Data Analysis in Pima Indians Diabetes Database

Student: Vanessa Soares Vieira

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.

Their inspiration is to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes.

Important questions:

1. What are the features we should use to describe our problem (predict whether or not the patients in the dataset have diabetes)? And what are the ones that has nothing to do with the outcome?
2. The zero values from a few variables are true or were they missing values filled with zero? If they're not true, how can we change them and not apply bias to the database?
3. Are older people most likely to be diabetic?
4. Why does the study uses indian woman from Pima?

Hypothesis:

- There are variables such as number of pregnancies that has anything to do with the fact that a person is diabetic or not.
- I believe it is impossible for a person to have 0 blood pressure, for example.

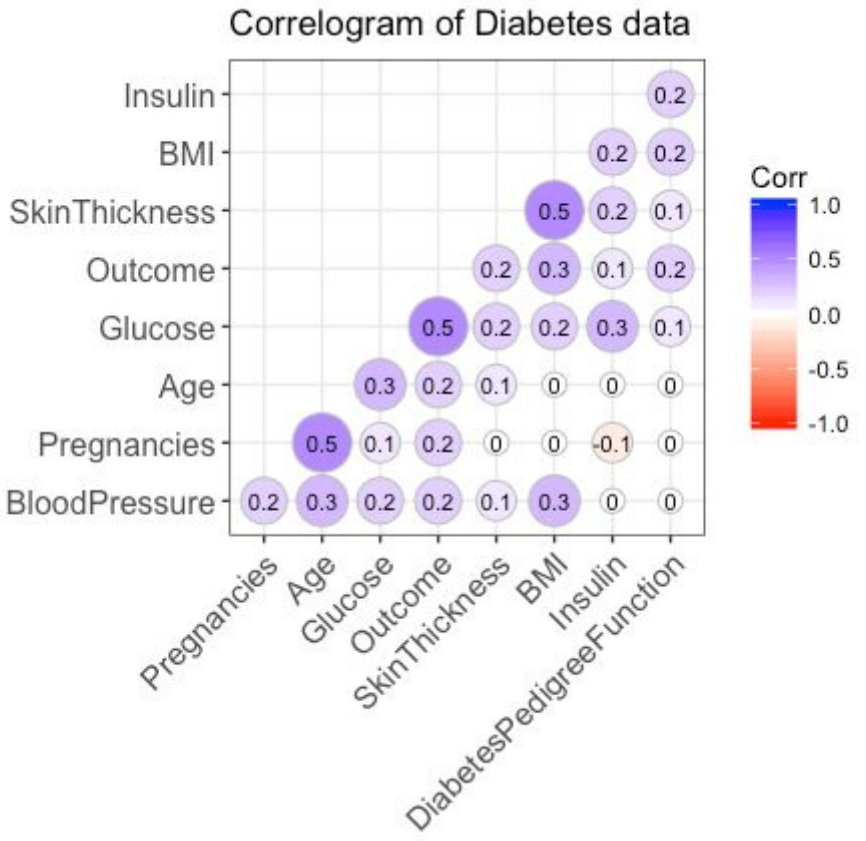
Answers (I guess):

1. The features that does not describe well the problem are definitely: Skin Thickness and Insulin, both their correlations with the Outcome are under 0.08. I know for sure that the Glucose feature is the most relevant of all features because it has the highest correlation with the Outcome, 0.40. Some of them I believe it's pretty relevant, like BMI (0.25) and Age (0.26). Some of them are yet unsure, such as Pregnancies (0.17), Blood Pressure (0.14) and Diabetes Pedigree Function (0.14). Here is the entire correlation matrix for the dataset:

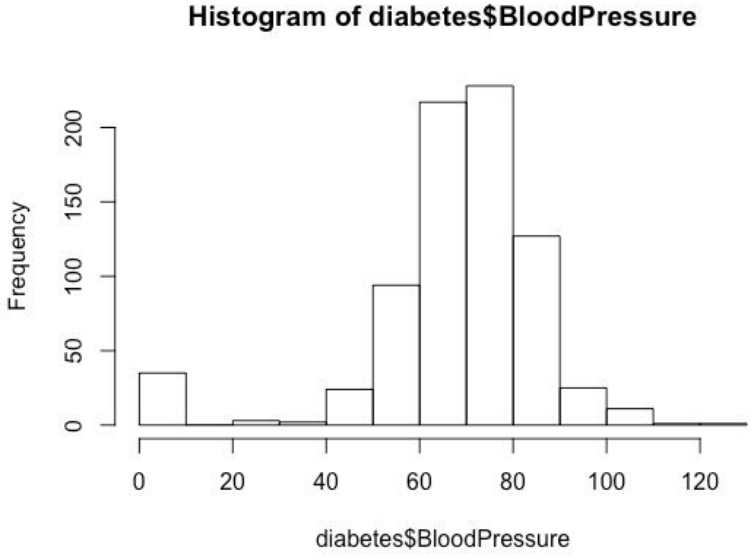
Console ~/

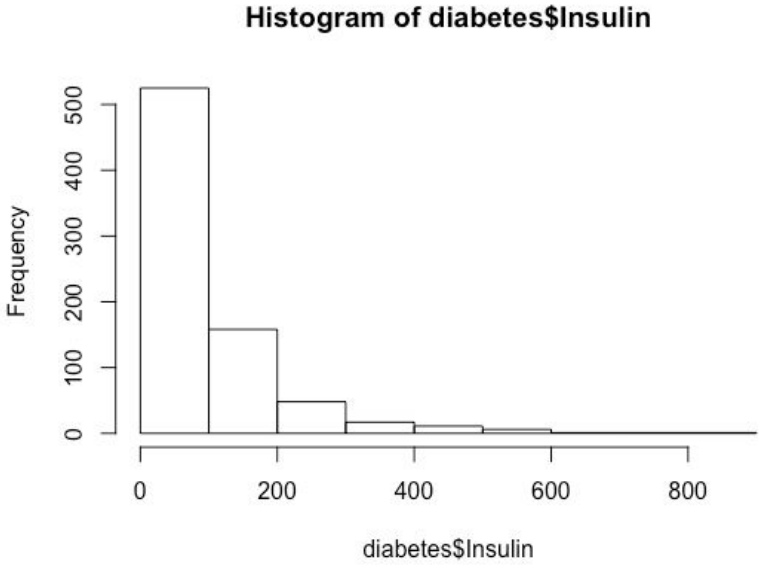
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
Pregnancies	1.00	0.09	0.14	-0.06	-0.10	0.00
Glucose	0.09	1.00	0.17	0.05	0.16	0.15
BloodPressure	0.14	0.17	1.00	0.04	-0.05	0.20
SkinThickness	-0.06	0.05	0.04	1.00	0.42	0.32
Insulin	-0.10	0.16	-0.05	0.42	1.00	0.13
BMI	0.00	0.15	0.20	0.32	0.13	1.00
DiabetesPedigreeFunction	-0.03	0.06	0.01	0.13	0.16	0.09
Age	0.46	0.19	0.26	-0.04	-0.08	0.08
Outcome	0.17	0.40	0.14	0.08	0.06	0.25

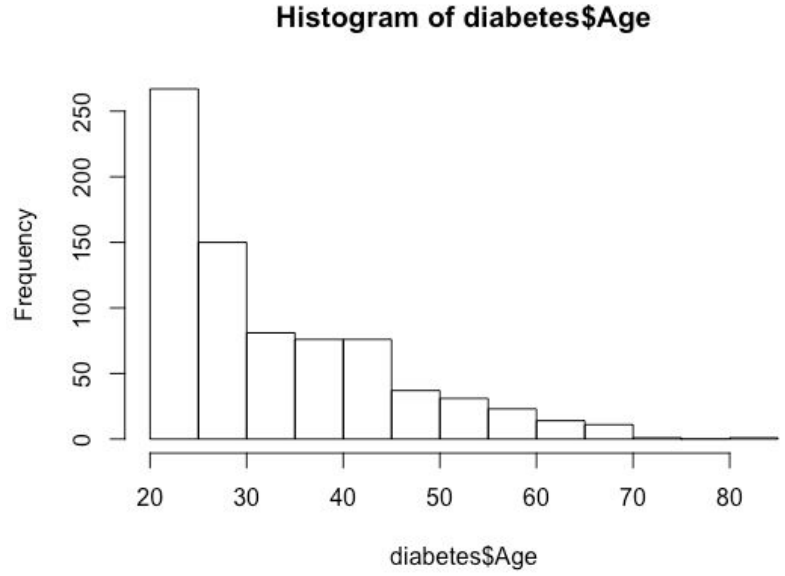
	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	-0.03	0.46	0.17
Glucose	0.06	0.19	0.40
BloodPressure	0.01	0.26	0.14
SkinThickness	0.13	-0.04	0.08
Insulin	0.16	-0.08	0.06
BMI	0.09	0.08	0.25
DiabetesPedigreeFunction	1.00	0.03	0.14
Age	0.03	1.00	0.26
Outcome	0.14	0.26	1.00

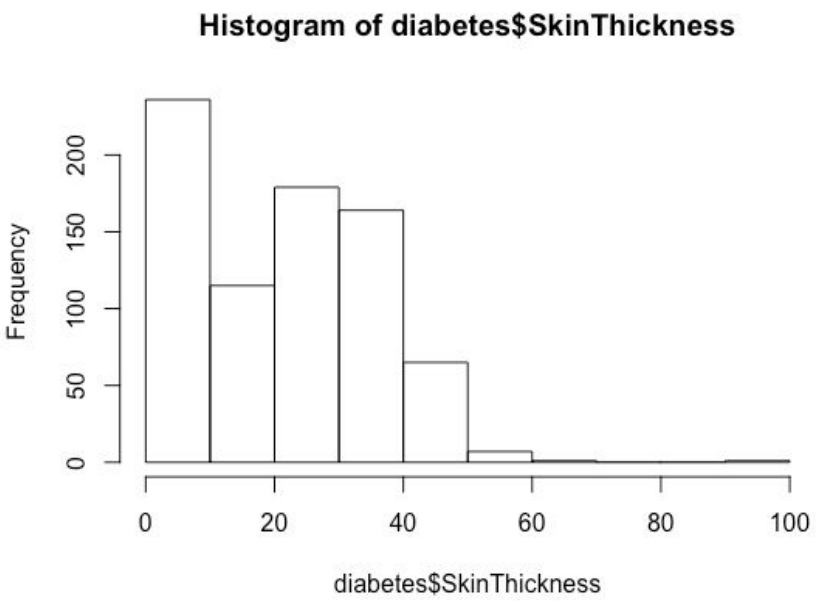


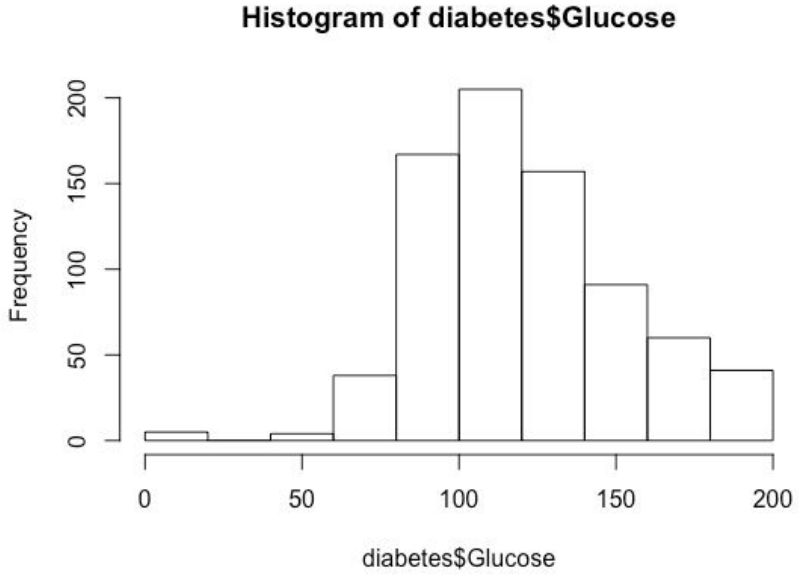
2. There were values such as 0 in Blood Pressure, Glucose, BMI and Skin Thickness that for what I could research, makes no sense. So I made an Histogram of all features and here are the results:

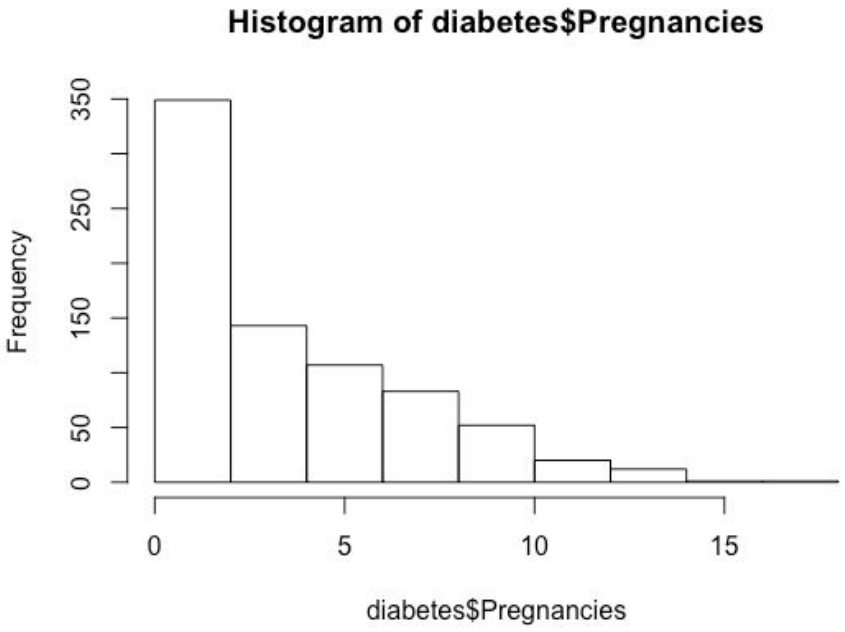

Histogram of diabetes$BloodPressure
Frequency
0
50
100
150
200
0
20
40
60
80
100
120
diabetes$BloodPressure

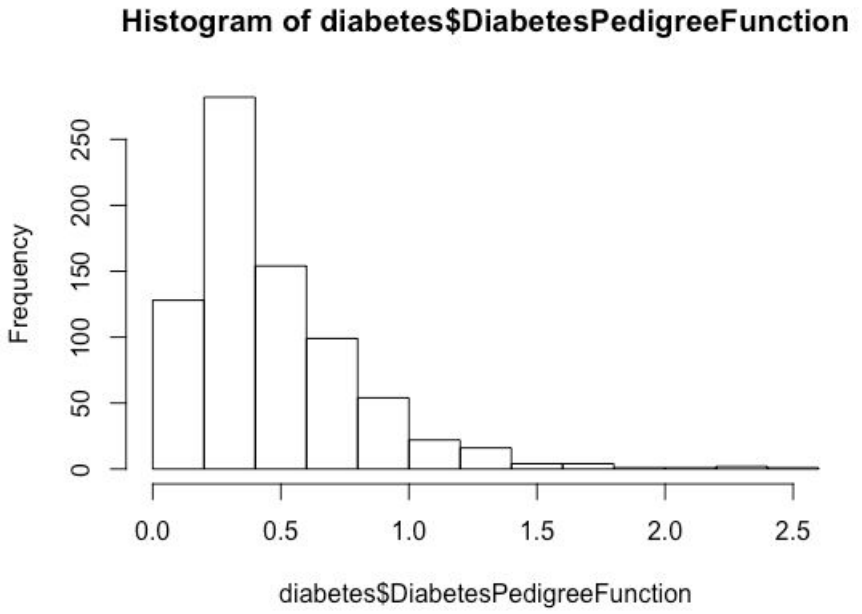

Histogram of diabetes$Insulin
Frequency
0
100
200
300
400
500
0
200
400
600
800
diabetes$Insulin

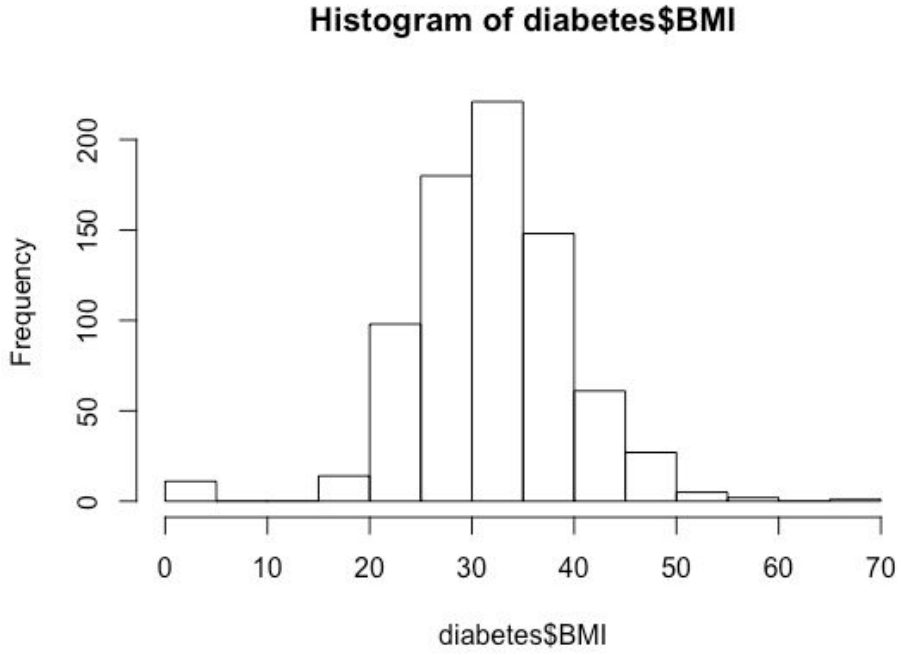

Histogram of diabetes$Age
Frequency
0
50
100
150
200
250
20
30
40
50
60
70
80
diabetes$Age


Histogram of diabetes$SkinThickness
Frequency
0
50
100
150
200
0
20
40
60
80
100
diabetes$SkinThickness

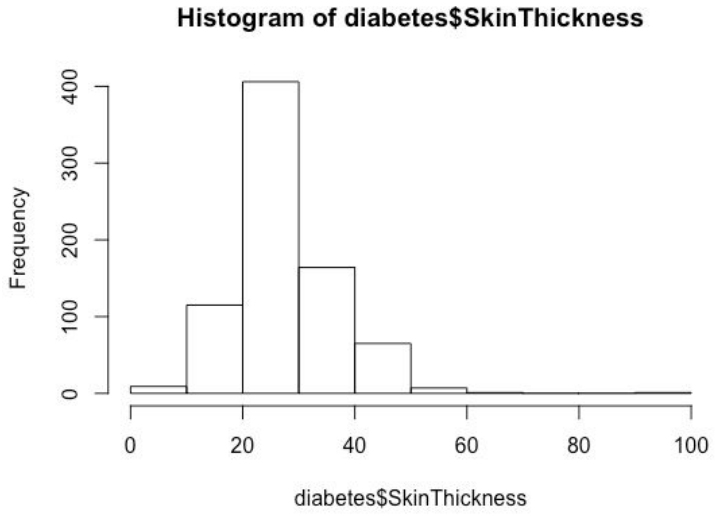

Histogram of diabetes$Glucose
Frequency
0
50
100
150
200
0
50
100
150
200
diabetes$Glucose

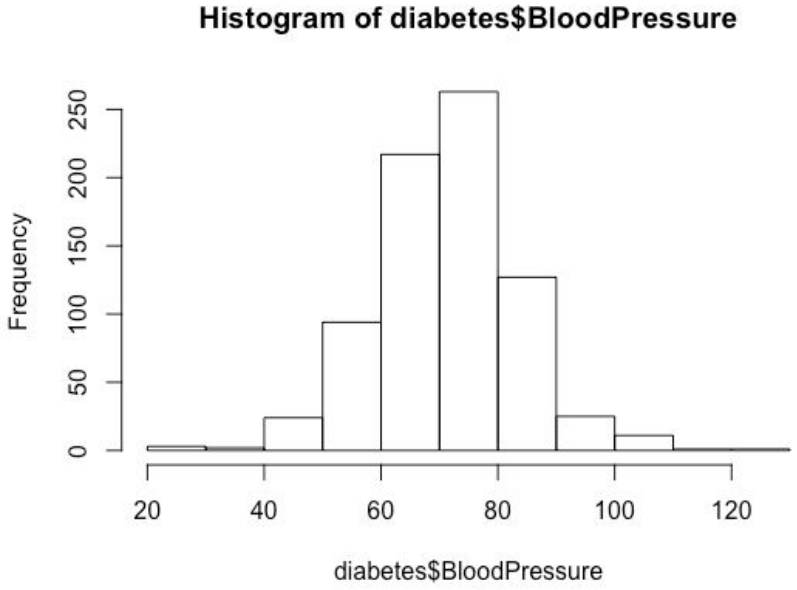

Histogram of diabetes$Pregnancies
Frequency
0
50
150
250
350
0
5
10
15
diabetes$Pregnancies

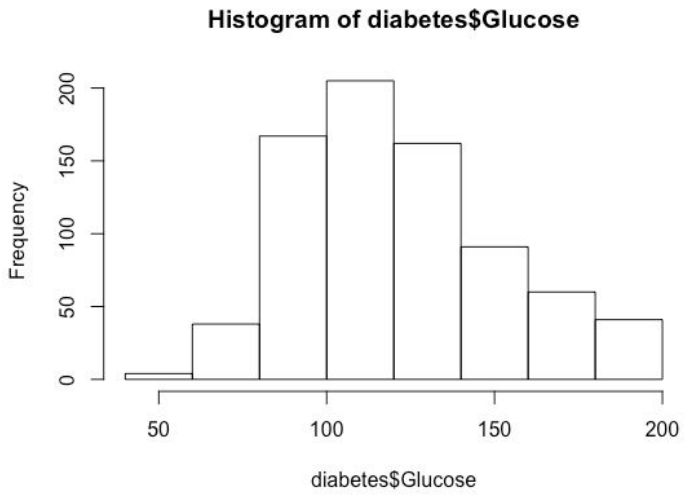


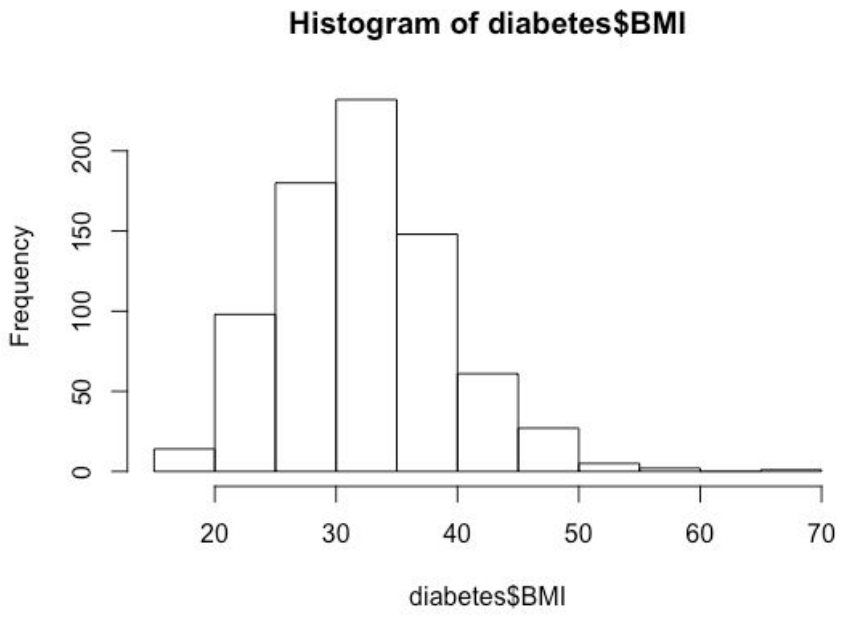


Some of the features follow some kind of a normal distribution and I thought that the zeros were just values that were missing. I replaced the ones that had the normal distribution with the median and the ones that had a different distribution with the mean. After that, here are the results for the features I've changed:

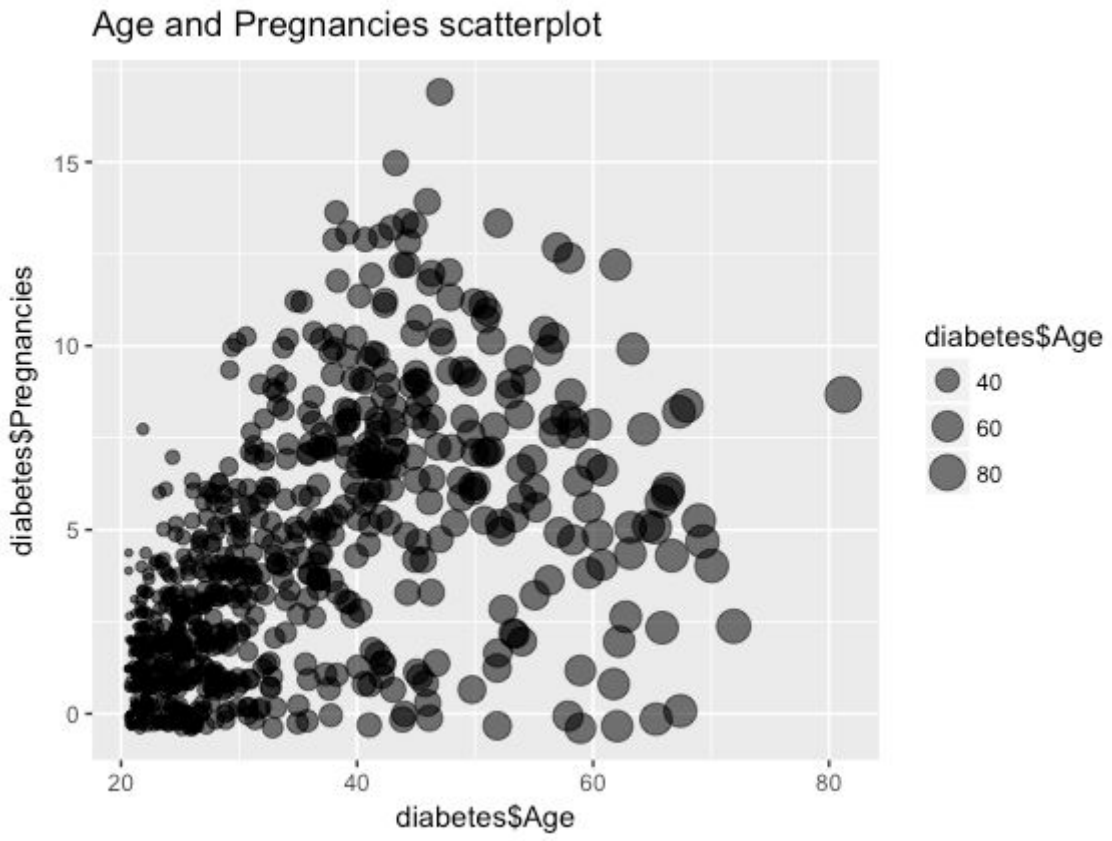

Histogram of diabetes$SkinThickness
Frequency
0
100
200
300
400
0
20
40
60
80
100
diabetes$SkinThickness


Histogram of diabetes$BloodPressure
Frequency
0
50
100
150
200
250
20
40
60
80
100
120
diabetes$BloodPressure





3. This is a question every old person wants to know the answer. I think the answer is not if older people are most likely to have diabetes but if there is any correlation between the variables. And yes. From the correlation table I showed we can see that the correlation is 0.47. I've made a scatterplot just to be safe. Here it is.



4. The study uses indian woman from Pima because they wanted people from a place where there is not much of a modern touch. They eat natural food, they don't eat industrial stuff. They're as natural as one can get and they're a great sample for what is healthy and what's not.