Comprehensive Analysis of Breast Cancer Statistics and Trends

A Project Report Submitted
in Partial Fulfillment of the Requirements
for the course of

DSCC 462 (Introduction to Computational Statistics)

by

Abhishek Sharma
Aashrith Maisa
Neha Rana
Ashika Kotia

under the guidance of

Prof. Anson Khang

to the

GEORGEN INSTITUTE OF DATA SCIENCE
UNIVERSITY OF ROCHESTER

Introduction

Our analysis leverages a detailed dataset encompassing a spectrum of breast cancer patients who underwent surgical procedures to excise their tumors. With 341 entries, the dataset meticulously documents a unique identifier for each patient (Patient_ID) and captures crucial demographic information, including age at the time of diagnosis and gender—with a noteworthy representation of 97% female patients, highlighting the gender prevalence in breast cancer incidence. It delves into the biological underpinnings by quantifying the expression levels of proteins (Protein1 and Protein2) that, despite being measured in undefined units, offer valuable insights into the molecular profile of the disease. These biomarkers, alongside other key features not limited to tumor stage and histological type, lay the groundwork for a nuanced exploration of the pathophysiological traits characterizing this condition. The dataset stands as a testament to the interdisciplinary approach required in oncological research, merging clinical, demographic, and molecular data to foster a comprehensive understanding of breast cancer dynamics.

```
brca <- read.csv("BRCA.csv")
library(psych)
summary(brca)
```

```
##   Patient_ID             Age           Gender             Protein1
##  Length:341         Min.   :29.00   Length:341         Min.   :-2.340900
##  Class :character   1st Qu.:49.00   Class :character   1st Qu.:-0.358888
##  Mode  :character   Median :58.00   Mode  :character   Median : 0.006129
##                     Mean   :58.89                      Mean   :-0.029991
##                     3rd Qu.:68.00                      3rd Qu.: 0.343598
##                     Max.   :90.00                      Max.   : 1.593600
##                     NA's   :7                          NA's   :7
##     Protein2          Protein3          Protein4         Tumour_Stage
##  Min.   :-0.9787   Min.   :-1.6274   Min.   :-2.025500   Length:341
##  1st Qu.: 0.3622   1st Qu.:-0.5137   1st Qu.:-0.377090   Class :character
##  Median : 0.9928   Median :-0.1732   Median : 0.041768   Mode  :character
##  Mean   : 0.9469   Mean   :-0.0902   Mean   : 0.009819
##  3rd Qu.: 1.6279   3rd Qu.: 0.2784   3rd Qu.: 0.425630
##  Max.   : 3.4022   Max.   : 2.1934   Max.   : 1.629900
##  NA's   :7         NA's   :7         NA's   :7
##   Histology          ER.status          PR.status          HER2.status
##  Length:341         Length:341         Length:341         Length:341
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Surgery_type       Date_of_Surgery    Date_of_Last_Visit Patient_Status
##  Length:341         Length:341         Length:341         Length:341
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

1.1 Age Distribution and Cancer Stage:

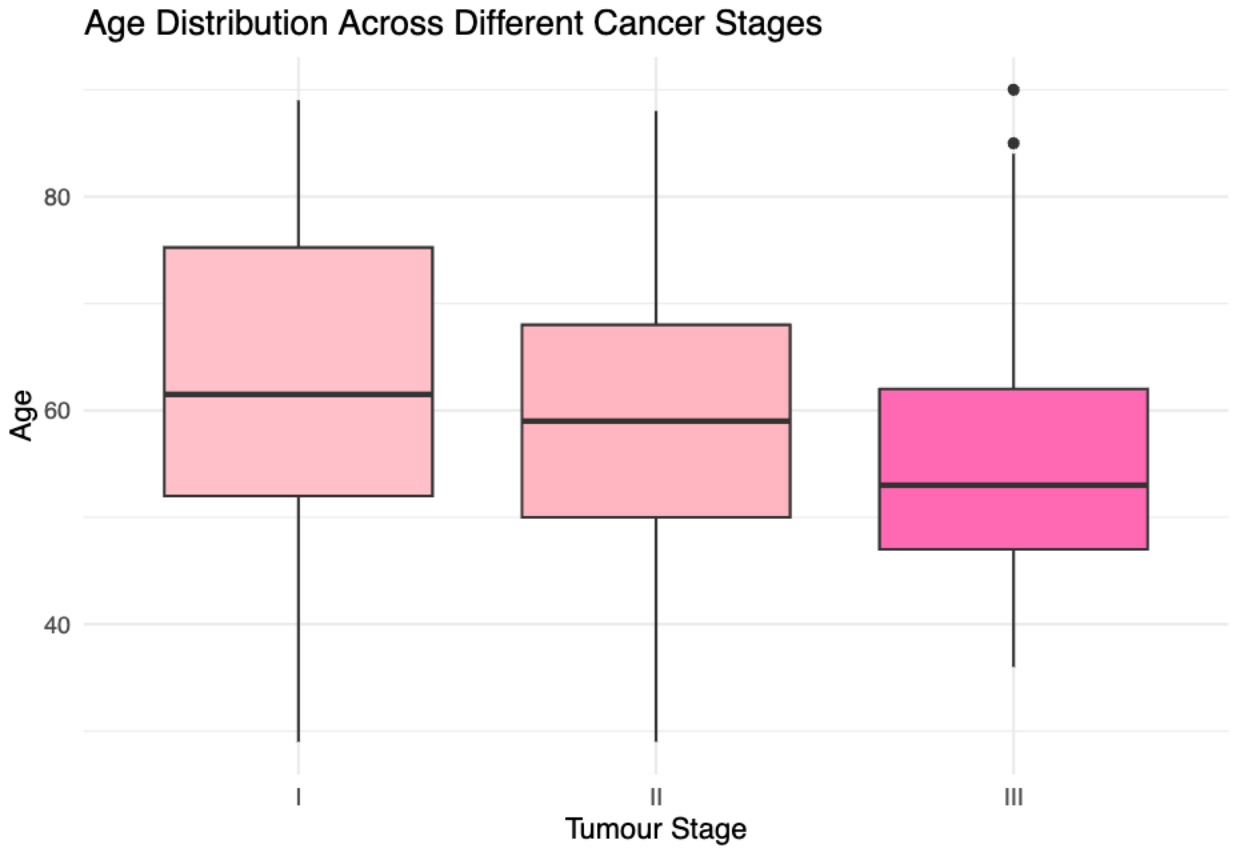
We will employ side-by-side box plots to represent the age distribution based on different cancer stages, providing insights into age variations across stages.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
cancer_data_cleaned <- na.omit(brca, cols = c("Age", "Tumour_Stage"))

pink_colors <- c("#FFC0CB", "#FFB6C1", "#FF69B4", "#FF1493", "#DB7093")

# Create the box plot
ggplot(cancer_data_cleaned, aes(x = Tumour_Stage, y = Age, fill = Tumour_Stage)) +
    geom_boxplot() +
    scale_fill_manual(values = pink_colors) +
    labs(title = "Age Distribution Across Different Cancer Stages",
         x = "Tumour Stage",
         y = "Age") +
    theme_minimal() +
    theme(legend.position = "none")
```



```r
# Display the plot
ggsave("Breast_Cancer_Age_Distribution.png", width = 10, height = 6)
```

The median age appears to decrease from cancer stage I to III. This suggests that in this dataset, younger patients tend to be more represented in later stages of cancer. The IQR for stage I is narrower compared to stage III, indicating a more consistent age range among early-stage cancer patients, while stage III shows greater age variability.

1.2 Histology Frequency:

2. Aiming to understand histology patterns, we will construct absolute and relative frequency tables to showcase the number of individuals with specific histological characteristics.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Absolute Frequency Table
absolute_freq_table <- cancer_data_cleaned %>%
  group_by(Histology) %>%
  summarise(Count = n())

# Relative Frequency Table
relative_freq_table <- absolute_freq_table %>%
  mutate(Percentage = (Count / sum(Count)) * 100)

# Print the tables
print("Absolute Frequency Table:")
```

```
## [1] "Absolute Frequency Table:"
```

```r
print(absolute_freq_table)
```

```
## # A tibble: 3 x 2
##   Histology                      Count
##   <chr>                          <int>
## 1 Infiltrating Ductal Carcinoma    233
## 2 Infiltrating Lobular Carcinoma    89
## 3 Mucinous Carcinoma                12
```

```r
print("Relative Frequency Table:")
```

```
## [1] "Relative Frequency Table:"
```

```r
print(relative_freq_table)
```

```
## # A tibble: 3 x 3
##   Histology                      Count Percentage
##   <chr>                          <int>      <dbl>
## 1 Infiltrating Ductal Carcinoma    233       69.8
```

```
## 2 Infiltrating Lobular Carcinoma    89      26.6
## 3 Mucinous Carcinoma                12       3.59
```

1.3 Protein Correlations

3. Utilizing scatter plots, we will explore correlations between different proteins, offering insights into potential relationships that may contribute to the understanding of breast cancer.

```
library(ggplot2)
```

```
brca <- read.csv("BRCA.csv")
```

```
brca_clean <- na.omit(brca)
correlation_coefficient <- cor(brca_clean$Protein1, brca_clean$Protein2, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] 0.2381413
```

```
ggplot(brca_clean, aes(x=Protein1, y=Protein2)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein1 and Protein2", x="Protein1", y="Protein2")
```

The correlation coefficient of approximately 0.223 suggests a weak positive linear relationship between the two columns, `Protein1` and `Protein2`. In other words, as the values of `Protein1` increase, the values of `Protein2` tend to also increase, but not very strongly.

```
correlation_coefficient <- cor(brca_clean$Protein1, brca_clean$Protein3, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] -0.1294824
```

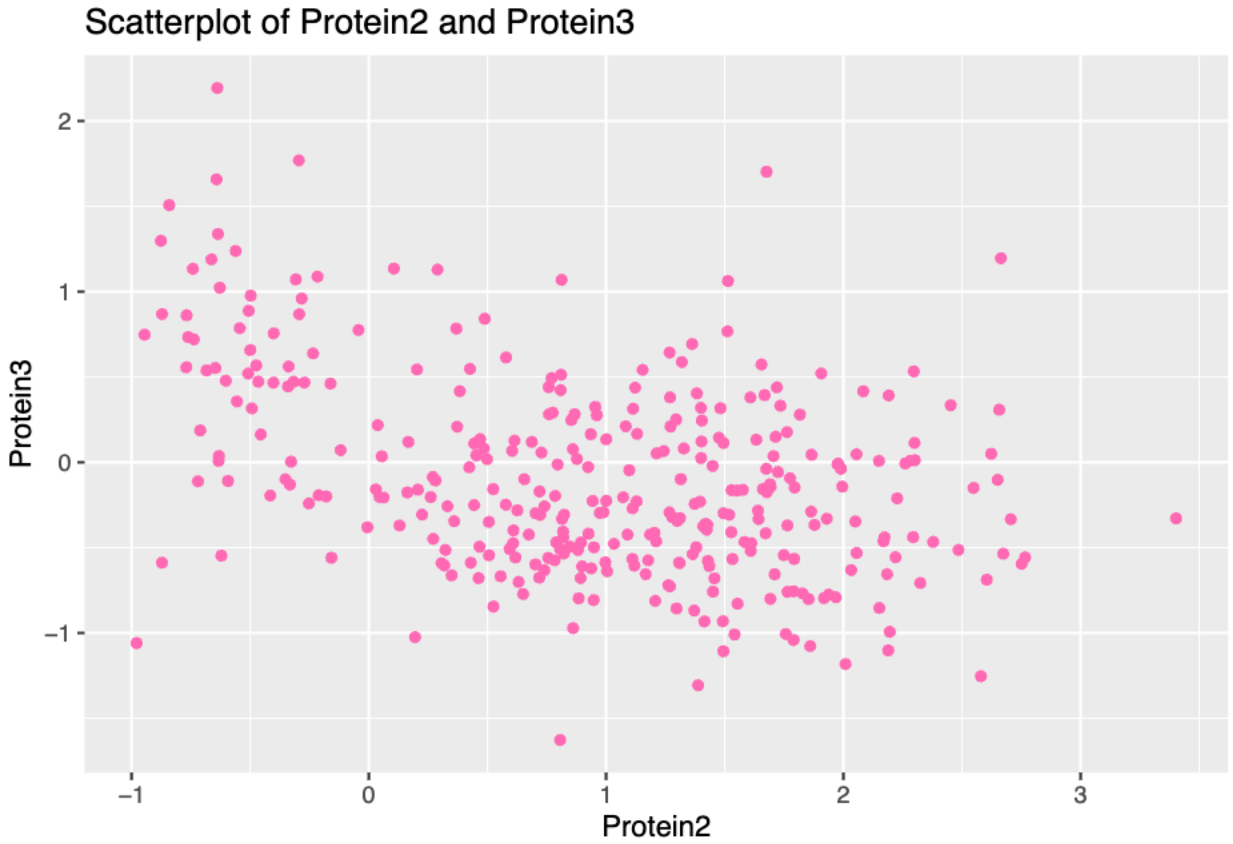
```
ggplot(brca_clean, aes(x=Protein1, y=Protein3)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein1 and Protein3", x="Protein1", y="Protein3")
```

The correlation coefficient of approximately -0.103 displayed in the screenshot indicates a very weak negative linear relationship between "Protein1" and "Protein3". This means that as the value of "Protein1" increases, the value of "Protein3" tends to decrease slightly, but not in a significant way.

```
correlation_coefficient <- cor(brca_clean$Protein2, brca_clean$Protein3, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] -0.4158248
```

```
ggplot(brca_clean, aes(x=Protein2, y=Protein3)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein2 and Protein3", x="Protein2", y="Protein3")
```

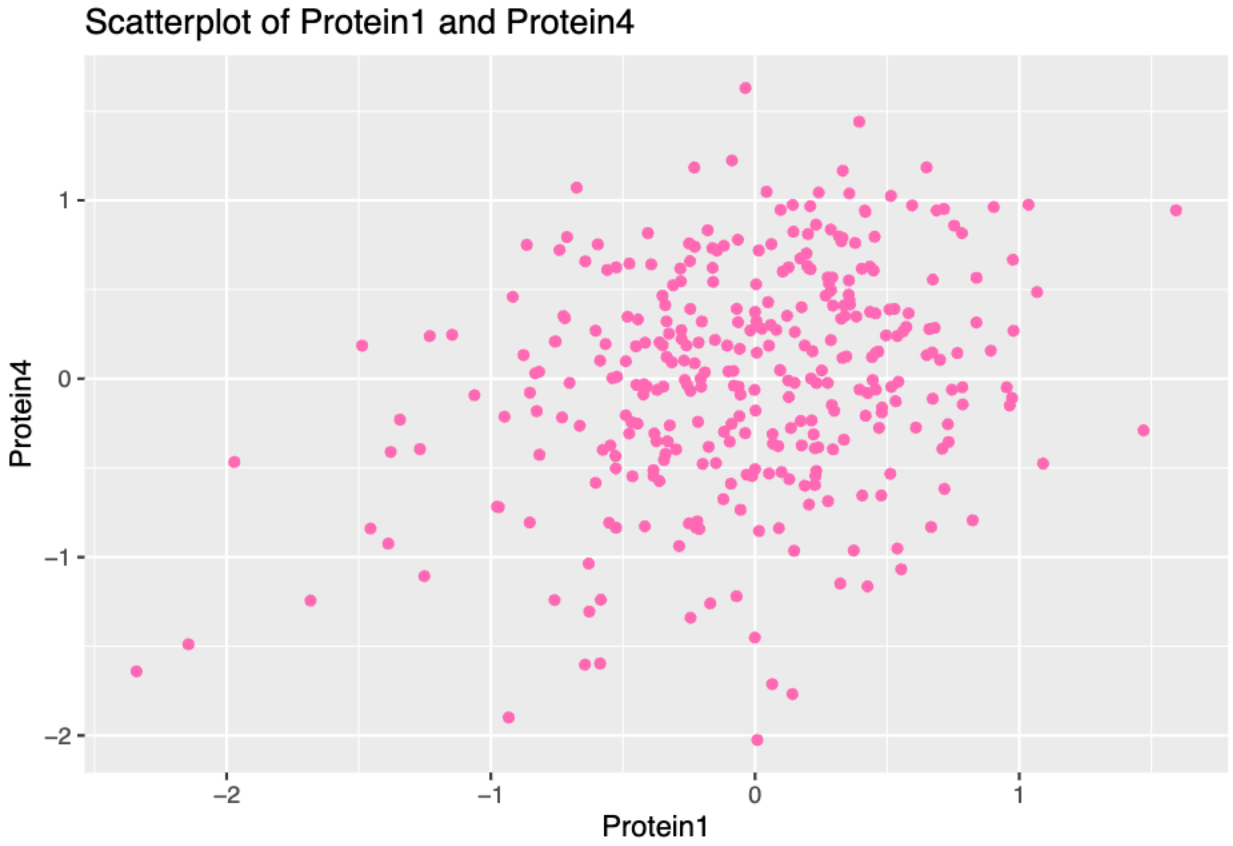


The correlation coefficient of approximately -0.407 suggests a moderate negative linear relationship between "Protein2" and "Protein3". This means that generally, as the values of "Protein2" increase, the values of "Protein3" tend to decrease, and vice versa.

```
correlation_coefficient <- cor(brca_clean$Protein1, brca_clean$Protein4, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] 0.2803431
```

```
ggplot(brca_clean, aes(x=Protein1, y=Protein4)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein1 and Protein4", x="Protein1", y="Protein4")
```

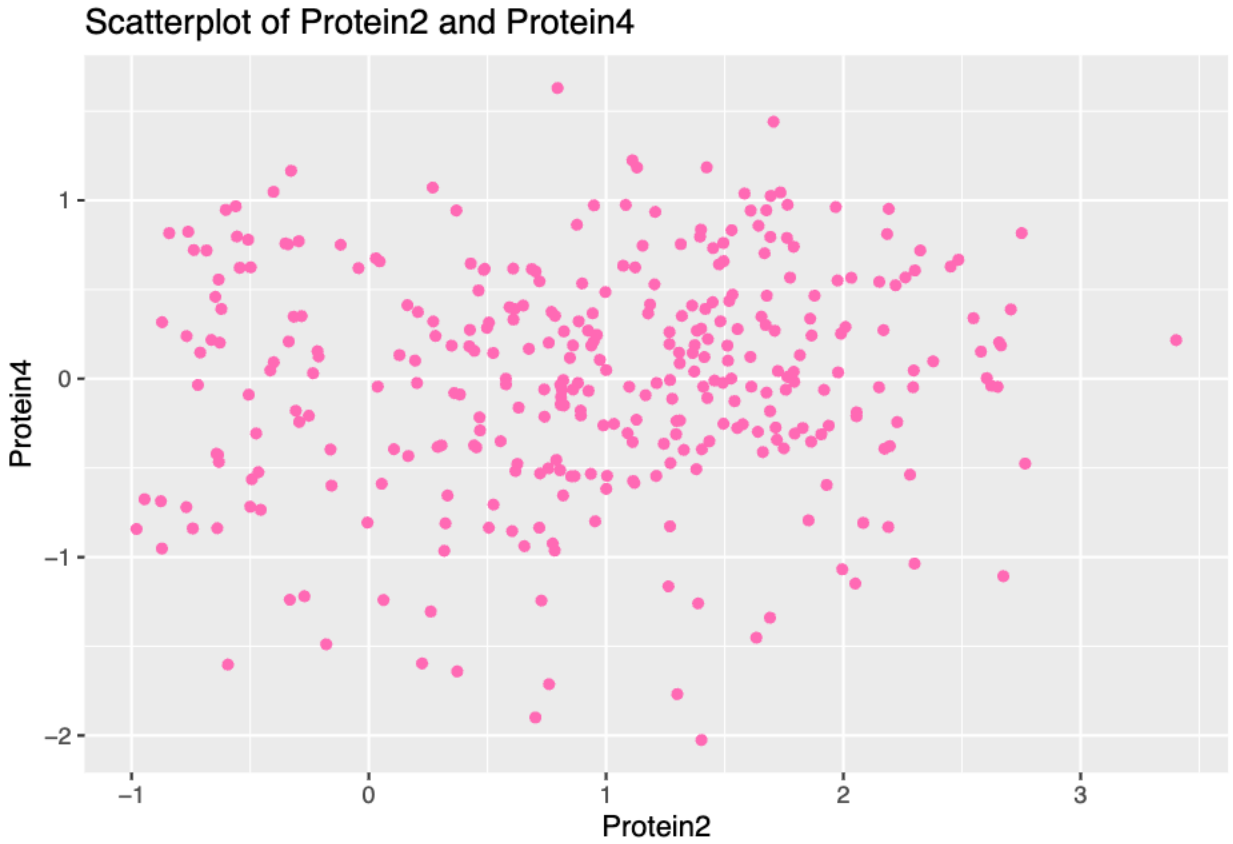


The correlation coefficient of approximately 0.259 indicates a low to moderate positive linear relationship between "Protein1" and "Protein4". This value suggests that as the value of "Protein1" increases, the value of "Protein4" also tends to increase, but not very strongly.

```
correlation_coefficient <- cor(brca_clean$Protein2, brca_clean$Protein4, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] 0.08823857
```

```
ggplot(brca_clean, aes(x=Protein2, y=Protein4)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein2 and Protein4", x="Protein2", y="Protein4")
```

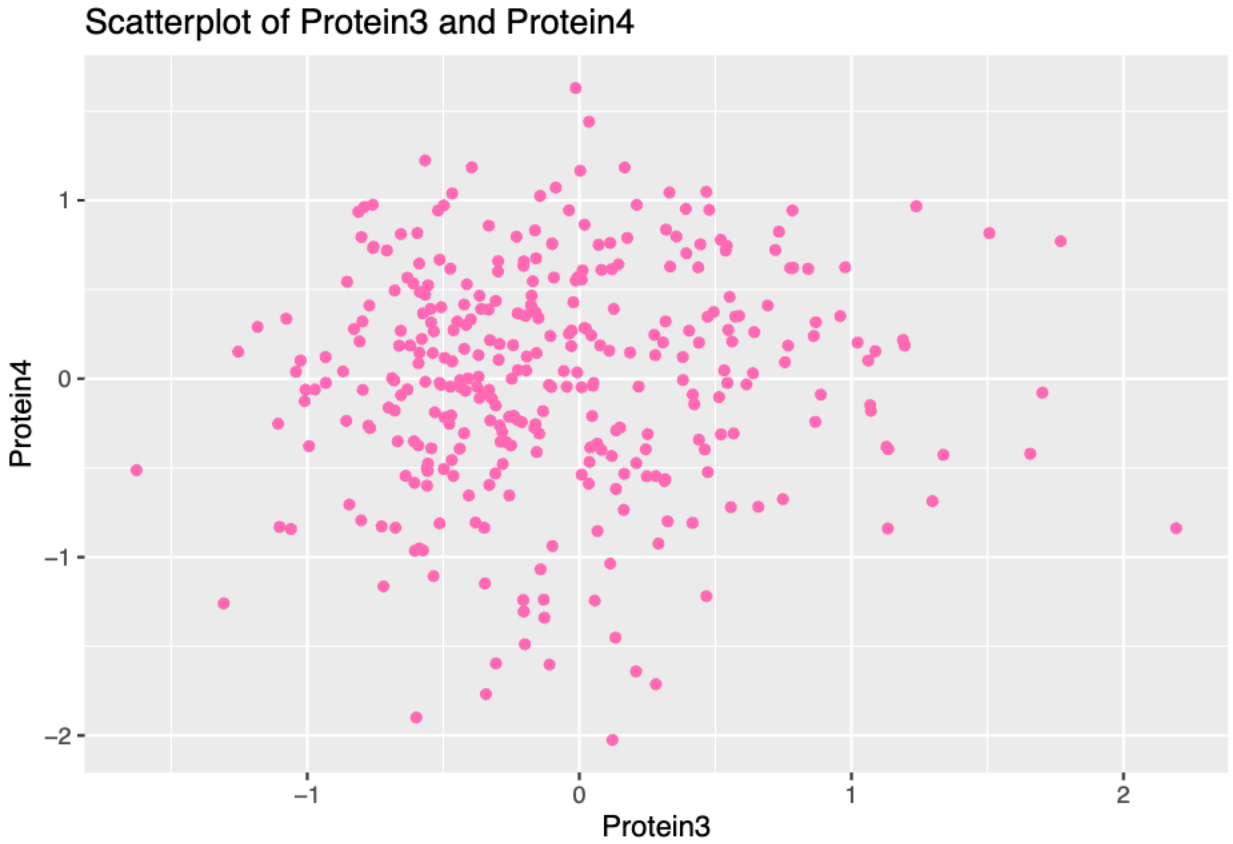


The correlation coefficient of approximately 0.084 suggests that there is a very weak positive linear relationship between "Protein2" and "Protein4". The positive sign indicates that, in general, as the value of "Protein2" increases, the value of "Protein4" also tends to increase slightly. However, the value of the correlation coefficient is very close to zero, which indicates that the linear relationship is negligible.

```
correlation_coefficient <- cor(brca_clean$Protein3, brca_clean$Protein4, use = "complete.obs")
print(correlation_coefficient)
```

```
## [1] 0.06531565
```

```
ggplot(brca_clean, aes(x=Protein3, y=Protein4)) +
  geom_point(color = "hotpink") +
  labs(title="Scatterplot of Protein3 and Protein4", x="Protein3", y="Protein4")
```

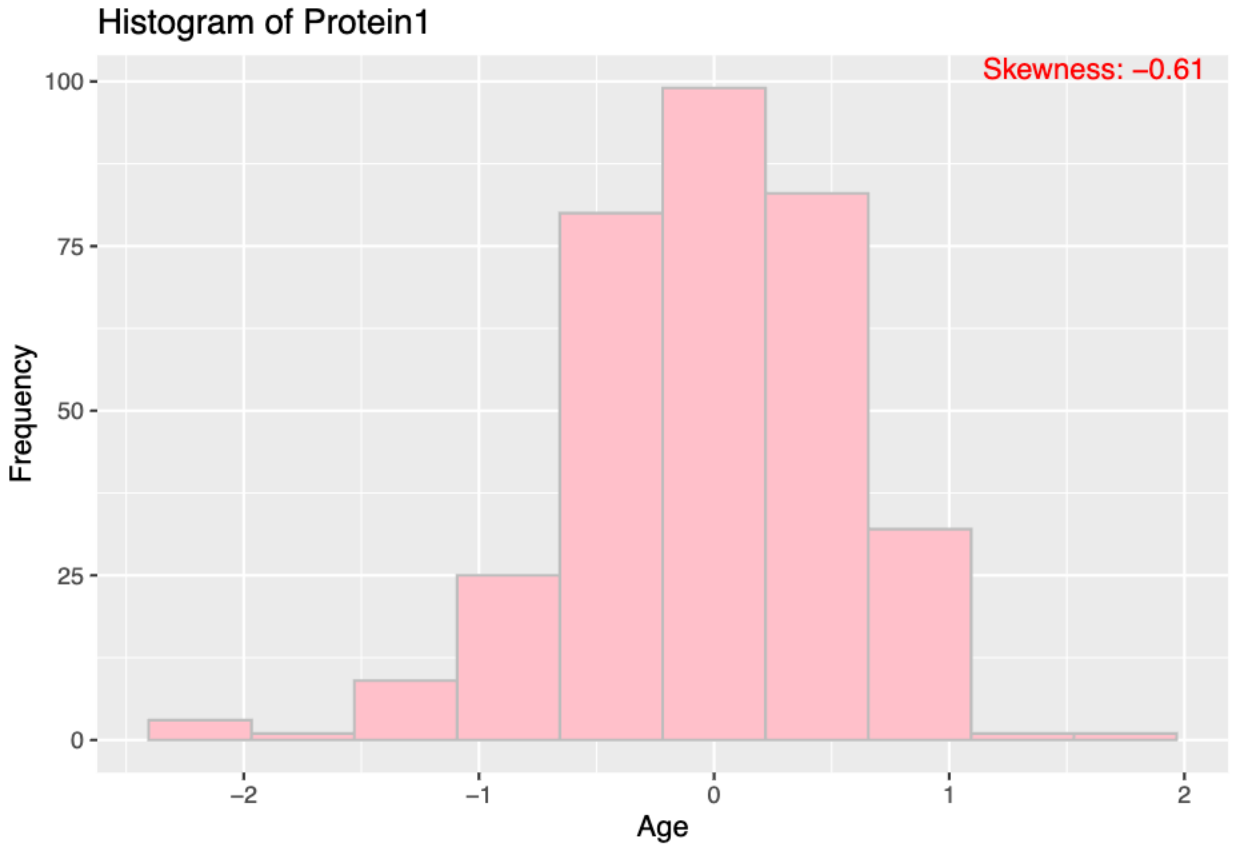


The correlation coefficient of approximately 0.0767 indicates a very weak positive relationship between "Protein3" and "Protein4". This means there is a slight tendency for "Protein4" to increase as "Protein3" increases, but the relationship is not strong.

```
N <- nrow(brca_clean)
num_bins <- ceiling(1 + log2(N))
num_bins
```

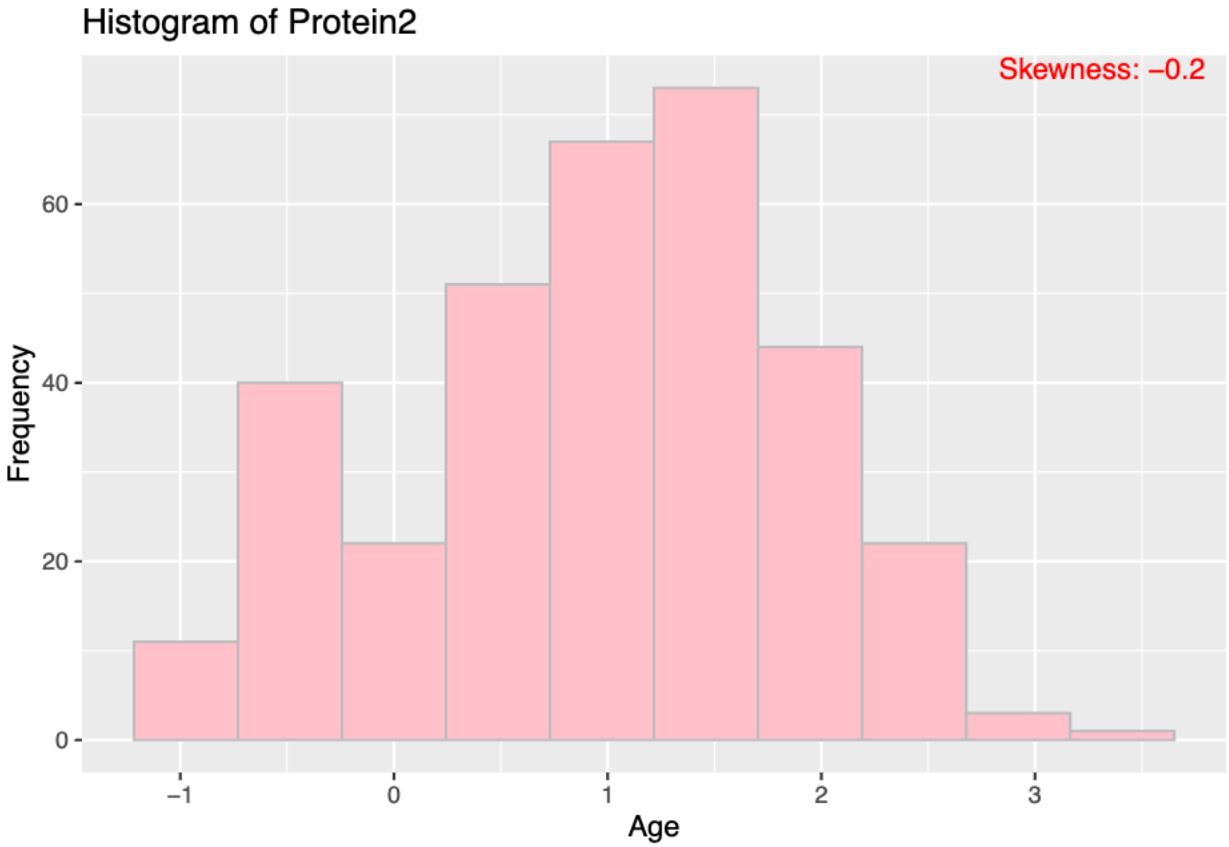
```
## [1] 10
```

```
library(e1071)
skewness_value <- skewness(brca_clean$Protein1, na.rm = TRUE)
protein1_hist <- ggplot(brca_clean, aes(x = Protein1)) +
  geom_histogram(bins = num_bins, fill = "pink", color = "grey") +
  labs(title = "Histogram of Protein1", x = "Age", y = "Frequency")
protein1_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



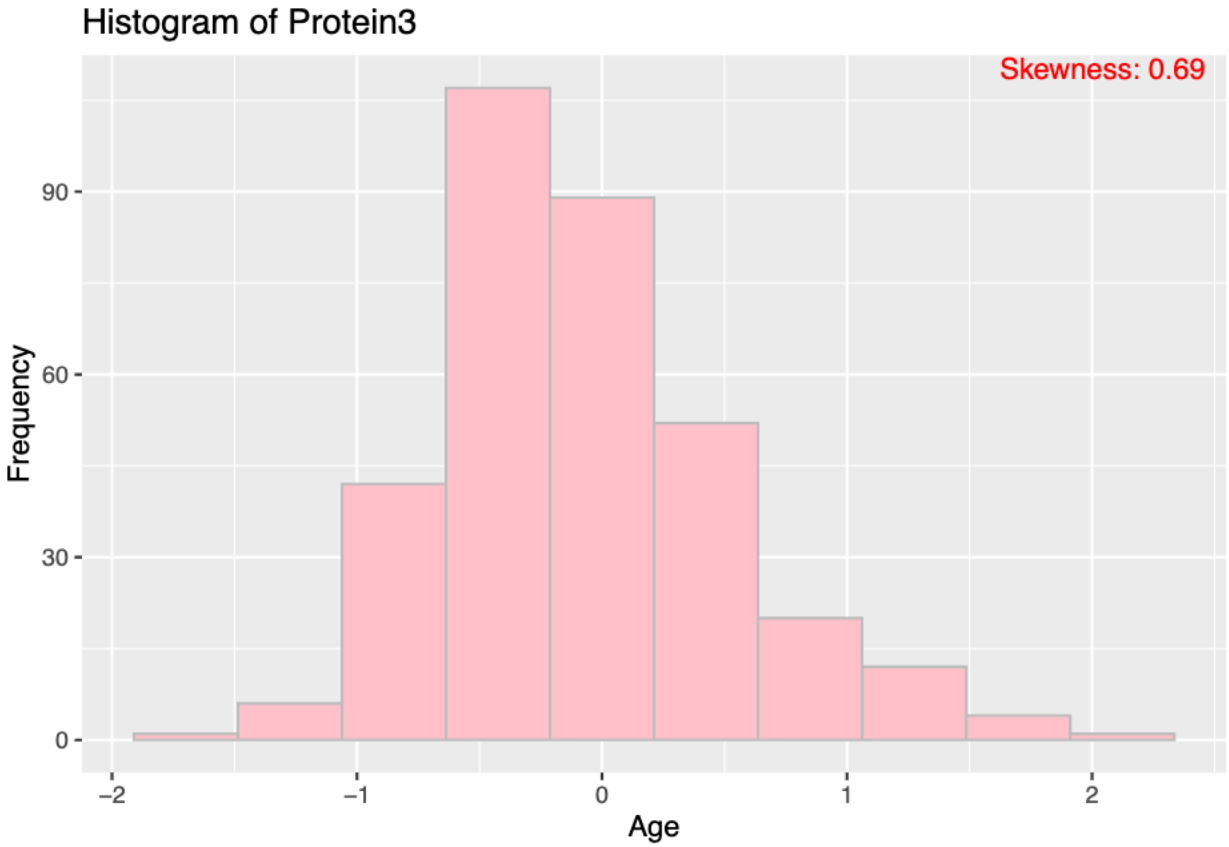
The skewness value of -0.61 suggests Protein1 is not symmetrical and standard statistical tests that assume normality may not be appropriate without transformation or the use of non-parametric methods.

```
skewness_value <- skewness(brca_clean$Protein2, na.rm = TRUE)
protein2_hist <- ggplot(brca_clean, aes(x = Protein2)) +
  geom_histogram(bins = num_bins, fill = "pink", color = "grey") +
  labs(title = "Histogram of Protein2", x = "Age", y = "Frequency")
protein2_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



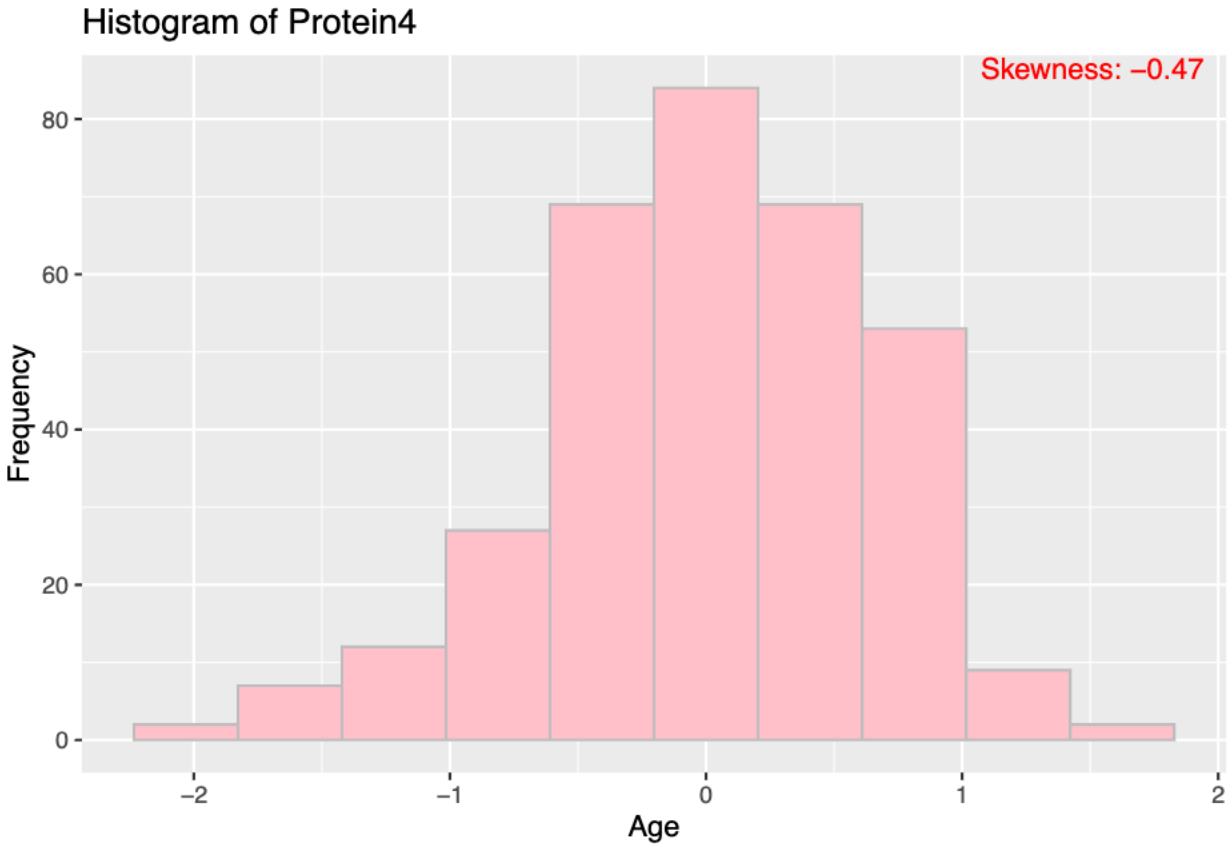
Since the skewness is very close to zero, it indicates that the distribution is almost symmetrical and closely resembles a normal distribution.

```
skewness_value <- skewness(brca_clean$Protein3, na.rm = TRUE)
protein3_hist <- ggplot(brca_clean, aes(x = Protein3)) +
  geom_histogram(bins = num_bins, fill = "pink", color = "grey") +
  labs(title = "Histogram of Protein3", x = "Age", y = "Frequency")
protein3_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



The skewness value of 0.69 suggests Protein3 is not symmetrical and standard statistical tests that assume normality may not be appropriate without transformation or the use of non-parametric methods.

```
skewness_value <- skewness(brca_clean$Protein4, na.rm = TRUE)
protein4_hist <- ggplot(brca_clean, aes(x = Protein4)) +
  geom_histogram(bins = num_bins, fill = "pink", color = "grey") +
  labs(title = "Histogram of Protein4", x = "Age", y = "Frequency")
protein4_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



Since the skewness is very close to zero, it indicates that the distribution is almost symmetrical and closely resembles a normal distribution.

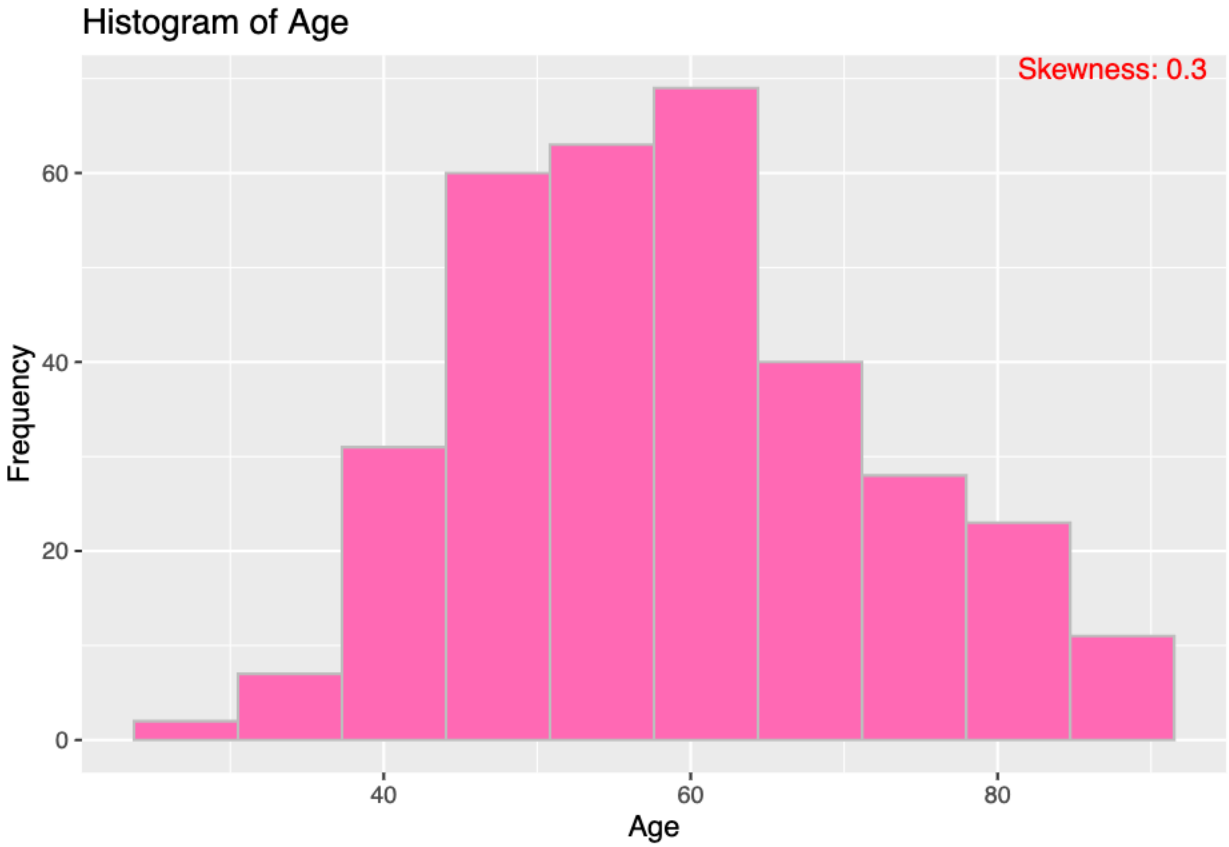
1.4 Age Distribution Visualization

4. Histograms will be employed to visualize the distribution of patient ages, providing a comprehensive overview with an appropriate number of bins.

```
N <- nrow(brca_clean)
num_bins <- ceiling(1 + log2(N))
num_bins
```

```
## [1] 10
```

```
skewness_value <- skewness(brca_clean$Age, na.rm = TRUE)
age_hist <- ggplot(brca_clean, aes(x = Age)) +
  geom_histogram(bins = num_bins, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Age", x = "Age", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



Since the skewness is very close to zero, it indicates that the distribution is almost symmetrical and closely resembles a normal distribution.

2.1 Inference on Mean

1. We will compare the mean age of patients who survived against those who did not. Our null hypothesis (H_0) is that the population mean of age for both the groups is same and alternate hypothesis (H_1) being that they are not. We will be testing this at a significance level α of 5%. We have already proved earlier that our data satisfies normality assumptions as well as CLT.

```r
# Reading the dataset and removing the outliers
brca <- read.csv("BRCA.csv")
brca[brca == ""] <- NA
brca <- brca[complete.cases(brca), ]
# brca <- na.omit(brca)

# Selecting the patients alive and dead
alive <- subset(brca, Patient_Status == "Alive")
dead <- subset(brca, Patient_Status == 'Dead')

# Sample mean and standard deviations
xbar_alive <- mean(alive$Age)
xbar_dead <- mean(dead$Age)
var_alive <- var(alive$Age)
var_dead <- var(dead$Age)

# Now that we have specified our significance level and sample means
# and sample averages. We will use a Welch t-test statistic
```

```r
# Computing sample pooled variance,
# See that our assumptions to use CLT are satisfied (n_alive, n_dead both >= 30)
n_alive <- nrow(alive)
n_dead <- nrow(dead)

# Assuming unequal variances for the two population groups
pooled_var <- ((n_alive-1)*var_alive + (n_dead-1)*var_dead)/(n_alive + n_dead-2)
dof <- n_alive + n_dead - 2                    # By Welch-Satterthwaite equation

t_statistic <- (xbar_alive - xbar_dead)/(sqrt(pooled_var*((1/n_alive) +(1/n_dead))))

# Computing the p-value corresponding to our t-statistic
p_val <- 2*(1-pt(t_statistic, dof))                  # Since 2 tailed t-test for t<0
p_val
```

```
## [1] 0.8258023
```

We can clearly see that the p-val corresponding to the t-statistic is > 0.05. Hence we do not have sufficient evidence to prove that the population mean age of the patients with breast cancer that are alive is same as the population mean for the patients who passed away.

2. ANOVA testing for patients mean age in different cancer stages. We will test this against the significance level $\alpha = 0.05$. As before our null hypothesis (H_0) is that: Mean age of patients in all the three stages is the same. Alternate hypothesis (H_1): Atleast one group mean age differs.

```r
xbar_stage1 <- subset(brca, Tumour_Stage == 'I')
xbar_stage2 <- subset(brca, Tumour_Stage == 'II')
xbar_stage3 <- subset(brca, Tumour_Stage == 'III')

# We will be using aov function directly
model <- aov(brca$Age~brca$Tumour_Stage)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: brca$Age
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## brca$Tumour_Stage    2   1088  543.96  3.3552 0.03616 *
## Residuals          314  50907  162.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can clearly see that p value is less than 0.05. hence the mean age groups are definitely different. Let's use posthoc tests with Scheffe's correction to see which group has different mean age.

```r
library(DescTools)
```

```
##
## Attaching package: 'DescTools'
```

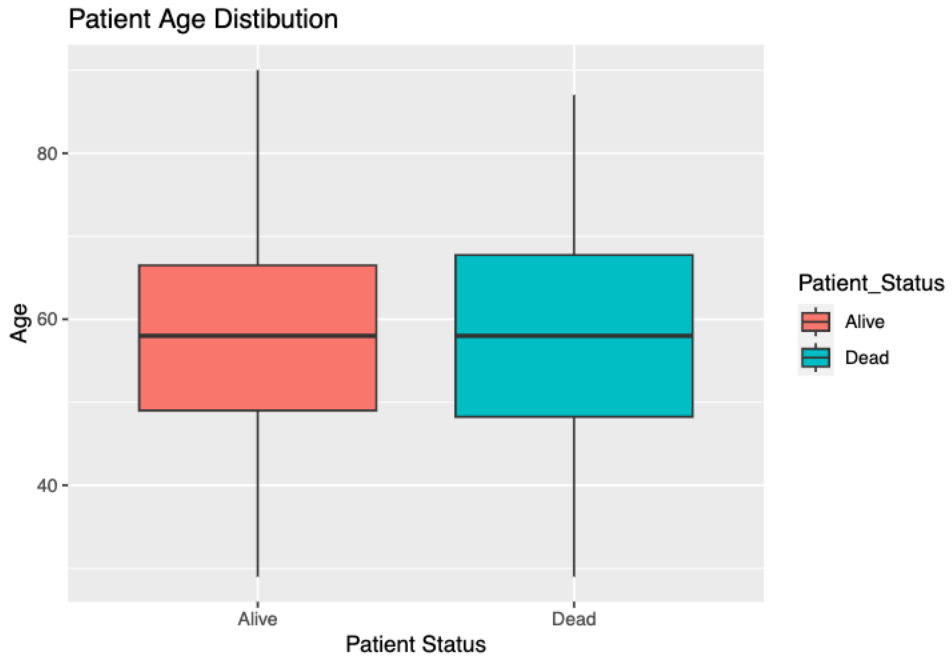
```
## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

```r
# Taking our previously created anova model as input
ScheffeTest(model, conf.level = 0.05)
```

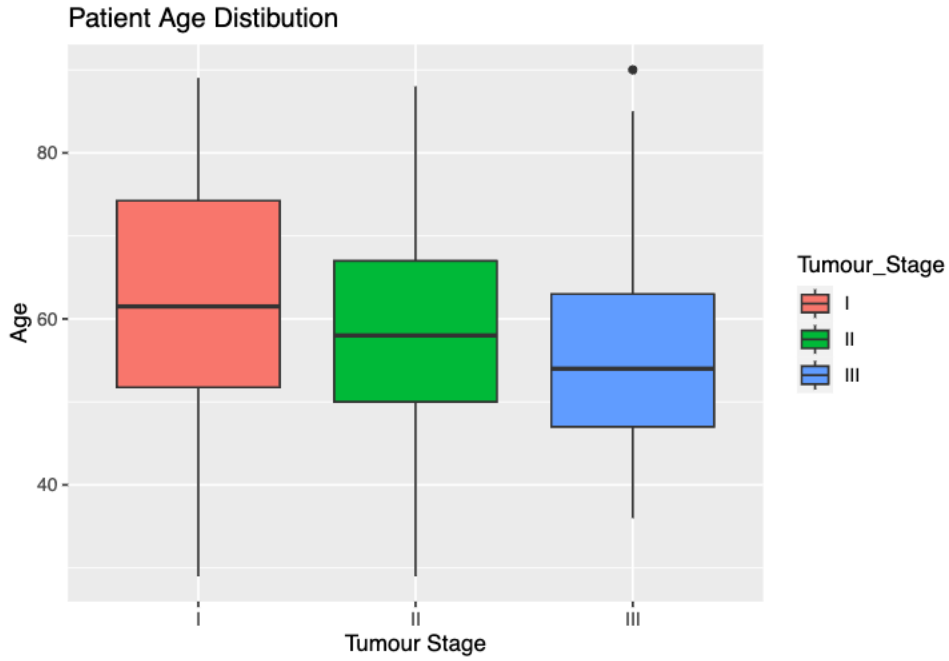
```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     5% family-wise confidence level
##
## $`brca$Tumour_Stage`
##             diff    lwr.ci    upr.ci   pval
## II-I   -3.016667 -3.624661 -2.408673 0.2842
## III-I  -5.673810 -6.376145 -4.971474 0.0364 *
## III-II -2.657143 -3.212523 -2.101762 0.3104
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result we can see that patients in with cancer stages-I and III share different mean ages! Now, let's make a boxplot to see visualize the sample age distribution according to Patient status and Tumour stages.

```
library(DescTools)
library(ggplot2)
box_plot1 <- ggplot(brca, aes(x = Patient_Status, y = brca$Age,
                              fill = Patient_Status)) + geom_boxplot()
box_plot1 <- box_plot1 + ggtitle("Patient Age Distibution")
box_plot1 <- box_plot1 + labs(y = "Age", x = 'Patient Status')

# Now for the tumour stages
box_plot2 <- ggplot(brca, aes(x = Tumour_Stage,
                              y = brca$Age,
                              fill = Tumour_Stage)) + geom_boxplot()
box_plot2 <- box_plot2 + ggtitle("Patient Age Distibution")
box_plot2 <- box_plot2 + labs(y = "Age", x = 'Tumour Stage')
plot(box_plot1)
```



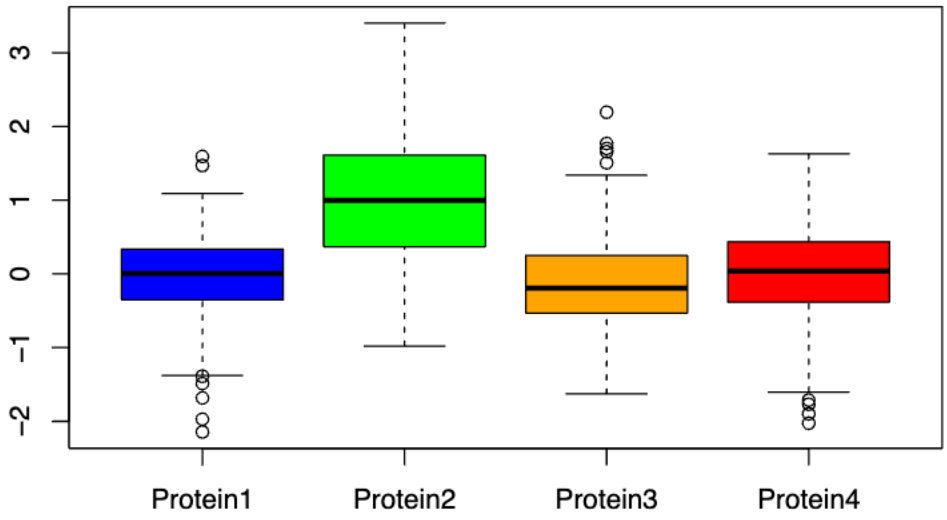

```
plot(box_plot2)
```



2.2 Inference about Variance

We will be comparing protein variances among patients using inference about variance methods. Before that we would like to plot boxplots for the different proteins present and see if we can visually spot any noticeable difference in the spreads.

```
# Getting the protein vectors

prot_1 <- brca$Protein1
prot_2 <- brca$Protein2
prot_3 <- brca$Protein3
prot_4 <- brca$Protein4

# Create a new dataframe
prot_dataframe <- data.frame(Protein1 = prot_1,
                             Protein2 = prot_2,
                             Protein3 = prot_3,
                             Protein4 = prot_4)


# Plot boxplots of each column in the same figure
boxplot(prot_dataframe, col = c("blue", "green", "orange", 'red'),
            main = "Protein Distribution")
```



We can see that the spread for the protein are not the same, there definetely seems to be some variation. Let us formally test this intuition. For this purpose, we will do various two-sample variance tests with corrected significance value using Bonferroni Correction such that our overall desired significance level α is 0.05. Our Null hypothesis (H_0) for all pairwise tests is that population variance for both the groups are same. Alternate hypothesis (H_1) therefore them being not the same. Overall we will be doing $C_2^4 = 6$ of these tests.

```
alpha <- 0.05/((nrow(brca))*(nrow(brca)-1)/2)                    #(aplha < 9.9e-7)

# Doing two-sample variance tests now

# Prot1 and Prot2
var.test(prot_1,prot_2, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_1 and prot_2
## F = 0.36022, num df = 316, denom df = 316, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
##  0.3602197 0.3602198
## sample estimates:
## ratio of variances
##          0.3602198
```

```
# Prot1 and Prot3
var.test(prot_1,prot_3, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_1 and prot_3
## F = 0.85251, num df = 316, denom df = 316, p-value = 0.1567
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
```

```
##  0.8525098 0.8525100
## sample estimates:
## ratio of variances
##          0.8525099
```

```
# Prot1 and Prot4
var.test(prot_1,prot_4, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_1 and prot_4
## F = 0.75487, num df = 316, denom df = 316, p-value = 0.01265
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
##  0.7548671 0.7548674
## sample estimates:
## ratio of variances
##          0.7548673
```

```
# Prot2 and Prot3
var.test(prot_2,prot_3, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_2 and prot_3
## F = 2.3666, num df = 316, denom df = 316, p-value = 4.796e-14
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
##  2.366638 2.366639
## sample estimates:
## ratio of variances
##           2.366638
```

```
# Prot2 and Prot4
var.test(prot_2,prot_4, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_2 and prot_4
## F = 2.0956, num df = 316, denom df = 316, p-value = 8.146e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
##  2.095574 2.095575
## sample estimates:
## ratio of variances
##           2.095574
```

```
# Prot3 and Prot4
var.test(prot_3,prot_4, ratio = 1, alt = 'two.sided', conf.level = alpha)
```

```
##
##  F test to compare two variances
##
## data:  prot_3 and prot_4
```

```
## F = 0.88546, num df = 316, denom df = 316, p-value = 0.2801
## alternative hypothesis: true ratio of variances is not equal to 1
## 9.98283e-05 percent confidence interval:
##  0.8854644 0.8854646
## sample estimates:
## ratio of variances
##          0.8854645
```

Even after a corrected significance level $\alpha^* = 9.9e{-}7$ we find that the p-value corresponding to the following pairs less than α^* -

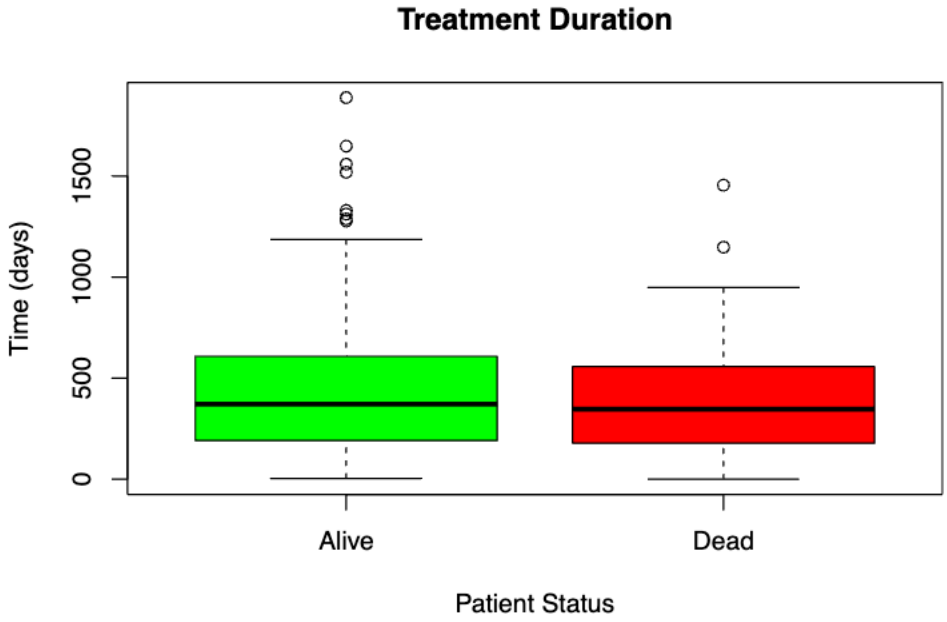
Protein Pair	p-value
I-II	$<$ 2.2e-16
II-III	4.796e-14
II-IV	8.146e-11

We see that Protein-II has a different variance in comparison to the rest and can visually confirm this.

2.3 Inference on Treatment Duration

We would like to find if the mean treatment time and the spread as same for the patients who are alive and who have passed away

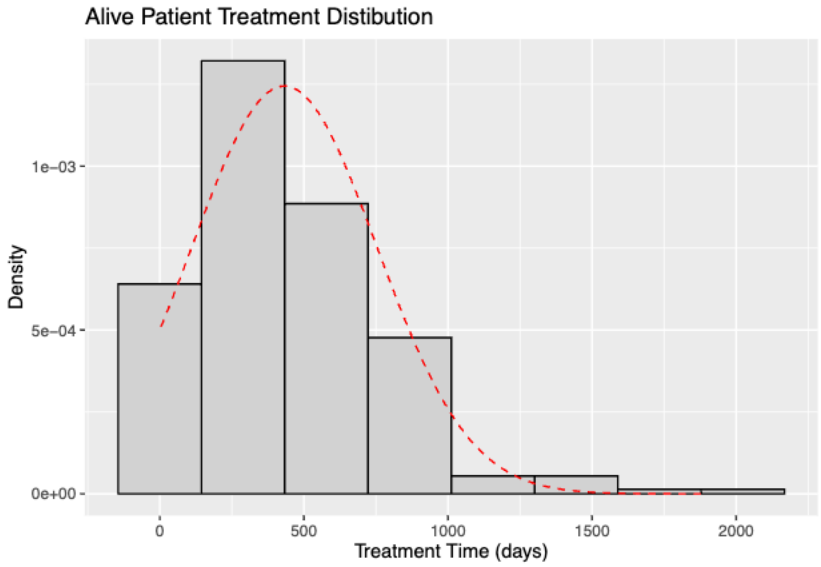
```r
# Removing rows that have last visit > today's date
today <- as.Date(format(Sys.Date()))
brca$Date_of_Surgery <- as.Date(brca$Date_of_Surgery, format = "%d-%b-%y")
brca$Date_of_Last_Visit <- as.Date(brca$Date_of_Last_Visit, format = "%d-%b-%y")
brca_mod <- brca[brca$Date_of_Last_Visit <= today, ]

alive <- subset(brca_mod, Patient_Status == "Alive")
dead <- subset(brca_mod, Patient_Status == 'Dead')
# Treatment duration

treatment_dur_alive <- as.numeric(alive$Date_of_Last_Visit - alive$Date_of_Surgery)
treatment_dur_dead <- as.numeric(dead$Date_of_Last_Visit - dead$Date_of_Surgery)

data <- list(Alive = treatment_dur_alive, Dead = treatment_dur_dead)

# Plot boxplots of each column in the same figure
boxplot(data, col = c("green", 'red'), main = "Treatment Duration",
        xlab = 'Patient Status', ylab = "Time (days)")
```



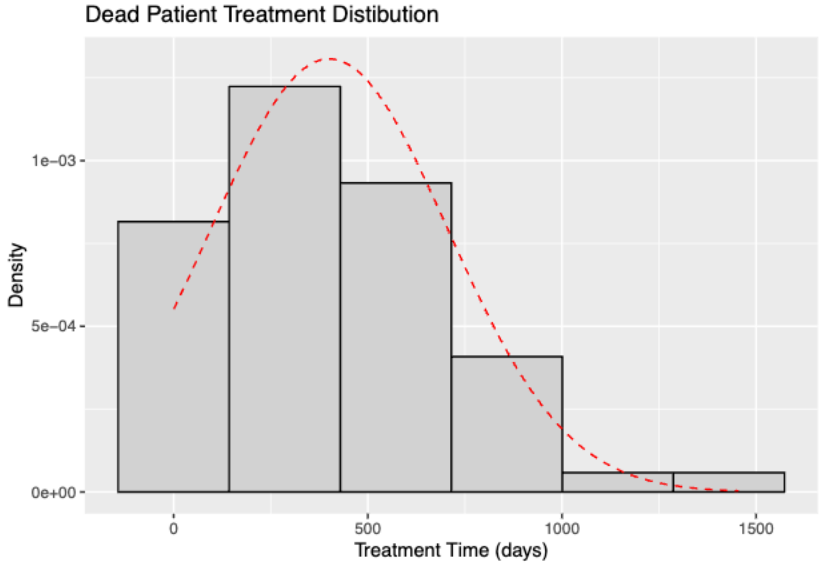

```
#
```

```
# For patients alive
binwidth <- ceiling((max(treatment_dur_alive) -
                       min(treatment_dur_alive))
                    / (1 + log(length(treatment_dur_alive))))
alive_data_hist <- data.frame(Alive = treatment_dur_alive)
alive_hist <- ggplot(alive_data_hist, aes(Alive)) +
  geom_histogram(binwidth = binwidth,
                 aes(y = ..density..), fill='lightgray', col='black')

alive_hist <- alive_hist + xlab("Treatment Time (days)") + ylab("Density")+
  stat_function(fun = dnorm, args = list(mean=mean(treatment_dur_alive),
                                         sd=sd(treatment_dur_alive)),
                col='red', lwd=0.5, lty='dashed')
alive_hist <- alive_hist + ggtitle("Alive Patient Treatment Distibution")
plot(alive_hist)
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# For patients not alive
binwidth <- ceiling((max(treatment_dur_dead) -
                     min(treatment_dur_dead))
                    / (1 + log(length(treatment_dur_dead))))
dead_data_hist <- data.frame(Dead = treatment_dur_dead)
dead_hist <- ggplot(dead_data_hist, aes(Dead)) +
  geom_histogram(binwidth = binwidth,
                 aes(y = ..density..), fill='lightgray', col='black')

dead_hist <- dead_hist + xlab("Treatment Time (days)") + ylab("Density")+
  stat_function(fun = dnorm, args = list(mean=mean(treatment_dur_dead),
                                         sd=sd(treatment_dur_dead)),
                col='red', lwd=0.5, lty='dashed')
dead_hist <- dead_hist + ggtitle("Dead Patient Treatment Distibution")
plot(dead_hist)
```



2.4 Inference about proportions

In this section, we intend to check the survival rate for the three tumour stages. We will first test to see if there is a difference between the survival rate for patients with different tumour stages. We will take the signficance value $\alpha = 0.05$ for all the tests.

```
library(DescTools)
data <- read.csv("BRCA.csv")
summary(data)
```

```
##   Patient_ID             Age           Gender             Protein1
##  Length:341         Min.   :29.00   Length:341         Min.   :-2.340900
##  Class :character   1st Qu.:49.00   Class :character   1st Qu.:-0.358888
##  Mode  :character   Median :58.00   Mode  :character   Median : 0.006129
##                     Mean   :58.89                      Mean   :-0.029991
##                     3rd Qu.:68.00                      3rd Qu.: 0.343598
##                     Max.   :90.00                      Max.   : 1.593600
##                     NA's   :7                          NA's   :7
##     Protein2          Protein3          Protein4         Tumour_Stage
##  Min.   :-0.9787   Min.   :-1.6274   Min.   :-2.025500   Length:341
##  1st Qu.: 0.3622   1st Qu.:-0.5137   1st Qu.:-0.377090   Class :character
##  Median : 0.9928   Median :-0.1732   Median : 0.041768   Mode  :character
##  Mean   : 0.9469   Mean   :-0.0902   Mean   : 0.009819
##  3rd Qu.: 1.6279   3rd Qu.: 0.2784   3rd Qu.: 0.425630
##  Max.   : 3.4022   Max.   : 2.1934   Max.   : 1.629900
##  NA's   :7         NA's   :7         NA's   :7
##   Histology          ER.status          PR.status         HER2.status
##  Length:341         Length:341         Length:341         Length:341
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Surgery_type       Date_of_Surgery    Date_of_Last_Visit Patient_Status
##  Length:341         Length:341         Length:341         Length:341
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
data <- data[data$Patient_Status != "", ]
x <- table(data$Tumour_Stage)
print(x)
```

```
##
##   I  II III
##  61 182  78
```

```
x1 <- nrow(data[data$Tumour_Stage == "I" & data$Patient_Status == "Alive", ])
x2 <- nrow(data[data$Tumour_Stage == "II" & data$Patient_Status == "Alive", ])
x3 <- nrow(data[data$Tumour_Stage == "III" & data$Patient_Status == "Alive", ])
n1 <- 61
n2 <- 182
```

```r
n3 <- 78
```

Firstly, we will check for the normality assumptions based on CLT to check if using prop test would be valid. From the data, $n_1p_1 = 51$ and $n_1(1-p_1) = 10$; $n_2p_2 = 144$ and $n_2(1-p_2) = 38$; and $n_3p_3 = 60$ and $n_3(1-p_3) = 18$. As all the values are greater than 5, we can use prop test on this data. We will perform two sample prop tests on different pairs of tumour stages data. The Null hypothesis H_0 will be that the both proportions will be equal and the alternative hypothesis H_1 will be that the both proportions will not be equal.

```r
x12 <- c(x1,x2)
n12 <- c(n1,n2)
x13 <- c(x1,x3)
n13 <- c(n1,n3)
x23 <- c(x2,x3)
n23 <- c(n2,n3)
prop.test(x12,n12,alternative = "two.sided", conf.level = 0.95, correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x12 out of n12
## X-squared = 0.33148, df = 1, p-value = 0.5648
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.07616953  0.16588310
## sample estimates:
##    prop 1    prop 2
## 0.8360656 0.7912088
```

```r
prop.test(x13,n13,alternative = "two.sided", conf.level = 0.95, correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x13 out of n13
## X-squared = 0.58045, df = 1, p-value = 0.4461
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.07958197  0.21325158
## sample estimates:
##    prop 1    prop 2
## 0.8360656 0.7692308
```

```r
prop.test(x23,n23,alternative = "two.sided", conf.level = 0.95, correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x23 out of n23
## X-squared = 0.053105, df = 1, p-value = 0.8177
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.09776574  0.14172178
## sample estimates:
##    prop 1    prop 2
## 0.7912088 0.7692308
```

From the above, we can see that p-value is much much greater than significance value for all the three tests, we fail to reject H_0 for all of them. Hence, there is no significance evidence to say that the survival rate for patients with different tumour stages is different for any pair of tumour stages.\

Now, we will try and see if the survival rate is greater when the tumor stage detected is early i.e, the survival rate for patients with tumour stage I is greater than that of tumour stage III. Hence, $H_0 : p_I \leq p_{III}$ and $H_1 : p_I > p_{III}$ We will perform one sided prop test see if the above claim is true.

```
prop.test(x13,n13,alternative = "greater", conf.level = 0.95,correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x13 out of n13
## X-squared = 0.58045, df = 1, p-value = 0.2231
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.05839044  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.8360656 0.7692308
```

From the above, we can see that the p-value is greater than the significance value, we fail to reject H_0. Hence, there is no significant evidence to say that the survival rate for patients with tumour stage I is greater than that of tumour stage III.

2.5 Inference on Independence

```
hist <- table(data$Histology)
print(hist)
```

```
##
##  Infiltrating Ductal Carcinoma Infiltrating Lobular Carcinoma
##                            226                             83
##             Mucinous Carcinoma
##                             12
```

Now, we will try to see if the Histology is associated with the tumour stage of the patients. For this, we will use χ^2 independence test. First, we collected the data from the dataset relating histology and tumour size, it is as follows:

Tumour stage	Infiltrating Ductal Carcinoma	Mucinous Carcinoma	Infiltrating Lobular Carcinoma
III	57	0	21
II	121	9	52
I	48	3	10

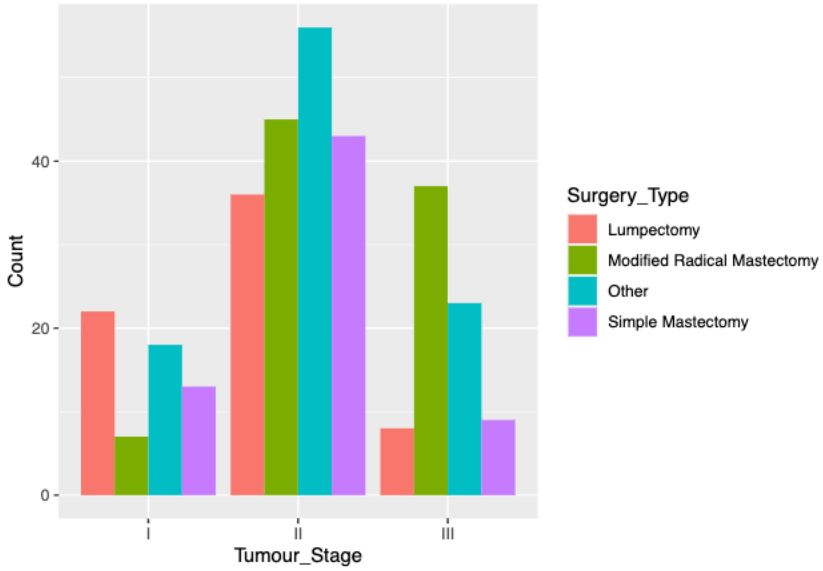
For this test, H_0: Histology is independent to the tumour stage and H_1: Histology is associated with the tumour stage. We will use $\alpha = 0.05$.

```
mat <- c()
for (i in unique(data$Histology)) {
  for (j in unique(data$Tumour_Stage)) {
    count <- nrow(data[data$Tumour_Stage == j & data$Histology == i, ])
    mat <- append(mat,count)
  }
}
mat1 <- matrix(mat, nrow = 3, ncol = 3)
chisq.test(mat1,simulate.p.value = TRUE, correct = F)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  mat1
## X-squared = 7.5888, df = NA, p-value = 0.1059
```

From the above, we can see that the p-value is greater the significance value, we fail to reject H_0. Hence, there is no significant evidence to say that the Histology is associated with the tumour stage of the patients. We intend to see whether surgery type varies with the tumor stage. For this purpose we used barplots.

```
# Tumour stage and surgery type
counts <- table(brca$Surgery_type, brca$Tumour_Stage)
# Converting the counts to a data frame
count_df <- as.data.frame(counts)
colnames(count_df) <- c("Surgery_Type", "Tumour_Stage", "Count")
plots <- ggplot(count_df, aes(x = Tumour_Stage, y = Count, fill = Surgery_Type))
plots <- plots + geom_bar(stat = "identity", position = "dodge")
plots
```



We can see no clear inference can be made other than the fact that patients in stage-II tend to receive surgery more than stage-I and stage-III patients. The dominant surgery type for the respective stages are summarized below. Note that other means that some other type of surgery was performed that the ones mentioned in the dataset.

Tumor Stage	Dominant Surgery Type
I	Lumpectomy
II	Other
III	Modified Radical Mastectomy

Now, we will try and see if above inferences are concrete by performing appropriate test. We will use the same significance value as before. Here, H_0: Surgery type is independent to the tumour stage and H_1: Surgery type is associated with the tumour stage. The data looks like:

Tumour stage	Modified Radical Mastectomy	Lumpectomy	Other	Simple Mastectomy
III	38	8	23	9
II	47	36	56	43
I	7	22	19	13

```
mat4 <- c()
for (i in unique(data$Surgery_type)) {
  for (j in unique(data$Tumour_Stage)) {
    count <- nrow(data[data$Tumour_Stage == j & data$Surgery_type == i, ])
    mat4 <- append(mat4,count)
  }
}
mat5 <- matrix(mat4, nrow = 3, ncol = 4)
chisq.test(mat5,simulate.p.value = TRUE, correct = F)
```

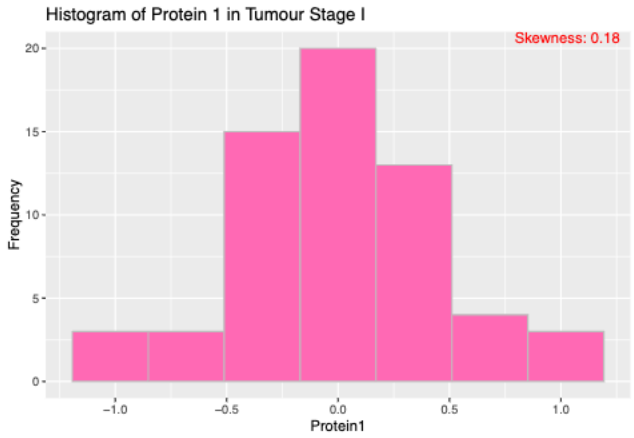
```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  mat5
## X-squared = 32.98, df = NA, p-value = 0.0004998
```

From the above, we can see that the p-value is less than the significance value, we reject H_0. Hence, there is significant evidence to say that the surgery type done is based on the tumour stage the patient is in.

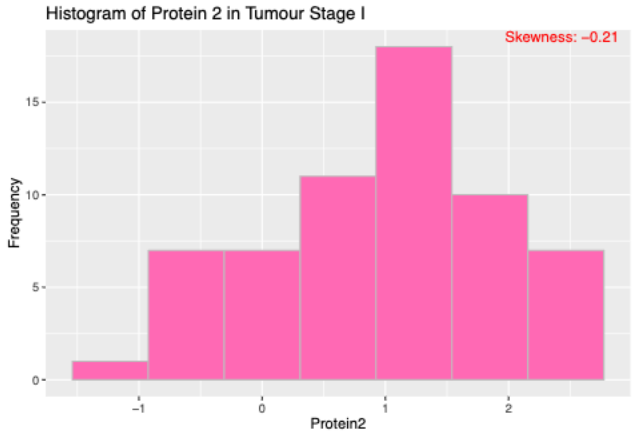
2.6 ANOVA

Here, we want to check if the mean value of different proteins is same for the patients at different tumour stages. Hence, we will perform ANOVA (as the tumour stages are greater than 2) to see what we can infer. To perform ANOVA, we need to meet three requirements such as the distribution of proteins should be normal across all the tumour stages, the variance of proteins should be equal across all the tumour stages and the protein values should be independent. We can almost say that the protein values are independent. We have to check for normality and equal variance. Below are different plots to let us know whether the requirements are met.

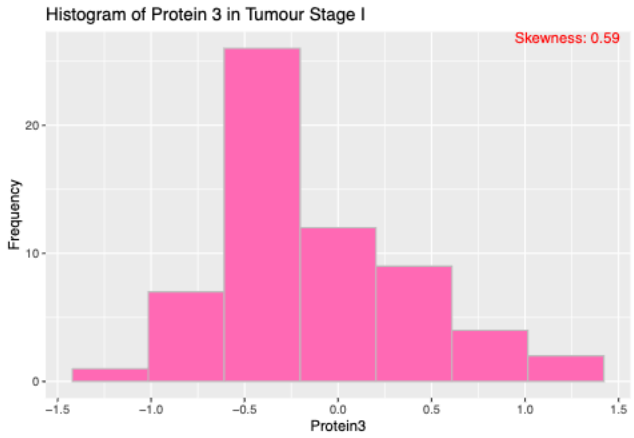
```
library(e1071)
library(ggplot2)
stage1 <- subset(data, Tumour_Stage == 'I')
stage2 <- subset(data, Tumour_Stage == 'II')
stage3 <- subset(data, Tumour_Stage == 'III')
N1 <- nrow(stage1)
num_bins_1 <- ceiling(1 + log2(N1))
skewness_value <- skewness(stage1$Protein1, na.rm = TRUE)
age_hist <- ggplot(stage1, aes(x = Protein1)) +
  geom_histogram(bins = num_bins_1, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 1 in Tumour Stage I", x = "Protein1", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



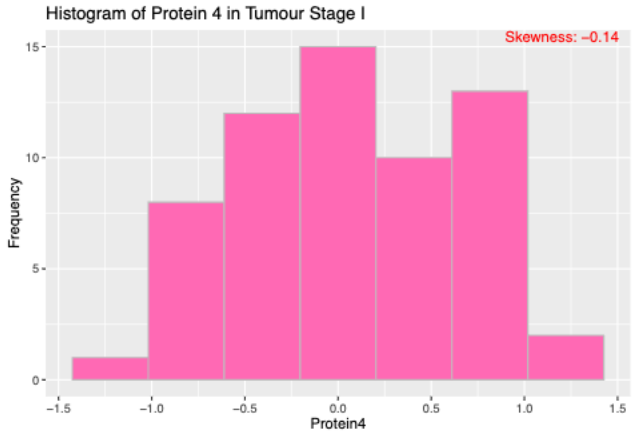

```
skewness_value <- skewness(stage1$Protein2, na.rm = TRUE)
age_hist <- ggplot(stage1, aes(x = Protein2)) +
  geom_histogram(bins = num_bins_1, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 2 in Tumour Stage I", x = "Protein2", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



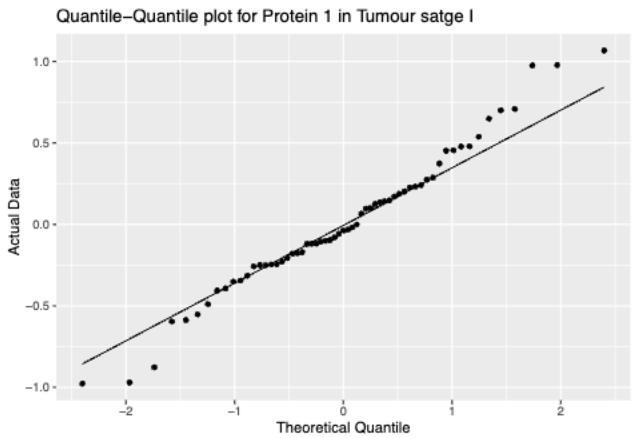

```
skewness_value <- skewness(stage1$Protein3, na.rm = TRUE)
age_hist <- ggplot(stage1, aes(x = Protein3)) +
  geom_histogram(bins = num_bins_1, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 3 in Tumour Stage I", x = "Protein3", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```




```
skewness_value <- skewness(stage1$Protein4, na.rm = TRUE)
age_hist <- ggplot(stage1, aes(x = Protein4)) +
  geom_histogram(bins = num_bins_1, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 4 in Tumour Stage I", x = "Protein4", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```

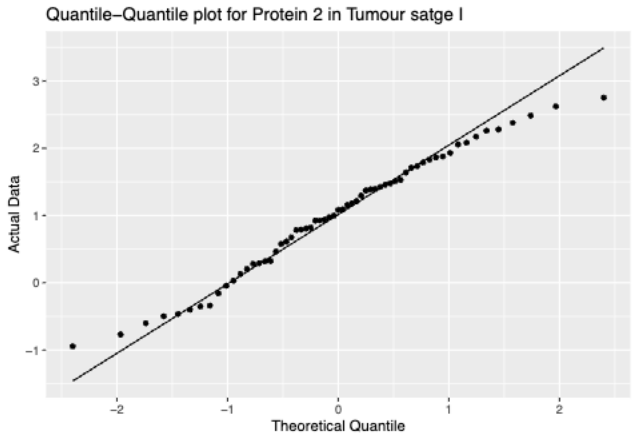



```
qq1 <- ggplot(stage1, aes(sample=Protein1)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq1
```

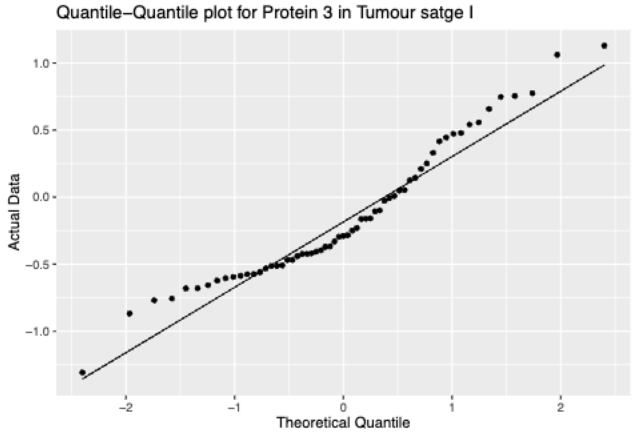



```
qq2 <- ggplot(stage1, aes(sample=Protein2)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
```

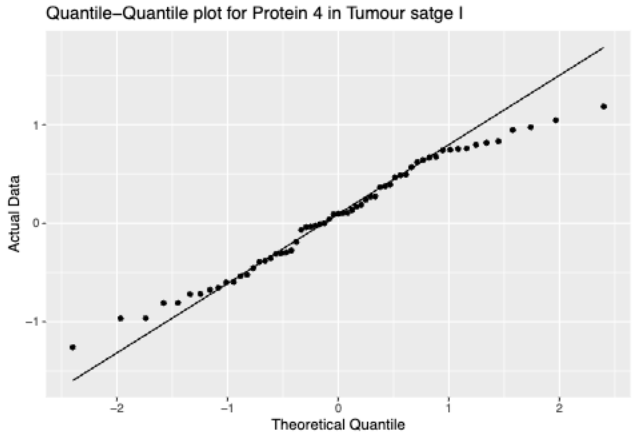
```
qq2
```




```
qq3 <- ggplot(stage1, aes(sample=Protein3)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq3
```

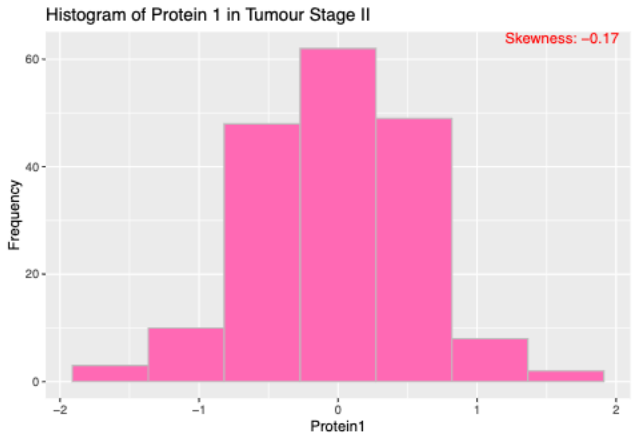



```
qq4 <- ggplot(stage1, aes(sample=Protein4)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq4
```

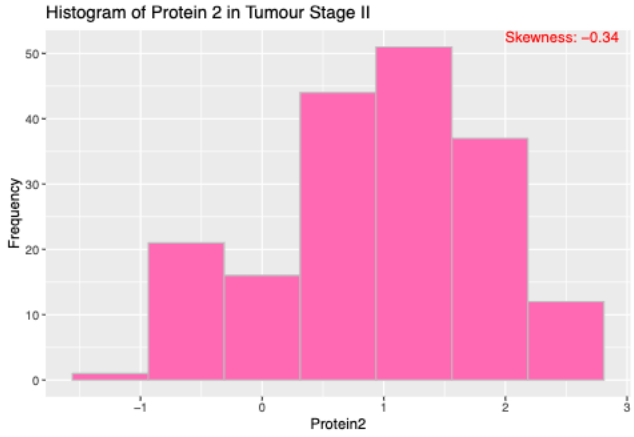



```
N2 <- nrow(stage2)
num_bins_2 <- ceiling(1 + log2(N1))
skewness_value <- skewness(stage2$Protein1, na.rm = TRUE)
age_hist <- ggplot(stage2, aes(x = Protein1)) +
```

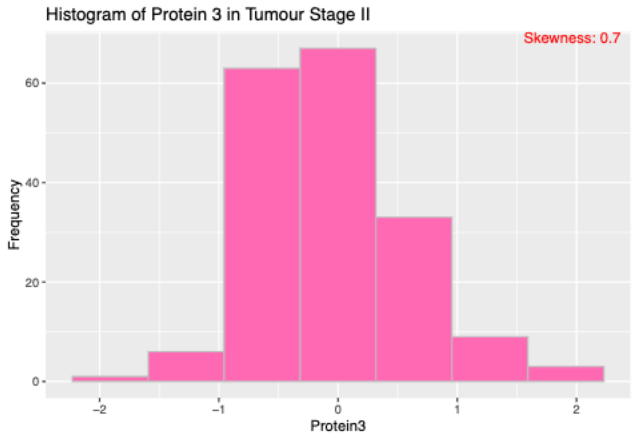
```
  geom_histogram(bins = num_bins_2, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 1 in Tumour Stage II", x = "Protein1", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



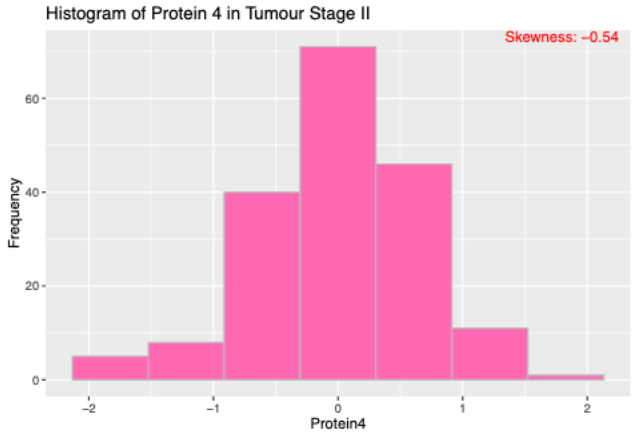

```
skewness_value <- skewness(stage2$Protein2, na.rm = TRUE)
age_hist <- ggplot(stage2, aes(x = Protein2)) +
  geom_histogram(bins = num_bins_2, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 2 in Tumour Stage II", x = "Protein2", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



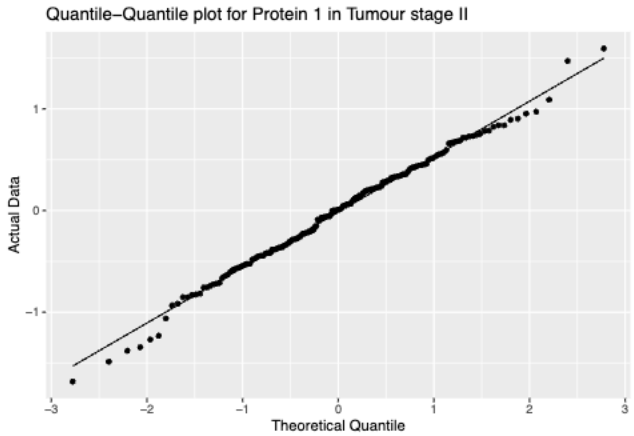

```
skewness_value <- skewness(stage2$Protein3, na.rm = TRUE)
age_hist <- ggplot(stage2, aes(x = Protein3)) +
  geom_histogram(bins = num_bins_2, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 3 in Tumour Stage II", x = "Protein3", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```




```
skewness_value <- skewness(stage2$Protein4, na.rm = TRUE)
age_hist <- ggplot(stage2, aes(x = Protein4)) +
  geom_histogram(bins = num_bins_2, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 4 in Tumour Stage II", x = "Protein4", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```

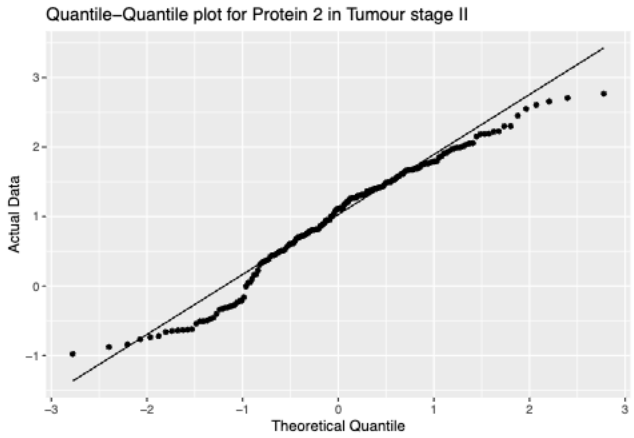



```
qq5 <- ggplot(stage2, aes(sample=Protein1)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq5
```

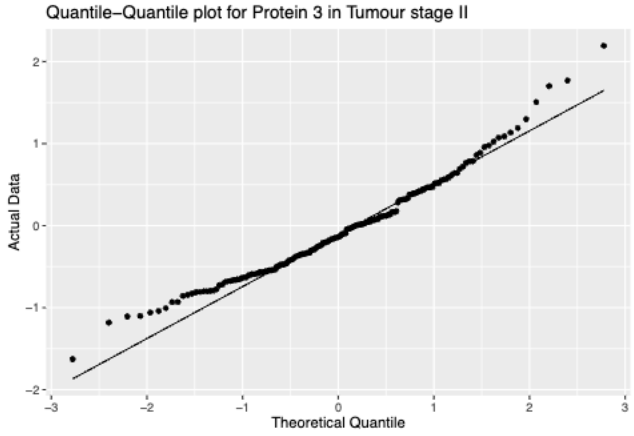



```
qq6 <- ggplot(stage2, aes(sample=Protein2)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
```

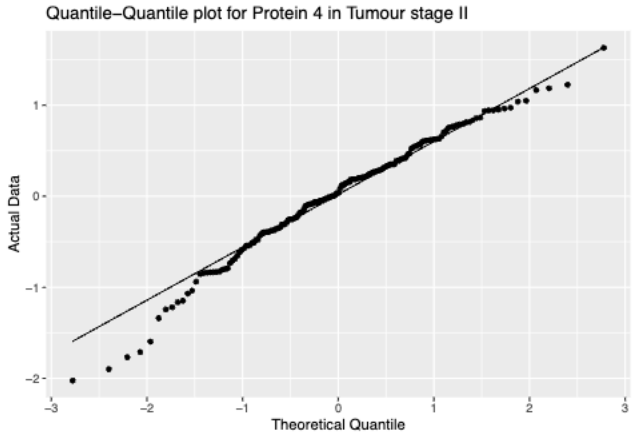
```
qq6
```




```
qq7 <- ggplot(stage2, aes(sample=Protein3)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq7
```

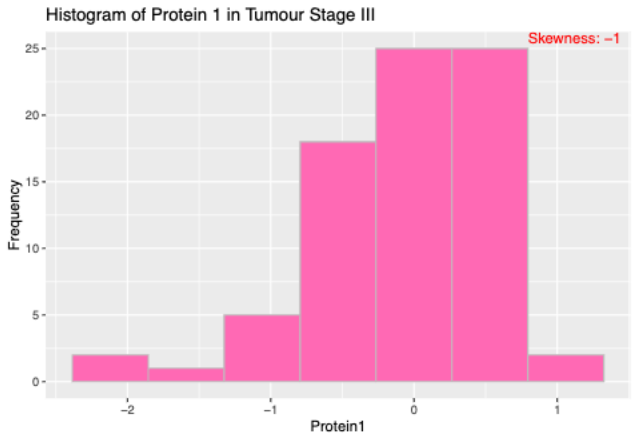



```
qq8 <- ggplot(stage2, aes(sample=Protein4)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq8
```

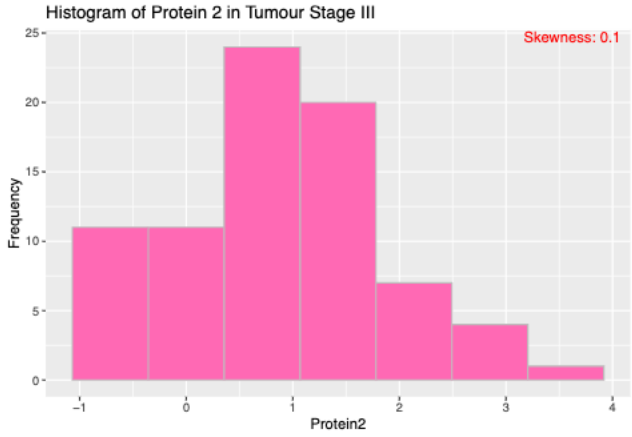



```
N3 <- nrow(stage3)
num_bins_3 <- ceiling(1 + log2(N1))
skewness_value <- skewness(stage3$Protein1, na.rm = TRUE)
age_hist <- ggplot(stage3, aes(x = Protein1)) +
```

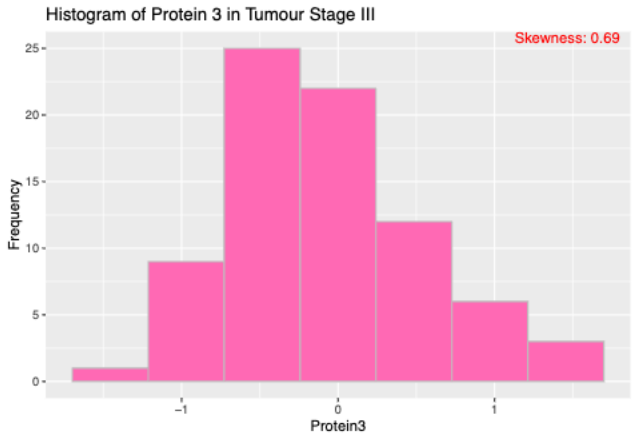
```
  geom_histogram(bins = num_bins_3, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 1 in Tumour Stage III", x = "Protein1", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



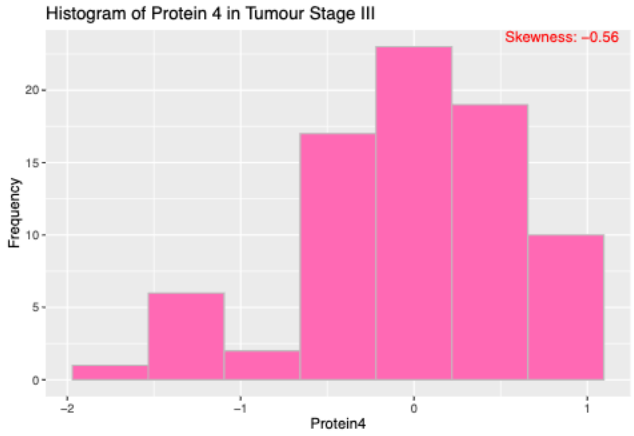

```
skewness_value <- skewness(stage3$Protein2, na.rm = TRUE)
age_hist <- ggplot(stage3, aes(x = Protein2)) +
  geom_histogram(bins = num_bins_3, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 2 in Tumour Stage III", x = "Protein2", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```



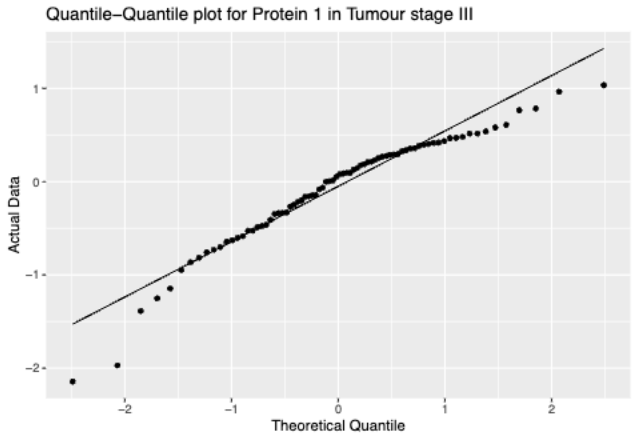

```
skewness_value <- skewness(stage3$Protein3, na.rm = TRUE)
age_hist <- ggplot(stage3, aes(x = Protein3)) +
  geom_histogram(bins = num_bins_3, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 3 in Tumour Stage III", x = "Protein3", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```




```
skewness_value <- skewness(stage3$Protein4, na.rm = TRUE)
age_hist <- ggplot(stage3, aes(x = Protein4)) +
  geom_histogram(bins = num_bins_3, fill = "hotpink", color = "grey") +
  labs(title = "Histogram of Protein 4 in Tumour Stage III", x = "Protein4", y = "Frequency")
age_hist +
  annotate("text", x = Inf, y = Inf, label = paste("Skewness:", round(skewness_value, 2)),
           hjust = 1.1, vjust = 1.1, color = "red")
```

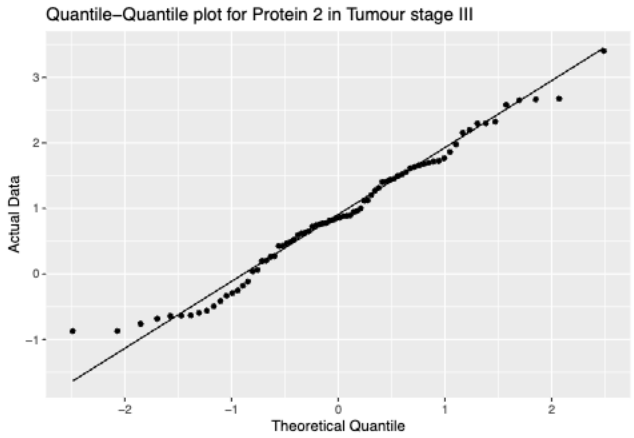



```
qq9 <- ggplot(stage3, aes(sample=Protein1)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantile
qq9
```

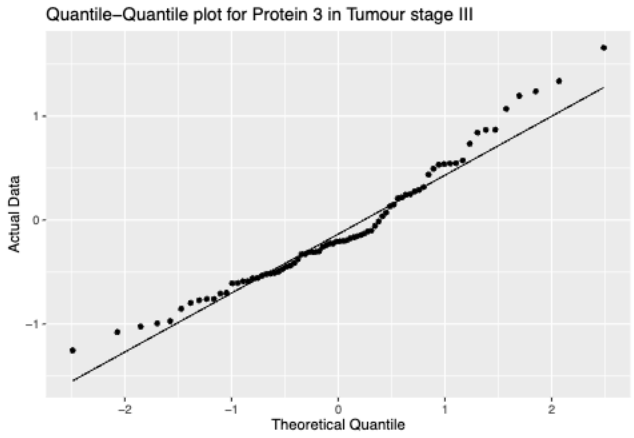



```
qq10 <- ggplot(stage3, aes(sample=Protein2)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantil
```

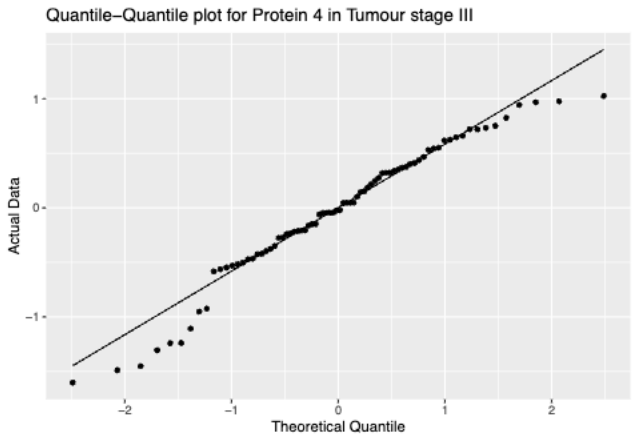
```
qq10
```




```
qq11 <- ggplot(stage3, aes(sample=Protein3)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantil
qq11
```




```
qq12 <- ggplot(stage3, aes(sample=Protein4)) + stat_qq() + stat_qq_line() + labs(x="Theoretical Quantil
qq12
```



From the above plots, we can see that there is no protein whose distribution is normal across all the tumour stages. But in case of protein 2, there is approximately same shape of distribution across all the tumour stages. Hence, we will perform Kruskal's test on it. The null hypothesis H_0 will be that the median values of protein 2 across all the tumour stages is equal and the alternate hypothesis H_1 will be that one of the

median values differ. We will use the same significance value.

```
#load library
library(FSA)
```

```
## ## FSA v0.9.5. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
##
## Attaching package: 'FSA'
## The following object is masked from 'package:psych':
##
##     headtail
```

```
kruskal.test(Protein2 ~ Tumour_Stage, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Protein2 by Tumour_Stage
## Kruskal-Wallis chi-squared = 0.93302, df = 2, p-value = 0.6272
```

```
dunnTest(Protein2 ~ Tumour_Stage, data = data,
         method="bonferroni")
```

```
## Warning: Tumour_Stage was coerced to a factor.
## Dunn (1964) Kruskal-Wallis multiple comparison
##   p-values adjusted with the Bonferroni method.
##   Comparison         Z   P.unadj P.adj
## 1     I - II 0.2467887 0.8050718     1
## 2    I - III 0.8710307 0.3837374     1
## 3   II - III 0.8302929 0.4063732     1
```

From the above, we can see that the p-value for kruskal's test is greater than the significance value, we fail to reject H_0. There is no significant evidence to say that the median value of protein 2 across all the tumour stages is different. To further evaluate this, we will do post-hoc analysis to see which pairs had similar or different medians. Using bonferroni correction which is very conservative, we will get p-values of all the pairs greater than α. Hence, we can surely say that the p-values will be greater than corrected significance value α^*. Hence, there are no pairs whose medians are different from each other. That means, there are no two tumour stages whose median value of protein 2 is significantly different.
Since, the requirements for ANOVA didn't meet, there is no point in using it on the protein data. We tried to do box-cox transformation and then perform ANOVA, but there are some negative values for proteins which prevented us from doing so. Just for a sample, below is ANOVA on one of the proteins.

```
fit2 <- aov(data$Protein1~data$Tumour_Stage)
summary(fit2)
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## data$Tumour_Stage   2   0.31  0.1571   0.524  0.593
## Residuals         318  95.32  0.2998
```

```
ScheffeTest(fit2)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
```

```
## $`data$Tumour_Stage`
##               diff     lwr.ci    upr.ci   pval
## II-I   -0.02491467 -0.2241217 0.1742924 0.9538
## III-I  -0.08823511 -0.3183779 0.1419077 0.6415
## III-II -0.06332044 -0.2455446 0.1189037 0.6944
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value is very high for ANOVA model indicating no evidence to say that the mean value of protein 1 is different across tumour stages. Using Scheffe post-hoc analysis, we can see that the p-values are very high for different pairs indicating no pair of tumour stages have different mean values of protein 1. These results are also not accurate because the assumption for ANOVA are not met.

2.6 Inference about correlation

We will now try to see if there is correlation between different proteins. For all of the tests performed below, the null hypothesis H_0 will be that there is zero correlation between the two proteins and the alternative hypothesis H_1 will be that there is a non-zero correlation between the two proteins. As the sample size is greater than 10, we can perform the correlation test with $\alpha = 0.05$ using spearman method because there maybe some outliers in the data.

```
cor.test(data$Protein1,data$Protein2,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein1 and data$Protein2
## S = 4231571, p-value = 2.61e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2323876
```

```
cor.test(data$Protein1,data$Protein3,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein1 and data$Protein3
## S = 6167459, p-value = 0.03338
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.118785
```

```
cor.test(data$Protein1,data$Protein4,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein1 and data$Protein4
## S = 4285894, p-value = 5.77e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2225333
```

```r
cor.test(data$Protein2,data$Protein3,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein2 and data$Protein3
## S = 7407286, p-value = 2.492e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3436912
```

```r
cor.test(data$Protein2,data$Protein4,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein2 and data$Protein4
## S = 4997149, p-value = 0.09442
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.09351074
```

```r
cor.test(data$Protein3,data$Protein4,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Protein3 and data$Protein4
## S = 5093592, p-value = 0.1743
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.07601585
```

Protein Pair	p-value
I-II	2.61e-05
I-III	0.03338
I-IV	5.77e-05
II-III	2.492e-10
II-IV	0.09442
III-IV	0.1743

From the above table, we can see the p-values for pairs I-II, I-III, I-IV and II-III are less than the significance value. Hence, there is significance evidence to say that there is non-zero correlation between the aforementioned pairs of proteins. Now, we will see if there is any correlation between the patients age with different proteins. H_0 will be that the age and protein have zero correlation and H_1 will be that the age and proteins have a non-zero correlation.

```r
cor.test(data$Age,data$Protein1,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
```

```
## data:  data$Age and data$Protein1
## S = 5683062, p-value = 0.581
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.03091481
```

```
cor.test(data$Age,data$Protein2,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Age and data$Protein2
## S = 5557688, p-value = 0.884
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## -0.008171688
```

```
cor.test(data$Age,data$Protein3,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Age and data$Protein3
## S = 5721578, p-value = 0.4986
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.0379016
```

```
cor.test(data$Age,data$Protein4,method="spearman",exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Age and data$Protein4
## S = 5017089, p-value = 0.1079
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.08989366
```

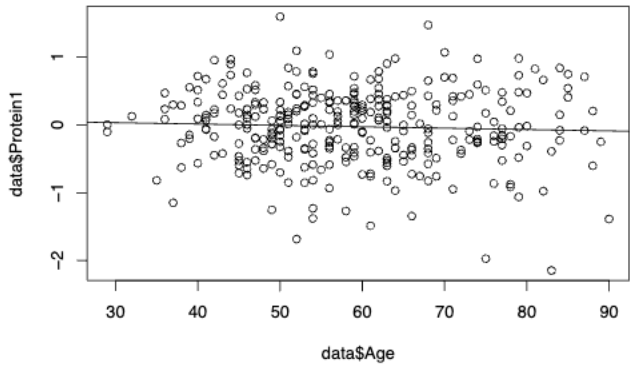
Protein Pair	p-value
age-I	0.581
age-II	0.884
age-III	0.4986
age-IV	0.4986

From the above table, we can see that there is no p-value which is less than the significance value. Hence, we can say that there is no significant evidence to say that there is non-zero correlation between age and any of the protein levels.

2.6 Regression Analysis

In this section, we intend to fit a regression model on age to predict different protein values. From the previous section, we got to know that there is no significant evidence to say that there is non-zero correlation between age and any protein values. Hence, performing regression fitting won't be much helpful and the predicted line won't be a proper fit. To visualize this, we have performed regression fitting for age to predict all the four proteins.

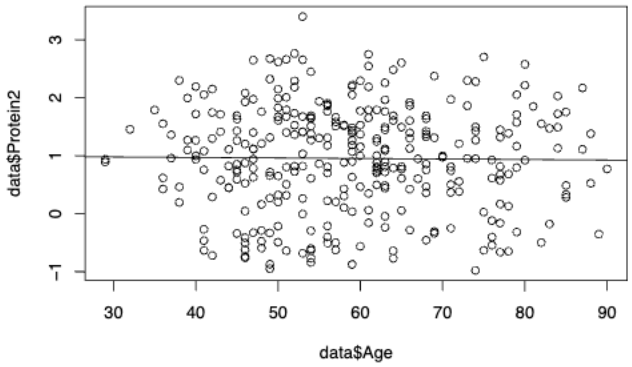
```
model <- lm(data$Protein1~data$Age)
plot(data$Age,data$Protein1)
abline(model)
```




```
summary(model)
```

```
##
## Call:
## lm(formula = data$Protein1 ~ data$Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06916 -0.32996  0.03109  0.34921  1.60207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.093010   0.142520   0.653    0.514
## data$Age    -0.002030   0.002365  -0.858    0.391
##
## Residual standard error: 0.5469 on 319 degrees of freedom
## Multiple R-squared:  0.002304,   Adjusted R-squared:  -0.0008234
## F-statistic: 0.7367 on 1 and 319 DF,  p-value: 0.3914
```

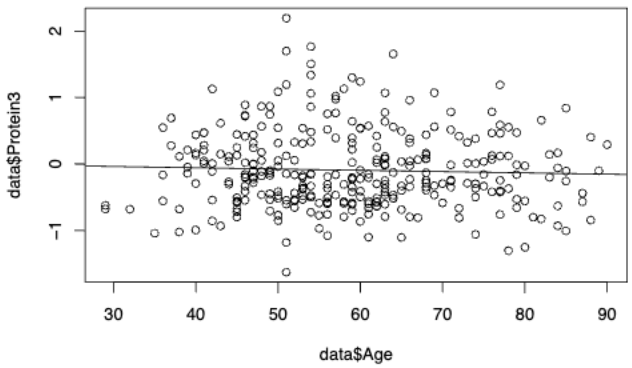
```
model1 <- lm(data$Protein2~data$Age)
plot(data$Age,data$Protein2)
abline(model1)
```




```
summary(model1)
```

```
##
## Call:
## lm(formula = data$Protein2 ~ data$Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91997 -0.57504  0.04673  0.66641  2.44246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.006432   0.237395   4.239 2.94e-05 ***
## data$Age    -0.000881   0.003939  -0.224    0.823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.911 on 319 degrees of freedom
## Multiple R-squared:  0.0001568,  Adjusted R-squared:  -0.002977
## F-statistic: 0.05003 on 1 and 319 DF,  p-value: 0.8231
```

```
model2 <- lm(data$Protein3~data$Age)
plot(data$Age,data$Protein3)
abline(model2)
```

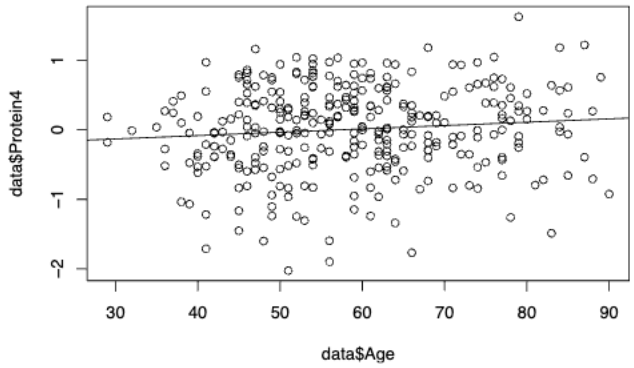



```
summary(model2)
```

```
##
## Call:
## lm(formula = data$Protein3 ~ data$Age)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5493 -0.4224 -0.1049  0.3403  2.2715
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.020192   0.153324   0.132    0.895
## data$Age    -0.001927   0.002544  -0.758    0.449
##
## Residual standard error: 0.5884 on 319 degrees of freedom
## Multiple R-squared:  0.001796,   Adjusted R-squared:  -0.001333
## F-statistic: 0.574 on 1 and 319 DF,  p-value: 0.4492
```

```
model3 <- lm(data$Protein4~data$Age)
plot(data$Age,data$Protein4)
abline(model3)
```



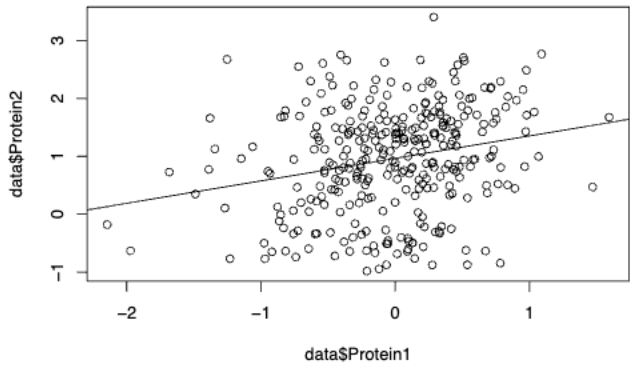

```
summary(model3)
```

```
##
## Call:
## lm(formula = data$Protein4 ~ data$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9964 -0.3787  0.0375  0.4358  1.5241
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.274776   0.161654  -1.700   0.0901 .
## data$Age     0.004817   0.002682   1.796   0.0734 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6203 on 319 degrees of freedom
## Multiple R-squared:  0.01001,    Adjusted R-squared:  0.006907
## F-statistic: 3.226 on 1 and 319 DF,  p-value: 0.07345
```

From all the above plots, we can see that the R^2 value for all the models is very less indicating the fit is bad. Moreover , the p-values for the age attribute in all the models is high indicating that age is not a good attribute to predict the protein levels.

Now, as we saw that there are some pairs of proteins whose correlation is non-zero, we will try to perform regression on those pairs.

```r
model4 <- lm(data$Protein2~data$Protein1)
plot(data$Protein1,data$Protein2)
abline(model4)
```

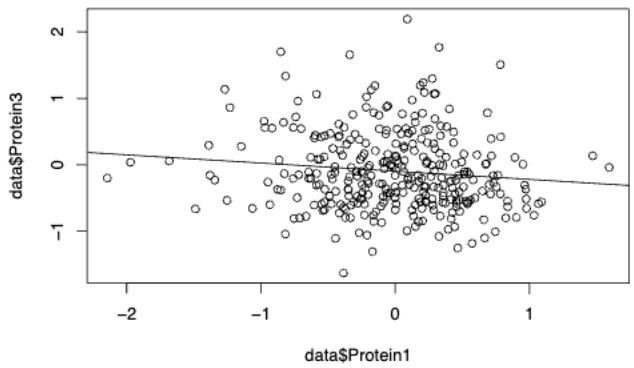



```r
summary(model4)
```

```
##
## Call:
## lm(formula = data$Protein2 ~ data$Protein1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11023 -0.58948  0.08096  0.61384  2.32593
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.96484    0.04951  19.489  < 2e-16 ***
## data$Protein1  0.38814    0.09059   4.285 2.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8859 on 319 degrees of freedom
## Multiple R-squared:  0.05441,    Adjusted R-squared:  0.05145
## F-statistic: 18.36 on 1 and 319 DF,  p-value: 2.428e-05
```

From the above regression fit, we can write Protein 2 = 0.96484 + 0.38814 * Protein 1. The R^2 value is still low but we can at least say they have the above relationship.

```r
model5 <- lm(data$Protein3~data$Protein1)
plot(data$Protein1,data$Protein3)
abline(model5)
```

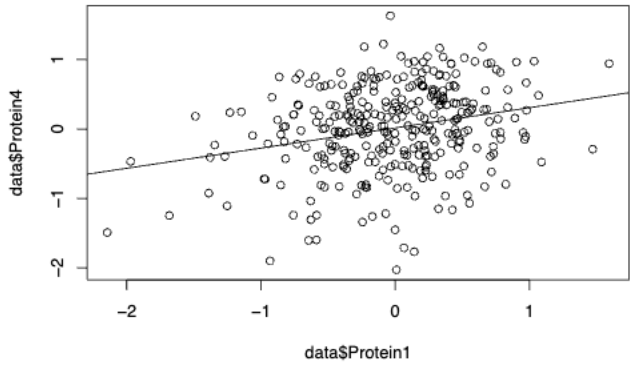


```
summary(model5)
```

```
##
## Call:
## lm(formula = data$Protein3 ~ data$Protein1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57819 -0.40272 -0.09244  0.35973  2.30107
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.09654    0.03269  -2.953  0.00338 **
## data$Protein1 -0.12304    0.05982  -2.057  0.04052 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.585 on 319 degrees of freedom
## Multiple R-squared:  0.01309,    Adjusted R-squared:  0.009994
## F-statistic:  4.23 on 1 and 319 DF,  p-value: 0.04052
```

From the above regression fit, we can write Protein 3 = -0.09654 - 0.12304 * Protein 1. The R^2 value is still low but we can at least say they have the above relationship.

```
model6 <- lm(data$Protein4~data$Protein1)
plot(data$Protein1,data$Protein4)
abline(model6)
```



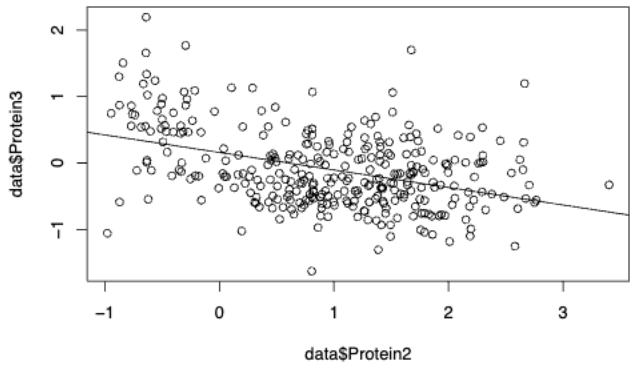

```
summary(model6)
```

```
##
## Call:
## lm(formula = data$Protein4 ~ data$Protein1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04447 -0.39295  0.05468  0.43092  1.62392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.01651    0.03368   0.490    0.624
## data$Protein1  0.29086    0.06164   4.719 3.55e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6028 on 319 degrees of freedom
## Multiple R-squared:  0.06525,    Adjusted R-squared:  0.06232
## F-statistic: 22.27 on 1 and 319 DF,  p-value: 3.554e-06
```

From the above regression fit, we can write Protein 4 = 0.01651 + 0.29086 * Protein 1. The R^2 value is still low but we can at least say they have the above relationship.

```
model7 <- lm(data$Protein3~data$Protein2)
plot(data$Protein2,data$Protein3)
abline(model7)
```




```
summary(model7)
```

```
##
## Call:
## lm(formula = data$Protein3 ~ data$Protein2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57328 -0.40614 -0.05814  0.36342  1.98545
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.15848    0.04354   3.640 0.000318 ***
## data$Protein2 -0.26374    0.03304  -7.982 2.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5377 on 319 degrees of freedom
## Multiple R-squared:  0.1665, Adjusted R-squared:  0.1639
## F-statistic: 63.71 on 1 and 319 DF,  p-value: 2.624e-14
```

From the above regression fit, we can write Protein 3 = 0.15848 - 0.263746 * Protein 2. The R^2 value is better than the above models and hence this is by far the best regression model because the p-value for Protein 2 is also very low indicating it can be very useful to predict Protein 3 levels.

##Conclusion

In conclusion, our examination of the BRCA dataset through extensive statistical analysis has yielded valuable insights into the intricate landscape of breast cancer characteristics and outcomes. Through meticulous dissection of a dataset encompassing a diverse array of variables, ranging from surgical histories and protein expression profiles to cancer stages based on age, we have significantly enhanced our comprehension of the factors pivotal to breast cancer prognosis and management. Our analytical approach involved conducting

hypothesis tests, independence tests, various inference tests, ANOVA, linear regression, and logistic regression on variables such as protein levels, age, and patient status, to name a few. The respective subsections above explain the individual findings derived from these analyses.