```
import pandas as pd
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

customers = pd.read_csv("Customers.csv")

products = pd.read_csv("Products.csv")

transactions = pd.read_csv("Transactions.csv")


customer_data = transactions.merge(customers[['CustomerID',
'Region']], on='CustomerID', how='left')
customer_data = customer_data.merge(products[['ProductID',
'Category']], on='ProductID', how='left')

customer_profile = customer_data.groupby('CustomerID').agg({
    'TotalValue': 'sum',
    'Quantity': 'sum',
    'Region': 'first'
}).reset_index()

customer_profile['Region'] =
customer_profile['Region'].astype('category').cat.codes

X = customer_profile.drop(columns='CustomerID')

cosine_sim = cosine_similarity(X)

def get_top_lookalikes(cust_id, cosine_sim_matrix, n=3):
    cust_index = customer_profile[customer_profile['CustomerID'] ==
cust_id].index[0]

    sim_scores = list(enumerate(cosine_sim_matrix[cust_index]))

    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = [score for score in sim_scores if score[0] !=
cust_index]

    top_similar_customers = sim_scores[:n]

    similar_customers = [(customer_profile.iloc[i[0]]['CustomerID'],
i[1]) for i in top_similar_customers]

    return similar_customers

lookalikes = {}
for i in range(1, 21):
    cust_id = f'C{str(i).zfill(4)}'
    lookalikes[cust_id] = get_top_lookalikes(cust_id, cosine_sim)
```

```
lookalike_df = []
for cust_id, similar_customers in lookalikes.items():
    for similar_customer in similar_customers:
        lookalike_df.append([cust_id, similar_customer[0],
similar_customer[1]])

lookalike_df = pd.DataFrame(lookalike_df, columns=['CustomerID',
'LookalikeID', 'SimilarityScore'])
lookalike_df.to_csv('Lookalike.csv', index=False)

print(lookalike_df.head())
```

```
  CustomerID LookalikeID  SimilarityScore
0      C0001       C0011              1.0
1      C0001       C0131              1.0
2      C0001       C0191              1.0
3      C0002       C0043              1.0
4      C0002       C0142              1.0
```

```
lookalikes_df = pd.DataFrame(

    [(k, v[0][0], v[0][1], v[1][0], v[1][1], v[2][0], v[2][1]) for k,
v in lookalikes.items()],

    columns=['CustomerID', 'Lookalike1', 'Score1', 'Lookalike2',
'Score2', 'Lookalike3', 'Score3']

)



lookalikes_df.to_csv('Lookalike.csv', index=False)

print("Lookalike results saved to Lookalike.csv")
```

```
Lookalike results saved to Lookalike.csv
```

```
pt = pd.read_csv("Lookalike.csv")

pt
```


```
{"summary":"{\n  \"name\": \"pt\",\n  \"rows\": 20,\n  \"fields\": [\n
{\n      \"column\": \"CustomerID\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 20,\n
\"samples\": [\n          \"C0001\",\n          \"C0018\",\n
\"C0016\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Lookalike1\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 20,\n        \"samples\":
[\n          \"C0011\",\n          \"C0068\",\n          \"C0096\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
```

```
}\n    },\n    {\n      \"column\": \"Score1\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
2.598397459498755e-07,\n        \"min\": 0.9999988298189876,\n
\"max\": 0.9999999999970868,\n        \"num_unique_values\": 18,\n
\"samples\": [\n          0.9999999917150764,\n
0.9999999978943408,\n          0.9999999910572124\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Lookalike2\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 20,\n        \"samples\": [\n
\"C0131\",\n          \"C0070\",\n          \"C0185\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Score2\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
2.887104142522175e-07,\n        \"min\": 0.999998701296768,\n
\"max\": 0.9999999997034684,\n        \"num_unique_values\": 20,\n
\"samples\": [\n          0.9999999906579886,\n
0.9999999923440126,\n          0.9999999965829188\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Lookalike3\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 20,\n        \"samples\": [\n
\"C0191\",\n          \"C0179\",\n          \"C0050\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Score3\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
3.0047382559618736e-07,\n        \"min\": 0.99999868530761,\n
\"max\": 0.9999999994532176,\n        \"num_unique_values\": 20,\n
\"samples\": [\n          0.9999999901222912,\n
0.9999999872217884,\n          0.9999999904854152\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"pt"}
```