**[Car\_Prediction](https://github.com/EsraaEzzathamam/Car_Prediction)**

**Introduction:**  
In this report, we will discuss how to make a code in Jupyter Notebook to preprocess a cars exhibition dataset in Kuwait. Preprocessing is an important step in data analysis as it helps to clean and prepare the data for further analysis. In this report, we will focus on the preprocessing steps of cleaning and removing duplicates.

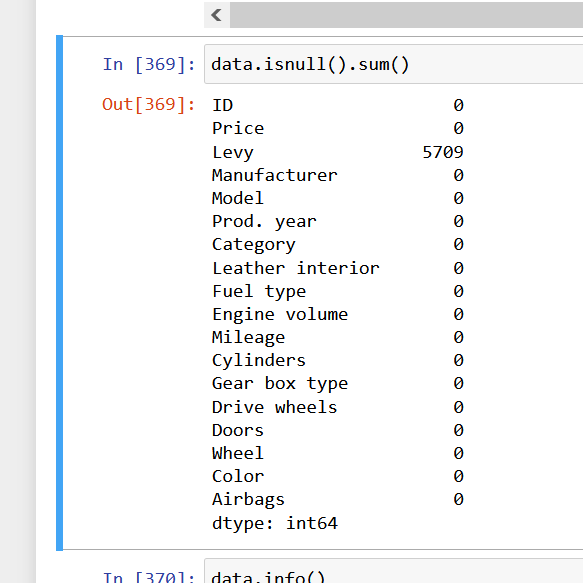
**Dataset:**  
The dataset used in this report is a cars exhibition dataset in Kuwait. The dataset contains information about various cars that were exhibited in the Kuwaiti car exhibition, including the type of car, its make, model, price, and other relevant information.

**Preprocessing:**

Preprocessing is an essential step in data analysis that involves cleaning and preparing the data for further analysis. In this report, we will discuss some common preprocessing techniques that we used in this project.

1. **Handling Missing Values:**  
   Missing values are a common issue in any dataset and can be caused by a variety of reasons such as data entry errors, sensor failures, or data loss during transmission.

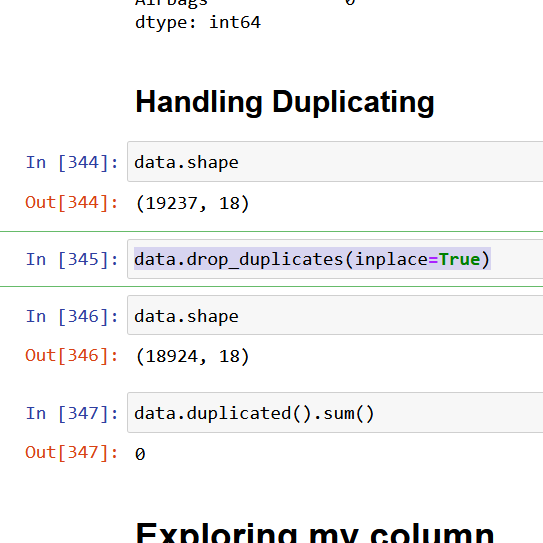
-example



Handling the null : Levy.fillna(Levy\_mean , inplace=True)

2-**Removing Duplicates:**  
 Duplicate data can cause problems during analysis and can lead to incorrect conclusions. Removing duplicates is a simple preprocessing technique that can help clean the data.

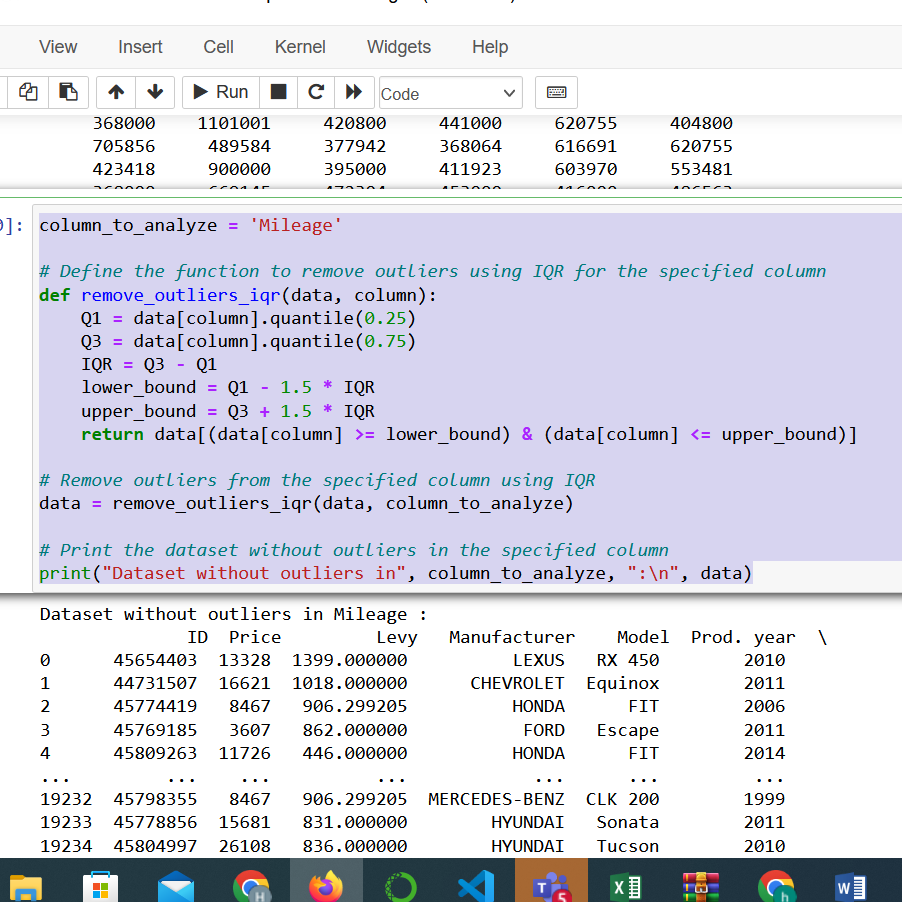
-example



3-**handling the outliers:**

Outliers are extreme values in a dataset that deviate significantly from other values and can have a disproportionate impact on the analysis results. Handling outliers is an important step in data analysis, as they can distort the distribution of the data, affect the accuracy of statistical models, and lead to incorrect conclusions. In this response.

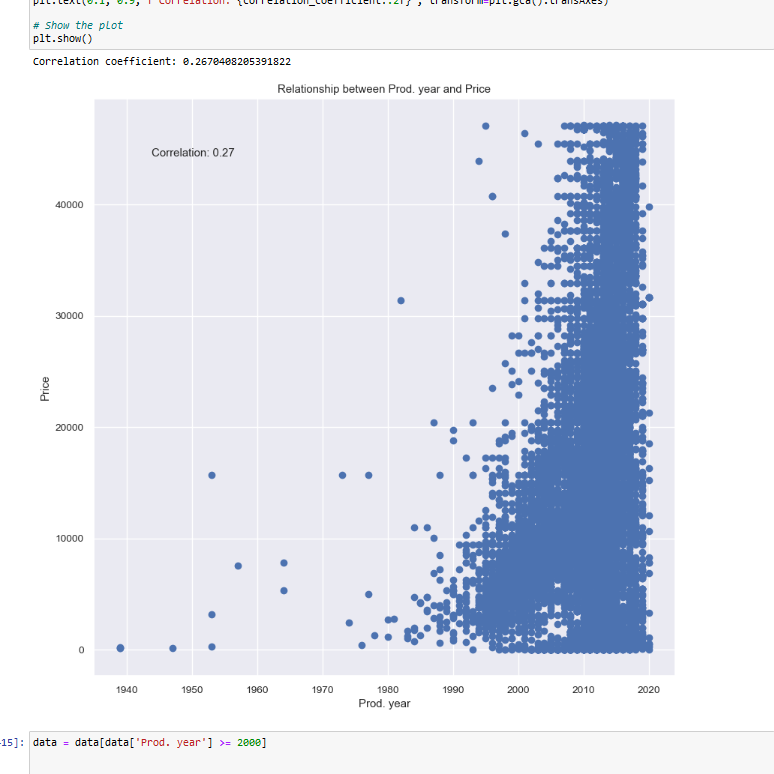
-Statistical methods can be used to identify outliers in a dataset. One of the most common methods is the z-score method, which calculates the number of standard deviations an observation is from the mean of the dataset. Observations with z-scores greater than a certain threshold, typically 3 are considered outliers.

Example: 

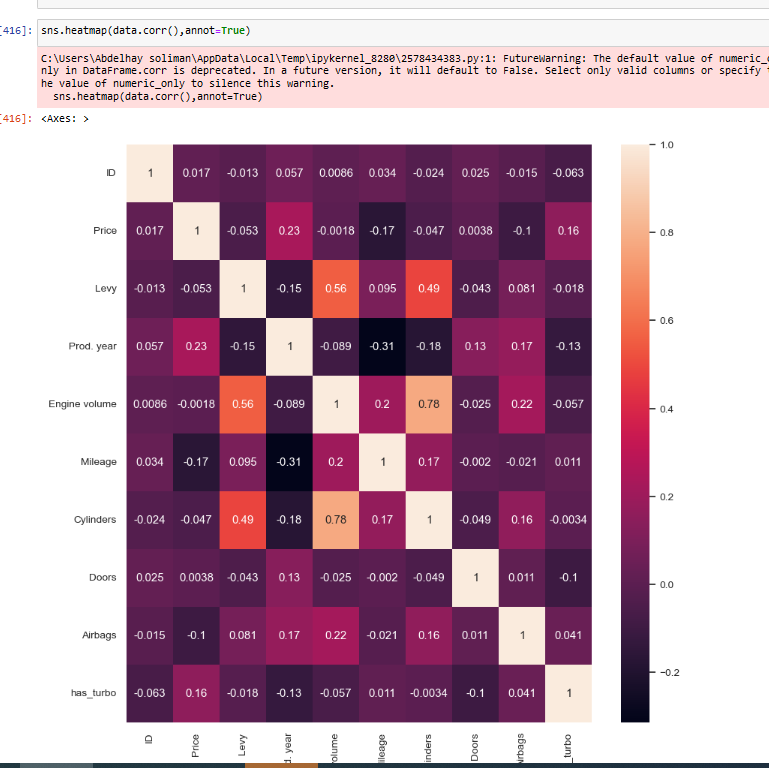
**Data visualization :**

is the process of representing data in a graphical or visual format. It is an essential component of data analysis as it allows us to communicate complex data insights in a clear and concise manner. In this response, I will discuss some common types of data visualizations and how they can be used to communicate insights from a dataset.

**1- Scatter Plots:**Scatter plots are a popular visualization technique used to represent the relationship between two continuous variables in a dataset. They are useful for identifying patterns or correlations between different variables in a dataset. Scatter plots can be simple or complex, depending on the number of variables and the level of detail required.



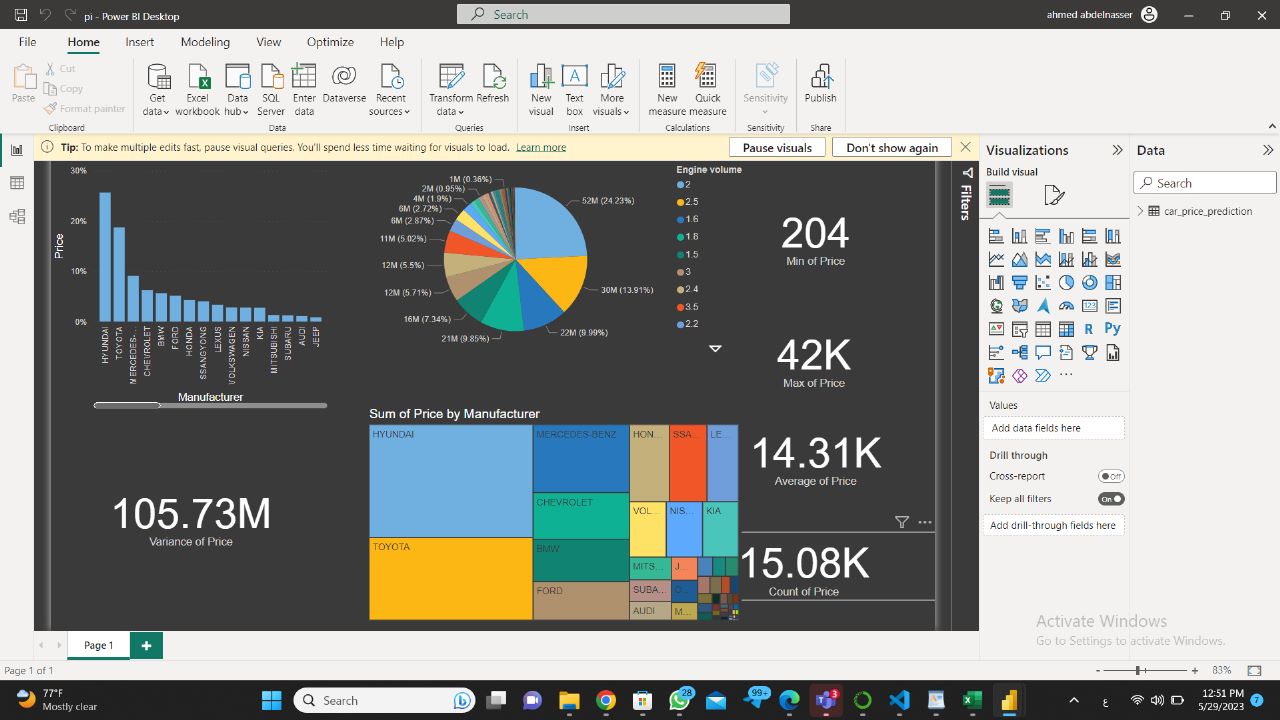
2-Heat Maps:  
Heat maps are a visualization technique used to represent large amounts of data in a condensed and visual format. They are useful for identifying patterns or correlations between different variables in a dataset. Heat maps can be simple or complex, depending on the number of variablesand the level of detail required.



3-Box Plots:  
Box plots are a visualization technique used to represent the distribution of a dataset. They are useful for identifying outliers, the range of values, and the median of a dataset. Box plots can be simple or complex, depending on the number of variables and the level of detail required.

Power PI:

Power BI is a business analytics service that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their reports and dashboards. It allows users to connect to different data sources, transform data, and create interactive visualizations, reports, and dashboards.



**A screenshot of a computer program

Description automatically generated with low confidence**