COSC 757 Data Mining Assignment 1

Mary Snyder
Department of Computer & Information Sciences
College of Science & Mathematics
Towson University
msnyde8@students.towson.edu

ABSTRACT
In this paper, I will be exploring a dataset to become more familiar with exploratory data analysis, data preprocessing, and regression analysis through the COSC 757 Data Mining Assignment 1.

Categories and Subject Descriptors
H.2.8 **[Database Management]** Database Applications – *Data mining*

Keywords
Regression Dataset; Continuous; Multi-Valued Discrete; Miles Per Gallon (MPG); Exploratory Data Analysis; Data Minimum; Data Maximum; Data Mean; Data Median; Standard Deviation; Data Symmetry; Distribution of Attributes; Histogram; Transformations; Scatterplot; Quadratic Decay; Scatterplot Matrix; Data Preprocessing; Normalization; Min-Max Normalization; Z-score Standardization; Decimal Scaling; Binning; K-means; Equal Width; Transformations to Achieve Normality; Skewness; Normal Probability Plot; Regression Analysis;

1. INTRODUCTION
I chose a dataset from the UCI Machine Learning Repository classified for the task of Regression Datasets. This dataset is maintained at Carnegie Mellon University as part of the StatLib library. The dataset contains information regarding auto-mpg data. There are 398 instances with 9 attributes: mpg (continuous), cylinders (multi-valued discrete), displacement (continuous), horsepower (continuous), weight (continuous), acceleration (continuous), model year (multi-valued discrete), origin (multi-valued discrete), and car name (string – unique for each instance). There are missing values for the mpg and horsepower field values in 6 of the instances.

2. EXPLORATORY DATA ANALYSIS
Included below in Table 1, is a summary of the variables for the dataset including the minimum, maximum, mean, median, and standard deviation for each the field values. The mean and median of acceleration are extremely close to each other (median of 15.50 and mean of 15.52), which is an indicator of possible symmetry. By the same token, the mean and median of displacement (median of 151 and mean of 194.8), horsepower (median 95 and mean 105.08), and weight (median 2822 and mean 2979) are fairly far apart from each other indicating they are not symmetric.

Table 1. Summary of Auto-Mpg Data

```
      mpg          cylinders      displacement
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0
 1st Qu.:17.50   1st Qu.:4.000   1st Qu.:105.0
 Median :23.00   Median :4.000   Median :151.0
 Mean   :23.51   Mean   :5.475   Mean   :194.8
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:302.0
 Max.   :46.60   Max.   :8.000   Max.   :455.0
 NA's   :8
   horsepower       weight      acceleration
 Min.   : 46.00   Min.   :1613   Min.   : 8.00
 1st Qu.: 75.75   1st Qu.:2226   1st Qu.:13.70
 Median : 95.00   Median :2822   Median :15.50
 Mean   :105.08   Mean   :2979   Mean   :15.52
 3rd Qu.:130.00   3rd Qu.:3618   3rd Qu.:17.18
 Max.   :230.00   Max.   :5140   Max.   :24.80
 NA's   :6
   model_year        origin                car_name
 Min.   :70.00   Min.   :1.000   ford pinto     :  6
 1st Qu.:73.00   1st Qu.:1.000   amc matador    :  5
 Median :76.00   Median :1.000   ford maverick  :  5
 Mean   :75.92   Mean   :1.569   toyota corolla:  5
 3rd Qu.:79.00   3rd Qu.:2.000   amc gremlin    :  4
 Max.   :82.00   Max.   :3.000   amc hornet     :  4
                                 (Other)        :377
```

2.1 Distribution of Attributes
Even if one was not aware beforehand, by looking at the histograms for cylinders (Figure 2), model year (Figure 7), and origin (Figure 8), one can tell that these variables are multi-valued, but discrete. There are only 5 different values for cylinders, the date range for model year is restricted to between 1970 (70) and 1982 (82), and the origin is one of 3 values. Another interesting distribution is the acceleration (Figure 6) values. These values seem to take on a nice bell curve without any data processing/transformations. The other remaining histograms for MPG (Figure 1), displacement (Figure 3), horsepower (Figure 4), and weight (Figure 5) all seem to have a right-skewed distribution. No histogram was completed for car name since it was specified as a unique identifier.

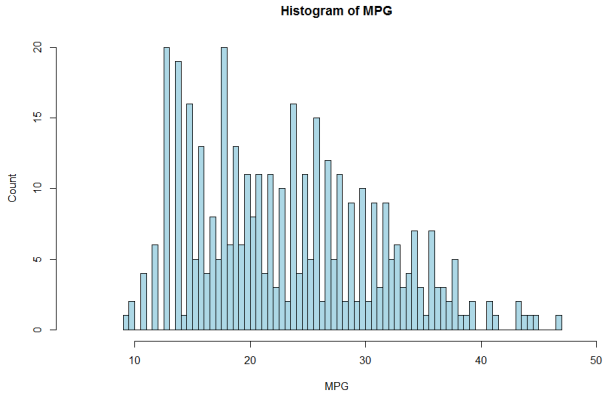

Figure 1. Histogram of MPG

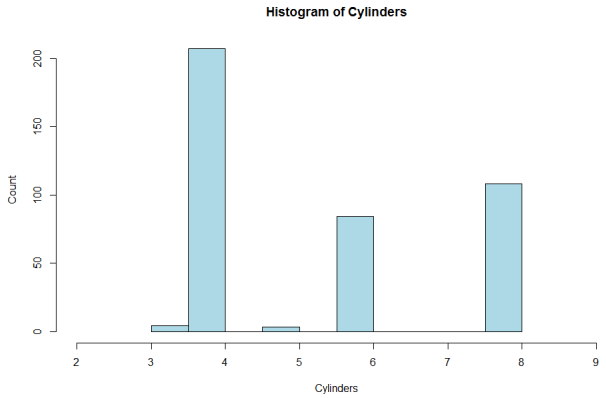

Figure 2. Histogram of Cylinders

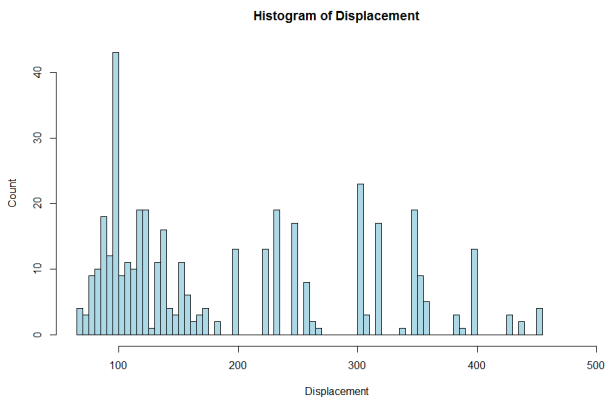

Figure 3. Histogram of Displacement

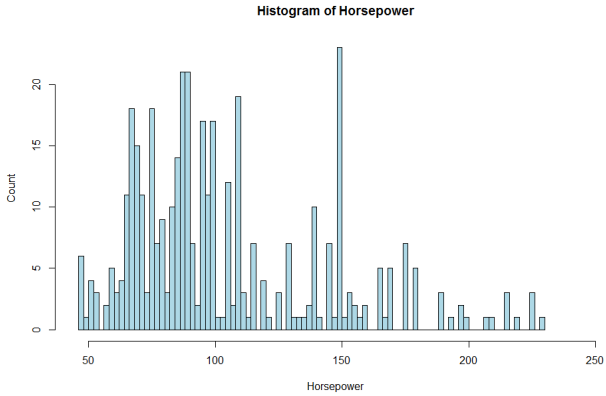

Figure 4. Histogram of Horsepower

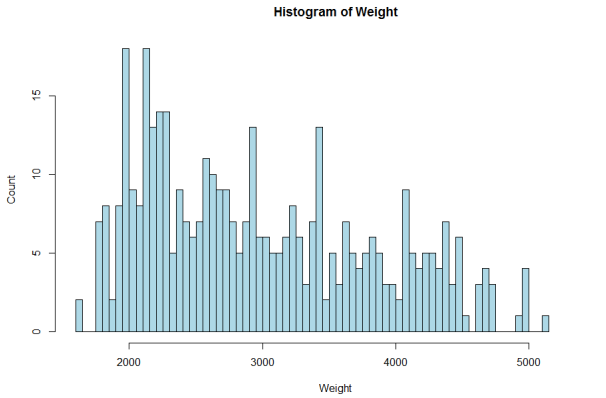

Figure 5. Histogram of Weight

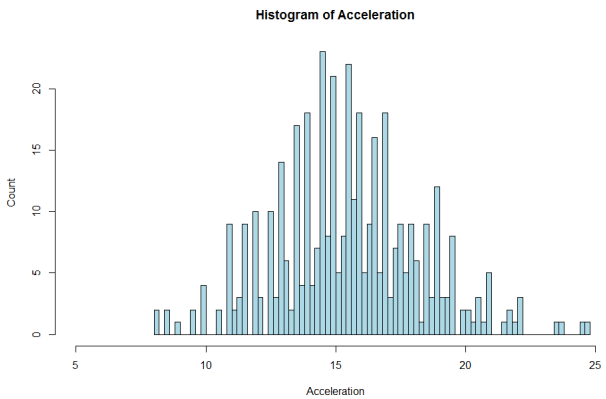

Figure 6. Histogram of Acceleration

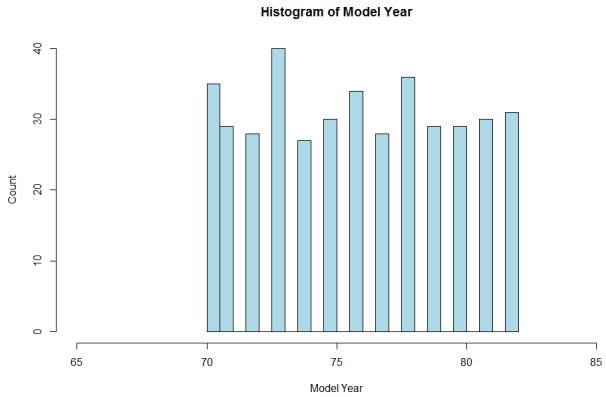

Figure 7. Histogram of Model Year

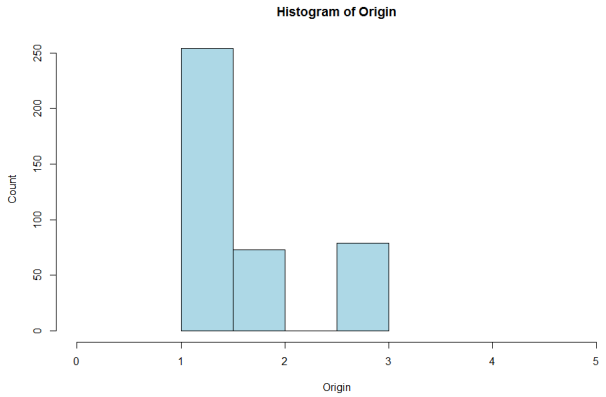

Figure 8. Histogram of Origin

2.2 Relationships Between Attributes

From my knowledge of cars, I chose to look at the relationship between mpg and the three variables displacement, horsepower, and weight. The best way to visualize the relationships was to use scatterplots. None of the three scatterplots for MPG vs displacement (Figure 9), MPG versus horsepower (Figure 10), as well as MPG versus weight (Figure 11) showed a linear or an exponential relationship, but all three scatterplots displayed a relationship of quadratic decay or a decreasing rate relative to the field value.

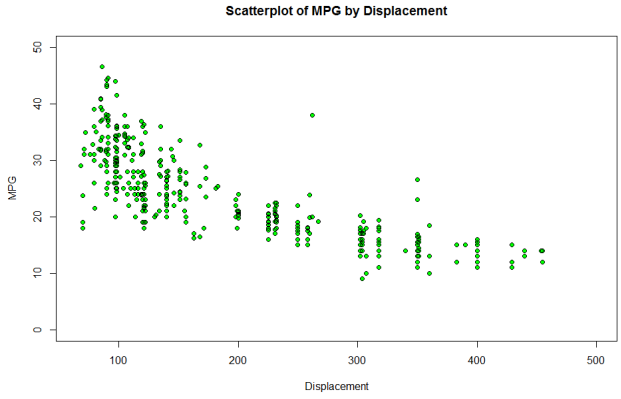

Figure 9. Scatterplot of MPG by Displacement

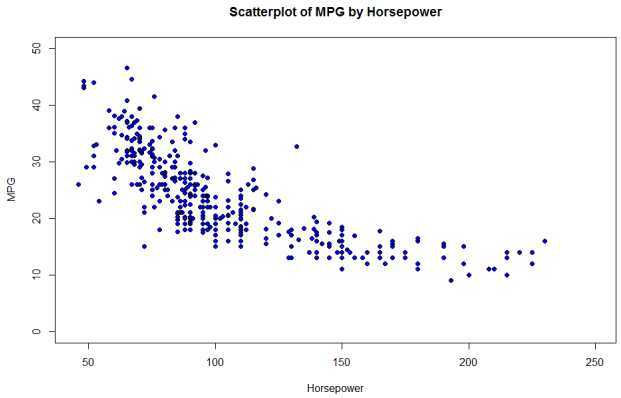

Figure 10. Scatterplot of MPG by Horsepower

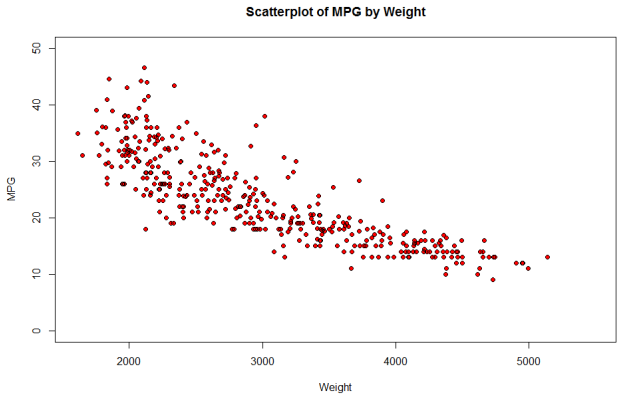

Figure 11. Scatterplot of MPG by Weight

2.3 More Analysis

Since the three variables I compared with MPG (displacement, horsepower, and weight) all seemed to have similar relationships to MPG I wondered what their relationships between each other were as well. I decided the best way to view any possible relationship was to use a scatterplot matrix (Figure 12). The scatterplot matrix shows a close to a linear relationship between horsepower and weight, weight and displacement, as well as horsepower and displacement. Considering the similarities between not only the histograms, but the scatterplots as well, this was not very surprising and confirmed what was seen in the other graphs.

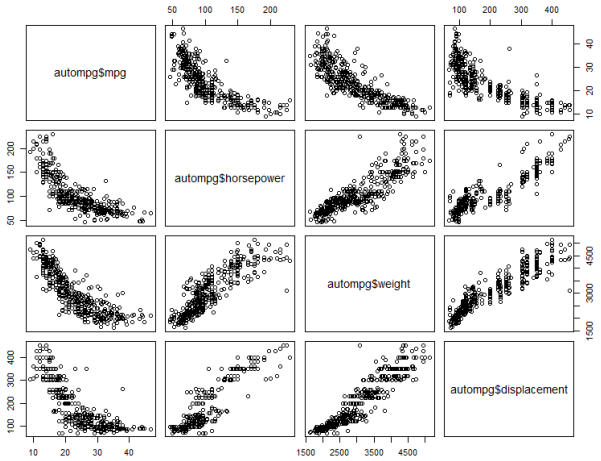

Figure 12. Scatterplot Matrix of MPG, Displacement, Horsepower, and Weight

3. DATA PREPROCESSING

3.1 Normalization

3.1.1 Min-Max Normalization

Min-Max normalization is used to determine how much greater a field value is than the minimum of the field values and scales this difference by the field's range. The formula for min-max normalization is as follows:

$$X^* = \frac{X - min(X)}{range(X)} = \frac{X - \min(X)}{max(X) - \min(X)}$$

A min-max normalization value of zero represents the minimum for the field values, a value of one represents the maximum for the field values, and a value of 0.5 represents the exact middle of the minimum and maximum.

For example, the displacement has a minimum of 68 and a range of 387 (maximum of 455); Therefore, a value like 383 would have a min-max normalized value of 0.8139535. Table 2 shows the summary of the original values as well as the min-max normalized values for mpg, displacement, horsepower, and weight.

Table 2. Field Value and Min-Max Normalized Summaries for MPG, Displacement, Horsepower, and Weight

```
MPG Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   17.50   23.00   23.51   29.00   46.60
MPG Min-Max Normalized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2261  0.3723  0.3860  0.5319  1.0000

Displacement Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   68.0   105.0   151.0   194.8   302.0   455.0
Displacement Min-Max Normalized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.09561 0.21450 0.32760 0.60470 1.00000

Horsepower Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   75.75   95.00  105.10  130.00  230.00
Horsepower Min-Max Normalized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1617  0.2663  0.3211  0.4565  1.0000

Weight Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1613    2226    2822    2979    3618    5140
Weight Min-Max Normalized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1739  0.3429  0.3874  0.5685  1.0000
```

3.1.2 Z-score Standardization

Z-score standardization takes the difference between the field value and the field value mean and scales the difference by the field's standard deviation. The formula for Z-score standardization is as follows:

$$X^* = \frac{X - mean(X)}{SD(X)}$$

Field values below the mean will have a negative Z-score, field values above the mean will have a positive Z-score, and field values that fall on the mean will have a Z-score of 0 (zero).

For example, the horsepower has a mean of 105.10 and a standard deviation of 38.76868; Therefore, a value like 198, which is above the mean, would have a positive Z-score of 2.396709. Table 3 shows the summary of the original values as well as the Z-score standardized values for mpg, displacement, horsepower, and weight.

Table 3. Field Value and Z-score Standardized Summaries for MPG, Displacement, Horsepower, and Weight

```
MPG Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   17.50   23.00   23.51   29.00   46.60
MPG Z-score Standardized Summary:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.85700 -0.76950 -0.06584  0.00000  0.70180  2.95400
Displacement Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   68.0   105.0   151.0   194.8   302.0   455.0
Displacement Z-score Standardized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2080 -0.8557 -0.4173  0.0000  1.0220  2.4800
Horsepower Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   75.75   95.00  105.10  130.00  230.00
Horsepower Z-score Standardized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.5240 -0.7566 -0.2601  0.0000  0.6427  3.2220
Weight Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1613    2226    2822    2979    3618    5140
Weight Z-score Standardized Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.6130 -0.8889 -0.1853  0.0000  0.7542  2.5510
```

3.1.3 Decimal Scaling

Decimal Scaling ensures that normalized values lie between -1 and 1. The formula for decimal scaling is as follows:

$$X^* = \frac{X}{10^d}$$

where *d* is the number of digits in the data value with the largest absolute value.

For example, the weight data has largest absolute value of |5140|, which would make *d* the value 4; therefore, a value like 3090 would have a decimal scaled value of 0.309. Table 4 shows the summary of the original values as well as the decimal scaled values for mpg, displacement, horsepower, and weight.

Table 4. Field Value and Decimal Scaled Summaries for MPG, Displacement, Horsepower, and Weight

```
MPG Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   17.50   23.00   23.51   29.00   46.60
MPG Decimal Scaled Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0900  0.1750  0.2300  0.2351  0.2900  0.4660
Displacement Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   68.0   105.0   151.0   194.8   302.0   455.0
Displacement Decimal Scaled Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0680  0.1050  0.1510  0.1948  0.3020  0.4550
Horsepower Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   75.75   95.00  105.10  130.00  230.00
Horsepower Decimal Scaled Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04600 0.07575 0.09500 0.10510 0.13000 0.23000
Weight Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1613    2226    2822    2979    3618    5140
Weight Decimal Scaled Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1613  0.2226  0.2822  0.2979  0.3618  0.5140
```

3.2 Binning

Data binning, sometimes referred to as bucketing, is a technique used in data pre-processing to accommodate algorithms that use categorical rather than continuous variables. The field values are each categorized into a bin representative of that field value. There are four common methods to bin field values: equal width binning, equal frequency binning, binning by clustering, and binning based on predictive value.

3.2.1 K-means

One method of binning is binning by clustering. In this method, a clustering algorithm is used to calculate automatically the "optimal" partitioning for the field values. In this case, I used the k-means clustering algorithm, which aims to partition the field value into *k* clusters. Each field value is binned into the cluster with the nearest mean. Figure 2 shows the weight values binned using a *k* of 3. This is one of the preferred methods of binning

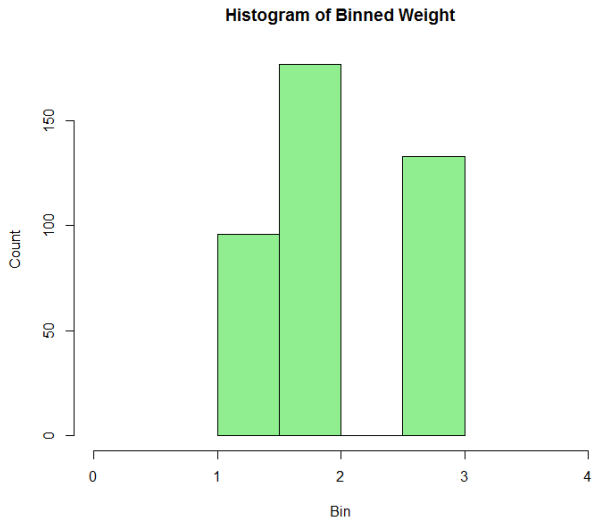

Figure 13. Histogram of K-means Binned Weight

3.2.2 Equal Width

A second method of binning is equal width binning. In this method, the field values are divided into *k* categories of equal width. Figure 3 shows the weight values binned using *k* of 3. This is not one of the preferred methods of binning since outliers may influence the width of the bins.

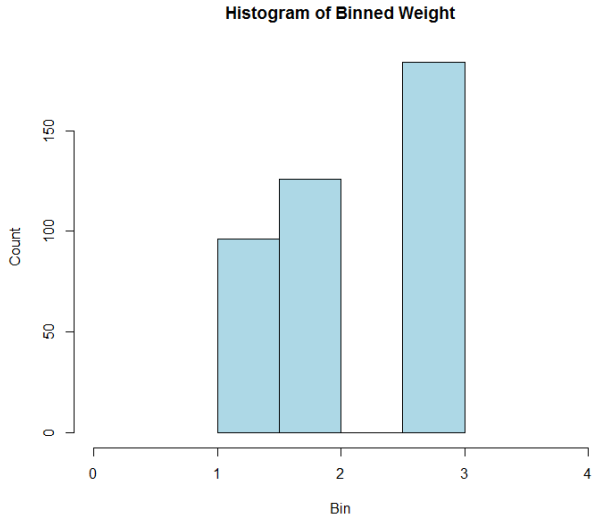

Figure 14. Histogram of Equal Width Binned Weight

3.3 Transformations to Achieve Normality

I chose the variable weight to use in achieving normality through transformations, since it does not have a normal distribution. I performed the natural log, square root, and inverse square root transformations to attempt to achieve normality. After calculating each transformation, I calculated the skewness to determine which of the transformations would be best in attempting to normalize the field values. The formula for skewness is as follows:

$$Skewness = \frac{3(mean - mediam)}{standard\ deviation}$$

The natural log transformation had a skewness value of 0.3608718. This value being positive indicates the mean is greater than the median and the data is right-skewed. The square root transformation had a skewness value of 0.1558764. This value is again positive indicating the data is right-skewed. The inverse square root transformation had a skewness value of 0.05198144. This value is the close to zero, which indicates the data is close to symmetric or at least the closest out of the three transformations.

Now that symmetry, or close to symmetry, has been achieved, I performed a check for normality. By plotting the normal probability plot, I would be able to see a normal distribution if the bulk of the points were in a straight line. As Figure 15 shows, the weight data did not adhere to a straight line indicating I was not successful in achieving normality.

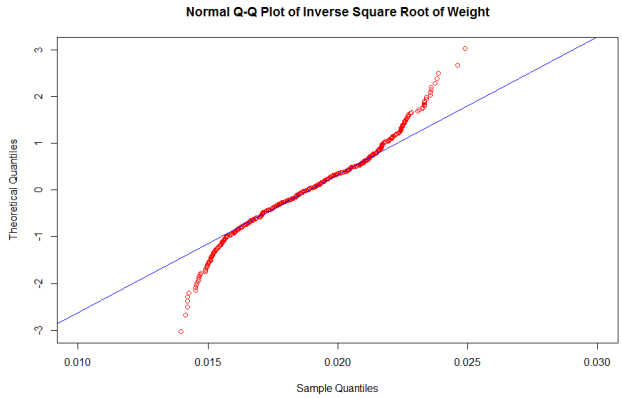

Figure 15. Normal Q-Q Plot of Inverse Square Root of Weight

4. REGRESSION ANALYSIS

From analyzing the data, there seemed to be an inverse relationship between horsepower and mpg as well as weight and mpg. In other words, as the horsepower increased the mpg decreased. Similarly, as weight increased the mpg seemed to decrease. For an even better understanding, regression analysis can be performed on the mpg, horsepower, and weight field values.

4.1 Regression Fit

4.1.1 MPG versus Horsepower

First, I used the R regression fit for mpg and horsepower. The results, in Table 5, have an R-squared value of 0.6059 and adjusted R-squared value of 0.649. This is a higher R-squared value, which indicates the model fits the data fairly well.

Table 5. Regression of MPG versus Horsepower

```
Call:
lm(formula = nonaall_mpg ~ nonaall_horsepower)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        39.935861   0.717499   55.66   <2e-16 ***
nonaall_horsepower -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

4.1.2 MPG versus Weight

Second, I used the R regression fit for mpg and weight. The results (Table 6) have an R-squared value of 0.6926 and adjusted R-squared value of 0.6918. This would be considered a higher R-squared value, which indicates a model that fits the data better.

Table 6. Regression of MPG versus Weight

```
Call:
lm(formula = nonaall_mpg ~ nonaall_weight)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9736  -2.7556  -0.3358   2.1379  16.5194

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    46.216524   0.798673   57.87   <2e-16 ***
nonaall_weight -0.007647   0.000258  -29.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.333 on 390 degrees of freedom
Multiple R-squared:  0.6926,    Adjusted R-squared:  0.6918
F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

4.2 Correlation Values

Using the p-value correlation values, one can determine the probability of observing the same results in the future as were observed in the field values. Since p-value is a probability it will fall between 0 and 1, with a small value being a predictor of rejecting the hypothesis; however, in this case since I am looking at the inverse relationship, a small p-value would be an indicator of accepting the hypothesis. The p-value for horsepower versus mpg was 7.031989e-81, close to zero, which would indicate a high correlation between increased horsepower and decreased mpg. The p-value for weight versus mpg was 6.015296e-102, which again is close to zero and would indicate a high correlation between decreased mpg for increasing weight. The calculated correlation vales for mpg, horsepower, and weight can be found in Table 7. While these calculated values do not necessarily confirm the hypothesis that increased horsepower will decrease mpg or that increased weight will decrease mpg, they do eliminate rejecting both hypotheses.

Table 7. Correlation Values for MPG, Horsepower, and Weight with p-values

	nonaall_mpg	nonaall_horsepower	nonaall_weight
nonaall_mpg	1.0000	-0.7784	-0.8322
nonaall_horsepower	-0.7784	1.0000	0.8645
nonaall_weight	-0.8322	0.8645	1.0000

5. REFERENCES

[1] Larose, Daniel and Larose, Chantal D. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience.