UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA
(FINAL PROJECT PAPER)

IZDELAVA SPLETNE APLIKACIJE ZA BIOSTATISTIKO ZA KVANTITATIVNO ANALIZO PODATKOV Z UPORABO R SHINY

(Creating Biostatistics Web Application for Quantitative Data Analysis Using R Shiny)

TANYA DENIZ TOLUAY

2022
ZAKLJUČNA NALOGA (FINAL PROJECT PAPER)
TOLUAY

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

Izdelava spletne aplikacije za biostatistiko za kvantitativno analizo podatkov z uporabo R Shiny

(Creating Biostatistics Web Application for Quantitative Data Analysis Using R Shiny)

Ime in priimek: Tanya Deniz Toluay
Študijski program: Bioinformatika
Mentor: Assist. Prof. Dr. Uroš Godnov

Koper, July 2022

Ključna dokumentacijska informacija

Ime in PRIIMEK: Tanya Deniz TOLUAY

Naslov zaključne naloge: Izdelava spletne aplikacije za biostatistiko za kvantitativno analizo podatkov z uporabo R Shiny

Kraj: Koper

Leto: 2022

Število listov: 49 Število slik: 15 Število tabel: 2

Število prilog: 3 Št. strani prilog: 12

Število referenc: 41

Mentor: Assist. Prof. Dr. Uroš Godnov

Ključne besede: R, R Shiny, spletna aplikacija za biostatistiko, kvantitativna analiza podatkov, hi-kvadrat test, t-test, ANOVA, MANOVA

Izvleček: Na raziskovalnih področjih se od raziskovalcev biologije in bioinformatike pričakuje, da uporabijo veliko različnih analiznih tehnik, pogosto z izvedbo cevovodov, pri čemer se zanašajo na več programske opreme. Podatki so običajno kvantitativni in so zbrani iz različnih laboratorijskih eksperimentov, raziskav ter opazovanj.

Priprava analiz je občutljiv postopek, ki zahteva skrb in pozornost, saj napaka v enem koraku lahko pokvari celotno raziskavo.

Ta diplomska naloga prikazuje pomen računalniških orodij za biostatistiko in izdelavo spletne aplikacije ponudi olajšanje pri statističnih analizah. Spletna aplikacija je zgrajena s pomočjo paketa R, Shiny, z namenom izvedbe kvantitativne analize podatkov, tako tabelarično kot s pomočjo grafikonov. Namen spletne aplikacije je tudi poenostaviti raziskovalceuporabo statističnih funkcij programa R, in sicer zbirno statistiko, test hi-kvadrat, t-test, ANOVA in MANOVA, s pomočjo grafičnega uporabniškega imena (GUI).

Eden izmed ciljev razvoja spletne aplikacije je tudi obstoj razumljive dokumentacije, ter možnost uvoza/izvoza podatkov. Na ta način lahko raziskovalci uporabljajo aplikacijo za analizo različnih naborov podatkov z interakcijo.

Key document information

Name and SURNAME: Tanya Deniz TOLUAY

Title of the final project paper: Creating Biostatistics Web Application for Quantitative Data Analysis Using R Shiny

Place: Koper

Year: 2022

Number of pages: 49 Number of figures: 15 Number of tables: 2

Number of appendix: 3 Number of appendix pages: 12

Number of references: 41

Mentor: Assist. Prof. Dr. Uroš Godnov

Keywords: R, R Shiny, biostatistics web application, quantitative data analysis, chi-square testing, student's t-test, ANOVA, MANOVA

Abstract: In research areas, it is expected from biology and bioinformatics researchers to perform many different analysis techniques, often by pipeline execution, relying on multiple software. The data is commonly quantitative and collected from various lab experiments, surveys, and observations. The pipelining of analysis is a delicate process that requires care and attention since a mistake in one step can be critically important, preventing from obtaining the actual results.

This thesis project recognises the importance of computational tools for biostatistics and aims to make the required operation easier. The web application has built using the R package, Shiny, with the purpose of quantitative data analysis, data exploration, calculating the statistics and creating meaningful plots. It also aims to simplify the process by providing a graphical user interface (GUI) for the researchers.

The desired implementation of the tools includes summary statistics, chi-square test, t-test, ANOVA and MANOVA. In addition, it should consist of easy to understand documentation, an example dataset format, and an option to download the output. This way, researchers can use the application for analysing various datasets by interacting with the application by using different tools efficiently.

ACKNOWLEDGEMENTS

Throughout the writing of this thesis, I have received a great deal of support and assistance.

I would like to thank my mentor, Assistant Professor Dr Uroš Godnov, for his valuable guidance throughout my studies, intellectually stimulating conversations, being there at every step of the process with support and providing me with the tools that I needed to choose the right direction and successfully complete my graduation thesis.

In addition, I would like to thank my family for their wise counsel, loving guide and being always there for me no matter what.

LIST OF CONTENTS

1 INTRODUCTION....1
1.1 Programming Background ....**Error! Bookmark not defined.**
1.2 Statistical Background 2
1.3 Purpose of the Study....4
2 METHODS....5
2.1 Statistics and Statistical Functions....5
2.2 Graphics and Graphical Functions....7
2.3 Construction of the Web Application 10
2.3.1 Server 10
2.3.1 UI & Visualization....**Error! Bookmark not defined.**1
3 RESULTS....13
3.1 Biostatistics Web Application....13
4 CONCLUSION....22
5 DALJŠI POVZETEK V SLOVENSKEM JEZIKU 23
6 REFERENCES 25

LIST OF TABLES

Table 1: Generated plots based on the statistical test chosen by the user 7
Table 2: Demonstration of the statistical test outputs with the model dataset 21

LIST OF FIGURES

Figure 1: Main page of the application 13
Figure 2: Quantitative data analysis tab after uploading the data.....14
Figure 3: Display of the output of the summary statistics from the dummy dataset.....15
Figure 4: Mosaic plot output for chi-square test (Sex ~ Smoking) 16
Figure 5: Bar plot output for chi-square test (Sex ~ Smoking) 16
Figure 6: Violin plot output for student's t-test (Weight ~ Group).....17
Figure 7: Box plot output for student's t-test (Weight ~ Group) 17
Figure 8: Scatter plot output for student's t-test (Weight ~ Group).....17
Figure 9: Violin plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex) 18
Figure 10: Box plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex).....18
Figure 11: Scatter plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex) 18
Figure 12: Q-Q plot output for ANOVA (Weight ~ Exercise Levels) 19
Figure 13: Dot plot output for ANOVA (Weight ~ Exercise Levels) 19
Figure 14: Q-Q plot output for MANOVA (Weight, Exercise Hours ~ Exercise Levels) ..20
Figure 15: Dot plot output for MANOVA (Weight, Exercise Hours ~ Exercise Levels) ...20

LIST OF APPENDICES

APPENDIX A *R Code*

APPENDIX B *User's Handbook Guide*

APPENDIX C *Supplemental Materials*

LIST OF ABBREVATIONS

ANOVA – analysis of variance test
MANOVA – multivariate analysis of variance test

1 INTRODUCTION

1.1 Programming Background

R is a programming language often used in statistical computing and data science (Vance, 2009). An essential factor for the popularity of R is the availability of numerous packages with different functionalities (Vance, 2009). R's standard functions are written in R itself, making it easy for users to follow the algorithmic choices (R Core Team, 2016). Furthermore, the capacity of R is extended through user-created packages, which allow specialised statistical techniques, graphical devices, import and export capabilities, and reporting tools (Wickham & Bryan, 2021). Another quality of R is static graphics, which can produce publication-quality graphs (Lewin-Koh, 2015). In addition, the R programming language offers excellent technological advancements to build a web application from scratch, such as the Shiny package (Chang et al., 2021). Shiny makes it easy to create interactive web apps straight from R, hosting standalone apps on a webpage or building dashboards, and can also be extended with CSS themes, JavaScript actions, and htmlwidgets (Vaidyanathan et al., 2020)(RStudio, 2013). Therefore, R was used to construct, deploy, and host the biostatistics web application with the extension of various packages.

R allows working with version control systems such as GitHub and GitLab. GitHub is an internet hosting provider for software development and version control using Git (Github, 2020). It offers the distributed version control and source code management functionality of Git while providing access control and various collaboration functions such as bug tracking, feature requests, task management, continuous integration and wikis for every project (Williams, 2012). GitLab is a web-based DevOps lifecycle tool that provides a Git repository manager that provides wiki functionality, issue tracking, and continuous integration and deployment pipeline using an open-source license (GitLab, 2020). It is constantly being expanded with new functions, offers pull requests and code reviews as well as package management, and at the same time enables simple code maintenance. GitLab has some notable advantages over GitHub; it gives developers an unlimited number of private repositories to use with an integrated continuous integration system (Vats, 2020). One of the main differences between GitHub and GitLab is the platform that showcases each philosophy. GitHub is more available and focused on infrastructure performance, while GitLab is more focused on offering a function-based system with a centralised and integrated platform for web developers (Vats, 2020). With respect to these differences, GitLab was used for storage, improvement, and maintenance of the code, as well as collaborating with the project mentor.

1.2 Statistical Background

Biostatistics, also known as biometry, is the application of statistical methods in biology studies and contains the planning of experiments, the gathering of knowledge, and, accordingly, the analysis and interpretation of data (Nature Portfolio, n.d.). In data analysis, two statistical methods are used: descriptive statistics, which summarise data using indexes such as mean and median, and inferential statistics, which draw conclusions from data using statistical tests (Mishra et al., 2019).

The crucial step in biostatistical inferential research is to have the main hypothesis and know which test is appropriate for the study. In some experiments, researchers want to explore the data and might not form any hypothesis, but there can be no statistical proof if there are no hypotheses. Therefore, a reasonable plan is to limit the number of confirmatory hypotheses severely (BMJ, n.d.). In the second step, the researcher has to decide whether data is paired (same subjects are measures at different time points or using different methods) or unpaired (each group have a different subject) (Mishra et al., 2019). In general, the analysis must reflect the design, so a paired design must be followed by a paired analysis and vice versa (BMJ, n.d.). For example, in most of the studies used in biostatistical research, the results of a crossover study or a case-control study in which controls were compared with cases such as for age, gender, and social class are not independent (BMJ, n.d.). Therefore, the test used should be determined from the data. The web application has been implemented with descriptive statistics summary and some of the most used inferential statistical testing methods (Yan et al., 2017); chi-square test, student's t-test, analysis of variance test and multivariate analysis of variance test.

A chi-square test is a tool for determining the group differences when measuring the dependent variable at the nominal level between the expected frequencies and the observed frequencies in one or more categories of a contingency table (Pearson, 1900). As with all nonparametric statistics, the chi-square is robust in terms of the distribution of the data. In particular, it does not require equal variance between study groups or homoscedasticity in the data (McHugh, 2013). It enables the evaluation of dichotomous independent variables as well as multiple group studies. Unlike many other nonparametric and parametric statistics, the calculations required to compute the chi-square provide considerable information about each of the groups performed in the study (McHugh, 2013). This wealth of detail enables the researcher to understand the results and thus derive more detailed information from these statistics. Even though the chi-square test is used for categorical variables, it has been included in the biostatistics web application due to being widely used by researchers and students for biological data analysis (Yan et al., 2017). The web application also takes Yate's correction for continuity into account. Yate's correction for

continuity corrects the error introduced by assuming that a continuous distribution can approximate the discrete probabilities of frequencies in the table (Yates, 1934; Salkind, 2010).

The student's t-test is appropriate for determining the statistically significant difference in the mean value between groups and checks if the difference happens by any chance or not (Gosset, 1908). The test is two-sided and uses the Welch (or Satterthwaite) approximation to the degrees of freedom (Satterthwaite, 1946; Welch, 1947). In general, a t-test can be bilateral, which indicates that the means are not equivalent, or it can be one-sided by indicating whether the observed mean is greater or less than the hypothetical mean. For example, researchers may use a student's t-test for independent, randomly sampled two normal populations, and the two independent groups have equal variances, but if there are more than two groups, it is better to use the analysis of variance test (SKP, n.d.).

The analysis of variance (ANOVA) test is appropriate for determining the variation among and between groups, including the differences between means (Girden, 1992). The test generalises the t-test beyond two means, and it is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation (Girden, 1992). In addition to looking at the variance within data groups, ANOVA also considers the sample size (the larger the sample, the lower the probability that outliers will be randomly selected for the sample) and the differences between the sample means (Girden, 1992). One way ANOVA compares the effects of an independent variable (a factor that affects other things) on multiple dependent variables, and researchers can use one way ANOVA to understand how different groups respond.

The multivariate analysis of variance (MANOVA) test is appropriate for comparing multivariate sample means, and it is an extension of ANOVA statistical testing (Friedrich & Pauly, 2018). As a multivariate procedure, it is used when there are two or more dependent variables and is often followed by significance tests involving individual dependent variables separately (Friedrich & Pauly, 2018). MANOVA extends this analysis by taking several continuous dependent variables into account and grouping them into a weighted linear combination or compound variable (Tabachnick et al., 2011). For instance, suppose a researcher is curious about the effect of various sorts of treatment (the IV; say, desensitisation, relaxation training, and a waiting-list control) on anxiety. In ANOVA, the researcher chooses one measure of hysteria from among many. With MANOVA, the researcher can assess several sorts of anxiety (say, test anxiety, anxiety in reaction to minor life stresses, and so-called free-floating anxiety) for testing the study (Tabachnick et al., 2011). The MANOVA compares whether or not the newly created combination differs according to the different groups or levels of the independent variables. In this way,

MANOVA essentially tests whether the independent grouping variable simultaneously explains statistically significant variance in the dependent variable (Tabachnick & Fidell, 2012).

1.3 Purpose of the Study

During the research and study of biological processes, the analysis and exploration of the biological datasets require the knowledge of programming, statistical background and a good sense of graphical understanding for visualizing the data. Therefore, this study aims to make data exploration and study easy and understandable to the researchers, students, and individuals interested in data analysis. Hence, the main objectives of developing a biostatistical web application for quantitative data analysis are; to minimize the time, knowledge and effort of numerous programming lines of code; to make the data exploration, analysis and statistical research easy to operate with a user-friendly and interactive interface as well as creating well-interpreting and meaningful graphical charts.

2 METHODS

A Shiny app is a directory containing two R scripts; one is ui.R, which controls the layout and appearance of the app (user interface), the other is server.R, which contains the back-end programming (Chang et al., 2021)(RStudio, 2013). In order to create a simpler, more user-friendly and efficient code, the application was divided into four different code snippets; statistical functions for calculating the chosen test, graphical functions for creating meaningful plots, server and UI to construct the web application. Even though the programming makes it very easy to calculate statistical testing and create plots, it is essential to understand the mechanisms behind what they mean and how to interpret the results logically. Therefore, in the methods part of the study, the statistical background will be considered together with the coding part.

2.1 Statistics and Statistical Functions

The inferential statistical functions are all collected under the getStats function, which calculates the results from the input dataset with chosen columns with the input desired statistical test. This function has been written separately and stored under *StatFunctions.R* file but later called in the server file of the web application (*Server.R*). The function is constructed with nested if-else logic; hence when the desired input statistical testing matches an if condition, the function will output the calculated statistics result.

The chi-square test (Pearson, 1900) can be calculated by the R function *chisq.test* (R Core Team, 2021), which takes an argument of two vectors, a logical indicator of whether to apply continuity correction when computing the test statistic and p-value cut off. The parameters are set as: correct = T, p = 0.05, which indicates Yate's continuity correction (Yates, 1934; Salkind, 2010) is applied with the p-value cut off equal to 0.05. This function outputs the calculated test statistic result, the degrees of freedom of the approximate chi-squared distribution of the test statistic, p-value, a character string indicating the type of test performed and name of the data, observed and expected counts. As well as the Pearson residuals [(observed - expected) / sqrt(expected)] and, the standardized residuals [(observed - expected) / sqrt(V)] where V is the residual cell variance (Pearson, 1900).

The R function *t.test* (R Core Team, 2021) can calculate the student's t-test (Gosset, 1908) by taking input arguments of two vectors, a character string specifying the alternative hypothesis, confidence level and a logical variable indicating whether to treat the two variances as equal. If set as TRUE, then the pooled variance is used to estimate the variance; otherwise, the Welch (or Satterthwaite) approximation to the degrees of freedom is used (Satterthwaite, 1946; Welch, 1947). The parameters are set as alternative: two-

sided, and the confidence level as 95%. The function outputs the result of the statistic, degrees of freedom, p-value, the confidence interval for the mean appropriate to the specified alternative hypothesis. It also outputs the estimated difference in means, the specified hypothesised value of the mean difference, the standard error of the mean (difference), a character string describing the alternative hypothesis and what type of t-test was performed.

In order to calculate the ANOVA and MANOVA tests, the functions used respectively as *oneway.test* (R Core Team, 2021) and *manova* (R Core Team, 2021) in R with input parameters as the desired input columns that the user has chosen. ANOVA tests whether two or more samples from normal distributions have the same means and returns to the value of the test statistic, the degrees of freedom, p-value and the method of the calculated test. The function manova calls the function aov (R Core Team, 2021), which calculates the analysis of variances and adds class manova to the result object for each stratum.

The written code can be seen below; the numbers indicate, respectively, calculation of the chi-square test (1), student's t-test (2), ANOVA test (3) and, MANOVA test (4). The function getStats gets two inputs df, as the data frame with chosen columns for calculation from the input dataset and input indicates the desired statistical test for calculation. Finally, the calculation result is returned as a transposed tidy (Robinson et al., 2012) data frame (5) to create a better visualised output for users.

```{R Code}
getStats <- function(df, input){
if(input$Model == "chisq"){ (1)
  res <- chisq.test(df$var1, df$var2, correct = T, p = 0.05)
}
if(input$Model == "ttest"){ (2)
  res <- t.test(var1 ~ var2, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=df)
}
if(input$Model == "anova"){ (3)
  res <- oneway.test(df$var1 ~ df$var2, var.equal = FALSE)
}
if(input$Model == "manova"){ (4)
  res <- manova(cbind(var1, var2) ~ Group, data = df)
}
```

```
  stat <-as.data.frame((t(tidy(res))))                                    (5)
  colnames(stat)<-"Value"
  return(stat)
}
```

2.2 Graphics and Graphical Functions

The graphical functions are all collected under the *getPlot* function, which is stored in *GraphicFunctions.R* file, and called in the server file of the web application. The function creates the appropriate graphs based on the chosen statistical test. In order to do so, ggplot2 (Wickham, 2016) was used. The ggplot2 allows programmers to create great visualisations to understand the data with only a few lines of code (Wickham, 2016), and it is a part of the tidyverse package (Wickham et al., 2019).

The tidyverse is a collection of several R packages (e.g. purrr, dplyr, tibble) designed for data science, introduced by Hadley Wickham and his team (2013) that share an underlying design philosophy and grammar. The generated plots in the web application consist of mosaic plot, bar chart, box plot, violin plot, scatter plot, Q-Q plot, and dot plot. The output plots change based on the chosen test and can be seen in more detail in the given Table 1.

Table 1: Generated plots based on the statistical test chosen by the user

Statistical Test	Generated Plots
Chi-square test	Mosaic plot, bar chart
Student's t-test	Box plot, violin plot, scatter plot
ANOVA	Q-Q plot, dot plot
MAVONA	Q-Q plot, dot plot

Due to the fact that chi-square testing has two specific purposes, to test the hypothesis of no association between two or more groups and to test how likely the observed distribution of data fits with the distribution that is expected (Sullivan, n.d.), two graphs are created with the user's chosen variables; a mosaic plot and a bar graph. A mosaic plot, also referred to as a Marimekko diagram, is a graphical representation of the contingency table, visualising data from two or more qualitative variables (Schlotzhauer, 2007). It gives a summary of the info and makes it possible to acknowledge relationships between different variables. Mosaic plots can be created in R with the package of ggmosaic (Jeppson, 2021), which was designed to create visualisations of categorical data.

A bar chart represents categorical data with rectangular bars with heights or lengths proportional to the values that they represent (Clagett, 1968). The bar chart is handy for

understanding the distribution of data points or comparing metric values across different subgroups of data. The bar chart has been created with the *geom_bar* function (Wickham et al., 2019).

The student's t-test has been implemented with the output graphs that would help users visualise the means, distributions and trends between the groups in the data. Therefore, created plots are box plot, violin plot, and scatter plot. The box plot is a method of graphing groups of numeric data by their quartiles (McGill et al., 1978). Box plots are not parametric: they represent the variation in the samples of a statistical population without making any assumptions about the underlying statistical distribution (Microsoft Academics, n.d.). The distances between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data. Box plots also show the variability outside the upper and lower quartile by lines and any outliers by whiskers. In addition to the points themselves, it can help visually estimate various L-estimators, notably the interquartile range, midhinge, range, mid-range, and trimean (Microsoft Academics, n.d.). Boxplots can be drawn in R with the function *geom_boxplot* (Wickham et al., 2019).

Violin plots are similar to box plots, except that they also show the probability density of the data at different values, with rectangular bars with heights or lengths proportional to the values that they represent (Hintze & Nelson, 1998). Hence, they are suitable for understanding the density trace in the data because it synergistically combines the box plot and the density trace into a single display that reveals the structure found within the data (Hintze & Nelson, 1998). In R, to create a violin plot, the function *geom_violin* can be used (Wickham et al., 2019).

A scatter plot uses Cartesian coordinates to display values and is very useful for plotting multivariable data to help determine the potential relationships among scale variables (Jarrell, 1994). It is used to see the overall trends and clusterings among variables, any displayed outliers from the overall trend, shape, and strength of the trend because the dots in a scatter plot report not only the values of individual data points but also patterns when the data are taken as a whole (Yi, 2019). The x and y coordinates for the quantiles are selected as the user's choice in the biostatistical web application. In order to create a scatter plot in R, the *geom_point* function has been used (Wickham et al., 2019).

For ANOVA and MANOVA tests, the web application has been implemented with a Q-Q plot and dot plot in order to help users see the variation within each group and the variation between the groups. The Q-Q plot, also known as the quantile-quantile plot, is a probability diagram that is a graphical method of comparing two probability distributions by plotting their quantiles against each other, by pairing the point values by their common

f-value (Wilk & Gnanadesikan, 1968). It is a great tool for comparing the shapes of distributions and provides a graphical representation of the similarity or differences in properties such as position, scale, and skewness in the two distributions (Ford, 2015). The pairing of values in a Q-Q plot is constructed from the ordering of values in each batch (Ford, 2015). The Q-Q graphs can be used to compare collections of data or theoretical distributions. Using Q-Q charts to compare two data samples can be viewed as a non-parametric approach to comparing the underlying distributions (Greenwood & Banner, n.d.). If the two distributions are similar, the points on the Q-Q graph are roughly on the line y = x (Wilk & Gnanadesikan, 1968). The x and y coordinates for the quantiles are selected as the user's choice in the biostatistical web application. Then, a point (x, y) in the graph corresponds to one of the quantiles of the second distribution (y-coordinate), plotted against the same quantile of the first distribution (x-coordinate). Drawing a Q-Q plot is achieved in programming in R by the *stat_qq* function and *stat_qq_line* (Wickham et al., 2019).

Dot plots, also known as density plots, are helpful for showing the distribution of a single scale variable (Wilkinson, 1999). The data are binned, but instead of one value for each bin, all of the points in each bin are displayed and stacked (Wilkinson, 1999). Dot plots are drawn in the R with the *geom_jitter* function (Wickham et al., 2019). The difference between the *geom_jitter* and *geom_dot* functions is the small amount of random noise to data. The *geom_jitter* is very useful to spread out points that would otherwise be overplotted (Wickham et al., 2019). Jittering the data into the whitespace allows the individual points to be seen. Using this method for ANOVA and MANOVA graphics, the web application will enable users with smaller datasets to see the data clearly, understand the behaviour and see the trends of grouping better.

The written code can be seen below; the numbers indicate, respectively, plots implemented for the chi-square test (1), student's t-test (2), ANOVA test and MANOVA test (3). The getPlot function is constructed with if-else nested logic and takes the input dataset with chosen columns and desired statistical test. In all the graphs, the x-axis and y-axis are left blank to give users the option to edit the name of variables in the plots later if desired. The function grid.arrange (R Core Team, 2021) is used for outputting multiple graphs at once. Finally, the *getPlot* function is called in the *Server.R* file, and the returned graphs are output by the *renderPlot* function (RStudio Inc, 2013).

```{R Code}
getPlot <- function(df, input){
  if(input$Model == "chisq"){ (1)
```

```
  p1 <-  ggplot(df) + geom_mosaic(aes(x=product(var1, Group),fill= Group)) + xlab("") +
ylab("") + ggtitle("Mosaic Plot")
  p2 <- ggplot(df, aes(x = var1, fill= Group)) + geom_bar() + xlab("") + ylab("") +
ggtitle("Bar Plot")
  plot <- grid.arrange(p1, p2, ncol=2, top= "Plots for Chi-Square Test")
 }
 if (input$Model == "ttest"){                                                    (2)
  p <- ggplot(df, aes(x= var2, y= var1, fill= Group))
  p1 <-  p + geom_violin(position = "dodge") + xlab("") + ylab("") + ggtitle("Violin Plot")
  p2 <- p + geom_boxplot() + xlab("") + ylab("") + ggtitle("Box Plot")
  p3 <- p + geom_dotplot(binaxis = "y", stackdir = "center", position = "dodge") +
xlab("") + ylab("") + ggtitle("Scatter Plot")
  plot <- grid.arrange(p1, p2, p3, ncol= 2, top= "Plots for Student's T-Test")
 }
 if(input$Model == "anova" | input$Model == "manova"){                           (3)
  p1 <- ggplot(df, aes(sample = var1, colour= Group)) + stat_qq() +
    stat_qq_line(aes(colour=Group)) + ggtitle("Q-Q Plot")
  p2 <- ggplot(df, aes(x = var1, y = var2)) + geom_jitter(aes(colour = Group), width =
0.1)  + xlab("") + ylab("") + ggtitle("Dot Plot")
  plot <- grid.arrange(p1, p2, ncol= 2, top= "Output Plots")
 }}
```

2.3 Construction of the Web Application

The code written for constructing the web application, server and UI, can be viewed in appendix A or on the GitLab repository (https://gitlab.com/tanyatoluay/finalthesis).

2.3.1 Server

Shiny Server is an open-source back end program that builds an internet server specifically designed to host Shiny apps (Chang et al., 2021)(RStudio Inc, 2013). Shiny Server will host each app at its web address and automatically start the app when a user visits the address. When the user leaves, Shiny Server will automatically stop the app (Chang et al., 2021). With Shiny Server, it is possible to host apps in a controlled environment. It also allows developers to start and stop the applications as required on a Linux server and provide a unique URL for every application (RStudio Inc, 2013). R also provides Shiny with a dashboard to support the activity on the developers' server, secure Shiny

applications using SSL (HTTPS), and regulate which users are allowed to access which applications (RStudio Inc, 2013).

The R Shiny Server has been used to create an interactive web application that allows users to input their dataset, choose desired columns as variables for statistical testing and the desired statistical test. In order to create the web page dynamic and user interactive, reactive expressions were used in the construction of the *Server.R*. To create reactive expressions, the function *reactive* (Chang et al., 2021) has been used for the dataset input, and the *observeEvent* (Chang et al., 2021) function has been used for allowing users to choose the desired column for further analysis. Reactive expressions are quicker than regular R functions as they cache the input values and know when they are outdated. After the first time running the expression, the expression saves the results in the memory of the computer. If the reactive function is called again, it can return the saved value without computation, allowing the web application to work faster. This procedure allows users to work with large datasets at ease. Another plus side is that the reactive expression will only return the saved result if it knows that the result's up-to-date. Suppose the reactive expression has learned that the result's obsolete; then the expression will recalculate. It then returns the new result and saves a replacement copy. The reactive expression will use this new copy until it too becomes out of date. In addition to these properties, *Server.R* has been used to calculate and output the results of the statistical testing and downloadable plots. For the deployment of the web application, the cloud shinyapps.io (RStudio Inc, 2013) has been used.

2.3.2 UI & Visualisations

The Shiny UI is responsible for building the dashboards and interactive user interfaces. The main code is written in R, but it can also be expanded with numerous packages (Chang et al., 2021)(RStudio Inc, 2013). In order to create a dynamic interface that is easy for users to navigate, *tabsets* (Chang et al., 2021) were used to organise the output. Tabsets are made by calling the *tabsetPanel* (Chang et al., 2021) function and require an input of different tabs, where each tab panel is also provided with a list of output elements. The primary three tabs in the web application are the quantitative analysis tab where users can view the uploaded dataset with the results, a handbook tab that describes how to use the web application, and the about tab that gives a short summary of the project with the information regarding the University of Primorska FAMNIT. Furthermore, the quantitative analysis tab is also divided into different tabs to categorise the outputs; dataset that shows uploaded dataset, a summary tab that shows descriptive statistics of the data, statistics result tab and the plots tab.

In order to make the web application easy to operate for users, a set of radio buttons and lists were added. The radio buttons can be used to indicate the separator of the dataset and if it has quotation marks or header and implemented by *radioButtons* (Chang et al., 2021) function. The lists are used for selecting the desired columns of the dataset for further analysis and implemented by the *selectInput* (Chang et al., 2021) function.

For enhancement of the design, the appropriate design icons are placed by each tab and drawn from the Font Awesome library (Fonticons Inc, 2018), as well as the spinners that indicate when output is recalculating with the function of the *withSpinner* (Sali & Attali, 2020). The library bslib package (Sievert & Cheng, 2021) provides tools for customising Bootstrap themes directly from R and providing easy access to pre-packaged Bootswatch themes (Park, n.d.). The bootstrap theme *cosmo* has been used to make the web application more aesthetically aligned with the design of the University of Primorska.

3 RESULTS

The web application is complete and functioning as intended. The code for the applications has been placed online in the GitLab repository hosting service, where it is freely available to view and download (https://gitlab.com/tanyatoluay/finalthesis). The user's handbook guide has also been placed on GitLab along with a model dataset as an example.

3.1 Biostatistics Web Application

The final application is hosted on the Shiny server that can be accessed by the web link https://tanyatoluay.shinyapps.io/finalthesis/. A model dataset has been created in order to test the web application. It consists of 60 individuals and their categorical and numerical data: 30 control and 30 case study group (nominal), sex (nominal), age (numerical), weight (numerical), weekly exercise hours (numerical), and information on smoking (nominal) and exercise levels (ordinal). Upon accessing the web application, quantitative data analysis, handbook, and about tabs can be seen (Figure 1). This functionality allows users to read about the project, learn how to navigate the web application from the handbook, and they can also download the model dataset.

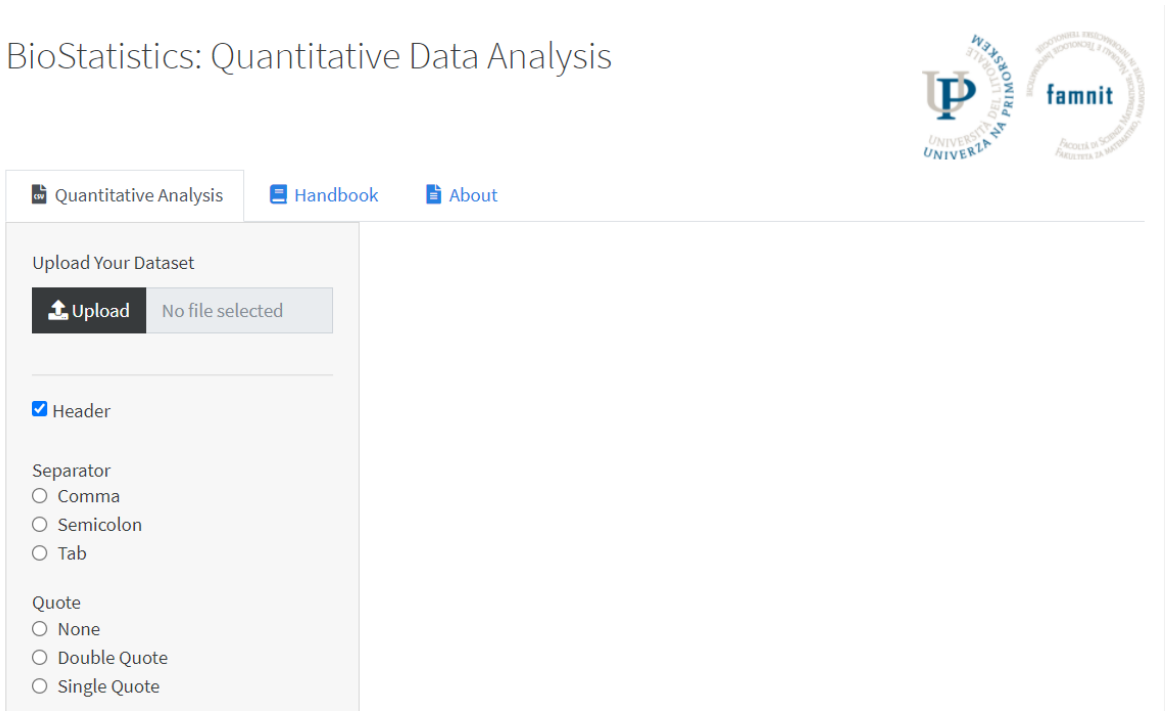

Figure 1: Main page of the application

The other tabs (data, summary statistics, statistics results and plots) can be seen after uploading the dataset (Figure 2). On the left side of the quantitative data analysis tab in the web application, users can enter the information about their datasets, such as indicating if the dataset has headers or quotes and the type of the separator. The default mode is set as the header exists, no quotes, and the separator is one tab. The uploaded dataset is visible in the data tab for making the choices easier for the users and also to provide feedback. Users can see the changes in the dataset after indicating the information about the dataset to make sure everything is correct. Afters users can choose desired columns from the data (first, second and group variables) and the statistical test for further analysis.

Dataset | Summary | Statistics | Plot

PatientID	Sex	Group	Age	Weight	WeeklyExerciseHours	Smoking	ExerciseLevel
1	Female	Control	22	50	7	Yes	Beginner
2	Female	Case	23	53	2	Yes	Intermediate
3	Female	Control	25	60	1	Yes	Advanced
4	Female	Case	30	64	10	Yes	Beginner
5	Female	Control	50	73	3	Yes	Beginner
6	Female	Case	21	63	2	Yes	Beginner
7	Female	Control	35	55	1	Yes	Beginner
8	Female	Case	31	57	5	Yes	Advanced
9	Female	Control	39	68	6	Yes	Advanced
10	Female	Case	37	53	2	Yes	Advanced
11	Female	Control	62	67	4	Yes	Advanced
12	Female	Case	26	68	6	Yes	Advanced
13	Female	Control	19	59	13	Yes	Advanced

Figure 2: Quantitative data analysis sub tabs after uploading the data

The created dummy dataset was uploaded to test and demonstrate the usage of the web application.

In the summary tab, users can view the descriptive statistics of the dataset. The output shows the number of rows and columns of the dataset, column type frequencies, for character variables shows the column name, missing and complete information rates, the minimum and maximum number of characters, number of unique values and number of white spaces. For the numerical values, it displays the column name, missing and complete information rates, mean and standard deviation values, and p0, p25, p50, p75, p100 values with a small histogram to see the distribution of the data. The summary tab for the descriptive statistics is a great tool for determining if the desired statistical test fits the data. Comparing mean, standard deviation values with the consideration of the distribution is

advised before applying the statistical test, as some statistical tests assume the distribution and some don't (such as student's t-test assume a normal distribution while the chi-square test does not). The output summary statistics results calculated from the model dataset can be seen below (Figure 3).

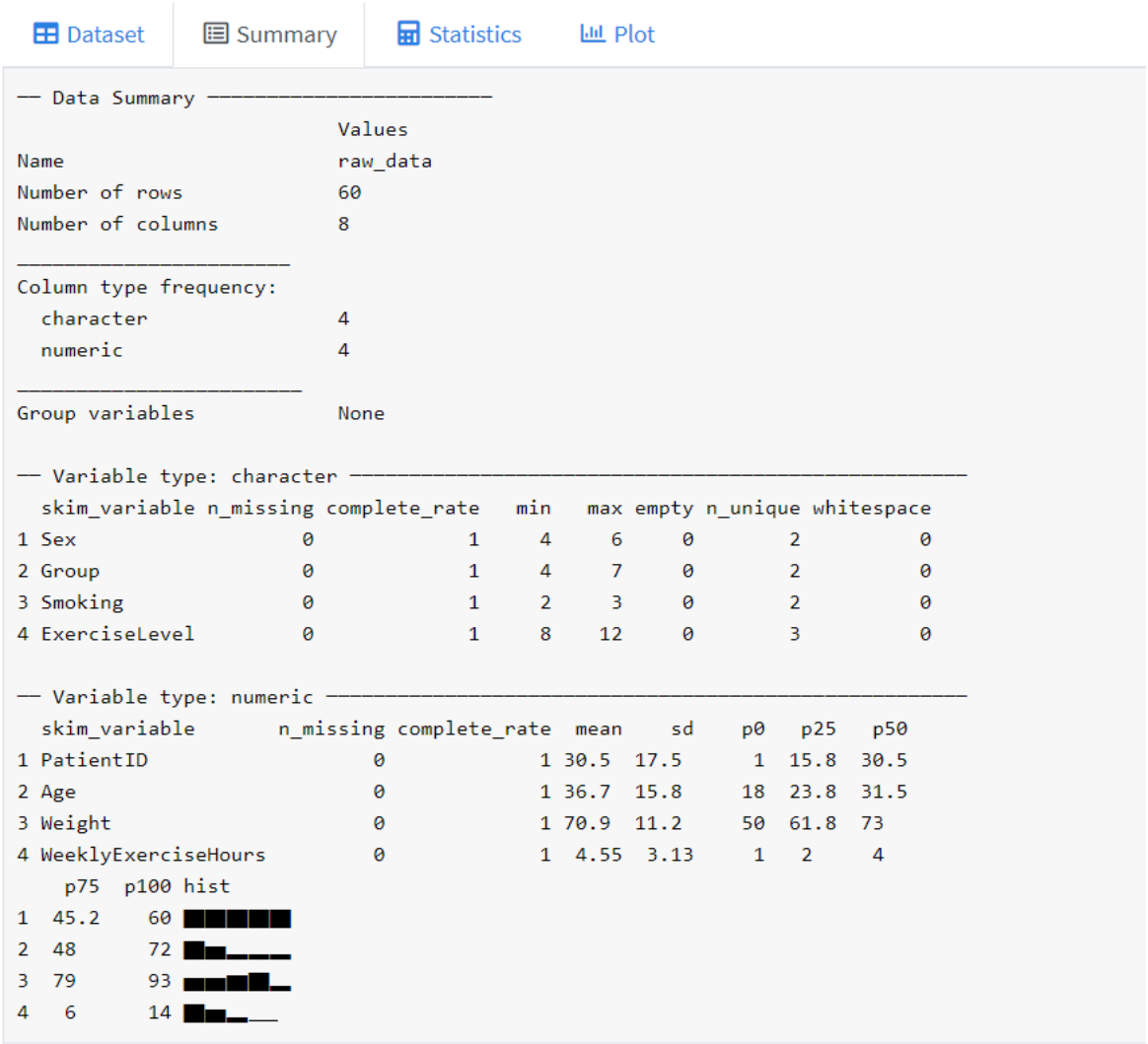

Figure 3: Display of the output of the summary statistics from the dummy dataset.

The statistical test has been tested over the model dataset with relevant information (Table 2). For example, assuming the studied topic is if smoking is related to sex. The null hypothesis would be that smoking and sex are not related, and the alternative hypothesis would be that smoking and sex are related. Therefore, the chi-square statistics can be chosen for statistical testing with the proper variables (first variable: smoking, second variable: sex, group: sex). The web application will output the results of statistics value, p-value, number of parameters and the used method in the statistics tab. In the plot tab, a mosaic plot and a bar chart will be created, and the output plots can be downloaded in png or pdf format, which the user can determine from the left sidebar (Figure 4, Figure 5).

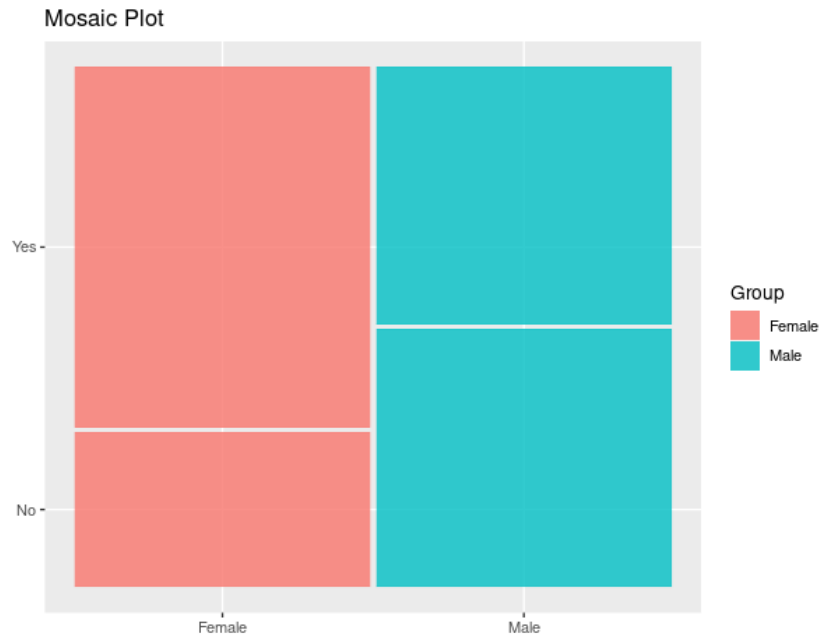

Figure 4: Mosaic plot output for chi-square test (Sex ~ Smoking).

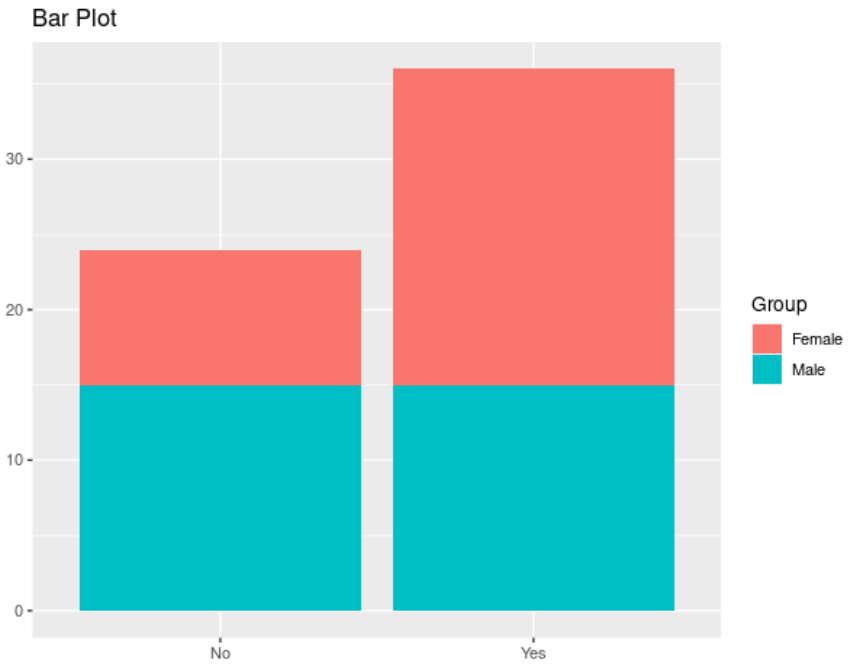

Figure 5: Bar plot output for chi-square test (Sex ~ Smoking).

In another example, the studied topic can be the weight relation with the group information among individuals (control group and case study). The null hypothesis would be that there is no difference in the weight means between the control healthy group and the case study group, and the alternative hypothesis would be that there is a difference in the weight means between the two groups. Therefore, the student's t-test can be chosen with the following variables; the first variable: weight, second variable: group, group: group. The web application will output the estimates, statistics, p-value, parameters, low and high

confidence intervals with the used method (which is set as two sampled in the web application). The output plots are as intended and downloadable: violin plot, box plot and scatter plot (Figure 6, Figure 7, Figure 8).

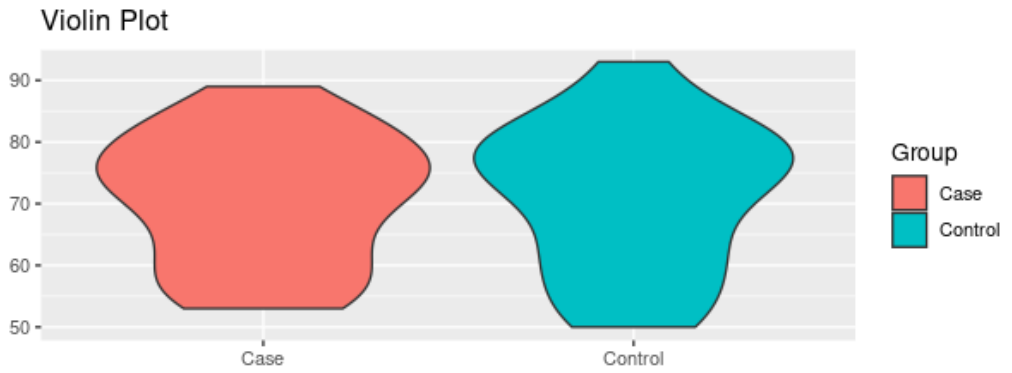

Figure 6: Violin plot output for student's t-test (Weight ~ Group).

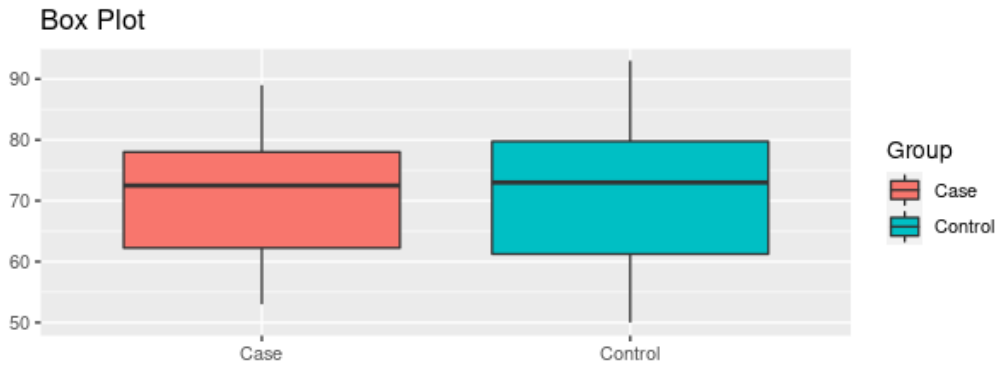

Figure 7: Box plot output for student's t-test (Weight ~ Group).

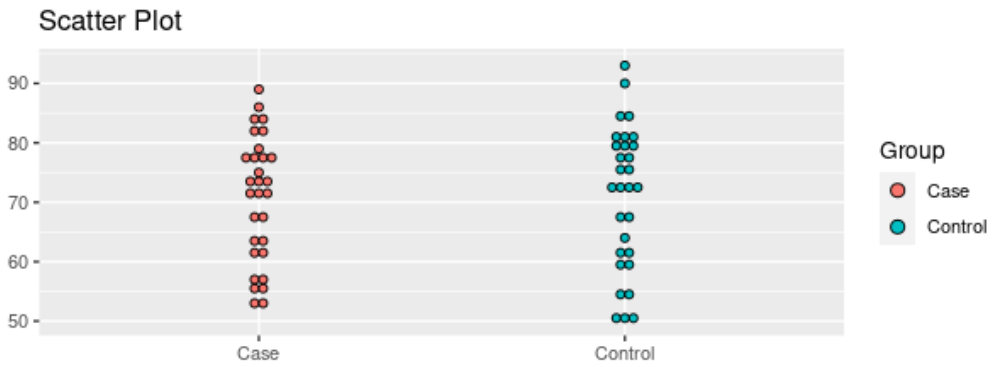

Figure 8: Scatter plot output for student's t-test (Weight ~ Group).

In the plots section for the student's t-test, the additional grouping option is available for users who want to see more differences and trends on the data with more grouping. This option has been demonstrated (Figure 9, Figure 10, Figure 11) by the variables chosen as the first variable: weight, second variable: group (control group vs case study group), but

the grouping variable is further chosen as sex to see the trend of weight~group data among the sexes.

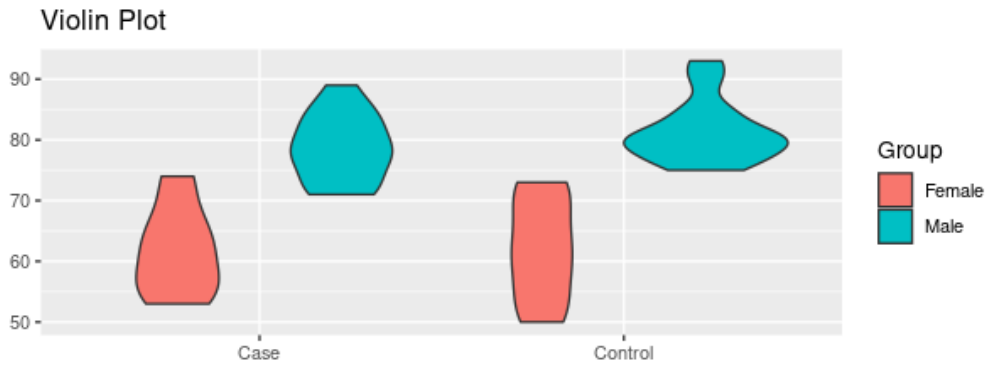

Figure 9: Violin plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex).

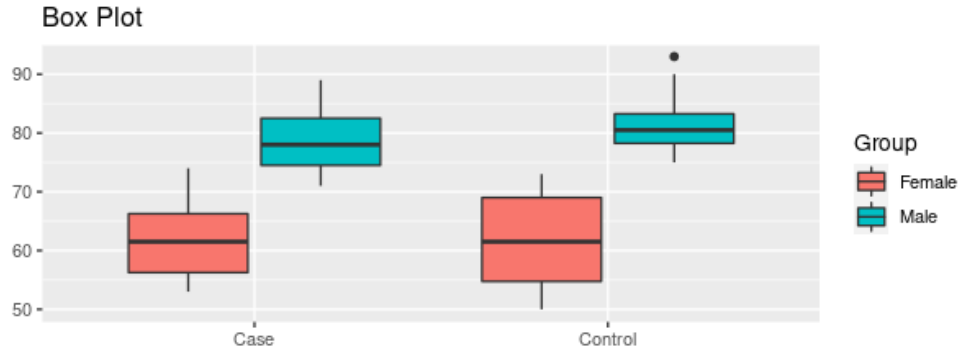

Figure 10: Box plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex).

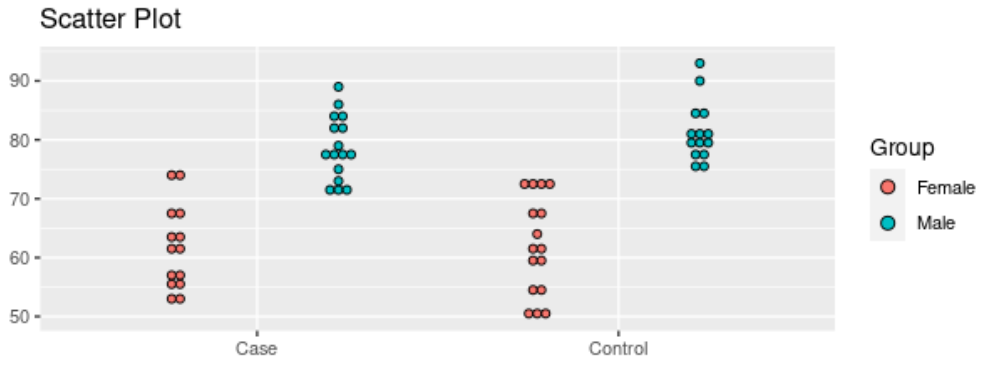

Figure 11: Scatter plot output for student's t-test with further grouping (Weight ~ Group, Coloured by: Sex).

The data of exercise level is divided into three categories: beginner, intermediate and advanced, which indicates the advancement of the exercise programme individual takes. In order to see the correlation between the means of weight with exercise level, the null hypothesis would be that there is no difference in the weight means among different exercise levels (hence, average weight is the same for individuals who does beginner or

intermediate or advanced level exercise). The alternative hypothesis would be that there is a difference in the weight means among different exercise levels. Therefore, the ANOVA test can be applied (the first variable: weight, second variable: exercise level, group: exercise level). The web application outputs the degrees of freedom, statistics result, p-value and the applied method (one-way analysis of means is applied as the method in the web application) as well as Q-Q plot and dot plot (Figure 12, Figure 13).

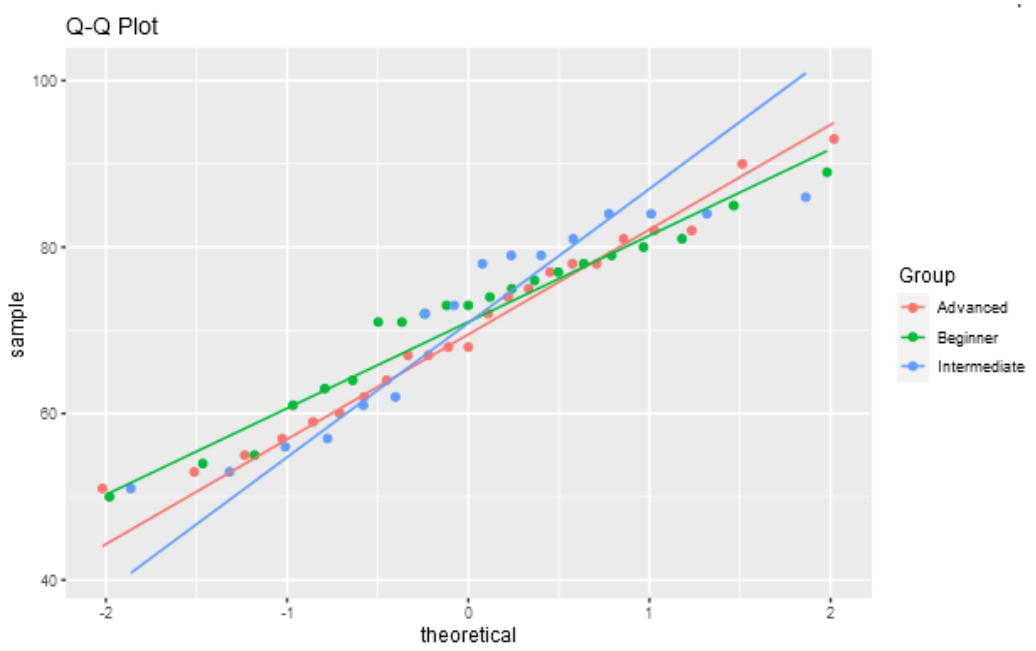

Figure 12: Q-Q plot output for ANOVA (Weight ~ Exercise Levels).

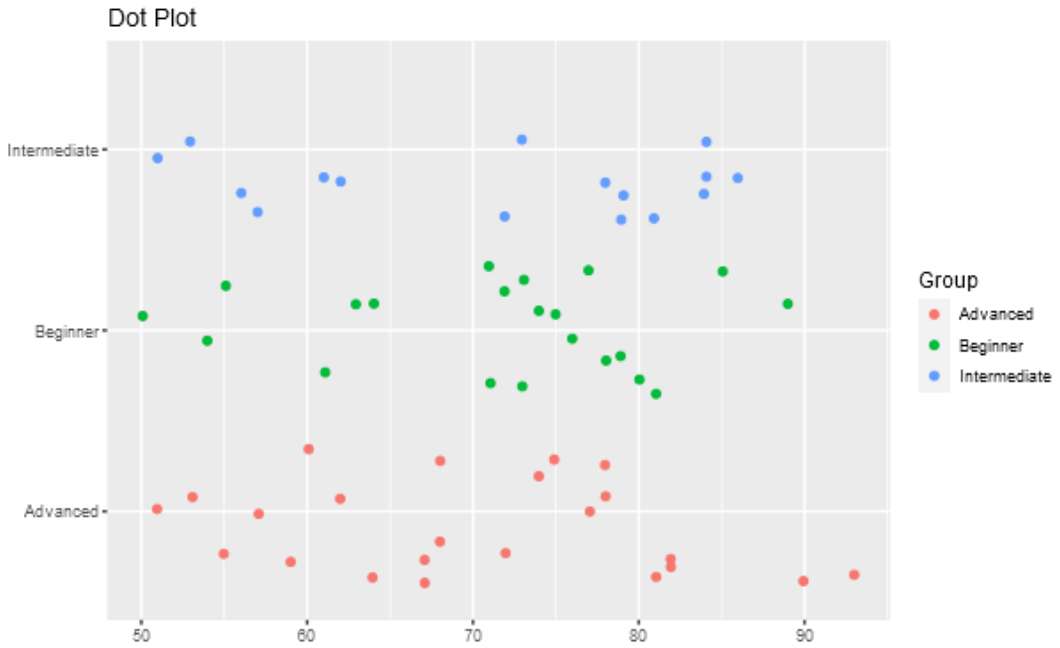

Figure 13: Dot plot output for ANOVA (Weight ~ Exercise Levels).

The MANOVA test, as an extension of the ANOVA test, will also take continuous dependent variables into account and grouping them into a weighted linear combination or compound variable (Tabachnick et al., 2011). So another example would be extending the analysis of weight correlation with exercise level by taking weekly exercise hours into the calculation (the first variable: weight, second variable: weekly exercise hours, group: exercise level). The output results will be Pillai, statistics results, degrees of freedom, number of degrees of freedom associated with the model errors and the p-value with the output graphs of Q-Q plot and dot plot (Figure 14, Figure 15).

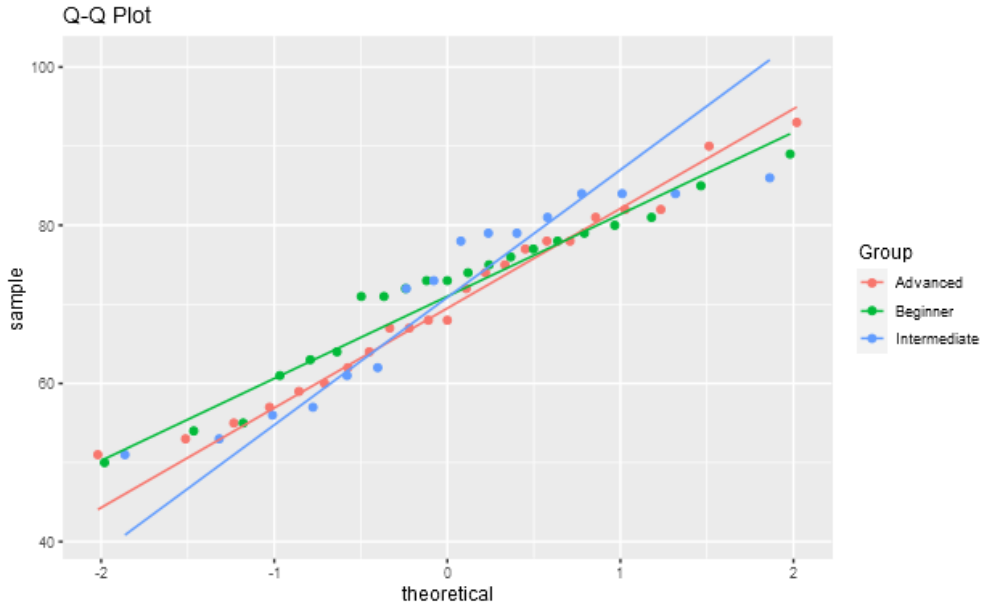

Figure 14: Q-Q plot output for MANOVA (Weight, Exercise Hours ~ Exercise Levels).

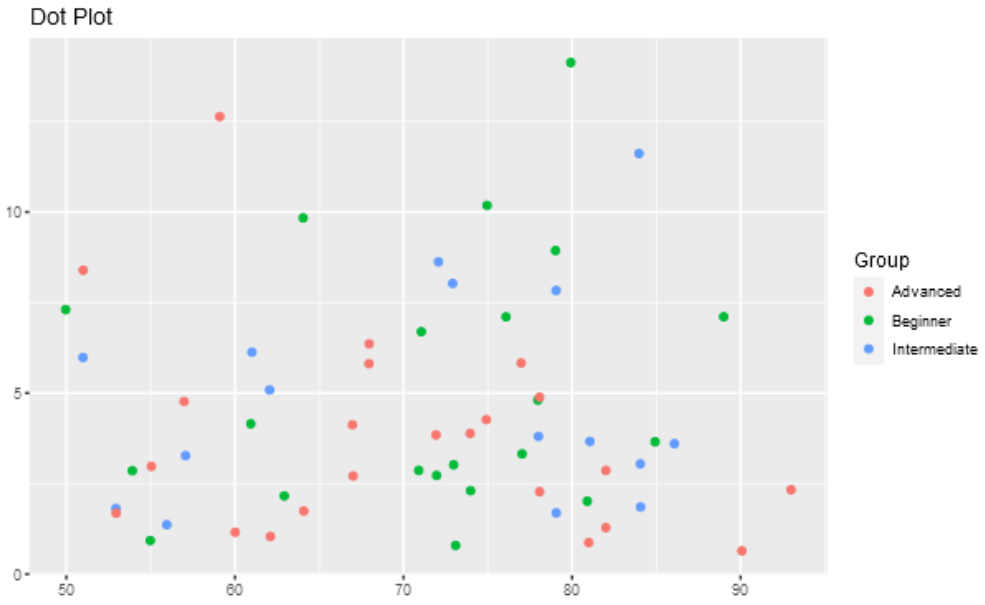

Figure 15: Dot plot output for MANOVA (Weight, Exercise Hours ~ Exercise Levels).

Table 2: Demonstration of the statistical test outputs with the dummy dataset

Test Statistics	First Variable	Second Variable	Group	Output Results	Output Plots
Chi-square test	Smoking	Sex	Sex	Statistic: 1.736111 P.value: 0.1876323 Parameter: 1 Method: Pearson's Chi-squared test with Yates' continuity correction	Figure 4, Figure 5
Student's t-test	Weight	Group	Group	Estimate: -0.2666667 Estimate1: 70.76667 Estimate2: 71.03333 Statistic: -0.09134658 P.value: 0.927532 Parameter: 58 Conf.low: -6.11025 Conf.high: 5.576916 Method: Two Sample t-test Alternative: two.sided	Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11
ANOVA	Weight	Exercise Level	Exercise Level	Num.df : 2 Den.df: 35.0155 Statistic: 0.08815461 P.value: 0.9158219 Method: One-way analysis of means (not assuming equal variances)	Figure 12, Figure 13,
MANOVA	Weight	Weekly Exercise Hours	Exercise Level	Group Residuals Df : 2 57 Pillai 0.041029 <NA> Statistic 0.59692 <NA> Num.df 4 <NA> Den.df 114 <NA> P.value 0.06655871	Figure 14, Figure 15

The calculated results have been double-checked in the R environment to make sure the web application is working correctly. All the outputs for the statistical results from the model dataset can be seen in Table 2.

4 CONCLUSION

The bioinformatics research techniques like biostatistics analysis have become more of standard practice in scientific research. Although, at the same time, the development of the technological tools allows researchers to gather and work on large datasets, therefore the necessity for tools that assist researchers in analysing or managing their data increases. R is a great programming language, especially for statistics and data analysis, but the steep learning curve limits its usefulness for several researchers. The created web application removes this barrier and supplies the extra benefit of interactive, dynamic visualisations. In addition, all the code has been uploaded to GitLab with documentation for everyone to view and download. Thus, it is open for further possibilities of collaboration and development with other bioinformaticians and data scientists. Moreover, given the extensive collection of R packages explicitly designed to be used in biological data analysis, this approach might increase their accessibility to a broader pool of users.

5 DALJŠI POVZETEK V SLOVENSKEM JEZIKU

Analiza biostatistike je običajna praksa v znanstvenih raziskavah. Razvoj tehnoloških orodij raziskovalcem omogoča zbiranje in delo z velikim naborom podatkov. Posledično se povečuje potreba po orodjih, ki raziskovalcem pomagajo pri analizi in upravljanju podatkov. Analiza bioloških podatkov zahteva znanje statističnih in grafičnih orodij ter programiranje. Za potrebe te naloge sta bili uporabljeni dve orodji: programski jezik R in orodje za nadzor različic – GitLab.

R je programski jezik, ki se pogosto uporablja v statistiki in analizi podatkov, vendar strma krivulja učenja omejuje njegovo uporabnost za mnoge raziskovalce. Cilj študije je ustvariti interaktivno in dinamično biostatistično spletno aplikacijo, ki je enostavna za uporabo za raziskovalce in študente, hkrati pa tudi dovolj preprosta za širšo populacijo. R ponuja različne pakete: na primer paket Shiny, ki uporabnikom omogoča izdelavo spletnih aplikacij. Spletna aplikacija v nalogi je ustvarjena z uporabo paketov R Shiny in drugih paketov R, ki so specializirani za analizo podatkov (na primer tidyverse).

Orodje za nadzor različic GitLab je bilo uporabljeno za shranjevanje in vzdrževanje kode ter sodelovanje z mentorjem študije. Spletna aplikacija uporabniku ponuja interaktivni uporabniški vmesnik za nalaganje nabora podatkov, izbiro želenih stolpcev za analizo in izbiro priljubljenih orodij za analiziranje podatkov v bioloških raziskavah: Hi-kvadrat test, študentov t-test, ANOVA in MANOVA.

Hi-kvadrat test je orodje za določanje skupinskih razlik pri merjenju odvisne spremenljivke na nominalni ravni med pričakovanimi in opaženimi frekvencami v eni ali več kategorijah preglednice nepredvidljivih dogodkov. Za razliko od mnogih drugih neparametričnih in parametričnih statistik, formule, ki so potrebne za izračun hi-kvadrata, dajejo znatne informacije o vseh skupinah v študiji. Te podrobnosti omogočajo raziskovalcu, da razume rezultate in tako iz teh statistik pridobi bolj podrobne informacije kot z drugimi vrstami statistik. Kljub temu, da se Hi-kvadrat test uporablja za kategorične spremenljivke, je bil vključen v spletno aplikacijo za analizo biostatistike. Raziskovalci in študenti ga namreč pogosto uporabljajo za analizo bioloških podatkov, kot so na primer klinični podatki.

Spletna aplikacija na hi-kvadrat test aplicira tudi Yatesov popravek za kontinuiteto, ki popravi morebitne napake pod predpostavko, da se lahko neprekinjena porazdelitev približa diskretnim verjetnostim frekvenc v tabeli. Mejna vrednost p je enaka 0,05. Ker ima hi-kvadrat testiranje dva namena (1) testirati hipotezo, ki nima nobene povezave med dvema ali več skupinami in (2) testirati verjetnost, da bo porazdelitev opazovanih podatkov ustrezala pričakovani porazdelitvi, se v aplikaciji ustvarita tudi dva tipa grafikonov: stolpični in mozaični grafikon.

Student-ov t-test se uporablja za določanje statistično pomembne razlike aritmetičnih vrednosti med skupinami in za preverjanje, ali se ta razlika zgodi naključno ali ne. Uporabljena raven zaupanja je enaka .95. Test je dvostranski in uporablja Welchov (ali

Satterthwaitov) približek stopnjam svobode. Student-ov t-test je bil implementiran v spletno aplikacijo s škatlastim, violinskim in razpršenim grafikonom.

Test ANOVA je primeren za določanje variance med skupinami in znotraj njih, in razlikami med njihovo povprečno vrednostjo. Preskus t-testa na splošno presega dva načina. Temelji na zakonu skupne variance, kjer je opažena varianca določene spremenljivke razdeljena na komponente, ki jih je mogoče pripisati različnim virom variacije. ANOVA poleg preučevanja variance znotraj podatkovnih skupin upošteva tudi velikost vzorca (večji kot je vzorec, manjša je verjetnost, da bodo deviacije naključno izbrane iz vzorca) in razlike med vzorčnimi povprečji. Spletna aplikacija podpira enosmerni test ANOVA.

Test MANOVA je primeren za primerjavo multivariatnih vzorčnih povprečji in je razširitev statističnega testiranja ANOVA. Kot multivariatni postopek se uporablja, kadar sta prisotni dve ali več odvisnih spremenljivk. Pogosto mu sledijo tudi preizkusi pomembnosti, ki ločeno vključujejo posamezne odvisne spremenljivke. MANOVA razširi to analizo tako, da upošteva več kontinuiranih odvisnih spremenljivk in jih združi v uravnoteženo linearno kombinacijo ali sestavljeno spremenljivko. Zato je grafični izhod za teste ANOVA in MANOVA v spletni aplikaciji Q-Q ali pikčasti grafikon.

Spletna aplikacija ponuja tudi priročnik, ki prikazuje njeno uporabo, in stran, kjer uporabnik lahko najde kratke informacije o projektu in Univerzi na Primorskem, FAMNIT. Poleg tega zavihek kvantitativne analize z majhnim histogramom prikaže celoten povzetek statistik, ta zajema dovodni set podatkov, ki odvaja podatke za manjkajočo in popolno informacijsko stopnjo, minimalne in maksimalne vrednosti, povprečje, sd, p0, p25, p50, p75 in p100. Histogram prikazuje distribucijo podatkov za vsak numerični stolpec. Ostala dva zavihka sta statistični zavihek, ki prikaže rezultate izbranega statističnega testa in zavihek ki omogoča prenos ustreznih grafov v png ali pdf obliki.

Koda je izvedena na štirih različnih mestih: (1) strežniku, ki gradi spletno aplikacijo, (2) UI, ki nadzoruje uporabniški vmesnik spletne aplikacije (3) datoteki statističnih funkcij in (4) datoteki grafičnih funkcij. Koda je napisana z dokumentacijo, projekt pa je dostopen za ogled in prenos vsem na GitLab portalu.

6 REFERENCES

[1] Chang, W., Cheng, J., Allaire, JJ., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert A., Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

[2] Clagett, M. (1968). Nicole Oresme and the Medieval Geometry of Qualities and Motions, Madison: Univ. of Wisconsin Press, pp. 85–99

[3] Ford, C. (2015, August 26). Understanding Q-Q Plots. Retrieved from https://data.library.virginia.edu/understanding-q-q-plots/

[4] Friedrich, S., Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. Journal of Multivariate Analysis, 165, 166-179.

[5] Girden, E. R. (1992). ANOVA: Repeated measures. Sage.

[6] Github. (2020). GitHub. Retrieved from https://github.com/

[7] GitLab. (2020). GitLab. Retrieved from https://about.gitlab.com/

[8] Gosset, W.S (1908), Probable error of a correlation coefficient. Biometrika. 6(2/3): 302-310.

[9] Greenwood, M., Banner, K. (n.d). ANOVA model diagnostics including QQ-plots. Retrieved from https://arc.lib.montana.edu/book/statistics-with-r-textbook/item/57#ANOVA%20model%20diagnostics%20including%20QQ-plots++3

[10] Hintze, J. L., Nelson, R. D. (1998) Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician 52, 181-184.

[11] Jarrell, S. B. (1994). Basic Statistics (Special pre-publication ed.). Dubuque, Iowa: Wm. C. Brown Pub. p. 492.

[12] Jeppson, H., Hofmann, H., Cook, D., Wickham, H., (2021, February 23). Retrieved from https://cran.r-project.org/web/packages/ggmosaic/ggmosaic.pdf

[13] JMP. (n.d.). Retrieved from https://www.jmp.com/en_us/statistics-knowledge-portal

[14] Lewin-Koh, N. (2015, July 1). CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization. Retrieved from https://cran.r-project.org/web/views/Graphics.html

[15] McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32(1), 12–16.

[16] McHugh M. L. (2013). The chi-square test of independence. Biochemia medica, 23(2), 143–149. https://doi.org/10.11613/bm.2013.018

[17] Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of appropriate statistical methods for data analysis. Annals of cardiac anaesthesia, 22(3), 297–301. https://doi.org/10.4103/aca.ACA_248_18

[18] Nature. (n.d.). Retrieved from https://www.nature.com/subjects/biostatistics

[19] Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), 157–175.
[20] R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/
[21] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. v3.6.2
[22] Robinson, D., Hayes, A., Couch, S. (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.6. https://CRAN.R-project.org/package=broom
[23] RStudio, Inc. (2013) Easy web applications in R. Retrieved from https://shiny.rstudio.com/
[24] Salkind, N. J. (2010). Encyclopedia of research design (Vols. 1-0). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288
[25] Sali, A., Attali, D. (2020). shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating. R package version 1.0.0. https://CRAN.R-project.org/package=shinycssloaders
[26] Schlotzhauer, S. D. (2007, April 1). Elementary Statistics Using JMP. SAS Institute. p. 407
[27] Sievert, C., Cheng, J. (2021). bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'. R package version 0.2.5.1. https://CRAN.R-project.org/package=bslib
[28] Sullivan, L. (n.d.). Hypothesis Testing - Chi Squared Test. Boston University School of Public Health. Retrieved from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare_print.html
[29] Tabachnick B.G., Fidell L.S. (2011) Multivariate Analysis of Variance (MANOVA). In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_394
[30] Tabachnick, B. G. & Fidell, L. S. (2012). Using multivariate statistics (6th ed.). Boston, MA: Pearson.
[31] Vaidyanathan, R., Xie, Y., Allaire, JJ., Cheng, J,. Sievert, C., Russell, K. (2020). htmlwidgets: HTML Widgets for R. R package version 1.5.3. https://CRAN.R-project.org/package=htmlwidgets
[32] Vance, A. (2009, June 6). Data Analysts Captivated By R's Power. Retrieved from https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html[3]
[33] Vats, R. (2020, May 27M). GitHub vs GitLab: Difference Between GitHub and GitLab. Retrieved from https://www.upgrad.com/blog/github-vs-gitlab-difference-between-github-and-gitlab/
[34] Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. Biometrika, 38, 330--336. 10.2307/2332579.

[35] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

[36] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. v1.3.1.

[37] Wilk, M.B.; Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. Biometrika. Biometrika Trust.

[38] Wilkinson, L. (1999). "Dot plots". The American Statistician. American Statistical Association. 53 (3): 276–281.

[39] Yan, F., Robert, M., & Li, Y. (2017). Statistical methods and common problems in medical or biomedical science research. International journal of physiology, pathophysiology and pharmacology, 9(5), 157–163.

[40] Yates, F. (1934). Contingency table involving small numbers and the $\chi 2$ test. Supplement to the Journal of the Royal Statistical Society 1(2): 217–235. JSTOR 2983604

[41] Yi, M. (2019, October 16). A Complete Guide to Scatter Plots. Retrieved from https://chartio.com/learn/charts/what-is-a-scatter-plot/

APPENDICES

APPENDIX A *R Code*

R code is being provided in this appendix for completeness but is better viewed after being downloaded from GitLab (See Appendix C: Supplemental Materials).

StatFunctions.R

```{R Code}
getStats <- function(df, input){

# Chi-square Test
if(input$Model == "chisq"){
  res <- chisq.test(df$var1, df$var2, correct = T, p = 0.05)
}

# Student's T-Test
if(input$Model == "ttest"){
  res <- t.test(var1 ~ var2, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=df)
}

# ANOVA Test (One Way)
if(input$Model == "anova"){
  res <- oneway.test(df$var1 ~ df$var2, var.equal = FALSE)
}

# MANOVA Test (One Way)
if(input$Model == "manova"){
  res <- manova(cbind(var1, var2) ~ Group, data = df)
}
  stat<-as.data.frame((t(tidy(res))))
  colnames(stat)<-"Value"
  return(stat)
}
```

GraphicFunctions.R

```{R Code}
getPlot <- function(df, input){
  if(input$Model == "chisq"){
    p1 <-  ggplot(df) + geom_mosaic(aes(x=product(var1, Group),fill= Group)) + xlab("") + ylab("") + ggtitle("Mosaic Plot")
    p2 <- ggplot(df, aes(x = var1, fill= Group)) + geom_bar() + xlab("") + ylab("") + ggtitle("Bar Plot")
    plot <- grid.arrange(p1, p2, ncol=2, top= "Plots for Chi-Square Test")
  }
  if (input$Model == "ttest"){
    p <- ggplot(df, aes(x= var2, y= var1, fill= Group))
    p1 <-  p + geom_violin(position = "dodge") + xlab("") + ylab("") + ggtitle("Violin Plot")
    p2 <- p + geom_boxplot() + xlab("") + ylab("") + ggtitle("Box Plot")
    p3 <- p + geom_dotplot(binaxis = "y", stackdir = "center", position = "dodge") + xlab("") + ylab("") + ggtitle("Scatter Plot")
    plot <- grid.arrange(p1, p2, p3, ncol= 2, top= "Plots for Student's T-Test")
  }
  if(input$Model == "anova" | input$Model == "manova"){
    p1 <- ggplot(df, aes(sample = var1, colour= Group)) + stat_qq() +
      stat_qq_line(aes(colour=Group)) + ggtitle("Q-Q Plot")
    p2 <- ggplot(df, aes(x = var1, y = var2)) + geom_jitter(aes(colour = Group), width = 0.1)  + xlab("") + ylab("") + ggtitle("Dot Plot")
    plot <- grid.arrange(p1, p2, ncol= 2, top= "Output Plots")
  }}
```

Server.R

```{R Code}
shinyServer(function(input,output, session){

    # Uploading the Dataset (Reactive)
    data1 <- reactive({
        inFile <- input$file1
        if(is.null(file)){return()} [1]
        read.csv(inFile$datapath, header=input$header, sep=input$sep, quote=input$quote)})
```

```
# Output the Dataset
output$contents <- renderTable({
   inFile <- input$file1
   if (is.null(inFile))
      return(NULL)
   read.csv(inFile$datapath, header=input$header, sep=input$sep, quote=input$quote)
})

#Choose Columns as Variables
observeEvent(input$file1,{
   varib1 <- updateSelectInput(session, "varlist",choices=c(colnames(data1())))
   varib2 <- updateSelectInput(session,"varlist2",choices=c(colnames(data1())))
   varib3 <- updateSelectInput(session,"varlist3",choices=c(colnames(data1())))
}
#Create a SubDataset
new_dataset <- function(){
   raw_data <- data1()
   df<-data.frame(
      var1 = raw_data[,input$varlist],
      var2 = raw_data[,input$varlist2],
      Group = raw_data[,input$varlist3]
   )
}

# Calculate Chosen Statistics
new_data <- function(){
   df <- new_dataset()
   statRes <- getStats(df, input)
   return(statRes)
}

# Output the Results of Statistics
output$stats <- renderPrint({
   new_data()
})

# Output of the Summary for the Data
output$summary <- renderPrint({
   raw_data <- data1()
```

```
    skim(raw_data)
  })

  # Create Graphs
  new_plot <- function(){
    df <- new_dataset()
    plot <- getPlot(df, input)
    return(plot)
  }

  # Output the Graphs
  output$myPlot <- renderPlot({
    new_plot()
  })

  # Download Handler
  output$savePlot <- downloadHandler(
    filename = function() {
      paste("myplot",input$type,sep=".")},
    content = function(file){
      if(input$type=="png") png(file)
      else pdf(file)
      print(new_plot())
      dev.off()
    })
})
```

UI.R

```{R Code}
  shinyUI(fluidPage(

    titlePanel(
      fluidRow(
        column(9, "BioStatistics: Quantitative Data Analysis"),
        column(3, list(tags$head(tags$style()),
```

```
                    HTML('<img
src="https://8ecm.si/system/admin/files/files/000/000/206/original/Logo_Famnit_-
_okrogel.png?1591952610",
          height="100px" style="float:right"/>','<p style="color:black"></p>')))
          )
      ),

        theme = bs_theme(version = 4, bootswatch = "cosmo"),
  tabsetPanel(
    tabPanel(
      title="Quantitative Analysis",
      icon = icon("file-csv"),
      fluid = TRUE,
  sidebarLayout(
    sidebarPanel(

      # Uploading the Dataset
      fileInput('file1', 'Upload Your Dataset',
                accept = c('text/csv',
                           'text/comma-separated-values,text/plain','.csv'), buttonLabel =
list(icon("upload")," Upload")),
      tags$hr(), checkboxInput('header', 'Header', TRUE),
      radioButtons('sep', 'Separator', c(Comma=',', Semicolon=';',
                                         Tab='\t'),'Comma'),
      radioButtons('quote', 'Quote',
                   c(None='', 'Double Quote'='"', 'Single Quote'="'"),
                   'Double Quote'),

      # Selection of Variables
      selectInput("varlist","First Variable:",choices=c(colnames(data()))),
      selectInput("varlist2","Second Variable:",choices=c(colnames(data()))),
      selectInput("varlist3","Group by:",choices=c(colnames(data()))),

      # Selection of Statistical Test
      selectInput("Model","Statistical Test:",
                  list("Chi-square test" = "chisq",
                       "T-test" = "ttest",
                       "Anova test" = "anova",
                       "Manova test" = "manova")),
```

```
        # Selection of Plot Saving
        radioButtons("type", "Select the plot file type", choices=list("png","pdf"))),

      # Tabs
      mainPanel(
        conditionalPanel(condition = "input.varlist!= "", withSpinner(
        tabsetPanel(id = "tabbar", #The conditional pannel for visibility of tabs functionality in the quantitative data analysis tab has been added by Dr Uros Godnov
tabPanel("Dataset", icon= icon("table"), withSpinner(tableOutput('contents'))),
          tabPanel("Summary", icon = icon("list-alt"),withSpinner(verbatimTextOutput("summary"))),
          tabPanel("Statistics",icon = icon("calculator"), withSpinner(verbatimTextOutput("stats"))),
          tabPanel("Plot", icon = icon("bar-chart-o"), withSpinner(plotOutput("myPlot")),
              downloadButton("savePlot", "Save the Plot")))

      ) #tabsetPanel(id = "tabbar"
      ) #conditional panel,
    ) #main panel
    ) #sidebarLayout
), #tabPanel
source("uiPart2.R",  local = TRUE)$value,
tabPanel("About", icon = icon("file-alt"),
      h2("About the Project"),
      p("This project is made for the bioinformatics BSc degree final year thesis in 2021."),
      p("The web application is built by the R package, Shiny, with the purpose of quantitative data analysis,
      data exploration, calculating the statistics and creating meaningful plots. It also aims to simplify the
       process by providing a graphical user interface (GUI) for the researchers"),
      p("The implemented statistical tools include summary statistics, chi-square test, t-test,
       ANOVA and MANOVA. The application outputs the appropriate plot with the option to download it, either in png or pdf format."),
      div("By: Tanya Deniz Toluay"),
      actionButton("Gitlab",
          label = "Gitlab",
          icon = icon("gitlab"),
```

```
            onclick = sprintf("window.open('%s')", "https://gitlab.com/tanyatoluay")),
        actionButton("Github",
            label = "Github",
            icon = icon("github"),
            onclick = sprintf("window.open('%s')", "https://github.com/tanyatoluay")),
        div("Mentor: Prof. Dr. Uroš Godnov"),
        actionButton("Gitlab",
            label = "Gitlab",
            icon = icon("gitlab"),
            onclick = sprintf("window.open('%s')", "https://gitlab.com/urosgodnov")),
        actionButton("Github",
            label = "Github",
            icon = icon("github"),
            onclick = sprintf("window.open('%s')", "https://github.com/urosgodnov")),
        br(),
        h2("About the Faculty"),
        h3("Univerity of Primorska: FAMNIT"),
        p("The Faculty of Mathematics, Natural Sciences and Information Technologies (UP
FAMNIT)
          is a member of the University of Primorska and was founded in 2006."),
        p("The Faculty offers courses at all three levels of higher education
          and conducts research in the fields of mathematics, computer science and
information technology, and in the natural sciences.
          The Faculty works closely with the University of Primorska, Andrej Marušič
Institute (UP IAM), where most of the academic
          staff of UP FAMNIT are currently conducting research activities."),
        p("The basic activities of the Faculty include education
          and research, while the Faculty is also the organiser and co-organiser of successful
international conferences and other
          scientific meetings."),
        h5("Contact"),
        div("University of Primorska"),
        div("Faculty of Mathematics, Natural Sciences and Information Technologies"),
        div("Glagoljaška 8"),
        div("SI-6000 Koper"),
        div("Slovenia"),
        br(),
        actionButton("FAMNIT",
            label = "FAMNIT: Home",
```

```
                icon = icon("home"),
                onclick = sprintf("window.open('%s')", "https://www.famnit.upr.si/en")),
        actionButton("FAMNIT2",
                label = "FAMNIT: Facebook",
                icon = icon("facebook"),
                onclick = sprintf("window.open('%s')",
"https://www.facebook.com/up.famnit"))
        )
) #tabsetPanel
) #fluidPage
)#shinyUI
```

APPENDIX B *User's Handbook Guide*

Quantitative Analysis

Dataset

The dataset format should be text-based such as csv, txt, tsv or tab. Please indicate the separator of the file as well as if the dataset has quotes and headings. An example dataset can be seen and downloaded on the webpage:
https://gitlab.com/tanyatoluay/finalthesis/-/blob/master/Dummy%20Dataset%20I.txt

Descriptive Statistics Summary

The output shows the number of rows and columns of the dataset, column type frequencies, for character variables shows the column name, missing and complete information rates, the minimum and maximum number of characters, number of unique values and number of white spaces. For the numerical values, it displays the column name, missing and complete information rates, mean and standard deviation values, and p0, p25, p50, p75, p100 values with a small histogram to see the distribution of the data. Comparing mean, standard deviation values with the consideration of the distribution is advised before applying a statistical test.

Inferential statistics

Statistical Tests & Variables

Chi-Square Test

The chi-square test is appropriate for determining the statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. It applies Yate's correction for continuity. The p-value cut off is equal to 0.05. Yate's correction for continuity corrects the error introduced by assuming that a continuous distribution can approximate the discrete probabilities of frequencies in the table. Select the first and second variables as desired columns to apply the statistical test. Select the group same as the second variable.

Such as in the example dataset:

First variable: Smoking

Second variable: Sex

Group by: Sex

Statistical test: Chi-square test

Student's T-Test

The student's t-test is appropriate for determining the statistically significant difference in the mean value between groups and checks if the difference happens by any chance or not. The confidence level is equal to .95. The test is two-sided and uses the Welch (or

Satterthwaite) approximation to the degrees of freedom. Select the first and second variables as desired columns to apply the statistical test regarding data types as numerical and categorical. Next, select the group variable as the second variable or another categorical variable for a further clustered graph.
Such as in the example dataset:
First variable: Weight
Second variable: Group
Group by: Group
Statistical test: T-test

Another Alternative for More Detailed Graphs:
First variable: Weight
Second variable: Group
Group by: Sex
Statistical test: T-test

ANOVA
ANOVA test is appropriate for determining the variation among and between groups, including the differences between means. The test generalises the t-test beyond means. It is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. Select the first variable as numerical, the second and group variable as categorical data.
Such as in the example dataset:
First variable: Weight
Second variable: ExerciseLevel
Group by: ExerciseLevel
Statistical test: Anova test

MANOVA
The MANOVA test is appropriate for comparing multivariate sample means. As a multivariate procedure, it is used when there are two or more dependent variables and is often followed by significance tests involving individual dependent variables separately. Select the first and second variable as numerical, the group variable as a column of categorical variables.
Such as in the example dataset:
First variable: Weight
Second variable: WeeklyExerciseHours
Group by: ExerciseLevel
Statistical test: Manova test

Plots

The output plots are created automatically from the chosen statistical test. The implemented output graphs include:

Chi-square test: mosaic plot and bar chart.

Student's t-test: violin plot, box plot, scatter plot

ANOVA & MANOVA: Q-Q plot, Dot plot

The output plots can be downloaded in png or pdf format. The format of the plot can be chosen in the left column in the data tab.

APPENDIX C *Supplemental Materials*

Code files, dummy dataset, and User's Handbook Guide for the biostatistics web application for quantitative data analysis may be downloaded from: https://gitlab.com/tanyatoluay/finalthesis

The URL of the web application can be reached from: https://tanyatoluay.shinyapps.io/finalthesis/