PORTFOLIO MILESTONE

Jennifer E Williams

SUID: 240356487

[Jwill131@syr.edu](mailto:Jwill131@syr.edu)

**TABLE OF CONTENTS**

1. Introduction
2. Demonstrate broad ranging practice areas
3. Collect and organize data and identify patterns in data via visualization, statistical analysis and data mining
   1. IST719 Data Visualization Final Project Poster
   2. IST718 Big Data Analytics Lab 1
4. Develop alternative strategies based on the data

IST659 Database Generation

1. Develop a plan of action to implement the business decision derived from the analysis

IST644 Reflection Paper

1. Demonstrate communication skills regarding data and its analysis/synthesize the ethical dimensions of data science practice
2. Conclusion

**1. Introduction:**

Documentation about this program and this degree describes the data science program as “a hybrid program, offering a cross-functional curriculum and student experience1.” Prospective students will likely navigate their way to the ischool website where the Applied Data Science Program is introduced as “As an interdisciplinary program, the master’s in Applied Data Science provides students the opportunity to learn in a broad range of areas related to data science2.” What the name Applied Data Science implies, but these descriptions do not explicitly articulate, is the truly applied nature of this degree and the instruction associated with its completion, blending theory and hands-on/real world experience.

Like all students, I approached this experience with my own lens – colored and shaped by decades in the workforce. My journey of data science had been largely circuitous, convoluted, and certainly not conventional. Nonetheless, I felt my wide ranging “real world” experience would serve me in spades. I found myself almost immediately out of my depth. Simply stated, my experiences represented specific niche elements of one element or application of data science but I acutely lacked the scope and understanding, both theoretically and hands-on, to understand how look at a data science problem and know what my next step would be. Through classes ranging from Business/Marketing and Financial Analytics to Quantitative Reasoning and Applied Machine Learning and a wide spectrum in between, I have come to appreciate the depth and breadth of data science as a field and get just a notion of how to approach a data science problem across that wide spectrum.

**2. Demonstrate broad ranging practice areas**

This final element proved key – how to approach a data science problem across a spectrum that represents the data science field. Armed with the wide ranging curriculum I embarked upon, these courses didn’t teach me what code to write in a specific situation but instead how to evaluate a data science problem from a multitude of angles and identify the way ahead. To demonstrate this **broad ranging practice areas**, I’ve selected projects from IST659 Data Admin Concepts and Database Management, IST644 Managing Data Science Projects, IST719 Information Visualization and IST718 Big Data Analytics. Collectively, these courses and some of their critical projects are representative of the programs learning goals. It is important to note, however, that these courses and these projects may underscore the holistic and synergistic element of the degree program and how the learning objectives of all the courses and assignments became bricks in a foundation that were all interconnected and all necessary.

**3. Collect and organize data and identify patterns in data via visualization, statistical analysis and data mining:**

Critical to the data science process is data collection, data understanding and then conveying the key points of that data to stakeholders. The final project for IST719 Data Visualization Poster and the first lab for IST718 Big Data Analytics typified the importance of data collection and organization with the identification of patterns and synergized the theoretical and real-world applications of data science.

**a. IST719 Data Visualization Final Project Poster** sought to identify a data topic and execute the process of data collection, organization and then optimized visualization. This assignment coalesced the elements of the course: cleaning the data, exploring the data with visualization techniques, data aggregation, simple design and information organization, and graphic presentation – into a single poster to convey the key points of a data problem or data set. Throughout my participation in the degree program, ideas for projects do not always come organically. My job is in the government and therefore utilizing or approximating that data is unrealistic and infeasible due to the nature and classification of the work. Around the time I took this particular course, I had been watching the TV series Dopesick and was curious about applying a data perspective to current understanding of the Opioid situation in the United States. The techniques and tools that I learned for manipulating and visualizing data in a way that will optimize the understanding of the message will serve me in my work and throughout my life. For me, a challenge was editing and finding ways to convey all the points that I felt were critically important without losing the message and the appropriate visual. If everything is important then nothing is and I must admit, my final project would have benefited from some additional editing. That realization reminded me of a critical point, just because the data is there and the work went into the reduction and visualization, does not mean that it is important enough to include.

**b. IST718 Big Data Analytics Lab 1** was quite the challenge. This assignment sought a data and machine learning solution to answer the question of what is the recommended salary for the Syracuse head football coach. Having worked the entire degree program until this final course in R, I had honestly been dubious of the switch to Python. Having watched some instructional videos but not yet gotten my hands dirty with coding yet, I was crestfallen when I realized the crawl phase presumed an already basic understanding of how to do all the things we had learned in R. As Dr. Fox had highlighted, the data set provided was inadequate and needed augmentation, ultimately taking five datasets and merging them into a singular set with sufficient data to answer the data problem. My data cleaning was not elegant nor was it optimized and admittedly, there are many efficiencies that I still need to learn. The process, however, of taking the collective skills that all the previous courses had taught and applying them to a new coding language was eye opening and euphoric. In a new language, I was able to collect and organize multiple datasets, clean and organize them and then identify the right visuals to convey my point. To me, this exemplified an application of the data science process as it took the understanding of how the process works and using that knowledge to learn something new. I will use not only the python application but the knowledge and confidence to apply data science constructs to new situations.

**4. Develop alternative strategies based on the data:**

This learning goal is slightly more nebulous for me than data organization and data visualization. For me, this is the comfort with the tools and the concepts to go where the data leads and where the data problem needs to go. Without an understanding of the underlying structure of the data and the database, it is difficult to manifest an alternate strategy and pivot as needed. For that reason, the final project for **IST659 Database Generation** with data manipulation and interface development gave me the context to be confident to develop alternative strategies in the future.

IST659 Final Project Database Generation sought to design and implement a database to solve a data management problem. This process included conceptual design of a database through to execution of a data project. This project proved pivotal to later assignments and my other data work because it showed me the mapping of interactions of data. By looking at the logic models of the data as a database or dataset as a whole, I could understand how the data interacted and how it should interact with outside data. That critical understanding enabled me to get the interplay of data better in subsequent assignments and in my work. That interplay became critical to problem solving and provided the lexicon for viewing the data from different perspectives.

**5. Develop a plan of action to implement the business decision derived from the analysis:**

The natural choice for me to address this learning goal is the **IST644 Reflection Paper**, intended to write to your future self and explain your thoughts and insights on data science project management. I approached this course with a slightly casual attitude, convinced that my years in my industry managing projects and my PMP certificate would render my organic knowledge largely sufficient. To say that I was incorrect is an understatement. To say that this course had the most significant short-term impact would also be an understatement.

Implementation of business decisions for me is the manifestation of data in the form of a project intended to answer a specific data related problem. Data projects are decidedly different than more conventional projects and therefore require their own very specific processes and lexicon. That nuance was lost on me until completing IST644. The notion of the Reflection Paper was to write a missive to your future self and explain your thoughts and opinions about managing data science projects based on the structured constructs. This exercise forced me to recognize the differences between data science projects and other projects, forced me to support and defend my positions on executing those data science projects, and enabled me to envision applications of the methodologies and approaches in my current and future work. In a very timely twist of fate, this course occurred immediately before a move courtesy of the miltiary. When applying for new jobs, almost all the concepts addressed in the Managing Data Science Projects course and discussed in the Reflection Paper became pivotal and yielded new job opportunities.

**6. Demonstrate communication skills regarding data and its analysis/synthesize the ethical dimensions of data science practice:**

Communications and ethics are pervasive concepts throughout data science and therefore should and are represented by each of the projects. Communications provides a lens to convey the “so what” of the data and corresponding analysis. What this degree program has taught me, and what these four projects represent, is that the method of communication is not a singular constant but rather dependent on the audience, the message, and the data.

IST659 presented a different way of understanding databases and data structures and that presented an extremely powerful mechanism for communicating data needs and data interactions. By communicating effectively the needs of the data analysts to the data engineers, the likelihood of project success increases. Without the ability to communicate or have a shared understanding and shared lexicon, the analyst will be far less likely to get the data they need. This conceptual database visualization provides a lexicon to communicate the data needs.

IST718 and IST19 looked at the graphic visualizations to communicate the “so what” but also how the language that is coded can impact understanding. From a psychology perspective, most individuals today identify as visual learners and therefore results in a visual form are going to be far more impactful and resonate in a way that verbal or auditory results may not. Identifying the right graphic is critical to conveying that “so what” and both the IST718 and IST719 looked directly at that problem and did so in different languages. Not only do the different languages provide different visualizations but people are more or less proficient in different languages so the ability to code and articulate and extract those data visualizations in languages that may or may not enable communication between analysts more effectively is key.

IST 644 looked less at the visualization or the data aspect of data science and more at the importance of conveying a clear and concise message in a business context. If the data engineers have a complete and thorough understanding of the analysts needs and the data visualizations are stunning but the analysis cannot be conveyed in a clear and concise manner to the stakeholders then the value of the work is lost. This is the summation of the application of data science communication in my current and future work. If I cannot adequately convey my data needs and extract my data, visualize it in a powerful manner and then coalesce that into a report that can articulate that message and the analysis to my stakeholders, my project is not a success.

But what of ethics? IST644 directly posed the question of how is ethics handled in data science projects and the observation therein highlights the observation here, it is and must be everywhere. While easier to say than to actually execute, the **IST644 Reflection Paper** discussed not only the overt aspect of ethics in the form of using data in an ethical and appropriate manner but also the element of bias in data, which is much more nuanced and difficult to identify and mitigate. All the projects completed in pursuit of this degree needed to be executed with an understanding that I was representing the data accurately, that I was conveying the analysis in a way that was both complete, correct and with a conscious of the implications on those reading and interpreting that analysis. It would be morally unconscionable to misrepresent data, only present a portion of data most suitable to my hypothesis or position, or allow my personal bias’s to cloud my use of data or analysis.

The optimal example of this could be the **IST719 Data Visualization Project** and its discussion of Opioid use in the United States. I attempted to remove my own bias from my data analysis and present an ethically sound presentation on Opioids but as I have stopped myself three times now from writing “Opioid Crisis” I must recognize that my own bias and interpretation influences the analysis. Critical to identifying and therefore mitigating bias is multiple viewpoints as I discussed in the reflection paper. By enabling projects with multiple individuals and viewpoints, these biases can be hopefully identified, mitigated and a more ethically sound report produced.

**7. Conclusion**

I mentioned that I had recently moved due to a miltiary change of primary duty station. Every couple of years our family moves and I search for a new job. Completing this dance repeatedly means that I had grown somewhat used to it. With an undergraduate in History and a Masters in Arts in Forensic Psychology, my education and job experience is admittedly a bit haphazard. This, however, was the first move where I had my projected completion of this Applied Data Science Degree on my resume. I entered this job hunting and the corresponding interviews with a sense of confidence imbued by the holistic application of data science with both theoretical knowledge and real world application. I felt empowered. It wasn’t one course or one project, or even four courses/projects for that matter. It was the totality of the courses, the instruction and the understanding on how to apply them as building blocks in my execution of data science in my future.

References

1. (n.d.). (publication). *Data Science@Syracuse Student Advising Packet* (pp. 1–4).
2. *Applied Data Science Master's degree - ischool: Syracuse University*. iSchool. (2022, June 2). Retrieved July 16, 2022, from https://ischool.syr.edu/academics/applied-data-science-masters-degree/