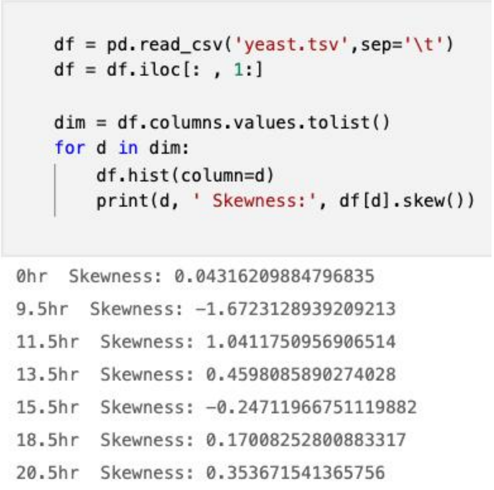
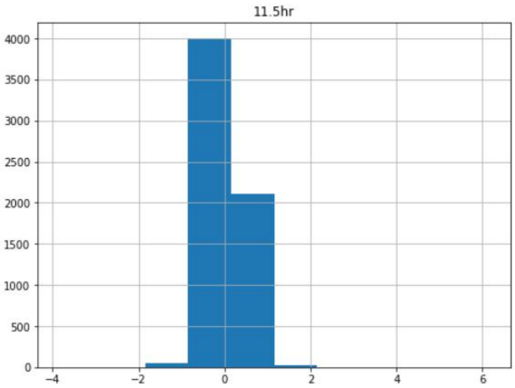
For data preprocessing: As we can see, the data is skewed.

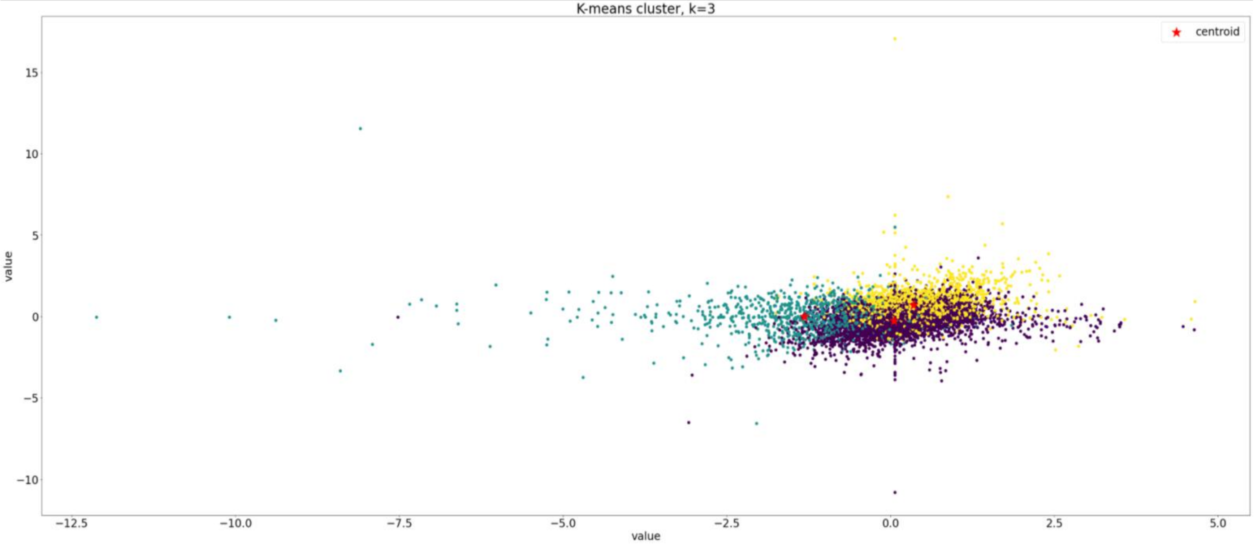


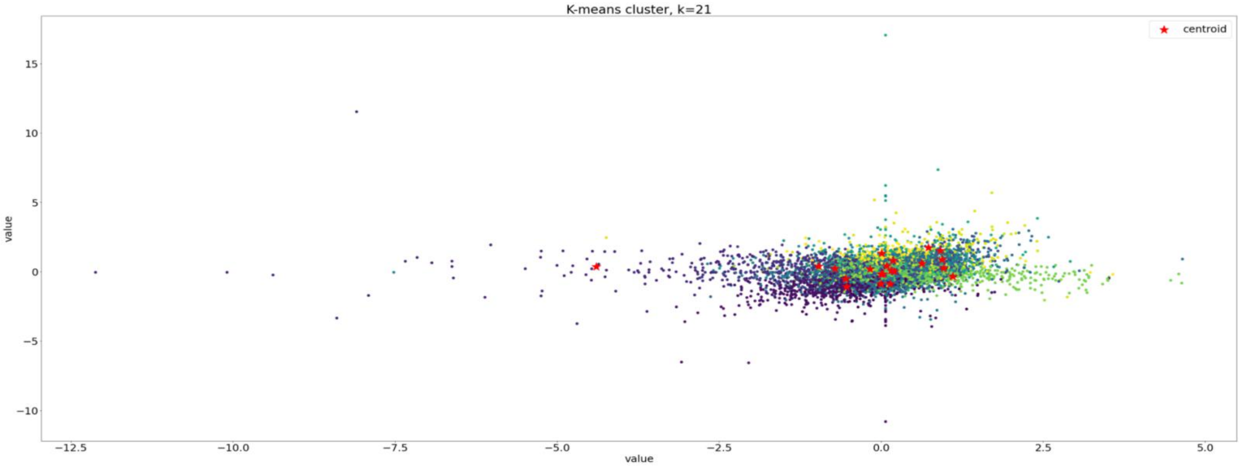
```
df = pd.read_csv('yeast.tsv',sep='\t')
df = df.iloc[: , 1:]

dim = df.columns.values.tolist()
for d in dim:
    df.hist(column=d)
    print(d, ' Skewness:', df[d].skew())
```

```
0hr  Skewness: 0.04316209884796835
9.5hr  Skewness: -1.6723128939209213
11.5hr  Skewness: 1.0411750956906514
13.5hr  Skewness: 0.4598085890274028
15.5hr  Skewness: -0.24711966751119882
18.5hr  Skewness: 0.17008252800883317
20.5hr  Skewness: 0.353671541365756
```

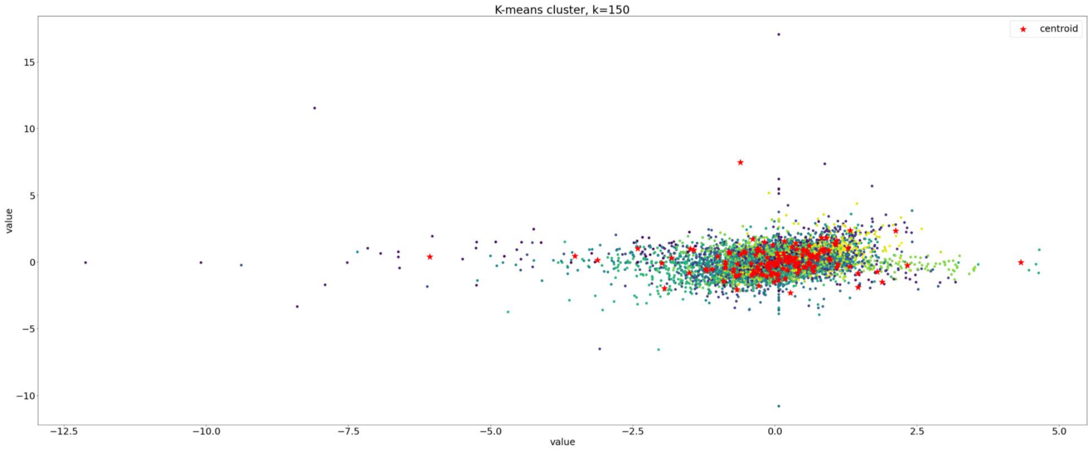


Result: as k increases, outliers tend to get a centroid and form cluster

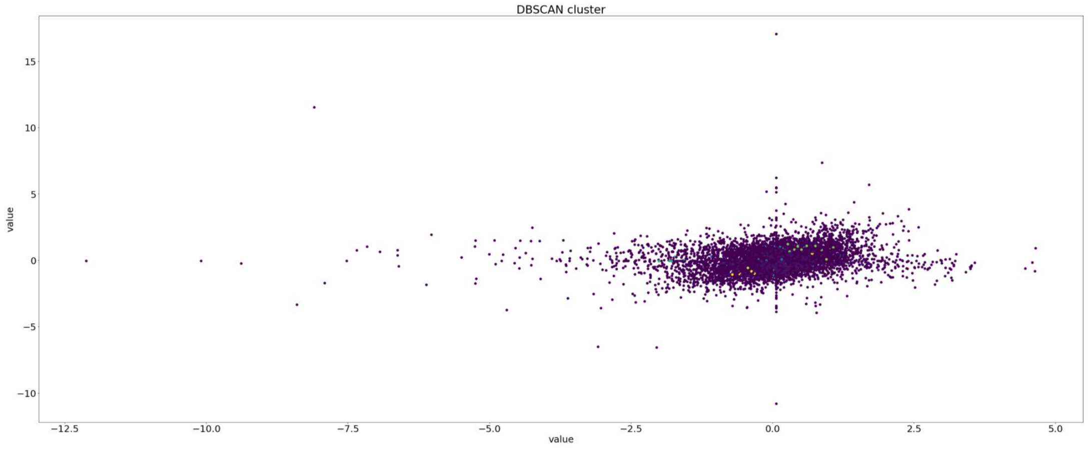




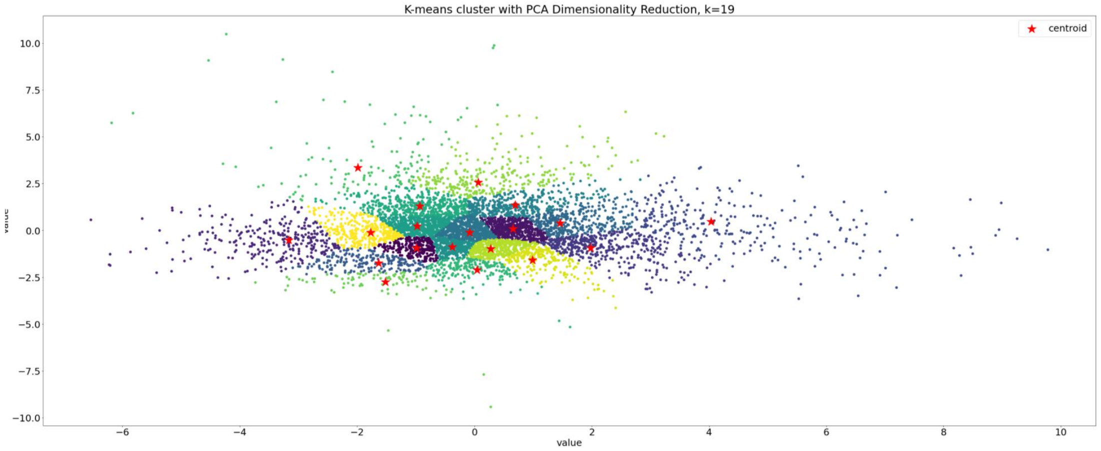
When k = 150



DBSCAN results



PCA visualization


K-means cluster with PCA Dimensionality Reduction, k=19
centroid
value
value