SQL - API Project

Hurricane harvey impact on flights' cancellations

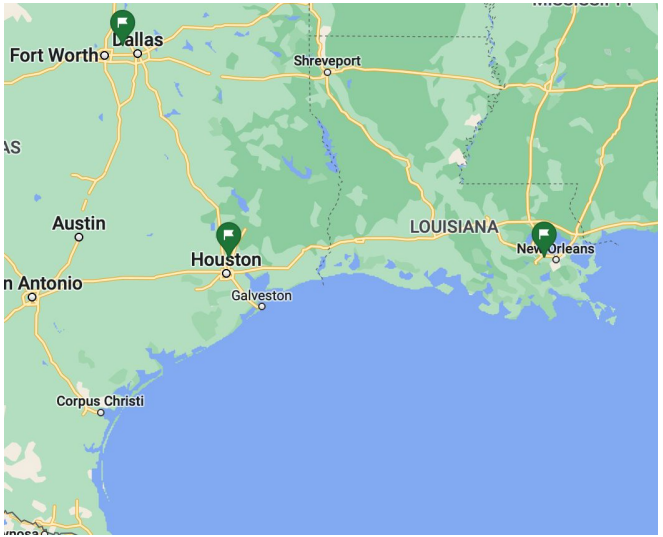
Julia, Trista and Daniel

Goal of our project

- Main goal: Did the Hurricane Harvey impact flights in August 2017?
- Getting the API data, working in jupyter notebook, dbeaver and combine with python

Defining our scope and problem

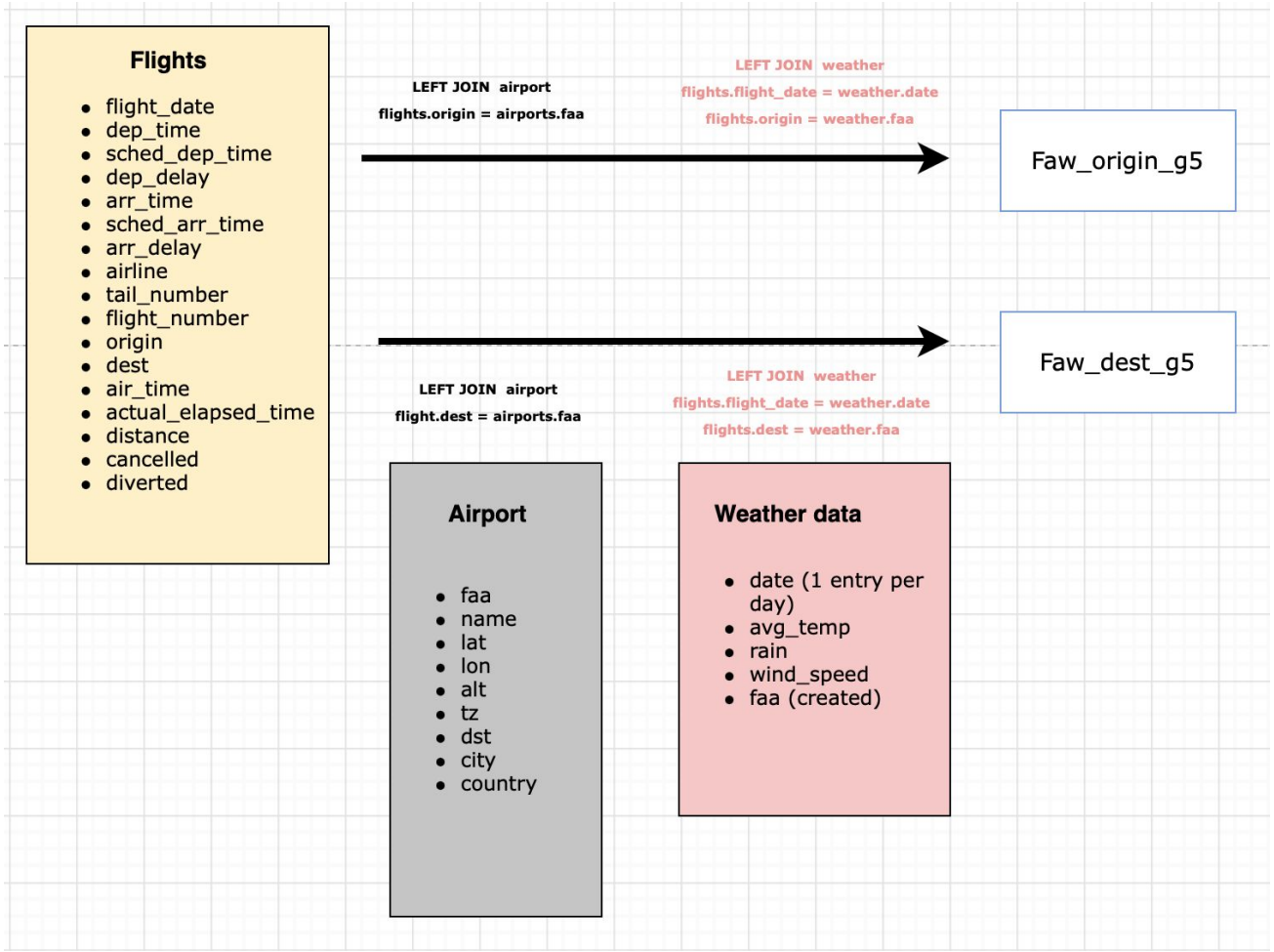
- Hurricane Harvey
  - Affected time: August 25 - 29, 2017
  - Category 4
  - One of the biggest hurricanes last 30 years
- Data
  - Weather: Daily weather report in August 2017
  - Flight: flights report in August 2017
  - Airports table
- Affected area & airports
  - Texas: IAH (Houston), DFW (Dallas)
  - Louisiana: MSY (New Orleans), ~~BTR (Baton Rouge)~~
  - New York: JFK (comparison)



Hypotheses

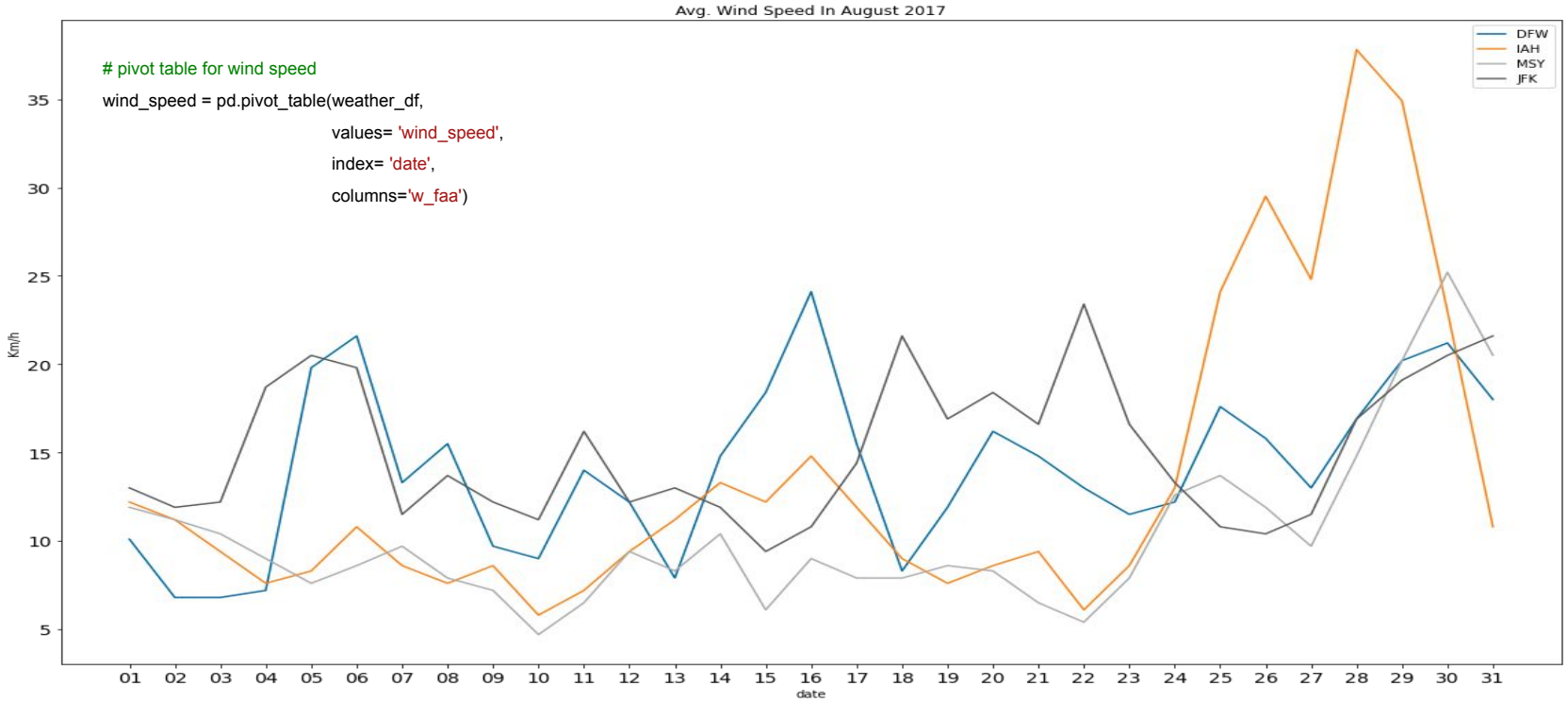
- The airports closer to the coast (IAH, MSY, BTR) will be more affected than the ones inland (DFW)
- Hurricane Harvey has caused flight diversions arriving in Texas and Louisiana
- JFK was not affected by Harvey like flights from/to Texas and Louisiana

Merging


Flights
flight_date
dep_time
sched_dep_time
dep_delay
arr_time
sched_arr_time
arr_delay
airline
tail_number
flight_number
origin
dest
air_time
actual_elapsed_time
distance
cancelled
diverted
LEFT JOIN airport
flights.origin = airports.faa
LEFT JOIN weather
flights.flight_date = weather.date
flights.origin = weather.faa
Faw_origin_g5
LEFT JOIN airport
flight.dest = airports.faa
LEFT JOIN weather
flights.flight_date = weather.date
flights.dest = weather.faa
Faw_dest_g5
Airport
faa
name
lat
lon
alt
tz
dst
city
country
Weather data
date (1 entry per day)
avg_temp
rain
wind_speed
faa (created)

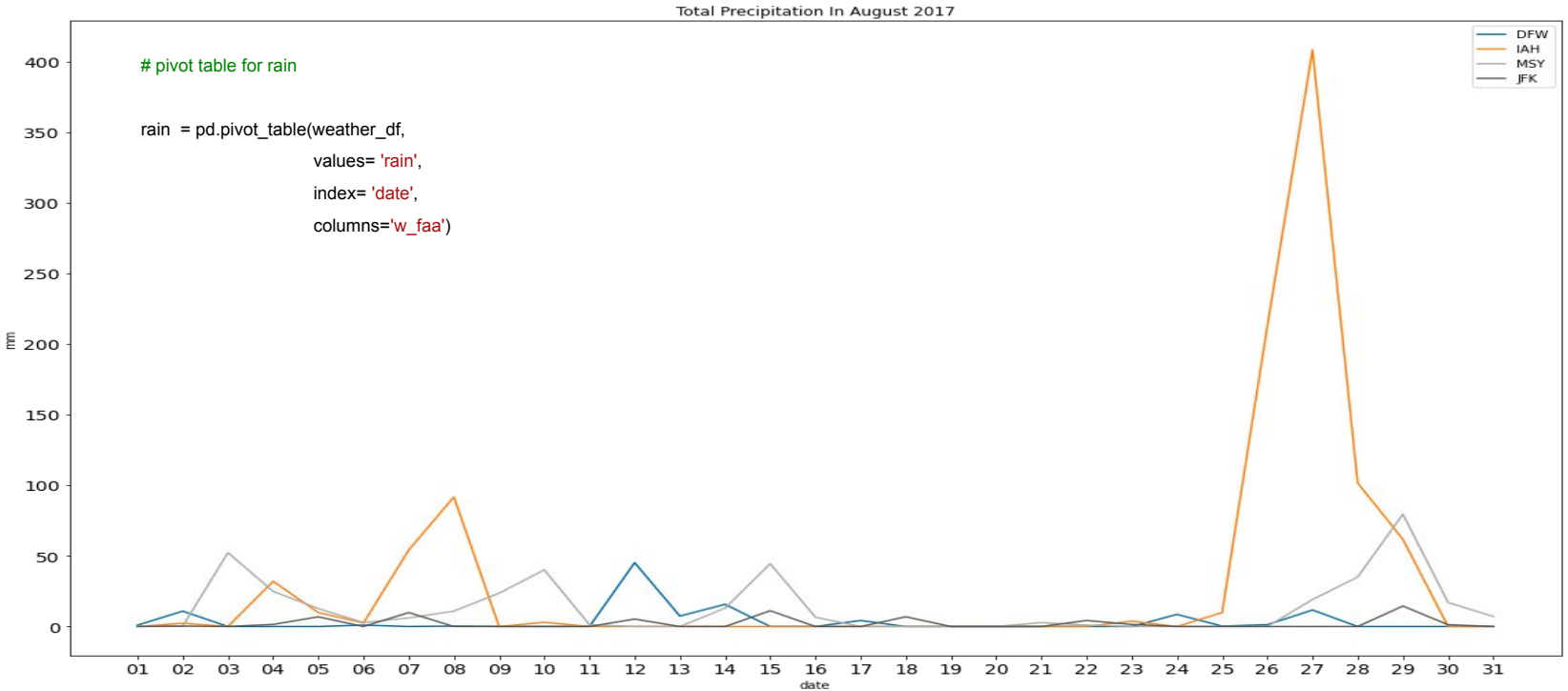
Wind speed significantly increased from 8/24 - 8/30

```
# pivot table for wind speed
wind_speed = pd.pivot_table(weather_df,
                            values= 'wind_speed',
                            index= 'date',
                            columns='w_faa')
```

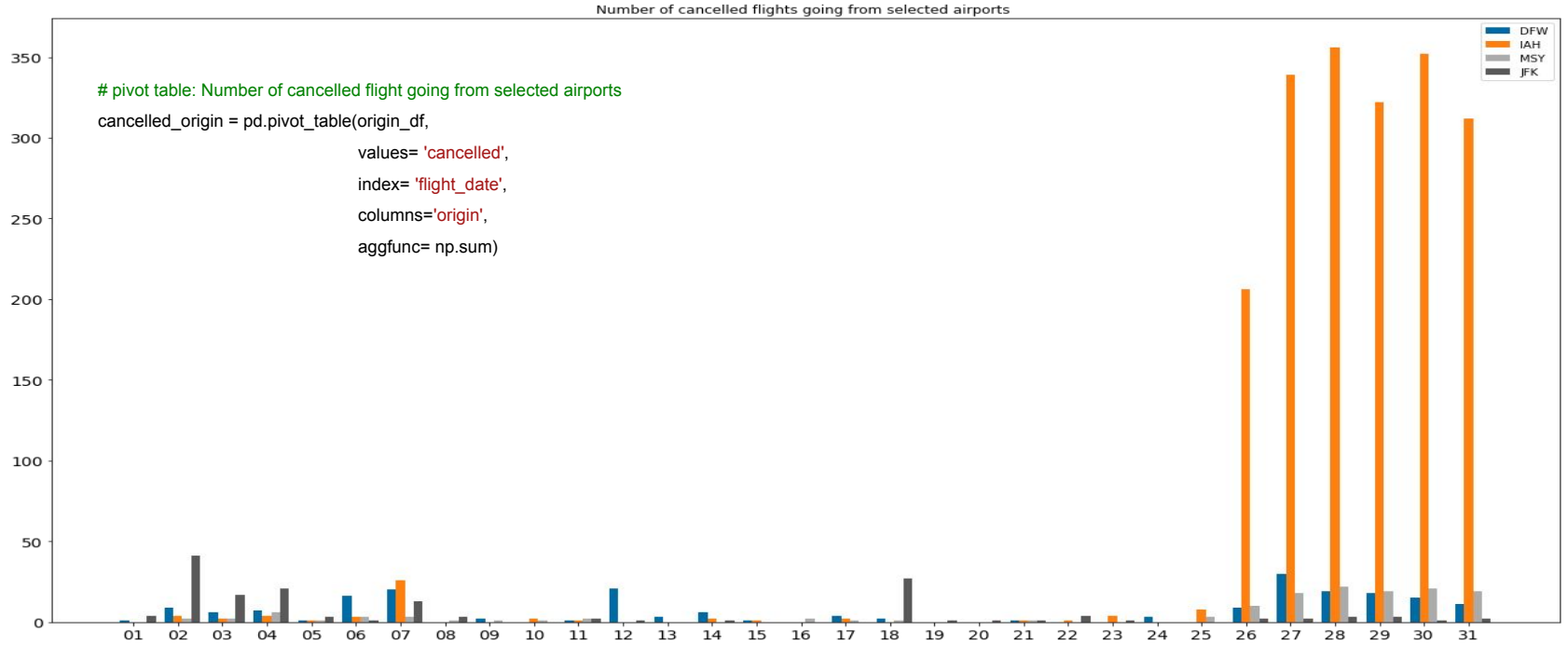


Total Rain significantly increased from 8/25 to 8/28 in Houston

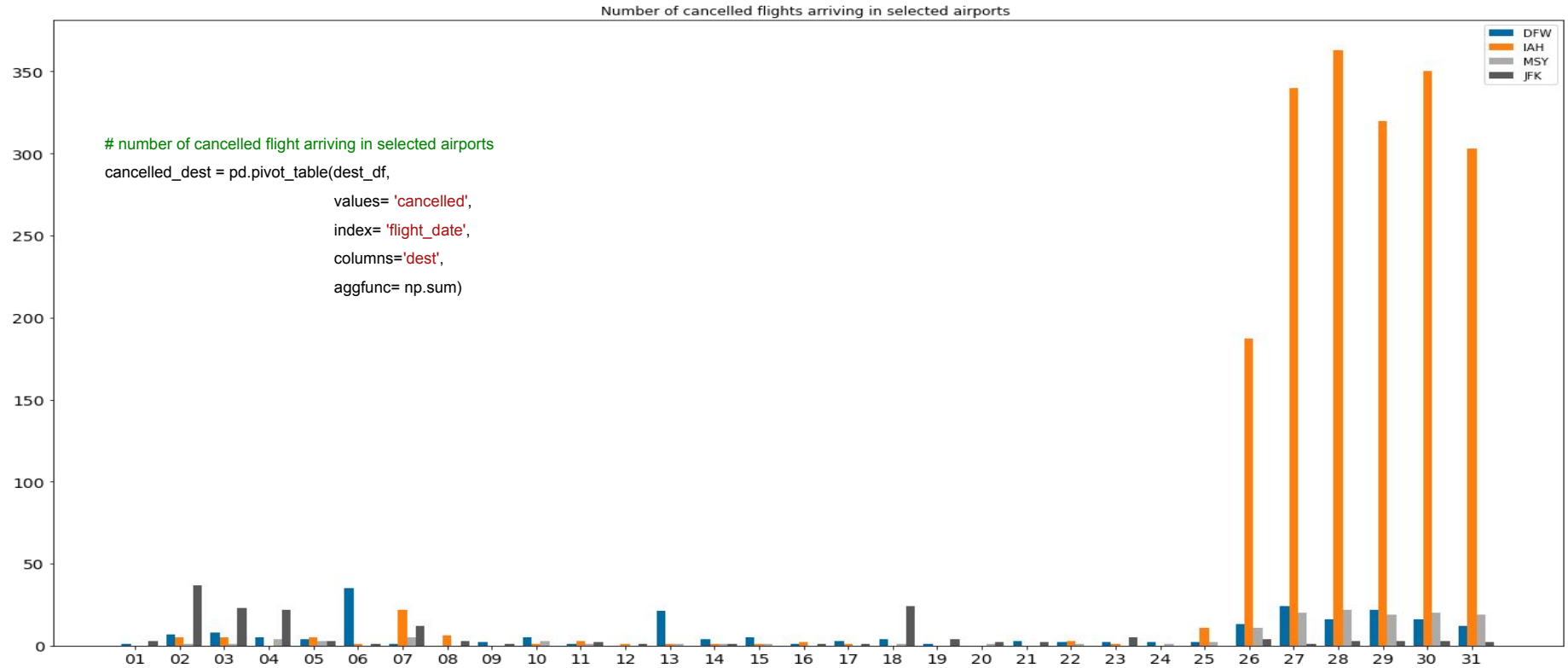
```
# pivot table for rain

rain  = pd.pivot_table(weather_df,
                       values= 'rain',
                       index= 'date',
                       columns='w_faa')
```



1: IAH (Houston) has significant increase of flights' cancellation than DFW, MSY, JFK



1: IAH (Houston) has the most cancelled flights arriving to its destination compared to DFW, MSY, JFK




```
# number of cancelled flight arriving in selected airports
cancelled_dest = pd.pivot_table(dest_df,
                                values= 'cancelled',
                                index= 'flight_date',
                                columns='dest',
                                aggfunc= np.sum)
```

```
# check correlation
# to merge three dataframe: cancelled_origin and wind_speed on date

cancelled_origin_wind = pd.merge(cancelled_origin, wind_speed, left_on='flight_date', right_on= 'date')
```

```
cancelled_origin_wind.corr(method= 'pearson')
```

✓ 0.4s

	DFW_x	IAH_x	JFK_x	MSY_x	DFW_y	IAH_y	JFK_y	MSY_y
DFW_x	1.000000	0.674916	0.079196	0.676911	0.150496	0.527803	-0.010818	0.454026
IAH_x	0.674916	1.000000	-0.139940	0.982049	0.369919	0.780681	0.223933	0.761779
JFK_x	0.079196	-0.139940	1.000000	-0.067268	-0.526227	-0.160555	0.069754	-0.021654
MSY_x	0.676911	0.982049	-0.067268	1.000000	0.370738	0.762629	0.267496	0.779962
DFW_y	0.150496	0.369919	-0.526227	0.370738	1.000000	0.410097	0.189999	0.372855
IAH_y	0.527803	0.780681	-0.160555	0.762629	0.410097	1.000000	-0.084260	0.624330
JFK_y	-0.010818	0.223933	0.069754	0.267496	0.189999	-0.084260	1.000000	0.239420
MSY_y	0.454026	0.761779	-0.021654	0.779962	0.372855	0.624330	0.239420	1.000000

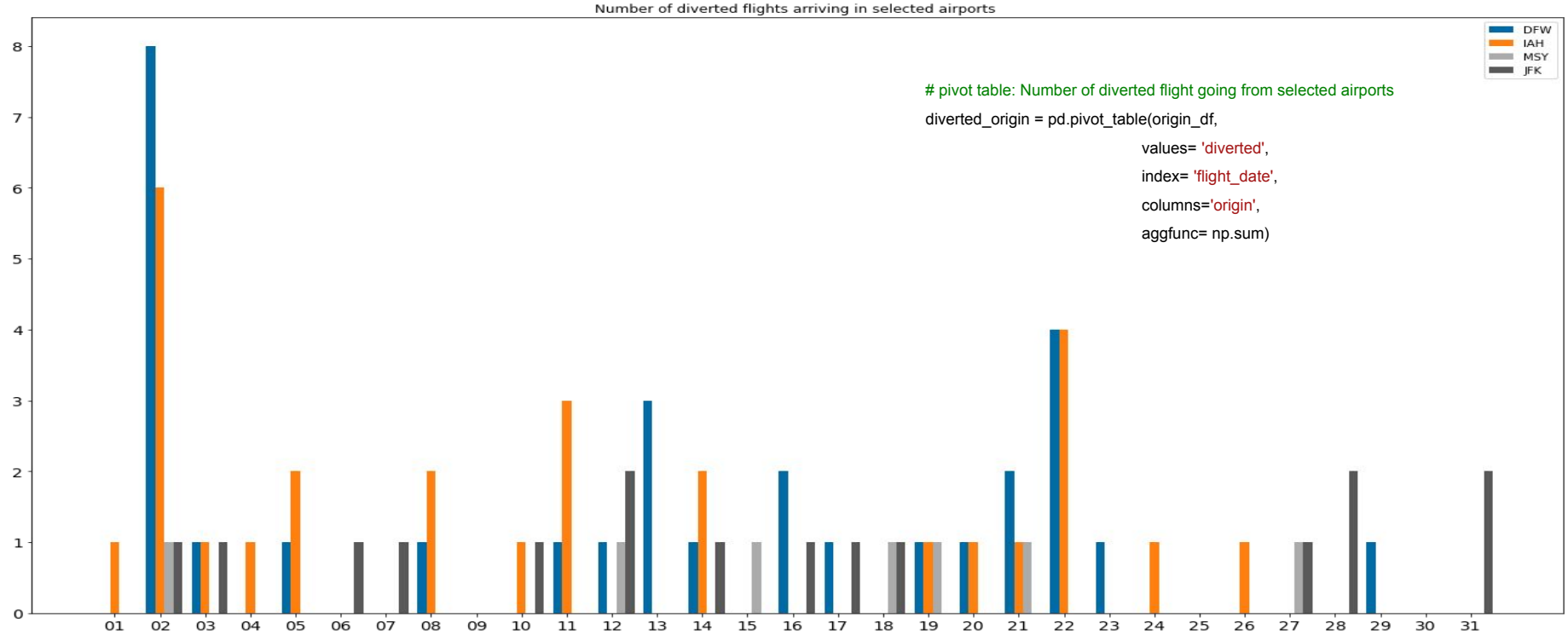
```
# to merge three dataframe: cancelled_origin and rain on date
cancelled_origin_rain = pd.merge(cancelled_origin, rain, left_on='flight_date', right_on= 'date')
cancelled_origin_rain
```

```
cancelled_origin_rain.corr(method= 'pearson')
```

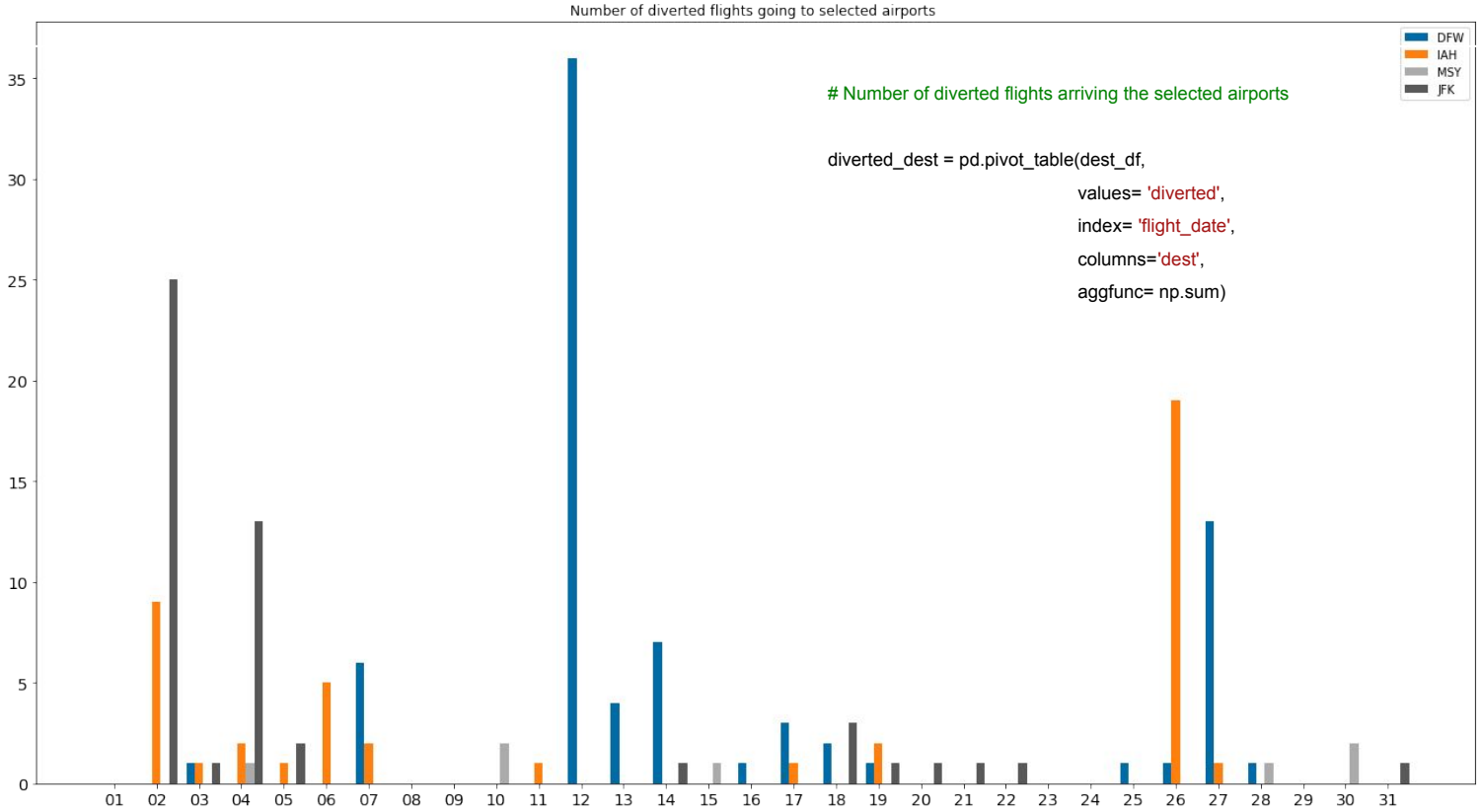
✓ 0.7s

	DFW_x	IAH_x	JFK_x	MSY_x	DFW_y	IAH_y	JFK_y	MSY_y
DFW_x	1.000000	0.674916	0.079196	0.676911	0.405139	0.590618	0.224040	0.251978
IAH_x	0.674916	1.000000	-0.139940	0.982049	-0.072775	0.559012	0.095822	0.381858
JFK_x	0.079196	-0.139940	1.000000	-0.067268	0.000486	-0.062419	0.119692	0.001721
MSY_x	0.676911	0.982049	-0.067268	1.000000	-0.121350	0.501403	0.085010	0.413690
DFW_y	0.405139	-0.072775	0.000486	-0.121350	1.000000	0.075741	0.021365	-0.187924
IAH_y	0.590618	0.559012	-0.062419	0.501403	0.075741	1.000000	-0.050206	0.121123
JFK_y	0.224040	0.095822	0.119692	0.085010	0.021365	-0.050206	1.000000	0.464538
MSY_y	0.251978	0.381858	0.001721	0.413690	-0.187924	0.121123	0.464538	1.000000

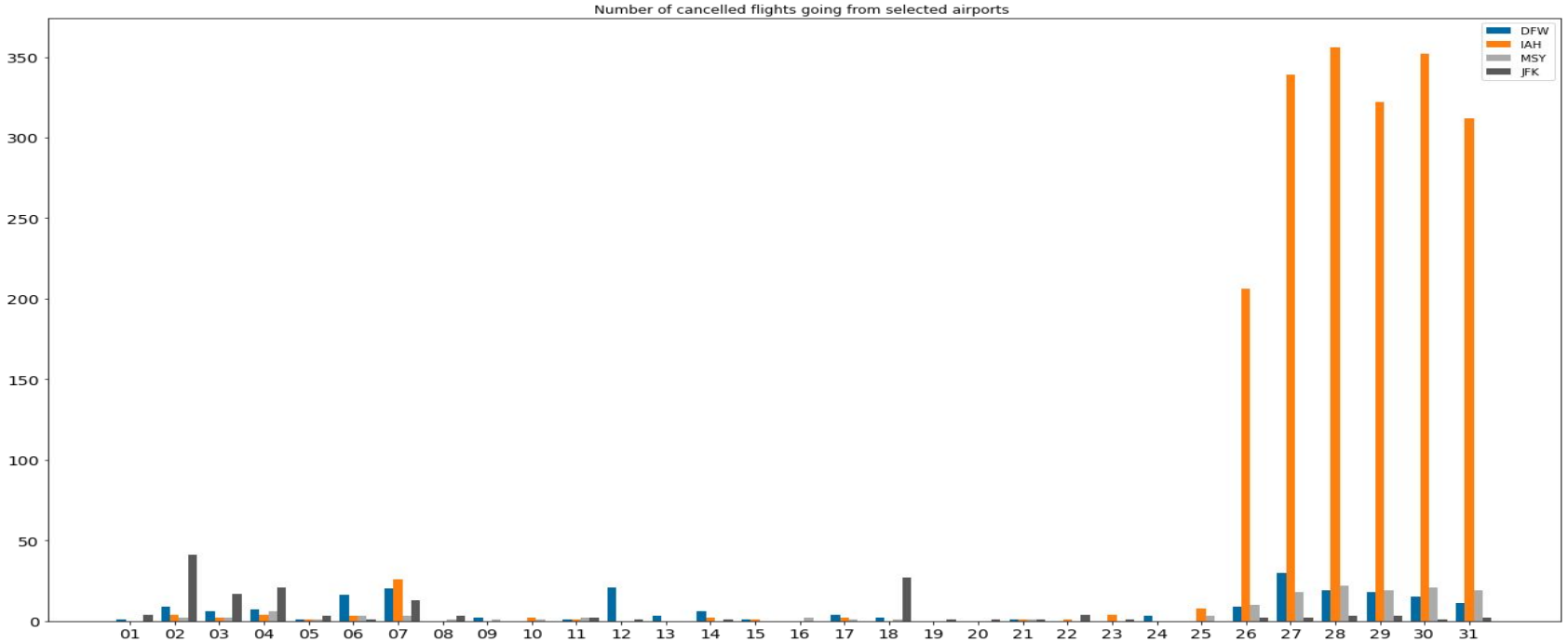
2: Number of diverted flights did not increase during Hurricane Harvey for origin airport



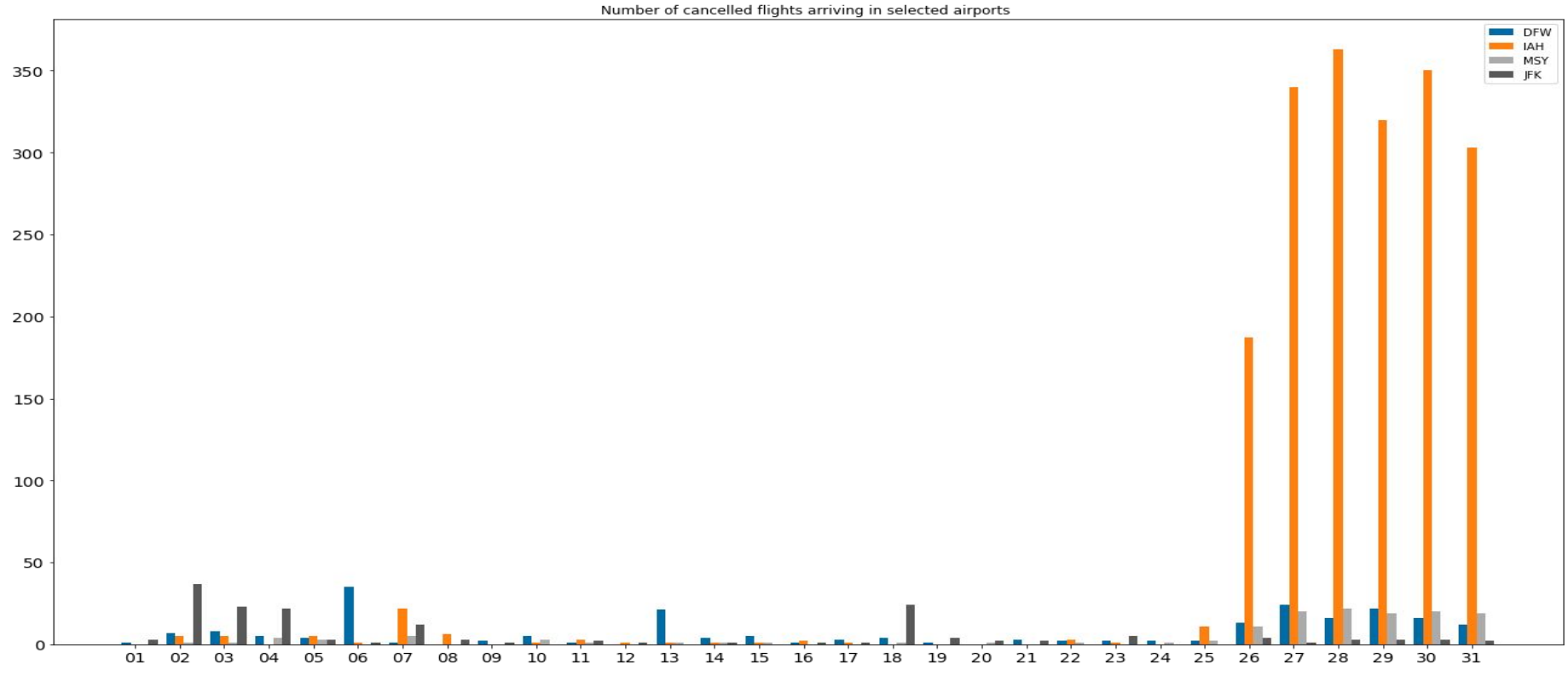
2: Number of diverted flights did not increase during Hurricane Harvey for destination airport



3: JFK(New York) was not affected by Harvey like Texas and Louisiana



3: JFK(New York) was not affected by Harvey like Texas and Louisiana



Summary of our insights

- Hurricane Harvey brought strong wind and more rain to Houston(IAH) and therefore caused more flight cancellations in IAH.
- However, the wind speed and total rain in Dallas and New Orleans did not have a significant increase during Hurricane Harvey.
- There was no correlation we could find between the graphs, and we reviewed our possibility of bias and corrected.

Limitation

Need more data on internal process: e.g. how airport and airline management team decide which flight should be cancelled

Time limitation: e.g. more time invested in this project help us polish our data analysing process, deeper-dive, finding new insights etc

Next Step

- Dig deeper into diverted flights
- Looking into more weather stations and airports

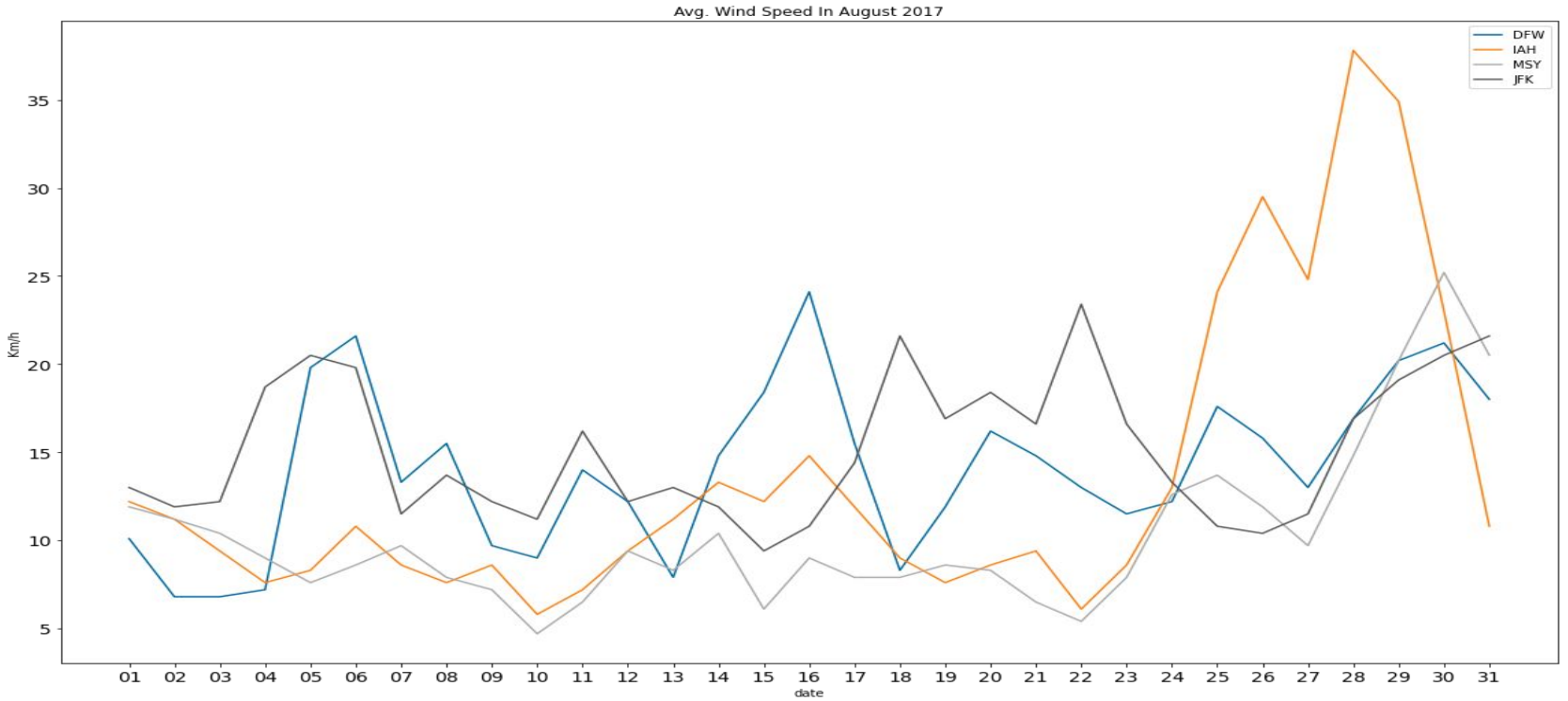
Reflections

- What did we learn during this project?
- What would you have done better?
- What are the ups and downs from this project?
- Did we meet our expectation of our study? And why?

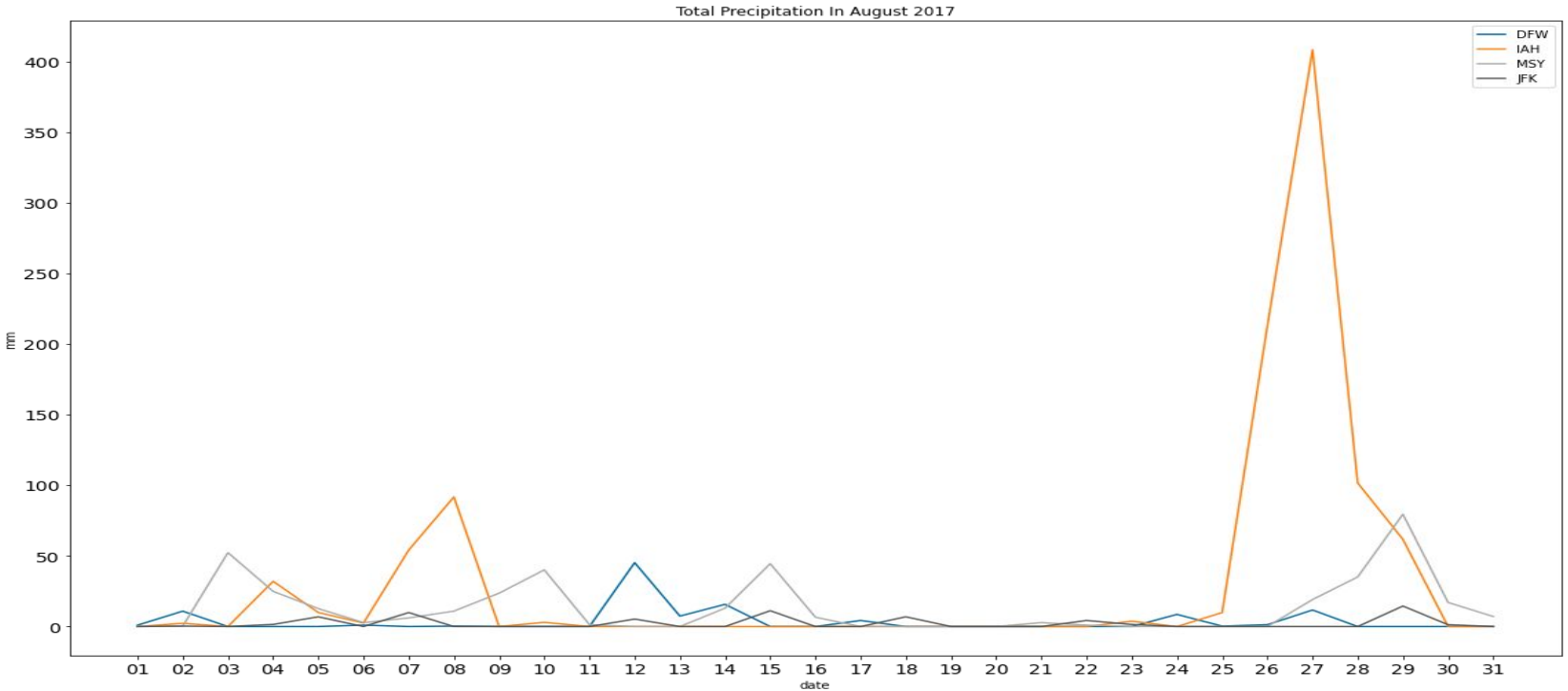
Dankeschön!

Reference for graphs

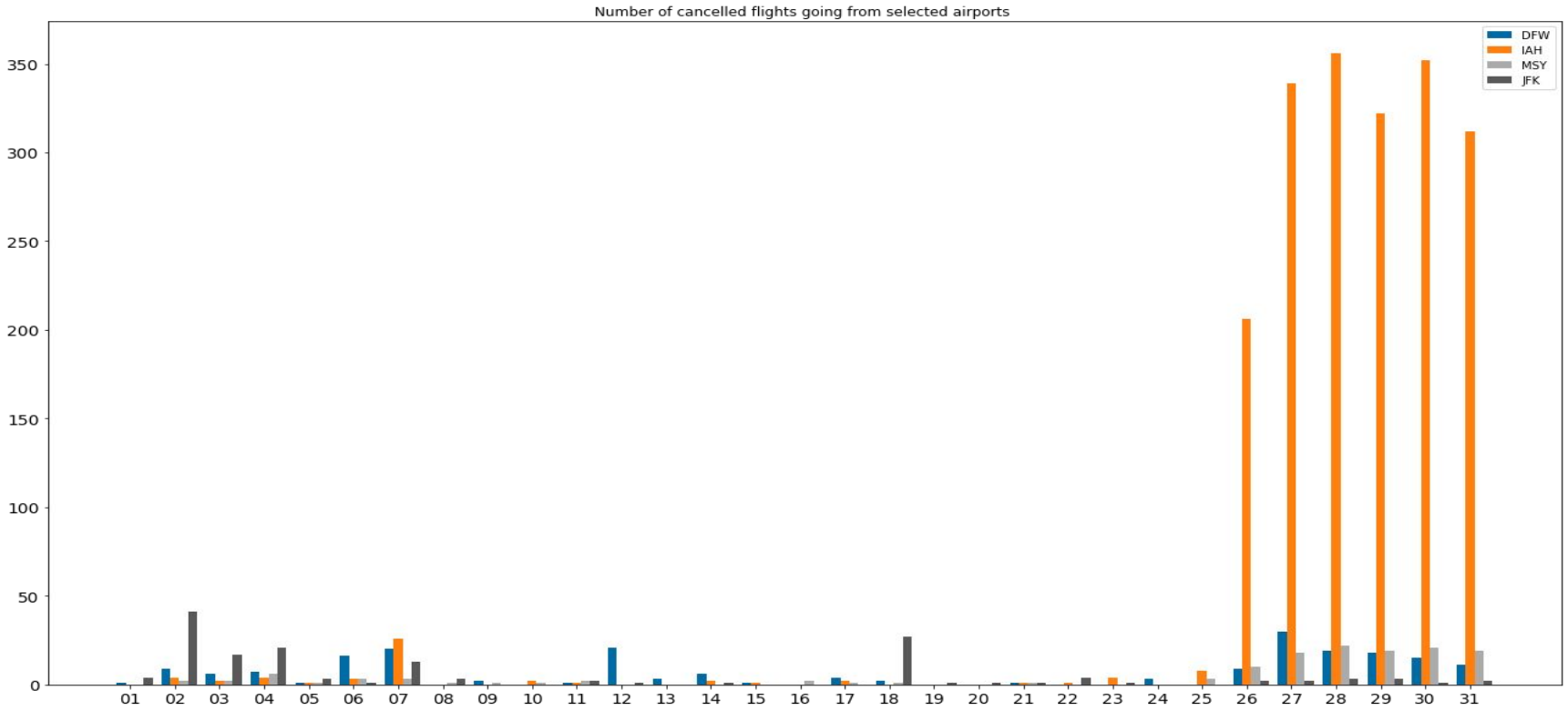
Graph 1 - Wind speed



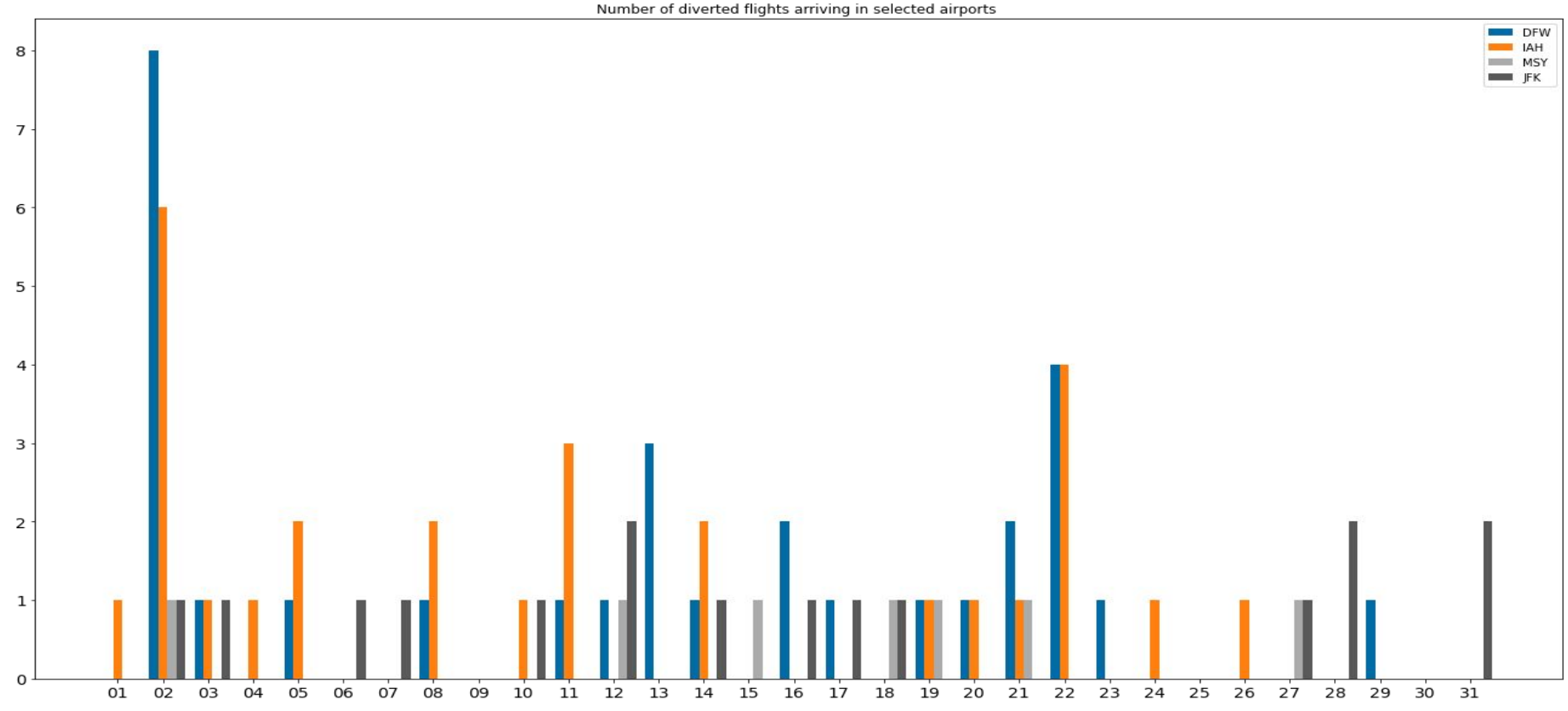
Graph 2 - Total precipitation



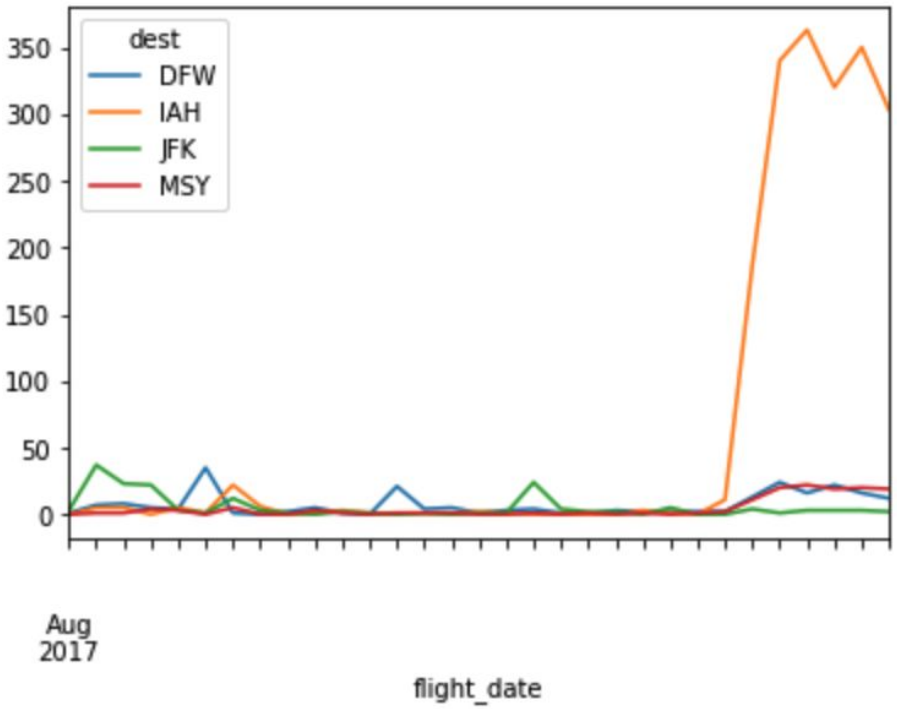
Graph 3 - Cancelled flights



Graph 4 - diverted flights



Graph 5 - Cancelled flights




```
# number of cancelled flight arriving in selected
airports

cancelled_dest = pd.pivot_table(dest_df,
                                values= 'cancelled',
                                index= 'flight_date',
                                columns='dest',
                                aggfunc= np.sum)
# plot
cancelled_dest.plot(xticks=cancelled_dest.index)
```

Reference for tables

Table 1

```
# pivot table: Number of cancelled flight going from
selected airports
cancelled_origin = pd.pivot_table(origin_df,
                                  values= 'cancelled',
                                  index= 'flight_date',
                                  columns='origin',
                                  aggfunc= np.sum)
```

origin	DFW	IAH	JFK	MSY
flight_date				
2017-08-01	1	0	4	0
2017-08-02	9	4	41	2
2017-08-03	6	2	17	2
2017-08-04	7	4	21	6
2017-08-05	1	1	3	1
2017-08-06	16	3	1	3
2017-08-07	20	26	13	3
2017-08-08	0	0	3	1
2017-08-09	2	0	0	1
2017-08-10	0	2	0	1
2017-08-11	1	1	2	2
2017-08-12	21	0	1	0
2017-08-13	3	0	0	0
2017-08-14	6	2	1	0
2017-08-15	1	1	0	0
2017-08-16	0	0	0	2
2017-08-17	4	2	0	1
2017-08-18	2	0	27	1
2017-08-19	0	0	1	0
2017-08-20	0	0	1	0
2017-08-21	1	1	1	1
2017-08-22	0	1	4	0
2017-08-23	0	4	1	0
2017-08-24	3	0	0	0
2017-08-25	0	8	0	3
2017-08-26	9	206	2	10
2017-08-27	30	339	2	18
2017-08-28	19	356	3	22
2017-08-29	18	322	3	19
2017-08-30	15	352	1	21
2017-08-31	11	312	2	19

Table 2

```
# number of cancelled flight arriving in selected
airports

cancelled_dest = pd.pivot_table(dest_df,
                                values= 'cancelled',
                                index= 'flight_date',
                                columns='dest',
                                aggfunc= np.sum)
```

dest	DFW	IAH	JFK	MSY
flight_date				
2017-08-01	1	0	3	0
2017-08-02	7	5	37	1
2017-08-03	8	5	23	1
2017-08-04	5	0	22	4
2017-08-05	4	5	3	3
2017-08-06	35	1	1	0
2017-08-07	1	22	12	5
2017-08-08	0	6	3	0
2017-08-09	2	0	1	0
2017-08-10	5	1	0	3
2017-08-11	1	3	2	1
2017-08-12	0	1	1	0
2017-08-13	21	1	0	1
2017-08-14	4	1	1	1
2017-08-15	5	1	0	1
2017-08-16	1	2	1	0
2017-08-17	3	1	1	0
2017-08-18	4	0	24	1
2017-08-19	1	0	4	0
2017-08-20	0	0	2	1
2017-08-21	3	0	2	0
2017-08-22	2	3	0	1
2017-08-23	2	1	5	0
2017-08-24	2	0	0	1
2017-08-25	2	11	0	2
2017-08-26	13	187	4	11
2017-08-27	24	340	1	20
2017-08-28	16	363	3	22
2017-08-29	22	320	3	19
2017-08-30	16	350	3	20
2017-08-31	12	303	2	19

Table 3

```
# flights that took off --> going to somewhere else,
no need to divert
diverted_origin = pd.pivot_table(origin_df,
                                 values= 'diverted',
                                 index= 'flight_date',
                                 columns='origin',
                                 aggfunc= np.sum)
```

origin	DFW	IAH	JFK	MSY
flight_date				
2017-08-01	0	1	0	0
2017-08-02	8	6	1	1
2017-08-03	1	1	1	0
2017-08-04	0	1	0	0
2017-08-05	1	2	0	0
2017-08-06	0	0	1	0
2017-08-07	0	0	1	0
2017-08-08	1	2	0	0
2017-08-09	0	0	0	0
2017-08-10	0	1	1	0
2017-08-11	1	3	0	0
2017-08-12	1	0	2	1
2017-08-13	3	0	0	0
2017-08-14	1	2	1	0
2017-08-15	0	0	0	1
2017-08-16	2	0	1	0
2017-08-17	1	0	1	0
2017-08-18	0	0	1	1
2017-08-19	1	1	0	1
2017-08-20	1	1	0	0
2017-08-21	2	1	0	1
2017-08-22	4	4	0	0
2017-08-23	1	0	0	0
2017-08-24	0	1	0	0
2017-08-25	0	0	0	0
2017-08-26	0	1	0	0
2017-08-27	0	0	1	1
2017-08-28	0	0	2	0
2017-08-29	1	0	0	0
2017-08-30	0	0	0	0
2017-08-31	0	0	2	0

Table 4

```
# Number of diverted flights arriving in selected
airports
diverted_dest = pd.pivot_table(dest_df,
                               values='diverted',
                               index='flight_date',
                               columns='dest',
                               aggfunc= np.sum)
```

dest flight_date	DFW	IAH	JFK	MSY
2017-08-01	0	0	0	0
2017-08-02	0	9	25	0
2017-08-03	1	1	1	0
2017-08-04	0	2	13	1
2017-08-05	0	1	2	0
2017-08-06	0	5	0	0
2017-08-07	6	2	0	0
2017-08-08	0	0	0	0
2017-08-09	0	0	0	0
2017-08-10	0	0	0	2
2017-08-11	0	1	0	0
2017-08-12	36	0	0	0
2017-08-13	4	0	0	0
2017-08-14	7	0	1	0
2017-08-15	0	0	0	1
2017-08-16	1	0	0	0
2017-08-17	3	1	0	0
2017-08-18	2	0	3	0
2017-08-19	1	2	1	0
2017-08-20	0	0	1	0
2017-08-21	0	0	1	0
2017-08-22	0	0	1	0
2017-08-23	0	0	0	0
2017-08-24	0	0	0	0
2017-08-25	1	0	0	0
2017-08-26	1	19	0	0
2017-08-27	13	1	0	0
2017-08-28	1	0	0	1
2017-08-29	0	0	0	0
2017-08-30	0	0	0	2
2017-08-31	0	0	1	0

Table 5

```
# pivot table for wind speed
wind_speed = pd.pivot_table(weather_df,
                            values= 'wind_speed',
                            index= 'date',
                            columns='w_faa')
```

w_faa	DFW	IAH	JFK	MSY
date				
2017-08-01	10.1	12.2	13.0	11.9
2017-08-02	6.8	11.2	11.9	11.2
2017-08-03	6.8	9.4	12.2	10.4
2017-08-04	7.2	7.6	18.7	9.0
2017-08-05	19.8	8.3	20.5	7.6
2017-08-06	21.6	10.8	19.8	8.6
2017-08-07	13.3	8.6	11.5	9.7
2017-08-08	15.5	7.6	13.7	7.9
2017-08-09	9.7	8.6	12.2	7.2
2017-08-10	9.0	5.8	11.2	4.7
2017-08-11	14.0	7.2	16.2	6.5
2017-08-12	12.2	9.4	12.2	9.4
2017-08-13	7.9	11.2	13.0	8.3
2017-08-14	14.8	13.3	11.9	10.4
2017-08-15	18.4	12.2	9.4	6.1
2017-08-16	24.1	14.8	10.8	9.0
2017-08-17	15.5	11.9	14.4	7.9
2017-08-18	8.3	9.0	21.6	7.9
2017-08-19	11.9	7.6	16.9	8.6
2017-08-20	16.2	8.6	18.4	8.3
2017-08-21	14.8	9.4	16.6	6.5
2017-08-22	13.0	6.1	23.4	5.4
2017-08-23	11.5	8.6	16.6	7.9
2017-08-24	12.2	13.0	13.3	12.6
2017-08-25	17.6	24.1	10.8	13.7
2017-08-26	15.8	29.5	10.4	11.9
2017-08-27	13.0	24.8	11.5	9.7
2017-08-28	16.9	37.8	16.9	14.8
2017-08-29	20.2	34.9	19.1	20.2
2017-08-30	21.2	23.0	20.5	25.2
2017-08-31	18.0	10.8	21.6	20.5

Table 6

```
# pivot table for rain

rain  = pd.pivot_table(weather_df,
                          values= 'rain',
                          index= 'date',
                          columns='w_faa')
```

w_faa	DFW	IAH	JFK	MSY
date				
2017-08-01	1.0	0.0	0.0	0.0
2017-08-02	10.9	2.3	0.3	0.3
2017-08-03	0.0	0.0	0.0	52.3
2017-08-04	0.0	32.0	1.5	24.9
2017-08-05	0.0	9.9	6.9	12.7
2017-08-06	1.0	2.5	0.0	2.8
2017-08-07	0.0	54.1	9.9	6.1
2017-08-08	0.3	91.7	0.0	10.9
2017-08-09	0.0	0.0	0.0	23.6
2017-08-10	0.0	3.0	0.0	40.1
2017-08-11	0.0	0.0	0.0	1.3
2017-08-12	45.2	0.0	5.3	0.0
2017-08-13	7.4	0.0	0.0	0.0
2017-08-14	15.7	0.0	0.0	13.0
2017-08-15	0.0	0.0	11.2	44.5
2017-08-16	0.0	0.0	0.0	6.6
2017-08-17	4.3	0.0	0.0	0.0
2017-08-18	0.0	0.0	6.9	0.0
2017-08-19	0.0	0.0	0.0	0.0
2017-08-20	0.0	0.0	0.0	0.0
2017-08-21	0.0	0.0	0.0	2.8
2017-08-22	0.0	0.0	4.3	1.0
2017-08-23	0.0	3.8	1.5	0.0
2017-08-24	8.6	0.0	0.0	0.0
2017-08-25	0.3	9.9	0.0	0.0
2017-08-26	1.3	212.6	0.0	0.0
2017-08-27	11.7	408.2	0.0	19.1
2017-08-28	0.0	101.6	0.0	35.1
2017-08-29	0.0	61.7	14.5	79.5
2017-08-30	0.0	0.0	1.3	17.0
2017-08-31	0.0	0.0	0.0	7.1