Waylon Abernathy

Portfolio Milestone

SUID: 330621920

Email: wabernat@syr.edu

Paper & GitHub Table of Contents:

Goal: Collect and Organize Data Folders (Pg. 3)

[IST 718: Zip code Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Zipcode%20Analysis)

[IST 687: Basketball MVP – Applied Data Science](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_687_Basketball%20MVP)

[IST 659: LOVIT Ultra Runner database creation – Database Management](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_659_LOVIT%20Ultra%20SQL)

Goal: Identify patterns in data via visualization, statistical analysis and data mining (Pg. 4)

[IST 718: Zip code Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Zipcode%20Analysis)

[IST 718: Coaches Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Coach%20Analysis)

Goal: Develop alternative strategies based on the data (Pg. 7)

[IST 707: Mental Health Project – Data Analytics](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_707_Mental%20Health%20Project)

Goal: Develop a plan of action to implement the business decisions derived from the analyses (Pg. 7)

[IST 707: Mental Health Project – Data Analytics](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_707_Mental%20Health%20Project)

Goal: Communication skills regarding data and its analysis for managers, IT, etc. (Pg. 8)

[MBC 638: Final Project – MBC Data Analysis and Decision Making](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/MBC_638_Final%20Project)

[IST 707: Mental Health Project – Data Analytics](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_707_Mental%20Health%20Project)

Goal: Synthesize the ethical dimensions of data science practice (Pg. 9)

[Privacy and the Public in the Technological Age – Information Policy](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_618_PrivacyPaper)

Each of the folder links are matched with the goals they satisfy. The written paper also includes the hyperlinks for the folders for easy navigation. Each folder contains the final submission for the professors of each course as well as the Python or R script. Only Python or R are the software required to view/run the projects. My professional resume is on the GitHub repository and has been included at the end of this document.

**Introduction / Broad Overview of Practice Areas in Data Science**

I began this program with zero technical skills and little knowledge about what it takes to be a data scientist in practice. I was only introduced to the field when artificial intelligence was deployed in the emergency department where I had worked as a nurse. I quickly realized that pursuing data science would not only allow me to make a positive impact on patients’ lives, but I could also create change within the struggling system of healthcare. The field of data science has quickly begun to shape our way of life through the use of data collection, statistical analysis, and machine learning/artificial intelligence. Each of these three areas can build upon and improve the other, and lacking knowledge and skill in one area tends to diminish the performance and impact of another. The program learning goals made by Syracuse University are designed to ensure that each student has exposure and practice in these major areas.

**Collect and organize data**

Data collection and organization should be regarded as the foundation of all data analysis. Data scientists can create new algorithms with great predictability power or image recognition; however, these new advances are lost on poor data. Data collection and organization requires the data scientist to carefully understand and review a dataset before diving into any project. Without proper review, the final product may be inaccurate, skewed, or simply incorrect. Outliers, missing data, and/or small data sizes can inaccurately represent the true story that is being told by the data. Below are three projects I have completed in this program that demonstrate the proper use of techniques in data collection and organization in three different languages: Python, R, and SQL.

[Zip code Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Zipcode%20Analysis)

This Big Data project involved collecting data in addition to the provided data on United States zip codes with the goal of choosing the best zip codes at which to live. The original data started with average home values from 1996 to 2017. It was possible to simply take the average home value increase by zip code to see what area had the greatest growth; however, there are many other factors that I wanted to consider when choosing a zip code: job growth, crime rate, income-growth rate, poverty, and wellness/life expectancy. I was able to accomplish this by pulling data from Wallethub’s listing of best places to live in the US. This allowed me to filter down the data to five potential zip codes in San Jose, Austin, Sioux Falls, San Diego, and Charleston. Finally, I used Facebook’s Prophet predictor to predict where the housing market value would be in 2018 to pick my zip codes. This project was very challenging due to the very large amount of data that needed to be collected and organized in a way to deliver meaningful results. The professor challenged us to pick from the entire U.S. listing of zip codes rather than filter down by a chosen state. I believe this project was a good example of how to use outside resources to narrow the scope on large datasets.

[Basketball MVP – Applied Data Science](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_687_Basketball%20MVP)

This project utilized 2017-2018 NBA statistics to find trends and meaningful variables to understand what separates players receiving MVP votes from the average NBP basketball player. To do so I utilized data from the basketball reference website to import into an excel document. This document was imported into R and data manipulation was required to add MVP votes received to the receiving players. Data cleaning was also required in way of adding columns and changing column names. This project would not have been possible without proper data collection and organization to reveal what it takes to be an MVP.

[LOVIT Ultra Runner database creation – Database Management](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_659_LOVIT%20Ultra%20SQL)

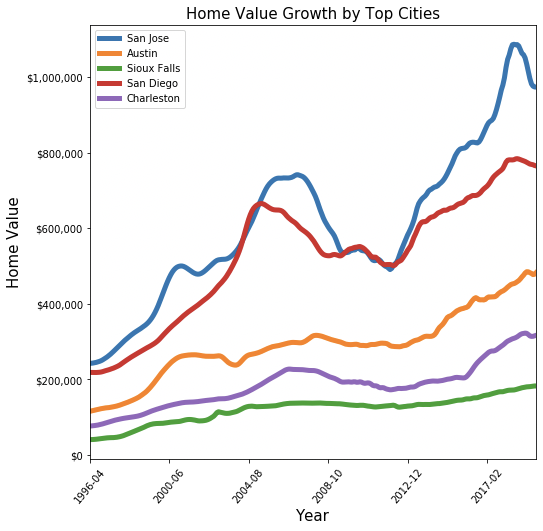
This project utilized SQL to build a database for the 100-mile Lake Ouachita Vista Trail Endurance Race. This project was especially challenging due to the fact that I needed to manually enter each runner’s information from 2015-2017 in order to build a proper database in SQL. The particularly difficult aspect of this project was finding a way to race times into a format that SQL would recognize – this task would have been no issue; however, these 100-mile races last longer than 24 hours and SQL’s time input won’t recognize anything over 24 hours. A great deal of coding was required in order to create a function that would organize these values properly.

**Identify patterns in data via visualization, statistical analysis and data mining**

Pattern identification is key to informing what story the data is telling. Visualization, statistical analysis, and data mining are all tools we can use to read that story. These techniques can also help identify anomalies in the data, and if the data has been properly cleaned for further analysis. Utilizing these skills can also help narrow the scope of a project by displaying which variables can play a role in a given outcome. The Zip code analysis project shows how important data visualization and mining can be in narrowing the scope of a dataset, while the Coaches project shines light on how different variables can help or hinder statistical analysis and prediction.

[Zip code Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Zipcode%20Analysis)

Visualization was a key component in understanding how home value growth has changed over time. It also helped reveal which cities may be a little more insulated from economic downturn in the future. Below, we see that San Diego and San Jose were hit quite hard in 2008 where Sioux Falls has remained on a slight, but steady increase through time. Visuals like this can help future home investors decide how much risk they are willing to accept when purchasing real estate.



[Coaches Analysis – Big Data](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_718_Coach%20Analysis)

The coach project took salaries from every D1 NCAA football coach to predict what a Syracuse coach’s salary might be based on variables such as stadium capacity, win-loss record, conference, etc. A simple boxplot was used to display the differences in pay one might expect when comparing conferences across the NCAA. We can quickly interpret pay differences among conferences (especially the SEC) visually.

A screenshot of a video game

Description automatically generated

I used a heatmap to understand what variables might be correlated with coach salary in order to help build my linear regression models.

A picture containing checker, object

Description automatically generated

Using this information, I was able to build a model using win percentage (Pct.), stadium capacity, and conference to run a model with a r-squared value of 0.807. This led to a predicted salary raise for the Syracuse head coach. My Jupyter notebook shows all of the results as well as a few models that did not work so well.

**Develop alternative strategies based on the data**

**&**

**Develop a plan of action to implement the business decisions derived from the analyses**

This program has taught me many things, but maybe one of the most beneficial is to not approach data with any preconceived outcomes. Sometimes we as data scientist want to tell the data where to go when we should simply let the data speak for itself. Forcing questions or answers on to data without physically altering the data rarely works. The mental health project completed for IST 707 is a clear example of how my group changed course based on what the data was giving us.

Using data to derive and inform business decisions are the reason why the field of data science exists. There would be no reason to collect and interpret data if there was not some positive benefit. Again, the mental health project is an example of how survey data can inform business of the importance of mental health in the workplace and the effect on job performance.

[Mental Health Project – Data Analytics](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_707_Mental%20Health%20Project)

This project utilized the dataset from “Open Sourcing Mental Illness, LTD” via Kaggle to further understand and possibly predict risk factors for tech/IT workers with mental illness. We were able to use data visualization to understand how tech/IT workers responded to a questionnaire regarding mental health.

A screenshot of a cell phone

Description automatically generated

The bar graph illustrates that a very large majority of respondents feel that mental health affects job performance. This is an example of how my group used exploratory data visuals to try to better understand the story the data was telling and how to focus the project. Armed with this information, we decided to use job performance as a way to stress the importance for businesses to provide mental health programs for employees. We wanted to see what roadblocks may be preventing employees from approaching their supervisors with regards to mental health issues. This was executed with the use of association rule analysis:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Association Rule Analysis Summary: (top three models)** | | | | | |
| **lhs** | **rhs** | **Support** | **Confidence** | **Lift** |
| anonymity=no, MH. Consequence=yes, age=twenties | supervisor=No | 0.0112 | 1.0000 | 4.2648 |
| coworkers=no, anonymity=yes, MH. Consequence=yes | supervisor=No | 0.0167 | 0.9600 | 4.0942 |
| wellness.program=no, coworker=no,MH.consequence=no | supervisor=No | 0.1221 | 0.7028 | 2.9974 |
| coworker=no, MH.hurt.career=’yes, I think it would’ | supervisor=No | 0.1172 | 0.7058 | 3.0105 |
| anonymity=no, MH.consequence=yes | supervisor=No | 0.2023 | 0.8055 | 3.4355 |
| anonymity=no, exp.neg.reponse=’yes, I observed’ | supervisor=No | 0.0105 | 0.7143 | 3.0463 |

We were able to ascertain that age as well as work environment can play a role in whether or not an employee decides to seek help via their supervisor. This project revealed through the data that 1) mental health is an issue in tech and 2) this issue can affect job performance and potentially drive down production. It is ethically important for a company to care about its employee’s mental health, but I believe that the information gathered from this project shows that mental health can also affect the bottom dollar. We started this project with the idea that we would use the data to identify and predict what employees might be at risk based on a survey that included identifying factors; however, the data was trying to tell a more important story. This project allowed me to realize that exploring data is a lot like floating down a river: you can try to force a narrative from data, but the data will ultimately take you where it wants to go.

**Communication skills regarding data and its analysis for managers, IT, etc.**

A disconnect in communication between data scientists and the average healthcare worker is largely how I arrived into this master’s program. Communication is vital to giving a project worth. A costly AI program was deployed on the ER at which I was employed. This program was predicting waiting room surges to a very high degree of accuracy; however, the usability of this program was not communicated effectively to the senior nursing staff. This led to a complete loss of benefits that the program provided, and it was eventually turned off. Data science has little use for the world if we cannot communicate effectively. I have provided two project examples for two different audiences: the business audience and the academic audience. Additionally, I speak to how IST 718 helped me pay attention to the small details and communicate effectively among fellow data scientists.

[Final Project – MBC Data Analysis and Decision Making](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/MBC_638_Final%20Project)

This project did a good job of teaching me how to create a story board as well as PowerPoint suitable for a business setting. The DMAIC style project was designed to understand what variables affect my personal running times (specifically the first mile of a run). The storyboard delivery taught me how to condense a large amount of information into one board that is easy to interpret and digest. Leading the PowerPoint with the storyboard was a good way to quickly introduce the basic concepts to the audience before getting a little more into the weeds of the project in the PowerPoint slides. This project was great for exposing me to quality improvement and communication; it also made me a significantly faster runner as a bonus.

[Mental Health Project – Data Analysis](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_707_Mental%20Health%20Project)

The final product of this project was submitted in an academic-styled report. I was lucky to be paired with another student (Brian Hogan) who’s day-job involves grading and editing academic papers and theses. Brian was a great resource to help me understand formatting and style when it comes to writing academic papers. I believe that the knowledge gained from this project will be a tremendous help when I hopefully pursue a PhD in the future.

All homework submissions - Big Data

Lastly, I just wanted to say that I learned a great deal in the Big Data course when it comes to communication skills and submitting Jupyter notebooks and visuals. Professor Fox pushed me to make my notebooks and visuals more coherent. He also adjusted my view to pay more attention to the little details such as adding dollar signs to a histogram plot or presenting information into tables for readability.

These three areas have given me enough exposure to feel proficient and confident in submitting a body of work in all different types of settings.

**Synthesize the ethical dimensions of data science practice**

Ethics in data science are paramount. We as data scientists hold a great responsibility to provide clear, precise, and unbiased work with everything we do. Major decisions are being made that impact millions of lives on a personal and professional level. I have provided a paper written for IST 618 that details the current state of privacy and data. In my opinion, Information Policy is an essential course. This course challenged the way in which I look at what I do and how it may impact other lives.

[Privacy and the Public in the Technological Age – Information Policy](https://github.com/wabernathy3/Portfolio-Milestone/tree/master/IST_618_PrivacyPaper)

I wrote a paper in Information Policy about the ethical implications regarding the NSA and the program known as PRISM. PRISM was used to acquire access to private communications of users on nine different Internet platforms including Google, Facebook, and Microsoft to essentially spy on whomever the NSA deemed a threat. We all know that the use of data and data science can shape the world for the better, but it can also be used in malicious ways. Privacy laws like those recently passed in California are small steps in protecting the public and giving ownership to private data. I personally believe that an ethical committee needs to be established to review and approve any new projects that utilize public data and information without a consumer’s full informed consent. Currently, we have to rely on the values, morals, and ethics of businesses, which does not always have the consumer’s best interest. I hope to carry forth the ethical dimensions I learned as a nurse into my practice as a data scientist. We can really change the world with insight from data, but we have to ensure that this change benefits everyone equally.

**Conclusion**

As I watch the first week of async in my final term for this program I am easily reminded of the rapid progress I have made in efforts to succeed in getting this degree. The transition from healthcare to data science was difficult; however, the perseverance I learned through working in intensive care settings exposed me to the hard work and determination required to learn and excel in the data science program. This program has renewed my hunger for knowledge, which is nearly worth as much as the technical skills and theory. There is much more that I need to master, and I look forward to continuing to learn and build upon the program learning goals. Data science does not reward stagnation, and I consider myself lucky to pursue a career that benefits those with curiosity and an eagerness to learn.