CMPSC 301
Data Analytics
Summer 2021

Course Final Project

Please submit your work to your GitHub repository by the due date.

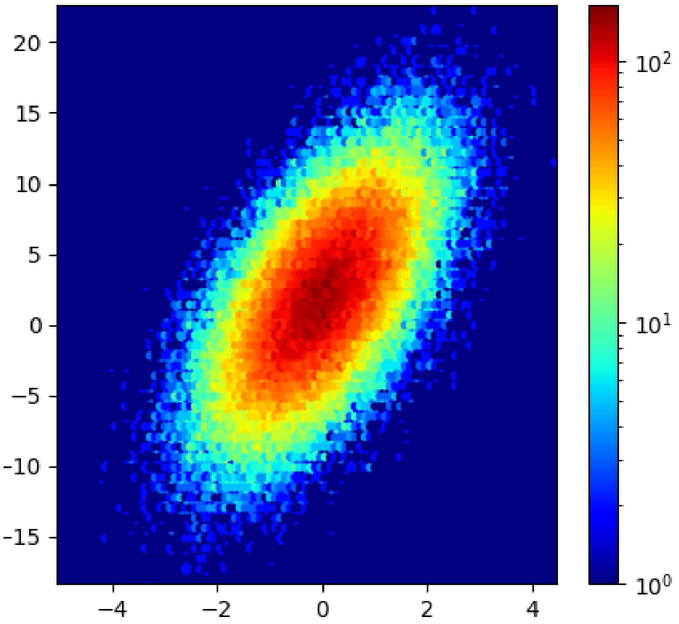

Figure 1: The data, when in textual form, is an unreadable script that generally tells us nothing of its story. However, by employing the actors 'color' and 'texture' to play out their script, the characters, 'patterns' and 'trends' become more developed to take the center-stage and steal the show.

Summary

The final project invites you to employ the methods explored in this course to conduct a comprehensive analysis of a real-world data set. You will select an application area and exploratory questions that are of interest to you, find an appropriate data set, conduct an in-depth analysis of this data set, produce plots, and examine your findings in the context of an application area. As you work, keep your exploratory questions in light of the issues of ethics, privacy, and power dynamics.

During the analysis process you will carry out the steps of data collection, cleaning and transformation (as necessary), wrangling, correlation, modeling as necessary and visualization to be able to tell a story from your data concerning some type of trend, as noted in Figure 1.

Since much of data analysis is to provide some type of *visually communicable* information to be used to change policy, or create awareness for some reason, your report is to argue for or against the continuance of a particular policy, either instated or potential. In other words, your report is to introduce its pieces of analysis as a way to influence a policy (of some type). You are at liberty to select a real-world policy to contest, or to provide the discussion of a potential policy that you believe to be a benefit after an analysis of its data.

Assignment Specifications

This project is broad and you may use whatever tools and skills that we have covered in this class to complete your study. For the project assignment you have to select one application area that is of interest to you from which you can obtain data (e.g., health, politics, economics, etc.). You should choose a broad exploratory question(s) to consider in this area. Then, while keeping in mind your selected area and questions you would like to explore, find a specific real-world data set that you can analyze. Finally, you are to conduct a comprehensive analysis of your selected data set, answering questions you have designed, creating new questions to ask, and comment on any issues with the data or its analysis.

You may use anything and everything we have learned (or will learn) in class and also you should research additional resources beyond of what we discussed in class. You may also extend any of the programs or concepts we have developed in the labs or in class. However, you are strongly encouraged to find new (publicly) available datasets for your study using online searches.

GitHub Starter Link

https://classroom.github.com/a/GKqmjZNW

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.
- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):

  ```
  – git commit <nameOfFile> -m ‘‘Your notes about commit here’’
  – git push
  ```

  Alternatively, you can use the following commands to add multiple files from your repository:

  ```
  – git add -A
  – git commit -m ‘‘Your notes about commit here’’
  – git push
  ```

Requirements

1. **Do your reading**: You are expected to consult our course textbooks Silge *et al.* [1] and Wickham *et al.* [2] to complete this work.

2. **Literature requirement**: Research relevant background and find at least two (2) **academic references** related to the selected area and your exploratory questions.

   Please do not use blogs or web sites at your references. Much of this text is likely to be unsubstantiated since it has not been subjected to an academic peer-review panel. Instead you are to use Google Scholar to locate peer-reviewed and scholarly articles which have been published by a reputable organization, and contain factual information.

3. **Scope of your study**: Determine what you would like to research. Isolate your research question into some manageable articulation that you will be able to address using an analysis of data. Try to be realistic in how you choose your research question: do not choose a topic which has too many smaller pieces that must be researched before your actual question may be addressed by analysis for discovery and conclusion.

4. **Data**: Select a **large-size**, **real-world** data set to investigate your phenomena. Your data must be free, public and available online. Your data should also be credible and originate from sources of good standing. Please perform necessary searches to locate public and credible data sets are able to be referenced in articles. To give you ideas, there is a list of sites (below) that specialize in providing publicly available data.

   - Pelletier Library at Allegheny College (online services): `https://allegheny.libguides.com/az.php`
   - World Health Organization: `http://www.who.int/`
   - The World Bank: `https://www.worldbank.org/` and `https://www.who.int/ncds/surveillance/en/`
   - Demographic and Health Surveys: `https://dhsprogram.com/`
   - Harvest Choice: `https://harvestchoice.org/`
   - Food and Agricultural Organization: `http://www.fao.org/home/en/`
   - World Population Prospects: `https://population.un.org/wpp/`
   - Centres for Disease Control and Prevention (CDC): `https://www.cdc.gov/`
   - US Food and Drug Administration Home Page: `https://www.fda.gov/`
   - The US Census: `https://www.census.gov`
   - Institute for Health Metrics and Evaluation: `www.healthdata.org/`
   - IBM's collection of opensource data sets: `https://developer.ibm.com/exchanges/data/`
   - Google's opensource data sets: `https://research.google/tools/datasets/`
   - Data.world: data for business-based questions: `https://data.world/`
   - Kaggle: `https://www.kaggle.com/`

  - Kaggle's Star Trek Scripts (Could be a cool idea!): `https://www.kaggle.com/gjbroughton/start-trek-scripts`
- And many more that you may conveniently find using online searches. Please remember to cite your data in your report; giving the name and its web address.

Wrangling: It may be necessary to clean and transform the data using functions such as `filter`(), `mutate`(), and similar from class. In addition, you will be asked to show the code and to justify all steps taken to treat (i.e., organize) your data.

5. **Analysis**: In your report, identify the method of your analysis: what will you measure and which techniques were required? How did you treat and detect this measurement?

6. **Design and development**: Develop computational techniques (i.e., R code and programs) using R's software libraries to conduct your analysis. Your analysis should include some of the basic statistics on the data to provide a global view, as well as, steps to explore the relationships between variables. If you are building a model to make predictions, please try to confirm/deny a hypothesis.

7. **Plots**: Making plots, summaries and interpretations of results. *You must have visualizations to show your results.* You must also address any data or inherent flaws and faults of the data which cannot be easily corrected (i.e., missing data entries, data collected on skewed population, too few data-points and etc.) You are to determine some of the reasons to explain biases, discrimination, stereotypes, etc. that may be present during collection, analysis, and reflect on the latent trends in real-world data sets.

Assignment Specifications and Due Dates

1. **Proposal** (at least one paragraph) Deadline: Monday, 2nd August by 5:00pm EST:
Here you are to write about 100 words in a markdown document to describe what you intend to study for your project. Be sure to include the data, its reference, and details about the question that you are asking of the data. If you are able, describe the form of analysis that you anticipate using.

2. **Presentation** Thursday, 5th August 2021, during class: In the presentation, you should describe the motivation, definition, challenges, approaches, and results and analysis. Rather than employing long sentences and complicated equations in your presentation, please be visual in your discussion. For this, please show plots and use a few bullet points to describe your results. The goal of the presentation is to convey the important high-level ideas and give intuition rather than be a formal specification of everything you did.

   Prepare for ∼ 7 minute presentation. Design at least five slides, including a slide with the title of your project and your name. If possible, please run your code after your presentation to show your analysis in action.

3. **Full Project Report and Code**: Friday, 6th August 5:00pm EST (report of at least 1000 words in length):

- Your code should run correctly and produce results discussed in the final report. Your final report should be clear, concise and, most importantly, well written, this includes no typos or grammatical errors. Your report should be written in a professional manner and should include explanation of all of the requirements outlined above.

Grading Rubric

1. **Proposal**: 10 points
2. **Presentation**: 20 points
3. **Final report and project implementation**: 70 points

For your deliverable, you are to submit Markdown files for your written work and your code in R. For your final report you are to submit any necessary and supplementary material. This includes programs, data sets, a *README.md* Markdown file documenting what everything is (i.e., a justification of the existence of the files that you have left for the instructor in your repository). Finally, for your code, you will need to write up documentation to instruct how the code is to be used and what its expected inputs and outputs should be. Please note that if you are creating extra software for some purpose, please provide documentation to instruct how to use it.

Required Deliverables

Note: Please remember to include your name on everything you submit for the class.

1. **Proposal**; File: `writing/proposal.md`: Your proposal should be about a paragraph.
2. **Presentation**: Your presentation is to be given in class and should not be more than seven or eight minutes in length. Please use slides (Google Docs, for example) to share with your discussion.
3. **Report**; File: `writing/report.md`: Your report should be about 1000 words in length and include graphics.
4. **Source code**; `src/analysis.r`: Your code that can be run to load the data files and to produce the plots and analysis of your work. Please add documentation to your code to help the instructor understand your thinking behind the code on a line-by-line basis.

Honor Code

In adherence to the Honor Code, students should complete this assignment while exclusively collaborating with the other member of their team. While it is appropriate for students in this class who are not in the same team to have high-level conversations about the assignment, it is necessary to distinguish carefully between the team that discusses the principles underlying a problem with another team and the team that produces an assignment that is identical to, or merely a variation on, the work of another team. Deliverables from one team that are nearly identical to the work of another team will be taken as evidence of violating Allegheny College's Honor Code. Do not be tempted to look online for possible problems and solutions, that institutes a violation to the Honor code! Please be original!

References

[1] Julia Silge and David Robinson. *Text mining with R: A tidy approach.* " O'Reilly Media, Inc.", 2017.

[2] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data.* " O'Reilly Media, Inc.", 2016.