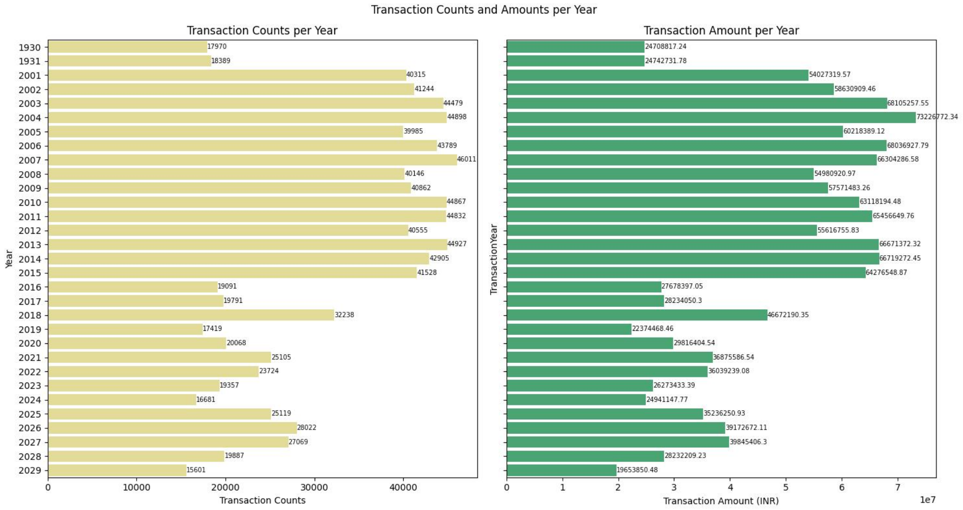
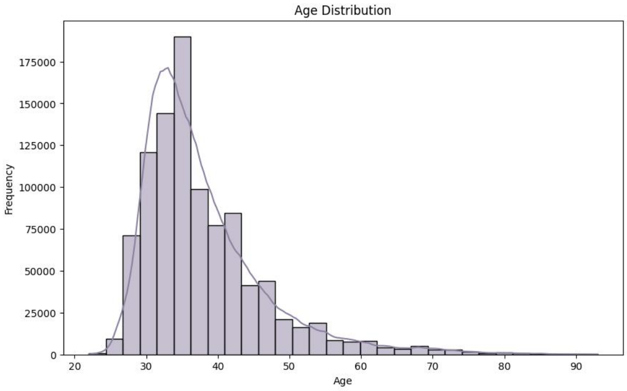
Q6 readme

a. explore the data table using visualization techniques

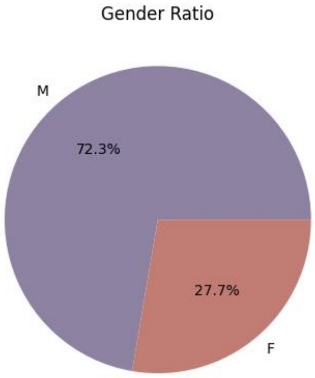
1) The transaction counts and transaction amounts per year show overall consistency, with a higher number observed between 2001 and 2015.



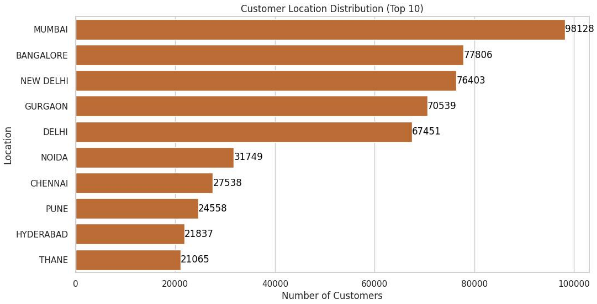
2) The Age Distribution analysis reveals that the largest transaction frequency occurs within the 30-40 age group, making them the primary target demographic for banking services. Personalized marketing strategies can optimize customer engagement and satisfaction.



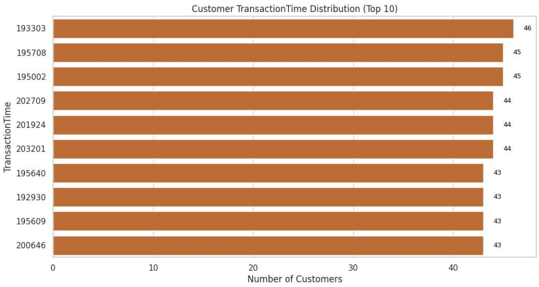
3) The Gender Ratio reveals a higher proportion of male customers conducting banking transactions. To maximize market potential, it is important to retain the male market while also tapping into the untapped female market through targeted marketing and tailored offerings.



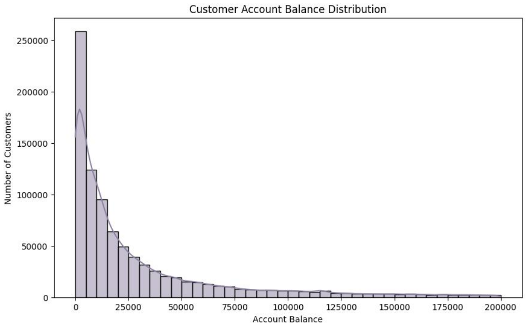
4) The analysis of Customer Location Distribution indicates that Mumbai, Bangalore, and New Delhi are the key markets for banking services. To meet the demand in these areas and enhance customer experience, it is advisable to allocate additional staff members to ensure efficient service delivery, reduce wait times, and provide personalized assistance to customers, ultimately fostering customer satisfaction and loyalty.



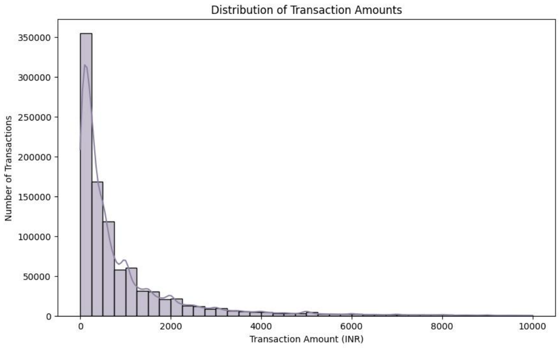
5) Customer Transaction Time Distribution reveals the peak transaction times. The top transaction times are 19:33:03, 19:57:08, and 19:50:02. By aligning staff schedules with these peak transaction times, banks can ensure adequate staffing levels to handle customer inquiries, provide prompt assistance.



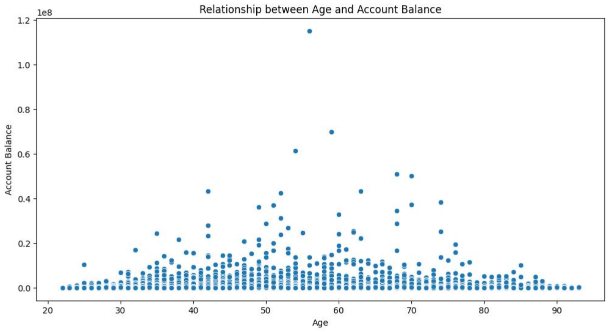
6) The Customer Account Balance Distribution shows that most customers have limited economic resources, with balances below 20,000. Banks are supposed to provide affordable and tailored financial solutions to meet the customers' limited economic capacity, fostering customer loyalty and financial well-being.



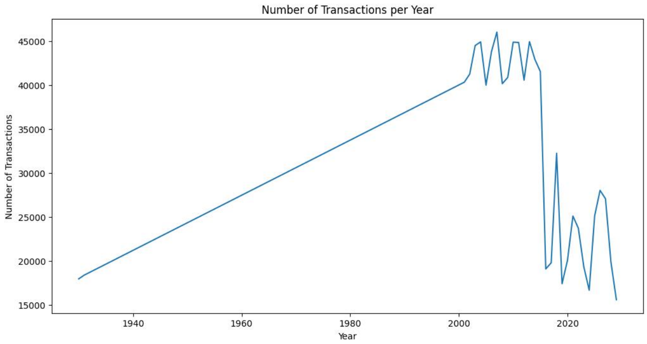
7) The analysis of Transaction Amount Distribution shows that the majority of transactions are below 2,000. This highlights the importance of providing tailored financial solutions for customers engaged in smaller value transactions, fostering loyalty and financial well-being.



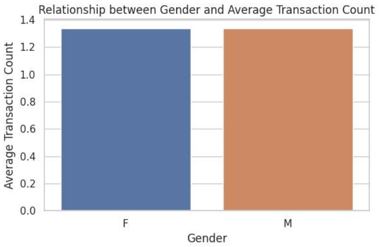
8) The analysis of age and account balances shows a concentrated scatter plot. While all age groups have potential as customers, middle-aged individuals exhibit stronger purchasing power. Targeting this demographic with tailored offerings can maximize revenue opportunities for banks.

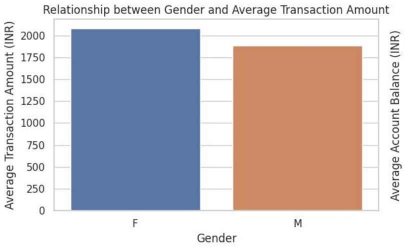


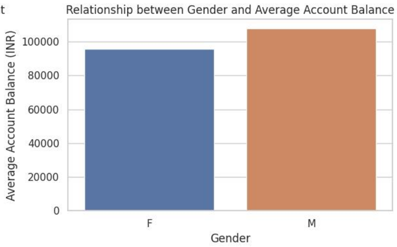
9) Number of Transactions per Year shows a rising trend, indicating an increasing transaction frequency over time before 2020. To increase transaction numbers in the future, banks should actively expand their business operations and implement strategies to attract more customers and drive revenue.



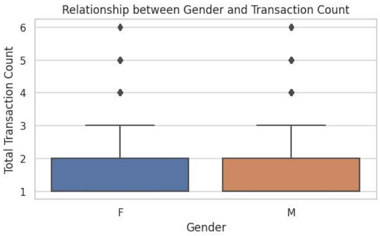
10) When examining the bar charts for transaction count, transaction amount, and account balance classified by gender, it appears that there are no substantial overall differences between males and females. This suggests that banks can conduct joint research and analysis to better understand the similarities and common needs between the two genders.

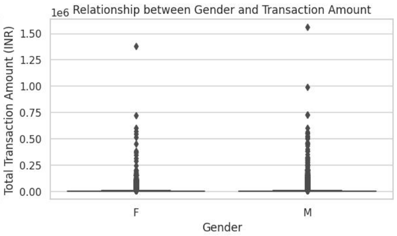


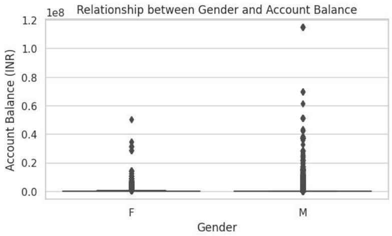




11) Classified by gender, box plots for transaction count, transaction amount, and account balance reveal significant differences within different gender groups. This emphasizes the importance of addressing the internal gaps within each gender. By recognizing these variations, banks can develop tailored strategies and services that cater to the unique needs and preferences of diverse segments within the male and female customer base.







b. use at least three different clustering algorithms to cluster customers.

Before clustering, I performed some data preprocessing steps. These steps included handling missing values, where for numerical features, I filled missing values with the mean, and for categorical features, I filled missing values with the mode. I also deleted rows with missing values in the "CustomerDateOfBirth" column. Then, I converted the date of birth into the customer's age and changed the transaction time to the transaction year.
Then, I selected the features and retained the top 100 locations with the highest customer count, labeling the remaining locations as "Other". I used a preprocessor that applied standard scaling to numerical features and one-hot encoding to categorical features. To reduce dimensionality, I used SVD and kept 50 components.

I have chosen three clustering algorithms: K-Means, HDBSCAN, and Birch. Here's a description of each algorithm:

1. K-Means

- Algorithm Description: K-Means is a widely used partition-based clustering algorithm. It aims to divide the data points into K clusters, such that the points within each cluster are as similar as possible (i.e., highly compact), while being as dissimilar as possible to points in other clusters.
- Applicable Scenarios: It's suitable for scenarios where the data dimensions are not excessively high, and the clusters are roughly circular or spherical in shape.
- Characteristics: Easy to understand and implement, with high computational efficiency. However, it requires pre-specifying the number of clusters (n_clusters) and is sensitive to the choice of initial centroids.

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
clusters_kmeans = kmeans.fit_predict(cluster_data_reduced)
data['kmeans_cluster'] = clusters_kmeans
```

2. HDBSCAN

- Algorithm Description: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. It's a variant of DBSCAN that can handle clusters of varying densities and does not require pre-specifying the number of clusters.
- Applicable Scenarios: Particularly effective for datasets where the cluster densities are uneven or the shapes of clusters are irregular.
- Characteristics: Performs well with noisy datasets. The parameters min_cluster_size and min_samples determine the way clusters are formed and their density.

```
import hdbscan
clusterer = hdbscan.HDBSCAN(min_cluster_size=50, min_samples=15)
labels = clusterer.fit_predict(cluster_data_reduced)
data['hdbscan_cluster'] = labels
```

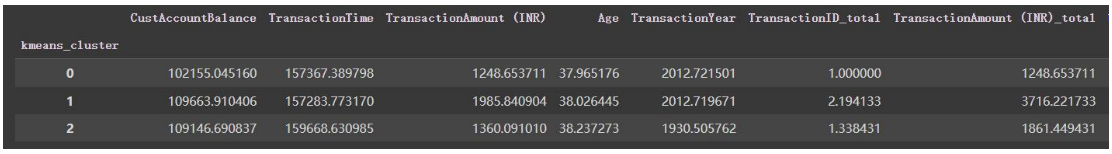
3. BIRCH

- Algorithm Description: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm designed specifically for large datasets. It progressively reduces the amount of data by building a CF (Clustering Feature) Tree.
- Applicable Scenarios: Especially suitable for large datasets, particularly those dominated by numerical features.
- Characteristics: Excellent performance on large datasets and high memory efficiency. The number of clusters formed can be controlled by the n_clusters parameter.

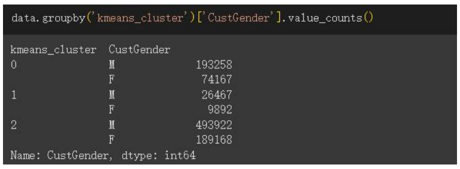
```
from sklearn.cluster import Birch
birch_clusterer = Birch(n_clusters=3)
clusters_birch = birch_clusterer.fit_predict(cluster_data_reduced)
data['birch_cluster'] = clusters_birch
```

c. explain the common characteristics shared by customers within the same cluster , as well as the differences among customers in different clusters.

1. K-Means



kmeans_cluster	CustAccountBalance	TransactionTime	TransactionAmount (INR)	Age	TransactionYear	TransactionID_total	TransactionAmount (INR)_total
0	102155.045160	157367.389798	1248.653711	37.965176	2012.721501	1.000000	1248.653711
1	109663.910406	157283.773170	1985.840904	38.026445	2012.719671	2.194133	3716.221733
2	109146.690837	159668.630985	1360.091010	38.237273	1930.505762	1.338431	1861.449431



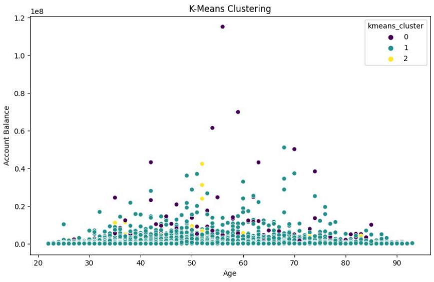
```
data.groupby('kmeans_cluster')['CustGender'].value_counts()

kmeans_cluster  CustGender
0               M             193258
                F              74167
1               M              26467
                F               9892
2               M             493922
                F             189168
Name: CustGender, dtype: int64
```

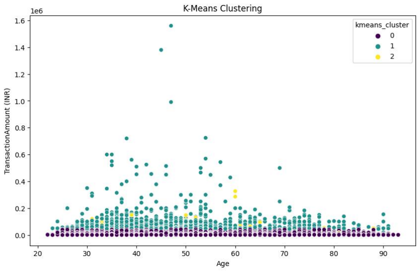
Cluster 0 comprises customers with moderate financial profiles, engaging in smaller transactions, mostly single transactions, with a higher male proportion.

Cluster 1 consists of customers with slightly better financial health, involved in larger transactions and more frequent trading, with a male-dominant but smaller overall group.

Cluster 2 features customers similar in financial status to Cluster 1, with a different pattern in transaction timing and the largest male customer group, but includes a potential data anomaly in the transaction year.

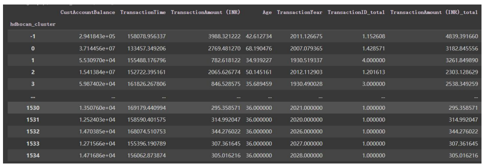
Evidence (from Financial Profile, Transaction Behavior and Gender Distribution): Cluster 0 seems to consist of customers with moderate financial profiles (lower account balances and transaction amounts), Cluster 1 includes customers who engage in larger transactions and have slightly higher account balances, and Cluster 2 resembles Cluster 1 in terms of financial status but with a different transaction time pattern. Customers in Cluster 1 engage in more transactions than those in Clusters 0 and 2. All clusters have more male customers, but the proportion varies, with Cluster 2 having the largest number.



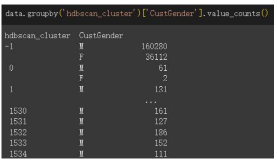


2. HDBSCAN

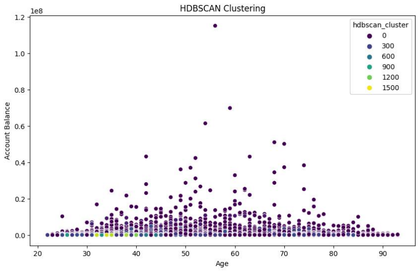
Due to HDBSCAN clustering resulting in 1535 clusters and noise points, the analysis method remains the same. We will only present the results here without further analysis.

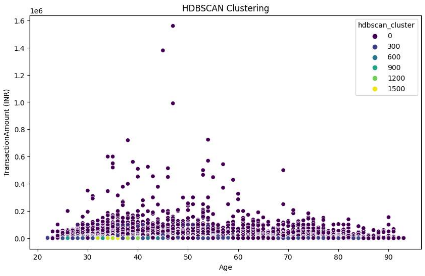


hdbscan_cluster	CustAccountBalance	TransactionTime	TransactionAmount (INR)	Age	TransactionYear	TransactionID_total	TransactionAmount (INR)_total
-1	2.941843e+05	158078.956337	3988.321222	42.612734	2011.126675	1.152608	4839.391660
0	3.714456e+07	133457.349206	2769.481270	68.190476	2007.079365	1.428571	3182.845556
1	5.530970e+04	155488.176796	782.618122	34.939227	1930.519337	4.000000	3261.849890
2	1.541384e+07	152722.395161	2065.626774	50.145161	2012.112903	1.201613	2303.128629
3	5.987402e+04	161826.267806	846.528575	35.689459	1930.490028	3.000000	2538.349259
...	...	...	...	...	...	...	...
1530	1.350760e+04	169179.440994	295.358571	36.000000	2021.000000	1.000000	295.358571
1531	1.252403e+04	158590.401575	314.992047	36.000000	2020.000000	1.000000	314.992047
1532	1.470385e+04	168074.510753	344.276022	36.000000	2026.000000	1.000000	344.276022
1533	1.271566e+04	155396.190789	307.361645	36.000000	2027.000000	1.000000	307.361645
1534	1.471686e+04	156062.873874	305.016216	36.000000	2028.000000	1.000000	305.016216

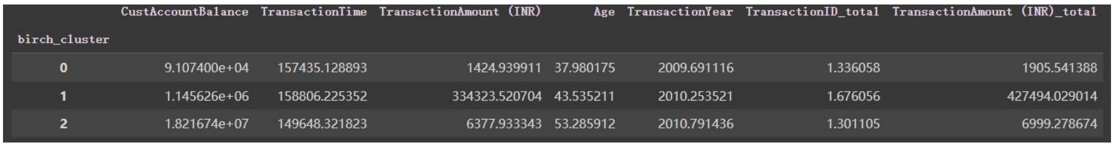


```
data.groupby('hdbscan_cluster')['CustGender'].value_counts()

hdbscan_cluster  CustGender
-1               M             160280
                 F              36112
 0               M                 61
                 F                  2
 1               M                131
                                  ...
 1530            M                161
 1531            M                127
 1532            M                186
 1533            M                152
 1534            M                111
```

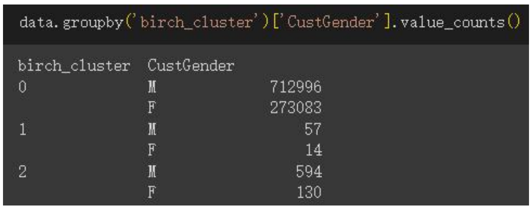




3. BIRCH



birch_cluster	CustAccountBalance	TransactionTime	TransactionAmount (INR)	Age	TransactionYear	TransactionID_total	TransactionAmount (INR)_total
0	9.107400e+04	157435.128893	1424.939911	37.980175	2009.691116	1.336058	1905.541388
1	1.145626e+06	158806.225352	334323.520704	43.535211	2010.253521	1.676056	427494.029014
2	1.821674e+07	149648.321823	6377.933343	53.285912	2010.791436	1.301105	6999.278674



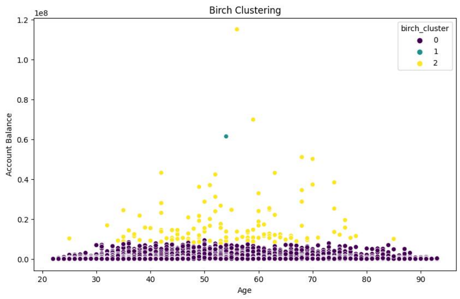
```
data.groupby('birch_cluster')['CustGender'].value_counts()

birch_cluster  CustGender
0              M             712996
               F             273083
1              M                 57
               F                 14
2              M                594
               F                130
```

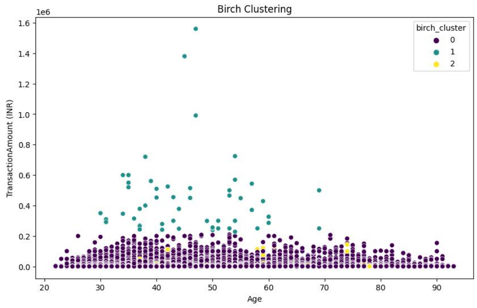
Cluster 0 represents ordinary consumers with average financial status and low transaction amounts.

Cluster 1 consists of high-net-worth individuals, with very high account balances and transaction amounts.

Cluster 2 probably includes ultra-high-net-worth individuals or corporate clients, characterized by extremely high account balances but moderate transaction amounts, reflecting more conservative capital management.

Evidence (from Financial Profile, Transaction Behavior and Gender Distribution):Cluster 0 consists of customers with moderate financial profiles, as indicated by lower account balances and transaction amounts. Cluster 1 includes customers engaged in larger transactions with slightly higher account balances. Cluster 2 resembles Cluster 1 in terms of financial status but exhibits a different pattern in transaction timing. Customers in Cluster 1 engage in more transactions compared to those in Clusters 0 and 2, indicating higher financial activity. All clusters predominantly feature male customers. However, the proportion varies across clusters, with Cluster 2 having the largest number of male customers.





Additionally, to enhance the specificity of spatial distance, we can replace location names with latitude and longitude coordinates. In Python, we can achieve this by utilizing the geopy library and specifically the Nominatim geocoder.

References：

[1] Parameter Selection for HDBSCAN：https://hdbscan.readthedocs.io/en/latest/parameter_selection.html

[2] A high performance implementation of HDBSCAN clustering: https://github.com/scikit-learn-contrib/hdbscan

[3] BIRCH 聚类算法: https://www.cnblogs.com/pinard/p/6179132.html

[4] Python 根据地址查询经纬度: https://www.jianshu.com/p/365ed2398fa3