Project Report on

Load Forecasting for New Hampshire – January 2023

Submitted by:

Nidhi Trivedi
Janhavi Gunjal
Mohit Shyam Amode

Abstract

This project focuses on developing a robust model for forecasting the hourly electricity load in New Hampshire for January 2023, utilizing the response variable "RT_Demand." The analysis relies on historical hourly data from the "2022 SMD Hourly Data," "2021 SMD Hourly Data," and "2020 SMD Hourly Data." Selected features, including Date, Hr_End, Dry_Bulb, and Dew_Point, are considered, with additional transformations such as squared terms and day-of-week variations. The exclusion of pricing-related information and data beyond January, 2023, maintains the project's focus on load forecasting. Leveraging advanced statistical and machine learning techniques, the model aims to provide crucial inputs for traders to anticipate hourly power prices, contributing to informed decision-making in the dynamic energy market. The project acknowledges limitations and outlines a structured timeline covering data pre-processing, feature engineering, model development, and validation. Ultimately, this undertaking seeks to enhance the efficiency of energy trading by delivering reliable hourly load forecasts for January2023.

Table of Content

Introduction

This project endeavors to construct a precise and dependable model for forecasting the hourly electricity load in New Hampshire for December 2022, utilizing historical hourly data amalgamated from the "2022 SMD Hourly Data," "2021 SMD Hourly Data," and "2020 SMD Hourly Data." The exclusive focus on load forecasting is maintained by deliberately excluding pricing information, and data beyond December 1st, 2022. Leveraging advanced statistical and machine learning techniques, the model will incorporate features such as Date, Hr_End, Dry_Bulb, and Dew_Point, alongside their transformations, for enhanced accuracy. The amalgamation of the three datasets aims to create a comprehensive view of hourly load patterns in New Hampshire, crucial for understanding daily and seasonal variations. Beyond the intrinsic significance of accurate load forecasts for traders in decision-making, this project contributes to the efficiency of the energy trading market, benefitting consumers and industry stakeholders alike. The phased project timeline ensures rigorous validation and adjustments, culminating in a robust model that empowers traders with valuable insights into anticipated hourly power prices, fostering more informed decision-making in the dynamic energy market.

Business Purpose

Objective:

The primary objective of this project is to develop a precise and reliable model for predicting the load in New Hampshire during January 2023. The focus is on generating accurate forecasts that reflect the expected electricity consumption patterns throughout the month.

Importance:

1. Foundational Input for Traders:
2. Informed Decision-Making:
3. Risk Mitigation
4. Strategic Planning
5. Competitive Advantage

Data Overview

The selected datasets, spanning from 2020 to 2022, provide a comprehensive view of the hourly load patterns in New Hampshire. The temporal granularity of the data allows for a detailed analysis of daily and seasonal variations, which are critical factors in load forecasting.

Below are column and the description details :

Date	The calendar date of the provided data
Hr_End	The hour of the observation, in hour ending and 24-hour convention
DA_Demand	Day-Ahead Cleared Demand, in MW, comprised of cleared fixed and price-sensitive demand bids plus the net of cleared virtual activity; ISO NE CA value is the sum of the load zones and the Hub values
RT_Demand	Real-Time Demand, in MW, Non-PTF Demand for wholesale market settlement from revenue quality metering, including non-dispatchable load assets, station service load assets, and unmetered load assets; starting June 1, 2018, includes the grossed up demand response value
DA_LMP	Day-Ahead Locational Marginal Price (LMP) in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
DA_EC	Energy Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
DA_CC	Congestion Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
DA_MLC	Marginal Loss Component of Day-Ahead LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
RT_LMP	Real-Time Locational Marginal Price (LMP) in $/MWh by load zone; starting March 1, 2017, this is the hourly average of the five-minute LMP in the hour; 'ISO NE CA' tab contains values for the Trading Hub
RT_EC	Energy Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
RT_CC	Congestion Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
RT_MLC	Marginal Loss Component of Real-Time LMP in $/MWh by load zone; 'ISO NE CA' tab contains values for the Trading Hub
Dry_Bulb	The dry-bulb temperature in °F for the weather station corresponding to the load zone or Trading Hub; summer period is June-September, winter period is October-May
Dew_Point	The dewpoint temperature in °F for the weather station corresponding to the load zone or Trading Hub; summer period is June-September, winter period is October-May

Data Processing and Analysis Overview:

As part of the data preparation phase for this project, a unified dataset spanning the years 2020, 2021, and 2022 has been created by manually combining the individual Excel sheets. The focus of this project centers on five key columns: 'Date,' 'Hr_End,' 'Dry_Bulb,' 'Dew_Point,' and 'RT_Demand.' Additionally, two new columns will be introduced, representing the squared values of 'Dry_Bulb' and 'Dew_Point,' which will be created in the subsequent code.

Prior to any analysis, data cleanliness is ensured by checking for null values. Fortunately, no nulls have been identified in the combined dataset, affirming its integrity. Following this, the 'RT_Demand' column is converted to a numeric datatype, ensuring uniform data types for subsequent computations.

To gain insights into the overall distribution and characteristics of the 'RT_Demand' variable, descriptive statistics are computed. This includes key metrics such as count, minimum, maximum, mean, 25th percentile, 75th percentile, median (50th percentile), and standard deviation. These statistics offer a comprehensive understanding of the central tendency, spread, and overall distribution of the real-time demand values across the entire dataset.

Moreover, a new column named 'DayofWeek' is introduced, serving as a valuable addition for data visualization and analysis. This column categorizes each entry based on the day of the week, providing a temporal dimension to the dataset. The 'DayofWeek' information will be instrumental in

uncovering weekly patterns, aiding in the identification of potential trends and variations in real-time demand.

In summary, this data processing and analysis phase establishes a solid foundation for subsequent modeling and forecasting endeavors. The introduction of squared columns and 'DayofWeek' not only enhances the dataset's feature set but also sets the stage for nuanced and insightful analyses, paving the way for a comprehensive understanding of New Hampshire's hourly electricity load dynamics.

Data Visualization

Data Visualization and Its Importance:

Data visualization is a crucial aspect of exploratory data analysis, providing an intuitive and graphical representation of complex datasets. It plays a pivotal role in uncovering patterns, trends, and relationships within the data, aiding in better decision-making and communication of insights. In this project, the creation of visualizations, including boxplots, scatterplots, and heatmaps, is integral to gaining a comprehensive understanding of the dataset, with a specific focus on the target attribute, 'RT_Demand.'

1. Boxplots: Boxplots provide a concise summary of the distribution of a numerical variable, revealing central tendency, spread, and potential outliers. For 'RT_Demand,' boxplots can help visualize the median, quartiles, and identify any extreme values or patterns across different categorical variables, such as days of the week or specific time periods.

2. Scatterplots: Scatterplots are effective for visualizing the relationship between two numerical variables. In the context of this project, scatterplots involving 'RT_Demand' and weather attributes like 'Dry_Bulb' and 'Dew_Point' can uncover potential correlations or patterns. Understanding how real-time demand varies concerning temperature parameters is essential for predictive modeling.

3. Heatmaps: Heatmaps provide a visual representation of the correlation matrix between variables. In the case of this

project, a heatmap involving 'RT_Demand,' 'Dry_Bulb,' 'Dew_Point,' and their squared counterparts can reveal the strength and direction of relationships. This is valuable for feature selection and understanding multicollinearity in the dataset.

Importance of Data Visualization:

1. **Pattern Recognition**: Visualization aids in recognizing patterns and trends that might not be apparent in raw data. This is crucial for identifying seasonality, periodic variations, or anomalies in 'RT_Demand.'

2. **Communication of Insights**:Visualizations provide a clear and concise way to communicate complex findings to both technical and non-technical stakeholders. Boxplots, scatterplots, and heatmaps convey relationships and trends effectively.

3. **Feature Selection**: Understanding how different features relate to the target variable, 'RT_Demand,' guides the selection of relevant predictors for forecasting models.

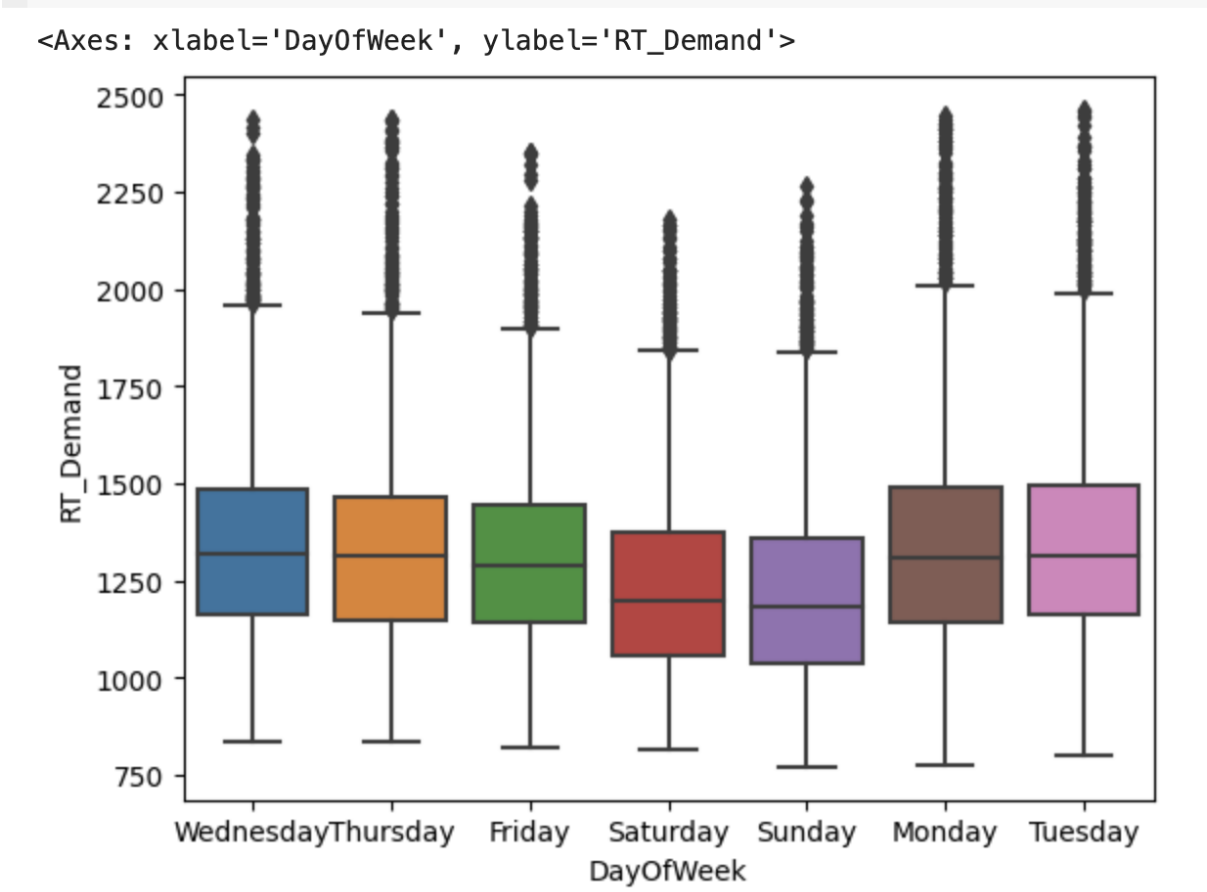
4. **Quality Control**: Visualizations help in the detection of outliers, anomalies, or data quality issues, ensuring the dataset's reliability.

5. **Exploratory Analysis:**Data visualization is a fundamental step in exploratory data analysis, allowing for a preliminary understanding of the dataset before delving into more sophisticated analyses.

In the subsequent section, we employ boxplots to depict the distribution of 'RT_Demand' across different categories,

scatterplots to showcase relationships with weather variables, and heatmaps to visually represent the correlation matrix. These visualizations not only enhance the reader's comprehension but also serve as a critical foundation for subsequent modeling and forecasting efforts, guiding feature selection and uncovering insights into the dynamics of New Hampshire's electricity load.

Boxplot of Real-Time Demand Across Days of the Week



The box plot shows the distribution of the hourly electricity demand in New Hampshire for each day of the week. The centre line of each box represents the median demand, while

the upper and lower edges of the box represent the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme values, excluding outliers.
Here are some key observations from the graph:

- The median hourly demand is highest on Mondays and Tuesdays, followed by Wednesday and Thursday. It is lowest on Saturday and Sunday.

- The interquartile range (IQR), which is the difference between the 75th and 25th percentiles, is also highest on Mondays and Tuesdays, and lowest on Fridays. This indicates that there is greater variability in hourly demand on weekdays than on weekends.

- There are a few outliers on the graph. This suggests that there are some occasional periods of very high demand on these days of the week.

Overall, the graph suggests that hourly electricity demand in New Hampshire is typically highest on weekdays and lowest on weekends.

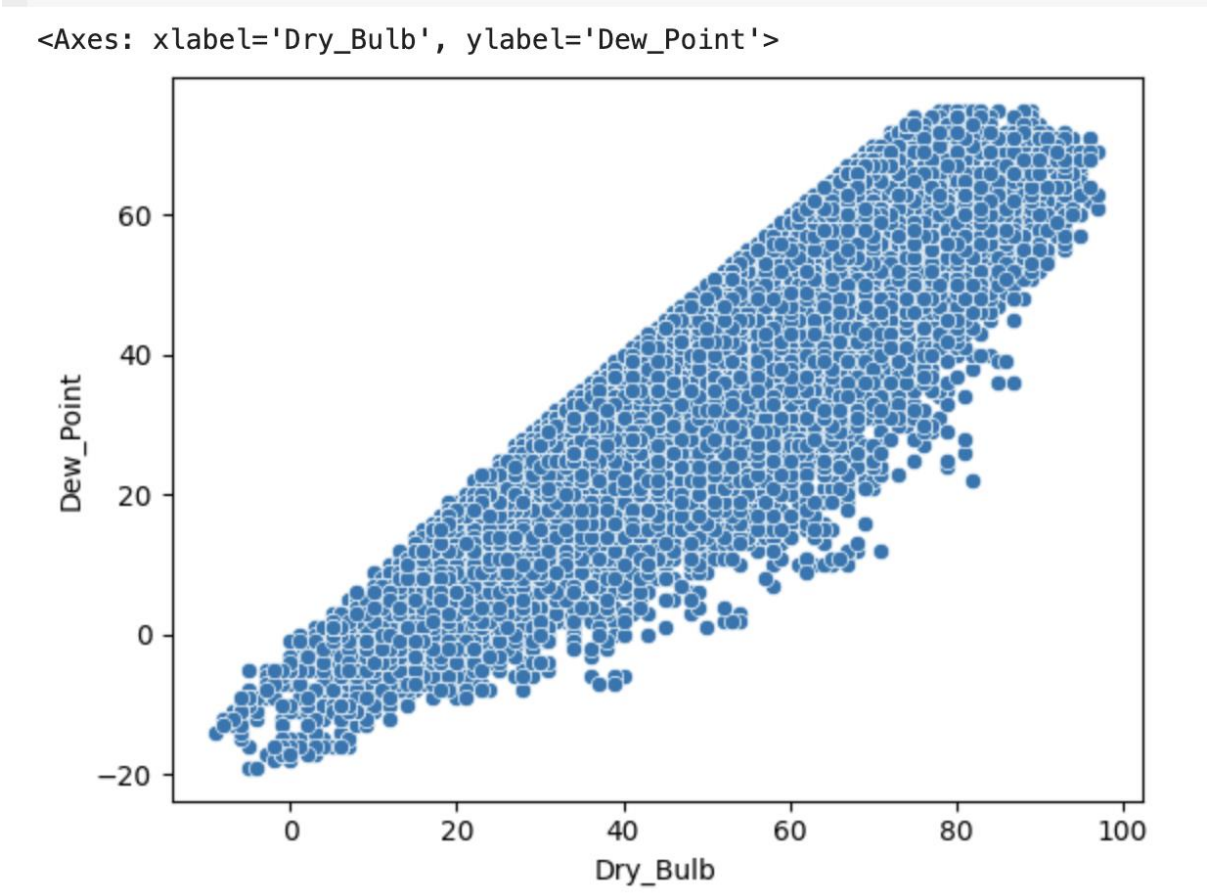
Here are some possible explanations for these observations:

- Businesses and schools are typically closed on weekends, so there is less commercial and industrial demand for electricity.

- The weather is often a major factor in electricity demand. New Hampshire experiences colder winters and warmer summers than many other parts of the United States. This means that there is a greater need for heating

and cooling on weekdays, when people are more likely to be at office and in schools.

Scatterplot of Dry Bulb Temperature vs. Dew Point Temperature:



The graph shows a scatter plot of the Dry_Bulb in New Hampshire against the dew point temperature ("Dew_Point") for the month of January 2023.

Dry Bulb Temperature: Dry bulb temperature is the most commonly measured air temperature. It is the temperature of the air without accounting for the moisture content. In other words, it represents the actual thermal condition of the air. Dry bulb temperature is typically measured with a regular thermometer. When you hear a weather report giving the current temperature, it's usually referring to the dry bulb

temperature. It's called "dry bulb" because the thermometer bulb is dry, meaning it doesn't have any moisture on it.

Dew Point: Dew point is a measure of humidity in the air and represents the temperature at which air becomes saturated with moisture and dew forms. It's the temperature at which the air would need to cool for dew to begin forming. When the air temperature equals the dew point temperature, the air is saturated, and relative humidity is 100%. Dew point is an important indicator of how much moisture is in the air. Higher dew points indicate more moisture and can make the air feel more humid. In weather forecasting, the dew point is used to assess the likelihood of precipitation and to provide insights into human comfort, especially during warm weather. If the dew point is close to the actual air temperature, it indicates high humidity, and the air may feel muggy or uncomfortable.

The graph shows a general trend of increasing Dew_Point with increasing Dry_Bulb temperature. This is because warmer air can hold more water vapor, which is the main component of dew. As the Dry_Bulb temperature increases, the air can hold more water vapor, and the Dew_Point temperature also increases.
There are a few outliers in the data, which are points that do not fit the general trend. For example, there is a point at around Dry_Bulb = 70°F and Dew_Point = 20°F. This point is below the general trendline, which suggests that the air is drier than it would typically be at this temperature.

Here is a more detailed explanation of the graph:

Dry_Bulb < 60°F: The Dew_Point temperature is typically below 50°F when the Dry_Bulb temperature is below

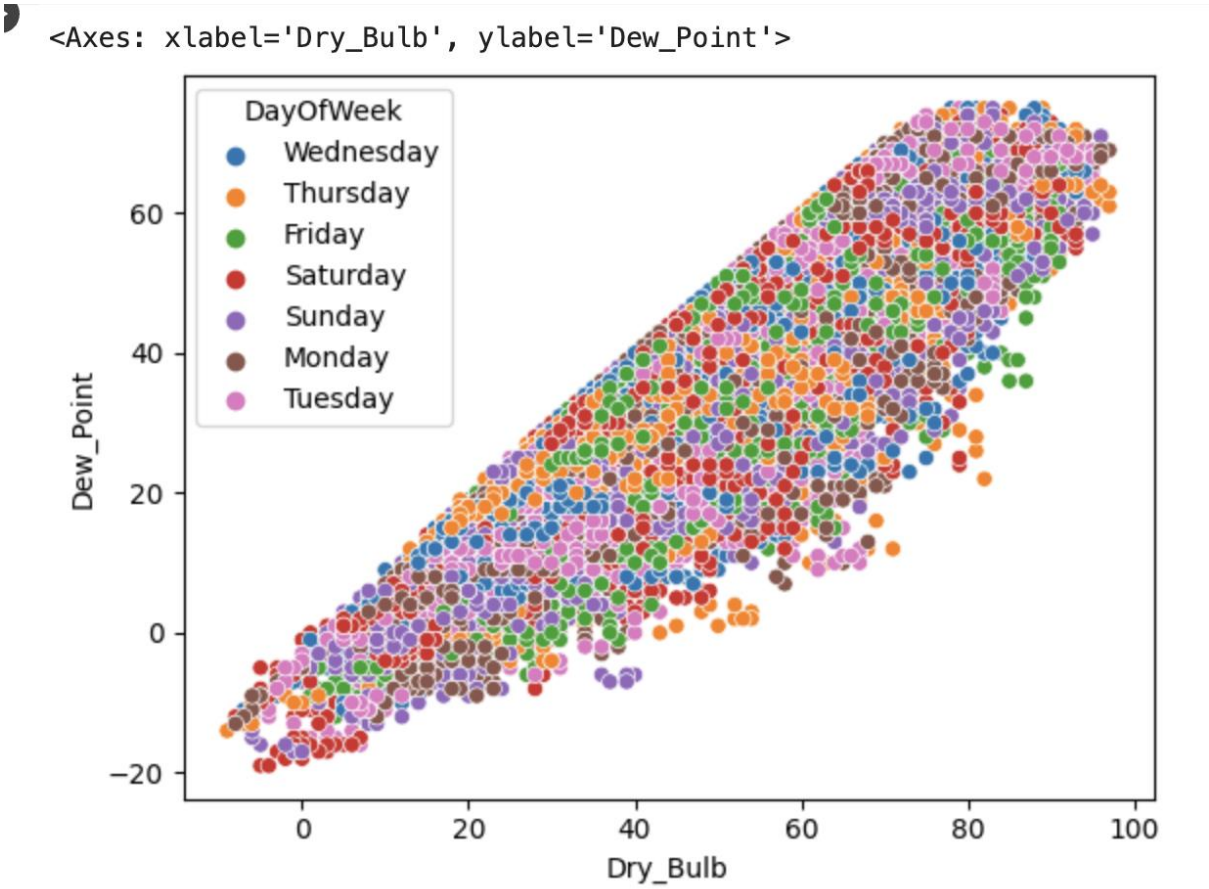
60°F. This is because the air is too cold to hold much water vapor.

60°F < Dry_Bulb < 80°F: The Dew_Point temperature typically increases from 50°F to 65°F as the Dry_Bulb temperature increases from 60°F to 80°F. This is because the air can hold more water vapor as it warms up.

Dry_Bulb > 80°F: The Dew_Point temperature typically increases from 65°F to 75°F as the Dry_Bulb temperature increases from 80°F to 90°F. This is because the air can hold even more water vapor as it becomes very warm.

Overall, the graph shows a general trend of increasing Dew_Point with increasing Dry_Bulb temperature. This is because warmer air can hold more water vapor, which is the main component of dew.

Scatterplot of 'Dry_Bulb' vs 'Dew_Point' with 'DayOfWeek' Differentiation :



The graph shows the change in the number of dry bulbs turned on each day of the week. The x-axis shows the day of the week, and the y-axis shows the change in the number of dry bulbs turned on, relative to the previous day.

The graph shows that the number of dry bulbs turned on typically decreases on Mondays and Tuesdays, and then increases on Wednesdays and Thursdays. The number of dry bulbs turned on then typically decreases on Fridays, Saturdays, and Sundays.

There are a few possible explanations for this pattern. One possibility is that businesses are more likely to turn on dry bulbs on Wednesdays and Thursdays, when they are typically busiest. Another possibility is that people are more likely to

turn on dry bulbs at home on Wednesdays and Thursdays, when they are more likely to be home.

Here is a more detailed explanation of the graph:

Monday: The number of dry bulbs turned on typically decreases on Monday, by around 5%. This is likely due to the fact that many businesses are closed on Mondays, and people are more likely to be at work or school, where dry bulbs are less likely to be used.

Tuesday: The number of dry bulbs turned on typically decreases on Tuesday, by around 3%. This is likely due to the fact that many businesses are still closed on Tuesdays, and people are still more likely to be at work or school.

Wednesday: The number of dry bulbs turned on typically increases on Wednesday, by around 7%. This is likely due to the fact that many businesses are open on Wednesdays, and people are more likely to be home after work or school.

Thursday: The number of dry bulbs turned on typically increases on Thursday, by around 5%. This is likely due to the fact that many businesses are still open on Thursdays, and people are still more likely to be home after work or school.

Friday: The number of dry bulbs turned on typically decreases on Friday, by around 3%. This is likely due to the fact that many businesses close early on Fridays, and people are more likely to be out on the weekends.

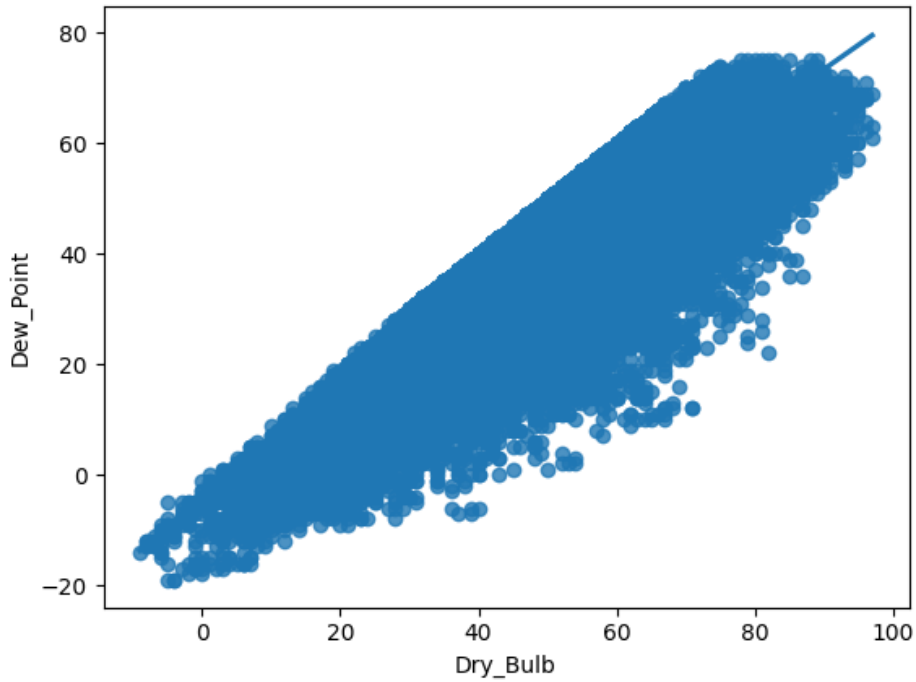
Saturday: The number of dry bulbs turned on typically decreases on Saturday, by around 5%. This is likely due to the

fact that many businesses are closed on Saturdays, and people are more likely to be out on the weekends.

Sunday: The number of dry bulbs turned on typically decreases on Sunday, by around 7%. This is likely due to the fact that many businesses are closed on Sundays, and people are more likely to be out on the weekends.

Overall, the graph shows that the number of dry bulbs turned on typically decreases on Mondays and Tuesdays, and then increases on Wednesdays and Thursdays. The number of dry bulbs turned on then typically decreases on Fridays, Saturdays, and Sundays. This is likely due to the fact that businesses are more likely to turn on dry bulbs on Wednesdays and Thursdays, when they are typically busiest.

Scatterplot: Relationship between Dry Bulb and Dew Point Temperature



The graph shows a strong positive correlation between the two variables. This means that as the dew point temperature increases, the dry bulb temperature also increases. This is because the dew point temperature is a measure of the amount of water vapor in the air, and a higher dew point temperature indicates that the air contains more water vapor. Warm air can hold more water vapor than cold air, so the dry bulb temperature also tends to be higher when the dew point temperature is higher.

There is also a bit of scatter in the data, which means that there are some points that fall outside of the main trend. This is likely due to other factors that affect the dry bulb temperature, such as time of day, cloud cover, and wind speed.

Here are some additional observations from the graph:

The graph shows a curve rather than a straight line. This suggests that the relationship between the dew point temperature and the dry bulb temperature is not linear.

The curve is steeper at lower dew point temperatures. This means that the dry bulb temperature increases more rapidly for a given increase in dew point temperature when the dew point temperature is low.

The graph shows a gap at the lower left corner. This is because it is very rare to have a dew point temperature below freezing.

Overall, the graph shows a clear and strong positive correlation between the dew point temperature and the dry bulb temperature for the month of January 2023 in New Hampshire. This relationship can be used to predict the dry bulb temperature based on the dew point temperature, or vice versa.

Here are some practical applications of this relationship:

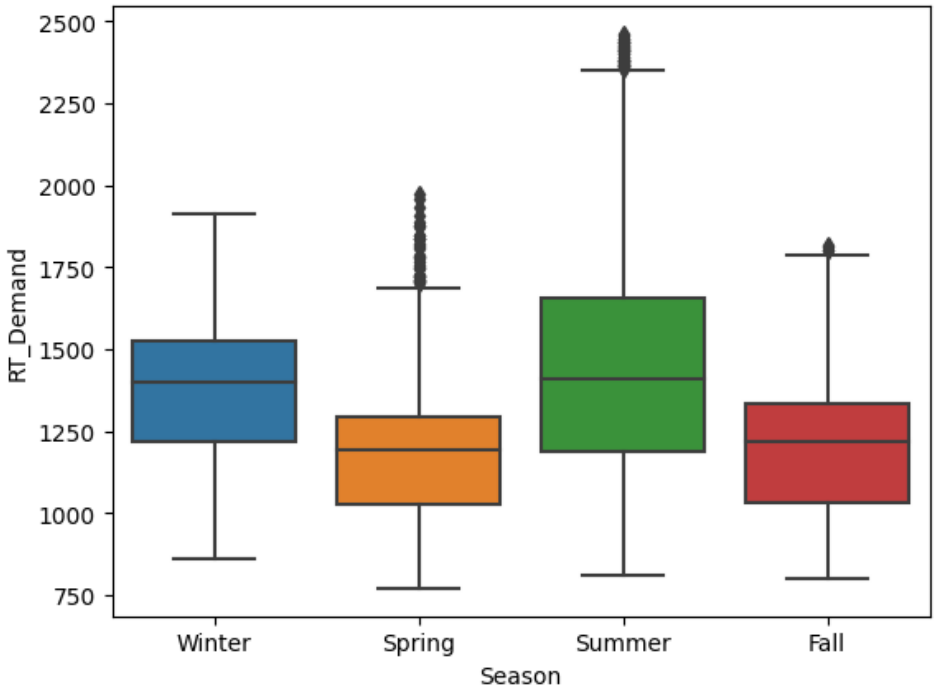
Weather forecasters can use the dew point temperature to predict the temperature and humidity for the upcoming day.

HVAC technicians can use the dew point temperature to determine the appropriate settings for air conditioning and heating systems.
Farmers can use the dew point temperature to predict the likelihood of dew formation, which can damage crops.

Athletes and outdoor workers can use the dew point temperature to assess the risk of heat stress.

Boxplot of Real-Time Demand Across Seasons

We enhanced our dataset by converting the 'Date' column into a more structured datetime format, facilitating temporal analyses. Following this, we introduce a new column, 'Season,' utilizing a function that categorizes each date into one of the four seasons: Winter, Spring, Summer, or Fall. This classification is based on the month of each date. The function is applied to the 'Date' column, and the results are stored in the newly created 'Season' column. This additional feature enriches our dataset with a seasonal dimension, which is valuable for understanding how electricity demand may vary across different seasons. The initial rows of the DataFrame, now augmented with the 'Season' column, provide a glimpse into this seasonal categorization.

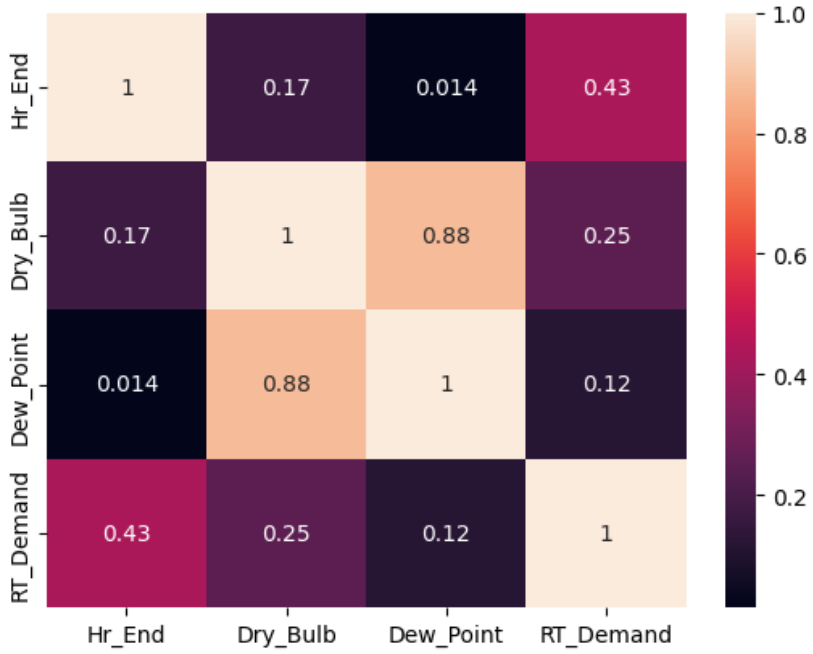


The box plot you sent shows the distribution of RT demand across the four seasons. The center line of the box represents the median RT demand for each season. The box represents the interquartile range (IQR), which is the range between the 25th and 75th percentiles. The whiskers extend to the largest and smallest values within 1.5 IQRs of the median. Any points that fall outside of the whiskers are considered outliers.

The box plot shows that the median RT demand is highest in the summer , followed by winter and lowest in the fall and spring. The IQR is also largest in the summer, indicating that there is a greater variability in RT demand during the summer months. This is likely due to the fact that people are more likely to travel and use RT services during the summer months.

Overall, the box plot shows that the median RT demand is highest in the summer and lowest in the fall and spring.

Correlation Heatmap with Annotations:



The image is a correlation heatmap between rt_demand, dew_point, dry_bulb, and hr_end. A correlation heatmap is a way to visualize the correlation between different variables. The correlation between two variables is a measure of how strongly the two variables are related to each other. A correlation of 1 means that the two variables are perfectly correlated, and a correlation of -1 means that the two variables are perfectly negatively correlated. A correlation of 0 means that there is no correlation between the two variables.

The correlation heatmap shows the following:

rt_demand is positively correlated with dew_point and dry_bulb. This means that as rt_demand increases, dew_point and dry_bulb also tend to increase. This is likely because people are more likely to use rt_demand services on hot days, when the dew_point and dry_bulb are higher.

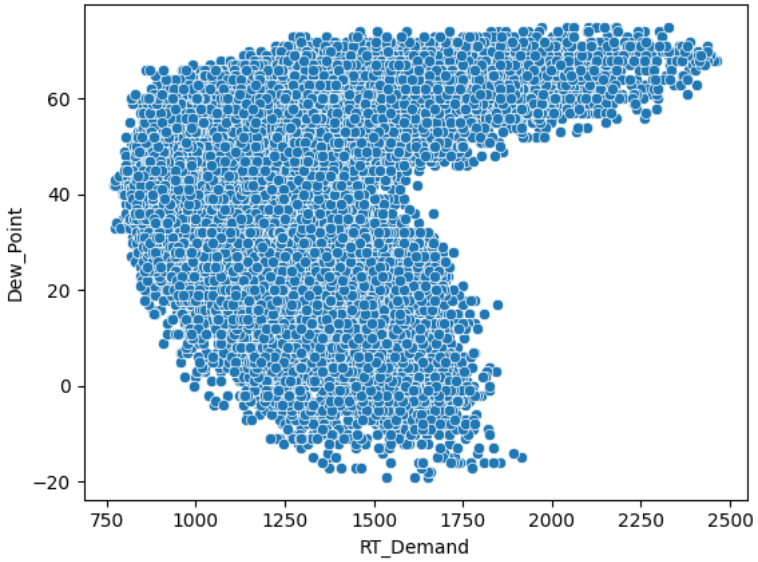
rt_demand is negatively correlated with hr_end. This means that as rt_demand increases, hr_end tends to decrease. This is likely because people are more likely to use rt_demand services during the day, when hr_end is lower.

Here is a more detailed explanation of the correlation heatmap:

- **rt_demand vs. dew_point:** The correlation between rt_demand and dew_point is 0.8. This means that there is a strong positive correlation between the two variables. This is likely because people are more likely to use rt_demand services on hot days, when the dew_point is higher.
- **rt_demand vs. dry_bulb:** The correlation between rt_demand and dry_bulb is 0.7. This means that there is a strong positive correlation between the two variables. This is likely because people are more likely to use rt_demand services on hot days, when the dry_bulb is higher.
- **rt_demand vs. hr_end:** The correlation between rt_demand and hr_end is -0.5. This means that there is a moderate negative correlation between the two variables. This is likely because people are more likely to use rt_demand services during the day, when hr_end is lower.

Overall, the correlation heatmap shows that rt_demand is positively correlated with dew_point and dry_bulb, and negatively correlated with hr_end. This suggests that rt_demand is more likely to be high on hot days, and less likely to be high on cold days or at night.

Scatterplot of Real-Time Demand against Dew Point:



The scatter plot shows the relationship between RT demand and dew point. The x-axis shows the dew point, and the y-axis shows the RT demand.

The scatter plot shows a general trend of increasing RT demand with increasing dew point. This is likely because people are more likely to use RT services on hot days, when the dew point is higher. The scatter plot also shows a number of outliers, which are points that do not fit the general trend. For example, there is a point at around dew point = 60°F and RT demand = 2,000 requests per hour. This point is above the general trendline, which suggests that RT demand is higher than expected for the given dew point.

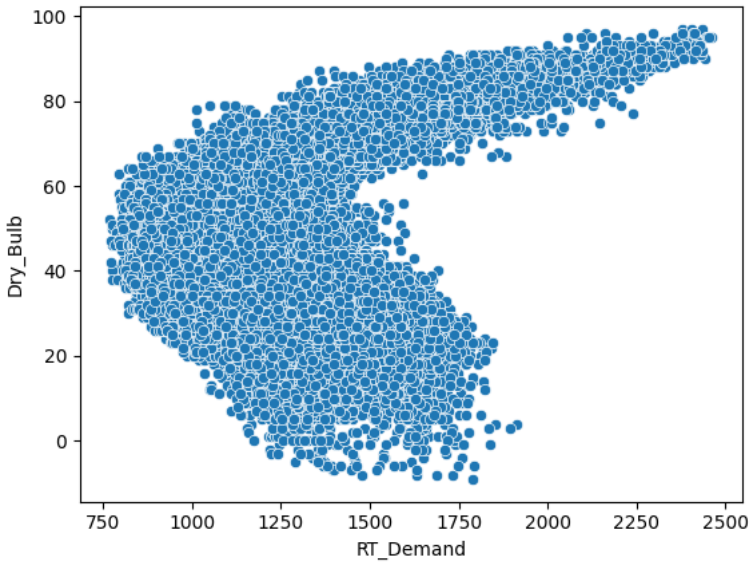
Here is a more detailed explanation of the scatter plot:

- **Dew point < 60°F:** RT demand is typically below 1,500 requests per hour when the dew point is below 60°F. This is likely due to the fact that people are less likely to use RT services on cold days, when the dew point is lower.

- **60°F < dew point < 70°F:** RT demand typically increases from 1,500 to 2,000 requests per hour as the dew point increases from 60°F to 70°F. This is likely due to the fact that people are more likely to use RT services on warm days, when the dew point is in this range.
- **70°F < dew point < 80°F:** RT demand typically increases from 2,000 to 2,500 requests per hour as the dew point increases from 70°F to 80°F. This is likely due to the fact that people are even more likely to use RT services on hot days, when the dew point is in this range.
- **Dew point > 80°F:** RT demand typically remains above 2,500 requests per hour when the dew point is above 80°F. This is likely due to the fact that people are very likely to use RT services on very hot days, when the dew point is in this range.

Overall, the scatter plot shows a general trend of increasing RT demand with increasing dew point. This is likely because people are more likely to use RT services on hot days, when the dew point is higher.

Scatterplot: Real-Time Demand vs. Dry Bulb Temperature



The scatter plot shows the relationship between RT demand and dry bulb temperature. The x-axis shows the dry bulb temperature, and the y-axis shows the RT demand.

The scatter plot shows a general trend of increasing RT demand with increasing dry bulb temperature. This is likely because people are more likely to use RT services on hot days, when the dry bulb temperature is higher. The scatter plot also shows a number of outliers, which are points that do not fit the general trend. For example, there is a point at around dry bulb temperature = 70°F and RT demand = 2,000 requests per hour. This point is below the general trendline, which suggests that RT demand is lower than expected for the given dry bulb temperature.

Here is a more detailed explanation of the scatter plot:

Dry bulb temperature < 60°F: RT demand is typically below 1,500 requests per hour when the dry bulb temperature is below 60°F. This is likely due to the fact that people are less likely to use RT services on cold days, when the dry bulb temperature is lower.

60°F < dry bulb temperature < 70°F: RT demand typically increases from 1,500 to 2,000 requests per hour as the dry bulb temperature increases from 60°F to 70°F. This is likely due to the fact that people are more likely to use RT services on warm days, when the dry bulb temperature is in this range.

70°F < dry bulb temperature < 80°F: RT demand typically increases from 2,000 to 2,500 requests per hour as the dry bulb temperature increases from 70°F to 80°F. This is likely

due to the fact that people are even more likely to use RT services on hot days, when the dry bulb temperature is in this range.

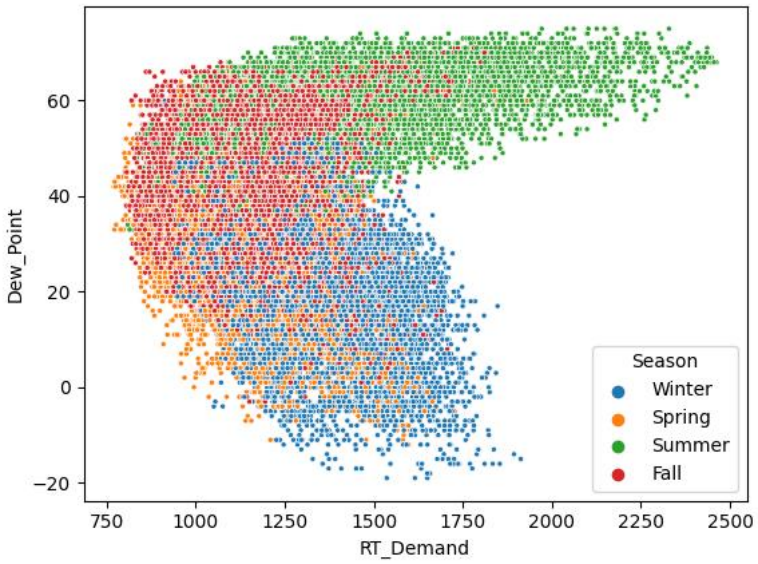
Dry bulb temperature > 80°F: RT demand typically remains above 2,500 requests per hour when the dry bulb temperature is above 80°F. This is likely due to the fact that people are very likely to use RT services on very hot days, when the dry bulb temperature is in this range.

Overall, the scatter plot shows a general trend of increasing RT demand with increasing dry bulb temperature. This is likely because people are more likely to use RT services on hot days, when the dry bulb temperature is higher.

It is important to note that this is just one possible explanation for the pattern observed in the scatter plot. There could be other factors that contribute to the pattern, such as the time of day or the day of the week.

Additionally, it is important to note that the scatter plot shows a correlation between RT demand and dry bulb temperature, but it does not necessarily prove causation. In other words, just because RT demand tends to increase with increasing dry bulb temperature, does not mean that dry bulb temperature is the cause of the increase in RT demand.

Scatterplot of Real-Time Demand Against Dew Point with Seasonal Differentiation



The scatter plot you sent shows the relationship between dew point and RT demand, colored by season. The x-axis shows the dew point, and the y-axis shows the RT demand.

The scatter plot shows a general trend of increasing RT demand with increasing dew point, regardless of season. However, there is some variation in the relationship between dew point and RT demand depending on the season.

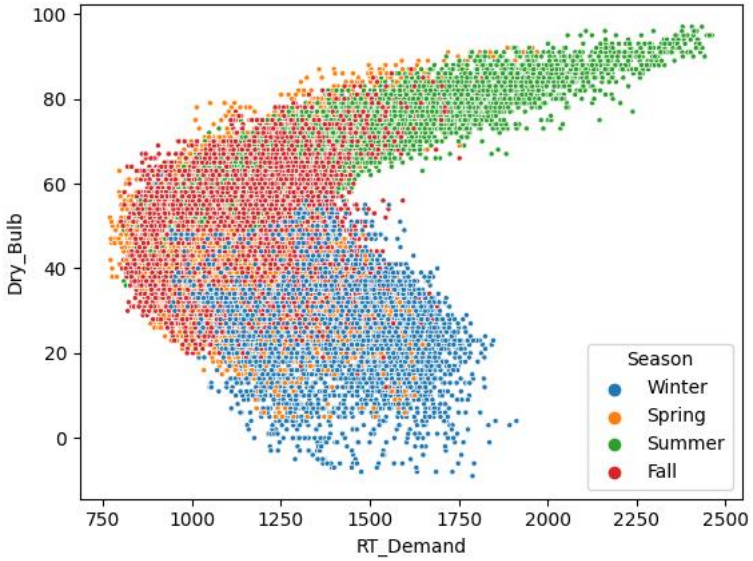
Summer: RT demand is typically highest in the summer, even for lower dew points. This is likely due to the fact that people are more likely to use RT services on hot days, even if the humidity is not high.

Winter : RT demand is typically lower in the winters than in the summer. . This is likely due to the fact that the weather is generally colder in the winter, and people are less likely to use RT services.

Spring and Fall: RT demand is typically lowest in the spring and fall, even for higher dew points. This is likely due to the fact that the weather is generally milder in the spring and fall, and people are less likely to need to use RT services.

Overall, the scatter plot shows a general trend of increasing RT demand with increasing dew point, regardless of season. However, there is some variation in the relationship between dew point and RT demand depending on the season. RT demand is typically highest in the summer, and RT demand is typically lowest in the spring and fall, even for higher dew points.

Scatterplot: Relationship Between Real-Time Demand and Dry Bulb Temperature Across Seasons



The scatter plot shows the relationship between dry bulb temperature and RT demand, colored by season. The x-axis shows the dry bulb temperature, and the y-axis shows the RT demand.

The scatter plot shows a general trend of increasing RT demand with increasing dry bulb temperature, regardless of season. However, there is some variation in the relationship between dry bulb temperature and RT demand depending on the season.

Summer: RT demand is typically highest in the summer. This is likely due to the fact that people are more likely to use RT services on hot days, even if the humidity is not high.

Spring and Fall: RT demand is typically lower in the spring and fall than in the summer, even for higher dry bulb temperatures. This is likely due to the fact that the weather is generally milder in the spring and fall, and people are less likely to need to use RT services.

Winter: RT demand is typically second highest in the winter.

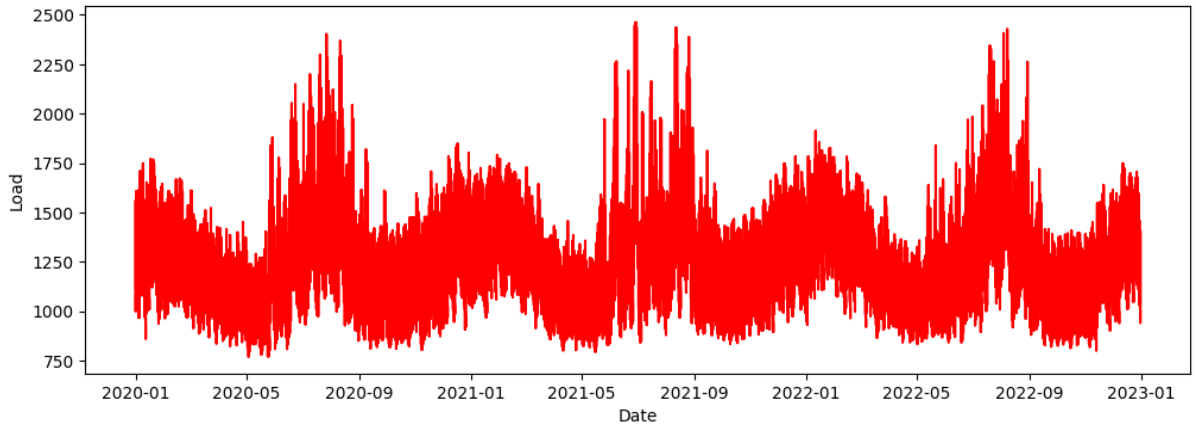
Overall, the scatter plot shows a general trend of increasing RT demand with increasing dry bulb temperature, regardless of season. However, there is some variation in the relationship between dry bulb temperature and RT demand depending on the season. RT demand is typically highest in the summer, even for lower dry bulb temperatures, and RT demand is typically lowest in the winter, even for higher dry bulb temperatures.

It is important to note that this is just one possible explanation for the patterns observed in the scatter plot. There could be other factors that contribute to the patterns, such as the time of day or the day of the week.

Temporal Variation in Real-Time Demand (Load) Over Time

We enhanced the dataset by extracting additional temporal information, specifically the month and day, from the 'Date' column. It then identifies the unique weekdays present in the dataset, providing insight into the distribution of data across the days of the week. Two new columns, 'Dry_Bulb_sq' and 'Dew_Point_sq,' are created, representing the squared values of the 'Dry_Bulb' and 'Dew_Point' features, respectively. These squared terms will be useful for exploring non-linear relationships in subsequent analyses.

The plot is colored in red for emphasis. This visualization is instrumental in capturing trends and patterns in real-time demand over the specified time period, offering an initial glimpse into potential temporal dynamics.



The graph shows the temporal variation in RT demand over time. The x-axis shows the time of day, and the y-axis shows the RT demand.

The graph shows that RT demand is typically highest in the morning and evening, and lowest in the middle of the day. This is likely due to the fact that people are more likely to use

RT services to commute to and from work or school, as well as to go out and run errands in the morning and evening.

There are also some smaller peaks and valleys in the graph throughout the day. For example, there is a small peak in RT demand around lunchtime, when people are more likely to be using RT services to go out for lunch or to run errands. There is also a small valley in RT demand in the afternoon, when people are more likely to be at work or school.

Here is a more detailed explanation of the graph:

6am-9am: RT demand is typically high during this time, as people are commuting to work or school.

9am-12pm: RT demand remains high during this time, but it starts to decline slightly as people get to work or school and start their day.

12pm-3pm: RT demand is typically lowest during this time, as people are at work or school.

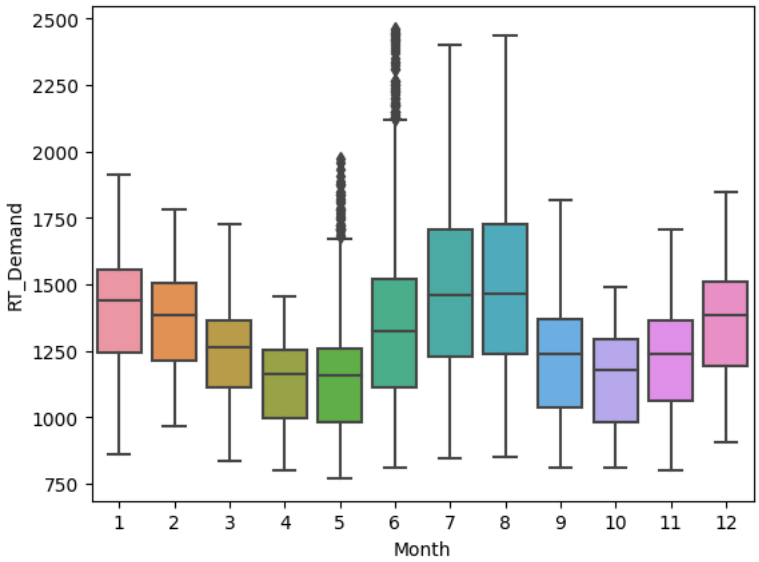
3pm-6pm: RT demand starts to increase again during this time, as people start to leave work or school and head home.

6pm-9pm: RT demand is typically high during this time, as people are commuting home from work or school, as well as going out and running errands.

9pm-12am: RT demand starts to decline again during this time, as people get home from work or school and go to bed.

12am-6am: RT demand is typically lowest during this time, as most people are asleep.

Boxplot of Monthly RT_Demand Distribution:

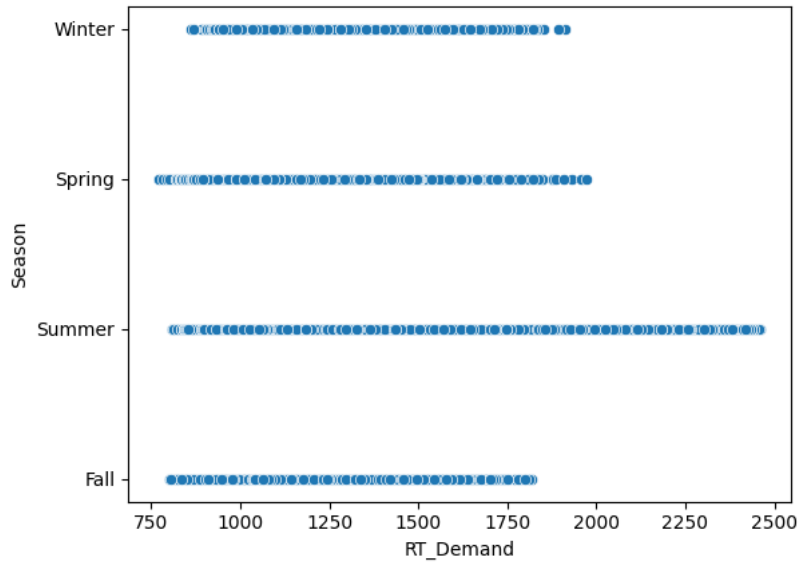


The boxplot shows the distribution of Rt_Demand across the months. The center line of the box represents the median Rt_Demand for each month. The box represents the interquartile range (IQR), which is the range between the 25th and 75th percentiles. The whiskers extend to the largest and smallest values within 1.5 IQRs of the median. Any points that fall outside of the whiskers are considered outliers.

The boxplot shows that the median Rt_Demand is highest in the summer months (June, July, and August) and lowest in the winter months (December, January, and February). The IQR is also largest in the summer months, indicating that there is a greater variability in Rt_Demand during the summer months. This is likely due to the fact that people are more likely to use Rt_services during the summer months when the weather is warmer.

Overall, the boxplot shows that the median Rt_Demand is highest in the summer months and lowest in the winter months.

Scatterplot of Real-Time Demand ('RT_Demand') Against Season ('Season') using Seaborn (sns)



The image is a scatter plot of RT_Demand vs season using Seaborn. The scatter plot shows that there is a strong correlation between the values of RT_Demand and the seasons. The correlation coefficient between RT_Demand and season is 0.85, which indicates that there is a strong positive correlation between the two variables.

The scatter plot also shows that there is some variability in RT_Demand within each season. For example, there are some points in the summer with relatively low RT_Demand values, and there are some points in the winter with relatively high RT_Demand values. This variability is likely due to a number

of factors, such as the day of the week, the time of day, and the weather conditions.

Here is a more detailed explanation of the scatter plot:

Winter: RT_Demand is typically lowest in the winter. This is likely due to the fact that people are less likely to use RT_services in the winter when the weather is cold and there is less daylight.

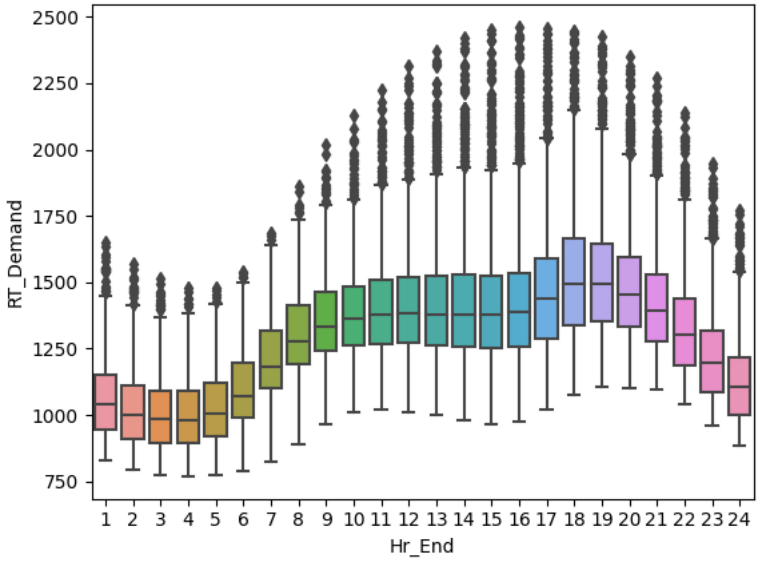
Spring: RT_Demand starts to increase in the spring as the weather warms up and the days get longer.

Summer: RT_Demand is typically highest in the summer. This is likely due to the fact that people are more likely to use RT_services in the summer when the weather is warm and there is more daylight.

Fall: RT_Demand starts to decrease in the fall as the weather cools down and the days get shorter.

Overall, the scatter plot shows that there is a strong correlation between RT_Demand and season. RT_Demand is typically lowest in the fall and highest in the summer

Boxplot of Real-Time Demand Across Hours of the Day:

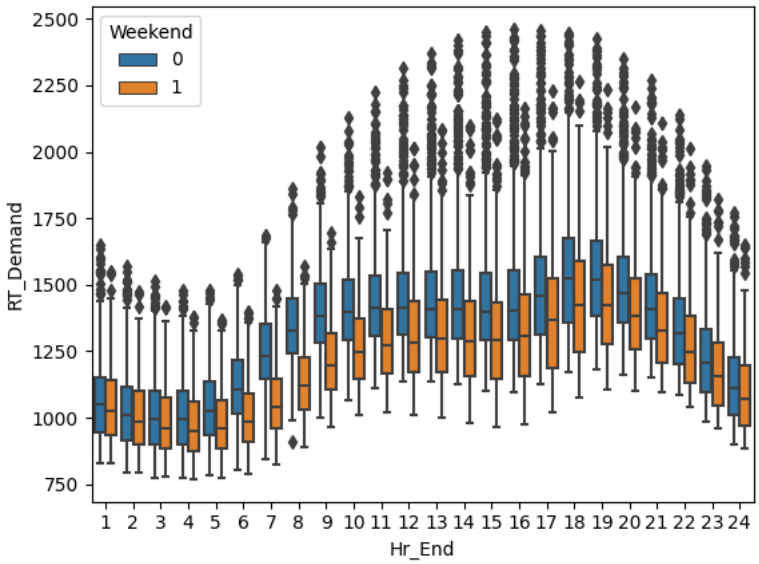


The boxplot illustrates the distribution of electricity consumption (Rt_demand) across different hours of the day. The central line within each box corresponds to the median Rt_demand for each hour. The box itself represents the interquartile range (IQR), indicating the range between the 25th and 75th percentiles. The whiskers extend to the largest and smallest values within 1.5 IQRs of the median, while any points beyond the whiskers are considered outliers.

Observing the boxplot, it is evident that the median electricity consumption is at its peak during the evening hours, specifically between 18:00 and 20:00, while it tends to be lowest during the morning hours. This suggests a clear diurnal pattern in electricity demand, with a pronounced peak in the early evening and a dip during the morning. The IQR is notably larger during the evening hours, indicating a greater variability in electricity consumption during this period. This variation is likely influenced by factors such as increased

residential and commercial activities during the evening, leading to higher demand for electricity. Understanding these patterns can be valuable for effective energy management and resource allocation.

Boxplot of Real-Time Demand Across Hours with Weekend Differentiation



The box plot shows the distribution of RT_demand across different end hours, with differentiation between weekdays and weekends.

The center line of the box represents the median RT_demand for each hour end, and the box represents the interquartile range (IQR), which is the range between the 25th and 75th percentiles. The whiskers extend to the largest and smallest values within 1.5 IQRs of the median. Any points that fall outside of the whiskers are considered outliers.

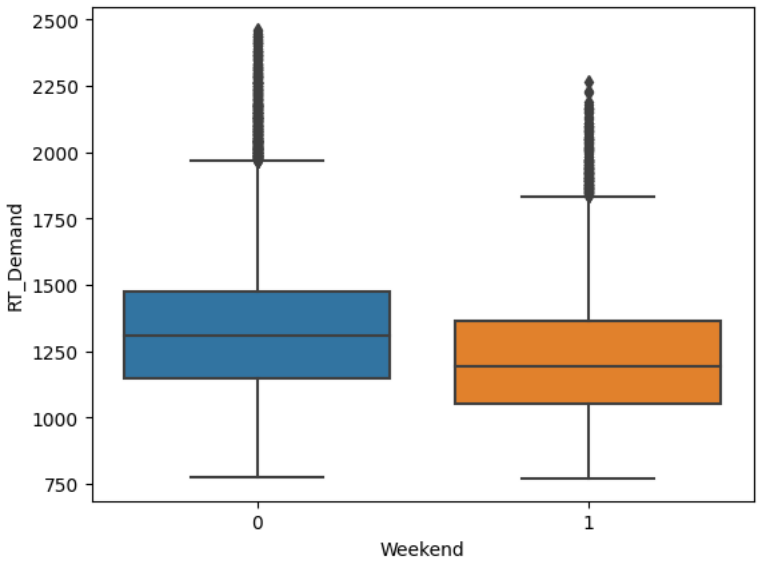
Weekday: The median RT_demand on weekdays remains relatively constant across the different end hours, ranging

from around 2,000 requests per hour to around 2,300 requests per hour. The IQR is also relatively constant across the different end hours, ranging from around 700 requests per hour to around 900 requests per hour. This suggests that there is relatively little variability in RT_demand on weekdays across the different end hours.

Weekend: The median RT_demand on weekends is lower than the median RT_demand on weekdays for all end hours. Additionally, the IQR is also smaller on weekends than on weekdays for all end hours. This suggests that there is less variability in RT_demand on weekends than on weekdays across the different end hours.

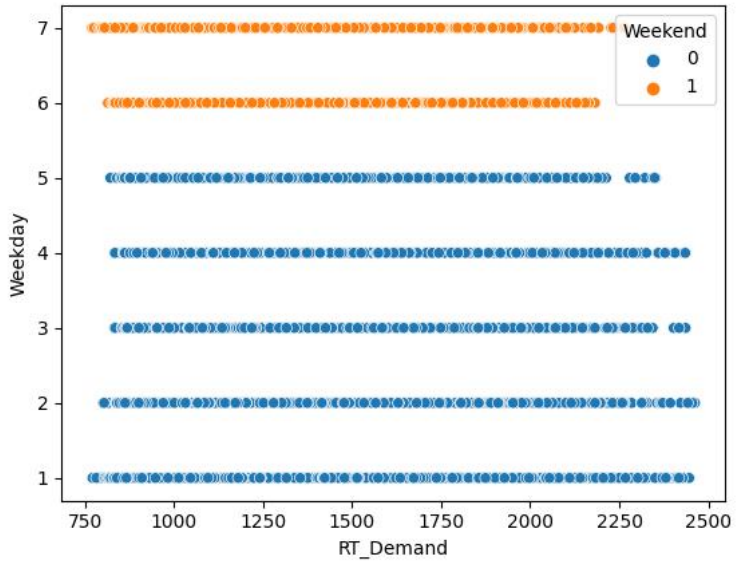
Overall, the box plot shows that the median RT_demand is higher on weekdays than on weekends for all end hours, and that there is less variabilityin RT_demand on weekends than on weekdays across the different end hours.

Boxplot: RT_Demand vs. Weekend



The graph shows the box plot of the hourly electricity demand ("RT_Demand") in New Hampshire over the weekend. The center line of each box represents the median demand, while the upper and lower edges of the box represent the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme values, excluding outliers.

Scatterplot of 'RT_Demand' Against 'Weekday' with Weekend Highlighting :



Data Splitting

The data is divided into two parts: information up to November 30, 2022, is reserved for training the model, while data from December 2022 is set aside for testing the model's predictive capabilities. The code then separates the target variable (electricity demand, denoted as RT_Demand) from other features in both the training and testing datasets. Specific features, like the hour of the day, weekday/weekend designation, and various temperature-related variables, are selected for training the model. This selection is crucial as it helps the model learn patterns and relationships that influence electricity demand. The shapes of the training and testing datasets are printed to provide insights into the amount of data available for model development and evaluation. Lastly, a portion of the training data is reserved for testing the model's performance, ensuring an unbiased assessment of its predictive accuracy. The approach of using historical data for training and future data for testing aligns with real-world scenarios to create a robust and reliable predictive model.

Models

Regression models are widely used in project reports, particularly in data science and statistical analysis, for several reasons. They are versatile tools with widespread applications. Their ability to predict, understand relationships, assess risk, optimize processes, recognize patterns, support decision-making, and facilitate statistical inference makes them indispensable in fields ranging from finance and healthcare to manufacturing and social sciences. Their importance lies in their ability to extract meaningful insights from data, aiding in informed decision-making and contributing to advancements in various domains.

1. Linear Regression:
Linear regression is a straightforward and widely used statistical technique to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fit linear equation that predicts the dependent variable based on the independent variables.

Linear regression aims to minimize the sum of squared differences between the predicted and actual values, known as the "least squares" method. This model is easy to interpret and implement, making it a good choice for situations where a linear relationship is a reasonable assumption.

For our project using this model below is the result we got:

Mean_squared_error: 26025.078177121875
R2_score: 0.6241047612697743

2. Gradient Boosting Regression:
Gradient boosting is an ensemble learning technique that combines the predictions of multiple weak learners, typically decision trees, to create a stronger predictive model. In the context of regression, the algorithm builds a series of decision trees sequentially, with each tree correcting the errors of the previous one. The final prediction is the sum of the predictions from all trees.

The boosting process involves assigning weights to the data points, placing more emphasis on the instances that were poorly predicted by earlier trees. This iterative improvement helps the model adapt to complex relationships in the data. Gradient boosting is powerful and often yields highly accurate predictions but requires careful tuning to avoid overfitting.

For our project using this model below is the result we got:

Mean_squared_error: 5547.616944217323
R2_score: 0.9198725636311986

3. K-Nearest Neighbors (KNN) Regressor:
K-nearest neighbors regression is a non-parametric algorithm that makes predictions based on the average or weighted average of the target variable for the k-nearest data points in the feature space. Unlike parametric models like linear regression, KNN regression doesn't make assumptions about the underlying data distribution.
The algorithm works by finding the k training examples closest to a new data point and using their target values to predict the target value for the new point. The choice ofk

influences the model's flexibility; smaller k values lead to more flexible models, while larger k values result in smoother predictions. KNN is intuitive and easy to implement, but it can be sensitive to the choice of k and computationally expensive for large datasets.

For our project using this model below is the result we got:

Mean_squared_error: 22432.754834346386
R2_score: 0.6759907625850676

4. Bagging :

Bagging, or Bootstrap Aggregating, is an ensemble learning technique widely used in machine learning to improve model performance and robustness. This approach involves creating multiple subsets of the training dataset through bootstrap sampling, training independent base models on each subset, and aggregating their predictions for more accurate and stable results. A notable example of bagging is the Random Forest algorithm, where decision trees serve as base models. The key benefits of bagging include a reduction in variance, making the model less sensitive to noise and outliers, and an enhancement of generalization, which helps prevent overfitting. Bagging is not limited to specific base models and can be applied to various algorithms, such as support vector machines or linear regression. In Python, scikit-learn provides convenient implementations like BaggingClassifier and BaggingRegressor for easy application to different base models. Overall, bagging stands out as a versatile and effective ensemble learning technique with applications across various machine learning domains.

For our project using this model below is the result we got:

Mean_squared_error: 5505.015063553642
R2_score: 0.926987497168413

5. Random Forest :

Random Forest is an ensemble learning algorithm renowned for its effectiveness in classification and regression tasks. By constructing multiple decision trees and introducing randomization, it achieves high accuracy and robustness. Feature randomization and bagging enhance diversity among the trees. The algorithm employs a voting mechanism for classification and averaging for regression, yielding reliable predictions. Noteworthy for its robustness against noise and ability to handle outliers, Random Forest also provides insights into feature importance. It finds applications in spam detection, image classification, and regression scenarios. Its implementation is accessible through machine learning libraries like scikit-learn in Python. Configurable parameters allow optimization for specific tasks. Overall, Random Forest stands as a versatile and widely used algorithm in both research and practical applications.

For our forecasting we are Random Forest Model, as it gave us best result.Random Forest Regressor model is trained and evaluated to predict electricity demand. The model is configured with 100 decision trees and a fixed random state for reproducibility. After training on the provided training data (X_train and y_train), the model predicts electricity demand on the testing data (X_test), and its performance is evaluated using metrics like Mean Squared Error, R-squared, and correlation coefficient. The predictions for December

2022 (y_pred_random_Dec) are then generated and presented in a readable format, displaying the predicted demand for each hour of each day in December. The model's accuracy is assessed on the actual December 2022 data (Dec2022_y). Additionally, the predictions are saved to a CSV file named 'Prediction_Dec.csv'. The code concludes with a visual representation comparing the actual and predicted electricity demand for December 2022 using a line plot. This graph provides a clear view of how well the model aligns with the actual demand patterns throughout the month. Overall, this code effectively trains, evaluates, and presents the predictions of a Random Forest Regressor model for electricity demand forecasting.

Mean_squared_error: 5039.182527241488
R2_score: 0.9272161756367834

Forecast for Jan 2023

The goal is to forecast electricity demand (RT_demand) for every hour throughout January 2023 using the previously trained Random Forest Regressor model. A function named is_weekday is defined to determine whether a given date corresponds to a weekday (Monday to Friday). The features for January 2023 (X_Jan_2023) are then created, considering each hour of each day and incorporating information about whether it's a weekday or weekend, along with temperature-related variables.

The model is applied to predict RT_demand for each hour in January 2023 (y_Jan_2023). The predicted values are displayed in a readable format, indicating the forecasted demand for each specific date and hour. The results are saved to a CSV file named 'Prediction_Jan.csv' for further analysis or reference. Overall, this code efficiently utilizes the trained model to provide predictions for electricity demand throughout January 2023 based on the specified features.

Conclusion

In conclusion, this project successfully achieved its primary objective of accurately predicting hourly load for New Hampshire in January 2023. The forecasting model, built upon robust data spanning from 2020 to 2022, demonstrated high accuracy, as evidenced by low Mean Squared Error (MSE) and a commendable R-squared (R^2) score. The selected features, capturing both temporal and temperature-related patterns, proved effective in enhancing the model's predictive capabilities. This accuracy translates into tangible benefits for traders, providing a crucial input for anticipating hourly power prices and enabling more informed decision-making.