Project part 1 information

Statistics 3080

Fall 2023

The **ENTIRE** project is individual, and can only be discussed with the instructor and course assistants. Personal project questions should be asked in office hours and should **NOT** be posted to Ed Discussion.

Purpose

The purpose of part 1 of the project is to gain experience with formulating research questions, collecting and exploring approprate data, and effectively communicating results.

Process

Overview

The process for this part of the project includes identifying a topic of interest, formulating at least three associated research questions, gathering data from primary source materials that is appropriate to answer at least one of the research questions, and clearly describing and summarizing this data.

Details

1. Consider a topic of interest and formulate three specific research questions that are directly related to that topic. One of these research questions will be analyzed further in part 2 of the project. The research questions can explore any type of parameter – center, spread, proportion, etc. – but needs to focus on one or two specific parameters that can be analyzed with a single, stand-alone test. Part 2 analysis will **NOT** include linear regression, logistic regression, time series, or any other predictive modeling method.

   Examples of specific, appropriate questions:

   - **Specific and appropriate:** Did NBA players taller than 6'2" have a higher starting salary than NBA players shorter than 6'2" in the 2021-2022 season? **VS Not specific:** Do taller players make more money?

- **Specific and appropriate:** Was the U.S. price of a dozen eggs higher in months when the U.S. price of corn was above average from 2020 to 2022? **VS**
  Not appropriate: What factors related to raising chickens influenced the price of a dozen eggs in the U.S. each month from 2020 to 2022?

2. Find and compile current data related to the topic of interest that reasonably answers the three research questions. The final data should include at least four variables that relate to the research questions – different variables can relate to different research questions, but each research question should use variables from **at least two primary sources**. There is not a minimum number of observations, but collecting at least 10 observations is recommended. Data from the last five years is considered current.

   Examples of primary data sources (more information can be found in the helpful resources section below):

   - Government agencies
   - Reputable global organizations (WHO, UN, World Bank, etc.)
   - Reputable news organizations (including FiveThirtyEight)
   - Reputable non-profit organizations
   - Reputable researchers and/or academic journals
   - Sports leagues

3. After appropriate data for the research questions is collected, explore and present the data. Make the story in the data come to life using up to eight numerical and graphical summaries, including at least two numerical summaries, at least two graphical summaries, and at least one summary (either numerical or graphical) for each research question. All but one of each type of summary (numerical or graphical) must be more complex than counts. **Do NOT perform any statistical inference.**

Report

The 7-page report, due on October 26, should clearly detail the context of the posed research questions and the appropriateness of the collected data, summaries, and interpretation. The report must be a PDF knit directly from the provided project template. The report should include the following labeled sections.

- **Introduction:** Introduce readers to the chosen topic and research questions.
  - Explain why the questions are relevant and interesting beyond personal interest.
  - Clearly list each question in bold or in a bulleted list.
- **Data summary:** Describe the data collection process.
  - Clearly describe each of the primary sources used and explain their trustworthiness.
  - If the data represent a population, explain how the data were collected.
  - If the data represent a sample, explain how the data were selected and collected.
  - Explain any data modifications after collection and the reasons behind these changes.
  - Discuss potential issues with the data and their possible impact.
  - Explain why the data is appropriate to answer each research question.
- **Data dictionary:** Create a table that includes the name and description of each variable in the data.
  - Name each variable as it will be referred to throughout the report. In general, these names should be different from the column names in the data.
  - Describe each variable, including units of measurement and a definition of all possible categories, when relevant, assuming that readers have no previous knowledge about the data.
- **Data exploration:** Display selected numerical and graphical summaries of key features that tell the story in the data.
  - Choose the summaries selectively and with purpose considering the requirements and limitations given above. Many different summaries can be created, but not all contribute to effectively telling the story in the data.
  - Display all of the chosen summaries summaries (numerical and graphical) in this section and within the 7-page report limit. The size of graphical summaries can be modified, but they need to stay readable.
  - Appropriately label and title all elements of all graphical summaries.
  - Note that numerical summaries can be displayed graphically by displaying the numeric values on a corresponding graphic. Two summaries combined in this way are considered two of the eight summaries.

  - Create all summaries in R and all graphical summaries using `ggplot2`.

- **Conclusions:** Interpret the chosen summaries and make initial conclusions based on the information displayed in them, including which of the three research questions may be most interesting to analyze further.
  - Interpret the values, trends, and patterns displayed in the chosen summaries statistically.
  - Explain what the values, trends, and patterns displayed in the chosen summaries may mean or imply in the context of the research questions.
  - **Do NOT perform any statistical inference.**
  - Summarize the story in the data in the context of the topic of interest.

- **Data appendix:** Display a readable table of the first 10-20 rows of the data.
  - This section is not considered part of the 7-page report length limit.
  - Begin this section on a new page.

- **References:** List all references used in any standard format (APA, MLA, etc.)
  - This section is not considered part of the 7-page report length limit.
  - Begin this section on a new page.
  - Include sources used for background research and the primary sources used for data collection.
  - In the body of the report, use any standard format for citing specific information from these sources.

Helpful resources

Finding primary source data

- UVA Library guide for undergraduate statistics courses
- UVA Library Stat Lab
- UVA Library Research Data Services

Data exploration and visualization

- Exploratory Data Analysis (from *R for Data Science*)
- Data Visualization (from *R for Data Science*)
- Data Visualization Cheat Sheet for ggplot2
- From Data to Viz