RFM Analysis

Muhammad Farkhan | Data Analyst Project - 2025

1 Business Understanding

Sebagai bisnis online retail dengan basis customer yang besar di UK, memahami perilaku setiap customer merupakan hal yang krusial. Pemahaman ini diperlukan untuk memastikan tercapainya:

1. Program loyalitas pelanggan yang efektif.
2. Strategi retensi tepat sasaran.

Oleh karena itu, RFM Analysis menjadi salah satu metode yang tepat digunakan untuk mencapai tujuan tersebut.

2 Data Understanding

Dataset ini diperoleh dari **UC Irvine Machine Learning Repository**, berisi data transaksional online retail (hadiah unik untuk segala acara) yang berbasis di **UK** dalam rentang waktu **01/12/2010 hingga 09/12/2011**. Dataset ini berjumlah **541.910 baris** dan beberapa kolom, yaitu:

- InvoiceNo
- StockCode
- Description
- Quantity
- InvoiceDate
- UnitPrice
- CustomerID
- Country

UC Irvine
Machine Learning
Repository

3 Data Preparation

Proses persiapan data dilakukan menggunakan MySQL (via DBeaver). Query ini mencakup dua langkah utama:

- Seleksi data: Hanya mengambil kolom yang relevan.
- Feature engineering: Membuat fitur baru, yaitu revenue.

```sql
select
    customerid,
    invoiceno,
    invoicedate,
    sum(quantity * unitprice) as revenue,
    country
from
    online_retail
group by
    customerid,
    invoiceno,
    invoicedate,
    country
having
    sum(quantity * unitprice) > 0
order by
    customerid asc,
    revenue desc
```

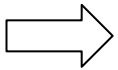
SQL - Extraction

Data cleaning melakukan beberapa hal, diantaranya:

1. **Handling missing values** (menghapus baris customerid null):
    - Customerid adalah kunci unik untuk segmentasi, sehingga tidak bisa diimputasi dengan mean atau median. Jumlahnya relatif kecil (7.1%) sehingga aman dihapus tanpa merusak representasi data. Tidak diimputasi dengan kode bisnis, karena tidak ada bisnis yang relevan dengan proyek.
2. **Deduplikasi data**: Mengecek dan menghapus data duplikat (jika ada) untuk menjaga akurasi jumlah transaksi.
3. **Konversi tipe data**: Mengubah customerid menjadi int dan invoicedate menjadi datetime (untuk menghitung nilai recency).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20002 entries, 0 to 20001
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   customerid   18562 non-null  float64
 1   invoiceno    20002 non-null  object
 2   invoicedate  20002 non-null  object
 3   revenue      20002 non-null  float64
 4   country      20002 non-null  object
dtypes: float64(2), object(3)
memory usage: 781.5+ KB
```

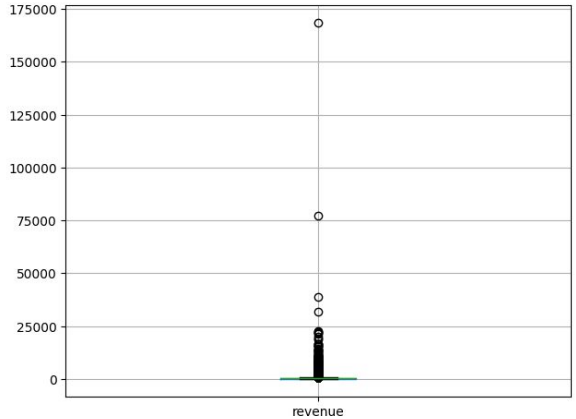
Before



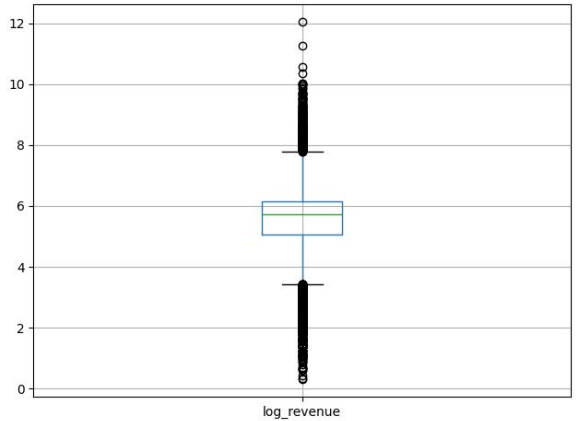
```
<class 'pandas.core.frame.DataFrame'>
Index: 18562 entries, 1440 to 20001
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   customerid   18562 non-null  int64
 1   invoiceno    18562 non-null  object
 2   invoicedate  18562 non-null  datetime64[ns]
 3   revenue      18562 non-null  float64
 4   country      18562 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(2)
memory usage: 870.1+ KB
```

After

Box plot pada tahap EDA menunjukkan distribusi right-skewed akibat adanya outlier. Nilai ini tetap dipertahankan karena valid dan merepresentasikan customer bernilai tinggi. Setelah dilakukan log transformation, pola asli terlihat jelas mayoritas customer memiliki transaksi menengah, sedangkan outlier teridentifikasi sebagai segmen spesifik yang sangat bernilai.



Before (Raw Data)



After (Log Transform)

Heatmap menunjukkan korelasi **0.55** antara Total Order dan Total Revenue. Ini menandakan hubungan positif, di mana customer yang sering bertransaksi cenderung memberikan revenue besar. Temuan ini menjadi dasar validasi untuk penyederhanaan model segmentasi di tahap selanjutnya.

Python - EDA

4 RFM Segmentation

Mentransformasi data transaksi menjadi metrik RFM setiap customer sebagai basis segmentasi.

	customerid	Recency	Frequency	Monetary
0	12346	326	1	77183.60
1	12347	3	7	4310.00
2	12348	76	4	1797.24
3	12349	19	1	1757.55
4	12350	311	1	334.40

Menggunakan quantile untuk membagi customer ke dalam 5 peringkat (skor 1-5). Skor ini kemudian digabungkan menjadi RFM_Score (contoh: 555) untuk memudahkan pengelompokan segmen.

	customerid	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
0	12346	326	1	77183.60	1	1	5	115
1	12347	3	7	4310.00	5	5	5	555
2	12348	76	4	1797.24	2	4	4	244
3	12349	19	1	1757.55	4	1	4	414
4	12350	311	1	334.40	1	1	2	112

Python - RFM

Tahap akhir pemrosesan data di Python adalah menerjemahkan skor RFM menjadi 9 segmen yang dapat ditindak lanjuti (actionable) menggunakan logika Regex. Berikut adalah beberapa kategorinya:

1. Top Tier: Champions & Loyal Customer (aset utama)
2. Potential: Potential Loyalist (F/M), New Customer, Promising (target pertumbuhan).
3. At Risk: At Risk & Needs Attention (prioritas retensi).
4. Churn: Lost (non prioritas).

	customerid	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score	Segment
0	12346	326	1	77183.60	1	1	5	115	Lost
1	12347	3	7	4310.00	5	5	5	555	Champions
2	12348	76	4	1797.24	2	4	4	244	At Risk
3	12349	19	1	1757.55	4	1	4	414	Potential Loyalists (M)
4	12350	311	1	334.40	1	1	2	112	Lost

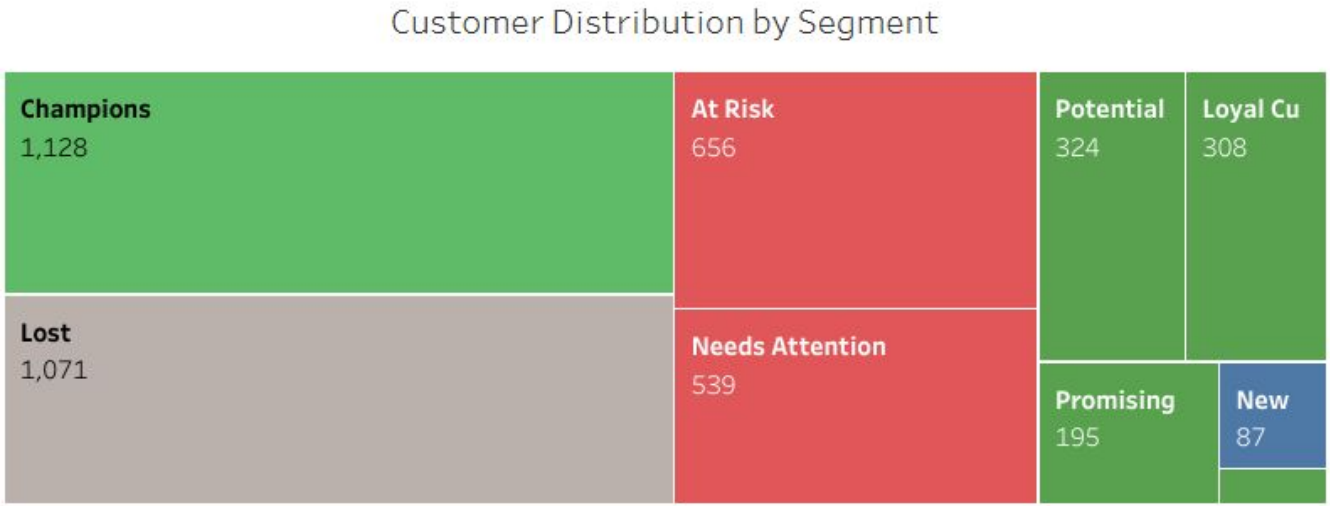
5 Visualization

Total Customers	Total Revenue	Total Orders	Average Revenue
4,338	£8,911,408	18,532	£2,054

Insight:

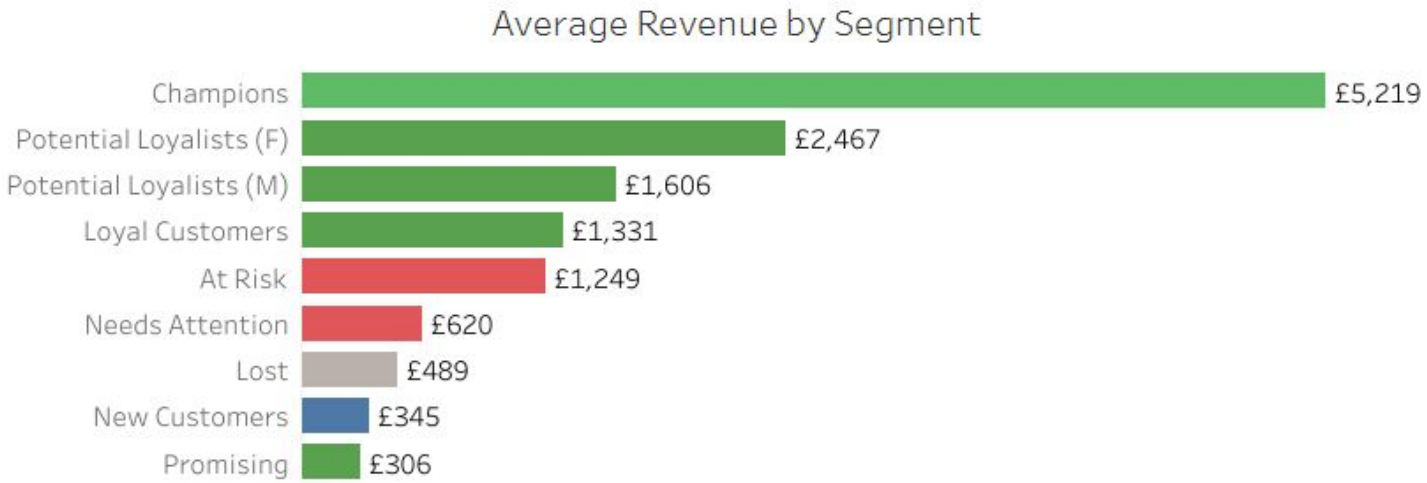
- Data mencakup 4.338 customer, jumlah yang sangat cukup untuk memberikan gambaran mengenai customer behavior saat ini.
- Dengan total pendapatan mencapai £8.9 Juta dari 18.000+ pesanan, bisnis menunjukkan aktivitas perdagangan yang sangat sehat dan aktif.
- Average revenue £2,054 ini menjadi sebuah acuan standar. Customer yang belanja di atas angka ini adalah aset prioritas, sedangkan yang jauh di bawahnya adalah peluang pertumbuhan.

Tableau - Visualization



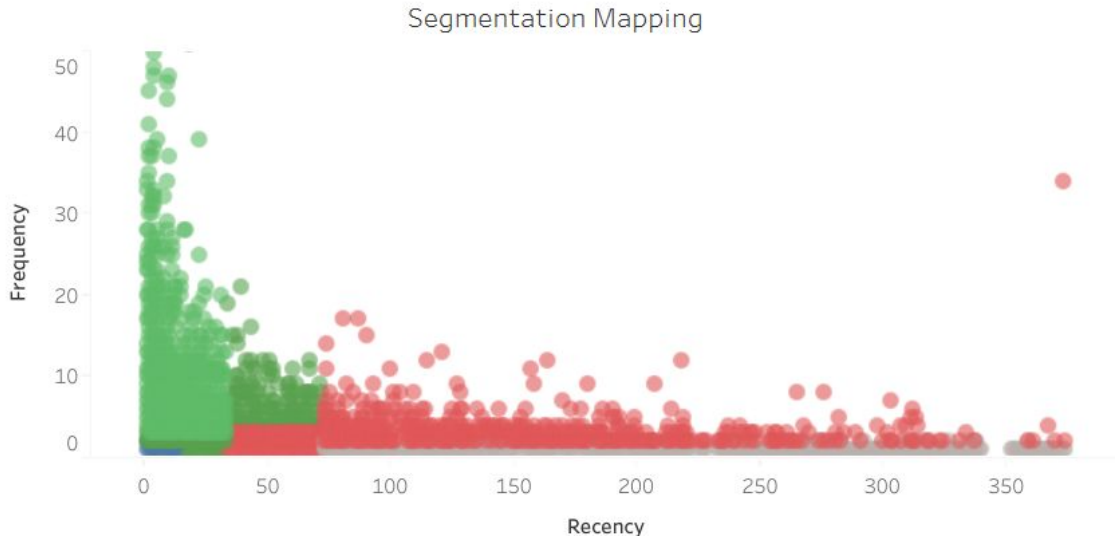
Insight:

- Basis customer didominasi oleh segmen Champions dan Lost, mengindikasikan kondisi polarisasi antara loyalitas tinggi dan tingkat churn yang besar.
- At Risk menempati urutan ketiga terbesar (656). Jumlah ini menjadi sinyal peringatan akan terjadinya churn lanjutan pada customer bernilai tinggi.
- Besarnya populasi At Risk menuntut prioritas strategi retensi untuk mencegah perpindahan ke segmen Lost.



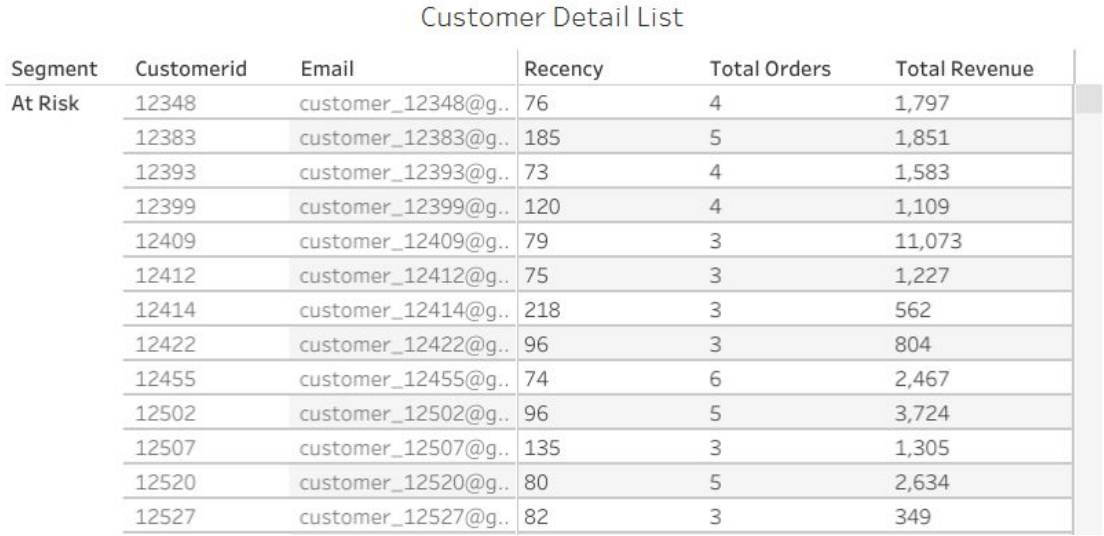
Insight:

- Segmen Champions memiliki rata-rata belanja tertinggi (£5,219), memvalidasi peran mereka sebagai penyumbang pendapatan utama perusahaan.
- Temuan krusial terlihat pada At Risk dengan rata-rata belanja £1,249. Nilai ini 3,6x lipat lebih tinggi dibandingkan rata-rata New Customers (£345).
- Data ini mengonfirmasi bahwa customer At Risk adalah aset bernilai tinggi. Kehilangan segmen ini akan berdampak finansial yang jauh lebih signifikan dibandingkan kegagalan mendapatkan customer baru.



Insight:

- Plot ini membuktikan bahwa logika segmentasi berhasil. Segmen Champions (hijau) dan Lost (abu-abu) terpolarisasi di sudut yang berlawanan.
- Segmen At Risk (merah) terlihat jelas di zona bahaya (kanan atas). Posisinya menunjukkan customer dengan frekuensi transaksi tinggi, namun recencynya semakin membesar (lama tidak kembali).
- Ketepatan plot ini memvalidasi bahwa segmentasi ini berhasil dan siap digunakan untuk decision making.



Customer Detail List

Segment	Customerid	Email	Recency	Total Orders	Total Revenue
At Risk	12348	customer_12348@g..	76	4	1,797
	12383	customer_12383@g..	185	5	1,851
	12393	customer_12393@g..	73	4	1,583
	12399	customer_12399@g..	120	4	1,109
	12409	customer_12409@g..	79	3	11,073
	12412	customer_12412@g..	75	3	1,227
	12414	customer_12414@g..	218	3	562
	12422	customer_12422@g..	96	3	804
	12455	customer_12455@g..	74	6	2,467
	12502	customer_12502@g..	96	5	3,724
	12507	customer_12507@g..	135	3	1,305
	12520	customer_12520@g..	80	5	2,634
	12527	customer_12527@g..	82	3	349

Visualisasi tabel ini bertujuan untuk menjadi alat kerja operasional dengan menyediakan data target yang siap dieksekusi.

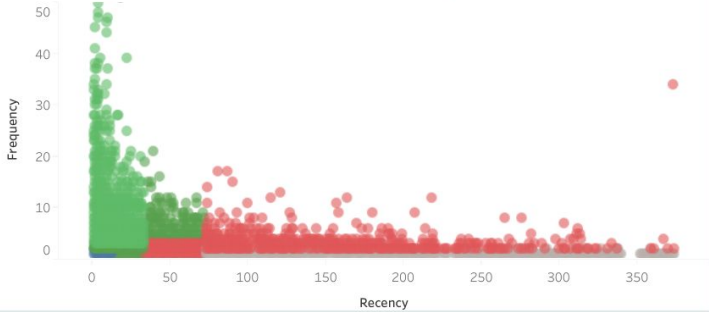
- Tabel secara dinamis hanya menampilkan customer dari segmen yang dipilih (misal: At Risk).
- Kolom email ditambahkan (berdasarkan customerid) untuk mensimulasikan data siap pakai bagi kebutuhan tim marketing atau lainnya.
- Daftar target dapat langsung diekspor ke Excel/CSV untuk eksekusi.

RFM Customer Segmentation Analysis

Total Customers	Total Revenue	Total Orders	Average Revenue
4,338	£8,911,408	18,532	£2,054

Customer Distribution by Segment

Segmentation Mapping



Customer Detail List

Segment	Customerid	Email	Recency	Total Orders	Total Revenue
At Risk	12348	customer_12348@g..	76	4	1,797
	12383	customer_12383@g..	185	5	1,851
	12393	customer_12393@g..	73	4	1,583
	12399	customer_12399@g..	120	4	1,109
	12409	customer_12409@g..	79	3	11,073
	12412	customer_12412@g..	75	3	1,227
	12414	customer_12414@g..	218	3	562
	12422	customer_12422@g..	96	3	804
	12455	customer_12455@g..	74	6	2,467
	12502	customer_12502@g..	96	5	3,724
	12507	customer_12507@g..	135	3	1,305
	12520	customer_12520@g..	80	5	2,634
	12527	customer_12527@g..	82	3	349

Average Revenue by Segment

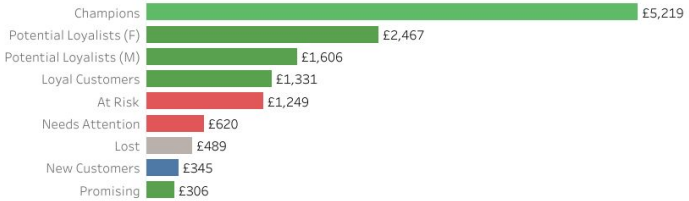


Tableau - Visualization

5 Recommendations

1. **Retention Strategy (At Risk & Needs Attention)**.
   - ➢ Tindakan: Melakukan kampanye win back yang personal dan agresif (diskon khusus atau notifikasi pengingat).
   - ➢ Tujuan: Mencegah churn pada customer bernilai tinggi.
2. **Appreciation Strategy (Champions & Loyal Customer)**.
   - ➢ Tindakan: Memberikan layanan VIP, akses awal produk baru, atau reward loyalitas (jangan berikan diskon).
   - ➢ Tujuan: Menjaga kepuasan supaya tidak pindah ke kompetitor.
3. **Growth Strategy (New Customer, Promising, & Potential Loyalists)**.
   - ➢ Tindakan: Berikan penawaran produk (cross-selling) dan insentif untuk pembelian berikutnya.
   - ➢ Tujuan: Mendorong frekuensi belanja agar naik kelas menjadi customer loyal.
4. **Efficiency Strategy (Lost)**.
   - ➢ Tindakan: Hentikan alokasi biaya marketing dan iklan untuk segmen ini.
   - ➢ Tujuan: Efisiensi anggaran untuk dialihkan ke segmen prioritas (retensi & apresiasi).

6 Conclusion

1. **Transformasi Data ke Strategi**

   Analisis ini berhasil mengubah data transaksi mentah menjadi 9 segmen customer yang jelas dan tervalidasi. Perusahaan kini tidak lagi bergantung pada strategi yang memukul rata semuanya.

2. **Efisiensi Anggaran**

   Dengan mengidentifikasi segmen At Risk dan Lost, perusahaan dapat mengalokasikan anggaran marketing hanya ke target yang memberikan ROI maksimal.

3. **Kesiapan Tindak Lanjut**

   Dashboard yang dibangun bukan sekadar laporan pasif, melainkan alat kerja operasional yang menyediakan data target (actionable list) untuk eksekusi kampanye CRM.

THANK YOU!

Muhammad Farkhan