**7PAM2000 Applied Data Science 1**

**Assignment 1: Visualization**

**Student Name: Pooja Nagarajan**

**Student id:**

**Link to Dataset:**

**Introduction:**

The USA Cars Dataset is a comprehensive collection of data about cars sold in the United States from the year 1990 to 2021. This dataset contains information on over 1 million car sales, including the make and model of the car, the year it was sold, the price of the car, the mileage, the type of transmission, and much more. The dataset is rich in detail and provides a wealth of information that can be used to gain insights into the US car market.

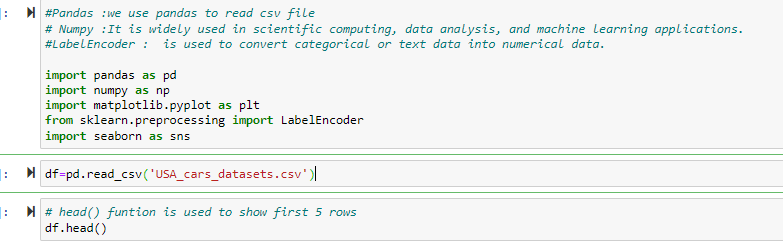
The goal of this dataset is to help you understand how to visualize and analyze data, identify trends and relationships, and make meaningful observations about the data. Through this process, you will learn how to create effective visualizations, choose the right chart types, and communicate your findings to others.

**Data Source:**

https://www.kaggle.com/code/gauravsahani/usa-cars-dataset-visualization-for-beginners/notebook

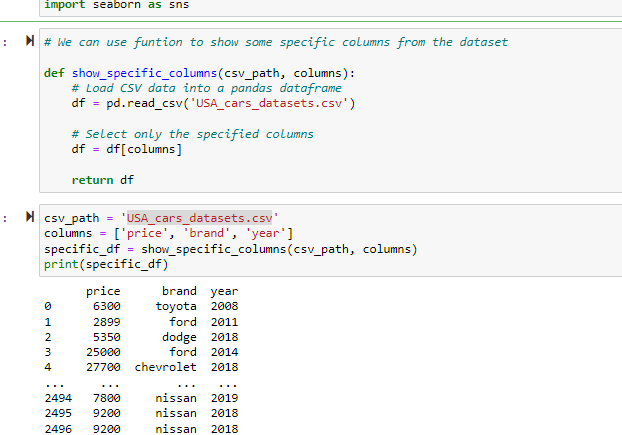
# **Libraries used**

* **pandas**: pandas is a powerful data manipulation library that provides data structures for efficiently storing and manipulating large datasets. It is often used for data analysis, data cleaning, and data preprocessing tasks.
* **numpy**: numpy is a numerical computing library for Python. It provides support for large, multi-dimensional arrays and matrices, as well as a large collection of high-level mathematical functions to operate on these arrays.
* **matplotlib**: matplotlib is a data visualization library for creating static, animated, and interactive visualizations in Python. It provides a wide variety of plotting functions for visualizing data in various formats, including line plots, scatter plots, bar plots, histograms, and more.
* **sklearn**.**preprocessing**: This is a module from the scikit-learn library that provides a collection of functions for preprocessing data. It includes functions for scaling, normalization, encoding, and imputation of missing values.
* **seaborn**: seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. It includes a wide range of plotting functions, including heatmaps, scatterplots, line plots, and histograms, among others.



# **Functions**

# **Show specific Coulmns**



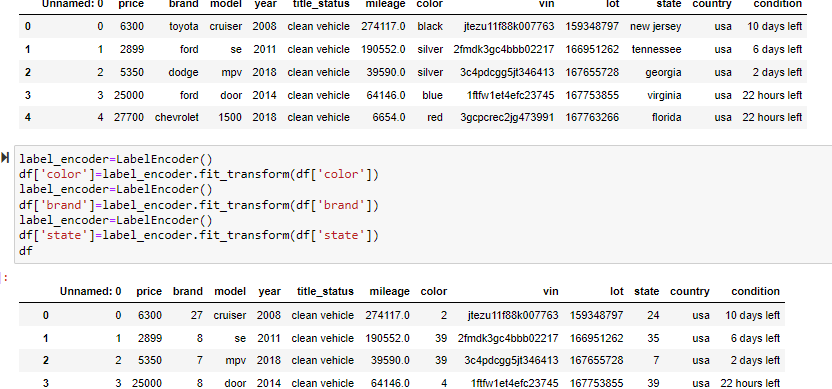
# **get\_column:**

# This function can be used to extract a single column from a CSV file and return it as a list



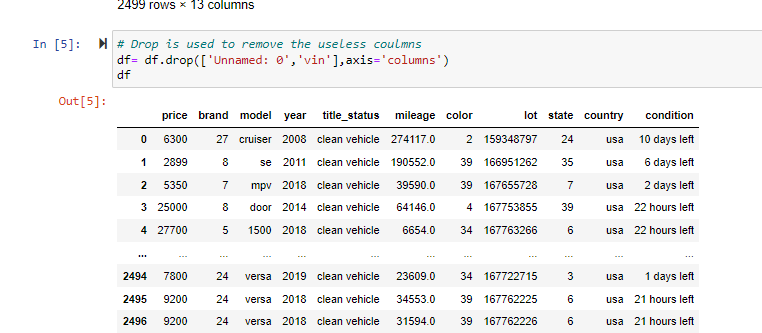
# **Use Label Encoder()**

Label encoder is a function in most programming languages and machine learning libraries. It is often provided as part of a data preprocessing toolkit or library that includes other functions for data cleaning, feature engineering, and data transformation



# **Drop Funtion()**

In machine learning, the .drop() function is often used to remove a specific column or row from a dataset. This function is commonly used in data preprocessing when we want to remove unnecessary or irrelevant data from our dataset.

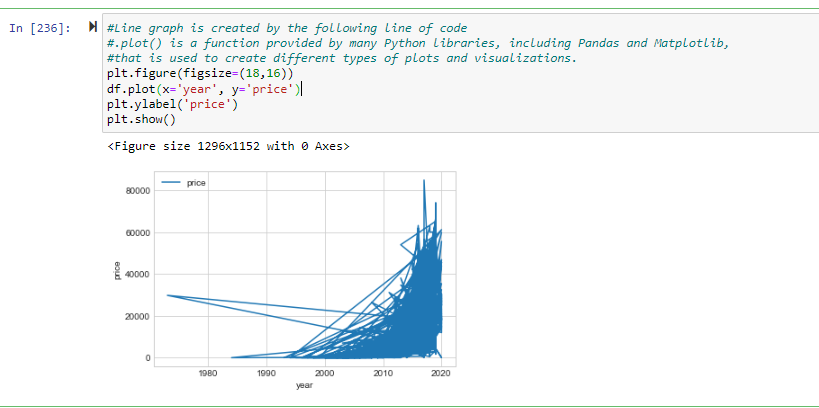


# **Some of the Visualization Method:**

* **Line plot**

A line plot, also known as a line chart or a time series plot, is a type of visualization that displays data points connected by lines to show trends or changes over time. Line plots are useful for analyzing data that changes continuously, such as stock prices, weather patterns, or sales figures over time.

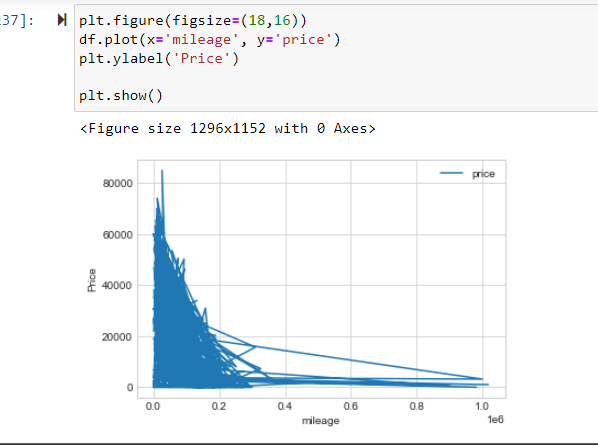
Line plots are particularly effective in showing trends over time because they allow you to easily see the changes in data points over a period of time

****

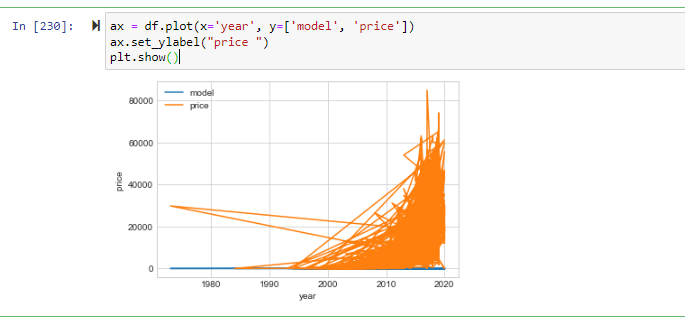
In the above figure, we have created a line plot between the column price and year.in which we see what happened in prices with time so in the beginning prices were too low but as time cross 2010 prices get higher and higher.

In the below figure, we are checking that mileage also matters with price. The mileage of a car is one of the factors that can affect its price. Generally, a car with higher mileage will have a lower value than a car with lower mileage, all other factors being equal. This is because a car with higher mileage has been driven more, which can result in more wear and tear on the engine, suspension, brakes, and other parts.

A car's mileage can also affect its resale value. As a car gets older and its mileage increases, its value will generally decrease. This is because potential buyers may be concerned about the car's reliability and how much longer it will last before needing major repairs or replacement parts.

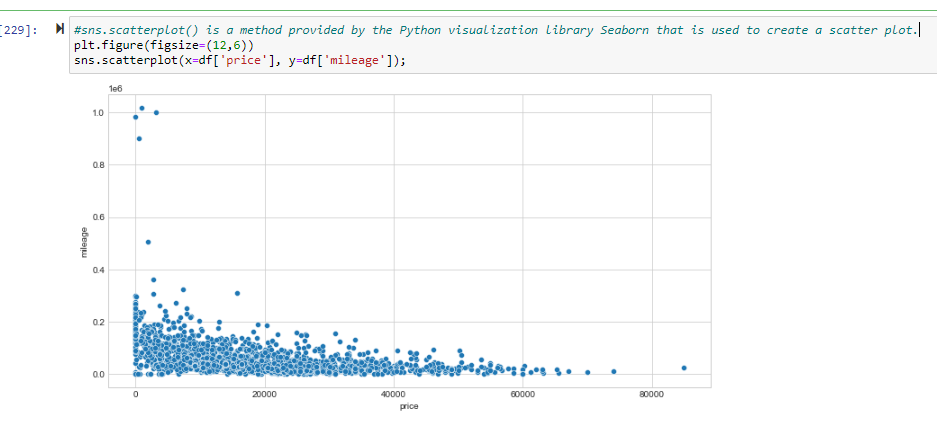


**Multiple line plot:**

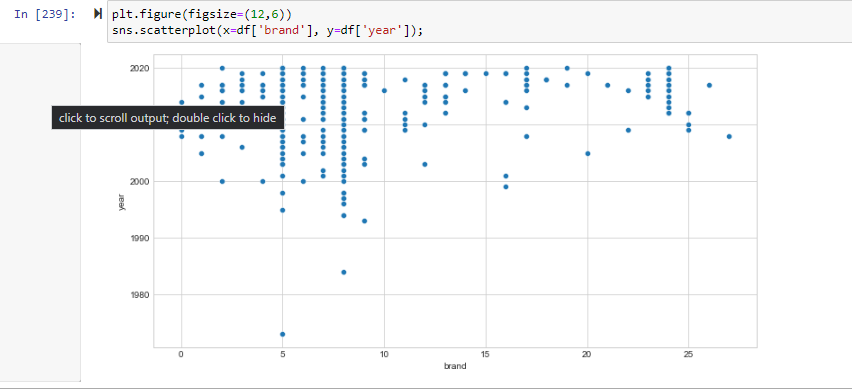
****

* **Scatter Plot:**

Scatter plots are a type of visualization commonly used to display the relationship between two variables. They are created by plotting one variable on the x-axis and another variable on the y-axis, with each data point represented by a dot or symbol. scatter plots are a versatile and valuable tool for data analysis and visualization. They can help to identify relationships, patterns, and outliers, and communicate important information about the data being analyzed.



In this figure we create a scatter plot with mileage on the y-axis and price on the x-axis. This would allow you to see whether there is a relationship between mileage and price, and whether cars with higher mileage tend to have lower prices. You could also color-code the data points by car make or model, which could help you identify any differences in the relationship between mileage and price for different types of cars.

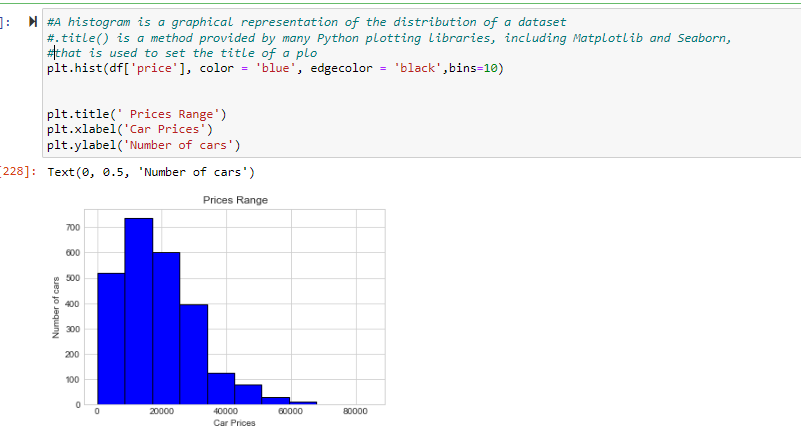


* **Histogram**

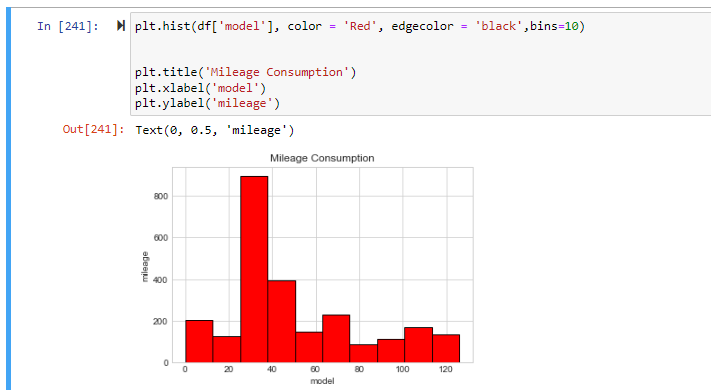
A histogram is a type of chart that is used to represent the distribution of a set of continuous data. It is a graph that consists of a set of vertical bars, where each bar represents a range of values, and the height of the bar represents the frequency or count of the number of observations that fall within that range.

Histograms are useful in visualization because they can help to quickly identify patterns in data, such as the presence of outliers, the overall shape of the distribution, and the location and spread of the data. Histograms can also be used to compare the distribution of two or more datasets by overlaying them plotting on the same chart or by them side-by-side

In this figure, plotting the car prices on the x-axis and the number of cars on the y-axis, a histogram can show the frequency of car prices and how they are distributed. This can help identify patterns or clusters of car prices, which can be useful in making business decisions such as setting prices or identifying customer segments.



In the below figure, histogram is used to visualize the distribution of car mileage for each car model. This can help identify any patterns or differences in mileage between different models.



.