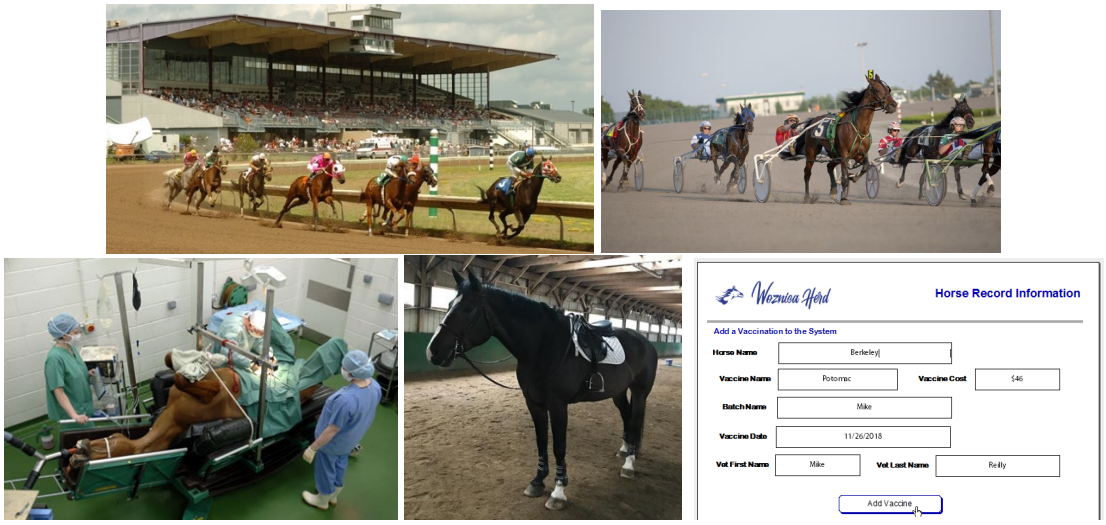
# Graduation Portfolio




## My Journey through Applied Data Analytics at Syracuse University

Joyce Woznica
Syracuse ID: 374633973

Date: September 7, 2021

# Table of Contents

# List of Figures

# Introduction

There are six short weeks remaining in the horse show season in 2006 — my daughters and their horses are trying to hang on to their ranking so that they will qualify for state finals and year-end awards for their efforts throughout the season. I have gathered data over the season for each show that they have attended which includes:

- the judge
- the classes entered by each daughter
- which horse was ridden in each class
- their placing in each class
- personal comments about the judge and what he/she was looking for in each class

Using this information, my daughters and I plan the remaining six weeks to maximize their potential to score additional points (each placing has an associated point value). The outcome: they made it to state finals and placed in the top two for their year-end awards in their respective divisions. Little did I know, I was doing rudimentary data analytics (without the assistance of programming and machine learning) to predict the outcome of the future shows. This "manual" method remained a mainstay throughout my daughters' amateur junior career in the saddle.

Fast forward a decade and both girls (and their horses) are in college, and I am looking at an empty nest as a single parent. I know that work and riding my own horse will not fill the gap of dealing with teenage girls over the past several years, so I had to come up with a plan for filling my days while they were gone — enter graduate school at the age fifty-seven. I selected the Syracuse Applied Data Analytics program because it addressed applying data analytics to specific business problems to achieve an outcome. This portfolio combines my obsession with numbers and data with my passion for horses and clearly represents my progression and acquired skills throughout the program.

## *Syracuse Applied Data Science Program*

The Syracuse Applied Data Science program is interdisciplinary providing students with a diverse program in a board range of areas.[1] This program focuses on seven main areas:

1. Describing a broad overview of the major practice areas of data science.
2. Collecting and organizing data.
3. Identifying patterns in data via visualization, statistical analysis, and data mining.

[1] *https://ischool.syr.edu/academics/applied-data-science-masters-degree/*

4. Developing alternative strategies based on the data.
5. Developing a plan of action to implement the business decisions derived from the analyses.
6. Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesizing the ethical dimensions of data science practice (e.g., privacy).

This paper provides details into my own personal journey in the Applied Data Science master's program at Syracuse University with concrete examples of the knowledge gained through various projects completed while I was a student at Syracuse University. All of this was completed while maintaining a focus around my passion – horses.

# Key Course Projects

## *Data Administration Concepts and Database Management*

*Project: Horse Records*
*Course Number: IST659*

One of the first topics addressed in the program is the importance of collecting data and the organization of that data in ways that will present that data properly for analysis with data analytics. Although I was familiar with some Standard Query Language (SQL) and databases from time in the workforce, I still entered IST659 with limited experience and exposure. Under the guidance of Chad Harper, I was able to construct a normalized database that linked information related to horse vaccination schedules and veterinary records. A key goal of this course was building a normalized database, understanding the relationship between database tables and then adding and extracting information from the database to answer questions.

As an owner of multiple horses that have been in different states with different veterinary options for vaccines and other routine medical work, it is very difficult to keep track of each horse's vaccine schedule and other pertinent information. By providing a single database of all my horses and their vaccine history, a solution was developed to maintain this information and provide reminders as well as recommendations for future routine work. For this project, I created a database that tracked of my horses' vaccines as well as all their dentist, chiropractor and other routine medical visits. The information in this database will allow me to answer simple inquiries:

- When is <HORSE> due for <VACCINE>?
- What vet did <HORSE>'s last <VACCINE>?
- When did <HORSE> get the last dose of <VACCINE>?

- How much money did I spent on vaccine(s) in <YEAR>?
- What is the average price for <VACCINE>?
- How much was spent on <HORSE> for vaccines in <YEAR>?

To create this database, data had to be collected and reviewed to determine the best way to represent the data in individual database tables and then note the relationships between the data in these tables. Once this was accomplished, data was gathered and imported into the database and various scripts and utilities for adding, modifying, and extracting data were created. Finally, visualizations of the data and possible user interfaces for querying the data and making inferences were developed.

The following figure shows the final normalized database showing the relationships between the tables as well as the keys within the tables. The planning and design of this layout was key to the overall project success.

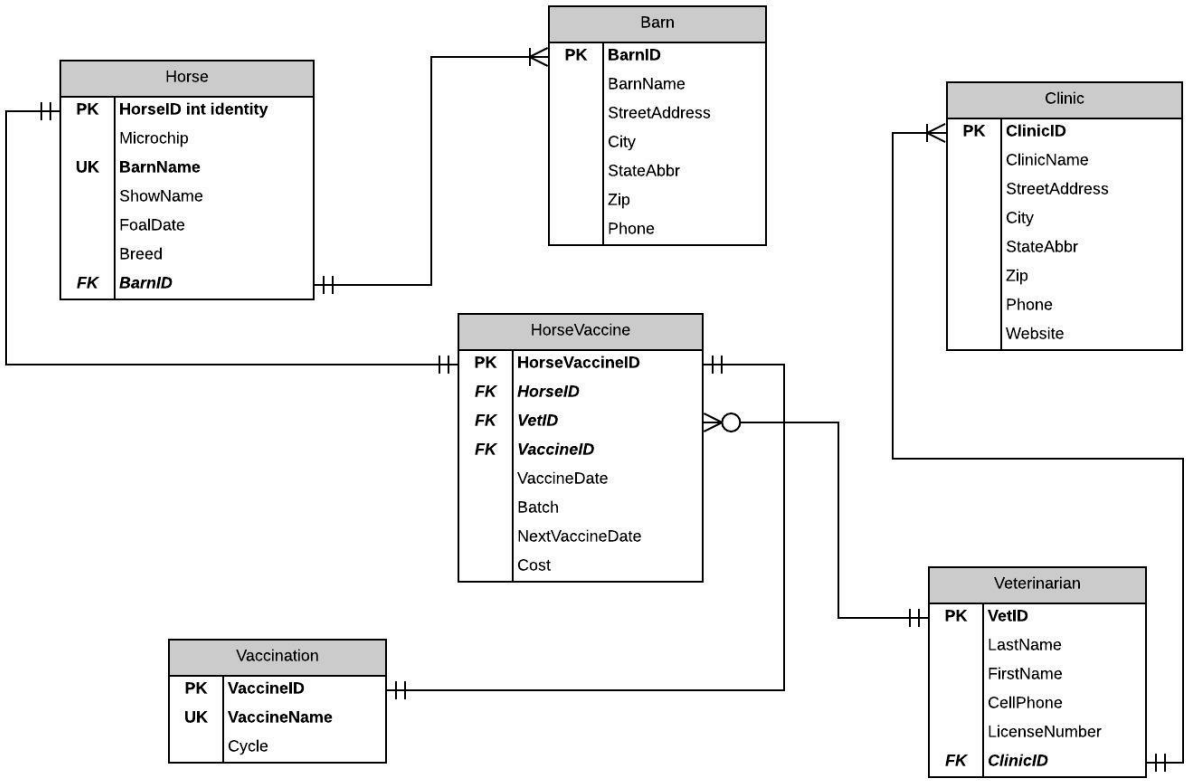



*Figure 1. Normalized Horse Records Database*

Private, universal, and foreign keys are displayed in each table along with the relationship (one to one, one to many, etc.) showing how the data is related in each table.

Once the database tables were properly loaded, gaining insight into the data was then available. This was accomplished by creating a user interface to query the database and dashboards to visualize the information. Examples of two query screens are provided in the following figure.

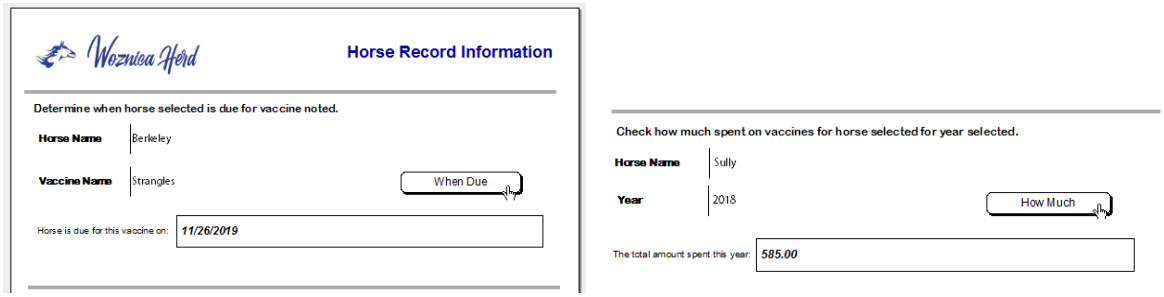



*Figure 2. Example Query User Interface Screens*

In addition to querying the database for answers to specific information, different views and visualizations of the data provided a more wholistic view of the information. Examples of these dashboards can be found in the next figure.

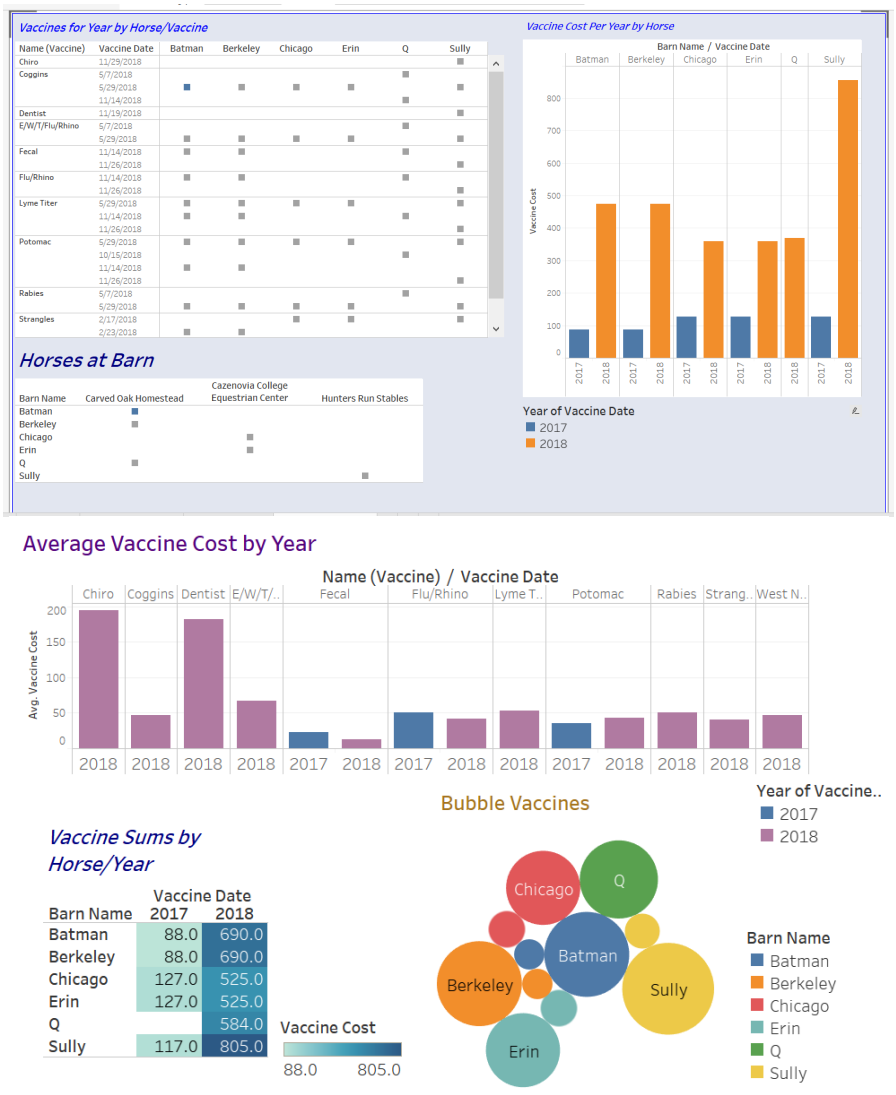



*Figure 3. Example Horse Vet Record Dashboards*

Using these dashboards, the business questions initially posed can be answered including responses to how much was spent on vaccines each year – by vaccine and by horse, the barn and location for each horse and when vaccines took place and are needed again.

*Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Collecting and organizing data.

  Data was in three different locations: Microsoft Word files, spreadsheets, invoices and my checkbook. The spreadsheets were combined and then expanded to include the information from all of these sources. Database tables were created using SQL scripts as per Figure 1.. This data was then loaded into the database tables by generating SQL "insert" scripts to populate the data in the correct table with the proper fields to achieve the relationships represented.

- Identifying patterns in data via visualization, statistical analysis, and data mining.

  After the database tables were created and populated, the data could then be visualized to gain insight into the information. This was accomplished, as mentioned, through various dashboards and querying techniques. Using these tools, the data could be viewed in a concise manner while providing details and insight into the horse record data.

- Developing a plan of action to implement the business decisions derived from the analyses.

  One of the key things learned and achieved in this course, was developing a plan from data collection to data visualization and insight. By carefully segmenting the information into individual database tables and providing the proper relationships between the tables, simple queries could be created to gather answers to questions posed to the database as well as visualized in the dashboards. Without a clear, thought-out plan, this would not have been possible.

- Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

  As with the majority of courses completed at Syracuse University, the skill of communicating findings to the appropriate audience was a key learning objective. The report requirement to document the work including an introduction, details and a conclusion increases the knowledge component of the course. In addition, the project for this course enabled the

creation of a marketable application for fellow equestrians and horse owners – something I plan to pursue in the future.

## *Data Analysis and Decision Making*

*Project: Horse Ride Preparation Process Improvements*
*Course Number: MBC638*

In Data Analysis and Decision Making, the class was asked to select a topic of interest for process improvement and then use statistical methods to determine if changes made in the process actually achieved improvement. This project very directly addressed how to use analyses to develop a plan of action to implement business decisions. A very interesting concept in this class was the idea of *Define – Measure – Analyze – Improve – Control* which was the approach taken for the final project. Staying with my original theme of equestrian/equine related topics, it was determined that this course provided the opportunity to solve a problem that I experienced with my own riding—streamlining the process of "preparing" to ride (or the grooming and tacking process). For the project, there needed to be a compelling reason with related impacts for lack of improvements; such as, costs, negative impacts (deterioration of the horse) among other things. In this example, a nonexercised horse suffers ailments similar to humans that are not physically fit – bad joints, swelling, arthritis and more. In addition, this results in pain for the animal as well as substantial veterinary bills.

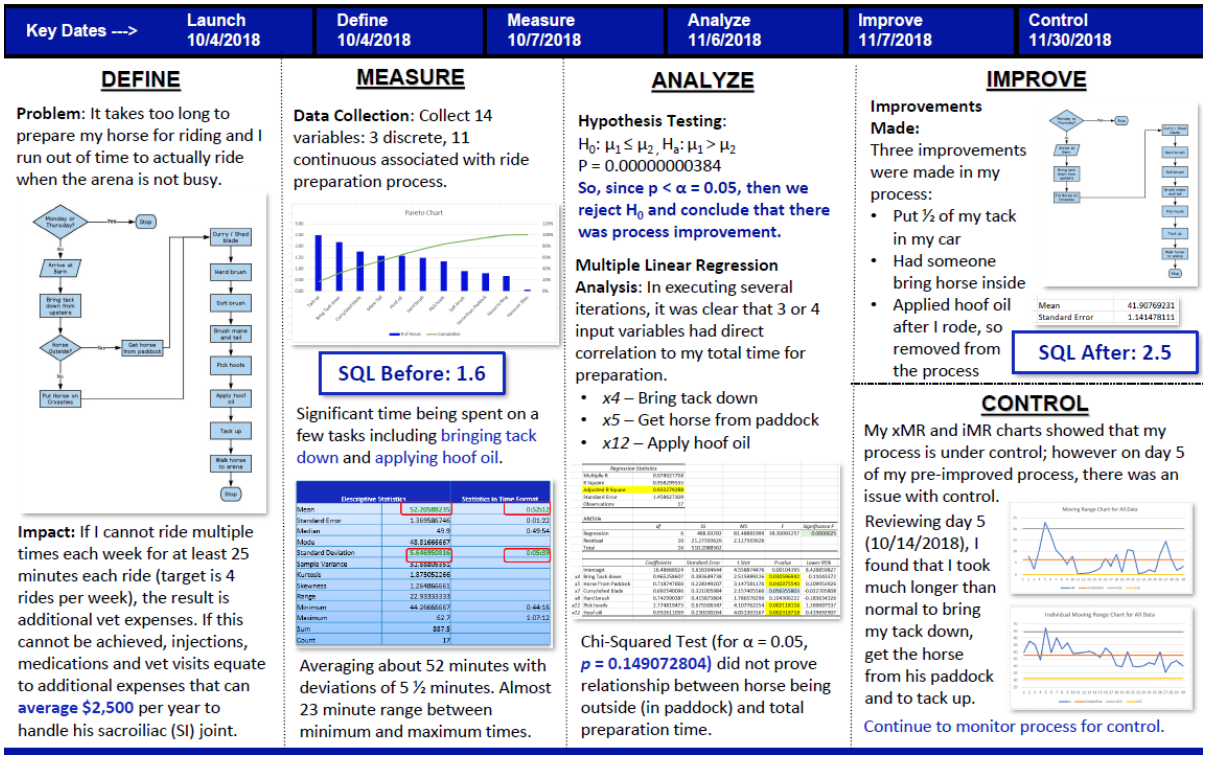



*Figure 4. Process Improvement Summary Slide*

The initial slide for the final presentation needed to portray all key data in a single view as shown in the previous figure. Then the subsequent presentation slides took a deeper look at individual components of the analysis.

By mapping the existing process using flowcharting tools, areas of improvement were immediately apparent. Using this process map as a guide, a data measurement plan was developed which included the specific data points to gather at each step in the process along with a time period for pre-improvement data collection. This was accomplished by researching tools to assist in setting tasks and then timing each task. The results were downloaded in spreadsheet form to be used in statistical analysis to attempt to improve the process. The data combined learning components of the course by utilizing different variable types: both discrete and continuous data were collected for the independent variables with a dependent variable as the overall preparation time for riding. Improvement goals were developed to measure success.

Then specific statistical methods were used to understand the data and to determine where improvements could be gained in the process. Summary statistics were reported to review the baseline pre-improvement time and how many standard deviations of improvement could be shaved off the initial process to minimize preparation time in order to achieve more time in the saddle.

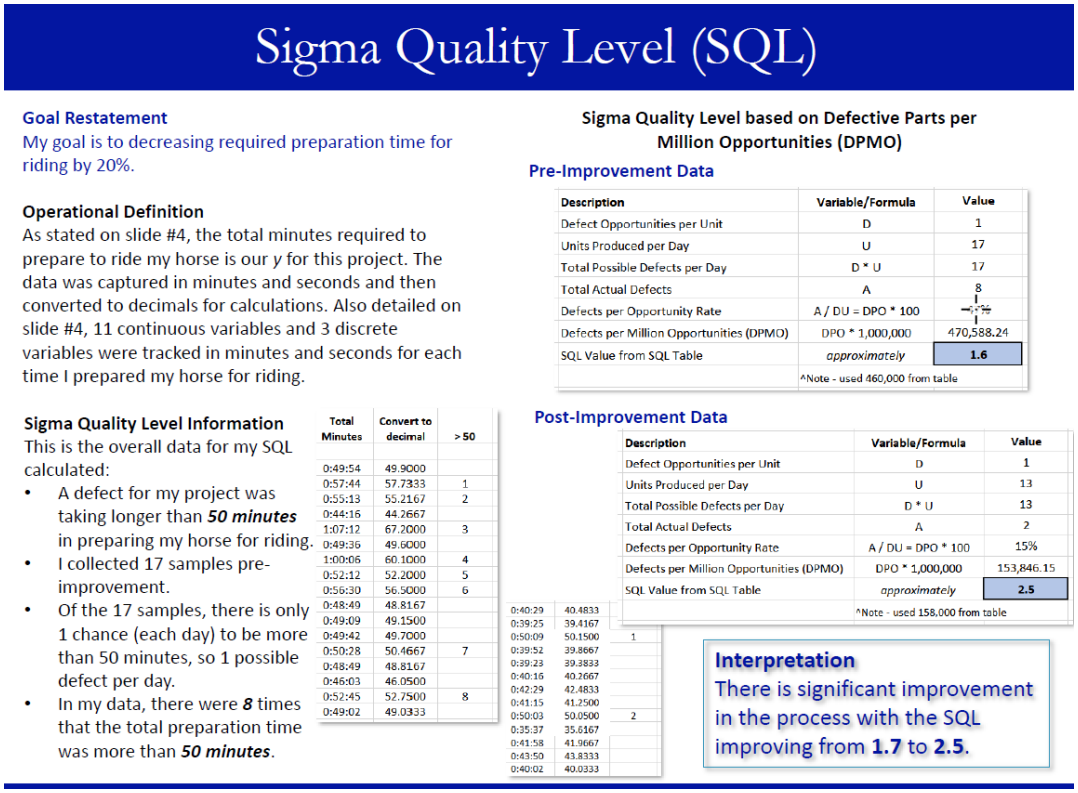



*Figure 5. Sigma Quality Level (SQL) Summary*

The previous figure shows the Sigma Quality Level which was used to compare pre and post improvement data and measure the success of the process improvement. As shown in the previous figure, with process improvements implemented, the Sigma Quality Level improved from 1.7 to 2.5.

Using Multiple Linear Regression analysis helped to narrow down specific areas that were directly affecting the time required to prepare to ride by reviewing the significance of each variable. Using this method and multiple iterations, it became clear what tasks should be modified to achieve improvements. The results of this analysis can be found in the following figure.

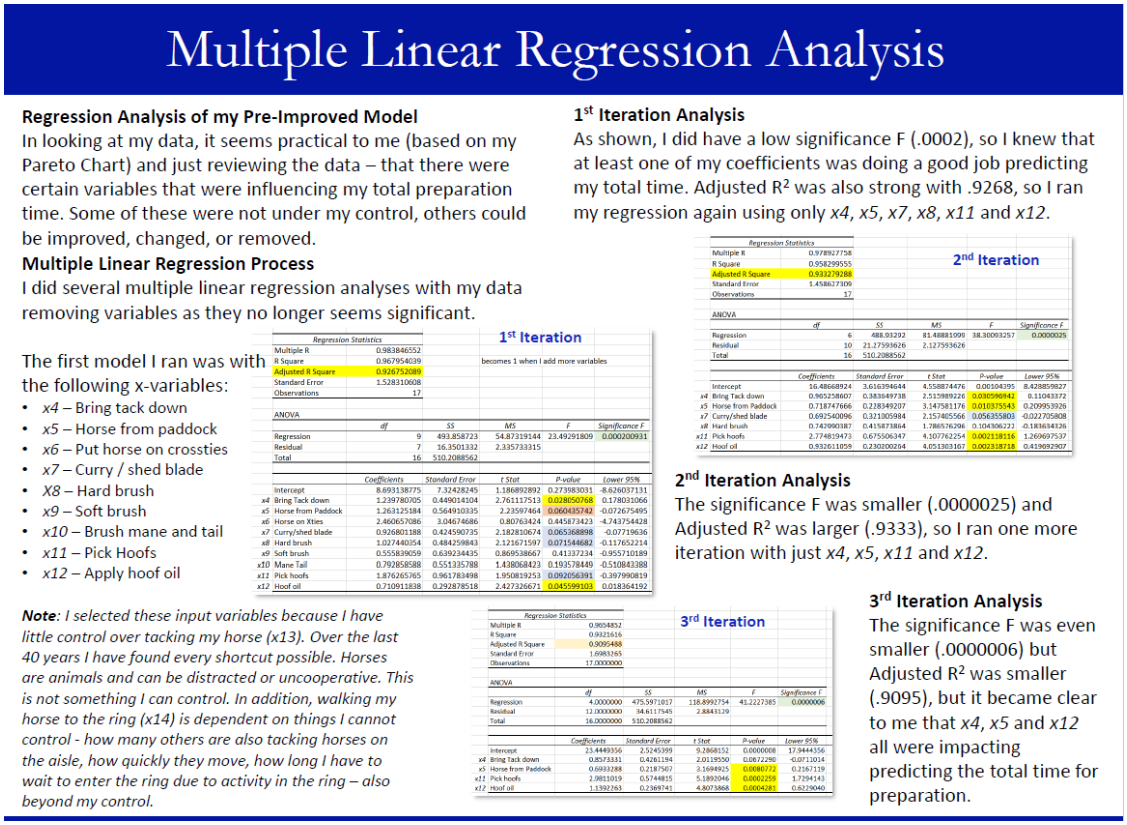



*Figure 6. Multiple Linear Regression Summary*

## *Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Describing a broad overview of the major practice areas of data science.

  Process improvement is a very important concept that businesses investigate in multiple areas of the operations. By articulating the methods used to identify and achieve process improvements, the concept of describing a broad overview of the major practice areas of data science was achieved. The final presentation for this course included the discussion of statistical methods and areas of data science that can be used to create, verify and prove

assumptions. For example, using the Sigma Quality Level and performing various statistical tests – hypothesis testing, multiple linear regression analysis, and then generating control charts are examples of utilizing practice areas of data science to achieve results.

- Collecting and organizing data.

  As with the previous course, data collection and organization were key to achieving the goals laid out for the final project. Learning new tools for data collection and organizing this data laid the foundation for this project. Using the "*ATracker*" program to collect the data was an excellent way to capture information for each possible task or component for improvement. This data was then easy to use for the process improvement project and it was a great way to keep the data organized.

- Developing a plan of action to implement the business decisions derived from the analyses.

  One component of the project was to implement different approaches and tests to determine as well as prove that improvement was achieved. This often required reviewing the data from a different perspective or taking a different direction. For example, using the *Define – Measure – Analyze – Improve – Control* approach showed a clear plan for determining if process improvement could be achieved in this scenario. Each step had an associated time period for capturing data related to this stage in the process. By clearly capturing each component, the improvement areas were easily distinguishable and business decisions could be derived from data analysis.

- Developing alternative strategies based on the data.

  A plan of action was developed to implement and test the business solutions made to improve the process as derived from the analysis completed. By documenting the individual tasks and how each could be modified, if applicable, the data presenting alternative strategies to narrow down improvement areas and to show that success was achieved. Determining defects and then testing the hypothesis that improvements could be made led to multiple linear regression and additional testing. Another example was the use of Pareto Charts and additional testing with the Chi Squared test for independence around horse location – all which were dictated by the data and looking at alternative strategies to determine how to improve the process.

- Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

  Finally, as with all courses I took in this program, documenting the approach and findings help to improve and fine tune communication skills. This project, using a PowerPoint presentation format that needed to play the role of a document summarizing the project, was new to me and allowed me to use a more creative side to lay out the slides in a readable manner as they had far too much text to actually be used in presentation form.

## *Data Analytics*

*Project: Predicting Horse Colic*
*Course Number: IST707*

Data Analytics provided a platform for understanding how to predict or determine if an outcome from past data outcomes. This became an excellent platform for working with equines again with the effort focused on determining if the outcome of a horse colic could be predicted. *Among domesticated horses, colic is the leading cause of premature death*. Equine colic is a relatively common disorder of the digestive system. Although the term colic, in the true definition of the word, simply means "abdominal pain," the term in horses refers to a condition of severe abdominal discomfort characterized by pawing, rolling, and sometimes the inability to defecate. This rolling can lead to the twisting of the over seventy feet of intestine which can lead to death of intestine sections or the entire intestine. There are a variety of different causes of colic, some of which can prove fatal without surgical intervention. Colic surgery is usually an expensive procedure as it is major abdominal surgery, often with intensive aftercare. A horse has about a 10% chance to colic over his/her lifetime. Between 25% and 30% of horses presented with colic symptoms are recommended to go to surgery. In those cases, between five and 10 percent will be humanely euthanized because of a poor prognosis or economic considerations (Woznica, *Classification of Racetrack Incidents*, 4).

This project was one of the most interesting and probably the biggest learning opportunity for me in the program and close to my heart, because I lost a horse to colic – she was euthanized on the operating table. I had a bit of an advantage with this dataset because I had some general knowledge in what area might be of interest and might be a contributing component to colic prognosis. The figure shown below looks specifically at the fecal output per outcome for the horses in the dataset. This is important because when a horse is suffering from colic, the lack of fecal output can mean that an intestine is twisted causing the pain. As can be seen in this figure, there are certain fecal status that seem to be related to the colic outcome (specifically, absent or decreased).

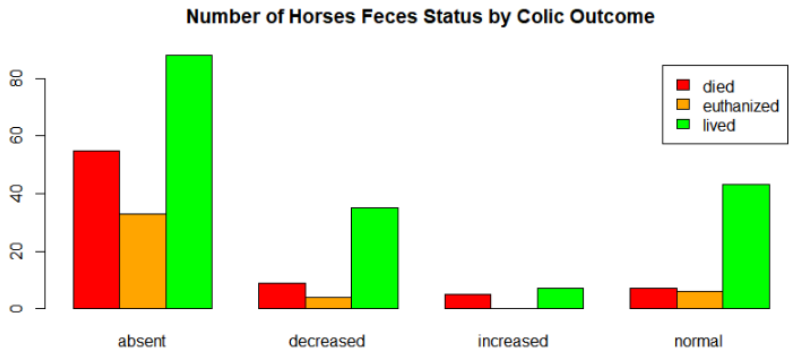



*Figure 7. Feces Status related to Colic Outcome*

With that information, it was also important to visualize if there were any specific lesions and where they were and how that might play a role in colic outcome as show in the following figure. This visualization, although similar, provides more insight into the data to prepare for modeling and prediction.


*Figure 8. Lesion Site related to Colic Outcome*

With this established information, it was time to begin looking into prediction and which models provided the best accuracy for predicting the colic outcome. This was a wonderful learning experience around selecting the appropriate statistical model, fine tuning the model appropriately, visualizing that model and the outcomes and then interpreting the results. Decision trees and random forests were discussed in class. The best decision tree can be found, but taking the time to try different models and components help to improve the accuracy. This was very important as accuracy can be improved, but you can use too much information making the model overly complex. There is a fine line between the simple or pruned model and the overly complex model. Keeping this in mind, a pruned model was developed as shown in the following figure.

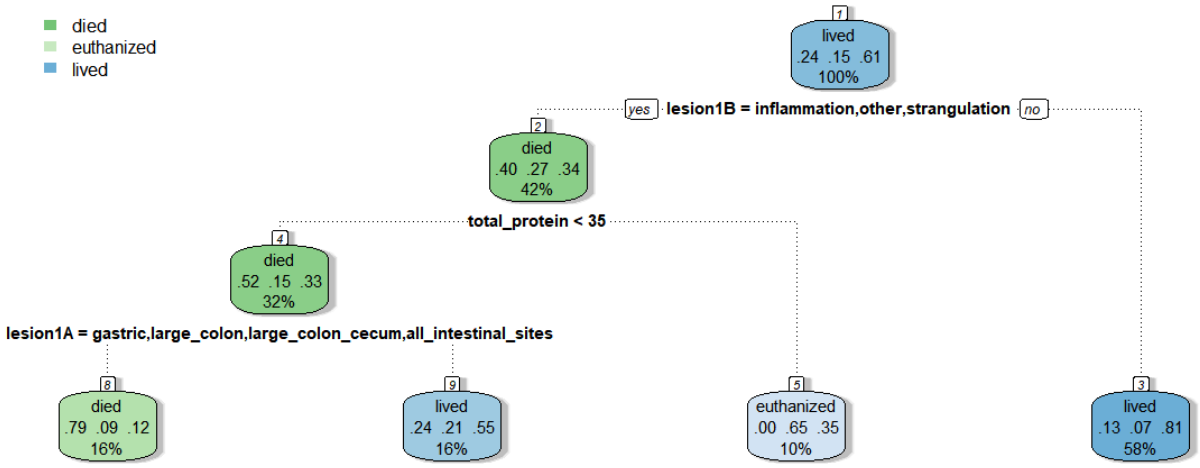



*Figure 9. Decision Tree for Predicting Horse Colic Outcome*

As per our initial visualization, the location of the lesion and type played an initial role in colic outcome prediction followed by blood protein and further lesion information. Using this simple model, the outcome of died, lived or euthanized could be determined with some level of accuracy. Neural networks were then reviewed to predict outcome. Two hidden nodes were introduced to obtain the following neural network with 75 steps and an error of 26.267969 as show in the figure below.

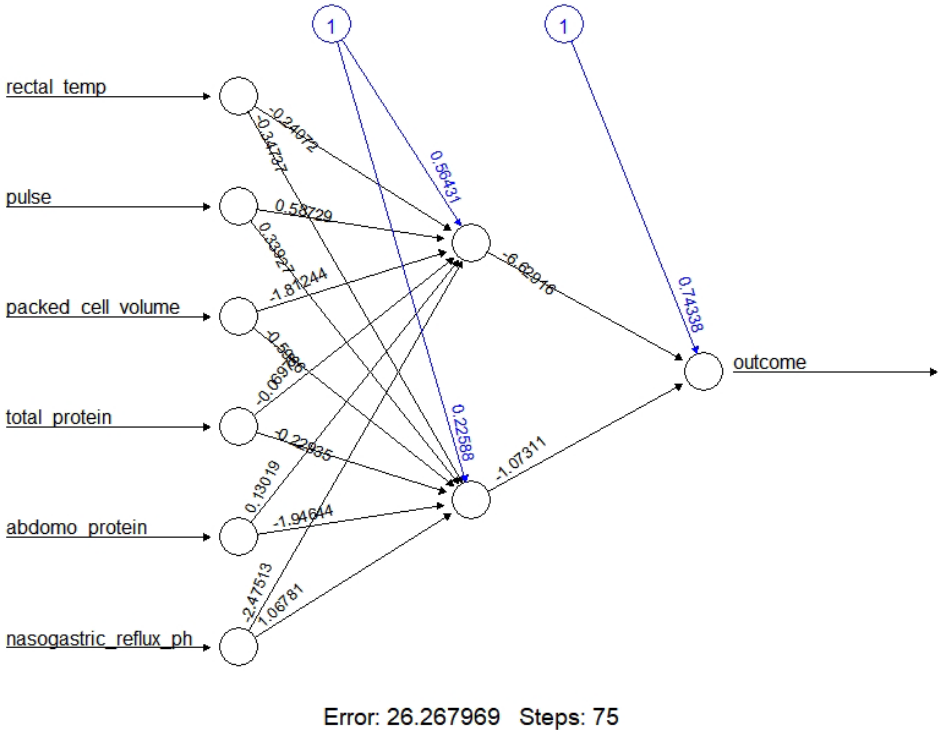



*Figure 10. Plotted Neural Network for Predicting Horse Colic Outcome*

Additional models were reviewed including Random Forests, Association Rule Mining and my introduction to Naïve Bayes and the importance cross-validation and the confusion matrix. The best results were found using Support Vector Machines which I was able to use in other classes with the knowledge gained in this course.

*Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Identifying patterns in data via visualization, statistical analysis, and data mining.

  By creating initial visualizations, patterns and other insight became apparent in the data set. Although I had some initial knowledge of the symptoms that are often seen in a horse that is experiencing a colic episode, the visualization of related variables was very helpful and uncovered additional patterns in the data by reviewing certain variable combinations with others. This information helped to determine which variables might be relevant when addressing the prediction with statistical modeling.

- Developing alternative strategies based on the data.

  As mentioned previously, many different models were executed to find the best model for predicting colic outcomes for the data set selected. For example, as shown with decision trees, trees were created and pruned to arrive at the optimal model. Working with neural networks provided a alternative look at the data from a new perspective. In addition, Support Vector Machines with different kernels were used to arrive on the best model and kernel for predicting the outcome. By reviewing the data and interpreting results of models, alternative strategies were used to arrive at a more optimal model.

- Developing a plan of action to implement the business decisions derived from the analyses.

  As part of this project, a plan was created, with Professor Bolton's guidance, to begin to understand the data and then attempt different models for colic prediction based on what was seen in initial observations. As shown in the conclusion for this project, understanding the key components or symptoms, veterinarians can approach colic incidents with a new insight. They can check these symptoms first and possibly make a more informed and faster decision that could possibly save horses lives. By understanding the results from the analyses, plans like this can be put into place for the future.

## *Information Visualization*

*Project: Tracking the Danger Zone – Horse Track Incidents*
*Course Number: IST719*

The course on Information Visualization might have been my favorite in the program. Learning how to properly present information to different audiences to assure that the findings are clear and obvious was of great value. There were very important topics covered in this course around drawing the eye of the consumer, how to present a consistent theme, simplifying graphics and focusing the reader on the areas of interest and more. The final project in this course was to prepare a full-sized poster for the project which was then presented to the class. A scaled version of this final poster can be found in the following figure.

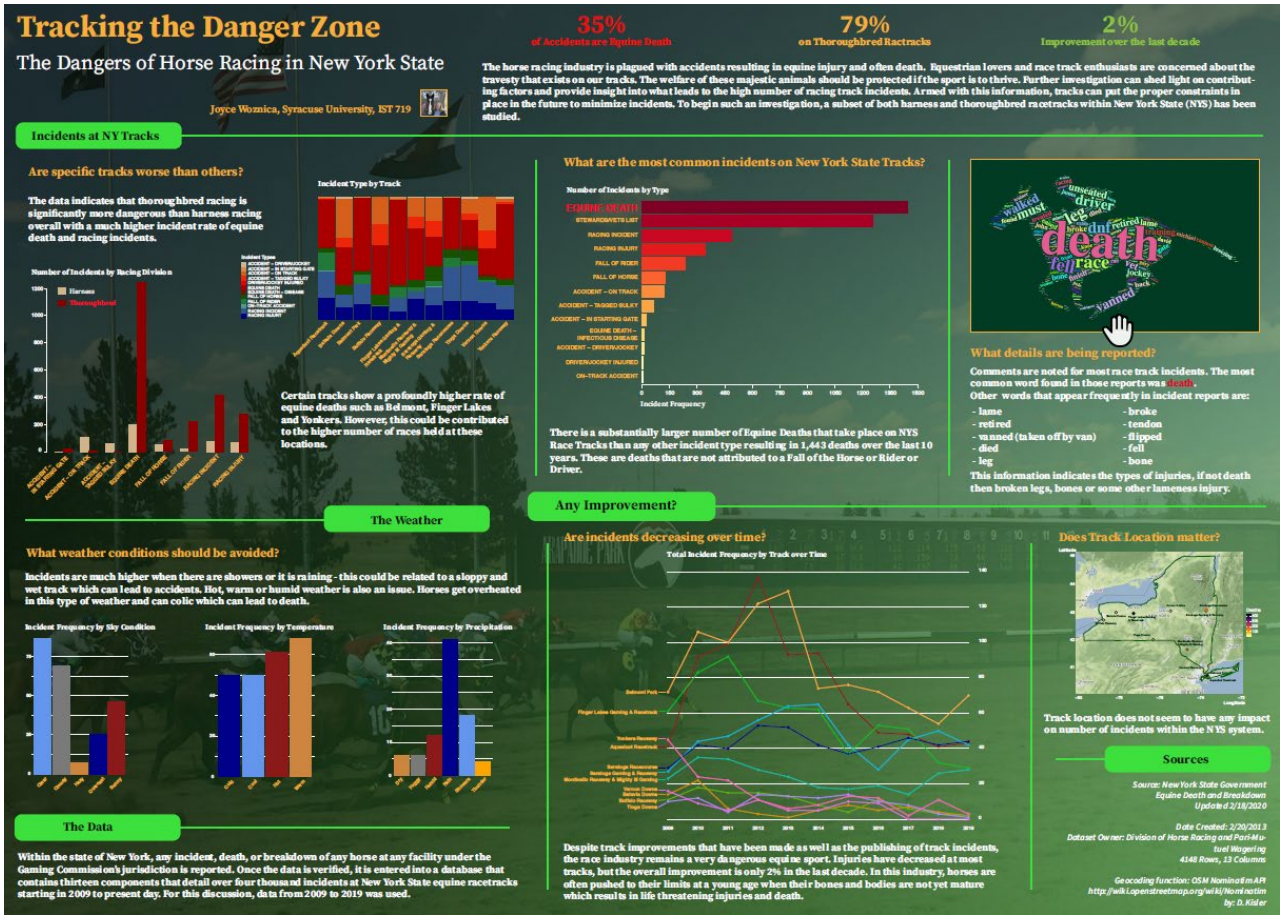



*Figure 11. Final Poster Project for Race Track Incidents*

As with other projects, I continued with a theme of horses, but focused on how to improve safety on the racetrack using a public set of data provided by the New York State Gaming Commission. Within the state of New York, any incident, death, or breakdown of any horse, as well as jockey and driver accidents, at any facility under the Gaming Commission's jurisdiction is reported to the Presiding Judge or Stewart at that facility. This individual then provides the details of the incident to the

Commission's main office in Schenectady, New York. The verified data is entered into a database that contains 13 components that detail over 4,000 incidents at New York State equine racetracks between March 4, 2009 to present day. This data is updated daily (Woznica, *Tracking the Danger Zone*, 4). Ironically, my eldest daughter now works for the Tioga Downs racetrack taking blood from the winners to test for illegal substances.

This course provided great insight into how people consume information and how to use that to draw attention to results in data analysis. For example, providing some key statistics in an easy-to-read manner and large font draws the consumers eye and provides immediate understanding into the data being analyzed. As shown below, some percentages of information about the data to bring the reader into the story being told and heighten their interest.

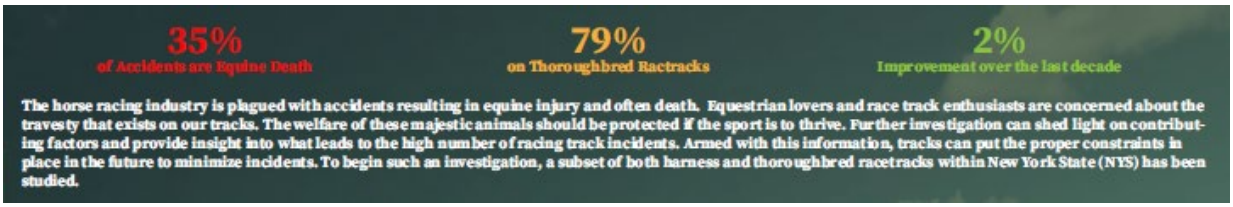



*Figure 12. Key Statistics concerning Race Track Incidents*

Another thing that was reviewed as part of this exercise was looking at words that could be found in the incident description and what could be determined by looking at this information. By presenting top words from the incident descriptions in the shape of a racing horse provided a way to draw the consumer's focus as well as a very apropos way to present the information. This exercise did show some insight into the most common ailments in the incident's reports.

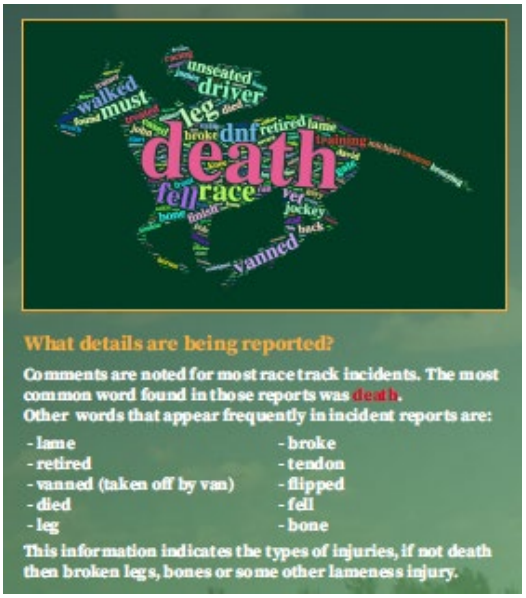



*Figure 13. Final Poster Project for Race Track Incidents*

In reviewing the data and applying some of my knowledge of horse injury and the strains on joints when galloping around a track at a young age, I started to look at information that I felt might have some influence on the racing incident outcome. The first of these was related to the weather. Running on a wet track can make it very slippery and can also affect the ability to see what is in front of you and running on a very cold or hard track due to cold temperatures are harder on the joints and can cause injury to the horse. This was detailed in the segment of the poster displayed in the following figure where weather conditions and racing outcomes were reviewed in visualization.


*Figure 14. Weather Conditions and Race Track Incidents*

By taking this course, it became very apparent how the important the display and dissemination of analysis results can be as part of the overall process of data analysis.

### *Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Identifying patterns in data via visualization, statistical analysis, and data mining.

  As this was a visualization course, the process of imagining the overall layout and how to present graphics and plots in the best way to showcase the results was stressed each week. We learned how to manipulate plots using Adobe Illustrator to make them more aesthetically pleasing and to clearly display the desired information. I feel this skill will be very useful in the future as it is a key component to communicating results to key stakeholders and executives.

- Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

  One of the most difficult components of data analytics, in my opinion, is to communicate the results of any analysis to managers, executives and other relevant professionals. By finding new ways to present graphics and plots and portray the information in appropriate visualizations, I was able to demonstrate how to communicate the results of data analysis to various individuals within an organization. At the end of this course, our final poster was presented to the class zooming in on certain areas to discuss specifics of the data visualizations. This is a skill that will be used throughout my career.

- Synthesizing the ethical dimensions of data science practice (e.g., privacy).

  One of the most difficult things in this course was finding a data set of interest that was public information that was available for analysis. So much information today is protected via privacy laws and more. It is key to know what protects personal information (PII) and laws around using this information is extremely important. For the past 20+ years, I have worked in electronic content management and governance, so this is a topic that is well understood by myself and customers. Understanding what can be used and not used is a key to data science. In this situation, I was lucky that New York State publishes this information readily and updates it daily. This is very rare as I could find no other information similar from any other tracks as they do not make this information public.

## *Scripting for Data Analytics*

*Project: Racetrack Infractions and Incidents*
*Course Number: IST652*

Scripting for Data Analytics introduced me to new concepts like accessing different data streams and capturing them using MongoDB, NoSQL and JSON files. Working with both structured and unstructured data in a more robust manner was an excellent way to deepen my knowledge in applied data science. Along with learning these new techniques and technologies, it provided new ways to visualize, interrogate and interpret data and results. Again, I was able to take advantage of additional information from the New York State Gaming Commission. Each race conducted at a New York Thoroughbred racetrack is observed by three stewards: one employed by the Racing and Wagering Board, one employed by the racing association and one employed by the Jockey Club. At the harness tracks, each race is observed by three judges who all serve as employees of the Racing and Wagering Board. The stewards' and judges' viewing stand is located near the finish line of each racetrack and is

equipped with several television monitors to permit the viewing of multiple angles of each race. The stewards and judges observe the races and the race grounds to ensure that all conduct is in accordance with rules and regulations.

By tracking infractions and incident, New York State builds transparency into their practices and into what happens on their track. If citing individuals that are breaking the rules will help to mitigate risk on the track – possibly racing incidents can also be minimized. With a deeper look at the infraction and incidents, contributing factors might provide insight into what leads to the high number of infractions. Armed with this information, tracks can put the proper constraints in place in the future to minimize infractions that could also minimize incidents (Woznica, *Racetrack Infractions and Incidents*, 4).

Before diving into the scripting related to social media, it was important to do some general visualizations, as with all projects, to begin to reveal patterns and areas of interest. In reviewing the data, it was clear that some specific occupations on the track were cited more often for infractions and, as expected, this was among the owners, trainers and drivers (the ones that had all three roles at the track). It is important to also note that this is expected. One would expect a trainer to more infractions that say Food Service or Security, for example. The results of this initial visualization can be seen in the following figure.

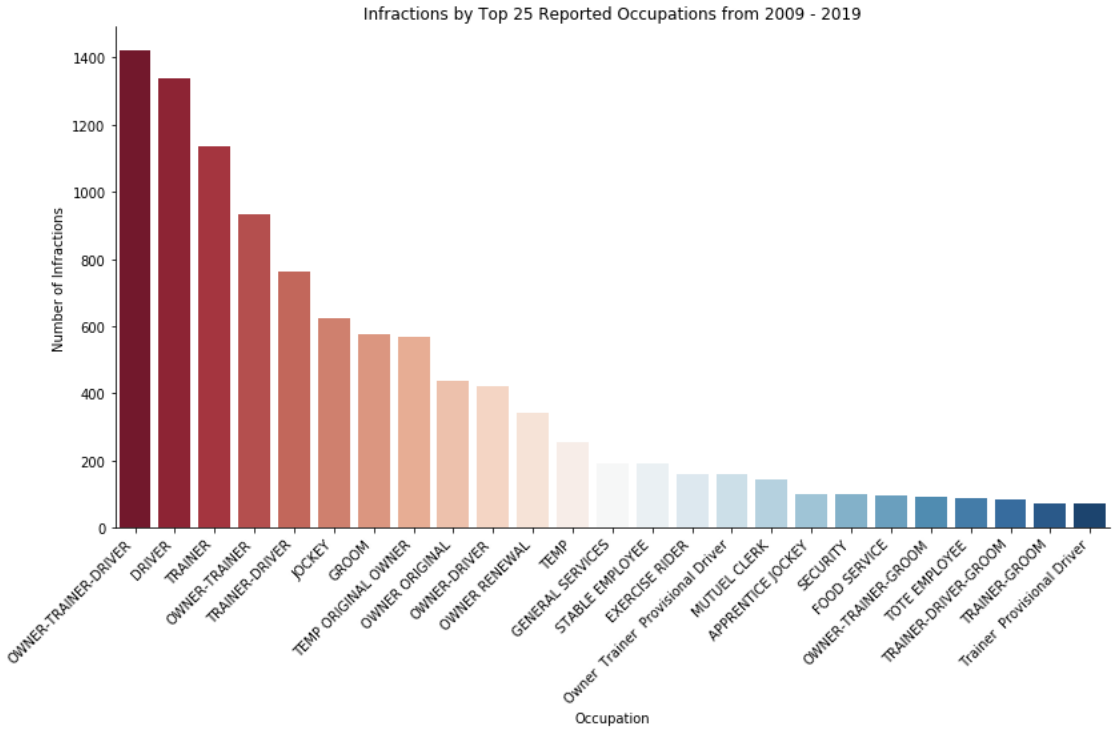



*Figure 15. Top Infractions by Occupation*

Before gathering more information on the infraction and incident descriptions, some additional visualization concerning the individual tracks and overlaps of infractions and incidents on the track. As can be seen below, there are tracks that have more infractions than others, but as was discovered through additional research – this can be contributed to just larger tracks with more races. One of the important things I learned in this class was the need for normalizing information for situations such as these.

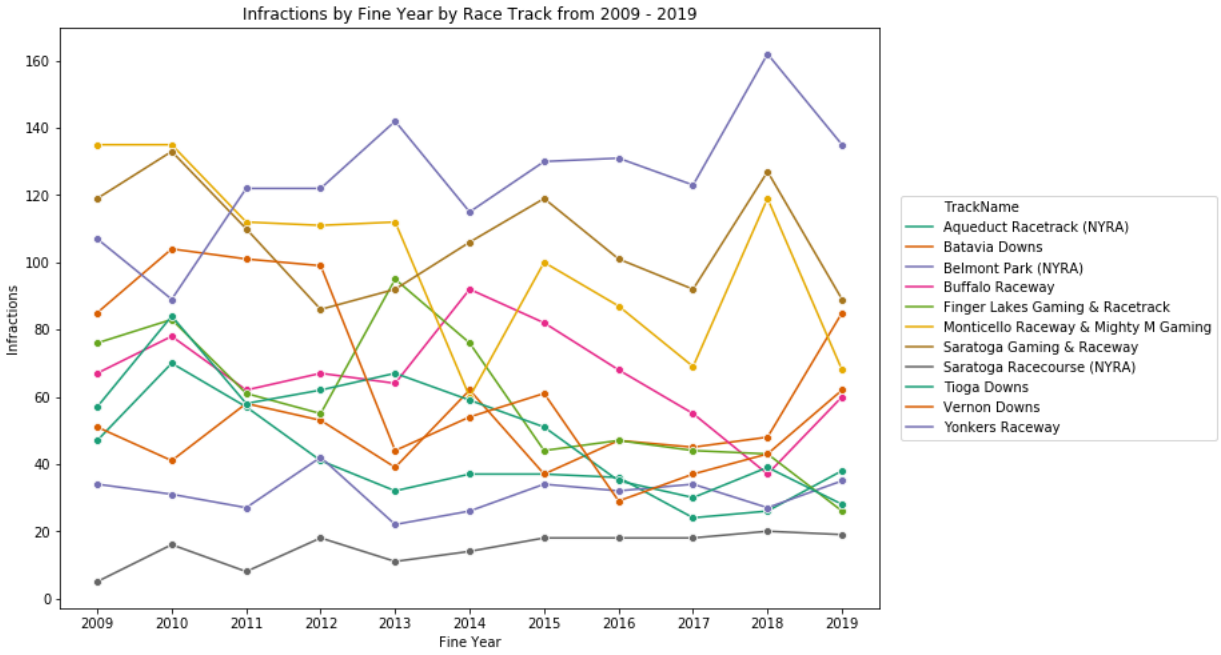



*Figure 16. Infractions by Fine Year by Race Track*

An important part of this project was to determine if there were specific incident types and compare that to any infraction the same day as an incident, I did a visualization on the 59 occurrences where overlap occurred.

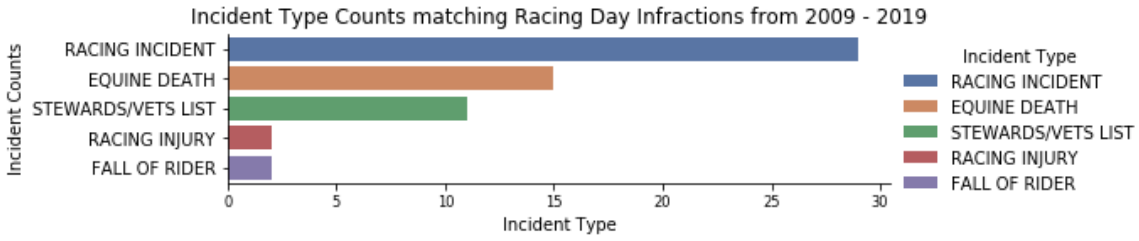



*Figure 17. Incident Types from Merged Data*

As seen in the previous section, "Racing Incidents" were most common. Unfortunately, "Equine Death" was the second most common incident type. To take this a step further, it was important to see if the incidents and violations were related in any way. For example, was it the same jockey or trainer that

was involved? The results of this investigation are shown in the simple print out generated showing the merged data in the following figure.

```
Matching Full Name and Jockey Driver
Year                                                          2019
Date                                                    08/07/2019
Track                                                 Batavia Downs
Division                                                    Harness
Occupation                                                  TRAINER
Full Name                                          KEVIN J. CUMMINGS
Trainer                                          ANTHONY C. CUMMINGS
Jockey Driver                                      KEVIN J. CUMMINGS
Horse                                              Good Lookin Woman
Incident Type                                        RACING INCIDENT
Type                                                            Fine
Ruling Text            While driving #7 Silver Arrow in the 9th race ...
Incident Description   H:Good Lookin Woman - D: Kevin Cummings - Hors...
Name: 18, dtype: object
Matching Full Name and Jockey Driver
Year                                                          2019
Date                                                    08/07/2019
Track                                                 Batavia Downs
Division                                                    Harness
Occupation                                                  TRAINER
Full Name                                          KEVIN J. CUMMINGS
Trainer                                          ANTHONY C. CUMMINGS
Jockey Driver                                      KEVIN J. CUMMINGS
Horse                                              Good Lookin Woman
Incident Type                                        RACING INCIDENT
Type                                                            Fine
Ruling Text            While driving #8 Ideal Ice in the 6th Race at ...
Incident Description   H:Good Lookin Woman - D: Kevin Cummings - Hors...
Name: 19, dtype: object
Matching Full Name and Trainer
Year                                                          2019
Date                                                    05/25/2019
Track                                        Saratoga Gaming & Raceway
Division                                                    Harness
Occupation                                                  TRAINER
Full Name                                          LISA M. ZABIELSKI
Trainer                                            LISA M. ZABIELSKI
Jockey Driver                                       JAY L. RANDALL
Horse                                                Love Yourself
Incident Type                                  STEWARDS/VETS LIST
Type                                                            Fine
Ruling Text            You failed fo timely file satisfactory documen...
Incident Description   Love Yourself - tr. Lisa Zabielski - lame in l...
```

*Figure 18. Infractions and Incidents with Same Individuals Involved*

As shown here, Kevin Cummings was the Driver on the same day at Batavia Downs when there was a Racing Incident. However, if you notice the text for the Incident and the Ruling – different horses were involved. The infraction was with a horse called “Silver Arrow” and again with a horse named “Ideal Ice.” However, the incident occurred with a horse called “Good Lookin’ Woman”. In the third matching set of data, Lisa Zabielski was involved as the trainer and the horse “Love Yourself” was lame and therefore put on the Stewards/Vet List. It is unclear if the ruling was for the same occurrence.

As previously, mentioned, this course taught students how to access social media accounts, in this case, Twitter, to find and analyze unstructured information. This was accomplished using standard python packages and storing the resulting JSON tweets in MongoDB for analysis. First, tweets from the racetracks themselves, were reviewed and a word cloud of the most common words found in

those tweets was compiled a displayed in the following figure. To create word clouds, the tweets had to be deconstructed by removing punctuation, tokenizing the tweets into individual words, and then creating a word bank of all the words and their frequency. This technique was then used in subsequent courses as well.


*Figure 19. Word Cloud of Commonly Appearing Words*

The information displayed in the tweets shown above were not unexpected as one would expect the race track to talk about wins, the weather, the beauty and the race park. To add a little more insight, additional tweets were found from "@racingwrongs" which is a twitter account that focuses on the dangers of horse racing and bringing them to the attention of the public. This provided an entirely different slant on what happens on a racetrack. The results of analysis o these tweets are shown in the figure below.


*Figure 20. Word Cloud of Commonly Appearing Words from @racingwrongs*

No additional modeling was done using these tweets, but the exercise to capture them, store and then extract them for use in data analysis was an excellent exercise and a wonderful skill to learn.

*Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Collecting and organizing data

  The collection and organization of data was a key part of this course. As mentioned, this course included accessing social media unstructured and storing this information in a structured way – in MongoDB. Using these new methods provided me with the skills to gather data from new sources and storing and organize this data in new ways. After storing these tweets in a NoSQL database (MongoDB), standard python functions were called to access this data and store in in a structured data frame that was used later to access the actual text of the tweets.

- Identifying patterns in data via visualization, statistical analysis, and data mining.

  Again, the final project in this course provided the option to create initial visualization of the data and review standard statistics to gain further insight into the data. For example. I was able to review the average fine by occupation or role and compare incidents on the track and where there was overlap between these incidents and the infractions that may have happened

at that same track on the same day. Then, by tokenizing the data found in the various tweets, word clouds were created to visualize the most common seen words in these tweets.

- Developing alternative strategies based on the data.

  In this particular project, I attempted to determine if there was some connection between an infraction at the track and any incident on that track. However, as I did further analysis, it was clear that this was not obvious. By working with melding and transposing techniques, the proper comparisons and overlaps could be explored to review if the overlap of infractions and incidents have anything in common.

- Developing a plan of action to implement the business decisions derived from the analyses.

  This is something that I found one of the most impressive parts of the Applied Data Science program at Syracuse University. As I reviewed the results of the analysis, one of the key components of any homework or project was "what does this mean?" and "how can these results help direct business decisions?" By learning to summarize the results and then find actionable insight that can be used to provide a plan of action or an approach for business to follow to take action. This skill is really what data analysis is all about. If we can gain insight into our processes, our data and what it means – we can take action to improve the business.

- Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

  As mentioned in the previous objective, documenting results is of the upmost importance. This includes documenting information for all levels the organization in a manner that is appropriate for the reader. Being able to make introductions and conclusions that reach a range of audiences with the tone and correct level can have a direct effect on the impact of the data, the analysis and the findings. This course instructed us to make sure that we provided these types of documentation to support our findings and reach the proper level.

- Synthesizing the ethical dimensions of data science practice (e.g., privacy).

  Taking a look at twitter and social media accounts opened up the discussion around who owns tweets and copyright infringement. In this case, the tweets were never published outside of this college paper. Entire tweets were never used in their entirety and only words and combinations of words were reviewed. Understanding privacy laws and what is public domain is very important when obtaining data from external and public sources like social media.

## *Natural Language Processing*

*Project: Classification of Racetrack Incidents*
*Course Number: IST664*

Finally, Natural Language Processing took over where Data Analytics left off by continuing the use of unstructured data and adding text analytics to predict outcomes. In this case, I revisited the race track and reviewed the incidents that took place on New York State tracks from 2009 through 2019 and the description of each incident and how this could be used to determine if a horse died or had to be euthanized as a result of a racetrack incident.

As with all projects, I started with initial visualizations of the data to help develop a plan of action for predicting the outcome on the track. First, it was important to gather words that appeared most often in the description of the incidents on the track. Using the same techniques that were used in Scripting for Data Analytics, the description for each incident was tokenize and then the top words were gathered into the word cloud displayed below.


*Figure 21. Word Cloud of Commonly Appearing Words*

To the casual observer, the information may not provide much insight, but to me – it gives a wealth of information. For example, words liked “ambulanced”, “vanned”, “retired”, and “collapsed” give details on the rider/driver and the horse. The horse could have collapsed and then was vanned (like a “horse ambulance”) because of some injury. Then there is additional information provided with words like “suspensory”, “fx”, ‘pastern”, “seasamoid” and “fetlock” which are words that denote injury

(suspensory and fx – or fracture) and the other talk about location on the horse "pastern", "sesamoid" and "fetlock." This information provided great insight into the incident descriptions.

To better understand how words were used in the descriptions is also important to better understand how combinations of words were used. This was important to derive which specific types of statistical methods should be used to predict incident outcomes. As shown in the following figure, looking at words that appear in pairs – like "fx, rf" which means fracture right front and "euthanized track" which implies the horse was euthanized on the track help to gain additional insight into incident results and similarities and commonalities among these incidents which will aid in predicting outcomes.

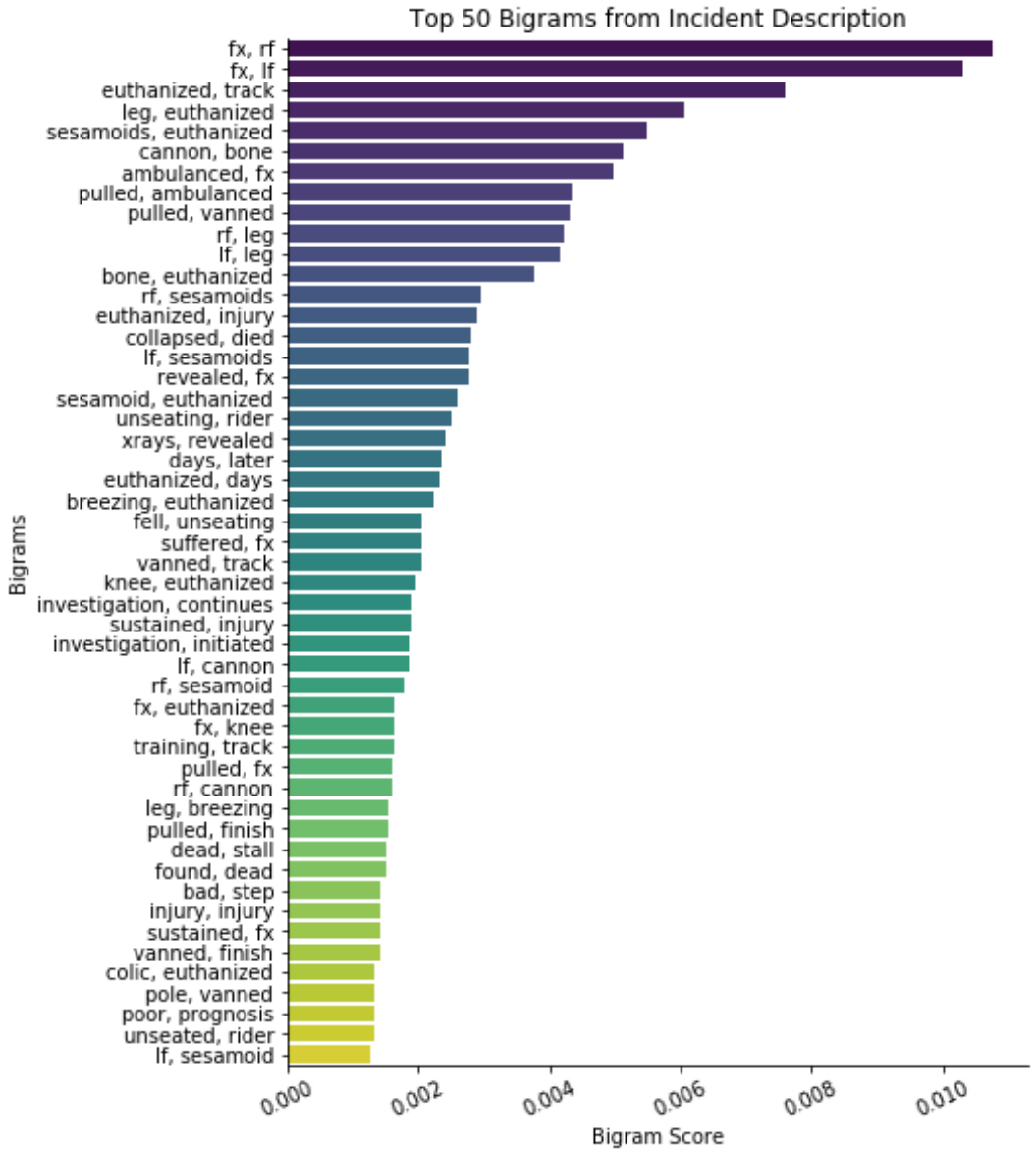



*Figure 22. Top 50 Bigrams in Incident Description (Normalized)*

Now that I had a better understanding of the data, it was time to focus on using the short description of the incident and use this information to classify the incident. As we learned in this class, in order to properly predict the outcome – we need have a balanced data set. This means that there should be a similar number of incident descriptions for each outcome to be predicted. In order to determine this, a quick frequency table was created by possible outcome as shown in the figure below.

| Frequency | Outcome |
|---|---|
| 1020 | Unknown |
| 909 | Steward's List |
| 778 | Equine Death |
| 580 | Euthanasia |
| 307 | Injury |
| 304 | Accident |
| 168 | Lameness |
| 109 | Equine Injury |
| 46 | Equine Injury / Equine Death |
| 3 | Death |
| 1 | Lame no death |
| 1 | death |

*Figure 23. Table of Incident Frequency by Outcome – Death or Injury*

As can be seen in this table, this is a very unbalanced set that is very skewed with certain outcomes having a significantly larger frequency. In order to remedy this situation, code was created to review the outcome and assign it to a consolidated set of possible outcomes:

- Death
- Euthanasia
- Injury
- Accident
- Lameness
- Other

As you can see from the visualization in the following figure, this data is still very skewed for "Other" incidents – which could be jockey incidents, horse falls without injury, horses that DNF (did not finish) the race without injury or some other type of incident on or around the race track grounds.

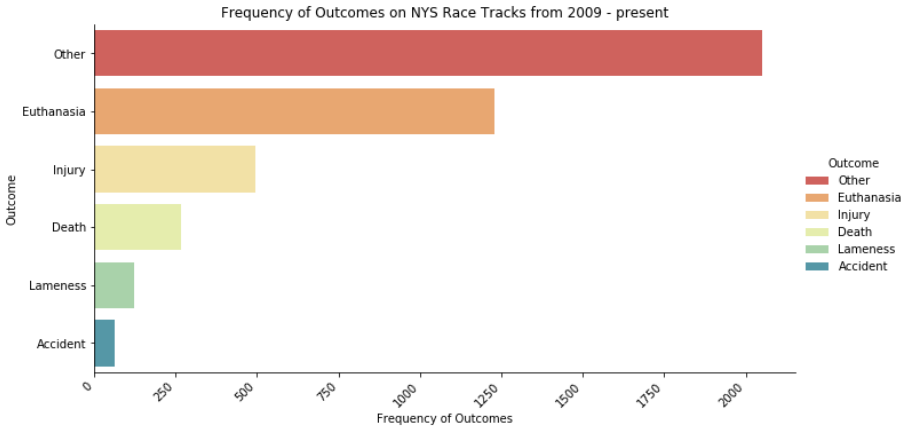



*Figure 24. Initial Outcomes for Classification*

In an attempt to properly balance this information, the resulting data frame was then condensed to group the "Injury", "Lameness" and "Accident" outcomes into a single "Injury" tag. In addition, the "Other" outcome incidents were removed from the dataset completely. The resulting data frame had less rows but with only three possible outcomes: "Euthanasia", "Death", "Injury".

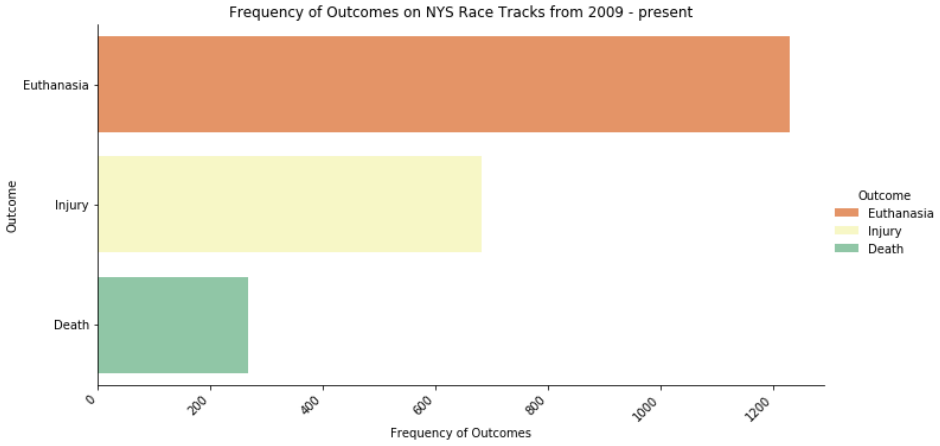



*Figure 24. Final Outcomes for Classification – Gold Standard*

Although this data was not completely balanced, it was a better representation for the purposes of this classification exercise. This initial work provided the proper backdrop to use various classifier features sets and the Natural Language Toolkit Naïve Bayes classifier. This course stressed the importance of defining the proper features sets, creating a "gold standard" list of correct classification labels, separating the test and training set and then running cross validation with multiple folds and averaging the fold accuracy.

As can be seen in the table below, several classification feature sets were used (unigrams, bigrams, trigrams and parts of speech) and the consolidated three (3) outcome data set along with a larger one with five (5) and one with six (6) outcomes. This table shows the accuracy and how the accuracy

improved with the consolidated more balanced data set and just using the most common words (bigrams) outperformed the other Naïve Bayes classifiers.

| *Data Set Outcome Count* | Classifier Feature Set | Mean Accuracy | Precision Micro-average | Recall Micro-average | F-Measure Micro-average |
|---|---|---|---|---|---|
| *Three Outcomes* | Unigrams | 94.93% | 91.0% | 90.5% | 90.7% |
| | Bigrams | 94.88% | 91.8% | 91.8% | 91.8% |
| | Trigrams | 94.75% | 88.8% | 88.6% | 88.7% |
| | Parts of Speech | 94.70% | 87.8% | 87.2% | 87.5% |
| *Five Outcomes* | Unigrams | 91.43% | 83.8% | 83.7% | 83.7% |
| | Bigrams | 91.38% | 85.9% | 85.7% | 85.7% |
| *Six Outcomes* | Unigrams | 82.68% | 82.6% | 84.0% | 83.0% |
| | Bigrams | 82.75% | 82.8% | 84.0% | 83.1% |

*Figure 25. Overall Experiment Summary Accuracy Table*

The concepts learned around classification and prediction as well as the need to properly prepare the data set in order to achieve optimal performance.

### *Learning Objectives Satisfied*

Among the concepts mastered in this course, the following were most closely mapped to the goals of the Syracuse Applied Data Science Program:

- Identifying patterns in data via visualization, statistical analysis, and data mining.

  As with other courses, the need to do primary visualization can help to guide a statistician down the path to the correct models and analysis. Data mining was an important concept in this class and one that was stressed in class often. By mining this data set down to a consolidated data set proved to be very important for the project.

- Developing alternative strategies based on the data.

  As mentioned, an important concept in this course was understanding your data and what strategy to take based on the layout of the information in the data set. In this case, the data led us to refine the data and consolidate outcomes so that the proper models could be used for prediction. Using this alternative strategy helped to improve the overall prediction accuracy and helped to lay out using different feature sets to aid in improving prediction accuracy.

- Demonstrating communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

  As before, I cannot stress the importance of this skill. I have worked as a technical writer, a technical marketing engineer, a solutions specialist and an education consultant in my career so far. Knowing your audience and speaking the proper level with the key information is what

can make our break your results. This course stressed again, the importance of explaining the data (in this case, the racetrack, and horses) so that the results can have the proper impact to the proper groups.

# Conclusion and Reflection

As reflected in this document, the learning objectives of the master's in Applied Data Science at Syracuse University were not only met but exceeded in the projects taken on to receive this degree. The selection of projects outlined in this document were just a few that helped to solidify my understanding of Data Science and allow me the flexibility to take on interesting projects while still permitting me to follow my passion.

Other courses and professors provided additional guidance and support that adding to the understanding of these learning objectives. For example, one key objective that I feel is so very important in the working world is that of being able to communicating about the data, the analysis and conclusions. Professor Ami Gates was one professor who stood out in this area and stressed that there should be no statistical jargon used in introductions or conclusions. These were for an executive or business audience and must convey the goals and the data at the proper level. This is such a strong skill and a very key strategy when putting papers together.

By taking time to reflect of the course projects embarked on in this program and how these courses helped to gain insight into how each of the seven learning objectives were satisfied and helped me to remember why I selected Syracuse for my second master's degree. This program provides many opportunities to learn skills that are specifically designed to applying these tools to applicable situations in the working world.

# Works Cited

"Syracuse Applied Data Science Master's Degree." *Applied Data Science Master's Degree*, Syracuse University, ischool.syr.edu/academics/applied-data-science-masters-degree/.

Woznica, Joyce L. *Classification of Racetrack Incidents: Analysis and Classification of New York State Racetrack Incidents between 2009 and present using Natural Language Processing*. Course IST 664, 2019.

Woznica, Joyce L. *Horse Ride Preparation Process Improvement*. Course MBC 638, 2018.

Woznica, Joyce L. *Joyce Woznica Horse Colic Project*. Course IST 707, 2019.

Woznica Joyce L. *Joyce Woznica Database Project: Horse Records*. Course IST 659, 2018.

Woznica, Joyce L. *Racetrack Infractions and Incidents*: *Analysis of New York State Racetrack Rulings and Incidents between 2009 and 2019*. Course IST 652, 2020.

Woznica, Joyce L. *Tracking the Danger Zone: The Dangers of Horse Racing in New York State.* Course IST 719, 2020.