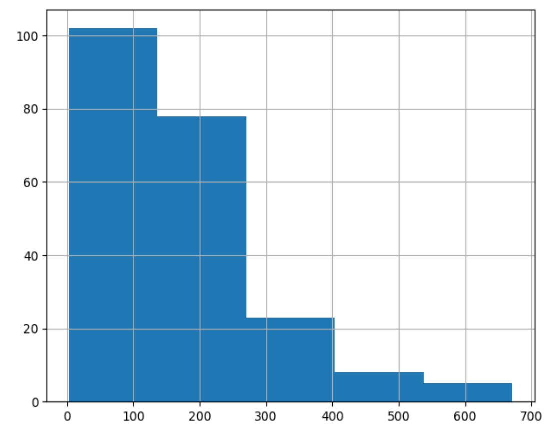
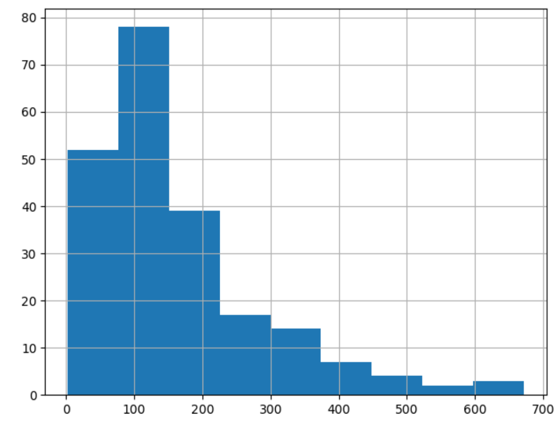
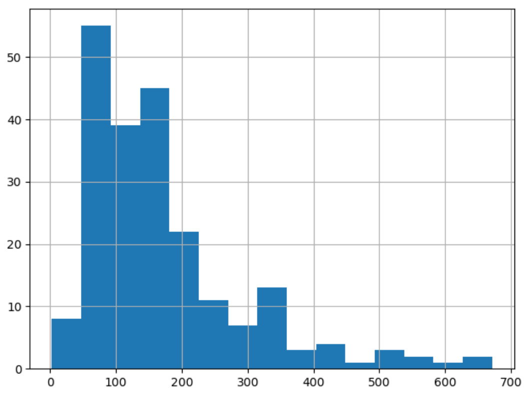
First, we use empirical formulas to estimate the bins and the result is 8~9.
To justify it, we choose 5, 9, 15, 20, 25 and 30 bins to draw 4 histograms and compare each histogram to select the most suitable one. The following is each histogram:



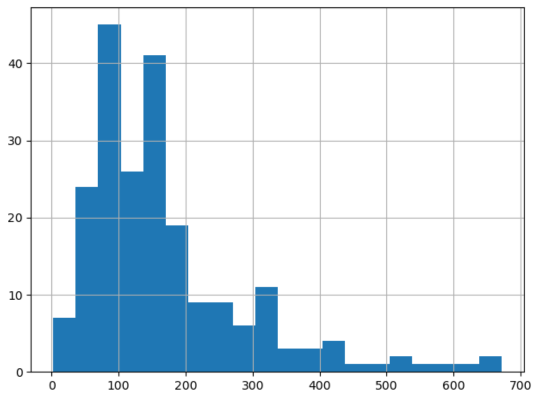
5bins



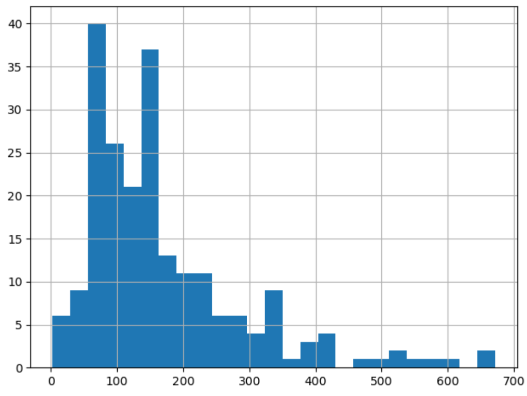
9bins



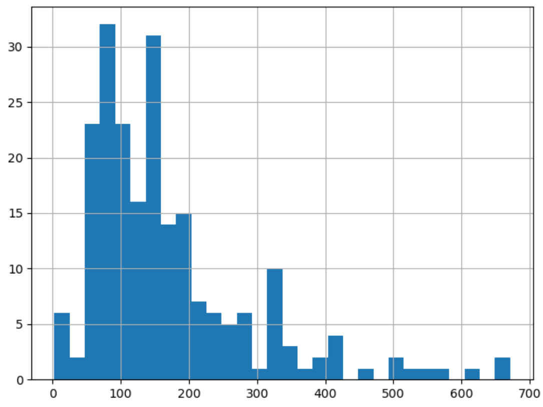
15bins



20bins



25bins



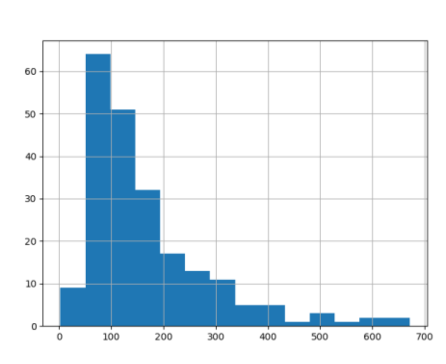
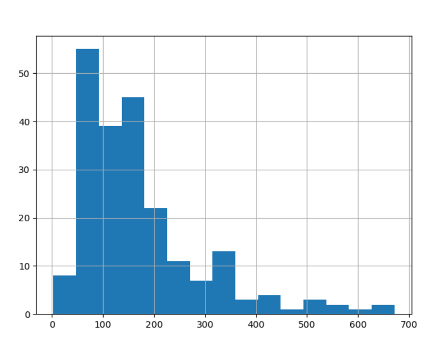
30bins

Then, we found the interval of bins may in [10,20]. This is because if the bins smaller than 10, the shape of first half part is not accurate. And it is no necessary to have bins more than 20, because the shape of 20, 25, 30 bins is similar and it will cost unnecessary loss

Second, we try to find more suitable bins in [10,20]. We think the unit of “diff” is hour, therefore, the width of bins uses multiple of 24 may be more suitable. And we found 2 numbers in [2,30] which is:

| max | min | bins |
|---|---|---|
| 671 | 3 | (max-min)//24 ≈ 27 |
| 671 | 3 | (max-min)//48 ≈ 14 |

According to the 1st step, we choose the 14 as the number of bins.
But we found the difference between 14 and 15 is large:

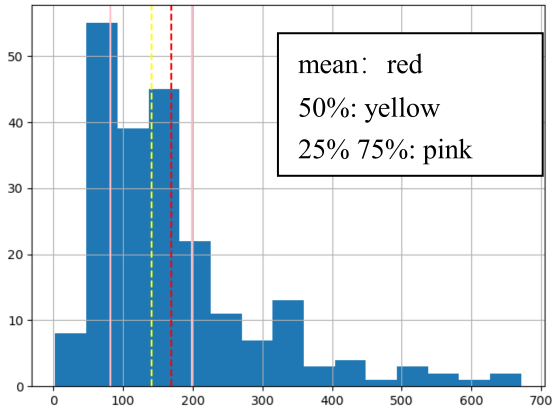
|  |  |
|---|---|
| 14bins | 15bins |

In order to make the shape more accurate(the 14 bins don’t shape the sudden increase between 100 and 200), we chose the 15 as the number of bins.

Third, we use the df.describe() to get the 8 attribute of the ‘diff’:

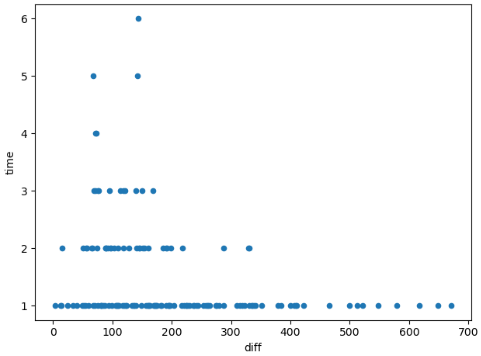
```
count       216.000000
mean          169.032407
std         122.340447
min              3.000000
25%             81.750000
50%            141.000000
75%            199.500000
max            671.000000
Name: diff, dtype: float64
```

And we mark each attribute on the histogram:



But we do not get some useful information from this histogram.

Therefore, we draw the scatter plot whose x is ‘diff’ and y is count of each ‘diff’ :



And we found the shape of this plot is similar to the histogram.
Therefore, we think the number of bins should be 15.