Statistical analyses and visualization in R (I): Project Report

Oscar Arandes Tejerina
Department of Physics, Stockholm University, AlbaNova University Center, 10691 Stockholm, Sweden

(Dated: January 15, 2026)

This report is part of the course Statistical Analyses and Visualization in R: I (15 credits) at Södertörn University. In the first part of this report, we analyze customer data collected by a Portuguese bank during a campaign between 2008 and 2010, aimed at encouraging clients to subscribe to a term deposit. In the second part, we explore the relationship between various lifestyle factors and sleep health, using a dataset that includes variables such as age, gender, occupation, sleep quality, and the presence of sleep disorders.

CONTENTS

I. PART 1. BANK MARKETING

A. Introduction

Banking plays a crucial role in the global economy, serving as the backbone for financial systems, facilitating transactions, providing loans and managing savings. In today's financial landscape, particularly with the rise of digital banking in recent years, institutions aim to focus more on customer relationships, personalized services and innovative ways to engage clients.

Direct marketing, for instance through phone calls, has been one such strategy conducted by a Portuguese banking institution. The goal in this case was to promote a specific financial product, a bank term deposit. This campaign primarily involved phone calls to potential customers between May 2008 and November 2010. The key outcome variable in the dataset is whether a customer accepted the term deposit offer (labeled as "yes") or rejected it (labeled as "no") [1]. Previous studies have been conducted on this dataset, such as the one in Ref. [2]. In this report, we focus on a closely related dataset available in the Kaggle repository [3].

B. Method

The dataset contains several variables related to customer demographics and banking interactions. Key variables include `age`, `job`, `marital` (divorced, married, single), `education`, and `balance`, which provide insights into the customer's profile. Variables such as `default`, `housing`, and `loan` indicate financial situations, while `contact`, `day`, and `month` describe the communication details of the marketing campaign. Others such as `duration`, `campaign`, and `previous` track the number and length of interactions with the customer, while `pdays` and `poutcome` reflect past contact history and the outcome of previous campaigns. The variable `deposit` indicates whether the customer subscribed to a term deposit.

The preprocessing of the dataset has been minimal. Since only a few missing values were present the corresponding rows were removed. Additionally, some nonsensical values in the `age` feature were also filtered out. Finally, the variable types were adjusted according to the nature of each feature, e.g., the `marital` variable was converted from a *character* to a *factor*. To begin analyzing the dataset, we first ask the following question:

Q1.1: Is there a correlation between a customer's age and their bank balance?

To measure the correlation between two numerical variables, we can compute the *Pearson product-moment correlation coefficient*, denoted as r. This coefficient takes values in the range from -1 to $+1$, where 1 indicates the strongest possible correlation and 0 indicates no correlation. It is important to note that Pearson's correlation assumes a linear relationship between the variables and that both variables follow a bivariate normal distribution. If these assumptions are not met, it is better to use *Spearman's rank correlation* (ρ), which can assess monotonic relationships, whether linear or not [4].

To further analysis the dataset we pose the following question:

Q1.2: Is there an association between client marital status and term deposit subscription?

To address this question, we first transform the data into a *contingency table*, summarizing the frequency of observations between different client marital status and terms deposit subscription. We then perform a Chi-square (χ^2) test. This type of test examines whether the observed counts differ significantly from the expected counts. Since our dataset does not include pre-specified expected frequencies, we will use a χ^2 test for *association* [4]. It is important to note that a χ^2 test is appropriate only if all expected counts are greater than 5 (general rule of thumb). Otherwise, a *Fisher's exact test* should be performed instead. Note that the greater the difference between the observed and expected counts, the larger the resulting test statistic will be. A final technical point is that we will not apply the *Yates' continuity correction*, which is used by default in `R` functions for the χ^2 test. This correction is a small adjustment applied almost exclusively to 2×2 contingency tables, compensating for the fact that the χ^2 distribution is continuous while the data (counts) are discrete. Since our contingency table is larger than 2×2 and the sample size is moderate, applying Yates' correction would unnecessarily make the test too conservative.

C. Result

We can start by reporting the following statistics for the variables `age` and `balance` in TABLE 1.

TABLE 1: Descriptive statistics for `age` and `balance` features.

Variable	**Mean ± SD**	**Median**
Age	40.94 ± 10.62	39
Balance	1362.56 ± 3045.19	449

To answer *Q1.1* we perform an exploratory scattering plot between the variables `age` and `balance`, shown in FIG. 1

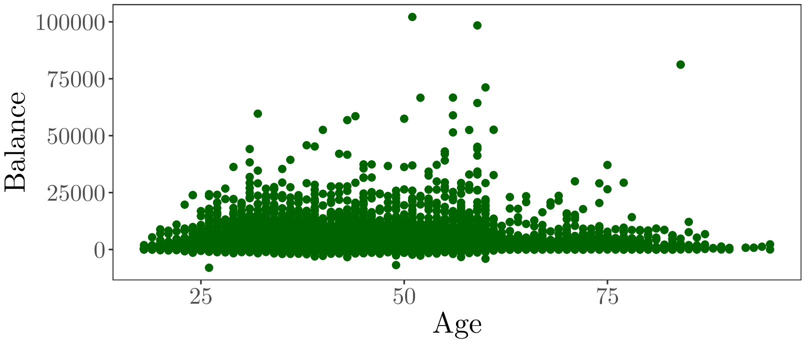

FIG. 1: Scatter plot between the variables `age` and `balance`.

It is clear that the linearity assumption is violated, and there exists no correlation between the variables whatsoever.

To further support this observation, we present the results of a Spearman's rank correlation test in TABLE 2, where a correlation coefficient close to `0` confirms the absence of correlation between the `age` and `balance` features.

TABLE 2: Spearman's rank correlation test between the variables `age` and `balance`.

Test	**S test statistic**	**ρ (Correlation Coefficient)**	**p-value**
Spearman's rank correlation	1.3905×10^{13}	0.0963	$< 2.2 \times 10^{-16}$

We can now address *Q1.2*. Since we are dealing with categorical variables, a bar plot is an effective way to visualize the data. This type of plot helps highlight the frequency or distribution of each category, making it easier compare categories. The bar plot is shown in FIG. 2.

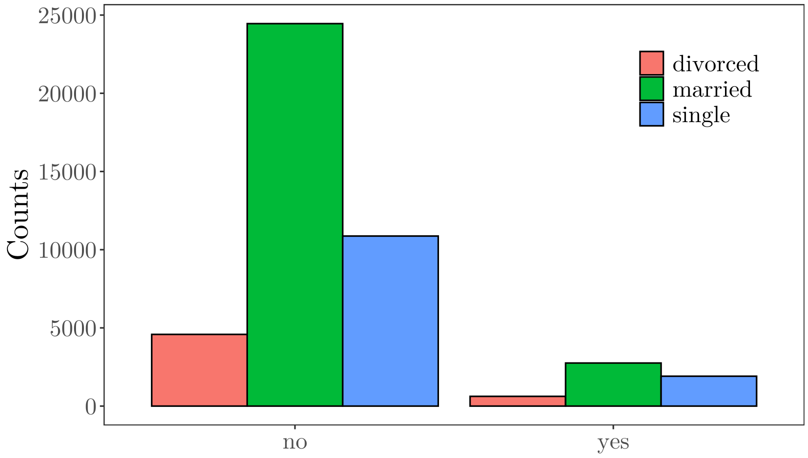

FIG. 2: Bar plot showing the distribution of marital status (`divorced`, `married`, and `single`) alongside whether the client subscribed to the term deposit (`yes`, `no`). This provides a visual representation of the *observed* counts.

After performing a χ^2 test, we can analyze and compare the *observed* and *expected* counts, which are reported in TABLE 3.

TABLE 3: Contingency table of the *observed* and *expected* counts after performing the χ^2 test for association between marital status and deposit subscription.

Marital Status	**No** (Observed)	**Yes** (Observed)	**No** (Expected)	**Yes** (Expected)
Divorced	4584	622	4597.12	608.88
Married	24454	2754	24025.83	3182.17
Single	10872	1910	11287.05	1494.95

By the rule of thumb mentioned in the first part of Section I B, the test is valid for interpretation because all expected frequencies are greater than 5. We can therefore confidently report the results of the test in TABLE 4.

TABLE 4: χ^2 test for association between marital status and deposit subscription.

Test	**χ^2 test statistic**	**df**	**p-value**
χ^2 test for association	196.06	2	$< 2.2 \times 10^{-16}$

We can deepen our understanding of the results by, on one hand, examining the residuals from the χ^2 test, presented in TABLE 5, which reveal the nature of the differences between observed and expected counts through the sign of their contributions. On the other hand, we can visualize the percentage of each individual contribution to the χ^2 test statistic in FIG. 3.

TABLE 5: Standardized residuals from the χ^2 test assessing the association between marital status and deposit subscription.

Marital Status	**No** (Residuals)	**Yes** (Residuals)
divorced	-0.1935	0.5317
married	2.7624	-7.5903
single	-3.9067	10.7347

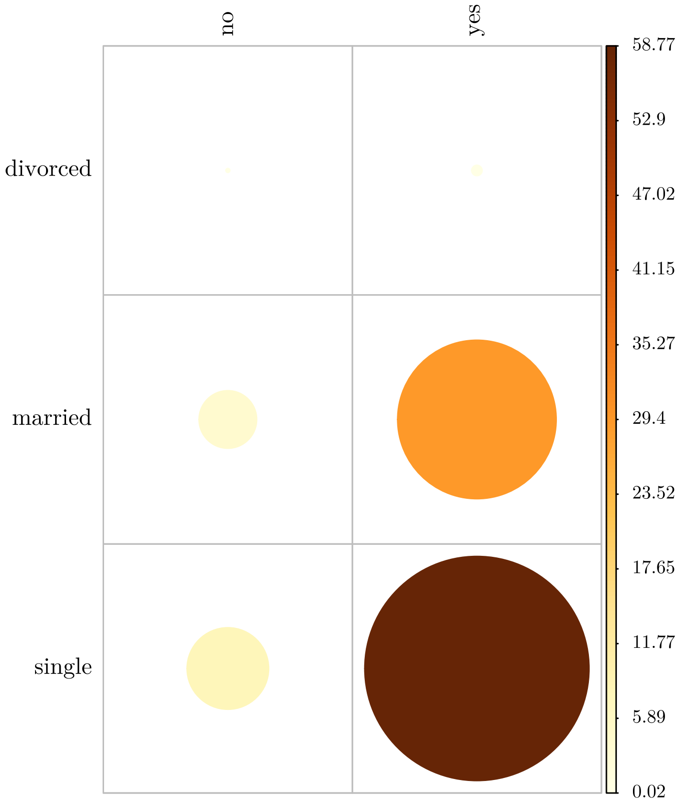

FIG. 3: Percentage of individual contributions to the χ^2 test statistic, illustrating which cells contribute most strongly to the overall deviation from independence.

D. Discussion

We are now ready to address our first research question *Q1.1*. We clearly find that in this dataset, there is no correlation whatsoever between the `age` and `balance` of the customers.

Regarding question *Q1.2*, we find a statistically significant association between clients' marital status (`divorced`, `married`, `single`) and term deposit subscriptions (`yes`, `no`). The larger residuals contributing to the χ^2 statistic, suggesting a significant relationship between marital status and term deposit subscriptions, come from the `single` customers, followed by the `married` ones, especially in the `yes` category. In the `single` group, we observe that customers are much more likely to subscribe to a term deposit (`yes`) than expected. In contrast, the `married` group shows the opposite behavior, with fewer `yes` responses than expected. The `divorced` category shows small residuals, indicating little deviation from the expected values, and suggesting a weak association with term deposit subscriptions.

II. PART 2. SLEEP HEALTH AND LIFESTYLE

A. Introduction

Sleep is a fundamental aspect of human health and well-being. It is during sleep that our bodies repair and regenerate, our brains consolidate memories, and our immune systems strengthen. Sleep health is linked to a range of physical and mental health issues, including cardiovascular disease, obesity or diabetes. In today's fast-paced society, improving sleep health has the potential to enhance quality of life, boost productivity, and prevent chronic health conditions.

In this study, we aim to explore the relationship between sleep patterns and lifestyle factors using a synthetic dataset created for illustrative purposes. This dataset is available for download on Kaggle, as referenced in Ref. [5].

B. Method

The dataset contains information on individuals' sleep quality and lifestyle factors. Key variables include `Age`, `Gender`, `Occupation`, `Sleep.Duration`, `Quality.of.Sleep`, and `Physical.Activity.Level`. It also includes health-related data such as `Stress.Level`, the body mass index (BMI) in `BMI.Category` (normal, overweight, obese), `Blood.Pressure`, `Heart.Rate`, and `Daily.Steps`. Additionally, the dataset captures the presence of `Sleep.Disorders` (none, insomnia, sleep apnea). Minimal preprocessing is applied to the dataset, adjusting the variable types for better consistency. For example, the `BMI.Category` variable was converted from a character to a factor.

In this second part, we start by aiming to answer the following research question:

Q2.1: Is there a significant difference in Sleep Duration between Male and Female participants?

When we want to study how a categorical (independent) variable influences a continuous (dependent) variable, we can use a two-sample t-test. This test helps us compare the means of two groups to see if they are significantly different from each other [4]. It is suitable for situations where the independent variable has two groups, such as `Gender` (Male and Female) in our case. As is typical in these tests, the null hypothesis assumes no difference in the means, while the alternative hypothesis being that there is a difference between the groups.

The assumptions for a t-test include: (1) the dependent variable must be continuous, (2) the observations should be independent, and (3) the dependent variable should be *approximately* normally distributed. The normality of the data can be visually inspected using a histogram or assessed through a *Shapiro-Wilk test*. Additionally, the assumption of *homoscedasticity* (i.e., the variance within each group is equal) must be met. If one or more of these assumptions are violated, a *Wilcoxon rank-sum* test (also called the Mann-Whitney U test) can be used as a non-parametric alternative [4].

To deepen our analysis of the dataset, we will also examine the following:

Q2.2: Does the quality of sleep differ between individuals with different BMI categories or those with different sleep disorders?

To address this question, we need to analyze the differences between BMI categories. The dependent variable is the `Quality.of.sleep`, which is continuous, while the values of the independent variable, the `BMI.category`, are categorical. Note that, in this case, the independent variable has more than two groups (normal, overweight and obese). The appropriate approach is then to perform an F-test using *analysis of variance (ANOVA)* to determine whether there are significant differences in the mean quality of sleep among people in different BMI categories. Here, the null hypothesis states that there is no difference between the group means.

This type of one-way ANOVA relies on the assumptions that the samples are independent, the residuals are normally distributed, and the residuals exhibit *homoscedasticity*, that is, they have constant variance across all levels of the independent variable, which in our case corresponds to the different BMI categories [6]. To assess whether the residuals are normally distributed, one can examine a *Q-Q plot* as part of the diagnostic tools. The homoscedasticity condition, on the other hand, can be evaluated using a *residuals vs fitted values plot*. As a general rule, this condition is considered satisfied if the spread of the residuals in any group does not exceed three times the spread of the group with the smallest residual variance. If those assumptions are not met, one can transform the data to meet the assumptions, or, as in our case, choose to use a non-parametric test such as the non-parametric *Kruskal-Wallis test*.

To correctly interpret the statistical significance of the differences between means, we extend our analysis by performing a post-hoc test. This can be done by the help of the `R` function `pairwise.wilcox.test()` to perform pairwise

Wilcoxon rank-sum tests for each pair of levels within the factor variables (`BMI Category` and `Sleep Disorder`). As a technical note, we apply *Holm's correction* for multiple comparisons, adjusting p-values to control the family-wise error rate. Alternatively, one can use the `kwAllPairsNemenyiTest()` function from the `PMCMRplus` package to perform a *Nemenyi post-hoc test* using *Tukey's distance* as a measure.

C. Result

Let us start with *Q2.1*. A visual inspection of the dependent variable `Sleep.Duration` through a histogram suggests that the feature is not normally distributed (see FIG. 4).

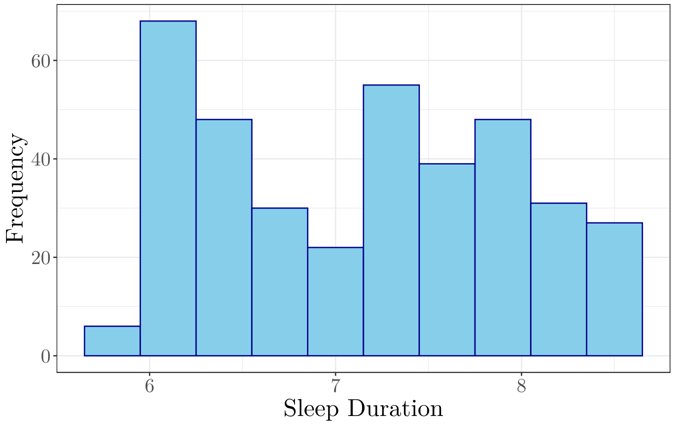

FIG. 4: Histogram of `Sleep.Duration`, showing the distribution of the feature. The shape indicates that the data is not normally distributed.

Consequently, we present the results of the Wilcoxon rank sum test in TABLE 6.

TABLE 6: Wilcoxon rank-sum test for `Sleep.Duration` by `Gender`.

Test	**W statistic**	**p-value**
Wilcoxon rank-sum test	20036	0.0144

We can also visualize the difference in means in FIG. 5.

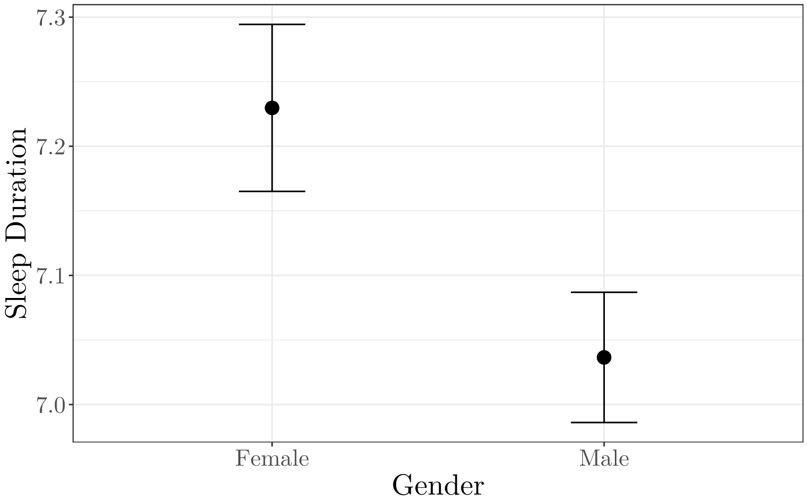

FIG. 5: Plot illustrating the difference in means of `Sleep.Duration` by `Gender`.

Finally, we present the results for *Q2.2*. Here we are going to perform two independent one-way ANOVAs for each categorial variable. Before analyzing the results of the one-way ANOVA test, it is important to first check if the test assumptions are satisfied. The diagnostic plots for the one-way ANOVA model, with `Quality.of.Sleep` as the dependent variable and `BMI.Category` as the independent variable, are shown in FIG. 6.

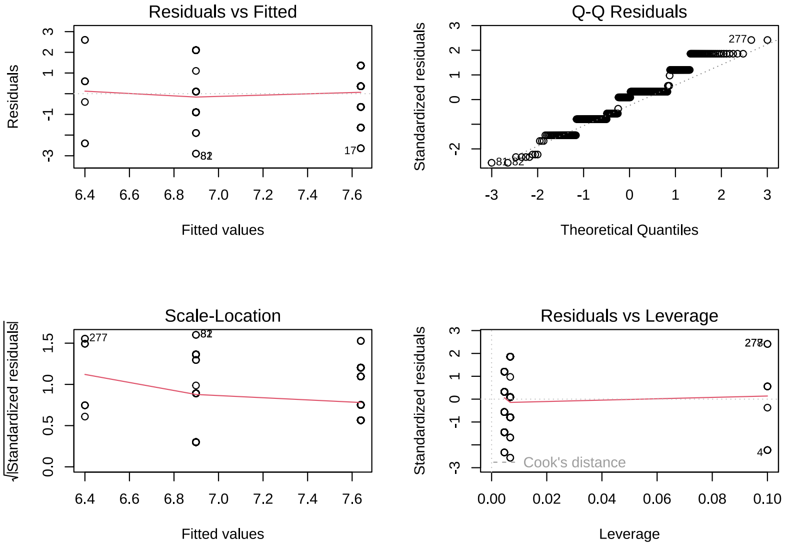

FIG. 6: Diagnostic plots for the one-way ANOVA examining `Quality.of.Sleep` by `BMI.Category`. The Q-Q plot shows that the residuals are not approximately normally distributed.

We can visually inspect that the residuals do not follow a normal distribution. Further evidence is provided by the Shapiro-Wilk test on the residuals (see TABLE 7), which confirms that the normality assumption for the residuals is violated. Therefore, we proceed by using the non-parametric alternative, the Kruskal-Wallis test, to analyze the data instead. We present the results for both the independent variables `BMI.Category` and `Sleep.Disorder`, as both exhibit similar diagnostic behaviors (see TABLE 8). We can extend our analysis and make a pair-wise analysis independently in both `BMI.Category` and `Sleep.Disorder` features. We perform our post hoc tests using both the `pairwise.wilcox.test` (see TABLE 9 and TABLE 10, respectively) and the `kwAllPairsNemenyiTest` functions (see TABLE 11 and TABLE 12, respectively).

TABLE 7: Shapiro-Wilk normality test for the residuals of the one-way ANOVA.

Test	**W statistic**	**p-value**
Shapiro-Wilk normality test	0.94668	2.335×10^{-10}

TABLE 8: Kruskal-Wallis rank-sum test for `Quality.of.Sleep` by `BMI.Category` and `Sleep.Disorder`.

Test	**Kruskal-Wallis χ^2 statistic**	**df**	**p-value**
`BMI Category`	37.895	2	5.904×10^{-9}
`Sleep.Disorder`	49.179	2	2.094×10^{-11}

Finally, these differences can also be visualized using boxplots, which are typically more appropriate for non-parametric tests. The boxplots are shown in FIG. 7.

TABLE 9: Pairwise comparisons using the Wilcoxon rank sum test with continuity correction between `BMI.Category` levels, based on `Quality.of.Sleep`.

Comparison	p-value
Obese - Normal	0.039
Overweight - Normal	6.7×10^{-9}
Overweight - Obese	0.672

TABLE 10: Pairwise comparisons using the Wilcoxon rank sum test with continuity correction between `Sleep.Disorder` levels, based on `Quality.of.Sleep`.

Comparison	p-value
None - Insomnia	8.5×10^{-15}
Sleep Apnea - Insomnia	0.091
Sleep Apnea - None	0.147

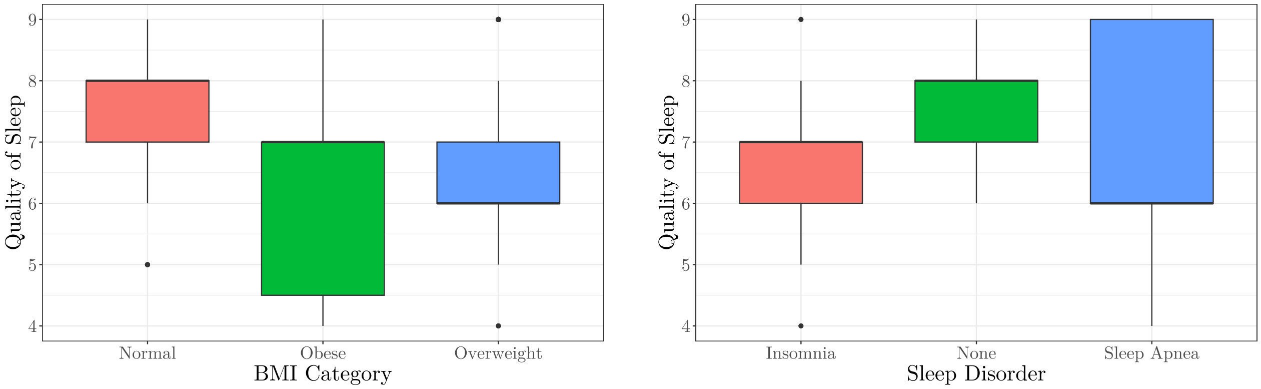

FIG. 7: Boxplots illustrating the differences in `Quality.of.Sleep` for the independent variables `BMI.Category` (left) and `Sleep.Disorder` (right).

D. Discussion

We are now ready to address our research question (*Q2.1*). We find that there is a significant difference in sleep duration between male and female participants, with females having a longer sleep duration compared to males.

Regarding research question (*Q2.2*), we find that the quality of sleep varies across different BMI categories (normal, overweight, and obese) as well as among different sleep disorders (none, insomnia, and sleep apnea). A pairwise comparison in both cases helps refine our analysis. For BMI categories, there is a clear statistically significant difference in sleep quality between the `Overweight` and `Normal` groups. However, the results between the `Obese` and `Normal` categories are somewhat inconclusive. The pairwise Wilcoxon test indicates a significant difference, which aligns more closely with intuition, whereas the Nemenyi's All-Pairs Rank Comparison Test shows no significant difference. Lastly, there is no clear difference between the `Overweight` and `Obese` groups in terms of sleep quality.

Regarding sleep disorders, we find a similar pattern. There are significant differences in sleep quality between `None` and `Insomnia`. However, inconclusive results are observed between `Sleep Apnea` and `Insomnia`, as the pairwise Wilcoxon test shows no difference, while the Nemenyi's All-Pairs Rank Comparison Test indicates a significant difference. Finally, there is no significant difference in sleep quality between `None` and `Sleep Apnea`.

For this dataset, we have used non-parametric tests for the analysis, as the assumptions of normality were not satisfied. While it has been suggested that non-parametric tests are less powerful than their parametric counterparts [7], the large sample size ensures the robustness of our results.

TABLE 11: Pairwise comparisons using the `kwAllPairsNemenyiTest` between `BMI.Category` levels, based on `Quality.of.Sleep`.

Comparison	q statistic	p-value
Obese - Normal	3.169	0.0645
Overweight - Normal	8.166	2.3203×10^{-8}
Overweight - Obese	0.470	0.9409

TABLE 12: Pairwise comparisons using the `kwAllPairsNemenyiTest` between `Sleep.Disorder` levels, based on `Quality.of.Sleep`.

Comparison	q statistic	p-value
None - Insomnia	9.582	3.7176×10^{-11}
Sleep Apnea - Insomnia	5.348	0.0005
Sleep Apnea - None	3.112	0.0711 .

Appendix A: Appendix. R Code

```r
###############################################################
#####---Statistical analyses and visualization in R (I)---#####
##########-----------------Project Work---------------#########
##########-----------Oscar Arandes Tejerina----------##########
###############################################################

#install.packages("tidyverse")
#install.packages("corrplot")
#install.packages("showtext")
#install.packages("psych")
#install.packages("PMCMRplus")
#install.packages("car")
library(tidyverse)
library(corrplot)
library(showtext)
library(psych)
library(PMCMRplus)
library(car)

# Add LM Roman font
font_path <- "K:/STOCKHOLM/Courses/Sodetorns Hogskola/Statistical analyses and visualization in R (I)/latin-modern-roman.mroman10-regular.otf"
font_add("LM Roman", font_path)
showtext_auto()
###############################################################################
##########----------------------------PART 1---------------------------########
###############################################################################
# Read the data
data_bank <- read.csv("bank_dataset.csv")
summary(data_bank)

# Preprocessing
data_bank <- data_bank |>
  mutate(
    Id = as.character(Id),
    age = as.numeric(age),
    job = as.factor(job),
    marital = as.factor(marital),
    education = as.factor(education),
    default = as.factor(default),
    housing = as.factor(housing),
    loan = as.factor(loan),
    contact = as.factor(contact),
    month = as.factor(month),
    poutcome = as.factor(poutcome),
    y = as.factor(y)
  ) |>
  rename(
    deposit = y                   # rename y" to "deposit"
  ) |>
  filter(!(age %in% c(999, -1))   # remove nonsense age values
  )

data_bank_clean <- na.omit(data_bank)

summary(data_bank_clean)

####################
### Question 1.1 ###
####################
# Is there a correlation between a customer's age and their bank balance?

# Statistics of the features (using "describe" function from "psych" library)
describe(data_bank_clean[, c("age", "balance")])

# Scatter Plot
data_bank_clean |>
  ggplot(mapping = aes(x = age, y =balance)) +
  geom_point(color = "darkgreen", size = 2) +
  labs(
    x = "Age",
    y = "Balance"
  ) +
  theme_bw() +
  theme(
    panel.grid = element_blank(),
    aspect.ratio = 0.4,
    text = element_text(size = 20, family = "LM Roman")
  )

# Perform a Spearman's Rank Correlation (non-parametric) test
cor.test(data_bank_clean$age,
         data_bank_clean$balance,
         method = "spearman")




####################
### Question 1.2 ###
####################
# Is there an association between client marital status and term deposit
# subscription?

# Perform a chi-2 test
cont_table <- table(data_bank_clean$marital, data_bank_clean$deposit)
result_chi2 <- chisq.test(cont_table, correct = FALSE) # Remove Yates+
                                                       # continuity correction

# Is the chi-2 test valid? Let's examine the Expected Frequencies
result_chi2$expected

# Explore the Individual Contributions to the chi-2test
round(result_chi2$residuals,4)
round(result_chi2$residuals^2,4)
```

```

# Visualization data
cont_table |>
  as.data.frame() |>
  setNames(c("Row", "Column", "Freq")) |>
  ggplot(aes(x = Column, y = Freq, fill = Row)) +
  geom_bar(stat = "identity", position = "dodge",color = "black") +
  labs(
    title = NULL,
    x = NULL,
    y = "Counts",
    fill = NULL
  ) +
  theme_bw() +
  theme(
    legend.position = c(0.85, 0.8),
    panel.grid = element_blank(),
    aspect.ratio = 0.6,
    text = element_text(size = 20, family = "LM Roman")
  )

# Visualization Individual Contributions to the chi-2test
contrib1 <- (result_chi2$residuals^2/result_chi2$statistic)*100
par(family = "LM Roman")
contrib1 |>
  corrplot(is.cor = FALSE,
           cl.align.text = "l",
           tl.col = "black")


################################################################################
##########-----------------------------PART 2---------------------------########
################################################################################
# Read the data
data_sleep <- read.csv("sleep_dataset.csv")
summary(data_sleep)

# Merge ’Normal’ and ’Normal Weight’ into one level
data_sleep <- data_sleep |>
  mutate(BMI.Category = ifelse(
    BMI.Category == "Normal Weight", "Normal", BMI.Category)
  )

data_sleep_clean <- data_sleep |>
  mutate(Person.ID = as.character(Person.ID),
         Gender = as.factor(Gender),
         Occupation = as.factor(Occupation),
         BMI.Category = as.factor(BMI.Category),
         Sleep.Disorder = as.factor (Sleep.Disorder),
         )

summary(data_sleep_clean)

####################
### Question 2.1 ###
####################
# Is there a significant difference in Sleep Duration between Male and
# Female participants?

# Two-samples t-test (independent samples t-test)?
# Check if the dependent variable (response variable) is normally distributed
data_sleep_clean |>
  ggplot(aes(x = Sleep.Duration)) +
  geom_histogram(bins = 10, fill = "skyblue", color = "darkblue") +
  labs(
    x = "Sleep Duration",
    y = "Frequency") +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )

# They are not so we better perform a Wilcoxon rank-sum test
# (also called the Mann-Whitney U test)
test_sleep <- wilcox.test(data_sleep_clean$Sleep.Duration ~ data_sleep_clean$Gender,
    alternative = "two.sided")

# Visualize difference means
summary_sleep <- data_sleep_clean |>
  group_by(Gender) |>
  summarise(
    mean_sleep = mean(Sleep.Duration),          # Mean
    se_sleep = sd(Sleep.Duration) / sqrt(n())  # Standard error of the mean
  )

summary_sleep |>
  ggplot(aes(x = Gender, y = mean_sleep)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymin = mean_sleep - se_sleep, ymax = mean_sleep + se_sleep),
                width = 0.2
  ) +
  labs(
    x = "Gender",
    y = "Sleep Duration"
  ) +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )

####################
### Question 2.2 ###
####################
# Does the quality of sleep differ between individuals with different BMI
# categories or those with different sleep disorders?

table(data_sleep_clean$BMI.Category, data_sleep_clean$Sleep.Disorder)

model_sleep <- data_sleep_clean |>
  lm(formula = Quality.of.Sleep ~ BMI.Category)

model_sleep_2 <- data_sleep_clean |>
  aov(formula = Quality.of.Sleep ~ BMI.Category)

summary(model_sleep)
summary(model_sleep_2)

# Is the model valid?
par(mfrow=c(2,2))
plot(model_sleep)                    # Graphical inspection
shapiro.test(residuals(model_sleep)) # Shapiro-Wilk test on the residuals

# Note that quality of sleep is already not normally distributed!
shapiro.test(data_sleep_clean$Quality.of.Sleep)
```

```

# It seems that normality is not satisfied. We perform the non-parametric
# version, i.e., the Kruskal-Wallis test for both variables separately
kruskal.test(Quality.of.Sleep ~ BMI.Category, data=data_sleep_clean)
kruskal.test(Quality.of.Sleep ~ Sleep.Disorder, data=data_sleep_clean)

# Perform a post hoc test (pairwise comparison) with "pairwise.wilcox.test" function
pairwise.wilcox.test(data_sleep_clean$Quality.of.Sleep,
                     data_sleep_clean$BMI.Category,
                     p.adjust.method = "holm")
pairwise.wilcox.test(data_sleep_clean$Quality.of.Sleep,
                     data_sleep_clean$Sleep.Disorder,
                     p.adjust.method = "holm")

# Perform a post hoc test (pairwise comparison) with "PMCMRplus" package
kruskal_posthoc_bmi <- kwAllPairsNemenyiTest(
                       data_sleep_clean$Quality.of.Sleep ~ data_sleep_clean$BMI.Category,
                       dist="Tukey",
                       data=data_sleep_clean)
summary(kruskal_posthoc_bmi)

kruskal_posthoc_disorder <- kwAllPairsNemenyiTest(
                            data_sleep_clean$Quality.of.Sleep ~ data_sleep_clean$Sleep.Disorder,
                            dist="Tukey",
                            data=data_sleep_clean)
summary(kruskal_posthoc_disorder)

# Summarize the data by BMI.Category to calculate mean and standard error of the mean
summary_bmi <- data_sleep_clean |>
  group_by(BMI.Category) |>
  summarise(
    mean_sleep = mean(Quality.of.Sleep),         # Mean of Quality of Sleep
    se_sleep = sd(Quality.of.Sleep) / sqrt(n())  # Standard error of the mean
  )

# Visualize difference in means
summary_bmi |>
  ggplot(aes(x = BMI.Category, y = mean_sleep)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymin = mean_sleep - se_sleep, ymax = mean_sleep + se_sleep), width = 0.2) +
  labs(
    x = "BMI Categoryr",
    y = "Quality of Sleep"
  ) +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )

# Visualize difference in medians
ggplot(data_sleep_clean, aes(x = BMI.Category, y = Quality.of.Sleep)) +
  geom_boxplot(aes(fill = BMI.Category), width = 0.7) +
  labs(
    x = "BMI Category",
    y = "Quality of Sleep"
  ) +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )

# Summarize the data by Sleep.Disorder to calculate mean and standard error of the mean
summary_disorder <- data_sleep_clean |>
  group_by(Sleep.Disorder) |>
  summarise(
    mean_sleep = mean(Quality.of.Sleep),          # Mean of Quality of Sleep
    se_sleep = sd(Quality.of.Sleep) / sqrt(n())   # Standard error of the mean
  )

# Visualize difference in means
summary_disorder |>
  ggplot(aes(x = Sleep.Disorder, y = mean_sleep)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymin = mean_sleep - se_sleep, ymax = mean_sleep + se_sleep),
                width = 0.2
   ) +
  labs(
    x = "Sleep Disorder",
    y = "Quality of Sleep"
  ) +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )

# Visualize difference in medians
ggplot(data_sleep_clean, aes(x = Sleep.Disorder, y = Quality.of.Sleep)) +
  geom_boxplot(aes(fill = Sleep.Disorder), width = 0.7) +
  labs(
    x = "Sleep Disorder",
    y = "Quality of Sleep"
  ) +
  theme_bw() +
  theme(
    aspect.ratio = 0.6,
    legend.position = "none",
    text = element_text(size = 20, family = "LM Roman")
  )
```

[1] S. Moro, P. Rita, and P. Cortez, Bank marketing, UCI Machine Learning Repository (2014).
[2] S. Moro, P. Cortez, and P. Rita, Decision Support Systems **62**, 22 (2014).
[3] A. Ahmedov, Predict term deposit, Kaggle (2021).
[4] N. J. Salkind, *Statistics for People Who (Think They) Hate Statistics*, 6th ed. (SAGE Publications, 2016).
[5] L. Tharmalingam, Sleep health and lifestyle dataset, Kaggle.
[6] L. Ståhle and S. Wold, Chemometrics and Intelligent Laboratory Systems **6**, 259 (1989).
[7] E. Whitley and J. Ball, Critical Care **6**, 509 (2002).