# 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.

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. Our group used R to perform the analysis with the assistance of numerous packages 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.

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 from 7 different data sets (Country, Match, League, Player, Player\_Attributes, Team, and Team\_Attributes).

R was used to perform the final predictive analysis that involved extensive data cleaning and feature engineering was used to combine the data from the different datasets into one final table that was used to create the 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.

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. The predictive model created using random forest were able to create a prediction model that was able to predict the outcome with 52.4%. 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.

This project helped 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.

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.

## 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.

To perform these analyses, the data had to 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)