Milestone 1

: Key Findings from DataSet Exploration

1. **Dataset Overview : Dataset** shape is 10002 records and 14 columns, and Exited column is our target column to predict client exited or not where 0 means not exited and 1 means exited.
2. **Missing Values :** (**We** removed all missing values because they are few values and will not affect the data).
    1- **Geography Column : Has** (1 missing value) .
    2- **Age column : Has** (1 missing value).
    3- **HasCrCard Column : Has** (1 missing value).
    4- **IsActiveMember column : Has** (1 missing value).
3. **Duplicates Columns : 2 Duplicated** records found (And We removed them).
4. **Unimportant Columns : Idcustomer** column, Surname column, and Rownumber column are remove from dataset because not important.
5. **Numerical Columns Summary :**
    1- **CreditScore Column : Has** ranges from **350 to** 850, has normal credit score range, has 15 Rows outliers from 383 to below and from 919 to above we did not remove them because them normal credit score range.
    2- **Age Column : Has** ranges from **18 to 92**, has 359 records outliers from age 14 to below and from age 62 to above, and we don't remove them because in normal range, and column needs scaling later, and convert its datatype later to integer datatype.
    3- **Balance Column : Has** a **median of ~97,000** but **25% of clients have a 0 balance** there count is 3616 Clients, 500 Clients left the bank and 3116 Clients did not left the bank, and column needs scaling later.
    4- **Estimatedslary Column : All** clients have normal estimatedsalary, and column needs scaling later.

6. **Categorical Columns Summary :**
    1- **Geograohy Column : We** most clients are French clients, then german clients, and we encode it later to hot encode.
    2- **Gender Column : We** most clients are male than female, and we encode it to 0 , 1 later where 0 means male and 1 means female.
    3- **Tenure Column : Most** client have 10 years in the bank and others have normal distribution.
    4- **HasCcard Column : Most** clients have creditcard and there count are above 7000 clients, and we convert hasccard column to integer datatype later.
    5- **Isactivemember Column : Most** clients are isactivemember and there count are above 5000 clients, and we convert isactivemember column to integer datatype later as it was float datatype.
    6- **NumOfProducts Column : Has** ranges from 1 to 4 most clients have 1 or 2 products and few have 3 or 4 products.
    7- **Exited Column** : **20.4%** from clients left the bank, meaning class imbalance might need handling later.
7. **Correlation Summary :**
    1- Columns of (Age, and Balance) have positive correlation, and Column of (Isactivemember) has negative correlation, and All have strong correlation with the target exited column.
    2- Columns of (Tenure, and Estimatedsalary) have positive correlation, and Columns of (Creditscore, Numofproduct, and Hasccard) have negative correlation, and All have weak correlation with the target exited column.
8. **Dataset Overview After Cleaning : Dataset** shape is 9996 records and 11 columns.