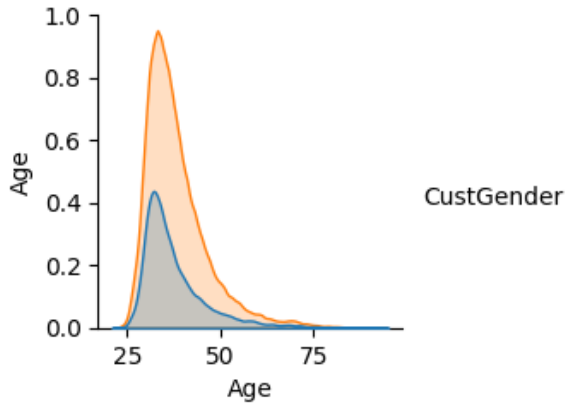
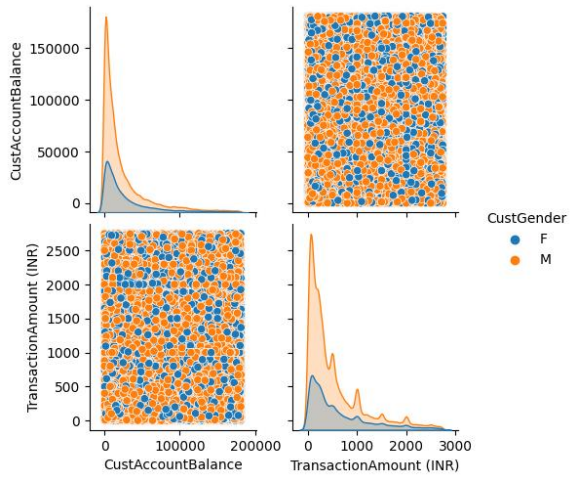
Q6. Bank Customer Clustering

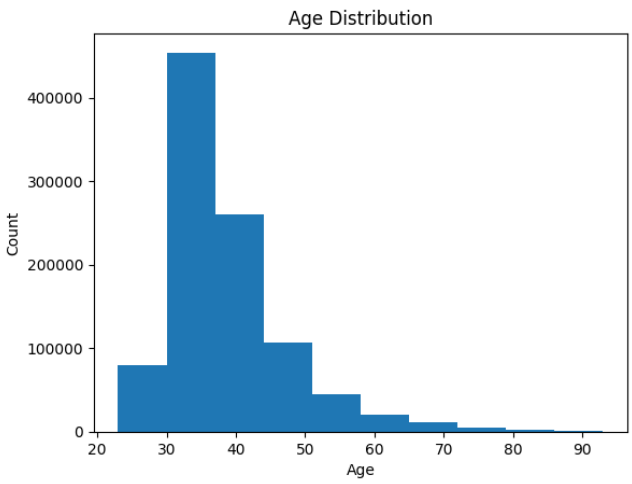
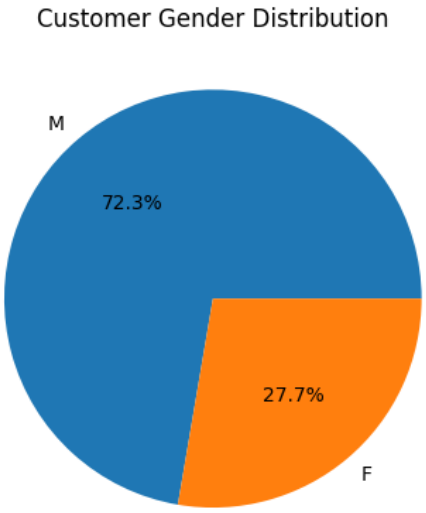
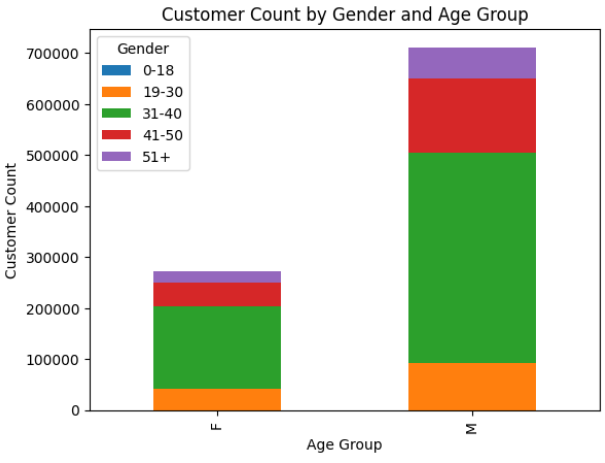
Just run the jupyter notebook of `Q6.ipynb`. It uses the given dataset.

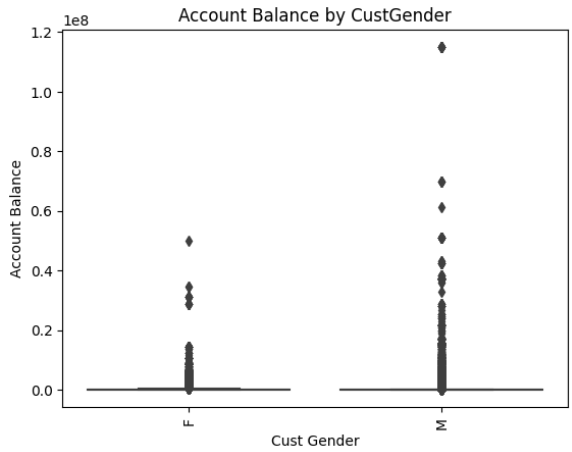
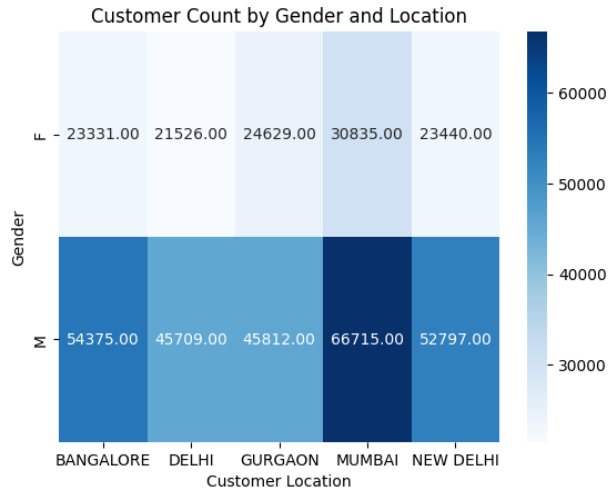
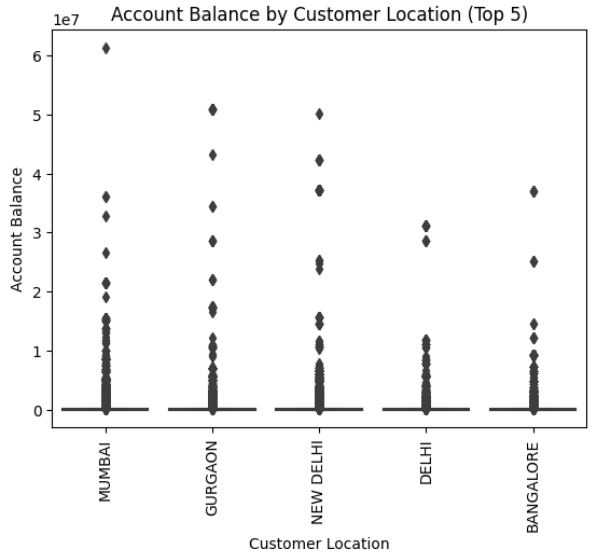
Preprocessing:

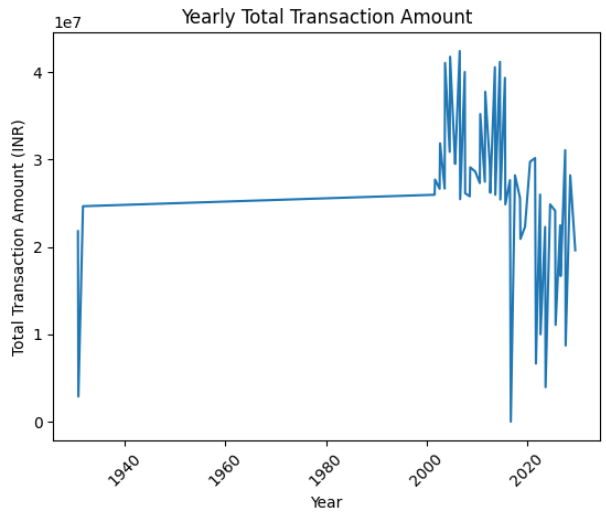
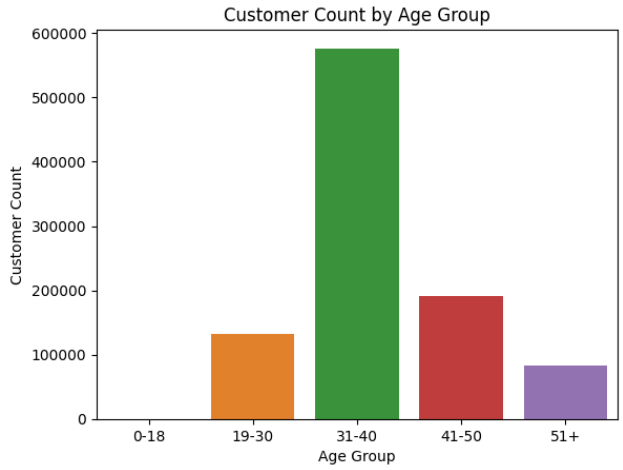
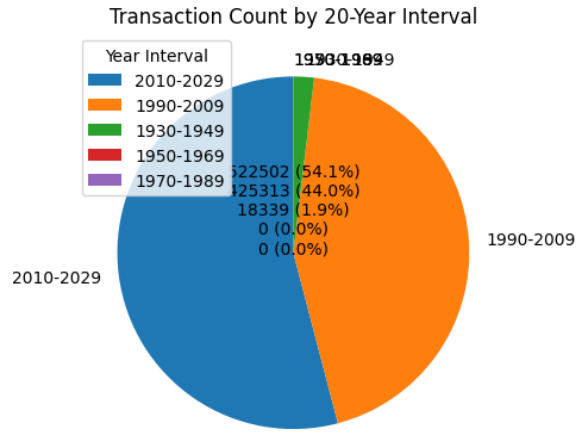
- Transfer [CustomerDateOfBirth] and [TransactionDate] to datetime type
- calculate the age from [CustomerDateOfBirth], calculate the hour from [TransactionDate]
- remove the NAN
- remove outlier of [CustAccountBalance] and [TransactionAmount (INR)]

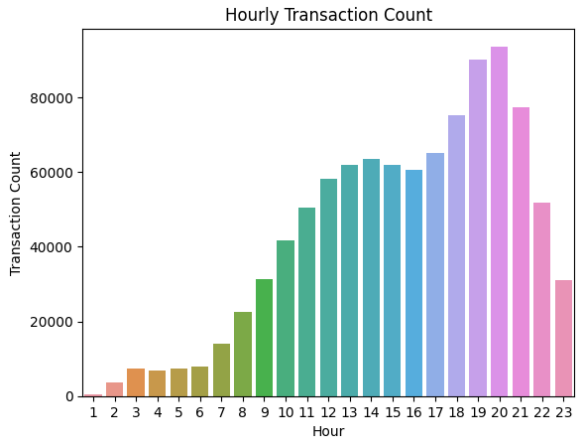
a. at least 10 pictures and at least 10 business insights

	1.The customers are between 20~80 years old. Female(blue) customers are younger than male (orange).
	2.Female(blue) customers have fewer finacial trasaction and deposite than male (orange).

	3.The customers between 30~40 have most counts which is more than 400000.
	4.Female(orange) customers are fewer than male (blue) which is around 72.3%.
	5.Both male and female have most customer in middle age.

	6.Male customers have more high-value customers.
	7.The biggest group of customers are in MUMBAI both for male and female.
	8.The top_5 account balance associate with location.

	9.In recent years, there has been significant volatility on total transaction amount, with the highest transaction amounts occurring between 2000 and 2010.
	10.Young adults between the ages of 31 and 40 have the highest proportion, followed by middle-aged adults between the ages of 40 and 50.
	11.In earlier years, the transaction volume was relatively small, but it has accounted for the largest proportion in the past 20 years.

	12.The user transactions are most active between 18:00 and 21:00, reaching a peak around 20:00.

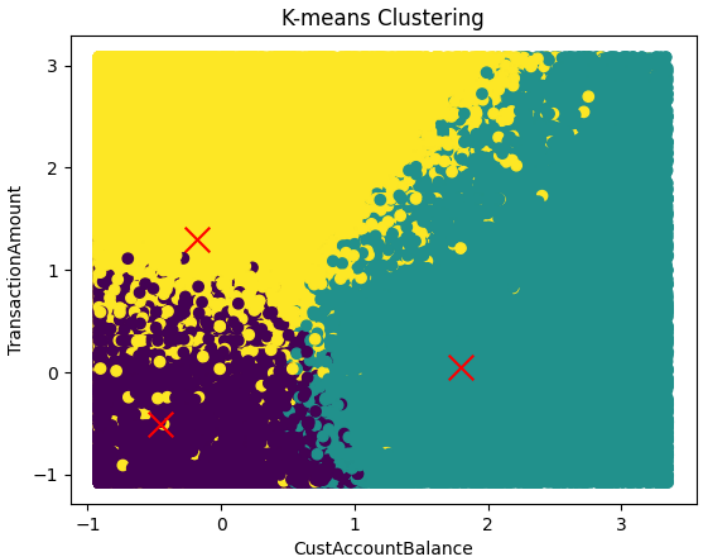
b. three different clustering algorithms to cluster customers

- **K-means**

  Use all data

  Number of cluster is 3

  Feature : ['CustAccountBalance', 'TransactionAmount (INR)', 'Age']

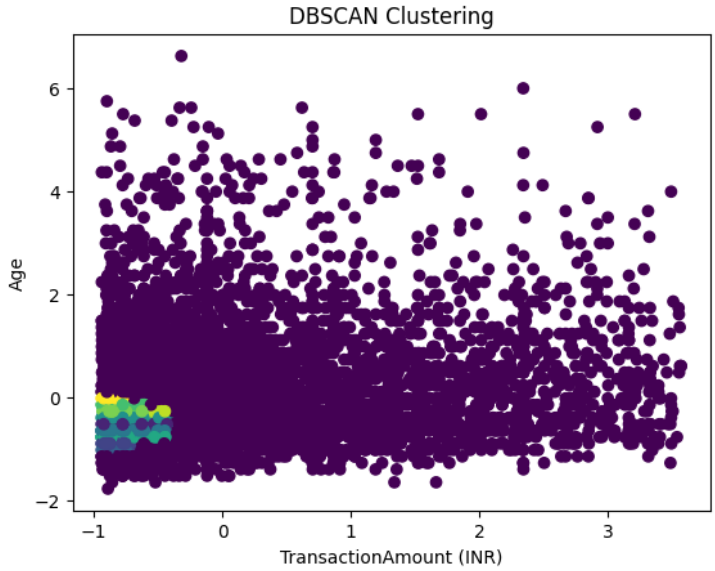


- **DBSCAN clustering**

  Use all data*0.01

  Number of cluster is 11

  Feature : ['TransactionAmount (INR)', 'Age']

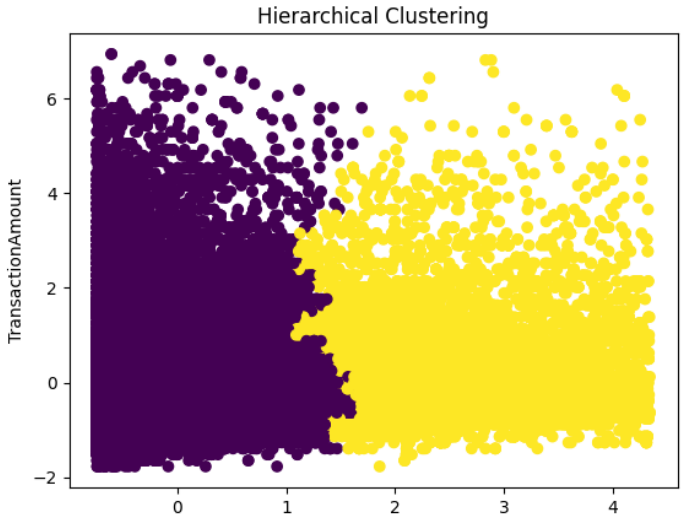


- **Hierarchical clustering(n_clusters=2)**

  Use all data*0.1

  Number of cluster is 2

  Feature : ['CustAccountBalance', 'Age']



c. common characteristics & difference

	Common characteristics	**difference**
Fig.1	Same distribution of CustAccountBalance and TransactionAmount (INR)	Difference distribution
Fig.2	Same distribution of CustAccountBalance and age	Difference distribution
Fig.3	Same distribution of TransactionAmount (INR) and age	Difference distribution