STAT7174 Applications of Computational Statistics: Assignment 1

Introduction

This assignment is part of the UQ course STAT7174.
The assignment is due at **3:00pm** on **Monday, April 8th 2024**

This assignment is designed to familiarise you with R, markdown, data structures, simple data visualisation and performing hypothesis testing. Your assignment should be submitted through Blackboard in the form of a HTML or PDF document. As you will need to submit a document that displays your R code along with comments and interpretation, it is recommended you use Rmarkdown (.Rmd) to create it. When including a figure and/or table, refer to it in your comments (you may like to include a caption for easier referencing). Marks will be allocated for correctness, clarity and completeness.

Here is an example question and answer with sufficient comments:

Mock Q. *What is a+b given a=1 and b=2?*

We are asked to find the sum of a and b.

```
a <- 1  # Note how variables are assigned instead of just printing 1+2
b <- 2
(c <- a + b) # Note that round brackets around an assignment automatically prints the o
utput of the assignment.
```

```
## [1] 3
```

```
c # Or by just typing the assigned variable name, you can also output its value.
```

```
## [1] 3
```

In conclusion, we found that a + b is 3.

Part 1: Data structures (5 marks)

Q1. *Create and print a data frame called "my_experiment" with 3 columns: Sample, Group and Value. Use the following details to populate your data frame:*

- *Sample* should contain sample numbers from 21 to 30;
- *Group* should have 'diseased' and 'healthy', 5 times each (not alternating).
- *Value* should contain a sample of normally distributed numbers with mean=5 and standard deviation of (sd=6) (after setting the seed at 12).

```
# Set seed for reproducibility
set.seed(12)

# Create the data frame
my_experiment <- data.frame(
  Sample = seq(21, 30),
  Group = rep(c("diseased", "healthy"), each=5),
  Value = rnorm(10, mean=5, sd=6)
)

# Print the data frame
print(my_experiment)
```

```
##    Sample    Group      Value
## 1      21 diseased -3.8834056
## 2      22 diseased 14.4630168
## 3      23 diseased -0.7404669
## 4      24 diseased -0.5200315
## 5      25 diseased -6.9858526
## 6      26  healthy  3.3662237
## 7      27  healthy  3.1079077
## 8      28  healthy  1.2304686
## 9      29  healthy  4.3612167
## 10     30  healthy  7.5680888
```

Q2. *Calculate the minimum, mean and maximum of the Value column*

```
# Minimum
min_value <- min(my_experiment$Value)

# Mean
mean_value <- mean(my_experiment$Value)

# Maximum
max_value <- max(my_experiment$Value)

# Printing the results
cat("Minimum of Value:", min_value, "\nMean of Value:", mean_value, "\nMaximum of Valu
e:", max_value)
```

```
## Minimum of Value: -6.985853
## Mean of Value: 2.196717
## Maximum of Value: 14.46302
```

Q3. *Calculate the minimum, mean and maximum of the Value column for each group*

```
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
```

```
# Use tidyverse dplyr to compute min, mean, and max, and organize them into separate co
lumns
stats <- my_experiment %>%
  group_by(Group) %>%
  summarise(Min_Value = min(Value),
            Mean_Value = mean(Value),
            Max_Value = max(Value))
print(stats)
```

```
## # A tibble: 2 × 4
##   Group    Min_Value Mean_Value Max_Value
##   <chr>        <dbl>      <dbl>     <dbl>
## 1 diseased     -6.99      0.467     14.5
## 2 healthy       1.23      3.93       7.57
```

Q4. *Create a new data frame "my_new_experiment" that contains only those samples in "my_experiment" with values > 5. What are the number of samples in each group in "my_new_experiment"?*

```
# Filtering samples where Value > 5 and printing the filtered data frame
my_new_experiment <- my_experiment %>%
  filter(Value > 5)
print(my_new_experiment)
```

```
##   Sample    Group     Value
## 1     22 diseased 14.463017
## 2     30  healthy  7.568089
```

```
# Counting the number of samples in each group in the new data frame
sample_counts <- table(my_new_experiment$Group)
print(sample_counts)
```

```
##
## diseased  healthy
##        1        1
```

Part 2: Data description and visualization (8 marks)

The questions in this part of the assignment require you to download the file “Melanoma_Subtypes.txt”, which comprises a subset of genomic data published in Newell et al. 2022 Cancer Discovery. In brief, this study is the largest whole-genome sequencing study of melanoma to date, with 570 tumors profiled. Melanoma is a cancer of melanocytes, with multiple subtypes based on body site location. Cutaneous melanoma is associated with skin exposed to ultraviolet radiation (e.g. from the sun); uveal melanoma occurs in the eyes; mucosal melanoma occurs in internal mucous membranes that line many tracts of the body, e.g. mouth and nose; and acral melanoma occurs on the palms, soles, and nail beds.

This dataset has been pre-processed to select the relevant information for the assignment, including clinical patient information, melanoma subtypes and the tumours’ mutational burden, which is calculated as the number of mutations per megabase sequenced.

NOTE Since this is a real dataset, you need to be wary of missing data. Make sure to check for NA values (using the *is.na()* function) and filtering them out at the appropriate steps. In your answers to questions 4 and 5 (Part 2), make note of how many patients were filtered out because they contained missing values.

Q1. *Load the dataset into R and ensure it is a data frame. What are the column names?*

```
# Read the dataset, ensuring it's treated as a data frame
melanoma_data <- read.table("Melanoma_Subtypes.txt", header = TRUE, sep = "\t", strings
AsFactors = FALSE)
colnames(melanoma_data) # Display the column names.
```

```
## [1] "Donor"                 "Gender"                "Age"
## [4] "Subtype"               "Mutations.per.megabase"
```

Q2. *Indicate the type of each variable (categorical, ordinal, continuous or discrete).*

```
# Inspect the structure of the dataset to classify variable types
str(melanoma_data)
```

```
## 'data.frame':    230 obs. of  5 variables:
##  $ Donor                 : chr  "MELA_0788" "MELA_0874" "MELA_0793" "MELA_0794" ...
##  $ Gender                : chr  "male" "female" "male" "female" ...
##  $ Age                   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Subtype               : chr  "Uveal" "Mucosal" "Uveal" "Uveal" ...
##  $ Mutations.per.megabase: num  0.805 1.44 0.622 0.45 0.435 0.144 1.8 2.12 0.475 0.8
74 ...
```

Q3. *How many patients with Cutaneous melanoma are in this dataset?*

```
# Filter the dataset for patients with Cutaneous melanoma and counting them
cutaneous_melanoma <- subset(melanoma_data, Subtype == "Cutaneous")
nrow(cutaneous_melanoma)
```

```
## [1] 69
```

Q4. *Illustrate the distribution of age in melanoma subtype patients using a plot of your choice. Interpret the plot. Is it normally distributed? (justify your answer)*

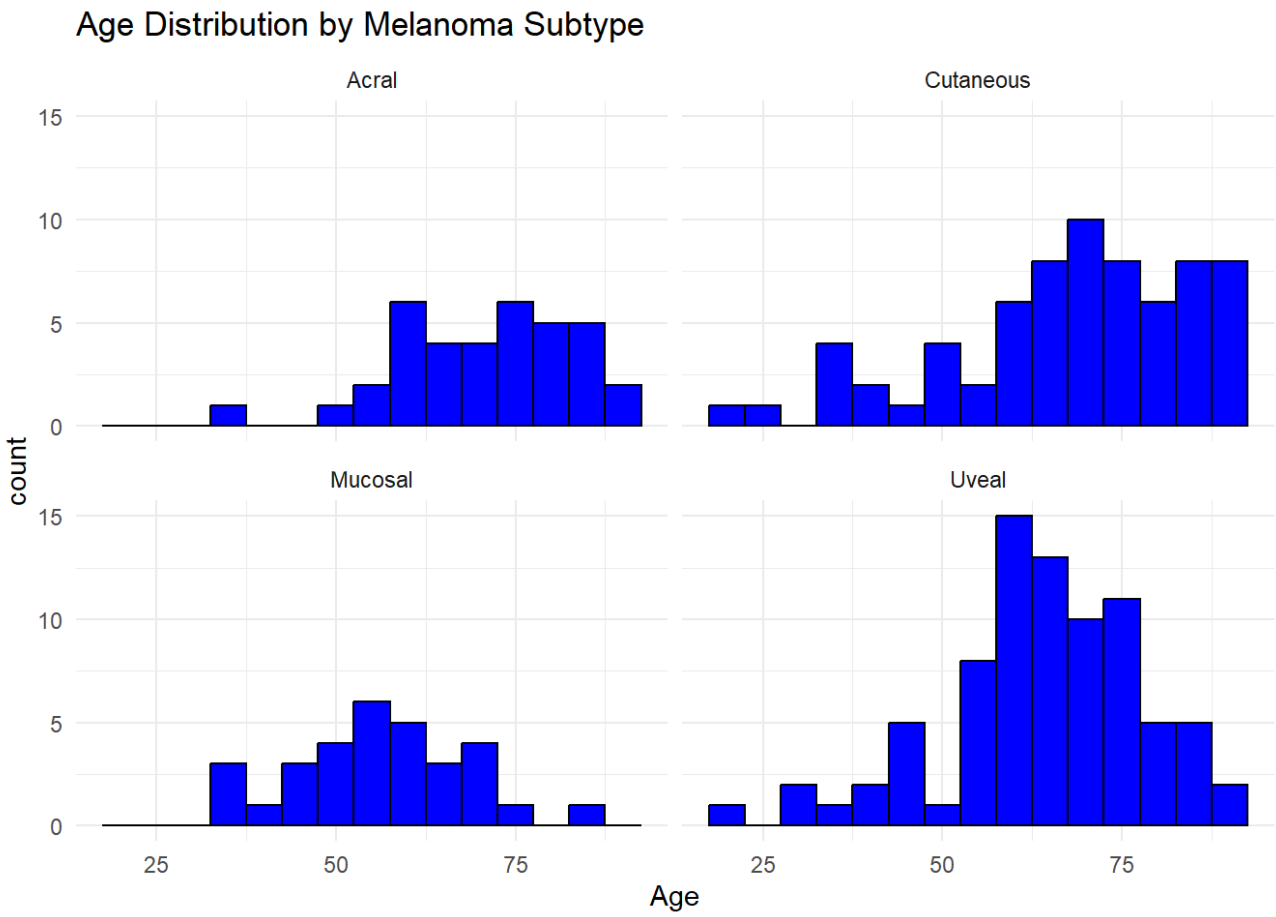
```
library(dplyr)
library(ggplot2) # A versatile plotting package for R
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(tidyr) # For handling missing data with drop_na
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
# Filter out NA values in 'Age' for accurate visualization
melanoma_data_filtered <- drop_na(melanoma_data, Age)

# Create a histogram to visualize age distribution across melanoma subtypes
ggplot(melanoma_data_filtered, aes(x=Age)) +
  geom_histogram(binwidth=5, fill="blue", color="black") + #Bins of width 5. The fill c
olor of the bars is set to blue, and the outline color of the bars is set to black.
  facet_wrap(~Subtype) + #Separate the data into different panels based on the Subtype
variable, allowing for comparison across the different melanoma subtypes.
  theme_minimal() + #Simplifies the background for a cleaner look.
  ggtitle("Age Distribution by Melanoma Subtype")
```




```
# Perform the Shapiro-Wilk test for normality on the Age distribution for each subtype
shapiro_tests <- melanoma_data_filtered %>%
  group_by(Subtype) %>%
  summarise(p_value = shapiro.test(Age)$p.value, #For each group, performs the Shapiro-Wilk test on the Age variable and extracts the p-value.
            sample_size = n()) %>%  #Counts the number of observations in each group to report the sample size.
  mutate(Normality_Rejected = p_value < 0.05) #Adds a new column that indicates whether the null hypothesis of normality is rejected for each group, based on the p-value.
print(shapiro_tests)
```

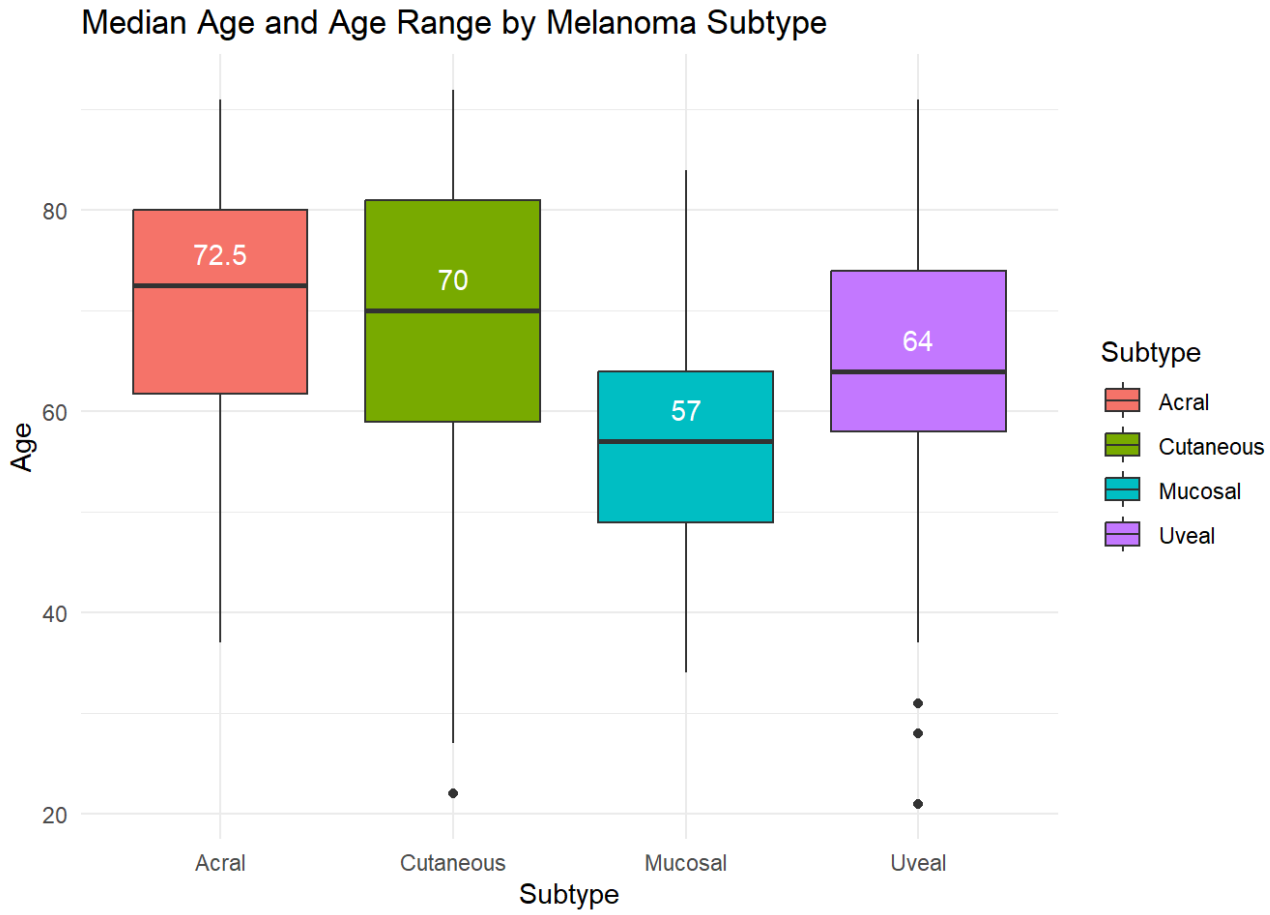
```
## # A tibble: 4 × 4
##   Subtype   p_value sample_size Normality_Rejected
##   <chr>       <dbl>       <int> <lgl>
## 1 Acral     0.330            36 FALSE
## 2 Cutaneous 0.00200          69 TRUE
## 3 Mucosal   0.805            31 FALSE
## 4 Uveal     0.0647           81 FALSE
```

```
#The Shapiro-Wilk test results indicate that for Acral, Mucosal, and Uveal melanoma pat
ients, there is insufficient evidence to reject the null hypothesis of normality as the
ir p-values are above the commonly accepted alpha level of 0.05. This suggests that the
age distributions for these subtypes do not significantly deviate from normality. Howev
er, for Cutaneous melanoma, the p-value is below 0.05, thus rejecting the null hypothes
is and conclude that the age distribution for this subgroup does not follow a normal di
stribution."
```

Q5. *Comment on the median age and age range for the individual melanoma subtypes using a plot of your choice.*

```
# Using ggplot2 to create boxplots, visualizing the median age and age range
ggplot(melanoma_data_filtered, aes(x=Subtype, y=Age, fill=Subtype)) +
  geom_boxplot() +  # Add a boxplot to visualize the distribution of 'Age' for each 'Subtype'.
  stat_summary(fun = median, geom = "text", aes(label = ..y..), vjust = -1, color = "white") +  # Overlay the median age on each boxplot.
  theme_minimal() +  # Use a minimal theme for a clean plot appearance.
  ggtitle("Median Age and Age Range by Melanoma Subtype")
```

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




```
#Acral and Cutaneous melanomas have higher median ages, pointing towards a greater prevalence in older populations, with Cutaneous melanoma showing the broadest age range, indicating its potential influence by external factors like UV exposure. Mucosal melanoma affect individuals at a younger age, and its age distribution is less variable. Uveal melanoma is also more common in older adults with a wide age range."
```

Q6. *Summarise the number of female and male patients in the melanoma subtypes.*

```
# Summary by gender and subtype
gender_summary <- table(melanoma_data$Subtype, melanoma_data$Gender)
print(gender_summary)
```

```
##
##             female male
##   Acral         17   19
##   Cutaneous     23   46
##   Mucosal       20   14
##   Uveal         48   43
```

Q7. *Is the proportion of males and females the same in the individual melanoma subtypes?*

```
# Perform chi-square test
chi_square_results <- chisq.test(gender_summary)
print(chi_square_results)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender_summary
## X-squared = 8.2899, df = 3, p-value = 0.04039
```

```
#Given that the p-value (0.04) is below the threshold of 0.05, we have sufficient evide
nce to reject the null hypothesis at a 5% significance level, indicating a statisticall
y significant difference in the distribution of genders among the melanoma subtypes. Th
is finding suggests that certain melanoma subtypes may be more prevalent in one gender
over the other, which could have implications for research tailored to these demographi
c differences."
```

Part 3: Comparing 2 groups (8 marks)

We hypothesise that Cutaneous melanomas on the skin are primarily due to sun exposure, causing DNA damage. We, therefore, would like to compare the mutations per megabase for sun-exposed Cutaneous vs non-sun-exposed Mucosal melanomas.

Q1. *Subset your melanoma data set, so it only contains Cutaneous and Mucosal melanomas.*

```
# Subset to include only Cutaneous and Mucosal melanomas
melanoma_subset <- melanoma_data %>%
  filter(Subtype %in% c("Cutaneous", "Mucosal"))
```

Q2. *State the null and alternative hypothesis. What are the assumptions of a t-test? Define the population parameters of interest either in words or mathematically. (no code required to answer this question)*

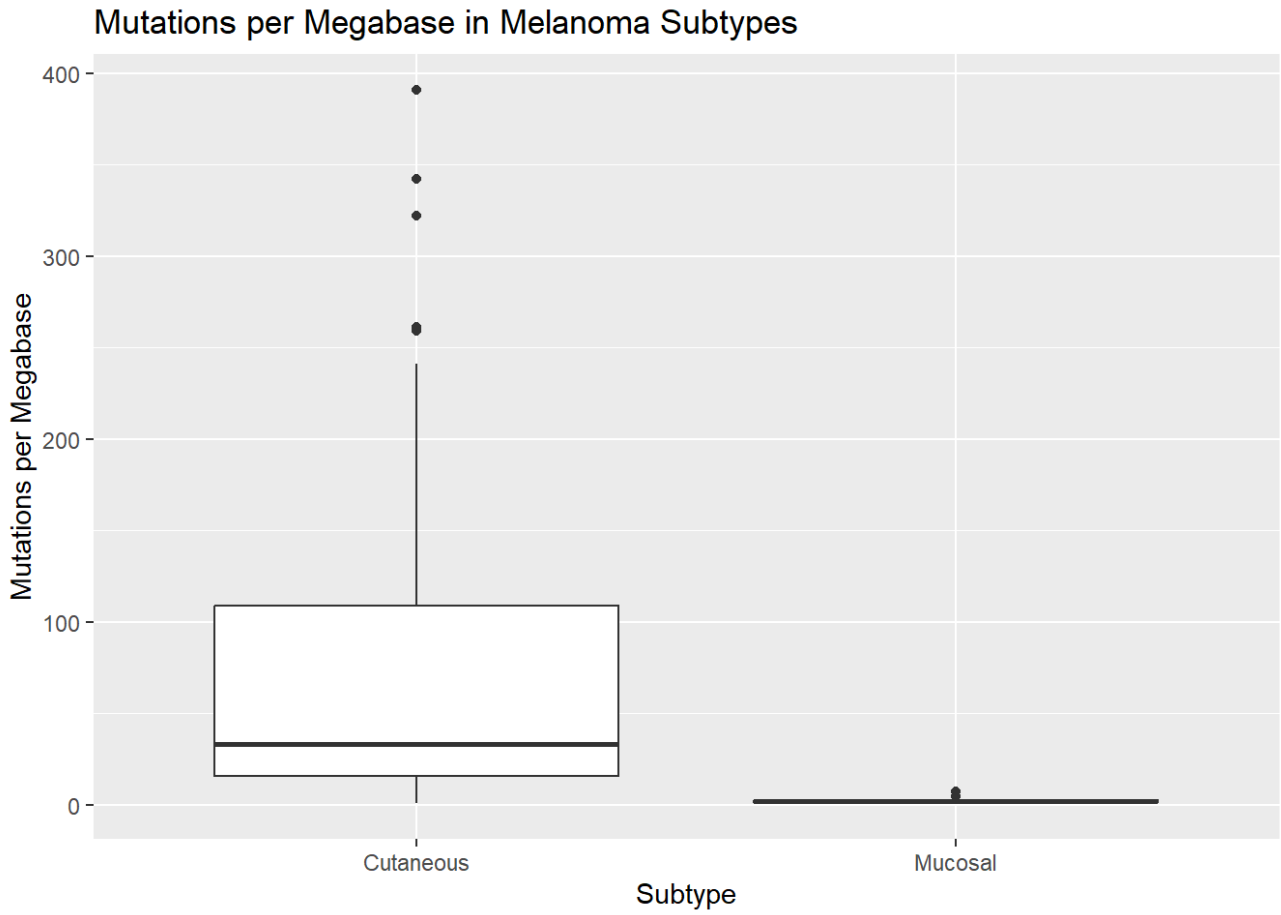
```
#Null Hypothesis (H0): The mean mutations per megabase for sun-exposed Cutaneous melano
mas are equal to that for non-sun-exposed Mucosal melanomas."
"Alternative Hypothesis (H1): The mean mutations per megabase for sun-exposed Cutaneous
melanomas are different from that for non-sun-exposed Mucosal melanomas."
```

```
## [1] "Alternative Hypothesis (H1): The mean mutations per megabase for sun-exposed Cu
taneous melanomas are different from that for non-sun-exposed Mucosal melanomas."
```

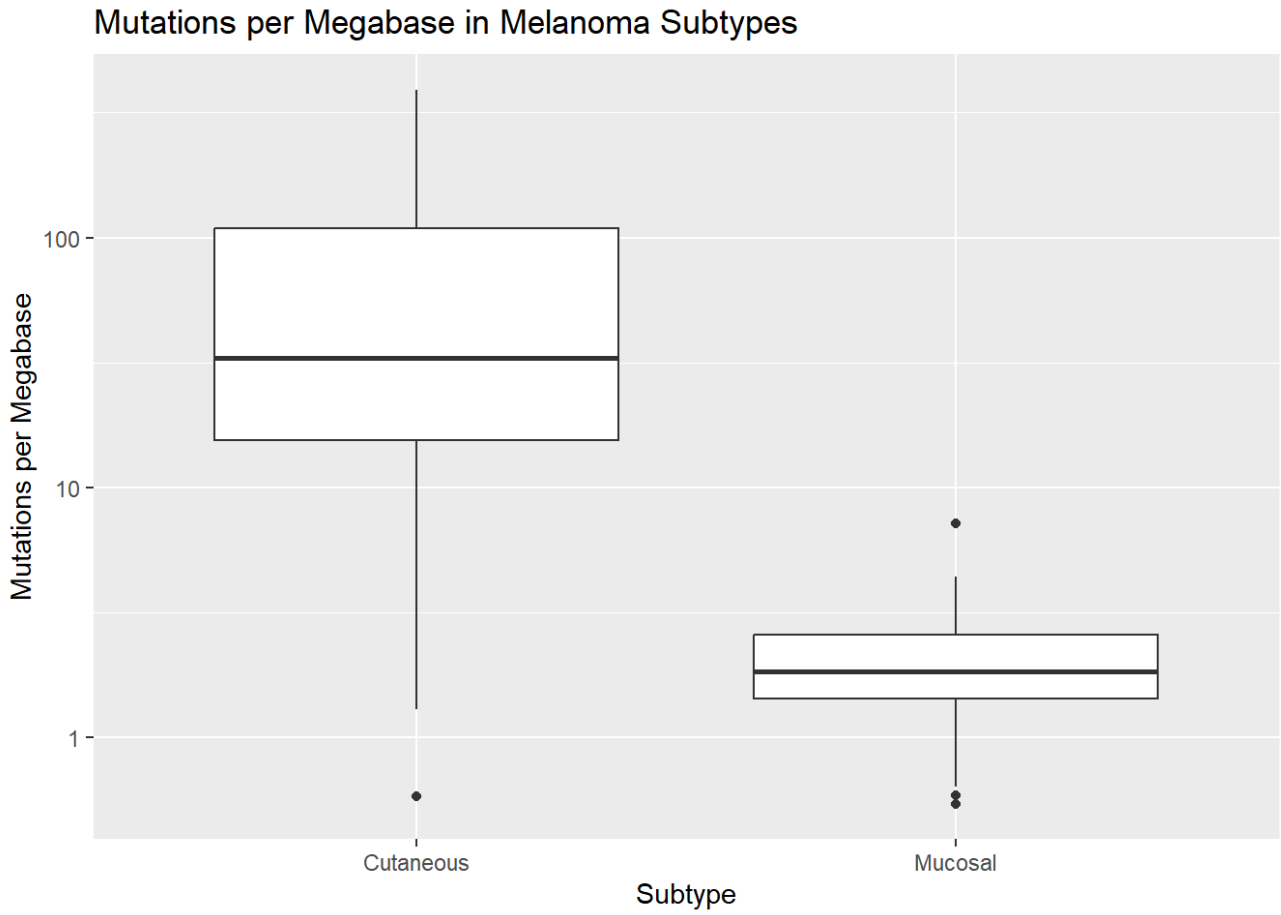
```
#For a t-test to be valid, certain assumptions about the data must be met:
#Data points are independent of each other.
#The distribution of the outcome variable (mutations per megabase) should be normal in
each group being compared.
#The variances in the two groups being compared should be equal."

#The population parameters of interest are the mean mutations per megabase in the Cutan
eous and Mucosal melanoma subtypes, represented as μCutaneous and μMucosal respectivel
y. The goal is to compare these means to investigate if there is a significant differen
ce in the mutational burden between melanomas that are typically exposed to the sun and
those that are not."
```

Q3. *Provide an appropriate plot to check the assumptions for a t-test for your melanoma data subset from Q1, and briefly explain why you believe the assumptions to be true or false.*

```
# Creating a boxplot
melanoma_subset %>%
  ggplot(aes(x = Subtype, y = Mutations.per.megabase)) +
  geom_boxplot() +
  ggtitle("Mutations per Megabase in Melanoma Subtypes") +
  xlab("Subtype") +
  ylab("Mutations per Megabase")
```




```
# Creating a boxplot (Log Scale)
melanoma_subset %>%
  ggplot(aes(x = Subtype, y = Mutations.per.megabase)) +
  geom_boxplot() +
  scale_y_log10() + # Apply log transformation
  ggtitle("Mutations per Megabase in Melanoma Subtypes") +
  xlab("Subtype") +
  ylab("Mutations per Megabase")
```




```
# Shapiro-Wilk test for Cutaneous melanomas
shapiro_cutaneous <- shapiro.test(melanoma_subset$Mutations.per.megabase[melanoma_subse
t$Subtype == "Cutaneous"])

# Shapiro-Wilk test for Mucosal melanomas
shapiro_mucosal <- shapiro.test(melanoma_subset$Mutations.per.megabase[melanoma_subset
$Subtype == "Mucosal"])

# Display the results
shapiro_cutaneous
```

```
##
##  Shapiro-Wilk normality test
##
## data:  melanoma_subset$Mutations.per.megabase[melanoma_subset$Subtype == "Cutaneou
s"]
## W = 0.77451, p-value = 6.35e-09
```

```
shapiro_mucosal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  melanoma_subset$Mutations.per.megabase[melanoma_subset$Subtype == "Mucosal"]
## W = 0.85253, p-value = 0.0003189
```

```
#Both subtypes have p-values well below 0.05. This indicates that the null hypothesis t
hat the data are normally distributed is rejected for both groups. Therefore, alternati
ve statistical approaches that do not assume normal distribution should be considered t
o compare the mutations per megabase between these two subtypes of melanoma accurately.

#Wilcoxon rank-sum test is an alternative approach that does not require the assumption
of normality or equal variances. This test would be suitable given that the normality a
ssumption is violated.
```

Q4. *Check if the variances are equal by performing an F-test (var.test), giving the probability value from the output and reporting your conclusion. Explain if you should perform a t-test or a Wilcox-test.*

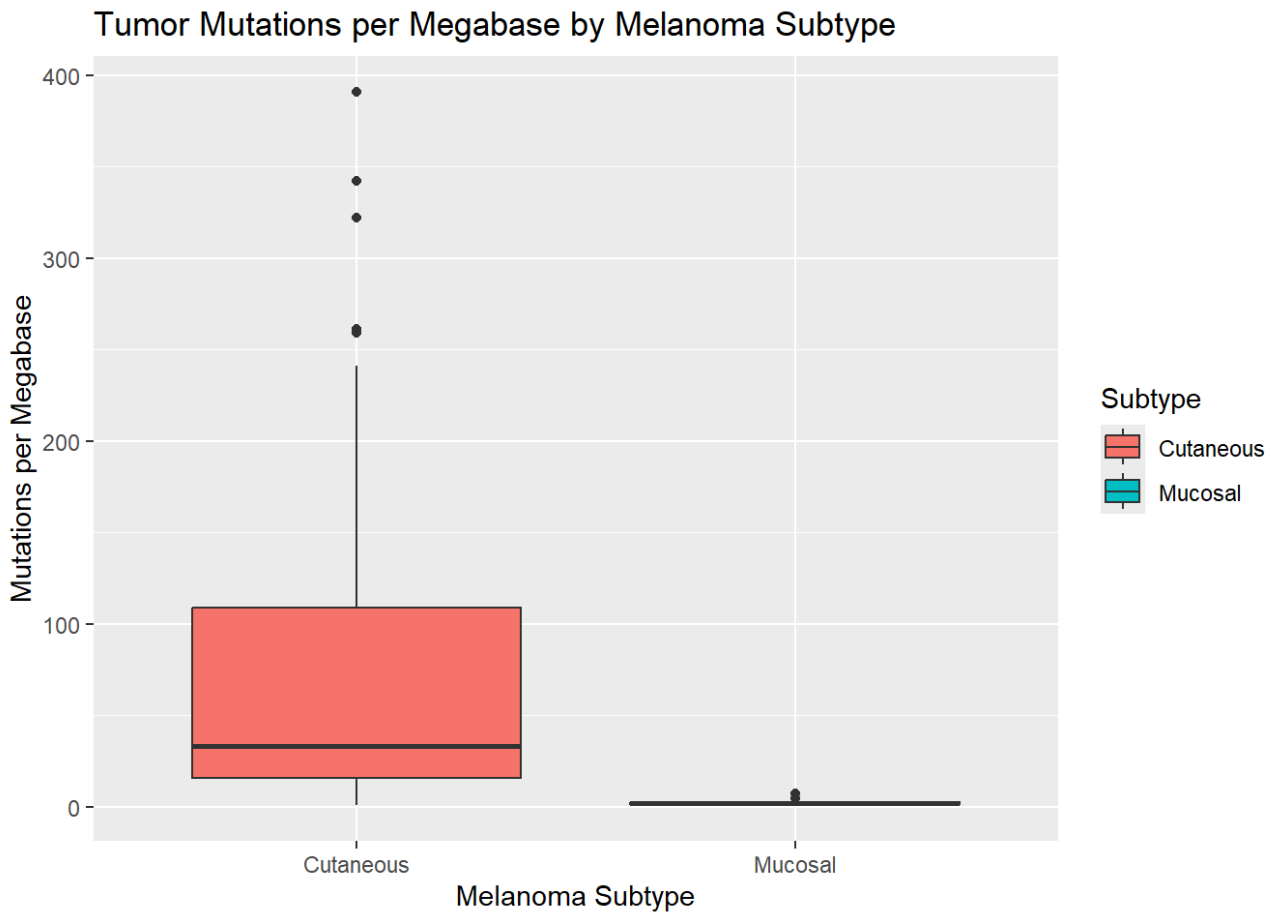
```
# Perform an F-test to check for equality of variances between Cutaneous and Mucosal me
lanomas
f_test_result <- var.test(Mutations.per.megabase ~ Subtype, data = melanoma_subset)
print(f_test_result)
```

```
##
##  F test to compare two variances
##
## data:  Mutations.per.megabase by Subtype
## F = 5098.1, num df = 68, denom df = 33, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2722.376 8972.509
## sample estimates:
## ratio of variances
##           5098.134
```

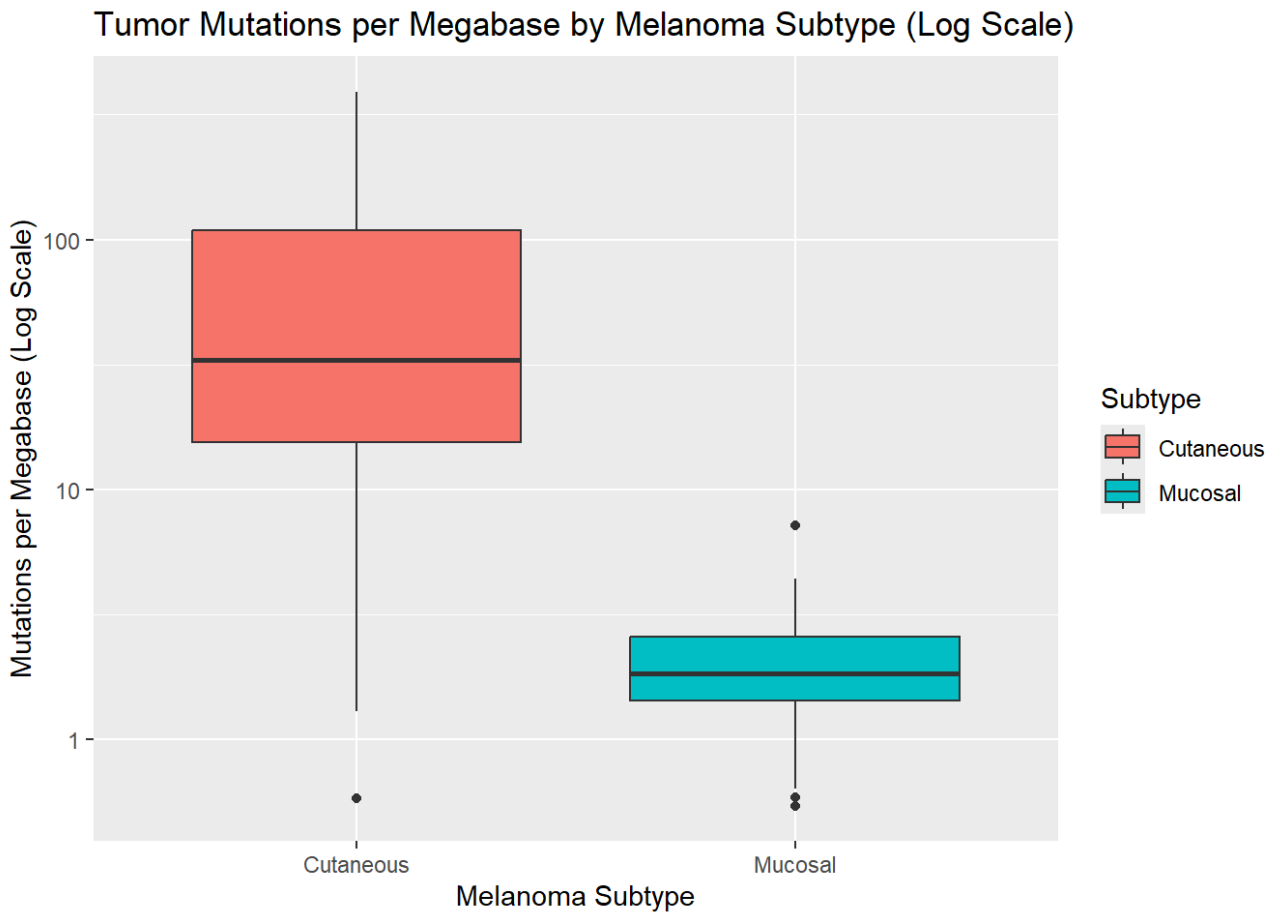
```
#The F-test results strongly suggest that there is a significant difference in the vari
ances of mutations per megabase between Cutaneous and Mucosal melanomas. Given the extr
emely small p-value and the ratio of variances being far from 1, the null hypothesis of
equal variances between the two groups is rejected.

#Wilcoxon rank-sum test is a non-parametric alternative that does not require the assum
ption of normality or equal variances. This test would be suitable given that the equal
variances assumption is violated.
```

Q5. *Visualise the tumour mutations per megabase for the melanoma subtypes as a boxplot. Calculate the mean and median mutations per megabase for each subtype.*

```
# Creating a boxplot
ggplot(melanoma_subset, aes(x = Subtype, y = Mutations.per.megabase, fill = Subtype)) +
  geom_boxplot() +
  ggtitle("Tumor Mutations per Megabase by Melanoma Subtype") +
  xlab("Melanoma Subtype") +
  ylab("Mutations per Megabase")
```




```
# Creating a boxplot (Log Scale)
ggplot(melanoma_subset, aes(x = Subtype, y = Mutations.per.megabase, fill = Subtype)) +
  geom_boxplot() +
  scale_y_log10() + # Apply log transformation
  ggtitle("Tumor Mutations per Megabase by Melanoma Subtype (Log Scale)") +
  xlab("Melanoma Subtype") +
  ylab("Mutations per Megabase (Log Scale)")
```




```
# Calculating mean and median
melanoma_subset %>%
  group_by(Subtype) %>%
  summarise(
    Mean_Mutations = mean(Mutations.per.megabase, na.rm = TRUE),
    Median_Mutations = median(Mutations.per.megabase, na.rm = TRUE)
  )
```

```
## # A tibble: 2 × 3
##   Subtype   Mean_Mutations Median_Mutations
##   <chr>              <dbl>            <dbl>
## 1 Cutaneous          79.4             32.9
## 2 Mucosal             2.19             1.84
```

Q6. *Perform the appropriate test to answer the biological question mentioned at the start of Part 3, giving the statistic value from the output and reporting your conclusion.*

```
# Perform the Wilcoxon rank-sum test
wilcox_test_result <- wilcox.test(Mutations.per.megabase ~ Subtype,
                                  data = melanoma_subset,
                                  alternative = "greater")

# Display the test result
print(wilcox_test_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Mutations.per.megabase by Subtype
## W = 2244.5, p-value = 2.932e-14
## alternative hypothesis: true location shift is greater than 0
```

```
#Given the p-value of 2.932e-14, which is significantly below 0.05, the null hypothesis is rejected in favor of the alternative. This means that there is statistically signifi cant evidence to suggest that the median mutations per megabase in the Cutaneous melano mas are greater than those in Mucosal melanomas.

#This significant difference clearly explains the impact of sun exposure on increasing mutations in Cutaneous melanomas, aligning with the hypothesis and suggesting further r esearch into preventative treatments.
```

Bonus point: This bonus point is assigned for formatting style and completeness of your report.

References

Newell, Felicity, et al. “Comparative genomics provides etiologic and biological insight into melanoma subtypes.” *Cancer discovery* 12.12 (2022): 2856-2879.