Chiyoung Jeffrey Lii

Fall 2020 | clii01@syr.edu

Applied Data Science Graduation Portfolio

A demonstration of the mastery of program goals

Contents

[Explanation of Program Learning Outcomes 2](#_Toc56693725)

[1. Describe a broad overview of the major practice areas of data science. 2](#_Toc56693726)

[2. Collect and organize data. 2](#_Toc56693727)

[3. Identify patterns in data via visualization, statistical analysis, and data mining. 2](#_Toc56693728)

[4. Developing alternative strategies based on the data. 2](#_Toc56693729)

[5. Develop a plan of action to implement the business decisions. 2](#_Toc56693730)

[6. Demonstrate communication skills regarding data and its analysis 3](#_Toc56693731)

[7. Synthesize the ethical dimensions of data science practice (e.g., privacy). 3](#_Toc56693732)

[Link to GitHub Repository 3](#_Toc56693733)

[IST 687 Intro to Data Science Final Project 4](#_Toc56693734)

[Using the data science life cycle: 4](#_Toc56693735)

[Identifying patterns via statistical analysis: 4](#_Toc56693736)

[Identifying patterns via visualization: 5](#_Toc56693737)

[Implementing various data strategies: 5](#_Toc56693738)

[Creating a plan of action and communicating results: 5](#_Toc56693739)

[IST 659 Data Administration Concepts and Database Management Final Project 6](#_Toc56693740)

[Creating a plan of action: 6](#_Toc56693741)

[Collecting Data: 6](#_Toc56693742)

[Organizing Data: 6](#_Toc56693743)

[Communicating Results: 7](#_Toc56693744)

[IST 623 Introduction to Information Security 10](#_Toc56693745)

[Analyzing the ethics of cyber warfare: 10](#_Toc56693746)

[The deontological viewpoint: 10](#_Toc56693747)

[The utilitarian viewpoint: 10](#_Toc56693748)

[Communicating Results: 11](#_Toc56693749)

[Conclusion 12](#_Toc56693750)

[Works Cited 13](#_Toc56693751)

# Explanation of Program Learning Outcomes

There are seven major Program Learning Outcomes for the Applied Data Science (ADS) program at the Syracuse University School of Information Studies. Each of the outcomes below is something a student should be able to do once graduated from the ADS program.

## Describe a broad overview of the major practice areas of data science.

A student should be able to elucidate the data science life cycle. This includes business understanding of the requirements, data collection, data cleaning, explanation, feature engineering, descriptive and predictive modeling, data visualization, and presenting an actionable business insight. For example, a graduate of the ADS program should be able to be a proactive member of a Fortune 500 company’s data analysis team, by suggesting and competently executing a consumer analysis plan.

## Collect and organize data.

A student should understand how to collect and organize data. Data collection entails things such as finding and vetting reputable data sources, proper data gathering procedures, and proper data governance. Data organization includes understanding and implementing data storage strategies such as operational relational and non-relational databases, and data warehouses. For example, a graduate should be able to gather Google Ad data and store them in a database.

## Identify patterns in data via visualization, statistical analysis, and data mining.

A student should be able to create data visualizations, statistical models, and mining models, such as bar charts, linear regressions, and association rules. Students should also be able to recognize trends and patterns such as positive linear correlations, n-space clusters, and item sets. For example, give consumer data, a graduate should be able to describe which kind of consumer is mostly likely to buy what products.

## Developing alternative strategies based on the data.

Students should be able to select an effective strategy when presented a dataset for analysis. This include correct visualizations, such as using word clouds for word sets and box plots for numerical data, as well as statistical models, such as linear, logistic, and quadratic regressions. For example, given a qualitative and a quantitative dataset, a graduate should be able to select separate and appropriate models to analyze the datasets, such as association mining for the qualitative dataset, and a linear regression model for the quantitative dataset.

## Develop a plan of action to implement the business decisions.

Students should be able to write a business plan of action based on their analysis. This includes clear understanding of business goals, the ability to set goals and create timelines. For example, a graduate should be able to break down an overarching business goal such as ‘improve revenue by 10%’ into smaller, actionable goals, such as ‘discover new markets to expand into’.

## Demonstrate communication skills regarding data and its analysis

Students should be able to present their data analysis in an effective and pleasing manner for a wide range of audiences. This includes the usage of tools such as Microsoft Excel, PowerPoint, Tableau, and Github. For example, a graduate should be able to write a paper documenting the past two years of their life in a humorous manner.

## Synthesize the ethical dimensions of data science practice (e.g., privacy).

Students should understand ethical systems such as consequentialism, deontology, and virtue ethics, and be able to make correct choices based on such systems. For example, if given the order to violate federal law and use personal medical records to increase insurance premiums without cause on employees, graduates should refuse and report his violation to Health and Human Services.

# Link to GitHub Repository

Here is a link to the GitHub Repository containing all my final projects:

<https://github.com/clii01/AppliedDataScienceGraduationPortfolio>

# IST 687 Intro to Data Science Final Project

Link to final report:

<https://github.com/clii01/AppliedDataScienceGraduationPortfolio/tree/main/IST%20687%20Intro%20to%20Data%20Science>

For our IST 687 final project, we wrote a paper (Chiyoung Lii, IST 687 Final Project Report.docx, 2019) analyzing customer satisfaction in the airline industry. The dataset given to us contained roughly 26 columns, of both the quantitative and qualitative variety. Example columns include: Satisfaction (integer in a range from 1 to 5), Age (of the customer), and Class (of the ticket).

The follow sections explain the program learning outcomes demonstrated when we wrote the project.

## Using the data science life cycle:

Over the course of the project we followed the data science life cycle very closely. For example, to start off, we wrote several business questions to answer. These are listed below:

1. What factors contribute to Customer Satisfaction? How can they be changed to increased customer satisfaction?
2. What factors contribute to Customer Dissatisfaction? How can they be changed to reduce customer dissatisfaction?
3. Which factors are more important than others? Rank the factors by level of impact. Is there a general trend that can be gleaned from the data?
4. What is the 0.5 and 0.95 percentile of the satisfaction levels? How is this important when determining and evaluating what actions to take?
5. How do the factors interact? What happens when multiple factors are prevalent in a single experience?

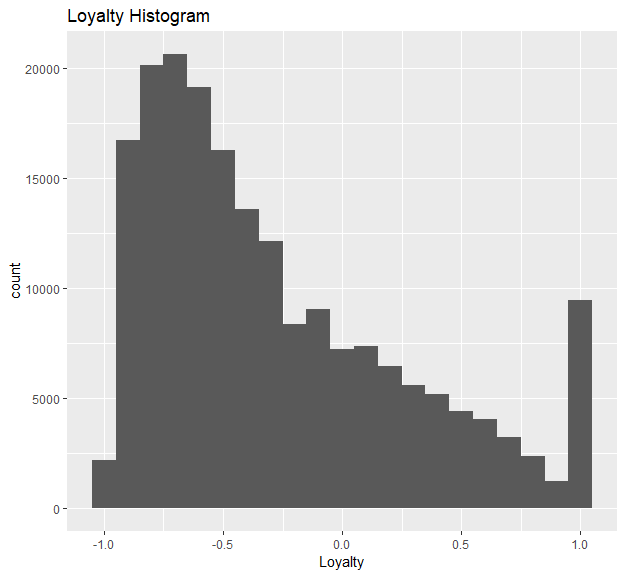
## Identifying patterns via statistical analysis:

Next, we applied descriptive statistics to our dataset. We loaded up the data set into R (Chiyoung Lii, IST 687 Final Project Code.R, 2019) and noted down the features with invalid fields. We did this via the summary() command and discovered the ‘person’ and ‘ecoplus’ and ‘flightcancel’ variables had NA’s. We also calculated the 5th and 95th percentile, the quartiles, and the mean and median of the numerical values. These statistics helped us understand and model the data set.

We also calculated the Pearson Correlation Coefficient (PCC) for the quantitative features. The PCC is a numerical value that shows the strength of the linear relationship between two numerical values. It ranges between one and negative one, with a higher absolute value indicating a stronger linear relationship. Using this value, we picked out several coefficients of interest to build our models on.

## Identifying patterns via visualization:

We also created various visualizations to help us with our exploratory data analysis. Below is an example, a histogram of the Customer Loyalty value (Chiyoung Lii, IST 687 Final Project Report.docx, 2019):



We also created boxplots and scatter plots for appropriate variables. These visualizations helped us understand the nature of the variables. For example, it is clear in the above histogram there are two peaks. This helped us decide on what kind of data strategy to use.

## Implementing various data strategies:

We explored several different data strategies for this final project. For example, we wrote several linear regression models with dummy variables, Support Vector Machine models and Association Rule mining models. These models helped us gather multiple insights into the dataset, such as the fact that loyalty is a stronger factor (0.23) than price sensitivity (0.09) when predicting customer satisfaction in a SVM model.

We considered how good the models were by looking at the R-squared values of the linear models and the confusion matrices of the SVM models. The R-squared values of the linear models were abysmal at 0.14, and the error rates of the SVM models were also quite high at around 20%.

## Creating a plan of action and communicating results:

Finally, after modelling and performance evaluation, we generated some actionable insights and conclusion to give to stakeholders. One of the ways we suggested to improve customer satisfaction was to improve customer loyalty using customer loyalty programs. Another was to reduce the effects of time delay (one of the major negative factors in predicting customer satisfaction). Our final suggestion was the reduce the effects of price sensitivity through a marketing campaign, as described by some marketing papers. These actionable insights were presented in a PowerPoint to the class and professor at the end of the class.

# IST 659 Data Administration Concepts and Database Management Final Project

Link to paper:

<https://github.com/clii01/AppliedDataScienceGraduationPortfolio/tree/main/IST%20659%20Data%20Administration%20Concepts%20and%20Database%20Management>

## Creating a plan of action:

For our IST 659 Final Project, we chose a topic to create a database and database interfaces for. For my project I chose to create a database that would provide stock market data for traders. I obtained my data from Yahoo finance. At the beginning of the semester, we were required to write a project proposal (Lii, Project Proposal 0.3.docx, 2020). Besides demonstrating a strong understanding of business values and requirements as part of the data science life cycle, the proposal contains a problem statement and a proposed solution that come together to demonstrate my ability to create a solid plan of action.

## Collecting Data:

For my project, I first looked for suitable historical stock market data from the internet. Eventually, I found a website called Alpha Vantage, which was backed by a well-known tech accelerator (Alpha Vantage Inc., 2020). I set up an account with them, and compared them to other sources to make sure their data was accurate, then downloaded a data set from them.

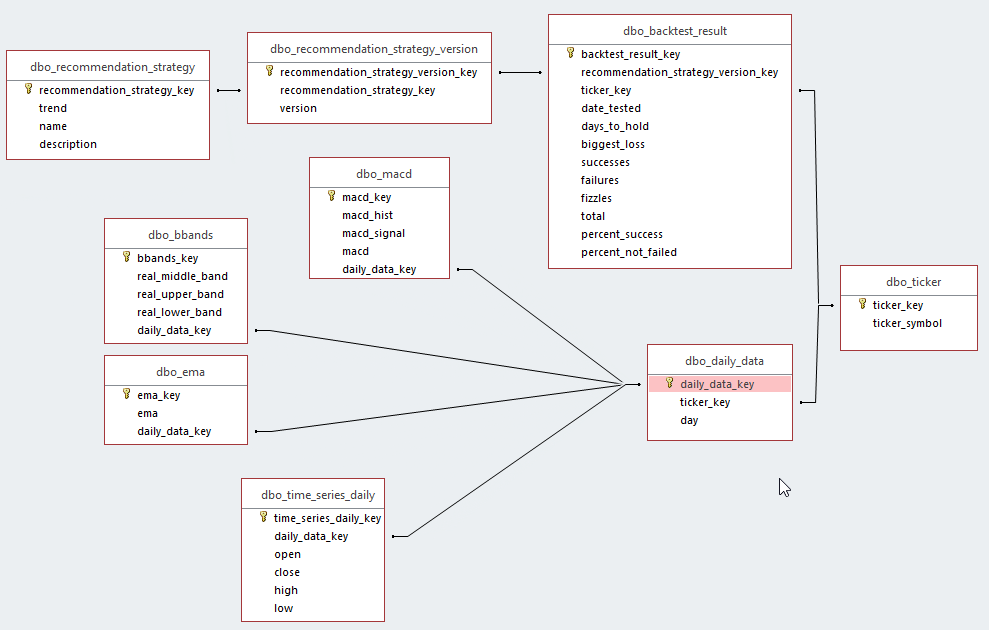
This data included the opening, closing, highest, and lowest price of a trading day in the form of time series data. I also grabbed several derived features from the Alpha Vantage API, including a day’s MACD, Bollinger band data, and Exponential moving averages.

## Organizing Data:

With the data in hand, I set about designing a database. First, I organized my database into historical and backtesting schemas. Next, I created a root table (dbo\_ticker) to store the ticker information. I partitioned the data into normalized forms and gave each different feature a separate table. The Entity-Relationship Diagram visualizing the schema I created is available on page 6 of my final project report (Lii, Project Final Report 0.2.docx, 2020), as well as below.

With the schema now created, I wrote a script to insert the data downloaded from Alpha Vantage into a SQL script. (Lii, insert\_data.sql, 2020). I then ran the data to insert the data into the database.

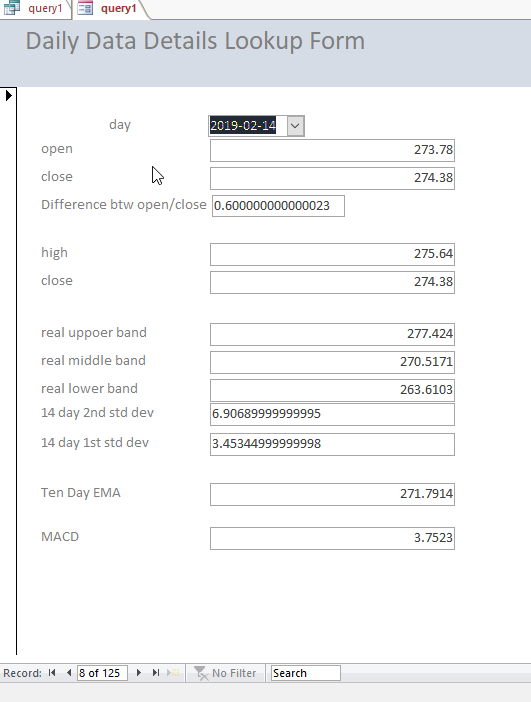
The final Entity-Relationship Diagram:



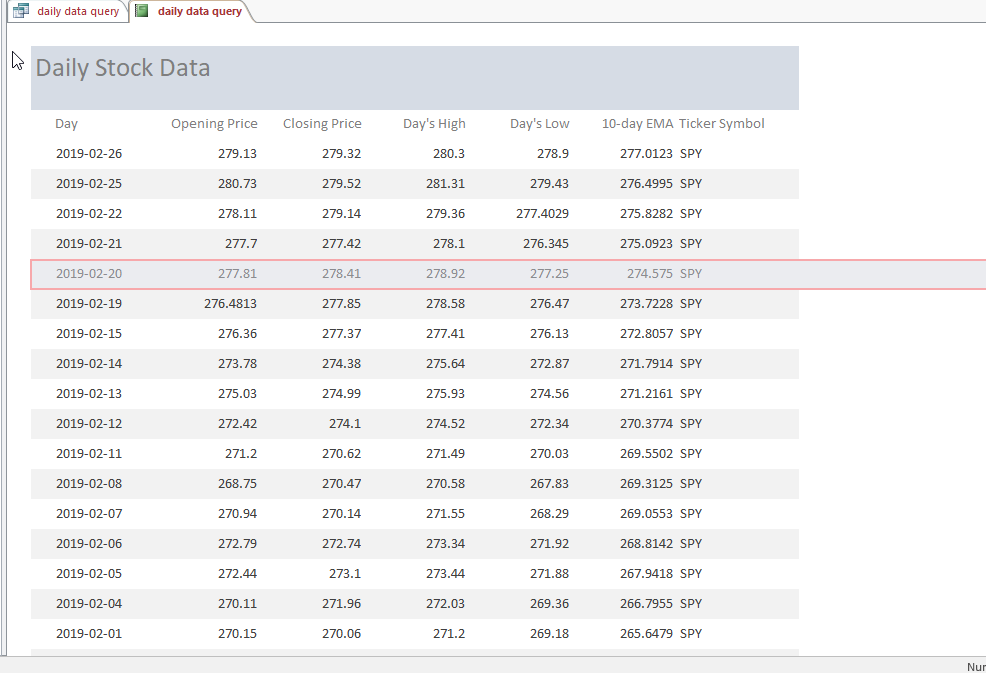
## Communicating Results:

As part of the final project, we were required to showcase our work with our classmates. We demonstrated our works in progress and gave recommendations while also receiving recommendations in return. We saved the recommendations and implemented them as part of our final report (Lii, Demo Recommendations.rtf, 2019).

Another requirement for the final project was the creation of forms and reports in Microsoft Access. The forms would allow people who were not well versed in SQL such as managers and salespeople to access the database through pre-written queries without having to learn SQL. An example of one of these forms is shown below:



This form allows users to look up a particular day’s data and modify it according to their needs. These forms are meant to be used in conjunction with forms that regularly sent data to users. An example of one of these reports is below:



In addition to showing the data over a period, the report also allows users to sort by ticker symbols, with many of the distracting fields removed. (such as standard deviation).

Through this Project I was able to practice and implement various ways of communicating my work in both data collection and organization to unfamiliar people in both personal and automated ways.

# IST 623 Introduction to Information Security

Link to paper: <https://github.com/clii01/AppliedDataScienceGraduationPortfolio/tree/main/IST%20623%20Introduction%20to%20Information%20Security>

For our IST 623 Final Project we wrote a paper analyzing modern cyber warfare (Siobhan Dunn, IST623 Cyberwar.docx, 2020).

## Analyzing the ethics of cyber warfare:

For this final project, we studied several implementations of modern information technology infrastructure. The three areas we studied were Cyberterror (how the ISIS uses the internet for recruiting and proliferating propaganda), Cyberwar (how the US was attached by Russia during the 2016 elections), and a couple of international computer viruses (NotPetya and Stuxnet).

## The deontological viewpoint:

Deontology is a system of moral philosophy that judges the correctness of a particular action based on whether it violates a set of rules. In the IST 623 paper we cover how the ISIS justifies their actions through a set of rules (Sharia law).

The ISIS believe that that all societies in existence that do not abide by Sharia law are unjust, immoral, unholy, and must be destroyed (Siobhan Dunn, IST623 Cyberwar.docx, 2020). They promote their actions by contacting potential recruits on social media and convincing them to join their army.

Fortunately, several groups have started using sophisticated programs to track recruitment efforts (Siobhan Dunn, IST623 Cyberwar.docx, 2020), in accordance with US law, another set of rules. Deontological frameworks are allowing people to do more with modern data science techniques.

## The utilitarian viewpoint:

Utilitarianism is another moral system that judges the correctness of an action based on either or not they benefited more people than it does not. Both systems were covered in IST 618 Information Policy. I use utilitarianism here to discuss the ethical aspects of Stuxnet, a computer virus covered in my IST 623 paper.

As discussed in the paper, Stuxnet was unique in that it was the first virus to cause physical damage. All the viruses up to the year it was discovered (2010) had been limited to the systems that it had been on. Stuxnet however, directly attached the tools of the Iranian nuclear program, destroying centrifuges by changing the rotation speed to dangerous frequencies.

Proponents of Stuxnet argue that it was justified, as Stuxnet only broke tools and neither hurt or killed anyone, in comparison to Operation Opera, a military strike on a research site that killed 11 soldiers and civilians (Singer, 2015). Therefore, going by the above definition of utilitarianism, Stuxnet, and weapons like Stuxnet is better than traditional operations because it kills less people. However, I speculate that the programmers of Stuxnet attempted to restrain themselves when writing the virus to limit damage. If a malignant actor decided to write a program to cause maximum damage, the concept of computer virus that can cause physical damage could cause thousands of deaths that would not have occurred otherwise, which is a strike against it from a utilitarian viewpoint.

## Communicating Results:

After writing the paper, our group gave a presentation where we explained the contents of the paper to our entire class (Siobhan Dunn, IST623 Assignment 2 Presentation Chiyoung Lii Julia Fontana Dennis Timmons Siobhan Dunn.pptx, 2020). This presentation demonstrated our ability to communicate our analysis results to a small to medium sized group of people.

# Conclusion

In conclusion, the three final projects I have covered above, (Chiyoung Lii, IST 687 Final Project Report.docx, 2019) (Lii, Project Final Report 0.2.docx, 2020) (Siobhan Dunn, IST623 Cyberwar.docx, 2020) clearly demonstrate my mastery of the ADS program learning goals.

Throughout the course of the Applied Data Science Program, not only have I learned the ins and outs of the data science life cycle, I have also mastered the application of the seven program learning outcomes, and repeatedly demonstrated this mastery through the authorship of multiple final projects every semester, for a wide variety of subjects and situations.

The IST 687 paper shows that I have the ability to describe and implement the data science lifecycle, identify patterns in data via visualization and statistical analysis, implement various data strategies, and develop a plan of action to implement insights gained from analysis.

The IST 659 paper shows the ability to create a plan of action, follow it through, collect and organize data, as well as communicate the results to a wide variety of people, via Microsoft Access forms and reports.

Finally, the IST 623 paper shows my ability to synthesize the ethical dimensions of data science practice via the analysis of cyber warfare, cyber terrorism, and computer viruses.

In the future, I look forward to applying these learning goals in a professional setting ethically, and efficiently.

# Works Cited

Alpha Vantage Inc. (2020, November 18). *Free Stock APIs in JSON & Excel*. Retrieved from Alpha Vantage: https://www.alphavantage.co/

Chiyoung Lii, D. A. (2019, November 16th). *IST 687 Final Project Code.R.* Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20687%20Intro%20to%20Data%20Science/IST%20687%20Final%20Project%20Code.R

Chiyoung Lii, D. A. (2019, November 16th). *IST 687 Final Project Report.docx.* Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20687%20Intro%20to%20Data%20Science/IST%20687%20Final%20Project%20Report.docx

Lii, C. J. (2019, November 18). *Demo Recommendations.rtf*. Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20659%20Data%20Administration%20Concepts%20and%20Database%20Management/Demo%20Recommendations.rtf

Lii, C. J. (2020, November 18). *insert\_data.sql.* Retrieved from insert\_data.sql: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20659%20Data%20Administration%20Concepts%20and%20Database%20Management/insert\_data.sql

Lii, C. J. (2020, November 16th). *Project Final Report 0.2.docx.* Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20659%20Data%20Administration%20Concepts%20and%20Database%20Management/Project%20Final%20Report%200.2.docx

Lii, C. J. (2020, November 18). *Project Proposal 0.3.docx.* Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20659%20Data%20Administration%20Concepts%20and%20Database%20Management/Project%20Proposal%200.3.docx

Singer, P. (2015). Stuxnet And Its Hidden Lessons On The Ethics of Cyberweapons. *Case Western Reserve Journal of International Law*, 79-84.

Siobhan Dunn, J. F. (2020, November 18). *IST623 Assignment 2 Presentation Chiyoung Lii Julia Fontana Dennis Timmons Siobhan Dunn.pptx*. Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20623%20Introduction%20to%20Information%20Security/IST623%20Assignment%202%20Presentation%20Chiyoung%20Lii%20Julia%20Fontana%20Dennis%20Timmons%20Siobhan%20Dunn.pptx

Siobhan Dunn, J. F. (2020, November 16th). *IST623 Cyberwar.docx.* Retrieved from AppliedDataScienceGraduationPortfolio: https://github.com/clii01/AppliedDataScienceGraduationPortfolio/blob/main/IST%20623%20Introduction%20to%20Information%20Security/IST623%20Cyberwar.docx