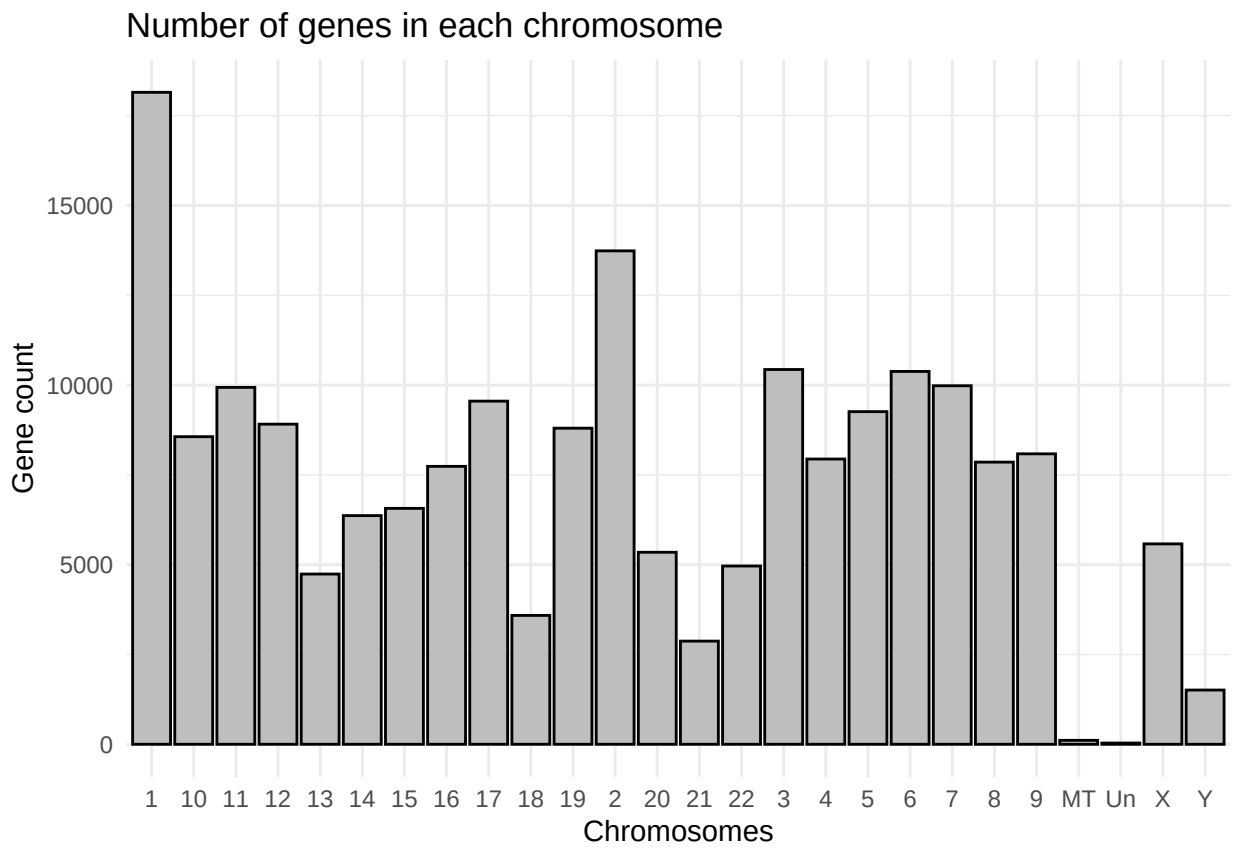
Problem 3

Maryam Gholampour

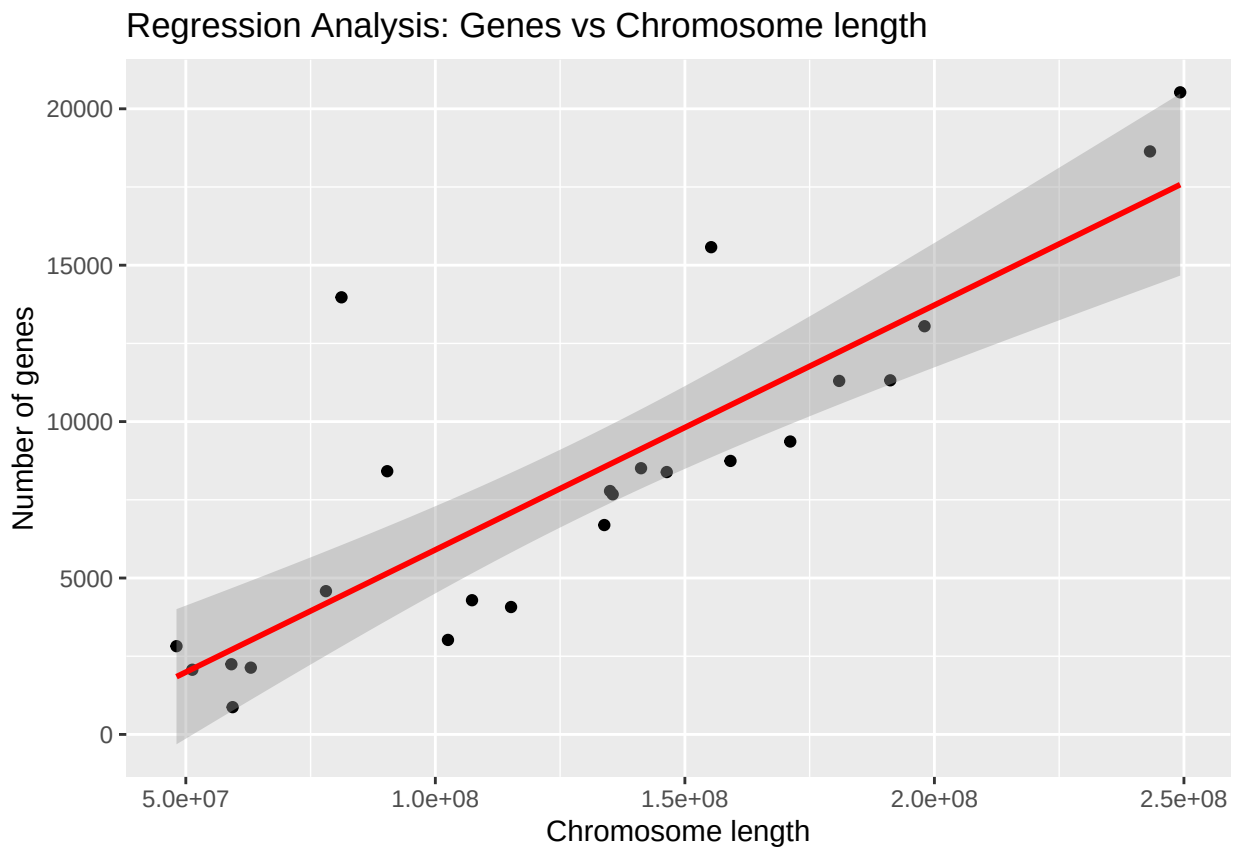
2023-12-11

```
#Plot

file_url <- "https://github.com/cb2edu/CB2-101-2023-assignment/raw/main/data/Homo_sapiens.gene_info.gz"

download.file(file_url, destfile = "Homo_sapiens.gene_info.gz")

data <- read.table("Homo_sapiens.gene_info.gz", header = TRUE, sep = "\t", quote
= "", comment.char = "", stringsAsFactors = FALSE)

subset_data <- data[, c(3, 7)]

gene_counts <- table(subset_data$chromosome)
gene_counts_df <- as.data.frame(gene_counts)
names(gene_counts_df) <- c("Chromosome", "GeneCount")

gene_counts_df_filtered <- gene_counts_df[!grepl("\\|", gene_counts_df$Chromosome), ]

gene_counts_df_filtered <- gene_counts_df_filtered[gene_counts_df_filtered$Chromosome != "-", ]

gene_counts_df_filtered$Chromosome <- factor(gene_counts_df_filtered$Chromosome,
levels = unique(gene_counts_df_filtered$Chromosome))

library(ggplot2)

gg_plot <- ggplot(gene_counts_df_filtered, aes(x = Chromosome, y = GeneCount)) +
  geom_bar(stat = "identity", fill = "gray", color = "black") +
  labs(title = "Number of genes in each chromosome",
       x = "Chromosomes",
       y = "Gene count") +
  theme_minimal()

ggsave("plot_output.pdf", plot = gg_plot, width = 8, height = 4)

print(gg_plot)
```




```
#Is there any correlation between number of genes and chromosome length?

library(ggplot2)

chromosome_lengths <- data.frame(
  chromosome = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14", "15", "16",
  length = c(249250621, 243199373, 198022430, 191154276, 180915260, 171115067, 159138663, 146364022, 14
  genes = c(20526, 18637, 13049, 11320, 11301, 9363, 8743, 8387, 8510, 7676, 7780, 6693, 4073, 4290, 30
)

linear_model <- lm(genes ~ length, data = chromosome_lengths)

summary_linear_model <- summary(linear_model)

ggplot(chromosome_lengths, aes(x = length, y = genes)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Regression Analysis: Genes vs Chromosome length",
       x = "Chromosome length",
       y = "Number of genes")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
print(summary_linear_model)
```

```
##
## Call:
## lm(formula = genes ~ length, data = chromosome_lengths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3078.6 -1801.1  -867.6   539.0  9542.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.920e+03  1.486e+03  -1.292     0.21
## length       7.822e-05  1.055e-05   7.415 2.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2930 on 22 degrees of freedom
## Multiple R-squared:  0.7142, Adjusted R-squared:  0.7012
## F-statistic: 54.98 on 1 and 22 DF,  p-value: 2.034e-07
```

```
chromosome_lengths$expected_genes <- predict(linear_model, newdata = chromosome_lengths)

chromosome_lengths$deviation <- chromosome_lengths$genes - chromosome_lengths$expected_genes
```

```r
print(chromosome_lengths[, c("chromosome", "genes", "expected_genes", "deviation")])
```

```
##    chromosome genes expected_genes   deviation
## 1           1 20526      17575.207  2950.79296
## 2           2 18637      17101.909  1535.09150
## 3           3 13049      13568.393  -519.39262
## 4           4 11320      13031.200 -1711.19976
## 5           5 11301      12230.355  -929.35484
## 6           6  9363      11463.832 -2100.83246
## 7           7  8743      10527.098 -1784.09768
## 8           8  8387       9527.929 -1140.92878
## 9           9  8510       9125.075  -615.07515
## 10         10  7676       8680.917 -1004.91672
## 11         11  7780       8639.601  -859.60112
## 12         12  6693       8549.292 -1856.29241
## 13         13  4073       7088.078 -3015.07792
## 14         14  4290       6476.410 -2186.40996
## 15         15  3021       6099.558 -3078.55837
## 16         16  8414       5147.162  3266.83779
## 17         17 13973       4430.748  9542.25171
## 18         18  4580       4186.877   393.12319
## 19         19  2243       2704.838  -461.83772
## 20         20  2134       3009.605  -875.60547
## 21         21  2821       1844.544   976.45626
## 22         22  2068       2092.851   -24.85072
## 23          X 15576      10224.554  5351.44610
## 24          Y   872       2723.968 -1851.96779
```

```r
confidence_interval <- confint(linear_model)
print(confidence_interval)
```

```
##                     2.5 %       97.5 %
## (Intercept) -5.002142e+03 1.162268e+03
## length       5.633931e-05 1.000907e-04
```

```r
#In earlier problem we calculated the frequency of disease variant in each gene in human genome. Can yo

download.file("https://github.com/cb2edu/CB2-101-2023-assignment/raw/main/data/humsavar.tsv.gz", destfil

data2 <- read.table(gzfile("humsavar.tsv.gz"), header = TRUE, sep = "\t", quote = "", comment.char = ""

colnames(data2) <- c("Symbol", "swiss_prot_ac", "ftid", "aa_changes", "variant", "dbsnp")

subset_data2 <- data2[, c("Symbol", "variant")]

merged_data <- merge(subset_data, subset_data2, by = "Symbol")

disease_variants <- merged_data[merged_data$variant == "LP/P", ]

chromosome_concentration <- table(disease_variants$chromosome)

chromosome_with_highest_concentration <- names(which.max(chromosome_concentration))
```

```r
cat("Chromosome with the highest concentration of disease variants:", chromosome_with_highest_concentra
```

```
## Chromosome with the highest concentration of disease variants: X
```