Assignment work:

Objective: To cluster the CIBIL(credit) score parameters together using different clustering algorithms

About Dataset:

Dataset that I have, is randomly self-generated one. It contains features like 'payment history', 'credit utilization' , 'length of credit history', 'mix of credit', 'new credit'. This dataset gives an insight on how different factors contribute to credit score of an individual.

The factors that affect our credit score are

- **Payment history –** The most important factor. How regular you are on your loan/credit card payments
- **Amounts owed/Credit Utilization –** Having very high debts or maxing out credit cards with dues continuing for many months will have a negative impact on your score
- **Length of credit history –** The longer the credit history, the higher the credit score
- **Credit mix –** Having multiple types of credit like personal loan, credit card and car loan shows that you can handle different type of credit efficiently and responsibly
- **New credit –** Taking out credits within short time negatively affects your credit score.
- **Credit Score** - A Credit Score is a three-digit numeric summary of an individual's credit history. Credit bureaus have all the information on you based on credit history and provide a background about a potential borrower to the lender (Bank or NBFC). So, high credit score indicates that you have managed your credit better and this increases the chance of your loan or credit card approved and get better offers in the future. It ranges from 300-900 in case of CIBIL.

Parameters considered to calculate Credit Score for existing users

PARAMETERS	WEIGHTAGE
PAYMENT PATTERN OF BORROWER	35%
AMOUNT OF MONEY HE OWES	30%
LENGTH OF CREDIT HISTORY	15%
MIX OF CREDIT	10%
NEW CREDIT APPLICATIONS	10%

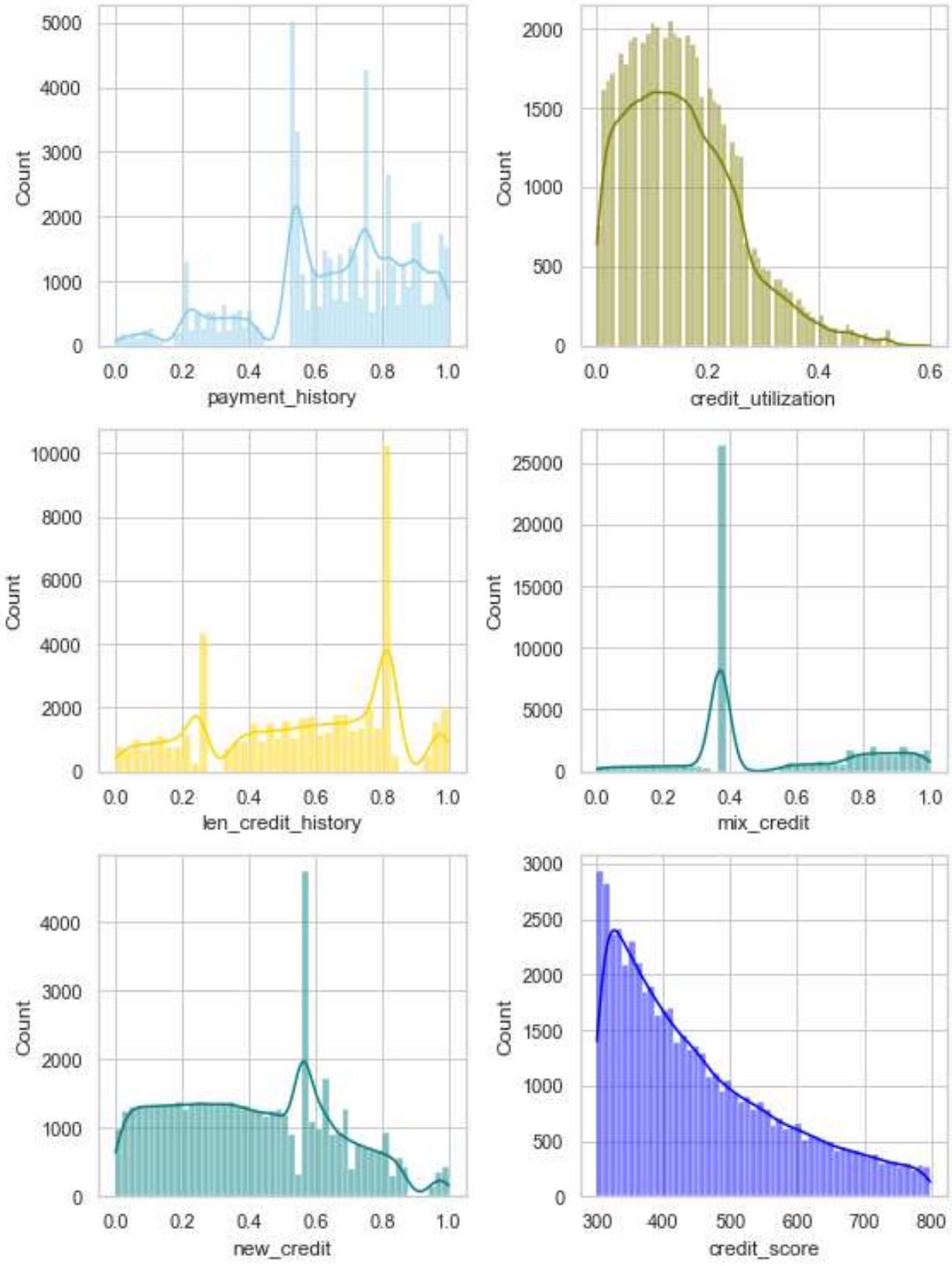
Source: [Reference Paper]

From above table, it is apparent that features with more weightage will impact credit score more than other features with lesser weightage and this can be shown through the correlation function.

	SSN	payment_history	credit_utilization	len_credit_history	mix_credit	new_credit	credit_score
SSN	1.000000	0.001720	0.001538	0.002458	-0.006638	-0.004953	0.001631
payment_history	0.001720	1.000000	0.344300	-0.092689	-0.078383	0.224179	0.075179
credit_utilization	0.001538	0.344300	1.000000	0.150759	0.091168	-0.454983	-0.424182
len_credit_history	0.002458	-0.092689	0.150759	1.000000	-0.023483	0.096539	0.036930
mix_credit	-0.006638	-0.078383	0.091168	-0.023483	1.000000	0.049136	0.022490
new_credit	-0.004953	0.224179	-0.454983	0.096539	0.049136	1.000000	-0.252627
credit_score	0.001631	0.075179	-0.424182	0.036930	0.022490	-0.252627	1.000000

The above figure describes the correlation function of features.

EDA of features:



Hypothesis Testing:

1.Null Hypothesis: loan/credit card payments does not impacts credit score

```
       credit_score  payment_hist
13908           542             1
37325           497             1
11471           538             1
15114           413             1
29241           349             1
p-value:  0.000255
Reject Null Hypothesis : Significant
```

2. Null Hypothesis: Having very high debts or maxing out credit cards with dues continuing for many months will have a negative impact on your score.

```
       credit_score  credit_utilisation
1269            545                   0
1917            505                   0
39746           454                   0
55645           344                   0
34096           587                   0
0.937466
Do not reject Null Hypothesis : Not Significant
```

3. Null Hypothesis: Taking out credits within short time negatively affects your credit score

```
       credit_score  new_cred
1269            545         0
1917            505         1
39746           454         1
55645           344         1
34096           587         1
0.937466
Do not reject Null Hypothesis : Not Significant
```

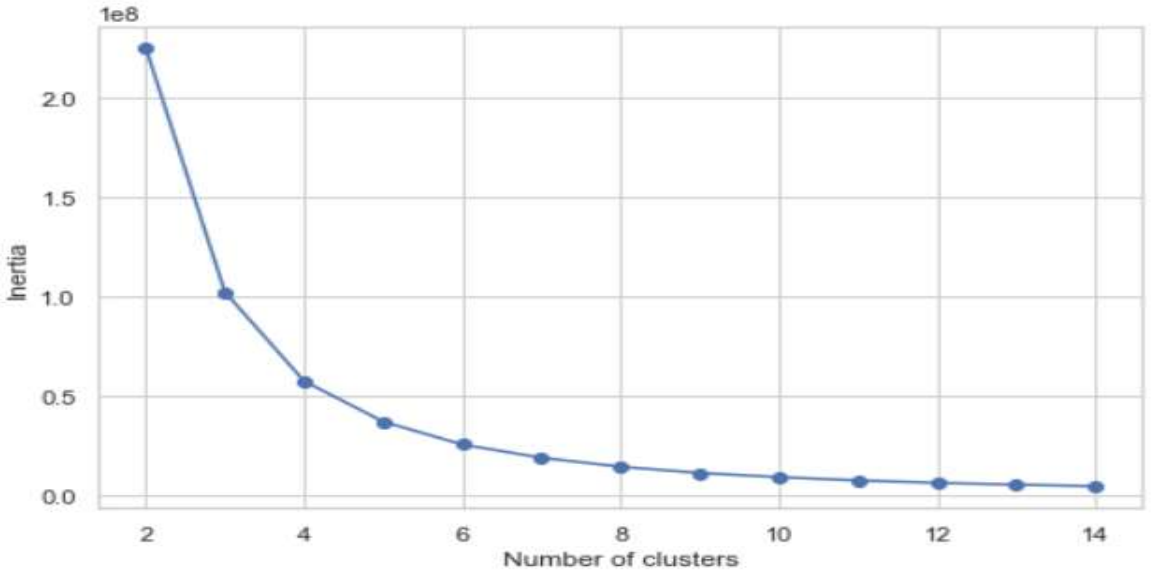
4.Null Hypothesis: length of credit history doesnt impacts the credit score

```
       credit_score  len_credit_history
13908           542                0.82
37325           497                0.25
11471           538                0.82
15114           413                0.96
29241           349                0.12
p-value:  0.062443
Do not reject Null Hypothesis : Not Significant
```

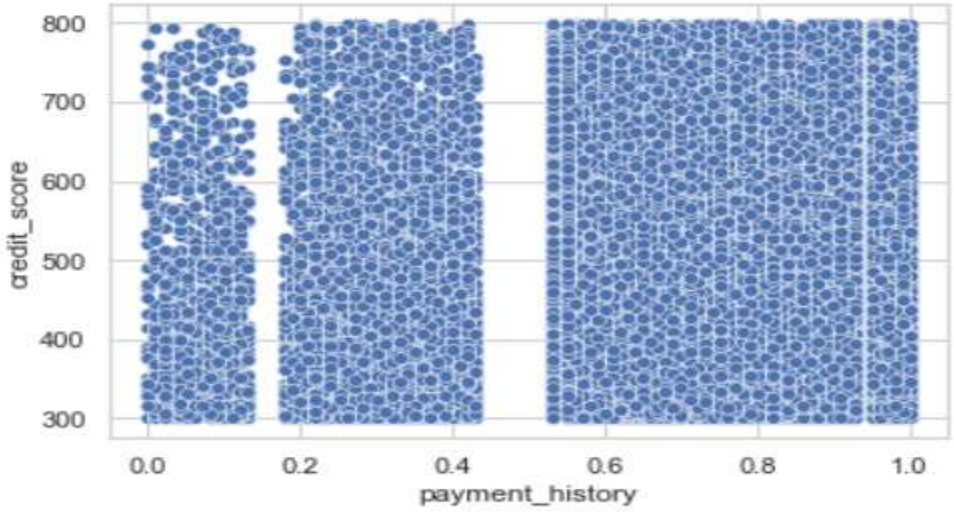
Clustering:

Here I have used K-Means algorithm to do clustering.

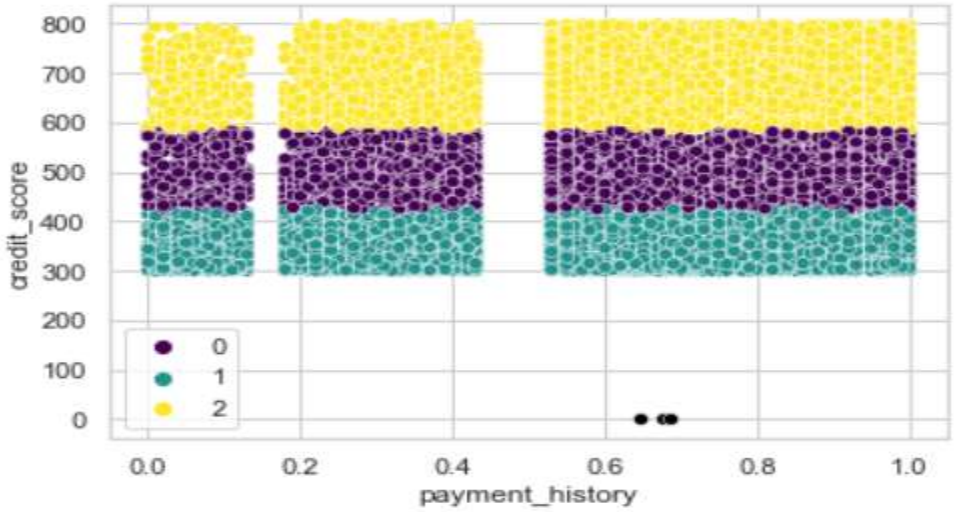
(Note: I could have used other algorithm but since I am working with random generated numbers so at the end they will perform in the same way as k-means algorithm is doing.)



```
Silhouette score for 2 clusters k-means : 0.649
Silhouette score for 3 clusters k-means : 0.602
Silhouette score for 4 clusters k-means : 0.582
Silhouette score for 5 clusters k-means : 0.566
Silhouette score for 6 clusters k-means : 0.559
Silhouette score for 7 clusters k-means : 0.552
```




```
[0 0 0 ... 1 1 0]
```



0,1,2 refer to different cluster in figure.