0.

```
Means:
Glucose: 121.6867627785059
Blood Pressure: 72.40518417462484
Skin Thickness: 29.153419593345657
Insulin: 155.5482233502538
BMI: 32.45746367239099

Standard Deviations:
Glucose: 30.53564107280403
Blood Pressure: 12.382158210105263
Skin Thickness: 10.476982369987212
Insulin: 118.77585518724514
BMI: 6.924988332105907
```

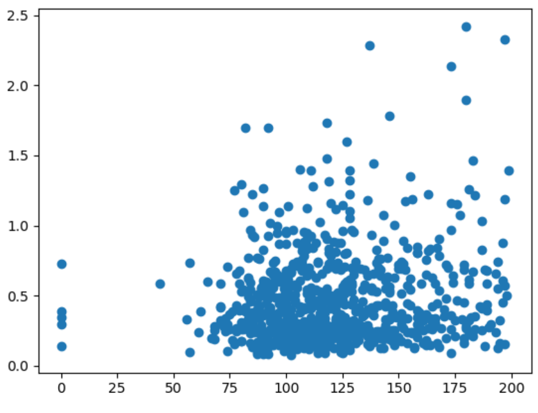
1.

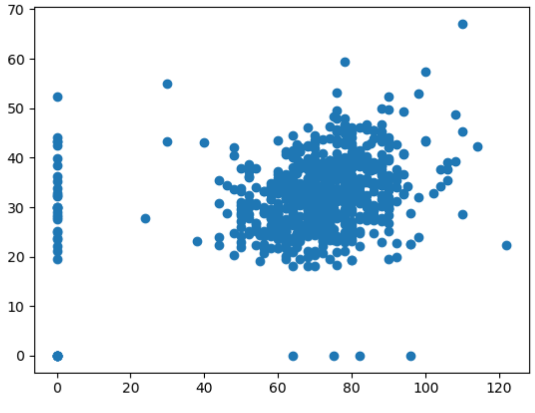
```
Covariance matrix:
                   Glucose  BloodPressure  SkinThickness       Insulin         BMI
Glucose         932.425376      84.811985      74.025750   2127.014566   49.355133
BloodPressure    84.811985     153.317842      29.240422    145.553584   24.644988
SkinThickness    74.025750      29.240422     109.767160    230.676780   46.725661
Insulin        2127.014566     145.553584     230.676780  14107.703775  190.422831
BMI              49.355133      24.644988      46.725661    190.422831   47.955463

Correlations:
Glucose, Blood Pressure: 0.22319177824954217
Glucose, Skin Thickness: 0.22804322678186384
Glucose, Insulin: 0.581186208912165
Glucose, BMI: 0.23277051103551327
Blood Pressure, Skin Thickness: 0.22683906740782173
Blood Pressure, Insulin: 0.09827229945465545
Blood Pressure, BMI: 0.28923034040466644
Skin Thickness, Insulin: 0.18488842018975898
Skin Thickness, BMI: 0.6482139430923206
Insulin, BMI: 0.228050157416571
```

Because covariance matrices are not standardized, we can only generalize the relationships between the attributes. Since the main diagonal of the matrix is large for Insulin (when compared to its’ standard deviation squared), we can say the standard deviation around the mean is largely spread. We can say this for other variables by comparing the standard deviation squared of each variable with the same variable in the matrix.
The values here for correlation are generally not very high, representing far linear relationships between each other. Some strongly correlated relationships, compared to others, to note are glucose and insulin, as well as skin thickness and BMI. Again, these are the most correlated attributes, and have the best linear relationships.

2.



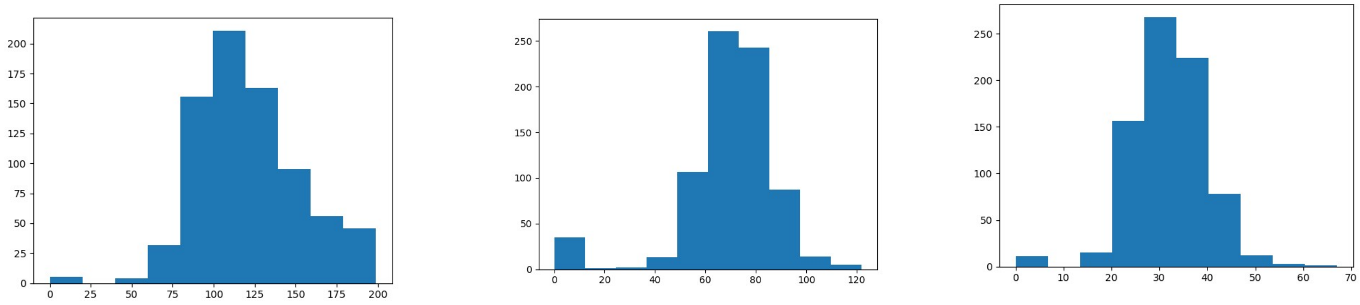


Left Attributes: 2, 7; Right Attributes: 3, 6

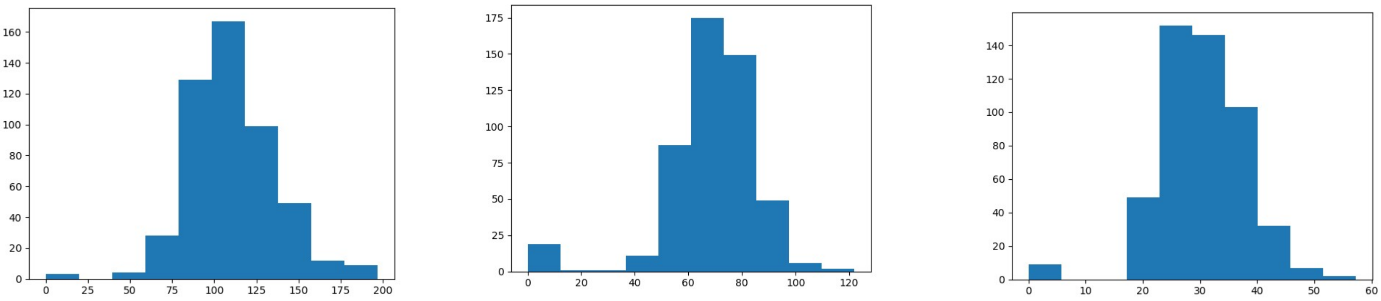
For attributes 2 and 7, the points appear to be most dense around value 100 to 125 on the x axis, with the values ranging from around 60 to 200. On the y axis, values range from around 0 to at most 1.5. For this scatter plot, there is no apparent trend or relationship between the two attributes.

For the scatter plot containing the attributes 3 and 6, the points are heavily centered around 60-80 on the x and 20-50 on the y axis. The trend here appears to be a linearly positive relationship.

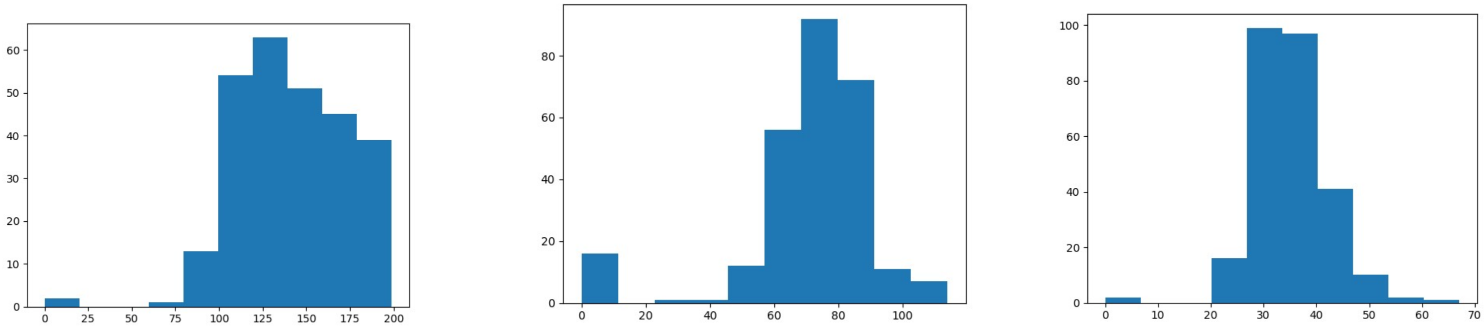
3.



Histograms for attributes: 2, 3, 6, respectively, for the whole data set



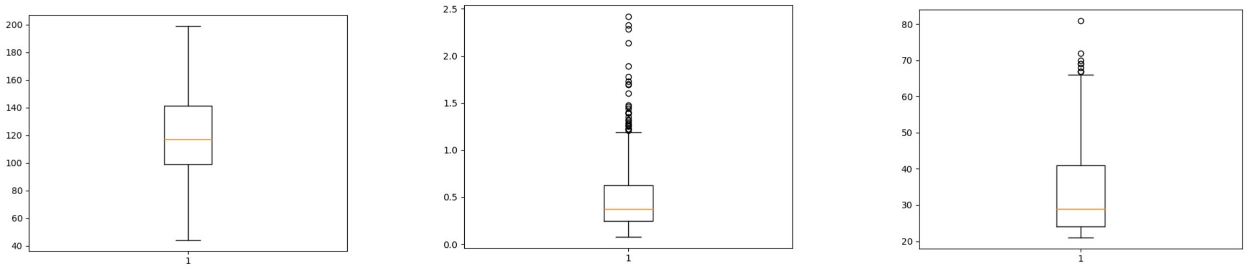
Histograms for attributes: 2, 3, 6, respectively, for outcome 0



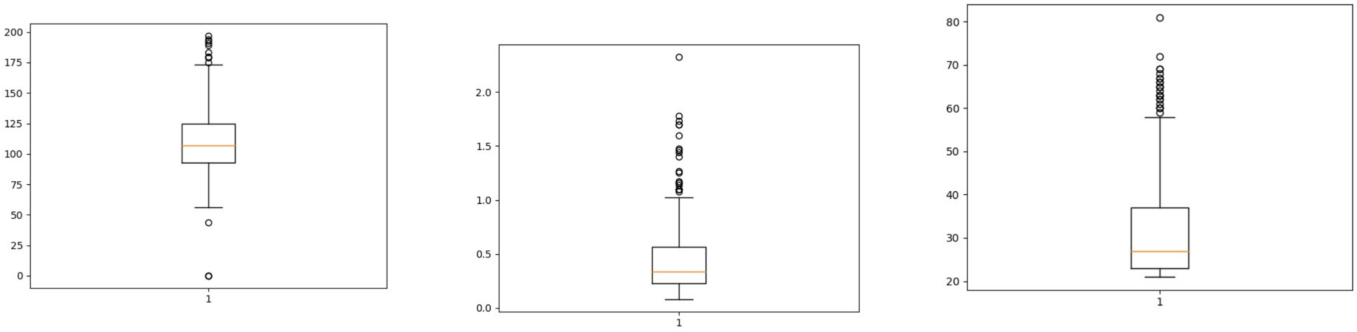
Histograms for attributes: 2, 3, 6, respectively, for outcome 1

Each histogram is generally Gaussian shaped, with sharper peaks for attributes 3 and 6 when the outcome was 0. All of the histograms displayed are unimodal. In this case, the differences are not very apparent. The only histogram that is skewed would be attribute 2 for outcome 0. I believe a larger bin size would allow for more information to be shown, while also revealing more differences between the classes.

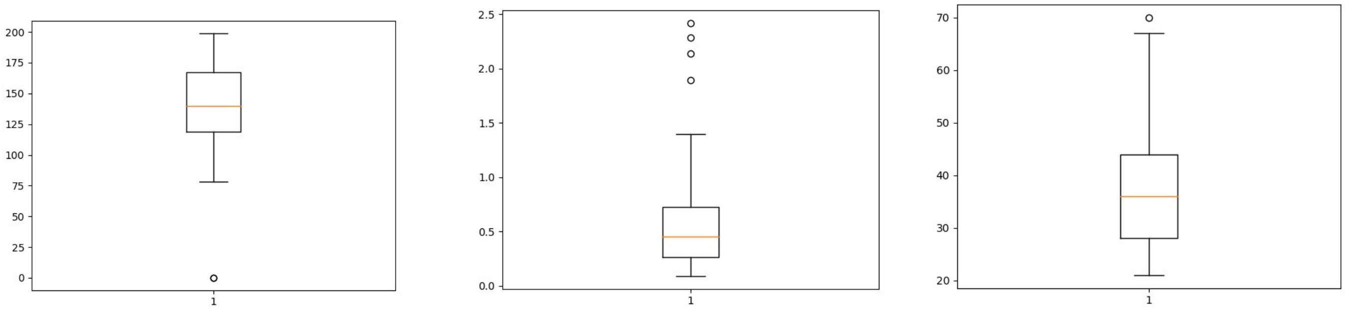
4.



Boxplots for attributes: 2, 7, 8, respectively, whole data set



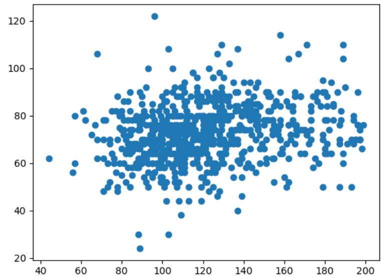
Boxplots for attributes: 2, 7, 8, respectively, for outcome 0

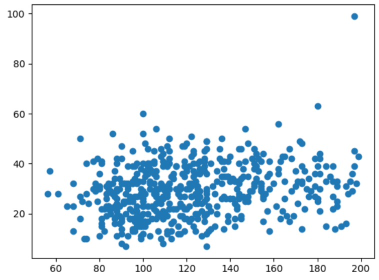


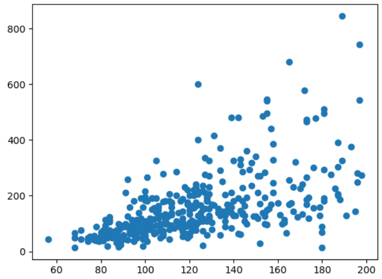
Boxplots for attributes: 2, 7, 8, respectively, for outcome 1

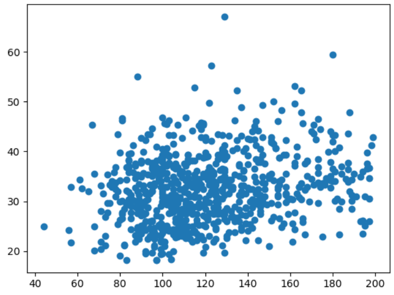
The data for each boxplot is centered at the median, and for the whole data set, the outliers are above the upper quartile. Also for the whole data set, the mean is generally skewed to the right. We can also note that these features also appear when the outcome is 0, as well as outcome 1. When looking at attribute 8, or age, we see that for those with diabetes have a higher median value, without much skew. On the other hand, those without had a lower median, with it being skewed to the right. There were also much more outliers present in those with an outcome of 0. Another attribute to observe between classes is the Glucose levels. Those with diabetes have a higher median, as well as higher 2nd and 3rd quartiles.

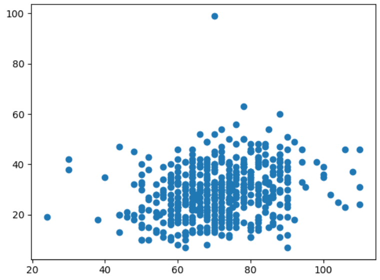
5.

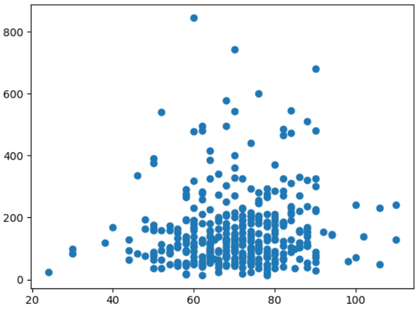


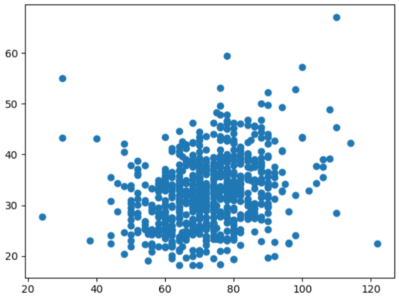


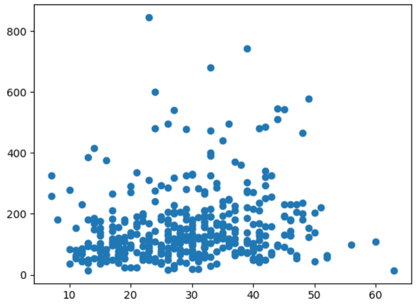


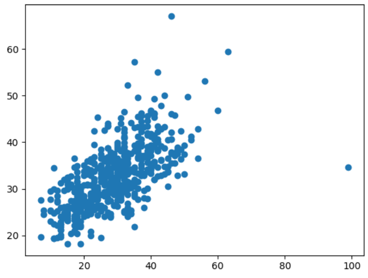


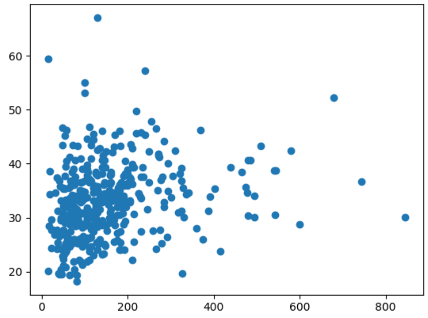






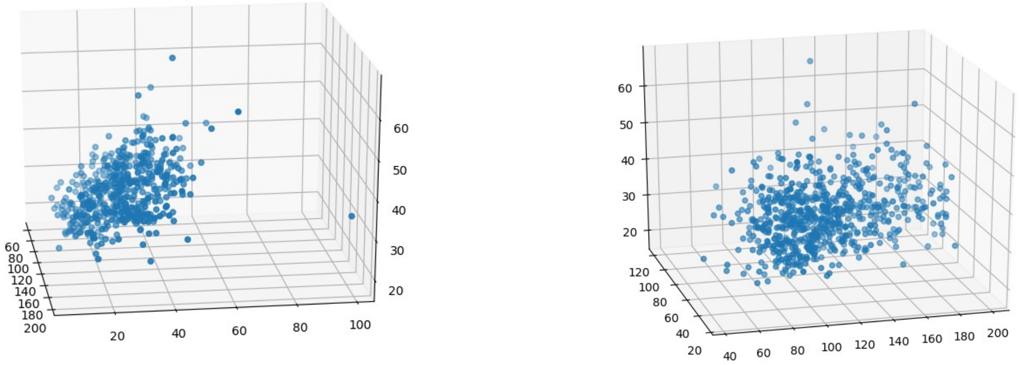






2D Scatter Plots Respective:
(2, 3), (2, 4), (2, 5)
(2, 6), (3, 4), (3, 5)
(3, 6), (4, 5), (4, 6)
(5, 6)

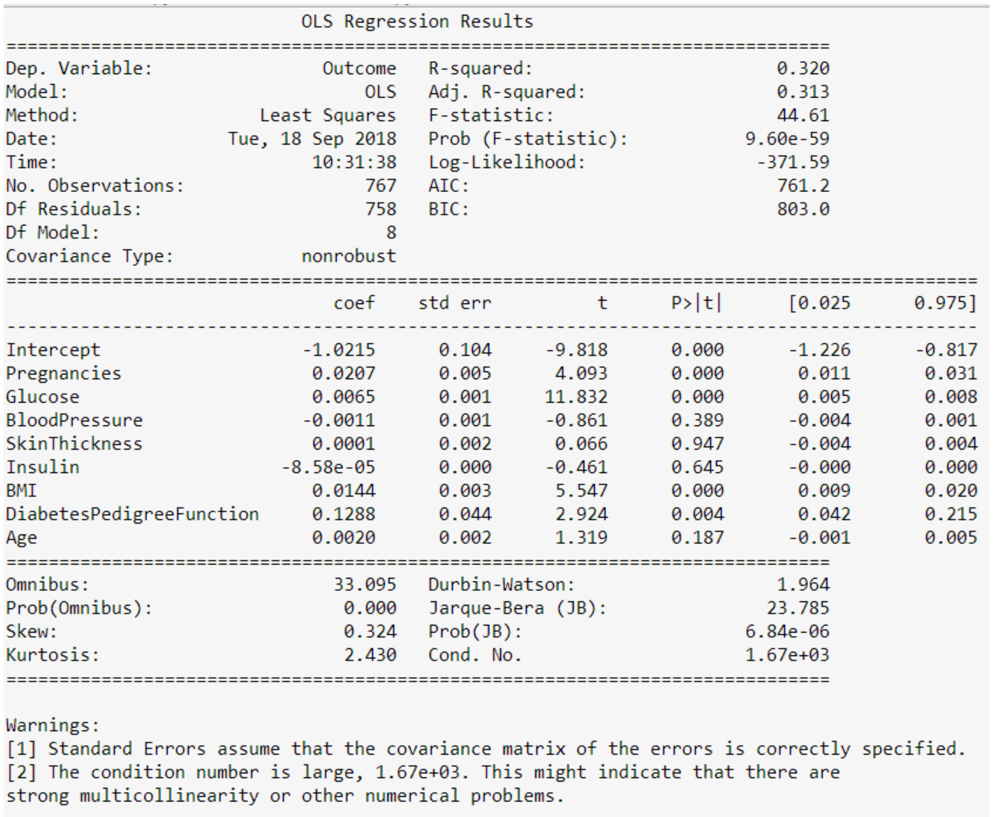
The three plots that appear to have a linear and positive relationship are with attributes (2, 5), (4, 6), (5, 6). Each graph has obvious outliers, and the 3 graphs that are particularly centered are with attributes (2, 6), (3, 4), (3, 5), (3, 6). Although debatable, these graphs have less variance at the mean when compared to plots (2, 4), and (4, 5). Because these plots have greater variance around the mean, the two attributes has a weaker relationship with each other, and may be a bad predictor of the outcome. The amount of attributes present does make determining if the patient has diabetes difficult, and for other projects with many more attributes, there will also be an increased amount of graphs to observe.



3D Scatter Plots: Attributes respective (2, 4, 6), (2, 3, 6)

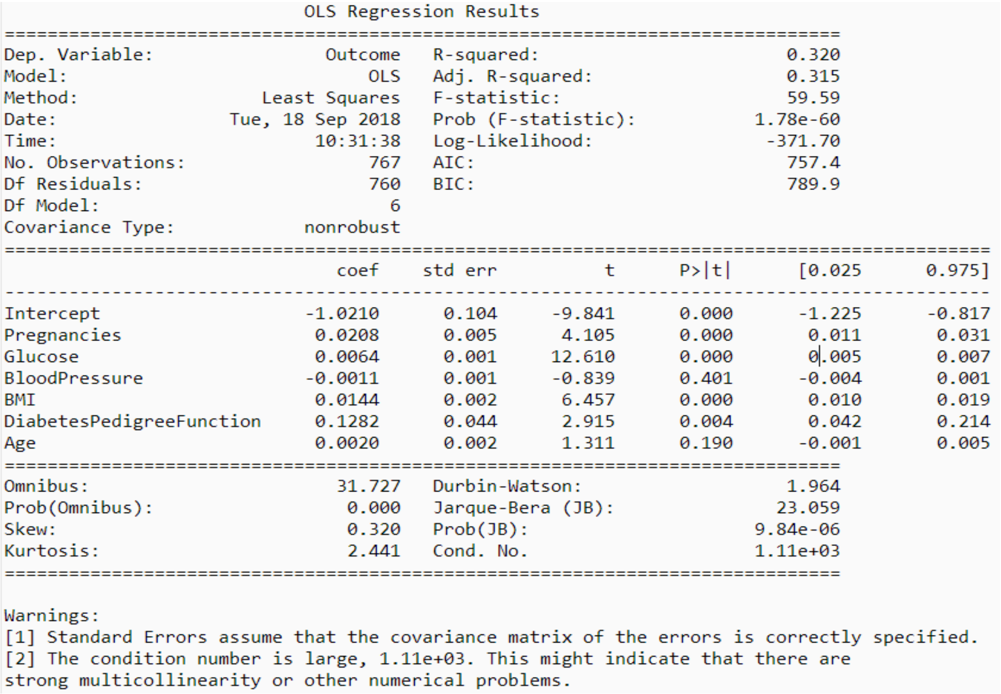
The difficulty with 3D plots is from multiple variables being displayed, cause issues observing the relationships in the 3D space. It can be seen that while both have some outliers, the plot on the left seems to have a positive relationship between the three attributes. On the other graph, the values are somewhat centered near the center of the space. They are generally spread out, due to high variance in these attributes. In this case, it was more difficult to describe the features of the 3 dimensional plot. On a positive note, a 3D space allows for stronger visualization of multivariate data, and may be helpful in determining whether relationships are present in this data.

7.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.320
Model:                            OLS   Adj. R-squared:                  0.313
Method:                 Least Squares   F-statistic:                     44.61
Date:                Tue, 18 Sep 2018   Prob (F-statistic):           9.60e-59
Time:                        10:31:38   Log-Likelihood:                -371.59
No. Observations:                 767   AIC:                             761.2
Df Residuals:                     758   BIC:                             803.0
Df Model:                           8
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   -1.0215      0.104     -9.818      0.000      -1.226      -0.817
Pregnancies                  0.0207      0.005      4.093      0.000       0.011       0.031
Glucose                      0.0065      0.001     11.832      0.000       0.005       0.008
BloodPressure               -0.0011      0.001     -0.861      0.389      -0.004       0.001
SkinThickness                0.0001      0.002      0.066      0.947      -0.004       0.004
Insulin                   -8.58e-05      0.000     -0.461      0.645      -0.000       0.000
BMI                          0.0144      0.003      5.547      0.000       0.009       0.020
DiabetesPedigreeFunction     0.1288      0.044      2.924      0.004       0.042       0.215
Age                          0.0020      0.002      1.319      0.187      -0.001       0.005
==============================================================================
Omnibus:                       33.095   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.785
Skew:                           0.324   Prob(JB):                     6.84e-06
Kurtosis:                       2.430   Cond. No.                     1.67e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.67e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Outcome   R-squared:                       0.320
Model:                            OLS   Adj. R-squared:                  0.315
Method:                 Least Squares   F-statistic:                     59.59
Date:                Tue, 18 Sep 2018   Prob (F-statistic):           1.78e-60
Time:                        10:31:38   Log-Likelihood:                -371.70
No. Observations:                 767   AIC:                             757.4
Df Residuals:                     760   BIC:                             789.9
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   -1.0210      0.104     -9.841      0.000      -1.225      -0.817
Pregnancies                  0.0208      0.005      4.105      0.000       0.011       0.031
Glucose                      0.0064      0.001     12.610      0.000       0.005       0.007
BloodPressure               -0.0011      0.001     -0.839      0.401      -0.004       0.001
BMI                          0.0144      0.002      6.457      0.000       0.010       0.019
DiabetesPedigreeFunction     0.1282      0.044      2.915      0.004       0.042       0.214
Age                          0.0020      0.002      1.311      0.190      -0.001       0.005
==============================================================================
Omnibus:                       31.727   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.059
Skew:                           0.320   Prob(JB):                     9.84e-06
Kurtosis:                       2.441   Cond. No.                     1.11e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.11e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

OLS Regression: Top picture is 1st run, bottom picture is 2nd run
Both models return similar results, even after removing blood pressure and skin thickness. The adjusted R-squared values only differed by 0.002, while other values also demonstrated very similar values. With regards to the coefficients, the DiabetesPedigreeFunction has the largest coefficients. It appears that when the coefficients are smaller or negative, they provide a larger p value, also providing less evidence against the null hypothesis. This can be seen for blood pressure, skin thickness, and insulin. On the other hand, when the coefficients are larger, the p value is generally smaller, equating to strong evidence against the null hypothesis. Because two values were removed due to high p values, and little changed in the R-squared values, it is obvious that their removal did not affect the result, and it may be safe to say that those attributes were not strongly related to the remaining attributes.

8.

Note: While trying to use Anaconda to install the graphviz package to show the decision trees, all of the dependencies became undetectable to my IDE, and thus I could not run any of the libraries afterwards like numpy/pandas/matplotlib/etc. Hours of troubleshooting were spent to no avail.

9.
In the linear regression, we observed that the removal of two attributes did not affect the results, thus we can conclude that the removed attributes were not strongly related to determining whether if the patient has diabetes. In the scatter plots, observing strong positive linear relationships between glucose and insulin, insulin and BMI, or skin thickness and BMI, could result in the patient having a higher chance of diabetes. The box plots determining age and diabetes were contradicting, but they did indicate higher glucose along with higher insulin levels, both pertaining to greater chances of diabetes. Overall, only a few attributes, such as the aforementioned insulin, glucose provides more indication of diabetes. It could also be seen that in the regression, that those with lower and negative correlations equated to less evidence against diabetes, while the opposite values in correlations provided greater evidence of diabetes.