Exploratory Analysis of Flight Cancellations From Weather in the U.S.

From 2015 - 2018

INDEX

PRESENTATION COMPONENTS:

- PROJECT DESCRIPTION
- DATA SOURCES
- DATA RETRIEVAL AND CLEANING
- DATA VISUALIZATION AND ANALYSIS
- CONCLUSIONS

INTRODUCTION

We looked at U.S. flight data from 2015-2018 and focused specifically on cancellations from weather. We looked at weather factors such as wind speed, precipitation, and temperature as well as time of year to make determinations around whether those factors had an effect on cancellation.

RESEARCH QUESTIONS

01 MONTH/YEAR
How does month of year impact flight cancellations in the U.S.?

02 WIND SPEED
How does wind speed affect U.S. flight cancellations by month of year?

03 TEMPERATURE
How does temperature affect U.S. flight cancellations by month of year?

04 PRECIPITATION
How does precipitation type affect U.S. flight cancellations by month of year?

OUR DATA

- LOCATION
- WIND SPEED
- TEMPERATURE
- PRECIPITATION
- TIME

DATA SOURCES

01 Airline Delay and Cancellation Data, 2009 - 2018 | Kaggle

02 Visual Crossing Weather | API

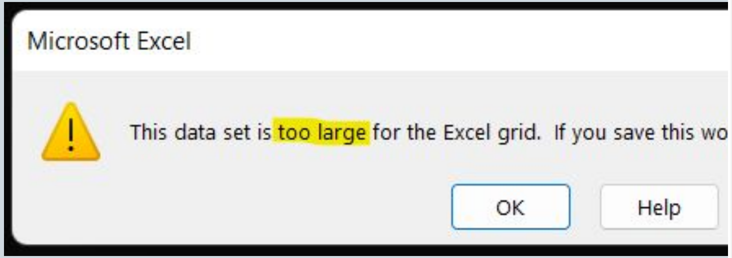
03 Geoapify | API

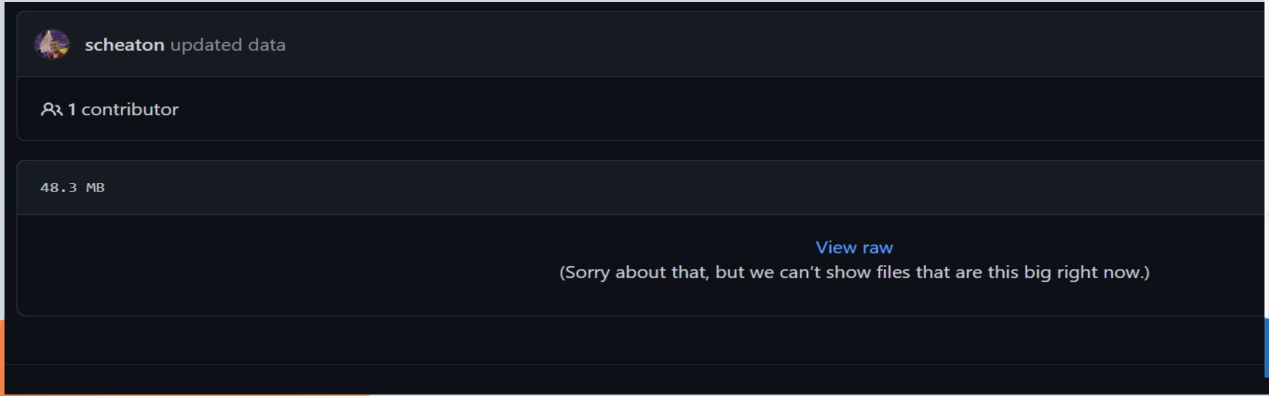
DATA EXPLORATION

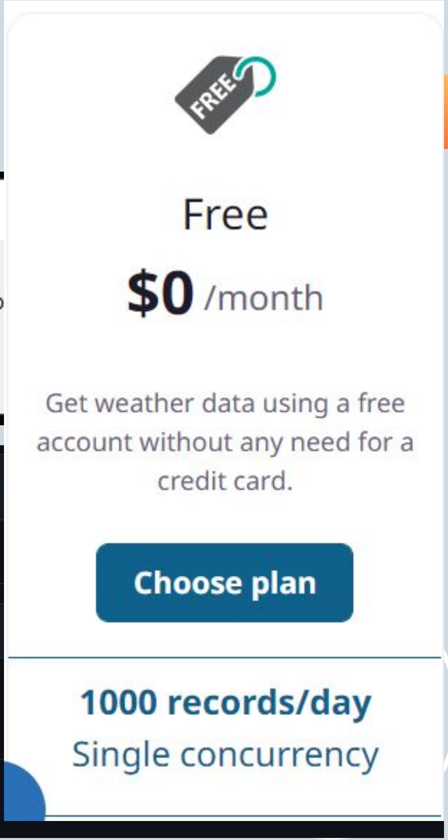
We compiled the following initial data for our exploration:

- Number of observations (raw dataset): Tens of millions
- Flight cancellation data included in our analysis:
  - Cancellations only due to weather
  - Cancelled flight data from 2009-2018
    - Narrowed data down to 2015 to 2018 to look at the most recent 4 years of data
    - Narrowed data to look at the top 5 airports by traffic
      - Airports: DFW, ATL, DEN, ORD, JFK
- Daily Weather Data from Historical Weather API for our selected date range and airport locations
- Longitude, Latitude information from Global Airport Database
- Airport details from the Geoapify API

DATA CLEANUP







```python
# Search for null values that d
original_df.isna().sum()
```

```
FL_DATE                   0
OP_CARRIER                0
OP_CARRIER_FL_NUM         0
ORIGIN                    0
DEST                      0
CRS_DEP_TIME              0
DEP_TIME              86153
DEP_DELAY             86153
TAXI_OUT              89047
WHEELS_OFF            89047
WHEELS_ON             92513
TAXI_IN               92513
CRS_ARR_TIME              0
ARR_TIME              92513
ARR_DELAY            105071
CANCELLED                 0
WEATHER_DELAY       4755640
NAS_DELAY           4755640
SECURITY_DELAY      4755640
LATE_AIRCRAFT_DELAY 4755640
Unnamed: 27         5819079
dtype: int64
```

```python
filtered_df = filtered_df[['Date', 'Origin', 'Destination', 'Expected Departure Time',\
                           'Expected Arrival Time', 'Distance', 'Weather Delay']]
return filtered_df
```

```python
def clean_flight_data(year):

    file = "raw_data/" + str(year) + ".csv"

    original_df = pd.read_csv(file, sep=",", header=0, index_col=False,
                              usecols=[0,3,4,5,6,7,12,13,14,16,21,23],
                              na_filter = True)\
                             .reset_index(drop=True)\
                             .fillna(0)

    filtered_df = original_df.loc[(original_df['ORIGIN'].isin(focused_airports) & \
                                   ((original_df['CANCELLATION_CODE'] == 0 )| \
                                    (original_df['CANCELLATION_CODE'] == 'B'))), :]

    filtered_df.loc[filtered_df['CANCELLATION_CODE'] == 'B', 'WEATHER_DELAY'] = 'CANCELLED'
    filtered_df = filtered_df.drop(columns = 'CANCELLATION_CODE')

    filtered_df = filtered_df.rename(columns={'FL_DATE'     : 'Date',
                                              'ORIGIN'       : 'Origin',
                                              'DEST'         : 'Destination',
                                              'CRS_DEP_TIME'  : 'Expected Departure Time',
                                              'DEP_TIME'      : 'Actual Departure Time',
                                              'DEP_DELAY'     : 'Departure Delay',
                                              'CRS_ARR_TIME'  : 'Expected Arrival Time',
                                              'ARR_TIME'      : 'Arrival Time',
                                              'ARR_DELAY'     : 'Arrival Delay',
                                              'DISTANCE'      : 'Distance',
                                              'WEATHER_DELAY' : 'Weather Delay'})

    filtered_df = filtered_df[['Date', 'Origin', 'Destination', 'Expected Departure Time',\
                               'Expected Arrival Time', 'Distance', 'Weather Delay']]
    return filtered_df
```

```python
for year in range(2015,2019):        ## Takes about 7 minutes ##
    output_path = 'clean_data/focused_airports_' + str(year) + '.csv'
    clean_flight_data(year).to_csv(output_path, index=False)
```

```python
# Loop through the sample dataframe
for index,row in source_df.iterrows():

    #----------Origin Latitude and Longitude
    lat=row["Latitude"]
    lng=row["Longitude"]
    datetime= row["Date"]
    # Build URL
    base_url ="https://weather.visualcrossing.com/VisualCrossingWebServices/rest/services/timeline/"
    query_1 = (f"{lat},{lng}/{datetime}/?key={weather_api_key}")
    query_2 = "&include=obs%2Cfcst%2Cstats%2Calerts%2Ccurrent%2Chistfcst"
    query_3 = "&elements=tempmax,precip,preciptype,windspeed"
    new_url= base_url + query_1 + query_2 + query_3
#     get the response

    # Use try and except to skip the missing data
    try:
        response = requests.get(new_url).json()
        source_df.loc[index,"Max Temp"]=response['days'][0]['tempmax']
        source_df.loc[index,"Precip"]=response['days'][0]['precip']
        precip_type_list = response['days'][0]['preciptype']
        precip_type_str=""
        if precip_type_list != None :
            for precip_type in precip_type_list:
                precip_type_str += precip_type+","
        else:
            precip_type_str = "NA"
        source_df.loc[index,"Precip Type"]=precip_type_str
        source_df.loc[index,"Wind Speed"]=response['days'][0]['windspeed']

    except (KeyError, IndexError, JSONDecodeError):
        print("Data not found... skipping.")
    except requests.Timeout:
        print("Request Timeout...")
    except requests.ConnectionError:
        print("ConnectionError...")
```

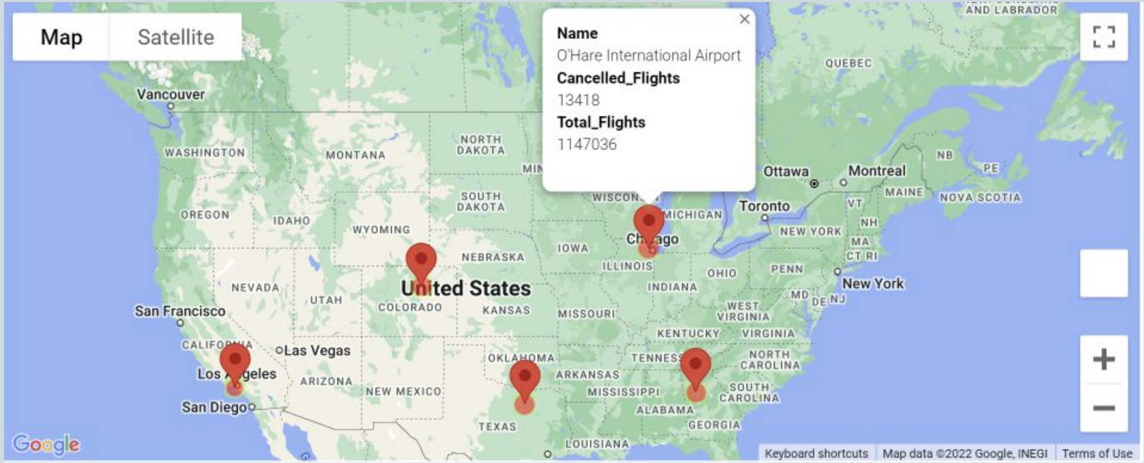
weather.visualcrossing.com/VisualCrossingWebServices/rest/services/timeline/32.896,-97.037/2015-01-01/?

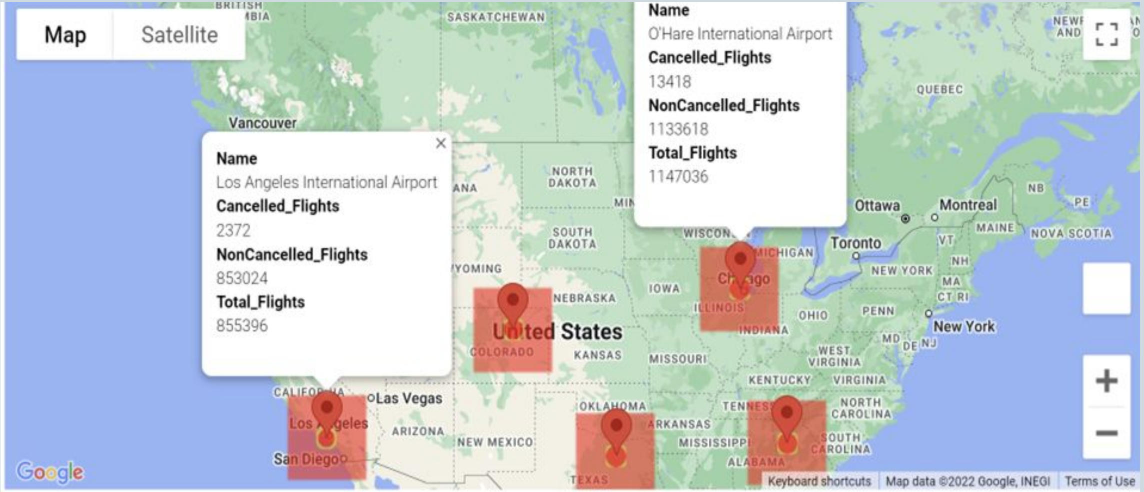

```
{
    "queryCost": 1,
    "latitude": 32.896,
    "longitude": -97.037,
    "resolvedAddress": "32.896,-97.037",
    "address": "32.896,-97.037",
    "timezone": "America/Chicago",
    "tzoffset": -6,
    "days": [
        {
            "tempmax": 36,
            "precip": 0.58,
            "preciptype": [
                "rain",
                "snow"
            ],
            "windspeed": 8,
            "normal": {
                "tempmax": [
                    29.9,
                    55.1,
                    82.1
                ],
                "precip": [
                    0,
                    0,
                    0.6
                ],
                "windspeed": [
                    8.1,
                    18.8,
                    28.6
                ]
            }
        }
    ]
}
```

VISUALIZATIONS AND ANALYSIS

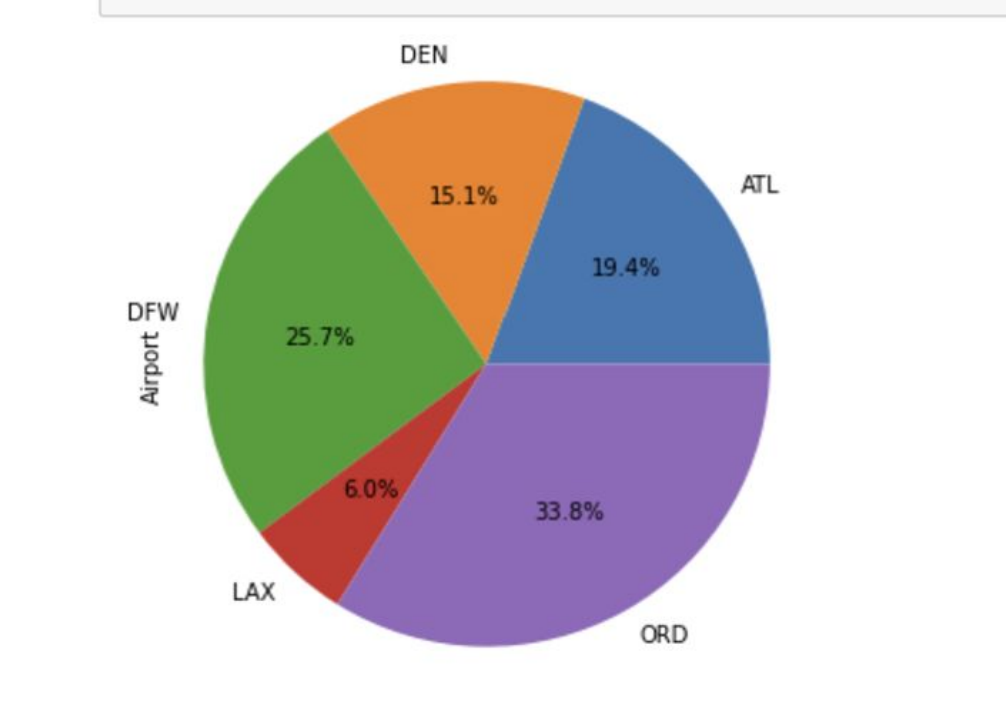
- LOCATION
- WIND SPEED
- TEMPERATURE
- PRECIPITATION
- TIME

LOCATION

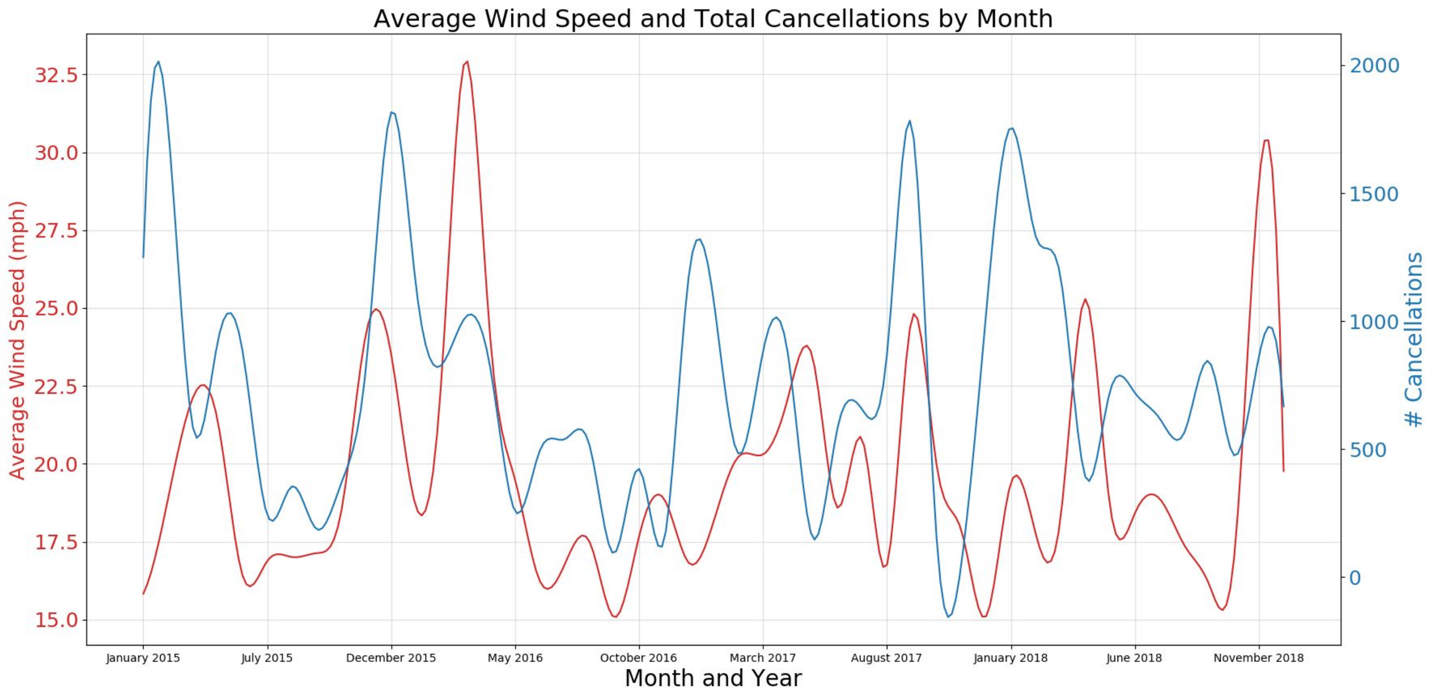

Map
Satellite
Name
O'Hare International Airport
Cancelled_Flights
13418
Total_Flights
1147036
United States
Vancouver
San Francisco
Los Angeles
San Diego
Las Vegas
Chicago
Ottawa
Montreal
Toronto
New York
Keyboard shortcuts
Map data ©2022 Google, INEGI
Terms of Use


Map
Satellite
Name
Los Angeles International Airport
Cancelled_Flights
2372
NonCancelled_Flights
853024
Total_Flights
855396
Name
O'Hare International Airport
Cancelled_Flights
13418
NonCancelled_Flights
1133618
Total_Flights
1147036
United States
Vancouver
Las Vegas
San Diego
Ottawa
Montreal
Toronto
New York
Keyboard shortcuts
Map data ©2022 Google, INEGI
Terms of Use

LOCATION

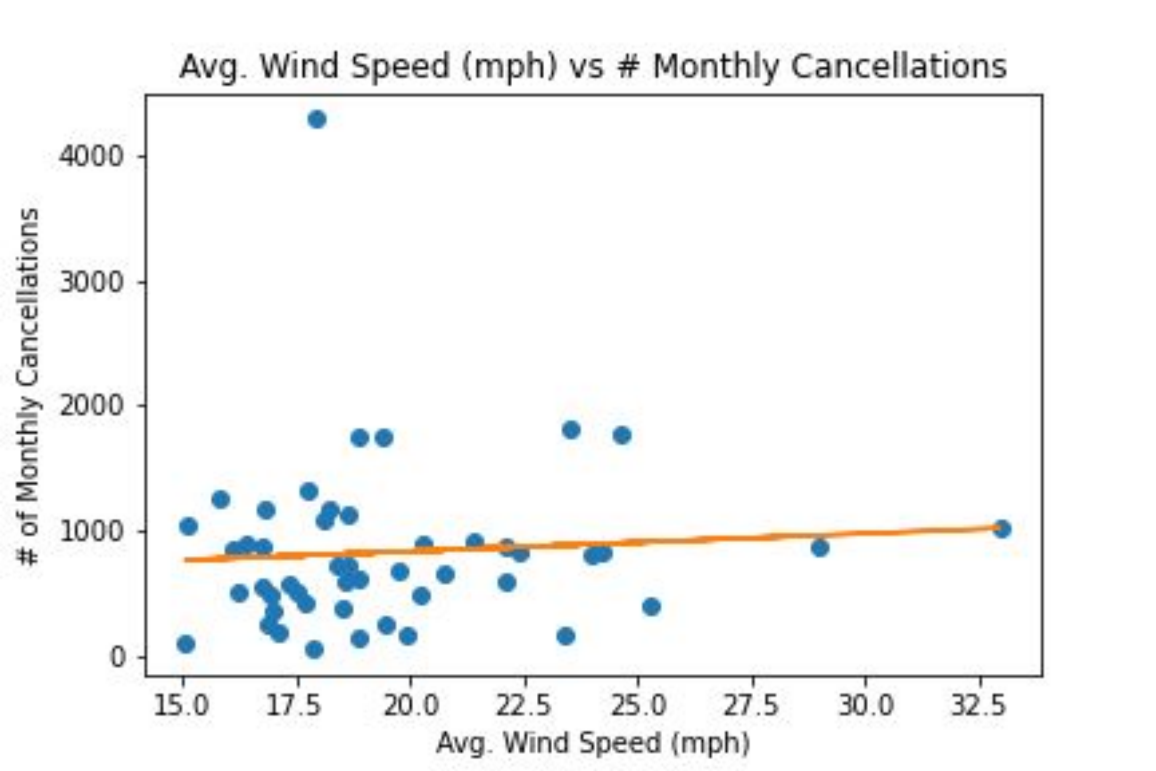


WIND SPEED



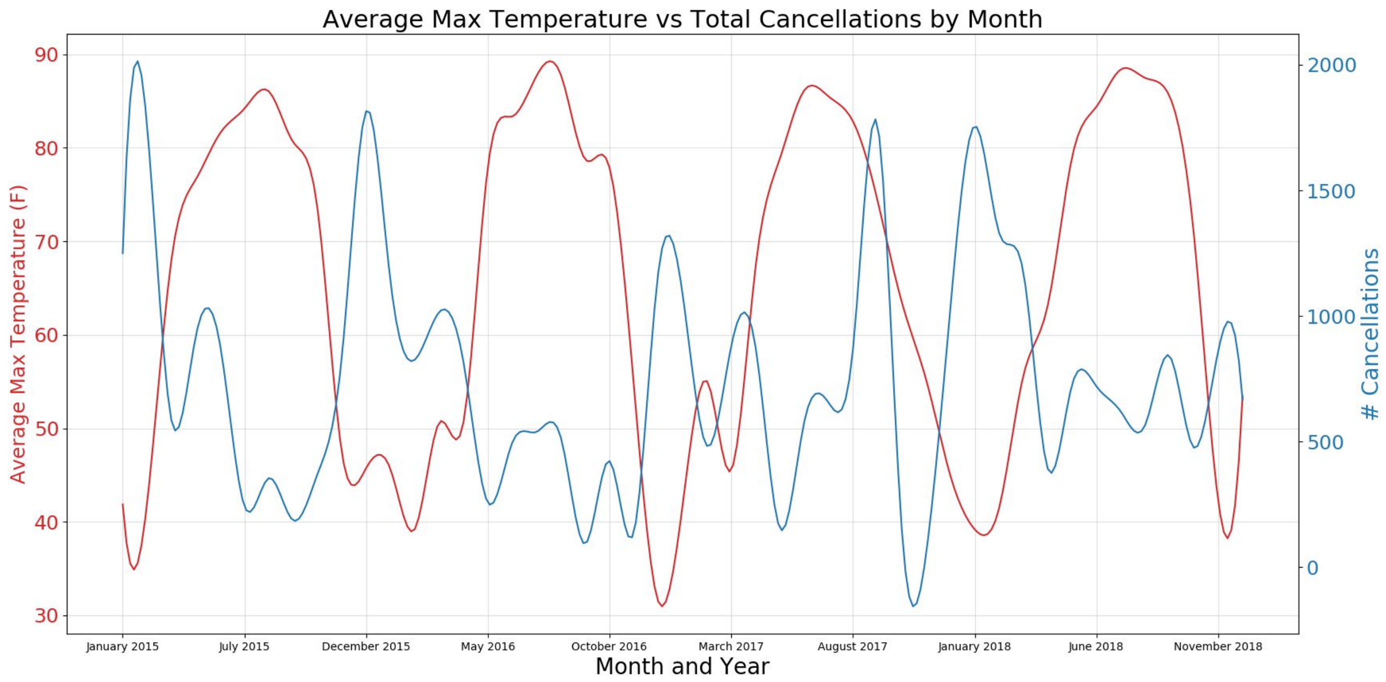
Average Wind Speed and Total Cancellations by Month
Average Wind Speed (mph)
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
Cancellations
2000
1500
1000
500
0
January 2015
July 2015
December 2015
May 2016
October 2016
March 2017
August 2017
January 2018
June 2018
November 2018
Month and Year

WIND SPEED

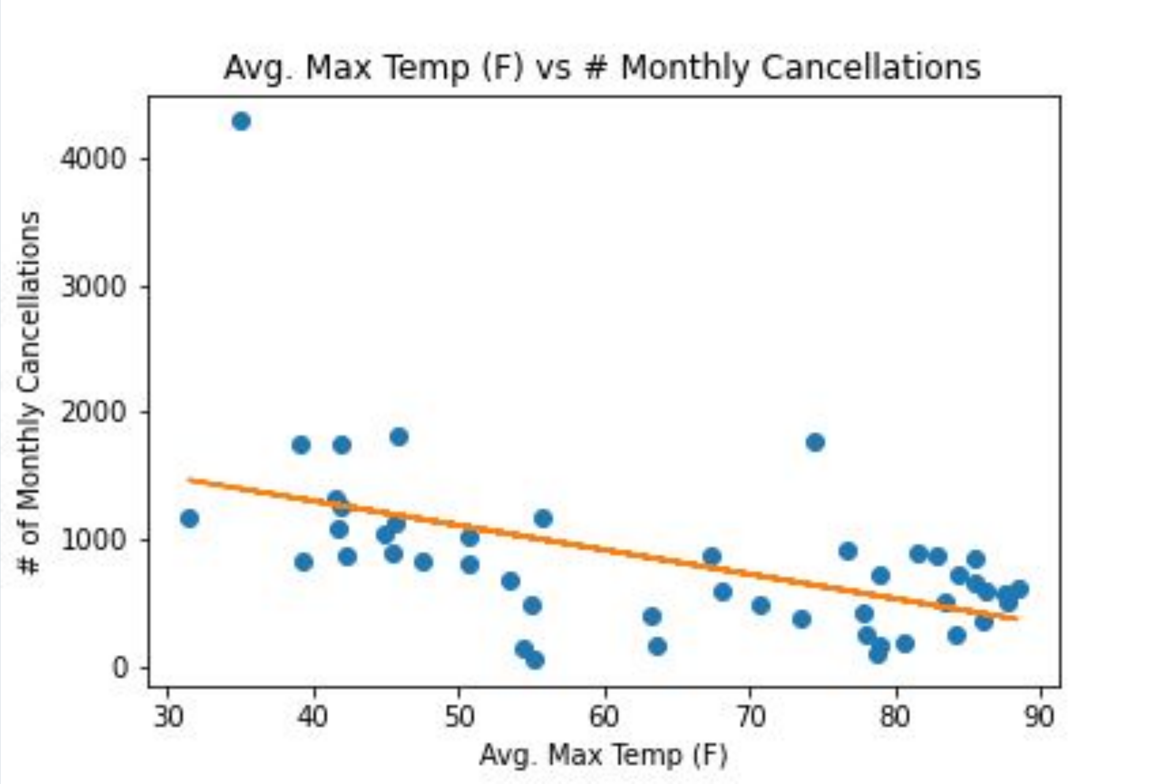


Avg. Wind Speed (mph) vs # Monthly Cancellations
# of Monthly Cancellations
0
1000
2000
3000
4000
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
Avg. Wind Speed (mph)

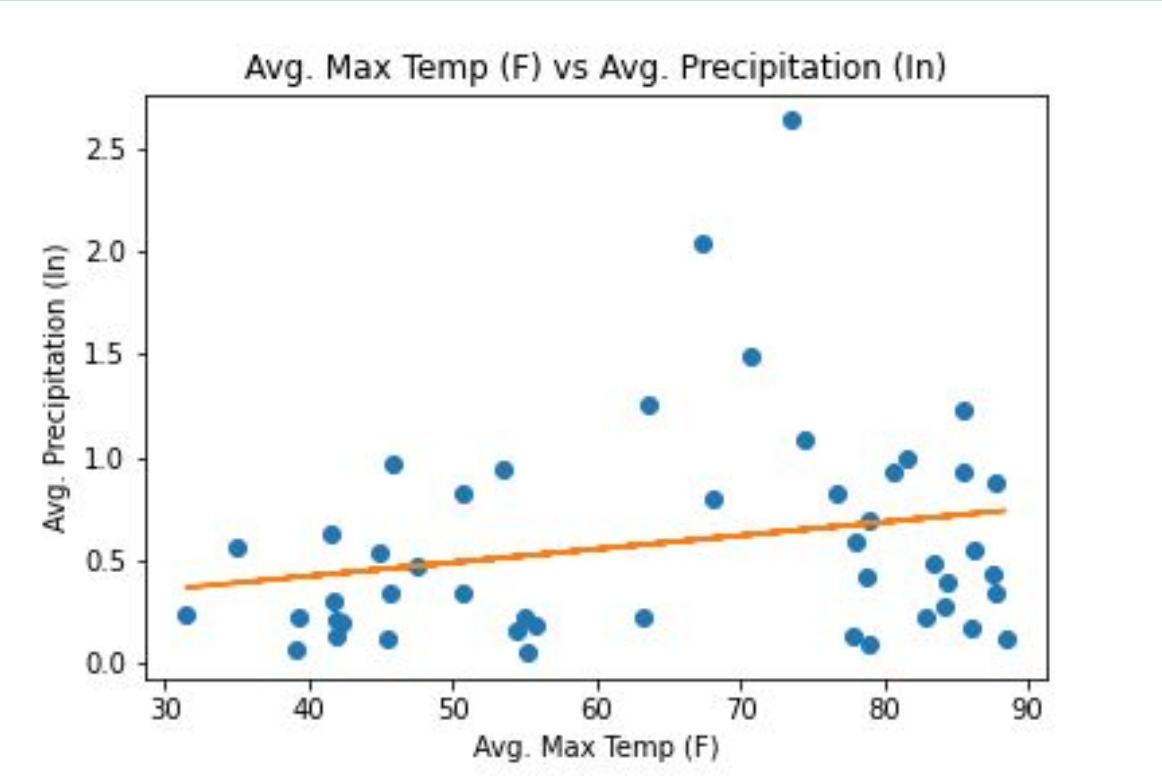
TEMPERATURE



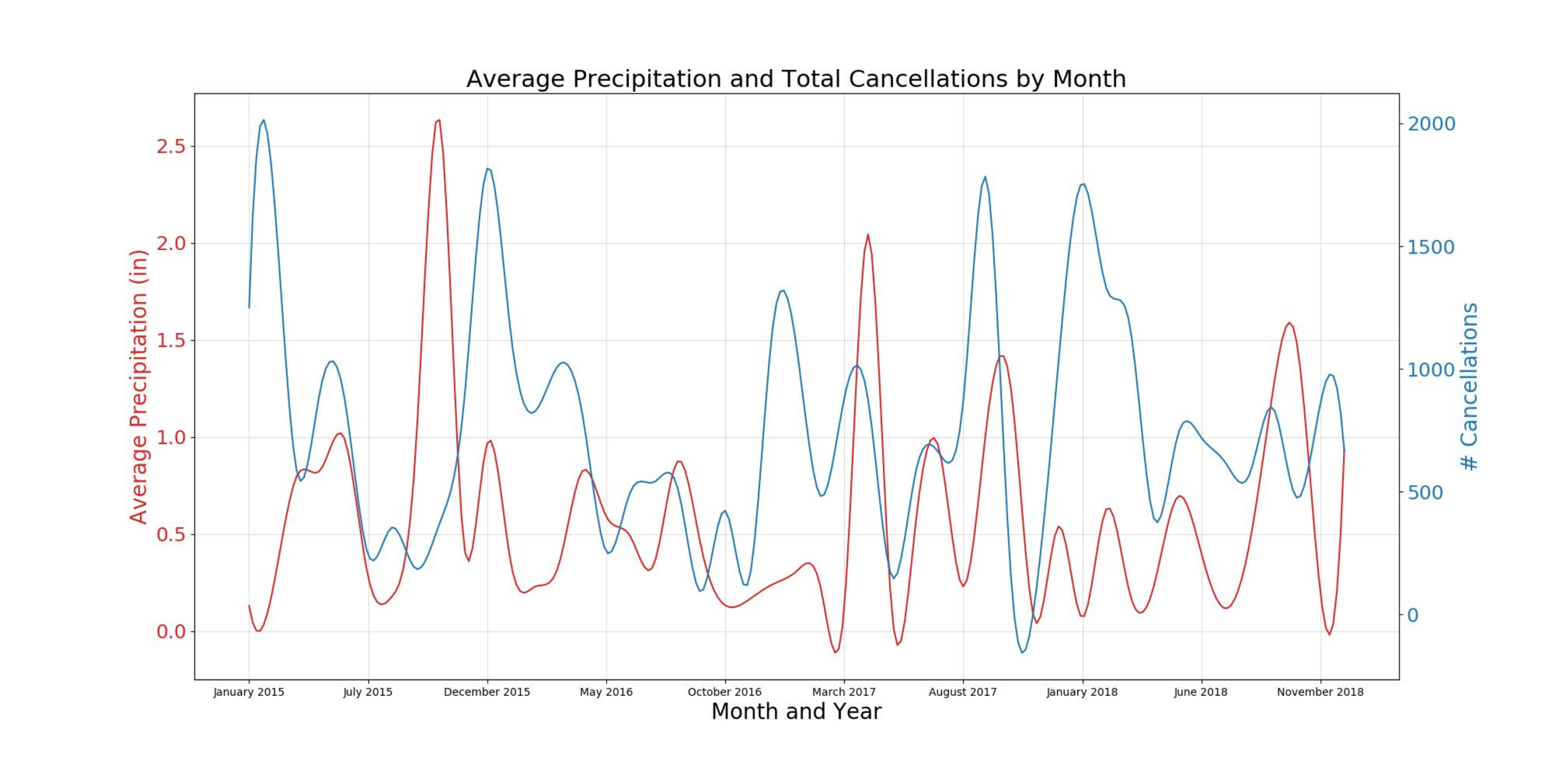
TEMPERATURE



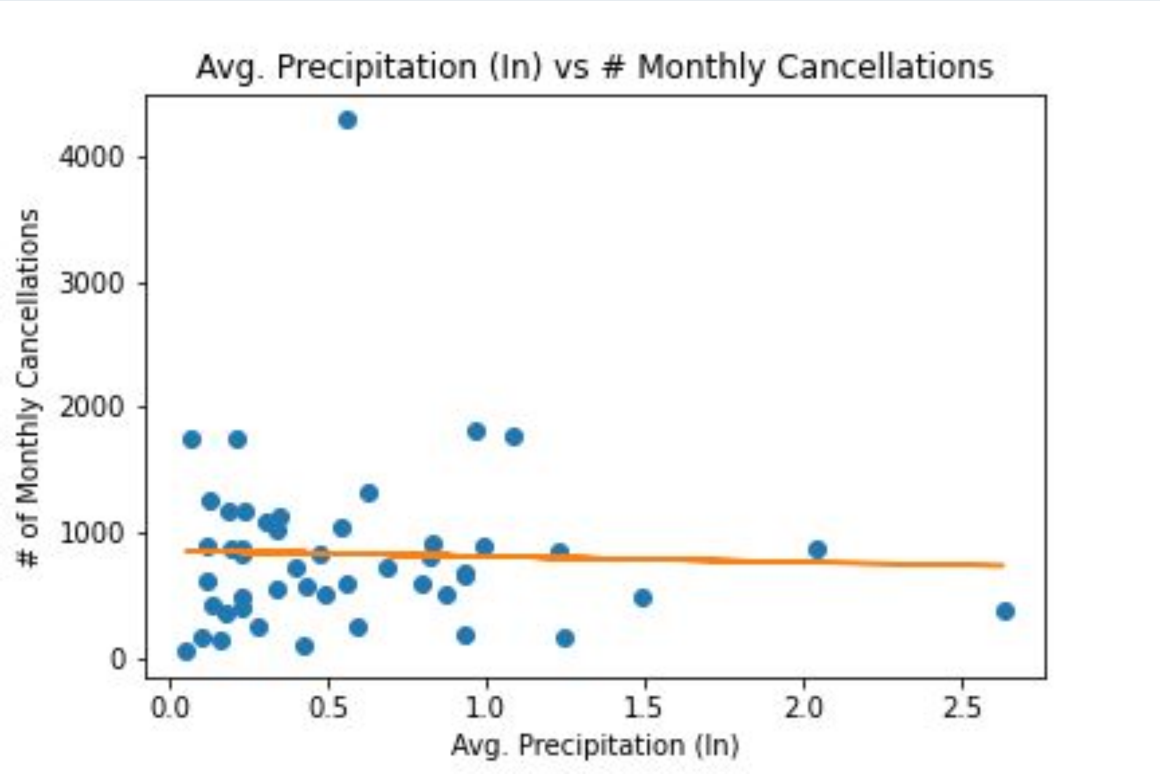
TEMPERATURE



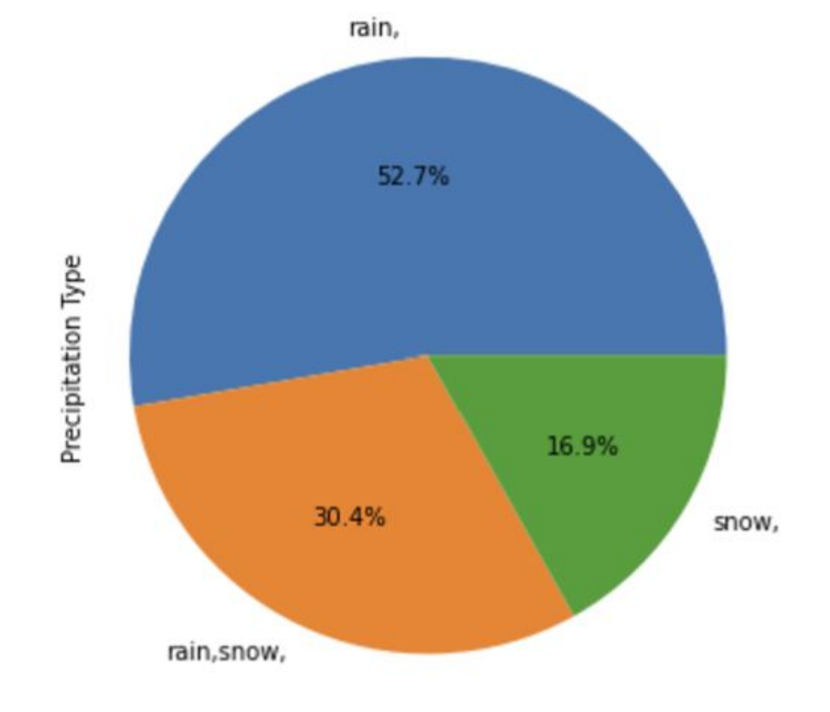
PRECIPITATION



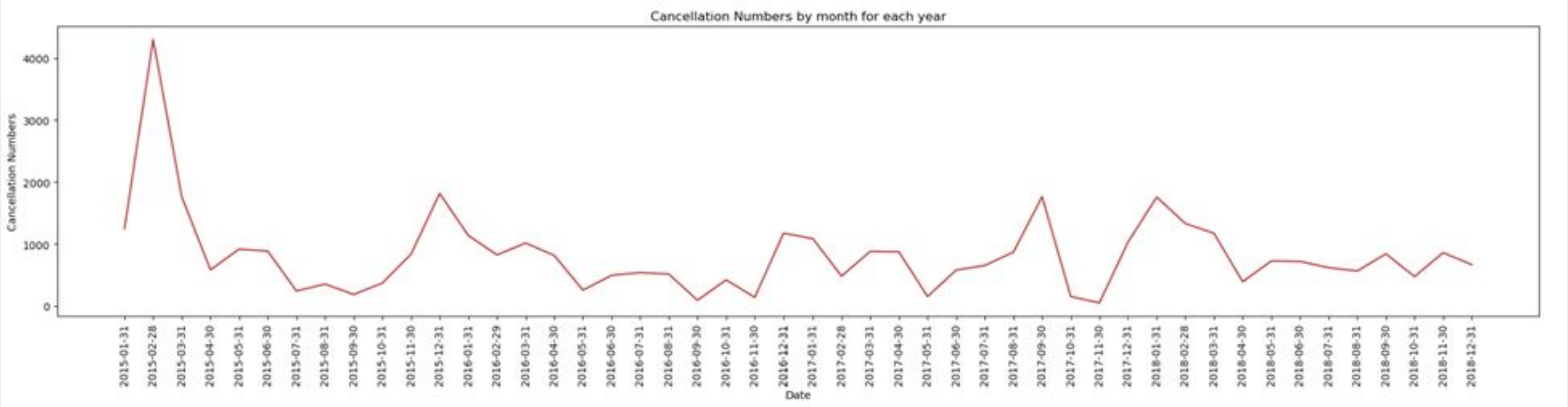
PRECIPITATION

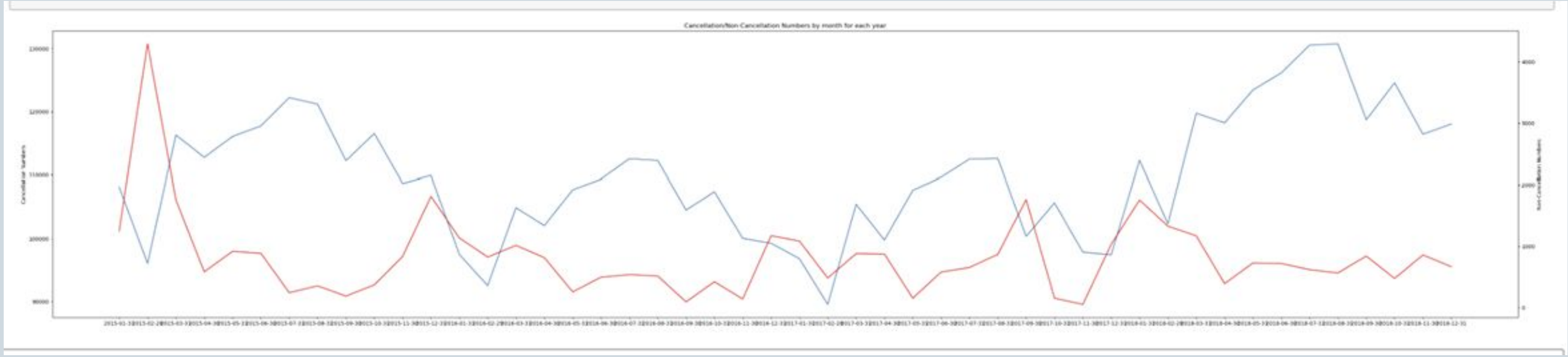


PRECIPITATION



Time


Cancellation Numbers by month for each year
Cancellation Numbers
0
1000
2000
3000
4000
Date
2015-01-31
2015-02-28
2015-03-31
2015-04-30
2015-05-31
2015-06-30
2015-07-31
2015-08-31
2015-09-30
2015-10-31
2015-11-30
2015-12-31
2016-01-31
2016-02-29
2016-03-31
2016-04-30
2016-05-31
2016-06-30
2016-07-31
2016-08-31
2016-09-30
2016-10-31
2016-11-30
2016-12-31
2017-01-31
2017-02-28
2017-03-31
2017-04-30
2017-05-31
2017-06-30
2017-07-31
2017-08-31
2017-09-30
2017-10-31
2017-11-30
2017-12-31
2018-01-31
2018-02-28
2018-03-31
2018-04-30
2018-05-31
2018-06-30
2018-07-31
2018-08-31
2018-09-30
2018-10-31
2018-11-30
2018-12-31


Cancellation/Non Cancellation Numbers by month for each year

Time

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
41	21.421	25.215	48.363	52.949	56.942	60.561	61.665	64.950	68.053	71.938	74.745
42	22.138	25.999	49.456	54.090	58.124	61.777	62.892	66.206	69.336	73.254	76.084
43	22.859	26.785	50.548	55.230	59.304	62.990	64.116	67.459	70.616	74.566	77.419
44	23.584	27.575	51.639	56.369	60.481	64.201	65.337	68.710	71.893	75.874	78.750
45	24.311	28.366	52.729	57.505	61.656	65.410	66.555	69.957	73.166	77.179	80.077
46	25.041	29.160	53.818	58.641	62.830	66.617	67.771	71.201	74.437	78.481	81.400
47	25.775	29.956	54.906	59.774	64.001	67.821	68.985	72.443	75.704	79.780	82.720
48	26.511	30.755	55.993	60.907	65.171	69.023	70.197	73.683	76.969	81.075	84.037
49	27.249	31.555	57.079	62.038	66.339	70.222	71.406	74.919	78.231	82.367	85.351
50	27.991	32.357	58.164	63.167	67.505	71.420	72.613	76.154	79.490	83.657	86.661
51	28.735	33.162	59.248	64.295	68.669	72.616	73.818	77.386	80.747	84.943	87.968
52	29.481	33.968	60.332	65.422	69.832	73.810	75.021	78.616	82.001	86.227	89.272
53	30.230	34.776	61.414	66.548	70.993	75.002	76.223	79.843	83.253	87.507	90.573
54	30.981	35.586	62.496	67.673	72.153	76.192	77.422	81.069	84.502	88.786	91.872
55	31.735	36.398	63.577	68.796	73.311	77.380	78.619	82.292	85.749	90.061	93.168
56	32.490	37.212	64.658	69.919	74.468	78.567	79.815	83.513	86.994	91.335	94.461
57	33.248	38.027	65.737	71.040	75.624	79.752	81.009	84.733	88.236	92.605	95.751
58	34.008	38.844	66.816	72.160	76.778	80.936	82.201	85.950	89.477	93.874	97.039
59	34.770	39.662	67.894	73.279	77.931	82.117	83.391	87.166	90.715	95.140	98.324
60	35.534	40.482	68.972	74.397	79.082	83.298	84.580	88.379	91.952	96.404	99.607

In [21]:

```
# Rename the columns
df.columns = ["observed","expected"]
df.head()
```

Out[21]:

	observed	expected
January 2015	1250	752
March 2015	1759	752
April 2015	585	752
May 2015	919	752
June 2015	886	752

In [19]:

```
# With four rows, the degree of freedom is 47-1 = 46
# With a p-value of 0.05, the confidence level is 1.00-0.05 = 0.95.
critical_value = st.chi2.ppf(q = 0.95, df = 46)
critical_value
```

Out[19]: 62.829620411408165

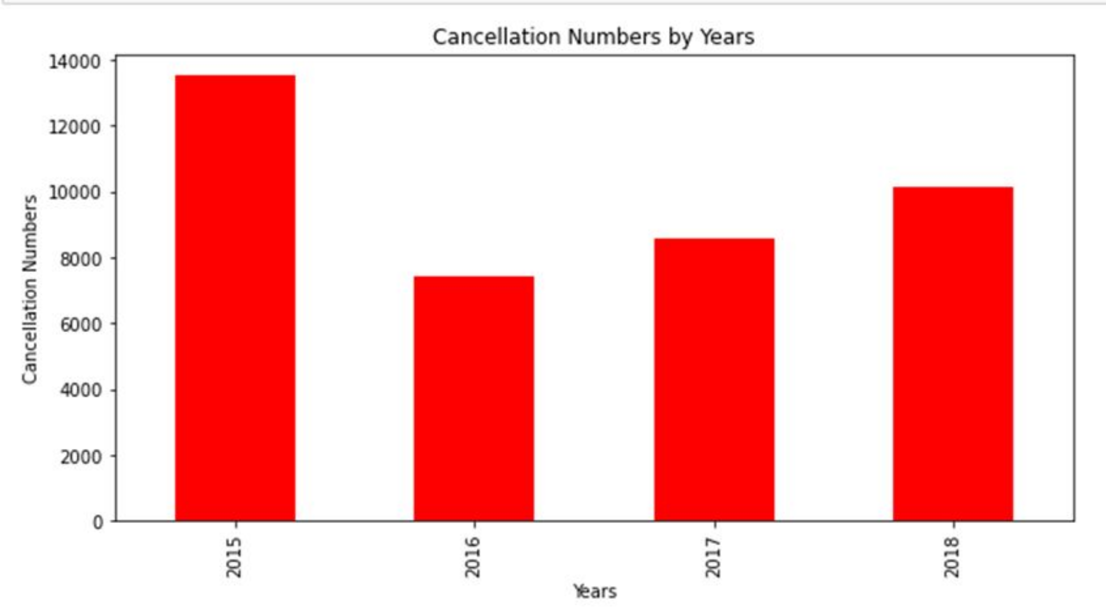
In [22]:

```
# Run the chi square test
st.chisquare(df['observed'], df['expected'])
```

Out[22]: Power_divergenceResult(statistic=12603.594414893618, pvalue=0.0)

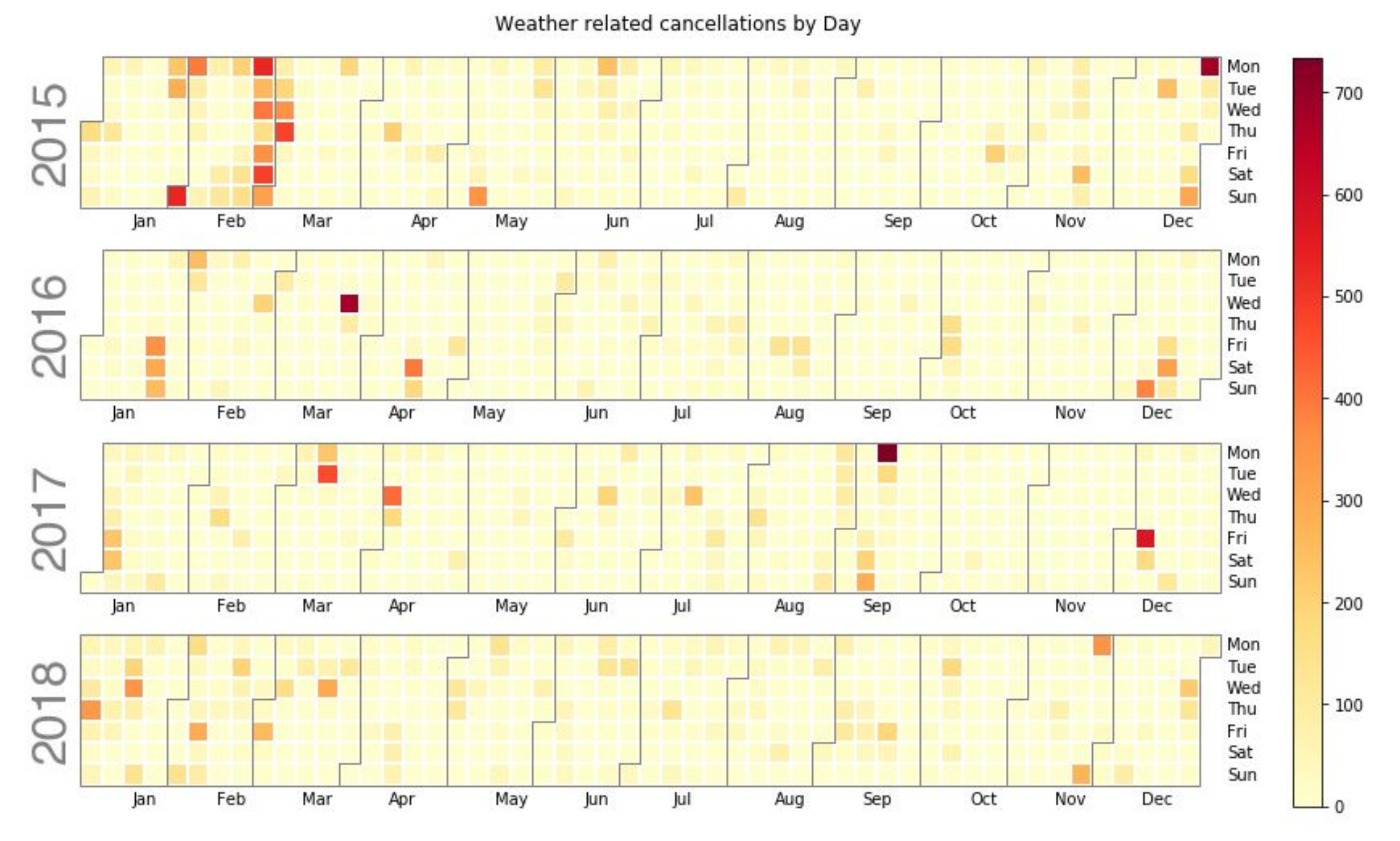
Since the chi-square value of 12603.59 at a confidence level of 95% exceeds the critical value of 62.83, we conclude that the differences seen in the number of cancellations by months of year are statistically significant.

Time



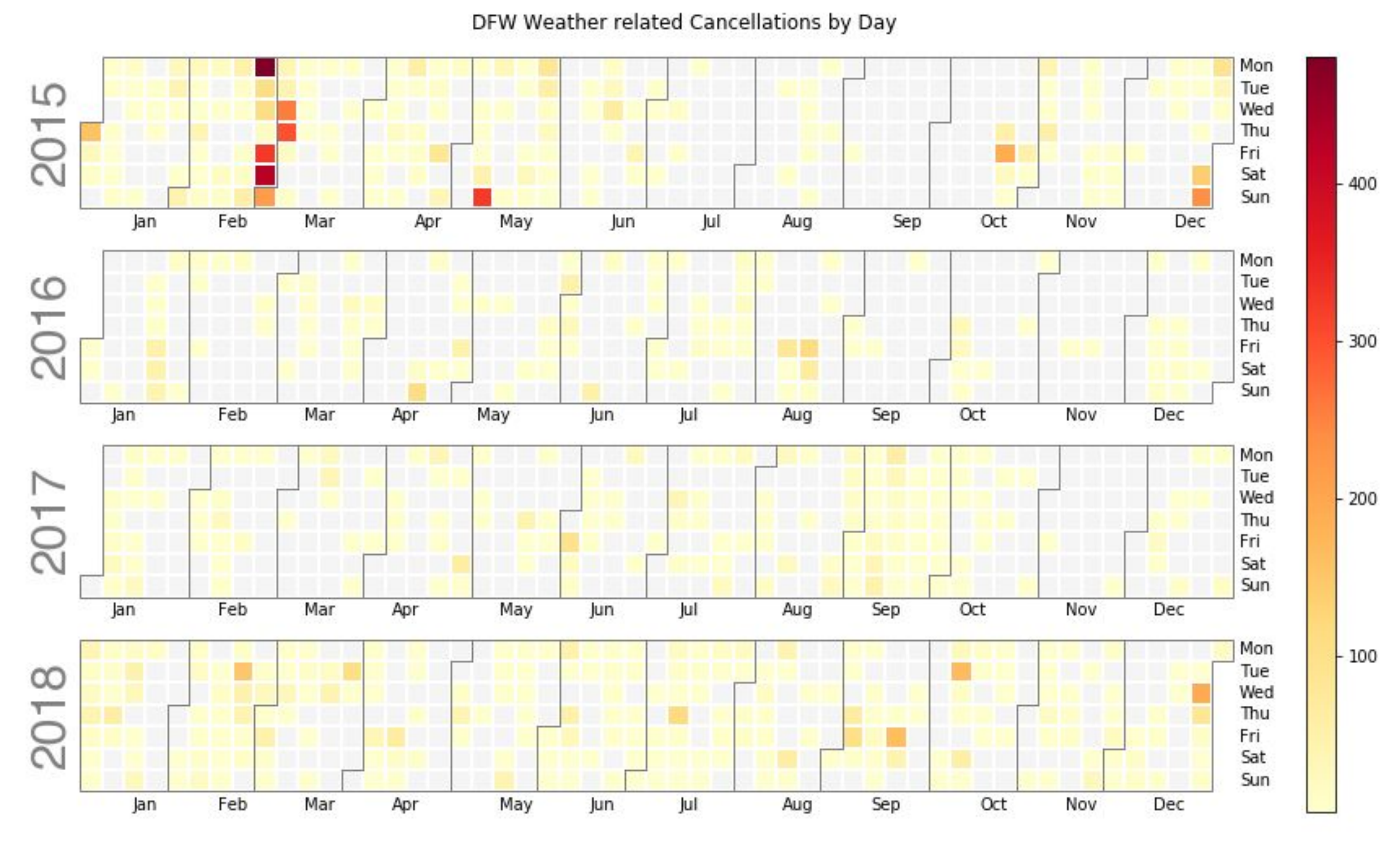
Cancellation Numbers by Years
14000
12000
10000
8000
6000
4000
2000
0
Cancellation Numbers
2015
2016
2017
2018
Years

Time

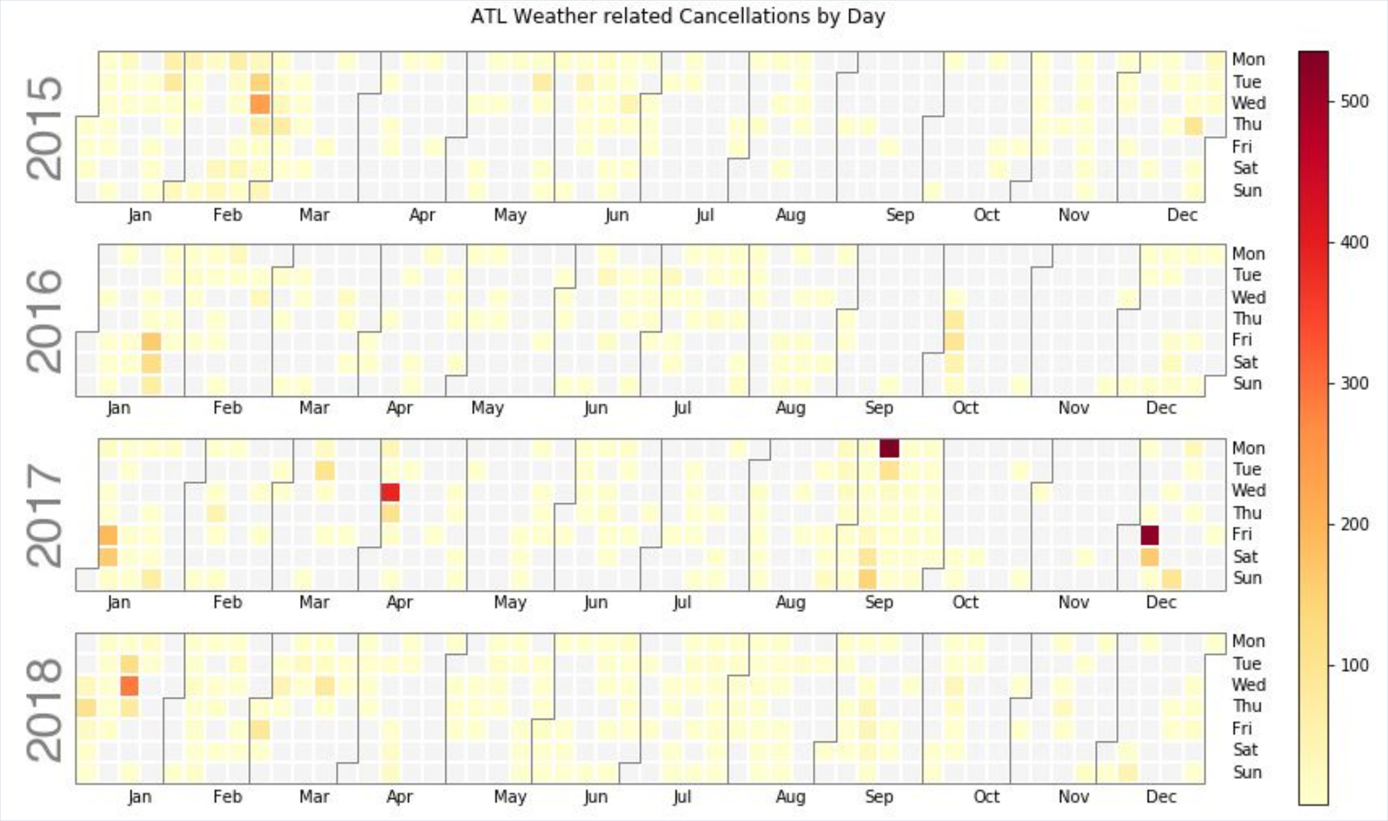


Weather related cancellations by Day
2015
2016
2017
2018
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0
100
200
300
400
500
600
700

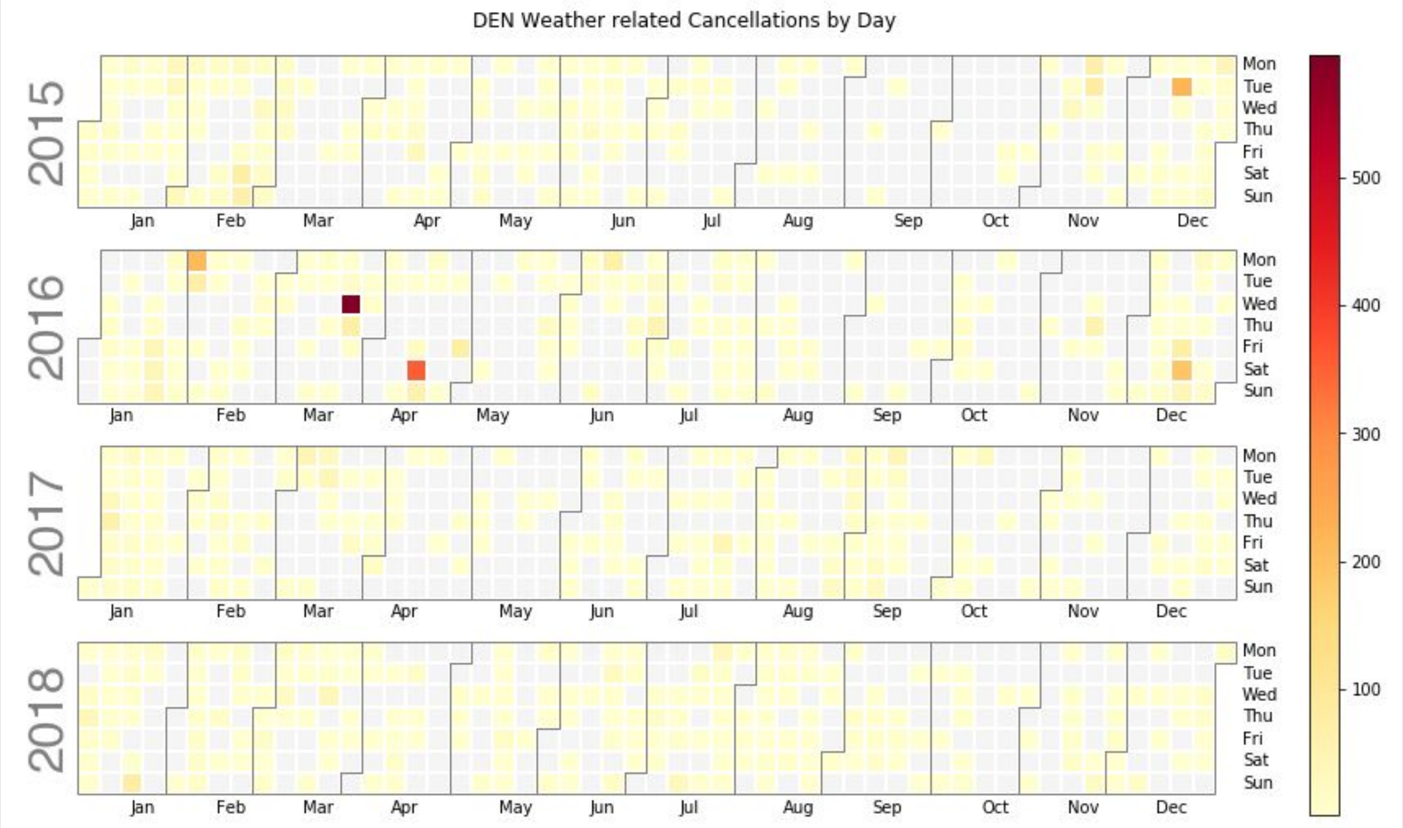
Time


DFW Weather related Cancellations by Day
2015
2016
2017
2018
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
400
300
200
100

Time

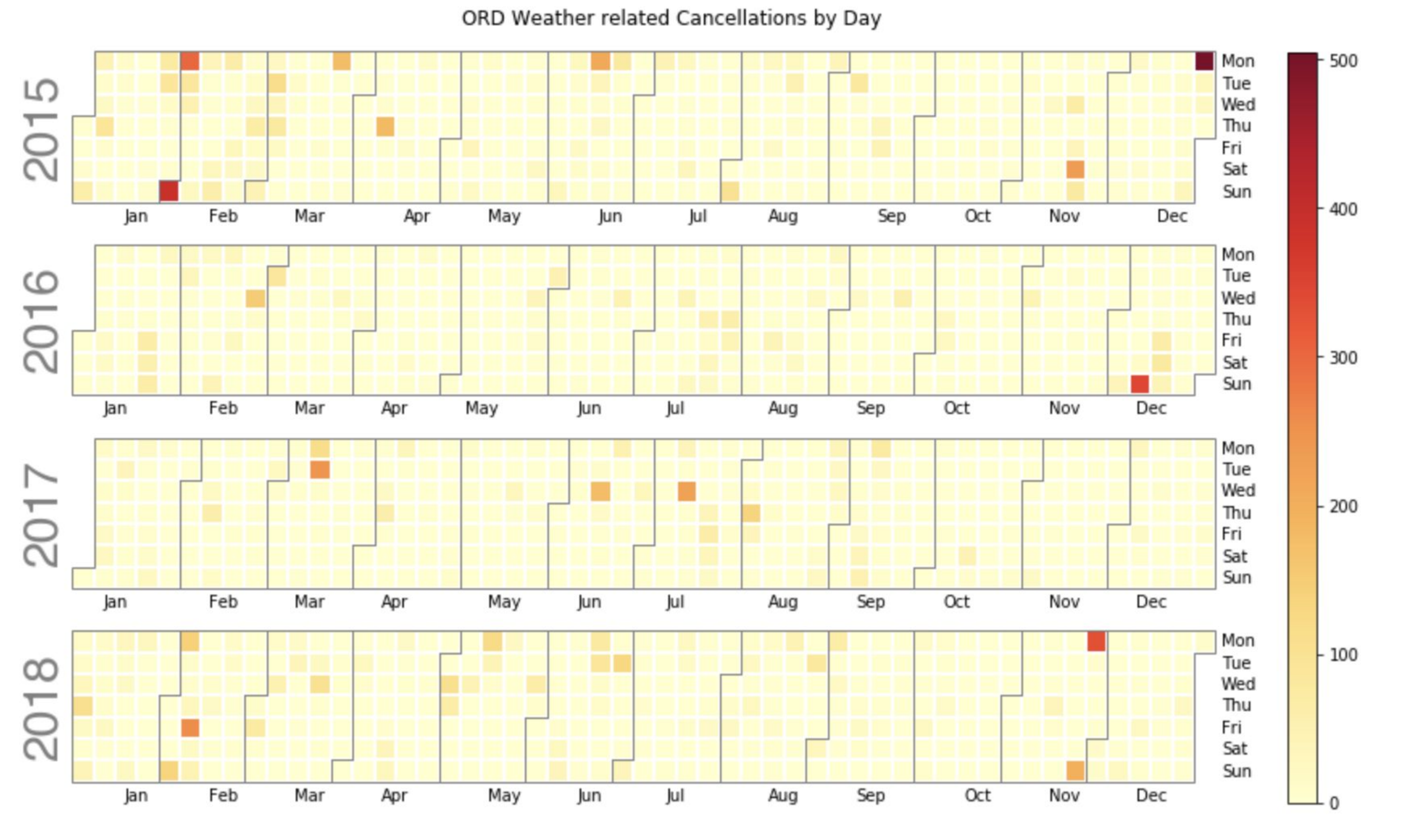

ATL Weather related Cancellations by Day
2015
2016
2017
2018
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
500
400
300
200
100

Time

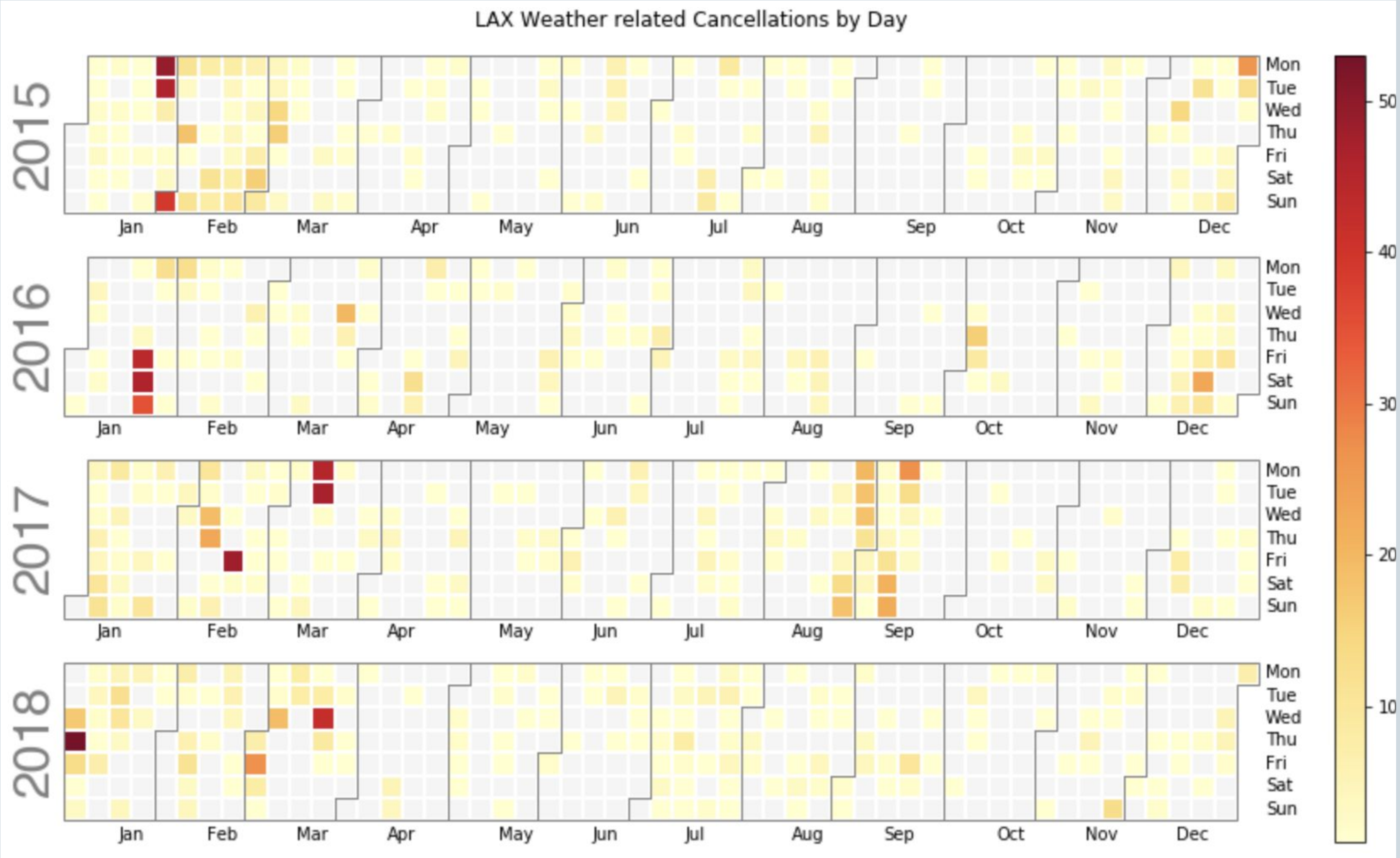


DEN Weather related Cancellations by Day
2015
2016
2017
2018
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Mon
Tue
Wed
Thu
Fri
Sat
Sun
100
200
300
400
500

Time


ORD Weather related Cancellations by Day
2015
2016
2017
2018
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Mon
Tue
Wed
Thu
Fri
Sat
Sun
500
400
300
200
100
0

Time


LAX Weather related Cancellations by Day
2015
2016
2017
2018
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Mon Tue Wed Thu Fri Sat Sun
50
40
30
20
10

OUR CONCLUSIONS

CONCLUSIONS

Our target variable was flight cancellations caused by weather. Our predictor variables are average temperature, wind speed, and precipitation. Our findings:

- How does month of year impact flight cancellations in the U.S.?
  - We reject the null hypothesis, weather factors impact flight cancellations by month of year
- How does wind speed affect U.S. flight cancellations by month of year?
  - We fail to reject the null hypothesis, wind speed will result in no impact on flight cancellations
- How does temperature affect U.S. flight cancellations by month of year?
  - We reject the null hypothesis, temperature impacts flight cancellations by month of year
- How does precipitation type affect U.S. flight cancellations by month of year?
  - We fail to reject the null hypothesis, precipitation will result in no impact on flight cancellations

IMPLICATIONS

- Weather cancellation factors are not mutually exclusive, there can be many weather factors impacting a single cancelled flight. Therefore, we are not able to draw strong correlations to across all observations to a single weather factor, however we did see a significant relationship of weather related cancellations to month and temperature.

- Our observations lead us to fail to reject the null hypothesis specific to wind speed and precipitation factors as they relate to the sample data for canceled flights by weather from top 5 U.S. airports during 2015 - 2018.

- These observations lead us to reject the null hypothesis for weather impact flight cancellations by month of year and average temperature based on our statistical analysis of data from the sample airports for the years 2015-2018.

- Based on these takeaways we uncovered while analyzing the data, considerations to continue this investigation would include an analysis of seasonality specific to weather related flight cancellations as well as a broader scope of locations such as regions or climates.

QUESTIONS?

Thank you!