|  |  |
| --- | --- |
| A picture of a winding road and trees  Kwasi Mensah’s Data Science Portfolio  **LinkedIn** | **Abstract**  The Master of Science in Applied Data Science program at Syracuse University equips students with a strong skill set in data science disciplines, emphasizing technical proficiency and ethical awareness. As Graduates, we emerge as competent data science practitioners with the ability to navigate the full data science life cycle. This portfolio showcases projects from the diverse courses taken at Syracuse University, highlighting the application of quantitative reasoning, scripting for data analysis, natural language processing, text mining, applied machine learning, and big data analytics. Through these projects, I demonstrated my capacity to collect, store, and access data, create actionable insights, apply visualization and predictive models, use programming languages effectively, and communicate findings ethically to various audiences.  **Kwasi Mensah**  **IST 782 - Portfolio**  **SU-Email: kwmensah@syr.edu**  **SUID: 897658917**  [**KwasMen/IST\_782\_Portfolio (github.com)**](https://github.com/KwasMen/IST_782_Portfolio) |

Table of Contents

[I. Introduction 3](#_Toc153098686)

[II. Portfolio Goals 3](#_Toc153098687)

[III. Project Gallery 3](#_Toc153098688)

[**1.** **IST 652 – Scripting for Data Analysis** 3](#_Toc153098689)

[**Course Description** 3](#_Toc153098690)

[**Project Description** 3](#_Toc153098691)

[**Project Synopsis** 3](#_Toc153098692)

[**Visualizations** 5](#_Toc153098693)

[**2.** **IST 664 – Natural Language Processing** 7](#_Toc153098694)

[**Course Description** 7](#_Toc153098695)

[**Project Description** 7](#_Toc153098696)

[**Project Synopsis** 8](#_Toc153098697)

[**Visualizations** 8](#_Toc153098698)

[**3.** **IST 718 - Big Data Analytics** 10](#_Toc153098699)

[**Course Description** 10](#_Toc153098700)

[**Project Description** 10](#_Toc153098701)

[**Project Synopsis** 11](#_Toc153098702)

[**Visualizations** 11](#_Toc153098703)

[**4.** **IST 736 – Text Mining** 13](#_Toc153098704)

[**Course Description** 13](#_Toc153098705)

[**Project Description** 14](#_Toc153098706)

[**Project Synopsis** 14](#_Toc153098707)

[**Visualizations** 15](#_Toc153098708)

[**5.** **IST 772 – Quantitative Reasoning for Data Science** 17](#_Toc153098709)

[**Course Description** 17](#_Toc153098710)

[**Project Description** 17](#_Toc153098711)

[**Project Synopsis** 17](#_Toc153098712)

[**Visualizations** 18](#_Toc153098713)

[IV. Team Projects 20](#_Toc153098714)

[**1.** **IST 659 – Data Administration Concepts and Database Management** 20](#_Toc153098715)

[**Course Description** 20](#_Toc153098716)

[**Project Description** 20](#_Toc153098717)

[**Project Synopsis** 20](#_Toc153098718)

[**Visualizations** 22](#_Toc153098719)

[**2.** **IST 664 – Introduction to Data Science** 24](#_Toc153098720)

[**Course Description** 24](#_Toc153098721)

[**Project Description** 24](#_Toc153098722)

[**Project Synopsis** 25](#_Toc153098723)

[**Visualizations** 26](#_Toc153098724)

[V. Conclusion 28](#_Toc153098725)

[VI. Resume 29](#_Toc153098726)

[SKILLS 29](#_Toc153098727)

[EXPERIENCE 29](#_Toc153098728)

[CERTIFICATIONS 31](#_Toc153098729)

[EDUCATION 31](#_Toc153098730)

# **Introduction**

The Master of Science in Applied Data Science program at Syracuse University’s School of Information Studies provides students with the opportunity to learn a wide range of data science disciplines and allow students the opportunity the practice these disciplines in various projects throughout the curriculum. As Graduate Students we learn to collect, store, and access data through various means, leveraging the current technology trends & tools. This program as a whole helps students with gaining technical proficiency in various skills needed in the data science field, and a deep understanding of ethical and social aspects of data science. Upon Graduation students will become competent data science practitioners, and potential data stewards with the skills and knowledge to make impactful decisions.

As students we gain an immense understanding of the full data science life cycle, thus allowing us to turn raw data into actionable insights. Through the use of various predictive models and visualization techniques and making use of industry standard software & packages in programming languages, we can deliver easy to comprehend insights that would be beneficial for any potential business client. In addition, we are able to communicate the insights gained to a broad range of technical and non-technical audiences when making use of visualizations, or programing languages with fluency such as R & Python for example. Most importantly, Graduate Students gain a deep-rooted understanding and foundation of ethics when dealing with data development, usage, and evaluation to avoid any biases in our data. In addition, we as students keep in mind fairness, transparency, and above all privacy when working with data.

The MS in Applied Data Science has serval learning objectives that align with the goal of this Portfolio. Specifically for this Portfolio we will highlight projects in courses such as Quantitative Reasoning for Data Science (IST 772), Scripting for Data Analysis (IST 652), Natural Language Processing (IST 664), Text Mining (IST 736~~),~~ Applied Machine Learning (IST 707), Big Data Analytics (IST 732) To demonstrate the ability to complete deliverables as a team the courses Introduction to Data Science (IST 687) & Database Administration (IST 659) will be focused on.

# **Portfolio Goals**

Below are the overall goals of the Project Portfolio is for students to demonstrate to the faculty expert overseeing the portfolio that the student is able to:

• 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 & Python to support the generation of actionable insight.

• Communicate insights gained via visualization and analytics to abroad range of audiences.

• Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

# **Project Gallery**

## **IST 652 – Scripting for Data Analysis**

### **Course Description**

This course, which is focused on the practical aspects of data science, delves into the art of scripting for data acquisition, access, and transformation across structured, semi structured, and unstructured data. The course emphasizes the development of scripting skills necessary to tackle a wide array of data-related challenges, including data wrangling. Students will learn to create data science pipelines, from data acquisition to preparation and transformation for subsequent analysis or visualization, employing the Python programming language. This course complements analytical and visualization skills learned in other data science courses.

### **Project Description**

In the final project, students will demonstrate their proficiency in writing Python scripts to access, aggregate, and utilize data from various types, including structured, semi structured, and unstructured data. We will showcase our ability to prepare and utilize data for generating summaries, lists, and other data structures. The final project will involve conducting experiments and presenting findings, illustrating the practical application of scripting skills acquired throughout the course.

### **Project Synopsis**

The project's primary aim was to conduct a comprehensive analysis of COVID-19 trends in the United States, drawing comparisons with trends in 14 other countries. The focus centered on vaccination rates and their correlation with the decline in COVID-19 cases. In addition, the project explored potential relationships between inflation and unemployment rates during the pandemic. This project involved thorough data collection and cleaning from various sources, including the Our World In Data website and government outlets, resulting in a dataset of over 300,000 rows and 67 columns, providing a strong foundation for in-depth analysis.

A distinguishing aspect of this project lay in the adept utilization of various Python libraries such as pandas, scikit-learn, requests, BeautifulSoup, and GitHub. These tools facilitated seamless data handling, exploration, and visualization. Additionally, inventive data handling functions were engineered to ensure smooth processing of datasets, encompassing functions for secure float conversion, acquiring direct web links to files, and writing/uploading files for our analysis.

The analysis yielded crucial insights. Notably, the trajectory of COVID-19 cases exhibited a consistent increase over time, particularly pronounced in the United States, which recorded the highest number of cases, followed by China, India, France, and Germany. By the close of 2022, approximately 65% of the US population had received COVID-19 vaccinations, likely contributing to the downturn in cases and fatalities. Pfizer/BioNTech spearheaded vaccine distribution in the US, trailed by Moderna and Johnson & Johnson. This project clearly displayed proficiency in data science and analysis skills, incorporating proficiency in data handling, exploration, and visualization through diverse Python libraries. In addition, the implementation of machine learning models, encompassing linear regression and random forest, provided valuable expertise in predictive modeling.

**Top of Form**

### **Visualizations**

A graph showing different colored lines

Description automatically generated

: This visualization shows a line graph of the trajectory of COVID-19 cases from 2019 to the present, highlighting the increasing trend in the 4 continents of Asia, Europe, North & South America .

A graph showing different colored lines

Description automatically generated

A line graph comparing the total COVID-19 cases in the United States with 14 other nations, providing a clear overview of the relative impact.

A blue and white rectangular object with text

Description automatically generated

A blue rectangle with white lines

Description automatically generated

A bar chart displaying the distribution of COVID-19 vaccines by manufacturer, emphasizing the most widely distributed vaccines around the world and the second is of the US specifically.

A screenshot of a video game

Description automatically generated

A scatter plot showcasing the states with the highest vaccination rates in 2021 and 2022, offering insights into regional trends.

A screenshot of a computer code

Description automatically generated

An example of modular code depicting three defined functions for visualization purposes. The entire code is available to view on GitHub.

## **IST 664 – Natural Language Processing**

### **Course Description**

This course explores the intersection of linguistics and computational techniques in natural language processing (NLP), equipping students with the skills needed for linguistic analysis of text at various levels. Emphasis is placed on practical application of computational techniques for NLP, enabling us to effectively analyze and interpret written text across diverse applications. The course covers objectives such as proficiency in processing text through different language levels with tools like NLTK and Python and gaining insights into real-world applications incorporating substantial NLP components. While original programming is not mandatory, those interested in computational techniques will have the chance to further develop their Python skills.

### **Project Description**

The final project requires students to focus on a text classification task, potentially involving sentiment analysis, and to conduct and report on a series of experiments. The code submitted must accurately reproduce the results discussed in the Final Project report. The development process should be clearly outlined, and the output should showcase results for two or more features, including comparisons with the baseline. Cross-validation code should be optimized for the task, and evaluation measures should be provided in terms of accuracy, precision, recall, F1, and confusion matrices. The conclusions should demonstrate a comprehensive understanding of feature engineering, evaluation measures, and the effect of cross-validation on result reliability, contextualized within the subject or business domain.

### **Project Synopsis**

The project focused on detecting spam emails within a dataset sourced from the Enron public email corpus, supplemented with additional spam emails, comprising 3672 ham emails and 1500 spam emails. The analysis contained sentiment and subjectivity analyses to determine between spam and ham emails. Furthermore, various classifiers were trained and tested to gauge their accuracy in prediction. The project's scope also extended to evaluating classifiers with new data from the "Enron2" dataset.

Throughout the project, several key insights were garnered. Notably, an average character count analysis revealed that spam emails tended to be longer on average than ham emails. However, sentiment analysis did not yield a definitive distinction in sentiment between the two categories. Similarly, subjectivity analysis did not serve as a reliable basis for differentiation. To ensure robust analysis, data preprocessing techniques like tokenization and removal of duplicates were employed, resulting in a refined corpus. Additionally, strategies for dataset balancing, including both down-sampling and up-sampling, were applied to mitigate classifier bias and overfitting.

The project also entailed training Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) classifiers using various vectorization techniques (Boolean, unigram and bigram term frequency, unigram TF-IDF). Notably, SVM classifiers consistently outperformed MNB classifiers, achieving higher accuracy scores across different vectorization options. Hyperparameter tuning using GridSearchCV for SVM models was employed, providing valuable insights for model improvement. When tested with new data, both SVM and MNB classifiers demonstrated reasonable accuracy in predicting spam and ham emails, with slight variations in performance. Overall, this project equipped me with valuable skills in data preprocessing, vectorization techniques, training and evaluating machine learning classifiers, hyperparameter tuning, and handling imbalanced datasets.

**Top of Form**

### **Visualizations**

**A blue and orange rectangular bars

Description automatically generated** A blue and orange rectangular bars

Description automatically generated A blue and orange rectangular bars

Description automatically generated

This visualization presents a bar plot displaying the distribution of spam and ham emails in the dataset after preprocessing. It provides a visual representation of the balance between the two classes, highlighting the number of spam and ham emails available for analysis.

**SVM Classifier vs. MNB Classifier**

**A screenshot of a computer error

Description automatically generated** **A screenshot of a number of numbers

Description automatically generated**

A comparison of precision, recall, and F1-scores for the best-performing classifiers. It offers insights into the balance between true positives, false positives, and false negatives. This helps evaluate the classifiers' overall effectiveness of the best-performing classifiers, when using a Support Vector Machine and Multinominal Bayes classifiers.

A screenshot of a computer code

Description automatically generated

A depiction of the impact of different hyperparameters on SVM classifier performance. This visual representation aids in understanding the hyperparameter tuning process and its impact on classifier performance.

**SVM Classifier vs. MNB Classifier**

A screenshot of a computer code

Description automatically generated A screenshot of a computer code

Description automatically generated

Comparison of the accuracy of our SVM and MNB classifier, to see which is the best model when introduced with new test data. We see that our SVM model gives us the best accuracy when introduced with new test data.

## **IST 718 - Big Data Analytics**

### **Course Description**

This course offers students a broad introduction to analytical processing tools and techniques. Upon completion, students will adeptly obtain and explain data, scrub data using scripting methods, explore data qualitatively, model data relationships, interpret findings, and select appropriate analytical methodologies for real-world problem-solving in business, science, and engineering.

### **Project Description**

This portfolio project is designed to demonstrate advanced proficiency in merging datasets and conducting insightful analyses for the identification of visual photos of clothing items. Specifically, the project aims to determine the most effective approach for classifying simple fashion images. Utilizing the Fashion MNIST dataset from Zalando Research, the project involves thorough data review, cleaning, and initial analysis.

Implementation of at least two classification approaches, such as Naïve Bayes, Neural Networks, Keras, Azure ML, IBM DSX, Boosted Trees, and/or Linear Classification will be employed. The key evaluation conditions includes the accuracy of each classification method, analysis of associated trade-offs, and a comprehensive assessment of computational performance. This project serves as a demonstration to the application of machine learning techniques in real-world scenarios, emphasizing sound data science methodologies and effective communication of findings.

### **Project Synopsis**

This portfolio project is focused on the classification of clothing items using machine learning techniques, utilizing the Fashion-MNIST dataset. The primary objective was to explore, model, and interpret the data to identify the most efficient algorithm for classifying simple fashion images. We employed seven distinct models, including Gaussian Naive Bayes, Categorical Naive Bayes, Random Forest, K-Nearest Neighbors, Linear Support Vector Machine, Neural Networks with TensorFlow using Keras, and a Multilayer Perceptron. Each model was evaluated, considering precision, recall, and F1 scores, providing insights into their performance characteristics.

Through thorough evaluation of each model, examining precision, recall, and F1 scores, several significant findings have emerged, providing insights into the performance characteristics of diverse machine learning approaches. The Gaussian Naive Bayes model demonstrated notable precision in accurately identifying ankle boots, while the Categorical Naive Bayes excelled in precision for trousers. Random Forest exhibited a balanced performance across various clothing categories, showcasing its versatility. The K-Nearest Neighbors model demonstrated high precision in identifying sandals. The Linear Support Vector Machine displayed precision in recognizing trousers and recall in identifying sandals. Neural Networks with TensorFlow using Keras, particularly the Multilayer Perceptron, exhibited competitive accuracy and reasonable compute time, indicating its effectiveness in image classification tasks. These findings emphasize the trade-offs between model accuracy, computational efficiency, and overall performance

### **Visualizations**

A picture containing clothing, dress, fashion, footwear

Description automatically generated

Displaying a sample image for each clothing type along with their respective labels and ID numbers provides an initial understanding of the dataset's diversity. The set of grayscale images showcase the diversity in design for different types of clothing, such as shoes, sneakers, sandals, and ankle boots, helps identify potential challenges in classification.

A picture containing text, sketch, black and white, screenshot

Description automatically generated

Examining misclassified images, especially those of pullovers and coats, sheds light on the model's limitations and areas for improvement.

A picture containing text, screenshot, number, software

Description automatically generated

A picture containing text, screenshot, colorfulness, number

Description automatically generatedA picture containing text, screenshot, diagram, number

Description automatically generated

A bar chart ranking the models by accuracy scores, along with visualizations of compute times and accuracy scores for training and testing sets, provides a comprehensive summary of each model's performance.

## **IST 736 – Text Mining**

### **Course Description**

This course introduces students to the concepts and techniques for extracting knowledge from large volumes of text data and applying text mining methods in various domains such as business intelligence, digital humanities, and social behavior analysis. The course provides a foundational understanding of text mining technologies, encompassing machine learning, natural language processing, and statistics.

We will gain proficiency in key areas of text mining, including document representation, information extraction, text classification and clustering, and topic modeling. In addition, we will also utilize standard corpora and both commercial and open-source text analysis and visualization tools to explore intriguing patterns.

### **Project Description**

For this project, students must select their own dataset. It can either be self-created or obtained from other sources such as Kaggle. The project will have several key components. Firstly, it should provide a comprehensive description of the real-world problem at hand, emphasizing its significance in the relevant domain. Next, the design should articulate the approach to problem modeling, which involves specifying the target categories for classification or anticipating the types of clusters expected in clustering tasks. Additionally, the rationale behind the selection of algorithms or exploratory analysis methods should be clearly elucidated, underscoring why they represent the most suitable solutions.

Moreover, the design should include pertinent details regarding data acquisition, encompassing aspects such as the duration of data collection, the quantity of examples gathered, and whether the dataset exhibits any skewness or balance. Furthermore, the chosen evaluation methods and metrics should be outlined, accompanied by a justification for their selection based on their appropriateness for the specific task. Lastly, the experiment design should also anticipate and identify potential challenges that may arise during the project.

### **Project Synopsis**

The project focused on applying advanced data analysis techniques to gain valuable insights from a dataset containing over 50,000 Amazon Fine Food reviews. The primary objectives were to first, to uncover meaningful topics and sentiments within the reviews through Topic Modeling, and second, to develop a predictive model using Support Vector Machines (SVM) to estimate rating scores for reviews lacking explicit ratings. The dataset, spanning more than a decade, presented a diverse range of reviews on various food products, making it a rich source for analysis.

The analysis began with a thorough data preprocessing phase, where columns deemed non-essential were removed to reduce noise. Due to computational constraints, a random sample comprising around 10% of the dataset was selected for analysis. Various techniques including TF-IDF vectorization, bigram creation, and lemmatization were employed to refine the corpus for both Topic Modeling and SVM modeling. Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) were utilized to extract topics, revealing distinct themes and sentiments within the reviews.

The project yielded noteworthy findings. Topic Modeling revealed prevalent themes such as food quality, flavors, and pet-related products, shedding light on popular review subjects. The SVM models, particularly when using TF-IDF vectorization, demonstrated impressive accuracy in predicting ratings for reviews without explicit scores. This capability holds significant potential for platforms seeking to implement star rating systems. The project displays a comprehensive approach to tackling intricate questions in the realm of sentiment analysis and rating prediction. Overall, the project not only showcased proficiency in data science and analysis but also demonstrated the value of leveraging advanced techniques to extract actionable insights from complex datasets.

### **Visualizations**

A text on a white background

Description automatically generated

Each cluster represents the topics derived from Latent Semantic Analysis (LSA). Each cluster highlights a specific theme within the reviews. This visualization offers a visual representation of the identified topics and their relatedness.

A graph with orange lines

Description automatically generated

A line plot displaying the coherence scores for various LDA models with different numbers of topics. This visualization aids in selecting the optimal number of topics for the LDA model, ensuring meaningful theme extraction from the reviews.

A screenshot of a graph

Description automatically generated

The LSA Topic Cluster plot represents the topics obtained through Latent Dirichlet Allocation (LDA). It visually displays distinct clusters, each corresponding to a specific theme or subject matter found in the reviews.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

A white background with black text

Description automatically generatedA screenshot of a computer code

Description automatically generated

Python code illustrating the accuracy of the SVM classifiers, indicating their effectiveness in predicting ratings for reviews with missing scores. This visualization provides a clear comparison of the performance of different SVM models.

## **IST 772 – Quantitative Reasoning for Data Science**

### **Course Description**

This course equips students with various strategies for inferential reasoning about quantitative data, emphasizing the importance of understanding uncertainty in data analysis. It delves into the process of statistical inference, enabling us to draw meaningful conclusions from samples of quantitative data. By the end of this course, we will possess the knowledge and skills to make informed choices in data collection, analysis, and interpretation, while also gaining proficiency in using statistical tools and platforms.

### **Project Description**

This project aims to provide evidence-based insights to a state legislator's office regarding vaccination rates in California school districts. Through comprehensive analyses of three datasets, we will assess U.S. vaccination trends, reporting proportions in public and private schools, and 2013 vaccination rates in California public schools. We will also explore the correlation between individual vaccine rates within districts. Furthermore, predictive analyses will identify factors influencing reporting completeness, up-to-date vaccination rates, and belief exceptions. The culmination of these findings will inform targeted financial allocation strategies, offering a roadmap for policymakers to enhance both vaccination rates and reporting compliance, ultimately ensuring the health and safety of students across the state.

### **Project Synopsis**

Through hands-on analysis, we strengthened our proficiency in data collection, cleaning, and management. I gained experience in leveraging programming languages like R and Python for in-depth data exploration and visualization. The project also provided an opportunity to apply predictive modeling techniques to generate actionable insights, enhancing my capabilities in statistical analysis. Moreover, we improved our ability to effectively communicate complex findings through clear visualizations and concise reports, a vital skill for conveying data-driven insights to diverse audiences.

In this project, three key discoveries emerged. Firstly, post-1990, U.S. vaccination rates displayed remarkable stability, with the Hepatitis B Birth Dose (HepB\_BD) vaccine showing an initial surge followed by a decline. Notably, the Diphtheria/Pertussis/Tetanus vaccine's (DTP1) first dose recorded the highest rate, while HepB\_BD had the lowest towards the project's end. Secondly, a noteworthy disparity was uncovered in reporting practices between public and private schools; approximately 76% of public schools provided vaccination data, in contrast to only 19% of private schools. Lastly, the alignment of California's 2013 vaccination rates with national averages was observed, with the Hepatitis B vaccination rate in California significantly impacting the overall U.S. rate for that specific year. These findings illuminate critical nuances in vaccination trends and reporting practices, offering valuable insights for informed decision-making.

In summary, this project not only provided crucial insights for policymakers but also further improved our data science skill set. It elevated our proficiency in data handling, programming, visualization, and predictive modeling. This project serves as a demonstration to the power of data-driven decision-making in informing critical public health strategies.

### **Visualizations**

**A graph of a vaccination rate

Description automatically generated**

The Time Series Plot of U.S. Vaccination Rates will show the trends in vaccination rates over time, highlighting any significant changes or patterns.

A red squares with black text

Description automatically generated

A heatmap displaying the contingency table results for public and private schools' reporting proportions, providing a clear visual representation that if a student is missing a vaccine they are likely missing all vaccinations.

**A screenshot of a computer

Description automatically generated**

This linear regression will show the relationship between predictors (Enrolled and TotalSchools) and reporting completeness, visually demonstrating their impact.

# **Team Projects**

## **IST 659 – Data Administration Concepts and Database Management**

### **Course Description**

This course covers the fundamentals of database management systems, including data structures, file organizations, and the principles of DBMS. Students will explore concepts such as data analysis, database design, and data modeling, gaining hands-on experience in database design and implementation through assignments and lab exercises. Additionally, the course delves into advanced topics like transaction management, distributed databases, client-server architectures, and query optimization techniques for extract, transform, load (ETL) processes. Through a combination of theory and practical application, students will develop essential skills in managing and utilizing databases for information systems.

Upon completing this course, we will have a solid grasp of fundamental data and database concepts, encompassing various storage models. We will be proficient in utilizing the database development life cycle and data models to address business problems, particularly in the context of the relational data storage model. We will demonstrate our ability to construct database objects and execute queries using SQL for problem-solving purposes. Additionally, we will be adept at identifying and implementing enhancements for data integrity and performance in existing database designs. The course will also equip us to evaluate and choose appropriate approaches for data migrations, temporal data management, and data normalization. Finally, we will be able to develop the capacity to critically assess the effectiveness of Database Management Systems (DBMS) within computer information systems.

### **Project Description**

In this course, students will engage in a team project focused on designing and implementing a functional system with a database. The project entails the formation of self-assembled teams consisting of two to three students, as well as the selection and pre-approval of a project idea, including the submission of a requirements document if it's an original concept. Teams will conduct data analysis to identify entities, attributes, and relationships within the data model, and create both conceptual and logical data model diagrams. Teams will also define the external data model and data logic, design the basic layout for all application screens, and generate diagrams for each screen used in the application. Additionally, teams will prepare SQL up/down scripts for implementing the internal model along with initial data, as well as for loading/migrating existing data and for data logic related to the external data model. The project culminates in the execution of the application's implementation. This comprehensive approach ensures a thorough understanding and practical application of database design principles.

### **Project Synopsis**

The project undertaken by “Data Models” for “Precise Management Production” was aimed at addressing the challenges faced by the company in managing their movie production information. The main objectives encompassed the design and implementation of a database system that would facilitate efficient storage, retrieval, and analysis of various movie-related data points, including genres, film locations, budgets, ratings, cast members, and directors. The new database system was intended to offer enhanced functionality, allowing users to conduct searches for movie data, actors, directors, as well as analyze movie performance and casting by year.

One of the primary goals of the project was to replace the existing Excel-based system with a robust database solution, which effectively mitigated issues related to clerical errors, data inconsistencies, and version control. The database was carefully designed to empower users with the ability to perform tasks such as researching current assignments of actors and directors, organizing movie castings, tracking ratings and revenue, and conducting in-depth statistical analyses on movie performances.

The provided SQL code on GitHub exemplifies a comprehensive database management system, starting with the removal of existing procedures, views, and tables to establish a clean slate. It then meticulously creates essential tables like 'roles', 'persons', 'genres', 'locations', 'prodcompany', 'movies', 'ratings', and 'moviecast', each with carefully defined attributes and constraints

Following table creation, views are then established to provide valuable insights. The 'budget' view calculates financial performance metrics, and 'budgetinfo' categorizes movies based on budget and profit status. Two additional views aggregate data on individuals' earnings and location rental rates. The code also implements stored procedures for seamless data manipulation, covering updates to person rates, location rental rates, fan, and critic ratings. Finally, search procedures enable precise information retrieval, such as finding film locations with rental rates matching specific values or persons with names matching given inputs. This SQL code showcases expertise in database design, manipulation, and optimization, making it an invaluable asset for any portfolio demonstrating SQL proficiency.

A distinctive aspect of this project was the innovative approach taken in the creation of conceptual and logical models, employing SQL and database management tools. These models played a pivotal role in identifying the underlying relationships between different entities, ultimately contributing to the efficient structuring of the database. Additionally, the team leveraged tools like Microsoft Excel and Tableau to provide visual representations of findings, offering a comprehensive view of the data.

The project significantly bolstered the team's expertise in database design, implementation, and management using SQL, and honed proficiency in utilizing tools such as Microsoft Excel and Tableau. It provided invaluable hands-on experience in creating conceptual and logical models and transforming these models into a functional and efficient database. The process of identifying foreign keys and gaining a deep understanding of the database lifecycle further enriched the team's proficiency in database development.

### **Visualizations**

**A diagram of a movie production

Description automatically generated** **A screenshot of a computer

Description automatically generated**

This is a visualization of our conceptual and logical model for the database that will be created using SQL for the movie database to be created.

**A screenshot of a computer program

Description automatically generated**

The "budget" view in the database compiles key financial metrics for movies, including initial and actual budgets, alongside theater and total revenues. This allows for a concise evaluation of a movie's financial success by comparing costs with revenue

**A graph of a number of locations

Description automatically generated**

A bar chart illustrating the frequency of movies filmed at different locations, generated using Tableau and SQL.

## **IST 664 – Introduction to Data Science**

### **Course Description**

This course provides students with practical exposure to the data science field, covering data collection, processing, transformation, management, and analysis. It introduces key concepts such as applied statistics, information visualization, text mining, and machine learning, with a focus on utilizing the widely used open-source tool, "R." Proficiency in R is highly regarded in the data analytics community, enhancing job prospects for aspiring data scientists. By the end of the course, we will have gained a solid grasp of essential data concepts, scripting/code development using R and R-Studio, data screening, cleaning, and linking practices, as well as effective communication of results to decision-makers. We will also be equipped to identify problems, perform basic computational scripting, manipulate data, and manage it throughout various stages of a project life cycle, while employing appropriate techniques for data analysis. This comprehensive curriculum equips us with valuable skills in the field of data science.

### **Project Description**

Teams in this course are tasked with selecting a dataset containing a minimum of 10,000 values, with flexibility to choose datasets from various sources, including those from their own companies (ensuring confidentiality is maintained). The emphasis is on applying advanced data science concepts rather than relying solely on descriptive statistics. Project updates will follow an Agile Kanban Data Science methodology, involving the use of a Kanban board and a concise one-page project summary. The chosen project, titled "Wisconsin Breast Cancer Diagnostic and Prognostic Outcomes," focuses on datasets related to breast cancer diagnostics and prognostics. Each dataset's structure is presented, and business goals range from determining key variables in cancer diagnoses to developing accurate prediction models. One anticipated challenge involves matching ids between different datasets. The data will be organized into 2D tables, subsequently converted into R data frames for manipulation, and ultimately exported as a .csv file for seamless sharing among team members.

### **Project Synopsis**

For this project, we delved into the Wisconsin Diagnostic and Prognostic Breast Cancer datasets, encompassing a total of 767 records and 66 attributes. Employing the R language within RStudio, we meticulously analyzed these datasets to extract meaningful insights. Our primary objectives were multifaceted, ranging from identifying pivotal variables in distinguishing benign from malignant breast cancer cases, to discovering the factors most influential in predicting recurrence/non-recurrence. Through rigorous data acquisition, cleaning, and transformation processes, we meticulously prepared the datasets for comprehensive analysis. Despite the absence of traditional demographic data, we uncovered intriguing trends, notably the significant imbalance between benign and malignant cases, which sparked further investigation..

One notable aspect of our approach was the careful data acquisition process, where we transformed raw data into structured data frames using the 'tidyverse' package in RStudio. We further enriched the dataset by adding descriptive column names. A key observation was the significant imbalance between benign and malignant cases, with approximately 37.3% classified as malignant, underscoring the need for nuanced analysis in such cases.

In the pursuit of meaningful insights, we employed correlation matrices to examine relationships within the data. Additionally, we identified the top ten variables with the highest correlations to both benign/malignant diagnoses and recurrence/non-recurrence prognoses. These findings laid the groundwork for subsequent modeling techniques. Leveraging Support Vector Machines and Decision Tree classifiers. The subsequent application of Support Vector Machine (SVM) and Decision Tree modeling provided us with deep insights into the predictive capabilities of the data. The SVM model demonstrated exceptional accuracy in diagnosing breast cancer cases, offering a vital tool for early detection. The Decision Tree approach shed light on the most influential factors affecting cancer diagnoses and prognoses.

This project equipped us with invaluable skills in data acquisition, cleansing, and transformation, laying the foundation for advanced analytical techniques. In addition, it improved our proficiency in data science and analysis, sharpening our ability to extract actionable insights from complex datasets. It highlighted the importance of methodical data acquisition, rigorous cleaning, and the thoughtful application of advanced modeling techniques. Despite some challenges in modeling the prognostic data, our approach illuminated critical insights into cancer recurrence prediction. This project not only strengthened our proficiency in data science but also highlighted the significance of robust analytical techniques in the field of healthcare.

### **Visualizations**

Chart, treemap chart

Description automatically generatedChart, pie chart

Description automatically generated

This visualization provides a clear visual representation of the distribution of benign and malignant cases in the diagnostic dataset, highlighting the prevalence of each category. The pie chart shows the percentage of cases that are benign and malignant in the diagnostic dataset. This highlights the data's class imbalance.

Graphical user interface, application

Description automatically generatedText

Description automatically generated

This visualization depicts the top 10 highest correlation between variables the relationships between variables in the diagnostic dataset, offering insights into potential influential factors in breast cancer diagnoses. In addition, the visualization presents the performance of the Support Vector Machine model in diagnosing breast cancer, showing the number of true positives, true negatives, false positives, and false negatives.

Graphical user interface, text, application

Description automatically generated Text

Description automatically generated

This visualization depicts the top 10 highest correlation between variables in the prognostic dataset, offering valuable insights into factors affecting recurrence/non-recurrence prognoses. In addition, the visualization presents the performance of the Support Vector Machine model in diagnosing breast cancer, showing the number of true positives, true negatives, false positives, and false negatives.

Text

Description automatically generated Table

Description automatically generated

Above are the most important diagnosis and prognosis features identified by the Decision Tree Classifier model. This provides insight into the factors influencing diagnoses and recurrence/non-recurrence prognoses.

# **Conclusion**

In this comprehensive portfolio, I have showcased a diverse range of projects undertaken during my studies in various Data Science, Business, and Technology courses. These projects span multiple domains, including data analysis, natural language processing, big data analytics, text mining, and quantitative reasoning for data science. Each project reflects a culmination of theoretical knowledge, hands-on application, and a commitment to addressing real-world challenges.

Throughout these courses, I have sharpened my skills in data acquisition, cleaning, transformation, and analysis. I have demonstrated proficiency in using programming languages such as Python and R, leveraged powerful libraries and frameworks for data manipulation and machine learning, and applied statistical and analytical techniques to extract meaningful insights from complex datasets. In addition, the projects highlight my ability to create clear and compelling visualizations, showcasing the importance of effective communication in the field of data science.

In addition, the team projects highlight my capacity to collaborate and contribute to the development of functional systems. These experiences have not only deepened my understanding of data science concepts but have also equipped me with practical skills that are crucial in today's data-driven landscape. Overall, this portfolio serves as a demonstration of my journey in data science and technologies alike.

# **References**

IST\_782\_Portfolio (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 782 Portfolio. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/Portfolio>

IST\_772\_Quantitative\_Reasoning\_for\_Data\_Science (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 772 Quantitative Reasoning for Data Science. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/IST_772_Quantitative_Reasoning_for_Data_Science>

IST\_718\_Big\_Data\_Analytics (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 718 Big Data Analytics. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/IST_718_Big_Data_Analytics>

IST\_664\_Natural\_Language\_Processing (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 664 Natural Language Processing. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/IST_664_Natural_Language_Processing>

IST\_659\_Data\_Administration\_Concepts\_and\_Database\_Management (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 659 Data Administration Concepts and Database Management. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/IST_659_Data_Administration_Concepts_and_Database_Management>

IST\_652\_Scripting\_for\_Data\_Analysis (n.d.). Retrieved from [GitHub link]

KwasMen. (n.d.). IST 652 Scripting for Data Analysis. GitHub. <https://github.com/KwasMen/IST_782_Portfolio/tree/main/IST_652_Scripting_for_Data_Analysis>

# **Resume**

**Kwasi Mensah**

Kwmensah@syr.edu | (646) 492-0012 | New York , New York | https://www.linkedin.com/in/kwasi-mensah/

# **SKILLS**

**Administrative**: Microsoft Office, Google Suite, RTG Bullhorn, ITIL, Process Redesign, Change Management

**Graphic Design:** Adobe XD, Figma, Google Site Maps, Adobe Creative Suite, Visual Studio, Tableau

**Analytics:** Microsoft Office Excel, R, R Studio, Data Cleaning, SQL, SQL Server, MySQL, Technical Reporting, Microsoft Access

**Technology:** Python, Windows OS, macOS, Android, iOS, Linux

**Soft Skills:** Communication, Problem-Solving, Customer Focus, Teamwork, Time Management, Adaptability, Attention to Detail, Multitasking, Conflict Resolution, Learning Agility, Cultural Sensitivity, Flexibility, Project Management, Networking.

# **EXPERIENCE**

**Deloitte 06/2023 – 07/2023**

*Data Analyst Intern – Risk and Financial Advisory*

*Regulatory and Legal Support Team member in Risk and Financial Advisory, focusing on Financial Industry Risk and Regulatory compliance.*

* Contributed to 3 impactful Risk and Financial Advisory projects, including an 8-week engagement with a multinational public US bank to lower market and fraud risk through extensive analysis and research.
* Optimized turnaround time for a time-sensitive project, implementing research-backed recommendations, resulting in enhanced efficiency for an 11-person team.
* Excelled in four data-driven case studies, showcasing proficiency in data analytics, management, and AI integration in the evolving economy.

**Vitamin Shoppe 08/2022 – 06/2023**

*IT Service Desk Analyst – IT Support & Infrastructure*

*Proven track record of providing exceptional customer service and technical support at Vitamin Shoppe. Skilled in troubleshooting and resolving customer inquiries, managing customer complaints, and providing product information.*

* Communicate between management, colleagues, store managers, and onsite tech to alleviate issues with the new network being installed to avoid future migration issues.
* Verified computers, registers, iPads, and printers were functional.
* Assisted in a project with Hughes LLC to migrate Vitamin Shoppe over to the Hughes Network, with 700+ stores networks upgraded during business hours without impacting customers in the various stores.
* Supported the VOIP project, including phone installation & disposal, setup, and testing, while collaborating effectively with a third-party vendor.

**RANDSTAD USA 11/2021 – 02/2022**

*Technical Recruiter - Financial Services; Strategic Vertical Group*

*Manage full cycle recruiting including preparing, sourcing, screening, interviewing, extending an offer and onboarding. Analyze statistical reports and other data to monitor applicant flow, selection, and turnover for compliance with legal regulations and with organization goals and policies. Establish and maintain relationships with Key Management, Hiring Managers, and clients to forecast and determine current and future company staffing goals while successfully developing proactive recruiting strategies to meet those needs.*

* Leverage multiple sourcing strategies and channels including Bullhorn, LinkedIn Recruiter, Monster, and Proprietary software, and internal referrals to provide high-quality IT candidates for open positions across the nation.
* Maintain active and passive candidate information in Bullhorn, as well as sustain data on recruitment activities, applicant flow, interviews, hires, transfers, promotions, and terminations.
* Develop national sourcing and recruiting strategies designed to identify qualified candidates through various recruiting tools.

**BANK OF AMERICA 10/2019 - 04/2021**

*IT Operations Support I – Trading Floor Support*

*Operated within a multi-faceted role, liaised with a dynamic team, delivered best-in-class technical support to over 1200 users +, including traders, seniors, directors, managing directors, & VPs. Increased user satisfaction by troubleshooting various software and network related issues concerning applications, PC performance, websites availability, shared drives/folders, VPN & remote desktop client connectivity, and internet connectivity.*

* Swiftly resolved 50-80+ tickets a week regarding user issues with desktops, laptops, printers, monitors, iOS and Android OS devices, Microsoft Office Suite, Web-Ex Computer hardware, Citrix, and Virtual Machines.
* Maximized operational efficiency by collaborating on a plethora of projects with cross-functional teams to mitigate complex issues, such as installing and overseeing the testing of new products, pushing out software and updates to computers.
* Leveraged exceptional technical acumen to troubleshoot and configure Bloomberg, software, and hardware peripherals.
* Ensured optimal level of service by providing remote technical support to users working from home due to COVID-19 pandemic.

# **CERTIFICATIONS**

Microsoft Office Specialist Excel Expert **|** Microsoft Certified Azure DevOps Engineer Expert

AWS Solutions Architect Associate**|** AWS Cloud Practitioner **|** AXELOS ITIL 4 Foundation

Microsoft Certified Azure Administrator Associate **|** CompTIA ITF+

# **EDUCATION**

**Syracuse University GPA:** 3.92

Applied Data Science (MS) Present

**Google Data Analytics, Coursera**

Certificate of Completion Mar 2022

**NPower Technology Fundamentals**

Certificate of Achievement Jul 2019

**University at Albany, State University of New York**

Rhetoric & Communication (B.A.) | Informatics (Cyber Security) (B.S.) Dec 2016