Waheed Anwar 4375.004 Portfolio Assignment: ML with sklearn

Read the Auto data

```
import pandas as pd
import numpy as np
import sklearn as sklearn

df = pd.read_csv('Auto.csv')

# output first few rows
print(df.head())
```

```
    mpg  cylinders  displacement  horsepower  weight  acceleration  year  \
0  18.0          8         307.0         130    3504          12.0  70.0
1  15.0          8         350.0         165    3693          11.5  70.0
2  18.0          8         318.0         150    3436          11.0  70.0
3  16.0          8         304.0         150    3433          12.0  70.0
4  17.0          8         302.0         140    3449           NaN  70.0

   origin                       name
0       1  chevrolet chevelle malibu
1       1          buick skylark 320
2       1         plymouth satellite
3       1              amc rebel sst
4       1                ford torino
```

```
# output dimensions of data
rows, cols = df.shape
print(f"Number of rows: {rows}")
print(f"Number of rows: {cols}")
```

```
Number of rows: 392
Number of rows: 9
```

Data exploration with code

```
# use describe() on the mpg, weight, and year columns
col_description = df[['mpg','weight','year']].describe()
```

```
print(col_description)
```

```
                mpg       weight        year
count  392.000000   392.000000  390.000000
mean    23.445918  2977.584184   76.010256
std      7.805007   849.402560    3.668093
min      9.000000  1613.000000   70.000000
25%     17.000000  2225.250000   73.000000
50%     22.750000  2803.500000   76.000000
75%     29.000000  3614.750000   79.000000
max     46.600000  5140.000000   82.000000
```

```
# Range of the columns:
#     mpg min = 9, mpg max = 46.6
#     weight min = 1613, weight max = 5140
#     year min = 70, year max = 82

# Average of the columns (rounded to two decimal places):
#     mpg avg = 23.45
#     weight avg = 2977.58
#     year avg = 76.01
```

Explore data types

```
# check the data types of all columns
print(df.dtypes)
```

```
mpg             float64
cylinders         int64
displacement    float64
horsepower        int64
weight            int64
acceleration    float64
year            float64
origin            int64
name             object
dtype: object
```

```
# change the cylinders column to categorical using cat.codes
```

```
# change the origin column to categorical without using cat.codes
df['origin'] = df['origin'].astype('category')
print(df.dtypes)
```

```
mpg              float64
cylinders          int64
displacement     float64
horsepower         int64
weight             int64
acceleration     float64
year             float64
origin          category
name              object
dtype: object
```

Dropping the NA

```
df.dropna()
```

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name

```
from sklearn.model_selection import train_test_split
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-76-bbe54bb64cc4> in <cell line: 1>()
----> 1 X = df.data
      2 y = df.target
      3 from sklearn.model_selection import train_test_split
      4 sklearn.model_selection.train_test_split(test_size=.80, train_size=.20)[df]

/usr/local/lib/python3.9/dist-packages/pandas/core/generic.py in __getattr__(self, name)
   5900         ):
   5901             return self[name]
-> 5902         return object.__getattribute__(self, name)
   5903
   5904     def __setattr__(self, name: str, value) -> None:

AttributeError: 'DataFrame' object has no attribute 'data'
```

SEARCH STACK OVERFLOW

0s completed at 1:26 PM