Project Check-In 2

Paulina Yao

2024-12-04

Project Check-In 2

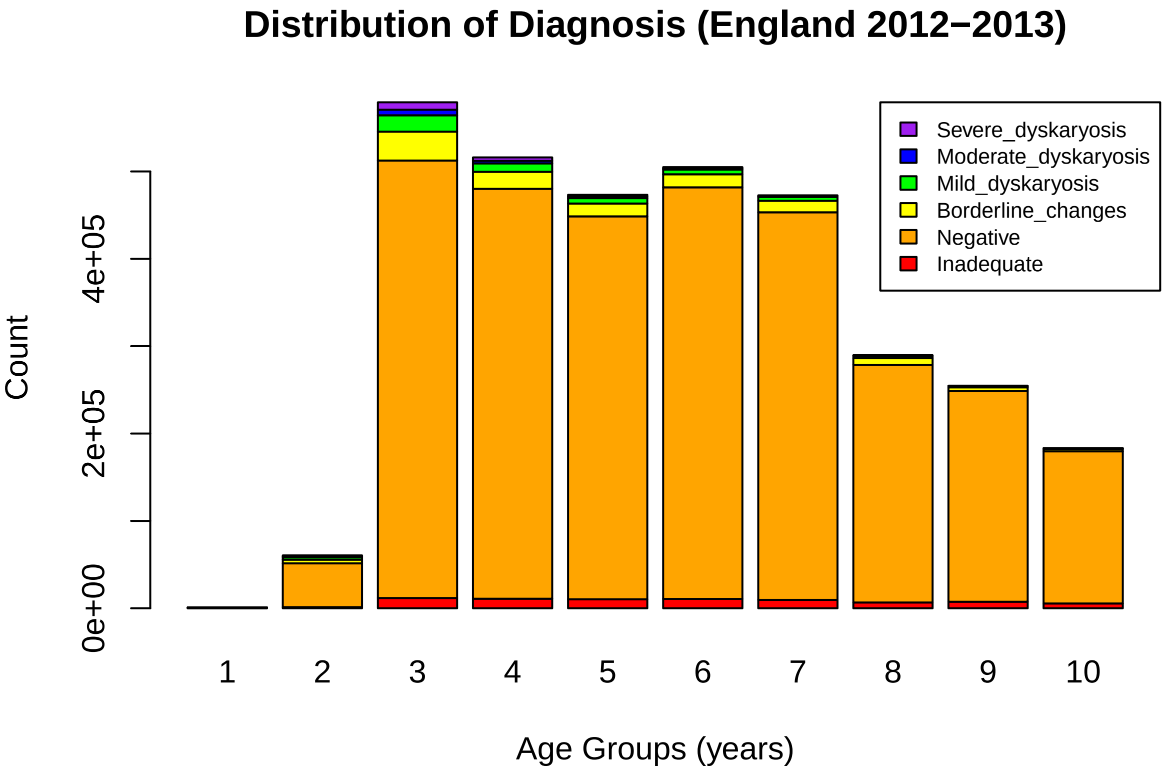
Paulina Yao

After shifting through the data and trying to understand the information better, I found that the data might be better interpreted through a different type of visualization. While I wasn't able to complete a regression analysis for this project check-in, I continued from my work cleaning the data to create barplots.

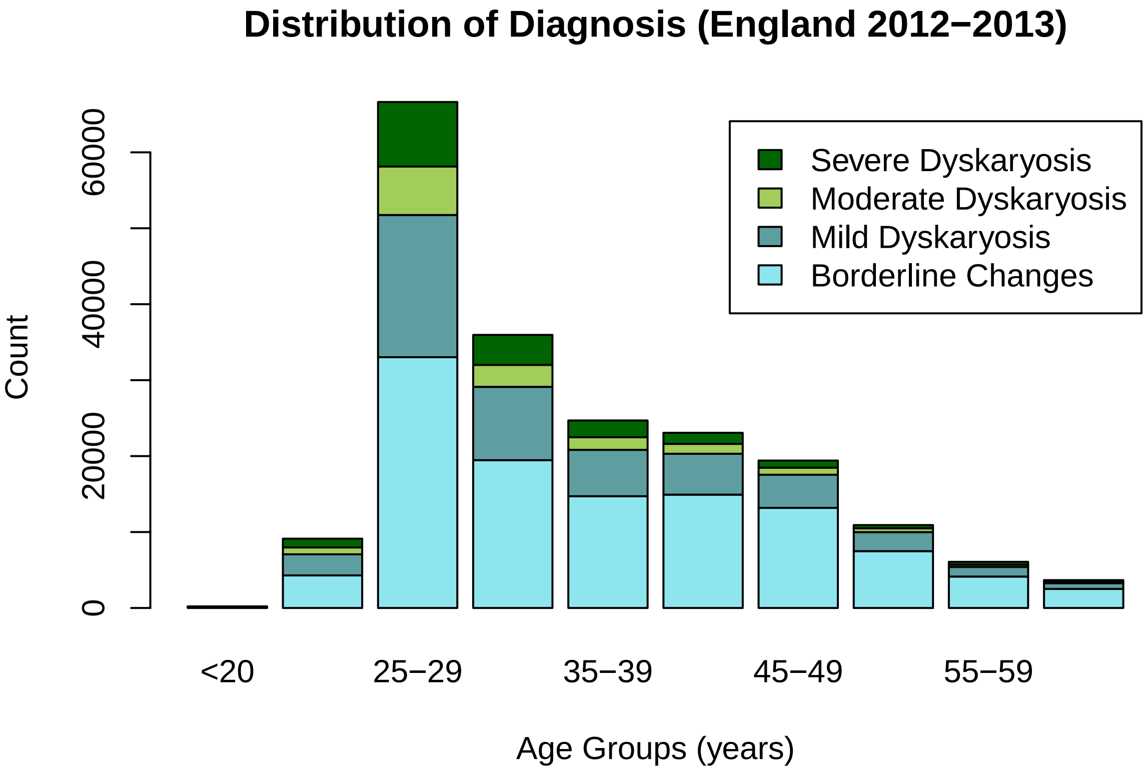
The barplots are able to better visualize the proportion of different diagnoses that the data is taken from.

After looking through the data, I found that for the England (national) data varied in the time frame that the data was taken; however, other data taken at the regional level are only taken from 2022-2023. The plots that were created may have been influenced from this discrepancy. I wanted to single out the data from each year frame.

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

source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)

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)

for (i in 1:length(uniq_ind)){
  df <- data.frame(source[source$Indicator == i, ])
  df_name <- paste0("df", i)
  assign(df_name, df)
}

severe_NA <- source[is.na(source$Severe_dyskaryosis) == FALSE, ]
severe_NA$Severe_dyskaryosis <- as.numeric(severe_NA$Severe_dyskaryosis)
# all code taken above are taken from initial ProjectCheckIn document
# all code was used for cleaning data

plot <- severe_NA[1:10, 9:14]
plot <- t(plot)
plot <- as.matrix(plot)

barplot(plot,
        col = c("red", "orange", "yellow", "green", "blue", "purple"),
        legend.text = row.names(plot),
        args.legend = list(x = "topright", cex = 0.7),
        main = "Distribution of Diagnosis (England 2012-2013)",
        xlab = "Age Groups (years)",
        ylab = "Count")
```



```
# As we can see the data has a lot of data for negative or inadequate data
# However, because of the large numbers of data, the plot is out of proportions
# The negative or inadequate was removed for a more conclusive plot

plot_1213 <- plot[-(1:2), ]
rownames(plot_1213) <- c("Borderline Changes", "Mild Dyskaryosis",
                                "Moderate Dyskaryosis", "Severe Dyskaryosis")
colnames(plot_1213) <- c("<20", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59",

barplot(plot_1213,
        col = c("cadetblue2", "cadetblue", "darkolivegreen3", "darkgreen"),
        legend.text = row.names(plot_1213),
        main = "Distribution of Diagnosis (England 2012-2013)",
        xlab = "Age Groups (years)",
        ylab = "Count")
```



This is a very preliminary evaluation of the data that is given. There will be more worked done to better visualize the raw data, and this might give a better idea of what models that need to be ran in the future.