Homework 1

2022-09-24

Contents

1 Requirements

- How to read the requirements? **Carefully**
- File format: For this Homework, you are required to submit both R Markdown and PDF files with your answers and codes in it. Make sure that Rmd file works, there won't be any errors when it is run and represent the same information as PDF Under each question (not in comments) write the code along with your interpretations. **Be sure to put your name at the top of your assignment (in the YAML header in front of the author).**
- Due date: 24.09.2022 19:09 **No late homework will be accepted.**
- Submission: **You need to upload files on Moodle**
- Suggestion: Start with creating a framework for your R Markdown file, retype the tasks and only then start to solve the tasks.
- Rule of thumb: If the number of data points is greater than 50, **do not print the whole data**. Use subsets. Try to show all outputs (do not just store an object as a variable). Also, try to avoid using the same name for variables in the file.
- Cheating: The purpose of tasks is to check your knowledge (rather than the ability of thinking). Please, try to solve without googling every exercise. Try not to discuss with your classmates and work only on your file. **Any similarities, which can be considered as cheated, will not be graded.**
- Packages: The suggested packages are: ggplot2, dplyr, MASS (+ggpubr/ggthemes), matplotlib, pandas, seaborn, numpy
- Note: Keep the order of sub-tasks. Pay attention to titles, legends and axes of graphs.

Good luck!

You will use wine.csv file which has following columns.

1. Wine: The type of wine, into one of three classes, 1 (59 obs), 2(71 obs), and 3 (48 obs)
2. Alchohol: alcohol
3. Malic: Malic acid
4. Ash: ash
5. Acl: Alcalinity of ash
6. Mg: magnesium
7. Phenols: total phenols
8. Flavanoids: flavanoids
9. Nonflavanoids: Nonflavanoid phenols
10. Proanth: Proanthocyanins
11. Color: Color intensity
12. Hue: hue
13. OD: D280/OD315 of diluted wines
14. Proline: proline

2 Use R for this section. (10 points)

1. Read the data and check the structure. Convert Type to factor. Check if there are any missing values.

2.1 Task 1:

Select all numeric variables from your data. Calculate correlation between variables and create correlation heatmap.

2.2 Task 2:

Using correlation matrix from task 1 extract two highly correlated variables. Create scatter plot of that two variables. Use type as color parameter. Define following colors for your graph: #248745, #874724, #000000. Use theme bw. Add appropriate title.

2.3 Task 3:

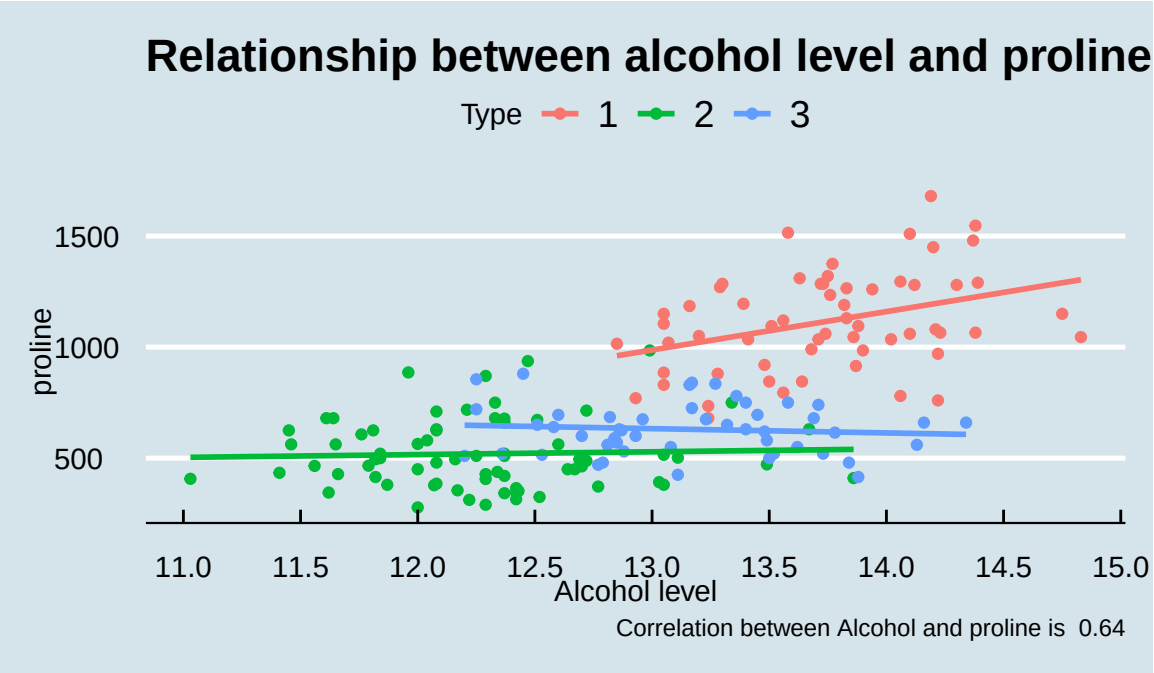
Using visualization techniques check if there is distributional difference between alcohol and type of wine. Use theme_classic for your graph(s). Using alpha add some transparency. Comment on your findings.

2.4 Task 4:

Reproduce following graph

Hints:

- Theme economist
- To add lines use geom_smooth with correct parameters configuration
- Make sure you didn't forget to add footnote to your graph.
- write correlation value and text of footnote using paste0 function
- Note that frequency of x axis ticks is changed



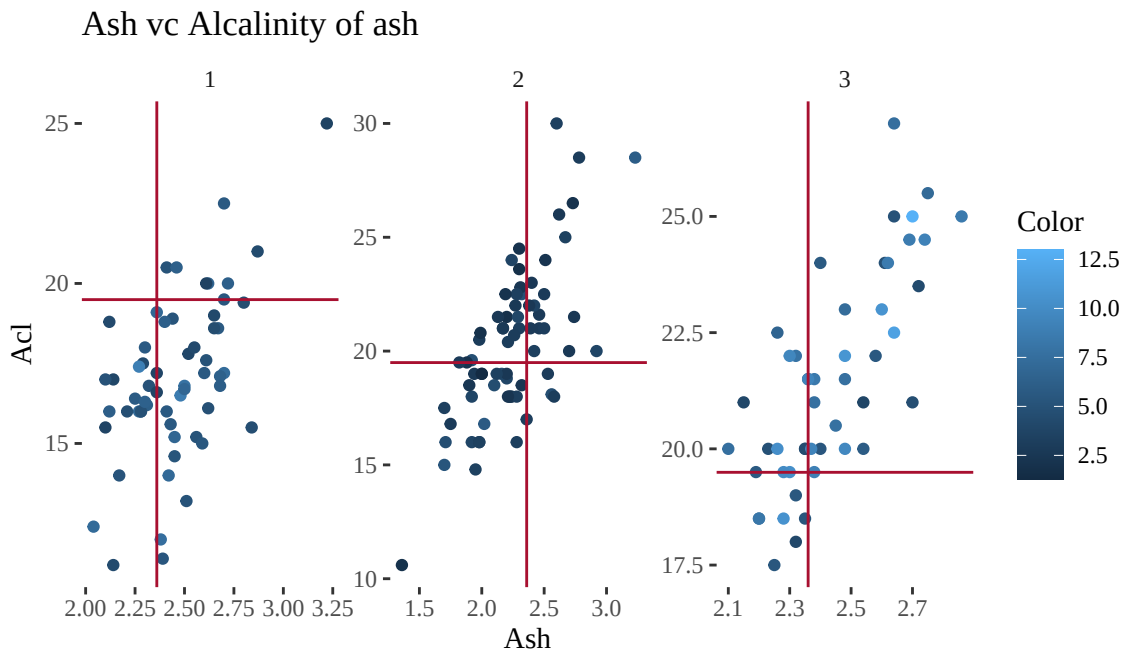
2.5 Task 5:

Bonus point for R section. (1 point)

Reproduce following graph.

Hint:

- Horizontal redline is average of Acl
- Vertical redline is median of Ash
- use theme_tufte()(everybody thinks the same about this theme:D)



3 Use python for this section (10 points)

For this section you have to use gpafactors_2.csv dataset with following columns:

1. studentid – ID of Respondent
2. surveydate – Survey conducting day
3. age – Age of Respondent
4. ehpw – Hours spent on extracurricular activities a week
5. hpw – Hours spent on studying a week
6. hsleep – Hours of sleep per day
7. GPA – Grand point average of student [0-100]
8. gender – 0:male, 1:female
9. job – 1:Respondent has a job, 0:Respondent does not have a job
10. type – 0: part-time, 1: full-time
11. marital.status – single – Respondent is single and has never been married, married – Respondent is married, divorced – Respondent is divorced or widowed
12. imp – Importance of getting/maintaining a high GPA (85 or greater)? 1: Not Important - 5: Very Important

Read the data and check if there are any missing values.

Convert all object variables to categorical.

3.1 Task 1:

Create histograms with binary variables vs grade point average of students. Use your own colors for each graph with appropriate title.

3.2 Task 2:

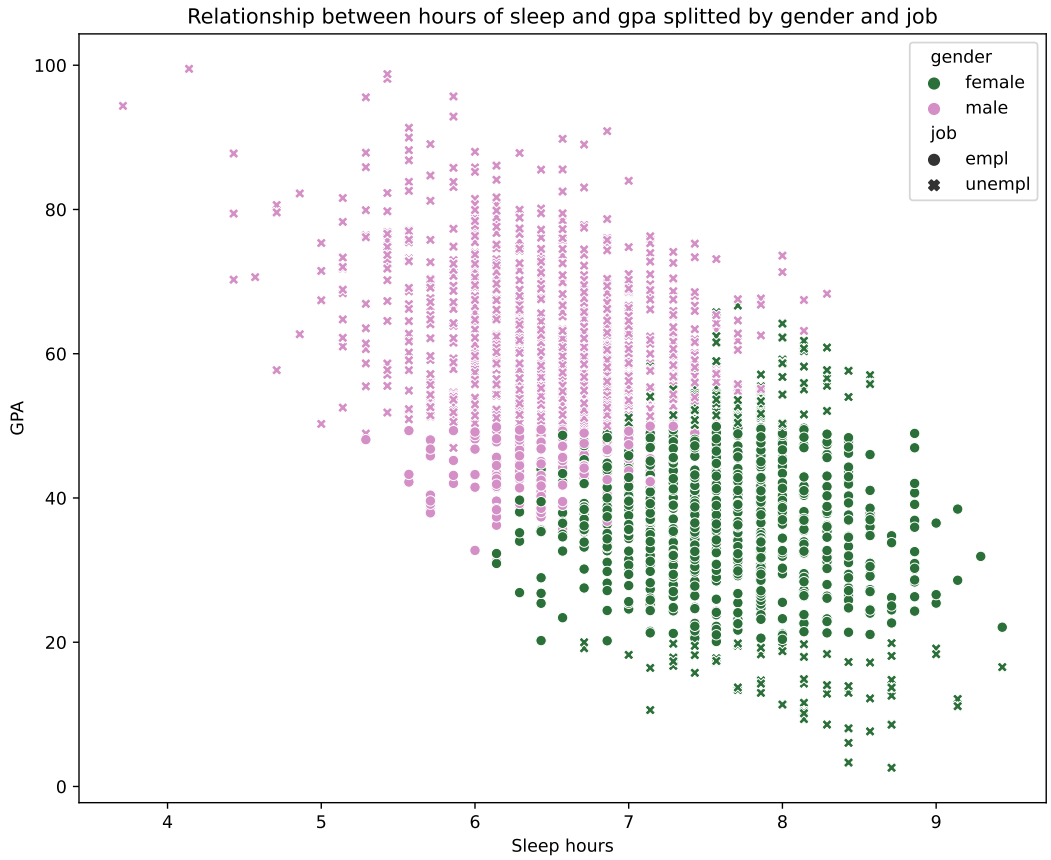
Using np.unique() and list() functions extract unique values of gender and job columns. Visualize relationship between age and gpa for each possible combination (create subplots). Set different colors and approrpiate titles for each subplot.

3.3 Task 3:

Using seaborn library reproduce following graph.

Hints:

- palette: cubehelix



3.4 Task 4:

Create correlation heatmap and show correlation values on the graph. Comment on the results.

3.5 Task 5:

Those students who will find the most important insight from the data will get extra 1 point in python section.

4 Use youtube for this section

Send good music :)