In [3]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
file_path='C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df
```

Out[4]:

| | 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

In [7]:
```
visa_df.columns
```

Out[7]:
```
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 [8]:
```
visa_df['prevailing_wage']  # as a series
```

Out[8]:
```
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 [9]:
```
visa_df['prevailing_wage'].values
```

Out[9]:
```
array([   592.2029,  83425.65  , 122996.86  , ..., 146298.85  ,
        86154.77  ,  70876.91  ])
```

- count
- min
- max
- mean
- median
- standerd deviation

In [5]:
```
dict1={'names':['Ram','Sita'],      # pure dictionary is not there so it ca
       'age':[25,20]}
pd.DataFrame(dict1)
```

Out[5]:

| | names | age |
|---|---|---|
| **0** | Ram | 25 |
| **1** | Sita | 20 |

In [6]:
```
dict2={'name':'Ram',              # here dictionary pure form so we are usi
       'age':25}

pd.DataFrame(dict2,index=['A'])
```

Out[6]:

| | name | age |
|---|---|---|
| **A** | Ram | 25 |

$Method - 1$

**using dictionary to make data frame**

In [11]:
```
dict1={}
wage_count=round(visa_df['prevailing_wage'].count(),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)

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=['prevailing_wage'])
```

Out[11]:

| | count | min | max | mean | median | std |
|---|---|---|---|---|---|---|
| **prevailing_wage** | 25480 | 2.14 | 319210.27 | 74455.81 | 70308.21 | 52815.94 |

In [12]:
```
dict1={}
wage_count=round(visa_df['prevailing_wage'].count(),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)
list1=[wage_count,wage_min,wage_max,wage_mean,wage_median,wage_std]
dict1['prevailing_wage']=list1
dict1
pd.DataFrame(dict1)
```

Out[12]:

| | prevailing_wage |
|---|---|
| **0** | 25480.00 |
| **1** | 2.14 |
| **2** | 319210.27 |
| **3** | 74455.81 |
| **4** | 70308.21 |
| **5** | 52815.94 |

$Using - List$

In [13]:
```
wage_count=round(visa_df['prevailing_wage'].count(),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)

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[13]:

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

**Seperation of categorical column and numerical column**

In [11]:
```
#step-1: numerical column list
dtypes=dict(visa_df.dtypes)
num=[i for i in dtypes if dtypes[i]!='O']
print(num)



# colume with numerical data

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[11]:
```
{'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 [15]:
```
df
```

Out[15]:

| | 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 [16]:
```
visa_df.describe()
```

Out[16]:

| | 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 [ ]:
```
# we implemented describe function with our own python skill
```

In [ ]:
```

```

In [17]:
```
visa_df['prevailing_wage'].mean()

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

Out[17]:
```
74455.81459209183
```

In [ ]:
```
# np.mean(<specific column data>)
```

In [12]:
```
np.mean(visa_df['prevailing_wage'])
np.median(visa_df['prevailing_wage'])
np.std(visa_df['prevailing_wage'])
np.min(visa_df['prevailing_wage'])
np.max(visa_df['prevailing_wage'])
```

Out[12]:
```
319210.27
```

*Percentile – Quantile*

- percentile ranges from 1 to 100
- quantile q1=25P q2=50p q3=75p
- np.percentile(<direct number between 1 to 100>,data)
- ex: np.percentile(75,data)
- np.quantile(,data)
- ex: np.quantile(0.75,data)

In [19]:
```
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
```

In [20]:
```
q1=round(np.quantile(visa_df['prevailing_wage'],0.25),2)
q2=round(np.quantile(visa_df['prevailing_wage'],0.50),2)
q3=round(np.quantile(visa_df['prevailing_wage'],0.75),2)
print(q1,q2,q3)
```

```
34015.48 70308.21 107735.51
```

In [21]:
```
#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 [22]:
```
df
```

Out[22]:

| | 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 [23]:
```
q1=round(np.percentile(visa_df['prevailing_wage'],25),2)
q1
```

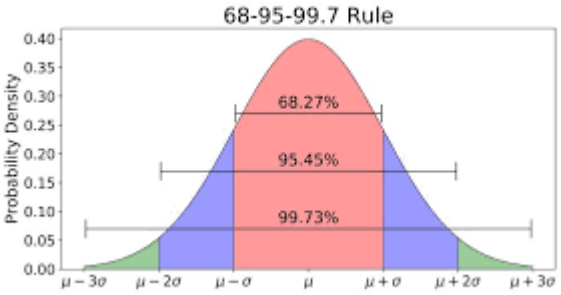
Out[23]: 34015.48

**What is the meaning of this**

In [ ]:
```
#25 percentage of observations from the total data
#have a value below 34015


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

#6370 people have wages less than 34015
```

In [24]:
```
len(visa_df[visa_df['prevailing_wage']<34015])

# how many True= 6370
```

Out[24]: 6370




- Select an Image
- Right click select insepect
- click on inspect
- 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 [25]:
```
wage_mean,wage_std
```

Out[25]:
```
(74455.81, 52815.94)
```

In [26]:
```
###################### 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]

In [27]:
```
68*25480/100
```

Out[27]:
```
17326.4
```

In [28]:
```
# 68%

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[28]:
```
0.673901098901099
```

In [29]:
```
# 95%
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[29]:
```
0.9647566718995291
```

In [30]:
```
# 99.7#

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[30]:
```
0.9884615384615385
```

In [ ]:
```
68-95-99.7
67-96-98
```

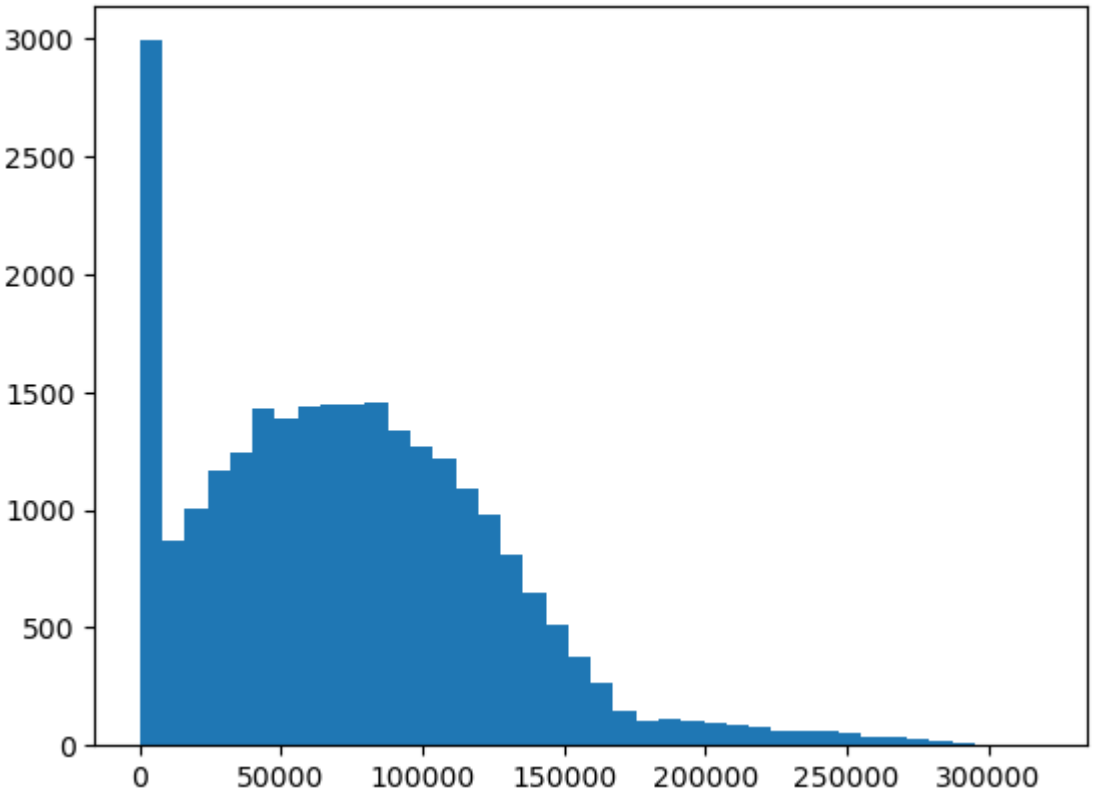
In [ ]:
```

```

In [31]:
```
# Import the packages
# Read the data

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

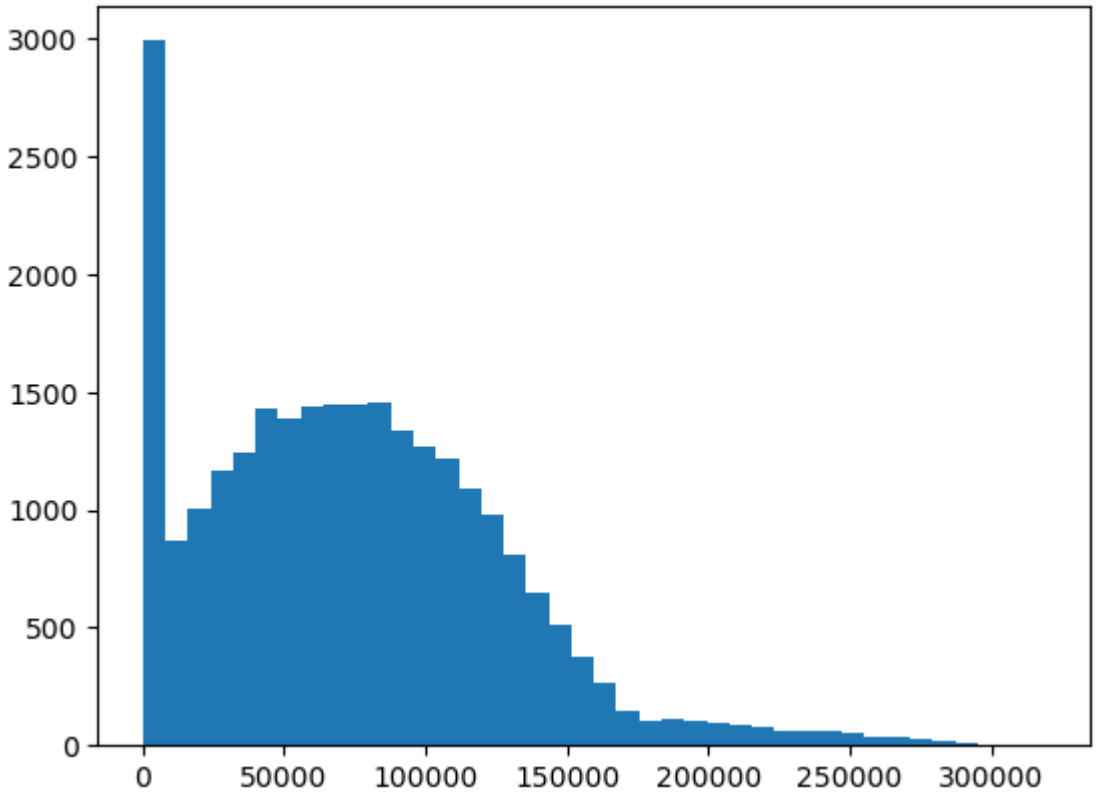
In [ ]:
```
file_path='C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df
```

- we read p_wage column
- we perform statistical analysis
- we perform empiricle rule analysis
- empiricle rule: 68-95-99.7
- p_wage 67-96-98
- wage_mean=74455
- wage_median=70308
- median<mean
- data might be right skewed
- but percentage of data almost valid with empiricle
- it is looks like normal distribution and slightly right skewed
- In order to see that darw histogram

In [32]:
```
plt.hist(visa_df['prevailing_wage'],
         bins=40)
plt.show()

# by defualt it will give as 10 intervals
# if you want increase the intervals
# argument name bins
```



In [33]:
```
frequency,interval,n=plt.hist(visa_df['prevailing_wage'],
         bins=40)

# frequency means number of observations are fall between an interval
# interval
# n= number of intervals
```




In [ ]:
```
# 2992 observations are between 2.13670000e+00, 7.98234003e+03
# 871 observations are between 7.98234003e+03, 1.59625434e+04
```

In [ ]:
```
len(frequency),len(interval),len(n)
```

In [34]:
```
# 2992 observations are between 2.13670000e+00, 7.98234003e+03
# verify

# step-1: write con1 = <col>>2.13
# step-2: write con2= <col><7982.3
# step-3: con1&con2
# step-4: col[con1&con2]
# step-5: len(col[con1&con2])

con1=visa_df['prevailing_wage']>2.13670000e+00
con2=visa_df['prevailing_wage']<7.98234003e+03
len(visa_df[con1&con2])
```

Out[34]: 2991

In [35]:
```
len(visa_df[visa_df['prevailing_wage'].between(2.13670000e+00,7.98234003e+0
```

Out[35]: 2992

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: