**ANOVA Project**

**STAT301-50**

**Spring 2025**

**SJ Zimmermann, Mike Earl, and Ian Wallin**

**An Analysis of Utility Meter Readings**

**Introduction:**

This dataset includes the usage data of electrical meter readings collected between 7/1/2024 and 7/31/2024, during the hours of 1 PM to 10 PM. These readings come from residential households in Minnesota that use central air and participated in a load control program.

The load control program was designed to limit electrical usage over peak times by cycling the air conditioners on and off at fifteen-minute intervals. The goal was to lower overall consumption during periods when the utility company faces higher demand. The dataset was limited to households that used over 700kWh per year to help eliminate seasonal usage properties. All meters experienced load control during the same periods. The data set also includes temperature and humidity readings for each of the time frames.

We aim to find out if temperature or humidity have a significant impact on electrical usage. Furthermore, we would like to find out if there is a significant difference in usage over the load control periods compared to uncontrolled periods to evaluate if this program is beneficial in reducing load over peak times. Reducing load during peak times is important as there is a demand charge for the highest usage hour of each month.

The dataset includes the following 10 variables:

* Timeframe – This string value is the date and start and end time of the reading
* StartTime – This POSIXt value is the date and start hour of the readings
* EndTime – This POSIXt value is the date and end hour of the readings
* MeterID – This integer value is the unique identifier of each meter in the dataset.
* Reading – This integer value is the total usage for each Timeframe in kWh (kilowatt-hour).
* Temperature – This integer value is the air temperature during the Timeframe in degrees Fahrenheit.
* Humidity – This integer value is the relative humidity during each Timeframe as a percentage.
* LoadControl – A factor field, “Y “for periods with LoadControl activated, “N” for periods when it was not.
* HourOfDay – This is an integer representing the ending hour of the readings, this is derived from the EndTime.
* DayOfWeek – This is a string, the day of the week corresponding to the date of the reading.

For our first question, we want to find out if temperature (A) or humidity (B) have a significant impact on usage using a 0.05 significance level. To do this we will break them into groups of high (h), medium (m) and low (l)

= = = = =

are different from each other

We will also test for significant difference between the interaction terms;

For our second question we want to find out if there is a significant difference between usage over the same hours when load control is enabled and when it is not.

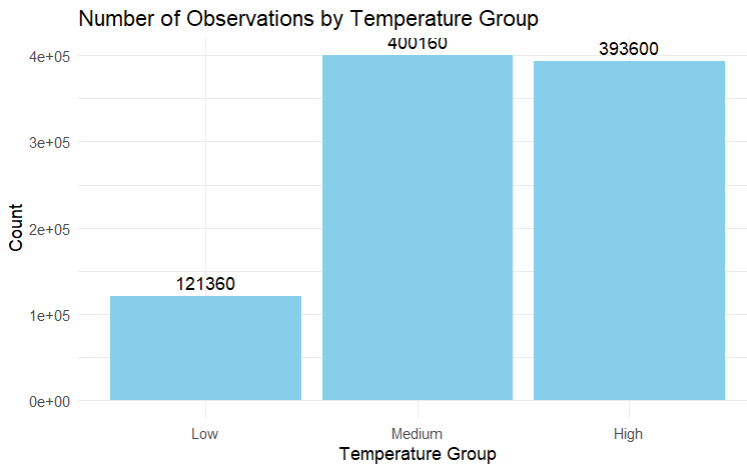
=

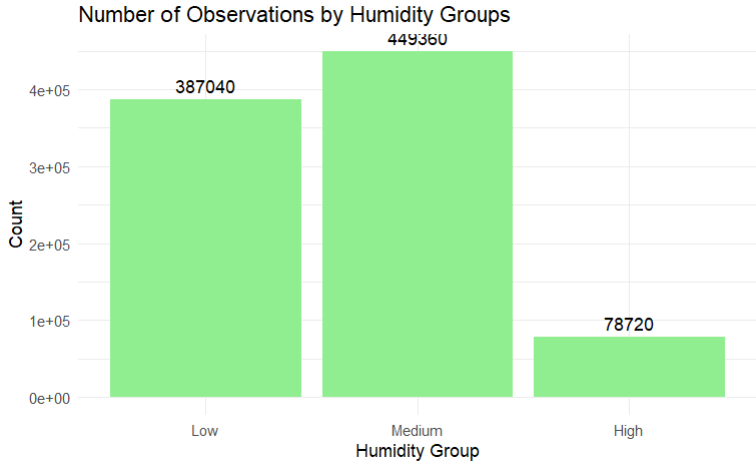
≠

**Methods:**

Before running the ANOVA Tests, we conducted exploratory data analysis. We removed any rows in which meter readings were missing (“NA”). This preserved 913,807 rows of our dataset. Missing values in temperature and humidity were imputed with the median value for each variable. To prepare for ANOVA, the independent variables Temperature and Humidity needed to be converted to categorical variables. Temperature was categorized as "Low" (60–70°F), "Medium" (70–80°F), and "High" (80–92°F) categories, while Humidity was grouped into "Low" (50–65%), "Medium" (65–80%), and "High” (80–100%) categories.

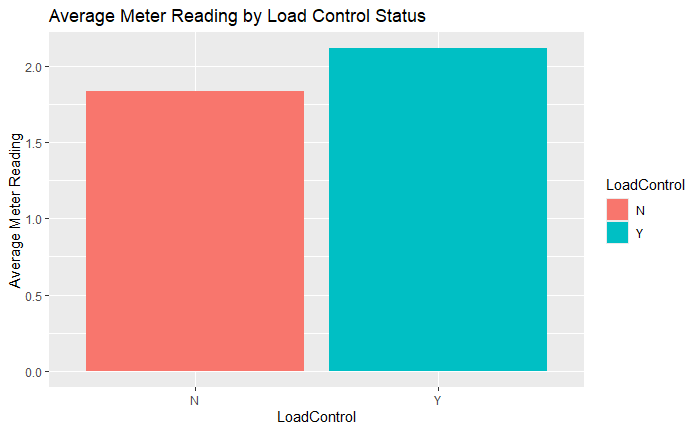
We chose to use ANOVA tests because we were comparing the mean electricity usage across three or more groups for both temperature and humidity. One way ANOVA is the appropriate test for determining whether there were significant differences in usage across temperature groups and humidity groups individually, while controlling for overall error rates. Additionally, we used two-way ANOVA to evaluate whether interactions between Temperature, Humidity and Load Control had a combined effect on electricity usage that would not be detected through one way analysis alone.



*Figure 1. Number of observations by Temperature group.*

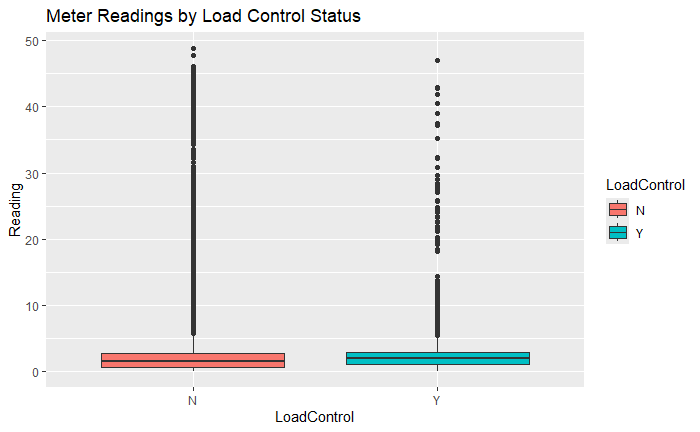
*Figure 2. Number of Observations by Humidity Group*

As shown above in Figure 1, most temperature observations fell into the medium and high categories, with fewer readings in the low group. Similarly, humidity readings are predominantly medium or low, with fewer readings categorized as high humidity, shown on Figure 2. The variable, LoadControl, which contains values of “Y” or “N” was converted to a factor for analysis.



*Figure 3. Average Meter Readings with and without load control.*

Shown above in Figure 3, there does appear to be some difference in the meter readings, with average meter readings being lower when load control is not enabled. However as shown in the boxplot in Figure 4, there is a high degree of outliers in the data that affect the variance, showing that household level differences can influence the variance.



*Figure 4. Meter Readings with and without load control.*

With our hypotheses determined and our data prepared, we created our models to test them.

**Conclusions:**

The results of our one-way ANOVA test are shown below in table 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | DF | Sum Sq | Mean Sq | F Value | Pr(>5) |
| LoadControl | 1 | 7,809 | 7,809 | 3,135 | <2e-16 |
| Residuals | 897,430 | 2.235.321 | 2.49 |  |  |

*Table 1. One-way ANOVA*

We observe a high F Value of 3135 as well as a low p-value of < 2e-16 indicating we would reject the null hypothesis. There is a significant difference in readings depending on the Load Control status. This can also be seen in the Tukey results with a p-value of 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | numDE | denDE | F-value | p-value |
| Intercept | 1 | 910522 | 8522.89 | <.0001 |
| TempRange | 1 | 910522 | 58160.79 | <.0001 |
| HumidityRange | 2 | 910522 | 133.80 | <.0001 |
| LoadControl | 1 | 910522 | 3832.12 | <.0001 |
| TempRange:HumidtyRange | 2 | 910522 | 111.23 | <.0001 |
| TempRange:LoadControl | 1 | 910522 | 238.27 | <.0001 |
| HumidityRange:LoadControl | 2 | 910522 | 153.24 | <.0001 |

*Table 2. Two-way ANOVA*

Shown above, are our results for a two-way fixed measure test to investigate the effect of Temperature, Humidity, LoadControl, and their interactions. As we can see all F-values are large, and all p-values are less than the 0.05 significance level; therefore, we will reject the null hypothesis. All terms are significant in their effect on the Reading variable for usage.

Our findings show that the given variables have a significant effect on the usage of central air. From our plots we can see that even under load control, the usage is still higher than the readings not under Load Control. This surprising finding is likely due to Load Control only being enabled during peak times, when air conditioning demand was naturally at its highest. If this experiment were to be run again, we may yield better results by having a control group of homes where Load Control is not enabled during peak times to compare the two groups under the same period and see how effective Load Control is at lowering usage.

GitHub repository: <https://github.com/IansHub/STAT301-ANOVA-Project>