**Syracuse University, School of Information Studies**

**M.S., Applied Data Science**

LMBALL@SYR.EDU

SUID: 244835217

Portfolio Milestone

Leland M. Ball

Table of Contents

[Resume 6](#_Toc67060088)

[Introduction 8](#_Toc67060089)

[Learning Goals 8](#_Toc67060090)

[Data Collection and Organization 8](#_Toc67060091)

[Data analysis and Visualization 10](#_Toc67060092)

[Strategy and decisions 12](#_Toc67060093)

[Implementation 13](#_Toc67060094)

[Conclusion 15](#_Toc67060095)

# Resume

# Introduction

This portfolio milestone marks the culmination of my learning and demonstrates my mastery of the key concepts that underpin the field of Data Science. Broadly described, Applied Data Science is a field of several facets, involving a combination of in-depth understanding of and ability to work with data, the application of statistics to thoroughly inspect information, and subject expertise not just in the data domain, but also in communications and ethics. These key ingredients are all required in order to provide results and predictions that are meaningful.

**Major practice areas**

Because of the modern ubiquity of data and computers, the field of Data Science can be applied in a myriad of ways, in almost any industry. Just a few of these relevant areas include: data warehousing, data mining, statistical analysis, distributed computing, business intelligence, machine learning, data visualization, finance and market related industries, and healthcare. While accomplishing this program, I have focused on Data Engineering and Data Mining aspects, without sacrificing attention elsewhere.

# Learning Goals

## Data Collection and Organization

Data collection is one of the first steps in the data science process. My achievement of this learning goal is demonstrated in several project deliverables from web scraping, to accessing publically available government APIs, to extracting data from storage formats in mobile phones. While every application of data science begins with data, it is often the case in an academic context that the data can be easily obtained or provided to the student. This data may be more academic in nature, easier to import and parse, and with fewer anomalies to address. The selected projects exemplify the messy and unexpected nature of data collection and how I was able to use what I have learned to accomplish the goals of data collection. .

**Achieving this learning objective**

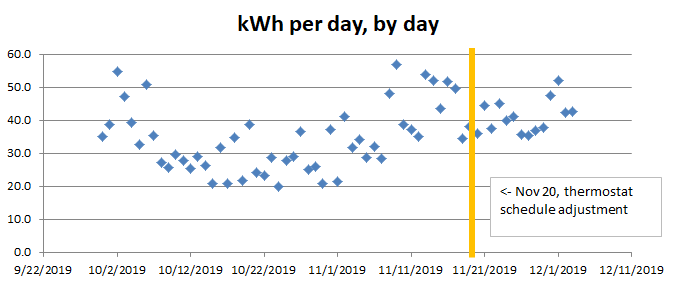
The very foundation of these projects is the data that was collected using tools, data formats, and repositories. Collecting data from these sources is a common and indisputably necessary step in the Data Science process. Python and R were used to acquire and format the data. Of particular interest is a side-project that I was able to incorporate into my electricity analysis project.. My side project kept track of the comings and goings of mobile phones within my house to an accuracy of a minute. My house has three residents besides myself, which resulted in a glut of data that I logged for some months in preparation to explore for significant effects. The electricity usage project required me to learn how to incorporate and combine multiple datasets (mobile phone presence, weather, electricity usage, and so on) into a cohesive product. Later analysis of this data reflected the impact one roommate had on electricity consumption through his manual use of his air conditioner in his room during the summer months.

Figure – Home Electricity Usage before and after process adjustments

The coursework in the Applied Data Science program also prepared me to interface with more well-formed sources of data. In my Police Incidents project, I interfaced with publicly available county government Geographic Information System (GIS) data and collections of police incident reports to visualize and analyze crimes that are reported in my county. The collected data was combined with other geo-relevant features like ZIP Code GIS shapes to filter and enrich this data, thereby showing another rather complex instance of data collection.

Figure -- Arlington police incidents

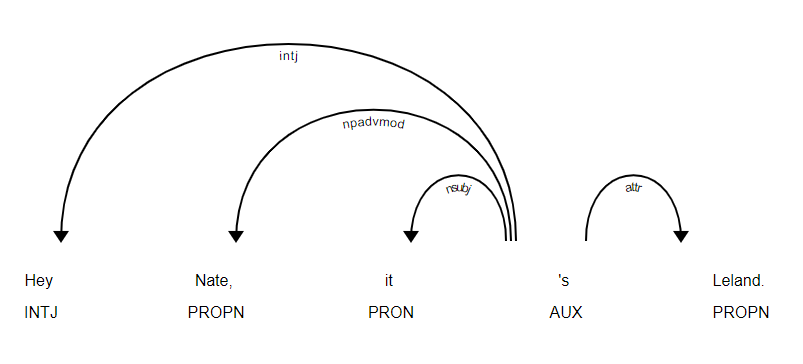
While the aforementioned projects existed largely in the realms of numeric data, another instance of data collection brought on by my projects in my degree program is exemplified in my natural language processing project to accomplish named entity resolution (NER) in text messages. This project sought to determine the efficacy of the commonly-used Spacey NER Python library against a baseline dictionary of proper nouns. The text messages were collected from my mobile phone and imported into a format that could be interacted with in Python, which involved transformation from different XML formats into a data frame shape.

Figure – Sentence parse tree for N.E.R. in text messages

**My chosen specialty**

My specialty in applied data science rests on the ability to collect and manipulate data. This necessitates acquiring the data from various sources including databases, APIs, and standard file formats, cleaning and formatting each variable to then gain an understanding of just what information is present in order to then provide actionable insights.

**Leveraging these studies in a program of life-long learning**

Going forward, what I have learned about data collection, cleaning, and importing will be employed in every data science project that I begin. These studies are proof that I am no stranger to this step of the data science workflow. The backbone of any good data science endeavor begins with obtaining data, scrubbing data, and exploring data for outliers, relevant features, and interesting relationships. I will continue to leverage my studies in my future work in the field by employing the techniques I’ve learned and demonstrated in these academic projects upon real-world problems.

## Data analysis and Visualization

Once data has been imported into an information system, it can then be analyzed. This step of the data science process includes statistical analysis, data visualization, and data mining. My ability to perform analysis on data is clearly demonstrated in the following projects: US Vaccine Analysis, Hotel Recommendations, and Hiring a Football Coach for Syracuse. In these academic endeavors, I showed the results of my learning in the employment and interpretation of statistical methods and other analysis techniques to gain further insight into the data. The techniques that I have learned in the data science program have given me the tools I need to ascertain the reality of what the data appears to be saying, and determine if there is any significance to those findings. From developing visualizations that help identify possible patterns, to the Frequentist and Bayesian approaches that verify whether such patterns are more than just appearances, I have demonstrably achieved this learning objective.

**Achieving this learning objective**

The projects mentioned above demonstrate that I can identify patterns in the data, perform visual analyses, and mine data for further insight in the following ways:

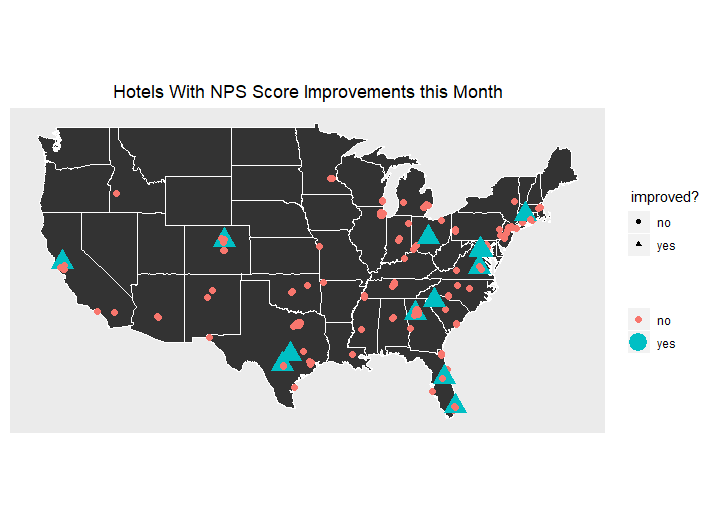
In the hotel recommendations report, my team took sanitized data from a nation-wide hotel chain and delivered observations and recommendations that C-Suite executives could use for the purposes of increasing revenue and driving enhanced customer engagement. Part of this work leveraged my understanding of statistics to identify which hotels were actual outliers when it came to performance. By employing these methods instead of a naïve sorting of all hotels, my team was able to give a more accurate picture of hotel performance.

Figure – Hotels showiing Net Promoter Score increases

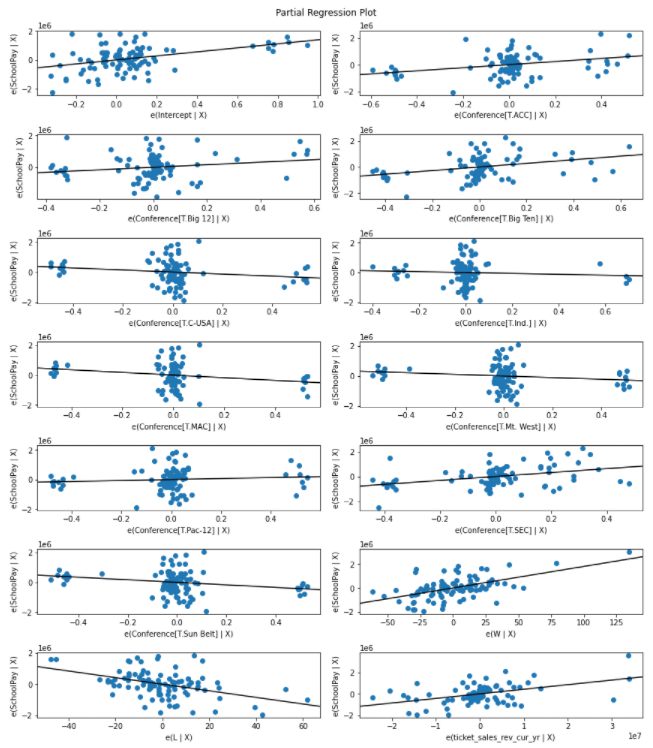
Finally, in my project to select a theoretical new football coach for Syracuse University, I demonstrated the achievement of the analysis learning objective by identifying and incorporating datasets relevant to the determination of a college football coach’s salary, developing visuals for triaging the importance of each variable among many, and employing statistics to truly determine if a given variable was actually relevant to a coach’s performance. The end result was a well-analyzed report for university staff to begin the selection of potential candidates, including what factors are most important to consider, and what salary range they should expect to pay.

Figure -- partial regressors for variables affecting coach salaries

**My chosen specialty**

An essential component to analysis in the field of data science is the ability to communicate findings to the intended audience. These projects are each representative of communication of my or team’s findings to persons with varying levels of background, with concerns ranging from the technical to the practical. Chiefly, the two categories that each of these projects display are my ability to analyze data and come to supportable conclusions, and then present such findings to persons of different backgrounds, adapting the technical details into language that is audience-appropriate.

**Leveraging these studies in a program of life-long learning**

Analysis will form an important part of my data science work going forward. The content I have learned over the course of this program will be a reference to guide my future work. The statistical knowledge I have learned has already been of great use when reading publications on the topic. Three grand categories of applicability stand out to me when considering how my studies will influence the rest of my life: The language of analysis, the impetus to measure progress (via baselines, objective and subjective model comparisons, as well as various accuracy measurements), and the steps of the analytical process. From a practical standpoint, this curriculum has given me practice in communicating the nuances of somewhat complicated topics to different audiences, a skill that will only be called upon more and more over time.

## Strategy and decisions

Data analysis benefits no one if the analysis is not published to the right people. Those people will ideally use the analysis of the data to come to an informed decision or to further develop a larger strategy. If those strategies are supported by the data, it can be best to suggest a course of action directly. During my studies at Syracuse I have produced the following deliverables that shows my accomplishment in the learning objective of developing alternative strategies based on the data: Electricity Usage and Hotel Recommendations. These projects highlight my ability to collect data, analyze it, and suggest strategies based on that analysis.

**Achieving this learning objective**

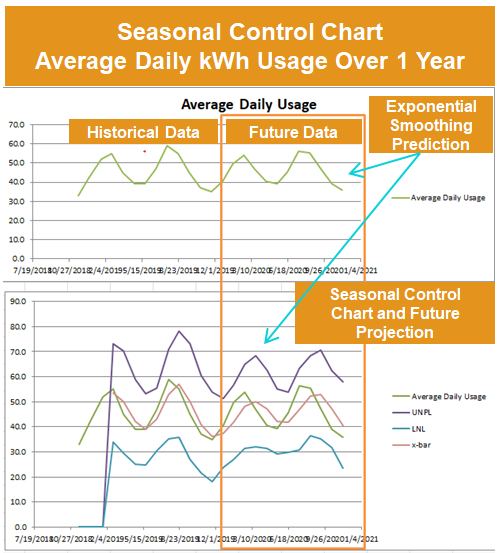
In my electricity usage project I analyzed my home electricity usage, taking into account the amount of electricity my household was using on an hourly basis, in addition to any changes made to the thermostat schedule which were indicative of a broken automated process. While the process changes put in place during the project did not result in statistically significant energy usage, I was able to measure an improvement to the process, and develop a process control chart that took seasonality into account to allow for future determination of process success at any time.

Figure -- Process control chart with seasonality

In addition, the hotel recommendations put forth in my team’s report also details actionable insights based upon data analysis, which aim to take a national hotel chain and increase the Net Promoter Score of their hotels. Our analysis used regression models to identify relevant variables that affected what guests liked about their visits, and from that analysis our recommended strategies included pursuing positive avenues to engage customers, and also suggested negative avenues that sounded promising, but that the data did not support pursuing further.

**My chosen specialty**

Without an attempt at action, data science will remain only academic. By using data analysis to shape strategies and inform decisions, I will bring the knowledge that I have gained from this program into the world in a real way. These deliverables are an example of real effects from actual data.

**Leveraging these studies in a program of life-long learning**

My studies have prepared me by focusing my mindset on the problem at hand. I appreciate the academic origins that are the genesis of data science, but my desire has been to more directly affect the world with the work that I do. To that end, I now have practice employing techniques that lead to actionable steps, and the knowledge necessary to tie data analysis to strategy again and again.

## Implementation

Long-term strategy is an important component of many organizations and successful endeavors, but strategy-in-the-moment is also needed when realizing overarching goals. In the realm of such implementation, I have a number of projects completed during my studies at Syracuse that show my ability to leverage data to develop plans of action, and implement business decisions in a sound way that is based upon a thorough understanding of the data. The projects that best display this focus are the US Vaccine Analysis, and Hiring a Football Coach for Syracuse.

**Achieving this learning objective**

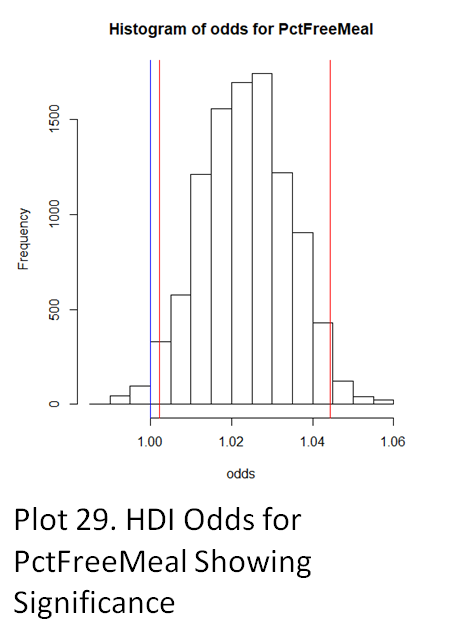
In the US vaccine analysis report, I employed both Frequentist and Bayesian statistical methods to thoroughly explore three datasets related to vaccinations of kindergarteners in the USA and specifically California. This coursework was the culmination of several courses in this program that I have taken that largely focused on statistics. In this paper, I was able to identify trends and statistical relationships between variables. This paper was written to a level that statistically knowledgeable persons could follow, but featured accurate conclusions that personnel in management positions could understand and act on. The conclusions detailed further research opportunities, but also recommended specific groups of schools to fund, in order to take more immediate effect.

Figure -- Highest Density Interval log-odds for variable showing significance

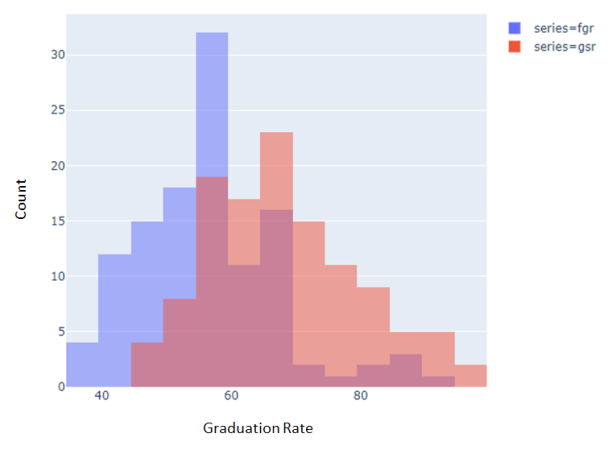
Finally, in my paper on hiring a football coach for Syracuse University, I was able to identify the variables most relevant to a competitive team (and by extension, their coach). I was able to extrapolate out from a given coach’s performance, conference the team played in, and revenue-generating histories of the team to create a model for more accurate salary starting points. The work produced a very actionable set of salary recommendations that would otherwise be hard to come by.

Figure -- Two methods of measuring graduation rate compared- further investigation showed no significance to coach salaries

**My chosen specialty**

The implementation phase of the data science process is especially important because it is in this phase that more actionable steps are recommended and undertaken. The steps leading up to implementation (data collection, data analysis, and strategizing) set the ground work for data science to affect the world. For this reason, these projects exemplify my ability to take data through this entire process and deliver a well-informed implementation plan.

**Leveraging these studies in a program of life-long learning**

When calling upon my data science skills in the future, I will be looking forward to the implementation stage from the beginning. My studies in this program have taught me that looking ahead to a possible final product is incredibly valuable. Not every problem will need a custom-built neural network, and many problems are not solved by an interactive model at all, but by careful consideration of the implementation that satisfies the need, the best solution can be arrived at. My experience problem-solving from the ground up in this degree program has given me practice in implementing needed solutions that I aim to apply to future learning throughout my life.

# Conclusion

This portfolio has demonstrated my achievement in the learning objectives outlined in the Masters of Science in Applied Data Science curriculum. Concerning the collection and organization of data from disparate locations, formats, and shapes (MBC-638, IST-652, IST-644, IST-718). Concerning identification of patterns in data via visualization, statistical analysis, and data mining using Frequentist and Bayesian statistical techniques, regression, and generalized linear models, model comparison techniques, and understanding of various accuracy measurements as seen in (IST-687, IST-772, IST-718). Relating to developing alternative strategies based on the data (MBC-638, IST-687, IST-718). Furthermore, this portfolio also addresses the learning objectives to develop the implementation of business decisions derived from analysis (MBC-638, IST-772, IST-718).

Throughout each deliverable a clear demonstration of communication skills regarding data and its analysis for their relevant audiences, be they managers or IT professionals, programmers or statisticians can be seen especially in these deliverables: (IST-772, IST-718). Finally, my understanding of the ethical dimensions of data science can be seen in how I addressed data privacy concerns in the withholding and sanitization of private information before use in some projects (IST-664).