# **BI Project- 19667**

## Tool Used: PowerBI

## Dataset:

<https://www.kaggle.com/datasets/taranvee/smart-home-dataset-with-weather-information>

The following points were summarized about the data after the data understanding phase:

1. The Dataset is of a smart home service product which is tasked with recording measurements/readings.
2. The data describes the electricity consumption of house components, which are appliance in use and power usage in different rooms, in kilo-Watts.
3. The dataset also has information related to the electricity generation being done by use of solar power, also in [kW].
4. After describing the consumption by the aforementioned components, the dataset gives extensive weather-related information such as:
   * 1. Categorical: How is the weather? Sunney? Rainy? Etc.
     2. Numerical: Weather indicators such as Temperature, humidity index, etc.
     3. Precipitation index and expectation
5. The columns related to weather information usually converge to give us the weather condition (Summary) and temperature, so we’ll use these two columns for our primary analysis.

## Problem Statement(s):

1. Make the Electricity Consumption and generation efficient using smart home technology.

2. Control the electricity consumption and bringing it down by focusing on individual elements like appliances/rooms, weather conditions, usage over time, etc.

## Data Wrangling:

**Preparation and Transformation:**

* After understanding the data, it was observed that the dataset had 32 columns and almost 500K rows.
* The Dataset was loaded to python, and we observed the rows and columns.
* Based on the wrangling technique taught in the course, the wrangling started in the following manner
* Observed the shape of data, head and tail of data rows and data types of the fields present.
* Missing Value replacement:

1. While observing the tail and shape, we saw that there was a ‘/’ in one row and no corresponding data
2. We used isna().sum() function to calculate missing value, and it generated 1 in all fields except one, thus confirming our assumption of having the slash in one row.
3. Best strategy to handle the missing value here was the deletion on that 1 row because deleting it will have no effect on the data, so we did it.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

* Taking care of the time column, which is the most important column for analysis:
  + - Time column had values as big integers, and the dataset description told us that the time was recorded by a 1-minute interval of first 350 days of 2016.
    - The pandas function ‘to\_DateTime()” does work on an integer range sso big, so it couldn’t be used.

Text

Description automatically generated with low confidence

* + - Instead of using the function mentioned above, we replace the column by a column of **our own,**
    - In this column that we make for Time, we will make our own time series accurately replicating the original one and will adjust it according to the description given in the description of dataset,
    - We start, according to description, the time from 1st Jan 2016 at 5:00 AM, and apply a 1-minute increment on every row, thus achieving an accurate result.
    - We discard the original time column and replace it with the new one.

Graphical user interface, text, application, email

Description automatically generated

* Next step is to take care of the “cloud cover” column.
* Although we didn’t use this column in our analysis, we should take care of this column before proceeding to data analysis.
* We delete the string variables in this column to null values, then used backfill to deal with these values to replace it all the way to the top.

Graphical user interface, text, application

Description automatically generated

* ‘Icon’ and ‘Summary’, ‘Solar [kW]’ and ‘gen [kW]’ and ‘use [kW]’ and ‘House Overall [kW]’ are the pairs of columns that are the same, so we drop one of these columns from the mentioned pairs.

Graphical user interface, text, application, email

Description automatically generated

* There were no cost columns in the dataset, whereas cost is one of the biggest indicators of consumption in such datasets,
* We create the costs columns by multiplying the total consumption of house by the **KE electricity price kW/minute (0.277 PKR) in 2016**.

Text

Description automatically generated

* We use box plots to analyze outliers. There were no apparent outliers in the data since consumption varies highly in extreme seasons, however there were a few instances when the consumption was very high and were back to normal the next minute, so it was decided to remove these rows as they were very rare instances and they could distort the analysis.

Text

Description automatically generated with medium confidence

* Inferential Analysis:

Since we have maximum numerical data and only a small number of categorical fields, and it is clear that appliances will not have a positive or negative correlation with each other since they are used independently, we only run anova test to confirm some of our assumptions about the data that only weather will impact the total generation or consumption of electricity. All the other correlations are just a circumstance of this. See below:

Text, application

Description automatically generated with medium confidence

**Strategy behind Data Transformation**:

***Note***:

*Electricity usage by appliances and rooms is in kilowatts.*

*Cost of electricity is in PKR according to KE electricity prices (2016).*

*‘Summary’ is a consequence of all the weather-related columns.*

The strategy used here is that we have transformed the data and made necessary changes to it to make it fool-proof when we begin analyzing the electricity usage and how it can be made efficient. We include major factors that are not present i.e. cost of electricity and dropped unwanted columns. The transformed data itself tells us most about how the analysis will be.

## **Dashboards:**

Dashboards and their analysis is attached with the file of this assignment. The three dashboards depict the following analysis in a nutshell:

How all the rooms and appliances consumed electricity throughout 2016, how much of the total consumption is done by every component.

Statistics related to electricity. How much electricity was consumed during each month, what was the cost spent on electricity and the effect of weather and temperature on usage of electricity. Also, since we are generating electricity too, how much was generated and what is the cost we saved as a result?

This is a combination of the first two points. It describes the temperature, usage and consumption by appliances and overall during a specific period of time, which the user can provide by the sliders at the upper left corner. In the snapshot of the dashboard, I have selected the first half of December in 2016.

Chart

Description automatically generated

Chart, treemap chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

## **KPIs:**

1. Electricity Usage in the smart home in kilowatts,
2. Cost of electricity used IN PKR,
3. Electricity being generated through solar panels in kilowatts,
4. Cost saved by electricity generation in PKR
5. Factors increasing or reducing Consumption and electricity generation,

# **Final Analysis and Findings:**

**Story**:

The three dashboard tell the following stories respectively:

1. Analysis during custom period:

We see that it provides the key indicators during a specific period. In the picture, first half of December is selected, so we see that due to winters, the consumption is low especially during the latter days. Mostly the weather is cold (snowy and overcast) and the minimum and average temperatures are low too. The total electricity consumed in just almost 19000 kW, which is less than 10% of the consumption. Also shows the rooms most consuming the electricity during this period.

1. Overall Electricity Utilization:

The third dashboard displays the following story, it shows that the electricity cost and generation was slowing down during the first months of 2016, which indicates a spring and an intermediate winter season. The cost and consumption went drastically up due to summers after June and it started coming down at a major rate after October. It also shows that during summers the electricity generated was the highest, that's because the generation method is solar energy. It also shows how much the consumption end generation were during different months. Then goes on to show the average consumptions during different weather conditions and we see that the weather condition mostly related to summers are on the right of the stack bar chart which means that during summer most of the electricity is consumed as indicated by the other visuals as well. Finally, we see the cost that is saved and the cost that is spent on electricity throughout the observation, and the number of units consumed.

1. Appliance/Room Consumption:

Some of the most appealing visuals have been added to this dashboard. This dashboard first and foremost tells us the consumption of **all** the categories throughout different months of 2016. It then goes on to compare the electricity consumption of different appliances and shows that how these appliances are used the most in different months, and when they are used the least. The dashboard also shows that the break-up of the total consumption used by different rooms and by different appliances though tree map and pie chart respectively. Thus, we can see which appliances/rooms used the most and which appliance/room uses the least amount of electricity. Finally, we see the total number of KW consumed during the analysis. And we can derive the percentages through that number

### Key Analysis:

We have figured out that the consumption of electricity throughout different months of 2016 rely heavily on the weather conditions. The peaks in graph show that electricity consumption has been significantly high during MAY-JULY which is the main time of summer. Along with this, the solar power generation that is being done to support the electricity consumption it is not enough as the readings taken in this data set are from a city which has a climate susceptible to rain throughout the year, thus the solar power generators does not get enough sunlight to match the needs of electricity consumption. We have also seen how different appliances affect the electricity consumption but more importantly if we do not consider the furnace used for a second, most consumption is saturated inside the Home Office, the living room which are the primarily used rooms/areas in a house.