PROCEEDINGS OF SCIENCE

PoS(ISGC2021)018

Piloting Data Science Learning Platforms through the Development of Cloud-based interactive Digital Computational Notebooks

Rajesh Kumar Gnanasekaran[a,*] **and Richard Marciano**[b]

[a,b] *The University of Maryland,*
4130 Campus Dr 4th floor, College Park, MD 20742, USA

E-mail: rgnanase@umd.edu, marciano@umd.edu

*Speaker

Physical learning, communication, and collaboration have taken a colossal hit in 2019 and 2020 with cascading lockdowns resulting from the spread of the COVID-19 pandemic. Virtual access has become the need-of-the-hour, and the uses of cloud-based course content delivery, distance learning, and document collaboration are becoming increasingly ubiquitous. This paper introduces a novel method to allow students and faculty in the Humanities, Arts, and Social Sciences (HASS) to collaborate and interact through data analytical technologies using "interactive Digital Computational Notebooks" (iDCNs). We demonstrate this approach using a digitized Legacy of Slavery (LoS) archival dataset collection from the Maryland State Archives (MSA) and illustrate the socio-technical challenges in establishing this learning environment. We provide a step-by-step process involved in accessing, developing, and integrating different infrastructure elements. The LoS [1] in Maryland is a major initiative of the MSA. The program seeks to preserve and promote the vast universe of experiences that have shaped the lives of Maryland's African American population. Over the last 18 years, some 420,000 individuals have been identified, and data assembled into 16 major databases. These databases contain information unique to enslaved people's lives, such as manumission records, certificates of freedom, census data, penitentiary records, etc. One of this paper's primary objectives is to enable the digital representation of these culturally rich and sensitive collections ready to be analyzed and studied through contemporary scholars' lenses. This project aims to achieve this goal by making these databases available and accessible so that users can generate individual stories, glean insights, and possibly recover "erased" memories of enslaved people. To achieve this goal, as a first step, unique dataset collections were prepared by downloading the databases and put through rigorous exploration, cleaning, and visualization process through coordination with interdisciplinary scholars composed of archivists, historians, computer scientists, and technology analysts. This project also illustrates the importance of a multidisciplinary approach to a unique set of digitized archival data with a specific focus on contextual aspects due to the data's historical value and sensitivity. The collaborative process used open-source and readily accessible tools to create meaningful visualizations as an arrangement that flows together conducive for educators to teach. The visualizations use the spatial and temporal characteristics of the datasets to produce graphs and charts for a graphical view of the datasets. The visualizations constructed are responsive to present the data by instant connections to the datasets dynamically. The integration of these digital artifacts obtained from each dataset was carried out through Jupyter Notebooks (JNs). These iDCNs are unlike the traditional *digital notebooks* that provide a space for students to take notes and collect clippings of text. Instead, the iDCNs developed in this project are a novel set of educational tools that allow text and software code to co-exist and be rendered in a single document coherently for instructors and students to follow the text with visual representations back-to-back. The iDCNs are also equipped with live examples of basic natural language processing on certain text-rich features of these dataset collections. The open-source nature of this project's setup and cloud-based distribution of these digital artifacts pave the way for students from under served communities to take advantage of a unique way of learning and to perform hands-on work on marketable software tools, preparing them for a successful career. The contributions of this paper to the fields of HASS and other non-STEM (Science, Technology, Engineering, and Mathematics) backgrounds lie in the idea of providing an "always-on" cloud-based pedagogical environment for aspiring students and researchers worldwide to analyze, learn and unearth stories through data science driven approach on a cultural dataset, in our case, the LoS dataset collection.

Keywords: *Computational Thinking, Interactive Digital Computational Notebooks, Computational Archival Science, Cloud-based digital learning*

International Symposium on Grids & Clouds 2021, ISGC2021 22-26 March 2021
Academia Sinica, Taipei, Taiwan (online)

1. Introduction

Sharing digital content online across the globe has become a necessity that has been well realized in the past year due to COVID-19. In [2], the authors indicate that the pandemic exposed the inadequate and inefficient digital infrastructure and setup at thousands of US institutions which struggled to adopt a fully online learning environment. Even before the pandemic, for the past two decades, the push from on-campus learning to online learning was taking place but not widely adopted by all institutions. In [2], Gallagher and Palmer point out that higher education has been lagging behind other industries in moving to a more digitally-driven business model. The authors identify the launch of a fully online Master's degree in Computer Science by Georgia Tech that has exceeded 10,000 enrollments in a semester in 2020 as an example. This moment in history is considered as an inflection point which measures higher education industry's ability to survive in the long term. In [3], the authors strongly suggest that higher education institutions work with the faculty and staff to promote creating quality digital content based on the needs of their students and the educational innovations. They consider that moving to digital-based learning is now an inevitable risk mitigation factor which the pandemic has clearly shown the need for, whereas just a year ago, digital transformation was primarily seen as achieving greater access, global reach, and personalized instruction.

In [4], Dunn points out the tremendous amount of change that the post-pandemic period will bring to teaching in these fields. Dunn describes that making digital content alone is not entirely sufficient to the students from these niche areas, instead there is an increasing need to transform these students themselves as digitally-aware citizens, and create digital ways of being in the pedagogical world that they know of. Dunn suggests that it is imperative that these students be taught digital methods for creating new content, thereby getting to know the "how" rather than "what". Knowledge on how the content is developed and the many other intricacies involved with the dissemination of information should be a necessary topic of discourse. In [5], the authors stress on the importance of the disruption taken place in the Archival world with the emergence of groundbreaking computing technologies that are now used in creation, sharing and storage of artifacts, thus making it even more necessary for aspiring archivists to learn and be equipped with the understanding of these new technologies. In [6], Underwood et al., argued that in the growing native digital space, there is an urgency for digital archivists to develop a stronger understanding of the characteristics and limitations of technology. In addition, Taylor et al., in [7] points out the importance of the changing dynamics of physical libraries where library resources are rapidly moving online and the increased interest from patrons in accessing the online content for the libraries. It is vital for the aspiring librarians to understand the computational terms to better meet the needs of patrons and address any issues or concerns that would arise. Thus the need to create a course that is not only available online but also educates students from the HASS, Cultural, Library, and Archival higher education degree programs to understand and perform computational operations on the digital content they would be managing and supporting in the future.

Our study's aim is to develop and contribute a novel set of digital educational course modules which would be available "always", made possible by cloud-computing technology and uses open-source content which would be published free for anybody with internet access. The course modules

are created using an open-source cloud-based web application tool called *Jupyter Notebooks* (JNs),[1] with the computational programming language *Python.*[2] These modules are named as "interactive Digital Computational Notebooks" (iDCNs) to differentiate from the traditionally used “digital notebooks” which students and professionals use to take notes, clip texts and keep them organized. These iDCNs get their name also because they are tailored to teach computer programming through a step-by-step data science driven analytical treatment of a dataset collection for analysis and visualization purposes. In addition, such a case study has not been conducted yet on a culturally rich dataset collection like the Legacy of Slavery (LoS) from the Maryland State Archives (MSA) in the USA. These iDCNs would form a new learning environment which would become part of courses taught for HASS and non-STEM higher education students across the country. To evaluate the acceptance of this novel educational tool among them, after the iDCN were created and tested, a user survey was distributed to current students and educators from these programs and the responses were studied and understood for future improvements to this unique learning tool. With these tasks ahead, the paper is divided into the following sections to address them: Related Work, Research Problem and Questions, Approach, Results, Discussion, Future Work and Conclusion.

2. Related Work

A review of the literature identifies several research studies that use JNs as a teaching tool to teach programming in higher education. In [8], Davies et al., created a set of JNs for bioscience and informatics education. The paper documents in detail the anatomy of the JNs with explanations and pictorial representations on how to set up the infrastructure and the environment for performing basic computational commands using them. The authors convey the benefits of using JNs as user-friendly collaboration tool, encouraging reproducibility, allowing sharing of the modules and how they could also be used as an assessment tool by the educators. The study also performs case studies to teach basic programming and uses coding commands in tutorial lessons, however, modified to suit the domain of bioscience education. Unlike our project, the study does not seem to conduct a data-driven analysis of a real dataset. It does, however, publish a survey of experiences gathered from the students as part of a 6-week course work in teaching them fundamentals of programming using JNs.

In another study [9], Reades conducts an evaluative research on how JNs performed in the teaching of programming to undergraduate students from geography, another STEM background subject. This study explains the creation of three JN modules calling them ‘geocomputational’ modules and evaluated the JNs based on factors such as minimal complexity, maximal flexibility, interactivity, utility, and maintainability. The authors published a detailed comparison of pros and cons of using JNs for teaching programming to their undergraduate students.

A similar study was conducted by [10] to prepare teaching materials for electronics undergraduates through JNs. The authors created modules to teach the subject Digital Signal Processing to the students using these modules. The authors used *Python* programming language to teach these STEM background students. The paper talks about the growth of JNs as an increasingly accepted

[1]Jupyter Notebook - https://jupyter-notebook.readthedocs.io/en/stable/notebook.html.
[2]Python Documentation - https://www.Python.org/doc/.

educational tool and how the supporting software libraries available for *Python* language makes it an attractive option for students and educators alike to incline towards these open-source resources.

There is one notable archival study [11] which uses JNs in the Archives domain, a non-STEM background. The authors highlight the ability of JNs to "see exactly the lines of code that have produced the data, aid transparency and supply the necessary contextual information for meaningful browsing." They discuss the development of JNs in combination with another software to enable archivists to explore, investigate and analyse the metadata in the Netherlands Institute for Sound and Vision (NISV) archive. The challenges faced by the authors in the development of these JNs are discussed. Our study differs from this work in how it introduces JNs to non-STEM students and its step-by-step data science educational focus. The iDCNs created by our project not only use the powerful infrastructure already offered by the JN project, but also follows a set of proven practices in performing a contextual data science driven approach on a culturally rich dataset collection.

In prior collaborative work [12], we conducted preliminary data science driven approaches to the MSA's dataset collections through a detailed study on how to derive "data background" from the contextual analysis of data with collaborations with experts from different disciplines. We used several tools, some of which were open source like *Open Refine,*[3] *R* [4] for data exploration and manipulation, and other industrial tools like *Tableau,*[5] and *Neo4j* [6] for data visualization purposes. We documented detailed results from these analyses of MSA datasets which also included the cross collection study on trying to link data between the two datasets. Our present study deals with a data-driven approach of one of these MSA datasets but in a different infrastructure. The *JupyterLab* [7] environment powered with *Python* programming language's support for software libraries allows us to perform most if not all of these operations in a homogenous environment unlike the prior study where different tools were used for different data operations. This was one of the important reasons for choosing JNs for the development of iDCNs as the idea of switching between different tools could become quickly distracting for beginners of programming from non-STEM backgrounds.

In [13], Wing coined the term "Computational Thinking" to stress the importance of the growth of computer science principles in every educational field and the importance for every child to nurture CT. Weintrop et al., in [14] developed on the views of [13] and presented their framework for applying CT practices to mathematics and science education. The four practices identified in the study are: data, modeling & simulation, computational problem solving, and systems thinking. The authors formulated these practices from detailed experimentation and external reviews. As mathematics and science education has been increasingly with the push of computing, these practices were created to enhance instruction and learning [14]. The authors presented their views and discussed a framework for applying Computational Thinking (CT) practices to mathematics and science problems in K - 12 schools. The practices aligned well with the objective of our own project and we have attempted to emulate them, as well as document how they applied to the development of iDCNs. To make the process of development of these iDCNs easy for us and to keep them logically arranged, there was a need to follow a set of well-established best practices.

[3]Open Refine - https://openrefine.org/documentation.html.
[4]R - https://www.r-project.org/other-docs.html.
[5]Tableau Documentation - https://help.tableau.com/current/pro/desktop/en-us/gettingstarted_overview.htm.
[6]Neo4j Documentation - https://neo4j.com/docs/.
[7]JupyterLab Documentation - https://jupyterlab.readthedocs.io/en/stable/.

3. Research Problems and Questions

In [5], Goudarouli indicates that advances found in emerging technologies, document creation, generation, storage, sharing and availability, have all but changed the nature of archival processing. These changes are believed to be widely felt across HASS, archives and libraries backgrounds. There is increasing demand for educators and students to stay abreast in learning and teaching these new technologies. In addition, it would be vital to understand the computational concepts behind the generation and dissemination of these digital artefacts. To equip themselves with this skill and to address the increasing demand for distance learning, course content and lesson plans appropriate to suffice these requirements need to be created. To solve these problems in our project, we formulated the following research questions:

- RQ1: Can a novel set of cloud-based iDCNs be developed with content that is used as a learning tool to teach computing for higher education students from non computational backgrounds?
- RQ2: Can the MSA's digitized CoF dataset collection be used to create the computational content as a case study through a step-by-step contextual data science driven analytical treatment?
- RQ3: Can the CT practices introduced in [14] by Weintrop et al., be extended and applied to non-STEM backgrounds in creating the computational content?
- RQ4: Can the resulting novel learning environment be useful for Library and Archival Science educators and students? How would they react to such changes to their pedagogical learning environment?

4. Approach

To address the research questions, we divided the project into two main parts: (1) an exploratory case study leading to the creation of a set of iDCNs using an open source tool, and JNs following CT practices from Weintrop's Taxonomy [14], and (2) a user survey to gather feedback from the end users of this educational tool (students and educators with Library and Archival Science backgrounds).

4.1 Creation of iDCNs - An exploratory case study

To address RQ1 and RQ2, we used one of the datasets from the MSA, (the CoF dataset), in an exploratory case study to build the iDCNs that teach step-by-step contextual data science driven analytical treatments. A Certificate of Freedom (CoF) [15] is a legal document issued from 1803 to 1865 in the state of Maryland to African American Enslaved persons, who were required to record proof of their free or emancipated status in the county court. The certificate was issued based on information documented and provided by the former slaveholder and a witness. The court or registar of wills would also use manumitting documents for verification. The CoFs were handwritten documents containing general biographic, demographic, and descriptive information about the enslaved person. In most cases, the data captured or documented for the CoF by the courts or court clerk followed a set of standards for documents across different counties which consisted of data features like county issuing the document, slave owner's first and last name, enslaved

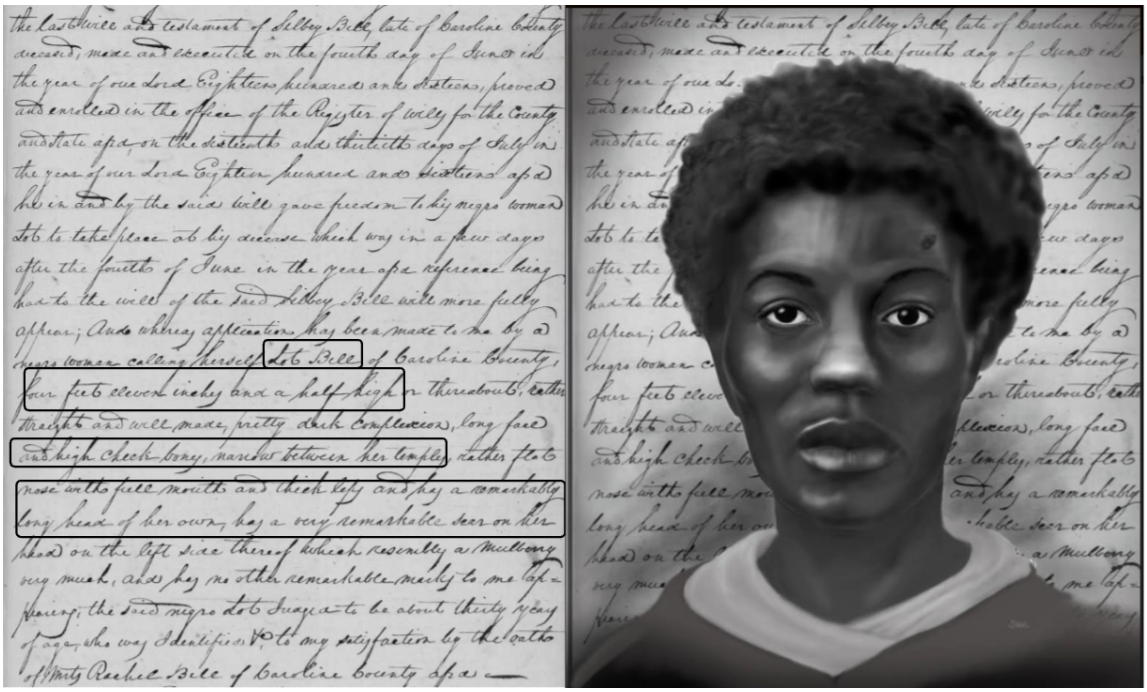


Figure 1: To the left, the 1816 CoF recorded for Lot Bell from the MSA scanned document; To the right, a forensic detective's rendition of the text data into a digital image. Image courtesy - Christopher Haley Director of the Study of the Legacy of Slavery in Maryland program, at the Maryland State Archives.

person's first and last name, gender, age, height, complexion, scars (for identification purposes), alias name, date of issue, witness name, prior status of the enslaved person, and a notes feature to enter comments/remarks. There are 23,655 records in the CoF dataset from the LoS collection. The dates of issuance range from 1806 to 1864, covering sixteen Maryland counties and one Maryland city.

This idea to use the CoF dataset collection was inspired from the long-standing efforts by MSA staff on several projects, including the Faces of Freedom Project [16]. MSA's Projects, in addition to digitizing the physical documents like CoF, have gone as far as turning data into people through stories and documenting African American enslaved people's lives and experiences as described in [17]. In the Faces of Freedom Project [16], the MSA approached a criminal forensic detective to use the text description identified from the CoF dataset collection as shown in Figure 1, to create a visual image of an enslaved, later freed, person named Lot Bell. This unique rendition of textual content by transformation into a digital image encouraged us to use this dataset. The iDCN modules were developed using a cloud-based open source web application, JNs as the digital infrastructure. *Python* programming language was leveraged for all the computational coding commands in the notebooks. Keeping this in mind, the coding commands in the notebook modules were written with care with basic to advanced commands gradually increasing in complexity to not overwhelm students being introduced to this new environment.

Papadakis et al., in [18] used a similar exploratory case study research method to perform research on the effect of using ScratchJr to teach basic programming concepts to pre-kindergarten students at a school in Greece. The exploratory case study method was well suited for this part

also because it allows us to focus on each of the historically and culturally rich dataset features or columns available from this specific dataset. As these data features hold tremendous historical value and possess human sensitivity, at every step possible, we worked closely with historical subject-matter experts and historical resources before manipulating the data.

4.1.1 Why the *Jupyter Notebook* (JN)?

Jupyter Notebook (JN) is an open source web application, created in 2014, which enables software developers and researchers to design, create, collaborate and publish their creative work in the form of narrative text like stories coexisting with computational coding snippets that perform specific tasks along with pictures and other media files. This web application is an outcome of *Project Jupyter,*[8] which is a non-profit open-source project. In addition to the central product, the Project community also has developed and implemented other products namely *JupyterLab*, which is an integrated development environment. The *JupyterLab* provides a user interface that allows developers to access resources like datasets, files, and use them across multiple digital JNs created in the environment. We used *JupyterLab* in this project to develop these notebook modules. Fully developed and tested JNs could be implemented as narrative HTML markdown files. In [8], Davies et al., have explained the anatomy of the JNs and they provide a number of training resources [19] for beginners to advanced users. An important feature of JNs is the ability to use them as a live editor to modify elements like Text and use different Markdown styling options to make the content appear as paragraphs with headings, indentations, bulleting and other text formatting. In [20], Smith details the usage of Markdown in development of JNs.

In [19], the authors discuss the benefits of using JNs in teaching and learning pedagogical processes. The authors indicate that by incorporating JNs, the course benefits from increased participation, engagement and real world understanding of the subject matter. They also point out that students could benefit from CT, active learning from wherever they are and whatever time they want to as the resources are free, open source and cloud-based. Instructors and educators could leverage these unique digital resources in several ways notably as a learning material, as demonstrative lectures, to perform online laboratory exercises, and also to conduct exit ticket quizzes. The authors also point out that the JNs are becoming popular not only STEM areas but also in Digital Humanities, Social Sciences, Writing, Music and Introduction to Programming courses. JNs allow developers to include coding commands from many computing programming languages like *Python*, *R*, however, we have used *Python* as it had supporting libraries for running a full suite of data analytical methods like exploration, manipulation, analyzing and visualization of data. In particular, it provides libraries that allow developers to create interactive visualizations for creating networks from data, to georeference areas by counties in the USA, to create word clouds from textual data, and to plot interactive charts and graphs on the cleaned data.

4.2 Following CT Framework

To answer RQ3, the approach is to adhere to the CT practices proposed by Weintrop et al., as shown in Figure 2, during the data driven analysis of the CoF dataset using these new iDCN learning modules and capture the list of practices that were followed for this specific project and

[8]Project Jupyter - https://jupyter.org/

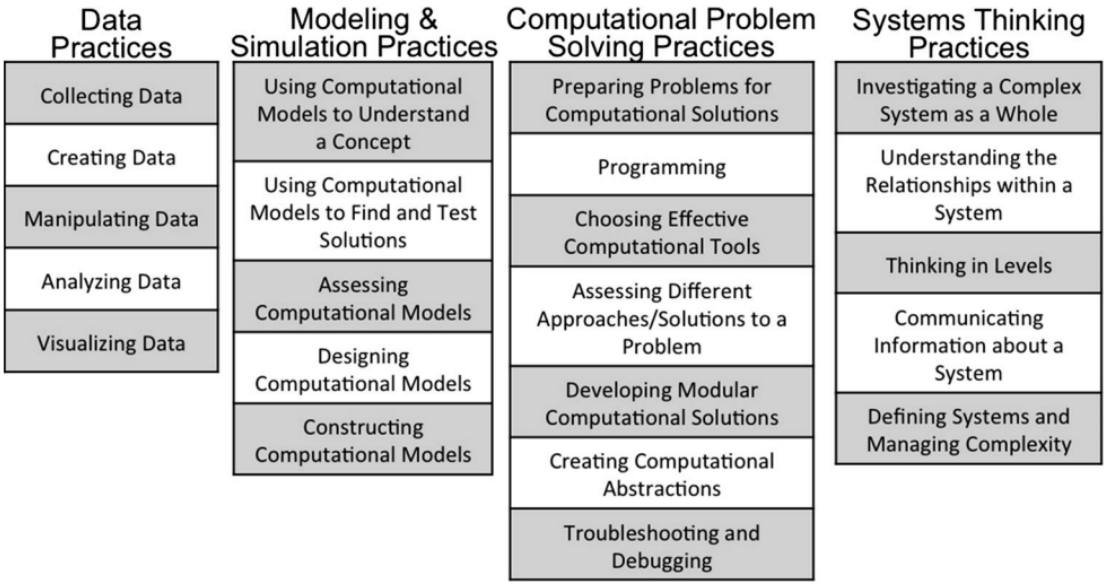


Figure 2: Computational Thinking (CT) Practices Taxonomy [14]

justification on how the steps relate to the practices, so the results could be extrapolated for further adoption of these CT practices in other areas to solve complex problems.

4.3 User Survey

To address RQ4, we performed a User Survey with graduate students and educators collaborating in a non-STEM Library and Information studies course at two higher education institutions in the USA. In [21] Sun and Oza proved that using a User Survey research method was effective in gauging the understanding and issues in using a collaborative online system among 260 users of the system. As these new cloud-based iDCNs would also be collaborative in nature and accessed online, we decided to use this method to gather feedback on the tool. The results are documented under the User Survey Results section.

5. Results

From the two main parts of the project, we had two outcomes: the iDCN learning modules and the user survey results.

5.1 Development of JN Learning Modules

We arranged the new iDCNs into four learning modules as follows:

- *Index Notebook*
- *Contextual Data Analysis and Manipulation - Part 1 Module*
- *Contextual Data Analysis and Manipulation - Part 2 Module*
- *Contextual Data Visualization Module*

We used *JupyterLab* as the development environment which let us create individual folders for the media files, datasets, notebooks as shown in Appendix Figure. 7. We used *GitHub,* [9] a cloud-based version control tool, and *JupyterLab* allowed us to update the files from the repository on the developer's PC directly which were later merged with the Master branch of the repository. Once merged with the master branch, the changes are instantly available and picked up at *https://cases.umd.edu*[10] which hosts a public cloud-based infrastructure for demonstration, sharing, and collaboration of these learning modules. Students and educators can also use the option to edit the notebooks for laboratory exercises using the Binder [11] software which renders an individualized sandbox environment as an open-source option to add or modify and run the individual elements from the iDCNs. Appendix Figure 6 shows a sample published iDCN module.

5.1.1 Index Notebook

As a starting point to this novel learning environment, the index notebook module was designed in such a way that it conveys two things. One through narrative text, it introduces the Legacy of Slavery Project at MSA, whose CoF dataset would be utilized for the exploratory data science driven case study with computational treatments. Another topic is to show the various formatting options available on JN using Markdown text editing such as Title, first, second and third level headings, creating a paragraph structure with bullet points, highlighting, bolding and italicizing abilities, options to provide hyperlinks to external resources, attaching images in line with the text, and other basic text editing features. Appendix Figure 8 shows an image of the index module's screenshot. One could choose to click the next module or choose a desired link from the available modules as shown in Appendix Figure 8. The theoretical text and pictorial representation of this module was designed keeping in mind to not overwhelm non-STEM users with computational coding commands up-front, but to provide a preface to the upcoming modules.

5.1.2 Contextual Data Analysis and Manipulation - Part 1 Module

Continuing from the index module, we created part 1 of the 2 module series. According to the CT framework, *data practices* include data creation, collection, manipulation, analysis and visualization. As the CoF dataset was digitally created from the physical scanned documents and data was collected by the MSA in their LoS database, the first two components of the *data practices* were already satisfied prior to this project. Using traditional database access tools, the MSA's LoS database was accessed to extract the CoF dataset in a commonly used comma-delimited file format called *CSV.*[12] From [14], data manipulation activities include sorting, filtering, cleaning, normalizing datasets, and data analysis pertains to looking for patterns or anomalies, defining rules to categorize data, identifying trends and correlations. We attempted to perform these data operations on the CoF dataset using *Python* programming language. Also the activities involved with these two practices appear to be closely related to each other as manipulation operations result in analysis which could lead to further manipulation and analysis. Hence, we decided to combine

[9] GitHub Documentation - https://docs.github.com/en

[10] Legacy Of Slavery iDCN Project - https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb.

[11] Binder Project Documentation - https://mybinder.readthedocs.io/en/latest/.

[12] CSV reading in Python - https://docs.python.org/3/library/csv.html.

them for the purpose of this project. By mastering these two practices, Weintrop claims that HASS students could be able to work with “big data", thereby equipping them to bring any large datasets into a desired set up or configuration and help them make confident claims on the analysis they make.

As discussed in the Introduction section, the CoF dataset has a number of culturally rich features. We chose to operate on four of these: CoF Issue Date, Enslaved person’s Prior Status before issuance of the CoF, Enslaved person’s Height, and Age. This decision was made for data manipulation and analysis purposes due to the categorical and quantitative nature of the features. As coding commands start to appear from this step, human interpretable comments were added to these commands to help the students understand the high level overview and functionality of them. To keep the data operations demonstrable and modular, the notebook was split into two parts. Part 1 processed CoF Issue Date, Enslaved person’s Prior Status features and Part 2 took care of Enslaved person’s Height, and Age features.

Appendix Figure 10 shows a screenshot image of the Part 1 module notebook which shows code commands to import software libraries accessed by the *Python* programming language to create dynamic data structures that would be used further for data operation purposes. The module begins with importing the CoF dataset which was uploaded to the project’s Dataset folder as shown in Appendix Figure 10. The dataset is then stored in *Python* programming memory into data structures called *‘data frames’* which are similar to database tables. Further commands are run to display the first 10 rows of the data frame stored, then the focus shifts to the CoF Issue Date feature. A number of data manipulations are run to first filter the unique formats and values, then to clean the date values into a standard format and substitute incorrect date values to a code, ‘NaT’, which could be parsed in the next step to isolate the erroneous records for further analysis or to perform corrections at the source data by sharing it with the MSA. Upon analysing the resulting values of error records, we were able to point out interesting patterns and anomalies with the data collected by the MSA for this date feature. In Appendix Figure 9, of the 657 erroneous date records, 2 of them which had a length of 6 digits were filtered for further analysis. This resulted in the display of two rows from the dataset as shown in Appendix Figure 9. Using this information, MSA’s database was searched to find the scanned CoF for the enslaved person named Jeremiah W. Brown, which showed that, on the physical recorded CoF, the day was not legibly visible as highlighted with blanks. This is a classic example of data manipulation and analysis steps which could be helpful for HASS students to quickly identify the root cause of problems with the source records. Working with the Enslaved’s Prior Status feature, which is a categorical data, similar interesting patterns were uncovered. There were different formats of data capture like ‘born free’, ‘free born’, ‘Born free’, etc which could be grouped into a single value, however, values like ‘Descendant of a white female woman’ as shown need to be fixed with the help of historical subject matter experts since it possesses contextual information. To resolve this problem, we consulted historians and they provided valuable insights who indicated that the person should be considered ‘Free’ by virtue of being born to a white female woman. This shows the importance of collaborating with the historians in performing certain key data manipulation and analysis tasks to the dataset collections pertaining to the field. The module follows with further code commands to transform and group the different categories into five of them. The ones which have missing values and unknown categories were shared with the MSA for correction at the source of record. After these data operations were completed, the in-memory data

frame was saved to an output *CSV* file which could be used in the next module, Part 2.

5.1.3 Contextual Data Analysis and Manipulation - Part 2 Module

The CoF dataset's Height and Age features were worked on for the Part 2 module whose input was the output dataset saved with new features from Part 1. Similar to the two features from Part 1, the Height feature had some anomalies. As the feature was originally collected as feet and inches for example as 5' 5", we anticipated that there would be transcription errors possible. Upon running descriptive analysis we identified several formatting errors. To run a data manipulation task to correct these errors, we had to write a custom *Python* function that parses each height value and converts all of them into one standard, inches, i.e, 5' 5" would be converted to 65.00 inches. This shows one of the strong abilities of *Python* to create and run custom functions and this step was created with appropriate comments to teach learners on the uses of this function so they could apply it for other problems in the future. Even after conversion to inches, we still found anomalies in the height data as shown in Appendix Figure 11 with a max height of 145 inches which seemed unreasonable. As this feature also required contextual historical data, we consulted subject matter experts who were able to provide their research results surrounding the average height of enslaved people referring us to a study by Margo and Steckel in 1982 [22], who performed an analysis of the height and age from the Enslaved Manifest data of around 50,000+ enslaved people shipped between 1811 and 1861 to ports like Baltimore, Richmond and other cities from the Port of Savannah in USA. According to this study, the average height of enslaved people was around 67 inches. In the same study where another set of Enslaved People's appraisal records showed the maximum height was found to be around 72 inches. This important collaborative insight led us to the path of identifying the outliers above a height of 80 inches and below a height of 5 inches and found that there were 4 entries with invalid values as shown in Appendix Figure 12. On performing a detailed analysis of looking up scanned documents of the records related to these error records, a result appeared as shown in Figure 3 where for the enslaved person Milly Farmer, the date was physically recorded as four feet eleven and three quarter inches however the data was digitally recorded as 4' 44.75" indicating an error. These entries were captured as incorrect value 'NaN' and shared with the MSA for source correction. With the Age feature, the issue was with how an enslaved person's age in months were recorded in year format. For example, in Appendix Figure 13, an 8-month old baby's age was recorded as 0.08 of an year which is an incorrect value. A custom *Python* function with comments was created to manipulate the data to correct age in year format and stored as a new feature. It should be noted that in all the data manipulation tasks where the source data has to be changed, the modified data was stored as a new feature without touching the source data in the feature. This allowed us to compare the original and transformed values as shown in Appendix Figure 13. This module finishes with saving its results into an output CSV file to be used in the next module, the contextual data visualisation.

5.1.4 Contextual Data Visualization Module

Weintrop [14] asserts that applying data visualisation practice to the computational treatments help students to effectively convey stories and insights gleaned from the previous data manipulation and analysis tasks. In this module, the key objective was to create meaningful context-based visualizations that highlight the findings as stories. The iDCN module begins by importing certain

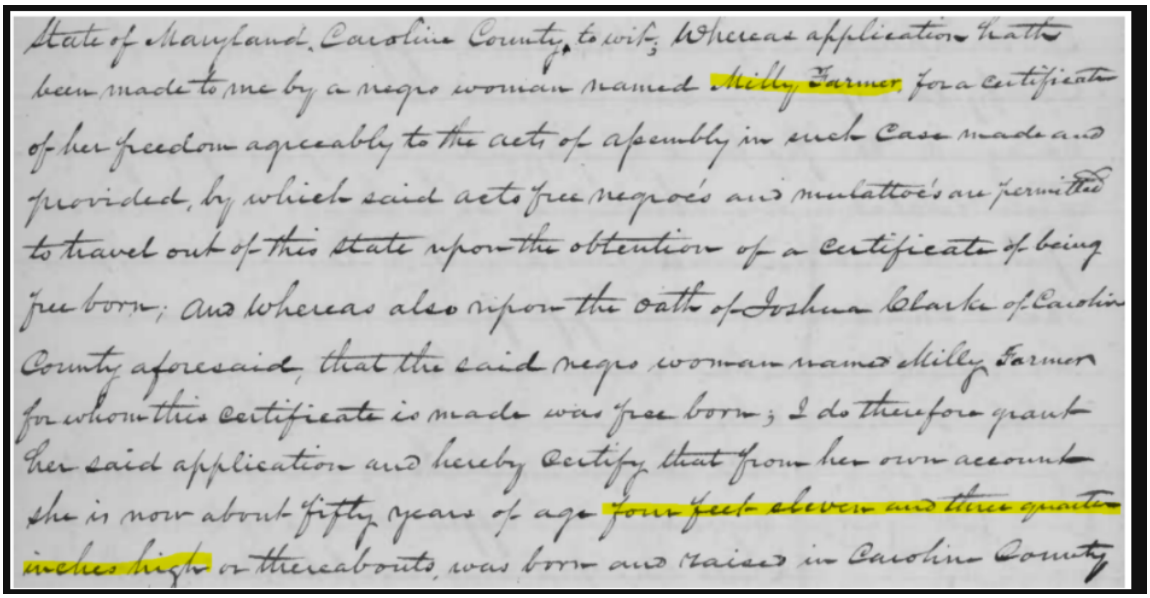

State of Maryland, Caroline County, to wit: Whereas application hath been made to me by a negro woman named Milly Farmer for a certificate of her freedom agreeably to the acts of assembly in such case made and provided, by which said acts free negroes and mulattoes are permitted to travel out of this state upon the obtention of a certificate of being free born; and whereas also upon the oath of Joshua Clarke of Caroline County aforesaid, that the said negro woman named Milly Farmer for whom this certificate is made was free born; I do therefore grant her said application and hereby certify that from her own account she is now about fifty years of age four feet eleven and three quarter inches high or thereabouts, was born and raised in Caroline County

Figure 3: Scanned CoF document of Milly Farmer highlighted with the height recorded at source

unique software libraries like *plotly*,[13] *bokeh*, [14] *networkx*, [15] *wordcloud*, [16] which are necessary to perform the visualisation tasks. As indicated in the Related Work section, the main difference between our prior work on the CoF dataset and this project is to harness the versatility of the *Python* programming language in its ability to support these different software libraries in one tool that adds interactiveness to the visualisations in a way other open source and industrial tools could not provide. With these appropriate libraries, *Python* acts like an all-in-one shop which lets users choose their own and customize their product to cater to their unique needs. All of this with the added benefit of being able to create dynamically changing graphs, charts, geomaps and network graphs. In Appendix Figure 14 and Figure 15, *plotly*'s iplot package was used to create basic histograms of the newly created CoF features Enslaved's Age and Height. It could be seen that most of the enslaved people were between the ages of 20 and 40, and within the height range of 60 - 75 inches which matches with the study by Margo and Steckel [22]. Similarly, *plotly*'s express package was used to create a simple pie chart of the counts of CoF issued for male and female gender. It could be noted that from Appendix Figure. 16 the percentage of CoF's issued to male was at about 93% and for females at about 5%. With this data, a plot between gender and the number of CoF's issued by year showed an interesting historical contextual finding as shown below in Figure. 4. As could be seen, we see a spike in the number of CoF's issued in the year 1832. By applying contextual historical analysis, historians identified that there were two key historical events that are believed to have taken place between 1831 – 1832 according to Maryland's state history as describe in [23] on page 29. Although we cannot make cause-effect relationships between these two factual findings, this kind of insight would not have been possible without the application of data science

[13] Plotly Documentation - https://plotly.com/python/.
[14] Bokeh Documentation - https://docs.bokeh.org/en/latest/index.html
[15] NetworkX Documentation - https://networkx.org/documentation/stable/index.html.
[16] Wordcloud Python - https://pypi.org/project/wordcloud/.

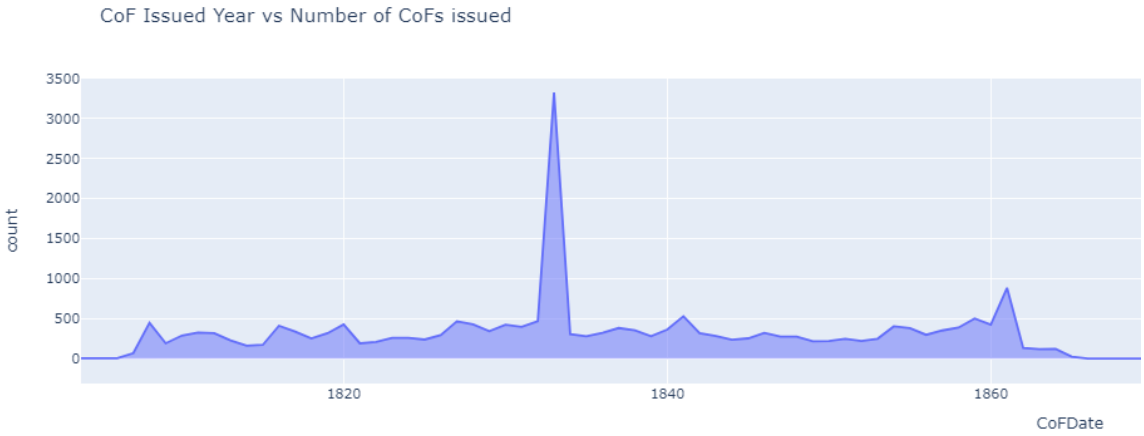


Figure 4: Visualization showing number of CoFs issued by Year from the CoF dataset

techniques followed in this project. These findings would help historians to identify certain patterns and make necessary connections that may have not been possible without a data-driven approach. Using Bokeh and networkx libraries, interactive network visualizations were also created as shown in Figure 17 which shows a network of Enslaved people owned by a Slave Owner indicated to be at the center with last name 'Atwell'. Another unique visualization of geo maps were plotted using the *plotly* express's choropleth map box technique which uses the FIPS code, an unique code given to each county in the USA, to plot the desired features on the map of the USA. We transformed the County feature from the CoF dataset into the list of Maryland state county FIPS codes and used. In Appendix Figure 18, we could see that an interactive map has been plotted that shows the color of the counties based on the weightage of the number of CoF's issued for them. This module also touches on the Natural Language Processing abilities of the wordcloud library supported by *Python*. From the CoF's Notes features, which included the handwritten text by the digital data transcribers at the MSA indicating any issues with the CoF record, we ran a coding command to create a word cloud that results in a graphical representation of the most commonly used words by the size. This type of visualization could be used as a starting point to do further research into the text oriented features of the dataset. These four iDCN modules together form a coherently arranged learning tool that could be harnessed by the educators and students for better understanding of digitized data as appropriate to their fields. In all of these modules, working with historical data needed help from the historians to research on the contextual impact. This is a very important finding from the entire project as cultural and sensitive datasets of this nature should not be put through run-of-the-mill data analytics tools and expect any meaningful results to be derived from those operations.

5.2 CT Practices

Of the 22 CT practices identified from Figure 2, we covered 13 of them in the development of iDCNs in our project as detailed below: This proves that RQ3 was addressed to meet the expectations of adhering to these CT practices and the taxonomy applies to a unique research idea of creating data science driven computational treatment of the MSA's CoF dataset.

- **Three Data Practices:** (1) *Manipulating Data* – Open source web application JN was used

with *Python* programming language to perform sorting, filtering, cleaning and exploring the data; (2) *Analyzing Data practice* – was employed using the same libraries in the notebook modules to apply rules, and categorize CoF features selected for analysis, identify anomalies and outliers, run them through contextual historical perspective, and glean insights from the results; and (3) *Visualizing Data* – New software packages and libraries were used to create meaningful visualizations and uncover interesting insights from the results of the analysis.

- **Seven Computational Problem Solving Practices:** (1) *Preparing Problems for Computational Solutions* – The CoF digitized data was exported as a *CSV* file from the MSA database using MS Spreadsheet and with *Python* programming language, the *CSV* file was imported into the project and individual features were extracted using coding commands. These features were then analyzed individually to identify problems with them and were provided solutions through computational treatments, (2) *Computer Programming* – By using *Python* programming language, individual data features were run through iterative, conditional logic to clean data features. (3) *Choosing effective Computational Tools – Python* programming language's versatility supports software libraries in performing most of the data science driven analytics within the JN ecosystem and not requiring us to look for external resources. (4) *Assessing Different Approaches/Solutions to a Problem* – This practice was realized during the development of the notebook modules where several solutions to a problem were applied before finally arriving at the best solution provided by the different software libraries supported by *Python*. (5) *Developing Modular Computational solutions* – This project's primary objective is to create a learning tool which is split into modules for easy understanding and demonstration purposes without overwhelming the beginners to computer programming. (6) *Creating Computational Abstractions* – Using software libraries, we created charts and graphs on the CoF data for abstracted results providing focused analysis on the visualizations without caring about the coding commands. (7) *Troubleshooting and Debugging* – This practice was also realized as part of the development process as it was done in an iterative manner, several sessions of troubleshooting had to be completed.
- **Three Systems Thinking Practices:** (1) *Investigating a Complex System as a Whole* and (2) *Communicating the information with stakeholders* – These practices were realized while developing the iDCNs that acts as an educational tool for beginners with computing. To provide a solution to this complex issue, we were able to understand the target audience and provide easy to understand solutions and not complicate them. To communicate the information to the stakeholders, appropriate comments and links to resources were made available, (3) *Understanding the Relationships within a System* – was made possible by looking at the enslaved and owner, dates, age, gender, places together with the help of contextual historical data research.

5.3 User Survey Results and Analysis

To conduct the User survey with the graduate students and educators with non-STEM backgrounds, the University of Maryland's Institutional Review Board (IRB) was contacted and presented with a package of the elements of the survey including providing a CONSENT form which was to be signed by survey participants at the beginning of the survey. The IRB members reviewed the package thoroughly for human-subject research determination and approved the survey for dis-

tribution. The survey was designed to be anonymous and no personal information was captured. The survey had three sections: a CONSENT form at the front, a Yes or No question requesting users to answer if they were able to review the iDCNs web page hosted publicly and a set of seven questions grouped together on a Likert scale as given below. The Likert scale questions were not all mandatory and the students could skip any questions if they did not feel comfortable answering them.

- Q1: Were the iDCN modules arranged logically and coherently?
- Q2: Does the idea of making text, pictures, visualizations along with computer programming code co-exist together, help make the iDCNs easy to understand?
- Q3: Do the iDCNs modules look appealing or inspire learning about the chosen dataset collection?
- Q4: Does learning from the iDCNs provide the same aesthetic value or more or less as studying from a paperback book / traditional book?
- Q5: Do the entire iDCNs modules follow the CT framework introduced and identified in the index module?
- Q6: Were these iDCNs educational?
- Q7: Can these iDCNs be used for further course development with other dataset collections?

These were on a 7-level Likert scale: Strongly disagree, Disagree, Somewhat disagree, Neutral, Somewhat agree, Agree, and Strongly Agree. The survey was distributed to two groups of students and educators, the vast majority of whom we believe were being introduced to this learning tool for the first time. It was sent to an overall count of n=37 participants of which 4 of them were educators and 33 were students. Of the 37 survey recruits, 31 were able to complete the survey at a response rate of 83.7%. 2 of them were unable to review the iDCNs and 4 of them were not able to complete the entire survey. Of those 31 who completed the survey, only 1 of them was not able to answer all of the 7 close-ended Likert scale questions. From Figure 5, which shows a chart of responses to these 7 questions, the results were analyzed holistically for overall feedback and key inferences were made. Although the sample size is not a big number to make generalizations from the survey evaluation, we believe the survey results capture first impressions and provide significant insights into the possible adoption of this novel educational tool for non-STEM educational areas. Survey responses were overwhelmingly positive on all 7 questions with some stronger reservations on questions Q3 and Q4 which probe on the look and feel side: contrasting iDCNs to traditional books, and asking about their appeal. This is valuable feedback, as one can be taken aback in the beginning by experiencing what may look like a "wall of text and code boxes." This suggests that particular care be given to not only telling an interactive computational story but also to its visual composition and representation. It was particularly encouraging to find extremely positive feedback for questions Q6 and Q7 which indicate that the iDCNs were educational in nature and that such courses could be developed in future for other dataset collections. This goes to show that we were able to achieve the primary objective of developing this learning tool. One of the students with historical background who was in the list of survey recruits sent a direct email to us with comments as quoted below. We were not sure if the student completed the survey due to anonymity of the data capture however, these comments surely indicated positive feedback. We could identify the yearning of powerful data tools that could work with historical data on a scale like this.

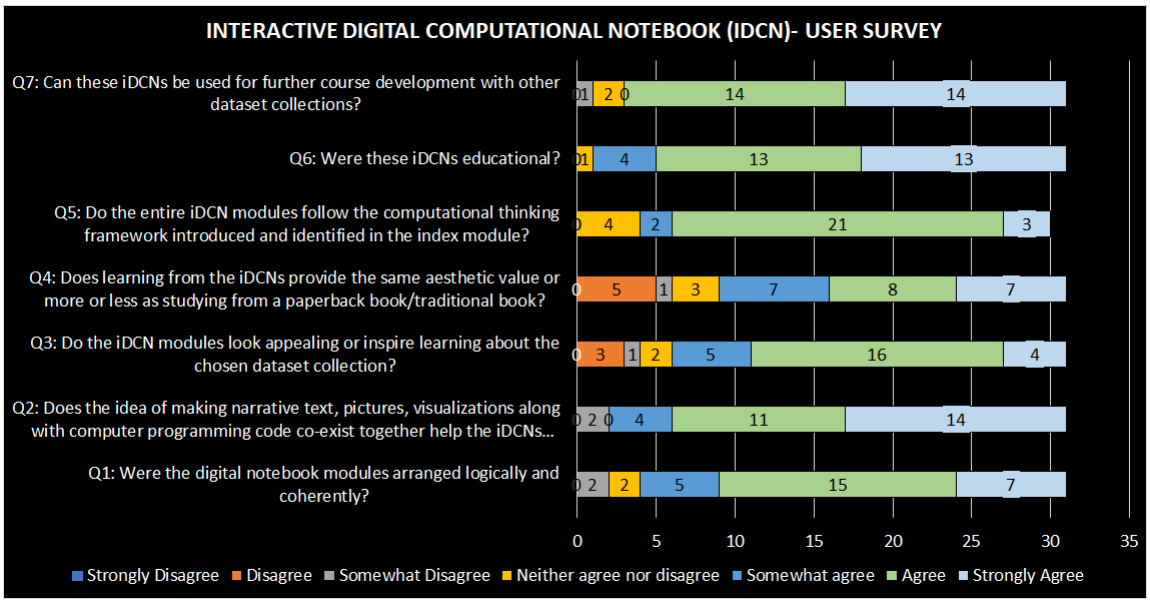


Figure 5: User Survey Results of Likert scale questions for the iDCNs, (n=31)

> "I think that some of the graphs that have already been created here do a good job of demonstrating the power and flexibility of these tools. The ability to deny or substantiate all sorts of historical claims using hard, quantitative data is of MASSIVE value to historians in the field. We've never been able to do this before on this scale, much less make it transferable to other scholars."

The survey results could be instrumental in deciding the future of a novel effort like this. To address the comments related to improving the aesthetic value of the iDCNs, consequently, one of our team members, with a background in art history and graphic design, is exploring disruptive ways of controlling the notebook layout to make it more visually appealing and less linear, through the use of extensions and widgets in particular. While very popular with data scientists and estimates several million JNs shared, their use in the library and archival space is very new. Adoption will take some time, but more importantly how they are created and look will most likely need to be adapted to non-STEM spaces.

6. Discussion

Though the objectives set for this project were achieved and the RQs were answered, it is understood that introducing a novel learning platform like this for traditional learners could be a challenge. The survey results, although predominantly positive and encouraging, do not capture the generalized reactions from the overall population of non-STEM educators and students. To bring back the point made by Levander and Decherney in [24], on how internet accessibility is a main issue with regards to underserved populations, the need to get online in order to access these iDCNs pose a significant challenge to reach all groups of students and educators. This challenge is particularly true for any online learning platform and needs to be addressed at an overarching higher educational level. Although the modular nature of the iDCNs developed provides a coherent

way to navigate within the course content, the impression as could be seen from some of the survey responses for Q1 indicates that this area could also be improved. This shows the limitations of working with an open-source tool designed and developed originally for purposes other than as a learning tool. Having said that, the non-profit organization managing the Jupyter suite of products are in the process of improving their tools with the introduction of new products that cater towards collaboration. In addition, due to the open-source nature of these tools, there are customized solutions built on top of these by different developers that provide solutions to these challenges. *Jupyter Books* [17] is one such project where several JNs could be combined together and presented as a coherent rendering of content which simulates the experience of pages in a book for easy navigation and with a digital table of contents.

7. Future Work and Conclusion

In this project, owing to the nature of it being an educational tool developed to reach students and educators from different levels of programming experience, the iDCNs only included basic functions supported by *Python* language. But this infrastructure could be developed to harness the power of integrating with software libraries to create iDCNs that could explore datasets on topics like Machine Learning (ML) concepts and advanced Natural Language Processing (NLP) processes. From the survey responses, it is clear that there is some approval to move ahead with similar data science driven approaches on other MSA's dataset collections [15] like Runaway Slave Ads, Manumissions, etc. By creating, sharing and making these open-source learning platforms available online for free, we hope these resources help students, educators and researchers across the world to leverage these and get benefitted from them. Especially, in today's world with the changing landscape of digitization in every field of study and work including HASS areas, acquiring knowledge of performing these unique data operations becomes truly essential. The iDCN modules have been published at the *https://cases.umd.edu.* [18] The source code is also published to *GitHub.* [19] A youtube video of the presentation of this paper at ISGC2021 is available at https://youtu.be/cNBc0AY-r-k

We wish to acknowledge funding from the 2020-2022 IMLS Laura Bush 21st Century Librarian Program: "Piloting an Online Collaborative Network for Integrating CT into Library and Archival Education and Practice." This grant supports the piloting of an online collaborative network of educators and practitioners to enable the sharing and dissemination of computational case studies and lesson plans through a JN interactive computational learning platform, called CASES [Computational Archival Science Educational System], see: https://cases.umd.edu. Special thanks to Christopher Haley Director of the Study of the Legacy of Slavery in Maryland program, at the Maryland State Archives, and the the IMLS Piloting project educators: Sarah Buchanan (U. Missouri), Karen Gracy (Kent State U.), and Joshua Kitchens (Clayton State U.).

[17] Books with Jupyter - https://jupyterbook.org/intro.html.
[18] Legacy of Slavery - https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb.
[19] Legacy Of Slavery GitHub project - https://github.com/cases-umd/Legacy-of-Slavery.

References

[1] C. Haley, "Maryland state archives - legacy of slavery home - [online]." http://slavery.msa.maryland.gov/, 2021.

[2] S. Gallagher and J. Palmer, "The pandemic pushed universities online. the change was long overdue." https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue, Sep, 2020.

[3] J. DeVaney, G. Shimshon, M. Rascoff and J. Maggioncalda, "Higher ed needs a long-term plan for virtual learning." https://hbr.org/2020/05/higher-ed-needs-a-long-term-plan-for-virtual-learning, Feb, 2021.

[4] S. Dunn, "What versus how: Teaching digital humanities after covid-19." https://blogs.kcl.ac.uk/ddh/2020/05/04/teaching-digital-humanities-after-covid19/, Nov, 2020.

[5] E. Goudarouli, A. Sexton and J. Sheridan, *The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK*, *Philosophy & Technology* **32** (2019) 173.

[6] W. Underwood, D. Weintrop, M. Kurtz and R. Marciano, *Introducing Computational Thinking into Archival Science Education*, pp. 2761–2765, 2018, DOI.

[7] N.G. Taylor, J. Moore, M. Visser and C. Drouillard, *Incorporating computational thinking into library graduate course goals and objectives*, *School Library Research* **21** (2018) .

[8] A. Davies, F. Hooley, P. Causey-Freeman, I. Eleftheriou and G. Moulton, *Using interactive digital notebooks for bioscience and informatics education*, *PLOS Computational Biology* **16** (2020) e1008326.

[9] J. Reades, *Teaching on Jupyter: Using notebooks to accelerate learning and curriculum development*, vol. 7 (01, 2020), DOI: 10.18335/region.v7i1.282.

[10] A. Zúñiga-López and C. Avilés-Cruz, *Digital signal processing course on Jupyter–Python Notebook for electronics undergraduates*, *Computer Applications in Engineering Education* **28** (2020) 1045.

[11] M. Wigham, L. Melgar and R. Ordelman, *Jupyter notebooks for generous archive interfaces*, in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2766–2774, 2018, DOI.

[12] L.A. Perine, R.K. Gnanasekaran, P. Nicholas, A. Hill and R. Marciano, *Computational treatments to recover erased heritage: A legacy of slavery case study (ct-los)*, in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1894–1903, 2020, DOI.

[13] J.M. Wing, *Computational thinking*, *Commun. ACM* **49** (2006) 33–35.

[14] D. Weintrop, E. Beheshti, M. Horn, K. Orton, K. Jona, L. Trouille et al., *Defining computational thinking for mathematics and science classrooms*, *Journal of Science Education and Technology* **25** (2016) 127.

[15] C. Haley, "Maryland state archives - legacy of slavery - dataset collection - [online]." http://slavery2.msa.maryland.gov/pages/Search.aspx, 2021.

[16] D.B. Driscoll, "The faces of freedom." https://bayweekly.com/the-faces-of-freedom/, Feb, 2020.

[17] C. Haley and M. Davis. https://bluetoadpublishing.co.uk/publication/?m=30305&i=572010&p=24, Mar, 2019.

[18] S. Papadakis, M. Kalogiannakis and N. Zaranis, *Developing fundamental programming concepts and computational thinking with ScratchJr in preschool education: a case study*, *International Journal of Mobile Learning and Organisation* **10** (2016) 187.

[19] L.A. Barba, L. Barker, D. Blank and A. Brown, "Teaching and learning with jupyter." https://jupyter4edu.github.io/jupyter-edu-book/why-we-use-jupyter-notebooks.html#why-do-we-use-jupyter, Dec, 2019.

[20] O. Smith, "Markdown in jupyter notebook." https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook, Aug, 2019.

[21] M. Sun and T. Oza, *User survey: The benefits of an online collaborative contract change management system*, *Electronic Journal of Information Technology in Construction* **15** (2010) 258.

[22] R. Margo and R. Steckel, *The heights of american slaves: New evidence on slave nutrition and health*, *Social science history* **6** (1982) 516.

[23] C. Haley, "A guide to the history of slavery in maryland." https://msa.maryland.gov/msa/intromsa/pdf/slavery_pamphlet.pdf, 2007.

[24] C. Levander and P. Decherney, "The covid-igital divide." https://www.insidehighered.com/digital-learning/blogs/education-time-corona/covid-igital-divide, Jun, 2020.

A. Appendix Section: Additional Figures supporting the main paper

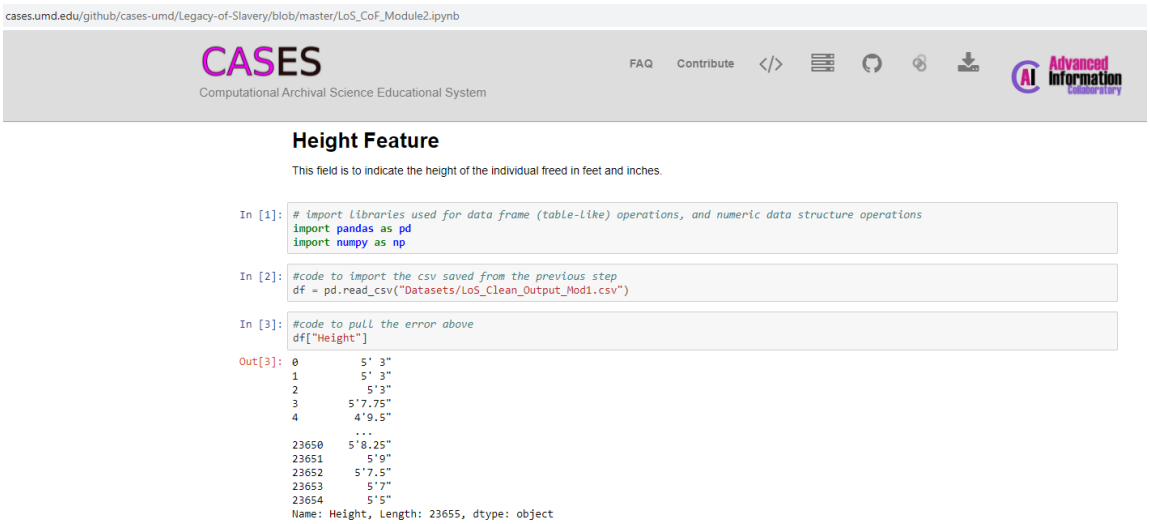


Figure 6: Section 5.1 - Sample published iDCN module hosted publicly on cases.umd.edu

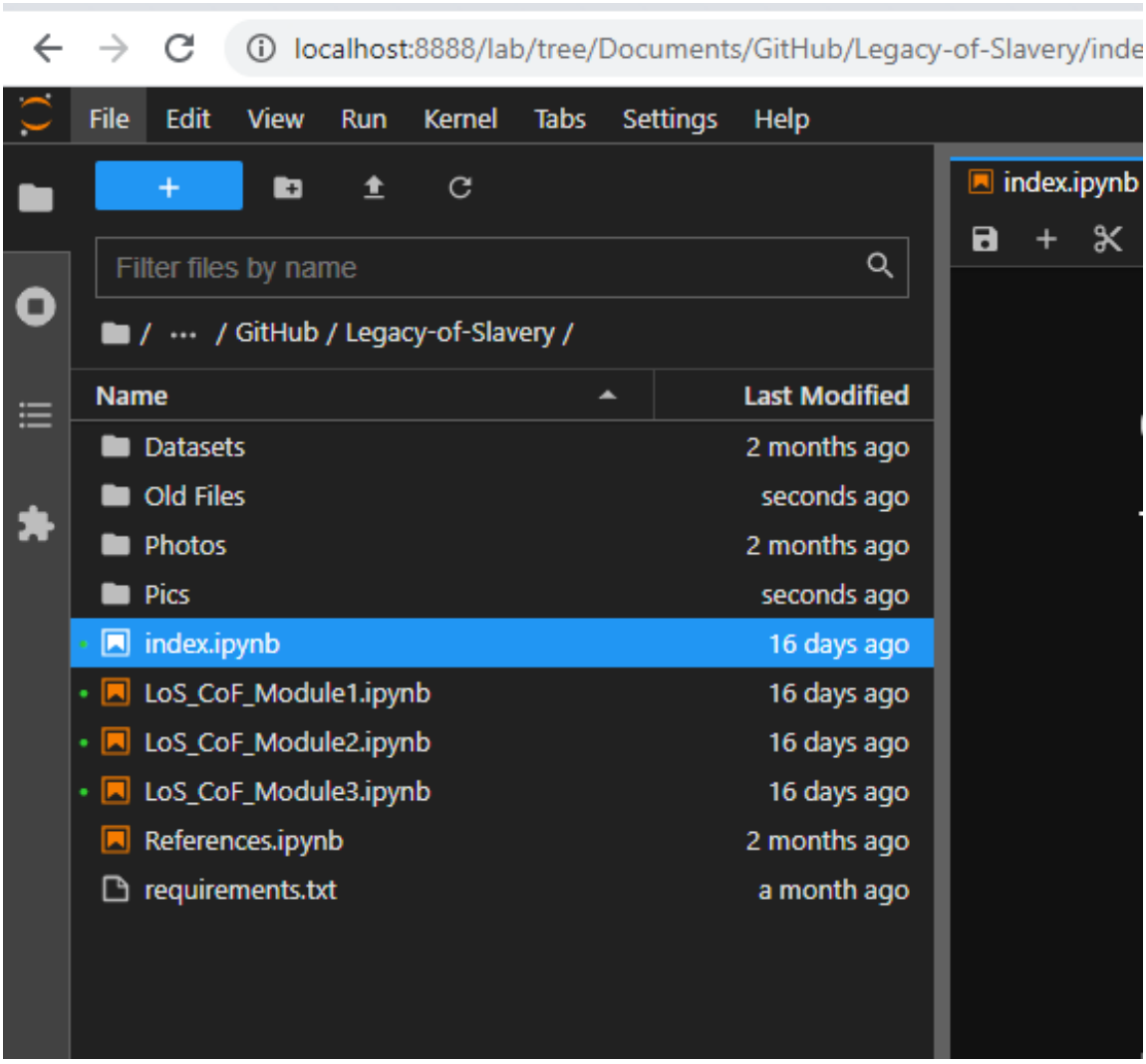


Figure 7: Section 5.1 - Screenshot showing *JupyterLab* Development Environment with iDCN Module Layout

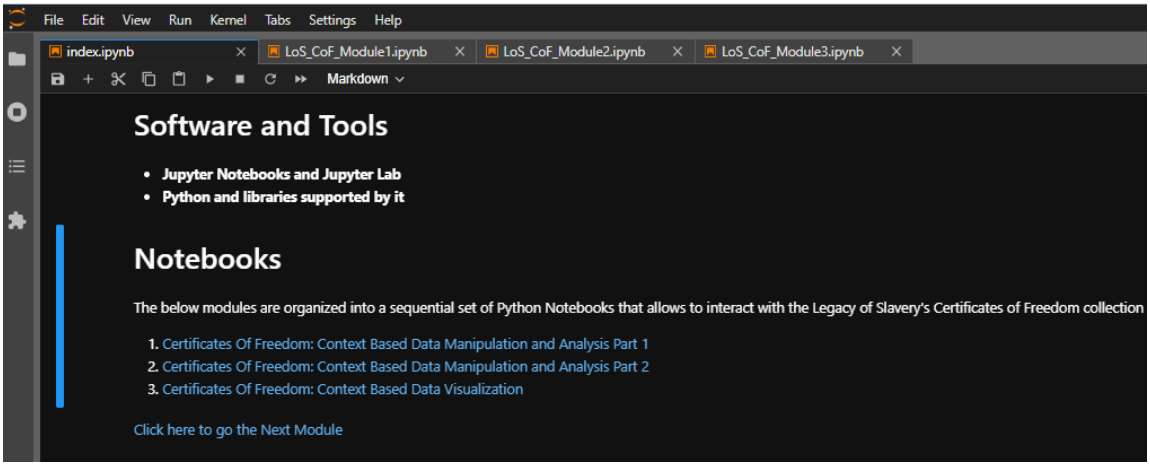


Figure 8: Section 5.1.1 - Screenshot showing a sample of Markdown edited text with Hyperlinks

```
# Below command displays the specific records that was identified as erroneously entered. The inner command 'df[
# in-built function to filter the records that match with the supplied criteria and the outer 'df[]' displays th
df[df['Date'].astype(str).str.strip()=="184006"]
```

	DataID	DataItem	County	Owner_FirstName	Owner_LastName	Witness	Date	Freed_FirstName	Freed_LastName	Alias
23307	AR7-46	23310	BA	Geo	Gillingham	NaN	184006	Jeremiah W.	Brown	Jerry
23308	AR7-46	23311	BA	NaN	NaN	NaN	184006	Rachael	Brown	NaN

2 rows × 29 columns

In two of the instances, as seen below, the day of issue has not been found to be legible or visible, hence the MSA transcriber may

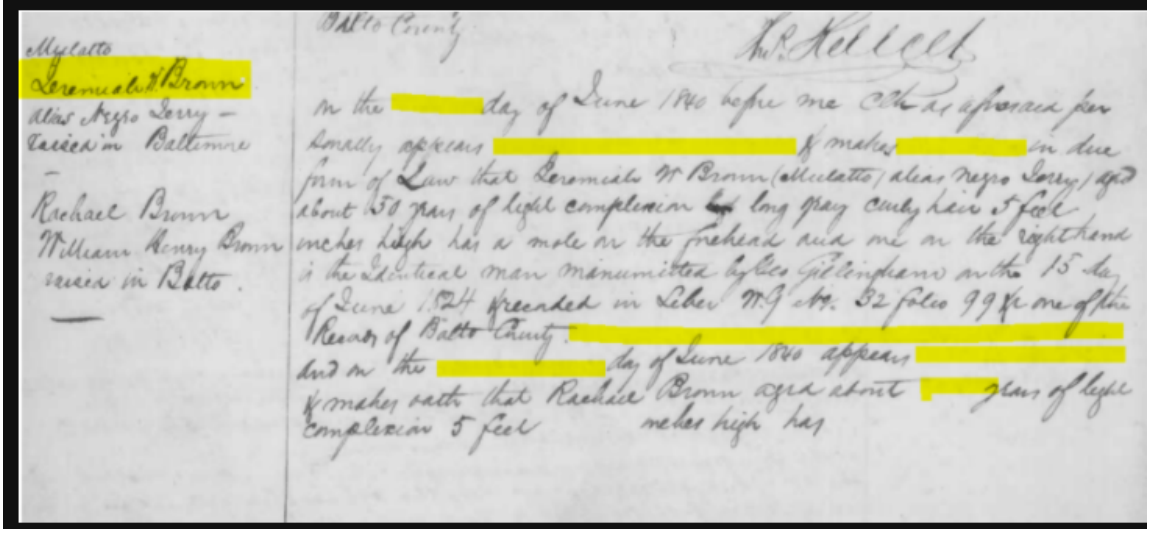
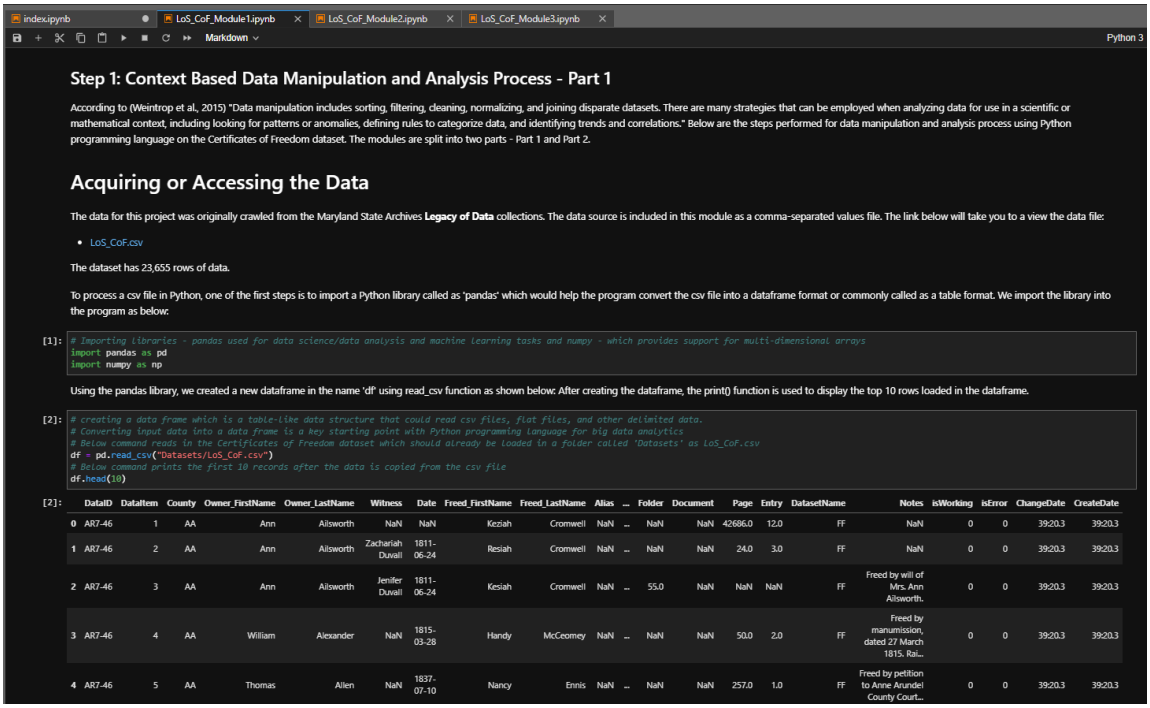


Figure 9: Section 5.1.2 - Screenshot showing data manipulation of erroneous date records from the CoF dataset



index.ipynb | LoS_CoF_Module1.ipynb | LoS_CoF_Module2.ipynb | LoS_CoF_Module3.ipynb

Markdown | Python 3

Step 1: Context Based Data Manipulation and Analysis Process - Part 1

According to (Weintrop et al., 2015) "Data manipulation includes sorting, filtering, cleaning, normalizing, and joining disparate datasets. There are many strategies that can be employed when analyzing data for use in a scientific or mathematical context, including looking for patterns or anomalies, defining rules to categorize data, and identifying trends and correlations." Below are the steps performed for data manipulation and analysis process using Python programming language on the Certificates of Freedom dataset. The modules are split into two parts - Part 1 and Part 2.

Acquiring or Accessing the Data

The data for this project was originally crawled from the Maryland State Archives **Legacy of Data** collections. The data source is included in this module as a comma-separated values file. The link below will take you to a view the data file:

- LoS_CoF.csv

The dataset has 23,655 rows of data.

To process a csv file in Python, one of the first steps is to import a Python library called as 'pandas' which would help the program convert the csv file into a dataframe format or commonly called as a table format. We import the library into the program as below:

[1]:
```
# Importing libraries - pandas used for data science/data analysis and machine learning tasks and numpy - which provides support for multi-dimensional arrays
import pandas as pd
import numpy as np
```

Using the pandas library, we created a new dataframe in the name 'df' using read_csv function as shown below: After creating the dataframe, the print() function is used to display the top 10 rows loaded in the dataframe.

[2]:
```
# creating a data frame which is a table-like data structure that could read csv files, flat files, and other delimited data.
# Converting input data into a data frame is a key starting point with Python programming language for big data analytics
# Below command reads in the Certificates of Freedom dataset which should already be loaded in a folder called 'Datasets' as LoS_CoF.csv
df = pd.read_csv("Datasets/LoS_CoF.csv")
# Below command prints the first 10 records after the data is copied from the csv file
df.head(10)
```

[2]:

	DataID	DataItem	County	Owner_FirstName	Owner_LastName	Witness	Date	Freed_FirstName	Freed_LastName	Alias	...	Folder	Document	Page	Entry	DatasetName	Notes	isWorking	isError	ChangeDate	CreateDate
0	AR7-46	1	AA	Ann	Ailsworth	NaN	NaN	Keziah	Cromwell	NaN	...	NaN	NaN	42686.0	12.0	FF	NaN	0	0	39:20.3	39:20.3
1	AR7-46	2	AA	Ann	Ailsworth	Zachariah Duvall	1811-06-24	Resiah	Cromwell	NaN	...	NaN	NaN	24.0	3.0	FF	NaN	0	0	39:20.3	39:20.3
2	AR7-46	3	AA	Ann	Ailsworth	Jenifer Duvall	1811-06-24	Kesiah	Cromwell	NaN	...	55.0	NaN	NaN	NaN	FF	Freed by will of Mrs. Ann Ailsworth.	0	0	39:20.3	39:20.3
3	AR7-46	4	AA	William	Alexander	NaN	1815-03-28	Handy	McCeomey	NaN	...	NaN	NaN	50.0	2.0	FF	Freed by manumission, dated 27 March 1815. Rai...	0	0	39:20.3	39:20.3
4	AR7-46	5	AA	Thomas	Allen	NaN	1837-07-10	Nancy	Ennis	NaN	...	NaN	NaN	257.0	1.0	FF	Freed by petition to Anne Arundel County Court...	0	0	39:20.3	39:20.3

Figure 10: Section 5.1.2 - Screenshot showing Part 1 module with *Python* code snippet to import csv and read first few records

```
# show descriptive statistics
df["Height_Inches"].describe()
```

```
count    22320.00000
mean        64.36112
std          4.13348
min         21.50000
25%         62.00000
50%         64.50000
75%         67.00000
max        145.00000
Name: Height_Inches, dtype: float64
```

Figure 11: Section 5.1.3 - Screenshot showing descriptive statistics for Height feature after conversion to inches (indicates outliers)

```
# code to show bad records
df.loc[(df["Height_Inches"]>80)|(df["Height_Inches"]<5),['DataItem','Height','Height_Inches']]
```

	DataItem	Height	Height_Inches
5032	5034	4' 44.75"	92.75
5625	5627	5'2 1.4"	81.40
6197	6199	5'85."	145.00
15694	15696	9'.75"	108.75

Figure 12: Section 5.1.3 - Screenshot showing erroneous values of Height feature after data manipulation operation

```
# code to find the converted Age records
df.loc[(df["Age"]<1),['DataItem','Age','AgeFormatted']]
```

	DataItem	Age	AgeFormatted
11876	11878	0.08	0.666667
13870	13872	0.18	1.500000
19625	19627	0.80	6.666667
22819	22822	0.09	0.750000
22849	22852	0.18	1.500000
22861	22864	0.03	0.250000
22890	22893	0.08	0.666667

Figure 13: Section 5.1.3 - Screenshot showing old and new Age features comparing values before and after data cleansing operation

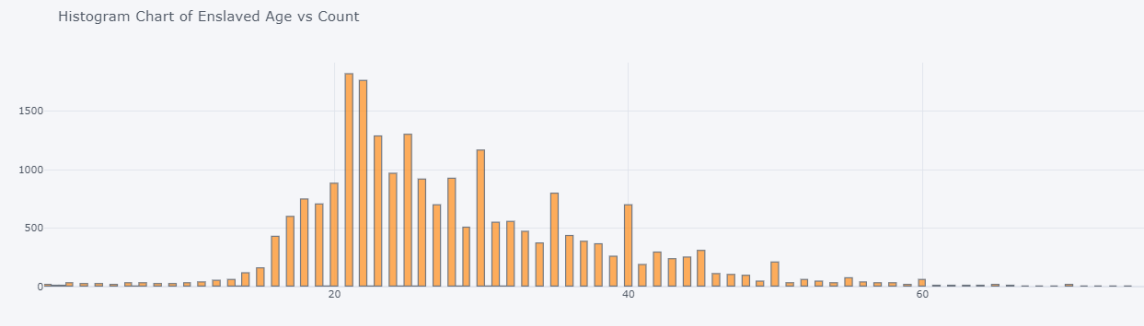


Figure 14: Section 5.1.4 - Screenshot showing a histogram chart between Age feature and the number of CoFs issued

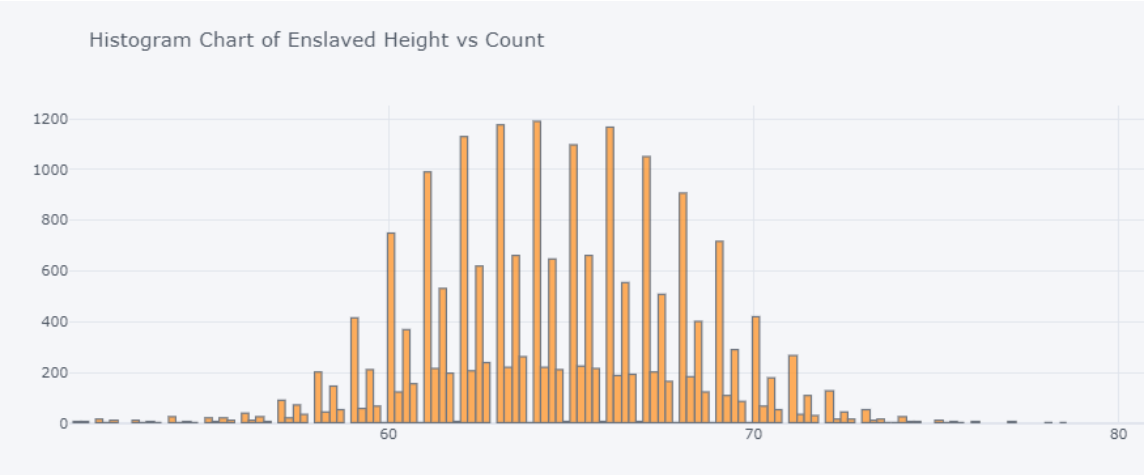


Figure 15: Section 5.1.4 - Screenshot showing a histogram chart between Height feature and the number of CoFs issued

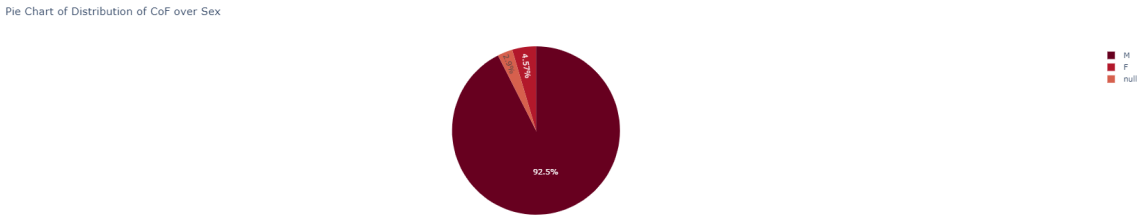


Figure 16: Section 5.1.4 - Screenshot showing a pie chart of percentage of CoFs issued by gender

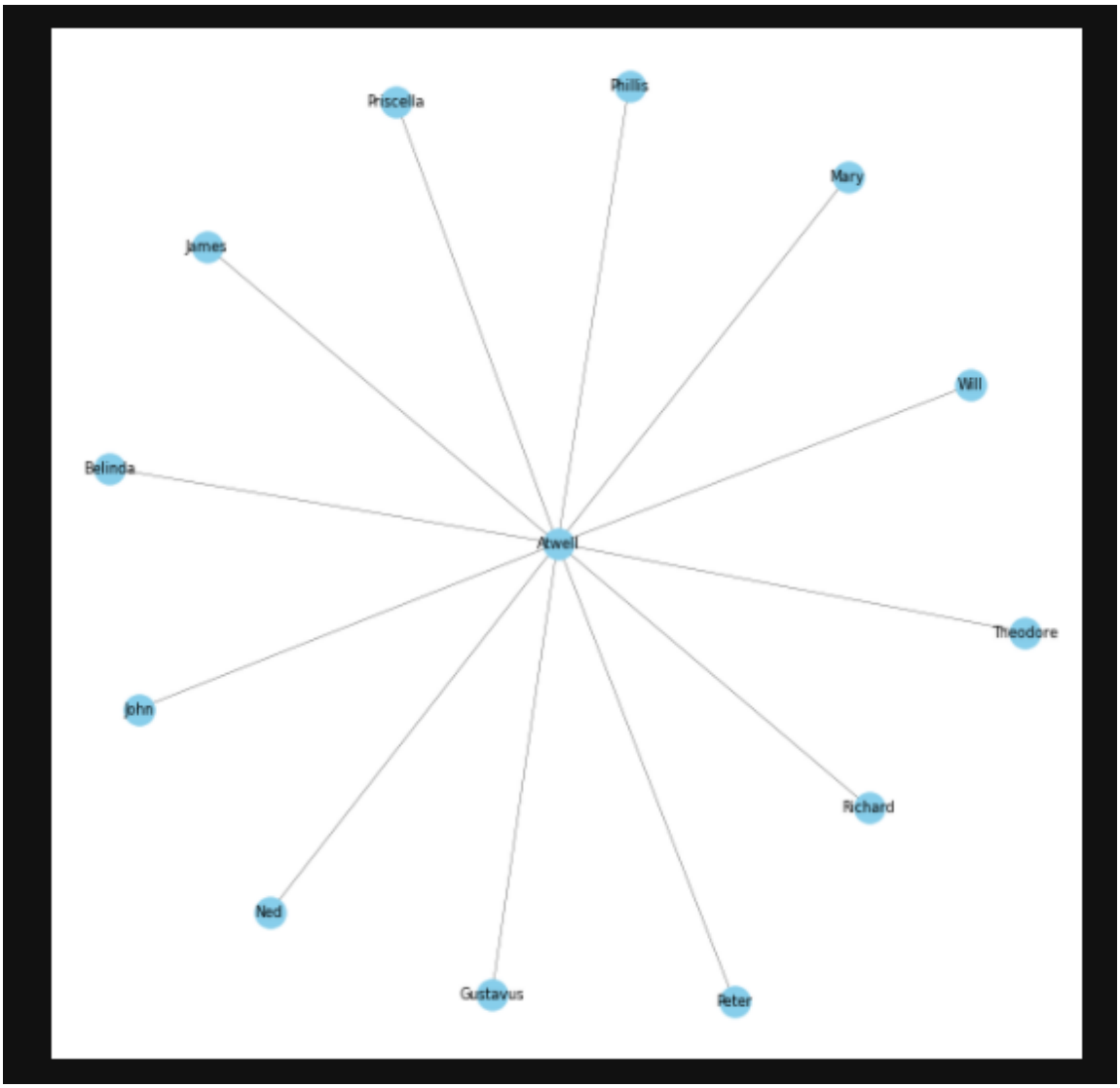


Figure 17: Section 5.1.4 - Screenshot showing a network diagram between Enslaved people and Slave owner (centre)

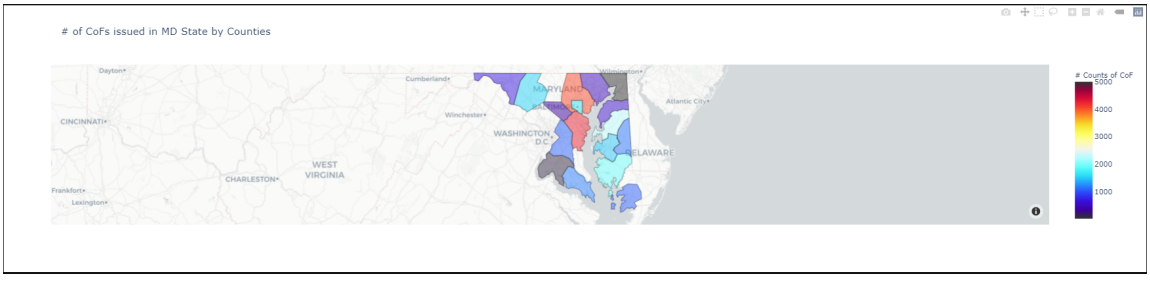


Figure 18: Section 5.1.4 - Screenshot showing a geomap diagram using FIPS County codes in MD state of the USA