Joyce Woznica

jlwoznic@syr.edu

**IST 707 – Jeremy Bolton**December 10, 2019

Joyce Woznica  
Horse Colic Project

Contents

[Introduction 4](#_Toc26729689)

[Types of Colic in Horses 4](#_Toc26729690)

[Goals/Questions 5](#_Toc26729691)

[Analysis 5](#_Toc26729692)

[The Data 5](#_Toc26729693)

[Data Load 6](#_Toc26729694)

[Data Cleanse 7](#_Toc26729695)

[Exploratory Data Analysis 11](#_Toc26729696)

[Descriptive Statistics 11](#_Toc26729697)

[Visualization 12](#_Toc26729698)

[General Visualization 12](#_Toc26729699)

[Detailed Visualizations 17](#_Toc26729700)

[Modeling 20](#_Toc26729701)

[Decision Trees 20](#_Toc26729702)

[Association Rule Mining (ARM) 26](#_Toc26729703)

[Naïve Bayes 31](#_Toc26729704)

[Support Vector Machine (SVM) 34](#_Toc26729705)

[Multiple Linear Regression 35](#_Toc26729706)

[Logit and Probit 41](#_Toc26729707)

[Neural Network 42](#_Toc26729708)

[Results 44](#_Toc26729709)

[Conclusions 44](#_Toc26729710)

[References 46](#_Toc26729711)

List of Figures

[Figure 1: Table of Horse Colic Variables 6](#_Toc26729712)

[Figure 2: Load and Initial Structure of Horse Colic Data 6](#_Toc26729713)

[Figure 3: Missing Data Plot from the Horse Colic Data 7](#_Toc26729714)

[Figure 4: Initial Summary of Colic Data 8](#_Toc26729715)

[Figure 5: Function to Replace Missing Numeric Data with Means 8](#_Toc26729716)

[Figure 6: Function to Replace Missing Factor Data with Modes 9](#_Toc26729717)

[Figure 7: Manual Updates of Modes for Special Cases 9](#_Toc26729718)

[Figure 8: Removal of Lesion 2 and Lesion 3 Data and Irrelevant Data 9](#_Toc26729719)

[Figure 9: Updated Data Frame with Lesion 1 Information 10](#_Toc26729720)

[Figure 10: Updated Data Frame by Removing incorrect Lesion 1 Rows 11](#_Toc26729721)

[Figure 11: Update Capillary Refill Time Rows 11](#_Toc26729722)

[Figure 12: Mix Up the Order of the Rows 11](#_Toc26729723)

[Figure 13: Summary Information from the Satisfaction Survey 11](#_Toc26729724)

[Figure 14: Describing the Horse Colic Data 12](#_Toc26729725)

[Figure 15: Summary Statistics of Some Horse Colic Continuous Variables 12](#_Toc26729726)

[Figure 16: Colic Outcomes 13](#_Toc26729727)

[Figure 17: Colic Variable Histograms 13](#_Toc26729728)

[Figure 18: Additional Bar Plots and Outcome Frequencies 14](#_Toc26729729)

[Figure 19: Lesion Site related to Colic Outcome 15](#_Toc26729730)

[Figure 20: Lesion Type related to Colic Outcome 16](#_Toc26729731)

[Figure 21: Lesion SubType related to Colic Outcome 16](#_Toc26729732)

[Figure 22: Lesion Code related to Colic Outcome 17](#_Toc26729733)

[Figure 23: Feces to Colic Outcome 17](#_Toc26729734)

[Figure 24: Box Plot of Outcome by Pulse Rate 18](#_Toc26729735)

[Figure 25: Box Plot of Total Protein and Outcome 18](#_Toc26729736)

[Figure 26: Packed Cell Volume by Outcome 19](#_Toc26729737)

[Figure 27: Packed Cell Volume by Outcome 19](#_Toc26729738)

[Figure 28: Outcome by Respiratory Rate and Extremity Temperature 20](#_Toc26729739)

[Figure 29: Decision Tree Code 21](#_Toc26729740)

[Figure 30: Decision Tree One 22](#_Toc26729741)

[Figure 31: Decision Tree Two 23](#_Toc26729742)

[Figure 32: Accuracy of Decision Tree Data Frame 23](#_Toc26729743)

[Figure 33: Summary of Classification Tree Five 23](#_Toc26729744)

[Figure 34: Decision Tree Five 24](#_Toc26729745)

[Figure 35: Pruned Tree 24](#_Toc26729746)

[Figure 36: Accuracy of Decision Tree Data Frame 25](#_Toc26729747)

[Figure 37: Random Forest R-Code 25](#_Toc26729748)

[Figure 38: Random Forest Accuracy Matrix 26](#_Toc26729749)

[Figure 39: ARM Data Breaks for Continuous Variables 26](#_Toc26729750)

[Figure 40: Functions for Association Rules 27](#_Toc26729751)

[Figure 41: Calling myARM Function 27](#_Toc26729752)

[Figure 42: First Twenty Rules with Support, Confidence and Lift, sorted by Lift 28](#_Toc26729753)

[Figure 43: First Twenty Rules with Support, Confidence and Lift, sorted by Confidence 28](#_Toc26729754)

[Figure 44: Interactive Plot of Top Eight Rules by Decreasing Support 29](#_Toc26729755)

[Figure 45: Interactive Plot of Top Eight Rules Sorted by Decreasing Confidence 29](#_Toc26729756)

[Figure 46: Interactive Plot of Top Eight Rules Sorted by Decreasing Confidence 30](#_Toc26729757)

[Figure 47: Scatterplot of Top Twenty Rules by Decreasing Confidence 30](#_Toc26729758)

[Figure 48: Rule 1 with Support=0.223, Confidence=0.97 and Lift=1.64 31](#_Toc26729759)

[Figure 49: Naïve Bayes with “train” and “naivebayes” functions 32](#_Toc26729760)

[Figure 50: Naïve Bayes R-Code for 4-fold Validation and Accuracy 33](#_Toc26729761)

[Figure 51: Naïve Bayes using naïve\_bayes function with Accuracy 34](#_Toc26729762)

[Figure 52: Naïve Bayes mlr R-Code 34](#_Toc26729763)

[Figure 53: SVM R-Code for Each Kernel 35](#_Toc26729764)

[Figure 54: SVM Kernel Accuracy Matrix 35](#_Toc26729765)

[Figure 55: R-Code to Convert Outcome to Numeric 36](#_Toc26729766)

[Figure 56: Regression on all Variables 36](#_Toc26729767)

[Figure 57: Regression Coefficients and Correlation Coefficient 37](#_Toc26729768)

[Figure 58: Variables with < .07 p-values 38](#_Toc26729769)

[Figure 59: Correlation Coefficient for Best Model 39](#_Toc26729770)

[Figure 60: R-Code for Regression Prediction and RSME 39](#_Toc26729771)

[Figure 61: Error by Predicted Outcome 40](#_Toc26729772)

[Figure 62: Error by Total Protein, Pain and Packed Cell Volume 40](#_Toc26729773)

[Figure 63: Error by Total Protein, Packed Cell Volume and Respiratory Rate 41](#_Toc26729774)

[Figure 64: Finding best logit Model for Colic Outcome Prediction 41](#_Toc26729775)

[Figure 65: RSME by Model Type 42](#_Toc26729776)

[Figure 66: Best Probit Model for Colic Prediction 42](#_Toc26729777)

[Figure 67: Neural Network – Used Only Continuous Data Variables 42](#_Toc26729778)

[Figure 68: Plotted Neural Network 43](#_Toc26729779)

[Figure 69: Neural Network Detailed Matrix 43](#_Toc26729780)

# Introduction

*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.

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. The incidence of colic in the general horse population has been estimated between four and ten percent over the course of their lifetime. Clinical signs of colic generally require treatment by a veterinarian. 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.

## Types of Colic in Horses

* **Gas Colic –** Gas colic occurs when there is excessive buildup of gas within the intestines of the horse. These horses can often have a lot of flatulence.
* **Spasmodic Colic –** Spasmodic colic is the result of intestinal cramps or spasms. This type of colic can also have intestinal hyper motility.
* **Impaction Colic –** Impaction colic accounts for 10% of all colics attended by veterinarians. These occur where partially digested feed, typically roughage, builds up in the large intestine of the horse and stops moving, resulting in a blockage or impaction. With impaction colic, the horse is not passing dung.
* **Sand Colic –** Sand colic occurs in horses living in sandy areas, or horses fed from sandy ground. Fine particle sand builds up in the large intestines resulting in colic.
* **Twisted Gut –** A twisted gut occurs where a portion of the intestine twists on itself (intestinal torsion) or where a portion of intestine inverts into itself (intussusception). This uncommon type of colic accounts for less than 4% of colics overall, but it is very serious and life threatening.
* **Displacement/Entrapment Colic –** Displacements occur when an area of the intestine moves from its normal location in the abdominal cavity to somewhere else, naturally this is not a common type of colic. When the displacements cannot freely move back to its original location, it becomes an entrapment. Displacements and entrapments are very serious because this change in location stretches the blood supply to the area of intestine and can result it being compressed or squashed.
* **Strangulation Colic –** Strangulation colic is very uncommon, but very serious. Strangulation colic occurs when the blood supply to an area of intestines is cut off (strangulated). Cutting off the blood supply, results in rapid death of the intestine wall, a serious life-threatening situation.

It is important to realize however, that the vast majority of colics never have their exact causation determined. Happily, this “unidentified type” of colic, also has a recovery rate of over 95%. This can be interpreted as; most horses get a mild form of colic, which is successfully treated by their veterinarian, making further investigation unnecessary.

## Goals/Questions

By reviewing the information provided in a publicly available data set with symptoms and presentation of horses experiencing colic episodes, the hope is to find patterns and possible predict outcomes of the prognosis for the horse. This has both medical and economic significance. If the outcome is not favorable, the owner may elect to pass on surgery especially if there is no effective insurance policy in force.

By looking at some of the information provided in this colic dataset (from the UCI Machine Learning Database at <http://archive.ics.uci,edu/ml/datasets/Horse-Colic>) to determine if some of the symptoms and characteristics of the horse have some sort of influence on its need for surgery and the outcome (either of surgery or the horse - died, lived, euthanized). This “illness” is a problem for horse owners and very difficult to predict. Personal experience has made this problem especially important.

Insurance companies use information like this - similar to any actuary for life and medical insurance - to determine if a horse can be insured for colic. Horses that have coliced traditionally are not insurable for at least another year. After a colic-free year, they can have this exclusion removed. This is important for horse owners as well to understand if surgery will actually give their horse a shot a survival or if it is not worth the expense.

This document provides details into the analysis completed on a colic data set, various prediction methods used and the conclusions.

# Analysis

## The Data

The data set available is a collection of symptoms and information for 299 horses that have had a colic episode. The data was provided in the following way.

| Variable | Definition |
| --- | --- |
| Surgery | yes or no |
| Age | adult (≥6 months), young (<6 months) |
| Hospital Number | Case number for hospital (only unique if horse treated one time only) |
| Rectal Temperature | Rectal temperature in degrees Celsius (normal=37.8) |
| Pulse | Heartrate in beats per minute (normal 30-40) |
| Respiratory Rate | Respiration rate (normal 8-10) |
| Extremities Temperature | Normal, warm, cool, cold |
| Peripheral pulse | Normal, increased, reduced, absent |
| Mucous membranes | Normal pink, bright pink, pale pink, pale cyanotic, bright red/injected, dark cyanotic (normal pink to bright pink is normal) |
| Capillary Refill Time | ≥ 3 seconds, < 3 seconds (< 3 seconds normal) |
| Pain | Alert (no pain), depressed, intermittent mild pain, intermittent severe pain, continuous severe pain |
| Peristalsis | Gut Activity. Hypermotile, normal, hypomotile, absent |
| Abdominal Distension | None, slight, moderate, severe |
| Nasogastric Tube | Gas coming from tube. None, slight, significant |
| Nasogastric reflux | None, ≥ 1 liter, < 1 liter |
| Nasogastric Reflux PH | Scale from 0 to 14 (normal = 7) |
| Rectal Examination (feces) | Normal, increased, decreased, absent |
| Abdomen | Normal, other, firm in large intestine, distended small intestine, distended large intestine |
| Packed Cell Volume | Number of red blood cells by volume (normal is 30-50) |
| Total Protein | Numeric. (normal is 6-7.5 (gms/dL) |
| Abdomen Fluid | Clear, cloudy, serosanguinous |
| Abdomen Fluid PH | Numeric |
| Outcome | Lived, died, euthanized |
| Surgical Lesion | Yes, no |
| Type of Lesion \* this is repeated 3 times Lesion1  Lesion2  Lesion3 | *1st Number* (site): 1-gastric, 2-small intestine, 3-large colon, 4-large colon and cecum, 5-cecum, 6-traverse colon, 7-rectum/descending colon, 8-uterus, 9-bladder, 11-all intestinal sites, 00-none  *2nd Number* (type): 1-simple, 2-strangulation, 3-inflammation, 4-other  *3rd Number* (subtype): 1-mechanical, 2-paralytic, 0-N/A  *4th Number* (code): 1- obturation, 2-intrinsic, 3-extrinsic, 4-adynamic, 5-volvulus/torsion, 6-intussuption, 7-thromboembolic, 8-hernia, 9-lipoma/selenic incarceration, 10-displacement, 0-N/A |
| Cp\_data | Pathology data present? Yes or no |

Figure 1: Table of Horse Colic Variables

### Data Load

The data was provided in a generic comma separated value (CSV) file. This was loaded using the *read.csv* function and then manipulation of the data was required.

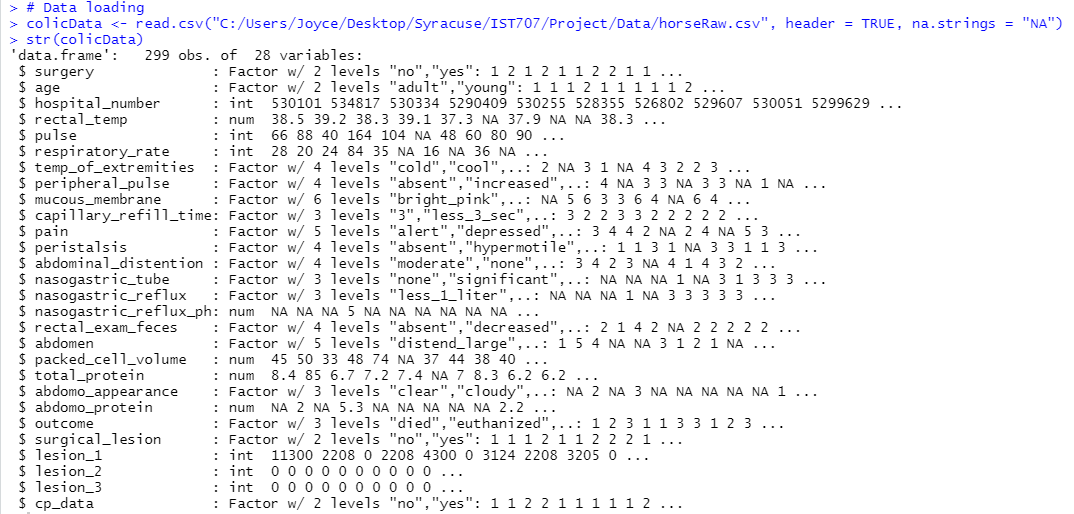


Figure 2: Load and Initial Structure of Horse Colic Data

### Data Cleanse

The dataset provided had 299 horses with 28 fields observed for each horse represented. Unfortunately, the data was missing from quite a bit of the data for each horse and decisions had to be made on how to handle the missing information. The total number of missing values was 1602 before any data munging was initiated. Missing data was addressed in a way that would minimize the skewing or misrepresentation of the data as much as possible.

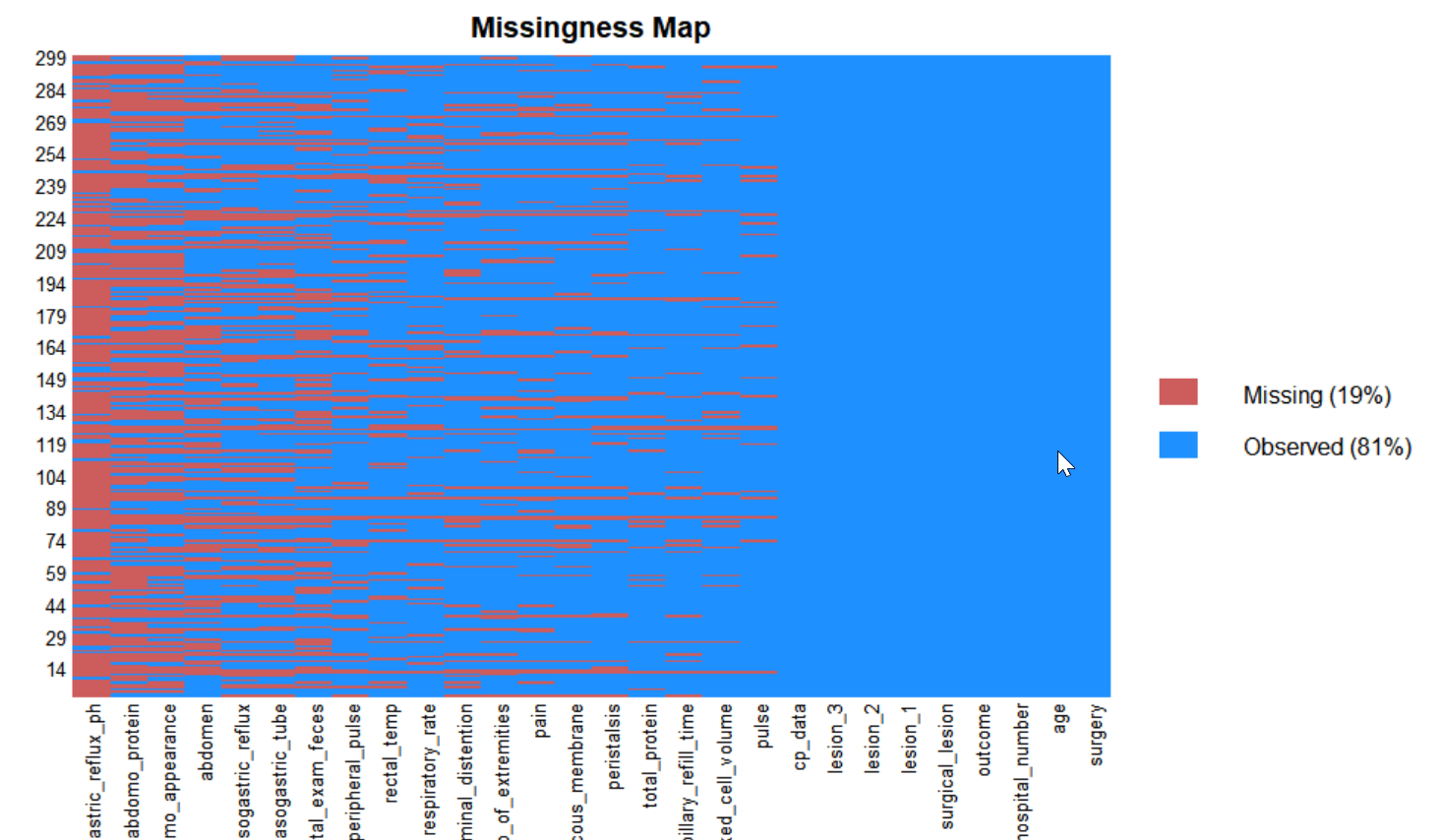


Figure 3: Missing Data Plot from the Horse Colic Data

To initially look at the missing data, a summary of the data after loading was executed as shown in Figure 4.

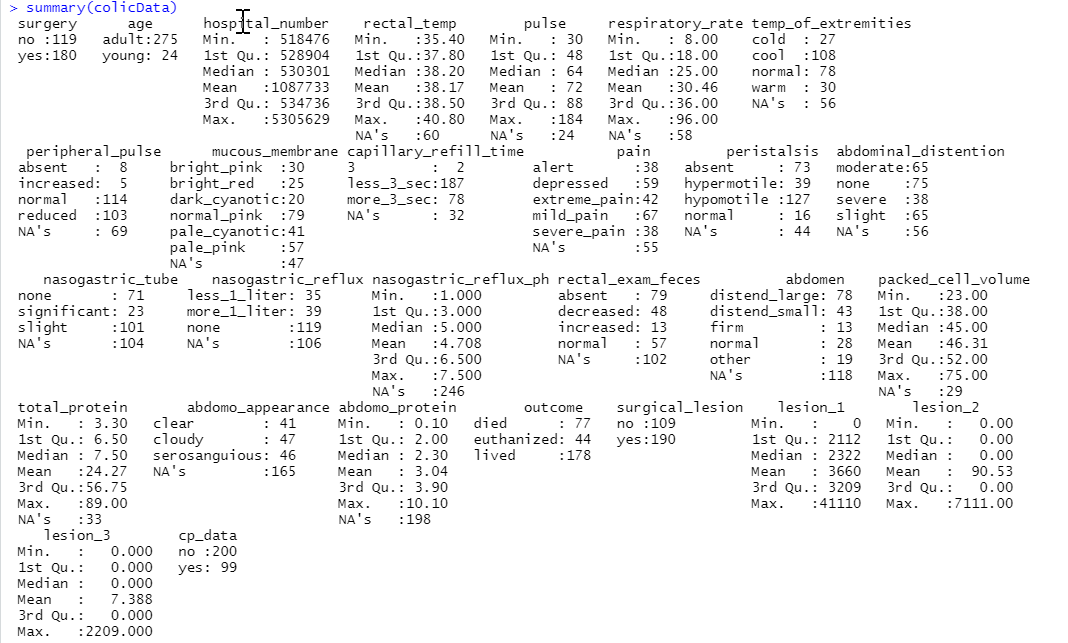


Figure 4: Initial Summary of Colic Data

First, a decision was made to use the mean for any numeric data missing from the data using the function and code below.

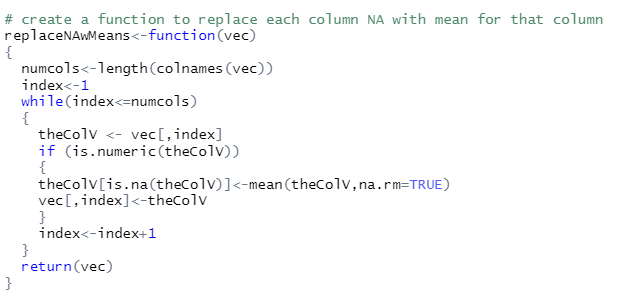
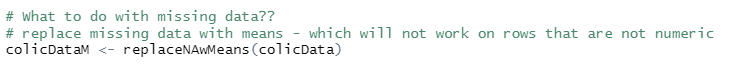
  


Figure 5: Function to Replace Missing Numeric Data with Means

The next step was how to address missing factor data. A function was created to replace missing factor data with the mean of each column.

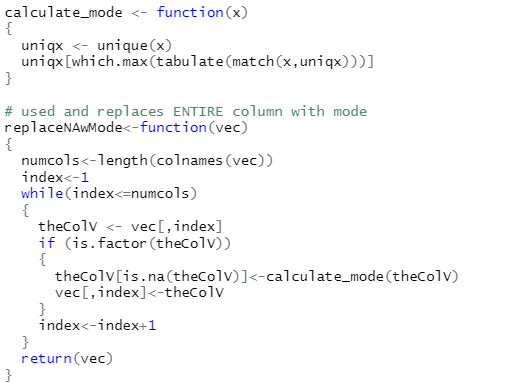
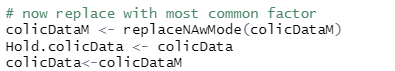
  


Figure 6: Function to Replace Missing Factor Data with Modes

This worked well except in cases where the mean was actually NA. Due to time constraints, those instances were modified manually in the code as special cases instead of modifying the mode replacement code to handle these situations. These initial replacements lowered the missing data from 1206 to 489, but this could be slightly skewed because some NAs could have been introduced with the mode function.

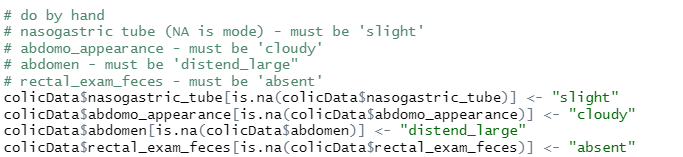


Figure 7: Manual Updates of Modes for Special Cases

These actions eliminated all missing data noted by “NA.” However, as shown in Figure 1 shown on page 7, the “lesion” fields collectively made up a significant set of data which was not easily accessible in the format provided. Some extensive replacement coding was done to replace the number combinations to represent the lesion data by location so that it could be properly used in the modeling.

First, it was determined that since the Lesion 2 (7 values for all 299 horses) and Lesion 3 (1 value for all 299 horses) data only have values for a very small amount of the horse population, that these columns could be eliminated from the modeling. In addition, the hospital number (unique case number) was also removed as it was determined to be irrelevant for modeling purposes.



Figure 8: Removal of Lesion 2 and Lesion 3 Data and Irrelevant Data

Since lesion 1 had 248 values for the 299 horses, it was determined that this column may shed some light on the colic outcomes. Temporary data frames were created and then these were modified to reflect the proper lesion location. In some cases, it was easier to manually complete the last few values that did not fall easily in the code created. This code can be found in the following figure.

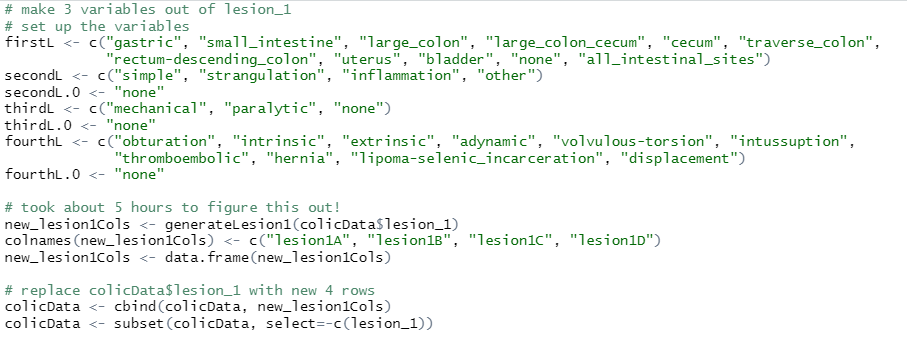
   


Figure 9: Updated Data Frame with Lesion 1 Information

There we six rows that had values in the lesion\_1 column that did not properly fit the digits as explained in the data dictionary. These rows were eliminated from the data set.

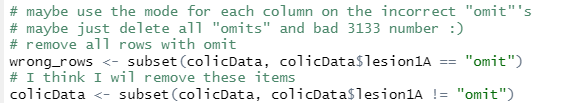


Figure 10: Updated Data Frame by Removing incorrect Lesion 1 Rows

Finally, there were a few rows that reflected bad capillary refill information and this was corrected with the normal value for this field so that the rest of the horse data for this equine could be used for modeling.

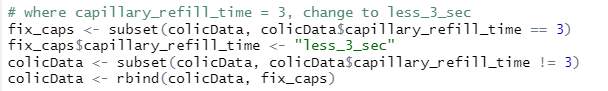


Figure 11: Update Capillary Refill Time Rows

Finally, the data needed to be properly randomized. This was done because the corrected lesion information was just appended to the original matrix each time.



Figure 12: Mix Up the Order of the Rows

## Exploratory Data Analysis

### Descriptive Statistics

The initial analysis of the data included some general study of the data. To accomplish this, generic descriptive statistic tools were used to find out a bit about the data before doing any extensive modeling and visualization.

For example, the frequency of various data in the horse colic data set as shown in the following figure.



Figure 13: Summary Information from the Satisfaction Survey

Using the *describe* function, general descriptive statistics were generated on the data set.

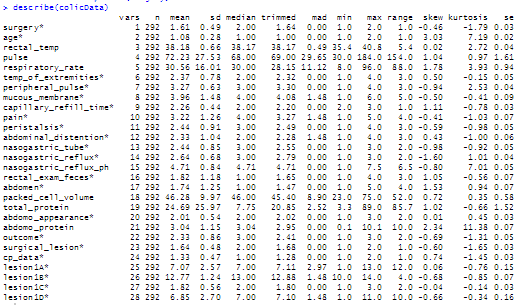


Figure 14: Describing the Horse Colic Data

Although there were only a few continuous variables in the colic data set, using the *stargazer* package, more descriptive statistics (mean, standard deviation, maximum and minimum) were displayed for these variables as shown in the following figure.

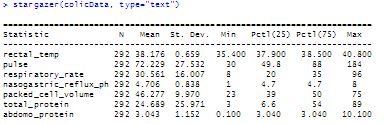


Figure 15: Summary Statistics of Some Horse Colic Continuous Variables

## Visualization

The next step was to do some initial visualizations of the data to better determine how the variables relate to the outcomes. This provides insight into which variables might be of interest for modeling.

### General Visualization

Initially, very generic plots were done on the data. For example, a simple bar graph of the outcomes and the frequency of these for the 292 horse subjects was plotted as shown in the figure on the following page, Figure 16.

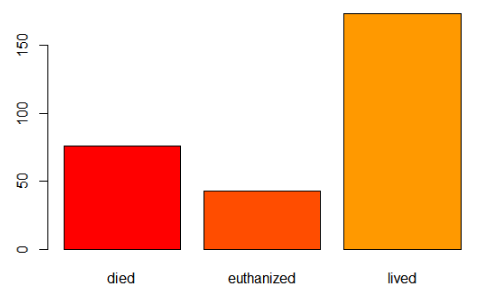


Figure 16: Colic Outcomes

Because of prior knowledge of some related variables to colic outcomes, histograms were run on several variables as shown in the plots in the figure below.

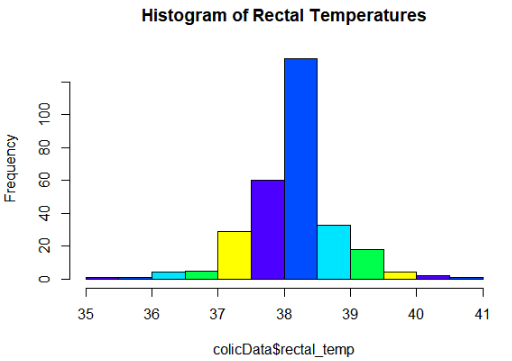
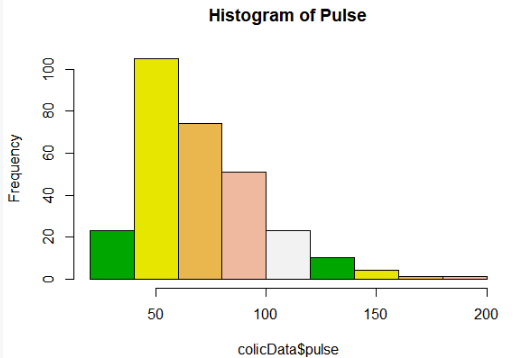
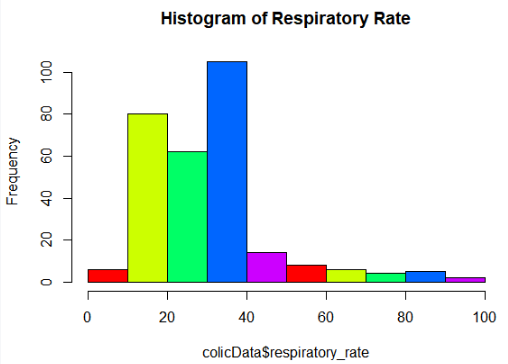
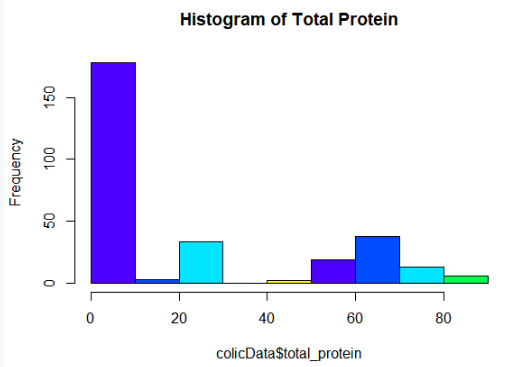
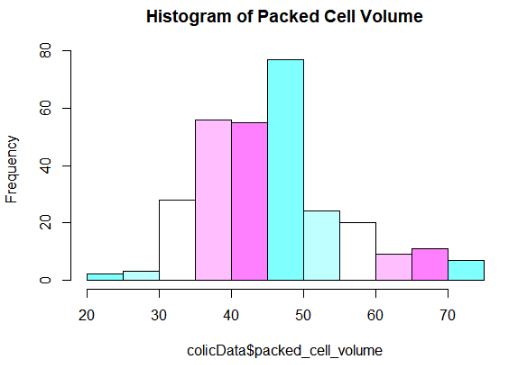
    

Figure 17: Colic Variable Histograms

Following these general bar plots, additional plots were done on non-continuous variables to better understand the relationship between these values and the colic outcome for the horse.

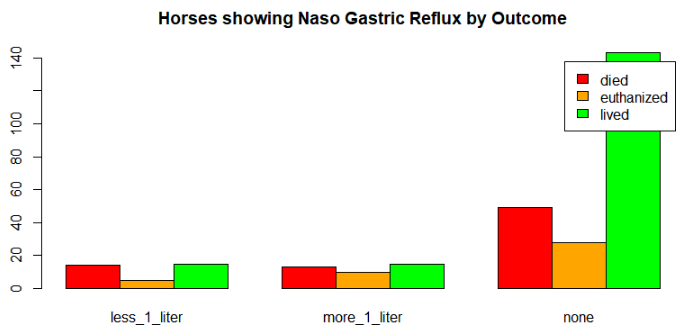
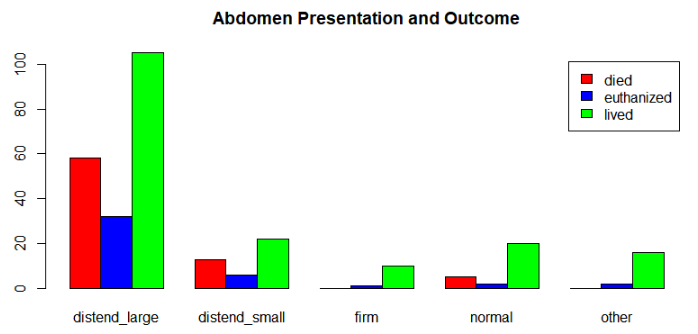
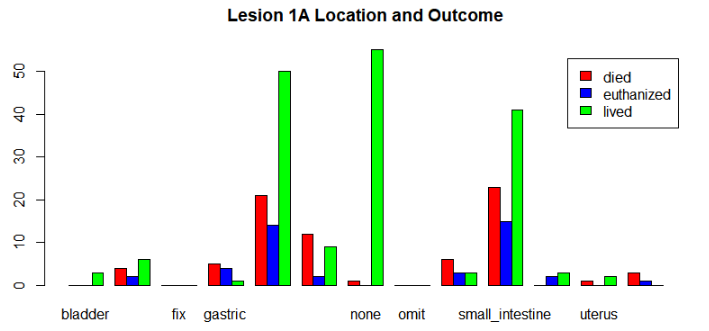
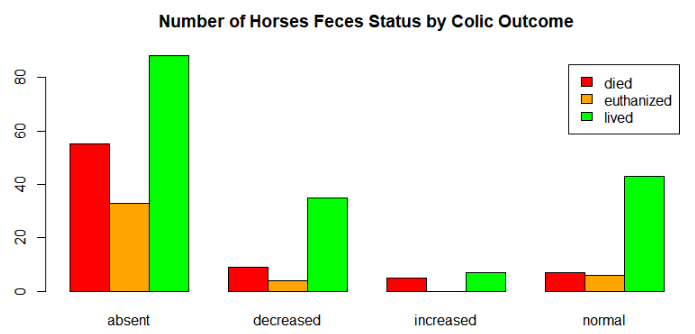
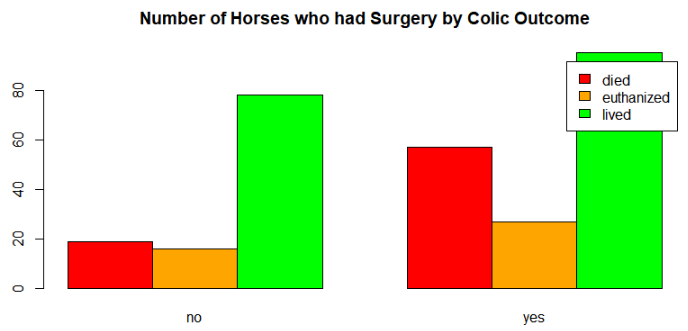
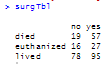
     

Figure 18: Additional Bar Plots and Outcome Frequencies

Reviewing this information provides some insight into which variables present higher risk of death for the horse. For example, more horses died or were euthanized after surgery: 84 out of 179 horses or approximately 47%. This is significantly different than those that did not face surgery: 35 out of 113 horses or approximately 31%. However, this could be attributed to the fact that those horses that face surgery are in a far greater state of distress or more advanced colic than those that do not go to surgery.

To further investigate the Lesion 1A locations and the relationship to the colic outcome, this particular plot was repeated with *ggplot* to reflect the information a bit more clearly as shown in the following figure.

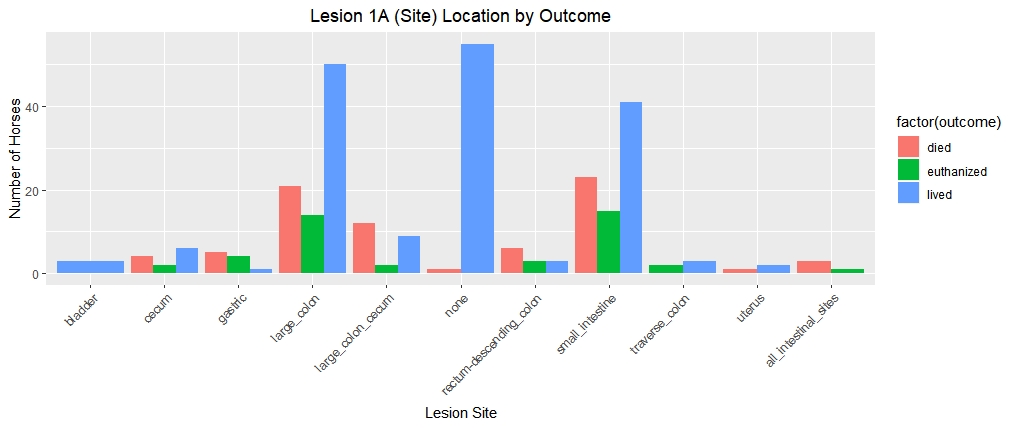


Figure 19: Lesion Site related to Colic Outcome

This bar graph shows that if the lesion is found in the large colon or small intestine, the horse is more likely to die than if the lesion is found elsewhere. However, by looking at the percentages for these lesion locations, a different pattern emerges. Although less common, a lesion in the cecum, the pouch connected to the junction of the small and large intestines, has a death and euthanasia percentage of 50% based on 12 horses presenting this location with 4 horse deaths and 2 euthanized horses. By looking the lesion in the large colon/cecum, the rate is quite high with about 61% death or euthanized out of the 23 horses presenting lesions in this location. Although both small and large intestine show high deaths and euthanized horses, the percentages are approximately 48% and 42%, respectively. The mortality rate for lesions in all intestinal locations is 100%. This graph proved very useful in reviewing the outcome related to lesion location.

This led to creating additional graphs based on the lesion information. Nothing surprising in the lesion type as strangulation type is the most common colic type that leads to death as shown in the following figure.



Figure 20: Lesion Type related to Colic Outcome

The subtype lesion information was less insightful as “none” was a prominent value and this does not provide any additional information.

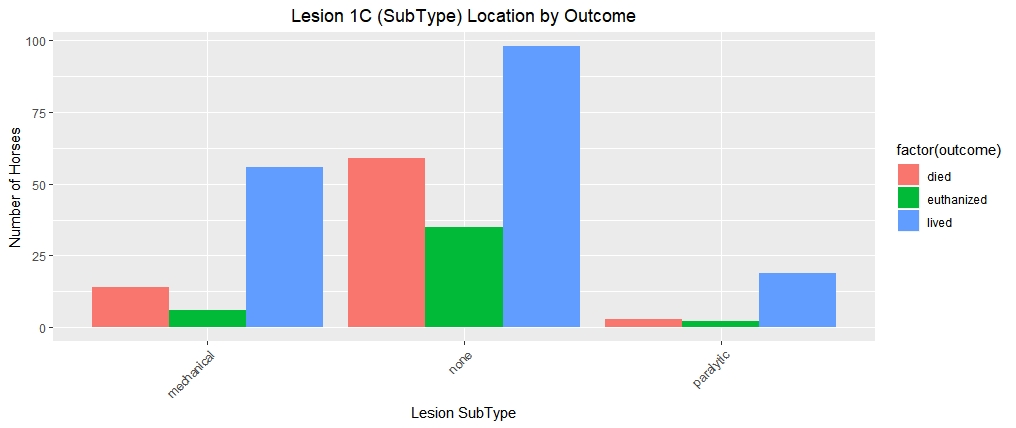


Figure 21: Lesion SubType related to Colic Outcome

The code provided some practical information. For example, volvulus-torsion, or when the intestine twists around itself causing obstruction, led to the most deaths and euthanized horses. This is not surprising as it requires surgery in most cases and depending on the timing, the obstruction can cause death of the intestine compromising its ability to function properly. In addition, a blood clot that breaks and is carried through the blood stream (a thromboembolic event) also had a high mortality rate. A fatty tumor incarcerated in the intestine also showed a significant mortality rate as shown in the figure on the following page, Figure 22.

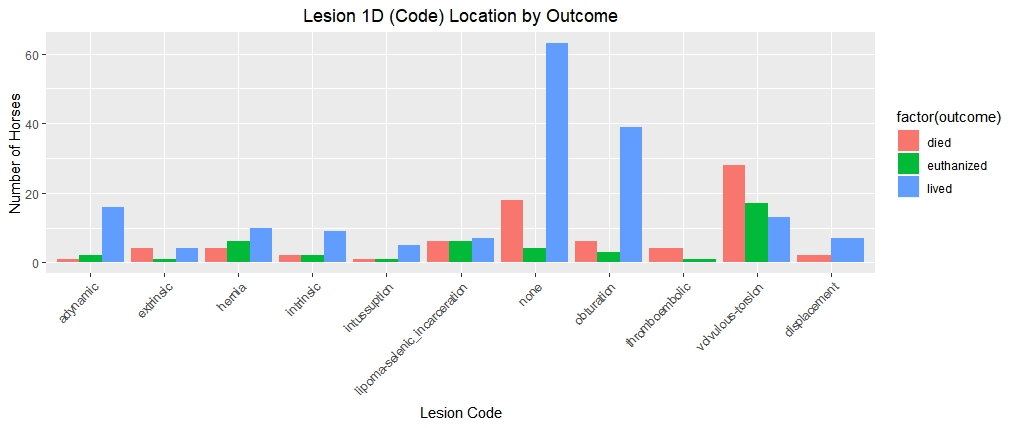


Figure 22: Lesion Code related to Colic Outcome

This insight into the individual variables continued by looking at the lack of or decreased feces in the horse showing that a blockage may be in the intestine as reflected in the following figure. Here it is easy to notice that the complete absence of feces is a very big red flag as related to mortality.

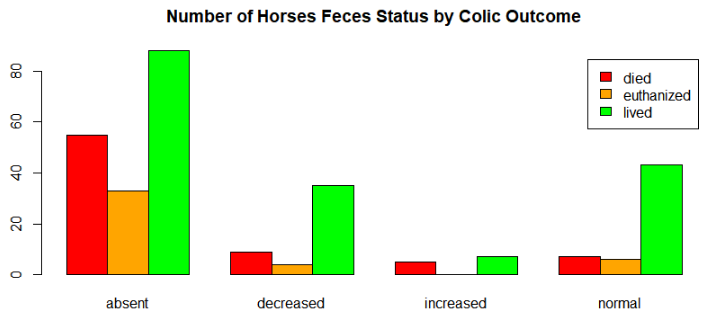


Figure 23: Feces to Colic Outcome

### Detailed Visualizations

Additional visualization proved useful as well. As shown in Figure 24, the higher the pulse rate indicating that the horse is likely in distress, the more likely the horse would die or be euthanized.



Figure 24: Box Plot of Outcome by Pulse Rate

Equally interesting was the total protein in the blood and the outcome of the colic incident. As shown in the following figure, the lower the protein, the more likelihood of death, but this was also not conclusive because low protein also resulted in life as well.

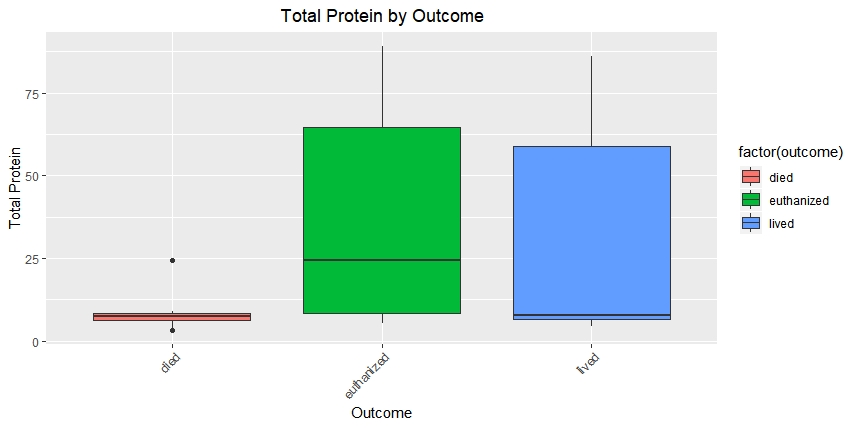


Figure 25: Box Plot of Total Protein and Outcome

Some additional visualizations were run to see if any other insight could be gained. The mortality rate peaks when the packed cell volume is around 60, as shown in Figure 26, as well as all outcomes peaked around 45. A normal packed cell volume is between 30 and 50, so this is not surprising, but clearly the higher the cell volume, the higher the likelihood of death as shown in the red line in the graph provided.

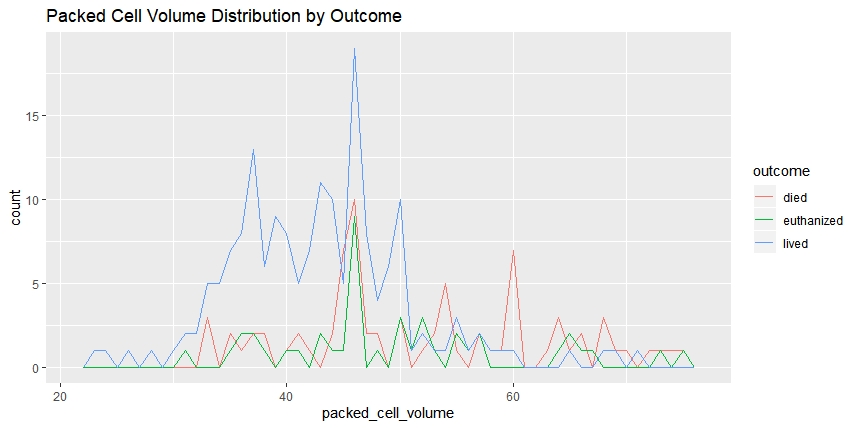


Figure 26: Packed Cell Volume by Outcome

To attempt to gain additional information on the total protein, another plot was done as shown in the figure below. In this figure, the higher death rate with protein under 15 is much more apparent.

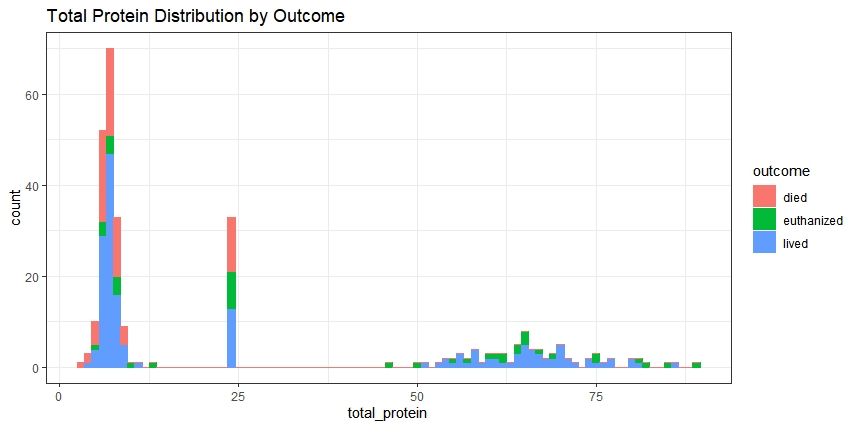


Figure 27: Packed Cell Volume by Outcome

To better understand how the blood is circulating through the extremities when the horse is experiencing colic, the following plot was run which also included the age of the horse.

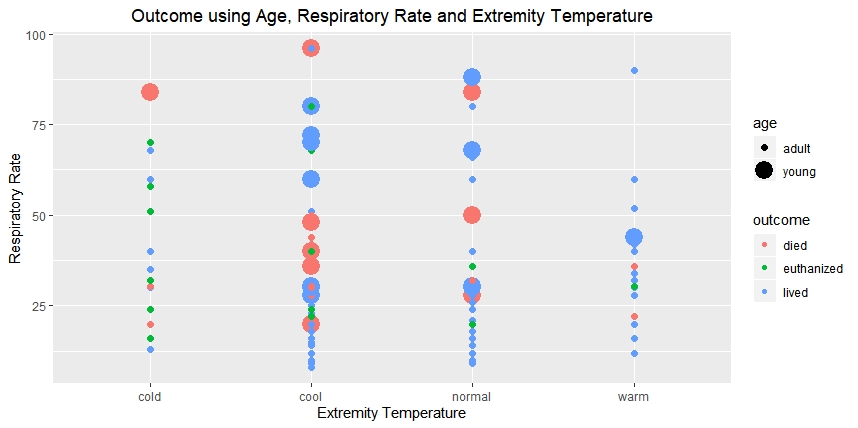


Figure 28: Outcome by Respiratory Rate and Extremity Temperature

Other graphs were run as per the code provided, but not provided in this report.

## Modeling

Several different modeling techniques were used to determine the best method for predicting the mortality outcome for the horse experiencing a colic incident. In this section, the different methods used, some code chunks as well as graphical representation, when possible, is reflected.

### Decision Trees

Several decision trees were run with differing depths and branches to determine the best possible tree. The accuracy for each run was stored in a vector so that a comparison could be made at the end of the runs to determine the best tree. An example of the code used for this can be found in Figure 29.

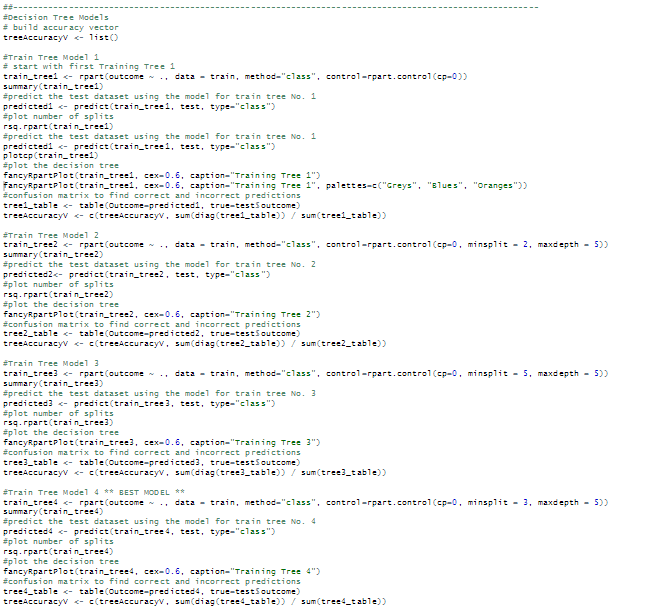


Figure 29: Decision Tree Code

As shown above, different tree splits and depths were done to provide an overall best possible tree. Some of these trees are shown in upcoming figures on the next few pages.

In the first tree, it is clear that the lesion1B or the type of lesion is used as the starting point (root node) of the tree. This is not surprising based on the information that found in the plots represented. Again “strangulation” appears. To understand this tree, the left track shows that if the lesion type is “inflammation,” “other,” or “strangulation”, then the next node represents the total protein being below 35. In this case, 52% will die, 15% will be euthanized and 33% will live. Otherwise, 0% will die, 65% will be euthanized and 35% will live. This continues showing a good depiction of the variables of importance:

* lesion type (lesion1B)
* total protein
* lesion site (lesion1A)
* lesion code (lesion1D)
* pain
* temperature of extremities
* surgery

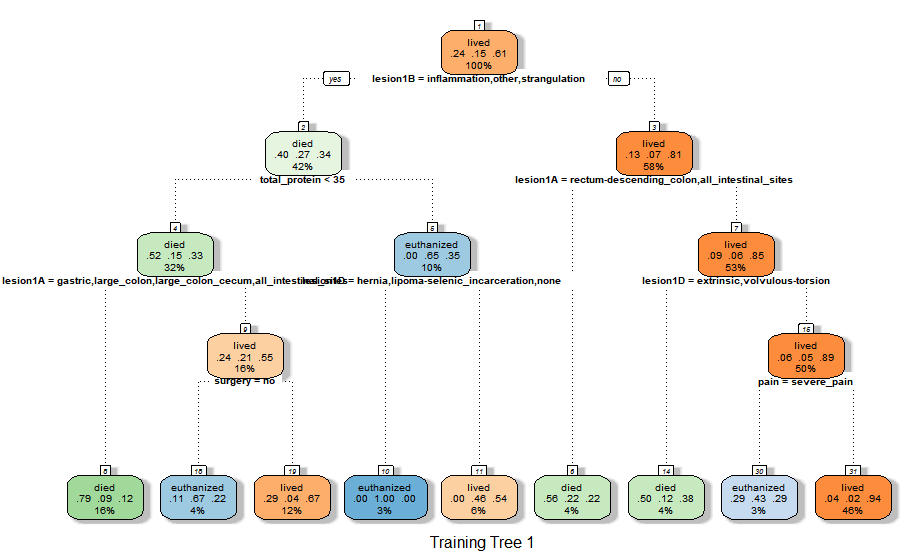


Figure 30: Decision Tree One

For the next tree, the minimum split and maximum depth were specified to be 2 and 5, respectively. This tree is far more complex than the initial tree above as shown Figure 31.

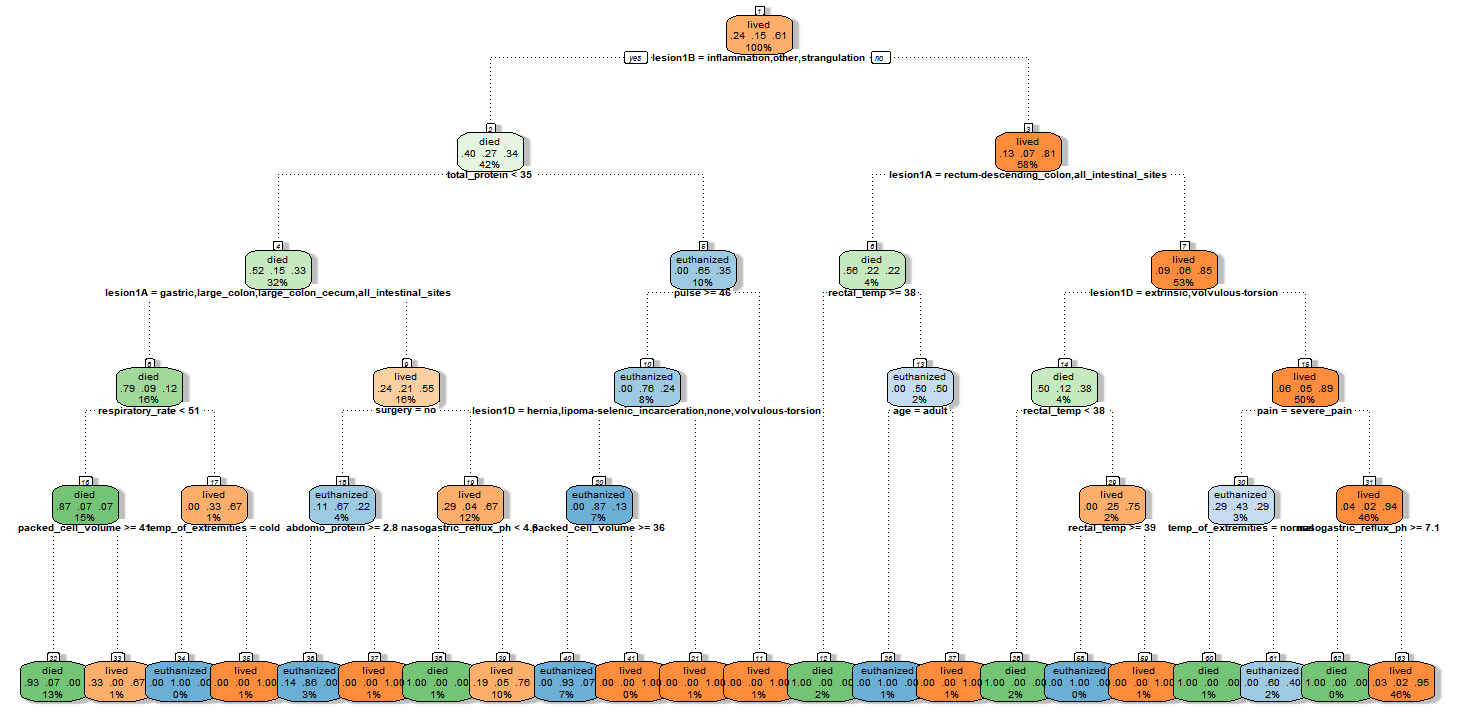


Figure 31: Decision Tree Two

As shown in the figure below, some of the trees provided the same level of accuracy. For this project, tree 5, as mentioned, was selected as the best tree for pruning based on the highest level of accuracy.



Figure 32: Accuracy of Decision Tree Data Frame

More information on the selected tree (the fifth) with accuracy of approximately 68% is reflected in the information below. This tree had a minimum split of 3 and a maximum depth of 4. The variables used in this model are slightly different as shown in the print of the model in the following figure.

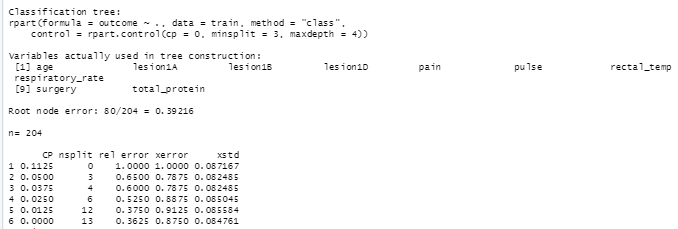


Figure 33: Summary of Classification Tree Five

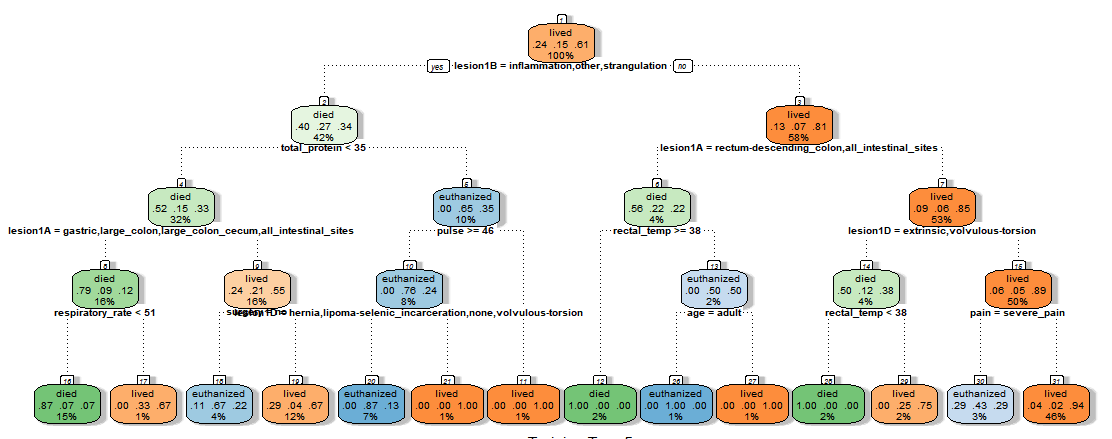


Figure 34: Decision Tree Five

This tree was pruned and the following representation was the result simplifying the tree. The pruned accuracy but the accuracy was decreased to approximately 65%.

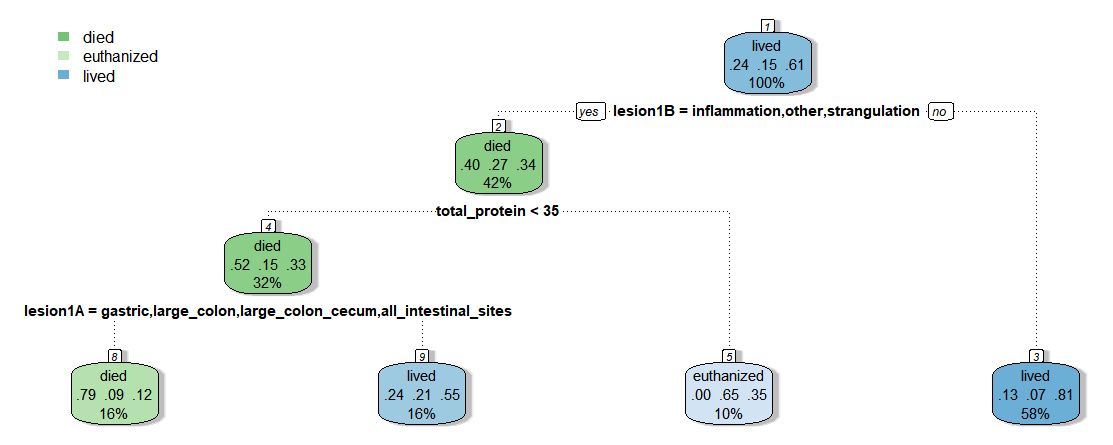


Figure 35: Pruned Tree

The decision trees were repeating using cross validation with 4 folds. The accuracy for certain splits and depths can be found in the table on the following page. Notice how similar the accuracy is on these varying models. Notice that the minimum split of 3 and maximum depth of 5 had the highest level of accuracy 70.89%.

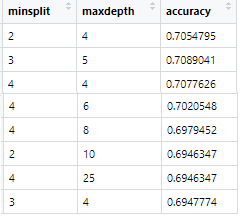


Figure 36: Accuracy of Decision Tree Data Frame

Finally, decision tree modeling with *randomforest* was run. To narrow down the best number of trees in the forest, several runs were done and an accuracy matrix was created as shown in the following figure, Figure 37.

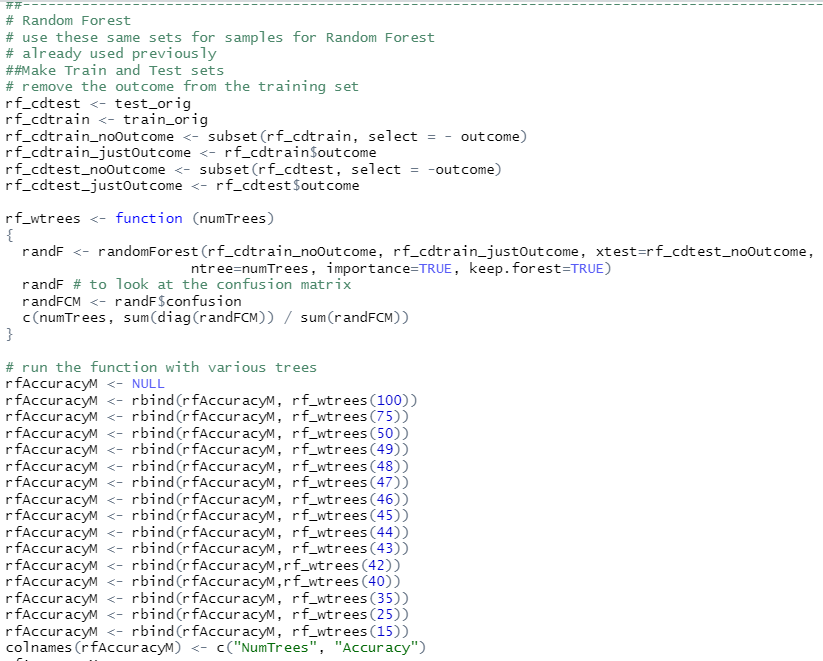


Figure 37: Random Forest R-Code

The best accuracy was at 44 trees with an accuracy of 74.56% as shown in Figure 38.

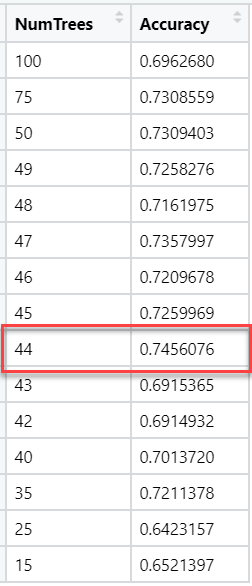


Figure 38: Random Forest Accuracy Matrix

### Association Rule Mining (ARM)

Additional data manipulation was required for ARM as shown in the code chunk in the following figure. This provided bins for various continuous variables in the horse colic data set.

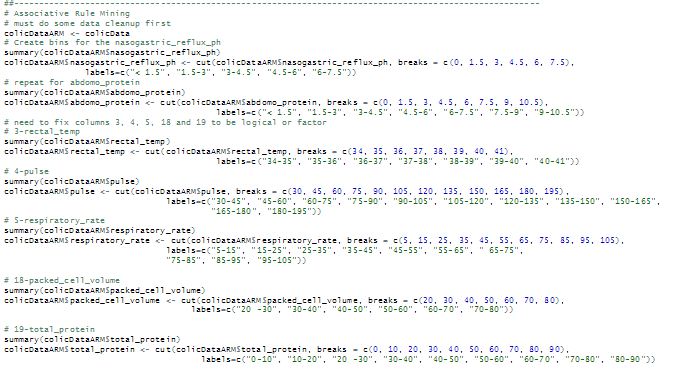
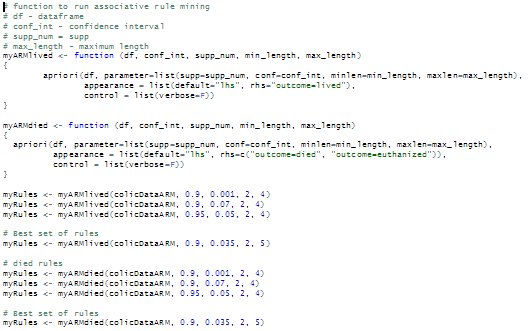


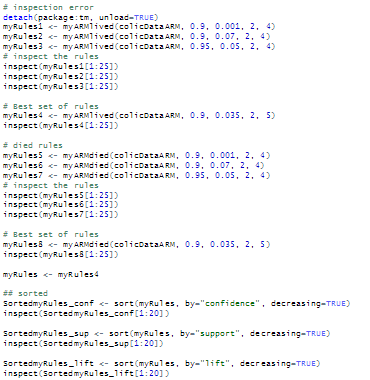
Figure 39: ARM Data Breaks for Continuous Variables

The next step was to run association rule discovery. In performing this analysis, different parameters were used to determine the strongest rules for this particular data set. Approaching the data set in a way to provide rules that will result in a positive outcome – a living horse, it is important to set the rules in in such a way that the “right hand side” of all rules is set to a result of “outcome=lived”. A function was also developed to reflect a negative outcome of died. As a result, the following function was created to return the rules of interest.



*Figure 40: Functions for Association Rules*

Using this method, a simple call could be made to manipulate the parameters for the *apriori* function to bring a different rule set. These results could then be inspected and/or sorted by confidences or support so that a better understanding of the rule could be obtained.

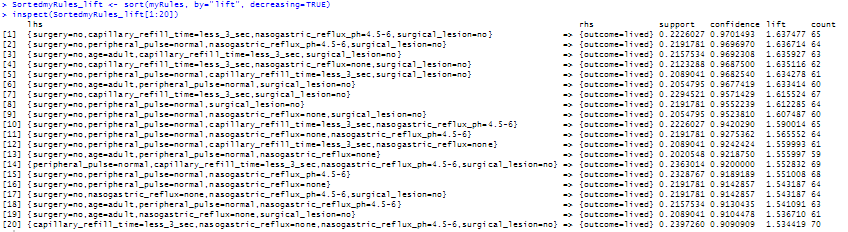


*Figure 41: Calling myARM Function*

Many different combinations of confidence, support, minimum length and maximum length were evaluated. In some cases, selecting a higher minimum with a higher confidence provided less than twenty rules for evaluation. In other cases, raising the minimum confidence and support numbers provided limited to no results.

For a solid set of rules, it was determined that using the following parameters gave strong rules for the combination of characteristics that would result in the life of the colicing horse:

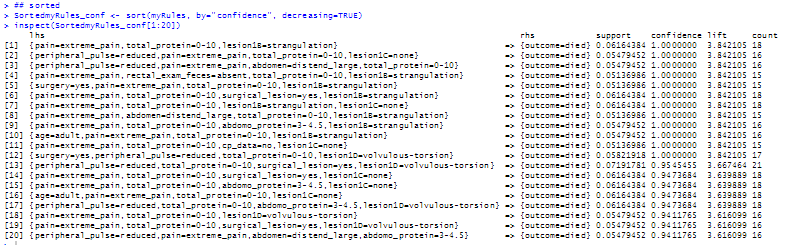
* Confidence minimum of 0.9
* Support minimum of 0.20
* Minimum number for ‘left hand side’ = 2
* Maximum number of ‘left hand side’ = 5



*Figure 42: First Twenty Rules with Support, Confidence and Lift, sorted by Lift*

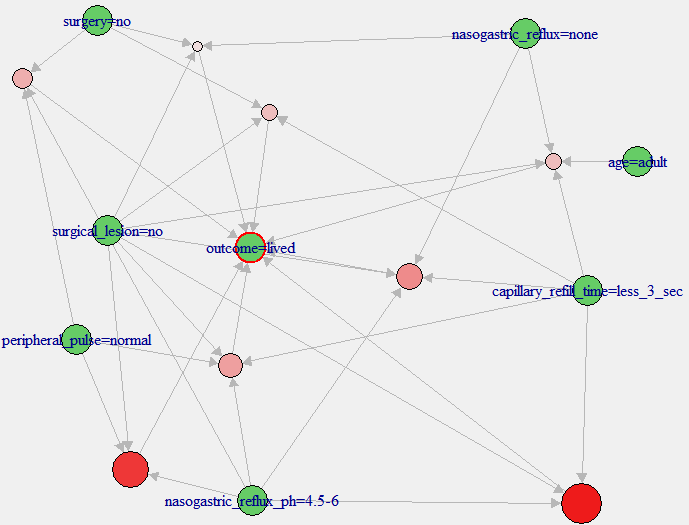
These rules show a greater correlation between the items on the left-hand side. The highest boasting a lift ratio of 1.637 and a high confidence of 97%. Using a higher support figure of at least .20 ruled out some less strong associations and provided good support at 0.22 for the best rules. This rule association indicates that not having surgery, showing a quick capillary refill time and normal reflux PH would indicate the horse would live.

Conversely, approaching the rule from the outcome being the death of the animal, a much lower support needed to be used. The lift ratio is significantly higher of 3.84 as shown in the following figure.

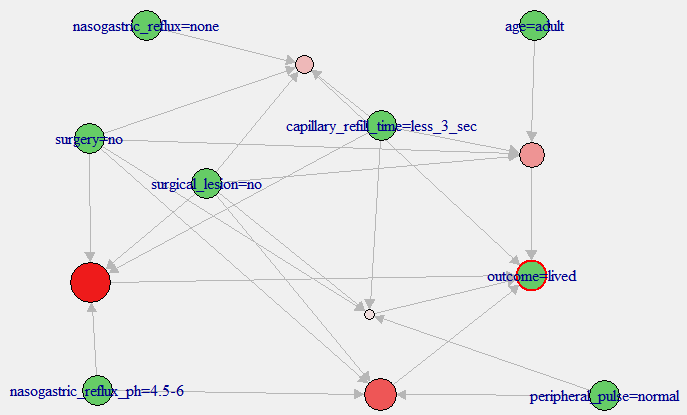


*Figure 43: First Twenty Rules with Support, Confidence and Lift, sorted by Confidence*

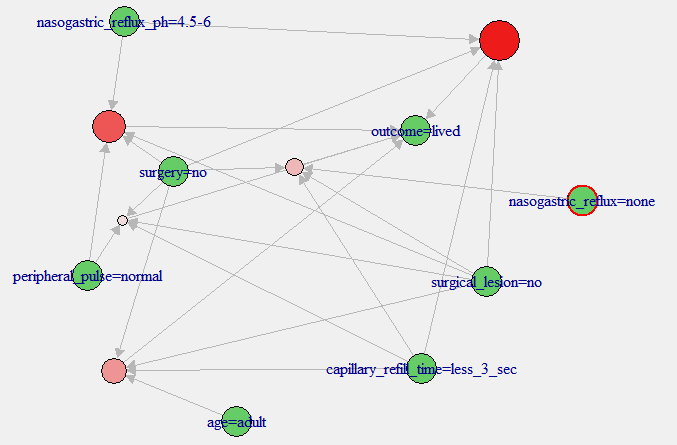
Some interesting plots were run on the rules to show some visualization of what was taking place by plotting interactive maps of the top eight (8) rules sorted by decreasing support, confidence and then lift.



*Figure 44: Interactive Plot of Top Eight Rules by Decreasing Support*



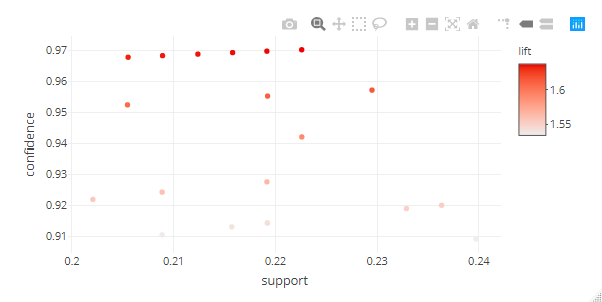
*Figure 45: Interactive Plot of Top Eight Rules Sorted by Decreasing Confidence*



*Figure 46: Interactive Plot of Top Eight Rules Sorted by Decreasing Confidence*

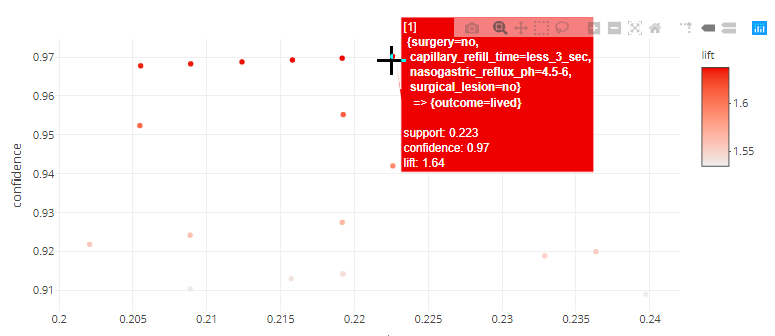
The similarity or reoccurrence of certain variables are very obvious even when sorted differently.

Visualizations were then done using a deprecated function that still provided the plot, *plotly\_arules* which shows a very interesting scatterplot of the rules. Each figure is color coded by lift getting lighter when the lift is smaller.



*Figure 47: Scatterplot of Top Twenty Rules by Decreasing Confidence*

Not only does this show an interesting grouping of the rules, but it is interactive as well. For example, by hovering over a point on the plot, the rules are displayed as shown in the following figure.



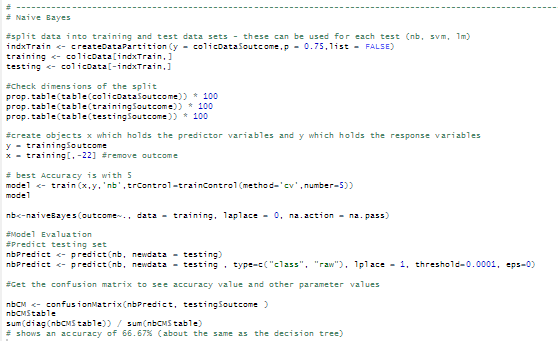
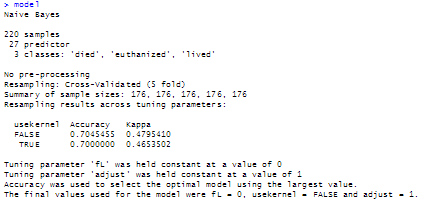
*Figure 48: Rule 1 with Support=0.223, Confidence=0.97 and Lift=1.64*

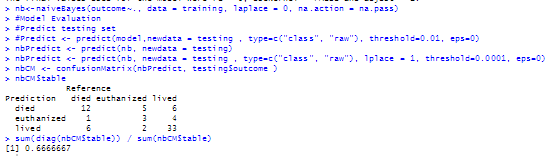
The information reflected in Figure 48 is not surprising. Essentially, if the horse does not go for surgery, has a quick capillary refill time (means blood is flowing properly), a normal range of PH in the reflux obtained – then the horse will likely live.

### Naïve Bayes

The next model selected was the Naïve Bayes Classifier. “Naïve Bayes classifier calculates the probabilities for every factor. Then it selects the outcome with the highest probability.”[[1]](#footnote-1)

Initially, the Naïve Bayes model was generated with a simple training and test set as shown in the code in Figure 49.



*Figure 49: Naïve Bayes with “train” and “naivebayes” functions*

Both the *train* using the “nb” classifier and the *naivebayes* function were both executed. The first providing an accuracy of 70% for both true and false classifications. Using the *naivebayes* function, this accuracy decreased to 66.67%. Both similar to that found using decision trees.

Then a similar approach was taken to run the Naïve Bayes method for k-fold validation by splitting the training data into 4 sets of random samples. Then the accuracy was determined from the final confusion matrix for this model.

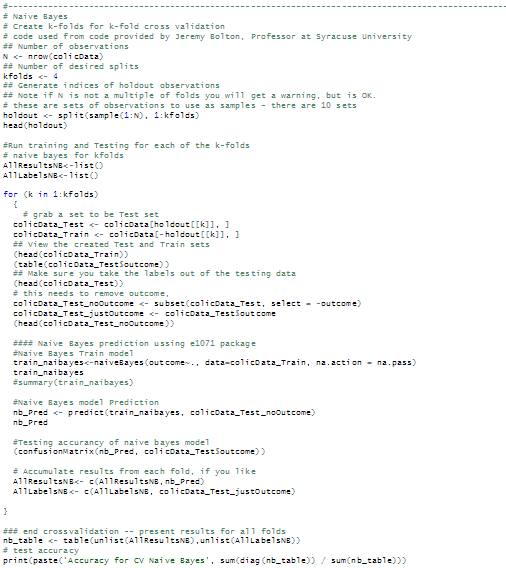
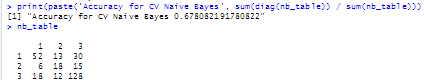
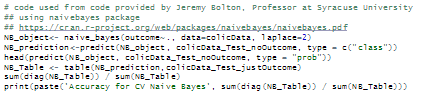
  


Figure 50: Naïve Bayes R-Code for 4-fold Validation and Accuracy

Using this cross-validation technique, the accuracy improved slight to 67.81% as shown in the following figure.

Following this attempt, some research was done on other ways to run Naïve Bayes that may provide a greater level of accuracy. Using *naïve\_bayes*, the result showed an accuracy of 65.75% which is less than the previous attempt with Naïve Bayes.



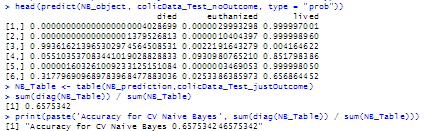


Figure 51: Naïve Bayes using naïve\_bayes function with Accuracy

Using the *mlr* package, another run using Naïve Bayes was run as shown using the code below.

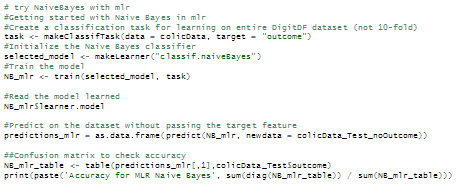
  


Figure 52: Naïve Bayes mlr R-Code

Using *mlr* for Naïve Bayes proved slight better than the *naïve\_bayes* function with an accuracy of about 68.5%.

### Support Vector Machine (SVM)

The next model used in modeling the data was the Support Vector Machine (SVM) which is a “discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.”[[2]](#footnote-2)

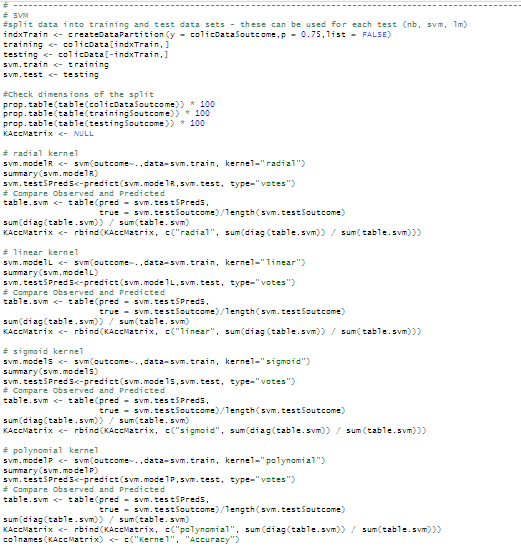


Figure 53: SVM R-Code for Each Kernel

Each possible kernel was testing for SVM with the “linear” kernel having the highest accuracy of 75%. This was an increase over all other methods attempted so far. In this case, the cost remained 0 and no other variables were modified to simply the modeling.

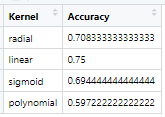


Figure 54: SVM Kernel Accuracy Matrix

### Multiple Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. It examines two things: do the set of independent variables do a good job predicting the dependent variable and which variables are significant in making this prediction. Multiple linear regression (MLR), also known simply as multiple regression, is a technique that uses more than one independent or descriptive variable to predict the response (dependent) variable.

In order to minimize the different models to be run, stepping through the models to find the best combination of independent variables was used. This provides the Akaike’s Information Criteria for the best model selection. This helps to select the most frugal model (or parsimonious model), essentially providing a measure of parsimony.

In order to run regression in R, the outcome (or dependent variable) had to be converted to numeric values. For this, the code below was used.

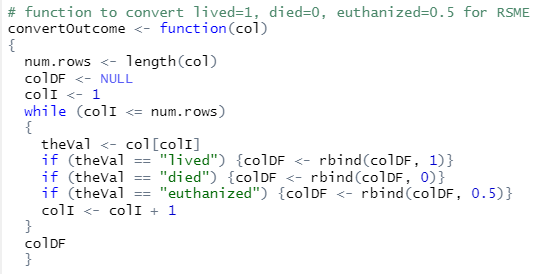


Figure 55: R-Code to Convert Outcome to Numeric

Once this was accomplished, a model was run with all variables as an initial baseline.

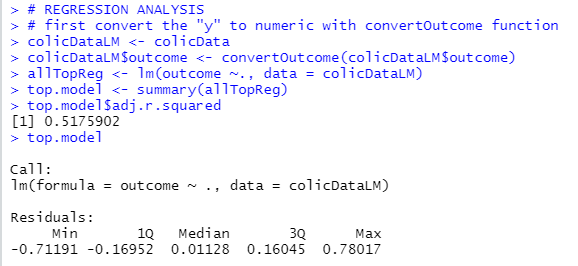


Figure 56: Regression on all Variables

Reviewing the detailed coefficients for each variable as shown in Figure 57, there are certain variables that have very low p-values and have a bigger impact on the model. These are noted by asterisks next to each variable.

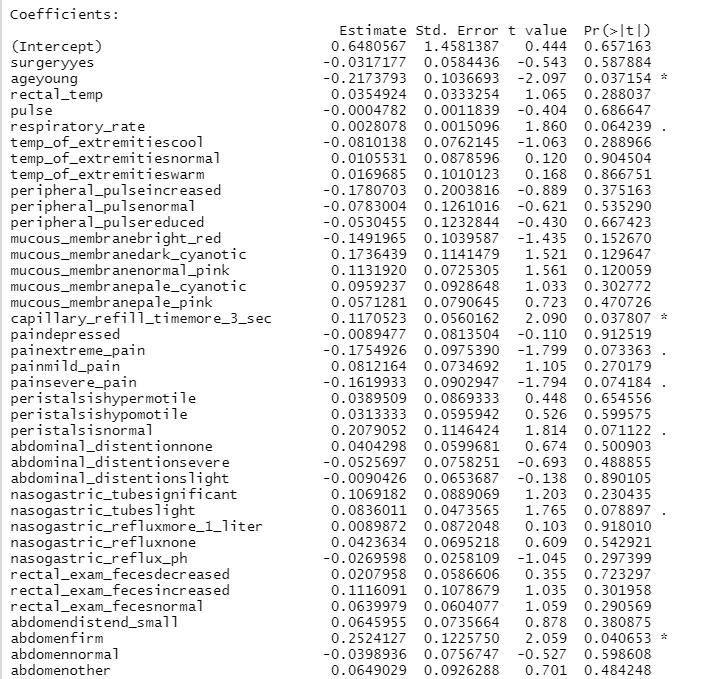
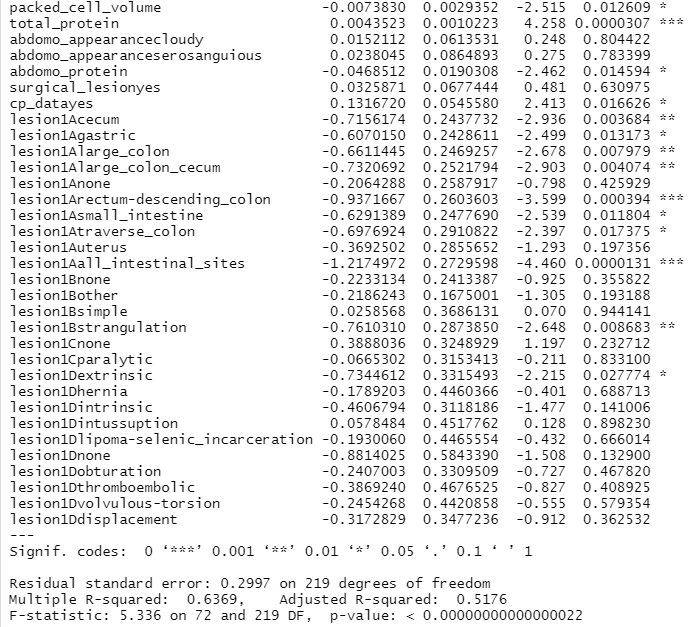
  


Figure 57: Regression Coefficients and Correlation Coefficient

These individual coefficients were extracted to form the list provided in the following figure.

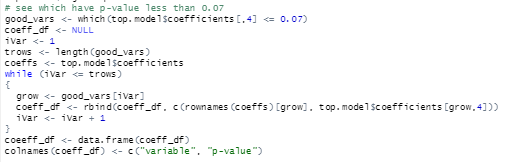
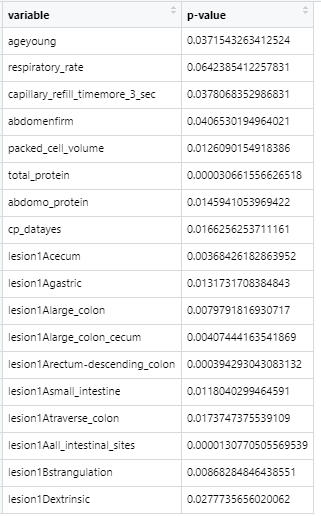
  


Figure 58: Variables with < .07 p-values

This table reflects that only certain values for factor variables like lesion1A, lesion1B and lesion1D are significant, but some have very strong significance such as if the lesion is found in in “all intestine sights” or in the “rectum descending colon.” It is clear that there are several continuous variables of significance and these are the same that were seen in the decision trees and association rules. However, the adjusted R-squared was not very high at .5176. A number closer to 1 is preferable.

Stepping through the model to find that with the smallest AIC provided a slightly better adjusted R-squared of .5356 as shown in the following figure. This model used only the following variables:

* Age
* Respiratory rate
* Temperature of extremities
* Mucous membrane color
* Capillary refill time
* Pain
* Packed cell volume
* Total protein
* Abdomen protein
* Lesion site
* Lesion type
* Lesion code
* Pathology data presence

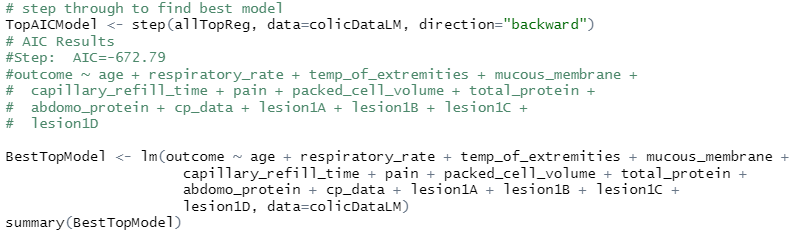
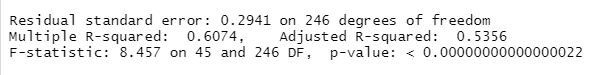
  


Figure 59: Correlation Coefficient for Best Model

This model was followed with predictions to determine the Root Mean Square Error (RSME) which is the standard deviation of the residuals or how far the residuals are spread out. This shows how concentrated the data is around what is the line of best fit.

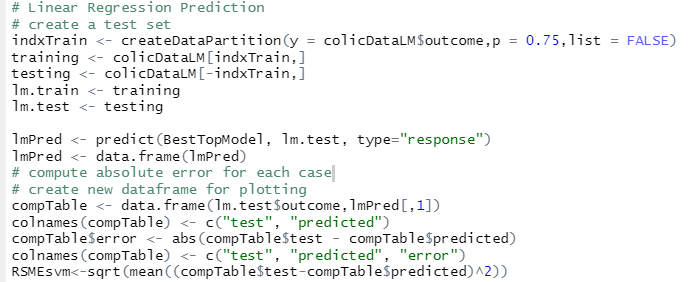


Figure 60: R-Code for Regression Prediction and RSME

Using this method, the RSME for the best model was .2754505. Plotting was done for this model showing the error associated with each prediction of live, euthanized or death.

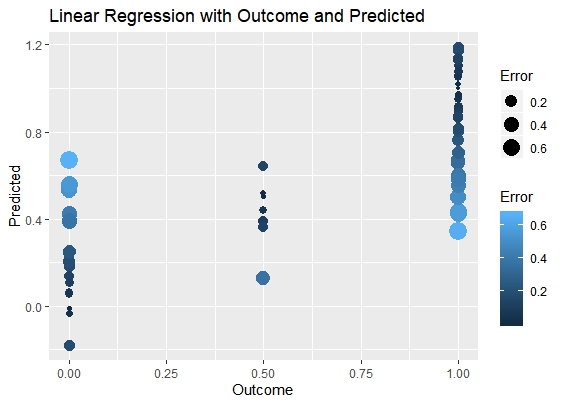


Figure 61: Error by Predicted Outcome

Since some variables were very significant in determining the outcome, some plots were made to help visualize this significance.

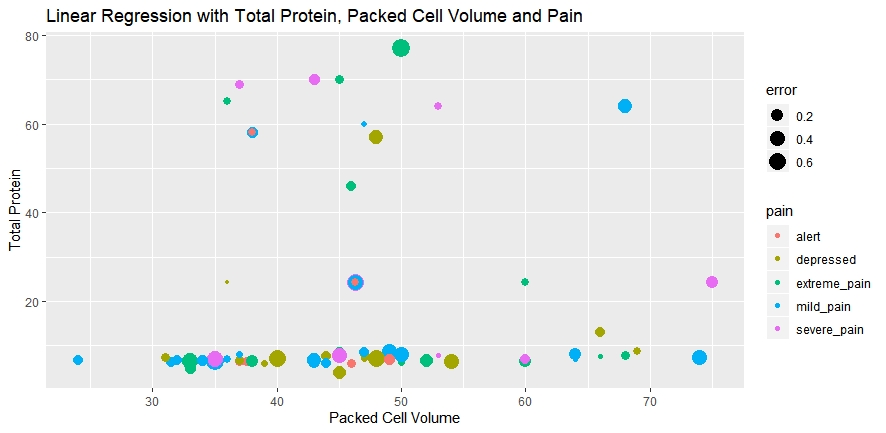


Figure 62: Error by Total Protein, Pain and Packed Cell Volume

In this case, certain prediction errors were large depending on the packed cell volume and the pain level. It was interesting to see that when the total protein was low, the errors were much smaller regardless of the pain level.

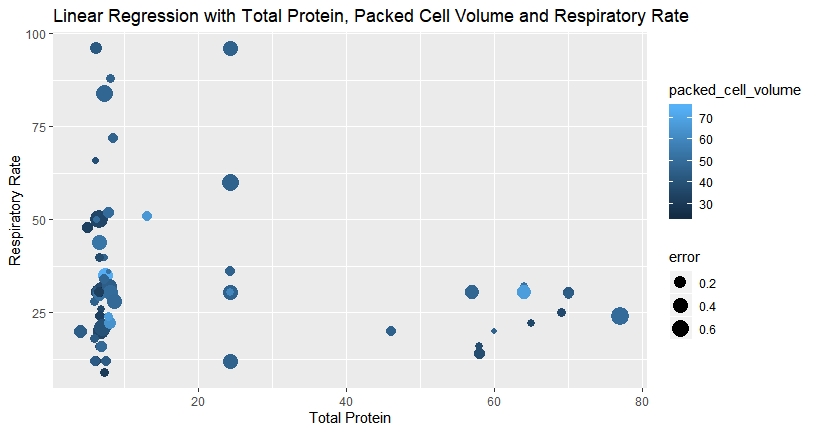


Figure 63: Error by Total Protein, Packed Cell Volume and Respiratory Rate

As before, the errors are smaller when the total protein value is also small or high with a few outliers.

### Logit and Probit

As a final modeling technique, both logit and probit models were run. In this case, the .

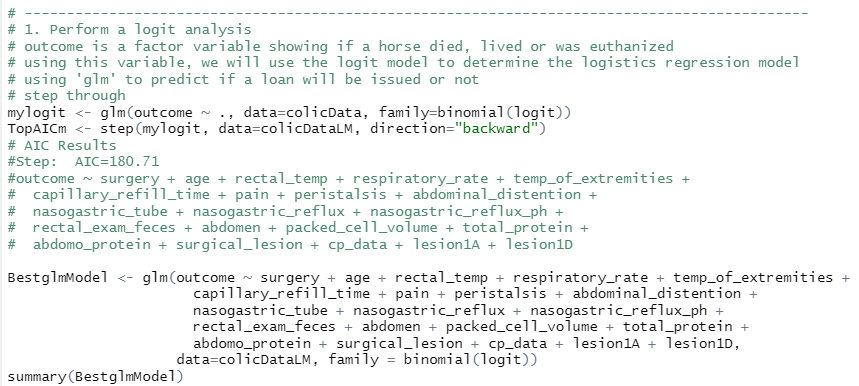


Figure 64: Finding best logit Model for Colic Outcome Prediction

After reviewing the variables that had very large p-values and running the model again without these values – the AIC was not lowered, but raised to 190.32. However, the RSME was also captured for each of the models run in logit and probit based on the outcome of the *step* through the variables to find the best model. That is reflected in the following figure.

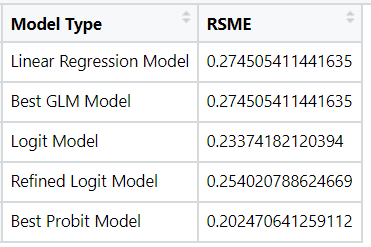


Figure 65: RSME by Model Type

As noted, the best model based on RSME was the probit model which is shown below.

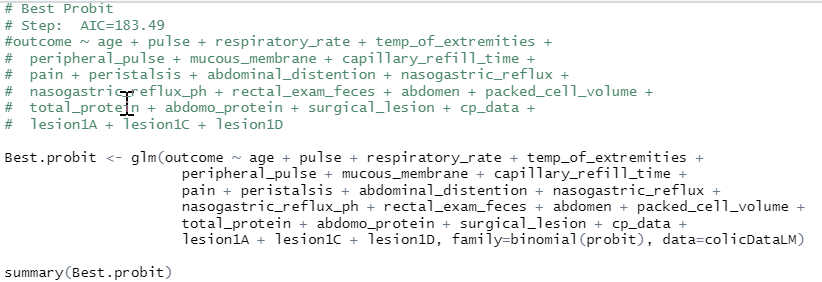


Figure 66: Best Probit Model for Colic Prediction

### Neural Network

Neural networks are use a set of algorithms that are modeled, loosely, after the human brain. These algorithms recognize patterns in data and them build a “path” to an outcome.

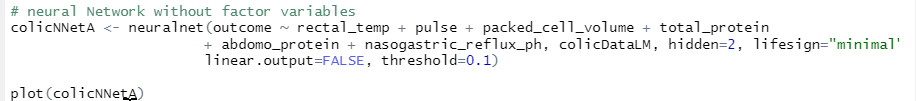


Figure 67: Neural Network – Used Only Continuous Data Variables

This neural network resulted in two hidden nodes, 75 steps and an error of 26.26 as shown in the following plot.

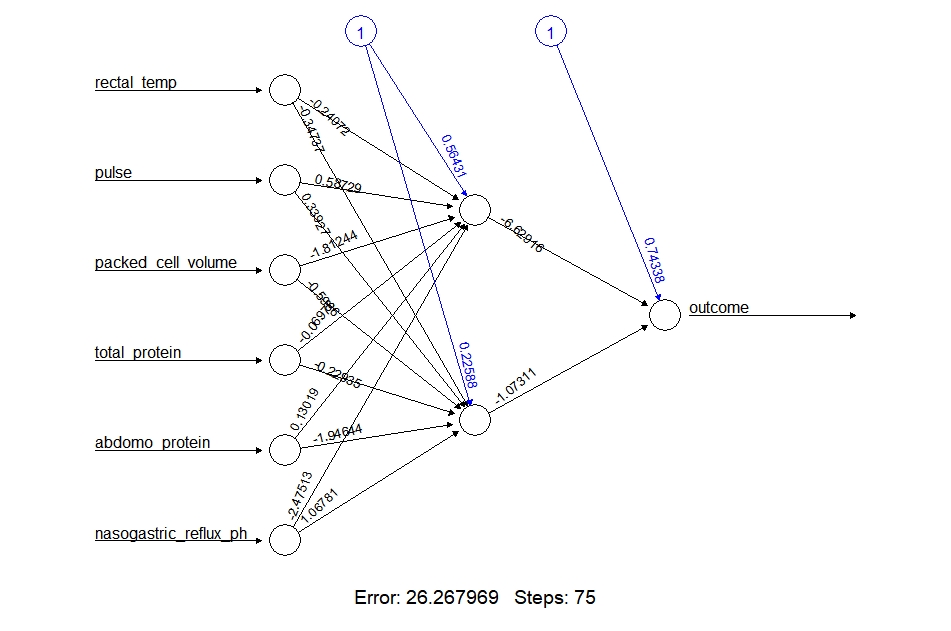


Figure 68: Plotted Neural Network

The details for each path can be found by reviewing the matrix of information provided by the function as shown in the following figure.

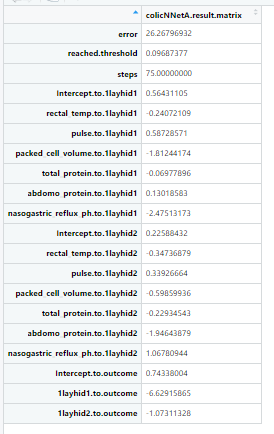


Figure 69: Neural Network Detailed Matrix

# Results

As shown in the previous section, a significant number of different models were run to determine the best for prediction of colic outcome when a horse presents specific symptoms. In general, the data set provided proved that is it not easy to predict the colic outcome based on these variables.

However, many models provided an accuracy between 60% and 75% which means there is some strength in the data. The linear kernel for the support vector machines proved to be the strongest based on this level of measure of accuracy at 75%.

It was clear that certain variables were of significance as they appeared in almost every model that was used on the data set.

* Lesion location
* Total Protein
* Respiratory Rate
* Capillary Refill Time
* Packed Cell Volume
* Abdomen Protein
* Lesion Type

Moving to linear regression proved useful, but the results are difficult to gage because the RSME is likely quite small because the outcome integers were 0, 0.5 and 1 which are also quite small. The adjusted R-squared for the multiple linear regression models was not very strong at .5356.

# Conclusions

As a result of the analysis done on this data, I would make the following conclusions and recommendations.

* Missing Data

There was a significant amount of missing data in this set. In fact, 1,602 data points out of a total of 3,872. This means that over 19% of the data was missing. In this case, means and modes were used to fill in missing data, but this may not have been the best recourse. Removing all the horses from the dataset which have left a far smaller dataset, but something to consider for further modeling.

* Old and Out of Date Data

This data is 30 years old. There is likely more up to date and more complete data sets; however, none could be located for this project.

* Additional Information

So much more can be captured now and that means that useful new data is not part of this original data set. For example, research has shown that monitoring the lactate levels in the peritoneal fluid can help predict if a horse that is colicing will require surgery. This data is recent in the last decade and would have proved very helpful in the dataset.

In addition, other information like finer granularity around age would be more appropriate. A six-month-old horse is not an adult; however, in this data that is how it was categorized. Horses continue to develop well into age 6 or 7, so a more definitive grouping of foal, yearling, young adult, adult, and geriatric would have been more helpful.

It is possible the horse’s sex could have proved useful. Mares in reproductive years may or may not experience colic more or less. They tend to respond more to stressful situations.

That leads to even more data that could be useful. For example, if the horse had previously coliced and had a previous colic surgery. These horses tend to be more prone to a second issue.

There are also medications, like doxycycline, tetracycline, and dexamethasone that can cause stomach upset inducing colic in horses. This information was not captured.

* Lack of Common Comparison Algorithm

Using different methods with very different determinations of effectiveness proved difficult. Without a common way to compare the accuracy or strength of model, it was difficult to make assumptions about which model would be best for predictions. Models had to be compared in “groupings” or like-models, so final “best model” was difficult to determine.

# 

# References

This dataset was originally published by the UCI Machine Learning Database: <http://archive.ics.uci.edu/ml/datasets/Horse+Colic>

* Title: Horse Colic database
* Source Information:
  + Creators: Mary McLeish & Matt Cecile  
    Department of Computer Science  
    University of Guelph  
    Guelph, Ontario Canada N1G 2W1
  + Donor: Will Taylor (taylor@pluto.arc.nasa.gov)  
    Date: 8/6/1989

Some additional R packages were used for this project.

* Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

1. Patel, Savan. Chapter 1: Supervised Learning and Naïve Bayes Classification – Part 1 (Theory). https://medium.com/machine-learning-101/chapter-1-supervised-learning-and-naive-bayes-classification-part-1-theory-8b9e361897d5 [↑](#footnote-ref-1)
2. Patel, Savan. Chapter 2: SVM (Support Vector Machine) – Theory. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [↑](#footnote-ref-2)