Data Collection and Preprocessing Phase

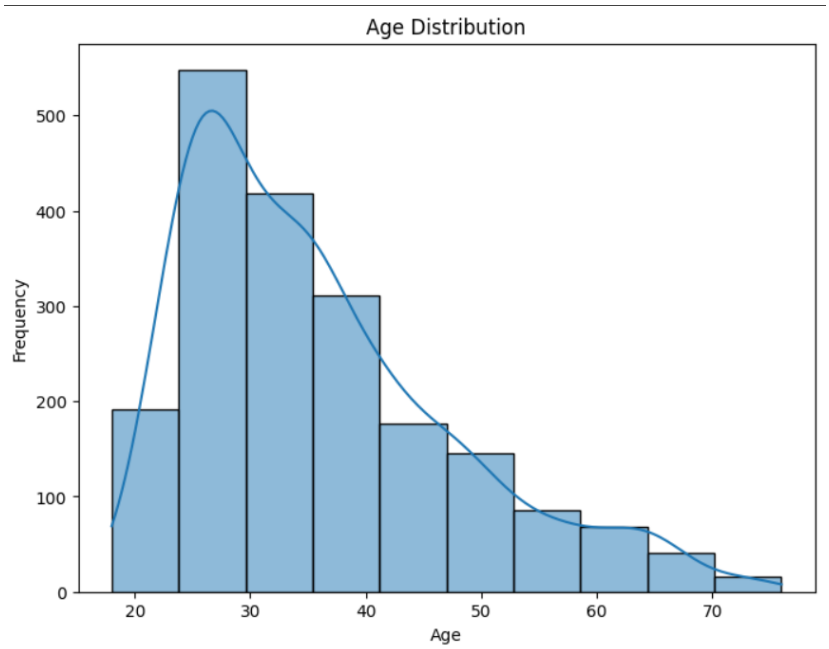
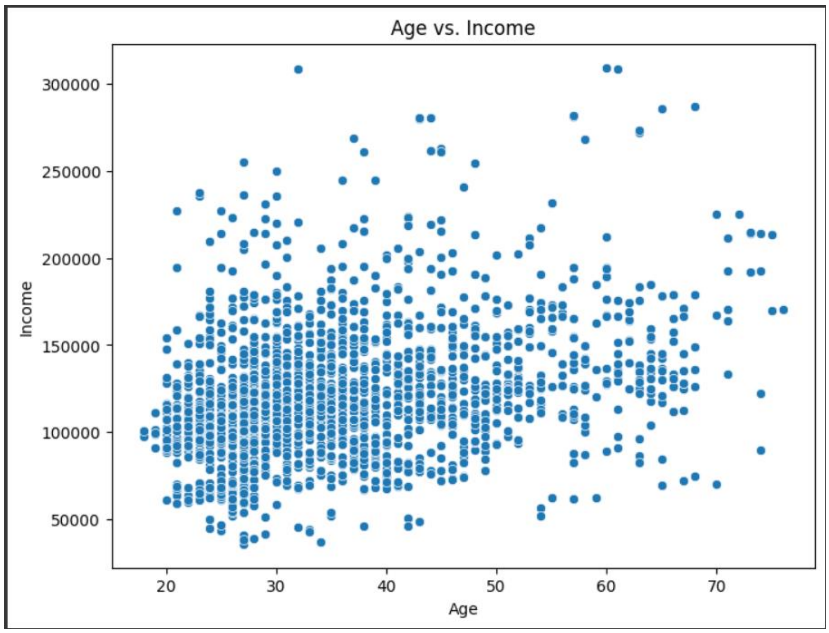
Date	23 September 2024
Team ID	LTVIP2024TMID25001
Project Title	Customer Segmentation Using Machine Learning
Maximum Marks	6 Marks

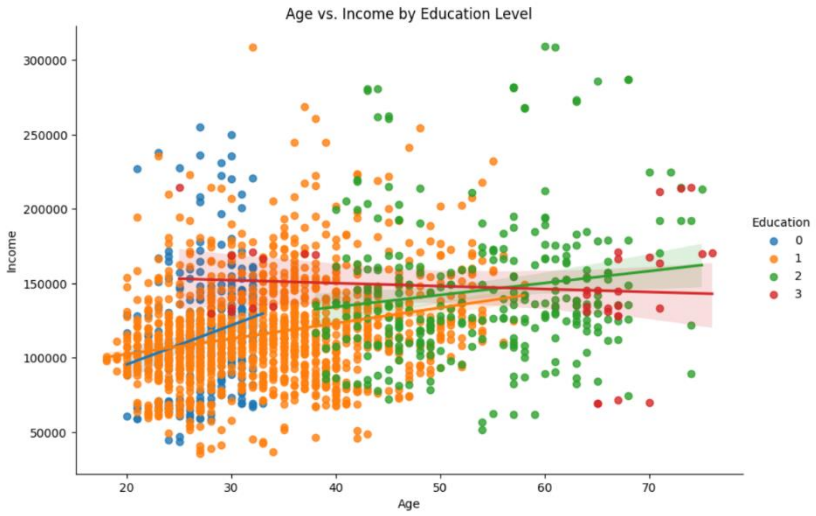
Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	2000RowsX8Columns

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

Univariate Analysis	
Bivariate Analysis	

Multivariate Analysis	
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data	

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

Handling Missing Data	

```
df.isnull().sum()
```

Data Transformation	

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
x=sc.fit_transform(x)
```

Feature Engineering	

```
# Feature Engineering - One-hot encode categorical variables (Sex, Marital status, Education, Occupation, Settlement size)
df = pd.get_dummies(df, columns=['Sex', 'Marital status', 'Education', 'Occupation', 'Settlement size'], drop_first=True)
```

Save Processed Data	``` import pickle pickle.dump(xgb_model, open("xgbmodel.pkl", 'wb')) from sklearn.preprocessing import StandardScaler sc = StandardScaler() pickle.dump(sc,open('scaling.pkl','wb')) ```