DATA EXPLORATION

Flight Data

In order to analyze flight cancellations, a fortunately rare occurrence, we would need information about LOTS of flights. Thankfully, a very large free dataset was found on kaggle.com.

At over 2GBs, the dataset appeared to have every domestic US flight from 2009 - 2018, and had all the information that one could reasonably expect such as expected time of departure/landing, cancellation reason, delay times, etc. There were, however, two issues with this dataset.

The first issue is that while the data has cancellation codes that determine between weather, airline issues, security delays, etc. there is no way to tell which airport caused the cancellation. That is to say, if a flight was weather-cancelled, one could not determine if it was cancelled because there was a storm at the origin airport, or the destination airport. For this reason, we would not be able to link each flight to a specific weather palette.

The second issue has to do with the size of the dataset. As a team, we needed to be able to share our clean filtered data through GitHub. This is not possible with 2GBs. In addition, making api calls for tens of millions of rows of data was not an option.

To answer both of these issues required us to greatly change our strategy. Instead of attempting to link each flight to a particular weather palette, we instead would focus on USA's 5 largest airports, and group flights and weather palettes by month for the latest 4 years of the dataset. The filtered data focused on the 5 largest airports parsed by year was still a large set of data with millions of rows, but it was just small enough that it could be shared on GitHub.

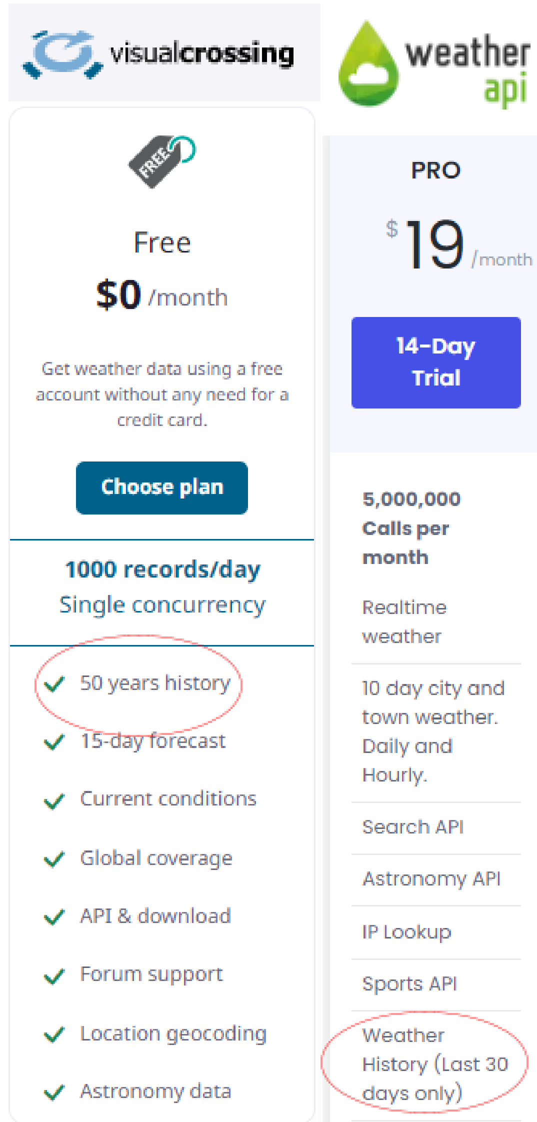
Location Data

When the project was first started, it was thought that we would use data from hundreds of airports across the United States, and that all we had was 3-letter codes. Those 3-letter airport codes would need to be translated into coordinates for calling weather data. To this end, a dataset of airport location data was retrieved from www.partow.net. GAD to csv.ipynb takes care to translate the data from a text file to a csv. This data was then used to manually create a new easy-to-read Airport_Data.csv using Microsoft Excel. Ultimately though, because we decided to only go with the 5 largest airports, this whole process could have been done in a few google maps searches.

Weather Data

For the weather data, we had a few API options. At first, we believed that weatherapi.com would be the way to go because the entire team had experience with this API. We assumed that we would have no problems getting the data we needed and went about getting all of the flight and location data tidied up and ready to get weather data for every single flight. A few days into the project however, one team member noticed that weatherapi.com's free version does not allow calls for historical weather data, only current data. On top of that, for the amount of flights we had data for, even a professional account didn't have that many API calls.

On the other hand, visualcrossing allows for historical weather data for free. As described above, the team's strategy changed a few days in and we decided to get information for days rather than each individual flight. For this amount of data, the team was able to pool our free API calls together and get all of the needed information.



SUMMARY TABLES

After our data exploration and cleaning, we merged weather data from **VISUAL CROSSING WEATHER | API** and flight data from **Airline Delay and Cancellation Data, 2009 - 2018 | Kaggle** into a new dataframe on the columns of "Date" and "Airport". This allowed us to align the weather specific to the airport and flight cancellation date.

```python
#Combine all flight data with weather data by day
all_data = pd.merge(new_summary, weather_data, on=["Date", "Airport"])
all_data
```

	Date	Airport	Destination	Expected Departure Time	Expected Arrival Time	Distance	Weather Delay	Latitude	Longitude	Max Temp	Precip	Precip Type	Wind Speed
0	2015-01-01	DFW	BWI	1342	1724	1217.0	0	32.896	-97.037	36.0	0.58	rain,snow,	8.0
1	2015-01-01	DFW	SAN	839	952	1171.0	0	32.896	-97.037	36.0	0.58	rain,snow,	8.0
2	2015-01-01	DFW	ATL	731	1032	731.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0
3	2015-01-01	DFW	MCI	1951	2117	460.0	0	32.896	-97.037	36.0	0.58	rain,snow,	8.0
4	2015-01-01	DFW	RSW	1020	1348	1017.0	0	32.896	-97.037	36.0	0.58	rain,snow,	8.0

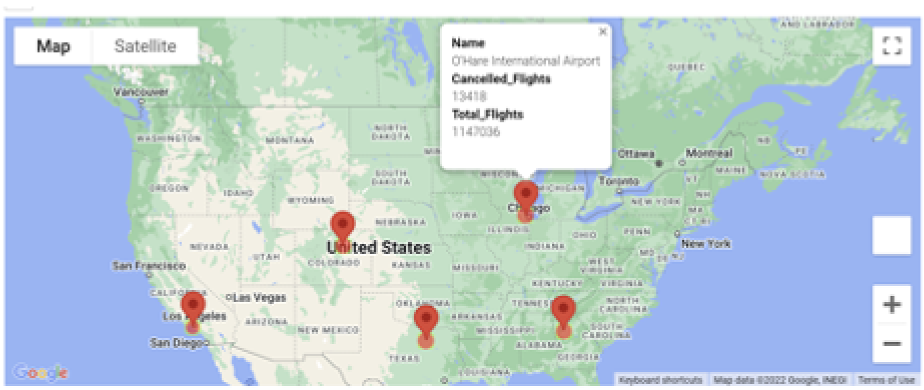
Date	Airport	Destination	Expected Departure Time	Expected Arrival Time	Distance	Weather Delay	Latitude	Longitude	Max Temp	Precip	Precip Type	Wind Speed	Month	Year	Month Year
2015-01-01	DFW	ATL	731	1032	731.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0	January	2015	January 2015
2015-01-01	DFW	TPA	818	1134	929.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0	January	2015	January 2015
2015-01-01	DFW	MSN	830	1038	821.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0	January	2015	January 2015
2015-01-01	DFW	LRD	1255	1418	396.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0	January	2015	January 2015
2015-01-01	DFW	DEN	1704	1813	641.0	CANCELLED	32.896	-97.037	36.0	0.58	rain,snow,	8.0	January	2015	January 2015

	Date	Expected Departure Time	Expected Arrival Time	Distance	Latitude	Longitude	Max Temp	Precip	Wind Speed	Weather Delay
0	2015-01-31	1462.730400	1624.560000	815.451200	37.330344	-94.808253	41.860480	0.130328	15.828560	1250
1	2015-02-28	1444.616387	1614.942970	700.278166	36.233835	-93.590573	34.982379	0.558464	17.933264	4296
2	2015-03-31	1360.388857	1536.847641	659.683911	36.074900	-94.303525	41.971063	0.215969	18.882661	1759
3	2015-04-30	1527.305983	1673.449573	547.136752	37.343065	-94.701993	68.020684	0.797641	22.124957	585
4	2015-05-31	1602.527748	1719.332971	601.388466	34.606746	-95.210214	76.612514	0.829902	21.383025	919
5	2015-06-30	1628.681716	1712.124153	583.067720	38.650205	-91.337823	81.573702	0.993758	16.423025	886
6	2015-07-31	1722.312757	1748.345679	511.637860	39.207798	-95.472481	84.199588	0.280864	16.882305	243
7	2015-08-31	1656.808451	1750.912676	501.002817	40.420746	-90.249144	86.053521	0.177803	17.006197	355
8	2015-09-30	1738.096257	1829.973262	372.219251	41.265594	-88.596246	80.624064	0.935508	17.134225	187
9	2015-10-31	1651.320755	1716.765499	673.444744	33.603523	-97.217879	73.540701	2.633342	18.529380	371
42	2018-07-31	1615.870968	1732.137097	721.916129	36.761398	-96.880968	88.394839	0.116871	18.877581	620
43	2018-08-31	1666.337456	1769.846290	790.749117	38.130850	-93.357804	87.478445	0.436290	17.378445	566
44	2018-09-30	1603.845606	1669.426366	749.273159	35.578025	-94.385004	85.537292	1.234050	16.119834	842
45	2018-10-31	1552.825630	1624.525210	618.838235	34.893731	-94.836017	70.778782	1.489349	16.969118	476
46	2018-11-30	1465.868757	1555.613240	724.054588	39.850828	-91.247846	42.283391	0.198118	29.009175	861
47	2018-12-31	1569.133433	1613.853073	644.142429	35.540753	-95.363150	53.434933	0.937181	19.765817	667

LOCATION

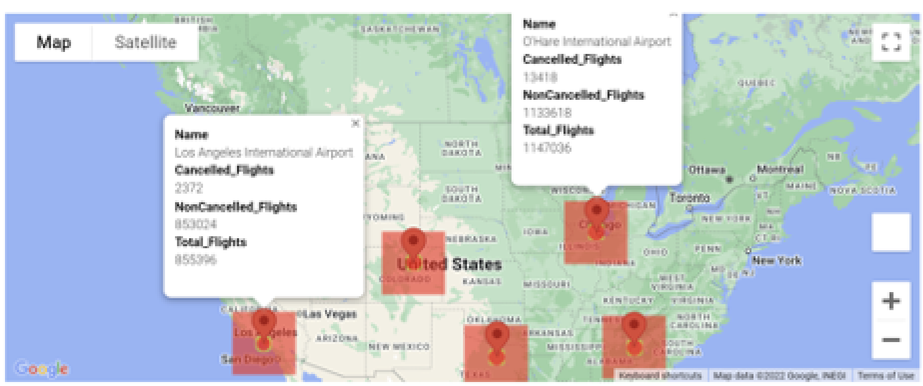
Heatmap Weighted by Cancellation Numbers

The map presented below shows the number of cancellations in each airport.



Heatmap Weighted by Canceled and Non-Cancelled Numbers

The map presented below shows the number of canceled,non-cancelled and total flights in each airport.The inner circle at each airport is weighted by the total canceled flights and the outer rectangle is weighted by the total non-cancelled flights.



Flight Cancellations by Airport

The sample data below reflects the percentage of flight cancellations caused by weather factors for five major US airports. The percentage of cancellations is greatest in Chicago (ORD) compared to other airports at 33.8% of all sample data retrieved.

Total Canceled Flights:

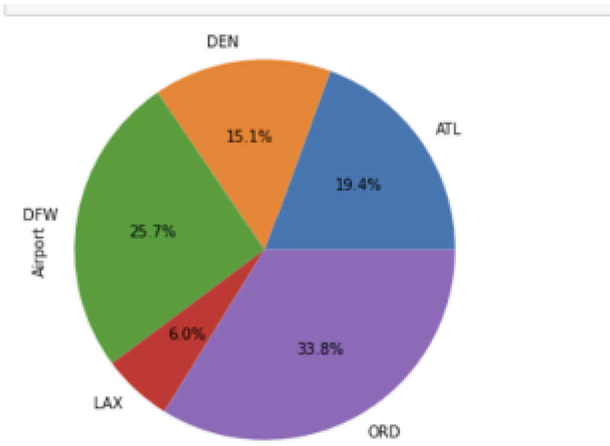
ORD: 13,418 Flight Cancellations

LAX: 2,372 Flight Cancellations

DFW: 10,181 Flight Cancellations

DEN: 5,989 Flight Cancellations

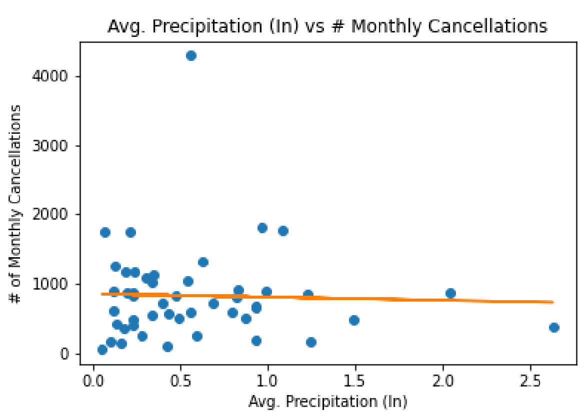
ATL: 7,701 Flight Cancellations



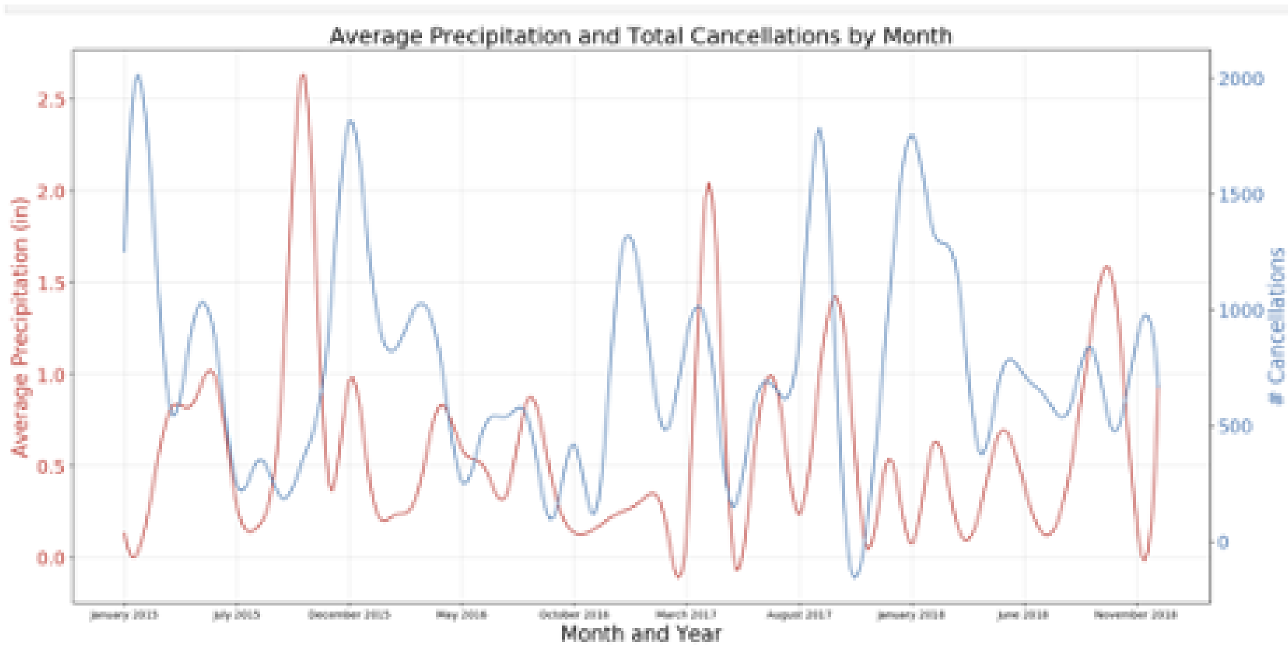
PRECIPITATION

Avg. Precipitation (in) by # Monthly Cancellations

- Visibility is a critical factor in flight safety. If the precipitation is too heavy, then the pilot's visibility can be impaired. When the amount of precipitation (in) is too high, it is too dangerous to takeoff, resulting in a flight cancellation.
- There was an increased # in monthly flight cancellations when the avg. precipitation was low, between 0.0 - 1.0 (in).
- The regressions line and the r-value indicate that precipitation was not a significant factor in the number of monthly flight cancellations. From the sampled data, we fail to reject the null hypothesis. Therefore, precipitation will result in no impact on flight cancellations.
- The correlation between both factors is -0.04
- The r-squared is: 0.001249106831586809



Average Precipitation vs. Total Cancellations by Month



The chart shows the average precipitation in comparison to the total canceled flights by month for the sample data set. The number of cancellations per month from January 2015 to December 2018 shows consistent spikes during the month January across all years. The average precipitation (in) from January 2015 to December 2018 shows a relatively consistent range falling between 0 - 1.0 (in) with the exception of spikes above 2 inches between July - December of 2015 and March to August of 2017.

Cancellations Due to Precipitation

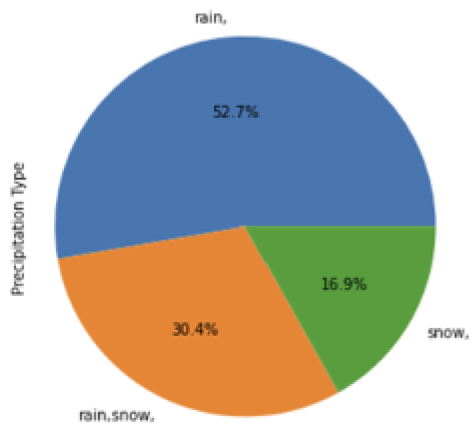
The sample data below reflects the breakdown of precipitation types (Rain, snow, rain/snow) that caused cancellations for five major US airports. The precipitation type that caused the greatest amount of flight cancellations in the sample data was rain at 52.7%, followed by rain/snow at 30.4%.

Precipitation type and the canceled numbers.

Rain: 14,592

Rain + Snow: 8,420

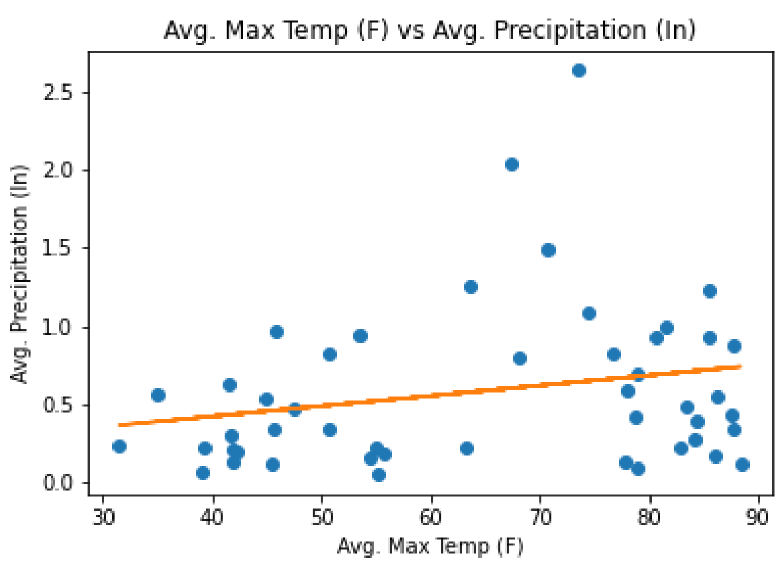
Snow: 4,676



TEMPERATURE

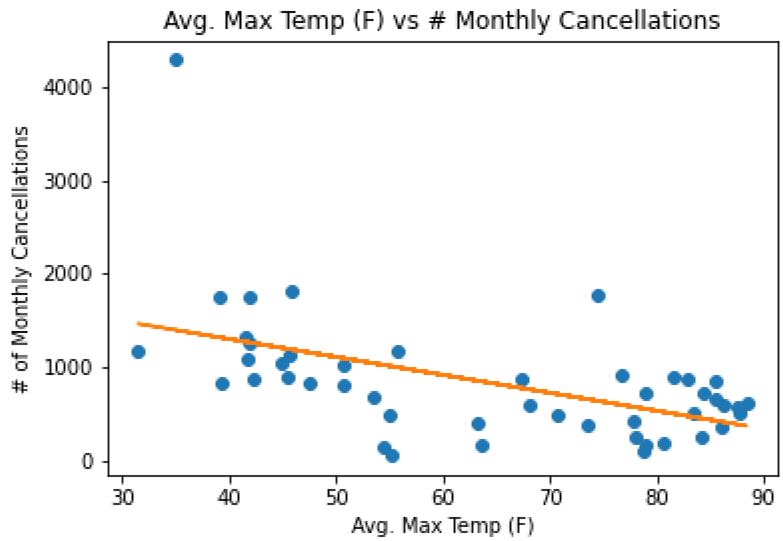
Avg. Max Temp (F) by Avg. Precipitation (In)

- As the average max temperature (F) increased, the average precipitation (In) increased.
- The correlation between both factors is 0.23
- The r-squared is: 0.053597775234979474
- The regressions line and the r-value indicate that there is a relationship between average max temp (F) and avg. precipitation (In).

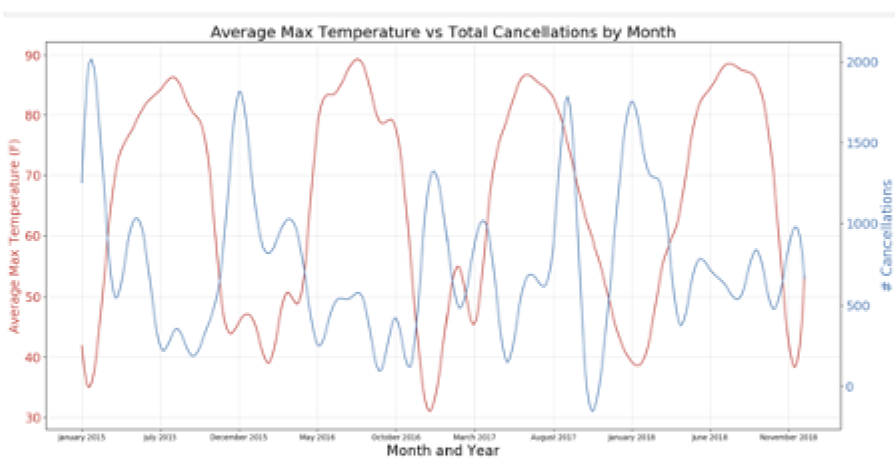


Avg. Max Temp (F) by Monthly Cancellations

- The correlation between both factors is -0.52
- The r-squared is: 0.26932427288672406
- During warmer weather, there were less total canceled flights by month. From the sampled data, we can reject the null hypothesis. Therefore, average maximum temperature will result in an impact on monthly flight cancellations.



Average Temperature vs Total Cancellations by Month



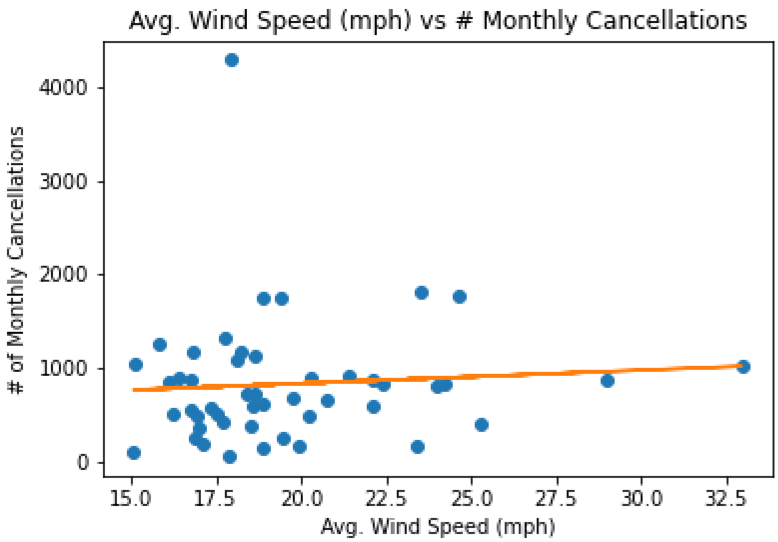
The chart shows the average maximum temperature in comparison to the total canceled flights by month for the sample data set. The number of cancellations per month from January 2015 to December 2018 shows consistent spikes during the month January across all years. The average temp (F) from the sample data set peaks each year in July at approximately ~85 (F).

WIND SPEED

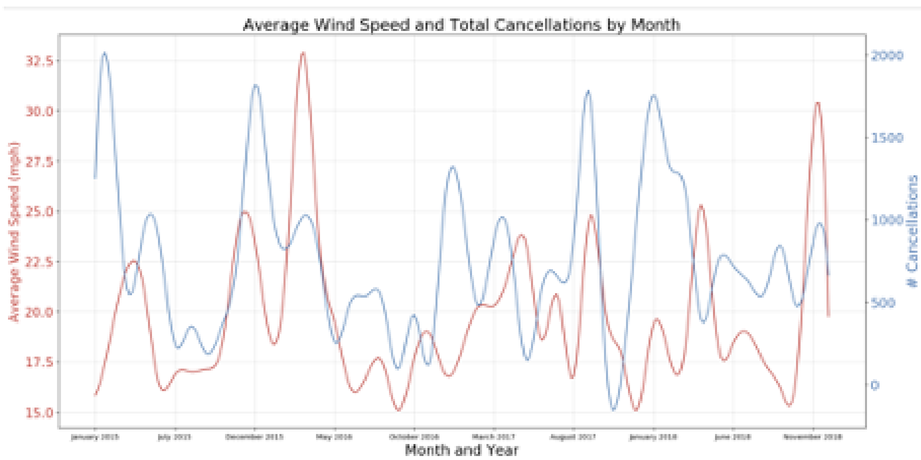
Avg. Wind Speed (mph) by Monthly Cancellations

- There were more months with flight cancellations when the average wind speed was between 15 mph and 20 mph. As wind speed increased, the number of monthly canceled flights decreased.

- The regressions line and the r-value indicate that avg. wind speed was not a significant factor in the number of monthly flight cancellations. From the sampled data, we fail to reject the null hypothesis. Therefore, wind speed will result in no impact on flight cancellations.
- The correlation between both factors is 0.07
- The r-squared is: 0.005518412470719093



Average Wind Speed vs Total Cancellations by Month

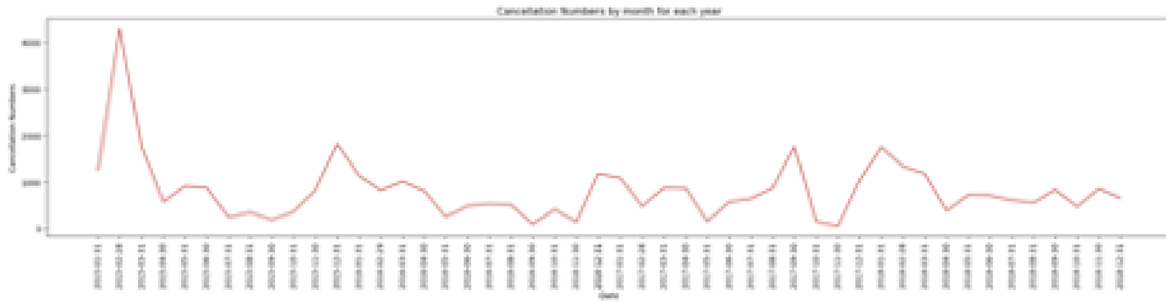


The chart shows the average wind speed (mph) in comparison to the total canceled flights by month for the sample data set. The number of cancellations per month from January 2015 to December 2018 shows consistent spikes during the month January across all years. The average wind speed (mph) from the sample data shows no correlation.

TIME

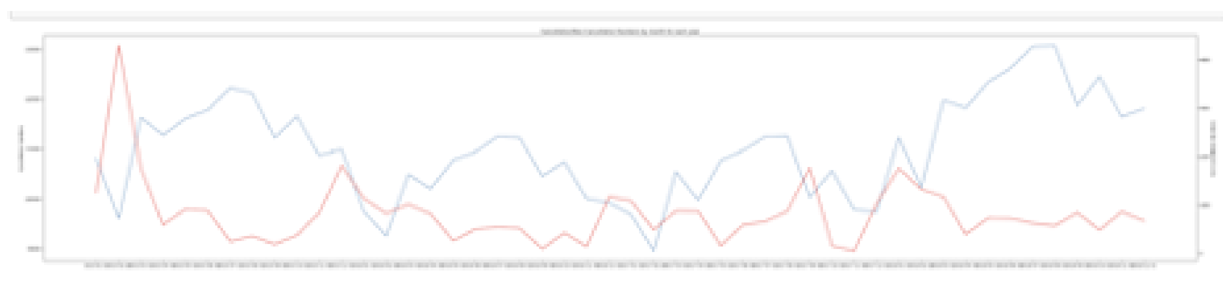
Cancellation Number by Months For Each Year

The below line graph shows the flight cancellation numbers along the months of the years (2015-2018). From the graph,the cancellations seem to be very high in the month of Feb (2015).



Cancellation / Non-Cancellation Numbers by Months For Each Year

Over the course of January 2015 - December 2018, there were a few instances in which canceled flights outnumbered non-cancelled flights. The below graph shows both the trends of cancellation and non-cancellations throughout the months of the year.



Weather Flight Cancellations From 2015 - 2018

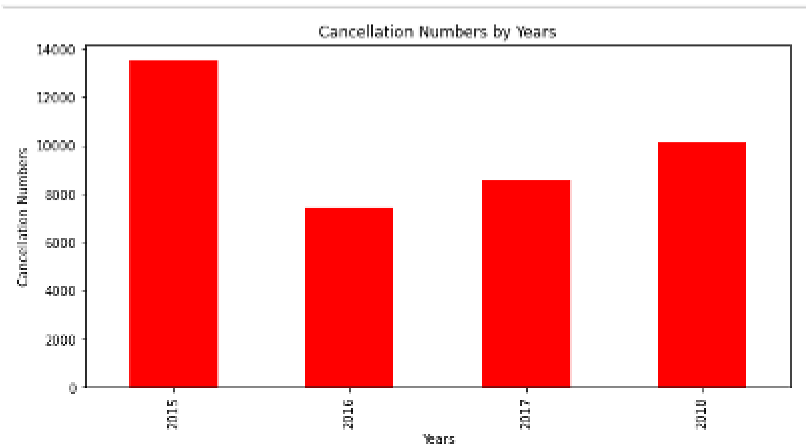
The below bar chart indicates the cancellations across the years of 2015-2018 from the sample data set. 2015 had the highest number of weather-specific flight cancellations at 34% compared to other years.

2015 Canceled Flights: 13,503

2016 Canceled Flights: 7,434

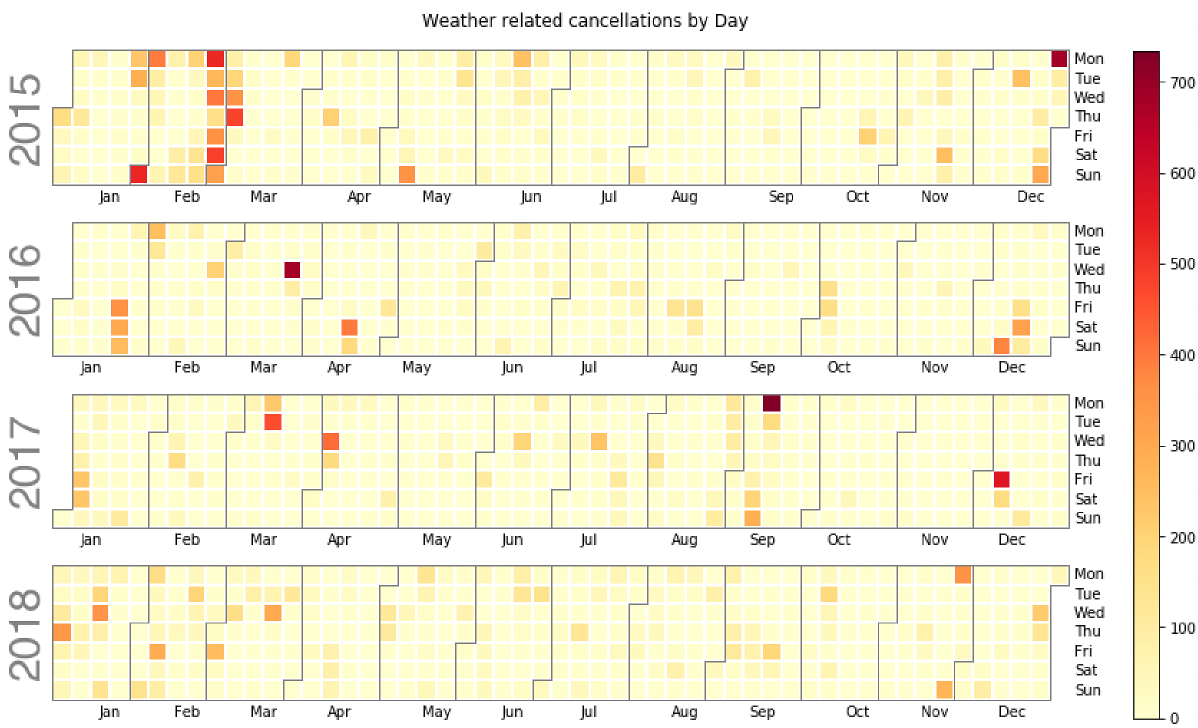
2017 Canceled Flights: 8,591

2018 Canceled Flights: 10,133



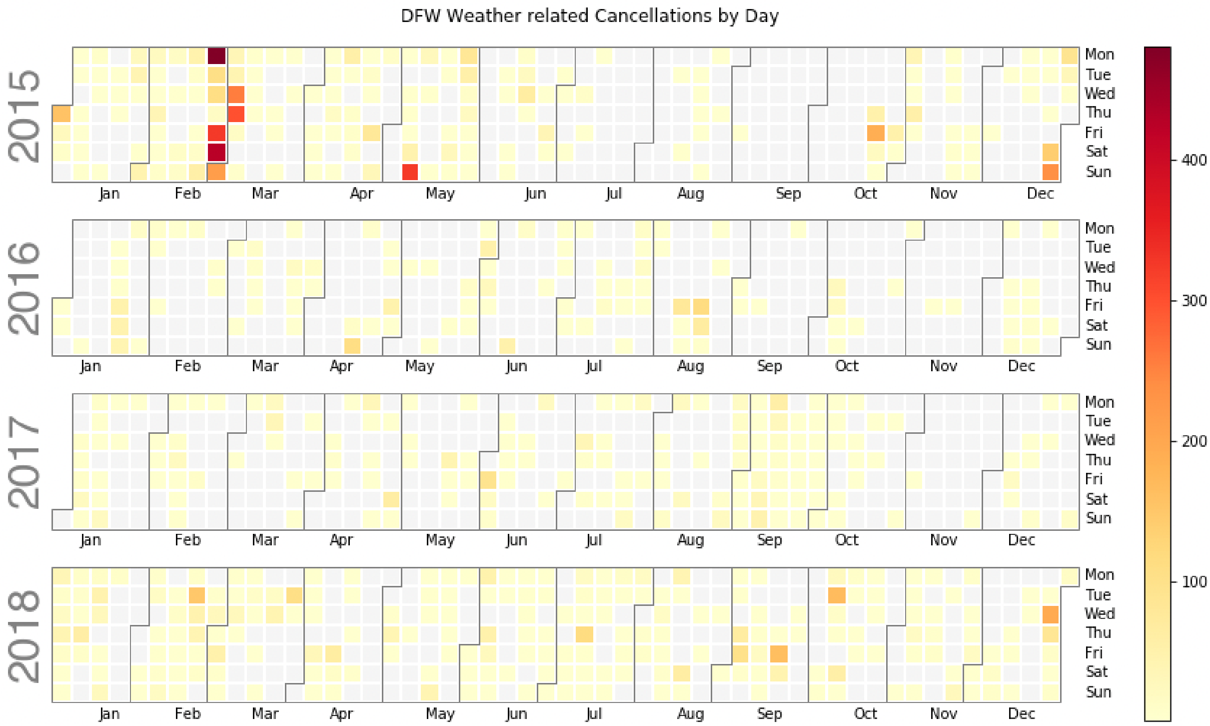
Heatmap Calendar 5 Top US Airport Weather Cancellations by Day

This heat map displays volume of weather related cancellations over the course of a calendar year and broken down by day. Overall, we can see the lower concentrations of cancellations are in the summer, and a higher concentration of cancellations in the earlier part of the year (Jan - Mar). To investigate if there are more weather-related cancellations in different areas, we zoom in on each airport.



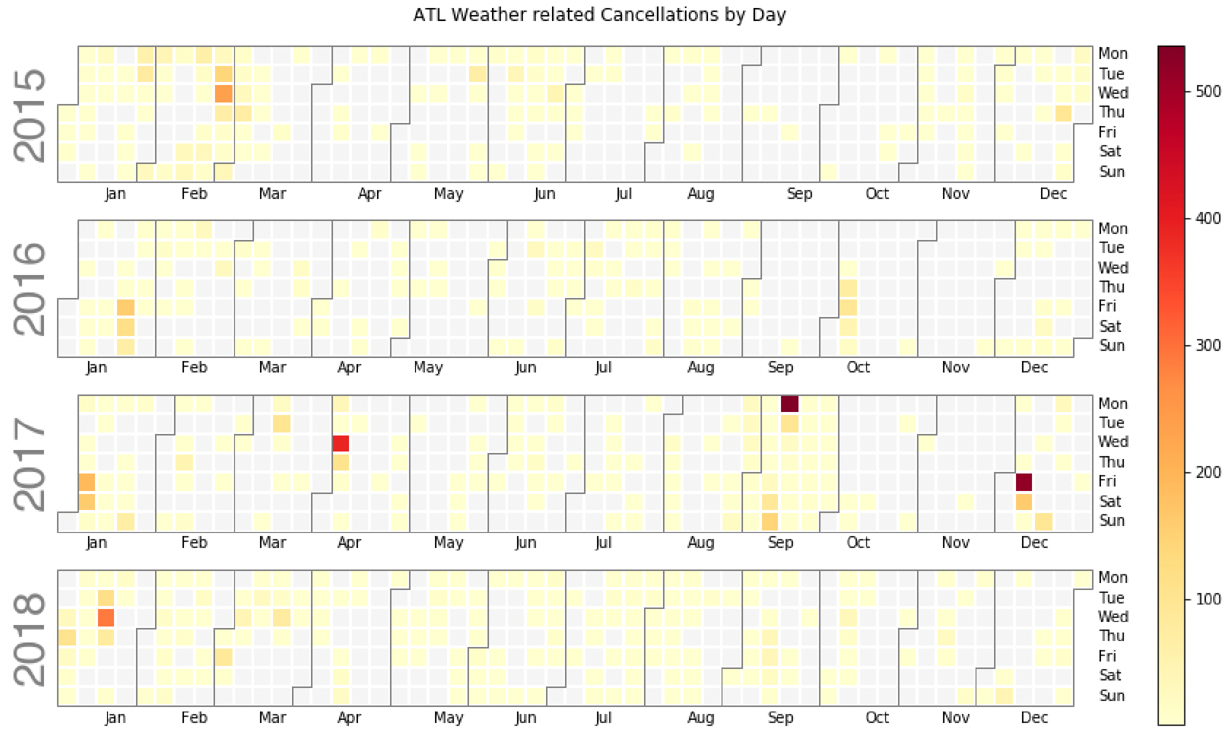
DFW Cancellation Calendar Heatmap

The heatmap calendar for DFW, a centrally located southern airport, we see higher concentrations of weather related cancellations in the winter months (Dec - Feb). Max cancellations for one day at DFW across 2015-2018 is 480.



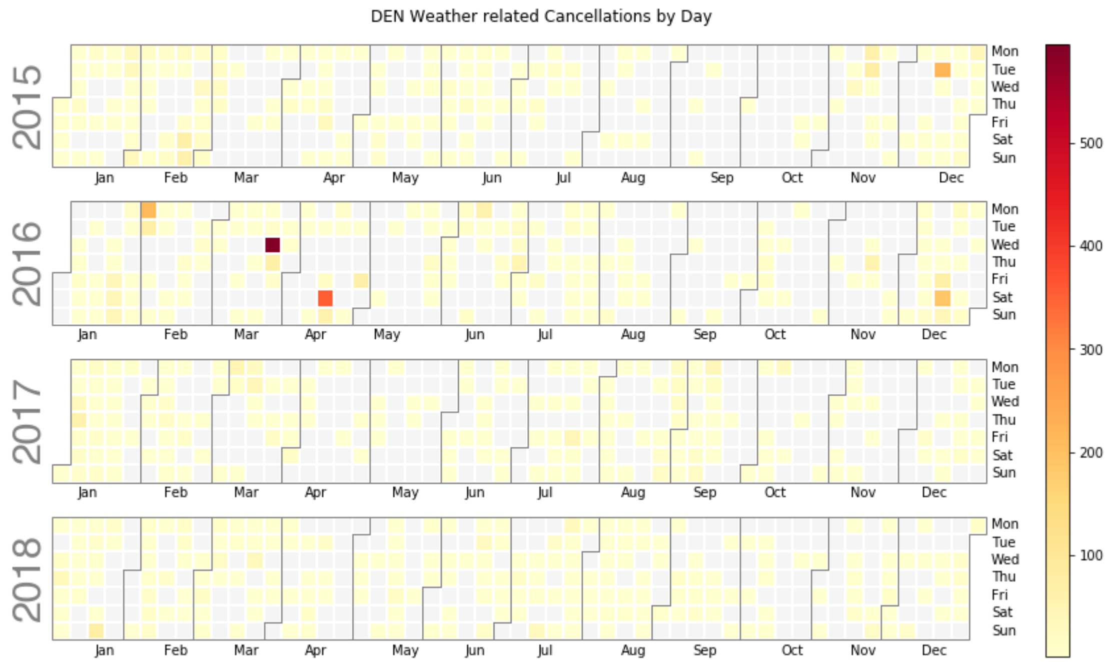
ATL Cancellation Calendar Heatmap

The heatmap calendar for ATL, we can observe concentrations of weather related cancellations in December and January. Max cancellations for one day at ATL across 2015-2018 is 536.



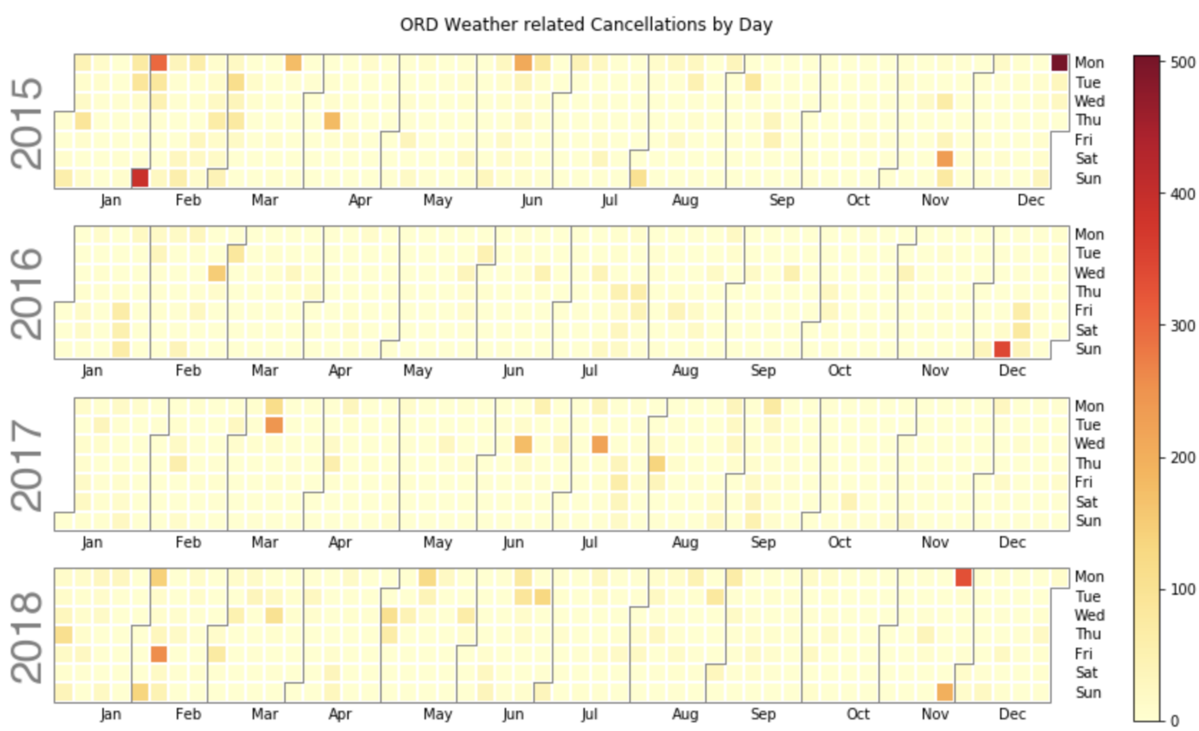
DEN Cancellation Calendar Heatmap

The heatmap calendar for DEN, we can observe concentrations of weather related cancellations in December and January. Max cancellations for one day at DEN across 2015-2018 is 596.



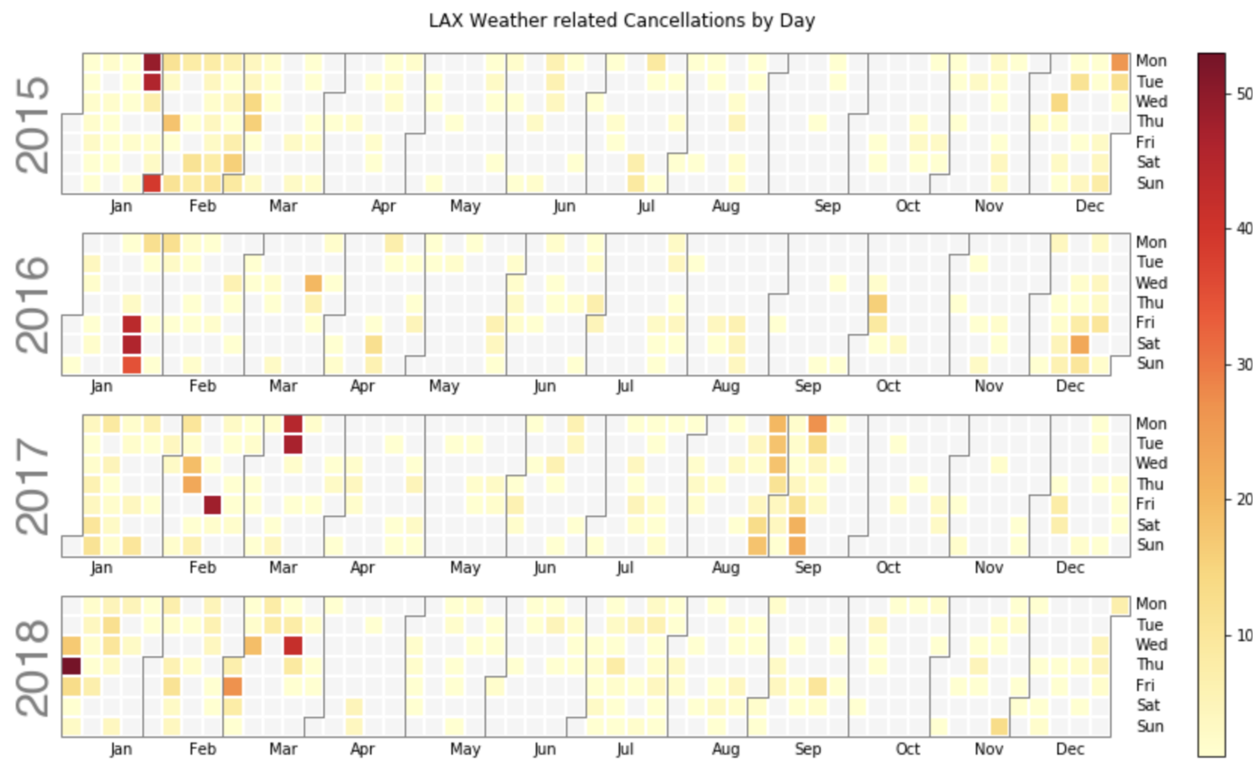
ORD Cancellation Calendar Heatmap

The heatmap calendar for ORD, we can observe concentrations of weather related cancellations between the months of Nov-Jan. Max cancellations for one day at ORD across 2015-2018 is 505.



LAX Cancellation Calendar Heatmap

The heatmap calendar for LAX shows the lowest numbers of weather related cancellations, we can observe concentrations of cancellations between the months of Dec-Mar. However, the max cancellations for one day at LAX across 2015-2018 is 53, showing that LAX experiences the least amount of weather related cancellations by day.



Chi-Square Test on Flight Cancellations from Weather

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.

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.