3-26-2013

A Statistical Analysis of the Factors that Potentially Affect the Price of A Horse

Caitlin Allen
The College at Brockport, caiteallen_812@yahoo.com

A Statistical Analysis of the Factors that Potentially Affect the Price of A Horse

A Senior Honors Thesis

Submitted in Partial Fulfillment of the Requirements
for Graduation in the College Honors Program

By
Caitlin Allen
Mathematics Major

The College at Brockport
April 26, 2013

Thesis Director: Dr. Hong Yin, Assistant Professor, Mathematics

Buying and selling a horse for a reasonable price is difficult. Horses can range in price from a couple hundred dollars all the way up to hundreds of thousands of dollars. The goal of this thesis is to investigate what factors determine the price of a horse. Once these factors are determined the goal is to come up with a regression model and equation that will allow future calculation of the price of a horse. If an equation can be established, when one is looking to buy a horse they could plug in the values and see if the selling price of the horse is reasonable. This would be very helpful because one wants to make sure that the horse is not overpriced. The equation would also be helpful when selling a horse, by easily determining the price to sell the horse at by plugging in the values of the factors into the equation to get a rough estimate of price. I have always been in the horse world. I grew up with horses and around horses. I grew up with two horses right at home. I soon became interested in riding and began riding at the age of nine years old. At the age of twelve I got my own horse, who I still have. Being involved in the horse world I was interested in what determined the price of a horse; was it the training, the breed, the color, or other factors. Not too long ago I thought about selling my horse and trying to figure out what price to sell her for was hard to come up with; it was difficult to come up with a price that I thought met her value.

The first step was to determine which factors I should focus on that would potentially determine the price, level of training, temperament, location, color, vices, markings, registration, and who was selling, a private individual or a barn. These twelve factors are the main factors one looks at when buying or selling a horse. Since the internet is widespread allowing more accessibility to a broader market, most people are listing the horses that are for sale on the internet through a variety of sites made just for selling horses, as well as Craig's list. Since the internet was easily accessible I used it to collect my data, looking through twenty different horse

selling sites to collect 120 different samples. The sites I used were either ones I already knew of or I simply searched for horse selling sites. These sites provided me with a wide variety of horses for sale, one example is a Lipizzaner. Some sites specialized in certain types of horses, such as Bigeq.com which specializes in very well known and high placing Hunter and Jumpers, meaning they were all very expensive. While many sites listed all different horses breeds sold by private sellers and/or barns. Collecting 120 samples gave me a large enough sample that would allow for errors, such as not all of the determining factors being listed. To collect my data I created a spread sheet with each factor listed across the top and noted what each sample had listed for each category. While collecting the data I was surprised to see the range of breeds I found as well as the range of prices I found. Breeds ranged from Miniature Pony all the way to Dutch Warmblood, with prices ranging from as little as $500 to $125,000. If the factor was not clearly written or not listed at all I left that factor blank for that specific sample. In order to get a range of data I collected data throughout the entire month of July. This allowed newer horses to be listed. From each site I would make sure to go to all different pages, to get a wide range, from those newly listed to those having been listed for a few months.

The factors I looked at were breed, sex, height, training, level of training, temperament, location, color, vices, markings, registration, and who was selling. There is an abundance of horse breeds from all over the world. There is a difference between pony and horse breeds. A breed generally has distinctive true-breeding characteristics over a number of generations. Some breeds can be cross referenced with another breed or are stemmed off of a main breed. For example, Hanoverians are a type of Warmblood. The sexes are listed as: gelding (castrated male), mare (female), and stallion (intact male). Subcategories of geldings, stallions, and mares

include colts and fillies. Age is the factor that determines colt/filly versus gelding/mare. A horse below the age of 3 is called a colt or a filly, but can also just be called mare and gelding.

The height of a horse is measured in hands abbreviated by hh. One hand is equivalent to four inches. Height determines whether animal is categorized as pony or horse. The cutoff determining height is 14.2hh where a pony is less than 14.2hh a horse is greater than 14.2hh. Within pony and horse there are 3 subcategories small, medium, and large. A small pony is 12.2hh and below, medium ranges from 12.3-13.2hh, large ponies range from 13.3-14.1hh. A small horse is in the range of 14.2-15hh, medium 15.1-15.3hh, and large is 16hh and up. There are two main disciplines that horses can be trained in English and Western. Under each there are a number of subcategories. The variations of English include Dressage, Eventing, Hunter, Jumper, Racing, and Saddleseat. The variations of Western include Reining, All Around, Barrel Racing, Cutting, and Calf Roping. Other disciplines include Trail Riding, which can be done either in English or Western, and Driving, which is when a cart is attached to a horse by a harness. The main difference between English and Western is the type of saddle used. A Western saddle has a deep seat with a high horn and cantle. An English saddle does not have a deep seat or high cantle or horn, English saddles tend to be smaller and allow the horse to move more freely.

How well a horse is trained in their discipline can be broken down into five different categories. Unbroke; meaning the horse has not been worked with by the owner on things such as brushing and basic skills and has never had a saddle put on their back. Green; the horse has been started, been sat on and has started to learn the basic commands of walk, trot, and canter. Well trained; the horse is trained in the specific components of the discipline and is doing well at them. Professionally trained is when a professional is hired to train the horse, these horses tend

to be more expensive. Show experience is also a determining factor on how well trained a horse is. Show experience means that the horse has competed at a horse show. The more show experience a horse has the more expensive they tend to be. If you are looking to enter the show circuit with your horse it is important for the horse to have show experience.

The temperament of a horse is graded on a scale from one to ten. Ten on the scale equals high strung and one equals very calm and quiet. Most horses range somewhere in the middle at around three, four or five. Location is self explanatory; I took down the City and State where each sample was being sold.

There are many ranges of colors for horses. There are patterned colors and non-patterned colors. Non-patterned colors include chestnut, bay, black, palomino, gray which includes dapple gray and flea-bitten gray, buckskin, dun, and roan, which includes, blue roans, red roans, and ray roans. The patterned color is pinto, which includes tobiano, overo, and piebald. There are many different markings a horse can have. Markings can be on the face or legs. Examples of markings on the face include star, blaze, and snip. A star is a small white patch between the eyes, a blaze is a wide white stripe down the center of the face from the forehead down to the muzzle, a snip is a small white patch on their muzzle. Examples of markings on the legs are sock, stocking, and fetlock. A fetlock is where white hair comes up to their fetlock, which is the lowest joint on their leg. A sock comes halfway up their cannon bone, and a stocking can go from halfway up to above their hock.

Vices, where horses are concerned, are any habit that is unwanted such as cribbing, when a horse swallows air, stall pacing, wood chewing, biting, and kicking to name a few. Vices can de-value a horse. There are different registration organizations for different breeds. Such as APHA, American Paint Horse Association, AQHA, American Quarter Horse Association, and

AMHA, American Morgan Horse Association. If a horse is registered to a specific Association the value of the horse is typically increased. When collecting data on who was selling I only looked at private sellers and barns.

After taking a close look at all of the different factors, I hypothesize that the factors that will have the most significance are breed, height, level of training, what they are trained in or discipline, and color. While I feel that temperament and location will be important I do not feel that those factors will be significant enough to affect the price. To determine if my hypothesis is correct or not I will first need to determine which factors are important and/or significant. In order to do this I need to run ANOVA on them. The assumptions that must be met in order to run ANOVA on your data is that each k population has a normal distribution, the variances of the k populations are equal, and that each k sample must be independent from the others. ANOVA stands for analysis of variance and uses a f-test to determine if the means of many different variations of a variable are significantly different. The f statistic is the ratio of mean squares for treatments, MST, which represents the average weighted squared deviation of the treatment mean from the grand mean, and the mean square errors, MSE.

MST=SST/(k-1), where SST is the sum of squares for treatments
$SST=\Sigma n_i(x_i-x)^2$
MSE=SSE/(N-k)
$SSE=\Sigma(x_{1i}-x_1)^2+\Sigma(x_{2i}-x_2)^2+\ldots\ldots\ldots+\Sigma(x_{ki}-x_{k)}{}^2$
F=MST/MSE

This f value is determined for you when you run ANOVA. For ANOVA you have a predetermined alpha significance level, which you choose, to compare to the p-value. Alpha can also be called the type one error. The type one error is the probability of falsely rejecting the null hypothesis when the null hypothesis is true. The alpha value I used is 0.05, which is a very common alpha value. When performing ANOVA you will obtain a p-value for each variable. A

p-value is the probability of obtaining a test statistic at least as extreme as the one observed. When comparing the p-value to the alpha chosen you reject the null hypothesis, stating that the means are all equal if the p-value is less than alpha. When you reject the null hypothesis you are stating there is a significant difference in means between variations in the variable. If the p-value is greater than the alpha chosen, you do not reject the null hypothesis, meaning there is not a significant difference between the means. The null hypothesis in our case is that the means of each variable are equal, and our alternate hypothesis is that at least one of them is not equal.

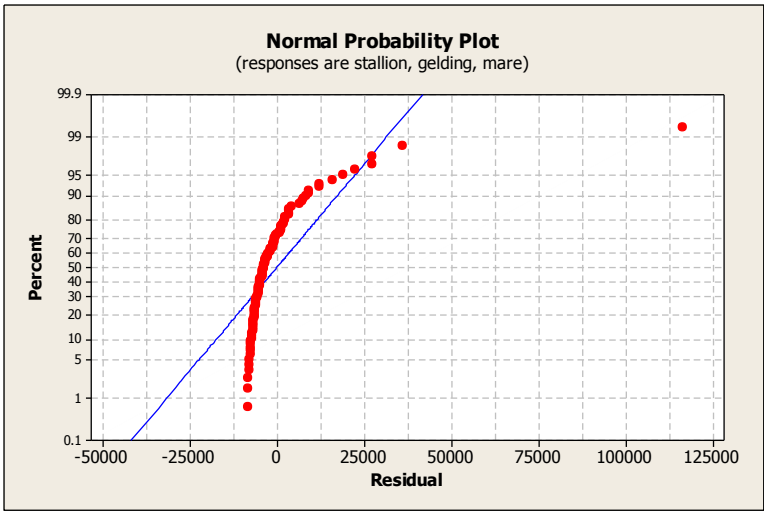
H_0: $\mu_1=\mu_2=\mu_3=\ldots\ldots=\mu_n$
H_a: at least one of the μ_i is different

To run ANOVA on my variables I used Minitab, which is a mathematical computer program, and ran One-Way unstacked ANOVA on each variable. However, before running ANOVA on every variable that I collected I first looked at the data collected for each. For the variables of sold by whom, vices, markings, and registration I did not have enough information on each one so I could not use these variables and had to disregard those variables right away.

Now that I had determined what variables I needed to run through ANOVA I started with sex. Before running ANOVA on sex I first combined colt with gelding and filly with mare. Running ANOVA on mare, gelding, and stallion I obtained the p value of 0.828. The f-statistic associated with this test was 0.19. For the N value, which is the sample size of each variation for the variable, stallion had 7 samples, gelding had the most at 56, and mare had 50 samples. Since the p-value of the f-statistic was significantly larger than 0.05, which was the chosen alpha value, I can determine that sex is not a significant factor when determining price of the horse. To determine if the ANOVA output is relevant one must look at the normality plot of the data to ensure that the data follows a relatively normal pattern. Looking at the normality plot inserted below, one sees that the data is relatively normal.

One-way ANOVA: Sex

```
Source   DF           SS         MS     F      P
Factor    2     70369384   35184692  0.19  0.828
Error   110  20450349643  185912269
Total   112  20520719027

S = 13635   R-Sq = 0.34%   R-Sq(adj) = 0.00%

                           Individual 95% CIs For Mean Based on
                           Pooled StDev
Level      N  Mean  StDev  -------+---------+---------+---------+--
stallion   7  5714   6800  (-----------------*----------------)
gelding   56  7598   7637               (-----*-----)
mare      50  8705  18607                (------*-----)
                           -------+---------+---------+---------+--
                                  0       6000      12000     18000
Pooled StDev = 13635
```

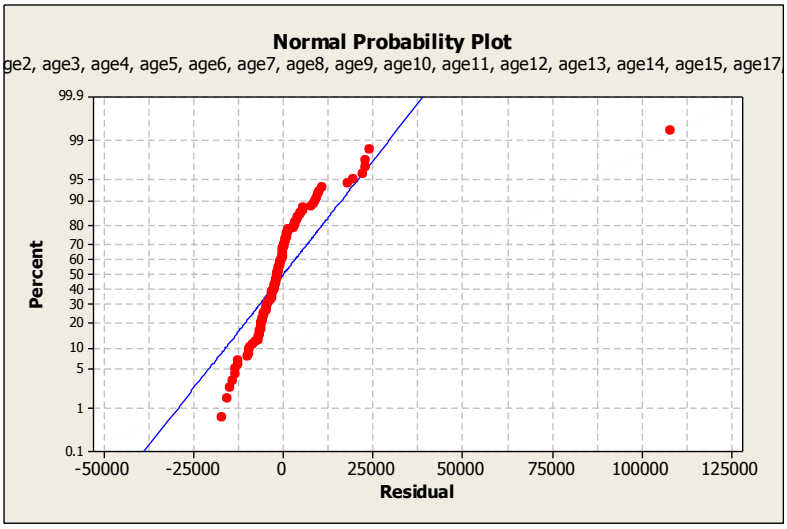


Next looking at age, running ANOVA gave me the following output.

One-way ANOVA: AGE

```
Source   DF           SS         MS     F      P
Factor   18   2448021172  136001176  0.72  0.780
Error    97  18242338721  188065348
Total   115  20690359892

S = 13714   R-Sq = 11.83%   R-Sq(adj) = 0.00%

                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level   N   Mean  StDev  ------+---------+---------+---------+---
age1    7   3193   3962            (-----*-----)
age2    6   4817   4040            (------*------)
age3    9   4033   4401             (-----*----)
age4   13   7119   7356                (---*----)
```

```
age5    11   8518   8124                    (----*----)
age6    13   6731   3358                  (----*----)
age7    11  17186  36164                          (----*----)
age8     8   7825   9692                  (-----*-----)
age9     6  11083  12820                    (------*------)
age10    8   5838   5018                (-----*-----)
age11    6   4167   2066              (------*------)
age12    6  12042  12387                     (------*-----)
age13    3  22167  20593                        (---------*---------)
age14    1   1500      *   (----------------*----------------)
age15    4   4013   5676             (--------*-------)
age17    1   4700      *     (----------------*----------------)
age18    1   1800      *   (----------------*----------------)
age19    1   1500      *   (----------------*----------------)
age20    1   1500      *   (----------------*----------------)
                           ------+---------+---------+---------+---
                            -16000          0      16000     32000

Pooled StDev = 13714
```



When looking at the output there are a wide range of sample sizes for each age, with the most being 13 for ages 4 and 6, and the least being 1 for ages 14, 17, 18, 19, and 20. Certain ages are not included in the output such as ages 16, and any age over 20, because there was no data for these specific age ranges. The f-statistic for this ANOVA is 0.72 with p-value of 0.78, which is well above 0.05. Before we can disregard age as a significant variable I have to look at the normality plot. Inserted above one can see that the data is relatively normal, showing that the results are relevant. To see if grouping the ages more reasonably would produce a better result I

decided to group the ages in ranges. The ranges included under 5 years old, 5 to 10, 10 to 14, 15 to 19, and over 20 years of age. Running ANOVA on this data I obtained the following output.

One-way ANOVA: Range of Age in Increments of 5

```
Source   DF           SS         MS     F      P
Factor    4    723937034  180984258  1.01  0.408
Error   111  19966422858  179877683
Total   115  20690359892

S = 13412   R-Sq = 3.50%   R-Sq(adj) = 0.02%


                              Individual 95% CIs For Mean Based on
                              Pooled StDev
Level        N   Mean  StDev  -------+---------+---------+---------+--
under 5     35   5146   5608                  (--*--)
ages 5-9    49  10191  18330                      (--*-)
ages 10-14  24   8831  10763                    (---*---)
ages 15-19   7   3436   4203               (-----*------)
over 20      1   1500      *  (-----------------*-----------------)
                              -------+---------+---------+---------+--
                                -15000         0     15000     30000

Pooled StDev = 13412
```

Looking at the output I noticed the p-value decreased significantly going from 0.78 down to 0.408. Noticing this I concluded that when grouped, the ages became more significant. Using this observation I grouped the ages differently once more, using two different groups, only ages 1 to 9 and 10 and up. The ANOVA output consisted of the f-statistic of 0.06 with the p-value of 0.812, which increased to above 0.05 also increasing above the p-value I first obtained when running ANOVA on ages. Studying this information from the three ANOVA outputs I can confidently conclude that age is not a factor that influences the value of a horse.

The next factor I looked at was breed, I had a large number of different breeds consisting of 24 breeds, and even though this is a large number of variations I did not change anything before running the first ANOVA.

One-way ANOVA: Breed

```
Source   DF           SS         MS     F      P
Factor   23   5072198446  220530367  1.44  0.115
```

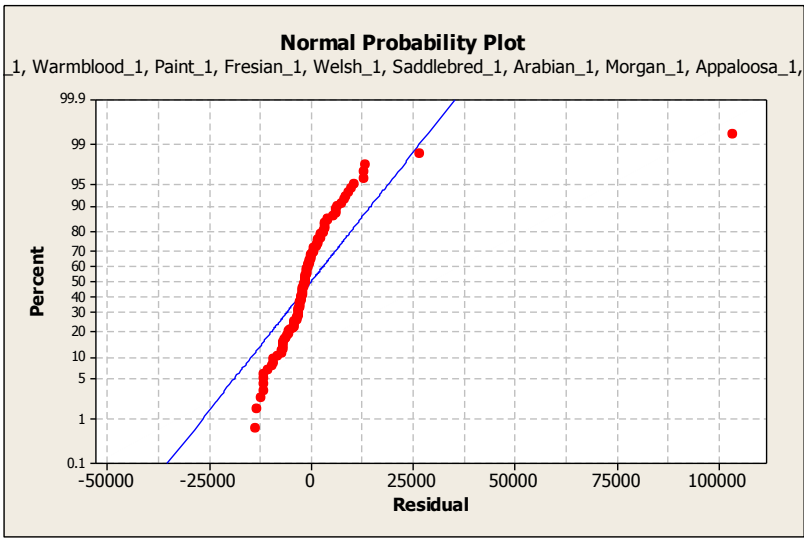
```
Error    92  14110353515  153373408
Total   115  19182551961

S = 12384   R-Sq = 26.44%   R-Sq(adj) = 8.05%

Level          N   Mean  StDev
Nokota         1  18000      *
Oldenburg      2  20250  10960
Quarter Horse 24   4273   3537
Thoroughbred  15   6267   4741
Saddlebred     2   4400   3677
Warmblood     12  23250  33241
Mini           1   1100      *
Pinto          4   4375   2136
Arabian        7   3179   1760
Fresian        5   5960   6126
Paint         15   3693   2796
Morgan         4   3050   1542
Appaloosa      5   2930   2219
Holsteiner     1  12500      *
Gypsy          2   2850   3748
Irish Sport    3  11800  11435
Lipizzaner       1   3500      *
Tennessee      4   4625   3750
Welsh          3  16333  17076
Hanoverian     1  15000      *
Trakehner      1  20000      *
Tocky Mtn      1   7500      *
Andalusian     1  17000      *
Shetland       1   1500      *

              Individual 95% CIs For Mean Based on Pooled StDev
Level            --+---------+---------+---------+-------
Nokota                    (-----------*-----------)
Oldenburg                     (--------*--------)
QH                          (-*--)
Thoroughbred                (--*--)
Saddlebred              (-------*--------)
Warmblood                           (---*--)
Mini              (------------*-----------)
Pinto                    (-----*-----)
Arabian                   (----*---)
Fresian                   (-----*----)
Paint                       (--*--)
Morgan                  (------*-----)
Appaloosa                 (----*-----)
Holsteiner              (-----------*------------)
Gypsy                  (-------*--------)
Irish Sport                   (------*------)
Lipizzaner          (-------------*-----------)
Tennessee                (-----*-----)
Welsh                         (------*------)
Hanoverian              (-------------*-----------)
Trakehner                  (-----------*-----------)
Rocky Mtn            (-------------*-----------)
Andalusian               (-------------*-----------)
Shetland           (-------------*-----------)
                   --+---------+---------+---------+-------
                -20000         0     20000     40000
Pooled StDev = 12384
```

The sample sizes for the different breeds had a large range, ranging from 1-24. The f-statistic for the test is 1.44 with p-value of 0.115. Since the p-value is only slightly larger than 0.05 I hypothesized that breed had a good chance of being significant if grouped more reasonably. Before running ANOVA again I combined like breeds together. Like breeds consisted of the sub breeds of Warmbloods, such as Oldenburg, Nokota, Hanoverian, Holsteiner, and Trakehner, or pony breeds such as Shetland, Gypsy, and Mini being combined with Welsh. I also combined Pinto with Paint, since they are both colored breeds. I also combined the Andalusians with the Fresians, and combined Rocky Mountain Horse with the Quarter Horses, because they are most alike. Combining the breeds together to get a smaller number of variations equaling 13 breeds, allowed me to get a more accurate output when I ran ANOVA, with the following result:

One-way ANOVA: Breed

```
Source   DF           SS          MS     F      P
Factor   12   4268515944  355709662  2.44  0.008
Error   103  15001648948  145647077
Total   115  19270164892

S = 12068   R-Sq = 22.15%   R-Sq(adj) = 13.08%


Level            N   Mean  StDev
QH_1            25   4822   4418
Thoroughbred_1  15   6267   4741
Warmblood_1     17  21588  27884
Paint_1         19   3837   2631
Fresian_1        6   7800   7095
Welsh_1          7   8186  12571
Saddlebred_1     2   4400   3677
Arabian_1        7   3179   1760
Morgan_1         4   3050   1542
Appaloosa_1      5   2930   2219
Irish Sport_1    3  11800  11435
Tennessee_1      4   4625   3750
Lipizzaner       2  10250   9546

                Individual 95% CIs For Mean Based on Pooled StDev
Level             ---+---------+---------+---------+------
QH_1                        (----*----)
Thoroughbred_1              (-----*-----)
Warmblood_1                                   (-----*----)
Paint_1                   (-----*----)
Fresian_1                 (---------*---------)
```

```
Welsh_1                          (--------*--------)
Saddlebred_1          (-----------------*-----------------)
Arabian_1                    (--------*--------)
Morgan_1                  (-----------*-----------)
Appaloosa_1               (----------*----------)
Irish Sport_1                      (-------------*-------------)
Tennessee_1               (----------*----------)
Lipizzaner                (----------------*----------------)
                  ---+---------+---------+---------+------
                -10000         0      10000     20000

Pooled StDev = 12068
```



The p-value for the f-statistic when grouping the breeds dropped significantly below 0.05 to a value of 0.008. Before determining that this value was accurate I took a look at the normality plot, inserted above. The data is relatively normal, which confidently showed that breed is a significant factor when determining price of a horse. After seeing that breed is a significant factor I took a look at the discipline the horses were trained in. Running ANOVA I got the following output.

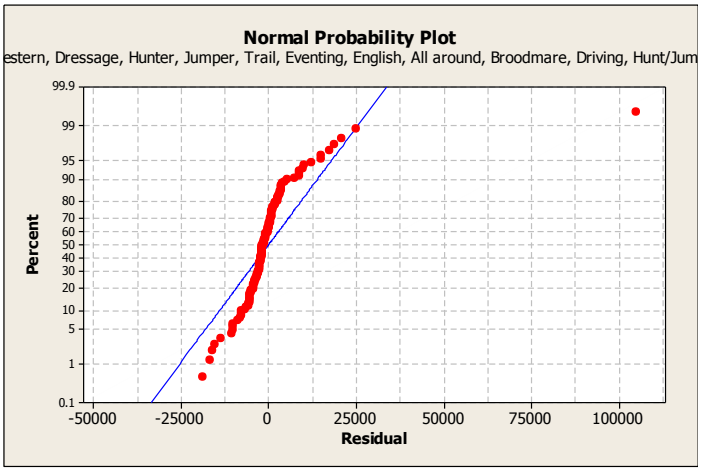
One-way ANOVA: Trained In

Source	DF	SS	MS	F	P
Factor	14	4173265817	298090416	2.27	0.008
Error	128	16807763728	131310654		
Total	142	20981029545			

```
S = 11459   R-Sq = 19.89%   R-Sq(adj) = 11.13%

                              Individual 95% CIs For Mean Based on
```

```
                                  Pooled StDev
Level          N    Mean   StDev  --------+---------+---------+---------+-
Reiner         3    3600    1637            (----------*----------)
Western       21    3655    2945                   (---*---)
Dressage      18    7139    7073                     (----*---)
Hunter        19   20076   27948                                (----*---)
Jumper        10   12350    7546                        (-----*-----)
Trail         19    3637    2119                   (---*---)
Eventing       4   11250    5895                    (--------*---------)
English       15    6040    5994                    (----*----)
All around     9    5828    5177                   (-----*-----)
Broodmare      2    4000     707           (------------*-------------)
Driving        5    6340    5876                  (-------*--------)
Hunt/Jump      4    9000    3851                   (---------*--------)
Racing         2    6500    7778            (------------*-------------)
Combo         11    4373    2460                  (-----*----)
Youth          1    1100       *   (-------------------*-------------------)
                                  --------+---------+---------+---------+-
                                     -12000          0      12000     24000

Pooled StDev = 11459
```

All together I had 15 different disciplines with sample sizes for each ranging from 21 in Western to only 1 for youth. The f statistic value for the ANOVA test was 2.27 with a p-value of 0.008. This is one of the factors that was significant right away. Before fully stating this factor is significant I had to look at the normality plot, inserted below. Looking at the plot the data is relatively normal showing that the f statistic and p-value are accurate.



Looking at the temperament data, the p-value was significantly larger than 0.05 with a f statistic of 0.44. The data included only temperaments from 0 to 7; with the largest sample size

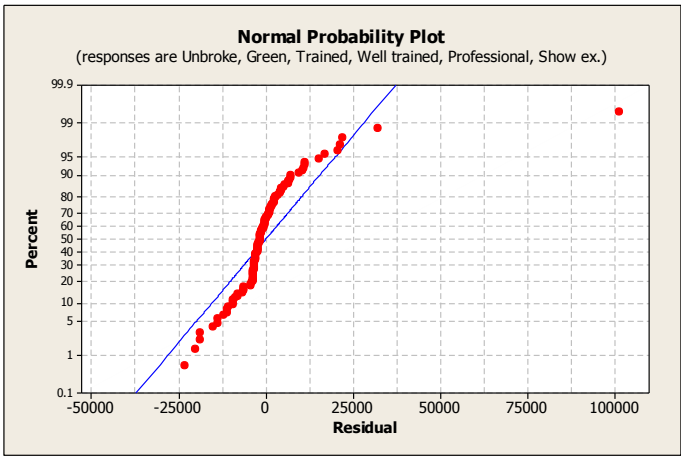
collected for a temperament of 3. Analyzing the results from ANOVA for temperament it is concluded that temperament is not a significant factor when determining the price of a horse.

Level of training was also significant without having to change the grouping of the data.

One-way ANOVA: Level of Training

```
Source   DF           SS          MS     F      P
Factor    5   5453759040  1090751808  7.22  0.000
Error   119  17969043960   151000369
Total   124  23422803000

S = 12288   R-Sq = 23.28%   R-Sq(adj) = 20.06%
```

```
                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level           N   Mean  StDev   ----+---------+---------+---------+-----
Unbroke        12   2983   3222   (------*------)
Green          29   4398   5194       (---*----)
Trained        10   5080   3603    (--------*--------)
Well trained   29   5103   3480       (---*----)
Professional   29  12805  10559                (----*---)
Show ex.       16  23563  30024                        (------*-----)
                                  ----+---------+---------+---------+-----
                                      0     10000     20000     30000

Pooled StDev = 12288
```



The f-statistic for the ANOVA is 7.22 with a p-value of 0 significantly below 0.05. I had 6 different categories ranging from unbroken up to show experience. Green, well trained, and professionally trained all had the same sample size of 29. I was pleased to see that my hypothesis was correct when I theorized that level of training would be a significant factor in

determining value. Looking at the normality plot the data is relatively normal so the p-value and f statistic can be said to be a valued number.

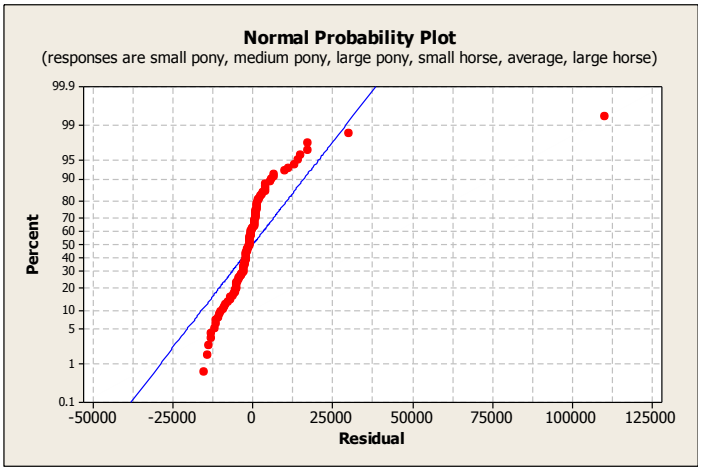
When collecting the data for physical location of a horse I collected the city and state where they were being sold. For ANOVA I changed the locations to include only the state which proved to be an easier grouping mechanism. Running ANOVA on the locations grouped by state I got a f statistic value of 0.68 with a p-value of 0.898 which is well above 0.05. Since grouping by state entailed 38 different groups I decided to group locations together by region. I believed that if I could get a smaller number of different groups it would change the significance of location. I grouped the states into the regions of Northeast, South, Midwest, and West. Running ANOVA on the regions grouped this way decreased the p-value to 0.743 reducing it by less than 0.1. Looking at the output and normality plot I can conclude that location is not a significant factor in determining value.

In looking at height I grouped the data by the groups mini, pony and horse. A mini is a horse that is less than 12.2hh, a pony is 12.2hh to 14.1hh and a horse is 14.2 above. Running ANOVA on the groups mini, pony, and horse gave the p-value of 0.876 for the f statistic of 0.13. Seeing that the p-value was so large I tried to group the heights in a more reasonable way to see if I could get a better p-value. In considering ponies and horses, the subcategories for height include small pony, medium, and large pony and small, medium, and large horse; I decided to group the heights in these 6 categories. Small pony was 12.2hh and under, medium is 12.3hh – 13.2hh, large is from 13.3-14.1hh. Small horse is 14.2 – 15hh, medium 15.1 – 15.3hh, which is where the average height of a horse lies at 15.2hh, large horse is 16hh and up. The ANOVA output for the 6 categories of heights gave the output of 3.84 for the f statistic with a p-value of 0.003. This p-value is significantly lower than 0.05 and a great improvement. Looking at the

normality plot we see that the data is relatively normal concluding that height is a significant value.

One-way ANOVA: Height

```
Source    DF           SS          MS      F      P
Factor     5   3100143829   620028766   3.84  0.003
Error    107  17294411481   161630014
Total    112  20394555310

S = 12713   R-Sq = 15.20%   R-Sq(adj) = 11.24%

Level          N   Mean  StDev
small pony     2   6800   8061
medium pony    5  17700  16266
large pony    15   2473   2224
small horse   20   2975   1443
average       39   5844   4383
large horse   32  14806  22239

              Individual 95% CIs For Mean Based on Pooled StDev
Level            -+---------+---------+---------+--------
small pony       (-----------------*-----------------)
medium pony                        (-----------*----------)
large pony              (-----*------)
small horse              (-----*-----)
average                       (---*---)
large horse                           (----*---)
                 -+---------+---------+---------+--------
               -10000         0     10000     20000

Pooled StDev = 12713
```



The last variable I tested was color. I had a wide range of colors with 14 different colors in total. Leaving the colors separated the way I collected them I ran ANOVA and got a very large p-value of 0.844 for a f statistic value of 0.61. Since the number of different colors was so large, at 14, I tried grouping the colors with other like colors to narrow down the number of

colors to only 9. Running ANOVA I got the p-value of 0.714 which was lower than the p-value that was collected before but still larger than 0.05 concluding that the color of a horse is not a factor that effects the price of a horse.

Overall the variables that I found significant were height, breed, training discipline and level of training. In order to make a regression model with these variables I first had to come up with a ranking system that ranked the variables based on their effect on pricing. A regression model is a model with multiple 'x' variables that will predict the 'y', namely the price. When running the regression each variable is assigned a coefficient that can either be negative or positive. An intercept coefficient also is created; this coefficient does not have an 'x' associated with it. When looking at the regression output it is important to check the significance of the overall regression model. In order to check this, the p-value of the f statistic of the model must also be below the alpha value of 0.05. In order to make sure the regression model is accurate you also have to make sure the overall model has a normal distribution which is determined by looking at the normal probability plot.

In order to rank the values I had to determine how the pricing of each different breed ranked with the other breeds, how the different disciplines ranked compared to eachother, and how training level would influence the pricing. Height is already a quantitative variable, so I left the height values as is when running the first regression. Ranking training level of a horse was an easy task. Again training level can be broken down into five subcategories; unbroke, green, well trained, professional trained, and showing experience. Unbroke is the least trained, if trained at all and professional and show experience are the highest, which meant that unbroke would rank the lowest with show experience and professional training ranking the highest. Unbroke was ranked at number 1, increasing to 5 with show experience. Breed and training

discipline were harder to rank because the price points are so close together, practically ranking at the same level. When ranking the disciplines I used the data I collected as a reference. By looking at the average value of each discipline I was able to determine if the rank assigned was accurate. The top ranking discipline, or most expensive was determined to be jumper, with eventing following close behind. I had 14 different disciplines to rank which caused jumper to be ranked at 14 and eventing to be ranked at 13. Racing is a popular sport and horses that are used for racing are typically priced very high, causing racing to be ranked as the next highest discipline, at 12. By checking the averages I was able to determine that dressage is the next most expensive discipline which caused it to be ranked at 11. Well trained and upper division level hunter and jumper can go for a very high price, making hunter/jumper the next ranked discipline at 10. I ranked hunter alone at number 9. To determine the discipline that ranked next I had to refer to the average values of the data that I collected. Looking at these averages I concluded that reiner was the next ranked discipline at number 8. After reiner the ranking of disciplines became very complicated. The rest could all be closely priced making the ranking very difficult. However, by comparing the averages of the data for the disciplines left I was able to determine that driving was next, ranked at 7, and "all around" was next putting it at 6. English and western are very close disciplines, determining their ranking was difficult even after referring to the averages; the averages differed by less than a hundred dollars. I decided to rank English at 5 and western at 4. The only disciplines left to rank were trail, youth, and broodmare. I ranked trail at 3, youth at 2 and broodmare at 1, because they would be the least expensive disciplines.

After determining the ranking of disciplines I had to determine how the different breeds would rank compared to the others. When I ran the ANOVA for breed I had 13 different breeds, when ranking them I decreased it to 12 because Lipizzaner had only one data sample to base the

result on. When ranking breed I compared the breed to what discipline they are primarily used in to help me determine the best ranking. Warmblood and Sport Horse are most commonly used for jumping and eventing, so I ranked them highest, Warmblood at 12 and Sport Horse at 11. Racing was ranked directly below jumping and eventing, Thoroughbreds are the most common horse used for horse racing causing them to be ranked at 10. To rank the rest of the breeds I used the averages of the data collected to get a better idea of how they should be ranked. Looking at the averages I noticed that Welsh ponies could be very highly priced. Welsh ponies are a very popular and well known pony breed, which can make their pricing high. Therefore I ranked Welsh next at 9. Quarter Horse and Paint were the next highest priced but their pricing values were very close together, however Quarter Horse was slightly higher than Paint so I ranked Quarter Horse at 8, and Paint at 7. Fresian's were the next highest average, but before ranking I decided to combine Lipizzaner with Fresian. I ranked Fresian at 6 and combined Lipizzaner with them. The last 5 breeds Saddlebred, Tennessee Walker, Arabian, Morgan, and Appaloosa, were all very close with each only being slightly higher than the next. With that said, Tennessee Walker ranked at 5, Saddlebred ranked at 4, with Arabians at 3, Morgan at 2, and Appaloosa ranking last at 1.

The first regression was run using the 12 ranked breeds,13 different disciplines, 5 ranks of training levels, and the height values left as their numerical values. Even though this was a large amount of data I ran the regression to see what results I would get. Excel has an add-in tool for data analysis that can create a regression when the data is inputted. Running the regression I was able to get an output with the overall significance below 0.05. Looking at the insert below the f-statistic for the regression is 9.072804 with the p-value of 0.00002126 which is well below 0.05. However, when looking at the coefficients and their p-values which will also

be used to determine their significance, two of the coefficients were not significant. The coefficient for height and discipline were not found to be significant.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.491399
R Square	0.241473
Adjusted R Square	0.214858
Standard Error	11737.21
Observations	119

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	5E+09	1.25E+09	9.072804	2.126E-06
Residual	114	1.57E+10	1.38E+08		
Total	118	2.07E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-26738.7	13729.07	-1.9476	0.053922	-53935.84485	458.4993325
breed	995.1238	416.0306	2.391948	0.018395	170.9703045	1819.277224
height	977.8037	925.8822	1.056078	0.293166	-856.3616326	2811.969101
discipline	71.00034	350.786	0.202404	0.839962	-623.9039925	765.9046738
Training level	3822.228	1018.745	3.751899	0.000278	1804.102002	5840.353709

After looking at the values and how some were not significant for the regression I decided to run the regression another time changing how I ranked breed, discipline, and height. Even though breed was found significant I decided that there were too many different ranks and decided to combine some together. When looking at the different breeds again I decided to combine like breeds and breeds that were closely priced or used for the same discipline. When ranking using these criterion I was able to narrow down the number of ranked breeds to 6. I combined Sport Horse and Warmblood together ranking them the highest with the value 6. Thoroughbred was left alone and not combined with any other breed and kept at the second

highest rank at value 5. Welsh was also not combined with any other breed because they were different than the rest and I kept them at the third highest rank being value 4. I determined that Quarter Horse, Paint, and Appaloosa could be combined into one group because they are like breeds often used for the same disciplines; this put them at rank number 3. I ranked Fresian at number 2. Lastly I grouped Saddlebred, Arabian, Morgan, and Tennessee Walker at the lowest level because they are breeds that are not very popular and overall had the lowest averages when compared to other breeds. I decided to rank what the horse is trained in, or the different disciplines by grouping them with other like disciplines and close pricing. By ranking under these criteria I was able to get the number of ranked disciplines down from 14 to 7. Ranked at the highest at number 7 I combined eventing, jumper, and dressage. These disciplines can have the highest priced horses which is why I ranked them at the top. I also decided to group these three together because they are all related, eventing combines jumping, dressage, and cross country into a three day event. Again racing is the next highest discipline ranking it at number 6. Since I had combined dressage with eventing and jumping the next discipline to follow racing was hunter which was ranked at 5. I decided to keep hunter separate from jumper so I no longer had the hunter/jumper discipline to rank. Before changing the criteria of how I ranked the disciplines reiner was ranked right below hunter which is why I ranked reiner combined with "all around" horse at number 4. Reiner and "all around" are closely related disciplines in Western riding. Driving was still ranked right below reiner which gave driving the new rank of number 3. I combined Western and English together because they are so closely priced and equally as popular. Combining these disciplines gave them the rank of 2. Lastly I combined the lowest priced disciplines, youth, trail, and broodmare, giving them the value of 1.

When running ANOVA on the height of the horses in order for the variable to be significant I had to rank the heights based on the subcategories of small pony, medium pony, large pony, small horse, medium horse, and large horse. Since height was found significant when combining them in these subgroups I ranked them by using these subcategories. However, since height is a quantitative variable I decided to take the average of the ranges of different height categories and rank them that way. In doing so small ponies average height was 12.2, medium pony 13.1, large pony 14, small horse 14.3, medium horse 15.2, and large horse 16.2. To rank the heights using the averages I took the values in the ranges and assigned the appropriate average value.

Creating a regression with the newly ranked variables proved to be not much better. Even though the overall model was significant with a p-value of 0.000000747 for the f statistic of 9.779, height and discipline's coefficients were still not significant and now the intercept coefficient was not significant. The p value for the intercept is 0.171198 which is close to 0.05 which leads me to believe that if I change the variables again I will get the intercept coefficient to be significant. However the coefficient for height's p-value is well above 0.05 at 0.488495 and discipline's is equally as high at 0.770456. I noted when comparing this regression to the previous one that the p-value of the overall regression decreased signifying that this is a better model than the first. Trying to improve this model I decided to rank the height in values from 1 to 6.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.505451
R Square	0.25548
Adjusted R Square	0.229357
Standard Error	11628.33

Observations	119

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	5.29E+09	1.32E+09	9.779713	7.74E-07
Residual	114	1.54E+10	1.35E+08		
Total	118	2.07E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-24422.5	17735.5	-1.37704	0.171198	-59556.4	10711.41
height	865.7599	1245.771	0.694959	0.488495	-1602.1	3333.622
breed	2308.659	797.6109	2.894468	0.004552	728.5984	3888.721
discipline	-178.449	610.1367	-0.29247	0.770456	-1387.12	1030.227
Training level	3882.062	1011.983	3.836096	0.000205	1877.332	5886.791

I decided to rank based solely on height, with the smallest height being ranked lowest and the tallest height being ranked the highest. Assigning the averages that I had previously determined to their appropriate ranking and running regression gave me relatively the same result. However, the p-value for the overall model again went down. The p-value for this model was 0.000000828. With the p-value being that low I can determine with confidence that this model is a good model. However, once again height's and discipline's coefficient values were not significant, with the p-values being relatively the same as they were before.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.504540015
R Square	0.254560627
Adjusted R Square	0.22840486
Standard Error	11635.5118
Observations	119

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*

Regression	4	5.27E+09	1.32E+09	9.732485	8.28158E-07
Residual	114	1.54E+10	1.35E+08		
Total	118	2.07E+10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-14408.8369	4948.252	-2.9119	0.004323	-24211.2856	-4606.388281
height	579.1064058	990.6505	0.584572	0.559991	-1383.36455	2541.577358
breed	2349.700724	790.7205	2.971595	0.003614	783.289546	3916.111901
discipline	-155.088086	608.0007	-0.25508	0.799122	-1359.53267	1049.356497
how well	3891.517821	1012.576	3.843185	0.0002	1885.612372	5897.42327

The final regression model was the best model created. I decided to run the regression in Minitab because Minitab enables me to create confidence and prediction intervals from that data, without having to calculate the formula by hand. Running the regression in Minitab gave me slightly different values for the regression equation. The regression equation I am going to use is:

Price = - 14409 + 579 height + 2350 breed - 155 discipline + 3892 training level.

The coefficients for height and discipline are still not significant but we will still consider them. The overall significance of the model is very good with p-value equaling 0. The output for Minitab is inserted below and one can see the regression equation and p-values for each coefficient and the overall model.

Regression Analysis: price versus height, breed, discipline, training level

```
The regression equation is
Price = - 14409 + 579 height + 2350 breed - 155 discipline + 3892 training level


Predictor     Coef  SE Coef      T      P
Constant    -14409     4948  -2.91  0.004
height         579      991   0.58  0.560
breed         2350      791   2.97  0.004
discipline    -155      608  -0.26  0.799
training le   3892     1013   3.84  0.000


S = 11635.5   R-Sq = 25.5%   R-Sq(adj) = 22.8%


Analysis of Variance
```

Source	DF	SS	MS	F	P
Regression	4	5270535433	1317633858	9.73	0.000
Residual Error	114	15433905366	135385135		
Total	118	20704440798			

In Minitab you can activate the option to create prediction intervals when you run the regression model on the data. Doing so, uses the equation found to create prediction values for data values you have already collected. For predictions it is important to come up with prediction intervals. I used a 95% prediction interval for both the coefficients for the regression model and for the price predictions. A 95% prediction interval implies that there is a 95% confidence that the prediction value will be within those values. For multiple regressions the prediction interval formula contains matrix algebra making it more complicated than a simple regression prediction interval. A 100(1-α)% prediction interval for Y when $x_1=x_1^*, x_2=x_2^*, \ldots\ldots x_k=x_k^*$ is the following:

$a'\beta \pm t_{\alpha/2}\, S(1+a'(\mathbf{X}'\mathbf{X})^{-1}a)^{(1/2)}$ where $a'=[1,x_1^*,x_2^*,\ldots\ldots.x_k^*]$

For our purposes, a′ are the values that we have collected for each variable with a column of 1's before each set of values. X is the matrix of variables with an additional column of 1's inserted for the first row. X′ is this matrix of variables transposed which is then multiplied by the matrix X. The matrix you get when you multiple these two matrices together is then inverted to get you the value of $(X'X)^{-1}$. The value $\beta = (X'X)^{-1}X'Y$. To determine what $t_{\alpha/2}$ should be the α value you have issued is divided by 2, I have decided to choose the value of 0.05, which when divided by 2 gives me 0.025, and used a T value table to find the appropriate t value, our t value is equal to 1.96. Some examples of the prediction values and prediction intervals created by Minitab are inserted below. Again the values used to predict the predicted prices are the observations that were observed and plugged into the regression equation obtained. The observation 1- 10 corresponds to the x values obtained in my data. When studying these prediction values, or "Fit"

as signified in the insert, I noticed that the values are not necessarily what I collected to be the actual price values for the associated x values. Also noted is that the prediction intervals have a very wide range including negative numbers. When talking about price, however, we can conclude that the price of a horse will never be at 0 or below.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	9902	1738	(6460, 13345)	(-13403, 33208)
2	17644	2221	(13245, 22044)	(-5822, 41110)
3	6476	1837	(2838, 10115)	(-16859, 29811)
4	11558	1853	(7888, 15229)	(-11782, 34899)
5	1777	2302	(-2783, 6337)	(-21720, 25273)
6	6476	1837	(2838, 10115)	(-16859, 29811)
7	6321	1459	(3431, 9211)	(-16909, 29551)
8	17644	2221	(13245, 22044)	(-5822, 41110)
9	8858	3212	(2496, 15220)	(-15054, 32770)
10	6900	1585	(3760, 10040)	(-16363, 30163)

Overall it can be concluded that the only factors that predict the price of a horse are height, breed, discipline they are trained in, and how well they are trained in said discipline. This is slightly different than what I hypothesized as I predicted that color would also influence the price. I am pleased to see that my hypothesis was mostly true and that what I assumed influenced price does indeed predict the price. By obtaining the regression model I can use the data that I have collected and predict what an appropriate selling price for a horse should be. This would be very beneficial in the horse world as it would enable one to set a reasonable price for that particular horse and would allow buyers to determine if the price is appropriate and understand the reason the horse is priced as it is.

Bibliography

"Buy. Sell. Learn. Horses!" *Eneigh.com*. eNeigh.com Inc, 2007-2013. Web. http://www.eneigh.com/. July 2012.

"Craiglist." *Craig's List Classifieds*. 2013. Web. http://www.craigslist.org/about/sites/. July 2012.

"Discovering Statistics [Hardcover]." *Discovering Statistics: James J. Hawkes, William H. Marsh: 9780918091864: Amazon.com: Books*. Quant Systems Inc; 2nd Edition. Pub. April 30, 2004. Print.

Dreamhorse.com. Dream Horse Classifieds, LLC, 1998. Web. http://www.dreamhorse.com/. July 2012.

Equinehits.com. Equinehits Horse Classifieds- Horses for Sale, 2002-2007. Web. http://www.equinehits.com/. July 2012.

Equispot. Equispot, LLC, 2011-2013. Web. http://www.equispot.com/. July 2012.

Horse Clicks. 2013. Web. http://www.horseclicks.com. July 2012.

"Horses for Sale." *BigEq.com First in Hunter/Jumper Sales Online*. Bigeq.com Online Horse Classifieds, 1999-2013. Web. http://www.bigeq.com/. July 2012.

"Horses for Sale." *BuyHorses.com*. 2013. Web. http://buyhorses.com/. July 2012.

"Horses for Sale." *Equine Now*. Equinenow Classifieds, 2005-2013. Web. http://www.equinenow.com. July 2012.

"Horse for Sale, Horse Classifieds, Pictures, and Horse Trailers." *Equine.com*. Cruz Bay Publishing, Inc, 201. Web. http://www.equine.com/index.html. July 2012.

"Horses for Sale." *Horsefinders.com*. 2005-2013. Web. http://horsefinders.com/. July 2012.

"Horses For Sale." *Horsetopic.com Classifieds, Horses for Sale*. Horsetopia LLC, 2002-2009. Web. http://www.horsetopia.com/. July 2012.

"Horses for Sale." *RanchWorldAds Horses for Sale- Ranch Classifieds.* 2005-2012. Web. http://www.ranchworldads.com/. July 2012.

"Horses for Sale." *The New Horsetrader.com*. California Horsetrader, Inc, 2013. Web. http://www.horsetrader.com/. July 2012.

Horseville. Horseville, LLC, 2001-2013. Web. http://horseville.com/. July 2012.

Kaps, Miroslav, and William R. Lamberson. *Biostatistics for Animal Science*. Wallingford, Oxfordshire: CABI Pub., 2004. Print.

MyHorseForSale.com. Front2Back Studio. Web. http://myhorseforsale.com/. July 2012.

Wackerly, Dennis D., William Mendenhall, and Richard L. Scheaffer. *Mathematical Statistics and Probability.* Andover: Brooks/Cole Cengage Learning, 2010. Print.