Rakamin Academy

UNSUPERVISED LEARNING

Homework 16
Kelompok UNO:
1. Daniel Rowin
2. Febby Maghfirani Aziz
3. Amodya Subagio
4. Nur Rahman Shalahudin
5. Fakhry Abdurrohman
6. Daviro Yota Nagasan Wahyudi
7. Putri Vina Fajriyani
8. Asmiyeni Islamiati

Estimasi Waktu Pengerjaan

 2 - 3 jam

Jumlah Soal

 4 Bagian

Total Point

 100 poin

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok**, dengan output berupa:
   a. File .ipynb yang berisi hasil analisis/modeling
   b. Sebuah presentasi .pdf yang berisi ringkasan dari poin-poin penting yang dijabarkan pada slide-slide selanjutnya

2. Homework ini berupa soal clustering **end-to-end pertanyaan bisnis** dimana teman-teman akan diberikan sebuah dataset mentah berisi data customer. Teman-teman diharapkan melakukan 4 hal berikut dengan menggunakan dataset tersebut
   - EDA
   - Feature Engineering
   - Modeling + Evaluasi
   - Interpretasi model + Rekomendasi

3. Upload hasil pengerjaanmu melalui LMS dengan format nama file sebagai berikut **Nama_Lengkap_Batch_XX** dalam format .html (cara save dalam format .html disini)

Dataset yang digunakan

Airline Customer Value Analysis Case

- **Deskripsi**:

  Dataset ini berisi data customer sebuah perusahaan penerbangan dan beberapa fitur yang dapat menggambarkan value dari customer tersebut.

- **Data**:

  Setiap baris mewakili customer, setiap kolom berisi atribut customer.

- **Link download disini**

Deskripsi Dataset

Code	Description
MEMBER_NO-b	: ID Member
FFP_DATE	: Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	: Tanggal Penerbangan pertama
GENDER	: Jenis Kelamin
FFP_TIER	: Tier dari Frequent Flyer Program
WORK_CITY	: Kota Asal
WORK_PROVINCE	: Provinsi Asal
WORK_COUNTRY	: Negara Asal
AGE	: Umur Customer
LOAD_TIME	: Tanggal data diambil
FLIGHT_COUNT	: Jumlah penerbangan Customer
BP_SUM	: Rencana Perjalanan
SUM_YR_1	: Fare Revenue
SUM_YR_2	: Votes Prices
SEG_KM_SUM	: Total jarak(km) penerbangan yg sudah dilakukan
LAST_FLIGHT_DATE	: Tanggal penerbangan terakhir
LAST_TO_END	: Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
AVG_INTERVAL	: Rata-rata jarak waktu
MAX_INTERVAL	: Maksimal jarak waktu
EXCHANGE_COUNT	: Jumlah penukaran
avg_discount	: Rata rata discount yang didapat customer
Points_Sum	: Jumlah poin yang didapat customer
Point_NotFlight	: point yang tidak digunakan oleh members

Instruksi Detil

1. Lakukan EDA pada dataset untuk mendapatkan pemahaman umum mengenai data dan memandu proses feature engineering (20 poin)

Langkah-langkah:

a. Pastikan setiap kolom dataset memiliki tipe data yang tepat, tidak ada data kosong, bebas dari duplikat, dan berada di *range* value yang tepat
b. Keluarkan statistik kolom baik numerik maupun kategorikal, cari bentuk distribusi setiap kolom (numerik), dan jumlah baris untuk setiap *unique* value (kategorikal)
c. Cari tahu apakah ada kolom-kolom yang berkorelasi kuat satu sama lain

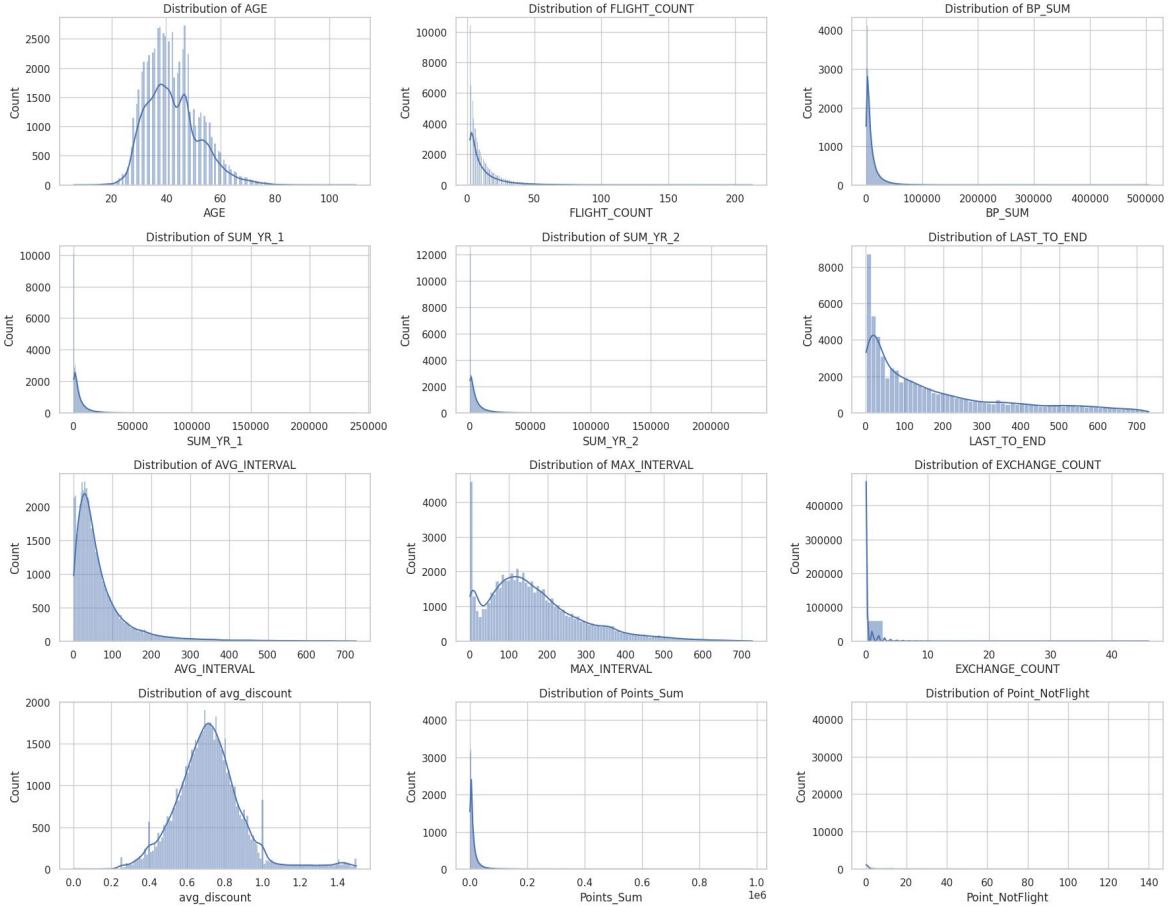
UnǦuk mempermudah kamu, yuk lihaǦ resource di bawah ini:

- Topic Machine Learning Preparation - EDA

30-45 menit

Exploratory Data Analysis

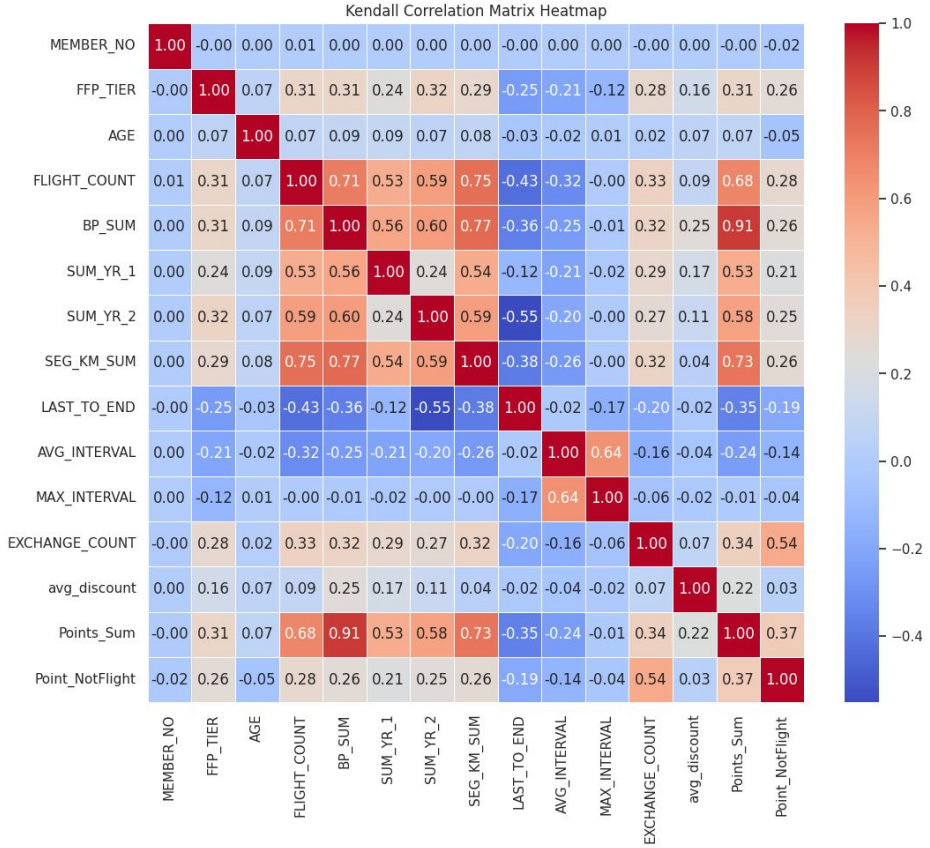
Rakamin Academy



Exploratory Data Analysis

- Analisis Statistik Kolom Numerik:
  - AGE: Mayoritas pelanggan berusia antara 35 dan 48 tahun.
  - FLIGHT_COUNT: Bervariasi luas, menunjukkan campuran penerbang sering dan sesekali.
  - SEG_KM_SUM (Jarak total terbang): Menunjukkan distribusi miring ke kanan, menyarankan sebagian besar pelanggan terbang jarak pendek.
  - AVG_INTERVAL (Waktu rata-rata antara penerbangan): Miring ke kanan, banyak pelanggan terbang tidak sering.
  - avg_discount: Sebagian besar berkumpul sekitar 0,6 hingga 0,8 berbentuk distribusi normal.
  - Points_Sum: Sangat bervariasi, dengan sebagian besar pelanggan memiliki poin lebih sedikit.
- Analisis Data Kategorik:
  - GENDER: Mayoritas adalah Pria (48,134) dibandingkan Wanita (14,851).
  - FFP_TIER: Sebagian besar pelanggan ada di Tingkatan 4 (58,066), diikuti Tingkatan 5 (3,409) dan Tingkatan 6 (1,513).
  - WORK_CITY, WORK_PROVINCE, WORK_COUNTRY: Nilai yang beragam, dengan guangzhou, beijing, dan shanghai sebagai kota yang umum.
- Validasi Data: Tidak ada catatan duplikat. Beberapa kolom memiliki nilai yang kosong, terutama WORK_CITY (2269 kosong), WORK_PROVINCE (3248 kosong), WORK_COUNTRY (26 kosong), dan AGE (420 kosong). SUM_YR_1 dan SUM_YR_2 juga memiliki nilai yang kosong (551 dan 138, masing-masing).

Exploratory Data Analysis



- **FFP_TIER** menunjukkan korelasi positif yang moderat dengan **FLIGHT_COUNT**, **BP_SUM**, **SUM_YR_1**, **SUM_YR_2**, **SEG_KM_SUM**, dan **Points_Sum**. Hal ini menunjukkan bahwa tingkatan yang lebih tinggi dalam program frequent flyer berhubungan dengan jumlah penerbangan yang lebih banyak, jarak tempuh yang lebih jauh, pendapatan tarif yang lebih tinggi, dan lebih banyak poin yang terakumulasi.
- **FLIGHT_COUNT** memiliki korelasi yang kuat dengan **BP_SUM**, **SUM_YR_1**, **SUM_YR_2**, **SEG_KM_SUM**, dan **Points_Sum**. Penumpang yang sering terbang cenderung memiliki pendapatan tarif yang lebih tinggi, lebih banyak poin, dan menjelajahi jarak yang lebih panjang.
- **BP_SUM**, **SUM_YR_1**, **SUM_YR_2**, dan **SEG_KM_SUM** memiliki korelasi internal yang kuat, menunjukkan bahwa anggota yang menghabiskan lebih banyak cenderung juga bepergian lebih banyak baik dalam hal jarak maupun jumlah penerbangan.
- **LAST_TO_END** dan **AVG_INTERVAL** memiliki korelasi negatif yang moderat dengan aktivitas penerbangan, mengindikasikan bahwa anggota yang belum terbang baru-baru ini atau yang memiliki interval waktu lebih lama antar penerbangan cenderung memiliki jumlah penerbangan dan poin yang terakumulasi lebih rendah.
- **EXCHANGE_COUNT** menunjukkan korelasi positif dengan tingkatan dan aktivitas penerbangan, menunjukkan bahwa anggota yang lebih aktif dalam program cenderung lebih sering menggunakan poin mereka untuk pertukaran.

2. **Pilih fitur-fitur yang menurut teman-teman masuk akal secara bisnis untuk digunakan sebagai fitur clustering. Lakukan feature engineering! (20 poin)**

Langkah-langkah:

a. Dari sekian banyak kolom yang ada, tentukan 3-6 fitur untuk digunakan sebagai fitur *clustering*. **Tulis alasan teman-teman memilih fitur tersebut.**

b. **Lakukan preprocessing dan feature engineering** (apabila fitur yang teman-teman pilih merupakan fitur baru yang dihasilkan dari fitur-fitur yang sudah ada).

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Machine Learning Preparation - Feature Engineering

30-45 menit

Feature Engineering

Feature yang digunakan:

1. **AGE** (umur pengguna): sebab umur itu merupakan **ciri-ciri pengguna** yang penting untuk diperhatikan dalam data
2. Untuk dari **sisi teknis penerbangan**, banyak aspek yang bisa dilihat, dari aspek waktu (AVG_INTERVAL), aspek jumlah (FLIGHT_COUNTS), aspek jarak (SEG_KM_SUM). Kita akan menggunakan **AVG_INTERVAL** dan **SEG_KM_SUM,** sebab lebih merepresentasikan dari segi waktu dan juga dari segi jarak**.**
3. Untuk dari **segi bisnis**, terdapat beberapa fitur yang bisa digunakan, namun akan digunakan **EXCHANGE_COUNT** dan **SUM_YR_1.** Hal ini sebab exchange count akan jelas mewakili jumlah transaksi yang terjadi, dan sum year 1 akan memberikan total revenue yang didapatkan dari masing-masing pengguna.

Rakamin Academy

3. Lakukan clustering K-means! Temukan jumlah cluster yang menurut teman-teman optimal dan evaluasi cluster yang dihasilkan dengan visualisasi dan silhouette score (30 poin)

Langkah-langkah:

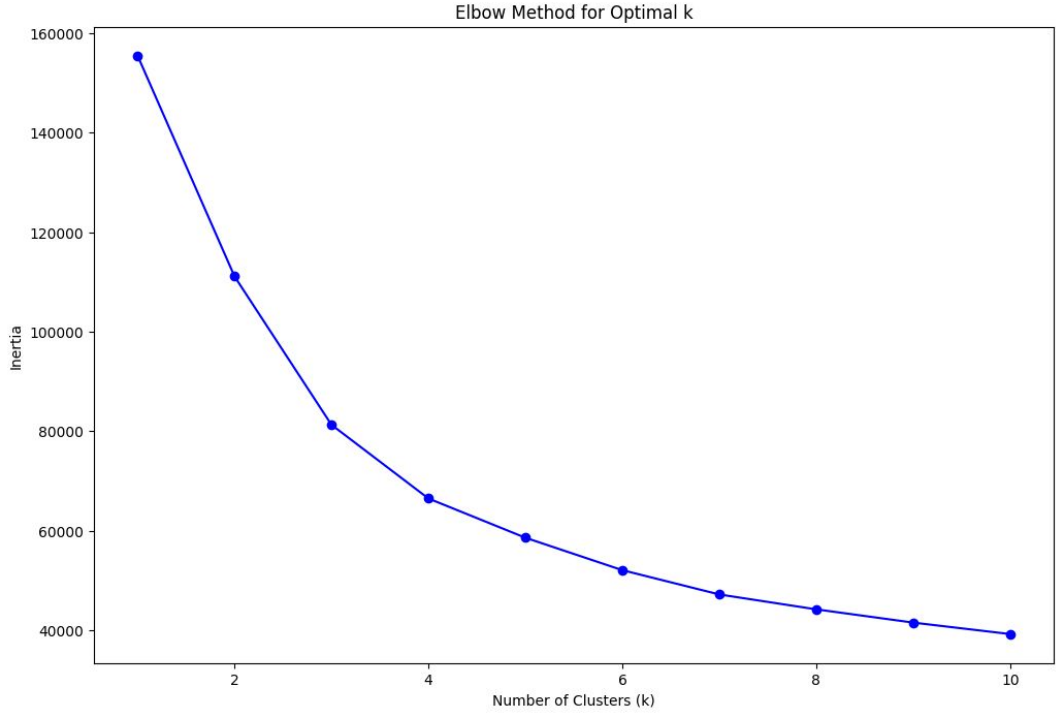
a. Temukan jumlah cluster yang optimal dengan menggunakan elbow method
b. Lakukan clustering menggunakan K-means
c. Evaluasi cluster yang dihasilkan dengan menggunakan visualisasi, gunakan PCA apabila diperlukan

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Unsupervised Learning - Clustering bagian KMeans dan Evaluasi

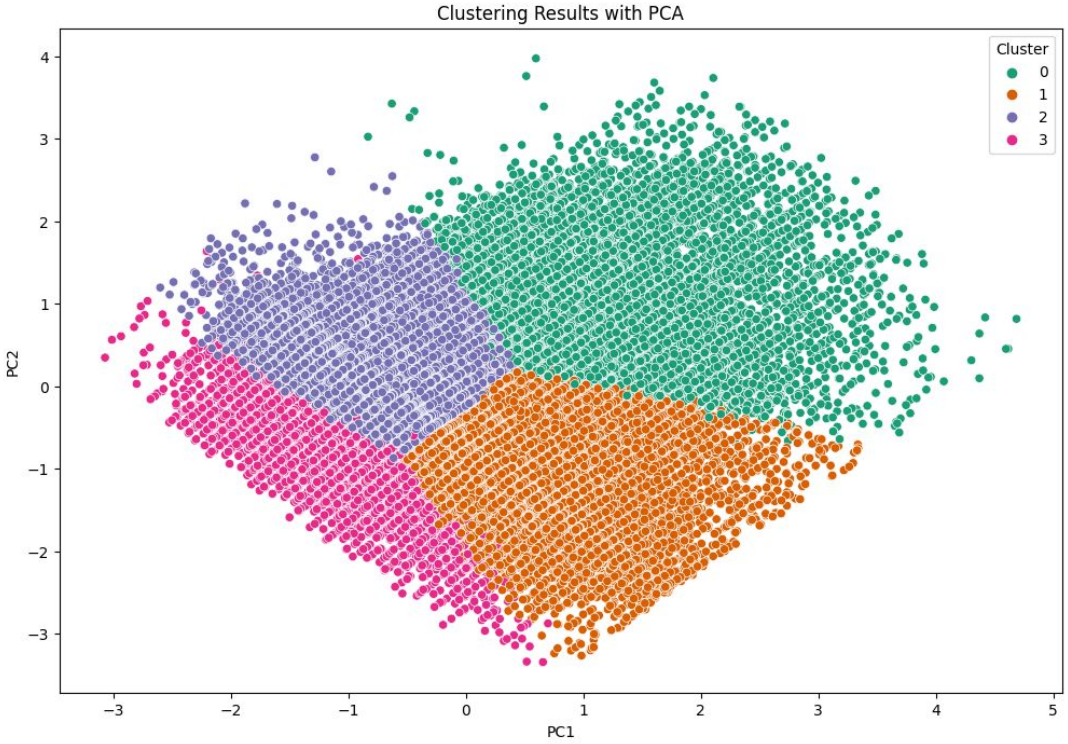
20-30 menit

Hasil Clustering K-Means



- Dengan menggunakan Elbow Method, maka jumlah cluster yang bisa digunakan adalah 4 cluster.

Hasil Clustering K-Means



- Visualisasi di samping dengan metode PCA, terlihat bahwasanya cluster telah terbagi menjadi 4 cluster yang dilabeli cluster 0 sampai 3.
- Silhouette Score yang didapatkan sebesar 0,31.

4. Interpretasi cluster yang dihasilkan secara bisnis dan berikan rekomendasi yang sesuai dengan cluster yang dihasilkan (30 poin)

Langkah-langkah:

a. Tempelkan kembali label yang dihasilkan ke dataframe asal, dan keluarkan statistik fitur dari setiap cluster

b. Deskripsikan secara kontekstual customer seperti apa yang ada di masing-masing cluster

c. Berdasarkan cluster tersebut, berikan 1-2 rekomendasi bisnis

Untuk mempermudah kamu, yuk lihat resource di bawah ini:

- Topic Unsupervised Learning - Clustering

20-30 menit

Interpretasi Bisnis

Cluster 0: Pelanggan Teratur yang Mapan

1. Ukuran: 13,086 anggota
2. Usia: Usia menengah (rata-rata 43 tahun), dengan rentang usia yang moderat.
3. Kebiasaan Penerbangan: Mereka memiliki interval rata-rata antara penerbangan yang relatif singkat dan terbang jarak jauh, menunjukkan kebiasaan terbang yang teratur.
4. Keterlibatan Finansial: Pendapatan tarif rata-rata yang tinggi, menunjukkan kontribusi finansial yang kuat terhadap maskapai.
5. Keterlibatan Loyalitas: Jumlah tukar poin yang moderat, menunjukkan keterlibatan yang wajar dengan program loyalitas maskapai.

Interpretasi Bisnis

Cluster 1: Pelanggan Tua yang Jarang Terbang

1. Ukuran: 18,407 anggota
2. Usia: Kelompok usia yang lebih tua (rata-rata 52 tahun). Kebiasaan Penerbangan: Mereka terbang kurang sering dengan interval sedang antara penerbangan dan menempuh jarak lebih pendek.
3. Keterlibatan Finansial: Pendapatan tarif yang lebih rendah, menunjukkan mereka rata-rata menghabiskan lebih sedikit untuk penerbangan.
4. Keterlibatan Loyalitas: Paling tidak terlibat dengan program loyalitas seperti yang ditunjukkan oleh jumlah tukar poin yang rendah.

Interpretasi Bisnis

Cluster 2: Pelanggan Muda yang Sering Terbang

1. Ukuran: 26,782 anggota
2. Usia: Demografi yang lebih muda (rata-rata 36 tahun), yang secara signifikan lebih rendah dari cluster lain.
3. Kebiasaan Penerbangan: Mereka memiliki interval reguler antara penerbangan tetapi menempuh jarak lebih pendek, yang bisa menunjukkan penerbangan jarak dekat yang sering.
4. Keterlibatan Finansial: Pendapatan tarif sedang, yang mungkin mencerminkan perilaku yang sadar anggaran atau mencari nilai.
5. Keterlibatan Loyalitas: Variabilitas jumlah tukar poin yang mengejutkan tinggi, berpotensi menunjukkan beberapa anggota yang sangat terlibat tetapi umumnya rendah keterlibatannya.

Interpretasi Bisnis

Cluster 3: Pelanggan Sesekali

1. Ukuran: 7,601 anggota
2. Usia: Muda hingga usia menengah (rata-rata 40 tahun), dengan rentang usia yang luas.
3. Kebiasaan Penerbangan: Mereka memiliki interval terpanjang antara penerbangan dan terbang jarak terpendek.
4. Keterlibatan Finansial: Pendapatan tarif terendah, yang sejalan dengan perilaku terbang yang jarang mereka lakukan.
5. Keterlibatan Loyalitas: Jumlah tukar poin sangat rendah, menunjukkan keterlibatan minimal dengan program loyalitas.

Rekomendasi Bisnis

Untuk **Cluster 0: Pelanggan Teratur yang Mapan**

Hadiah Fokus: Tawarkan hadiah yang mendorong peningkatan belanja yang berkelanjutan, seperti diskon volume atau peningkatan tier.
Pemasaran Personalisasi: Gunakan strategi pemasaran yang dipersonalisasi untuk mempertahankan tingkat keterlibatan dan pengeluaran mereka yang tinggi.

Untuk **Cluster 1: Pelanggan Tua yang Jarang Terbang**

Kampanye Retargeting: Implementasikan kampanye untuk menarget ulang demografi ini, mungkin dengan penawaran yang menarik bagi pelanggan tua, seperti peningkatan kenyamanan atau paket liburan.
Kesadaran Program Loyalitas: Tingkatkan kesadaran tentang manfaat program loyalitas untuk mendorong keterlibatan yang lebih besar.

Rekomendasi Bisnis

Untuk **Cluster 2: Pelanggan Muda yang Sering Terbang**

Penawaran Pelanggan Muda: Buat penawaran yang melayani pelanggan muda (di bawah 39 tahun), seperti diskon untuk perjalanan petualangan atau kemitraan dengan merek yang berorientasi pada pemuda.
Program Keterlibatan: Dorong keterlibatan program loyalitas dengan tantangan yang digamifikasi atau promosi media sosial.

Untuk **Cluster 3: Pelanggan Sesekali**

Inisiatif Frekuensi: Kembangkan insentif bagi pelanggan ini untuk terbang lebih sering, seperti menawarkan penerbangan gratis setelah jumlah perjalanan tertentu.
Komunikasi Khusus: Gunakan komunikasi langsung tentang bagaimana program loyalitas dapat memberi manfaat bahkan bagi pelanggan sesekali.

Rakamin Academy

TERIMA KASIH