

CIS 5250 – 01 VISUAL ANALYTICS

SAS PROJECT

TITLE: CAR SALES

TEAM: Sirisha Mahesh & Spencer Asahi

SAS Project: CAR Sales

[A] Introduction

This Project Analyzes a Car Sales dataset Obtained from Kaggle, Which Provides extensive information on various car features to predict Car Sales. The Primary Objective is to identify the features that significantly impact Car Sales and to leverage these insights to build a predictive model. Using SAS Studio for Visual Analytics, this Project will Utilize advanced Analytical Techniques to understand relationships within the data and optimize the predictive capabilities of a classifier.

Project Plan

The Project is divided into two main tasks:

Feature Analysis: By Examining each feature in relationship to car sales, I will determine which factors hold the most predictive power for Sales Performance.

Model Training and Prediction: A Classifier will be trained using these impactful features to predict Car Sales outcomes. Model accuracy will be assessed to ensure its reliability for predictive Analysis.

Outcome Analyses

Sales Impact by Manufacture and Model:

Top Manufactures: The top three manufacturers are Dodge, Ford and Chevrolet contribute a large share of the sales. Analyzing the features common among high-selling models from these brands like pricing, horsepower or resale values could help pinpoint strategies these brand use to drive sales.

Model-wise Distribution: Specific models like the ford F-Series and Chevrolet Cavalier show high sales figures, potentially influenced by factors such as price, engine size and vehicle type.

Impact of Vehicle Type on Sales and Resale:

Passenger vs Car Segment: Passenger vehicles dominate sales at 74%, while cars make up, 26%. The higher percentage in passenger vehicles suggests a user preference for comfort, utility, or features offered by this segment. Analyzing the factors that might influence this preference such as fuel efficiency, engine size or width could provide insights.

Car Sales Trend by Vehicle Type: By Visualizing trends in sales for specific types of cars sedans, SUV's etc could forecast demand in certain Segments.

Feature Impact on Sales:

Engine Size and Horsepower: Vehicles with larger engines and higher horsepower such as Dodge Viper with 8L and 850 HP could be preferred for power but have niche sales due to price or fuel efficiency concerns, Evaluating the impact of horsepower and engine size on sales might reveal threshold where performance outweighs costs or fuel drawbacks.

Wheelbase and Width: As attribute tied to vehicle stability and passenger comfort, wheel base and width might influence sales positivity in higher-end, luxury models where comfort is prioritized.

Resale Value Predictors:

Resale by Model & Brand: Creating a model that uses brand, vehicle type, price, and features as predictors can help determine which models are most likely to retain value. This can be especially useful for dealerships in marketing and sales pitches.

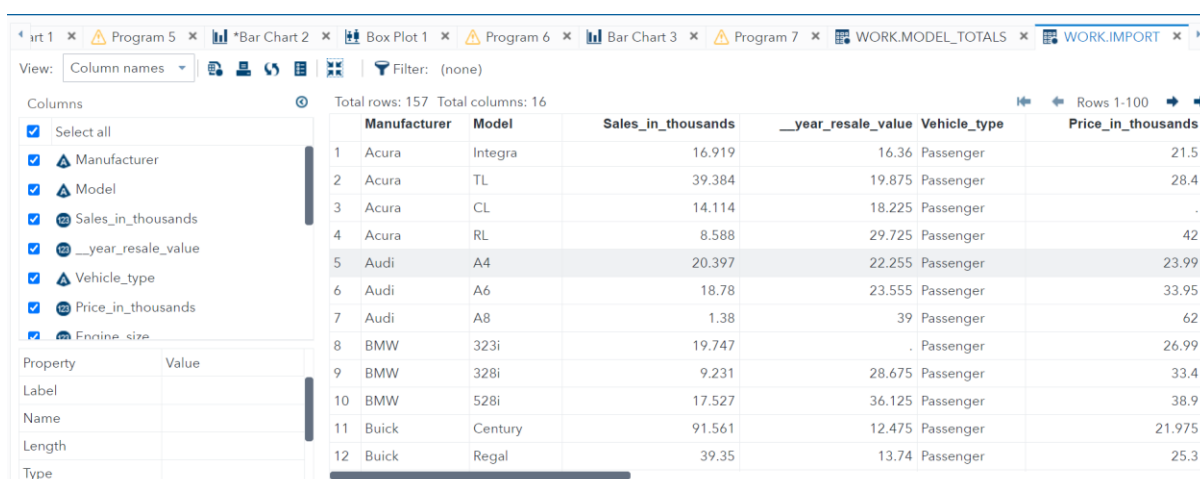
Resources Referred: <https://www.kaggle.com/datasets/gagandeep16/car-sales>

[I]Keim, D. A., Mansmann, F., Schneidewind, J., & Thomas, J. (2008). *Visual Analytics: Scope and Challenges*. Springer.

[II]Wexler, S., Shaffer, J., & Cotgreave, A. (2017). *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. Wiley.

[III] Matignon, R. (2007). *Data Mining Using SAS Enterprise Miner*. Wiley.

[IV]SAS Institute. *Getting Started with SAS Visual Analytics*. Available on the SAS support site for insights into creating effective visuals.



	Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands
1	Acura	Integra	16.919	16.36	Passenger	21.5
2	Acura	TL	39.384	19.875	Passenger	28.4
3	Acura	CL	14.114	18.225	Passenger	.
4	Acura	RL	8.588	29.725	Passenger	42
5	Audi	A4	20.397	22.255	Passenger	23.99
6	Audi	A6	18.78	23.555	Passenger	33.95
7	Audi	A8	1.38	39	Passenger	62
8	BMW	323i	19.747	.	Passenger	26.99
9	BMW	328i	9.231	28.675	Passenger	33.4
10	BMW	528i	17.527	36.125	Passenger	38.9
11	Buick	Century	91.561	12.475	Passenger	21.975
12	Buick	Regal	39.35	13.74	Passenger	25.3

Table Data:

Data Columns	Descriptions
Manufacturer	The Company brand that produces the vehicle.
Model	The Specific name or a designation of a vehicle produced by the Manufacturer.
Sales in thousands	The total number of units sold, expressed in thousands.
Year Resale Value	The estimated value of the vehicle when resold after a specific number of years.
Vehicle Type	The classification of the vehicle, such as sedan, SUV etc but in Our dataset its given as Passenger and Car
Price in thousands	The retails price of the vehicle, expressed in thousands of Currency units.
Engine Size	The total volume of the engine cylinders, typically measured in Liters.
Horsepower	A measure of the engine's power output, indicating the vehicle's Performance potential.
Wheelbase	The distance between the front and rear axles of the vehicle, Affecting stability and handling.

Width	The measurement across the vehicle from side to side, impacting Interior space and maneuverability.
Length	The measurement from the front to the back of the vehicle, Influencing storage capacity and passenger space.
Curb Weight	The total weight of the vehicle including all fluids and equipment But excluding passengers and cargo.
Fuel Capacity	The maximum amount of fuel the vehicle's tank can hold, usually Measured in gallons or liters.
Fuel Efficiency	The distance a vehicle can travel per unit of fuel, typically Expressed in miles per gallon (MPG)
Latest Launch	The most recent year or date when a new model or version of the Vehicle was introduced.
Power Per Factor	A metric that relates the power output of the vehicle to a specific Factor, which may include weight or size, helping to evaluate Performance efficiency.

Excel Data Screen Shot:

1	Manufacturer	Model	Sales_in_thousands	_year_resale_val	Vehicle_type	Price_in_thousa	Engine_size	Horsepower	Wheelbase	Width	Length	Curb_weight	Fuel_capacity	Fuel_efficiency	Latest_launch	Power_pdr_factor
2	Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67.3	172.4	2.639	13.2	28	2/2/2012	58.28015
3	Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70.3	192.9	3.517	17.2	25	6/3/2011	91.37078
4	Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70.6	192	3.47	17.2	26	1/4/2012	
5	Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71.4	196.6	3.85	18	22	3/10/2011	91.38978
6	Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68.2	178	2.998	16.4	27	10/8/2011	62.77764
7	Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76.1	192	3.561	18.5	22	8/9/2011	84.56511
8	Audi	A8	1.38	39	Passenger	62	4.2	310	113	74	198.2	3.902	23.7	21	2/2/2012	134.6569
9	BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68.4	176	3.179	16.6	26	6/28/2011	71.19121
10	BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68.5	176	3.197	16.6	24	1/29/2012	81.87707
11	BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70.9	188	3.472	18.5	25	4/4/2011	83.99872
12	Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72.7	194.6	3.368	17.5	25	11/2/2011	71.18145
13	Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72.7	196.2	3.543	17.5	23	9/3/2011	95.6367
14	Buick	Park Avenue	27.851	20.19	Passenger	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5	24	3/23/2012	85.82841
15	Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73.5	200	3.591	17.5	25	7/23/2011	84.25453
16	Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5	22	2/23/2012	113.8546
17	Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	75	201		18.5	22	4/29/2011	115.6214
18	Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75.5	200.6	3.843	19	22	11/27/2011	113.7659
19	Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70.3	194.8	3.77	18	22	9/28/2011	83.48309
20	Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	77	201.2	5.572	30	15	4/17/2012	109.5091
21	Chevrolet	Cavalier	145.519	9.25	Passenger	13.26	2.2	115	104.1	67.9	180.9	2.676	14.3	27	8/1/2011	46.36335
22	Chevrolet	Malibu	135.126	11.225	Passenger	16.535	3.1	170	107	69.4	190.4	3.051	15	25	3/19/2012	67.31446
23	Chevrolet	Lumina	24.629	10.31	Passenger	18.89	3.1	175	107.5	72.5	200.9	3.33	16.6	25	5/24/2011	69.9914
24	Chevrolet	Monte Carlo	42.593	11.525	Passenger	19.39	3.4	180	110.5	72.7	197.9	3.34	17	27	12/22/2011	72.03092
25	Chevrolet	Camaro	26.402	13.025	Passenger	24.34	3.8	200	101.1	74.1	193.2	3.5	16.8	25	10/23/2011	81.11854
26	Chevrolet	Corvette	17.947	36.225	Passenger	45.705	5.7	345	104.5	73.6	179.7	3.21	19.1	22	5/12/2012	141.1412
27	Chevrolet	Prizm	32.299	9.125	Passenger	13.96	1.8	120	97.1	66.7	174.3	2.398	13.2	33	9/11/2011	48.29764
28	Chevrolet	Metro	21.855	5.16	Passenger	9.235	1	55	93.1	62.6	149.4	1.895	10.3	45	4/13/2012	23.27627
29	Chevrolet	Impala	107.995		Passenger	18.89	3.4	180	110.5	73	200	3.389	17	27	6/18/2011	71.83804
30	Chrysler	Sebring Coupe	7.854	12.36	Passenger	19.84	2.5	163	103.7	69.7	190.9	2.967	15.9	24	11/17/2012	65.95718
31	Chrysler	Sebring Conv.	32.775	14.18	Passenger	24.495	2.5	168	106	69.2	193	3.332	16	24	11/1/2011	69.52136
32	Chrysler	Concorde	31.148	13.725	Passenger	22.245	2.7	200	113	74.4	209.1	3.452	17	26	6/6/2012	80.02378
33	Chrysler	Cirrus	32.306	12.64	Passenger	16.48	2	132	108	71	186	2.911	16	27	10/6/2011	53.5662
34	Chrysler	LHS	13.462	17.325	Passenger	28.34	3.5	253	113	74.4	207.7	3.564	17	23	5/8/2012	101.3293
35	Chrysler	Town & Country	53.48	19.54	Car										7/13/2011	
36	Chrysler	300M	30.696		Passenger	29.185	3.5	253	113	74.4	197.8	3.567	17	23	2/10/2012	101.6552
37	Dodge	Neon	76.034	7.75	Passenger	12.64	2	132	105	74.4	174.4	2.567	12.5	29	12/1/2011	52.0849
38	Dodge	Avenger	4.734	12.545	Passenger	19.045	2.5	163	103.7	69.1	190.2	2.879	15.9	24	7/1/2012	65.65051
39	Dodge	Stratus	71.186	10.185	Passenger	20.23	2.5	168	108	71	186	3.058	16	24	10/31/2011	67.87611
40	Dodge	Intrepid	88.028	12.275	Passenger	22.505	2.7	202	113	74.7	203.7	3.489	17		6/7/2012	80.83147
41	Dodge	Viper	0.916	58.47	Passenger	69.725	8	450	96.2	75.7	176.7	3.375	19	16	8/7/2011	188.1443
42	Dodge	Ram Pickup	227.061	15.06	Car	19.46	5.2	230	138.7	79.3	224.2	4.47	26	17	3/6/2012	90.2117

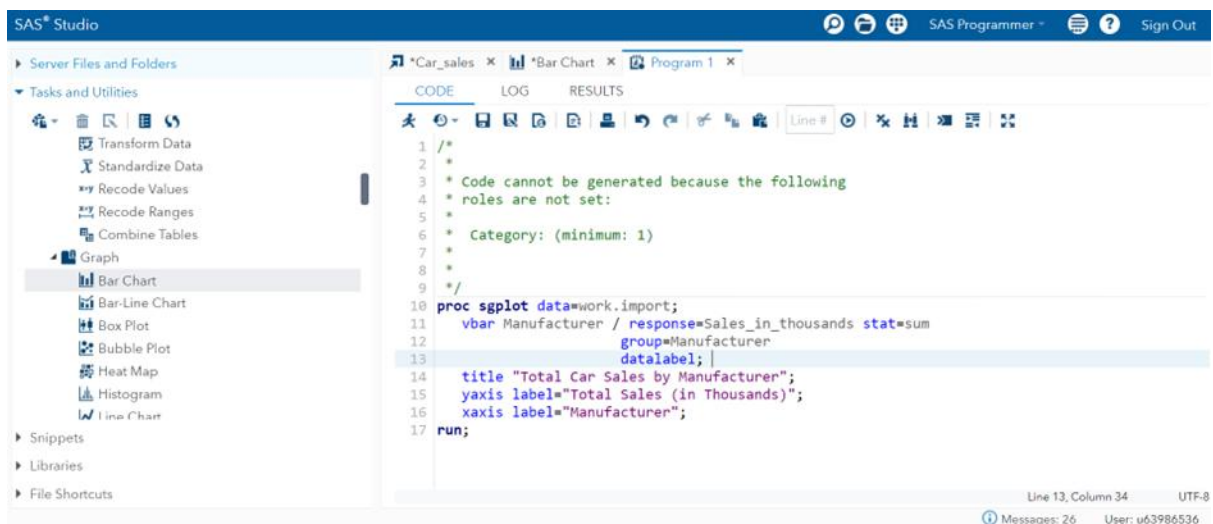
Questions:

- 1) What is the distribution of car sales by manufacturer?
- 2) How does the average car price vary by vehicle type?
- 3) Is there a relationship between engine size and horsepower?
- 4) Which manufacturer has the highest/lowest average fuel efficiency?

[B]Data Visualizations

1)What is the distribution of car sales by manufacturer?

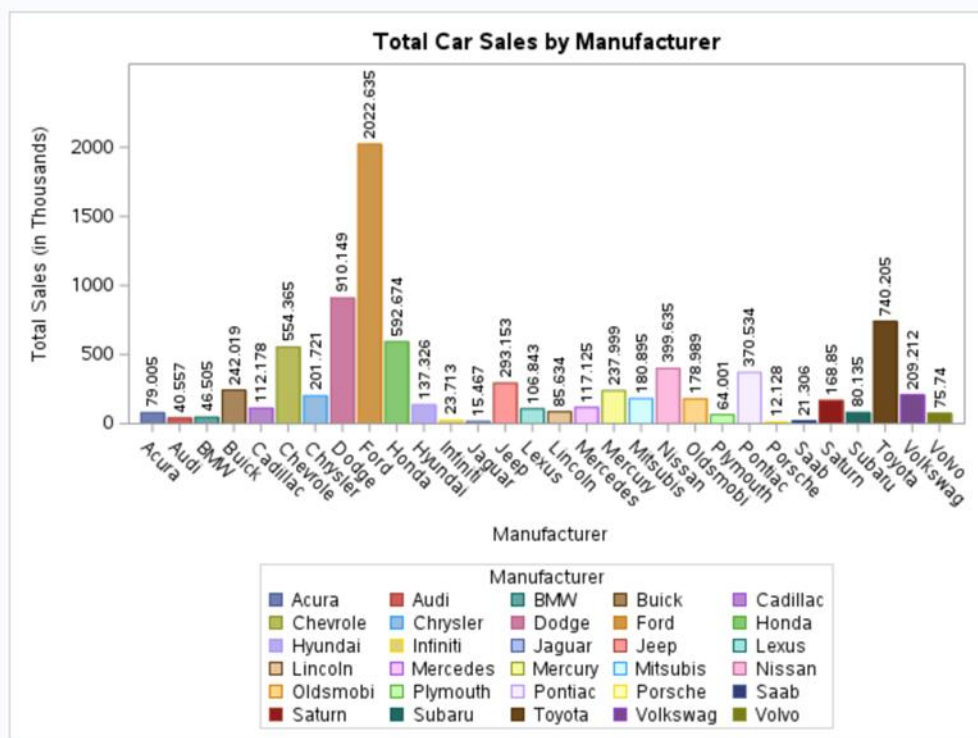
Car sales are usually done in great volume by reliability and the best manufacturer of that area. We can see with the graph there are predominantly American, European, and Japanese cars. These numbers represent car sales and within all 3 of these categories and you can see a hike in the data for some of these results. For example Ford and Toyota would be the highest. Ford represents the highest in the American car segment and Toyota in the Japanese car segment. Domestically American car sales dominate in this segment because even other American cars have higher sales than the highest Japanese car sales (Toyota).



The screenshot shows the SAS Studio interface. On the left, the 'Tasks and Utilities' pane is open, with 'Graph' selected and 'Bar Chart' highlighted. The main editor displays a SAS program with the following code:

```
1 /*  
2 *  
3 * Code cannot be generated because the following  
4 * roles are not set:  
5 *  
6 * Category: (minimum: 1)  
7 *  
8 *  
9 */  
10 proc sgplot data=work.import;  
11   vbar Manufacturer / response=Sales_in_thousands stat=sum  
12                       group=Manufacturer  
13                       datalabel; |  
14   title "Total Car Sales by Manufacturer";  
15   yaxis label="Total Sales (in Thousands)";  
16   xaxis label="Manufacturer";  
17 run;
```

The status bar at the bottom indicates 'Line 13, Column 34' and 'UTF-8'. The message pane shows 'Messages: 26' and 'User: u63986536'.



Top Sellers: The top-selling manufacturers in 2000 were Ford, Toyota, and General Motors (which includes Chevrolet, Buick, Cadillac, etc.).

Low Sales: Several manufacturers had very low sales figures, including Saab, Porsche, and Saturn.

Market Dominance: Ford and Toyota had a significant market share, with their bars being much taller than most others.

It's important to note that this data is from 2000, and the automotive market has changed significantly since then. Some of the manufacturers mentioned may no longer exist or have merged with other companies.

2)How does the average car price vary by vehicle type?

Yes but ever so slightly. The results below show a significant price difference in passenger vehicles compared to “car” vehicles which this data represents as everything else. So that “car” segment would be your typical hatchback, truck, or even sports cars. Passenger cars have a similar average to “cars” but there are much more expensive offerings due to the nature of being a passenger car. We being Americans almost every vehicle sold is going to be a passenger car which makes manufacturers cater and focus on producing more of these. Which entails more research and development and better passenger cars which gives us a wide distribution of cheaper passenger cars and nicely priced average passenger cars to quite expensive ones.

SAS® Studio

Server Files and Folders

Tasks and Utilities

- Transform Data
- Standardize Data
- Recode Values
- Recode Ranges
- Combine Tables
- Graph
 - Bar Chart
 - Bar-Line Chart
 - Box Plot
 - Bubble Plot
 - Heat Map
 - Histogram
 - Line Chart
- Snippets
- Libraries
- File Shortcuts

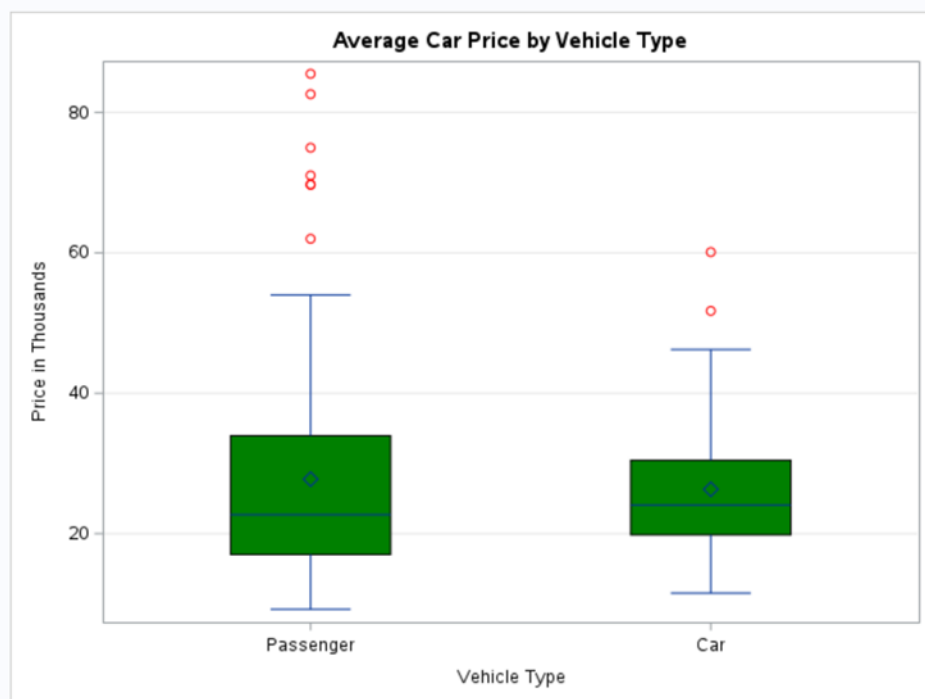
*Car_sales x *Bar Chart x Program 1 x Box Plot x Program 2 x

CODE LOG RESULTS

```
3 * Code cannot be generated because the following
4 * roles are not set:
5 *
6 * Analysis variable: (minimum: 1)
7 *
8 *
9 */
10
11 proc sgplot data=work.import;
12   vbox Price_in_thousands / category=Vehicle_type
13     fillattrs=(color=green)
14     outlierattrs=(symbol=circle color=red)
15     lineattrs=(color=black thickness=1);
16
17   title "Average Car Price by Vehicle Type";
18   yaxis label="Price in Thousands" grid;
19   xaxis label="Vehicle Type" discreteorder=data;
20 run;
21
22
```

Line 16, Column 2 UTF-8

Messages: 34 User: u63986536



The box for Passenger vehicles is slightly wider than the box for Car vehicles, indicating a larger spread in prices for Passenger vehicles.

Both distributions have outliers, represented by the red dots above the boxes. These outliers suggest the presence of some very expensive vehicles within each category.

While the median prices are similar, the distribution of prices differs. Passenger vehicles tend to have a slightly higher median price and a wider range of prices compared to Car vehicles.

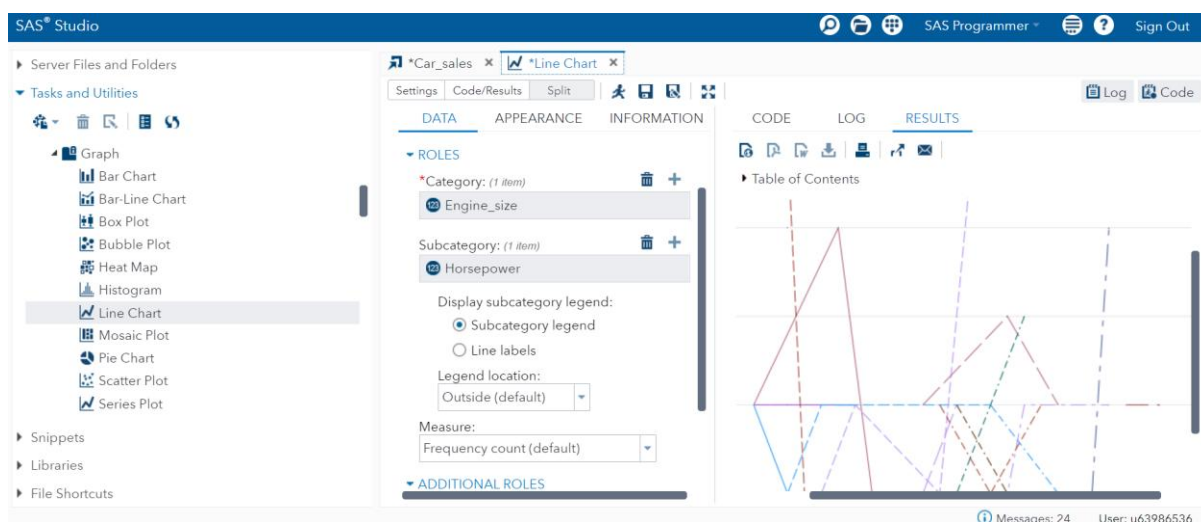
The whiskers on the box plots represent the minimum and maximum values within the data, excluding outliers.

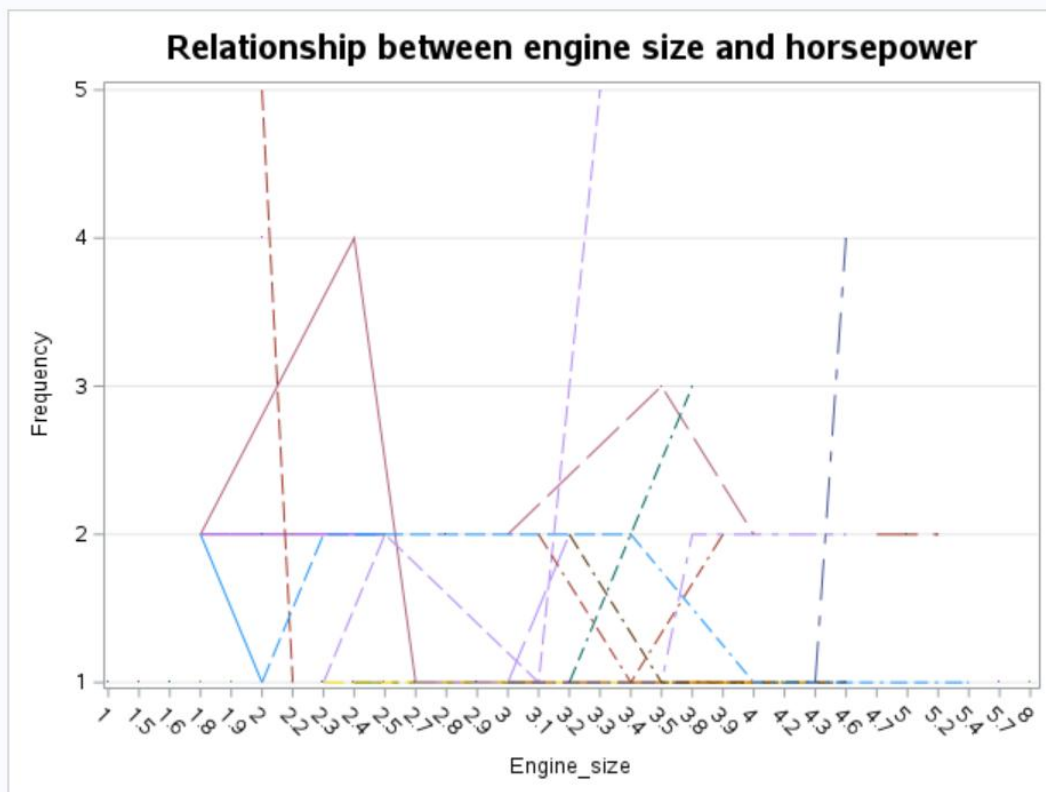
The line inside the box indicates the median price.

The box itself represents the interquartile range (IQR), which contains 50% of the data.

3) Is there a relationship between engine size and horsepower?

Yes definitely. To put it simply as engine size goes up, power as well. This is due to the nature of the engine. A bigger engine is able to pump more air and fuel and have a stronger space for a more powerful combustion which then creates more horsepower. Majority of vehicle manufacturers know this and putting an engine with a larger displacement will create more power. That's why many trucks and suvs as well as sports cars use a slightly bigger engine than your average economy car. You'll see horsepower numbers climb as engines by liter size goes higher up. Majority of cars are under 250 horsepower but we see plenty more reaching to the 400 horsepower range. You also see quite a lot under 200 horsepower. To give you perspective the under 200 horsepower vehicles would be economy cars and around 250 would be passenger vehicles or regular everyday sedans. They need 250 to carry family, friends, and to drive distances where most economy cars wouldn't be that comfortable. And the higher horsepower vehicles would be your sports cars or trucks.





Engine Size Distribution: The plot shows the distribution of engine sizes in the dataset. Some engine sizes occur more frequently than others.

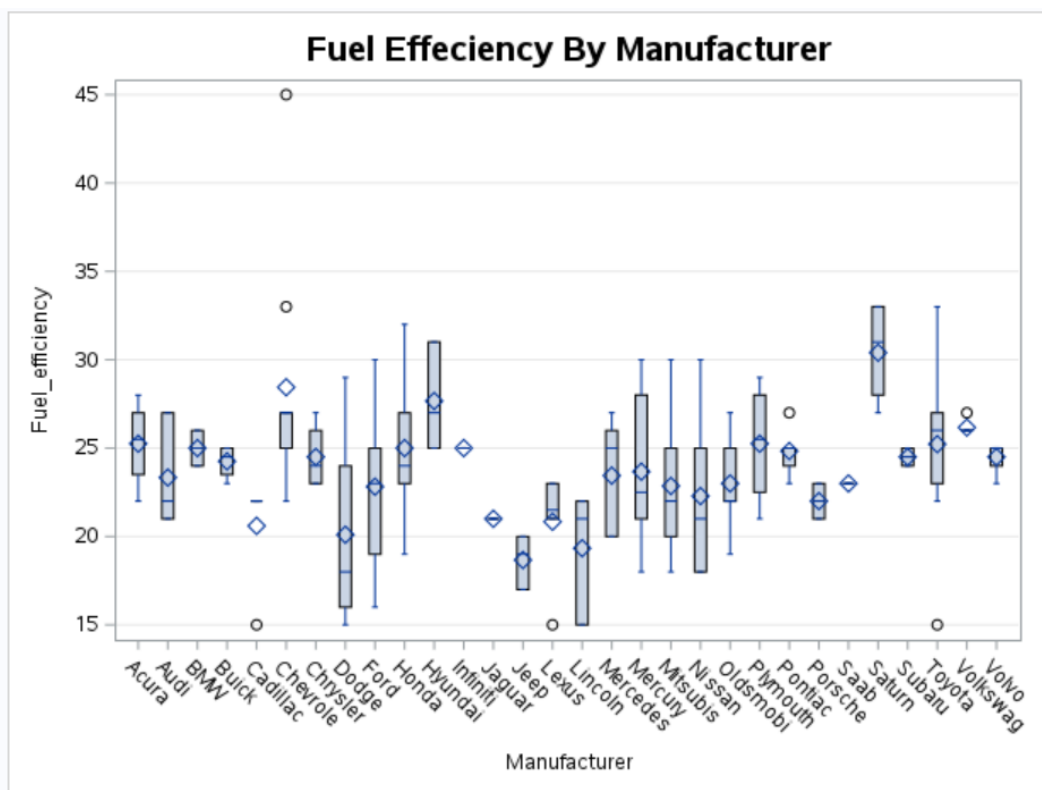
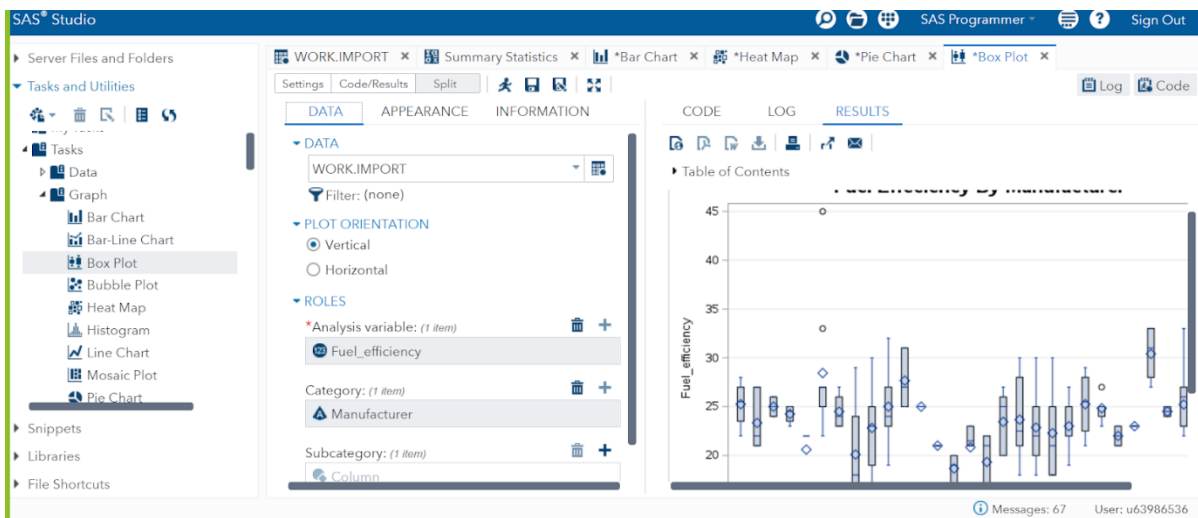
Frequency Peaks: There are a few engine sizes with higher frequencies, indicating that these sizes are more common in the dataset.

Variation in Frequency: The frequency of different engine sizes varies, suggesting a diverse range of engines in the dataset.

No Specific Values: The exact engine size values are not clearly labeled, making it difficult to pinpoint specific sizes and their frequencies.

Overlapping Lines: The overlapping lines can make it challenging to distinguish the frequencies for certain engine sizes.

4) Which manufacturer has the highest/lowest average fuel efficiency?



The box itself represents the middle 50% of the data, from the first quartile (Q1) to the third quartile (Q3). The line inside the box indicates the median (Q2), which divides the data into two halves. The lines extending from the box are called whiskers. They represent the range of the data, excluding outliers. The end of each whisker typically indicates the minimum and maximum values. Data points that fall outside the whiskers are considered outliers. They are often represented by individual dots or symbols.

Spread: The length of the box indicates the spread of the middle 50% of the data. A longer box suggests a wider spread.

If the median is closer to the bottom of the box, the data is skewed to the right (positively skewed).

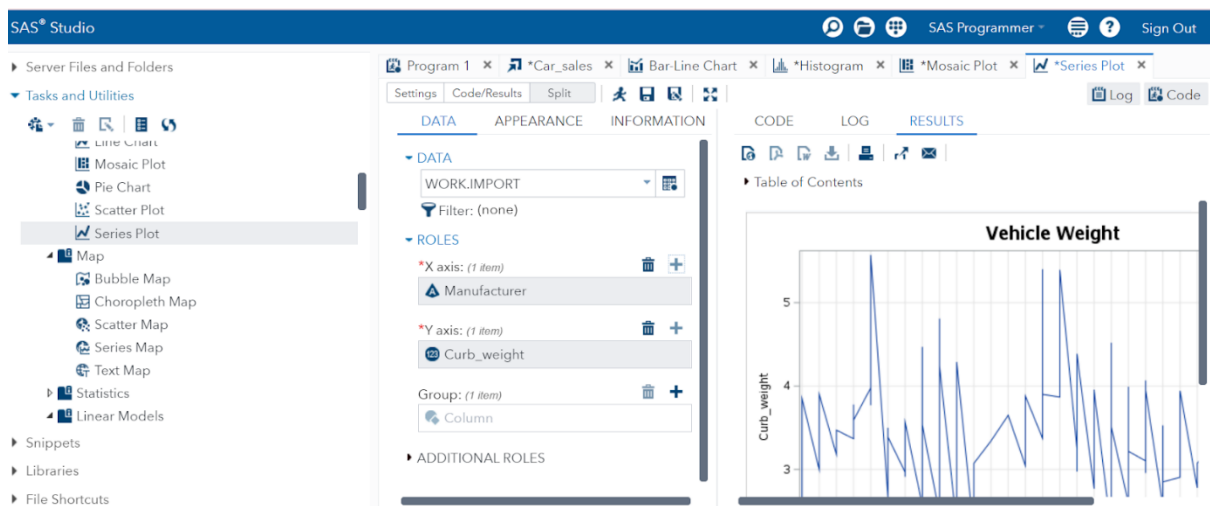
If the median is closer to the top of the box, the data is skewed to the left (negatively skewed).

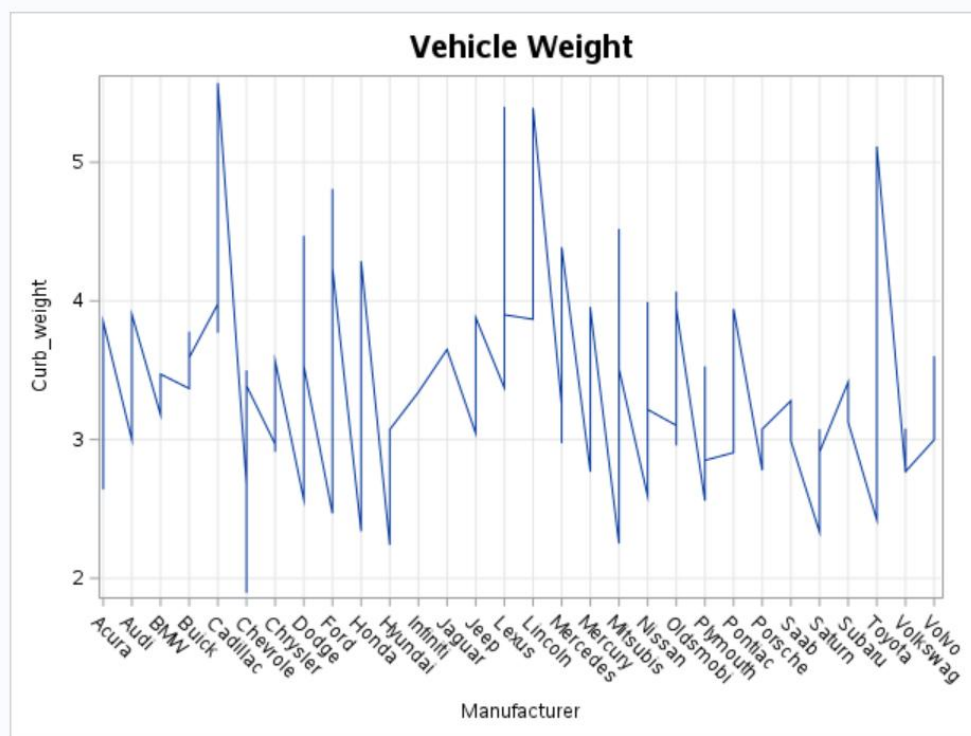
A symmetrical distribution has the median in the middle of the box.

Each box plot represents the distribution of fuel efficiency for a specific manufacturer. By comparing the box plots, we can:

- Identify manufacturers with higher or lower median fuel efficiency.
- Assess the variability in fuel efficiency within each manufacturer.
- Detect any outliers in fuel efficiency.
- Compare the overall distribution of fuel efficiency across different manufacturers.

5) Vehicle Weight By Manufacturer





There's significant variation in curb weight across different manufacturers.

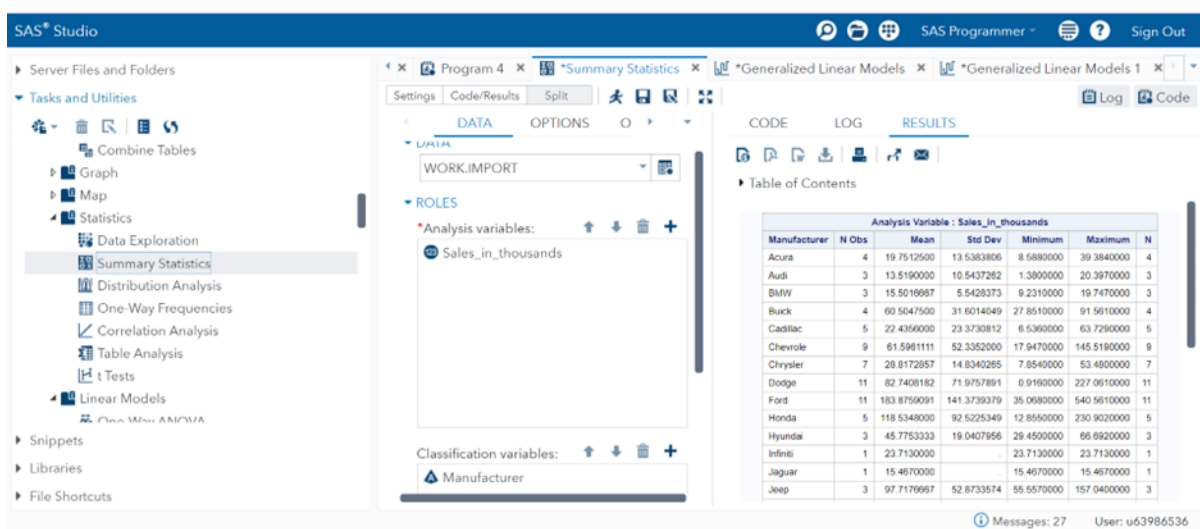
Clusters: Some manufacturers seem to have similar curb weights, potentially indicating similar vehicle types or segments.

Outliers: There might be a few manufacturers with unusually high or low curb weights.

[C] Statistical Summary:

The statistical summary you will see is a number under N OBS, that is the data set option and we have significantly more American vehicles which are Chevrolet, Chrysler, Dodge, and Ford. The mean for Ford and Honda are the highest and Honda would be the highest mean for the Japanese auto manufacturers. This means that Honda is one of the better performing Japanese car brands despite Toyota selling a lot more. It is the center of the entire distribution and the most fair priced that means. This is because Honda has a very good offerings in their line of vehicles that are similar to ford. Both sell good sedans and passenger vehicles such as the Ford Fusion and Taurus. Honda has the Civic and Accord which are competitors to earlier stated ford vehicles. But Ford does sell more in majority of the US due to the pride of American manufacturing and Fords stake in the American population as “the American” car to get is very popular.

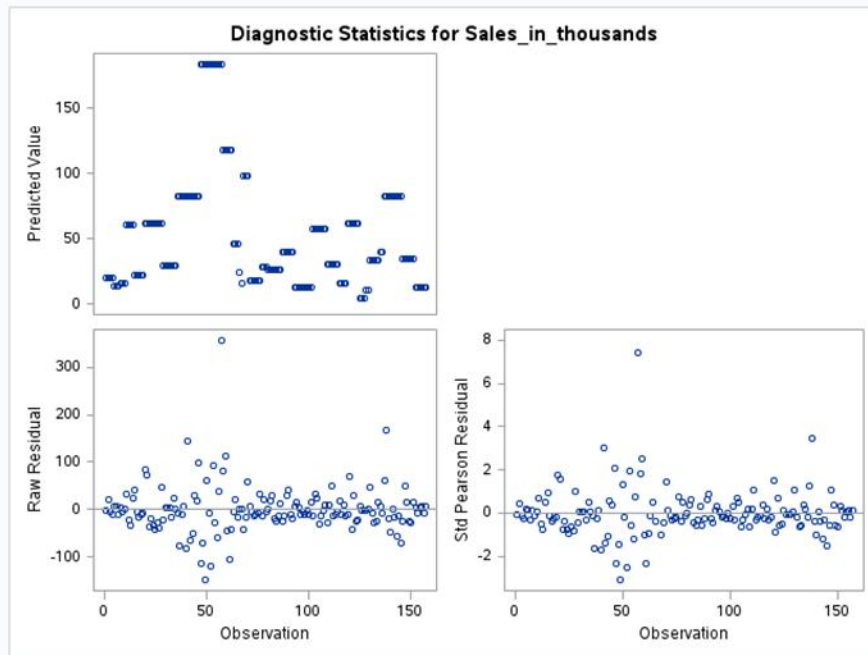
So a smaller AIC which is a Akalike Information Criterion adjusts for the small sample sizes and since it is quite low, this suggest that there is actually a good balance between the data such as manufacturers and the other factors like prices and engine sizes. Another thing to note is a larger BIC which means the parameters were a bit more than needed but still effective. Our data was simple but there were so many intricacies and numbers such as engine liters, curb weight, as well as horsepower in a plethora of manufacturers and models within them. There is an advantage because the deviance and Pearson Chi square you see at the top is exactly the same. This displays that there is no underdispersion or over. Secondly there is confidence with the predictions and stats we show. Another reason why having both of the numbers the same is great is because there is simplicity in our data which keeps the statistical model and graphs more interpretable



The screenshot shows the SAS Studio interface. The left pane displays the 'Tasks and Utilities' menu, with 'Summary Statistics' selected. The main window is divided into three panes: 'DATA' (showing 'WORK.IMPORT'), 'ROLES' (showing 'Sales_in_thousands' as the analysis variable), and 'RESULTS'. The 'RESULTS' pane displays a table of statistics for various manufacturers.

Manufacturer	N Obs	Mean	Std Dev	Minimum	Maximum	N
Acura	4	19.7512500	13.5383806	8.5880000	39.3840000	4
Audi	3	13.5190000	10.5437282	1.3800000	20.3970000	3
BMW	3	15.5019987	5.5428373	9.2310000	19.7470000	3
Buick	4	60.5047500	31.6014049	27.8510000	91.5610000	4
Cadillac	5	22.4356000	23.3730812	6.5360000	63.7290000	5
Chevrolet	9	61.5961111	52.3352000	17.9470000	145.5190000	9
Chrysler	7	28.8172857	14.8340285	7.8540000	53.4800000	7
Dodge	11	82.7408182	71.9757891	0.9190000	227.0610000	11
Ford	11	183.8750091	141.3739379	35.0680000	540.5610000	11
Honda	5	118.5348000	92.5225349	12.8550000	230.9020000	5
Hyundai	3	45.7753333	19.0407956	29.4500000	66.6920000	3
Infiniti	1	23.7130000		23.7130000	23.7130000	1
Jaguar	1	15.4670000		15.4670000	15.4670000	1
Jeep	3	97.7176987	52.8733574	55.5570000	157.0400000	3

Analysis Variable : Sales_in_thousands						
Manufacturer	N Obs	Mean	Std Dev	Minimum	Maximum	N
Acura	4	19.7512500	13.5383806	8.5880000	39.3840000	4
Audi	3	13.5190000	10.5437262	1.3800000	20.3970000	3
BMW	3	15.5016667	5.5428373	9.2310000	19.7470000	3
Buick	4	60.5047500	31.6014049	27.8510000	91.5610000	4
Cadillac	5	22.4356000	23.3730812	6.5360000	63.7290000	5
Chevrolet	9	61.5961111	52.3352000	17.9470000	145.5190000	9
Chrysler	7	28.8172857	14.8340265	7.8540000	53.4800000	7
Dodge	11	82.7408182	71.9757891	0.9160000	227.0610000	11
Ford	11	183.8759091	141.3739379	35.0680000	540.5610000	11
Honda	5	118.5348000	92.5225349	12.8550000	230.9020000	5
Hyundai	3	45.7753333	19.0407956	29.4500000	66.6920000	3
Infiniti	1	23.7130000	.	23.7130000	23.7130000	1
Jaguar	1	15.4670000	.	15.4670000	15.4670000	1
Jeep	3	97.7176667	52.8733574	55.5570000	157.0400000	3
Lexus	6	17.8071667	17.8801677	3.3340000	51.2380000	6
Lincoln	3	28.5446667	18.2185670	13.7980000	48.9110000	3
Mercedes	9	13.0138889	10.6618711	0.9540000	28.9760000	9
Mercury	6	39.6665000	27.7632462	14.3510000	81.1740000	6
Mitsubishi	7	25.8421429	20.9192886	0.1100000	55.6160000	7
Nissan	7	57.0907143	21.8214101	27.3080000	88.0940000	7
Oldsmobile	6	29.8315000	27.5642408	1.1120000	80.2550000	6
Plymouth	4	16.0002500	14.8538296	1.8720000	32.7340000	4
Pontiac	6	61.7556667	41.8487703	19.9110000	131.0970000	6
Porsche	3	4.0426667	4.2876111	1.2800000	8.9820000	3
Saab	2	10.6530000	2.0675802	9.1910000	12.1150000	2
Saturn	5	33.7700000	31.6147934	5.2230000	80.6200000	5
Subaru	2	40.0675000	9.9553564	33.0280000	47.1070000	2
Toyota	9	82.2450000	73.1785678	9.8350000	247.9940000	9
Volkswagen	6	34.8686667	31.5924777	5.5960000	83.7210000	6
Volvo	6	12.6233333	7.1524206	3.4930000	18.9690000	6



1. Predicted Value vs. Observation Plot:

Pattern: The points are scattered randomly, without any clear pattern or trend. This is a good indication that the model's predictions are not systematically biased.

Interpretation: The model appears to be making unbiased predictions across the range of observations.

2. Raw Residual vs. Observation Plot:

Pattern: The residuals are centered around zero, with some fluctuations above and below. There's no discernible pattern or trend.

Interpretation: The model's residuals are randomly distributed, suggesting that the error terms are independent and have constant variance. This is a crucial assumption for valid inference.

3. Standardized Pearson Residual vs. Observation Plot:

Pattern: Similar to the raw residual plot, the standardized residuals are scattered randomly around zero.

Interpretation: The standardized residuals also exhibit no discernible pattern, further confirming the assumption of constant variance and independent errors.

[D]Statistical Tests:

One -way Frequency:

What is One-Way Frequency Analysis?

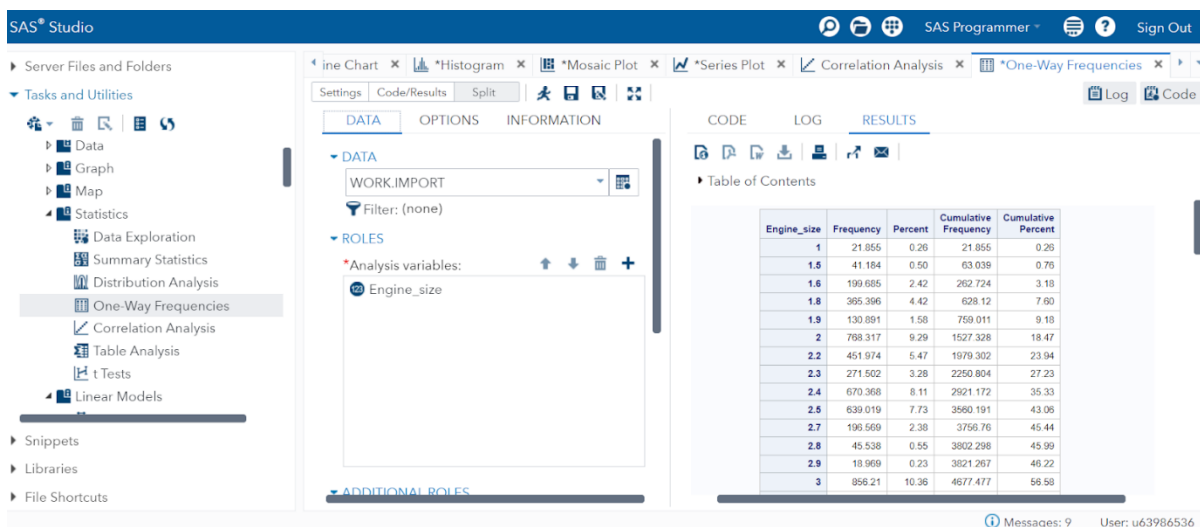
One-way Frequency Analysis is a Statical technique used to summarize categorical data. It helps us to understand the distribution of a single categorical variable by counting the frequency number of occurrences of each category.

Frequency: The number of times each category occurs. Ex: there are 21,855 cars with an engine size of 1.

Percentage: The proportion of each category relative to the total number of observations. Ex: For instance 21.06% of the cars have ab engine size of 1.

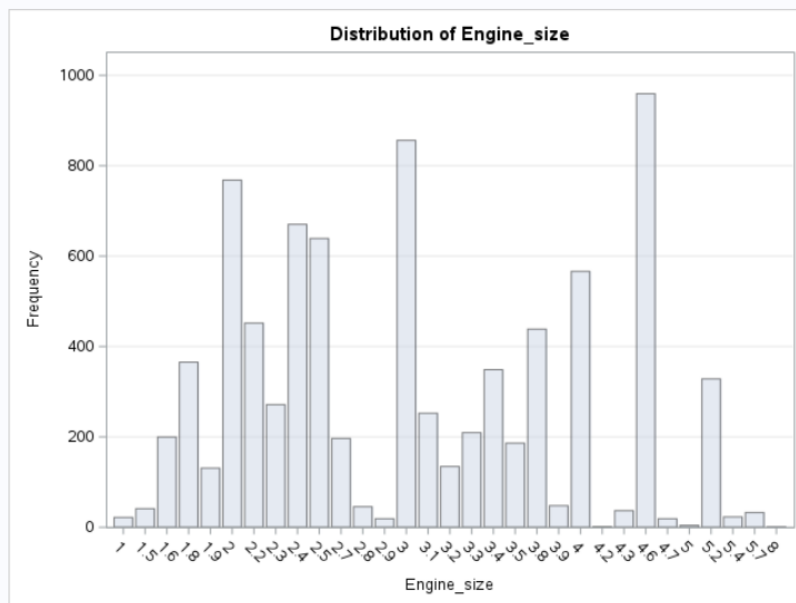
Cumulative Frequency: The running total of frequencies as you moved down the table. It shows how many observations have a value less than or equal to the current category.

Cumulative Percentage: The running total of percentages. It shows the proportion of observations with a value less than or equal to the current category.



We can compare the proportions of different engine sizes. For ex: cars with an engine size of 1.5 are less common than those with an engine size of 1.6.

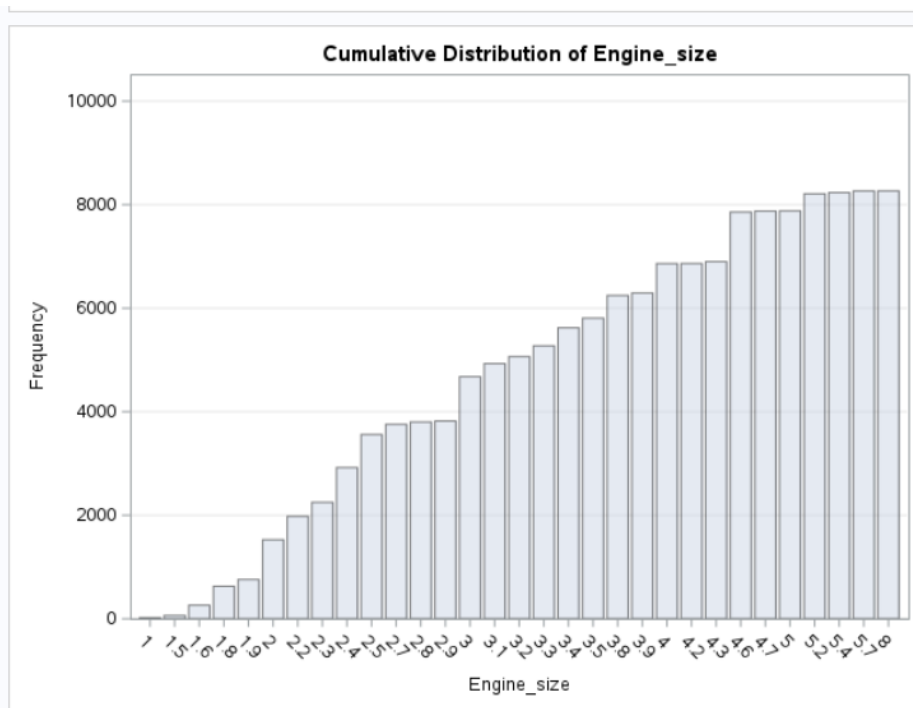
Engine_size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21.855	0.26	21.855	0.26
1.5	41.184	0.50	63.039	0.76
1.6	199.685	2.42	262.724	3.18
1.8	365.396	4.42	628.12	7.60
1.9	130.891	1.58	759.011	9.18
2	768.317	9.29	1527.328	18.47
2.2	451.974	5.47	1979.302	23.94
2.3	271.502	3.28	2250.804	27.23
2.4	670.368	8.11	2921.172	35.33
2.5	639.019	7.73	3560.191	43.06
2.7	196.569	2.38	3756.76	45.44
2.8	45.538	0.55	3802.298	45.99
2.9	18.969	0.23	3821.267	46.22
3	856.21	10.36	4677.477	56.58
3.1	252.428	3.05	4929.905	59.63
3.2	134.523	1.63	5064.428	61.26
3.3	209.425	2.53	5273.853	63.79
3.4	348.764	4.22	5622.617	68.01
3.5	186.249	2.25	5808.866	70.26
3.8	438.449	5.30	6247.315	75.57
3.9	47.805	0.58	6295.12	76.15
4	566.351	6.85	6861.471	83.00
4.2	1.38	0.02	6862.851	83.01
4.3	36.791	0.45	6899.642	83.46
4.6	959.393	11.60	7859.035	95.06
4.7	18.961	0.23	7877.996	95.29
5	4.265	0.05	7882.261	95.34
5.2	328.384	3.97	8210.645	99.32
5.4	22.925	0.28	8233.57	99.59
5.7	32.732	0.40	8266.302	99.99
8	0.916	0.01	8267.218	100.00
Frequency Missing = 53.48				



We can see how the engine sizes are distributed across different categories. For instance the most common engine size is 2.5, with 7.73% of the cars falling into this category.

The distribution appears to be centered around an engine size of approximately 2.5. This means that a significant portion of the cars likely have engines in this range.

The distribution is moderately spread out, with engine size ranging from 1 to 3. This indicates some variability in engine size across the dataset.



The Cumulative columns helps you understand the distribution in terms of how many cars have an engine size less than or equal to a certain value.

For Ex: 43.06% of the cars have an engine size of 2.5 or less.

The plot shows an overall ascending trend, indicating that as the engine size increases, the cumulative frequency also increases. This is expected in an cumulative distribution.

The plot reaches a plateau around an engine size of 2.5, suggesting that a majority of the cars have engine size less than or equal to 2.5.

We can use the cumulative plot to identify threshold for engine sizes, such as the engine size at which 50% or 90% of the car below.

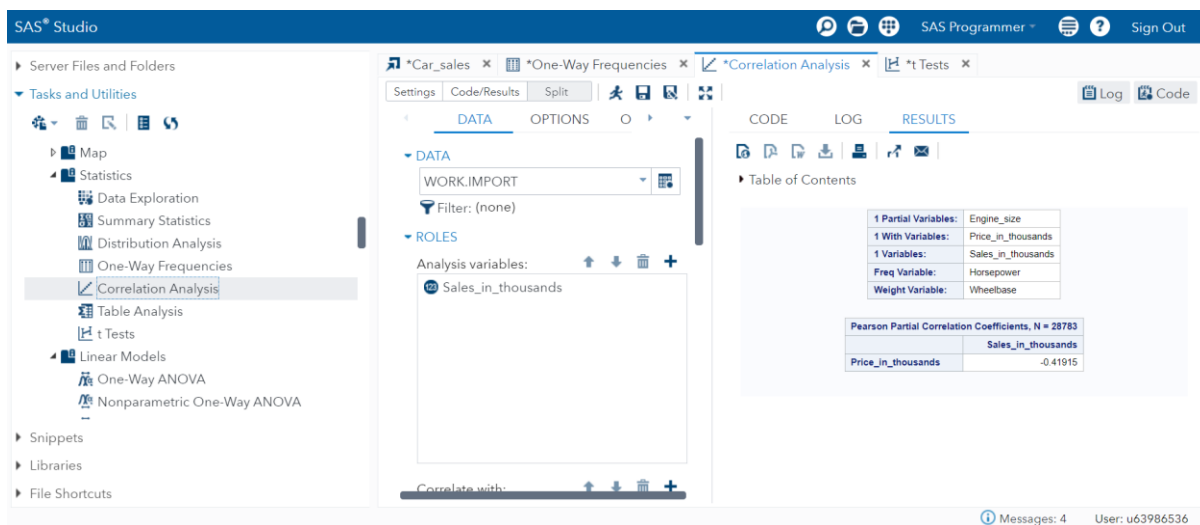
Adding a density curve to the cumulative distribution plot can help visualize the underlying distribution more clearly. Calculating Percentiles and quantiles can provide numerical summaries of the distribution.

Correlational Analysis

What is Correlation Analysis?

Correlation Analysis is a statistical method used to measure the strength and direction of a linear relationship between two variables. In simpler terms, it tells us how much two variables change together.

The graph below provided appears to be correlational matrix, which is a table showing correlation coefficients between multiple variables.



Some of the variables from the above pictures are

Sales in thousands: This is likely represents the number of cars sold in thousands of units.

Price in thousands: This is likely represents the average price of cars in thousands of dollars.

Engine Size: This is likely represents the average engine size of the cars.

Horsepower: This is likely represents the average horsepower of the cars.

Wheelbase: This is likely represents the average wheelbase length of the cars.

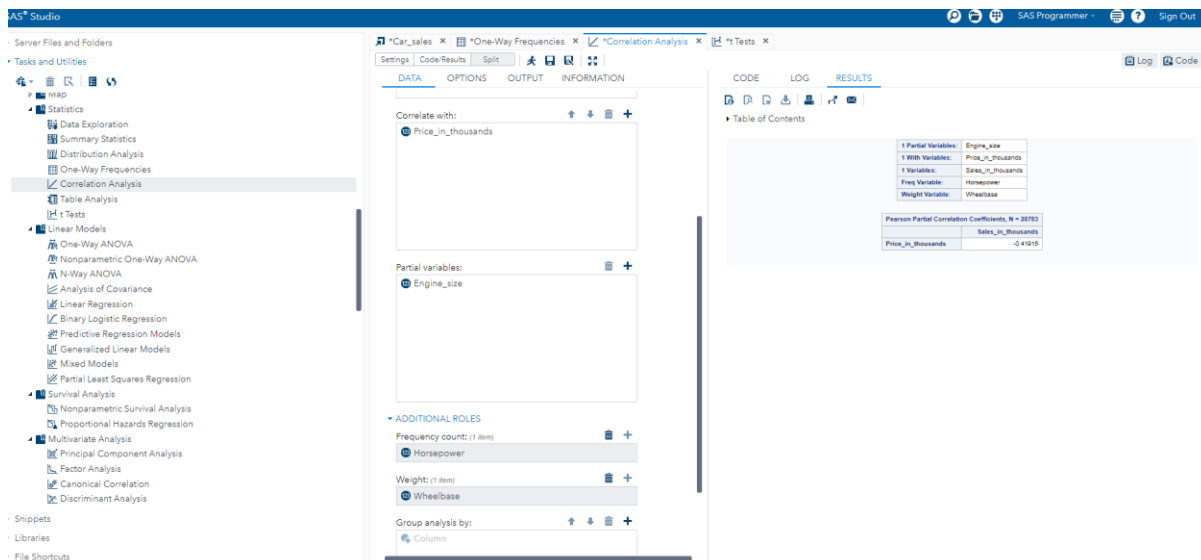
The number table range from -1 to 1.

Positive Value indicates a positive correlation, meaning that as one variable increases, the other also tends to increase.

Negative Value indicates a negative correlation, meaning that as one variable increases, the other also tends to decrease.

Values closer to 1 or -1 indicate a strong correlation.

Values closer to 0 indicate a weaker correlation or no correlation.



1 Partial Variables:	Engine_size
1 With Variables:	Price_in_thousands
1 Variables:	Sales_in_thousands
Freq Variable:	Horsepower
Weight Variable:	Wheelbase

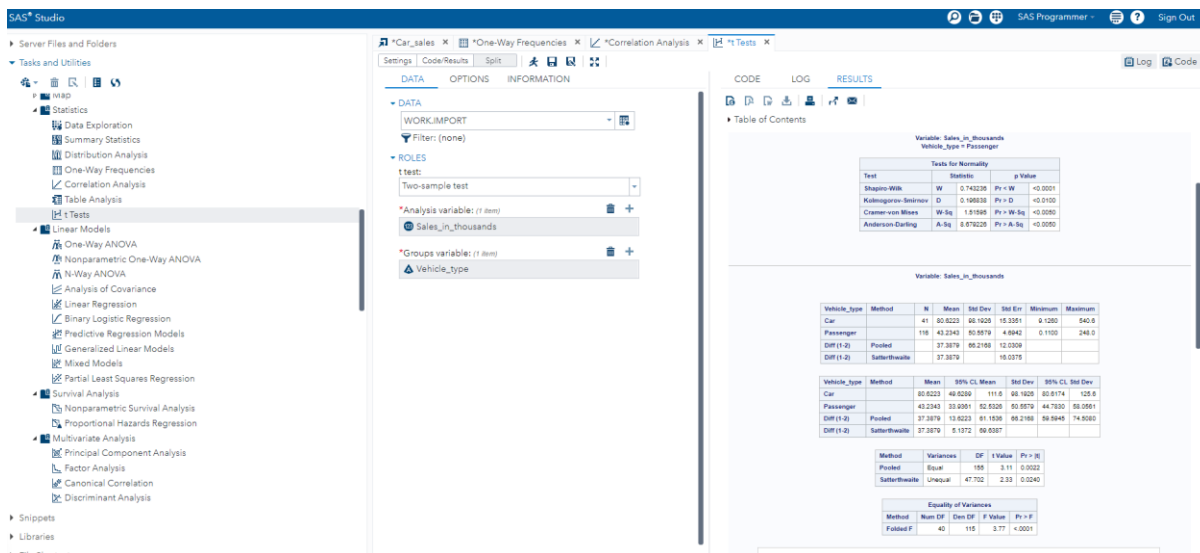
Pearson Partial Correlation Coefficients, N = 28783	
	Sales_in_thousands
Price_in_thousands	-0.41915

Sales in thousands and Price in thousands: The correlation coefficient is 0.41915. This indicates a moderate positive correlation. This means that, generally as the price of cars increases, the sales tend to increase as well. However this relationship is not very strong.

If the p-value associated with a correlation coefficient is less than 0.05, it is considered statistically significant, meaning that the correlation is unlikely to be due to chance.

T-Test

The Summary statistics show that on average, passenger vehicles have higher sales than pickup vehicles. However, there is also more variability in sales among passenger vehicles.



The provided tests (Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling) are used to assess whether a given dataset follows a normal distribution.

P-Value: All p-value less than 0.05 for both vehicle types (Car and Passenger)

Significance Level: A p-value less than 0.005 indicates that we reject the null hypothesis that the data is normally distributed.

Variable: Sales_in_thousands
Vehicle_type = Car

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.865083	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.242024	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.733187	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.867958	Pr > A-Sq	<0.0050

Variable: Sales_in_thousands
Vehicle_type = Passenger

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.743236	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.196838	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.51595	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	8.679226	Pr > A-Sq	<0.0050

Variable: Sales_in_thousands

Vehicle_type	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Car		41	80.6223	98.1926	15.3351	9.1260	540.6
Passenger		116	43.2343	50.5579	4.6942	0.1100	248.0
Diff (1-2)	Pooled		37.3879	66.2168	12.0309		
Diff (1-2)	Satterthwaite		37.3879		16.0375		

Vehicle_type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Car		80.6223	49.6289	111.6	98.1926	80.6174	125.6
Passenger		43.2343	33.9361	52.5326	50.5579	44.7830	58.0561
Diff (1-2)	Pooled	37.3879	13.6223	61.1536	66.2168	59.5945	74.5080
Diff (1-2)	Satterthwaite	37.3879	5.1372	69.6387			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	155	3.11	0.0022
Satterthwaite	Unequal	47.702	2.33	0.0240

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	40	115	3.77	<.0001

Car: N = 41, Mean = 80.6223, Std Dev = 98.1926

Passenger: N = 116, Mean = 43.2343, Std Dev = 50.5579

Difference (Car – Passenger) : Mean = 37.3879, Std Dev(pooled) = 66.2168, Std Dev(Satterthwaite) = 16.0375

t-test for Equal Variances(Pooled):

t-value = 3.11

p-value = 0.0022

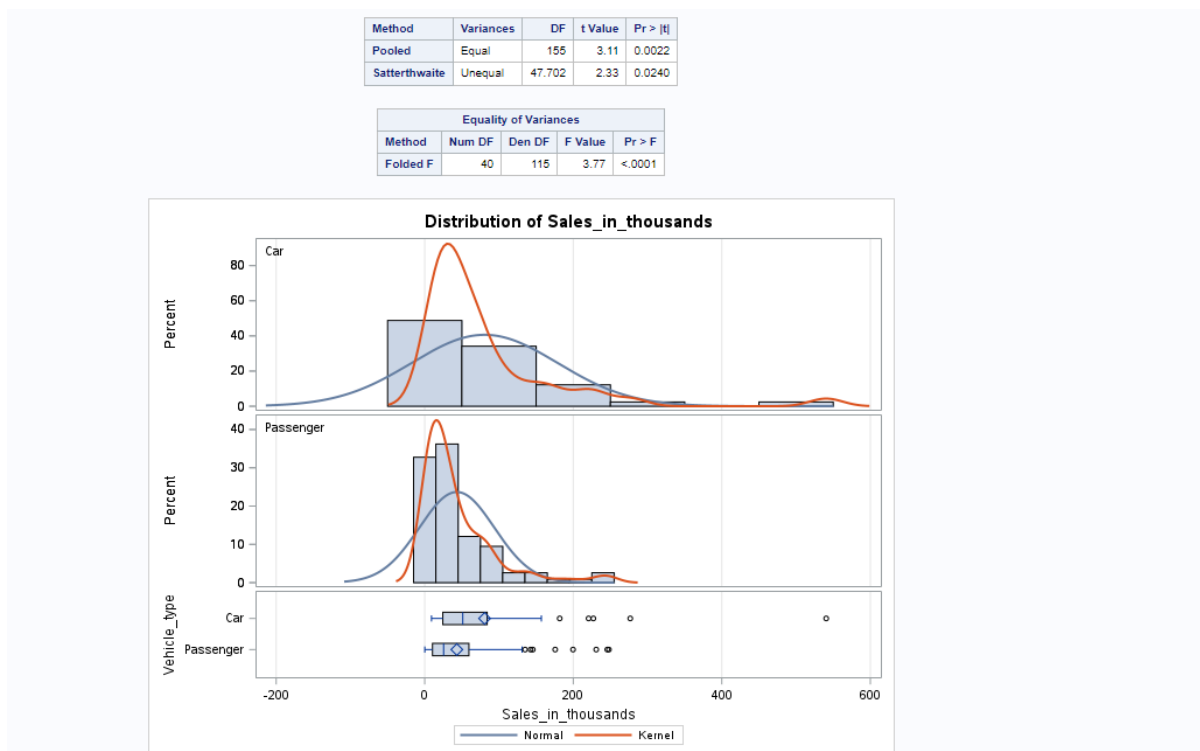
This indicates a significant difference between the means of the groups assuming equal variances.

t-test for Unequal Variances(Satterthwaite):

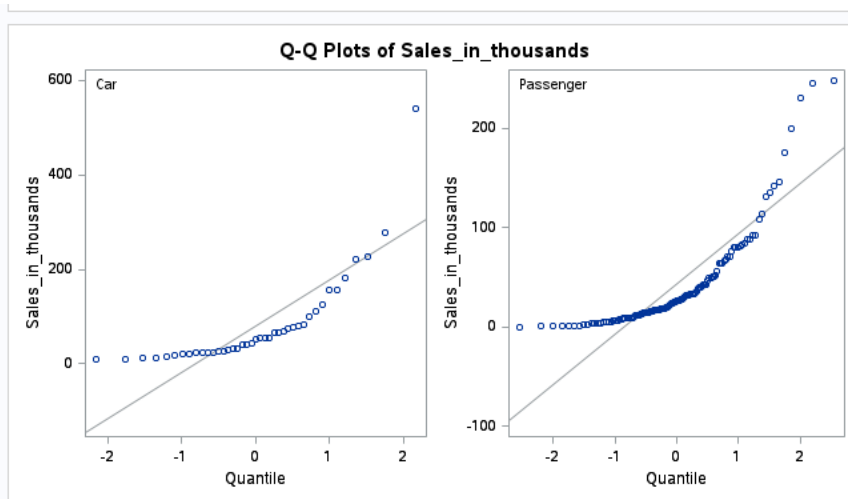
t-value = 2.33

p-value = 0.0240

This also indicates a significant difference between the means, but the conclusion is based on the assumption of unequal variances.



Both t-tests suggest a significant difference in sales between “Car” and “Passenger” Vehicle types. The **average sales for car vehicles are significantly higher than those for passenger vehicles**



Q-Q Plots Quantile-Quantile plots are used to visually assess whether a dataset follows a specific distribution, in this case a normal distribution

Car Sales: The Point in the Q-Q plot for “Car” sales deviate significantly from the straight line, especially in the tails. This indicates that the distribution of “Car” sales is not normal. It has heavier tails than a normal distribution, suggesting the presence of outliers or extreme values.

Passenger Sales: The points in the Q-Q plot for "Passenger" sales also deviate from the straight line, but to a lesser extent. This suggests that the distribution of "Passenger" sales is not perfectly normal, but it's closer to a normal distribution compared to "Car" sales.

Possible Alternatives

These tests do not assume normality. The Mann-Whitney U test is a suitable alternative to the t-test for comparing two independent groups.