MAKALAH PENERAPAN PCA (*principal component analysis)* PADA PENGELOMPOKKAN PRODUK DARI DATA RETAIL

**Dosen Pengampu : Fitrah Maharani Humaira M.Kom**

**Disusun untuk Memenuhi Tugas Mata Kuliah:**

**PRAKTIKUM PEMODELAN STATISTIKA TERAPAN**

**Disusun Oleh:**

**Aisyah Nailah Mufidah 3324600035**

**Indy Nailuvar Hamro’ Faqi 3324600037**

**Hafizh Muhammad Ilham 3324600043**

**Dea Aulia Rifda Yuniar 3324600047**

**Tegar Maulana Bhakti K. 3324600053**

**PROGRAM STUDI D4 SAINS DATA TERAPAN**

**DEPARTEMEN TEKNIK INFORMATIKA DAN KOMPUTER**

**POLITEKNIK ELEKTRONIKA NEGERI SURABAYA**

# 1. Pendahuluan

1.1 Latar Belakang

Industri ritel menghadapi tantangan dalam menganalisis data pelanggan yang kompleks dengan banyak variabel. Setiap kategori produk memiliki pola konsumsi yang berbeda, dan tidak semua variabel dalam dataset memiliki dampak yang sama dalam menjelaskan variasi perilaku pelanggan (Baihaqie, 2023).

Salah satu tantangan besar dalam analisis data penjualan adalah jumlah variabel yang besar, yang sering kali membuat analisis menjadi kompleks dan sulit diinterpretasikan (Dio, Arifin, & Septian, 2023). Dalam dunia bisnis, khususnya di sektor ritel dan grosir, data transaksi pelanggan mencerminkan berbagai variabel yang menggambarkan perilaku konsumsi. Misalnya, pelanggan mungkin membeli berbagai kategori produk seperti Fresh (produk segar), Grocery (sembako), Frozen (makanan beku), dan Detergents Paper (produk kebersihan). Namun, tidak semua variabel ini memiliki dampak yang sama dalam menjelaskan pola konsumsi pelanggan.

Untuk menyederhanakan struktur data tanpa kehilangan banyak informasi, Principal Component Analysis (PCA) digunakan sebagai teknik reduksi dimensi. PCA membantu mengidentifikasi kombinasi variabel yang paling berpengaruh terhadap variabilitas data dengan mengubah banyak variabel asli menjadi sejumlah principal components (PCs) yang tetap mempertahankan variasi terbesar dalam data (Anshary, Soesanto, & Ayatullah, 2022).

Dalam penelitian ini, PCA diterapkan pada Wholesale Customers Dataset, yang berisi data pengeluaran pelanggan grosir dalam berbagai kategori produk. Tujuan utama dari penggunaan PCA adalah untuk:

* Mengurangi dimensi data menjadi beberapa komponen utama yang lebih mudah dianalisis.
* Menentukan variabel yang paling berkontribusi terhadap variabilitas data pelanggan.
* Menganalisis jumlah optimal Principal Components berdasarkan explained variance.

Beberapa penelitian sebelumnya menunjukkan bahwa PCA efektif dalam mengekstrak pola utama dalam data ritel. Misalnya, penelitian oleh Susanto & Bachtiar (2023) menggunakan PCA untuk mengurangi kompleksitas data transaksi ritel dan memahami keterkaitan antar variabel produk.

Melalui penelitian ini, diharapkan PCA dapat membantu mengidentifikasi struktur utama dalam data pelanggan grosir, sehingga dapat memberikan wawasan mengenai keterkaitan antar variabel tanpa kehilangan banyak informasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, beberapa pertanyaan yang ingin dijawab dalam penelitian ini adalah:

1. Bagaimana PCA dapat digunakan untuk mereduksi dimensi data pelanggan wholesale tanpa kehilangan terlalu banyak informasi?

2. Bagaimana Principal Components (PCs) dapat merepresentasikan variasi dalam dataset?

3. Seberapa banyak informasi dalam dataset yang dapat dijelaskan oleh beberapa PCs?

4. Apa saja variabel yang memiliki kontribusi paling besar dalam komponen utama hasil PCA

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

* Menerapkan PCA untuk mereduksi dimensi data pelanggan retail tanpa kehilangan informasi yang signifikan.
* Menganalisis struktur hubungan antar variabel berdasarkan hasil PCA.
* Menentukan jumlah optimal Principal Components (PCs) berdasarkan explained variance.
* Menginterpretasikan makna dari PCs yang terbentuk, termasuk variabel mana yang paling berkontribusi dalam masing-masing komponen.

# 2. Tinjauan Pustaka

2.1 Definisi PCA

Metode principal component analysis (PCA) atau analisis komponen utama merupakan salah satu metode analisis data multivariat (Sari, 2023). Menurut Manullang et al. (2023), PCA merupakan suatu teknik seleksi data multivariat (multivariabel) yang mengubah atau mentransformasi suatu matriks data original menjadi suatu kumpulan kombinasi homogen yang lebih sedikit namun menyerap sejumlah besar varian dari data awal. Sedangkan menurut Susanto (2025), Principal Component Analysis (PCA) adalah metode statistik yang digunakan untuk mengurangi dimensi data namun tetap mempertahankan informasi yang penting.

Berdasarkan pengertian tersebut, dapat disimpulkan bahwa Principal Component Analysis adalah teknik statistik yang digunakan untuk mereduksi dimensi data dengan mengubah variabel asli yang mungkin saling berkorelasi menjadi sekumpulan variabel baru yang tidak berkorelasi, yang disebut komponen utama.

2.2 Prinsip Kerja PCA

Pada dasarnya, tujuan utama dari PCA adalah untuk menyederhanakan variabel yang diamati dengan cara mereduksi dimensinya. Ini dilakukan dengan menghapus korelasi antar variabel bebas melalui transformasi variabel bebas awal menjadi variabel baru yang sepenuhnya tidak berkorelasi, yang dikenal sebagai Principal Component (PC). Peranan paling penting dalam proses transformasi dipegang oleh nilai eigenvector dan eigenvalue yang besar. Oleh karena itu mereduksi dimensi dengan cara membuang eigenvector dan eigenvalue yang bernilai sangat kecil tidak akan membuat kehilangan data yang penting (Smith, 2002).

Adapun, terdapat beberapa cara untuk mendefinisikan vektor eigen dan nilai eigen. Pendekatan yang paling umum adalah mendefinisikan vektor eigen dari matriks **A** sebagai vektor **U** yang memenuhi persamaan berikut:

**Au = λu**

Yang apabila dituliskan ulang, persamaannya menjadi:

**(A - λI)u = 0**

di mana **λ** merupakan sebuah skalar yang disebut nilai eigen yang terkait dengan vektor eigen.

Dengan cara yang sama, kita juga dapat mengatakan bahwa vektor **u** adalah vektor eigen dari matriks **A** jika panjang vektor tetapi bukan arahnya) diubah ketika dikalikan dengan **A** (Abdi, 2007).

Sifat – sifat nilai eigen dan vector eigen (Susanto, 2025):

• Sebuah matriks memiliki invers jika dan hanya jika λ =0 bukan nilai eigen.

• Matriks diagonal memiliki elemen diagonal berupa nilai eigen. Matriks A dapat didiagonalisasi jika ada matriks P sehingga P-1

AP adalah matriks diagonal dengan kolom P adalah basis ruang eigen matriks A

2.3 Aplikasi PCA dalam berbagai bidang

PCA telah diterapkan dalam berbagai bidang untuk menyederhanakan data dan mengungkap pola tersembunyi. Berikut adalah beberapa aplikasi PCA dalam berbagai sektor:​

1. Analisis Data Retail

Dalam industri retail, PCA digunakan untuk menyederhanakan pola konsumsi pelanggan dengan mengidentifikasi variabel utama dalam transaksi. Dengan mereduksi dimensi data transaksi yang kompleks, perusahaan dapat memahami perilaku belanja pelanggan dan mengelompokkan mereka ke dalam segmen yang berbeda. Hal ini memungkinkan peritel untuk menyesuaikan strategi pemasaran dan promosi sesuai dengan kebutuhan masing-masing segmen pelanggan.​

1. Pengenalan Pola & Computer Vision

PCA banyak digunakan dalam pengenalan wajah dan analisis citra. Dengan mereduksi dimensi data citra wajah, PCA memungkinkan sistem untuk mengenali individu dengan lebih efisien. Misalnya, dalam penelitian yang menggunakan metode Haar Cascade untuk mendeteksi kantuk pada pengemudi, PCA digunakan untuk mereduksi dimensi fitur gambar, sehingga meningkatkan akurasi sistem deteksi (Lubis et al, 2023). ​

1. Bioinformatika

Dalam bidang bioinformatika, PCA digunakan untuk menemukan pola dalam ekspresi gen. Dengan menganalisis data ekspresi gen yang kompleks, PCA dapat mengidentifikasi komponen utama yang mempengaruhi variasi genetik. Ini membantu peneliti dalam memahami mekanisme biologis dan mengidentifikasi gen yang berperan penting dalam kondisi atau penyakit tertentu.​

1. Keuangan

Dalam sektor keuangan, PCA digunakan untuk mengurangi jumlah variabel dalam analisis risiko pasar. Dengan mereduksi dimensi data keuangan yang kompleks, seperti harga saham atau indikator ekonomi, PCA membantu analis dalam mengidentifikasi faktor utama yang mempengaruhi risiko dan return. Ini memungkinkan pengelolaan portofolio yang lebih efektif dan pengambilan keputusan investasi yang lebih baik.

2.4 Kelebihan dan kekurangan PCA

PCA memiliki berbagai kelebihan dan kelemahan yang perlu dipertimbangkan saat menggunakannya dalam analisis data.​

1. Kelebihan PCA
2. Mengurangi Kompleksitas Data tanpa Kehilangan Banyak Informasi

PCA secara efektif mengurangi dimensi kumpulan data kompleks dengan mengubahnya menjadi ruang berdimensi lebih rendah sambil mempertahankan informasi yang paling penting. Ini memungkinkan representasi data yang disederhanakan tanpa kehilangan fitur yang relevan secara signifikan.

1. Menghilangkan Multikolinearitas antar Variabel

Salah satu keunggulan PCA adalah kemampuannya untuk membersihkan data dari korelasi tanpa harus mengurangi jumlah variabel asli. Proses PCA mengubah variabel asli yang berkorelasi menjadi komponen utama yang orthogonal atau tidak berkorelasi, sehingga mengatasi masalah multikolinearitas dalam analisis regresi atau model statistik lainnya.

1. Mempermudah Visualisasi Data Berdimensi Tinggi

PCA memungkinkan kita untuk visualisasi data dalam ruang berdimensi lebih rendah, sehingga memudahkan kita dalam pemahaman struktur data. Dengan mereduksi data ke dalam 2 atau 3 dimensi, kita dapat membuat plot yang membantu dalam mengidentifikasi pola, kluster, atau anomali dalam data.

1. Kelemahan PCA
2. PCA Hanya Menangkap Hubungan Linear, sehingga Tidak Efektif untuk Pola Non-Linear

PCA adalah teknik linier, sehingga mungkin tidak dapat menangkap hubungan non-linier dalam data. Jika data memiliki struktur non-linear yang signifikan, PCA mungkin tidak efektif dalam mereduksi dimensi atau mengungkap pola yang ada. Dalam kasus seperti itu, teknik lain seperti Kernel PCA atau metode non-linear lainnya mungkin lebih sesuai.

1. Komponen Utama Sulit untuk Diinterpretasikan secara Langsung

Setiap Principal Component (PC) atau Komponen Utama (KU) merupakan kombinasi linier dari semua variabel. Artinya, setiap KU merupakan kombinasi linier dari semua variabel dengan beban (nilai loading) tertentu, yang dapat menyulitkan interpretasi langsung dari komponen tersebut dalam konteks variabel asli.

1. Sensitif terhadap Skala Data, sehingga Perlu Standarisasi Terlebih Dahulu

PCA sensitif terhadap skala variabel, sehingga variabel dengan skala yang lebih besar dapat mendominasi komponen utama yang dihasilkan. Oleh karena itu, penting untuk melakukan standarisasi atau normalisasi data sebelum menerapkan PCA untuk memastikan bahwa setiap variabel berkontribusi secara proporsional dalam analisis.

2.5 Standarisasi Data

Standarisasi data adalah melakukan proses penskalaan ulang fitur-fitur numerik dalam dataset. Penskalaan bertujuan untuk membuat data memiliki format dan skala yang konsisten. Kami melakukan penskalaan ulang memastikan bahwa semua variabel memiliki rata-rata 0 dan standar deviasi 1. Standarisasi ini dikenal juga dengan Z-score Normalization atau Standard Scaler. Formulasi Standard Scaler sebagai berikut :

adalah data setelah standarisasi

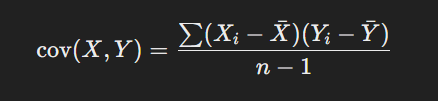
adalah nilai asli

adalah nilai rata-rata

adalah nilai varians

2.6 Matriks Kovarians dalam PCA

Matriks kovarians digunakan mengukur seberapa erat hubungan antar variabel dalam dataset. Nilai positif menunjukkan hubungan searah (misalnya, jika pembelian Grocery naik, Detergents\_Paper juga naik), sedangkan nilai negatif menunjukkan hubungan berlawanan (misalnya, jika pembelian Fresh naik, Grocery turun).

Cara Menghitung Matriks Kovarians:

Matriks kovarians digunakan untuk mendapatkan nilai eigen (eigenvalues) dan vektor eigen (eigenvectors) yang membentuk Principal Components.

2.7 Cumulative Explained Variance dalam PCA

Cumulative explained variance dalam Principal Component Analysis (PCA) mengacu pada proporsi total variansi dalam dataset yang dijelaskan secara kumulatif oleh sejumlah komponen utama tertentu. Ini membantu menentukan jumlah optimal komponen yang diperlukan untuk mewakili data secara efektif tanpa kehilangan informasi signifikan.​

2.7.1 Tujuan Cumulative Explained Variance

Tujuan utama dari cumulative explained variance adalah untuk mengidentifikasi jumlah komponen utama yang diperlukan untuk menjelaskan sebagian besar variansi dalam data. Dengan memilih jumlah komponen yang tepat, kita dapat mengurangi dimensi data sambil mempertahankan informasi penting, sehingga mempermudah analisis dan visualisasi data.​

2.7.2 Cara Kerja Cumulative Explained Variance dalam PCA

Dalam PCA, data asli ditransformasikan menjadi sejumlah komponen utama yang tidak berkorelasi, diurutkan berdasarkan variansi yang dijelaskan. Cumulative explained variance dihitung dengan menjumlahkan proporsi variansi yang dijelaskan oleh setiap komponen utama secara bertahap. Biasanya, grafik cumulative explained variance digunakan untuk menentukan jumlah komponen yang optimal; titik di mana penambahan komponen baru tidak lagi memberikan peningkatan signifikan dalam variansi yang dijelaskan sering disebut sebagai "elbow point".

2.8 Konsep Reduksi Dimensi Dalam Analisis Data

Masalah pada data yang memiliki fitur yang banyak adalah data sulit divisualisasikan, sulit diolah, dan sulit diinterpretasikan, masalah ini disebut juga dengan istilah “Curse of Dimensionality” . Data dengan dimensi yang tinggi juga dapat menyebabkan redudansi informasi akibat korelasi tinggi antar fitur.

Teknik reduksi dimensi hadir sebagai solusi untuk masalah ini. Reduksi dimensi bertujuan untuk mengurangi jumlah dimensi sambil mempertahankan informasi fitur sebanyak mungkin. Keuntungan pada melakukan reduksi dimensi antara lain :

* Mempercepat proses komputasi
* Mempermudah visualsi
* Meningkatkan Efektivitas Model Analitik

# 3. Implementasi

A. Dataset

Dataset yang diambil dari makalah ini diambil dari website UCI Machine Learning Repository. UCI Machine Learning Repository adalah kumpulan basis data, teori domain, dan generator data yang digunakan oleh komunitas pembelajaran mesin untuk analisis empiris algoritma pembelajaran mesin yang dibuat oleh University of California Irvine. Untuk makalah ini, data yang digunakan adalah Wholesale Customers Dataset (Cardoso, 2013). Data berisi 440 baris dan 8 kolom yaitu Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents\_Paper, dan Delicassens. Berikut informasinya:

1. Channel : Label tipe pelanggan yaitu 1 untuk Horeca (Hotel/Restaurant/Café) dan 2 untuk Retail , tipe data nominal
2. Region : Label lokasi geografis pelanggan yaitu 1 untuk Lisbon, 2 untuk Oporto, dan 3 untuk wilayah selain Lisbon & Oporto, tipe data nominal
3. Fresh : Pengeluaran tahunan untuk produk segar (misalnya sayur & buah), tipe data kontinu
4. Milk : Pengeluaran tahunan untuk produk susu dan turunanya, tipe data kontinu
5. Grocery : Pengeluaran tahunan untuk produk sembako dan bahan makanan pokok, tipe data kontinu
6. Frozen : Pengeluaran tahunan untuk produk makanan beku, tipe data kontinu
7. Detergents\_Paper : Pengeluaran tahunan untuk produk kebersihan (deterjen, tisu, dll.), tipe data kontinu
8. Delicassen : Pengeluaran tahunan untuk makanan olahan atau mewah, tipe data kontinu

B. Penggunaan *Python*

*Python* adalah bahasa pemrograman yang umum digunakan untuk pembelajaran mesin dan analisis data. Pemilihan Python sebagai alat yang kita gunakan untuk implementasi kasus ini dikarenakan Python memiliki banyak *Library* (pustaka) berisi modul-modul yang memudahkan penggunanya untuk melakukan memproses data dan visualisasi data. Berikut adalah *Library* yang digunakan :

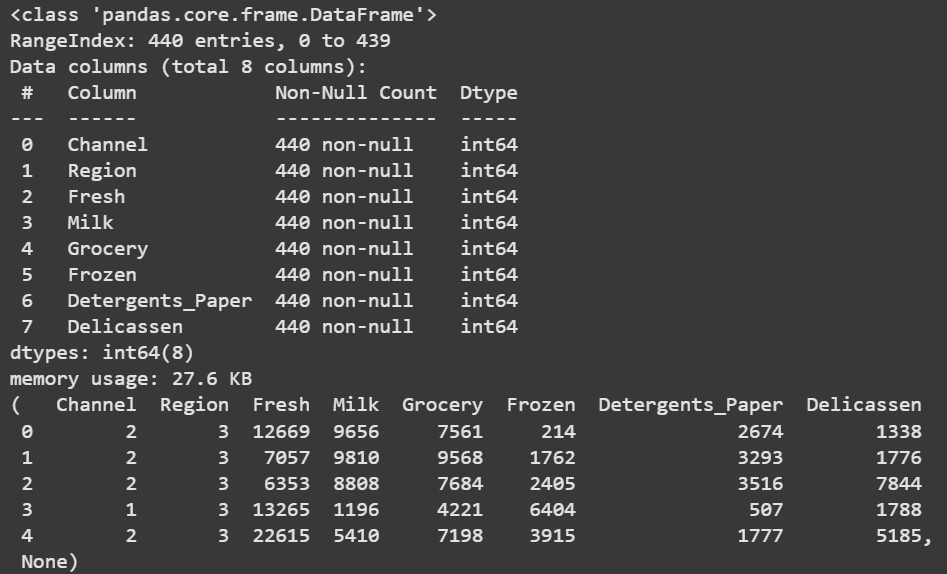
* pandas : Digunakan untuk membaca dan mengelola dataset Wholesale Customers.
* numpy : Digunakan untuk perhitungan matriks kovarians, eigenvalues, dan eigenvectors.
* matplotlib.pyplot : Digunakan untuk visualisasi hasil PCA, seperti Scree Plot.
* sklearn.preprocessing.StandardScaler : Digunakan untuk melakukan standarisasi data sebelum PCA diterapkan.
* sklearn.decomposition.PCA : Modul utama yang digunakan untuk melakukan PCA dan memperoleh Principal Components.

| import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  from sklearn.preprocessing import StandardScaler  from sklearn.decomposition import PCA |
| --- |

C. *Load data*

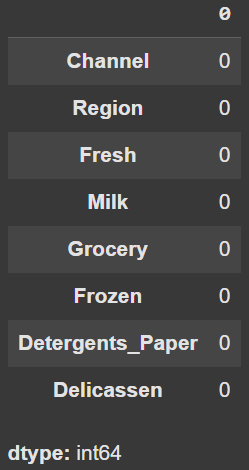
Berikut adalah kode untuk mengimpor data diikuti dengan keluaran untuk melihat informasi dan isi dataset :

| df = pd.read\_csv(path + "/Wholesale customers data.csv")  df.head(), df.info() |
| --- |



langkah ini digunakan untuk membaca suatu data dari .csv yang diperoleh dari dataset tersebut. Untuk memproses data kita mengambil data .csv yang telah diunduh sebelumnya, dengan menggunakan library pandas.

D. Mengecek Nilai Kosong

*Missing value* (nilai kosong) biasanya ada karena ketidaksempurnaan pemasukan data saat penyusunan dataset. *Missing value* bisa sangat mempengaruhi proses analisis data, oleh karena itu kita akan cek apakah ada atau tidak: 

| df.isna().sum() |
| --- |

Hasilnya pada setiap kolom tidak ada nilai yang hilang, ini berarti dataset sudah siap di olah ke tahap berikutnya.

E. Pemilihan Fitur

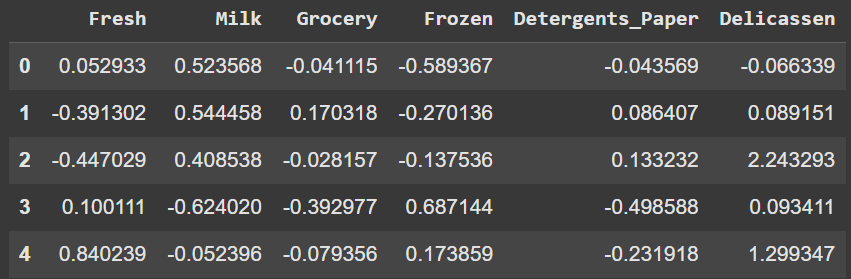
Kita akan mengolah fitur Fresh, Milk, Grocery, Frozen, Detergents\_Paper, dan Delicassen karena Channel dan Region hanya label. Kita drop fitur Channel dan Region

| df\_num = df.drop(columns=["Channel", "Region"]) |
| --- |

F. Standarisasi Data

Melakukan standarisasi data menggunakan Standard Scaler dari library Python.

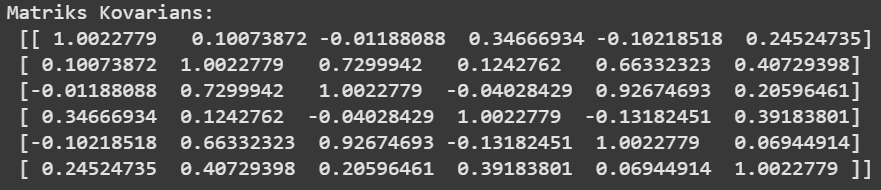
| # Standarisasi data  scaler = StandardScaler()  df\_scaled = scaler.fit\_transform(df\_num)  # Konversi kembali ke DataFrame  df\_scaled = pd.DataFrame(df\_scaled, columns=df\_num.columns)  # Cek hasil standarisasi  df\_scaled.head() |
| --- |



Bisa kita lihat hasil yang didapatkan adalah rentang nilai pada fitur-fitur yang dipilih sebelumnya terdiri dari ratusan dan ribuan, sekarang rentang nilainya adalah -1 hingga 1. Semua fitur yang dipilih memiliki distribusi dengan rata-rata 0 dan standar deviasi 1**.**

G. Membuat Matriks Kovarians

| # Hitung matriks kovarians  cov\_matrix = np.cov(df\_scaled.T)  # Tampilkan bentuk matriks kovarians  print("Matriks Kovarians:\n", cov\_matrix) |
| --- |



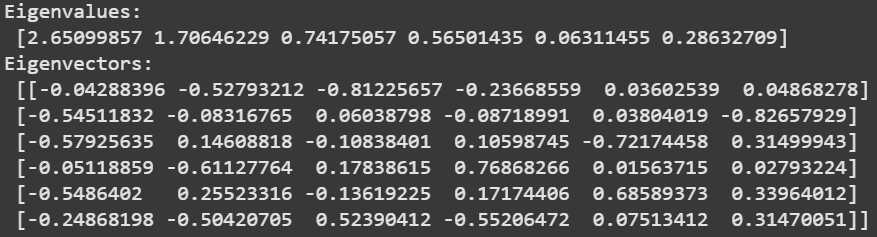
Dari data yang sudah distandarisasi tadi, kita bentuk matriks kovarians. Matriks kovarians mengukur seberapa besar perubahan pada satu variabel berkaitan dengan perubahan variabel lain. Jika kovarians positif, berarti kedua variabel cenderung bergerak ke arah yang sama. Jikanegatif**,** berarti keduanya bergerak ke arah yang berlawanan.

Diagonal utama menunjukkan varians dari masing-masing variabel. Nilai-nilai ini sekitar 1.002, yang sesuai dengan hasil standarisasi. Elemen non-diagonal menunjukkan kovarians antara variabel yang berbeda. Milk & Grocery (0.7299) memiliki ovarians positif tinggi, menunjukkan hubungan erat antara kedua variabel.Detergents\_Paper & Grocery (0.9267) Hubungan sangat kuat, menunjukkan bahwa pelanggan yang membeli Detergents\_Paper cenderung membeli Groceryjuga. Frozen & Milk (-0.0402) kovarians hampir nol, menunjukkan bahwa Frozen dan Milk memiliki hubungan yang sangat lemah atau tidak berkorelasi.

H. Eigenvalue & Eigenvector

Setelah mendapatkan matriks kovarians, sekarang kita akan mencari eigenvalue dan eigenvector menggunakan numpy.

| # Hitung eigenvalues dan eigenvectors dari matriks kovarians  eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)  print("Eigenvalues:\n", eigenvalues)  print("Eigenvectors:\n", eigenvectors) |
| --- |



Eigenvalues (λ) menunjukkan jumlah variansi yang dijelaskan oleh masing-masing Principal Component (PC). Eigenvectors (v) mewakili arah dari komponen utama, yaitu kombinasi linear dari fitur asli yang membentuk setiap PC. Menurut eigenvalues, pada PC ke 3 sudah cukup menjelaskan variansi data sekitar 0.74. Menurut eigenvectors, Milk, Grocery dan Detergents\_Paper memiliki kontribusi besar terhadap PC1, menunjukkan bahwa konsumsi ketiga produk ini memiliki hubungan kuat. Frozen dan Fresh lebih dominan di PC2, yang menunjukkan bahwa mereka memiliki pola konsumsi berbeda dibandingkan dengan Milk dan Grocery. Memang masih sulit untuk membaca jumlah variansi yang dijelaskan pada PC.

I. Menghitung *Cumulative Explained Varianc*e

Jika hanya eigenvalues dan eigenvectors masih sulit untuk menentukan PC, kita akan menggunakan c*umulative explained variance* :

| # Urutkan eigenvalues dari yang terbesar ke terkecil  sorted\_eigenvalues = sorted(eigenvalues, reverse=True)  # Hitung cumulative explained variance  explained\_variance\_ratio = np.cumsum(sorted\_eigenvalues) / np.sum(sorted\_eigenvalues)  # Tampilkan eigenvalues yang telah diurutkan  print("Cumulative Explained Variance:\n", explained\_variance\_ratio) |
| --- |

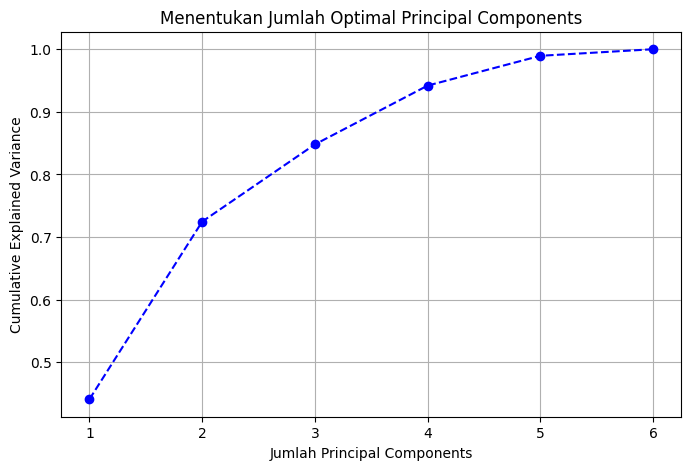


*Cumulative Explained Variance* menunjukkan berapa persen dari total variansi dalam data yang dapat dijelaskan oleh sejumlah PC. Cumulative Explained Variance dapat dicari dengan mencari nilai jumlah kumulatif dari eigenvalues dibagi dengan jumlah eigenvalues. Hasil yang didapatkan adalah PC1 menjelaskan sekitar 44% dari total variansi data, PC1 + PC2 menjelaskan sekitar 72% variansi, PC1 hingga PC3 sudah menjelaskan lebih dari 85% variansi, Semua PC (PC1 hingga PC6) menjelaskan 100% variansi, tetapi menambahkan lebih dari 4 PC mungkin tidak memberikan banyak manfaat tambahan. Untuk bisa lebih yakin jumlah PC yang digunakan visualisasi dengan plot akan membantu.

J. Menentukan Jumlah PC

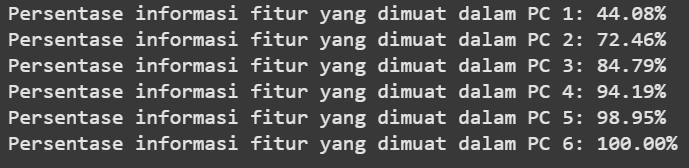
Kita akan membuat line plot jumlah *Principal Components* berbanding dengan *Cumulative Explained Variance:*

| # Plot explained variance untuk menentukan jumlah optimal PC  plt.figure(figsize=(8,5))  plt.plot(range(1, len(df\_scaled.columns)+1), explained\_variance\_ratio, marker='o', linestyle='--', color='b')  plt.xlabel('Jumlah Principal Components')  plt.ylabel('Cumulative Explained Variance')  plt.title('Menentukan Jumlah Optimal Principal Components')  plt.grid()  plt.show() |
| --- |



Ini adalah hasil dari tampilan plot, terlihat titik siku berada di sekitar PC 2 hingga 4, untuk melihat lebih jelas, kita akan melihat persentase *cumulative explained variance* :

| # Cek berapa persen informasi yang tercakup dalam PC  for i, j in enumerate(explained\_variance\_ratio):  print(f"Persentase informasi fitur yang dimuat dalam PC {i+1}: {j:.2%}") |
| --- |

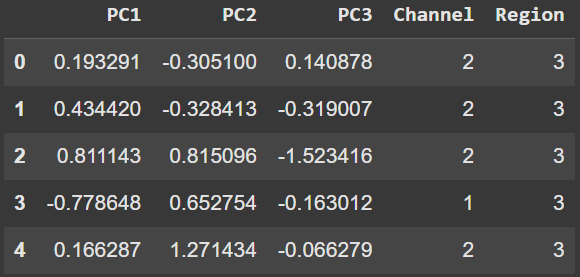


Disini lebih jelas informasi fitur yang dimuat pada PC berdasarkan *cumulative explained variance* . Pada PC 3 variance yang dijelaskan ada 84.79%, ini menunjukkan informasi yang dimuat hampir mencapai 85%, sehingga ini merupakan titik optimal untuk menjaga keseimbangan antara kompleksitas dan akurasi representasi data. Penambahan PC4 hanya meningkatkan sedikit informasi, sehingga tidak terlalu signifikan. Menggunakan lebih banyak PCs justru menambah kompleksitas tanpa manfaat informasi yang besar. Keputusan akhir kita akan menggunakan 3 PC

K. Menerapkan PCA

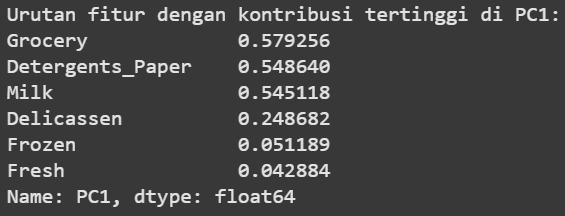
Menerapkan PCA dengan 3 komponen, lalu kita membuat kolom untuk masing-masing PC:

| pca = PCA(n\_components=3)  df\_pca = pca.fit\_transform(df\_scaled)  df\_pca = pd.DataFrame(df\_pca, columns=['PC1', 'PC2', 'PC3'])  df\_pca[['Channel', 'Region']] = df[['Channel', 'Region']]  df\_pca.head() |
| --- |

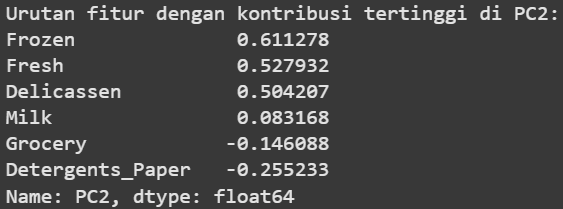


Dataset dengan fitur PC1, PC2, dan PC3 sudah dibuat, sekarang kita akan melihat fitur dengan kontribusi paling tinggi ke rendah pada masing-masing PC:

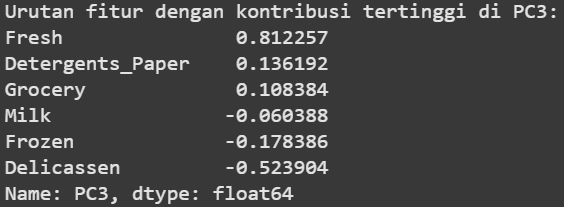
| # Cek kontribusi setiap fitur terhadap PC1, PC2, dan PC3  loadings = pd.DataFrame(pca.components\_.T, columns=['PC1', 'PC2', 'PC3'], index=df\_num.columns)  # Urutkan loadings untuk melihat fitur dominan dalam tiap PC  for pc in ['PC1', 'PC2', 'PC3']:  print(f"\nUrutan fitur dengan kontribusi tertinggi di {pc}:")  print(loadings[pc].sort\_values(ascending=False)) |
| --- |



Grocery, Detergents\_Paper, dan Milk memiliki kontribusi terbesar dalam PC1. PC1 tampaknya merepresentasikan kategori produk kebutuhan rumah tangga sehari-hari, yang terdiri dari produk sembako dan kebersihan. Frozen dan Fresh memiliki kontribusi paling rendah dalam dimensi ini.



Frozen, Fresh, dan Delicassen memberikan kontribusi terbesar dalam PC2. PC2 lebih banyak menangkap variasi dari produk yang lebih spesifik seperti makanan beku, makanan segar, dan makanan olahan mewah. Grocery dan Detergents\_Paper memiliki kontribusi negatif, menunjukkan bahwa PC2 menangkap dimensi yang berbeda dari PC1.



Fresh menjadi fitur dominan dalam PC3, diikuti oleh Detergents\_Paper dan Grocery. Delicassen dan Frozen memiliki kontribusi negatif, menunjukkan bahwa PC3 lebih menggambarkan dimensi perbedaan konsumsi Fresh dengan kategori lainnya.

Berdasarkan Cumulative Explained Variance, tiga komponen pertama sudah menjelaskan 84.79% variasi data. PC1 menangkap pola konsumsi produk kebutuhan rumah tangga, PC2 menangkap pola makanan segar dan beku, sedangkan PC3 memberikan variasi tambahan untuk Fresh dibandingkan kategori lainnya. Oleh karena itu, kita dapat menggunakan tiga Principal Components (PC1, PC2, dan PC3) sebagai representasi dari data pelanggan tanpa kehilangan terlalu banyak informasi.

**4. Kesimpulan**

Berdasarkan hasil penelitian dan implementasi Principal Component Analysis (PCA) pada Wholesale Customers Dataset, beberapa kesimpulan dapat diambil sebagai berikut:

1. PCA berhasil mereduksi dimensi data tanpa kehilangan informasi yang signifikan.
   * Dari enam variabel asli, analisis menunjukkan bahwa tiga Principal Components (PCs) pertama sudah cukup untuk menjelaskan 84.79% dari total variansi data.
   * Penggunaan lebih dari tiga PC tidak memberikan peningkatan informasi yang signifikan, sehingga keputusan untuk menggunakan tiga PC merupakan pilihan optimal.
2. Setiap Principal Component (PC) menangkap pola tertentu dalam konsumsi pelanggan:
   * PC1 merepresentasikan produk kebutuhan rumah tangga seperti Grocery, Milk, dan Detergents\_Paper, yang menunjukkan bahwa produk-produk ini memiliki pola konsumsi yang mirip.
   * PC2 merepresentasikan makanan segar dan beku, dengan kontribusi tertinggi dari Frozen, Fresh, dan Delicassen, yang menunjukkan pola konsumsi yang berbeda dibandingkan PC1.
   * PC3 berfokus pada Fresh, dengan kontribusi tertinggi dari variabel Fresh, sedangkan produk lainnya memberikan pengaruh negatif atau rendah dalam dimensi ini.
3. PCA membantu mengidentifikasi variabel yang paling berkontribusi dalam pola konsumsi pelanggan.
   * Hasil analisis menunjukkan bahwa Grocery dan Detergents\_Paper memiliki korelasi tinggi, sehingga dapat digunakan untuk memahami perilaku pelanggan yang cenderung membeli kedua produk tersebut secara bersamaan.
   * Frozen dan Fresh memiliki pola yang lebih independen, sehingga kemungkinan besar target pelanggan yang membeli produk ini berbeda dari pelanggan yang membeli Grocery dan Detergents\_Paper.
4. Penerapan PCA dalam analisis data retail memberikan manfaat dalam menyederhanakan kompleksitas data.
   * Dengan mengurangi dimensi data menjadi tiga komponen utama, analisis dapat lebih fokus pada pola konsumsi pelanggan tanpa perlu mempertimbangkan seluruh variabel secara langsung.
   * Hasil PCA dapat digunakan untuk analisis lanjutan seperti clustering pelanggan atau strategi pemasaran berbasis data.

5. Untuk penelitian selanjutnya, dapat dilakukan eksplorasi lebih lanjut dengan teknik clustering (misalnya K-Means) menggunakan hasil PCA untuk mengelompokkan pelanggan berdasarkan pola konsumsi mereka. Selain itu, metode non-linear seperti Kernel PCA dapat dicoba untuk melihat apakah ada pola lain yang tidak terdeteksi oleh PCA berbasis linear.

**5. Daftar Pustaka**

**Abdi, H. (2007). *Eigenvalues and eigenvectors*. University of Texas at Dallas.** [**https://personal.utdallas.edu/~herve/Abdi-EVD2007-pretty.pdf**](https://personal.utdallas.edu/~herve/Abdi-EVD2007-pretty.pdf)

**Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202, 1.** [**https://doi.org/10.1098/rsta.2015.0202**](https://doi.org/10.1098/rsta.2015.0202)

**Syah, W. F., Sudiarjo, A., & Hartono, R. (2024). Analisa dan rekomendasi tampilan antarmuka pada aplikasi Budiman Mobile menggunakan metode Kansei Engineering. *THEMATIC: Innovative Research Science Information Technology, 1*(1), 36–46.** [**Thematic (Innovative Research Science Information Technology)**](https://e-journal.uncip.ac.id/index.php/thematic)

**Fanisa, A., Hendra, R., (2024). Pemasaran Berbasis Data Cara Memanfaatkan Analitik Untuk Menjangkau Yang Tepat. *Jurnal Ilmiah Menajemen, Ekonomi dan Akuntasi, 4*(2), 16-23.** [**https://doi.org/10.55606/jurimea.v4i2.849**](https://doi.org/10.55606/jurimea.v4i2.849)

**Siti, A., Sofyan, S., & Maryah. (2020). Strategi Pemasaran Hni-Hpai Dalam Meningkatkan Jumlah Konsumen Studi Analisis Konsep Pemasaran Syariah. *Jurnal Syariah, 8*(2).**

**Faisal, A., Indrianti, I., Meiga, I, M., & Edy, Widodo. (2020). Analisis Faktor Pelanggan Dalam Pemilihan Toko Ritel Modern Dan Toko Ritel Tradisional Di Kota Yogyakarta. *Prosiding Konferensi Nasional Penelitian Matematika dan Pembelajarannya*.** [**https://proceedings.ums.ac.id/knpmp/article/view/1920/1875**](https://proceedings.ums.ac.id/knpmp/article/view/1920/1875)

**M, S, Noya, V, D., A, Z, Wattimena., & S, D, Saputri. (2017). Penggunaan Metode Analisis Komponen Utama Untuk Mereduksi Faktor-Faktor Inflasi Di Kota Ambon. *Jurnal Ilmu Matematika dan Terapan, 11*(2), 109-118.** [**https://doi.org/10.30598/barekengvol11iss2pp109-118**](https://doi.org/10.30598/barekengvol11iss2pp109-118)

**Fauzan, A, M., Cahyo, A, L., Dinda, P, A, N., Verina, A, M., & Desta, S, P. (2021). Pengaruh *Pricipal Component Analysis* terhadap Akurasi Model Machine Learning dengan Algoritma *Artificial Neural Network* untuk Prediksi Kebangkrutan Perusahaan. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasi (SENAMIKA)*. ISBN 978-623-93343-4-5**

**Sugoro, B, S., Subagyo., Aliq, Z., & Fitra, L, N. (2017). Peramal Kausal Berbasis Integrasi Princial Component Analysis dan Jaringan Syaraf Tiruan. *Jurnal Sains, Teknologi dan Industri, 14*(2), pp.117-125. ISSN 2407-0939**

**E, Srivishnu, H., Musa, H., & Nurheni, S, P. (2011). Kajian Perilaku Konsumen terhadap Strategi emasaran Teh Herbal di Kota Bogor. *Manajemen IKM, 6*(2), pp. 143-151. ISSN 2085-8418**

**Ilyas, M., Mardila, G, F, A. (2018). Pengambilan Keutuan Multi Kriteria: Kajian Teoritis Metode dan Pendekatan Dalam Pemilihan Pemasok. *Jurnal Ilmiah Teknik Industri,17*(1).** [**https://doi.org/10.23917/jiti.v17i1.5389**](https://doi.org/10.23917/jiti.v17i1.5389)

**Fahmi, A., Hamdan., Sugiyanto., & Ilham, W, S., (2022). Klasifikasi Customer Relationship Management Menggunakan Dataset KDD Cup 2009 dengan Teknik Dimensi. *Komputika: Jurnal Sistem Komputer,11*(2), 193-202.** [**https://doi.org/komputika.v11i2.6498**](https://doi.org/komputika.v11i2.6498)

**Lu, Z., & Zhang, Y. (2009). An augmented Lagrangian approach for sparse principal component analysis. *Computational Optimization and Applications, 46*(3), 427-444.** [**https://doi.org/10.1007/s10589-009-9261-0**](https://doi.org/10.1007/s10589-009-9261-0)

**Lubis, A. A. R., Purnama, S. I., & Afandi, M. A. (2023). Sistem Pendeteksi Kantuk Berbasis Metode Haar Cascade Untuk Aplikasi Computer Vision. Techno. Com, 22(3), 589-598.**

**Manullang, S., Simangunsong, B. A., Syahputra, W. I., & Sihombing, A. L. (2023). Penerapan Principal Component Analysis (PCA) dalam Menentukan Faktor Kepuasan Mahasiswa FMIPA Universitas Negeri Medan pada Pembelajaran Daring Menggunakan SIPDA E-Learning UNIMED. Innovative: Journal Of Social Science Research, 3(2), 14123-14138.**

**Mujahid, I. A., & Trianasari, N. (2021). Analisis faktor yang mempengaruhi milenial dalam memilih *smartphone* Samsung di Indonesia. *e-Proceeding of Management, 8*(2), 793.** [**https://repository.telkomuniversity.ac.id/pustaka/167422/analisis-faktor-yang-mempengaruhi-milenial-dalam-memilih-smartphone-samsung-di-indonesia.html**](https://repository.telkomuniversity.ac.id/pustaka/167422/analisis-faktor-yang-mempengaruhi-milenial-dalam-memilih-smartphone-samsung-di-indonesia.html)

**Marinda, E, L., Ibnu, A., & Indra, L, S. (2023). Penerapan PCA (*Principal Component Analysis*) pada Deteksi Outlier untuk Data Text. *e-Proceeding of Engineering, 10*(3). ISSN : 2355-9365**

**Maria, A, K, W., Satria, F, P., & Ninditya, N. (2021). Analisis Faktor Kepuasan Pelanggan terhadap Layanan Perusahaan Daerah Air Minum. *Jurnal Sains Dan Seni ITS, 10*(1), 2337-3520.** [**https://doi.org/10.12962/j23373520.v10i1.60071**](https://doi.org/10.12962/j23373520.v10i1.60071)

**Sari, D. R. P. (2023). Metode Principal Component Analysis (PCA) sebagai Penanganan Asumsi Multikolinearitas. PARAMETER: Jurnal Matematika, Statistika dan Terapannya, 2(02), 115-124.**

**Smith;, Rahmawati, T. (2014). Aplikasi Principal Component Analysis (PCA) untuk mereduksi faktor-faktor yang berpengaruh dalam peramalan konsumsi listrik. Teknomatika, 7(1), 31-42.**

**Susanto, S. M. (2025). Penerapan Principal Component Analysis (PCA) pada Segmentasi Pelanggan dengan Metode RFM. Makalah IF2123 Aljabar Linier dan Geometri, Program Studi Teknik Informatika, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung.**