In [12]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as pt
import seaborn as sns
```

In [13]:
```
visadf=pd.read_csv('C:/Users/Anuja_PC/OneDrive/Documents/dataFiles/Visadataset.c
visadf
```

Out[13]:

	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	

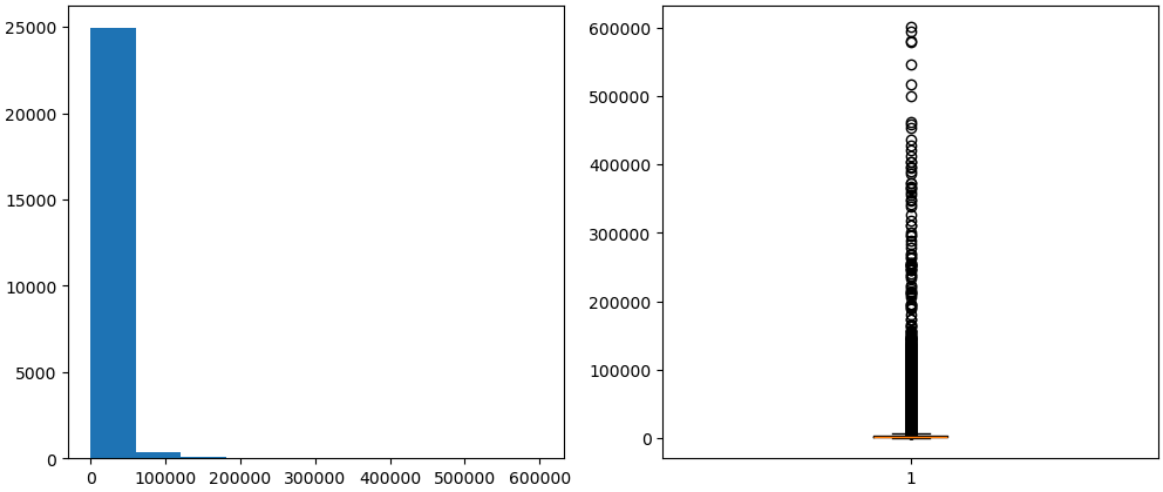
25480 rows × 12 columns

In [14]:
```
empDf = visadf["no_of_employees"]
empDf
```

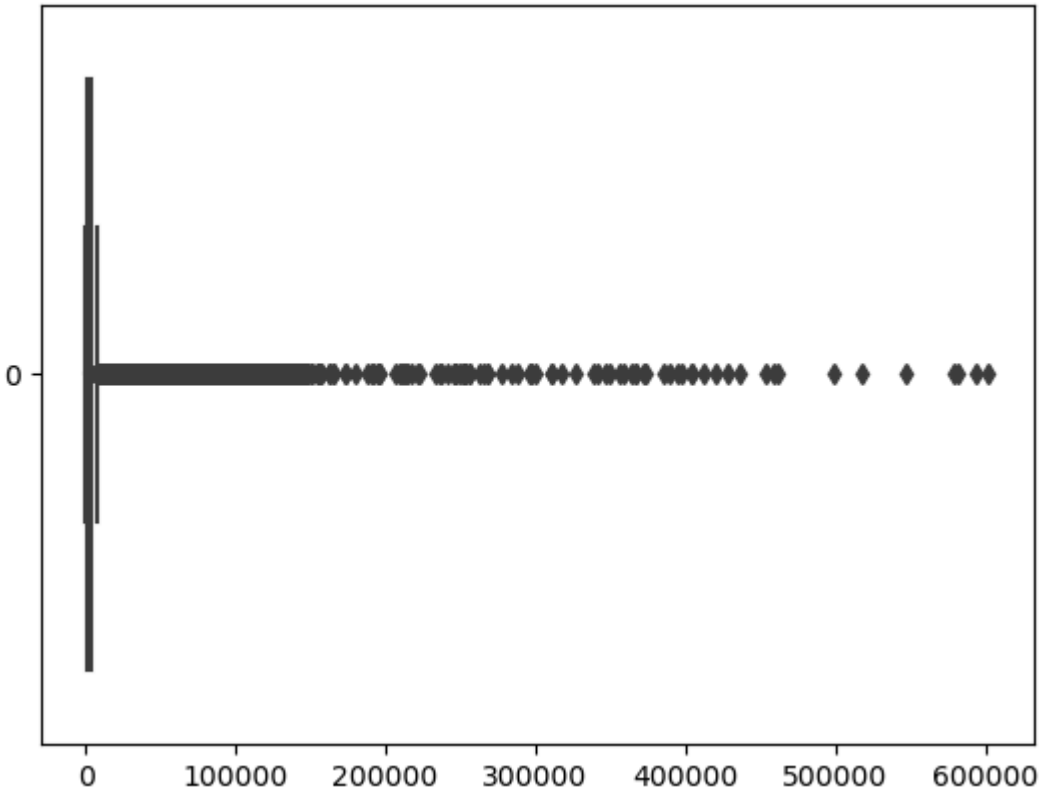
Out[14]:
```
0        14513
1         2412
2        44444
3           98
4         1082
         ...
25475     2601
25476     3274
25477     1121
25478     1918
25479     3195
Name: no_of_employees, Length: 25480, dtype: int64
```

Box Plot

In [45]:
```
pt.figure(figsize=(12,5))
pt.subplot(1,2,1).hist(empDf,bins=10)
pt.subplot(1,2,2).boxplot(empDf)
pt.show()
```



In [25]:
```
sns.boxplot(empDf,orient='h')
pt.show()
```



Finding the outliers

In [36]:
```
q1=np.quantile(empDf,0.25)
q3=np.quantile(empDf,0.75)

q1,q3
```

Out[36]: (1022.0, 3504.0)

In [37]:
```
IQR = q3-q1
IQR
```

Out[37]: 2482.0

In [38]:
```
LB = q1-1.5*IQR  #Lower bound
UB = q3+1.5*IQR  #upper bound

LB,UB
```

Out[38]: (-2701.0, 7227.0)

In [40]:
```
cond1 = empDf < LB
cond2 = empDf > UB

outliersData=visadf[cond1 | cond2]
outliersData
```

Out[40]:

	case_id	continent	education_of_employee	has_job_experience	requires_job_t
0	EZYV01	Asia	High School	N	
2	EZYV03	Asia	Bachelor's	N	
12	EZYV13	Asia	Bachelor's	Y	
14	EZYV15	Asia	Master's	Y	
16	EZYV17	Europe	Master's	Y	
...	...	...	...	...	
25441	EZYV25442	Asia	Master's	N	
25443	EZYV25444	Africa	Bachelor's	N	
25455	EZYV25456	South America	Bachelor's	N	
25464	EZYV25465	Asia	Master's	N	
25471	EZYV25472	Asia	High School	N	

1556 rows × 12 columns

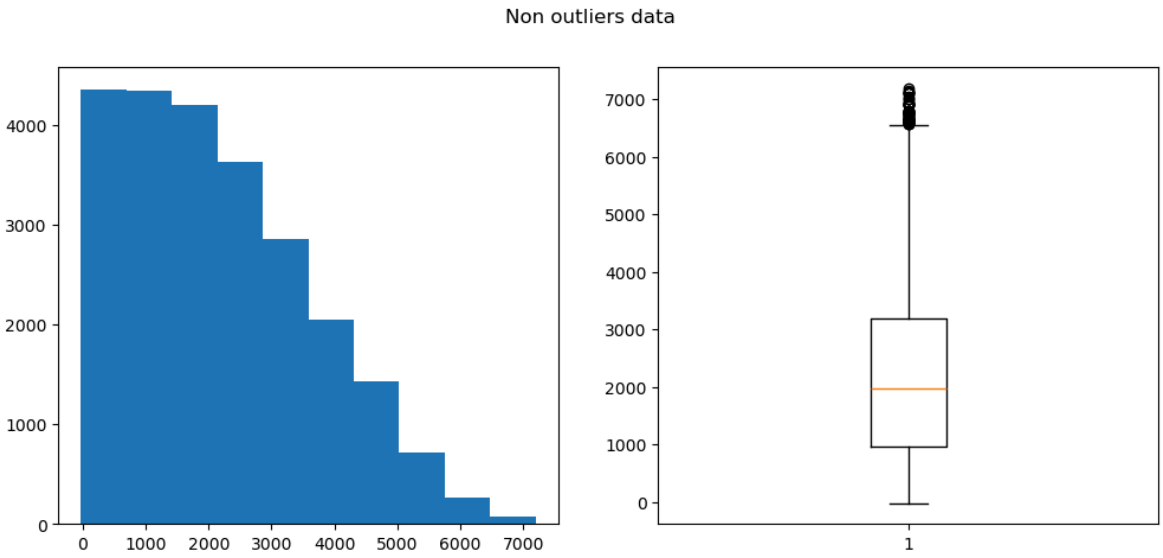
In [41]:
```
cond1 = empDf> LB
cond2 = empDf < UB

non_outliersData=visadf[cond1 & cond2]  # non outliers data
non_outliersData
```

Out[41]:

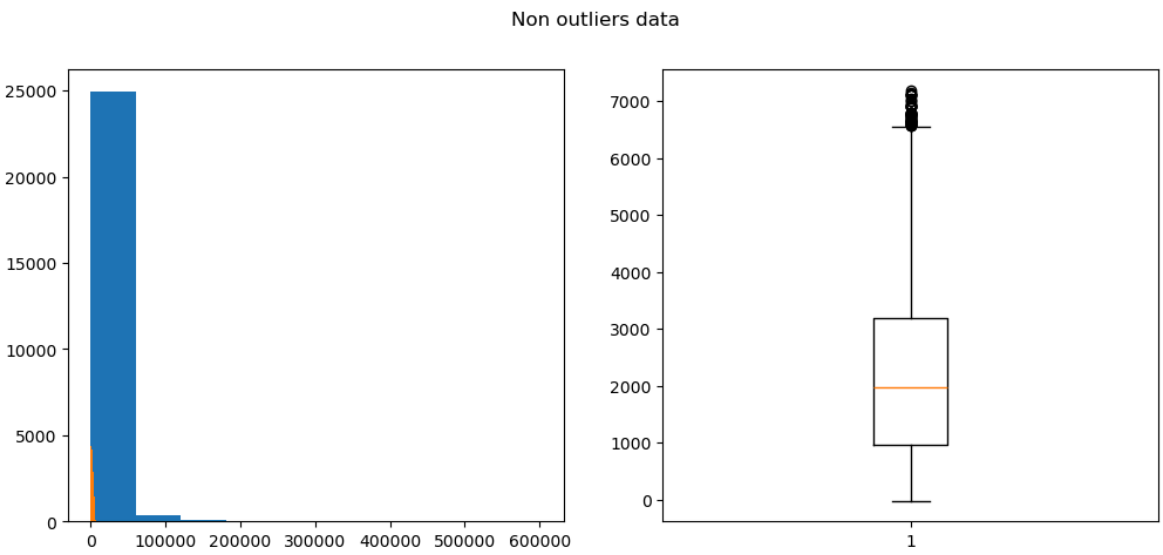
	case_id	continent	education_of_employee	has_job_experience	requires_job_t
1	EZYV02	Asia	Master's	Y	
3	EZYV04	Asia	Bachelor's	N	
4	EZYV05	Africa	Master's	Y	
5	EZYV06	Asia	Master's	Y	
6	EZYV07	Asia	Bachelor's	N	
...	...	...	...	...	
25475	EZYV25476	Asia	Bachelor's	Y	
25476	EZYV25477	Asia	High School	Y	
25477	EZYV25478	Asia	Master's	Y	
25478	EZYV25479	Asia	Master's	Y	
25479	EZYV25480	Asia	Bachelor's	Y	

23924 rows × 12 columns

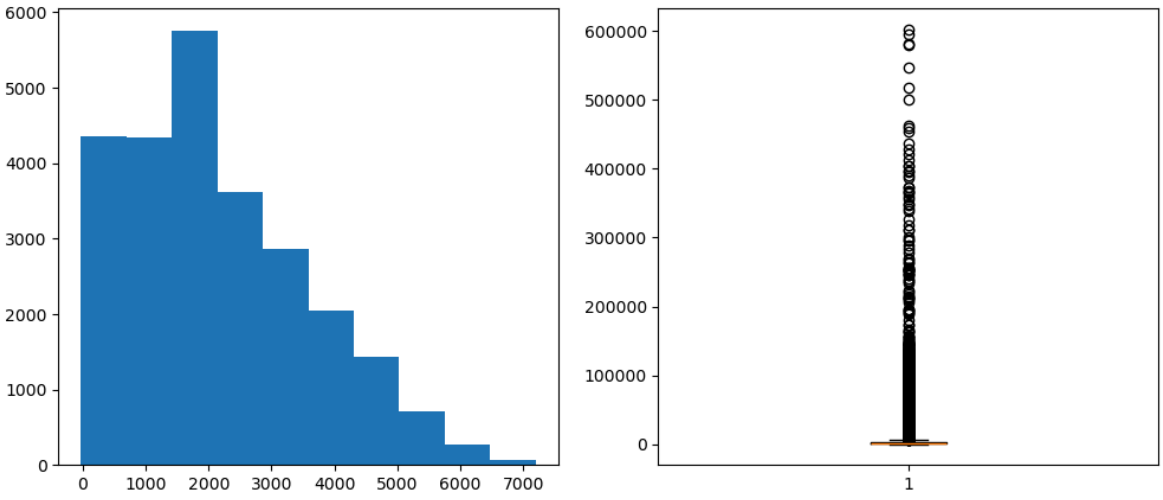
In [42]:
```
pt.figure(figsize=(12,5))
pt.suptitle("Non outliers data")
pt.subplot(1,2,1).hist(non_outliersData["no_of_employees"])
pt.subplot(1,2,2).boxplot(non_outliersData["no_of_employees"])
pt.show()
```



In [44]:
```
pt.figure(figsize=(12,5))
pt.suptitle("Non outliers data")
pt.subplot(1,2,1).hist(empDf)
pt.subplot(1,2,1).hist(non_outliersData["no_of_employees"])
pt.subplot(1,2,2).boxplot(non_outliersData["no_of_employees"])
pt.show()
```



In [47]:
```
l1= []  #create emply list
mediann=empDf.median() # calculate median
  # iterate through values
for value in empDf.values:
   if value < LB or value > UB:
        l1.append(mediann)
   else:
        l1.append(value)

empdf_copy = empDf.copy()
empdf_copy = l1

pt.figure(figsize=(12,5))
pt.subplot(1,2,1).hist(empdf_copy)
pt.subplot(1,2,2).boxplot(empDf)
pt.show()
```



In [49]:
```
con1 = empDf< LB
con2 = empDf>UB
cond = con1 | con2
mediann=empDf.median()
l=np.where(cond,mediann,empDf)
l
```

Out[49]:
```
array([2109., 2412., 2109., ..., 1121., 1918., 3195.])
```

In [ ]:

In [ ]: