databricks TEAM_4_1_w261_final_project_raw_EDA

(https://databricks.com)

Team Storage Code Block

Table

	path	name
1	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_4153839644541642250	_committed_41538396
2	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_4230289246598713164	_committed_42302892
3	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_5738663123897603291	_committed_57386631
4	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_7056931953596314476	_committed_70569319
5	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_8764298161920156867	_committed_87642981
6	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_committed_vacuum5449533034058546270	_committed_vacuum54
7	wasbs://fpteam41container@fpteam41.blob.core.windows.net/TP/_started_8764298161920156867	_started_87642981619

10 rows

Set up

Know your mount

Here is the mounting for this class, your source for the original data! Remember, you only have Read access, not Write! Also, become familiar with `dbutils` the equivalent of `gcp` in DataProc

Table

	path	name	size	modificationTime
1	dbfs:/mnt/mids-w261/HW5/	HW5/	0	0
2	dbfs:/mnt/mids-w261/OTPW_12M/	OTPW_12M/	0	0
3	dbfs:/mnt/mids-w261/OTPW_1D_CSV/	OTPW_1D_CSV/	0	0
4	dbfs:/mnt/mids-w261/OTPW_36M/	OTPW_36M/	0	0
5	dbfs:/mnt/mids-w261/OTPW_3M/	OTPW_3M/	0	0
6	dbfs:/mnt/mids-w261/OTPW_3M_2015.csv	OTPW_3M_2015.csv	1500620247	1679772070000

10 rows

Table

	path	name
1	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/_SUCCESS	_SUCCESS
2	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/_committed_6166100735690431139	_committed_61661007
3	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/_started_6166100735690431139	_started_61661007356
4	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/part-00000-tid-6166100735690431139-8bc743e4-6a75-4b12-abb3-d11b3074da5b-10783-1-c000.snappy.parquet	part-00000-tid-616610
5	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/part-00109-tid-6166100735690431139-8bc743e4-6a75-4b12-abb3-d11b3074da5b-10773-1-c000.snappy.parquet	part-00109-tid-616610
6	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/part-00110-tid-6166100735690431139-8bc743e4-6a75-4b12-abb3-d11b3074da5b-10786-1-c000.snappy.parquet	part-00110-tid-616610
7	dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_weather_data_1y/part-00111-tid-6166100735690431139-8bc743e4-6a75-4b12-abb3-d11b3074da5b-10778-1-c000.snappy.parquet	part-00111-tid-616610

42 rows

Data for the Project

For the project you will have 4 sources of data:

1. Airlines Data: This is the raw data of flights information. You have 3 months, 6 months, 1 year, and full data from 2015 to 2019. Remember the maxima: "Test, Test, Test", so a lot of testing in smaller samples before scaling up! Location of the data?

`dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_airlines_data/`, `dbfs:/mnt/mids-w261/datasets_final_project_2022/parquet_airlines_data_1y/`, etc. (Below the dbutils to get the folders)

2. Weather Data: Raw data for weather information. Same as before, we are sharing 3 months, 6 months, 1 year
3. Stations data: Extra information of the location of the different weather stations. Location `dbfs:/mnt/mids-w261/datasets_final_project_2022/stations_data/stations_with_neighbors.parquet/`
4. OTPW Data: This is our joined data (We joined Airlines and Weather). This is the main dataset for your project, the previous 3 are given for reference. You can attempt your own join for Extra Credit. Location `dbfs:/mnt/mids-w261/OTPW_60M/` and more,

Table

	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	OP_UNIQUE_CARRIER	OP_CARRIER_AIR
1	1	2	19	4	2015-02-19	AA	19805
2	1	2	20	5	2015-02-20	AA	19805
3	1	2	21	6	2015-02-21	AA	19805
4	1	2	22	7	2015-02-22	AA	19805
5	1	2	23	1	2015-02-23	AA	19805
6	1	2	24	2	2015-02-24	AA	19805
7	1	2	25	3	2015-02-25	AA	19805
8	1	2	26	4	2015-02-26	AA	19805
9	1	2	27	5	2015-02-27	AA	19805
10	1	2	28	6	2015-02-28	AA	19805
11	1	2	1	7	2015-02-01	AA	19805
12	1	2	2	1	2015-02-02	AA	19805
13	1	2	3	2	2015-02-03	AA	19805
14	1	2	4	3	2015-02-04	AA	19805

3,505 rows | Truncated data

Table

	usaf	wban	station_id	lat	lon	neighbor_id	neighbor_name
1	690020	93218	69002093218	36	-121.233	69002093218	JOLON HUNTER LIGGETT MIL RES
2	690020	93218	69002093218	36	-121.233	69007093217	FRITZSCHE AAF
3	690020	93218	69002093218	36	-121.233	69014093101	EL TORO MCAS
4	690020	93218	69002093218	36	-121.233	70027127506	BARROW POINT BARROW
5	690020	93218	69002093218	36	-121.233	70045027512	LONELY
6	690020	93218	69002093218	36	-121.233	70063027403	OLIKTOK POW 2
7	690020	93218	69002093218	36	-121.233	70063526465	GALBRAITH LAKE AIRPORT
8	690020	93218	69002093218	36	-121.233	70063627405	PRUDHOE BAY
9	690020	93218	69002093218	36	-121.233	70104626418	CENTRAL AIRPORT

10,000 rows | Truncated data

Table

	STATION	DATE	LATITUDE	LONGITUDE	ELEVATION	NAME	REPORT
1	52652099999	2015-01-01T02:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12
2	52652099999	2015-01-01T05:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12
3	52652099999	2015-01-01T08:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12
4	52652099999	2015-01-01T11:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12
5	52652099999	2015-01-01T14:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12
6	52652099999	2015-01-01T17:00:00	39.0833333	100.2833333	1462.0	ZHANGYE, CH	FM-12

2,668 rows | Truncated data

Example of EDA

EDA on Raw Data

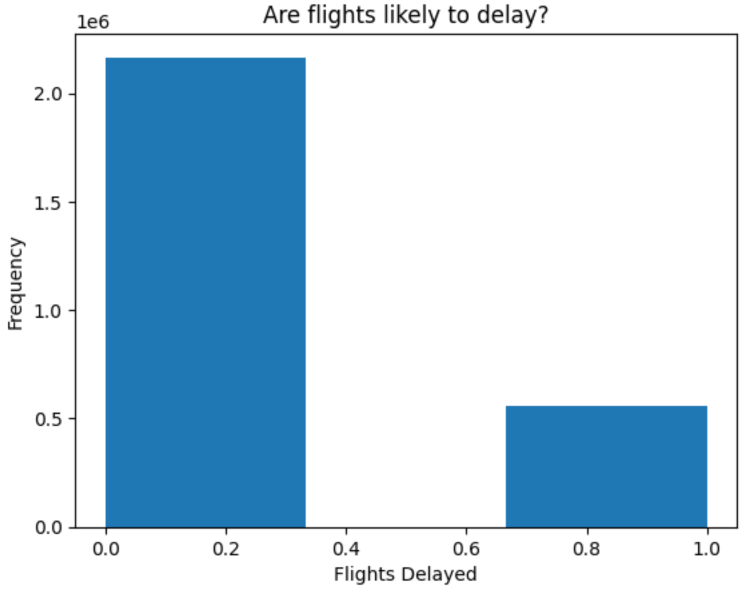
Flights

3 Months sample

```
[Row(QUARTER=1, MONTH=2, DAY_OF_MONTH=19, DAY_OF_WEEK=4, FL_DATE='2015-02-19', OP_UNIQUE_CARRIER='AA', OP_CARRIER_AIRLINE
_ID=19805, OP_CARRIER='AA', TAIL_NUM='N520AA', OP_CARRIER_FL_NUM=323, ORIGIN_AIRPORT_ID=15016, ORIGIN_AIRPORT_SEQ_ID=1501
603, ORIGIN_CITY_MARKET_ID=31123, ORIGIN='STL', ORIGIN_CITY_NAME='St. Louis, MO', ORIGIN_STATE_ABR='MO', ORIGIN_STATE_FIP
S=29, ORIGIN_STATE_NM='Missouri', ORIGIN_WAC=64, DEST_AIRPORT_ID=11298, DEST_AIRPORT_SEQ_ID=1129803, DEST_CITY_MARKET_ID=
30194, DEST='DFW', DEST_CITY_NAME='Dallas/Fort Worth, TX', DEST_STATE_ABR='TX', DEST_STATE_FIPS=48, DEST_STATE_NM='Texa
s', DEST_WAC=74, CRS_DEP_TIME=901, DEP_TIME=949, DEP_DELAY=48.0, DEP_DELAY_NEW=48.0, DEP_DEL15=1.0, DEP_DELAY_GROUP=3, DE
P_TIME_BLK='0900-0959', TAXI_OUT=23.0, WHEELS_OFF=1012, WHEELS_ON=1130, TAXI_IN=5.0, CRS_ARR_TIME=1058, ARR_TIME=1135, AR
R_DELAY=37.0, ARR_DELAY_NEW=37.0, ARR_DEL15=1.0, ARR_DELAY_GROUP=2, ARR_TIME_BLK='1000-1059', CANCELLED=0.0, CANCELLATION
_CODE=None, DIVERTED=0.0, CRS_ELAPSED_TIME=117.0, ACTUAL_ELAPSED_TIME=106.0, AIR_TIME=78.0, FLIGHTS=1.0, DISTANCE=550.0,
DISTANCE_GROUP=3, CARRIER_DELAY=0.0, WEATHER_DELAY=5.0, NAS_DELAY=0.0, SECURITY_DELAY=0.0, LATE_AIRCRAFT_DELAY=32.0, FIRS
T_DEP_TIME=None, TOTAL_ADD_GTIME=None, LONGEST_ADD_GTIME=None, DIV_AIRPORT_LANDINGS=0, DIV_REACHED_DEST=None, DIV_ACTUAL_
ELAPSED_TIME=None, DIV_ARR_DELAY=None, DIV_DISTANCE=None, DIV1_AIRPORT=None, DIV1_AIRPORT_ID=None, DIV1_AIRPORT_SEQ_ID=No
ne, DIV1_WHEELS_ON=None, DIV1_TOTAL_GTIME=None, DIV1_LONGEST_GTIME=None, DIV1_WHEELS_OFF=None, DIV1_TAIL_NUM=None, DIV2_A
IRPORT=None, DIV2_AIRPORT_ID=None, DIV2_AIRPORT_SEQ_ID=None, DIV2_WHEELS_ON=None, DIV2_TOTAL_GTIME=None, DIV2_LONGEST_GTI
ME=None, DIV2_WHEELS_OFF=None, DIV2_TAIL_NUM=None, DIV3_AIRPORT=None, DIV3_AIRPORT_ID=None, DIV3_AIRPORT_SEQ_ID=None, DIV
3_WHEELS_ON=None, DIV3_TOTAL_GTIME=None, DIV3_LONGEST_GTIME=None, DIV3_WHEELS_OFF=None, DIV3_TAIL_NUM=None, DIV4_AIRPORT=
None, DIV4_AIRPORT_ID=None, DIV4_AIRPORT_SEQ_ID=None, DIV4_WHEELS_ON=None, DIV4_TOTAL_GTIME=None, DIV4_LONGEST_GTIME=Non
e, DIV4_WHEELS_OFF=None, DIV4_TAIL_NUM=None, DIV5_AIRPORT=None, DIV5_AIRPORT_ID=None, DIV5_AIRPORT_SEQ_ID=None, DIV5_WHEE
LS_ON=None, DIV5_TOTAL_GTIME=None, DIV5_LONGEST_GTIME=None, DIV5_WHEELS_OFF=None, DIV5_TAIL_NUM=None, YEAR=2015, DEP_DELA
Y_GT_15=1)]
```

```
2806942
```

Delay Plots



Note that we have a huge class imbalance.

Let's also take a look at the delays by carrier, departing airport, and time periods.

Table

	OP_UNIQUE_CARRIER	Delay_Count	Delay_Ratio
1	UA	59394	0.2566724286949006
2	NK	13158	0.2555546923555003
3	AA	50266	0.20094664715805968
4	EV	55650	0.195139911634757
5	B6	32138	0.2628144319780184
6	DL	63980	0.16268307567127746
7	OO	53046	0.19057438889447742

14 rows

Table

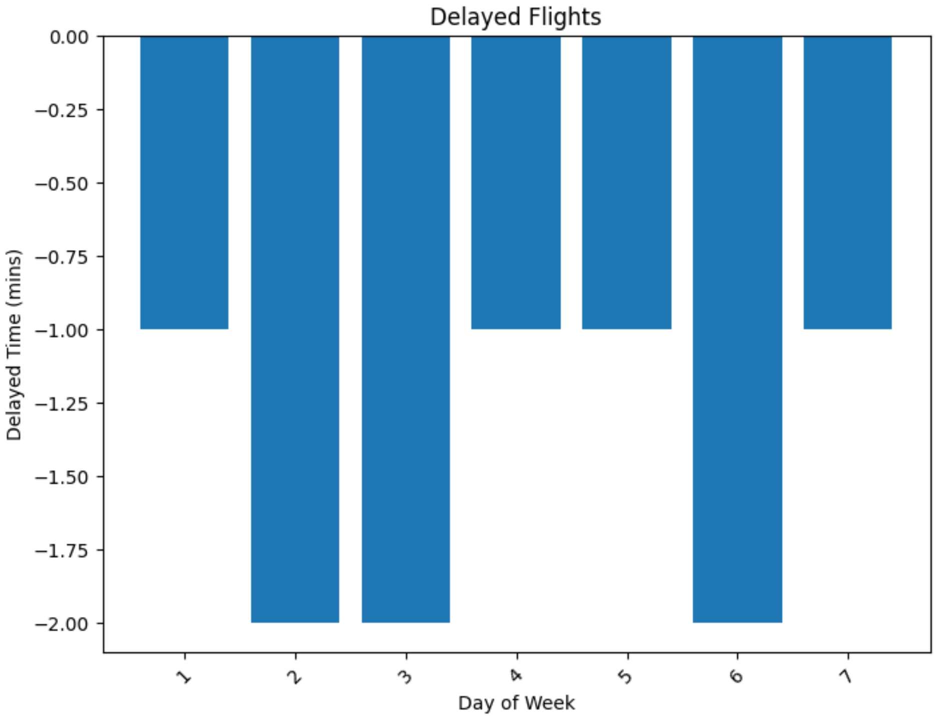
	ORIGIN	ORIGIN_CITY_NAME	Delay_Count	Delay_Ratio
1	COS	Colorado Springs, CO	584	0.16666666666666666
2	SDF	Louisville, KY	1020	0.18653986832479882
3	CLL	College Station/Bryan, TX	192	0.17777777777777778
4	MSN	Madison, WI	814	0.1726039016115352

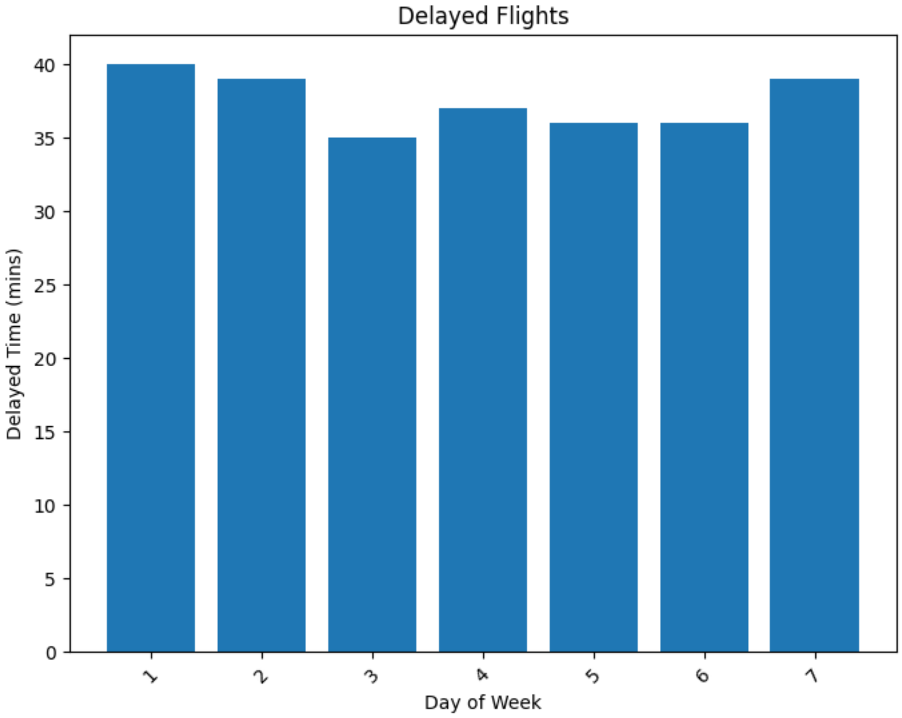
6	CMI	Champaign/Urbana, IL	256	0.24150943396226415

315 rows

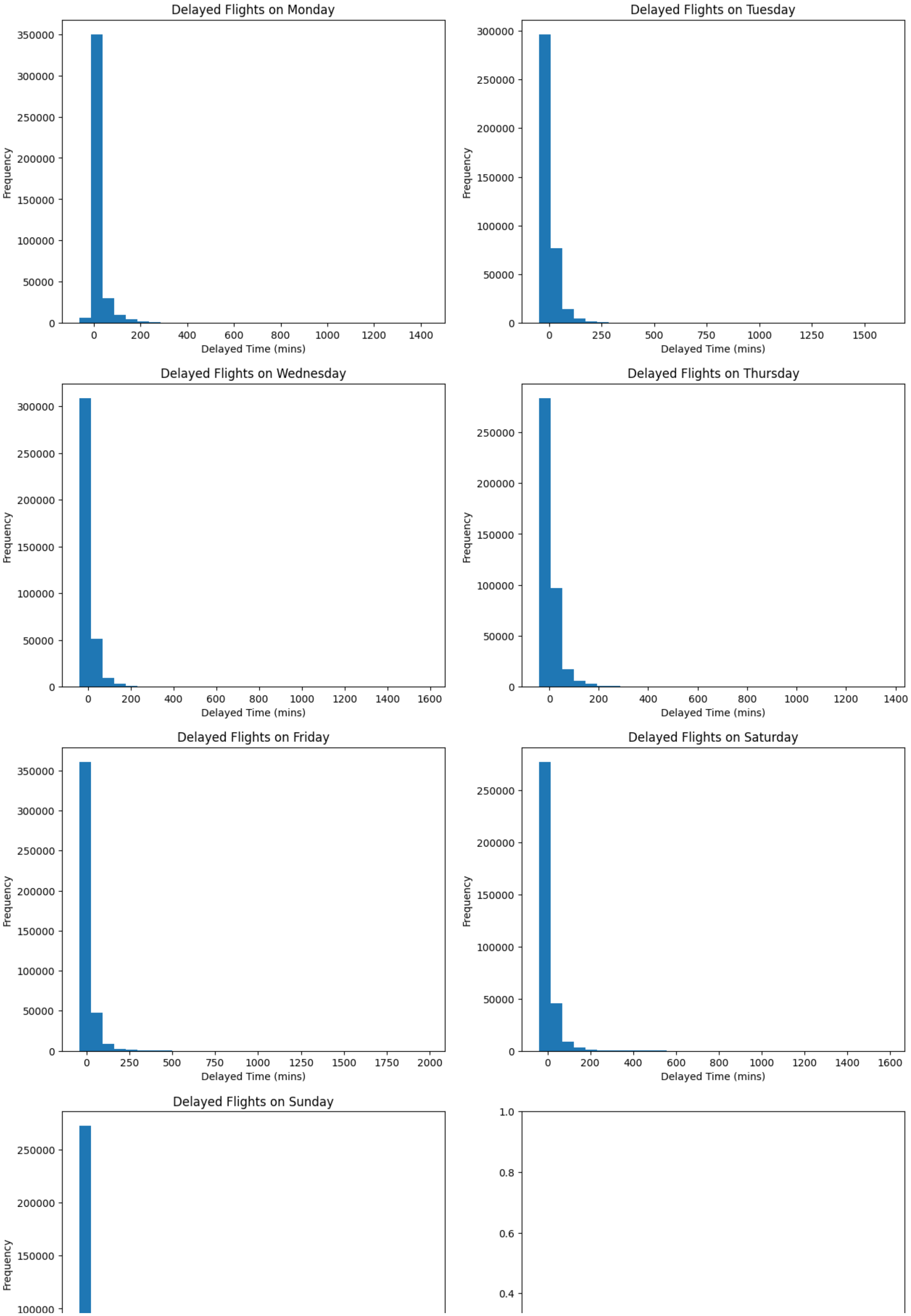
By carrier, the proportion of delayed flights ranges from 8% (HA) to 30% (F9).

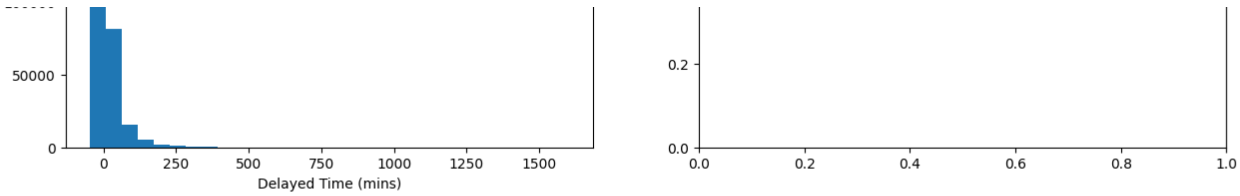
By departing airport, the proportion of delayed flights ranges from 1.7% (BTM in Butte, MT) to 37% (ILG in Wilmington, DE)





Looks like we don't have much delays, but out of the delayed flights, the median time delayed is about 35 to 40 minutes for most days.


Delayed Flights on Monday
Frequency
0
50000
100000
150000
200000
250000
300000
350000
0
200
400
600
800
1000
1200
1400
Delayed Time (mins)
Delayed Flights on Tuesday
Frequency
0
50000
100000
150000
200000
250000
300000
0
250
500
750
1000
1250
1500
Delayed Time (mins)
Delayed Flights on Wednesday
Frequency
0
50000
100000
150000
200000
250000
300000
0
200
400
600
800
1000
1200
1400
1600
Delayed Time (mins)
Delayed Flights on Thursday
Frequency
0
50000
100000
150000
200000
250000
0
200
400
600
800
1000
1200
1400
Delayed Time (mins)
Delayed Flights on Friday
Frequency
0
50000
100000
150000
200000
250000
300000
350000
0
250
500
750
1000
1250
1500
1750
2000
Delayed Time (mins)
Delayed Flights on Saturday
Frequency
0
50000
100000
150000
200000
250000
0
200
400
600
800
1000
1200
1400
1600
Delayed Time (mins)
Delayed Flights on Sunday
Frequency
100000
150000
200000
250000
0.4
0.6
0.8
1.0



```
# Look at weather delays only flight_3m_wk_wd = df_flights.select(['DAY_OF_WEEK' ...
```

Show cell

Features:

QUARTER=1, MONTH=2, DAY_OF_MONTH=19, DAY_OF_WEEK=4, FL_DATE='2015-02-19', OP_UNIQUE_CARRIER='AA', OP_CARRIER_AIRLINE_ID=19805, OP_CARRIER='AA', TAIL_NUM='N520AA', OP_CARRIER_FL_NUM=323, ORIGIN_AIRPORT_ID=15016, ORIGIN_AIRPORT_SEQ_ID=1501603, ORIGIN_CITY_MARKET_ID=31123, ORIGIN='STL', ORIGIN_CITY_NAME='St. Louis, MO', ORIGIN_STATE_ABR='MO', ORIGIN_STATE_FIPS=29, ORIGIN_STATE_NM='Missouri', ORIGIN_WAC=64, DEST_AIRPORT_ID=11298, DEST_AIRPORT_SEQ_ID=1129803, DEST_CITY_MARKET_ID=30194, DEST='DFW', DEST_CITY_NAME='Dallas/Fort Worth, TX', DEST_STATE_ABR='TX', DEST_STATE_FIPS=48, DEST_STATE_NM='Texas', DEST_WAC=74, CRS_DEP_TIME=901, DEP_TIME=949, DEP_DELAY=48.0, DEP_DELAY_NEW=48.0, DEP_DEL15=1.0, DEP_DELAY_GROUP=3, DEP_TIME_BLK='0900-0959', TAXI_OUT=23.0, WHEELS_OFF=1012, WHEELS_ON=1130, TAXI_IN=5.0, CRS_ARR_TIME=1058, ARR_TIME=1135, ARR_DELAY=37.0, ARR_DELAY_NEW=37.0, ARR_DEL15=1.0, ARR_DELAY_GROUP=2, ARR_TIME_BLK='1000-1059', CANCELLED=0.0, CANCELLATION_CODE=None, DIVERTED=0.0, CRS_ELAPSED_TIME=117.0, ACTUAL_ELAPSED_TIME=106.0, AIR_TIME=78.0, FLIGHTS=1.0, DISTANCE=550.0, DISTANCE_GROUP=3, CARRIER_DELAY=0.0, WEATHER_DELAY=5.0, NAS_DELAY=0.0, SECURITY_DELAY=0.0, LATE_AIRCRAFT_DELAY=32.0, FIRST_DEP_TIME=None, TOTAL_ADD_GTIME=None, LONGEST_ADD_GTIME=None, DIV_AIRPORT_LANDINGS=0, DIV_REACHED_DEST=None, DIV_ACTUAL_ELAPSED_TIME=None, DIV_ARR_DELAY=None, DIV_DISTANCE=None, DIV1_AIRPORT=None, DIV1_AIRPORT_ID=None, DIV1_AIRPORT_SEQ_ID=None, DIV1_WHEELS_ON=None, DIV1_TOTAL_GTIME=None, DIV1_LONGEST_GTIME=None, DIV1_WHEELS_OFF=None, DIV1_TAIL_NUM=None, DIV2_AIRPORT=None, DIV2_AIRPORT_ID=None, DIV2_AIRPORT_SEQ_ID=None, DIV2_WHEELS_ON=None, DIV2_TOTAL_GTIME=None, DIV2_LONGEST_GTIME=None, DIV2_WHEELS_OFF=None, DIV2_TAIL_NUM=None, DIV3_AIRPORT=None, DIV3_AIRPORT_ID=None, DIV3_AIRPORT_SEQ_ID=None, DIV3_WHEELS_ON=None, DIV3_TOTAL_GTIME=None, DIV3_LONGEST_GTIME=None, DIV3_WHEELS_OFF=None, DIV3_TAIL_NUM=None, DIV4_AIRPORT=None, DIV4_AIRPORT_ID=None, DIV4_AIRPORT_SEQ_ID=None, DIV4_WHEELS_ON=None, DIV4_TOTAL_GTIME=None, DIV4_LONGEST_GTIME=None, DIV4_WHEELS_OFF=None, DIV4_TAIL_NUM=None, DIV5_AIRPORT=None, DIV5_AIRPORT_ID=None, DIV5_AIRPORT_SEQ_ID=None, DIV5_WHEELS_ON=None, DIV5_TOTAL_GTIME=None, DIV5_LONGEST_GTIME=None, DIV5_WHEELS_OFF=None, DIV5_TAIL_NUM=None, YEAR=2015, DEP_DELAY_GT_15=1

Correlation Matrix and heatmap

```
DEP_DELAY              float64
DEP_DEL15              float64
DISTANCE               float64
CARRIER_DELAY          float64
WEATHER_DELAY          float64
NAS_DELAY              float64
SECURITY_DELAY         float64
```

```
LATE_AIRCRAFT_DELAY    float64
dtype: object
```

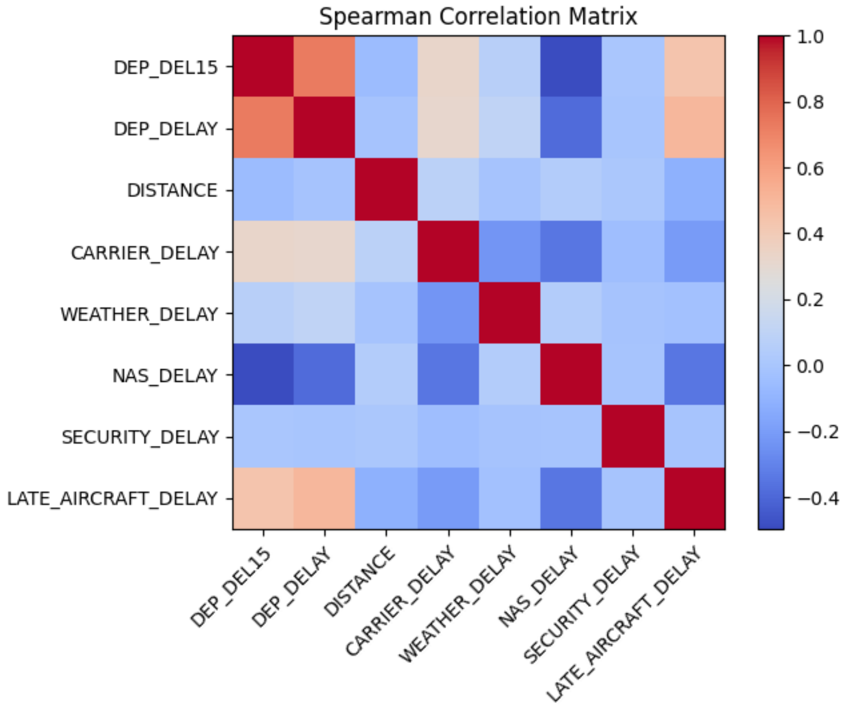
```
#pd.plotting.scatter_matrix(flight_num_features_df, diagonal = 'hist')
```

Show cell

```
flight_num_features = df_flights.select(['DEP_DEL15','DEP_DELAY','DISTANCE','CAR ...
```

Show cell

```
[[ 1.          0.730949   -0.0566369   0.32540508  0.06842479 -0.49600024
   0.00431242  0.43581934]
 [ 0.730949    1.         -0.01615092  0.31378081  0.10018604 -0.38850939
  -0.00388942  0.49812293]
 [-0.0566369  -0.01615092  1.          0.07837497 -0.01327107  0.04245024
   0.01204205 -0.11535873]
 [ 0.32540508  0.31378081  0.07837497  1.         -0.23206464 -0.34414598
  -0.04390785 -0.20519372]
 [ 0.06842479  0.10018604 -0.01327107 -0.23206464  1.          0.04521814
  -0.01225962 -0.02531354]
 [-0.49600024 -0.38850939  0.04245024 -0.34414598  0.04521814  1.
  -0.00998186 -0.34622726]
 [ 0.00431242 -0.00388942  0.01204205 -0.04390785 -0.01225962 -0.00998186
   1.         -0.01029319]
 [ 0.43581934  0.49812293 -0.11535873 -0.20519372 -0.02531354 -0.34622726
  -0.01029319  1.        ]]
```



Cross-Referencing "Cancelled" and "Dep_Del15"

```
Total Cancelled Flights: 87002
Cancelled and Delayed Flights: 1310
All Cancelled Flights are Delayed: False
```

Total Cancelled Flights: 87002

Cancelled and Delayed Flights: 1310

All Cancelled Flights are Delayed: False

IMPORTANT: Some canceled flights are marked as delayed.

How many flights are cancelled due to weather?

The CancelationCode column specifies the reason for cancellation and has the following codes:

A. Carrier

B. Weather

C. National Air System

D. Security

Do we need to pay attention to canceled flights due to weather and take them into account for delay prediction?

Flights cancelled due to weather

```
Total Weather Cancellations: 58662
Total Weather Cancellations Minus Weather Cancellations Marked as Delayed: 58170
```

```
Weather Cancellations Marked as Delayed: 756
Weather Cancellations Not Marked as Delayed: 492
```

Total Weather Cancellations: 58662

Total Weather Cancellations Minus Weather Cancellations Marked as Delayed: 58170

Weather Cancellations Marked as Delayed: 756

Weather Cancellations Not Marked as Delayed: 492

Checking for missing values

```
   QUARTER  MONTH  DAY_OF_MONTH  ...  DIV5_WHEELS_OFF  DIV5_TAIL_NUM  YEAR
0        0      0             0  ...          2806942        2806942     0

[1 rows x 109 columns]
```

Total number of features with missing values

```
   QUARTER  MONTH  DAY_OF_MONTH  ...  DIV5_WHEELS_OFF  DIV5_TAIL_NUM  YEAR
0        0      0             0  ...          2806942        2806942     0

[1 rows x 109 columns]
Total number of entries: 2806942
Total number of features with missing values: 71
```

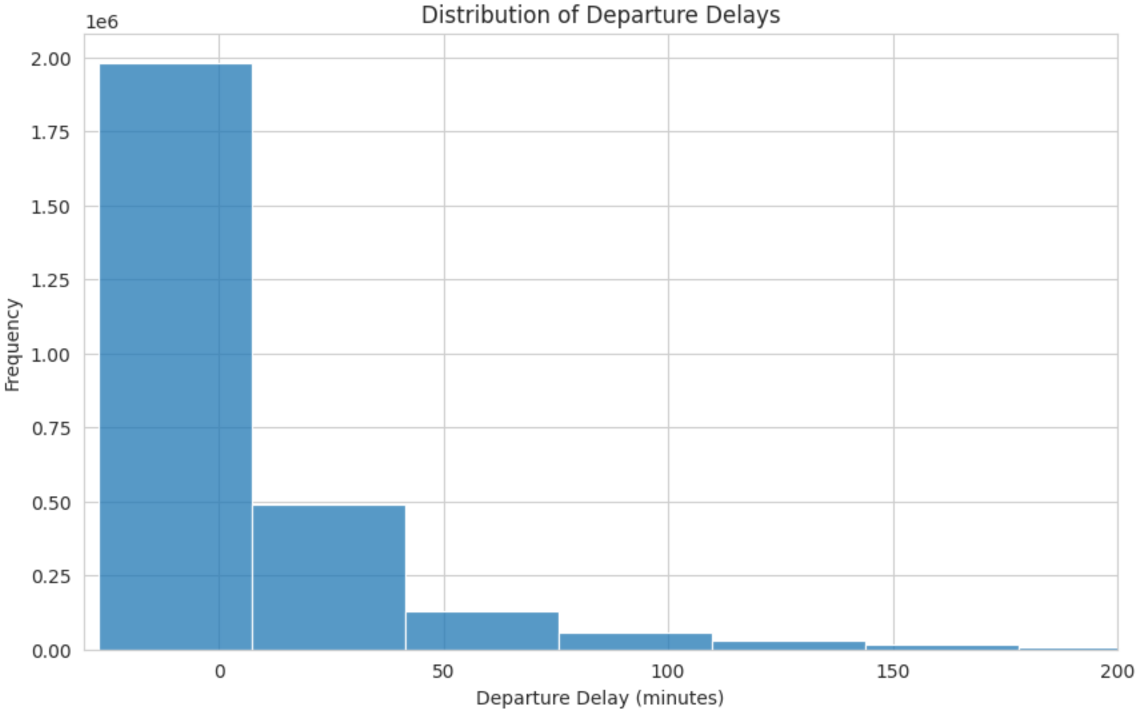
Table of missing value percentages for each feature

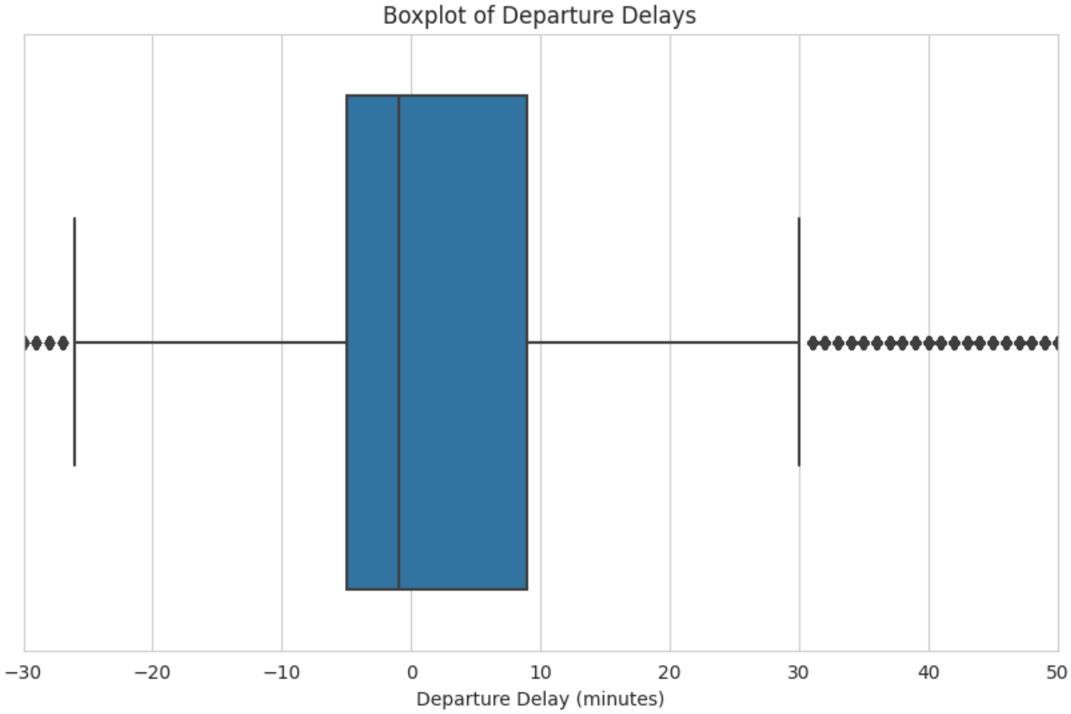
```
+-----------+------------------+-------------------+
|Num missing|Perc missing      |Variable           |
+-----------+------------------+-------------------+
|2719940    |96.90047033390786 |CANCELLATION_CODE  |
|2233778    |79.58048295974766 |CARRIER_DELAY      |
|2233778    |79.58048295974766 |WEATHER_DELAY      |
|2233778    |79.58048295974766 |NAS_DELAY          |
|2233778    |79.58048295974766 |SECURITY_DELAY     |
|2233778    |79.58048295974766 |LATE_AIRCRAFT_DELAY|
|93314      |3.324400717934321 |ACTUAL_ELAPSED_TIME|
|93314      |3.324400717934321 |ARR_DELAY          |
|93314      |3.324400717934321 |AIR_TIME           |
|93314      |3.324400717934321 |ARR_DELAY_GROUP    |
|93314      |3.324400717934321 |ARR_DELAY_NEW      |
|93314      |3.324400717934321 |ARR_DEL15          |
|88736      |3.1613050786229286|WHEELS_ON          |
|88736      |3.1613050786229286|ARR_TIME           |
|88736      |3.1613050786229286|TAXI_IN            |
|86342      |3.0760165332949523|TAXI_OUT           |
|86342      |3.0760165332949523|WHEELS_OFF         |
|84710      |3.0178749685600916|DEP_DEL15          |
```

Variable	Num Missing	Perc Missing
CANCELLATION_CODE	2,719,940	96.90047033390786
CARRIER_DELAY	2,233,778	79.58048295974766
WEATHER_DELAY	2,233,778	79.58048295974766
NAS_DELAY	2,233,778	79.58048295974766
SECURITY_DELAY	2,233,778	79.58048295974766
LATE_AIRCRAFT_DELAY	2,233,778	79.58048295974766
ACTUAL_ELAPSED_TIME	93,314	3.324400717934321
ARR_DELAY	93,314	3.324400717934321
AIR_TIME	93,314	3.324400717934321
ARR_DELAY_GROUP	93,314	3.324400717934321
ARR_DELAY_NEW	93,314	3.324400717934321

Variable	Num Missing	Perc Missing
ARR_DEL15	93,314	3.324400717934321
WHEELS_ON	88,736	3.1613050786229286
ARR_TIME	88,736	3.1613050786229286
TAXI_IN	88,736	3.1613050786229286
TAXI_OUT	86,342	3.0760165332949523
WHEELS_OFF	86,342	3.0760165332949523
DEP_DEL15	84,710	3.0178749685600916
DEP_TIME	84,710	3.0178749685600916

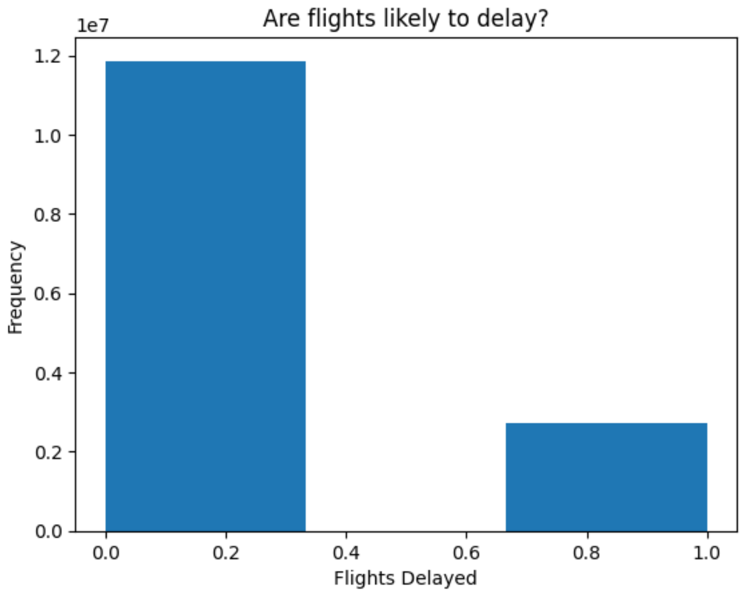
Distribution and Boxplot of Departure Delays





1 Year sample

Delay Plots



Table

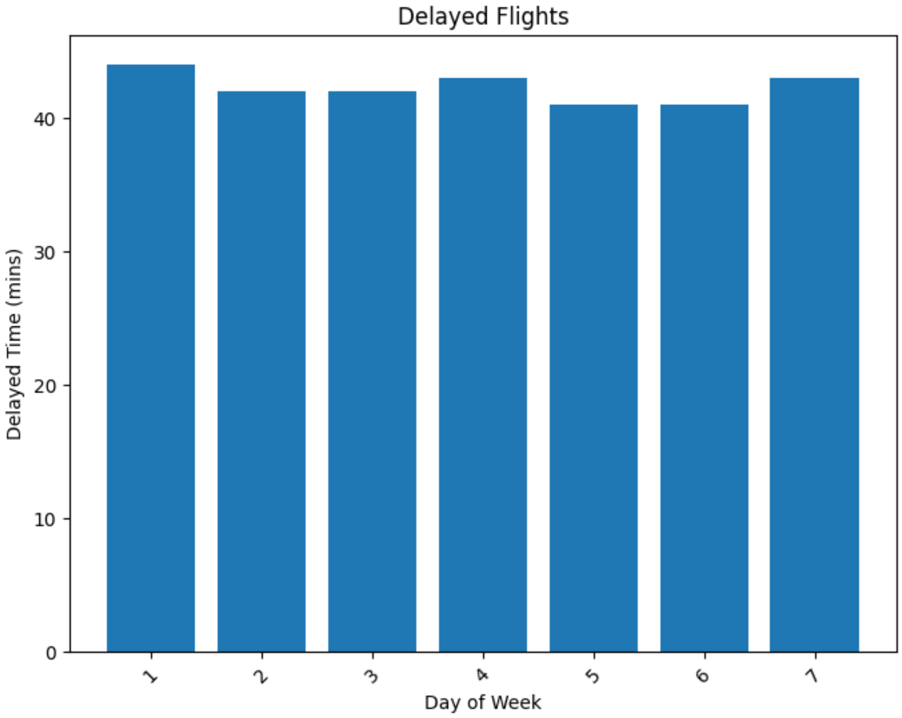
	OP_UNIQUE_CARRIER	Delay_Count	Delay_Ratio
1	F9	68828	0.25798954967651977
2	B6	151032	0.2570661183751417
3	EV	56418	0.2191654170972178
4	WN	551740	0.20732783304950106
5	AA	371316	0.2001815735221813
6	UA	241716	0.19469140595424692

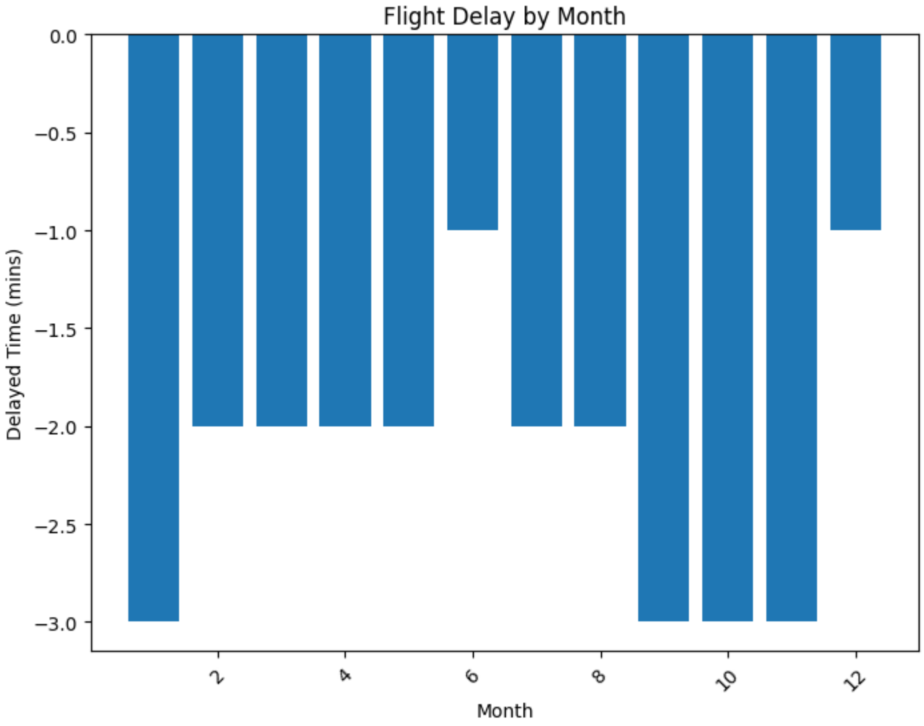
17 rows

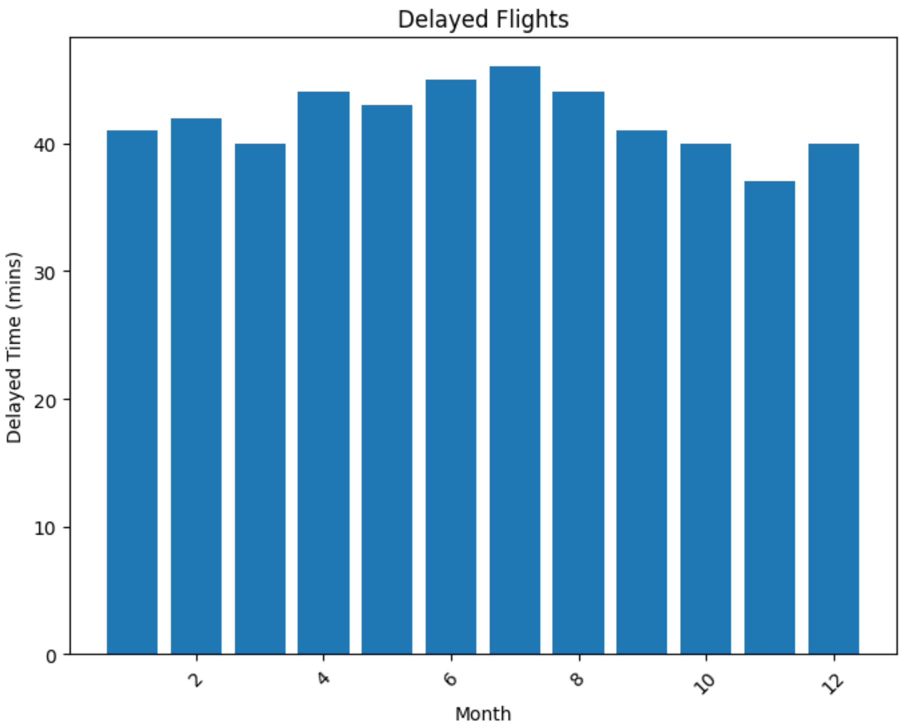
Table

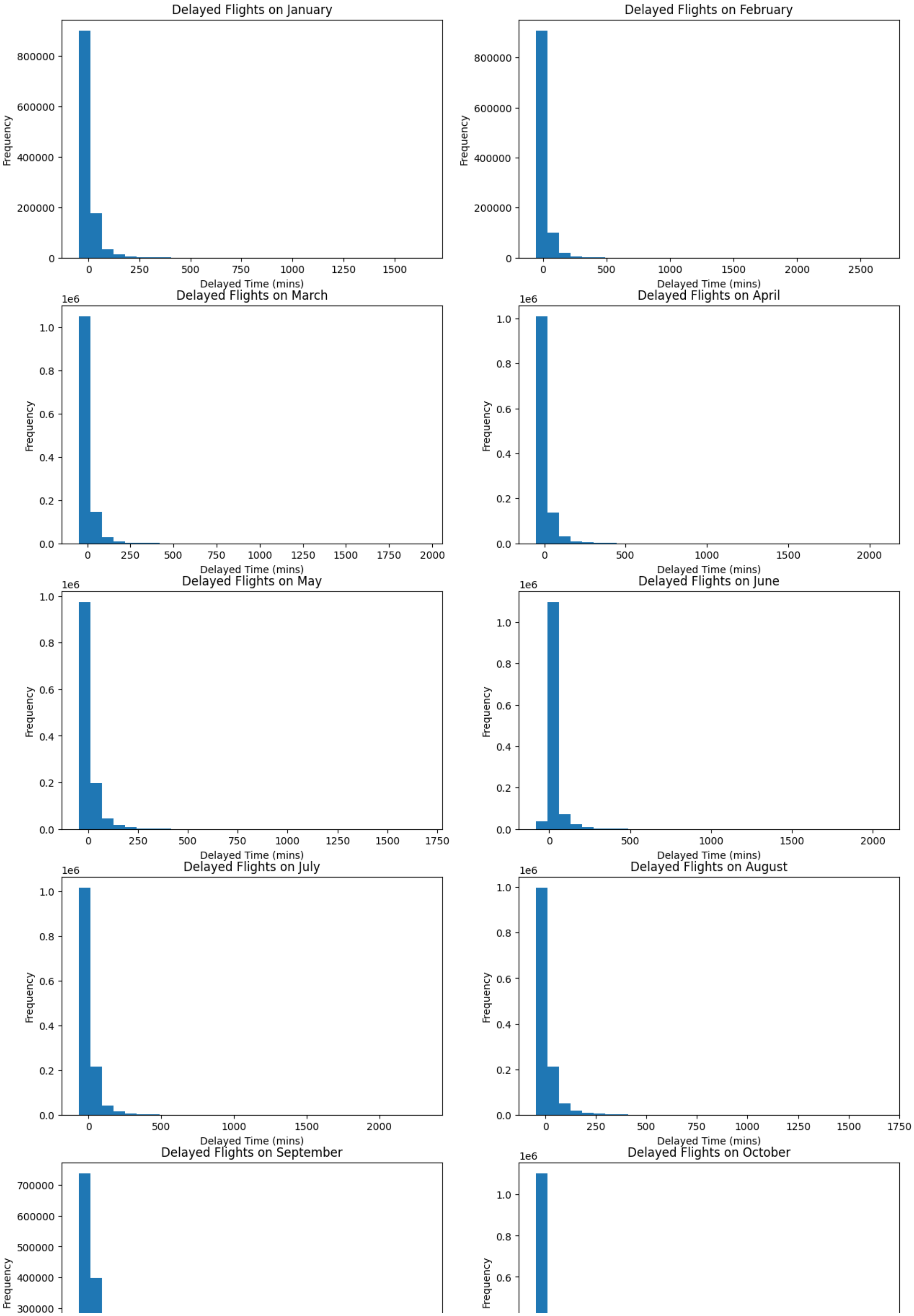
	ORIGIN	ORIGIN_CITY_NAME	Delay_Count	Delay_Ratio
1	ADK	Adak Island, AK	76	0.3877551020408163
2	HYA	Hyannis, MA	62	0.37349397590361444
3	OTH	North Bend/Coos Bay, OR	252	0.3490304709141274
4	OGD	Ogden, UT	70	0.33653846153846156
5	HGR	Hagerstown, MD	108	0.2967032967032967
6	MMH	Mammoth Lakes, CA	264	0.2926829268292683

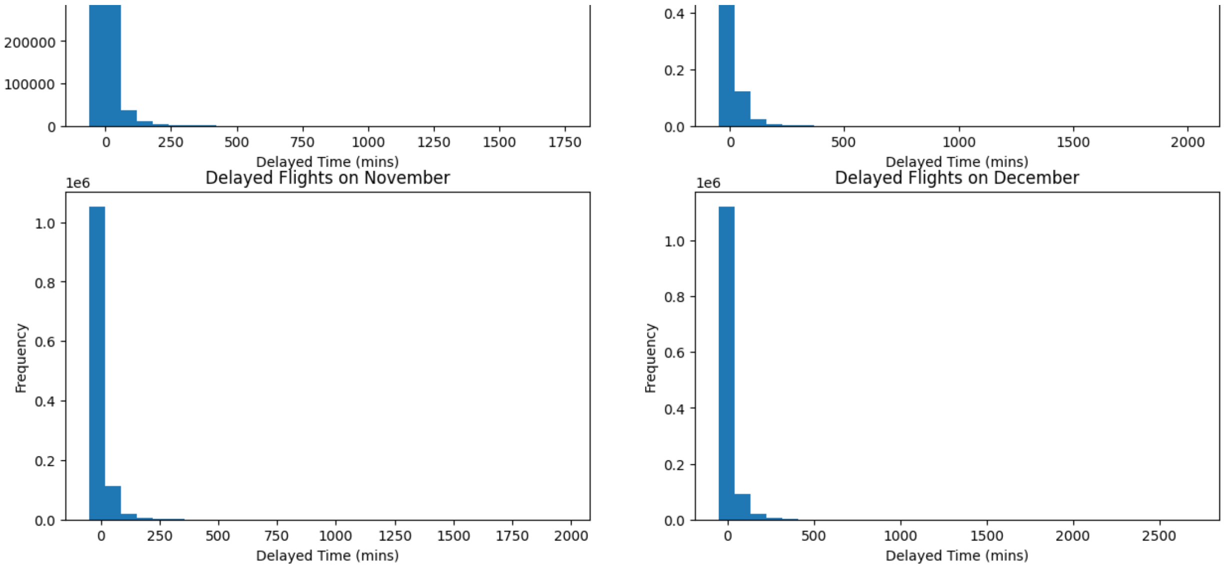
360 rows




Flight Delay by Month
Delayed Time (mins)
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
Month
2
4
6
8
10
12

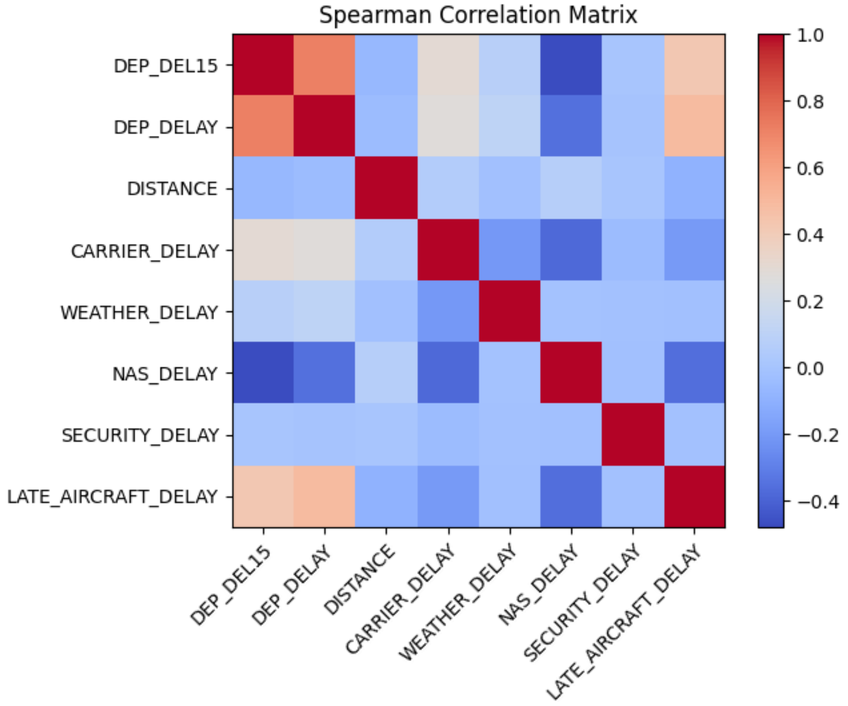

Delayed Flights
Delayed Time (mins)
0
10
20
30
40
2
4
6
8
10
12
Month


Delayed Flights on January
Frequency
0
200000
400000
600000
800000
0
250
500
750
1000
1250
1500
Delayed Time (mins)
Delayed Flights on February
Frequency
0
200000
400000
600000
800000
0
500
1000
1500
2000
2500
Delayed Time (mins)
Delayed Flights on March
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
1750
2000
Delayed Time (mins)
Delayed Flights on April
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Delayed Time (mins)
Delayed Flights on May
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
1750
Delayed Time (mins)
Delayed Flights on June
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Delayed Time (mins)
Delayed Flights on July
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Delayed Time (mins)
Delayed Flights on August
1e6
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
1750
Delayed Time (mins)
Delayed Flights on September
Frequency
300000
400000
500000
600000
700000
Delayed Flights on October
1e6
Frequency
0.6
0.8
1.0



Correlation Matrix and heatmap

```
[[ 1.          0.7145913  -0.06488022  0.29774828  0.07539183 -0.47962337
   0.00929733  0.42255695]
 [ 0.7145913   1.         -0.04314132  0.272305    0.10728592 -0.35502221
  -0.00138753  0.4860969 ]
 [-0.06488022 -0.04314132  1.          0.05441163 -0.02182299  0.0657451
   0.00840496 -0.09772084]
 [ 0.29774828  0.272305    0.05441163  1.         -0.20758261 -0.37778887
  -0.04392561 -0.19802927]
 [ 0.07539183  0.10728592 -0.02182299 -0.20758261  1.         -0.0110232
  -0.01322771 -0.02035012]
 [-0.47962337 -0.35502221  0.0657451  -0.37778887 -0.0110232   1.
  -0.01948326 -0.35936137]
 [ 0.00929733 -0.00138753  0.00840496 -0.04392561 -0.01322771 -0.01948326
   1.         -0.01459867]
 [ 0.42255695  0.4860969  -0.09772084 -0.19802927 -0.02035012 -0.35936137
  -0.01459867  1.        ]]
```



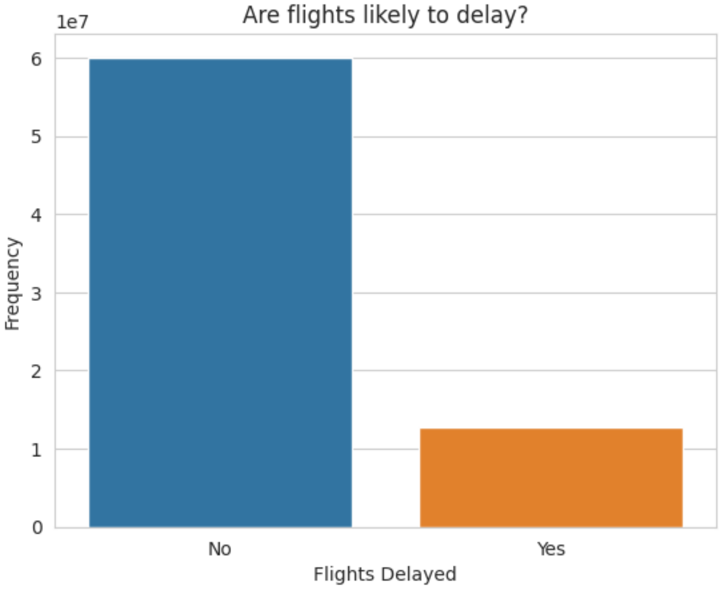
Full Data

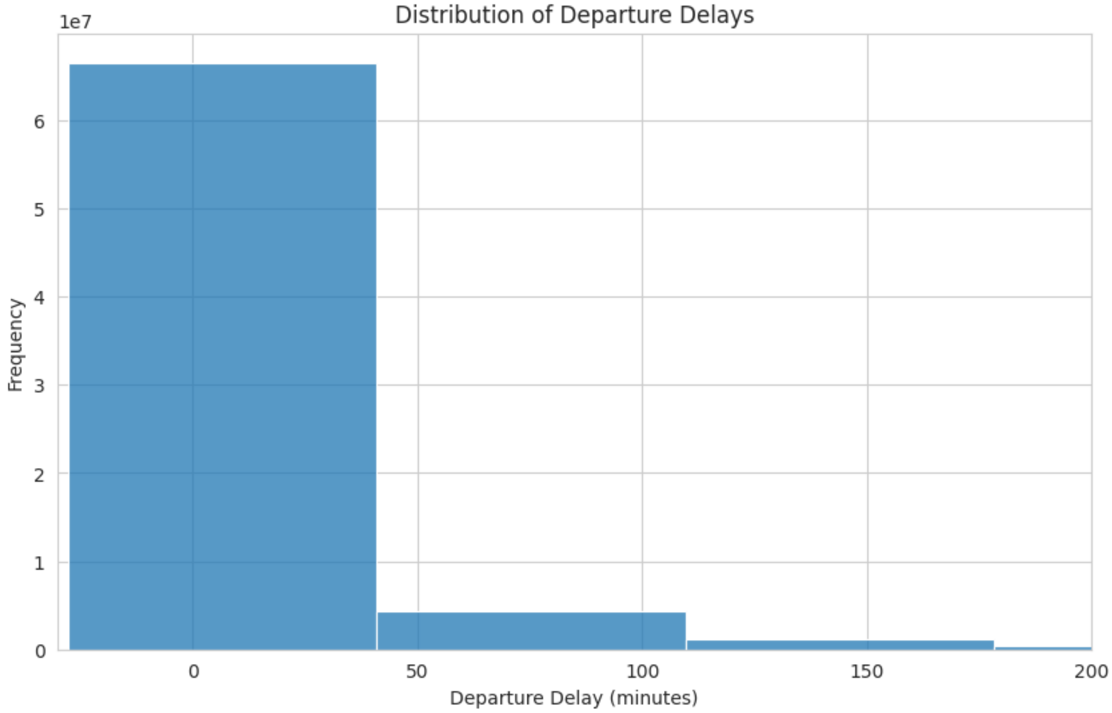
```
[Row(QUARTER=2, MONTH=6, DAY_OF_MONTH=4, DAY_OF_WEEK=2, FL_DATE='2019-06-04', OP_UNIQUE_CARRIER='YX', OP_CARRIER_AIRLINE_
ID=20452, OP_CARRIER='YX', TAIL_NUM='N206JQ', OP_CARRIER_FL_NUM=5927, ORIGIN_AIRPORT_ID=14100, ORIGIN_AIRPORT_SEQ_ID=1410
005, ORIGIN_CITY_MARKET_ID=34100, ORIGIN='PHL', ORIGIN_CITY_NAME='Philadelphia, PA', ORIGIN_STATE_ABR='PA', ORIGIN_STATE_
FIPS=42, ORIGIN_STATE_NM='Pennsylvania', ORIGIN_WAC=23, DEST_AIRPORT_ID=10721, DEST_AIRPORT_SEQ_ID=1072102, DEST_CITY_MAR
KET_ID=30721, DEST='BOS', DEST_CITY_NAME='Boston, MA', DEST_STATE_ABR='MA', DEST_STATE_FIPS=25, DEST_STATE_NM='Massachuse
tts', DEST_WAC=13, CRS_DEP_TIME=1640, DEP_TIME=1627, DEP_DELAY=-13.0, DEP_DELAY_NEW=0.0, DEP_DEL15=0.0, DEP_DELAY_GROUP=-
1, DEP_TIME_BLK='1600-1659', TAXI_OUT=36.0, WHEELS_OFF=1703, WHEELS_ON=1803, TAXI_IN=6.0, CRS_ARR_TIME=1814, ARR_TIME=180
9, ARR_DELAY=-5.0, ARR_DELAY_NEW=0.0, ARR_DEL15=0.0, ARR_DELAY_GROUP=-1, ARR_TIME_BLK='1800-1859', CANCELLED=0.0, CANCELL
ATION_CODE=None, DIVERTED=0.0, CRS_ELAPSED_TIME=94.0, ACTUAL_ELAPSED_TIME=102.0, AIR_TIME=60.0, FLIGHTS=1.0, DISTANCE=28
0.0, DISTANCE_GROUP=2, CARRIER_DELAY=None, WEATHER_DELAY=None, NAS_DELAY=None, SECURITY_DELAY=None, LATE_AIRCRAFT_DELAY=N
one, FIRST_DEP_TIME=None, TOTAL_ADD_GTIME=None, LONGEST_ADD_GTIME=None, DIV_AIRPORT_LANDINGS=0, DIV_REACHED_DEST=None, DI
V_ACTUAL_ELAPSED_TIME=None, DIV_ARR_DELAY=None, DIV_DISTANCE=None, DIV1_AIRPORT=None, DIV1_AIRPORT_ID=None, DIV1_AIRPORT_
SEQ_ID=None, DIV1_WHEELS_ON=None, DIV1_TOTAL_GTIME=None, DIV1_LONGEST_GTIME=None, DIV1_WHEELS_OFF=None, DIV1_TAIL_NUM=Non
e, DIV2_AIRPORT=None, DIV2_AIRPORT_ID=None, DIV2_AIRPORT_SEQ_ID=None, DIV2_WHEELS_ON=None, DIV2_TOTAL_GTIME=None, DIV2_LO
NGEST_GTIME=None, DIV2_WHEELS_OFF=None, DIV2_TAIL_NUM=None, DIV3_AIRPORT=None, DIV3_AIRPORT_ID=None, DIV3_AIRPORT_SEQ_ID=
None, DIV3_WHEELS_ON=None, DIV3_TOTAL_GTIME=None, DIV3_LONGEST_GTIME=None, DIV3_WHEELS_OFF=None, DIV3_TAIL_NUM=None, DIV4
_AIRPORT=None, DIV4_AIRPORT_ID=None, DIV4_AIRPORT_SEQ_ID=None, DIV4_WHEELS_ON=None, DIV4_TOTAL_GTIME=None, DIV4_LONGEST_G
TIME=None, DIV4_WHEELS_OFF=None, DIV4_TAIL_NUM=None, DIV5_AIRPORT=None, DIV5_AIRPORT_ID=None, DIV5_AIRPORT_SEQ_ID=None, D
IV5_WHEELS_ON=None, DIV5_TOTAL_GTIME=None, DIV5_LONGEST_GTIME=None, DIV5_WHEELS_OFF=None, DIV5_TAIL_NUM=None, YEAR=2019)]
```

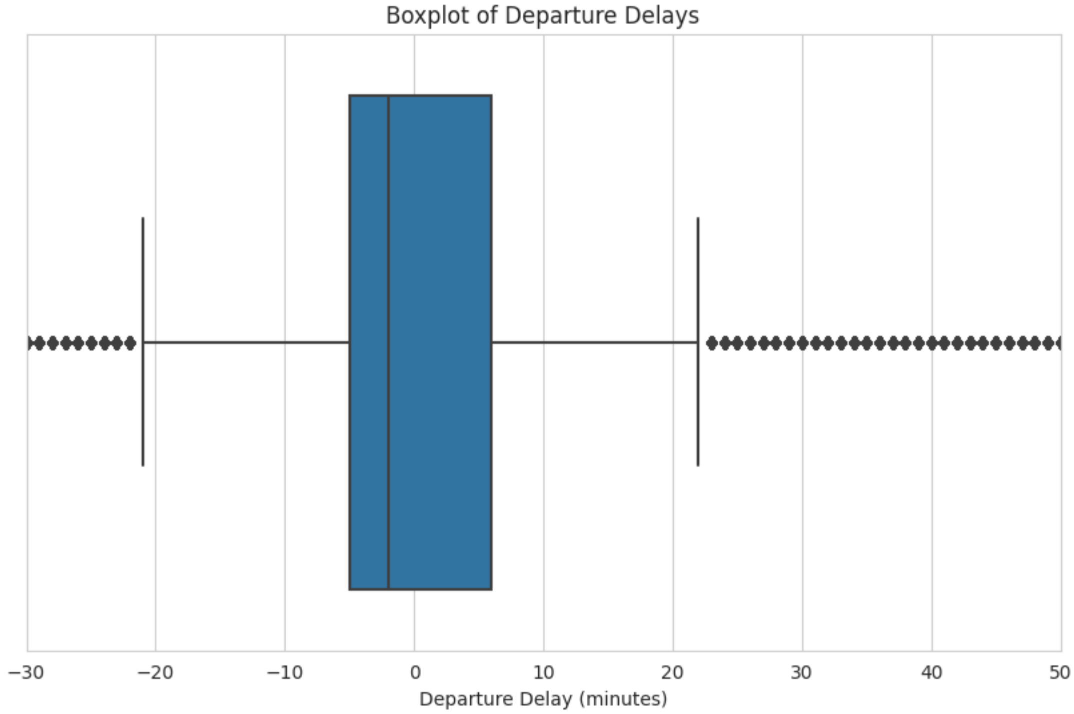
```
QUARTER , IntegerType()
MONTH , IntegerType()
DAY_OF_MONTH , IntegerType()
DAY_OF_WEEK , IntegerType()
FL_DATE , StringType()
OP_UNIQUE_CARRIER , StringType()
OP_CARRIER_AIRLINE_ID , IntegerType()
OP_CARRIER , StringType()
TAIL_NUM , StringType()
OP_CARRIER_FL_NUM , IntegerType()
ORIGIN_AIRPORT_ID , IntegerType()
ORIGIN_AIRPORT_SEQ_ID , IntegerType()
ORIGIN_CITY_MARKET_ID , IntegerType()
ORIGIN , StringType()
ORIGIN_CITY_NAME , StringType()
ORIGIN_STATE_ABR , StringType()
ORIGIN_STATE_FIPS , IntegerType()
ORIGIN_STATE_NM , StringType()
ORIGIN_WAC , IntegerType()
DEST_AIRPORT_ID , IntegerType()
DEST_AIRPORT_SEQ_ID , IntegerType()
```

```
74177433
```

Delay Plots


Are flights likely to delay?
1e7
6
5
4
3
2
1
0
Frequency
No
Yes
Flights Delayed





Table

	OP_UNIQUE_CARRIER	Delay_Count	Delay_Ratio
1	B6	783600	0.24603165632429458

2	F9	301789	0.2327189473132243
3	VX	92484	0.21301235455073106
4	WN	3093934	0.20711619292233927
5	G4	123261	0.2065901834424155
6	NK	369164	0.19682479765109456

20 rows

Table

	ORIGIN	ORIGIN_CITY_NAME	Delay_Count	Delay_Ratio
1	YNG	Youngstown/Warren, OH	4	1
2	ENV	Wendover, UT	2	1
3	TKI	Tokeen, AK	2	1
4	BIH	Bishop, CA	20	0.5714285714285714
5	CDB	Cold Bay, AK	106	0.43621399176954734
6	ILG	Wilmington, DE	115	0.3709677419354839

389 rows

Table

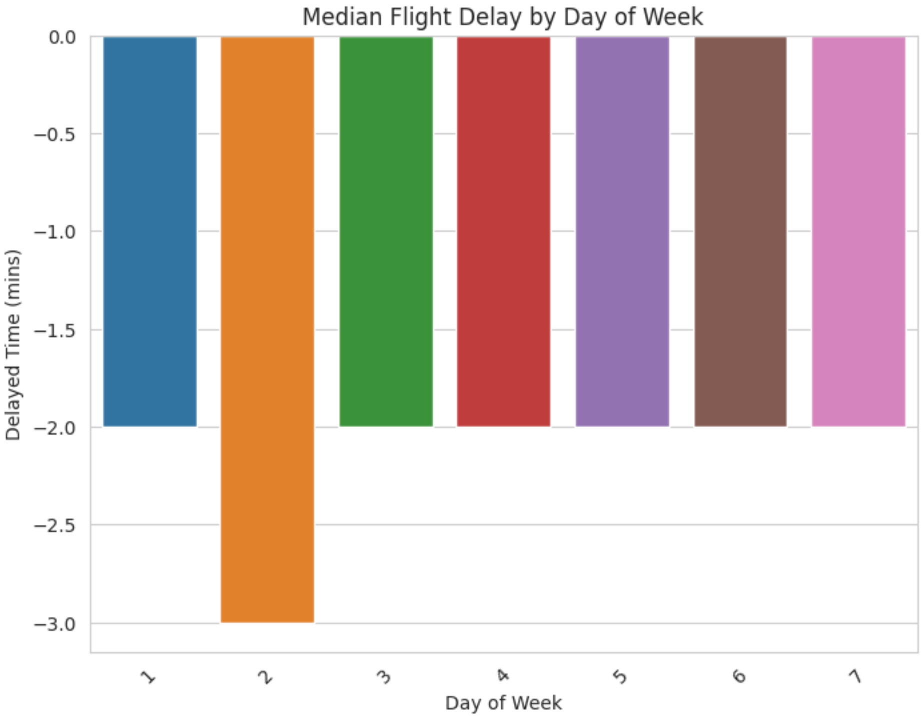
	DEST	DEST_CITY_NAME	Delay_Count	Delay_Ratio
1	BIH	Bishop, CA	19	0.5428571428571428
2	YNG	Youngstown/Warren, OH	2	0.5
3	PSE	Ponce, PR	2650	0.31958514230583696
4	BQN	Aguadilla, PR	6352	0.313679012345679
5	OTH	North Bend/Coos Bay, OR	1175	0.30567117585848075
6	CDB	Cold Bay, AK	69	0.2727272727272727

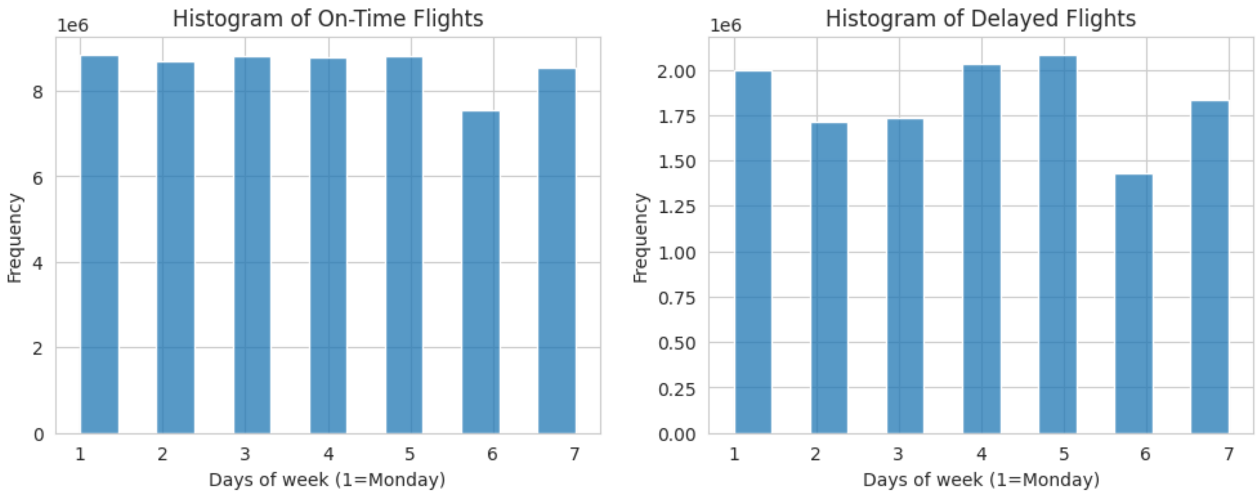
387 rows

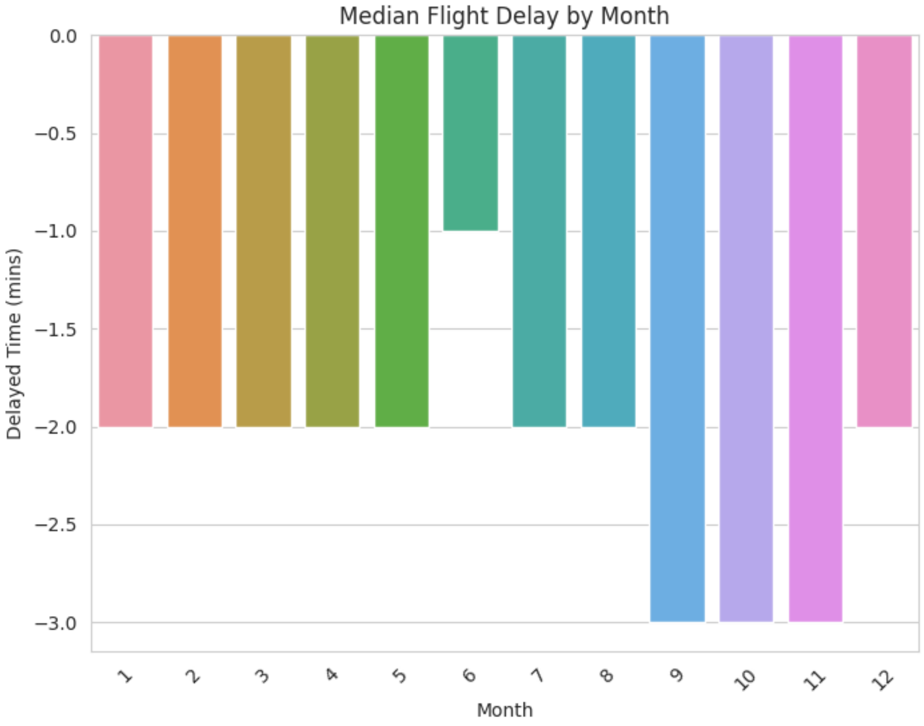
Table

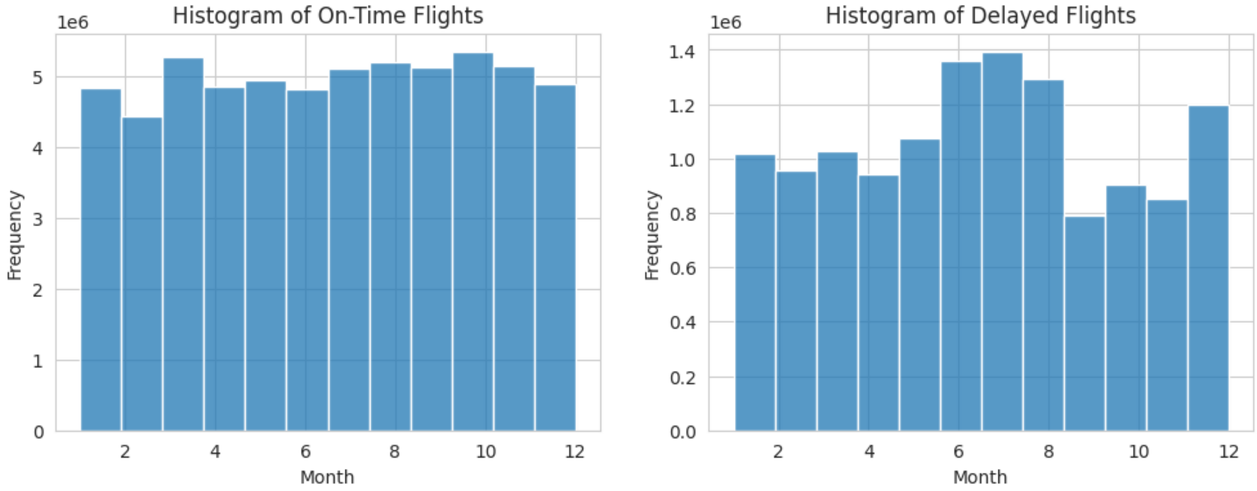
	YEAR	Delay_Count	Delay_Ratio
1	2015	2115108	0.18447019898739317
2	2019	2724328	0.1868043934065714
3	2016	1907086	0.17167956800994993
4	2018	2612870	0.18410315249882614
5	2017	2027690	0.1812289271287555
6	2020	401152	0.09099605190480566

7 rows


Median Flight Delay by Day of Week
Delayed Time (mins)
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1
2
3
4
5
6
7
Day of Week


Histogram of On-Time Flights
1e6
Frequency
0
2
4
6
8
1
2
3
4
5
6
7
Days of week (1=Monday)
Histogram of Delayed Flights
1e6
Frequency
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1
2
3
4
5
6
7
Days of week (1=Monday)

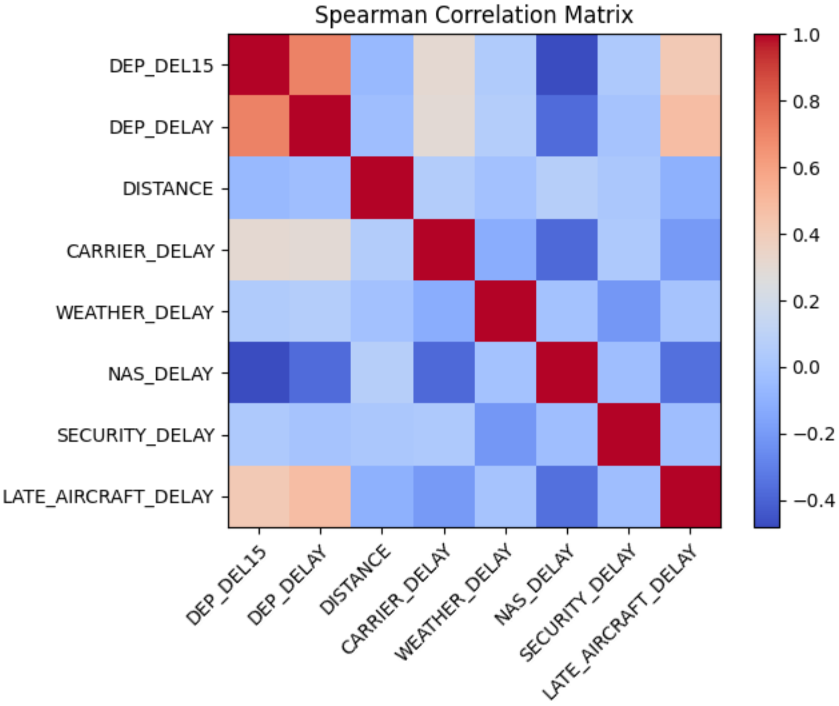

Median Flight Delay by Month
Delayed Time (mins)
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
1
2
3
4
5
6
7
8
9
10
11
12
Month

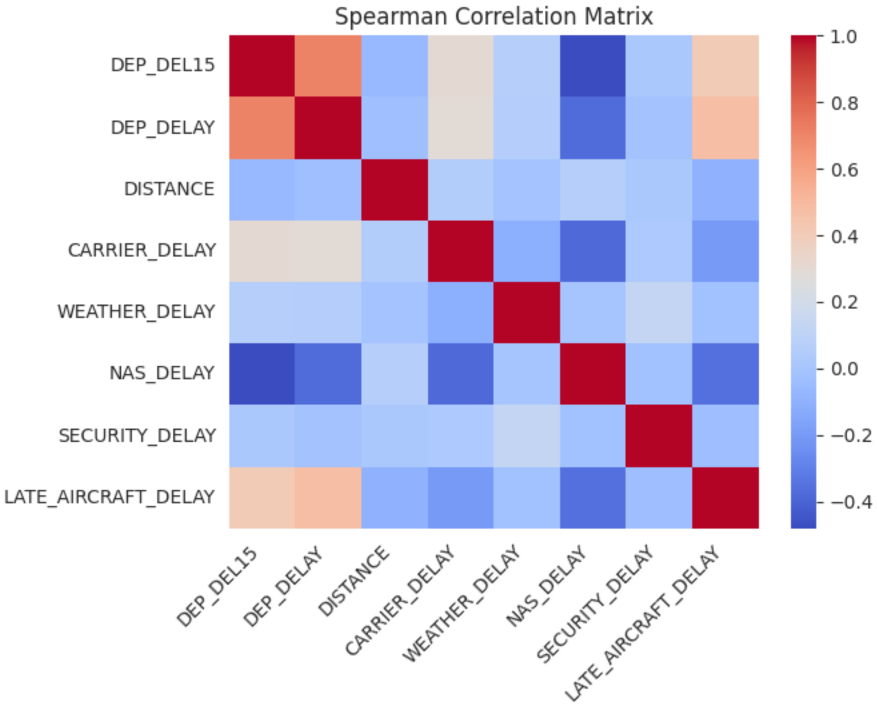


Correlation Matrix and heatmap

```
[[ 1.          0.70959726 -0.0634397   0.30245325  0.06639912 -0.48139925
   0.01973156  0.40895162]
 [ 0.70959726  1.         -0.03458421  0.28979412  0.05686575 -0.37404534
  -0.01625542  0.47392281]
 [-0.0634397  -0.03458421  1.          0.05405613 -0.00855174  0.06520499
   0.01829658 -0.10004021]
 [ 0.30245325  0.28979412  0.05405613  1.         -0.11573552 -0.38202021
   0.02233521 -0.20117023]
 [ 0.06639912  0.05686575 -0.00855174 -0.11573552  1.         -0.00545482
   0.12730953 -0.02316097]
 [-0.48139925 -0.37404534  0.06520499 -0.38202021 -0.00545482  1.
  -0.02327225 -0.35677416]
 [ 0.01973156 -0.01625542  0.01829658  0.02233521  0.12730953 -0.02327225
```

```
  1.         -0.03670589]
[ 0.40895162  0.47392281 -0.10004021 -0.20117023 -0.02316097 -0.35677416
 -0.03670589  1.        ]]
```





Checking for missing values

Table

	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	OP_UNIQUE_CARRIER	OP_CARRIER_AIRL
1	0	0	0	0	0	0	0

1 row

Table

	ORIGIN_AIRPORT_ID	ORIGIN_AIRPORT_SEQ_ID	ORIGIN_CITY_MARKET_ID	ORIGIN	ORIGIN_CITY_NAME	ORIG
1	0	0	0	0	0	0

1 row

Table

	DEST_AIRPORT_SEQ_ID	DEST_CITY_MARKET_ID	DEST	DEST_CITY_NAME	DEST_STATE_ABR	DEST_STAT
1	0	0	0	0	0	0

1 row

Table

	DEP_DELAY	DEP_DELAY_NEW	DEP_DEL15	DEP_DELAY_GROUP	DEP_TIME_BLK	TAXI_OUT	WHEELS_OFF
1	1335235	1335235	1335235	1335235	0	1355826	1355816

1 row

Table

	ARR_TIME	ARR_DELAY	ARR_DELAY_NEW	ARR_DEL15	ARR_DELAY_GROUP	ARR_TIME_BLK	CANCELLED
1	1390099	1547237	1547237	1547237	1547237	0	0

1 row

Table

	ACTUAL_ELAPSED_TIME	AIR_TIME	FLIGHTS	DISTANCE	DISTANCE_GROUP	CARRIER_DELAY	WEATHER_
1	1542041	1542041	0	0	0	61136954	61136954

1 row

Table

	FIRST_DEP_TIME	TOTAL_ADD_GTIME	LONGEST_ADD_GTIME	DIV_AIRPORT_LANDINGS	DIV_REACHED_DEST
1	73711726	73711743	73711744	95	73999356

1 row

Table

	DIV1_AIRPORT_SEQ_ID	DIV1_WHEELS_ON	DIV1_TOTAL_GTIME	DIV1_LONGEST_GTIME	DIV1_WHEELS_OFF	D
1	73991099	73991102	73991099	73991099	74024559	7

1 row

Table

	DIV2_TOTAL_GTIME	DIV2_LONGEST_GTIME	DIV2_WHEELS_OFF	DIV2_TAIL_NUM	DIV3_AIRPORT	DIV3_AIRPC
1	74175859	74175859	74176789	74176789	74177419	74177419

1 row

Table

	DIV3_WHEELS_OFF	DIV3_TAIL_NUM	DIV4_AIRPORT	DIV4_AIRPORT_ID	DIV4_AIRPORT_SEQ_ID	DIV4_WHEELS
1	74177431	74177431	74177433	74177433	74177433	74177433

1 row

Table

	DIV5_AIRPORT	DIV5_AIRPORT_ID	DIV5_AIRPORT_SEQ_ID	DIV5_WHEELS_ON	DIV5_TOTAL_GTIME	DIV5_LONG
1	74177433	74177433	74177433	74177433	74177433	74177433

1 row

```
+-----------+------------------+-------------------+
|Num missing|Perc missing      |Variable           |
+-----------+------------------+-------------------+
|61136954   |82.41988368618796 |CARRIER_DELAY      |
|61136954   |82.41988368618796 |WEATHER_DELAY      |
|61136954   |82.41988368618796 |NAS_DELAY          |
|61136954   |82.41988368618796 |SECURITY_DELAY     |
|61136954   |82.41988368618796 |LATE_AIRCRAFT_DELAY|
|1547237    |2.085859455395282 |ARR_DELAY          |
|1547237    |2.085859455395282 |ARR_DELAY_NEW      |
|1547237    |2.085859455395282 |ARR_DEL15          |
|1547237    |2.085859455395282 |ARR_DELAY_GROUP    |
|1542041    |2.0788546295475068|ACTUAL_ELAPSED_TIME|
|1542041    |2.0788546295475068|AIR_TIME           |
|1390121    |1.874048404991313 |WHEELS_ON          |
|1390121    |1.874048404991313 |TAXI_IN            |
|1390099    |1.8740187463753295|ARR_TIME           |
|1355826    |1.8278146670295263|TAXI_OUT           |
|1355816    |1.8278011858404428|WHEELS_OFF         |
|1335235    |1.8000555505877374|DEP_DELAY          |
|1335235    |1.8000555505877374|DEP_DELAY_NEW      |
```

Check carriers that disappeared over time

```
Carrier 9E has the following missing months:
2015 - 1
2015 - 2
2015 - 3
2015 - 4
2015 - 5
2015 - 6
2015 - 7
2015 - 8
2015 - 9
2015 - 10
2015 - 11
2015 - 12
2016 - 1
2016 - 2
2016 - 3
2016 - 4
2016 - 5
2016 - 6
2016 - 7
2016 - 8
```

Weather

3 Months sample

```
Row(STATION='52652099999', DATE='2015-01-01T02:00:00', LATITUDE='39.0833333', LONGITUDE='100.2833333', ELEVATION='1462.
0', NAME='ZHANGYE, CH', REPORT_TYPE='FM-12', SOURCE='4', HourlyAltimeterSetting=None, HourlyDewPointTemperature='-1', Hou
rlyDryBulbTemperature='3', HourlyPrecipitation=None, HourlyPresentWeatherType=None, HourlyPressureChange='+0.04', HourlyP
ressureTendency='7', HourlyRelativeHumidity='83', HourlySkyConditions=None, HourlySeaLevelPressure='30.72', HourlyStation
Pressure='25.45', HourlyVisibility=None, HourlyWetBulbTemperature='2', HourlyWindDirection='290', HourlyWindGustSpeed=Non
e, HourlyWindSpeed='4', Sunrise=None, Sunset=None, DailyAverageDewPointTemperature=None, DailyAverageDryBulbTemperature=N
one, DailyAverageRelativeHumidity=None, DailyAverageSeaLevelPressure=None, DailyAverageStationPressure=None, DailyAverage
WetBulbTemperature=None, DailyAverageWindSpeed=None, DailyCoolingDegreeDays=None, DailyDepartureFromNormalAverageTemperat
ure=None, DailyHeatingDegreeDays=None, DailyMaximumDryBulbTemperature=None, DailyMinimumDryBulbTemperature=None, DailyPea
kWindDirection=None, DailyPeakWindSpeed=None, DailyPrecipitation=None, DailySnowDepth=None, DailySnowfall=None, DailySust
ainedWindDirection=None, DailySustainedWindSpeed=None, DailyWeather=None, MonthlyAverageRH=None, MonthlyDaysWithGT001Prec
ip=None, MonthlyDaysWithGT010Precip=None, MonthlyDaysWithGT32Temp=None, MonthlyDaysWithGT90Temp=None, MonthlyDaysWithLT0T
emp=None, MonthlyDaysWithLT32Temp=None, MonthlyDepartureFromNormalAverageTemperature=None, MonthlyDepartureFromNormalCool
ingDegreeDays=None, MonthlyDepartureFromNormalHeatingDegreeDays=None, MonthlyDepartureFromNormalMaximumTemperature=None, 
MonthlyDepartureFromNormalMinimumTemperature=None, MonthlyDepartureFromNormalPrecipitation=None, MonthlyDewpointTemperatu
re=None, MonthlyGreatestPrecip=None, MonthlyGreatestPrecipDate=None, MonthlyGreatestSnowDepth=None, MonthlyGreatestSnowDe
pthDate=None, MonthlyGreatestSnowfall=None, MonthlyGreatestSnowfallDate=None, MonthlyMaxSeaLevelPressureValue=None, Month
lyMaxSeaLevelPressureValueDate=None, MonthlyMaxSeaLevelPressureValueTime=None, MonthlyMaximumTemperature=None, MonthlyMea
nTemperature=None, MonthlyMinSeaLevelPressureValue=None, MonthlyMinSeaLevelPressureValueDate=None, MonthlyMinSeaLevelPres
sureValueTime=None, MonthlyMinimumTemperature=None, MonthlySeaLevelPressure=None, MonthlyStationPressure=None, MonthlyTot
alLiquidPrecipitation=None, MonthlyTotalSnowfall=None, MonthlyWetBulb=None, AWND=None, CDSD=None, CLDD=None, DSNW=None, H
```

```
124
```

```
30528602
```

```
+-------+--------------------+-------------------+------------------+------------------+------------------+--------------
------+-----------+------------------+----------------------+-------------------------+------------------------+---------
-----------+------------------------+--------------------+----------------------+----------------------+-----------------
--+----------------------+---------------------+-----------------+------------------------+-------------------+----------
---------+----------------+-----------------+------------------+-------------------------------+-------------------------
-----+----------------------------+----------------------------+---------------------------+-----------------------------
-+---------------------+----------------------+------------------------------------------+----------------------+--------
----------------------+------------------------------+----------------------+------------------+-------------------+-----
-------------+-------------------+---------------------------+-----------------------+------------+----------------+-----
---------------------+--------------------------+-----------------------+-----------------------+----------------------+-
----------------------+--------------------------------------------+-------------------------------------------+---------
----------------------------------+--------------------------------------------+-----------------------------------------
---+---------------------------------------+--------------------------+---------------------+-------------------------+--
----------------------+----------------------------+-----------------------+---------------------------+-----------------
--------------+-----------------------------------+-----------------------------------+-------------------------+--------
--------------+-------------------------------+-----------------------------------+-----------------------------------+--
-----------------------+-----------------------+----------------------+-------------------------------+------------------
--+--------------+------------------+------------------+------------------+------------------+------------------+--------
----------+-----------------------+-----------------------+-----------------------+-----------------------+--------------
---------+-----------------------+-----------------------+----------------------+------------------------+---------------
--------+-----------------------+-----------------------+-----------------------+-----------------------+----------------
```

Column	Count	Mean	StdDev	Min	Max
STATION	2999268	5.996474120192107E10	3.275264218070831E10	01001099999	A5125600451
DATE	2999268	NULL	NULL	2015-01-01T00:00:00	2015-03-31T23:59:00
LATITUDE	2975744	37.85545689381726	21.314414505655474	-0.0166667	9.993861
LONGITUDE	2975744	-38.27010466933147	78.97939165678359	-0.005456	99.9666666
ELEVATION	2975744	356.02773612922283	530.6554767451264	-1.0	999.1
NAME	2975744	NULL	NULL	068 BAFFIN BAY PO...	ZYRYANKA, RS
REPORT_TYPE	2999268	NULL	NULL	CRN05	SY-MT
SOURCE	2999268	5.011725895419656	1.4000126213423996	1	O
HourlyAltimeterSetting	1616217	30.057367522798064	0.28781085986893973	27.79	31.09
HourlyDewPointTemperature	2472222	30.625930187209857	21.878629209732722	*	9s
HourlyDryBulbTemperature	2935018	39.48445492381933	23.03253595100165	*	9s
HourlyPrecipitation	386534	0.007989249217319316	0.04585078874871604	0.00	T
HourlyPresentWeatherType	385600	NULL	NULL	* * *	* * *
HourlyPressureChange	827955	0.001452065075291...	0.048390030114599414	+	1.18
HourlyPressureTendency	857356	4.854734789282398	2.7477097436639304	0	9

Column	Count	Mean	StdDev	Min	Max
HourlyRelativeHumidity	2471443	72.99516732807767	20.126855338899222	*	99
HourlySkyConditions	1577523	29.622613964013286	27.13229561567899	*	X:10s 0s
HourlySeaLevelPressure	1086638	30.027782892252237	0.3422377743347005	27.70	32.16
HourlyStationPressure	1523887	28.84685306566797	1.6814026837599156	15.69	31.78
HourlyVisibility	1960802	8.458200940556457	5.561604918489748	*	95.00
HourlyWetBulbTemperature	1498568	34.89997382387743	19.43629638159066	*	98
HourlyWindDirection	2571550	[illegible]	[illegible]	000	VRB

Table

	STATION	DATE	LATITUDE	LONGITUDE	ELEVATION	NAME	REPORT_TYPE	SOURCE	Hou
1	0	0	23524	23524	23524	23524	0	0	138

1 row

```
['STATION', 'DATE', 'LATITUDE', 'LONGITUDE', 'ELEVATION', 'NAME', 'REPORT_TYPE', 'SOURCE', 'HourlyAltimeterSetting', 'Hou
rlyDewPointTemperature', 'HourlyDryBulbTemperature', 'HourlyPrecipitation', 'HourlyPresentWeatherType', 'HourlyPressureCh
ange', 'HourlyPressureTendency', 'HourlyRelativeHumidity', 'HourlySkyConditions', 'HourlySeaLevelPressure', 'HourlyStatio
nPressure', 'HourlyVisibility', 'HourlyWetBulbTemperature', 'HourlyWindDirection', 'HourlyWindGustSpeed', 'HourlyWindSpee
d', 'Sunrise', 'Sunset', 'DailyAverageDewPointTemperature', 'DailyAverageDryBulbTemperature', 'DailyAverageRelativeHumidi
ty', 'DailyAverageSeaLevelPressure', 'DailyAverageStationPressure', 'DailyAverageWetBulbTemperature', 'DailyAverageWindSp
eed', 'DailyCoolingDegreeDays', 'DailyDepartureFromNormalAverageTemperature', 'DailyHeatingDegreeDays', 'DailyMaximumDryB
ulbTemperature', 'DailyMinimumDryBulbTemperature', 'DailyPeakWindDirection', 'DailyPeakWindSpeed', 'DailyPrecipitation',
 'DailySnowDepth', 'DailySnowfall', 'DailySustainedWindDirection', 'DailySustainedWindSpeed', 'DailyWeather', 'MonthlyAver
ageRH', 'MonthlyDaysWithGT001Precip', 'MonthlyDaysWithGT010Precip', 'MonthlyDaysWithGT32Temp', 'MonthlyDaysWithGT90Temp',
 'MonthlyDaysWithLT0Temp', 'MonthlyDaysWithLT32Temp', 'MonthlyDepartureFromNormalAverageTemperature', 'MonthlyDepartureFro
mNormalCoolingDegreeDays', 'MonthlyDepartureFromNormalHeatingDegreeDays', 'MonthlyDepartureFromNormalMaximumTemperature',
 'MonthlyDepartureFromNormalMinimumTemperature', 'MonthlyDepartureFromNormalPrecipitation', 'MonthlyDewpointTemperature',
 'MonthlyGreatestPrecip', 'MonthlyGreatestPrecipDate', 'MonthlyGreatestSnowDepth', 'MonthlyGreatestSnowDepthDate', 'Monthl
yGreatestSnowfall', 'MonthlyGreatestSnowfallDate', 'MonthlyMaxSeaLevelPressureValue', 'MonthlyMaxSeaLevelPressureValueDat
e', 'MonthlyMaxSeaLevelPressureValueTime', 'MonthlyMaximumTemperature', 'MonthlyMeanTemperature', 'MonthlyMinSeaLevelPres
sureValue', 'MonthlyMinSeaLevelPressureValueDate', 'MonthlyMinSeaLevelPressureValueTime', 'MonthlyMinimumTemperature', 'M
onthlySeaLevelPressure', 'MonthlyStationPressure', 'MonthlyTotalLiquidPrecipitation', 'MonthlyTotalSnowfall', 'MonthlyWet
Bulb', 'AWND', 'CDSD', 'CLDD', 'DSNW', 'HDSD', 'HTDD', 'NormalsCoolingDegreeDay', 'NormalsHeatingDegreeDay', 'ShortDurati
onEndDate005', 'ShortDurationEndDate010', 'ShortDurationEndDate015', 'ShortDurationEndDate020', 'ShortDurationEndDate03
0', 'ShortDurationEndDate045', 'ShortDurationEndDate060', 'ShortDurationEndDate080', 'ShortDurationEndDate100', 'ShortDur
```

```
+--------------------+----------+-----------+
|            Variable|NumMissing|PercMissing|
+--------------------+----------+-----------+
| HourlyWindGustSpeed|   2786532|      92.86|
| HourlyPrecipitation|   2614781|      87.13|
|HourlyPressureTen...|   2143969|      71.44|
|HourlySeaLevelPre...|   1913792|      63.77|
| HourlySkyConditions|   1423176|      47.42|
|    HourlyVisibility|   1039363|      34.63|
|HourlyDewPointTem...|    527689|      17.58|
|     HourlyWindSpeed|    398261|      13.27|
|HourlyDryBulbTemp...|     63966|       2.13|
+--------------------+----------+-----------+
```

Variable	NumMissing	PercMissing
HourlyWindGustSpeed	2,786,532	92.86%
HourlyPrecipitation	2,614,781	87.13%
HourlyPressureTendency	2,143,969	71.44%
HourlySeaLevelPressure	1,913,792	63.77%
HourlySkyConditions	1,423,176	47.42%
HourlyVisibility	1,039,363	34.63%
HourlyDewPointTemperature	527,689	17.58%
HourlyWindSpeed	398,261	13.27%

```
Column: HourlyWindSpeed, Data Type: string
Column: HourlyVisibility, Data Type: string
Column: HourlySkyConditions, Data Type: string
Column: HourlyPrecipitation, Data Type: string
Column: HourlyDewPointTemperature, Data Type: string
Column: HourlySeaLevelPressure, Data Type: string
Column: HourlyDryBulbTemperature, Data Type: string
Column: HourlyPressureTendency, Data Type: string
Column: HourlyWindGustSpeed, Data Type: string
```

```
+---------------+----------------+-------------------+-------------------+-------------------------+---------------------
-+------------------------+----------------------+-------------------+
|HourlyWindSpeed|HourlyVisibility|HourlySkyConditions|HourlyPrecipitation|HourlyDewPointTemperature|HourlySeaLevelPressur
e|HourlyDryBulbTemperature|HourlyPressureTendency|HourlyWindGustSpeed|
+---------------+----------------+-------------------+-------------------+-------------------------+---------------------
-+------------------------+----------------------+-------------------+
|              4|            NULL|               NULL|               NULL|                        6|                 30.6
4|                      16|                     2|               NULL|
|              9|            NULL|               NULL|               NULL|                        4|                 30.1
9|                      18|                     7|               NULL|
|              4|           18.64|               NULL|               NULL|                        1|                 30.2
6|                      10|                     7|               NULL|
|              7|            9.32|               NULL|               NULL|                        0|                 30.7
2|                       6|                     2|               NULL|
|             11|           18.64|             FEW:02|               NULL|                        2|                 30.7
3|                      14|                     4|               NULL|
|              2|           18.64|               NULL|               NULL|                        0|                 30.7
1|                       8|                     2|               NULL|
|              2|            NULL|               NULL|               NULL|                       -3|                 30.3
2|                      16|                     2|               NULL|
|              2|            NULL|               NULL|               NULL|                        1|                 30.6
```

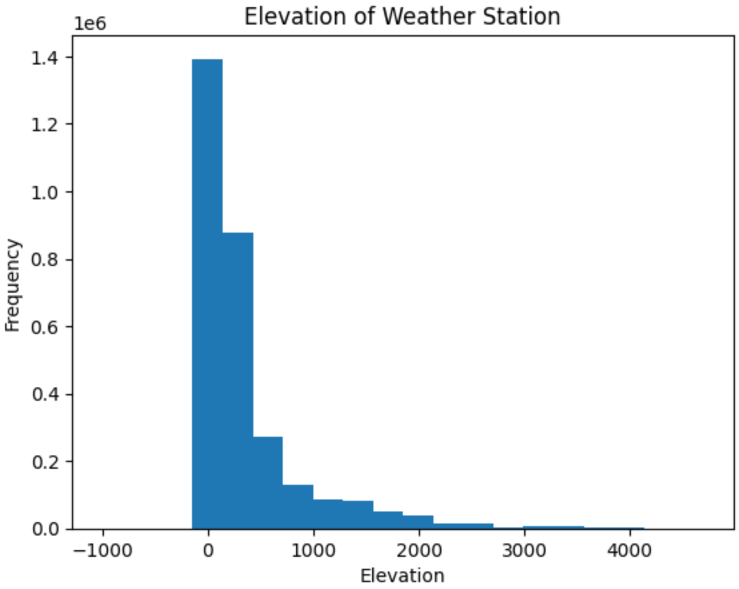
```
[('STATION', 'string'),
 ('DATE', 'string'),
 ('LATITUDE', 'string'),
 ('LONGITUDE', 'string'),
 ('ELEVATION', 'string'),
 ('NAME', 'string'),
 ('REPORT_TYPE', 'string'),
 ('SOURCE', 'string'),
 ('HourlyDryBulbTemperature', 'string'),
 ('HourlyWindDirection', 'string'),
 ('HourlyWindSpeed', 'string'),
 ('REM', 'string'),
 ('YEAR', 'int')]
```

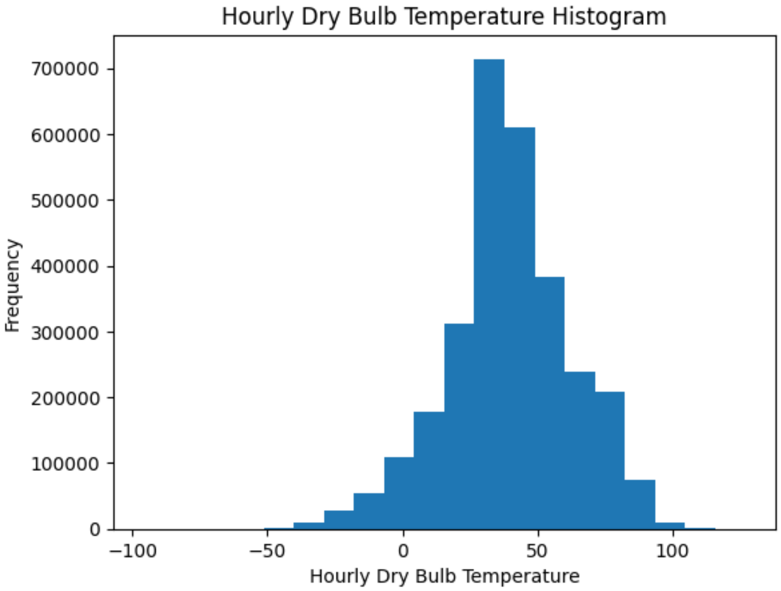
```
[('STATION', 'string'),
 ('DATE', 'string'),
 ('LATITUDE', 'string'),
 ('LONGITUDE', 'string'),
 ('ELEVATION', 'double'),
 ('NAME', 'string'),
 ('REPORT_TYPE', 'string'),
 ('SOURCE', 'string'),
 ('HourlyDryBulbTemperature', 'double'),
 ('HourlyWindDirection', 'double'),
 ('HourlyWindSpeed', 'double'),
 ('REM', 'string'),
 ('YEAR', 'int')]
```

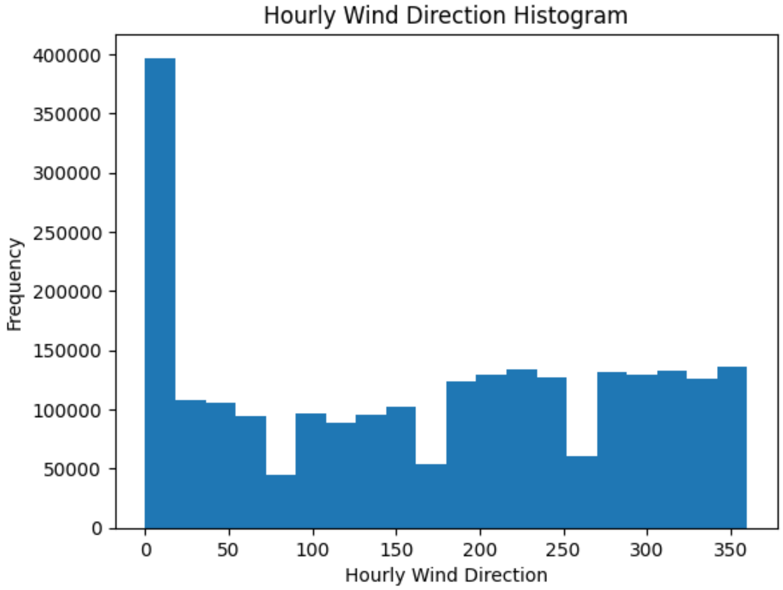
```
+-----------+-------+
|REPORT_TYPE|  count|
+-----------+-------+
|      FM-13|  86136|
|      FM-15|1503761|
|      CRN05| 380512|
|      FM-12| 771090|
|      FM-16| 125964|
|      SHEF |   7297|
|      SAO  | 107638|
|      SOM  |    317|
|      SOD  |  15064|
|      SAOSP|      2|
|      SURF |   1225|
|      SY-MT|     26|
|      FM-14|    236|
+-----------+-------+
```

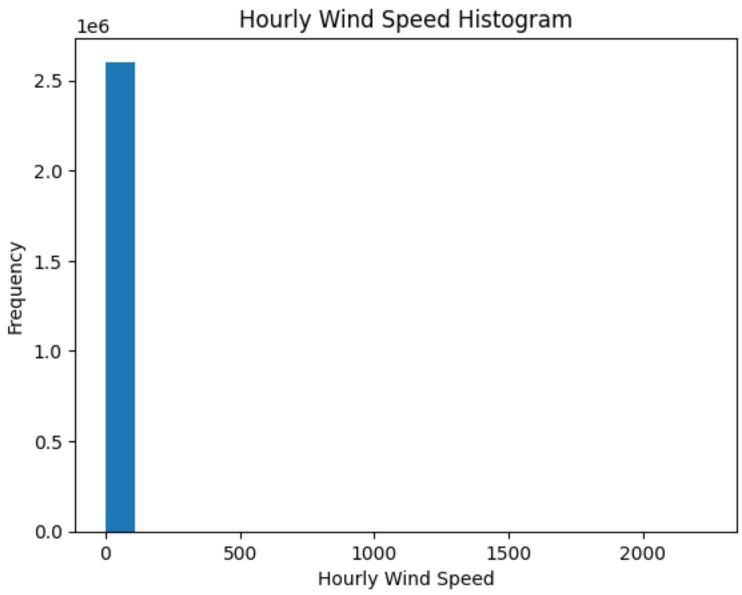
```
+------+-------+
|SOURCE|  count|
+------+-------+
|     7| 834722|
|     8|    392|
|     6|  68204|
|     O|   4604|
|     I| 381954|
|     4|1708155|
|     1|     11|
|     K|   1225|
|     2|      1|
+------+-------+
+----+-------+
|YEAR|  count|
+----+-------+
|2015|2999268|
+----+-------+
```

```
+--------------------+-----+
|                NAME|count|
+--------------------+-----+
|           TISKA, AG|  255|
|        PAVELETS, RS|   66|
|     BLAGODARNYJ, RS|   37|
|WINSTON SALEM REY...|  288|
|  VELYKYI BURLUK, UP|   68|
|CALDWELL ESSEX CO...|  265|
|        VESLJANA, RS|   66|
|BANGALURU INTERNA...|  462|
| RUDNAJA PRISTAN, RS|    1|
|           THYNA, TS|  265|
|        SHIZUOKA, JA|  210|
|         LANGRES, FR|  220|
|    VELIKIE LUKI, RS|   72|
|           LESKO, PL|  202|
|CLINTON MUNICIPAL...|  647|
|BOULDER MUNICIPAL...|  634|
|        AITUTAKI, CW|  192|
|          RAMADI, IZ|    9|
```


Elevation of Weather Station
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
−1000
0
1000
2000
3000
4000
Elevation


Hourly Dry Bulb Temperature Histogram
700000
600000
500000
400000
300000
200000
100000
0
Frequency
−100
−50
0
50
100
Hourly Dry Bulb Temperature


Hourly Wind Direction Histogram
Frequency
400000
350000
300000
250000
200000
150000
100000
50000
0
0
50
100
150
200
250
300
350
Hourly Wind Direction



1 Year Data

Stats and Tables

```
131937550
```

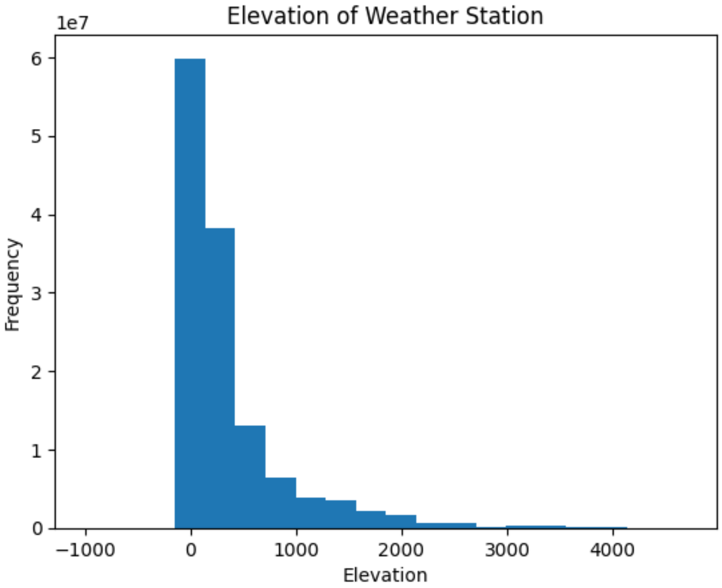
Table

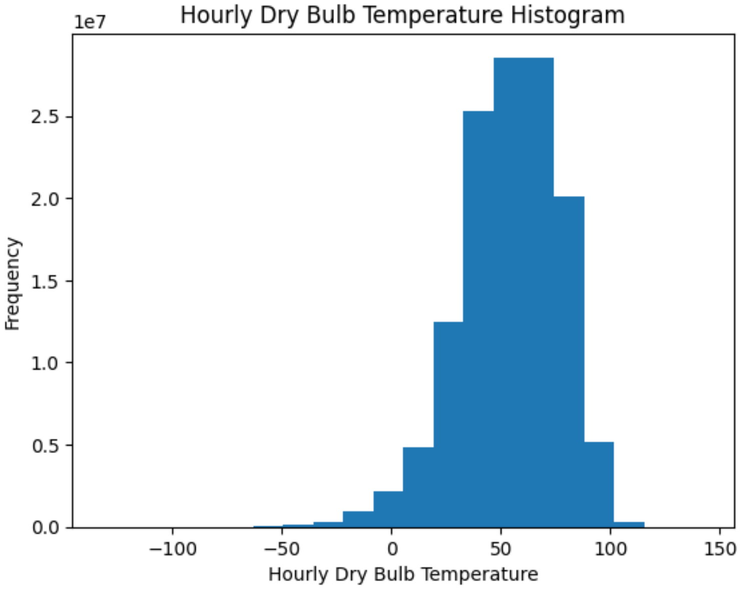
	STATION	DATE	LATITUDE	LONGITUDE	ELEVATION	NAME	REPORT_TYPE	SOURCE	Hou
1	0	0	829165	829165	829165	829165	0	0	645

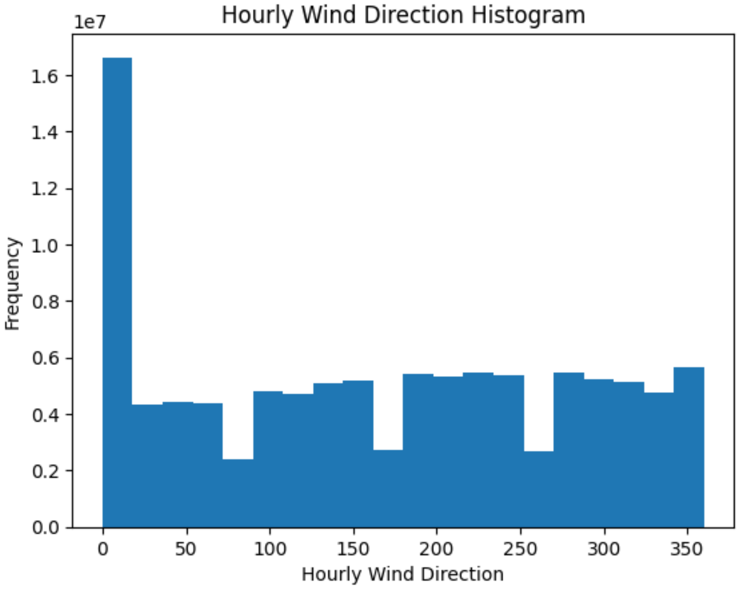
1 row

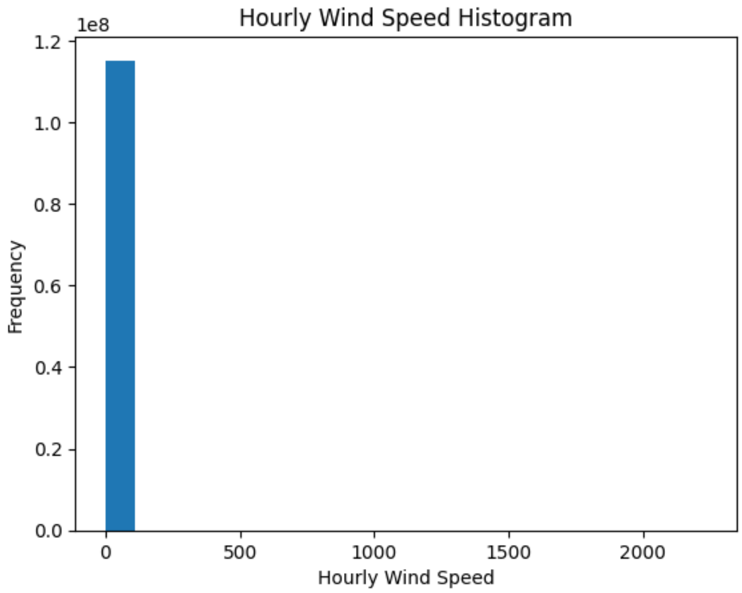
```
[('STATION', 'string'),
 ('DATE', 'string'),
 ('LATITUDE', 'string'),
 ('LONGITUDE', 'string'),
 ('ELEVATION', 'double'),
 ('NAME', 'string'),
 ('REPORT_TYPE', 'string'),
 ('SOURCE', 'string'),
 ('HourlyDryBulbTemperature', 'double'),
 ('HourlyWindDirection', 'double'),
 ('HourlyWindSpeed', 'double'),
 ('REM', 'string'),
 ('YEAR', 'int')]
```

Histograms




Hourly Dry Bulb Temperature Histogram
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Frequency
−100
−50
0
50
100
150
Hourly Dry Bulb Temperature


Hourly Wind Direction Histogram
1e7
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
50
100
150
200
250
300
350
Hourly Wind Direction



Full Data

Stats and Tables

```
898983399
```

```
null_counts = df_weather.select([count(when(isnan(c) | col(c).isNull(), c)).alia ...
```

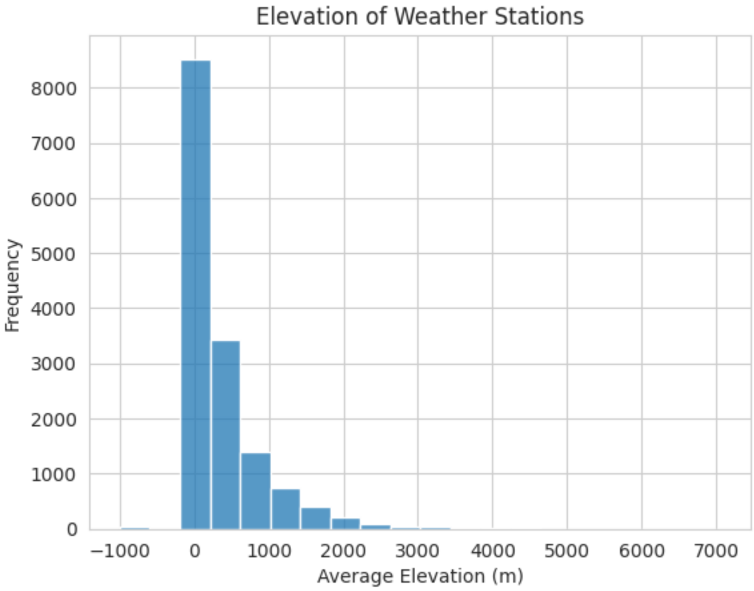
Show cell

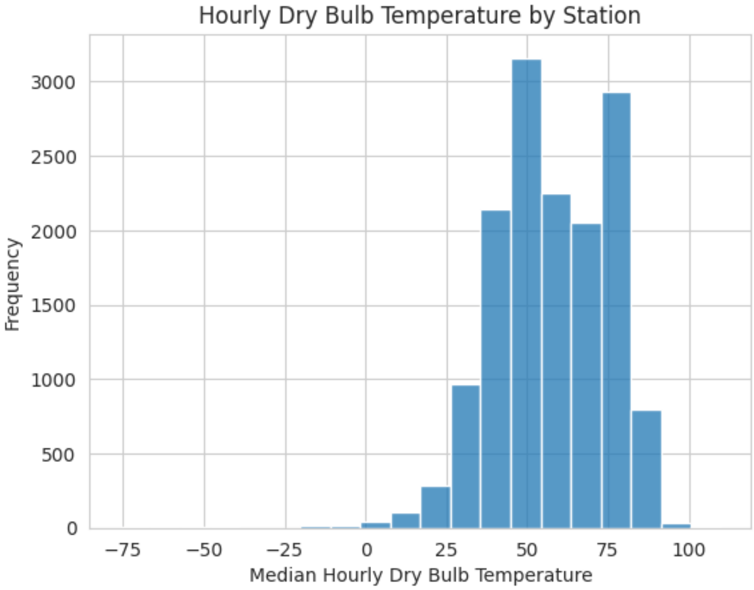
```
+--------------------+----------+-----------+
|            Variable|NumMissing|PercMissing|
+--------------------+----------+-----------+
| HourlyWindGustSpeed| 827361398|      92.03|
| HourlyPrecipitation| 784617131|      87.28|
|HourlyPressureTen...| 631593536|      70.26|
|HourlySeaLevelPre...| 546547716|       60.8|
| HourlySkyConditions| 461348098|      51.32|
|    HourlyVisibility| 315169298|      35.06|
|HourlyDewPointTem...| 153150719|      17.04|
|     HourlyWindSpeed| 116109419|      12.92|
|HourlyDryBulbTemp...|  19395099|       2.16|
+--------------------+----------+-----------+
+--------------------+---------+----------+
|            Variable|NumMissing|PercMissing|
+--------------------+---------+----------+
|                 REM| 228945826|     25.47|
| HourlyWindDirection| 125475322|     13.96|
|     HourlyWindSpeed| 116109419|     12.92|
|HourlyDryBulbTemp...|  19395099|      2.16|
|            LATITUDE|   6388837|      0.71|
|           LONGITUDE|   6388837|      0.71|
|                NAME|   6388837|      0.71|
|           ELEVATION|   6392201|      0.71|
|                DATE|         0|       0.0|
|             STATION|         0|       0.0|
|                YEAR|         0|       0.0|
|              SOURCE|         0|       0.0|
|         REPORT_TYPE|         0|       0.0|
+--------------------+---------+----------+
```

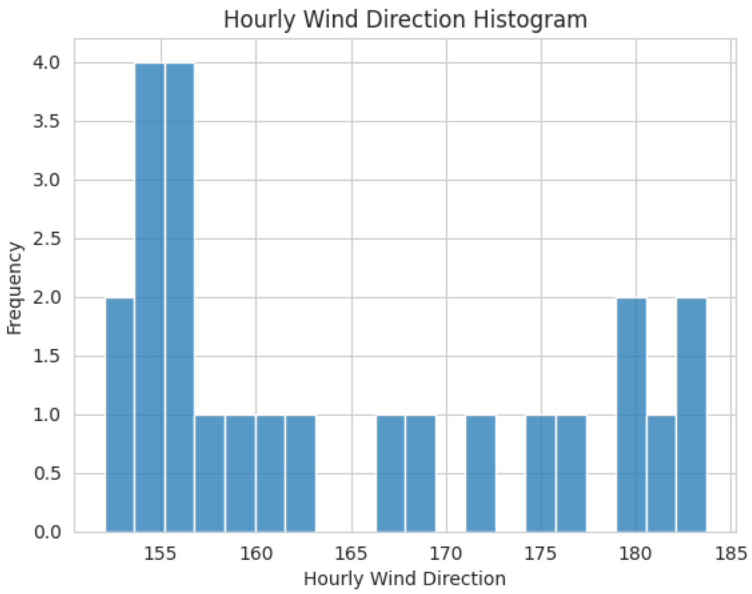
```
+-----------+---------+
|REPORT_TYPE|    count|
+-----------+---------+
|      FM-13| 20341676|
|      MESOW|    68035|
|      FM-15|434002331|
|      CRN05|112816978|
|      SY-MT|    16197|
|      SURF |   309431|
|      FM-12|246229213|
|      FM-16| 32530933|
|      SHEF |  1631142|
|      SAO  | 31249365|
|      SOM  |    90006|
|      SOD  |  4569553|
|      FM-14| 15124826|
|      SAOSP|     3713|
+-----------+---------+
+------+---------+
|SOURCE|    count|
+------+---------+
|     K|   309431|
|     7|224875430|
|     8|   180073|
|     6| 21148045|
|     O|  1442375|
|     1|     1855|
|     I|113221529|
|     4|537803624|
|     2|     1037|
+------+---------+
```

```
+----+---------+
|YEAR|    count|
+----+---------+
|2020|130219890|
|2016|125155076|
|2017|129697625|
|2018|129080303|
|2015|123856083|
|2021|129036872|
|2019|131937550|
+----+---------+
```

Histograms


Elevation of Weather Stations
Frequency
0
1000
2000
3000
4000
5000
6000
7000
8000
−1000
0
1000
2000
3000
4000
5000
6000
7000
Average Elevation (m)


Hourly Dry Bulb Temperature by Station
Frequency
0
500
1000
1500
2000
2500
3000
−75
−50
−25
0
25
50
75
100
Median Hourly Dry Bulb Temperature



Stations

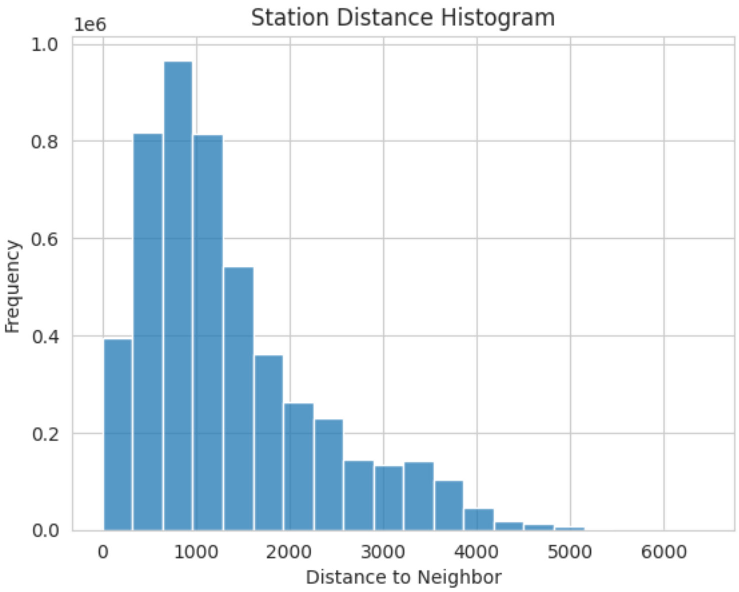
Table

	usaf	wban	station_id	lat	lon	neighbor_id	neighbor_name
1	690020	93218	69002093218	36	-121.233	69002093218	JOLON HUNTER LIGGETT MIL RES
2	690020	93218	69002093218	36	-121.233	69007093217	FRITZSCHE AAF
3	690020	93218	69002093218	36	-121.233	69014093101	EL TORO MCAS
4	690020	93218	69002093218	36	-121.233	70027127506	BARROW POINT BARROW
5	690020	93218	69002093218	36	-121.233	70045027512	LONELY
6	690020	93218	69002093218	36	-121.233	70063027403	OLIKTOK POW 2

10,000 rows | Truncated data

```
StructType([StructField('usaf', StringType(), True), StructField('wban', StringType(), True), StructField('station_id', S
tringType(), True), StructField('lat', DoubleType(), True), StructField('lon', DoubleType(), True), StructField('neighbor
_id', StringType(), True), StructField('neighbor_name', StringType(), True), StructField('neighbor_state', StringType(),
True), StructField('neighbor_call', StringType(), True), StructField('neighbor_lat', DoubleType(), True), StructField('ne
ighbor_lon', DoubleType(), True), StructField('distance_to_neighbor', DoubleType(), True)])
```

```
5004169
```



OTPW

Link to OTPW analysis: https://adb-4248444930383559.19.azuredatabricks.net/?o=4248444930383559#notebook/4034308520565772/command/4034308520565786 (https://adb-4248444930383559.19.azuredatabricks.net/?o=4248444930383559#notebook/4034308520565772/command/4034308520565786)

Link to Data Cleaning/Manipulation: https://adb-4248444930383559.19.azuredatabricks.net/?o=4248444930383559#notebook/2798664673112925/command/2798664673114658 (https://adb-4248444930383559.19.azuredatabricks.net/?o=4248444930383559#notebook/2798664673112925/command/2798664673114658)