Data Collection and Preprocessing Phase

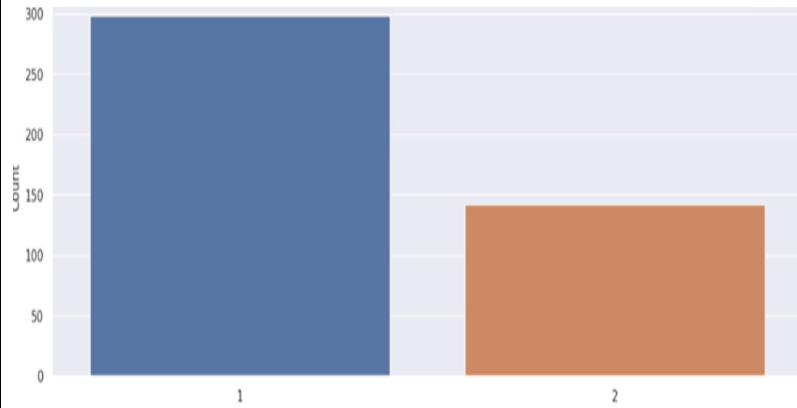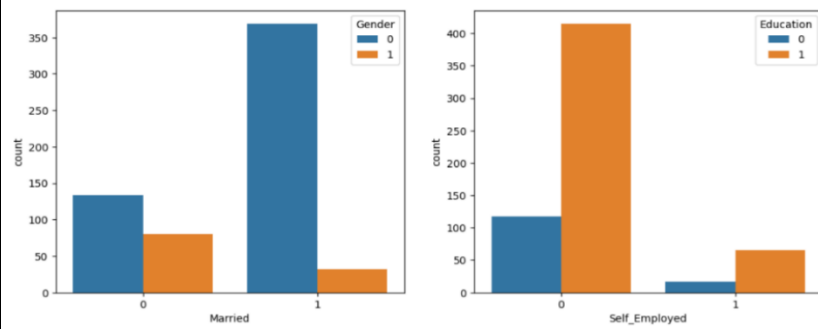| Date	July, 2024
Team ID	739656
Project Title	Zombie detection using Machine Learning
Maximum Marks	6 Marks

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	

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
count	2000.000000	2000.000000	2000.000000	2000.00000	2000.000000	2000.000000	2000.000000
mean	0.457000	0.496500	35.909000	1.03800	120954.419000	0.810500	0.739000
std	0.498272	0.500113	11.719402	0.59978	38108.824679	0.638587	0.812533
min	0.000000	0.000000	18.000000	0.00000	35832.000000	0.000000	0.000000
25%	0.000000	0.000000	27.000000	1.00000	97663.250000	0.000000	0.000000
50%	0.000000	0.000000	33.000000	1.00000	115548.500000	1.000000	1.000000
75%	1.000000	1.000000	42.000000	1.00000	138072.250000	1.000000	1.000000
max	1.000000	1.000000	76.000000	3.00000	309364.000000	2.000000	2.000000

Univariate Analysis	
Bivariate Analysis	
Multivariate Analysis	
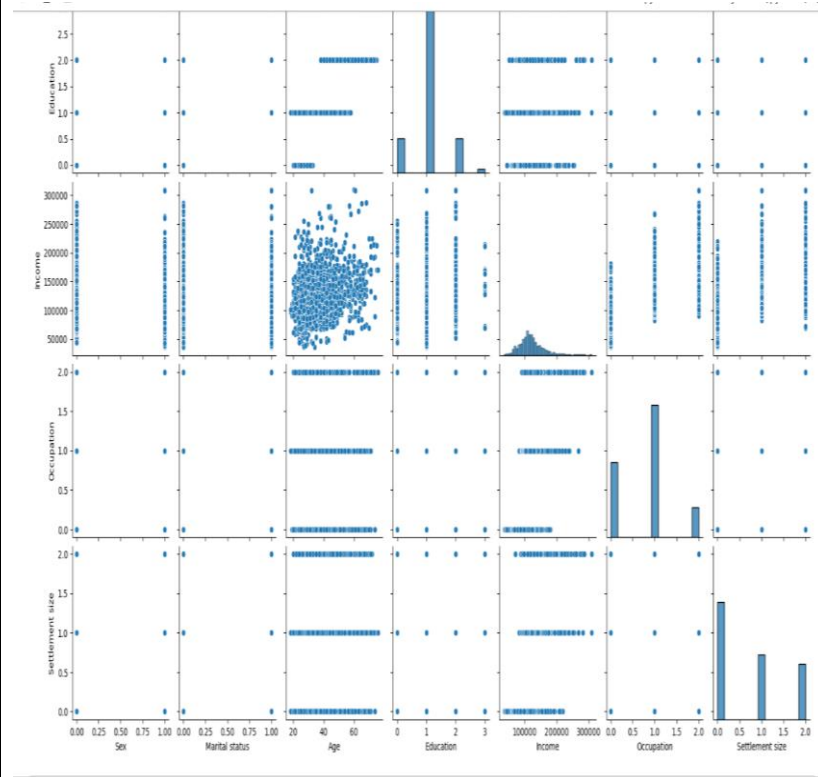
Outliers and Anomalies	
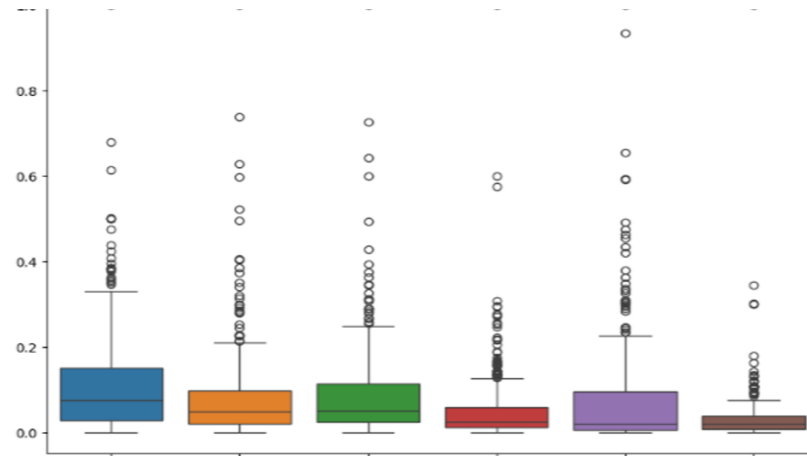
Data Preprocessing Code Screenshots

| Loading Data | ```
3]: os.chdir(r"C:/Users/kusur/Downloads")
    data = pd.read_csv('segmentation data.csv',header='infer')
    data.head()
```<br><br>3]: <table><tr><td></td><td>ID</td><td>Sex</td><td>Marital status</td><td>Age</td><td>Education</td><td>Income</td><td>Occupation</td><td>Settlement size</td></tr><tr><td>0</td><td>100000001</td><td>0</td><td>0</td><td>67</td><td>2</td><td>124670</td><td>1</td><td>2</td></tr><tr><td>1</td><td>100000002</td><td>1</td><td>1</td><td>22</td><td>1</td><td>150773</td><td>1</td><td>2</td></tr><tr><td>2</td><td>100000003</td><td>0</td><td>0</td><td>49</td><td>1</td><td>89210</td><td>0</td><td>0</td></tr><tr><td>3</td><td>100000004</td><td>0</td><td>0</td><td>45</td><td>1</td><td>171565</td><td>1</td><td>1</td></tr><tr><td>4</td><td>100000005</td><td>0</td><td>0</td><td>53</td><td>1</td><td>149031</td><td>1</td><td>1</td></tr></table> |
| Handling Missing Data | ```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Sex              2000 non-null   int64
 1   Marital status   2000 non-null   int64
 2   Age              2000 non-null   int64
 3   Education        2000 non-null   int64
 4   Income           2000 non-null   int64
 5   Occupation       2000 non-null   int64
 6   Settlement size  2000 non-null   int64
dtypes: int64(7)
memory usage: 109.5 KB
``` |

| Data Transformation | ```python
data = minmax_scale(data,feature_range=(0,1))

import pickle

pickle.dump(data,open("scale.pk2",'wb'))

names = ['Sex','Martial status','Age','Education','Income','Occupation','Settlement size']
data = pd.DataFrame(data,columns=names)

wcss = []
for i in range(1, 11):
    kmeans = sk.cluster.KMeans(n_clusters=i, init='k-means++', random_state=0)
    kmeans.fit(data)
    wcss.append(kmeans.inertia_)
``` |
|---|---|
| Feature Engineering | Attached the codes in final submission |
| Save Processed Data | - |