport_df[dest_airport_df.DEST_AIRPORT_NAME.str.contains('Tri Cities')]
dest_airport_df.loc[dest_airport_df["DEST"] == "PSC", "DEST_AIRPORT_NAME"] = 'Tri Cities
dest_airport_df[dest_airport_df.DEST_AIRPORT_NAME.str.contains('Tri Cities')]
```

Out[25]:

	DEST	STATUS	COUNT	PERCENTAGE	DEST_AIRPORT_NAME
40	PSC	On-Time	439	88.15	Tri Cities(PSC)
125	TRI	On-Time	310	83.56	Tri Cities
610	TRI	Delayed	57	15.36	Tri Cities
680	PSC	Delayed	59	11.85	Tri Cities(PSC)
985	TRI	Cancelled	4	1.08	Tri Cities

Since the chart has many airports to fit, filtering the list to get the top 10 destination airports with most delays

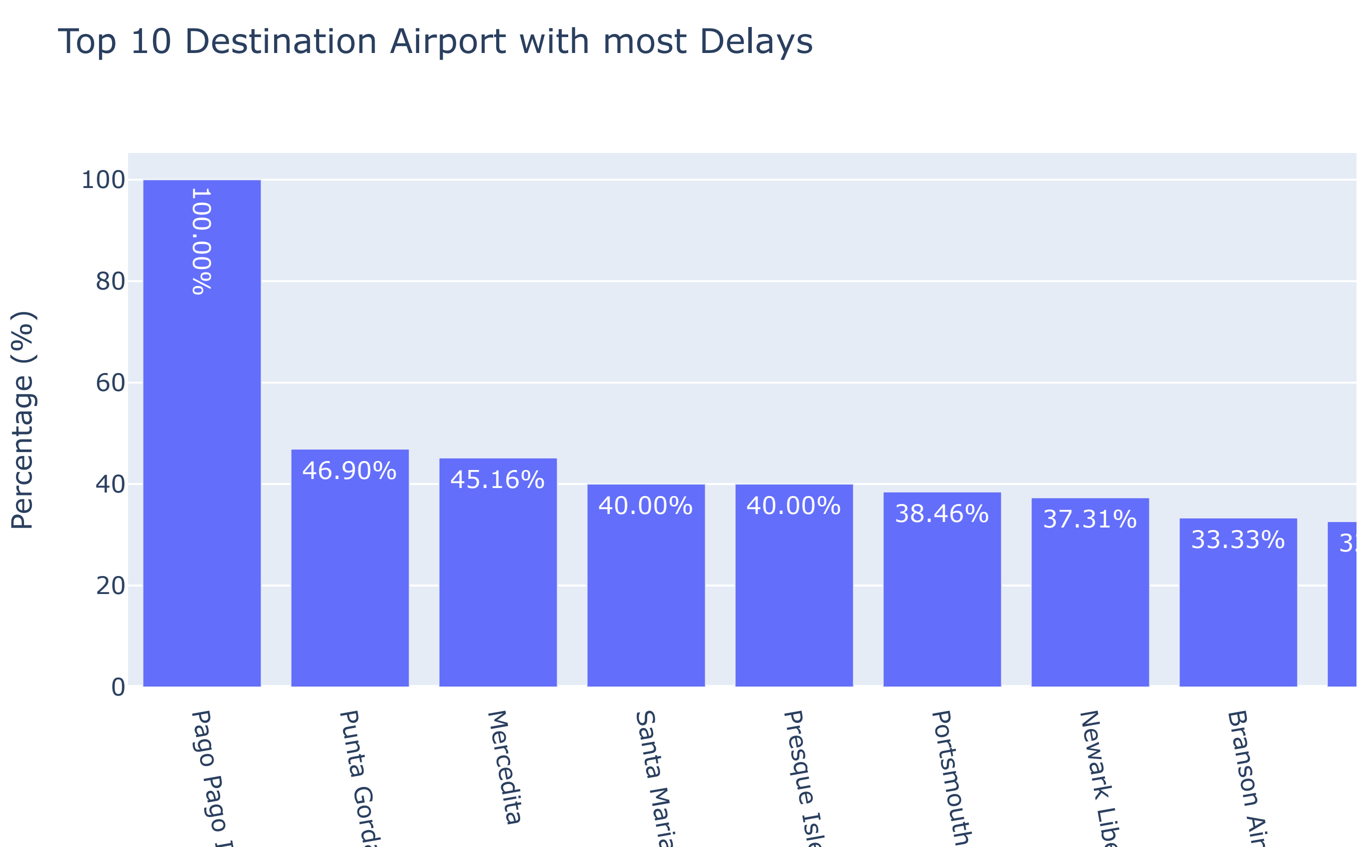
In [26]:
```
top_10_dest_delay_airports = dest_airport_df[dest_airport_df.STATUS=='Delayed'].head(10)
top_10_dest_delay_airports
```

Out[26]:

	DEST	STATUS	COUNT	PERCENTAGE	DEST_AIRPORT_NAME
2	PPG	Delayed	3	100.0	Pago Pago International
368	PGD	Delayed	212	46.9	Punta Gorda Airport
369	PSE	Delayed	42	45.16	Mercedita
372	SMX	Delayed	4	40.0	Santa Maria Public/Capt. G. Allan Hancock Field
373	PQI	Delayed	20	40.0	Presque Isle International
374	PSM	Delayed	10	38.46	Portsmouth International at Pease
375	EWR	Delayed	5097	37.31	Newark Liberty International
376	BKG	Delayed	3	33.33	Branson Airport
377	SCK	Delayed	15	32.61	Stockton Metro
378	USA	Delayed	24	32.43	Concord Padgett Regional

Bar chart for Destination Airprot with most delays

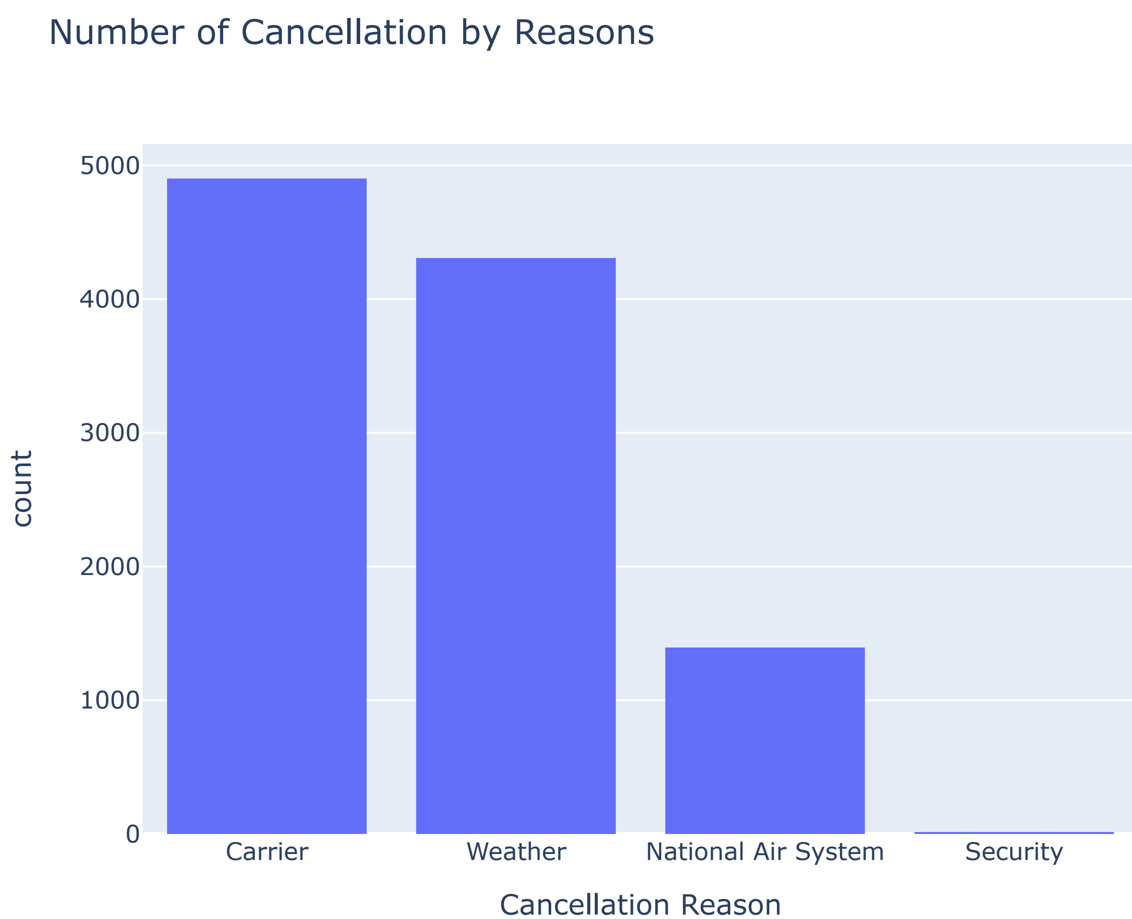
In [27]:
```
fig = px.bar(top_10_dest_delay_airports[top_10_dest_delay_airports.STATUS=="Delayed"], x
             title="Top 10 Destination Airport with most Delays",
             text=top_10_dest_delay_airports[top_10_dest_delay_airports.STATUS=="Delayed
             labels=dict(DEST_AIRPORT_NAME="Destination Airport", PERCENTAGE="Percentage
fig.update_xaxes(tickangle=80)
fig.update_layout(autosize=False,width=900, height=700)
fig.show()
```



All flights flying into Pago Pago Internation airport hasare delayed.

Histogram for Overall cancellations by cancellation reason

In [28]:
```
#Using CANCELLED==1 to filter the dataframe for only cancelled rows.
fig = px.histogram(flight_data_df[flight_data_df.CANCELLED==1], x="CANCELLATION_REASON",
                   title="Number of Cancellation by Reasons",
                   labels=dict(CANCELLATION_REASON="Cancellation Reason"))
fig.show()
```



From the chart, we can see that most cancellations in May'22 were due to carriers followed by weather

MILESTONE 2 - Data Preparation

Dataframe at this time has most of the required features.

In [29]:
```
flight_data_df.columns
```

Out[29]:
```
Index(['YEAR', 'QUARTER', 'MONTH', 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'FL_DATE',
       'MKT_UNIQUE_CARRIER', 'OP_UNIQUE_CARRIER', 'ORIGIN_AIRPORT_ID',
       'ORIGIN', 'ORIGIN_CITY_NAME', 'ORIGIN_STATE_ABR', 'ORIGIN_STATE_NM',
       'ORIGIN_WAC', 'DEST_AIRPORT_ID', 'DEST', 'DEST_CITY_NAME',
       'DEST_STATE_ABR', 'DEST_STATE_NM', 'DEST_WAC', 'DEP_DELAY',
       'DEP_DELAY_NEW', 'TAXI_OUT', 'TAXI_IN', 'ARR_TIME', 'ARR_DELAY',
       'ARR_DELAY_NEW', 'CANCELLED', 'CANCELLATION_CODE', 'DIVERTED',
       'ACTUAL_ELAPSED_TIME', 'AIR_TIME', 'FLIGHTS', 'DISTANCE',
       'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY',
       'LATE_AIRCRAFT_DELAY', 'CANCELLATION_REASON', 'STATUS', 'ARR_DELAYED',
       'DELAY_REASON'],
      dtype='object')
```

1. Drop any features that are not useful for your model building and explain why they are not useful.

Dropping null rows was performed in Milestone 1 - Data Transformation (Step 1). Some additional features that can be dropped are as follows.

i. Dropping the year, month, day, day of month as these details are in the FL_DATE feature which is in a DateTime format.

In [30]:
```
flight_data_df.drop(['YEAR', 'QUARTER', 'MONTH'], axis=1, inplace=True)
```

In [31]:
```
flight_data_df.head(5)
```

Out[31]:

	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	MKT_UNIQUE_CARRIER	OP_UNIQUE_CARRIER	ORIGIN_AIRPORT_ID
0	1	7	5/1/2022 12:00:00 AM	AA	AA	10140
1	1	7	5/1/2022 12:00:00 AM	AA	AA	10140
2	1	7	5/1/2022 12:00:00 AM	AA	AA	10140
3	1	7	5/1/2022 12:00:00 AM	AA	AA	10140
4	1	7	5/1/2022 12:00:00 AM	AA	AA	10140

5 rows × 40 columns

ii. Dropping Origin Airport ID, Destination Airport ID, Origin_WAC, DEST_WAC as these are not significant for this project.

In [32]:
```
flight_data_df.drop(['ORIGIN_WAC', 'DEST_WAC', 'ORIGIN_AIRPORT_ID', 'DEST_AIRPORT_ID'],
```

In [33]:
```
flight_data_df.columns
```

Out[33]:
```
Index(['DAY_OF_MONTH', 'DAY_OF_WEEK', 'FL_DATE', 'MKT_UNIQUE_CARRIER',
       'OP_UNIQUE_CARRIER', 'ORIGIN', 'ORIGIN_CITY_NAME', 'ORIGIN_STATE_ABR',
       'ORIGIN_STATE_NM', 'DEST', 'DEST_CITY_NAME', 'DEST_STATE_ABR',
       'DEST_STATE_NM', 'DEP_DELAY', 'DEP_DELAY_NEW', 'TAXI_OUT', 'TAXI_IN',
       'ARR_TIME', 'ARR_DELAY', 'ARR_DELAY_NEW', 'CANCELLED',
       'CANCELLATION_CODE', 'DIVERTED', 'ACTUAL_ELAPSED_TIME', 'AIR_TIME',
       'FLIGHTS', 'DISTANCE', 'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY',
       'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY', 'CANCELLATION_REASON',
       'STATUS', 'ARR_DELAYED', 'DELAY_REASON'],
      dtype='object')
```

2. Perform any data extraction/selection steps.

Most of the data extraction and selection steps are already performed in Milestone 1 before plotting. I will update this section should there be any additional extraction/selection steps required.

3. Transform features if necessary.

i. Replacing values in a column

I realised the below step is not required for this project. Hence, commenting it.

Day of week is mentioned as numbers, updating numbers to days of week.

flight_data_df.DAY_OF_WEEK = np.where(flight_data_df.DAY_OF_WEEK==1, 'Sunday', np.where(flight_data_df.DAY_OF_WEEK==2, 'Monday', np.where(flight_data_df.DAY_OF_WEEK==3, 'Tuesday', np.where(flight_data_df.DAY_OF_WEEK==4, 'Wednesday', np.where(flight_data_df.DAY_OF_WEEK==5, 'Thursday', np.where(flight_data_df.DAY_OF_WEEK==6, 'Friday', np.where(flight_data_df.DAY_OF_WEEK==7, 'Wednesday','')))))))
flight_data_df.groupby(['DAY_OF_WEEK'])['DAY_OF_WEEK'].count().sort_index()

Arrival and departure delays in the '*_New' column for flights departing/arriving earlier than schedule are updated to 0. For this project we are considering flights arriving 15 minutes or later as delayed. Updating arr_delay to 0 for 15 minutes or less.

In [34]:
```
flight_data_df.loc[flight_data_df.ARR_DELAY<=15, 'ARR_DELAY'] = 0
flight_data_df.ARR_DELAY.unique()
```

Out[34]:
```
array([   0.,    17.,    20., ..., 1830.,   658.,   651.])
```

4. Engineer new useful features.

Features such as DELAY_REASON,ARR_DELAYED, CANCELLATION_REASON and STATUS were engineered and created in Milestone 1, data transformation section. These features were required for plotting.

5. Deal with missing data (do not just drop rows or columns without justifying this).

In [35]:
```
flight_data_df_null = flight_data_df.isnull().sum() / len(flight_data_df) #Calculate % o
flight_data_df_null
```

Out[35]:
```
DAY_OF_MONTH           0.000000
DAY_OF_WEEK            0.000000
FL_DATE                0.000000
MKT_UNIQUE_CARRIER     0.000000
OP_UNIQUE_CARRIER      0.000000
ORIGIN                 0.000000
ORIGIN_CITY_NAME       0.000000
ORIGIN_STATE_ABR       0.000000
```

```
ORIGIN_STATE_NM          0.000000
DEST                     0.000000
DEST_CITY_NAME           0.000000
DEST_STATE_ABR           0.000000
DEST_STATE_NM            0.000000
DEP_DELAY                0.000000
DEP_DELAY_NEW            0.016958
TAXI_OUT                 0.017551
TAXI_IN                  0.017902
ARR_TIME                 0.017902
ARR_DELAY                0.000000
ARR_DELAY_NEW            0.020256
CANCELLED                0.000000
CANCELLATION_CODE        0.000000
DIVERTED                 0.000000
ACTUAL_ELAPSED_TIME      0.020256
AIR_TIME                 0.020256
FLIGHTS                  0.000000
DISTANCE                 0.000000
CARRIER_DELAY            0.000000
WEATHER_DELAY            0.000000
NAS_DELAY                0.000000
SECURITY_DELAY           0.000000
LATE_AIRCRAFT_DELAY      0.000000
CANCELLATION_REASON      0.000000
STATUS                   0.000000
ARR_DELAYED              0.000000
DELAY_REASON             0.000000
dtype: float64
```

In [36]:

```
#Look for any features with over 40% missing data
missing_features = flight_data_df_null[flight_data_df_null > 0.40].index #Identify colum
missing_features
```

Out[36]:

```
Index([], dtype='object')
```

There are no features with missing data over 40%.

6. Create dummy variables if necessary.

At this point, I don't belive I need dummy variables. Will revisit this section based on outcomes from Model creations.

Conclusion

It appears we have enough information at this time to apply different models on the dataframe to be able to predict a flights performance. From the different charts, we can also see the best performing carriers, airports with most delays for arrivals and departures and reasons for cancellations.

On futher analysis, it appears we do not need any external data at this time for the model prediction. I target on predicting which airline and/or airport are more likely to be delayed based on the arrival delay feature.

MILESTONE 3 - Model building and Evaluation

In [37]:

```
from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree

from sklearn.metrics import accuracy_score, roc_curve, roc_auc_score,confusion_matrix, c
from imblearn.over_sampling import SMOTE
from imblearn.combine import SMOTETomek
from sklearn.model_selection import train_test_split, cross_val_score,KFold

from IPython.display import Image

import warnings
from pandas.errors import SettingWithCopyWarning
```

In [38]:
```
warnings.simplefilter(action="ignore", category=SettingWithCopyWarning)
```

In [39]:
```
flight_data_df.head(5)
```

Out[39]:

	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	MKT_UNIQUE_CARRIER	OP_UNIQUE_CARRIER	ORIGIN	ORIGIN_CITY
0	1	7	5/1/2022 12:00:00 AM	AA	AA	ABQ	Albuquerq
1	1	7	5/1/2022 12:00:00 AM	AA	AA	ABQ	Albuquerq
2	1	7	5/1/2022 12:00:00 AM	AA	AA	ABQ	Albuquerq
3	1	7	5/1/2022 12:00:00 AM	AA	AA	ABQ	Albuquerq
4	1	7	5/1/2022 12:00:00 AM	AA	AA	ABQ	Albuquerq

5 rows × 36 columns

In [40]:
```
#Select columns for modeling
model_df = flight_data_df[(flight_data_df.STATUS=='On-Time')|(flight_data_df.STATUS=='De
flight_model_df = model_df[['DAY_OF_WEEK','OP_UNIQUE_CARRIER', 'ORIGIN','DEST', 'DEP_DEL
                            'FLIGHTS', 'DISTANCE', 'CARRIER_DELAY', 'WEATHER_DELAY', 'NA
                            'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY', 'STATUS']]


flight_model_df.head(5)
```

Out[40]:

	DAY_OF_WEEK	OP_UNIQUE_CARRIER	ORIGIN	DEST	DEP_DELAY	ARR_DELAY	FLIGHTS	DISTANCE	CARRIER_D
0	7	AA	ABQ	DFW	-9.0	0.0	1.0	569.0	
1	7	AA	ABQ	DFW	-8.0	0.0	1.0	569.0	
2	7	AA	ABQ	DFW	-5.0	0.0	1.0	569.0	
3	7	AA	ABQ	DFW	-5.0	0.0	1.0	569.0	
4	7	AA	ABQ	DFW	0.0	0.0	1.0	569.0	

In [41]:

```
#Check the data types for features to split into categorical and numerical fields
#for standaridizing and creating dummy variables
flight_model_df.dtypes
```

Out[41]:

```
DAY_OF_WEEK              int64
OP_UNIQUE_CARRIER       object
ORIGIN                  object
DEST                    object
DEP_DELAY              float64
ARR_DELAY              float64
FLIGHTS                float64
DISTANCE               float64
CARRIER_DELAY          float64
WEATHER_DELAY          float64
NAS_DELAY              float64
SECURITY_DELAY         float64
LATE_AIRCRAFT_DELAY    float64
STATUS                  object
dtype: object
```

In [53]:

```
#Converting data types to avoid memory issues while executing the model fit.
cols=['DEP_DELAY','ARR_DELAY','FLIGHTS','DISTANCE','CARRIER_DELAY','WEATHER_DELAY','NAS_
flight_model_df[cols] = flight_model_df[cols].astype('float16') #Converting float64 to f

flight_model_df['DAY_OF_WEEK'] = flight_model_df['DAY_OF_WEEK'].astype(np.uint8)   #Conve

obj_cols = ['OP_UNIQUE_CARRIER','ORIGIN','DEST']
flight_model_df[obj_cols] = flight_model_df[obj_cols].astype('category') # #Converting o
```

In [43]:

```
#Filtering categorical and numeric fields
X_cat = flight_model_df[['OP_UNIQUE_CARRIER', 'ORIGIN', 'DEST']]
X_num = flight_model_df.drop(['OP_UNIQUE_CARRIER', 'ORIGIN','DEST','STATUS'], axis=1)
```

In [44]:

```
#Intial Status was created for all types, cancelled, diverted, delayed and on-time. For 
#to on-time and delayed, I'm dropping this column and recreating the same with 0's and 1
flight_model_df.drop(columns = ['STATUS'], axis = 1, inplace = True)
```

In [45]:

```
flight_model_df.loc[flight_model_df.ARR_DELAY <= 15, 'STATUS'] = 0
flight_model_df.loc[flight_model_df.ARR_DELAY > 15, 'STATUS'] = 1
```

In [46]:

```
flight_model_df.groupby(['STATUS'])['STATUS'].count().sort_index()
```

Out[46]:

```
STATUS
0.0    469752
1.0    119624
Name: STATUS, dtype: int64
```

In [47]:

```
#Creating dummy variables for categorical columns
X_cat = pd.get_dummies(X_cat, drop_first=True)
```

In [48]:

```
# Create standardizer
scaler = StandardScaler()

# Standardize features
scaler.fit(X_num)
X_scaled = scaler.transform(X_num)
X_scaled = pd.DataFrame(X_scaled, index=X_num.index, columns=X_num.columns)
```

In [49]:

```
#Concatenate the categorical and scaled numeric fields and set the features
X = pd.concat([X_scaled, X_cat], axis=1)
X.isna().sum()
```

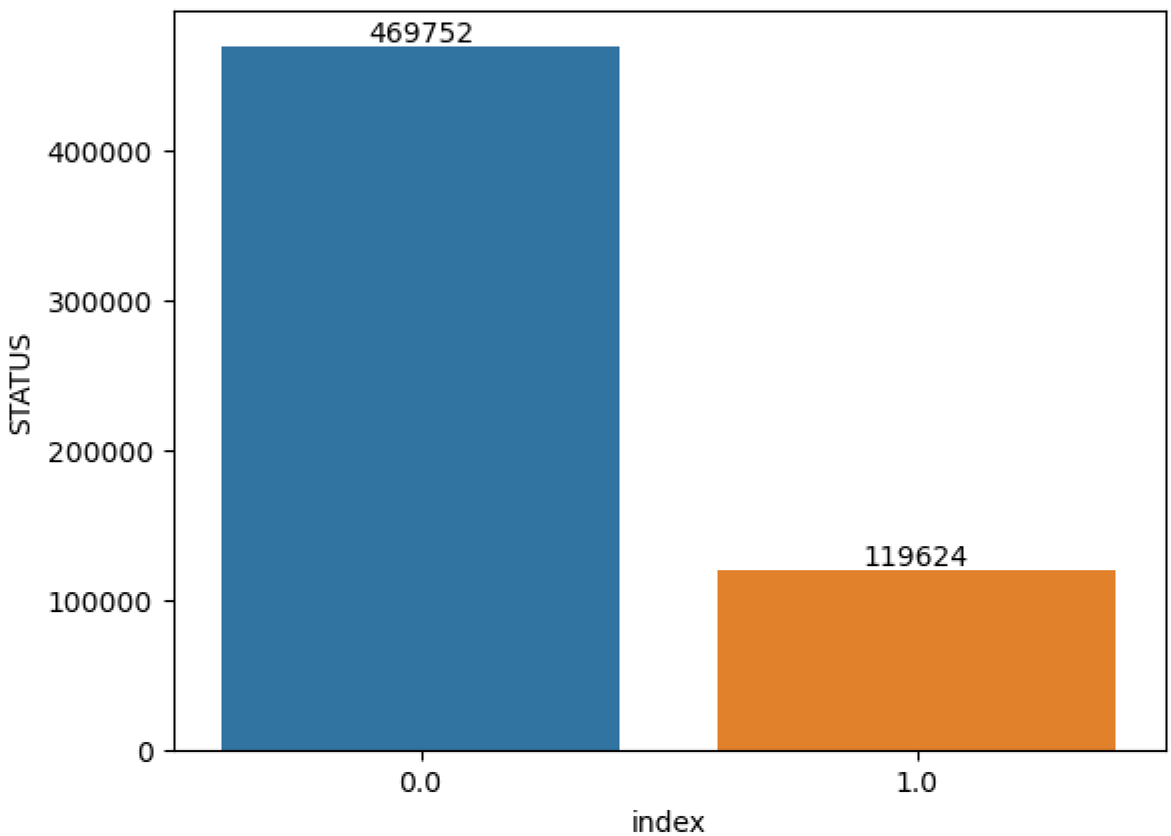
Out[49]:
```
DAY_OF_WEEK     0
DEP_DELAY       0
ARR_DELAY       0
FLIGHTS         0
DISTANCE        0
               ..
DEST_XNA        0
DEST_XWA        0
DEST_YAK        0
DEST_YKM        0
DEST_YUM        0
Length: 768, dtype: int64
```

In [50]:
```
#Set the training data in Y
Y = flight_model_df['STATUS'].astype('int')
```

In [51]:
```
#Adding this step to clear memory, to avoid memory issues suring execution
import gc

gc.collect()
```

Out[51]:
```
743
```

In [57]:
```
#Check the balance of test and train data
xx = flight_model_df['STATUS'].value_counts().reset_index()
ax = sns.barplot(x="index",y="STATUS",data=xx)
for i in ax.containers:
    ax.bar_label(i,)
```



We can clearly see that the data is not balanced. The number of delayed flights in the dataset are very low in comparison to the flights on-time. Building models with this data could give inaccurate results.

Testing a model with unbalanced data with RandomForestClassifier

In [58]:
```
# Split dataset into random train and test subsets:
```

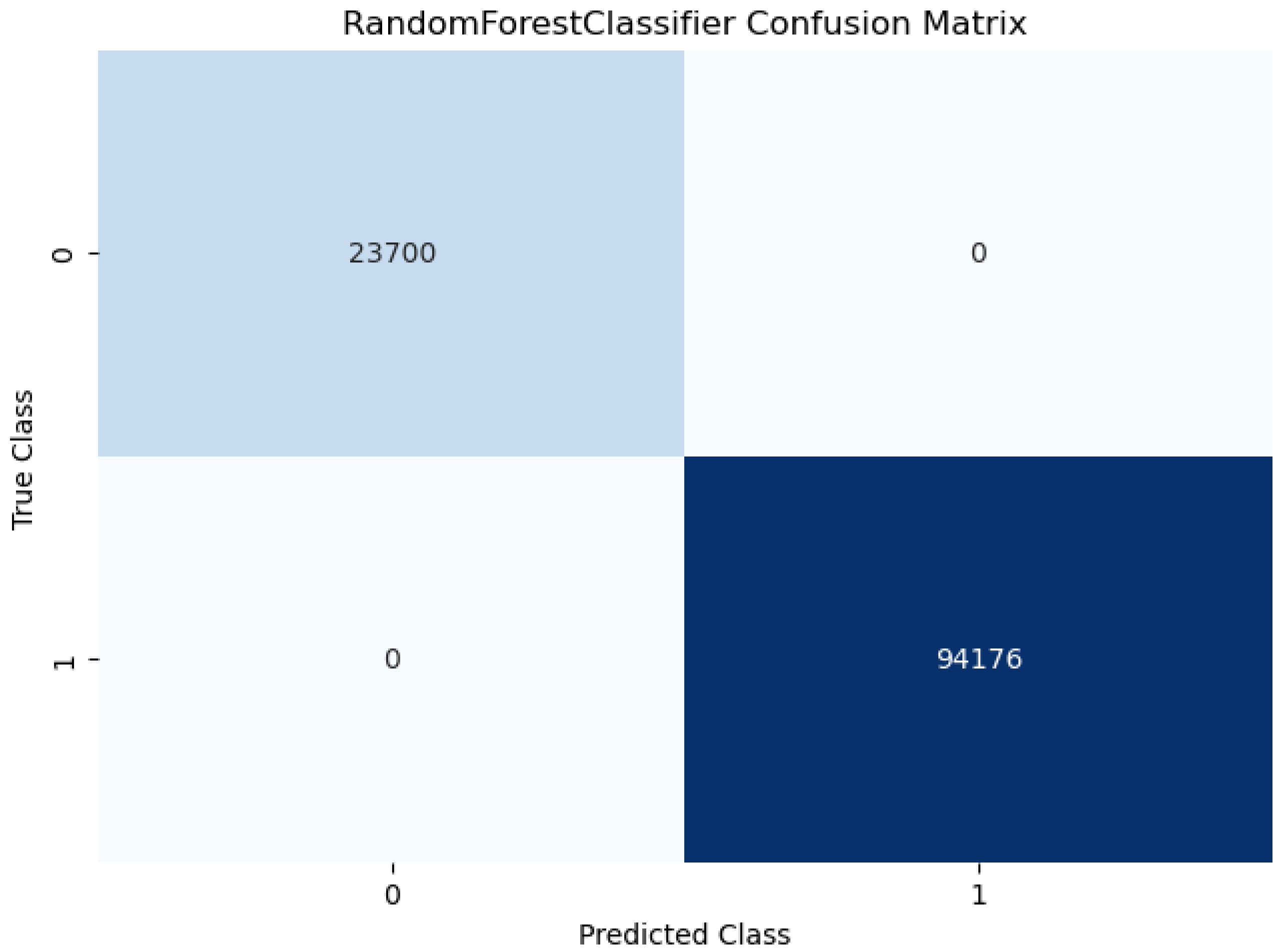
```python
x_train_ub, x_test_ub, y_train_ub, y_test_ub = train_test_split(X,Y,test_size = 0.2)
```

In [59]:
```python
#Use RandomForestClassifier to fit the unbalanced data
rfc = RandomForestClassifier()
rfc_model = rfc.fit(x_train_ub,y_train_ub)
#Predict y data with classifier:
y_pred_ub = rfc_model.predict(x_test_ub)

#Print model results
print(classification_report(y_test_ub, y_pred_ub))
print(confusion_matrix(y_test_ub, y_pred_ub))
print(f'ROC-AUC score : {roc_auc_score(y_test_ub, y_pred_ub)}')
print(f'Accuracy score : {accuracy_score(y_test_ub, y_pred_ub)}')
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     94176
           1       1.00      1.00      1.00     23700

    accuracy                           1.00    117876
   macro avg       1.00      1.00      1.00    117876
weighted avg       1.00      1.00      1.00    117876

[[94176     0]
 [    0 23700]]
ROC-AUC score : 1.0
Accuracy score : 1.0
```

In [60]:
```python
#Build the confusion matrix
matrix = confusion_matrix(y_test_ub, y_pred_ub, labels=[1,0])

print(matrix)

# Create pandas dataframe
df = pd.DataFrame(matrix)

# Create a heatmap
sns.heatmap(df, annot=True, cbar=None, cmap="Blues",fmt='.0f')
plt.title("RandomForestClassifier Confusion Matrix"), plt.tight_layout()
plt.ylabel("True Class"), plt.xlabel("Predicted Class")
plt.show()
```

```
[[23700     0]
 [    0 94176]]
```



Data looks unbalanced for delayed and on-time flights, as expected. Using SMOTE to balance data and re-build the model.

Smote oversampling for imbalanced classification

In [61]:
```
smote = SMOTE()
x, y = smote.fit_resample(X, Y)
```

In [62]:
```
smote_df = pd.concat([pd.DataFrame(x), pd.DataFrame(y)], axis=1)
smote_df.head(5)
```

Out[62]:

	DAY_OF_WEEK	DEP_DELAY	ARR_DELAY	FLIGHTS	DISTANCE	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY
0	1.494141	-0.403564	-0.266357	0.0	-0.377686	-0.148438	-0.05722	-0.166138
1	1.494141	-0.384766	-0.266357	0.0	-0.377686	-0.148438	-0.05722	-0.166138
2	1.494141	-0.328125	-0.266357	0.0	-0.377686	-0.148438	-0.05722	-0.166138
3	1.494141	-0.328125	-0.266357	0.0	-0.377686	-0.148438	-0.05722	-0.166138
4	1.494141	-0.234009	-0.266357	0.0	-0.377686	-0.148438	-0.05722	-0.166138

5 rows × 769 columns

In [63]:
```
#Split the smote (balanced) data into random train and test subsets:
x_train_sm, x_test_sm, y_train_sm, y_test_sm = train_test_split(x,y,test_size = 0.2)
```

In [64]:
```
print(x.shape)
print(y.shape)
print(x_train_sm.shape)
```

```
print(y_train_sm.shape)
print(x_test_sm.shape)
print(y_test_sm.shape )
```

```
(939504, 768)
(939504,)
(751603, 768)
(751603,)
(187901, 768)
(187901,)
```

RandomForestClassifier

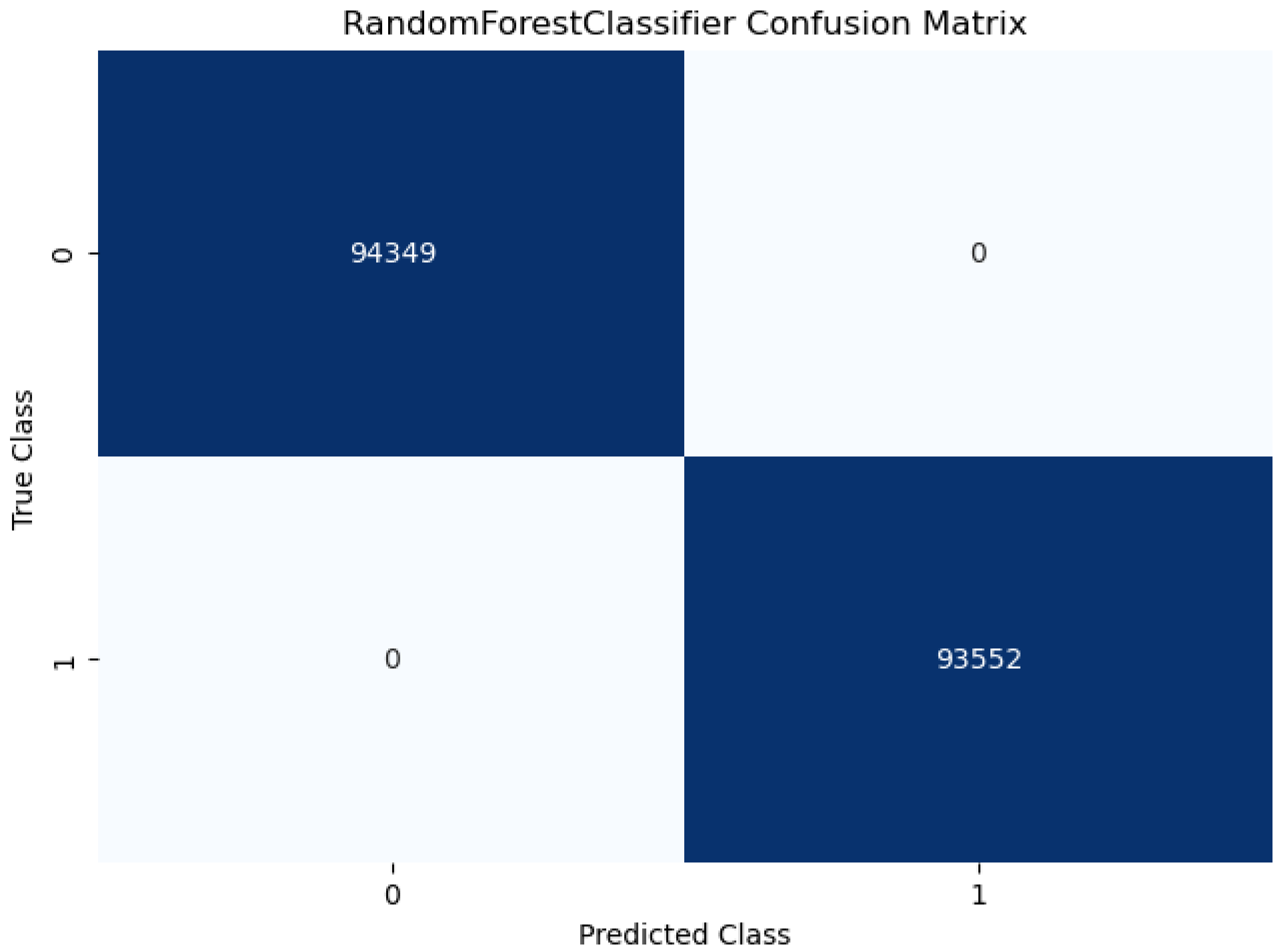
In [65]:
```
# Use the RandomForestClassifier to fit balanced data
rfc = RandomForestClassifier()
rfc_model = rfc.fit(x_train_sm,y_train_sm)
```

In [66]:
```
#Predict y data with classifier:
y_pred_rfc = rfc_model.predict(x_test_sm)

print(classification_report(y_test_sm, y_pred_rfc))
print(confusion_matrix(y_test_sm, y_pred_rfc))
print(f'ROC-AUC score : {roc_auc_score(y_test_sm, y_pred_rfc)}')
print(f'Accuracy score : {accuracy_score(y_test_sm, y_pred_rfc)}')
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93552
           1       1.00      1.00      1.00     94349

    accuracy                           1.00    187901
   macro avg       1.00      1.00      1.00    187901
weighted avg       1.00      1.00      1.00    187901

[[93552     0]
 [    0 94349]]
ROC-AUC score : 1.0
Accuracy score : 1.0
```

In [67]:
```
#Build the confusion matrix
matrix = confusion_matrix(y_test_sm, y_pred_rfc, labels=[1,0])

print(matrix)

# Create pandas dataframe
df = pd.DataFrame(matrix)

# Create a heatmap
sns.heatmap(df, annot=True, cbar=None, cmap="Blues",fmt='.0f')
plt.title("RandomForestClassifier Confusion Matrix"), plt.tight_layout()
plt.ylabel("True Class"), plt.xlabel("Predicted Class")
plt.show()
```

```
[[94349     0]
 [    0 93552]]
```



DecisionTreeClassifier

In [68]:
```
#save the feature name and target variables
feature_names = x.columns
labels = y.unique()

#split the dataset into train and test
x_train_dt, x_test_dt, y_train_dt, y_test_dt = train_test_split(x,y,
                                                                 test_size = 0.4,
                                                                 random_state = 42)
# Use the DecisionTreeClassifier to fit data
clf = DecisionTreeClassifier(max_depth =3, random_state = 42)
clf.fit(x_train_dt, y_train_dt)
```

Out[68]:
```
▾ DecisionTreeClassifier
DecisionTreeClassifier(max_depth=3, random_state=42)
```

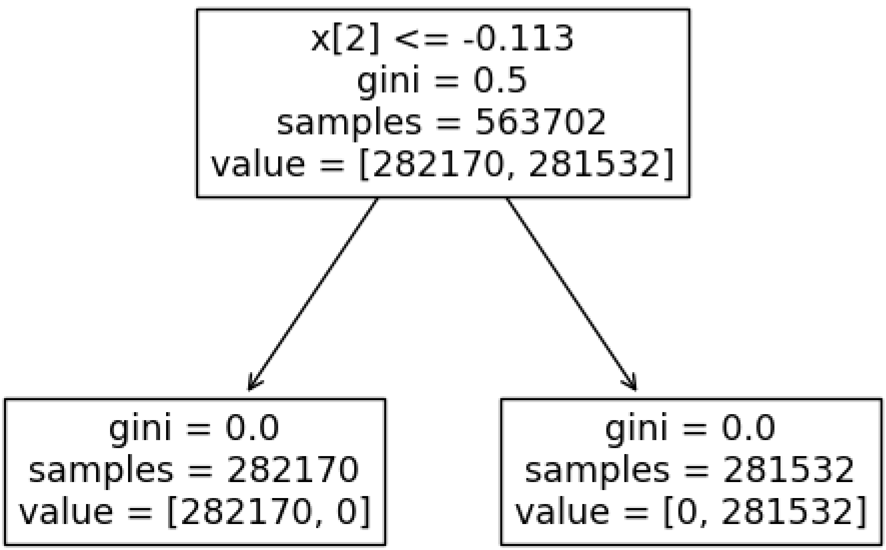
In [69]:
```
#Predict y data with classifier:
y_pred_dtc = clf.predict(x_test_sm)

#Print results
print(classification_report(y_test_sm, y_pred_dtc))
print(confusion_matrix(y_test_sm, y_pred_dtc))
print(f'ROC-AUC score : {roc_auc_score(y_test_sm, y_pred_dtc)}')
print(f'Accuracy score : {accuracy_score(y_test_sm, y_pred_dtc)}')
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93552
           1       1.00      1.00      1.00     94349
```

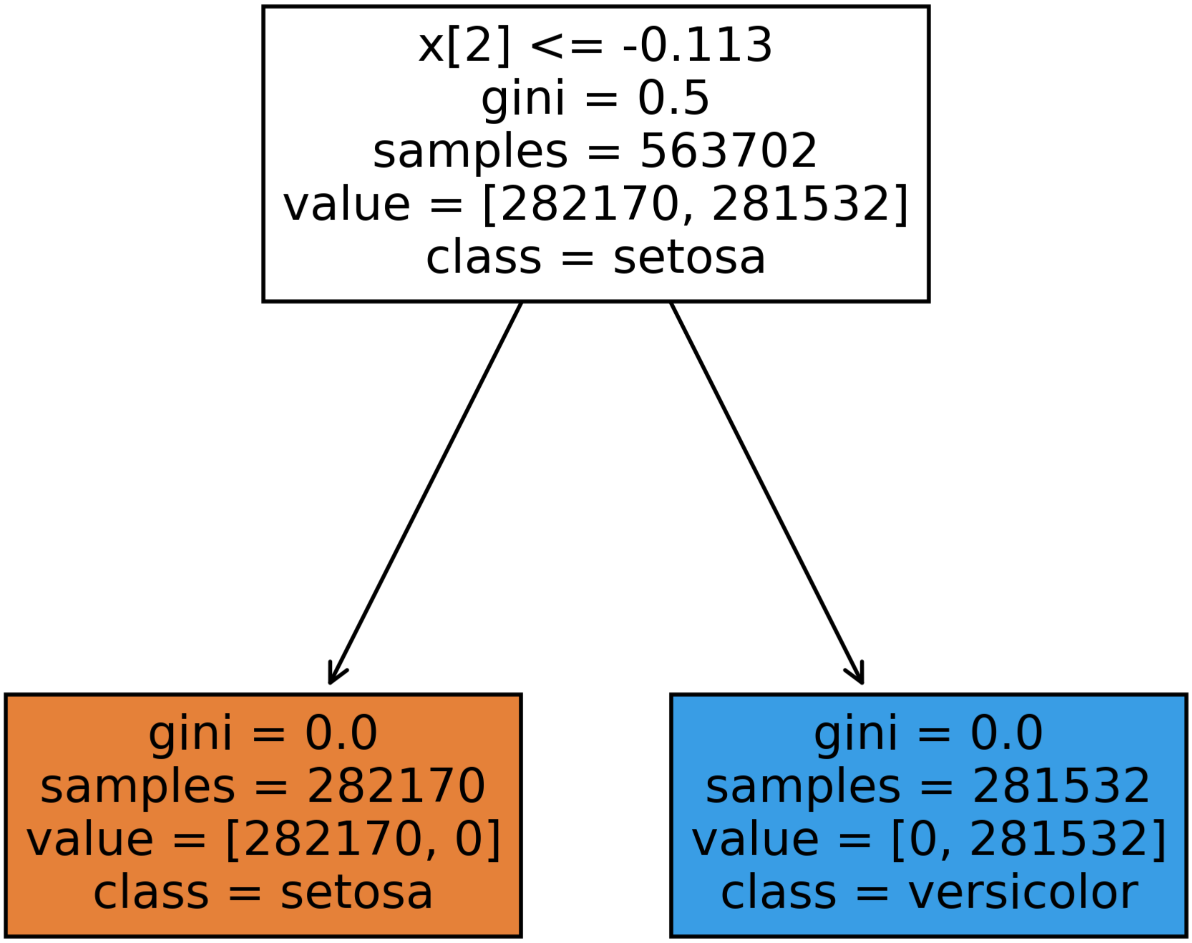
```
    accuracy                           1.00    187901
   macro avg       1.00      1.00      1.00    187901
weighted avg       1.00      1.00      1.00    187901

[[93552     0]
 [    0 94349]]
ROC-AUC score : 1.0
Accuracy score : 1.0
```

In [70]:
```python
# Draw graph
tree.plot_tree(clf)
```

Out[70]:
```
[Text(0.5, 0.75, 'x[2] <= -0.113\ngini = 0.5\nsamples = 563702\nvalue = [282170, 28153
2]'),
 Text(0.25, 0.25, 'gini = 0.0\nsamples = 282170\nvalue = [282170, 0]'),
 Text(0.75, 0.25, 'gini = 0.0\nsamples = 281532\nvalue = [0, 281532]')]
```



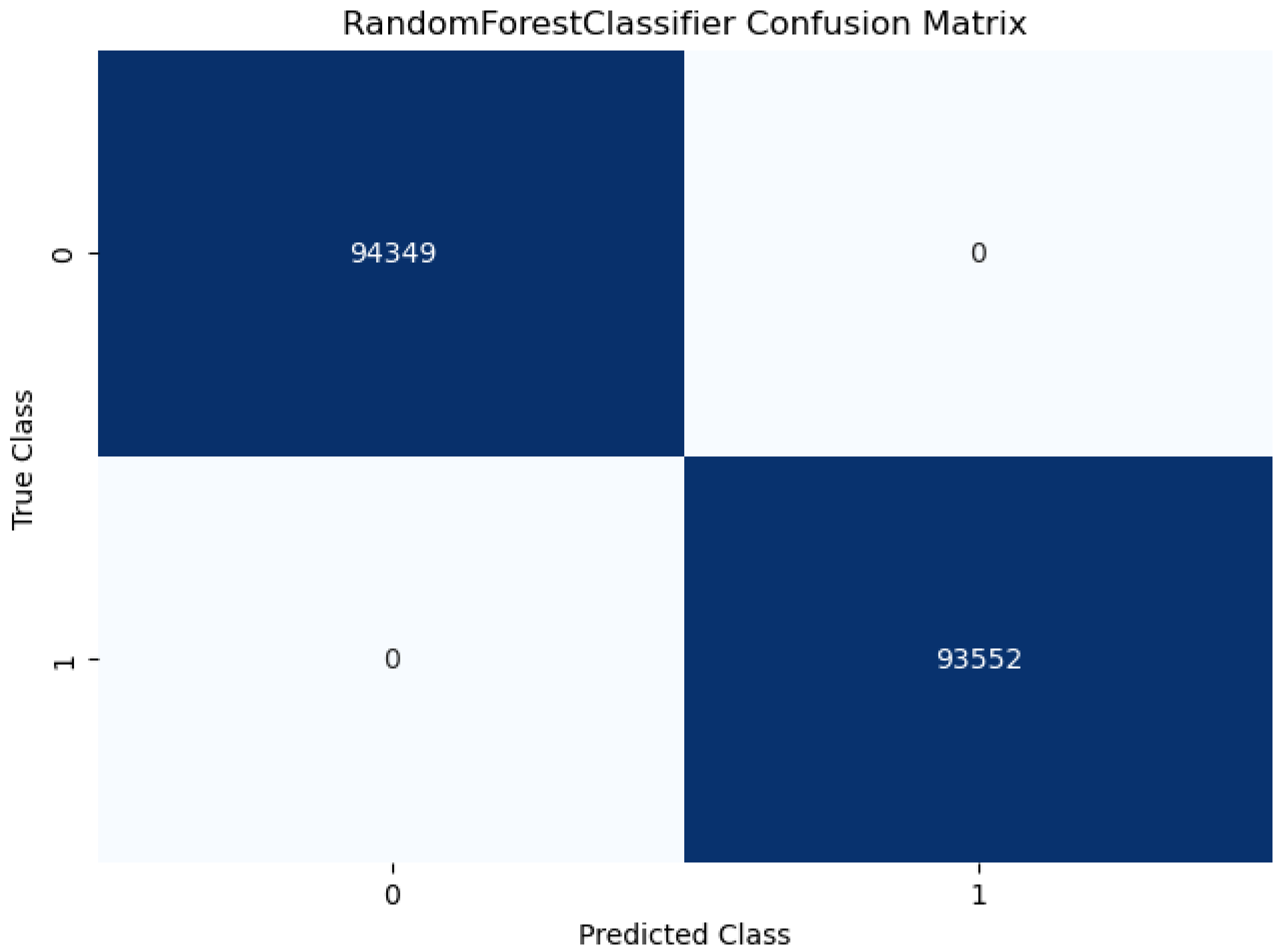
In [71]:
```python
# Draw graph

cn=['setosa', 'versicolor', 'virginica']
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (6,6), dpi=400)
tree.plot_tree(clf,
               class_names=cn,
               filled = True);
fig.savefig('dtc.png')
```



In [72]:

```
#Build the confusion matrix
matrix = confusion_matrix(y_test_sm, y_pred_dtc, labels=[1,0])

print(matrix)

# Create pandas dataframe
df = pd.DataFrame(matrix)

# Create a heatmap
sns.heatmap(df, annot=True, cbar=None, cmap="Blues",fmt='.0f')
plt.title("RandomForestClassifier Confusion Matrix"), plt.tight_layout()
plt.ylabel("True Class"), plt.xlabel("Predicted Class")
plt.show()
```

```
[[94349     0]
 [    0 93552]]
```



LogisticRegression

In [73]:
```
# Use the LogisticRegression to fit data:
lr_model = LogisticRegression(solver='liblinear')
model = lr_model.fit(x_train_sm, y_train_sm)
```

In [74]:
```
#Predict y data with classifier:
y_pred_lr = model.predict(x_test_sm)

#Pritn results
print(classification_report(y_test_sm, y_pred_lr))
print(confusion_matrix(y_test_sm, y_pred_lr))
print(f'ROC-AUC score : {roc_auc_score(y_test_sm, y_pred_lr)}')
print(f'Accuracy score : {accuracy_score(y_test_sm, y_pred_lr)}')
```

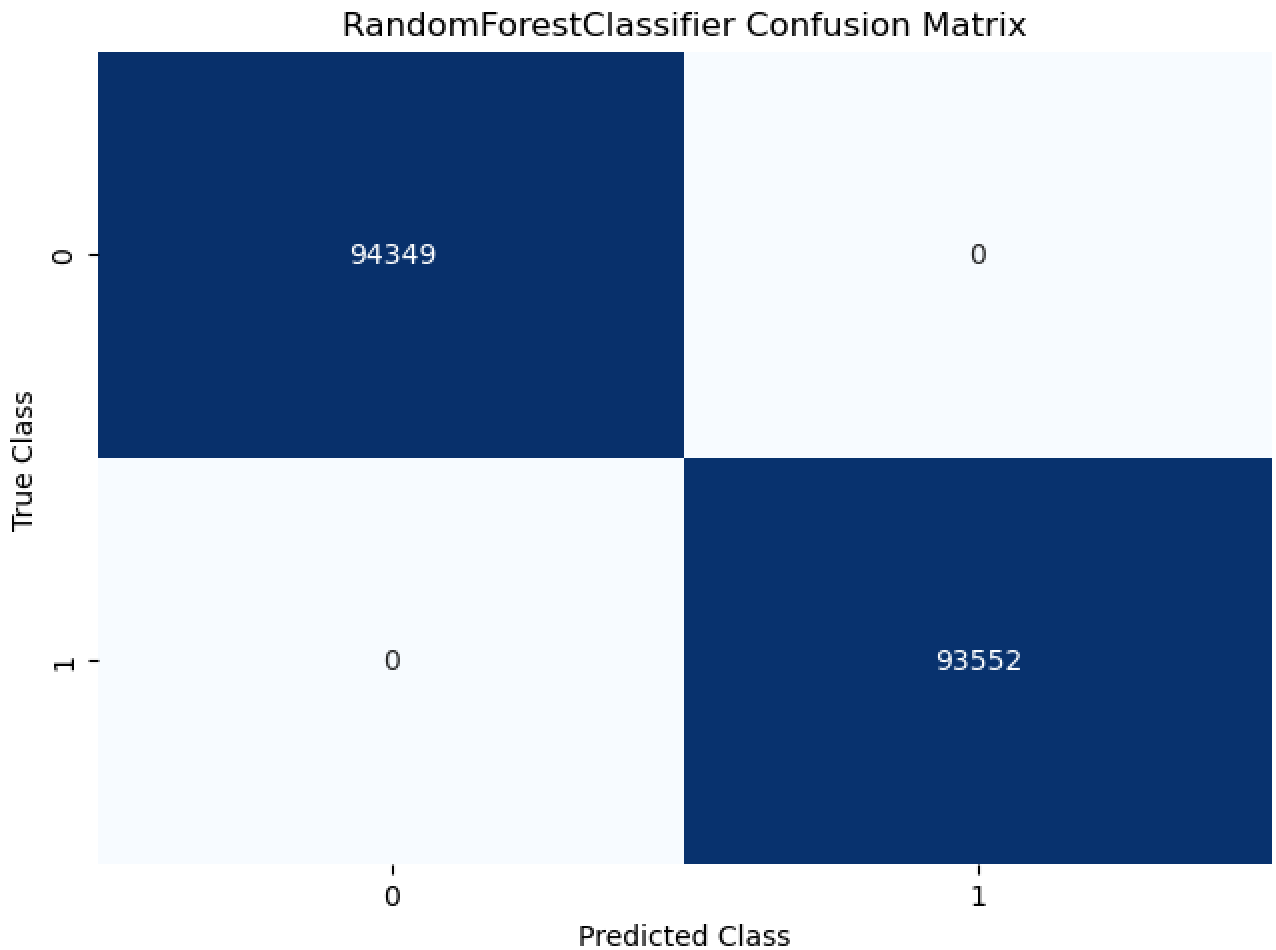
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93552
           1       1.00      1.00      1.00     94349

    accuracy                           1.00    187901
   macro avg       1.00      1.00      1.00    187901
weighted avg       1.00      1.00      1.00    187901

[[93552     0]
 [    0 94349]]
ROC-AUC score : 1.0
Accuracy score : 1.0
```

In [75]:
```
# Cross-validate model using accuracy
cross_val_score(lr_model, x, y, scoring="accuracy")
```

Out[75]:
```
array([1., 1., 1., 1., 1.])
```

In [76]:
```
#Build the confusion matrix
matrix = confusion_matrix(y_test_sm, y_pred_lr, labels=[1,0])

print(matrix)

# Create pandas dataframe
df = pd.DataFrame(matrix)

# Create a heatmap
sns.heatmap(df, annot=True, cbar=None, cmap="Blues",fmt='.0f')
plt.title("RandomForestClassifier Confusion Matrix"), plt.tight_layout()
plt.ylabel("True Class"), plt.xlabel("Predicted Class")
plt.show()
```

```
[[94349     0]
 [    0 93552]]
```



from yellowbrick.classifier import DiscriminationThreshold lr_visualizer = DiscriminationThreshold(lr_model) lr_visualizer.fit(x,y) # Fit the data to the visualizer lr_visualizer.show() # Finalize and render the figure

Using class_weight='balaced' to check if the imbalance get's any better.

In [77]:
```
logit = LogisticRegression( solver='liblinear', class_weight='balanced')
model_logit = logit.fit(x_train_sm, y_train_sm)

y_pred_logit = model_logit.predict(x_test_sm)

print(classification_report(y_test_sm, y_pred_logit))
print(confusion_matrix(y_test_sm, y_pred_logit))
print(f'ROC-AUC score : {roc_auc_score(y_test_sm, y_pred_logit)}')
print(f'Accuracy score : {accuracy_score(y_test_sm, y_pred_logit)}')

# Cross-validate model using accuracy
cross_val_score(logit, x, y, scoring="accuracy")
```

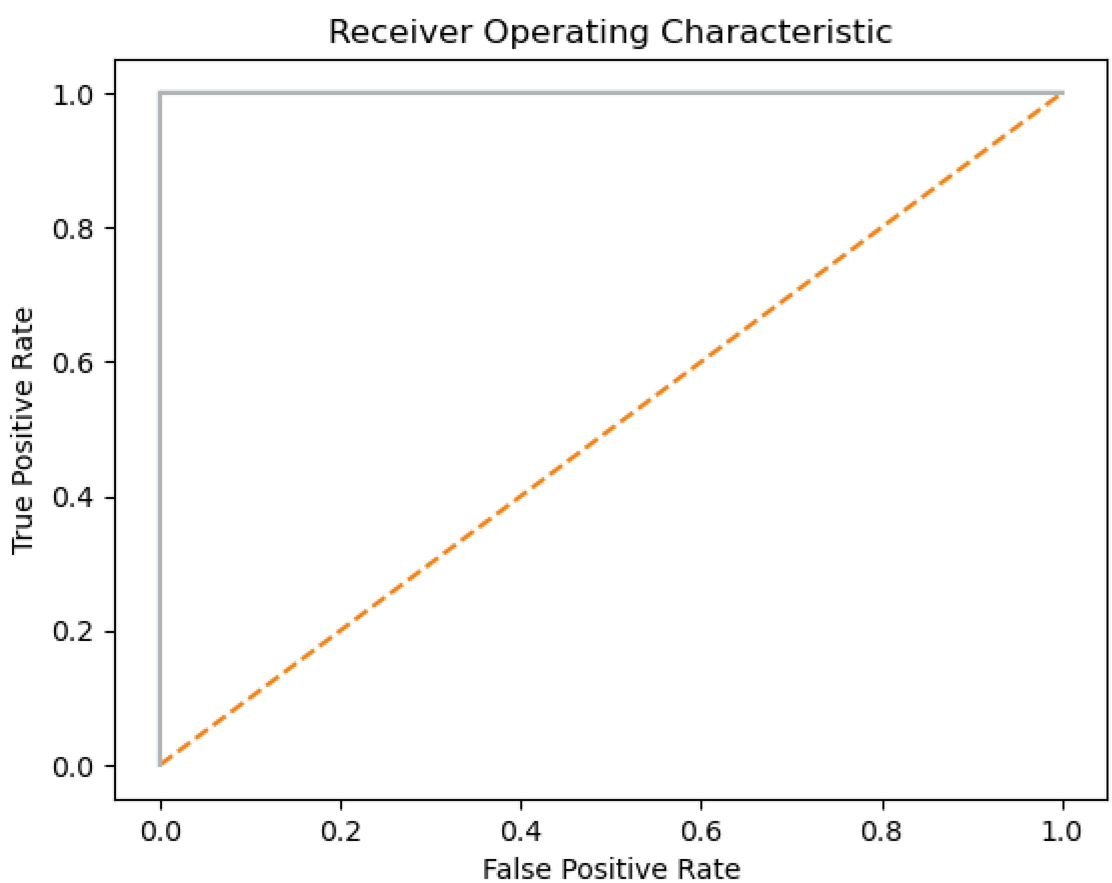
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93552
           1       1.00      1.00      1.00     94349

    accuracy                           1.00    187901
   macro avg       1.00      1.00      1.00    187901
weighted avg       1.00      1.00      1.00    187901

[[93552     0]
 [    0 94349]]
ROC-AUC score : 1.0
Accuracy score : 1.0
```

Out[77]:

```
array([1., 1., 1., 1., 1.])
```

In [78]:

```
logit.fit(x_train_sm, y_train_sm)

# Get predicted probabilities
target_probabilities = logit.predict_proba(x_test_sm)[:,1]
# Create true and false positive rates
false_positive_rate, true_positive_rate, threshold = roc_curve(y_test_sm,target_probabil

# Plot ROC curve
plt.title("Receiver Operating Characteristic")
plt.plot(false_positive_rate, true_positive_rate)
plt.plot([0, 1], ls="--")
plt.plot([0, 0], [1, 0] , c=".7"), plt.plot([1, 1] , c=".7")
plt.ylabel("True Positive Rate")
plt.xlabel("False Positive Rate")
plt.show()
```



This is a perfect model, which means there's a clear way to tell whether any given record is in the training or test sets. This is a violation of the assumption that our training and test sets are identically distributed.

In [79]:

```
print("Threshold:", threshold[116])
```

```
print("True Positive Rate:", true_positive_rate[116])
print("False Positive Rate:", false_positive_rate[116])
```

```
Threshold: 0.9999999999999709
True Positive Rate: 0.4126275848180691
False Positive Rate: 0.0
```

Conclusion

The following 4 models were built and evaluated with theBTS airline performance data

- LogisticRegression
- RandomForestClassifier
- DecisionTreeClassifier

Due to imbalance in the data I am getting a perfect accuracy score for all models. I tried using different datasets from the BTS website (May, June, entire 2022 data from Jan through Nov, and the current execution with Dec'22 data), and I am getting the same accuracy score, precision, recall and f1-score for all datasets from the website.

Following lines can be ignored

SMOTE + TOMEK

print(X.shape) print(y.shape) print(X_cat.shape, X_num.shape) print(len(flight_model_df['STATUS']))
y_st = flight_model_df['STATUS'].astype('int')
smotetomek = SMOTETomek() x_tomek, y_tomek = smotetomek.fit_resample(X, y_st)
smote_tomek_df = pd.concat([pd.DataFrame(x_tomek), pd.DataFrame(y_tomek)], axis=1) smote_tomek_df.head(5)
x_train_st, x_test_st, y_train_st, y_test_st = train_test_split(x,y_st,test_size = 0.2)
print(x.shape) print(y.shape) print(x_train.shape) print(y_train.shape) print(x_test.shape) print(y_test.shape)
KNeighborsClassifier
Train a KNN classifier with 5 neighbors and Predict y data with classifier: y_pred_knn = KNeighborsClassifier(n_neighbors=5, n_jobs=-1).fit(x_train_sm, y_train_sm).predict(x_test_sm)
Print results: print(classification_report(y_test_sm, y_pred_knn)) print(confusion_matrix(y_test_sm, y_pred_knn)) print(f'ROC-AUC score : {roc_auc_score(y_test_sm, y_pred_knn)}') print(f'Accuracy score : {accuracy_score(y_test_sm, y_pred_knn)}')