**TRANSFORMATIONS**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name:**

**Batch Id:**

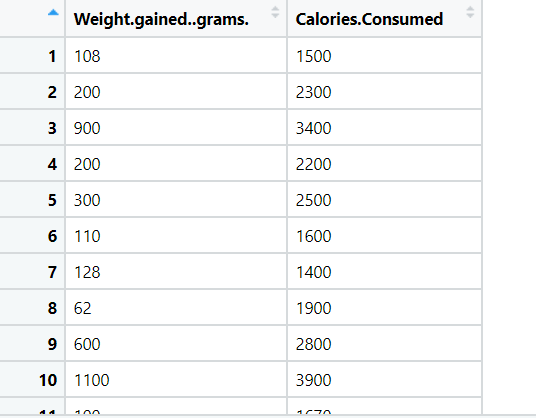
**Topic: Preliminaries for Data Analysis**

**Problem Statement:**

Everything will revolve around the data in Analytics world. Proper data will help you to make useful predictions which improve your business. Sometimes the usage of original data as it is does not help to have accurate solutions. It is needed to convert the data from one form to another form to have better predictions. Explore on various techniques to transform the data for better model performance. you can go through this link:

<https://360digitmg.com/mindmap-data-science>

1. Prepare the dataset by performing the preprocessing techniques, to have the data which improve model performance.



**Answer:**

**ANSWER:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Weight gained (grams) | About the gaining of weight | Quantitative, Continuous, Ratio | Relevant |
| Calories Consumed | How much calories consumed | Quantitative, Continuous, Ratio | Relevant |

**Objective**: To improve model performance

**Constraint:** Sometimes the usage of original data may not help to have accurate solutions

**Working with the Data**

* In order to start working with the dataset, we need to import the data first and the necessary libraries
* We use either Spyder, Jupyter, Colab, as per your system requirement, ease of use and availability.
* We import Pandas to import the data

**Pre processing and EDA**:

* We check the data description using “info” or “describe”
* Additionally checking the shape shows that the data consists of 14 rows and 2 columns/ features
* We check the data for missing values using the function “isna()” and summing it up with “sum()”
* It describes that there are no missing values present in the data
* To know the tendency of the dataset we opted first business moment such as mean, median, mode
* To know the measure of dispersion we check in second business moment by using range function
* We check the skew ness and kustosis in third and fourth business moment respectively.
* To know the distribution of the dataset we use histogram plot.
* We use scatter plot to know data dispersion
* Now for transforming the data we use **np.log function** and **sqrt function.**
* Hence the data is now pre-processed and EDA performed. It is ready for further processing!

**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Hint: Refer to calories\_consumed.csv dataset
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function