PROJECT PROPOSAL ON DIABETICS PREDICTION USING LOGISTIC REGRESSION

Submitted by

Huong Huynh (Claire)

Prathiba Swamykannu

Radhika Sood

Rekha Raj Ravindra

INTRODUCTION

The goal of this project is to predict whether a person is having diabetics or not. Several constraints were placed on the selection of these instances from a larger database. All patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

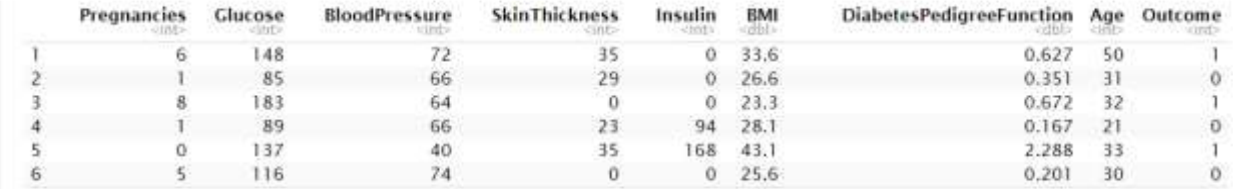
1.In your data set, how many covariates do you have? What are they? How many observations do you have? How did you collect them?

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database. Our project focuses on factors which affect the likelihood of person getting the diabetes.

We begin our assessment by considering the range of variables that are available in the dataset. There are a total of 768 rows and 9 columns in the dataset. By performing this analysis, we plan to reduce the unnecessary regressor variables one by one and end up with a simplified logistic regression model that will predict whether a person has been affected with diabetes or not.

There are a total of 8 covariates and 1 response variable. They are

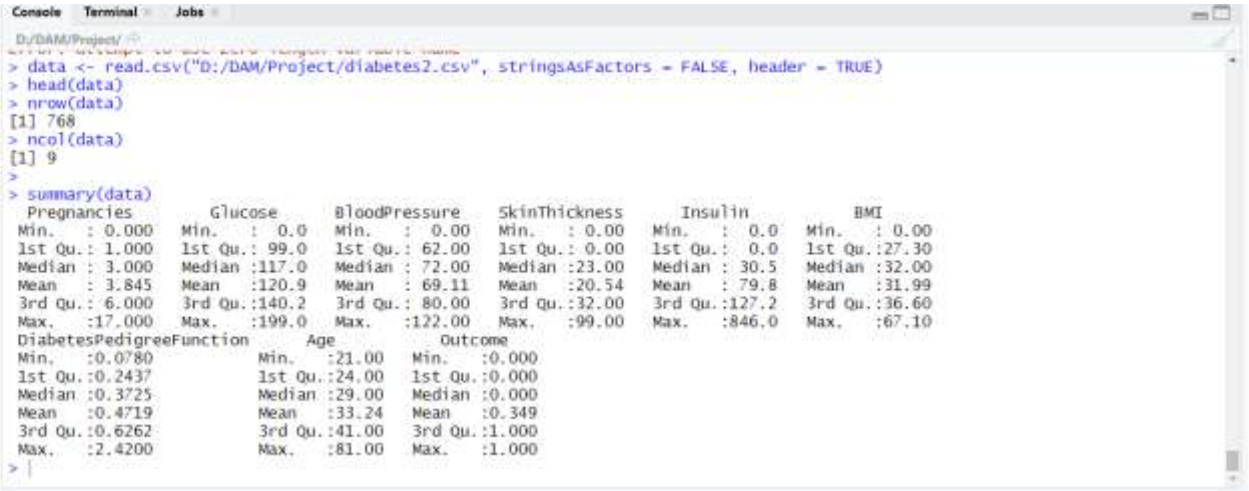
1. Pregnancies
2. Glucose
3. BloodPressure
4. SkinThickness
5. Insulin
6. BMI
7. DiabetesPedigreeFunction
8. Age
9. Outcome



	Pregnancies <int>	Glucose <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	BMI <dbl>	DiabetesPedigreeFunction <dbl>	Age <int>	Outcome <int>
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

2. Include a short exploratory data analysis of the collected data, i.e., figures and tables.

a. Summary of the data



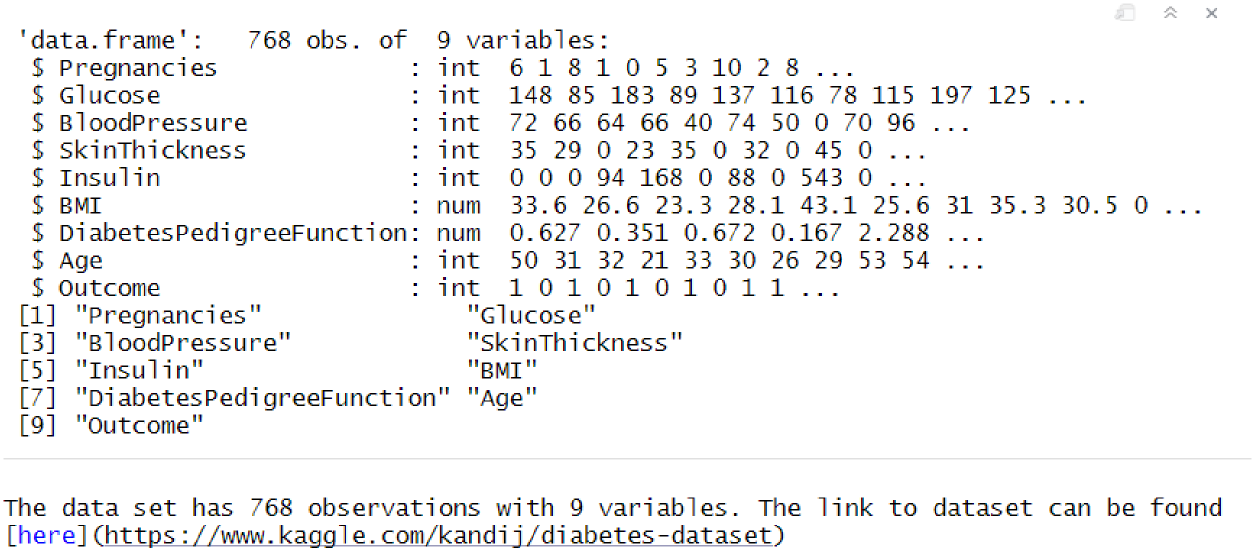
```
> data <- read.csv("D:/DAM/Project/diabetes2.csv", stringsAsFactors = FALSE, header = TRUE)
> head(data)
> nrow(data)
[1] 768
> ncol(data)
[1] 9
>
> summary(data)
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin          BMI       
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10  
 DiabetesPedigreeFunction      Age           Outcome     
 Min.   :0.0780           Min.   :21.00   Min.   :0.000  
 1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000  
 Median :0.3725           Median :29.00   Median :0.000  
 Mean   :0.4719           Mean   :33.24   Mean   :0.349  
 3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000  
 Max.   :2.4200           Max.   :81.00   Max.   :1.000  
> 
```

There are 9 columns in the dataset, within those columns we have 8 covariates and 1 response variable 'Outcome'.

The total of observation number is 768.

There are no missing values in any of the variables so we can safely proceed with the same. Additionally, we check for null values. The dataset is clean and contains no null values. So we can move on to the next phase of the analysis.

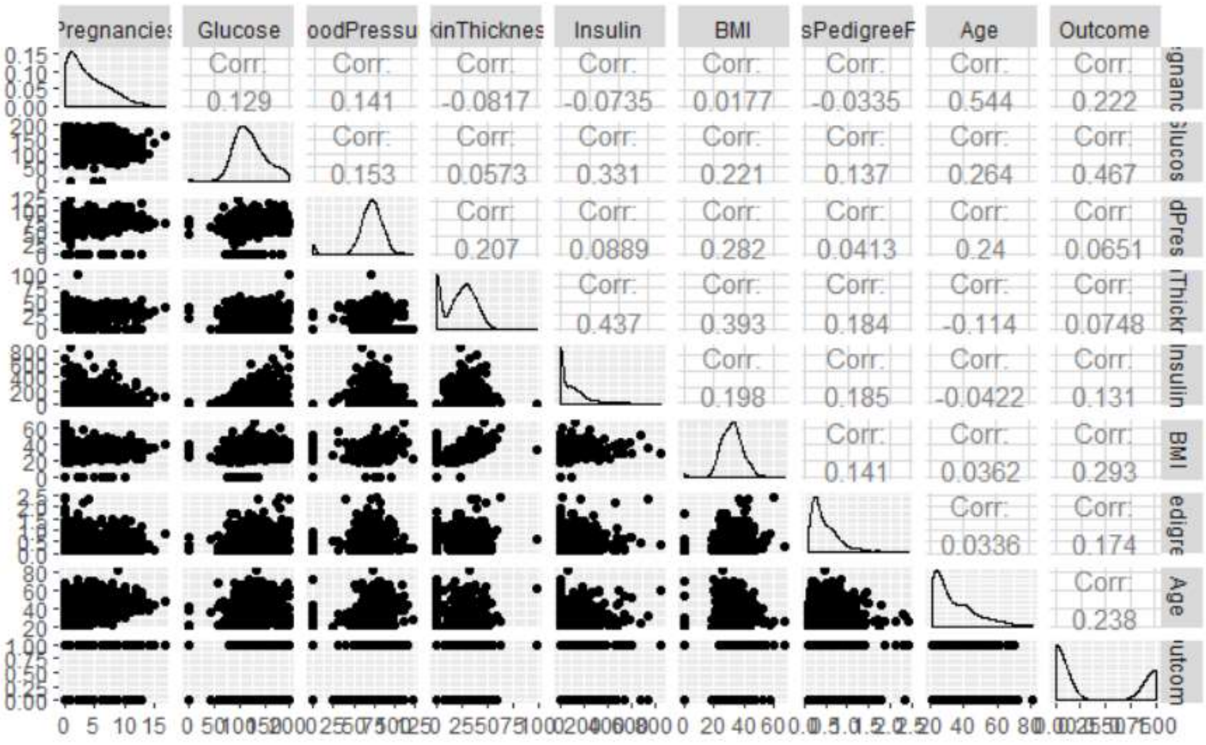
The covariates are



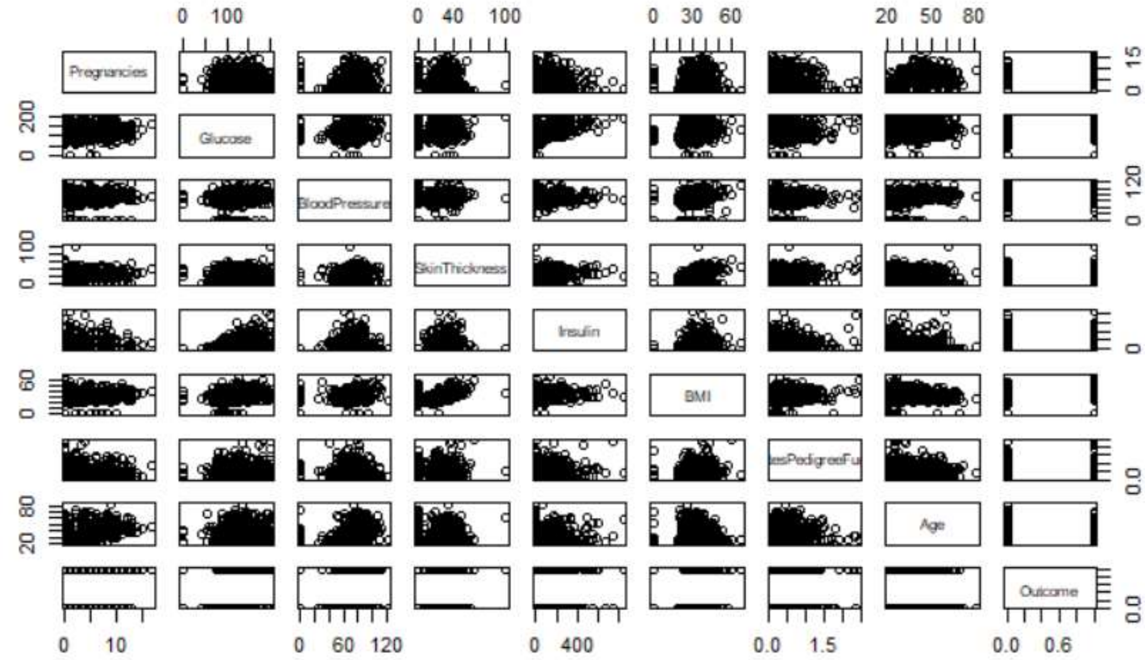
```
'data.frame':	768 obs. of  9 variables:
 $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 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 ...
[1] "Pregnancies"              "Glucose"                 
[3] "BloodPressure"            "SkinThickness"           
[5] "Insulin"                  "BMI"                     
[7] "DiabetesPedigreeFunction" "Age"                     
[9] "Outcome"                 
```

```
The data set has 768 observations with 9 variables. The link to dataset can be found
[here](https://www.kaggle.com/kandij/diabetes-dataset)
```

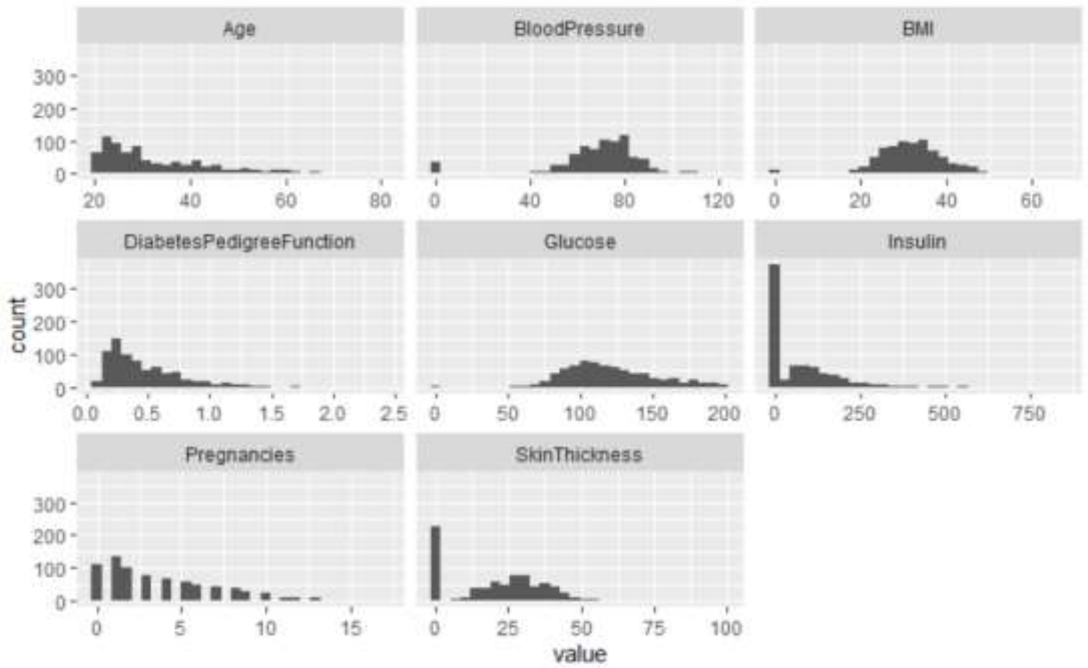
Correlation between the variables



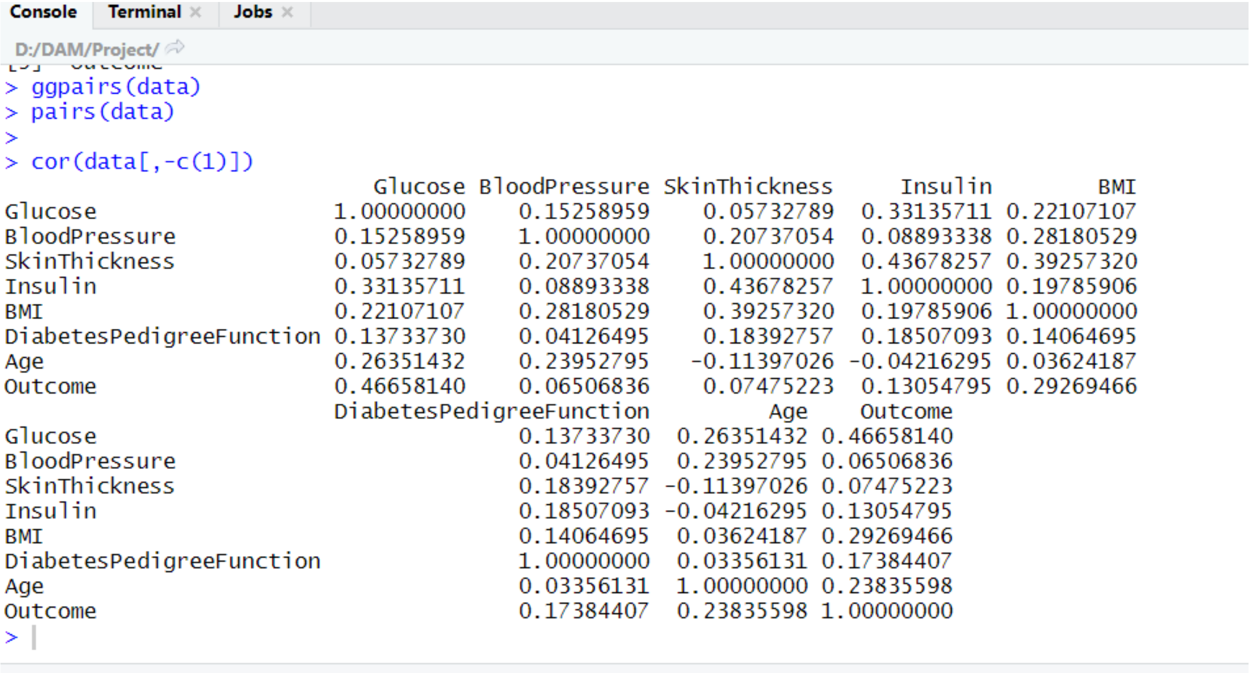
Pregnancies
Glucose
oodPressu
kinThicknes
Insulin
BMI
sPedigreeF
Age
Outcome
Corr:
0.129
0.141
-0.0817
-0.0735
0.0177
-0.0335
0.544
0.222
0.153
0.0573
0.331
0.221
0.137
0.264
0.467
0.207
0.0889
0.282
0.0413
0.24
0.0651
0.437
0.393
0.184
-0.114
0.0748
0.198
0.185
-0.0422
0.131
0.141
0.0362
0.293
0.0336
0.174
0.238
gnanc
Glucos
dPres
Thickr
Insulin
BMI
edigre
Age
utcom
0.15
0.10
0.05
0.00
0 5 10 15



Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
esPedigreeFu
Age
Outcome
0 100
0 40 100
0 30 60
20 50 80
15
0
200
0
120
0
100
0
0
60
0
0.0
80
20
0.0
0 10
0 60 120
0 400
0.0 1.5
0.0 0.6



From the above visualization, we can see that there are outliers in BloodPressure, Insulin and SkinThickness. Other covariates have closely spaced values.



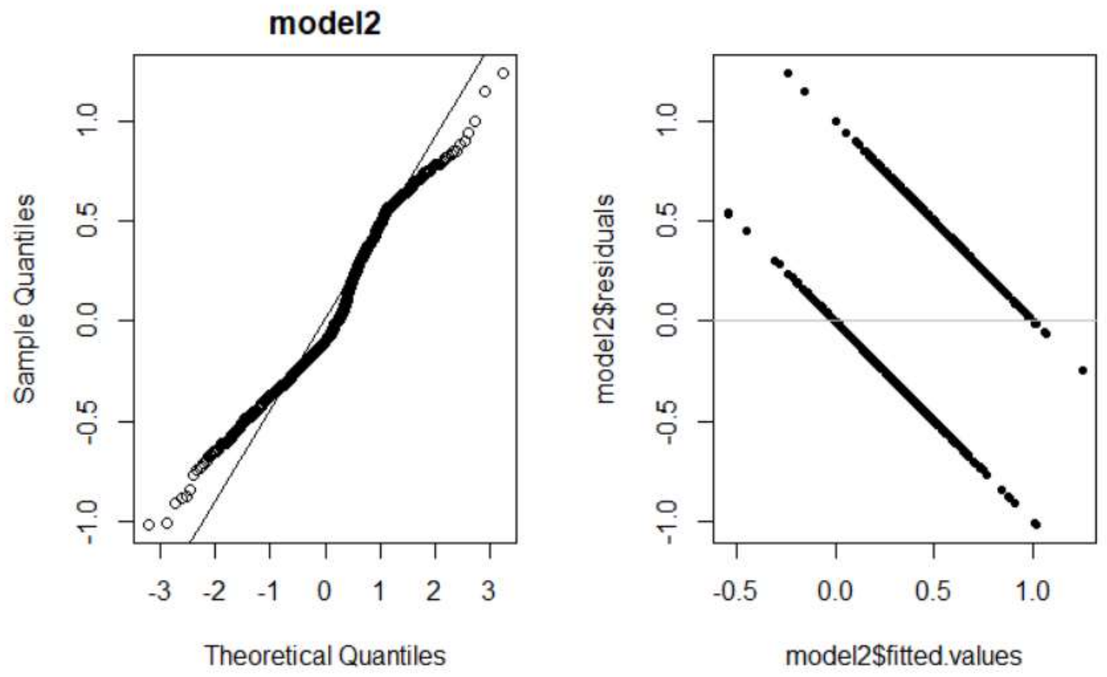
```
> ggpairs(data)
> pairs(data)
>
> cor(data[,-c(1)])
                            Glucose BloodPressure SkinThickness     Insulin        BMI
Glucose                  1.00000000    0.15258959    0.05732789  0.33135711 0.22107107
BloodPressure            0.15258959    1.00000000    0.20737054  0.08893338 0.28180529
SkinThickness            0.05732789    0.20737054    1.00000000  0.43678257 0.39257320
Insulin                  0.33135711    0.08893338    0.43678257  1.00000000 0.19785906
BMI                      0.22107107    0.28180529    0.39257320  0.19785906 1.00000000
DiabetesPedigreeFunction 0.13733730    0.04126495    0.18392757  0.18507093 0.14064695
Age                      0.26351432    0.23952795   -0.11397026 -0.04216295 0.03624187
Outcome                  0.46658140    0.06506836    0.07475223  0.13054795 0.29269466
                         DiabetesPedigreeFunction         Age    Outcome
Glucose                                0.13733730  0.26351432 0.46658140
BloodPressure                          0.04126495  0.23952795 0.06506836
SkinThickness                          0.18392757 -0.11397026 0.07475223
Insulin                                0.18507093 -0.04216295 0.13054795
BMI                                    0.14064695  0.03624187 0.29269466
DiabetesPedigreeFunction               1.00000000  0.03356131 0.17384407
Age                                    0.03356131  1.00000000 0.23835598
Outcome                                0.17384407  0.23835598 1.00000000
>
```

3. What kind of models are you using? What techniques are you using, for example, indicator variables, polynomial regression, transformation and so on?

For this project, we have decided to go ahead with logistic regression algorithm to predict if the person will suffer from diabetes. This algorithm will help us build our classification model. With the data we have binary response variable 'Outcome' values 1 and 0 which can represent “success” and “failure” respectively. This module is often used for biopharmaceutical field, clinical trials.

After deciding on choosing the type of model, we will fit logistic regression to data and interpret the output and make prediction.

4. What are the potential problems/issues in your model? For example, skewness, nonnormality, nonlinearity, multicollinearity, heteroscedasticity, dummy variables, outliers and/or the data simply has very weak signal?



The results of the residual analysis show that L, N, E (Linearity, Normality, Equal Variance) assumptions are violated. We can also see outliers in few variables like BMI which may or may not influence our model.

5. What kind of remedies are you proposing to use to solve the potential issues?

For problems pertaining to nonlinearity, nonnormality and unequal variances, transformation techniques would be used.

If multicollinearity really happen to exist in our model, we will resolve it using Ridge Regression.

Outliers will be treated based on the results of Cook’s Distance.