Final Portfolio Milestone

Caitlin Rookey

SU ID: 333629736

<https://github.com/cadyannn/portfolio>

Table of Contents

[Introduction 1](#_Toc1512607431)

[IST 659: Database Administration 1](#_Toc1378754050)

[Project Description 1](#_Toc474463552)

[Learning Outcome 1](#_Toc1092744023)

[IST 707: Applied Machine Learning 1](#_Toc865528164)

[Project Description 2](#_Toc1677264323)

[Learning Outcome 2](#_Toc2142084844)

[IST 652: Scripting for Data Analysis 2](#_Toc1650759492)

[Project Description 2](#_Toc968517849)

[Learning Outcome 2](#_Toc1025267421)

[IST 623: Introduction to Information Security 2](#_Toc1443139294)

[Project Description 2](#_Toc1792241577)

[Learning Outcome 2](#_Toc965032045)

[Conclusion 2](#_Toc203393366)

# Introduction

I entered the Syracuse University Master's in Applied Data Science program in Fall 2022. My background is in Library Science, and I was looking for a change that suited my love of data and data management. The School of Information’s Data Pipelines & Platforms track within the program intrigued me, and I have found pursuing this degree incredibly fulfilling.

The Applied Data Science program at Syracuse University provides students the opportunity to collect, manage, and analyze data from multiple sources using various tools and methods.

The School of Information’s Applied Data Science Program learning objectives:

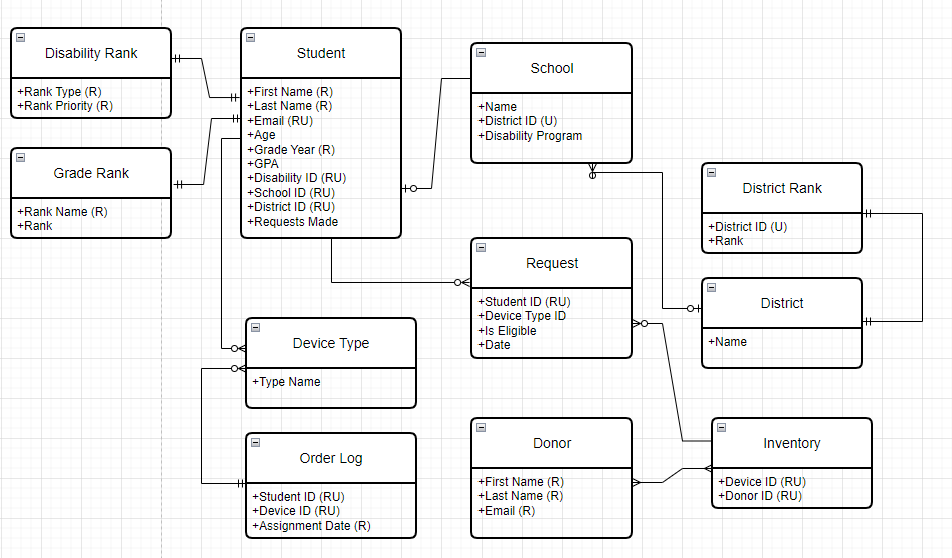
* Collect, store, and access data by identifying and leveraging applicable technologies.
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle.
* Apply visualization and predictive models to help generate actionable insight.
* Use programming languages such as R and Python to support the generation of actionable insight.
* Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).
* Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

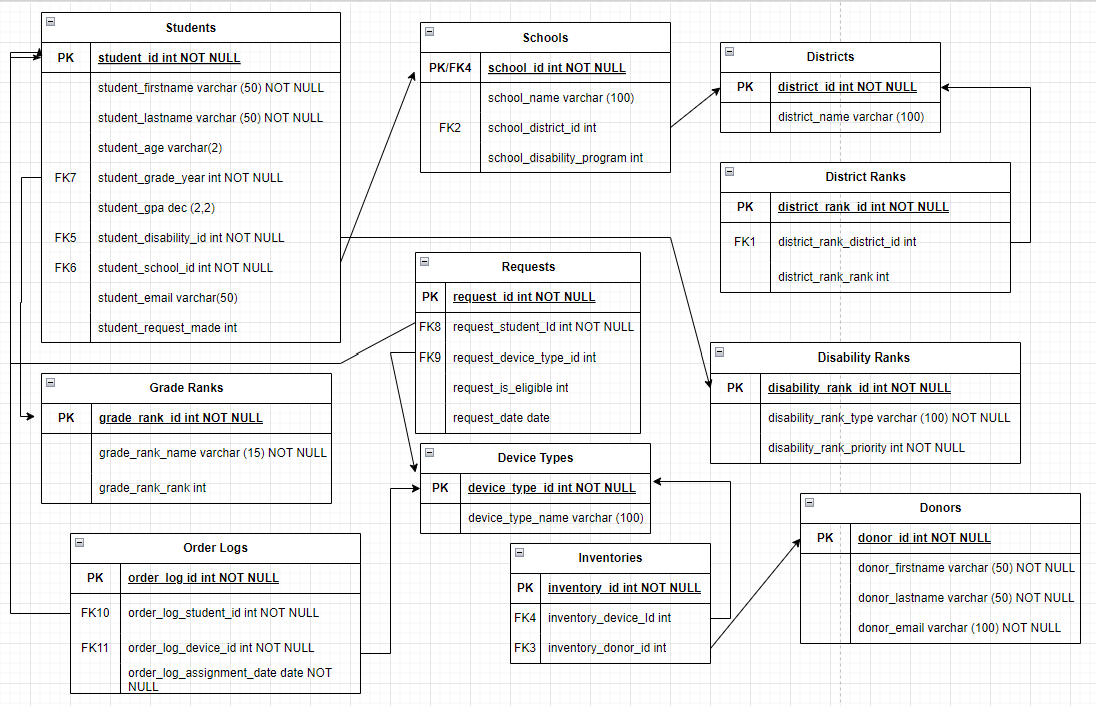
# IST 659: Database Administration

## Project Description

My final project team’s fictitious foundation, TechForAll, solicits technology donations for students across a city. The devices are distributed by criteria determined by the foundation’s board of directors. The CompU database tracks the requests made by students to receive a device from the charity and the inventory of devices waiting to be distributed. (Rookey, “IST 659,” 2023).

The team delegated roles and responsibilities based on our strengths. I developed the models and assisted with SQL trigger generation. These tasks were challenging and required multiple meetings to ensure our work aligned with the model. We communicated often to apply updates to the model when errors or unexpected changes occurred. The models were created using Draw.io and SQL was written in Microsoft SQL Server Management Studio.

**Fig. 1: Conceptual Model (Rookey, ”IST 659,” 2023).**

  
**Fig. 2: Logical Model (Rookey, ”IST 659,” 2023).**

## Learning Outcomes

The course and project contributed to the successful application of the learning outcome “collect, store, and access data by identifying and leveraging applicable technologies”. We used the technology provided in the course to develop a data management system that stored and delivered data to the user. Building the system gave us a view of where data scientists retrieve data within an organization. The proper organization and implementation directly impact the analysis of the data.

Completing the course and final project encouraged me to pursue the Data Pipelines and Platforms track. Skills obtained in this course are currently being utilized in IST 722: Data Warehousing, where I am utilizing SQL knowledge from IST 659 on the basic troubleshooting for our ETL process buildouts.

# IST 707: Applied Machine Learning

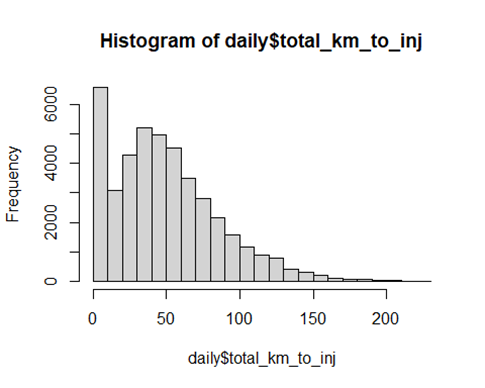
## Project Description

Our final project attempted to use Garmin running data to predict injuries in professional runners. Our motivation was to try to use some of the techniques we learned in class to predict injuries and optimize training for runners based on their training and injury history. As we progressed through the project, we narrowed our scope to one specific problem: determining if a variable in the data impacted overuse injury.

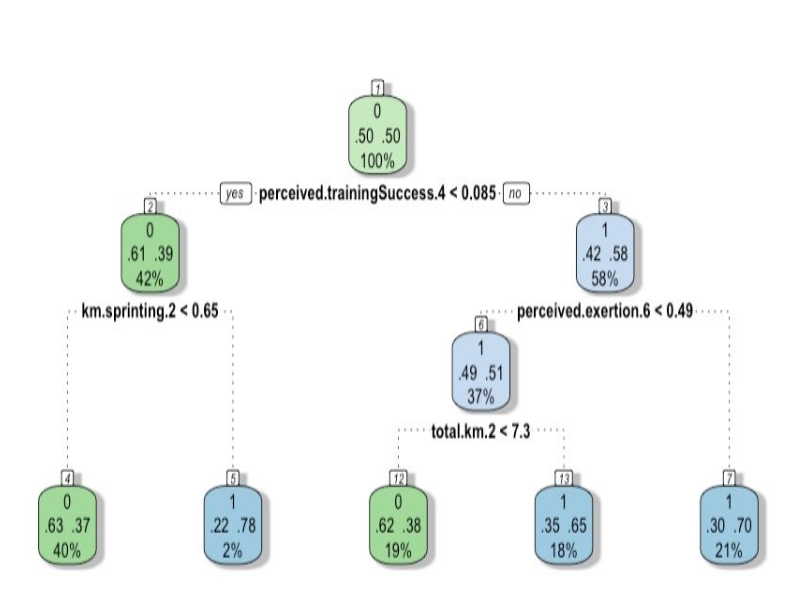
The Kaggle dataset provided a daily and weekly dataset from professional runners' Garmin and journal entries. We decided to use the daily approach as this approach gave more accurate results in the original experiment. The dimensions for this dataset: 42,766 records and 73 attributes. All attributes are numeric except for Athlete ID, injury, and date.

Each day consists of 10 attributes: Number of sessions [0, 2]: training sessions the athlete completed that day. The dataset became unwieldy as the project unfolded but gave us great troubleshooting experience. We struggled to normalize and balance the data, which impacted the success of our predictive models.

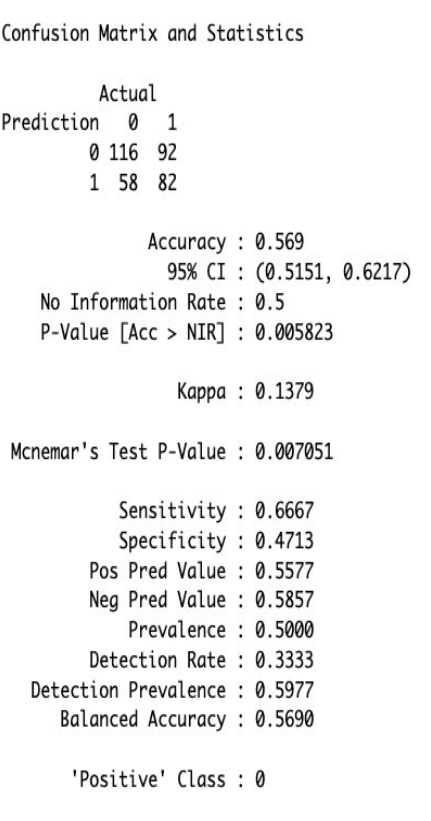
The project required utilizing various data mining techniques to perform regression, classification, and clustering analysis. The dataset required cleaning, preprocessing, and many attempts to normalize the data for more efficient results. After several attempts we found a sample size that was balanced enough to run the predictive models: Naive Bayes, Random Forest, and SVM.



**Fig. 3: Histogram of kilometers run daily before injury (Rookey, ”IST 707,” 2023).**



**Fig. 4: Decision Tree Model (Rookey, ”IST 707,” 2023).**



**Fig. 5: Decision Tree Confusion Matrix (Rookey, ”IST 707,” 2023).**

During the semester, my partner and I conducted an immense amount of research, as we were both relatively new to the program. This was my second term in the program, so I was still familiarizing myself with R. By the end of the course, I felt exponentially more confident in my skills with the language.

This project was far from a success, but the experience and skills it gave me are invaluable. Dr. Bolton was an incredible professor, who explained the concepts well and provided great feedback and guidance. In the end, this dataset was just too challenging, and we could have garnered results by utilizing a neural network, which was not part of the syllabus.

## Learning Outcomes

This project was unique, because I had spent the previous term working with more business-centered data and analysis. Being able to choose a unique dataset and problem, which I was interested in, further engaged me in the learning process. In that way I believe I fulfilled the objective to “create actionable insight across a range of contexts, using data and the full data science life cycle.”

Though the project did not solve the problem statement, we were able to “apply visualization and predictive models to help generate actionable insight” which was that further analysis was needed with more advanced techniques. The course honed my skills using R and introduced me to valuable applied machine learning techniques. It also challenged my troubleshooting on a large project, as we adjusted our scope and attempted to wrangle the data.

# IST 652: Scripting for Data Analysis

## Project Description

Our final project group proposed to analyze various Airbnb properties offered throughout New York City from the dataset available at Kaggle (<https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>). We identified neighborhoods with the highest rate of return for potential property managers to invest in as well as the lowest priced neighborhoods for future customers looking to rent a property. We used multiple measures of analysis as well as additional datasets such as New York City citywide crime statistics that we found relevant to rental prices in neighborhoods as reported in the New York City Open Data site (<https://opendata.cityofnewyork.us/>).

Our results showed that the correlation between Airbnb listing prices and the rate of criminal complaints in each neighborhood is a multifaceted and complex relationship. While some degree of correlation may exist, it is important to recognize that numerous factors can influence both Airbnb prices and crime rates independently. Neighborhoods with higher crime rates may, in some cases, experience lower Airbnb prices due to decreased demand for short-term rentals. Conversely, neighborhoods with lower crime rates might attract more visitors, leading to higher Airbnb prices. However, it is essential to exercise caution when drawing direct causal relationships, as numerous other socioeconomic, cultural, and environmental factors can also impact both Airbnb pricing and criminal activity within a neighborhood. A comprehensive analysis would require a nuanced examination of various contributing variables to gain a more accurate understanding of this correlation.



**Fig. 6: Violin Plot of Price Distribution in NYC Boroughs (Rookey, ”IST 652,” 2023).**

## Learning Outcomes

This final project enhanced my then-novice Python skills and explored packages for visualizing our findings. I felt well prepared to enter the Big Data Analytics course the following term, to continue to “use programming languages such as R and Python to support the generation of actionable insight.”

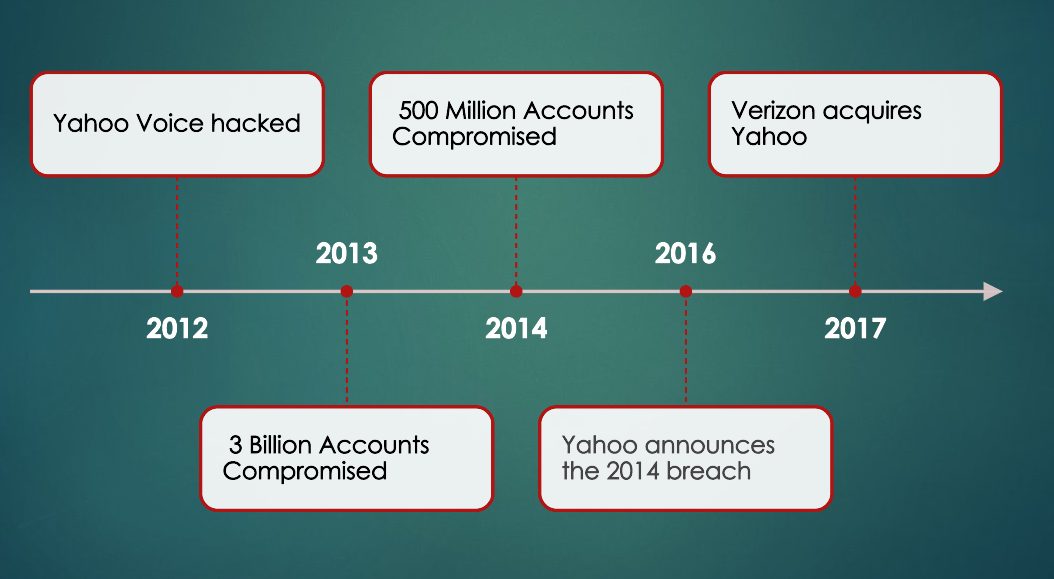
This project was my first project transitioning from coding in R to Python. After researching many packages, such as Pandas and Seaborn, I generated many of the visuals for our presentation. This involved research and troubleshooting, but in the end, I was able to elevate the bar graphs we started with to a more interesting visualization of price distribution. I was also able to learn from my teammates, who had a versatile background and experience using Python to generate heat maps.

# IST 623: Introduction to Information Security

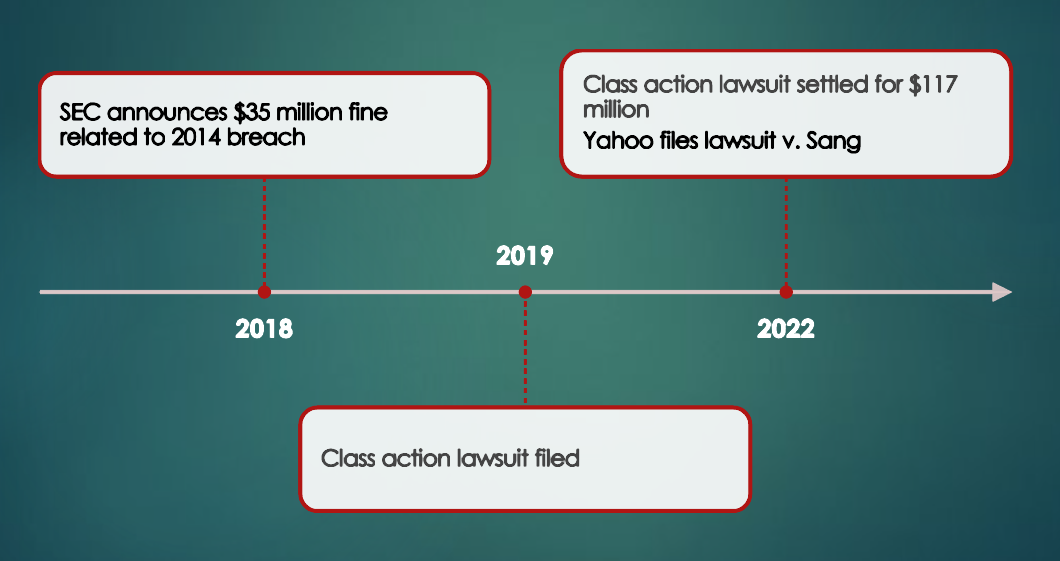
## Project Description

The final project for IST 623 required our group to research a data breach of our choosing. Our team agreed to breakdown the multiple breaches that occurred at Yahoo starting in 2012. Their questionable data handling and security practices made for a robust and engaging project. Our research involved untangling numerous breaches at Yahoo over multiple years and how it impacted the company and users. There was significant crossover with the Information Policy course I completed the previous semester. I was able to more critically review the court filings and SEC actions and wonder how more proactive policy and regulation might have prevented the events.

We used timelines to illustrate the breaches, response, and aftermath.



**Fig. 6: Yahoo Data Breach Timeline (Rookey, ”IST 623,” 2024).**



**Fig. 7: Yahoo Data Breach Timeline cont. (Rookey, ”IST 623,” 2024).**

## Learning Outcome

The Introduction to Information Security and Introduction to Information Policy courses have directly impacted the actions I take and the decisions I make in my current role as Assistant Director of Information Management at a college. Protecting data throughout its lifecycle in an organization is imperative to ensure its quality and privacy.

Applying ethics in the development, use and evaluation of data and predictive models protects the data and the organization from legal consequences. These ethical data practices contribute to the long-term sustainability of data-driven technology and culture.

# Conclusion

Syracuse University’s School of Information gives students the opportunity to integrate data collection, management, and analysis. Skills learned in the program have given me the knowledge to deliver actionable insights using various techniques and nurtured strategies to improve organizational efficiency. As the program draws to a close, I find myself equipped with the skills to address a wide range of problems and the resources to explain observations to diverse audience of stakeholders.