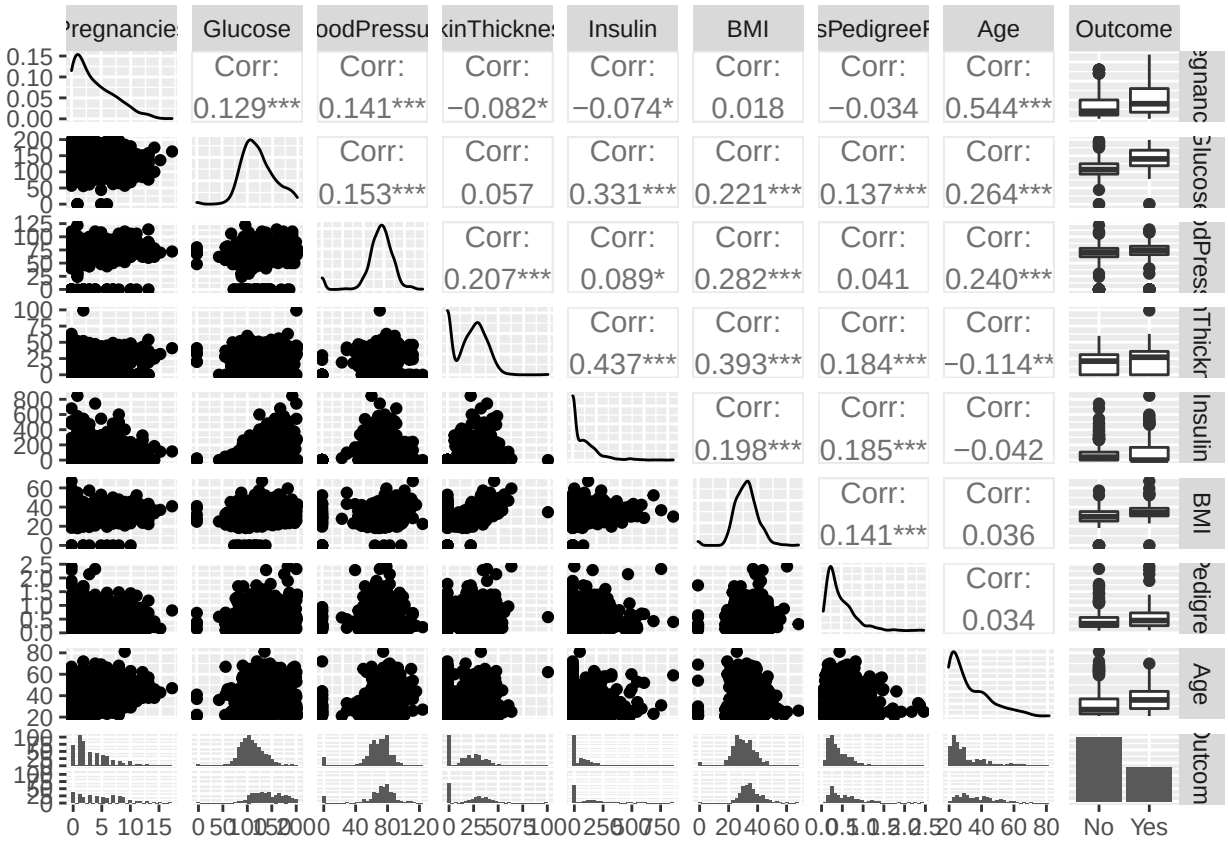
Multiple Linear Regression First Run

Uyen Nguyen

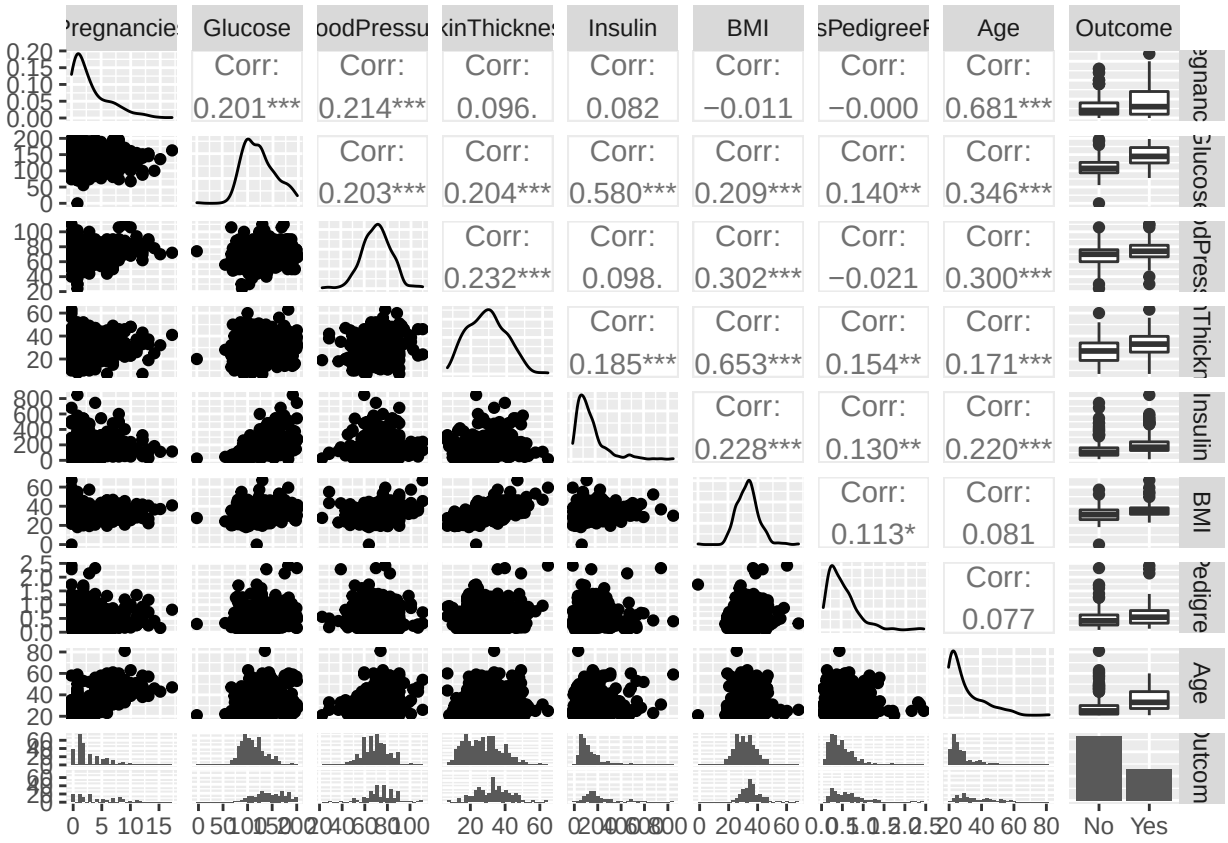
The data cleaning will ensure ggplot will color correctly for variable Outcome

```
# Scatter plot using GGpairs
GGally::ggpairs(diabetes)
```



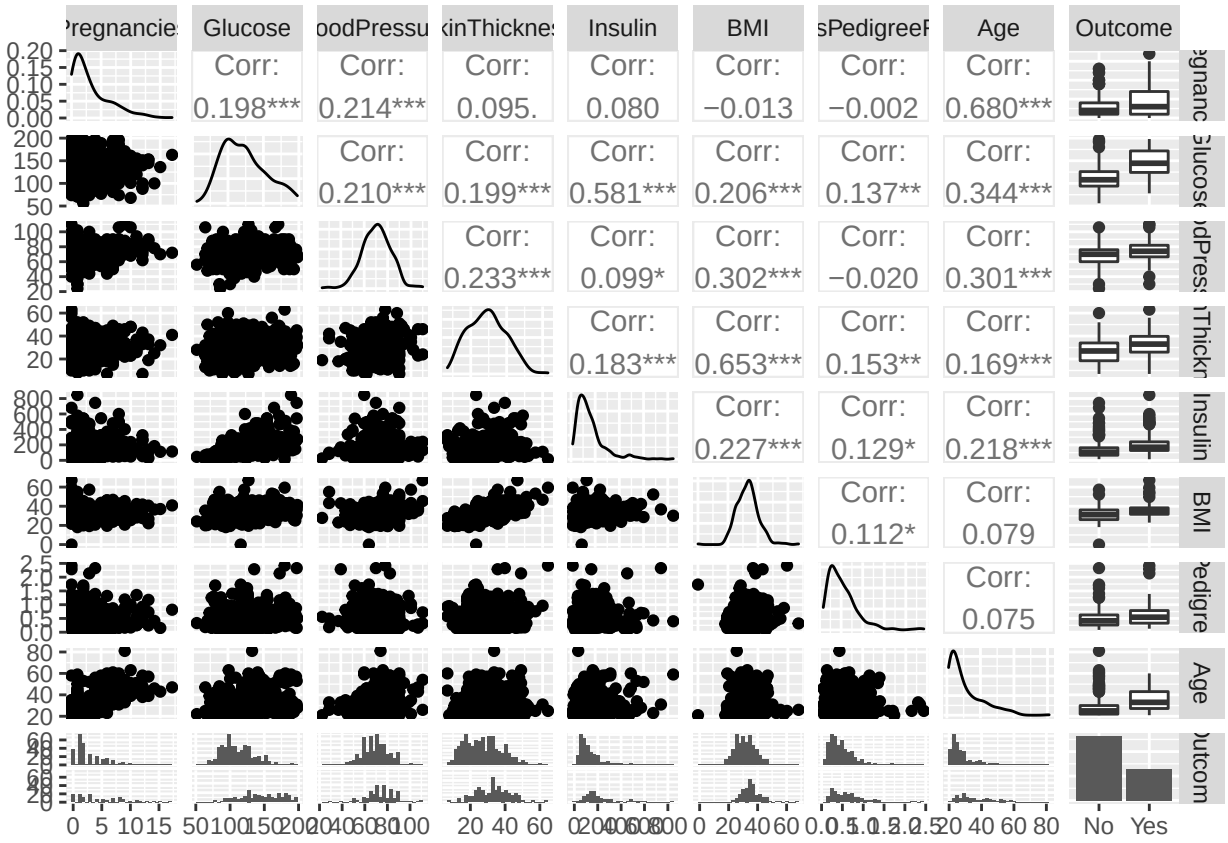
There were a lot of 0s in the plots and they might hurt the correlation so I removed them by Insulin then Glucose.

```
# Filter out 0 values in Insulin and plot pairs
noNullIns <- diabetes %>% filter(Insulin != 0)
GGally::ggpairs(noNullIns)
```



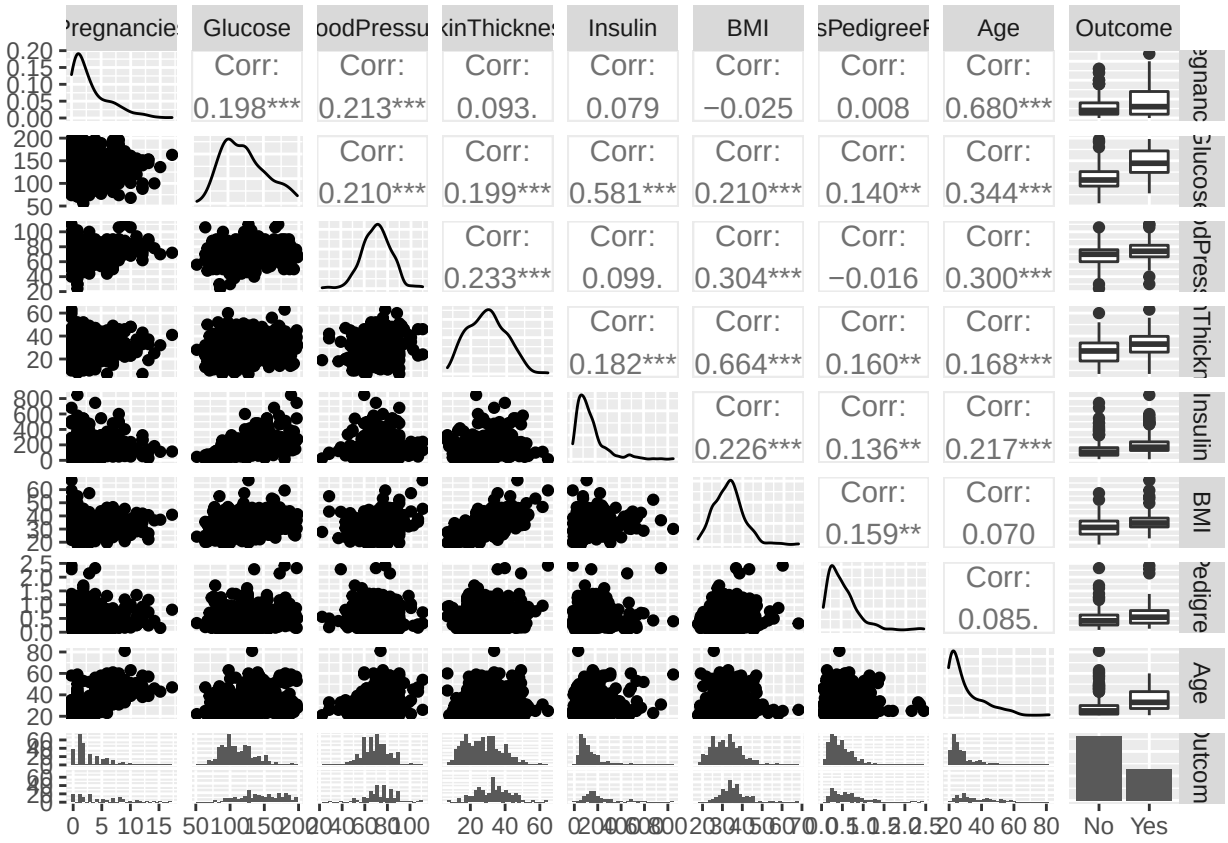
Correlation and graphs look better after taking out 0s from Insulin!

```
# Filter out 0 value in Glucose and plot pairs
noNullInsGlu <- noNullIns %>% filter(Glucose != 0)
GGally::ggpairs(noNullInsGlu)
```



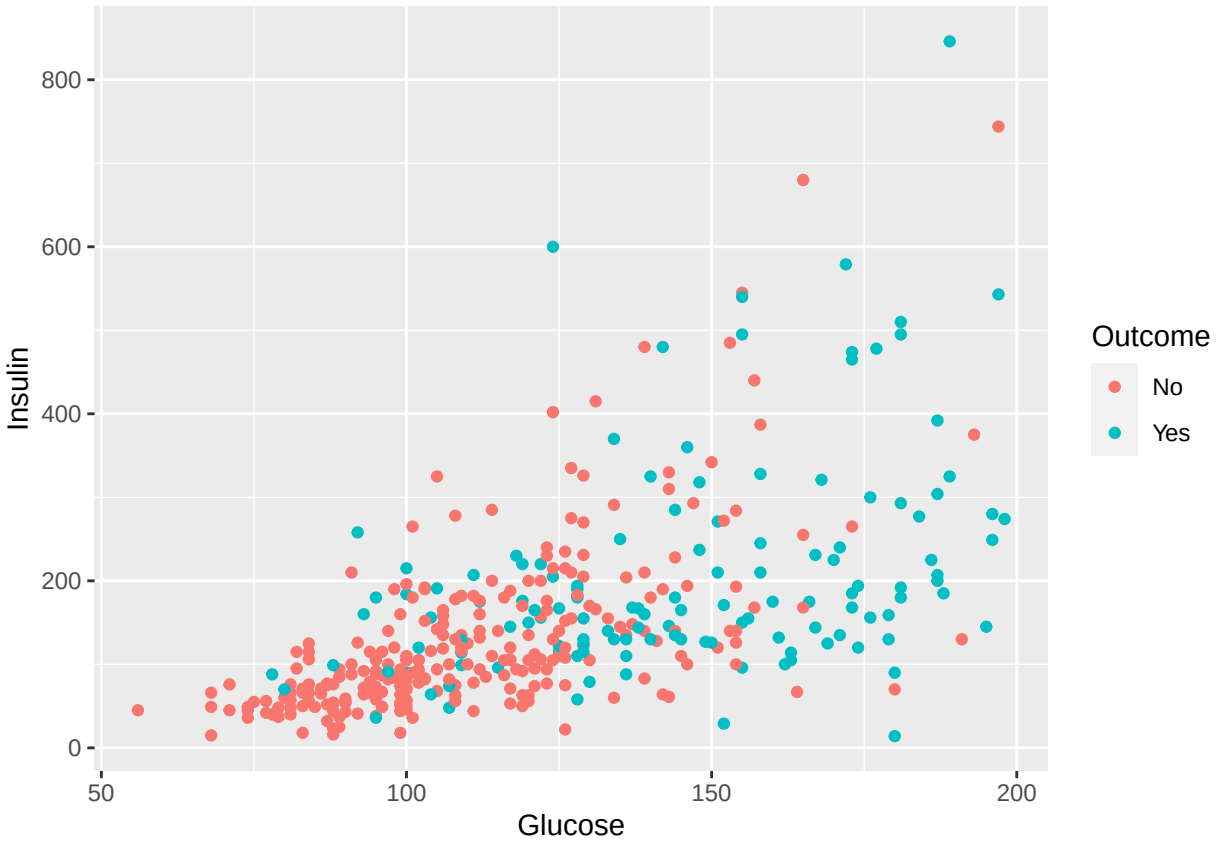
Some correlations dropped after taking the 0 from Glucose, but graphs look better.

```
# Filter out 0 value in Glucose and plot pairs
noNullInsGluBMI <- noNullInsGlu %>% filter(BMI != 0)
GGally::ggpairs(noNullInsGluBMI)
```



I noticed some 0s in BMI so I removed them as well.

```
# Scatterplot of Glucose and Insulin with Outcome
ggplot(noNullInsGluBMI, aes(Glucose, Insulin, color = Outcome)) +
  geom_point()
```



The variance was non-constant so Idk.

```
# Full model and summary
result <- lm(Glucose ~ Pregnancies + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeFu
summary(result)
```

```
##
## Call:
## lm(formula = Glucose ~ Pregnancies + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age, data = noNullInsGluBMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.185 -15.558  -3.087  11.847  74.331
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               60.03002    8.44119   7.112 5.65e-12 ***
## Pregnancies                0.07383    0.52315   0.141 0.887848
## BloodPressure              0.21341    0.10769   1.982 0.048219 *
## SkinThickness              0.07433    0.15769   0.471 0.637628
## Insulin                    0.13321    0.01084  12.293  < 2e-16 ***
## BMI                        0.13038    0.24389   0.535 0.593239
## DiabetesPedigreeFunction   4.17855    3.62412   1.153 0.249635
## Age                        0.57734    0.17182   3.360 0.000857 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.1 on 384 degrees of freedom
## Multiple R-squared:  0.4012, Adjusted R-squared:  0.3903
## F-statistic: 36.76 on 7 and 384 DF,  p-value: < 2.2e-16
```

Pregnancies, skin thickness, BMI, and diabetes pedigree function didn't look significant based on p-values in the presence of the other predictors. Partial F test will be conducted to see if we can drop these variables.

```
# Checking for multicollinearity
faraway::vif(result)
```

```
##              Pregnancies            BloodPressure            SkinThickness
##                 1.900621                 1.219344                 1.851701
##                  Insulin                      BMI DiabetesPedigreeFunction
##                 1.116662                 1.978124                 1.055661
##                      Age
##                 2.068613
```

Checking for multicollinearity signs in this model and it looks fine. Everything is definitely under 5 so we're in the clear.

```
reduced <- lm(Glucose ~ BloodPressure + Insulin + Age, data=noNullInsGluBMI)
summary(reduced)
```

```
##
## Call:
## lm(formula = Glucose ~ BloodPressure + Insulin + Age, data = noNullInsGluBMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.576 -15.015  -3.763  12.144  78.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.4659     7.2187   9.069  < 2e-16 ***
## BloodPressure   0.2423     0.1022   2.371   0.0182 *
## Insulin         0.1372     0.0105  13.063  < 2e-16 ***
## Age             0.6036     0.1276   4.730 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.07 on 388 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3915
## F-statistic: 84.87 on 3 and 388 DF,  p-value: < 2.2e-16
```

Everything is significant in the reduced model so we'll run with this.

```
# Checking for multicollinearity
faraway::vif(reduced)
```

```
## BloodPressure       Insulin           Age
##      1.100343      1.050805      1.143554
```

No signs of multicollinearity here either.

```
# Conducting partial F test
anova(reduced, result)
```

```
## Analysis of Variance Table
##
## Model 1: Glucose ~ BloodPressure + Insulin + Age
## Model 2: Glucose ~ Pregnancies + BloodPressure + SkinThickness + Insulin +
##     BMI + DiabetesPedigreeFunction + Age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    388 224845
## 2    384 222975  4      1870 0.8051 0.5224
```

Insignificant p-value so we failed to reject the null and favor the reduced model.