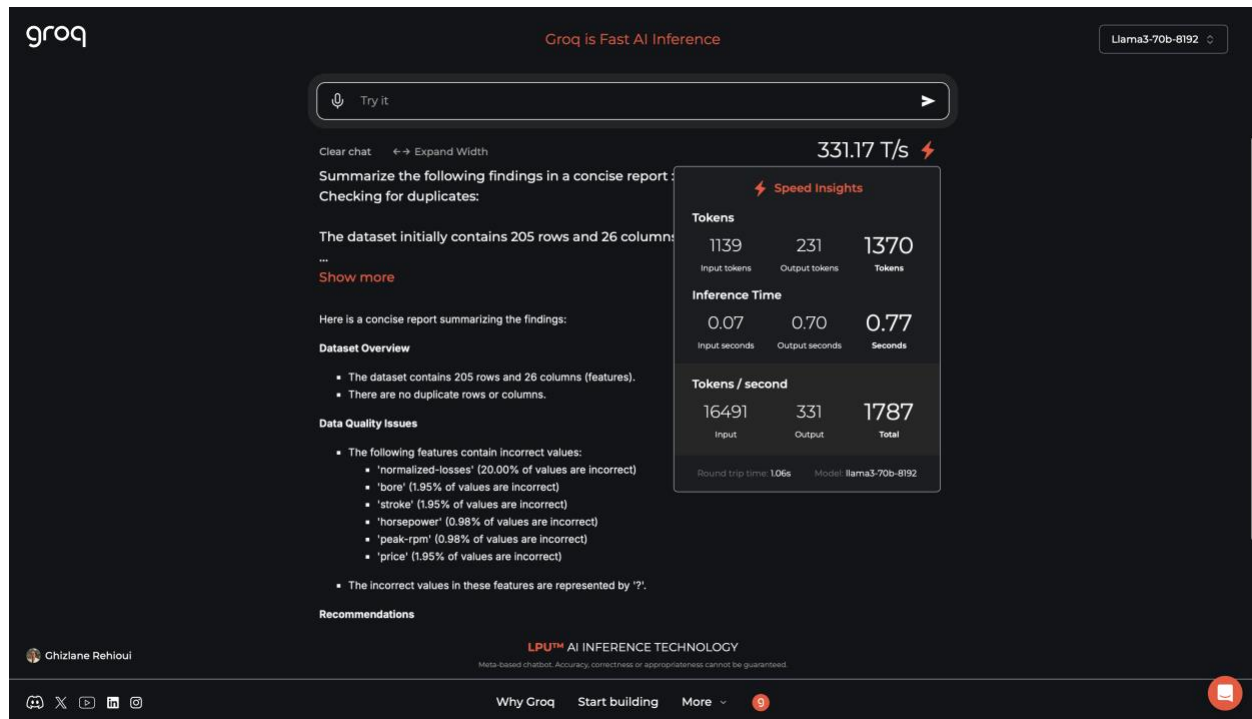Output

Here is a concise report summarizing the findings:

Dataset Overview

- The dataset contains 205 rows and 26 columns (features).
- There are no duplicate rows or columns.

Data Quality Issues

- The following features contain incorrect values:
    - 'normalized-losses' (20.00% of values are incorrect)
    - 'bore' (1.95% of values are incorrect)
    - 'stroke' (1.95% of values are incorrect)
    - 'horsepower' (0.98% of values are incorrect)
    - 'peak-rpm' (0.98% of values are incorrect)
    - 'price' (1.95% of values are incorrect)
- The incorrect values in these features are represented by '?'.

Recommendations

- Clean and preprocess the data by replacing the incorrect values in the above-mentioned features.
- Consider converting the data types of 'normalized-losses', 'bore', 'stroke', 'horsepower', 'peak-rpm', and 'price' to numeric types (e.g., float64) to ensure consistency and accuracy in analysis.

Model: Llama 3-70b-8192

Prompt:
Summarize the following findings in a concise report :

Checking for duplicates:

The dataset initially contains 205 rows and 26 columns (features).

    - The number of duplicate rows in this dataset is 0.
    - The number of duplicate columns in this dataset is 0.

--> There are no duplicates to drop!
There are 26 colummns in the dataset.

Looking at each feature:

- The feature: 'symboling'
- Datatype: int64
- Potential Data Inconsistencies:

- The feature: 'normalized-losses'
- Datatype: object
- Potential Data Inconsistencies:
      - This feature contains numbers, either an integer or a float. Since the code is unable to catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

- The feature: 'make'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'fuel-type'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'aspiration'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'num-of-doors'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'body-style'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'drive-wheels'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'engine-location'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'wheel-base'
- Datatype: float64
- Potential Data Inconsistencies:

- The feature: 'length'
- Datatype: float64
- Potential Data Inconsistencies:

- The feature: 'width'
- Datatype: float64
- Potential Data Inconsistencies:

- The feature: 'height'
- Datatype: float64
- Potential Data Inconsistencies:

- The feature: 'curb-weight'
- Datatype: int64

- Potential Data Inconsistencies:

- The feature: 'engine-type'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'num-of-cylinders'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'engine-size'
- Datatype: int64
- Potential Data Inconsistencies:

- The feature: 'fuel-system'
- Datatype: object
- Potential Data Inconsistencies:

- The feature: 'bore'
- Datatype: object
- Potential Data Inconsistencies:
        - This feature contains numbers, either an integer or a float. Since the code is unable to
catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

- The feature: 'stroke'
- Datatype: object
- Potential Data Inconsistencies:
        - This feature contains numbers, either an integer or a float. Since the code is unable to
catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

- The feature: 'compression-ratio'
- Datatype: float64
- Potential Data Inconsistencies:

- The feature: 'horsepower'
- Datatype: object
- Potential Data Inconsistencies:
        - This feature contains numbers, either an integer or a float. Since the code is unable to catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

- The feature: 'peak-rpm'
- Datatype: object
- Potential Data Inconsistencies:
        - This feature contains numbers, either an integer or a float. Since the code is unable to catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

- The feature: 'city-mpg'
- Datatype: int64
- Potential Data Inconsistencies:

- The feature: 'highway-mpg'
- Datatype: int64
- Potential Data Inconsistencies:

- The feature: 'price'
- Datatype: object
- Potential Data Inconsistencies:
        - This feature contains numbers, either an integer or a float. Since the code is unable to catch that the data type is neither an integer nor a float, there is at least one incorrect value:
        - The incorrect values are: ['?']

The columns/features and values to fix in the entire dataset are: {'normalized-losses': ['?'], 'bore': ['?'], 'stroke': ['?'], 'horsepower': ['?'], 'peak-rpm': ['?'], 'price': ['?']}

Percentage of incorrect values for normalized-losses: 20.00%
Percentage of incorrect values for bore: 1.95%
Percentage of incorrect values for stroke: 1.95%
Percentage of incorrect values for horsepower: 0.98%
Percentage of incorrect values for peak-rpm: 0.98%
Percentage of incorrect values for price: 1.95%