Capstone 2: Final Report

Problem Statement and Context

Based on electric grid outage data from 2018 - 2023, in which geographic areas and in which infrastructure should investments be made to have the greatest positive impact on grid reliability in the next 2 decades?

Electric grid resilience and reliability are vital for the US economy and for the well-being of Americans. Without reliable electricity, human health and safety would suffer, our economic prosperity would be threatened, and national security would be at risk. Electricity is fundamental to keeping life-saving hospital equipment functioning, communication systems operating, maintaining safe temperatures and adequate ventilation in buildings, among a whole host of other necessities.[1] As the Secretary of Energy stated in a cover letter of the Department of Energy's most recent electric reliability report, "a reliable and resilient electric grid is critical not only to our national and economic security, but also to the everyday lives of American families."[2,5] To put this in monetary terms, American consumers paid ~$419B for electricity in 2021[3] - representing ~1.8% of total US GDP.

In November of 2022, the government allocated $13B to improve electric grid resilience and reliability.[4] The main threats to reliability are seen as aging infrastructure (~70% of America's electric grid is over 25 years old[4]), climate change / severe weather, a changing energy generation mix (i.e., higher reliance on renewable energy sources), increasing energy consumption (e.g., expected adoption of electric vehicles will greatly increase electrical demand), and cyber attacks (which will be outside the scope of this analysis). It is estimated that "the US needs to expand electricity transmission systems by 60% [above Nov 2022 levels] by 2030 and may need to triple current capacity by 2050 [above Nov 2022 levels] to accommodate the country's rapidly increasing supply of cheaper, cleaner energy and meet increasing power demand."[4]

This analysis seeks to use data relating to major power outages across the US spanning from 2018- 2023, electrical demand and generation data spanning from 2018 - 2023, and weather data to assess which factors have the greatest impact on grid reliability, how we can expect these factors to contribute to future grid reliability, and assess where these recently allocated funds would best be used to improve electric grid reliability.

[1] https://www.energy.gov/eere/energy-reliability#:~:text=It%20is%20vital%20to%20human,good%20ventilation%2C%20and%20much%20more.
[2] https://www.energy.gov/sites/prod/files/2017/08/f36/Secretary%20Perry%20Grid%20Study%20Cover%20Letter.pdf
[3] https://www.eia.gov/todayinenergy/detail.php?id=57320#:~:text=Petroleum%20products%20made%20up%20the,a%203%25%20increase%20from%202020.
[4] https://www.energy.gov/articles/biden-harris-administration-announces-13-billion-modernize-and-expand-americas-power-grid
[5] https://www.energy.gov/sites/prod/files/2017/08/f36/Staff%20Report%20on%20Electricity%20Markets%20and%20Reliability_0.pdf

Data Wrangling

This analysis pulled data from the following sources:

1. **Outage Data**: Major US electrical outages from 2018 - 2023 as recorded by the Department of Energy (DOE) (https://www.oe.netl.doe.gov/OE417_annual_summary.aspx)
    - This dataset consisted of separate excel files for each year. The code imports each excel file into a pandas dataframe and concatenates them into one dataframe
    - This dataset included outage information from 2000 – 2023, however since energy data was only available from 2018 onward, the years prior to 2018 were dropped
    - The dataset had to be cleaned since the different years recorded data in slightly different ways. Some modifications included: converting column titles for consistency, converting dates to the proper format and datatype, converting numeric columns to numeric datatypes, assigning each outage to the proper state (some information was recorded by county, or power company and therefore had to be manually adjusted to assign it to the affected state)

2. **Weather Data**: Local weather at the time of the outage event as recorded by the National Oceanic and Atmospheric Administration (NOAA) (https://www.ncdc.noaa.gov/cdo-web/webservices/v2#gettingStarted)
    - This data was pulled from the NOAA API which was very finnicky and would periodically return an error. The code to pull from this API includes a try / except to wait 30 seconds if a 503 error is returned in order to try again.
    - This code follows this general process:
        a. For each event, look up the affected state and match the state to the associated FIPS ID (FIPS ID is a numeric code that NOAA assigns to each state), look up the start date (end date is set equal to the start date since I only care about the weather at the time the event started),
        b. Then, for each event request the information for average temperature, average windspeed, precipitation amount, and snowfall amount for the affected state. Each state has numerous stations, so the request will pull data from each station within that state and then average the amount for each metric over the whole state. This average amount is then converted to the proper units and added to a dictionary.
        c. Since the API was finnicky, I kept track of the last index updated and used an if statement to skip over the index of the outage data if the weather data was already pulled. This helped to complete the data pull in a timely manner

d. The resulting dictionary was then converted to a pandas data frame and then concatenated to the outage data frame. For entries with no data, I converted them to null values.

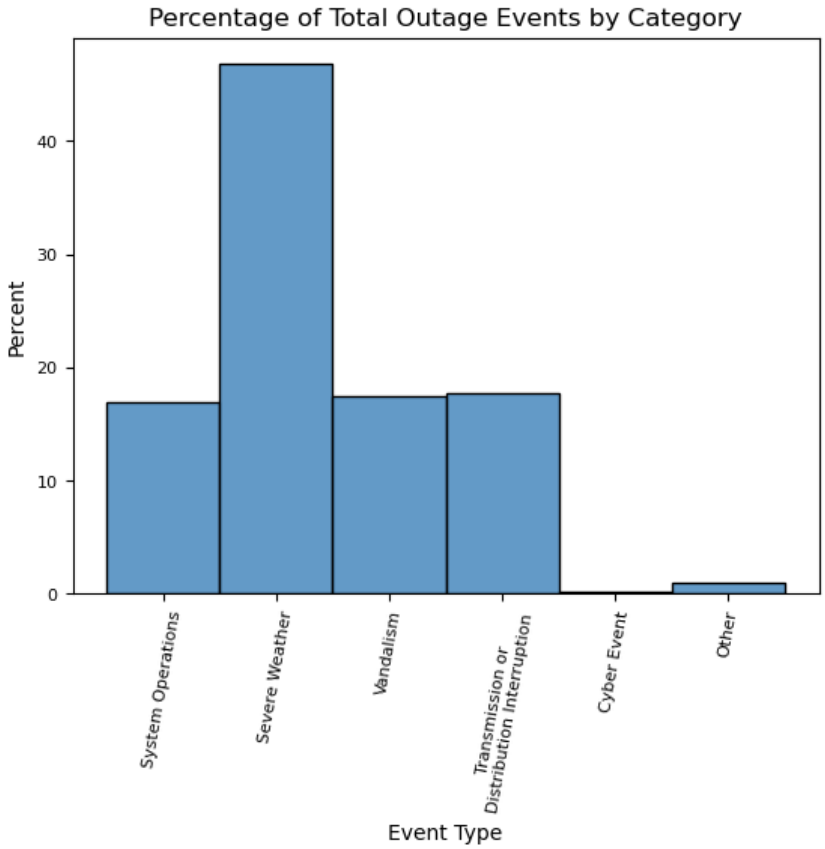
3. **Energy Demand Data**: Levels of energy demand at the time and reliability region of the outage event as recorded by the North American Electric Reliability Corporation (NERC) (https://www.nerc.com/pa/RAPA/ESD/Pages/default.aspx)
    - The energy demand data from NERC is separated into geographic regions that span multiple states. The outages dataset already has a column denoting which NERC region the outage occurred in, so I was able to download energy generation and demand data in excel from NERC directly.
    - I then created a script that would pull the energy demand and generation information for the applicable NERC region during the month of the outage event from the NERC data set (daily information was not available)
    - I incorporated this information into the outage dataset
    - Note: I had originally planned to use data from EIA to have daily energy demand and generation information, however, that dataset was incomplete and contained many null values. I decided that the benefits gained by having less null values are greater than the drawbacks due to the lack of specificity.

4. The resulting dataset contained information regarding 419 outages (i.e., 419 rows) and 12 features for each outage (i.e., 12 columns). For each outage I had the following features:
    - Datetime Event Began
    - State affected
    - NERC Region
    - Alert Criteria (free text describing cause of outage)
    - Event type (categorical variable classifying the outage type)
    - Demand Loss (MW)
    - Number of Customers affected
    - State Avg Temp (F)
    - State Avg Windspeed (mph)
    - State Avg Precipitation (mm)
    - Monthly Net Energy for Load (GWh) – i.e., energy supply within the NERC region
    - Monthly Peak Hour Demand (MW) – i.e., energy demand within the NERC region

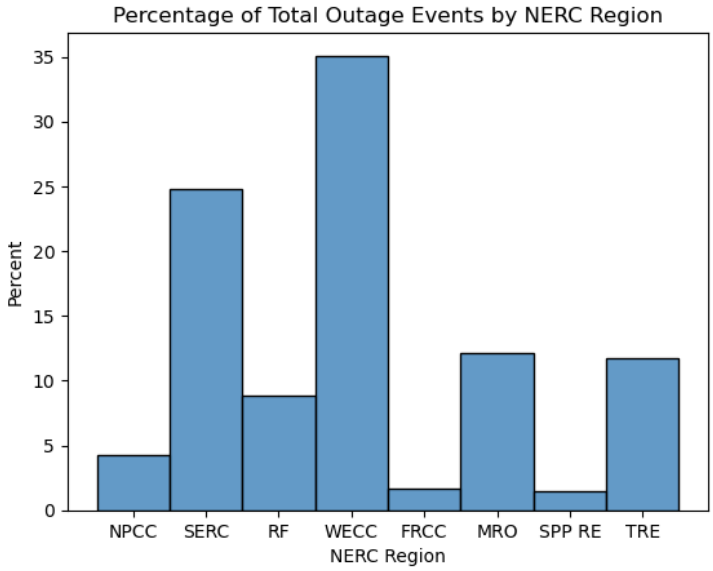
Exploratory Data Analysis

To get a sense of the data and the characteristics of the outages, I made the following visualizations and had the following key takeaways:

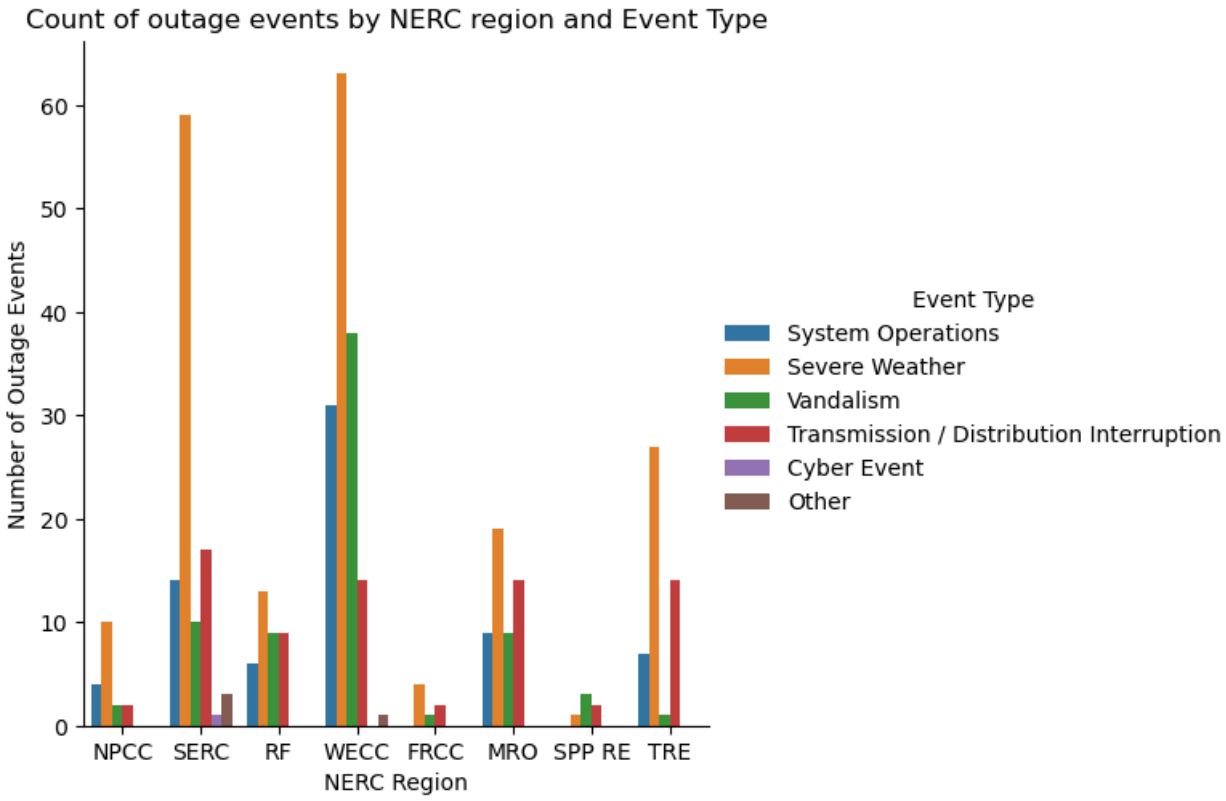
1. **Histogram of outage events by type:** by far, the most common cause of a severe outage in the US since 2018 has been severe weather.



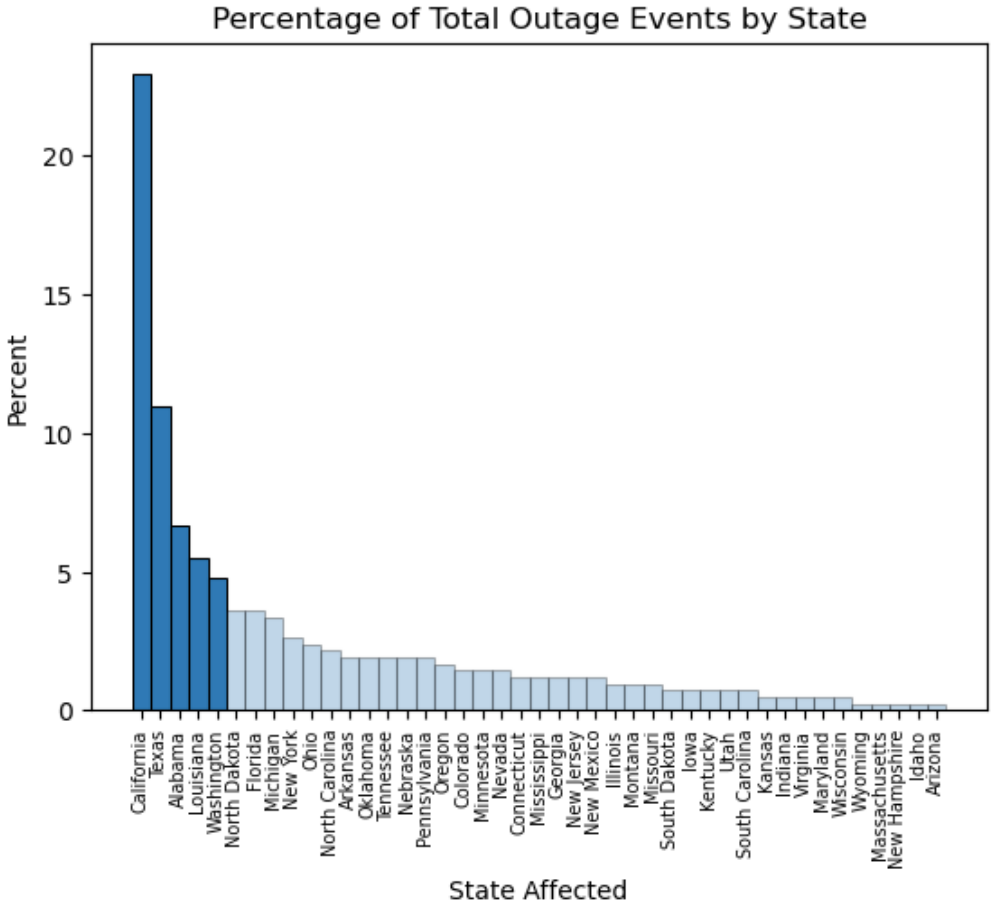
2. **Histogram of outage events by NERC region:** the WECC region makes up nearly 35% of all outages. WECC is by far the largest geographic region, covering most of the American west, so perhaps this isn't surprising. Additionally, the American west tends to be extremely hot and has a history of large wildfires. These factors could contribute to its outsized share of outages.



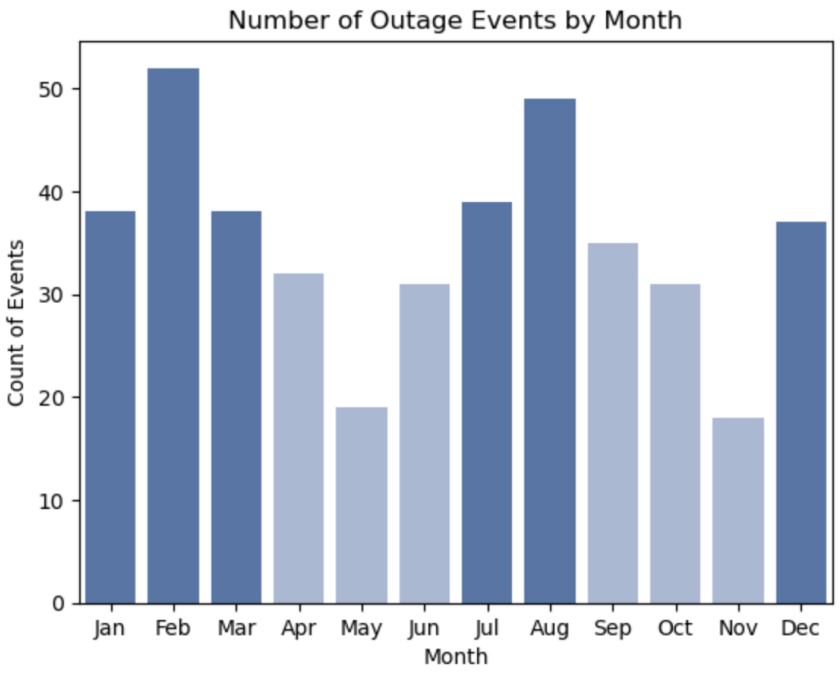
3. **Histogram of outage events by NERC Region/Event Type:** WECC seems much more likely than other regions to have system operation issues as well as vandalism events. This could be interesting to explore. Maybe WECC needs more investment to improve system operations and security to prevent vandalism.



4. **Countplot of outage events by state:** California and Texas by far have the most outages. This isn't totally surprising given their large geographic areas. What is surprising is Alabama, Louisiana, and Washington have a surprising amount of outages. Given their relatively high number of outage events compared to their geographic area, an investment in infrastructure resiliency here may have a high impact compared to investments elsewhere. **These top 5 states account for >50% of all outages reported.**



5. **Countplot of outage events by month:** There is some seasonality associated with outages. There seem to be peaks in February (typically associated with cold winter weather) as well as August (typically associated with hot summer weather). The mild seasons of spring and fall typically have fewer outages. This reinforces the above insight that severe weather is the most predominant cause of outages.



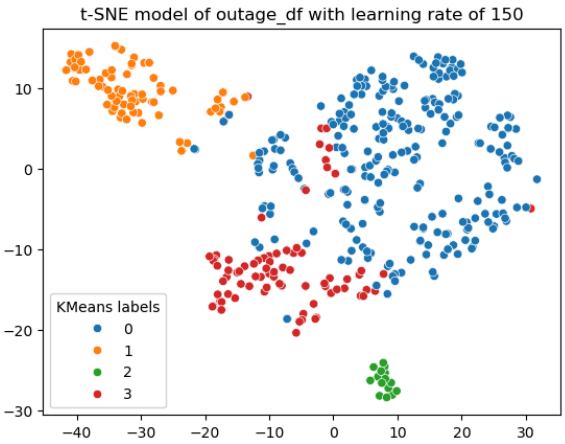
6. Other key takeaways:
    a. 2020, 2021, and 2022 had significantly more outage events than the other years. Maybe this could be due to stay at home orders and a corresponding shift in electrical energy demand?
    b. A lot of the outages seem to occur at higher temperatures. This may be due to increased demand on the electrical grid during high temperature times. When demand is near supply without extra capacity to bring online, a weather event is more likely to cause a significant outage.
    c. The following features seem to be consistently highly correlated:
        i. Monthly net energy for load and monthly peak hour demand: this makes sense as power generators bring on more supply as demand increases
        ii. Avg temp and peak hour demand: This is also expected as higher temperatures require more energy to run air conditioning units
        iii. Avg windspeed and precipitation: It seems as if precipitation storms also bring higher windspeeds
        iv. Demand loss and number of customers affected: As more demand is lost, more customers will lose power

Modeling

I decided to pursue unsupervised clustering to gather additional insights about the data. I executed the following models: KMeans, MeanShift, Hierarchical, and DBScan.

1. KMeans:
    a. Hyperparameters adjusted: n_clusters. I varied n_clusters for the KMeans model from 3 to 20 and plotted the resulting model inertia on a lineplot. This allowed me to visualize which was the ideal number of clusters. From the plot, it seems as if 5 or 6 clusters is ideal, however no clear 'elbow' exists and inertia continues to drop quite consistently as the number of clusters increases.
2. MeanShift:
    a. I fit a MeanShift model to the dataset. This model will output the number of clusters it found. In this case, the number of clusters formed was 5.
3. Hierarchical Clustering:
    a. I created a dendrogram of the dataset, clustering based on euclidean distance. This visualization helps to see how the data is clustered. I then extracted the cluster labels based on the height of the dendrogram that I believed explained a significant portion of the dataset, in this case at a height of 6.
4. DBScan Clustering:
    a. Hyperparameters adjusted: eps. I varied eps from 0.5 to 10.5 and analyzed how many clusters were formed for each. 2.0 seemed the ideal eps as it formed 5 clusters whereas most of the other eps values formed only 1 cluster.

I then performed dimensionality reduction to help me visualize the clusters. I did two forms of dimension reduction, t-SNE and PCA. For t-SNE, I varied the learning time while for PCA I plotted the number of components against variability explained. I then plotted the clusters against the t-SNE dimensions in order to view the clusters. This visualization technique allowed me to choose which clustering method produced the best results. As, can be seen below, KMeans clustering produced 4 distinct clusters and are separated pretty well in the t-SNE dimensions. I therefore believe KMeans with n_clusters = 4 is the clustering model that provides the best insight.



Cluster Interpretation

I used the crosstab function and descriptive statistics to analyze the characteristics of the KMeans clusters.

KMeans labels crosstab table against NERC Region: most of the western outages are in cluster 0. Cluster 1 contains all the outages from Texas, Florida, and SPP (which is a combination of Arkansas, Iowa, Kansas, Louisiana, Minnesota, Missouri, Montana, Nebraska, New Mexico, North Dakota, Oklahoma, South Dakota, Texas, and Wyoming).

NERC Region	*FRCC*	*MRO*	*NPCC*	*RF*	*SERC*	*SPP*	*TRE*	*WECC*
KMeans Labels								
0	0	36	3	26	60	0	0	123
1	7	0	11	1	0	6	49	1
2	0	2	0	2	0	0	0	11
3	0	13	4	8	44	0	0	12

KMeans labels crosstab table against Event Type: vandalism seems to be included in cluster 0, while distribution interruptions are primarily in cluster 0 and 1. Severe weather seems spread evenly throughout.

Event Type	*Cyber Event*	*Other*	*Severe Weather*	*System Operations*	*Distribution Interruption*	*Vandalism*
KMeans Labels						
0	1	4	85	49	50	59
1	0	0	40	9	20	6
2	0	0	11	2	0	2
3	0	0	60	11	4	6

The descriptive statistics for each feature grouped by the KMeans clusters revealed the following insights:

- **Demand Loss:** By far, the largest demand losses are concentrated in cluster 3, however this seems to be due to a major outlier with an extremely high value of demand loss. The medians are much closer together.

- **Number of Customers Affected:** Again, outliers cause the highest mean of this metric to be in cluster 1. However, the medians are much more closely distributed.
- **State Avg Temp and State Avg Snowfall:** Cluster 2 seems to contain colder temperatures and larger amounts of snowfall. Perhaps this is the defining characteristic of cluster 2.
- **State Avg Windspeed:** Cluster 3 seems to contain the outages with the highest associated windspeed.
- **State Avg Precipitation:** Cluster 2 seems to contain the highest amount of associated precipitation. This is in line with this cluster being associated with the largest amount of snowfall. Cluster 3 also seems to have large amounts of precipitation, although not as much as cluster 2.
- **Net Energy for Load:** Cluster 1 has the lowest load. The others are pretty similar.
- **Energy Demand:** Cluster 1 has the lowest demand as well. Other clusters are pretty similar.

Takeaways and Next Steps

1. Severe weather is the most common cause of major outages in the US. Investing in infrastructure that can withstand weather events (e.g., underground distribution, higher supply capacity) may provide a high ROI in terms of outages prevented.
    a. Relatively few outages are caused by heavy snowfall (cluster 2) or high winds (predominantly cluster 3). In fact, many outage events for severe weather occur during elevated temperatures. It seems like severe weather may be more of an indirect cause of outages (e.g., high temperatures causing more people to run their AC, overloading the electrical system) vs. a direct cause (e.g., snowfall causing a transmission line to go down). **Further studies should examine how best to combat this.**

2. Vandalism is also a common cause of major outages, accounting for nearly 20% of all events. Making investments to provide better security of critical transmission / distribution sites may be a better investment than purchasing new equipment. **A more thorough cost/benefit analysis would be warranted here.**

3. 5 states account for over 50% of major outage events. Making infrastructure investments in these 5 states may prove to be more beneficial to the resilience of the electrical grid than investments elsewhere. **Further research should be done on these five states to determine if investments in these regions would be optimal.**