W203 Lab 1: EDA

W203 Instructional Team

Overview

The goal of this lab is to provide you with an opportunity to use R and gain experience performing exploratory data analysis (EDA). In this lab, you will be asked to find new insight into a data set by assessing the underlying structure, evaluating the variables, detecting outliers and anomalies, and so on.

This is a group lab. Each team will be assigned a cancer dataset to work on. In your assigned folder, you will find a file containing background information on your data, along with a research objective and any instructions that are specific to your team.

Please note that you will be working with real data, but it may have been modified by your instructors to test your abilities.

Although your assigned topic may be the focus of an existing literature, we recommend that you do not spend your time researching what others have done, or gaining significant domain expertise. The purpose of the lab is to see how well you can apply exploratory techniques. Moreover, the background we have provided in your assignment should be sufficient to guide your analysis.[1]

Assignment

Generate an exploratory analysis to address the goals found in your assigned folder.

Be sure to follow the guidelines we covered in class. Remember that you are to use descriptive tools (no inference), but note any features you find that you think would be relevant to statistical modeling.

Your analysis should be thorough, but limit your report to a maximum of 25 pages. This means that you will have to make choices about what variables and relationships to focus on (and justify those choices).

To assist with evaluation, we are providing the following outline for your report. As you work, you may fill in each section with your analysis.

[1] We also do not want you to be led astray by the bad advice that is common on the internet.

Introduction (20 pts)

State the research question that motivates your analysis.

Load your data set into R.

Describe your data set. What types of variables does it contain? How many observations are there?

Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues.

Explain whether any data processing or preparation is required for your data set.

Univariate Analysis of Key Variables (20 pts)

Use visualizations and descriptive statistics to perform a univariate analysis of each key variable. Be sure to describe any anomalies, coding issues, or potentially erroneous values. Explain how you respond to each issue you identify. Note any features that appear relevant to statistical analysis. Discuss what transformations may be appropriate for each variable.

Analysis of Key Relationships (30 pts)

Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bivariate relationship.

What transformations can you apply to clarify the relationships you see in the data? Be sure to justify each transformation you use.

Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Explain how these variables affect your understanding of the data.

Conclusion (20 pts)

Summarize your exploratory analysis. What can you conclude based on your analysis?

Evaluation

We will evaluate your report for technical correctness, but also clarity and overall effectiveness. A point distribution is provided with the above outline. In addition to these point totals, we will impose penalties for output dumps, unclear language, and other errors.

Submission

Only one student in the team needs to submit via the ISVC. Make sure that you include the names of all group members in your report.

You must turn in

1. Your pdf report. In this report, do not suppress the R code that generates your output.
2. The source file you use to generate your report (i.e. your .ipynb/.Rmd file)

Use the following naming convention for your files:

```
lastname1_lastname2_lab1.pdf
lastname1_lastname2_lab1.whatever
```

Due Date

This lab is due 24 hours before the week 4 live session.