**Executive Summary**

Stakeholders: DOE executive and managerial leadership.

Assumptions and Limitations:

* Excluded Canada, DOE only has jurisdiction over U.S. states and territories.
* Not all U.S. territories are represented in reporting.
* No reporting from Alaska. The DOE does get some reporting from an alternative source as can be seen in this [DOE report](https://www.energy.gov/sites/prod/files/2015/05/f22/AK-Energy%20Sector%20Risk%20Profile.pdf).
* No direct contact with DOE. Unable to get clarity on questions and concerns.

Overview: A DOE-commissioned dashboard for review of past events and continued monitoring of future events. The dashboard allows a review of key concerns: demand loss, duration, customers affected, and event types. Filtering can be done by year, month, NERC Region, state, and event type. The dashboard is intended to be used by all levels of leadership.

Recommendations:

1. Require the reporting entity to complete missing data or confirm after the event that the data is not available. DOE manual or automated follow-up on incomplete data. With many old "ongoing" events or missing resolution data/time, it is clear that reporting is not prioritized by reporters. Stressing the need for the missing data may yield more complete data.
2. Create reporting personnel standards. Data inconsistencies indicate that underqualified or undertrained personnel are reporting outages.
   * Create a standard of expectation on the qualification of reporting personnel.
   * Review instruction and documentation for any areas that may be unclear. Then distribute updated documentation to reporting entities.
     + Consider adding criteria on how to estimate information as an additional option and/or mandate it on reporting of event resolution.
     + Mandate [NERC region](https://www.nerc.com/AboutNERC/keyplayers/Pages/default.aspx) and state identification and provide a separate column for [NERC sub-regions](https://www.eia.gov/electricity/data/eia411/#tabs_NERC-1).
   * Review lines of communication to ensure reporting agencies can seek clarification.
3. Create a data repository for all data reporting of outages.
   * Consider standardizing how and what data is collected through different channels.

**Review data, documents, excel code, and Python code in** [this GitHub](https://github.com/gregorywmorris/Maven-Power-Outage-Challenge).

**Dashboard Design Decisions**

* Style guide determined by:
  + The DOE logo was used as a color style guide due to a lack of specifications.
    - NOTE: colors are not color-blind friendly. For accessibility, white space is used to separate colors and color distinction is not necessary for interpretation.
  + [DOE Logo and Branding](https://www.energy.gov/department-energy-logo-and-branding-guidelines)
  + [DOE Style Guide Full Text](https://www.energy.gov/eere/communicationstandards/style-guide-full-text#abbreviations)
* Interactivity: Field selection allows for a quick user-friendly way to filter data as well as increasing accessability.
* Allow high-level overview with drill-down capabilities so leadership at any level can access key information.
* The expected use is DOE internal. The design utilized 'Tableau automatic' and is best viewed on a desktop or laptop and is functional on a tablet. Phone use is not recommended.
* Simple design: Leadership-focused, avoid any distracting images or too many colors. For consumer-facing, a flashier design would be appropriate.

**Data Cleaning**

* Links throughout reference sources were used to assist in data cleaning decisions.
* Used the Data Dictionary as a guide for expected values.
* Assumed that the DOE will accept filled missing data for analysis purposes. Alternatively, I would have considered row deletion or simply left as is if desired by the DOE; the latter would likely have a greater negative effect on analysis. The ideal scenario would be to contact reporting entities and request missing data.
* Assumed Puerto Rico data is desired. PR region created.
* Assumed Hawai data is desired. HI region created.

Overview TLDR

Three-stage data cleaning processes in Excel and Python.

1. Stage one: Merge and formatting focused, light data cleaning.
2. Stage two: Deep dive into Excel data cleaning and subject matter research.
3. Stage Three: Final Excell data cleaning, Python to identify and/or clean, machine learning implementation to be considered for filling missing data.

Full Detail  
Stage One

Excel

* Formatted data tables:
  + Format to use 2022's column names as this format best fits the majority of the data.
  + For Alert criteria before 2015, listed as 'unknown'.
  + Split combined date time columns for years before 2011.
  + Formatted the correct type for date, time, and integer columns.
* Dates transposed (showing as finishing before the start date), corrected the date and month then moved to the correct sheet when appropriate.
* Removed 'ongoing' from the date column.
* Remove gridlines and white background color.
* Sheet 2023: Removed Temp column; no data.
* Sheet 2016: Only had data up to October 31st.
* Sheet 2011: the years 2077 (well into the future) and 2001 (before data collection) were corrected to 2011 to match the start date.
* Sheet 2004: Removed record with no start and end dates. Unable to confirm actuals.
* Sheet 2002 and 2003: Merge and center data cells by date across all columns where appropriate.

Stage Two

Excel

* Combined into a single sheet. Shape (3913,11).
* Merge any identified duplicates.
* Format time to resolve a.m./p.m. issue.
* Convert to time data type.
* Merge date and time.
* Convert restoration time based on hours and correct resolution dates.
  + Date Event Began: Format to date type, correct transposed dates i.e. ends before it starts.
  + Combined date and time as "Date Event Began".
* Time Event Began: Removed extra spaces and words. Transitioned to time ##:## AM/PM format, Changed 5:70 to 5:00, N/A and blank changed to 1600 as this is the most common time per [the Red Cross](https://perryco.org/wp-content/uploads/2020/07/poweroutage.pdf).
* Date of Restoration: Ongoing and blank dates are filled with the start dates as that is the only day we can confirm the outage occurred. Formatted to date data type.
* Time of Restoration:
  + Removed extra spaces and words.
  + Transitioned to time ##:## AM/PM format, N/A and blank end times determined by average time per year based on reporting from [US Energy Information Administration](https://www.eia.gov/todayinenergy/).
  + No reliable date before 2008, used an average time of 3.5 hours for 2009 and all earlier years. [Eaton Blackout Tracker](https://www.eaton.com/content/dam/eaton/products/backup-power-ups-surge-it-power-distribution/backup-power-ups/blackout-tracker-/eaton-blackout-tracker-annual-report-2009.pdf) began in early 2008 but only the 2009 report is available.
  + No reliable date after 2021, used an average time of 7 hours in 2021 for all later years.
  + Combined date and time as "Time of Restoration".
* Area Affected: Puerto Rico: to Puerto Rico, blank to Unkown
* Created NERC Region: Puerto Rico areas changed to PR [they do not fall into a specific region](https://19january2017snapshot.epa.gov/energy/north-american-reliability-corporation-nerc-region-representational-map_.html). Filled blanks by looking at similar areas affected.
* Event Type:
* Note 1: Format so that similar items are grouped for analysis. Many nuanced details fail to provide additional insight and hinder analysis often due to the low number of records. Grouping related instances will allow better general analysis and is more appropriate for a dashboard.
* Note 2: Public appeals are a request for public energy conservation and not a cause. It is done when the power grid is unable to supply the power needed; Generation Inadequacy.
  + Format as Text.
  + Cyber Event: All cyber attacks/events and telecommunication attacks. Excludes computer hardware.
  + Equipment Failure: All variants of equipment/line/generator/switch/hardware/cable/substation/exciter/breaker faulted/failure/malfunction/tripped/error/loss/shutdown, complete system failure, Operational Failure of Electrical System. Unless the cause is listed i.e. fire, severe weather, etc., we only know the equipment failed.
  + Fire/Wildfire: All variations of fire listed, all variants of wildfire and Brushfire (a type of wildfire); supersede equipment failure.
  + Generation Inadequacy Load/Fuel/Supply: Generation inadequacy, inadequate electric resources to serve load, pubic appeals, high loads, all to Fuel Supply Deficiency, loss of power from a wholesaler, CAISO Initiated Interruption: All [CAISO (California Independent System Operator)](https://www.caiso.com/Documents/Rotating-Power-Outages-Fact-Sheet.pdf) variants converted. Unless the cause is listed i.e. fire, severe weather.
  + Natural Disasters - Earthquake/Hurricane/Tornado/Tropical: All variants of hurricanes, tornados, and tropical depressions.
  + Natural disasters could encompass more than listed here depending on the source definition. A decision was made to group other weather events separately as it is likely unique decisions can be made for them.
  + Other: Rare and unique events. Low-flying helicopter, Voltage Reduction (System Test), Made Public Appeal - System Drill, and other.
  + Physical Attack/Vandalism: all variants of physical attack, vandalism, and suspicious activity.
  + Severe Weather - Heat Wave: Heat storm, heat wave, high temperatures.
  + Severe Weather - Lightning/Thunderstorm: Lightning storms, lightning strikes, lightning, thunderstorms and [Hail](https://www.nssl.noaa.gov/education/svrwx101/winter/types/).
  + Severe Weather - Rain/Wind/Flooding: All variants of rain, and flooding, with or without Wind.
  + Severe Weather - Wind: Nor'easter, high winds, Severe Storm with High Wind Gusts, dust storm.
  + Severe Weather - Winter/Snow/Ice: All variants of winter, snow, ice, cold weather, freezing rain, and winter storm events. NOAA's National Severe Storms Laboratory [groups these as winter storms](https://www.nssl.noaa.gov/education/svrwx101/winter/types/), Public Appeal due to Severe Weather - Cold.
  + Severe Weather - Unspecified/Other: May or may not include high winds, severe weather, severe/major storms, weather, and fog.
  + System Operations: [System operations](https://www.pjm.com/markets-and-operations/ops-analysis), operational failure of electrical system.
  + Unkown/Unspecified: Unknown \*, - Unknown, Distribution Interruption - Unknown Cause, [Load shedding](https://www.techtarget.com/searchdatacenter/definition/load-shedding), shed firm load, public appeal (no cause listed), load reduction, interruption of firm power, Electrical System Separation/[Islanding](https://en.wikipedia.org/wiki/Islanding). Unless the cause is listed i.e. fire, severe weather, etc.
* NERC Region
  + Convert to current [NERC regions](https://www.nerc.com/AboutNERC/keyplayers/Pages/default.aspx) due to the limited sub-region reporting.
  + Electricity Information Sharing and Analysis Center (E-ISAC) converted to the appropriate NERC region based on the criteria listed below:
  + [NERC Atlas for NERC identification](https://atlas.eia.gov/maps/nerc-regions) with Google Maps to identify locations not in the atlas.
  + NERC based on the Area Affected Column.
  + Convert delimiters to ",".
  + Corrected spellings.
  + Purto Rico: PR
  + Hawaii: HI
  + Indeterminate NERC membership: List nearest NERC.
* Demand Loss(MW):
  + The expected value is a number or leave blank if unknown. All strings were removed.
  + Deleted: 'NA', 'unknown', '-', descriptive text.
  + For ranges, only accept the highest estimate.
  + Approx, greater/Less than converted to just the number given.
  + Customers affected moved to the correct column, deleted peak and kept actual (and removed strings).
  + Removed dates and times.
  + Error: Number stored as text, converted to a number.
  + Formated number with one decimal place to maintain the accuracy of estimates given.
* Number of Customers:
  + The expected value is a number or left blank if unknown.
  + Deleted descriptive text.
  + Error: Number stored as text, converted to a number.
  + Deleted: 'NA' and 'unknown' strings, date, '-'.
  + Approx, greater/Less than converted to just the number given.
  + Converted utilities and industrial to just the number given.
  + Formatted to a whole number.

Stage Three

Excel

* Validate and complete state and province identification.
* Correct NERC Region based on state identification.
* Manually correct states post Python processing.
* Google city, area, and county locations.
* Unknown locations left blank, 18 total.
* Carolina = ['North Carolina','South Carolina']
* Identify states through researching these listed Area Affected:
  + [Midcontinent Independent Operator (MISO)](https://www.misoenergy.org/), [service area](https://en.wikipedia.org/wiki/Midcontinent_Independent_System_Operator).
  + [Delmarva Power service territory](https://www.delmarva.com/AboutUs/Pages/CompanyInformation.aspx)
  + [Southwestern Region of Service Territory](https://www.swepco.com/company/about/)
  + [Mid-Altantic Region of PJM](https://www.pjm.com/about-pjm/who-we-are.aspx)
  + [Duke Energy](https://www.duke-energy.com/partner-with-us/economic-development/the-carolinas)
  + [Dominion Energy](https://en.wikipedia.org/wiki/Dominion_Energy)
  + [ComEd](https://www.exeloncorp.com/companies/comed)
  + [BGE](https://www.bge.com/AboutUs/Pages/CompanyInformation.aspx)
  + [A.D. Edmonston Pumping Plant](https://www.watereducation.org/aquapedia/ad-edmonston-pumping-plant)
  + [CSWS-AEP West](https://www.aep.com/about/businesses/opcos)
  + [Southern Company](https://www.southerncompany.com/about/our-business/energyisessential.html)
  + [Entergy System](https://www.entergy.com/about/)
  + [TVA Service Territory](https://www.enelx.com/n-a/en/resources/brochures/tennessee-valley-authority-demand-response)
    - [Government archives](https://www.archives.gov/milestone-documents/tennessee-valley-authority-act)
  + [Balancing Area](https://www.eia.gov/electricity/gridmonitor/about)
  + [Southeastern Power Administration (SEPA)](https://www.federalregister.gov/agencies/southeastern-power-administration)
* Duplicates are identified by area affected and time, then merged.
  + Different areas affected in the same state, the same event type, the same cause, and the same start and end times will be combined.
  + If the same reporting area, the highest number is kept. If one has additional reporting then numbers are combined.
  + Duplicate special cases:
    - These are things like slight variance in times but the same specific reporting area such as California: Butte County.
    - 2007-9-18 5:15 and 9-18 5:14 events.
    - 2010-6-17 0930, all 3 keep the latest resolution reporting.
    - 2007-10-22 14:01:00 and 2007-10-22 14:05:00 merge, keep highest reporting.
    - 2018-08-07 01:22:00: latest time and highest reporting kept.2020-01-09 23:07:00 one-minute variance time of restoration.
    - 2021-02-15 01:54:00 and 2021-02-15 02:51:00 combined Texas: Travis County.
    - 2021-02-15 18:00:00 kept latest time.
    - 2021-02-16 06:48:00 state reported as separate and combined with another state, highest customer number kept.
    - 2022-01-01 12:36:00 highest time of restoration kept.
    - 2022-02-03 12:56:00 highest time of restoration kept.
* Deleted rows that have ALL of the following as unknown: area affected, event type, power, and customer.
* Clean up extra spaces and quotation marks.

Python

* Identify negative datetimes
  + For datetime64 correction: in the case of 00:00 to early morning, (0400) it is considered a wrong day issue as it is common for Americans to transition to the AM as if it is the same day in common talk. This is further supported by the times often starting in the late evening or near midnight.
  + In all other cases, the dates will be treated as if they are transposed and swapped accordingly.
* Compare machine learning algorithms and rule-based logic to determine the best fit for missing data.
  + Determined rule-based solution better mirrored U.S. population. See [GitHub repository](https://github.com/gregorywmorris/Maven-Power-Outage-Challenge) for us\_population.xlsx and Python code.
* Rule-based solution:
  1. Fill in blanks based on Event Type and State averages.
  2. Fill remaining based on Event Type and NERC Region averages.
  3. Fill remaining by just Event Type.
* Identify US states and Canadian provinces.
* Identify duplicate dates, and export the list to Excel for manual comparison to confirm if a merge is necessary.
* Values are assumed to be Missing Completely at Random (MCAR).
* Demand loss MW
  + Filled blanks where specified in Alert Criteria, accepting the highest if a range is given.
  + In the case of "Uncontrolled loss of (various numbers provided) Megawatts or more...", blanks filled in as 100.
* Customers
  + Fill in blanks where specified in Alert Criteria, accepting the highest if a range is given.
  + In the case of "Loss of electric service to more than 50,000 customers...", blanks filled in as 50000.
  + In some instances, the number reported may be less than suggested in the Alert Criteria. No correction was made.
* Column names to all upper case.
* Save as 'DOE\_final.xlxs'.