WEEK 4

Title: Auto-Mpg Data

Sources: (a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. (c) Date: July 7, 1993

Attribute Information:

mpg: continuous cylinders: multi-valued discrete displacement: continuous horsepower: continuous weight: continuous acceleration: continuous model year: multi-valued discrete origin: multi-valued discrete car name: string (unique for each instance)

Removing outliers/Missing values

In [1]:
```
import pandas as pd
df = pd.read_csv('C:/Users/saile/Desktop/IoT/datasets/auto-mpg.csv')
```

In [2]:
```
df
```

Out[2]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

In [4]:
```
#replacing  question marks with  null values in horse power
for col in df.columns:
    df[col].replace('?',None, inplace=True)
df
```

Out[4]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

In [6]:
```
#identifying missing values
df.isnull().sum()
```

Out[6]:
```
mpg             0
cylinders       0
displacement    0
horsepower      0
weight          0
acceleration    0
model year      0
origin          0
car name        0
dtype: int64
```

Detect and remove outliers using percentiles-method_1

In [10]:
```
#identiy the outliers in dispalcement
max_threshold=df['cylinders'].quantile(0.95)
min_threshold=df['cylinders'].quantile(0.05)
df[(df['cylinders']>max_threshold) |(df['cylinders']<min_threshold)]
```

Out[10]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
71	19.0	3	70.0	97	2330	13.5	72	3	mazda rx2 coupe
111	18.0	3	70.0	90	2124	13.5	73	3	maxda rx3
243	21.5	3	80.0	110	2720	13.5	77	3	mazda rx-4
334	23.7	3	70.0	100	2420	12.5	80	3	mazda rx-7 gs

In [14]:
```
#dataframe after removing outliers

df[(df['cylinders']>max_threshold) |(df['cylinders']<min_threshold)]
df
```

Out[14]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

outlier detetcion using standard deviation method

In [18]:
```
#identifying outliers in mpg
df.acceleration.mean()
df.acceleration.std()
max1=df.acceleration.mean() + 3*df.acceleration.std()
min1=df.acceleration.mean() - 3*df.acceleration.std()
print("the max value is", max1)
print("the min value is", min1)
```

```
the max value is 23.841157241699317
the min value is 7.295023662823265
```

In [20]:
```
#removing outliers
df[(df['acceleration']>max1) |(df['acceleration']<min1)]
```

Out[20]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
299	27.2	4	141.0	71	3190	24.8	79	2	peugeot 504
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup

Outlier detection using z=method

In [27]:
```
df['zscore'] = df.acceleration- df.acceleration.mean()/df.acceleration.std()
df.head(5)
```

Out[27]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-score	
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.35466	6
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.85466	5
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.35466	5
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.35466	6
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.85466	4

In [31]:
```
df[(df.zscore<-3) |(df.zscore >3)]
```

Out[31]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.3546
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.8546
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.3546
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.3546
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.8546
...	...	...	...	...	...	...	...	...	...	.
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl	9.9546
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup	18.9546
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage	5.9546
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger	12.9546
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10	13.7546

395 rows × 11 columns

In [32]:
```
df[df['zscore'] >3]
```

Out[32]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.3546
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.8546
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.3546

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.3546
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.8546
...	...	...	...	...	...	...	...	...	...	.
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl	9.9546
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup	18.9546
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage	5.9546
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger	12.9546
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10	13.7546

395 rows × 11 columns

outlier detection using IQR

In [34]:
```
#identifying outliers
Q1= df.mpg.quantile(0.25)
Q3= df.mpg.quantile(0.75)
IQR= Q3-Q1
lower= Q1 -1.5*IQR
upper=Q3 +1.5*IQR
df[(df.mpg<lower)|(df.mpg>lower)]
```

Out[34]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.3546
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.8546
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.3546
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.3546
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.8546

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
...	...	...	...	...	...	...	...	...	...	.
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl	9.9546
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup	18.9546
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage	5.9546
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger	12.9546
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10	13.7546

398 rows × 11 columns

In [35]:
```
#removing outliers
df_no_outlier = df[(df.mpg<lower)|(df.mpg>lower)]
df_no_outlier
```

Out[35]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.3546
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.8546
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.3546
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.3546
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.8546
...	...	...	...	...	...	...	...	...	...	.
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl	9.9546
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup	18.9546
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage	5.9546
396	28.0	4	120.0	79	2625	18.6	82	1	ford	12.9546

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
									ranger	
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10	13.7546

398 rows × 11 columns

In [36]:
```
newdf=df.dropna()
newdf
```

Out[36]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-scor
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	6.3546
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	5.8546
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	5.3546
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	6.3546
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino	4.8546
...	...	...	...	...	...	...	...	...	...	.
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl	9.9546
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup	18.9546
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage	5.9546
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger	12.9546
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10	13.7546

398 rows × 11 columns

observations

Missing values are found in horsepower columns and they are remived from the dataframe using dropna() method Outliers are identified using various methods and in columns mpg,cylinder, and acceleration

Inputing standard values

In [ ]:
```
std =df[(df['cylinders'] <= max_threshold) | (df['cylinders']>=min_threshold)]
print('standard value',std)
```

In [41]:
```
df[(df['cylinders']>max_threshold) |(df['cylinders']<min_threshold)]
```

Out[41]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-score	
71	19.0	3	70.0	97	2330	13.5	72	3	mazda rx2 coupe	7.85466	7
111	18.0	3	70.0	90	2124	13.5	73	3	maxda rx3	7.85466	7
243	21.5	3	80.0	110	2720	13.5	77	3	mazda rx-4	7.85466	7
334	23.7	3	70.0	100	2420	12.5	80	3	mazda rx-7 gs	6.85466	6

observations

Outliers are imputed with standard mean value which were identified using threshold values.

capping of values

In [43]:
```
df['acceleration'].clip(min1,max1, inplace=True)
df[(df.acceleration >max1) |(df.acceleration<min1)]
```

Out[43]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name	z-score	zscore

observations

1.outliers in acceleration are capped with maximum and minimum values respectively using clip method in pandas. 2.Clip methods replaces the value with lower limit which is less than it and similarly works for values greater tahn upper limit

In [ ]: