



UNIVERSITAS  
NEGERI SURABAYA

# OPTIMASI MODEL ANALISIS REGRESI PENJUALAN PRODUK BERDASARKAN PENDAPATAN KOTOR SUPERMARKET

presented by group 2A



# OUR TEAM



rakha alcander

22031554052



ilham warmandev

22031554049



azaria syahla

22031554019



dhani aditya

22031554038



# INTRODUCTION

Supermarket merupakan salah satu industri ritel yang tentunya perlu memahami perilaku dari konsumen dan faktor-faktor yang memengaruhi keputusan pembelian yang akan dijadikan suatu indikator untuk meningkatkan hasil penjualan dan kepuasan pelanggan. Salah satu cara untuk menganalisis perilaku konsumen adalah melalui penilaian atau rating yang diberikan oleh pelanggan terhadap produk yang mereka beli. Rating customer ini tidak hanya mencerminkan tingkat kepuasan konsumen, tetapi juga dapat digunakan sebagai indikator kualitas produk dan layanan yang disediakan oleh supermarket. Oleh karena itu, pada projek ini kami akan mengeksplorasi data penjualan dan rating customer untuk mengidentifikasi pola dan faktor-faktor utama yang



# KAJIAN PUSTAKA

## PCA (PRINCIPAL COMPONENT ANALYSIS)

Principal Component Analysis (PCA) adalah metode statistik yang digunakan untuk mereduksi dimensi data dengan cara mengubah sejumlah besar variabel menjadi beberapa variabel utama yang disebut komponen utama. Metode ini berguna untuk mengidentifikasi pola dalam data yang memiliki korelasi tinggi antar variabel. PCA tidak hanya mengurangi jumlah variabel tetapi juga mempertahankan sebanyak tersebut.

## DECISION TREE REGRESSION

DT regression merupakan salah satu teknik dalam machine learning yang digunakan untuk memprediksi nilai kontinu dengan cara membagi data ke dalam subset yang lebih kecil, membentuk struktur pohon yang setiap cabangnya merepresentasikan keputusan berdasarkan atribut tertentu. DT regression bekerja dengan membagi dataset berdasarkan nilai atribut yang memaksimalkan perbedaan antara subset data pada setiap langkah pembagian. Setiap node internal dalam pohon merepresentasikan tes pada atribut, setiap cabang adalah hasil dari tes tersebut, dan nantinya setiap daun (leaf) akan ditentukan berdasarkan subset data pada node tersebut.



# KAJIAN PUSTAKA

## ANALISIS KORELASI PEARSON

Korelasi Pearson adalah salah satu metode korelasi statistik yang ditemukan oleh Karl Pearson dengan tujuan untuk menentukan nilai keidentikan dua variabel yang terbentang secara linear. Metode ini kerap digunakan untuk menentukan sejauh mana signifikansi perubahan satu variabel terhadap variabel lain dengan menerka bagaimana nilai variabel tersebut ditarik terhadap garis linear.

## ANALISIS KORELASI SPEARMAN

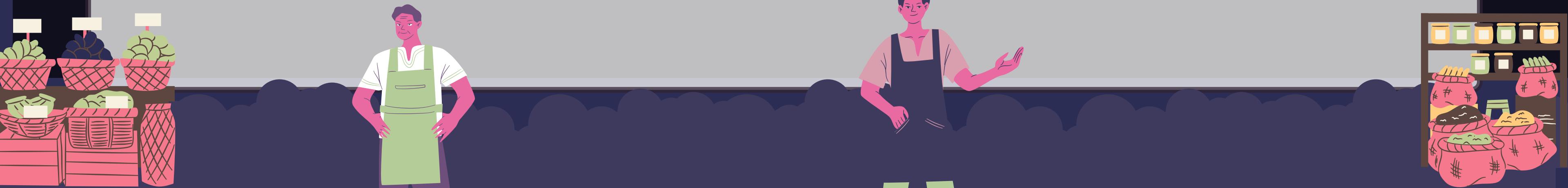
Korelasi Spearman adalah salah satu metode korelasi yang ditemukan oleh Charles Spearman dengan tujuan untuk menentukan nilai keidentikan dua variabel tidak menggunakan nilai raw data dan memfokuskan korelasi tersebut terhadap nilai peringkat. Pengaplikasian metode tersebut dinilai lebih baik pada keadaan semacam variabel yang non-parametrik dan diasumsikan tidak mengikuti distribusi tertentu.



# KAJIAN PUSTAKA

## REGRESI LINEAR

Regresi dalam statistik adalah alat analitis yang digunakan untuk memahami dan memperkirakan hubungan antara variabel dependen (Y) dan variabel independen (X). Dengan mengeksplorasi bagaimana perubahan pada variabel independen mempengaruhi variabel dependen, analisis regresi membantu mengidentifikasi variabel-variabel yang berpengaruh signifikan. Hubungan ini biasanya diekspresikan dalam bentuk fungsi linier. Selain itu, metrik evaluasi seperti RMSE (Root Mean Squared Error) dan MAE (Mean Absolute Error) digunakan untuk menilai akurasi model prediksi yang dikembangkan, memastikan prediksi yang lebih akurat.



# METODE PENELITIAN

## SELEKSI FITUR DAN REDUKSI DIMENSI

- PCA
- Analisis Korelasi (pearson & Spearman)

## TRANFORMASI DATA

- Normalisasi data

## MODEL REGRESI

- Decision Tree Regression
- Regresi Linear



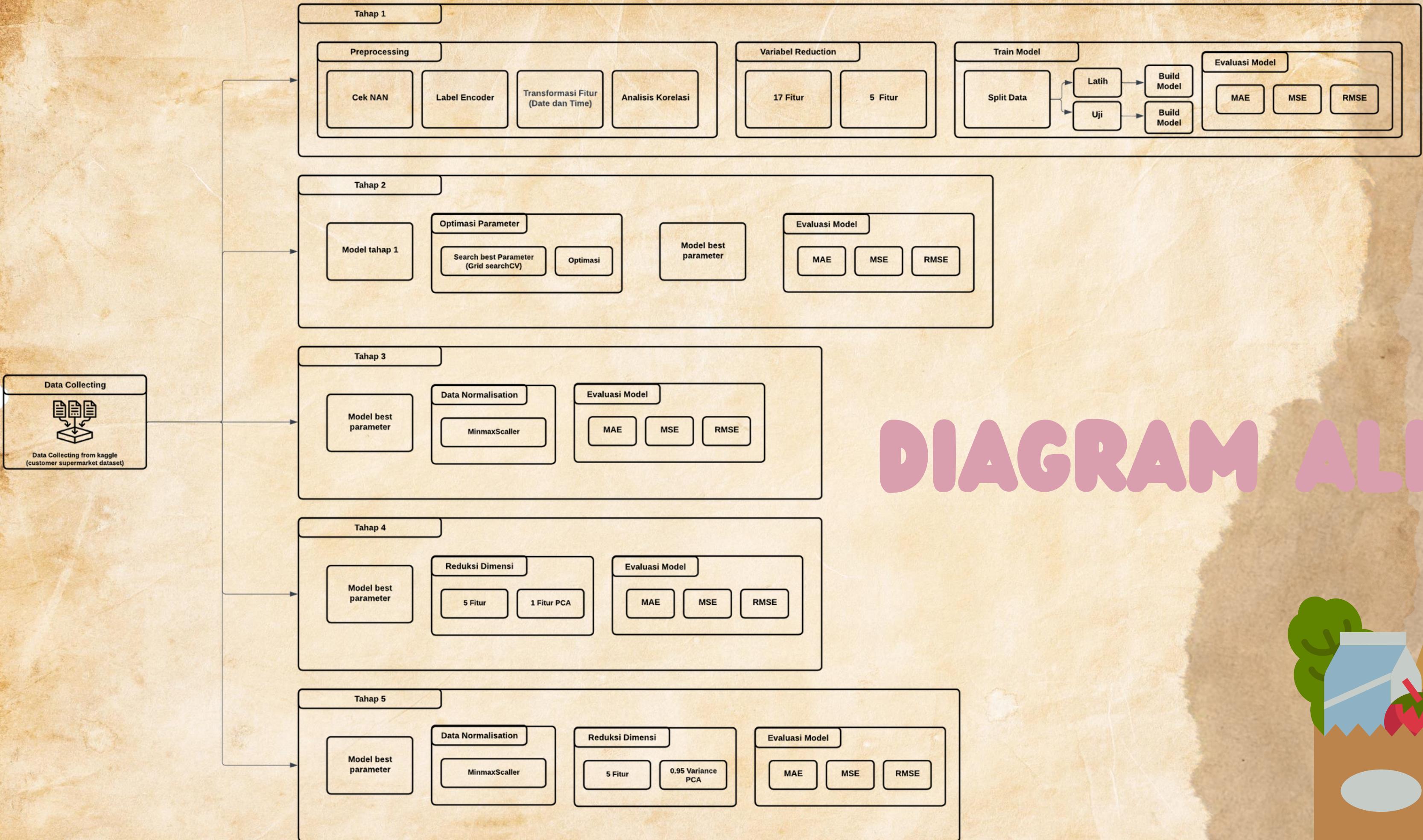
# DATASET

**CLICK ME**

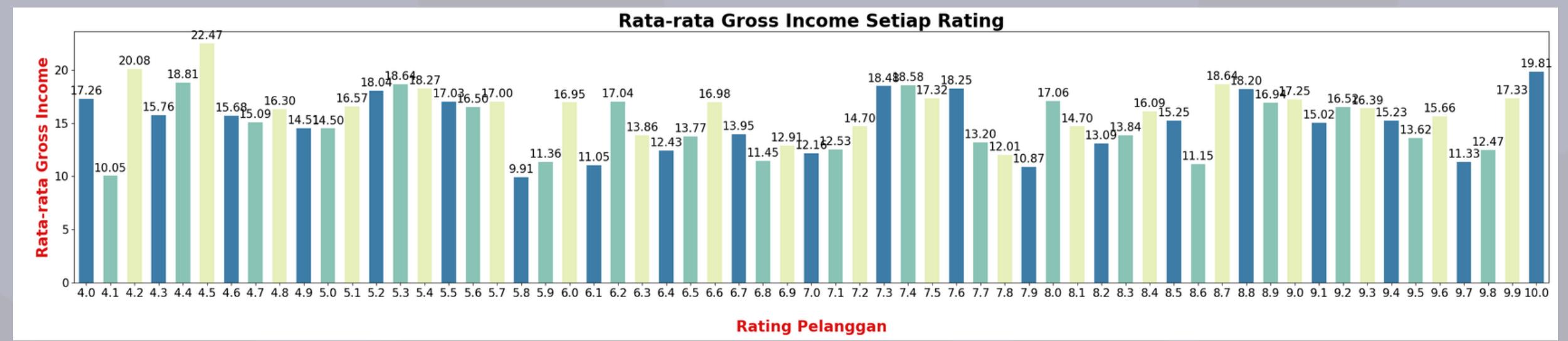
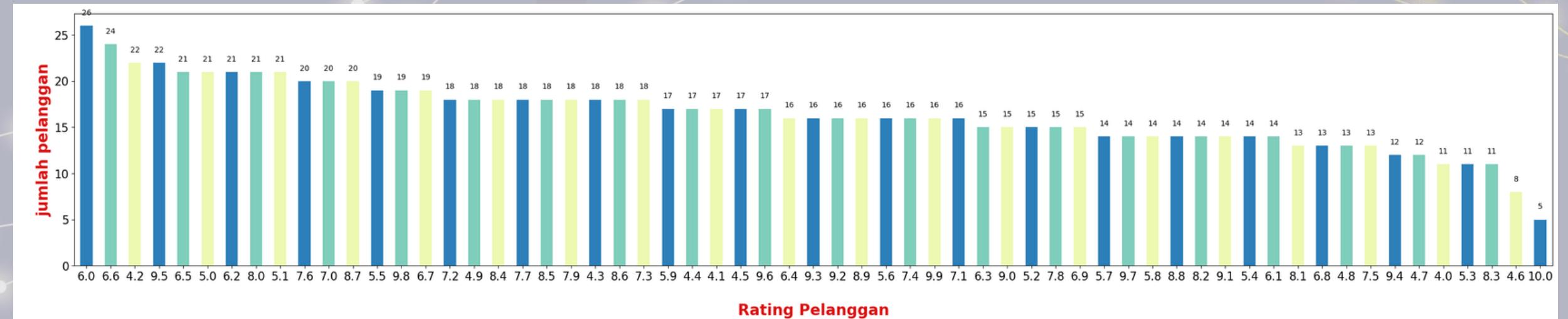
Historical  
record of  
sales data in  
3 different  
supermarkets



dataset ini terdiri dari 1000 baris dan 17 kolom (Invoice ID, Branch, Customer type, Product line, Unit Price, Quantity, Tax 5%, Total, Date, Time, Payment, Cogs, margin Gross Gross Income, dan Rating).



# EXPLORATORY DATA ANALYSIS



The average gross income is: 15.379368999999999  
The minimum gross income is: 0.5085  
The maximum gross income is: 49.65

df.isnull().sum()
Invoice ID
Branch
City
Customer type
Gender
Product line
Unit price
Quantity
Tax 5%
Total
Date
Time
Payment
cogs
gross margin percentage
gross income
Rating
dtype: int64

# ANALISIS KORELASI

## DATA CONTINUE

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0 750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1

## DATA KATEGORIKAL

hari	bulan	jam	menit
5	1	13	8

## DATA KATEGORIKAL

## KORELASI SPEARMAN

	Branch	City	Customer type	Gender	Product line	Payment	hari	bulan	jam	menit
gross income	0.019624	-0.005247	-0.016998	-0.051948	0.035857	-0.011271	0.002232	-0.023287	0.006832	-0.019308

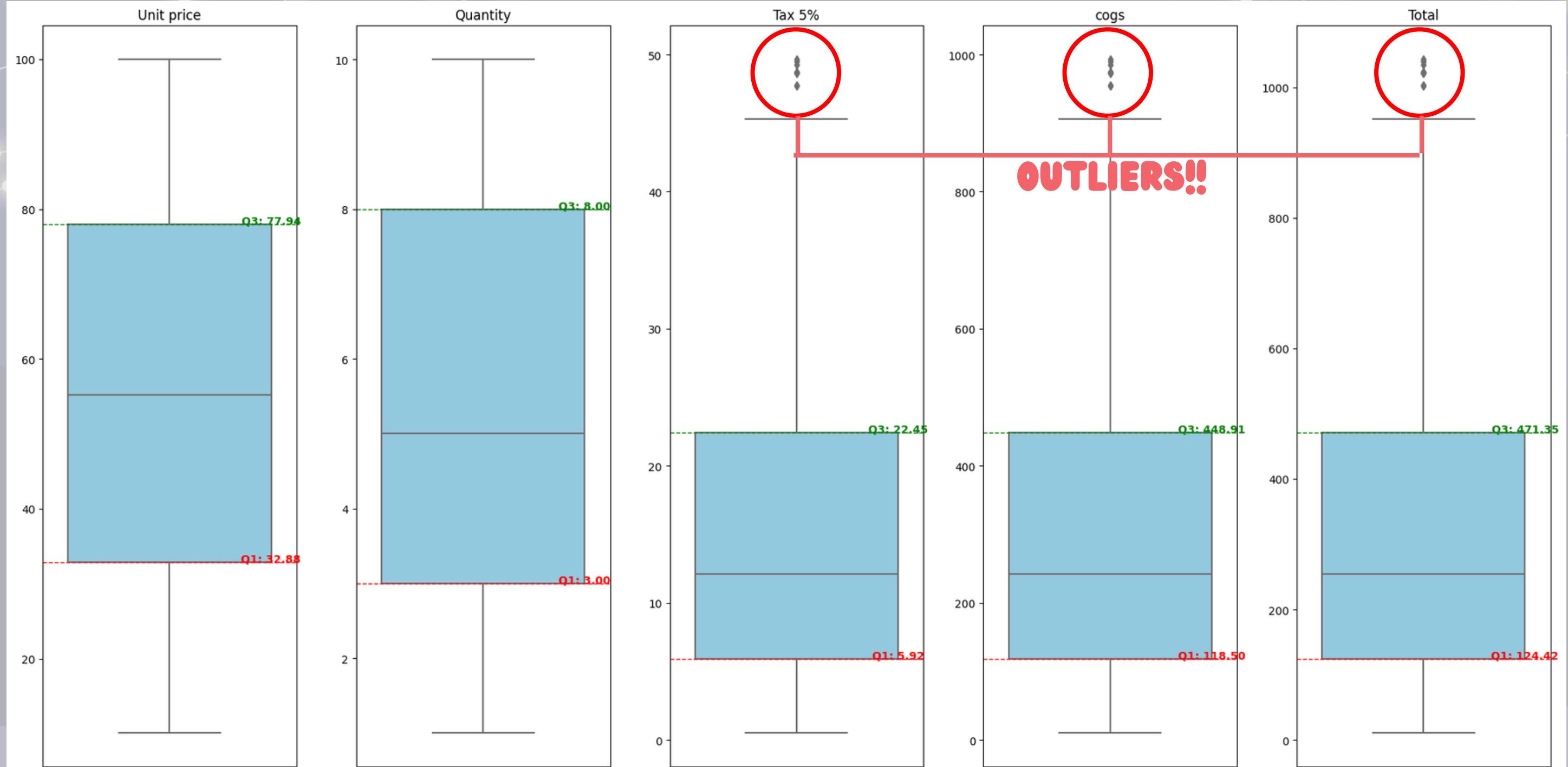
## KORELASI PEARSON

	Unit price	Quantity	Tax 5%	Rating	cogs	Total
gross income	0.633962	0.70551	1.0	-0.036442	1.0	1.0

Selected columns:

'unit price', 'quantity', 'tax 5%', 'cogs', 'total'

# ANALISIS OUTLIERS



# ANALISIS REGRESI LINEAR

## SIGNIFIKASI PERTAMA

Variabel	t hitung	t0,975 ; 999	Pvalue	Keputusan
Unit Price	1.920	0.975 ; 999	0.055	Terima H0
Quantity	-4.660		0.000	Tolak H0
Tax 5%	8.26e+15	1.96234	0.000	Tolak H0
Cogs	8.26e+15		0.000	Tolak H0
Total	8.26e+15		0.000	Tolak H0

## SIGNIFIKASI KEDUA

Variabel	t hitung	t0,975 ; 999	Pvalue	Keputusan
Quantity	0.233		0.816	Terima H0
Tax 5%	1.68e+16	1.96234	0.000	Tolak H0
Cogs	1.68e+16		0.000	Tolak H0
Total	1.68e+16		0.000	Tolak H0

## SIGNIFIKASI TERAKHIR

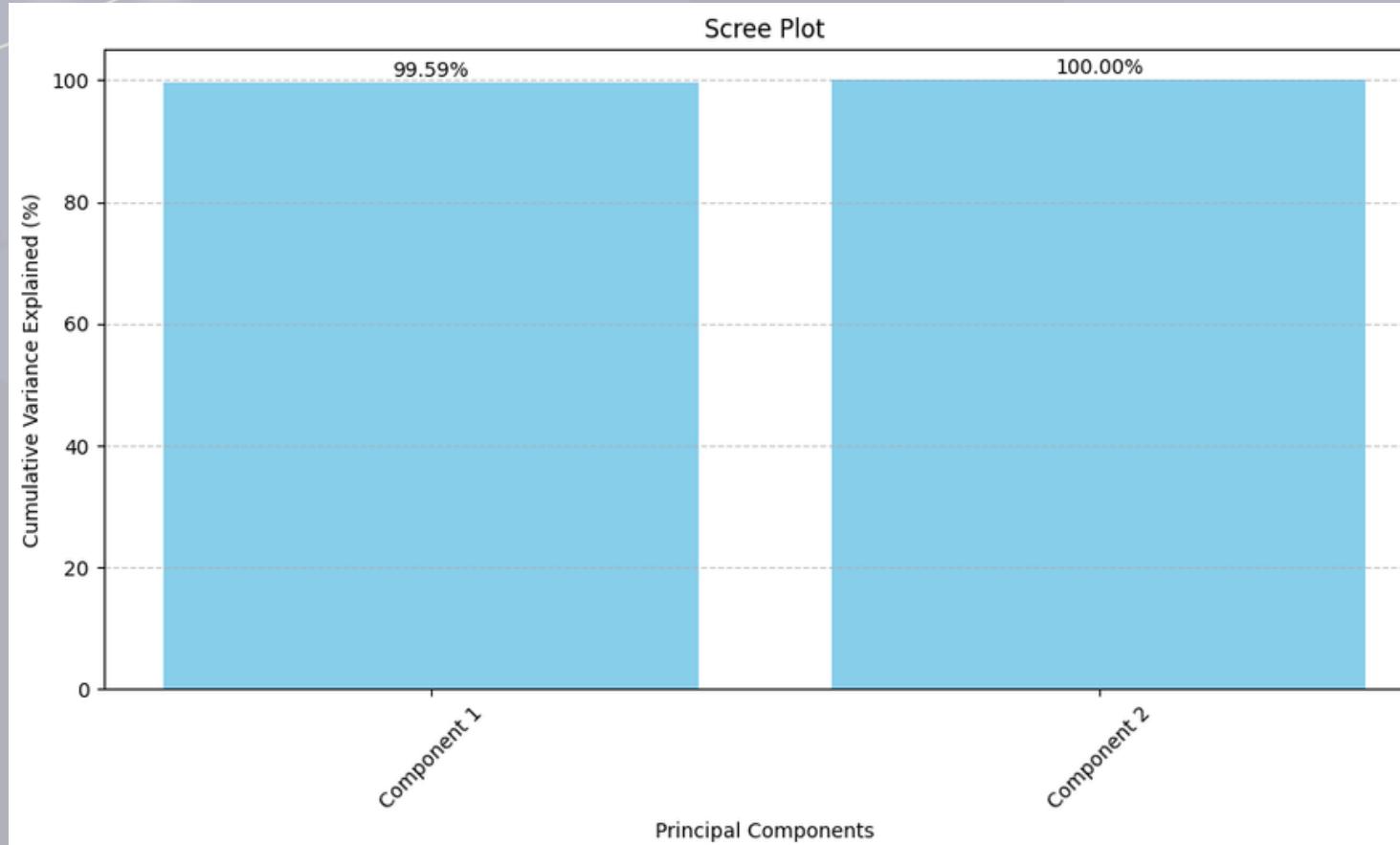
Variabel	t hitung	t0,975 ; 999	Pvalue	Keputusan
Tax 5%	5.89e+16	1.96234	0.000	Tolak H0
Cogs	5.89e+16		0.000	Tolak H0
Total	5.89e+16		0.000	Tolak H0

Selected columns:  
'tax 5%'. 'cogs', 'total'

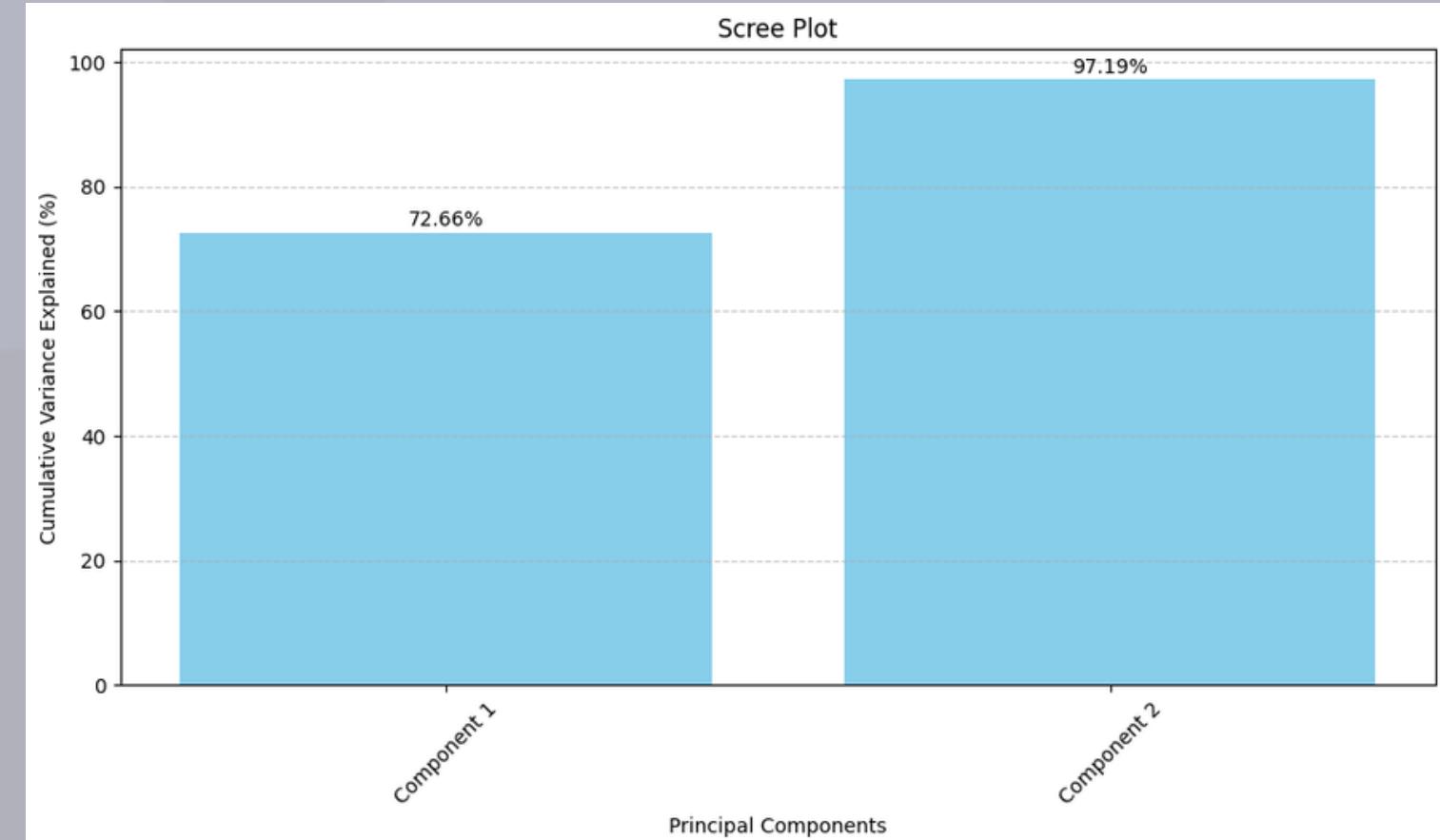


# ANALISIS VARIANSI PCA

DATA ASLI



DATA NORMALISASI



# HASIL EVALUASI MODEL

## DECISION TREE

Tabel Evaluasi Model Analisis Multivariat

Metode	Decision Tree Regression			
	MAE	MSE	RMSE	AVG
Tahap Pertama (Seleksi Fitur Analisis Korelasi)	0.37	0.21	0.46	0.346667
Tahap 2 (Optimasi Parameter GridsearchCV)	0.056	0.020	0.143	0.073
Tahap 3 (Best parameter model + Data normalisasi)	0.056	0.020	0.143	0.073
Tahap 4 (Best parameter model + PCA)	0.056	0.005	0.07	0.043667
Tahap 5 (Best parameter model + Data normalisasi + PCA)	0.35	3.523	1.877	1.916667

## REGRESI LINEAR

Tabel Evaluasi Model Analisis Multivariat

Ekstraksi Fitur	Linear Regression			
	MAE	MSE	RMSE	AVG
Tahap Pertama (Seleksi Fitur Analisis Korelasi)	1.37	2.19	1.48	1.68
Tahap kedua (Penggunaan Variabel berdasarkan signifikan atau tidaknya dengan regresi linear)	1.46	2.42	1.55	1.81
Tahap ketiga (Penggunaan Variabel berdasarkan signifikan atau tidaknya dengan regresi linear)	6	3.94	6.27	5.40333
Tahap Keempat (PCA)	6	3.94	6.27	5.40333
Tahap Kelima (Data normalisasi + PCA)	6.3	5.16	7.18	6.21333

# KESIMPULAN

Dari kelima variabel yang telah diamati, setelah dilakukan analisis regresi linear diperoleh 3 faktor yang mempengaruhi secara signifikan terhadap jumlah pendapatan kotor, yaitu dihasilkan pada variabel 'tax 5%', 'cogs', dan 'total'. Dimana, sebelum mendapatkan hasil signifikasi akhir tersebut, telah dilakukan signifikasi tahap pertama terhadap kelima variabel tersebut sehingga menyisakan 4 faktor, yaitu 'quantity', 'tax 5%', 'cogs', dan 'total'. Pada tahap tersebut variabel quantity tidak berpengaruh karena pvalue kurang dari 0.05, sehingga tidak digunakan pada tahap signifikasi selanjutnya. Setelah itu uji komparasi model di lakukan pada kedua model regresi dengan perhitungan rata rata dari 3 perhitungan MAE,MSE, dan RMSE di dapatkan model terbaik berupa model regresi decision tree pada tahap kedua dengan hasil akhir penilaian rata-rata terkecil bernilai 0.073.





**THANK YOU**

