Project-Step1 (20 points)

Saad Mouti

10/22/23

The purpose of this document is to provide guidance for structuring your project. The deliverables will consist of an R markdown in .Rmd to be submitted on Canvas and knitted pdf version to be submitted on Gradescope. **The deadline for this step is October 22nd, 2023**.

Data Description & Descriptive Statistics

Goal

The first objective is to understand the variables in your dataset and their relationships. Your task is to choose an appropriate dataset, describe it, and perform some descriptive statistical analyses. You should carefully consider the observational units in the dataset to ensure they are independent.

The final report should include the following:

- The name and/or source of your data, a description of all relevant variables, and the observational unit (i.e., row).
- Appropriate summary statistics with adequate explanation and interpretation. You should examine binary or other relationships, as well as describing individual distributions. Consider making an appropriate table of some kind. The package skimr has a function called skim, which is great for summarizing data succinctly.
- Graphical displays with adequate explanation and interpretation. These should effectively summarize your data and point out any interesting features. You do not need a picture or table for every variable. Be careful with the word "normal". Unless you actually check that the data are normal, stick with words like "symmetric" or "bell-shaped".
- A comment on anything of interest that occurred during the project. Were the data approximately what you expected, or did some of the results surprise you? How did the sampling go? Do you think you got a representative sample of your population?

Data Limitations

- The dataset must include at least 10 variables, with at least 3 independent quantitative and at least 2 independent categorical variables. (Label/ID/name does not count as a variable because it cannot be used in a model.) Please use full variable names/descriptions in your sentences (or make your abbreviation clear to your reader).
- If you are unable to find one with these specifications, take one with continuous variables and create categorical variables by picking some of these variables and splitting them. Example, you can pick BMI and split it into very low, low, mediam, high, very high by using common sense of percentiles.
- The dataset should have at least 100 **independent** cases/observations (ideal number of observations is 200-500). Be wary of missing observations! If you have too many, pick randomly 500 observations from the dataset. That way, scatter plots will not seem overwhelmingly charged.
- Be very careful with time (e.g., year) as a variable because it can indicate that your observational units are not independent.
- Because you will be doing hypothesis testing later, you need to indicate what population your data describes. If it is a census, then maybe it is representative of an even larger population? (For example, a census of state information from 2015 might be somewhat representative of 2016? Is it?) Also, discuss the limitations of describing a larger population.
- If you plan on choosing the dataset yourselves, consider using the following data sources:

    1. UCI Machine Learning Repository: This is a large collection of datasets maintained by the University of California, Irvine. Many of the datasets are suitable for linear regression analysis https://archive.ics.uci.edu/ml/index.php.
    2. Kaggle: Kaggle is a platform for data science competitions and hosts a wide range of datasets. Many of the datasets are suitable for linear regression analysis https://www.kaggle.com/datasets.
    3. OpenML: OpenML is an open-source platform for sharing and organizing machine learning data and experiments. It also provides a large collection of datasets for linear regression analysis https://www.openml.org/home.
    4. R datasets package: The R programming language comes with a collection of datasets built-in to the base installation. These datasets can be loaded directly into R and used for linear regression analysis.

Format

Do:

- Use captions for every plot, e.g., in the chunk command give the caption: "'{r fig.cap = "here is the caption"}.
- Use complete sentences.
- Annotate everything that the reader sees.
- Keep the file to 4 or fewer pages (mostly graphics).
- submit in both the .Rmd and .pdf file on gradescope. Make sure the .Rmd file can knit!

DON'T:

- Print any warning or error messages. Only print code that is interesting and relevant to the reader (e.g., use echo=FALSE).
- do not print lists of data.
- no overplotting (use boxplots instead of scatterplots when appropriate; use alpha=0.1 for transparent plotting symbols).
- **no linear models** for this part; **no hypothesis tests or inference** of any kind is expected. If you are curious about relationships in your data, it is possible you could run a t-test or a chi-squared test, but most people will not have a hypothesis test of any kind.
- do not include any tables, output, or graphs which are unannotated.
- do not be tempted to turn in everything you do. Only turn in the interesting parts of the analysis. One of the hardest parts of being a consultant is figuring out what to tell the client.

Groups

Try to form groups of different majors. I also encourage you to work the homework in the same group even though the submission must be individual. This way you can also help each other and make the learning more thorough. In industry and academia you will have to work in groups so managing different backgrounds and levels will be very rewarding. Anyway, working in groups has learning and communication benefits, working alone has other benefits. You can choose either but the grading criteria will be same. You need to keep the same group throughout the quarter for all the submissions. The project is quite heavy for a single person and the grading will be consistent (will not take into account if you are by yourself). If the class has an odd number of students then one group could have 3 students in it (exceptionally).

List of some dataset suggestions:

0. Boston Housing dataset (UCI)
1. Concrete Compressive Strength dataset (UCI)
2. Airfoil Self-Noise dataset (UCI)
3. Energy Efficiency dataset (UCI)
4. Bike Sharing dataset (UCI)
5. Red Wine Quality dataset (UCI)
6. Student Performance dataset (UCI)
7. Forest Fires dataset (UCI)
8. Auto MPG dataset (UCI)
9. Baseball Databank dataset (Kaggle)
10. Adult Census Income dataset (Kaggle)
11. Ames Housing dataset (Kaggle)
12. Forest Cover Type dataset (Kaggle)
13. Melbourne Housing dataset (Kaggle)
14. California Housing dataset (Kaggle)
15. Superconductivity dataset (Kaggle)
16. Diamonds dataset (Kaggle)
17. Advertising dataset (Kaggle)
18. Bike Sharing Demand dataset (OpenML)
19. SkillCraft1 Master Table dataset (OpenML)
20. Parkinsons Telemonitoring dataset (OpenML)
21. EEG Eye State dataset (OpenML)
22. Online News Popularity dataset (OpenML)

Final remarks:

Data is not supposed to give "perfect" results. The goal is not to build the perfect model but to apply what we learn in class. Take time to select your data, and for that, ask your instruction team what they think of the data you selected (after doing some quick graphs). You can also ask them during office hours if they have a dataset to suggest (although, this exercise is for you, first but in case you ran into trouble). Make sure to read the description of the data first (some link specify the predicted variable and predictors, as well as whether the target (predicted variable) is continuous or categorial (we don't want a categorical Y, that's for classification, not regression).