Exploratory Data Analysis Project

Time to put everything you learned in this class into action. In an exploratory data analysis (EDA) you are just looking at (exploring) the data and learning about the data and possible relationships between variables. This is not a formal statistical analysis, you cannot make any claims about groups being statistically different. This is just descriptive. You are allowed and encouraged to hypothesize why you observe certain relationships or data characteristics, just be sure not to draw any conclusions from the data.

Instructions

Using your data set of choice, pose a brief research question that explores the relationship between 2-3 variables. Use markdown headers to make the following sections

1. **Introduction**: A short introduction/description of the data.
    - Specifically mention the 2-3 variables you are going to explore.
    - What is your research question? What are you interested in finding out more about?
2. **Univariate Exploration**: Describe each of the variables under consideration.
    - This means calculate some summary statistics (N(%) or mean(sd)) and make a graphic
3. **Bivariate Exploration**: Comparison between two variables of interest.
    - Calculate grouped summary statistics as appropriate. **This is often the most often forgotten part**
    - You can go further and explore more than two variables at a time using paneling, but be sure to explain what you learn from each graph.
4. **Conclusion**: What did you find? If you had a prior hypothesis, does the data seem to support it? Remember this is NOT a statistical analysis.

All descriptions (univariate and bivariate) must be done using graphics, summary statistics, and words.

This is a very vague set of instructions for a reason. I want you to explore and choose a pair of variables that you find interesting. Create tables, graphics, grouped summary statistics (mean of the continuous variable across levels of the categorical variable). Whatever you need to do to understand the relationship between these two measures.

Use the grading rubric at the end of this document for guidance as to what you should present, in what order, and level of detail you need to present.

Data

You have a choice here. If you are currently working on some data that you would like to explore, talk with your instructor to get your data set approved. As long as it has more than a few variables in it, and at least 30 observations it should be fine.

If you do not have your own data, you can choose from one of the following data sets, all of which can be downloaded from the Data page of Dr. D's teaching course website. Here are some viable choices:

- **Email Spam**: Characteristics of emails used to predict if the email is spam or not.
- **HIV**: Data on adolescent children living with HIV positive parents.
- **Depression**: Level of depression (`cesd`), health care, and demographic characteristics.

- **High School and Beyond**: Educational, vocational, and personal development of elementary and high school students.
- **Police Shootings**: Characteristics of individuals killed by police in 2015.

Any other data sets require instructor approval. Data sets such as the `plants` or `arm strength` data sets are insufficiently complex for this project. Also you can't use the `dsmall`, `diamonds`, or `NCbirths` data sets because we've used them too much already.

How to submit

- To allow for adequate time for peer grading the submission deadline is a strict cutoff. Really late assignments won't be accepted.
- You must name your file **EDA_username** and knit to PDF.
  - **If you knit to HTML, you must open & save the resulting file as PDF so it can be commented on.**
- Upload your final project to this [Google Drive folder] by the due date.

Peer Review

After the submission deadline, your analysis projects will be randomly assigned to two other people to peer review and score. This means you will also score and provide feedback on 2 reports. Your instructor will also score all projects for your class section.

How to do your reviews:

1. The morning after the due date you will be emailed a link to the peer review spreadsheet in Google Drive. This is where you can find the names of the people you are assigned to review.
2. Go to the [project page] and download the projects you are assigned to review.
3. Using the commenting feature provide 4 comments for each project.
   - Two positive: What specific features did they include that you liked or found helpful?
   - Two improvements: What can they do different or better next time? Did you find a bug in their code?
4. Score each project using this [Google Form].
   - This data entry form follows the scoring rubric printed below in this document.
   - You will upload your review to this form as well.

   How to add comments

   - MS Word
   - Adobe Reader

Guidelines

- Knit early and often. As often as every time you include a new R code chunk.
- Spell check your report prior to submission using RStudio.
- Re-read your report and edit for clarification and removing duplicated information.
- Remove superfluous code and output (i.e. printing a data set to the screen).
- This is to be independent work. Papers that are too similar will receive no credit.
- Look at the grading rubric to help you decide the level of detail required.

Grading

- Your final project grade will be a weighted average of 30% peer reviews ($\bar{PR}$) and 70% instructor review (IR): $(.3 * \bar{PR} + .7 * IR)$.
- Your submission is worth 20 points, the peer review is 5 pts.

Scoring Rubric

The criteria below is what you will be graded on. Below each criteria is an example of the points awarded for the level of competency. Use this criteria when you score your peers reports.

1. **Data Description**: Provide a description of the data set and the variables of interest.
    - (Novice) There is no description or the description is a copy of the help file.
    - (Competent) There is a minor description of the data but not enough to understand what is being measured or compared.
    - (Proficient) The data description is clear and concise, it is clear to me what data is being analyzed and where it was obtained.
2. **Univariate Description**: Fully describe the distribution of *each variable* by itself
    - (Novice) There are no numerical or graphical summaries provided.
    - (Competent) Only numeric or only graphical summaries were created, but no textual description.
    - (Proficient) The variable was fully described using both numeric and graphical summary methods. This information was summarized below the output in a paragraph form.
3. **Bivariate Comparison**: Describe the relationship between the two chosen variables.
    - (Novice) No comparison was made, or the variables were compared, but inappropriate graphics or summary statistics were created.
    - (Competent) The variables were compared using appropriate graphical methods and grouped summary statistics were created, but nothing was discussed.
    - (Proficient) The variables were compared using appropriate graphical methods and a short textual explanation of what the summaries showed.
4. **Organization / Grammar**: How well does the report read? How well organized is it? Was it checked for grammar and spelling mistakes?
    - (Novice) Only R code, output is present. There is no discussion of results. Tons of extra R code that is not relevant to the discussion is present. Markdown headers were not used.
    - (Competent) An attempt was made to discuss the results, but the explanations are not in a report format or there are some large grammar and/or spelling problems. Some R code that is not relevant to the analysis question at hand is being displayed. Markdown headers were used to create sections.
    - (Proficient) The report was spell written in well edited, full English sentences, and spell checked prior to submission. The report flowed well and followed the required order of discussion topics with markdown headers used successfully.