****

Subject:- ML

**Weekly Report Submission 3**

**Group:- DATA DYNAMOS**

Submitted to faculty: Prof. Mehul Raval

Date of Submission: 04/03/2023

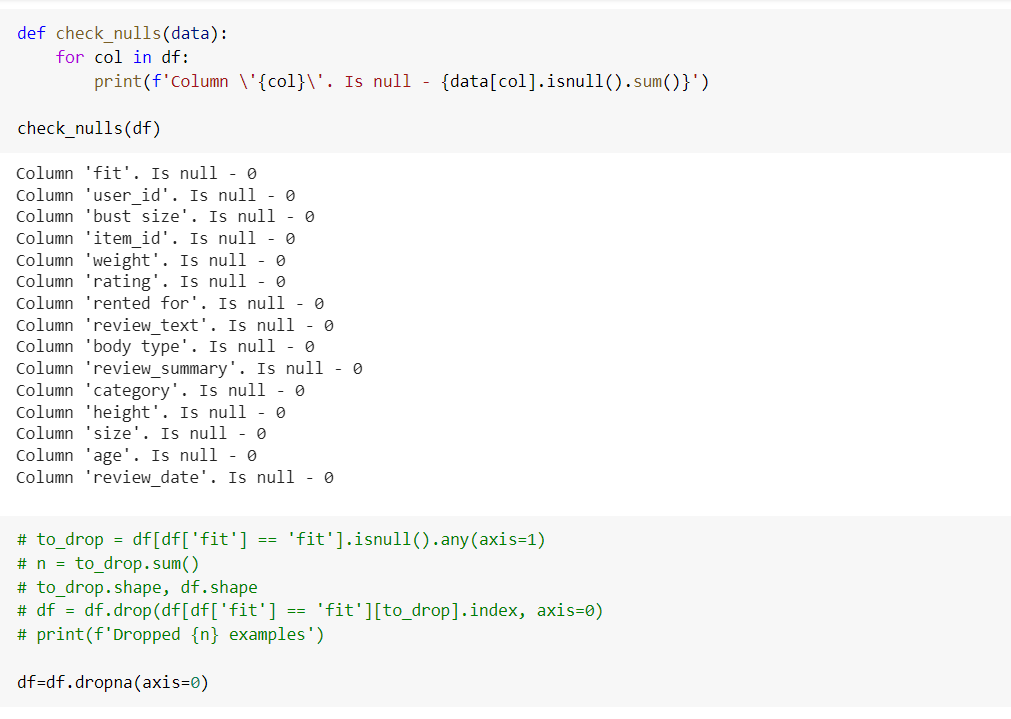
**Student Details**

| Roll No. | Name of the Student | Name of the Program |
| --- | --- | --- |
| AU2040122 | Aditi Vasa | Btech CSE |
| AU2040002 | Shrey Somani | Btech CSE |
| AU2040196 | Vandan Shah | Btech CSE |
| AU2040048 | Ronit Shah | Btech CSE |

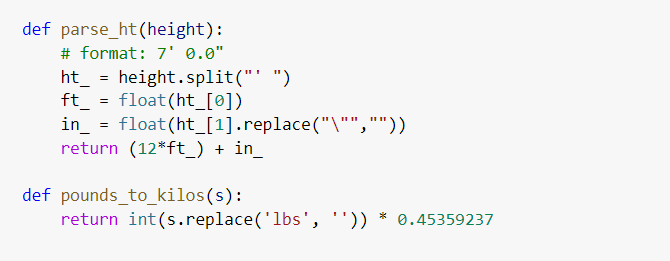
2022-2023 (Winter Semester)

**Report**

**Dropping all null values**

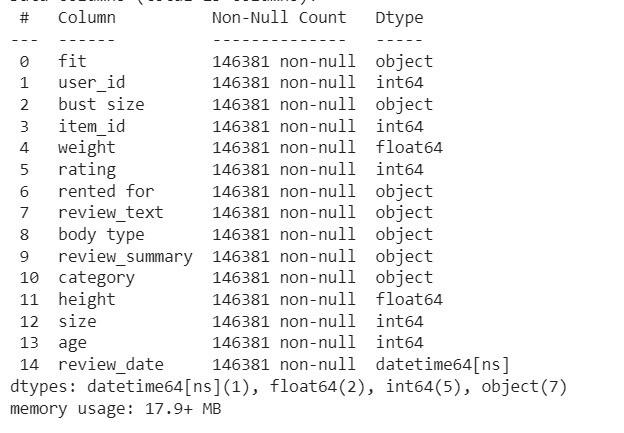
****

**Converting height into inches and weight into pounds**

****

**Changing data type of various parameters into their respective data type as all parameters were initially objects**

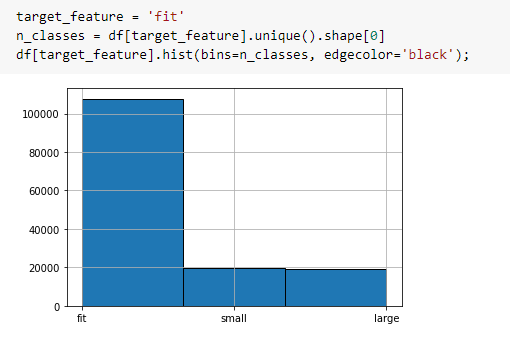




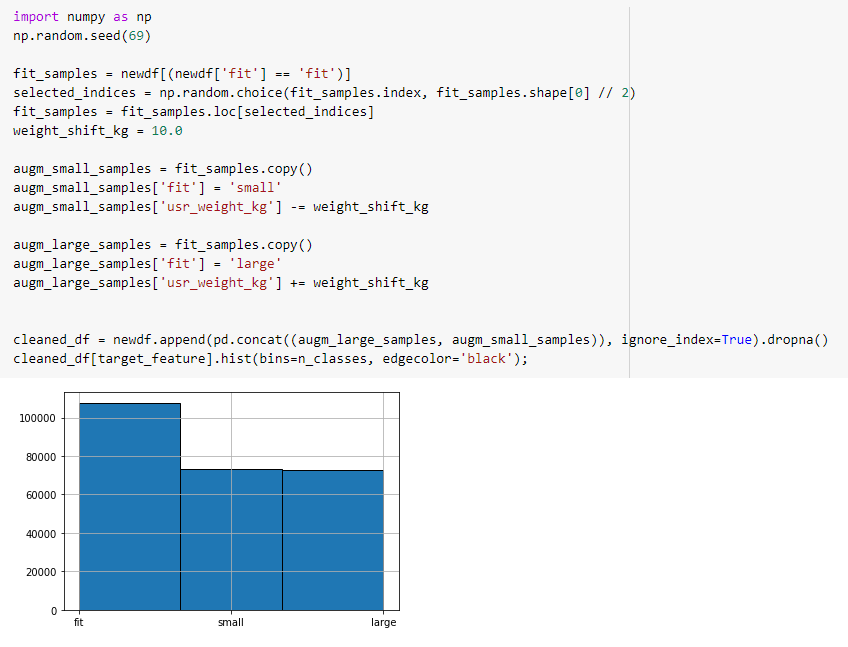
**Changing variable name for better reference**



**Data before data imputation**



**Data after imputation by considering shift in weight for +10 kg (large fit) or -10 kg (small fit)**



**SUMMARY**

After understanding the dataset, we moved on towards the data cleaning part. While looking through data, we found there were around 10,000 null values in our dataset. So, we decided to drop them. Next, we observed that all the parameters were of object data type and realized it would be difficult to perform EDA on them as it will be going to be our next step. This week, we created functions that would convert height into inches and weight into pounds. Next we tried to plot and see our target feature: FIT, we realized that we have a little biased data. After inferring with TA, we decided to perform data imputation. That was done by adding +- 10 kg weight to make fit as large fit or small fit respectively. This helped us to scale our data for the initial level.

Finally, since the target feature or the output was in the form of ordinal data, we did encoding and categorized “fit” as 0, “small” as -1 and “large” as 1.

**TASK IN UPCOMING WEEK:-**

* In the upcoming week we are plotting a heat map and will find correlation between each variable.
* We plan to do encoding on more variables
* We also plan to check if is there any better way to do data imputation of “fit”
* And also plotting box and whisker plot of each variable.