

TABLE OF CONTENT

No.	Content	Page
1.	INTRODUCTION	1
2.	METHODOLOGY	3
3.	PROJECT OBJECTIVES	3
4.	RESULT AND DISCUSSION	4
5.	PROJECT OUTPUT	18
6.	CONCLUSION	20
7.	REFERENCE	
8.	APPENDIX	

INTRODUCTION

1.1 Project Description

The goal of this project is used generalized linear model to predict the cause of diabetes. The independent variables that will be used in the model are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The response variable, y will be outcome.

For this project, we use dataset from Kaggle that includes variables such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome. Using this dataset, we will do data preprocessing and analysis using Rstudio. We clean the data first by checking for missing values and duplicated rows. Then, we build the generalized linear model, use variance inflation factor (VIF) to check multicollinearity and ANOVA to check for significant variables. steps were used for models 2 and 3. Model adequacy checking was done, and detection of outliers for extreme values using Cooks' distance was done.

1.2 Problem Statement

To predict diabetes can be quite challenging as there are many factors to be considered before confirming whether a person has diabetes or not. It is very crucial to take note of every factor that can affect the result whether one has diabetes or not. If the factors are not carefully considered, the diagnosed result might be inaccurate or invalid. When the result become inaccurate or invalid, it may cause a person to lose the chance to get earlier treatment for their diabetes. Therefore, the goal of this study is to find the best generalized linear model to determine what factors affect the outcome whether a person has diabetes or not.

1.3 Data Description

The dataset that we are using in this project is from Kaggle which the author got from the National Institute of Diabetes and Digestive and Kidney Diseases.

Variables	Data description	Data Type
Pregnancies	Number of pregnancies	int
Glucose	Glucose level in blood	int
BloodPressure	Blood pressure measurement	int
SkinThickness	Thickness of the skin	int
Insulin	Insulin level in blood	int
BMI	Body mass index	float
DiabetesPedigreeFunction	Percentage of diabetes	float
Age	Age of the patient	int
Outcome	Determine diabetes or not.	int

2.0 METHODOLOGY

2.1 Tools Used

The libraries used in this project are in the table below:

Library	Decription
ggplot2	Ggplot2 is a system for creating graphics like maps. It can also include themes for personalizing charts.
corrplot	Corrplot package is used to display a correlation matrix confidence interval. It also contains some algorithms to do matrix reordering.
png	Plots in PNG can be easily converted to many other bitmap formats. It is loseless and best for line diagrams and blocks of solid colour.
car	Provides functions and tools for regression analysis.
tidyverse	Tidyverse library is used to help to transform and better present data. It helps with data import, tidying, manipulation and data visualization.
lmtest	A collection of tests, data sets, and examples for diagnostic checking in linear regression models.
pairsD3	Creates an interactive scatterplot matrix using the D3 JavaScript library.
Hmisc	Contains many functions useful for data analysis, high level graphics, utility operations, functions for computing sample size and power, simulation, importing and annotating dataset and many more functions.

3.0 PROJECT OBJECTIVES

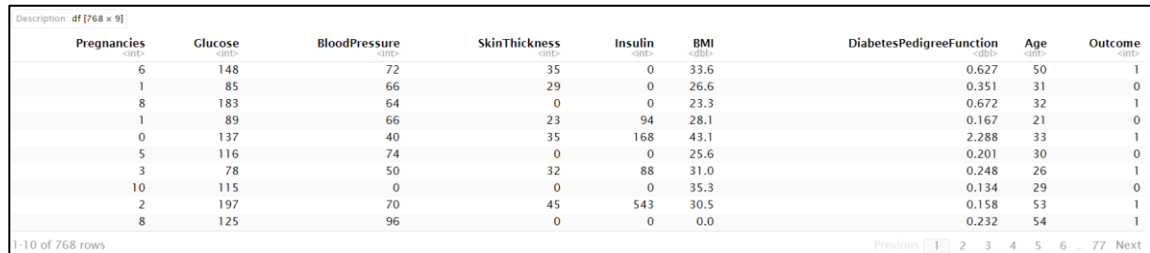
The project objectives are as follows:

1. To observe the most significant factor that affect the outcome of diabetes.
2. To develop the best generalized linear model to predict diabetes.
3. To develop a website that could help healthcare services to easily key in data of diabetes patients to keep track of their record.

4.0 RESULTS AND DISCUSSION

4.1 Data Analysis

Extract and load the "diabetes.csv" dataset into R studio.



Pregnancies <int>	Glucose <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	BMI <dbl>	DiabetesPedigreeFunction <dbl>	Age <int>	Outcome <int>
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
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31.0	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0.0	0.232	54	1

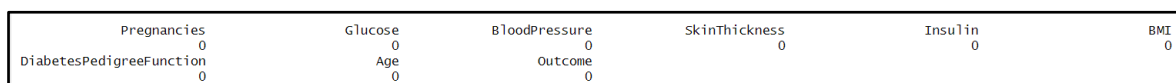
Figure 4.1 Import the dataset

After importing the data, check the dataset to see whether missing values exist. The result “FALSE” shows that the dataset does not have any missing values.

```
[1] FALSE
```

Figure 4.2 Check missing values

Next, check the dataset to see whether missing values exist in each column. The results show “0” in each column which mean no missing values in each column.



Pregnancies 0	Glucose 0	BloodPressure 0	SkinThickness 0	Insulin 0	BMI 0	DiabetesPedigreeFunction 0	Age 0	Outcome 0
------------------	--------------	--------------------	--------------------	--------------	----------	-------------------------------	----------	--------------

Figure 4.3 Check missing value in each column

Then, we check the dataset to see whether duplicate values exist in each row. Based on the figure below the dataset shows that there are no duplicate values in the dataset.



Description: df [0 x 9]
0 rows 1-6 of 9 columns

Figure 4.4 Check duplicate rows

Figure 4.5 displays the dataset's structure, which contains the number of observations and variables and the details of each variable. The result shows there are 768 observations and 9 variables. “BMI” and “DiabetesPedigreeFunction” are float while the other 7 variables are integer.

```
'data.frame': 768 obs. of 9 variables:
 $ Pregnancies      : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose          : int 148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure    : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness    : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin          : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI              : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...
 $ Age              : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome          : int  1 0 1 0 1 0 1 0 1 1 ...
```

Figure 4.5 Display the structure of the dataset

Figure 4.6 displays the summary of the dataset which includes the minimum, maximum, median, mean, 1st quartile and 3rd quartile for each variable.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Min. : 0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 0.0
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.: 0.00	1st Qu.: 0.0
Median : 3.000	Median :117.0	Median : 72.00	Median :23.00	Median : 30.5
Mean : 3.845	Mean :120.9	Mean : 69.11	Mean :20.54	Mean : 79.8
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:32.00	3rd Qu.:127.2
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.0
BMI	DiabetesPedigreeFunction	Age	Outcome	
Min. : 0.00	Min. :0.0780	Min. :21.00	Min. :0.000	
1st Qu.:27.30	1st Qu.:0.2437	1st Qu.:24.00	1st Qu.:0.000	
Median :32.00	Median :0.3725	Median :29.00	Median :0.000	
Mean :31.99	Mean :0.4719	Mean :33.24	Mean :0.349	
3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00	3rd Qu.:1.000	
Max. :67.10	Max. :2.4200	Max. :81.00	Max. :1.000	

Figure 4.6 Summary of the dataset

Figure 4.7 show the histogram and density curve for pregnancies and glucose, respectively. The distribution of the histogram for Pregnancies looks right skewed meanwhile the distribution of the histogram for glucose looks normal distributed.

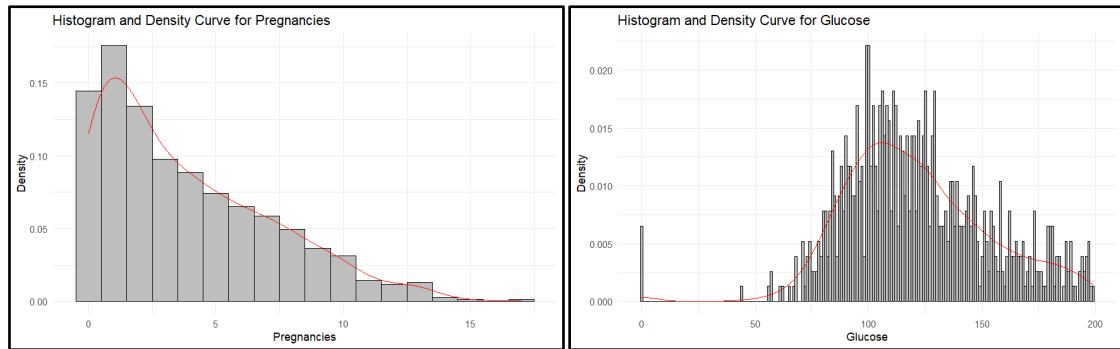


Figure 4.7 Histogram of Pregnancies and Glucose

Figure 4.8 shows the histogram and density curve for blood pressure and skin thickness, respectively. For the histogram of blood pressure, we can see that it is bell curved shape so we can assume that the blood pressure is normally distributed. The histogram of skin thickness appears to be normally distributed.

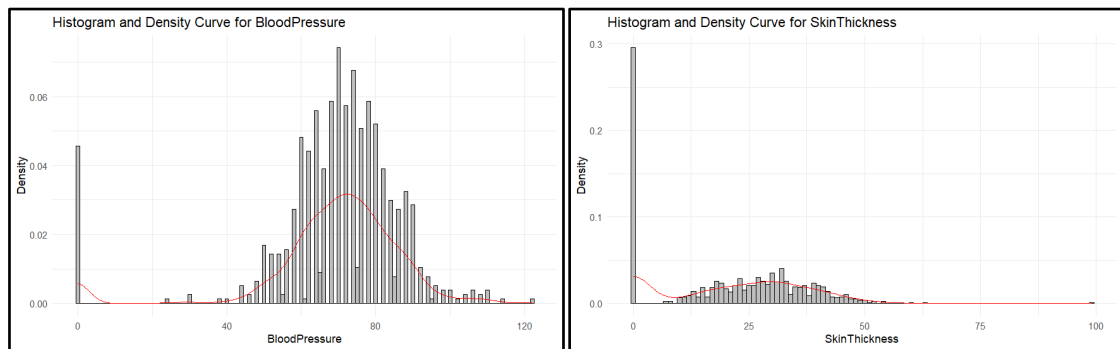


Figure 4.8 Histogram of Blood Pressure and Skin Thickness

Figure 4.9 shows Insulin and BMI's histogram and density curves, respectively. For Insulin's histogram, the distribution's shape seems right skewed, although it can't look clearly. Meanwhile, the histogram for BMI appears to be normally distributed.

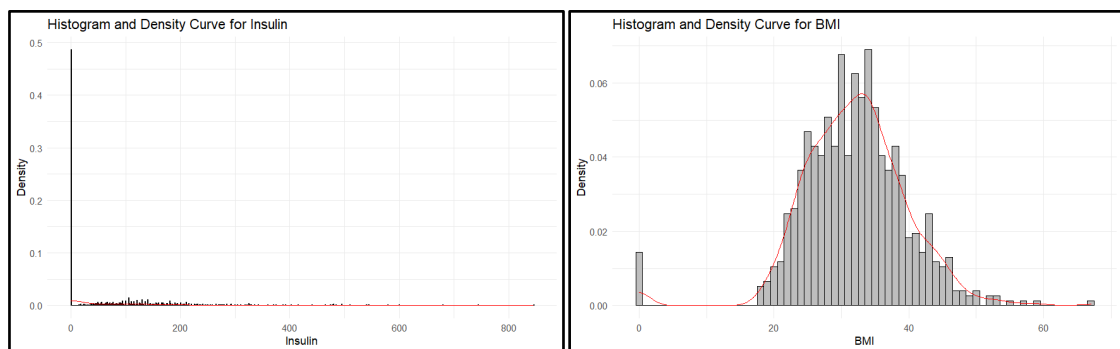


Figure 4.9 Histogram of Insulin and BMI

Figure 4.10 shows histogram and density curve for Diabetes Pedigree Function and Age, respectively. For the Age, the data is right skewed, implying that there is a greater concentration of results at the lower end of the Diabetes Pedigree Function scale. Meanwhile, the histogram for Age appears to be right skewed.

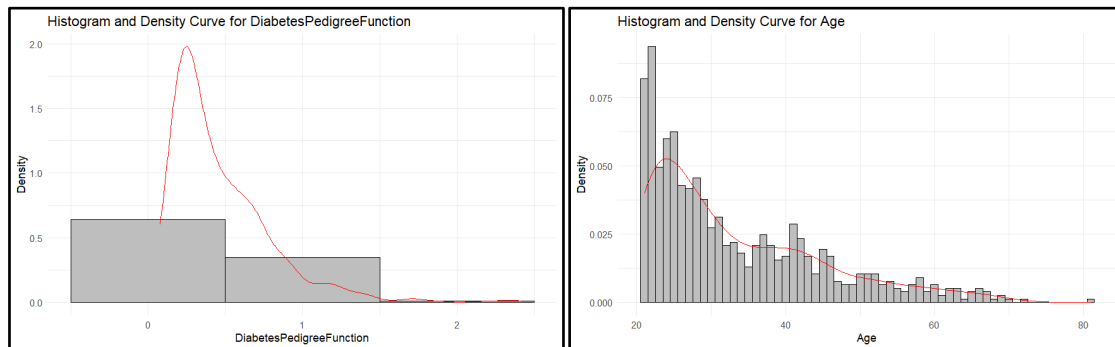


Figure 4.10 Histogram of Diabetes Pedigree Function and Age

Figure 4.11 shows the frequency of the outcome. Since 0 represents patients without diabetes and 1 refers to patients with diabetes. So, the number of patients who have diabetes is less than those without diabetes.

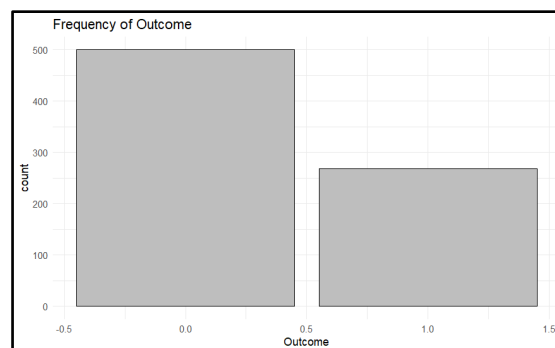


Figure 4.11 Frequency of Outcome

Figure 4.12 illustrates the correlation heatmap between the 9 variables. There is a moderate correlation coefficient between Pregnancy and Age, around 0.54. In other words, as age increases, the number of pregnancies also tends to increase. At the same time, the data shows a significant positive correlation of 0.47 between glucose levels and outcome, suggesting that higher glucose levels are linked to an increased probability of diabetes diagnosis. The correlation coefficient between Skin Thickness and Age in the heatmap is -0.11, indicating a

weak negative correlation. This means that as age increases, skin thickness slightly decreases, but the relationship is not strong.

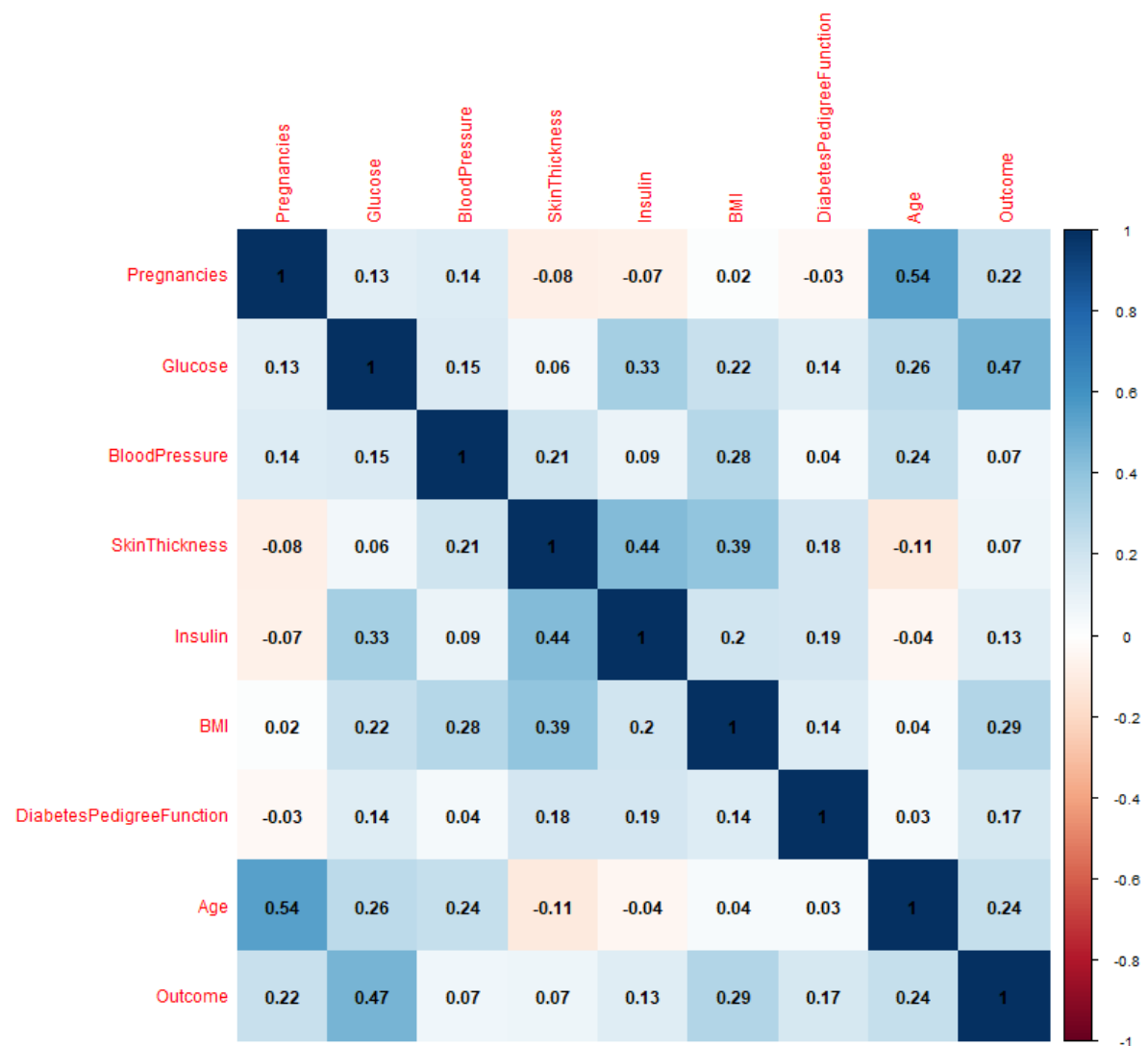


Figure 4.12 Correlation plot

4.2 Modelling

The variables that used in the first time generalized linear model (glm) is “Pregnancies”, “Glucose”, “BloodPressure”, “SkinThickness”, “Insulin”, “BMI”, “DiabetesPedigreeFunction” and “Age”. From the measures of the goodness of fit of a null model, a lower residual deviance indicates a better fit of the model to the data. In this case, the residual deviance is 723.45, which is lower than the null deviance of 993.48, indicating that the model is a better fit to the data than the null model.

From Figure 4.13, the values for null deviance and residual deviance are as follows:

Null deviance: 993.48 on 767 degrees of freedom

Residual deviance: 723.45 on 759 degrees of freedom

The Chi-Square statistics of the mode is calculated as follows:

$$\chi^2 = \text{Null deviance} - \text{Residual deviance}$$

$$\chi^2 = 993.48 - 723.45$$

$$\chi^2 = 270.03$$

The predictor variables degrees of freedom, p:

$$p = 767 - 759$$

$$p = 8$$

```
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
     data = diabetes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5566  -0.7274  -0.4159   0.7267   2.9297

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.4046964  0.7166359 -11.728 < 2e-16 ***
Pregnancies     0.1231823  0.0320776   3.840 0.000123 ***
Glucose         0.0351637  0.0037087   9.481 < 2e-16 ***
BloodPressure  -0.0132955  0.0052336  -2.540 0.011072 *
SkinThickness   0.0006190  0.0068994   0.090 0.928515
Insulin        -0.0011917  0.0009012  -1.322 0.186065
BMI             0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 **
Age            0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

Figure 4.13 First generalized linear model

Based on Figure 4.14 the VIF checking shows that all the variables have no multicollinearity between the variables.

```
Warning: package 'car' was built under R version 4.2.3 Loading required package: carData
Warning: package 'carData' was built under R version 4.2.3
SkinThickness 1.408434 1.214367 1.175283 1.522040
Insulin 1.467918 1.220416 1.034318 1.502069
Pregnancies
Glucose
BloodPressure
```

Figure 4.14 VIF checking

Based on Figure 4.15, the ANOVA table shows that pregnancies, glucose, Skin Thickness, BMI and DiabetesPedigreeFunction are significant since their p-values is less than $\alpha=0.05$. These significant variables will be used for our second generalized linear model due to its significance.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Outcome

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			767	993.48	
Pregnancies	1	37.274	766	956.21	1.026e-09 ***
Glucose	1	171.260	765	784.95	< 2.2e-16 ***
BloodPressure	1	0.888	764	784.06	0.3460418
SkinThickness	1	3.999	763	780.06	0.0455212 *
Insulin	1	1.972	762	778.09	0.1602210
BMI	1	41.243	761	736.85	1.344e-10 ***
DiabetesPedigreeFunction	1	10.880	760	725.97	0.0009719 ***
Age	1	2.522	759	723.45	0.1122535

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.15 ANOVA table

The variables that used in the second generalized linear model (glm) is “Pregnancies”, “Glucose”, “Skin Thickness”, “BMI”, and “DiabetesPedigreeFunction”. From the measures of the goodness of fit of a null model, a lower residual deviance indicates a better fit of the model to the data. In this case, the residual deviance is 733.06, which is lower than the null deviance of 993.48, indicating that the model is a better fit to the data than the null model.

From Figure 4.16, the values for null deviance and residual deviance are as follows:

Null deviance: 993.48 on 767 degrees of freedom

Residual deviance: 733.06 on 762 degrees of freedom

The Chi-Square statistics of the mode is calculated as follows:

$$\chi^2 = \text{Null deviance} - \text{Residual deviance}$$

$$\chi^2 = 993.48 - 733.06$$

$$\chi^2 = 260.42$$

The predictor variables degrees of freedom, p:

$$p = 767 - 762$$

$$p = 5$$

```
glm(formula = Outcome ~ Pregnancies + Glucose + SkinThickness +
    BMI + DiabetesPedigreeFunction, family = binomial(link = "logit"),
    data = diabetes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7498  -0.7417  -0.4326   0.7418   2.9115

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.471433   0.661394  -12.808 < 2e-16 ***
Pregnancies     0.139757   0.027159   5.146 2.66e-07 ***
Glucose         0.033790   0.003343  10.107 < 2e-16 ***
SkinThickness  -0.006715   0.006012  -1.117  0.2640
BMI             0.083831   0.014800   5.664 1.48e-08 ***
DiabetesPedigreeFunction 0.944013   0.295557   3.194  0.0014 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 733.06  on 762  degrees of freedom
AIC: 745.06

Number of Fisher Scoring iterations: 5
```

Figure 4.16 Second generalized linear model

Based on Figure 4.17 the VIF checking shows that all the variables have no multicollinearity between the variables.

Pregnancies	Glucose	SkinThickness	BMI
1.026442	1.002796	1.182071	1.164280
DiabetesPedigreeFunction			
1.027884			

Figure 4.17 VIF checking

Based on Figure 4.18, the ANOVA table shows that pregnancies, glucose, BMI and DiabetesPedigreeFunction are significant since their p-values is less than $\alpha=0.05$. These significant variables will be used for our third generalized linear model due to its significance.

Analysis of Deviance Table

Model: binomial, link: logit

Response: Outcome

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			767	993.48	
Pregnancies	1	37.274	766	956.21	1.026e-09 ***
Glucose	1	171.260	765	784.95	< 2.2e-16 ***
SkinThickness	1	3.030	764	781.92	0.081761 .
BMI	1	38.321	763	743.60	6.003e-10 ***
DiabetesPedigreeFunction	1	10.541	762	733.06	0.001167 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 4.18 ANOVA table

The variables that used in the third generalized linear model (glm) is “Pregnancies”, “Glucose”, “BMI”, and “DiabetesPedigreeFunction”. From the measures of the goodness of fit of a null model, a lower residual deviance indicates a better fit of the model to the data. In this case, the residual deviance is 734.31, which is lower than the null deviance of 993.48, indicating that the model is a better fit to the data than the null model.

From Figure 4.19, the values for null deviance and residual deviance are as follows:

Null deviance: 993.48 on 767 degrees of freedom

Residual deviance: 734.31 on 763 degrees of freedom

The Chi-Square statistics of the mode is calculated as follows:

$$\chi^2 = \text{Null deviance} - \text{Residual deviance}$$

$$\chi^2 = 993.48 - 734.31$$

$$\chi^2 = 259.17$$

The predictor variables degrees of freedom, p:

$$p = 767 - 763$$

$$p = 4$$

```
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
    family = binomial(link = "logit"), data = diabetes)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7581	-0.7349	-0.4264	0.7580	2.9008

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-8.415851	0.656908	-12.811	< 2e-16	***
Pregnancies	0.141926	0.027105	5.236	1.64e-07	***
Glucose	0.033826	0.003345	10.112	< 2e-16	***
BMI	0.078097	0.013771	5.671	1.42e-08	***
DiabetesPedigreeFunction	0.901294	0.291696	3.090	0.002	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 993.48 on 767 degrees of freedom
Residual deviance: 734.31 on 763 degrees of freedom
AIC: 744.31

Number of Fisher Scoring iterations: 5

Figure 4.19 Third generalized linear model

Based on Figure 4.20 the VIF checking shows that all the variables have a moderate correlation to outcome of diabetes.

Pregnancies	Glucose	BMI	DiabetesPedigreeFunction
1.022292	1.002622	1.018577	1.009126

Figure 4.20 VIF checking

Based on Figure 4.21, the residuals vs fitted plot is used to check the linear relationship. From the plot, we can see that the linear relationship does exist. A linear relationship is indicated by a horizontal line that lacks unique patterns, which is positive.

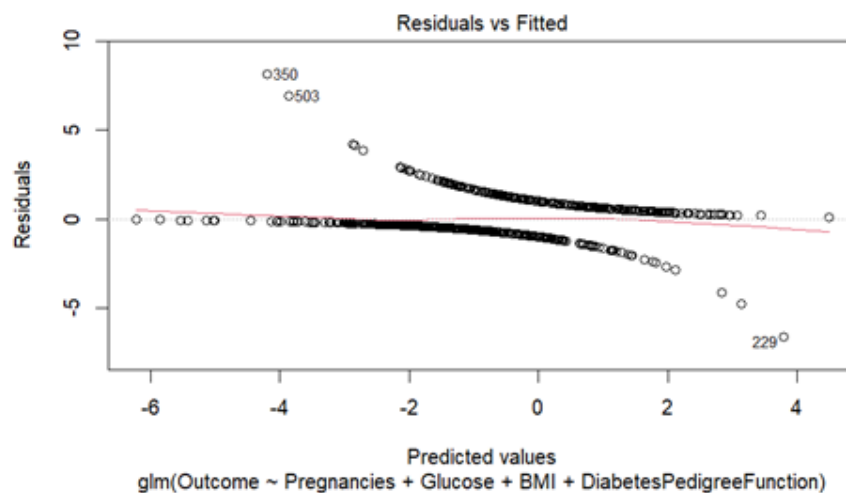


Figure 4.21 Residual vs Fitted Plot

Based on the Figure 4.22, the residuals are not normally distributed because the residuals point not follow the straight dashed line. To ensure that the plot is not normally distributed, Shapiro-Wilk test to prove.

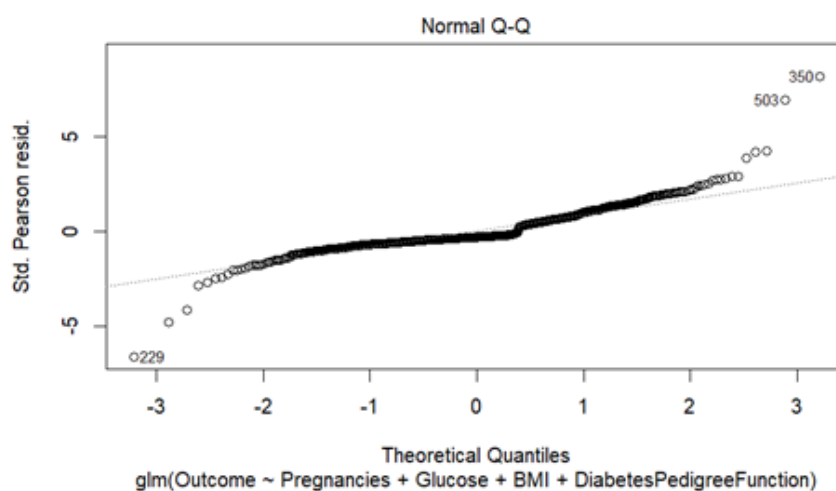


Figure 4.22 Normal Q-Q Plot20 VIF checking

Based on Figure 4.23 which is normality test using Shapiro-Wilk Test, it shows that the p-value is lower than $\alpha=0.05$. Which mean that we will reject the null hypothesis. We have sufficient evidence to say that the residuals are not normally distributed.

```
Shapiro-wilk normality test

data: residuals
W = 0.93146, p-value < 2.2e-16
```

Figure 4.23 Normality Test using Shapiro-Wilk Test

Based on Figure 4.24, the scale location plot is used to check the homogeneity of variance of the residuals. A horizontal line with equally spread points is a good indication of homoscedasticity. From the scale location plot, we are not sure whether heteroscedasticity exists or not therefore we conduct the Breusch-Pagan test.

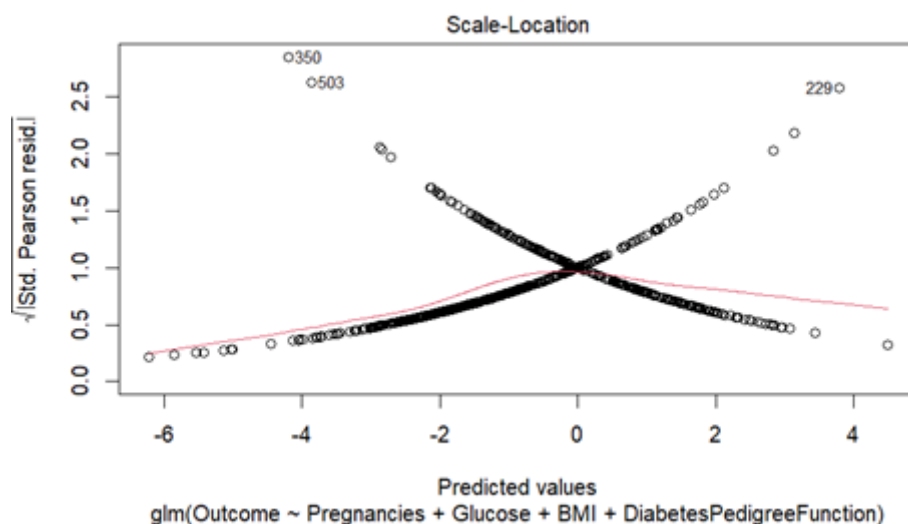


Figure 4.24 Scale Location Plot

Based on the BP test show in figure 4.25, we can see that the p-value is 0 which is less than $\alpha=0.05$. Then, we do not reject null hypothesis. We have sufficient evidence to say that heteroscedasticity is present in the model.

```
studentized Breusch-Pagan test

data: model3
BP = 37.983, df = 4, p-value = 1.129e-07
```

Figure 4.25 BP Test

Based on Figure 4.26, the residuals vs leverage plot is used to identify influential cases. The extreme values that might influence the regression when included or excluded from the analysis. The residuals vs leverage plot shows that there exists influential point in the plot.

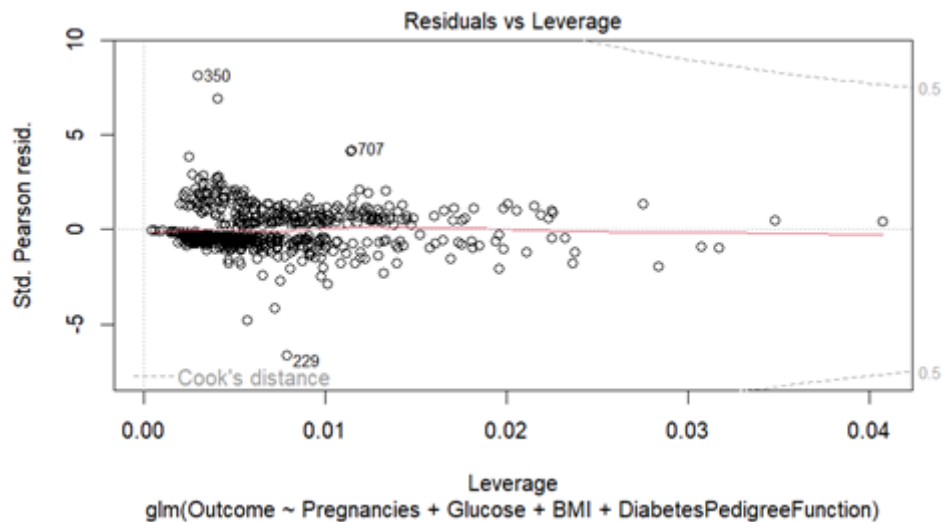


Figure 4.26 Residuals vs Leverage Plot

Based on the Figure 4.27 above, there exists a lot of cook's distance greater than 4 times the mean may be classified as influential.

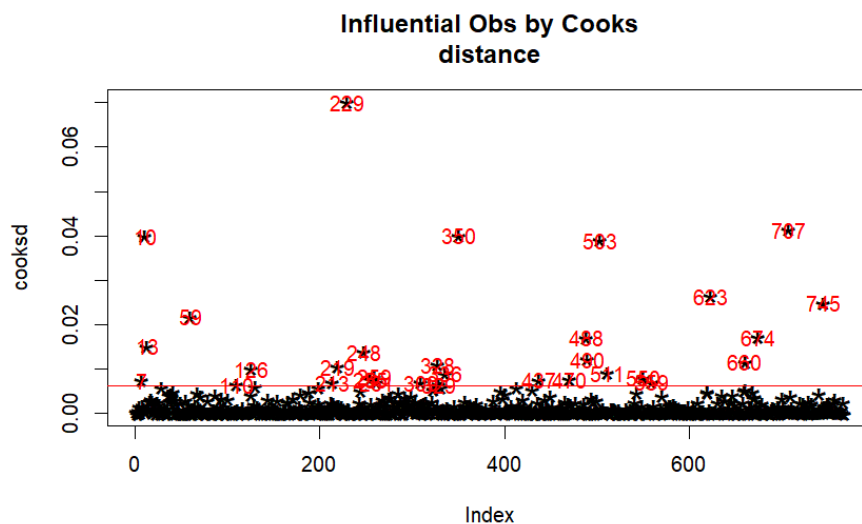


Figure 4.27 Influential Observations by Cooks Distance

Figure 4.28 depicts the pie chart for the number of diabetes outcomes. We know that the number of patients without diabetes is more than that of patients has diabetes.

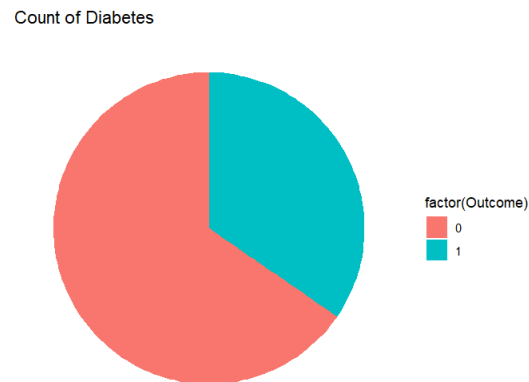


Figure 4.28 Pie chart for diabetes outcome

Figure 4.29 shows a scatterplot between BMI and Glucose with the color of the points indicating the diabetes status of the individuals. The plot shows a positive correlation between glucose and BMI.

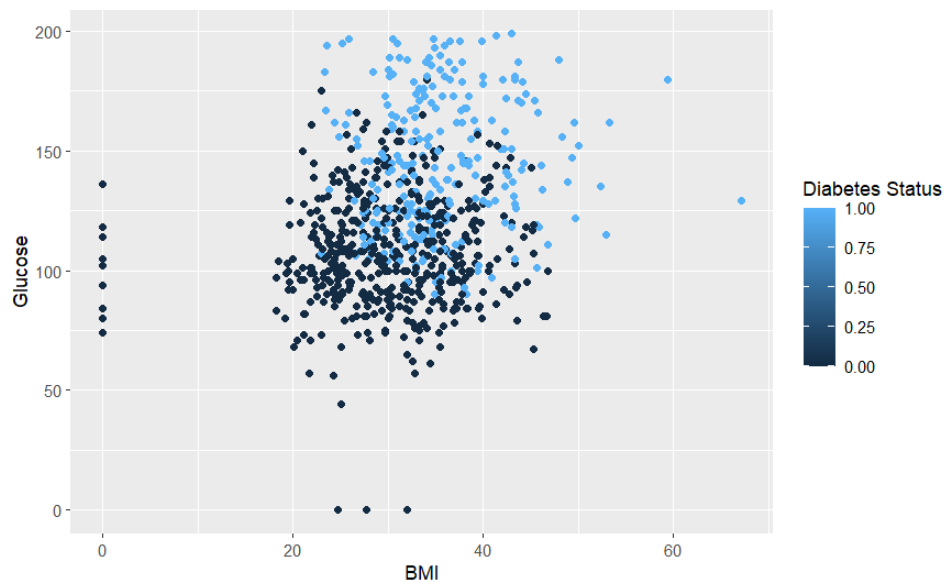
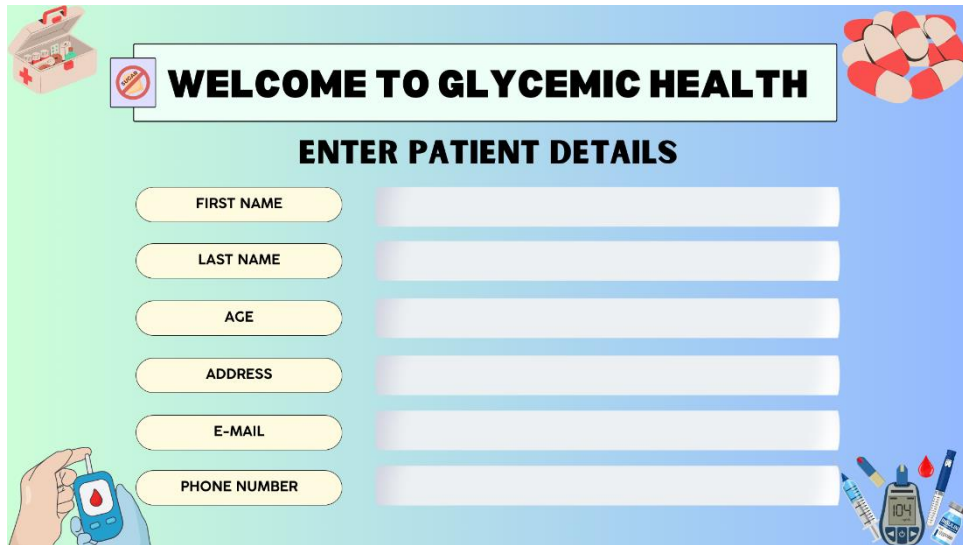


Figure 4.29 BMI and Glucose scatter plot

5.0 PROJECT OUTPUT

5.1 Website Interface



The image shows a web interface for 'GLYCEMIC HEALTH'. At the top, there's a header with a first aid kit icon on the left, a 'NO SMOKING' sign, the title 'WELCOME TO GLYCEMIC HEALTH', and a cluster of red and white pills on the right. Below the header is a section titled 'ENTER PATIENT DETAILS'. This section contains seven input fields, each with a label in a yellow rounded rectangle: 'FIRST NAME', 'LAST NAME', 'AGE', 'ADDRESS', 'E-MAIL', and 'PHONE NUMBER'. The background is a light blue gradient. Decorative elements include a hand holding a glucose meter on the bottom left and various medical supplies like syringes and a glucose meter on the bottom right.

WELCOME TO GLYCEMIC HEALTH

ENTER PATIENT DETAILS

FIRST NAME

LAST NAME

AGE

ADDRESS

E-MAIL

PHONE NUMBER

Figure 5.1 Welcome page with details



The image shows a web interface for 'GLYCEMIC HEALTH' with a section titled 'FILL IN MORE DETAILS'. This section contains eight input fields, each with a label in a yellow rounded rectangle: 'GLUCOSE', 'BLOOD PRESSURE', 'SKIN THICKNESS', 'INSULIN', 'WEIGHT', 'HEIGHT', 'BMI', and 'DIABETES PEDIGREE FUNCTION'. The background is a light blue gradient. Decorative elements include a cluster of colorful pills on the top left, a syringe and insulin bottle on the top right, and cartoon characters representing blood cells and a glucose meter on the bottom left. An illustration of a doctor and a patient is on the bottom right.

GLUCOSE

BLOOD PRESSURE

SKIN THICKNESS

INSULIN

WEIGHT

HEIGHT

BMI

DIABETES PEDIGREE FUNCTION

Figure 5.2 Fill in more details

DETAILS

FIRST NAME: SUZANNE	INSULIN: 30 G
LAST NAME: KIM	WEIGHT: 78 KG
AGE: 50	HEIGHT: 150 CM
ADDRESS: 75, ORCHARD ROAD, 27834	BMI: 34.7
EMAIL: SUZANNEKIM37@GMAIL.COM	DIABETES PEDIGREE FUNCTION: 0.25
PHONE NUMBER: 012571009	OUTCOME: DIABETIC PATIENT
GLUCOSE: 196 MG/DL	
BLOOD PRESSURE: 135 MM HG	
SKIN THICKNESS: 20.74 MM	

Figure 5.3 Full details with diagnosed result

Based on figure 5.1, it requires the doctor to fill up the details of the individual same goes to figure 5.2. Based on figure 5.3, the details of the individual will be sum up in one page and the outcome will determine whether the patient is a diabetic patient or not.

6.0 CONCLUSION

In conclusion, the best generalized linear model will be the third generalized linear model after reducing the factors. We use ANOVA to check for the significance of the variables to the outcome of diabetes. From the p-value, we develop a new model have significance value from the ANOVA table. Then, we continue checking with VIF and develop a new model with reducing the number of variables that can affect the outcome of diabetes. We did until the third generalized linear model where all the variables are significant. Therefore, the variables that used in the third generalized linear model (glm) is “Pregnancies”, “Glucose”, “BMI”, and “DiabetesPedigreeFunction” will affect the outcome of diabetes. Although there exist many factors that could affect the result whether an individual has diabetes or not but from our project we found that pregnancies, glucose, BMI and diabetes pedigree function are factors that affect diabetes.

Pregnancies affect women if they were overweight during pregnancy as the hormonal changes can affect glucose metabolism and insulin resistance. Glucose level must be monitored regularly to help identify if the individual is at risk of getting diabetes. BMI has effect on the result as well because when a person is obese or overweight, it is a risk factor for diabetes. Diabetes pedigree function refers to the history of diabetes in an individual’s family. If the diabetes patient has a family history of diabetes, it can increase the risk of getting diabetes.

From the view of business, we want to help to create a screening tool that can be useful to detect diabetes disease as it will be able to bring benefits to a lot of parties such as government, healthcare sector and the individual themselves. With the model that we had developed, it will be useful to help healthcare providers and individuals to try their best to do prevention or managing the diabetes disease.

REFERENCE

Introduction. (n.d.). R-Packages. Retrieved January 22, 2024, from [https://cran.r-](https://cran.r-hub.io/web/packages/corrplot/vignettes/corrplot-intro.html#:~:text=The%20corrplot%20package%20is%20a)

[hub.io/web/packages/corrplot/vignettes/corrplot-](https://cran.r-hub.io/web/packages/corrplot/vignettes/corrplot-intro.html#:~:text=The%20corrplot%20package%20is%20a)

[intro.html#:~:text=The%20corrplot%20package%20is%20a](https://cran.r-hub.io/web/packages/corrplot/vignettes/corrplot-intro.html#:~:text=The%20corrplot%20package%20is%20a)

irshadahma. (2023, July 31). *Car package in R*. GeeksforGeeks.

<https://www.geeksforgeeks.org/car-package-in-r/>

Lauren, N. (2023). *What is Tidyverse?* Study.com.

[https://study.com/academy/lesson/tidyverse-in-r-programming-definition-](https://study.com/academy/lesson/tidyverse-in-r-programming-definition-functions.html#:~:text=Tidyverse%20is%20an%20R%20programming)

[functions.html#:~:text=Tidyverse%20is%20an%20R%20programming](https://study.com/academy/lesson/tidyverse-in-r-programming-definition-functions.html#:~:text=Tidyverse%20is%20an%20R%20programming)

Madhugiri, D. (2022, March 7). *A Comprehensive Guide on ggplot2 in R*. Analytics Vidhya.

[https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-guide-on-ggplot2-](https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-guide-on-ggplot2-in-r/#:~:text=The%20ggplot2%20in%20R%20package)

[in-r/#:~:text=The%20ggplot2%20in%20R%20package](https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-guide-on-ggplot2-in-r/#:~:text=The%20ggplot2%20in%20R%20package)

R: JPEG and PNG graphics devices. (2020). Psu.edu.

<https://astrostatistics.psu.edu/su07/R/html/grDevices/html/png.html>

Tarr, G., htmlwidgets /lib, M. B. (d3 js library and much of pairsD3 code in, &

<https://d3js.org>). (2022, June 6). *pairsD3: D3 Scatterplot Matrices*. R-Packages.

<https://cran.rstudio.com/web/packages/pairsD3/index.html>