Practical Lab Assignment 2

Pre-processing of missing values

Problem Statement: Replace the missing values for given automobile dataset “imports-85.data” with the mean, median and mode value of numeric attributes.

Dataset:
https://github.com/nyuvis/datasets/blob/master/auto/imports-85.data

Dataset Information :
https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

Replacing the missing value with Mean

Implementation:

```
#importing the libraries
import numpy as np
import pandas as pd

#importing the dataset
df = pd.read_csv("https://raw.githubusercontent.com/nyuvis/datasets/
master/auto/imports-85.data")

#replacing missing values with NaN
df.replace("?",np.nan,inplace=True)

#now convert the object type to float
df['normalized-losses'] = df['normalized-losses'].astype(float)
df['bore'] = df['bore'].astype(float)
df['stroke'] = df['stroke'].astype(float)
df['horsepower'] = df['horsepower'].astype(float)
df['peak-rpm'] = df['peak-rpm'].astype(float)
df['price'] = df['price'].astype(float)
#now after converting the object in float, calculate the mean
col_mean = df.mean()
```

```
symboling              0.834146
normalized-losses    122.000000
wheel-base            98.756585
length               174.049268
width                 65.907805
height                53.724878
curb-weight         2555.565854
engine-size          126.907317
bore                   3.329751
stroke                 3.255423
compression-ratio     10.142537
horsepower           104.256158
peak-rpm            5125.369458
city-mpg              25.219512
highway-mpg           30.751220
price              13207.129353
dtype: float64
```

#Replace the NaN with mean

df = df.fillna(df.mean())

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width
0	3	122.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1
1	3	122.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1
2	1	122.0	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5
3	2	164.0	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2
4	2	164.0	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4
...	...	...	...	...	...	...	...	...	...	...	...	...
200	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	188.8	68.9
201	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	188.8	68.8
202	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	188.8	68.9
203	-1	95.0	volvo	diesel	turbo	four	sedan	rwd	front	109.1	188.8	68.9
204	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	188.8	68.9

205 rows × 26 columns

Replacing the missing value with Median

Implementation:

```
#import libraries
import numpy as np
import pandas as pd

#import dataset
df=pd.read_csv("https://raw.githubusercontent.com/nyuvis/datasets/
master/auto/imports-85.data")

#replacing the missing values with Nan
df.replace("?",np.nan, inplace= True)

#convert the object type to float
df['normalized-losses'] = df['normalized-losses'].astype(float)
df['bore'] = df['bore'].astype(float)
df['stroke'] = df['stroke'].astype(float)
df['horsepower'] = df['horsepower'].astype(float)
df['peak-rpm'] = df['peak-rpm'].astype(float)
df['price'] = df['price'].astype(float)

#sort the values
df['normalized-losses'] = df['normalized-losses'].sort_values()
df['bore'] = df['bore'].sort_values()
df['stroke'] = df['stroke'].sort_values()
df['horsepower'] = df['horsepower'].sort_values()
df['peak-rpm'] = df['peak-rpm'].sort_values()
df['price'] = df['price'].sort_values()
```

```
#find the median
med = df.median()
```

```
symboling              1.00
normalized-losses    115.00
wheel-base            97.00
length               173.20
width                 65.50
height                54.10
curb-weight         2414.00
engine-size          120.00
bore                   3.31
stroke                 3.29
compression-ratio      9.00
horsepower            95.00
peak-rpm            5200.00
city-mpg              24.00
highway-mpg           30.00
price              10198.00
dtype: float64
```

```
#replacing nan with medians
df.replace(np.nan,df['normalized-losses'].median(), inplace= True)
df.replace(np.nan,df['bore'].median(), inplace= True)
df.replace(np.nan,df['stroke'].median(), inplace= True)
df.replace(np.nan,df['horsepower'].median(), inplace= True)
df.replace(np.nan,df['peak-rpm'].median(), inplace= True)
df.replace(np.nan,df['price'].median(), inplace= True)
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height
0	3	115.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8
1	3	115.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8
2	1	115.0	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4
3	2	164.0	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3
4	2	164.0	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...
200	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	188.8	68.9	55.5
201	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	188.8	68.8	55.5
202	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	188.8	68.9	55.5
203	-1	95.0	volvo	diesel	turbo	four	sedan	rwd	front	109.1	188.8	68.9	55.5
204	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	188.8	68.9	55.5

205 rows × 26 columns

Replacing the missing value with Mode

Implementation:

```
#import libraries
import numpy as np
import pandas as pd

#import dataset
df=pd.read_csv("https://raw.githubusercontent.com/nyuvis/datasets/master/auto/imports-85.data")

#replacing the missing values with Nan
df.replace("?",np.nan, inplace= True)

#convert the object type to float
df['normalized-losses'] = df['normalized-losses'].astype(float)
df['bore'] = df['bore'].astype(float)
df['stroke'] = df['stroke'].astype(float)
df['horsepower'] = df['horsepower'].astype(float)
df['peak-rpm'] = df['peak-rpm'].astype(float)
df['price'] = df['price'].astype(float)

#mode
df['normalized-losses'] = df['normalized-losses'].fillna(df['normalized-losses'].mode()[0])
df['bore'] = df['bore'].fillna(df['bore'].mode()[0])
df['stroke'] = df['stroke'].fillna(df['stroke'].mode()[0])
df['horsepower'] = df['horsepower'].fillna(df['horsepower'].mode()[0])
df['peak-rpm'] = df['peak-rpm'].fillna(df['peak-rpm'].mode()[0])
df['price'] = df['price'].fillna(df['price'].mode()[0])
```

```
print(df['normalized-losses'].mode())
print(df['bore'].mode())
print(df['stroke'].mode())
print(df['horsepower'].mode())
print(df['peak-rpm'].mode())
print(df['price'].mode())
```

```
0    161.0
dtype: float64
0    3.62
dtype: float64
0    3.4
dtype: float64
0    68.0
dtype: float64
0    5500.0
dtype: float64
0    5572.0
dtype: float64
```

```
#replacing the values with mode
df.replace(np.nan,df['normalized-losses'].mode(), inplace= True)
df.replace(np.nan,df['bore'].mode(), inplace= True)
df.replace(np.nan,df['stroke'].mode(), inplace= True)
df.replace(np.nan,df['horsepower'].mode(), inplace= True)
df.replace(np.nan,df['peak-rpm'].mode(), inplace= True)
df.replace(np.nan,df['price'].mode(), inplace= True)
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors
0	3	161.0	alfa-romero	gas	std	two
1	3	161.0	alfa-romero	gas	std	two
2	1	161.0	alfa-romero	gas	std	two
3	2	164.0	audi	gas	std	four
4	2	164.0	audi	gas	std	four
...	...	...	...	...	...	...
200	-1	95.0	volvo	gas	std	four
201	-1	95.0	volvo	gas	turbo	four
202	-1	95.0	volvo	gas	std	four
203	-1	95.0	volvo	diesel	turbo	four
204	-1	95.0	volvo	gas	turbo	four

205 rows × 26 columns