**Data Science II Project 1st Milestone**

**Dataset: US Weather Events**

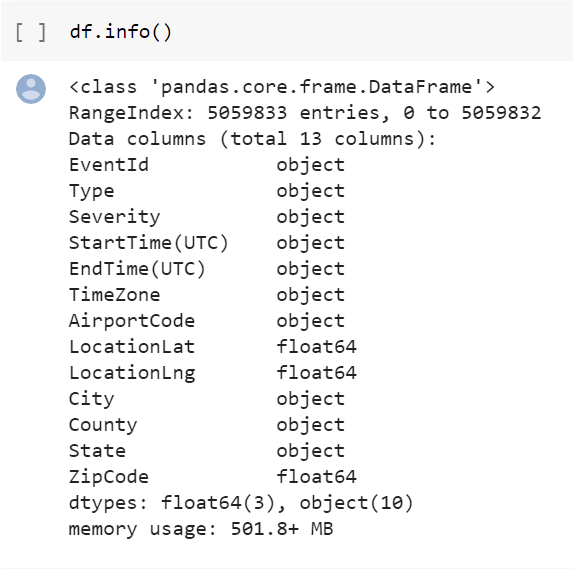
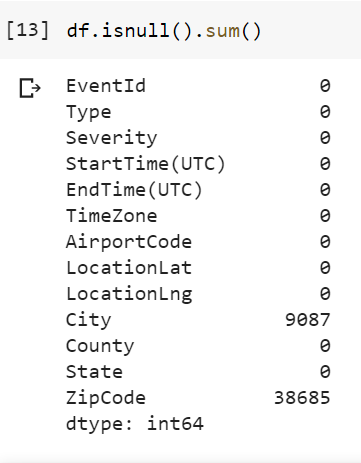
**Deependra Bassnet, Vyas Ramankulangara, Chantha Mak**

**Introduction**

The US Weather Events dataset provides weather information throughout all of the United States. The dataset covers weather data for 48 states and has over 5 million weather events that occurred from January 2016 to December 2019. The weather events that the dataset utilized includes server-cold, fog, hail, rain, snow, storm and other precipitation.

The main objection of this project is to find the weather trend over the span of 4 years from 2016 to 2019 for various cities across the United States. We are hoping to find interesting trends with the provided data.

**Additional Dataset Information**

**Dataset Features and type:**

Our dataset contains 13 columns   with 3 of the columns that are type float and 10 of the columns being object which mean they are type string.

From the second report above, we are seeing that the city has 9087 entries of empty values and zip code with 38685 entries of empty values. We will go over how we will handle the empty values in the next section.

**Feature Description:**

Below are the column(feature) description from the data description.docx given to us

1. Severe-Cold: The case of having extremely low temperature, with temperature below -23.7 degrees Celsius.
2. Fog: The case where there is low visibility condition as a result of fog or haze.
3. Hail: The case of having solid precipitation including ice pellets and hail.
4. Rain: The case of having rain, ranging from light to heavy.
5. Snow: The case of having snow, ranging from light to heavy.
6. Storm: The extremely windy conditions, where the wind speed is at least 60 km/h.
7. Other Precipitation: Any other type of precipitation which cannot be assigned to previously described event types.

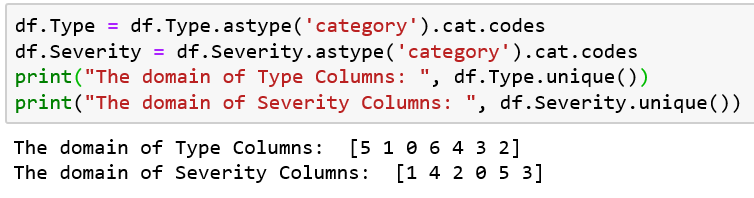
**Data Clean up and Dimensionality Reduction**

The factors that we are looking for while we preprocess the data:

1. Dealing with missing data

Empty values are a problem to machine learning models and as we can see from the data report earlier, we have some empty data in zip code and city column that we need to take care of. Once of the common way to deal with empty values is to fill them up with the mean of the column that contains the empty value but since the empty resides in zip code and city column, it wouldn’t make sense to fill the empty values with the mean value of the column since the zip code and city are location based attributes. Alternatively, we will remove the empty values and it will not negatively impact the dataset that much since the empty values that we encountered are just a small fraction of the 5 million total entries.

1. Dealing/Converting Categorical to Nominal

In our dataset, we have 2 categorical attributes that needed to be changed to nominal attributes since machine learning model works best with nominal attributes. In order to do so, we will be converting the domain of the categorical attribute to integers using numerical encoding. Below is the result of the encoding.

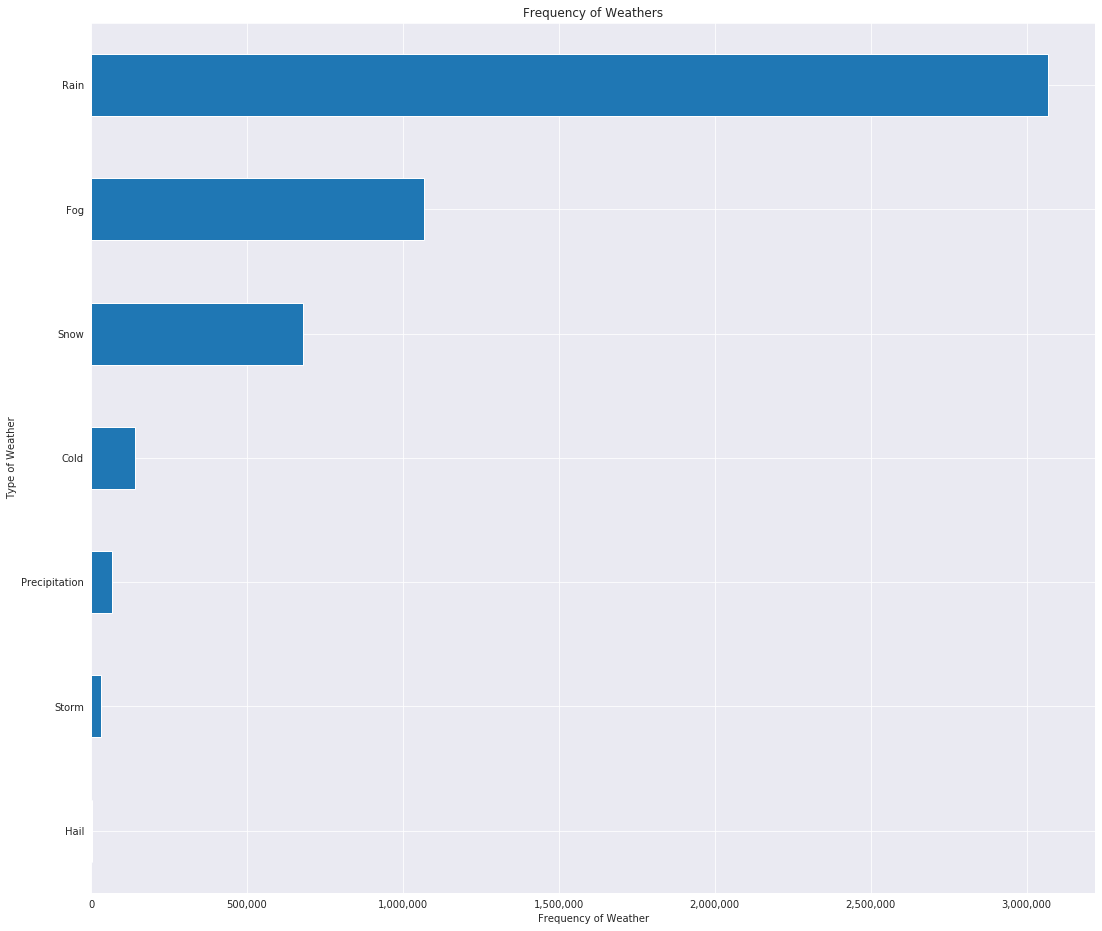
1. Dropping Unnecessary/Irrelevant Features

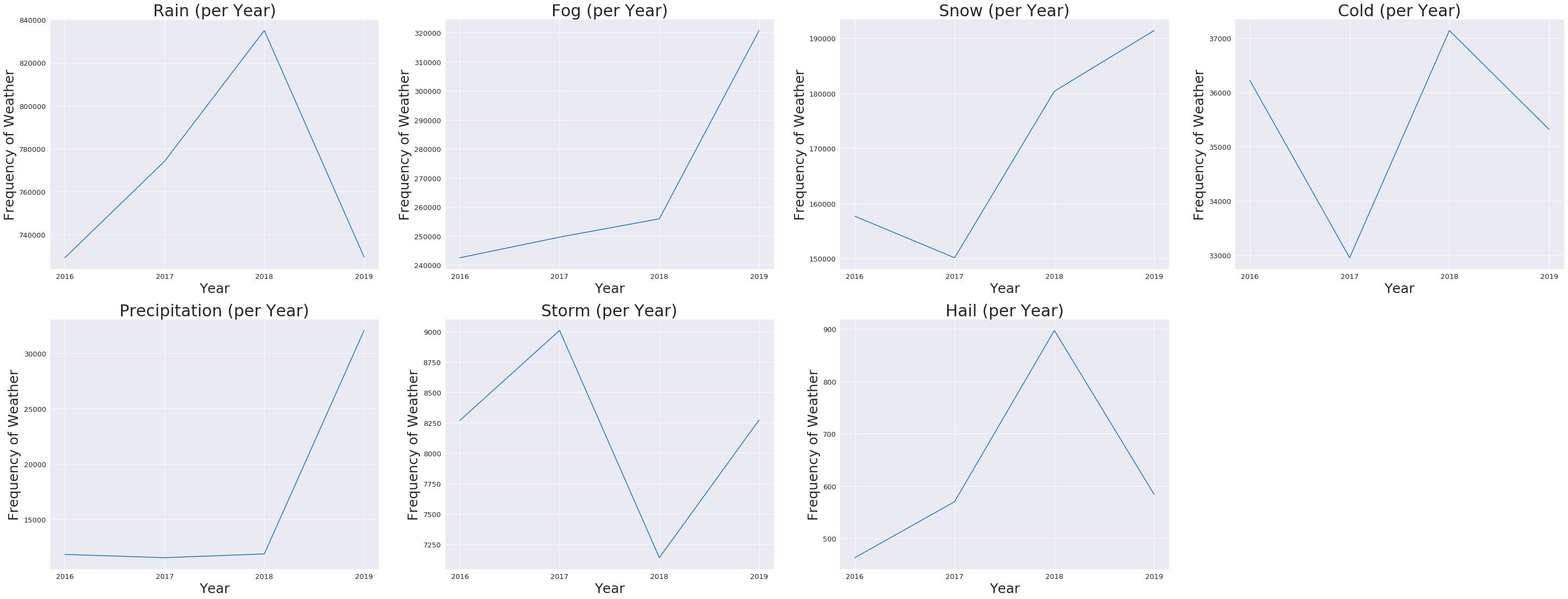
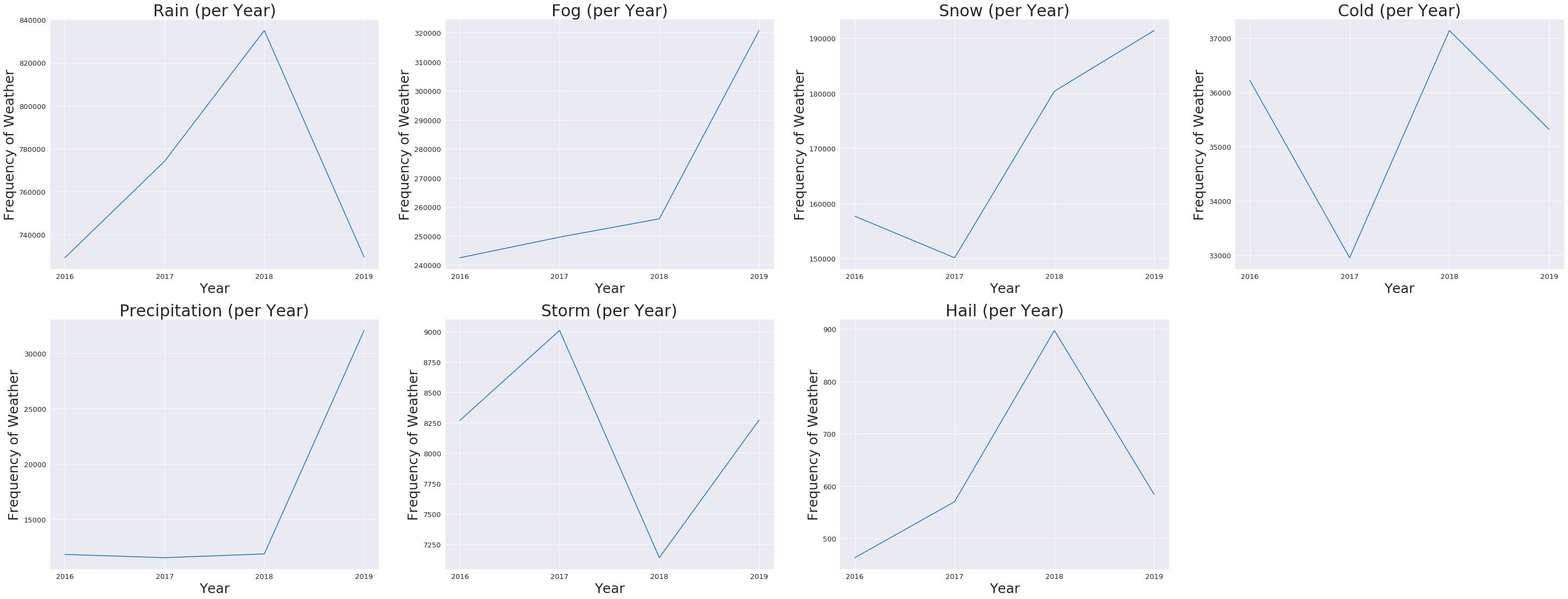
Having a lot of features increases the chance of overfitting since the model will more complex. In this case, we will be dropping irrelevant/unnecessary features. In this US weather dataset, we have a few features that would not contribute or help out with our model or prediction. We will be dropping the zip code and city since these two features have a lot of missing values in its field. Additionally, we will be dropping the latitude and longitude features since it’s too specific that it will not be beneficial for our machine learning model. Furthermore, we will be dropping Time Zone feature since we will be using UTC time format instead to keep the time consistent throughout the data. Finally, we will be dropping the EventID column since it’s just a unique identifier for every entry in our dataset and it would interfere with the prediction of our model.

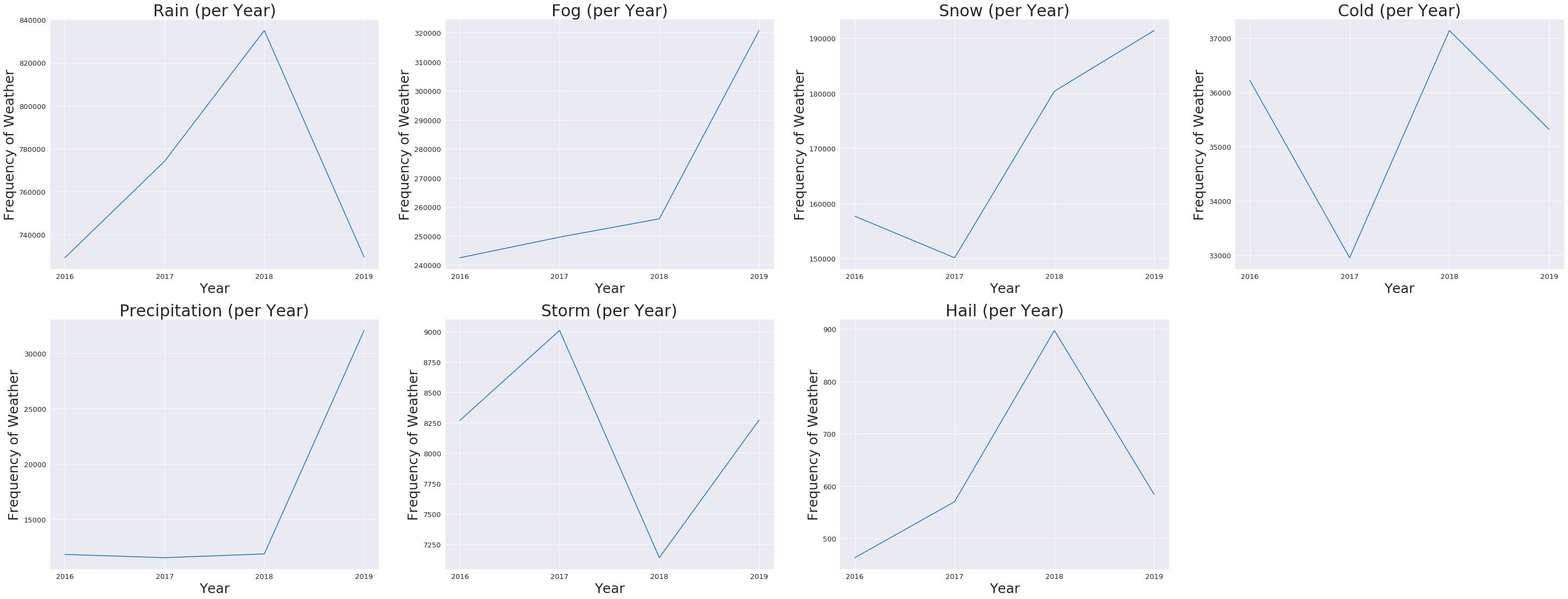
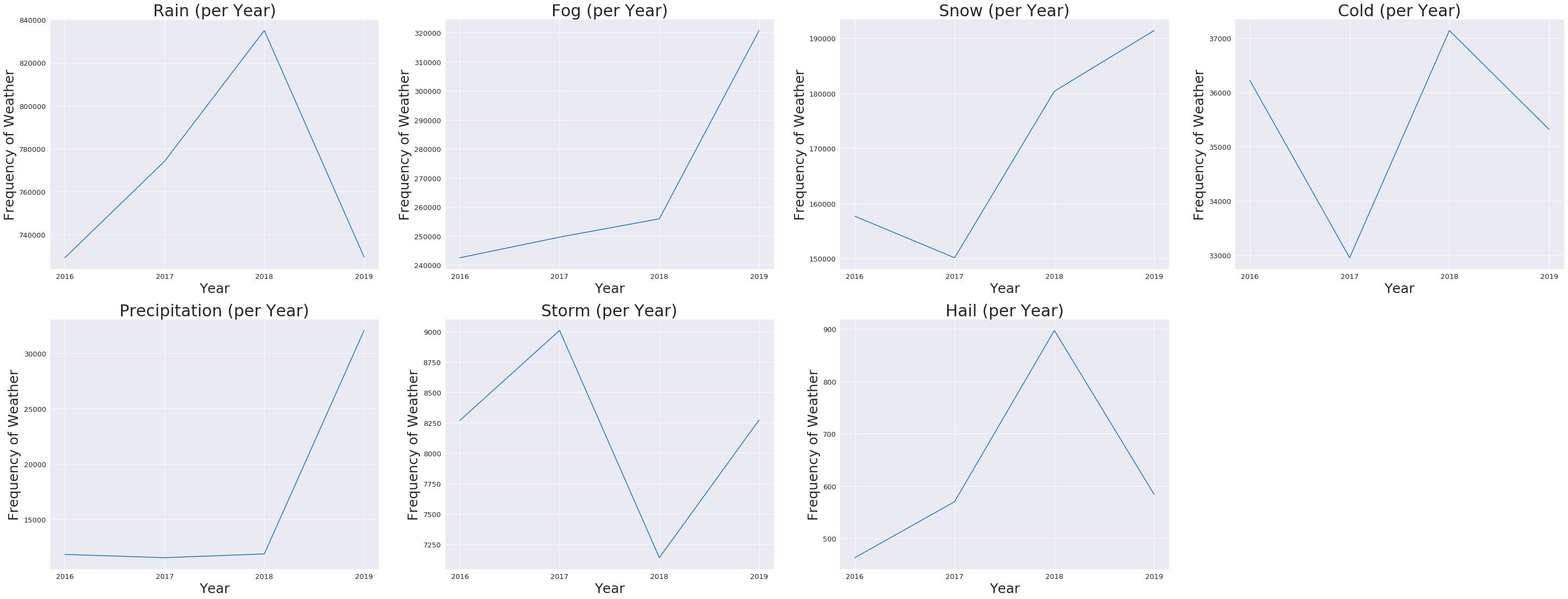
1. Extracting additional features

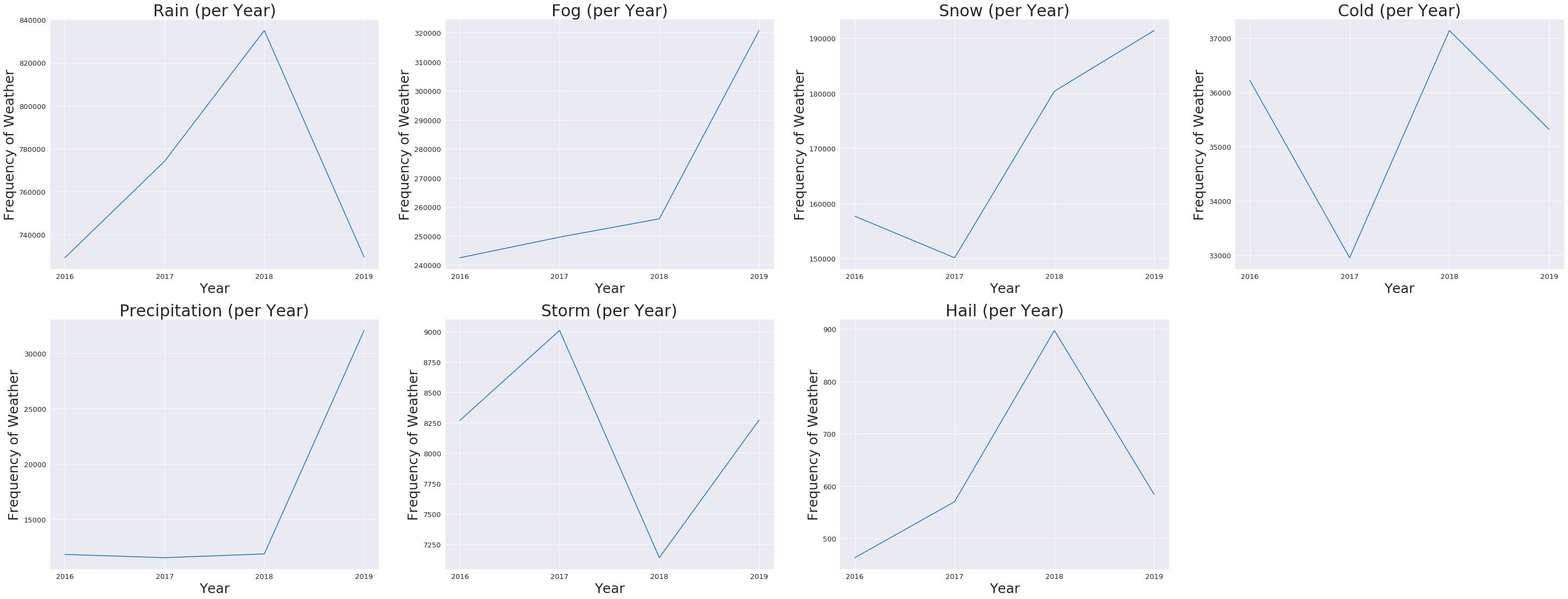
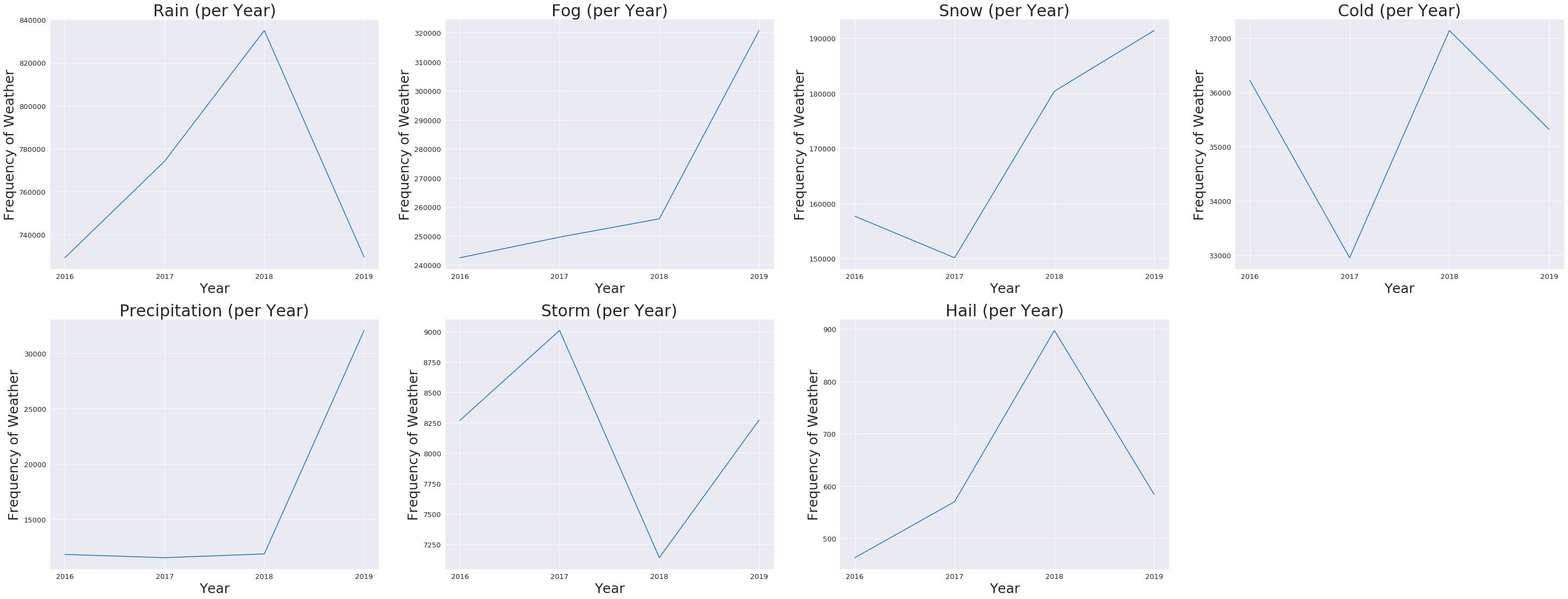
For the ease of data visualization, we will be extracting the hour, month and year from the time zone feature.

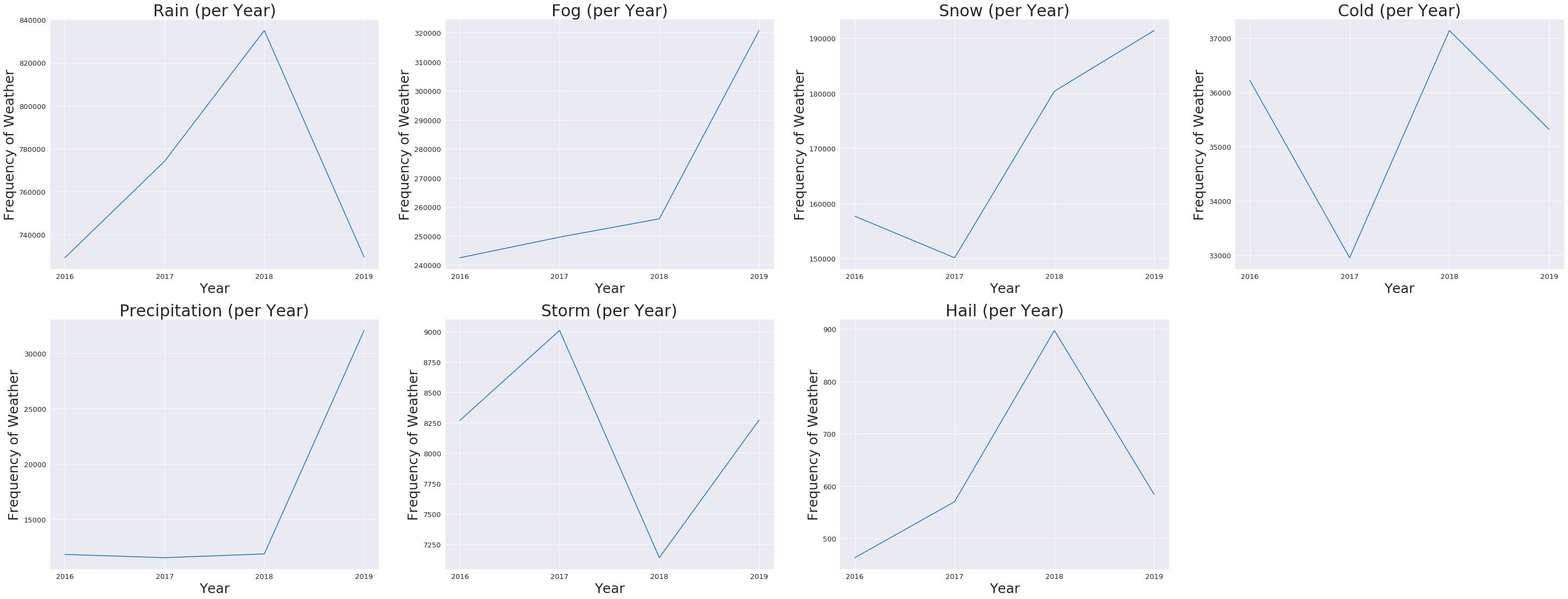
**Data Distribution and Visualization**

Below is a comparison of weather frequency distribution based on the type of weather from 2016-2019. As we can see from the histogram below, We can the most occurrence of rain event in our dataset as we get 3.1 million records and the least occurrence of hail event in our dataset as we get less than 1000 records.

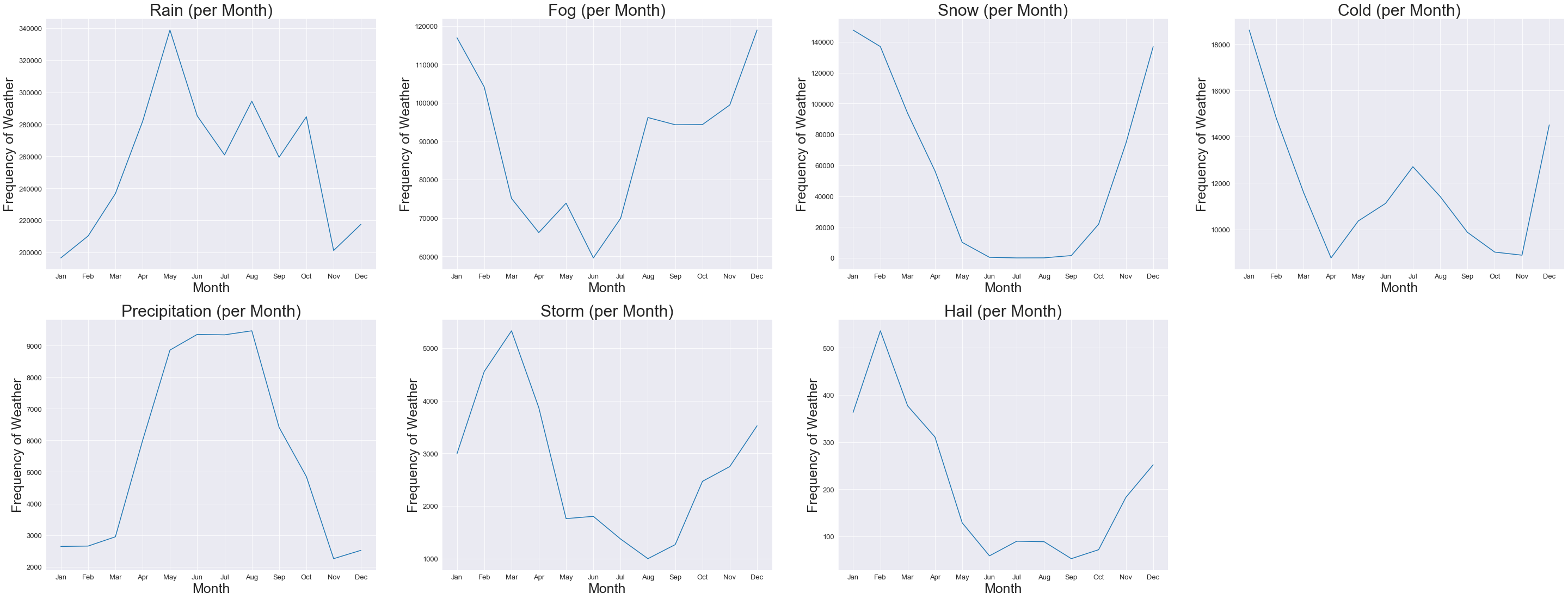
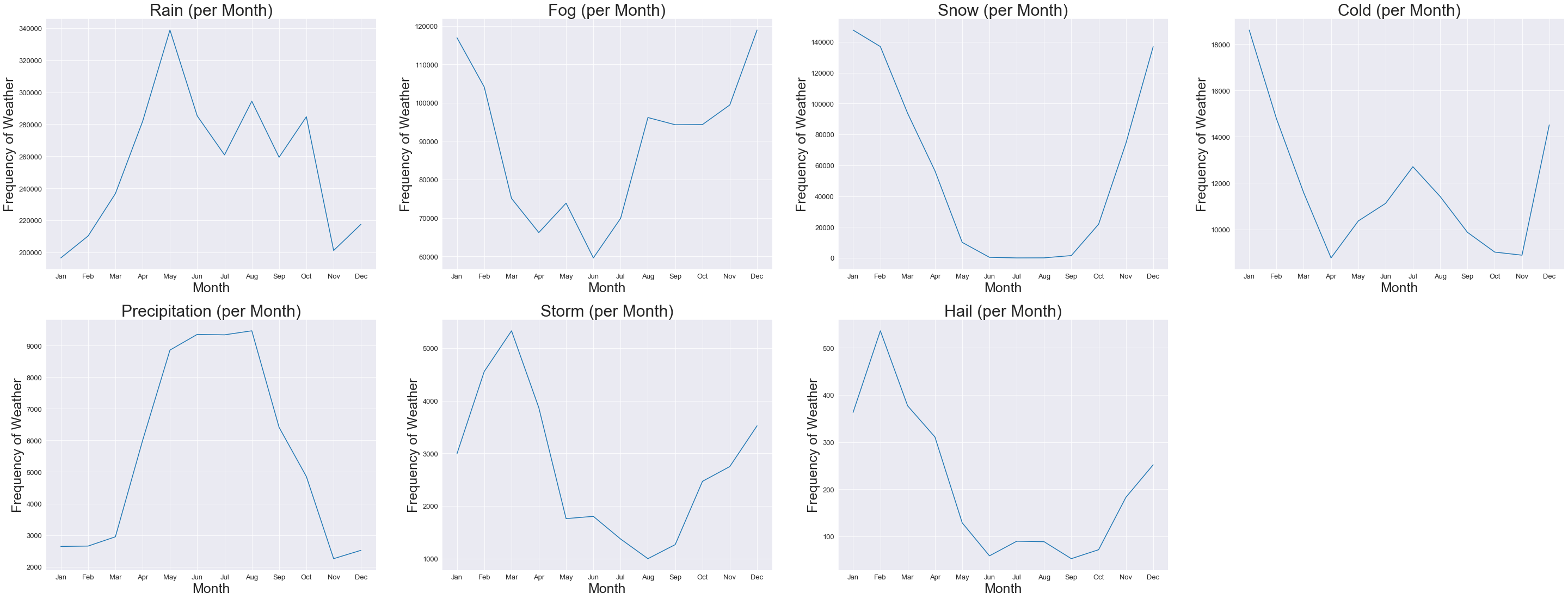
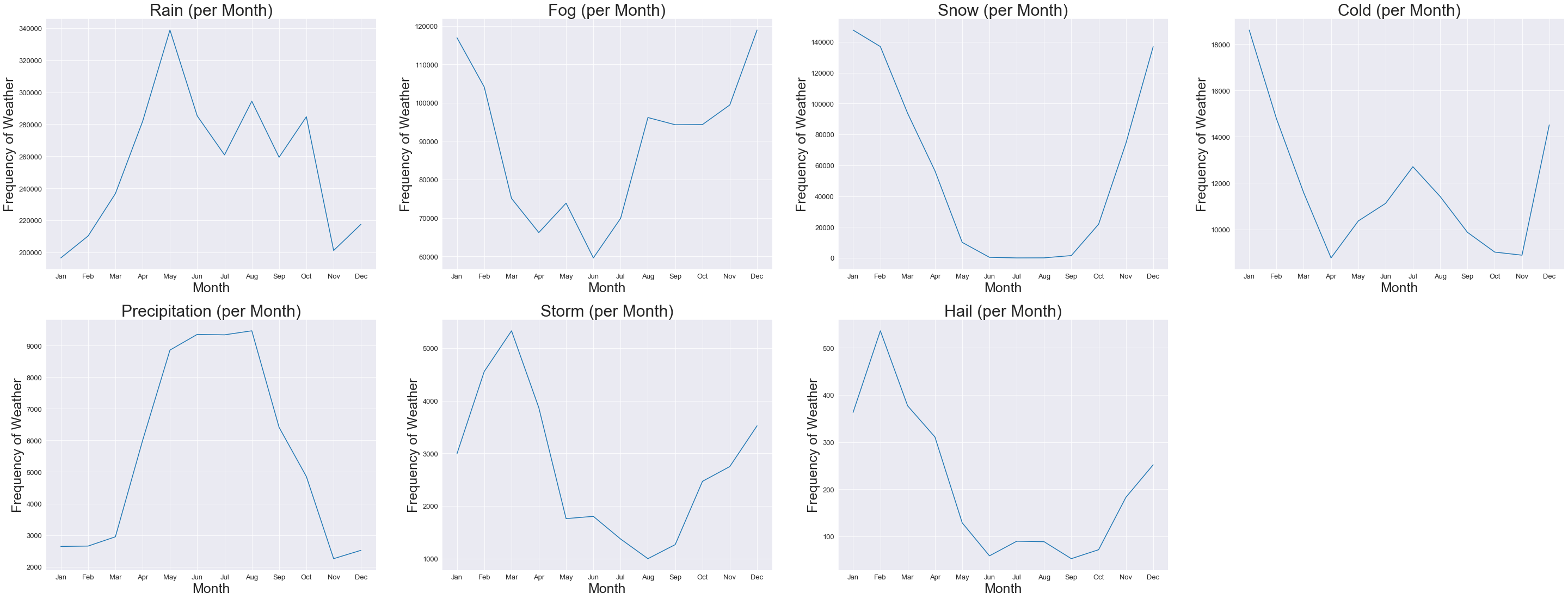
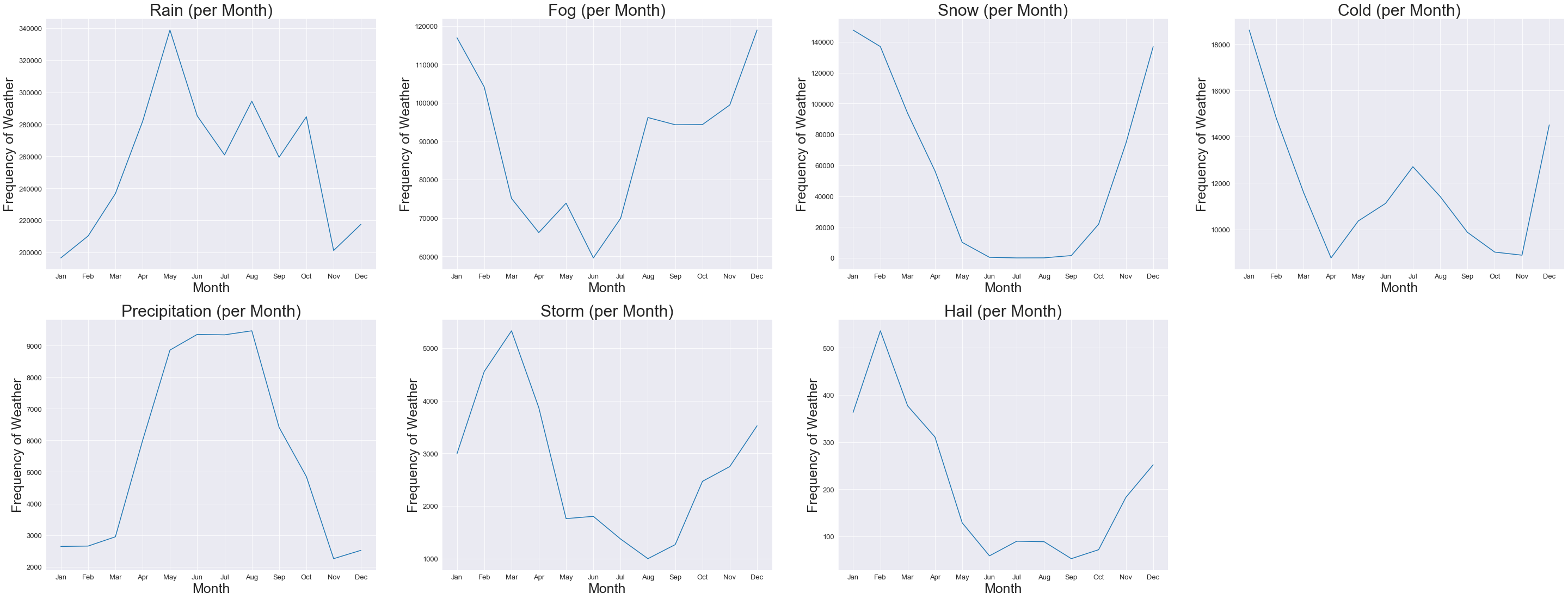
The 7 figures below represent the frequency of each type of weather from 2016 to 2019

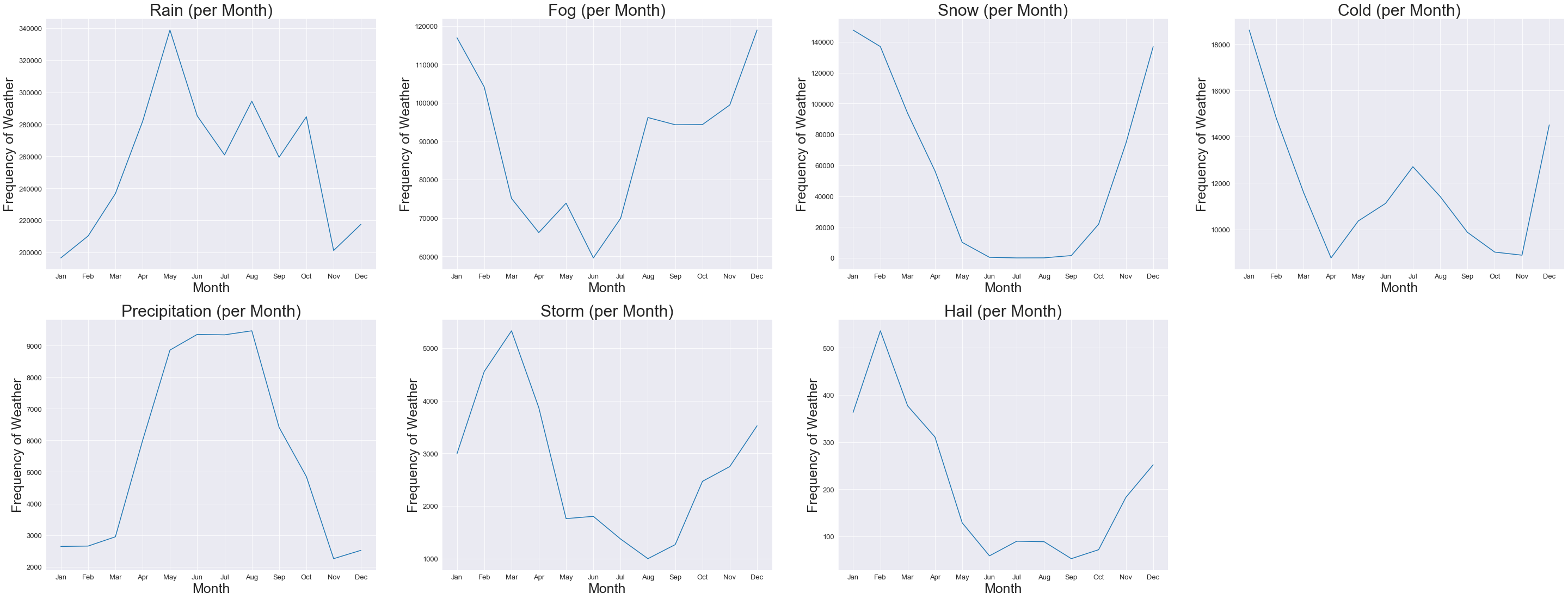


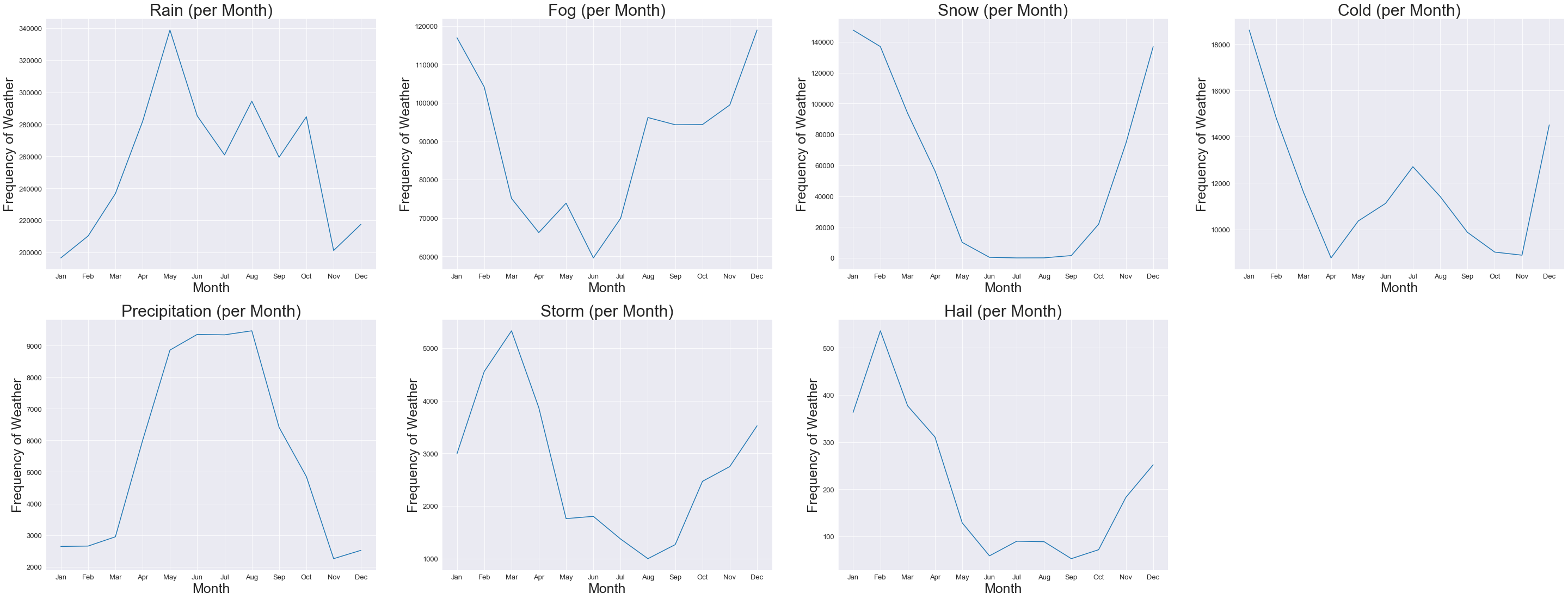


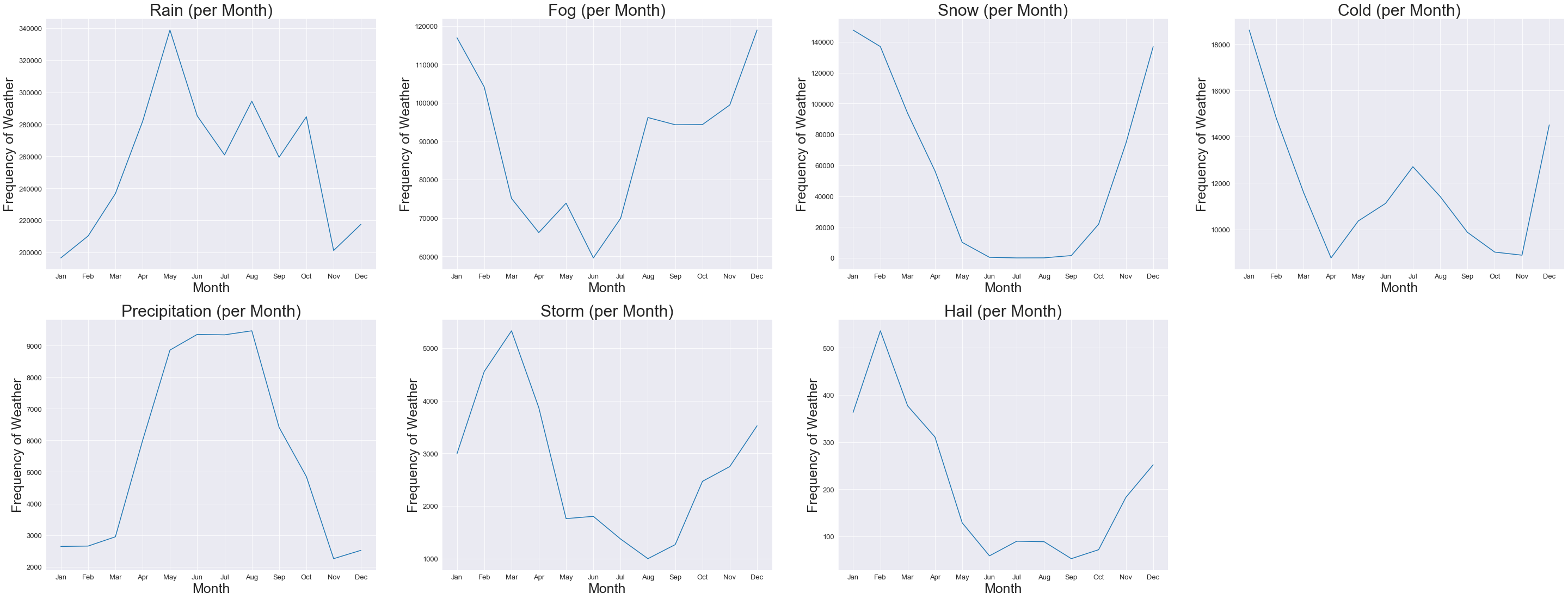


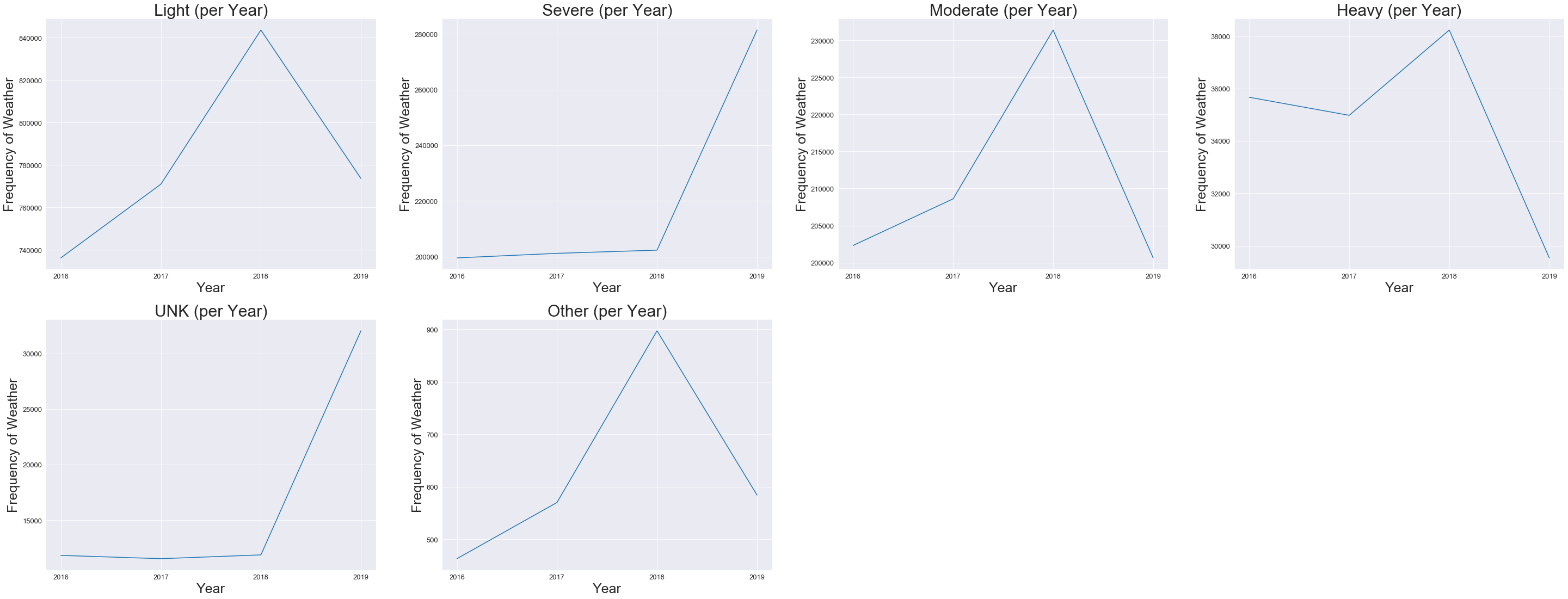
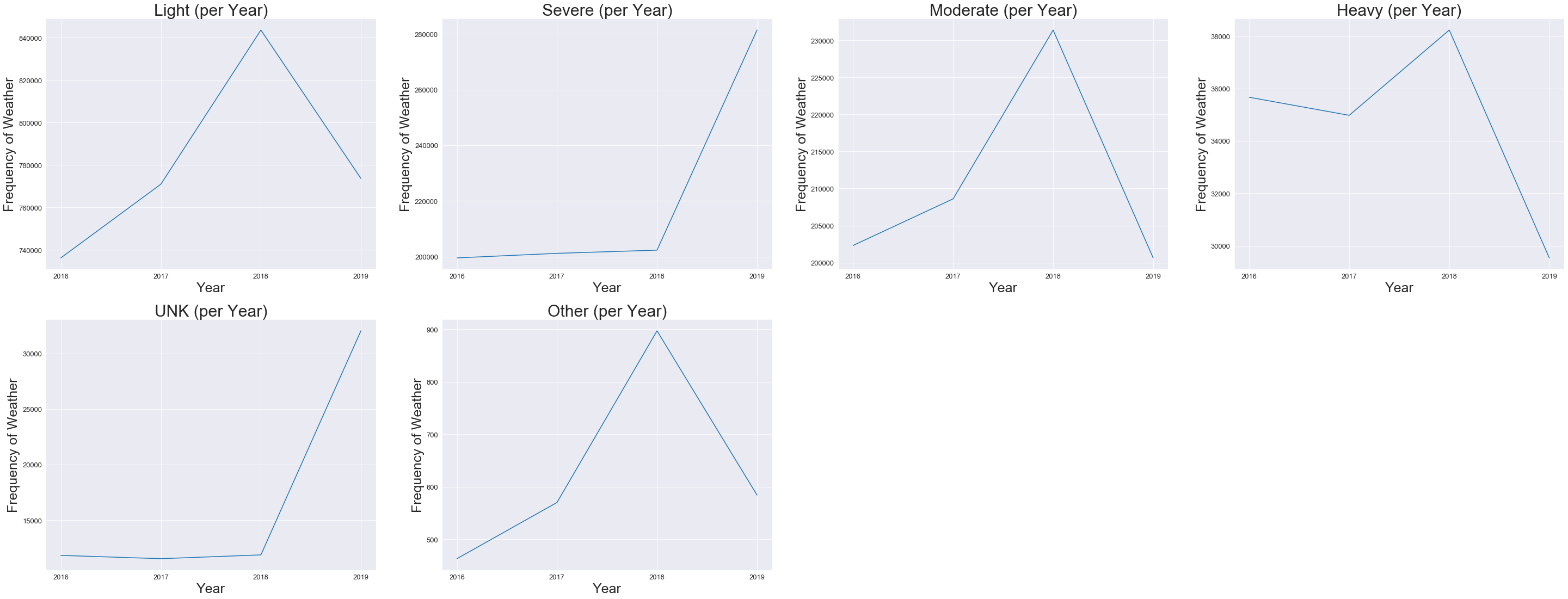
The 7 figures below shows the occurrence frequency of each weather events on a montly basis from 2016-2019

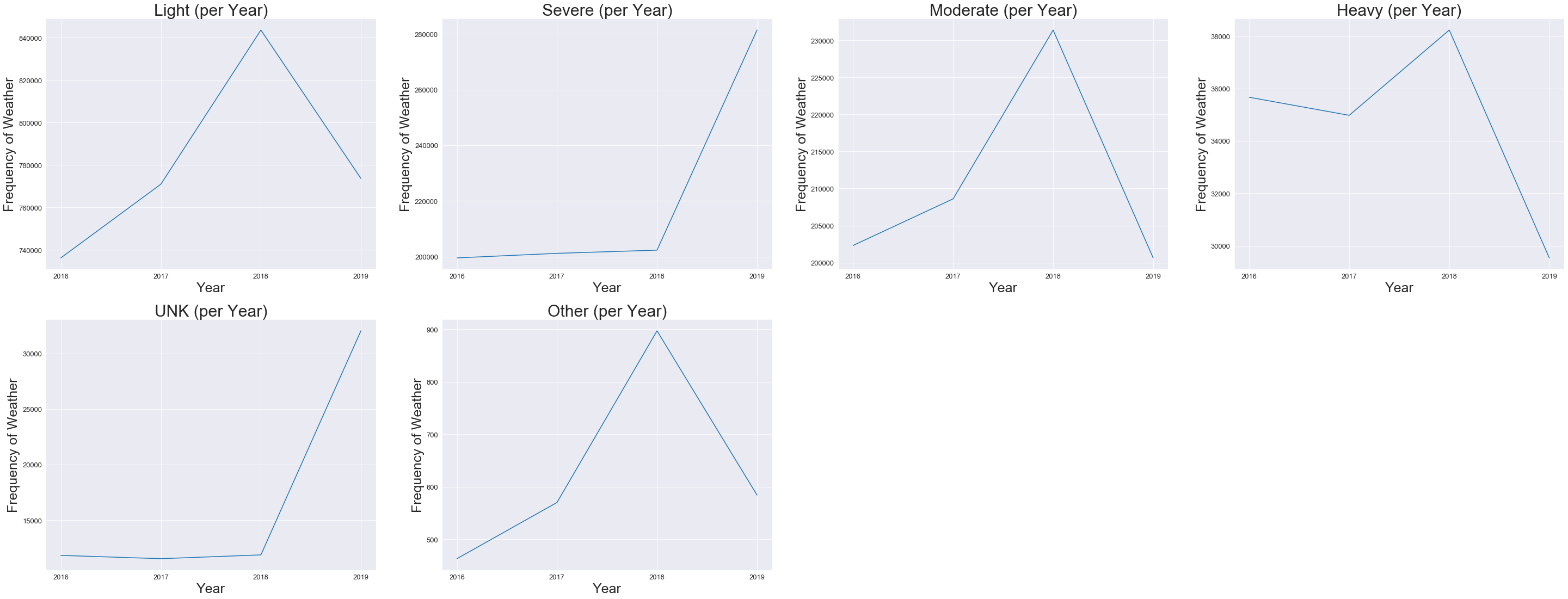
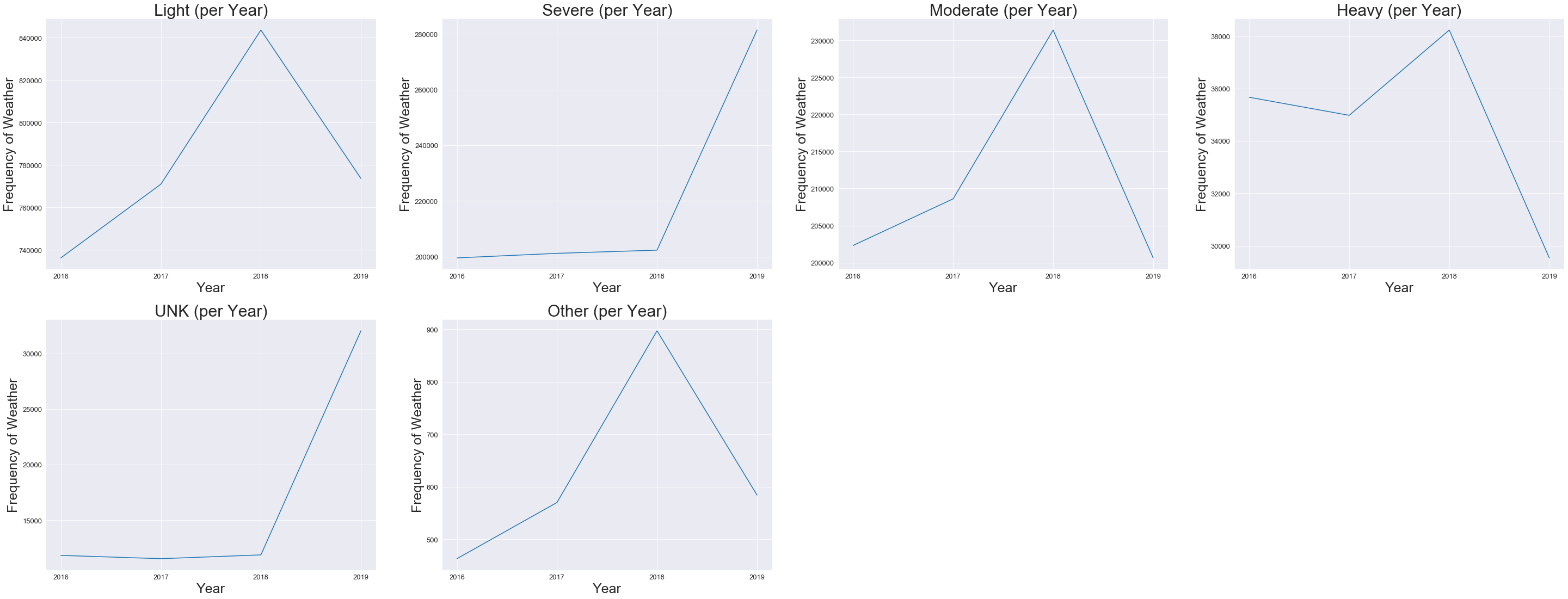


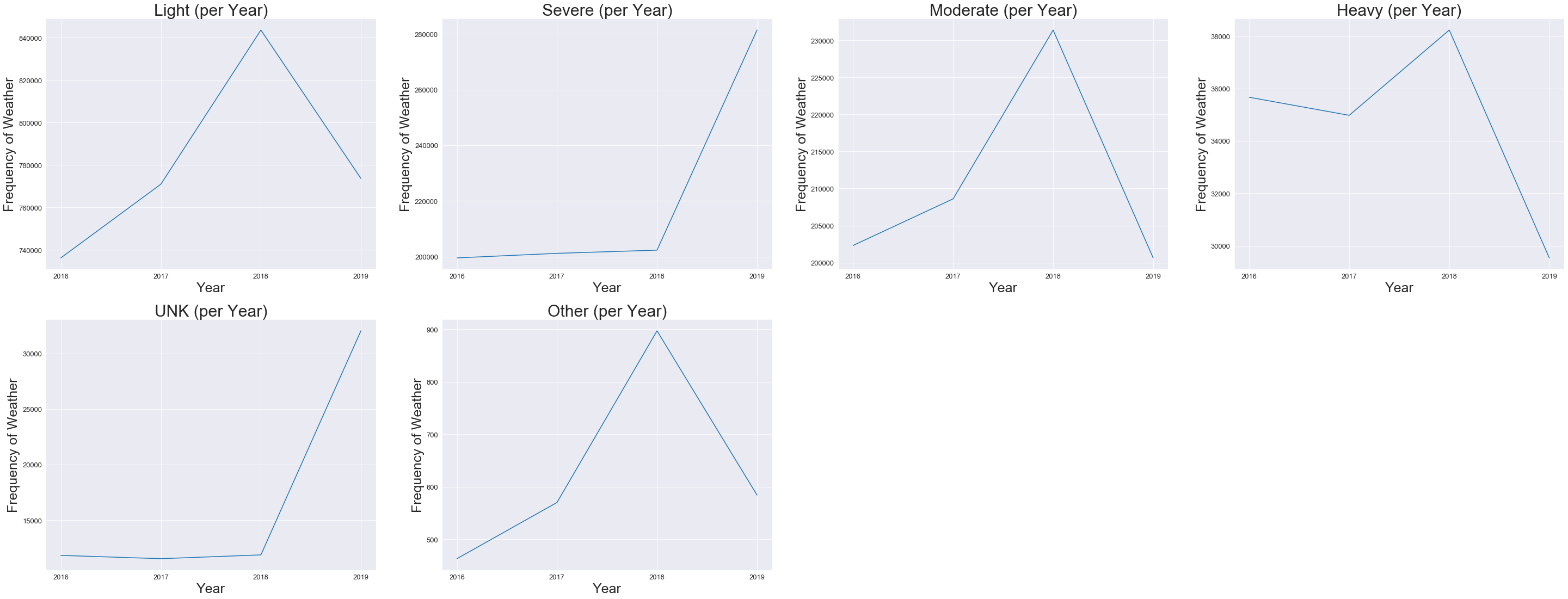
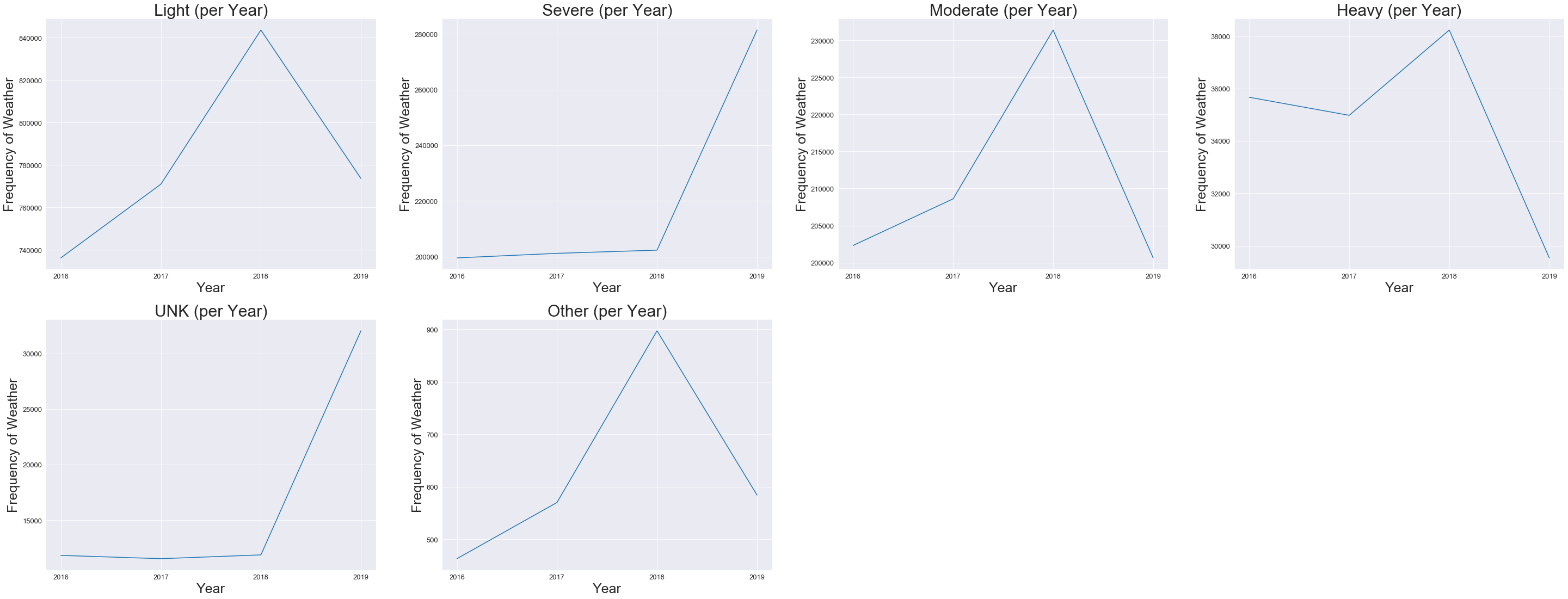






The 6 figures below represent the severity for each weather event that were happening from 2016 to 2019





**Result**

From our data preprocessing, below are the data description after we applied feature engineering. We eliminated multiple features along the way and extracted 3 additional. In the end, we only have the relevant features that we are planning to implement in our later model.

