MCIS6273 Data Mining (Prof. Maull) / Fall 2021 / HW0

This assignment is worth up to 20 POINTS to your grade total if you complete it on time.

Points Possible	Due Date	Time Commitment (estimated)
20	Wednesday, Sep 1 @ Midnight	*up to* 20 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Familiarize yourself with the JupyterLab environment, Markdown and Python
- Familiarize yourself with Github and basic git
- Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework
- Listen to the Talk Python[‘Podcast’] from June 25, 2021: A Path to Data Science Interview with Sanyam Bhutani
- Explore Python for data munging and analysis, with an introduction to CSV and Pandas

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw0`. Put all of your files in that directory. Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hw0_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(0%) Familiarize yourself with the JupyterLab environment, Markdown and Python

As stated in the course announcement Jupyter (https://jupyter.org) is the core platform we will be using in this course and is a popular platform for data scientists around the world. We have a JupyterLab setup for this course so that we can operate in a cloud-hosted environment, free from some of the resource constraints of running Jupyter on your local machine (though you are free to set it up on your own and seek my advice if you desire).

You have been given the information about the Jupyter environment we have setup for our course, and the underlying Python environment will be using is the Anaconda (https://anaconda.com) distribution. It is not necessary for this assignment, but you are free to look at the multitude of packages installed with Anaconda, though we will not use the majority of them explicitly.

As you will soon find out, Notebooks are an incredibly effective way to mix code with narrative and you can create cells that are entirely code or entirely Markdown. Markdown (MD or `md`) is a highly readable text format that allows for easy documentation of text files, while allowing for HTML-based rendering of the text in a way that is style-independent.

We will be using Markdown frequently in this course, and you will learn that there are many different "flavors" or Markdown. We will only be using the basic flavor, but you will benefit from exploring the "Github flavored" Markdown, though you will not be responsible for using it in this course – only the "basic" flavor. Please refer to the original course announcement about Markdown.

§ **THERE IS NOTHING TO TURN IN FOR THIS PART.** Play with and become familiar with the basic functions of the Lab environment given to you online in the course Blackboard.

§ **PLEASE *CREATE A MARKDOWN DOCUMENT* CALLED `semester_goals.md` WITH 3 SENTENCES/FRAGMENTS THAT ANSWER THE FOLLOWING QUESTION:**

- **What do you wish to accomplish this semester in Data Mining?**

Read the documentation for basic Markdown here. Turn in the text `.md` file *not* the processed `.html`. In whatever you turn in, you must show the use of *ALL* the following:

- headings (one level is fine),
- bullets,
- bold and italics

Again, the content of your document needs to address the question above and it should live in the top level directory of your assignment submission. This part will be graded but no points are awarded for your answer.

(0%) Familiarize yourself with Github and basic git

Github (https://github.com) is the *de facto* platform for open source software in the world based on the very popular git (https://git-scm.org) version control system. Git has a sophisticated set of tools for version control based on the concept of local repositories for fast commits and remote repositories only when collaboration and remote synchronization is necessary. Github enhances git by providing tools and online hosting of public and private repositories to encourage and promote sharing and collaboration. Github hosts some of the world's most widely used open source software.

If you are already familiar with git and Github, then this part will be very easy!

§ **CREATE A PUBLIC GITHUB REPO NAMED `"mcis6273-F21-datamining"` AND PLACE A README.MD FILE IN IT.** Create your first file called `README.md` at the top level of the repository. You can put whatever text you like in the file (If you like, use something like lorem ipsum to generate random sentences to place in the file.). Please include the link to **your** Github repository that now includes the minimal `README.md`. You don't have to have anything elaborate in that file or the repo.

(0%) Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework

The Linux console in JupyterLab is a great way to perform command-line tasks and is an essential tool for basic scripting that is part of a data scientist's toolkit. Open a console in the lab environment and familiarize yourself with your files and basic commands using git as indicated below.

1. In a new JupyterLab command line console, run the `git clone` command to clone the new repository you created in the prior part. You will want to read the documentation on this command (try here https://www.git-scm.com/docs/git-clone to get a good start).
2. Within the same console, modify your `README.md` file, check it in and push it back to your repository, using `git push`. Read the documentation about `git push`.
3. The commands `wget` and `curl` are useful for grabbing data and files from remote resources off the web. Read the documentation on each of these commands by typing `man wget` or `man curl` in the terminal. Make sure you pipe the output to a file or use the proper flags to do so.

§ **THERE IS NOTHING TO TURN IN FOR THIS PART.**

(30%) Listen to the Talk Python[‘Podcast’] from June 25, 2021: A Path to Data Science Interview with Sanyam Bhutani

Data science is one of the most important and “hot” disciplines today and there is a lot going on from data engineering to modeling and analysis.

Bhutani is one of the top Kaggle leaders and in this interview shares his experience from computer science to data science, documenting some of the lessons he learned along the way.

Please listen to this one hour podcast and answer some of the questions below. You can listen to it from one of the two links below:

- Talk Python[‘Podcast’] landing page
- direct link to mp3 file

§ PLEASE ANSWER THE FOLLOWING QUESTIONS AFTER LISTENING TO THE PODCAST:

1. List 3 things that you learned from this podcast?
2. What is your reaction to the podcast? Pick at least one point Sanyam brought up in the interview that you agree with and list your reason why.
3. After listening to the podcast, do you think you are more interested or less interested in a career in Data Science?

(70%) Explore Python for data munging and analysis, with an introduction to CSV and Pandas

Python’s strengths shine when tasked with data munging and analysis. As we will learn throughout the course, there are a number of excellent data sources for open data of all kinds now available for the public. These open data sources are heralding the new era of transparency from all levels from small municipal data to big government data, from transportation, to science, to education.

To warm up to such datasets, we will be working with an interesting dataset from the US Fish and Wildlife Service (FWS). This is a water quality data set taken from a managed national refuge in Virginia called Back Bay National Wildlife Refuge, which was established in 1938. As a function of being managed by the FWS, water quality samples are taken regularly from the marshes within the refuge.

You can (and should) learn a little more about Back Bay from this link, since it has an interesting history, features and wildlife.

- https://www.fws.gov/refuge/Back_Bay/about.html

The data we will be looking at can be found as a direct download from data.gov, the US data repository where many datasets from a variety of sources can be found – mostly related to the multitude of US government agencies.

The dataset is a small water quality dataset with several decades of water quality data from Back Bay. We will be warming up to this dataset with a basic investigation into the shape, content and context of the data contained therein.

In this part of the assignment, we will make use of Python libraries to pull the data from the endpoint and use Pandas to plot the data. The raw CSV data is readily imported into Pandas from the following URL:

- FWS Water Quality Data 12/20/2020

Please take a look at the page, on it you will notice a link to the raw CSV file:

- https://ecos.fws.gov/ServCat/DownloadFile/173741?Reference=117348

We are going to explore this dataset to learn a bit more about the water quality characteristics of Bay Bay over the past couple decades or so.

§ WRITE THE CODE IN YOUR NOTEBOOK TO LOAD AND RESHAPE THE COMPLETE CSV WATER QUALITY DATASET:

You will need to perform the following steps:

1. **use pandas.read_csv() method to load the dataset** into a Pandas DataFrame;
2. **clean the data so that the range of years is restricted to the 20 year period from 1999 to 2018**
3. **store the entire dataset back into a new CSV** file called back_bay_1998-2018_clean.csv.

HINTS: *Here are some a code hints you might like to study and use to craft a solution:*

- study pandas.DataFrame.query()] to learn how to filter and query year ranges
- study pandas.DataFrame.groupby() to understand how to group data

§ USE PANDAS TO LOAD THE CSV DATA TO A DATAFRAME AND ANSWER THE FOLLOWING QUESTIONS:

1. How many and what are the names of the columns in this dataset?
2. What is the mean Dissolved Oxygen (mg/L) over the entire dataset?
3. Which year were the highest number of AirTemp (C) data points collected?
4. Which year were the least number of AirTemp (C) data points collected?

To answer these questions, you'll need to dive further into Pandas, which is the standard tool in the Python data science stack for loading, manipulating, transforming, analyzing and preparing data as input to other tools such as Numpy (http://www.numpy.org/), SciKitLearn (http://scikit-learn.org/stable/index.html), NLTK (http://www.nltk.org/) and others.

For this assignment, you will only need to learn how to load and select data using Pandas.

- **LOADING DATA** The core data structure in Pandas is the DataFrame. You will need to visit the Pandas documentation (https://pandas.pydata.org/pandas-docs/stable/reference/) to learn more about the library, but to help you along with a hint, read the documentation on the pandas.read_csv() method.

- **SELECTING DATA** The tutorial here on indexing and selecting should be of great use in understanding how to index and select subsets of the data to answer the questions.

- **GROUPING DATA** You may use DataFrame.value_counts() or DataFrame.groupby() to group the data you need for these questons. You will also find DataFrame.groupby()](https://pandas.pydata.org/pandas-docs/s and [DataFrame.describe()‘ very useful.

CODE HINTS

Here is example code that should give you clues about the structure of your code for this part.

```
import pandas as pd

df = pd.read_csv('your_json_file.csv')

# code for question 1 ... and so on
```

§ EXPLORING WATER SALINITY IN THE DATA

The Back Bay refuge is on the eastern coast of Virginia and to the east is the Atlantic Ocean. Salinity is a measure of the salt concentration of water, and you can learn a little more about salinity in water here.

You will notice that there is a Site_Id variable in the data, which we will find refers to the five sampling locations (see the documentation here) of (1) the Bay, (2) D-Pool (fishing pond), (3) C-Pool, (4) B-Pool and (5) A-Pool.

The ppt in Salinity is the percent salinity, and so 1 ppt is equivalent to 10000 ppm salinity. Use this information to answer the following questions.

1. Which sampling location has the highest mean ppt? What is the equivalent ppm?
2. When looking at the mean ppt, which location would you infer is furthest from the influence of ocean water inflows? (Assume that higher salinity correlates to closer proximity to the ocean.)
3. Dig a little deeper into #2, and write why there may be some uncertainty in your answer? (hint: certainty is improved by consistency in data)
4. Use the data to determine the correlation between Salinity (ppt) and pH (standard units). Use the DataFrame.corr(). You just need to report the correlation value.