Flight Delay Analysis and Prediction

TEAM MEMBERS

DURGAVENKATA PRAVEEN REDDY CHAPPIDI
NITHIN BOLLA

HTTPS://DRIVE.GOOGLE.COM/FILE/D/1UVV0SEC9MKBWXBK5NO2ISRDCZFLH09FI/VIEW?USP=SHARING

PROBLEM STATEMENT

To Analyze and Understand Delays in Airplane On-Time Performance

FACTORS CAUSING FLIGHT DELAY

Extreme Weather

Late Arrival Aircrafts

Carrier Issue

Security

National Aviation System

TO FIND ACCURACY??

Logistic Regression

METHODS

01

FINDSPARK :
USED TO CREATE A FILE AND USE THE SPARK FEATURES TO EXTRACT THE FLIGHT DELAY ANALYSIS

02

FINDSPARK.INIT() :
INIT METHOD IS USED TO INITIALIZE THE SPARK ENVIRONMENT

03

SPARKCONF() :
USED TO CREATE A FILE AND USE THE SPARK FEATURES TO EXTRACT THE FLIGHT DELAY ANALYSIS

04

SPARKCONTEXT() :
USED TO SET THE PROPERTIES OF SPARK AND CONNECT TO CLUSTER AND CONFIG FILE CAN BE PASSED AS ARGUMENT

05

SPARKSESSION() :
IS USED TO CREATE DATA FRAME FROM THE DATASET

06

SQLCONTEXT:
SQL CONTEXT LIBRARY IS USED IN THE CODE AS WE DEAL WITH CREATION, QUERYING AND ANALYSIS OF DATA FRAMES AND TABLES.

07

LABELEDPOINT() :
THE METHOD IS USED FOR REGRESSION AND IT ACCEPTS LABEL AND FEATURE AS PARAMETERS.

08

CREATEDATAFRAME() :
USED TO CREATE A DATAFRAME AND PERFORM OPERATIONS

09

WITHCOLUMN () :
MODIFYING THE DATAFRAMES BY ADDING A NEW COLUMN WHICH STORES THE DELAY TIME

10

TOPANDAS () :
USED FOR CONVERSION OF DATA FRAMES TO PANDAS

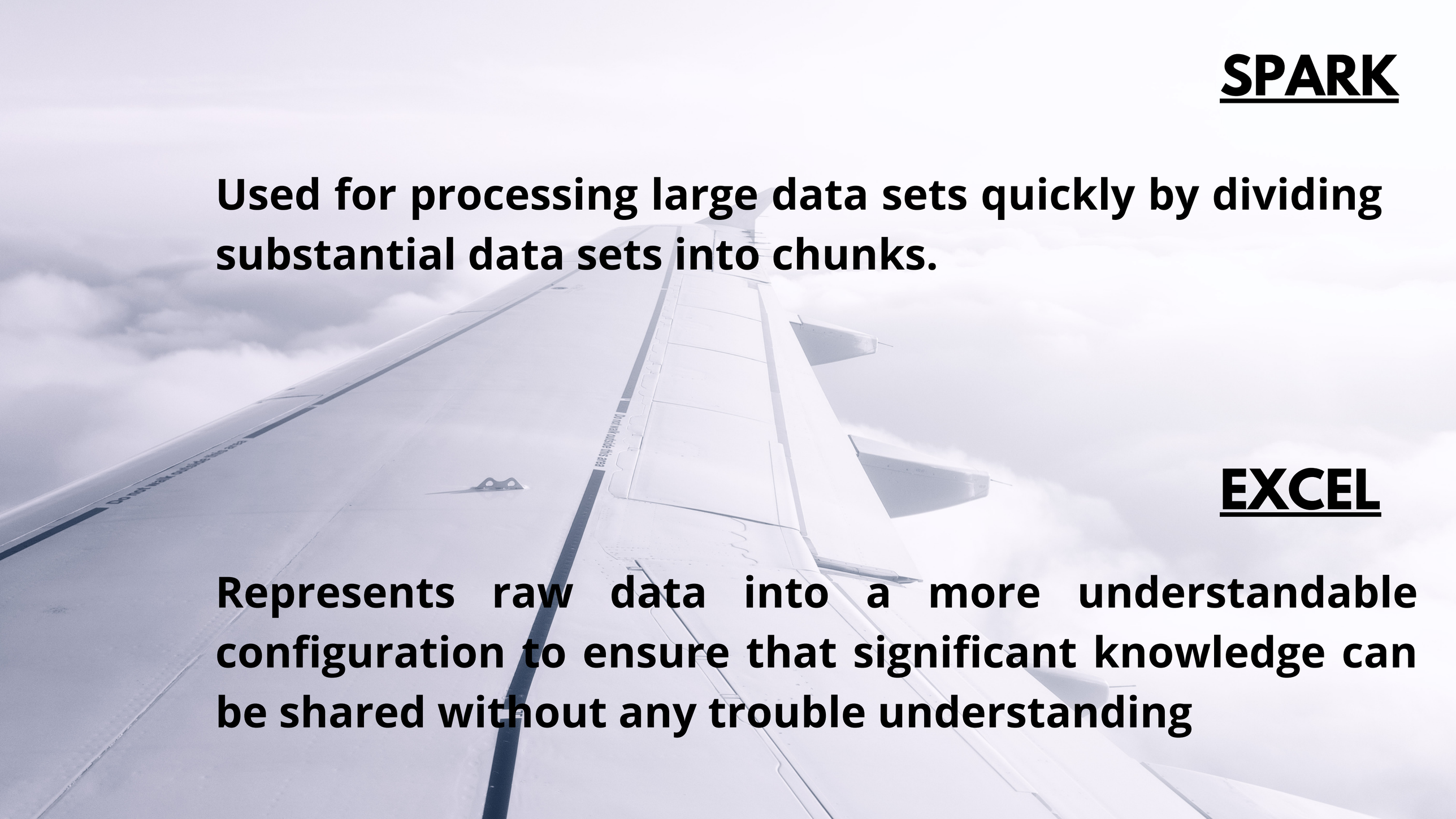
TOOLS USED

PYTHON

Use of python during the processing of large data sets makes the processing easy because the syntaxes and data handling is also simple

R

We use several built-in functions and R packages for statistical computing and graph representations



SPARK

Used for processing large data sets quickly by dividing substantial data sets into chunks.

EXCEL

Represents raw data into a more understandable configuration to ensure that significant knowledge can be shared without any trouble understanding

RESULTS

```
In [9]: cause_delay = sqlContext.sql("SELECT sum(WeatherDelay) Weather,sum(NASDelay) NAS,sum(SecurityDelay) Security,sum(Late
                                      FROM airlineDF ")
```

```
In [10]: df_cause_delay = cause_delay.toPandas()
         df_cause_delay.head()
```

Out[10]:

	Weather	NAS	Security	lateAircraft	Carrier
0	578051	1476657	5353	3243417	2384451

1. Primary cause for flight delay

2. Airports which have the Most Delays

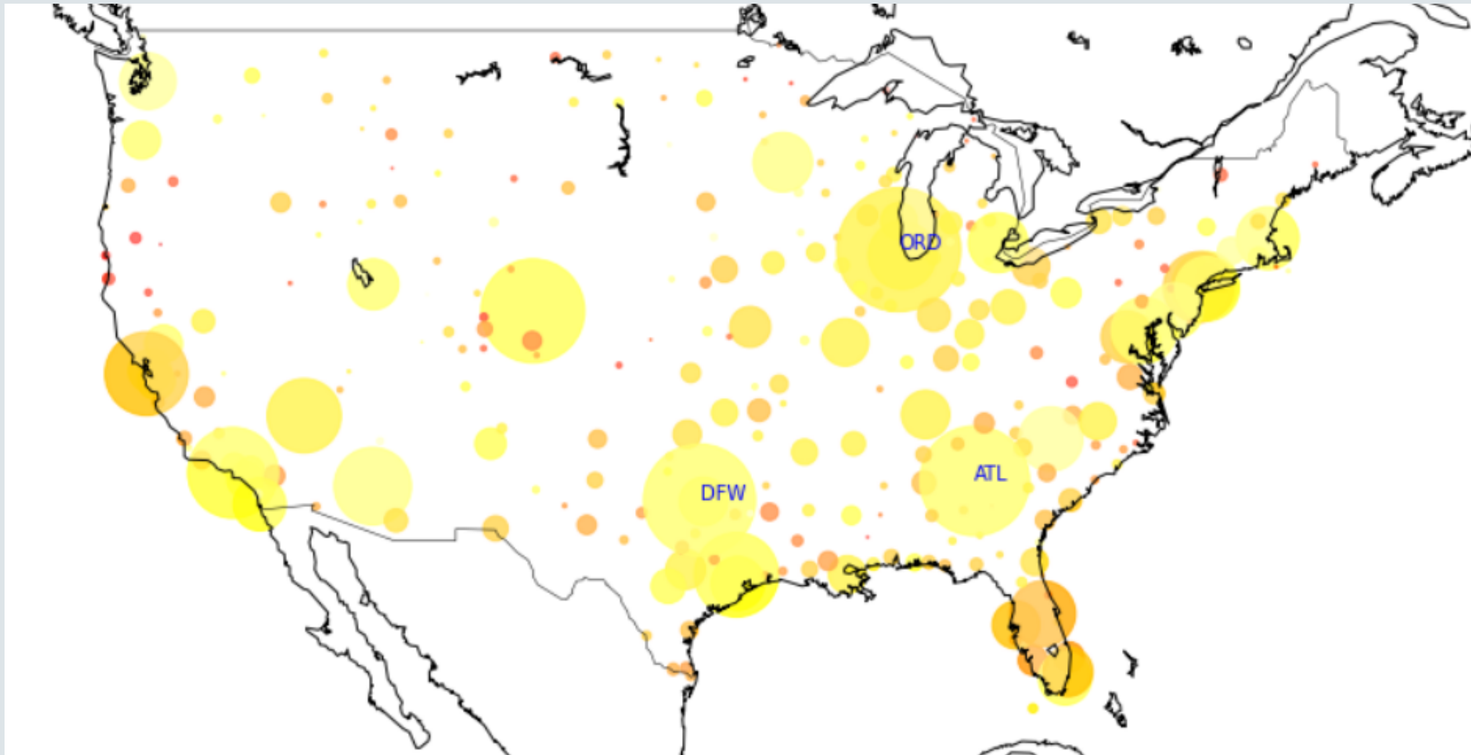
```
In [11]: groupedDelay = sqlContext.sql("SELECT Origin, count(*) conFlight,avg(DepDelay) delay \
                                        FROM airlineDF \
                                        GROUP BY Origin")

         df_origin = groupedDelay.toPandas()
         df_origin.sort_values('delay',ascending=0).head()
```

Out[11]:

	Origin	conFlight	delay
273	BGR	9	161.000000
42	CDV	6	125.166667
282	YAK	8	125.125000
271	SIT	13	110.615385
253	ALO	17	105.823529

Sketch for airports which have most delay



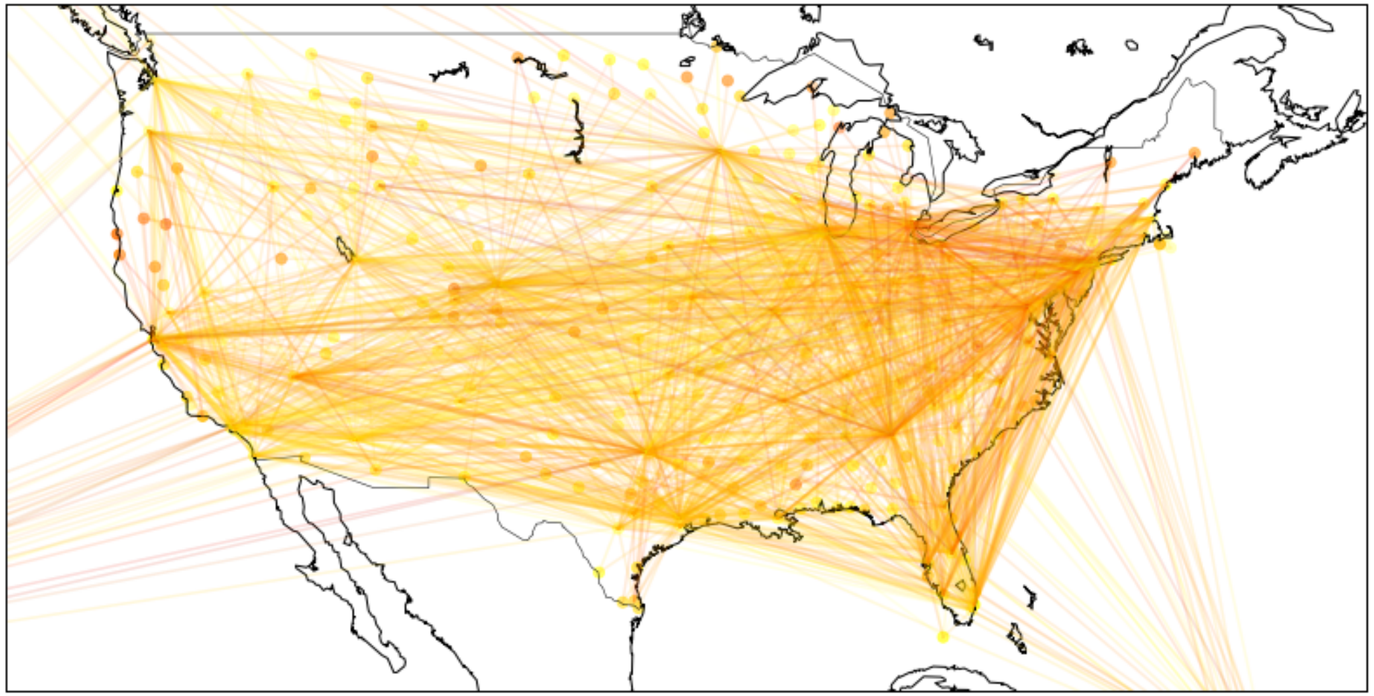
3.Routes which have most delay

```
In [22]: grp_rout_Delay = sqlContext.sql("SELECT Origin, Dest, count(*) traffic,avg(Distance) avgDist,\
                                          avg(DepDelay) avgDelay\
                                        FROM airlineDF \
                                        GROUP BY Origin,Dest")
         rout_Delay = grp_rout_Delay.toPandas()
```

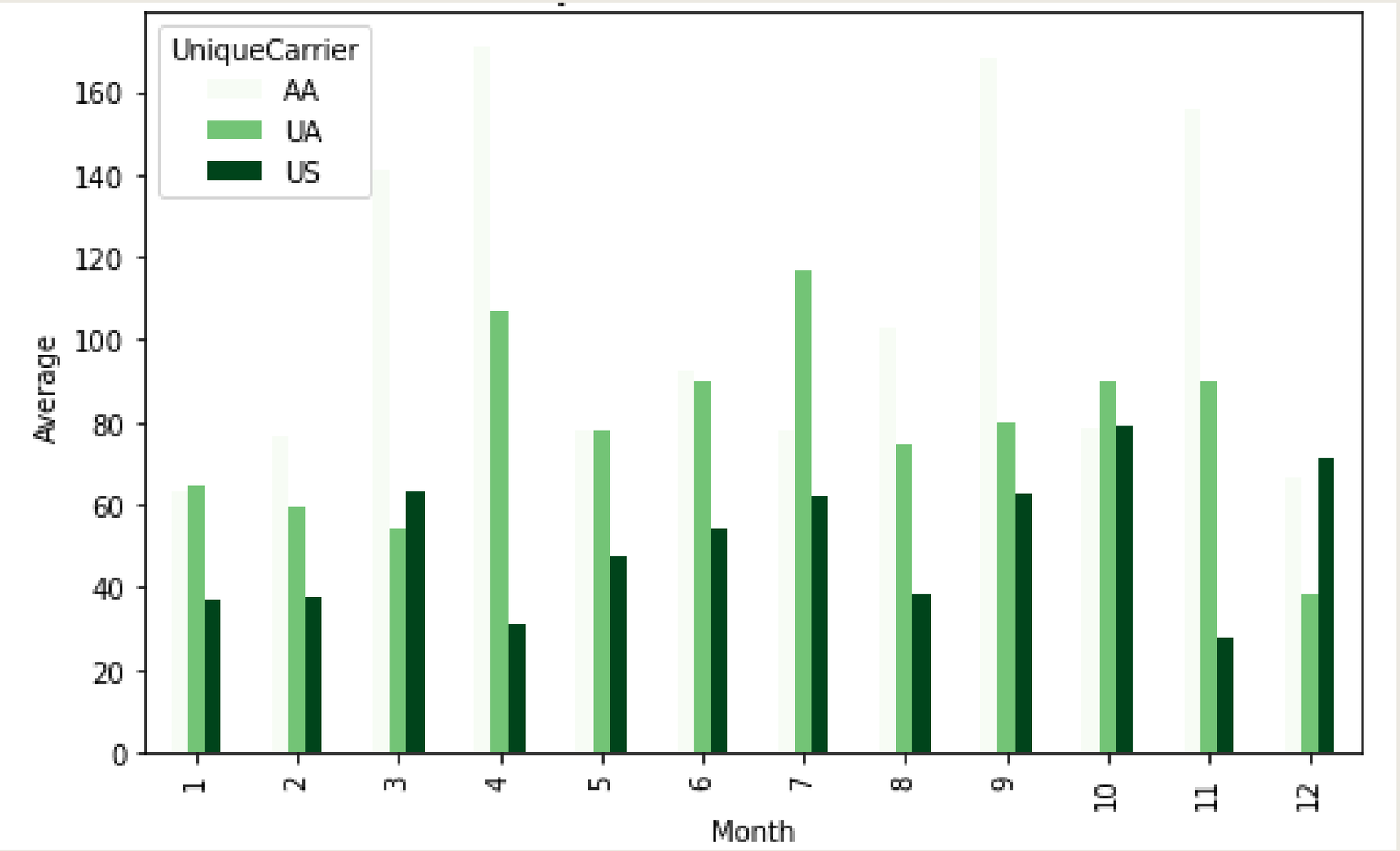
```
In [23]: df_airport_rout1 = pd.merge(rout_Delay, df, left_on = 'Origin', right_on = 'IATA')
         df_airport_rout2 = pd.merge(df_airport_rout1, df, left_on = 'Dest', right_on = 'IATA')
         df_airport_rout = df_airport_rout2[["Origin","lat_x","lng_x","Dest","lat_y","lng_y",\
                                             "avgDelay", "traffic"]]
         df_airport_rout.sort_values('avgDelay',ascending=0).head()
```

Out[23]:

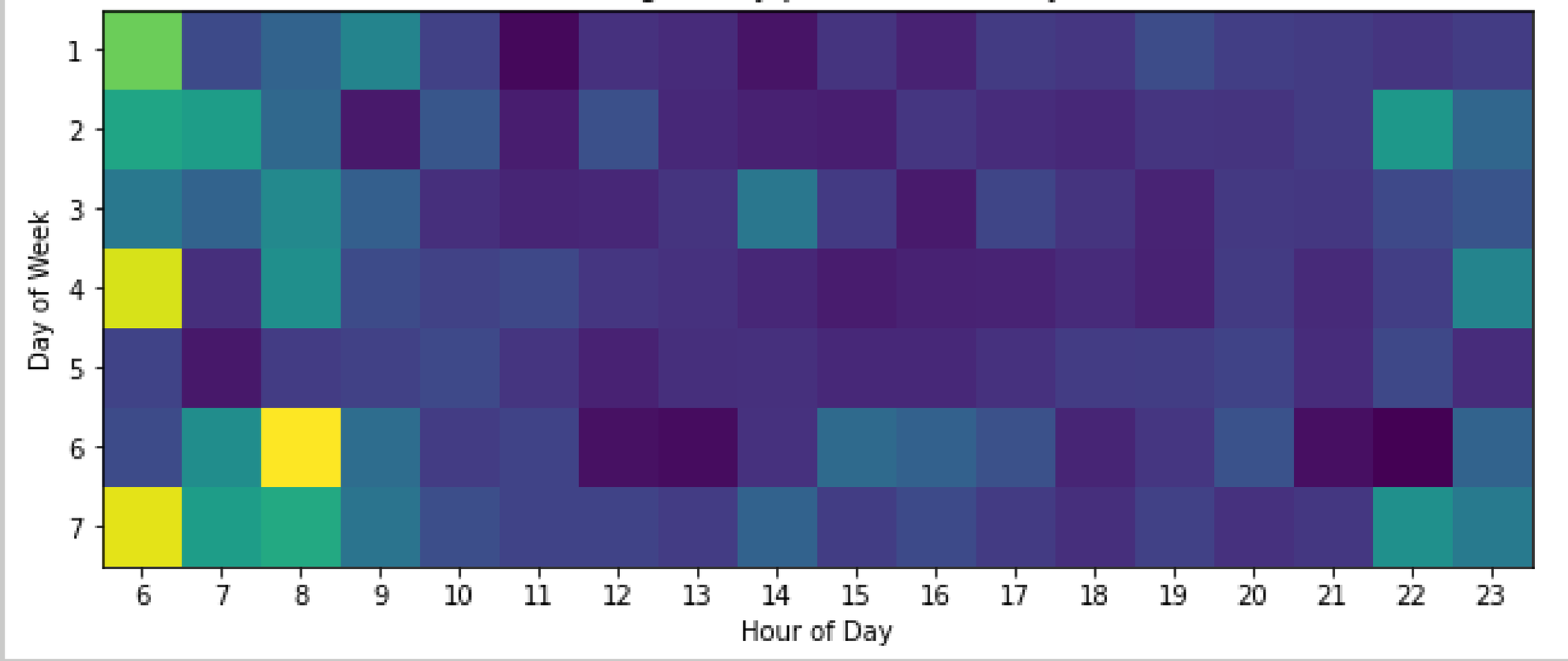
	Origin	lat_x	lng_x	Dest	lat_y	lng_y	avgDelay	traffic
2679	BNA	36.124500	-86.678200	BHM	33.562901	-86.753502	402.0	1
2242	PIT	40.491501	-80.232903	PBI	26.683201	-80.095596	330.0	1
3776	JFK	40.639801	-73.778900	JAC	43.607300	-110.737999	322.0	1
3983	SYR	43.111198	-76.106300	BTV	44.471901	-73.153297	257.0	1
1936	GTF	47.481998	-111.371002	MSP	44.882000	-93.221802	256.5	2



Sketch for Routes which have most delay



4.Airport Origin delay per month



5.Airport Origin delay per day/hour

FINAL ANALYSIS

1. **Which airports have the highest delays**
2. **Which air traffic routes are the most delayed**
3. **When is the best time of day/day of week/time to fly to minimize delays**
4. **Which aircraft's carriers have the most delays**
5. **Do older planes suffer more delays**

CHALLENGES

Tried to find out the prediction for old flights and could'nt find the exact dataset

Lessons Learned

01

From this project we were able to predict, what day and time of the week has the highest probability of a flight getting delayed

02

By querying the datasets, we were also able to come to a conclusion on the list of airports that has the highest probability of an airline getting delayed

03

By analyzing the data, we found that the major reasons for an airline delay is weather, NAS, security, late aircraft and carrier delay


QATAR
القطرية
Thank you!