

**CMPSC 390  
Data Analytics  
Fall 2017**

**Course Project  
Given: 13<sup>th</sup> October, 2017**

## **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.

For your final project, you can work individually or in groups of two or three. 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. If you work in a group, you have to create a new project repository, named `cs390f2017-groupName`, that can be shared between your team members and your instructors. Here, 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 not allowed to use the data sets that we have explored during the labs or in class.

## **Requirements**

1. Research relevant background and find at least 5 references related to the selected area and your exploratory questions.
2. Select a **large-size, real-world data** set to investigate your phenomena. 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 not taken. Note: There is much free and

public data available online. Please perform necessary searches to locate public and credible data sets are able to be referenced in articles (such as the data from the WHO and similar).

3. Identify the method of your analysis: what will you measure and which techniques will be required.
4. Develop computational techniques (e.g., 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.
5. Summarize and interpret your results. You must have visualizations. 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.

### Timeline: Deliverables

1. **Proposal** (at least one page) **Deadline: Monday, October 30<sup>th</sup>, 2017 by 5:00pm:**  
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.
2. **Progress report** (3-4 pages) **Deadline: Friday, 17<sup>th</sup> November, 2017 by 2:30pm:**  
Describe everything you have done so far in your progress report. 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. Describe anything new that you have learned so far and any unexpected challenges that you have encountered.
3. **Presentation** **Friday, 8<sup>th</sup> December, 2017, during the lab session (another date may be added, depending on the number of projects):** In the presentation, you should describe the motivation, definition, challenges, approaches, and results and analysis. Use diagrams and a few bullet points rather than long sentences and equations. 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 ~ 10 minute presentation. Design at least five slides, including a slide with the title of your project and group members' names. Every member of the group needs to contribute to the presentation talk. At the end of the presentation give a demonstration of your project.
4. **Final report** (6 or more pages) **Deadline: Friday, 18<sup>th</sup> December, 2017 by 9am:** 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

10 points: **Proposal**

20 points: **Progress report**

25 points: **Presentation and demonstration**

45 points: **Final report and project implementation?**

For each deliverable, you are to submit a PDF file with your report (or presentation slides). For your final report you are to submit any necessary and supplementary material. This includes programs, data sets, a *README.txt* 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.

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!