# Assignment 2

1. EDA
2. Summary statistics:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Mean Radius | 198 | 17.4121 | 3.1572 | 10.95 | 15.0525 | 17.28 | 19.58 | 27.22 |
| Mean Texture | 198 | 22.3190 | 4.2831 | 10.38 | 19.5175 | 21.795 | 24.655 | 39.28 |
| Mean Perimeter | 198 | 114.8566 | 21.3834 | 71.9 | 98.16 | 113.7 | 129.65 | 182.1 |
| Mean Area | 198 | 970.0409 | 352.1492 | 361.6 | 702.525 | 929.1 | 1193.5 | 2250.0 |
| Mean Smoothness | 198 | 0.1027 | 0.0125 | 0.07497 | 0.0939 | 0.1019 | 0.1110 | 0.1447 |
| Mean Compactness | 198 | 0.1426 | 0.0499 | 0.04605 | 0.1102 | 0.1318 | 0.1722 | 0.3114 |
| Mean Concavity | 198 | 0.1562 | 0.0706 | 0.02398 | 0.10685 | 0.1513 | 0.2005 | 0.4268 |
| Mean Concave Points | 198 | 0.0868 | 0.0339 | 0.02031 | 0.06367 | 0.0861 | 0.1039 | 0.2012 |

1. total rows = 198

unique values = ['N' 'R']

top value = N

top value frequency = 151

1. Yes, we can encode the categorical variable ‘outcome’ to numerical data type. As the column only has two unique values and we are performing binary classification using logistic regression, I am going to use ‘*label encoding*. For complex classification we can also use OHE (One-hot Encoding).
2. Yes, there are multiple redundant features in the dataset. Those are the features that have high correlation or same data. For example, in the given dataset, mean\_radius and mean\_perimeter have correlation of 1. There are also other features which have high correlation. Since the features having high correlation doesn’t affect the outcome of the data, removing one of the features having high correlation with the other won’t impact the analysis.

One interesting observation is that there are different variables that have correlation of 1 even though the data is different. There are also many variables with high correlations suggesting many redundant features in the data.

Another interesting observation is that based on the statistics computed among the given variables, mean\_radius, mean\_texture, mean\_perimeter, mean\_area, mean\_smoothness, mean\_compactness, mean\_concavity, and mean\_concave\_points, median is close to the mean but slightly less in all the variables suggesting slight right skewness.

1. correlation between ‘mean\_perimeter’ and ‘se\_perimeter’ = 0.6099643781634989