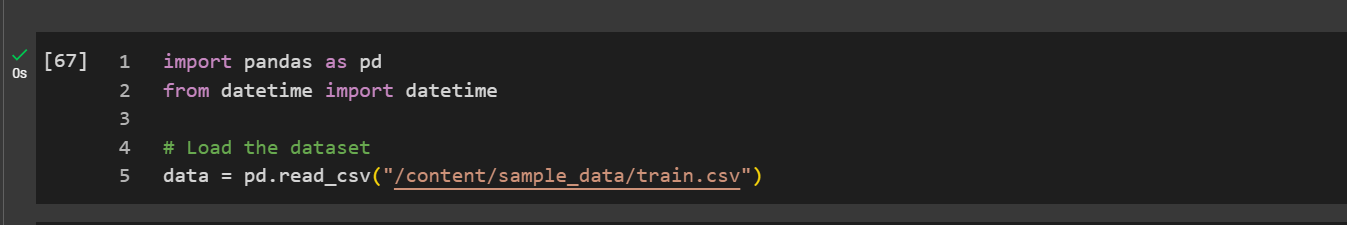
**Principles of Data Science (5530)-Assignment 2**

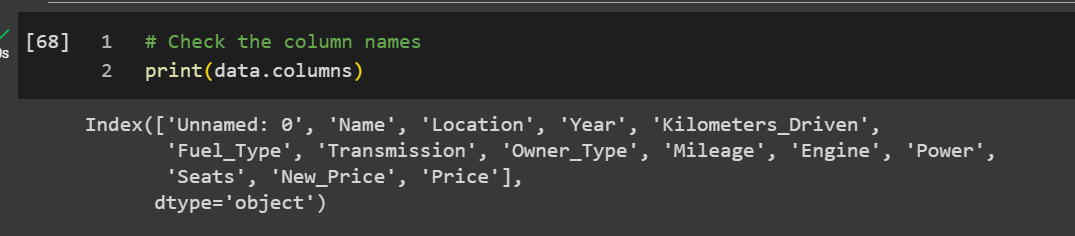
**Name: Akhila Reddyrajula**

**ID: 16346391**

**Used Car Dataset Analysis**

The provided dataset contains information about used cars, including attributes such as make and model, location, year of manufacture, mileage, fuel type, transmission type, and price. The objective of this analysis is to preprocess the data and perform various operations to gain insights into the dataset.

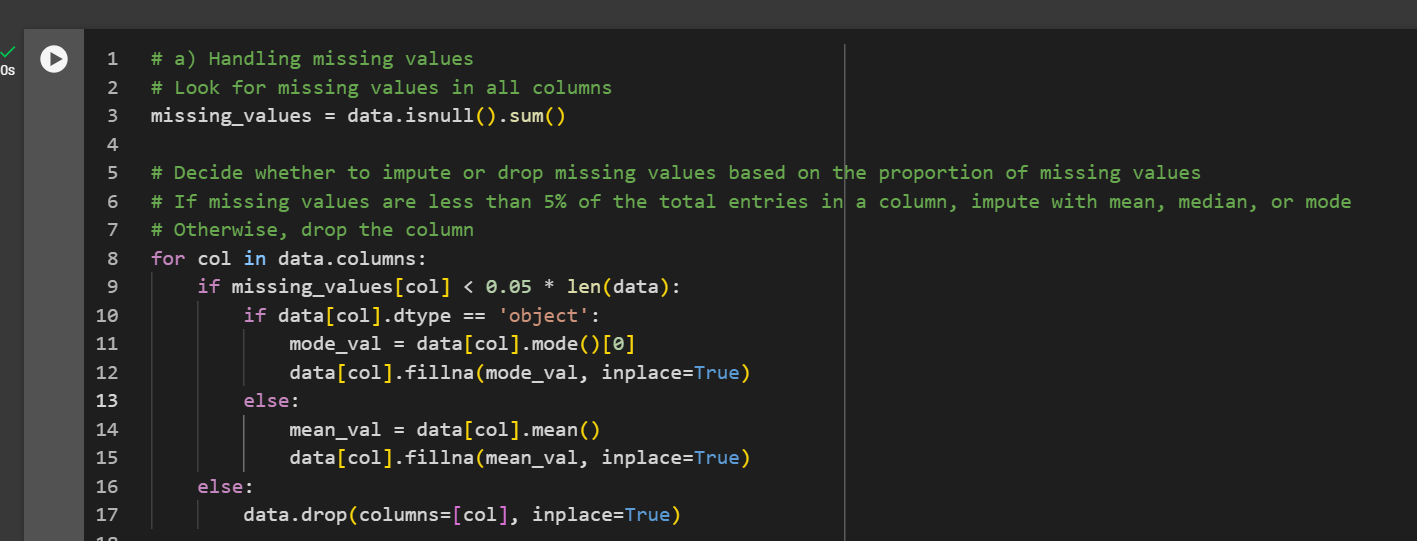




**Preprocessing Steps**

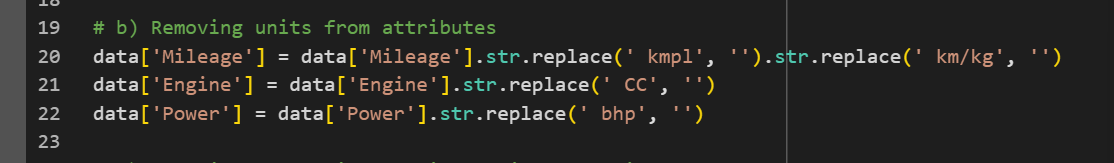
**a) Handling Missing Values**

Missing values are identified in all columns, and a decision is made whether to impute or drop them. For columns where missing values constitute less than 5% of the total entries, they are imputed using mean, median, or mode, depending on the data type. Otherwise, the entire column is dropped.



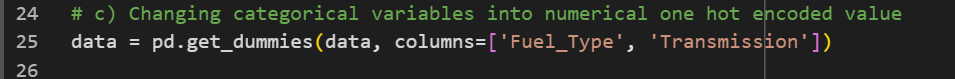
**b) Removing Units from Attributes**

Units are removed from certain attributes to convert them to numerical values. For instance, units such as 'kmpl' from 'Mileage', 'CC' from 'Engine', and 'bhp' from 'Power' are removed, leaving only the numerical values.



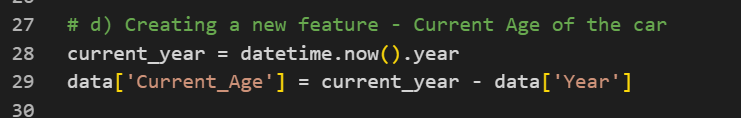
**c) Changing Categorical Variables to One-Hot Encoded Values**

Categorical variables like 'Fuel\_Type' and 'Transmission' are converted into numerical one-hot encoded values using the pandas get\_dummies() function. This process allows for the inclusion of categorical data in machine learning models.



**d) Creating a New Feature - Current Age of the Car**

A new feature named 'Current\_Age' is created by subtracting the 'Year' of manufacture from the current year. This feature provides information about the age of the car at the time of analysis.



**e) Performing Select, Filter, Rename, Mutate, Arrange, and Summarize Operations**

Several operations are performed on the dataset:

**- Select:** Specific columns are selected for analysis, such as 'Name', 'Location', 'Year', 'Mileage', etc.

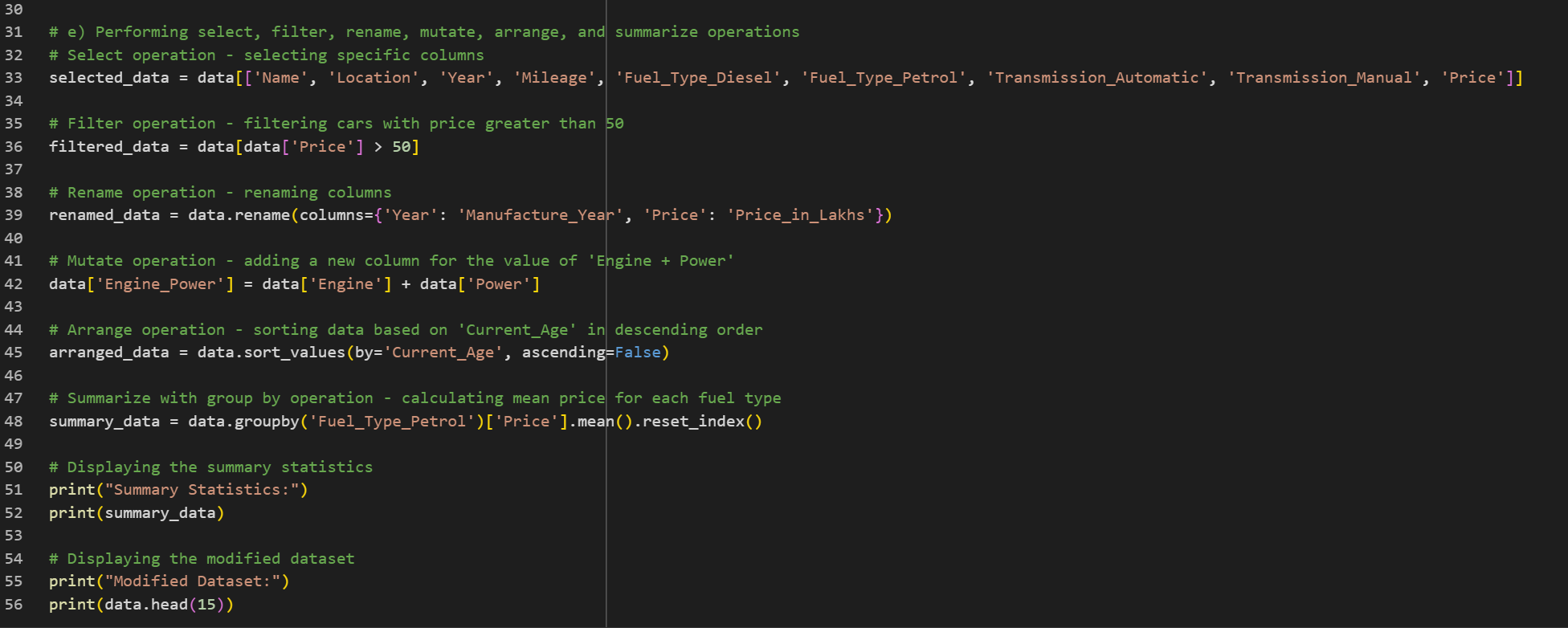
**- Filter:** Cars with a price greater than 50 are filtered out for further analysis.

**- Rename:** Columns are renamed to provide more descriptive names, such as renaming 'Year' to 'Manufacture\_Year' and 'Price' to 'Price\_in\_Lakhs'.

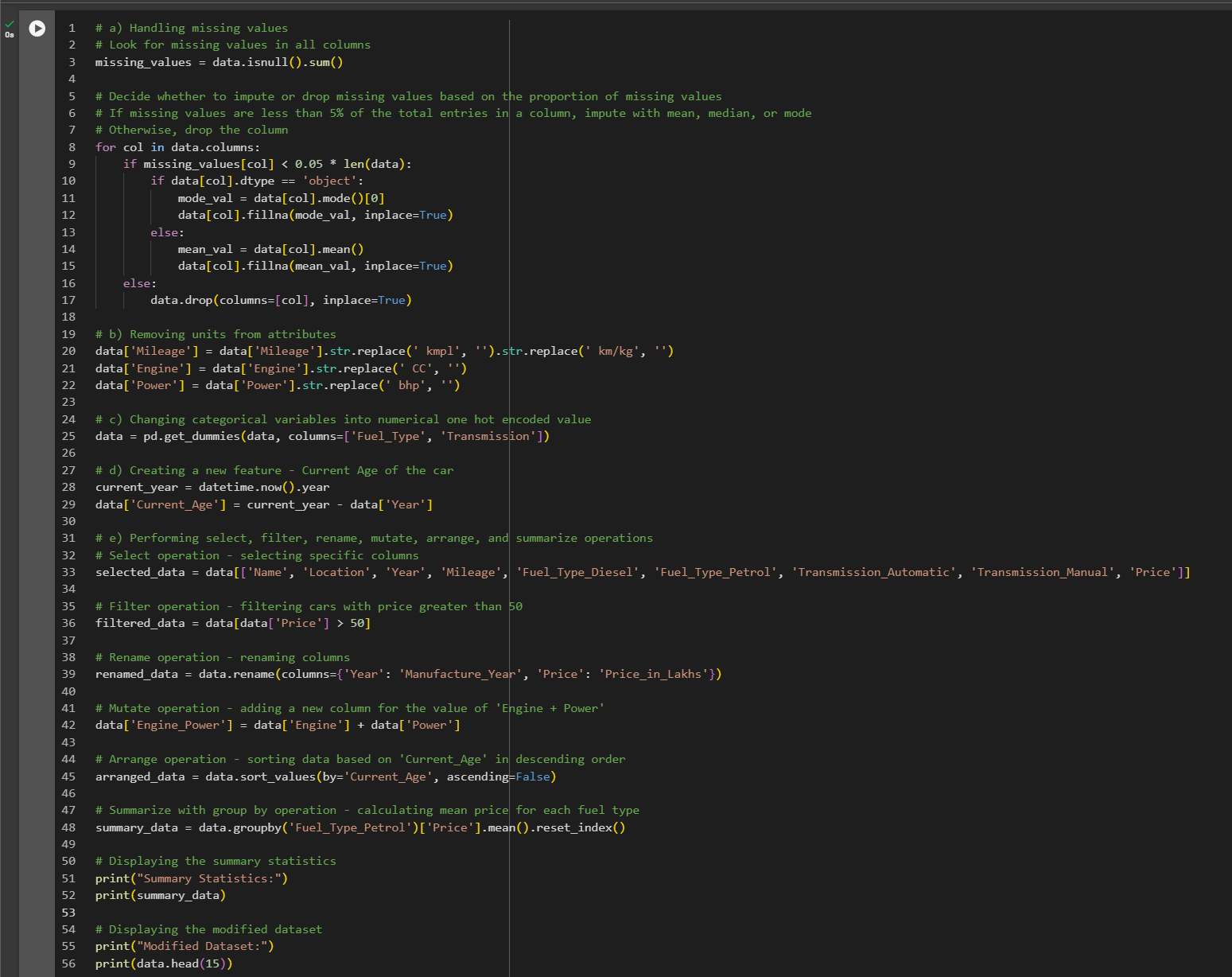
**- Mutate:** A new column named 'Engine\_Power' is added by combining the values of 'Engine' and 'Power'.

**- Arrange:** Data is sorted based on 'Current\_Age' in descending order to understand the distribution of car ages.

**- Summarize:** Summary statistics are calculated, such as the mean price for each fuel type ('Fuel\_Type\_Petrol').



This analysis provides valuable insights into the used car dataset, including preprocessing steps to handle missing values, converting categorical variables, and creating new features.



**Output:**

