AIT 580 – Final Project

Submission to BOTH your **Github** and **Blackboard**

Data

Select one dataset from the list below and create a Github-based Jekyll Website that shows the meaningful data analysis results for key stakeholders.

1. **Consumer Complaint Database** (Data contains complaints received about financial products and services such as bank account, credit card, loans etc. Good for NLP as well. https://catalog.data.gov/dataset/consumer-complaint-database
2. **Department of the Treasury Income Tax Return** (This study provides detailed tabulations of individual income tax return data at the state and ZIP code level. You can use the most recent year data only for analysis, or you can compare two consecutive years in your analysis) https://catalog.data.gov/dataset/zip-code-data
3. **College Scorecard Data** (College Scorecards make it easier for students to search for a college that is a good fit for them. They can use the College Scorecard to find out more about a college's affordability and value so they can make more informed decisions about which college to attend) https://catalog.data.gov/dataset/college-scorecard
4. **Local Data for Better Health** (This dataset includes estimates for 27 measures of chronic disease related to unhealthy behaviors, health outcomes and use of preventive services for 500 cities. These estimates can be used to identify emerging health problems and to inform development and implementation of effective, targeted public health prevention activities) https://www.cdc.gov/places/index.html
5. **NOAA Storm Database** (Contains chronological listing, by state, of hurricanes, tornadoes, thunderstorms, hail, floods, drought conditions, lightning, high winds, snow, temperature extremes, statistics on personal injuries and damage estimates. Storm Data covers the United States of America) https://www.ncdc.noaa.gov/stormevents/

Website Structure

Please follow the structure below in your website and make sure to include all the sections with detailed briefing on each of them. Each section can be either headings in the same page, or menu items – the design of the websites is up to the student teams. Depending on the website design, tweaking the title of each section is allowed (e.g., if "Brief Introduction" is not a good stylish choice for the web page, you can just say "Background" etc.). The important thing is that the websites need to include all the components specified here (even though the wordings can be different).

(0) Title of the Website or Project

(1) Brief Introduction (about overall project)

- Briefly explain what problem you're trying to solve through data analysis.
- Who can benefit from your data analysis (i.e., who are stakeholders)? Specify detailed

justifications on why your analysis will be beneficial to particular groups of people, researchers, or organizations.

(2) Nature of the Data Curation

- Who (company, agency, organization) collected the data?
    - Who they are, what do they do?
- Why did they collect the data (purpose)?
- What is the nature of the data given the purpose of the data collection (e.g., any bias)?
    - Usually, many datasets are NOT collected for data scientists themselves, but as a byproduct of organizational process. Because of this reason, it is important to understand the nature of the data.
- Given the nature of the data, how can you adjust and leverage the data (i.e., what are pros and cons of the data and how can you overcome it)?
- Is there any privacy, quality, or other issues with this data?

(3) Questions

- What are the main questions of your interests that can be answered through the data that you chose?
    - List some specific questions, and be sure to answer them in your analysis.
    - Provide justifications on why your question(s) are important for stakeholders.

(4) Requirements and Resources needed

- What software and hardware resources you have used in this project?
- What kinds of pre-processes were needed to make use of the data, and why?
- What are the advantages and limitations of the target dataset in answering your questions?

(5) Descriptive Analysis

- Briefly describe the dataset
- Prepare and describe relevant metadata (types of attributes/variables in the dataset)
- Provide some descriptive statistics so readers can understand the data without actually looking into it (e.g., mean, SD, frequency, distribution, network structure, etc.)
- Explore the dataset using relevant tools discussed in the course such as R, SQL, Python, Tableau etc., and prepare relevant descriptive statistics and visualizations either as static images (e.g., graphs) or dynamic visualizations on the web. However, you don't need to analyze all the items in the dataset – just focus on descriptive statistics that is necessary for the main audience in understanding your data.

(6) Results/Findings

- Analyze data using R or Python so that you can provide answers to your questions.
- Provide justifications on how and why each of your analysis answers your question(s).
    - If you have hypothesis, provide hypothesis testing results and justifications on why they do or don't make sense.
- Include the result for each of the analyses. You can add more than one analysis for each of them but at least one analysis on each of them is required. Your results could be represented as one or more of the followings (but not limited to these):
    - Scatterplot
    - Boxplot
    - Correlation analysis (visualization or table)

    - Regression analysis (visualization or table)
    - Hypothesis test (table)
    - Network visualization (color coding or size of node/edge)
    - Tree map
    - Map-based visualization (geospatial data)
- Interpret the results
    - What conclusions can be supported for each analysis?
    - This should reflect answers to the specific questions specified in the "Questions" section.
- Graphics must follow good visualization practices discussed in course lectures (e.g., 6 principles).

(7) Limitations

- Discuss limitations of your analysis.
- Provide the future work that can improve your analysis to answer your question(s).

(8) References

- Provide appropriate citations and references (if any)
    - Include citation for the dataset as well: http://infoguides.gmu.edu/citingdata
    - Other references such as books and articles should follow APA format: https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_overview_and_workshop.html
    - Making general arguments or combining supplemental data without citation could be seen as plagiarism.

(9) Explain/define Terms

- Include explanation of any technical terms relevant to the project

Guideline for the Websites

- Provide the names of the team members on the website.
- The website files need to be hosted under any of the team members' assignment Github repository (Grading is done not only based on the websites themselves but also for the code).
- The websites should not be text-heavy, rather it should be easily readable with many graphics and visualizations. In other words, the websites should not be a text-heavy report; it should be appealing to the audience.

Submission Guideline

1. **Blackboard Submission**
    a. Submit the URL & the screenshot of your Final Project website to Blackboard.
2. **Github Submission**: Your data analysis source code
    a. Make a folder named "`docs`" in your Github repository. Jekyll-based Github websites are automatically loaded from this folder.
    b. Put ALL the scripts and Jekyll files that you generated for your final project in this folder. For the data analysis scripts, make a separate folder under the "analysis" folder, which should be located under the "docs" folder.
    c. Anybody should be able to reproduce your work. If necessary, provide explanations on how to run your scripts such as the script execution order in the "README.md" file.

Grading Criteria

- Did the team present the website and data analysis well in the final presentation, with clear presentation of the problems, solutions, and data analytic methods? [20%]
- Did the team members comment on other team members' presentations at least once, with a constructive feedback? (Constructive feedback means some comments that are helpful for improving the projects, rather than simple compliments) [5%]
- Does the website running well on Github using Jekyll? [10%]
- Does the website make it clear about what the problem is, who the main stakeholder is, and why proposed data analysis results would benefit the key stakeholders? [10%]
- Are the question(s) well-developed and are conclusions reasonable? [10%]
- Are the analysis plans/methods well-designed to answer the question(s) [5%]
- Have the data analysis been conducted correctly to answer the question(s) [10%]
- Do the websites provide reasonable interpretations of the analysis results? [10%]
- Are the data analysis code reproducible (i.e., can anybody run the scripts in Github without any problems/errors)? [10%]
- Do the websites include all the required components without errors? [10%]
- Late-submission policy
    - When the submission is late, 5 points will be deducted every day from the deadline.
    - For example, if you miss the deadline and submitted it within 24 hours after the deadline, the final score will be: your report score – 5. If your submission were late by 2 days after the deadline, it will be 10 points deducted from your original score, and so on.
    - Accordingly, if you submit the final report 20 days later after the deadline, your score becomes 0 regardless of your submission.