CUSTOMER SEGMENTANTION USING K-MEANS CLUSTERING

LINK GITHUB :

HTTPS://GITHUB.COM/HPRADITYA/PH_ML

SCENARIO

THIS DATA-SET CONTAINS TRANSACTION DATA IN A SUPERMARKET. YOU AS A DATA SCIENTIST AT THE RETAIL COMPANY ARE EXPECTED TO BE ABLE TO PROCESS THE DATA THAT HAS BEEN COLLECTED SO THAT IT BECOMES VALUABLE INSIGHT.

1. WHAT ITEMS CUSTOMERS BUY THE MOST AND ARE THERE ANY ITEMS WE SHOULD IGNORE?

- Untuk item yang customernya paling banyak beli, pertama groupby dahulu berdasarkan barang lalu dijumlahkan kuantitas demand tiap barang. Berikut adalah 5 barang dengan jumlah demand terbanyak :

Barang	Jumlah
MEDIUM CERAMIC TOP STORAGE JAR	78033.0
WORLD WAR 2 GLIDERS ASSTD DESIGNS	55047.0
JUMBO BAG RED RETROSPOT	48474.0
WHITE HANGING HEART T-LIGHT HOLDER	37891.0
POPCORN HOLDER	36761.0

1. WHAT ITEMS CUSTOMERS BUY THE MOST AND ARE THERE ANY ITEMS WE SHOULD IGNORE?

- Untuk item yang harus diabaikan, cari dulu nilai transaksi dari tiap barang dengan membuat column baru : Nilai_Transaksi berisi Jumlah * Harga_Satuan. Barang yang akan diabaikan adalah barang yang nilai transaksinya berada di urutan 5 terendah. Berikut adalah daftar barang yang akan diabaikan :

Barang	Nilai_Transaksi
PADS TO MATCH ALL CUSHIONS	0.003
HEN HOUSE W CHICK IN NEST	0.420
SET 12 COLOURING PENCILS DOILEY	0.650
VINTAGE BLUE TINSEL REEL	0.840
PINK CRYSTAL GUITAR PHONE CHARM	0.850

2. WHAT IS OUR STRATEGY TO INCREASE SALES BASED ON OUR EXPORT DESTINATION COUNTRIES?

Untuk melihat mana negara yang prospektif untuk export, perlu dilihat berapa jumlah customer yang ada di negara tsb. dan berapa nilai transaksi yang dihasilkan dari negara tsb. Berikut adalah table jumlah pelanggan dan nilai transaksi dari tiap negara.

Negara	Jumlah Pelanggan
United Kingdom	3919
Germany	94
France	88
Spain	30
Belgium	25

Negara	Nilai Transaksi
United Kingdom	8798801
Netherlands	284024
Ireland	281813
Germany	227911
France	209382

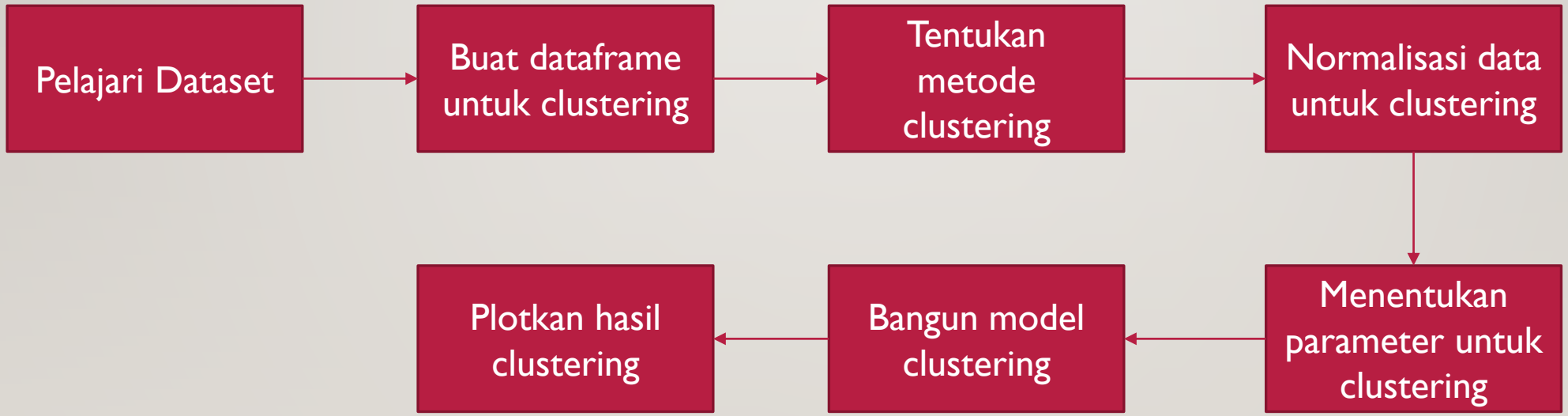
2. WHAT IS OUR STRATEGY TO INCREASE SALES BASED ON OUR EXPORT DESTINATION COUNTRIES?

Berdasarkan table diatas, didapatkan bahwa meskipun jumlah customer di Netherland dan Ireland tidak mencapai 5 besar, namun nilai transaksi yang dihasilkan dari customer di Netherland dan Ireland mencapai 5 besar.

Oleh karena itu saya sarankan untuk strategy salesnya perlu di expand untuk menambah jumlah customer di Netherland dan Ireland sambal tetap mempertahankan customer di United Kingdom, yang merupakan basis customer terbesar

3. DO CUSTOMER SEGMENTATION APPROPRIATELY. PLEASE EXPLAIN IN DETAIL AND COMPLETELY

Tahapan yang saya lakukan untuk customer segmentation adalah sebagai berikut :



3. DO CUSTOMER SEGMENTATION APPROPRIATELY. PLEASE EXPLAIN IN DETAIL AND COMPLETELY

Dari dataset ini didapatkan bahwa :

1. Ada Jumlah yang nilainya tidak positif
2. Ada row yang tidak terparsing dengan sempurna
3. Pembayaran yang sama bisa berisi barang berbeda dari customer yang sama
4. Sebagian besar customer berasal dari United Kingdom

Untuk clustering saya menggunakan data numeris berupa Jumlah dan Harga_Satuan, karena jumlah negara terlalu banyak untuk menjadi data category

3.DO CUSTOMER SEGMENTATION APPROPRIATELY. PLEASE EXPLAIN IN DETAIL AND COMPLETELY

Untuk customer segmentation ini saya menggunakan K-means clustering karena ini adalah salah satu metode clustering yang mudah untuk diimplementasikan.

Sebelum membangun model clustering, saya menormalisasikan datanya agar memiliki range dan distribusi yang sama.

Parameter yang saya gunakan untuk K-means clustering adalah :

1. jumlah cluster = 3
2. Jumlah iterasi = 10

3. DO CUSTOMER SEGMENTATION APPROPRIATELY. PLEASE EXPLAIN IN DETAIL AND COMPLETELY

Berikut adalah scatterplot Jumlah dan Harga_Satuan berdasarkan cluster :

3. DO CUSTOMER SEGMENTATION APPROPRIATELY. PLEASE EXPLAIN IN DETAIL AND COMPLETELY

- Dari scatterplot diatas, terlihat bahwa clusteringnya mengikuti besarnya nilai transaksi yang berupa Jumlah * Harga_Satuan, secara umum dapat dikatakan bahwa :
- Cluster 0 = nilai transaksi rendah, berisi customer dari 38 negara
- Cluster 1 = nilai transaksi tinggi , berisi customer dari 1 negara
- Cluster 2 = nilai transaksi sedang, berisi customer dari 11 negara

GET THE HIDDEN INSIGHT FROM THE DATA

Cluster_Num	Jumlah	Harga_Satuan	Nilai_Transaksi
0	9.776962	3.668951	18.234101
1	74215.000000	1.040000	77183.600000
2	606.585132	325.945755	1927.962110

4. GET THE HIDDEN INSIGHT FROM THE DATA

Berikut adalah table perbandingan rata-rata jumlah barang dibeli, harga satuan tiap barang, dan nilai transaksi dari ketiga cluster hasil K-Means clustering .

Cluster_Num	Jumlah	Harga_Satuan	Nilai_Transaksi
0	9.776962	3.668951	18.234101
1	74215.000000	1.040000	77183.600000
2	606.585132	325.945755	1927.962110

Dari table diatas, dapat terlihat bahwa meskipun harga satuan dari cluster 1 lebih rendah dari cluster 0. Namun karena jumlahnya yang sangat banyak, nilai transaksi di cluster 1 dapat melebihi rata-rata nilai transaksi cluster 2

4. GET THE HIDDEN INSIGHT FROM THE DATA

Cluster 1 hanya berisi 1 transaksi dari 1 customer. Berikut adalah profil dari customer di cluster 1. Skala jumlah pembelian yang sangat banyak ini patut menjadi perhatian. Customer ini perlu dihubungi lebih lanjut karena berpotensi menjadi saluran ke market baru yang berupa grosir. Sehingga perusahaan bisa menggarap market grosir dari customer ini dan juga market retail dari customer lainnya

Column	Value
Kode_Bayar	541431
Kode_Barang	23166
Barang	MEDIUM CERAMIC TOP STORAGE JAR
Jumlah	74215
Tanggal_Transaksi	1/18/2011 10:01
Harga_Satuan	1.04
Kode_Pelanggan	12346
Negara	United Kingdom
Nilai_Transaksi	77183.6
Cluster_Num	1