In [1]:

```
# Import packages
# read the data

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

- In every dataset we have different columns has different units
- In every dataset we have different columns has values varies from -inf to inf
- It is very important standardize the data, make sure all the column values under same range
- To achieve this we have two methods
  - Normalization
  - standardization

Normalization:

- min max scalar

$$\overset{\text{new value}}{X'} = \frac{\overset{\text{original value}}{x}-\min(x)}{\max(x)-\min\,(x)}$$

Standedization:

- Z-score

$$Z = \frac{x-\mu}{\sigma}$$

In [ ]:

```
# step-1: calcaulate min value of p_Wage= min_wage
# step-2: calculate max value of p_wage = max_wage
# step-3: Dr= max_wage-min_wage
# step-4: Nr= p_wage-min_wage
# step-5: Nr/Dr
```

In [8]:
```
############################# Read the data ###############################

file_path="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df


min_wage=visa_df['prevailing_wage'].min()
max_wage=visa_df['prevailing_wage'].max()
dr=max_wage-min_wage
nr=visa_df['prevailing_wage']-min_wage
visa_df['prevailing_wage_norm']=nr/dr
```

In [9]:
```
visa_df[['prevailing_wage','prevailing_wage_norm']]
```

Out[9]:

	prevailing_wage	prevailing_wage_norm
0	592.2029	0.001849
1	83425.6500	0.261345
2	122996.8600	0.385312
3	83434.0300	0.261371
4	149907.3900	0.469616
...	...	...
25475	77092.5700	0.241505
25476	279174.7900	0.874579
25477	146298.8500	0.458311
25478	86154.7700	0.269895
25479	70876.9100	0.222033

25480 rows × 2 columns

In [10]:
```
visa_df['prevailing_wage_norm'].max(),visa_df['prevailing_wage_norm'].min()
```

Out[10]:
```
(1.0, 0.0)
```

In [11]:
```
visa_df['prevailing_wage'].max(),visa_df['prevailing_wage'].min()
```

Out[11]:
```
(319210.27, 2.1367)
```

In [12]:
```
max_id=visa_df['prevailing_wage_norm'].idxmax()
min_id=visa_df['prevailing_wage_norm'].idxmin()
max_id,min_id
```

Out[12]:
```
(21077, 20575)
```

In [13]:
```
visa_df[['prevailing_wage','prevailing_wage_norm']].iloc[[max_id,min_id]]
```

Out[13]:

	prevailing_wage	prevailing_wage_norm
21077	319210.2700	1.0
20575	2.1367	0.0

MinMaxscaler

- MinMaxScalar is a method from sklearn preprocessing
- Read the packages
- Save the package
- Apply fit transform

In [18]:

```
############################### Read the data ###########################

file_path="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df

# step-1:

from sklearn.preprocessing import MinMaxScaler
#step-2:

v2=MinMaxScaler()

#step-3:
visa_df['prevailing_wage_norm1']=v2.fit_transform(visa_df[['prevailing_wage
```

In [19]:

```
visa_df[['prevailing_wage_norm1','prevailing_wage']]
```

Out[19]:

	prevailing_wage_norm1	prevailing_wage
0	0.001849	592.2029
1	0.261345	83425.6500
2	0.385312	122996.8600
3	0.261371	83434.0300
4	0.469616	149907.3900
...	...	...
25475	0.241505	77092.5700
25476	0.874579	279174.7900
25477	0.458311	146298.8500
25478	0.269895	86154.7700
25479	0.222033	70876.9100

25480 rows × 2 columns

In [15]:

```
v1=np.array([1,2,3,4])
v1.ndim
```

Out[15]:

```
1
```

Note:

- inside minmaxscaler pass dataframe not serise

Z-score

In [ ]:

```
# step-1: calculate mean
# step-2: calculate std
# step-3: Nr= x-mean
# step-4: Nr/Std
```

In [20]:

```
mean_wage=visa_df['prevailing_wage'].mean()
std_wage=visa_df['prevailing_wage'].std()
nr=visa_df['prevailing_wage']-mean_wage
visa_df['prevailing_wage_zscore']=nr/std_wage
```

In [21]:

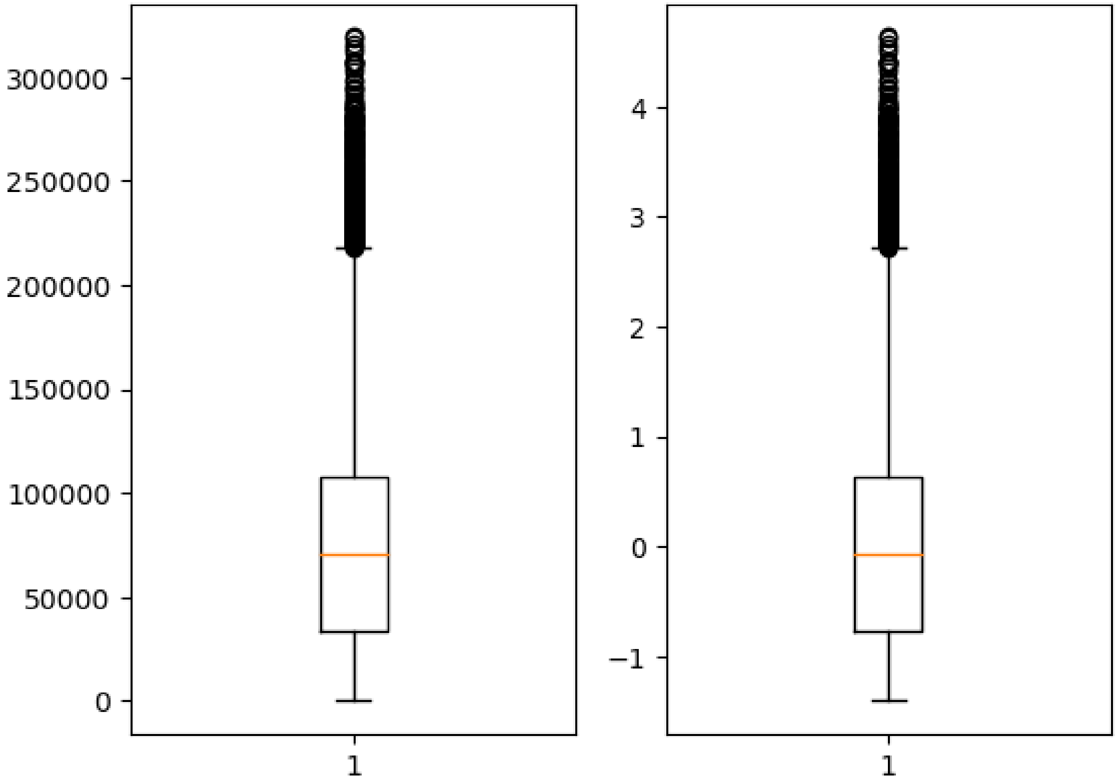
```
visa_df[['prevailing_wage','prevailing_wage_zscore']]
```

Out[21]:

	prevailing_wage	prevailing_wage_zscore
0	592.2029	-1.398510
1	83425.6500	0.169832
2	122996.8600	0.919060
3	83434.0300	0.169991
4	149907.3900	1.428576
...	...	...
25475	77092.5700	0.049923
25476	279174.7900	3.876083
25477	146298.8500	1.360253
25478	86154.7700	0.221504
25479	70876.9100	-0.067762

25480 rows × 2 columns

In [22]:
```
plt.subplot(1,2,1)
plt.boxplot(visa_df['prevailing_wage'])
plt.subplot(1,2,2)
plt.boxplot(visa_df['prevailing_wage_zscore'])
plt.show()
```



StandardScalr

In [23]:
```
file_path="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df

from sklearn.preprocessing import StandardScaler

v3=StandardScaler()

v3.fit_transform(visa_df[['prevailing_wage']])
```

Out[23]:
```
array([[-1.39853722],
       [ 0.1698353 ],
       [ 0.91907852],
       ...,
       [ 1.36027953],
       [ 0.22150859],
       [-0.06776315]])
```

In [ ]:

In [ ]:

In [ ]:

In [ ]: