DIABETES

ABOUT THE DATASET:

This dataset contains:

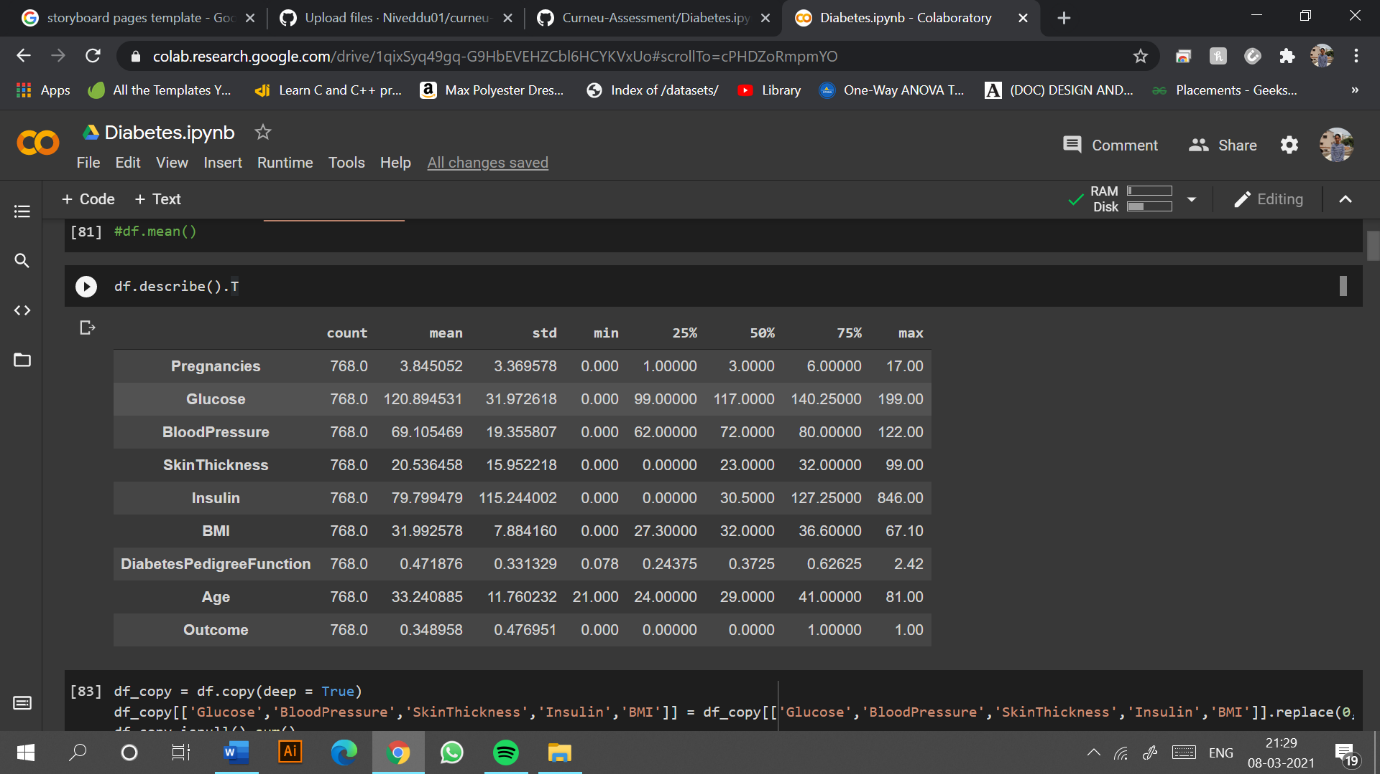
* Pregnancies
* Glucose
* Blood\_Pressure
* Skin\_Thickness
* Insulin
* BMI
* Diabetes\_Pedigree\_Function
* Age
* Outcome: If the above results in Diabetes or not for a person (The target class)

KNN function:

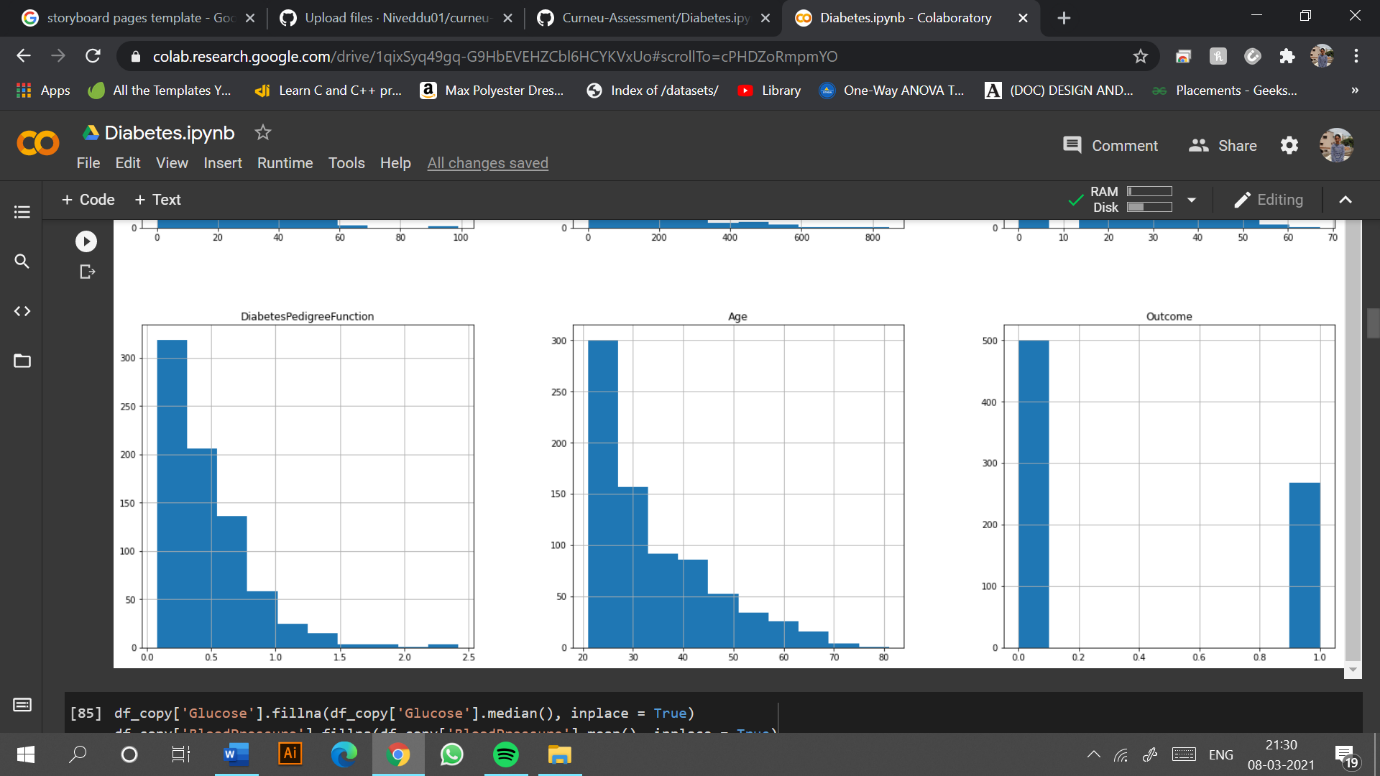
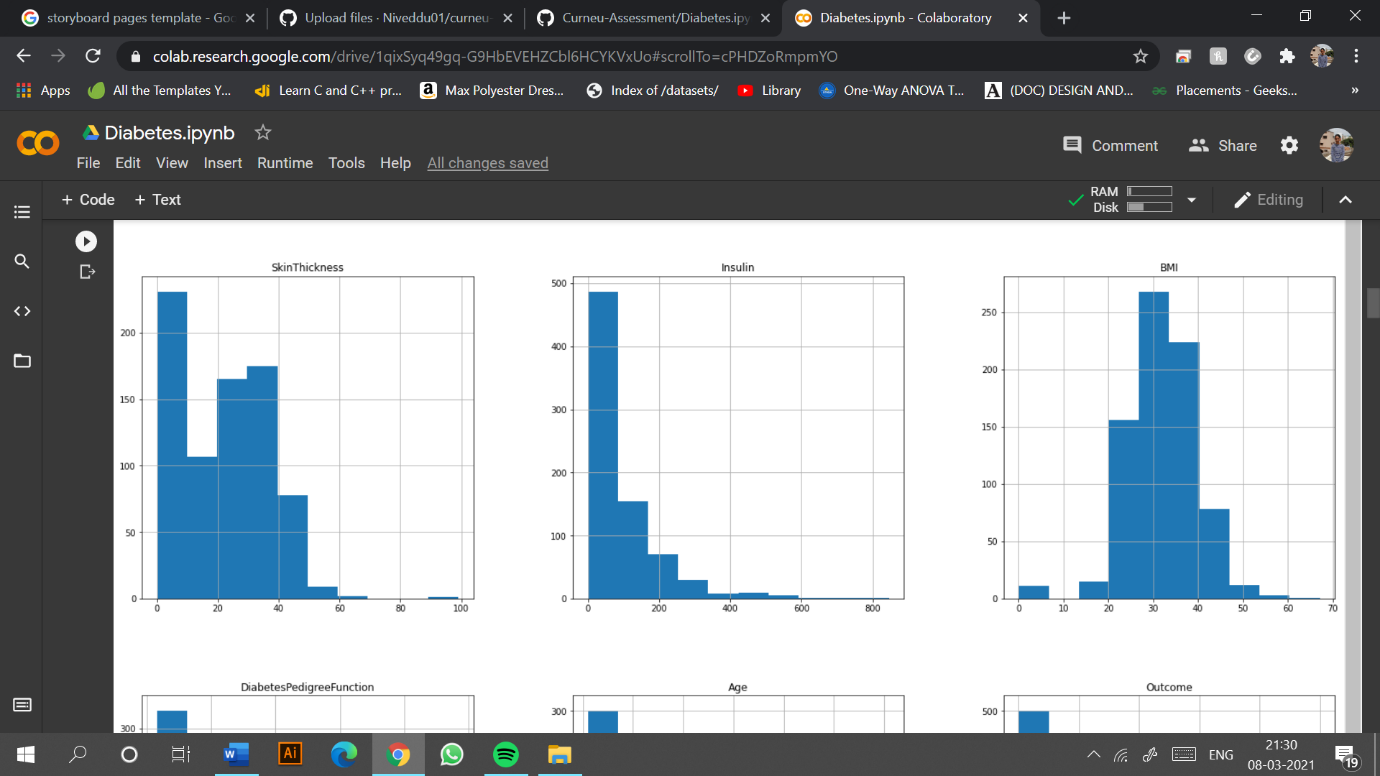
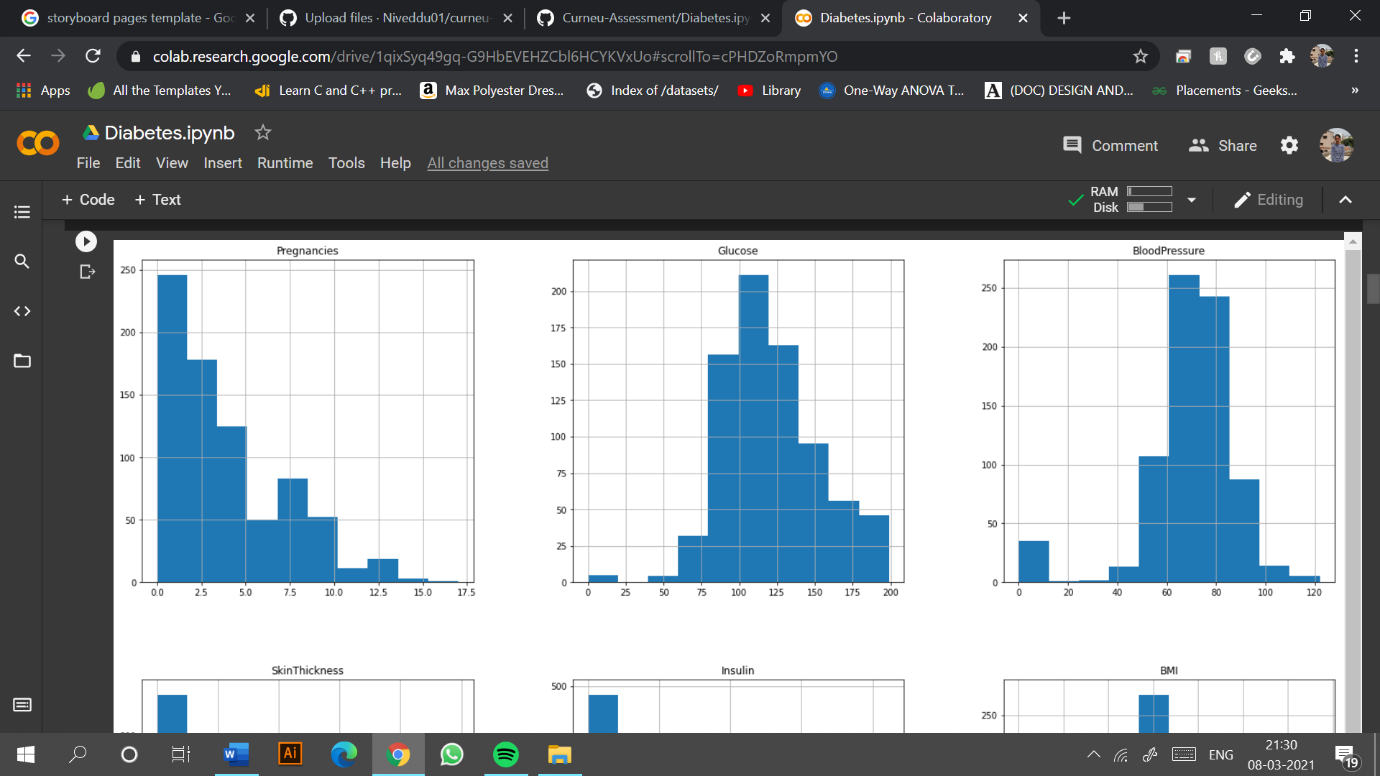
KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

EDA:

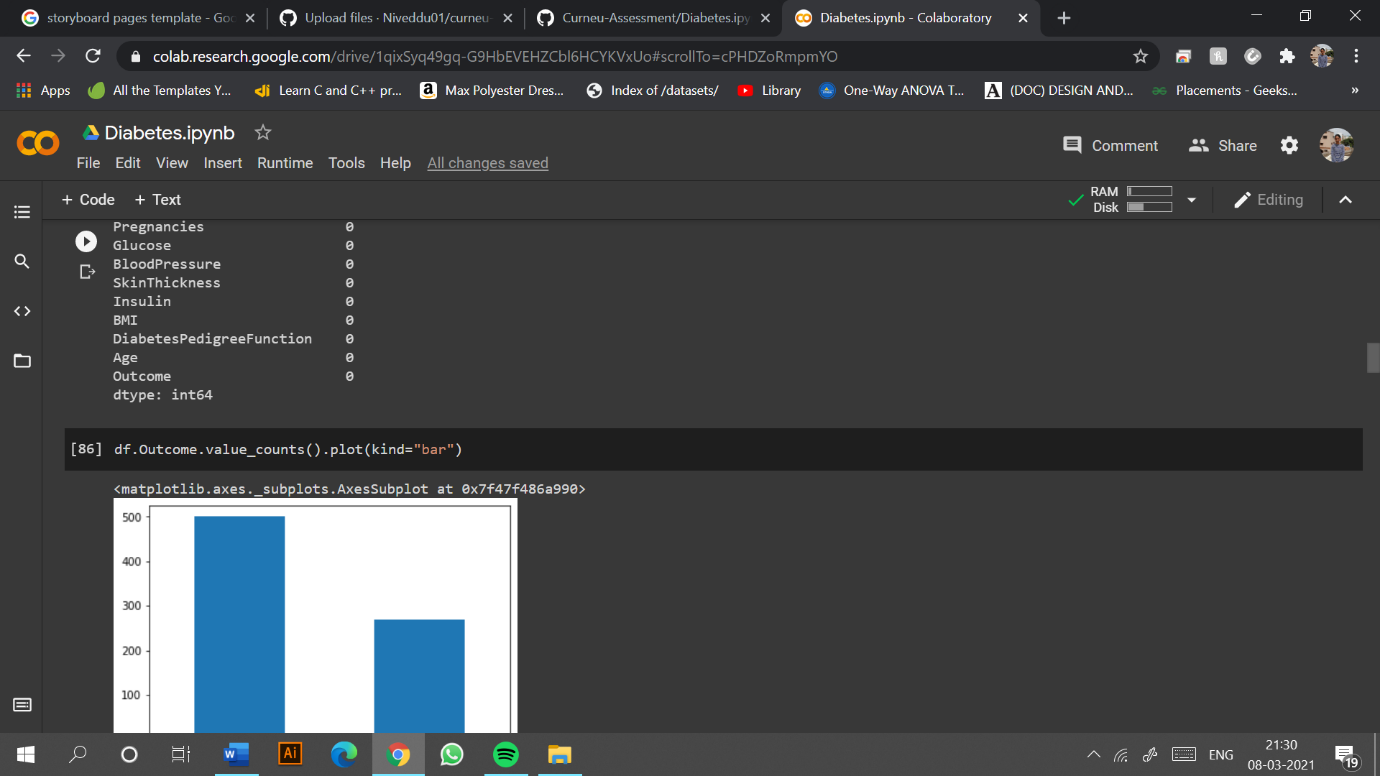
Given the data is in numerical format, no encodings are required. There are no irrelevant columns and hence no cleaning is also required. But there are some missing values and outliers that affect the accuracy of the model and must be rectified. After rectification, the basic analysis is done.

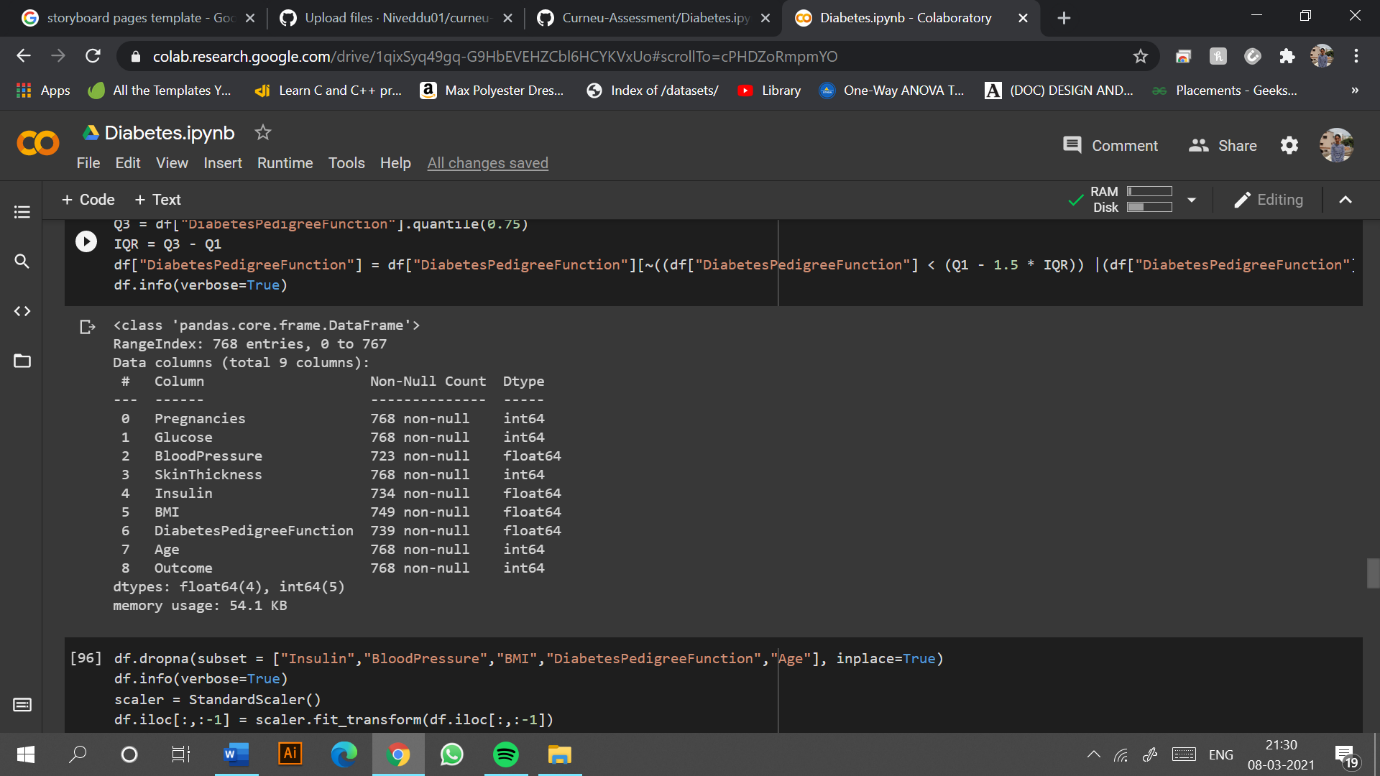


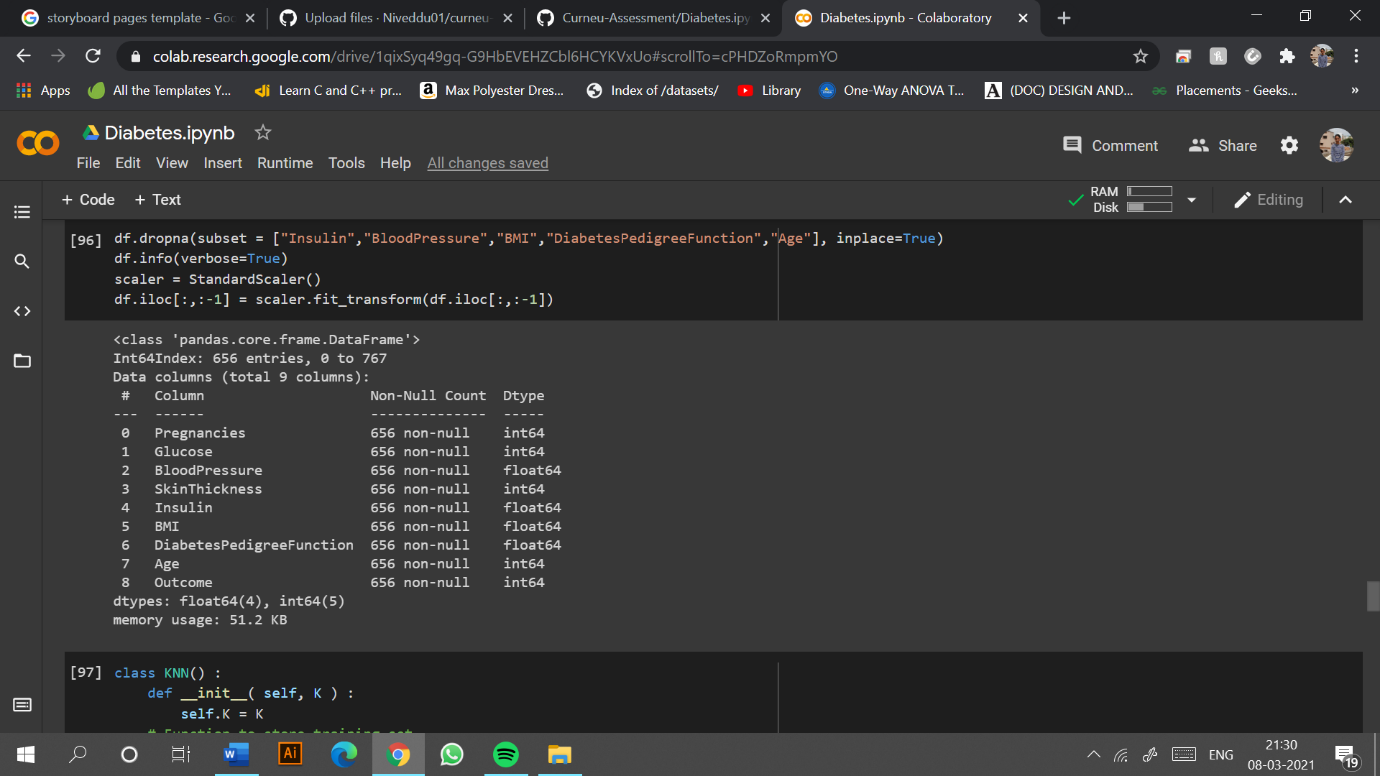
The Corresponding plots are made for each and every attribute in the dataset to display its range.



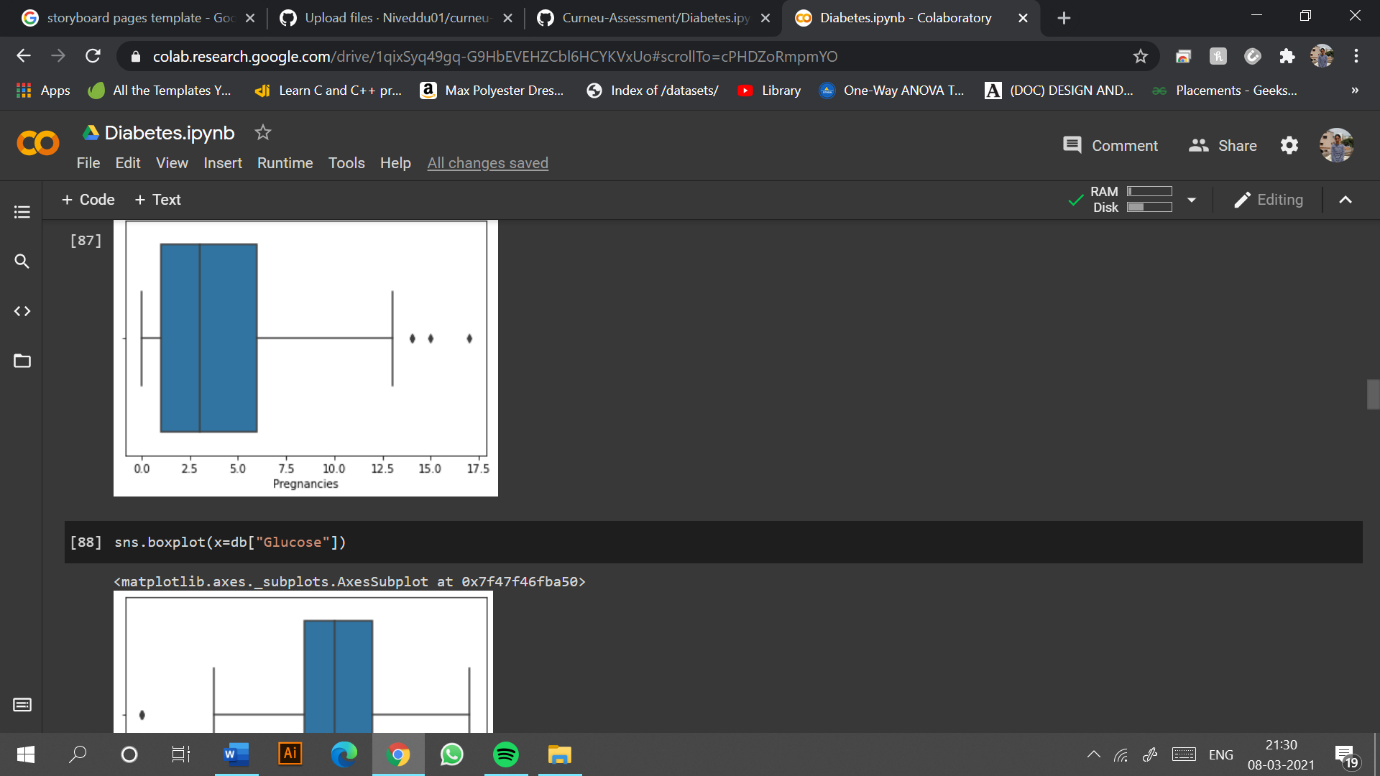
The values here show that there are no null values and hence the data is complete. The barplot below describes that the majority of women do not have any complaints of diabetes.







Further, individual boxplots are plotted to detect and show the outliers.



The entire dataset is split into train and test on one hand and a KNN function defined. The KNN function is again called into the main function resulting an accuracy score of 76% approximately, which is assumed to be very low.

