Content Page

1. Pre-Project Plan
2. Monitor
3. Introduction
4. Data Cleaning
5. Exploratory Data Analysis

1. Further Insights
2. Data Modeling
3. Conclusion
4. Reflection
5. References (if any)

## Pre-Project Plan

|  |
| --- |
| **Goal Setting** |
| I aim to complete my project by 24*/12/2021.*  I shall take initiative to find out the information needed.  I shall check the project rubric to ensure all items are done before submission. |

My data set is Appliances Energy – March and May.

My preliminary questions that I will answer from my data set:

1. Which day of the month do we have the lowest average temperature?
2. Which day of the month do we have the most ideal average relative humidity (30-50%)?
3. What is the relationship between relative humidity and temperature?
4. Is there a correlation between relative humidity and dew point temperature? A low dew-point temperature of 16 is ideal so that peoples’ perspiration would be at its fastest rate.
5. How does pressure affect windspeed? Are there other factors which strongly affect windspeed?
6. What is the correlation between energy consumption and the surrounding temperature and humidity, and how to minimize energy consumption?
7. How effective is the data collected from the one floor, and is it reliable enough to predict the data collected from the other floor?
8. Are there any factors that are independent (not reliant on other data)?

**Monitor**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task/Milestone** | **By When** | **Actual Completed Date** | **Comment**  (On-time/delay/early) |
| Download the data.  Understand the rows and columns. | **21 Nov 2021** | **19 Nov 2021** | **Early** |
| Background research of delivery mode, function of weather. | **28 Nov 2021** | **18 Nov 2021** | **Delay** |
| Perform data cleaning. | **3 Dec 2021** | **18 Dec 2021** | **Delay** |
| Perform data transformation. | **8 Dec 2021** | **19 Dec 2021** | **Delay** |
| Exploratory Data Analysis | **17 Dec 2021** | **20 Dec 2021** | **Delay** |
| Submit Report 1 | **24 Dec 2021** | **23 Dec 2021** | **Early** |
| Answer my preliminary questions | **9 Jan 2022** |  |  |
| Data modelling | **10 Jan 2022** |  |  |
| Final report conclusion and reflection | **14 Jan 2022** |  |  |
| Create Dashboard | **17 Jan 2022** |  |  |

1. **Introduction**

I would be answering the following questions:

* Provide some background information.
* Refer to the question description.
* Include some other research that are related to the topic.
* Could be why the data was collected?
* Or how the data was collected?
* Or what can we learn from the data?
* Or any past analysis done before?
* Include your research references in section 10 of the report.
* Give an overview of the data structure.

This data set is recorded in 10-minutes intervals from March to May in Belgium. To monitor the temperature and humidity, a wireless sensor network is used in a 2-storey house. Weather data has been garnered from the weather station in the proximity. The data was collected during the Spring season in Belgium, where the weather was warmer.

Weather data would benefit people as they would be able to plan their schedule ahead of time. There is a growing trend in maximizing productivity. Having certainty in the upcoming weather is an important factor which goes into planning for the day’s activities. This data contains information on temperature, pressure, and relative humidity for the months of March through May. The company aims to gather some insights on weather prediction to effectively guide through people’s plans for the day.

This data contains information on the temperature, pressure, and relative humidity for the months of March and May respectively. For any weather data, knowing the geographical location would be useful as well so that we would know what season that region is experiencing. We would be assessing the data based on the populace satisfaction i.e. to find weather conditions that people would prefer. We would also predict the future weather conditions and check how accurate our results are to the actual data.

Table

Description automatically generated

There are 8317 observations (rows of data). Each observation is a sensor output. There are 27 variables (columns headers).

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Variable | Data Type | Minimum Value | Maximum Value |
| date | Categorical | 1/3/2016 | 31/3/2016 |
| Appliances | Numerical, Discrete | 20 | 880 |
| lights | Numerical, Discrete | 0 | 50 |
| T1 | Numerical, Continuous | 18.6 | 26.26 |
| RH\_1 | Numerical, Continuous | 27.02333333 | 54.66666667 |
| T2 | Numerical, Continuous | 16.2 | 29.85666667 |
| RH\_2 | Numerical, Continuous | 20.46333333 | 54.09 |
| T3 | Numerical, Continuous | 19.1 | 29.236 |
| RH\_3 | Numerical, Continuous | 28.76666667 | 47.69333333 |
| T4 | Numerical, Continuous | 17.29 | 26.2 |
| RH\_4 | Numerical, Continuous | 28.716 | 50.76 |
| T5 | Numerical, Continuous | 17.23 | 25.795 |
| RH\_5 | Numerical, Continuous | 29.815 | 90.0 |
| T6 | Numerical, Continuous | -3.59 | 28.29 |
| RH\_6 | Numerical, Continuous | 1.0 | 86.96666667 |
| T7 | Numerical, Continuous | 17.39 | 26.0 |
| RH\_7 | Numerical, Continuous | 23.39 | 48.29 |
| T8 | Numerical, Continuous | 19.0 | 27.23 |
| RH\_8 | Numerical, Continuous | 29.79 | 55.0 |
| T9 | Numerical, Continuous | 17.26 | 24.5 |
| RH\_9 | Numerical, Continuous | 29.85 | 51.536 |
| T\_out | Numerical, Continuous | -3.0 | 26.1 |
| Press\_mm\_hg | Numerical, Continuous | 738.8 | 769.9 |
| RH\_out | Numerical, Continuous | 24.0 | 100.0 |
| Windspeed | Numerical, Continuous | 0.0 | 13.0 |
| Visibility | Numerical, Continuous | 1.0 | 65.0 |
| Tdewpoint | Numerical, Continuous | -3.5 | 15.5 |

1. **Data Cleaning**

For the data cleaning process, I have first combined the two separate data sets by using the Concatenate Node. I checked for missing and have recorded them in a table below. Then, I identified missing values by using the Missing Value Column Filter Node. I checked once again to ensure I have not missed any missing data. Lastly, I have converted the data using the String to Date & Time node. The data is now clean and ready to be evaluated in the next section.

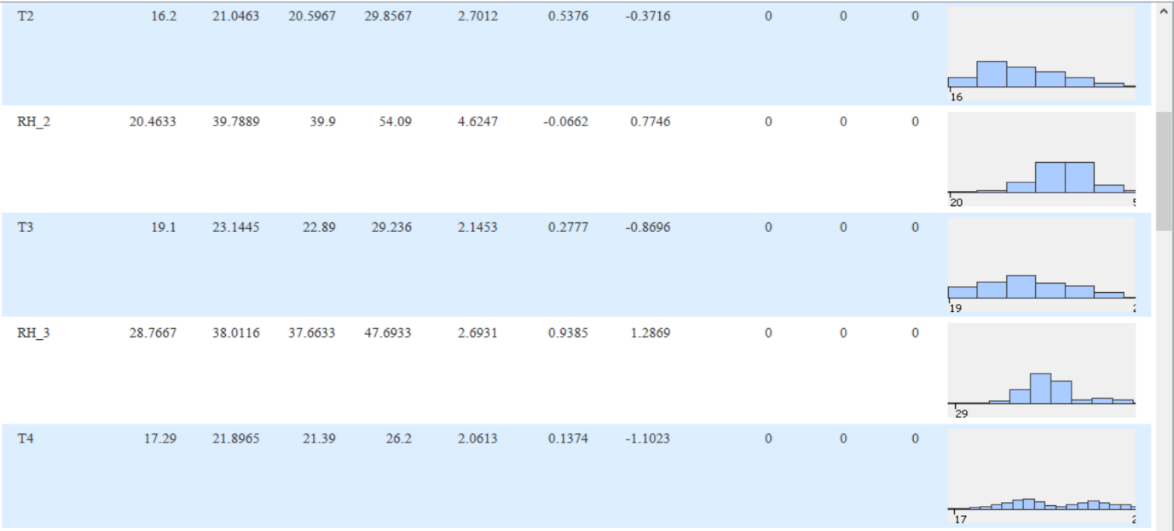
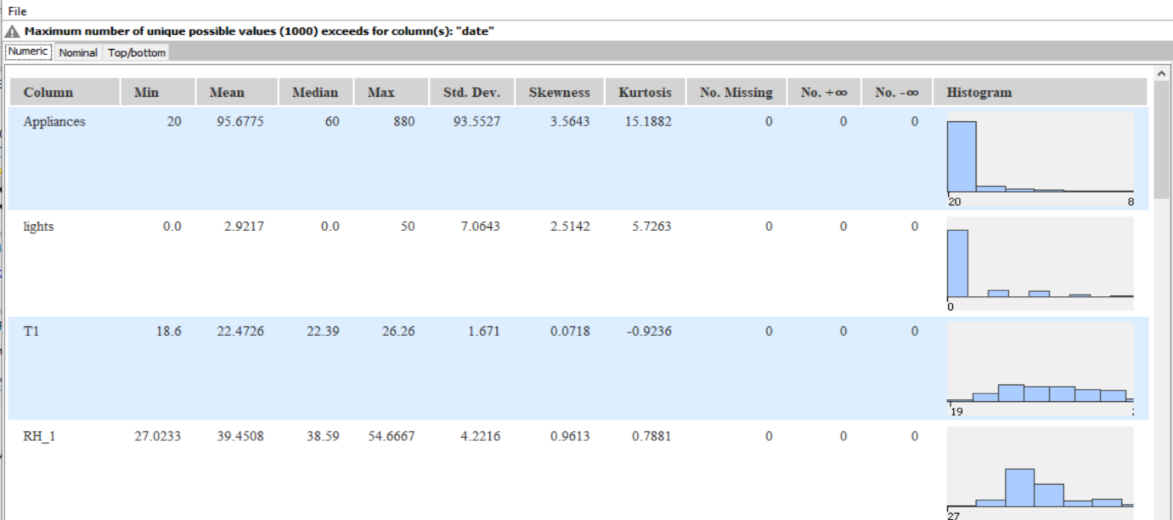
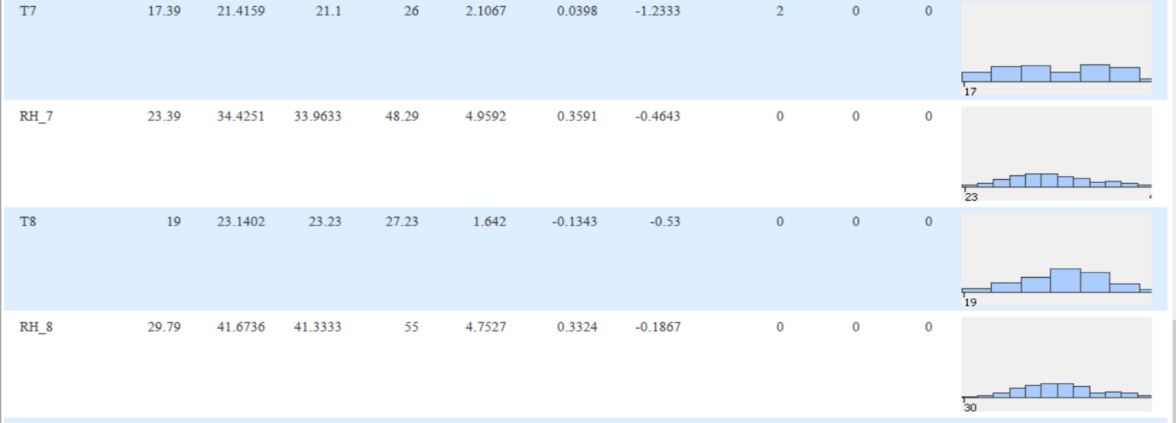
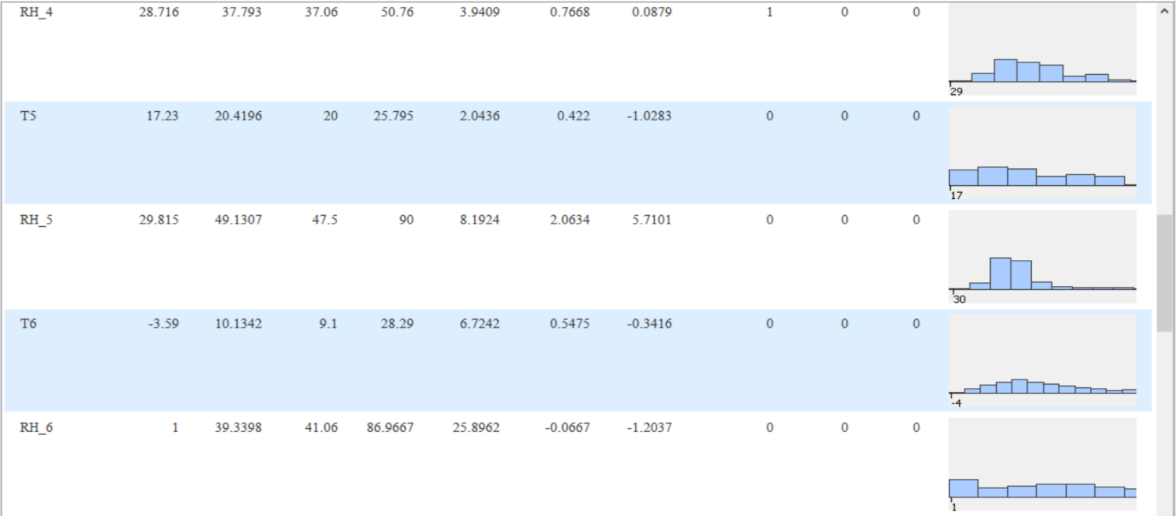
|  |  |
| --- | --- |
| Name of the variable | Number of missing values |
| RH\_4 | 1 |
| T7 | 2 |

Using the missing value node, the data have been reduced from 8317 to 8314, losing 3 rows of observations. We use a statistics node to ensure no missing values after this step.

Table

Description automatically generated with low confidence

**From the Statistics node, there was missing values in “RH\_4” and “T7”. These values were found to be blanks in the cell in the CSV file given.**



**If you are given more than one data set, you will need to either append them or merge them into one single file. Explain how you do it.**

Use the Concatenate node to append data. Under the Column Handling section, select “use union of columns” to append the 2nd data to the first. In my project, I have appended the data from the May CSV file to the end of the data from the March CSV file.

**State if there is any cleaning needed?**

Data cleaning would be needed if there is missing data read from the File Reader node. We must check through all the data by using the Statistics node. If there are missing values found under the “No. Missing” column, we would have to clean the data. Data cleaning would also be carried out if the data type needs to be changed. An example of this would be a column with both string and numbers in its cells. To convert this data, we would need to use the Rule Engine node to convert the categorical data from strings to a given value so that all the data has the same data type.

**Describe how do you know if there are any missing values in your data?**

I would use the statistic node and look up the “No. missing” column to find out how many missing values exist and which column they belong in. Additionally, the location of the missing values can be found using the Extract Missing Value Cause node, should you want to find out more information about the specific cells.

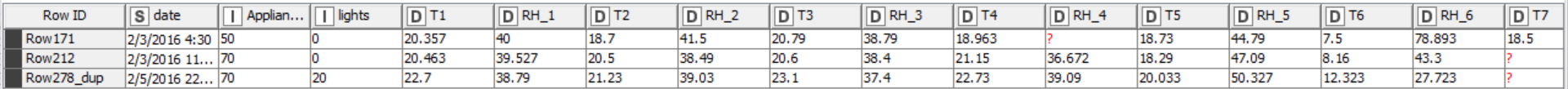
**How do you perform the cleaning?**

To solve it we can use the missing values node and select remove row. When data is missing there is an option to either modify or remove the row. Missing data occurs when they do not match the type given to its respective column. To modify missing data, we could set it to a fixed value that we recognize to be a null value. We also have the option to replace that cell with a previous or next cell’s value. After cleaning the data, it is a good practice to use the Statistics node to ensure we do not miss other missing data.

**How do you know if your data is clean?**

After using the missing value node, we use the statistic node and we look up the “No. Missing” column to check whether all the values are zero, which means that the data is clean. Furthermore, if all the columns are given the correct data type and would not need to be changed.

**Any reason for the missing value?**

The cells at their respective locations are left blank. This might have been caused by a sensor issue while recording the data. Another reason for this might be a corrupted csv file or a cell with invalid characters. Lastly, the missing value might have been a human error. The CSV is not read only and a user might have accidentally deleted the cells value without noticing it. The missing data in my project is shown below for reference.

**Data Transformation**

**Do you need to transform the date-time format? How?**

There is a need to transform the date-time format such that it is consistent and the graph that we plot will be more meaningful as it reflects any changes in the date and time. In my project, I have used regex to create a new column with date and time values from the date column.

To convert it, we use the String to Date Time Node, in this manner: d[.dd]/M[.MM]/yyyy H[.HH]:mm

The following legend applies:

|  |  |
| --- | --- |
| **Symbol** | **Meaning** |
| d | Date (1 digit) |
| dd | Date (2 digits) |
| M | Month (1 digit) |
| MM | Month (2 digits) |
| yyyy | Year |
| H | Hour using 24-hour clock (1 digit) |
| HH | Hour using 24-hour clock (2 digits) |
| mm | Minutes (2 digits) |

Table

Description automatically generated

**Do you need to create new variables? (e.g., Day of week)**

Yes, I would create new variables to test my hypothesis . For example, a column to represent the date of the month for each row. This would be useful when collecting data for each day by using the Group By node. An example of an application of this node would be to find out the mean temperature change throughout the month to determine whether the weather is getting warmer or colder and by how much.

**Do you need to find co-relation between which variables?**

Yes. Co-relation is important when we are using the predictors models like regression. Co-relation is especially useful for big data, so that the user can identify the relationships between data and perform further analysis.

**Do you need to use Aggregation? Group by month? Group by different type of users?**

Yes. Aggregation can be used to find standard deviation and mean between a large data set for each group. Groups are a method operation performed for a similar given variable. These variables are chosen based on the criteria that we are comparing for e.g., dates of the month, date of the year and energy consumption. For example, there is a class of boys and girls. A simple use of aggregation would be to find the mean height of boys and girls respectively so that we could make a comparison. Another use of grouping using the same scenario would be to group classes by date to find out during which year of study would the students most likely experience a change in height and by how much on average.

1. **Exploratory Data Analysis**

Statistical analysis of all the numerical variables

From this data, I found that appliances have the most skewed data. This can be shown using a histogram. The reason for this skewness would be that a large majority (87.5%) of the values are less than 140 kWh followed by a small amount of data with values more than 140kWh. Another way to approach this would be that the mean is 95.69 while the max is 880, meaning that the data is highly skewed.

Another finding would be that the standard deviation of T1 is the lowest among the data set. This would mean that there is little variation between the values in each observation. This can be further observed using a graphical plot and viewing the change in values over time. This can also be seen from the minimum and maximum values having about an 8 degrees difference.

Graphical user interface, timeline

Description automatically generated

Timeline

Description automatically generatedGraphical user interface, timeline

Description automatically generated

Timeline

Description automatically generatedGraphical user interface, application, table, timeline, Excel

Description automatically generated

Data Transformation

A picture containing table

Description automatically generated

I have extracted the date value portion from the date column. This cannot be done using indexing only as date values can range from one to two digits. The date column is a string and must be parsed by finding the index of the ‘/’ character and slicing the string data. This process can be repeated with slight adjustment to separate the month, year and time values as well, if needed for my project.

The expression used is: substr($date$, 0, indexOf($date$, "/"))

The expression would be explained in the table below:

|  |  |
| --- | --- |
| substr | Slices a part of a string for all data in a column |
| 1st Input ($date$) | Select which column to slice from |
| 2nd input (0) | Starting point of result. In this case, zero refers to the start of the string. |
| 3rd input ( indexOf($date$, "/")) ) | Ending point of result. This would be defined as a numeric value. |
| indexOf | This function is used to look for the first instance of a selected string. In this case, we are looking for the ‘/’ in the date column. In other words, we must change the index depending on the position of the last digit of the date value. |

Chart, histogram

Description automatically generatedData Mining of Pie Chart and Histogram

Describe the percentage of probability

Since March used 54.39% of the total appliances, it is more likely for higher energy consumption in March than in May. This might mean that heating systems were used during the colder periods but was not used as often when the weather became warmer. Another reason for lower energy consumption would be that May might have been less busy month or that the house was not used as often due to other commitments during this period compared to the month of March.

Describe your findings

For the histogram, the graph is right skewed, since the bulk is on the left edge of the graph. The data represents the outside air temperature from the nearby weather station. As such, it would be reliable data of the surrounding air temperature change over time due to the change in the season. Since, the data is right skewed, it would mean that the temperature increased with a steep gradient then decreased with a gentler gradient. This might be important because if the temperature change was too large, the public might not be ready for it or might even suffer heatstroke from it. In this case, the temperature peaked at 26. 1 degrees so the heat might not be an issue. However, the temperature dipped to -4 degrees which could also affect the people’s health. Other than the formation of snow, which people would have to rake off, cold weather would leave people prone to dry or cracked skin and in some cases, asthma attacks.

There is only one mode, which is at 8 degrees Celsius, which would be a good gauge for the most common temperature than would be expected. There is a large sample size of 8314 observations, which means that the data is comprehensive and reliable. Lastly, there is one outlier at -4 degrees Celsius. An outlier may be due to variability in the measurement, or it may indicate experimental error; the latter are sometimes excluded from the data set.

1. Further Insights

1. Data Modeling
2. Conclusion
3. Reflection
4. References

<https://seasonsyear.com/Belgium>