Final Report

We have given the raw data and we can able to find many missing values which may hinder our further analysis.

To identify the missing data values the important steps to follow

1. Identify missing data
2. deal with missing data
3. correct data format

**Identify & Handle Missing Values**

Convert "?" to NaN

In the car dataset, missing data comes with the question mark "?". We replace "?" with NaN (Not a Number), which is Python's default missing value marker

**Evaluating for Missing Data**

The missing values are converted to Python's default. We use Python's built-in functions to identify these missing values. There are two methods to detect missing data:

1. **.**isnull()
2. .notnull()

The output is a boolean value indicating whether the value that is passed into the argument is in fact missing data.

**Count missing values in each column**

Using a for loop in Python, we can quickly figure out the number of missing values in each column. In the body of the for loop the method ".value\_counts()" counts the number of "True" values.

**to deal with missing data**

1. drop data  
   a. drop the whole row  
   b. drop the whole column
2. replace data  
   a. replace it by mean  
   b. replace it by frequency  
   c. replace it based on other functions

Whole columns should be dropped only if most entries in the column are empty. In our dataset, none of the columns are empty enough to drop entirely. We have some freedom in choosing which method to replace data; however, some methods may seem more reasonable than others. We will apply each method to many different columns:

**Replace by mean:**

* "normalized-losses": 41 missing data, replace them with mean
* "stroke": 4 missing data, replace them with mean
* "bore": 4 missing data, replace them with mean
* "horsepower": 2 missing data, replace them with mean
* "peak-rpm": 2 missing data, replace them with mean

**Replace by frequency:**

* "num-of-doors": 2 missing data, replace them with "four".
  + Reason: 84% sedans is four doors. Since four doors is most frequent, it is most likely to occur

**Drop the whole row:**

* "price": 4 missing data, simply delete the whole row
  + Reason: price is what we want to predict. Any data entry without price data cannot be used for prediction

**Correct data format**

The last step in data cleaning is checking and making sure that all data is in the correct format (int, float, text or other).

In Pandas, we use

**.dtype()** to check the data type

**.astype()** to change the data type