# Applied Data Science Portfolio Documentation

The Applied Data Science program at Syracuse University greatly strengthened my skills as a data scientist and enabled me to go and leverage those skills in a meaningful way in my career. The most critical of these skills include:

* Collecting, storing, and accessing data with applicable technologies
* Creating actionable insights for many different contexts
* Applying visual and predictive models to generate insights
* Communicating insights to a broad range of audiences, both technical and business minded
* Applying ethics in the development, use and evaluation of data and models
* Demonstrating the ability to gain insight from data and models with industry standard software packages, including R, Python, and SQL

The entire curriculum provided engaging material to learn and discuss with the faculty and my peers, but three courses stand out as foundational in the skills that I am taking away from the program:

* IST 718: Big Data Analytics
* IST 707: Data Analytics
* IST 719: Information Visualization
* IST 722: Data Warehouse

# IST 718: Big Data Analytics

IST 718 was my favorite course in the entire program, as I felt myself truly grow while learning the material. I will be referring to the Labs in this course, instead of the final project, as I actually learned more from doing the labs, where as the final project was more of an exercise in running a powerful machine learning algorithm.

In Lab 1 (Zelazny, IST 718 Lab 1, 2019), we were tasked with creating a model to recommend a salary for the head coach of the Syracuse Football program. Most other projects in other classes provided CSV files or links to existing datasets to use, but I was forced to collect 100% of the data myself for this assignment. This included scraping data websites including Wikipedia for details on stadium capacity, and the NCAA website for historical records of wins for various college programs. We then had to build a model using Python out of the data collected to predict a coach’s salary based on a given set of factors, with the historical win-rate of the coach and the expected level of competition (i.e., conference) being the largest contributors.

In Lab 2 (Zelazny, IST 718 Lab 2, 2019) we leveraged a powerful Python package called fbprophet[[1]](#endnote-1) to build time-series models of residential housing prices throughout the United States and then identify several markets as areas for possible investment in real estate. This assignment demonstrated the power that we have at our fingertips in the various machine learning packages, and also enabled me to learn new methods of visualization and to communicate the insights gained to a technical audience.

In Lab 3 (Zelazny, IST 718 Lab 3, 2019) we used the popular MNIST Fashion dataset [[2]](#endnote-2)to compare the effectiveness of various classification models. This assignment was incredibly technical, requiring the ingesting of visual data and converting it into a format usable for machine learning, as well as constructing several different models including naïve-bayes, neural networks, and decision trees. A key learning moment for me was when I took feedback from the instructor on the previous assignment and ensured that I included a summary of my results in the first section of the report, which he referred to as “bottom-line-up-front”. This advice in communication was immediately transferrable to my current work as a propulsion systems engineer and has served me well.

# IST 707: Data Analytics

While typically earlier in the curriculum, I completed IST 707 near the end of my studies in this program.

The final project for this class (Zelazny, IST 707 Project: Student Performance, 2021) is similar to much of what I had learned in IST 718, gathering and cleaning data, building and comparing models, drawing insights and communicating them. However, the analysis for IST 707 was completed in R instead of Python. Many other courses in the program teach the use of the R program, but being more comfortable with Python before starting the program I wasn’t truly comfortable with the language until IST 707, and I now have a greater appreciation of what other software packages are capable of in the data science field.

# IST 719: Information Visualization

IST 719 expanded my skills not by teaching me a new machine learning package, but by teaching me how to create strong and eye-catching visualizations, that are organized and easy to follow. The final project for the class was a poster (Zelazny, Global Warming: How Hot Is It Really?, 2020) that was created in Adobe Illustrator, a program that I had never used before the class. What has stuck with me the most from this class and project where learning how to draw the reader’s attention in with an eye-catching visual, get them to quickly and easily understand your question or your point, and then follow-up with more detailed analysis to back up your initial point.

Another key point of IST 719 was the ethics of data science. When communicating our results, it can be very easy to make a flashy graphic that can appear to say one thing, while upon closer inspection it may actually show something else or be attempting to deceive the reader. This course impressed upon me the necessity to always present the “truth of the data”, and to not try and twist the facts of what the data is telling us to fit a narrative that we or our sponsors may be pushing for or expecting.

While presented as a visualization course, the takeaway for me was more on learning how to communicate to a wide audience, which was immediately valuable to me in my current career, and will continue to be so in the future.

# IST 722: Data Warehouse

IST 722 again taught me skills in multiple software packages, primarily SQL Server and PowerBI. In the final project (Gendron, Zelazny, Fisher, & Vanleuvan, 2020) we built a data warehouse that consolidated data from two existing SQL databases, would regularly pull and update data from both sources, and then present that data in the form of a PowerBI dashboard that can access and query the data to enable non-SQL users to gain their own insights into the data. This was a massive project, but also one of my favorites in the entire program because I truly started to appreciate the meaning of “big data” and how it can help make business decisions. It also covered the entire life-cycle of data science, from collecting and cleaning, storing and maintain accurate records, and enable actionable insights to be made by anyone with access to the platform.

# Conclusion

I came into the Applied Data Science program at Syracuse University because I saw the opportunity and the need for the advertised skills to be used at my current company, but I didn’t have the skillset to jump into a role that would use them. I initially thought that I would just learn how to feed data into different machine learning models, but I soon learned there was much more than that. The life-cycle of data, knowing where it came from, and how to store it, are just as important as knowing the different methods to process it to gain insights. Even more important, is the ability to properly communicate the results and conclusions, and to know how to present to different audiences. The message and content that you present in a technical paper will be vastly different to what you present to a CEO, and even different again to what you would share with the general public. While not covered in the sections above, an article that we read in SCM 651 Business analytics (Ferguson, 2014) regarding GE and their culture of analytics really reinforced this point for me, that a successful career in data science requires proficiency not just in the technical aspects but it the communication and presentation as well. Knowing your audience is just as important as knowing the data, and if you can’t properly communicate then no one is going to listen to you. The Applied Data Science program at Syracuse University has taught me these skills and more, and I am prepared to go forth and leverage them in my career as a systems engineer and data scientist.

# References

Ferguson, R. B. (2014, January). GE and the Culture of Analytics. *MIT Sloan Management Review*.

Gendron, C., Zelazny, A., Fisher, R., & Vanleuvan, C. (2020). *IST 722 Group 4 - Final Project.*

Zelazny, A. (2019). *IST 718 Lab 1.*

Zelazny, A. (2019). *IST 718 Lab 2.*

Zelazny, A. (2019). *IST 718 Lab 3.*

Zelazny, A. (2020). *Global Warming: How Hot Is It Really?*

Zelazny, A. (2021). *IST 707 Project: Student Performance.*

1. <https://facebook.github.io/prophet/> [↑](#endnote-ref-1)
2. <https://github.com/zalandoresearch/fashion-mnist> [↑](#endnote-ref-2)