Mini-Project 1

Stat 212: Interim 2021

General Instructions

This assignment is due **Friday, January 8 at 11:59 PM**. You will submit one pdf file for the whole group, knitted from an RMarkdown document, to Moodle with all contributing group member names. Your file name should include the last name of everyone in the group and the project number, `PotterGrangerWeasley_Project1.pdf`. You do NOT need to submit your RMarkdown .Rmd file, *however* it needs to be saved in your "Submit" folder in the R server so I can access it if I need to look at your code more closely. For this first project your document can show all code and output to answer the questions. Eventually, you will hide the R code and have a neater final report. You may use the RMarkdown file that created this pdf (available as `MiniProject1.Rmd` in the R server) as a template for your analysis (but please delete this and other instructional paragraphs).

There is not a great way to simultaneously code and share the same document with your group members in the R server. You may want to have a shared Google doc open to copy and paste text/code with your partners. You can share a single RMarkdown file by saving it in the "Project" folder, but be aware that everyone in the class can also access files in this folder.

Are all pockets created equally?

You will be exploring the wide world of the fashion industry for the first mini-project. We will use the `measurements` dataset to answer a few questions about how jeans are priced and sized. The data consist of several measurements for different brands and styles of blue jeans in 2018. You can find more information about the data through **this link**.

The first thing you need to do is read in the dataset. The code below will read the data directly from the website (a little different from importing it from our R server). After you read in the data, complete numbers 1 - 3 by writing any code in the R chunk and writing your comments in well structured paragraphs directly below the code. You may refer to the EDA guide posted on Moodle and the `SleepScript.Rmd` file for help with the code, in many cases all you need to do is change the variable and dataset names.

```r
library(readr) ## need to load a special library to import data from a website
measurements <- read_csv("https://raw.githubusercontent.com/the-pudding/data/master/pockets/measurements
## click the play arrow to run the code in this chunk
## type measurements on a new line and click run to see the data
measurements
```

```
## # A tibble: 80 x 16
##    brand style menWomen name  fabric price maxHeightFront minHeightFront
##    <chr> <chr> <chr>    <chr> <chr>  <dbl>          <dbl>          <dbl>
##  1 Ariz~ skin~ women    Fave~ 78% c~  42             14.5           15
##  2 Ariz~ stra~ women    Perf~ 78% c~  42             14.5           14
##  3 Ralp~ skin~ women    Mode~ 92% c~  89.5           13             13.5
##  4 Ralp~ stra~ women    Prem~ 92% c~  89.5           13             13.5
##  5 Uniq~ skin~ women    Skin~ 87% c~  39.9           13             13
##  6 Uniq~ stra~ women    High~ 98% c~  39.9           15.5           12
##  7 Calv~ skin~ women    Midr~ 98% c~  79.5           12             12
##  8 Calv~ stra~ women    Stra~ 85% c~  69.5           14             11.2
##  9 Lucky skin~ women    Ava ~ 69% c~  99             13             14.5
## 10 Lucky stra~ women    Swee~ 96% c~  79.5           15             16.5
## # ... with 70 more rows, and 8 more variables: rivetHeightFront <dbl>,
## #   maxWidthFront <dbl>, minWidthFront <dbl>, maxHeightBack <dbl>,
## #   minHeightBack <dbl>, maxWidthBack <dbl>, minWidthBack <dbl>,
## #   cutout <lgl>
```

1. The Data and Variables

Before we jump into any visualization or summarization of the data, we need to understand the data as a whole. Knowing the intended population, sample size, and variables will help us to better formulate research questions we can answer statistically and make sure we know exactly to whom we can generalize our results. (Hint: you may use R to answer some of these questions, but the link above will also provide very useful information. The functions `dim()` and `name()`, among others may be useful here.)

```
## Use this blank R chunk to write any code you may need to answer questions below.
## You may also create new R chunks by pressing Ctrl+Alt+I, new chunks help to organize your analysis
```

a. Understanding the sample and population.

- What does each observation represent in this data? Each observation represents a pair of pants.
- What do you think the intended population is for this sample? The population is "20 of the US' most popular blue jeans brands".
- Do you think the sample is representative of the intended population? In what ways might it be biased or misrepresentative? The sample does not represent all brands of jeans in the US. There might be selection bias of the researchers to choose brands they have anecdotally experienced pants with small pockets.

b. What are the *dimensions* of this dataset? In other words, how many observations (rows) and how many variables (columns) have been collected.

```
dim(measurements)
```

```
## [1] 80 16
```

80 rows, 16 column.

c. Create a **variable codebook** for the data. This is a summary of all variable names (as they appear in the raw data), what they represent, how they are measured, and what the units of measurement are. I've started the table below, add in additional lines for each remaining variable. | **Variable name** | **Original name** | **Description** | **Type** | **Levels/Encoding** | | ———— | ———— | ———— | ——— | —————- | | Brand | `brand` | Full brand name | categorical | identifier | | Style | `style` | The cut of each type of jean | categorical | boot-cut, skinny, slim, straight, regular |

Maximum back pocket height | `maxHeightBack` | Height (cm) of the longest axis of the back pocket | numeric | none |

|man or woman pants|`menWomen`|intended sex|categorical|men, women| |name|`name`|name of a pair of jeans on the tag and style|categorical|identifier|

|fabric|`fabric`| |price|`price`| |maximum height of front pocket|`maxHeightFront`| |minimum height of front pocket|`minHeightFront`| |depth of front pocket rivet|`rivetHeightFront`| |max Width of Front pocket|`maxWidthFront`| |width of the top of the pocket hole|`minWidthFront`| |maxHeight of Back pocket|`maxHeightBack`| |minHeight of back pocket|`minHeightBack`| |maxWidth of Back pocket|`maxWidthBack`| |minWidth of Back pocket|`minWidthBack`|

2. Variable Exploration

Use the exploratory data analysis skills we've practiced to look for patterns/trends/relationships in individual variables as well as combinations of variables. A good EDA should always include the following:

- Numeric summaries or tables (depending on variable type) of the individual variables of interest
- Visual plot of the individual variable
- Numeric summaries or two-way tables of variable pairs (this is the most interesting part where we can look for relationships)

- Visual plots of the variable pairs (this is the most interesting part where we can look for relationships)

Never leave the summaries and plots by themselves. Include a brief description of the variable (remember to CUSS). Consider whether you see any interesting similarities or differences across groups or over time. You do not need to perform any formal statistical analysis on this data. Your statements can rely on visual or casual inspection of the data, in other words, just things you notice about the plots or summaries. Provide some EDA for the following variables from this data:

a. Price of jeans overall, and price grouped by gender

b. Maximum front pocket depth overall, grouped by style, and by gender

c. Minimum front pocket height and minimum back pocket height (try to include gender in your visualization for 2 extra credit points)

3. Create your own research question

One important skill for any researcher/statistician is the ability to create a *useful* research question that can be answered with the available data. For example, "Are pocket heights different in men's and women's jeans?" is an okay question to ask with this data. A better question is, "What is the average difference, in cm, between front pocket height on men's and women's jeans?" Try to be specific, and try to quantify (if possible) what you would like to answer. Good research questions usually include more than one variable as well, comparing groups or looking for patterns as values change. "Do women prefer smaller pockets?" is not a great research question because we don't have any data measuring satisfaction in the pockets provided for the wearer and there is no comparison to make.

a. Create your own research question based on variable combinations not used in part 2. Be sure to:

    - Clearly state your specific question
    - Provide some summary statistics for the variables you are using
    - Plot the variables and comment on any patterns (or lack thereof) you notice

b. Identify any potential limitations in the data for answering your research question. Is there any additional data you would like to see collected in the future to better explain/explore possible trends?