Midterm EDA Project for STA305

Miao Yu

2023-10-21

Goals

The purpose of the project is to give you an opportunity to work on a real-world data set to exercise EDA analysis that we have learned. You should submit a pdf report with at least 10 pages, along with your R markdown file.

The project is expected to take approximately two to three times as much work as a normal homework assignment. You are expected to spend some time **polishing your final report** to make it more concise and accurate in language. Actually, the amount of effort is not necessarily positively correlated with the length of the final report - here are a couple of quotes for your reference:

"I have made this letter longer than usual, because I lack the time to make it short." - Blaise Pascal

"If I am to speak for ten minutes, I need a week for preparation; if fifteen minutes, three days; if half an hour, two days; if an hour, I am ready now." - Woodrow Wilson

Briefly speaking, I hope you to start to exercise how to find the balance between length and quality of a report through this project.

Data set

In this project, you should use R to explore a data set studying obesity. Click here to see the description of the data set and download link.

To understand the meaning of each variable and how the data are collected and created, you are required to **carefully read** the short article here before you start.

Task list

There are two parts. For required tasks, every student must finish them strictly following the instructions. For customized tasks, you may choose the questions that you hope to answer.

1. Required tasks (50%)

(a) Load the data set into Rstudio and name it `obesity`, then create a new column named `BMI` following the formula given in the article. After creating `BMI` (you must do it correctly), execute the following code to reorder the categories of obesity level and change its name to `obesity_level`.

```
obesity %>%
  mutate(NObeyesdad = fct_reorder(NObeyesdad, BMI)) %>%
  rename(obesity_level = NObeyesdad) -> obesity
```

The numeric BMI and the categorized obesity_level that measure the level of obesity will serve as our target variables.

(b) Explore the distribution properly of **each** variable in the data set.

(c) Explore the relationship between the target variable (you may pick BMI or obesity_level or both based on your judgement) and **every other** variable.

2. Customized tasks (50%)

(a) Use graphs or tables to explore any question that you are interested to answer from the data set. You are required to explore at least 3 different questions or answer one question in depth with at least 3 pages of analysis (including graphs) in length.

(b) Build a model with BMI as your target variable. Properly choose independent variables and state why. You must include an error analysis in your report.

Rubrics

- For part 1, the grading requirement is same as previous assignments. Organize your analysis in a well structured format, and for each part clearly state the variable or the relationship to be explored, the graph/table that shows the result, and a concise description of your findings.
  - Completeness: 20%
  - Correctness: 40%
  - Appearance: 20%
  - Accuracy: 20%
- For part 2a and 2b, you will exercise writing a technical report. The rubrics is
  - Professionalism: 30% (free of typos, bad visualization and other “low-level” mistakes that can be avoided by carefulness)
  - Technical Rigor: 40% (correctness and accuracy of your problem definition, codes, graphs and analysis)
  - Logical Structure: 30% (the overall logical flow of your report, or whether the connection between different parts of the report is logically sound)
- For 2a, your report should contain
  - The question(s) to be answered and why you are interested in it (motivation).
  - The graphs/tables (code included) to answer the question.
  - “Storytelling” of how your graphs/tables help give us insights in the question of interest. The writing style should not concise and clear to a general reader with common knowledge but not knowledge in data science or coding.
- For part 2b, you need to
  - show the final model that you build in terms of mathematical formula.
  - show an error analysis of your model.
  - explain how you make decision in choosing the variables included in your model (meanwhile ignoring others).
- For this report, you will be given a round of revision to improve its quality after I give comments before the final grade is given.