**Temperature Prophecy**

1) The data was loaded from the runtime files. We can also mount the drive and load the data from the mounted drive. (import Pandas library)

2) Then we can see the dimensions of the data by its shape, different columns, checking if any missing values (NaN or None) exist.

3) Then we import important libraries:

**NumPy** - a high-performance multidimensional array object, and tools, for efficient numerical operations, as a base for other numerical libraries like pandas and scikit-learn.

**holoviews** - a dynamic data visualization library. It is built on top of other visualization libraries like matplotlib and bokeh.

**matplotlib** - a plotting library. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits

**sklearn** – a machine learning library. It It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits

**os** – a portable way of using operating system dependent functionality like reading or writing to the filesystem, starting a subprocess, etc.

**prophet** – a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.

4) Then we loaded the dataset and pre-processed it.

* We dropped the **'room\_id/id'** column since it has only one value(Room Admin), so we don't need this column for analysis.
* We also renamed the columns **'noted\_date'** to **'date'** and **'out/in'** to **'place'**
* The Date and Time have been categorized and distributed as follows:
  + Date to the format Day, Month, Year, Hour, Minute='%d-%m-%Y %H:%M:'

5) The Seasonal information was changed as follows:

* Winter : December to February
* Summer : March to May
* Monsoon : June to September
* Post-monsoon : October to November

And the Timing information as:

* Night : 22:00 - 23:59 / 00:00 - 03:59
* Morning : 04:00 - 11:59
* Afternoon : 12:00 - 16:59
* Evening : 17:00 - 21:59

6) Then after checking whether any record is duplicated, it turned out that there were duplicate records. So, we put duplicate records into one unique record.

7) It was identified that column **'id'** seems to have some information related to 'date' column.

Column **'date'** does not have seconds information, so **'id'** may have seconds information or some uniqueness of when the data was collected.

8) The **'id'** column, it seemed to have unique values and two decomposable components, numeric and alpha-numeric.

In the case of **'id'** of **'\_\_export\_\_.temp\_log\_101144\_ff2f0b97'**, it can be decomposed into two parts.

numeric part : **101144**

alpha-numeric part : **ff2f0b97**

Alpha-numeric part looks impossible to understand, but numeric part may indicate uniqueness or sortability of each records, for example seconds information.

There also gaps found in **‘id’**

So, it can be said that 'id' column is not related to second information, but it can be used as a unique identifier of each record.

9) EDA - Exploratory data analysis

We plotted a bar graph to show how many times we have readings for each month in a year. The x-axis represents "Month" and y-axis represents "Percentage".

I also tried plotting the same from the point of view of "Seconds" as well.

The Univariate Analysis for "Monthly Readings", "Temperature", "Place", "Season", and "Timings" were done.

[I am still trying to fix the code for "Temperature", "Season", and "Timings"]