CMPSC 301
Data Analytics
Spring 2020

Course Final Project

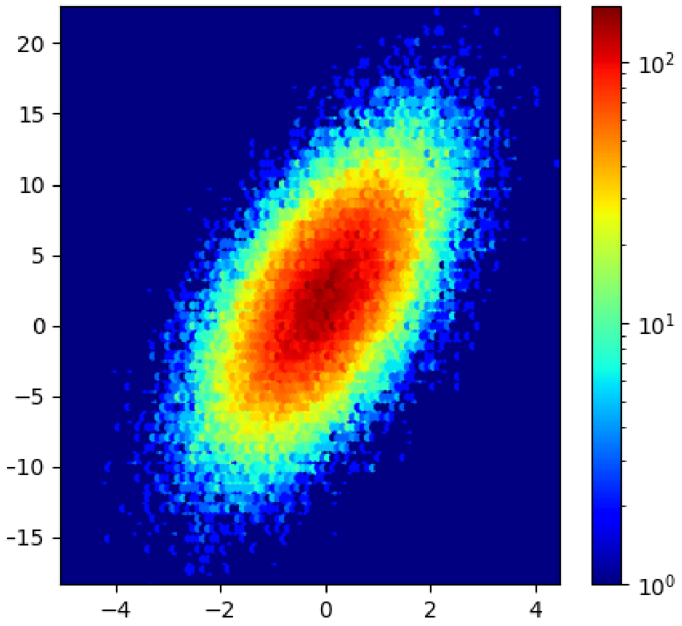

Figure 1: The data, when in textual form, is an unreadable script that generally tells us nothing of its story. However, by employing the actors ‘color’ and ‘texture’ to act out this script, the characters, ‘pattern’ and ‘trending’ become more developed to take the center-stage and steal the show.

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, and examine your findings in the context of the application area and your exploratory questions while keeping in mind 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 and modeling, if necessary, and visualization to be able to tell a story from your data, as noted in Figure 1. Note: This project should contain the same amount of work as about three labs and will not include a presentation.

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 your group believes to be a benefit after an analysis of its data.

All of your project deliverables should be submitted through a project repository after you or your group has accepted the project assignment. For your final project, you can work individually or in groups of up to four people. If you decide to work in a group, each member of the group will be evaluated separately based on his or her contributions to the project. This evaluation will be determined largely from the feedback of the group members. **As always, please be sure to include all the names of the group members.**

Assignment Specifications

For the project assignment you have to select one application area that is of interest to you from which you can draw 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 commenting 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 requested to find new datasets to for your study which were not covered in class.

GitHub Starter Link for Groups

STOP! STOP!
Not everyone will be clicking this link at this time!
Only the team leader will be clicking the link to create the repository!!
`https://classroom.github.com/g/uMbUXOhT`

Creating Your Repository

If you would like work as a group, then the option is open to you. If you would like to work individually, then you will be in a group of your own on GitHub Classroom for this assignment. For group assignments **only one person will be creating the team while the other team member will join that team.** Please form a team of **no more than two people** and select one person to create the repository.

The selected person of the team should go into the link to the lab in the assignment sheet. Copy this link and paste it into your web browser. Now, you should accept the laboratory assignment and create a new team with a unique and descriptive team name (under “Or Create a new team”).

Now the other members of the team can click on the assignment link and select their team from the list under “Join an Existing Team”. When other team members join their group in GitHub Classroom, a team is created in our GitHub organization. Every team member will be able to push and pull to their team’s repository.

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 locally
- As you are working on your lab, you are to commit and push regularly. The commands are the following.
  - `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 five (5) academic references related to the selected area and your exploratory questions. **Please do not use blogs or web sites for this work. 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 library resources or 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 question into some manageable articulation that your group and 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 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.

Wrangling: It may be necessary to clean and transform the data. In addition, you will be asked to justify all steps taken to treat your data, or to explain why such steps were, or were not, taken.

5. Identify the method of your analysis: what will you measure and which techniques will be required. How will you treat and detect this measurement?

6. Develop computational techniques (i.e., R code and programs) to conduct your analysis. Your analysis must include basic statistics on the data, as well as exploration of the relationships between variables and/or modeling of data. Your analysis may try to discover new features in the data or try to confirm/deny a hypothesis.

7. Summarize and interpret your 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 datapoints 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. Project ideas Friday, 3rd April by midnight:

   - Begin your project by considering three ideas. For each of your ideas, submit one or two sentences to explain the idea in some detail. You will get some feedback on your Issue Tracker from the instructor.

2. Proposal Friday, 10th April by midnight (at least 500 words):

- Develop an idea for your project including preliminary research on the importance of the questions you decided to consider and data availability. Your proposal should include at least five references that motivate the importance of your selected area of exploration. You do not need to include any specifications on how exactly you will analyze at this point, however you should discuss the data set that you will analyze and potential techniques/tools you may be able to utilize for your project.

3. Progress Report Friday, 17th April by midnight (at least 500 words):

    - Here you are to describe what you have been working on so far, and to discuss some of the challenges that you have encountered and how these challenges were met. By this point, you should have conducted necessary research on the background, examined in detail the data set you have selected, decided on the approach you will use to analyze it, and made a significant progress towards implementation. Be sure to include discussion of issues concerning data, programming and similar for your concept demonstration.

4. **Presentation**: None required for this semester due to the college closing as a result of the COVID-19 health concern.

5. Full Project Report Thursday, 30th April by midnight (at least 2000 words):

    - Incorporate any feedback from the progress report and the presentation session. 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. **Ideas**: 5 points
2. **Proposal**: 10 points
3. **Progress report**: 25 points
4. **Final report and project implementation**: 60 points

For each deliverable, you are to submit Markdown files for written work. 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.

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.