Columns were removed using df.drop() that were not going to be used in analysis: “Fema\_delcaration\_string, “fips”, “last\_ia\_filing\_date”, etc.

Data was filtered to only include declared disasters from 1980 to 2020.

Outlier incidents were removed from the data set, as well as man-made disasters, so analysis could be conducted on natural declared disasters only using df.query(). US territories and commonwealths were removed from the dataset due to some areas, such as PR, causing imbalance to the dataset because of higher number of disasters using df.query().

Date strings were changed to date format using pd.DatetimeIndex(df[column]).date

Duplicate rows for the same incidents in different areas were removed from the dataset using df.drop\_duplicates(subset=[column]) - There were numerous incident reports for the same disaster in different areas that were taken out of the dataset so there would only be one row per disaster.

A new column was created for the month of occurrence using pd.DatetimeIndex(df[column]).month

A new column was created for incident duration by subtracting the begin date from the end date. The data type was also changed from an object type to an integer type: df[new\_column] = df[“end date”].sub(df[“begin date”], axis=0). The zero values in this column were then all replaced with ones, as the zeros did not mean zero but accounted for disasters less than 1 day in length.

A dictionary was created that matched the US region to the state and a new column was created using .apply: def function(state) return dictionary[state] df[new column] = df[“state”].apply(function). This same method was used to create a new column for the seasons each month of occurrence fell into.

The programs declared columns were added together and a new column was created that housed the totals

Data needed to be removed from the dataset: the original data set contained 63,701 rows and after cleaning, only 1,772 rows remained (61,929 rows dropped)