Database redesign of home sensors and weather instruments for simplified data analysis

Data Management (12741) Fall 2020

Yash Shailendra Gokhale
Dept. of Chemical Engineering,
Carnegie Mellon University,
Pittsburgh, PA, USA
ysg@andrew.cmu.edu | yashsgokhale@gmail.com

ABSTRACT

Energy usage of the house is an important parameter of consideration while modeling energy systems or studying the demand-supply dynamic of electricity supply. Accurate study of energy use for households serve as a critical aspect for energy supply corporations to develop their plan of action. Temperature and humidity sensors are the most commonly used type of sensors to monitor environmental conditions for the area under focus and can affect the energy usage for the area. These devices are used to monitor extreme air conditions or to detect fluctuations in the atmospheric conditions [1]. Further, the increasing variability in climatic conditions poses a challenge in understanding the relation of energy usage. To facilitate long-term electricity planning and short-term electricity management, it is necessary to be able to provide a satisfactory quantitative representation [2]. This project aims to put forth a framework, which would aid in efficiently managing data from multiple sources and create chunks of new datasets for simplified data analysis.

1. MOTIVATION

As per the standard form of data collection, time-series data from temperature and humidity sensors are combined with atmospheric parameters measured at the same time instant. Although such a framework provides easier access to this data at a single source and simplifies the look-up process, it is often tedious to derive useful insights from the data. In this project, the existing database has been redesigned to a lucid format, with additional categorical variables being introduced, correlating the sensor and weather data. Additionally, custom SQL queries have been implemented to identify situations with high and low energy usage with its correlation to atmospheric data. Thus, the proposed database redesign schema aids energy analysts to develop complex predictive models based on these developed databases. These prediction models can be useful in a number of applications such as adequate sizing of energy storage systems, for demand side management (DSM) and demand side response (DSR), etc. [3].

2. DATA SOURCES

The time-series data, sourced from [3], is from a house located in Stambruges, Belgium with data (appliance load, light load, temperature and humidity) logged every 10 minutes. Additionally, weather data from the nearest airport weather station (Chievres Airport, Belgium) is merged with the sensor data, for every time measurement.

The data file is in form of a '.csv' file, downloaded from the public repository on Kaggle [4]. Data from nine from the house, viz., Kitchen area, Living Room, Laundry Room, Office Room, Bathroom, North building side, Ironing Room, Teenager Room and Parent's Room, have been recorded. Additionally, parameters that are recorded from the weather station are: Air temperature, Air Humidity, Wind Speed, Humidity, Visibility and Dew point. Appliance usage (kWh) and Lights usage (kWh) are the two dependent variables that are added to the data file. The two random variables that are added to the original dataset are ignored, as they are mainly used for modeling and are out of the scope of this project. In total, there is a time variable, 18 sensor variables, two energy variables and nine weather parameters, contributing to 27 attributes.

3. PRELIMINARY DATA ANALYSIS

Due to the a significant number of attributes for the data, it is important to understand the distribution of certain key variables, which would pave way for further restructuring of the data.

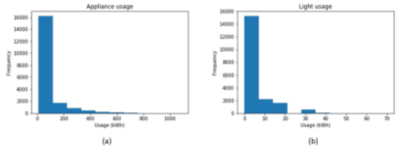

Figure 1: Frequency distributions of appliance and light use in kWh, which are the dependent variables for energy modeling

From Figure 1, it can be observed that, over the period of the entire dataset, the house operates on a low energy use (~0-200 kWh) and a low light use (~0-20 kWh).

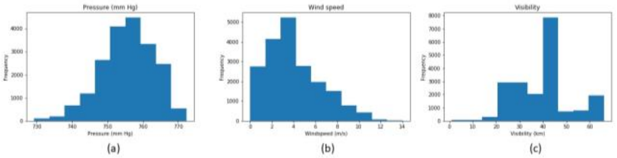

Figure 2: Frequency distributions of atmospheric variables derived from the weather station.

It is important to understand the degree to which values of the atmospheric variables influence the energy usage in the house. The atmospheric variables, which are plotted on Figure 2 can thus be categorized into different sections to aid understanding the data on a broad scale. For instance, the atmospheric pressure can be categorized into two categories as: If the pressure is less than 760 mm Hg, the pressure is below atmospheric pressure. Similarly, if the pressure is equal to or above 760 mm Hg, the pressure is equal to or above atmospheric pressure. Similar categorizations can be done for the other variables, the study of which has been done in the subsequent sections of the paper.

4. OBJECTIVES

Based on the preliminary analysis in Section 3, correlating energy usage to atmospheric parameters is an important aspect of study for energy analysts. As the data is also efficient in capturing temperature and humidity data for nine locations across the house, it is also equally important to compare the data from these sensors to temperature and humidity in the atmosphere. The primary objective of this project is to thus, run customized SQL queries to analyze the dependence of these variables on each other. For instance, a sample query could be to check what the average energy usage was, if the pressure was higher than the atmospheric pressure and there was a high wind speed in the atmosphere. Answering such questions by studying the trend in the data forms the objective of the project. Thus, developing a look-up table using MySQL is to be achieved for the scope of this project.

It can be observed that, for this particular dataset, the single table contains a huge amount of attributes, leading to confusion in reading the data and deciphering the hidden insights. To enable easier querying of data and further organizing the data into multiple tables, a database restructure has been proposed, through the introduction of a new database schema and defining the corresponding functional depedencies.

5. DATA INSPECTION USING MySQL

Apart from the sensor data in the nine rooms in the house, the atmospheric variables to be inspected are: Atmospheric pressure, visibility, wind speed, external temperature and humidity. Each of these atmospheric variables have been categorized into two sub-divisions based on the range as seen in Figure 2, as the upper division and the lower devision. The two decision variables, Appliance usage (in kWh) and lights usage (in kWh) are categorized as following:

Table 1: Categorization of Decision Variable Data

Variable	Division
Appliance usage (A), kWh	$A < 100$: Low Usage $100 \leq A < 300$: Moderate Usage $A > 300$: High Usage
Lights usage (L), kWh	$L = 0$: Lights Not in Use $0 < L \leq 10$: Low Usage $10 < L \leq 25$: Moderate Usage L>25: High Usage

Moreover, it was observed that 77% of the values in the dataset had a zero usage for lights. This prompts a separate study of the data for cases when light was in use and cases when light was used. The results for the same are tabulated, indicating the percentage of data present for each of the subcases.

Table 2: Look-up table for appliance usage in the house

	Low Energy	**Moderate Energy**	**High Energy**	**Avg T**	**Avg RH**
Low Pressure	0.56	0.11	0.04	8.15	80.24
High Pressure	0.23	0.04	0.016	5.58	78.53
Low Visibility	0.29	0.05	0.019	6.74	82.27
High Visibility	0.5	0.11	0.04	7.78	78.35
Low Wind	0.58	0.11	0.036	7.1	80.68
High Wind	0.2	0.05	0.021	8.23	77.27

Table 2, which was compiled using customized SQL queries, is an effective way of analyzing the composition data. The look-up table, thus is effective in conveying a lot of information in a customized way. For instance, by looking up the keys in the table, it can answer the question of: 'What is the major usage pattern when atmospheric pressure is low?', which answers that, around 56% of the cases correspond to a low-energy and low-pressure situation. Another interesting question which the look-up table can solve is, 'In which case is the average atmospheric temperature higher, for low windspeed or high windspeed?', to which the answer is 'High Windspeed' as the average temperature is 8.23^{o} C for high windspeed whereas 7.1^{o} C for low windspeed, all the values which are derived from Table 2. Thus, development of look-up table using MySQL serves to answer a variety of questions through simplistic comparisons.

6. DATABASE RESTRUCTURING

The original database is a "Mega" database, containing all the attributes and features in a single table. In order to enable efficient querying of data and for futher insertion of data, a new database schema has been proposed in this paper.

The entity relationship diagram, developed in Figure 3 enables effectively managing the data, obtained from multiple sources. Each of the entities in the E-R diagram are representative of a section of the data from the original database and thus highlights the robustness of the new schema. The tables: Sample, Rooms, Weather, Energy, Independent Variables and Correlated Variables make use of existing attributes in the original schema.

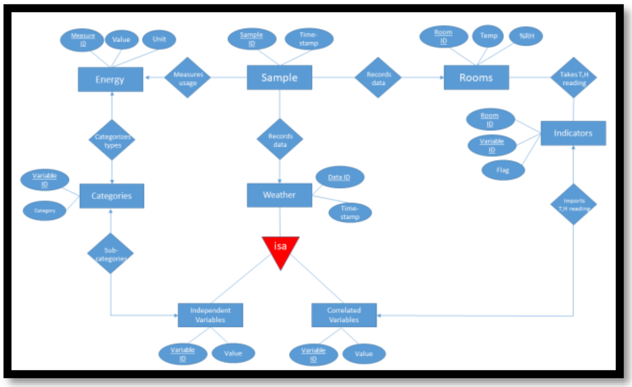

Figure 3: Entity-Relationship Diagram for the new proposed simplified database schema instead of the current database

This restructuring thus enables segregating complicated data into multiple tables, thereby enabling easier addition of new entries into the table and updating the data. Another accomplishment of the new database design is that, it simplifies dealing with missing data, as it is immune to failure of either the home sensors or inability to record weather data. As per the earlier database design, since all the attributes were recorded into a single table, missing values would yield a 'NULL' value by default. On the other hand, in case of the new schema, data from the home sensors could be added independently of the weather gauge data and vice-versa.

In addition to the six new tables in the new schema, adding a number of derived attributes are recommended, which would simplify the data analysis process. Two tables, 'Categories' and 'Indicators' are proposed to be added to the database schema, which would transform the data into categorical variables and also correlate data from the home sensors and weather station. Based on this, the following situations are analyzed and answers to the following questions are put forth: (Queries to the questions can be found in the appendix)

1. How does the temperature of a particular room compare to the outdoor temperature?

By running a customized query, trend of temperature in a particular room can be analyzed. Running the query on 'Kitchen' room, which uses an inner join operation on the Temperature and Correlated variables tables, it is observed that in 99.91% of the readings, the room temperature of the kitchen was higher than the outside temperature. Moreover, the average temperature for the kitchen was 21.68 ^{o}C whereas the average outdoor temperature was 7.95^{o} C. Based on the output of this query, it can be generalized that the kitchen is always warmer than the outside temperature. Thus, the new database schema is effective in deriving qualitative answers to routine questions.

2. Can dew be expected at a particular moment?

Dew formation is an important environmental process and is an important parameter in study of atmospheric systems. Through an inner join on the Independent and Correlated variables, the likelihood of dew can be estimated by comparing the atmospheric temperature and dew point. In general, dew is formed when the dew point temperature is greater than or equal to the air temperature [5]. Using a combination of Case, Select, Where and Inner join functions in SQL, the answer to this question can be found out. Based on the output of the customized query run on the new database schema, it was observed that dew was likely to be absent for 98.3% of the cases. This helps in understanding the likelihood of dew present in the atmosphere and running deeper customized queries, correlating other variables would help in understanding the conditions conducive for dew formation.

3. Creating a table using existing data with categorical variables

New tables with categorical variables are created in SQL, using a customized query which utilizes a combination of Create, select and case statements in MySQL. Creating a new table with a categorical emedding can be used in conveying insights in the dataset to a wider audience. For instance, part of the table 'Categories' in Figure 3 has been recreated using the query, which is indicative of the distribution of appliances and lights.

4. Comparing new categorical variables to existing data

Data from the newly created tables, such as 'Categories' can be used to answer certain practical questions. One such question is: What could be the average visibility when light energy use is high and when the light energy use is moderate? By using the customized query, this question can be answered effectively and could explain the need for usage of lights. It can be observed that the average visibility is 38.83 units for high light usage whereas 38.49 units for low light usage. In this manner, the new categorical variables can be used to provide further qualitative results.

7. SUMMARY & CONCLUSIONS

Sensor data and weather measurements are of extreme importance to civil engineers and energy analyst to be able to effectively model energy systems and design future models. Through this project, an effective method of developing look-up tables has been discussed in the paper, with a practical example of categorizing the data based on the values of the variables has been highlighted, which can be seen in Table 2. Developing such look-up tables developed using MySQL provides non-technical professionals to derive useful insights from the otherwise clustered and messy data.

The original database system follows a 'Mega' table approach wherein a single table is capable of storing the entire information. Although this might be useful in storing the data concisely, it tends to be complex and needs to be converted into a simplified form. Keeping this in mind, a new database schema was conceptualized, as put forth by the E-R diagram in Figure 3: Entity-Relationship Diagram for the new proposed simplified database schema instead of the current database This new database schema is simpler to use, allowing easier insertion of data and capable of bypassing 'NULL' values in the new entries. This powerful schema is capable of answering a varied set of questions, as highlighted in Section 6. The new schema is effective in correlating variables with ease and also paves way for adding data in sections. For instance, the new database schema is capable of updating data from the room sensors and weather gauges indepently. Failure in recording data from either of the two sources does not lead to missing values in the new database schema. Correspondingly, it can be thus stated that the new schema is more robust than the original schema and is capable of providing deeper insights into the data.

8. ACKNOWLEDGEMENTS

I would like to thank Professor Mario Berges for teaching the key concepts of data management such as ER diagram, SQL queries, etc and providing me this opportunity to work on such a novel and interesting project. I would also like to thank the Teaching Assistant, Bingqing Chen for clearing my conceptual doubts and providing feedback on the developed codes. Lastly, I'd like to thank the scientific community of SQL coders for providing solutions to common syntax problems encountered in using SQL language through online platforms such as Stack Overflow.

9. REFERENCES

[1] Comptus, 'The Importance of Temperature and Humidity Sensors', 19th December 2014, https://www.comptus.com/blog/the-importance-of-temperature-and-humidity-sensors

[2] Guan et al., 'Incorporating residual temperature and specific humidity in predicting weather-dependent warm-season electricity consumption', Environmental Research Letters, 20th February 2017, https://iopscience.iop.org/article/10.1088/1748-9326/aa57a9

[3] Candanedo et al., 'Data driven prediction models of energy use of appliances in a low-energy house', Energy and Buildings, https://doi.org/10.1016/j.enbuild.2017.01.083

[4] Kaggle, Appliances Energy Production, Gokagglers, https://www.kaggle.com/loveall/appliances-energy-prediction

[5] WeatherQuestions.com, 13th December 2019, https://weatherstreet.com/weatherquestions/What_is_dewpoint_temperature.htm

10. APPENDIX

SQL scripts:
[1] Data Inspection: [GitHub]
[2] Data Redesign: [GitHub]
[3] Dataset: [GitHub]

Database Schema:
SAMPLE (sampleID, timestamp)
TEMPERATURE (sampleID, kitchen, livingroom, laundryroom, officeroom, bathroom,northbuilding, ironingroom, teenageroom, parentsroom)
HUMIDITY(sampleID, kitchen, livingroom, laundryroom, officeroom, bathroom,northbuilding, ironingroom, teenageroom, parentsroom)
INDVARS (measureID, pressure, visibility, windspeed, dewpoint)
CORRVARS (measureID, Tout, RHout)
ENERGYUSE (measureID, Appliances, Light)