In [ ]: In [1]:

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

In [2]:

```
# Read the packages

# Read the data
```

```
file_path="C:\\Users\\omkar\\OneDrive\\Do
cuments\\Data science\\Naresh IT\\
visa_df=pd.read_csv(file_path)
visa_df
```

Out[2]: **case_id continent education_of_employee has_job_experience requires_job_traini** **0** EZYV01 Asia High

School N

1 EZYV02 Asia Master's Y

2 EZYV03 Asia Bachelor's N

3 EZYV04 Asia Bachelor's N

4 EZYV05 Africa Master's Y

... ...

25475 EZYV25476 Asia Bachelor's Y

25476 EZYV25477 Asia High School Y

25477 EZYV25478 Asia Master's Y

25478 EZYV25479 Asia Master's Y

25479 EZYV25480 Asia Bachelor's Y

25480 rows × 12 columns

```
visa_df.col
umns
```

In [3]:

```
Out[3]: Index(['case_id', 'continent', 'education_of_employee', 'has_job_experienc
e',
    'requires_job_training',   'no_of_employees',   'yr_of_estab',
  'region_of_employment',     'prevailing_wage',     'unit_of_wage',
  'full_time_position', 'case_status'],
   dtype='object')
```

In [4]:

```
visa_df['prevailing_wage
'] # as a series
```

Out[4]:

```
0 592.2029
 1 83425.6500
 2 122996.8600
 3 83434.0300
 4 149907.3900
   ...
 25475 77092.5700
 25476 279174.7900
 25477 146298.8500
 25478 86154.7700
 25479 70876.9100
 Name: prevailing_wage, Length: 25480, dtype: float64
```

In [5]:

```
visa_df['prevailing_
wage'].values
```

Out[5]:

```
array([ 592.2029, 83425.65 , 122996.86 , ..., 146298.85 ,
  86154.77 , 70876.91 ])
```

count

min

max

mean

median

In [29]:

```
dict1={'names':['Ram
','Sita'],
'age':[25,20]}
pd.DataFrame(dict1)
```

```
dict2={'name':'Ram',
 'age':25}
pd.DataFrame(dict2,i
ndex=['A'])
```

Out[29]: **name age A** Ram

25

In [21]:

```
dict1={}
wage_count=round(visa_df['prevai
ling_wage'].count(),2)
wage_min=round(visa_df['prevaili
ng_wage'].min(),2)
wage_max=round(visa_df['prevaili
ng_wage'].max(),2)
wage_mean=round(visa_df['prevail
ing_wage'].mean(),2)
wage_median=round(visa_df['preva
iling_wage'].median(),2)
wage_std=round(visa_df['prevaili
ng_wage'].std(),2)

dict1['count']=wage_count
dict1['min']=wage_min
dict1['max']=wage_max
dict1['mean']=wage_mean
dict1['median']=wage_median
dict1['std']=wage_std
pd.DataFrame(dict1,index=['preva
iling_wage'])
```

Out[21]: **count min max mean median std prevailing_wage** 25480 2.14 319210.27

74455.81 70308.21 52815.94

In [32]:

```
dict1={}
wage_count=round(visa_df['prevailing_w
age'].count(),2)
wage_min=round(visa_df['prevailing_wag
e'].min(),2)
wage_max=round(visa_df['prevailing_wag
e'].max(),2)
wage_mean=round(visa_df['prevailing_wa
ge'].mean(),2)
wage_median=round(visa_df['prevailing_
wage'].median(),2)
wage_std=round(visa_df['prevailing_wag
e'].std(),2)
list1=[wage_count,wage_min,wage_max,wa
ge_mean,wage_median,wage_std]
dict1['prevailing_wage']=list1
dict1
pd.DataFrame(dict1)
```

Out[32]: **prevailing_wage 0**

25480.00

1 2.14

2 319210.27

3 74455.81

4 70308.21

5 52815.94

����������-��������

In [30]:

```
wage_count=round(visa_df['prevailing_w
age'].count(),2)
wage_min=round(visa_df['prevailing_wag
e'].min(),2)
wage_max=round(visa_df['prevailing_wag
e'].max(),2)
wage_mean=round(visa_df['prevailing_wa
ge'].mean(),2)
wage_median=round(visa_df['prevailing_
wage'].median(),2)
wage_std=round(visa_df['prevailing_wag
e'].std(),2)
```

```
list1=[wage_count,wage_min,wage_max,wage_mean,wage_median,wage_std]
pd.DataFrame(list1,
```

```
 columns=['prevailing_wage'],

index=['count','min','max','mean','median','std'])
```

Out[30]: **prevailing_wage** **count**

25480.00

min 2.14

max 319210.27

mean 74455.81

median 70308.21

std 52815.94

In [46]:

```
#step-1: numerical column list
dtypes=dict(visa_df.dtypes)
num=[i for i in dtypes if
dtypes[i]!='O']
print(num)

dict1={}
for i in num:
 count=round(visa_df[i].count(),2)
 MIN=round(visa_df[i].min(),2)
 MAX=round(visa_df[i].max(),2)
 mean=round(visa_df[i].mean(),2)
 median=round(visa_df[i].median(),2)
```

```
 std=round(visa_df[i].std(),2)

 list1=[count,MIN,MAX,mean,median,std]
 dict1[i]=list1

 df=pd.DataFrame(dict1,

index=['count','min','max','mean','median','std'])


dict1
```

```
['no_of_employees', 'yr_of_estab',
'prevailing_wage']
```

```
Out[46]: {'no_of_employees': [25480, -26, 602069, 5667.04, 2109.0, 22877.93],
         'yr_of_estab': [25480, 1800, 2016, 1979.41, 1997.0, 42.37],
          'prevailing_wage': [25480, 2.14, 319210.27, 74455.81, 70308.21,
         52815.9 4]}
```

In [47]:

```
df
```

Out[47]: **no_of_employees** **yr_of_estab** **prevailing_wage** **count**

25480.00 25480.00 25480.00 **min** -26.00 1800.00 2.14 **max** 602069.00

2016.00 319210.27 **mean** 5667.04 1979.41 74455.81 **median** 2109.00

1997.00 70308.21 **std** 22877.93 42.37 52815.94

In [48]:

```
visa_df.describe()
```

Out[48]: **no_of_employees** **yr_of_estab** **prevailing_wage** **count**

25480.000000 25480.000000 25480.000000 **mean**

5667.043210 1979.409929 74455.814592

std 22877.928848 42.366929 52815.942327 **min** -26.000000

1800.000000 2.136700 **25%** 1022.000000 1976.000000

34015.480000 **50%** 2109.000000 1997.000000 70308.210000

75% 3504.000000 2005.000000 107735.512500 **max**

602069.000000 2016.000000 319210.270000

In [ ]: In [49]:

```
visa_df['prevailing_wage'].mean()
```

- we implemented describe function **with** our own python skill

```
# Reading a specific column
# we have a mean method
```

Out[49]: 74455.81459209183

In [ ]: In [51]:

```
np.median(visa_df['prevailing_wage'])
np.std(visa_df['prevailing_wage'])
np.min(visa_df['prevailing_wage'])
np.max(visa_df['prevailing_wage'])
```

```
#np.mean(<specific column data>)

np.mean(visa_df['prevailing_wage'])
```

Out[51]: 319210.27

��������������������-����������������

In [ ]:

- quantile q1=25P q2=50p q3=75p

- np.percentile(<direct number between 1 to 100>,data)

- ex: np.percentile(75,data)

- np.quantile(<decimal number>,data)

- ex: np.quantile(0.75,data)

In [54]: In [55]: In [58]:

```
q1=round(np.percentile(visa_df['prevailing_wage'],25),2)
q2=round(np.percentile(visa_df['prevailing_wage'],50),2)
q3=round(np.percentile(visa_df['prevailing_wage'],75),2)
print(q1,q2,q3)
```

```
34015.48 70308.21 107735.51
```

- percentile ranges **from** 1 to 100

```
q1=round(np.quantile(visa_df['prevailing_w
age'],0.25),2)
q2=round(np.quantile(visa_df['prevailing_w
age'],0.50),2)
q3=round(np.quantile(visa_df['prevailing_w
age'],0.75),2)
print(q1,q2,q3)
```

```
34015.48 70308.21 107735.51
```

```
#step-1: numerical column list
dtypes=dict(visa_df.dtypes)
num=[i for i in dtypes if dtypes[i]!='O']
print(num)

dict1={}
for i in num:
 count=round(visa_df[i].count(),2)
 MIN=round(visa_df[i].min(),2)
 MAX=round(visa_df[i].max(),2)
 mean=round(visa_df[i].mean(),2)
 median=round(visa_df[i].median(),2)
 std=round(visa_df[i].std(),2)
 ###################################
 q1=round(np.percentile(visa_df[i],25),2)
 q2=round(np.percentile(visa_df[i],50),2)
 q3=round(np.percentile(visa_df[i],75),2)


list1=[count,MIN,MAX,mean,median,std,q1,q2
,q3]
 dict1[i]=list1

 df=pd.DataFrame(dict1,


index=['count','min','max','mean','median'

,'std','25%',
```

```
['no_of_employees', 'yr_of_estab',
'prevailing_wage']
```

In [59]:

```
df
```

Out[59]: **no_of_employees yr_of_estab prevailing_wage** **count**

25480.00 25480.00 25480.00 **min** -26.00 1800.00 2.14 **max** 602069.00

2016.00 319210.27 **mean** 5667.04 1979.41 74455.81 **median** 2109.00

1997.00 70308.21 **std** 22877.93 42.37 52815.94 **25%** 1022.00 1976.00

34015.48 **50%** 2109.00 1997.00 70308.21 **75%** 3504.00 2005.00 107735.51

In [60]:

```
q1=round(np.percentile(visa_df['
prevailing_wage'],25),2) q1
```

Out[60]: 34015.48

what is the meaning of this

In [ ]: In [ ]:

```
#total_obsrvations=25480
#25percentagae(25480)
#25*25480/100= 6370

#6370 people have wages less
than 34015
```

In [64]:

```
#25 percentage of
observations from the total
data #have a value below
34015
```

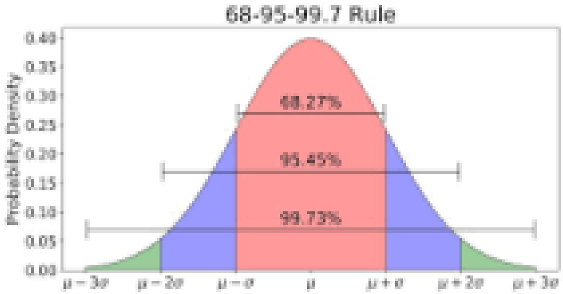
```
len(visa_df[visa_df['prevaili

ng_wage']<34015]) # how many
```

```
True= 6370
```

```
Out[64]: 6370
```

������������������-��������



Select an Image
Right click select insepect

Right side you will img src
Right click on img src and
select Edit as HTML dont move
your curosr
CTRL+A
CTRL+C
CTRL+V
ESC+M
SHIFT+ENTER

when data follows a normal distribution

u-1*sigma to u+1*sigma : 68%
u-2*sigma to u+2*sigma : 95%
u-3*sigma to u+3*sigma : 99.7%

```
In [66]:
```

click on inspect

```
wage_mean,wage_std
```

```
Out[66]: (74455.81, 52815.94)
```

In [68]: In [70]:

```
##################### 68% ###################
val_minus_1=round(wage_mean-1*wage_std,2)
val_plus_1=round(wage_mean+1*wage_std,2)

#################### 95%########################
val_minus_2=round(wage_mean-2*wage_std,2)
val_plus_2=round(wage_mean+2*wage_std,2)

#################### 99.7%########################
val_minus_3=round(wage_mean-3*wage_std,2)
val_plus_3=round(wage_mean+3*wage_std,2)

print(val_minus_1,val_plus_1,val_minus_2,val_plus_2,val_minus_3,val_plus_3)
```

```
21639.87 127271.75 -31176.07 180087.69 -83992.01 232903.63
```

68 percentage of observations have values between [21639.87,127271.75] 95 percentage of observations have values between [-31176.07,180087.69] 99.7 percentage of observations have values between [-83992.01,232903.63]

```
68*25480/100
```

Out[70]: 17326.4

In [76]:

```
con1=visa_df['prevailing_wage']>val_minus_1
con2=visa_df['prevailing_wage']<val_plus_1
len(visa_df[con1&con2])
len(visa_df[con1&con2])/len(visa_df)
```

Out[76]: 0.673901098901099

In [77]:

```
con1=visa_df['prevailing_wage']>val_minus_2
con2=visa_df['prevailing_wage']<val_plus_2
len(visa_df[con1&con2])
len(visa_df[con1&con2])/len(visa_df)
```

Out[77]:

```
0.9647566718995291
```

In [78]:

```
con1=visa_df['prevailing_wage']>val_minus_3
con2=visa_df['prevailing_wage']<val_plus_3
len(visa_df[con1&con2])
len(visa_df[con1&con2])/len(visa_df)
```

Out[78]:

```
0.9884615384615385
```

In [ ]:

```
68-95-99.7
67-96-98
```