In [23]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math
```

In [2]:
```
data = pd.read_csv("Lab1_data_group3.txt",delimiter="\t")
```

In [4]:
```
#1
data.describe()
```

Out[4]:

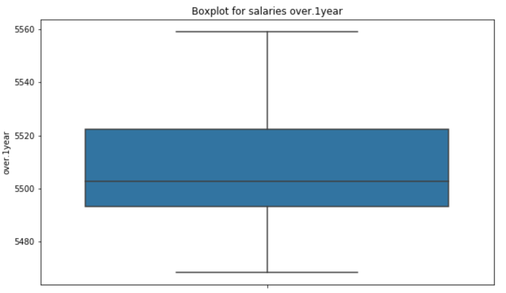
	over.1year	not.1year
count	50.000000	50.000000
mean	5505.673600	5489.336600
std	21.821165	19.668898
min	5468.110000	5454.290000
25%	5493.145000	5478.755000
50%	5502.745000	5488.380000
75%	5522.272500	5499.477500
max	5559.170000	5532.190000

over.1year: mean:5505.673600, median:5502.745000, lower quantile:5493.145000, upper quantile:5522.272500,

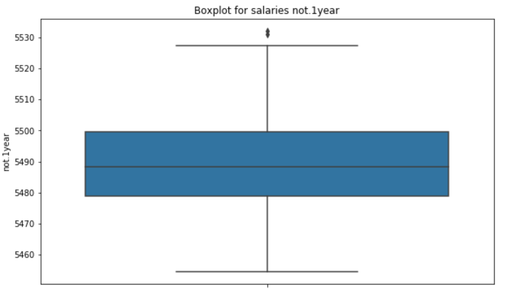
not.1year: mean:5489.336600, median:5488.380000, lower quantile:5478.755000, upper quantile:5499.477500,

Commentary: Both groups have got similar salaries. For both std is around ~ 20 and distributions of salaries is close to

In [18]:
```
#2
plt.figure(figsize=(10,6))
sns.boxplot(y=data["over.1year"])
plt.title("Boxplot for salaries over.1year")
plt.show()
```

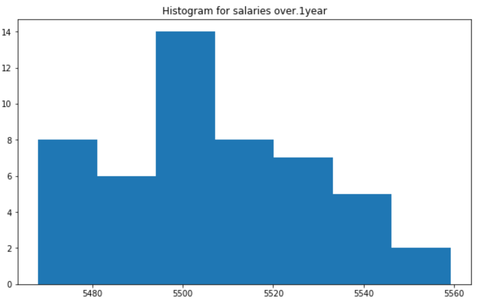


In [19]:
```
plt.figure(figsize=(10,6))
sns.boxplot(y=data["not.1year"])
plt.title("Boxplot for salaries not.1year")
plt.show()
```

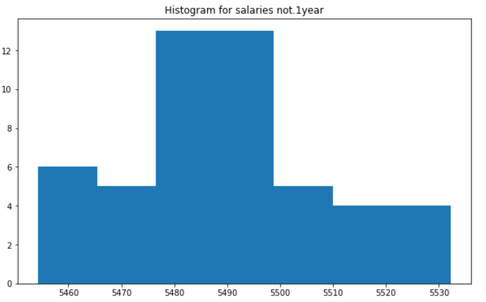


Commentary: There are visible diffrence between boxplots. In case of boxplot for salaries 'not.1year' the median is closer to the middle of boxplot and also we can see few outliers. In case for boxplot 'over.1year' the distribution is also a little bit right

In [25]:
```
#3
plt.figure(figsize=(10,6))
x = len(data["over.1year"])
binsizes = math.sqrt(x)
plt.hist(data["over.1year"], bins = int(binsizes))
plt.title('Histogram for salaries over.1year')
plt.show()
```



In [32]:
```
plt.figure(figsize=(10,6))
x = len(data["not.1year"])
binsizes = math.sqrt(x)
plt.hist(data["not.1year"], bins = int(binsizes))
plt.title('Histogram for salaries not.1year')
plt.show()
```



In case of histogram for 'not.1year' there is significantly more salaries in range 5480-5500.

In [52]:
```
#4
#Variances:
print('Variances over.1year: {}, not.1year: {}'.format(data['over.1year'].var(),data['not.1y
ear'].var()))
```

```
Variances over.1year: 476.1632480000002, not.1year: 386.8655412653049
```

In [51]:
```
#5
#Hypothesis
#H0: Average payment in the group of employees 'over.1year' is the same as in the group of
 'not.1year'.

results = stats.ttest_ind(data['over.1year'],
                          data['not.1year'])
```

```
Ttest_indResult(statistic=3.9322821289595757, pvalue=0.00015706141373177958)
```

Commentary: The p value is lower than p < 0,05, so we can reject the null hypothesis.