Joyce Woznica

jlwoznic@syr.edu

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

Joyce Woznica  
Horse Colic Project

Contents

[Introduction 4](#_Toc26390731)

[Types of Colic in Horses 4](#_Toc26390732)

[Goals/Questions 5](#_Toc26390733)

[Analysis 5](#_Toc26390734)

[The Data 5](#_Toc26390735)

[Data Load 6](#_Toc26390736)

[Data Cleanse 7](#_Toc26390737)

[Exploratory Data Analysis 11](#_Toc26390738)

[Descriptive Statistics 11](#_Toc26390739)

[Visualization 12](#_Toc26390740)

[General Visualization 12](#_Toc26390741)

[Detailed Visualizations 17](#_Toc26390742)

[Modeling 20](#_Toc26390743)

[Clustering 20](#_Toc26390744)

[Decision Trees 21](#_Toc26390745)

[Associative Rule Mining (ARM) 26](#_Toc26390746)

[Visualization 29](#_Toc26390747)

[Linear Regression (LM) 32](#_Toc26390748)

[Support Vector Machine (SVM) 38](#_Toc26390749)

[Kernel Support Vector Machine (KSVM) 40](#_Toc26390750)

[Naïve Bayes 40](#_Toc26390751)

[Results 41](#_Toc26390752)

[Conclusions 42](#_Toc26390753)

[References 44](#_Toc26390754)

List of Figures

[Figure 1: Table of Horse Colic Variables 6](#_Toc26390755)

[Figure 2: Load and Initial Structure of Horse Colic Data 7](#_Toc26390756)

[Figure 3: Missing Data Plot from the Horse Colic Data 7](#_Toc26390757)

[Figure 4: Initial Summary of Colic Data 8](#_Toc26390758)

[Figure 5: Function to Replace Missing Numeric Data with Means 8](#_Toc26390759)

[Figure 6: Function to Replace Missing Factor Data with Modes 9](#_Toc26390760)

[Figure 7: Manual Updates of Modes for Special Cases 9](#_Toc26390761)

[Figure 8: Removal of Lesion 2 and Lesion 3 Data and Irrelevant Data 9](#_Toc26390762)

[Figure 9: Updated Data Frame with Lesion 1 Information 10](#_Toc26390763)

[Figure 10: Updated Data Frame by Removing incorrect Lesion 1 Rows 11](#_Toc26390764)

[Figure 11: Update Capillary Refill Time Rows 11](#_Toc26390765)

[Figure 12: Mix Up the Order of the Rows 11](#_Toc26390766)

[Figure 13: Summary Information from the Satisfaction Survey 11](#_Toc26390767)

[Figure 14: Describing the Horse Colic Data 12](#_Toc26390768)

[Figure 15: Summary Statistics of Some Horse Colic Continuous Variables 12](#_Toc26390769)

[Figure 16: Colic Outcomes 13](#_Toc26390770)

[Figure 17: Colic Variable Histograms 13](#_Toc26390771)

[Figure 18: Additional Bar Plots and Outcome Frequencies 14](#_Toc26390772)

[Figure 19: Lesion Site related to Colic Outcome 15](#_Toc26390773)

[Figure 20: Lesion Type related to Colic Outcome 16](#_Toc26390774)

[Figure 21: Lesion SubType related to Colic Outcome 16](#_Toc26390775)

[Figure 22: Lesion Code related to Colic Outcome 17](#_Toc26390776)

[Figure 23: Feces to Colic Outcome 17](#_Toc26390777)

[Figure 24: Box Plot of Outcome by Pulse Rate 18](#_Toc26390778)

[Figure 25: Box Plot of Total Protein and Outcome 18](#_Toc26390779)

[Figure 26: Packed Cell Volume by Outcome 19](#_Toc26390780)

[Figure 27: Packed Cell Volume by Outcome 19](#_Toc26390781)

[Figure 28: Outcome by Respiratory Rate and Extremity Temperature 20](#_Toc26390782)

[Figure 29: Clustering by Manhattan Distance 21](#_Toc26390783)

[Figure 30: Clustering by Euclidean Distance 21](#_Toc26390784)

[Figure 31: Decision Tree Code 22](#_Toc26390785)

[Figure 32: Decision Tree One 23](#_Toc26390786)

[Figure 33: Decision Tree Two 24](#_Toc26390787)

[Figure 34: Accuracy of Decision Tree Data Frame 24](#_Toc26390788)

[Figure 35: Summary of Classification Tree Five 25](#_Toc26390789)

[Figure 36: Decision Tree Five 25](#_Toc26390790)

[Figure 37: Pruned Tree 26](#_Toc26390791)

[Figure 38: Accuracy of Decision Tree Data Frame 26](#_Toc26390792)

[Figure 39: ARM Data Breaks for Continuous Variables 27](#_Toc26390793)

[Figure 40: Functions to run Association Rules 27](#_Toc26390794)

[Figure 41: Calling myARM Function 28](#_Toc26390795)

[Figure 10: Ruling down the Rules 28](#_Toc26390796)

[Figure 11: First Twenty Rules obtained with Support, Confidence and Life 28](#_Toc26390797)

[Figure 12: Interactive Plot of Top Ten Rules by Decreasing Support 29](#_Toc26390798)

[Figure 13: Interactive Plot of Top Ten Rules Sorted by Decreasing Confidence 30](#_Toc26390799)

[Figure 14: Scatterplot of Top Twenty Rules by Decreasing Confidence 31](#_Toc26390800)

[Figure 15: Rule 13 with Support=0.075, Confidence=0.938 and Lift=2.05 31](#_Toc26390801)

[Figure 23: Satisfaction based on Total Flight Minutes Linear Regression Model 32](#_Toc26390802)

[Figure 24: Satisfaction Model based on Status, Total Number of Flights, Price Sensitivity 33](#_Toc26390803)

[Figure 25: AIC Best Linear Regression Model for Top Three Airlines and All Airlines 33](#_Toc26390804)

[Figure 26: Summary of the Best Linear Regression Model for Top Three and All Airlines 34](#_Toc26390805)

[Figure 27: Summary of the Final and Best Linear Regression Model 35](#_Toc26390806)

[Figure 28: Best LM showing Price Sensitivity and Status with Prediction Error 35](#_Toc26390807)

[Figure 29: Best LM showing Price Sensitivity, Travel Type and Status with Prediction Error 36](#_Toc26390808)

[Figure 30: Best LM showing Departure and Arrival Delays with Prediction Error 37](#_Toc26390809)

[Figure 31: Zoomed Best LM showing Departure and Arrival Delays with Prediction Error 37](#_Toc26390810)

[Figure 32: Zoomed Again Best LM showing Departure and Arrival Delays with Prediction Error 38](#_Toc26390811)

[Figure 33: SVM showing Price Sensitivity and Status with Prediction Error 39](#_Toc26390812)

[Figure 34: SVM showing Price Sensitivity, Travel Type and Status with Prediction Error 39](#_Toc26390813)

[Figure 35: Zoomed SVM showing Departure and Arrival Delays with Prediction Error 40](#_Toc26390814)

[Figure 36: Root Mean Squared Error by Model Type 41](#_Toc26390815)

[Figure 37: Final Variables Determined in Modeling 42](#_Toc26390816)

# 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. I have had one horse that suffered from colic 2 times that required surgery. She survived the first and lived 2 more years. However, she was euthanized the second time she was on the table. I have also had other horses that have had various degrees of colic that were able to recover with pain medication, diet change and rest.

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.

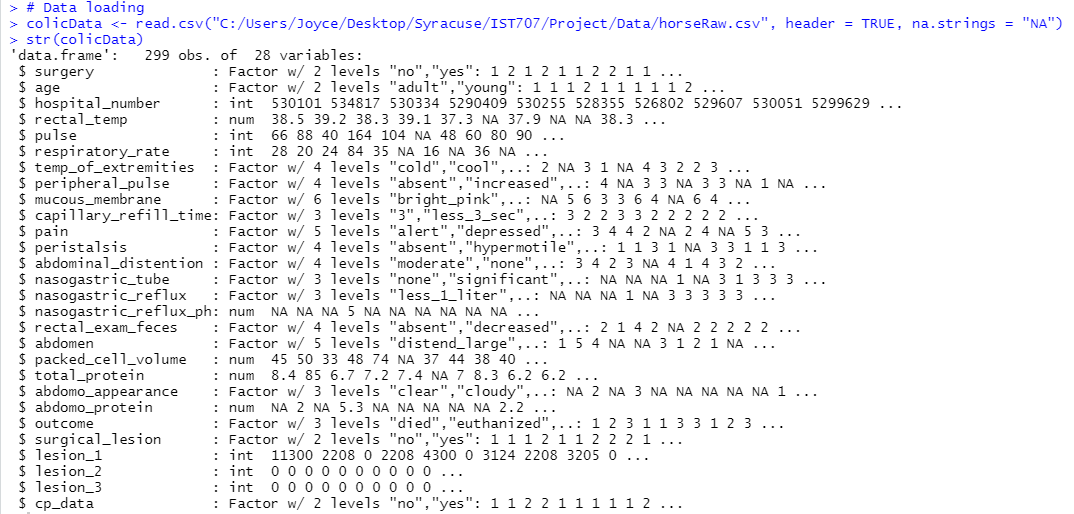


Figure 2: Load and Initial Structure of Horse Colic Data

### 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.

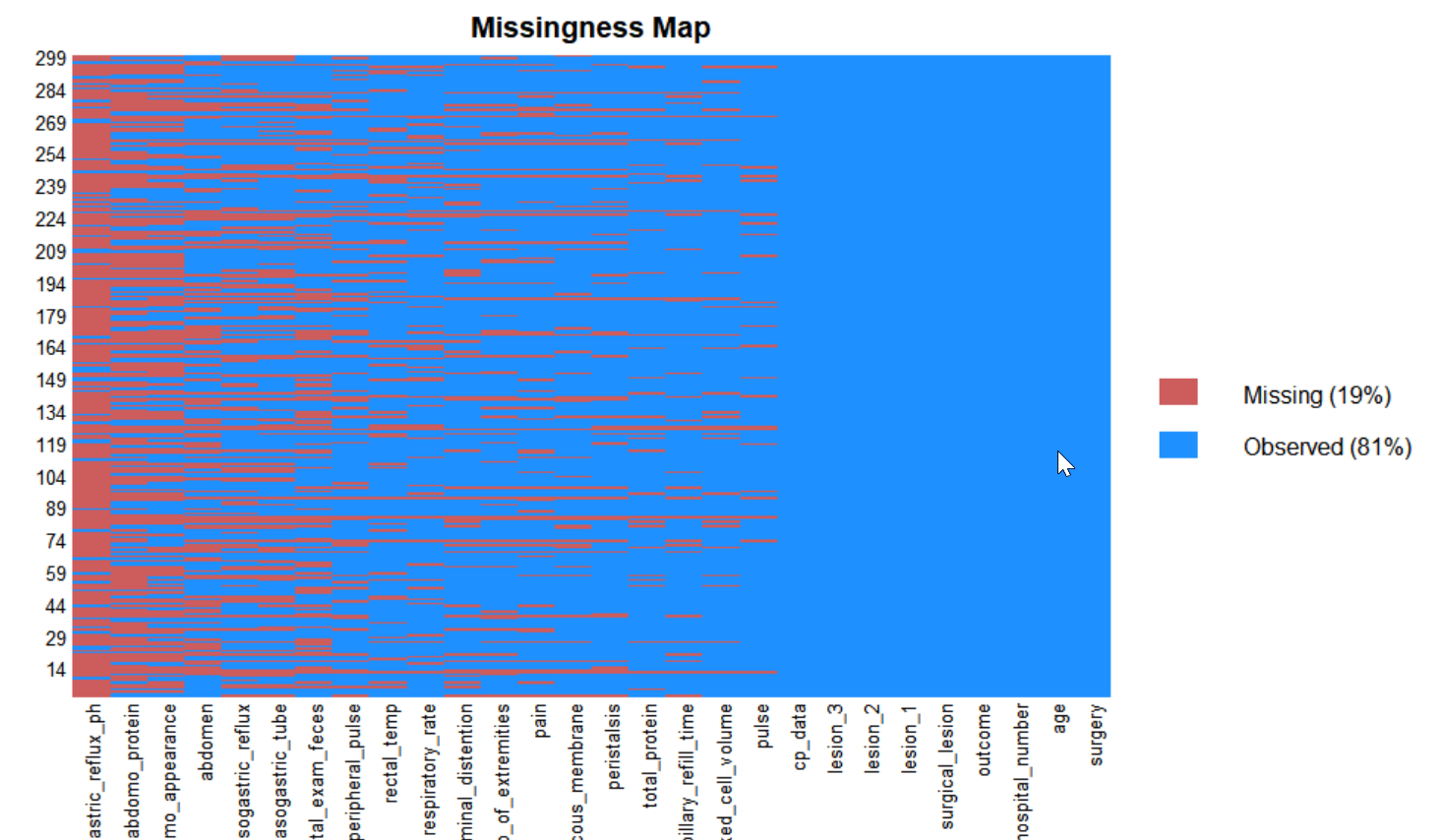


Figure 3: Missing Data Plot from the Horse Colic Data

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

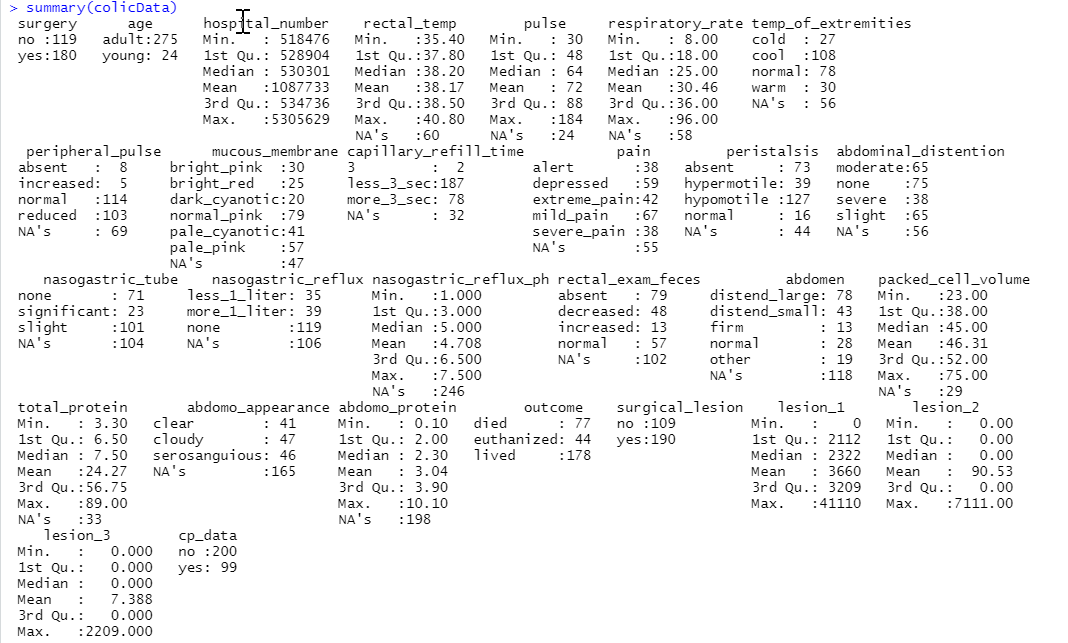


Figure 4: Initial Summary of Colic Data

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

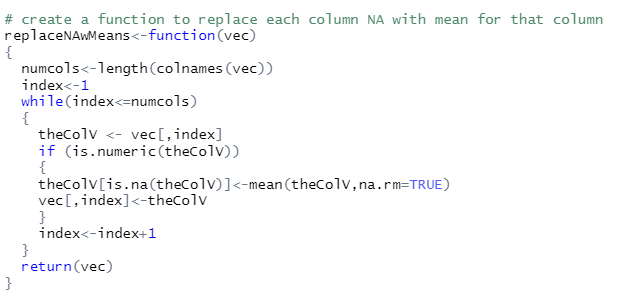
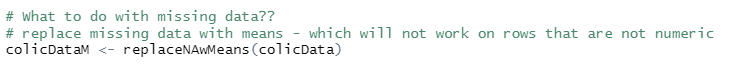
  


Figure 5: Function to Replace Missing Numeric Data with Means

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.

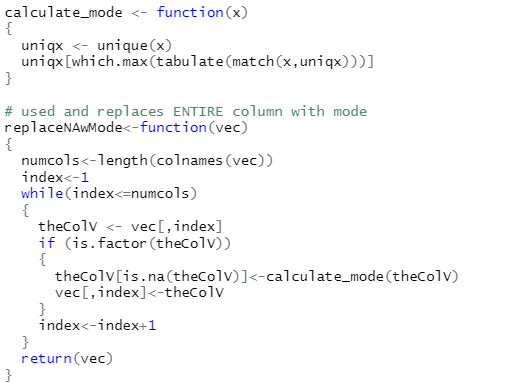
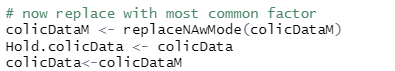
  


Figure 6: Function to Replace Missing Factor Data with Modes

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.

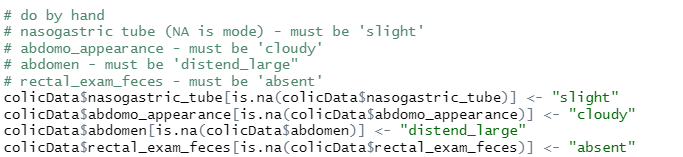


Figure 7: Manual Updates of Modes for Special Cases

These actions eliminated all missing data noted by “NA.” However, as shown in Figure 1 shown on page 5, 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.

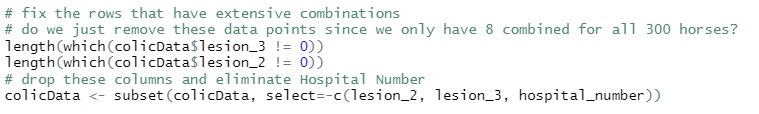


Figure 8: Removal of Lesion 2 and Lesion 3 Data and Irrelevant Data

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.

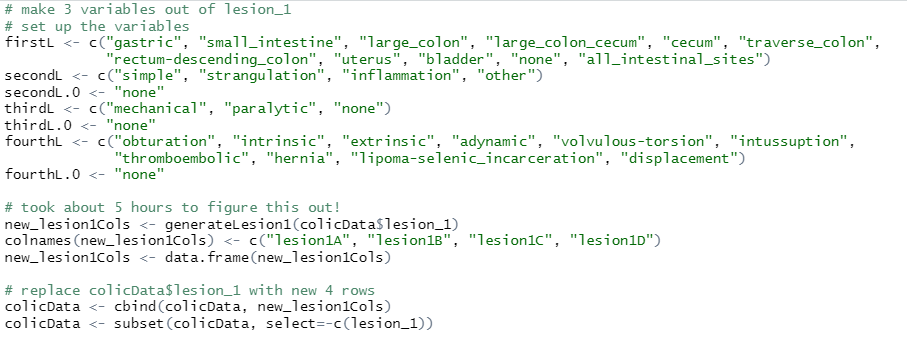
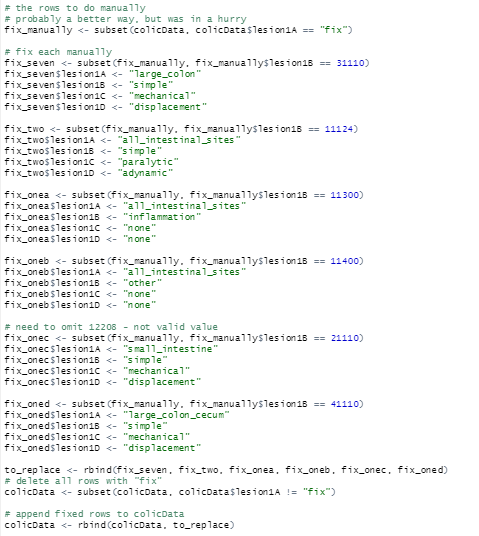
   


Figure 9: Updated Data Frame with Lesion 1 Information

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.

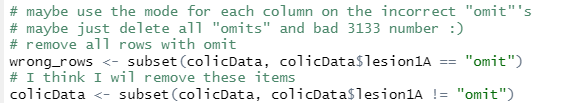


Figure 10: Updated Data Frame by Removing incorrect Lesion 1 Rows

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.

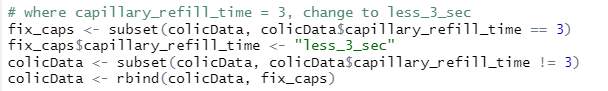


Figure 11: Update Capillary Refill Time Rows

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.

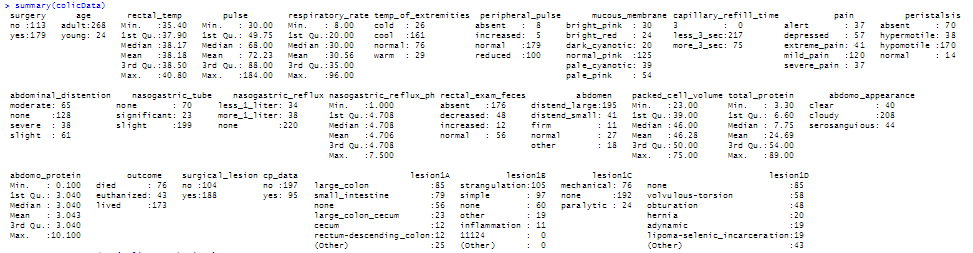


Figure 13: Summary Information from the Satisfaction Survey

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

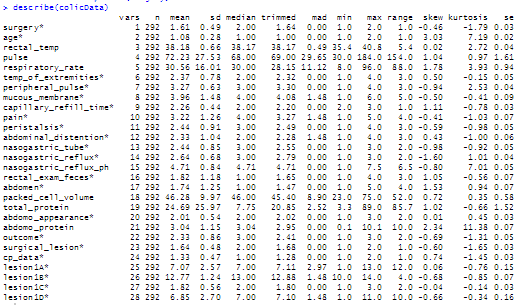


Figure 14: Describing the Horse Colic Data

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.

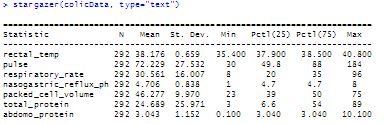


Figure 15: Summary Statistics of Some Horse Colic Continuous Variables

## 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 following figure.

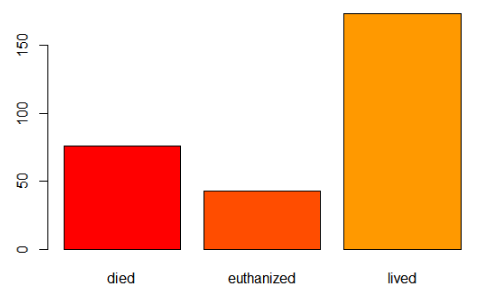


Figure 16: Colic Outcomes

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.

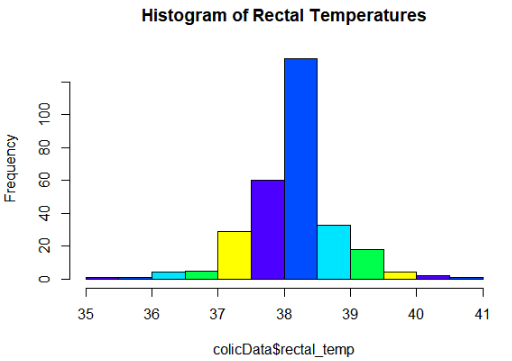
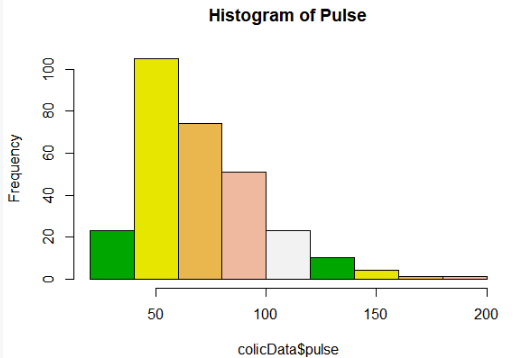
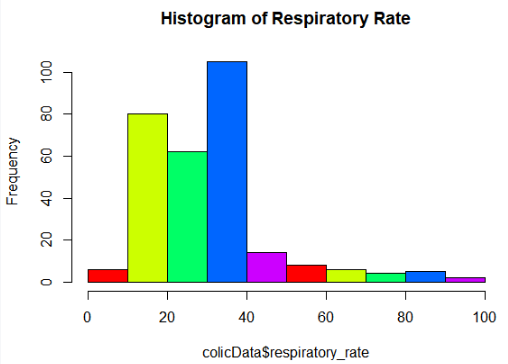
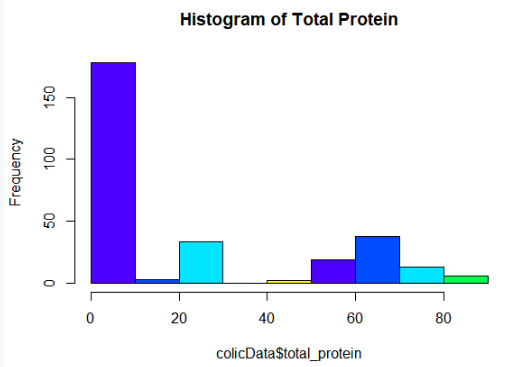
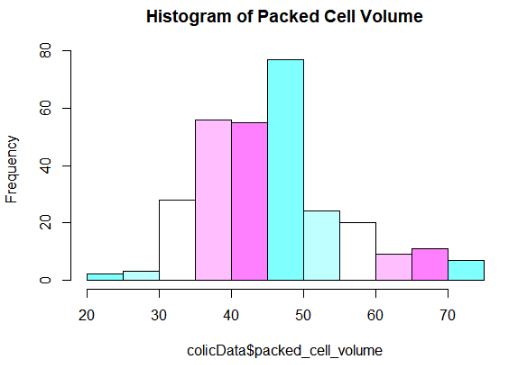
    

Figure 17: Colic Variable Histograms

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.

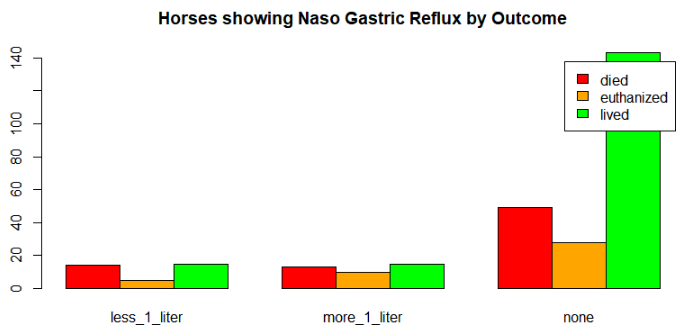
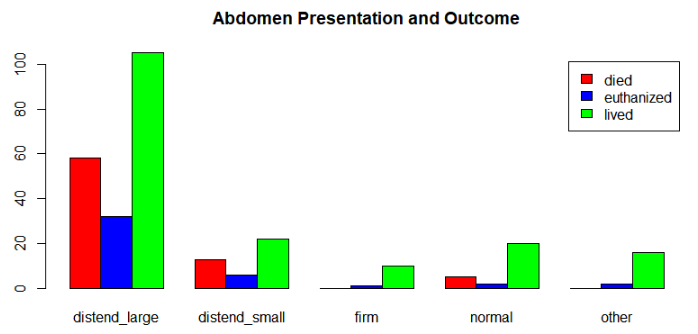
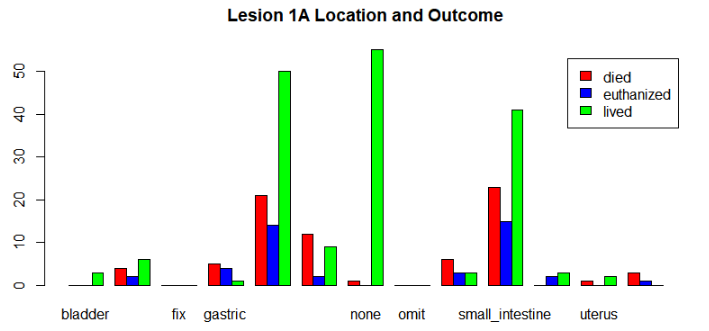
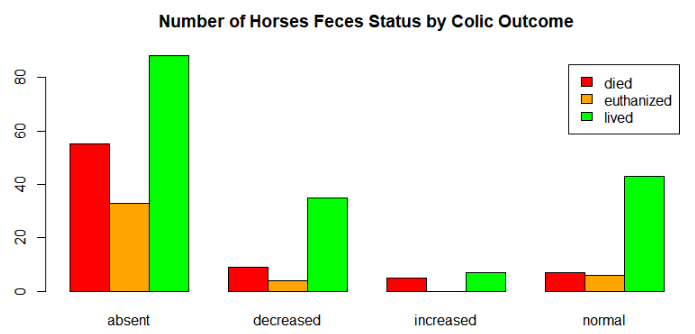
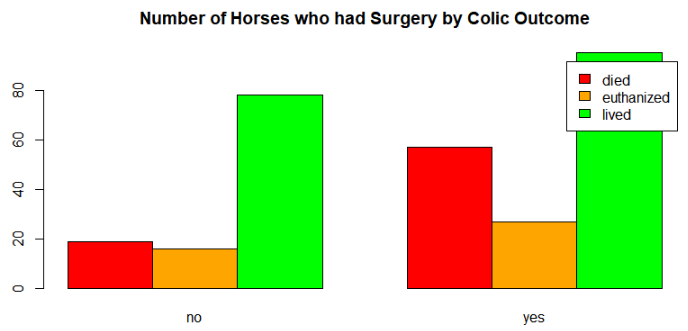
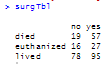
     

Figure 18: Additional Bar Plots and Outcome Frequencies

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.

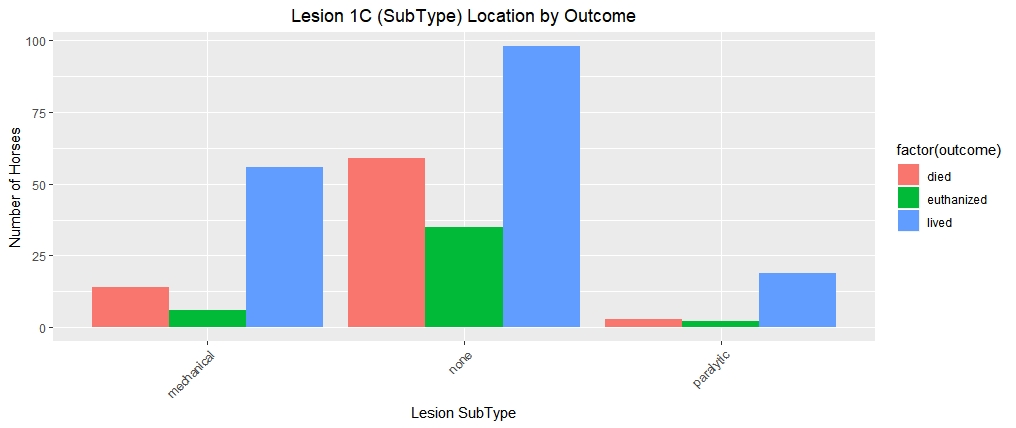


Figure 21: Lesion SubType related to Colic Outcome

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 following figure.

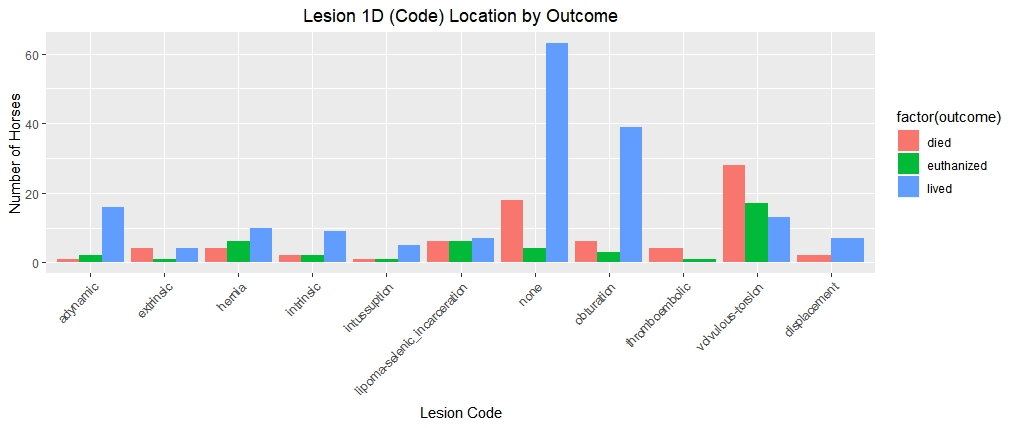


Figure 22: Lesion Code related to Colic Outcome

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.

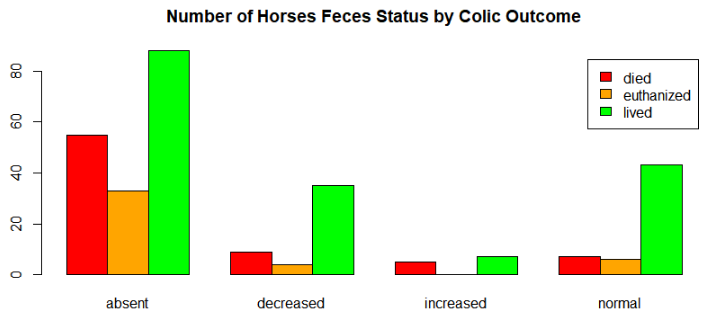


Figure 23: Feces to Colic Outcome

### Detailed Visualizations

Additional visualization proved useful as well. As shown in the following figure, 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.

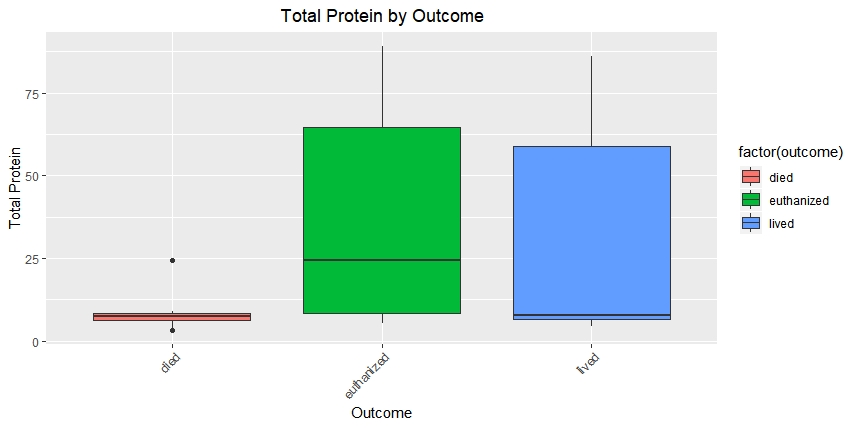


Figure 25: Box Plot of Total Protein and Outcome

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 the following figure 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.

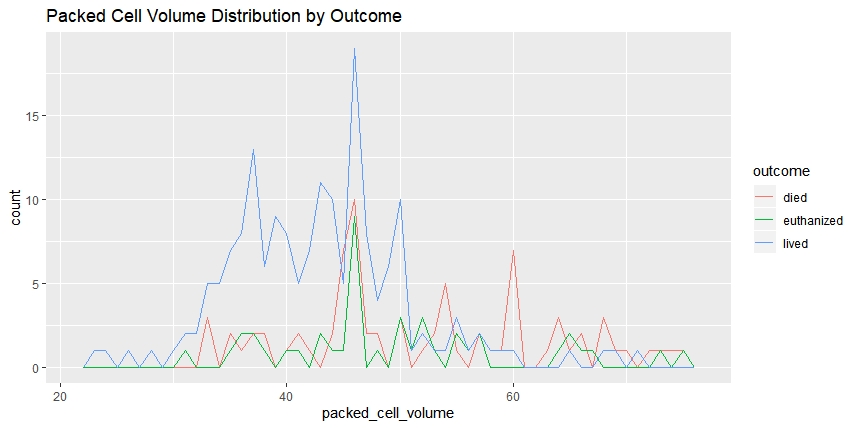


Figure 26: Packed Cell Volume by Outcome

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.

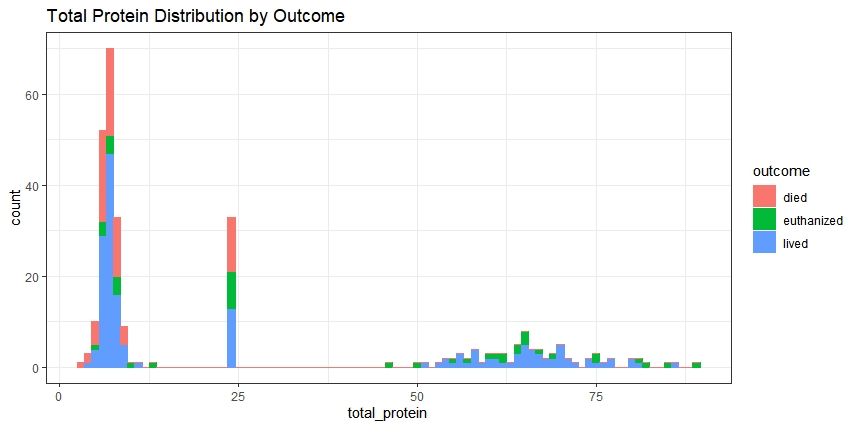


Figure 27: Packed Cell Volume by Outcome

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.

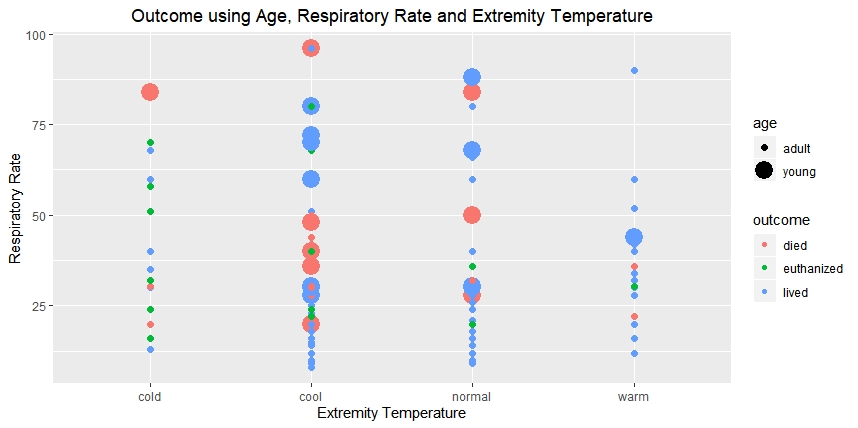


Figure 28: Outcome by Respiratory Rate and Extremity Temperature

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.

### Clustering

The clustering approach deemed very useless due to the large amount of information collected as shown in the plots below.

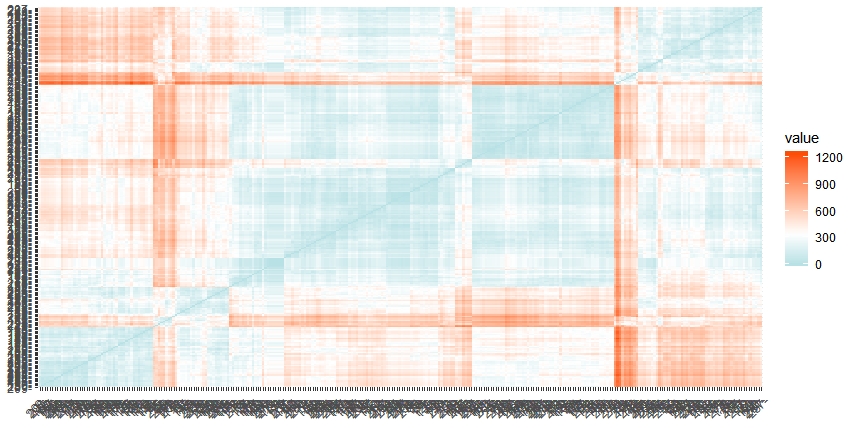


Figure 29: Clustering by Manhattan Distance

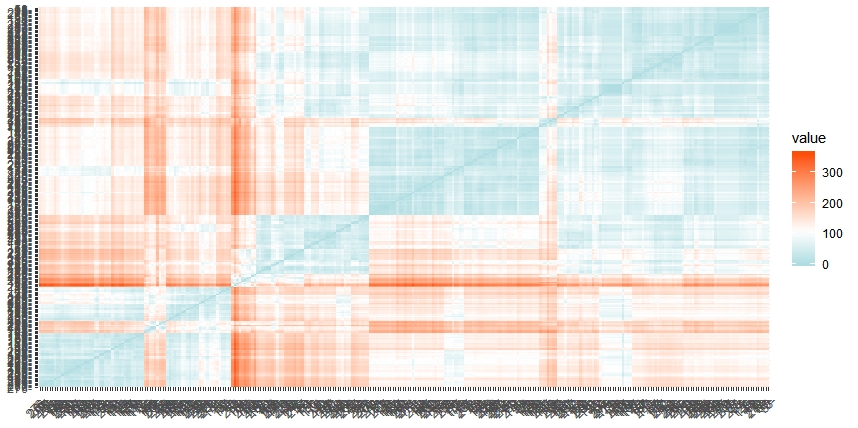


Figure 30: Clustering by Euclidean Distance

### 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 the following figure.

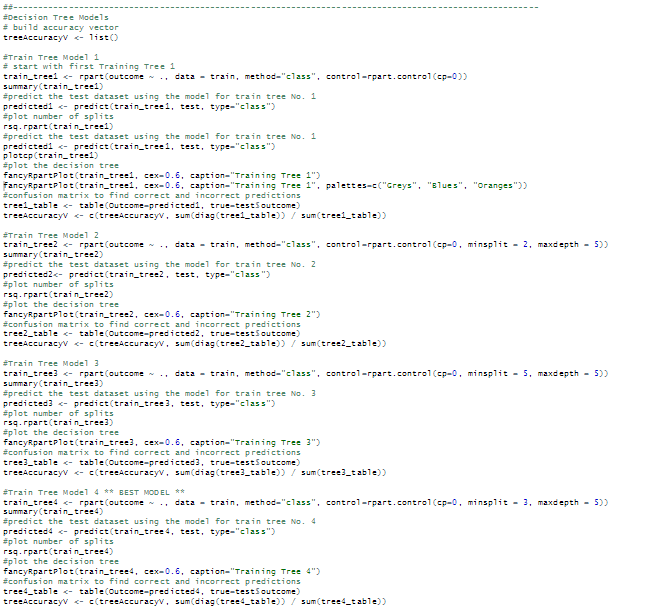


Figure 31: Decision Tree Code

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

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 “other” or “strangulation”, then the next node represents the total protein being below 9.5. In this case, 41% will die, 26% will be euthanized and 33% will live. Otherwise, 9% will die, 51% will be euthanized and 40% will live. This continues showing a good depiction of the variables of importance:

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

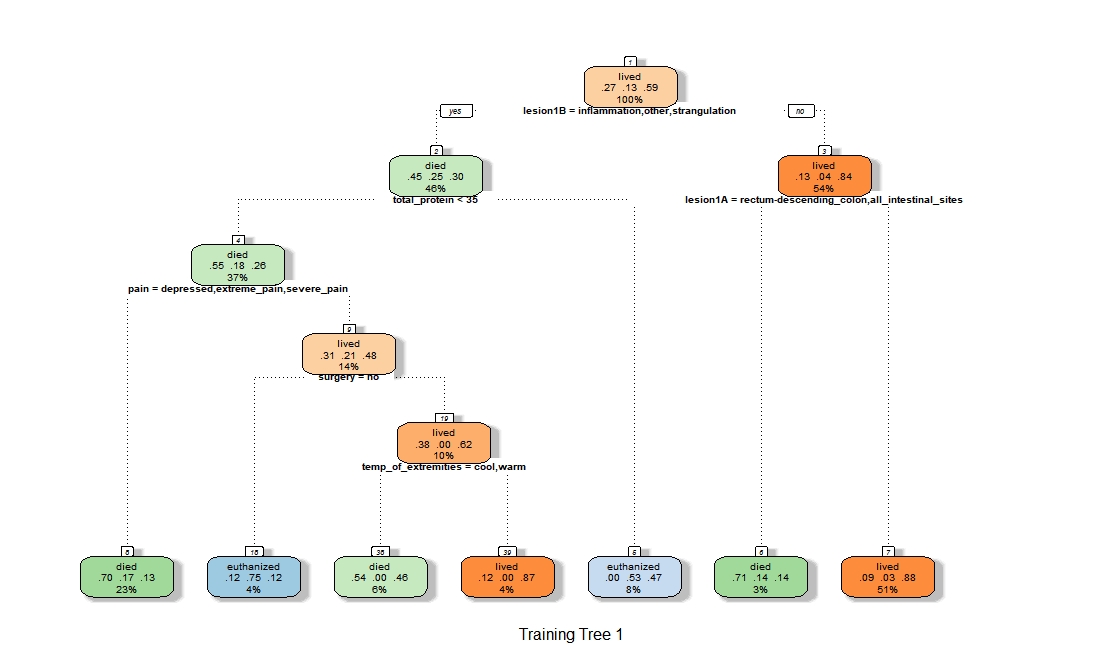


Figure 32: Decision Tree One

For the next tree, the minimum split and maximum depth were specified to be 2 and 5, respectively. This tree is far more complex as shown in the following figure.

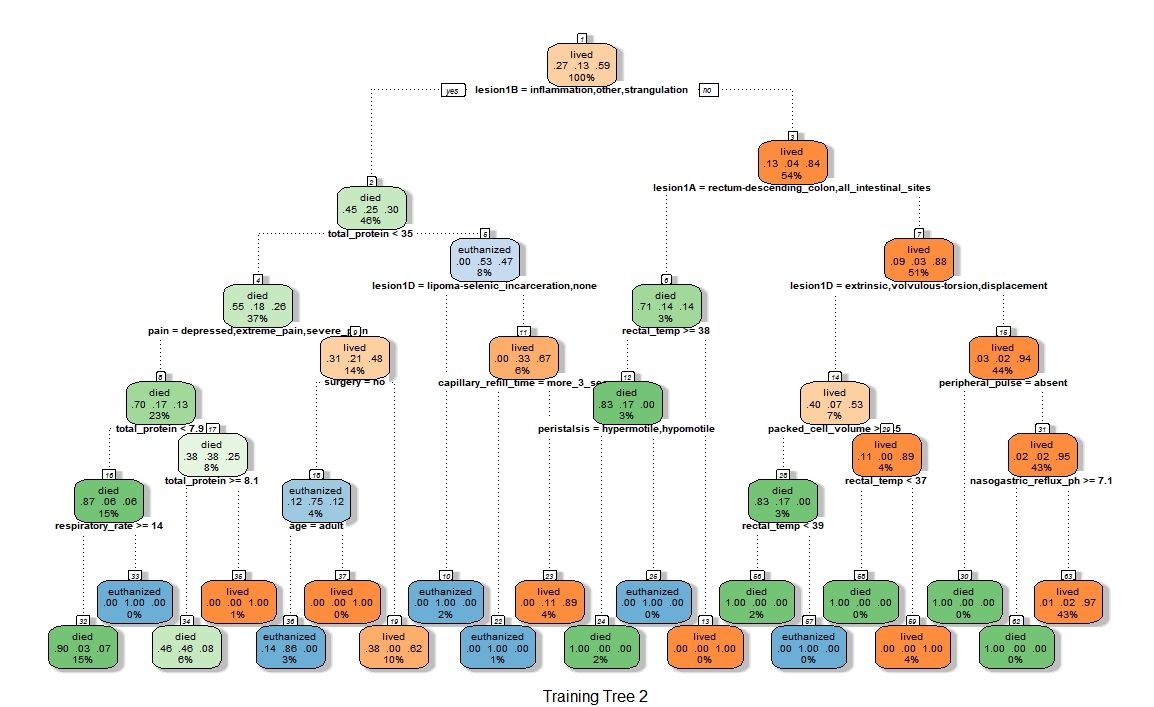


Figure 33: Decision Tree Two

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.



Figure 34: Accuracy of Decision Tree Data Frame

More information on the selected tree (the fifth) with accuracy of approximately 73% is reflected in the following information. 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.

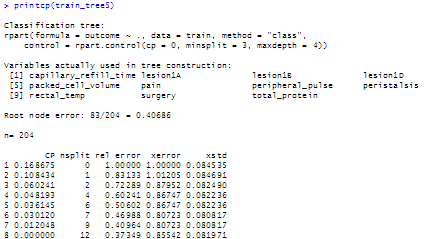


Figure 35: Summary of Classification Tree Five

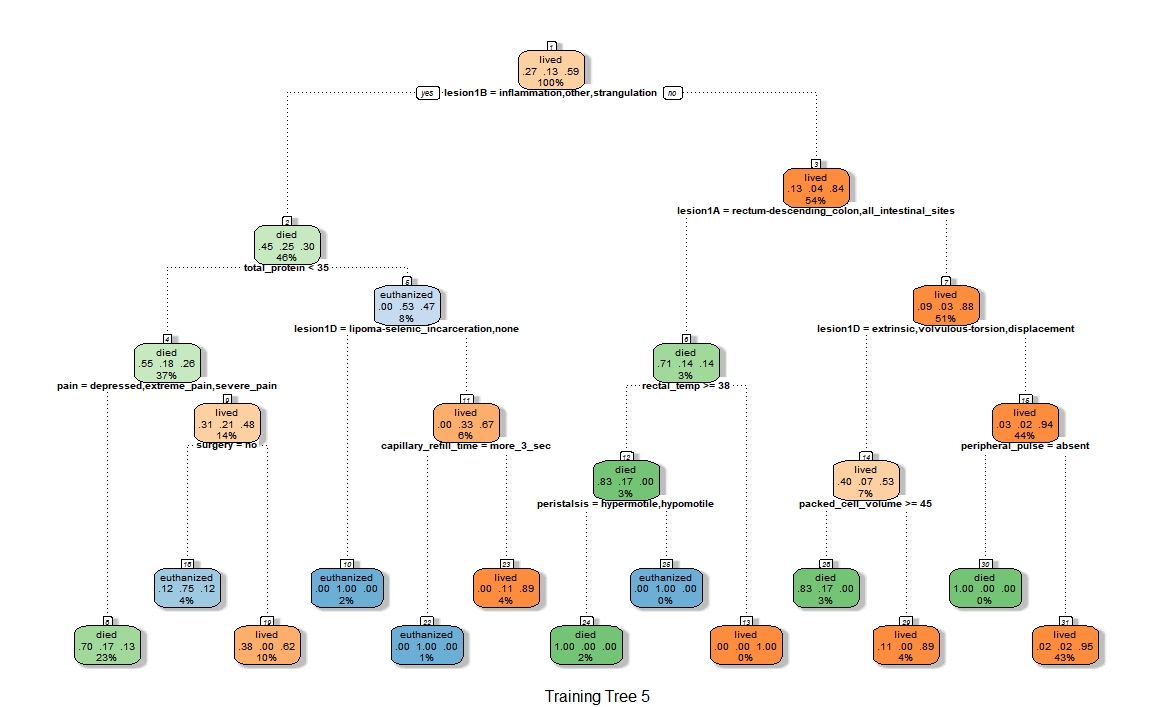


Figure 36: Decision Tree Five

This tree was pruned and the following representation was the result simplifying the tree. However, the pruned accuracy was slightly less than the originally accuracy at 70.45%. This is an acceptable modification because the tree was significantly simplified through pruning.

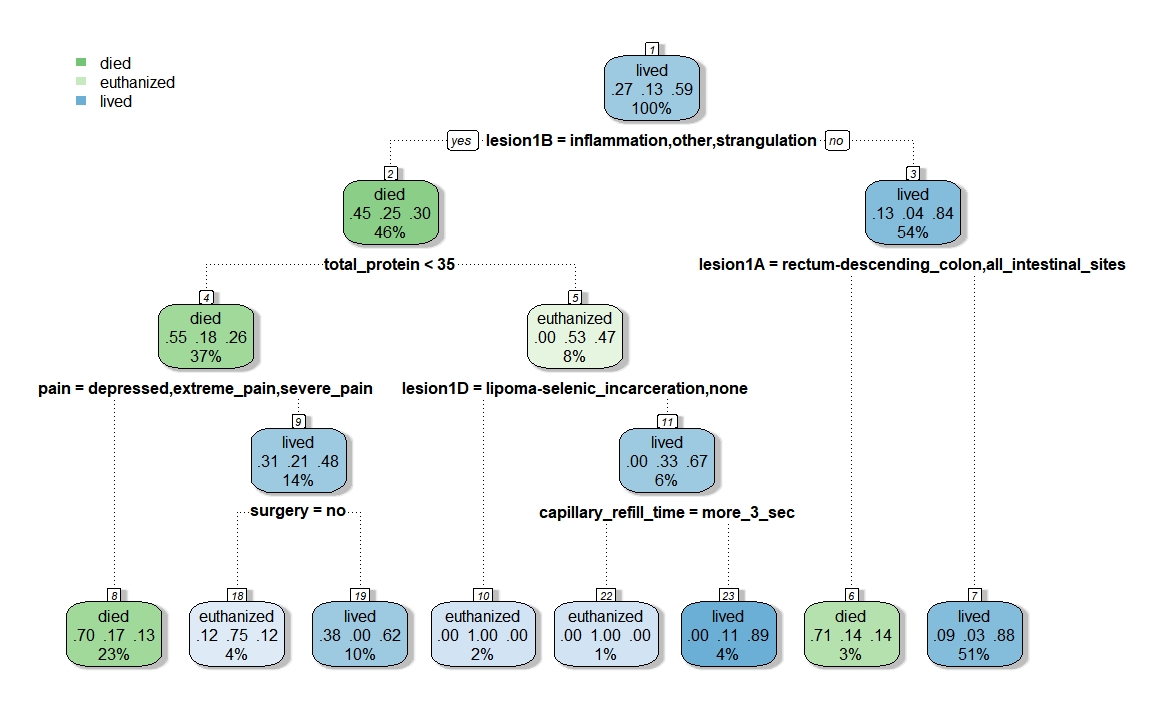


Figure 37: Pruned Tree

The decision trees were repeating using cross validation with 4 folds. The accuracy for certain splits and depths can be found in the following table. Notice how similar the accuracy is on these varying models.

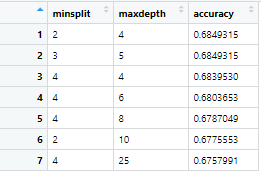


Figure 38: Accuracy of Decision Tree Data Frame

### Associative 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.

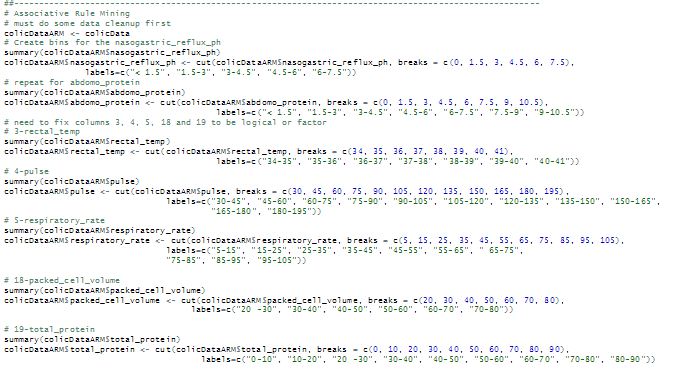
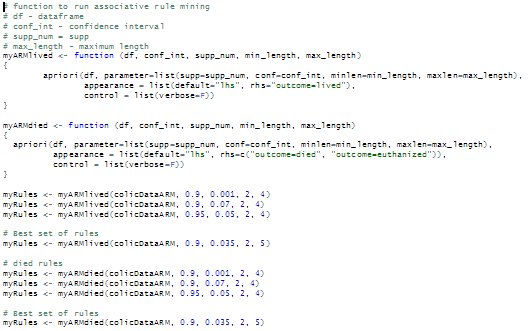


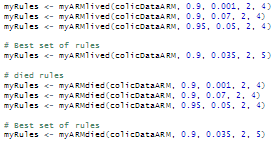
Figure 39: ARM Data Breaks for Continuous Variables

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.



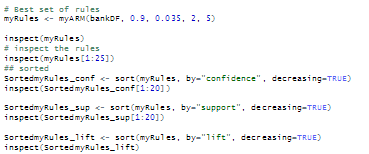
*Figure 40: Functions to run Association Rules*

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*

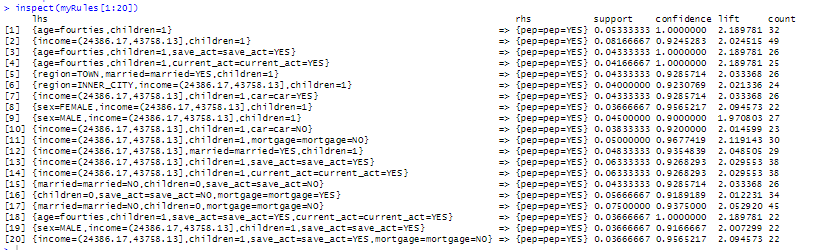
Many different combinations of confidence, support, minimum length and maximum length were evaluated. In some cases, selecting a higher minimum with a higher confidence provided less than twenty rules for evaluation. In other cases, raising the minimum confidence and support numbers provided limited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



*Figure 10: Ruling down the Rules*

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 a customer purchasing a Personal Equity Plan.:

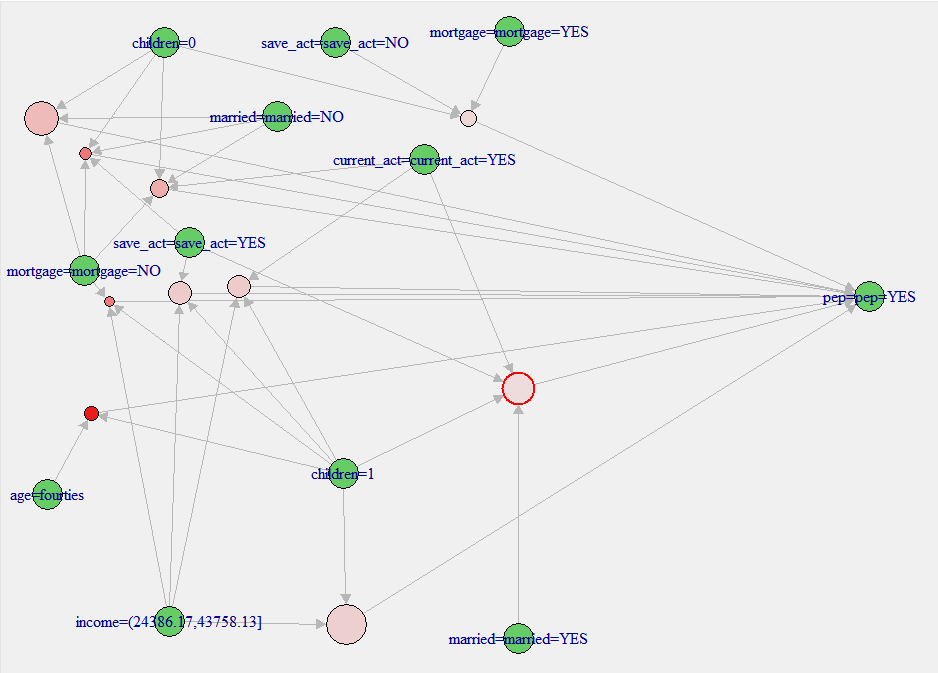
* Confidence minimum of 0.9
* Support minimum of 0.035
* Minimum number for ‘left hand side’ = 2
* Maximum number of ‘left hand side’ = 5



*Figure 11: First Twenty Rules obtained with Support, Confidence and Life*

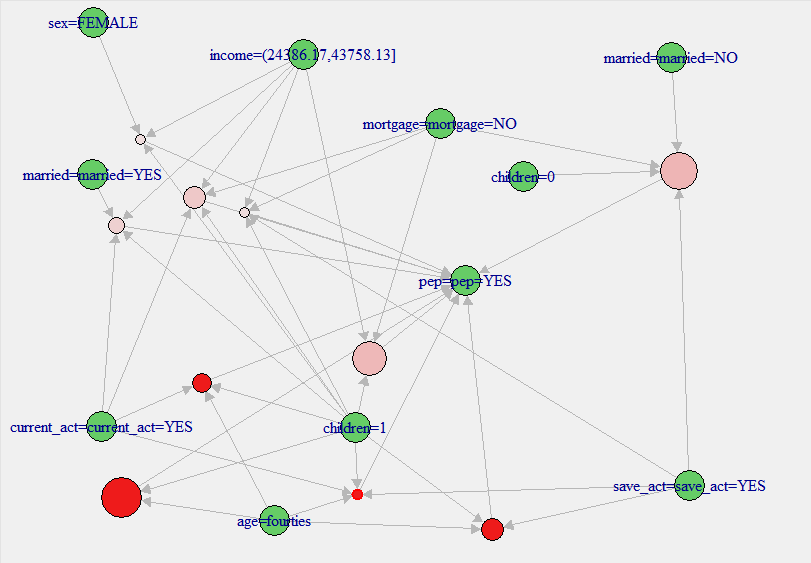
Visualization

Some interesting plots were run on the rules to show some visualization of what was taking place. The first was plotting the top ten sorted rules based on support in an interactive map.



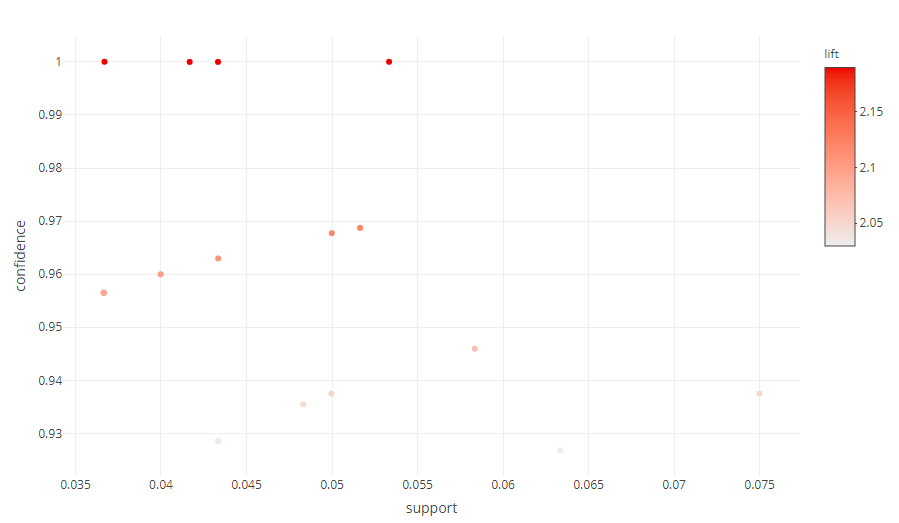
*Figure 12: Interactive Plot of Top Ten Rules by Decreasing Support*

This was followed by a similar visualization of the top ten rules sorted by decreasing confidence which clearly showed some very different results.



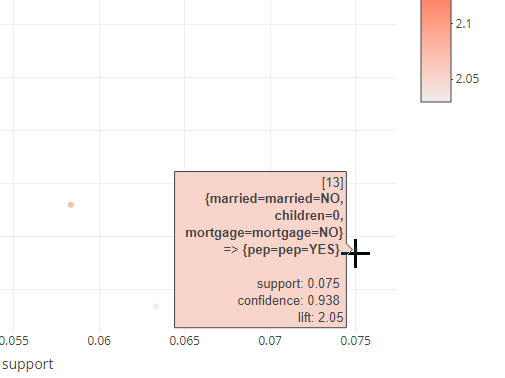
*Figure 13: Interactive Plot of Top Ten Rules Sorted by Decreasing Confidence*

A visualization was then done using a deprecated function that still provided the plot, *plotly\_arules* which shows a very interesting scatterplot of the rules. The following figure shows the top twenty rules sorted by decreasing confidence displayed with support on the *x*-axis and Confidence on the *y­*-axis and color coded by lift.



*Figure 14: 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 15: Rule 13 with Support=0.075, Confidence=0.938 and Lift=2.05*

This final plotting method served useful in selecting rules of interest that are discussed in the Results section.

### Linear Regression (LM)

The first model helped me to determine which, if any variables are influencing the Satisfaction Rating as well as determine how much influence they have. This is determined by looking at the Adjusted R Squared Value (to show the relevance/strength of the model) and the p-values which should be < 0.05 to show that the individual independent variable has significant influence on the Satisfaction Rating (the dependent variable).

Initially, I looked as some simple modeling and information to rule out certain variables. For example, I completed a quick linear regression model with just how long a flight took and that correlation to flight satisfaction as shown in the following figure.

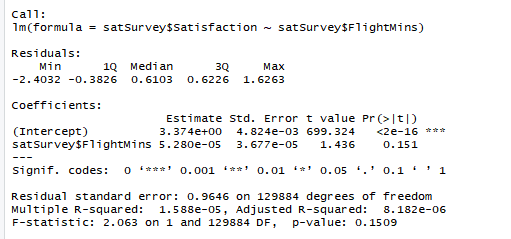


Figure 23: Satisfaction based on Total Flight Minutes Linear Regression Model

As shown in the details from this linear regression model, the p-value value was quite a bit higher than 0.05 (0.151) and the Adjusted R-squared value is very low indicating no correlation between these two pieces of data.

I completed some more simple models and found that although the Adjusted R-squared value was low, the p-value was low for the Status (Silver, Gold, Platinum) with the airline. The same was true of the number of flights taken be the individual.

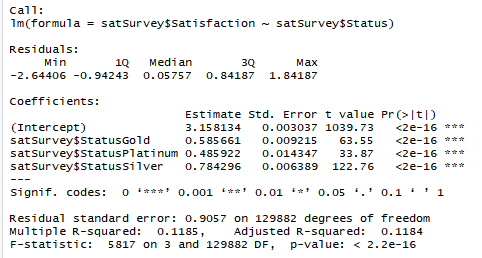
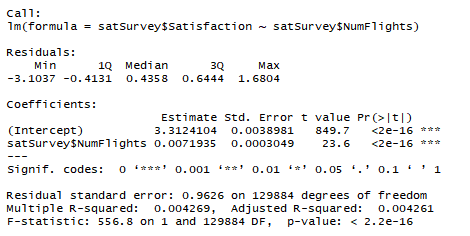
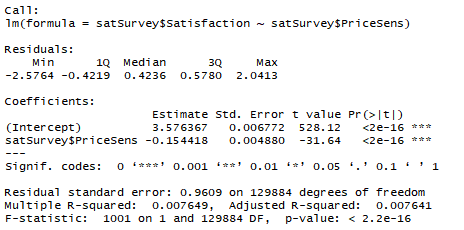
  

Figure 24: Satisfaction Model based on Status, Total Number of Flights, Price Sensitivity

To gather some additional data on the satisfaction, I decided to narrow down the survey and create a new data frame of just the top 3 airlines based on their mean satisfaction value:

* West Airways Inc (HA)
* Cool&Young Airlines Inc. (VX)
* FlytoSun Airlines Inc. (AS)

This can then be used to look at how we might be able to improve the satisfaction of Southeast Airlines Inc. (US). I obtained the Airline Codes for these by filtering the satSurvey data with RStudio.

After creating this new data, I created a linear regression on these airlines and found that the variables listed in the following figure were the ones that influenced satisfaction for these three airlines. I ran the AIC technique against all the ***all*** the airlines and got the following for the best model which appears to the right of the best airline model.

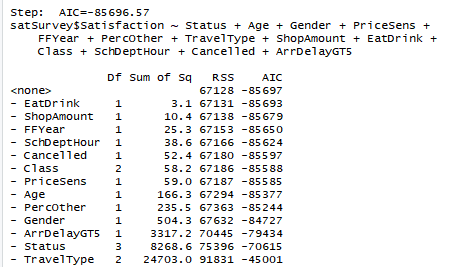
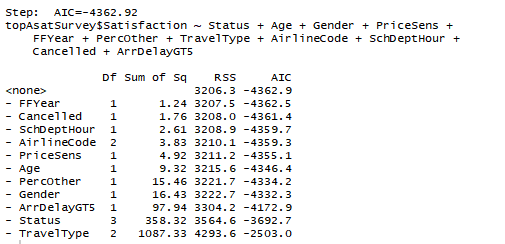


Figure 25: AIC Best Linear Regression Model for Top Three Airlines and All Airlines

You will see that these models have quite a few variables in common:

* Scheduled Departure Hour
* Flight Cancelled or Not
* Price Sensitivity
* Gender
* Arrival Delay Greater than 5 minutes
* Percent of travel on other airlines
* Type of Travel – business, personal, mileage award
* Age of traveler
* Year of First Flight

The variables that appear in only one model are as follows:

* Eating or Drinking in the airport
* Shopping Amount in the airport
* Class of Travel
* Airline Code – only in the top three airlines model

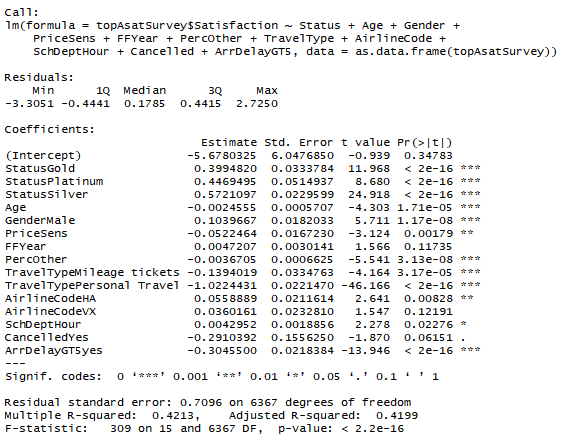
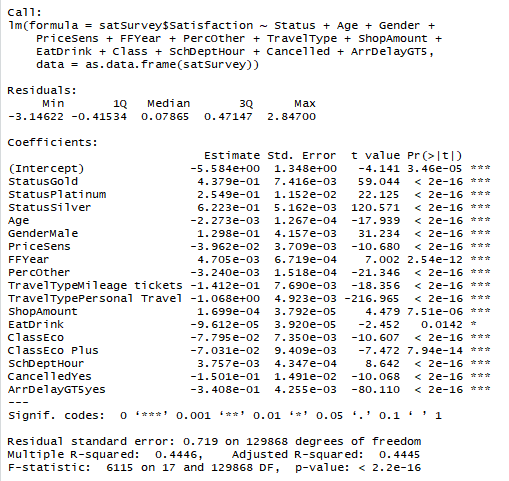
 

Figure 26: Summary of the Best Linear Regression Model for Top Three and All Airlines

The figure above provides a summary of the variables for each of the models (top 3 airlines and all airlines). As you can see in the model on the right, eating and drinking in the airport had the largest p-value, so I elected not to use that variable in the final linear regression model. In addition, the airline code had a larger p-value in the top three airline model and I elected not to allow this to interfere in the model. The final model chosen had a union of the two models shown above without EatDrink and AirlineCode.

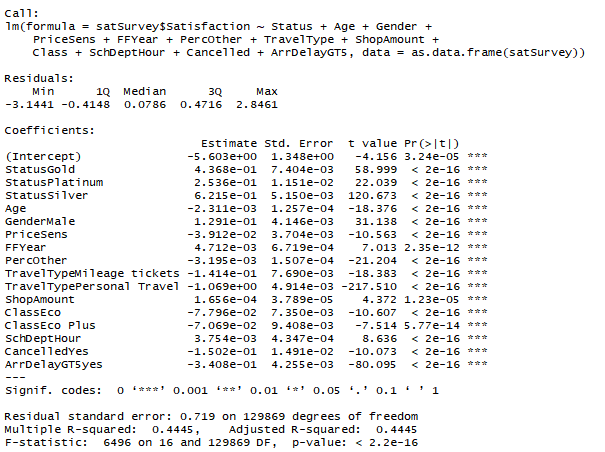


Figure 27: Summary of the Final and Best Linear Regression Model

As I reviewed the coefficients associated with the variables of influence, you can see how much these individual variables increase or decrease satisfaction. For example, we can see that Price Sensitivity (PriceSens) have a negative coefficient, so it brings the overall satisfaction down if you are price sensitive. In addition, if the traveler has flown often on other airlines, is traveling Economy or Economy Plus status, the flight is cancelled, or the arrival delay greater than 5 minutes – satisfaction is likely to be decreased. However, predicted satisfaction is likely to increase if the traveler is male traveling that holds Gold, Platinum or Silver status. In addition, if he/she has been shopping before his/her flight, satisfaction is higher.

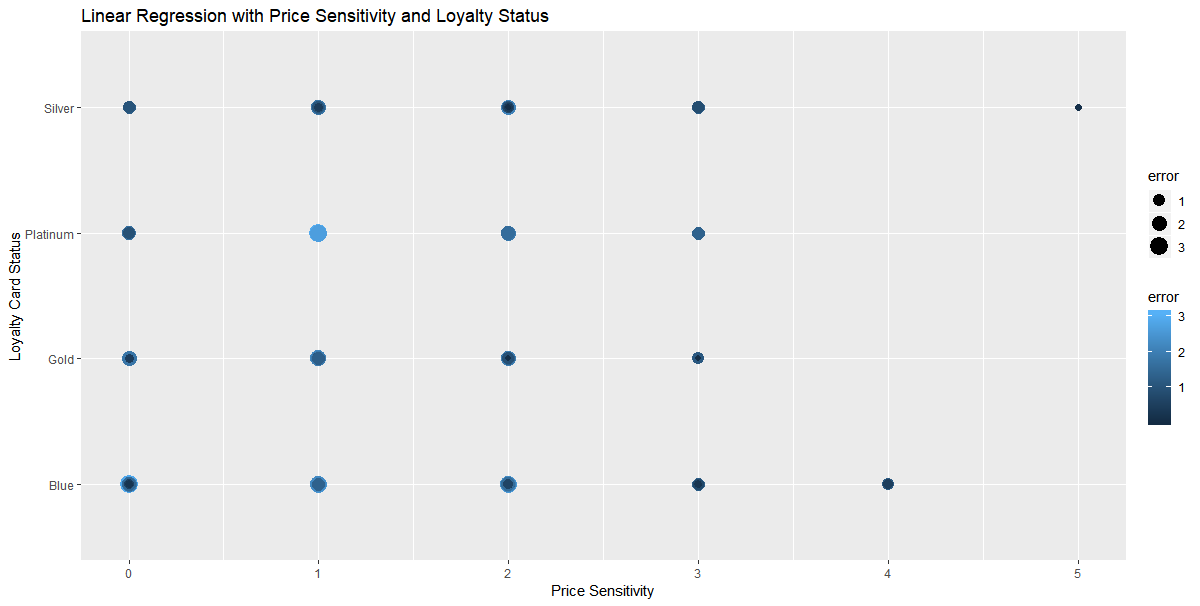


Figure 28: Best LM showing Price Sensitivity and Status with Prediction Error

As shown in the plot above, our linear regression model does a fairly good job predicting when the Status is Silver, Gold or Platinum with a higher level of Price Sensitivity.

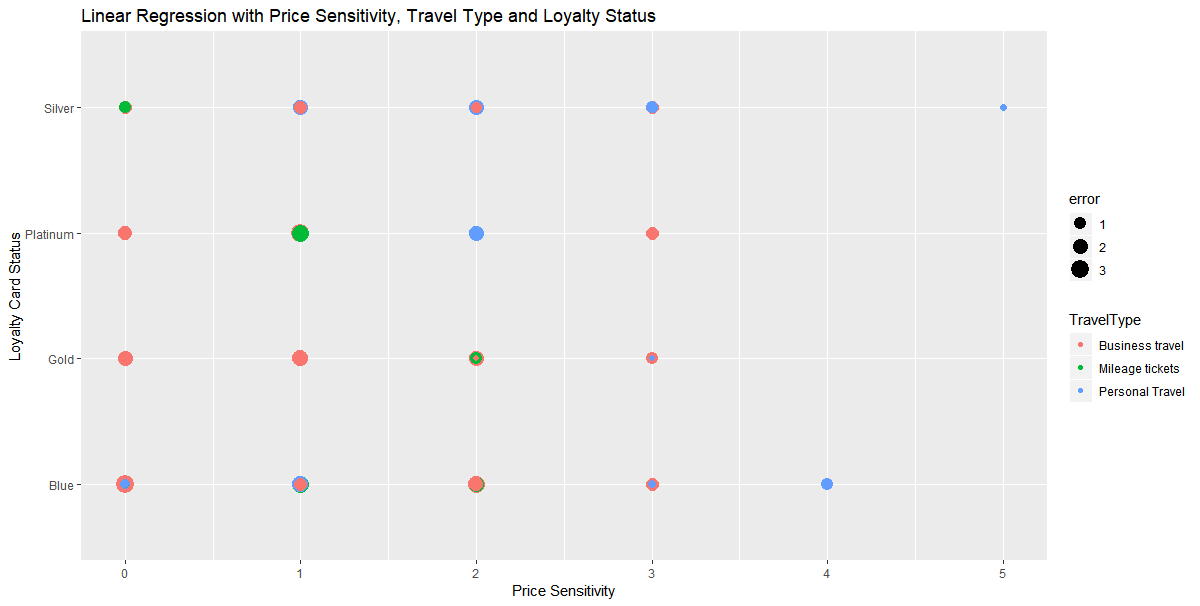


Figure 29: Best LM showing Price Sensitivity, Travel Type and Status with Prediction Error

As reflected in the figure above, again the model shows low errors between our predictions with the model and the actual values when we are looking at the Silver and Gold members with high price sensitivity.

With a linear regression model and our data, the plots do now show much of a pattern. This is due to the finite number of values like Price Sensitivity of 1 through 5 only, or Status of 4 different values, etc. There is very little continuous data except delays that can be found in our data, but that led me to the next plot.

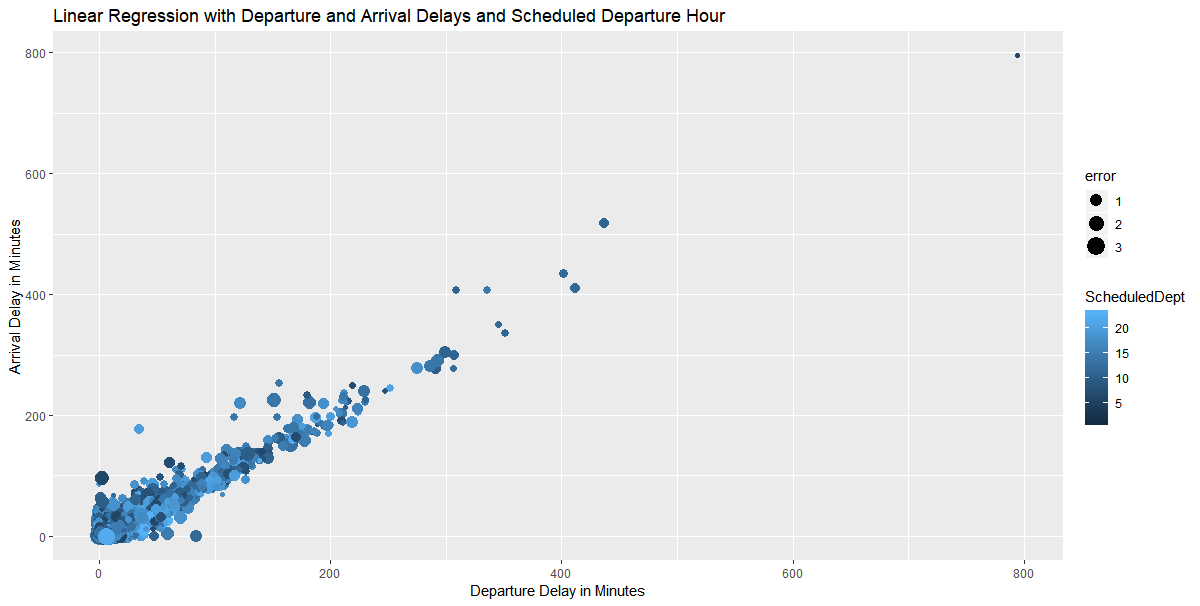


Figure 30: Best LM showing Departure and Arrival Delays with Prediction Error

As shown above, we can predict poor satisfaction when the delays are extensive. In fact, the errors get pretty obvious (very small) as we move to longer arrival delays and longer departure delays. However, it is harder to predict the satisfaction when the delays are more minimal.

In an attempt to see this a bit better, I changed the scale to blow up the plot. Although we missed 2 points taking this approach, this helps us to see the information at the shorter arrival and departure delays. It is clear that the predictions are better the longer the delays.

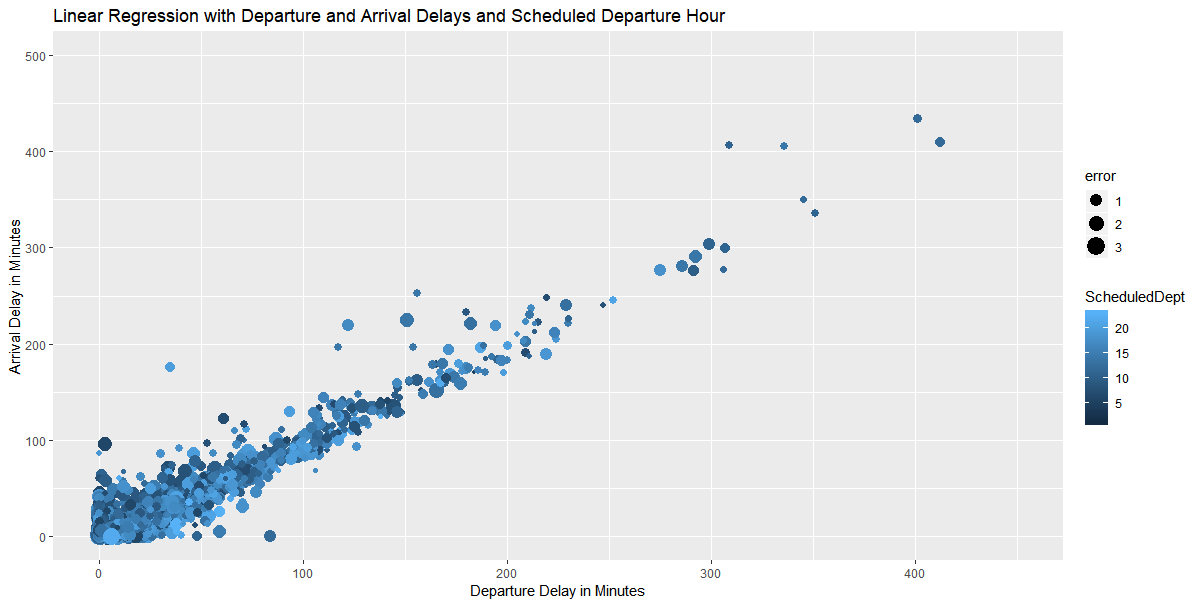


Figure 31: Zoomed Best LM showing Departure and Arrival Delays with Prediction Error

Zooming in as shown in the plots above and below show there is definitely a pattern, but as we look at the errors, the size of the dots are smaller as the delays get larger meaning our model is doing a better job at predicting in these circumstances. This next zoom reflects an area with a drop of 41 observations.

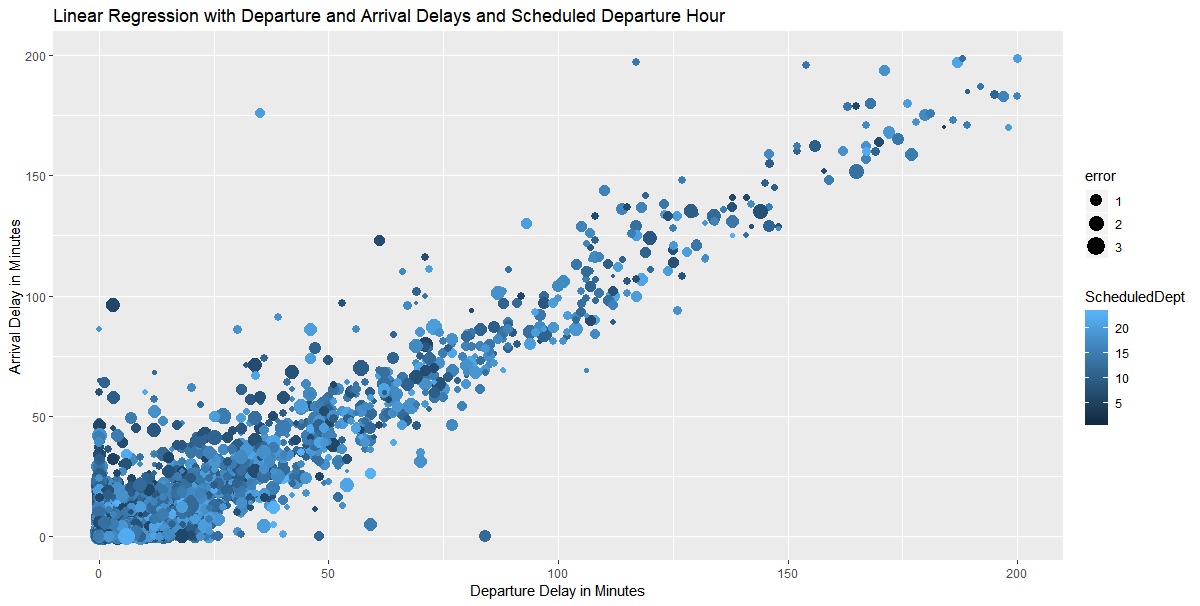


Figure 32: Zoomed Again Best LM showing Departure and Arrival Delays with Prediction Error

### Support Vector Machine (SVM)

The next model that I used in reviewing 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.”[[1]](#footnote-1)



Figure 33: SVM showing Price Sensitivity and Status with Prediction Error

In reviewing the plot above, we see similar results to that in our LM model and that as our errors are smaller when both Price Sensitivity is higher and the loyalty card is Silver or Gold.

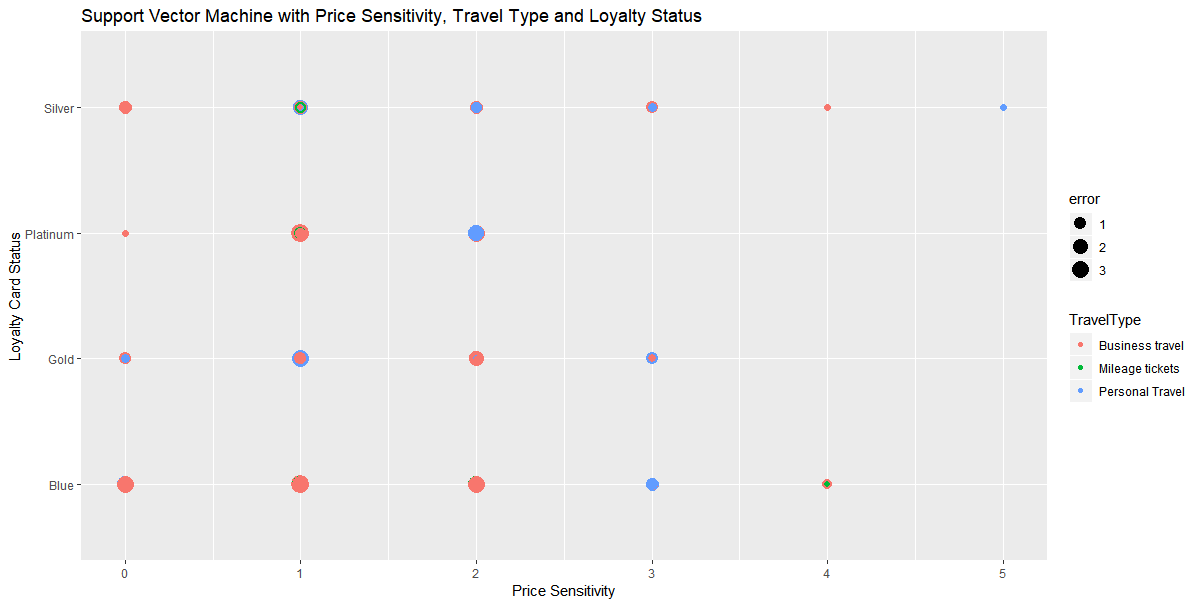


Figure 34: SVM showing Price Sensitivity, Travel Type and Status with Prediction Error

Similarly, our model does a better job prediction Satisfaction when Price Sensitivity is higher. This may mean that Price Sensitivity has a significant bearing on the overall satisfaction rating.

Zooming on a similar plot to that done of our Linear Regression Model This next zoom reflects an area with a drop of 33 observations from our test data. You will see that as the Arrival Delay increases and/or the Departure Delay increases, we do a better job prediction satisfaction (reflected by the smaller dot for a smaller error).

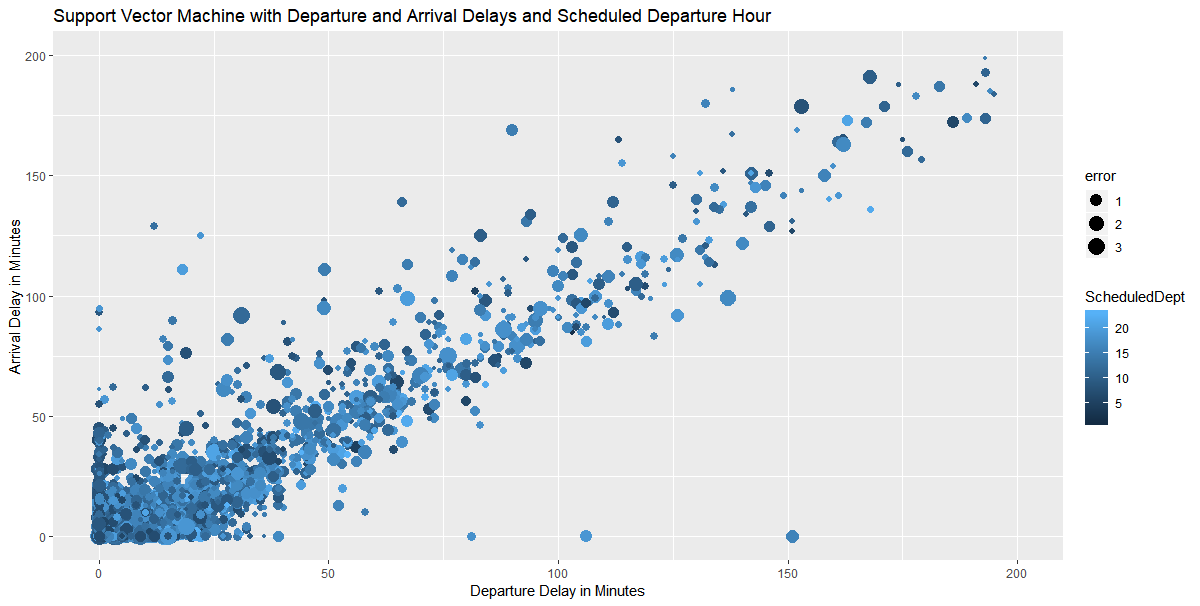


Figure 35: Zoomed SVM showing Departure and Arrival Delays with Prediction Error

### Kernel Support Vector Machine (KSVM)

The next model selected was the Kernel Support Vector Machine (KSVM), which is like SVM, but supports the well-known C-svc, nu-svc, (classification) one-class svc (novelty), eps-svr, nu-svr (regression) formulations along with native multi-class classification formulations and the bound-constraint SVM formulations. KSVM also supports class-probabilities output and confidence intervals for regression.[[2]](#footnote-2)

For KSVM, I just calculated the Root Squared Mean Error and did not produce additional plots for this model.

### Naïve Bayes

The final model selected was the Naïve Bayes Classifier. “Naive Bayes classifier calculates the probabilities for every factor. Then it selects the outcome with highest probability.”[[3]](#footnote-3)

Unfortunately, I had some difficulties with this model and was unable to product results for this final project.

#### Summary of Modeling

As I review the models (Linear Regression, Support Vector Machines and Naive Bayes), I review the Root Mean Squared Error (RSME) for each of the models. These were as follows:

* For our best Linear Regression model, RSME = 0.7146178
* For our SVM model, RSME = 0.7500606
* For our KSVM model, RSME = 0.7250396
* For Naïve Bayes, RSME = *TBD*

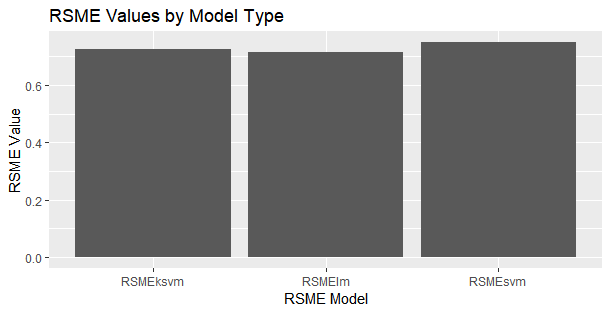


Figure 36: Root Mean Squared Error by Model Type

Using this as an evaluation, I would conclude that our Linear Regression model is the best predictor of Satisfaction based on the fact that it has the lowest Root Mean Squared Error.

# Results

As shown in the previous section, there are quite a few variables that have a direct affect on how a traveler will rate his/her overall satisfaction with a specific trip. For example, the loyalty card or airline status (Blue, Silver, Gold, Platinum) has a strong effect on the satisfaction rating. One would assume the perks associated with those cards, has an impact on the satisfaction rating provided by most flyers. This is expected.

Ironically, on our modeling, we found that some unexpected items such as shopping amount in the airport before the flight as well as the eating or drinking in the airport before the flight had influence on the overall rating given to the airline by the flyer. The latter had much less influence, but still played a role in the overall satisfaction rating.

We can also see in our graphs in the visualization section that the destination and origination location have some effect on our overall satisfaction, but they are not of a great influence. We can also see that even though departure and arrival delays can help us predict satisfaction, they are not something that contributes (when reviewing our linear regression model) greatly to the overall satisfaction rating. That being said, that is a good thing since delays in most cases cannot be predicted or even minimized.

After all of our modeling and analysis, the final group of variables determined for our best model are found in the following table.

| **Variable** | **Definition** |
| --- | --- |
| **Satisfaction** | **Rating from 1 (low) to 5 (high) of satisfaction** |
| Airline Status | Status with airline (frequent flyer) |
| Age | Age of traveler |
| Gender | Gender of traveler – male or female |
| Scheduled Departure Hour | Hour the Flight was scheduled to depart |
| Class | Class of travel (first, business, etc.) |
| Price Sensitivity | Sensitivity to ticket price |
| Arrival Delay Greater Than 5 Minutes | Arrival Delay greater than 5 minutes |
| Type of travel | Business, personal, mileage award |
| Percentage on other airlines | Percentage of travel on other airline carriers |
| Year of First Flight | The year the first airline flight was taken |
| Shopping Amount | Amount of money spent shopping in the airport |
| Flight Cancelled | If flight was cancelled (yes/no) |

Figure 37: Final Variables Determined in Modeling

# Conclusions

As a result of the analysis done on this data, I would recommend the following for Southeast Airlines:

* Target Male Flyers

This can be done by sending specials deals to male travelers or creating a campaign to target getting more male travelers to fly with Southeast Airlines.

* Increase Loyalty Members

Since it is clear that Silver, Gold, and Platinum flyers are more satisfied by our plots and the positive coefficients in our model, it would be good for Southeast Airlines to provide more ways to increase loyalty members. This could be done by finding ways to bump flyers up to the next loyalty level. For example, offering double miles for certain fights or decrease the requirements to obtain the next loyalty level.

* Pricing

Since sensitivity to price affects satisfaction, offering reduced pricing for certain legs or flights or possibly discounted pricing for different types of travel would give Southeast Airlines the ability to increase their overall satisfaction rating for those customers that are more sensitive to price.

* Shopping

I found this particular variable to be unexpected, but since it has some influence – it should be addressed. Southeast Airlines could provide coupons for shopping in the airports before a flight as a perk. For example, shop the XYZ Store for 15% off before your flight.

* Delays

Obviously, minimizing delays would help the airline increase satisfaction. This might be as easy as extending the arrival time by 5 or 10 minutes so that it is a more attainable goal so that it seems that the flight is closer to the scheduled time.

Missing Data

1. How long ago was the horse’s last colic?

1. Mare or gelding?
2. Any recent stressors? Move, Show, Pregnancy, etc.
3. Breaking up age into more groups: under 6 months, yearling, 1 <= 3 years, 4 <= 10 years, 10 <= 15 years, 15 <= 20 years, over 20
4. How long experiencing colic symptoms
5. On any medication? Doxyclicine, tetracycline, etc.
6. Any reflux or not (indicates impaction).

# 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 2: SVM (Support Vector Machine) – Theory. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [↑](#footnote-ref-1)
2. KSVM. https://www.rdocumentation.org/packages/kernlab/versions/0.9-27/topics/ksvm [↑](#footnote-ref-2)
3. 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-3)