Portfolio Milestone Project

Allison R. Deming

SUID: 394248454

Syracuse University

Master of Science in Applied Data Science

Spring 2024

(Insert Link Here to Github)

Portfolio Milestone Project

Table of Contents

Introduction

Data science, an emergent discipline, amalgamates insights from three foundational domains: statistics, computer science, and individual areas of expertise. This expertise can span diverse fields, showcasing the synthesis of statistical methodologies and computational techniques to augment knowledge within that domain. For instance, my background in cognitive experimental psychology, though unconventional in data science, has been instrumental in shaping my trajectory. After a fulfilling yet challenging tenure as a Psychology Professor spanning two decades, I embarked on a career transition driven by a profound interest in data science. This shift allowed me to reinvigorate the statistical acumen I cultivated during my Ph.D. program while also mastering advanced programming languages beyond my proficiency with Excel and SPSS.

My professional endeavors revolve around a Higher Education Resource Services (HRSA) grant and student services, ventures seemingly distant from data science. Nevertheless, data analysis remains a cornerstone of my responsibilities with the grant, and I firmly believe that integrating data science methodologies into student services can yield substantial benefits. Post-grant conclusion and degree attainment, my aspirations converge on securing a role that seamlessly merges my newfound data science prowess with the pedagogical expertise accrued during my academic tenure, fostering student success through data-informed initiatives.

My journey led me to choose the Syracuse Master of Applied Data Science program offered by Syracuse, drawn by its comprehensive curriculum tailored to equip students with what they need for the workforce. The program's objectives encompass a spectrum of proficiencies, including but not limited to:

- Providing a comprehensive overview of key *practice domains* in data science.

- Proficient *data collection* and organization techniques.

- Discerning patterns through data visualization, *statistical analysis*, and data mining.

- Crafting *alternative strategies* grounded in data insights.

- Formulating actionable plans to actualize *business decisions* derived from analyses.

- Effectively *communicating* data insights to diverse stakeholders within organizations.

- *Ethically* navigating the complexities of data science practice.

This portfolio is a testament to the culmination of skills honed over the past two years within the iSchool's data science program. Spanning diverse projects encompassing programming tasks in R and Python, utilizing Excel, and exploring ethical dimensions through traditional papers, these endeavors encapsulate my unique journey and skillset, emblematic of my experience within the program.

Project Descriptions

There were many courses that I could have chosen for this portfolio. Narrowing it down was quite hard and was a learning experience in and of itself. I picked five classes I felt would best show what I have learned over the past two years. They also showcase various means and methods, programming languages, and writing skills. The first course that I picked was IST 618. This course in information policy is more of a non-applied course with papers and presentations, along with one debate that helped me grow my communication skills. I also learned about many ethical dilemmas in data analysis, and how those affect information policy. The first paper was on access and affordability; the second paper was on "Big Tech," the third paper was on the right to repair one's technological devices such as cell phones, and the last paper was on cell

phone tracking technology along with other GPS tracking technologies and the ethics behind them.

The subsequent course I enrolled in was IST 652, an introductory Python course. One notable aspect of this class was a group project centered on analyzing Airbnb data, where we explored three distinct hypotheses. Moreover, we incorporated an additional data set sourced from the city of New York, which provided crime statistics, enabling us to conduct a comprehensive analysis. By amalgamating these datasets, we examined various aspects, including the frequency of terms used in listing descriptions, pricing trends, and the correlation with crime statistics.

(It's worth mentioning that this portfolio features two distinct projects utilizing Airbnb data. The richness and versatility of this dataset allowed for its utilization across multiple projects without redundancy in analysis or code overlap. Prior clearance from instructors confirmed that the reuse of the same topic was permissible if the datasets differed in the time frame and analytical approach. While this may limit the breadth of showcased variety, the consideration of future course requirements wasn't immediately apparent during enrollment. Notably, the first project (IST 652) entailed a concise Python analysis and concluded with a comprehensive, multifaceted analysis executed in R (IST 707).)

My next selection for representation from my tenure here was IST 687, Introduction to Data Science. This course delved into data exploration and analysis utilizing the versatile R programming language. For this project the agile method was used along with a Kanban board always to track the project's status. This was a new process for me, which I found very enlightening. Our project focused on the pivotal task of forecasting school placement, homing in on the factors that influence employment prospects for junior and high school students. Leveraging the wealth of data available within the student database, we embarked on an investigative journey to unearth post-graduation job placement status determinants. Through

applying fundamental R programming techniques, we unearthed actionable insights and furnished recommendations to optimize academic programs and enhance student success initiatives. This endeavor facilitated a deeper understanding of data science methodologies and underscored the transformative potential of data-driven decision-making in shaping educational outcomes.

IST 707 presented a formidable challenge for my fourth endeavor compared to its predecessors. In this Data Analytics course, we explored the Airbnb dataset comprehensively, employing the R programming language as our tool of choice. Unlike previous projects, the objectives here were distinctly business-oriented and exploratory, elevating our analysis's complexity.

Our project aimed to uncover intricate relationships within the dataset, spanning five key areas of inquiry. Initially, we sought to discern the impact of various amenities on the pricing of Airbnb listings. Subsequently, we analyzed the correlation between review sentiments, proximity to major tourist attractions, and Airbnb pricing, necessitating minor Python coding and integrating Google Maps API for geographical insights.

Furthermore, we investigated the seasonal variability in Airbnb pricing and reviewed frequency in New York City, probing into the influence of time on these metrics. Lastly, we endeavored to identify the geographical distribution of affordable Airbnb listings across the five boroughs of New York City, alongside predicting their availability across different seasons.

This multifaceted project not only stretched our analytical capabilities but also underscored the practical relevance of data analytics in informing strategic business decisions and forecasting trends within dynamic markets like the hospitality industry. Visualization is key when explaining complicated statistical concepts to those at all levels of knowledge. For our

analysis, we used decision trees, K-means clustering, Sentiment analysis, and multiple other basic methods for visualization of the big and small data.

For my culminating project, I opted for IST 718, delving into Big Data Analytics. Rather than centering my focus solely on the larger group project at the course's conclusion, I was drawn to the inaugural homework assignment, which entailed an intriguing analysis: assessing the potential earnings of the Syracuse Football Coach if they were to coach in alternative conferences, specifically the SEC or ACC.

This endeavor demanded the amalgamation of multiple datasets from various online repositories, facilitating a comprehensive comparison of recent compensation packages offered to college football coaches across different leagues, factoring in similar levels of experience and tenure. Subsequently, the analysis expanded to encompass a broader examination, probing into metrics such as win rates, participation in bowl games, and academic success.

By meticulously scrutinizing graduation rates among football program students, we sought to glean insights into the multifaceted dynamics influencing coaching salaries and performance metrics across diverse collegiate football landscapes. This nuanced exploration underscored the intricacies of data analytics and shed light on the multifaceted considerations shaping the collegiate athletics ecosystem.

Collecting Data
Analyzing Data
Alternative Strategies with Data
Communication
Ethical Dimensions

In data science, ensuring privacy remained a paramount priority throughout the project. Specifically, adherence to the General Data Protection Regulation (GDPR) was a primary concern for our teams (IST 618, IST 687), particularly given my role in handling sensitive user

data within my current position at a university. Throughout the program, I deliberately avoided analyses or datasets containing sensitive data (IST 652, IST 687, IST 707, IST 718), such as medical information, personally identifiable data, or financials. This proactive approach was implemented at the outset of each project to preemptively sidestep the necessity of scrubbing sensitive information from datasets. This resulted in a lack of instances wherein such actions were required. It's crucial to acknowledge that while this approach may not always be feasible in real-world scenarios, it underscores the necessity of vigilance in safeguarding the privacy and integrity of individuals potentially impacted by our research endeavors.

Conclusion

The compilation presented herein attests to the proficient attainment of Syracuse University's Applied Data Science program's seven learning objectives while also showcasing a nuanced comprehension of key practice domains within the field of data science, namely business analysis, computer science, and data analysis (IST 718). Data procurement entailed the utilization of diverse sources, including Kaggle and open-source databases, subsequently structured and processed utilizing either R or Python (IST 652, IST 687, IST 707, IST 718). Through machine learning algorithms, visual representations, and statistical modeling techniques, discernible patterns within the dataset were elucidated (IST 718, IST 707). Subsequent analyses culminated in formulating alternative strategies and actionable plans, with a. emphasis on the business implications and option of running an Airbnb in New York City (IST 718). Syracuse University's School of Information Studies empowers students to cultivate essential competencies in data analysis, business analysis, computer science, and communication, equipping them to furnish actionable insights to diverse audiences.

The Applied Data Science program places paramount importance on bridging the divide between business stakeholders and IT experts, a pivotal aspect in tackling data-related challenges, addressing business needs, and enhancing operational effectiveness. Furthermore,

the program fosters a reflective consideration of the ethical dimensions inherent in data management and analysis, recognizing the volume of data in our world. It underscores the imperative for data scientists to exercise vigilance in safeguarding the privacy of individuals and organizations while also ensuring that the data utilized in analysis remains unbiased and representative. I want to thank the faculty, student services staff, and technology support employees for their help throughout this program. I know that without their help, I would not have made it to the point that I am at where I can have the chance to showcase my work and learning.