DATA203 Foundational Python (Prof. Maull) / Fall 2025 / HW3

Points Possible	Due Date	Time Commitment (estimated)
25	Thursday, November 13 @ midnight	*up to* 15 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

- Explore a digital humanities dataset in Python and Pandas.
- Develop statistical summaries of the data.
- Build a Seaborn heatmap for the monthly publication counts for all years.

WHAT TO TURN IN

You will enjoy the highest benefits of the starter notebook if you clone the HW Github repository from your Jupyter Hub terminal with the command:

```
git clone https://github.com/kmhuads/f25_data203.git
```

This will ensure you have the most updated files and starter notebook.

Once you have cloned the repository, you can edit the starter notebook with your solution code.

When you are done with your work, it will be best to zip your `hw2` folder and all sub-folders with the terminal command (one level outside your notebook folder):

```
zip -r data203_hw3_maull.zip ./hw3
```

This will produce the file with all necessary supporting files (notebooks, data output, etc.) then download it from the Jupyter Hub to your local machine, then upload the `.zip` to Teams.

If are confused on how to do this, please ask, or visit one of the many tutorials 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

(30%) Explore a digital humanities dataset in Python and Pandas.

As a data scientist, we must remain curious.

In lecture, we learned about the importance of expertise and the relationships between math/statistics and coding/"hacking" (see the Data Science venn diagram in slides).

Working with any dataset, large or small, and becoming curious about the contents and questions that can be asked about *any* data is a skill that will cultivate as it serves your future growth.

We are going to do just that in this assignment.

It is always suprising to learn about the variety of data that is openly available online.

One such dataset is a curated digital anthology of African American Poetry. *African American Poetry: A Digital Anthology (1870-1928)* is such a dataset. Curated by Dr. Amardeep Singh, an English scholar at Lehigh University, it contains an open-access

dataset of poetry that is now out of copyright – at a time when the *digital* tools available to explore it are beyond imaginable just 50 years ago.

The interactive website for this data is here:

- **website →** African American Poetry: A Digital Anthology by Amardeep Singh

Full citation:

> Amardeep Singh, Ed. African American Poetry: A Digital Anthology. Lehigh University, 2024. https://scalar.lehigh.edu/african-american-poetry-a-digital-anthology/index.

The site includes interesting information about the project rationale, the contents of the data and even some research analyses that might be performed.

§ Task: 1.1 Learn about the African American Poetry: A Digital Anthology website and research project.

1. Browse the site and summarize the contents in your own words. please keep your summary to under 250 words (in other words, *be brief* about what you learned while browsing).
2. In late 1944 Dorothy Porter, a librarian from Howard University, extended and revised a 1916 compilation of books of poetry by African American authors. Her preface and checklist titled *North American Negro Poets: A Bibliographical Checklist 1760-1944* can be seen here:

   - Dorothy Porter, "North American Negro Poets: A Bibliographical Checklist of Their Writings, 1760-1944" (1945)

   Read the preface written by Ms. Porter and respond to her statement:

   > It must be mentioned that one of the principal objectives of this bibliography is to afford an index to the relative distribution of books and other published materials of Negro poetry among our libraries and thus indirectly to reflect the richness of American holdings in this sphere.

   by answering the question:

   > *Consider the tools available to us in 2025. How does this new Digital Anthology help accomplish the original goals set forth in 1944?*

3. Explore the research questions presented by the Anthology here (in the number section 2, there is a bulleted list of questions):

   - Dr. Amardeep's preliminary research questions.

   Answer the question:

   > *Which of these research questions interests you and why?*

§ Task: 1.2 Load and explore the primary dataset from the Digital Anthology.

You will find the primary dataset for the Anthology here:

- African American Periodical Poetry (1900-1928) at Responsible Datasets

Use Pandas to perform the following in your Jupyter notebook:

1. Load the dataset from the raw data URL on the page using only `pandas.read_csv()`.
2. Give the number of rows and columns in the dataset.
3. Learn how to use `DataFrame.value_counts()` and report the counts of the number of works from 1900-1928.
4. Use `value_counts()` to report the number of `Magazine Type` publications which were `Predom. Black`.
5. Provide the percent of total `Magazine Type` which are `Predom. Black` (you'll need to sum all `Magazine Type` and use that to find the ratio or percent).

(50%) Develop statistical summaries of the data.

We will extend the work we started in the first part and continue to explore the data.

§ Task: 2.1 Provide a table which shows the percent of publications by year.

- You will need to get all the counts of publications by year (use the `year` column) and divide all counts by the total number of publications.

§ Task: 2.2 Use the provided function `get_monthly_pub_counts()` to create the Series for 1904, 1905 and 1906.

- You can just run `get_monthly_pub_counts()` on the three years in separate cells.

§ Task: 2.3 Continue using `get_monthly_pub_counts()` to produce a new DataFrame with counts for ALL years, 1900-1928.

The new DataFrame will have the year as the index and the ordered months as the columns. Each value in the DataFrame will include the counts for that year and month.

Your final DataFrame will look something like:

	January	February	…	November	December
1900	0	0	…	1	2
…	…	…	…	…	…
1929	2	5	…	3	9

(20%) Build a Seaborn heatmap for the monthly publication counts for all years.

We learned in lecture that visualizing data makes everything better – most of us and our peers do not want to see thousands of data points in a table, they want to have a picture of that data.

We are going to use a library called Seaborn to make a heatmap of the data table you produced in the previous section.

§ Task: 3.1 Build a heatmap of all the publication counts from 1900-1928.

Study the documentation for Seaborn heat maps here:

- `seaborn.heatmap`

Don't over think this, you will not need to do much work to make the heatmap. If you have more than 2 lines of code, you are on the wrong track. You can do it in 1 line of code once you have the complete DataFrame from the previous part.

§ Task: 3.2 Please look at the heatmap and make 3 observations about what you see.

You may include observations about the months and years which are most frequent, trends, data gaps, etc.