CS181: Data Systems

Project 1: Raccoon Creek

1 Overview

For the tabular data unit project we will conduct a data-driven investigation of the Raccoon Creek Watershed. The specific Raccoon Creek is the one that originates Southwest of Johnstown, OH (as Kiber Run) and runs East by Southeast through Granville and Newark until it conjoins the South Fork of the Licking River just South of Newark, OH. There are multiple Raccoon Creeks in Ohio, so be sure to identify the correct one.

The USGS maintains a river flow gauge on Raccoon Creek, near Granville, OH. The data from this gauge is publicly available through the USGS websites. Part of your dataset is a file that contains data from this gauge.

The NOAA maintains weather stations across Ohio (and the greater United States). Again, the data from these weather stations is publicly available on the appropriate government websites. A second part of your dataset comes from this website.

Your job in this assignment is to use these two data sources to create a *data story* about the connection between the weather data and the creek data. You will create a jupyter notebook using your knowledge of python to facilitate the analysis of the data story.

2 Evaluation Criteria

The quality of the data story that students tell is based on many aspects, several of which are non-technical. This course is about data systems. Obviously the project is a place to demonstrate the technical skills students have learned. But students often concentrate only on the technical and fail to address the important non-technical aspects of the data story.

2.1 Technical Content

This project is an opportunity to apply our tabular data skills in this setting. Here are some suggestions for things to consider:

- Appropriate reading of data files.
- Appropriate use of data structures to store information.
- Appropriate operations to manipulate data to uncover data story.
- Using appropriate python procedures to perform data manipulations efficiently.
- Attention to missing data values.

- Adherence to correct normalization of tabular data (Tidy Data format).
- Demonstrating good python programming skills.
- Arriving at a data story based on sound analysis of the data.
- Creation of quality visual products to illustrate the data story.

2.2 Background Research

A data story is richer if it provides context for your reader. How might you anticipate and handle questions that naturally arise in your readers?

- Where is Raccoon Creek? What is the Raccoon Creek watershed? What is the nature of the creek (size, length, depth, etc..)?
- Where are the sensors that record the data?
- Who owns and maintains these sensors?
- Where is the data publicly available?
- What are the measurements that are being provided in the data story? How can you explain or relate these measurements to non-technical (non geoscience) people?
- How will you set the background in your notebook story? What graphics might you include in your story?

2.3 Writing Quality

It can be an effective strategy to think of a data story (notebook) as a research paper. The same qualities that make great research papers (thesis, organization, audience, style, mechanics) will make great data stories.

- What is your thesis? You probably won't know this until you complete the investigation of the data. What single compelling discovery drives the data story? How will you communicate your thesis in the notebook format?
- Who is your audience? Hint: it is not your course instructor. Think about publishing or releasing this data story to a general science audience. Who will read this? What diversity can you expect in their backgrounds and levels of expertise? These people will not know about "the assignment", so avoid any references to the assignment. Your data story will be much stronger as a stand-alone product.

- Organization? The notebook format both presents both some challenges and some advantages over a traditional research paper's organization. What structure will you use to tell your data story? How will you use the notebook format to present this story?

- Style? The characteristics that make good science writing in expository papers will likely translate smoothly over to the notebook format. Think about tone. Wordchoice. Phrase composition. Sentence structures and variations.

- Mechanics. This is a published product and should adhere to the conventions of good writing mechanics (grammar, punctuation, conventions, etc.). How will you ensure your data story exhibits good writing mechanics? Writing Center appointments? Reviews?

2.4 Reproducability

Traditionally published research is built on a ”trust me” model. The author basically implies that any reader should ”trust me, I wrote the software correctly, interpreted the results correctly, and have reached valid conclusions.” There are reviewers – technically sound people who review the research before it is accepted for publication. But these reviewers often rely on the integrity of the conclusions without being able to verify them personally. Rarely does someone publish work that is intentionally deceptive in its conclusions, but it does happen. More often, mistakes are made in the analysis that lead to faulty conclusions.

There is a move underfoot to make published results *reproducible*. In addition to releasing the published papers, researchers will also release their datasets and their code. These can then be subject to public scrutiny for interested readers. And readers can reproduce the results for themselves. Because journals are published in a paper format (or electronic paper), the additional materials – the code and the data – are often stored separately, such as in a github repository.

The jupyter notebook format allows for published reproducibility in a wholistic way. The notebook functions as both a research paper and as a computational device. Both the written story and the code that produced the story are present. And they are woven together seamlessly by effective journalism. Think about how you can take advantage of the notebook format to tell a data story – an active version of a written research paper.

3 Logistics

- This is a group project. You must share the work equitably with your partner; each person must contribute to the project.

- Complete the project in a single notebook. Submit your notebook to notebowl.

- You should use data from the two source files provided: `RaccoonCreekFlowData.txt` and `PrecipData.csv`. You are not required to, but you are permitted to augment these data sources with other related data that you find. If you desire to make substitutions to the original datasets, please seek permission from the instructor.