 Due by 9:30am EST, 2^{nd} August

CMPSC 301
Data Analytics
Summer 2021

Lab 2: Analysis of Data From Psychology

Please submit your work to your GitHub repository by the due date.

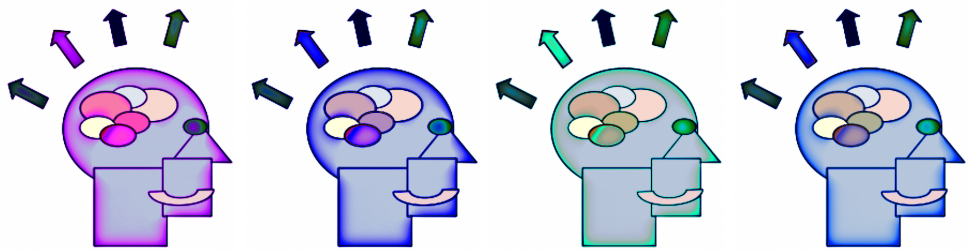

Figure 1: In psychology, data about what we feel, how we think and what makes us react, comes from personality testing. The results of such a tests may enable one to get to know oneself better and to allow one to understand why decisions and actions were taken. In a personality test, simple random-looking questions are put to participants and data is gained by their responses to the many questions. The data that is then processed by statistical tools and rigorous analysis so that conclusions can be drawn. By this form of analysis, investigations in Psychology help us understand some of the issues related to feelings, thoughts, moods so that one can enjoy better health by making better choices.

Objectives

To determine a how to use correlations, t-tests and linear models during research of data from psychological investigation. To gain experience using the computational methods from the R `Psych` library package. In particular, to be able to research how to use new R-statistics software packages and apply them to particular contexts for which they were designed. To extract knowledge from the produced visualizations and extracted interpretation of results.

Reading Assignment

Please review the technical manual for the *Psych* package for use with R. It can be found at the following link: `http://personality-project.org/r/overview.pdf`. This document will be used to guide your coding for testing and visualization.

GitHub Starter Link

https://classroom.github.com/a/PQ8grhRq

Handed out: 27^{th} July 2021

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.
- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git commit <nameOfFile> -m ''Your notes about commit here''`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:
  - `git add -A`
  - `git commit -m ''Your notes about commit here''`
  - `git push`

Psychology Data Analysis

As described in Figure 1, research in Psychology help one to find strategies for better mental and physical health. These strategies for better living are drawn from relationships that may be found in data collected after interviewing participants. In this lab, you to find and argue for the existence of studied relationships in a data sets that originate from actual experiments in psychology.

Your data will come from the *Open Psychometrics* website at link: `https://openpsychometrics.org/_rawdata/`. On the website, data is offered from *actual* experiments and concerns personality-type assessments such as the Big Five Personality Test, Anxiety (Taylor Manifest Anxiety Scale), as well as other prominent tests.

The Existence of Relationships Between Variables?

Once you have selected and obtained your data from the above-mentioned *Open Psychometrics* website, you are to run your own tests between columns of the data **to argue that there is a relationship between at least two of the variables of your selected data set.**. In order to make this assessment, please be sure to spend some time to study all variables of interest to determine potential relationships that may exist. Using code, you will determine a correlation that might indicate a relationship between any two variables that you believe is significant (you could find relationships between more than two variables, if you are interested in the challenge!

Use Your Judgement to Run Your Analysis

The relationship(s), for which you argue to exist(s), is left to your judgment. If you decide that you find no such relationships in your chosen dataset, you will still use your tests, plots and graphics

to make this argument in your report. Note, please use your notes and slides from class to guide your code for each of these tests which are listed below.

Your tests are the following for this work.

1. Correlations (see below)
2. T-tests (see class material)
3. Linear models (see class material)
4. A new approach or test that you will choose and implement from the *psych* software library (more detail given below).

Research a New Approach

Often in data analytics work, you will have to read library documentation to determine how to use a new procedure or method of approach for your work. The last item on the above list is a test that you will have to determine for yourself using the *Psych* library documentation (link: `http://personality-project.org/r/overview.pdf`. Here, you are to read over the library to choose a relevant method of plotting a result, or an analysis to contribute to your report to support your argument of a relationship. Below, we give code and describe an analysis by `pairs.panel` correlation which is discussed in the *Psych* library. While you are invited to use the code for your own correlation study, your researched approach cannot be the `pairs.panel` correlation which is described below.

Correlations

The below is the code that produced the graphic of Figure 2.

```
rm(list = ls()) # remove variables stored in memory.
# install.packages("psych")
library(psych)

# The below is sample code to get you started using pairs.panels function. Remember, for your c

data(sat.act) #load data

# detect outliers in the data
d2 <- outlier(sat.act,cex=.8)

# combine the d2 statistics from before with the sat.act data.frame
sat.d2 <- data.frame(sat.act,d2)

# detect any correlations using pairs.panels() and some parameters
pairs.panels(sat.d2,bg=c("yellow","blue")[(d2 > 25)+1],pch=21)
View(sat.d2)
```

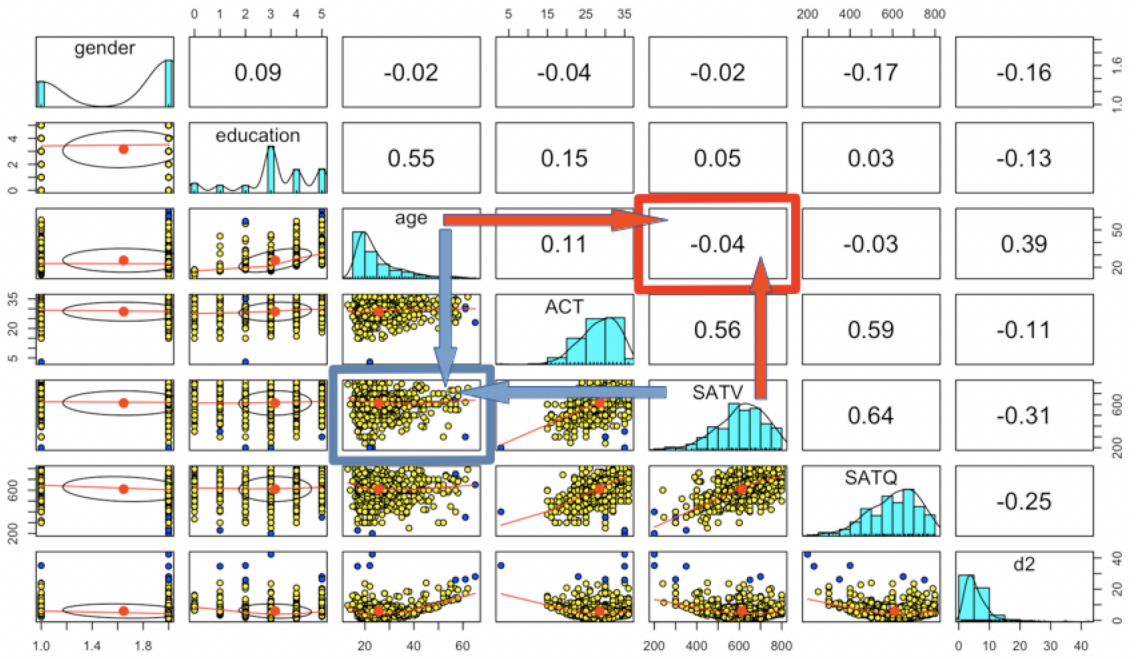

Figure 2: In this `pairs.panel` correlation, all columns of data in a table have been compared to all others in the entire data set. The number values on the right (see red square) are the Pearson correlations and the plots on the left (see blue squares) describe a visual on how the points plot. The correlation values and the plots are found at the intersections of data sets. For example, the correlation between the `age` and `SATV` sets is -0.04, as denoted by the intersection of the red arrows, and the plot is denoted by the intersection of the blue arrows.

T-tests and Models

Once you have found your correlation, you are to use at least one t-test to argue that some relationship may exist. In addition, you are to use a linear model (regression) to further explore this relationship. You are invited to use as many t-tests and linear models as you desire. Whether you find a relationship or not, you must describe the output of your tests to inform the reader of your work in your report.

Details to Consider For The Report

1. **Correlations**: Using your code, notes and skills, uncover correlations in your chosen dataset. As you are working, think about what these correlations might indicate between two columns of data (of your chosen set). Ask yourself how to interpret the output score of the correlation. Are the correlations *high correlations*, or are they *low-correlations* between variables? You could also investigate using `pairs.panels()` to help you discover the correlations but it is expected that you will use other types of code as well to indicate any relationships.

2. **Justify Your Correlation Results**: Once you have found correlations in your data set,

try to justify why the correlation exists. Remember that the data will not answer the *why* part of your work and so you may need to investigate academic articles to provide grounded theories to explain your results. You are to use logical reasoning to give plausible reasoning to provide explanations about these correlation(s). Note: If no correlations were found in your data, then you could go back to pick new columns to run the test again or, you could explain why you think that correlations are not present.

3. **t-Tests and linear models**: You are to use the `t.test()` or `lm()` (linear model regression, to be discussed in more detail in class) to provide insight into your analysis of correlations from above to argue for the existence of relationships. In your report, spend a few lines of text help the reader to understand your work by discussing these tests (i.e., `t.test`() and `lm`() and the variables involved. For correlating variables, how did the t-tests or linear models help you to find new evidence of a relationship, other than that of your correlation study? What were the statistical values (i.e., p-values, R-squared values, and etc.) to note? Please write your conclusions for each of your tests.

Required Deliverables

Note: Please remember to include your name on everything you submit for the class.

1. **Report**; `writing/report.md`: Your report should not be more than two or three pages of text (not counting graphics), and is to have the following parts.
   (a) **Citations**: Please add a reference to your obtained data in your report. Please give a title, link and other information about the data as a reference.
   (b) **Plots an screenshots**: Show and explain your plots and important screenshots to support your argument of a relationship. Explain why you choose your variables featured in your plots and screenshots. You are also to explain how the plots are relevant to your study. Note: there is markdown code included with your report document to help you add graphics to your work.
2. **Dataset**; `data/*`: You are to store your downloaded version of the dataset in this directory to preserve it with the project.
3. **Source code**; `src/analysis.r`: Your code that can be run to load the data files and to produce the plots and analysis of your work. Please add documentation to your code to help the instructor understand your thinking behind the code on a line-by-line basis.