Machine Learning Capstone Project

UnSupervised Learning : Because the labels are not known at first.Only input data is present.

Type of Machine Learning: Clustering

Dataset : https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation

1.Data Collection :

In this phase data is collected and basic information about the dataset is studied using

dataset.Info() - this gives range index,columns, and memory usage.

Dataset.describe()- gives the mean, median, mode and percentile values .

Dataset.isna().sum()- gives the count of null values of every columns.

2.Data Preprocessing:

In this phase, data is cleaned and preprocessed to make it usable for the model so that the accuracy will be good.

1.Handling TransactionTime & TransactionAmount (INR) Column.

Transaction amount null value will not help in clustering the customers. Hence dropping the row with null value in TransactionAmount (INR) column.

The transaction time column had value in the format 16:45:36.

Sliced this value to have only Hour and minutes in a new column 'HourMinute'.

Sliced it to have new columns 'Hour' & 'Minutes'

Also binned the time of the day with this Hour column to a new column 'Time of the day' which holds value like 'Morning/Afternoon/Night'

2. Handling CustAccountBalance Column

The missing values of this column is replaced with mean value.

3.Handling CustGender Column & CustLocation

Replacing the custgender and CustLocation column with the mode value

4.Handling CustomerDOB column

This column has value in the format 1994-01-25. The age is calulated based on today's date from this value and 'Age' column is created.

The null values are replaced with median values in age.

The rows with Age in negative is dropped.

The rows with Age above 110 which is not valid is also dropped.

3.1.Univariate Analysis:

1.Separation of quantitative and qualitative columns

```
In [3]:  quan

Out[3]: ['CustAccountBalance', 'TransactionAmount (INR)', 'Hour', 'Minute', 'Age']

In [4]:  qual

Out[4]: ['CustomerDOB',
         'CustGender',
         'CustLocation',
         'TransactionDate',
         'TransactionTime',
         'TimePeriod',
         'HourMinute']
```

2.Calculating Measures of Central Tendency

Out[6]:

	CustAccountBalance	TransactionAmount (INR)	Hour	Minute	Age
Mean	79061.1	1290.27	10.4375	29.1234	37.3111
Median	14842.5	398	10	29	37
Mode	115537	100	10	20	35
Q1:25th	4181.37	146	5	14	33
Q2:50th	14842.5	398	10	29	37
Q3:75th	47929.8	1007	15	44	41
Q:99th	1.08432e+06	15418.7	23	58	49
Q4:100th	2.79796e+07	1.56003e+06	23	59	50
skewness	24.5017	85.6969	0.253728	0.00559909	0.407808
kurtosis	1043.52	17637.9	-0.92626	-1.19774	-0.291623
variance	1.29759e+11	3.20802e+07	41.7952	299.622	26.8742
std_deviation	360220	5663.94	6.46492	17.3096	5.18403

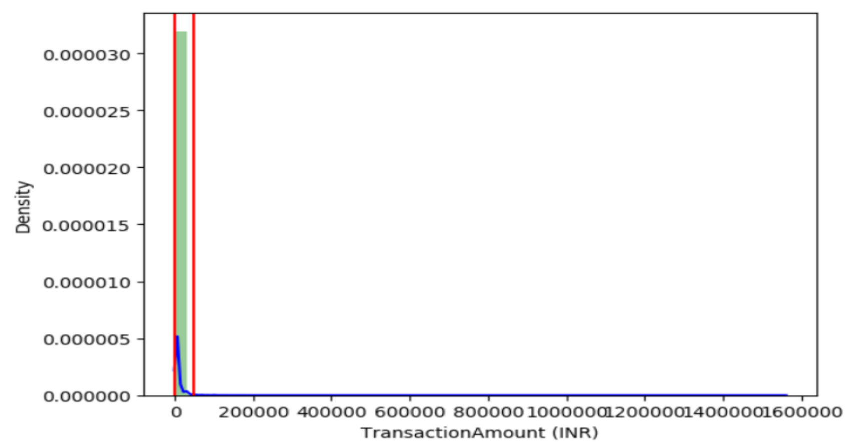
3.Calculating Interquartile range and Outliers

Out[10]:

	CustAccountBalance	TransactionAmount (INR)	Hour	Minute	Age
Q1:25th	4181.37	146	5	14	33
Q2:50th	14842.5	398	10	29	37
Q3:75th	47929.8	1007	15	44	41
Q4:100th	2.79796e+07	1.56003e+06	23	59	50
IQR	43748.4	861	10	30	8
1.5rule	65622.6	1291.5	15	45	12
lesser_outlier	-61441.2	-1145.5	-10	-31	21
greater_outlier	113552	2298.5	30	89	53
min	0	0	0	0	0
max	2.79796e+07	1.56003e+06	23	59	50

4.Probability Density Function

Out[17]: 0.5204715180919586



5.Frequency Table

```
Out[22]:
```

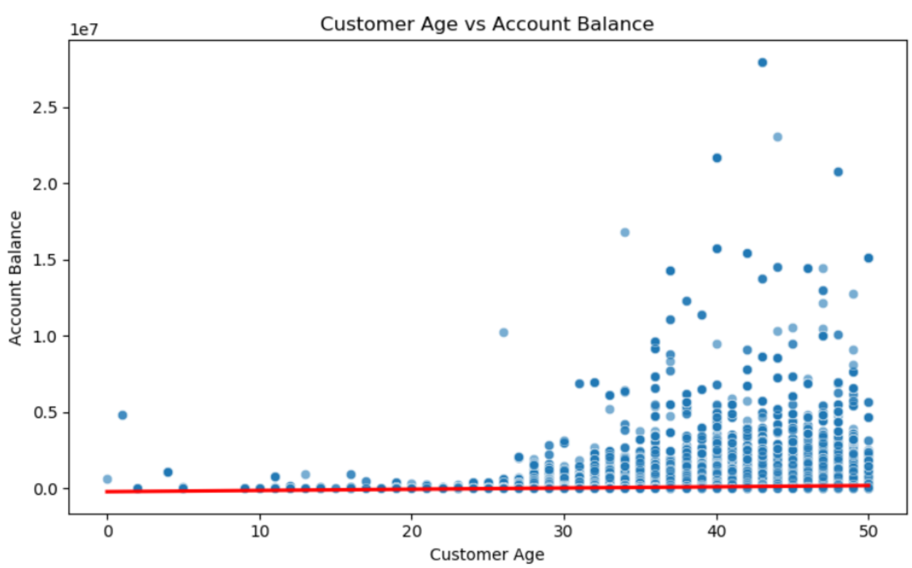
	Unique_Values	Frequency	Relative_Frequency	Cumulative_Frequency
0	115537.260095	1568	0.013232	1.781905e+04
1	0.000000	1122	0.009468	3.569349e+04
2	45856.240000	393	0.003316	4.240792e+04
3	10238.630000	354	0.002987	4.338138e+04
4	25256.280000	262	0.002211	1.384569e+05
...	...	...	...	...
118495	29305.100000	1	0.000008	9.150146e+09
118496	2456.190000	1	0.000008	9.150252e+09
118497	23350.160000	1	0.000008	9.150326e+09
118498	863.250000	1	0.000008	9.150346e+09
118499	51800.630000	1	0.000008	9.150346e+09

118500 rows × 4 columns

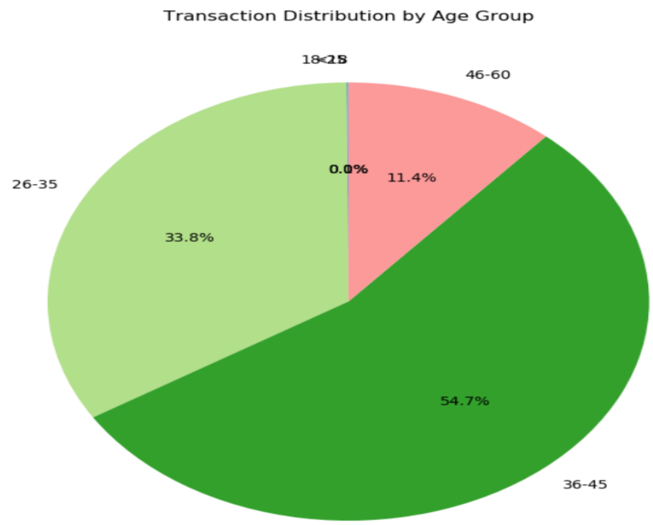
3.2. Bivariate Analysis:

1.Is there a correlation between Customer Age and Account Balance?

```
Correlation between Age and Account Balance: 0.12
```



2.What age group has done most transactions?

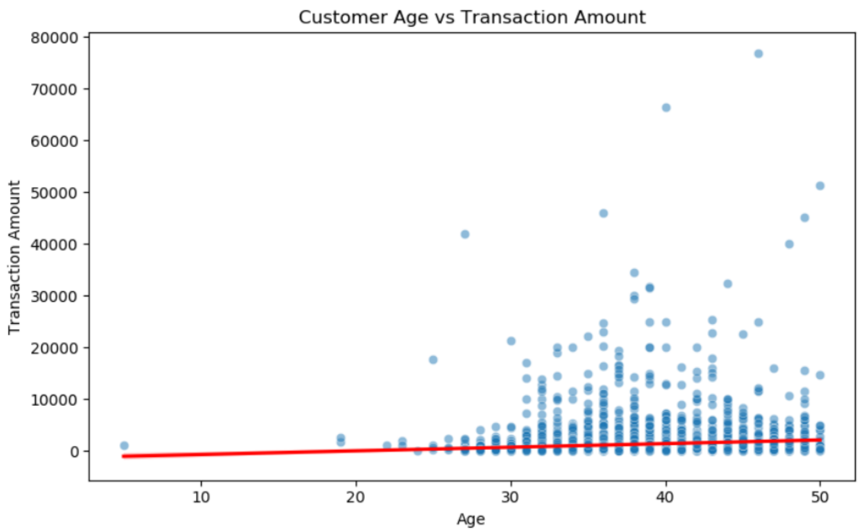


3. Does Transaction Amount differ by CustGender?

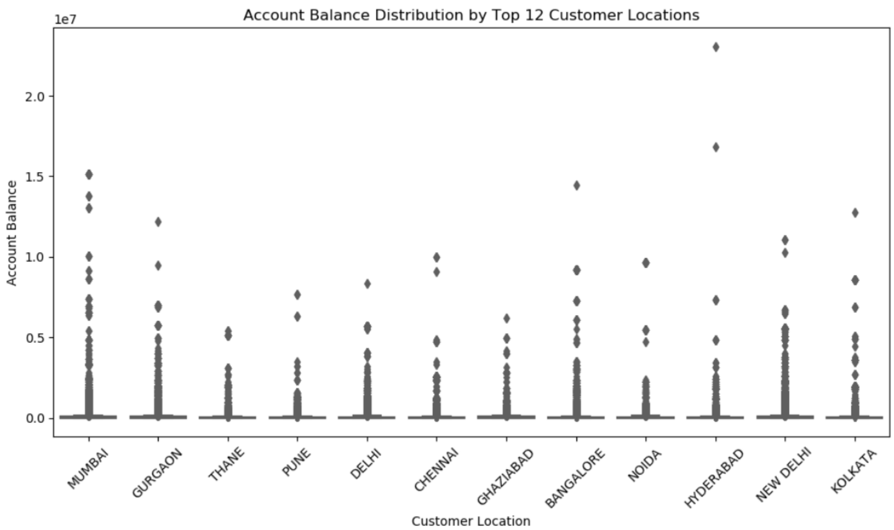
```
CustGender
F           1380002.88
M           1560034.99
```

4.Is there a relationship between Customer Age and Transaction Amount?

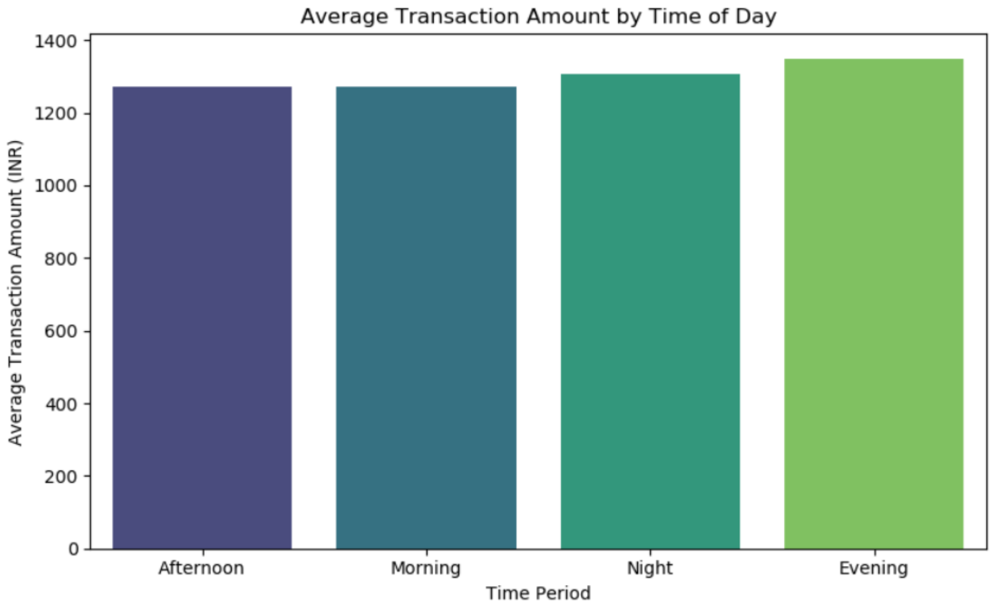
```
Correlation between Age and Transaction Amount: 0.07
```



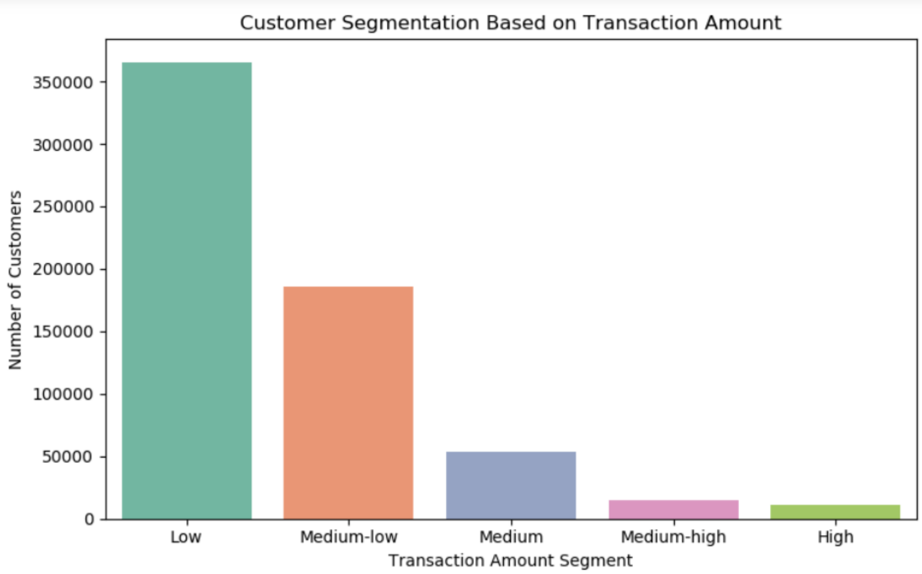
5.How does CustAccountBalance vary across different CustLocations?



6.Does the time of day (TransactionTime) affect average Transaction Amount?

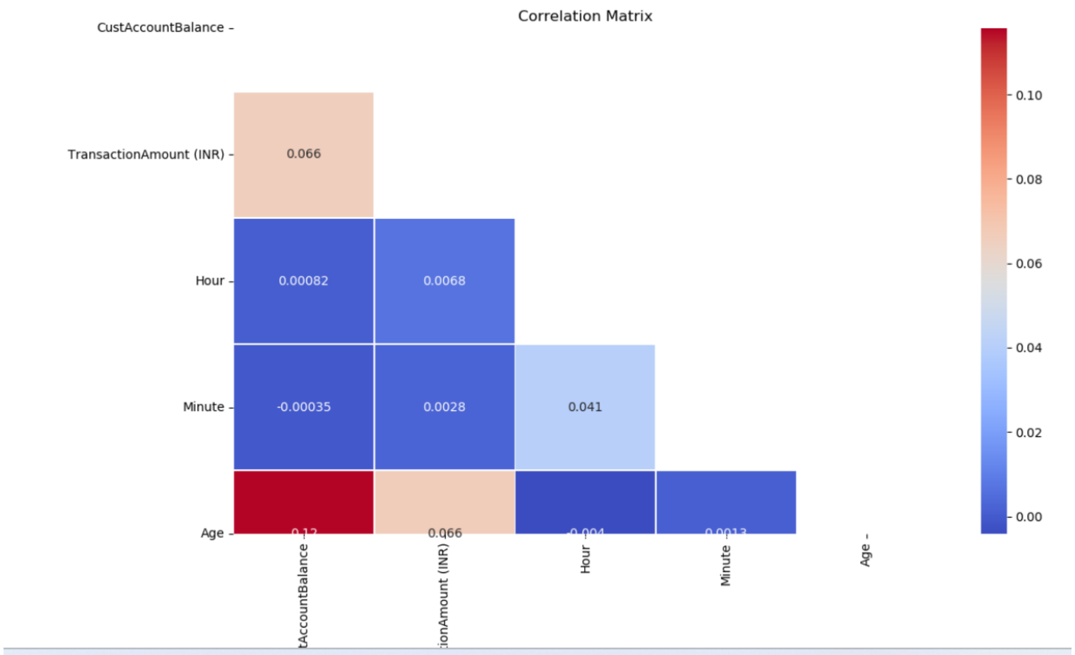


7.Give to 5 customer segmentation with respect to transaction amount?



4.1.Feature Selection :

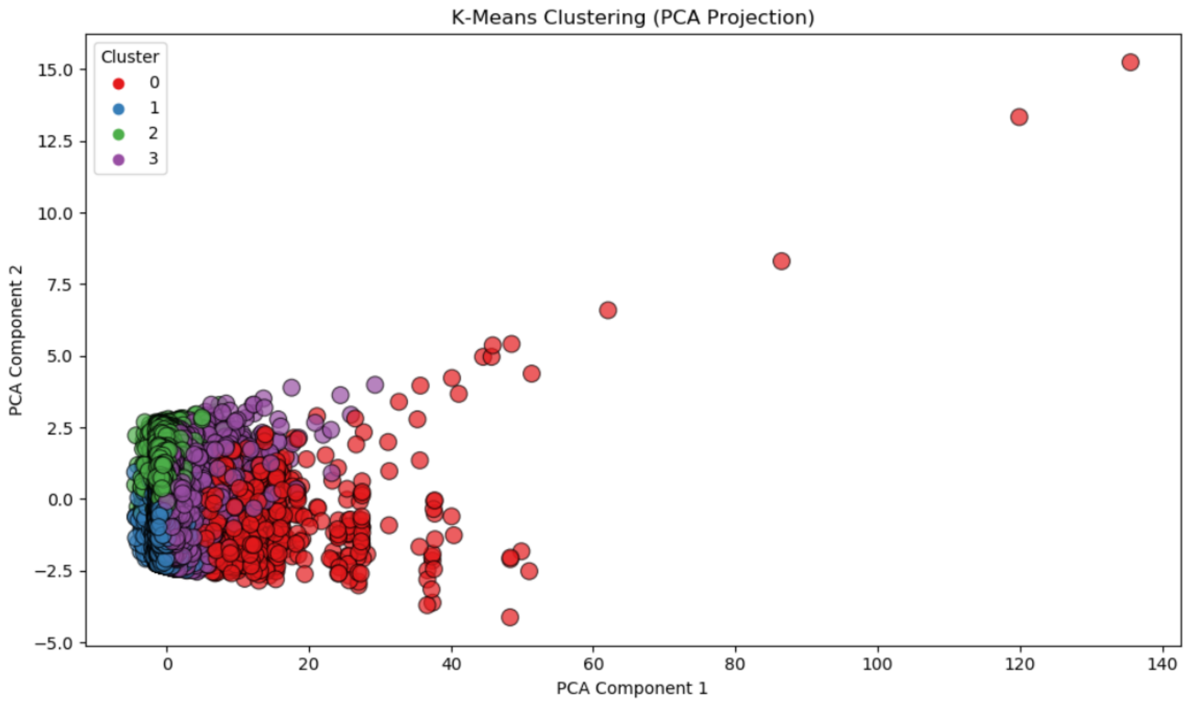
Here we are using Feature selection using Correelation



The result Series([], dtype: float64) indicates that there are no highly correlated features with a correlation above the threshold you set (0.9). This means that your features are not highly correlated with each other, which is good because it ensures that there is no multicollinearity in your data for clustering.

4.2.Model Creation :

We are using KNN means algorithm for clustering.




```
Cluster 0:
  • Transaction Amount Range: ₹0 - ₹1560035
  • Customer Age: Avg ≈ 42.6 years (Range: 1-50)
  • Dominant Location: MUMBAI
------------------------------------------------------------
Cluster 1:
  • Transaction Amount Range: ₹0 - ₹100742
  • Customer Age: Avg ≈ 34.9 years (Range: 2-43)
  • Dominant Location: MUMBAI
------------------------------------------------------------
Cluster 2:
  • Transaction Amount Range: ₹0 - ₹100000
  • Customer Age: Avg ≈ 34.9 years (Range: 0-43)
  • Dominant Location: MUMBAI
------------------------------------------------------------
Cluster 3:
  • Transaction Amount Range: ₹0 - ₹346004
  • Customer Age: Avg ≈ 44.1 years (Range: 30-50)
  • Dominant Location: MUMBAI
------------------------------------------------------------
```

4.3.Model Evaluation :

```
✅ Silhouette Score        : 0.2256 (Higher is better)
✅ Davies-Bouldin Score    : 1.2541 (Lower is better)
✅ Calinski-Harabasz Score : 116970.68 (Higher is better)
```

5.Model Deployment :

User Input

```
# ----------------------------------------
# Step 1: Define user input
# ----------------------------------------
user_input = {
    'TransactionAmount (INR)': 5000,
    'TransactionTime': '1994-10-01 15:45:00',
    'CustAccountBalance': 100000,
    'CustGender': 'Male',
    'CustLocation': 'Mumbai',
    'CustomerDOB': '1985-07-15',
    'TransactionDate': '2025-07-20',
}
```

Preprocess user input

```
# ----------------------------------------
# Step 2: Preprocess user input
# ----------------------------------------
user_input['TransactionTime'] = pd.to_datetime(user_input['TransactionTime'])
user_input['Hour'] = user_input['TransactionTime'].hour
user_input['Minute'] = user_input['TransactionTime'].minute

def map_time_period(hour):
    if 5 <= hour < 12:
        return 'Morning'
    elif 12 <= hour < 17:
        return 'Afternoon'
    elif 17 <= hour < 21:
        return 'Evening'
    else:
        return 'Night'

user_input['TimePeriod'] = map_time_period(user_input['Hour'])
user_input['HourMinute'] = user_input['TransactionTime'].strftime('%H:%M')
user_input['CustomerDOB'] = pd.to_datetime(user_input['CustomerDOB'])
today = pd.Timestamp.today()
user_input['Age'] = (today - user_input['CustomerDOB']).days // 365

user_df = pd.DataFrame([user_input])

# Select same features used in training
features = ['TransactionAmount (INR)', 'CustAccountBalance', 'Age', 'Hour', 'Minute']
user_input_scaled = scaler.transform(user_df[features])
```

Predicted cluster

```
# ----------------------------------------
# Step 3: Predict user cluster
# ----------------------------------------
predicted_cluster = kmeans.predict(user_input_scaled)[0]
print(f"\n🔍 Predicted Cluster: {predicted_cluster}\n")
```