Outages

By Saroop Samra, Data Science Major UCSD, e: sksamra@ucsd.edu

Date : 20th Feb 2020

Summary of Findings

Introduction

This dataset relates to major electricity outage patterns and characteristics for U.S. states from the time period Jan 2000-July 2016. The definition of a major outage is as defined by the Department of Energy (DOE), at least 50,000 customers or loss of at least 300 MW, and each observation in the dataset is a major outage event as described. The fundamental data variables includes the reason of the outage (e.g. natural hazard), demand loss, customers affected and duration of the outage. Together the dataset also includes state based data climate characteristics, electical consumption, economic indicators, land use characteristics of each state at the time of the observation. This dataset was used in the paper, "A Multi-Hazard Approach to Assess Severe Weather-Induced Major Power Outage Risks in the U.S." (Mukherjee et al., 2018) [1]. The paper is based on a study that uses this dataset to develop a two-stage hybrid risk estimation model by using machine learning algorithms on this data. The model uses the various factors such of type of natural hazards, rural versus urban data etc, to develop a model that could help state regulators to make risk informed reslinace investment decisions.

The variables used in this analysis will be described shortly as they are used for EDA and analysis. However, in total there are 1536 observations and 56 variables and the broad characteristics of variables:

- GENERAL INFORMATION
- REGIONAL CLIMATE INFORMATION
- OUTAGE EVENTS INFORMATION
- REGIONAL ELECTRICITY CONSUMPTION INFORMATION
- REGIONAL ECONOMIC CHARACTERISTICS
- REGIONAL LAND-USE CHARACTERICS

The data was acquired by different publicly available datasets from the following agencies:

- DOE's Office of Electricity Delivery and Energy Reliability
- U.S. Energy Information Administration
- National Oceanic and Atmospheric Administration (NOAA) and National Climatic Data Center (NCDC)
- U.S. Department of Labor; Bureau of Labor Statistics
- U.S. Census Bureau.

For example, the DOE produces a monthly data report (OE-417 Electric Emergency and Disturbance Report). The raw data is available here: https://www.oe.netl.doe.gov/OE417_annual_summary.aspx (https://www.oe.netl.doe.gov/OE417_annual_summary.aspx). The Form OE-417 is a mandatory emergency form filed by state electricity utility companies. This leads to numerous different companies and hundreds of individual people who fill out this report. In consequence, the data entry relies on humans and is a potential problem in accurate and consistent data generation. For example, in the paper from Mukherjee et al, they mention out of the recorded disruptions "10 observations had wrong inputs of date and time (or both) for either or both the Event Start time and/or Restoration end time".

The summary of the missingness showed that there were 26 variables that had missing values and 1493 observation had at least one data variable missing. The missing variables included ones which were missing by design (e.g. Hurricane Names has 1462 missing values), as well as variables that were missing at random (e.g. Demand Loss had 705 missing values), as well as potentially non ignorable missing data. The initial approach to establishing missingness was to run descriptive statistics and then chart the variables that were missing. Finally, permutation tests were run to find if missing data dependent and in those cases to do conditional imputation. The details of this will be descibed below, however, overall the missingness could have an affect to answer questions about the dataset, in particular the demand loss variable is one of the important variables in my analysis and that had 705 missing values. As a result, I ran my hypothesis with the imputed data as well as the original data and this allowed me to be more confident that I did not reach different conclusions after the imputation step.

The key question for my analysis was how major power outages have changed over time, in particular with events caused by natural disaster. This is a vital question today as the topic of global warming and it's effect on the planet and human civilization is a controversial topic. The EDA analysis helped to frame and narrow this broad question to a meaningful question that I could attempt to analyze. Based on this I narrowed the question to the state of California:

The Null Hypothesis is as follows: "The California's demand loss from major outages in the recent (3 full years) times are equal to demand loss seen historically, any differences are due to random chance and is not statistically significant".

The Alternatve Hypothesis is as follows: "California's demand loss from major outages in the recent (3 full years) times is signficantly greater than the historical demand loss".

The test significance is 5% and the conclusions that could be drawn are that either we cannot reject that recent demand loss is equal to historical demand loss. Or we could conclude we can reject the null hypothesis and there is increased demand loss that we have seen in recent times. It is vital to make sure that this does NOT mean there is causation with these results, i.e. obviously we CANNOT conclude that global warming is the cause.

Cleaning and EDA

The data was loaded from the original excel file and required us to skip columns and rows to remove, for example the subheaders. The first step of cleaning the data was to create some new columns out of the given ones to make the analysis easier and more efficient. Here are the new columns that were added:

- OUTAGE.START.DATE and OUTAGE.START.TIME into a new column called OUTAGE.START
- OUTAGE.RESTORATION.DATE and OUTAGE.RESTORATION.TIME into a new column called OUTAGE.RESTORATION
- Location columns latitude and longitude. This used Nominatim feature from the geopy module. Unfortunately the data did not have the exact location of the power outage, so uses the US State state capital city.
- A "Natural Disaster" True/False column that is a convenience that allows analysis to quickly ask questions about natural disasters (which were originally marked as "severe weather" values in the CAUSE.CATEGORY column.

The next step was to clean up data:

- Dropped the OBS column, which was a duplicate of the index

- Dropped the observations for missing MONTH, these were 9 observations. These correspond to the "wrong inputs of date and time (or both) for either ot both the "Event Start time" and/or "Restoration end time" as reported in the paper and were due to incorrect data inputted.
- Cleaned up names of category events and details, these were messy fields that had numerous duplcates (e.g. "snow/ice storm", "snow/ice" etc). This condensed the number of potential values to 52 to 30. We also remove spaces and capatilized names.

I also replaced data that should be missing with NaN values. For example, when the DEMAND.LOSS.MW was greater than zero yet the CUSTOMERS.AFFECTED were reported as zero, the customers affected was set to NaN. There were additional 5 cases which also replaced NaN's relating to ensure that DEMAND.LOSS.MW, OUTAGE.DURATION and CUSTOMERS.AFFECTED had no illogical cases. The final step was to optimize some of the columns to a more efficient type. For example, changing MONTH and YEAR from strings to integers, CLIMATE.CATEGORY which is an ordinal that was represented by strings to integers.

Before starting EDA charting, the overall description of the data (using describe) was run to establish the overall statistics for each variable. The following plots were made for univariate analysis:

- Figure 1: Distribution of Category of Events plotted with a Pie Chart. This indicates that the majority of events were caused by Severe weather (49.8%) followed by Intentional attack (27.4%). This data indicates that natural disasters (severe weather) are an important area for us to further explore.
- Figure 2: Natural Disaster Details Summary. This plot was to narrow down on what are the frequency for the details for severe weather events. This indicates that storms was the largest cause with over 200 events and winter caused approx. 150 events.
- Figure 3: Monthly Electricity Price Residential, Commercial and Industrial Sectors. Next, we look at the price costs for residential, commercial and industrial sectors. The boxplot shows that for the fifty percentile, residential customers pay the most with over 10 cents/KWHour, commercial costs are lower and industrial costs are the lowest with close to half the price. This makes sense based on the volume pricing that is likely given to large industrial customers. There are some major outliers, for example for residential costs, 35 cents/KWHour which may indicate seasonal fluxuations in pricing.
- Figure 4: Electricity Consumption in Residential, Commercial and Industrial Sectors. Next, we look at the consumption for residential, commercial and industrial sectors. The boxplot shows for the fifty percentile, residential and commercial customers above 0.25 MGWHour whereas the industrial customers are below this figure. This is somewhat counter intuitive to Figure 3 as one could expect that customers who consume more would get better pricing. We also see major outliers for residential consumption which may be related to seasons, such as using air conditioning in the summer and heating in the winter.
- Figure 5 and Figure 6: Outage Duration Summary. The boxplot and histogram for outage duration together with the describe data shows that mean is 2637 minutes, but there is a large standard deviation (5953 minutes). We can see this in the boxplot and histogram as there are numerous major outliers, it is heavily right skewed: the average outages are less than 1 hour, but the worst case is 43 hours!
- Figure 7 and Figure 8: Customers Affected Summary. The boxplot and histogram for customers affected together with the describe data shows that mean is 167K customers for each outage, but there is a large standard deviation (300K customers). We can see this in the boxplot and histogram as there are numerous major outliers, it is heavily right skewed.
- Figure 9 and Figure 10: Demand Loss Affected Summary. The boxplot and histogram for demand loss (MW) affected together with the describe data shows that mean is 649 MW for each outage, but there is a large standard deviation (2408 MW). We can see this in the boxplot and histogram as there are numerous major outliers, it is heavily right skewed: the maximum loss is 4.178 GW, which is enough to power 3 million homes for a year(!) (Ref https://www.quora.com/How-many-homes-can-one-gigawatt-in-energy-capacity-provide-for (https://www.quora.com/How-many-homes-can-one-gigawatt-in-energy-capacity-provide-for))

- Figure 11: Top Ten States with Outages (Frequency). This shows the major outages are in California (210) and Texas (126).
- Figure 12: Top Ten Reasons for California Outages. This barchart shows that Fire (24 occurences) is the most frequent cause. The month of September has the most frequent occurences of fire events in California.
- Figure 13: Top Ten Reasons for Texas Outages. This barchart shows that Storm (19 occurences) is the most frequent cause. The month of May has the most frequent occurences of fire events in Texas.
- Figures 14: States Distribution of Top Ten 50 Largest Outages (Duration). We do the same analysis here but with durations rather than frequency of occurences. Now we see that New York is the state with the largest duration (10) occurences. Figure 15 shows the breakdown for New York and that fuel supply emergency was the largest cause (7) and the month this is most frequent is in February.
- Figures 16-18: Climate Regions with Most Outages Frequency/Mean Duration/Max Demand Loss. These three charts show that different climate regions have higher outages depending on the measurement, Northeast region (349) has the most outage events by frequency, East North Central region (5352 mins) has the mean outage duration and the West region (41788 MW) has the maximum demand loss. Overall the picture shows there is no one region that dominates what could be regarded as all the major outage measurements.
- Figures 19-20: Climate Category with Most Outages Frequency/Max Demand Loss. We repeat analysis now based on climate category (warm,cold,normal), two charts show that different climate categories have higher outages depending on the measurement, Normal climate category (744 occurences) has the most outage events by frequency yet Warm climate category (41788 MW) maximum demand loss. Overall the picture shows there is no one climate category that dominates what could be regarded as major outage measurements.
- Figures 21-22: El Nino Anomaly Level Distrubution Summary. The boxplot and histogram for El Nino Anomaly Level distributions duration together with the describe data shows that mean is -0.096852, but there is a large standard deviation (0.739957). We can see this in the boxplot and histogram as there are numerous major outliers, it is heavily right skewed.
- Figures 23: Outage Frequency each Month. The plot shows the month of June with the most events (195 occurences). This analysis is course grained and is expanded later with conditional analysis to provide a more complete picture.
- Figure 24: State utility sector's income New York vs California. This shows a different picture for these two important states, in California the income went up and in New York it went down. These might also be related to how much money the utility companies could invest in prevention of outages (such as cleaning up dry scrub to prevent fires).

Before we look into bivariate analysis, we do a scatter matrix (Figure 25) plot against variables that indicate the seriousness of the outage (duration, demand Loss, customers affected) together with economic and environment state variables (gross state product, urban population, residential price etc). This is a useful chart to then decide what bivariate analysis we should start with. The following plots were made for bivariate analysis:

- Figure 26: Duration vs State GSP. The outliers show short duration when State GSP is high and long durations when GSP is low. The majority of plots are less than 2000 mins and $8000 per capital GSP.
- Figure 27: Duration vs Urban Population. Short duration outages when Urban Population is High. Long durations outliers when Urban Population is low (higher rural population). The majority of data falls between urban population between 65-95%.
- Figure 28: Duration vs Residential Price. Short duration outages when Residential Price is Low. Long durations outliers when Residential Price is Higher. The majority of the data is under 20 cents/KWh.
- Figure 29: Duration vs Demand Loss MW. here is a trend but the outliers are short duration with large demand loss (cities). Long durations with small demand loss (rural). These show two interesting extreme

cases, one with little power loss that lastest 8000 minutes. And a major power loss even of 4000 MW that lasted a short time.
- Figure 30: Duration vs Demand Loss MW (outliers removed). With outliers removed there is a little more of an obvious positive trend. The most frequent data is in the lower quartet of the data, less than 500 minutes and less than 500 MW loss.
- Figure 31: Urban Population vs State Population with Demand Loss (size). Positive trend and demand loss is also higher. The state populations with greater than 4 million residents have the large urban populations.
- Figure 32: Residential Consumption vs State Population with Outage Duration (size). Positive trend and outage duration tends also higher. There are 4 clusters, for example one cluster is between 0.5-1.0 residential consumption and 3-4 million population.
- Figure 33: Land vs State Population with Outage Duration (size). Still a positive trend, the duration sizes do though vary in low and high percentage of land usage. Greater than 80 percent state land area compared to continental US is with populations of 4 million
- Figure 34: Natural Disaster year vs Demand Loss. Not a clear trend, this is something that needs more categorical analysis.

As highlighted in Figure 34, we now need to establish conditional distributions using pivot tables with the impact of the causes (demand loss, customers impacted, outage duration, frequency) over time (YEAR). The following plots were made:

- Figure 35: Frequency of major outages by Cause. A striking result shown is that Intentional attacks substantially increased in 2011 (121 occurences). This maybe related to heightened security risks after 9/11.
- Figure 36: Duration of major outages by Cause. Fuel Supply was a major cause of duration outages in 2014 (226177 minutes). As we have found out earlier this was a major issue in the state of New York.
- Figure 37: Customers Affected by major outages by Cause. Severe Weather is the major cause for largest customers affected events. Over 134,717,133 customers incidents have been reported in the dataset.
- Figure 38: Demand Loss by major outages by Cause. 2014 was an unusual year as it had a massive number loss associated with system operability disruption (53646).
- Figure 39: Frequency of severe weather events. Storms have been the largest cause with 25 cases reported in 2008.
- Figure 40: Duration of severe weather events. Hurricane/tornadoes have been the largest cause 2018 as the year with the highest loss because of these events with over 166 days lost.
- Figure 41: Customers affected for severe weather events. Earthquakes are seldom a cause of major outages, 2008 was the largest customer (350886) impact due to earthquakes.
- Figure 42: Duration of severe weather events. Winter 2014 was a major impact with 18 days lost.

To conclude this section, we do develop visualization of geospatial data by using Folium geospatial plotting library. Note, as we do not have exact geospatial data for each incident, the geospatial is of a somewhat limited application:

- Figure 43: Top 20 Outages vs Top 20 Demand Loss. We see the top 20 outages based on demand loss are in the west and north east. However, the midwest is a region with impact due to outage duration. Texas also has been impacted in both duration and demand loss.
- Figure 44: Top 5 Longest Duration Outages by Severe Event Type. We can see fire events in the west and wind and winter events in the north east.

Finally, the EDA has been immensely useful and in particular the trivariate analysis of using pivot tables for aggregated charts, leading to a potential hypothesis question relating to demand loss trends. Together, the severe weather is a major cause overall in outages and California is a state that has major outages. This leads to

an intersting question if major power outages have changed over time, in particular with events caused by natural disaster. In particular, has California's demand loss from major outages in the recent times been signficantly greater than the historical demand loss. We will have to establish what we mean by "recent", which is explored later in the hypothesis testing section.

Assessment of Missingness

In this section we first plot the missingness of variables that have missing data. The focus initially is on what are the important columns that have lots of things missing and could thus impact our ability to do accurate analysis. Figure 45 shows the major missing values per column, with major being defined as having 40 or more values missing with greater than 40 things missing. Hurricane names clearly stand out as the obvious longpole with 1453 missing values. However, as mentioned in the dataset description, "If the outage is due to a hurricane, then the hurricane name is given by this". This means that the dataset was intentionally designed to have this column missing. I performed a query to check these missing values and see if any CAUSE.CATEGORY.DETAIL values were due to hurricanes, and there was none. This means we can rule that "HURRICANE.NAMES" is Missing by Design (MD) and no further action is warranted.

The next set of values that have very high missing values are the Demand Loss (DEMAND.LOSS.MW) with 844 missing values and the Customers Affected (CUSTOMERS.AFFECTED) with 595 missing values. These columns are critical for the analysis on impact of the outage events and warrants a high priority to resolve. For the demand loss, the following potential dependent columns (see below) were used to do a permutation test to verify if demand loss was MAR dependent on them. Some of them are not obvious why they would be MAR dependent but at this stage our goal is to try multiple columns and see if there are any dependencies:

- MAR DEMAND.LOSS.MW check dependency against YEAR pvalue= 0.873
- MAR DEMAND.LOSS.MW check dependency against MONTH pvalue= 0.216
- MAR DEMAND.LOSS.MW check dependency against U.S._STATE pvalue= 0.026
- MAR DEMAND.LOSS.MW check dependency against NERC.REGION pvalue= 0.306
- MAR DEMAND.LOSS.MW check dependency against CLIMATE.REGION pvalue= 1.0
- MAR DEMAND.LOSS.MW check dependency against CLIMATE.CATEGORY pvalue= 1.0
- MAR DEMAND.LOSS.MW check dependency against CAUSE.CATEGORY pvalue= 0.864

The DEMAND.LOSS.MW column is not for example MAR dependent on CLIMATE.CATEGORY or CLIMATE.REGION and it doesn't appear the MONTH or YEAR have any correlation to missingness. However, the state (U.S._STATE) has a p-value of 0.026 and is concluded to be the dependent column for the demand loss. In this case we did an imputation based on mean of groups, we could have done a probablistic distribution but there is no data to indicate if the non missing demand loss data is representative. This does reduce variation, however we do not expect it to affect the mean. Figure 46 and 47 show the histogram plot of DEMAND.LOSS.MW before and after cleaning. Note, that the hypothesis test that is later executed was tested with and without the imputation to ensure this modification did not overall change our conclusion.

For the customer affected, the following potential dependent columns (see below) were used to do a permutation test to verify if customers affected was MAR dependent on them:

- MAR CUSTOMERS.AFFECTED check dependency against YEAR pvalue= 0.147
- MAR CUSTOMERS.AFFECTED check dependency against MONTH pvalue= 0.77
- MAR CUSTOMERS.AFFECTED check dependency against U.S._STATE pvalue= 0.471
- MAR CUSTOMERS.AFFECTED check dependency against NERC.REGION pvalue= 0.954
- MAR CUSTOMERS.AFFECTED check dependency against CLIMATE.REGION pvalue= 0.265
- MAR CUSTOMERS.AFFECTED check dependency against CLIMATE.CATEGORY pvalue= 0.887

- MAR CUSTOMERS.AFFECTED check dependency against CAUSE.CATEGORY pvalue= 0.93

In this case we do not find any dependencies. These are all categorical variables and the quantative variables were also tried using a binning approach to allow a distribution against missingness to be calculated. However, there was no case that resulted in statistically significant p-values. The demand loss which also has high amounts of missing data did result in a small p-value and that we have determined is MAR dependent on U.S._STATE. But is it reasonable to use the argument to extend the dependency of states to customers affected too? Or should the interpretation hinge on should we treat missing data for customers affected as MCAR or NMAR? There could be potential limitless number of reasons to come up why this is NMAR but it is really important to establish some reasonableness to be able to qualify this as NMAR. As mentioned in the summary, the data collection process involves filling out Form OE-417 is a mandatory emergency form filed by state electricity utility companies. This leads to numerous different companies and hundreds of individual people who fill out this report. In consequence, the data entry relies on humans and is a potential problem in accurate and consistent data generation. This would imply that this data should be treated as MCAR and is imputed with a unconditional probabilistic replacement. Figure 47 and 48 show the histogram plot of CUSTOMERS.AFFECTED before and after cleaning. Note, that the hypothesis test that is later executed was tested with and without the imputation to ensure this modification did not overall change our conclusion.

Hypothesis Test

The key question for my analysis was how major power outages have changed over time, in particular with events caused by natural disaster. This is a vital question today as the topic of global warming and it's affect on the planet and human civilization is a controversial topic.

The Null Hypothesis is as follows: "The California's demand loss from major outages caused by natural disaster in the recent (3 full years) times are equal to demand loss seen historically, any differences are due to random chance and is not statistically significant".

The Alternatve Hypothesis is as follows: "California's demand loss from major outages caused by natural disaster in the recent (3 full years) times is signficantly greater than the historical demand loss".

The test significance is 5 percent and the conclusions that could be drawn are that either we cannot reject that recent demand loss due to natural hazards in California is equal to historical demand loss. Or we could conclude we can reject the null hypothesis and there is increased demand loss that we have seen in recent times.

One obvious question is how to determine "recent", as indicated this is for a duration of 3 years. Climate is the long-term average of weather, typically averaged over a period of 30 years (https://en.wikipedia.org/wiki/Climate (https://en.wikipedia.org/wiki/Climate)). Unfortunately we have limited historical data, Jan-2000-July 2016. This results in a ratio of 3.5 years out of 15 years total, approx 20 percent is a reasonable amount of data in the recent past. We could select more years as "recent", however, we will have diminishing returns as the analysis would treat "recent" as more common and potentially smooth out any historical trends.

We perform the hypothesis test using a difference of means. We use the data set as representing the population and the years 2013-2016 where there were natural disaster caused outages (severe weather) in California. We simulate 10,000 null hypothesis tests using empirical sampling with replacement as false. The result is shown in Figure 50 and has a p-value of 0.0139. Thus, we can reject our Null Hypothesis. As we did perform imputation on the DEMAND.LOSS.MW column, I performed the same experiment without imputation and the resultant p-value was also statistically significant (0.0107). Thus, our imputation has not impacted our conclusion.

Finally it is vital to make sure that this does NOT mean there is causation with these results, i.e. obviously we CANNOT conclude that global warming is the cause.

Addendum: How would our conclusion change if we used customers affected column instead of demand loss? This is a valid experiment to run and Figure 51. This does not reject the null hypothesis as the p-value is 0.5212. This although on appearance is counterintutive with our main conclusion, shows that the demand loss although increased, the number of customer who are impacted has been reduced. This could be due to a variety of factors such as investment made by utility companies to quickly reliable power. This would be an interesting follow up study.

Citations

[1] A Multi-Hazard Approach to Assess Severe Weather-Induced Major Power Outage Risks in the U.S." (Mukherjee et al., 2018) https://www.sciencedirect.com/science/article/pii/S0951832017307767 (https://www.sciencedirect.com/science/article/pii/S0951832017307767)

Code

In [104]:
```
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import seaborn as sns
%matplotlib inline
%config InlineBackend.figure_format = 'retina'  # Higher resolution figu
res

import folium
import folium.plugins
import geojson
from geopy.geocoders import Nominatim
import branca.colormap as cm
import datetime

%load_ext autoreload
%autoreload 2
import util
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

Cleaning and EDA

In [105]:
```
# Clean Data: load the Outages, and skip first few rows
outages_fp = os.path.join("outages_data", "outage.xlsx")
outages = pd.read_excel(outages_fp, skiprows=5)
# Drop the sub header row
outages = outages.drop(0)
# Drop the empty variables column
outages = outages.drop("variables", axis=1)
outages.head()
```

Out[105]:

	OBS	YEAR	MONTH	U.S._STATE	POSTAL.CODE	NERC.REGION	CLIMATE.REGION	ANOMAI
1	1.0	2011.0	7.0	Minnesota	MN	MRO	East North Central	
2	2.0	2014.0	5.0	Minnesota	MN	MRO	East North Central	
3	3.0	2010.0	10.0	Minnesota	MN	MRO	East North Central	
4	4.0	2012.0	6.0	Minnesota	MN	MRO	East North Central	
5	5.0	2015.0	7.0	Minnesota	MN	MRO	East North Central	

5 rows × 56 columns

In [106]:
```
# Clean Data: get date time information from date and time columns

# Utility function to reduce code duplication, takes date and time and c
ombines into one column
def convert_datetime(df, date_col, time_col, datetime_col):
    df[datetime_col] = pd.to_datetime(df[date_col]) + pd.to_timedelta(df
[time_col].astype(str))
    df = df.drop([date_col, time_col], axis=1)
    return df

# Clean Data: Combine OUTAGE.START.DATE and OUTAGE.START.TIME into a new
column called OUTAGE.START.
outages = convert_datetime(outages, "OUTAGE.START.DATE", "OUTAGE.START.T
IME", "OUTAGE.START")

# Clean Data: Combine OUTAGE.RESTORATION.DATE and OUTAGE.RESTORATION.TIM
E into a new column called OUTAGE.RESTORATION
outages = convert_datetime(outages, "OUTAGE.RESTORATION.DATE", "OUTAGE.R
ESTORATION.TIME", "OUTAGE.RESTORATION")

outages[["OUTAGE.START", "OUTAGE.RESTORATION"]].head(1)
```

Out[106]:

	OUTAGE.START	OUTAGE.RESTORATION
1	2011-07-01 17:00:00	2011-07-03 20:00:00

In [107]:
```
# Clean Data: add GPS Latitude and Longitude columns based on State
geolocator = Nominatim(user_agent='myapplication')
def add_lat_long(x):
    state = x.iloc[0]["U.S._STATE"]
    location = geolocator.geocode(state)
    x['latitude'] = location.latitude
    x['longitude'] = location.longitude
    return x

outages = outages.groupby("U.S._STATE").apply(add_lat_long)
outages[["U.S._STATE", "latitude", "longitude"]].head(1)
```

Out[107]:

	U.S._STATE	latitude	longitude
1	Minnesota	45.989659	-94.611329

In [108]:
```
# Clean Data: Add Natural Disaster Column which is True/False
outages["Natural Disaster"] = outages["CAUSE.CATEGORY"]  == "severe weat
her"
outages[["U.S._STATE", "Natural Disaster", "CAUSE.CATEGORY"]].head(1)
```

Out[108]:

	U.S._STATE	Natural Disaster	CAUSE.CATEGORY
1	Minnesota	True	severe weather

In [109]:

```
# Clean Data: TODO : Uncomment Drop OBS Column
outages = outages.drop("OBS", axis=1)

# Clean Data: Drop these: 9 Missing Months, Outages Start, Duration & Re
storation, ANOMALY.LEVEL, CLIMATE.CATEGORY, RES.PRICE-IND.PERCEN
outages = outages[outages["MONTH"].notnull()]

# Clean Data: Cleanup messy field to aggerate types of severe weather ev
ents
outages = outages.replace({"thunderstorm; islanding" : "thunderstorm",
                           "wind storm" : "wind",
                           "wind/rain" : "wind",
                           "heavy wind" : "wind",
                           "winter storm" : "winter",
                           "snow/ice storm" : "winter",
                           "snow/ice " : "winter",
                           "snow" : "winter",
                           "hailstorm" : "winter",
                           "uncontrolled loss" : "other",
                           "public appeal" : "other",
                           "fog" : "other",
                           "heatwave" : "fire",
                           "wildfire" : "fire",
                           "hurricane" : "hurricane/tornadoes",
                           "hurricanes" : "hurricane/tornadoes",
                           "tornadoes" : "hurricane/tornadoes",
                           "thunderstorm" : "storm",
                 })

# Clean Data: Cleanup messy field remove spaces at start
outages["CAUSE.CATEGORY.DETAIL"] = outages["CAUSE.CATEGORY.DETAIL"].str.
strip(" ")
```

In [110]:
```
# Clean Data: Replace Nan for CUSTOMERS.AFFECTED when is zero but DEMAN
D.LOSS.MW>0
outages.loc[(outages["DEMAND.LOSS.MW"] > 0) & (outages["CUSTOMERS.AFFECT
ED"]==0), "CUSTOMERS.AFFECTED"] = np.nan
# Clean Data: Replace Nan for OUTAGE.DURATION when is zero but DEMAND.LO
SS.MW>0
outages.loc[(outages["DEMAND.LOSS.MW"] > 0) & (outages["OUTAGE.DURATION"
]==0), "OUTAGE.DURATION"] = np.nan
# Clean Data: Replace Nan for DEMAND.LOSS.MW when is zero but OUTAGE.DUR
ATION>0
outages.loc[(outages["OUTAGE.DURATION"] > 0) & (outages["DEMAND.LOSS.MW"
]==0), "DEMAND.LOSS.MW"] = np.nan
# Clean Data: Replace Nan for CUSTOMERS.AFFECTED when is zero but OUTAG
E.DURATION>0
outages.loc[(outages["OUTAGE.DURATION"] > 0) & (outages["CUSTOMERS.AFFEC
TED"]==0), "CUSTOMERS.AFFECTED"] = np.nan
# Clean Data: Replace Nan for DEMAND.LOSS.MW when is zero but CUSTOMERS.
AFFECTED>0
outages.loc[(outages["CUSTOMERS.AFFECTED"] > 0) & (outages["DEMAND.LOSS.
MW"]==0), "DEMAND.LOSS.MW"] = np.nan
# Clean Data: Replace Nan for OUTAGE.DURATION when is zero but CUSTOMER
S.AFFECTED>0
outages.loc[(outages["CUSTOMERS.AFFECTED"] > 0) & (outages["OUTAGE.DURAT
ION"]==0), "OUTAGE.DURATION"] = np.nan
```

In [111]:
```
# Clean Data: Optimize types for MONTH convert to int
outages["MONTH"] = outages["MONTH"].astype(int)
# Clean Data: Optimize types for YEAR convert to int
outages["YEAR"] = outages["YEAR"].astype(int)
# Clean Data: CLIMATE.CATEGORY set to Ordinal
outages["CLIMATE.CATEGORY"] = outages["CLIMATE.CATEGORY"].replace({"Col
d" : 0, "Normal" : "1", "Warm" : 2})
```

In [112]:
```
# Clean Data : Capilitize names to make reading easier
outages["CAUSE.CATEGORY"] = outages["CAUSE.CATEGORY"].str.capitalize()
outages["CAUSE.CATEGORY.DETAIL"] = outages["CAUSE.CATEGORY.DETAIL"].str.
capitalize()
```

In [113]:
```
# Clean Data : Data after cleaning
outages.head(5)
```

Out[113]:

	YEAR	MONTH	U.S._STATE	POSTAL.CODE	NERC.REGION	CLIMATE.REGION	ANOMALY.LEVE
1	2011	7	Minnesota	MN	MRO	East North Central	-0.
2	2014	5	Minnesota	MN	MRO	East North Central	-0.
3	2010	10	Minnesota	MN	MRO	East North Central	-1.
4	2012	6	Minnesota	MN	MRO	East North Central	-0.
5	2015	7	Minnesota	MN	MRO	East North Central	1.

5 rows × 56 columns

In [114]:
```
# EDA : use describe at this stage to better understand the data
outages.describe()
```

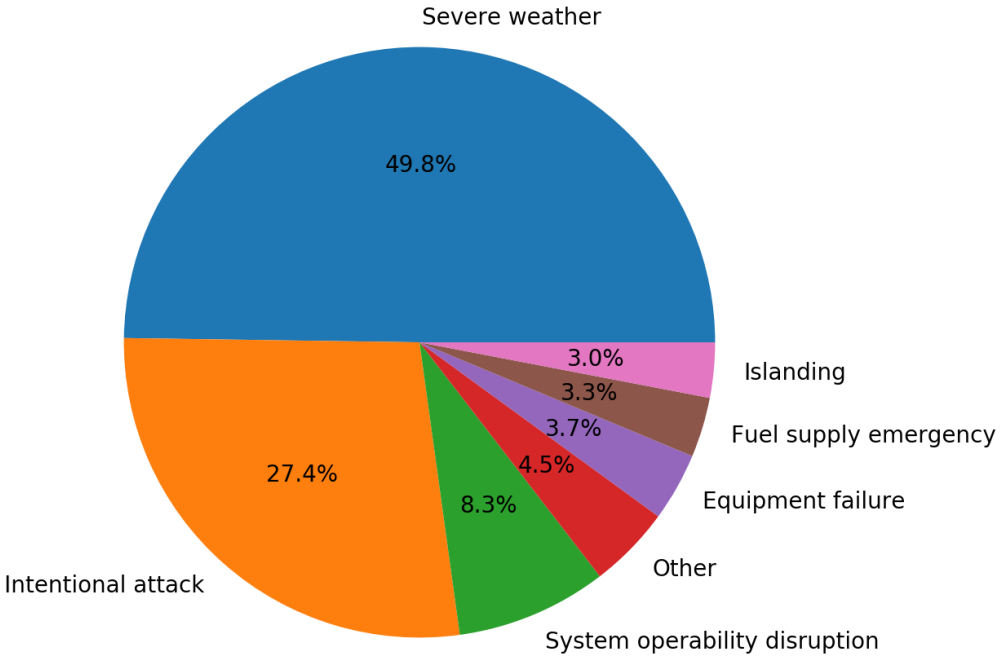
Out[114]:

	YEAR	MONTH	ANOMALY.LEVEL	OUTAGE.DURATION	DEMAND.LOSS.MW	CUST
count	1525.00000	1525.000000	1525.000000	1469.000000	681.000000	
mean	2010.17377	6.234754	-0.096852	2637.908781	649.020558	
std	3.76412	3.254510	0.739957	5953.861841	2408.040033	
min	2000.00000	1.000000	-1.600000	0.000000	0.000000	
25%	2008.00000	4.000000	-0.500000	105.000000	84.000000	
50%	2011.00000	6.000000	-0.300000	720.000000	225.000000	
75%	2013.00000	9.000000	0.300000	2880.000000	451.000000	
max	2016.00000	12.000000	2.300000	108653.000000	41788.000000	

8 rows × 45 columns

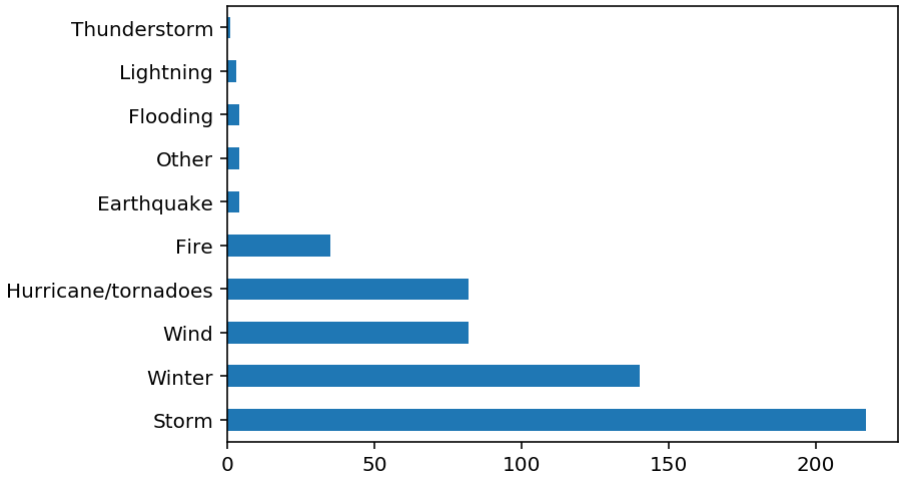
In [14]:
```
fignum = 1
# EDA : Univariate Analysis: Distribution of Cause of Events
# Conclusion: The majority major outages, 49.7%, are related to severe w
eather events
causes = 100*outages["CAUSE.CATEGORY"].value_counts() / outages["CAUSE.C
ATEGORY"].value_counts().sum()
causes.name = ""
fig = plt.figure(figsize=(3,3), dpi=200)
ax = plt.subplot(111)
causes.plot(kind='pie', ax=ax, autopct='%1.1f%%', fontsize=5)
title = plt.title("Fig. "+str(fignum)+" : "+"Distribution of Category of
Events", loc='center', pad=None)
fignum += 1
```

Fig. 1 : Distribution of Category of Events



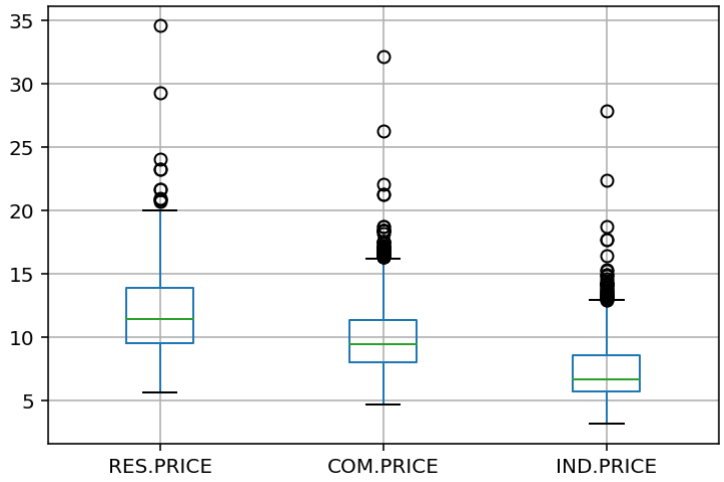
In [15]:
```
# EDA : Univariate Analysis: Natural Disaster Details Summary
# Conclusion: The majority of natural disasters are Storm and Winter, li
ghting and earthquakes are rare events
outages[outages["Natural Disaster"]]["CAUSE.CATEGORY.DETAIL"].value_coun
ts().plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Natural Disaster Details Sum
mary", loc='center', pad=None)
fignum += 1
```

Fig. 2: Natural Disaster Details Summary



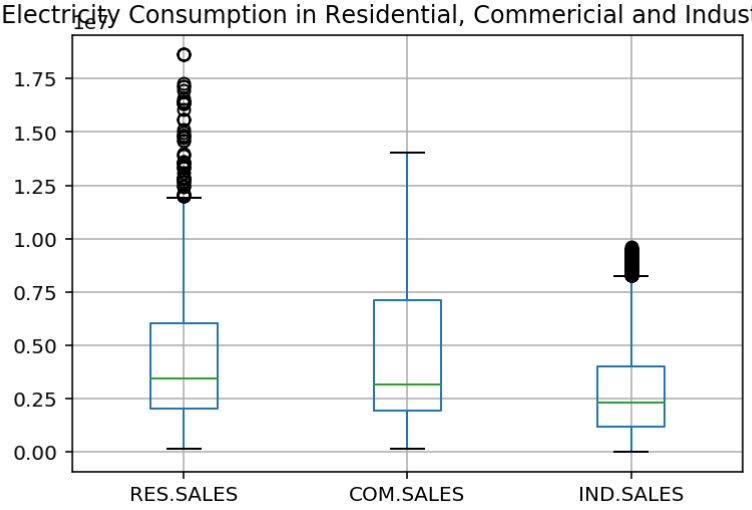
In [16]:
```
# EDA : Monthly electricity price Residential, Commericial, Industrial S
ummary
# Conclusion: Industrial price is lowest and Residential Price is higher
# There are lots of large outliers, which might imply prices are high in
unusual events?
boxplot = outages.boxplot(column=['RES.PRICE', 'COM.PRICE', 'IND.PRICE'
])
title = plt.title("Fig. "+str(fignum)+": "+"Monthly Electricity Price Re
sidential, Commericial and Industrial Sectors", loc='center', pad=None)
fignum += 1
```

Fig. 3: Monthly Electricity Price Residential, Commericial and Industrial Sectors



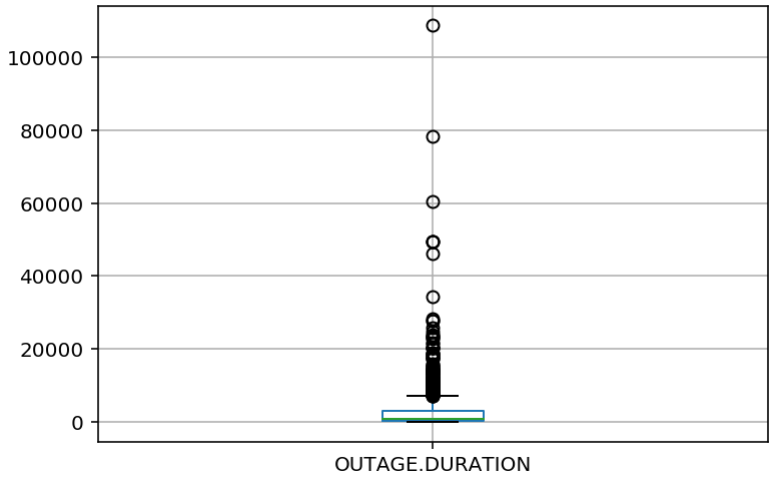
In [17]:
```
# EDA : Electricity consumption in Residential, Commericial, Industrial
 Summary
# Conclusion: Commercial comsuption is highest then followed by resednti
al and industrial
# However, there are a huge number of outliers for resendential customer
s, is this extreme weather (summer/winter) events?
boxplot = outages.boxplot(column=['RES.SALES', 'COM.SALES', 'IND.SALES'
])
title = plt.title("Fig. "+str(fignum)+": "+"Electricity Consumption in R
esidential, Commericial and Industrial Sectors", loc='center', pad=None)
fignum += 1
```

Fig. 4: Electricity Consumption in Residential, Commericial and Industrial Sectors



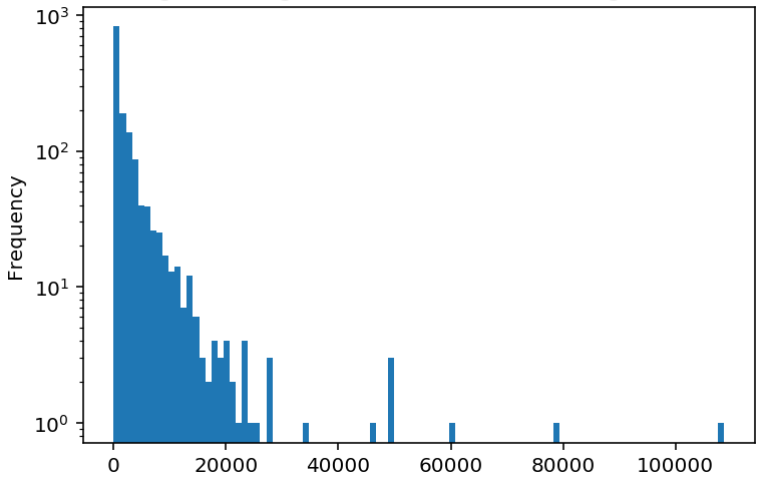
In [18]:
```
# EDA : Outage Duration Summary
# Conclusion: The duration is small but huge amount of large outliers
boxplot = outages.boxplot(column=['OUTAGE.DURATION'])
title = plt.title("Fig. "+str(fignum)+": "+"Outage Duration Summary", lo
c='center', pad=None)
fignum += 1
```

Fig. 5: Outage Duration Summary



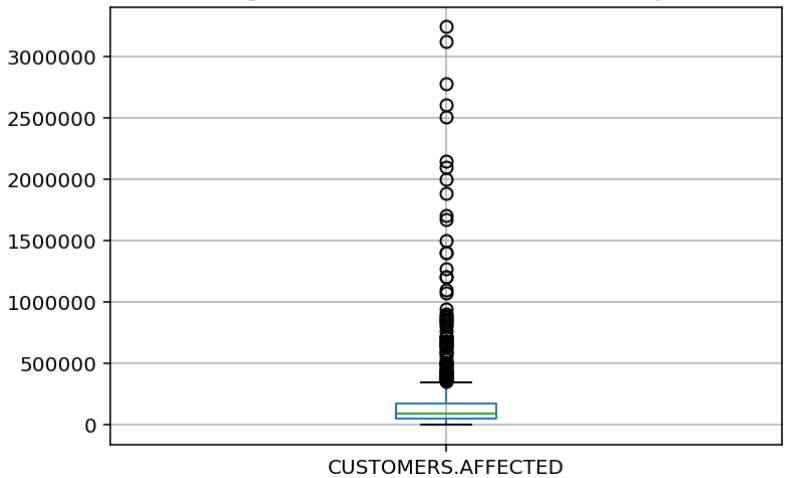
In [19]:
```
# EDA : Outage Duration Distribution (using log scale)
# Conclusion: Looking at more detail Outage Duration indeed has a lot of
outliers and is right skewed
outages['OUTAGE.DURATION'].plot(kind='hist', bins=100, logy=True)
# TODO : Plot mean
title = plt.title("Fig. "+str(fignum)+": "+"Outage Duration Distribution
(log scale)", loc='center', pad=None)
fignum += 1
```

Fig. 6: Outage Duration Distribution (log scale)



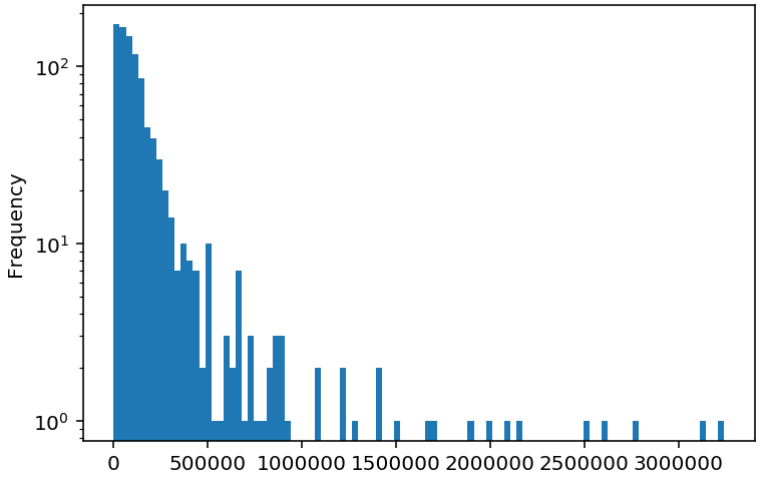
In [20]:
```
# EDA : Customers Affected Summary
# Conclusion: The Customers Affected is small but huge amount of large o
utliers
boxplot = outages.boxplot(column=['CUSTOMERS.AFFECTED'])
title = plt.title("Fig. "+str(fignum)+": "+"Customers Affected Summary",
loc='center', pad=None)
fignum += 1
```

Fig. 7: Customers Affected Summary



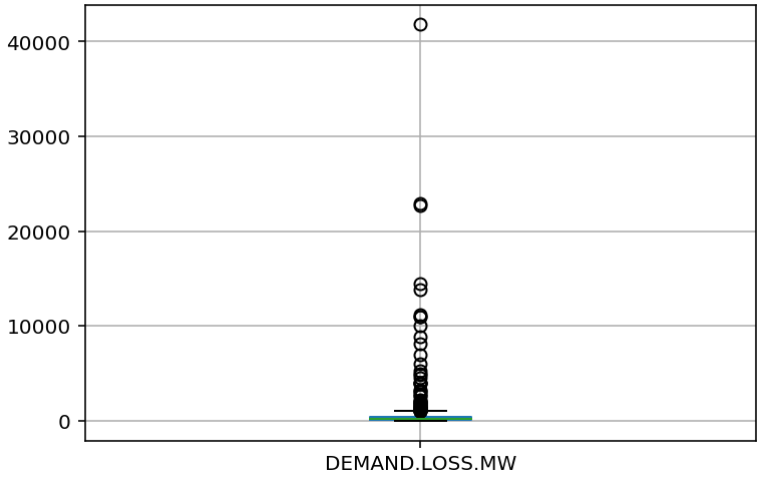
In [21]:
```
# EDA : Customers Affected Distribution (using log scale)
# Conclusion: Looking at more detail Customers Affected indeed has a lot
of outliers and is right skewed
outages['CUSTOMERS.AFFECTED'].plot(kind='hist', bins=100, log=True)
# TODO : Plot mean
title = plt.title("Fig. "+str(fignum)+": "+"Customers Affected Distribut
ion (log scale)", loc='center', pad=None)
fignum += 1
```

Fig. 8: Customers Affected Distribution (log scale)



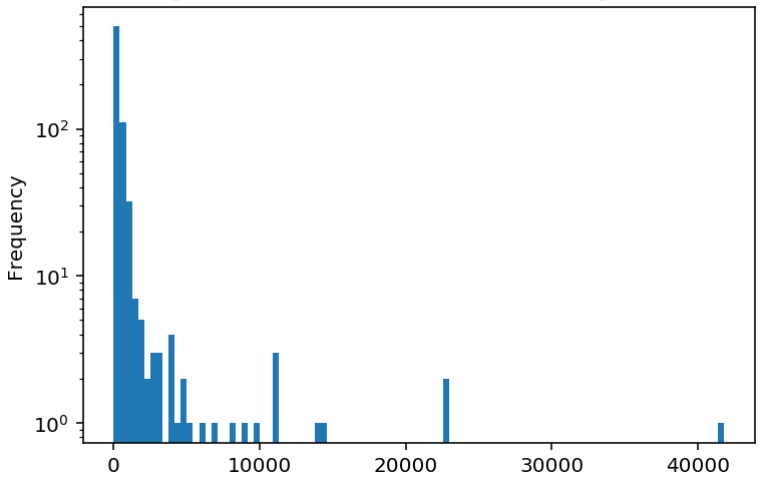
In [22]:
```
# EDA : Demand Loss Summary
# Conclusion: The Demand Loss is small but huge amount of large outliers
boxplot = outages.boxplot(column=['DEMAND.LOSS.MW'])
title = plt.title("Fig. "+str(fignum)+": "+"Demand Loss Affected Summar
y", loc='center', pad=None)
fignum += 1
```

Fig. 9: Demand Loss Affected Summary



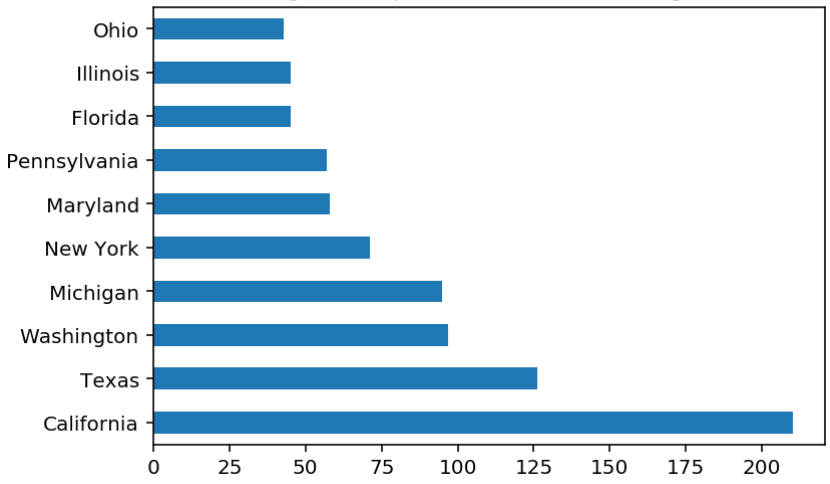
In [23]:
```
# EDA : Demand Loss Distribution (using log scale)
# Conclusion: Looking at more detail Demand Loss indeed has a lot of out
liers and is right skewed
outages['DEMAND.LOSS.MW'].plot(kind='hist', bins=100, log=True)
# TODO : Plot mean
title = plt.title("Fig. "+str(fignum)+": "+"Demand Loss Distribution (lo
g scale)", loc='center', pad=None)
fignum += 1
```

Fig. 10: Demand Loss Distribution (log scale)



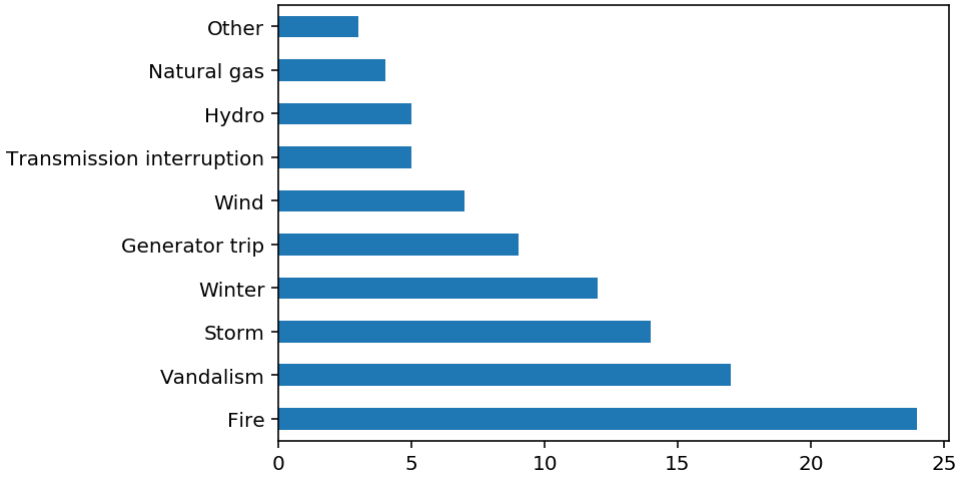
In [24]:
```
# EDA : Top Ten States with Outages (Frequency)
# Conclusion: California has the largest number of outages, followed by
 Texas
outages["U.S._STATE"].value_counts().iloc[:10].plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Top Ten States with Outages"
, loc='center', pad=None)
fignum += 1
```

Fig. 11: Top Ten States with Outages



In [25]:
```
# EDA : Top Ten States Reasons for California Outages
# Conclusion: California Outages are dominated by Fire
ca_outages = outages[outages["U.S._STATE"]=="California"]
ca_outages["CAUSE.CATEGORY.DETAIL"].value_counts().iloc[:10].plot(kind=
'barh')
title = plt.title("Fig. "+str(fignum)+": "+"Top Ten California Outage Re
asons", loc='center', pad=None)
fignum += 1
```

Fig. 12: Top Ten California Outage Reasons

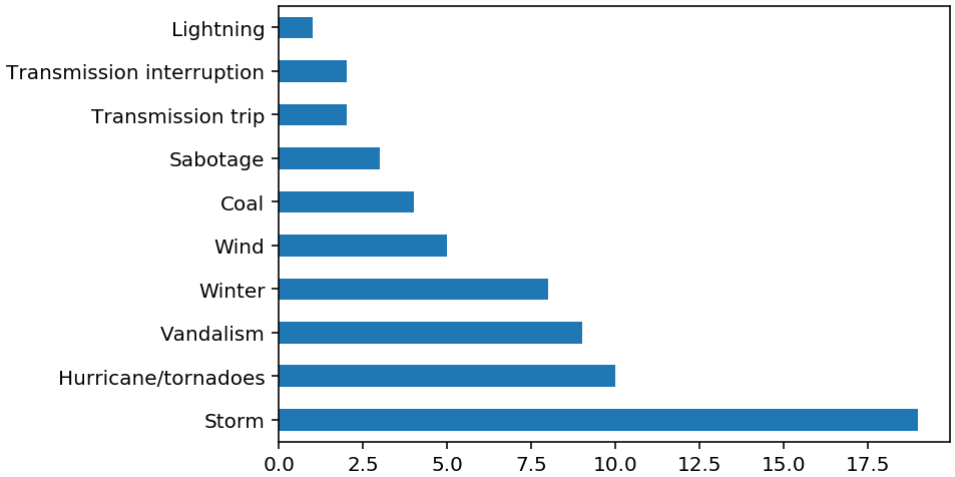


In [26]:
```
# EDA : Which Month has most Fire Outages in California
# Conclusion: September is California's worst month for Outage (frequenc
y) caused by Fire
monthinteger = ca_outages[ca_outages["CAUSE.CATEGORY.DETAIL"] == "Fire"]
["MONTH"].value_counts().iloc[0]
month = datetime.date(1900, monthinteger, 1).strftime('%B')
month
```

Out[26]:
```
'September'
```

In [27]:
```
# EDA : Top Ten Reasons for Texas Outages
# Conclusion: Texas Outages are dominated by Storm events
texas_outages = outages[outages["U.S._STATE"]=="Texas"]
texas_outages["CAUSE.CATEGORY.DETAIL"].value_counts().iloc[:10].plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Top Ten Texas Outage Reasons", loc='center', pad=None)
fignum += 1
```

Fig. 13: Top Ten Texas Outage Reasons

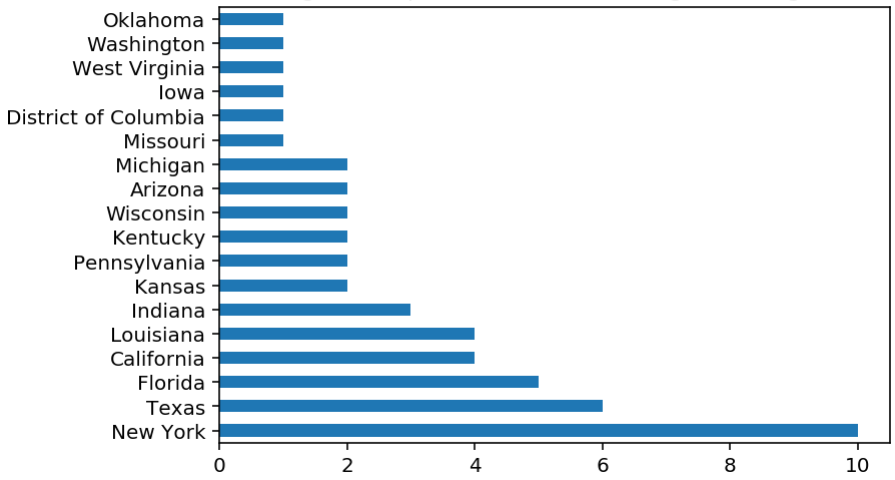


In [28]:
```
# EDA : Which Month has most Storm Outages in Texas
# Conclusion: May is Texas' worst month for Outage (frequency) caused by Storm
monthinteger = texas_outages[texas_outages["CAUSE.CATEGORY.DETAIL"] == "Storm"]["MONTH"].value_counts().iloc[0]
month = datetime.date(1900, monthinteger, 1).strftime('%B')
month
```

Out[28]: 'May'

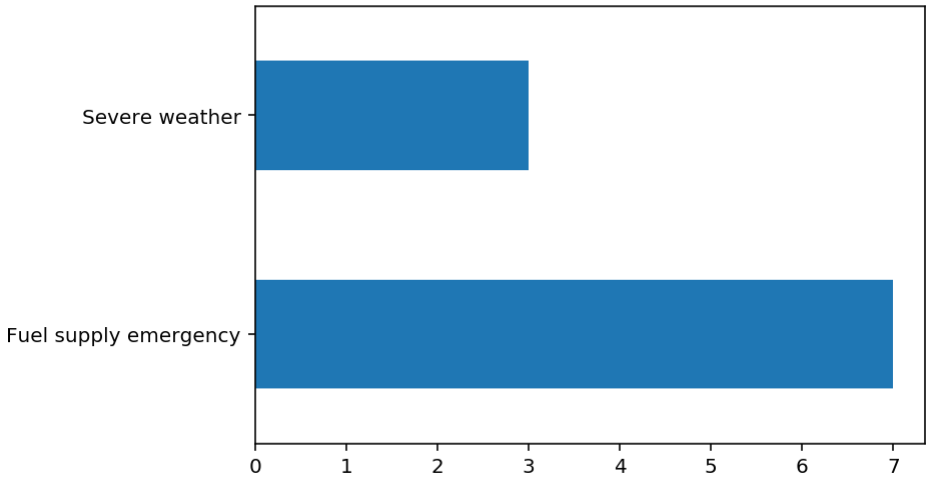
In [29]:
```
# EDA : States Distribution of Top Ten 50 Largest Outages (Duration)
# Conclusion: Paints a different pictur,  New York is has most longest d
uration outages
longest_outages = outages.sort_values("OUTAGE.DURATION", ascending=False
).iloc[:50]
longest_outages["U.S._STATE"].value_counts().plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Top Ten States with Longest
 Outages", loc='center', pad=None)
fignum += 1
```

Fig. 14: Top Ten States with Longest Outages



In [30]:
```
# EDA : Reasons for New York Longest Duration Outages
# Conclusion: New York Largest Duration Outages are dominated by Fuel su
pply emergency
ny_longest_outages = longest_outages[longest_outages["U.S._STATE"] == "N
ew York"]
ny_longest_outages["CAUSE.CATEGORY"].value_counts().plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Reasons for New Tork Longest
Duration Outages", loc='center', pad=None)
fignum += 1
```

Fig. 15: Reasons for New Tork Longest Duration Outages

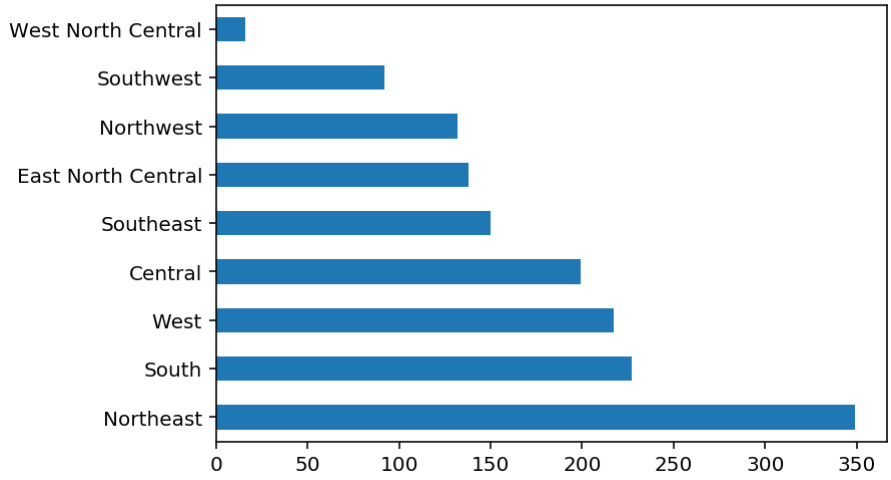


In [31]:
```
# EDA : Which Month has most Fuel Supply Emergency Outages in Texas
# Conclusion: February is Texas' worst month for Outage (duration) cause
d by Fuel Supply Emergency
monthinteger = ny_longest_outages[ny_longest_outages["CAUSE.CATEGORY"] =
= "Fuel supply emergency"]["MONTH"].value_counts().iloc[0]
month = datetime.date(1900, monthinteger, 1).strftime('%B')
month
```

Out[31]:
```
'February'
```

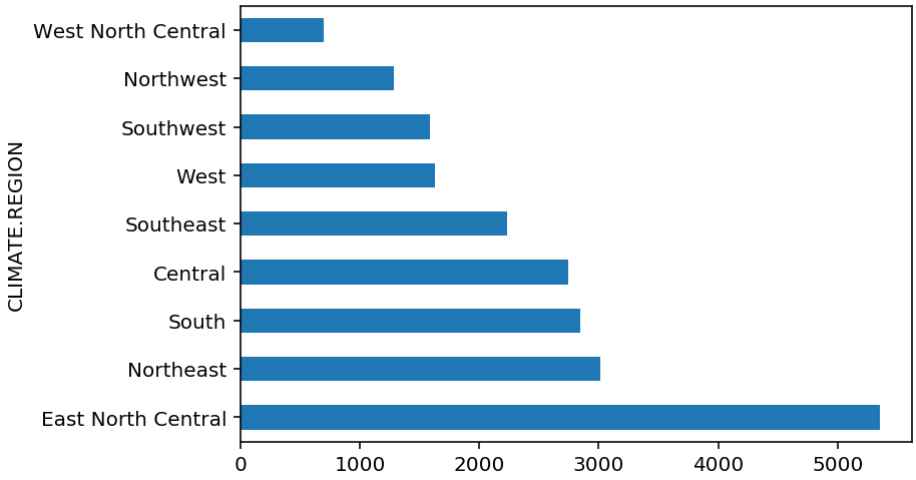
In [32]:
```
# EDA : Climate regions with Outages (Frequency)
# Conclusion: Northeast has the most outages
outages["CLIMATE.REGION"].value_counts().plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Climate Regions with Most Ou
tages", loc='center', pad=None)
fignum += 1
```

Fig. 16: Climate Regions with Most Outages



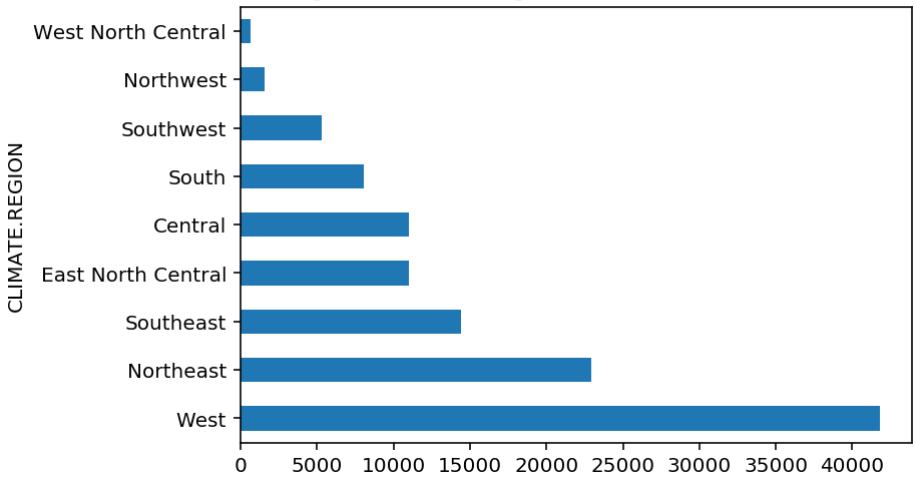
In [33]:
```
# EDA : Climate regions with Mean Outage Duration
# Conclusion: East North Central/Northeast has the most outages
outages.groupby("CLIMATE.REGION")["OUTAGE.DURATION"].mean().sort_values(
ascending=False).plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Climate Regions with Mean Du
ration Outages", loc='center', pad=None)
fignum += 1
```

Fig. 17: Climate Regions with Mean Duration Outages



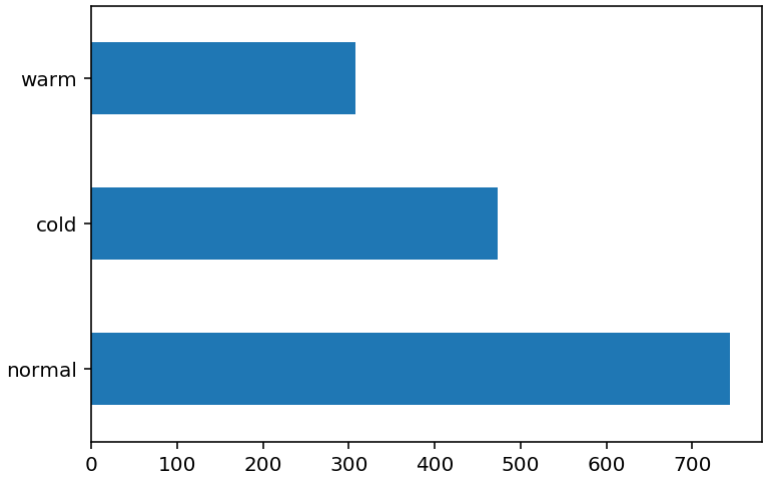
In [34]:
```
# EDA : Climate regions with Max Demand Loss
# Conclusion: However, West region has larger MW loss
outages.groupby("CLIMATE.REGION")["DEMAND.LOSS.MW"].max().sort_values(as
cending=False).plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Climate regions with Max Dem
and Loss", loc='center', pad=None)
fignum += 1
```

Fig. 18: Climate regions with Max Demand Loss



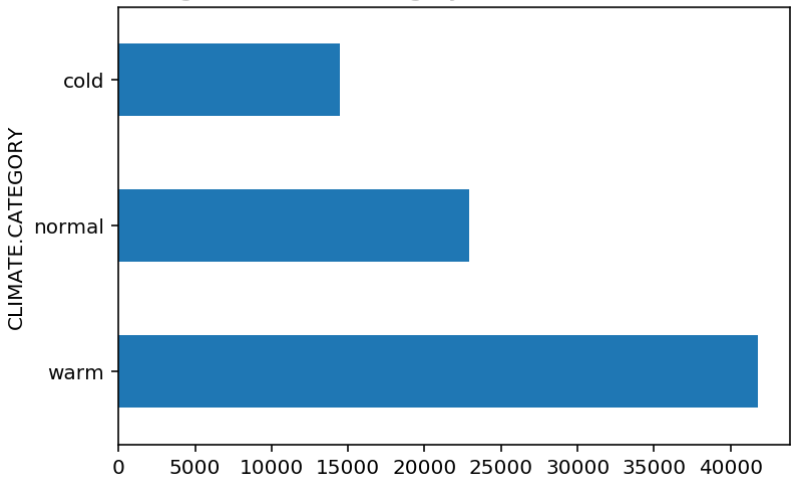
In [35]:
```
# EDA : Climate category with Outages (Frequency)
# Conclusion: Normal climate has the most
outages["CLIMATE.CATEGORY"].value_counts().plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Climate category with Most O
utages", loc='center', pad=None)
fignum += 1
```

Fig. 19: Climate category with Most Outages



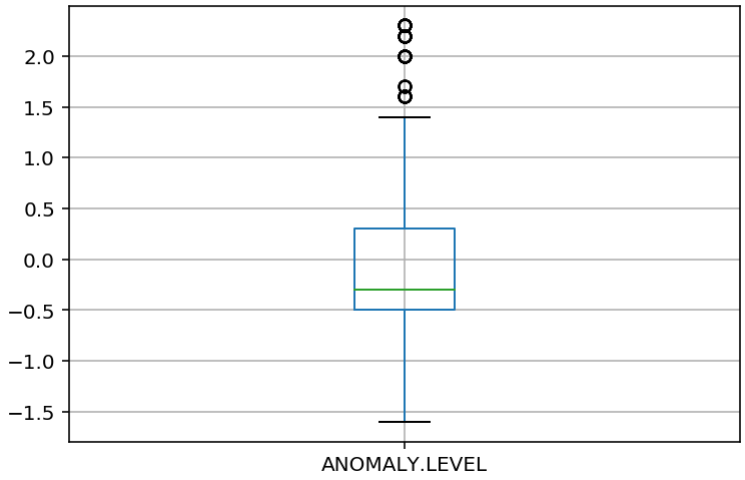
In [36]:
```
# EDA : Climate category with Max Demand Loss
# Conclusion: Warm climate has the most
outages.groupby("CLIMATE.CATEGORY")["DEMAND.LOSS.MW"].max().sort_values(
ascending=False).plot(kind='barh')
title = plt.title("Fig. "+str(fignum)+": "+"Climate category with Max De
mand Loss", loc='center', pad=None)
fignum += 1
```

Fig. 20: Climate category with Max Demand Loss



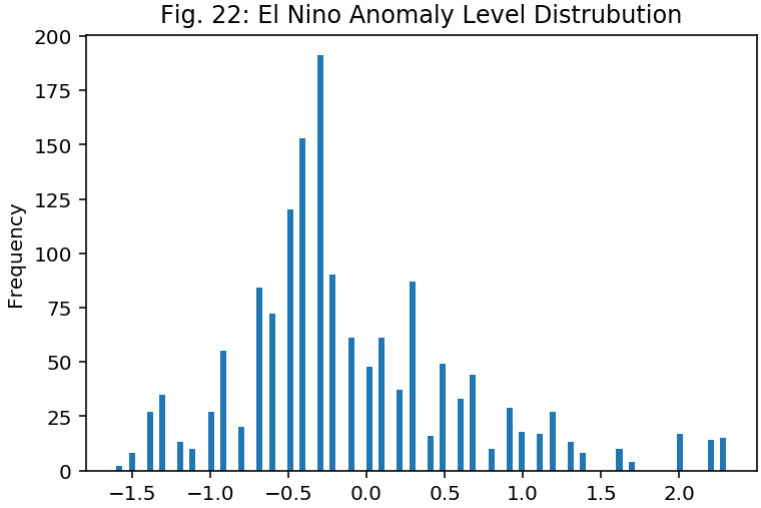
In [37]:
```
# EDA : El Nino Anomaly Level Distrubution
# Conclusion: Expected to be centered around zero, but long tail and whi
skers implying unusual weather events
boxplot = outages.boxplot(column=['ANOMALY.LEVEL'])
title = plt.title("Fig. "+str(fignum)+": "+"El Nino Anomaly Level Distru
bution", loc='center', pad=None)
fignum += 1
```

Fig. 21: El Nino Anomaly Level Distrubution



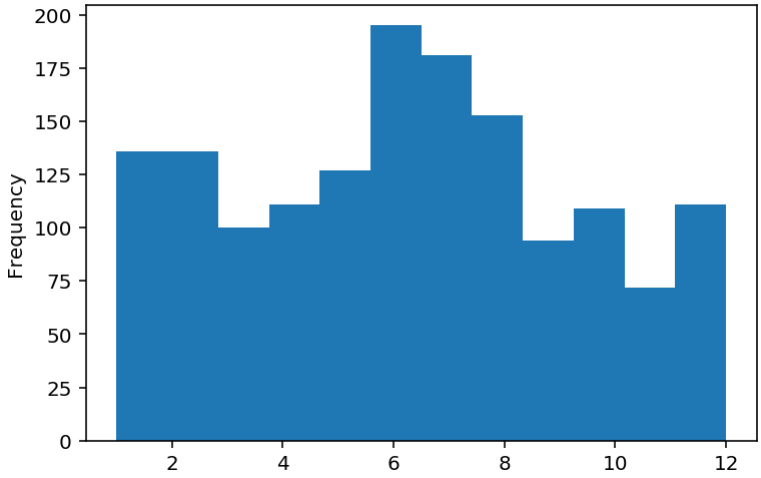
In [38]:
```
# EDA : El Nino Anomaly Level Distrubution
# Conclusion: Expected to be centered around zero, but long tail and whi
skers implying unusual weather events
outages['ANOMALY.LEVEL'].plot(kind='hist', bins=100)
# TODO : Plot mean
title = plt.title("Fig. "+str(fignum)+": "+"El Nino Anomaly Level Distru
bution", loc='center', pad=None)
fignum += 1
```

Fig. 22: El Nino Anomaly Level Distrubution



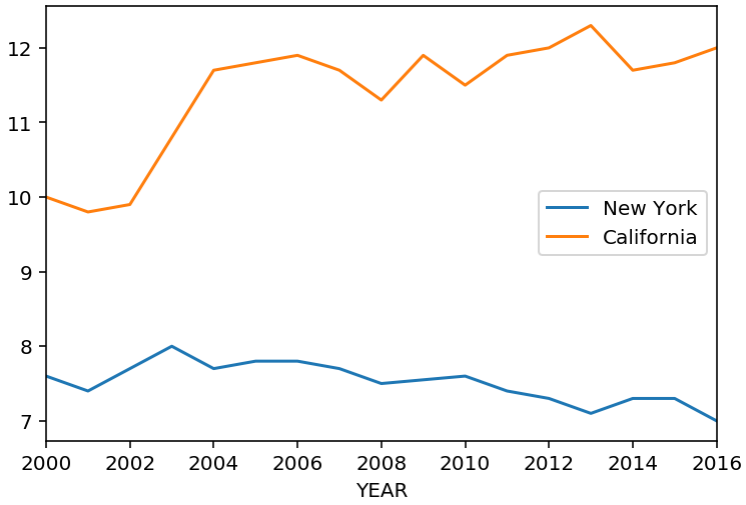
In [39]:
```
# EDA : Outage Frequency each Month
# Conclusion: Most frequent outages are in the summer
outages['MONTH'].plot(kind='hist', bins=12)
# TODO : Change Labels
title = plt.title("Fig. "+str(fignum)+": "+"Outage Frequency each Month"
, loc='center', pad=None)
fignum += 1
```

Fig. 23: Outage Frequency each Month



In [40]:
```
# EDA : State utility sector's income New York vs California over time
# Conclusion: New Tork has fallen income but California has increased
outages[outages["U.S._STATE"]== "New York"].groupby("YEAR")["PI.UTIL.OFU
SA"].mean().plot()
outages[outages["U.S._STATE"]== "California"].groupby("YEAR")["PI.UTIL.O
FUSA"].mean().plot()
title = plt.title("Fig. "+str(fignum)+": "+"State utility sector's incom
e New York vs California", loc='center', pad=None)
plt.legend(["New York", "California"])
fignum += 1
```

Fig. 24: State utility sector's income New York vs California



In [41]:
```
# EDA : Scatter Matrix Plot
# Conclusion: Duration has interesting link to the Per capita GSP, Urban Population and Residential Price etc
p = pd.plotting.scatter_matrix(outages[['OUTAGE.DURATION', 'DEMAND.LOSS.MW', 'CUSTOMERS.AFFECTED', 'PC.REALGSP.STATE', 'POPPCT_URBAN', 'RES.PRICE', 'ANOMALY.LEVEL']], figsize=(15,15))
title = plt.title("Fig. "+str(fignum)+": "+"EDA Scatter Matrix Plot", loc='center', pad=None)
fignum += 1
```

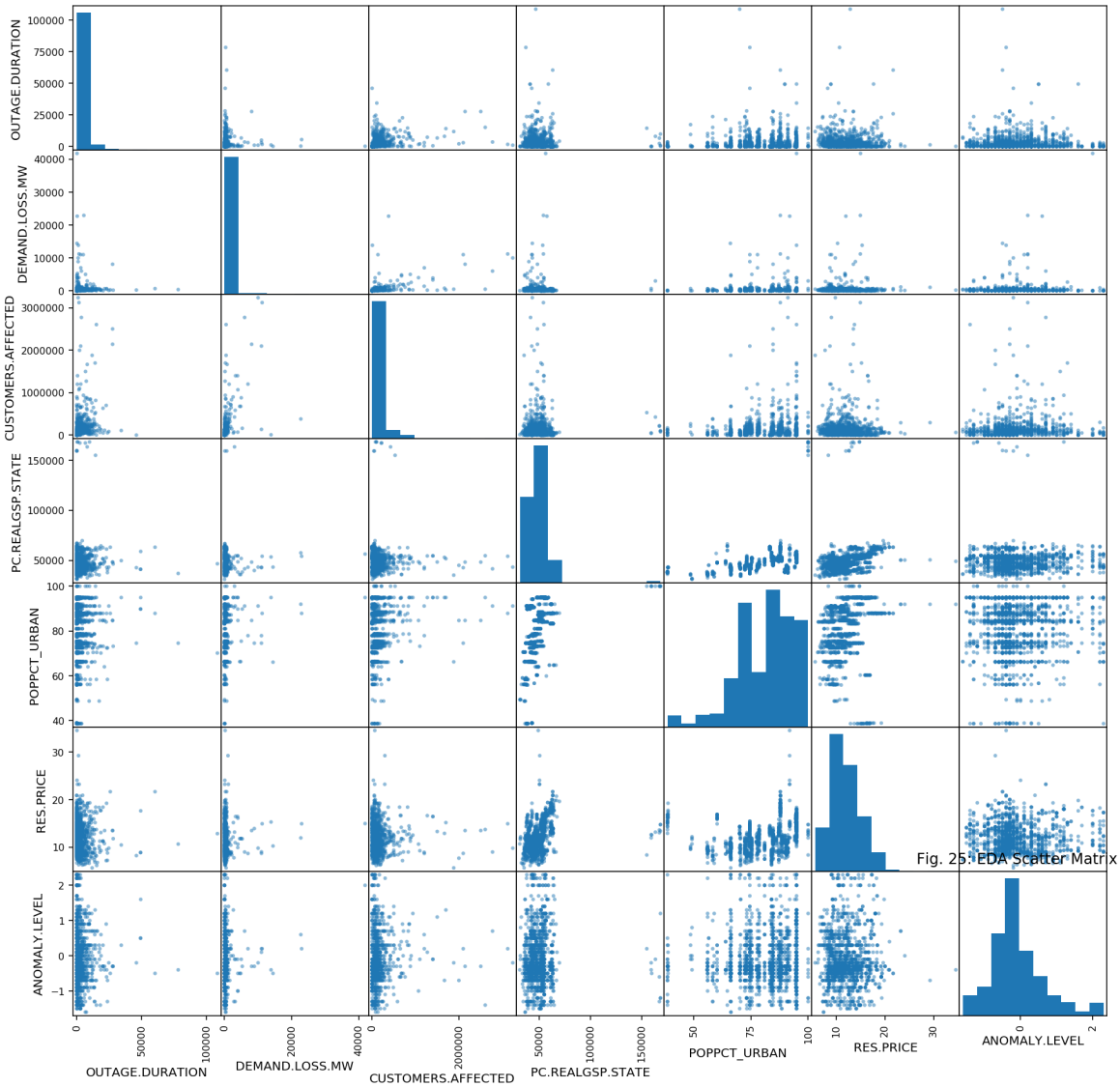
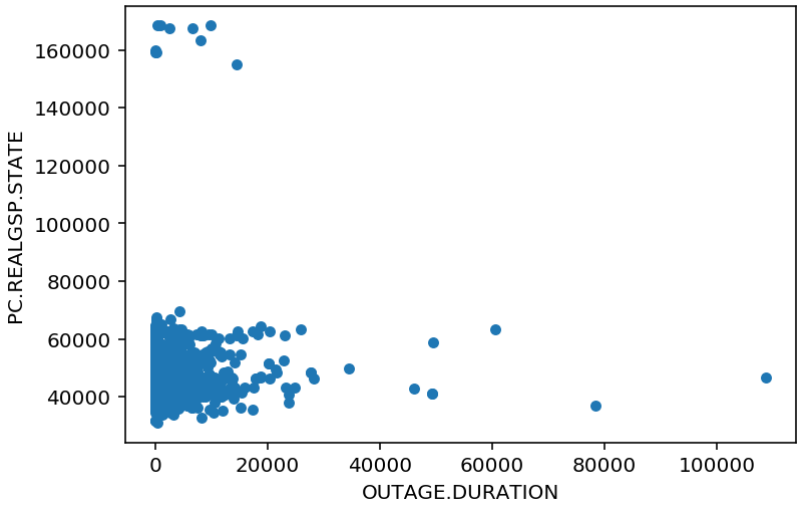


Fig. 25: EDA Scatter Matrix Plot

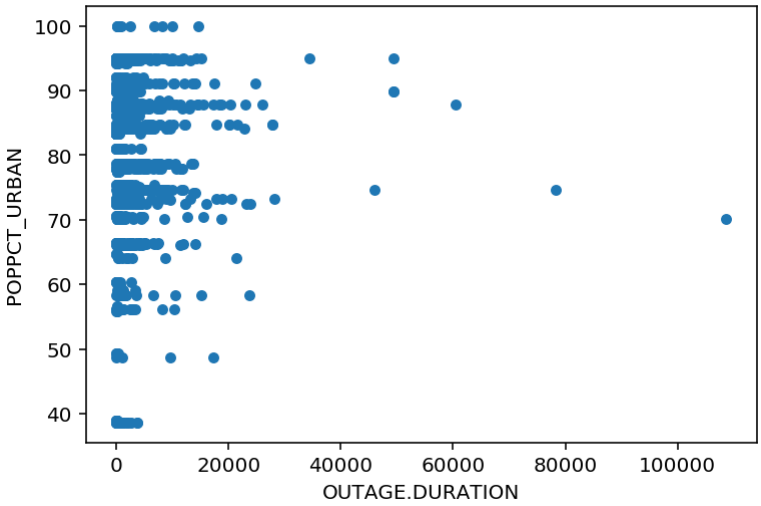
In [42]:
```
# EDA : Scatter Plot of Duration vs State GSP
# Conclusion: Outliers show short duration when State GSP is high and lo
ng durations when GSP is low
p = outages.plot(kind='scatter', x='OUTAGE.DURATION', y='PC.REALGSP.STAT
E')
title = plt.title("Fig. "+str(fignum)+": "+"Outage Duration vs State GS
P", loc='center', pad=None)
fignum += 1
```

Fig. 26: Outage Duration vs State GSP



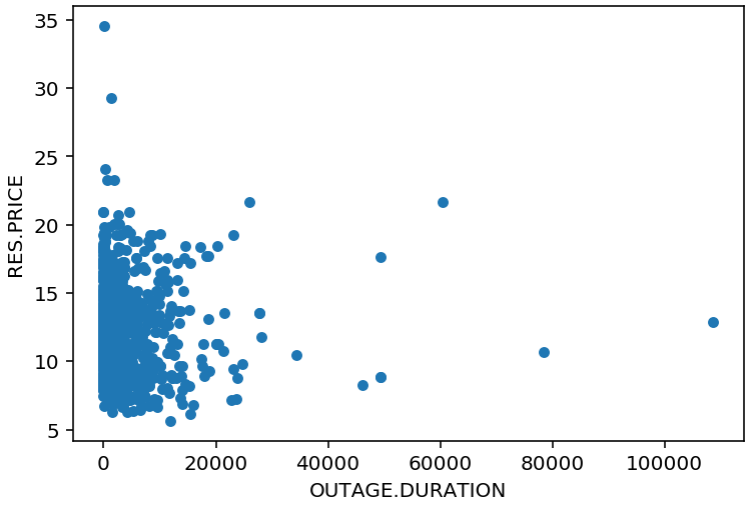
In [43]:
```
# EDA : Scatter Plot of Duration vs Urban Population
# Conclusion: Short duration outages when Urban Population  is High
#             Long durations outliers when Urban Population is low (high
er rural population)
p = outages.plot(kind='scatter', x='OUTAGE.DURATION', y='POPPCT_URBAN')
title = plt.title("Fig. "+str(fignum)+": "+"Outage Duration vs Urban Pop
ulation", loc='center', pad=None)
fignum += 1
```

Fig. 27: Outage Duration vs Urban Population



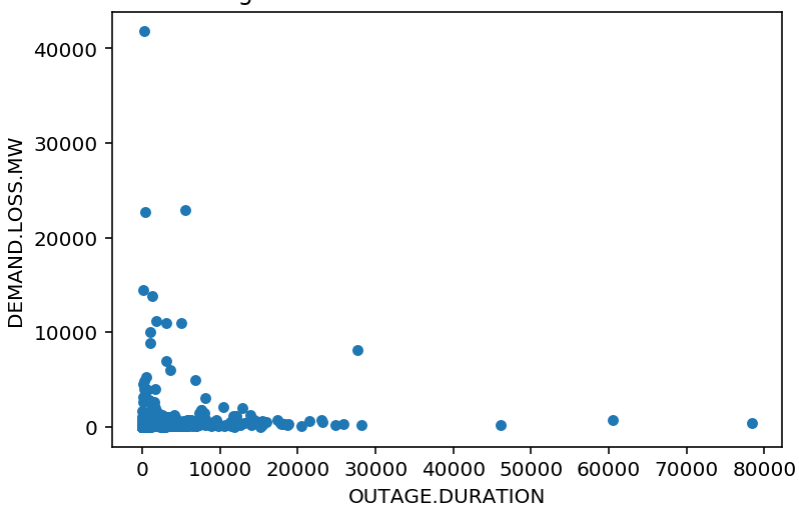
In [44]:
```
# EDA : Scatter Plot of Duration vs Residential Price
# Conclusion: Short duration outages when Residential Price  is Low
#             Long durations outliers when Residential Price is Higher
p = outages.plot(kind='scatter', x='OUTAGE.DURATION', y='RES.PRICE')
title = plt.title("Fig. "+str(fignum)+": "+"Outage Duration vs Residenti
al Price", loc='center', pad=None)
fignum += 1
```

Fig. 28: Outage Duration vs Residential Price



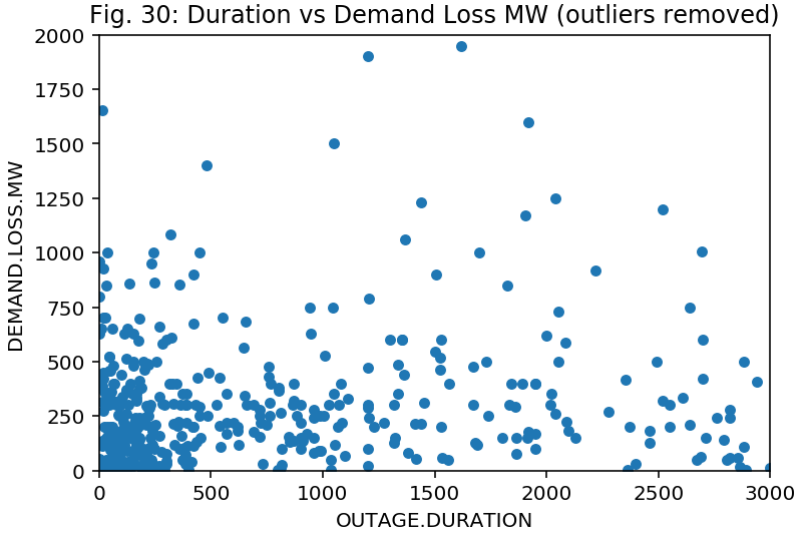
In [45]:
```
# EDA : Scatter Plot of Duration vs Demand Loss MW
# Conclusion: There is a trend but the outliers are short duration with
 large demand loss (cities)
#             Long durations with small demand loss (rural)
p = outages.plot(kind='scatter', x='OUTAGE.DURATION', y='DEMAND.LOSS.MW'
)
title = plt.title("Fig. "+str(fignum)+": "+"Duration vs Demand Loss MW",
loc='center', pad=None)
fignum += 1
```

Fig. 29: Duration vs Demand Loss MW



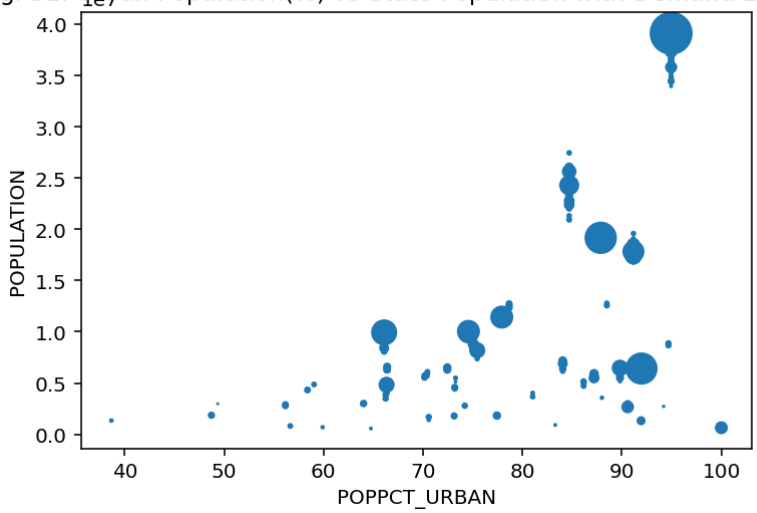
In [46]:
```
# EDA : Scatter Plot of Duration vs Demand Loss MW, outliers removed
# Conclusion: With outliers removed there is a little more of an obvious positive trend
p = outages.plot(kind='scatter', x='OUTAGE.DURATION', y='DEMAND.LOSS.MW')
plt.xlim(0, 3000)
plt.ylim(0, 2000)
title = plt.title("Fig. "+str(fignum)+": "+"Duration vs Demand Loss MW (outliers removed)", loc='center', pad=None)
fignum += 1
```

Fig. 30: Duration vs Demand Loss MW (outliers removed)



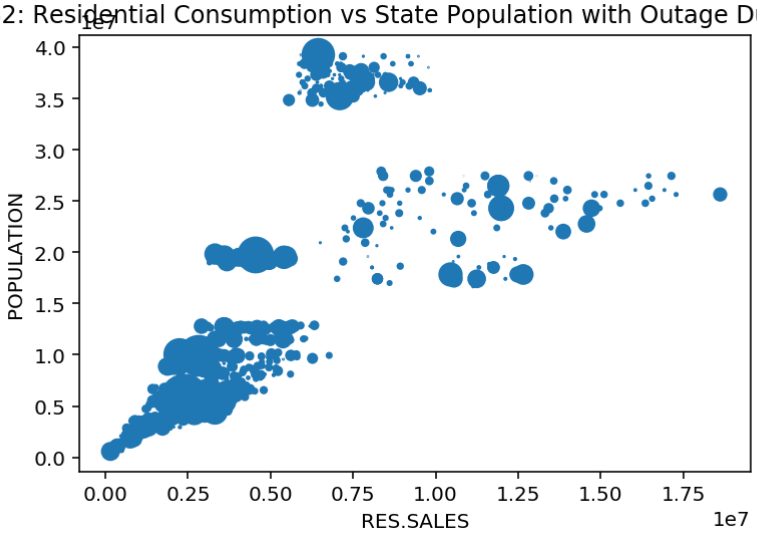
In [47]:
```
# EDA : Scatter Plot of Urban Population(%) vs State Population with Dem and Loss Size
# Conclusion: Positive trend and demand loss is also higher.
p = outages.plot(kind='scatter', x='POPPCT_URBAN', y='POPULATION', s=outages['DEMAND.LOSS.MW'].abs()/100)
title = plt.title("Fig. "+str(fignum)+": "+"Urban Population(%) vs State Population with Demand Loss Size", loc='center', pad=None)
fignum += 1
```

Fig. 31: Urban Population(%) vs State Population with Demand Loss Size



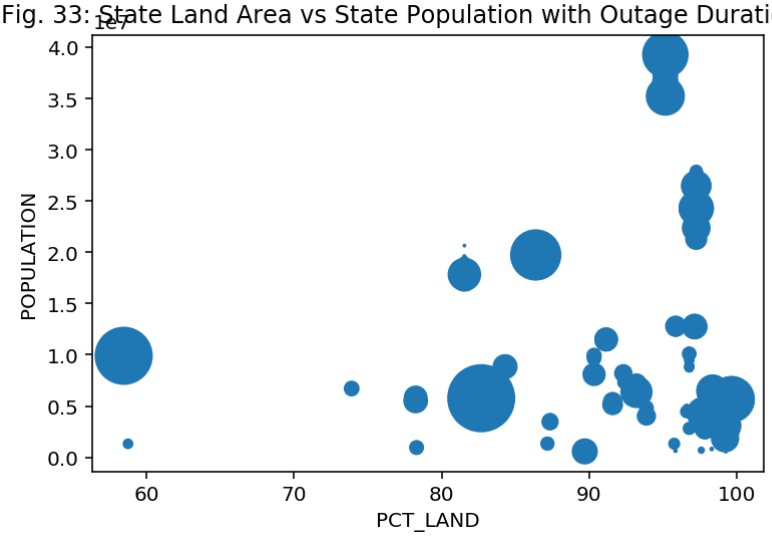
In [48]:
```
# EDA : Scatter Plot of Residential Consumption vs State Population with
Outage Duration size
# Conclusion: Positive trend and outage duration tends also higher.
p = outages.plot(kind='scatter', x='RES.SALES', y='POPULATION', s=outage
s['OUTAGE.DURATION'].abs()/200)
title = plt.title("Fig. "+str(fignum)+": "+"Residential Consumption vs S
tate Population with Outage Duration size", loc='center', pad=None)
fignum += 1
```

Fig. 32: Residential Consumption vs State Population with Outage Duration size



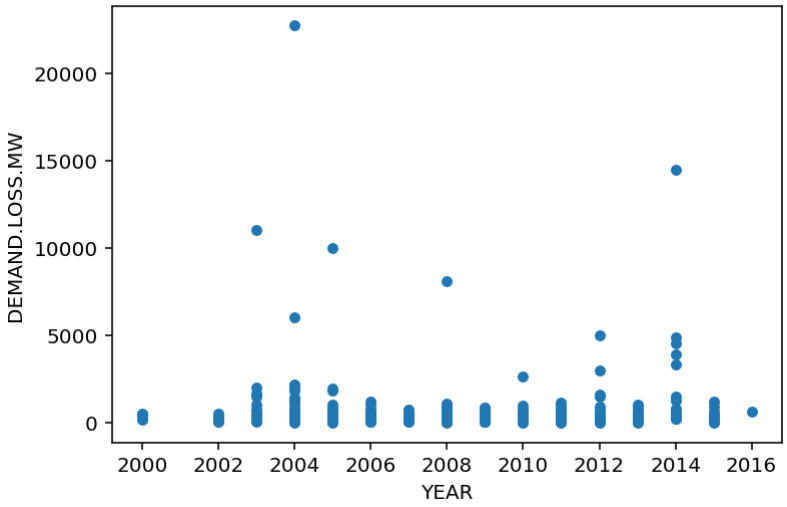
In [49]:
```
# EDA : Scatter Plot of State Land Area vs State Population with Outage
 Duration size
# Conclusion: Still a positive trend, the duration sizes do though vary
 in low and high percentage of land usage
p = outages.plot(kind='scatter', x='PCT_LAND', y='POPULATION', s=outages
['OUTAGE.DURATION'].abs()/100)
title = plt.title("Fig. "+str(fignum)+": "+"State Land Area vs State Pop
ulation with Outage Duration size", loc='center', pad=None)
fignum += 1
```

Fig. 33: State Land Area vs State Population with Outage Duration size



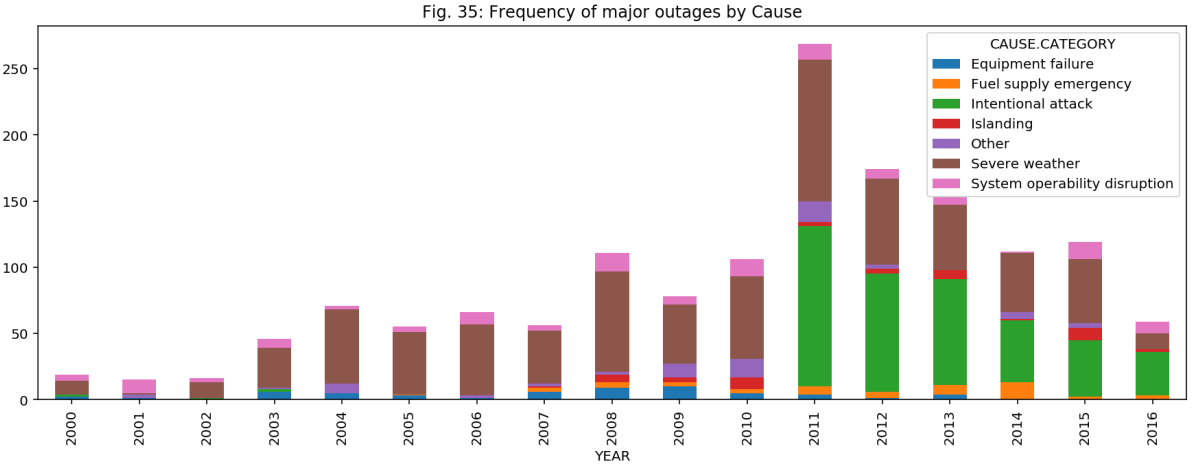
In [50]:
```
# EDA : Scatter Plot of Natural Disaster year vs Demand Loss
# Conclusion: Not a clear trend, this is something that needs more categ
orical analysis
p = outages[outages["Natural Disaster"]].plot(kind='scatter', x='YEAR',
y='DEMAND.LOSS.MW')
title = plt.title("Fig. "+str(fignum)+": "+"Natural Disaster Demand Loss
vs year", loc='center', pad=None)
fignum += 1
```

Fig. 34: Natural Disaster Demand Loss vs year



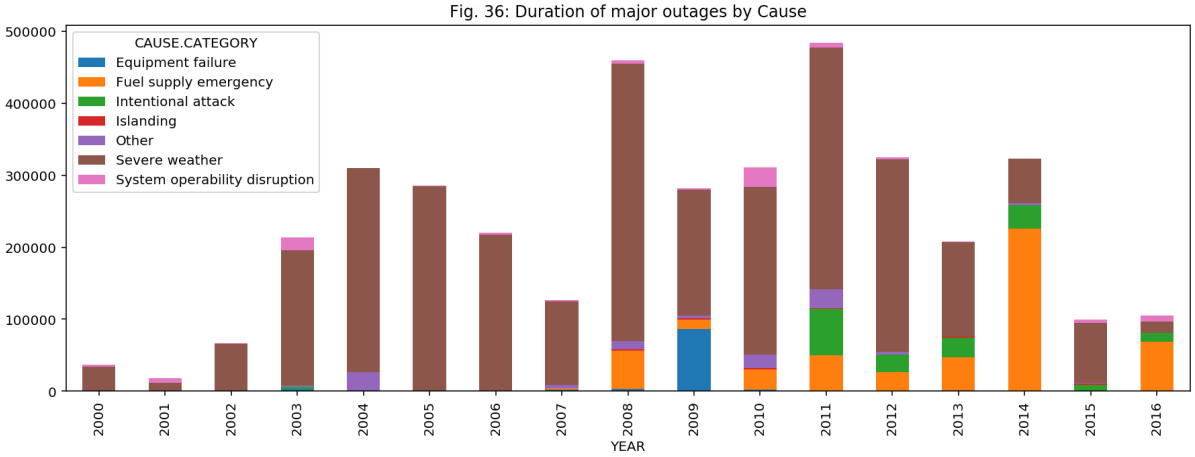
In [51]:
```
# EDA : Frequency of major outages by Cause
# Conclusion: International (terrorist) attacks substantially increased
 in 2011, almost certainly related to 9/11
df = outages.pivot_table(index="YEAR", columns="CAUSE.CATEGORY", values=
"MONTH", aggfunc="count", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Frequency of major outages b
y Cause", loc='center', pad=None)
fignum += 1
```

Fig. 35: Frequency of major outages by Cause



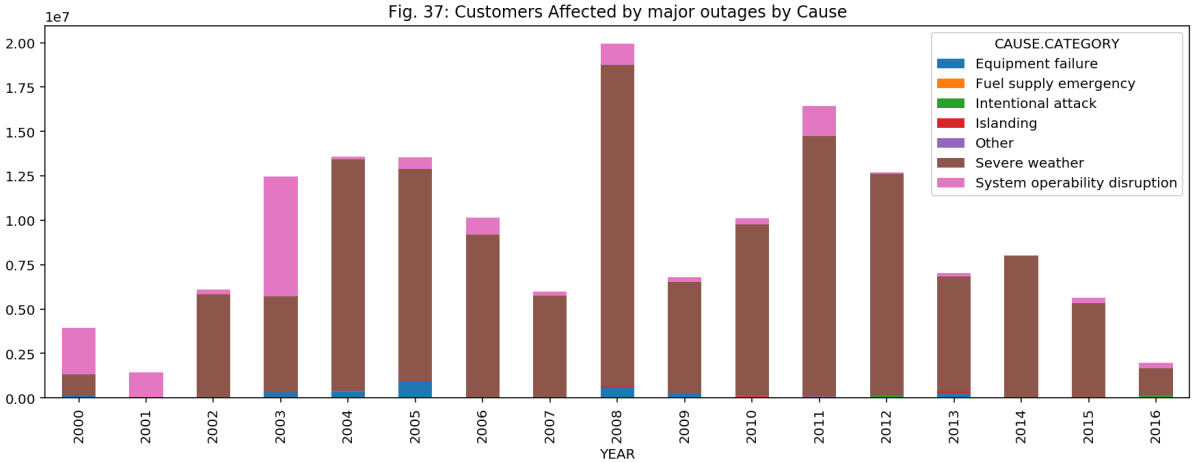
In [52]:
```
# EDA : Duration of major outages by Cause
# Conclusion: Fuel Supply was a major cause of duration outages in 2014
df = outages.pivot_table(index="YEAR", columns="CAUSE.CATEGORY", values=
"OUTAGE.DURATION", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Duration of major outages by
Cause", loc='center', pad=None)
fignum += 1
```

Fig. 36: Duration of major outages by Cause



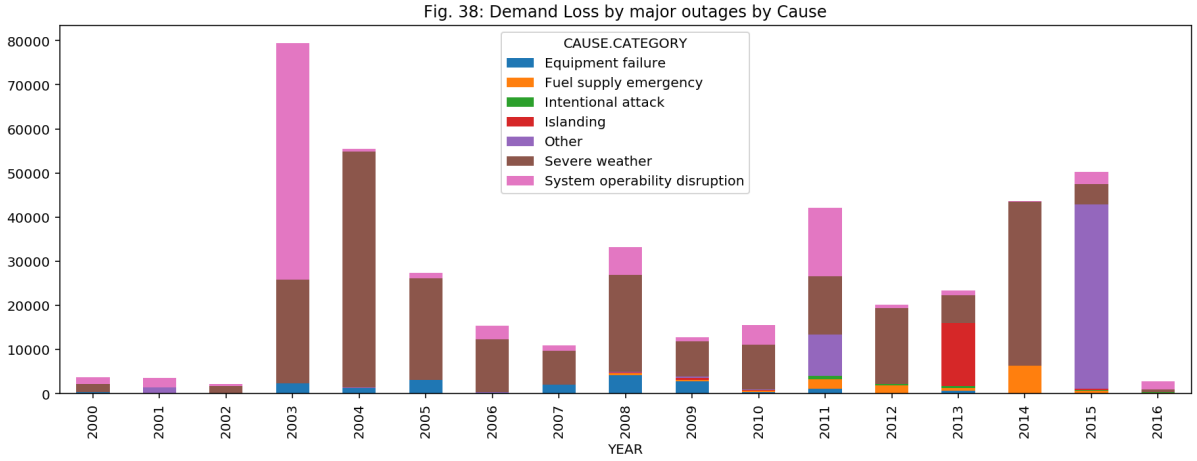
In [53]:
```
# EDA : Customers Affected by major outages by Cause
# Conclusion: Severe Weather is the major cause for largest customers af
fected events
df = outages.pivot_table(index="YEAR", columns="CAUSE.CATEGORY", values=
"CUSTOMERS.AFFECTED", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Customers Affected by major
 outages by Cause", loc='center', pad=None)
fignum += 1
```

Fig. 37: Customers Affected by major outages by Cause



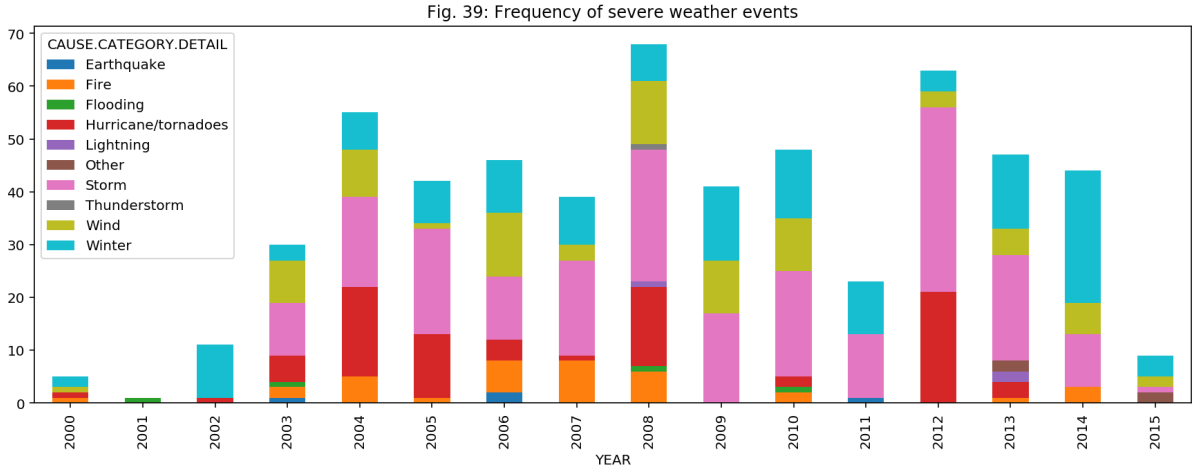
In [54]:
```
# EDA : Demand Loss by major outages by Cause
# Conclusion: Severe Weather is the major cause for largest customers af
fected events
df = outages.pivot_table(index="YEAR", columns="CAUSE.CATEGORY", values=
"DEMAND.LOSS.MW", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Demand Loss by major outages
by Cause", loc='center', pad=None)
fignum += 1
```

Fig. 38: Demand Loss by major outages by Cause



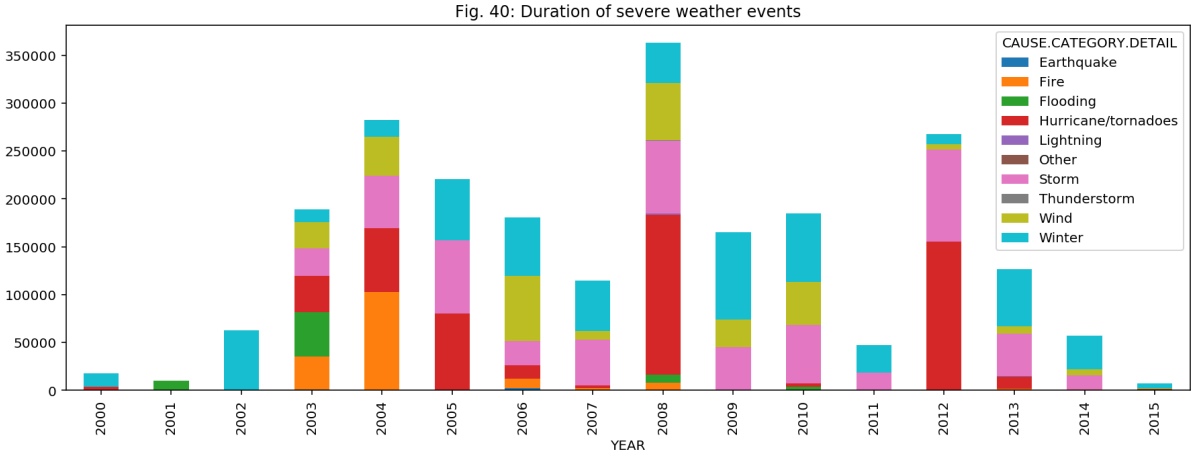
In [55]:
```
# EDA : Frequency of severe weather events
# Conclusion: Storms have been the largest cause with 25 cases reported
 in 2008.
df = outages[outages["Natural Disaster"]]
df = df.pivot_table(index="YEAR", columns="CAUSE.CATEGORY.DETAIL", value
s="MONTH", aggfunc="count", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Frequency of severe weather
 events", loc='center', pad=None)
fignum += 1
```

Fig. 39: Frequency of severe weather events



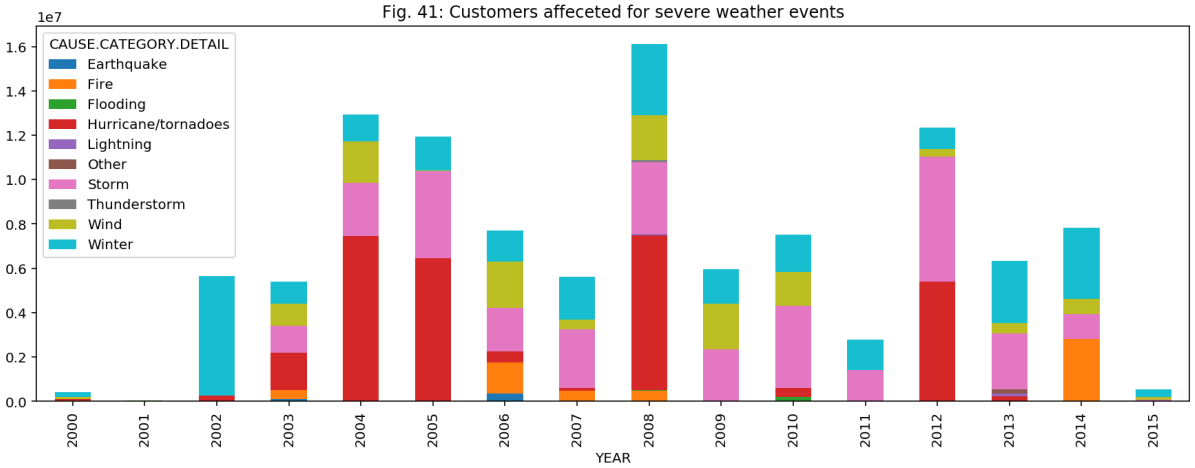
In [56]:
```
# EDA : Duration of severe weather events
# Conclusion: Hurricane/tornadoes have been the largest cause 2018 as th
e year with the highest loss because of these events with over 166 days
 lost.
df = outages[outages["Natural Disaster"]]
df = df.pivot_table(index="YEAR", columns="CAUSE.CATEGORY.DETAIL", value
s="OUTAGE.DURATION", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Duration of severe weather e
vents", loc='center', pad=None)
fignum += 1
```

Fig. 40: Duration of severe weather events



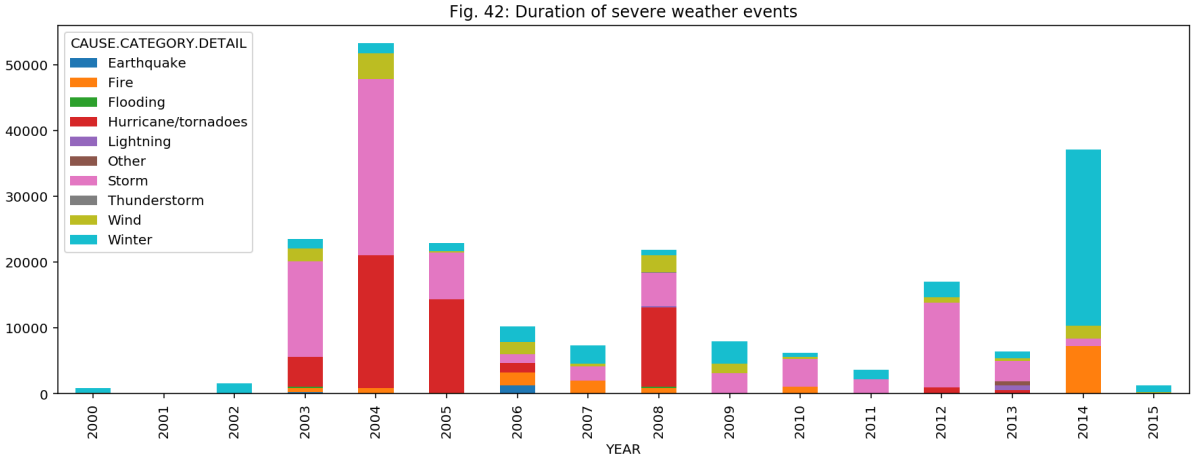
In [57]:
```
# EDA : Customers affeceted for severe weather events
# Conclusion: Earthquakes are seldom a cause of major outages, 2008 was
 the largest customer (350886) impact due to earthquakes.
df = outages[outages["Natural Disaster"]]
df = df.pivot_table(index="YEAR", columns="CAUSE.CATEGORY.DETAIL", value
s="CUSTOMERS.AFFECTED", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Customers affeceted for seve
re weather events", loc='center', pad=None)
fignum += 1
```

Fig. 41: Customers affeceted for severe weather events



In [58]:
```
# EDA : Duration of severe weather events
# Conclusion: Winter 2014 was a major impact with 18 days lost.
df = outages[outages["Natural Disaster"]]
df = df.pivot_table(index="YEAR", columns="CAUSE.CATEGORY.DETAIL", value
s="DEMAND.LOSS.MW", aggfunc="sum", fill_value=0)
df.plot(kind="bar", figsize=(15,5), stacked=True)
title = plt.title("Fig. "+str(fignum)+": "+"Duration of severe weather e
vents", loc='center', pad=None)
fignum += 1
```

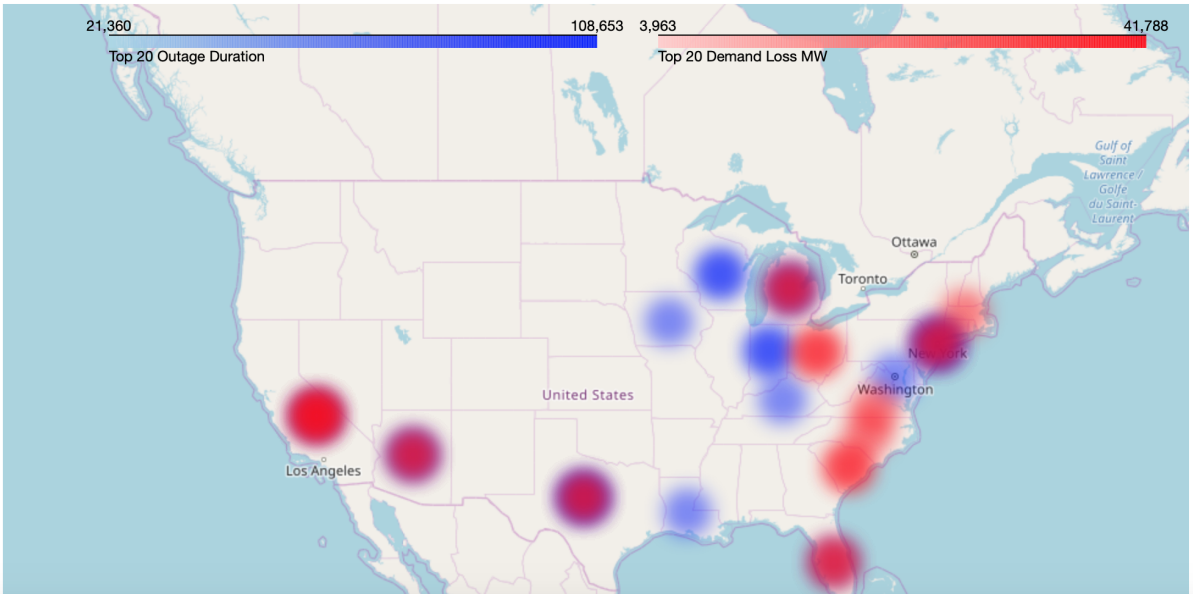
Fig. 42: Duration of severe weather events



In [139]:
```
# EDA: Top 20 Outages vs Top 20 Demand Loss
# Conclusion: We see the top 20 outages based on demand loss are in the
 west and north east.
# However, the midwest is a region with impact due to outage duration.
# Texas also has been impacted in both duration and demand loss.
heat_map = folium.Map(location=[39.8283, -98.5795], zoom_start=4.3, zoom
Control= False)
util.show_geospatial_1(heat_map, outages)
```

Out[139]:

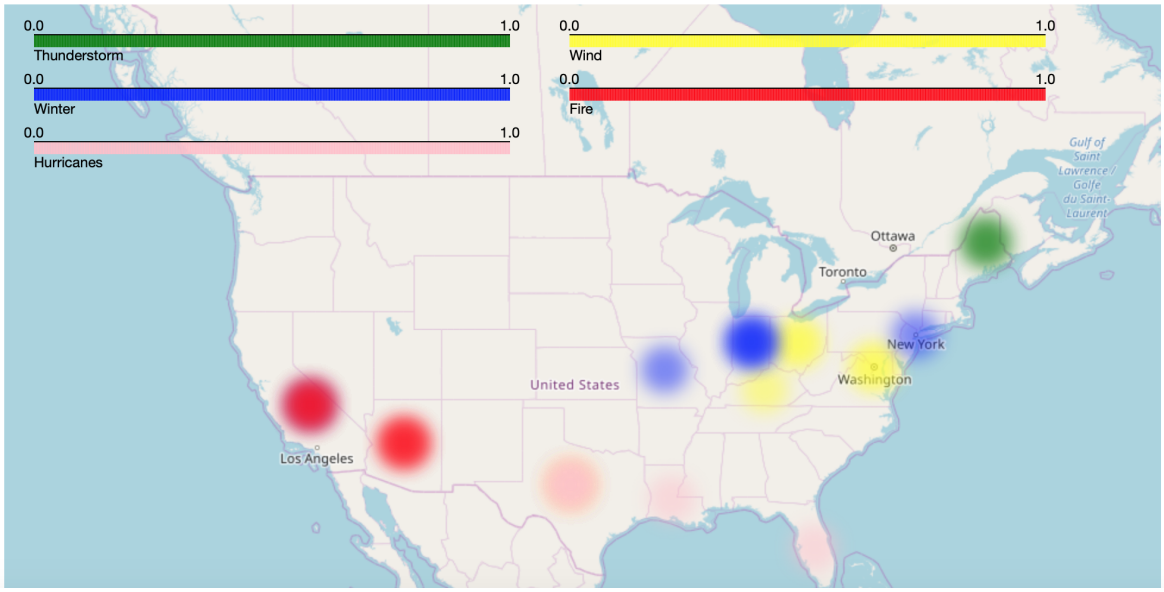
Fig. 43: Top 20 Outages vs Top 20 Demand Loss



In [140]:
```
# EDA: Top 5 Longest Duration Outages by Severe Event Type
# Conclusion: We can see fire events in the west and wind and winter eve
nts in the north east
heat_map = folium.Map(location=[39.8283, -98.5795], zoom_start=4.3, zoom
Control= False)
util.show_geospatial_2(heat_map, outages)
```

Out[140]:

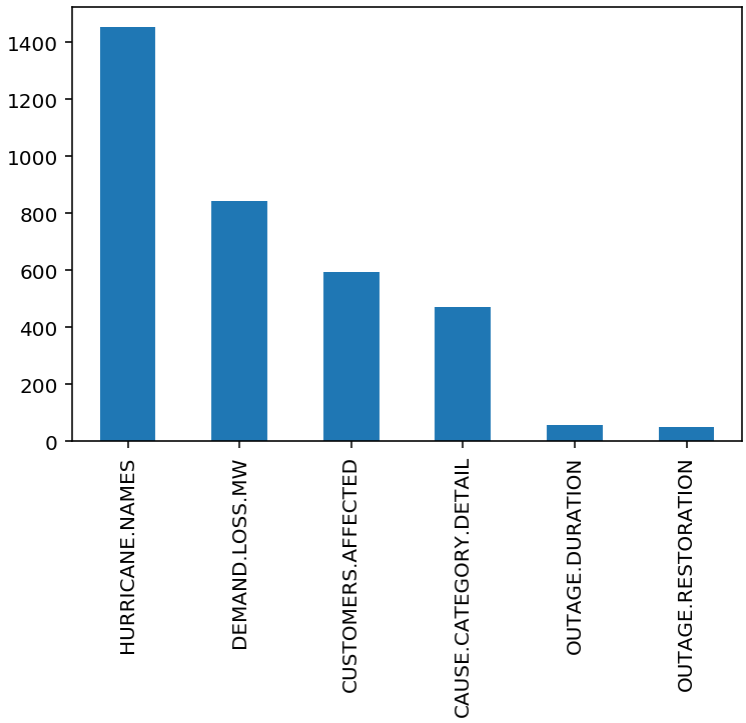
Fig. 44: Top 20 Longest Duration Outages by Severe Event Type



Assessment of Missingness

In [354]:
```
# Plot the missingness of things with greater than 40 things missing
# Conclusion: Hurricane Names should be MAR dedepdent on Cause, we can v
erify this
# Conclusion: First Priority: Demand Loss and Customers Affected are hig
h NaN's and are important to see if we can immute
# Conclusion: Second Priority: Outage Duration and Restoration are mediu
m NaN's
missing_sums = (outages.isnull().sum() > 40)
outages.iloc[:,missing_sums.to_numpy()].isnull().sum().sort_values(ascen
ding=False).plot(kind='bar')
title = plt.title("Fig. "+str(fignum)+": "+"Major missing values per col
umn", loc='center', pad=None)
fignum += 1
```

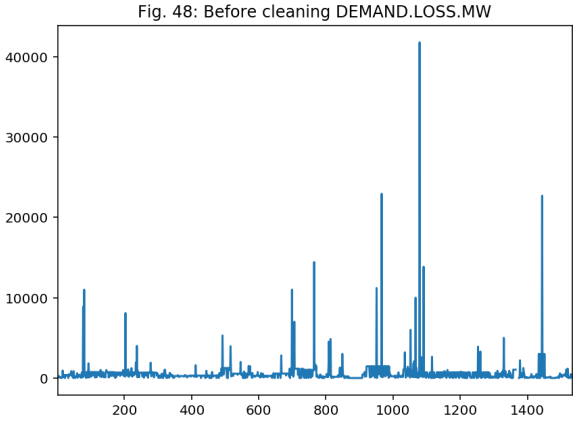
Fig. 45: Major missing values per column

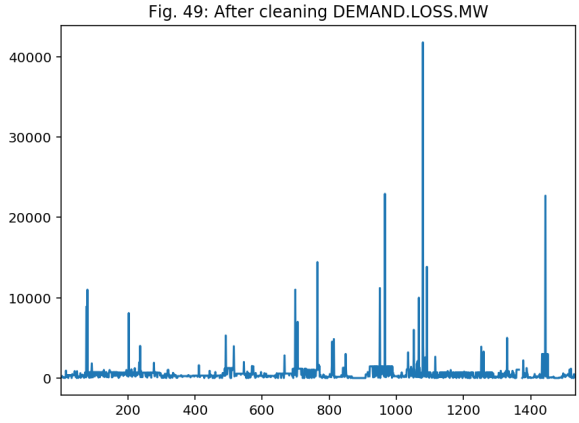


In [355]:
```
# MAR dependency permutation test for DEMAND.LOSS.MW
potential_dep_columns = ['YEAR', 'MONTH', 'U.S._STATE', 'NERC.REGION',
'CLIMATE.REGION', 'CLIMATE.CATEGORY', 'CAUSE.CATEGORY']

for col in potential_dep_columns:
    print("MAR DEMAND.LOSS.MW check dependency against", col, 'pvalue=',
util.perm4missing(outages, 'DEMAND.LOSS.MW', col, 1000))
```

```
MAR DEMAND.LOSS.MW check dependency against YEAR pvalue= 0.872
MAR DEMAND.LOSS.MW check dependency against MONTH pvalue= 0.211
MAR DEMAND.LOSS.MW check dependency against U.S._STATE pvalue= 0.02
MAR DEMAND.LOSS.MW check dependency against NERC.REGION pvalue= 0.315
MAR DEMAND.LOSS.MW check dependency against CLIMATE.REGION pvalue= 1.0
MAR DEMAND.LOSS.MW check dependency against CLIMATE.CATEGORY pvalue= 1.
0
MAR DEMAND.LOSS.MW check dependency against CAUSE.CATEGORY pvalue= 0.87
3
```

In [357]:
```
# Conclusion is that Demand Loss is MAR conditionally dependent on State
# Impute it by using group of means
before_cleaning = outages['DEMAND.LOSS.MW'].copy()
after_cleaning = util.impute_by_mean_groups(outages, 'DEMAND.LOSS.MW',
'U.S._STATE')

outages['DEMAND.LOSS.MW'] = after_cleaning

fis, axes = plt.subplots(1,2, figsize=(15,5))
before_cleaning.plot(ax=axes[0], title="Fig. "+str(fignum)+": "+'Before
 cleaning DEMAND.LOSS.MW')
fignum+=1
after_cleaning.plot(ax=axes[1], title="Fig. "+str(fignum)+": "+'After cl
eaning DEMAND.LOSS.MW')
fignum+=1
```



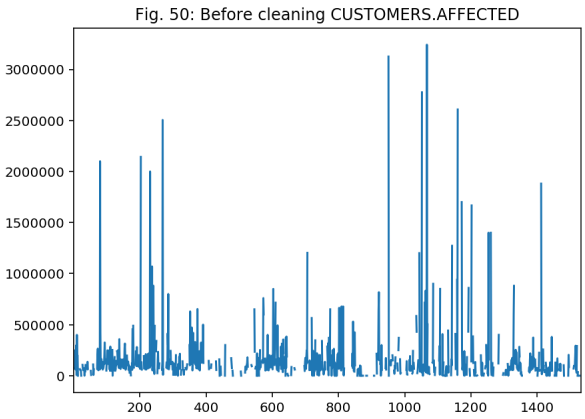


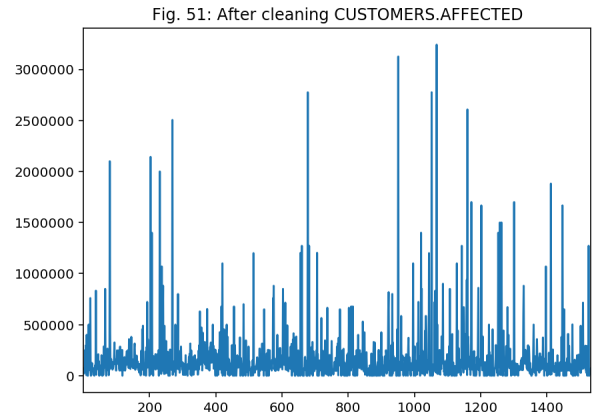
In [255]:
```
# MAR dependency test for CUSTOMERS.AFFECTED
for col in potential_dep_columns:
    print("MAR CUSTOMERS.AFFECTED check dependency against", col, 'pvalu
e=', util.perm4missing(outages, 'CUSTOMERS.AFFECTED', col, 1000))
```

```
MAR CUSTOMERS.AFFECTED check dependency against DEMAND.LOSS.MW pvalue=
0.0
MAR CUSTOMERS.AFFECTED check dependency against MONTH pvalue= 0.777
MAR CUSTOMERS.AFFECTED check dependency against U.S._STATE pvalue= 0.47
4
MAR CUSTOMERS.AFFECTED check dependency against NERC.REGION pvalue= 0.9
52
MAR CUSTOMERS.AFFECTED check dependency against CLIMATE.REGION pvalue=
0.25
MAR CUSTOMERS.AFFECTED check dependency against CLIMATE.CATEGORY pvalue
= 0.889
MAR CUSTOMERS.AFFECTED check dependency against CAUSE.CATEGORY pvalue=
0.908
```

In [358]:
```
# Conclusion is that Demand Loss is MCAR
# Impute with unconditional probablistic distribution
before_cleaning = outages['CUSTOMERS.AFFECTED'].copy()
after_cleaning = util.impute_by_probablistic_uncond(outages['CUSTOMERS.A
FFECTED'])

outages['CUSTOMERS.AFFECTED'] = after_cleaning

fis, axes = plt.subplots(1,2, figsize=(15,5))
before_cleaning.plot(ax=axes[0], title="Fig. "+str(fignum)+": "+'Before
 cleaning CUSTOMERS.AFFECTED')
fignum+=1
after_cleaning.plot(ax=axes[1], title="Fig. "+str(fignum)+": "+'After cl
eaning CUSTOMERS.AFFECTED')
fignum+=1
```

```
595 0
```



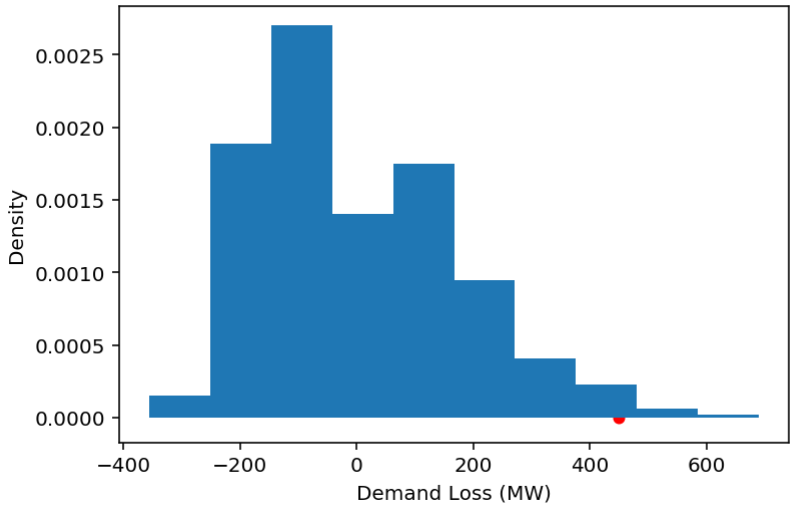


Hypothesis Test

In [359]:
```
# Perform Hypothesis testing using difference of means
def calc_statistic(df, population_mean, hypothesis_axis):
    return df[hypothesis_axis].mean() - population_mean
def simulate_experiment(num_experiments, hypothesis_axis, year_start, ye
ar_end):
    natural_disater = outages[outages["Natural Disaster"]]
    natural_disater_ca = natural_disater[natural_disater["U.S._STATE"]==
"California"]
    later_period = natural_disater_ca[(natural_disater_ca["YEAR"] >= yea
r_start) & (natural_disater_ca["YEAR"] <= year_end)]
    population_duration = natural_disater_ca[hypothesis_axis].mean()
    num_samples = later_period.shape[0]
    observed_stat = calc_statistic(later_period, population_duration, hy
pothesis_axis)
    sample_means = []
    for _ in range(num_experiments):
        df_sample = natural_disater_ca.sample(n=num_samples, replace=Fal
se)
        sample_means.append(calc_statistic(df_sample, population_duratio
n, hypothesis_axis))
    p_value = np.count_nonzero(sample_means>=observed_stat)/num_experime
nts
    return p_value, sample_means, observed_stat
```

In [360]:
```
# Hypothesis Test
# Conclusion: Reject Null
p_value, sample_means, observed_stat = simulate_experiment(10000, "DEMAN
D.LOSS.MW", 2013, 2016)
pd.Series(sample_means).plot(kind='hist', title="Fig. "+str(fignum)+": "
+'Hypothesis Test - Demand Loss (MW) Historical vs Recent 3 Years, pvalu
e=' + str(p_value), density=True)
plt.xlabel('Demand Loss (MW)')
plt.ylabel('Density')
plt.scatter([observed_stat], [0], c='r', s=25);
fignum += 1
```

Fig. 52: Hypothesis Test - Demand Loss (MW) Historical vs Recent 3 Years, pvalue=0.0139



In [263]:
```
# Hypothesis Test (II) : Changing to use Customers Affected, in this sce
nerio we CANNOT reject the NULL
p_value, sample_means, observed_stat = simulate_experiment(10000, "CUSTO
MERS.AFFECTED", 2013, 2016)
pd.Series(sample_means).plot(kind='hist', title="Fig. "+str(fignum)+": "
+'Hypothesis Test - Customers Affected Historical vs Recent 3 Years, pva
lue=' + str(p_value), density=True)
plt.xlabel('Customers Affected')
plt.ylabel('Density')
plt.scatter([observed_stat], [0], c='r', s=25);
fignum += 1
```

Fig. 51: Hypothesis Test - Outage Duration Historical vs Recent 3 Years, pvalue=0.5212