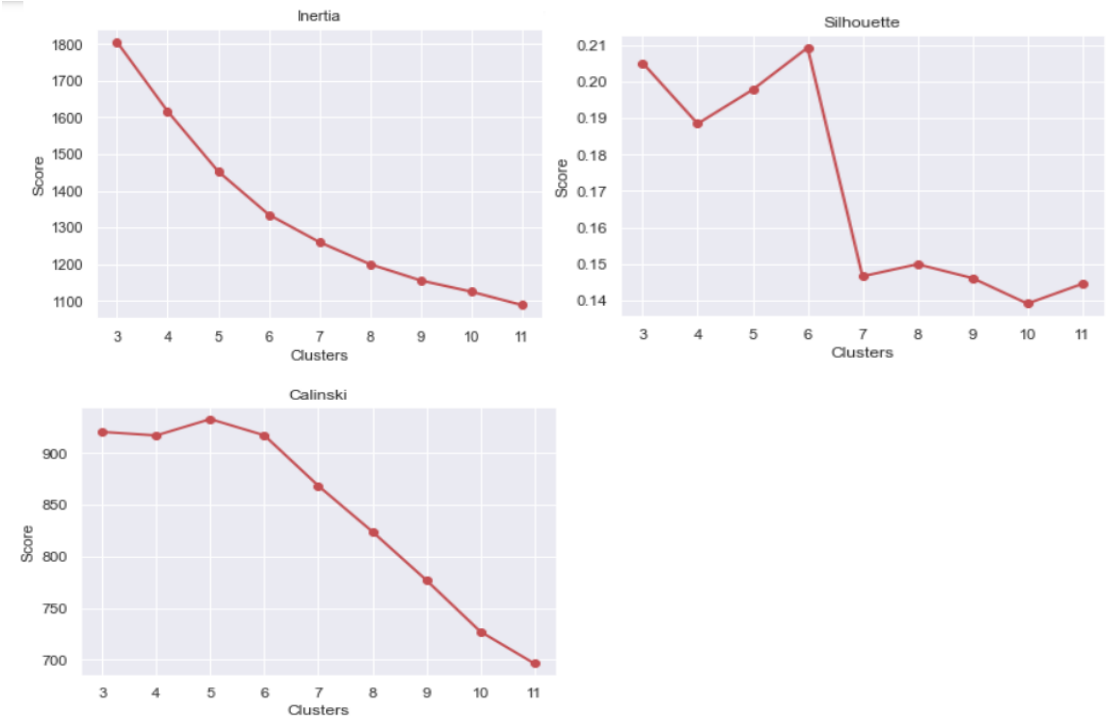
MSDS-422-SEC56 Quick Assignment KMeans Daniel Zaremba

KMeans Data Observations



Notes & Observations

After changing the random state number a couple times i.e the starting point to see if the clusters were consistent and looking at the above plots I settled on k = 6. The clusters could not be used to predict Loss Amount. Cluster 0 is the riskiest and Cluster 5 is the least risky. Cluster 0 has the highest TRUNC_IMP_DEROG (derogatory marks on their credit report) while cluster 5 has the highest TRUNC_IMP_YOJ (years on job) which makes sense.

```
CLUSTER  TARGET_BAD_FLAG
0        1                158
         0                105
1        0                571
         1                180
2        0               2238
         1                422
3        0                862
         1                117
4        1                228
         0                192
5        0                803
         1                 84
Name: TARGET_BAD_FLAG, dtype: int64
```

CLUSTER	TARGET_BAD_FLAG
0	0.600760
1	0.239680
2	0.158647
3	0.119510
4	0.542857
5	0.094701