Stat 250 Project 3

Import, Transform, and Visualize

Due Apr. 26, 2023

Project Description

So far you've enjoyed (or endured) the data examples that I have chosen in class and in the homework. Now I get to enjoy the data examples that you choose! Your task in this project is to find, import, clean/transform and visualize data on a topic of your choosing.

Project Instructions

- Unlike the other two projects, this project is to be done individually.
- As in the other projects, you should do this project in a quarto file (.qmd), interspersing your code with commentary about your data, code and why you made the decisions you made.
- Find two or more related data sets. These cannot be from the same source (i.e. you cannot get sports statistics from 2 different NBA pages). There are many resources to find data. BYU has compiled a list of data resources at `https://guides.lib.byu.edu/c.php?g=216399&p=7978371`. One provider here that may be very useful is data planet: `https://dbs.lib.byu.edu/data-planet`. BYU has a subscription to access these otherwise hard to obtain data. You could also get data from a site that aggregates data such as Kaggle. It could also just be from a web page that you find through googling!
- Import the data sets into R in a code chunk in the .qmd. Do this in such a way that your work is reproducible (i.e. don't just use the import wizard in R without providing any code).
- Clean/transform the data sets. I would like you to use all 5 of the primary `dplyr` functions we discussed in class: `filter()`, `arrange()`, `mutate()`, `select()`, and `summarize()`. Here are some examples of ways that you might clean/transform the data sets:
  - Create new variables as a functions of other variables in the data sets
  - Create a summaries of the data set on a factor variable of interest
  - Remove missing values (unless you feel like there is a reason to keep them)
  - Sort the data such that it makes it easier to make comparisons in a tabular view of the data

  - Change data types of variables (e.g. make sure numeric and character variables are the data type they should be)
  - Combine factor levels for variables with many small factor levels
  - Ensure dates and times are correctly labeled
  - Remove any duplicate or unnecessary columns

- Using joins, combine the cleaned data sets to create one data set.
- Using the `kable` package, create a visually appealing table of your joined, cleaned, and transformed data set. The table should only show 10 rows of the data set. If you need to omit some columns to make sure the table fits within the width of the page that is perfectly okay.
- Create two separate graphics to visualize interesting features/relationships in the data set(s). For each plot, choose two or more variables and create a ggplot graphic using the principles you have learned in class. These can be any type of graphs, however, they should have at least one element more than a simple plot. For example, a simple bar plot would not achieve full points, but a bar plot on one variable that had a fill color of a separate variable would be fine. Scatterplots should have an additional feature or a facet wrap. The plots should be well-labeled so that a reader not familiar with the data sets could understand all aspects of the plots. Ideally, I'd like these plots to illustrate interesting/meaningful relationships, or answer a question that you might have had when choosing your data sets.

Deliverables

There are four deliverables:

1. Files of the raw data sets. If you scrape data directly from the web, you don't need this. You may instead note the URL(s) in your quarto document (make sure the urls are formatted as links in your quarto document if you do this, so that I can click the links on the pdf and be taken to the data sets).

2. Your `.qmd` script with the code used to import, clean/transform, and visualize the data set and the corresponding markdown in the report. Mention the sources used to find the data in your markdown commentary, and why you decided to use the data that you did. Your `.qmd` should be able to be completely reproducible by me without me having to alter anything in your script, provided that I have the raw data files in my working directory. You may assume that I have all necessary packages installed (but still include the `library` commands in the .qmd code chunks).

3. A csv file(s) of your final cleaned/transformed data set(s). If your two graphics use different transformation operations—so that you have two separate data sets on which your two graphics rely—please submit separate csv files for both transformed data sets. The purpose of this file(s) is so that if I fail to read your raw data into R for some reason, I can read in this csv file(s) and reproduce your graphics if needed. Note that the code to write your cleaned data set(s) to file should be included in code chunks in the .qmd. (Hint: `?write_csv` in the tidyverse)

4. A pdf rendered from the .qmd file containing your report for the project. Include your code as code chunks in the report. Intersperse your code with formatted text to describe your code and figures. Make sure to include your name and section number in an author line on the report.

Put the files from the four steps above in a folder called `'<YOUR>_<NAME>_proj_3'`, then compress (zip) the folder and submit the resulting compressed file on Canvas. For example, in my case, this folder would be titled `nathan_sandholtz_proj_3`.

Grading

The project will be graded according to the following rubric. Late work will be deducted a penalty according to how late it was.

Grading Item	Points
Following instructions and code that runs without errors	15
Good presentation (including well-commented code), clear description of your data, the choices you made when cleaning/transforming, and your figures	5
Total	20

Following Instructions refers to following all the items laid out in the Project Instructions and Deliverables sections above.