# Learning Goal Deliverables

During my time within the Syracuse Applied Data Science program, I have been able to experience a wide spectrum of tools and information relating the data science. The below sections will detail an array of different projects completed with reference to how each has contributed to my data science education in the program.

## 2008 Arrival Delay Analysis: American Airlines vs. Everyone Else

One of the first projects completed in the Applied Data Science program was in IST 687 – Applied Data Science. This class introduced the fundamentals about data and methods for organizing, managing, curating, and using data. It was all conducted through R programming and provided one of my first insights into this powerful programming language.

The final project for this course gave an opportunity to use all these skills by analyzing 2008 airline delay data. This project involved extensive data cleaning and formatting to extract insights from the data and display the results in a visual format.

Numerous packages in R were used to assist in this project, such as ggplot2 and dplyr. Machine learning models such as linear regression and logistic regression were also used to predict length of delays and whether there would be a delay, respectively. Results were given using the ggplot2 package to visualize the results through numerous plots, including geographic plots.

These results were given in the perspective of American Airlines and included recommendations on how they could focus their attention to help best improve their customer experiences. By focusing the project from the perspective on American Airlines, our group was able to help focus on recommendations to give during our presentation and helped train us to understand that the goal of any analysis project is to give recommendations that can improve a company or product.

The skills learned in this class have helped to provide a strong foundation on how to handle a small dataset and extract information from it. Additionally, it helped build skills on how to collaborate on a project with a team of people. GitHub was used extensively to store updates to our script in a centralized manner.

## European Soccer Match Analysis

As the final project for my IST 565 – Data Mining class I did a predictive analysis of European soccer matches. This project utilized data collected from Kaggle with 7 different data sets (Country, Match, League, Player, Player\_Attributes, Team, and Team\_Attributes) that spanned from 2008-2015 for all European soccer leagues. However, this data was then subset to only include 5 different countries (England, Spain, France, Italy, Germany) to keep the data set to a more manageable size, while still keeping the data for the top 5 leagues in Europe.

The need to subset the data to help run the analyses in R, provided a valuable lesson to me on managing data. It became evident, that additional tools and resources are needed to perform some big data analyses, such as Hadoop and Spark. Yet, even despite this data limitation, numerous meaningful insights can be collected from a subset dataset when done in a reasonable manner.

R was used to perform the final predictive analyses, as well as the extensive data cleaning and feature engineering that was needed to combine the data from the different datasets into multiple new tables that were each able to answer a specific data question or used to create the predictive models. Prior to creating the final models, a thorough exploratory data analysis process was conducted to gain insights into the data that helped to influence how the final models were created. Some of the interesting findings included a histogram on the age where soccer players reach their peak (30), a bar plot showing the home and away goals by league (on average the home team scores ~35% more goals than away team), and how the average player rating by league has fluctuated from 2008-2015.

Numerous approaches were used to give analytical insights into the data such as association rule mining, random forest, and clustering analysis. Each different model was used to provide a different angle to understanding the dataset, but it also helped me learn how many different ways one single data set could be analyzed and gave me tools to implement them.

The final predictive model created using random forest, which beat out the support vector machine model, was able to create a prediction model that was able to predict the outcome of a European soccer match with 52.4% accuracy. Although unimpressive a first glance, Las Vegas bookies odds were at 53% accuracy, which is due to the large amount of uncertainty in soccer results.

This project did show some weaknesses in my R coding abilities though. During part of my script I was unable to complete the data transformations needed through R after an extensive period of trials. To get around this, I output two dataframes from R to Excel where I did the data transformations needed through VLOOKUP. This resulted in a new dataframe that I imported back into R to complete my analysis.

## Shelter Animal Outcome Analysis

My first experience using Python came during IST 652 – Scripting for Data Analysis class. During the final project for this class, I utilized the Python skills taught to predict whether shelter animals would be adopted or euthanized with 93% accuracy.

Adapting to using Python, as opposed to R, provided a challenge as this was a new syntax that had to be learned. This led to a steepened learning curve to understand how to clean the data, visualize it, and perform analysis in Python. But, because of this project, I was able to strengthen my data analysis skills by providing another tool with which to perform analysis. The ability to complete a data science analysis with Python has helped to give me a wider range of tools that I am able to leverage moving forward.

Additionally, this project tested me and taught me many lessons in the experience of having subject matter expertise on a subject. This dataset had a column of dog breeds, but it was set at a very high level that was too granular to make many meaningful insights. Without the knowledge of these dog breeds and the time to do extensive research to group them, there was a whole column of data that I was unable to get any value from. Having this subject matter expertise could have helped me improve the feature engineering of the dataset and helped make clear the importance of subject matter expertise in any project.

During this project I also realized my weakness in plotting through Python. Creating visualizations through matplotlib proved much more time-consuming than making similar plots in R through ggplot2. Having this knowledge may impact the decisions I make when I need to make visualizations in the future, but I still feel comfortable using either, if needed.

This project also differed from many other projects I have completed as I was working by myself on this one. Working by myself on this project helped strengthen my overall data science learning though, as I was unable to rely on group members to fill in missing pieces of my knowledge.

## Spam Email Classification

The explosion of data in the world has largely come from unstructured data sources that come from text sources. Thus, the ability to extract information from structured data sources is not enough. IST 664 – Natural Language Processing has helped to provide a strong groundwork in analyzing text data. The final project for this class used these skills to make a model that can classify spam vs. ham (non-spam) emails. These models were created using Python and leaning heavily on the NLTK package for formatting the data and creating the features. This project was a first for me in working with unstructured data and helped teach me how to work with it and opened my eyes to the wide number of additional projects that could be started or improved upon by using this unstructured data.

To perform these analyses, the data had to first be converted from text to a structured format that could be used to create a model. This class taught how to convert the text to a structured format and how to create additional features from the data to improve model performance. The data pre-processing steps used to perform this analysis is very different than those performed with structured data. First, a corpus of words was created from the emails with a frequency distribution to grab the 2,000 most frequently sed words. Numerous filters were created to clean this corpus including a stop word filter to remove filler words and a non-alphabetic character filter to remove punctuations and symbols. Additional features were then created such as unigram features, bigrams, part-of-speech (POS) tags, and sentiment analysis. By combining the different features and filters with different modeling packages (sci-kit learn and NLTK).

This project was invaluable in learning how to work with unstructured data sources and taught me a very valuable package, NLTK, that proved to be exceptional in transforming and formatting the data and applying machine learning model to analyze it.

# Learning Goal Outcomes

The MS in Applied Data Science is meant to provide students with an opportunity to learn a broad range of topics related to data science. Below are some of the learning outcome goals with comments on which projects demonstrated my competencies in the area:

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

My work on the “European Soccer Match Analysis” demonstrated the broad overview of the major practice areas of data science. It touches on all the stages of the data science pipeline. Numerous data tables were pulled from a SQL database into R, cleaned and organized for analysis, and multiple analytic approaches were applied to acquire multiple new insights. After the data was obtained and cleaned, extensive exploratory data analysis (EDA) was performed to help visualize trends in the data that informed future decisions on the final analysis. After this EDA, numerous machine learning algorithms (association rule mining, clustering, random forest, support vector machine) were applied to the data to extract insights. These insights were then presented to my class through a presentation and a written paper. The whole process worked through the entire data science pipeline and demonstrated a broad overview of the major practice areas of data science.

1. *Collect and organize data.*

The “European Soccer Match Analysis” proves an example of the competency developed and collecting and organizing data. Through this project, several different data tables had to be combined, cleaned, and modified to create new tables that were able to answer specific data questions. Using both R and Excel, numerous table joins were performed, new columns were created based on existing data, and data was cleaned and organized into structured data frames. This involved extensive data cleaning and demonstrated an ability to collect and organize data in a way to provide analytics.

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

My competency in data visualization, statistical analysis, and data mining can be exemplified from my project on “Shelter Animal Outcome Analysis”. As part of the EDA process, I created histograms, bar charts, box plots, and line graphs that showed trends in the data. This involved statistical analysis as means of the data were compared to each other and regressions were run to display trends in the data. Finally, random forest models were applied to predict the outcomes and discover the most important features in the data. Overall, this project demonstrates my capabilities in data visualization, statistical analysis, and data mining.

1. *Develop alternative strategies based on the data.*

The “Spam Email Classification” project helped developed alternative strategies based on the data as this project dealt with unstructured text data. To adjust to this, the data had to be formatted out of the unstructured format into a more structured one so analysis could be conducted. After the data was in a data frame format, numerous approaches were taken to find which performed best. To find the best model, multiple different machine learning algorithms were performed on slightly modified tables that included different features. By cycling through these different machine learning algorithms and data tables, the best model was found to help answer the question. This process of transforming the data and iterating through different data and algorithms demonstrated my ability to take alternative strategies based on the data.

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

My “2008 Arrival Delay Analysis: American Airlines vs. Everyone Else” demonstrates my competency to implement the business decisions derived from the analysis. This entire project was taken from the perspective as employees of American Airlines. The analysis performed used the entire dataset to compare performance from American Airlines to their competitors to see where they had weaknesses that should be improved on, and where they had strengths to maintain performance on. From the analysis, we were able to discover what leads to the majority of arrival delays and recommend some steps to improve upon their arrival stats.

1. *Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professional in their organization.*

All the projects completed in this program have helped demonstrate my competency in communications skills regarding data and its analysis. For each project I completed, the results were presented to my class through a formal presentation. The data and analysis shown during this step was kept at a high level to deliver this information to those who did not have any subject matter expertise in the matter. Additionally, the analysis shown did not give overwhelming details on the data cleaning or analyses performed, but instead kept this information at a level appropriate for those in the audience. The written papers, on the other hand, went into more extensive details on the data cleaning and analyses performed. They gave rationale on the algorithms used, including reasonings for the algorithms that were used for each analysis. These papers included details that were appropriate for more data-literate members of the organization. By understanding what information to convey to different audiences, these projects have helped demonstrate my communication skills regarding data and its analysis.

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

In my most recent project on “Recordable Incidents and Near-Misses in the Workplace”, my team applied the ethical dimensions of a data science practice. Since this project involved company data for multiple companies, our team took steps prior to any analysis to scrub the company info from the analysis. In our presentation to our class, all companies discussed were listed as aliases (e.g. Company D) so no company data was leaked.