Final Project

Diabetes Dataset Report

Srijesh Reddy Yarram

15th October 2022

Executive summary:

I discovered that the diabetic.csv dataset contained enough evidence as to say that most of the women in the dataset have less than 5 pregnancies. More the insulin or more the BMI more is the chance of getting diabetes. The people who got affected by diabetes has more glucose in their bodies. Most of the women who have diabetes are above the age of 30.

Definition:

The dataset is originally from the national Institute of Diabetes and Digestive and Kidney Diseases. This dataset has 9 variables which the institute felt they can be used to predict whether a patient has diabetes. Here we would learn about how the attributes are contributing towards learning whether the patient has diabetes.

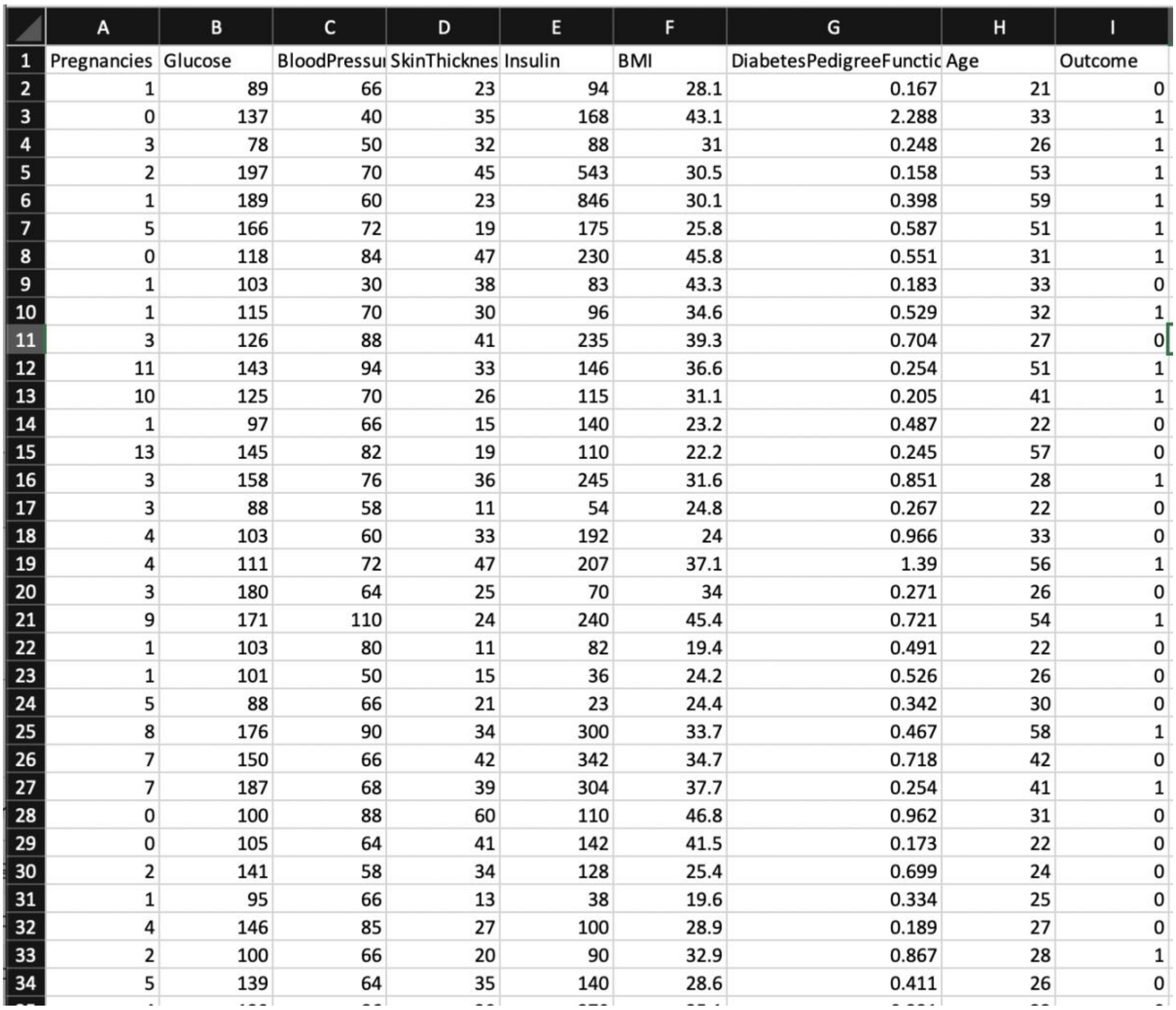
Preparing for the data:

The link for the data set is (https://www.kaggle.com/datasets/mathchi/diabetes-data-set) which contains 768 instances of the people who have been diagnosed and the people who haven't been diagnosed with diabetes. Table 1.0 summarizes the variables contained in the dataset.

Table 1.0 Description of the variables in the diabetes Dataset

Variable Name	Variable Type
Glucose	continuous
Number of times pregnant	multi-valued discrete (1-17)
Blood Pressure	continuous
Skin thickness	continuous
Insulin	continuous
BMI	continuous
Diabetes Pedigree Function	continuous
Age	discrete
Outcome	multi-valued discrete 0-no diabetes, 1 - diabetes

The table contains the variables and the variable type, and here in fig1.1 we can find the snippet pf the dataset.



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressu	SkinThicknes	Insulin	BMI	DiabetesPedigreeFunctic	Age	Outcome
2	1	89	66	23	94	28.1	0.167	21	0
3	0	137	40	35	168	43.1	2.288	33	1
4	3	78	50	32	88	31	0.248	26	1
5	2	197	70	45	543	30.5	0.158	53	1
6	1	189	60	23	846	30.1	0.398	59	1
7	5	166	72	19	175	25.8	0.587	51	1
8	0	118	84	47	230	45.8	0.551	31	1
9	1	103	30	38	83	43.3	0.183	33	0
10	1	115	70	30	96	34.6	0.529	32	1
11	3	126	88	41	235	39.3	0.704	27	0
12	11	143	94	33	146	36.6	0.254	51	1
13	10	125	70	26	115	31.1	0.205	41	1
14	1	97	66	15	140	23.2	0.487	22	0
15	13	145	82	19	110	22.2	0.245	57	0
16	3	158	76	36	245	31.6	0.851	28	1
17	3	88	58	11	54	24.8	0.267	22	0
18	4	103	60	33	192	24	0.966	33	0
19	4	111	72	47	207	37.1	1.39	56	1
20	3	180	64	25	70	34	0.271	26	0
21	9	171	110	24	240	45.4	0.721	54	1
22	1	103	80	11	82	19.4	0.491	22	0
23	1	101	50	15	36	24.2	0.526	26	0
24	5	88	66	21	23	24.4	0.342	30	0
25	8	176	90	34	300	33.7	0.467	58	1
26	7	150	66	42	342	34.7	0.718	42	0
27	7	187	68	39	304	37.7	0.254	41	1
28	0	100	88	60	110	46.8	0.962	31	0
29	0	105	64	41	142	41.5	0.173	22	0
30	2	141	58	34	128	25.4	0.699	24	0
31	1	95	66	13	38	19.6	0.334	25	0
32	4	146	85	27	100	28.9	0.189	27	0
33	2	100	66	20	90	32.9	0.867	28	1
34	5	139	64	35	140	28.6	0.411	26	0

fig 1.1 contains the portion of the dataset regarding diabetes

Handling Missing Data:

The attribute glucose has 4 instances of null values, Blood pressure has 35 null values, skin thickness has 227 null values, insulin has 374 null values, BMI has 11 null values. we use the relevant data by clicking on the ‘edit filter’ and we uncheck the box ‘include null values’ Or we can also filter the data in the excel sheet and delete the rows in the null values and then save it back in the .csv file.

Data Outliers:

There are no data outliers in this Data.

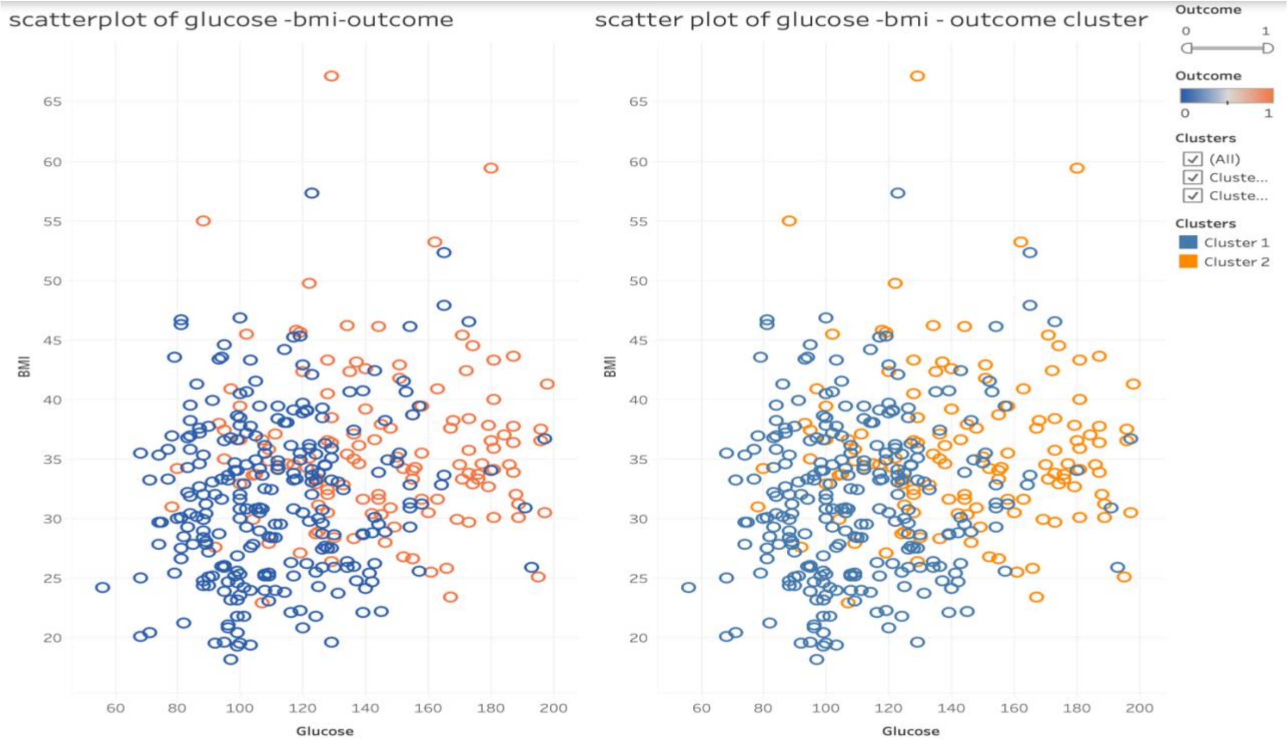
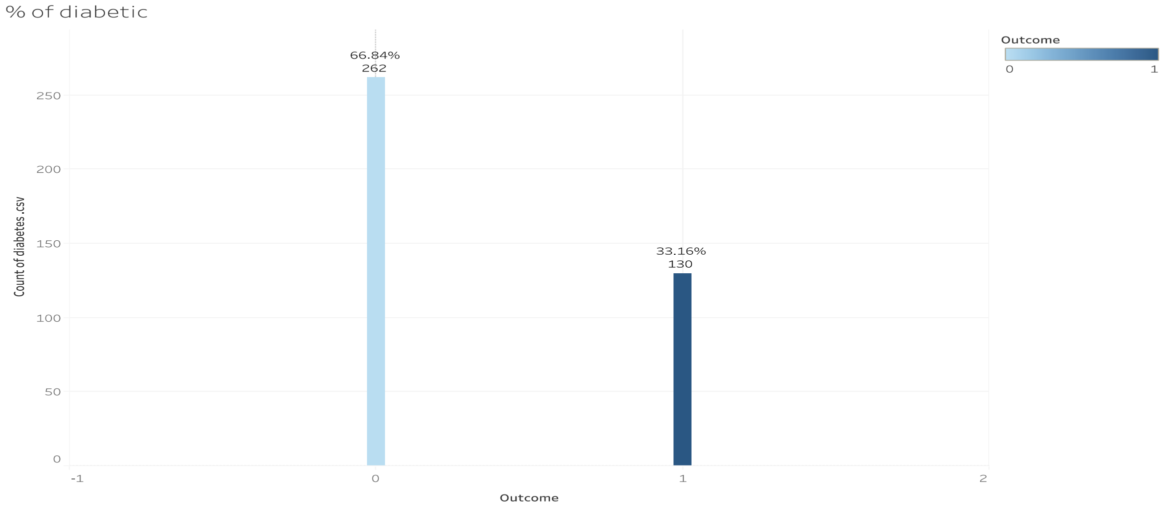


Figure 1.2 dashboard of scatterplot representing Glucose/BMI with filter as outcome and other with clustering technique.

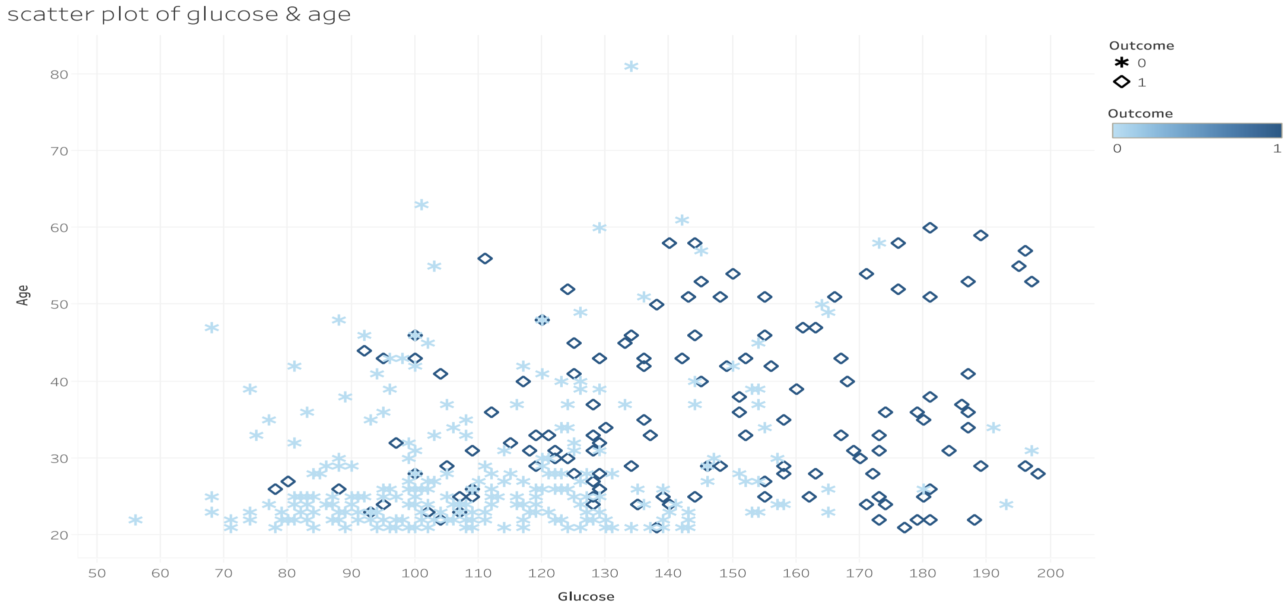


The plot of count of diabetes .csv for Outcome. Color shows details about Outcome. The marks are labeled by % of Total Count of diabetes .csv and count of diabetes .csv. The data is filtered on Pregnancies, which ranges from 0 to 17.

Figure 1.3 showing the bar-graphs of the number of people who haven't and have diabetes.

Here in both figure 1.2 we can observe that there is just a slight difference in clustered and filtered techniques both look almost similar. And also in figure 1.3 we can see that out of 392 women in the dataset 262(66.84%) doesn't have diabetes and 130(33.16%) people have diabetes.

Analysis:



Glucose vs. Age. Color shows details about Outcome. Shape shows details about sum of Outcome.

Fig 1.4 showing the scatterplot of glucose/age with filter as outcome

Here we can observe that the women with higher glucose levels are prone to diabetes more than the women with lower levels of glucose levels. And with the graph we can see that the age of the person doesn't matter when it comes to diabetes but we can see that in this dataset most of the people are above the age of 30 years.

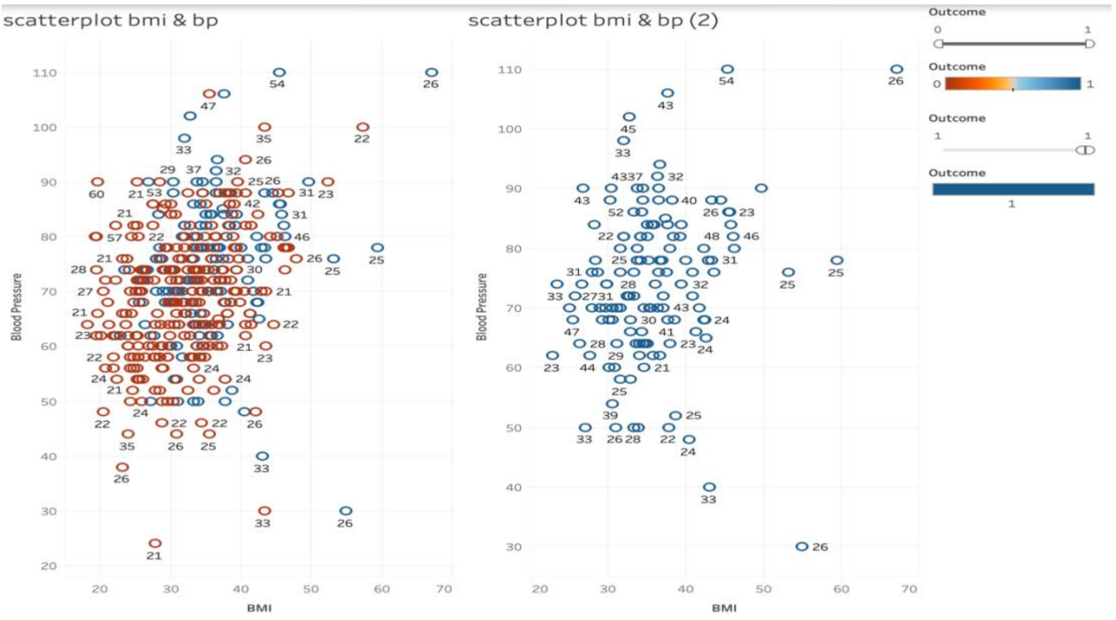
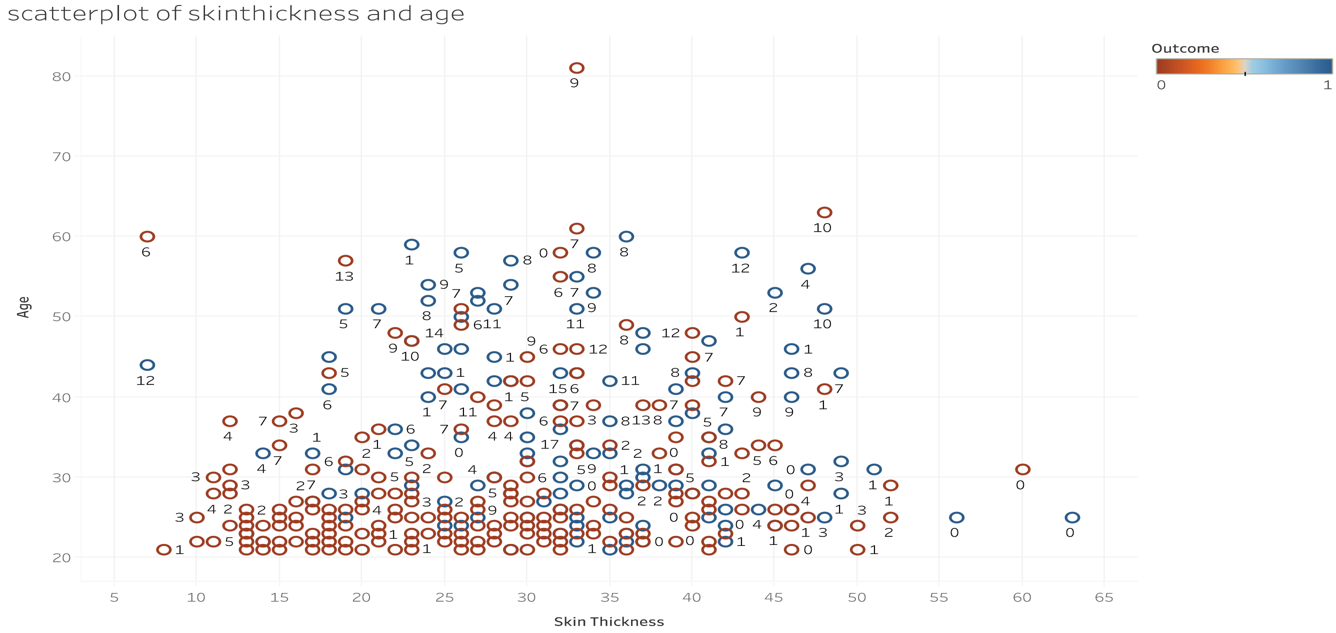


Figure 1.5 dashboard of scatterplot representing BP/BMI with filter as outcome. The right one representing only diabetic.

Here from this scatterplot we can say that the BMI of the body hasn’t got much to do with diabetes and most of the people who have and who doesn’t have diabetes have less than 45 as their BMI. But we can also observe most of the people who have diabetes has higher blood pressure, Which makes it a significant factor in testing for diabetes.



Skin Thickness vs. Age. Color shows details about Outcome. The marks are labeled by sum of Pregnancies. The data is filtered on Pregnancies, which ranges from 0 to 17. The view is filtered on Outcome and Age. The Outcome filter ranges from 0 to 1. The Age filter ranges from 21 to 81.

Fig 1.6 showing the scatterplot of skin thickness/age

Here we can see that most of the people who are not diabetic lies in the age group of 20-30 years and the people with in this age range has a skin thickness ranging from (10-32mm). and according to the dataset the thickness of the skin has got nothing to do with the diabetes but it looks like the age does. We can see more blue than red circles from the age group of 28 and above and From Figure 1.2 we can say that the more the glucose in the body the more the chances of being diabetic.

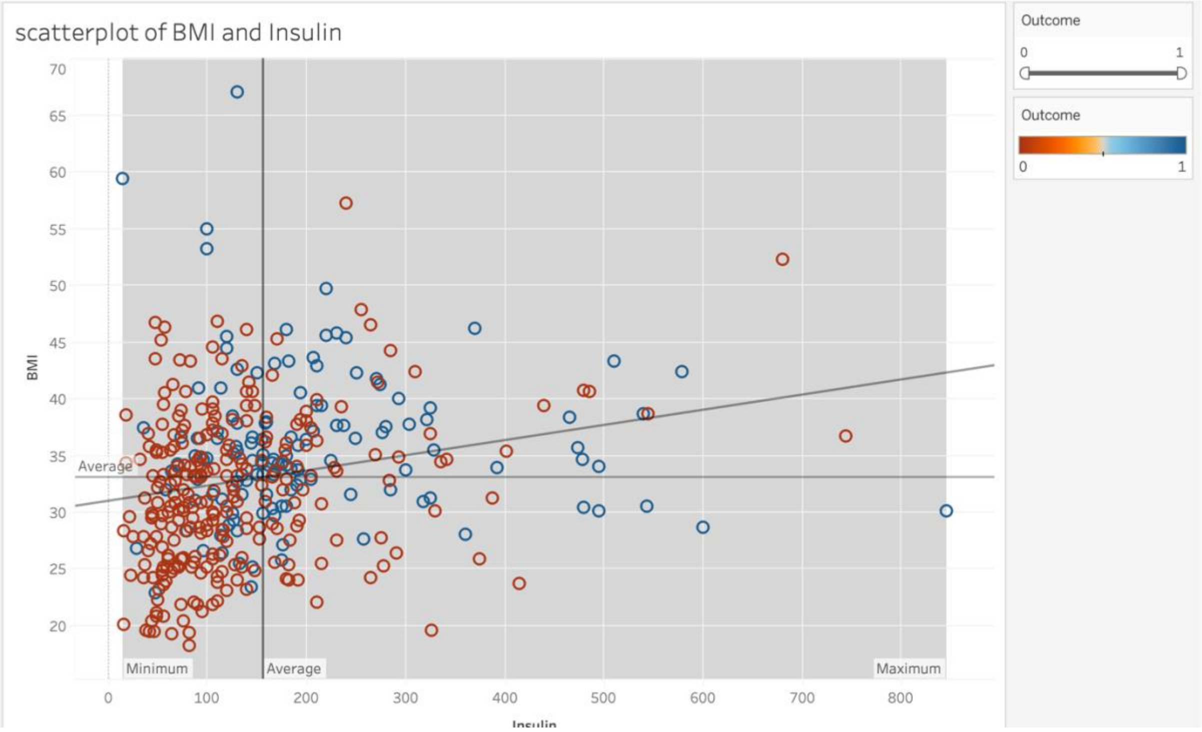
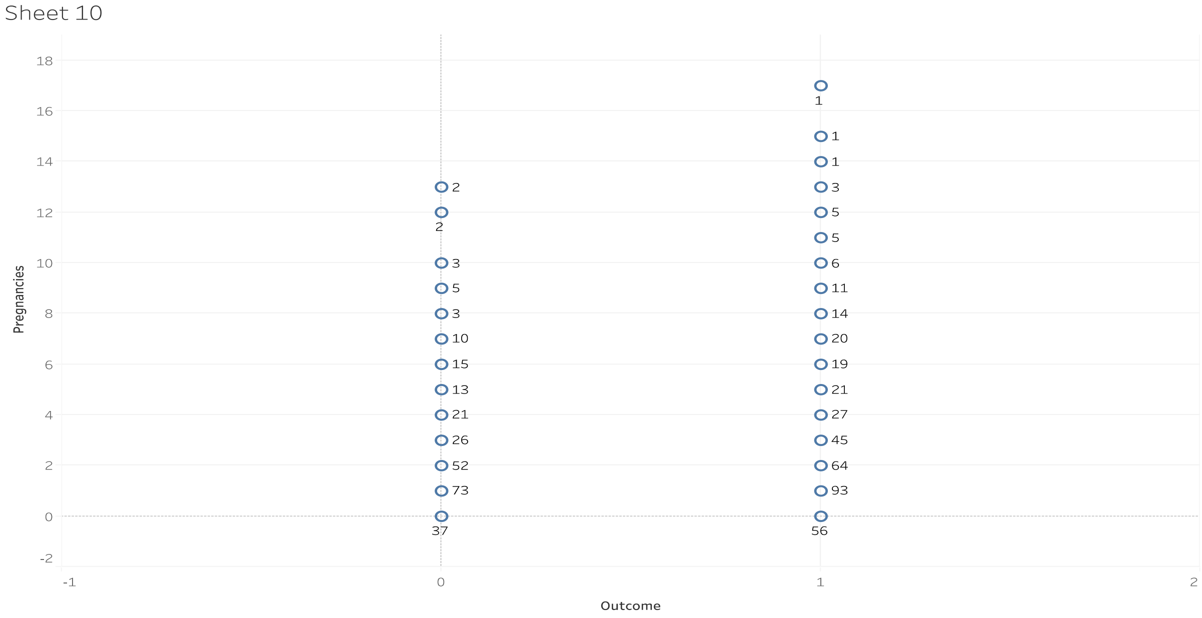


Fig 1.7 showing the scatterplot of skin Insulin/BMI

The more the insulin in the body and the more the BMI in the body the more are the chances that a women is diabetic. But it may vary with the pregnancies that a women had.



Outcome vs. Pregnancies. The marks are labeled by Running Sum of Count of diabetes .csv.

Fig 1.8 showing the scatterplot of number of Pregnancies/outcome

From the graph I could say that pregnancy can be one of the major factors for diabetes.

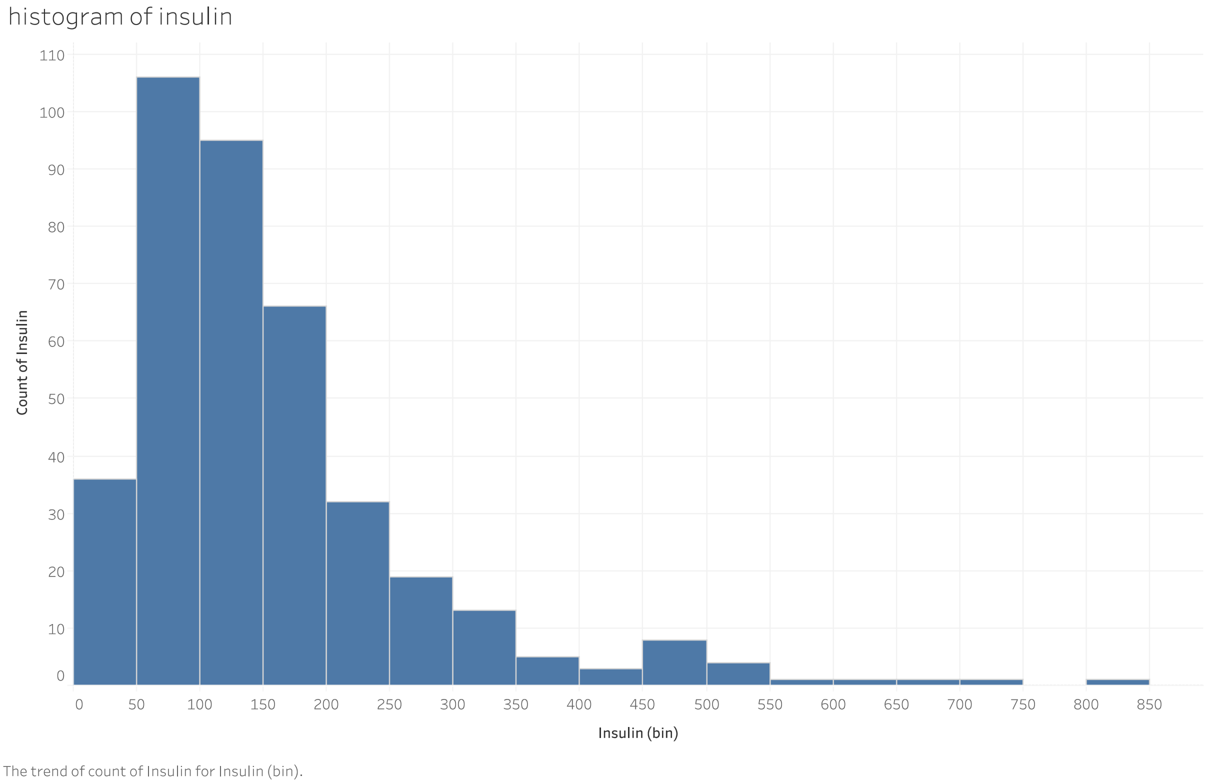


Fig 1.9 showing histogram of Insulin.

The dataset is skewed towards the left which means that the dataset is partial towards the people who are having less insulin. This can be clearly observed in the fig 1.4 where most of the data is accumulated in the less than average insulin levels. The highest is in the range of 50-100 insulin(bin) where the number of women are 106.

Conclusion:

From this Analysis I can conclude that, most of the women in this dataset has less than 5 pregnancies and the more the glucose, age, pregnancies, insulin and blood pressure the higher are the chances that the woman is suffering with diabetes. And the BMI and the Skin thickness has got not much to do with the occurrence of diabetes in the woman.

References:

Myatt, G. J., & Johnson, W. P. (2009, February 3). *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications* (1st ed.). Wiley.

Sleeper, R. (2018). *Practical Tableau: 100 tips, tutorials, and strategies from a tableau zen master*. O'Reilly.