customer-segmentation

March 10, 2025

```
[ ]: from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ]: file_path='/content/customer segmentation.csv'
```

```
[ ]: import pandas as pd
     df=pd.read_csv(file_path)
```

```
[ ]: df.head()
```

```
[ ]:          ID  Sex  Marital status  Age  Education  Income  Occupation  \
     0  100000000    0               0   27          0  302122           1
     1  100000001    1               0   45          3  228035           0
     2  100000002    1               1   37          0  126914           2
     3  100000003    1               0   75          1   58989           2
     4  100000004    1               1   75          3  156718           1

        Settlement size
     0                1
     1                0
     2                2
     3                0
     4                2
```

```
[ ]: df
```

```
[ ]:              ID  Sex  Marital status  Age  Education  Income  Occupation  \
     0      100000000    0               0   27          0  302122           1
     1      100000001    1               0   45          3  228035           0
     2      100000002    1               1   37          0  126914           2
     3      100000003    1               0   75          1   58989           2
     4      100000004    1               1   75          3  156718           1
     …          …  …             …  …         …     …         …
     89995  100089995    0               0   36          2   43672           1
     89996  100089996    0               1   56          2   74230           2
     89997  100089997    1               1   39          2   61334           2
```

```
89998  100089998    0               1   55           2  178610           1
89999  100089999    1               1   71           2  299329           1

       Settlement size
0                    1
1                    0
2                    2
3                    0
4                    2
…                …
89995                0
89996                0
89997                1
89998                0
89999                0

[90000 rows x 8 columns]
```

```
[ ]: df.isnull()
```

```
[ ]:          ID    Sex  Marital status    Age  Education  Income  Occupation  \
0      False  False           False  False      False   False       False
1      False  False           False  False      False   False       False
2      False  False           False  False      False   False       False
3      False  False           False  False      False   False       False
4      False  False           False  False      False   False       False
…      …    …             …    …        …     …         …
89995  False  False           False  False      False   False       False
89996  False  False           False  False      False   False       False
89997  False  False           False  False      False   False       False
89998  False  False           False  False      False   False       False
89999  False  False           False  False      False   False       False

       Settlement size
0                False
1                False
2                False
3                False
4                False
…                …
89995            False
89996            False
89997            False
89998            False
89999            False

[90000 rows x 8 columns]
```

```
[ ]: df.isnull().sum()
```

```
[ ]: ID                 0
     Sex                0
     Marital status     0
     Age                0
     Education          0
     Income             0
     Occupation         0
     Settlement size    0
     dtype: int64
```

```
[ ]: df.dropna(inplace=True)
```

```
[ ]: df
```

```
[ ]:               ID  Sex  Marital status  Age  Education  Income  Occupation  \
     0      100000000    0               0   27          0  302122           1
     1      100000001    1               0   45          3  228035           0
     2      100000002    1               1   37          0  126914           2
     3      100000003    1               0   75          1   58989           2
     4      100000004    1               1   75          3  156718           1
     …           …  …            …  …        …     …          …
     89995  100089995    0               0   36          2   43672           1
     89996  100089996    0               1   56          2   74230           2
     89997  100089997    1               1   39          2   61334           2
     89998  100089998    0               1   55          2  178610           1
     89999  100089999    1               1   71          2  299329           1

            Settlement size
     0                    1
     1                    0
     2                    2
     3                    0
     4                    2
     …                 …
     89995                0
     89996                0
     89997                1
     89998                0
     89999                0

     [90000 rows x 8 columns]
```

```
[ ]: df.shape
```

```
[ ]: (90000, 8)
```

[ ]:

```
df.info()
```

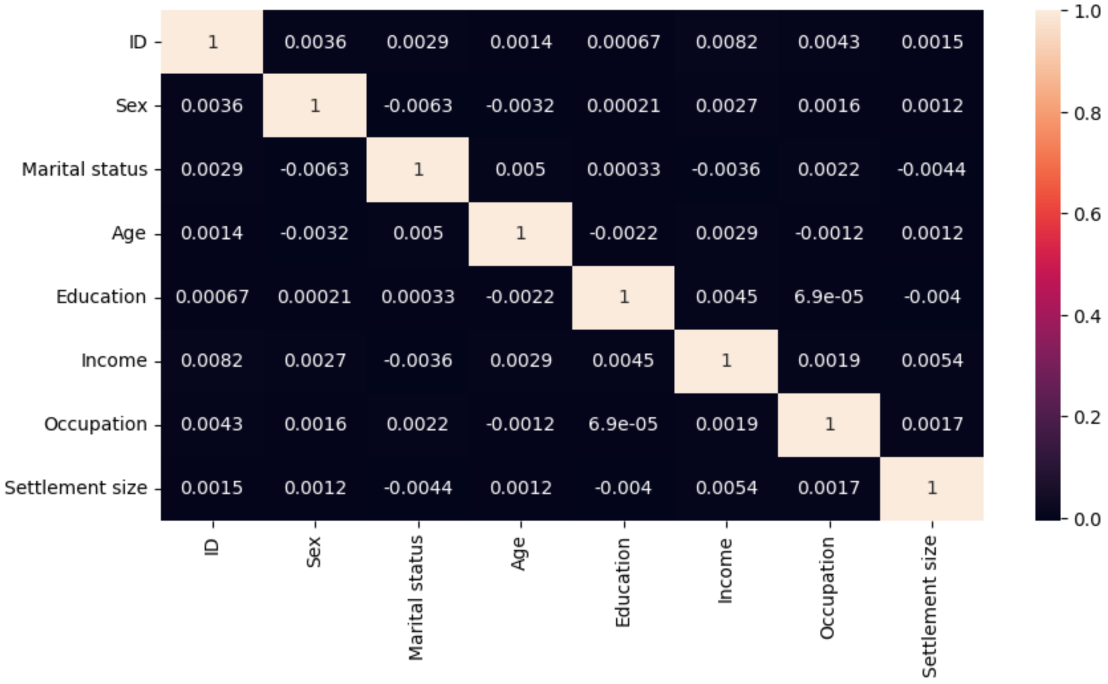
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90000 entries, 0 to 89999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               90000 non-null  int64
 1   Sex              90000 non-null  int64
 2   Marital status   90000 non-null  int64
 3   Age              90000 non-null  int64
 4   Education        90000 non-null  int64
 5   Income           90000 non-null  int64
 6   Occupation       90000 non-null  int64
 7   Settlement size  90000 non-null  int64
dtypes: int64(8)
memory usage: 5.5 MB
```

[ ]:

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,5))
sns.heatmap(df.corr(),annot=True)
```

[ ]:

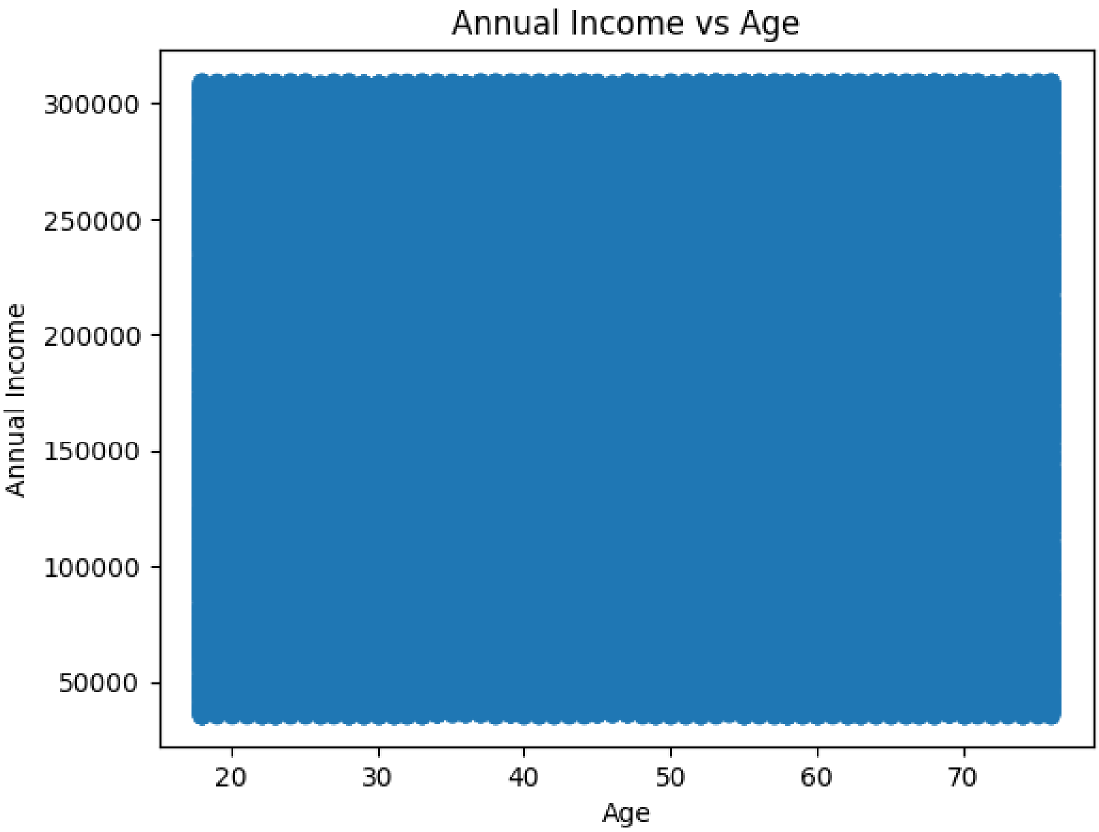
```
<Axes: >
```



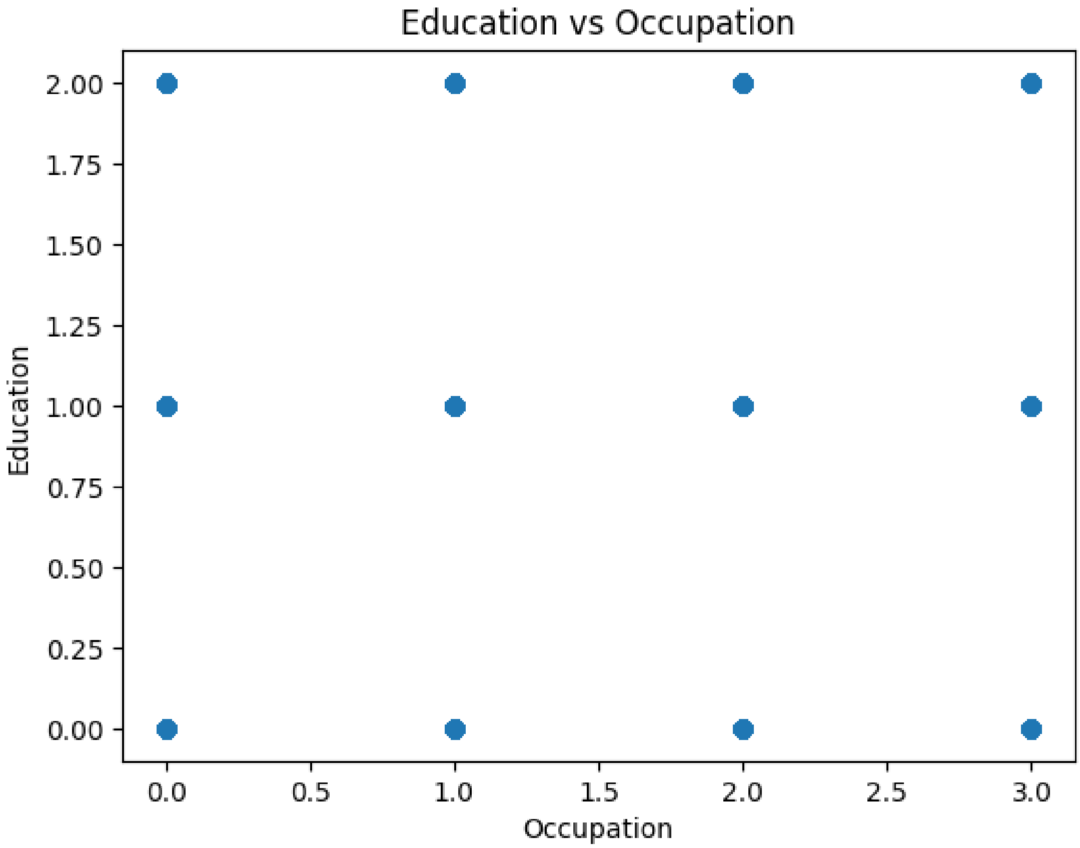
```
[ ]: get_standard_values = lambda x: round(x, 2)
     get_standard_values
```

```
[ ]: <function __main__.<lambda>(x)>
```

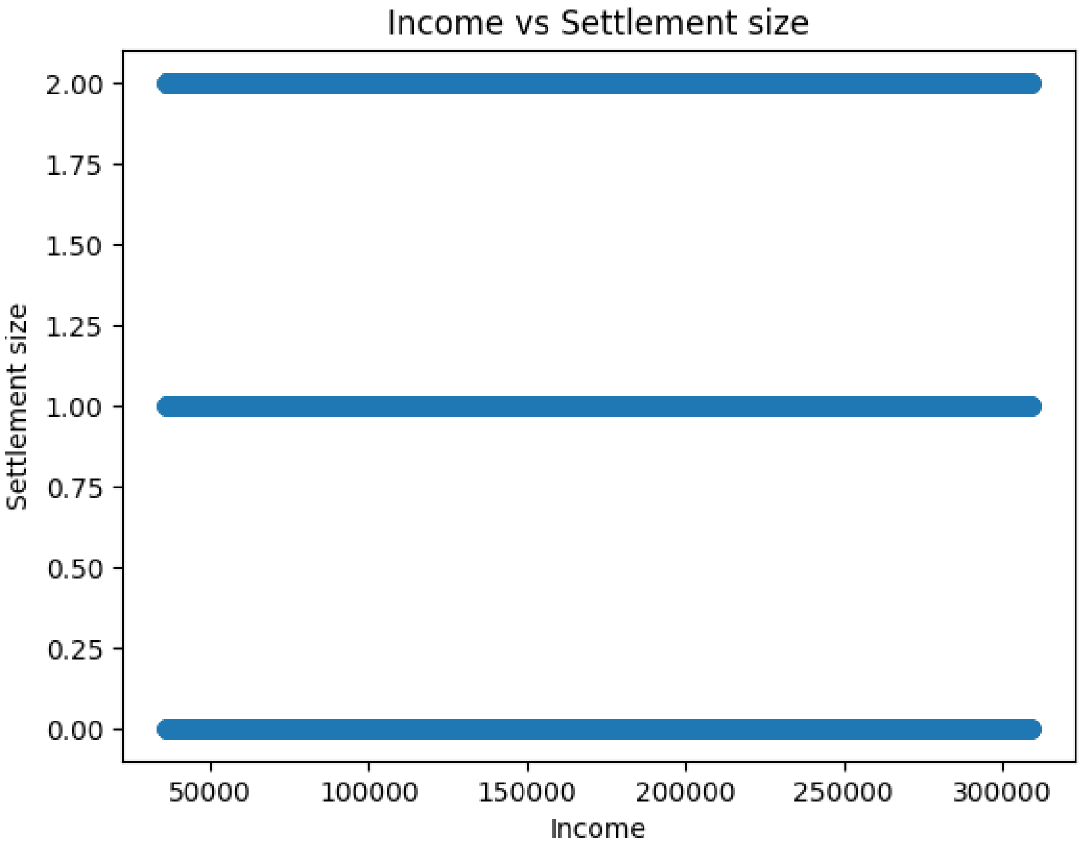
```
[ ]: plt.scatter(get_standard_values(df['Age']),get_standard_values(df['Income']))
     plt.title('Annual Income vs Age')
     plt.xlabel('Age')
     plt.ylabel('Annual Income')
     plt.show()
```



```
[ ]: plt.
      ↪scatter(get_standard_values(df['Education']),get_standard_values(df['Occupation']))
     plt.title('Education vs Occupation')
     plt.xlabel('Occupation')
     plt.ylabel('Education')
     plt.show()
```



```
[ ]: plt.
       ↪scatter(get_standard_values(df['Income']),get_standard_values(df['Settlement␣
       ↪size']))
     plt.title('Income vs Settlement size')
     plt.xlabel('Income')
     plt.ylabel('Settlement size')
     plt.show()
```



```
[ ]: df.corr()
     df
```

```
[ ]:               ID  Sex  Marital status  Age  Education  Income  Occupation  \
     0      100000000    0               0   27          0  302122           1
     1      100000001    1               0   45          3  228035           0
     2      100000002    1               1   37          0  126914           2
     3      100000003    1               0   75          1   58989           2
     4      100000004    1               1   75          3  156718           1
     …          …  …             …  …         …     …         …
     89995  100089995    0               0   36          2   43672           1
     89996  100089996    0               1   56          2   74230           2
     89997  100089997    1               1   39          2   61334           2
     89998  100089998    0               1   55          2  178610           1
     89999  100089999    1               1   71          2  299329           1

            Settlement size
     0                    1
     1                    0
     2                    2
```

```
3                    0
4                    2
…                …
89995                0
89996                0
89997                1
89998                0
89999                0

[90000 rows x 8 columns]
```

```
[ ]: x_train=df.drop(['ID','Age'],axis=1)
     x_test=df['Age']
     y_train=df['Age']
     y_test=df['Age']
```

```
[ ]: x_train,x_test,y_train,y_test
```

```
[ ]: (       Sex  Marital status  Education  Income  Occupation  Settlement size
      0        0               0          0  302122           1                1
      1        1               0          3  228035           0                0
      2        1               1          0  126914           2                2
      3        1               0          1   58989           2                0
      4        1               1          3  156718           1                2
      …    …             …        …     …         …              …
      89995    0               0          2   43672           1                0
      89996    0               1          2   74230           2                0
      89997    1               1          2   61334           2                1
      89998    0               1          2  178610           1                0
      89999    1               1          2  299329           1                0

      [90000 rows x 6 columns],
      0        27
      1        45
      2        37
      3        75
      4        75
               ..
      89995    36
      89996    56
      89997    39
      89998    55
      89999    71
      Name: Age, Length: 90000, dtype: int64,
      0        27
      1        45
      2        37
      3        75
      4        75
               ..
      89995    36
      89996    56
      89997    39
      89998    55
      89999    71
      Name: Age, Length: 90000, dtype: int64,
      0        27
      1        45
      2        37
      3        75
      4        75
               ..
      89995    36
      89996    56
      89997    39
      89998    55
      89999    71
      Name: Age, Length: 90000, dtype: int64)
```

[ ]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x_train, y_train,␣
 ↪test_size=0.2, random_state=42)
print(x_train.shape, x_test.shape, y_train.shape, y_test.shape)
```

```
(72000, 6) (18000, 6) (72000,) (18000,)
```