Contents

[1. Objective 2](#_Toc93875495)

[1.1. Dataset 2](#_Toc93875496)

[2. Character Variables 2](#_Toc93875497)

[2.1. Extract brand name and model from Car name 2](#_Toc93875498)

[2.2. Checking for Spelling errors 3](#_Toc93875499)

[2.3. Changing Date format 4](#_Toc93875500)

[3. Numeric Variables 4](#_Toc93875501)

[3.1. Descriptive Statistics 4](#_Toc93875502)

[3.2. Imputing missing values 4](#_Toc93875503)

[3.3. Detecting Outliers 5](#_Toc93875504)

[3.4. Treating Outliers 6](#_Toc93875505)

[3.5. Logarithm Transformation 7](#_Toc93875506)

[4. Conclusion 8](#_Toc93875507)

# Objective

Data cleaning and preparation using SAS to prepare it for regression.

## Dataset

The dataset has been downloaded from <https://archive.ics.uci.edu/ml/datasets/Auto+MPG>.

The task associated with the dataset is regression. Since, statistical methods assume data to be normally distributed and have equal variance, we will analyse all numeric variables for outliers, missing values, check kurtosis and skewness and apply the suitable data transformation technique.

The dataset has 398 observations

Metadata for the dataset is as follows:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Description** |
| Mpg | Continuous numeric | Vehicle fuel consumption in mpg |
| Cylinders | Discrete numeric | Number of cylinders |
| Displacement | Continuous numeric | Engine displacement in cc |
| Horsepower | Continuous numeric | Vehicle Horsepower |
| Weight | Continuous numeric | Vehicle weight in lbs |
| Acceleration | Continuous numeric | Vehicle acceleration |
| Model year | Discrete numeric | Model year |
| Origin | Discrete numeric | Country of Origin |
| Car name | Character string | String contains Brand and model of vehicle |

# Character Variables

## Extract brand name and model from Car name

The variable **Car name**, is a string. The table below is an example. The first word of the string is the vehicle Brand name followed by the model.

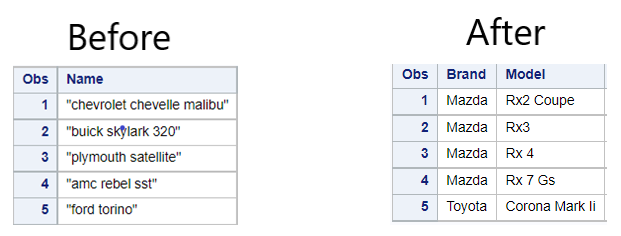


Figure 1: Car name string

First, task is to use propcase to change the first alphabet of each word to a capital letter. Next, an array is used to extract each word from the string and assign it to temporary variables.

The first word is assigned to a new variable called Brand. The following words are concatenated and assigned to the variable Model.

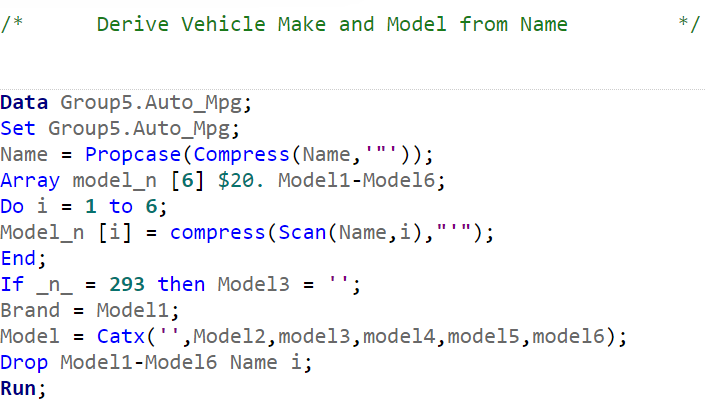


Figure 2: Using array to extract Brand and Model from Car Name

## Checking for Spelling errors

Proc freq in SAS, is used to display the number of observations under each category for categorical variables. We can use this procedure to quickly check if there are any spelling errors in Brand and rectify them.

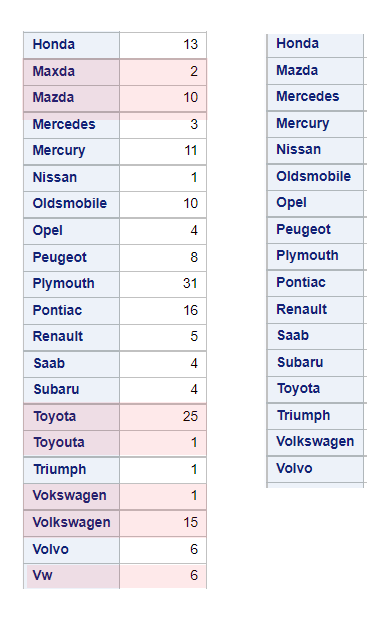


Figure 3: Spelling errors highlighted in red, in left table. Corrected entries in right table.

## Changing Date format

In the raw data, Date is a 2-digit integer (YY). SAS calculated dates as number of days from 1st January 1960. Hence, to convert date we first concatenate 02/01 with date. Using the concatenate function converts it to character type. Using the input function, it is converted to ddmmyy format and only the year part is extracted.

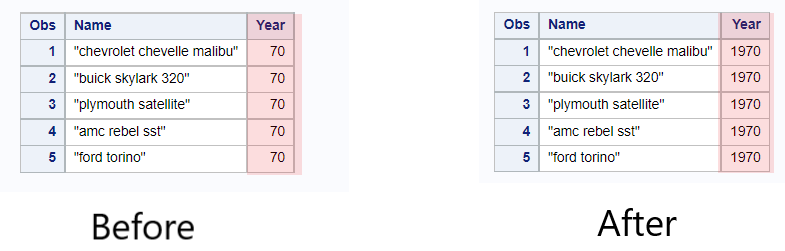


Figure 4: Date changed from an integer type to YYYY

# Numeric Variables

## Descriptive Statistics

We start by checking the descriptive statistics like mean, median, minimum, maximum, standard deviation and number of missing values for continuous numeric variables. There are 6 missing values for horsepower.

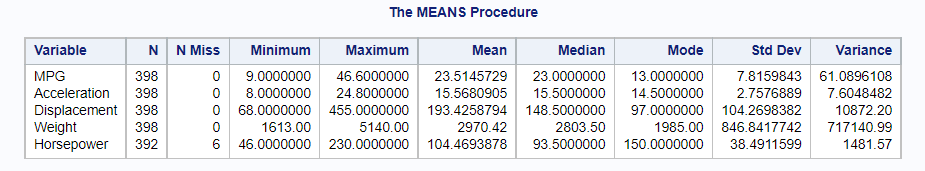


Figure 5: Checking descriptive statistics

## Imputing missing values

Since we have a categorical variable **Cylinders**, we use this variable to calculate the mean horsepower for each category and impute the missing values based on the category they fall under.

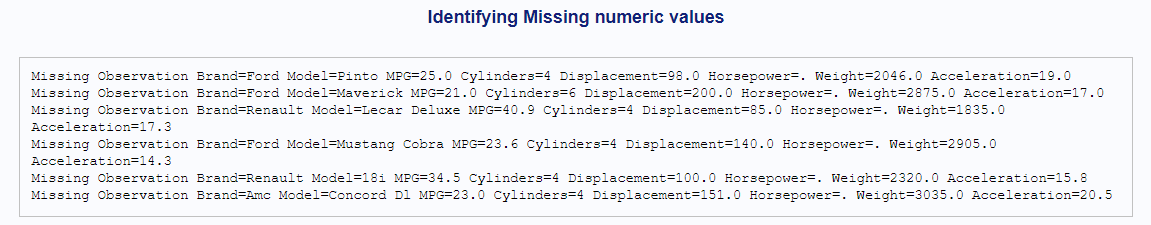


Figure 6: Identifying missing values of horsepower

We can observe that we have missing horsepower values for vehicles in the 6 and 4 cylinders categories.

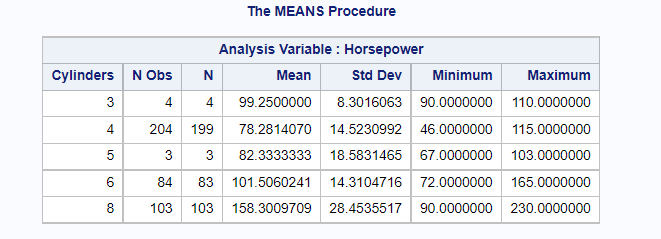


Figure 7: Calculating mean for Cylinder categories

We impute the missing values of 4 cylinder cars with 78.28 and 6 cylinder cars with 101.50 respectively.

## Detecting Outliers

Outliers are values that are extremely large or abnormal. They can influence the mean, median, standard deviation and variance by inflating them. Thus, by using the box plots and Q-Q plots, we can identify outliers.

Power to weight ratio is a derived variable. It is calculated by dividing Horsepower by Weight. From its box plot we can observe that it has outliers.

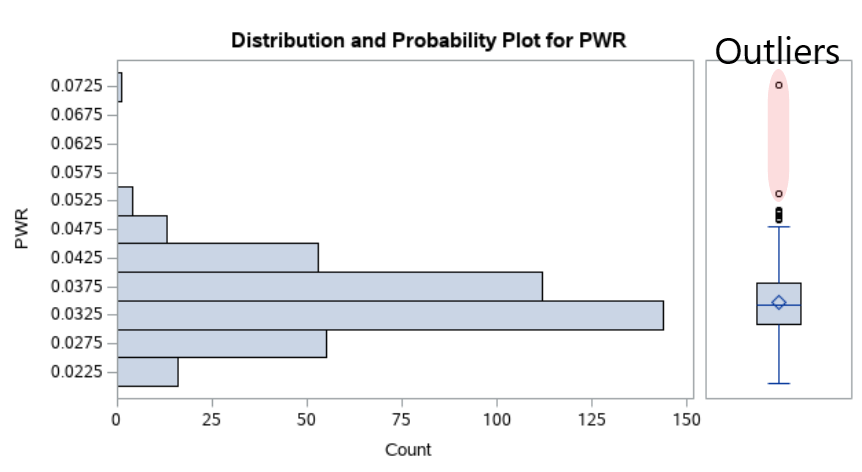


Figure 8: Power to weight ratio box plot

## Treating Outliers

To treat outliers, new minimum and maximum are calculated. All values falling outside these new limits can be deleted. There are 2 methods for calculating these limits.

If the data is normally distributed, mean and standard deviation are used to calculate the new limits as follows:

Minimum = Variable mean – 2\*Standard deviation

Maximum = Variable mean + 2\*Standard deviation

Acceleration has a skewness of 0.163 and by observing the box plot and Q-Q plot we can conclude that it is normally distributed. We use the above method to treat outliers for acceleration.

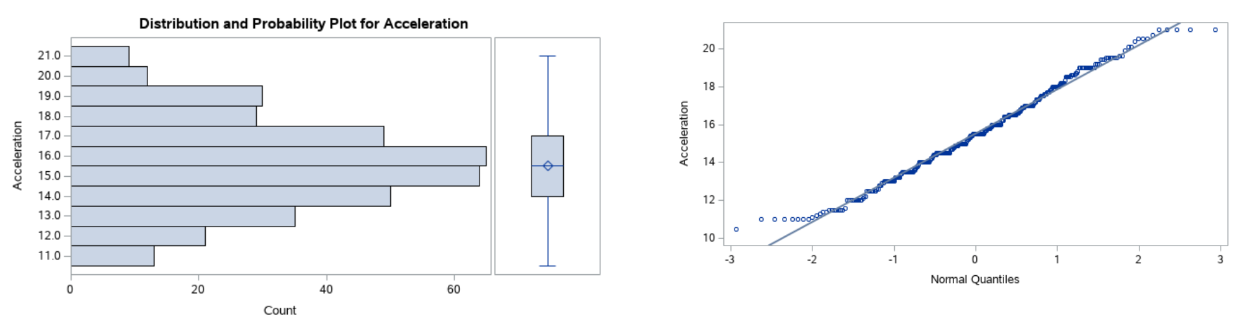


Figure 9: Distribution for Acceleration

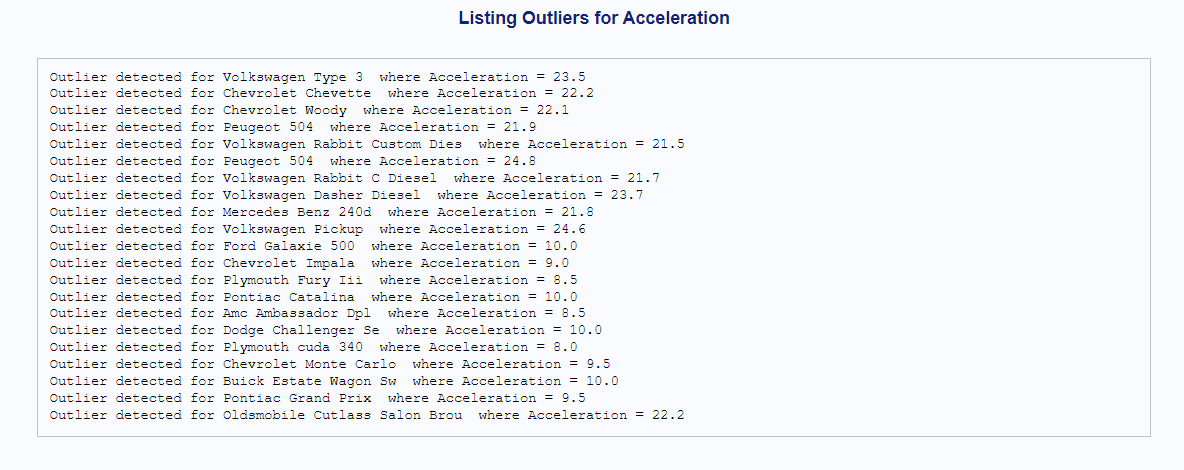


Figure 10: Outliers for acceleration calculated using mean and standard deviation

For variables that are not normally distributed, inter quartile range is used to calculate the new minimum and maximum as follows:

Minimum = Quartile1 – 1.5\*inter quartile range

Maximum = Quartile3 + 1.5\*inter quartile range

Power to weight ratio is not normally distributed, hence, we use this method to detect outliers.

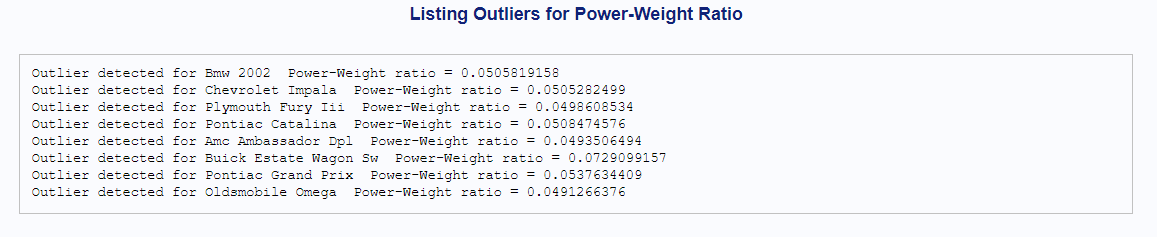


Figure 11: Outliers for power to weight ratio calculated using inter quartile range

## Logarithm Transformation

Statistical methods reply on the assumption that data is normally distributed. Log transformation are seldom used to lower skewness and normally distribute data. Horsepower has a skewness of 1.05 and by observing the histogram, we can conclude that it is not normally distributed. We perform a log transformation on it.

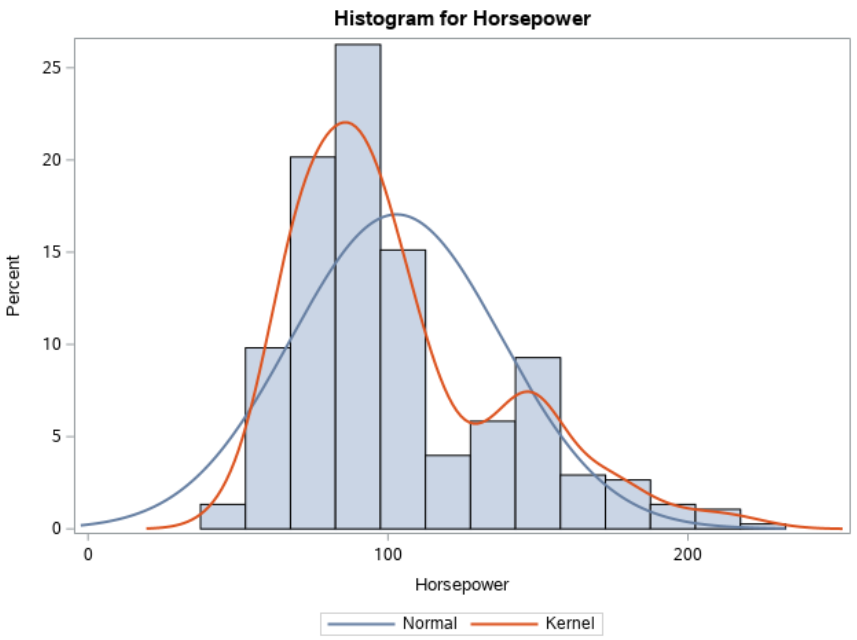


Figure 12: Horsepower distribution before log transformation

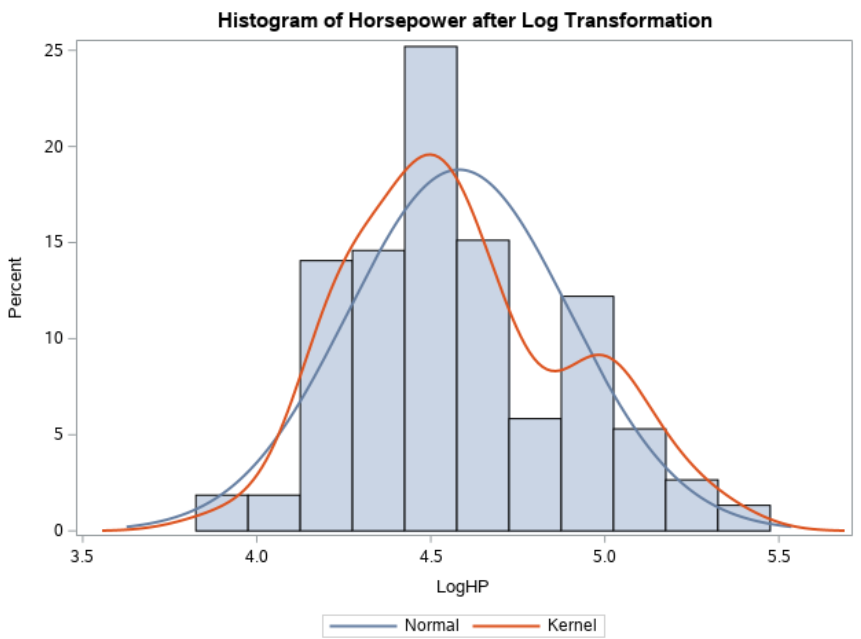


Figure 13: Horsepower distribution after log transformation

We can observe, that post log transformation, the skewness for variable Horsepower has reduced to 0.403 and is nearly normally distributed.

# Conclusion

The data is now ready for regression. These are some of the steps usually employed in preparing data for statistical analysis. The data cleaning would depend on the task associated with the data set.