In [1]:
```
# import the packages
# read the data
# divide into cat num coumns

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

path='Visadataset.csv'
visa_df=pd.read_csv(path)
cat=visa_df.select_dtypes(include='object').columns
num=visa_df.select_dtypes(exclude='object').columns
```

In [2]:
```
num
```

Out[2]:
```
Index(['no_of_employees', 'yr_of_estab', 'prevailing_wage'], dtype='object')
```

prevailing_wage

- len
- max
- min
- mean
- median
- 25percentile
- 50 percentile
- 75p

In [5]:
```
wage_data=visa_df['prevailing_wage']
len(wage_data)
```

Out[5]:
```
25480
```

In [6]:
```
min(wage_data)
```

Out[6]:
```
2.1367
```

In [7]:
```
wage_data.min()
```

Out[7]:
```
2.1367
```

In [8]:
```
np.min(wage_data)
```

Out[8]:
```
2.1367
```

In [9]:
```
wage_data=visa_df['prevailing_wage']
min(wage_data),wage_data.min(),np.min(wage_data)
```

Out[9]:
```
(2.1367, 2.1367, 2.1367)
```

In [10]:
```
max(wage_data),wage_data.max(),np.max(wage_data)
```

Out[10]:
```
(319210.27, 319210.27, 319210.27)
```

In [11]:
```
wage_data.mean(),np.mean(wage_data)
```

Out[11]:
```
(74455.81459209183, 74455.81459209183)
```

In [12]:
```
wage_data.median(),np.median(wage_data)
```

Out[12]:
```
(70308.20999999999, 70308.20999999999)
```

qunatile-percentile

In [14]:
```
p_25=np.percentile(wage_data,25)
p_25
```

Out[14]:
```
34015.479999999996
```

In [26]:
```
p_50=np.percentile(wage_data,50)
p_50
```

Out[26]:
```
70308.20999999999
```

In [28]:
```
p_75=np.percentile(wage_data,75)
p_75
```

Out[28]:
```
107735.51250000001
```

In [30]:
```
q_25=np.quantile(wage_data,0.25)
q_50=np.quantile(wage_data,0.50)
q_75=np.quantile(wage_data,0.75)
```

In [32]:
```
np.quantile(wage_data,0.25)
```

Out[32]:
```
34015.479999999996
```

In [34]:
```
wage_data=visa_df['prevailing_wage']
wage_count=len(wage_data)
wage_min=round(wage_data.min(),2)
wage_mean=round(wage_data.mean(),2)
wage_med=round(wage_data.median(),2)
wage_25p=round(np.percentile(wage_data,25))
wage_50p=round(np.percentile(wage_data,50))
wage_75p=round(np.percentile(wage_data,75))
wage_max=round(wage_data.max(),2)
l=[wage_count,wage_min,wage_mean,wage_med,
   wage_25p,wage_50p,wage_75p,wage_max]
Id=['Count','Min','Mean','Median','25p','50p',
     '75p','Max']
```

```
pd.DataFrame(l,columns=['prevailing_wage'],index=Id)
#pd.DataFrame(l,Id,['prevailing_wage'])
```

Out[34]:

	prevailing_wage
Count	25480.00
Min	2.14
Mean	74455.81
Median	70308.21
25p	34015.00
50p	70308.00
75p	107736.00
Max	319210.27

In [36]:

```
L=[]
for i in num:
    data=visa_df[i]
    count=len(data)
    Min=round(data.min(),2)
    mean=round(data.mean(),2)
    med=round(data.median(),2)
    p_25=round(np.percentile(data,25))
    p_50=round(np.percentile(data,50))
    p_75=round(np.percentile(data,75))
    Max=round(data.max(),2)
    l=[count,Min,mean,med,
       p_25,p_50,p_75,Max]
    L.append(l)
    Id=['Count','Min','Mean','Median','25p','50p',
        '75p','Max']

pd.DataFrame(L,columns=Id,index=num).T
```

Out[36]:

	no_of_employees	yr_of_estab	prevailing_wage
Count	25480.00	25480.00	25480.00
Min	-26.00	1800.00	2.14
Mean	5667.04	1979.41	74455.81
Median	2109.00	1997.00	70308.21
25p	1022.00	1976.00	34015.00
50p	2109.00	1997.00	70308.00
75p	3504.00	2005.00	107736.00
Max	602069.00	2016.00	319210.27

In [38]:

```
visa_df.describe()
```

Out[38]:

	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

percentile concept

- 25p: 25percentage of data below 25p value
- wage_25p: 34015.48
  - there 25*(25480)/100=6370 applicants have salary less tha 34015
- wage_50p: 70308.12
  - there 50*(25480)/100=12740 applicants have salary less tha 70308.12
- wage_75p: 107735.51
  - there 75*(25480)/100=19110 applicants have salary less tha 107735.51

In [41]:
```
con=visa_df['prevailing_wage']<wage_25p
len(visa_df[con]),len(visa_df)*25/100
len(visa_df[con])==len(visa_df)*25/100
```

Out[41]:
```
True
```

In [43]:
```
len(visa_df)*25/100
```

Out[43]:
```
6370.0
```

In [45]:
```
# **Emperical rule**

# u-1*sigma  to  u+1*sigma 68%

# step-1: wage data mean (u)
# step-2: wage data std (sigma)
# step-3: LB=u-1*sigma
# step-4: UB=u+1*sigma
# step-5: con1=visa_df['prevailing_wage']>LB
# step-6: con2=visa_df['prevailing_wage']<UB
# step-7: con=con1 & con2
# step-8: len(visa_df[con])
```

In [47]:
```
wage_data=visa_df['prevailing_wage']
U=wage_data.mean()
sigma=wage_data.std()
```

```
LB=U-1*sigma
UB=U+1*sigma
con1=visa_df['prevailing_wage']>LB
con2=visa_df['prevailing_wage']<UB
con=con1 & con2
len(visa_df[con])==68*25480/100
```

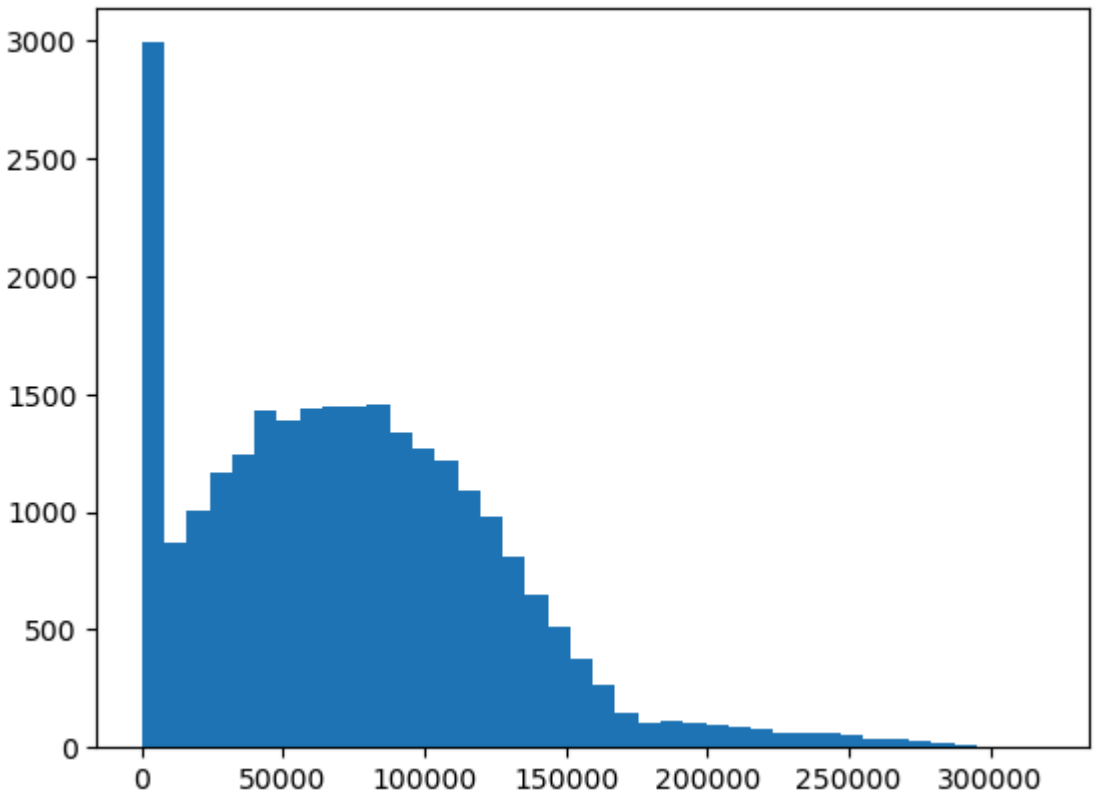
Out[47]: False

Conclusion: Emperical rule failed wage data does not following Normal distribution

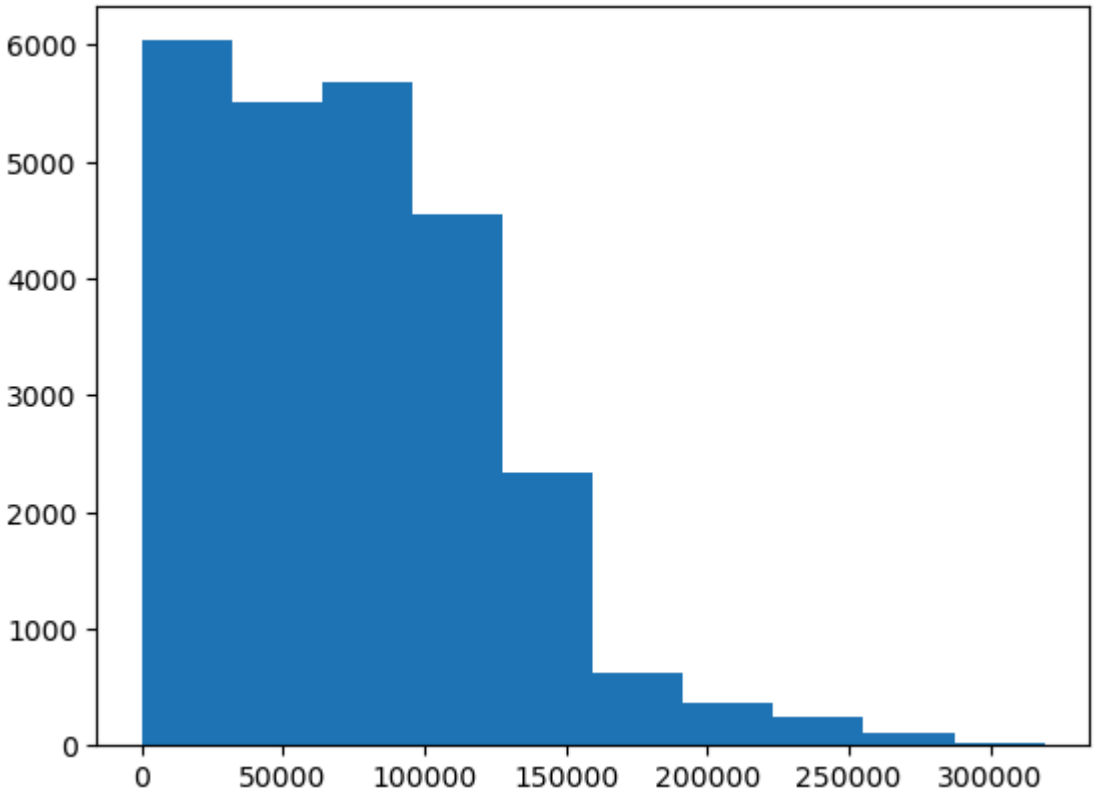
Histogram

In [51]:
```
plt.hist(wage_data,bins=40)
```

Out[51]:
```
(array([2992.,  871., 1005., 1170., 1242., 1434., 1385., 1443., 1444.,
        1445., 1457., 1335., 1268., 1217., 1088.,  978.,  807.,  645.,
         509.,  373.,  264.,  144.,  105.,  111.,  107.,   99.,   88.,
          79.,   65.,   64.,   58.,   53.,   33.,   33.,   29.,   19.,
           7.,    3.,    6.,    5.]),
 array([2.13670000e+00, 7.98234003e+03, 1.59625434e+04, 2.39427467e+04,
        3.19229500e+04, 3.99031534e+04, 4.78833567e+04, 5.58635600e+04,
        6.38437634e+04, 7.18239667e+04, 7.98041700e+04, 8.77843734e+04,
        9.57645767e+04, 1.03744780e+05, 1.11724983e+05, 1.19705187e+05,
        1.27685390e+05, 1.35665593e+05, 1.43645797e+05, 1.51626000e+05,
        1.59606203e+05, 1.67586407e+05, 1.75566610e+05, 1.83546813e+05,
        1.91527017e+05, 1.99507220e+05, 2.07487423e+05, 2.15467627e+05,
        2.23447830e+05, 2.31428033e+05, 2.39408237e+05, 2.47388440e+05,
        2.55368643e+05, 2.63348847e+05, 2.71329050e+05, 2.79309253e+05,
        2.87289457e+05, 2.95269660e+05, 3.03249863e+05, 3.11230067e+05,
        3.19210270e+05]),
 <BarContainer object of 40 artists>)
```



In [52]:
```
inter_count,inter_vals,n=plt.hist(wage_data,bins=10)
```



In [55]:
```
inter_count
```

Out[55]:
```
array([6038., 5504., 5681., 4551., 2334.,  624.,  373.,  240.,  114.,
         21.])
```

In [57]:
```
inter_vals
```

Out[57]:
```
array([2.13670000e+00, 3.19229500e+04, 6.38437634e+04, 9.57645767e+04,
       1.27685390e+05, 1.59606203e+05, 1.91527017e+05, 2.23447830e+05,
       2.55368643e+05, 2.87289457e+05, 3.19210270e+05])
```

In [59]:
```
2.13670000e+00, 3.19229500e+04,
```

Out[59]:
```
(2.1367, 31922.95)
```

In [61]:
```
3.19229500e+04
```

Out[61]:
```
31922.95
```

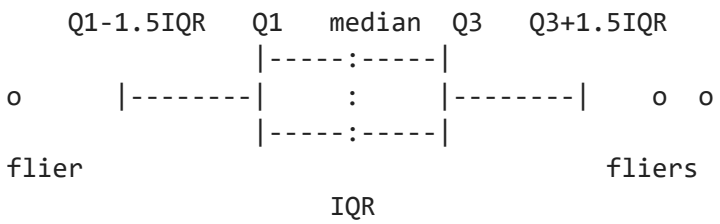
In [63]:
```
wage_data=visa_df['prevailing_wage']
lb=2.13670000e+00
ub=3.19229500e+04
con1=wage_data>=lb
con2=wage_data<ub
con=con1 & con2
len(visa_df[con])
```

Out[63]:
```
6038
```

Outlier analysis

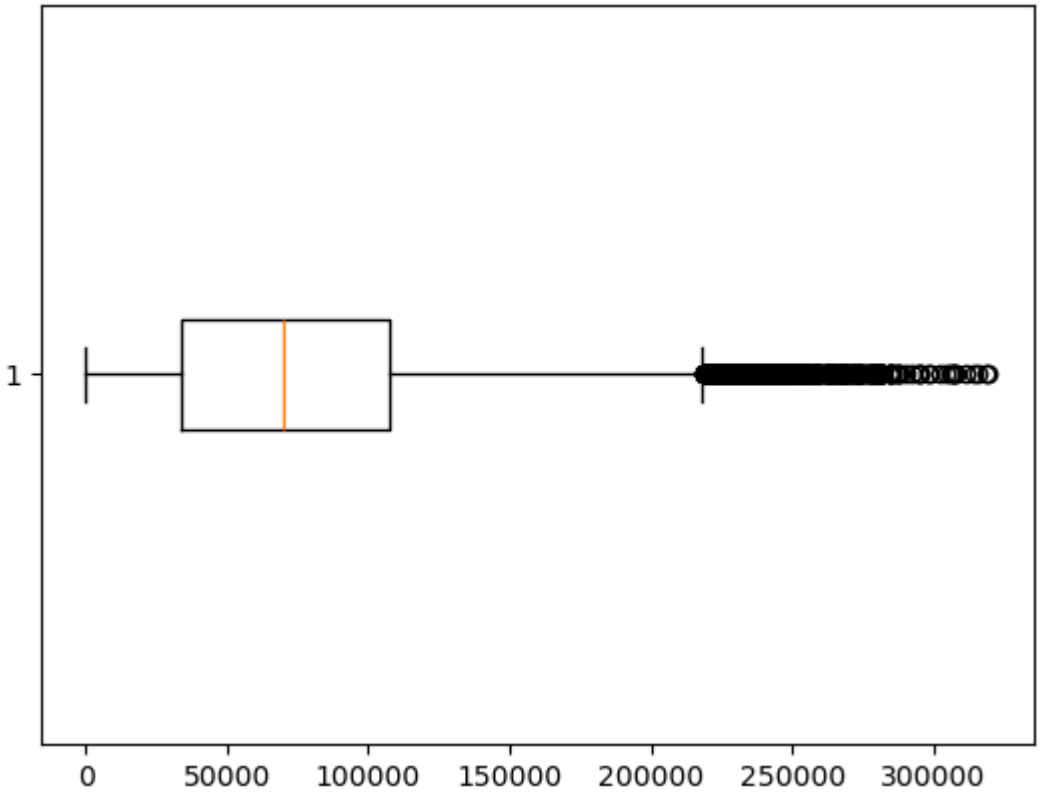
- we understood emperical rule failed

- Then we plotted histogram its slightly deviated
- So it indicates the data has outliers
- Outliers can be show by **box plot**

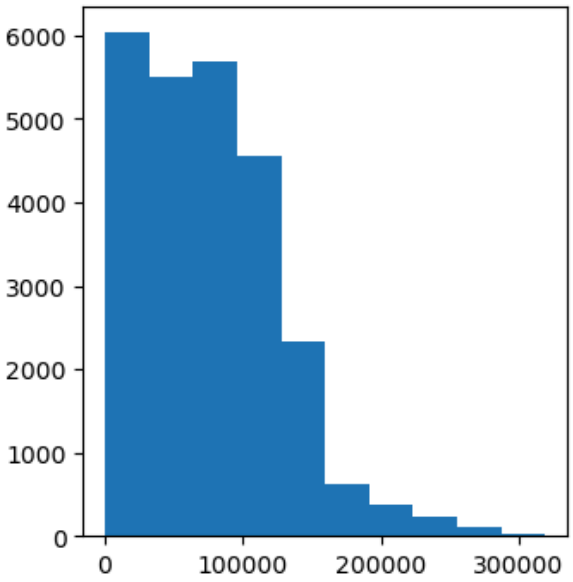


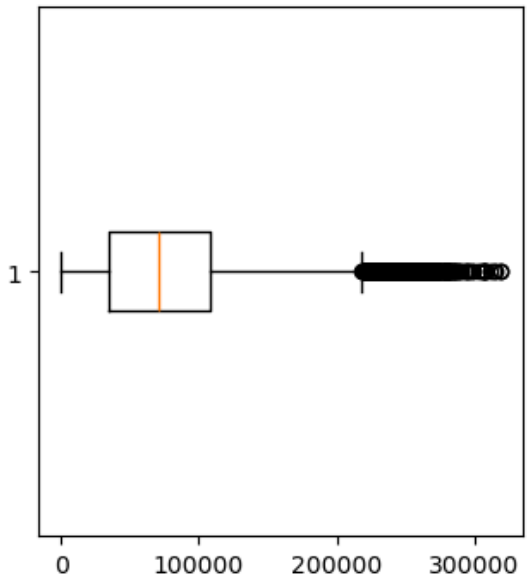
boxplot

In [78]:
```
wage_data=visa_df['prevailing_wage']
plt.boxplot(wage_data,vert=False)
plt.show()
```



In [92]:
```
plt.figure(figsize=(8,4))
plt.subplot(1,2,1).hist(wage_data)
plt.subplot(1,2,2).boxplot(wage_data,vert=False)
plt.show()
```





outliers data

In [ ]:
```
# q1
# q2
# q3
# IQR=q3-q1
# UB=q3+1.5*IQR
# LB=q3-1.5*IQR
# con1= wage >UB
# con2= wage<LB
# con= con1 | con2  outliers

# con1= wage<UB
# con2= wage > LB
# con3= con1 & con2
```

In [99]:
```
q1=np.percentile(wage_data,25)
q2=np.percentile(wage_data,50)
q3=np.percentile(wage_data,75)
IQR=q3-q1
UB=q3+1.5*IQR
LB=q1-1.5*IQR
con1= wage_data >UB
con2= wage_data<LB
con= con1 | con2
outliers_df=visa_df[con]
len(outliers_df)
```

Out[99]: 427

In [131…
```
outliers_df
```

Out[131…

	case_id	continent	education_of_employee	has_job_experience	requires_job_t
14	EZYV15	Asia	Master's	Y	
34	EZYV35	Asia	Master's	N	
130	EZYV131	South America	High School	N	
216	EZYV217	Asia	Master's	Y	
221	EZYV222	North America	Doctorate	Y	
...	...	...	...	...	
25191	EZYV25192	Asia	Master's	N	
25195	EZYV25196	North America	Master's	Y	
25468	EZYV25469	Asia	Bachelor's	N	
25469	EZYV25470	North America	Master's	Y	
25476	EZYV25477	Asia	High School	Y	

427 rows × 12 columns

In [101…

```
q1=np.percentile(wage_data,25)
q2=np.percentile(wage_data,50)
q3=np.percentile(wage_data,75)
IQR=q3-q1
UB=q3+1.5*IQR
LB=q1-1.5*IQR
con1= wage_data<UB
con2= wage_data>LB
con= con1 & con2
non_outliers_df=visa_df[con]
len(non_outliers_df)
```

Out[101…

```
25053
```

In [103…

```
25053+427==25480
```

Out[103…

```
True
```

fill the outliers

- drop the outliers
- fill with median
- winsorization: fill with LB and UB

In [125…

```
# Fill the outliers using median
outliers_df
```

```
outliers_df['prevailing_wage']
outliers_df['prevailing_wage'].values
outliers=outliers_df['prevailing_wage'].values.tolist()
l=[]
wage_data=visa_df['prevailing_wage']
wage_med=wage_data.median()
for value in wage_data:
    if value in outliers:
        l.append(wage_med)
    else:
        l.append(value)
visa_df['prevailing_wage1']=l
visa_df
```

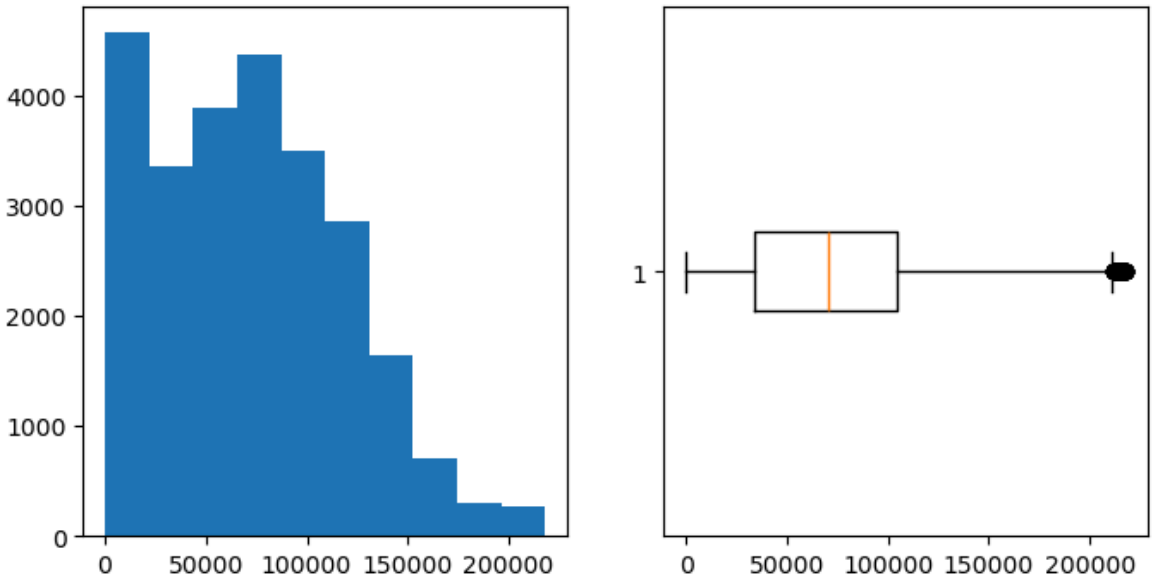
Out[125…

	case_id	continent	education_of_employee	has_job_experience	requires_job_t
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 × 13 columns

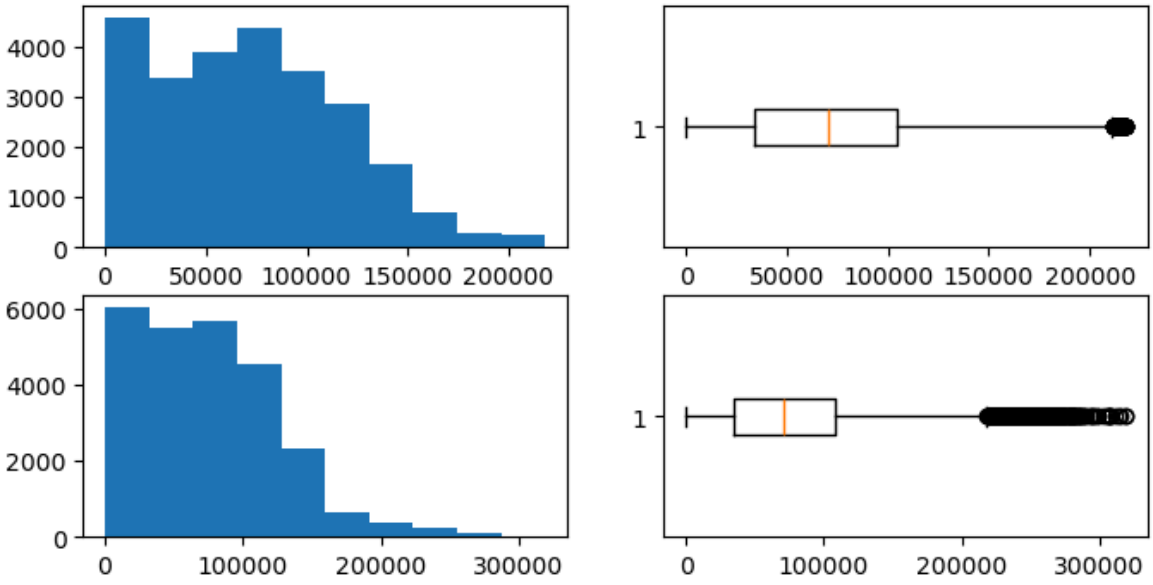
In [127…

```
plt.figure(figsize=(8,4))
plt.subplot(1,2,1).hist(visa_df['prevailing_wage1'])
plt.subplot(1,2,2).boxplot(visa_df['prevailing_wage1'],vert=False)
plt.show()
```



In [129…

```
plt.figure(figsize=(8,4))
plt.subplot(2,2,1).hist(visa_df['prevailing_wage1'])
plt.subplot(2,2,2).boxplot(visa_df['prevailing_wage1'],vert=False)
plt.subplot(2,2,3).hist(wage_data)
plt.subplot(2,2,4).boxplot(wage_data,vert=False)
plt.show()
```



In [133…

```
wage_data=visa_df['prevailing_wage']
q1=np.percentile(wage_data,25)
q2=np.percentile(wage_data,50)
q3=np.percentile(wage_data,75)
IQR=q3-q1
UB=q3+1.5*IQR
LB=q1-1.5*IQR
visa_df['prevailing_wage2']=wage_data.clip(lower=LB,upper=UB)
visa_df
```

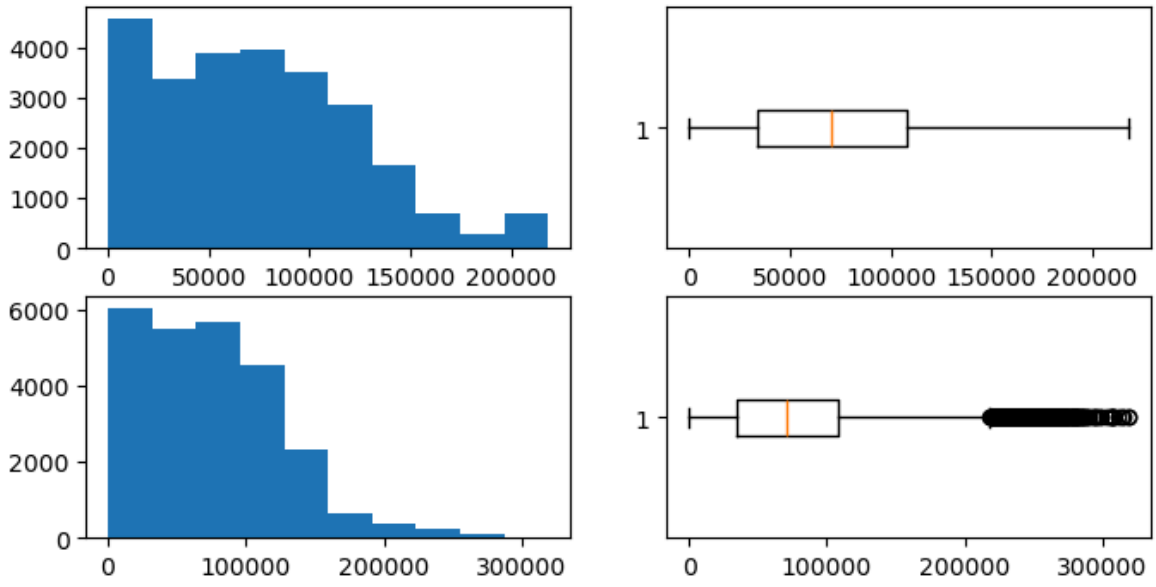
Out[133…

	case_id	continent	education_of_employee	has_job_experience	requires_job_t
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 × 14 columns

In [135…

```
plt.figure(figsize=(8,4))
plt.subplot(2,2,1).hist(visa_df['prevailing_wage2'])
plt.subplot(2,2,2).boxplot(visa_df['prevailing_wage2'],vert=False)
plt.subplot(2,2,3).hist(wage_data)
plt.subplot(2,2,4).boxplot(wage_data,vert=False)
plt.show()
```



In [ ]: