Guidelines for Assignment #1

CODE

STA 522: Applied Statistical Machine Learning

This is the first part of the first project on data preparation for machine learning. The assignment focuses on visual exploratory data analysis and must be submitted to D2L before the deadline.

Some general requirements

- Provide a clear and descriptive title for your work. Generic titles like *Assignment 1, Project 1*, or *STA552 Homework 1* are not acceptable.
- Ensure the inclusion of essential graphical elements, such as axis labels, a meaningful title, numbered captions, legends, and relevant annotations.
- Both base R plots and ggplot plots are acceptable. However, the use of interactive plots is strongly encouraged.
- For 3D plots, utilize interactive graphical libraries or functions, such as `plot_ly` or `ggplotly()`.
- Adhere to the principles of graphical design to ensure clarity, readability, and effective communication of information.

1. Working Data Set

This data set will be used for project #1 on data preparation. The following are basic requirements

- **Data Size and Feature variables**
  - Data size must be at least 1000.
  - Number of feature variables must be bigger than 10 (among them, at least 3 categorical and 3 numerical features)
- **Missing values**
  - Consider choosing a data set with some missing values in both categorical and numberical feature variables
  - *If you choose a data set with no missing values, but you are particularly interested in working with that data set*, you can created some missing values for some categorical and numerical features. See the following code showing how to randomly replace values in individual feature variables with missing values.

HIDE

```
# create a data frame
WorkingData <- data.frame(letter = letters,
                          gpa = rnorm(100, mean=2.5, sd = 0.5)[1:26])
# create random observation ID and replace the corresponding obs with missing
ltr.missing.id <- sample(1:26, 6, replace = FALSE)
gpa.missing.id <- sample(1:26, 8, replace = FALSE)
WorkingData$letter[ltr.missing.id] <- NA
WorkingData$gpa[gpa.missing.id] <- NA
print(WorkingData)
```

```
   letter      gpa
1       a       NA
2       b 3.167320
3       c 2.427115
4    <NA>       NA
5       e 2.786401
6       f 3.062430
7       g 2.286247
8    <NA> 2.195960
9    <NA> 2.214618
10      j       NA
11   <NA> 3.078095
12      l 2.449709
13   <NA> 2.784375
14   <NA>       NA
15      o 2.895830
16      p       NA
17      q 2.356203
18      r 2.953314
19      s 2.372893
20      t       NA
21      u 2.246788
22      v       NA
23      w 2.611601
24      x 2.562126
25      y 2.946807
26      z       NA
```

- **Where to find data**: There several sites you can explore to find your data set.
    - My teaching data repository (see the link in the top navigation panel of the course web page)
    - UCI machine learning data repository
    - Kaggle site
    - any other sites you can find.

2. Introduction

- **Description of Data**
    - *Purpose of Data Collection* - Provide a clear and concise explanation of why the data is being collected, highlighting the specific objectives and intended use of the data.
    - *Description of the Data Collection or Generation Process* - Outline the methods used to collect or generate the data, including any tools, technologies, or protocols followed. Specify the time frame

and location, if applicable.
    - *Sample Size and Number of Feature Variables* - State the total sample size and the number of feature variables included in the dataset, providing context for the scope and representativeness of the data.
    - *Itemized List of Feature Variables* - Present a detailed list of feature variables, including:
        - Definition/Description: Provide a brief explanation of what each variable represents.
        - Data Types: Specify the type of data (e.g., categorical, numerical, boolean, text).
- **Purposes of Using This Data Set**
    - Outline the analytical tasks for this project
    - *Problem Statements* – Ensure problem statements are specific, accurate, explicit, and concise.

3. Distribution of Individual Features

- **Opening paragraph** - outlines the analytic tasks required to prepare and analyze the data effectively. Each task will be detailed in individual subsections, providing clear guidance on the specific actions needed to address data preparation, exploration, and visualization requirements. The goal is to ensure the data is well-suited for subsequent modeling and analysis.
- **Individual analytic tasks** - describe each individual analytic task in detail. Each subsection begins with a brief introductory paragraph and highlights key features requiring additional analytic actions, such as imputation or feature engineering.
    - open a subsection with a short opening paragraph to provide an overview of the task and summarize relevant features that demand additional pre-processing steps. These may include handling missing values, addressing outliers, or engineering features to enhance model performance.,
    - summarize only those features that require additional analytic actions such as imputation and feature engineering in subsequent sections.
    - all visual representations will include the following components:
        - *Observations*: A description of key patterns, trends, or anomalies observed in the data.
        - *Implications*: The practical and analytical significance of the observations, highlighting how they might influence the data's usability or the modeling approach.
        - *Analytical Actions*: Specific steps to address insights drawn from the visualization, such as performing transformations and other feature engineering procedures in ML data preparation workflows.

4. Relationship between Features

- **Opening Paragraph of the section** - outlines the analytic tasks in this section.
- **visualize relationship between two feature variables** - The goal is to select appropriate visual representations and ensure all necessary components are included in each visualization. Specifically, the tasks involve the relationship between:
    - Two numerical feature variables
    - Two categorical feature variables
    - One numerical and one categorical feature variables
- **Three or More Feature Variables** (optional) - Representing high-dimensional data using marks, channels, or other model-based methods.