Contents

1. Objective

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

1.1. 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

2. Character Variables

2.1. 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.

Before

Obs	Name
1	"chevrolet chevelle malibu"
2	"buick skylark 320"
3	"plymouth satellite"
4	"amc rebel sst"
5	"ford torino"

After

Obs	Brand	Model
1	Mazda	Rx2 Coupe
2	Mazda	Rx3
3	Mazda	Rx 4
4	Mazda	Rx 7 Gs
5	Toyota	Corona Mark Ii

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.

```
Data Group5.Auto_Mpg;
Set Group5.Auto_Mpg;
Name = Propcase(Compress(Name,'"'));
Array model_n [6] $20. Model1-Model6;
Do i = 1 to 6;
Model_n [i] = compress(Scan(Name,i),"'");
End;
If _n_ = 293 then Model3 = '';
Brand = Model1;
Model = Catx('',Model2,model3,model4,model5,model6);
Drop Model1-Model6 Name i;
Run;
```

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

2.2. 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.

Honda	13
Maxda	2
Mazda	10
Mercedes	3
Mercury	11
Nissan	1
Oldsmobile	10
Opel	4
Peugeot	8
Plymouth	31
Pontiac	16
Renault	5
Saab	4
Subaru	4
Toyota	25
Toyouta	1
Triumph	1
Vokswagen	1
Volkswagen	15
Volvo	6
Vw	6

Honda
Mazda
Mercedes
Mercury
Nissan
Oldsmobile
Opel
Peugeot
Plymouth
Pontiac
Renault
Saab
Subaru
Toyota
Triumph
Volkswagen
Volvo

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

2.3. 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.

Obs	Name	Year
1	"chevrolet chevelle malibu"	70
2	"buick skylark 320"	70
3	"plymouth satellite"	70
4	"amc rebel sst"	70
5	"ford torino"	70

Before

Obs	Name	Year
1	"chevrolet chevelle malibu"	1970
2	"buick skylark 320"	1970
3	"plymouth satellite"	1970
4	"amc rebel sst"	1970
5	"ford torino"	1970

After

Figure 4: Date changed from an integer type to YYYY

3. Numeric Variables

3.1. 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.

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum	Mean	Median	Mode	Std Dev	Variance
MPG	398	0	9.0000000	46.6000000	23.5145729	23.0000000	13.0000000	7.8159843	61.0896108
Acceleration	398	0	8.0000000	24.8000000	15.5680905	15.5000000	14.5000000	2.7576889	7.6048482
Displacement	398	0	68.0000000	455.0000000	193.4258794	148.5000000	97.0000000	104.2698382	10872.20
Weight	398	0	1613.00	5140.00	2970.42	2803.50	1985.00	846.8417742	717140.99
Horsepower	392	6	46.0000000	230.0000000	104.4693878	93.5000000	150.0000000	38.4911599	1481.57

Figure 5: Checking descriptive statistics

3.2. 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.

Identifying Missing numeric values

```
Missing Observation Brand=Ford Model=Pinto MPG=25.0 Cylinders=4 Displacement=98.0 Horsepower=. Weight=2046.0 Acceleration=19.0
Missing Observation Brand=Ford Model=Maverick MPG=21.0 Cylinders=6 Displacement=200.0 Horsepower=. Weight=2875.0 Acceleration=17.0
Missing Observation Brand=Renault Model=Lecar Deluxe MPG=40.9 Cylinders=4 Displacement=85.0 Horsepower=. Weight=1835.0
Acceleration=17.3
Missing Observation Brand=Ford Model=Mustang Cobra MPG=23.6 Cylinders=4 Displacement=140.0 Horsepower=. Weight=2905.0
Acceleration=14.3
Missing Observation Brand=Renault Model=18i MPG=34.5 Cylinders=4 Displacement=100.0 Horsepower=. Weight=2320.0 Acceleration=15.8
Missing Observation Brand=Amc Model=Concord Dl MPG=23.0 Cylinders=4 Displacement=151.0 Horsepower=. Weight=3035.0 Acceleration=20.5
```

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.

The MEANS Procedure

Analysis Variable : Horsepower						
Cylinders	N Obs	N	Mean	Std Dev	Minimum	Maximum
3	4	4	99.2500000	8.3016063	90.0000000	110.0000000
4	204	199	78.2814070	14.5230992	46.0000000	115.0000000
5	3	3	82.3333333	18.5831465	67.0000000	103.0000000
6	84	83	101.5060241	14.3104716	72.0000000	165.0000000
8	103	103	158.3009709	28.4535517	90.0000000	230.0000000

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.

3.3. 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.

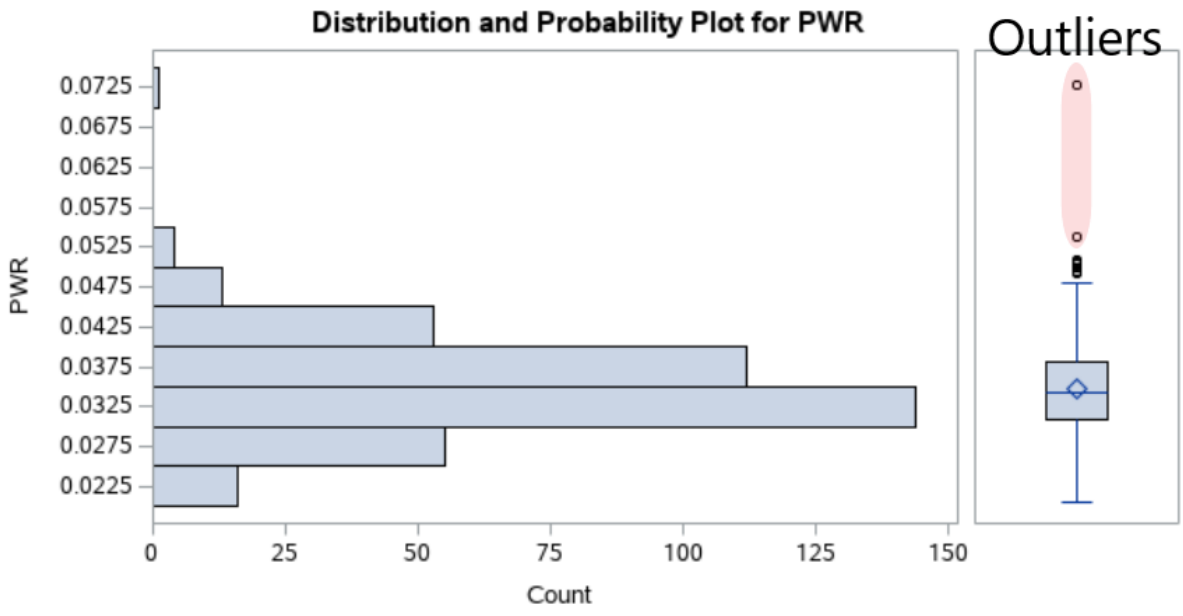

Figure 8: Power to weight ratio box plot

3.4. 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.

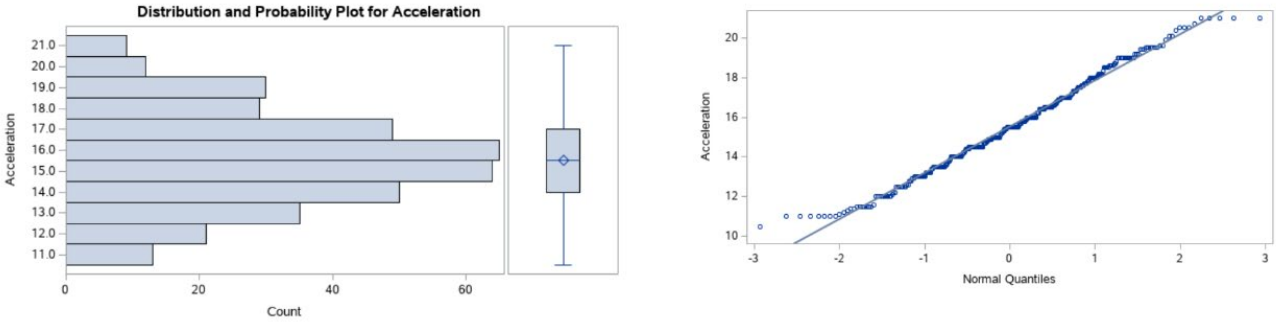

Figure 9: Distribution for Acceleration

Listing Outliers for Acceleration

```
Outlier detected for Volkswagen Type 3  where Acceleration = 23.5
Outlier detected for Chevrolet Chevette  where Acceleration = 22.2
Outlier detected for Chevrolet Woody  where Acceleration = 22.1
Outlier detected for Peugeot 504  where Acceleration = 21.9
Outlier detected for Volkswagen Rabbit Custom Dies  where Acceleration = 21.5
Outlier detected for Peugeot 504  where Acceleration = 24.8
Outlier detected for Volkswagen Rabbit C Diesel  where Acceleration = 21.7
Outlier detected for Volkswagen Dasher Diesel  where Acceleration = 23.7
Outlier detected for Mercedes Benz 240d  where Acceleration = 21.8
Outlier detected for Volkswagen Pickup  where Acceleration = 24.6
Outlier detected for Ford Galaxie 500  where Acceleration = 10.0
Outlier detected for Chevrolet Impala  where Acceleration = 9.0
Outlier detected for Plymouth Fury Iii  where Acceleration = 8.5
Outlier detected for Pontiac Catalina  where Acceleration = 10.0
Outlier detected for Amc Ambassador Dpl  where Acceleration = 8.5
Outlier detected for Dodge Challenger Se  where Acceleration = 10.0
Outlier detected for Plymouth cuda 340  where Acceleration = 8.0
Outlier detected for Chevrolet Monte Carlo  where Acceleration = 9.5
Outlier detected for Buick Estate Wagon Sw  where Acceleration = 10.0
Outlier detected for Pontiac Grand Prix  where Acceleration = 9.5
Outlier detected for Oldsmobile Cutlass Salon Brou  where Acceleration = 22.2
```

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.

Listing Outliers for Power-Weight Ratio

```
Outlier detected for Bmw 2002  Power-Weight ratio = 0.0505819158
Outlier detected for Chevrolet Impala  Power-Weight ratio = 0.0505282499
Outlier detected for Plymouth Fury Iii  Power-Weight ratio = 0.0498608534
Outlier detected for Pontiac Catalina  Power-Weight ratio = 0.0508474576
Outlier detected for Amc Ambassador Dpl  Power-Weight ratio = 0.0493506494
Outlier detected for Buick Estate Wagon Sw  Power-Weight ratio = 0.0729099157
Outlier detected for Pontiac Grand Prix  Power-Weight ratio = 0.0537634409
Outlier detected for Oldsmobile Omega  Power-Weight ratio = 0.0491266376
```

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

3.5. 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.

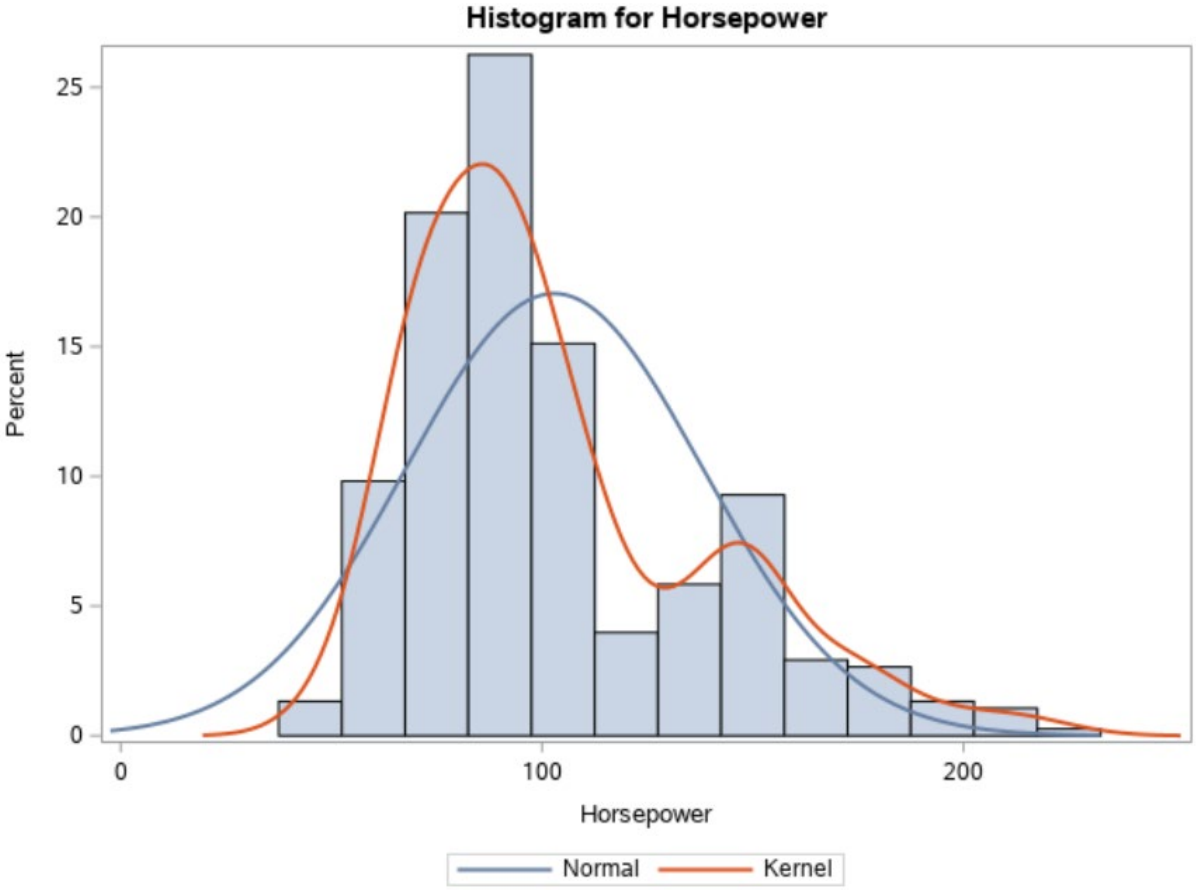

Figure 12: Horsepower distribution before log transformation

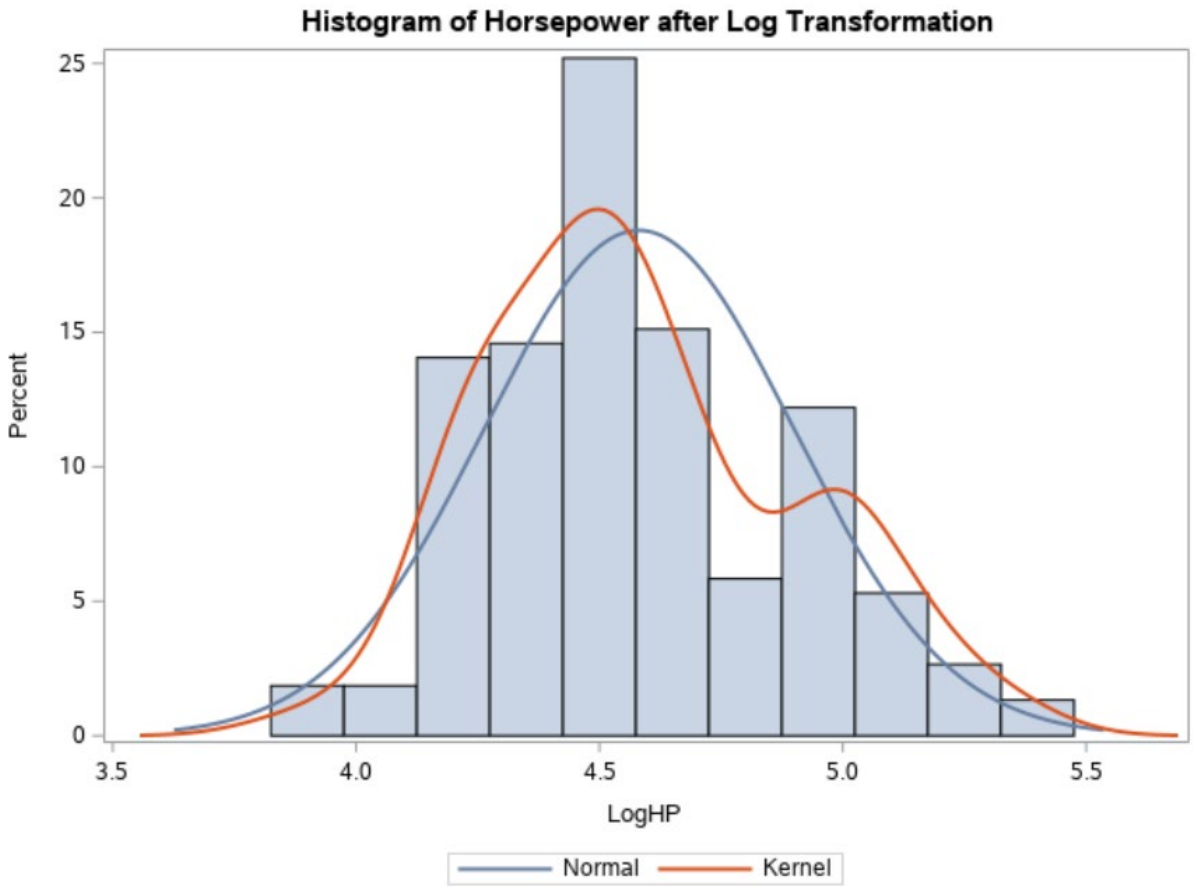

Figure 13: Horsepower distribution after log transformation

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

4. 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.