**Data Wrangling**

1. Project Objective
2. Selecting Data
3. Collecting Data
4. Compiling Data
5. Data Quality Problems and Remedies  
    Missing Data  
    Outliers

**1. Project Objective**Predict the 5 highest electrical power usage days in Ontario, in a given year

**2. Selecting Data**The body that manages electrical power supply in Ontario is the Independent Electricity System Operator (IESO). Their publication “Methodology to Perform Long Term Assessments”1 identifies the major factors that drive Ontario power usage:

1. Power Demand Data – Available in various formats
2. Calendar Variables – Weekdays, Weekends, Public Holidays
3. Weather – temperature, humidity, etc. Acute weather conditions drive peak demand
4. Economic Data – GDP – stronger economic activity drives power demand higher
5. Conservation – higher energy efficiency drives lower demand

**3. Collecting Data** **Historical Power Demand Data:** Available for every hour in yearly .csv files on the Independent Electricity System Operator’s (IESO) Website2. 19 files were manually downloaded from the website.

**Calendar Variables:** Ontario historical holiday dates are available in a python package – “holidays3”

**Weather:** This was a bit more complicated. There are hundreds of weather stations across Ontario, each collecting different data at different frequencies, with each station identified by a unique station number. Furthermore, over the years, stations may change the data collected, the collection frequency, and their station number. The Canadian Ministry of Environment provides a csv file containing all this metadata. The relevant station numbers and time periods were manually extracted. These were then used as parameters in a command line script4 to download the .csv files. The script ran 6 times to download 816 files.

**Economic Data:** Historical data by quarter available in an Excel Spreadsheet downloadable from the Ontario Ministry of Finance website5

**Conservation:** No aggregated publicly available data

**4. Compiling Data** The power data was in yearly csv files. These were indexed on a pandas datetime index, and compiled into a single multi-year file. The pre-2002 data was formatted slightly differently to the rest, requiring an enhanced data processing function. The date and time data was stored across multiple fields, requiring compilation to a single datetime field. The compiled dataset comprised 222,096 rows of data, with each row containing the total Ontario power usage every hour.

The weather data was compiled in a similar manner. This was slightly easier in that the field names were consistent across the whole data set. The compiled dataset comprised 582,248 rows of data for 9 variables, where each row represents an hour, and the variables include temperature, relative humidity, and wind speed.

**5. Data Quality Problems & Remedies**

**Missing Data:** The power data had a single full month of missing data out of a total of 300 months, or 0.32%.Unfortunately the missing month was December; a critical month in terms of peak power load. Fortunately, there are fairly consistent patterns from year to year, month to month, and day to day. Therefore, I searched for the closest in time (to keep economic conditions most similar) December data where the average monthly temperature of the replacement data was closest to the average monthly temperature of the missing data. The replacement data also needed to be offset so that the Christmas holidays lined up across the data sets. For repeatability, the processing data was captured in a code variable.

There was a small amount of missing data within the temperature data. Out of 582,433 data points there were 185 missing points, representing 0.03% of the total data. The missing data was randomly scattered through the dataset. Therefore, I felt comfortable imputing the missing data using linear interpolation.

**Outliers:** In August 2003, Ontario and the Northeast USA suffered one of the largest power outages in history6. Power was lost for several days. In the period following the outage, demand was artificially reduced through a public appeal to consumers. If ever there was an outlier, this is it. I considered removing this data, but thought a more representative strategy would be to fill across the 6 days with data from the closest equivalent days. For repeatability, this processing was captured in code variable.

I also investigated the impact of the 911 terrorist attacks in 2001, when all North American civil aviation was grounded for 3 days. Although power usage appeared to be slightly lower, there did not appear to be a significant impact relative to normal variation.

**Notes**

1. <http://www.ieso.ca/-/media/files/ieso/document-library/planning-forecasts/18-month-outlook/methodology_rtaa_2017mar.pdf>
2. <http://www.ieso.ca/Power-Data/Data-Directory>
3. <https://pypi.org/project/holidays/>
4. for year in `seq 2001 2019`;do for month in `seq 1 12`;do wget --content-disposition "http://climate.weather.gc.ca/climate\_data/bulk\_data\_e.html?format=csv&stationID=31688&Year=${year}&Month=${month}&Day=14&timeframe=1&submit=Download+Data" ;done;done
5. https://www.fin.gov.on.ca/en/economy/ecaccts/oea\_hist.xlsx
6. <https://en.wikipedia.org/wiki/Northeast_blackout_of_2003>

David MacGillivray – Springboard Capstone Project 1 - 27 June 2019