Jack Foster
3/8/2025
DS-401

Data Science Reflection Paper

The data science minor has been an incredible experience during my time at Washington and Lee. Coming into college, I had a love of statistics from my AP Stats class in high school and hoped to further my knowledge of the subject as much as I could. Luckily, that's just what I was able to accomplish during my time at W&L. The first data science course I was able to take was Introduction to Statistics in the fall of my sophomore year. This class was a perfect intro to data science as it covered a wide basis of both statistical analysis tests that I ended up using throughout various other classes as well, and the basics of coding with R, another skill that was put to good use for many projects and assignments. With this being my first data science class, it provided the perfect background knowledge needed to be successful in classes taken after it, while also providing a challenging and engaging experience in its own right. This class is also where I completed the first of the three projects that I plan on including in my portfolio. My final project in this class was a solo statistical analysis done on the three-point attempt tendencies of the Golden State Warriors in the playoffs compared to the regular season. I chose this project specifically as it was the starting point in my journey in data science projects and offers a benchmark to see how far I have come in terms of my ability to complete statistical analyses. For the project, I had to find publicly available datasets online that were useful for my hypothesis in addition to trying to reduce all possible bias, which I did by using a random number generator to pick out 10 random regular season games to compare to the playoff average. I was then able to do a chi-squared test to calculate a p-value of .000267, allowing for me to come to the conclusion that the warriors did indeed end up changing the amount of three-pointers taken during the playoffs. I found this conclusion to be a great success due to the p-value lying quite below the .05 value set, showing that my hypothesis was worth testing. Overall, Math-118 represented my first steps towards completing the minor while teaching valuable skills that came in handy for various classes that followed.

The next class I took for the minor was Statistics for Biology and Medicine (BIOL 201). This class was a perfect follow up to Math-118, as it showed how the various analyses and tests taught could be applied to real world applications. The main focus of this course was how to analyze and covert the results of medical statistical analyses to words and what that could mean for either research or diagnoses. Considering I major in Biochemistry, I found this class to be fascinating in the way it combined all of the information I was learning in the medical field with data science, showing how the two could be blended to better understand how both of these fields work together to get the best success. I also ended up choosing the final project from this class to add to my portfolio, which was a meta-analysis to determine whether exercise or diet had greater impact on weight loss. This process involved a lot of research to find research papers that both correlated with what we were attempting to find, while providing data that we could use to run tests of our own. While challenging, this project did a great job of teaching us how to find useful resources to use for our own benefit, while also showing that running tests from available data could lead to you coming to conclusions that others may not have seen. For example, using data from 4 different papers, I found that individuals who were only on a diet lost an average of 5 more pounds than those who only did exercise, leading to the conclusion that diet had the greater impact on weight loss. This project also had us create various graphs and visualizations on a poster once done, further teaching us how to communicate statistical findings in a concise and easy to understand method.

The next class I took for the minor was Linear Algebra (Math 222). This class deviated from analyzing data specifically but taught me a whole new style of mathematics. Linear algebra was very different from the Calculus and Statistic classes I had been taking before, but still offered a new perspective on how mathematics can help in the real world. Particularly, the use of matrices and their properties opened a new way to visualize numbers and formulas, all of which can be used to analyze data in completely new ways. Overall, I was glad to see how linear algebra could be applied to data science through its various aspects such as vector calculations and matrices.

Statistics in Korean Music was the next class I took for data science. While the most important aspect of this class for me was living abroad and immersing myself into Korean culture, we still learned lots of data science applications and was able to see firsthand how data science could be used on music. By studying the Benford distribution, we were able to learn of the various wide-ranging topics that it shows up in. With this information, we studied Korean music and worked to see if the note frequencies of the many styles it has contains the Benford distribution. This class combined learning coding, music, statistics, and culture in a perfect blend that will be an experience I will never forget.

Next up was Probability (Math 309). This class put good emphasis on studying probability with various formulas and tests and used coding for a good portion of the class to run the tests for us. I find probability as a whole to be a very interesting topic due to how it applies to every situation you could be faced with. It seemed like every test or homework question applied to a different industry, showing just how wide-ranging our work could be. The projects we completed, coding blackjack and poker, provided an entertaining view into how gambling could be maximized with working on solving the probability of winning each hand dealt. All in all, Math 309 really focused in on how we could apply statistics rather than learning the methods themselves, a good change of pace due to the amount of data science classes taken previously.

For my coding requirement for the minor, I took Database Management for Business (BUS 315). In most of my previous classes, R was used to code due to its ability to easily run statistical tests with datasets added. For this class, we used SQL due to its ability to create models that can be used for businesses specifically. Learning SQL was a bit of a struggle due to the need to understand how to both model a database, and how to query that model to produce desired analyses. With time however, I was able to see the various advantages that SQL offers due to its flexibility for what a business might need. This was most prevalent in the last project I will use for my portfolio, creating an example database for CVS. This project had us start from scratch, needing to model both an efficient and usable model for CVS pharmacies. This had us think of multiple sets of data a CVS might need to function at maximum efficiency, such as customer information, store information, and how to store payments. We next had to use our model to create 10 queries that could be needed for the store, combining both the coding and modeling sections of the class. By working on a project that could easily be needed for thousands of companies around the country, we got first-hand experience in how businesses both store and analyze their data. This class taught me yet another example of the importance of data science as a whole due to the necessity of SQL or other similar programs to the success of many businesses around the world.

The last class I am taking for the minor is Data Science in Medicine (DS 181), which I am currently attending. This class is a great pairing with BIOL 201 due to both of their connections with the medical field. This class has us look at how data science is used in medicine and the various current events that are better understood with statistical analyses. We specifically have looked at the pandemic and the many research papers that have arisen from its impact. From this, we have learned various analyses used for research and the ethical implications that come along with completing these tests. While I still haven't finished this class, what I have learned from this class is just how important work is surrounding the data of humans and how all medical discoveries couldn't have been done without data science at its core.