

Praktikum Data Mining

UAS

Faris Saifullah
3124640034

```
#no.1
```

```
import pandas as pd
```

```
dataset = pd.read_csv('transaction.csv')
```

```
dataset
```

	InvoiceNo	StockCode	Qty	InvoiceDate	CustomerID	Country
0	537626	22725	830	12/7/2010 14:57	12347	Iceland
1	537626	22729	948	12/7/2010 14:57	12347	Iceland
2	537626	22195	695	12/7/2010 14:57	12347	Iceland
3	542237	22725	636	1/26/2011 14:30	12347	Iceland
4	542237	22729	536	1/26/2011 14:30	12347	Iceland
...
10541	543911	21700	455	2/14/2011 12:46	17829	United Arab Emirates
10542	543911	22111	578	2/14/2011 12:46	17829	United Arab Emirates
10543	543911	22112	163	2/14/2011 12:46	17829	United Arab Emirates
10544	564428	23296	545	8/25/2011 11:27	17844	Canada
10545	564428	23294	643	8/25/2011 11:27	17844	Canada

Analisa :

menggunakan library pandas untuk membaca file CSV bernama transaction.csv dan menyimpannya ke dalam variabel dataset dalam bentuk DataFrame. Setelah itu, isi DataFrame ditampilkan untuk melihat data yang telah dimuat.

#no.2

```
dataset['InvoiceDate'] = pd.to_datetime(dataset['InvoiceDate'])
dataset['month'] = dataset['InvoiceDate'].dt.month
dataset['year'] = dataset['InvoiceDate'].dt.year
data = dataset[(dataset['year'] == 2011)][['InvoiceNo', 'Country', 'month', 'year']]
data.head()
```

	InvoiceNo	Country	month	year
3	542237	Iceland	1	2011
4	542237	Iceland	1	2011
5	542237	Iceland	1	2011
6	542237	Iceland	1	2011
7	542237	Iceland	1	2011

Analisa :

memproses data dari kolom InvoiceDate dalam DataFrame dataset. Kolom tersebut diubah menjadi format datetime menggunakan `pd.to_datetime`, lalu diekstrak bulan (month) dan tahun (year) dari data tersebut. Setelah itu, filter diterapkan untuk mengambil data hanya dari tahun 2011, dengan memilih kolom tertentu: InvoiceNo, Country, month, dan year.

#no.3
data_bulan = data[(data['month'] >= 1) & (data['month'] <= 12)]
data_bulan

	InvoiceNo	Country	month	year
3	542237	Iceland	1	2011
4	542237	Iceland	1	2011
5	542237	Iceland	1	2011
6	542237	Iceland	1	2011
7	542237	Iceland	1	2011
...
10541	543911	United Arab Emirates	2	2011
10542	543911	United Arab Emirates	2	2011
10543	543911	United Arab Emirates	2	2011
10544	564428	Canada	8	2011
10545	564428	Canada	8	2011

Analisa :
memfilter data pada DataFrame data untuk memastikan hanya baris dengan nilai month antara 1 dan 12 yang diambil (representasi bulan valid dalam setahun).

```
#no.4
transaksi = data_bulan.groupby('Country')['InvoiceNo'].nunique()
transaksi.head()
```

	InvoiceNo
Country	
Australia	41
Austria	12
Bahrain	1
Belgium	79
Brazil	1

Analisa :

mengelompokkan data dalam data_bulan berdasarkan kolom Country dan menghitung jumlah unik dari nomor faktur (InvoiceNo) untuk setiap negara.

```
#no.5
import pandas as pd
from scipy.cluster.hierarchy import linkage, fcluster
from sklearn.preprocessing import StandardScaler
transaksi_df = transaksi.reset_index()
scaler = StandardScaler()
scaled_data = scaler.fit_transform(transaksi_df[['InvoiceNo']])
cluster_result = linkage(scaled_data, method='average')
k = 3
transaksi_df['cluster'] = fcluster(cluster_result, k, criterion='maxclust')
transaksi_df.head()
```

	Country	InvoiceNo	cluster
0	Australia	41	1
1	Austria	12	1
2	Bahrain	1	1
3	Belgium	79	1
4	Brazil	1	1



Analisa :

melakukan klasterisasi hirarkis pada data transaksi. Data InvoiceNo dinormalisasi menggunakan StandardScaler, lalu diolah dengan metode linkage untuk menghasilkan matriks klaster. Data dikelompokkan menjadi tiga klaster (k=3) menggunakan fcluster, dan hasil klasterisasi disimpan dalam kolom baru cluster.

```
#no.6
centroid = transaksi_df.groupby('cluster')['InvoiceNo'].mean()
centroid
```

InvoiceNo	
cluster	
1	18.28125
2	340.00000
3	213.00000

Analisa :

menghitung rata-rata nilai InvoiceNo untuk setiap klaster dalam DataFrame transaksi_df. Data dikelompokkan berdasarkan kolom cluster, lalu rata-rata dihitung menggunakan fungsi mean().

```
#no.7  
sorted_centroid = centroid.sort_values(ascending=True)  
sorted_centroid
```

InvoiceNo	
cluster	
1	18.28125
3	213.00000
2	340.00000

Analisa :
mengurutkan nilai centroid dari setiap klaster dalam urutan naik (ascending) menggunakan sort_values().


```
#no.8
transaksi_tinggi = sorted_centroid.index[-1]
print(f"Cluster dengan transaksi tinggi: {transaksi_tinggi}")

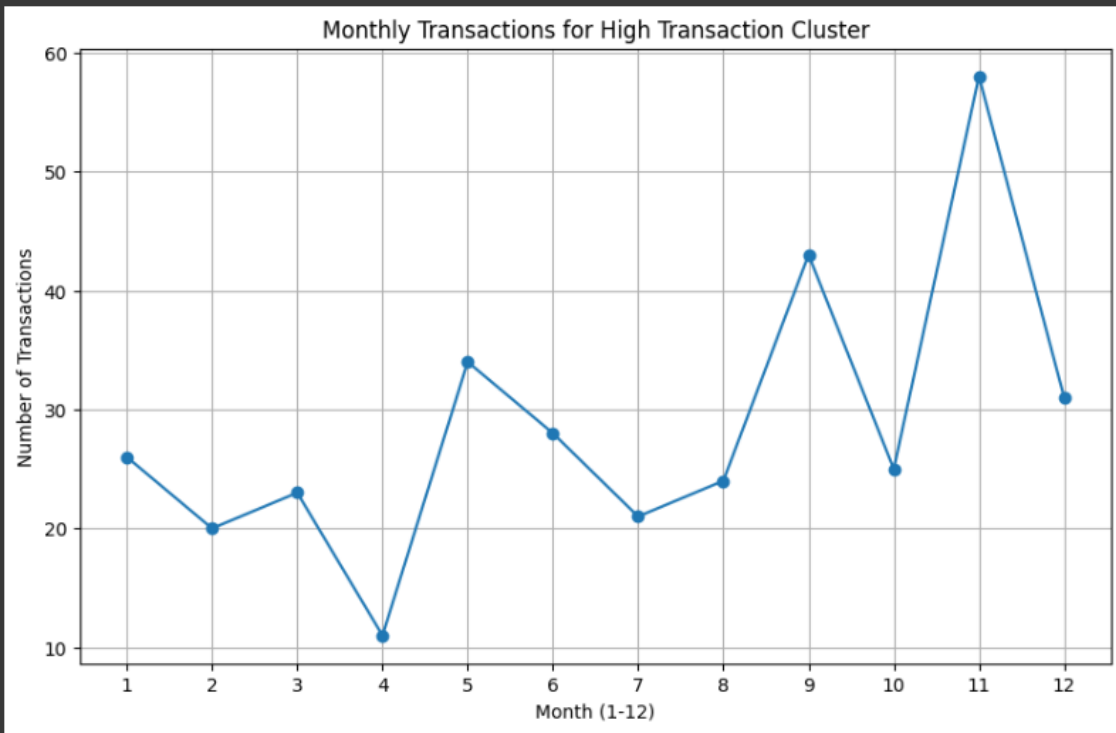
Cluster dengan transaksi tinggi: 2
```

Analisa :

mengambil indeks klaster terakhir dari `sorted_centroid`, yang mewakili klaster dengan rata-rata transaksi tertinggi. Indeks tersebut disimpan dalam variabel `transaksi_tinggi` dan ditampilkan menggunakan `print`.



```
#no.9
import matplotlib.pyplot as plt
monthly_transactions = dataset[dataset['Country'] == transaksi_df.loc[transaksi_df['cluster'] == transaksi_tinggi, 'Country'].iloc[0]].groupby('month')['InvoiceNo'].nunique()
plt.figure(figsize=(10, 6))
plt.plot(monthly_transactions.index, monthly_transactions.values, marker='o')
plt.xlabel('Month (1-12)')
plt.ylabel('Number of Transactions')
plt.title('Monthly Transactions for High Transaction Cluster')
plt.xticks(range(1, 13))
plt.grid(True)
plt.show()
```



Analisa :

menghasilkan grafik yang menunjukkan jumlah transaksi per bulan untuk negara dengan klaster transaksi tertinggi. Data difilter berdasarkan negara tersebut, kemudian dihitung jumlah transaksi unik per bulan. Dengan menggunakan matplotlib, kode ini menampilkan grafik garis yang memvisualisasikan tren transaksi bulanan, lengkap dengan label sumbu, judul, dan grid.

```
#no.10
import pandas as pd
from sklearn.linear_model import LinearRegression
X = monthly_transactions.index.values.reshape(-1, 1)
y = monthly_transactions.values
model = LinearRegression()
model.fit(X, y)
PredictedVal = model.predict([[1]])
print(f"Prediksi transaksi tertinggi untuk bulan januari 2012: {PredictedVal[0]}")
```

Prediksi transaksi tertinggi untuk bulan januari 2012: 18.3974358974359

Analisa :

menggunakan regresi linier untuk memprediksi jumlah transaksi pada bulan Januari 2012 berdasarkan data transaksi bulanan. Bulan (1-12) digunakan sebagai fitur, dan jumlah transaksi per bulan sebagai target. Model yang dilatih kemudian digunakan untuk memprediksi jumlah transaksi di bulan Januari 2012, dan hasil prediksi ditampilkan.