Data Analysis for Diabetes Prediction

1st Doga Yagmur Yilmaz
Galatasaray University, Introduction to Data Analysis
Report I
Istanbul,Turkey
dogaylmz8@hotmail.com

***Abstract*—In this document, we will see a detailed introduction of a Diabetes dataset. The high-level question that we ask to draw conclusion from dataset and low-level questions (technical questions) to help us answer the high-level question. Also, Hypothesis tests will be formed, and we will explain the methods which will be used for answering the questions.**

***Index Terms*—Data, sample, instance, feature, dimension, middle, variation, frequency, mean, median, mode, range, percentiles, interquartile range, boxplot, standard deviation, histogram, bar chart.**

Basic Data Science and ML Pipeline(OSEMN Pipeline)

- **O - Obtaining our data**
- **S - Scrubbing / Cleaning our data**
- **E - Exploring / Visualizing our data will allow us to find patterns and trends**
- **M - Modeling our data will give us our predictive power as a wizard (we will not apply this ML step)**
- **N - Interpreting our data**

I. Introduction

Diabetes is a chronic disease that occurs when the pancreas is no longer able to make insulin, or when the body cannot make good use of the insulin it produces. As we all know, this is a major problem worldwide. In general opinion, diabetes is very common in poor countries or in countries where people eat more fast food. However, diabetes can occur in humans due to many factors. In India, it is expected that by 2030 the number of people living with diabetes will rise to 101,2 million. There are reportedly 77.2 million people with prediabetes. In 2012, nearly 1 million people in India died of diabetes. Indians get diabetes 10 years before their Western counterparts on average. Changes in lifestyle lead to physical decreases: Increased fat, sugar and activity calories and higher insulin-cortisol levels, obesity and vulnerability... In 2011, India cost around 38 billion dollars annually because of diabetes.[1]

II. About This Project

The objective of this project is to classify whether someone has diabetes or not. So, the high-level question is "Can we be able to predict if someone suffers from diabetes before diagnosing her/him?" Dataset consists of several Medical Variables (Independent) and one Outcome Variable (Dependent). The independent variables in this data set are: 'Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes Pedigree Function', 'Age' The outcome variable value is either 1 or 0 indicating whether a person has diabetes (1) or not (0).

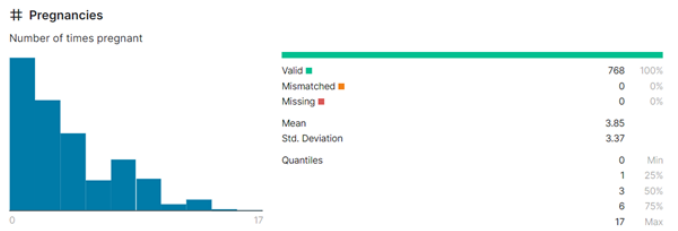
III. The Dataset

Pima Indians Diabetes Database (Predict the onset of diabetes based on diagnostic measures)[2]

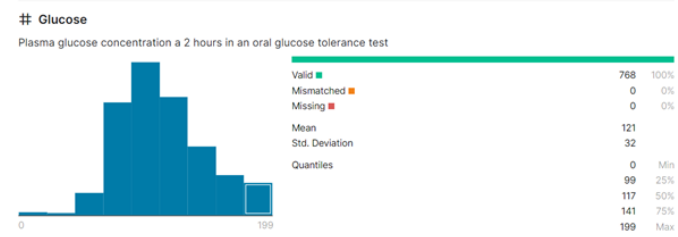
The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

A. About the Dataset

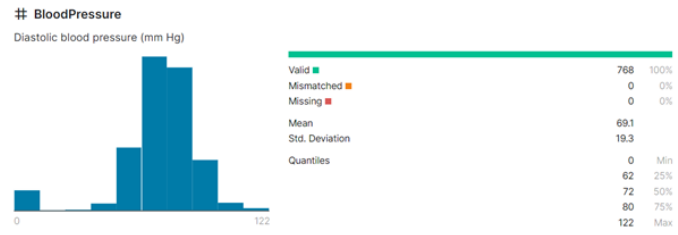
- Pregnancies:



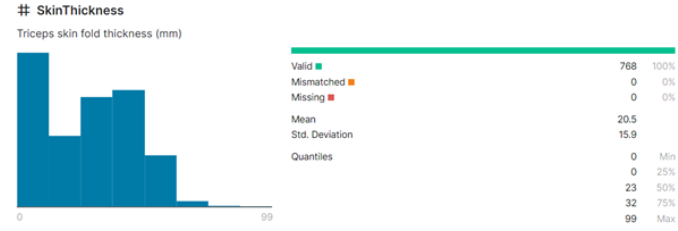
- Glucose:



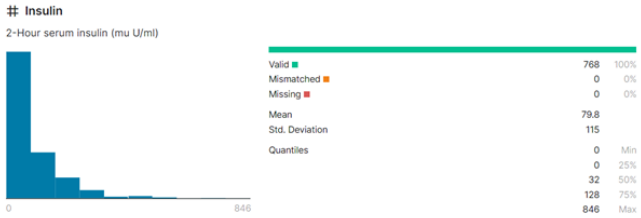
- Blood Pressure:



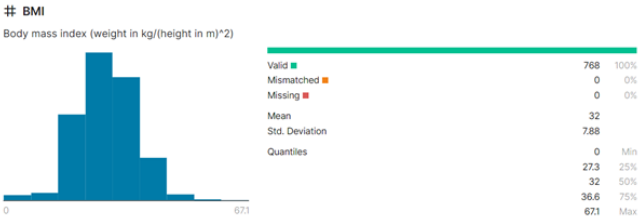
- Skin Thickness:



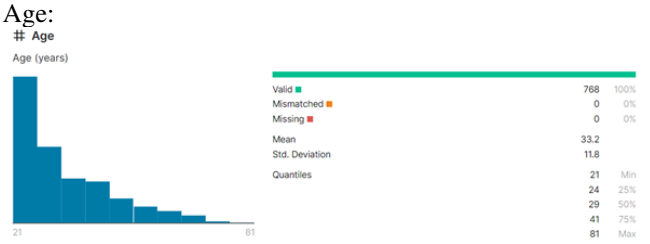
- Insulin:



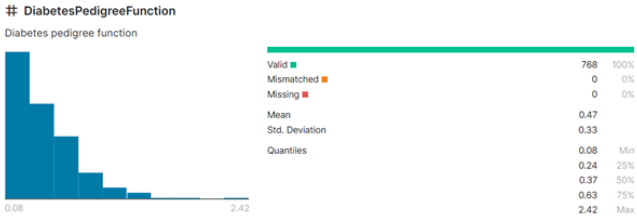
- Body Mass Index(BMI):



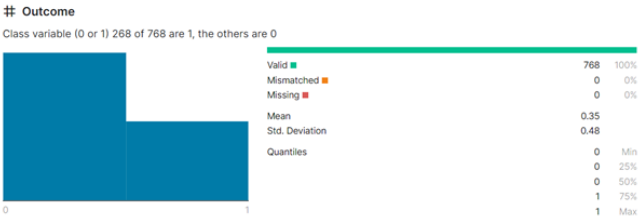
- Age:



- Diabetes Pedigree Function: Scores likelihood of diabetes based on family history.



- Outcome: 0(doesn't have diabetes) or 1 (has diabetes).



B. Context

This dataset comes from the Diabetes and Digestive and Kidney Disease National Institutes. The purpose of this dataset is to diagnose whether or not a patient is diabetic, on the basis of certain diagnostic measures in the dataset. The selection of these instances from a larger database was subject to several restrictions. All patients are women from the Indian heritage of Pima, at least 21 years old.

C. Structure

-Descriptive Data Analysis

- 1. Importing Required Libraries
- 2. Loading the Dataset
- 3. Exploratory Data Analysis
- a. Understanding the dataset
- -Head of the dataset
- -Shape of the dataset
- -Types of columns
- -Information about dataset
- -Summary of the dataset
- b. Data Cleaning
- -Dropping duplicate values
- -Checking NULL values
- -Checking for 0 value
- 4. Data Visualization
- - Count Plot: to see if the dataset is balanced or not
- -Histograms: to see if data is normally distributed or skewed
- -Box Plot: to analyze the distribution and see the outliers.
- -Scatter plots :to understand relationship between any two variables.
- Heat Map: to find out the strength of the correlation between two features.

IV. Questions

As we have the outcome data (someone has diabetes or not) which is target data and several medical variables which are predictor variables for us, we would like to "predict if someone suffers from diabetes before diagnosing her/him." It would be interesting to look into the factor like Body Mass Index, Glucose level or Blood Pressure that may affect the outcome. In this point we can ask some technical questions in order to answer the high-level question. These would be just about one feature:

- Are Predictors normally distributed?
- Can high glucose level be a sign of having diabetes?
- Can high number of pregnancy times be a sign of having diabetes?
- Can low blood pressure be a sign of having diabetes?
- Can low skin thickness be a sign of having diabetes?
- Can low insulin level be a sign of having diabetes?
- Can body mass index be a sign of having diabetes?
- Can Diabetes Pedigree Function be a sign of having diabetes?

,or about relations between the predictor variables:

- Do diabetic patients have high glucose levels and high number of pregnancy times?
- How is the age and pregnancy distribution for diabetic patients?
- Is there a Collinearity in all Predictors?

V. Hypothesis

- 1st Hypothesis:
- H0: All independent variables are normally distributed.
- H1:All independent variables are not normally distributed.
- For Solution:
- We will check each of predictors at " Normalizing Data":

- At first, we will check in section 'A.Checking for normality of predictors' Then we will check after scaling the data in section 'B. Handling Outliers'.
- 2nd Hypothesis:
- H0: High glucose level can't be a sign of having diabetes.
- H1: High glucose level can be a sign of having diabetes.
- For Solution:
- We will check the correlation between glucose level and outcome. If we get a high relationship between them, we will reject the null hypothesis.
- 3rd Hypothesis:
- H0: High number of pregnancy times can't be a sign of having diabetes.
- H1: High number of pregnancy times can be a sign of having diabetes.
- For Solution:
- We will check the correlation between pregnancies and outcome. If we get a high relationship between them, we will reject the null hypothesis.
- 4th Hypothesis:
- H0: Low blood pressure can't be a sign of having diabetes
- H1: Low blood pressure can be a sign of having diabetes
- For Solution:
- We will check the correlation between BloodPressure and outcome. If we get a high relationship in the opposite direction between them, we will reject the null hypothesis.
- 5th Hypothesis:
- H0: Body mass index can't be a sign of having diabetes.
- H1: Body mass index can be a sign of having diabetes.
- For Solution:
- We will check the correlation between BMI and outcome. If we get a high relationship between them, we will reject the null hypothesis.

VI. Descriptive Data Analysis

A. Importing Required Libraries

B. Loading the Dataset

```
In [8]: df=pd.read_csv(r"C:\Users\Monster\Downloads\diabetes.csv")
```

C. Exploratory Data Analysis

- a. Understanding the dataset
- -Head of the dataset

```
In [10]: df.head() #get familier with dataset, display the top 5 data records
Out[10]:
```

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

- -Shape of the dataset

```
In [11]: df.shape #getting to know about rows and columns we're dealing with - 768 rows , 9 columns
Out[11]: (768, 9)
```

- -Types of columns

```
In [12]: df.columns #learning about the columns
Out[12]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
                'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
               dtype='object')
```

- -Information about dataset

```
In [14]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Based on the above we can see our dataframe has 9 columns of which all are of type integers or float.
It doesn't appear we have any categorical variables with the exception of our response variable outcome.

- -Summary of the dataset

df.describe() method generates descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values. This method tells us a lot of things about a dataset. One important thing is that the describe() method deals only with numeric values. It doesn't work with any categorical values. So, if there are any categorical values in a column the describe() method will ignore it and display summary for the other columns unless parameter include="all" is passed. But it's okay for us because we have any categorical variable. Now, let's understand the statistics that are generated by the describe() method which help us to understand how data has been spread across the table.

- - Count: the number of non-empty rows in a feature.
- - Mean: mean value of that feature.
- - std: Standard Deviation Value of that feature.
- - min: minimum value of that feature.
- - max: maximum value of that feature.
- - 25, 50, and 75 are the percentile/quartile of each features.

```
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
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	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

CONCLUSION: We observe that min value of some columns is '0' which could not be possible medically. Hence in the data cleaning process we'll have to replace them with median/mean value depending on the distribution. Also, in the max column we can see insulin levels as high as "846" We must treat outliers.

- b. Data Cleaning
- -Dropping duplicate values "df=df.drop.duplicates()"
- -Checking NULL values

```
In [17]: df.isnull().sum()

Out[17]: Pregnancies                 0
         Glucose                     0
         BloodPressure               0
         SkinThickness               0
         Insulin                     0
         BMI                         0
         DiabetesPedigreeFunction    0
         Age                         0
         Outcome                     0
         dtype: int64
```

There are no null values in the dataset.

- -Checking for 0 value
  It isn't medically possible for some data record to have "0" value such as Blood Pressure, Glucose levels, Skin Thickness, Insulin levels and BMI. Hence, we replace them with the mean value of that particular column. We don't adjust Insulin and Skin Thickness as zero observations make up 40 per cent so adjusting zeros to the mean would significantly change the distribution of these variables. Age and DiabetesPedigreeFunction don't have minimum "0" value so no need to replace, also number of pregnancies as 0 is possible as observed in df.describe().

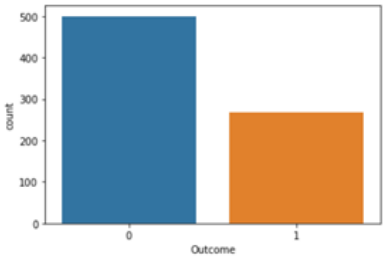
```
In [18]: print(df[df['BloodPressure']==0].shape[0])
         print(df[df['Glucose']==0].shape[0])
         print(df[df['SkinThickness']==0].shape[0])
         print(df[df['Insulin']==0].shape[0])
         print(df[df['BMI']==0].shape[0])

         35
         5
         227
         374
         11
```

Some of the columns have a skewed distribution, so the mean is more affected by outliers than the median. Glucose and Blood Pressure have normal distributions hence we replace 0 values in those columns by mean value. Skin Thickness, Insulin, BMI have skewed distributions hence median is a better choice as it is less affected by outliers.

D. Data Visualization

- - Count Plot:

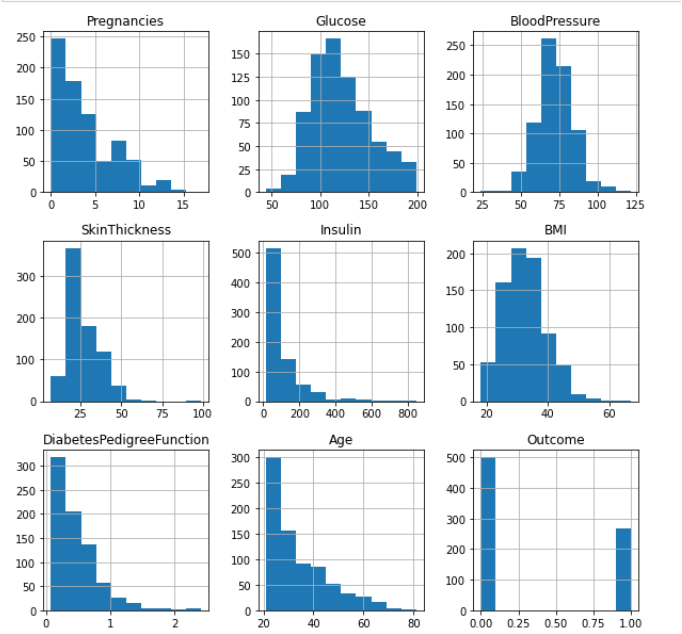
```
In [20]: sns.countplot('Outcome',data=df)
Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x1e820a07588>
```



We observe that number of people who don't have diabetes is more than people who has diabetes. This situation indicates that our data is imbalanced.

- -Histograms:

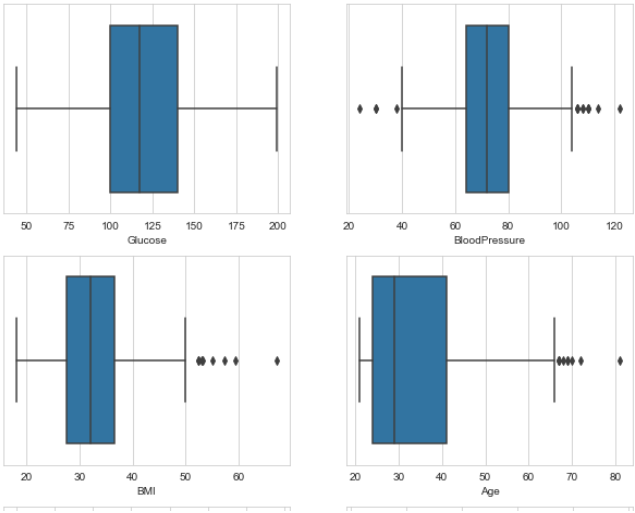
```
In [18]: df.hist(bins=10,figsize=(10,10))
         plt.show()
```

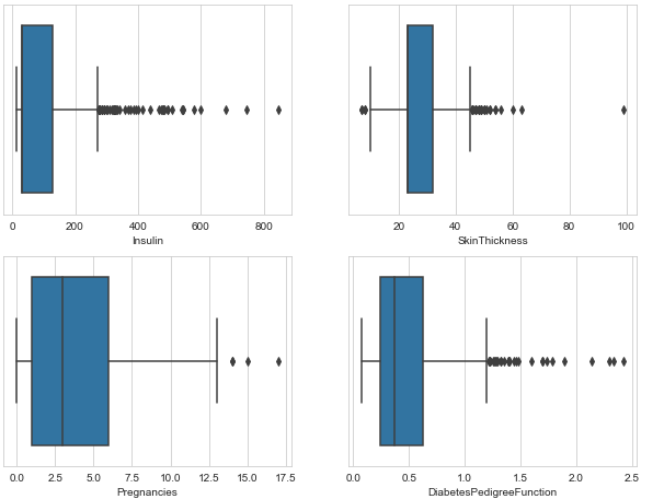


Conclusion : We observe that only glucose and Blood Pressure are normally distributed. Others are skewed and have outliers.

- -Box Plot:

```
In [22]: plt.figure(figsize=(16,12))
         sns.set_style(style='whitegrid')
         plt.subplot(3,3,...)
         sns.boxplot(x='...',data=df)
```

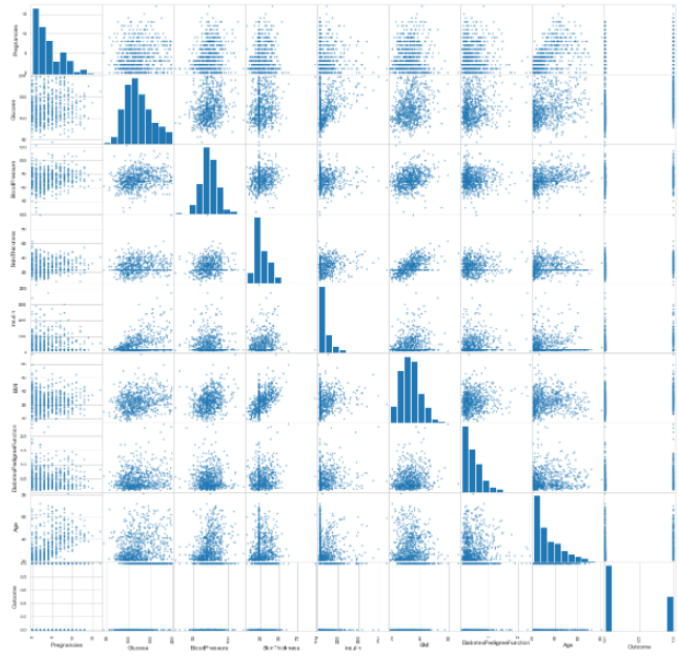




Outliers are unusual values in the dataset, and they can deform statistical analyses and break their assumptions. Hence it is of high importance to deal with them. In this case removing outliers can cause data loss so we must deal with it using various scaling and transformation techniques.

- -Scatter plots:

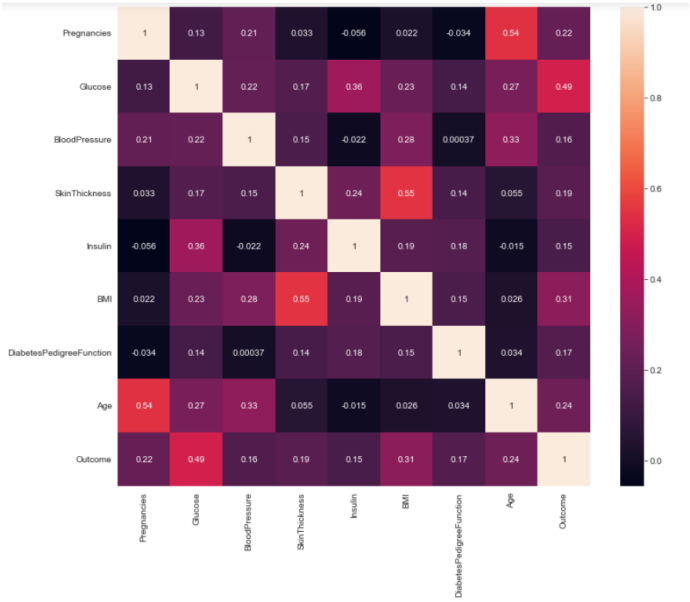
```
In [23]: from pandas.plotting import scatter_matrix
         scatter_matrix(df,figsize=(20,20));
         # we can come to various conclusion looking at these plots for example
         #if you observe 5th plot in pregnancies with insulin, you can conclude that
         #women with higher number of pregnancies have lower insulin
```



- -Heat Map: Diabetes Correlation of Features

Correlation Coefficient: Helps us to find out the relationship between two features. The value of Correlation Coefficient can be between -1 to +1.It shows the strength of the correlation between two properties. 1 means that they are highly correlated and 0 means no correlation. A heat map is a two-dimensional representation of information with the help of colors.

```
In [29]: correlation = df.corr()
         plt.figure(figsize = (12,10))
         sns.heatmap(correlation,annot = True)
```

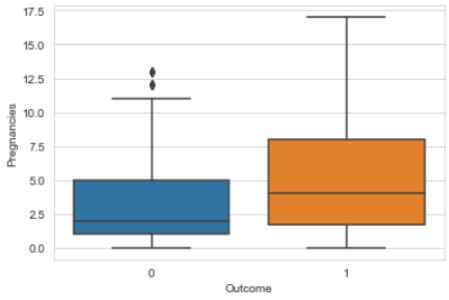


CONCLUSION :We can observe that Glucose, BMI and Age are the most correlated with Outcome. BloodPressure, Insulin, DiabetesPedigreeFunction are the least correlated. We can see also a strong correlation between pregnancies and age; between BMI and SkinThickness.

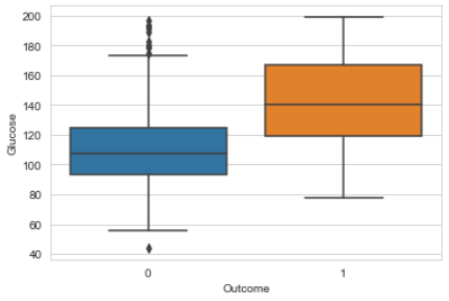
VII. For Answering Technical Questions

- In order to see the distributions more easily, let's project box-plots that divide all the other features into two parts according to the outputs of the 'outcome' column: '0' and '1'.

```
[34]: sns.boxplot(x='Outcome',y='Pregnancies',data=df)
      plt.show()
```

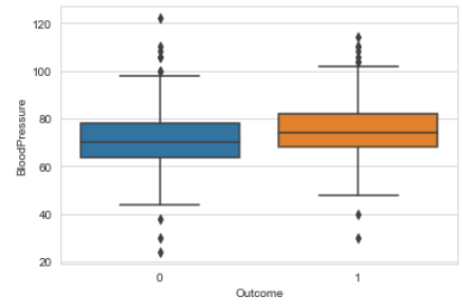



```
[35]: sns.boxplot(x='Outcome',y='Glucose',data=df)
      plt.show()
```

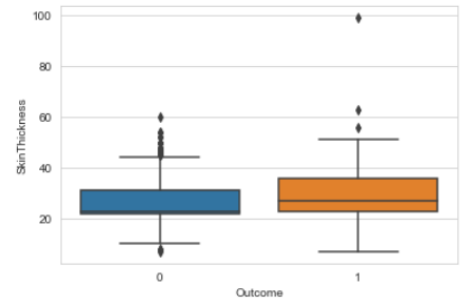


- In the first image, we can say that people with diabetes have a higher average number of pregnancy. However, there is no definite correlation between them. As can be seen in the second one, people with diabetes have a higher glucose level than those who do not have diabetes.This means that there is a high correlation between outcome and glucose. We have seen this relationship in the heat map as well.

```
[36]: sns.boxplot(x='Outcome',y='BloodPressure',data=df)
      plt.show()
```

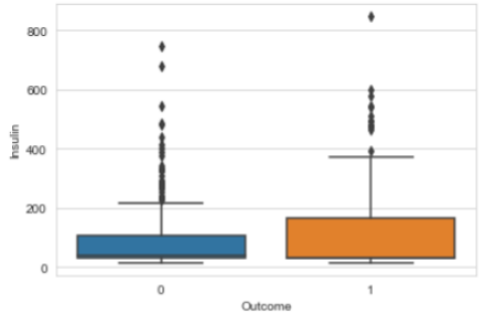



```
[37]: sns.boxplot(x='Outcome',y='SkinThickness',data=df)
      plt.show()
```

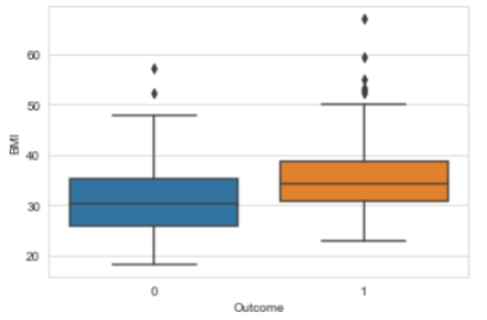


- In the third picture, we couldn't see any noticeable difference. However, we can say that the blood pressure of people with diabetes is a little higher. Also the box-plot of people without diabetes , has more scattered outliers. For the fourth image, we can't make any inferences.

```
[38]: sns.boxplot(x='Outcome',y='Insulin',data=df)
      plt.show()
```

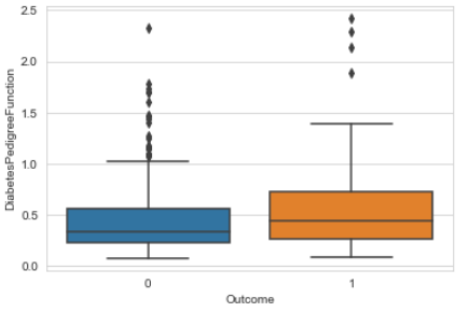



```
[39]: sns.boxplot(x='Outcome',y='BMI',data=df)
      plt.show()
```

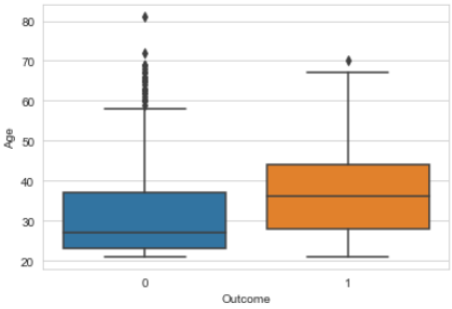


- As for the BMI box-plot, people with diabetes have higher body mass index.This means that; there is a strong correlation between outcome and BMI. We have seen this relationship in the heat map as well.

```
[40]: sns.boxplot(x='Outcome',y='DiabetesPedigreeFunction'
      plt.show()
```

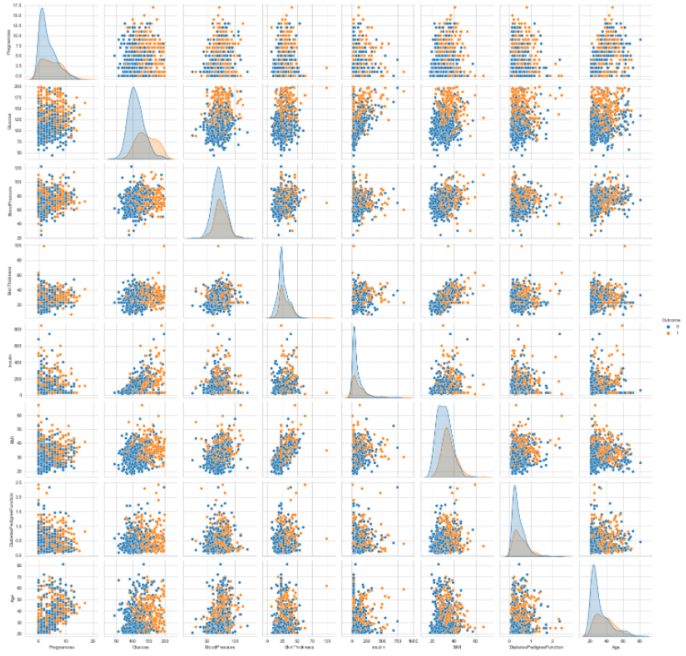



```
[41]: sns.boxplot(x='Outcome',y='Age',data=df)
      plt.show()
```



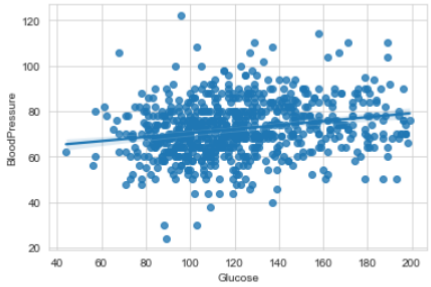
- For the last image, we can say that people who have diabetes are older than the people without diabetes. So there is a correlation between outcome and age as we saw in the heat map.

Also, we can project the scatter matrix according to 'outcome' outputs. Thus we can see all in one.

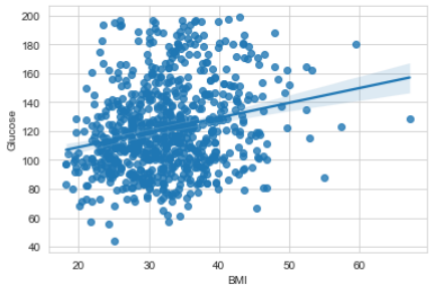


We can look into scatter plot between features with regression line. We can draw a graph between blood pressure and glucose.Also; Since Glucose and BMI are related to outcome, let's plot a linear regression between the two.

```
[43]: # scatter plot between glucose and  Blood pressure with r
      sns.regplot(x = "Glucose",y = "BloodPressure",data = df)
      plt.show()
```




```
[44]: sns.regplot(x='BMI', y= 'Glucose', data=df)
      plt.show()
```



VIII. Normalizing Data

A. Checking for normality of predictors

Pregnancies
1.60925711873724e-21
Glucose
1.73290310284045e-11
BloodPressure
5.25543634152065e-06
SkinThickness
1.75175408046787e-21
Insulin
7.91433227770875e-34
BMI
6.2474430694009e-09
DiabetesPedigreeFunction
2.47750565380547e-27
Age
2.40227398977845e-24

Above we can see that all of the p.values returned from our shapiro wilks test are less than 0.05. Based on this we reject the null hypothesis and conclude that are independent variables are not normally distributed.

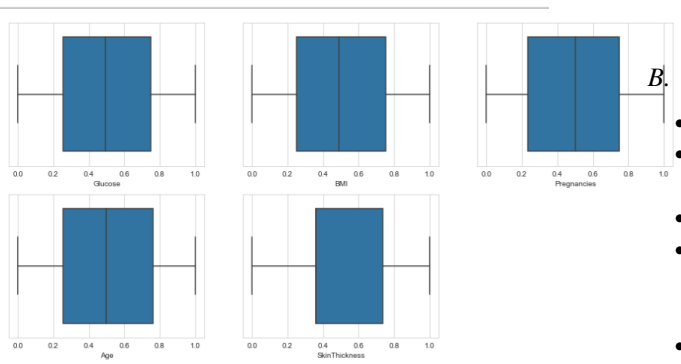
B. Handling Outliers

BloodPressure, Insulin, DiabetesPedigreeFunction are the least correlated with outcome. Hence we can drop them for looking into the related ones. Now, we will scaling the data. when i use quantile transformer, i reached better solutions than standard scaler (z test). So i will apply quantile transformer for the realisation of first hypothesis. Quantile Transformer :This method transforms the features to follow a uniform or a normal distribution. Therefore, for a given feature, this transformation

tends to spread out the most frequent values. It also reduces the outliers. Here are the results:

[58]:

	Pregnancies	Glucose	SkinThickness	BMI	Age	Outcome
0	0.747718	0.810300	0.801825	0.591265	0.889831	1.0
1	0.232725	0.091265	0.644720	0.213168	0.558670	0.0
2	0.863755	0.956975	0.357888	0.077575	0.585398	1.0
3	0.232725	0.124511	0.357888	0.284224	0.000000	0.0
4	0.000000	0.721643	0.801825	0.926988	0.606258	1.0

- 

Checking to see results of shapiro test on transformed data:

Pregnancies
7.43861016640543e-17

Glucose
0.00023608142273188

SkinThickness
1.75175408046787e-21

BMI
0.167398142826797

Age
2.40227398977845e-24

In our transformed dataset we can see that BMI now fails to reject the null hypothesis(meaning we conclude it is normally distributed) however the rest of the predictor variables still aren't.

IX. SUMMARY

A. Answers For Hypothesis

- 1st Hypothesis:
- We reject the Null Hypothesis. All independent variables are not normally distributed.
- 2nd Hypothesis:
- We reject the Null Hypothesis.High glucose level can be a sign of having diabetes. We have a high relationship between glucose level and outcome.
- 3rd Hypothesis:
- We reject the Null Hypothesis.
- H1: High number of pregnancy times can be a sign of having diabetes. We get a high relationship between pregnancies and outcome. People with diabetes are usually have multiple births.
- 4th Hypothesis:
- We can't reject the Null Hypothesis. Low blood pressure can't be a sign of having diabetes. There is no strong correlation between BloodPressure and outcome.
- 5th Hypothesis:
- H0: We reject the Null Hypothesis.
- H1: Body mass index can be a sign of having diabetes.People with diabetes have higher body mass index.This means that: there is a strong correlation between outcome and BMI.

B. Answers for Technical Questions

- Are Predictors normally distributed?
- We observe that only glucose and Blood Pressure are normally distributed. Others are skewed and have outliers.
- Can high glucose level be a sign of having diabetes?
- people with diabetes have a higher glucose level than people without diabetes.This means that there is a high correlation between outcome and glucose.
- Can high number of pregnancy times be a sign of having diabetes?
- we can say that people with diabetes have a higher average number of pregnancy. However, there is no definite correlation between them.
- Can low blood pressure be a sign of having diabetes?
- We can't make such an inference.(no)
- Can low skin thickness be a sign of having diabetes?
- We can't make such an inference.(no)
- Can low insulin level be a sign of having diabetes?
- There is no obvious correlation between them.(no)
- Can body mass index be a sign of having diabetes?
- People with diabetes have higher body mass index.This means that: there is a strong correlation between outcome and BMI.
- Can Diabetes Pedigree Function be a sign of having diabetes?
- We can't make such an inference.(no)
- Do diabetic patients have high glucose levels and high number of pregnancy times?
- We could say that people who have multiple births and have high glucose level are usually suffer from diabetes.
- How is the age and pregnancy distribution for diabetic patients?
- People with diabetes are usually older and have multiple births.
- Is there a Collinearity in all Predictors?
- Obviously not.

C. Conclusion

As a result, we analyzed our data to answer our high-level question ("Can we be able to predict if someone suffers from diabetes before diagnosing her/him?". Although we found some medical features that were highly correlated with diabetes(Glucose, BMI and Age), we could not conclude that the one with a high value is definitely diabetic. However, in the results we obtained, if the factors affecting the state of

diabetes are present in the individual at the same time (ex.: an elderly person with a high glucose level and who has given multiple) birth, it is possible to keep the risk of diabetes.

X. References

[1] http://www.arogyaworld.org/wp-content/uploads/2010/10/ArogyaWorld-IndiaDiabetes-FactSheets-CGI2013-web.pdf

[2] Dataset Source: UCI Machine Learning—Repository:https://www.kaggle.com/uciml/pima-indians-diabetes-database

[3] https://towardsdatascience.com/pima-indians-diabetes-prediction-knn-visualization-5527c154afff

[4] https://www.kaggle.com/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed/notebook

[5] https://github.com/boosuro/diabetes-prediction-with-knn/blob/master/diabetes-prediction-with-knn.ipynb