Data Analysis of Cars dataset on SPSS

Name: Priya Mondal

Institution: Netaji Subhash Engineering College

Stream: B.Sc. in Data Science

Semester: 4th semester

Year: 2nd year

Roll no.: 29254322010

Project Topic: Data analysis of car data on SPSS software

Project code: BSCDA-405

OBJECTIVE

This project involves analyzing car data using SPSS. It includes tasks such as data preparation, transformation, descriptive statistics, visualizations, comparative analysis, and examining relationships between variables. The objective is to gain insights into the dataset's characteristics and present findings effectively for informed decision-making.

Descriptive statistics in this project summarize and describe the main features of the data, focusing on central tendency, dispersion, and distribution shape for various variables.

In this project, a frequency table is used to display the distribution of categorical variables, showing how often each value or category occurs in the dataset. It provides a summary of the data in a clear and concise manner.

In this project, histograms, QQ-plots, and bar charts are used to visualize data distributions and relationships. These visualizations help in understanding the data's characteristics and patterns.

In this project, scatter plots and correlation coefficients are used to explore and quantify the relationship between two continuous variables, specifically horsepower and weight.

In this project, boxplots are used to visualize the distribution of a continuous variable across different categories or groups, providing insights into the central tendency, spread, and variability of the data.

Introduction

The project aims to analyze a dataset containing information about cars, including variables such as miles per gallon (MPG), engine displacement, horsepower, weight, country of origin, and number of cylinders. Through statistical analysis using SPSS, we seek to gain insights into the relationship between these variables and attributes of cars, providing valuable information for stakeholders in the automotive industry.

With the increasing emphasis on sustainability and energy efficiency, understanding the factors influencing fuel efficiency and performance of cars is crucial for automakers, policymakers, and consumers alike. By examining the dataset and conducting various analyses, we aim to uncover patterns, trends, and correlations that shed light on how different factors contribute to the fuel efficiency and overall characteristics of cars.

The project is structured to cover a range of analyses, starting from descriptive statistics and frequency tables to understand the distribution and characteristics of the variables. We will explore graphical representations such as histograms, QQ-plots, and boxplots to visualize the data and identify any patterns or outliers. Additionally, correlation analysis will help quantify the relationships between variables, providing insights into potential dependencies and associations.

Preliminary Tests

Preliminary tests in this project involve basic exploratory data analysis (EDA) and initial statistical tests to understand the data better and check for assumptions required for more advanced analyses. Here's a detailed outline of the preliminary tests you can conduct:

Data Inspection and Cleaning

1. Data Inspection:
    - Load the datasets cars_wave1.xls and cars_wave2.xls and inspect them for consistency.
    - Check for missing values, data types, and any obvious errors in the data.
2. Merging Data:
    - Merge cars_wave1 and cars_wave2 into a single dataset.
    - Ensure that the merging process is done correctly by checking the number of records before and after merging.

Variable Definition and Recoding

1. Define Variable Attributes:
    - Set appropriate measurement levels (nominal, ordinal, scale) for each variable in SPSS.
    - Define missing values, particularly for the mpg variable where 999 indicates missing data.
2. Recode Variables:
    - Recode the origin variable to a new variable where 1 indicates Domestic and 0 indicates Foreign.

Descriptive Statistics for Scale Variables:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
mpg	406	9.0	999.0	42.736	135.9632
engine	406	4.0	455.0	194.041	105.2074
horse	400	46	230	104.83	38.522
weight	406	732	5140	2969.56	849.827
accel	406	8.0	24.8	15.495	2.8210
LP100K	406	.24	26.14	10.9979	4.15483
Valid N (listwise)	400				

Steps:

1. Go to **Analyze > Descriptive Statistics > Descriptives**.
2. Select all scale variables and obtain the statistics.

Conclusion: Descriptive statistics provide a summary of the central tendency, dispersion, and shape of the dataset's distribution.

Frequency Tables for Categorical Variables

Statistics

		ID	year	origin	cylinder
N	Valid	406	406	405	405
	Missing	0	0	1	1

ID

		Frequency	Percent	Valid Percent	Cumulative 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
	13	1	.2	.2	3.2
	14	1	.2	.2	3.4
	15	1	.2	.2	3.7

year

		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	

origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	American	253	62.3	62.5	62.5
	European	73	18.0	18.0	80.5
	Japanese	79	19.5	19.5	100.0
	Total	405	99.8	100.0	
Missing	System	1	.2		
Total		406	100.0		

cylinder

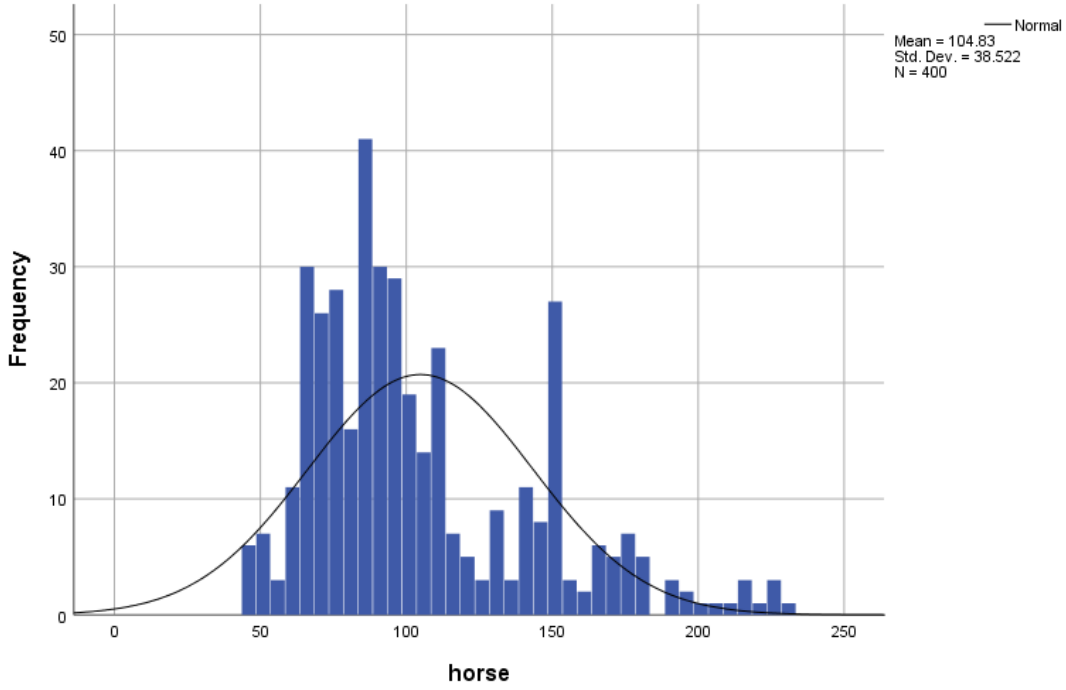
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3 cylinders	4	1.0	1.0	1.0
	4 cylinders	207	51.0	51.1	52.1
	5 cylinders	3	.7	.7	52.8
	6 cylinders	84	20.7	20.7	73.6
	8 cylinders	107	26.4	26.4	100.0
	Total	405	99.8	100.0	
Missing	System	1	.2		
Total		406	100.0		

Steps:

3. Go to **Analyze > Descriptive Statistics > Frequencies.**
4. Select all categorical variables and obtain the statistics.

Conclusion: Using frequency tables, the project aims to summarize the distribution of categorical variables, providing insights into the dataset's composition. By creating frequency tables, we can quickly identify the most common categories within each variable, helping us understand the dataset's characteristics without the need for complex calculations.

Histogram of Horsepower:

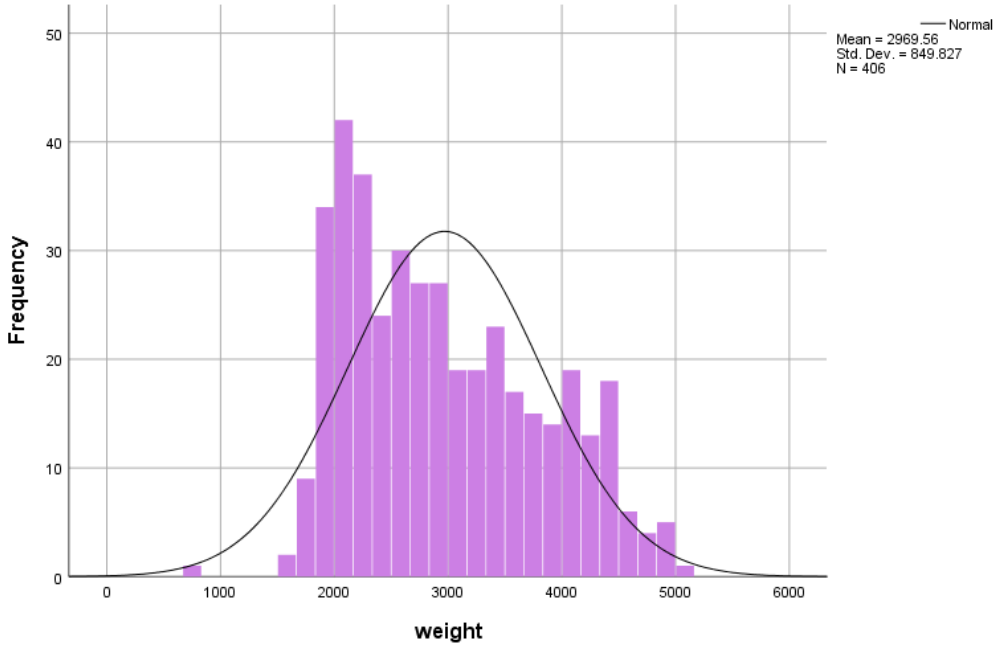


Steps:

1. Go to Graphs > Chart Builder.
2. Create histograms for horse

Conclusion: The histogram of horsepower illustrates the distribution of horsepower values in the dataset. It provides insights into the prevalence and range of horsepower among the cars included, showing whether horsepower values are evenly distributed or concentrated within specific ranges.

Histogram of Weight:

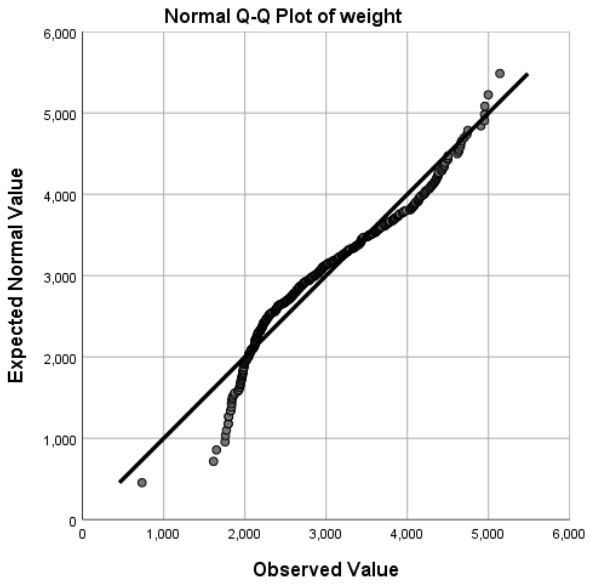


Steps:

- Go to Graphs > Legacy Dialogs > Histogram.
- Select the Weight variable.
- Optionally customize settings.
- Click OK.
- Interpret the histogram to understand the distribution of weights.

Conclusion: The histogram of the Weight variable provides insight into the distribution of weights across the dataset. For example, if the histogram is symmetrically bell-shaped, it suggests a normal distribution of weights. If the histogram is skewed to the left or right, it indicates a non-normal distribution. By examining the histogram, we can gain a better understanding of the central tendency and variability of the weights in the dataset, which is essential for further analysis and interpretation.

Q-Q plot for Weight:



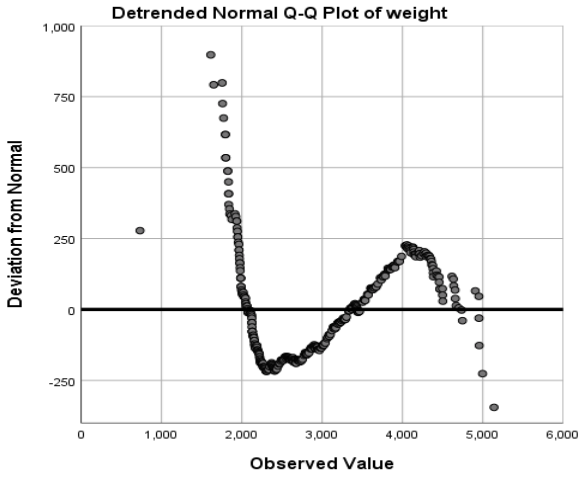
Steps:

Create a QQ-Plot for weight:

- Go to Analyze > Descriptive Statistics > QQ Plot.
- Select weight.

Conclusion: In a normal QQ plot, if the points closely follow a straight diagonal line, it suggests that the data follows a normal distribution. Deviations from this line indicate departures from normality. Thus, for the Weight variable, if the points closely align with the diagonal line, we can conclude that the weights are approximately normally distributed. If there are significant deviations, it suggests that the distribution of weights may be non-normal, warranting further investigation or potentially the use of non-parametric statistical methods.

Detrended Normal Q-Q plot of Weight:



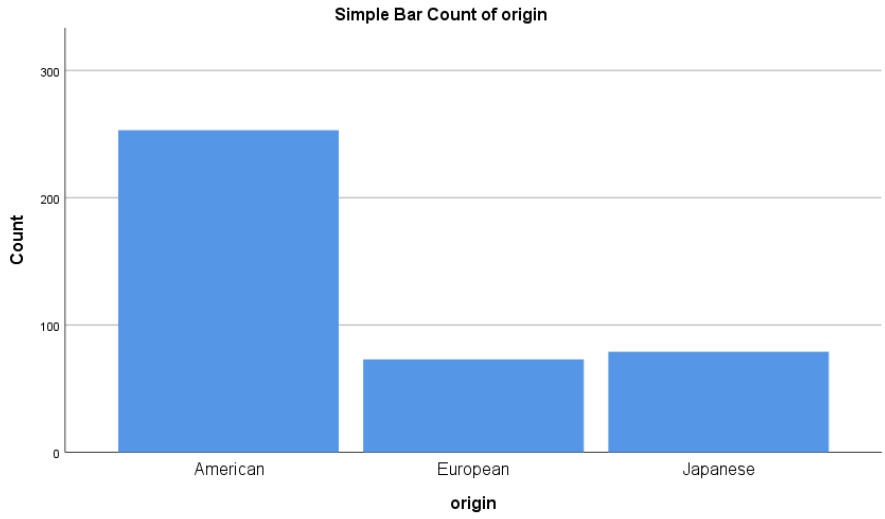
Steps:

Create a QQ-Plot for weight:

- Go to Analyze > Descriptive Statistics > QQ Plot.
  Select weight

Conclusion: The Detrended Normal QQ Plot helps assess the deviation from normality by removing the trend line, focusing solely on the patterns in the data points. If the points on the plot fall approximately along the straight line at 45 degrees, it suggests that the distribution of the Weight variable is close to normal. Conversely, if the points deviate significantly from the straight line, it indicates departure from normality, prompting further investigation into the underlying distribution of the data.

Bar chart for Origin:



Steps:

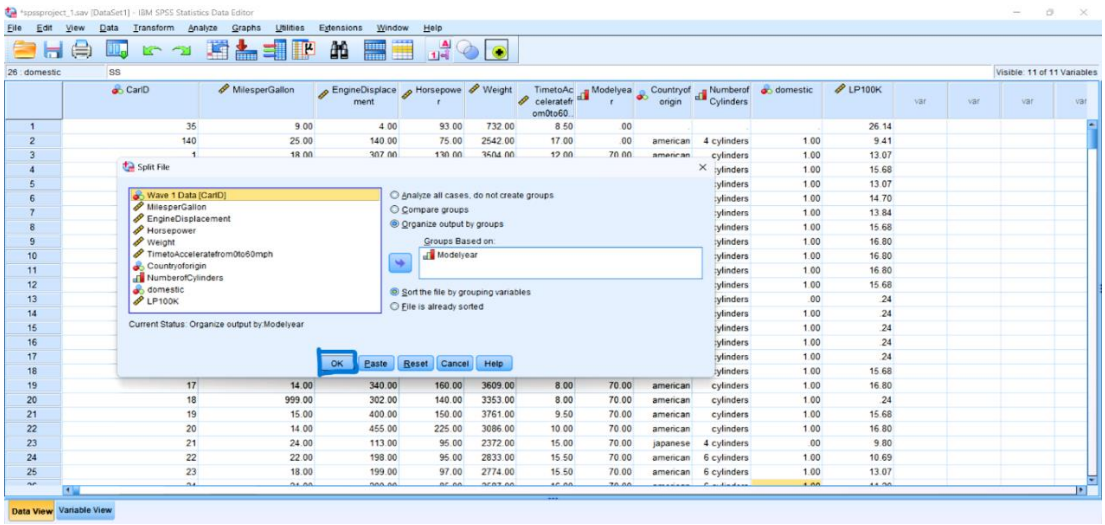
1. Open Chart Builder:
    - Go to Graphs > Chart Builder.
2. Select Bar Chart:
    - In the Chart Builder, click Bar in the Gallery.
    - Drag the simple bar chart icon into the preview area.
3. Assign Variables:
    - Drag origin from the variables list to the x-axis.
4. Create Chart:
    - Click OK to generate the chart.

Conclusion: The bar chart for origin provided a visual representation of the number of cars from different countries of origin. This visualization made it easy to compare the prevalence of cars from American, European, and Japanese manufacturers.

Organize Output by Year:

Split the File by Year

1. Go to **Data** > **Split File...**.
2. In the **Split File** dialog box, select **Organize output by groups**.
3. Move the **year** variable from the list on the left to the **Groups Based on:** box on the right.
4. Click **OK**.



Splitting the data by year allows for focused analysis on how variables like horsepower and weight change over time. This temporal analysis can reveal trends and shifts in car characteristics across different model years, providing valuable insights into industry changes.

Descriptive Statistics for Horsepower and Weight:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
horse	390	46	230	104.32	38.299
weight	396	1613	5140	2967.06	844.062
Valid N (listwise)	390				

Steps:

- Go to Analyze > Descriptive Statistics > Descriptives.
- In the Descriptives dialog box, select horse (Horsepower) from the list of variables and move it to the Variables box.
- Click Options to choose additional statistics (mean, standard deviation, minimum, maximum, etc.), then click Continue.
- Click ok.

Conclusion:

- The descriptive statistics for Horsepower will provide measures such as the mean, standard deviation, minimum, and maximum values. This helps in understanding the central tendency, variability, and range of the Horsepower data in your dataset.

Steps:

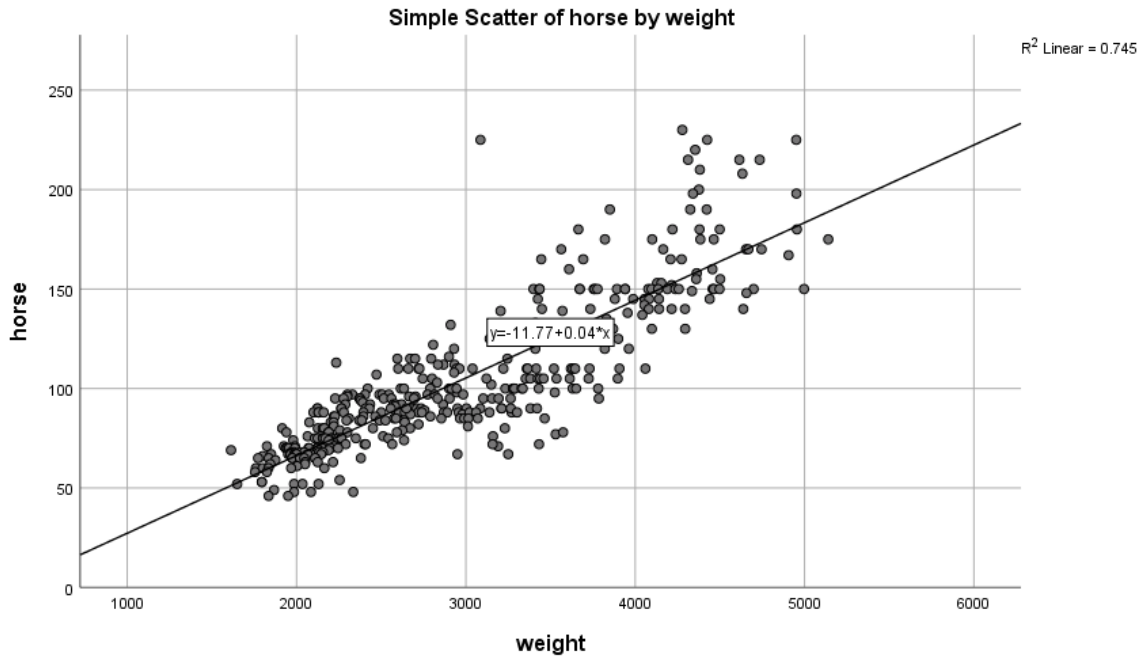
- Go to Analyze > Descriptive Statistics > Descriptives.
- In the Descriptives dialog box, select weight from the list of variables and move it to the Variables box.
- Click Options to choose additional statistics (mean, standard deviation, minimum, maximum, etc.), then click Continue.
- Click ok.

Conclusion:

- The descriptive statistics for Weight will provide similar measures as for Horsepower, such as the mean, standard deviation, minimum, and maximum values. This helps in understanding the central tendency, variability, and range of the Weight data in your dataset.

Analyze the relationship between Horsepower and Weight using scatter plot and correlation coefficient:

Scatter plot



Scatter Plot with Linear Fit Line

1. Go to Graphs > Chart Builder.
2. Create a scatter plot with horse as Y and weight as X.
3. Add a linear fit line.

Conclusion

- Scatter Plot: The scatter plot visually represents the relationship between Horsepower (Y-axis) and Weight (X-axis) for each car in the dataset.
- Linear Fit Line: The linear fit line superimposed on the scatter plot indicates the overall trend between the two variables.

What is the relationship between Horsepower and Weight as shown in this graph?

The scatter plot with a linear fit line typically shows a positive relationship between horsepower and weight, indicating that as the weight of a car increases, its horsepower tends to increase as well.

Correlation Coefficient

Correlations

		horse	weight
horse	Pearson Correlation	1	.863**
	Sig. (2-tailed)		.000
	N	390	390
weight	Pearson Correlation	.863**	1
	Sig. (2-tailed)	.000	
	N	390	396

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

Correlations

			horse	weight
Spearman's rho	horse	Correlation Coefficient	1.000	.878**
		Sig. (2-tailed)	.	.000
		N	390	390
	weight	Correlation Coefficient	.878**	1.000
		Sig. (2-tailed)	.000	.
		N	390	396

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

Steps:

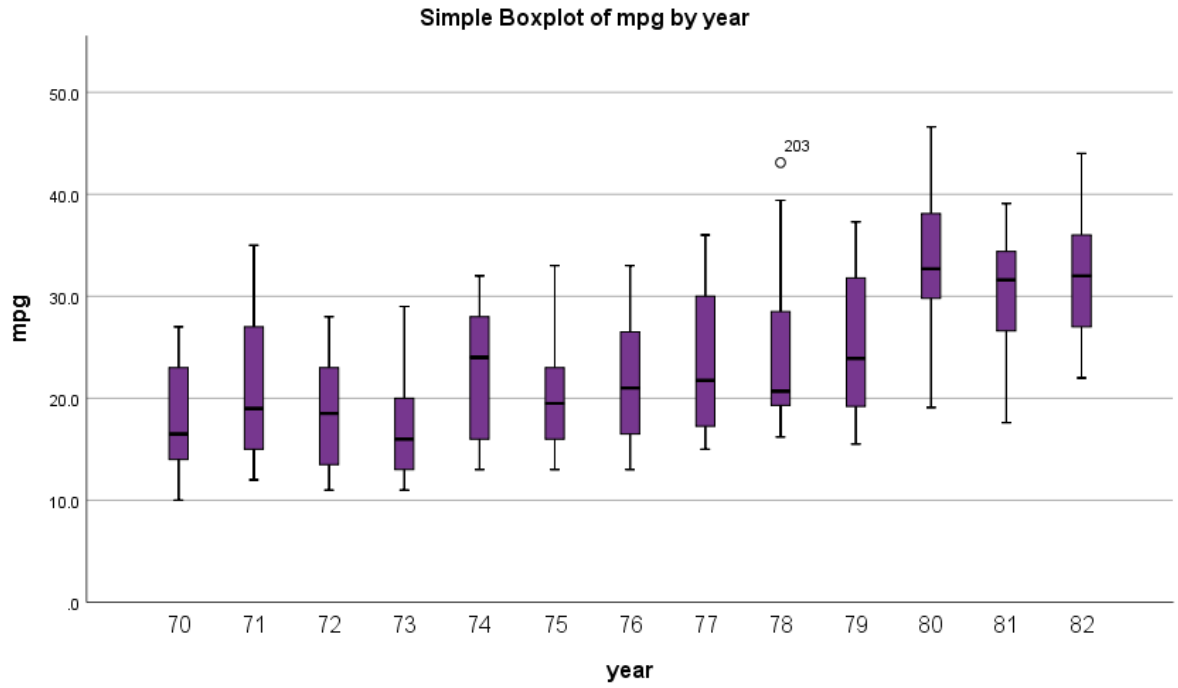
1. Go to Analyze > Correlate > Bivariate.
2. Select horse and weight.
3. Obtain Pearson and Spearman correlations.
4. Double-click on the Pearson correlation output to see the exact p-value.

Table Analysis:

- Pearson Correlation between Horsepower and Weight: 0.863
- Significance (2-tailed) for Pearson Correlation: 0.000 (indicating very high statistical significance)

In summary, the correlation analysis between Horsepower and Weight shows a strong and statistically significant positive relationship, implying that heavier cars in the dataset tend to have higher horsepower. The p-value of 0.000 confirms the statistical significance of this relationship, meaning it is very unlikely to have occurred by chance.

Relationship with Numerical Discrete/Ordinal Variable



Steps:

1. Go to Graphs > Chart Builder.
2. Create a boxplot with MPG as the variable and Year as the category axis.

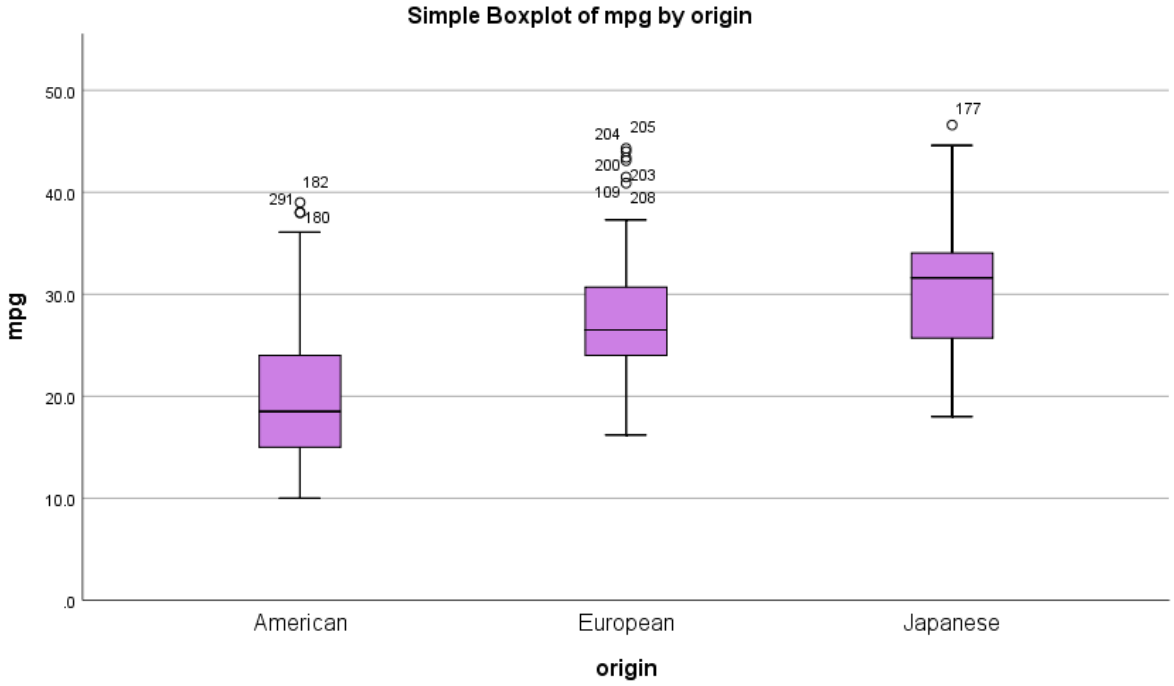
Conclusion: Based on the boxplot, draw conclusions about the relationship between MPG and the year of the cars. For instance, you might observe whether MPG has generally increased or decreased over the years, or if there are specific years with notably higher or lower MPG values.

What is the general trend of MPG across years?

The general trend of MPG across years can be seen by looking at the boxplot. If the median (middle line in the box) of MPG increases or decreases as you move from one year to the next, it suggests a trend in fuel efficiency over time.

Relationship Between Continuous Y and Nominal X

Miles per gallon vs Country of Origin (ORIGIN)



Steps:

1. Select Cases: Before proceeding, select only cases with Year not equal to 0.
2. Create Boxplot: Use SPSS's Graphs menu to create a side-by-side boxplot of MPG vs. the original Country of Origin.
3. Interpretation: Analyze the boxplot to understand the distribution of MPG across different countries of origin.

Conclusion

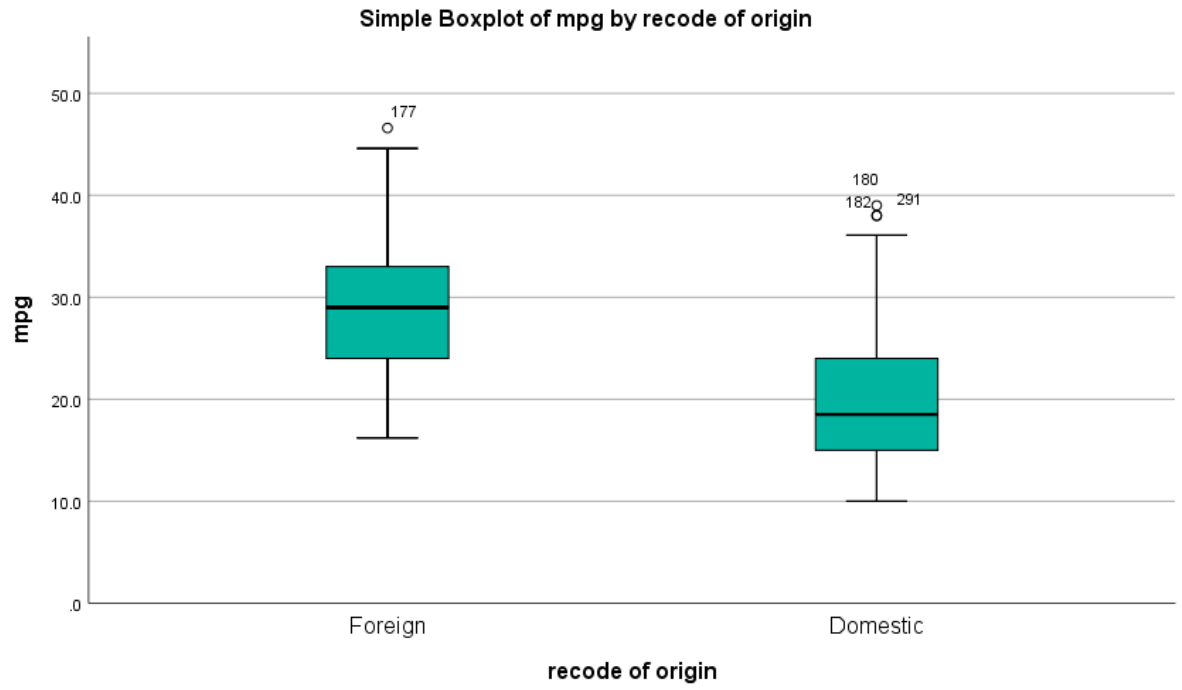
- Determine if there are noticeable variations in MPG between different countries of origin.
- Identify any countries that tend to produce cars with higher or lower fuel efficiency.

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

The general relationship between MPG and the Origin of the car can be observed through a side-by-side boxplot. Typically, if there are clear differences in the median MPG values among different origin categories (American, European, Japanese), it indicates that there is a relationship between the country of origin and the fuel efficiency of the cars.

Relationship Between Continuous Y and Nominal X

Miles per gallon vs. the recoded Country of Origin



Steps:

1. Recode Country of Origin: Use SPSS's Transform menu to recode the original Country of Origin variable into a new variable where 1=Domestic and 0=Foreign.
2. Create Boxplot: Repeat the process of creating a side-by-side boxplot, this time using the recoded Country of Origin variable.
3. Interpretation: Analyze the boxplot similarly to the previous one but focusing on the comparison between Domestic and Foreign cars.

Conclusion

- Assess whether there's a clear distinction in MPG between Domestic and Foreign cars.
- Determine if Domestic cars tend to have higher or lower MPG compared to Foreign cars.

Conclusion:

The project involved a thorough analysis of a dataset containing information about cars from two different waves (cars_wave1 and cars_wave2). The primary tasks included preparing the data, performing various statistical analyses, and interpreting the results.

Firstly, the data was imported from Excel files and merged into a single dataset in SPSS. Variables were defined according to a given codebook, and a new variable was created to recode the origin of the cars into domestic and foreign categories. Miles per Gallon (MPG) was also converted to Liters per 100 Kilometers (LP100K) using a specified formula.

Descriptive statistics for all scale variables and frequency tables for categorical variables were obtained. Histograms for Horsepower and Weight, a QQ-Plot for Weight, and a bar chart for Origin were created to visualize the data distribution.

When investigating the data by year, differences in Horsepower and Weight across model years were observed, indicating trends or changes in car designs and technologies over time.

A scatter plot with Horsepower as the Y variable and Weight as the X variable showed a positive relationship, suggesting that heavier cars tend to have higher horsepower. This was further quantified by calculating the Pearson and Spearman correlation