April 30, 2023

The results below are generated from an R script.

````
---
title: "Assignment 05 and Student Survey"
author: "Madhavi Ghanta"
date: '2023-04-28'
output:
pdf_document: default
html_document: default
word_document: default

---
# --- Assignment 05 ---

```{r include = FALSE}
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/mghan/Documents/dsc520")

## Load the data "data/r4ds/heights.csv"
heights_df <- read.csv("data/r4ds/heights.csv")


tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
```
## Using 'cor()' compute correlation coefficients
Height Correlations
```{r include = TRUE}
cor(heights_df$height, heights_df$earn)
cor(heights_df$age, heights_df$earn)
cor(heights_df$ed, heights_df$earn)
```
Tech Spending and Suicide Correlation
```{r include= TRUE}
cor(tech_spending, suicides)
```

# --- Student Survey ---
```{r include = FALSE}
library(ggplot2)
library(ggm)
setwd("C:/Users/mghan/Documents/dsc520")
ss_df <- read.csv("data/student-survey.csv")
```
## Part I
````

A. Calculate covariance of survey variables
```{r include= TRUE}
cov(ss_df[, c(1:4)])
```
B. Why use this calculation?
Covariance in R is calculated by using the cov() function

C. What do the results indicate?
1.Time of reading is negatively related to Time of watching TV,
2.Time of reading is negatively related to Happiness.
3.Time of watching TV is positively related to Happiness.
4.As gender is represented as integer, we can ignore the covariance associated with gender.

Part II
A. What measurement is being used for the variables?
TimeReading: time, in hours (rounded to whole hr)
TimeTV: time, in minutes (rounded to nearest 5 min)
Happiness: looks like percent. It looks like Happiness index varies from 45.67% to 89.52%
Gender - By looking at the values, I assumed that the Gender is measure in boolean.It is not specified t

B. Explain what effect changing the measurement being used for the variables
would have on the covariance calculation.
Because time is measured in two different ways, out initial covariance calculation
is not accurate in terms of calculation, but it should still be accurate in terms
of showing positive or negative covariance.

C. Would this be a problem? Explain and provide a better alternative if needed.
Differences in how a type of variable is calculated can be problematic. This is
true not only for time (min v hours) but also for length (in v ft or imperial v metric)
and any other measurement. We should compare like to like to ensure accuracy.

Here we have the corrected covariance after altering the data so that all time
is represented in minutes:
```{r include = FALSE}
TimeReadingMin <- ss_df$TimeReading*60
ss_edit_df <- cbind(ss_df[-c(1)], TimeReadingMin)
ss_edited_df <- ss_edit_df[,c(4,1:3)]
```
```{r include = TRUE}
cov(ss_edited_df[, c(1:4)])
```

Part III
A. Choose the type of correlation test to perform.
B. Why this test?
C. Make a prediction as to whether or not yield +/- correlation.

checking normality of data

```{r include = FALSE}
ggplot(ss_edited_df, aes(sample=TimeReadingMin)) + stat_qq() + stat_qq_line()
ggplot(ss_edited_df, aes(sample=TimeTV)) + stat_qq() + stat_qq_line()
ggplot(ss_edited_df, aes(sample=Happiness)) + stat_qq() + stat_qq_line()
```

```
By looking at plots, I can confirm that data is normally distributed.
We can used Perason's correlation coefficient to check the correlation between variables.

```{r echo = FALSE}
cor.test(ss_edited_df$TimeReadingMin,ss_edited_df$TimeTV)
cor.test(ss_edited_df$TimeReadingMin,ss_edited_df$Happiness)
cor.test(ss_edited_df$Happiness,ss_edited_df$TimeTV)
```

Part IV

Correlation analysis of:
A. All variables
```{r echo = FALSE}
cor(ss_edited_df[, c(1:4)])
```
B. A single correlation between two of the variables
```{r echo = FALSE}
cor(ss_edited_df$TimeReadingMin, ss_edited_df$TimeTV)
```
C. Repeat above, but set confidence interval at 99%
```{r echo = FALSE}
cor.test(ss_edited_df$TimeReadingMin, ss_edited_df$TimeTV, conf.level = 0.99)
```
D. Describe what the calculations in the correlation matrix suggest about the relationship between the v
The variables TimeReadingMin and TimeTV are negatively correlated.
That is, the more time students spend reading the less time they spend watching TV.

Part V
A. Calc correlation coefficient and coefficient of determination.
Earlier, we looked at correlation between the variables TimeReadingMin and TimeTV and found they were ne
```{r echo = FALSE}
cor(ss_edited_df)
```
```{r echo = FALSE}
cor(ss_edited_df)^2
```

B. Describe what you conclude about the results.
The coefficient of determination is a measurement used to explain how
much variability of one factor can be caused by its relationship to another related factor. This correla

Looking at coeffiecient of determination between TimeTV and Happiness shared variability is about 40% wh

Part VI
A. Based on analysis, does watching more TV cause students to read less?
```{r echo = FALSE}
cor(ss_edited_df$TimeReadingMin, ss_edited_df$TimeTV)^2
```
Looking at coefficient of determination (r^2) we can say that variability in TimeReading can cause upto
There could be other variables that may cause 23% variability in TimeTV.
```

```
# Part VII
A. Pick 3 variable and perform a partial correlation.
I will select TimeReadingMin, TimeTV, and Happiness. Run partial correlation between TimeTV and Happines
```{r echo = FALSE}
ss_edited_df2 <- ss_edited_df[,1:3]
```

B. Be sure to document which variable you are "controlling."
Run partial correlation between TimeTV and Happiness while controlling TimeReading
```{r echo = FALSE}
pcor(c("TimeTV","Happiness","TimeReadingMin"), var(ss_edited_df2))
pcor(c("TimeTV","Happiness","TimeReadingMin"), var(ss_edited_df2))^2
```

C. Does this change your interpretation? How, or why not?
If we keep TimeReading controlling , the correlation coefficient between TV time and happiness decrease

Error: attempt to use zero-length variable name
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggm_2.5       ggplot2_3.4.1
##
## loaded via a namespace (and not attached):
##  [1] pillar_1.9.0     compiler_4.2.2   highr_0.10       tools_4.2.2      digest_0.6.31
##  [6] evaluate_0.20    lifecycle_1.0.3  tibble_3.2.1     gtable_0.3.3     pkgconfig_2.0.3
## [11] rlang_1.1.0      igraph_1.4.2     cli_3.6.1        rstudioapi_0.14  yaml_2.3.7
## [16] xfun_0.38        fastmap_1.1.1    withr_2.5.0      dplyr_1.1.1      knitr_1.42
## [21] generics_0.1.3   vctrs_0.6.1      grid_4.2.2       tidyselect_1.2.0 glue_1.6.2
## [26] R6_2.5.1         fansi_1.0.4      rmarkdown_2.21   purrr_1.0.1      farver_2.1.1
## [31] magrittr_2.0.3   scales_1.2.1     htmltools_0.5.5  rsconnect_0.8.29 colorspace_2.1-0
## [36] labeling_0.4.2   tinytex_0.45     utf8_1.2.3       munsell_0.5.0

Sys.time()

## [1] "2023-04-30 13:42:55 PDT"
```