Data Preprocessing and Exploratory Data Analysis on Automobile Data

William G. Hatcher
Department of Computer and Information Sciences
Towson University, Maryland, USA 21252, USA
Emails: whatch2@students.towson.edu

***Abstract*—Big Data is being increasingly generated by Internet of Things technologies, driving an growing need for highly skilled data scientists and highly accurate and applicable data analytics. Data Mining, as a broad field that encapsulates the various stages of data discovery, data cleaning, data analysis, model generation, and implementation, also promotes a deeper understanding of statistical mechanisms and the proper application thereof. In this work, we apply these data mining techniques to reprocess and analyze data for further regression analysis. Using a real-world dataset of automobile attributes, we attempt to achieve accurate prediction of mile per gallon data, as well as understand the inherent relationships between the various attributes. Through detailed experimentation, we demonstrate the applicability of data mining for prediction, and avoid the pitfalls and shortcomings of unplanned and non-incremental analytic techniques.**

***Index Terms*—Data mining, Machine learning, Regression**

I. Introduction

In the current era, the emergence of Big Data has engendered an arms race in data analytics. Particularly, as data exceeds the traditional boundaries of volume, veracity, velocity, value, etc., becoming a new and basic commodity, the computational techniques to understand this data, and to extract detailed insight and prediction from this data have matured rapidly. Data Mining, as the holistic process of understanding and generating value from data, encapsulates the processes of data preprocessing, reduction, analysis, insight, projection, and application. Moreover, with recent advances in machine learning technologies, the prospects of data mining have never been better.

At a fundamental level, data mining is directed toward finding useful patterns in and approximations of data. The term *useful* is critical, as the implication is that data mining is applied toward a purpose, whether it be through applied business analytics and providing better services or pushing the boundaries of scientific exploration and medical diagnosis.

In this work, we consider historical trends in automotive manufacturing, and assess the complexities and interrelationships between vehicle and engine properties, and gasoline usage. To be more concise, we analyze data publicly shared by UC Irvine that describes various automobiles and their properties spanning the model years 1970 through 1982, focusing in particular on the measurements of miles per gallon in relation to other vehicle properties, such as displacement and number of cylinders among others.

The remainder of this paper is as follows. In Section II, we provide a basic description of the datasets introduced. In Section IV we conduct an exploratory data analysis and make some preliminary results. In Section III we outline the preprocessing steps taken. In Section V, we conduct a regression analysis and develop a predictive model. Finally, in Section VI, we provide concluding remarks.

II. Data Description

We now describe the dataset utilized from the perspectives of the included Attributes, the state of the Unprocessed Data, and the Statistical Properties of the raw data.

A. Attributes

The dataset consists of 398 instances, each having nine attributes: mpg (miles per gallon), cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name. Of these, five are continuous value attributes (*mpg*, *displacement*, *horsepower*, *weight*, and *acceleration*), three are multi-valued discrete attributes (*cylinders*, *model year*, and *origin*), and the last attribute is a unique string or identifier representing the make and model of the vehicle (*car name*).

Considering the discrete-valued attributes, we note that the *cylinder* attribute denotes the number of cylinders in the vehicle's engine, represented by integer values of 3, 4, 5, 6, or 8. Additionally, the *model year*, as noted above, ranges from 1970 to 1982, while the *origin* attribute represent integer values from one to three.

B. State of Unprocessed Data

Regarding the state of the unprocessed data, there are a total of six missing values, each from a different instance, and each belonging to the *horsepower* attribute. These instances have been removed in total due to lack of particular domain knowledge or expertise. More detail is provided below in Section III.

C. Extended Description

As we can see from Table I, a preliminary analysis provides the minimum, maximum, and quantile values of each attribute, excluding *car name*, along with mean, median, mode and variance. For instance, the mpg data has a minimum value of 9 and a maximum of 46.6, with a mean of 23.5. Likewise, displacement falls within a range of 68 to 455 with a median value of 151.

TABLE I. Grid Search Parameters

Measurement	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Min	9	3	68	46	1613	8.0	70	1
25% Quantile	17	4	105	75	2225.3	13.8	73	1
50% Quantile (Median)	23	4	151	93.5	2803.5	15.5	76	1
75% Quantile	29	8	276	126	3614.8	17.0	79	2
Max	46.6	8	455	230	5140	24.8	82	3
Mean	23.4	5.5	194.4	104.5	2977.6	15.5	76.0	1.6
Mode	13	4	97	150	2130	14.5	73	1
Variance	60.8	2.9	10922.4	1477.8	719644	7.6	13.5	0.6

Note that, due to lack of information from the data provider, there is some lack of clarity regarding several attributes. For instance, it is unclear if the unit of measure of the *displacement* attribute is in units of cubic inches or cubic centimeters. Likewise, the *weight* is assumed to be in lbs, but no description is provided. More concerning, the attribute of *origin* is not well defined, and can only be guessed at.

III. Data Preprocessing

Data preprocessing is necessary on raw datasets to properly conduct analysis and evaluation. Especially in the context of Big Data, data preprocessing and cleaning help to combat unintended errors in collected data, help to remove personally identifying information, and control for outliers and anomalies. As we consider the generation of big data by devices to be massive, the probability of data errors increases apace. In this section, we consider the methods for data cleaning, as well as normalization for comparative analysis.

A. Data Cleaning

As discussed above, we have noted missing values from data instances in our data description section (Section II). Cleaning datasets helps to reduce errors in observation and analysis due to anomalies. For instance, missing data that has a null or zero value may severely bias an analysis if left unresolved. Likewise, data that is of the incorrect type may cause similar problems, either being dropped, or mishandled, potentially interfering with system operation.

In our case, we have resolved to remove the erroneous data points. This decision was based on the lack of clear information as to the nature of attribute derivation or calculation, as well as lack of specific domain knowledge. While this is precautionary, we feel the inference of the missing data, specifically under the *horsepower* attribute, would be reasonably predictable. This is because of the clear relationships between mpg, displacement, horsepower, and weight, as we see in Figures 1 and 2. Specifically, we see a strong positive correlation between weight, displacement, and horsepower, and a negative correlation between these three and mpg.

B. Normalization

Given the simple plots in Figures 1 and 2, we can see that, despite the clear trends, the data is quite noisy, and requires reduction and transformation. Moreover, we can see from the figures and Table I that the scales of each set of

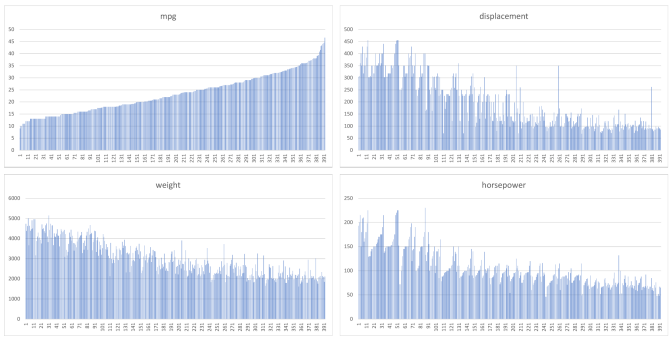

Fig. 1. Bar charts of mpg, displacement, weight, and horsepower, sorted by mpg.

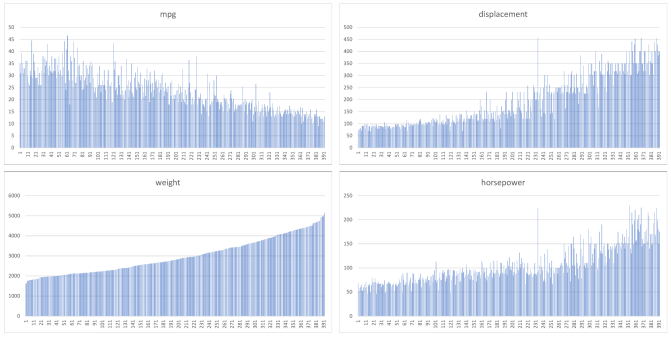

Fig. 2. Bar charts of mpg, displacement, weight, and horsepower, sorted by weight.

continuous data are very different, potentially giving rise to bias in a later regression analysis. In addition, considering Figure 4, we can observe the distributions of our dataset in top two level histograms, as well as the kernel density estimation (KDE) graphs on the bottom. Note that the two top level histograms cover the same original dataset, but the first set has a fixed binning of 30 bins, while the second set is generated automatically (fixed at 10 bins). This gives us two different views to better understand the properties of the dataset as a whole. Additionally, note that KDE is a mechanism for viewing data distributions similar to a histogram, typically using a Gaussian kernel.

Observing these trends in our data, we consider Min-Max, Z-Score, and Decimal Scaling normalization methods, as depicted in Figure 4. Primarily, we can see in rows three and four of the figure the histograms for Min-Max and Z-

Fig. 3. Histograms (Top to Bottom) of Original data with 30 bins, the Original data with 10 bins, Min-Max Normalization, Z-Score, and Kernel Density Estimation transformations for: mpg, weight, displacement, horsepower, and acceleration.

Score respectively, making note of the change in x-axis scale. Note that no noticeable change occurred to the distributions visually from the application of these techniques.

Considering methods of binning data, aside from the two different binnings (30 and 10) of the top two rows in Figure 4 mentioned above, we have additionally investigated the binning of the Horsepower attribute, as evident in Figure 5. We can see in the left subfigure the results of the Equal Frequency binning method, and notice the small clustering of many data points into the first and second groups. Clearly, these groups have much narrower dispersions than the third group, which appears to deviate more severely. In the rightmost subfigure, we can observe the Equal Width binning mechanism.

Returning our focus to normalization techniques, we have applied logarithmic, inverse square root, and square root normalization transformations to our dataset, as seen in Figure 6. Again, note that the top row is our original dataset binned in 10 bins. In this case, we can clearly see the effect of the transformations on the data. Particularly, mpg and weight attributes become much more normalized, and the horsepower attribute appears to become less skewed.

Despite these effects, we can observe from the normal probability plots in Figure 7, which match the plots in Figure 6, that none of our attributes, aside from Acceleration, have a very normal distribution. Indeed, while we have boxed in red those transformations that appear to have the best affect of normalization, these are still not very close to normal.

IV. Exploratory Data Analysis

Having considered all of this information provided, we can make several assessments about our data. First, we can see that much of our data does not follow a normal distribution. Second, we have observed strong positive correlations between the attributes of weight, horsepower, and displacement. This is clearly visible in Figure 8. In addition, these three are strongly inversely correlated with mpg. Additionally, though Acceleration displays a strong normal distribution, it does not appear to be highly correlated to the attributes of mpg, displacement, horsepower, or weight of the vehicles. Finally, we can consider our attempts at normalization to be relatively unsuccessful.

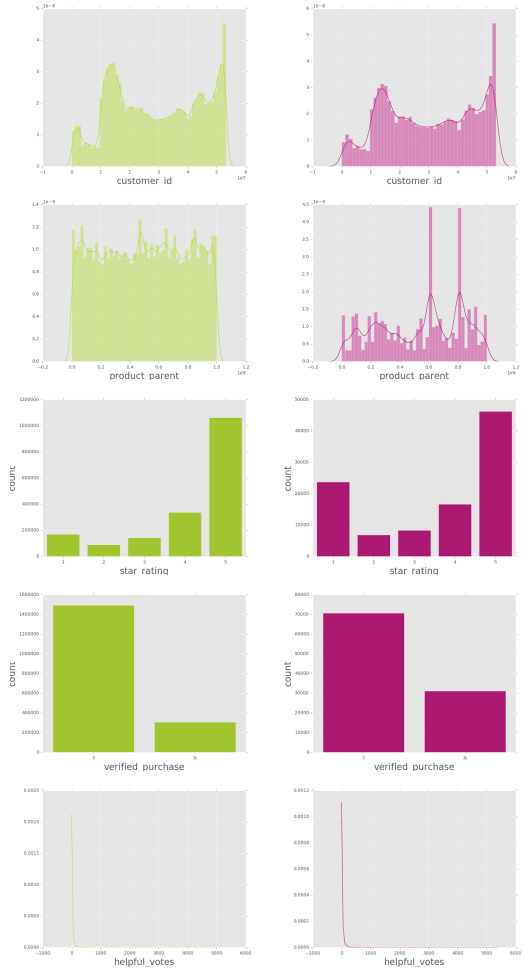

Fig. 4. Histograms (Top to Bottom) of Original data with 30 bins, the Original data with 10 bins, Min-Max Normalization, Z-Score, and Kernel Density Estimation transformations for: mpg, weight, displacement, horsepower, and acceleration.

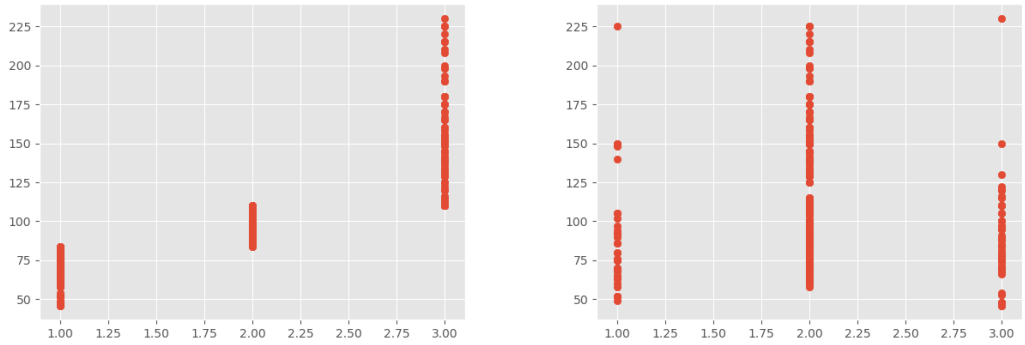

Fig. 5. Plots of two binning methods on the Horsepower attribute both using 3 bins. These methods are Equal Frequency binning (Left) and Equal Width Binning (Right).

V. Regression Analysis

Having applied a variety of statistical techniques we now consider the data for regression analysis. In particular, we have already observed strong correlations between the three attributes of weight, horsepower, and displacement. In addition, as the stated goal of this exploratory research, we are interested in predicting the mpg of new cars based on the generic attributes of the vehicle. Thus we consider two groups of predictions that we want to investigate and believe will likely be successful. The first considers the strong prediction of weight, horsepower, or displacement based on one of the two. In this case we selected weight as the attribute to predict

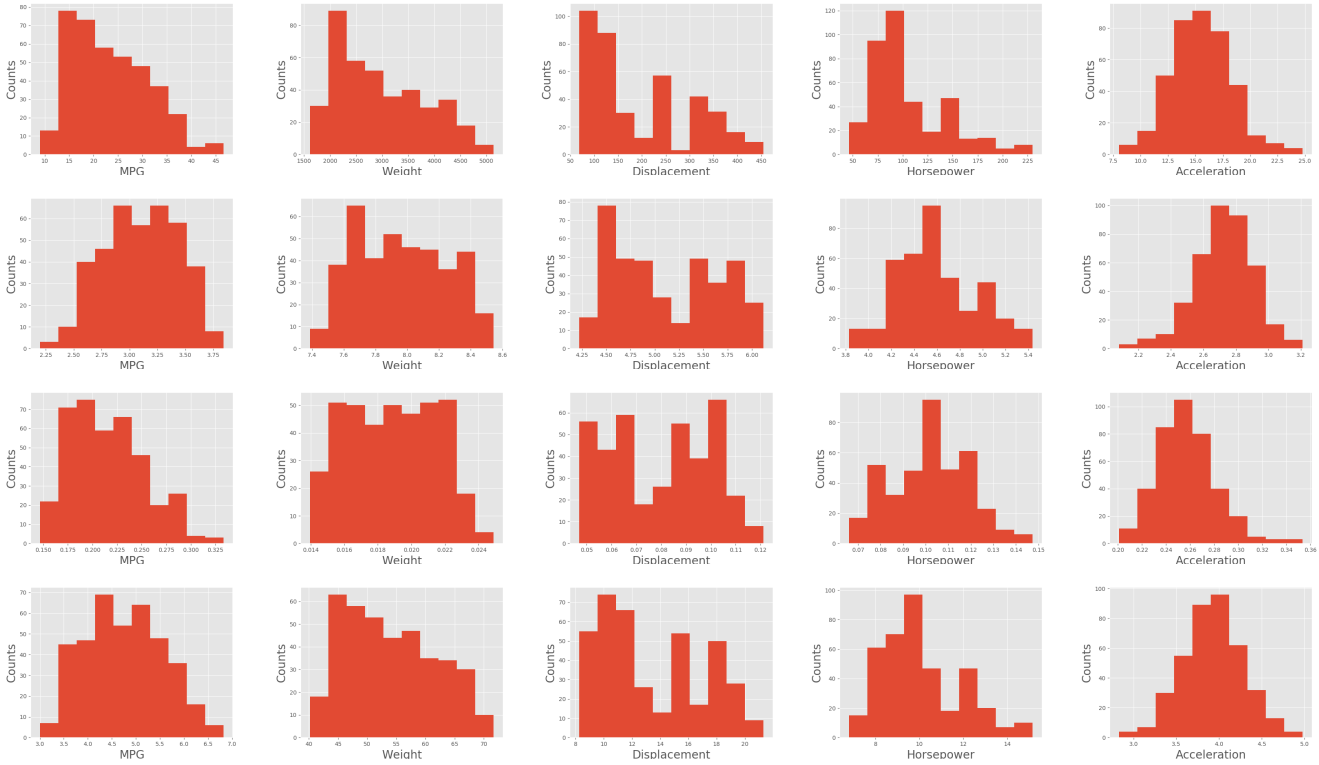

Fig. 6. Histograms (Top to Bottom) of the Original data separated into 10 bins, Natural Log transformation, Inverse Square Root transformation, and Square Root transformation for: mpg, weight, displacement, horsepower, and acceleration.

based on the horsepower and displacement. The second relies on the strong inverse correlations of the mpg with the three mentioned attributes. Finally, we also consider both weight and mpg predicted based upon acceleration, which we believe will likely fail.

In Figure 9 we can see the results of our regression analysis. Note that the plots display the data relationship denoted on the x- and y-axes, and the trend lines represent the result of linear regression. As predicted, the regression of mpg, while not the most accurate, still provides some meaningful approximate prediction. The caveats we should note are the likely better prediction that would be achieved by a polynomial regression, rather than a linear one. Here, we can clearly see the problem of extrapolating our prediction too far past the range of our dataset.

Even better than the regressions for the mpg, we can clearly see very accurate trends between weight, displacement, and horsepower. Indeed, this is a very good result, and only a heteroscedastic plot would better encapsulate these attribute relationships. Finally, as predicted, the acceleration attribute was not very useful for prediction. This was clear, however, from the significant difference in distributions.

VI. Conclusion

Data mining provides practitioners with powerful tools to assess and extract insight from highly divers and correlational data of all types. Further still, the data mining requires first the preprocessing of noisy and anomalous data before effective analysis can occur. Likewise, the ability to visualize data can provide an intuitive understanding of the data that may otherwise be lost. In this work, we applied these principles to accurately predict the outcome of our regression analysis. We first considered our data in detail, and assessed the potential and need for transformation of the data for analysis. In addition, we conducted preprocessing and exploratory analysis to determine the likely best predictive attributes. Finally, our predictions bore out, as our highly correlated weight, displacement, and horsepower attributes indeed, through linear regression, achieved strong predictive potential.

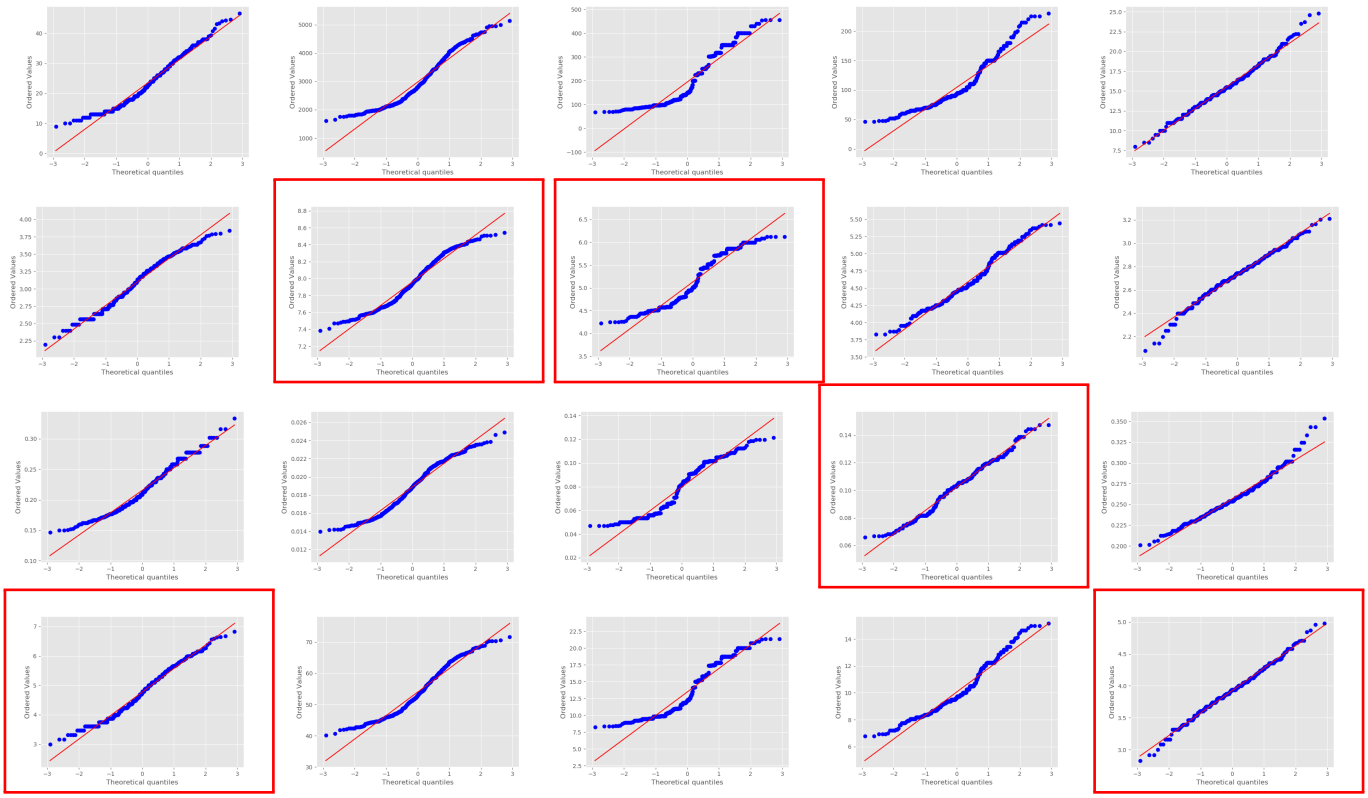

Fig. 7. Normal probability plots (Top to Bottom) for the Original data separated into 10 bins, the Natural Log transformation, the Inverse Square Root transformation, and Square Root transformation for : mpg, weight, displacement, horsepower, and acceleration. Red frames indicate the best observed normalization transformations.

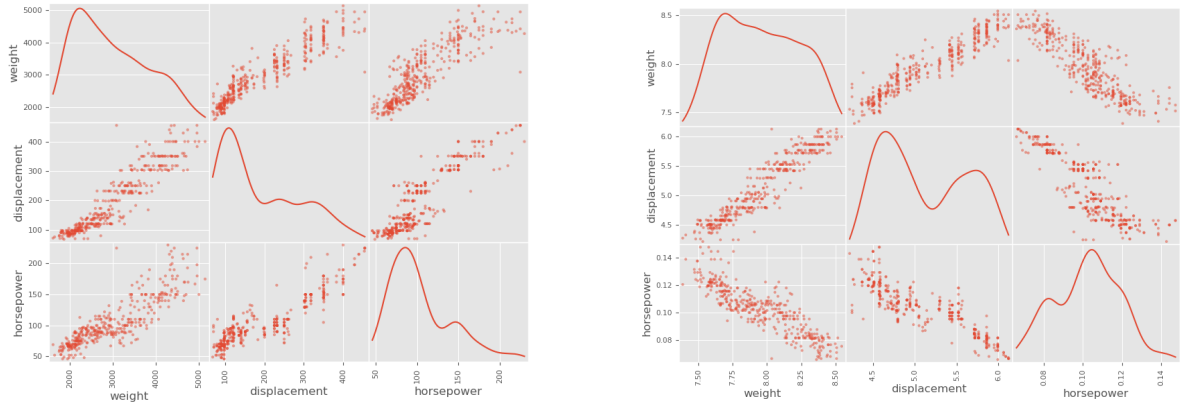

Fig. 8. Original data (left) and normalized data (right) scatter matrix plots of weight, displacement, and horsepower.

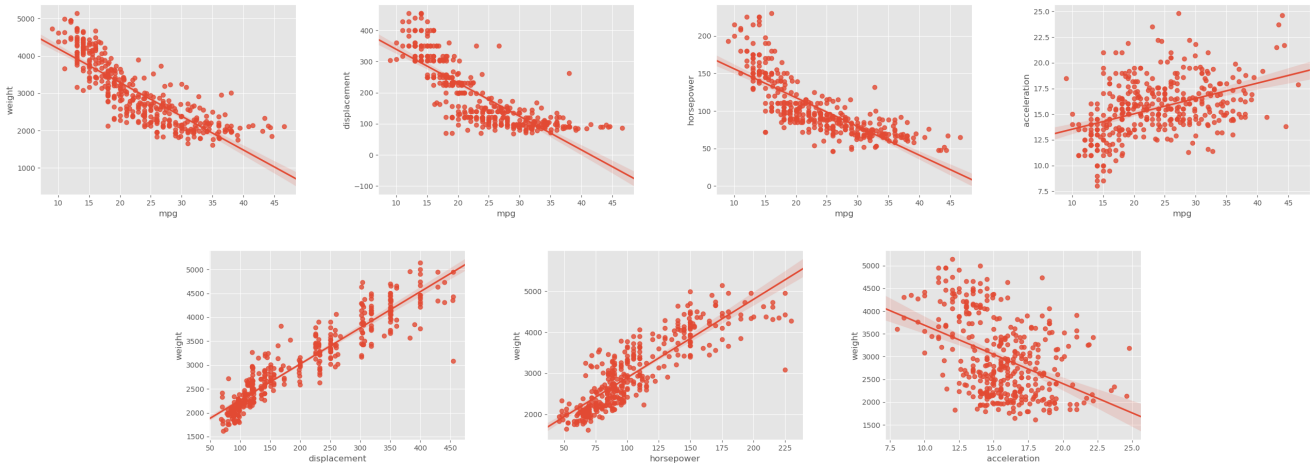

Fig. 9. Binary linear regression analysis of mpg (top) and weight(bottom).