Hi this is Dinesh from Team 8 , we started cleaning the data by generating descriptive statics of mean ,SD, min max and quartile deviation of variables. While doing this we notices minimum value for many variables are 0 ,as biological parameters like glucose , BP, skin thickness , insulin and BMI cannot have zero values , we can infer that null values have been coded as 0

Checking for 0 count , we see Glucose , Blood Pressure and BMI have less than 5% 0 values , hence dropping them would not significantly impact the sample size.

However Insulin with 374 and Skin thickness 227 in a data set 768 is almost 50% , hence we need to impute them with suitable Mean or Median

We can see from the box plot for insulin against Class variable that there are many outliers which are skewing the mean by around 30 points and there is also a significant 70 point difference between the mean and median . we decided to replace the zero values of ND with median 102.5 and D with 169.5

In the Box plot for skin thickness and class variable we see only marginal difference between mean and median hence we decided to replace 0 with their respective Mean

This is the code to impute 0 with suitable Mean Median for insulin and skin thickness . we dropped the 0 value rows for low impact Glucose , BP and BMI.

We also renamed the long variable names with shorter Alias to help visualization through graphs

You can see we have been able to successfully impute the 0 values with only 6% sample size reduction.

Moving on to Data exploration we started with Univariate analysis . Pie Plot of resultant class variable shows that 65% of sample population are non diabetic

we plotted the Histogram and density graph for other variables.

Insulin and Pedigree are right skwed with long tail

Glucose and results are left skewed and rest of the variables are more or less centered

Interestingly no bi modal series were observed ,so there could be no proper demarcation between groups , let us further check using Bi Variate analysis

We used pair plot to compare inter relation between variables

The diagonal shows the main effect of the variable on itself in sub group of D /ND. For example the density plot of 2nd row glucose shows distinct peak point for D /ND , but there is also large over lap between the groups .

The interaction effect of glucose with other variable shows some distinct grouping in scatter plots. Let us further explore using pearson co relation coefficient and heat maps

If we look horizontally in the last low , we can see resultant variable is most co related to Glucose followed by others . Incidentally Glucose is also equally co related to insulin

Above is the ranking of resultant co relation , as seen from the heat map even though insulin is 2nd after glucose only one of this variable is recommended for use in model to avoid any inflation variation.

Let us further deep dive into Data Visualization and Analysis.