Exploring the Pima Indians Diabetes Dataset: A Statistical Analysis and Visualization Study

Mohammed Arslaan Kola

Table of Contents

Introduction

Diabetes is a chronic disease that causes a person's blood sugar level to become abnormally high and is a lifelong condition (NHS, 2019). Diabetes affects approximately 422 million people worldwide, and diabetes is directly responsible for 1.5 million deaths each year. Diabetes has been steadily increasing in both the number of cases and the prevalence over the last few decades (Diabetes, 2022). The report statistically analyses relationships between clinical factors and diabetes using the dataset provided.

About the Data

The dataset provides insights into the number of Pima Indian women aged above 21 affected by diabetes. It comprises data of 768 females, out of whom 268 have been diagnosed with diabetes. Eight variables are included in different columns, namely, the Number of Pregnancies, Glucose levels, Insulin levels, Age, BMI, Skin thickness, Diabetes Pedigree Function and Blood Pressure. A binary classifier, "Outcome" serves as the dataset's response variable and shows whether the person has been diagnosed with diabetes or not. The value of 1 indicates that the person has diabetes, while a value of 0 indicates otherwise. A preview of data (First 6 rows including headers) is shown in Table 1.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Table 1: Diabetes Dataset

Problems with Missing Data in the Dataset

"There are missing entries throughout the dataset. The dataset shows a 0 where that entry is missing. Describe why this is problematic for this dataset."

Using 0 as a placeholder for missing data in the dataset can be problematic for several reasons, including:

- **It can affect the accuracy of the analysis**: By using 0 to represent missing data, it will be assumed that the missing data has a value of 0. This can lead to inaccurate analysis because it distorts the actual values of the data (Tay, 2021). For instance, if there are missing values in a column representing Blood Pressure, and these are replaced with 0s, it will appear as if the Blood Pressure value for the entry is 0, which can skew overall Blood Pressure figures.

- **It can impact statistical analysis**: Using 0 as a placeholder for missing data can affect statistical analysis such as mean, standard deviation, and correlation calculations. Since the missing data is replaced with 0, it can impact the distribution of the data, leading to incorrect results. For example, if the missing data is replaced with 0s in a column representing Glucose, it will artificially lower the average Glucose levels of the dataset.

- It can lead to incorrect conclusions: If 0 is used to replace missing data, it can result in incorrect conclusions being drawn. For example, if there are missing data entries in a column representing the BMI of the individual and these are replaced with 0s, it may appear that the individuals have a BMI of 0, which is impossible. This may lead to incorrect conclusions being drawn, such as underestimating the prevalence of obesity in a population, which can have negative consequences.

To avoid these issues, it is important to use appropriate techniques for handling missing data. These strategies can include the elimination of missing data, imputing values based on other data, or estimating missing values statistically.

Data Cleaning and Preparation for Statistical Analysis

"Proceed with the assumption that there is no missing data in the Outcome variable. Discard the Pregnancies variable. Clean the data so that missing entries in other variables become NaN values (this will allow you to perform statistical tests without issues)."

To discard the Pregnancies Variable from the dataset following R-Script is used:

```
diabetesDataset<- read.csv("diabetes.csv")  #Load the dataset
diabetesDataset$Pregnancies<- NULL  #Remove the Pregnancies Variable
```

To replace the missing values with NaN following R-Script is used, considering the dataset is already loaded in "diabetesDataset":

```
# Select all the columns other than "Outcome" and replace 0 with NaA
diabetesDataset[ ,names(diabetesDataset) != "Outcome"][diabetesDataset[
,names(diabetesDataset) != "Outcome"] == 0] <- NA
# Save in the new file named "cleaned_dataset"
write.csv(diabetesDataset, "cleaned_dataset.csv", row.names = FALSE)
```

The code loads the dataset from the "diabetes.csv" file into R using the "read.csv()" function. The column consisting of details regarding pregnancy is dropped from the dataset and the 0 values from all the columns other than the Outcome column are changed to NaN. Table 2 shows the preview the first 6 rows of the updated dataset stored in "cleaned_dataset.csv" file.

Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
148	72	35	NA	33.6	0.627	50	1
85	66	29	NA	26.6	0.351	31	0
183	64	NA	NA	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1

Table 2: Updated Diabetes Dataset

VISUALIZING AND SUMMARIZING THE DISTRIBUTIONS OF THE VARIABLES

“Visualise the distributions of each variable and provide their summary statistics.”

For statistically analysis, visualizing the distribution of each variable and calculating their summary statistics is an important step to understand the dataset. By doing this, we can get a sense of the data's shape, spread, and possible outliers. The “ggplot()” and “theme_minimal()” function is used in the script to display the plots with no background annotations (Heiss, 2021). Running this script in R provides series of plots as depicted below.

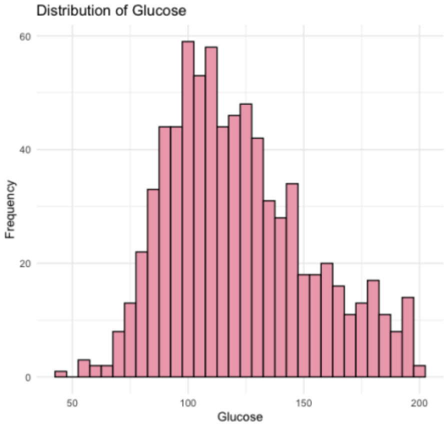


Figure 1: Plot for Distribution of Glucose

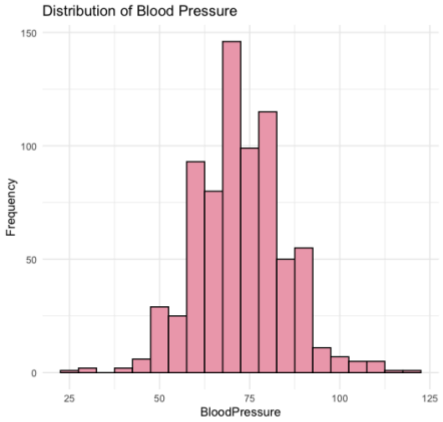


Figure 2: Plot for Distribution of Blood Pressure

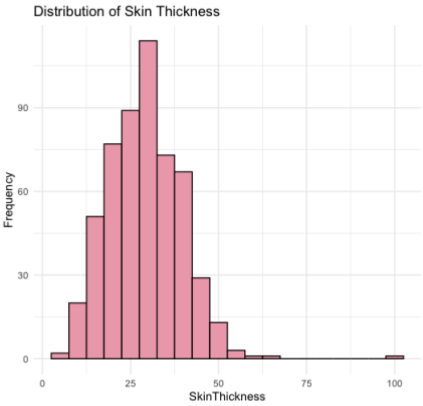


Figure 3: Plot for Distribution of Skin Thickness

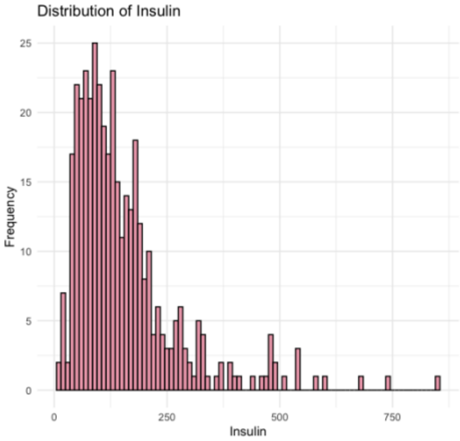


Figure 4: Plot for Distribution of Insulin

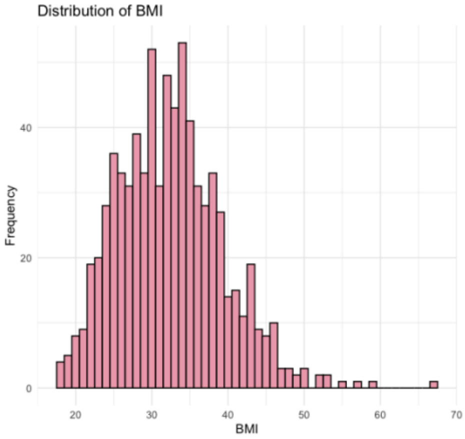


Figure 5: Plot for Distribution of BMI

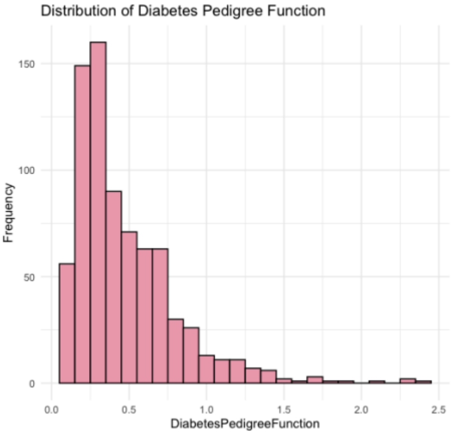


Figure 6: Plot for Distribution of Diabetes Pedigree Function

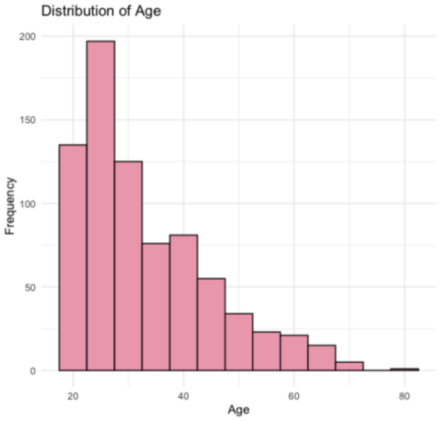


Figure 7: Plot for Distribution of Age

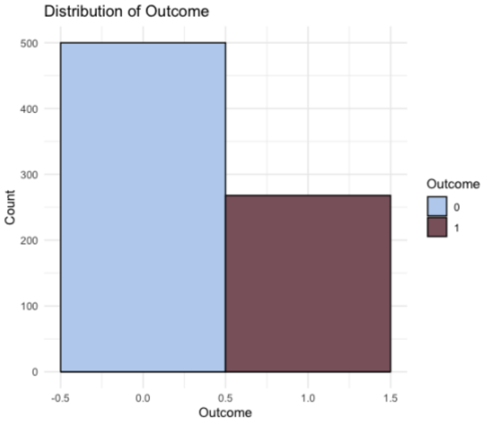
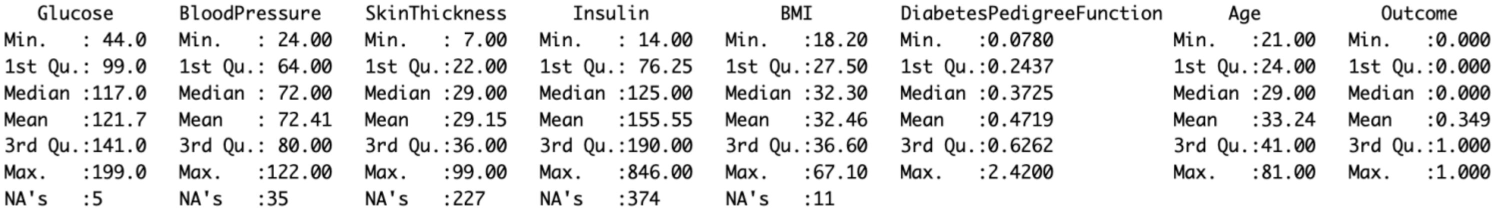


Figure 8: Plot for Distribution of Outcome

Next, the “summary()” function was used to calculate the summary statistics of each variable in the dataset. This function provides with minimum and maximum values, the first and third quartiles, the median, and the mean for each variable. Figure 9 depicts the output generated when using the summary function on the dataset.



```
   Glucose       BloodPressure    SkinThickness      Insulin           BMI        DiabetesPedigreeFunction       Age           Outcome
Min.   : 44.0   Min.   : 24.00   Min.   : 7.00   Min.   : 14.00   Min.   :18.20   Min.   :0.0780          Min.   :21.00   Min.   :0.000
1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00   1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437          1st Qu.:24.00   1st Qu.:0.000
Median :117.0   Median : 72.00   Median :29.00   Median :125.00   Median :32.30   Median :0.3725          Median :29.00   Median :0.000
Mean   :121.7   Mean   : 72.41   Mean   :29.15   Mean   :155.55   Mean   :32.46   Mean   :0.4719          Mean   :33.24   Mean   :0.349
3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00   3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262          3rd Qu.:41.00   3rd Qu.:1.000
Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00   Max.   :67.10   Max.   :2.4200          Max.   :81.00   Max.   :1.000
NA's   :5       NA's   :35       NA's   :227     NA's   :374      NA's   :11
```

Figure 9: Summary statistics of all Variables in the Dataset

Following insights can be drawn from the above histograms:

- Patients who tested negative are represented by class 0, which has 500 examples, whereas patients who tested positive are represented by class 1, which has 268 instances. Around 65 percent of patients tested negative, which suggests that the data set is tiny and biased. This can serve as a study constraint.
- Variables Age, DiabetesPedigreeFunction, Insulin, are all strongly skewed to the right. Although BMI, BloodPressure, and Glucose seem to have a normal distribution.

Statistical Tests and Visualizations for Assessing Differences in Central Tendencies

"Firstly, assuming all predictor variables are independent of one another, perform statistical tests assess differences in central tendencies of your predictor variables with respect to diabetes outcome. Visualise your results appropriately.”

The aim of this analysis is to investigate whether there are differences in central tendencies of the predictor variables with respect to the diabetes outcome. The dataset consists of 768 samples, where the Outcome variable takes on the value of 0 for non-diabetic and 1 for diabetic individuals. At Some of the predictor variables, including BMI, BloodPressure, and SkinThickness, were considered normally distributed, while others are not. The above hypothesis was substantiated by a Q-Q plot as shown in figure 10.

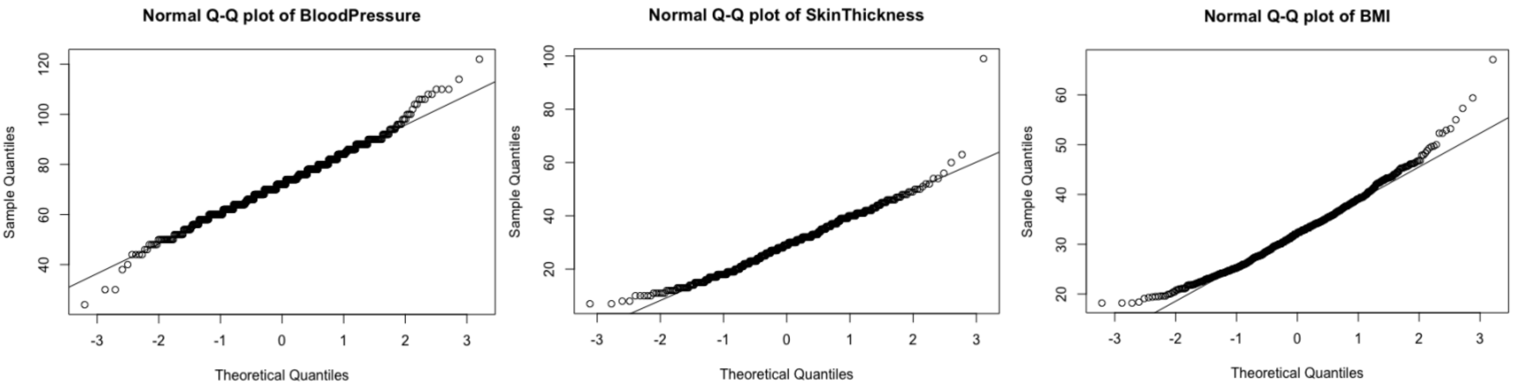
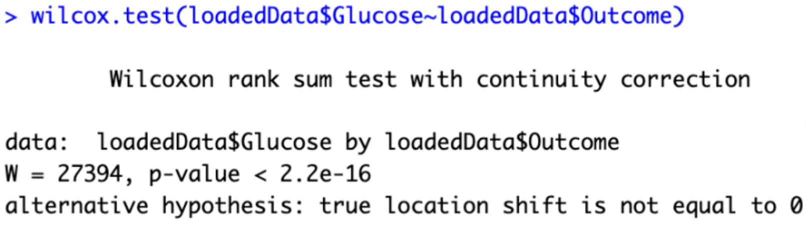


Figure 10: Q-Q plots for BloodPressure, SkinThickness and BMI

Next, appropriate statistical tests are performed for the normally distributed variables, we used a t-test, and for the non-normally distributed variables, we used a Wilcoxon rank-sum test.

Following are the results obtained from the test conducted and boxplots obtained,

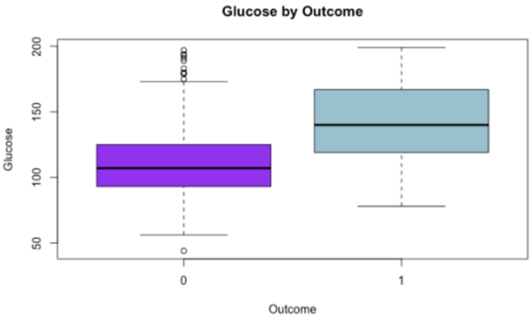
- **Glucose**



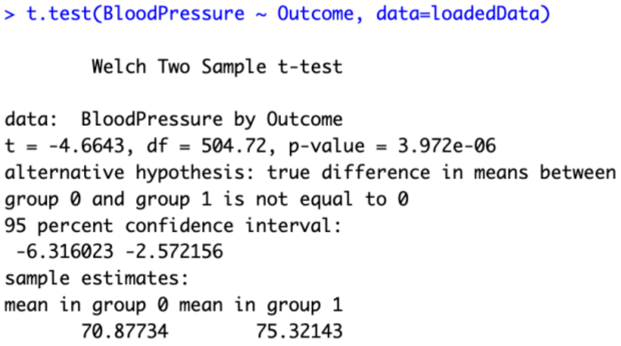
```
> wilcox.test(loadedData$Glucose~loadedData$Outcome)

        Wilcoxon rank sum test with continuity correction

data:  loadedData$Glucose by loadedData$Outcome
W = 27394, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```



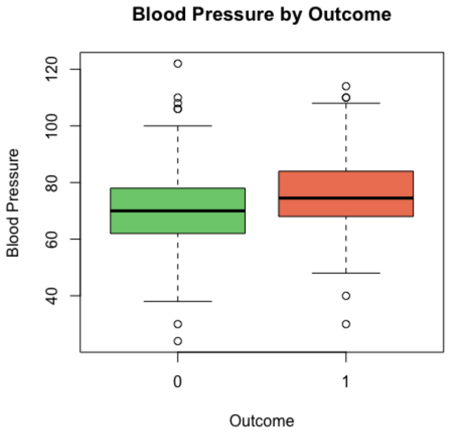
- **BloodPressure**



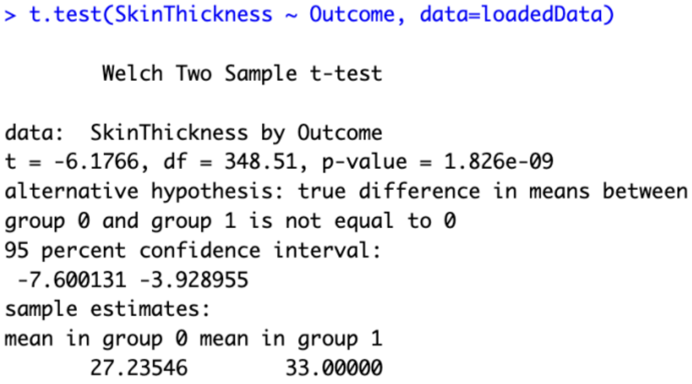
```
> t.test(BloodPressure ~ Outcome, data=loadedData)

        Welch Two Sample t-test

data:  BloodPressure by Outcome
t = -4.6643, df = 504.72, p-value = 3.972e-06
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -6.316023 -2.572156
sample estimates:
mean in group 0 mean in group 1
       70.87734        75.32143
```



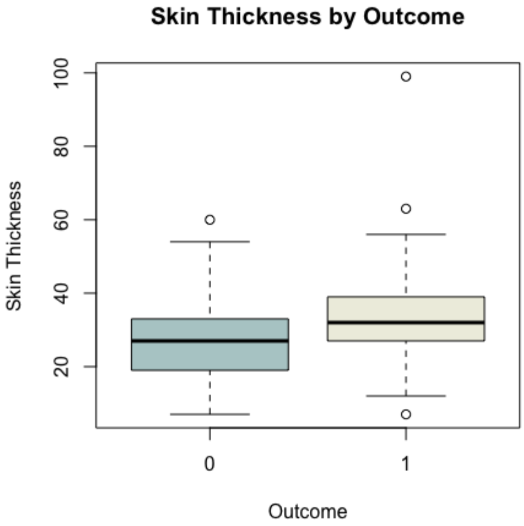
- **SkinThickness**



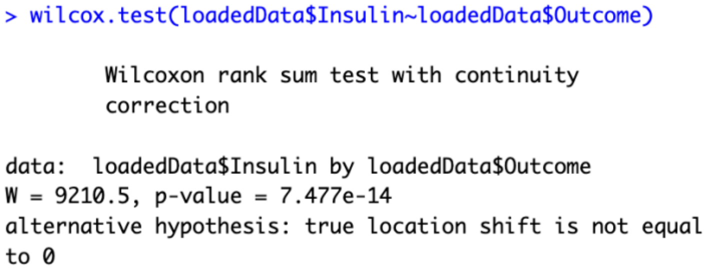
```
> t.test(SkinThickness ~ Outcome, data=loadedData)

        Welch Two Sample t-test

data:  SkinThickness by Outcome
t = -6.1766, df = 348.51, p-value = 1.826e-09
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -7.600131 -3.928955
sample estimates:
mean in group 0 mean in group 1
       27.23546        33.00000
```



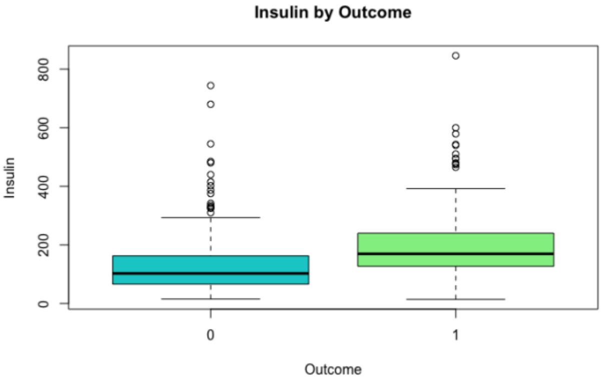
- **Insulin**



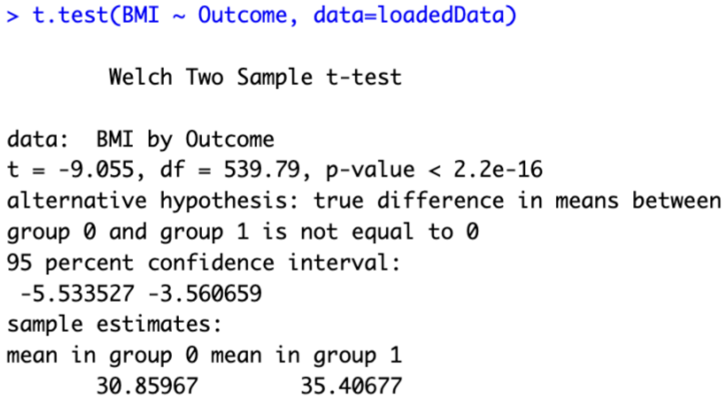
```
> wilcox.test(loadedData$Insulin~loadedData$Outcome)

        Wilcoxon rank sum test with continuity
        correction

data:  loadedData$Insulin by loadedData$Outcome
W = 9210.5, p-value = 7.477e-14
alternative hypothesis: true location shift is not equal
to 0
```



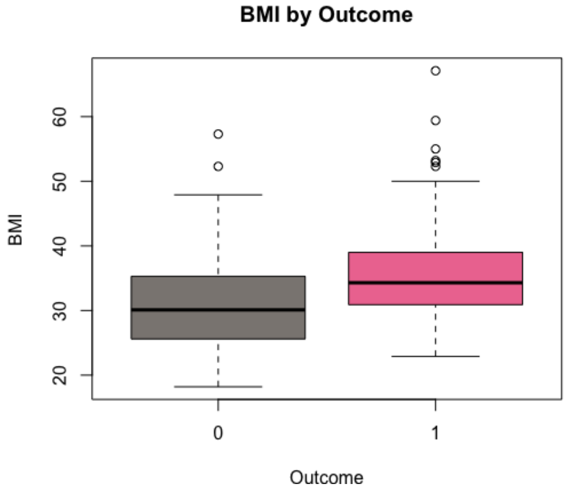
- **BMI**



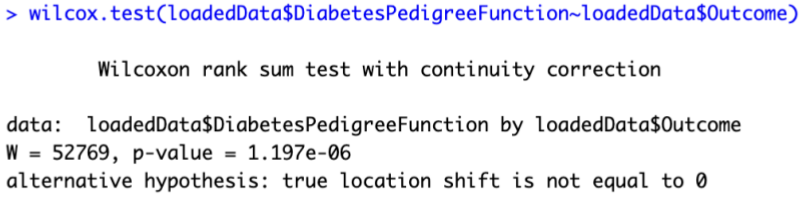
```
> t.test(BMI ~ Outcome, data=loadedData)

	Welch Two Sample t-test

data:  BMI by Outcome
t = -9.055, df = 539.79, p-value < 2.2e-16
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -5.533527 -3.560659
sample estimates:
mean in group 0 mean in group 1
       30.85967        35.40677
```



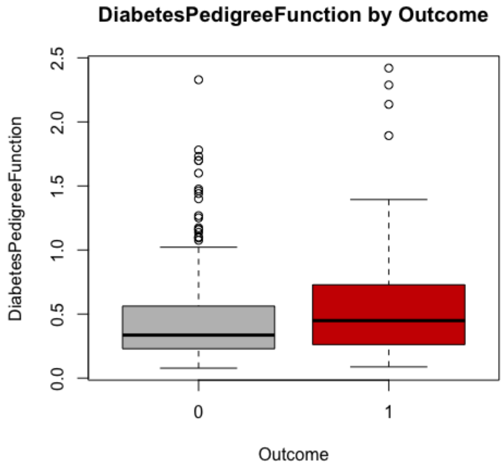
- **DiabetesPedigreeFunction**



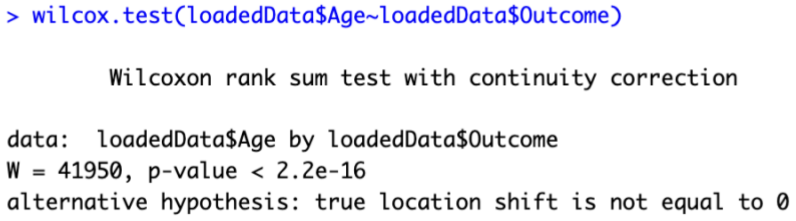
```
> wilcox.test(loadedData$DiabetesPedigreeFunction~loadedData$Outcome)

	Wilcoxon rank sum test with continuity correction

data:  loadedData$DiabetesPedigreeFunction by loadedData$Outcome
W = 52769, p-value = 1.197e-06
alternative hypothesis: true location shift is not equal to 0
```



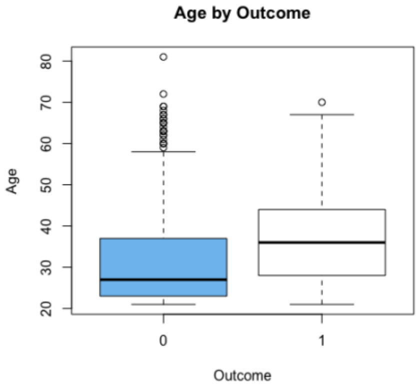
- **Age**



```
> wilcox.test(loadedData$Age~loadedData$Outcome)

	Wilcoxon rank sum test with continuity correction

data:  loadedData$Age by loadedData$Outcome
W = 41950, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```



Based on the given p-values, all predictor variables (Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age) are statistically significant predictors of the outcome variable, which is diabetes. This means that the mean or median values of these variables differ significantly between individuals with and without diabetes. In particular, BMI and Glucose show the strongest statistical significance among all the predictor variables, with p-values of 2.2e-16 in both t-test and Wilcoxon test, indicating that they are the most important predictors of diabetes outcome in this dataset.

Testing Independence of Predictor Variables

"Now, test the assumption that all predictor variables are independent using correlation coefficients. Visualise your results using scatter plots or otherwise."

To test the assumption that all predictor variables are independent, a correlation matrix of all the predictor variables is created and the correlation coefficients are examined. The association between each variable in the "diabetes" dataset is depicted visually using a correlation plot. The target "Outcome" and all of the variables are plotted against one another to look for any potential correlations. For this purpose, "corrplot" function is used as it serves as a visual exploratory tool on a correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables (Wei, 2021). Upon analysing the dataset in the first section using summary function, it was observed that there were missing values in every variable as shown in table 3.

Variable Name	No of Zeros
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0

Table 3: Number of Zeros in each variable

Variables like BloodPressure, Glucose, SkinThickness, BMI, and Insulin can not be zero. To overcome this, the mean of each variable is substituted in place of the missing values. The code snippet for replacing the mean of Glucose with the missing values in Glucose is shown below,

```
#Calculte the mean of Glucose column
meanGlucose <- mean(diabetesDataset$Glucose[diabetesDataset$Glucose > 0])

#Replace Zeros with mean value in the Glucose column
loadedData$Glucose <- ifelse(diabetesDataset$Glucose == 0, round(meanGlucose,0),
diabetesDataset $Glucose)
```

The same syntax is used to replace the missing values with their respective means in the BloodPressure, SkinThickness, Insulin, BMI column. Figure 11 shows the plot obtained for the diabetes dataset using the code below.

```
#Correlation Plot
library(corrplot) #Load the corrplot library
numVars <- unlist(lapply(diabetesDatasetMean, is.numeric))
diaNums <- diabetesDatasetMean[ , numVars]
diaCorr <- cor(diaNums)
corrplot(diaCorr, method="number")
corrplot.mixed(diaCorr,tl.pos = "lt",number.cex = 0.80,tl.cex=0.75,tl.col="black")
```

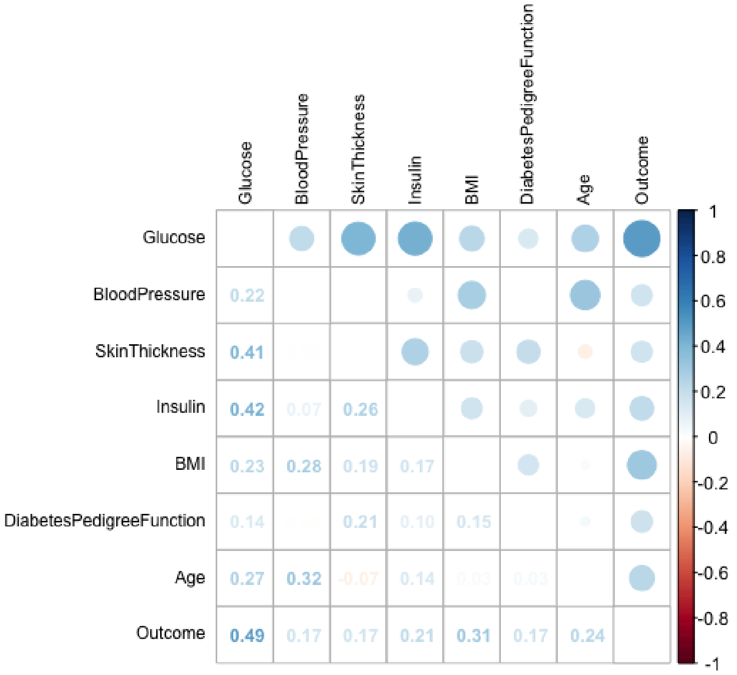


Figure 11: Correlation plot between all the variables of the dataset

To achieve this plot, the self-correlation values are eliminated, the tl.pos is set to "lt" to print the variable names outside the matrix, font size is set to 0.8%. The correlation matrix and map plotted in this analysis provided insight into the relationships between the predictor variables. The heatmap indicates the strength of the correlation between each pair of variables with brighter colors indicating a greater correlation. From the map, it can be inferred that there are significant correlations between Glucose, BMI, and outcome variable.

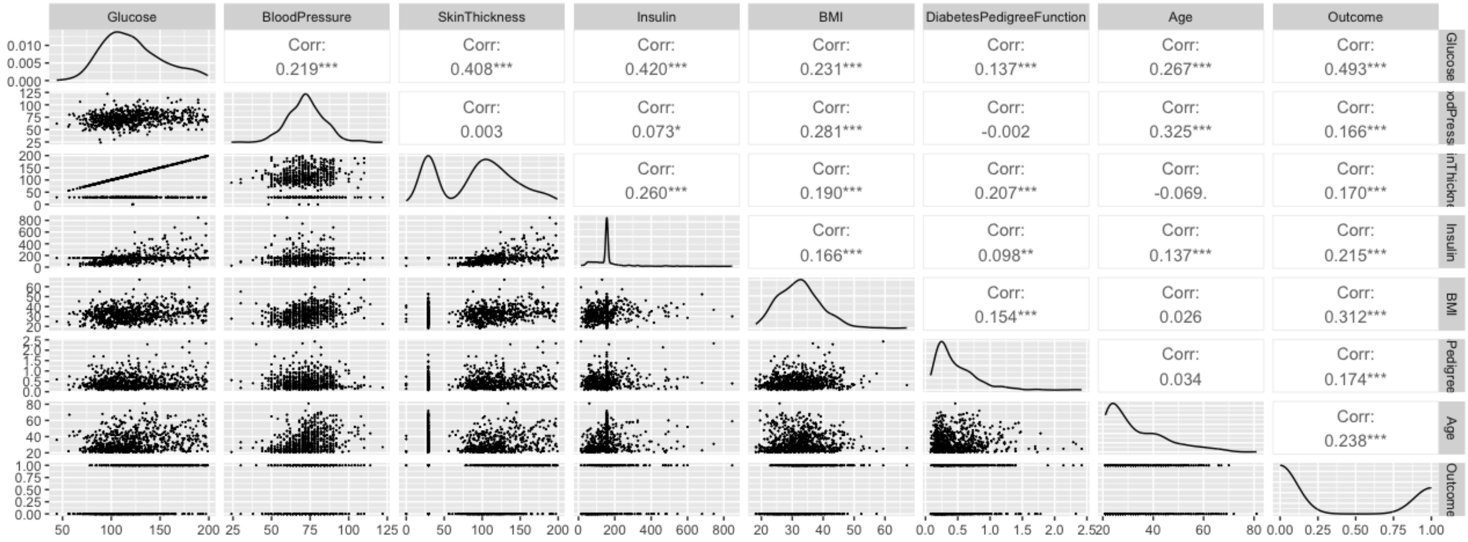


Figure 12: Scatter Plot diagram of all variables in the dataset

The best predictor of the outcome variable is glucose levels, with a correlation coefficient of 0.49 indicating a positive relationship. This suggests that as glucose levels increase, the probability of having diabetes increases. Similarly, BMI also shows a linear relationship with the outcome variable, with a correlation coefficient of 0.31. This indicates that as BMI increases, the probability of having diabetes also increases but not to great extent. It can also be inferred that BMI above 30 and high level of glucose together increase the risk of diabetes.

The correlation of 0.42 between Insulin and Glucose, also correlation of 0.41 between SkinThicknesss and Glucose indicates that these two variables are moderately positively correlated. This suggests that as Insulin increases, Glucose is likely to increase as well and the correlation coefficient of 0.32 between Age and Blood pressure also indicates that Females tend to have more blood pressure as their age increases. There is also a negative correlation of 0.07 between Age and Skin thickness. Overall, the correlation matrix and scatter plot provides useful information about the relationships between the predictor variables and the outcome variable, which can be used to guide further analysis and modelling.

Identification of Significant Predictors for Diabetes using logistic Regression Modelling

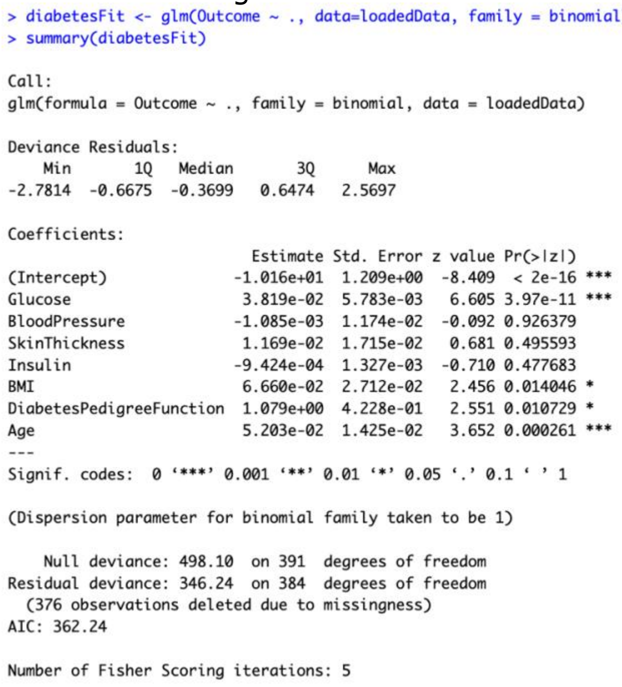
"Using linear regression modelling or otherwise, describe the variables which you believe have influence on diabetes. Note, you may consider other models beyond taught material which may be more appropriate for modelling the outcome variable."

The regression process is used to determine whether there is a relationship between all the variables and the Outcome, i.e., whether we can predict whether an individual has diabetes or not based on the data set. Note that when the dependent variable is continuous, linear regression is used, whereas logistic regression is used when the dependent variable is binary (Arya, 2022). As a result, logistic regression is used in this analysis.
The following equation represents logistic regression:

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}} \tag{1}$$

Here,
X = input value, y = predicted output, b_0 = bias or intercept term,
b_1 = coefficient for input (X)

In R, 'glm' function is used to fit a general linear model to the data set. To indicate that the regression model includes all the independent variables, use the shorthand notation '.' instead of explicitly writing out all the independent variables in the data set. The 'family = binomial' specifies that we want a logistic regression model. Following is the result obtained:



```
> diabetesFit <- glm(Outcome ~ ., data=loadedData, family = binomial)
> summary(diabetesFit)

Call:
glm(formula = Outcome ~ ., family = binomial, data = loadedData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7814  -0.6675  -0.3699   0.6474   2.5697

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.016e+01  1.209e+00  -8.409  < 2e-16 ***
Glucose                   3.819e-02  5.783e-03   6.605 3.97e-11 ***
BloodPressure            -1.085e-03  1.174e-02  -0.092 0.926379
SkinThickness             1.169e-02  1.715e-02   0.681 0.495593
Insulin                  -9.424e-04  1.327e-03  -0.710 0.477683
BMI                       6.660e-02  2.712e-02   2.456 0.014046 *
DiabetesPedigreeFunction  1.079e+00  4.228e-01   2.551 0.010729 *
Age                       5.203e-02  1.425e-02   3.652 0.000261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 346.24  on 384  degrees of freedom
  (376 observations deleted due to missingness)
AIC: 362.24

Number of Fisher Scoring iterations: 5
```

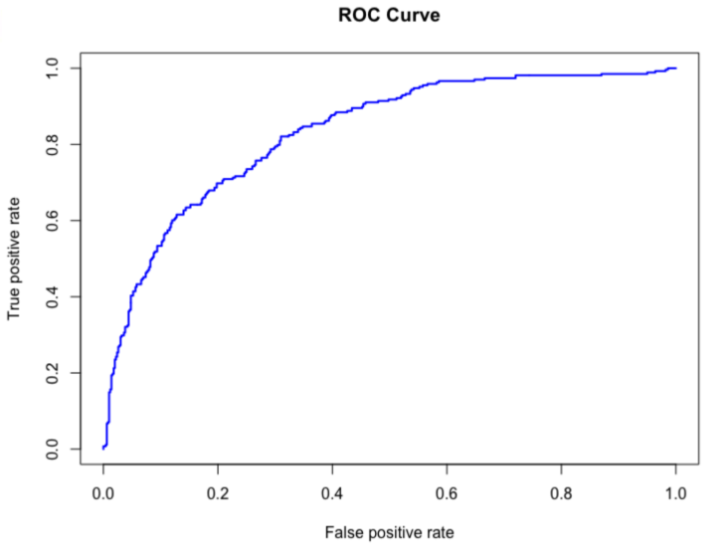


Figure 13: ROC curve for the Diabetes dataset

The logistic regression model was fit using the “cleaned_dataset” with the Outcome variable as the response variable and the remaining variables as predictor variables. The summary of the model shows the following results:

- Deviance Residuals: The minimum value of the residual is -2.7814, and the maximum value is 2.5697.
- Coefficients: The coefficients for the predictor variables are shown in the table, along with their estimates, standard errors, z-values, and p-values.
- The intercept is -10.16, and it is statistically significant (p-value < 0.001).
- Null and residual deviances: The null deviance, representing the deviance of the model with no predictors, is 498.10 with 391 degrees of freedom. The residual deviance, representing the deviance of the model with predictors, is 346.24 with 384 degrees of freedom. The AIC of the model is 362.24.

The coefficient for "Glucose" is 0.038, which indicates that for each one unit increase in glucose level, the log odds of diabetes increase by 0.038 units, holding all other variables constant. This means that as glucose level increases, the risk of diabetes also increases.

The coefficient for "BloodPressure" is -0.001, "SkinThickness" is 0.012 and "Insulin" is -0.001, which is not statistically significant (p-value of 0.926, 0.496 and 0.478 respectively). This suggests that blood pressure may not be a significant predictor of diabetes in this model.

The coefficient for "BMI" is 0.067, which is statistically significant (p-value of 0.014). This suggests that for each one unit increase in BMI, the log odds of diabetes increase by 0.067 units, holding all other variables constant. This means that as BMI increases, the risk of diabetes also increases.

The coefficient for "DiabetesPedigreeFunction" is 1.079, which is statistically significant (p-value of 0.011). This suggests that for each one unit increase in the diabetes pedigree function, the log odds of diabetes increase by 1.079 units, holding all other variables constant. This means that as the diabetes pedigree function increases, the risk of diabetes also increases.

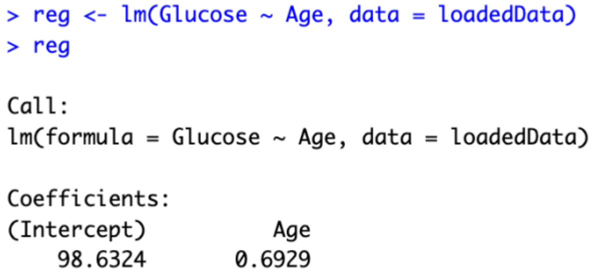
The coefficient for "Age" is 0.052, which is statistically significant (p-value of 0.0003). This suggests that for each one unit increase in age, the log odds of diabetes increase by 0.052 units, holding all other variables constant. This means that as age increases, the risk of diabetes also increases.

On a whole, the variables Glucose, BMI, DiabetesPedigreeFunction, and Age appear to have a significant influence on diabetes in this model. Blood pressure, Insulin level and Skin Thickness may not be significant predictors of diabetes in this model.

Prediction of Glucose Levels for Missing Entries Based on Age using Linear Regression Modelling

"Using linear regression modelling or otherwise, provide predictions of Glucose levels for the missing entries (those who have a Glucose entry of 0), assuming that Glucose depends only on Age".

There are 5 missing entries of Glucose as per the dataset for the corresponding Age values: 21,22,22,37,41. Considering the condition that Glucose level is dependent only on Age factor, the linear regression model for the same is designed using the following code:



```
> reg <- lm(Glucose ~ Age, data = loadedData)
> reg

Call:
lm(formula = Glucose ~ Age, data = loadedData)

Coefficients:
(Intercept)          Age
    98.6324       0.6929
```

The output of the "lm" function indicates that the regression equation for predicting Glucose from Age is:

$$Glucose = 98.6324 + 0.6929 * Age \tag{2}$$

This means that for every one-unit increase in Age, we would expect Glucose to increase by 0.6929 units, on average, holding all other variables constant.

To predict missing glucose values, the regression equation can be used by plugging in the missing ages into the equation and solving for the corresponding glucose values. For example, a missing glucose value for a person who is 21 years old:

$$Glucose_{(Age:21)} = 98.6324 + 0.6929 * Age = 113.18 \tag{3}$$

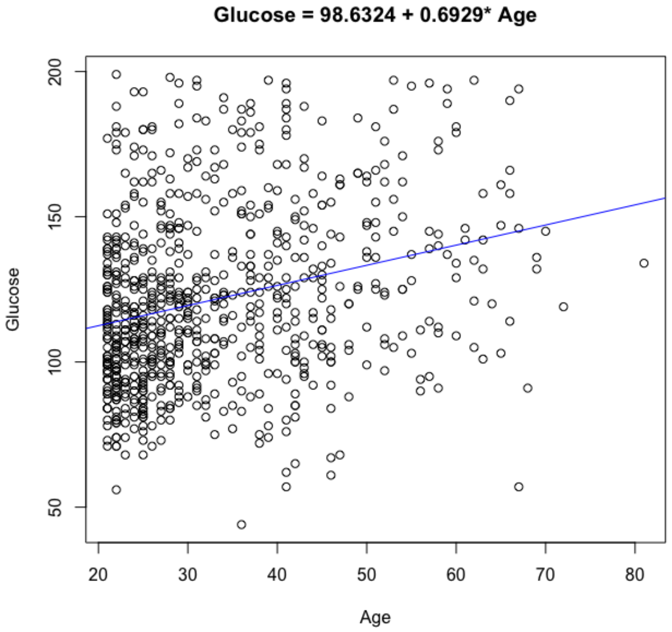


Figure 14: Linear plot between Glucose and Age Variables

Figure 13 shows the relationship graphically between Glucose and Age. Table 4 lists all the missing values obtained for Glucose by repeating the above mentioned procedure.

Age	Predicted Glucose Values
21	113.18
22	113.87
22	113.87
37	124.27
41	127.04

Table 4: Predicted Glucose Values from Age Variable

CONCLUSION

Based on our analysis of the diabetes dataset, it can be concluded several variables such as Age, BMI, and Diabetes pedigree function had a significant influence on the likelihood of a person developing diabetes. The logistic regression model also allowed to predict the probability of a patient developing diabetes given their specific characteristics.

Furthermore, linear regression was used to predict the missing values of glucose levels based on age. The analysis showed a positive correlation between glucose levels and age. This information can be useful in predicting glucose levels for patients with missing values, allowing for more accurate treatment and management of diabetes.

Overall, the analysis provides valuable insights into the factors contributing to diabetes and can be used to develop strategies for prevention, early detection, and treatment of this disease.

References

NHS (2019) NHS choices. Diabetes. Available at: https://www.nhs.uk/conditions/diabetes/ (Accessed: February 27, 2023).

Diabetes (2022) World Health Organization. World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed: February 27, 2023).

Tay, H.F. (2021) When is it OK to impute missing values with a zero?, Medium. Towards Data Science. Available at: https://towardsdatascience.com/when-is-it-ok-to-impute-missing-values-with-a-zero-6d94b3bf1352 (Accessed: February 28, 2023).

Heiss, A. (2021) Themes, Data Visualization. Available at: https://datavizs21.classes.andrewheiss.com/lesson/05-lesson/ (Accessed: March 2, 2023).

Wei, T. and Simko, V. (2021) An introduction to corrplot package. Available at: https://cran.microsoft.com/web/packages/corrplot/vignettes/corrplot-intro.html (Accessed: March 9, 2023).

Arya, N. (2022) Linear vs logistic regression: A succinct explanation, KDnuggets. Available at: https://www.kdnuggets.com/2022/03/linear-logistic-regression-succinct-explanation.html (Accessed: March 13, 2023).