Non-Graded Assignment 2

Cafer Bakac

11 05 2023

Description of the Exercises

This week we have dealt with data preparation and visualization and the exercises will also be about the respective points. In the process of solving the exercises, we will use the data I have collected from you. As you might remember, I collected data from you via a survey platform called Soscisurvey and have previously created an API key to extract the data from the platform. As some of you have indicated problems with soscisurvey package, instead of using the package to get data, you will work with the data I uploaded (and am constantly updating) the data on Moodle. Please download the data and load it on your devices with the following line of code. Make sure that you either set your working directory where the data are located at or download the data into your working directory where your r code is located at.

I am also providing rhe code to get the data using the soscisurvey so that those of you who are interested in using the API to get the data could do so.

I want you to work with ggplot and plotly package for data visualization to gain some experience in the package as I think package could be very handy for you guys to create 'beautiful' plots. In the following we will first of all get the data.

– get all of the libraries needed

```r
library(soscisurvey)
library(plotly)
library(ggplot2)
library(dplyr)
```

– get the data (run it as code). This is the soscisurvey way of getting the data as you already know.

```r
# This is an API and using this, one can address where the data is located.
path <- "https://www.soscisurvey.de/RCOURSESS23/?act=aicmMlFDWe9TWBfEwRscP8mG"
# this reads the data from the path we have specified above.
data <- read_sosci(path)
```

this is the read.csv() way of getting the data

```r
# read the data from the csv file on your working directory.
data <- read.csv("Soscisurvey_Data.csv", fileEncoding = "UCS-2LE")
```

Data Cleaning, Processing and Exploration

– I want you to check the data and see that some columns (e.g., the first four columns and some columns at the end of the dataset) are not needed to be kept in our dataset as they are not related to the constructs that we wanted to measure. The only columns we are interested in are AM01_01, AM01_02 … UMS_30. Exclude all of the other columns from the dataset and keep only the column names specified above in the dataset.

– Now that you have cleaned the dataset, we know which items we have in the dataset. Please be reminded from the classes that the items CS01_01, CS01_02,CS01_03 … CS01_13 measure Affective Preferences, CS02_01, CS02_02,CS02_03 … CS02_13 measure Cognitive Preferences and finally, CS03_01, CS03_02,CS03_03 … CS03_08 measure Perceived Abilities. Also, realize in the dataset there are items such as AM01_01, AM01_02, AM01_03…. AM01_12 to measure Academic Motivation. What I want you here is to change the column names to more understandable ones. Use the following names for renaming: for Affective Preferences: AffPreference_1, AffPreference_2 and AffPreference_3; for Cognitive Preferences: CogPreference_1, CogPreference_2 and CogPreference_3; for Perceived Abilities: PerAbilities_1, PerAbilities_2 and PerAbilities_3; finally, for Academic Motivation: AcademicMotivation_1, AcademicMotivation_2 … AcademicMotivation_12. Also, realize that there is a column named as SC01. Find the position of this column using grep() function and change the column name to InformedConsent.

– I want you, here, to store the items corresponding to different constructs in different datasets (i.e. for affective preferences, cognitive preferences, perceived abilities and academic motivation). These datasets should not include anything other than the items for a specific construct. After doing so, I want you to check the descriptive statistics of these scales. Check if there is any item that has a low mean and if so, you might checked if it is a reverse coded (i.e., negatively worded item) and if so, you can (for now) exclude that item from the main dataset (if there are any).

– Using sapply package I want you to detect if there are any missing values across all of the items.

it seems there are some missing data in all of the items. It might be the case that some participants did not answer any of the items and could be excluded. Please check the data visually. However, for the sake of this data, let us kick out the datapoints with missing values based on the first cognitive preferences item.

– If there are no missing values, in the main dataset, take the mean of items that measure certain constructs and store the values in variables in accordance with construct names (e.g, affective preferences (AffPreference_1, AffPreference_2 …. AffPreference_13) items should be averaged and the result should be stored in a variable with a name like AffPreference_Mean)

– Using piping function in dplyr package, create a pipe that starts from the data, groups the data by gender and using summarise take the mean, standard deviations, median, min and max values of AffPreference_Mean and AcademicMotivation_Mean Store the results in a r object and inspect the results by opening it up. Compare the means and standard deviations across gender.

—- also, you might consider writing the results from these into a .txt file on your working directory

Data Visualization

– Create a boxplot with AcademicMotivation_Mean on y-axis and gender set to color in plotly. Realize that there is only one box for gender but we were expecting two: one for each gender. Why could that be the reason? Think about how the gender data is? Is it numeric or something else? See what happens if transform gender variable into a factor variable and replot. Compare the differences.

– The gender variable includes 1s and 2s. Change 1s to Female and 2s to Male.

– Using the plot above, change the colors of the boxes to black for females and grey for males. Also, name the x-axis as "Gender" and y-axis as "Achievement Motive".

– Using a scatterplot, check the relationship between AffPreference_Mean and AcademicMotivation_Mean. AcademicMotivation_Mean should be plotted on y-axis and AffPreference_Mean on x-axis. Also, add gender as a symbol variable so that the points are shown with separate shapes for males and females. Change the axis labels accordingly.

– Using help from ggplot package, for each gender add an lm() line to a scatterplot like the one above so that it shows the strength of the relationship between AffPreference_Mean and AcademicMotivation_Mean. differently for males and females.

– Create the relationship between AffPreference_Mean and AcademicMotivation_Mean having gender variable as facet. That means, there will be different plots for different genders that show the relationship between the variables. You could get help from ggplot and/or use dplyr package for solving this problem. See, for example, https://stackoverflow.com/questions/58103196/how-to-facet-a-plot-ly-chart