DIABETES PROJECT DATA ANALYSIS

Submitted to Ms.Shalini Kumari

Submitted By:
Namratha K
Batch Number: 4868
Banglore

TABLE OF CONTENTS

Abstract:- Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying Techniques. Provide better result for prediction by constructing models from datasets collected from patients

Introduction

Data analysis is all about getting and overall understanding of data. It is mainly done to find it's properties , patterns and visualizations

In this blog we are using python as our programming language for the analysis purpose. Python has a wide variety of libraries like pandas , seaborn , numpy , matplotlib which can be used for this purpose. We are using the Pima Indians Dataset where it shows the various diagnostic factors influencing the diabetes.

Brief Introduction to the used libraries

As discussed above , we are going to use the following libraries to perform different operations on the data

- Numpy
  - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
  - http://www.numpy.org/
- Pandas
  - pandas is a python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal.
  - https://pandas.pydata.org/
- Seaborn
  - Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
  - https://seaborn.pydata.org/
- Matplotlib
  - Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
  - https://matplotlib.org/

Understanding the data :-

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

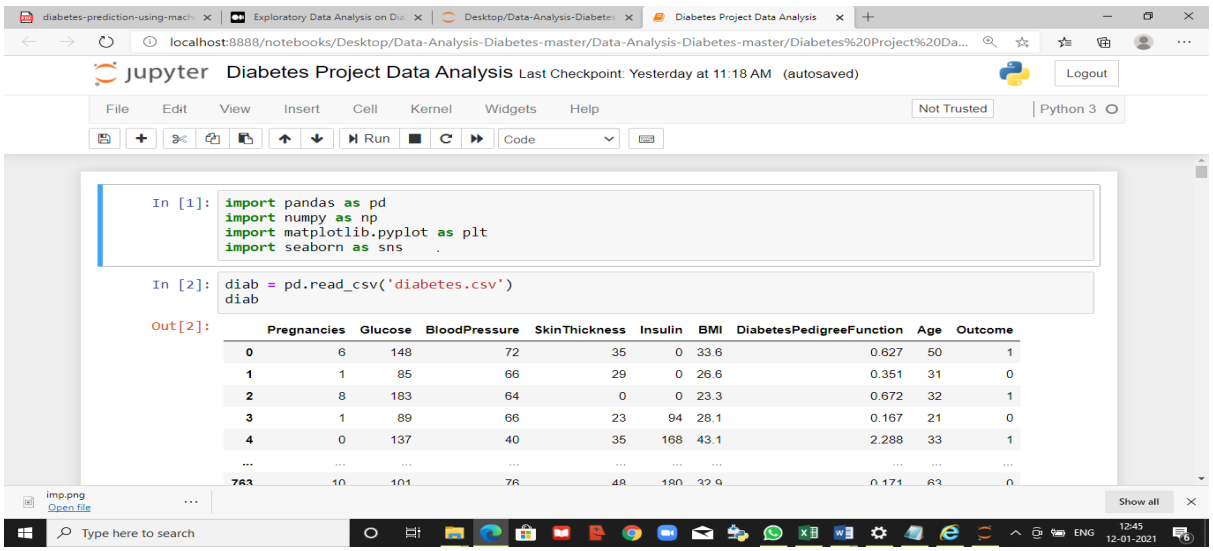
Content

The datasets consists of several medical predictor variables and one target variable, Outcome. Columns are following :-

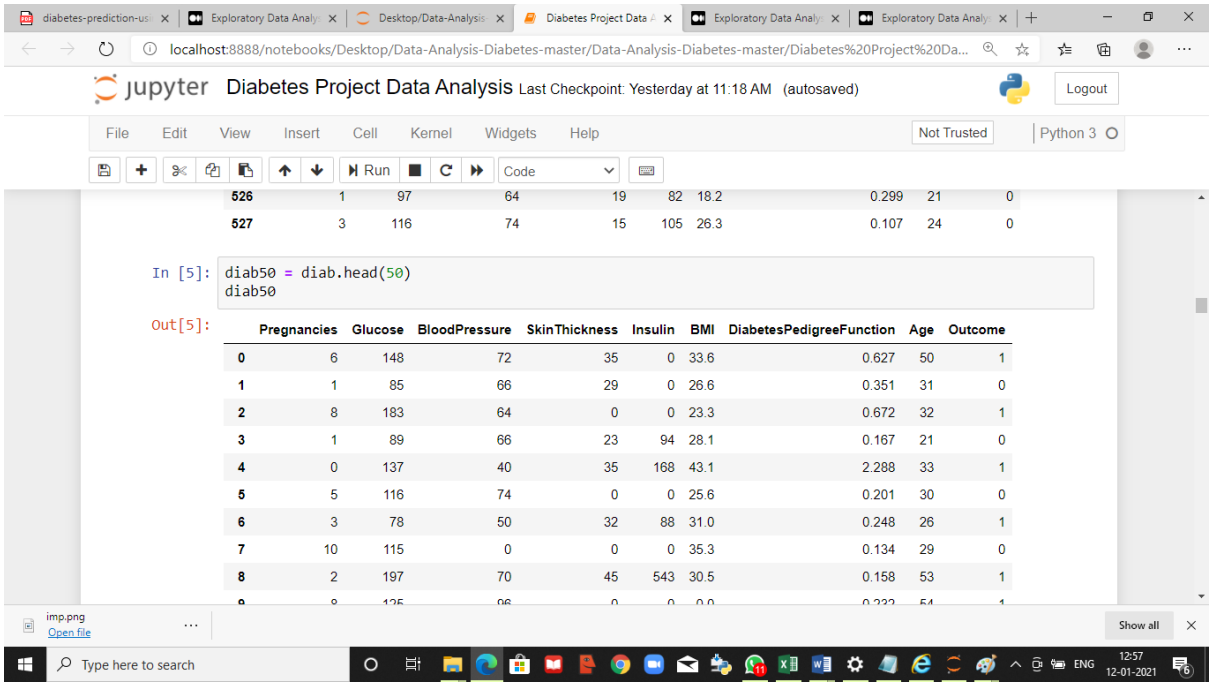
1. Pregnancies :- Number of times pregnant
2. Glucose:- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Blood Pressure:- Diastolic blood pressure (mm Hg)
4. Skin Thickness:- Triceps skin fold thickness (mm)
5. Insulin:- 2-Hour serum insulin (mu U/ml)
6. BMI:- Body mass index (weight in kg/(height in m)^2)
7. Diabetes Pedigree Function:- Diabetes pedigree function
8. Age:-Age in years
9. Outcome:- Class variable (0 or 1) 268 of 768 are 1, the others

are 0

1) Importing the necessary libraries and Loading the dataset



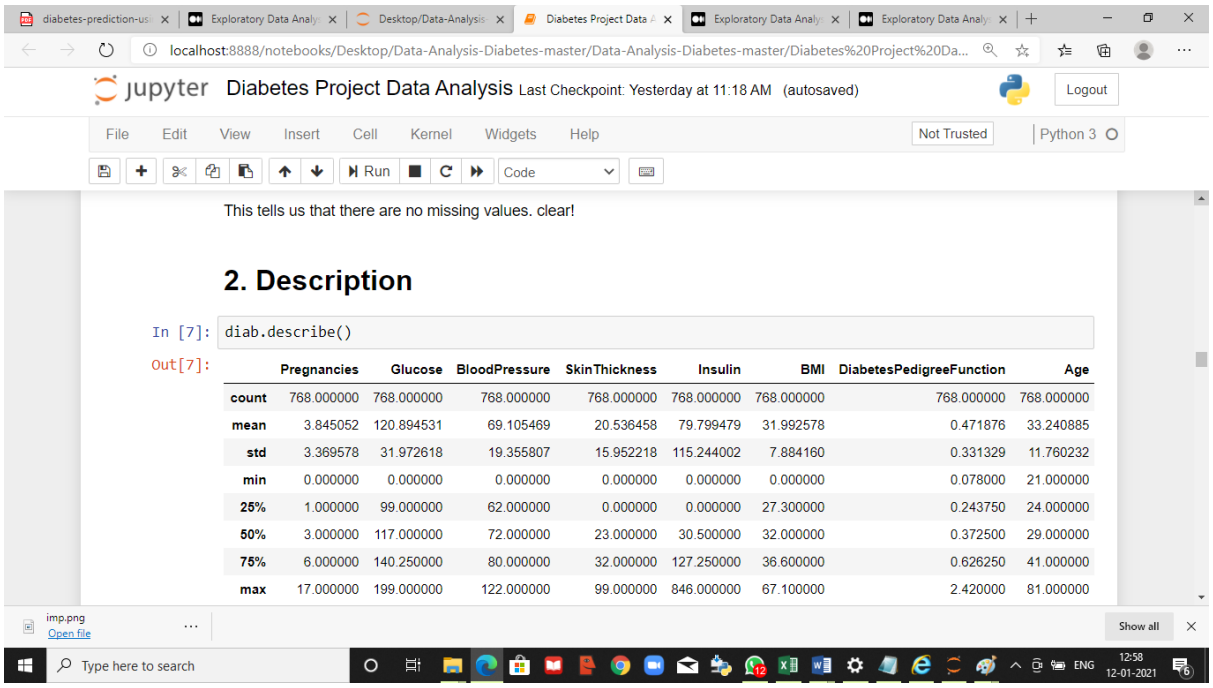
2) Collecting Basic Information about the Data



	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
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1

It shows that there are eight independent variables (Pregnancies,Glucose,Blood Pressure , SkinThickness ,Insulin, BMI, DiabetesPedigreeFunction, Age) and one dependent variable (Outcome).

3)Descriptive statistics using describe method as shown below



This tells us that there are no missing values. clear!

2. Description

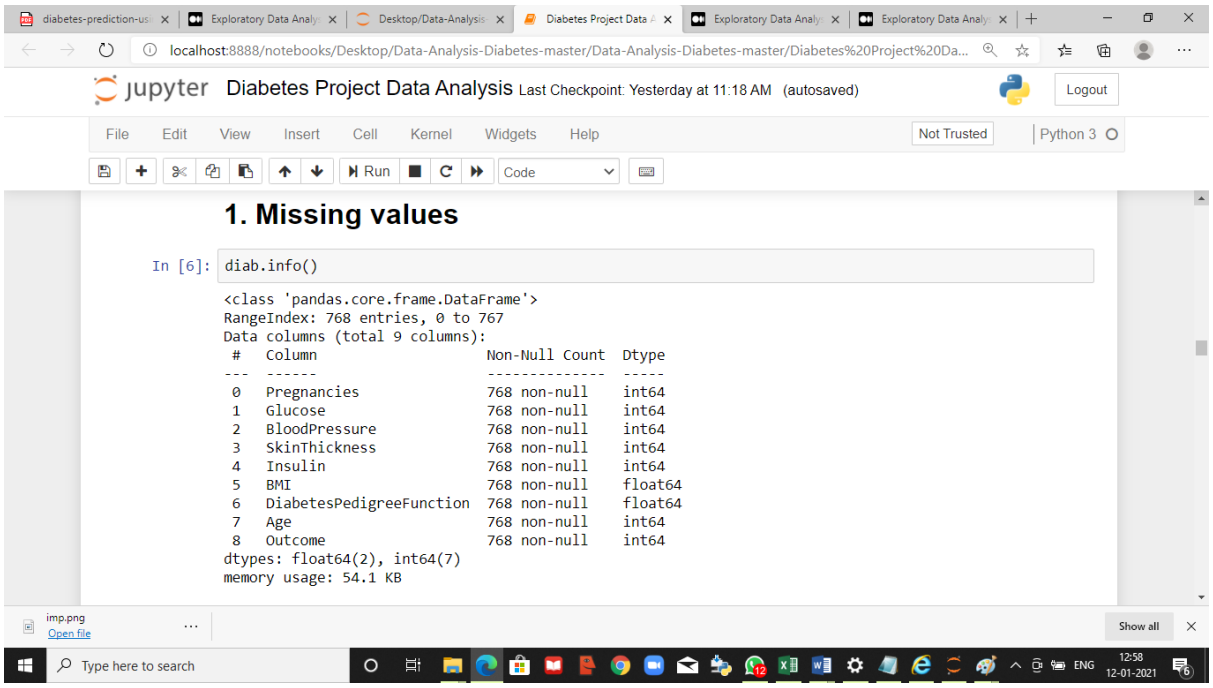
In [7]:
```
diab.describe()
```

Out[7]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	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
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

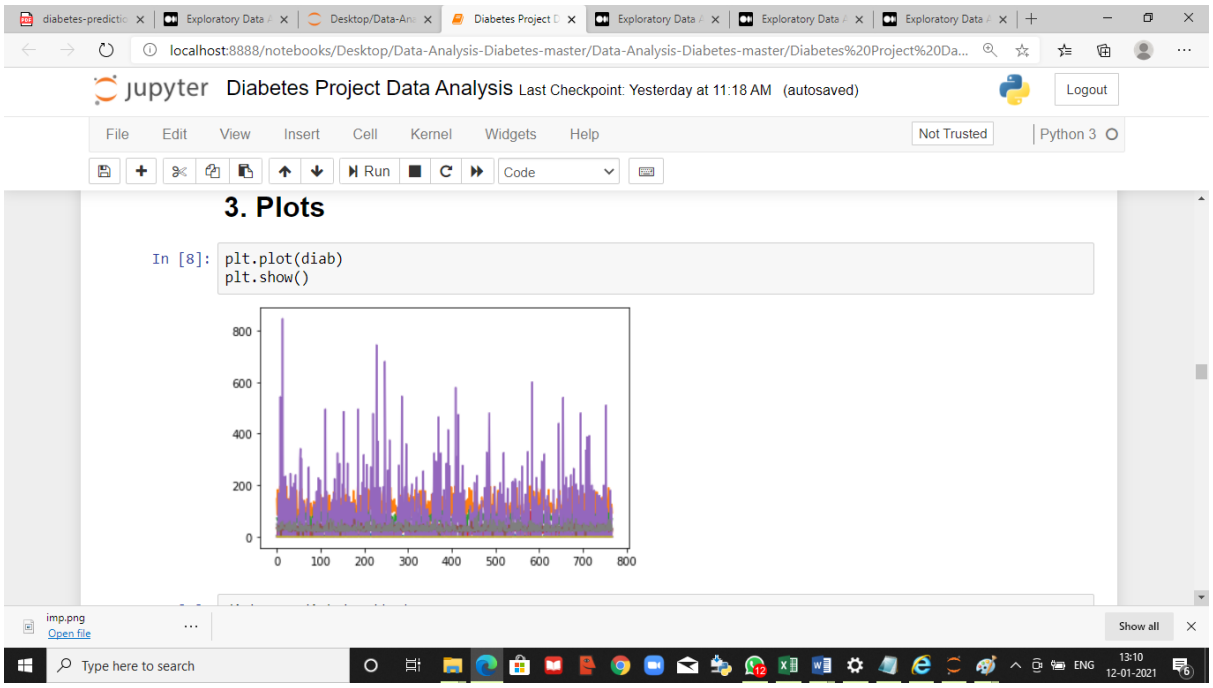
4) Handling the missing values

We need to check the presence of the missing values and need to replace them with mean, median and mode. Sometimes we have null values in the form of 0 , so we need to convert them to NaN and then replace them accordingly. The missing values can be removed also but it should be less than 5 percent of the whole dataset



5)Data Visualization

Visualizing data in different type of graphs will provide us with greater insights into our data. We will explore different options on visualizing our data and find out any patterns between within it



6) Analysis of each parameter w.r.t outcome parameter

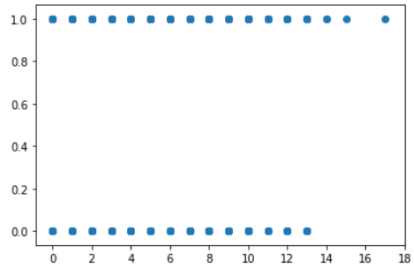
“Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships. It’s major purpose is to describe; it takes data, summarizes that data and finds patterns in the data. This means here we deal and with only one variable or column of the data and try to find out it’s nature

Analysis of ‘Pregnancies’ parameter

jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

A. Between parameters and the target.

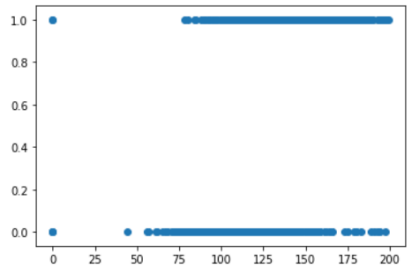
In [11]:
```
# pregnancies and outcome
plt.scatter(diab['Pregnancies'], diab['Outcome'])
plt.xticks([0,2,4,6,8,10,12,14,16,18])
plt.show()
```



jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

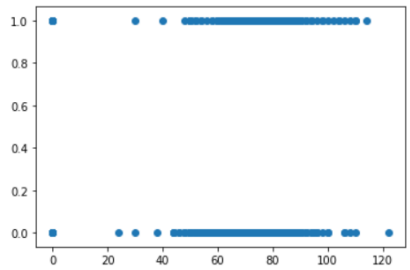
In [12]:
```
# diab['Pregnancies'].isin([15])
```

In [13]:
```
# glucose and outcome
plt.scatter(diab['Glucose'], diab['Outcome'])
plt.show()
```



jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

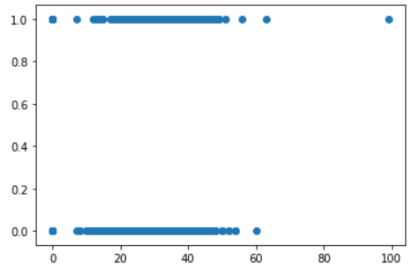
In [14]:
```
# BloodPressure and outcome
plt.scatter(diab['BloodPressure'], diab['Outcome'])
plt.show()
```



In [15]:
```
# SkinThickness and outcome
```

jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

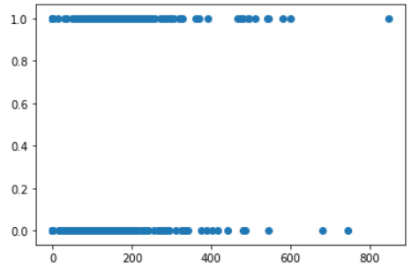
In [15]:
```
# SkinThickness and outcome
plt.scatter(diab['SkinThickness'], diab['Outcome'])
plt.show()
```



In [16]:
```
# Insulin and outcome
```

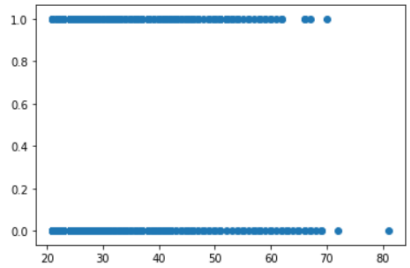
jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

In [16]:
```
# Insulin and outcome
plt.scatter(diab['Insulin'], diab['Outcome'])
plt.show()
```

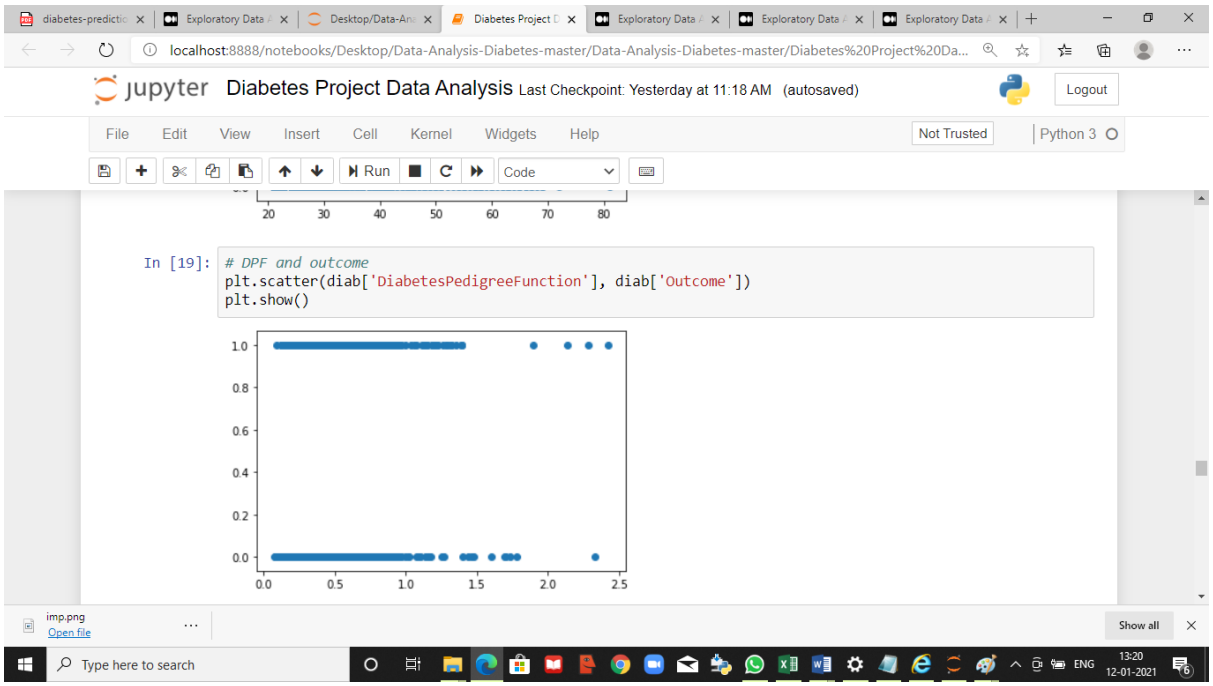


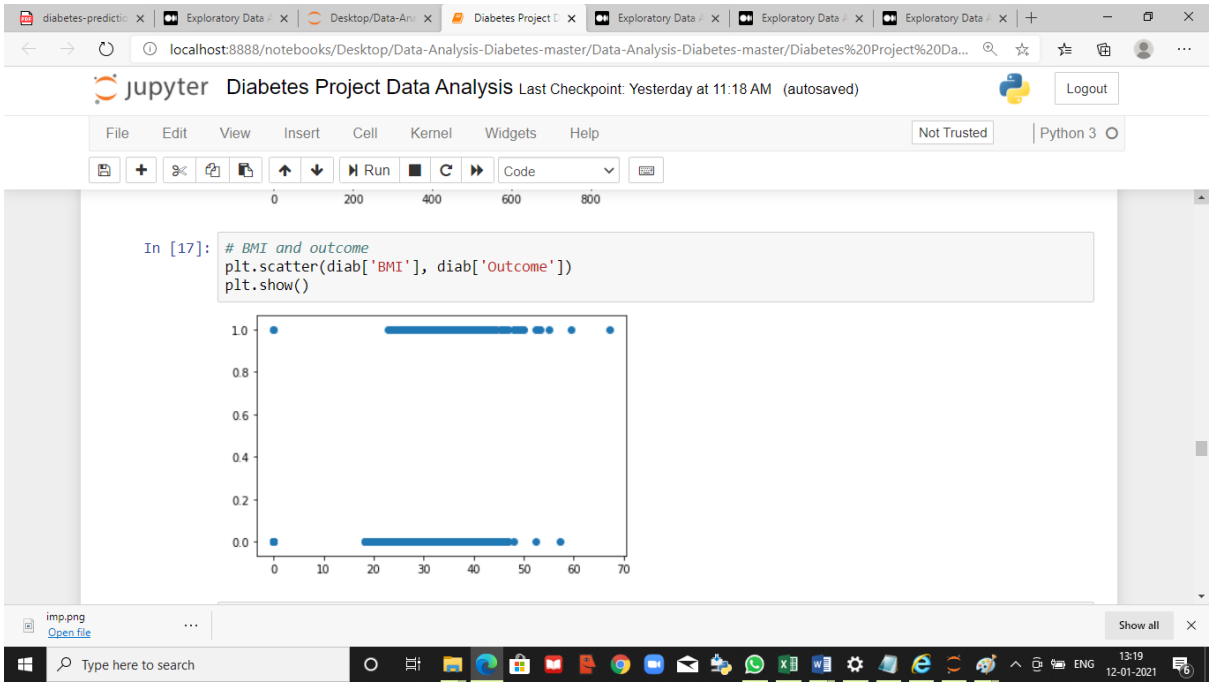
jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

In [18]:
```
# Age and outcome
plt.scatter(diab['Age'], diab['Outcome'])
plt.show()
```



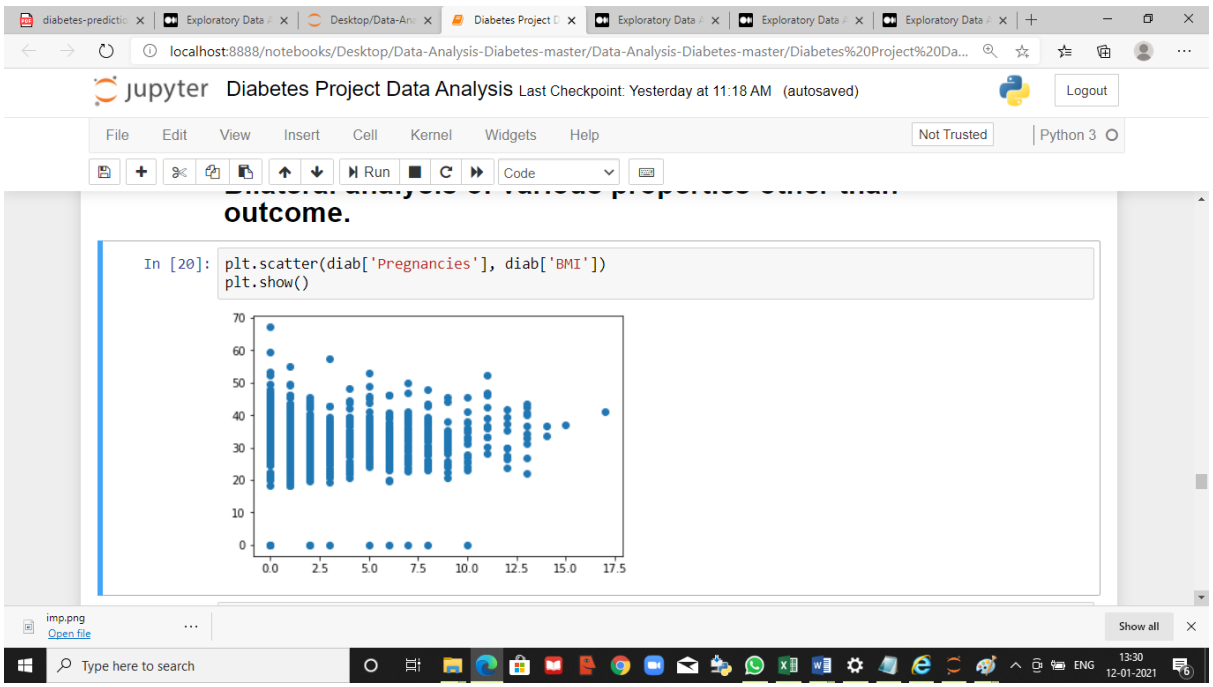
In [19]:
```
# DPF and outcome
plt.scatter(diab['DiabetesPedigreeFunction'], diab['Outcome'])
```

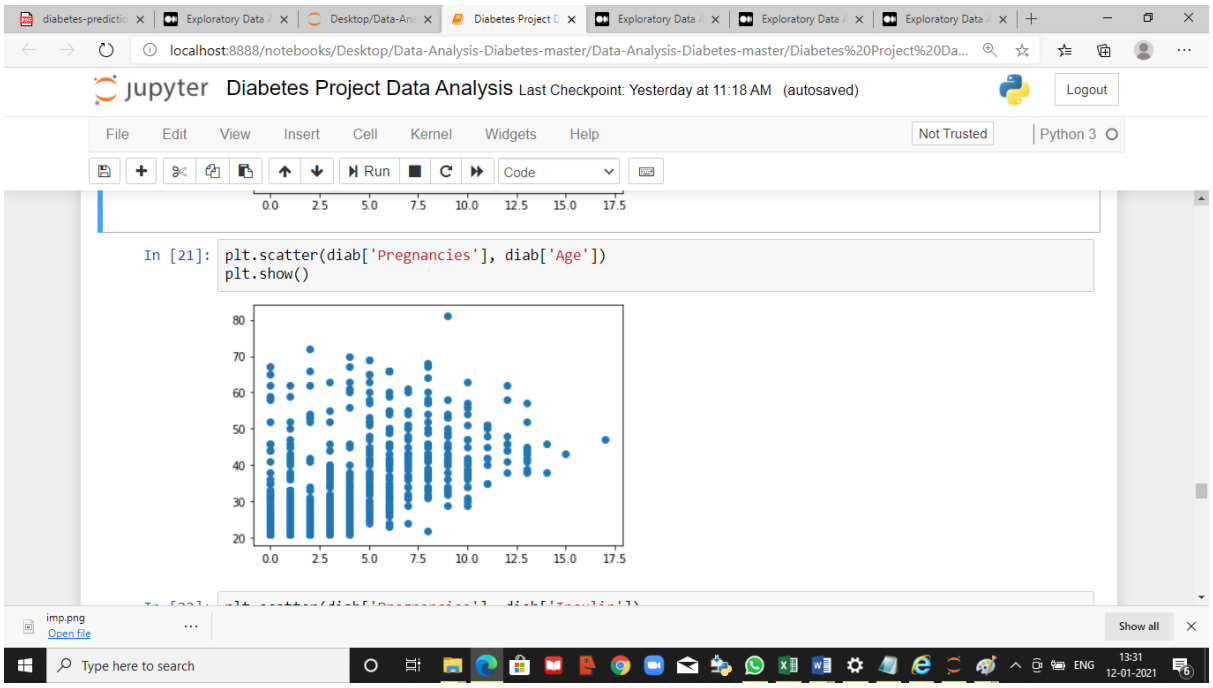

jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [19]:
# DPF and outcome
plt.scatter(diab['DiabetesPedigreeFunction'], diab['Outcome'])
plt.show()

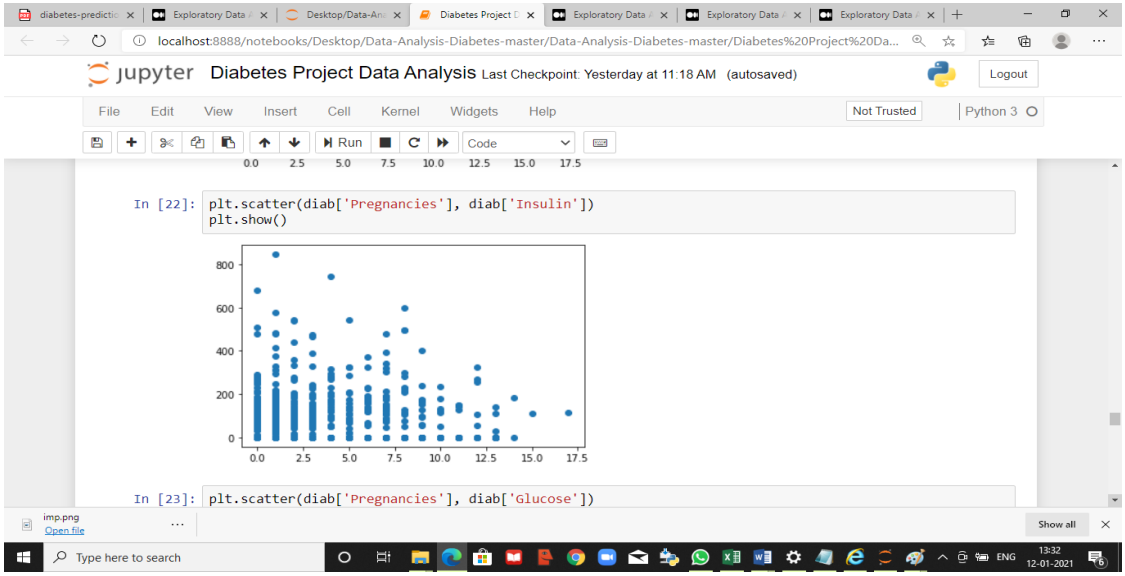

jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [17]:
# BMI and outcome
plt.scatter(diab['BMI'], diab['Outcome'])
plt.show()

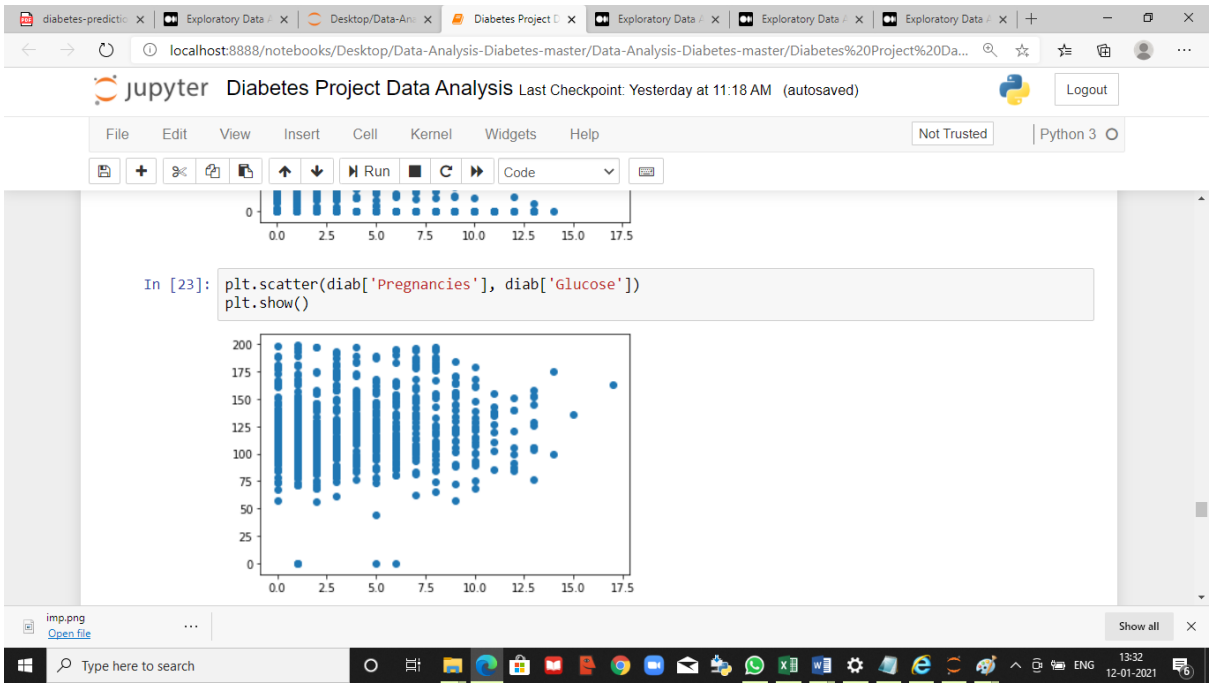
1. Pregnancies and age have some kind of a linear line.
2. BloodPressure and age have little relation. Most of the aged people have BloodPressure.
3. Insulin and Glucose have some relation.

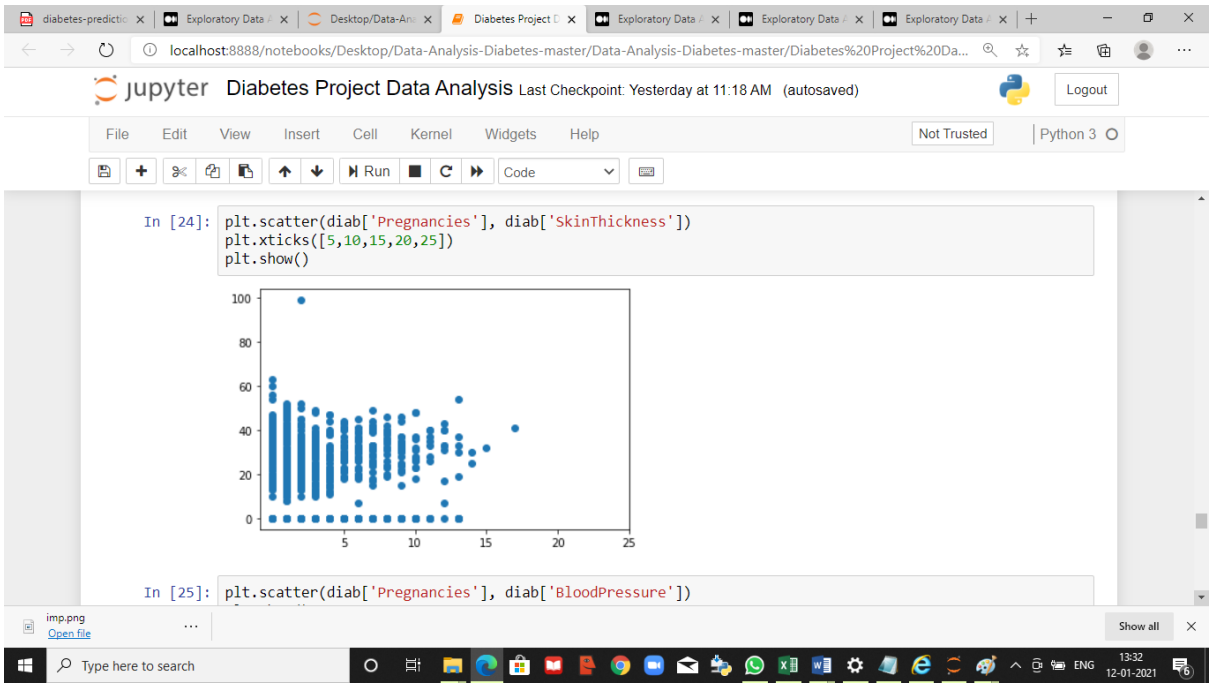
7) Bilateral analysis of various properties other than outcome

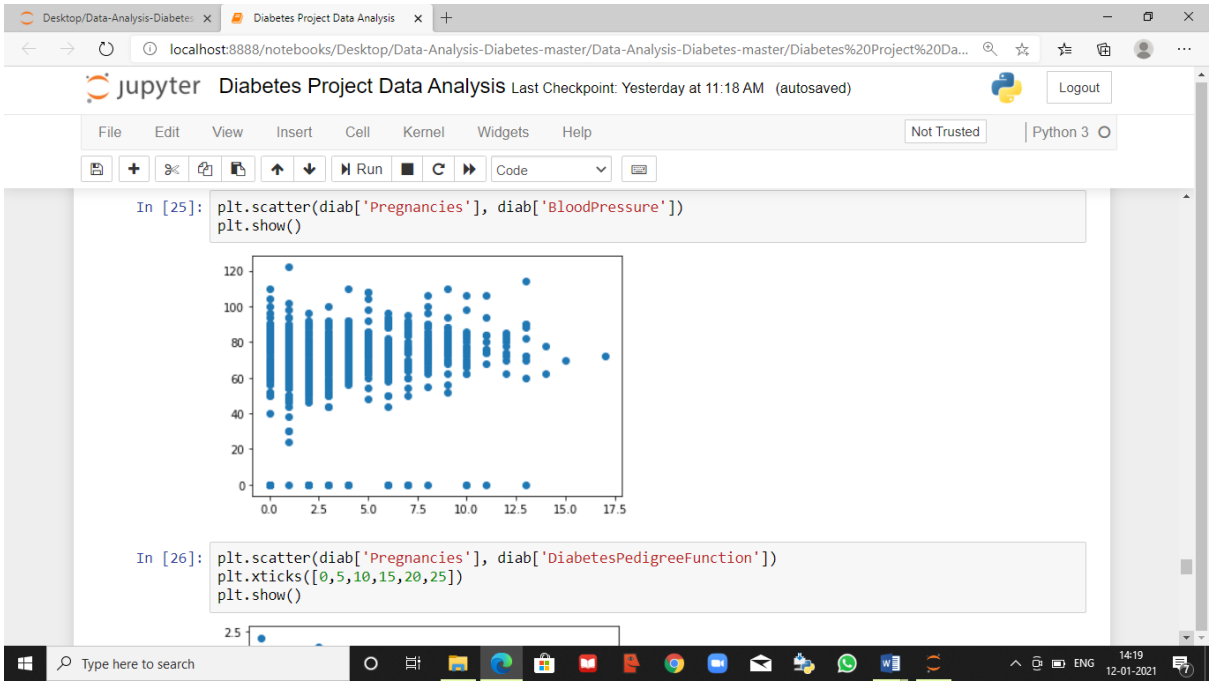


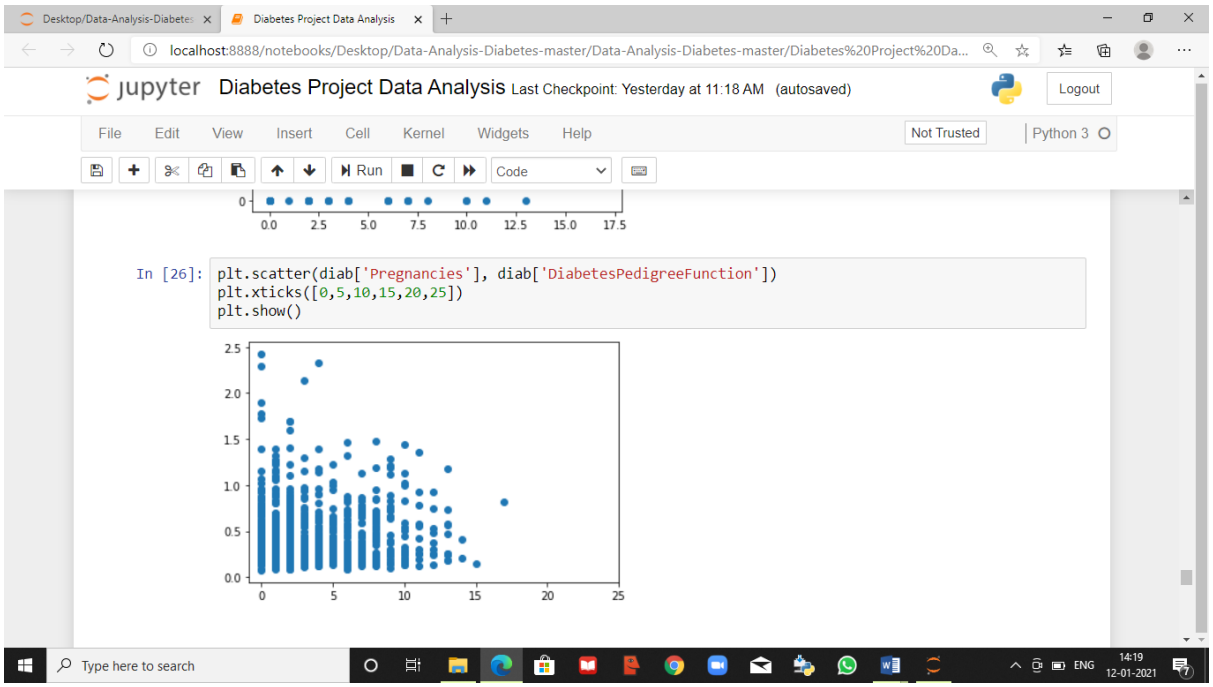

jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [21]:
plt.scatter(diab['Pregnancies'], diab['Age'])
plt.show()


jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [22]:
plt.scatter(diab['Pregnancies'], diab['Insulin'])
plt.show()
In [23]:
plt.scatter(diab['Pregnancies'], diab['Glucose'])


jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
In [23]:
plt.scatter(diab['Pregnancies'], diab['Glucose'])
plt.show()


jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
In [24]:
plt.scatter(diab['Pregnancies'], diab['SkinThickness'])
plt.xticks([5,10,15,20,25])
plt.show()
In [25]:
plt.scatter(diab['Pregnancies'], diab['BloodPressure'])


Desktop/Data-Analysis-Diabetes
Diabetes Project Data Analysis
localhost:8888/notebooks/Desktop/Data-Analysis-Diabetes-master/Data-Analysis-Diabetes-master/Diabetes%20Project%20Da...
jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [25]: plt.scatter(diab['Pregnancies'], diab['BloodPressure'])
plt.show()
In [26]: plt.scatter(diab['Pregnancies'], diab['DiabetesPedigreeFunction'])
plt.xticks([0,5,10,15,20,25])
plt.show()
Type here to search
ENG
14:19
12-01-2021

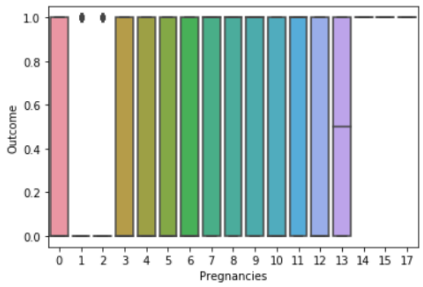

Desktop/Data-Analysis-Diabetes
Diabetes Project Data Analysis
localhost:8888/notebooks/Desktop/Data-Analysis-Diabetes-master/Data-Analysis-Diabetes-master/Diabetes%20Project%20Da...
jupyter Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [26]: plt.scatter(diab['Pregnancies'], diab['DiabetesPedigreeFunction'])
plt.xticks([0,5,10,15,20,25])
plt.show()
Type here to search
ENG
14:19
12-01-2021

8) Box Plot

Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

In [27]:
```
sns.boxplot(x='Pregnancies',y='Outcome',data= diab)
```

Out[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1be62950>
```



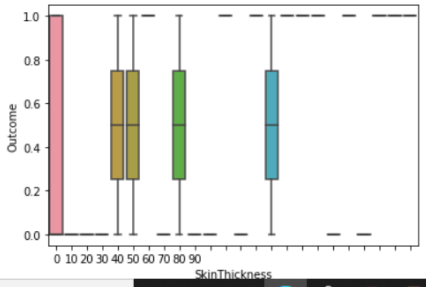
In [28]:
```
# list = np.arange(1,200,1)
# glucose = sns.boxplot(x='Glucose',y='Outcome',data= diab)
# glucose.set(xticklabels=list)
# glucose
```

Diabetes Project Data Analysis Last Checkpoint: Yesterday at 11:18 AM (autosaved)

In [ ]:
```
# sns.boxplot(x='BMI',y='Outcome',data= diab)
```

In [31]:
```
list = np.arange(0,100,10)
st = sns.boxplot(x='SkinThickness',y='Outcome',data= diab50)
st.set(xticklabels=list)
st
```

Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1c147310>
```



Conclusion

Hence we can say from the above that EDA is very important as it helps us in achieving the following :-

1. It helps in detection of mistakes (like missing values and outliers)
2. It determines relationships between explanatory variables
3. Assessing the direction and rough size of relationships between explanatory and outcome variables.
4. It makes our data ready for machine learning algorithm