Sabrina Crowe

MIS 110

Professor Kevin Ryan

12 December 2022

Analyzing the Potential Causes of Type II Diabetes in Women

Project Overview

This project's dataset uses quantitative methods to measure risk factors linked to developing Type II Diabetes such as BMI, blood pressure, and blood glucose levels in women ages 21 and above. Each data entry is accompanied by a '0' or '1' to indicate whether the woman ended up having diabetes – 0 for no, 1 for yes.

The aim of my python program was to visualize the data in this .csv file and to create a Logistic Regression using this data. I chose a Logistic Regression because the outcome variable is categorical, being either a 0 or 1. The data visualization uses packages from seaborn and pandas, while the regression was created using sklearn. The sklearn package also allowed me to display the confusion matrix and accuracy of the logistic regression model.

Data Analysis

Before analyzing and visualizing the data, I noticed that the .csv file contained zeroes under categories such as blood pressure, BMI, insulin, etc., which were unlikely to have such values. For this reason, I cleaned up the data by only including rows that had values greater than 0 in columns that would reasonably have non-zero values.

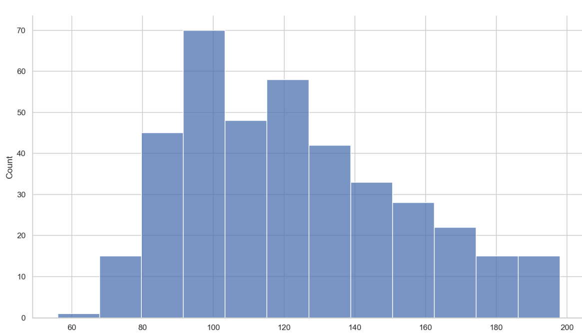
```
#cleaning up data

df = df[df.Glucose != 0]
df = df[df.BloodPressure != 0]
df = df[df.SkinThickness != 0]
df = df[df.Insulin != 0]
df = df[df.BMI != 0]
```

I then modeled the distributions of the variables included in the .csv files using seaborn's displot() function and used the summary() function to get a broader idea of the sample's shape and values.

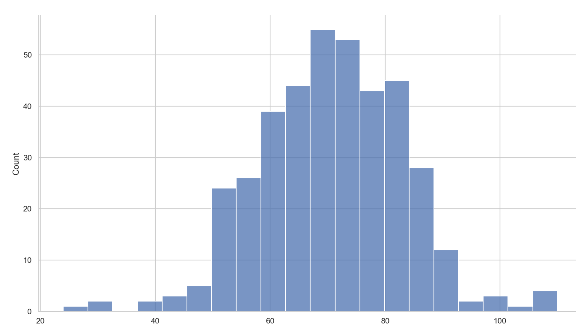
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	392.000000	392.000000	392.000000	392.000000	392.000000
mean	3.301020	122.627551	70.663265	29.145408	156.056122
std	3.211424	30.860781	12.496092	10.516424	118.841690
min	0.000000	56.000000	24.000000	7.000000	14.000000
25%	1.000000	99.000000	62.000000	21.000000	76.750000
50%	2.000000	119.000000	70.000000	29.000000	125.500000
75%	5.000000	143.000000	78.000000	37.000000	190.000000
max	17.000000	198.000000	110.000000	63.000000	846.000000

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	392.000000	392.000000	392.000000	392.000000
mean	33.086224	0.523046	30.864796	0.331633
std	7.027659	0.345488	10.200777	0.471401
min	18.200000	0.085000	21.000000	0.000000
25%	28.400000	0.269750	23.000000	0.000000
50%	33.200000	0.449500	27.000000	0.000000
75%	37.100000	0.687000	36.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

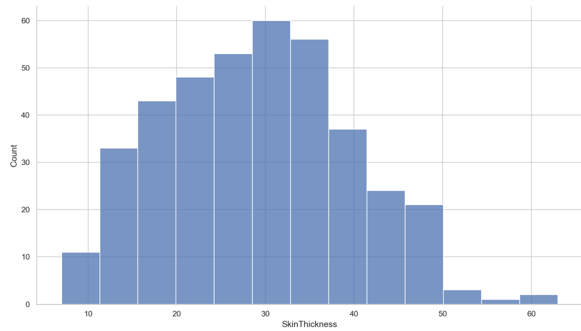
Glucose



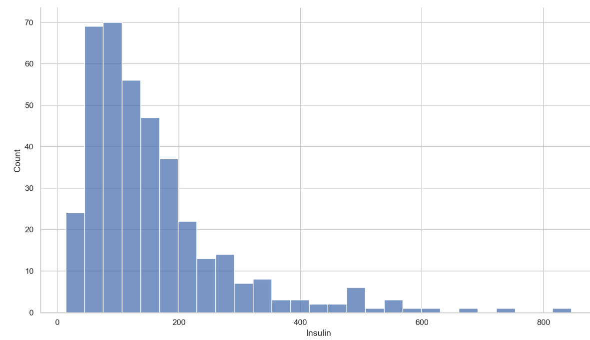
Blood Pressure



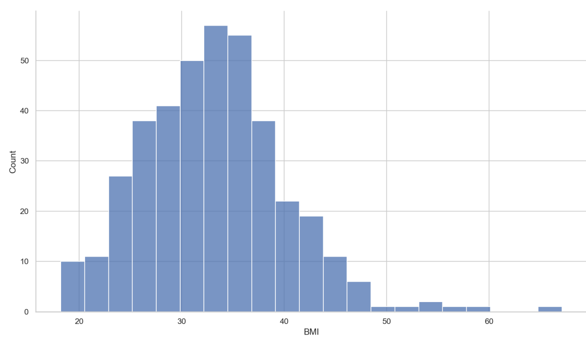
Skin Thickness



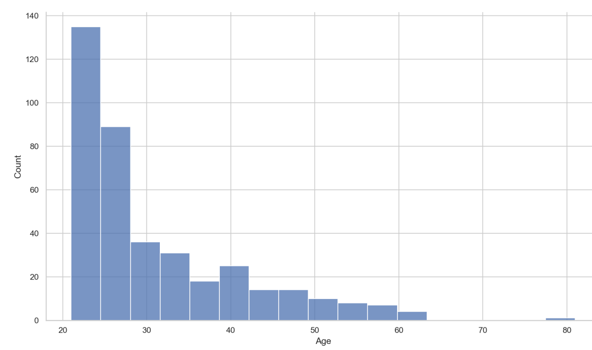
Insulin



BMI



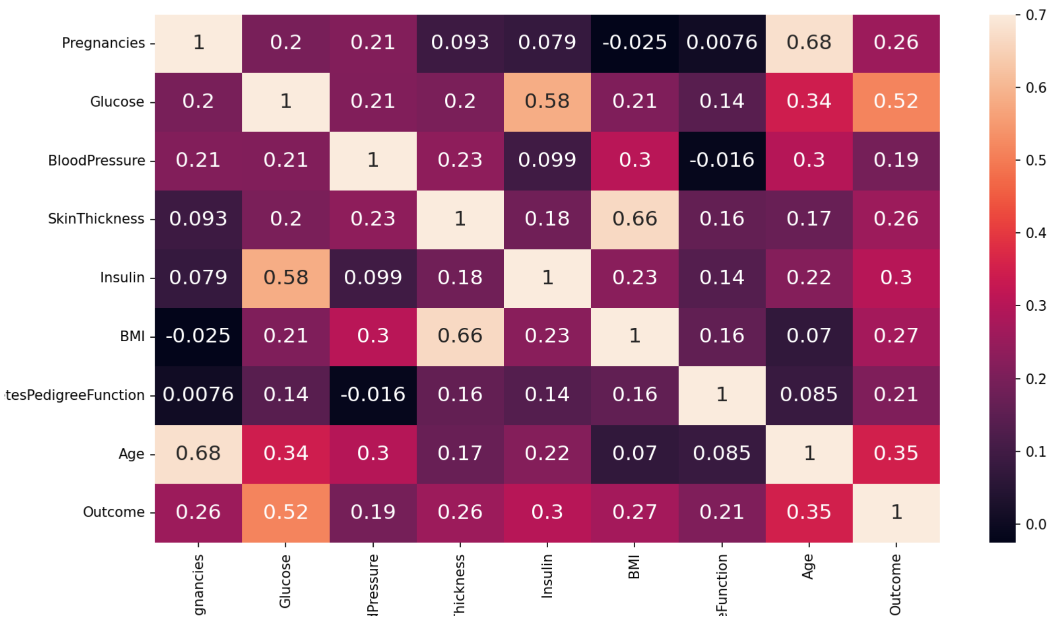
Age



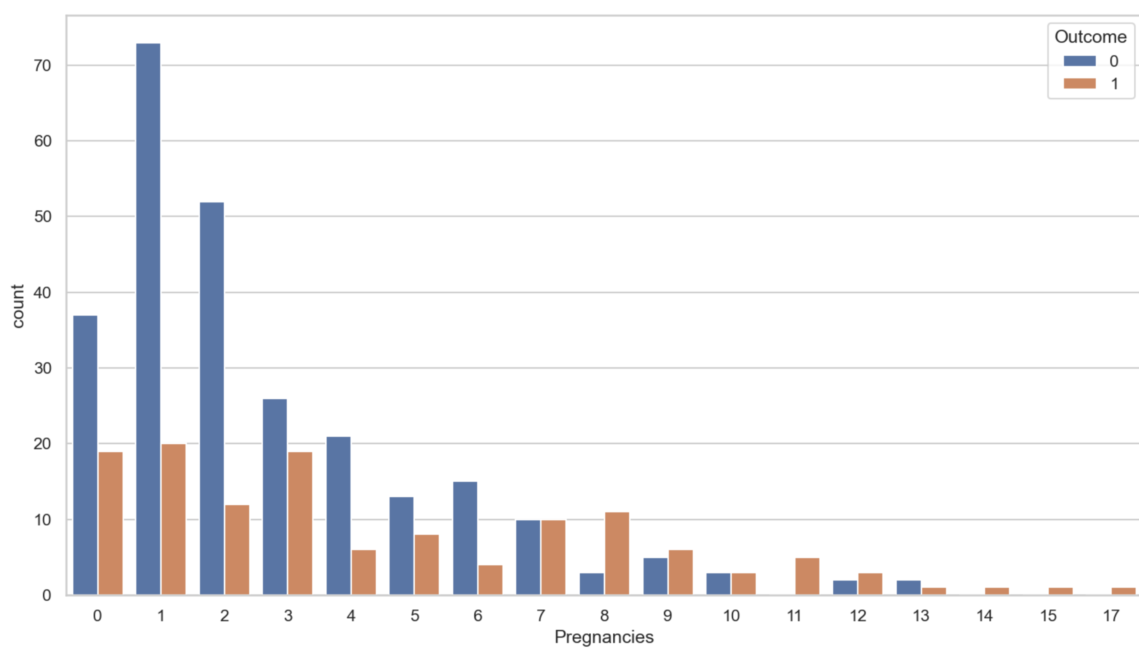
I decided not to plot the number of pregnancies because I didn't think the distribution shape would lend much insight. From these distributions, we can see that our sample skews younger in age with roughly normal distributions of skin thickness, BMI, glucose, and blood pressure. Insulin has a few outliers that give it a right-skewed distribution.

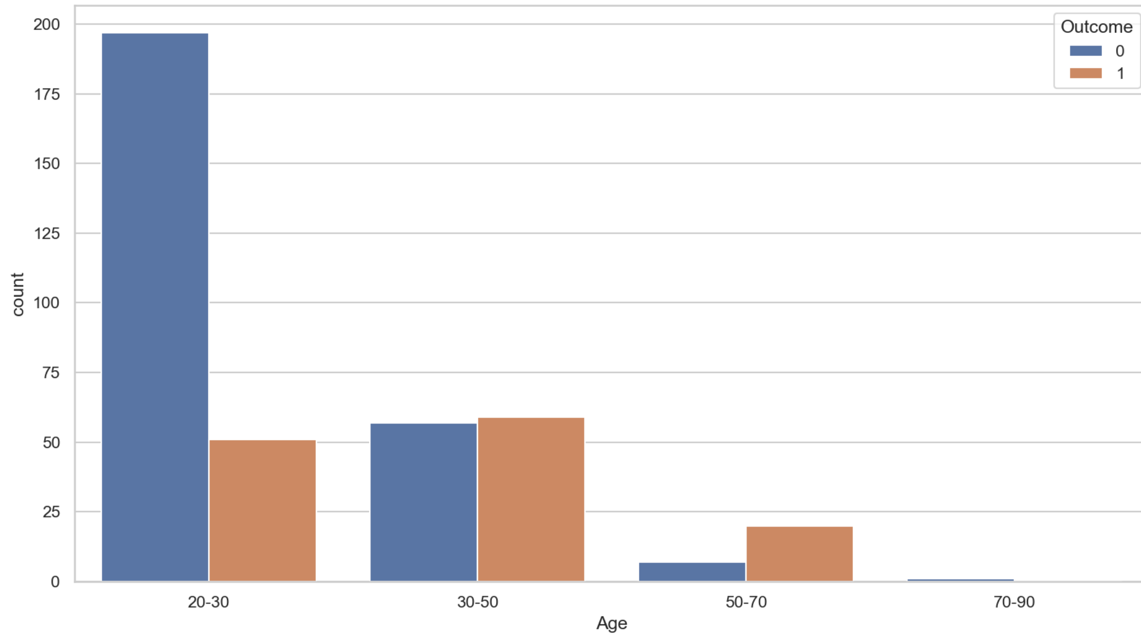
To see if there was a clear correlation between any single one of these variables and the outcome, I created a correlation matrix and visualized it using a heatmap.

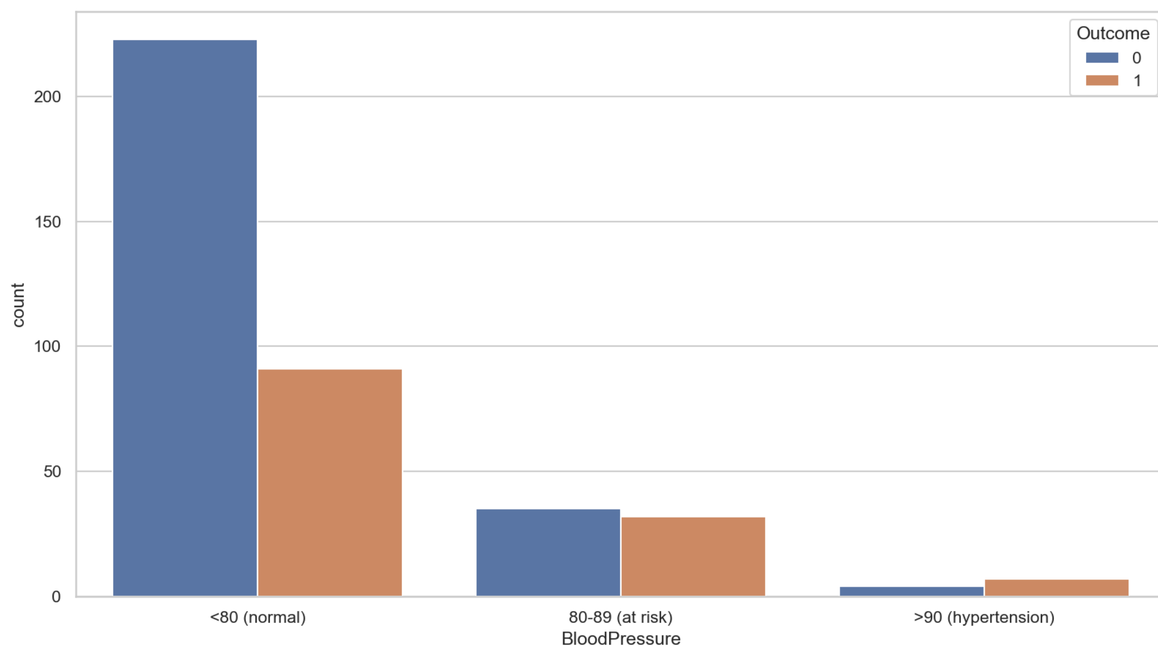
```
                          Pregnancies   Glucose  ...       Age   Outcome
Pregnancies                  1.000000  0.198291  ...  0.679608  0.256566
Glucose                      0.198291  1.000000  ...  0.343641  0.515703
BloodPressure                0.213355  0.210027  ...  0.300039  0.192673
SkinThickness                0.093209  0.198856  ...  0.167761  0.255936
Insulin                      0.078984  0.581223  ...  0.217082  0.301429
BMI                         -0.025347  0.209516  ...  0.069814  0.270118
DiabetesPedigreeFunction     0.007562  0.140180  ...  0.085029  0.209330
Age                          0.679608  0.343641  ...  1.000000  0.350804
Outcome                      0.256566  0.515703  ...  0.350804  1.000000
```

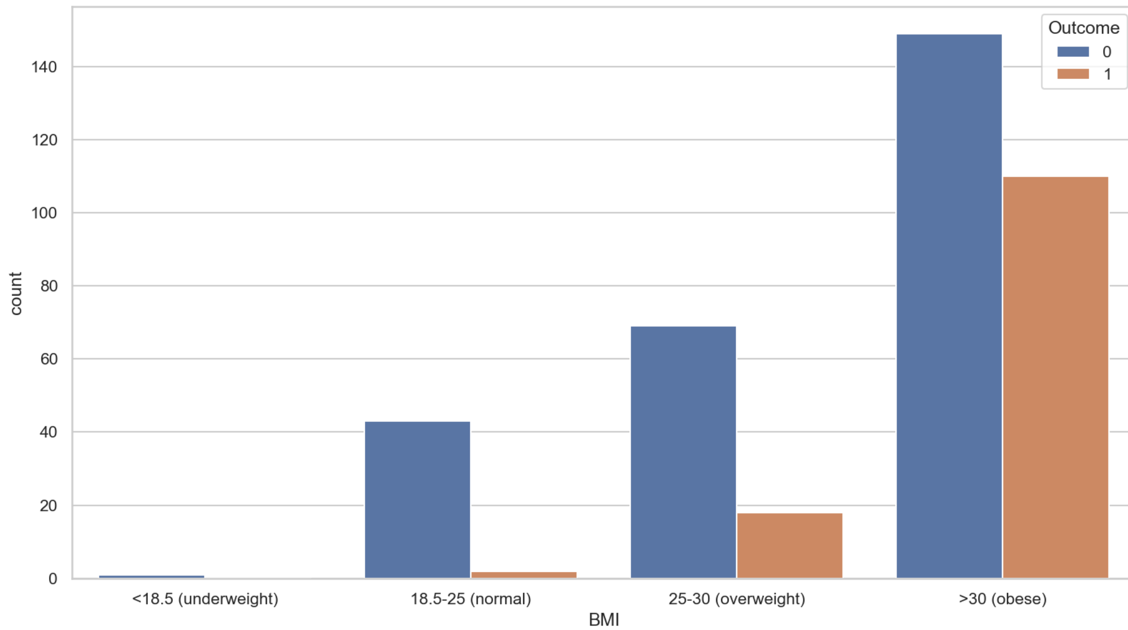


As indicated by the heatmap, there is no strong correlation between outcome and any of the variables. However, using count plots, one can get a better idea of the attributes a woman diagnosed with diabetes tends to have.




Outcome
0
1
count
200
175
150
125
100
75
50
25
0
20-30
30-50
50-70
70-90
Age


Outcome
0
1
count
200
150
100
50
0
<80 (normal)
80-89 (at risk)
>90 (hypertension)
BloodPressure

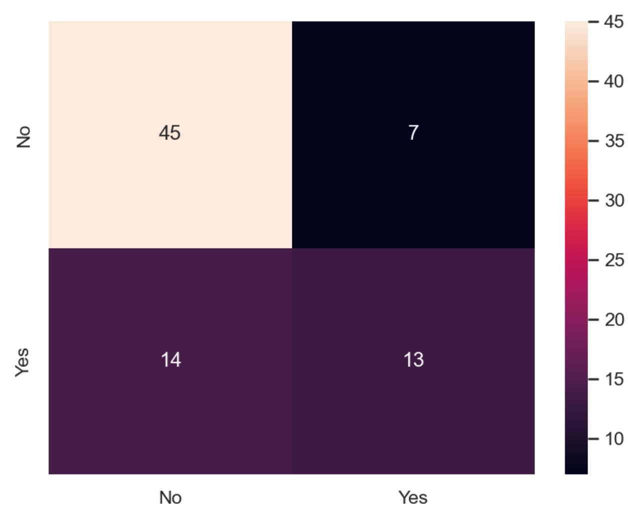


A higher proportion of people classified as having an obese BMI, an at risk and above blood pressure, or above 30 years old are diagnosed with diabetes.

Logistic Regression

In the logistic regression, 20% of the data is used in the test split. The x variable doesn't include outcome or the diabetes pedigree function because of its ambiguous value. After testing and training the data, I had my program output the confusion matrix and accuracy score of the regression. The confusion matrix is visualized through a heat map.

```
[[45  7]
 [14 13]]
0.7341772151898734
```



The regression has an accuracy score of 73.4%, however the confusion matrix indicates that the model is more accurate at predicting whether a woman does not have diabetes rather than whether a woman does.

Future Extensions / Reflection

Further extensions of this project could include inputting a value for one or several of the variables and receiving a probability of being diagnosed with diabetes; however, this feature is beyond my current skill level. I would also look into ways of having the model be more accurate in predicting whether a woman does have diabetes. This program could further be expanded by using a more varied data set, including data from men and women from different countries; the current data set is exclusively women of one ethnicity. Countries could be analyzed to find a correlation between the location's prevalence of diabetes and the likelihood of having diabetes by taking into account average BMI and a "walkability" score for each country. Overall, the project is useful as an educational tool for visualizing the comorbidities of type II diabetes and identifying patterns in diagnoses.