Block 4 Week 5: Fitness & Health Report

1 Introduction

The data available at https://uoepsy.github.io/data/fitnessProgram.csv were provided by a Personal Trainer who specializes in fitness programs, specifically running programs. Their dataset contained information on 8 variables - the number of days their clients have been enrolled on the program (`daysInProgram`), the maximum distance they can run (in miles; `maxDistance`), their base fitness level (`baseFitness`), their cardiovascular health rating (`cvHealth`), their average resting pulse rate (`pulse`), their age (in years; `age`), their difficulty rating of the program (on a scale of 1-5; `Difficulty`), and their overall rating of satisfaction with the program (on a scale of 1-5; `Satisfaction`).

1.1 Research Questions

- RQ1: Is there a significant association between the number of days one spends in the program and one's cardiovascular health?
- RQ2: Is there a significant association between satisfaction and difficulty ratings?

2 Analysis

2.1 Research Question 1

Total time enrolled in the program was moderately positively associated with cardiovascular health, and this association was statistically significant ($r = .46, t(498) = 11.64, p < .001$). This association is visually presented in Figure 1. These results suggested that a greater number of days in the program was positively associated with better cardiovascular health.

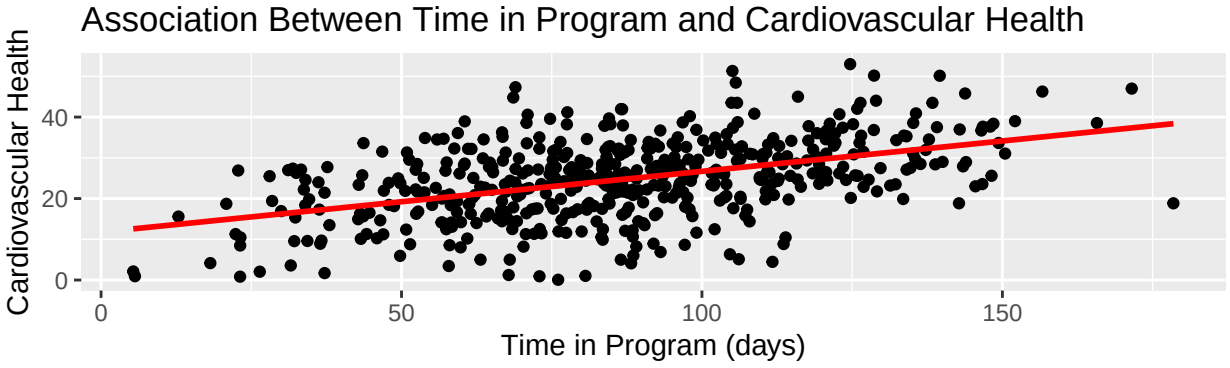

Figure 1: Association Between Time in Program and Cardiovascular Health

Both days in program ($W = 0.998, p = .725$) and cardiovascular health ($W = 0.996, p = .163$) were normally distributed. From Figure 1, we can see that there are no extreme outliers, the association between the two variables is linear, and there is no evidence of heteroscedasticity.

2.2 Research Question 2

We used Spearman's ρ to determine whether there was a significant association between difficulty and satisfaction ratings. We found a significant weak negative association between difficulty and satisfaction ($\rho = -.21, p < .001$), such that those who found the program more difficult reported lower satisfaction with the program overall.

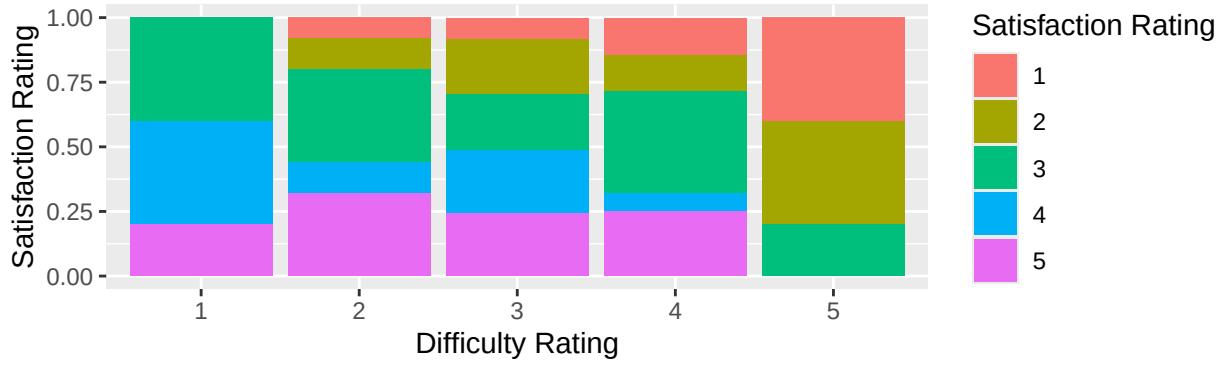

Figure 2: Association Between Difficulty and Satisfaction

3 Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
######## LOAD LIBRARIES & DATA ########
library(tidyverse)
library(kableExtra)
library(psych)
library(patchwork)
library(kableExtra)


dat <- read_csv("https://uoepsy.github.io/data/fitnessProgram.csv")

###### DATA CHECKS ######

str(dat)
summary(dat)


###################### RQ1 ########################

####### ASSUMPTION CHECKS #######

# Days in programme
shapiro.test(dat$daysInProgram)

hd_viz_time <- ggplot(dat, aes(x = daysInProgram, y = after_stat(density))) +
  geom_histogram() +
  geom_density()
```

```r
hd_viz_time

q_viz_time <- ggplot(dat, aes(sample = daysInProgram)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "QQPlot - Days in Program")
q_viz_time

#CV health
shapiro.test(dat$cvHealth)

hd_viz_cv <- ggplot(dat, aes(x = cvHealth, y = after_stat(density))) +
  geom_histogram() +
  geom_density()
hd_viz_cv

q_viz_cv <- ggplot(dat, aes(sample = cvHealth)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "QQPlot - CV Health")
q_viz_cv

####### VISUALISATION #######

#Because we are interested in running a hypothesis test on the correlation between `daysInProgram` and

viz_time_cv <- ggplot(dat, aes(x = daysInProgram, y = cvHealth)) +
  geom_point() +
  geom_smooth(method = 'lm', colour = 'red', se = F) +
  labs(x='Time in Program (days)', y = 'Cardiovascular Health', title = "Association Between Time in Pr

####### CORRELATION #######

cor(dat$daysInProgram, dat$cvHealth)

cor.test(dat$daysInProgram, dat$cvHealth,
         alternative = "two.sided")

###################### RQ2 ########################

####### VISUALISATION #######

# we can see that a scatterplot is not informative for Likert scale data:
ggplot(dat, aes(x = Difficulty, y = Satisfaction)) +
  geom_point() +
  geom_smooth(method = 'lm', colour = 'red', se = F)

#boxplots or barplots may be more informative

ggplot(dat, aes(x = factor(Difficulty), y = Satisfaction)) +
  geom_boxplot()

viz_diff_sat <- ggplot(dat, aes(x = factor(Difficulty), fill = factor(Satisfaction))) +
```

```
  geom_bar(position = "fill") +
  labs(x = "Difficulty Rating", y = "Satisfaction Rating", fill = "Satisfaction Rating")
viz_diff_sat

####### CORRELATION #######

cor.test(dat$Difficulty, dat$Satisfaction,
         alternative = "two.sided",
         method = "spearman")

#To suppress warning message, use the following `exact = FALSE` argument witin cor.test

###################### BONUS CONTENT ########################


#There are a few ways that you can plot correlation data. When you are looking at many correlation valu

#Instead of running separate correlations for each variable pair, it's much similar to create a correla

corDat <- round(cor(dat[1:5]),2)
corDat

####### CORR PLOT #######

library(corrplot)

#You can make a correlogram with the numeric correlation values:
corrplot(corDat, method = 'number')

#or with representative colors:
corrplot(corDat, method = 'color')

#You can mix numbers and colors with the `corrplot.mixed` function. I've also added the `tl.col` argume
corrplot.mixed(corDat, lower='number', upper='color', tl.col='black')

#the text is cut off in the diagonal, so I've added the `tl.pos` argument to set the text position to t
corrplot.mixed(corDat, lower='number', upper='color', tl.col='black', tl.pos = 'lt')

####### CORRR #######

library(corrr)
# Need to use dataset - so using columns 1-5 from dat dataset in this example

#Network plot - variables that are more highly correlated appear closer together and are joined by stro
dat[1:5] |>
  correlate() |>
  network_plot()

#shave() retain only one part of correlation matrix - upper or lower
#rplot() the correlations with shapes in place of the values - bigger shapes = larger associations
dat[1:5] |>
  correlate() |>
  shave() |>
```

```
  rplot()


# present in well formatted table
dat[1:5] |>
  correlate() |>
  shave() |>
  fashion() |>
  kable(caption = "Correlations", digits = 2) |>
  kable_styling()

viz_time_cv
viz_diff_sat
```