Forecasting Electricity Demand and Prices in PJM and MISO Markets

Data Manipulation & Descriptive Statistics

A Data-Driven Approach for Efficient Utility Management

Group 5

Team members :

Renee Chang - 1009640051
Ke Han Bei - 1009338236
Zhanglin Liu - 1009315974
Miaoyan Qi - 1008786388
Ba Minh Dang, Le - 1006136272
Yian Yan - 1009103239

Data-Cleaning: Transforming Data

Cleaning Data

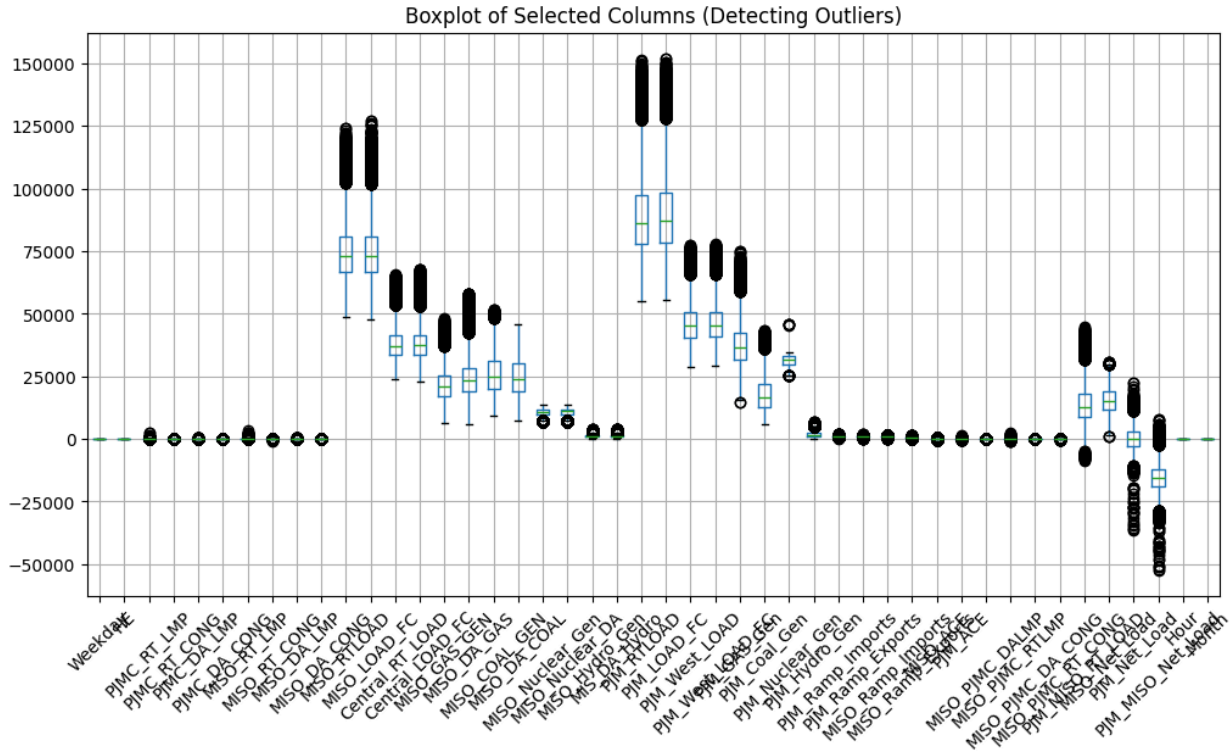
Prior to importing the dataset into python, we perform a simple data cleaning process with the original csv in excel to remove all the colons in the original dataset to make it easier to work with. Then we pre-configure the data type of each column in python to avoid any data type conflicts during the analytics

Upon investigation we realized that the dataset has more than enough data so we decided to remove all missing data points to ensure that the missing values will not affect the analytic process. We treat any records with missing values and incomplete records. In addition, The code will count the missing rows in any column and automatically remove these rows from the dataset. After the data cleaning process, the dataset still has more than 40 thousand records.

Outliers

We use Boxplot to visualize the data and IQR to find outliers. In the analysis of electricity market data, Boxplot can be used to observe abnormal changes in load demand (such as rapid rises or drops in electricity demand during specific periods) as well as detect unexpected variations in electricity prices (such as changes in LMP prices in PJM and MISO markets). Furthermore, it can assist us in determining whether ACE contains outliers to comprehend market stability.

Our goal is to predict daily and hourly demand to improve grid performance and reduce the risk of power outages. In addition, we also want to identify seasonal and annual usage trends, especially in extreme weather conditions, to improve infrastructure development and capacity planning.



From the visualization results of the boxplot, multiple variables exhibit significant outliers, among which:

- **Electricity price related variables (PJMC_RT_LMP, MISO-RT_LMP):** There are a large number of outliers, indicating that market prices fluctuate greatly, especially during certain periods when extreme price changes may occur, such as sudden increases in electricity demand or surges in electricity prices caused by transmission network congestion.
- **Load data (MISO-RTLOAD, PJM-RTLOAD):** Outliers might be caused by seasonal changes, extreme weather, or sudden surges in demand.
- **Regional control errors (MISO-ACE, PJM-ACE):** Outliers show significant imbalances in the power grid during certain periods, which may pose challenges to power dispatch.

After using the IQR method to calculate outliers, we selected real-time marginal price (LMP), transmission congestion cost (RT_CONG), real-time load (RTLOAD), and area control error (ACE) as the key outlier data to focus on:

LMP Outlier:

- Unusual fluctuations in electricity prices may be caused by factors such as extreme weather, transmission bottlenecks, and sudden increases in demand.
- The lack of LMP outliers may lead to increased market prediction errors because market prices themselves have relatively high volatility.
- Retaining abnormal LMP values can help improve the adaptability of the model in extreme situations, which can assist market operators in better predicting price fluctuations.

RT_GONG Outlier:

- The cost of transmission congestion can reflect the bottleneck and load scheduling problems of the power grid.
- The cost of extreme transmission congestion can correspond to peak loads or emergency situations in the power grid. If it is missing, there may be a probability of masking important market information.
- If the model cannot identify extreme cases of transmission congestion, the predicted results may deviate significantly from reality. This will lead to market participants making incorrect decisions.

RTLOAD Outlier:

- Real demand variations in the market are represented by RTLOAD outliers. For example, severe weather (such cold waves), changes in industrial activity, or seasonal peaks (higher heating load in winter, higher air conditioning load in summer).
- It could show how supply and demand are out of balance in the market. If RTLOAD outliers are removed, load demand during peak hours may not be correctly predicted, affecting power supply planning.
- It can also reflect potential problems in power grid operation, such as sudden load surges that may lead to power shortages. This may affect market prices.

ACE Outlier:

- Regional Control Error (ACE) represents the ability of power grid scheduling to match actual load demand. The ACE outlier indicates a mismatch between power grid dispatch and market demand. This may lead to frequency fluctuations or insufficient power supply.
- A high ACE may be due to delayed response from generator sets, transmission network congestion, or deviation in demand forecasting. These pieces of information are crucial for optimizing the stability of the power grid.
- If the ACE outlier is removed, it may affect the optimization effect of power dispatch.

We finally decided to keep these outliers:

- **LMP (PJMC_ST_LMP, MISO-RT_LMP):** Real time marginal electricity prices (LMP) have many outliers, indicating severe fluctuations in short-term electricity prices in the market. This may be due to a shortage of power generation resources, sudden surge in demand, or transmission network bottlenecks.

- **RT_CONG (PJMC_CNT-CONG, MISO-RT_CONG):** A large number of abnormal values in the transmission congestion cost (RT_CONG) indicate that the power grid often encounters transmission bottlenecks. Therefore, the market might need to make further changes to the grid load allocation during specific times.

- **RTLOAD (MISO-RTLOAD, PJM-RTLOAD):** A large amount of real-time load data outliers may be due to seasonal variations or adverse weather conditions.

- **ACE (MISO-ACE, PJM-ACE):** At times, there may be significant differences between load dispatch and the actual demand of the power grid. For this, additional scheduling strategy optimization may be required.

Creating and Transforming Data

Upon investigation, we decided to add the following attributes on the existing data in hope of getting further insights.

1. **Day_Type:** This attribute categorizes each date as either a "Weekday" or a "Weekend," based on the day of the week.
    a. This attribute can be used to analyze weekly electrical load and price variability
    b. Segment the data by weekend and weekday can provide deeper understanding in typical behavioural differences in electricity usage between these period during any given week
2. **Hour_category:** This attribute categorizes each date as either a "Weekday" or a "Weekend," based on the day of the week.
    a. We will use this attribute to further study the patterns in electricity demands through any given day.
    b. By categorizing the hours, we can more precisely analyze demand and price fluctuations during different parts of the day.
    c. We will also use this attribute to study the electrical Import and Export loads in different parts of a date to finds the load pattern of each grid
3. **Season**
    a. We extracted the month from the Date_Time column and categorized it into four seasons, storing the result in a new column.
    b. This step is crucial because electricity demand and pricing are highly seasonal. Summer & Winter: Higher energy demand due to cooling and heating. Spring & Fall: Transition seasons where electricity usage patterns shift, leading to the electricity consumption being less predictable.
    c. Grouping by seasons helps us observe seasonal variations in electricity demand, congestion cost and price changes to improve forecasting accuracy. It also allows us to adjust future machine learning models by incorporating seasonal adjustment factors.

Adding Columns: LMP Price Difference (Real-Time vs. Day-Ahead)

One of our target variables is Real-Time Locational Marginal Pricing (RT LMP), which reflects real-time electricity prices. To evaluate how accurately the market predicts these prices, we created two new columns that calculate the difference between Real-Time (RT) and Day-Ahead (DA) LMP for PJM and MISO markets. It measures how well the market predicts electricity prices and possible adjustments are needed.

A large LMP difference indicates large forecasting errors, suggesting there might be unexpected demand surges, congestion, or generator outages. Analyzing the price difference can enhance forecasting accuracy. This contributes to grid performance optimization by predicting potential demand spikes that aren't included in day-ahead markets, reducing the risk of blackouts.

Adding Columns: Net Export (Numerical & Categorical Variables)

To capture the electricity trade dynamics between PJM and MISO, we created two columns calculating net exports for PJM and MISO markets.

$$\textit{Net Exports} = \textit{Ramp Exports} - \textit{Ramp Imports}$$

We also created a categorical column Export Status that classifies each market as Net Exporter, Net Importer or Balanced. As a potential independent variable, net export is important to our analysis, because it influences electricity pricing and grid stability.

- If a region is a net importer, there may be an energy shortage, probably due to insufficient local generation or higher congestion costs.
- If a region is a net exporter, there might be an energy surplus, potentially lower prices.

It helps us to analyze how net exports correlate with extreme price movements. Also, it improves price forecasting by incorporating grid dependency into the model.

Adding New Columns: Congestion Price Difference (Real-Time VS. Day-Ahead)

Difference between day ahead and real-time congestion costs for PJM and MISO Analyzing the congestion costs allows us to examine the additional charges incurred when there is not enough transmission capacity to meet high demand. The source of these costs are from energy generators that incur a higher cost as current transmission has reached its capacity. This aligns with our objective of forecasting electricity demand in the PJM and MISO electricity

markets to focus on trends in demand, and to have a better idea of how the capacity is used. Congestion costs are also a part of marginal pricing, therefore analyzing it will improve our forecast of demand and prices to understand when demand exceeds transmission capacity.

We have created two columns to measure the difference between real-time congestion and day-ahead congestion for PJM and MISO to analyze if the day-ahead congestion costs are higher, which indicates an overestimation as congestion costs are forecasted the day before in the day-ahead market. The difference will present how the actual costs compare to the forecast, showing either a possible demand rise or overestimating costs. To interpret our data, the mean, standard deviation, minimum, and maximum were calculated to illustrate what the statistics are telling us.

Adding columns: Detecting for Anomalies

To put it in simple terms, the primary goal for performing anomaly detection in the electricity load, by comparing Real-Time Load (RTLOAD) with Forecasted Load (LOAD_FC) for both markets, was to test out if there were any unusual patterns in the electricity demands. Analyzing and detecting anomalies is crucial for determining whether the actual demands for electricity deviate from the forecasted which could cause problems such as building excessive power grid (Wasted electricity, and operational inefficiency), excessive cost that may result in profit or revenue lost, and ratepayers unable to make the end meets if unnecessary cost incurred from the inefficient forecast of the electricity demand. These new columns created by using differences between actual and forecasted using rolling z-stats (Attributes) would give us insight into the anomalies related to the electricity demand that occurred over the period, thus, we can adjust if there is any deviation from the forecast value.

In percentage form	MISO_Anomaly	PJM_Anomaly
Normal	94.63	95.13
Drop	2.37	2.87
Surge	2.50	2.50

We constructed the new attribute(‘Anomaly’ for PJM and ‘Anomaly1’ for MISO) by using the difference between actual and predicted load, and based on the rolling mean and standard deviation, we computed the rolling z-score to identify whether the difference lies within the 97.5th percentile, if data lies beyond the threshold, we would consider the data’s to be unusual values that does not occur frequently. From the results we can conclude that although the skewness of anomalies (using rolling z-scores) for MISO (skewness: -0.05) and PJM (skewness: 0.15) is slightly left skewed, it does not deviate far away from the center, therefore indicating a precise prediction of the electricity load for both markets. As we also focused on the aspect of proportion of ‘surge’, ‘drop’, and ‘normal’ (in terms of electricity load), we can conclude that ‘normal’ consists of up to MISO (95.13%) and PJMC (94.63%) of the total proportion of the anomalies for both markets. In contrast, drop and surge only consist of up to 2.57% (on average) of the overall proportion in the anomalies column. In summary, the results show that the actual data load for electricity does not deviate much from the predicted values, meaning the expected load is accurately forecasted.

Filtering the data

We will focus on analyzing the impact of different factors on trends in electricity demand and use, both in terms of time and place. First, in terms of the time factor, we will look at the columns of ON_OFF, Weekday, and seasons to see if these categorical variables are related to MISO / PJ's electricity demand.

We will use Columns: PJMC_RT_LMP, MISO_RT_LMP, ON_OFF, Hour_Category, and by filtering Hour_Category by value (morning, afternoon, evening, and nighttime) and ON/OFF time period, we can conclude that prices in the afternoon are consistently high, indicating that electricity demand is strong during these times of the day, while overnight demand remains low. Analyzing electricity demand at different points in time and during peak periods can help energy suppliers optimize pricing strategies and ensure that supply effectively meets demand. By filtering Day_Type (weekdays vs. weekends), we can conclude that demand is usually higher on weekdays, while weekends may see a decrease in demand.

We know that electricity demand fluctuates greatly with the seasons, and by filtering the Season column (Winter, Spring, Summer, Fall) we can conclude that loads are consistently higher in the winter months, which can be attributed to the high demand for heating in cold weather; hot days in the summer months see a spike in demand, which can be attributed to the increased demand for refrigeration and air conditioning; and demand is more stable in the spring and fall months.

Grouping Data

We think the following groupings combination will be useful

1. **Group by ON/OFF and Day_Type:** This grouping can helps to understand the dynamics of electricity pricing in relation to demand patterns influenced by time of the day and week

		PJMC RT LMP		MISO RT LMP	
ON_OFF	Day_Type	min	max	min	max
OFF	Weekday	-20.18	432.03	-21.28	459.36
	Weekend	-22.76	312.89	-13.59	2540.00
ON	Weekday	-45.62	2213.58	-88.93	3430.14
	Weekend	-6.53	917.37	-81.14	1048.36

Insight: Both markets exhibit wider price ranges during weekdays compared to weekends regardless of whether its peak usage periods or not.

2. **Grouping by ON/OFF and Hour_Category:** This grouping combinations reveals interesting insights about electricity pricing dynamics during different times of the day and under varying demand conditions

	PJMC_RT_LMP	MISO_RT_LMP

ON_OFF	Hour_Category	mean	std	mean	std
OFF	Afternoon	N/A	N/A	N/A	N/A
	Evening	29.53	21.93	27.71	17.85
	Morning	28.25	19.99	27.73	42.43
	Overnight	25.59	17.20	24.89	47.87
ON	Afternoon	39.83	46.00	40.05	55.71
	Evening	42.02	45.19	40.38	52.48
	Morning	35.88	33.52	34.39	32.20
	Overnight	N/A	N/A	N/A	N/A

Insights:

The N/A suggest that there are no OFF peaks usage during afternoon and no ON peaks electricity usage during Overnight which means that Afternoon is an always high demand period and Overnight is an always low demand period in both markets.

However it's important to notice that while the average electricity price seems to not vary much across different times during the day, the standard deviation tells us a different story. We can see that the volatility level in both grids is much higher during peak hours except for MISO overnight off peaks.

Sorting Data

Sorting is not considered to be necessary since the dataset is already in time order and our analysis relies on grouping rather than sequential ordering. We focus on transforming, filtering, and grouping the data to provide meaningful insights for forecasting electricity demand and price volatility.

Descriptive Statistics

Summary of descriptive statistics for PJMC and MISO Real Time Locational Marginal Price per MWh.

Unit: $/MWh	PJMC Real Time LMP	MISO Real Time LMP
Mean	35.00	34.14
Median	25.02	24.77
Mode	20.27	21.13
Standard Deviation	36.39	46.62
Minimum	-45.62	-88.93
25%	19.84	18.35
50%	25.02	24.77
75%	39.17	39.61
Maximum	2213.58	3430.14

Skewness Values for PJMC and MISO Real Time Locational Marginal Price per MWh.

	PJMC Real Time LMP	MISO Real Time LMP
Skew value	16.95	32.07

Insight on PJMC and MISO Real Time LMP

- **Mean:** The average price of electricity in the PJMC market over the analyzed period. The average price is $35.00 and $34.14/MWh
- **Median:** The middle value of all electricity prices. At $25.02 and $24.77, the median is significantly lower than the mean, suggesting a positively skewed distribution
- **Mode:** This is the most frequently quoted price at $20.27 and $21.13, further supporting the skewness towards lower prices.
- **Standard Deviation:** Measures how wide the spread in electricity is. A high standard deviation of $36.39 and $46.62 illustrates considerable volatility in PJMC's electricity prices, indicative of frequent and significant deviations from the average price.
- **Min/Max:** The minimum price of -$45.62 and -$88.93 suggests instances of negative pricing indicating periods of extremely low demand or high renewable energy generation. The maximum price of $2213.58 and $3430.14 reflects extreme spikes in prices where demands are high and electricity supply is low or even disrupted. These extremes could indicate specific events or anomalies such as extreme weather events.

- **Quartiles (25/50 and 75th percentiles):** The first quartile and the third emphasize the concentration of prices in lower prices range but with a significant spread towards higher prices.
- **Skewness Value:** Skew values of 16.95 and 32.07 are extremely high, emphasizing extreme positive distribution with long right-tailed. This extreme skewed tails significantly influences the average price.

Summary of descriptive statistics for PJMC and MISO Real Time Load in MW

Unit: MW	PJMC Real Time Load	MISO Real Time Load
Mean	88791.98	75018.86
Median	86320.05	73225.44
Standard Deviation	15816.19	11860.32
Minimum	54936.83	48827.33
25%	77657.07	66669.41
50%	86320.05	73225.44
75%	97550.45	80960.48
Maximum	151570.04	124229.33

Skewness Values for PJMC and MISO Real Time Locational Marginal Price per MWh.

	PJMC Real Time LMP	MISO Real Time LMP
Skew value	16.95	32.07

Insight on PJMC and MISO Real Time Load

- **Mean:** The average Real Time Load in both grids are 88791.98 and 75018.86 MW
- **Median:** At 86320.05 and 73225.44 MW, the median is slightly less than the mean suggesting a relatively symmetrical load distribution with a slight tilt towards lower values
- **Standard Deviation:** High Standard Deviation at 15816.19 and 11860.32 MW indicates considerable variability in electrical load in both grids. In addition, the wide standard deviation suggests operational conditions and demand can vary significantly from day to day.

- **Min/Max:** The minimum real time load in both grids are 54936.83 and 48827.33 MW. while the maximum real time load is 151570.04 and 124229.33 MW. This wide range indicates extreme variations, potentially reflecting changes due to seasonal demand, weather conditions, or operational incidents.
- **Quartiles (25/50 and 75th percentiles):** The first quartile and the third emphasize a wide spread in the middle of 50% of data
- **Mode:** The mode indicates most frequent occurring loads on both grids suggesting that these are common operational points on the grids
- **Skewness:** Both grids have skewness values between 0.5 and 1 suggest a moderate right skew in distribution.

Statistical Analysis of LMP Differences (Real-Time vs Day-Ahead)

	PJM_LMP_Diff	MISO_LMP_Diff
mean	-0.57	-0.28
std	28.95	39.69
min	-196.98	-273.01
25%	-5.45	-5.26
50% (median)	-1.95	-1.47
75%	0.44	1.41
95%	15.07	16.74
max	2105.50	3272.79

- The mean is -0.57 and -0.28, respectively. It's close to zero, probably due to distribution in both positive and negative sides, indicating both overestimation and underestimation for electricity price.
- The standard deviation is 28.95 and 39.71, respectively. It might be due to a large price difference.

- The 5th, 25th, 50th, 75th, 95th percentile indicates that the majority of price differences are small, suggesting that day-ahead electricity prices usually align with real-time predictions.
- The min and max value are very large, meaning that huge spikes still exist and are impactful. Forecasting models should focus on predicting those extreme cases, as they can cause severe market disruptions.

Statistical Analysis of Congestion Cost:

Statistics	PJM Congestion Cost Difference	MISO Congestion Cost Difference
Mean	0.076019	-0.635066
Standard Deviation	6.416086	12.249336
Minimum	-203.0	-743.27
Maximum	180.07	157.14

The mean difference of PJM congestion cost difference is approximately 0.076019 and the standard deviation is 6.416086. The mean difference of MISO congestion cost difference is approximately -0.635066 and the standard deviation is 12.249336. The minimum value for PJM congestion cost difference is -203, and for MISO is -743.27. The maximum value for PJM congestion cost difference is 180.07, and for MISO is 157.14.

The positive mean difference indicates that overall, the real time congestion costs for PJM is slightly higher than the forecasted number which means demands exceed capacity on a somewhat higher occasion. Meanwhile, MISO has a negative mean difference for congestion costs, which means its real-time congestion costs on average is lower than the predicted value. This implies an overestimation of day-ahead congestion costs for MISO. The standard deviation for MISO is much higher than PJM, meaning that the difference of congestion costs for MISO fluctuate significantly, while PJM has a more stable cost difference. This may imply that demand is more volatile for MISO that causes the fluctuations and the variance. Additionally, MISO also has a more extreme value of -743.27 as a minimum difference, which means the day-ahead

congestion cost was overestimated by $743.27/MWh. This suggests that MISO may need to develop their forecasting of congestion costs as this impacts the price heavily, which also hurts our long term capacity planning.

Distribution of Net Export:

	PJM_Export_Status	MISO_Export_Status
Net Exporter	46	12
Net Importer	112	932
Balanced	43670	42884

	PJM_Net_Exports	MISO_Net_Exports
Fall	0.6	-497.90
Spring	-0.21	-496.53
Summer	-0.37	-496.25
Winter	0.10	-496.89

Period with balanced status for PJM and MISO is 43,670 and 42,884, respectively. It indicates that the market is relatively stable, most of the time local supply meets local demand.

MISO is a Net Importer in 932 periods, significantly higher than PJM (112 periods). MISO also has negative amounts of net export for all seasons, indicating that MISO depends more on external electricity sources, which might make it more sensitive to price spikes during periods of high demand. In contrast, PJM has net exports nearly zero for all seasons, suggesting it's more self-sustaining with more efficient electricity generation.