Final Project Report

Raynah Cheng

2024-12-07

Abstract:

The incidence of cervical cancer is a widespread problem that affects many women across the world. Additionally, this is one of the most easily preventable cancers through the use of different screening tests such as Pap smears and liquid-based cytology. Prediction of severity of cancer neoplasms is useful to determining the stage of cancer and can provide important information regarding treatments needed and incidence rates. Using data from the NHS in England, linear regressions, a BIC/MIC model, a LASSO model, and a parametric model were constructed as candidate models in order to predict various aspects of cervical cancer incidence and neoplasms. The final model chosen was the LASSO model, but all of the models provide a different aspect of clinical significance.

Introduction:

Cervical cancer is a type of cancer that occurs in the cervix, specifically the cells lining the cervix wall. It is the fourth most common cancer in women globally with around 660,000 new cases and 350,000 deaths in 2022, according to the WHO. Our project contains data on patients that have cervical cancer, taken from the NHS England. There were many datasets, but the one that we focused on was titled k61_sample_result_age_source. In this dataset, there are 330 observations, with 5 variables of focus. We chose on this one because it contained data on cervical cancer changes based on age and screening. The relevant variables we are looking at are mild, moderate, and severe dyskaryosis, age, and borderline changes. These terms are broad and hard to understand, so defining them is useful. Dyskaryosis is the change in appearance of cells that line the cervix. Mild, moderate, and severe are the classifications that this dataset uses. Mild is little change, moderate is moderate change, and severe is severe change. These changes are defined based on a 3-year screening. Borderline changes are abnormal changes in the cervix that are often pre-cancerous. These could indicate cervical cancer.

Data Cleaning:

The dataset for Cervical Screening Program was taken from the NHS England, published on November 23, 2023 and covers the national screening data as well as local and clinical commissioned data. Out of the documents that were provided by the NHS, the data frame including the ages and results of diagnoses was selected for this project (`kc61_sample_result_age_Source.csv`). This set of data includes the information that was of most interest, including age and severity of diagnosis.

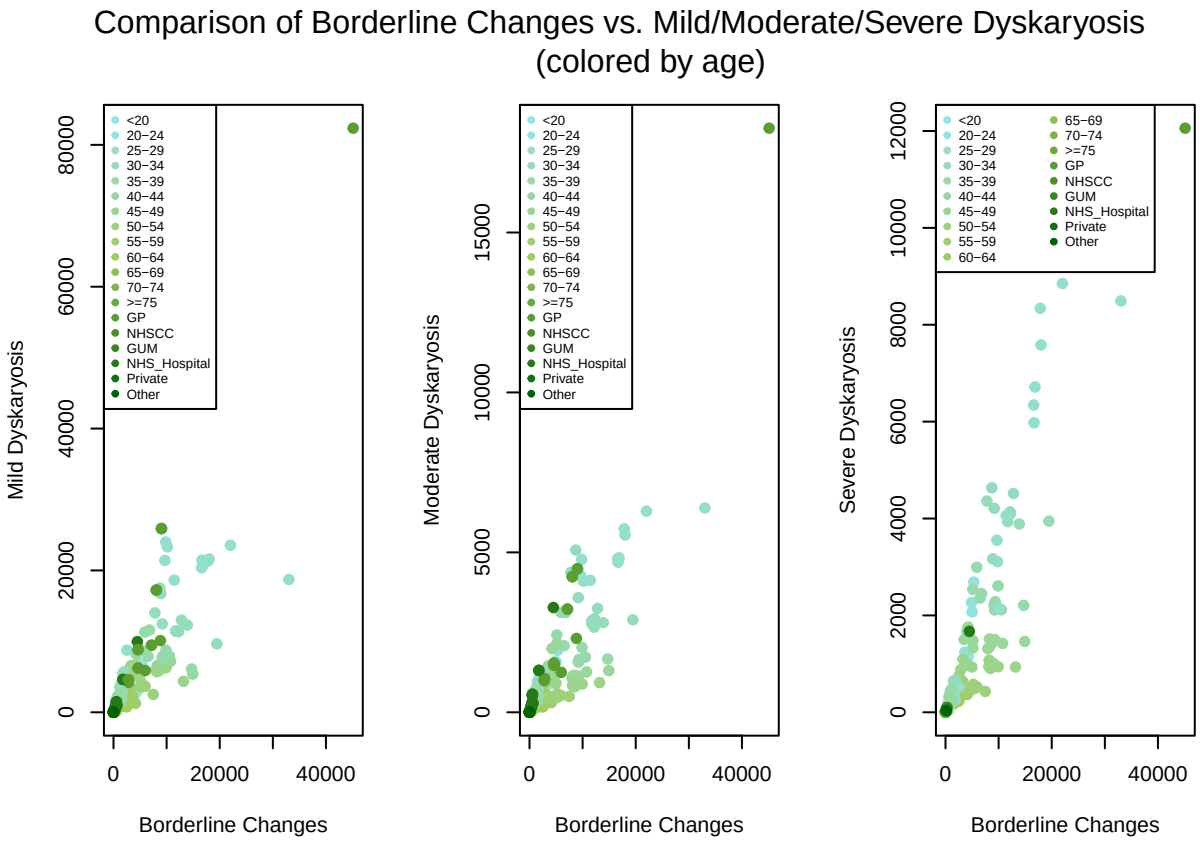
The initial data frame had 330 observations for 17 variables, and the interested variables were recorded in a multitude of fashions. The `Indicator` variable was recorded as characters for age ranges. The severity of the diagnosis (`Borderline Changes`, `Mild Dyskaryosis`, `Moderate Dyskaryosis`, and `Severe Dyskaryosis`) were recorded as numeric variables, explaining the number of diagnoses that were observed for each age range, and ome variables had missing values recorded as “not included”.

In order to properly use the data, the "not included" observations were replaced with `NA` values and then removed from the data frame. The `Indicator` variable was then renamed to be a number from 1 through 19. Each number corresponds to a specific age range or source of sample.

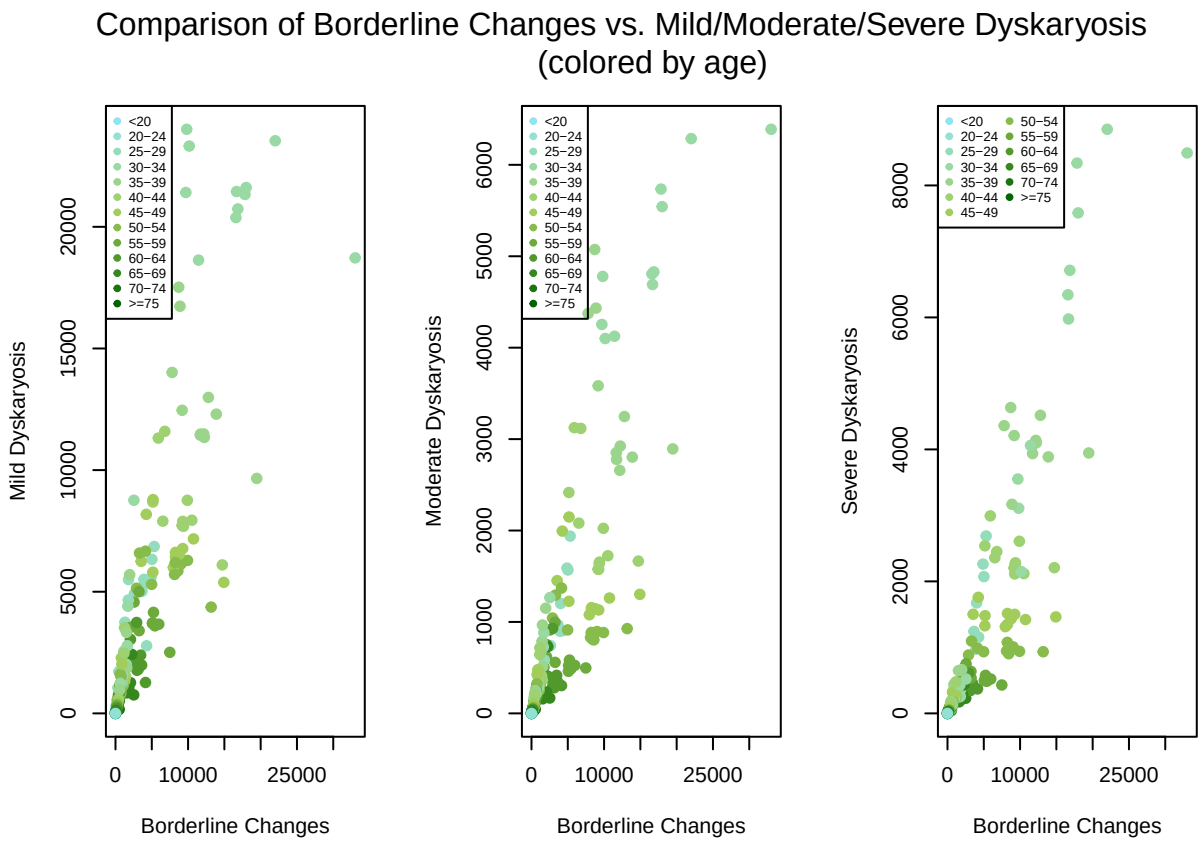
- 1: <20 years old (non-inclusive)
- 2: 20-24 years old (inclusive)
- 3: 25-29 years old (inclusive)
- 4: 30-34 years old (inclusive)
- 5: 35-39 years old (inclusive)
- 6: 40-44 years old (inclusive)
- 7: 45-49 years old (inclusive)
- 8: 50-54 years old (inclusive)
- 9: 55-59 years old (inclusive)
- 10: 60-64 years old (inclusive)
- 11: 65-69 years old (inclusive)
- 12: 70-74 years old (inclusive)
- 13: >=75 years old (inclusive)
- 14: data taken from GP (General Practitioners)
- 15: data taken from NHSCC (NHS Clinical Commissioners)
- 16: data taken from GUM (Genitourinary Medicine)
- 17: data taken from NHS Hospital
- 18: data taken from Private sources
- 19: data taken from Other sources

Using the new named indicators, each group was separated to be its own data frame. This concluded the general data cleaning, data was further cleaned for the purpose of running plots and models.

Exploratory Data Analysis

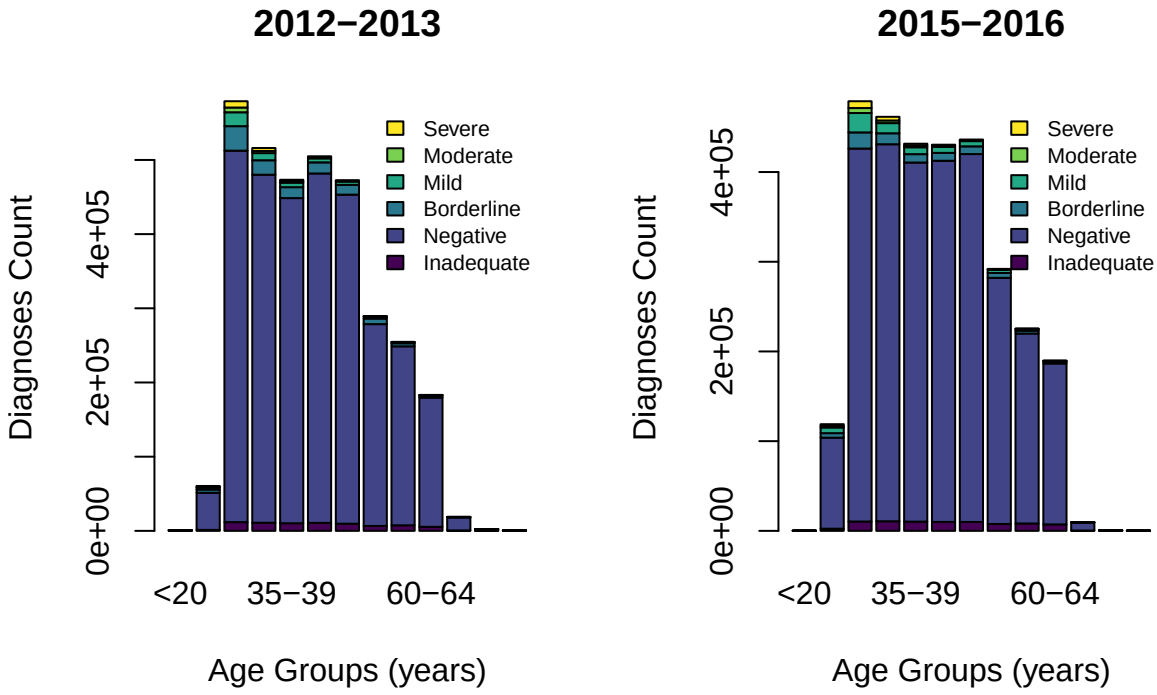


The inital scatterplots were made to observe any correlations between the number of borderline changes and the mild, moderate, or severe dyskariosis changes. The plot was organized by the indicator, sorting the different age groups and data sources. Looking at this scatterplot, the bulk of the data is clustered around the lower left corner of the graph, having a general positive trend. There looks to be one outlier at the upper right corner of the graph. The weird distribution of the data prompted the dataset to be reexamined and the observations taken from the separate data sources showed to be throwing off the data. Because there isn't consistency or details about how the data was recorded at each of these individual sources, those variables were taken out.

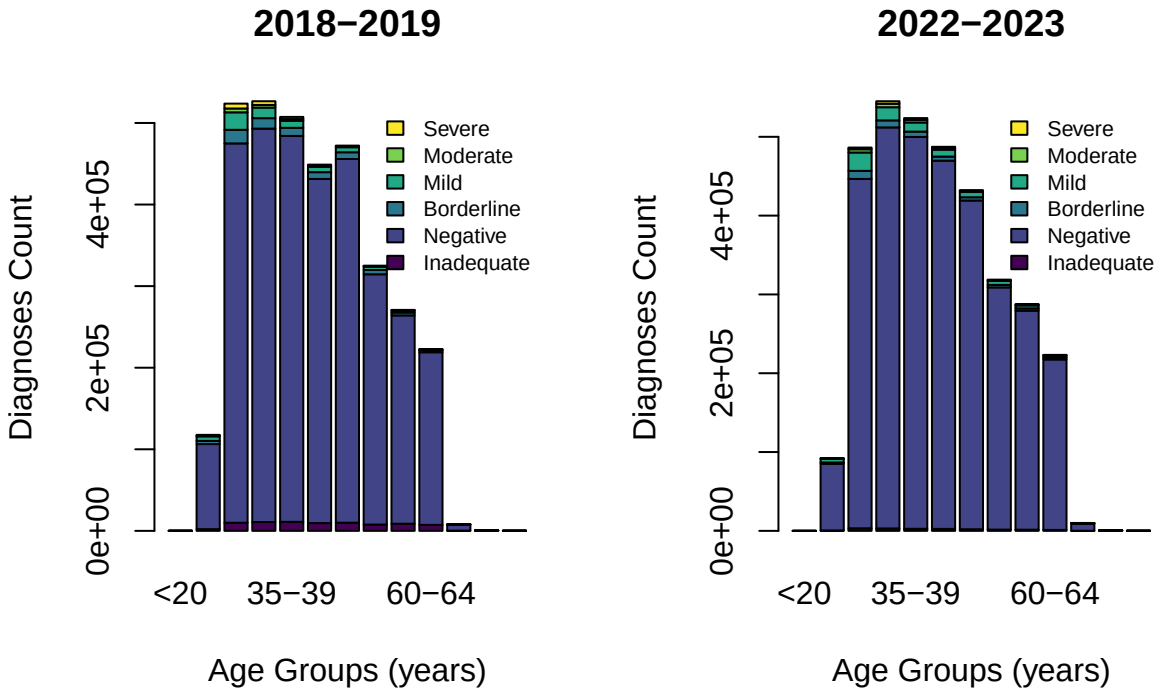


This scatterplot gives a better idea of how the data changing for each age group, however it is still difficult to visualize the difference between each variable group. The data was once again reexamined to see if there were any confounding variables that caused the distribution of the data. Looking back at the data, it was discovered that the information was separated between national, regional, and private data. National data contained information of data taken between 2012-2023, while regional and private data were taken only from 2022-2023. The descrepancies in the timeframe the data was collected promted to change vizualization of the data to bar plots, looking only at the National data with specific timeframes.

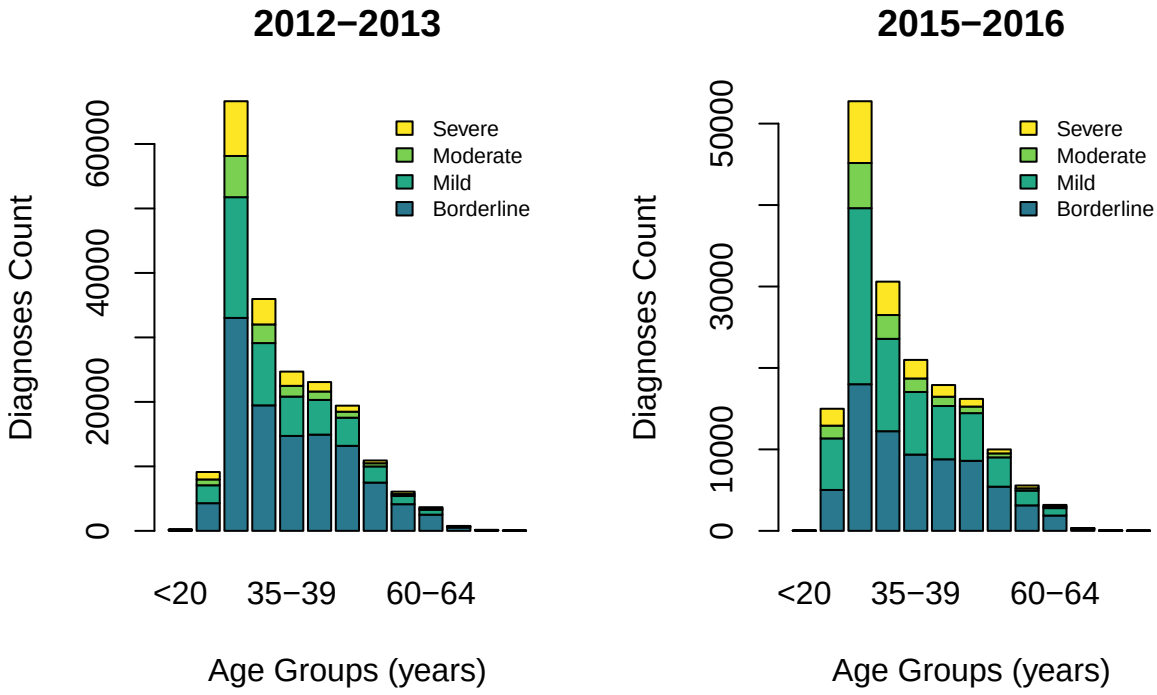
Cervical Screening Distribution (England)

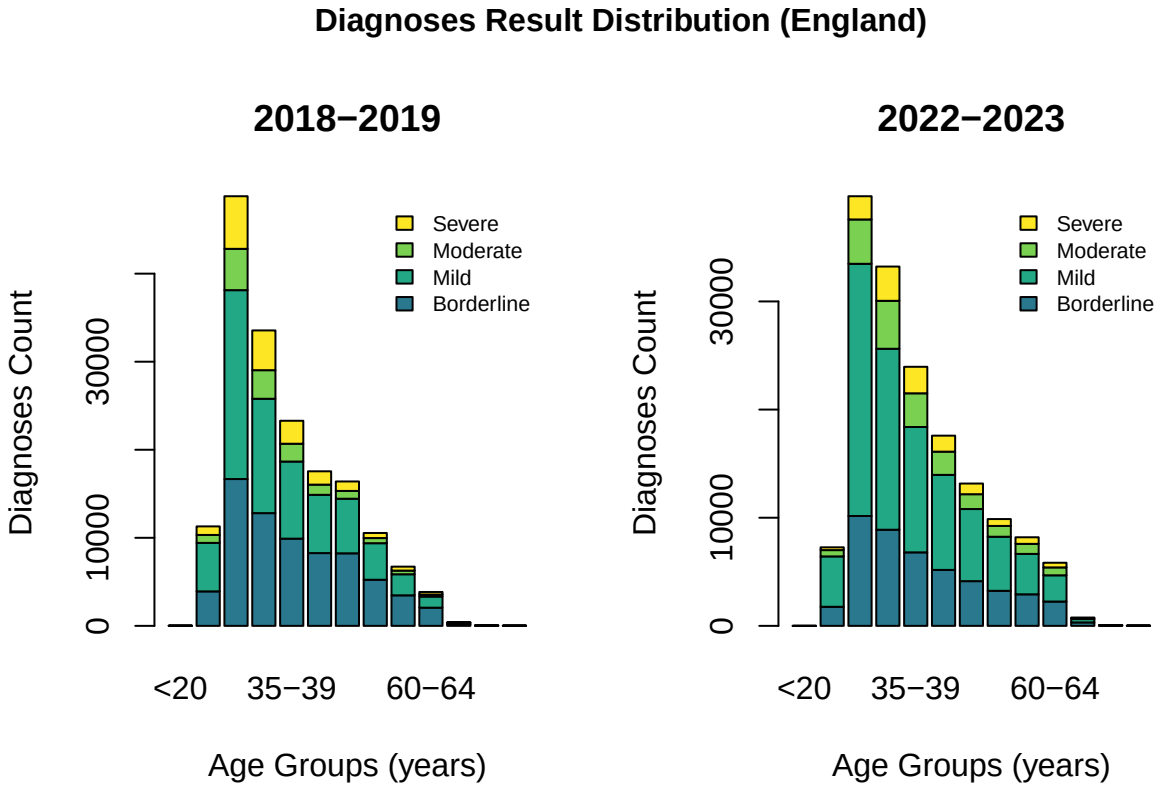


Cervical Screening Distribution (England)



Diagnoses Result Distribution (England)

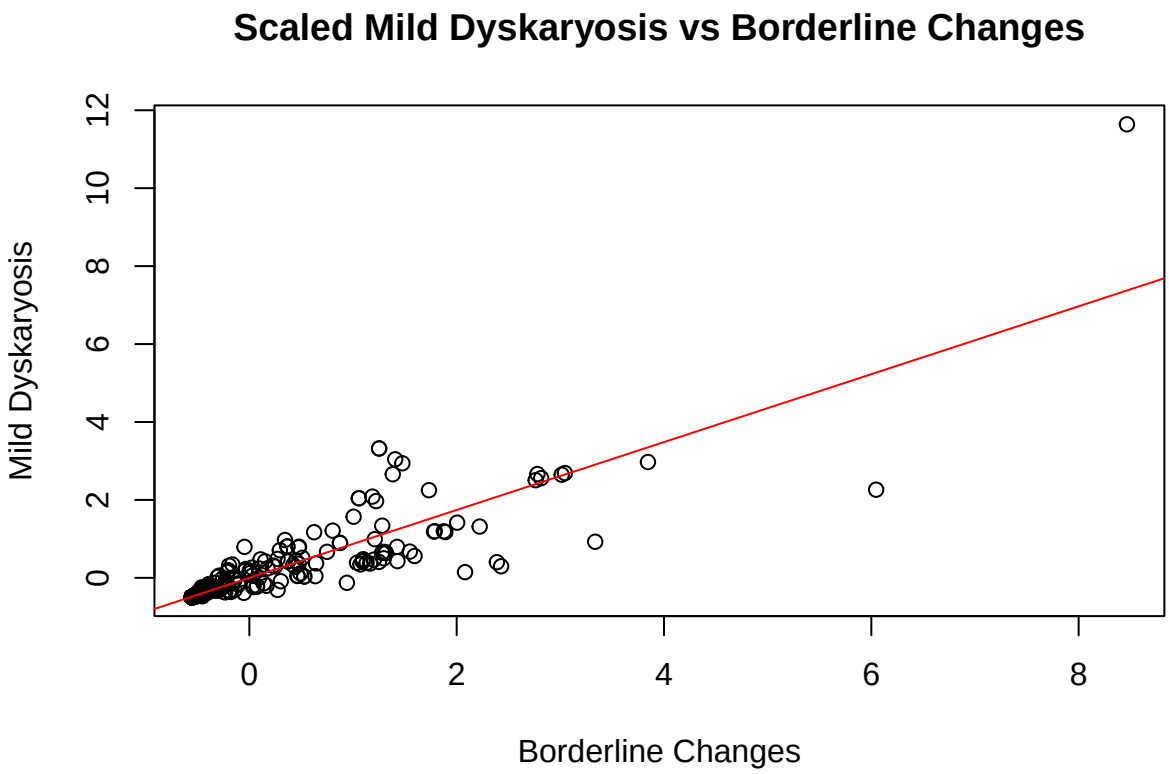




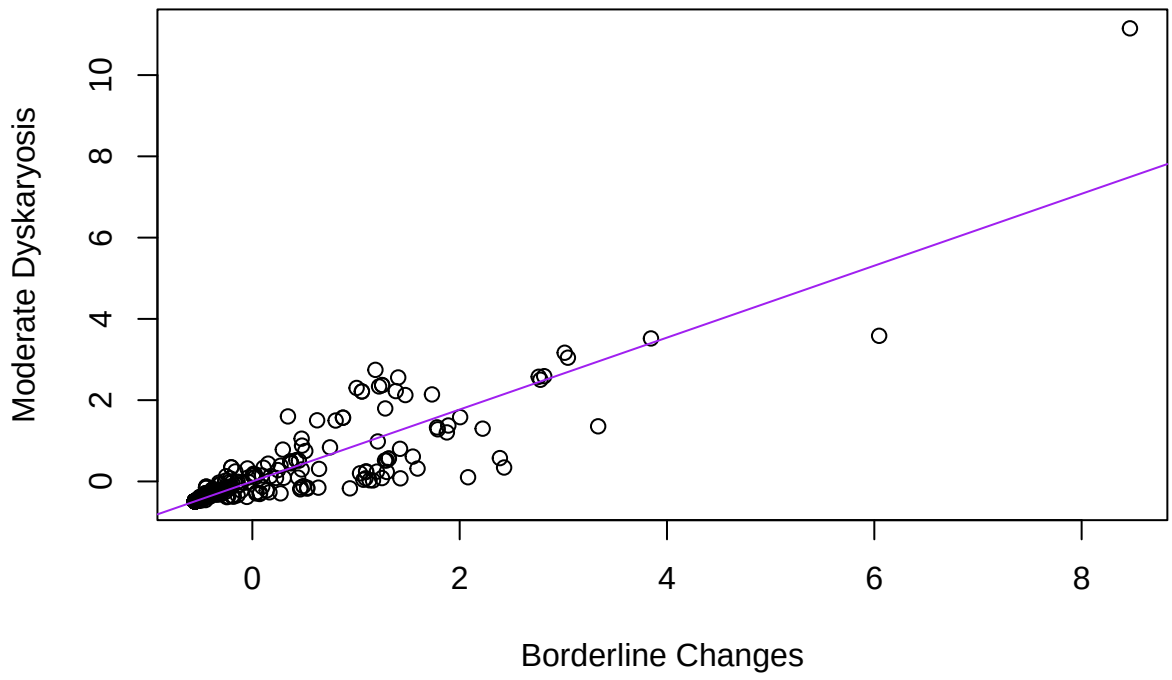
From these bar graphs, we can observe that the general trend of diagnoses has decreased throughout the years. This means that there is an absolute decrease in the amount of patients that are diagnosed for any type of cervical tissue changes. We also see that the number of diagnoses for severe dyskaryosis has decreased as well. This change could be explained by the increasing administration of the HPV vaccine.

Results:

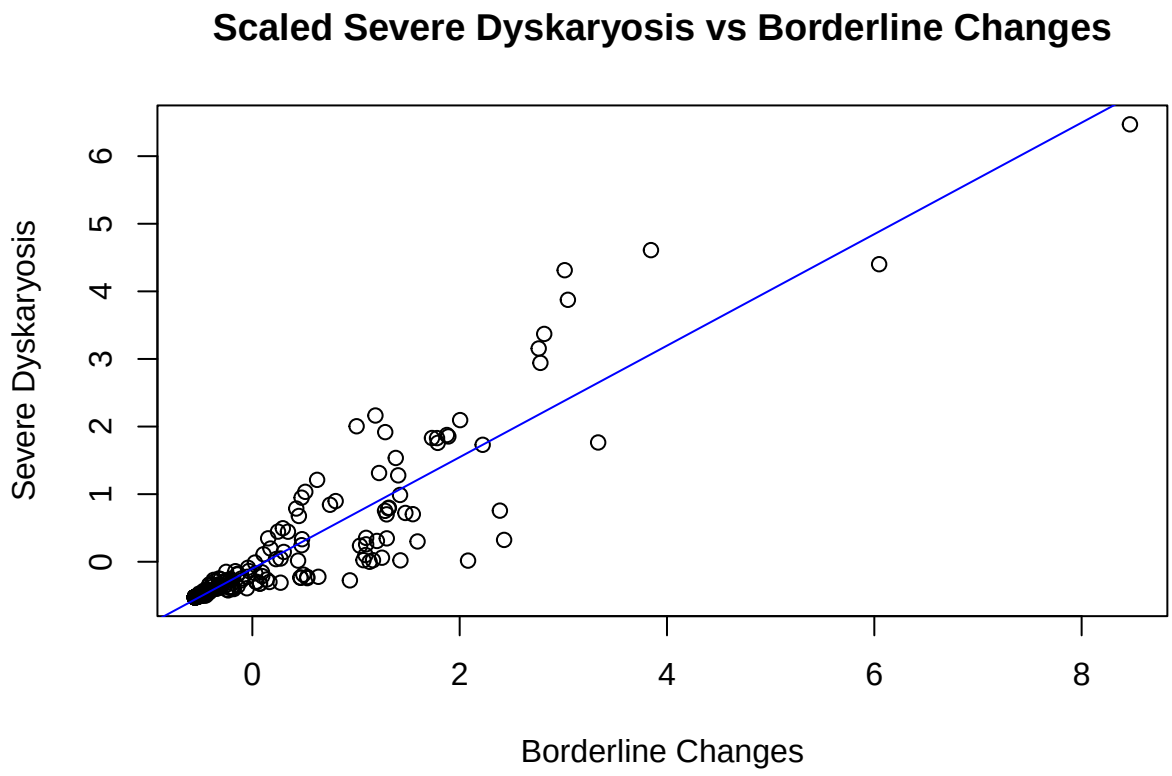
Scaled Linear Regression:




```
##
## 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)        2.213e-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
```




```
##
## 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)        2.621e-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
```




```
##
## 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
```

Borderline changes are defined by visual changes in the borders of the cancer neoplasm. In this dataset, the number of people who have presented with an abnormally shaped neoplasm are counted based on clinic site and age range and recorded in borderline changes. Dyskaryosis by definition is the abnormal appearance of a cell which is due to having an abnormal nucleus. The nucleus of the cell is where all the DNA is stored, so irregularities seen in the nucleus imply that there are issues with the DNA as well. DNA structural issues

and chromosomal issues imply cancer, which is why dyskaryosis is tested for using pap smears and liquid cytology tests. In this dataset, the counts of mild, moderate, and severe dyskaryosis were all measured by clinic site and age range.

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, indicating 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. Therefore, it is reasonable to assume a relationship between a higher number of patients presenting with borderline changes as well as a larger number of patients presenting with severe dyskaryosis.

BIC MIC Model Comparison:

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ ., data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes     0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis   1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16


##
## Call:
```

```
## lm(formula = Severe_dyskaryosis ~ Borderline_changes + Mild_dyskaryosis +
##     Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes     0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis   1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = Severe_dyskaryosis ~ Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2625.70   -49.93    56.48    68.43  1635.61
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -61.79307   48.85798  -1.265    0.208
## Moderate_dyskaryosis   1.17963    0.02726  43.274   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 508.9 on 163 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic:  1873 on 1 and 163 DF,  p-value: < 2.2e-16
```

The BIC model chooses all variables: borderline changes, mild dyskaryosis, and moderate dyskaryosis, with about equal levels of significance. However, we can see that the moderate dyskaryosis t-value is higher, at 15.163. For our BIC model, 95.16% of the variance can be explained by our model. The MIC model only chooses the moderate dyskaryosis variable, at an alpha level of <.001, with a much higher t-value than the BIC model. However, the variance that can be explained by the model goes down with a value of 91.99%.

Model	R^2	RMSE	MAE
MIC	.997	1128.59	248.29
BIC	.996	806.2	194.54

The BIC model has a lower R^2 value (.996), but the RMSE and MAE values are lower (806.2, 194.54). The MIC model has a higher R^2 value (.997), but the RMSE and MAE values are higher (1128.59, 248.29). Both models have a high RMSE value but low MAE value, which indicate that there are a few cases where the models performs poorly, such as when outliers are present in the data, inflating the RMSE value.

Overall, the BIC model seems to be a candidate for a good model, however, because it selects all 3 variables, this model might not have reasonable clinical application in predicting severe dyskaryosis levels in cervical cancer.

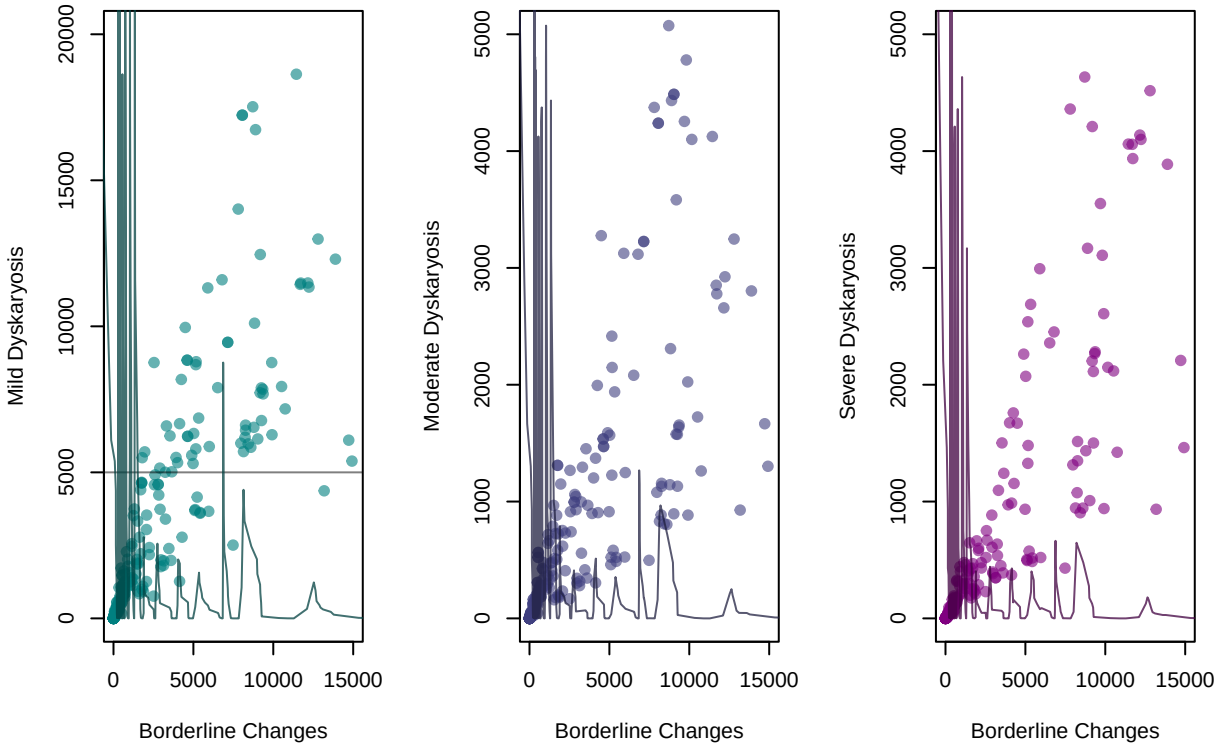
Parametric Linear Model:

```
##                Model:All Model:Polynomial Order 2
##                0.8671448                0.8968722
```

```
##
## Call:
## lm(formula = comparison_data$Borderline_changes ~ x + I(x^2) +
##     y + I(y^2) + z + I(z^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6628.5  -388.7  -305.0   302.7  9810.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.196e+02  1.515e+02   2.110   0.0359 *
## x            1.793e+00  2.098e-01   8.544 1.71e-15 ***
## I(x^2)      -4.809e-05  6.983e-06  -6.887 5.26e-11 ***
## y           -9.649e+00  1.054e+00  -9.152  < 2e-16 ***
## I(y^2)       1.147e-03  1.526e-04   7.518 1.20e-12 ***
## z            5.378e+00  4.967e-01  10.826  < 2e-16 ***
## I(z^2)      -3.314e-04  4.907e-05  -6.755 1.13e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1807 on 233 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8942
## F-statistic: 337.7 on 6 and 233 DF,  p-value: < 2.2e-16
```

The R^2 values for the two models are 0.7831351 and 0.8968722. The parametric model accounted for 89% of the variation of the data analysed. Plotting the model along with the data is a good sanity check to see if the model is over fitting or properly modeling the data.

The coefficients all are significant, yet the only coefficients that have an absolute value greater than 1 are the non-squared terms. This means the model was mostly only using the value of the `Mild_dyskaryosis` (`x` in the summary), `Moderate_dyskaryosis` (`y`), and `Severe_dyskaryosis` (`z`).



It is clear from these graphs that the model is over fitting to the data. The model may be fitting accurately to data points with lower amounts of borderline changes while inaccurately fitting to the data points with higher amounts of borderline changes. The high density of data at the lower end would allow the model to keep a high average R^2 value while being inaccurate at the higher end of the graphs.

As the model is clearly inaccurate, by over fitting or by the data's incompatibility with a parametric model, the model should not be used in any case.

Lasso Model:

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                    s1
## (Intercept)              26.250736649
## Inadequate                0.016693953
## Negative                 -0.004662753
## Borderline_changes        0.065960881
## Mild_dyskaryosis          .
## Moderate_dyskaryosis      0.706554226
## Severe_dyskaryosis_Inv   19.651314391
## Glandular_neoplasia       2.509415319
```

For the first lasso model, the independent variables acting as name tags for the data were removed so the model could be simplified and fitted to the best predictors. The goal of this model is to see if any variables within this data would be a good predictor of a severe dyskaryosis diagnosis. Looking at the R-squared of the lasso model, we see that the model explains 93.9% of the variability within the data which means the

model is fitted very well to the data. However, looking at the RMSE and MAE, we see that the model is very lacking in its predictions. This can be because of the odd way the data is organized and collected.

Some data within this dataset is taken across multiple year frames, while others are taken within the same year. The data also varies widely between the number of observations that are taken at each "location". The largely variability and inconsistency of the data could contribute to a high RMSE and MAE seen within the lasso model.

Because of the data, a portion of the data is singled out to be evaluated again. The data taken from 2012-2023 in England for all ages were subsetted into a new data frame, and the lasso model was ran again under these restrictions.

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                s1
## (Intercept)            10.6538755
## Inadequate              .
## Negative                .
## Borderline_changes      .
## Mild_dyskaryosis        .
## Moderate_dyskaryosis    0.4684427
## Severe_dyskaryosis_Inv  .
## Glandular_neoplasia     3.1208777
```

The R-squared, RMSE, and MAE have all improved from the previous model. However, since the data does have the quality of time attached to it, it may still be an issue in terms of predictions. This may explain the large RMSE and MAE values.

Model	R^2	RMSE	MAE
lasso 1	0.939	421.24	187.10
lasso 2	0.970	46.52	29.34

Chosen Model:

Lasso Model, but linear regression and BIC shows strong correlation between severe dyskaryosis and borderline changes.

Appendix:

Data Cleaning:

```
source <-
  read.csv("cervical-programme-annual-2022-23-csvs/kc61_sample_result_age_source.csv")

dim(source)
```

```
## [1] 330  17
```

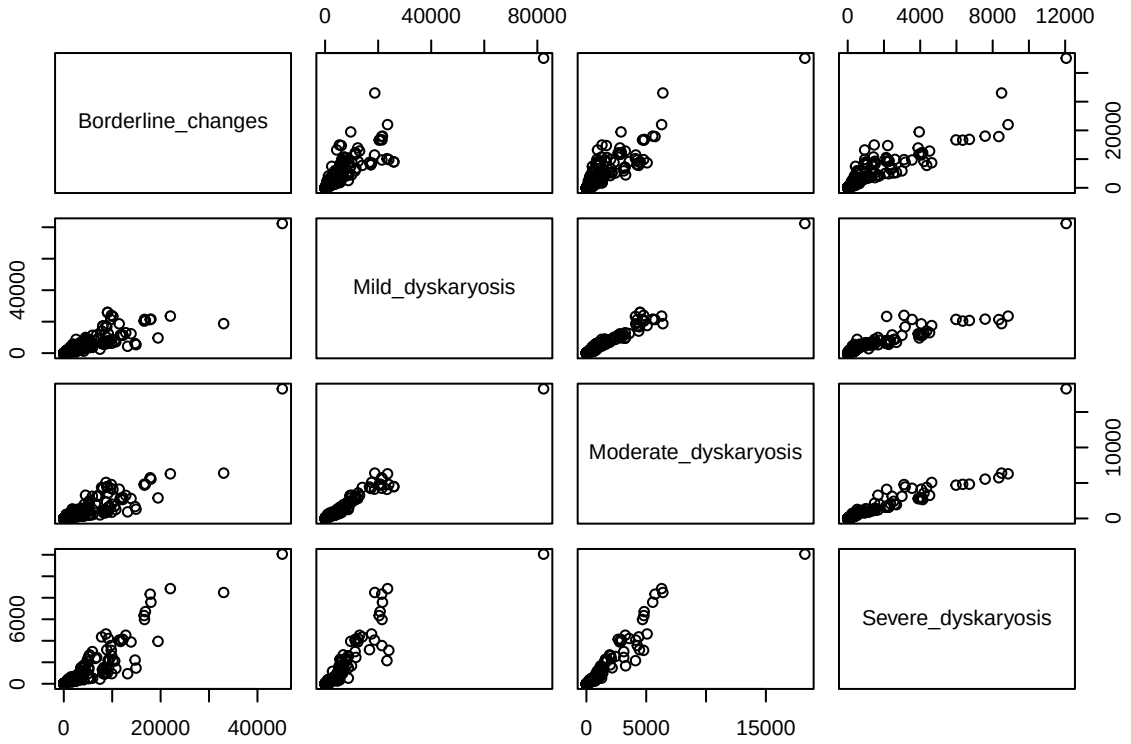
```r
#columns of interest: Indicator (age), Borderline changes, mild, moderate,
#and severe dyskaryosis

#only column that has NA's is severe dyskaryosis; sub in values for NA to clean
source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)
sum(is.na(source$Severe_dyskaryosis) == TRUE)
```

```
## [1] 90
```

```r
uniq_ind <- unique(source$Indicator)
for (i in 1:length(uniq_ind)){
  source$Indicator <- gsub(uniq_ind[i], i, source$Indicator)
}
source$Indicator <- as.numeric(source$Indicator)
#changing each age group to correspond with a number from 1-19. for example, <20 is 1

#removing NAs and converting to numeric values
severe_NA <- source[is.na(source$Severe_dyskaryosis) == FALSE, ]
severe_NA$Severe_dyskaryosis <- as.numeric(severe_NA$Severe_dyskaryosis)

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

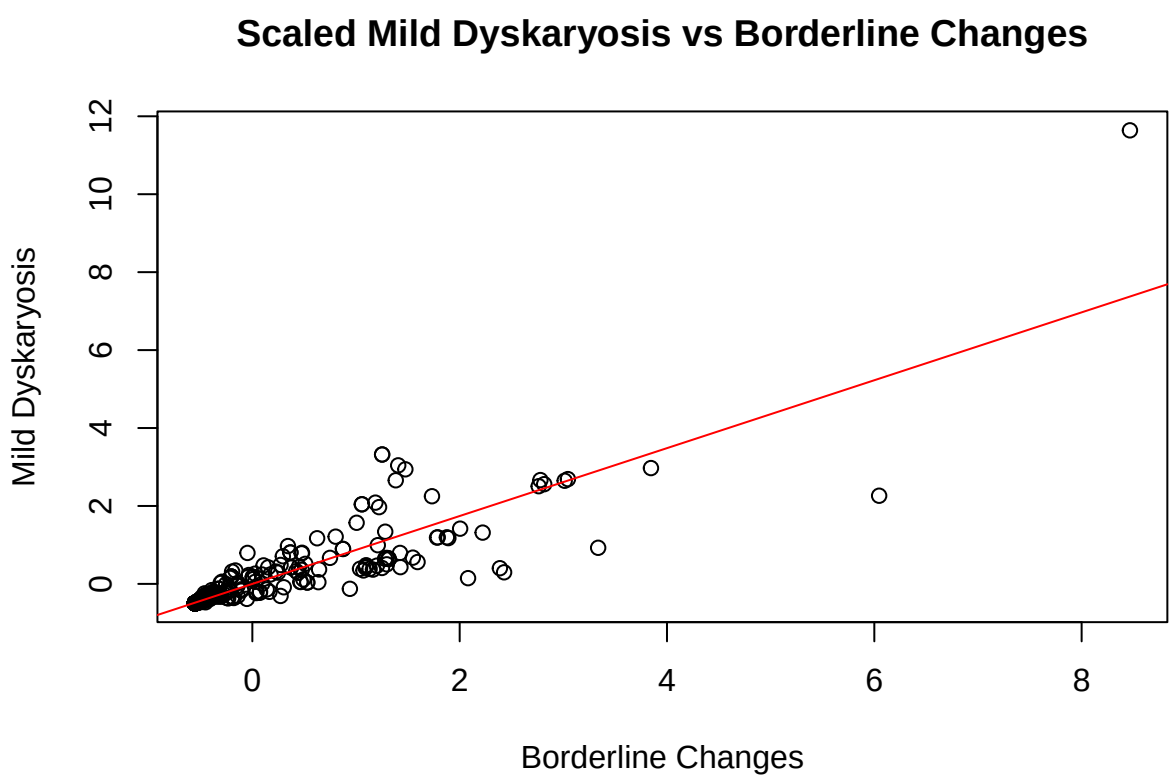
Subsetting Source Data:

```r
## 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)
```



Scaled Linear Regression:

```
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')
```

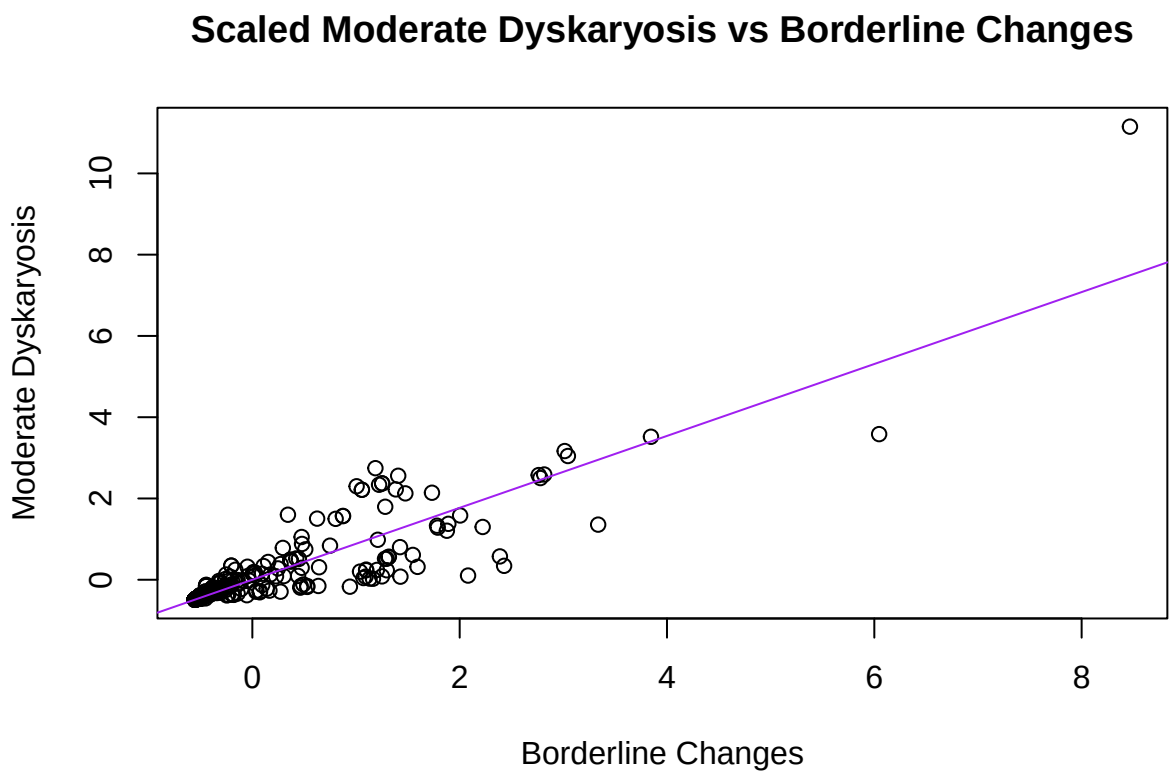



```
summary(scaled_mild_model)
```

```
##
## 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)        2.213e-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",
```

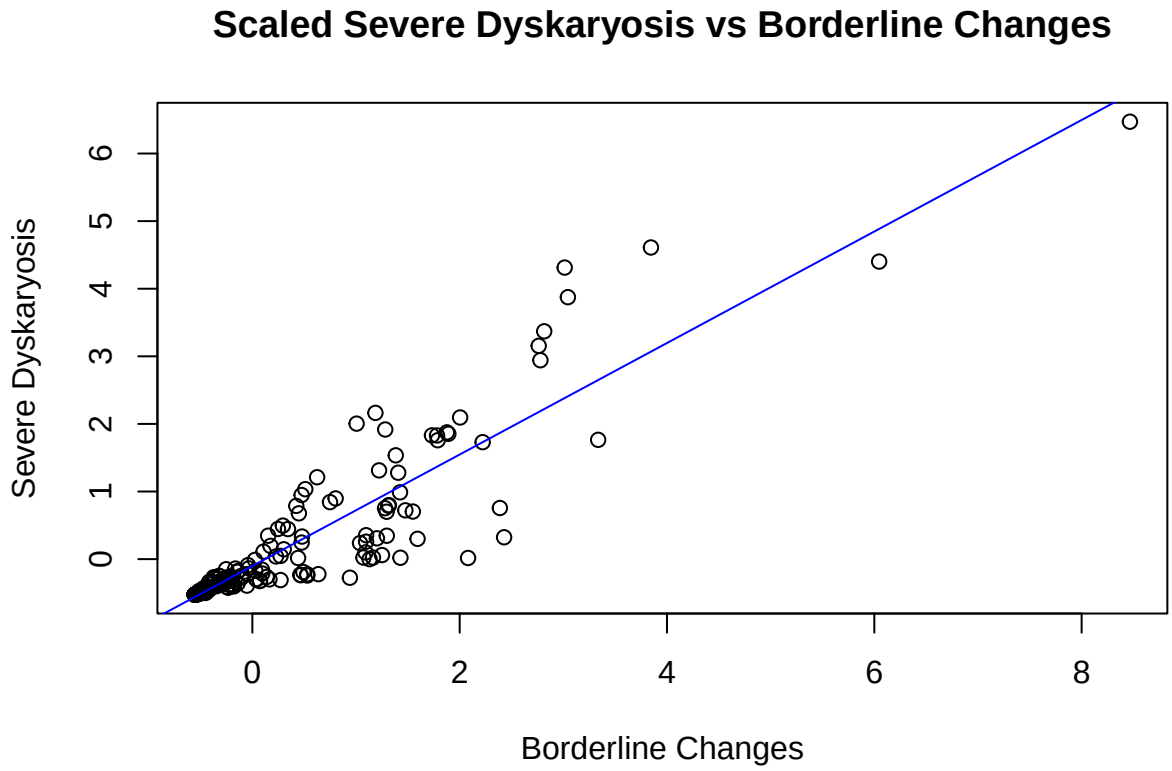
```
     xlab = "Borderline Changes",
     ylab = "Moderate Dyskaryosis")
abline(scaled_mod_model, col = 'purple')
```




```
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)        2.621e-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
```

BIC MIC Model Comparison:

```
#splitting data into test and train
train_index <- 1:165
test_index <- 166:nrow(comparison_data)

train_frame <- comparison_data[train_index,]
test_frame <- comparison_data[test_index,]


train_model <- lm(Severe_dyskaryosis ~ ., data = train_frame)
summary(train_model)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ ., data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes     0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis   1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16
```

```
back_BIC <- step(train_model, direction = "backward",
                 k = log(nrow(train_frame)), trace = 0)
summary(back_BIC)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes + Mild_dyskaryosis +
##     Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -92.17422   41.51175  -2.220   0.0278 *
## Borderline_changes     0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis   1.45536    0.09598  15.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic:  1055 on 3 and 161 DF,  p-value: < 2.2e-16
```

```
back_MIC <- step(train_model, direction = "backward",
                 k = nrow(train_frame), trace = 0)
summary(back_MIC)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2625.70   -49.93    56.48    68.43  1635.61
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -61.79307   48.85798  -1.265    0.208
## Moderate_dyskaryosis   1.17963    0.02726  43.274   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 508.9 on 163 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic:  1873 on 1 and 163 DF,  p-value: < 2.2e-16
```

```
test_MIC <- predict(back_MIC, test_frame)
test_BIC <- predict(back_BIC, test_frame)

cor(test_MIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_MIC model
```

```
## [1] 0.9966765
```

```
cor(test_BIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_BIC model
```

```
## [1] 0.9958021
```

```
errors_MIC <- test_frame$Severe_dyskaryosis - test_MIC
errors_BIC <- test_frame$Severe_dyskaryosis - test_BIC

sqrt(mean(errors_MIC^2, na.rm = TRUE)) #RMSE for back_MIC
```

```
## [1] 1128.586
```

```
sqrt(mean(errors_BIC^2, na.rm = TRUE)) #RMSE for back_BIC
```

```
## [1] 806.2014
```

```
mean(abs(errors_MIC), na.rm = TRUE) #MAE for back_MIC
```

```
## [1] 248.288
```

```
mean(abs(errors_BIC), na.rm = TRUE) #MAE for back_BIC
```

```
## [1] 194.5406
```

Lasso Model:

```
row_select <- sample(1:nrow(severe_NA), nrow(severe_NA) / 2)
col_exclude <- c(1:6, 8, 17)

source_train <- severe_NA[row_select, -col_exclude]
source_test <- severe_NA[-row_select, -col_exclude]

suppressMessages(library(glmnet))

cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis" | cn == "CollectionYearRange")
X <- as.matrix(source_train[ , -exclude])

suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                                      s1
## (Intercept)            184.166697689
## Indicator              -17.570410743
## Inadequate               0.025048886
## Negative                -0.002866436
## Borderline_changes       0.122552225
## Mild_dyskaryosis         .
## Moderate_dyskaryosis     0.638453181
## Severe_dyskaryosis_Inv   .
## Glandular_neoplasia      4.398363311
```

```
Y <- as.matrix(source_test[ , -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##                 [,1]
## lambda.1se 0.9651329
```

```r
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 348.8481
```

```r
mean(abs(source_errors))
```

```
## [1] 208.346
```

```r
lasso_data <- severe_NA[severe_NA$CollectionYearRange == "2022-23" &
                          suppressWarnings(severe_NA$Indicator == c(1:13)), ]
row_select <- sample(1:nrow(lasso_data), nrow(lasso_data) / 2)
col_exclude <- c(1:6, 8, 17)

source_train <- lasso_data[row_select, -col_exclude]
source_test <- lasso_data[-row_select, -col_exclude]

cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis")
X <- as.matrix(source_train[ , -exclude])

suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                                    s1
## (Intercept)            36.86518375
## Indicator               .
## Inadequate              .
## Negative                .
## Borderline_changes      0.03025916
## Mild_dyskaryosis        0.01450963
## Moderate_dyskaryosis    0.26035293
## Severe_dyskaryosis_Inv  .
## Glandular_neoplasia     1.84803219
```

```r
Y <- as.matrix(source_test[ , -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##                 [,1]
## lambda.1se 0.9791545
```

```r
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 194.4096
```

```r
mean(abs(source_errors))
```

```
## [1] 90.16544
```

Sources:

Acknowledgements:

Data Cleaning: Paulina Yao

Scaled Linear Regression: Anika Dachiraju

BIC MIC Model Comparison: Raynah Cheng

Lasso Model: Paulina Yao

Bibliography: