Final Project

Pubhlth 390R: Introduction to data science using R

Ben Rogers
UMass Amherst

Due Dec 19th 11:59 pm

The goal of the project is for you to demonstrate that you have mastered the data visualization, transformation, and wrangling skills that we have studied in this course. You will do a data exploratory analysis by formulating questions, acquiring and analyzing data, and writing up your findings in the form of a brief report. You need to submit the report and all supporting R code and other files used to conduct your analysis.

The report should contain:

1. a description of your dataset, including necessary background, data source(s) and important variables;
2. questions of interest for your exploratory analysis;
3. a description of the analysis you performed using those data;
4. a discussion of the results of your analysis;
5. a summary of the conclusion(s) you have drawn on the basis of those results.

A detailed rubric which will be used to grade your submission is provided at the end of this document.

Rules

- This assignment may be completed either individually or in groups of two.
- This is an open book and take home project. Though discussion is encouraged, each student/group should finish the project on their own effort. Students can read and search related materials from both offline and online sources. Other authors' intellectual contributions (e.g. language, codes, figures, thoughts, ideas, expressions etc.) should be be properly cited if they appear in your project report. Note that it is easy for us to detect code plagiarism. It is also extremely easy for us to search the internet and see if the problems were actually posted online. If found, we will report this to the Dean's office and investigation will be made. Please adhere by your academic honor code and attempt the problem by yourself.
- The report should not exceed **5 pages** including everything (e.g. text, equations, tables, figures etc.) except the list of references. The list of references should be added at the end of report. The font size should be no smaller than 11 pt and the line space should be at least single spaced. The report should be in the format of pdf (you can save figures produced in RStudio and attach them in the report) or html. All tables and figures included in the report should be properly numbered. The sections and subsections should also be properly numbered. You must also upload all supporting R code (in an R script) and other files used to conduct your analysis. The quality and presentation of your code factors into your grade.
  - **Do not include the code in the report.** Instead upload it in a different R script file.
  - Do not copy output from RStudio. Instead, embed them into your sentences to explain the results. For example, you should write the average arrival delay is 20 minutes" instead of displaying the output from running your code.

- If any part of rule or project problem seems ambiguous to you, please contact the instructor for further clarification. You are encouraged to meet and discuss your final project with instructor or TA during office hours or making an appointment.

Advice

- Carefully choose which results are central to answer your questions and only use figures to present these results. Once you know what results to present, then choose the most effective graphical methods.
- Create figures such that readers who are unfamiliar with your research can immediately understand the results. Make your figures self-explanatory, e.g., label each axis clearly, use annotation to highlight important information. Also, make your figures beautiful, e.g., choose an appropriate binwidth for a histogram, avoid having too many lines in one figure.
- Do not add filler to get up to 5 pages. A lesser amount of high-quality content is better than a large amount of low-quality content.

Here is a brief example for the question "how do different factors of a diamond affect its price?"

We first explore the relationship between cut and price. Figure 1 is a boxplot for price split by cut. We find that the average price under the fair cut is the highest among all the cut levels. (You can add numbers here, e.g. the fair cut diamonds have an average price of xxxx, while the ideal cut diamonds have an average price of xxxx) This is counterintuitive because a better cut should correspond to a higher price. To investigate why this is the case, we explore the relationship between cut and carat in Figure 2, and the relationship between carat and price in Figure 3. From Figure 2, we find that a better cut is associated with a smaller value of carat (you can add numbers here); from Figure 3, we find that a larger value of carat is associated with a higher price (you can add numbers here). Therefore, the fair cut diamonds have larger size than the idea cut diamonds, and the size of a diamonds is positively associated with the price. This explains the counterintuitive phenomenon in Figure 1.

Data Sources

There are many potential sources of data available for you to analyze. A selection of packages and websites you may find useful:

- Kaggle: This is an online community of data scientists and machine learning practitioners. You can find tons of datasets. You may need to register first.
- General social survey: Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of US society.
- Lahman 100+ years of major league baseball data.
- Global health observatory: Global health statistics from the WHO.
- rnoaa Download data from NOAA.
- UNdata: Demographic data from the United Nations.
- pwt: This is an R package *pwt8*. The Penn World Table provides purchasing power parity and national income accounts converted to international prices for 189 countries for some or all of the years 1950-2010.

Of course, you are not limited to just these datasets, and are encouraged to find other sources of data on your own.

Examples of questions

Below I provide some examples of questions you may choose to investigate. You can come up with your own questions based on the data you choose. Some of the questions are (intentionally) vague and broad; it is not possible to fully answer them in the space allotted. It is up to you to narrow down to specific questions that you will then use data to answer. For example, if I would like to know "What perpetuates social inequality?", a specific question I could address based on a social survey dataset would be "How does access to education affect lifetime earnings?" I would then break this question into several steps:

1. What are the distributions (summary statistics) of education and lifetime earnings? Are there missing values of outliers, how should I deal with those?
2. Try different graphs to investigate the relationship between education and earnings, decide the best one to use in the report. Polish the graph.
3. Think about what further questions can be asked, e.g., is the relationship the same across different genders (racial groups etc.)? Does the relationship change over time?

Examples of questions

- What areas or groups of people were most affected by the Great Recession?
- How much did 9/11 cost the U.S./global economy?
- In what region(s) of the United States is income inequality the greatest, and why?
- What is the most employable major for someone graduating college in xxxx?
- Why do obesity rates vary greatly between different regions of the U.S.?
- What countries or regions have the most extreme life expectancy at birth, and why?
- What is the main determinant of whether someone is happy or not?
- In terms of weather, where is the safest/most dangerous place to live in the U.S.?
- What is the biggest potential natural disaster threatening America?
- Are environmental conditions in the United States improving or getting worse?
- What is an appropriate rate of compensation for an NCAA athlete?
- Who is the greatest (pick any sport) player of all time and why?
- How does income and education affect voting behavior?
- What determines a president's approval rating?

Rubric

- Content and presentation (8 pts.) This category judges the appropriateness and difficulty of the question you choose to answer, the overall strength of the argument you put forth to answer it, and the quality and correctness of your writing.
  - 0-2 pts:
    - Overly simplistic or unmotivated question.
    - Unclear what question is being answered, or what the conclusion(s) are.
    - Incoherent or disorganized writing. Large numbers of typos or grammatical mistakes.
    - Explanation is illogical or incoherent.
    - Data do not support conclusion.
  - 3-5 pts:
    - Appropriately motivated question and level of sophistication.

    - Question and conclusions are clearly stated.
    - Writing coherent and adequate to convey argument. May contain some spelling or grammar errors.
    - Explanation is correct and convincing.
    - Conclusion follows from data analysis, but may lack some evidence or be partially complete.
  - 6-8 pts:
    - Well-motivated, insightful, and/or interesting question.
    - Question and conclusions are clearly stated.
    - Clear, efficient writing. Almost no spelling or grammar errors.
    - Explanation is correct, complete, and elegant.
    - Conclusion follows strongly from the data analyzed.

- Technical correctness (12 pts.) This category scores the choice of analyses and visuals (plots and tables) that you use to support your argument.
  - 0-3 pts:
    - Visuals contain obvious errors.
    - Wrong plot types for data analyzed.
    - Visuals do not convey useful information; are unclear, unlabeled, uninterpreted in text; do not contribute to conclusion.
    - Statistical analyses, if used, are inappropriate, extraneous, or do not support conclusions.
  - 4-8 pts:
    - Visuals are free of obvious errors, and are appropriate for data analyzed.
    - Visuals convey information which is relevant to conclusion, but may lack context or interpretation.
    - Visuals are clearly labeled.
    - Statistical analysis supports conclusion.
  - 9-12 pts:
    - Visuals are correct, appropriately labeled and organized, and contribute integrally to argument.
    - Elegant and correct use of statistics to support conclusion.

- Code quality and reproducibility (10 pts.) This category scores the correctness, legibility, and efficiency of the code you use to conduct your analysis.
  - 0-2 pts:
    - Results cannot be reproduced and/or code does not run.
    - Code is illegible, uncommented, poorly organized, and/or contains major bugs.
  - 3-7 pts:
    - Code runs and reproduces the plots, tables, and results discussed in text.
    - May contain minor, non-obvious bugs.
    - Code has some comments and is organized, and contains little or no unused code.
  - 8-10 pts:
    - Code is clean, commented, compact, bug-free, and reproduces all results with minimal effort.
    - No irrelevant or distracting code.
    - Has meaningful object names that illustrate their purpose.
    - Grader can easily read your code and understand what it does without even having to run it.

- Innovation (5 pts.) This category scores originality/inventiveness of your analysis.
  - 0-1 pts:

    - Question, conclusion, and/or analysis are copied from book/lectures/online sources, potentially with minor modifications.
    - Sources have not been provided.
    - Data sets used have been extensively analyzed in book, class, homework, or online: mpg, flights, diamonds, who, etc.
    - Level of sophistication of plots, tables, and statistical analysis is below that of problem sets and lecture.
- 2-3 pts:
    - Student has come up with their own question, and answered it using original ideas.
    - Sources provided.
    - Analysis include new sources of data that we did not thoroughly analyze in class
    - Level of sophistication of plots, tables, data manipulation, and statistical analysis equals that of problem sets and lecture.
- 4-5 pts:
    - Student has successfully answered a unique, interesting, and/or ambitious question.
    - Student has learned to correctly use new types of plots, statistical techniques, or data sources not covered in class.