Import packages

In [1]:

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

Read the data

In [2]:

```
visa_df=pd.read_csv(r"C:\Users\omkar\OneDrive\Documents\Data science\Naresh IT\N
visa_df.head(2)
```

Out[2]:

	case_id	continent	education_of_employee	has_job_experience	requires_job_training
0	EZYV01	Asia	High School	N	N
1	EZYV02	Asia	Master's	Y	N

Select the numerical coulmns

In [3]:

```
visa_df.select_dtypes(exclude='object').columns
```

Out[3]:

```
Index(['no_of_employees', 'yr_of_estab', 'prevailing_wage'], dtype='object')
```

prevailing_wage

- len
- max
- min
- mean
- median
- std
- 75%
- 50%
- 25%

count or len

In [5]:

```
len(visa_df['prevailing_wage'])
```

Out[5]:

```
25480
```

max

```
In [6]: max(visa_df['prevailing_wage']) # Keyword
```

```
Out[6]: 319210.27
```

```
In [7]: visa_df['prevailing_wage'].max() # Pandas
```

```
Out[7]: 319210.27
```

```
In [8]: np.max(visa_df['prevailing_wage']) # numpy
```

```
Out[8]: 319210.27
```

min

```
In [9]: min(visa_df['prevailing_wage'])
```

```
Out[9]: 2.1367
```

```
In [10]: visa_df['prevailing_wage'].min()
```

```
Out[10]: 2.1367
```

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

```
Out[11]: 2.1367
```

```
In [ ]: #instead of len can we use nunique ?
        #how many uniques values different
        #how many total values different
```

mean

```
In [12]: visa_df['prevailing_wage'].mean()
```

```
Out[12]: 74455.81459209183
```

```
In [13]: np.mean(visa_df['prevailing_wage'])
```

```
Out[13]: 74455.81459209183
```

median

```
In [14]: visa_df['prevailing_wage'].median()
```

```
Out[14]: 70308.20999999999
```

```
In [15]: np.median(visa_df['prevailing_wage'])
```

```
Out[15]: 70308.20999999999
```

std

```
In [19]: visa_df['prevailing_wage'].std()
```

```
Out[19]:  52815.94232687357
```

```
In [20]:  np.std(visa_df['prevailing_wage'])
```

```
Out[20]:  52814.90589711402
```

Mode is not good option because it is numerical variable

```
In [24]:  ##All together
          wage_count=round(len(visa_df['prevailing_wage']),2)
          wage_min=round(visa_df['prevailing_wage'].min(),2)
          wage_max=round(visa_df['prevailing_wage'].max(),2)
          wage_mean=round(visa_df['prevailing_wage'].mean(),2)
          wage_median=round(visa_df['prevailing_wage'].median(),2)
          wage_std=round(visa_df['prevailing_wage'].std(),2)
          list_values=[wage_count,wage_min,wage_max,
                       wage_mean,wage_median,wage_std]
          index_val=['count','min','max','mean','median','std']
          pd.DataFrame(list_values,
                       columns=['prevailing_wage'],
                       index=index_val)
```

Out[24]:

	prevailing_wage
count	25480.00
min	2.14
max	319210.27
mean	74455.81
median	70308.21
std	52815.94

Percentile and Quantile

- Percentile:
    - np.percentile()
    - It will take two arguments
        - data :a
        - percentile: q the values varies from 0 to 100
        - if you want 50P data q=50
- Quantile:
    - np.quantile()
    - It will take two arguments
        - data :a

    - percentile: q the values varies from 0 to 1

    - if you want 50p q=0.5

25p-50p-75p

In [28]:

```
wage_25p=round(np.percentile(visa_df['prevailing_wage'],25),2)
wage_50p=round(np.percentile(visa_df['prevailing_wage'],50),2)
wage_75p=round(np.percentile(visa_df['prevailing_wage'],75),2)

print(f"the 25% data is {wage_25p}")
print(f"the 50% data is {wage_50p}")
print(f"the 75% data is {wage_75p}")
```

```
the 25% data is 34015.48
the 50% data is 70308.21
the 75% data is 107735.51
```

In [29]:

```
345.89678
```

Out[29]:

```
345.89678
```

In [30]:

```
round(345.89678,2)
```

Out[30]:

```
345.9
```

In [33]:

```
wage_25p=round(np.quantile(visa_df['prevailing_wage'],0.25),2)
wage_50p=round(np.quantile(visa_df['prevailing_wage'],0.50),2)
wage_75p=round(np.quantile(visa_df['prevailing_wage'],0.75),2)

print(f"the 25% data is {wage_25p}")
print(f"the 50% data is {wage_50p}")
print(f"the 75% data is {wage_75p}")
```

```
the 25% data is 34015.48
the 50% data is 70308.21
the 75% data is 107735.51
```

Understand the percentiles

- defination of 25percentile

    - there 25% of employees has salary less than 34015

    - total employees= 25480

    - 25% of employees= 25*25480/100= 6370

    - 6370 employees salary less than 34015

In [38]:

```
con=visa_df['prevailing_wage']<34015
len(visa_df[con])
```

Out[38]:

```
6370
```

In [39]:

```
con=visa_df['prevailing_wage']<wage_25p
len(visa_df[con])
```

Out[39]:
```
6370
```

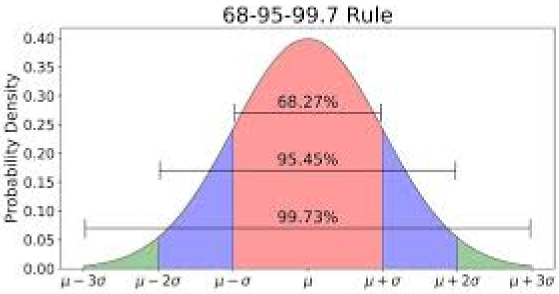
In [41]:
```
con=visa_df['prevailing_wage']<wage_50p
len(visa_df[con])

#50*25480/100
```

Out[41]:
```
12740.0
```

In [42]:
```
con=visa_df['prevailing_wage']<wage_75p
len(visa_df[con])
```

Out[42]:
```
19110
```

Emperical rule (68-95-99.7)



- First calculate mean value
- Second calculate std value
- Con1: mean-1*std
- Con2: mean+1*std
- If you apply above conditions on wage data , the output count should be equal to 68percentile data
- 68% of total employees: 17326

In [60]:
```
v1=wage_mean-1*wage_std
v2=wage_mean+1*wage_std
con1=visa_df['prevailing_wage']>v1
con2=visa_df['prevailing_wage']<v2

count1=len(visa_df[con1 & con2])
###############################################################
count1 ,68*25480/100
```

Out[60]:
```
(17171, 17326.4)
```

In [62]:
```
v1=wage_mean-2*wage_std
v2=wage_mean+2*wage_std
con1=visa_df['prevailing_wage']>v1
con2=visa_df['prevailing_wage']<v2
```

```
count1=len(visa_df[con1 & con2])
###############################################################
count1 ,95*25480/100
```

Out[62]:

```
(24582, 24206.0)
```

In [63]:

```
v1=wage_mean-3*wage_std
v2=wage_mean+3*wage_std
con1=visa_df['prevailing_wage']>v1
con2=visa_df['prevailing_wage']<v2

count1=len(visa_df[con1 & con2])
###############################################################
count1 ,99.7*25480/100
```

Out[63]:

```
(25186, 25403.56)
```

In [ ]: