

## **Data Analysis on SPSS**

B.Sc. Data Science

**- SHAHIL BISWAS**

Roll Number: 29254322008

SEMESTER IV

Supervisor(s):

Mr. Ashok Gupta



NETAJI SUBHASH ENGINEERING COLLEGE, KOLKATA

&

BSE (BOMBAY STOCK EXCHANGE) INSTITUTE, KOLKATA

2022- 2025

## Report Approval

This report entitled “Data Analysis of Cars on SPSS” by Shahil Biswas is approves for the degree of B.Sc. in Data Science.

Examiners:

---

---

---

Supervisor(s):

---

---

---

Date: \_\_\_\_\_

Place: \_\_\_\_\_

**CERTIFICATE OF AUTHRNTICITY**  
**NETAJI SUBHASH ENGINEERING COLLEGE & BSE (BOMBAY STOCK EXCHANGE)**  
**INSTUTUTE**  
**KOLKATA**

**CERTIFICATE**

To whomever it may concern

This is to certify that the work entered in this journal is the work of

**SHAHIL BISWAS**, has successfully completed a project report on the

**Data Analysis of Cars dataset on SPSS** topic terms of the **2022-2025**, in the institute as laid down by the  
institute authority.

Internal Guide \_\_\_\_\_

Project Co-Ordinator \_\_\_\_\_

Examiner \_\_\_\_\_

Date: \_\_\_\_\_

## **Declaration**

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and reference the sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---

Signature

---

Name of the Student

---

Roll No.

Date: \_\_\_\_\_

## **Acknowledgements**

I would like to take the opportunity to give a vote of thanks to all those who supported me with this report project. Thanking all my batch mates who extended their support to the project through cooperation and providing information. I would like to express my gratitude to my mentor, **Mr. Ashok Gupta**, for being the go-to person for information on any doubts or queries that I had while working on this project. I enjoyed myself while doing this project report where I acquired immense amount of knowledge regarding my own topic that I would not even have imagined would correlate with the same.

**Thank You.**

## TABLE OF CONTENTS

SL no.	Topic	Page No.	Remarks
1	Introduction	1	
1(a)	Need	1	
1(b)	State of problem	1	
1(c)	Significance	1	
1(d)	Objective	2	
1(e)	Abstract	2	
1(f)	Limitation	2	
1(g)	Key words	2	
2	Investigation of Dataset	4	
3	Data Analysis on SPSS	8	
4	Conclusion	38	

## **Introduction**

In today's data-driven world, the ability to analyze and interpret data is an essential skill across various fields, including business, healthcare, social sciences, and engineering. This project aims to introduce students to the fundamentals of data analysis using SPSS (Statistical Package for the Social Sciences) software. SPSS is a powerful tool widely used by researchers and data scientists to perform complex statistical analyses and derive meaningful insights from data. By working on this project, students will gain hands-on experience in data manipulation, statistical testing, and result interpretation, preparing them for real-world data analysis challenges.

## **Need**

In an era where data is abundant and increasingly integral to decision-making processes, understanding how to analyze and interpret data is crucial. SPSS, a widely-used statistical analysis tool, provides an accessible platform for students to learn these skills. This project is necessary to bridge the gap between theoretical knowledge and practical application, helping students to develop proficiency in data analysis and to prepare them for careers in research, data science, and various other fields where data plays a pivotal role.

## **State of Problem**

The problem addressed by this project is the lack of practical experience in data analysis among students. While many students are familiar with theoretical statistical concepts, they often lack the hands-on experience needed to apply these concepts using software tools like SPSS. This project aims to address this gap by providing structured guidance and practical exercises that will enable students to become proficient in data analysis and interpretation using SPSS.

## **Significance**

The significance of this project lies in its potential to enhance students' analytical capabilities and prepare them for real-world challenges. By mastering SPSS, students will be able to conduct thorough data analyses, contributing to research and evidence-based decision-making in their respective fields. Additionally, the skills gained from this project are highly transferable and sought after in the job market, giving students a competitive edge in their careers.

## **Objective**

The primary objectives of this project are:

- To familiarize students with SPSS software and its capabilities.
- To teach the basics of data analysis, including data cleaning, descriptive statistics, and inferential statistics.
- To apply SPSS to analyze a given dataset, drawing meaningful conclusions from the results.
- To understand the role of data analysis in research and its application in various fields.
- To develop critical thinking and analytical skills through practical data analysis exercises.

## **Abstract**

This project provides an in-depth exploration of data analysis using SPSS software. Through practical application, students will learn to navigate SPSS, perform essential data manipulations, and execute statistical tests. The project will cover topics such as data cleaning, descriptive statistics, correlation analysis, and hypothesis testing. By analyzing a provided dataset, students will draw conclusions and present their findings. This hands-on experience aims to equip students with the necessary skills to perform independent data analyses and understand the importance of statistical tools in research and decision-making.

## **Limitation**

While this project provides a comprehensive introduction to data analysis using SPSS, it is not exhaustive. The scope is limited to fundamental data analysis techniques, and more advanced topics such as multivariate analysis, time series analysis, and machine learning are beyond the scope of this project. Additionally, the project relies on a single dataset, which may limit the exposure to different types of data and analysis techniques. Future learning should include exposure to a variety of datasets and more complex statistical methods.

## **Keywords**

**SPSS:** Statistical Package for the Social Sciences, a software used for data management and statistical analysis.

**Data Analysis:** The process of inspecting, cleansing, transforming, and modeling data to discover useful information and support decision-making.

**Descriptive Statistics:** Statistical methods that summarize and describe the features of a dataset.



**Inferential Statistics:** Techniques that allow conclusions to extend beyond an immediate data set; for example, making inferences about a population based on a sample.

**Hypothesis Testing:** A statistical method used to determine whether there is enough evidence in a sample to infer that a certain condition holds for the entire population.

**Correlation Analysis:** A method used to evaluate the strength and direction of the linear relationship between two quantitative variables.

**Data Cleaning:** The process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset

# Investigation of Dataset ( Data Cleaning, transformation, Loading)

## Investigating the sample data-sets:


We were provided with two files:

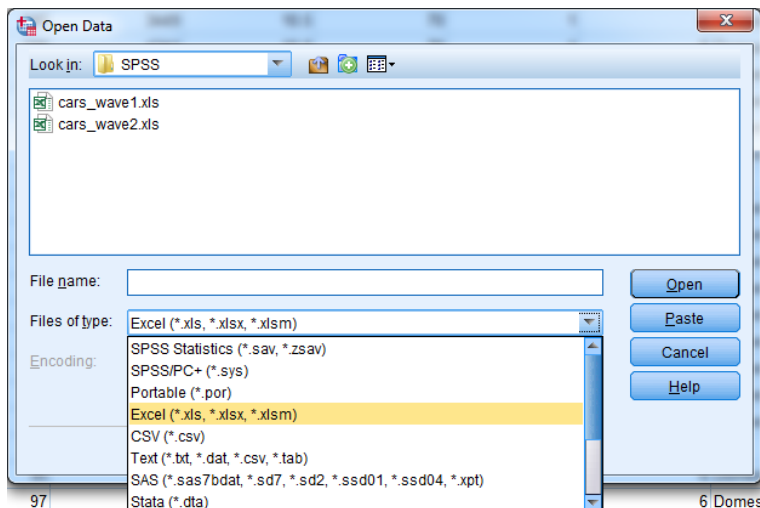
- cars\_wave1.xls
- cars\_wave2.xls

First, you need to clean the data, such as filling up the missing values with some necessary data. Then delete the unwanted rows with no values or unwanted spaces. This is how you can prepare the data for presenting it in SPSS and conduct the further analysis.

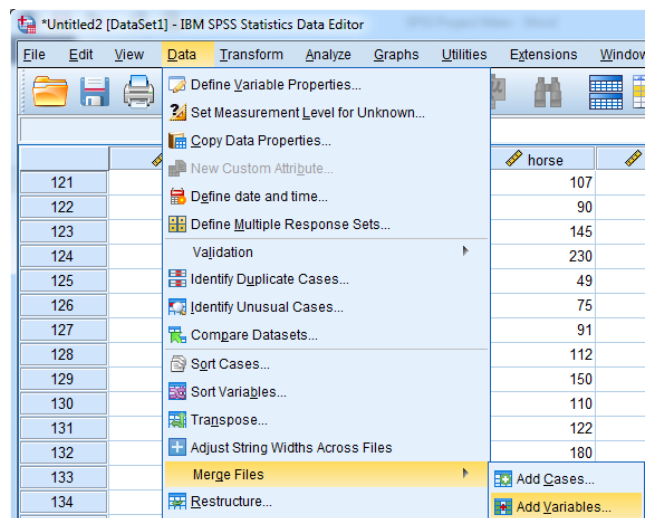
## Import the two files and merge them:

The following steps will help you in importing the two files and merging them:

1. Open the SPSS software and select the  icon from the top left corner. A new window will pop up.
2. Select the folder where you have saved the file and also select the file type.

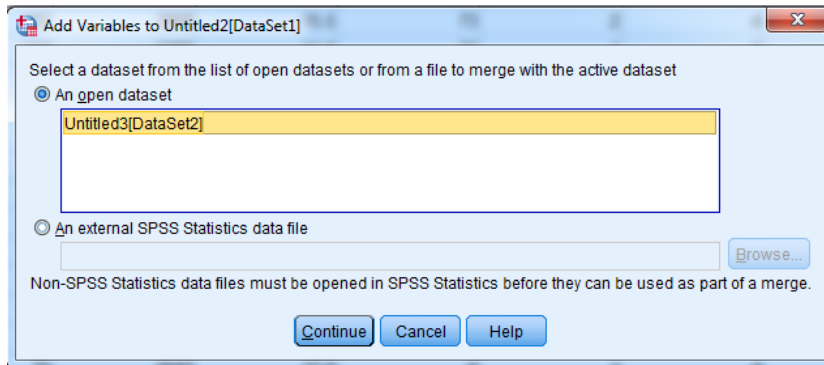


3. Select the file and then click on **Open**.
4. Then repeat the same process for opening the second file.
5. Open the first file on SPSS, then select **Data > Merge Files > Add Variables**.



A new window will pop up.

6. Select the second file and then click on **Continue**.

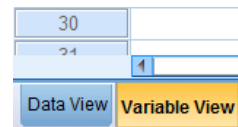


Hence, the two files are merged successfully.

7. Save the new data-set in your desired folder.

### Defining the proper attributes:


1. Select **Variable View** from the bottom left corner.
2. Then click on the cells and define the attributes under the **Label** section.



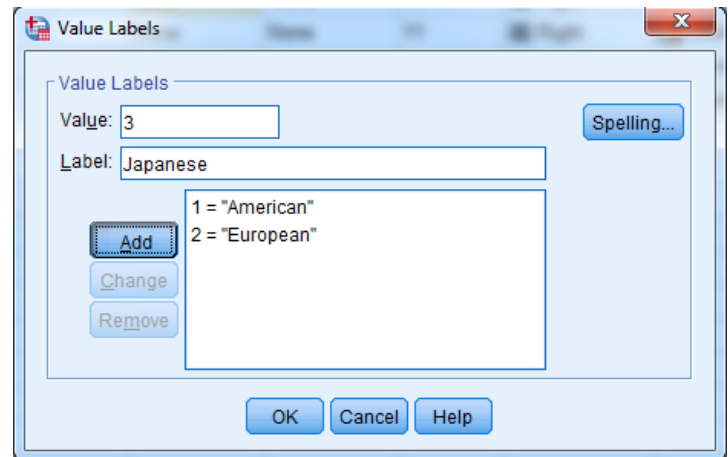
	Name	Type	Width	Decimals	Label	Values
1	ID	Numeric	11	0	Car ID Number	None
2	mpg	Numeric	8	1	Miles Per Gallon	None
3	engine	Numeric	8	1	Engine Displac...	None
4	horse	Numeric	11	0	Horse Power	None
5	weight	Numeric	8	2	Weight (lbs.)	None
6	accel	Numeric	8	1	Time to Acceler...	None
7	year	Numeric	11	0	Model Year (m...	None
8	origin	Numeric	11	0	Country Origin	{1, America...
9	cylinder	Numeric	11	0	Number of Cyli...	{3, 3 Cyli...
10	LP100K	Numeric	8	2		None

3. Also change the **Measure** section according to the given data.

Now, we also need to assign Labels for the code numbers of origin (1,2 and 3). Follow the following steps perform it:

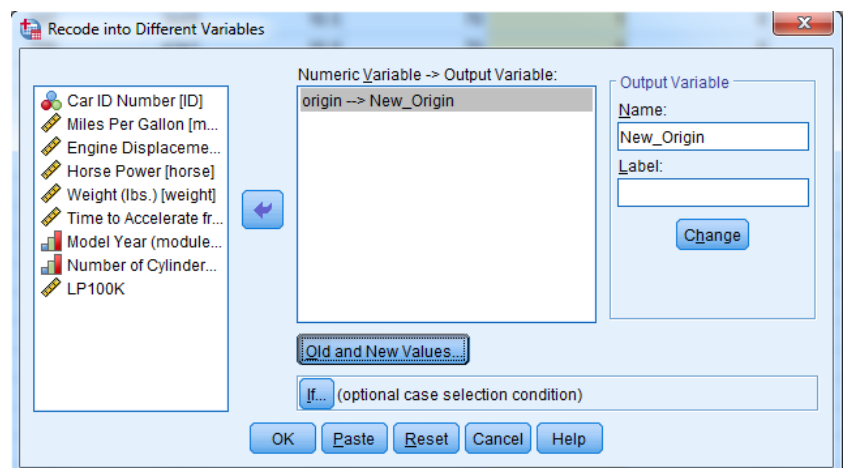
1. Under the Values section, in the row of **origin**, click  on icon. A new window will pop up.

2. Enter the number in the **Value** box and the assign a Label in the **Label** box, as shown below.
3. Click on **Add** after each insertion, and then click on **OK**.
4. Repeat the same process for the row of **cylinder**.

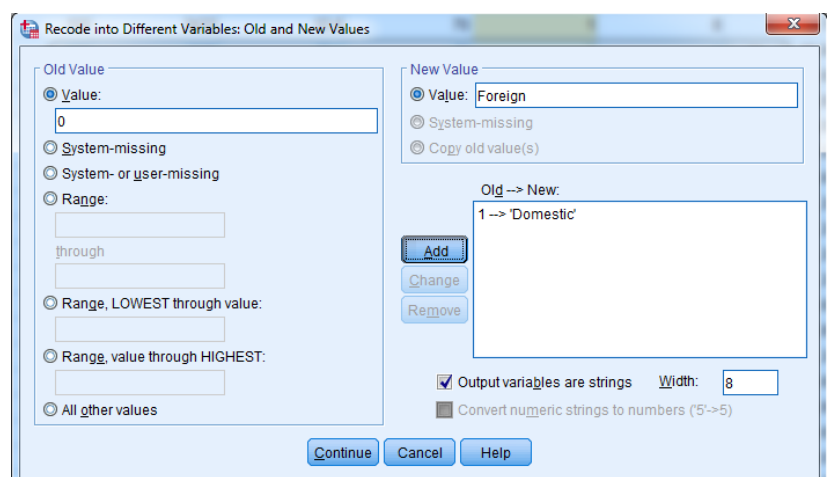


**Recoding origin such that 1=Domestic, 0=Foreign:**

1. Click on **Transform > Recode into Different Variable**. A new window will pop up.
2. Select **origin** and then import it in the **Numeric Variable -> Output Variable** box.
3. Under the **Output Variable** section, write a new name as shown above, and click on **Change**.



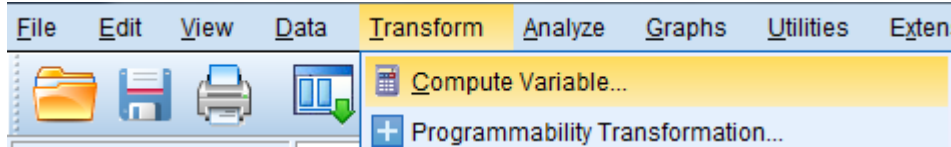
4. Click on **Old and New Values**. A new window will appear.
5. Check the box named **Output variables are strings**.
6. Now, under the Old Value type the value and under the New Value type the name, as shown above.
7. Click on **Add**. Repeat the process for adding the second value.
8. Click on **Continue**, then **OK**.



You will see, a new column appears with the assigned values.

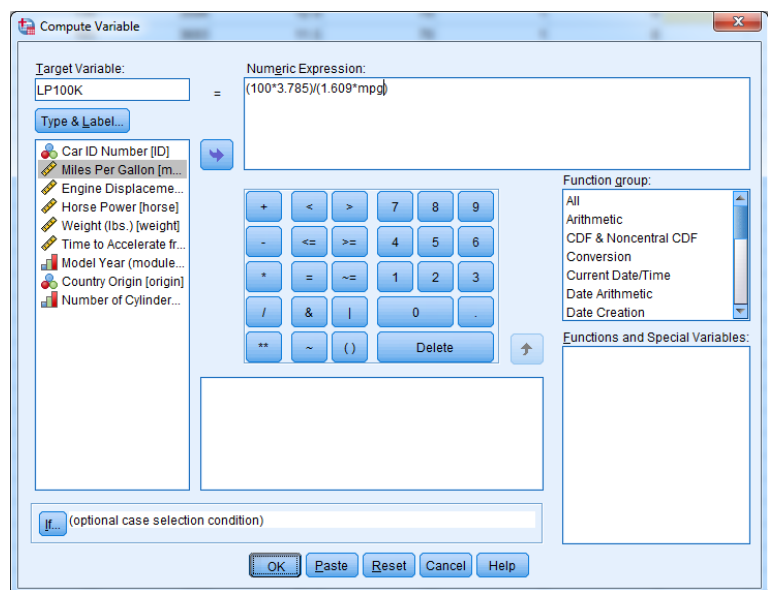
## Converting Miles Per Gallon (mpg) to Liters Per 100 Kilometers:

1. Select **Transform > Compute Variable**. A new menu will appear.



1. Under the **Target Variable** section, write **LP100K**. Then under the **Numeric Expression** section, enter the formula:  $(100 * 3.785) / (1.609 * \text{mpg})$

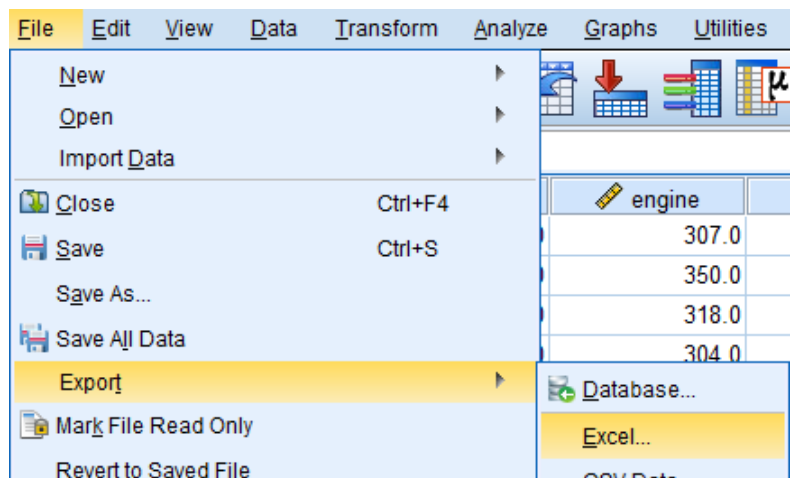
2. Click on **OK**. Hence, you can see a new column as **LP100K**.



Now you can export the data in excel file. Follow the following steps to do so:

1. Select **File > Export > Excel**
2. Select the desired folder and then click on **Save**.

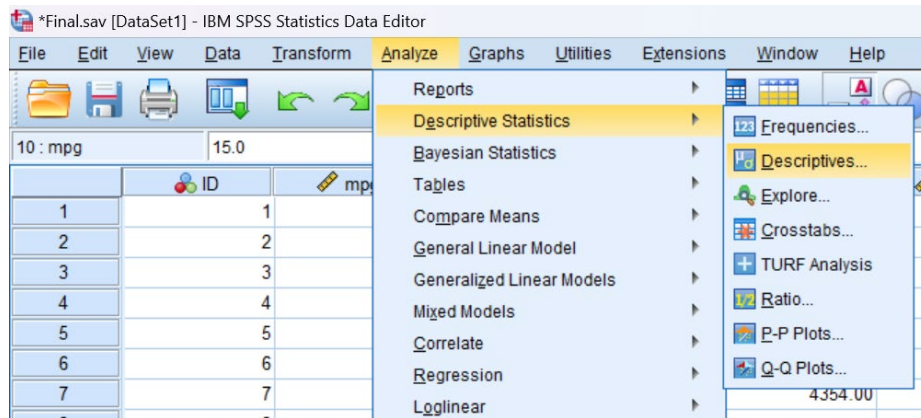
It is always good to have a backup of your file.




## Data Analysis on SPSS

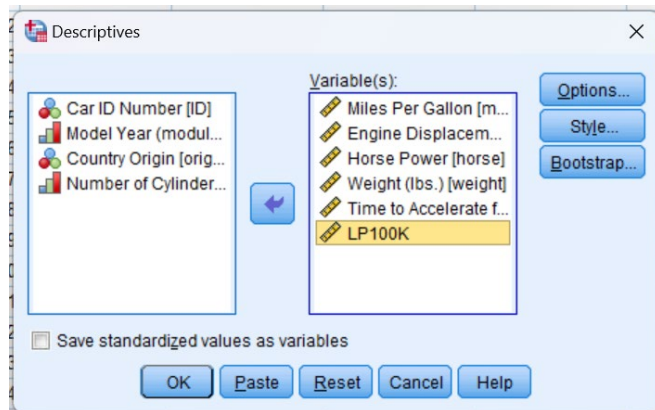
1. Get descriptive statistics for all scale variables in the data set.

- Select on **Analyze** from the Ribbon tab
- Then go to **Descriptive Statistics**
- Then **Descriptives...** and click on it



Then a new window will pop up -

- Transfer all the scale variables to the variable(s) box by selecting and clicking on  button.
- After that click OK.



Variable	Position	Label	Measurement Level	Missing Values
mpg	2	Miles per Gallon	Scale	
engine	3	Engine Displacement (cu. Inches)	Scale	
horse	4	Horsepower	Scale	
weight	5	Weight (lbs.)	Scale	
accel	6	Time to Accelerate from 0 to 60 mpg (sec)	Scale	

After that new output file will be created with all **Descriptive Statistics** will be generated in it.

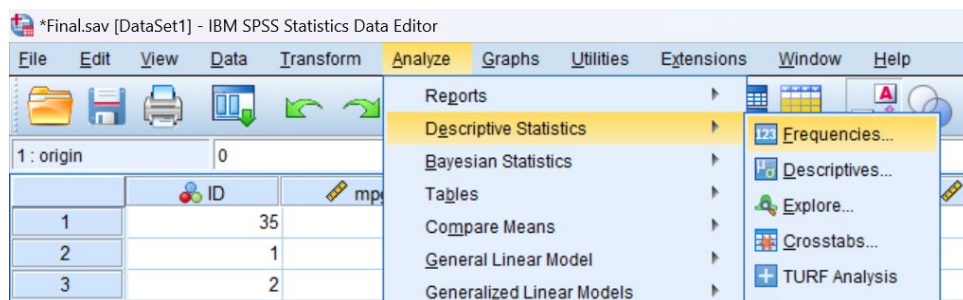
DESCRIPTIVES VARIABLES=mpg engine horse weight accel LP100K  
/STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS.

## Descriptives


Descriptive Statistics											
	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error
	c	c	c	c	Statistic	Statistic	Error	Statistic	Statistic	c	Error
Miles Per Gallon	406	990.0	9.0	999.0	17350.8	42.736	6.7477	135.9632	18485.989	46.035	.242
Engine Displacement (cu.inches)	406	451.0	4.0	455.0	78780.5	194.041	5.2214	105.2074	11068.589	-.791	.242
Horse Power	400	184	46	230	41933	104.83	1.926	38.522	1483.949	.591	.243
Weight (lbs.)	406	4408.00	732.00	5140.00	1205642.00	2969.5616	42.17621	849.82717	722206.212	-.752	.242
Time to Accelerate from 0 to 60 mpg (sec)	406	16.8	8.0	24.8	6291.0	15.495	.1400	2.8210	7.958	.389	.242
LP100K	406	25.90	.24	26.14	4465.15	10.9979	.20620	4.15483	17.263	.419	.242
Valid N (listwise)	400										

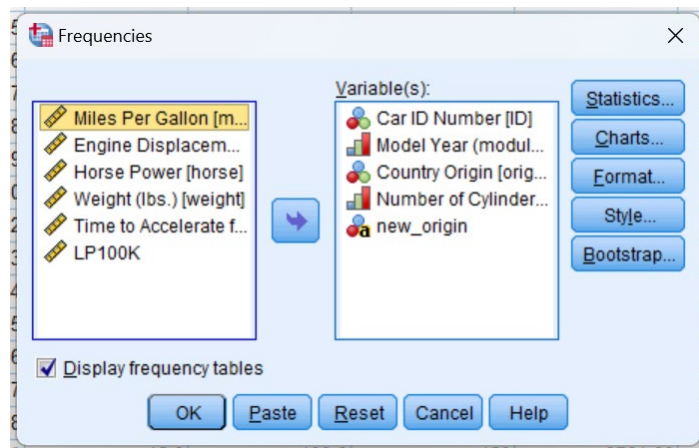
2. Get frequency tables for all categorical variables (ordinal or nominal) in the data set.

- Select on Analyze from the Ribbon tab
- Then go to **Descriptive Statistics**
- Then click on **Frequencies...**



Then a new window will pop up -

- Transfer all the scale variables to the variable(s) box by selecting and clicking on  button.
- After that click OK.



Categorical values are :

Variable	Position	Label	Measurement Level	Missing Values
ID	1	Car ID Number	Nominal	
year	7	Model Year (module 100)	Ordinal	
origin	8	Country of Origin	Nominal	
cylinder	9	Number of Cylinders	Ordinal	

Output –

FREQUENCIES VARIABLES=ID year origin cylinder new\_origin  
/ORDER=ANALYSIS.

## Frequencies

		Statistics				
		Car ID Number	Model Year (module 100)	Country Origin	Number of Cylinders	new_origin
N	Valid	406	406	406	406	406
	Missing	0	0	0	0	0



## Frequency Table

		Car ID Number			Cumulative Percent
		Frequency	Percent	Valid Percent	
Valid	1	1	.2	.2	.2
	2	1	.2	.2	.5
	3	1	.2	.2	.7
	4	1	.2	.2	1.0
	5	1	.2	.2	1.2
	6	1	.2	.2	1.5
	7	1	.2	.2	1.7
	8	1	.2	.2	2.0
	9	1	.2	.2	2.2
	10	1	.2	.2	2.5
	11	1	.2	.2	2.7
	12	1	.2	.2	3.0
	15	1	.2	.2	3.7
	403	1	.2	.2	99.3
	404	1	.2	.2	99.5
	405	1	.2	.2	99.8
	406	1	.2	.2	100.0
	Total	406	100.0	100.0	

[All cases are purposely not shown here in the above table since it's a huge data, so only the starting and ending portion been shown here]

		Model Year (module 100)			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	.5	.5	.5
	70	34	8.4	8.4	8.9
	71	29	7.1	7.1	16.0
	72	28	6.9	6.9	22.9
	73	40	9.9	9.9	32.8
	74	26	6.4	6.4	39.2
	75	30	7.4	7.4	46.6
	76	34	8.4	8.4	54.9
	77	28	6.9	6.9	61.8
	78	36	8.9	8.9	70.7
	79	29	7.1	7.1	77.8
	80	29	7.1	7.1	85.0
	81	30	7.4	7.4	92.4

82	31	7.6	7.6	100.0
Total	406	100.0	100.0	

### Country Origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	.2	.2	.2
	American	253	62.3	62.3	62.6
	European	73	18.0	18.0	80.5
	Japanese	79	19.5	19.5	100.0
	Total	406	100.0	100.0	

### Number of Cylinders

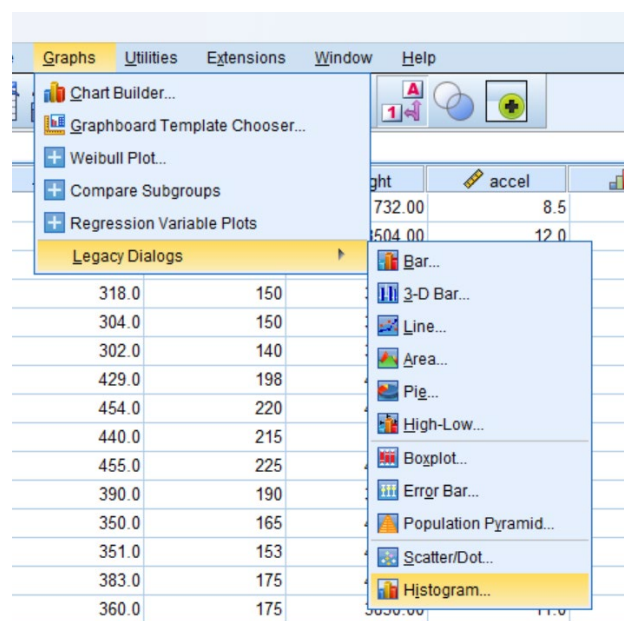
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3 Cylinders	5	1.2	1.2	1.2
	4 Cylinders	207	51.0	51.0	52.2
	5 Cylinders	3	.7	.7	53.0
	6 Cylinders	84	20.7	20.7	73.6
	8 Cylinders	107	26.4	26.4	100.0
	Total	406	100.0	100.0	

### new\_origin


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		152	37.4	37.4	37.4
	Domestic	253	62.3	62.3	99.8
	Foreign	1	.2	.2	100.0
	Total	406	100.0	100.0	

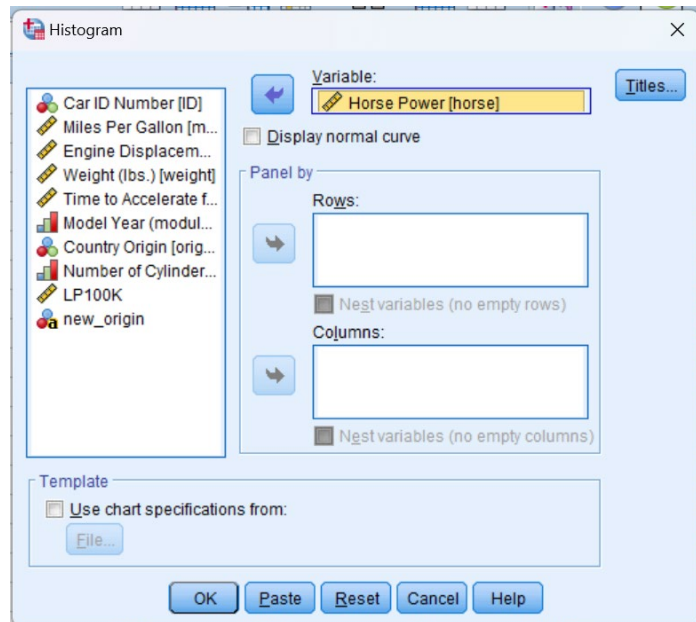
### 3. Create a histogram of Horsepower.

- First we need to go to **Graphs** option on the ribbon bar.
- Then go to **Legacy/Dialogs**.
- Then **Histogram** and click on it.



After that new window will open-

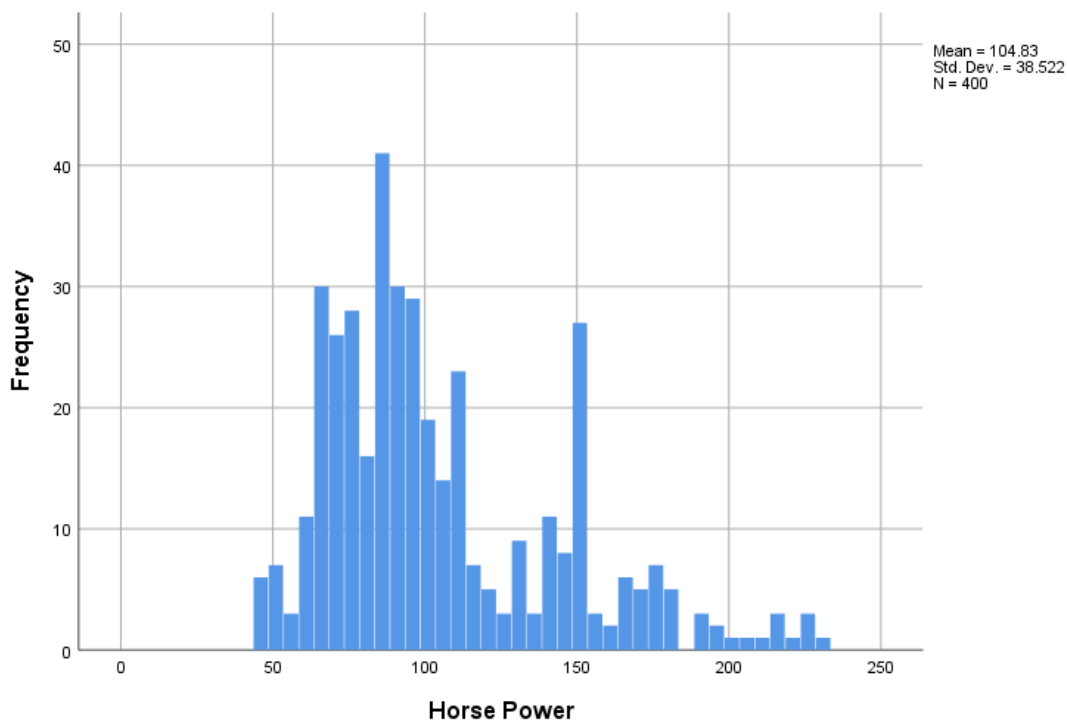
- Select **Horse Power** variable and move it to variable section with help of  button.
- Keep even as default and then click OK.



Output :

GRAPH  
/HISTOGRAM=horse.

## Graph



## Statistics

Horse Power


N	Valid	400
	Missing	6

#### 4. Create a histogram of Weight.

The process is same like question no.3

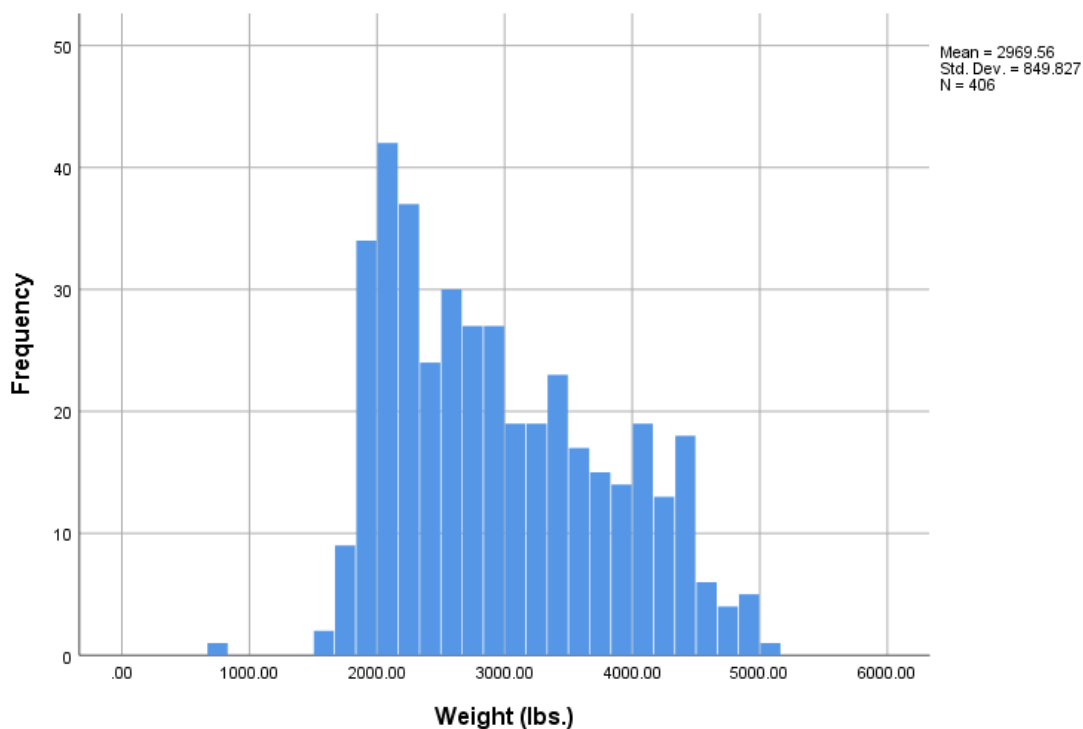
- First we need to go to **Graphs** option on the ribbon bar.
- Then go to **Legacy/Dialogs**.
- Then **Histogram** and click on it.

After that new window will open-

- Select **Weight (lbs.)** variable and move it to variable section with help of  button.
- Keep even as default and then click OK.

GRAPH  
/HISTOGRAM=weight.

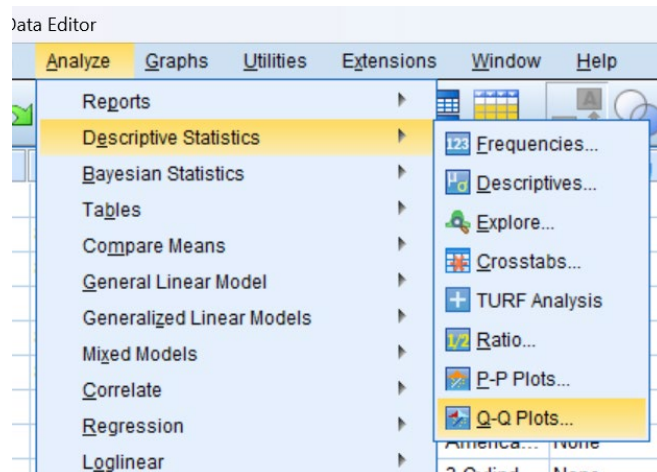
#### Graph



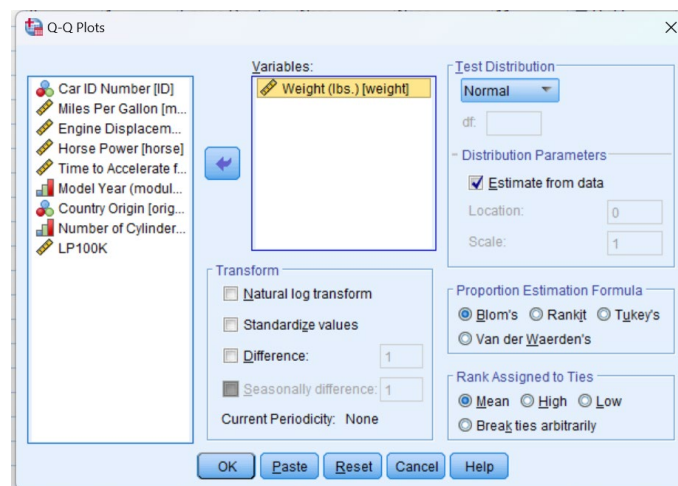
5. Create a QQ-Plot for Weight (Analyze Descriptive Statistics QQ Plot Select Weight, leave others as default settings OK)


[The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.]

- Go to **Analyze**
- Then **Descriptive Statistics**
- Then **Q-Q Plots**



New window will open like shown below :



- Select **Weight(lbs.)[Weight]** and move to Variables box by using  button.
- Then press OK.

Output:

```
PLOT
/VARIABLES=weight
/NOLOG
/NOSTANDARDIZE
/TYPE=Q-Q
/FRACTION=BLOM
/TIES=MEAN
/DIST=NORMAL.
```

## PPlot

### Model Description

Model Name		MOD_1
Series or Sequence	1	Weight (lbs.)
Transformation		None
Non-Seasonal Differencing		0
Seasonal Differencing		0
Length of Seasonal Period		No periodicity
Standardization		Not applied
Distribution	Type	Normal
	Location	estimated
	Scale	estimated
Fractional Rank Estimation Method		Blom's
Rank Assigned to Ties		Mean rank of tied values

Applying the model specifications from MOD\_1

### Case Processing Summary

		Weight (lbs.)
Series or Sequence Length		406
Number of Missing Values in the	User-Missing	0
	System-Missing	0

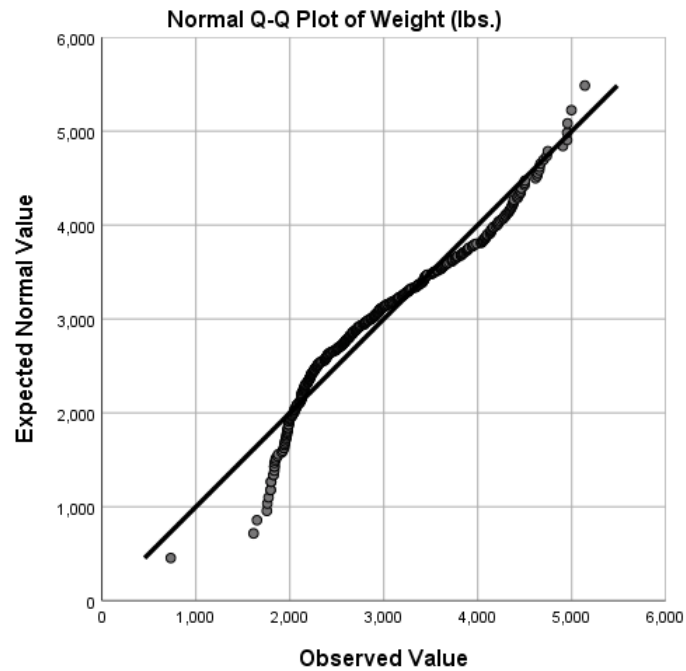
The cases are unweighted.

### Estimated Distribution Parameters

		Weight (lbs.)
Normal Distribution	Location	2969.5616
	Scale	849.82717

The cases are unweighted.

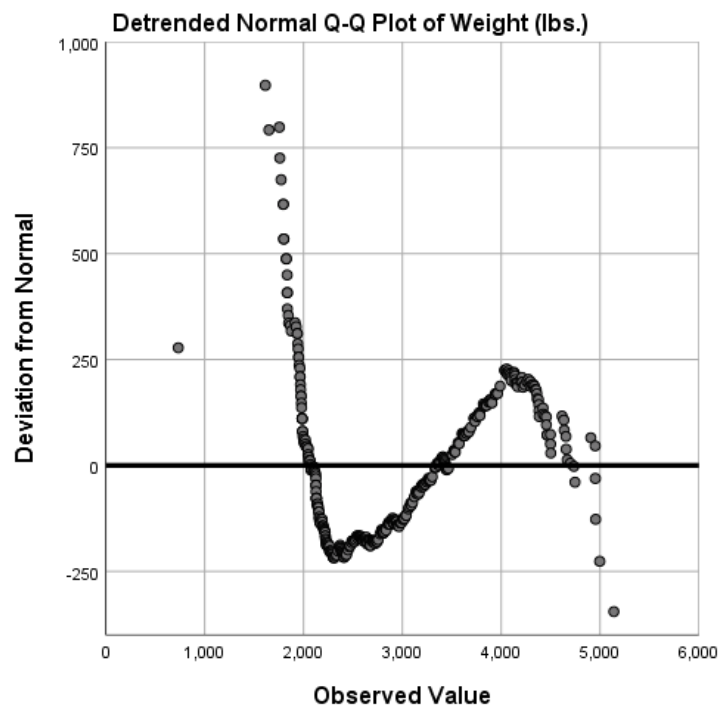
## Weight (lbs.)



Fig(1.1)

Points on the Normal QQ plot provide an indication of univariate normality of the dataset. If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.

So from seeing the fig(1.1) its evident that the sample is not Normally distributed because as most of the data points of the dataset are away from the reference line.



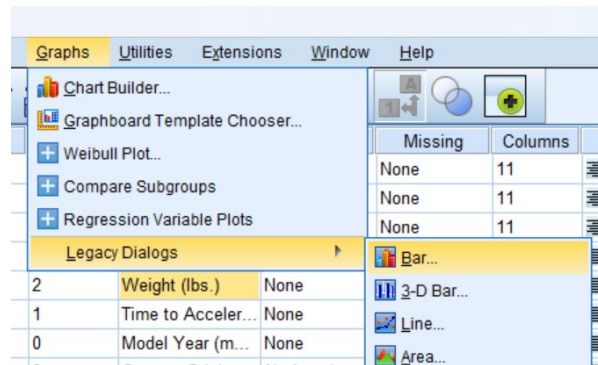
Fig(1.2)

The detrended normal Q-Q plot depicts the differences between the observed and the predicted values. When the distribution of the values of the dependent variable is normal then the values of the difference between observed and predicted fall randomly about the zero line.

Here from the above fig(1.2) we can easily conclude that the sample is not normally distributed since most of the points are away from the zero line

6. Create a bar chart for Origin.

- Select **Graphs**
- Then **Legacy Dialogs**
- And lastly **Bar...**

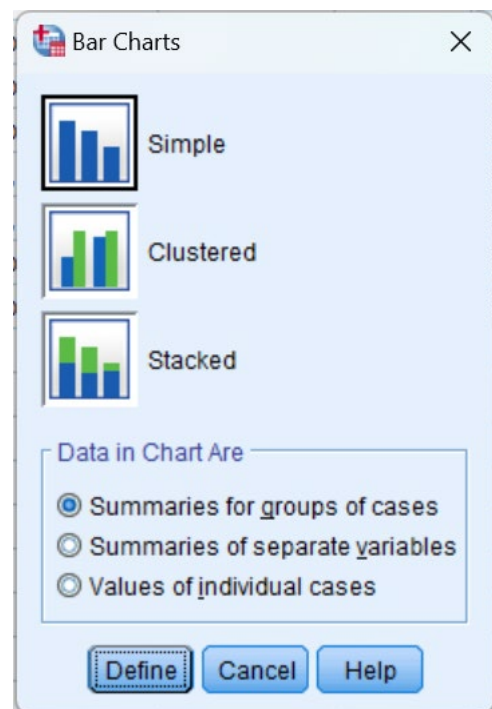


New window will open like show below:

- Keep everything as default and click **Define**.

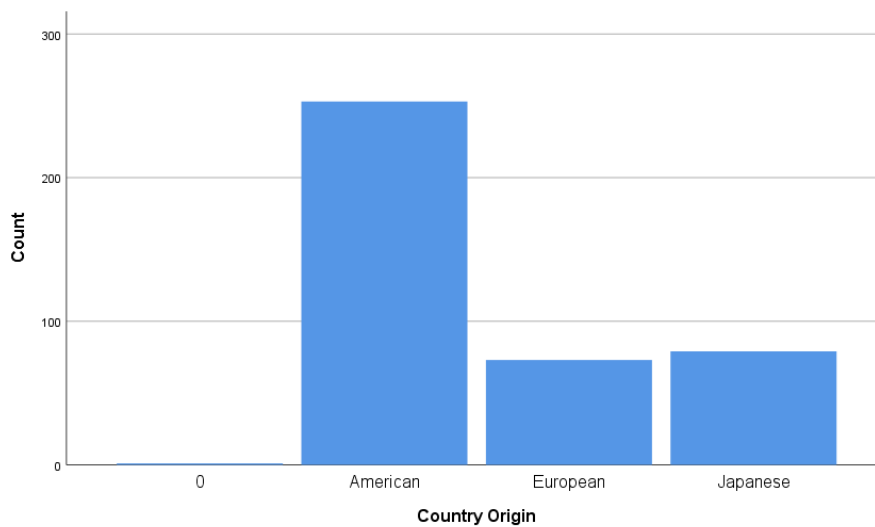
Output:

GRAPH  
/BAR(SIMPLE)=COUNT BY origin.





## Graph



FREQUENCIES VARIABLES=origin  
/ORDER=ANALYSIS.

## Frequencies

### Statistics

Country Origin

N	Valid	406
	Missing	0

### Country Origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	.2	.2	.2
	American	253	62.3	62.3	62.6
	European	73	18.0	18.0	80.5
	Japanese	79	19.5	19.5	100.0
	Total	406	100.0	100.0	

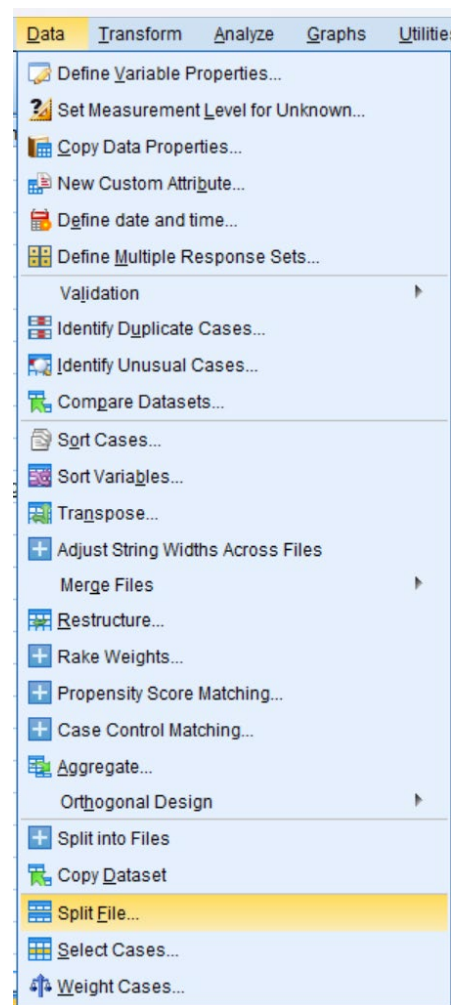
**7. Organize the output by Year (Analyzing groups of cases separately, compare groups). Before proceeding, select only cases with Year not = 0.)**

- Investigate Horsepower (descriptive statistics)
- Investigate Weight (descriptive statistics)
- What do you see?

Before we proceed with this question, we need to do some prior changes and setup.

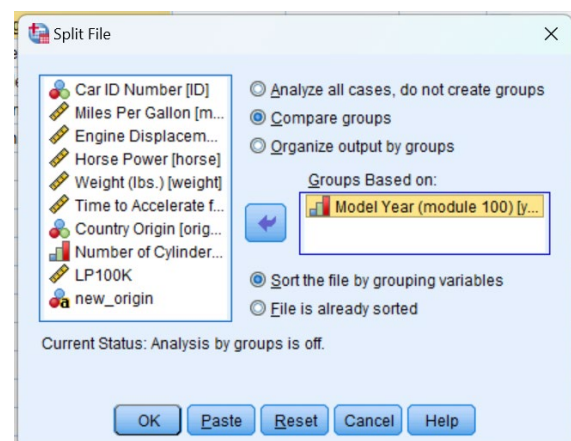
Firstly we need to **split file**, so that we can analyzing groups of cases separately, here it would be **Model year[year]**Steps:

- Go to the **Data** option of the Ribbon bar.
- Then go for **Split File** at the bottom.



New window will open:

- Now from here select **Compare groups**.
- And then select **Year** move it to **Groups Based on**
- Rest keep default
- And click OK.



SORT CASES BY year.  
SPLIT FILE LAYERED BY year.

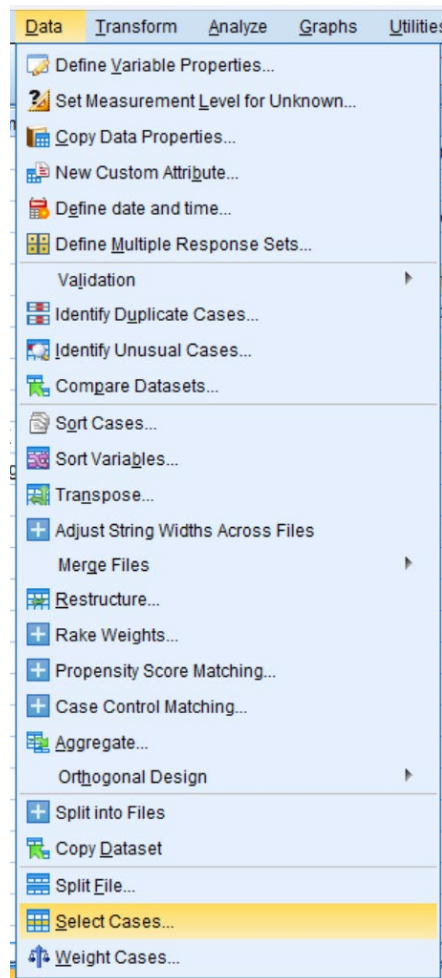
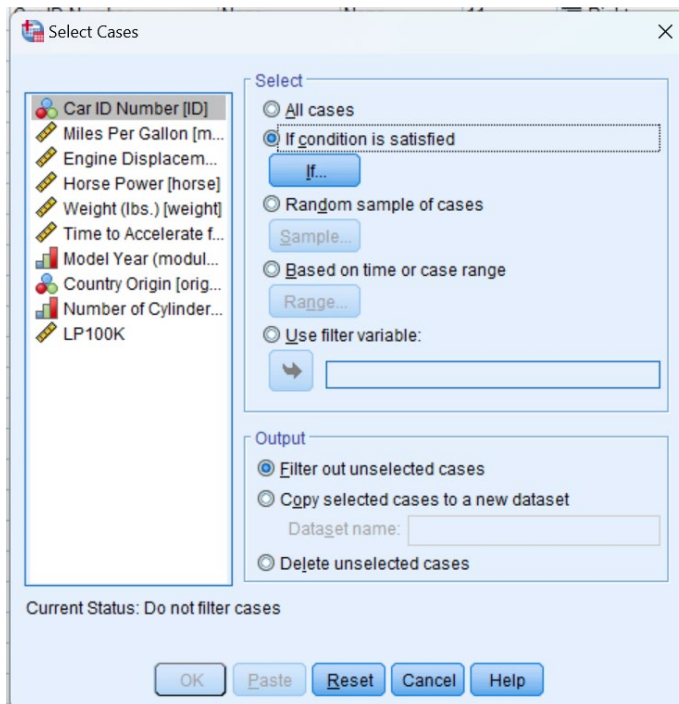
The file has been **sorted by year** and been split

Now next step, we need to **select cases...** like shown in the picture below:

- Go to the **Data** option of the Ribbon bar.
- Then go for **Select Cases..** at the bottom

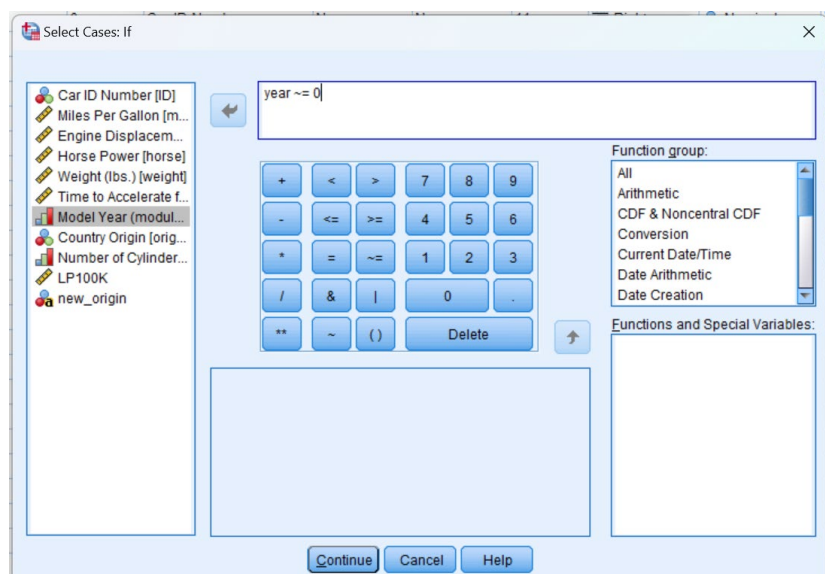
Then new window opens

- First select the “**If condition is satisfied**”
- And click **If** button



Again new window opens shown below:

- Select **Year**.
- And type in the box – **year ~= 0** [ it means : year not equal to zero ]
- Then **Continue**



SORT CASES BY year.  
 SPLIT FILE LAYERED BY year.  
 USE ALL.  
 COMPUTE filter\_\$(year ~= 0).  
 VARIABLE LABELS filter\_\$(year ~= 0 (FILTER)).

VALUE LABELS filter\_\$ 0 'Not Selected' 1 'Selected'.  
 FORMATS filter\_\$ (f1.0).  
 FILTER BY filter\_\$.  
 EXECUTE.

Now we are ready to analysis,

First Analyze Horsepower

Steps :

- Like we have done pervious, go to **Analyze**
- Then go to **Descriptive Statistics**
- Then select **Descriptives...**

Again new window,

- Select **horsepower** and move it to variable(s) and press **OK**.

Output:

DESCRIPTIVES VARIABLES=horse  
 /STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS SKEWNESS.

## Descriptives

Descriptive Statistics													
		N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis
Model Year (module 100)		Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic
70	Horse Power	34	179	46	225	5017	147.56	8.535	49.770	2477.042	-.045	.403	-1.062
	Valid N (listwise)	34											
71	Horse Power	28	132	48	180	2938	104.93	7.456	39.455	1556.735	.809	.441	-.605
	Valid N (listwise)	28											
72	Horse Power	28	154	54	208	3365	120.18	7.771	41.121	1690.967	.371	.441	-.981
	Valid N (listwise)	28											
73	Horse Power	40	184	46	230	5219	130.48	7.338	46.412	2154.102	.315	.374	-.557

	Valid N (listwise)	40												
74	Horse Power	25	98	52	150	2375	95.00	6.007	30.033	902.0 00	.736	.464	-.536	.902
	Valid N (listwise)	25												
75	Horse Power	30	117	53	170	3032	101.0 7	4.852	26.577	706.3 40	.792	.427	.657	.833
	Valid N (listwise)	30												
76	Horse Power	34	128	52	180	3438	101.1 2	5.562	32.431	1051. 743	.634	.403	-.385	.788
	Valid N (listwise)	34												
77	Horse Power	28	132	58	190	2942	105.0 7	6.821	36.095	1302. 884	.935	.441	.090	.858
	Valid N (listwise)	28												
78	Horse Power	36	117	48	165	3589	99.69	4.739	28.436	808.6 18	.268	.393	-.451	.768
	Valid N (listwise)	36												
79	Horse Power	29	90	65	155	2935	101.2 1	5.284	28.456	809.7 41	.365	.434	- 1.260	.845
	Valid N (listwise)	29												
80	Horse Power	27	84	48	132	2092	77.48	3.497	18.173	330.2 59	.942	.448	1.782	.872
	Valid N (listwise)	27												
81	Horse Power	29	62	58	120	2379	82.03	3.449	18.575	345.0 34	.660	.434	-.762	.845
	Valid N (listwise)	29												
82	Horse Power	30	60	52	112	2444	81.47	2.428	13.297	176.8 09	.159	.427	.363	.833
	Valid N (listwise)	30												

**Now similarly we analyze for Weight ,**

Following the same steps like the pervious one

Steps :

- Like we have done pervious, go to **Analyze**
- Then go to **Descriptive Statistics**
- Then select **Descriptives...**

Again new window,

- Select **Weights** and move it to variable(s) and press **OK**.

Outputs:

DESCRIPTIVES VARIABLES=weight

/STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS SKEWNESS.

## Descriptives

Descriptive Statistics													
		N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis
Model Year		Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic
(module 100)		tic	tic	tic	tic	ic	ic	Error	Statistic	Statistic	tic	Error	tic
70	Weight (lbs.)	34	2780.00	1835.00	4615.00	115714.00	3403.3529	135.01741	787.28001	619809.811	-.264	.403	-1.092
	Valid N (listwise)	34											
71	Weight (lbs.)	29	3527.00	1613.00	5140.00	85850.00	2960.3448	196.77684	1059.67573	1122912.663	.663	.434	-.802
	Valid N (listwise)	29											
72	Weight (lbs.)	28	2533.00	2100.00	4633.00	90656.00	3237.7143	184.16715	974.52096	949691.101	.143	.441	-1.887
	Valid N (listwise)	28											
73	Weight (lbs.)	40	3130.00	1867.00	4997.00	136761.00	3419.0250	154.13086	974.80913	950252.846	.107	.374	-1.335
	Valid N (listwise)	40											
74	Weight (lbs.)	26	3050.00	1649.00	4699.00	75162.00	2890.8462	189.38633	965.68461	932546.775	.629	.456	-1.003
	Valid N (listwise)	26											
75	Weight (lbs.)	30	2873.00	1795.00	4668.00	95304.00	3176.8000	139.70208	765.17978	585500.097	.383	.427	-.301
	Valid N (listwise)	30											

	Valid N (listwise)	30												
76	Weight (lbs.)	34	2585. 00	1795. 00	4380. 00	10467 7.00	3078.7 353	140.86 405	821.371 48	674651. 110	-.082	.403	- 1.396	.788
	Valid N (listwise)	34												
77	Weight (lbs.)	28	2510. 00	1825. 00	4335. 00	83926. 00	2997.3 571	172.50 788	912.825 90	833251. 127	.255	.441	- 1.604	.858
	Valid N (listwise)	28												
78	Weight (lbs.)	36	2280. 00	1800. 00	4080. 00	10302 5.00	2861.8 056	104.33 732	626.023 91	391905. 933	-.169	.393	- 1.111	.768
	Valid N (listwise)	36												
79	Weight (lbs.)	29	2445. 00	1915. 00	4360. 00	88605. 00	3055.3 448	138.87 811	747.881 50	559326. 734	-.091	.434	- 1.267	.845
	Valid N (listwise)	29												
80	Weight (lbs.)	29	1546. 00	1835. 00	3381. 00	70663. 00	2436.6 552	80.264 12	432.235 49	186827. 520	.498	.434	-.708	.845
	Valid N (listwise)	29												
81	Weight (lbs.)	30	1970. 00	1755. 00	3725. 00	75965. 00	2532.1 667	96.171 75	526.754 35	277470. 144	.522	.427	-.473	.833
	Valid N (listwise)	30												
82	Weight (lbs.)	31	1070. 00	1965. 00	3035. 00	76060. 00	2453.5 484	63.629 98	354.276 71	125511. 989	.049	.421	- 1.369	.821
	Valid N (listwise)	31												

### Relationship Between Continuous Y (Horsepower) and Continuous X (Weight)

1. Create a Scatter Plot with Horsepower as the Y variable and Weight as the X variable.
  - a. Add a Linear fit line.
  - b. What is the relationship between Horsepower and Weight as shown in this graph?

Before we begin, we need to turn off the **split file mode**.

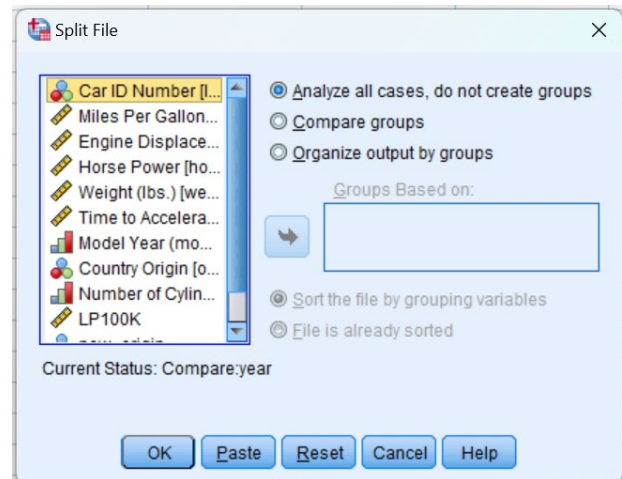
Steps :

- Just go to **Data** option on Ribbon bar

- Then go and select **Split File**
- After that just click on **Reset**

And Split File would be turned off.

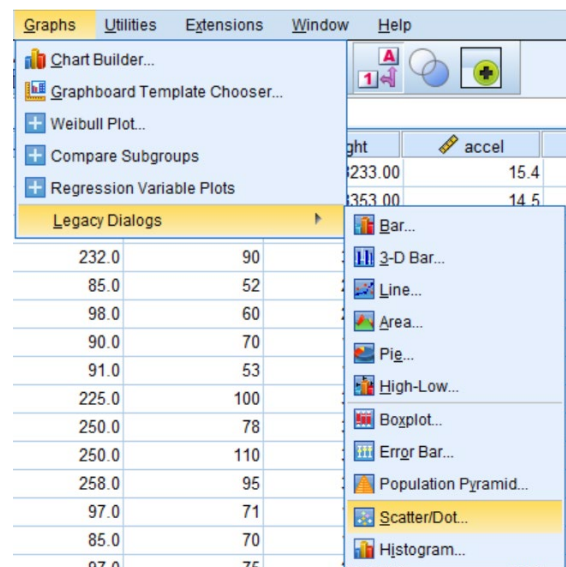
SORT CASES BY year.  
 SPLIT FILE LAYERED BY year.  
 USE ALL.  
 COMPUTE filter\_\$=(year ~= 0).  
 VARIABLE LABELS filter\_\$ 'year ~= 0 (FILTER)'.  
 VALUE LABELS filter\_\$ 0 'Not Selected' 1 'Selected'.  
 FORMATS filter\_\$ (f1.0).  
 FILTER BY filter\_\$.  
 EXECUTE.  
 SPLIT FILE OFF.



Now we are good to go,

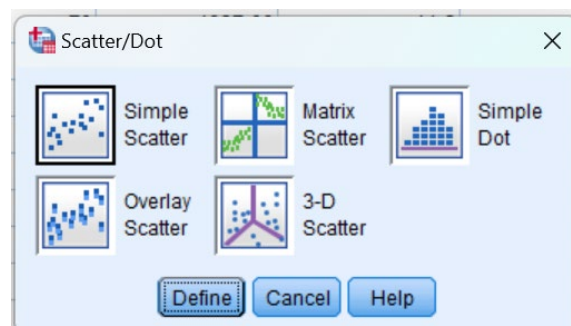
Next step :

- Go to **Graphs**
- Then **Legacy Dialogs** → **Scatter/Dot...**



New window,

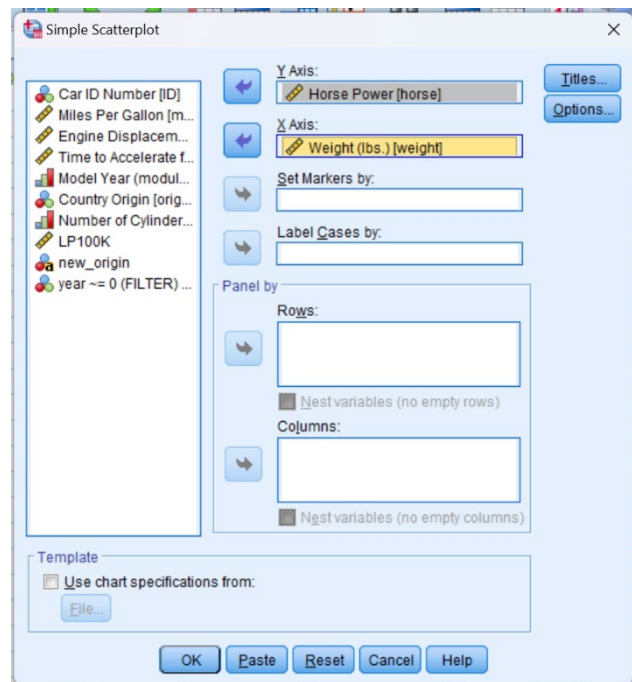
- Make sure you choose **Simple Scatter**
- And click **define**.



Again, new window



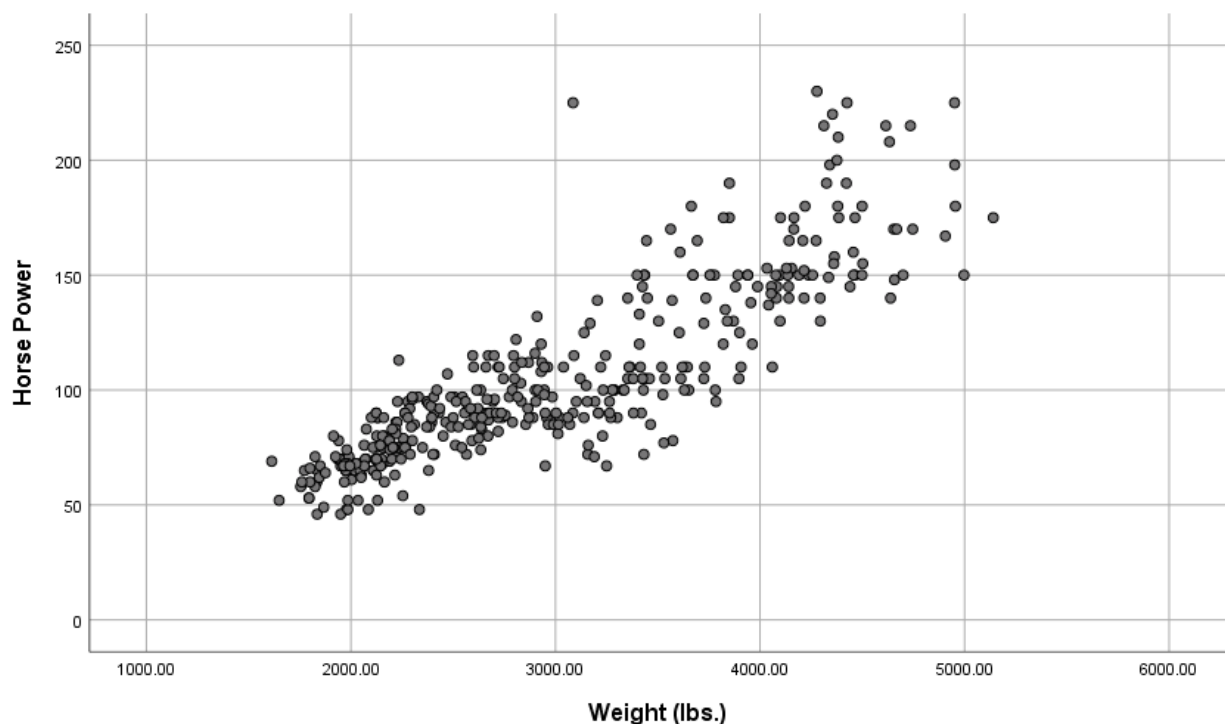
- Here, **Select Horse power** and move it to **Y- Axis**
- And, **Weight** to **X-Axis** as shown below picture
- And lastly press **OK**.



Output :

GRAPH  
 /SCATTERPLOT(BIVAR)=weight WITH horse  
 /MISSING=LISTWISE.

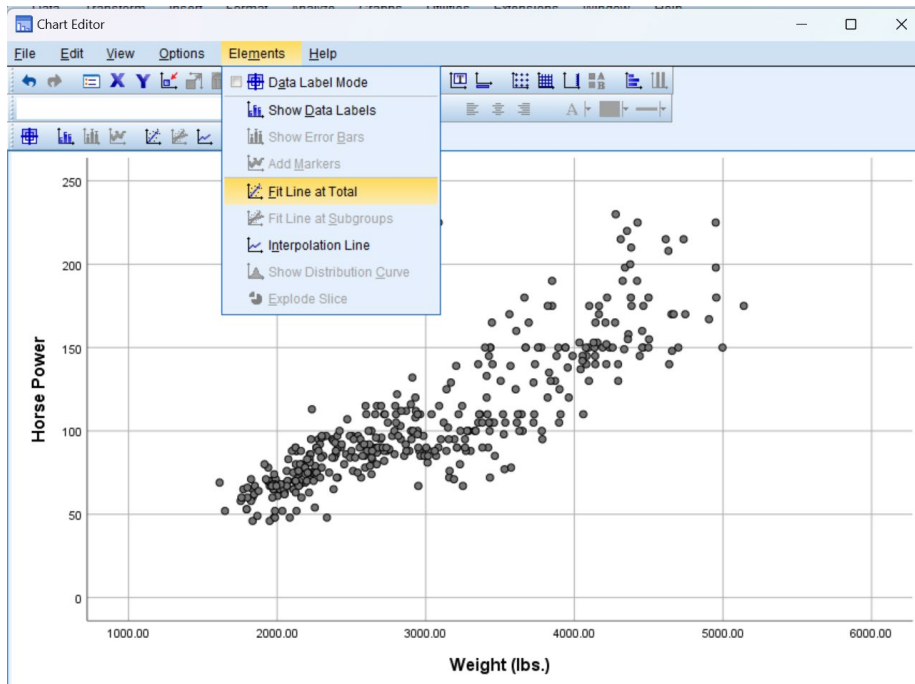
## Graph



Also we have to fit line on the graph:-

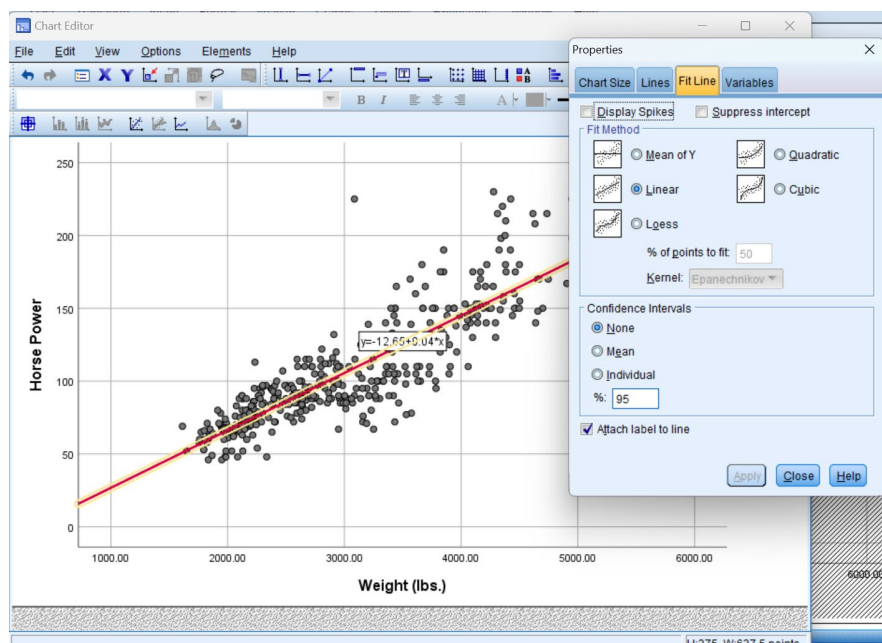
Steps:

- Double Click on the **Graph** & a new window will open
- Then go to **Elements** options and select
- Then **Select Fit Line at Total**

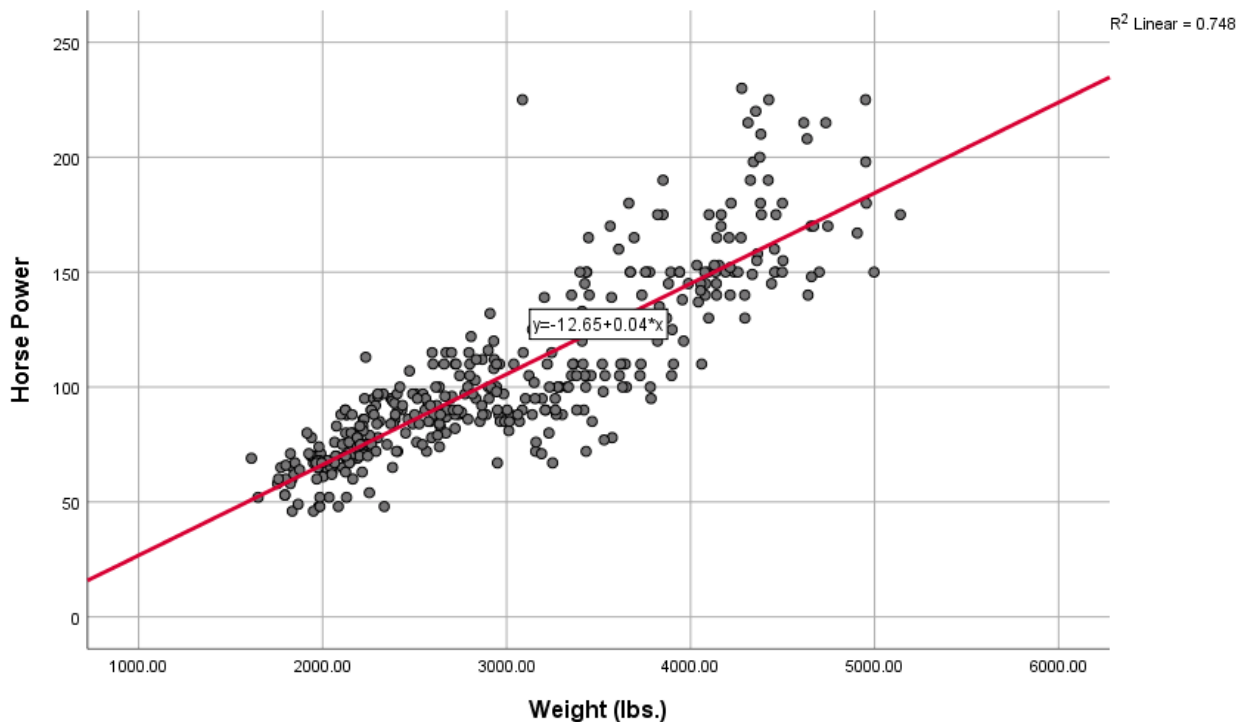


New window,

- After selecting, an line automatically will on the graph
- If we want can edit the line style & colour, we can find options in the properties box



Final Output :



The scatter plot provided shows the relationship between Horsepower and Weight of cars. Here are the key points observed from the graph:

1. **Positive Correlation:** There is a positive correlation between Horsepower and Weight, meaning as the weight of a car increases, its horsepower tends to increase as well.
2. **Linear Relationship:** The red line represents a linear regression line fitting the data points. The equation of the line is  $(y = -12.65 + 0.04x)$ , where  $(y)$  is the Horsepower and  $(x)$  is the Weight. This indicates that for every additional pound of weight, the horsepower increases by approximately 0.04 units.
3. **R-squared Value:** The  $(R^2)$  value of 0.748 suggests that approximately 74.8% of the variance in Horsepower can be explained by the Weight of the car. This indicates a strong linear relationship between the two variables.

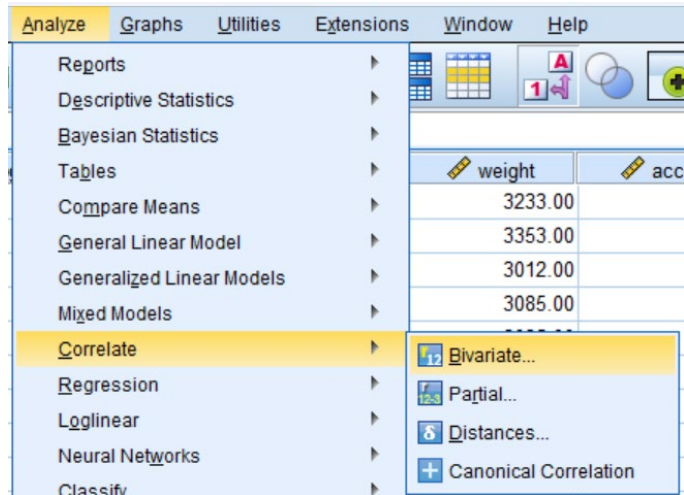
In summary, the scatter plot illustrates a strong positive linear relationship between the weight of a car and its horsepower, with heavier cars generally having higher horsepower.

2. **Calculate the Pearson and Spearman Correlation coefficients for the relationship between Horsepower and Vehicle Weight.**
  - c. What is the p-value for the Pearson correlation?
  - d. What is the actual p-value, as opposed to the p-value that is displayed? To display the actual p-value for the Pearson correlation, double-click on the Pearson correlation output table and double-

click on the p-value. (Remember, p-values cannot actually be equal to zero. The p-value you will see displayed, after double-clicking, will be in scientific notation.)

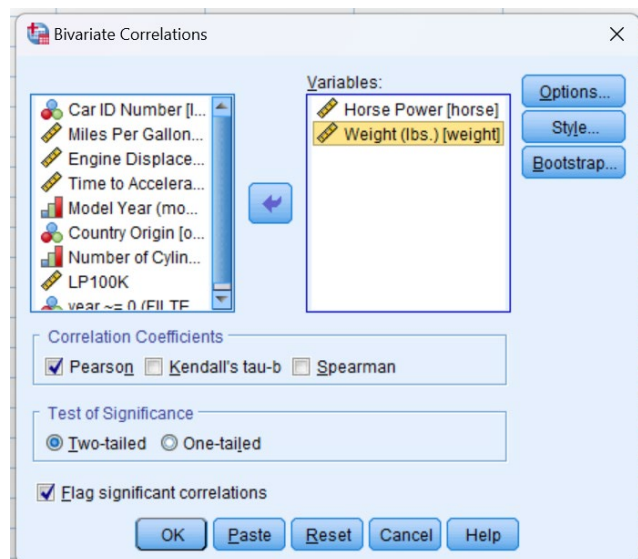
Steps :

- Select **Analyze**
- Then **correlate → Bivariate...**



New window will open,

- Select **Horsepower & Weight** and move them to variables
- In the **Correlation Coefficients** Section, tick **Person & Spearman**
- Rest keep default
- And lastly press OK.



Output:

```
CORRELATIONS
/VARIABLES=horse weight
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
```

## Correlations

### Correlations

		Horse Power	Weight (lbs.)
Horse Power	Pearson Correlation	1	.865**
	Sig. (2-tailed)		.000
	N	398	398
Weight (lbs.)	Pearson Correlation	.865**	1
	Sig. (2-tailed)	.000	
	N	398	404

\*\*. Correlation is significant at the 0.01 level (2-tailed).

```

NONPAR CORR
/VARIABLES=horse weight
/PRINT=SPEARMAN TWOTAIL NOSIG
/MISSING=PAIRWISE.

```

## Nonparametric Correlations

Correlations			Horse Power	Weight (lbs.)
Spearman's rho	Horse Power	Correlation Coefficient	1.000	.881**
		Sig. (2-tailed)	.	.000
		N	398	398
	Weight (lbs.)	Correlation Coefficient	.881**	1.000
		Sig. (2-tailed)	.000	.
		N	398	404

\*\* . Correlation is significant at the 0.01 level (2-tailed).

In order to get the P – Value :

- Double click on the **corelation table**
- Then click on **Sig. (2-tailed )** value
- It will show the Scientific value of **P- value** , which is **1.1807E-120**

### Correlations

Correlations			
		Horse Power	Weight (lbs.)
Horse Power	Pearson Correlation	1	.865**
	Sig. (2-tailed)		1.1807E-120
	N	398	398
Weight (lbs.)	Pearson Correlation	.865**	1
	Sig. (2-tailed)	.000	
	N	398	404

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Therefore the P- value is **1.1807E-120**

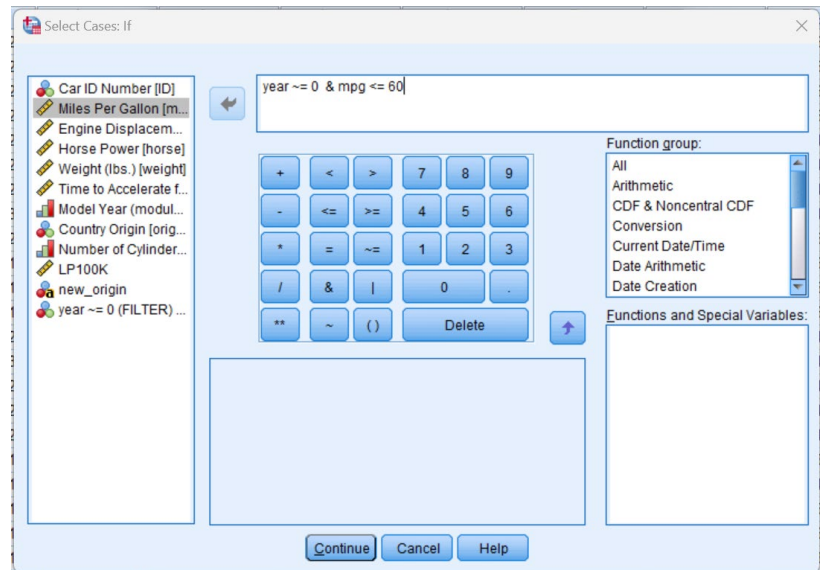
-----

## Relationship Between Continuous Y and Numerical Discrete/Ordinal X

1. Before doing any analyses, select only cases with Year not = 0.
2. Create a side-by-side boxplot of MPG vs. Year. Choose MPG as the “variable” and Year as the “category axis”.

Steps:

- Like all the pervious steps, here also we have to do the same
- First Go to Data options
- Then select Select Cases..
- Select the “If condition is satisfied”
- And click If button
- Now new window opens
- And type “year~= 0 & mpg <=0”
- Lastly press Continue



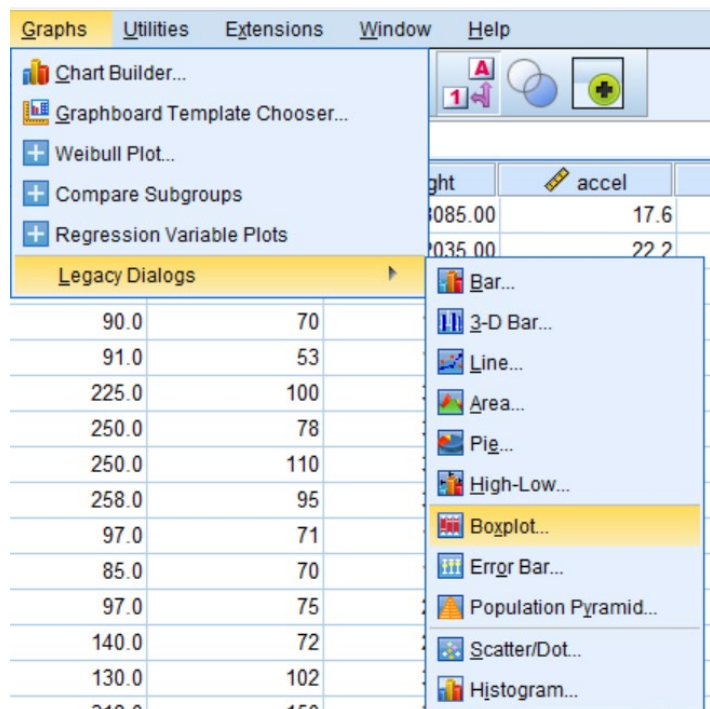
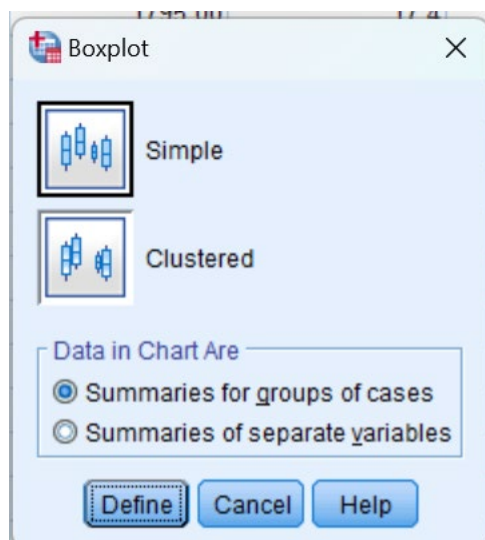
Now we are all set for creating boxplot :

Steps :

- Go to graphs → Legacy Dialogs → Boxplot..

New window opens,

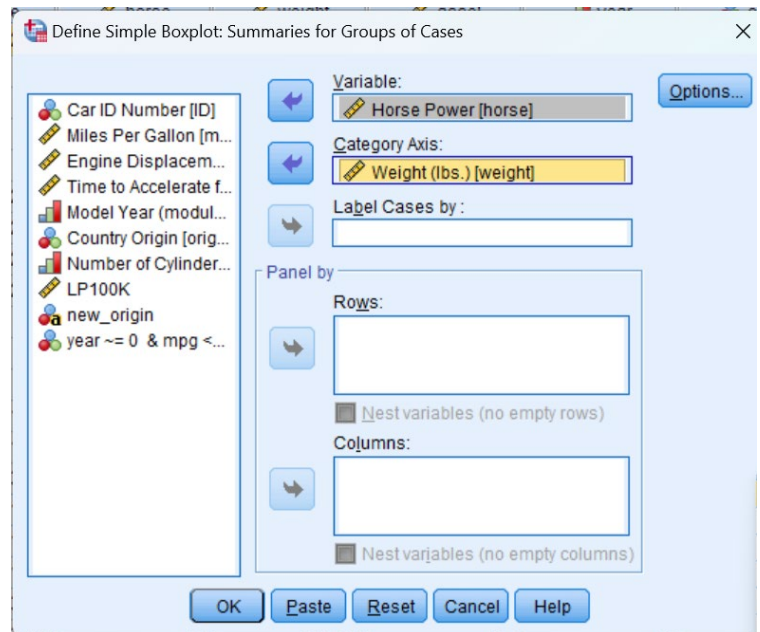
- Keep everything as it is.
- And click Define.



Again new window,

- Select horse power and move it to variables
- Select Weight and move it to Category Axis
- Click OK

Output :



USE ALL.

COMPUTE filter\_\$(year ~= 0 & mpg <= 60).

VARIABLE LABELS filter\_\$(year ~= 0 & mpg <= 60 (FILTER)).

VALUE LABELS filter\_\$(0 'Not Selected' 1 'Selected').

FORMATS filter\_\$(f1.0).

FILTER BY filter\_\$(.

EXECUTE.

EXAMINE VARIABLES=mpg BY year

/PLOT=BOXPLOT

/STATISTICS=NONE

/NOTOTAL.

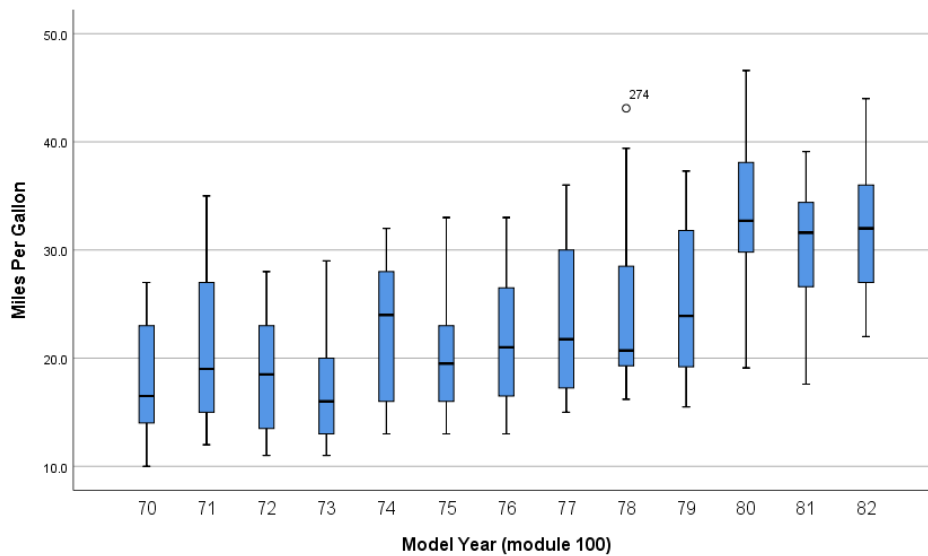
## Model Year (module 100)

### Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Model Year (module 100)	N	Percent	N	Percent	N	Percent
Miles Per Gallon	70	28	100.0%	0	0.0%	28	100.0%
	71	28	100.0%	0	0.0%	28	100.0%
	72	28	100.0%	0	0.0%	28	100.0%
	73	40	100.0%	0	0.0%	40	100.0%
	74	26	100.0%	0	0.0%	26	100.0%
	75	30	100.0%	0	0.0%	30	100.0%
	76	34	100.0%	0	0.0%	34	100.0%
	77	28	100.0%	0	0.0%	28	100.0%
	78	36	100.0%	0	0.0%	36	100.0%
	79	29	100.0%	0	0.0%	29	100.0%
	80	29	100.0%	0	0.0%	29	100.0%
	81	29	100.0%	0	0.0%	29	100.0%
	82	31	100.0%	0	0.0%	31	100.0%



## Miles Per Gallon



What is the general trend of MPG across years?

The box plot illustrates the distribution of miles per gallon (MPG) for cars across various model years from 1970 to 1982. The general trend in MPG over these years can be summarized as follows:

1. Increasing Trend:

- There is a noticeable upward trend in MPG over the years. The median MPG increases steadily from 1970 to 1982.
- The lowest median MPG is seen in the early 1970s, and the highest median MPG is observed in the early 1980s.

2. Variability:

- The variability in MPG also seems to change over the years. Early in the 1970s, there is higher variability (wider interquartile ranges and longer whiskers), indicating a broader range of fuel efficiencies among cars.
- As we move towards the late 1970s and early 1980s, the variability slightly decreases, suggesting more consistency in fuel efficiency improvements across different car models.

3. Outliers:

- There are a few outliers in certain years (e.g., 1978 and 1980), but they do not significantly alter the general trend of increasing MPG.
- The presence of outliers decreases in later years, indicating that extremely low or high MPG values become less common as overall fuel efficiency improves.

4. Range:

- The overall range of MPG increases over the years. While the minimum MPG values remain relatively stable, the maximum MPG values increase significantly, reflecting advancements in fuel efficiency technology.

In summary, the general trend across the model years from 1970 to 1982 is a steady increase in miles per gallon, indicating that cars have become more fuel-efficient over this period.



## Relationship Between Continuous Y and Nominal X

1. Create a side-by-side boxplot of Miles per gallon vs Country of Origin (ORIGIN). (Note: even though Origin is numeric in the data set, its values are **nominal**: American, European, Japanese).
2. What is the general relationship between MPG and the Origin of the car?
3. Create a side-by-side Boxplot of Miles per gallon vs. the recoded Country of Origin (1=Domestic, 0=Foreign).

Steps :

- Same like previous question, steps are similar
- Just in the place to Category Axis put Origin
- Rest all steps are same

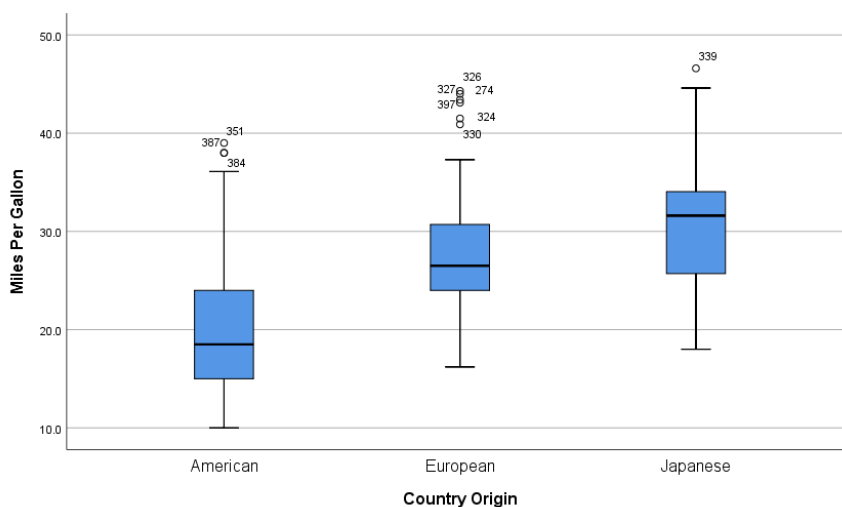
EXAMINE VARIABLES=mpg BY origin  
 /PLOT=BOXPLOT  
 /STATISTICS=NONE  
 /NOTOTAL.

## Country Origin

Case Processing Summary

		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Miles Per Gallon	American	247	100.0%	0	0.0%	247	100.0%
	European	70	100.0%	0	0.0%	70	100.0%
	Japanese	79	100.0%	0	0.0%	79	100.0%

## Miles Per Gallon



What is the general relationship between MPG and the Origin of the car?

The box plot shows the distribution of miles per gallon (MPG) for cars originating from three different regions: American, European, and Japanese. Here are some key observations:

**1. Median MPG:**

- Japanese cars have the highest median MPG, indicating they are generally more fuel-efficient compared to American and European cars.
- European cars have a lower median MPG than Japanese cars but higher than American cars.
- American cars have the lowest median MPG.

**2. Interquartile Range (IQR):**

- The IQR, which represents the middle 50% of the data, is widest for American cars, indicating greater variability in fuel efficiency among these cars.
- European cars have a relatively narrow IQR, suggesting less variability in their MPG.
- Japanese cars also show moderate variability but less than American cars.

**3. Outliers:**

- American cars have a few outliers on the higher end, indicating some cars with very high MPG.
- European cars have multiple outliers, both high and low, suggesting the presence of both very fuel-efficient and less fuel-efficient models.
- Japanese cars have one notable high outlier.

**4. Range:**

- The overall range of MPG is widest for American cars and narrowest for European cars.
- Japanese cars fall in between with a relatively moderate range.

In summary, Japanese cars generally have higher fuel efficiency, followed by European cars, with American cars being the least fuel-efficient on average. However, American cars show the greatest variability in fuel efficiency.

Also need to Create a side-by-side Boxplot of Miles per gallon vs. the recoded Country of Origin(1=Domestic, 0=Foreign).

Follow the previous question's steps

**Steps will be exactly same just put new\_Origin(the recoded Country of Origin (1=Domestic, 0=Foreign) in Category Axis. That's it.**

```
SORT CASES BY new_origin (A).
SORT CASES BY new_origin (D).
EXAMINE VARIABLES=mpg BY new_origin
/PLOT=BOXPLOT
/STATISTICS=NONE
/NOTOTAL.
```

**new\_origin**

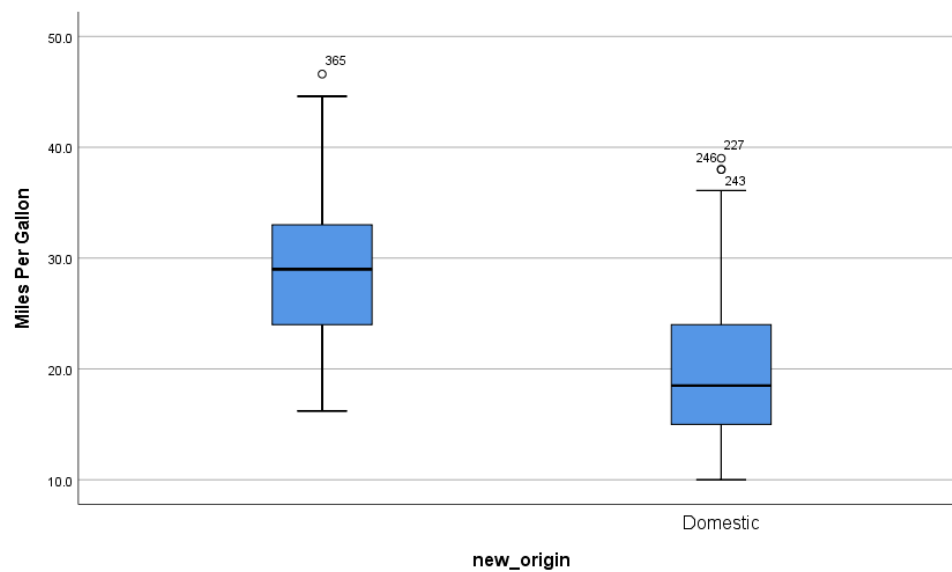
new\_origin

**Case Processing Summary**

Cases

		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Miles Per Gallon		149	100.0%	0	0.0%	149	100.0%
	Domestic	247	100.0%	0	0.0%	247	100.0%

## Miles Per Gallon



## Conclusion

Throughout the analysis, we employed various statistical techniques (on provided dataset of cars\_wave1.xls ,cars\_wave2.xls) including descriptive statistics, inferential statistics, and predictive modeling. Descriptive statistics provided an overview of the dataset, highlighting key features such as central tendencies, dispersion, and the distribution of variables. Inferential statistics, including hypothesis testing and correlation analysis, allowed us to infer properties about the population from our sample and understand relationships between variables. Also needed to deal with data cleaning and transformation of the dataset for carrying the data analysis properly on SPSS software.

Key findings from our analysis include:

- **Descriptive Insights:** The central tendencies of key variables revealed.
- **Correlations:** Significant positive/negative correlations were observed between Horsepower and Weight , indicating positive correlation .
- **Q-Q plot :** from the Q-Q plotting we discover Weight(lbs) variables are not Normally Distributed
- **Box Pot :** with the help of box plotting we analysed the trends of mpg (miles per gallon) with respect to year and origin.

These results provide a solid foundation for understanding the dynamics within the dataset and suggest areas for further investigation. The insights derived can be leveraged for making informed decisions, optimizing processes, and identifying areas for improvement.

In conclusion, the SPSS data analysis of the given dataset has been successful in uncovering critical insights and establishing a basis for future research and practical application. The methodological approach adopted in this project ensures robustness and reliability of the findings, paving the way for ongoing data-driven decision-making.

