Analysis of American Flights in 2020

January 7, 2021

0.1 Analysis of American Flights in 2020

0.2 Table of Contents

- Introduction
- Gathering data
- Assessing data
- Cleaning data
- visualization

Introduction

In this project i will analyse American flights during the year 2020. The aim is to answer these questions :

What percentage of flights has been cancelled?

Which Airline has had the most flight?

How many primary airports?

Number of Airlines?

Which airport has the most departing flights? least departing flights?

What Percentage of all flights depart from or arrive at Atlanta International Airport?

Number of Flights diverted?

Relationship between states based on Arrival and Departure.

Relationship between states based on Arrival and Departure with travel time. Data was provided by BUREAU OF TRANSPORTATION STATISTICS, U.S. Department of Transportation. https://www.transtats.bts.gov/Fields.asp?Table_ID=236 To get started, let's import our libraries.

[11]:

```
import pandas as pd
import numpy as np
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
```

[12]:

```
flights_df = pd.read_csv('715514522_T_ONTIME_REPORTING.csv')
```

[13]:

```
flights_df.head()
```

[13]:

```
   YEAR  MONTH  DAY_OF_MONTH  DAY_OF_WEEK OP_UNIQUE_CARRIER TAIL_NUM  \
0  2020      1             1            3                WN   N951WN
1  2020      1             1            3                WN   N467WN
2  2020      1             1            3                WN   N7885A
3  2020      1             1            3                WN   N551WN
4  2020      1             1            3                WN   N968WN

   OP_CARRIER_FL_NUM  ORIGIN_AIRPORT_ID  ORIGIN_AIRPORT_SEQ_ID  \
0               5888              13891                1389101
1               6276              13891                1389101
2               4598              13891                1389101
3               4761              13891                1389101
4               5162              13891                1389101

   ORIGIN_CITY_MARKET_ID  … DIVERTED CRS_ELAPSED_TIME  AIR_TIME  DISTANCE  \
0                  32575  …      0.0             95.0      74.0     363.0
1                  32575  …      0.0             90.0      71.0     363.0
2                  32575  …      0.0             70.0      57.0     333.0
3                  32575  …      0.0             75.0      63.0     333.0
4                  32575  …      0.0             80.0      57.0     333.0

   CARRIER_DELAY WEATHER_DELAY NAS_DELAY  SECURITY_DELAY  LATE_AIRCRAFT_DELAY  \
0            8.0           0.0      27.0             0.0                 33.0
1            NaN           NaN       NaN             NaN                  NaN
2            NaN           NaN       NaN             NaN                  NaN
3            NaN           NaN       NaN             NaN                  NaN
4            NaN           NaN       NaN             NaN                  NaN

   Unnamed: 36
0          NaN
1          NaN
2          NaN
3          NaN
4          NaN

[5 rows x 37 columns]
```

[14]:

```
flights_df.describe()
```

[14]:

```
           YEAR     MONTH  DAY_OF_MONTH    DAY_OF_WEEK  OP_CARRIER_FL_NUM  \
count  607346.0  607346.0  607346.000000  607346.000000      607346.000000
mean     2020.0       1.0      16.014354       3.955735        2622.365261
std         0.0       0.0       8.990719       1.910205        1822.545302
min      2020.0       1.0       1.000000       1.000000           1.000000
25%      2020.0       1.0       8.000000       2.000000        1070.000000
50%      2020.0       1.0      16.000000       4.000000        2177.000000
75%      2020.0       1.0      24.000000       5.000000        4108.000000
max      2020.0       1.0      31.000000       7.000000        6860.000000

       ORIGIN_AIRPORT_ID  ORIGIN_AIRPORT_SEQ_ID  ORIGIN_CITY_MARKET_ID  \
count      607346.000000           6.073460e+05          607346.000000
mean        12657.389167           1.265743e+06           31761.273269
std          1524.407203           1.524405e+05            1308.052641
min         10135.000000           1.013506e+06           30070.000000
25%         11292.000000           1.129202e+06           30713.000000
50%         12889.000000           1.288903e+06           31453.000000
75%         14027.000000           1.402702e+06           32467.000000
max         16869.000000           1.686901e+06           35991.000000

       DEST_AIRPORT_ID  DEST_AIRPORT_SEQ_ID  …       DIVERTED  \
count    607346.000000         6.073460e+05  …  607346.000000
mean      12657.196320         1.265724e+06  …       0.001893
std        1524.279269         1.524277e+05  …       0.043473
min       10135.000000         1.013506e+06  …       0.000000
25%       11292.000000         1.129202e+06  …       0.000000
50%       12889.000000         1.288903e+06  …       0.000000
75%       14027.000000         1.402702e+06  …       0.000000
max       16869.000000         1.686901e+06  …       1.000000

       CRS_ELAPSED_TIME       AIR_TIME       DISTANCE  CARRIER_DELAY  \
count     607346.000000  599268.000000  607346.000000   82285.000000
mean         144.583689     112.187437     798.022341      24.696324
std           72.688861      70.629553     587.282639      72.972359
min          -77.000000       8.000000      31.000000       0.000000
25%           92.000000      61.000000     369.000000       0.000000
50%          127.000000      94.000000     641.000000       1.000000
75%          175.000000     142.000000    1037.000000      22.000000
max          700.000000     698.000000    5095.000000    2489.000000

       WEATHER_DELAY     NAS_DELAY  SECURITY_DELAY  LATE_AIRCRAFT_DELAY  \
count   82285.000000  82285.000000    82285.000000         82285.000000
mean        4.594944     14.262733        0.091062            20.561658
std        39.180258     33.736783        2.308003            50.370818
min         0.000000      0.000000        0.000000             0.000000
25%         0.000000      0.000000        0.000000             0.000000
50%         0.000000      2.000000        0.000000             0.000000
```

```
75%         0.000000     19.000000         0.000000            22.000000
max      1525.000000   1408.000000       188.000000          2228.000000

       Unnamed: 36
count          0.0
mean           NaN
std            NaN
min            NaN
25%            NaN
50%            NaN
75%            NaN
max            NaN

[8 rows x 30 columns]
```

[15]:

```
totalNumberOfFlights = 607346.0
```

What percentage of flights has been cancelled?

[16]:

```
len(flights_df[flights_df.CANCELLED != 0]) / float(len(flights_df))
```

[16]:

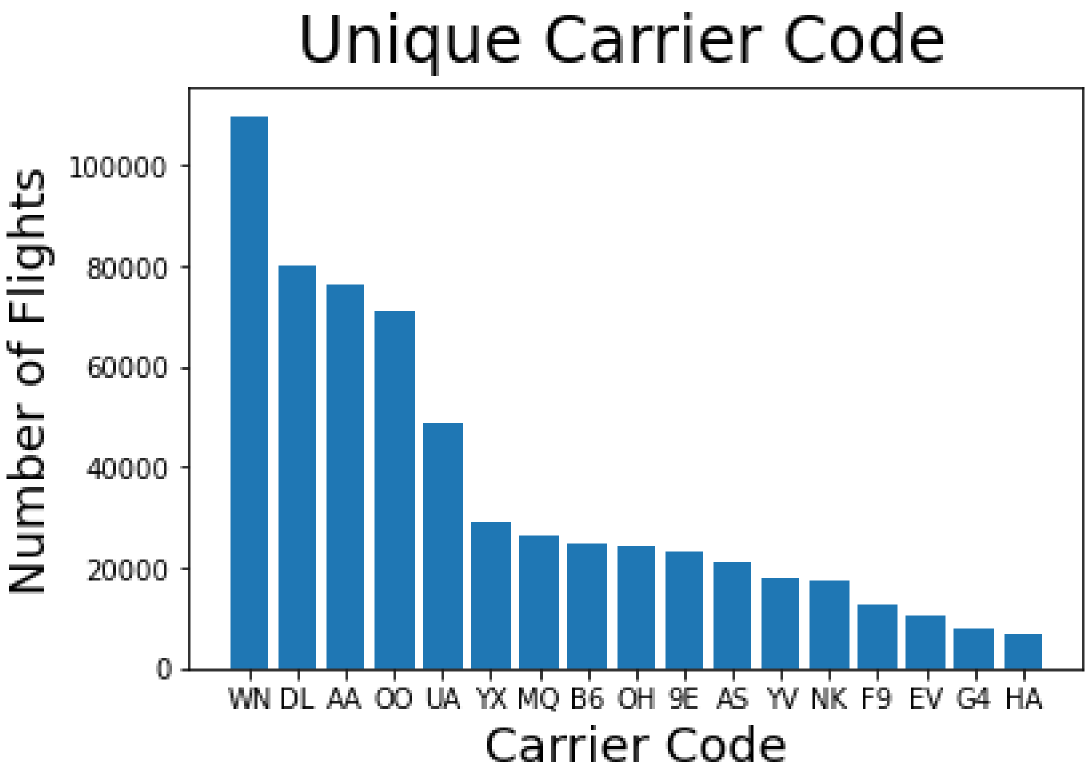
```
0.011407006879110095
```

1.14% of the total flights have been cancelled in the year 2020 across all states.

Which Airline has had the most flight?

[17]:

```
uniqueCarrierCode = flights_df.OP_UNIQUE_CARRIER.value_counts()
plt.suptitle('Unique Carrier Code', fontsize=24)
plt.xlabel('Carrier Code', fontsize=18)
plt.ylabel('Number of Flights', fontsize=18)
plt.bar(uniqueCarrierCode.index, uniqueCarrierCode);
```



Unique carrier code WN,belonging to Southwest Airlines, has the most flights in the year 2020.

How many primary airports?

```
[18]: flights_df['ORIGIN_AIRPORT_ID'].nunique()
```

```
[18]: 351
```

There are 351 primary airports.

Number of Airlines?

```
[19]: flights_df['OP_UNIQUE_CARRIER'].nunique()
```

```
[19]: 17
```

There are 17 Airlines.

Which airport has the most departing flights? least departing flights?

```
[20]: depstate = flights_df.ORIGIN.value_counts()
      plt.figure(figsize=[180,8])
      plt.suptitle('Departing flights per Airport', fontsize=24)
      plt.xlabel('Airport Code', fontsize=18)
      plt.ylabel('Number of Flights', fontsize=18)
      plt.bar(depstate.index, depstate);
```

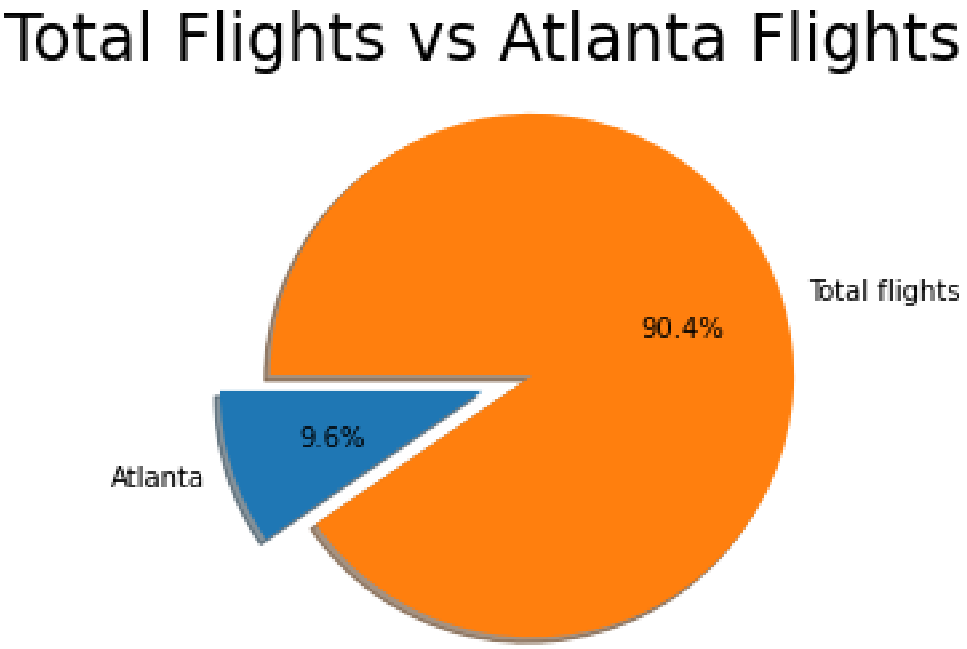
Hartsfield Jackson Atlanta International Airport (IATA Code ATL) has the most departing flights out of all airports in the US.

Yellowstone Regional Airport (IATA Code COD) has the least amount of departing flights int he US.

What Percentage of all flights depart from or arrive at Atlanta International Airport?

[21]:
```
atlFlightsD = len(flights_df.query('DEST == "ATL"'))
atlFlightsO = len(flights_df.query('ORIGIN == "ATL"'))
allATLFlights = atlFlightsD + atlFlightsO
```

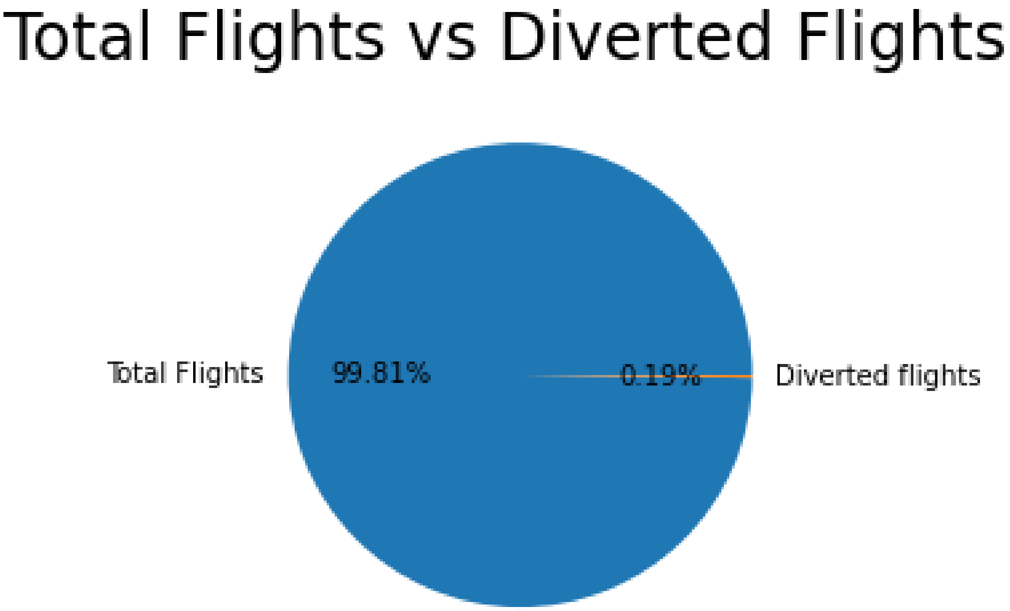
[22]:
```
labels = 'Atlanta', 'Total flights'
sizes = [allATLFlights, totalNumberOfFlights]
explode = (0.2, 0)

fig1, pieat = plt.subplots()
pieat.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=180)
pieat.axis('equal')
plt.suptitle('Total Flights vs Atlanta Flights ', fontsize=24)
plt.show()
```



9.6% of all flights int the US either arrive or depart from Atlanta International Airport.

Number of Flights diverted?

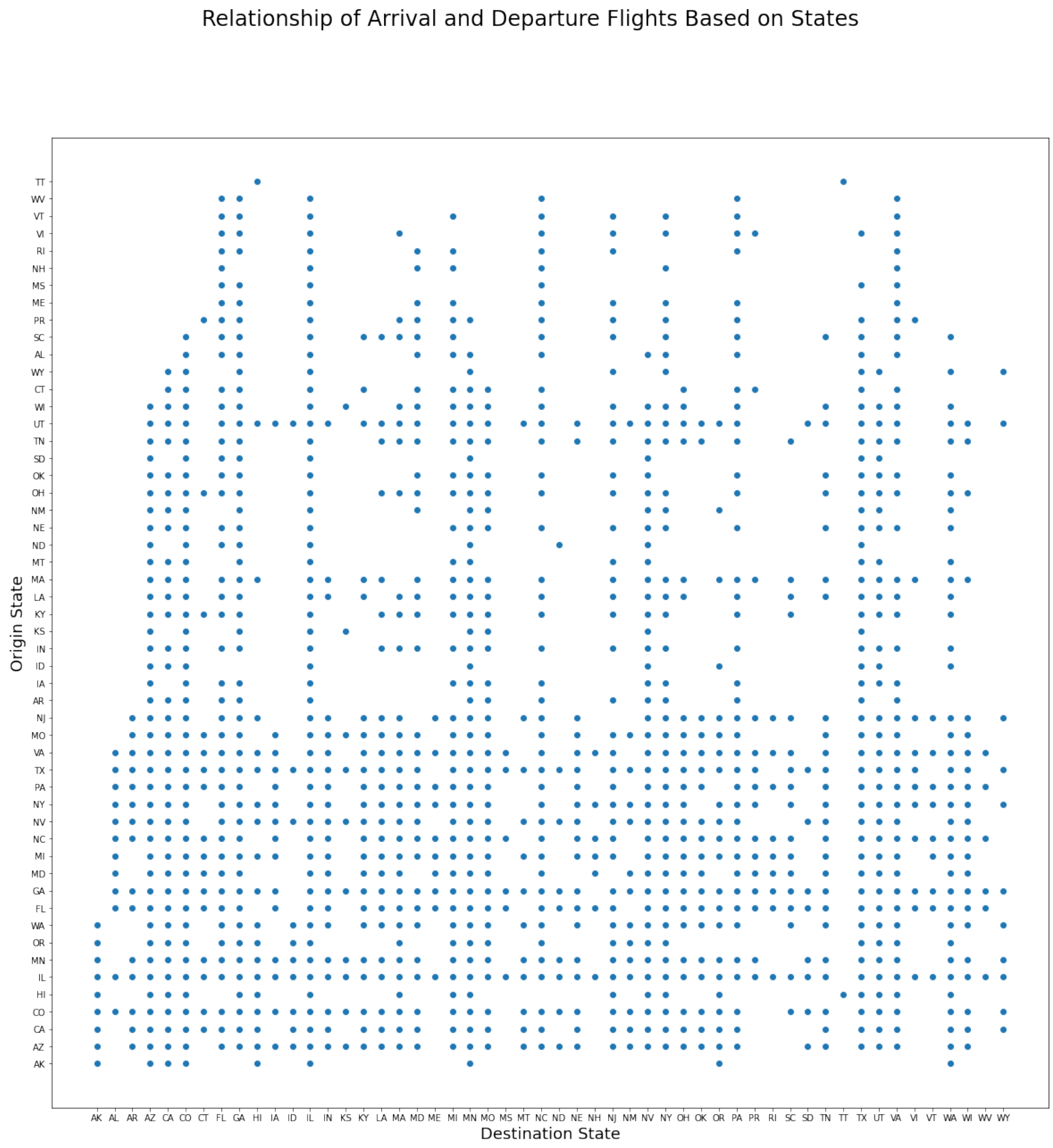
```
[23]: labels = 'Total Flights', 'Diverted flights'
plt.suptitle('Total Flights vs Diverted Flights', fontsize=24)
plt.pie(flights_df.DIVERTED.value_counts(),labels=labels, autopct='%.2f%%');
```



0.19% of all flights were diverted

Relationship between states based on Arrival and Departure.

```
[24]: state_airlines = flights_df.groupby(['ORIGIN_STATE_ABR', 'DEST_STATE_ABR'],␣
 ↪as_index=False).count()
state_airlines = state_airlines[['ORIGIN_STATE_ABR', 'DEST_STATE_ABR']]
plt.figure(figsize=[20,20])
plt.suptitle('Relationship of Arrival and Departure Flights Based on States ',␣
 ↪fontsize=24)
plt.xlabel('Destination State', fontsize=18)
plt.ylabel('Origin State', fontsize=18)
plt.scatter(state_airlines.ORIGIN_STATE_ABR, state_airlines.DEST_STATE_ABR);
```

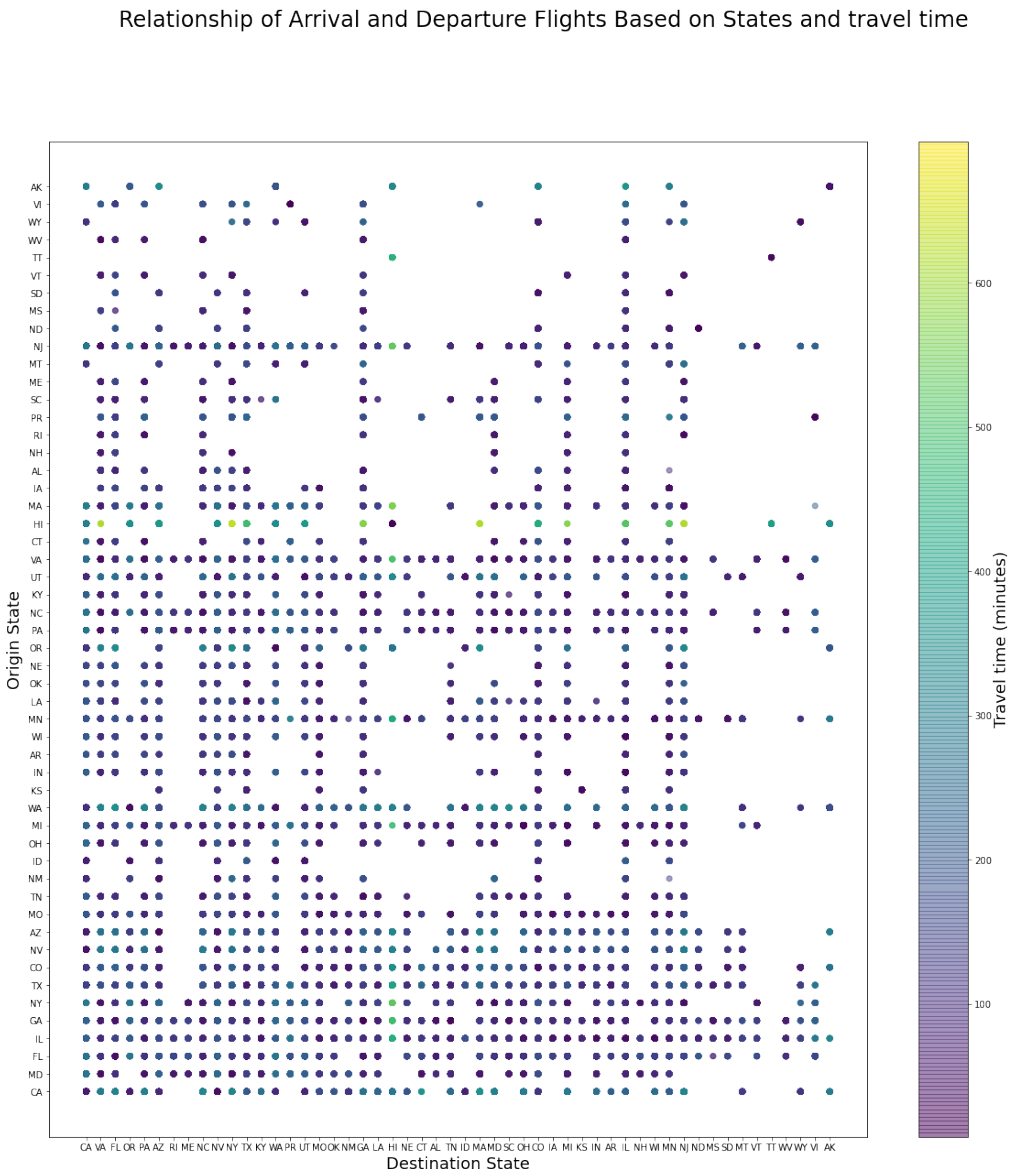


Using scatter plot we can see the relationships of all states based on flight paths. The state of Illinois(IL) has had flights departing to all states and arriving form all states except for Trust Territory(TT).

Relationship between states based on Arrival and Departure with travel time.

[25]:
```
airTimeMean = flights_df.AIR_TIME.mean()
```

[27]:
```
def airTimeColor(x):
    aT_max, aT_min = airTimeMean, -(airTimeMean)
    if(x>aT_max):
        x=aT_max
    elif(x<aT_min):
        x=aT_min
    return x
flights_df.DepDelay = flights_df.AIR_TIME.apply(airTimeColor)
plt.figure(figsize=[20,20])
plt.suptitle('Relationship of Arrival and Departure Flights Based on States and␣
 ↪travel time', fontsize=24)
plt.xlabel('Destination State', fontsize=18)
plt.ylabel('Origin State', fontsize=18)
plt.scatter(x=flights_df.ORIGIN_STATE_ABR, y=flights_df.DEST_STATE_ABR, alpha=.
 ↪5, c=flights_df.AIR_TIME)
colorBar = plt.colorbar();
colorBar.set_label("Travel time (minutes)",fontsize=18, labelpad=+2)
plt.show()
```



From this multivariate scatter plot we can see that all flights arriving to or departing from the state of Hawaii(HI) have high travel time.

ref: https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/pie_features.html#sphx-glr-gallery-pie-and-polar-charts-pie-features-py