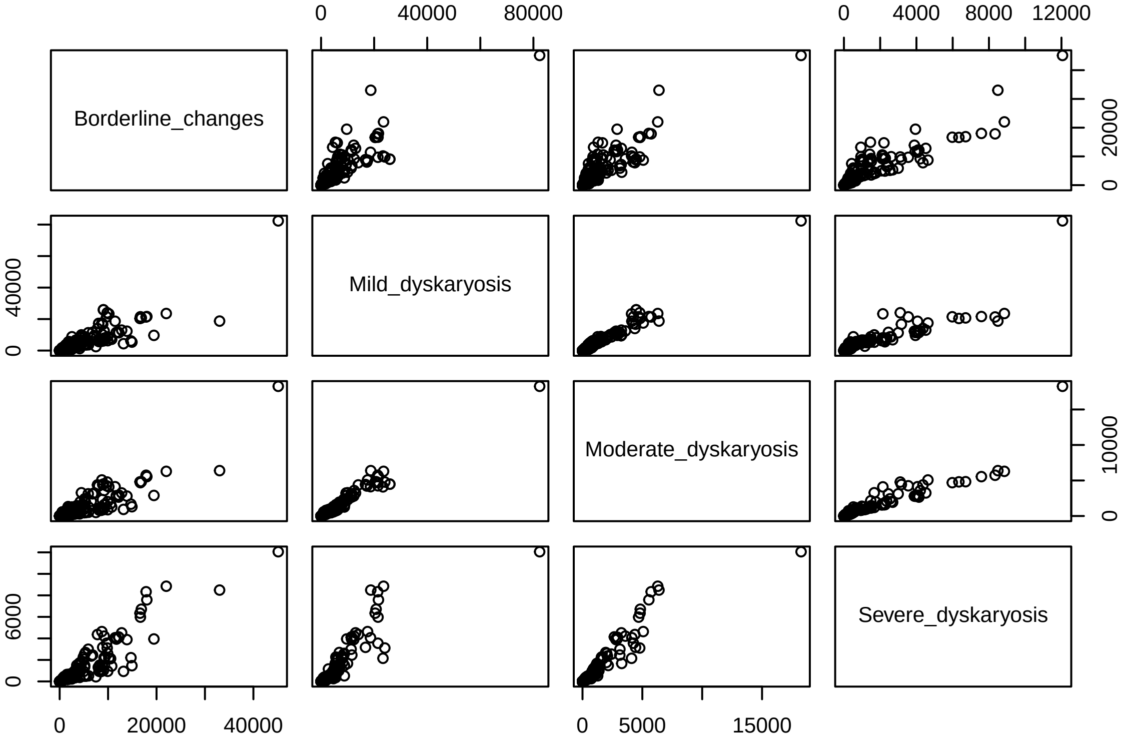
The above was all copied from the project check in 1 document. It will not appear in the knitted file because it is suppressed.

First, we will subset the data for borderline changes, mild, moderate, and severe dyskaryosis into a separate data frame. Then the pairs function will be run to determine any reasonable relationships.

```
## Subset source data for only borderline changes and mild, moderate, severe dyskaryosis

comparison_data <- source[,c("Borderline_changes", "Mild_dyskaryosis", "Moderate_dyskaryosis","Severe_dy
comparison_data$Severe_dyskaryosis <- as.numeric(comparison_data$Severe_dyskaryosis)

pairs(comparison_data)
```



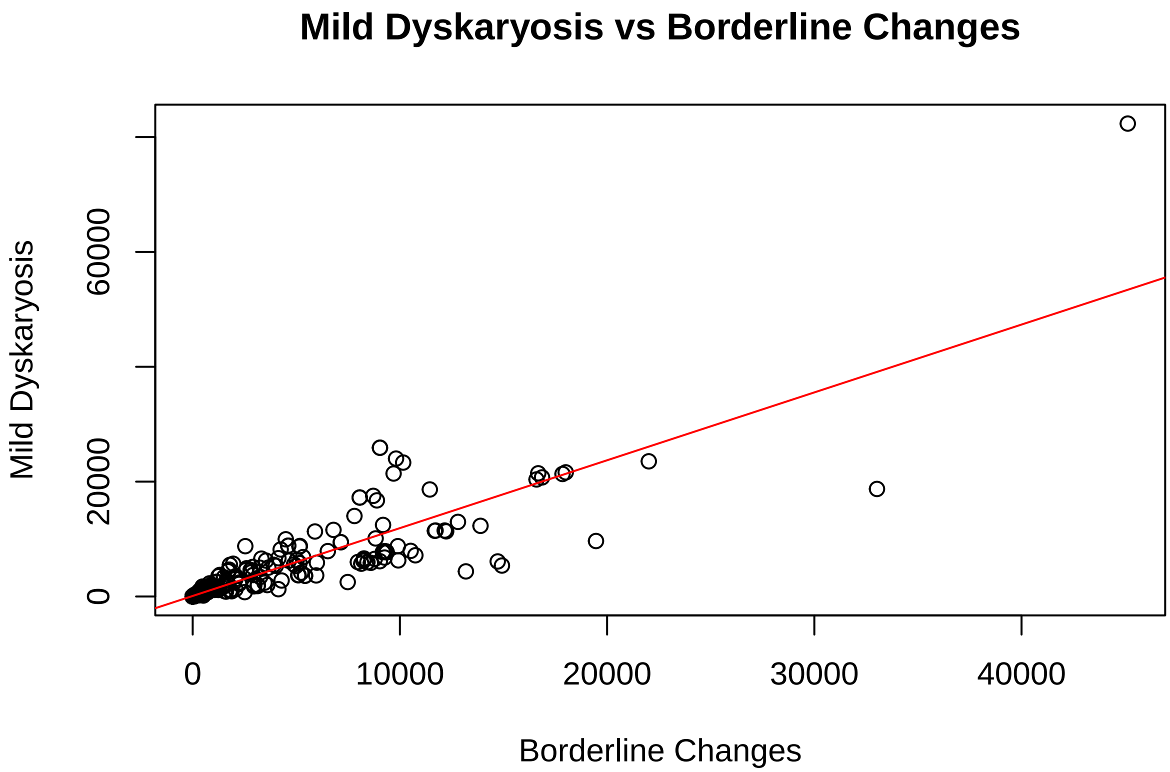
First, we will run a linear regression for all of these relationships as they appear to have decently strong relationships.

Unscaled plots

```
# Mild Dyskaryosis
lm_model_mild <- lm(Mild_dyskaryosis ~ Borderline_changes, data = comparison_data)
summary(lm_model_mild)
```

```
##
## Call:
## lm(formula = Mild_dyskaryosis ~ Borderline_changes, data = comparison_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20385.1   -115.6    -84.5    190.5  28938.7
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         88.63154  210.22556   0.422    0.674
## Borderline_changes   1.18161    0.03679  32.117   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3338 on 328 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.758
## F-statistic:  1031 on 1 and 328 DF,  p-value: < 2.2e-16
```

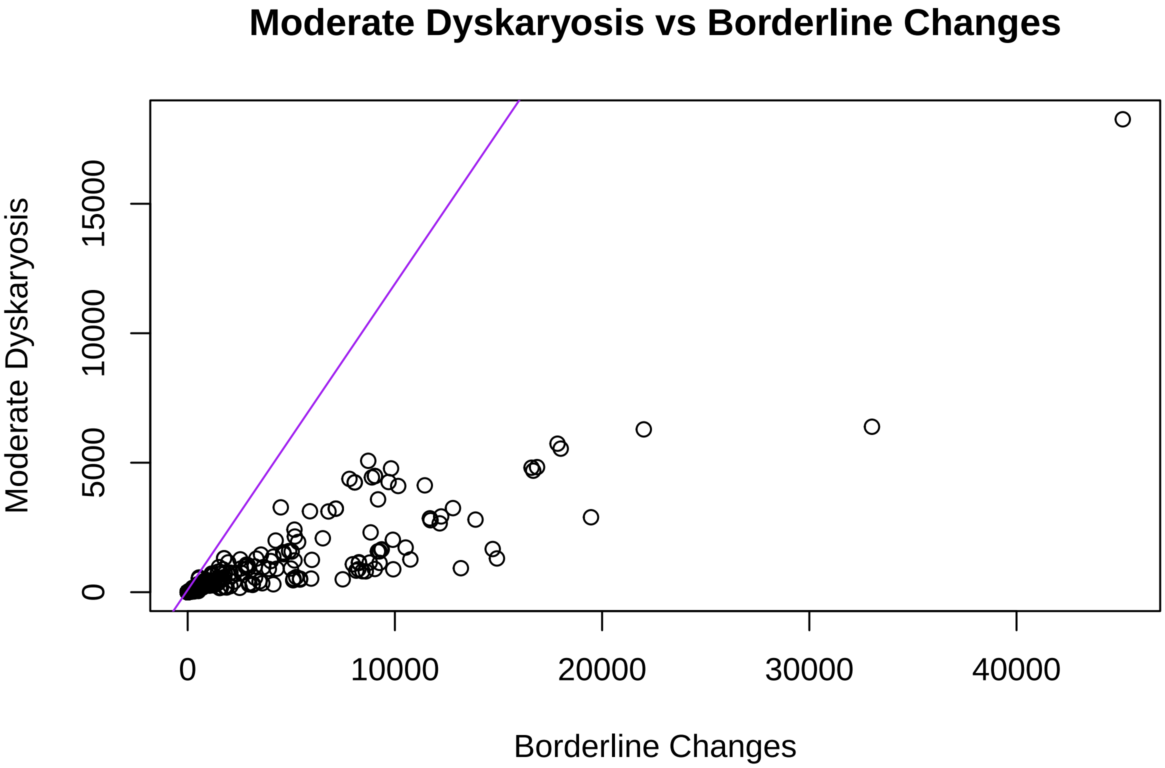
```
plot(comparison_data$Borderline_changes, comparison_data$Mild_dyskaryosis,
     main = "Mild Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Mild Dyskaryosis")
abline(lm_model_mild, col = 'red')
```



```
# Moderate Dyskaryosis
lm_model_mod <- lm(Moderate_dyskaryosis ~ Borderline_changes, data = comparison_data)
summary(lm_model_mod)
```

```
##
## Call:
## lm(formula = Moderate_dyskaryosis ~ Borderline_changes, data = comparison_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2835.2    -1.0     6.0    80.4  5741.2
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.414850  46.102620  -0.117    0.907
## Borderline_changes  0.277578   0.008068  34.403   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 732.1 on 328 degrees of freedom
## Multiple R-squared:  0.783,  Adjusted R-squared:  0.7823
## F-statistic:  1184 on 1 and 328 DF,  p-value: < 2.2e-16
```

```
plot(comparison_data$Borderline_changes, comparison_data$Moderate_dyskaryosis,
     main = "Moderate Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Moderate Dyskaryosis")
abline(lm_model_mild, col = 'purple')
```

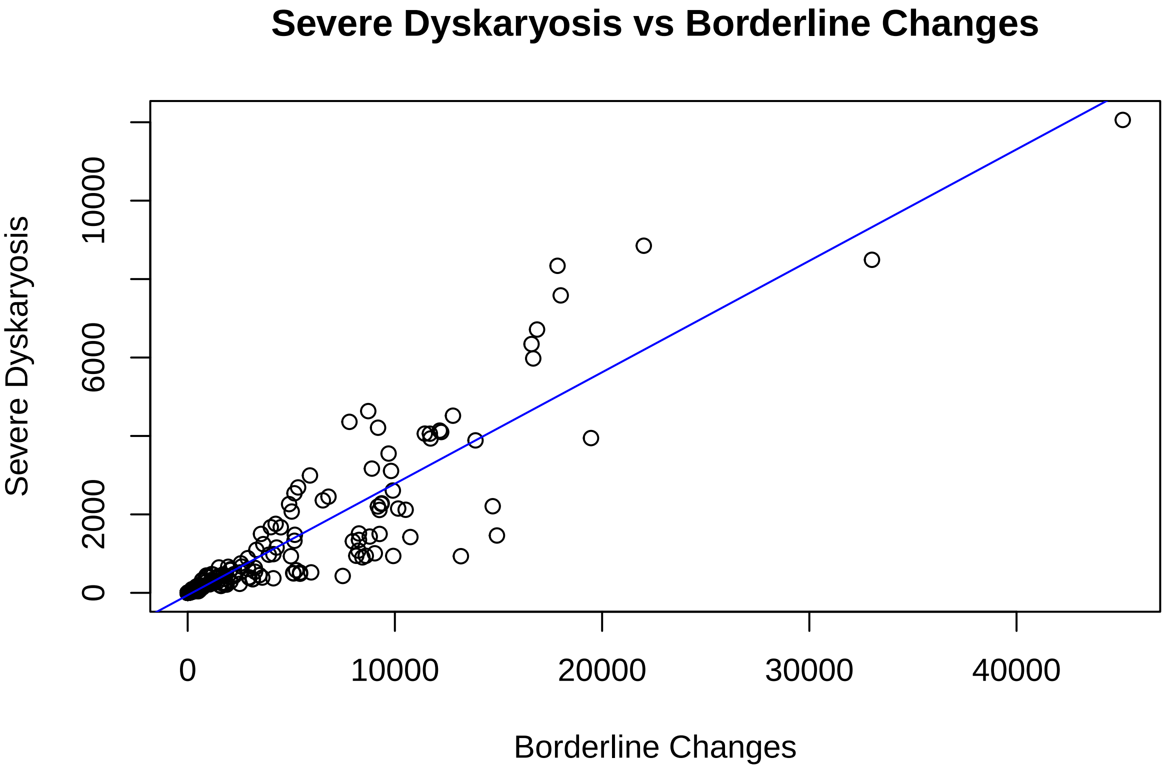


```
# Severe Dyskaryosis
lm_model_severe <- lm(Severe_dyskaryosis ~ Borderline_changes, data = comparison_data)
summary(lm_model_severe)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes, data = comparison_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2752.9   -48.3    51.9    61.2  3327.0
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -60.208265  52.390598  -1.149    0.252
## Borderline_changes   0.284191   0.008057  35.271   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

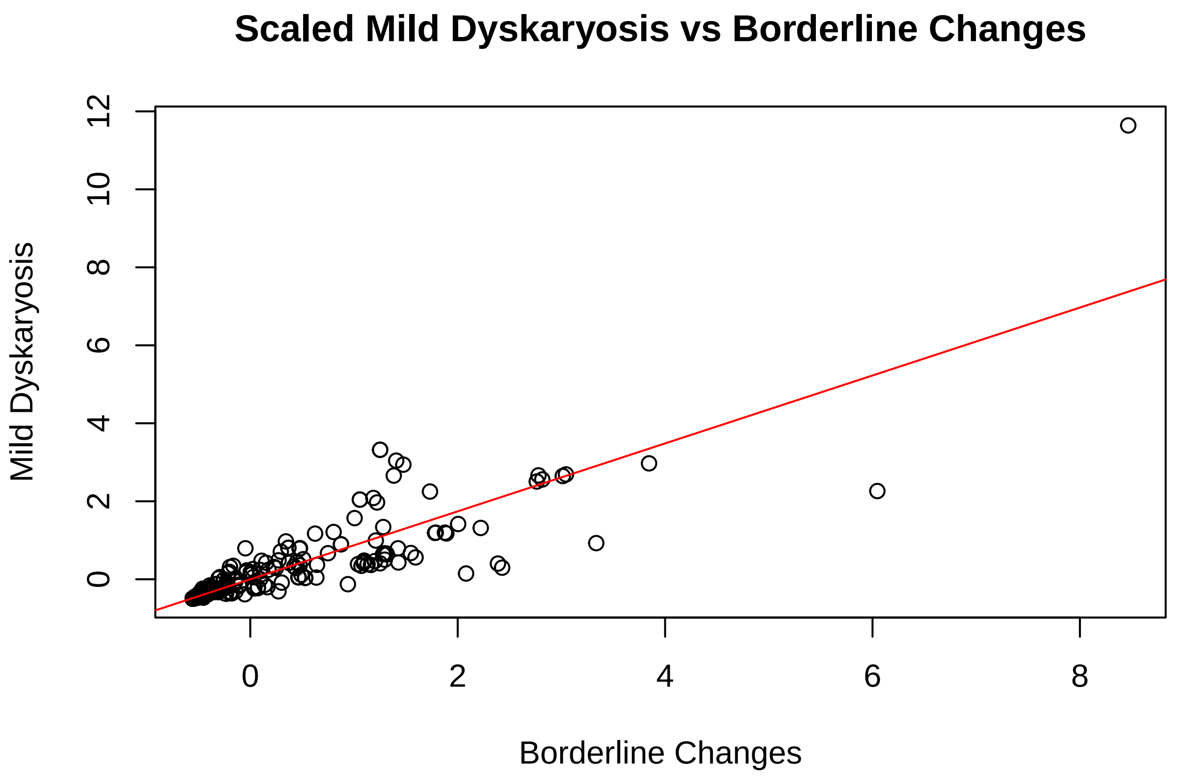
```
## Residual standard error: 692.2 on 238 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8387
## F-statistic:  1244 on 1 and 238 DF,  p-value: < 2.2e-16
```

```
plot(comparison_data$Borderline_changes, comparison_data$Severe_dyskaryosis,
     main = "Severe Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Severe Dyskaryosis")
abline(lm_model_severe, col = 'blue')
```



Scaled Plots

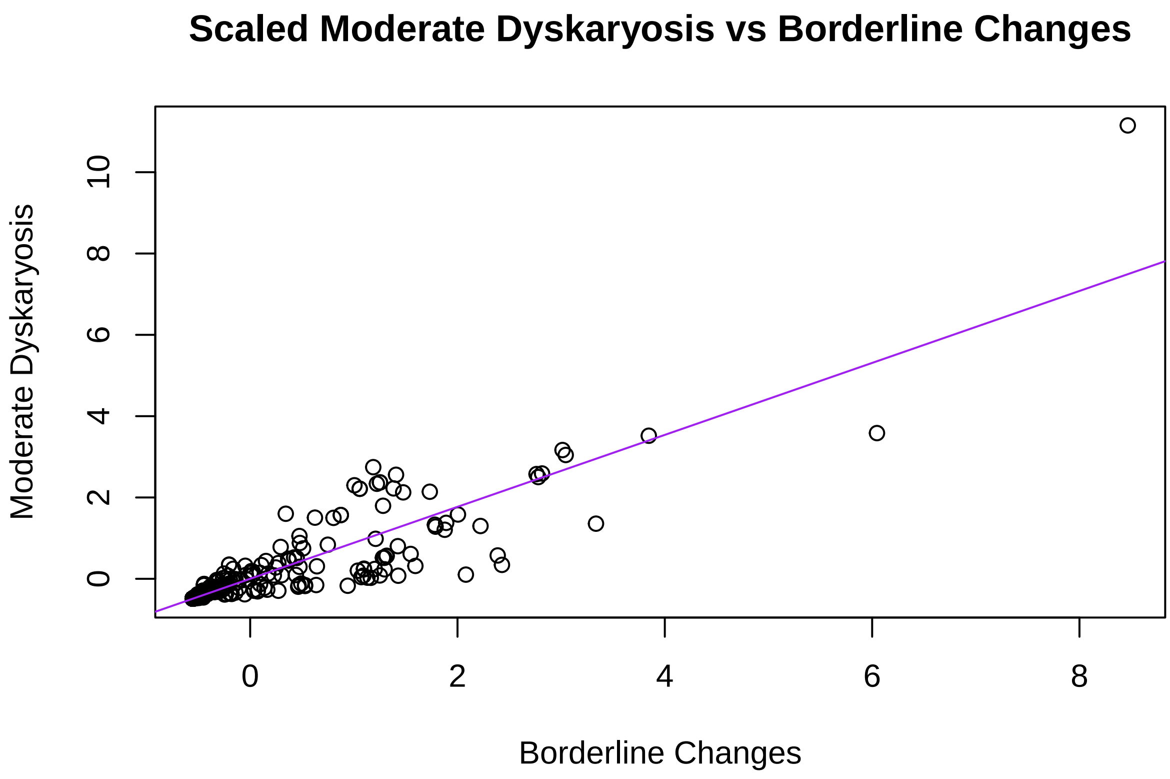
```
scaled_data <- as.data.frame(scale(comparison_data))

# mild dyskaryosis scaled
scaled_mild_model <- lm(Mild_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Mild_dyskaryosis,
     main = "Scaled Mild Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Mild Dyskaryosis")
abline(scaled_mild_model, col = 'red')
```



```
summary(scaled_mild_model)
```

```
##
## Call:
## lm(formula = Mild_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0040 -0.0170 -0.0125  0.0281  4.2644
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.798e-16  2.708e-02    0.00        1
## Borderline_changes 8.711e-01  2.712e-02   32.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4919 on 328 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.758
## F-statistic:  1031 on 1 and 328 DF,  p-value: < 2.2e-16
```

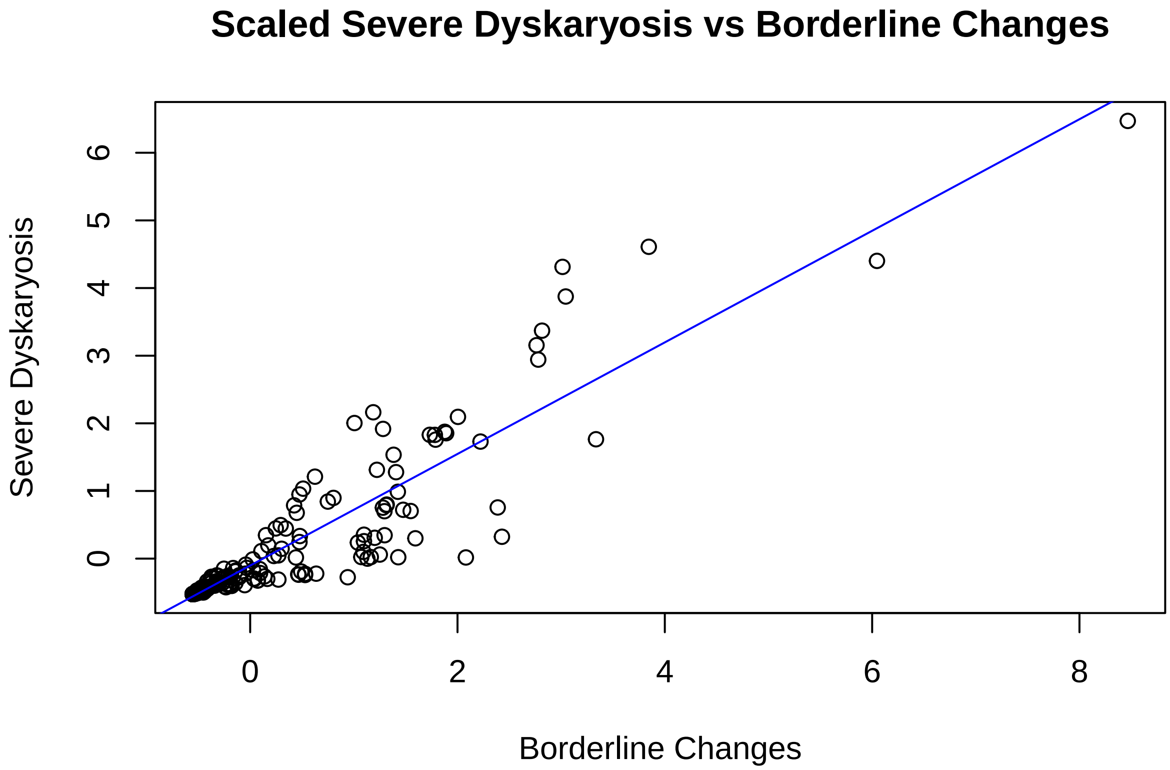
```
# moderate dyskaryosis scaled
scaled_mod_model <- lm(Moderate_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Moderate_dyskaryosis,
     main = "Scaled Moderate Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Moderate Dyskaryosis")
abline(scaled_mod_model, col = 'purple')
```



```r
summary(scaled_mod_model)
```

```
##
## Call:
## lm(formula = Moderate_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8067 -0.0007  0.0038  0.0512  3.6586
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.967e-16  2.568e-02     0.0        1
## Borderline_changes 8.849e-01  2.572e-02    34.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4665 on 328 degrees of freedom
## Multiple R-squared:  0.783,  Adjusted R-squared:  0.7823
## F-statistic:  1184 on 1 and 328 DF,  p-value: < 2.2e-16
```

```r
# severe dyskaryosis scaled
scaled_severe_model <- lm(Severe_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Severe_dyskaryosis,
     main = "Scaled Severe Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Severe Dyskaryosis")
abline(scaled_severe_model, col = 'blue')
```



```r
summary(scaled_severe_model)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59706 -0.02801  0.03012  0.03551  1.93013
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.10225    0.02608   -3.92 0.000116 ***
## Borderline_changes  0.82478    0.02338   35.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4016 on 238 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8387
## F-statistic:  1244 on 1 and 238 DF,  p-value: < 2.2e-16
```

For the first linear regression model of Mild Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001. The intercept of the line was not significant at any level. The adjusted R-squared value is 0.758, indicating that 75.8% of the variance in the data can be explained by the model. The F statistic is large with a p-value < 0.05, indicating that this regression is significant.

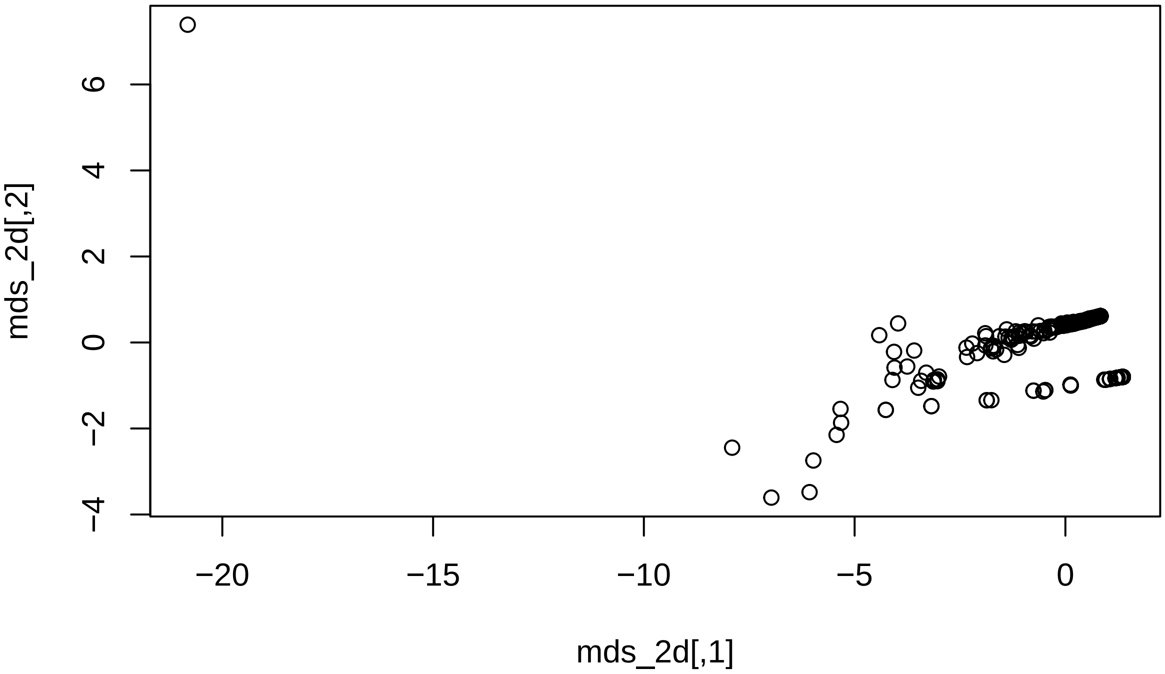
For the second linear regression model of Moderate Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001. The intercept of the line was not significant at any level. The adjusted R-squared value is 0.7823, inidcating that 78.23% of the variance in the data can be explained by the model. The F statistic is large with a p-value < 0.05, indicating that this regression is significant.

For the third linear regression model of Severe Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001. The intercept of the line was not significant at any level. The adjusted R-squared value is 0.8387, indicating that 83.87% of the variance in the data can be explained by the model. The F statistic is large with a p-value of < 0.05, indicating that this regression is significant.

The regression model of Severe Dyskaryosis vs Borderline Changes appears to be the best at predicting incidence of severe dyskaryosis from borderline changes due to it having the largest F-statistic value (1244) as well as the largest adjusted R-squared value (0.8387). This could be because borderline changes in a neoplasma will typically be larger and more easily detectable in later stage cancers, meaning that large borderline changes would have a higher correlation with severe dyskaryosis.

Clustering

```r
pairwise_dist <- dist(scaled_data)
mds_2d <- cmdscale(pairwise_dist, k = 2)
plot(mds_2d)
```


mds_2d[,2]
mds_2d[,1]