EXPLORATORY DATA ANALYSIS (EDA)

of E-Commerce Transactions in the UK

14 May, 2025

Digital Skill Fair 39.0 - Data Science

HELLO!

Saya Hilma Zahra Qorina

Saya seorang mahasiswa Universitas Negeri Surabaya, jurusan Teknik Informatika. Saya memiliki semangat untuk terus belajar dan mengembangkan diri di bidang Data Science, khususnya dalam analisis data dan pengembangan berbagai model berbasis Machine Learning.

Saya percaya bahwa data adalah aset berharga, sehingga saya terus mendalami eksplorasi, preprocessing, dan implementasi machine learning untuk menghasilkan model yang efektif dan andal. Melalui portofolio ini, saya ingin membagikan antusiasme saya terhadap Data Science.

Digital Skill Fair 39.0 - Data Science

ibimbing

AGENDA OVERVIEW

Digital Skill Fair 39.0 - Data Science

01 Data Overview

02 Handling Missing Value

03 Handling Duplicate Data

04 Results

05 Visualization

E-COMMERCE UK DATASET

Dataset E-Commerce UK merupakan dataset yang disediakan oleh Atharva Arya di Kaggle yang berisi kumpulan data transaksi dari sebuah perusahaan retail online yang berbasis di Inggris. Data ini mencakup aktivitas penjualan yang terjadi antara bulan Desember 2010 hingga Desember 2011.

Dataset ini berisi lebih dari 500.000 baris data dan 8 kolom.

Dataset ini sering digunakan untuk latihan Exploratory Data Analysis (EDA) karena berisi data transaksi nyata yang cukup lengkap dan bervariasi. Dengan informasi seperti tanggal transaksi, jumlah barang, harga, dan negara asal pelanggan, dataset ini memungkinkan analisis berbagai aspek bisnis seperti tren penjualan, perilaku pelanggan, dan produk terlaris.

Selain itu, adanya data yang tidak bersih seperti nilai yang hilang atau duplikat juga menjadikannya cocok untuk latihan pembersihan dan visualisasi data.

TOOLS

Sumber Dataset

kaggle

Bahasa Pemrograman

Library

pandas

matplotlib

DATA OVERVIEW

5 baris pertama dan terakhir dari dataset. Dimulai dari Desember 2010 hingga Desember 2011.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

DATA OVERVIEW

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
df['CustomerID'] = df['CustomerID'].astype(str)
```

✓ 0.3s

	Quantity	InvoiceDate	UnitPrice
count	541909.000000	541909	541909.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114
min	-80995.000000	2010-12-01 08:26:00	-11062.060000
25%	1.000000	2011-03-28 11:34:00	1.250000
50%	3.000000	2011-07-19 17:17:00	2.080000
75%	10.000000	2011-10-19 11:27:00	4.130000
max	80995.000000	2011-12-09 12:50:00	38970.000000
std	218.081158	NaN	96.759853

Mengubah datatype ‘InvoiceDate’ menjadi datetime, dan datatype ‘CustomerID’ menjadi string.

df.describe()

Untuk memahami karakteristik dataset sebelum melakukan pemodelan lebih lanjut, perlu dilakukan statistik deskriptif. Output dari df.describe() menampilkan berbagai metrik penting, seperti mean (rata-rata), std (standar deviasi), min & max (nilai minimum dan maksimum), serta 25%, 50%, dan 75% yang mewakili persentil atau kuartil.

DATA OVERVIEW

df.info()

Dataset memiliki 541.909 baris dan 8 kolom.

Tipe data kolom bervariasi: ada object (teks), numerik (int64 dan float64), serta datetime.

- Kolom InvoiceNo, StockCode, Description, dan Country bertipe object karena berisi data teks.
- Kolom Quantity bertipe int64, berisi jumlah produk yang dibeli.
- Kolom UnitPrice dan CustomerID bertipe float64, berisi angka desimal.
- Kolom InvoiceDate bertipe datetime64[ns], karena berisi tanggal dan waktu transaksi.

Terdeteksi Missing Value pada kolom Description.

Dataset menggunakan sekitar 33.1 MB memori.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   541909 non-null  object
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.1+ MB
```

HANDLING MISSING VALUE

df.isna().sum()

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID         0
Country            0
dtype: int64
```

Data pada kolom Description tidak memiliki nilai (null atau kosong) sebanyak 1454 baris.

Mengisi data yang hilang pada setiap kolom dalam DataFrame.

```python
for column in df.columns:
    if df[column].dtype == 'object':
        df[column].fillna(df[column].mode()[0], inplace=True)
    else:
        df[column].fillna(df[column].mean(), inplace=True)
```

✓ 0.2s

Jika kolom bertipe objek (teks atau kategori), nilai kosong diisi dengan nilai yang paling sering muncul (modus). Jika kolom bertipe numerik, nilai kosong diisi dengan rata-rata (mean). Tujuannya adalah untuk memastikan data bersih sebelum dianalisis atau digunakan dalam model.

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Setelah penangannan Missing Value, sudah tidak ada data yang tidak memiliki nilai (null atau kosong).

HANDLING DUPLICATE DATA

```
check_duplicate = df.duplicated().sum()

print(f"Jumlah data yang duplikat = {check_duplicate}")
```

✓ 0.2s

```
Jumlah data yang duplikat = 5268
```

Cek duplikat data pada dataset, terdeteksi 5268 data duplikat

```
df = df.drop_duplicates()
```

✓ 0.2s

Hapus data duplikat

```
Jumlah data yang duplikat = 0
```

Setelah menghapus semua data duplikat

RESULTS

df.info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 536641 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    536641 non-null  object
 1   StockCode    536641 non-null  object
 2   Description  536641 non-null  object
 3   Quantity     536641 non-null  int64
 4   InvoiceDate  536641 non-null  datetime64[ns]
 5   UnitPrice    536641 non-null  float64
 6   CustomerID   536641 non-null  object
 7   Country      536641 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 36.8+ MB
```

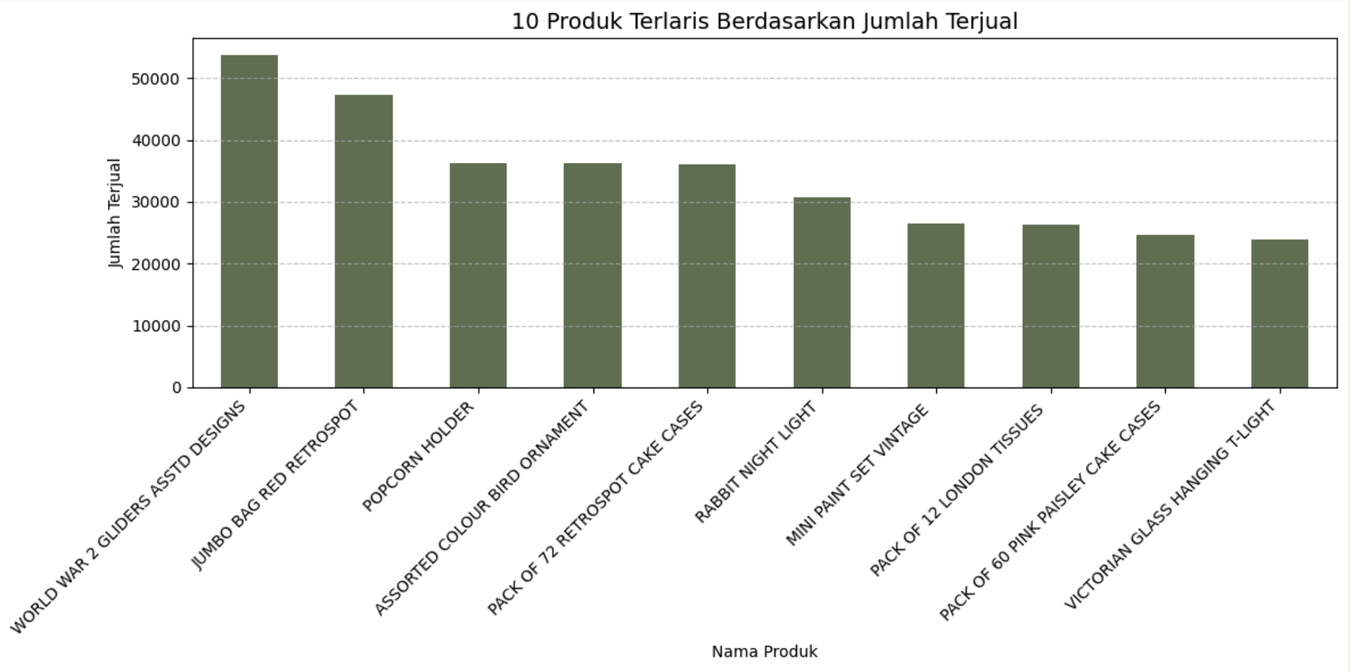
df.describe()

	Quantity	InvoiceDate	UnitPrice
count	536641.000000	536641	536641.000000
mean	9.620029	2011-07-04 08:57:06.087421952	4.632656
min	-80995.000000	2010-12-01 08:26:00	-11062.060000
25%	1.000000	2011-03-28 10:52:00	1.250000
50%	3.000000	2011-07-19 14:04:00	2.080000
75%	10.000000	2011-10-18 17:05:00	4.130000
max	80995.000000	2011-12-09 12:50:00	38970.000000
std	219.130156	NaN	97.233118

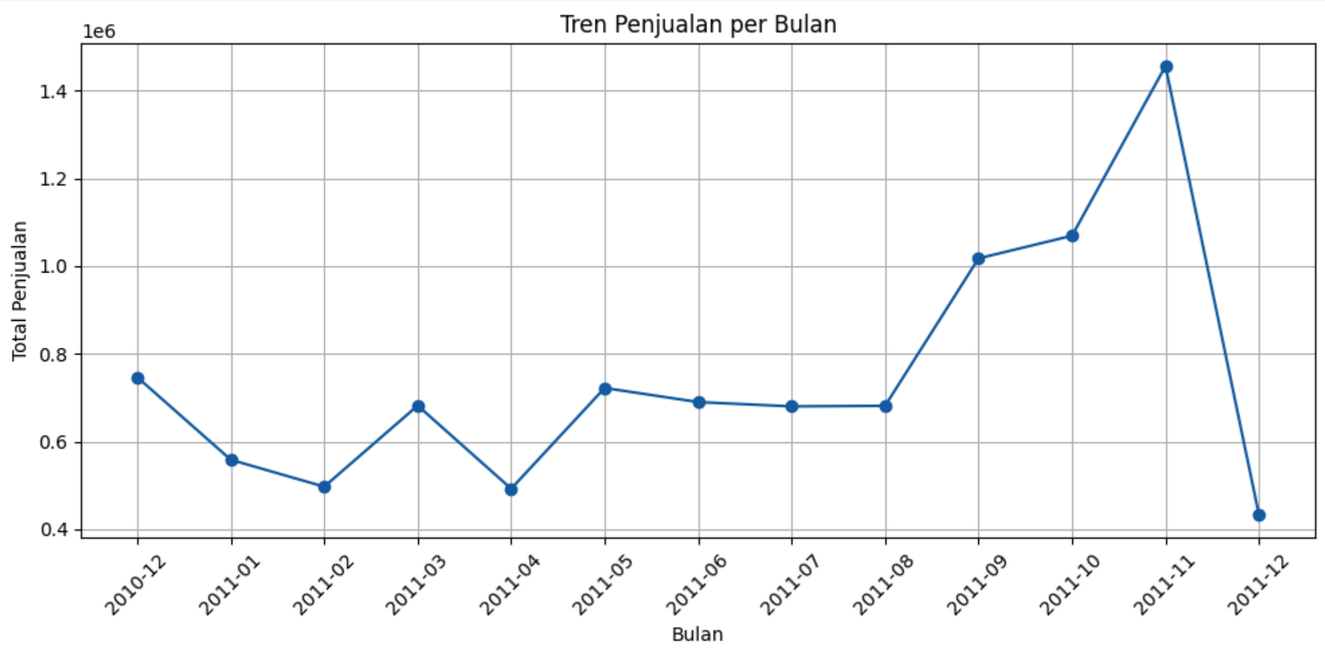
Jumlah data menjadi 536641 baris tanpa missing value dan duplikat

PRODUK TERJUAL TERBANYAK

Grafik menampilkan 10 produk dengan jumlah penjualan tertinggi. Produk paling laris adalah WORLD WAR 2 GLIDERS ASSTD DESIGNS yang terjual sebanyak 53.847.

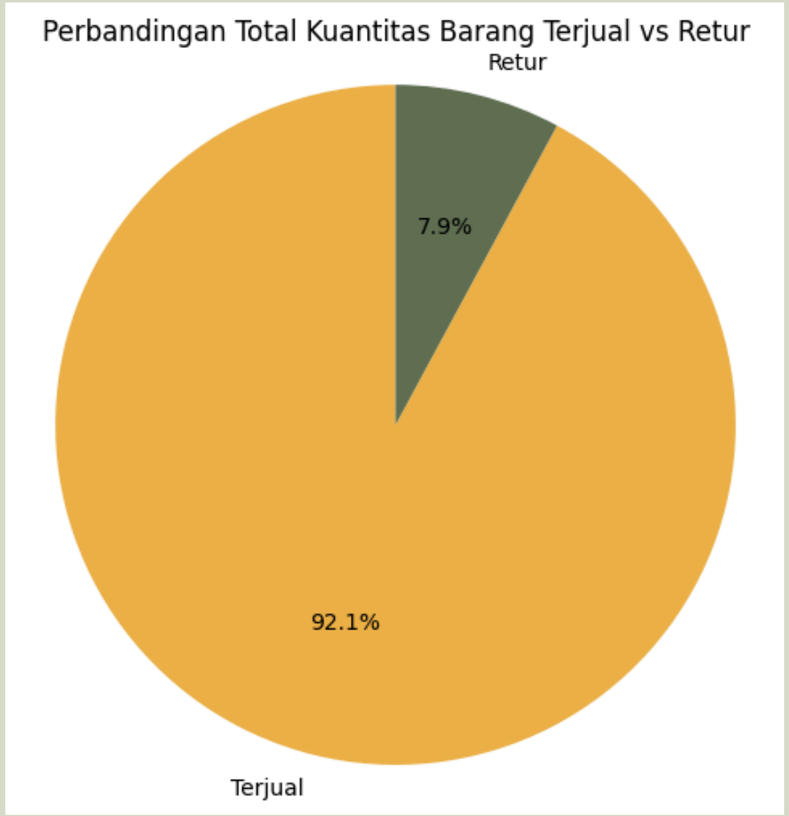


TREN PENJUALAN



Grafik menunjukkan pola penjualan per bulan dari Desember 2010 hingga Desember 2011 yang cenderung fluktuatif. Di awal tahun 2011 terjadi penurunan, lalu naik turun dengan lonjakan kecil pada Maret dan Mei. Penjualan mulai meningkat signifikan sejak September, mencapai puncak tertinggi pada November 2011. Namun, terjadi penurunan tajam di Desember 2011 ke titik terendah sepanjang periode.

TERJUAL VS RETUR

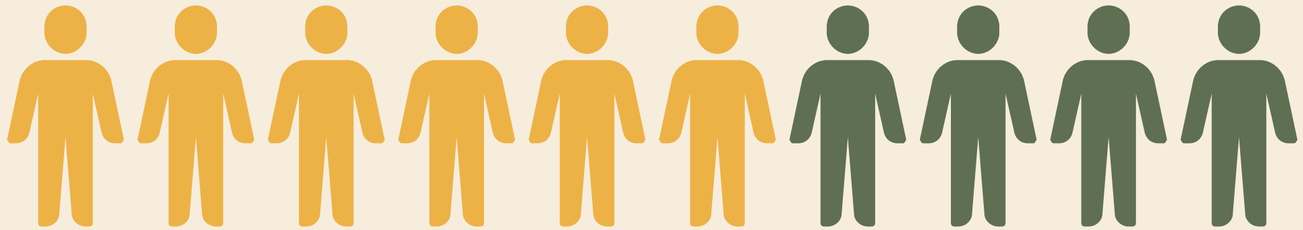


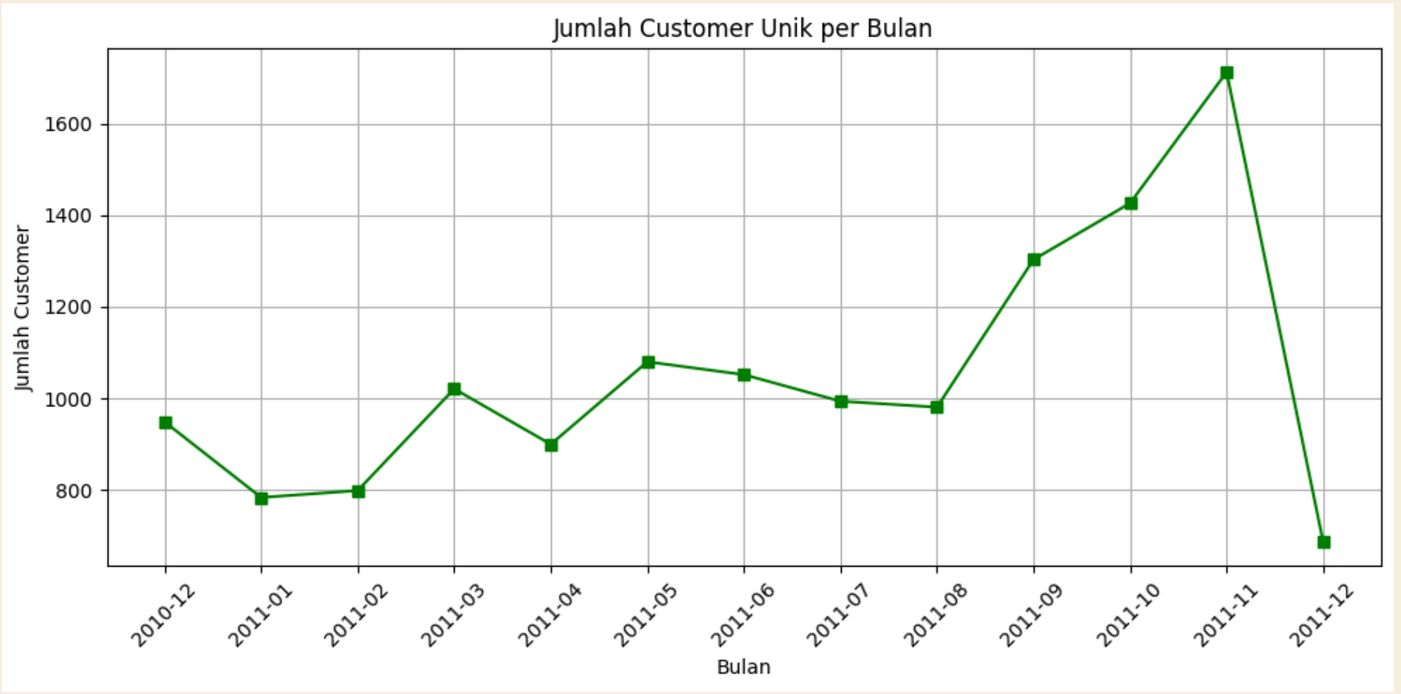
Sebagian besar barang (lebih dari 90%) yang dicatat dalam transaksi adalah barang yang berhasil dijual.
Ini menandakan penjualan berjalan dengan baik dan sebagian besar transaksi adalah penjualan.

Namun, sekitar 7.9% dari total kuantitas adalah barang yang diretur oleh pelanggan.
Ini cukup kecil, tapi tetap penting diperhatikan karena retur bisa menunjukkan adanya masalah seperti kualitas produk, kesalahan pengiriman, atau ketidaksesuaian pesanan.

TREN PELANGGAN

Grafik menunjukkan pola fluktuatif jumlah customer unik per bulan dari Desember 2010 hingga Desember 2011. Di awal tahun 2011 terjadi penurunan, namun mulai Maret hingga Agustus jumlah customer cenderung meningkat meskipun tidak stabil. Lonjakan signifikan terjadi dari September hingga mencapai puncak tertinggi pada November 2011. Namun, pada Desember 2011 jumlah customer turun drastis ke titik terendah sepanjang tahun.





ibimbing

Digital Skill Fair 39.0 - Data Science

THANK YOU

14 May, 2025