**Portfolio Milestone 1:**

**Written Report**

Author: Connor Gendron

NetID: cgendron

Student Number: 96559-2353

Project Data: <https://github.com/cjgendro/Connor-Gendron-Data-Science-Portfolio>

**Table of Contents**

1. Introduction ------------------------------------------------------------------------------ 3
2. IST 707: Data Analytics ---------------------------------------------------------------- 4
   1. Project Overview ------------------------------------------------------------------ 4
   2. Satisfied Learning Goals --------------------------------------------------------- 4
3. IST 736: Text Mining ------------------------------------------------------------------- 5
   1. Project Overview ---------------------------------------------------------------- 5
   2. Satisfied Learning Goals --------------------------------------------------------- 5
4. IST 718: Big Data Analytics. ------------------------------------------------------------- 6
   1. Project Overview ----------------------------------------------------------------- 6
   2. Satisfied Learning Goals -------------------------------------------------------- 7
5. IST 659: Data Admin Concepts & Database Management ----------------------------- 7
   1. Project Overview ------------------------------------------------------------------ 7
   2. Satisfied Learning Goals --------------------------------------------------------- 8
6. Conclusions --------------------------------------------------------------------------------- 8

**Introduction**

The Applied Data Science Master’s program at Syracuse University rigorously prepares its students for a future in the field through underlying theories in addition to hands on work. The program introduces students to a variety of data science disciplines and the appropriate tools for each job. In the past eighteen months working through my classes I have gotten exposure to several tools for the first time, such as Access, SQL, Python, R, Idea, and Weka, while honing my skills in more familiar tools like Excel. In addition to working with these tools in the iSchool classes, the dual nature of the degree also has students make sense of data in a business sense through the Whitman School of Management coursework.

Particular courses such as Data Analytics (IST 707), Big Data Analytics (IST 718), Text Mining (IST 736), and Database Admin Concepts & Management (IST 659) all demonstrated different methods for exploring, analyzing, and explaining data in varying forms and degrees of complexity. The skills learned I in these courses best exemplify the learning objectives for the Applied Data Science degree and have prepared me to look at new information as a Data Scientist.

Completing the Applied Data Science degree requires all candidates successfully demonstrate their mastery of the 7 core learning objectives. Using projects from the previously outlined classes I will highlight particular examples of the following learning objectives:

* Describe a broad overview of the major practice areas in data science
* Collect and organize data
* Identify data via visualization, statistical analysis, and data mining
* Develop alternative strategies based on the data
* Develop a plan of action to implement the business decisions derived from the analyses
* Demonstrate communication skills regarding and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
* Synthesize the ethical dimensions of data science practice(e.g. privacy)

**IST 707: Data Analytics**

1. Project Overview: Fantasy Football Analysis  
     
    Data Analytics was primarily an R based class with a focus on following the whole process of data science, from collection, to cleaning, visualization, mining, and analytics. Over the quarter we gradually learned how to complete these tasks in R with the end goal being to display our proficiency with a final project. We were tasked with finding a data set to work with, generate some insight from our analysis, and present our findings.  
    My group decided to work with a football statistics data set, it was decided that weekly fantasy football would be an interesting and useful application for the data. Initially a lot of cleaning and general restructuring needed to be done for the data to be useful, certain positions needed to be grouped together while others had to be dropped from the analysis altogether. After the data was cleaned and organized correctly, we were able to translate the raw statistics into fantasy football points and explore the dataset more.  
    The final analysis attempted to provide fantasy point predictions for players by using techniques including support vector machine, naïve bayes, and decision trees. In terms of descriptive analysis clustering and association rule mining were conducted to find relationships within the data.
2. Satisfied Learning Goals:  
   1. **Identify data via visualization, statistical analysis, and data mining:**Throughout the process in this project data visualizations were used for exploring data, highlighting interesting findings and finally in expressing analytic results. Statistical analysis was crucial to many of the insights we generated and informed decisions made with more complex analytics. Finally, association rule mining among other data mining techniques were used to identify noteworthy trends/relationships within the data.
   2. **Demonstrate communication skills regarding and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization:**Part of the final project was an in-class presentation of the most important findings that our group made. During the presentation we were able to effectively express useful insight in a manner that could be equally valuable to managers and programmers. Using a combination of visualizations and relevant technical information we were able to get our message across.

**IST 736: Text Mining**

1. Project Overview: Presidential Nomination Speeches  
     
    Text mining is becoming increasingly important as greater amounts of new text content is being generated every day, each with potentially valuable information. The goal of the class was to develop the necessary python skills to conduct useful text mining, and analysis. Our final project was to find a text collection, clean, mine, conduct analysis, and present our findings.  
    We chose a dataset of presidential nomination speeches for each major party from 1900-2016. After finding the data we had to clean it using regular expression methods, stopwords, frequency counts, etc. Some analyses conducted included LDA topic modeling, hierarchical and k-means clusters. Adding labels also allowed us to conduct svm analysis for predicting things like whether or not the speaker won the presidency.
2. Satisfied Learning Goals:  
   1. **Develop alternative strategies based on the data:**Early on in the data gathering phase we were planning to focus our project on whether or not text analysis on nomination speeches could be used to predict presidential elections. Later on we expanded our predictions to include different labels such as presidential party, poll status and economic indicators. In the end our most promising results came when predicting economic indicators like the unemployment rate.
   2. **Ethical dimensions of data science practice:**  
      When working with political data it is important to avoid biases and present the data objectively. To this end we wanted to have as much representation from each party as possible by using nomination speeches for both major parties. Furthermore, party affiliation was not considered in most computer analyses. Finally, the results were presented through an apolitical lens where the focus was on success rates for our predictive analyses and other interesting finds.

**IST 718: Big Data Analytics**

1. Project Overview: Kaggle NFL Big Data Bowl  
      
    Big Data Analytics builds off the foundational skills learned in the Data Analytics course by applying them to more complicated datasets. Big data is defined by the volume, velocity, and variety of the data which can present unique challenges. The final project was to find a sufficiently “Big” dataset and use the methods we had learned in python to generate actionable information.  
    Our group decided to work on the NFL big data bowl project, where the goal is to predict the yards gained on any given play. The original data set was difficult for the computer to interpret so we generated new features and restructured the data to make our analyses. Then we attempted to predict yards gained using several algorithms including: Naïve Bayes, Neural Net, and Random Forest.
2. Satisfied Learning Goals:  
   1. **Describe a broad overview of the major practice areas in data science:**  
       Most of the projects for this program involved many of the major practice areas within data science, however, this project required the most intense application of them. The data engineering was extensive, we had to reformat the entire dataset, and generate new features from essentially the raw components. The exploratory analysis and some of our most useful insights came from our data mining and statistical analysis. Our data mining results were incorporated into an interactive python application intended to aid in play calling strategy. Using data visualization techniques, we were able to capture and express our most important findings in impactful ways.
   2. **Develop a plan of action to implement the business decisions derived from the analyses:**  
       After the initial results it was clear we needed a way to implement our findings in a practical way for them to have any noteworthy impact. We were able to combine the results from our best performing algorithm with insights we gained while data mining to create an interactive python application. The idea of the algorithm is to provide the stakeholder, in this case a football coach, instantaneous feedback and play suggestions given their current circumstances. The implementation of this application takes the best of our insights and can give relevant actionable information to the stakeholder in real time, responding to real criteria.

**IST 659: Data Admin Concepts & Database Management**

1. Project Overview: Video Game Database  
      
    The Data Admin Concepts & Database Management program covered the basic theories and SQL skills necessary to curate a relational database. The course gradually built up from establishing conceptual data models up until ultimately implementing and populating the database. Our final project was to come up with our own dataset, complex enough to justify the relational database format.  
    My dataset was compiled from a video game collection split into multiple attributes such as genre and publisher. Each game’s data was collected by hand and coded into the database using SQL script. In the end queries were generated using the relational model to derive useful insights on the collection such as which year accounts for the most games in the collection.
2. Satisfied Learning Goals:   
   1. **Collect and organize data:** Most of the data for this project had to be classified, organized and copied over by hand. The raw data, however, came in multiple media as some games had physical copies while others only existed digitally. The collection process involved manually scanning game cases and digital descriptions for the appropriate data points, additionally online resources were used to find unlisted information. Once the underlying data was gathered it had to be organized to fit the relational model. This database format allows for the most complete control of the data collection and organization process, including the datatype, data rules, and format necessary for data entry.

**Conclusions:**

This portfolio has been a collection of data science projects that best exemplify the 7 Core learning objective for the Applied Data Science program. Throughout the document specific examples are cited during each project where a particular learning objective was satisfied. These are the most demonstrable examples, but the core learning objectives were integral parts of other classes, assignments, and projects. The 7 Core objectives were layered into every component of the data science degree, and necessary skills for the candidates.

Now reaching the end of my program I have seen a broad overview of data science fields, including data engineering, data mining, database management, and data visualization, all of which have been implemented in one project or another. The data collection and organization for Data Admin Concepts & Database Management was particularly in depth but data collection and organization were necessary for almost all classes in the program. Data visualization, statistics and data mining were also a necessity for much of the program, being extensively used in projects including the Fantasy Football project for Data Analytics. Throughout the program there were many times when the data was not equipped to answer your questions, this requires a data scientist to rethink what it is they want to glean from the data. When our Text Mining project wasn’t yielding interesting results, we were able to come up with new questions, changing the direction of our project.

The Applied Data Science curriculum was focused on teaching data science alongside relevant business applications from the Whitman School. This style informed the idea that our projects should set out to solve “business problems” and required candidates to consider what is most useful to the stakeholders. In our Big Data Analytics project, we set out to solve a football problem where the stakeholder is a coach who needs to know which play to run. We implemented our findings by creating an interactive app that can give play call suggestions to our stakeholder. The program also places a large emphasis on presentations that test communication skills, each class requires a presentation of some sort that shows how effectively candidates can express their findings to people of various professional backgrounds. Lastly, the ethical dimensions of data science are not only important to the program but are increasingly becoming pressing issues in the world writ large. Considering the political biases of nomination speeches for the Text Mining project was only one of many times where the program has confronted me with ethical considerations. Whether it is privacy, bias, exclusivity, etc. data scientists need to be cognizant of these potential issues when moving forward with any project.

This portfolio shows that the Applied Data Science program through Syracuse University prepares its candidates to meet and master the 7 outlined core learning objectives. Throughout the program each learning objective is applied repeatedly to ensure responsible data practices, from collection through to presentation and even implementation. These methods foster an environment that equips new data scientists with the skills necessary to make their own findings and express them in meaningful ways.