Untitled

Or gabay 314923681 & Daniel levy 208150433 & Shachar oron 322807231

2023-06-25

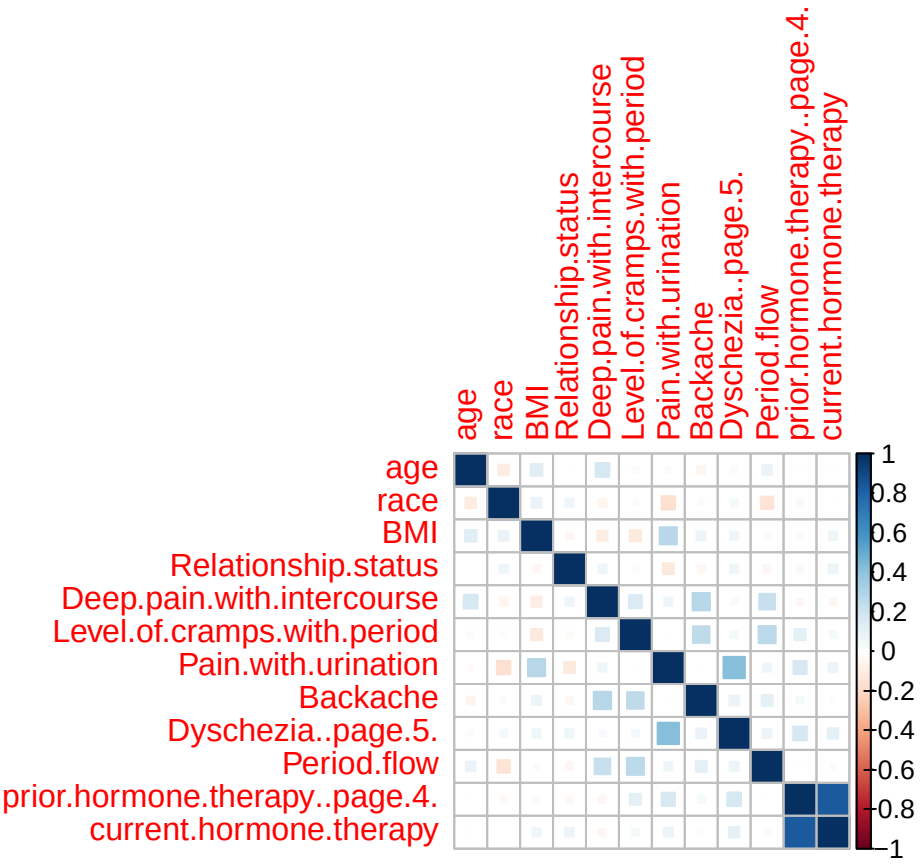
Data Structure

first we get the data from the csv file

```
mydata <- read.csv("../FinalProjectML/imputed_data.csv")
```

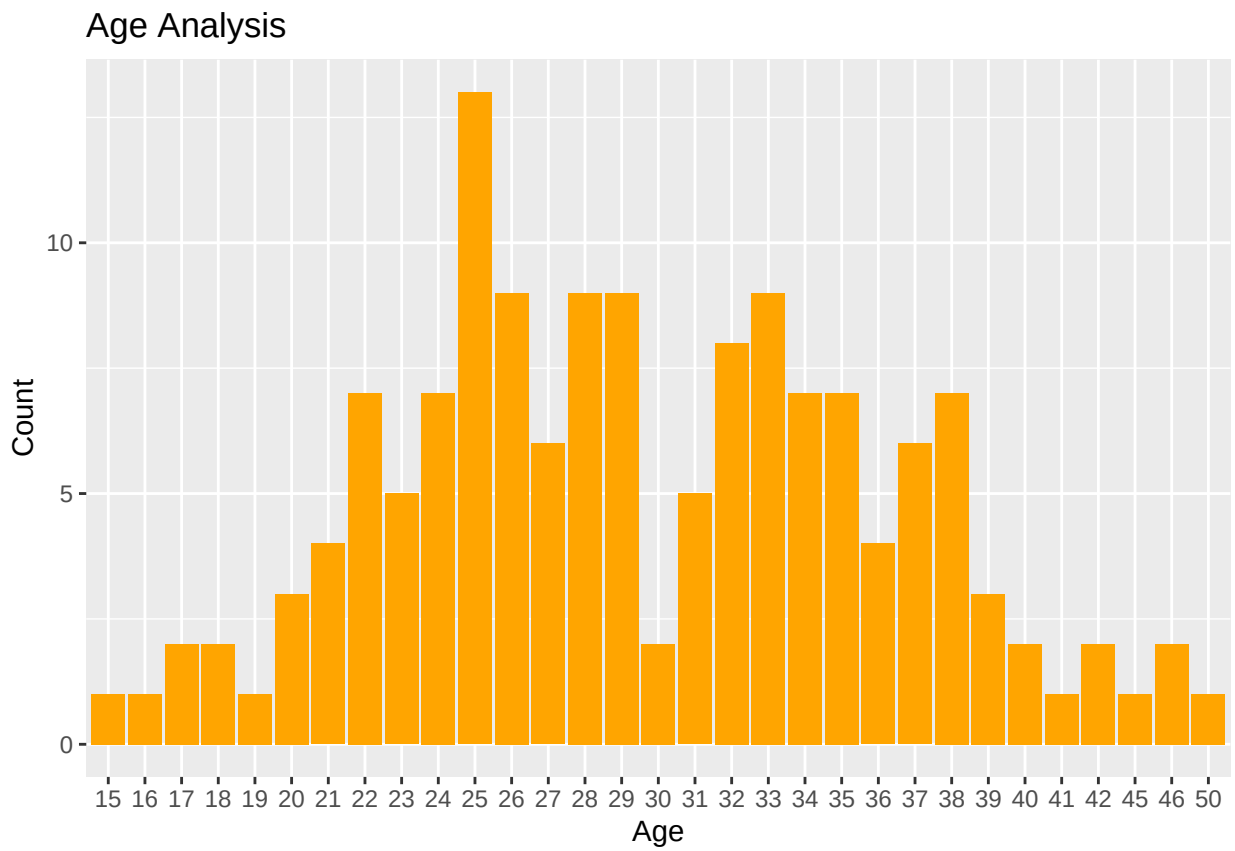
correlation plot

```
# Select the desired columns
selected_columns <- mydata[, c("age", "race", "BMI", "Relationship.status", "Deep.pain.with.intercourse
                               "Level.of.cramps.with.period", "Pain.with.urination", "Backache",
                               "Dyschezia..page.5.", "Period.flow", "prior.hormone.therapy..page.4.", "

# Compute the correlation matrix
cor_matrix <- cor(selected_columns)

# Create the correlation plot
corrplot(cor_matrix, method = "square")
```



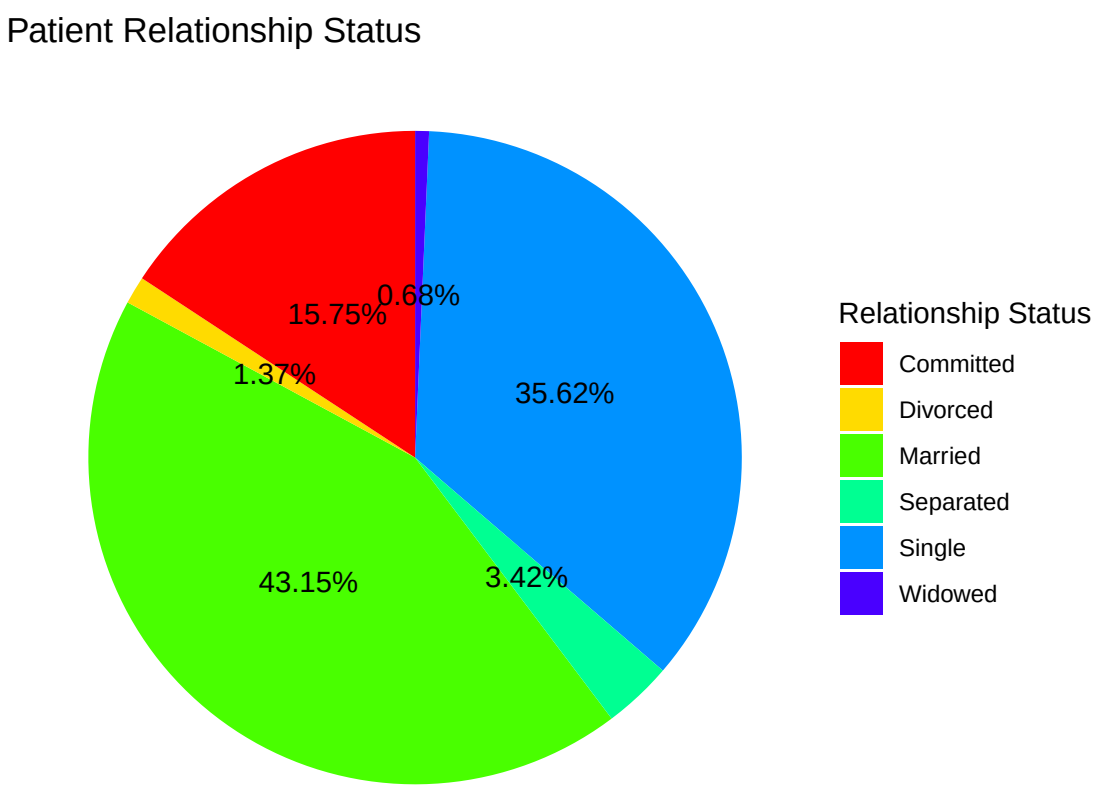
It can be concluded that there is a strong relationship between dyschezia and pain with urination in endometriosis patient. The strong relationship suggests a connection between endometriosis lesions affecting the gastrointestinal tract and pelvic discomfort. Dyschezia, a common symptom in endometriosis, is associated with inflammation, adhesions, or scarring caused by these lesions, leading to pain during bowel movements and potentially pain with urination. Further investigation is needed to establish causation and consider individual factors. In addition, it can be seen that there is a strong connection between the blood flow and the menstrual pain of the patients, which makes a lot of sense. Moreover, it can be seen that the previous hormonal treatment of the patients is correlated with the current treatment of the patients ## Exploring the data

```
#Age Analysis
mydata %>%
  group_by(age) %>%
  count() %>%
  ggplot(aes(x = as.factor(age), y = n)) +
  geom_bar(stat = "identity", fill = "orange") +
  ggtitle("Age Analysis") +
  xlab("Age") +
  ylab("Count")
```



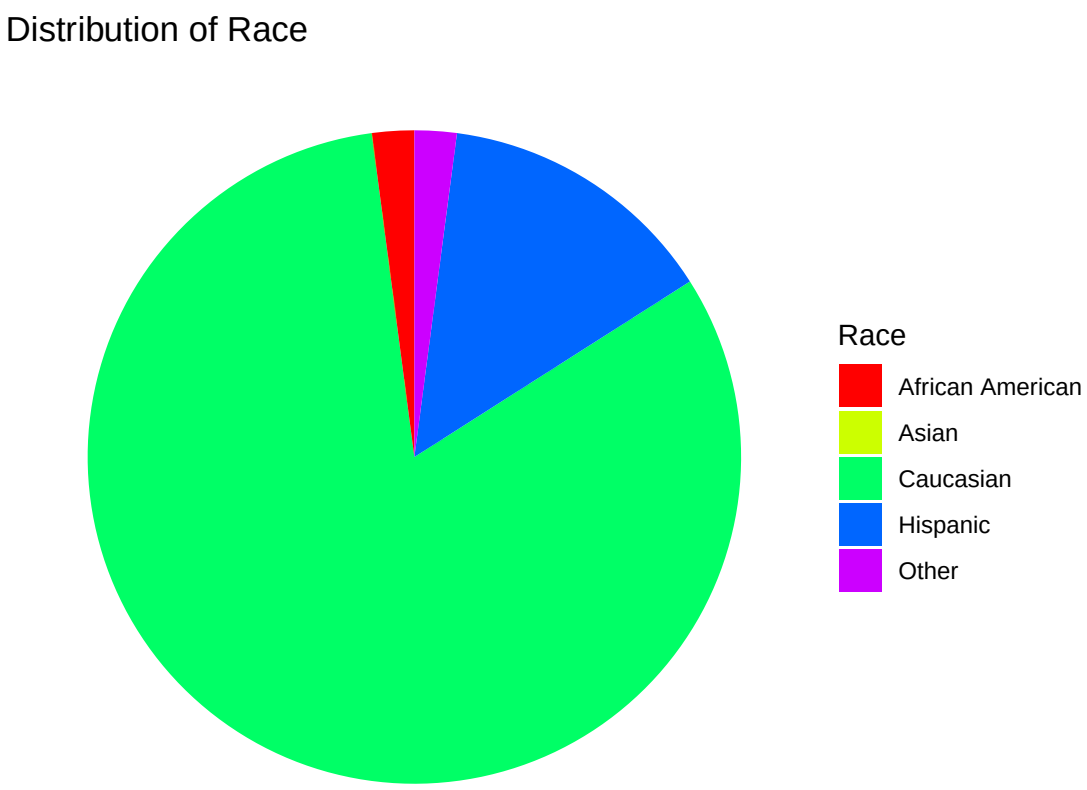

```
#relationship stusus
# Create a data frame for the relationship statuses and their corresponding labels
relationship_labels <- c("Single", "Married", "Widowed", "Remarried", "Separated", "Divorced", "Committe
relationship_data <- data.frame(status = c(0, 1, 2, 3, 4, 5, 6), label = relationship_labels)

# Count the number of patients for each relationship status
relationship_counts <- mydata %>%
  count(Relationship.status) %>%
  left_join(relationship_data, by = c("Relationship.status" = "status")) %>%
  filter(!is.na(label))

# Calculate the percentages
relationship_counts <- relationship_counts %>%
  mutate(percentage = round(n / sum(n) * 100, 2))

# Create the pie chart with percentages
pie_chart <- ggplot(relationship_counts, aes(x = "", y = n, fill = label)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(fill = "Relationship Status",
       title = "Patient Relationship Status",
       x = NULL,
       y = NULL) +
  scale_fill_manual(values = rainbow(length(relationship_labels))) +
  theme_void() +
  geom_text(aes(label = paste0(percentage, "%")), position = position_stack(vjust = 0.5))
```

```
pie_chart
```



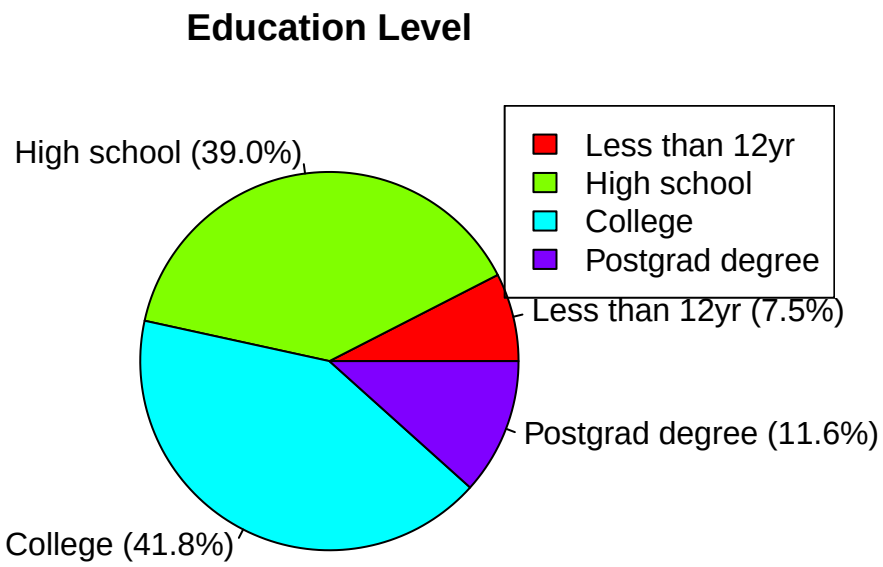

```
#race
# Create a copy of the data with the race column
race_data <- mydata[, "race"]

# Define the labels for the race categories
race_labels <- c("Caucasian", "African American", "Hispanic", "Asian", "Other")

# Convert the race column to a factor with the labels
race_data <- factor(race_data, levels = 0:4, labels = race_labels)

# Calculate the count of each race category
race_counts <- table(race_data)

# Calculate the percentages
race_percentages <- race_counts / sum(race_counts) * 100

# Create a data frame for the pie chart
race_df <- data.frame(Race = race_labels, Count = as.numeric(race_counts), Percentage = race_percentages

# Plot the pie chart
pie_chart <- ggplot(race_df, aes(x = "", y = Count, fill = Race)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
```

```
  labs(fill = "Race", x = NULL, y = NULL, title = "Distribution of Race") +
  scale_fill_manual(values = rainbow(length(race_labels))) +
  theme_void()

# Display the pie chart
pie_chart
```




```
#  Education level
# Create a data frame for education levels and their counts
education_df <- data.frame(
  Education = c("Less than 12yr", "High school", "College", "Postgrad degree"),
  Count = table(mydata$Education.level)
)

# Fix the count values in the education_df data frame
education_df$Count <- education_df$Count.Freq

# Calculate the percentage
education_df$Percentage <- percent(education_df$Count / sum(education_df$Count))

# Generate the pie chart with percentages
pie(education_df$Count, labels = paste0(education_df$Education, " (", education_df$Percentage, ")"), co

# Add a legend
legend("topright", legend = education_df$Education, fill = rainbow(length(education_df$Count)))
```

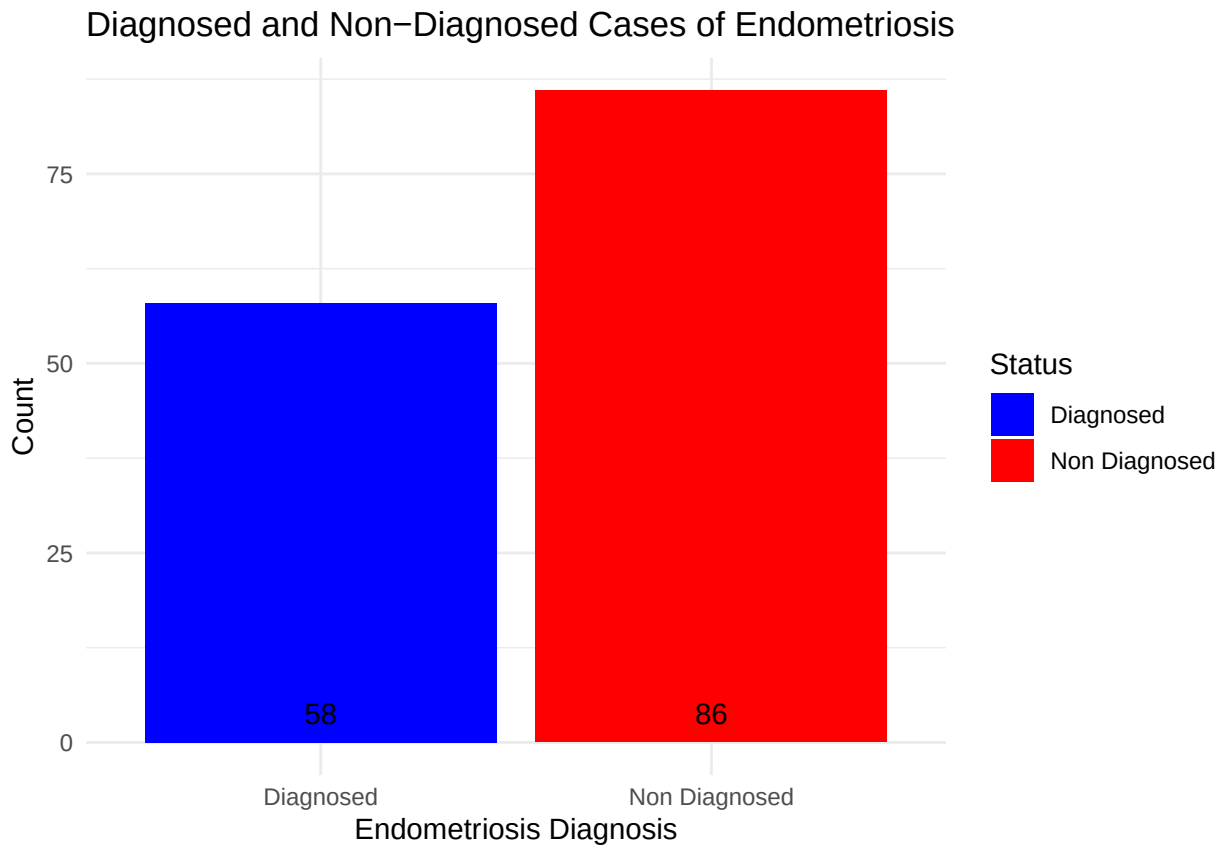
```r
# Add a title
title("Education Level")
```



It can be seen that: 1)the patients are aged 15-50. 2)most of them are married (43%). 3)the origin of most of them is Caucasian. 4) most of them went to college (41.8%) ## Diagnosed Vs non diagnosed patients Our data contains 144 patients, some of them are Diagnosed with Endometriosis but not every one. We would like to check the relationship between the number of those diagnosed and those who are not so that we can know things to continue such as how to normalize the data.

```r
# Filter the data for diagnosed patients
diagnosed_data <- filter(mydata, `Diagnosed.with.endo..ie.at.least.one.biopsy.positve.` == 1)

# Filter the data for non-diagnosed patients
non_diagnosed_data <- filter(mydata, `Diagnosed.with.endo..ie.at.least.one.biopsy.positve.` == 0)

# Create a new column for diagnosed status
diagnosed_data$Status <- "Diagnosed"
non_diagnosed_data$Status <- "Non Diagnosed"

# Combine the diagnosed and non-diagnosed data
combined_data <- rbind(diagnosed_data, non_diagnosed_data)

# Calculate the counts for each diagnosis status
count_data <- combined_data %>%
  group_by(Status) %>%
  summarise(Count = n())
```

```r
# Plot the data
ggplot(combined_data, aes(x = Status, fill = Status)) +
  geom_bar() +
  geom_text(data = count_data, aes(label = Count), vjust = -0.5, color = "black", size = 4, stat = "cou
  xlab("Endometriosis Diagnosis") +
  ylab("Count") +
  ggtitle("Diagnosed and Non-Diagnosed Cases of Endometriosis") +
  scale_fill_manual(values = c("Diagnosed" = "blue", "Non Diagnosed" = "red")) +
  theme_minimal()
```



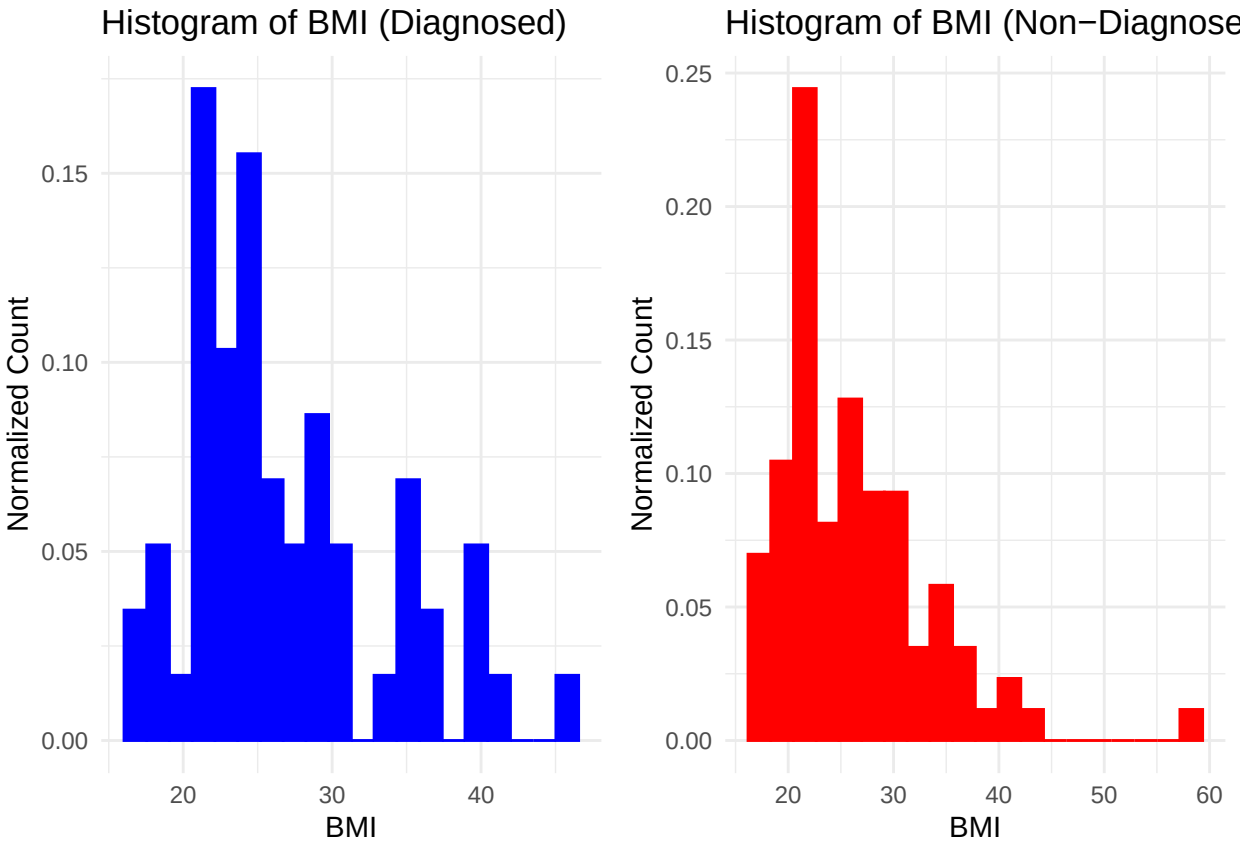
It can be seen that there are 58 diagnosed with endometriosis compared to 86 who are not diagnosed.

Correlation between BMI and Endometriosis

We needed to normalize the counts in the histogram for better comparison between diagnosed and non-diagnosed patients, by using the position = "identity" argument in the geom_histogram() function and divide the counts by the total number of patients in each group.

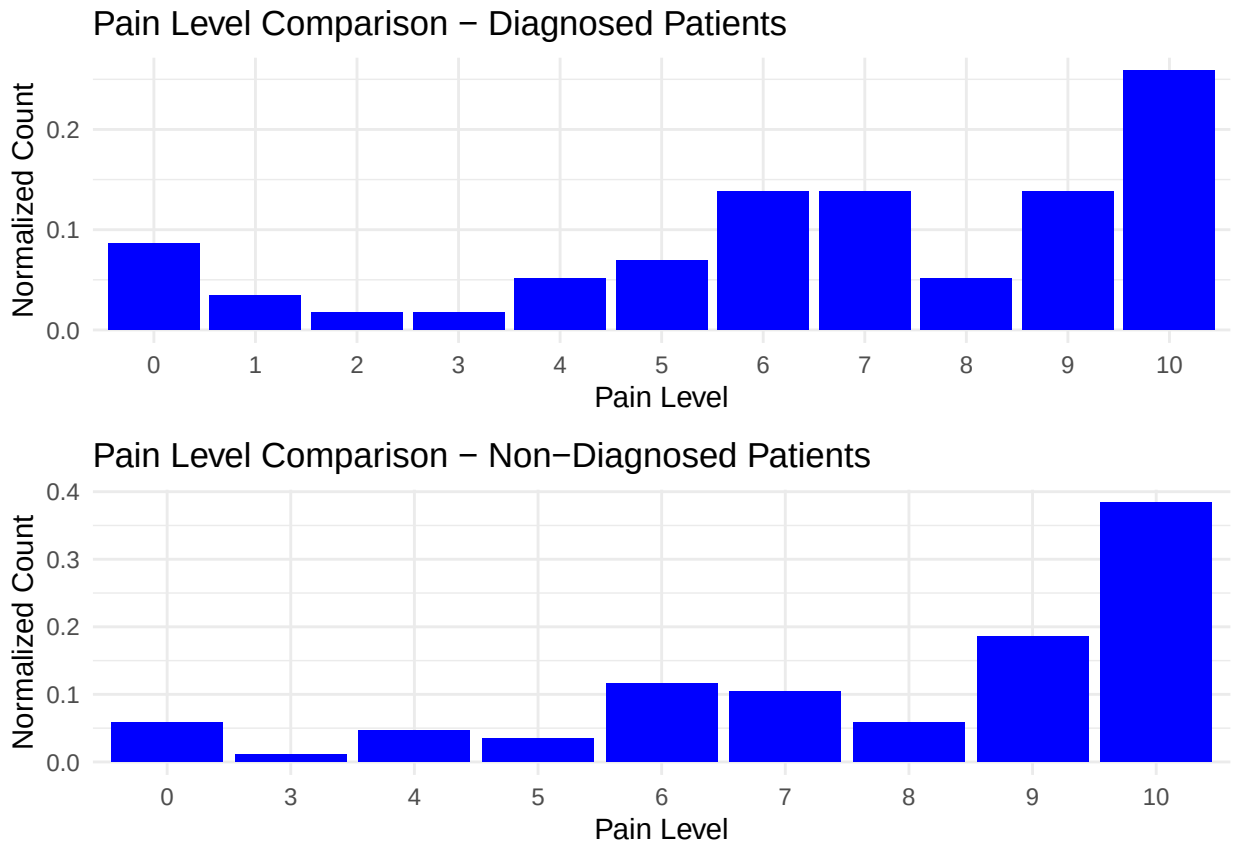
```r
# Calculate the total number of diagnosed and non-diagnosed patients
total_diagnosed <- nrow(diagnosed_data)
total_non_diagnosed <- nrow(non_diagnosed_data)

# Create a histogram of BMI for diagnosed patients with normalized counts
histogram_diagnosed <- ggplot(diagnosed_data, aes(x = BMI)) +
  geom_histogram(aes(y = after_stat(count) / total_diagnosed), color = "blue", fill = "blue", bins = 20
```

```
  labs(x = "BMI", y = "Normalized Count", title = "Histogram of BMI (Diagnosed)") +
  theme_minimal()

# Create a histogram of BMI for non-diagnosed patients with normalized counts
histogram_non_diagnosed <- ggplot(non_diagnosed_data, aes(x = BMI)) +
  geom_histogram(aes(y = after_stat(count) / total_non_diagnosed), color = "red", fill = "red", bins = 1
  labs(x = "BMI", y = "Normalized Count", title = "Histogram of BMI (Non-Diagnosed)") +
  theme_minimal()

# Combine the histograms into a single plot
combined_plot <- cowplot::plot_grid(histogram_diagnosed, histogram_non_diagnosed, ncol = 2)

# Display the combined plot
print(combined_plot)
```



Explanation about the BMI values: Underweight - BMI less than 18.5. Normal weight - BMI ranges from 18.5 to 25. Overweight - BMI ranges from 25 to 30. Obesity - BMI greater than 30. we can infer that both diagnosed and non diagnosed BMI values are mostly normal. The first thought was that most endo patients will have an abnormal BMI due to the effects the disease has on the digestive system.In conclusion, compared to healthy people, the BMI of endo patients is not much different. ## Pain Level of cramps Comparison - bar plots Endo patients are known to suffer from pain before and during menstruation. We wanted to check the evidence given by endometriosis patients about their pain compared to those not diagnosed with endometriosis. We needed to normalize the counts in the histogram for better comparison between diagnosed and non-diagnosed patients.

```r
# Calculate the total number of diagnosed and non-diagnosed patients
total_diagnosed <- nrow(diagnosed_data)
total_non_diagnosed <- nrow(non_diagnosed_data)

# Create a function to normalize counts
normalize_counts <- function(counts, total) {
  counts / total
}

# Normalize counts for diagnosed patients
diagnosed_data <- diagnosed_data %>%
  group_by(`Level.of.cramps.with.period`) %>%
  summarize(count = n()) %>%
  mutate(normalized_count = normalize_counts(count, total_diagnosed))

# Normalize counts for non-diagnosed patients
non_diagnosed_data <- non_diagnosed_data %>%
  group_by(`Level.of.cramps.with.period`) %>%
  summarize(count = n()) %>%
  mutate(normalized_count = normalize_counts(count, total_non_diagnosed))

# Create a bar plot for diagnosed patients
diagnosed_plot <- ggplot(diagnosed_data, aes(x = factor(`Level.of.cramps.with.period`, levels = c("0", "
  geom_bar(fill = "blue", stat = "identity") +
  labs(x = "Pain Level", y = "Normalized Count", title = "Pain Level Comparison - Diagnosed Patients") +
  theme_minimal()

# Create a bar plot for non-diagnosed patients
non_diagnosed_plot <- ggplot(non_diagnosed_data, aes(x = factor(`Level.of.cramps.with.period`, levels =
  geom_bar(fill = "blue", stat = "identity") +
  labs(x = "Pain Level", y = "Normalized Count", title = "Pain Level Comparison - Non-Diagnosed Patients
  theme_minimal()

# Combine the plots into a single plot
combined_plot <- plot_grid(diagnosed_plot, non_diagnosed_plot, nrow = 2)

# Display the combined plot
print(combined_plot)
```



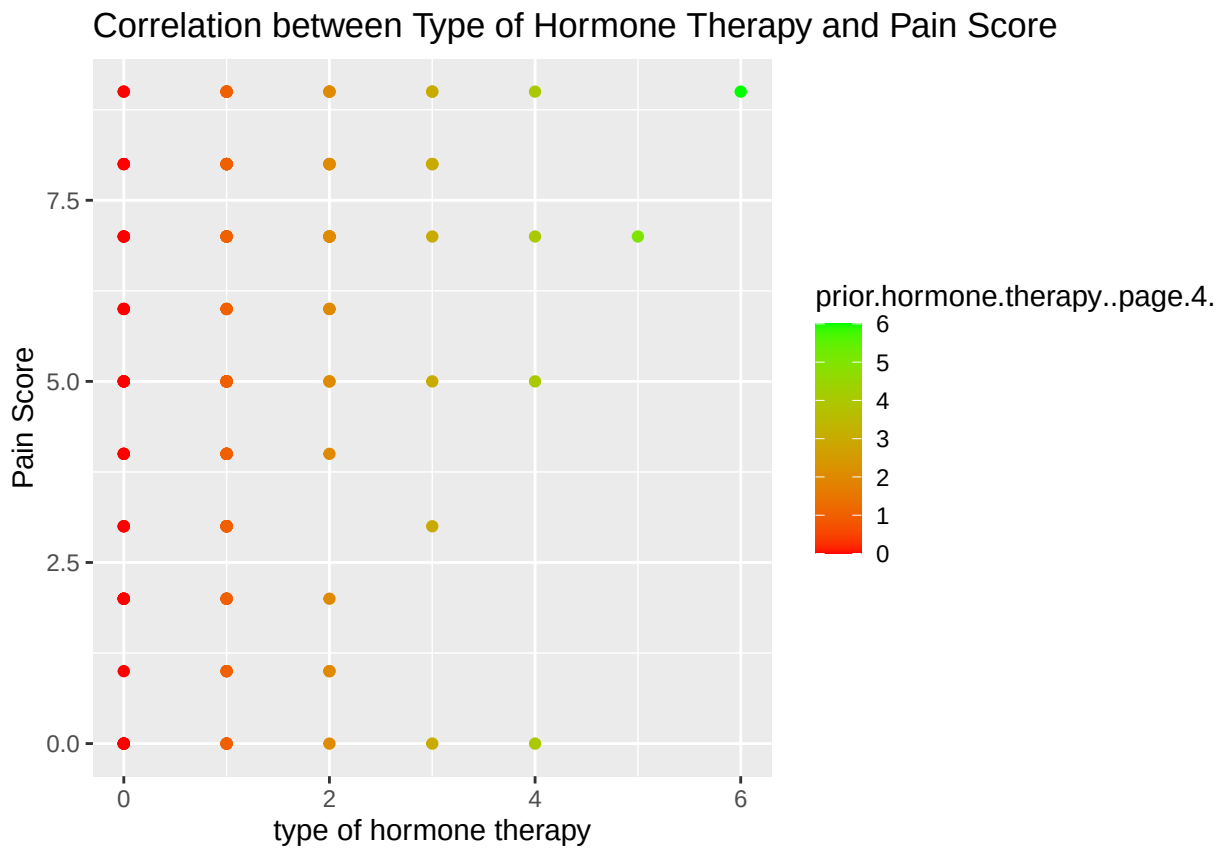
Since the data is normalized, it can be seen that 0.4 of the patients diagnosed with endo testified to severe menstrual pain compared to the evidence of women not diagnosed with endo (0.25). In addition, most of the diagnosed patients testify to severe menstrual pain at strong levels compared to those who are not diagnosed, for whom the range of pain is wider.

Different treatments

first of all we would like to check the effectiveness of the hormonal treatment.In our data there is a column 'Initial.visual.analog.score'. A visual analog scale (VAS) is a measurement tool commonly used in medical and research settings to assess subjective experiences or perceptions, such as pain intensity.

```
# scatter plot:
# create a frequency table of Company.Location
freq_table <- table(mydata$Initial.visual.analog.score)

# create a dataframe from the frequency table
freq_df <- data.frame(location = names(freq_table), frequency = as.numeric(freq_table))

# order the dataframe by frequency and select top 40 countries
top_locations <- freq_df[order(freq_df$frequency, decreasing = TRUE), ][1:40, "location"]

# filter the data to only include the top 40 countries
mydata_filtered <- mydata[mydata$Initial.visual.analog.score %in% top_locations, ]

# create the scatter plot
ggplot(data = mydata_filtered, aes(x = prior.hormone.therapy..page.4., y = Initial.visual.analog.score,
```

```
geom_point() +
scale_color_gradient(low = "red", high = "green") +
xlab("type of hormone therapy") +
ylab("Pain Score") +
ggtitle("Correlation between Type of Hormone Therapy and Pain Score")
```



The scatter plot will display the "type of hormone therapy" on the x-axis and the "Pain Score" on the y-axis. Each data point represents an individual observation, and the color of the data points represents the different categories of "type of hormone therapy". The color scale used ranges from red to green. the different type of hormones are: 0=none 1=COC/estrogen method- The combined oral contraceptive pill (COC) is a tablet that contains two hormones, progestogen and estrogen, and is taken daily to prevent pregnancy. 2=dep provera - An injectable contraceptive. This is a hormonal treatment based on progestin, which is injected into the muscle and is released slowly, thus delaying ovulation in the uterus. 3=IUD - Intrauterine device 4=Lupron - is used to prevent premature ovulation in cycles of controlled ovarian stimulation for in vitro fertilization (IVF).

We can infer important information from this plot. for example: 1) The ones that are using the 4th type of hormone therapy, Lupron, have a strong pain score. From the research we have done, we discovered that this drug has serious side effects so in order to use the drug, the medical condition is probably really serious. 2) There is no difference in the distribution of the different pain scores between the different hormonal treatments. In all of them the pain levels are varied in the same way. 3)The majority of individuals in the dataset have either "none" (0) or "COC/estrogen method" (1) as their type of hormone therapy.