**Question 1.**

**a)**

*(note: the first row of this table is a fictional example—not based on a variable in our actual dataset)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Variable name** | **Data type that R incorrectly assigned** | **Data type that R SHOULD HAVE assigned** | **R code to assign proper data type** |
| Example | time1 | integer | character | attach(practicedata)  #treataschar() is a fictional function, i.e. it won’t work  treataschar(time1) |
| Variable 1 | participant | integer | numeric | heartrate <- as.integer(heartrate) |
| Variable 2 | drinking | integer | factor | Drinking <- as.factor(drinking) |

**b) Variable:** weight

**c) Reason:** The outliers probably exist because the distribution of the weights of samples is more extreme than normal (i.e. those are random variation within the sampling frame).I suspect that as my reason as the data that contains those outliers are consistent for the 3 time-points of the same participant, suggesting a small likelihood of reporting mistake. Besides, although rare, those data are still within the normal weight range of humans based on my personal experience.

**d) Method:** After check the original datasheet, I would find whether those outliers are reporting mistakes or random variation. If those are reporting mistakes, I can correct them with the original data obtained. If not, I will drop the outlier record and conduct my analysis without those values.

**e) R code(s) + output:** R code: boxplot.stats(labdata$weight)$out output: [1] 74.6 75.8 75.1 74.8 74.9

**Question 2.**

**a) format:** long format

**elements:** multiple rows for each subject (aka. participant) to store their data

**b) mock-up: A screen shot of a computer

Description automatically generated**

**Question 3.**

**a) table:**

*(note: the first row of this table is a fictional example (not based on a variable in our actual dataset) that addresses that the range of intelligence (via IQ score) in the Japanese population is from 98 to 157)*

|  |  |  |  |
| --- | --- | --- | --- |
| **characteristic** | **Statistic value(s) from sample** | **Description of statistic(s)** | **R code used to compute statistics** |
| Intelligence | Minimum: 89  Maximum: 123  Range: 34 | The range of intelligence for the sample of adults is 34 IQ units (with minimum IQ of 89 and maximum IQ of 123). | attach(practicedata)  #iqscore is the fictional variable in my fictional dataset, “practicedata”  #I used summary() to find the max and min iq scores  summary(iqscore)  #compute the range of the iq scores  range(iqscore) |
| Gender | Female: 90.2%  Male: 9.8% | 90.2% of the sample is female. | # proportions of the two sex are saved in a table called protortions\_sex  proportions\_sex <-  table(labdata$sex)/length(labdata$sex)  # the proportions calculated are turned into percentages by multiplying them by 100  percentages\_gender <- proportions\_sex\*100 |
| Age | Age > 65: 7.84%  Age < 65: 92.16% | 7.84% of the sample is 65 years or older. | # proportion of people who is 65 or older is calculated  proportions\_age <- sum(labdata$age > 65)/length(labdata$age)  # the proportion calculated are turned into percentage by multiplying by 100  percentages\_gender <- protortions\_age \*100 |
| Sports | Number of people habitually participates in exercise based on their age:  (20, 30]: 18  (30, 40]: 27  (40,50]: 31  (50,60]: 27  (60,70]: 19  (70,80]: 1 | The number of people that habitually participate in exercise is higher for lower age groups. | # make a subset with all the people that habitually participate in exercise (i.e. have their exercise data as 2)  exercisedata <- subset(labdata, labdata$exercise==2)  # label breaks with an interval of 10  breaks <- c(20,30,40,50,60,70,80)  # cut the sample with exercise habit based on their age by the breaks  exercise\_t <- cut(exercisedata$age, breaks=breaks)  # make a summary out of it  summary(exercise\_t) |

**b) Compare and discuss:** The sample taken has consistent characteristics with the general Japanese population in terms of the relationship between the frequency of participation in sport and age. For the gender can age components, however, this sample varies quite a lot from our population of interest. The sampling error is probably caused by the narrow sample frame, small sample size and the sampling method used.

The sampling frame of this study is those who can read flyers and newspapers that live in communities in/near Matsuyama. This would leave out those who are too young to read or participate in the study and thus may lower the percentage of age 65- participants in this study.

The sample, which is made up of 51 (57 – 2 – 4) individuals, is obviously a rather small sample considering the number of people in the entire Japanese population. The sample error could have been reduced if more people are studied and that may make the results more consistent with the general population. As the sample variation exists, such a small sample may also be a naturally occurred “outlier” and have those statics different from the whole population.

As voluntary response sampling was conducted in this study, it could be more biased. Since this study is related to aromatherapy foot massage, which is generally considered to be more attractive to women, males may be less likely to volunteer and thus are undercovered for the sample.

**Question 4.**

**a) graph plus figure legend:** (*use space needed so your graph is legible*)

**A screenshot of a cell phone

Description automatically generated**

**Fig 1.** Body mass index (BMI) obtained from subjects directly before beginning the aromatherapy foot massage compared with sex and treatment group.

**b) R code:** # R code to see if the body mass index at baseline differ within the individuals in the study or differ across sexes

# create a subset that contains data took directly before beginning the aromatherapy foot massage from the two treatment groups

baselinedata <- subset(labdata, (labdata$group == "A" & labdata$timepoint == "baseline")|(labdata$group == "B" & labdata$timepoint == "week 4"))

# calculate the BMI for each participant based on the baselinedata

bmi <- baselinedata$weight/baselinedata$height\*\*2

# create a data frame containing information of the treatment group, gender, and BMI for each participant

bmidata <- data.frame(group = baselinedata$group, sex = baselinedata$sex, bmi=bmi)

# make a boxplot based on the data frame bmidata

boxplot(bmi~sex:group, data = bmidata, ylab = "BMI kg/m^2", col = c("#FCE6E8","#B9DFFE"), at = c(1,2,4,5))

**c) explanation:** Sex and treatment groups are two independent variables I used because those are the factors being asked in the question. As the timepoint is described as “directly before beginning the aromatherapy foot massage treatment” in the question, data recorded at baseline and week 4 are selected for treatment groups A and B, respectively. To reduce the complexity of the data frame used for plotting, a subset containing only group treatment, sex, and BMI was then created. As the graph would show the association between two categorical and one quantitative variable (sex: categorical; treatment group: categorical; BMI: quantitative), both side-by-side boxplot and strip chart should be appropriate for the graph. I chose the boxplot because it can provide more information on the median and the range, which better demonstrates the dispersion of the data and thus seems to be more appropriate to the question.