Project 1

Dr. Christopher Brown

9/25/2020

What Will You Be Doing?

So far in the course we've looked at importing, cleaning, and exploring data. You're going to be putting all of that into practice. You (and your team, if you choose to work with one) will be choosing a data set of interest to you, building or cleaning it, exploring it to find some preliminary results of interest, and then communicating your results.

Okay, But What Are The Steps?

1. Form a team of four or fewer students, or decide to go it alone. I'll create a thread in the discussion board in class, **lfg** (looking for group, not lease finance group or let's … umm, freakin' … go).

2. Choose a data set or decide to build one.

Choosing a data set? Try Kaggle (https://www.kaggle.com/ (https://www.kaggle.com/)), the UCI Machine Learning site (https://archive.ics.uci.edu/ml/index.php (https://archive.ics.uci.edu/ml/index.php)), or data.gov (https://www.data.gov/ (https://www.data.gov/)) (Bureau of Labor Statistics is also a solid choice, but more specialized).

You should not choose just *any* data set of interest. You'll need the format to be a `.csv` or an spreadsheet file that can be saved as `.csv` from Excel or Libre Office, you'll need to have at least a few numerical variables and a few factor or text variables, you'll need to be able to access it, and you'll need to be able to understand how it was collected. You're welcome to use data you've collected in research projects for classes or other experiences, but you'll need to get written consent from everyone else involved in the group that carried out the research that collected the data.

Building a data set? There are a variety of ways to do this; an easy one is to collect data from websites (for example, textbook prices from Amazon). Be warned, there are a few things to avoid in data collection. First, **don't create and administer a survey** to get your data! It can be difficult to obtain good data without good survey design, and that's hard to do. Second, **don't hypothetically plan to collect some data and build your project plan around that**; actually go out and collect some of the data you're planning to collect, measure how much time and effort it takes you, and decide whether your data collection is feasible.

3. Import and clean your data. (If you've entered your data yourself, hopefully you won't have to clean it, so instead you should describe how you collected it.) There are lots of decisions to make here. Make sure to document your steps in an RMarkdown document, and give brief descriptions of *why* you made the decisions you did. Remember: this will not always be obvious to someone reading your work! It's your job to make clear what you did and why to a reasonable reader.

4. Write a data dictionary. Once you have imported and cleaned your data set, make a list of the variables. For each, write the name of the variable, write the intended type (numeric, factor, datetime?), and write a very brief description of the variable.

5. Explore your data. Create lots of tables and plots. Look for meaningful, or at least potentially meaningful, relationships among variables.

6. Write everything up. Focus on **clear**, **concise**, and **correct** communication.

You will probably want the following sections: Introduction, Import and Cleaning, Data Dictionary, Exploratory Data Analysis, Conclusions and Questions for Further Study.

For project 1, you will probably have few *results* and you will hopefully have *many questions raised*; since raising questions for further study is really the point of an exploratory data analysis, you've done your job if you've generated questions…so you can stop there! (Foreshadowing: there is a second project in the course so what might be in that?…)

Things to avoid when writing up your results? (1) You can speculate about causality, but don't claim it unless it is very clear. (2) *Don't show every table and graph that you generated to explore your data set.* Show those that led to the questions you raised. Your work will be graded heavily on your ability to communicate your results, and a large part of that is exercising your judgement to decide what is relevant to your write-up and what is not … and then getting rid of what's not. (In all possible ways, I'm trying to nicely say this: You know that person that told you that whenever you see a question on an exam that you don;t know how to answer, you should just write down everything you know and hope for partial credit? Don't do that in this project.) Clarity is key, and a component of clarity is relevance.

How Should I Turn This In? How Are You Going To Grade This?

Your written project should be submitted as a `.pdf` file to BlackBoard. I do not need to see your dataset.

Each person on your team should submit the *same* `.pdf` as their assignment. The `.pdf` of your write-up should include all team members' names as authors. (This is really a record-keeping issue.) Everyone on the team will receive the same grade.

The grading rubric (total of 150 points):

- Importing and cleaning (or collection) … 25 points
- Data dictionary … 25 points
- Exploration … 25 points
- Communication - Clarity (readable, understandable, could I replicate your work based on your write-up? Relevant work?) … 25 points
- Communication - Correctness (is your work correct? did you use correct types of graphs for variable types? are your R manipulations correct?) … 25 points
- Communication - Concision (are you keeping your work concise, no fluff added?) … 25 points