Syracuse University, School of Information Studies

M.S. Applied Data Science

Portfolio Milestone

Amanda Austin

https://github.com/AmandaLeigh03/SU_Portfolio

Table of Contents

Introduction 2
IST 659: Data Administration Concepts & Database Management 2
Project Description 2
Reflection & Learning Goals 4
IST 623: Introduction to Information Security 5
Project Description 5
Reflection & Learning Goals 5
IST 719: Information Visualization 6
Project Description 6
Reflection & Learning Goals 7
IST 718: Big Data Analytics 7
Project Description 7
Reflection & Learning Goals 10
Conclusion 11

Introduction

The Applied Data Science Program at Syracuse University's School of Information Studies challenges students to study the entire data pipeline: data collection, data management, data cleaning, data analysis, and data visualization. A wide variety of tools and techniques are provided to develop insights into the data, then communicate these insights as actional recommendations for strategic decision making. SQL Server Management Studio, Microsoft Access, R Studio, Adobe Illustrator, and Python were utilized for reports and presentations in courses such as Data Administration Concepts & Data Management (IST 659), Introduction to Information Security (IST 623), Information Visualization (IST 719), and Big Data Analytics (IST 718). The techniques

Syracuse's Applied Data Science Program has seven learning objectives, which we're achieved through coursework and projects:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualizations, statistical analysis, and data mining.
4. Develop alternative strategies based on data.
5. Develop a plan of action to implement the business decisions derived from analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in the organization (managers, IT, programmers, business users, etc.).
7. Synthesize the ethical dimensions of data science practice.

IST 659: Data Administration Concepts & Database Management

Project Description

In Data Administration Concepts & Database Management (IST 659), a SQL database was built based off a real-life Access database used by Broadridge Financial Solutions, a global fin-tech company that also

provides clients with mailing solutions. The data was collected from the existing Access database and anonymized as to protect the names of Broadridge employees and company information such as clients, vendors, and specific mailing solutions.

In production operations, physical mail is printed, inserted, and tendered to either the United States Postal Service (USPS) directly or to a third party presort mailing vendor. This database was developed to allow the logistic departments to record all mail that is tendered to presort vendors, provide reports for financial reconciliation, and for manages to make business decisions. The tables (Image 1), indexes, views, and stored procedures were created in SQL Server Management Studio, then the database was connected to Microsoft Access using ODBC to provide end-users with an easy-to-use interface with forms to input data (Image 2) and standard reports (Image 3). User permissions were also implemented to control access to sensitive information, as well as protect data integrity.

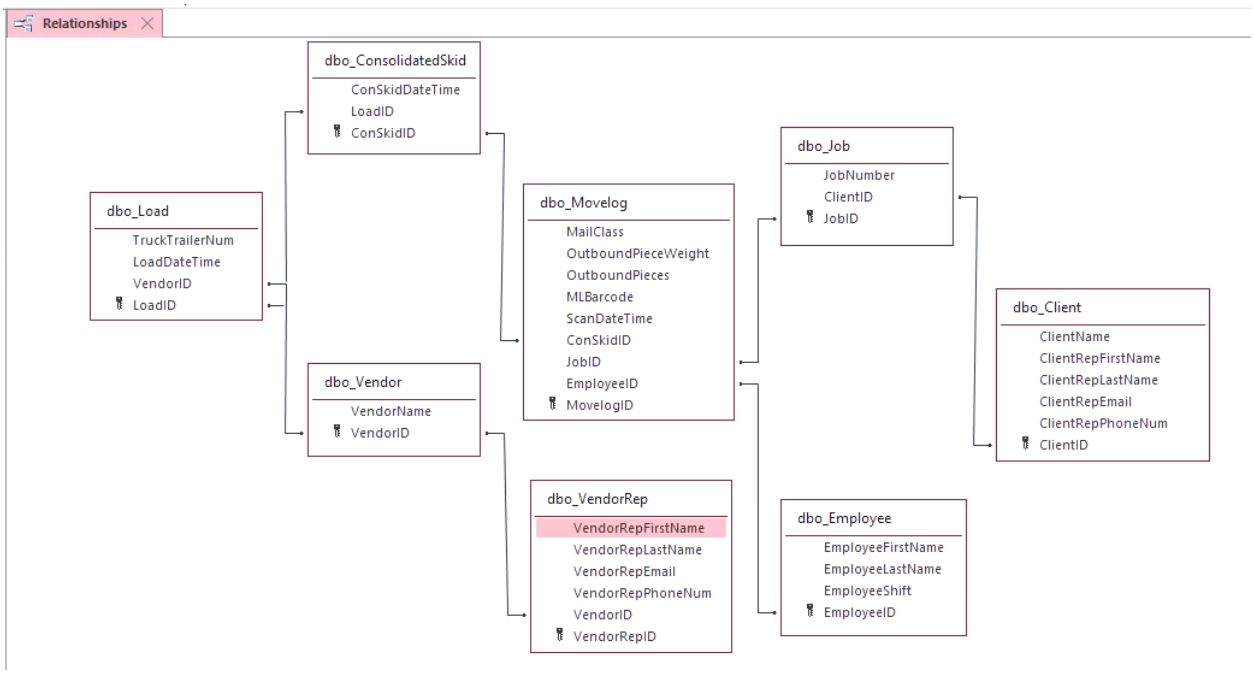

Image 1 – Logical Model in Microsoft Access

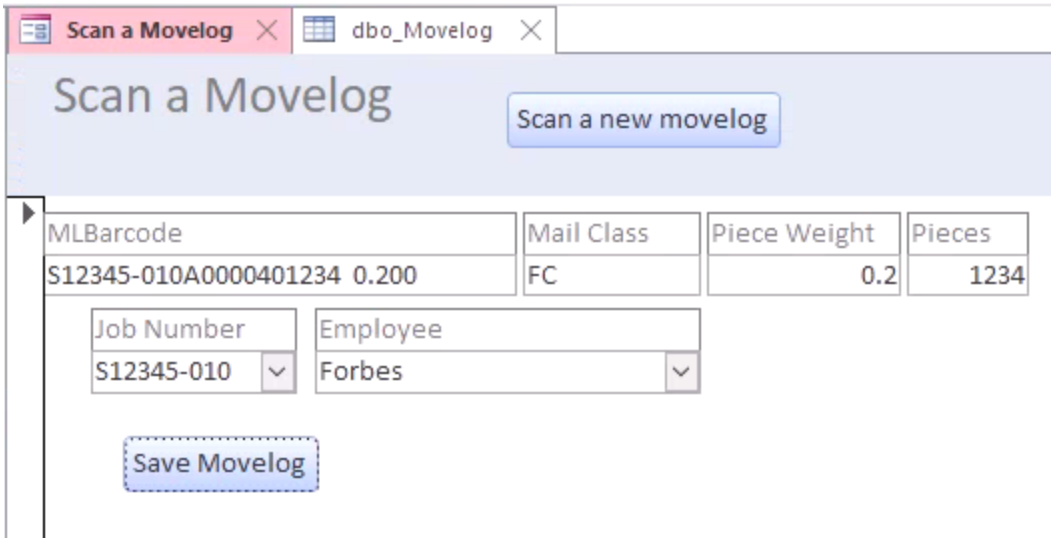

Image 2 – Entering data into the database

Image 3 – Report containing Proof of Mailing for Client Services

Reflection & Learning Goals

IST 659 and the corresponding project emphasized how crucial data management and integrity are to an organization's success. Additional courses in the Applied Data Science Program such as Advanced Big Data Management (IST 769) and Data Warehouse (IST 722) further solidify this lesson by introducing

more advanced data management solutions and architectures for larger data sets from disparate sources. Manual tasks were automated with the use of Microsoft Access forms – the database allows end users to scan barcodes into Access directly rather than typing in information, drastically reducing mail processing times and ensuring client satisfaction in Broadridge Mailing Solutions.

This project achieved learning goal 2, collect and organize data. The data was collected from an existing real-life Microsoft Access solution, loaded into the SQL database, then organized into SQL tables. Forms and reports were created in Access for end-users to seamlessly input and read data.

IST 623: Introduction to Information Security

Project Description

In Introduction to Information Security (IST 623), identity theft and the dark web were researched to identify solutions to protect personally identifiable information. Group member, David Ladd, provided a case study outlining how his identity was stolen from a prescription pill bottle found. He shared the emotional and financial implications of identity theft, and steps he now takes to protect his identity. Identity theft data was analyzed to find trends in how victims discover identity theft, theft resolution versus emotional distress, and actions victims now take to prevent secondary instances. Lastly, the dark web was researched, including its origins, initial intended purpose and benefits, and its use today for criminal purposes.

Reflection & Learning Goals

This case study and the other material presented in IST 623 underscores the importance of protecting personally identifiable information (PII), both at the individual level and within business practices. PII should only be shared with those who require it, and records should be destroyed as soon as the data is no longer of use. In addition, companies must protect their client, employee, and company data to

ensure safety from identity theft and other data breeches. These lessons achieve learning goal seven, synthesize the ethical dimensions of data science practice.

IST 719: Information Visualization

Project Description

In the course Information Visualization (IST 719), R was used to collect, clean, analyze, and visualize data on political donations made by American sports team owners and commissioners. The data was sourced through Kaggle, then loaded into R for analysis and generation of the alluvial plot, scatter plots, and bar charts. Adobe Illustrator was utilized to present the findings in the form of a poster, containing polished versions of the R generated plots, as well as insights into the findings from the analysis. Image 4 shows the section of the poster that contains a data description, the alluvial plot, and corresponding explanation of the alluvial plot. The intended audience is American sports fans who are interested in the potential political affiliations of leagues or teams they support. As such, the information is presented in a concise and easy to comprehend manner to keep the audience engaged.

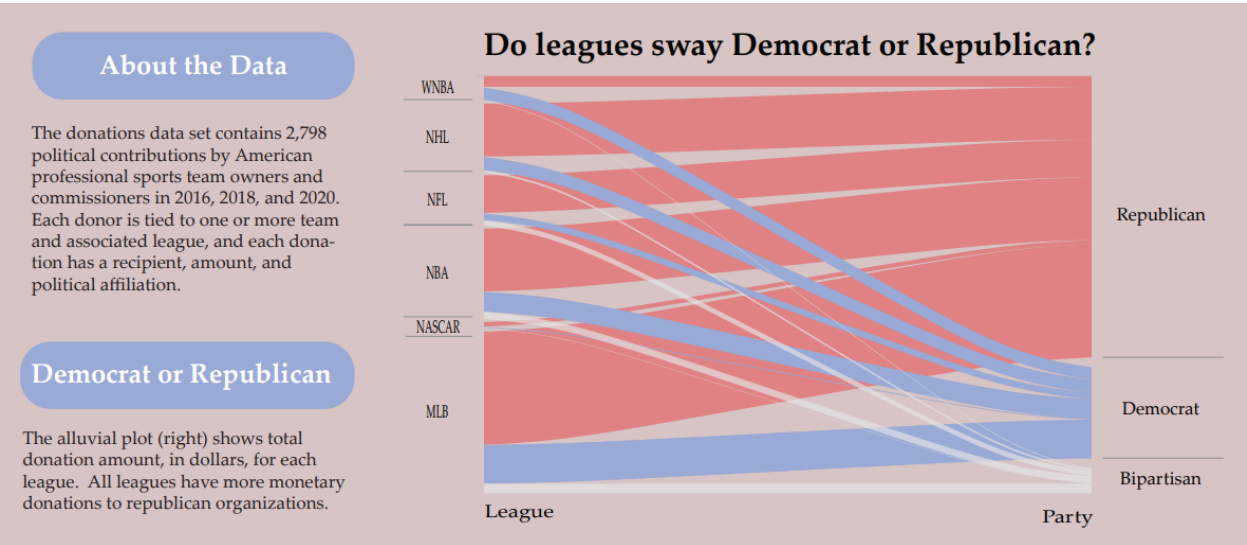

Image 4 – Portion of Poster (IST 719)

Reflection & Learning Goals

The poster project from IST 719 had students explore advanced techniques of presenting data visually using R and Adobe Illustrator. Data visualization skills are crucial for capturing the audience and presenting findings in a way that is suitable for the intended audience. Typically, data scientists are not presenting to other data scientist, but instead to business users, managers, and executive leaders to make strategic business decisions. It is crucial that the findings of the analysis are conveyed in a non-technical manner, and that visualizations are clear and capture attention.

Learnings goals two, three, and six were achieved during IST 719 and through completing this poster project. Data was collected from Kaggle and loaded into R to be cleaned and organized. Then, patterns in the data were found through an exploratory data analysis and data visualizations. Finally, the analysis findings were communicated to an audience of sports fans, who are assumed to not be data scientists or analysts.

IST 718: Big Data Analytics

Project Description

Big Data Analytics (IST 718) challenges students to utilize all tools and techniques studied during the Applied Data Science program. This includes collecting and organizing structured and unstructured data from disparate sources, advanced analytical techniques such as machine learning, data visualization, communicating analysis results, and providing actionable insight to the audience.

Mental health in the technology industry was studied for the final project. Survey data from Open Sourcing Mental Illness (OSMI), Google Trends data on search terms, and Twitter data were sourced to provide recommendations to companies looking to better support their employee's mental health, especially during the unprecedented time of COVID-19. The purpose of the analysis was to determine

what companies are currently doing to support employee mental health, how COVID-19 affected mental health, and the sentiment around mental health.

An exploratory analysis of the OSMI survey data provided visualizations to show distributions, geographical trends (Image 5), correlations between survey question. For the OSMI survey data, a response variable "score" was calculated to determine how supported the respondent feels in terms of mental health. This score was the summation of the responses for "overall, how much importance does your employer place on mental health?" and "overall, how well do you think the tech industry supports employees with mental health issues?" If the respondent scored highly on these two questions, it can be presumed that they generally feel supported at work.

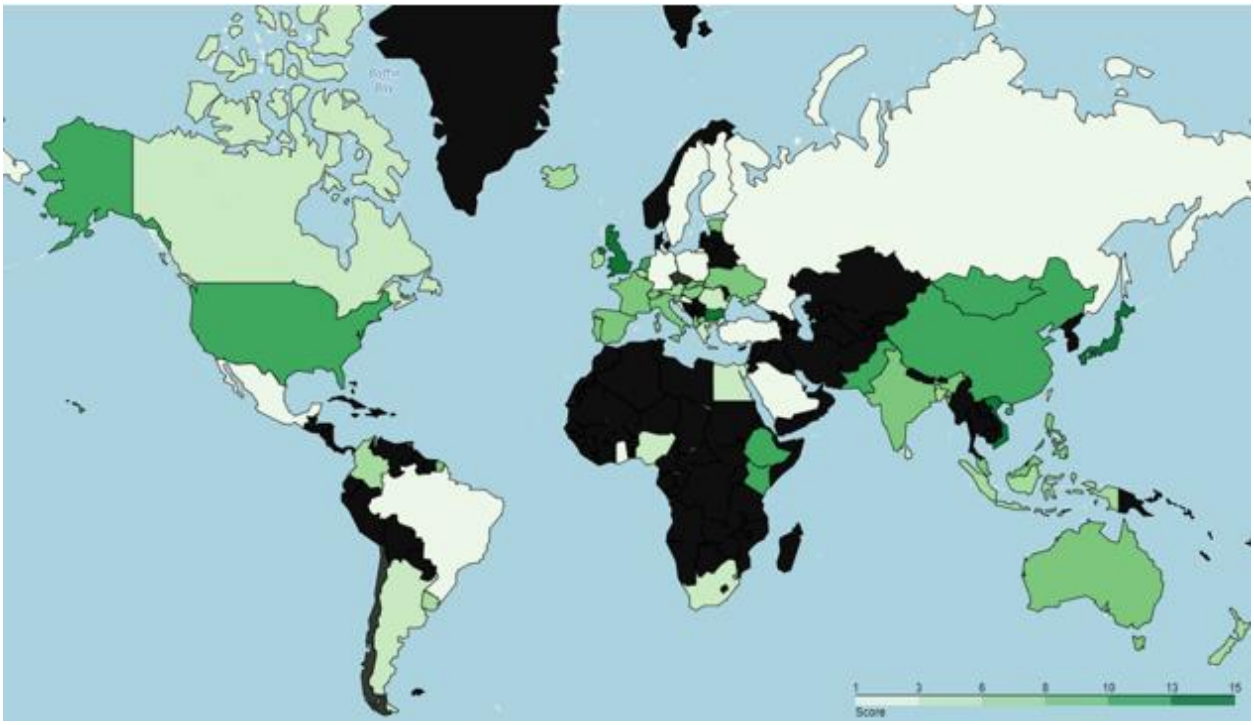

Image 5 – Countries Respondents Work In (Score)

Following the exploratory data analysis, a regression analysis was preformed to determine the questions that positively and negatively contributed to respondents score. The regression model accounted for 53.8% of the variability in the data (Image 6). It was found that a work culture which allows employees

to open discuss mental health and clear policy that supports employees with mental health disorders positively contribute to score. Companies and human resource departments can look at these specific survery questions and implement policy that is in line with employee values. For example, allowing employees to take time off of work for mental health reasons and providing educational resources to employees regularily. A random forest model was also run, which found similar results with 67.7% accuracy and a mean absolute error of 1.63 degrees.

OLS Regression Results

Dep. Variable:	Score	R-squared:	0.538
Model:	OLS	Adj. R-squared:	0.530
Method:	Least Squares	F-statistic:	67.04
Date:	Sat, 26 Mar 2022	Prob (F-statistic):	2.80e-220
Time:	18:04:14	Log-Likelihood:	-3105.7
No. Observations:	1463	AIC:	6263.
Df Residuals:	1437	BIC:	6401.
Df Model:	25		
Covariance Type:	nonrobust		

Image 6 – Regression Results

Google Trends data on search terms “mental health,” “depression,” “anxiety”, and “therapy” was pulled from Google and run through a time series analysis. This analysis found that searches for “anxiety,” “therapy,” and “mental health” have increased over time, especially after the first lock down for COVID-19, while searches for “depression” decreased.

Lastly, Twitter API was used to scrape Tweets with the hastags “depression,” “anxiety,” and “mental health.” The text of these Tweets were then put into word clouds and analysed for sentiment. Overall, there is a positive sentiment for depression and mental health, possibly indicating that the stigma around mental health has decreased, and that Twitter users are more likely to speak positively about

overcomming struggles with mental health, specifically depression. However, the tweets pertaining to anxiety had a negative sentiment (image 7).

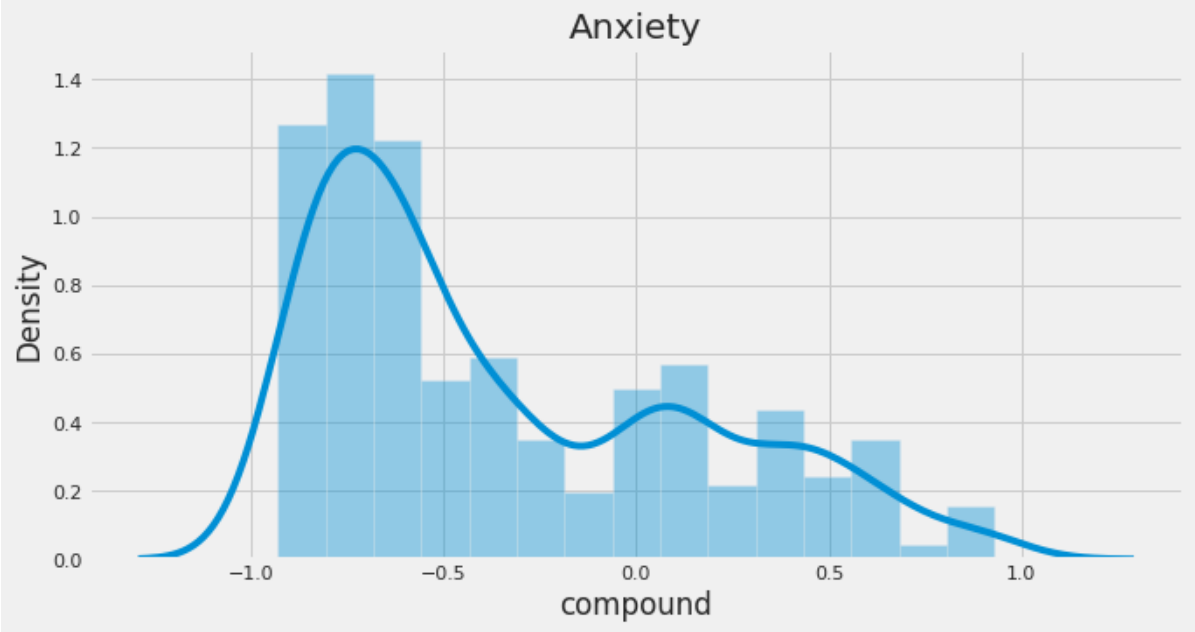

Image 7 – Sentiment Analysis for “Anxiety”

Reflection & Learning Goals

Nearly all learning goals were achieved in the analysis for IST 718. Big Data Analysis and the corresponding project allows students to utilize all skills gained throughout the Applied Data Science program, achieving learning goal one. Business analysis was demonstrated in helping companies with business decisions relating to employee mental health and well-being, programming in Python for the analysis displayed computer science skills, and mathematics & data analysis were utilized to find trends in the data using advanced analytics and machine learning techniques.

Data was collected and organized from three different sources: structured data from the OSMI survey results and Google Trends, and semi-structured data from Twitter. Patterns were found in the data

using word clouds, sentiment analysis, time series, regression, random forest models, and correlation matrices. These methods achieve learning goals two and three.

The data analysis led to alternative strategies to support employee mental health, as well as alternative means of analysis based on the data that was available. For example, the OSMI survey data consisted of multiple years with different sets of questions, the questions needed to be evaluated for commonality, and the data combined to create one data set for the analysis. This alternative strategy for ingesting the data into Python achieved learning goal four.

Lastly, learning goals five and six were achieved through providing actionable insight for technology companies and human resource departments, then communicating these insights from an executive leadership perspective: happier employees are more engaged and productive, which leads to better profits. Actionable insights were made clear, for example: provide mental health coverage with the company insurance plan and provide employees with additional mental health resources such as events, webinars, newsletters, and external resources.

Conclusion

This portfolio has demonstrated the successful implementation of the seven learning objectives for Syracuse University's Applied Data Science program, as well as an understanding of the major practice areas of data science: business analysis, computer science, mathematics, and data analysis (IST 718). Data was collected using real-life data, Kaggle, and Twitter API, then organized in either SQL, R or Python (IST 695, IST 719, IST 718). Patterns in the data were identified using machine learning, visualizations, and statistical modeling techniques (IST 718, IST 719). From these analyses, alternative strategies and plans of action were developed, specifically regarding actionable insight for companies looking to support employee mental health (IST 718).

Communication skills were demonstrated when presenting to executives and human resource departments on the importance of employee mental health (IST 718), as well as to American sports fans who want to be more informed about the teams and leagues they support (IST 719). The insight from these analyses were presented in terms which the respective audience would understand, as to allow the audience to act upon these insights. Lastly, the ethical dimensions of data science were emphasized while studying the dark web and the effects of identity theft (IST 623). In addition, all analyses were performed with privacy at the forefront. For example, all data from the SQL database was anonymized as to protect individual and company privacy, and the OSMI data was provided without personally identifiable information. The projects outlined in this portfolio demonstrate the successful execution of the Applied Data Science learning objectives and the development of the necessary skills for practice in the field of data science.

Syracuse University's School of Information Studies allows student to develop the data analysis, business analysis, computer science, and communication skills necessary to provide actionable insight to a variety of audiences. The Applied Data Science program's focus on bridging the gap between business users and IT professionals is crucial when solving data problems, addressing business concerns, and improving organizational efficiency. In addition, the program encourages students to consider the ethical implications of data management and analysis as the amount of data available in our world continues to grow. Data scientists should remain diligent to protect individuals and companies' private information, as well as ensure the data used for analysis does not introduce bias into the model due to lack of representation.