Applied Data science project portfolio milestone

Evidence of Completion of Learning Goals

Bryan D’Amico

Fall 2023

## Program Outline and Description of Projects

In the following sections, I will briefly introduce projects I completed during each quarter of the program that I will be using as evidence towards my achievement of the learning goals.

### 1st Quarter – Summer 2022

Starting the program in the summer of 2022 marked my first real experience in learning how to leverage a programming language, R, to more efficiently apply mathematical and statistical concepts in solving problems. During this quarter I learned the basics of coding in R, how to perform statistical analyses in R, and how to tackle business problems from an analytical perspective.

#### IST 687: Introduction to Data Science

In this course, I worked on a team to analyze data from the music streaming service Spotify. The goal of this project was to use analytics to uncover what makes a song popular among Spotify users and to attempt to build a model that could use song characteristics to predict how popular a song would be.

### 2nd Quarter – Fall 2022

My second quarter marked the beginning of a deep dive into both data and advanced predictive techniques. Learning how to create a database and query data using SQL helped me understand the back-end of data engineering. Knowing the proper techniques and practices on how to store data has helped me to more efficiently work with data to produce insights. Learning about how to clean and prepare data in new ways and apply many different unsupervised and supervised machine learning techniques to that data vastly improved my array of skills and has helped me to be more creative in the ways I approach problems.

#### IST 707: Applied Machine Learning

In this course, I used a comprehensive suite of machine learning tools to analyze a data set containing the physiochemical properties of wine along with their quality ratings as assigned by professional tasting judges. I prepared and explored the data before using unsupervised and supervised machine learning techniques to predict the quality of the wine based on its characteristics.

#### IST 659: Data Administration Concepts and Database Management

In this course, I worked on a team to create a database solution that would improve the operations of a business. We created a fictional winery, named Chateau Cuse, that grows their own grapes, produces wine, and then sells the wine as an online retailer. The database was comprised of three main branches – production, sales, and HR. By using the database, the company would be able to track any bottle of wine through the entire process from when it was still grapes on a vine to the point when it was sent to a customer. Also, my team presented how this database could be linked to a dashboard which reveals important insights to business owners and key stakeholders.

### 3rd Quarter – Winter 2023

The third quarter was particularly exciting for me because it was my first opportunity to fully explore my passion of data visualizations. Learning how to combine programming with actual artistic design software was extremely rewarding and has equipped me with skills that not all data scientists have. I feel confident that I can now not only make informative visualizations, but ones that are interesting to general audiences and easy for them to understand. This quarter also marked my first foray into programming in Python. This was a much smoother transition for me than I expected, with my skills in R being very transferrable to this new language. This experience also helped me realize the advantages and disadvantages of each language related to the type of data I am working with or the task I need to perform.

#### IST 719: Information Visualization

In this course, I created an infographic poster based on a data set that recorded the global and regional sales figures of video games released over a thirty-year period. By cleaning and joining additional data to the data set, I was able to reveal important insights and patterns in a visually striking format that would garner the attention of those in the gaming industry and casual observers as well.

#### IST 652: Scripting for Data Analysis

In this course, I worked on a team to answer a set of key questions regarding movie and TV show data from a variety of streaming services such as Netflix and Hulu. The main goal of this project was to create actionable insights through data munging, visualization, and tools such as associate rule mining, for users to decide what streaming service would be the most worthwhile for them to subscribe to.

### 4th Quarter – Spring 2023

The fourth quarter was an opportunity for me to explore a related area of interest while beginning my program specialization. Since I do not have a computer science background, having the opportunity to learn about how cloud technologies work has helped me at work to better navigate how data is accessed and shared across servers and other technologies such as Git. I was also excited to start working with text data. Previous courses had almost exclusively focused on numerical data stored in neat rectangular formats, so this was a huge departure for me. Learning how to process text data using techniques such as regular expressions and then model that data in different ways has given me the skills needed to approach difficult problems that require multiple forms of data to solve.

#### IST 615: Cloud Management

In this course, I worked on a team to develop a business proposal and presentation for a cloud-based that could be used to attract investors. This product, which was named FLDR, would be a cloud-based software as a service that would allow schools to migrate all their student data to a cloud platform that would be more secure than keeping records saved locally on computers or printed out on paper. It would also provide more streamlined access to the information that teachers, administrators, nurses, and others would need to best meet the needs of their students.

#### IST 664: Natural Language Processing

In this course, I worked on a team to create a sentiment classification model for a data set of movie reviews. The main goal of this project was to compare the methods of using different feature sets along with Naïve Bayes models with alternative model types such as SVM or random forest, along with more advanced approaches such as BERT and word embeddings.

### 5th Quarter – Summer 2023

This quarter was a continuation of my program specialization in working with text data. The foundation I built in IST 664, helped me to dive more deeply into the assignments and project in IST 736. Since I was now comfortable with both processing text data and working in Python, I was able to focus on trying new things such as directly working with an API and using different types of predictive models. I was also able to expand my analyses to include topic modeling and topic recognition to uncover more insights hidden within text data.

#### IST 736: Text Mining

In this course, I used machine learning to create a game development tool for the popular trading card game Magic: The Gathering. The main goal of this project was to collect the data through an API, clean and prepare the data for modeling, then use the written text on the cards to create a classification model that could predict the color that should be assigned to the card. Additionally, I performed a topic modeling analysis to uncover different storylines in the Magic: The Gathering universe and a sentiment analysis to uncover if certain card colors contained more positive or negative text than the others.

### 6th Quarter – Fall 2023

Now, during my sixth quarter, I am currently taking IST 718: Big Data Analytics. While I am still in the middle of this course, it has already been a great opportunity to translate skills that I had only learned in R or Excel, such as modeling data with different types of regression or working with time series data, over to Python. I feel that this course will be particularly helpful for the types of tasks I complete at work and provide me with new ideas or techniques that I can share with my coworkers to help improve our current processes.

## Discussion of Learning Goals

In the following sections I will discuss my experience with each learning goal during the Applied Data Science programs and relate specific results, outcomes, and deliverables from the projects I completed as evidence towards my achievement of those learning goals.

### Learning Goal 1:

### Collect, store, and access data by identifying and leveraging applicable technologies.

Learning to work with data is the crux of this program. When I began my first classes in the summer of 2022, I had a math and statistics background, but having chosen a career as a teacher, I had not worked with data in a significant way or in a professional capacity. The Applied Data Science program has taught me how to properly collect data, whether that includes producing it myself or finding it from various sources online, that can be used to solve a business or research problem. I also learned how to store data, especially in IST 659, where I was introduced to the full process of building a database from scratch and designing it in a relational manner. Furthermore, I then learned how to access and leverage that data to produce reports, generate visualizations, and answer questions. The rest of my classes taught me how to work with data from various sources such as csv, JSON, raw text, and more. With the variety of skills I have learned, I can now confidently approach problems involving any data source, even if I have not worked with it before.

While almost all my classes focused on this learning goal, I believe the projects from IST 659 and IST 736 best demonstrate my achievement of this goal during my time in the program.

The database creation project from IST 659 did not involve collecting data, but rather creating our own data. Data was created that recorded what types of grapes were grown and harvested at different times from a vineyard, the machines used to process, store, age, and bottle the wine made from those grapes, the employees that worked on different steps in that process, and the orders that customers made to buy the wine produced. This data needed to be created in such a way that it followed the principles of design for a relational database. Once it was created, the data was stored in various tables and linked using keys so it could be accessed by analysts, and in the case of the data on customer orders, used to relate order and customer information to facilitate the purchasing and shipping process. My team also showed how this data could be brought into Power BI to create an interactive dashboard that would help the business leverage the information contained in the database. The business can investigate how much product they have in stock, track revenue, summarize employee and pay information, and uncover how well each of their wine products are selling.

The text mining project from IST 736 demonstrated my ability to collect and store data on my own for use in a machine learning analysis. All the data for this project was collected using API calls through Python to access an online database of Magic: The Gathering cards. This gave me the opportunity to import JSON data into Python, think about what data would be best to use and then extract the information on the cards that I wanted for the project. Once the data was gathered, I saved it out as a csv file. After storing the csv locally, I was able to bring it back into Python as a data frame which allowed me to more efficiently prepare the data for the analysis.

### Learning Goal 2:

### Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle.

The data science life cycle was a very influential conceptual work on how I approached a given task. I used that framework to complete many projects that spanned a range of topics. My projects from IST 707 and IST 652 showed my ability to create insights for general audiences to improve their entertainment. My projects from IST 659 and IST 687 showed how I was able to create actionable insights for a business. Finally, my project from IST 736 showed how I was able to work towards improving the efficiency of a development team.

The results of the analysis of the project from IST 652 can help a person select the streaming service that they feel would be the best use of their money. The analysis uncovered the genres of shows and movies offered by each service along with their IMDB ratings, if the releases on each service were increasing or decreasing in quality over time, if the IMDB scores could be biased in assigning higher scores to programs that are more popular in the mainstream, and if the descriptions provided by the service accurately reflect the age ratings of their programs.

The results of the analysis of the project from IST 687 were intended for a leadership team at a record company. By uncovering what makes a song popular on Spotify, the record company can better select what new artists they want to support and what types of songs they want to produce in the hopes of garnering a large audience of listeners on Spotify. Popularity on a streaming platform can lead to increased sales of concert tickets, vinyl records, and merch, which all bring more revenue to the company.

The results of the analysis of the project from IST 736 were intended to support the game development teams working on Magic: The Gathering. This card game has been around for over 30 years, with new cards having been released every year. This means that it is a difficult task both to come up with cards that do new things and to make sure they fit in with the existing cards to ensure a good player experience. By building a model that can predict the color of the card based on its text, designers are able to take more creative risks with the abilities and powers they design since they have a tool that can help them determine the best color those new powers will fit into.

### Learning Goal 3:

### Apply visualization and predictive models to help generate actionable insight.

While most people familiar with data science and analytics understand that predictive models are used to create actionable insights, visualizations are an important tool in the process as well. Instead of requiring users to comb through complex tables and numerical outputs, a data scientist can make visualizations that neatly summarize and explore different facets of the data that anyone can understand. Visualizations are a very powerful tool in uncovering patterns and trends that can lead to insights all on their own, even if a model is not developed afterwards. Creating clear and informative visualizations became one of my most passionate areas during this program and a focus of all my projects. Many of them included the development of various types of predictive models as well, but the two projects that I feel best demonstrate my success with this learning goal are from IST 719 and IST 707.

In IST 719, I combined the principles of data analytics and artistic design to create an infographic poster revealing insights in the video game industry over the course of a thirty-year period. The trends in sales in various regions around the world as they relate to the genre of game, the console the game was released for, and the year of release can be leveraged by those in the gaming industry to understand what genres are gaining and losing popularity, and in what parts of the world certain types of games tend to be more popular. For example, a small indie studio may choose to only release their game in certain regions to limit the number of languages the game needs to support. This would reduce their budget and development time, while maximizing their revenue by targeting the countries where that genre of game is most popular.

In the project for IST 707, I used a large variety of predictive models, including SVM, decision trees, random forest, Naïve Bayes, k-Nearest Neighbor, along with other techniques such as clustering and association rule mining to see if the scores given to a wine by professional judges could be predicted by the wine’s physiochemical properties. The insights from this project could be used by wine manufacturers to tune their recipes to give them the best chance of getting high scores from tasting judges. This project gave me the opportunity to compare the results of many different models, while considering their accuracy metrics, their interpretability, and their efficiency. When ranking the models, I had to decide whether it would be more important to focus on getting just the highest accuracy possible, or if it would be better to use a model that more clearly outlines why a wine would be given a certain score.

### Learning Goal 4:

### Use programming languages such as R and Python to support the generation of actionable insight.

Upon beginning the Applied Data Science program, I had only very little coding experience. I did not learn how to code during my undergraduate studies or my previous career in education, so when preparing to start at Syracuse I completed some introductory projects and tutorials in R to learn how to do simple tasks such as creating a data frame, making some visualizations, and writing functions. Now that I am at the end of the program, I can use both R and Python to read and write data, create complex and interactive visualizations, filter, aggregate, and manipulate data frames, process raw text data, and create, train, and test predictive models.

During the first half of the program, my classes focused on providing me a strong foundation in R. I learned to work with data using both the base R functions along with the dplyr functions in the tidyverse package. I visualized the data using ggplot2 and created interactive plots with R Shiny. I also learned how to run inference tests, create predictive models such as generalized linear models, and run machine learning analyses. Many times, my final output was created as an R Markdown file.

My project for IST 687 was coded in R and involved data exploration techniques, visualization, statistical inference tests, and a basic R Shiny App. My project for IST 707 was done entirely in R, including importing the data, data preparation, and modeling. The report was generated in R Markdown. The basis for my project for IST 719 was created in R. Before creating the actual infographic poster using Adobe Illustrator, all the visualizations were created in R using the tidyverse package with ggplot2.

The second half of the program involved a shift to learning how to code in Python as well. Since I was now already familiar with the main principles of data science, I was able to focus on working with different types of data, such as JSON and messy text data, creating new types of visualizations using packages like seaborn, and employing new techniques such as those found in the field of natural language processing.

The project for IST 652 was completed in Python using Jupyter notebooks. This allowed me to combine my code and results into a single file where the insights directly flowed from the visualizations, calculations, and tables. While my team’s project for IST 664 involved using other tools such as Hugging Face, a large portion of the project was completed in Python. This project gave me the opportunity to clean messy text data, turn it into feature sets, and use those feature sets to create insights from predictive models. My project for IST 736 was completed entirely in Python and allowed me to better understand how to import, store, clean, and leverage data all within a Jupyter notebook.

### Learning Goal 5:

### Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsor and technical team leads).

Throughout my classes, I was able to produce visualizations and analytics to communicate insights to a variety of audiences. In particular, one of the main evaluative criteria for the project for IST 719 was to create informative visualizations that would capture the attention of anyone who might see it. It was designed to be seen both digitally and in person as a printed poster at any distance. Principles of artistic design such as color schemes, layout, fonts, contrast, and accessibility were combined to draw viewers in closer to learn more. They would first notice the large main graphic that communicated the main information on the project, then walk closer to be able to read the title and see the accompanying plots, and then finally approach the poster fully to be able to read the fine details about the story the data was telling.

All my projects included a presentation component that was designed for a particular audience. Projects such as the ones for IST 664 and 707 were crafted for more technical audiences. These included discussions of data processing, model development and results, and strategies used to make predictions. Conversely, projects such as the ones for IST 736 and 659 were for general audiences and focused on the implications of the results of the projects, business impacts, and how end users would benefit from what was developed.

The audience for the project from IST 687 was the C-suite of a record company and my team’s presentation was intended to show them how data analytics could lead them to choosing the next top artist and song. The project from IST 652 also had to do with developing insights based on entertainment media, but this time the results were intended for anyone in the general public who was interested in learning more about the streaming services available to them. The analytics and questions answered would help them save money by understanding how to choose just one or two services that would give them the most enjoyment, instead of feeling compelled to subscribe to them all.

### Learning Goal 6:

### Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

Throughout my time in the program, I learned the importance of applying ethics to data science. Models must not be chosen purely on raw accuracy metrics, but to fit the needs of the situation they are trying to model. Part of the responsibility of the data scientist is to understand where bias might come into play when developing a model and take steps to avoid it. Building a model that predicts on features such as race or gender could be inappropriate or even illegal given the circumstance. The type of model used also contributes to how fair and transparent the story that the data scientist will craft from their analyses will be. A model such as a decision tree is clear and interpretable, but an SVM model, for example, operates more like a black box. This would make it very difficult to explain to business partners who are using it to make decisions, and also increases the risk of the model generating conclusions that could be biased against groups of people depending on what data were used as inputs.

In the project for IST 659, we applied the principles of protecting sensitive data by designing our database to hold that information separately from company information and sales orders. In the project for IST 652, we directly addressed any underlying biases in our data sources by exploring the relationship between overall IMDB scores and the number of users who contribute to the scores. We wanted to make sure that the shows and movies that were identified as having the best quality, were not just experiencing inflated scores due to so many people having watched them. In IST 615, our product design and proposal involved leveraging cloud technology to store sensitive information about students more securely. This included personally identifiable information, transcripts, standardized test scores, and medical information. Our product would allow teachers and school staff to access and update this information without relying on riskier methods such as having it on loose papers that could be lost or stolen or having it sent through less secure electronic sources like email.

A few of my projects also had a focus on fairness and transparency in model development. In my IST 707 project, I weighed the positives and negatives of using a black box model versus a transparent model to achieve the best predictive results possible. In the project for IST 664, we explored model results not just in terms of accuracy, but also with measures of precision, recall, and F1 score to fully understand the strengths and weaknesses of each model and its associated set of input features.

Based on the types of projects and assignments I completed in the program, I did not have too much opportunity to work with data that contained personally identifiable information or sensitive variables. From a conceptual standpoint, I learned the best practices and cautions that I should take, but this is an area that I am looking forward to pursuing more in the future to continue to build my skillset as a data scientist.

## Influential Conceptual Works

The data science life cycle and OSEMiN, Bayesian thinking, Nathan Yau’s Visualize This, Edward Tufte’s works on data visualization, and the current research being performed in natural language understanding and generation are all works that have inspired the way that I perform data science.

The data science life cycle was the guiding framework for how I approached all the projects I completed during this program. Following the cycle allowed me to plan and organize each task I would need to complete during a project and helped me to keep on schedule and meet all my deadlines. I am particularly fond of the cycle because of the flexibility it encourages when moving between stages. Data science is not a linear process. When discoveries are made in the data cleaning and exploration stages, they could require one to take a step backwards and collect more data. Also, after creating several predictive models, the insights of those results could be used to perform different types of feature engineering techniques on the variables before making another batch of models. The open-mindedness that the life cycle encourages helped me to stay motivated during my projects and not get discouraged if things didn’t work out perfectly on the first try. I knew it was natural for even the most experienced data scientists to move back and forth between the stages as they got a better handle on the data. In my final quarter, I am now learning about the OSEMiN framework and see it as a useful supplement to the thinking promoted by the data science life cycle. This framework has fewer distinct steps and I feel that it is very useful for organizing tasks into different buckets to create a more accurate timeline of completion dates to keep things moving along smoothly, while accounting for those times when I will need to go back to previous steps in order to truly move forward.

In my first quarter I was formally exposed to the full extent of Bayesian thinking for the first time. Previously, I was very familiar with Bayes’ formula through classes I took on probability, but I did not realize there was a whole movement of reasoning that was breaking away from the traditional statistical inference tests. While I do not discount the insights that can be gleaned from those traditional inference tests, the process of using prior probabilities to inform posterior probabilities felt more natural to me. I also found it rewarding to create and consider more outputs than just a t-test. In IST 772 I learned how to leverage the visualizations of a highest density interval, craft more meaningful explanations of differences, or the lack of, between distributions, and consider more types of test statistics such as the Bayes Factor generated from an ANOVA analysis. This course made statistics more exciting for me to delve into than in the past and this is something I would like to continue to learn more about in the future.

There are several conceptual works related to visualizations by Nathan Yau and Edward Tufte that are very influential on how I strive to represent data in powerful and intuitive ways. Yau’s Visualize This contains a lot of the inspiration behind how I crafted my final project for IST 719. His book helped me to recognize different ways to use visualizations and how many distinct pieces of information can be combined into a graphic that is still easy to understand for all audiences. Now, when I approach visualizations, I think about selecting the best color scheme, plot space, font, text size, shapes, and labels for the data. Previously, I do not think I would have focused on something like choosing a font that matches the tone and message of the data, but I now firmly believe that any visualizations I create are more engaging if I do. I have also begun looking into some of Tufte’s works such as Beautiful Evidence and Envisioning Information. Tufte examines almost every facet of information visualization in great detail. The examples he provides span a vast range of topics and involve many different types of data. Learning more about the history of information visualization, how viewers experience seeing visualizations for the first time and the adjustments that can be made to improve that experience, and how graphics of all different types and sizes can be extremely impactful towards making the understanding of data more achievable by all. These works will continue to provide me with more things to consider and more strategies to employ when visualizing the data I am working with.

While taking the two classes in my chosen specialty track, Text Mining and Natural Language Processing, I was exposed to the current research and applications of natural text understanding and generation. In particular, I was inspired by the research and development being performed by companies such as Microsoft and Apple, who are using this technology to create a more inclusive world for people with disabilities. As one example, both companies are working on apps that allow a phone’s camera to read a document or interpret a scene occurring in real life and auditorily relay that information to a user. Most people are familiar with how companies will use data science in various capacities to improve profits, but not everyone is aware of how data science can be used to improve the human experience. Accessibility has been important to me for many years, influenced by my time working as a teacher in special education classes, but now I am inspired by these types of applications and would love to find the ability to incorporate these types of data science applications at work. Developing these kinds of life-changing models would be an extremely rewarding experience and is something that I hope I am able to pursue in the future.

## Areas of Strength

Now that I am at the end of the Applied Data Science program, I can identify many areas of strength in relation to the field. Although I previously had almost no experience with coding, I can now point to that as one of my professional skills. I feel that I have a very strong foundation in both R and Python, and I am confident completing tasks in either language. This program has also equipped me to be able to find, understand, and apply documentation related to either language so that I am able to successfully tackle new tasks with my coding abilities. As a recent example, I did not take any courses that focused on R Shiny, but I was able to use the available documentation and resources to make a working app for a task I was working on for my job.

I have also learned a variety of techniques related to data cleaning and preparation. With these techniques I am comfortable approaching various types of structured and unstructured data regardless of how they are stored. I am confident applying best practices in terms of cleaning and preparing data as it applies to a particular scenario. There is no one right way to impute missing values or transform variables for use in a model, but I now understand the underlying conceptual factors that allow me to diagnose the situation and make appropriate adjustments to the data.

Another area of strength is related to visualizations. Throughout all the courses, and especially in IST 719, I made visualizations a primary component of my assignments and projects. I believe that visualizing data is almost always the best way to present it to an audience, so I challenged myself to create not only informative graphs, but interesting and appealing ones as well. The design principles I learned in IST 719 now allow me to make visualizations that are clear to technical and non-technical audiences alike, and employ appropriate color schemes and contrast to accommodate viewers with visual disabilities. I have recently begun incorporating interactivity into my visualizations, as well using tools such as R Shiny and Python packages like plotly express.

Throughout this program, the majority of my classes focused on using different machine learning techniques to build models that could solve a given business problem. I am now confident that I can choose the appropriate model type based on the requirements of task and the level of interpretability required. I know the proper techniques to prepare data for various models, such as performing transformations or using strategies such as principal component analysis, to both improve the performance of a model and increase its efficiency. Throughout my courses I also learned the many ways that each model can be tuned to try to get the best performance possible given the data I am working with. Understanding the strengths and weaknesses of the models allows me to leverage them to their maximum potential. Combining insights from unsupervised and supervised models can also lead to a more comprehensive understanding of the data and identify areas of improvement in successive model runs.

After completing the courses in text mining and natural language processing, I now considering working with text data one of my strengths as well. I can collect, organize, process, and leverage raw text to discover insights, such as through topic modeling or parts of speech tagging, and create predictive models. I see text data as an exciting opportunity and will continue to look for creative ways to incorporate it with other data to provide additional information. I feel that these skills are a competitive advantage as not all data scientists and analysts are able to work as comfortably with text data.

## Areas of Challenge and Moving Forward

I have also identified some areas of challenge that I will be continuing to work on after finishing this program. While I had many opportunities to develop visualizations during the program, I did not frequently work with dedicated dashboarding tools. In the future, I would like to be very proficient using these tools at my job to improve my team’s ability to easily investigate the data and see the results of any analyses that are completed. I feel that having extensive knowledge in relation to preparing dashboards is a very important skill for a data scientist to have. Stakeholders and business partners are likely to prefer examining results in this type of interactive and visual setting rather than reading a technical report or viewing a slide deck. Dashboards also give them the ability to explore the data on their own even if they do not know a programming language.

While my classes heavily focused on advanced machine learning models such as SVM and random forest, there was not as much of a focus on employing more traditional models such as generalized linear models. Presently, I do not feel as comfortable deciding on all the best parameters and variables to use in such as model as I do with the machine learning models. In my current job, I am going to be mainly working with generalized linear models of different families and link functions combined with techniques such as gradient boosting. I feel that this experience will help me to continue developing as a data scientist. Understanding what types of models are best for my field of work while keeping current with advancements in the overall field of data science will help me to identify what new techniques could be incorporated to improve our business processes.

Another area of challenge arises when I need to incorporate techniques that are more skewed towards computer science or those that involve other types of programming languages that I do not have any experience with. After the program, I plan on becoming more familiar with HTML and web site design to improve my ability to leverage web scraping packages. This knowledge will also help me in designing more advanced applications that are accessed through a web page such as those made in R Shiny. Now that I will be working in data science in a more collaborative fashion, on a large team of coworkers, instead of in a small group of classmates, I need to become more familiar with using R Studio and Jupyter on a server instead of on my local hardware. This includes creating and using virtual environments, using terminal commands, and sharing and receiving work through Git. These concepts are mostly unfamiliar to me, and while I am gaining some confidence with them, I would like to become very proficient so I am able to independently troubleshoot any issues that occur or even help others when they encounter a problem.

As part of my continuing journey as a life-long learner I will continue to learn more about different applications and techniques within the field of data science. I am planning on taking self-paced courses in Tableau, Power BI, SQL, and cloud platforms like AWS to expand my current toolset. In addition to this, I will also be getting involved in my company’s training programs as a teaching assistant before transitioning to hosting trainings as an instructor. I believe that teaching others is one of the best ways to continue learning. Teaching the content will expose me to new questions and scenarios I have not thought of before that will deepen my own understanding of the content. This will also mean that I will stay very current with new developments and will have to think critically about how to incorporate them into existing business practices to improve performance and productivity.

## conclusion

Looking back at my time during the Applied Data Science program has been very rewarding. When I first started, I felt anxious about changing careers and wondered if I would be successful in this program. The classes I took and the projects I completed have helped me develop into a confident data scientist starting a new professional journey. I am very grateful for all I have learned over the last year and a half, and I am very proud of all the new professional skills I have acquired. I have come to develop a great passion for data science and look forward to continuing to learn even more. Being able to work with data allows me to learn new things about almost any topic, and with the skills I have learned during this program, I can now craft a story from what I learn and share that with others.