HEC MONTRÉAL

MATH60638A- Forecasting Methods

Report Part I:

Presented to:

Prof. Debbie Dupuis

Presented by:

Hamid Aliakbarlou– 11291818

Negar Aminpour – 11305212

Dhaval Patel– 11296446

Billy Vuong – 11169018

February 2023

1 Introduction

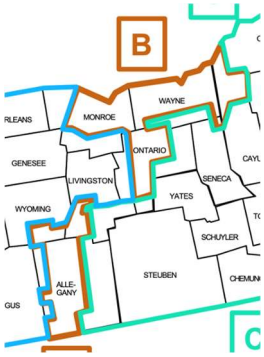

Figure 1.1 : Zone B Genese

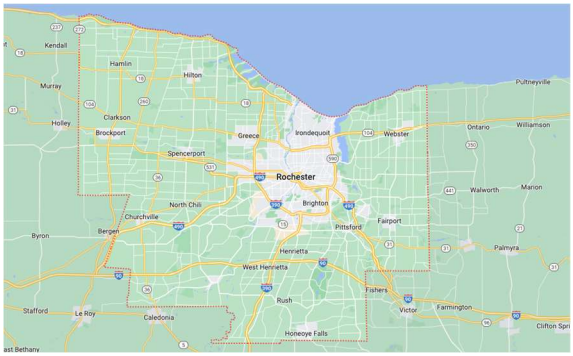

Figure 1.2 : Monroe County

This report is part 1 of 3 series to forecast the daily peak electricity load for Zone B, Genese, operated by NYISO. We explore the hourly demand dataset, evaluate naïve methods, and establish explanatory variables leading to peak demands. In Fig. 1.1, Zone B overlaps the following counties: Monroe, Wayne, Ontario, Allegany, Wyoming, Livingston, and Cayuga. Respectively their populations from Census 2020 are 759 443, 91 283, 112 458, 46 456, 40 531, 61 834 and 76 248. Monroe represents 64% of the population. Its growth is 2.02% from 2010 population of 744 344. We observed population growth in Monroe and Ontario from 2020 versus 2010 Census population. Further analysis uses Monroe County as the representation of Zone B due to its population importance and development [1]. We observe more roads and constructions against other counties in Fig. 1.2. On population characteristics, 18.3% are over 65 years, 81.7% are under 65 years and 20.5% are under 18. The major NAICS industry by employment goes to Health Care and Social Assistance 15.77%, Retail Trade 9.27%, Manufacturing 8.10%, Local Government 7.72%, Professional, Scientific, and Technical Services 7.41%, and Educational Services 7.25%. They account for 55.52% of employment sectors [2].

2 Exploratory data analysis

The hourly dataset ranges from 2015 to 2022 with 1 missing value replaced by the average of the prior and next hour to keep overall distribution. A dataset is reduced to peak daily load for manipulation.

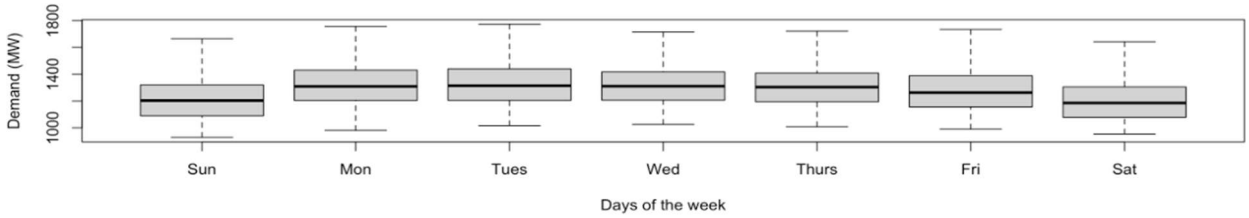

Figure 2.1: Zone B, Daily peak demands by days of the week in MW (2015-2022)

Starting the week, demand is low. A sharp average demand increases from Sunday to Monday and gradually decrease from Wednesday to Saturday. We assume higher buildings and services used during weekdays due to labour activities with 61.2% of population age between 18 and 65.

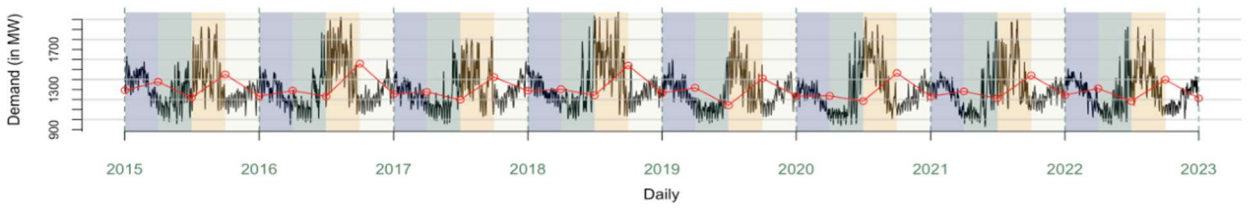

Figure 2.2: Zone B, Daily peak demand in MW (2015-2022)

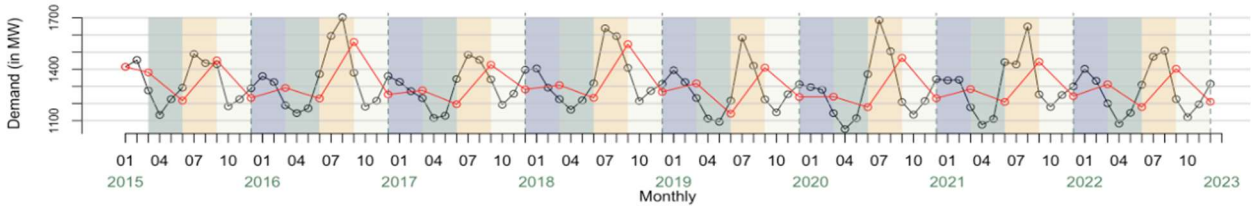

Figure 2.3: Zone B, Monthly average of daily peak demands in MW (2015-2022)

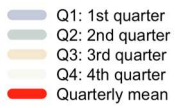

Fig. 2.2 shows a seasonal pattern with increases in Q1s and Q3s, and decreases in Q2s and Q4s. It is present in all years.

In the summer 2016, there is a stronger average monthly demand due to excessive temperature in the central and eastern U.S. caused by a heat dome [3]. In Fig. 2.3, 2019 Q2 has the worst monthly average demand since 2015, the first year to floor pass 1100MW. After this incident, subsequent

Q2s and Q3s in 2020-2022 have seen lower monthly average demands near 1100MW compared to previous years in 2015-2018. In general, 2019-2022 period has more monthly average demands dropping near 1100MW. The average quarterly means are relatively stable from one year to another compared to the 2015-2018 period. This might be due to changes in the economic cycle by COVID-19's worldwide impact, inhibiting human activities, businesses, and the global economy [4].

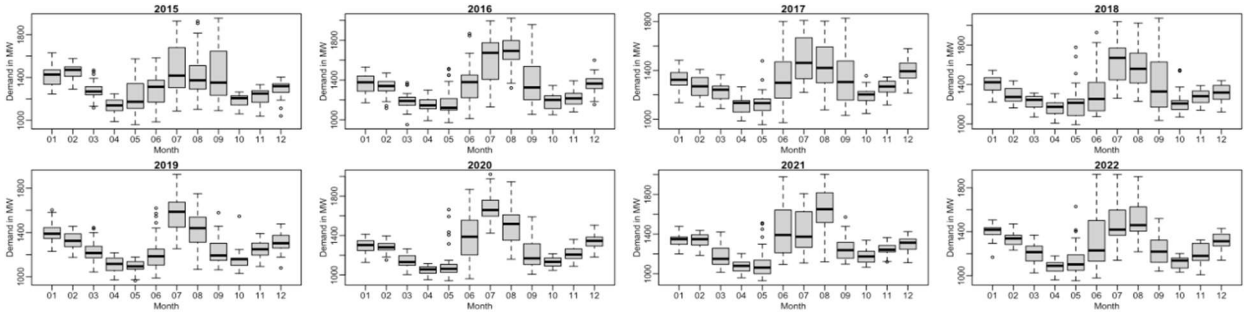

Figure 2.4: Zone B, Daily peak demands by month in MW (2015-2022)

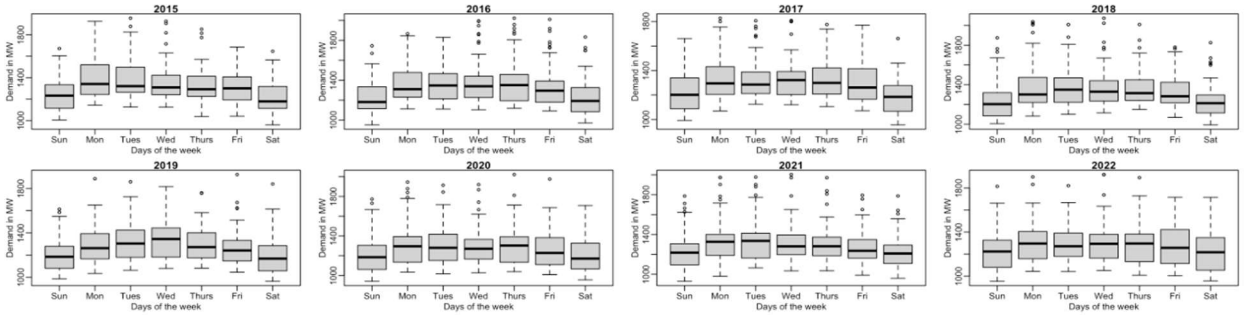

Figure 2.5: Zone B, Daily peak demands by days of the week in MW (2015-2022)

In Fig. 2.4, the summer season has the widest interquartile range, suggesting lower accuracy in forecasting. May has the most outliers, followed by March or June. In Fig. 2.5, the weekly pattern is the same as in Fig. 2.1 with fewer outliers in 2019 and 2022.

3 Train and test

We divide the dataset into train, validation, and test covering respectively 2015-2019, 2020-2021, and 2022. The training set is based on the recognition of yearly seasonal patterns and to capture

part of the COVID-19 effects prior to the Great Lockdown [4]. The validation set is given more weight in the remaining data for creating naïve benchmarks to compare with models or methods.

4 Evaluation of naive methods

The forecasting methods that we use are 1) Naïve (Random Walk) for its straightforward nature and ease of use because it assumes the future forecast will be similar to the current demand. 2) Seasonal Naïve method because it considers the seasonal patterns in the demand from previous years. 3) 7-day moving average is used because it uses the average of the 7-day period for the next period, and it can smooth out the fluctuation in electricity demand. 4) we also consider 3-day moving average.

Evaluating forecasts involves comparing the forecasted values to the actual values to determine the accuracy of the forecast on the validation set. Table 1 shows the results of 4 different naïve forecasts on the validation set, we can see that the Random Walk was the best in comparison to the others based on all criteria. But for further analysis, the mean absolute percentage error (MAPE) criteria were chosen because it does not depend on the scale.

Table 1:Methods evaluation

Methods	RMSE	MAE	MPE	MAPE
Naïve (Random Walk)	1945.02	1432.47	-0.25	5.33
Seasonal Naïve	3183.08	2305.13	-0.29	8.42
3-day moving average	2493.26	1886.43	-0.51	7.04
7-day moving average	2569.49	1879.04	-0.65	6.94

Moreover, based on the observed and forecast plot (Fig. 4.1), the Naive random walk method is found to be the best as it shows the closest prediction to the actual values. This method used the last day's observed value for the next day's forecast. The seasonal naive forecast is less accurate

here compared to other naive forecasting methods because it does not account for unpredictable fluctuations in demand, like a pandemic COVID 19, which will result in inaccurate forecasts [4]. Furthermore, it assumes that the demand for the specific season will be the same as it was during the same season in the previous year, instead of taking the most recent value. This "naïve" assumption does not take into account economic conditions and trend changes [5].

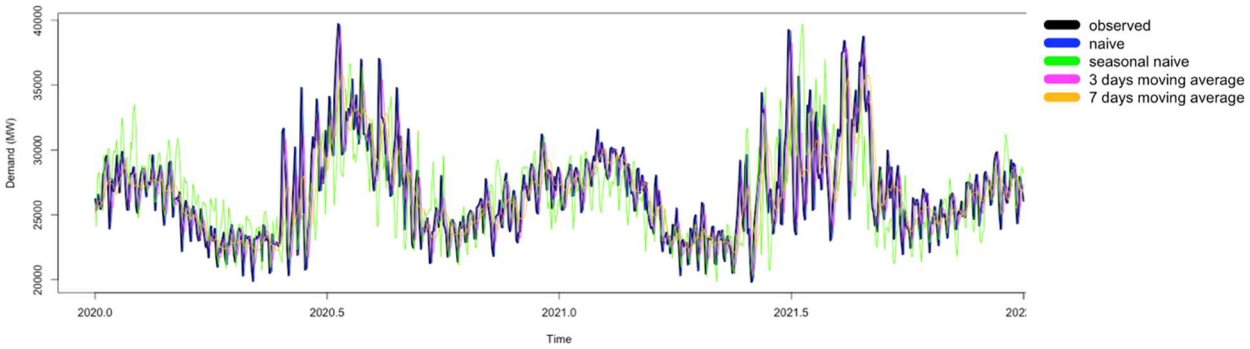

Figure 4.1: Naïve result plot

5 Explanatory variables

5.1 Selection of data stations

We decided to use weather data from the NOAA (National Oceanic Atmospheric Administration) website [6] as recommended by the project guidelines. This website contained a lot of weather data from various stations across the United States. The selection of said weather stations is based on the NYISO Day-Ahead Scheduling Manual [7]. It provides accurate weight representation for each zone base on-demand usage. Three weather stations located in airports are considered: Elmira, Rochester, and Syracuse. Among all available weather stations in the zone, airport stations are the most reliable, having fewer missing values. This might be because the airport is rarely vacant and thus defective sensors are noticed quickly. The station's names, IDs, and their associated weights are shown in table 1. These weights proportionally represent the contribution

of the stations in the weather data for the GENESE zone. We have thus combined the data originating from each of the three stations to make one single dataset. The location of the stations and the corresponding numbers are shown in the descriptive Fig. 3.1.

Zone	Station name	Station characteristics	Station weight factor	Location number (Fig 3.1)
GENESE	Elmira	ELMIRA CORNING REGIONAL AIRPORT, NY US View Full Details, Station ID: GHCND:USW00014748	5%	5
	Rochester	ROCHESTER GREATER INTERNATIONAL, NY US View Full Details, Station ID: GHCND:USW00014768	85%	18
	Syracuse	SYRACUSE HANCOCK INTERNATIONAL AIRPORT, NY US View Full Details, Station ID: GHCND:USW00014771	10%	20

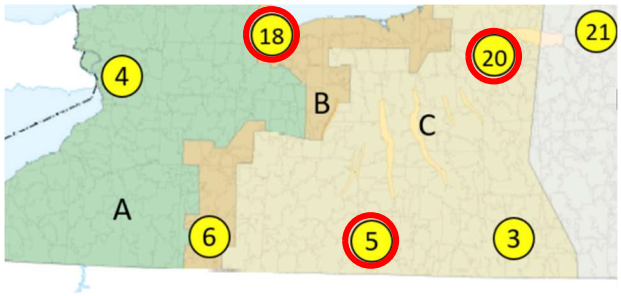

Figure 3.1: Location of the stations for GENESE zone

5.2 Explanatory Variables Analysis

After plotting the potential explanatory variables from the weather data extraction against the electricity demand from the NAOO [6], we concluded that the following variables (due their direct relation or the relation of variables resulting from their transformations with demand) are the most pertinent in our attempt to forecast peak daily electricity load:

T_{max} = Maximum temperature (Fahrenheit)	T_{min} = Minimum temperature (Fahrenheit)
HDD = Heating Degree Day	CDD = Cooling Degree Day
AWND = Average daily wind speed	CP = Wind Chill

We also investigated a potential relationship for Demand vs Precipitation and Demand vs Snowfall. However, we did not see a relationship between them. Hence, we omitted their investigated graphs.

5.2.1 Maximum Temperature (T_{max}) & Minimum Temperature (T_{min}) of a day

When it gets cold, we turn on the heating and when it gets hot we turn on the air conditioning. Thus, intuitively, we argue that one of the important candidates for influencing electricity consumption is temperature and in this section, its relationship with electricity demand is investigated.

In the features dataset, we have access to the average, maximum and minimum temperatures. However, to forecast daily peak demand, our preference is to use maximum and minimum temperatures. On the hot days of the year (summer), the maximum temperature is a better indicator of the usage of air conditioning and, as a result, explains better the peak demand. In the same way, in the cold days of the year (winter), the minimum temperature results in high heating and therefore indicates the peak electricity demand better.

As shown in Fig. 3.2 (shown in the next part), there is an almost quadratic relationship between demand and minimum temperature. The same relationship is identified with the maximum temperature and demand shown in Fig. 3.3. However, the two "tails" of the scatterplot do not show enough symmetry for this to be considered a quadratic relationship. Hence, to reveal some sort of underlying relationship beyond what we see initially, we need to treat the maximum and minimum temperature variables in an adequate way to segregate the hot and cold "ends" versus electricity demand. This is done in the following part.

5.2.2 Heating Degree Day (HDD) & Cooling Degree Day (CDD)

As important climate indicators, HDD and CDD are commonly used in investigating the impact of climate change on forecasting future energy demand [8]. The formula to calculate these two indicators are as follow:

$$HDD = Max\big(\,(T_{ref} - T_{min})\,,0\,\big) \qquad CDD = Max\big(\,(T_{max} - T_{ref})\,,0\,\big)$$

Based on the daily linear relationship between the maximum temperature (T_{max}) and peak demand starting from 65 °F (18.3 °C) onward, the reference temperature (T_{ref}) for CDD is chosen. This behavior can be seen in Fig. 3.2. We can infer that people on average start using cooling facilities in the GENESE region of New York state at temperatures above 65 °F and as temperature increases, their demand also increases linearly. With the same logic, T_{ref} for HDD is considered to be 45 °F (7.2 °C). This choice is based on the linear relationship between daily demand and daily minimum temperature (T_{min}) at temperatures below 45 °F which is shown in Fig. 3.3.

By treating and transforming temperature into HDD and CDD, the linear trend between the peak demand and these two indicators is revealed and shown in Fig. 3.4 and Fig. 3.5.

We did the same plots for CDD lag 1 and lag 2 and it shows a linear relationship with daily peak demand as seen in Fig. 3.6 and Fig. 3.7.

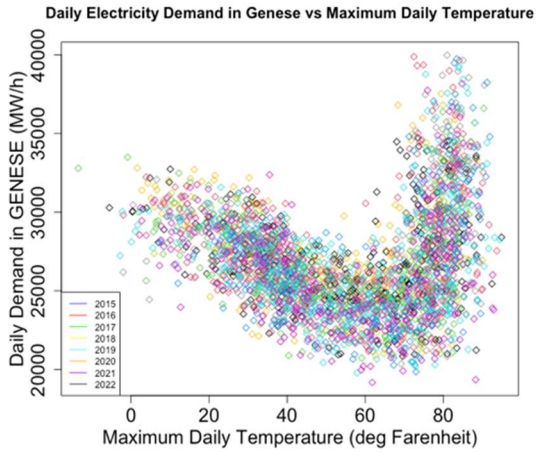

Figure 3.2: Min Temperature vs Demand

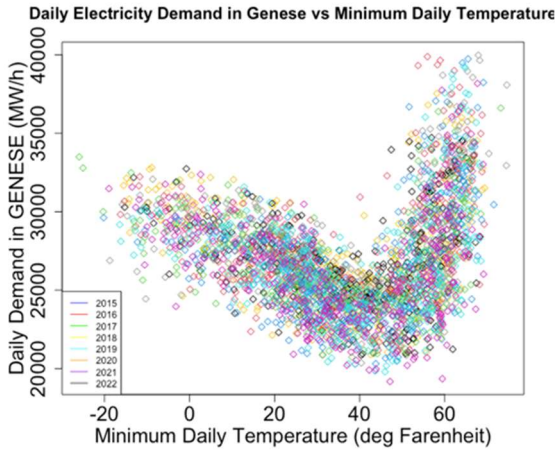

Figure 3.3 Max Temperature vs Demand

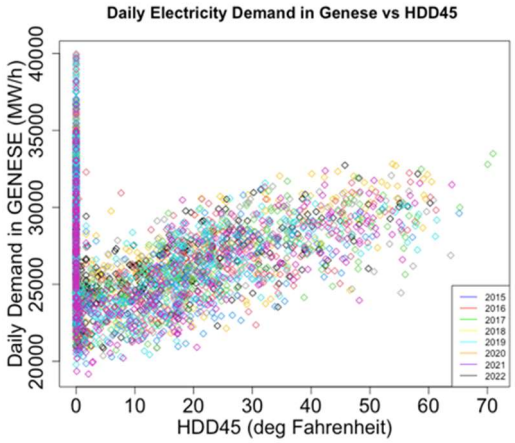

Figure 3.4: HDD vs Demand

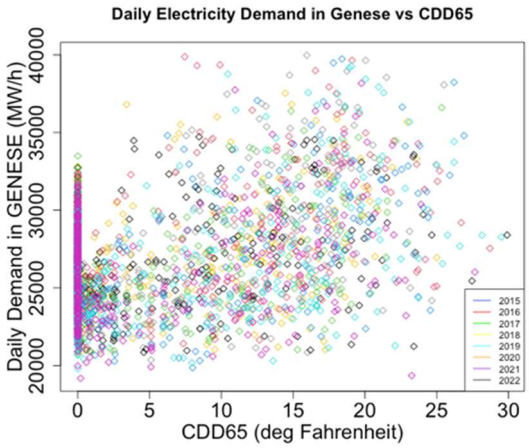

Figure 3.5: CDD vs Demand

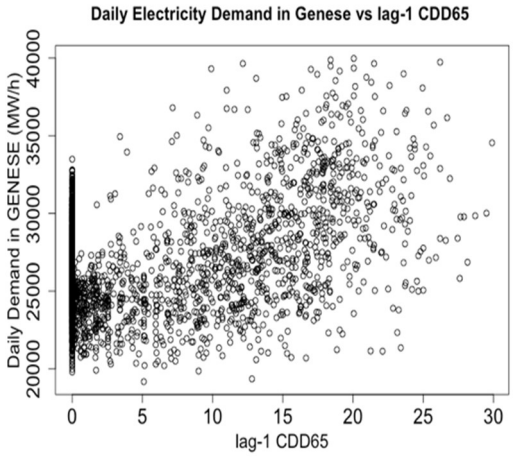

Figure 3.6: CDD lag-1 vs Demand

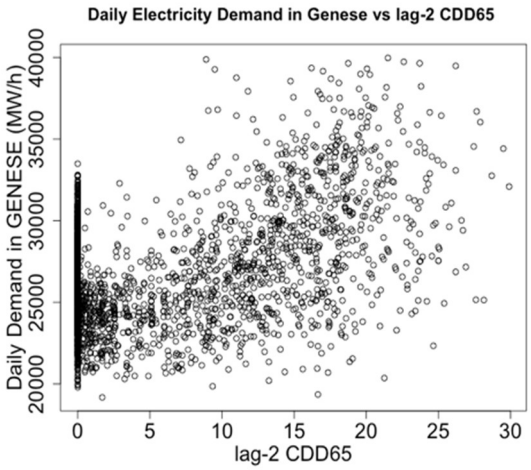

Figure 3.7: CDD lag-2 vs Demand

5.2.3 Wind Chill (CP_t)

Wind chill is based on the fact that with rise of wind speed, the heat loss also increases, thereby making the air "feel" colder. Wind chills can result in a demand rise for electricity as people attempt to heat their homes and businesses in cold weather. To investigate its significance and effect on load demand, CP vs Demand is plotted in Fig. 3.8. It shows us a linear relationship with more windchill resulting in higher peak load.

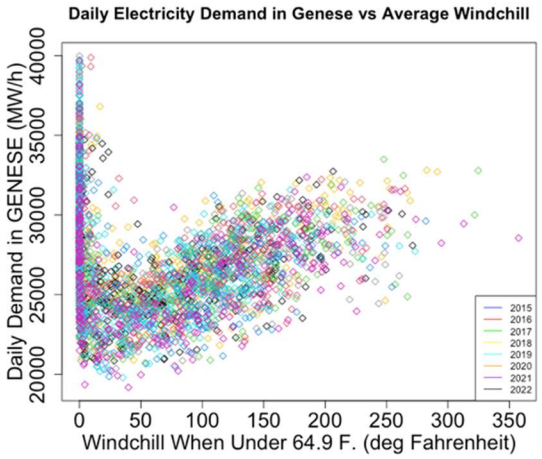

Figure 3.8: Wind chill vs Demand

5.2.4 Dummy Variable for Weekdays vs. Weekends

We consider creating a dummy variable to better capture the difference of weekdays vs. weekends when it comes to electricity demand. This consideration is backed by Fig. 2.5, where we see in the boxplots that year to year, weekdays seem to have a higher electricity demand than weekends. Beyond the boxplot analysis, we can not say anything else about this potential variable as we do not know how to test this relationship further at this point in the class.

6 References

[1] U.S. Census Bureau QuickFacts: United States. (n.d.). Retrieved February 7, 2023, from https://www.census.gov/quickfacts/fact/table/US/PST045222

[2] Monroe County Employment Structure & Growth by Major Industry, 1969-2021. (n.d.). New York Regional Economic Analysis Project. Retrieved February 9, 2023, from https://new-york.reaproject.org/analysis/industry-structure/industries_by_region/employment/tools/360055/

[3] Hersher, R. (2016, July 22). "Heat Dome" Causing Excessive Temperatures In Much Of U.S. NPR. https://www.npr.org/sections/thetwo-way/2016/07/22/487031278/heat-dome-causing-excessive-temperatures-in-much-of-u-s

[4] COVID-19 recession. (2023). In Wikipedia. https://en.wikipedia.org/w/index.php?title=COVID-19_recession&oldid=1136843107

[5] Hyndman, Rob J., and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018.

[6] Climate Data Online (CDO) | National Climatic Data Center (NCDC). (n.d.). Retrieved February 10, 2023, from https://www.ncdc.noaa.gov/cdo-web/search

[7] NYISO. (2022). Day-Ahead Scheduling Manual. https://www.nyiso.com/documents/20142/2923301/dayahd_schd_mnl.pdf/0024bc71-4dd9-fa80-a816-f9f3e26ea53a

[8] Cawthorne D, de Queiroz AR, Eshraghi H, Sankarasubramanian A, DeCarolis JF. The Role of Temperature Variability on Seasonal Electricity Demand in the Southern US. Frontiers in Sustainable Cities. 2021 Jun 2;3:644789.