Master of Science in Applied Data Science

Portfolio Milestone

Syracuse University

School of Information Studies

Joe Hernandez | SUID #627926839 | jahern01@syr.edu

**Executive Summary**

**Objective:**

Reflect abilities specified in program learning outcomes with selected assignments and projects worked on during the M.S in Applied Data Science program at Syracuse University School of Information Studies.

**Summary:**

This graduation portfolio reflects the learning goals I met after completing all 32 semester hour courses for the M.S. in Applied Data Science Program at Syracuse University School of Information Studies. My organization’s needs for a data science professional requires me to gain comprehension in the collecting, processing, presenting, and discussing of information and their outputs for environmental studies and civil works construction projects for economic justification. Currently, the methods of acquiring data are limited by the format that is available and the inflexible software applications ability to process data limiting integration of inputs from other disciplines. Throughout the course of the program, I have presented the goal to my organization to leverage the lessons and skills with the information collected to determine effective data science strategies with the flexibility of coding to develop our statistical insight, machine learning, and automation to improve the way we process and convey our findings for projects.

The courses available provided both comprehension and application of data science concepts through statistical inference, scripting, and communication. Learning the fundamentals of the data lifecycle and the database development cycle helped me understand how to effectively prepare data for processing and practical exercises with scripting deployed effective and creative methods for me to produce meaningful output. Understanding effective interpretation, communication, and visualization of data aided me in conveying my findings to the target audience with comprehensive language for them to make the best decisions. Each course shared in value a distribution of the learning goals of the Applied Data Science Program the focus from all of them centered around the following: identifying data problems and user needs, performing computational scripting, modeling data, and determining appropriate statistic, visual, and algorithmic techniques. After completing 32 semester hours, I spent time digging deeper into solidifying my comprehension of data science, statistics, and scripting where the time was spent reading, coding, and interacting to learn more about this emerging discipline.

This report shows my knowledge and confidence in the Data Science field gained from Syracuse University School of Information Studies. The skills gained helped me collect, process, and present data with a greater understanding of statistical inference and the techniques used to share my findings. The projects provided me resolve to continue to pursue data engineering while maintaining the data science flexibility in a project management role, where, after graduating from the program, I plan to use many of the data science solutions at my organization through expanded statistical and scripting literacy and assembling a data science team to improve our analysis of projects. Overall, I learned a significant amount from the Applied Data Science program with the value of using data to explain output through storytelling as the major takeaway and I am grateful for the content, personnel, and professionalism the Syracuse University School of Information Studies has provided.

**Resume:**

My latest resume can be found in the following link with a copy found in the main report:

<https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone/blob/main/2.%20Curriculum%20Vitae/Resume%20Joe%20Hernandez%202022.docx>

**Author Information:**

* Author: Joe Hernandez
* Repository: <https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone>
* SUID#: 627926839
* E-mail: jahern01@syr.edu
* Program: M.S. in Applied Data Science, Syracuse University, School of Information Studies (iSchool)

**Referenced Projects & Aligned Learning Goals:**

**Course** **Name**

Statistical Methods in Information Science and Technology

Data Mining

Scripting for Data Analysis

Financial Analytics

**Project Overview & Learning Outcomes**

**Course Assignments**

Major practice areas in Data Science; Collect and organize data; Identify patterns in data; Develop alternative strategies based on data; Develop plan of action to implement business decisions from analyses; Demonstrate communication skills regarding data and its analysis

**Kickstarter Project Success Project**

Major practice areas in Data Science; Collect and organize data; Identify patterns in data; Develop plan of action to implement business decisions from analyses; Demonstrate communication skills regarding data and its analysis; Synthesize ethical dimensions of data science practice

**San Francisco Human Waste Map Project**

Major practice areas in Data Science; Collect and organize data; Identify patterns in data; Develop alternative strategies based on data; Develop plan of action to implement business decisions from analyses; Demonstrate communication skills regarding data and its analysis; Synthesize ethical dimensions of data science practice

**Course Assignments**; **PGE Infrastructure Improvement Analysis Project**

Major practice areas in Data Science; Collect and organize data; Identify patterns in data; Develop alternative strategies based on data; Develop plan of action to implement business decisions from analyses; Demonstrate communication skills regarding data and its analysis; Synthesize ethical dimensions of data science practice

**Skills**

R, Statistical Inference, Data Cleaning & Wrangling, dplyr, BEST, BayesFactor

Concept description, Association rule mining, Classification, Clustering

Python, Pandas, Matplotlib, Numpy, Plotly, Matplotlib, Plotly.io

Shiny, knitr, ggplot2, xtable, MASS, R markdown, ACF, PACF,

flexdashboard

**Objective:**

Reflect abilities specified in program learning outcomes with selected assignments and projects worked on during the M.S in Applied Data Science program at Syracuse University School of Information Studies.

**Context:**

This Graduation Portfolio Milestone reflects meeting the learning goals after completing 32 course hours for the M.S. in Applied Data Science program at Syracuse University School of Information Studies. The program was brought to my attention by a work colleague who recommended a modernization in how our profession collects, processes, presents, and discusses information to decision makers. As an Economist at the U.S. Army Corps of Engineers, I collect data and present the relevant information to inform whether a project is economically justifiable. Some major hurdles my colleagues and I dealt with regarded sourcing the information and effectively extracting the data from embedded sources where, during my early days, the most optimal way to insert information was to manually record every cell in a table. Additionally, the provided software applications available were single method ways to process data and required specific formatting and file types to process statistical information. The statistical information we use determines potential damages to structures, National Economic Development, and critical infrastructure, quantified by dollars spent to restore the area, as well as, life loss, quantified by number of people trapped, within a floodplain. We saw potential in the Applied Data Science Program at Syracuse University as an opportunity to expand our capabilities and improve our organization’s practices and integrate our capabilities in Economics, Hydrology, and Geology to collect data from various departments, process the data into a usable format, present the data with improved visualizations, and to address the data we show in an understandable format for decision makers both within the organizations and those in their local communities can understand the recommendations made by myself and my colleagues without any obfuscation. The projects I present, will summarize the process toward reaching the output as well as the observations and recommendations explained and presented to the target audience.

Throughout the program the courses have provided me comprehension and ability of Data Science concepts and practice. I learned the fundamental concepts of data and databases including their life and development cycles. The scripting and coding exercises have helped in preparing data through screening, cleaning, and linking data to develop effective insights from improved data management tools and techniques. Regardless, the only way to make sense of the output as well as the development process is through effective communication by recommending and justifying strategies for managing data and professionally presenting clear results with reception to critique. From an ethical perspective, the data collected, analyzed, and presented must comply with digital and information policies for an objective and unbiased conclusion. Through course assignments and projects, I was able to, first, identify data problems and user needs through discipline-specific means and noticing any further information requests prior to data preparation. Second is performing computational scripting, collecting, and developing the data into usable data structures and elements to include scrubbing, debugging, and manipulating data. Next is processing and modeling relationships between data, thus, transforming the data with linking, aggregating, summarizing, and searching methods to meet the information needs. Lastly, the courses in this program provided me the skills to determine the appropriate techniques to analyze data by designing data analyses using data modeling, normalizing, and statistical techniques. To explore the data, I must use quantitative techniques using descriptive statistics, summarization, and visualizations applied throughout the program to interpret and analyze data models and the analyses and findings to make practical decisions regarding the reliability and accuracy of the sources, processing, and output. With full comprehension and ability from the Applied Data Science program courses, I have the capacity to present and communicate my findings and provide effective data and recommendations to users, clients, and decision makers.

While I finished my courses at the end of Winter 2020, I took the time to familiarize myself with Data Science, R, and Python further. The reading material that helped in this pursuit for further knowledge came from series such as the For Dummies and O’Reilly Cookbooks provided simplified summaries of advanced concepts and additional practical exercises in conjunction to the reading material suggested in the program courses. Additionally, practical books regarding statistical literacy and manipulation supplemented the transmission and reception of data output through verbal, visual, and written communications. From a networking standpoint, I follow various social media topics related to Data Science, R, and Python. Additionally, there are many interactive sources to further supplement data science and statistics comprehension such as videos and academic sites I have used as well such as Kahn Academy and DataCamp. Throughout the program, I have presented many of my Data Science findings and skills to my workplace proposing Data Science solutions such as Dashboard interfaces, 3D visualizations, and alternative collecting and processing solutions to automate some processes even for an inferential observation to see if these innovations match up with the software applications we use. Overall, this program has enabled me to assist with the improvement of my organization with the buildup of a Data Science Team in the short-term future.

**Course Assignments and Projects**

**Statistical Methods in Information Science and Technology (Course Assignments)**

<<https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone/tree/main/3.%20Course%20Material/Statistical%20Methods%20in%20IST>>

To start the portfolio off, I wanted to begin with demonstrating my knowledge of inferential statistics through data science analytics with R as the data management platform. The practice skills within the course assignments helped me familiarize myself with statistical vocabulary and concepts in the realm of descriptive statistics. Additionally, the scripting exercises in RStudio helped me manage data and execute calculations and visualizations. The exercise begins with uploading data sets and producing basic statistical summary for island land mass data which is followed by the visualization of the distribution for the weight of chicks with a histogram. Both examples show a Poisson distribution where the bulk of the data is in a positive skew. The following assignment, I looked at the basic probability of flipping a coin, where, based on the outcome of seven heads and two tails replicated 100,000 times would be reproduced only 6,980 times based on the set seed to replicate in a normally distributed bar plot. Additionally, from statistic test and home repossession data sets, I created contingency tables to explain the marginal pass/fail between high school and college students and Repo/No Repo ratio for homes in the United Kingdom. The marginal outputs were isolated and normalized to see the probability of each event only to find out that the high school passing rate of statistics testing is 66% and that approximately 1.138% of customers who fail the screening test default on their mortgage. The next assignment uses the chick weight data set to explore the sampling distribution over the long run with 10,000 replications of events transforming the graph and statistics of the distribution from a positive skew to a skew approaching zero explained by increasing the number of samples from the chick weights from the raw data and finding the average for each set. In the next assignment, the data from plant growth was used in performing statistical inference from a small set of raw data where the summary statistics shows, when compared to the control group, applying Treatment 1 created larger variability but a tighter interquartile range for less growth and applying Treatment 2 lessened the variability and had an interquartile range tighter than both the control and Treatment 1 with higher growth then running a Two Sample t-test for a confidence interval around the mean difference between each treatment and the control. The assignment after continues with the plant growth analysis analyzing the t-test upper and lower bounds taking careful note that the interval may or may not contain the true population value and comparing it to the Bayesian approach of using prior data through Markov chain Monte Carlo simulation. I found that between the t-test and the Bayesian approach, the t-test displays information that aids but not proves the significance in the values whereas the Bayes approach provides information based on density intervals and can be used without caveats of uncertainty. For the next assignment, I run an analysis of variance to interpret data on insect spray using the variance to understand the mean differences where we find the means overlap between three sprays in two sets. In the next assignment, I ran a correlation test on rock samples through Pearson product moment and Bayes Factor methods for comparison followed by chi-square data collection and analysis for data regarding college admission between genders to compare the proportion distribution disparity between them. In the following exercise, I ran multiple regression and linear prediction for car data create a prediction equation influencing the miles per gallon as an output with frequentist and Bayesian hypothesis testing using between- and within-group variance and F-distribution. The penultimate assignment continues the use of the car dataset to determine the quality of a model using logistic regression finding horsepower as the better predictor coefficient. The use of Bayesian estimation of logistic regression was applied to a Chilean voting dataset and determined that age was a minor coefficient that affected the election result of retaining Chilean authoritarian leader Augusto Pinochet in power. Lastly, the final assignment explored data from previous datasets with the inclusion of a new data set for air passengers where a time series analysis was created to explore cyclical patterns and explore the nonindependence between observations. The feedback from the course proved fruitful in assessing my ability to communicate results, where, in future analyses I will express a deeper reasoning interpreting distribution visualizations and uncertainty using concise wording. Overall, this course laid out some great groundwork integrating statistics and coding where the course assignments allowed for me to perform an independent analysis with diverse data and use inference to explain the data and the results from the output.

**Data Mining (Kickstarter Project)**

<<https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone/tree/main/3.%20Course%20Material/Data%20Mining>>

I transition over to a project from the Data Mining course to work more with real-world data from a repository and provide a business level observation. Throughout the course, there was an emphasis on applying data mining methods to extract, translate, and convey knowledge from the data through data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis to a business problem. For the project, I proposed the problem regarding whether the crowdfunding platform Kickstarter should continue to support unsuccessful categories that fail to meet their funding goal by finding the criteria that allow for successful categories to meet or exceed their funding goal. This analysis uses descriptive statistics, predictive analysis, visualization, and data mining techniques such as classification and clustering to make that determination. In the analysis, my date preparation plan began with loading the data and filtering the dataset to campaigns launched in 2016 as it had the most complete year of records of approximately 50,000 campaigns and observing the attributes from Category, Main Category, Goal, USD Pledged, State, and Backers. This ensured that the data used was collected, used, and displayed ethically for simply analysis purposes. When exploring the data for anomalies, missing entries, and other rarities within the records, I decided the most simplified solution to handle them was by exclusion as much of the data with anomalies had missing USD Pledged entries that could not be retrieved and is a critical determinant for the success of a campaign. The cluster analysis used kMeans to prevent overlapping with campaigns put into one of 15 clusters where I could find the means for the goal, pledged, backers, USD pledged, and the profit. I noticed a high sum of square error and, after plotting a linear model, noticed the shape created between components that would explain only 87.07% of the point variability. Another technique was used to explore this data is through decision tree analysis using J48 creating a tree diagram categorizing the data based on Kickstarter campaigns producing correctly classified instances and a confusion matrix to identify those campaigns that were canceled, failed, live, successful, and suspended. Lastly, applying code to predict the classified state of campaigns which categories would become successful. The conclusion from this project showed that using kMeans helped look at the data based on grouping while the decision tree helped look at the data based on classification with the latter showing predictions for the success and failure of campaigns based on their main category. The feedback received was positive with a request for more discussion regarding the decision tree analysis to better illustrate the conclusion and will continue to pursue with more execution in that type of analysis. Overall, the course and project were helpful in assembling data and using clustering analysis with the integrated use of summary statistics.

**Scripting for Data Analysis (San Francisco Human Waste Map Project)**

<<https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone/tree/main/3.%20Course%20Material/Scripting%20for%20Data%20Analysis>>

In this course, I was to use access information and process different data types to solve problems with data wrangling by writing scripts that would transform data for analysis and visualization for effective data summaries. For the project, I would demonstrate my ability by analyzing the human waste reporting and responses to allocate cleanup personnel for expedited cleanup in critical streets and neighborhoods within San Francisco. This analysis stemmed from a web article exploring all the human waste reports in San Francisco in a graphic, which, plotted nearly every spot in the city. This article was a starting point that helped me generate three questions that the city would need to explore when addressing this issue through sanitary work. First, I wanted to know what the reporting frequencies were based on times of day, neighborhood, and street followed by the long-term trend in human waste reporting with the final question of what reporting integration strategies has the city implemented from this data. The data used in the analysis came from San Francisco’s 311 non-emergency reporting open-source csv file that is constantly updated with increasing record numbers with over 10 years of data that the city encourages to use for data science analyses. This would explain the graphic from the article essentially covering the city with plots where my analysis strategy would look at the data by dissecting it based on time of day, year recorded, neighborhood and street. The biggest challenge from this project was the use of Python and the descriptive statistics and plotting packages. Loading the data into Python took quite some time as this uses big data with the file size in gigabytes, but filtering helped limit the number of entries by using records in the Human Waste category and excluding incomplete entries where the dispatcher did not provide complete information for the date and time of both reporting and completion. Additionally, the attributes in some categories required some adjustments to match format such as date and time and latitude and longitude coordinates using floats with the inclusion of an additional attribute called Duration to account for the time it took from reporting to cleanup completion. This analysis used descriptive statistics observing the trends using the population data to capture all the records to develop a trend analysis of the time of day and a time-series analysis for a yearly observation. As a result, I produced a map plotting of human waste reports based on time of day, a raw frequency bar chart of these reports by both neighborhood and street, descriptive statistics of response time duration for each year with a boxplot visualization for the beginning, midpoint, and end year observations, and a time-series plot of different reporting methods daily and a cumulative trend graph showing the diverse methods the city uses to report human waste. The conclusions drawn from this project showed that even though the daytime hours accounted for most of the reporting, the morning hour reporting can be explained by the lack of activity and adequate lighting at night as the night reporting is extremely low. Additionally, nearly all the top 25 streets with the highest reporting frequency were also among those in the top three neighborhoods with the highest reporting frequency. Each year there has been a declining trend in the average and median response time duration and is transforming into a positive distribution. Lastly, while the city has landline phones as the dominant source of reporting, mobile technology and the internet has become increasingly popular, and the innovation allows for graphics to be integrated. When I presented this topic to the class my peers enjoyed the presentation along with the interactive map that presented information about each plot while sparing them the graphical elements of the newer reporting methods for reports with pictures. At the very least, I did provide links to the pictures because graphics like that do tell the story and leave a lasting impact at how serious this matter is for the city. Overall, I enjoyed this project pursuing a deeper explanation of the original article using statistics and visuals and has encouraged me to pursue visualizations like this at work for Flood Risk Management studies.

**Financial Analytics (Team Projects & PGE Project)**

<<https://github.com/JoeAHernandez/SU-MSADS-PortfolioMilestone/tree/main/3.%20Course%20Material/Financial%20Analytics>>

The final course consists of two sets of projects with one focusing on team project assignments and a personal project. The course leveraged much of the lessons from previous courses throughout the program with an emphasis on data analysis and mathematical thinking and modeling into a managerial finance perspective with the focus on solving business problems such as cash flow, present value, risk and return, portfolios, and markets. All the projects were assessed based on how well organized the summaries were regarding prose and position, the precision of numerical summaries accounting for any uncertainty, informative visualizations, formatted code that is organized for others to understand, clear explanations of models and rationale, and answering the analytical questions with reasoning from the results. For the team projects, I rotated with my colleagues as the data engineer, data scientist, and business analyst diversifying my role and supporting and learning from them. The first team project we used descriptive statistics to review heating oil price volatility that is impacting earnings answering questions regarding the nature of heating oil returns and the range of those returns by direction and the frequency in positive and negative movements. To further characterize the movement additional questions to visually show the differences in the shape of ups and downs of heating oil based on risk tolerance and the regularity and reliability of the price movements and a simulation of future movements using predictive analysis were asked. To find the nature of heating oil returns we used log returns as the best method to capture the volatility where we identified a few outliers that upon further investigation was explained by major current events at the time. When converted to absolute value and compared to the returns it shows that volatility had a period of little volatility but recently it has risen. Based on the descriptive statistics the measures of central tendency showed the change in heating oil price was close to 0 with a negative skew that is heavily influenced by the outlier. Through a pivot table we looked at the varying size of returns where 47.52% were negative, 48.86% were positive, and 3.63% were zero expressing the volatility of heating oil. We used cumulative relative frequency where we found that our tolerable risk is 3.6% and we can regularly and reliably analyze heating oil price movements by creating upper and lower bounds for predicting earnings. As a result, our findings suggest that the company should consider hedging heating oil price volatility by procuring another commodity and purchase between commodities based on their business cycle. For the next team project, we explore the exposure to exchange rates where our functional currency is in USD with customers based in the UK, EU, and Japan where our cash flow is affected by the volatility of exchange rates. Similar questions regarding the nature of exchange rates and their up and down movement, descriptive statistical insight, and risk tolerance are asked with the addition of knowing the exchange rate history correlation and how they are related to volatility. For the same questions a similar approach was used like from the previous project with changes to the visualization showing percent changes in the exchange rates between the USD and each currency to show volatility for each except the Chinese yuan which is pegged along with the use of autocorrelation to visualize the lag between the USD and each currency. We determined the tolerable risk to be at 1.47% in a similar format to the previous project. To explore the history of the exchange rate market correlation we used lag to find that at plus and minus 18 the correlation would appear more positive than negative. With regards to the correlation between USD and other currencies we found through fitting a correlation between the USD with currencies like the Euro or GBP but show no correlation between currencies such as the Chinese yuan and Japanese yen. Based on these findings we recommended to the company to keep a diversified portfolio of currencies to limit volatility and have available capital. The last two team projects explore a company’s entry into the nickel market with us answering their business questions regarding the timing affecting the size of shipping, the value of the new shipping arrangements affecting value with customers and allocating current resources to the new market. Within the dashboard, the team ran the descriptive statistics and correlations to explore the metals market and see how the metal commodities associate with one another, then added an analysis with visualizations regarding expected shortfalls and value at risk for further data insights. From the analysis we concluded that the autocorrelation showed a two-unit lag on each of the metals market and trade would be based on that specific metal. Additionally, running a correlation for shipping cost, commodity demanded, and commodity volatility can help explain the value of the business and finding a partner already in this commodity market and transportation market can help hedge off the effects of the market cycle spreading the risk between the partnership. In the final team project, we assembled the exploratory descriptive statistics analysis into a single tab in the dashboard and added exclusive tabs to assess the market risk of metals using rolling correlation, within sample correlation and volatility, and in separate tabs, performed confidence and risk measures and Markowitz tangency modeling. The team opted for the Markowitz tangency model since we are new to the nickel market with a recommendation to invest more into nickel once the business is more familiar with the market. As a result, we found that the nickel market is more volatile than the copper and aluminum market suggesting less than 10% of capital invested into the nickel market.

Finally, I exercised my personal comprehension of the course performing a similar analysis to that of the metals market, which, integrated the lessons throughout the course in a single dashboard. At the time, California was experiencing severe wildfire risk with Pacific Gas and Electric, the energy utility company, accountable due to their poor maintenance of their infrastructure. This negligence resulted in the company forced to payout to customers for damages and declare bankruptcy. In the rebuilding a massive infrastructure renovation is required with me exploring the various commodity markets to analyze and recommend their investment portfolio for the energy and metals market and the risks of these new markets. The business questions I investigated regarded the timing procurement for the purchase and sale of the commodities, managing the allocation of their existing resources, the impact from the wildfire liability in future investments, value of the energy market, and scalability in allocating resources to the customers. After the background information, the following three tabs looked at commodity prices for electricity, copper, and natural gas with additional insight on the company’s stock price and treasury bill as the last item is the safest asset. I added features such as faceting to get a better visualization of the value changes of each commodity with consideration toward percent changes in size and direction of the change, a correlation matrix plot with numerical values of how strong or weak the correlation is between commodities and added a slider to adjust risk measure quantiles if more risk would like to be taken producing graphs with adjusting expected shortfall and value at risk for each commodity. Observing the market risk, I looked at the rolling correlation of each commodity and found an inverse correlation between copper and electricity and a positive correlation between electricity and natural gas along with copper and natural gas. I ran a Markowitz model for optimization to find that the company should not short on any of the commodity trading. The conclusion I reached in this project showed the allocation of working capital heavily into the copper market, buying in to more electricity, and selling more from natural gas to implement a procurement polity of these commodities.

The feedback from all my projects emphasized applying skills for data insight and business summarization associated to the figures for both observation and clarity with an emphasis on numbers of interest to address cash flow and to specify the models picked and mention any similarities between models with and without constraints. The projects in this course integrated all the program coursework while using innovative procedures at presenting data in a dashboard which I presented to my supervisors gaining their interest in developing something like this in our operations.

**Conclusion**

During the years spent honing my skills in Data Science, this reflection serves as an accurate testament conveying my aptitude and confidence toward the M.S. Degree in Applied Data Science. The acquired skills in R and Python served as a great introduction toward data engineering. From a Data Science perspective, complementing the programming languages with lessons on statistical literacy served as a boon to effectively communicating the quantitative elements in the output and providing understandable visualizations. Additionally, the program contributed to practical skills such as teamwork by understanding different perspectives both personally and professionally with collaborating strategies to achieve the same goals. Lastly, the analytical thinking skills from the program expanded my personal understanding with interpreting and explaining data in addressing correlation of inputs, externalities affecting calculated predictions, and uncertainty. Overall, the collective skills gained from the program allowed for me to observe and predict from my models; however, accept the real-life outcomes with consideration to improve in the process.

Throughout the course of the program, I regularly informed my supervisor about the courses and the similarities I discovered with our methods of handling and processing data. We developed a modernization team with the focus toward enhancing programming and algorithmic prowess within the discipline. I have hosted workshops to enhance literacy to better understand coding and effective notation when developing code. A medium-term goal we have is to develop a dashboard analysis as a one stop shop location for Flood Risk Management modeling while extracting the necessary data from open-source archives regardless of format. Ultimately, we want to develop the team where their time allocated is toward interpreting data output, improving visualizations, and developing programming expertise to better automate our reports. In this Data Science role, I feel comfortable since participating in the program and the supplementary material and experience where I can take on roles as a Statistician, Data Scientist, or Business Analyst when working with a Data Science Team. I have networked with organizations such as the U.S. Forestry Service and the Hydrologic Engineering Center for a better understanding of their data management strategy and the lessons learned when proposing and implementing data driven goals.

Ultimately, the Applied Data Science courses taught me the importance of collecting and processing data and the interpretation and storytelling, thus, allowing others to follow along with my vision. The program was very flexible with the course layout and patient in the duration of the program. I look forward to conveying my competence as a representative alumnus of Syracuse University and the School of Information Studies.

**References/Bibliography**

Stanton, Jeffrey M. *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*. The Guilford Press, 2017.

Severance, Charles R. *Python for Everybody: Exploring Data Using Python 3*. CreateSpace Independent Publishing, 2016.

Tan, Pang-Ning, et al. *Introduction to Data Mining*. Pearson, 2015.

Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.

**Supplemental Reading Bibliography**

Pierson, Lillian, and Jake Porway. *Data Science for Dummies, 2017*. John Wiley and Sons, 2017.

Vries, Andrie de, and Joris Meys. *R For Dummies, 2nd Edition*. John Wiley & Sons, 2015.

Schmuller, Joseph. *R Projects for Dummies*. John Wiley & Sons, 2018.

Schmuller, Joseph. *Statistical Analysis with R for Dummies*. John Wiley & Sons, 2017.

Bari, Anasse, et al. *Predictive Analytics for Dummies*. John Wiley and Sons, 2017.

Dubossarsky, Eugene. *Data Science Strategy for Dummies*. John Wiley & Sons, 2019.

Bruce, Peter C., and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly, 2018.

Teetor, Paul. *R Cookbook*. O'Reilly, 2011.

Chang, Winston. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly, 2020.

Gillespie, Colleen, and Robin Lovelace. *Efficient R Programming: A Practical Guide to Smarter Programming*. O'Reilly, 2017.

Paulos, John Allen. *Innumeracy*. Viking, 1989.

Gigerenzer, Gerd. *Calculated Risks: How to Know When Numbers Deceive You*. Simon & Schuster, 2002.

Huff, Darrell, and Irving Geis. *How to Lie with Statistics*. W.W. Norton & Co., 2006.

**Resume**

**Joe Alexander Hernandez**

6671 Topaz Dr., Pollock Pines, CA 95726 | 786-422-4667 | [joe.a.hernandez@usace.army.mil](mailto:joe.a.hernandez@usace.army.mil)

EDUCATION

Syracuse University, Syracuse, NY

**Master of Science: Applied Data Science 2017 - Present**

Coursework: Data Analysis & Decision Making, Business Analytics, Financial Analytics,

Data Mining, Statistical Methods in Information Science and Technology.

**Bachelor of Science: Economics, Finance 2009**

Coursework: Economic Ideas and Issues, Corporation Finance, Financial Econometrics,

Economic Growth, Mathematical Economics, Investments.

Miami Dade College, Miami, FL

**Associate of Arts: Business Administration 2006**

Coursework: Business Calculus, Principles of Business, Business Law, Basic Business Statistics,

Business Writing, Financial Accounting, Managerial Accounting.

EXPERIENCE

U.S. Army Corps of Engineers, Sacramento, CA

**Economist – GS-11 CESPK June 2018 – Current**

* Lead regional economist for Army and Air Force MILCON projects and trained district economists to conduct MILCON Economic Analysis independently undertaking quality control role for their reports. Locations added: Camp Parks, Davis Monthan, Edwards AFB, Ft Hunter Liggett, Ft Huachuca, Luke AFB, and Presidio of Monterey.
* Economist for levee screening for nearly 100 sites mostly located in the California Delta region incorporating consequence analysis to demographics, evacuation effectiveness, and critical infrastructure impacts due to flooding.
* Economist for Semi-Quantitative Risk Assessment for California Delta region and mentoring junior economist to formulate a report based on hydrologic and levee screening data using ArcMAP GIS tabular and visual information.
* Performed regional on-site structure inventory for Sutter, Lower Cache Creek, and Pine Flat Dam Flood Risk Management Studies adding programmed randomized sampling of residential, commercial, and industrial buildings for economic updates.
* Completed the following PROSPECT courses:
  + Leadership Development Program (Level 2)

**Economist – GS-09 CESPK June 2017 – June 2018**

* Performed modeling for levee rehabilitation project for PL 84-99 sites supplementing Appendix with economic data, application to exceedance probability-damage curves, expected annual damage estimates, and methodology used in the report.
* Became lead regional economist for Air Force MILCON projects with formal training in Decision Support Analysis by the Air Force Financial Management Center of Expertise and adding an Army MILCON project using EconPACK software. Locations added: Edwards AFB and Yuma Proving Ground.
* Completed the following PROSPECT courses:
  + USACE Economic Analysis
  + Leadership Development Program (Level 1)
  + Risk Analysis in Water Resources Policy and Management

**Economist – GS-07 CESPK June 2016 – June 2017**

* Assisted in data preparation for the Lower Santa Cruz River flood risk management site damage economic analysis updating demographic information with census data and land usage and determining value through commodity prices for agricultural crops based on acreage for flood damage assessment using HEC-FDA to address regional economic development associated with the region and other social effects associated with flood risk.
* Supported Consequence Analysis Economic Update for New Hogan Dam focusing on incremental life loss based on breach and non-breach scenarios and developing breach parameters, modeling assumptions and discharges, and defining the area impacted by inundation with focus on population at risk, critical infrastructure, and economic losses.
* Pioneered Economic Analysis reports for United States Air Force MILCON projects determining justification through personal interviews, life cycle cost, benefit, uncertainty, and risk analysis using Air Force Financial Management Center of Expertise reporting products. Locations include: Nellis AFB and Creech AFB.
* Completed the following PROSPECT courses:
  + CW Project Development (PCC1)
  + Planning Essentials (PCC2)
  + Plan Formulation Capstone (PCC3)

Technical Skills

Disciplines: Economics, Statistics, Financial Management, & Computer Modeling

Programs: R & Python

Tools: ArcGIS, HEC-FDA, Levee Screening Tool, & EconPACK