Statistics and Machine Learning 1

Section I – Introduction and Description of the Data

This report explores the PimaDiabetes dataset, obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, and reflects on the diagnostic measures of 750 women of Pima Indian heritage (Smith et al, 1988). Recognized for their heightened diabetes prevalence, this population provides a focused framework for assessing the impact of various health indicators on diabetes outcomes (Smith et al, 1988). The report encompasses an overview of the data's origin and quality issues, EDA, the relationship between the number of pregnancies and the outcome, and the development of a final predictive model for Diabetes in Pima Indian women. However, the dataset's specificity to the Pima Indian community limits its application to a wider population.

Section II – Exploratory Data Analysis

1 – Descriptive statistics

Figure 1 - Descriptive statistics

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome
count	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000
mean	3.844000	120.737333	68.982667	20.489333	80.37[illegible]	31.959067	0.473544	33.166667	0.346667
std	3.370085	32.019671	19.508814	15.918828	115.0[illegible]198	7.927399	0.332119	11.708872	0.476226
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	36.500000	32.000000	0.377000	29.000000	0.000000
75%	6.000000	140.750000	80.000000	32.000000	129.750000	36.575000	0.628500	40.750000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2 - Missing Values

```
Pregnancies           0
Glucose               5
BloodPressure        35
SkinThickness       221
Insulin             362
BMI                  11
DiabetesPedigree      0
Age                   0
Outcome               0
dtype: int64
```

There are some 0 values in insulin, SkinThickness, BMI, and Glucose, which isn't possible. Counting them tells us about 48% of the values in the insulin column are missing, and 29% in the SkinThickness column.

2 – Distribution of the Data

Figure 3 – Histogram of PimaDiabetes Feature Distribution

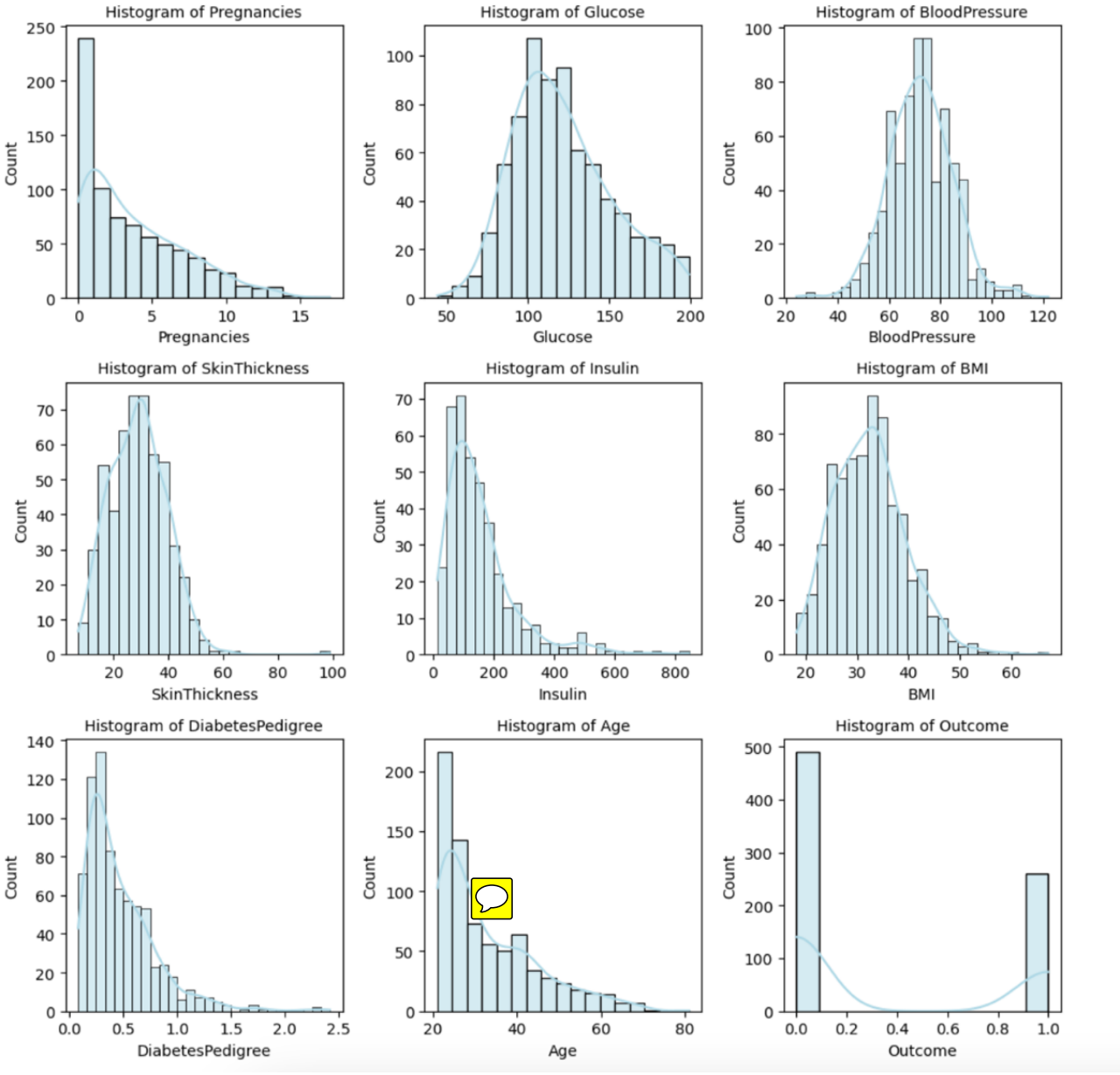
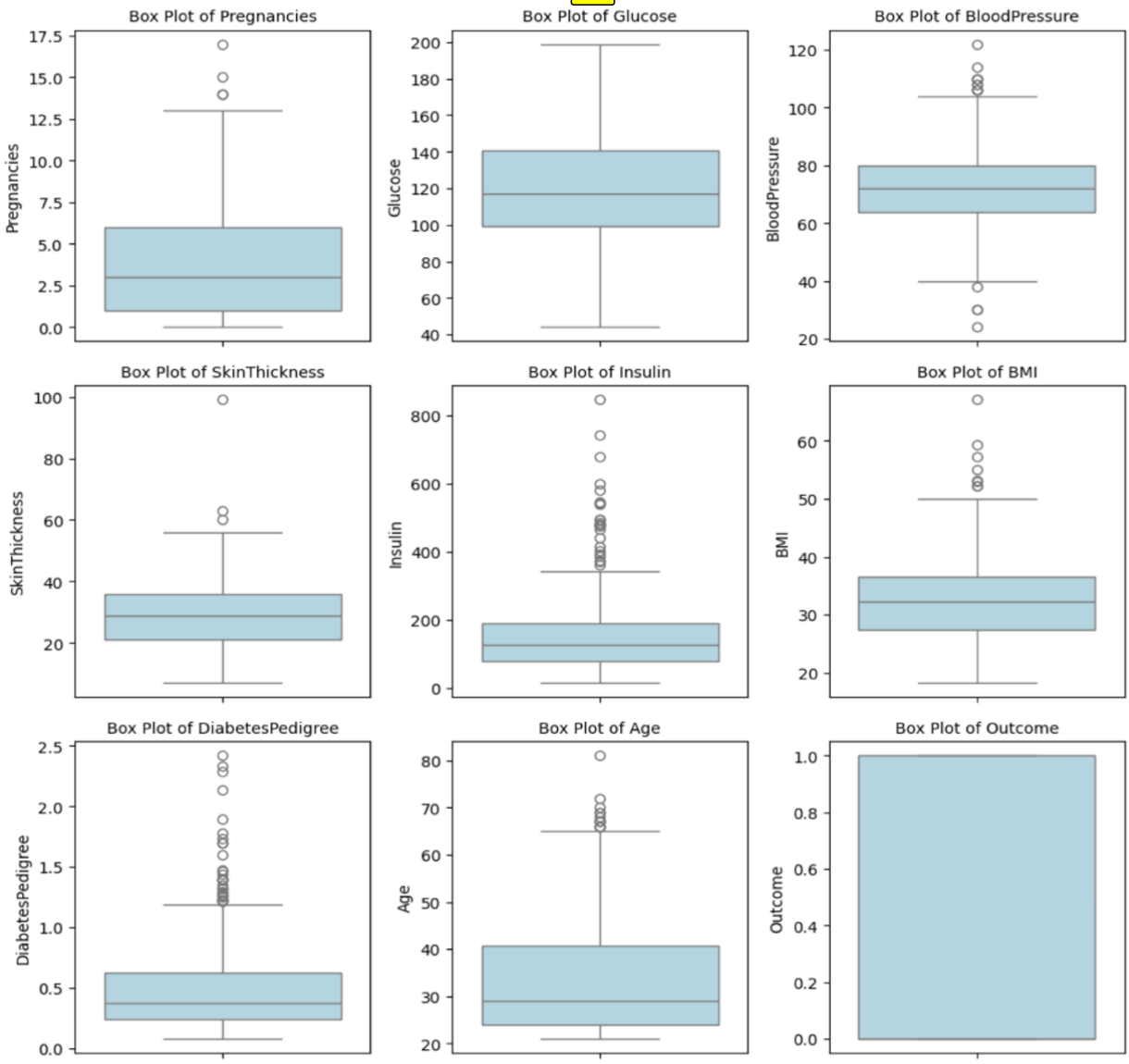


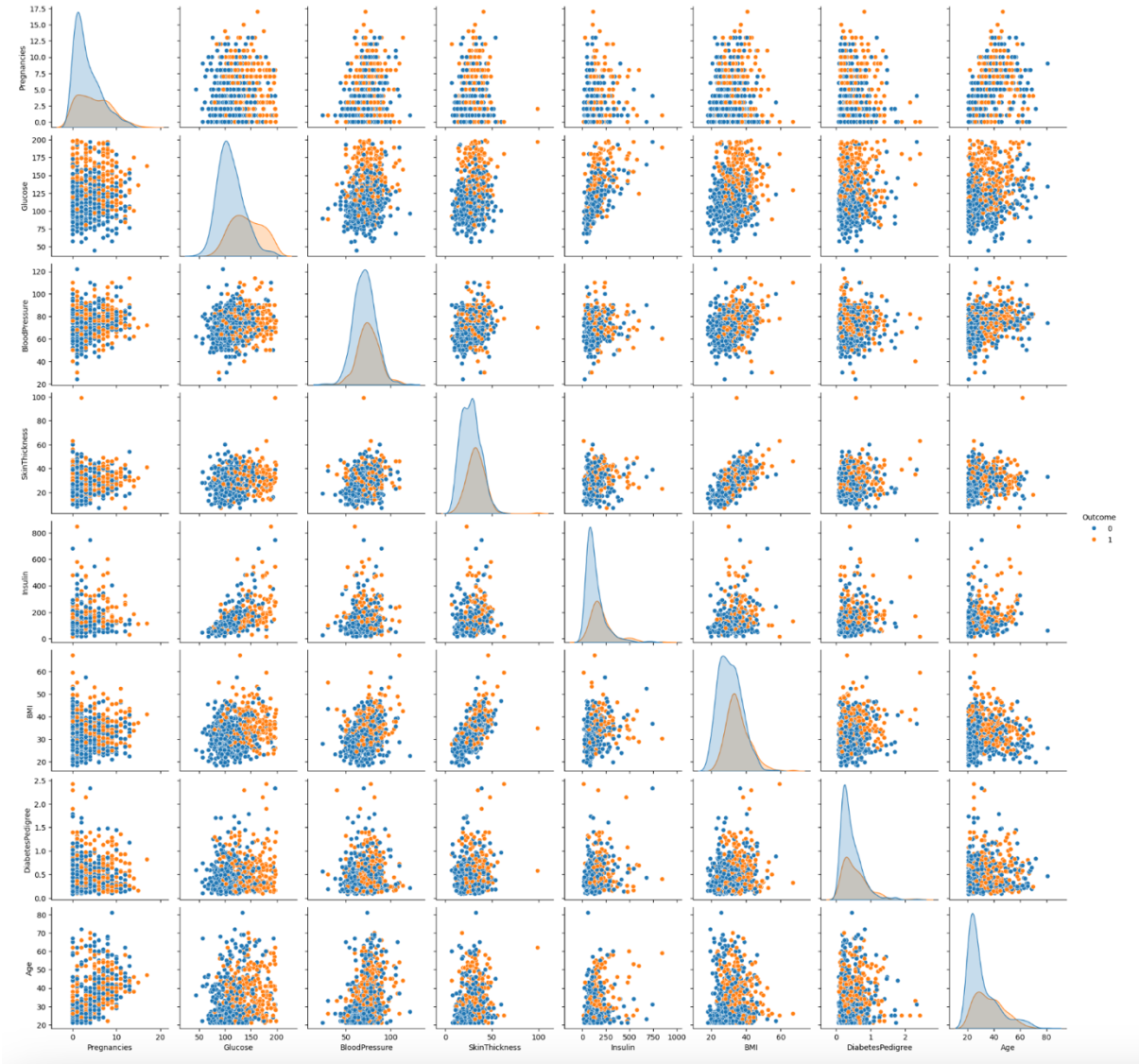
Figure 4 - Box Plot of PimaDiabetes Feature Distribution



There is skewness and outliers in variables such as Insulin and Skin Thickness, with Age and Pregnancy counts leaning towards younger individuals and fewer pregnancies.

3 – Relationships between variables

Figure 5 - Pairplot of PimaDiabetes Feature Relationships



The pair plots reveal that diabetic women score higher across most predictors.

4 – Grouping by Outcome

Figure 5 – Line plot of Mean Feature Values per Outcome

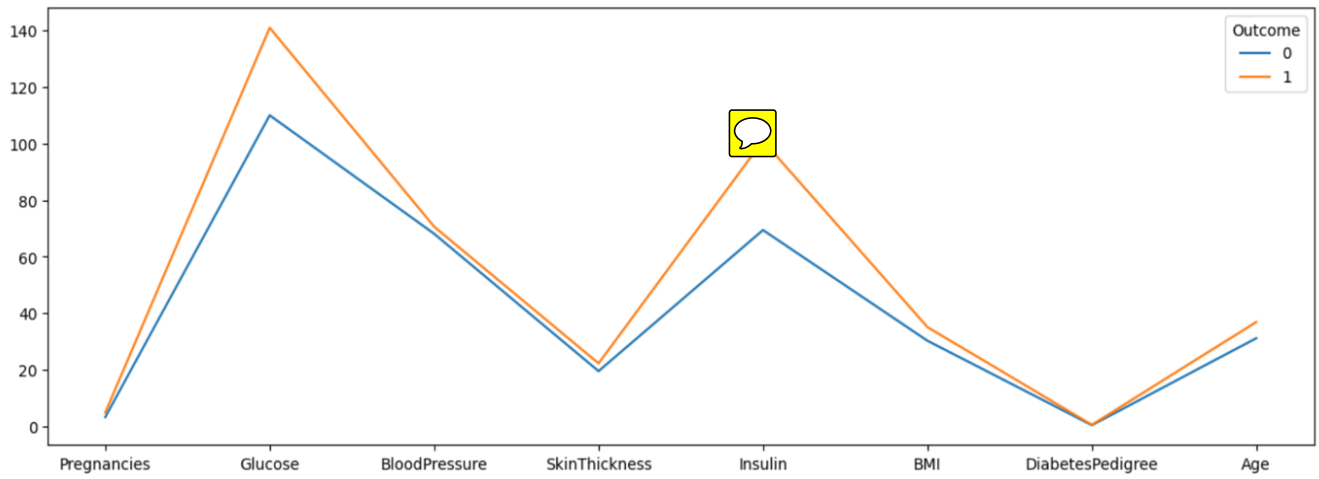


Figure 6 - Distribution of Outcome

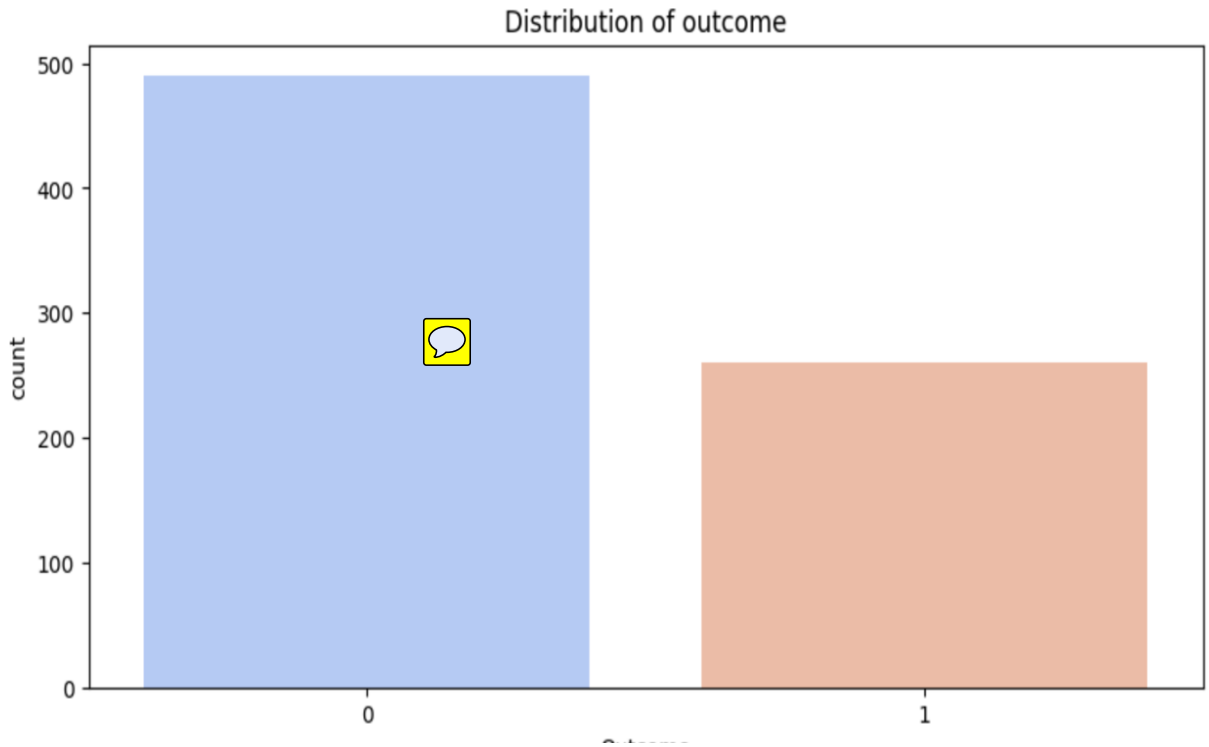
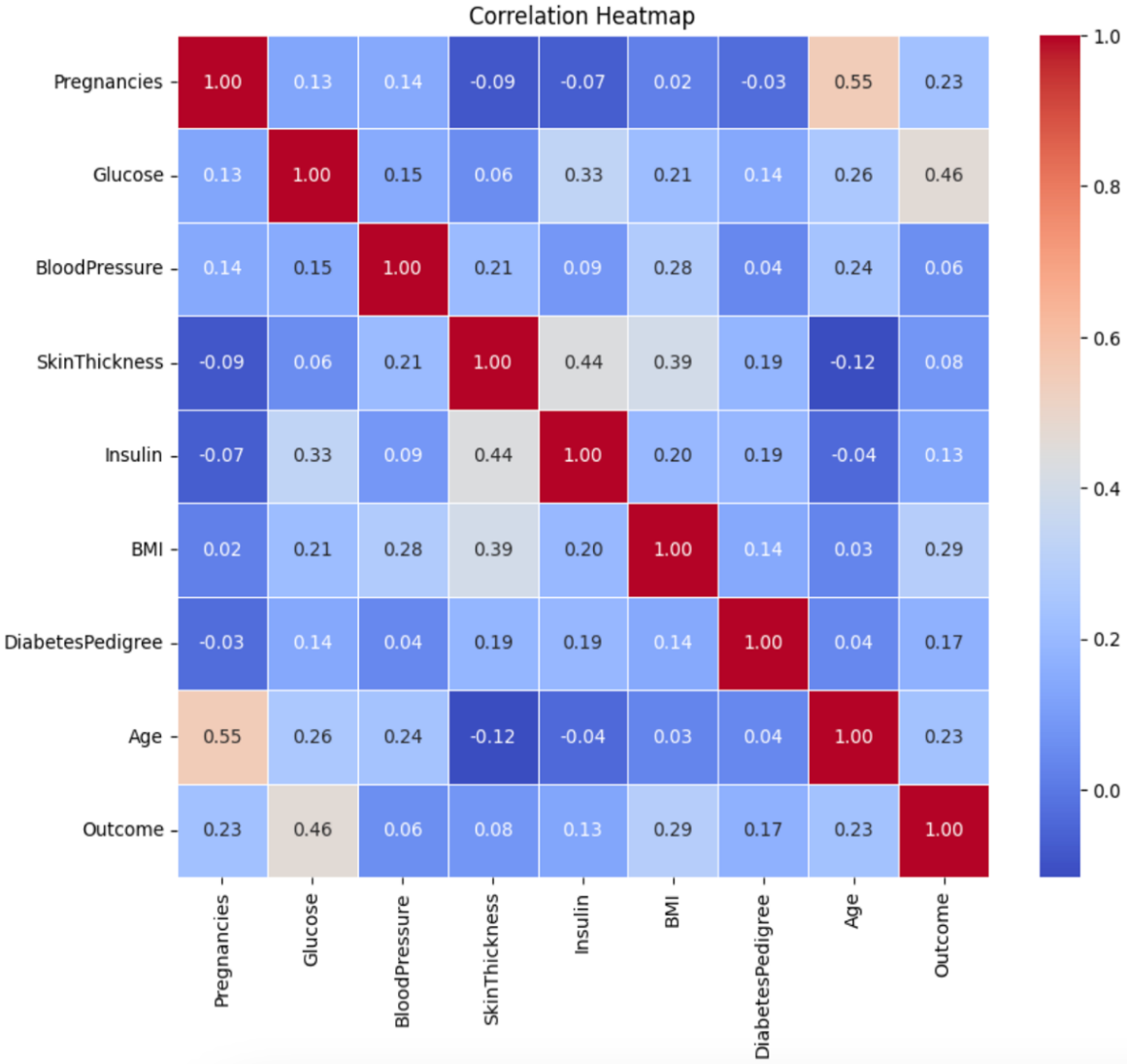


Figure 6 highlights an instance of class imbalance.

5 – Correlation analysis

Figure 7 - Correlation Matrix of PimaDiabetes Features

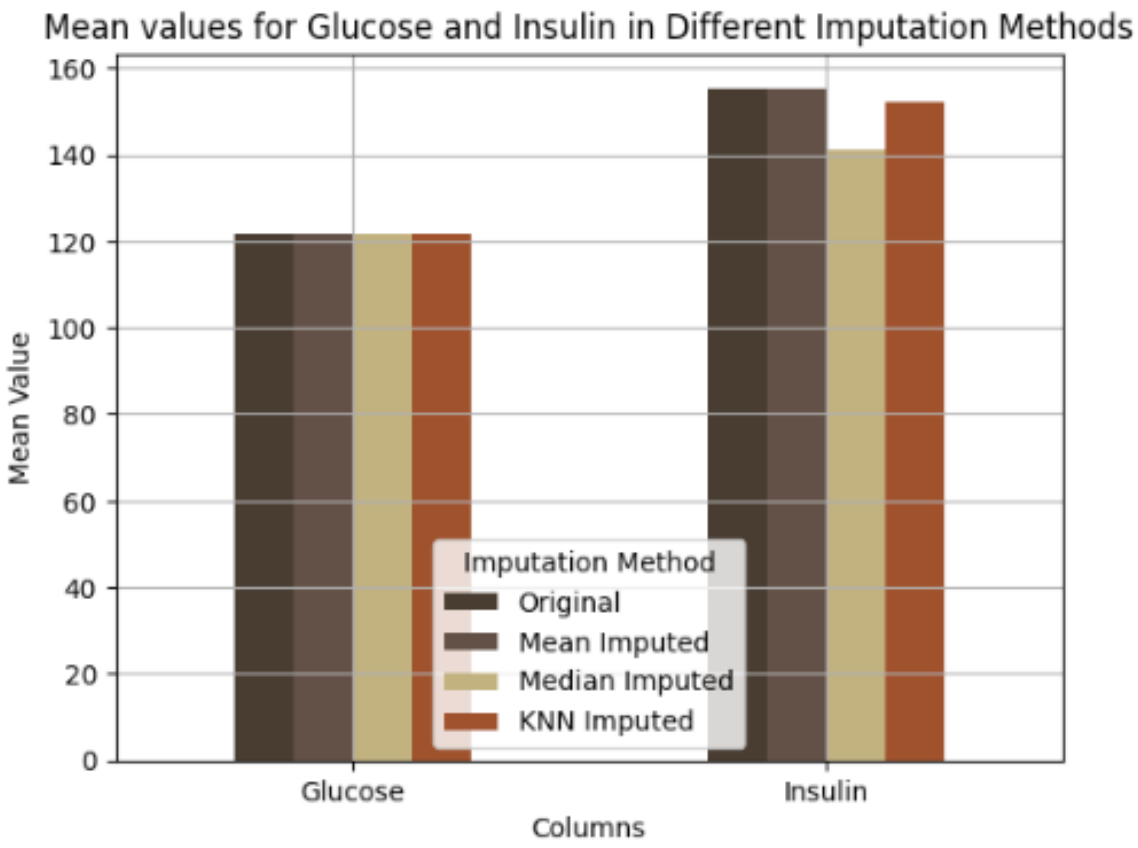


There is correlation between the number of pregnancies and Age as expected. But Glucose levels are the most significant predictor of diabetes, followed by Age and Insulin.

6 – Computing the missing values

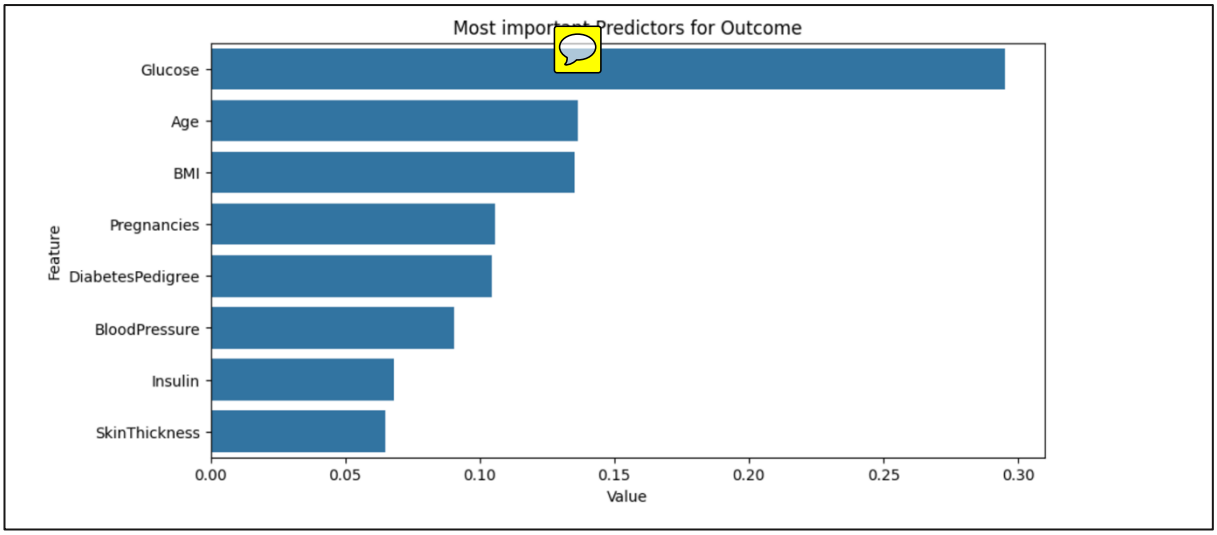
Addressing missing values, especially in key variables like Glucose and Insulin, was essential as there were lots. Mean and median fills seemed a bit simplistic may have missed crucial inter-variable relationships, especially with so much data missing. K-Nearest Neighbours (KNN) imputation preserved data structure and relationships, enhancing model accuracy by using physiologically correlated data points to get us the most accurate imputation. As this is a small dataset, the number of neighbours is set at 25 and managed to maintain computational efficiency.

Figure 8 - Bar Chart of mean Glucose and Insulin after Imputation



7 – Finding the strongest predictors of outcome

Figure 9 - Bar plot of most important predictors of Outcome

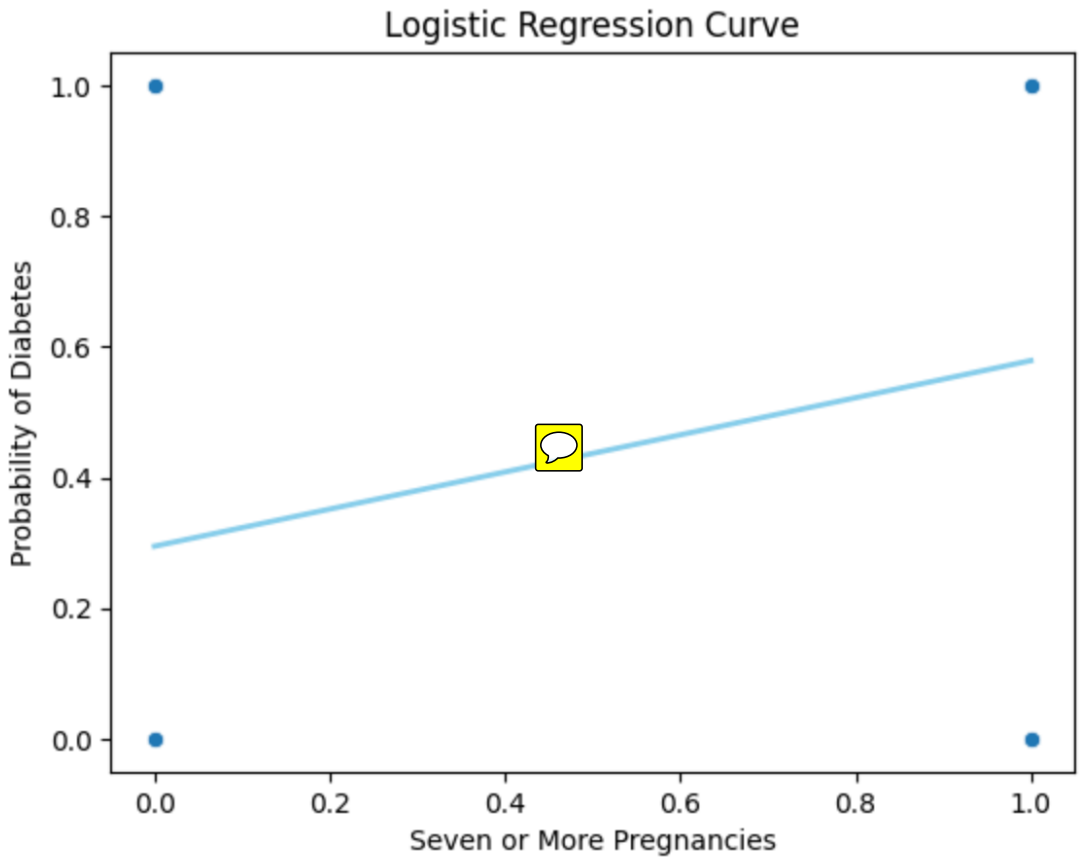


This shows Glucose is the strongest predictor of the outcome.

Section III – Influence of Pregnancies on Outcome

We added a column "SevenOrMorePregnancies" to indicate if a woman had 7+ pregnancies (1) or fewer (0) and fitted a logistic regression model where this new variable was a single predictor for the outcome. Women with six or fewer pregnancies have a 28% chance of developing diabetes, which rises to % for those with seven or more pregnancies.

Figure 10 - Effect of SevenOrMorePregnancies outcome



Section IV – Predictive Model

The aim of this section is to build a robust predictive model for diabetes amongst the women in the ToPredict dataset. Logistic Regression seems like the most suitable choice as it is applicable to binary outcomes such as diabetes, and it is also quite interpretable. The model of choice, Model 5, introduces an interaction term between Age and SevenOrMorePregnancies. This decision is the result of our EDA insights, revealing a nuanced relationship between these two predictors and the risk of diabetes. Most importantly, this model strikes a balance between precision and recall, which is vital for medical predictions where both false negatives and false positives carry significant implications. The performance of all the models was evaluated with Scikit-Learn's classification report.

Model 5's precision for positive diabetes detection, an important metric given the significant implications of undiagnosed diabetes stands at 65% with a recall of 67%. During the modelling, there was very persistent precision-recall trade-off. Model 5 optimises for this better than its predecessors, However, there is a ceiling effect observed in model 4 at 69% precision for a positive outcome, and efforts to exceed this severely compromised other important metrics.

The process involved iterative modifications to models, each informed by the preceding model's performance. Earlier models varied in handling class imbalance and feature selection. The introduction of polynomial features in Model 4, for instance, aimed to capture more complex relationships but resulted in a trade-off between precision and recall. Model 5, however, achieved a more harmonized predictive ability without overfitting, a testament to the efficacy of including biologically plausible interaction terms.

Some interesting findings include that, Model 1, built on the findings of EDA and aimed to account for class imbalance (more negatives than positives in the dataset), however it only achieved a 57% accuracy for class 1 and 83 for 0. Model 2 dropped diabetes pedigree and didn't account for class imbalance, improving class 1 precision to 62%, which is a very interesting result. With model 4, we took a different a approach using polynomial regression. This achieved 69% precision for class 1 at the expense of recall.

Figure 11 - Model Performance on our metrics of interest

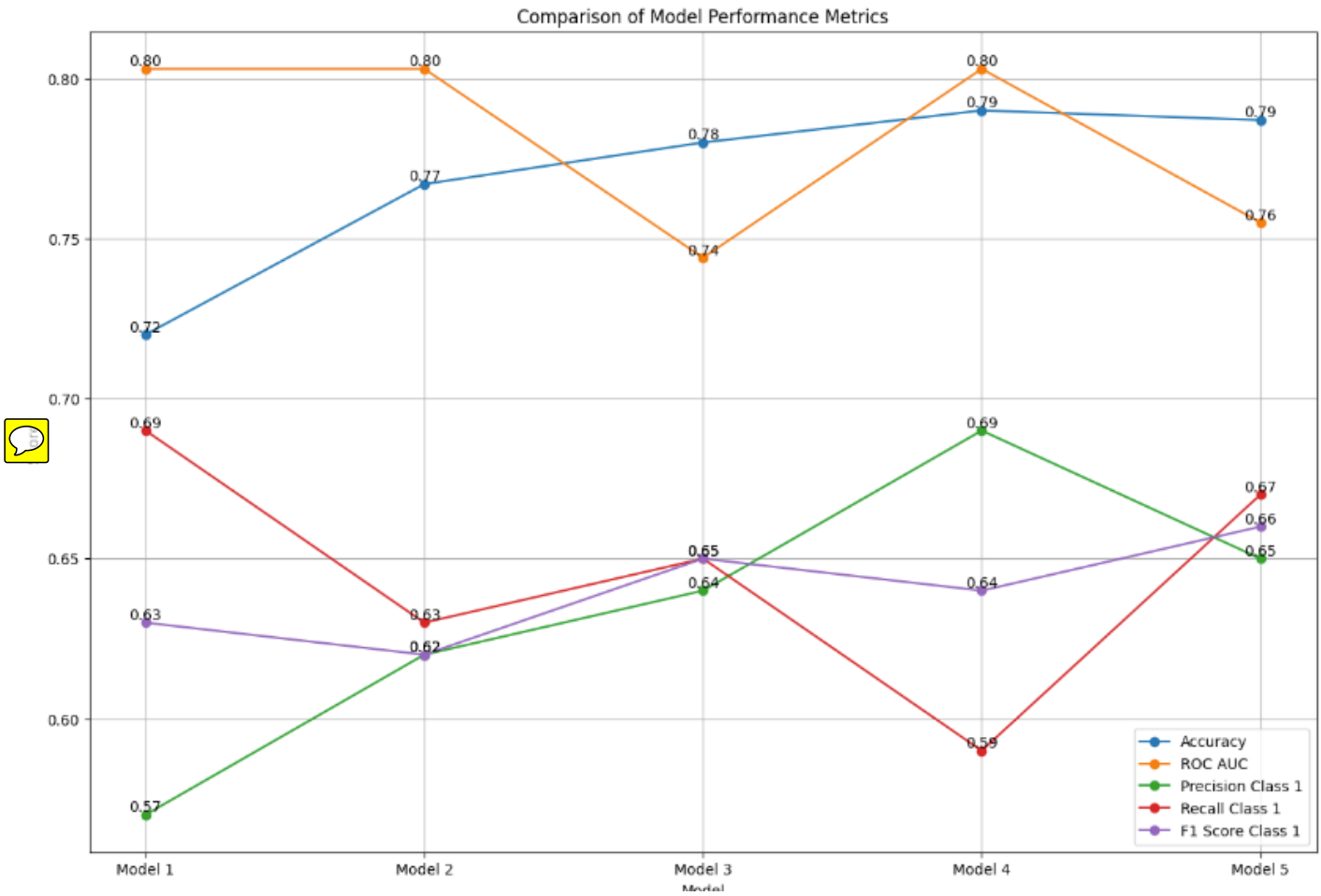


Figure 12 - Overall Performance Metrics of models 1 to 5

	Model	Accuracy	Precision (Class 0)	Recall (Class 0)	F1 Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1 Score (Class 1)	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1 Score
0	Model 1	0.77	0.83	0.74	0.78	0.57	0.69	0.63	0.74	0.72	0.73
1	Model 2	0.77	0.83	0.83	0.83	0.62	0.63	0.62	0.77	0.77	0.77
2	Model 3	0.78	0.84	0.84	0.84	0.64	0.65	0.65	0.78	0.78	0.78
3	Model 4	0.79	0.83	0.88	0.86	0.69	0.59	0.64	0.79	0.79	0.79
4	Model 5	0.79	0.85	0.84	0.84	0.65	0.67	0.66	0.79	0.79	0.79

Applying Model 5

There were several pre-processing steps to apply to the ToPredict dataset. There were also missing values in columns SkinThickness and Insulin, which are biologically impossible. As a result, these 0 values have been changed to NaN so that the KNN Imputer trained earlier in the assesmeent could be applied for optimal accuracy. During data processing, several errors revealed that we had to play around with the ToPredict dataset to run the model on it. For example, we had to add a dummy column "Outcome" with zero values so that the model would see this column and run, as it wouldn't otherwise.

Figure 13 - Predicted Probability of Diabetes

	Pregnancies	Glucose	BloodPressure	BMI	DiabetesPedigree	Age	Predicted Probability of Diabetes
0	4.0	136.0	70.0	31.2	1.182	22.0	0.479348
1	1.0	121.0	78.0	39.0	0.261	28.0	0.367996
2	3.0	108.0	62.0	26.0	0.223	25.0	0.107717
3	0.0	181.0	88.0	43.3	0.222	26.0	0.881452
4	8.0	154.0	78.0	32.4	0.443	45.0	0.762077

This assessment has demonstrated the practical challenges and considerations in predictive modelling for medical data. Iterative model refinement highlighted a precision-recall trade-off, with a notable ceiling effect at 69% precision for positive diabetes detection. Another approach to generate of additional data that mimics the intricate the relationships between the original variables. This would enable analysis using more advanced methods such as neural networks.

BIBLIOGRAPHY:

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O'reilly, p.277.

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O'reilly, p.92.

Smith, J., Everhart, J., Dicksont, W., Knowler, W. and Johannes, R. (1988). *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf.

EDA

1. Descriptive Statistics

```
> summary(Diabetes)
  Pregnancies      Glucose     BloodPressure   SkinThickness     Insulin        BMI           DiabetesPedigreeFunction      Age          Outcome
 Min.   : 0.000   Min.   :  0.000   Min.   :  0.000   Min.   : 0.000   Min.   :  0.000   Min.   : 0.000   Min.   : 0.078   Min.   : 21.00   Min.   :0.000
 1st Qu.: 1.000   1st Qu.: 99.000   1st Qu.: 62.000   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:27.375   1st Qu.: 0.244   1st Qu.: 24.00   1st Qu.:0.000
 Median : 3.000   Median :117.000   Median : 72.000   Median : 23.000   Median : 30.500   Median :32.000   Median : 0.372   Median : 29.00   Median :0.000
 Mean   : 3.845   Mean   :120.895   Mean   : 69.105   Mean   : 20.536   Mean   : 79.799   Mean   :31.992   Mean   : 0.472   Mean   : 33.241   Mean   :0.349
 3rd Qu.: 6.000   3rd Qu.:140.250   3rd Qu.: 80.000   3rd Qu.: 32.000   3rd Qu.:127.250   3rd Qu.:36.600   3rd Qu.: 0.626   3rd Qu.: 41.000   3rd Qu.:1.000
 Max.   :17.000   Max.   :199.000   Max.   :122.000   Max.   : 99.000   Max.   :846.000   Max.   :67.100   Max.   : 2.420   Max.   : 81.000   Max.   :1.000
> sd(Diabetes$Glucose)
[1] 31.97262
> sd(Diabetes$Age)
[1] 11.76023
> sd(Diabetes$BloodPressure)
[1] 19.355807
> sd(Diabetes$Insulin)
[1] 115.244002
> sd(Diabetes$BMI)
[1] 7.884160
> sd(Diabetes$Pregnancies)
[1] 3.369578
> sd(Diabetes$DiabetesPedigreeFunction)
[1] 0.331329
> sd(Diabetes$SkinThickness)
[1] 15.958863
> sd(Diabetes$Outcome)
[1] 0.476951
> table(Diabetes$Outcome)

  0   1
500 268
> table(Diabetes$Pregnancies)

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  17
111 135 103 75  68  80  72  43  52  38  24  11  9   10  1   1   1
> table(Diabetes$Age)

21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 65 66 67 69 70 72 81
63 72 62 48 35 32 36 39 29 21 24 19 20 16 22 18 19 11 13 14 15 10 9  8  13 12 9  9  8  8  7  6  5  5  4  6  3  3  2  2  2  1  1  1  1  1  1  1  1  1
> table(Diabetes$Outcome)

  0   1
500 268
> table(Diabetes$Glucose)

  0   44  56  57  61  62  65  67  68  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 193 194 195 196 197 198 199
  5   1   1   1   1   2   1   1   1   2   4   1   1   2   1   1   1   1   2   1   1   1   2   2   3   1   1   2   2   1   1   1   1
```