Fall 2022

CIS: 5250 Visual Analytics

SAS Project on Car Sales Analysis

Yadav, Pratiksha V

Table of Contents

A. Introduction:

In this fast-paced world, managing your personal and professional lives can be hectic. If you don't have your own personal mode of transportation. And owning a car is very important since it provides the opportunity to travel long distances due to a lack of public transport, especially in densely populated areas. And one of the main reasons why cars have become more prevalent and important is because it is an easy mode of transportation *[2]*. But not everyone can afford a brand-new car. Most people prefer buying a vehicle that is previously used. This project uses a dataset from Kaggle and performs analysis on Car Sales in the US *[1]*. SAS Studio for academics this software has been used for creating visualizations for this project. The SAS studio allows the user to create reports to collaborate, involving discussions that can run deeper insights and construct better decisions. With its easy-to-use features, users can gain precise insights after using the SAS studio tool. The Car sales dataset from Kaggle contains information related to the car's make with respect to the year its price. Through our analysis, we want to find out which are the most popular brands of cars in the US. Through the analysis, we aim to get to know the Used and New car Market in the US better. Furthermore, this analysis will be helpful for businesses to identify the market opportunities and areas where they could include volume. It can be useful for managers to make better data-driven decisions moving forward.

For analysis purposes, 10 years of data in the US from the year 2012 to 2022 has been used. It contains a total of 8362 rows and 11 usable columns.

Dataset URL: https://www.kaggle.com/datasets/chancev/carsforsale

B. Data Description:

Data is an important source of information in any kind of project as it allows the analyst to build relationships between what is happening in different measures and categories.

Field Name	Data Description	Example Values
Year	Car Make year Data Type: Integer	2012, 2013, 2019, etc.
Make	Brand of the car Data Type: Char	Acura, Honda, Toyota, etc.
Used/New	Whether the car being sold is Used or New Data Type: Char	Used or New
Price	The listing price of the Car in dollar amount Data Type: Integer	39998, 49985, 53000, etc.
Consumer Rating	Average consumer rating based on submitted consumer reviews. Rating Range: 1-5 Data Type: Float	3, 4, 4.5, 5, etc.
Performance Rating	How consumers rated the car's performance	3, 4, 4.5, 5, etc.

	Rating Range: 1-5 Data Type: Float	
Seller Rating	Average seller ratings based on submitted seller reviews Rating Range: 1-5 Data Type: Float	1, 2, 3, 3.5, 5, etc.
State	State of the seller's location Data Type: Char	FL, CA, TX, etc.
Fuel Type	Type of fuel that the car uses. Data Type: Char	Gasoline, Electric, Hybrid, etc.
Drive Train	The Drive train type of the car Data Type: Char	AWD, FWD, RWD, 4WD
Mileage	Number of miles on the car Data Type: Integer	105469, 10458, 14940

C. Data Cleaning:

The data which we get from the official sites is usually raw data. This data needs to be cleaned to make it easier to analyze data. We need to remove incorrect, duplicate, or incomplete data. Various data cleaning methods can be used for Data cleaning, these may vary from Dataset to dataset. Listed below are some of the Data Clearing methods which have been used to clean the Cars Dataset. The Data has been cleaned using Microsoft Excel.

1. Missing Values:

Data Cleaning can be handled in many ways, one way is fixing missing values. The Price column values are missing. I have imputed appropriate values based on the make and model of the car. The screenshot below highlighted in Yellow displays the missing values.

Before cleaning

Year	Make	Model	Used/New	Price
2019	Honda	Civic Sport	Used	
2019	BMW	540 i xDrive	Used	
2019	Jeep	Cherokee Limited	Jeep Certified	
2019	Honda	Civic Sport	Used	
2020	Honda	CR-V LX	Used	

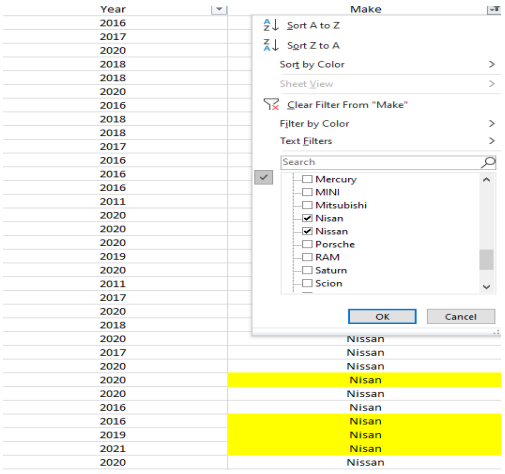
After cleaning

Year	Make	Model	Used/New	Price
2019	Honda	Civic Sport	Used	25979
2019	BMW	540 i xDrive	Used	39950
2019	Jeep	Cherokee Limited	Jeep Certified	23891
2019	Honda	Civic Sport	Used	25979
2020	Honda	CR-V LX	Used	26500

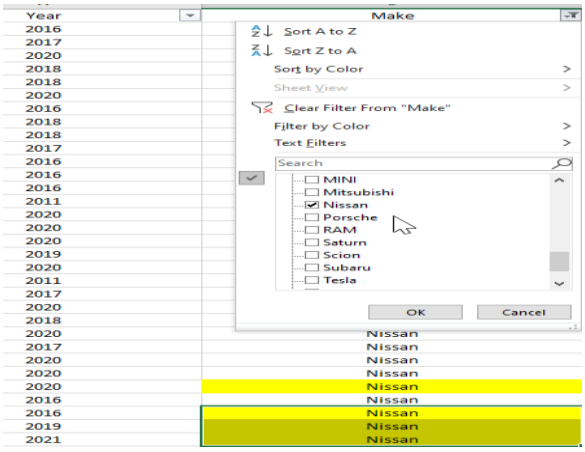
2. Misspellings:

It is very important to fix the Misspellings in a data set, as this may cause ambiguity. As shown in the screenshot below the category ‘Make’- **Nisan** is spelled Incorrectly. Hence replaced the incorrect value of **Nisan** with the Correct Value, for the whole Column- ‘Make’. After correcting the Misspelling value, we can see only one filter with the ‘Make’-**Nissan** in the after-cleaning figure.

Before Cleaning



After cleaning



3. Inconsistent Data:

As we can see from the screenshot below, there are multiple abbreviations of one state. Hence that would not give be appropriate to analyze and visualize it. Hence, I will change the ‘AZ-101’ to ‘AZ’ to match the data consistently throughout.

Before Cleaning

ConsumerRating	State
4.8	AZ
4.1	AZ-101
4.7	AZ
4.8	AZ
4.9	AZ
4.8	AZ
4.5	AZ-101
4.8	AZ
4.8	AZ
4.4	AZ-101

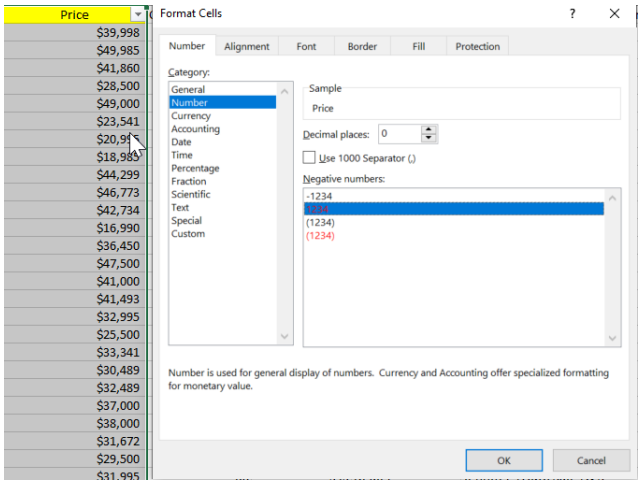
After cleaning

ConsumerRating	State
4.8	AZ
4.1	AZ
4.7	AZ
4.8	AZ
4.9	AZ
4.8	AZ
4.5	AZ
4.8	AZ
4.8	AZ
4.4	AZ

4. Remove Currency Formatting

The Price column has the currency as $. It is important to convert it into an Integer for better analysis and visualization. In the excel workbook, we can remove the currency using the Format option in the home tab. This helps us to better display the result.

Before cleaning



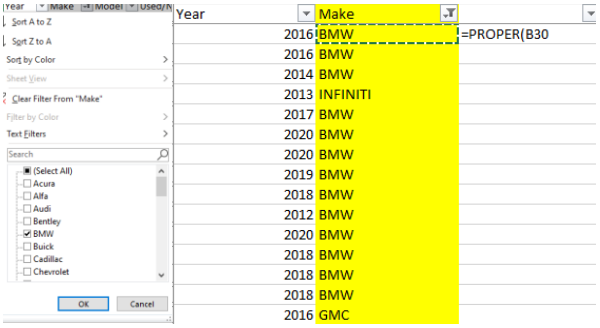
After Cleaning

Make	Model	Used/New	Price
Toyota	Sienna SE	Used	39998
Ford	F-150 Lariat	Used	49985
RAM	1500 Laramie	Used	41860
Honda	Accord Sport SE	Used	28500
Lexus	RX 350	Used	49000
Toyota	4Runner SR5	Used	23541
Honda	HR-V LX	Used	20995
Mercedes-Benz	E-Class E 350	Used	18985
Honda	Pilot Touring 8-Pas	Used	44299
Dodge	Charger Scat Pack	Dodge Certified	46773
Lexus	RC 350 Base	Used	42734
Lexus	GS 350 Base	Used	16990
Lexus	RX 350 RX 350	Used	36450
Dodge	Charger Scat Pack	Dodge Certified	47500
Dodge	Charger Scat Pack	Used	41000
Lexus	RC 350 Base	Used	41493
Ford	Explorer XLT	Used	32995
Subaru	Crosstrek Premium	Used	25500
Ford	Explorer XLT	Used	33341
Ford	Explorer XLT	Used	30489
Ford	Explorer XLT	Used	32489

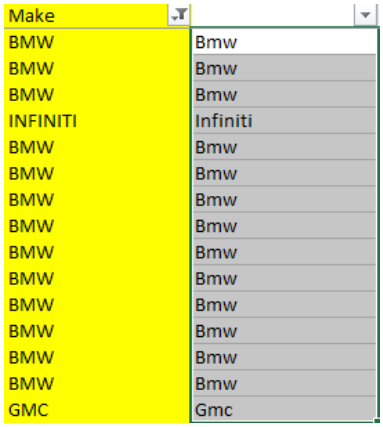
5. **Changing the case of text:**

It is very important to have a sense of characters how to look and act and maintain uniformity throughout. The Make column in the dataset has the brand names of the car and all the names do not have uniformity amongst them. Hence, In the excel sheet, we are changing the Upper case to the lower case using the PROPER function which converts the text to lower case. First, I have inserted a temporary column next to the column that contains the text we want to convert. Likewise, I followed the quickest way to fill other columns by using the drag option to reflect it.

Before Cleaning



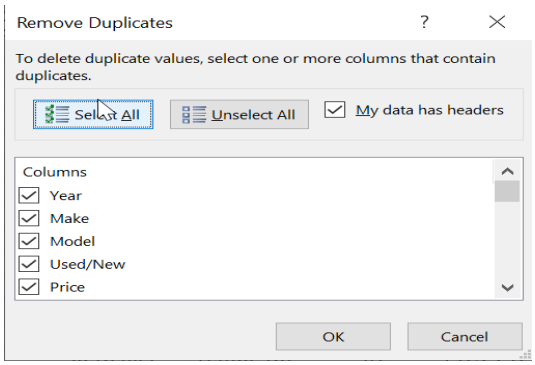
After Cleaning



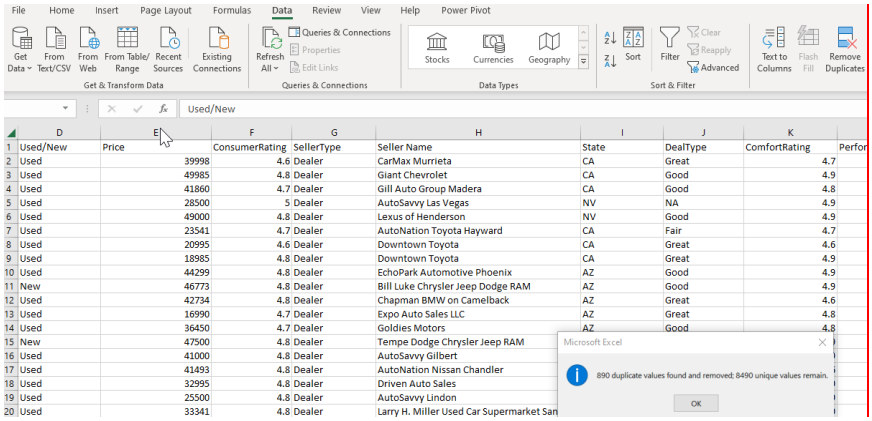
6. Duplicate Data:

Duplicate records cause confusion. It also causes unnecessary wastage of memory, hence duplicate rows and columns should be removed from a data set. It is necessary to check for duplicates in the excel sheet. Hence, I deleted the duplicate data from the dataset using Remove Duplicates in the Data Tools group. As shown below.

Before Cleaning



After cleaning



	D	E	F	G	H	I	J	K
1	Used/New	Price	ConsumerRating	SellerType	Seller Name	State	DealType	ComfortRating
2	Used	39998	4.6	Dealer	CarMax Murrieta	CA	Great	4.7
3	Used	49985	4.8	Dealer	Giant Chevrolet	CA	Good	4.9
4	Used	41860	4.7	Dealer	Gill Auto Group Madera	CA	Good	4.8
5	Used	28500	5	Dealer	AutoSavvy Las Vegas	NV	NA	4.9
6	Used	49000	4.8	Dealer	Lexus of Henderson	NV	Good	4.9
7	Used	23541	4.7	Dealer	AutoNation Toyota Hayward	CA	Fair	4.7
8	Used	20995	4.6	Dealer	Downtown Toyota	CA	Great	4.6
9	Used	18985	4.8	Dealer	Downtown Toyota	CA	Great	4.9
10	Used	44299	4.8	Dealer	EchoPark Automotive Phoenix	AZ	Good	4.9
11	New	46773	4.8	Dealer	Bill Luke Chrysler Jeep Dodge RAM	AZ	Good	4.9
12	Used	42734	4.8	Dealer	Chapman BMW on Camelback	AZ	Great	4.6
13	Used	16990	4.7	Dealer	Expo Auto Sales LLC	AZ	Great	4.8
14	Used	36450	4.7	Dealer	Goldies Motors	AZ	Good	4.8
15	New	47500	4.8	Dealer	Tempe Dodge Chrysler Jeep RAM			
16	Used	41000	4.8	Dealer	AutoSavvy Gilbert			
17	Used	41493	4.8	Dealer	AutoNation Nissan Chandler			
18	Used	32995	4.8	Dealer	Driven Auto Sales			
19	Used	25500	4.8	Dealer	AutoSavvy Lindon			
20	Used	33341	4.8	Dealer	Larry H. Miller Used Car Supermarket San			

D. Analysis and Data Visualization:

1. What is the percentage of different Drivetrains for the cars sold?

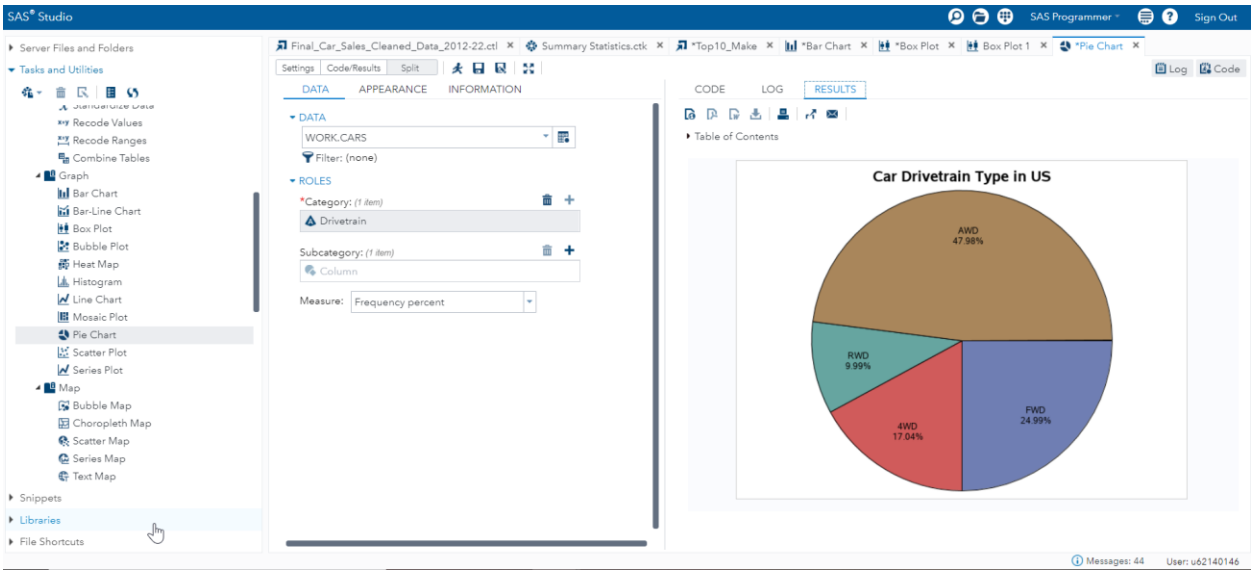


Figure 1.1 Drivetrain types of the cars sold in the US from the year 2012-2022

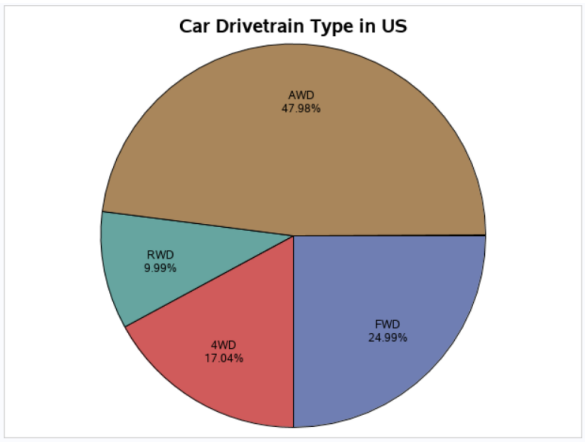


Figure 1.2 Drivetrain types of the cars sold in the US from the year 2012-2022

Data Roles:

Category: Drivetrain

Measure: Frequency percent

In figure 1.2, the pie chart above depicts the distribution of 4 distinct drivetrain types in the US. The drivetrain type is a significant statistic when buying a new car or used, the type of the powertrain is an incredibly important decision *[3]*. Based on the dataset, AWD (All-Wheel Drive) vehicles, represent the first Brown slice, which has just less than 50% of the cars sold. The most beneficial thing about driving an AWD vehicle is the fact that they are dynamic and adaptable to almost any kind of terrain or road conditions. Secondly, FWD (Front Wheel Drive) marks blue on the chart with about 25%. Under normal conditions, FWD vehicles can serve well. 4WD (Four Wheel Drive) shows about 17% of the cars sold. Four-wheel drive is preferred in deep snow mud, rough or rocky terrain as well as sharp inclines or declines. Since 4WD systems tend to be more robust than AWD systems, they can handle more abuse and power. It's also a good idea for drivers that live in isolated areas or routinely drive out into remote zones to have access to a 4WD vehicle. While the RWD (Rear Wheel Drive) has 10% population. These vehicles are high in a performance like muscle cars or sports cars. This analysis helps us to know that choosing the appropriate drivetrain is important when buying a car depending on Driving preferences, Weather conditions where you live, Typical Road conditions, daily driving routine, and how you plan to use your vehicle.

2. How is the trend of New and Used Car Sales in the US?

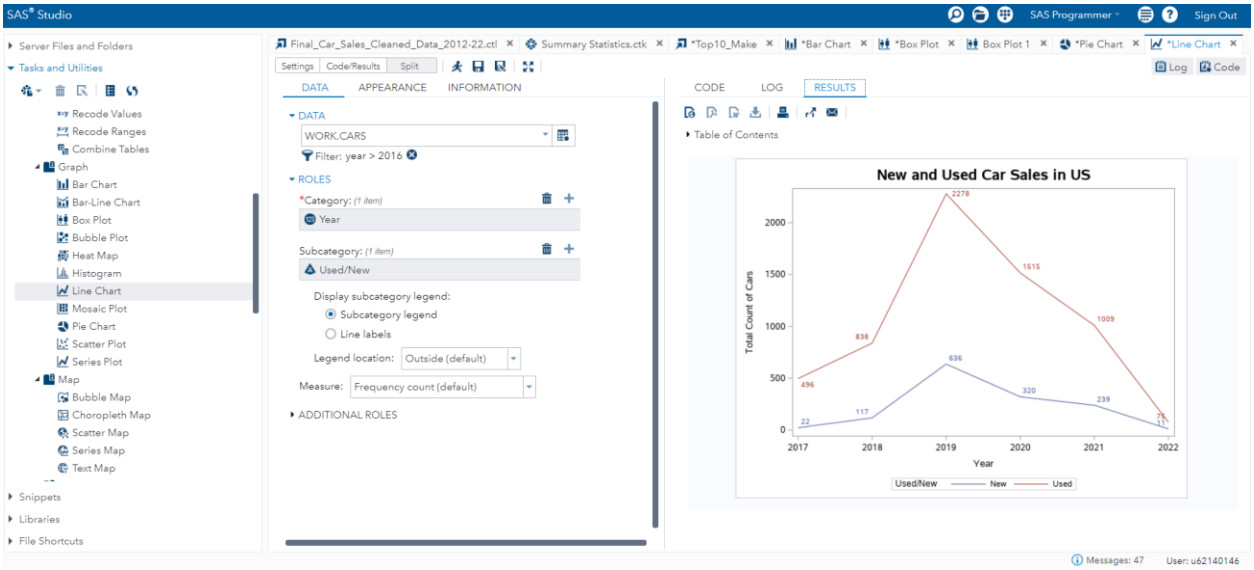


Figure 2.1 Sales of New and Used cars in the US from 2017 to 2022

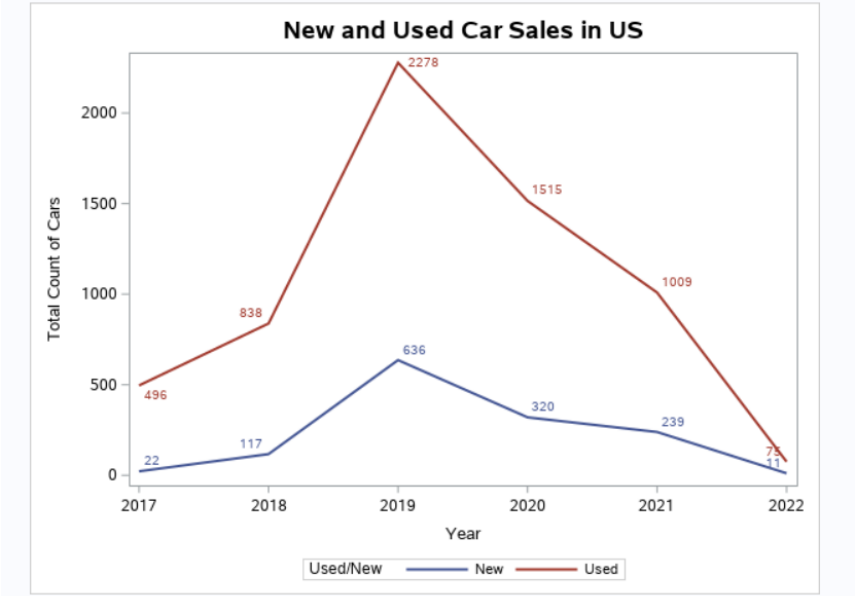


Figure 2.2 Sales of New and Used cars in the US from 2017 to 2022

Data Roles:

Category: Year

Subcategory: Used/New

Measure: Frequency Count

Figure 2.2 Line graph illustrates the New and Used car sales trend in the US from the year 2017 to 2022. Two different colors have been used for analysis. The Blue color represents the number of new cars, and the Blue is the number of Used cars. At the beginning of the year 2017, we could see a gradual rising trend in New and Used car sales. During the year 2018 to 2019, there was a sudden rise in car sales both New and Used. Car sales reached their peak in the year 2019.

The decrease in the trend of **New Car** after the year 2019 shows that fewer cars were manufactured in 2020 which can be due to the pandemic crisis, the decrease in the year 2021 onwards can be attributed to temporal proximity to the data collected. This may also be due to the supply chain shortage in the market. As the cars are already newer, users are less likely to sell them, this can be confirmed by the decrease in the number of **Used Cars** after the year 2019.

We can see that used car sales are always higher than new car sales. This could be because not everyone would have a budget for purchasing new cars which are expensive comparatively. These insights would be helpful for the business to maintain the inventory of the cars.

Let us now look at the ratings that the Used Car and New Cars sellers hold in the US from the year 2012-22 in below figure 2.4

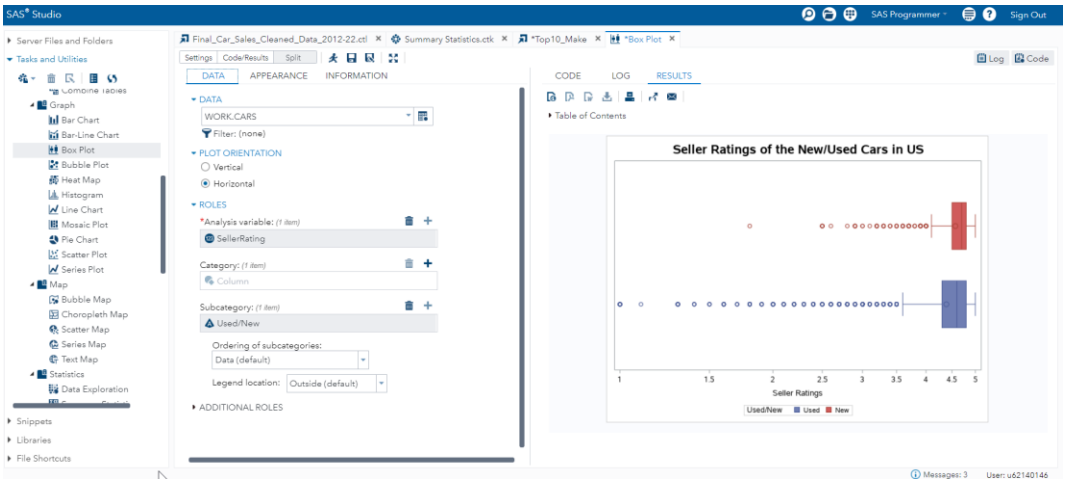


Figure 2.3 Used/New cars Sellers Ratings

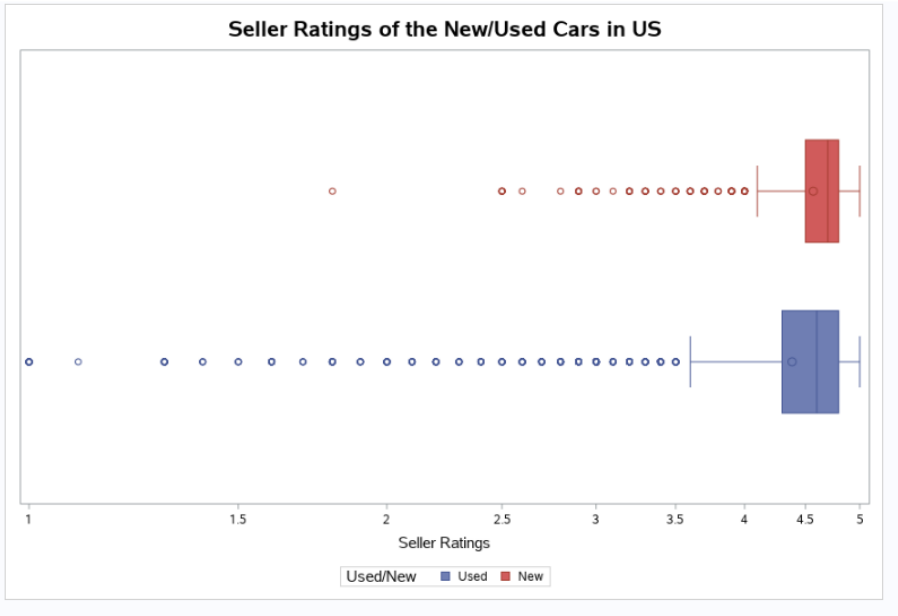


Figure 2.4 Used/New cars Sellers Ratings

Data roles:

Analysis variable: Seller ratings

Subcategory: Used/New Cars

As we can see from Figure 2.4 that the median value of the seller's ratings for new cars in the US is around 4.7, with 80% of the sellers rating above 4.6. Likewise, surprisingly the used car sellers' ratings are around 4.6, with 75% of the sellers having a rating of 4.7. This clearly shows that the sellers are doing a pretty good job when it comes to customer satisfaction for both Used and New cars, be it a car seller or a guy buyer. Both parties are rating the sellers quite high in general which is great for these sellers considering how competitive and technical the domain is *[3]*.

3. How is the Distribution of Car Makers associated with their Fuel type?

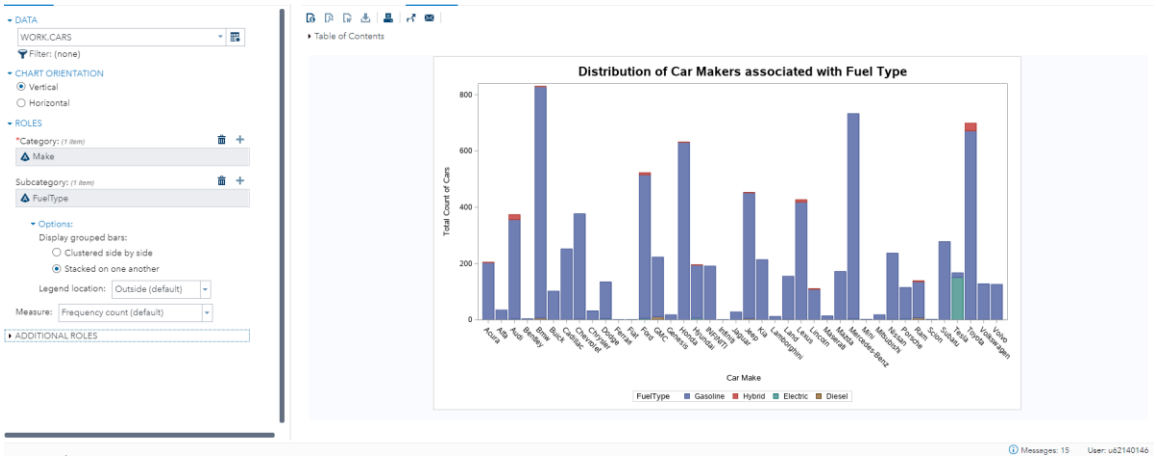


Figure 3.1 Distribution of Car Makers associated with their Fuel Type

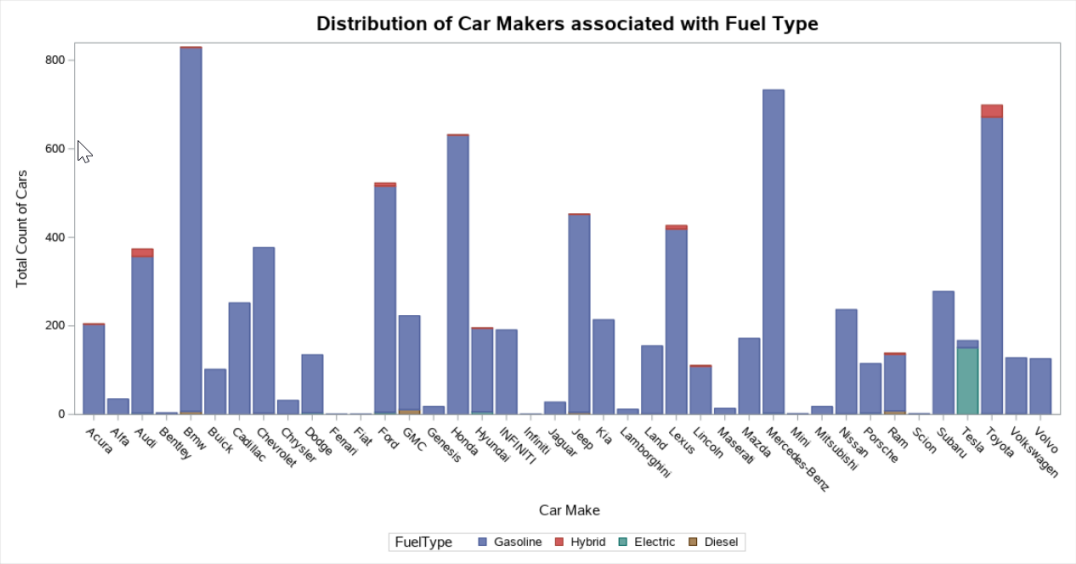


Figure 3.2 Distribution of Car Makers associated with their Fuel Type

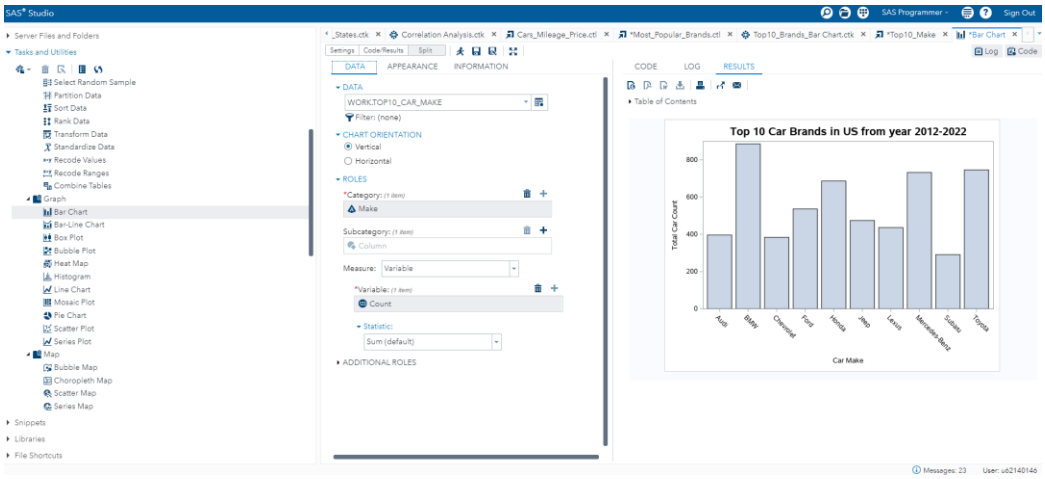


Figure 3.3 Top 10 Car Brands in the US from the year 2012-2022

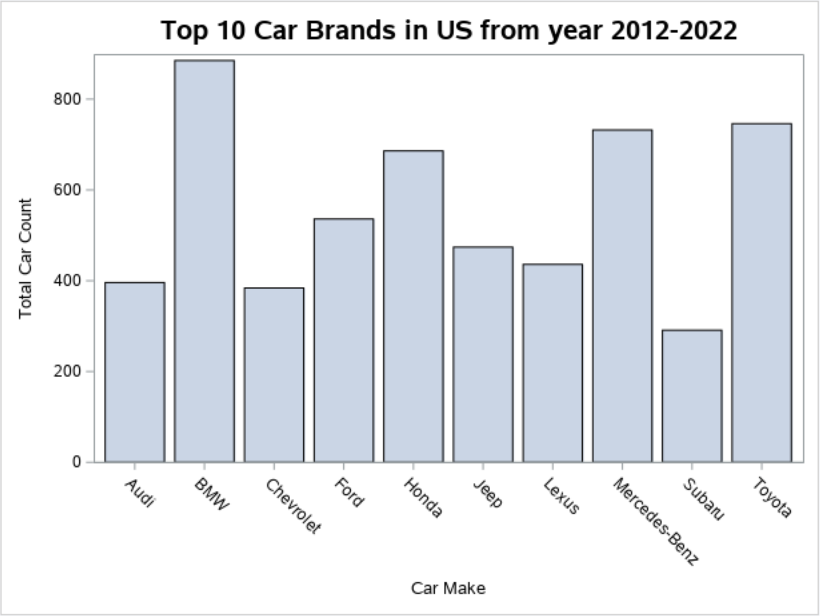


Figure 3.4 Top 10 Car Brands in the US from the year 2012-2022

Data roles:

Category: Make

Subcategory: Fuel Type

Measure: Frequency Count

Figure 3.2 illustrates the distribution of different types of car brands sold in the US. We can see that there are few luxury car makers like Porsche, Ferrari, and Lamborghini whose sales are very less as compared to the other car makers. This may be due to the high price which only a few people can afford to purchase. Most of the electric cars are sold by Tesla as shown in the graph. More and more people are inclined towards electric vehicles nowadays since gasoline prices have skyrocketed. Furthermore, we have visualized the top 10 car brands sold in the US from the year 2012-2022

Data roles:

Category: Make

Measure: Variable

Variable: Count

Figure 3.4 bar graph shows the Top 10 cars sold in the US from the year 2012 - 2022. We can see that BMW has the highest-sold cars followed by Toyota and Mercedes-Benz. BMW being at a slightly higher price point are still the number one selling brand. This could be due to the well-known German build quality which most users prefer. Toyota on the other hand is known for its reliability and economical daily use and stands at number two. Subaru has the least sold cars overall. This might be because Subaru is popular for its sporty and powerful engine which consumes a high amount of fuel. Most consumers prefer economical cars in terms of fuel and daily maintenance costs.

E. Statistical Summary:

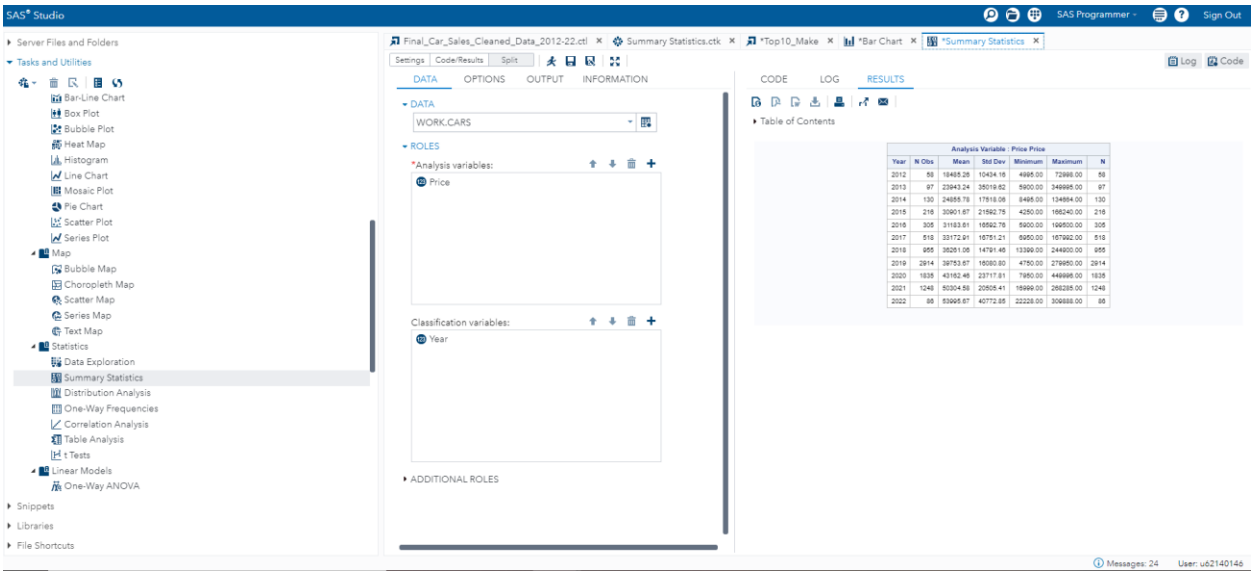


Figure 4.1 Summary Statistics of the car sales price in the US from the year 2012 to 2022

Analysis Variable : Price Price								
Year	N Obs	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss
2012	58	18485.26	10434.16	4995.00	72998.00	17250.00	58	0
2013	97	23943.24	35019.62	5900.00	349995.00	17995.00	97	0
2014	130	24855.78	17518.06	8495.00	134664.00	20845.00	130	0
2015	216	30901.67	21592.75	4250.00	166240.00	24332.00	216	0
2016	305	31183.61	16592.76	5900.00	199500.00	28434.00	305	0
2017	518	33172.91	16751.21	6950.00	167992.00	29992.00	518	0
2018	955	36261.06	14791.46	13399.00	244900.00	33494.00	955	0
2019	2914	39753.67	16080.80	4750.00	279950.00	36000.00	2914	0
2020	1835	43162.46	23717.81	7950.00	449996.00	39981.00	1835	0
2021	1248	50304.58	20505.41	16999.00	268285.00	45905.50	1248	0
2022	86	53995.67	40772.85	22228.00	309888.00	44993.50	86	0

Figure 4.2 Summary Statistics of the car sales price in the US from the year 2012 to 2022

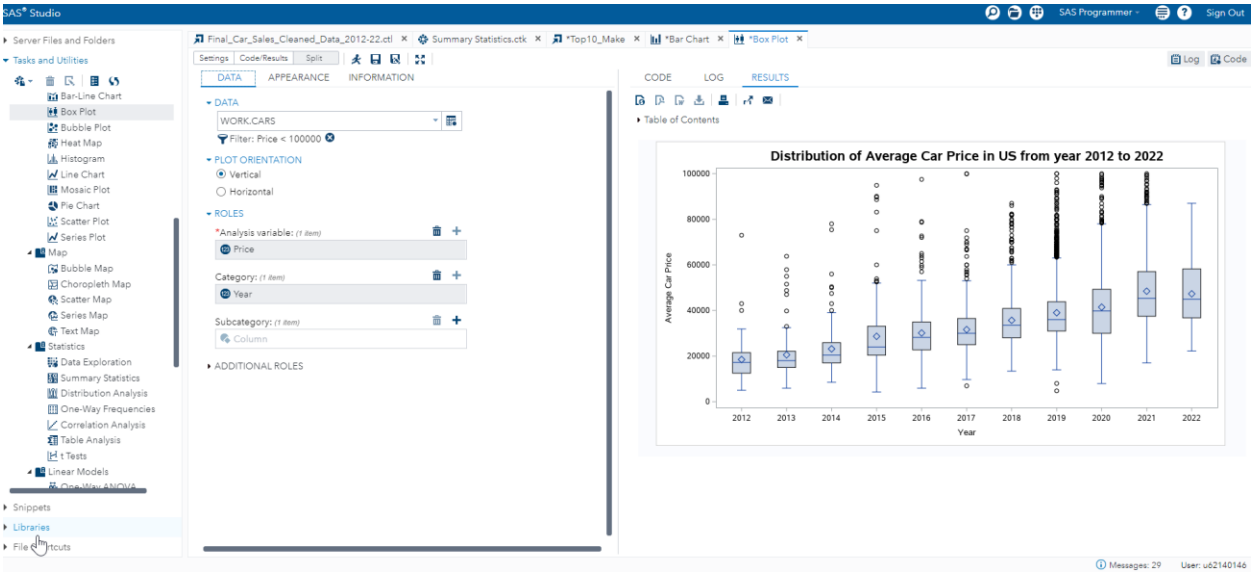


Figure 4.3 Summary Statistics of the car sales price in the US from the year 2012 to 2022

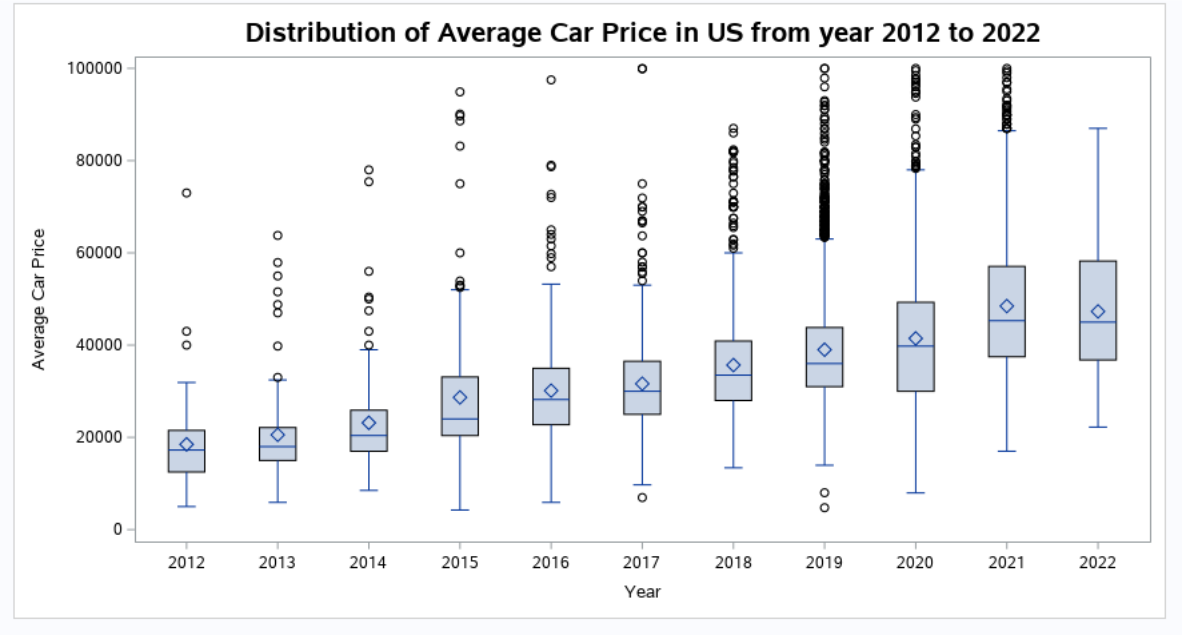


Figure 4.4 Summary Statistics of the car sales price in the US from the year 2012 to 2022

Data Role:

Analysis Variable: Price

Classification Variable: Year

The above table in figure 4.2 describes the summary statistics of the car sales price in the US from the year 2012 to 2022 for a total of 8362 observations. The average price of the cars is the Mean column. We could see the increasing trend in the prices of cars per year. The average price of the car is observed highest in the year 2022 at 53995.67 which describes the Mean and the Minimum price of the car observed as 22228 approximately. The maximum price is observed as 309888 approximately *[4]*. The standard deviation shows the average variation in the price, and it is observed to be 40772.85 for the year 2022. The Median is the middle value in the dataset for the year 2022 which is around 44993.50. The overall trend in figure 4.4 shows the rise in the prices of cars in the given dataset. Figure 4.3 explains that due to inflation and other external factors with each passing year the average prices of cars increased.

F. Statistical Tests:

One-way frequency

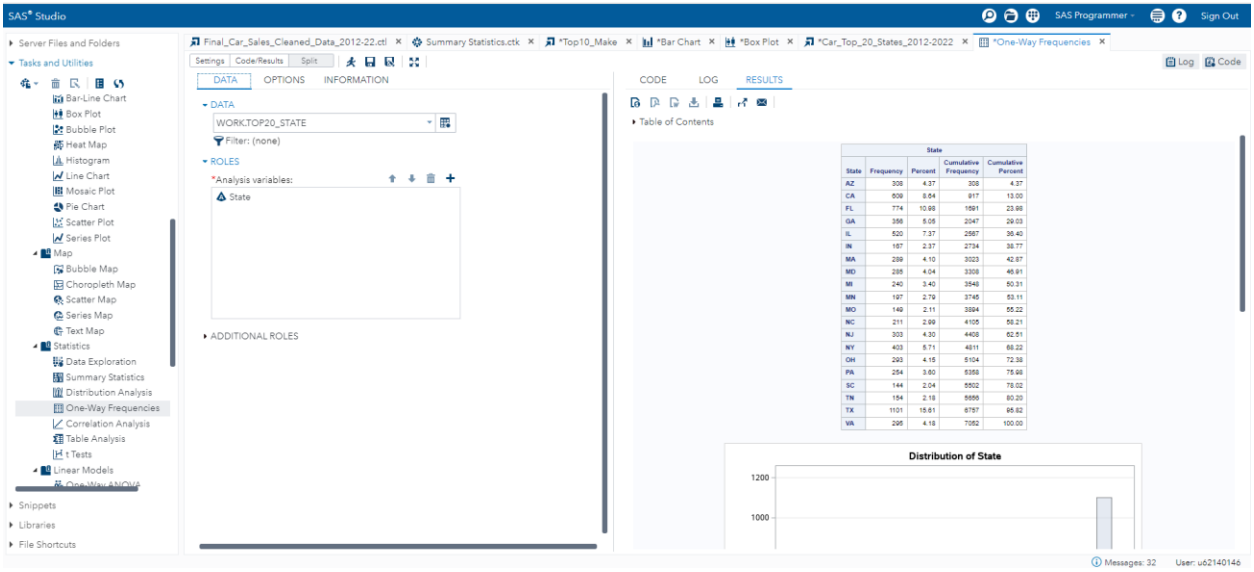


Figure 5.1 One-way Frequency of Top 20 states in the US from the year 2012 to 2022

State				
State	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AZ	308	4.37	308	4.37
CA	609	8.64	917	13.00
FL	774	10.98	1691	23.98
GA	356	5.05	2047	29.03
IL	520	7.37	2567	36.40
IN	167	2.37	2734	38.77
MA	289	4.10	3023	42.87
MD	285	4.04	3308	46.91
MI	240	3.40	3548	50.31
MN	197	2.79	3745	53.11
MO	149	2.11	3894	55.22
NC	211	2.99	4105	58.21
NJ	303	4.30	4408	62.51
NY	403	5.71	4811	68.22
OH	293	4.15	5104	72.38
PA	254	3.60	5358	75.98
SC	144	2.04	5502	78.02
TN	154	2.18	5656	80.20
TX	1101	15.61	6757	95.82
VA	295	4.18	7052	100.00

Figure 5.2 One-way Frequency of Top 20 states in the US from the year 2012 to 2022

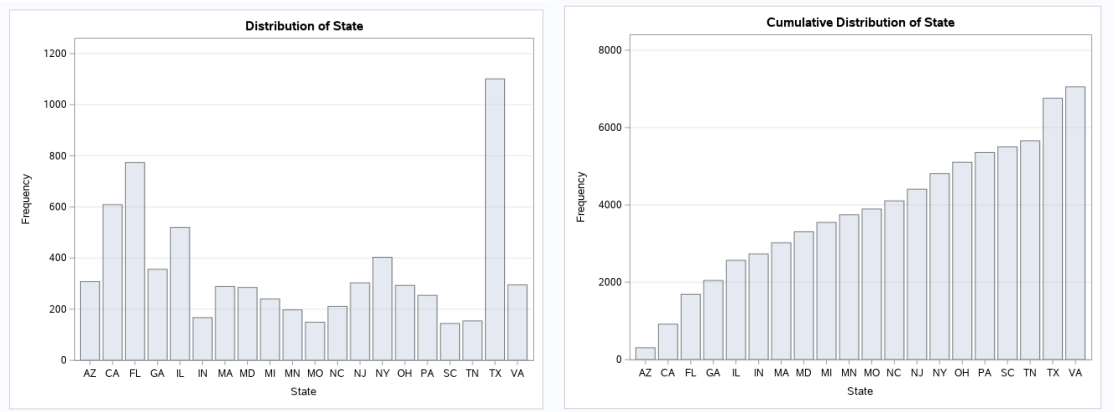


Figure 5.3 One-way Frequency of Top 20 states in the US from the year 2012 to 2022

Data Role:

Analysis Variable: State

The one-way frequency refers to a tabulation of data that only examines one categorial variable at a time. The table reports the Frequency, percent, cumulative frequency, and cumulative percent of the categorial variable ‘State’. These are the top 20 listed states in the US with the highest number of cars sold. By observing this table one can determine that Texas has the highest number of cars sold amongst all other states with the highest frequency of 1101, which makes up 15.61% of the distribution. The cumulative frequency is 6757 and the percentage is 96.82%. On the other hand, the least car sales were in the state of Arizona with a frequency of 308 which is 4.37% of the distribution.

Correlation analysis

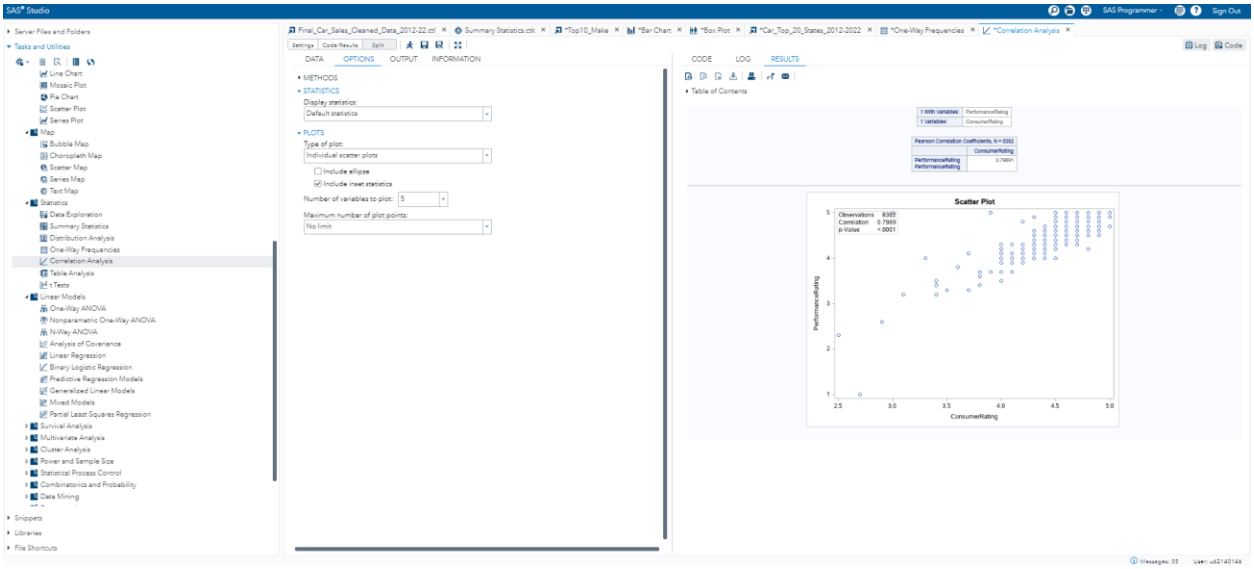


Figure 6.1 Correlation Analysis of Top Consumer Ratings based and Performance Rating

1 With Variables:	PerformanceRating
1 Variables:	ConsumerRating

Pearson Correlation Coefficients, N = 8362	
	ConsumerRating
PerformanceRating PerformanceRating	0.79891

Figure 6.2 Correlation Analysis of Top Consumer Ratings based and Performance Rating

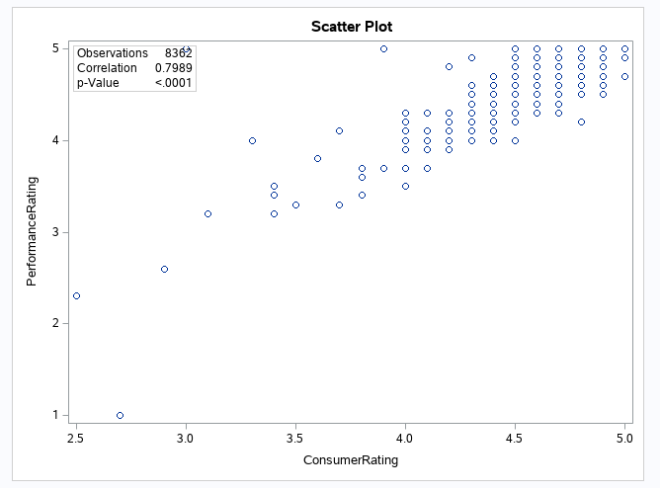


Figure 6.3 Correlation Analysis of Top Consumer Ratings based and Performance Rating

Data Roles:

Analysis variable: Consumer Rating

Correlate with: Performance Rating

Correlation analysis is used to check if there is a relationship between two variables in the dataset and how strong that relationship can be. In terms of car sales correlation analysis is used to identify whether there are any significant patterns, trends, or insights for the business operation team to estimate the consumer's ratings and vehicle performance ratings. This will help the company to know the consumer market better and grow in terms of sales. Essentially, correlation analysis is used for spotting patterns within datasets. A positive relation means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

In above figure 6.3. The scatter plot analyses consumer ratings based on the performance ratings. Here the analysis variable Consumer Rating is an independent variable, and we are correlating it

with the dependent variable Performance Rating. It indicates the positive relationship between two variables with a Pearson Correlation value of 0.7989 is observed in figure 6.2. Any value from +0.5 to +1 indicates a very strong positive correlation, which means that they both increase at the same time *[5]*. The plots follow an upward trend and indicate the positive correlation between Consumer Rating and Performance Rating.

Linear Regression

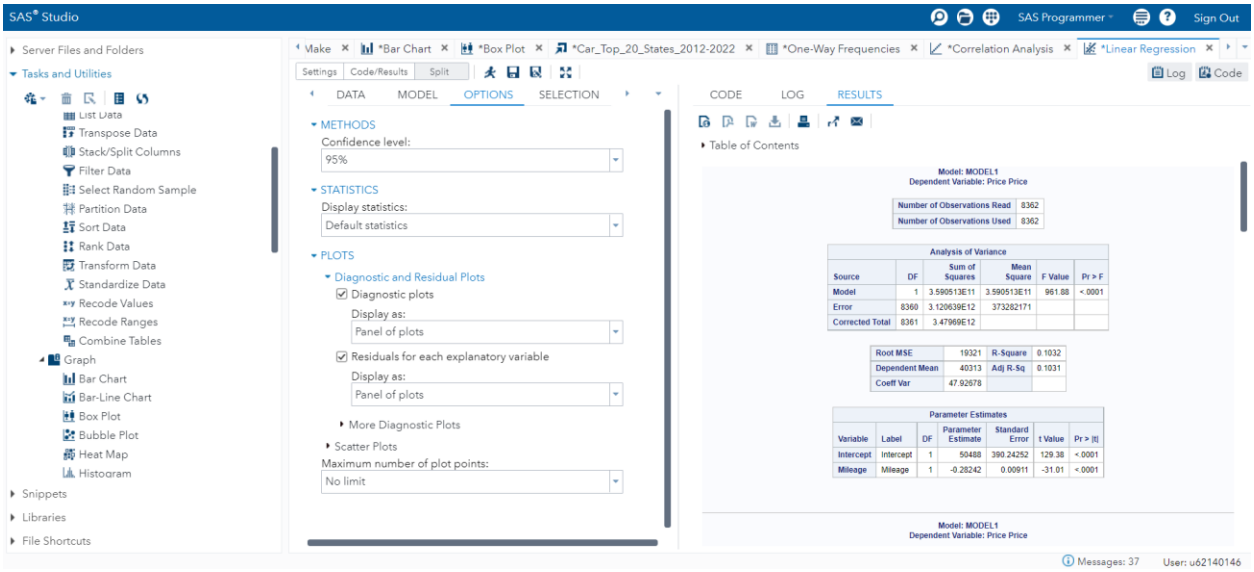


Figure 7.1 Linear Regression Model for Price and Car Mileage

Model: MODEL1
Dependent Variable: Price Price

Number of Observations Read	8362
Number of Observations Used	8362

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.590513E11	3.590513E11	961.88	<.0001
Error	8360	3.120639E12	373282171		
Corrected Total	8361	3.47969E12			

Root MSE	19321	R-Square	0.1032
Dependent Mean	40313	Adj R-Sq	0.1031
Coeff Var	47.92678		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	50488	390.24252	129.38	<.0001
Mileage	Mileage	1	-0.28242	0.00911	-31.01	<.0001

Figure 7.2 Linear Regression Model for Price and Car Mileage

Data Roles:

Dependent Variable: Price

Continuous Variable: Mileage

The above linear model is performed using 2 variables the Mileage and Price of the car. The Price is the dependent variable and the Mileage is the independent variable. It further demonstrates

whether there are co-efficiently significant or not. R square is a primary measure of how well a regression model fits the data.

The estimated line of regression can be written as **y = ax + b**

Where b is the intercept and a is the coefficient of the slope.

From the above linear regression expression, the slope value is **a = -0.28242** and **x variable is the mileage of the car** and the intercept value is **b = 50488**. The table shows the R^2 value is **"0.1032 (10.32%)"** which represents there is a weak correlation between Mileage and Price *[6]*.

In total there are 8362 observations and we do not notice any missing observations. As the assumed value of p = 0.05 is greater than the Probability value = 0.0001 and the R^2 value is 0.1032, which indicates a 10.32% variation in the price can be explained by the mileage of the car. It means that this model doesn't explain much of the variation of the data but it is significant.

G. References:

1. V, C. (2022, July 21). *Carsforsale*. Kaggle. Retrieved October 31, 2022, from https://www.kaggle.com/datasets/chancev/carsforsale.

2. Houghton, L. (2022, May 2). *Why cars are important and why you should invest in one*. Global Cars Brands. Retrieved November 1, 2022, from https://www.globalcarsbrands.com/why-cars-are-important-and-why-you-should-invest-in-one/.

3. Gupta, H. (2021, September 8). *Published in Jovian—Data Science and Machine Learning*. Exploratory Data Analysis of Used Cars in the United States. Retrieved November 1, 2022, from https://blog.jovian.ai/understanding-used-cars-market-in-usa-52489e10d551.

4. *Car Data Statistical Analysis*. myassignmenthelp.com. (2016, November 22). Retrieved November 1, 2022, from https://myassignmenthelp.com/free-samples/car-data-statistical-analysis.

5. James, E. (2022, April 8). *What is correlation analysis? A definition and explanation*. What is Correlation Analysis? A Definition and Explanation. Retrieved November 1, 2022, from https://blog.flexmr.net/correlation-analysis-definition-exploration.

6. Shahbaz, F. (2021, November 5). *Relationship between R-squared and p-value in a regression*. Medium. Retrieved November 1, 2022, from https://medium.com/@faisalshahbaz/relationship-between-r-squared-and-p-value-in-a-regression-2378b1e2e9ce.