Exploratory Data Analysis and Unsupervised Learning Methods

Alexander McLain

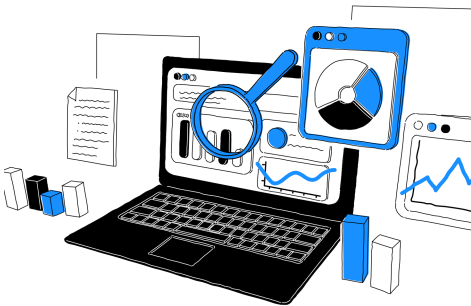
June 17, 2025

Introduction to EDA

- Definition: EDA involves analyzing datasets to summarize their main characteristics, often using visual methods.
- Purpose:
  1. **Identify Patterns:** Understand data distributions and relationships.
  2. **Spot Anomalies:** Detect outliers and errors.
  3. **Test Hypotheses:** Formulate and test initial hypotheses.

Objectives of EDA

- **Data Understanding:** Gain insights into data distributions, structures, and patterns.
- **Hypothesis Testing:** Test assumptions and identify potential relationships.
- **Anomaly Detection:** Identify outliers, errors, and anomalies in the data.
- **Feature Selection:** Determine which variables are important for further analysis or modeling.



Types of EDA

- **Univariate Analysis:** Analysis of a single variable to understand its distribution and characteristics.
- **Bivariate Analysis:** Analysis of two variables to understand relationships and dependencies.
- **Multivariate Analysis:** Analysis of more than two variables to understand complex relationships and interactions.
- **Visual EDA:** Using graphs and plots to visually explore data patterns and trends.

Univariate Analysis

- **Central Tendency:** Measures like mean, median, and mode that indicate the central point of the data.
- **Variability:** Measures such as range, variance, and standard deviation that indicate the spread of the data.
- **Data Distribution:** Understanding the distribution of the data using histograms, box plots, and other visual tools.

EDA Example

Student ID	Gender	Math Score	English Score	Science Score	Study Hours
1	Female	85	88	90	10
2	Male	78	75	80	8
3	Female	92	95	98	15
4	Male	70	65	72	6
5	Female	88	90	94	12

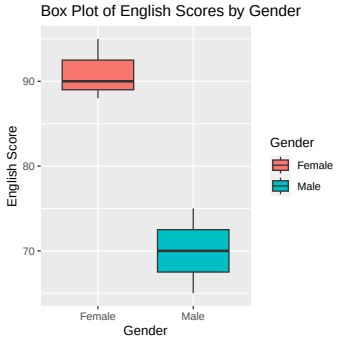
EDA Example

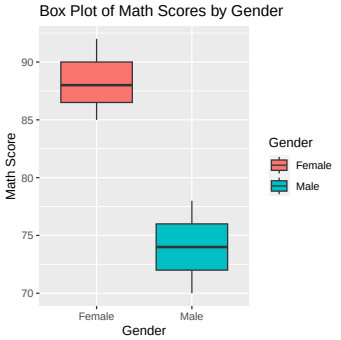
Gender	Statistic	Math Score	English Score	Science Score	Study Hours
n=3 (Female)	Mean	82.6	82.6	86.8	10.2
n=2 (Male)	Median	85.0	88.0	90.0	10.0
	StDev	8.7	12.3	10.6	3.4

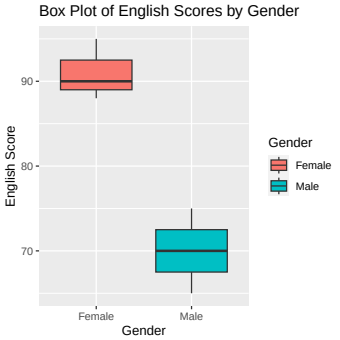
Bivariate Analyses

- **Definition:** Analysis of two variables to understand the relationship between them.
- Techniques:
  1. **Scatter Plots:** Visual representation of the relationship between two variables.
  2. **Correlation Coefficient:** Measures the strength and direction of the relationship.
  3. **Cross-tabulation:** Summarizes data for two categorical variables.
- Purpose: Identifies patterns, relationships, and potential causations between variables.

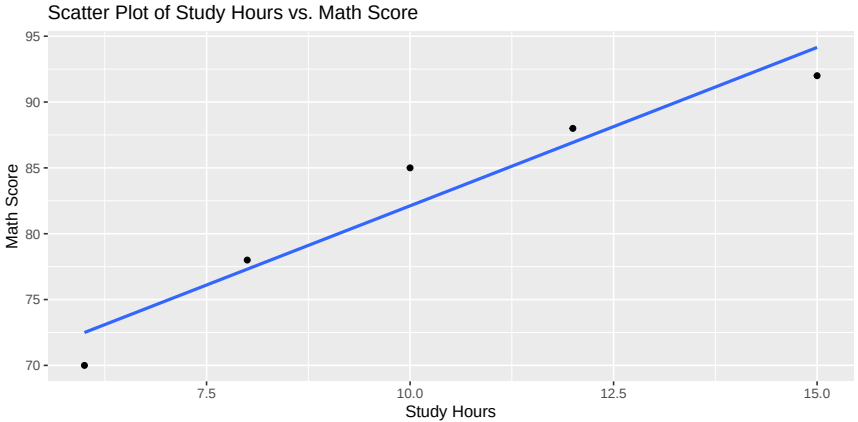
EDA Example


Box Plot of English Scores by Gender
English Score
90
80
70
Female
Male
Gender
Gender
Female
Male


Box Plot of Math Scores by Gender
Math Score
90
85
80
75
70
Female
Male
Gender
Gender
Female
Male


Box Plot of English Scores by Gender
English Score
90
80
70
Female
Male
Gender
Gender
Female
Male

EDA Example


Scatter Plot of Study Hours vs. Math Score
Math Score
95
90
85
80
75
70
7.5
10.0
12.5
15.0
Study Hours

EDA Example

	Math Score	English Score	Science Score	Study Hours
Math Score	1.000	0.995	0.998	0.965
English Score	0.995	1.000	0.996	0.945
Science Score	0.998	0.996	1.000	0.963
Study Hours	0.965	0.945	0.963	1.000

Correlation matrix

Multivariate Analyses

- **Definition:** Analysis involving multiple variables to understand their interactions and combined effects.
- Techniques:
  1. Principal Component Analysis for dimensionality reduction.
  2. Linear Discriminant Analysis for classification tasks.
  3. Cluster Analysis for grouping similar data points.
- **Applications:** Used in identifying patterns, segmenting data, and predictive modeling.

What is Unsupervised Learning?

- Most models you've seen predict something: disease, score, label.
- But what if there's no "right answer" to learn from?
- Unsupervised learning **finds structure** in unlabeled data.
- It lets us explore data, discover patterns, and form new hypotheses.

Why Is It Useful?

- **Public Health:**
  - Discover hidden subgroups of patients
  - Segment behaviors (e.g., diet patterns)
  - Summarize imaging or genomic data
- **No labels? No problem.**
- Unsupervised learning is like being a data detective.

Two Major Tools in Unsupervised Learning

- **Clustering:** groups similar individuals
  - Example: *K-means, hierarchical clustering*
- **Dimensionality Reduction:** simplifies complexity
  - Example: *PCA – principal component analysis*

We'll explore these in detail with hands-on examples.

What is K-Means Clustering?

- **Goal:** Group similar data points into k clusters
- **Unsupervised learning:** No labels, just features
- Useful for:
  - Patient subgroup discovery
  - Market segmentation
  - Dimensionality reduction (as a preprocessing step)

How Does K-Means Work?

1. Choose the number of clusters k
2. Randomly initialize k centroids
3. Assign each point to the **nearest centroid**
4. Recompute each centroid as the mean of its assigned points
5. Repeat steps 3–4 until assignments stop changing

What Does K-Means Optimize?

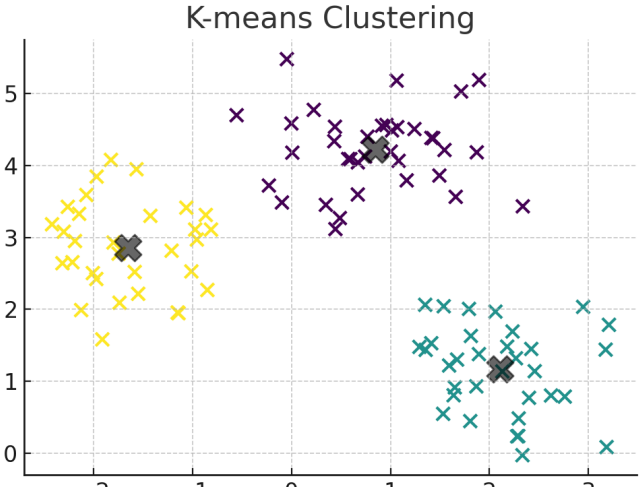
Objective: Minimize total within-cluster variation

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- μ_i = centroid of cluster i
- C_i = set of points assigned to cluster i
- The algorithm finds cluster assignments that minimize this squared distance

Limitations of K-Means

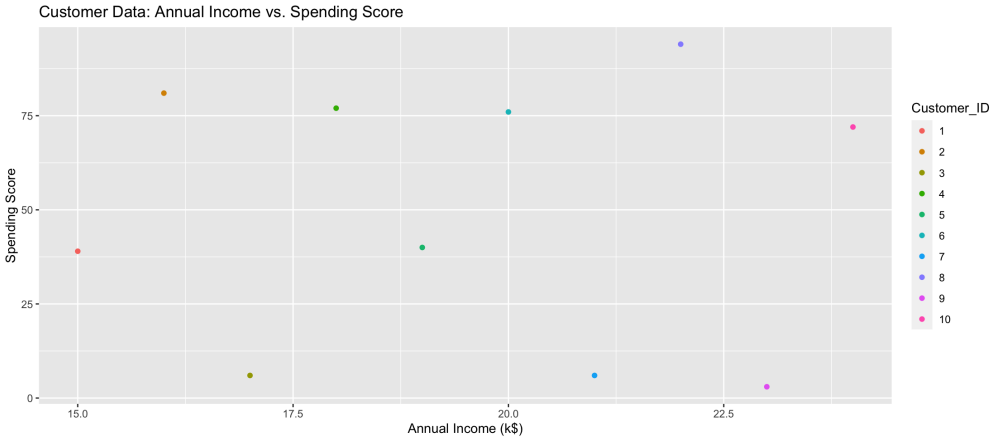
- You must choose k in advance
- Sensitive to initial centroid placement
- Assumes spherical, equal-sized clusters
- Not suitable for clusters of varying density or shape
- Can be improved using **k-means++** initialization or run multiple times

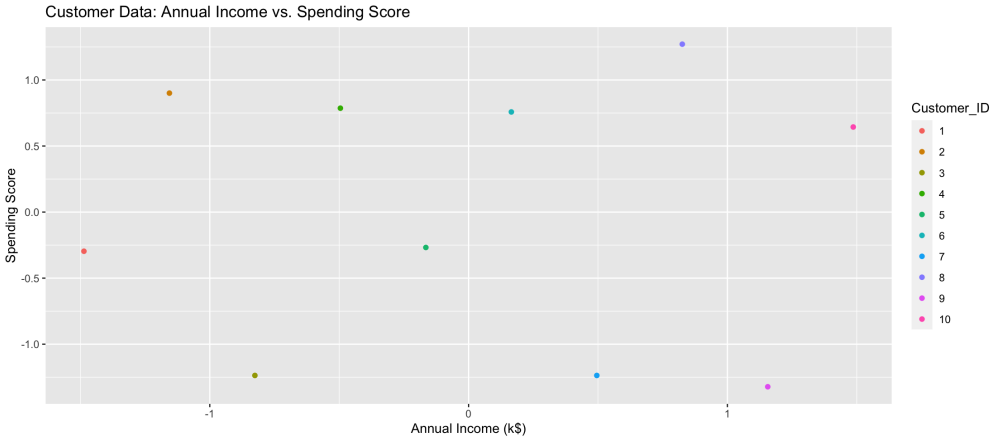

K-means Clustering
5
4
3
2
1
0
−2
−1
0
1
2
3

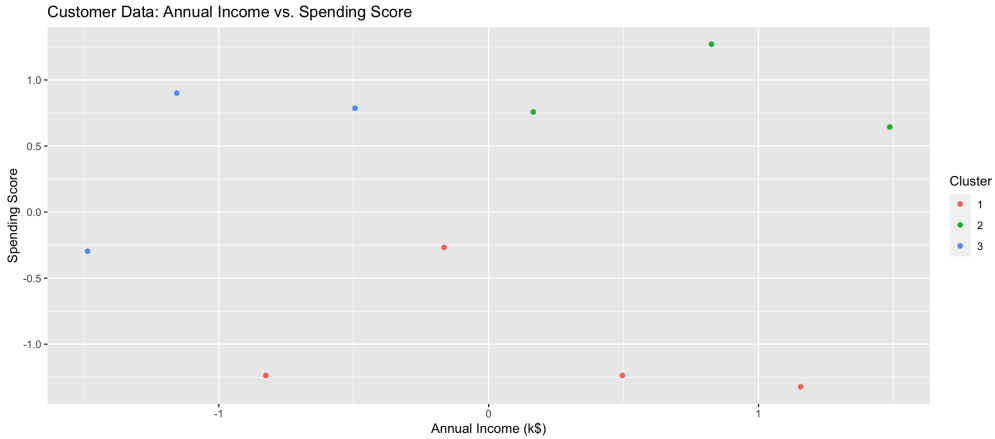
K-means Example

Customer ID	Annual Income	Spending Score
1	15	39
2	16	81
3	17	6
4	18	77
5	19	40
6	20	76
7	21	6
8	22	94
9	23	3
10	24	72

Let's start with observations 3, 4, and 10 as the "centers"


Customer Data: Annual Income vs. Spending Score
Spending Score
75
50
25
0
15.0
17.5
20.0
22.5
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Cluster
1
2
3

What is Hierarchical Clustering?

- A clustering method that builds a hierarchy of clusters
- Does not require you to pre-specify the number of clusters
- Output is typically visualized as a **dendrogram**
- Two main types:
  - **Agglomerative** (bottom-up)
  - **Divisive** (top-down)

Agglomerative Clustering (Bottom-Up)

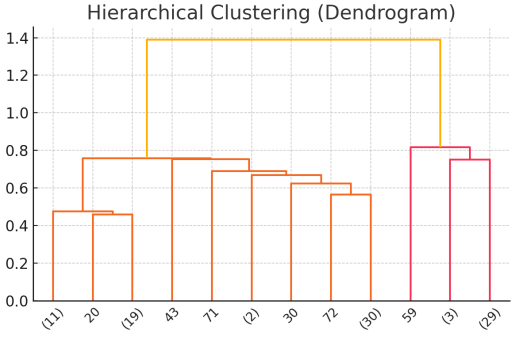
1. Start with each point as its own cluster
2. Compute distances between all pairs of clusters
3. Merge the two closest clusters
4. Repeat until all points belong to one cluster

Linkage criteria (how distances between clusters are measured):

- Single linkage (minimum distance)
- Complete linkage (maximum distance)
- Average linkage (mean distance)

Understanding the Dendrogram

- A dendrogram shows how clusters are merged step-by-step
- The height of each merge indicates the distance between clusters
- You can "cut" the dendrogram at a given height to define k clusters



Strengths and Weaknesses

Strengths:

- No need to pre-specify number of clusters
- Produces a full hierarchy (multi-scale clustering)

Weaknesses:

- Computationally expensive for large datasets
- Sensitive to distance metric and linkage method
- Once a merge is made, it cannot be undone

Hierarchical Example

Customer ID	Annual Income	Spending Score
1	15	39
2	16	81
3	17	6
4	18	77
5	19	40
6	20	76
7	21	6
8	22	94
9	23	3
10	24	72

Example

	1	2	3	4	5	6	7	8	9
2	42.01								
3	33.06	75.01							
4	38.12	4.47	71.01						
5	4.12	41.11	34.06	37.01					
6	37.34	6.4	70.06	2.24	36.01				
7	33.54	75.17	4	71.06	34.06	70.01			
8	55.44	14.32	88.14	17.46	54.08	18.11	88.01		
9	36.88	78.31	6.71	74.17	37.22	73.06	3.61	91.01	
10	34.21	12.04	66.37	7.81	32.39	5.66	66.07	22.09	69.01

Table: Dissemilarity matrix for example data

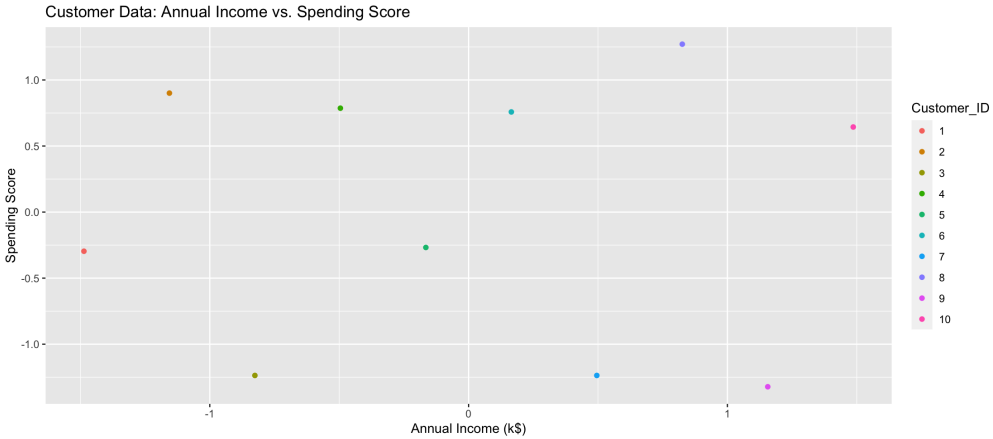
Hierarchical Example

Customer ID	Annual Income (Z-score)	Spending Score (Z-score)
1	-1.49	-0.30
2	-1.16	0.90
3	-0.83	-1.24
4	-0.50	0.79
5	-0.17	-0.27
6	0.17	0.76
7	0.50	-1.24
8	0.83	1.27
9	1.16	-1.32
10	1.49	0.64

Example

	1	2	3	4	5	6	7	8	9
2	1.24								
3	1.15	2.16							
4	1.47	0.67	2.05						
5	1.32	1.53	1.17	1.10					
6	1.96	1.33	2.23	0.66	1.08				
7	2.19	2.70	1.32	2.25	1.17	2.02			
8	2.79	2.02	3.00	1.41	1.83	0.84	2.53		
9	2.83	3.21	1.98	2.68	1.69	2.30	0.67	2.61	
10	3.12	2.65	2.98	1.99	1.89	1.33	2.13	0.91	1.99

Table: Dissemilarity matrix for example data using Z-scores


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Example

	1	2	3	4/6	5	7	8	9
2	1.24							
3	1.15	2.16						
4/6	1.71	1.00	2.14					
5	1.32	1.53	1.17	1.09				
7	2.19	2.70	1.32	2.14	1.17			
8	2.79	2.02	3.00	1.12	1.83	2.53		
9	2.83	3.21	1.98	2.49	1.69	0.67	2.61	
10	3.12	2.65	2.98	1.66	1.89	2.13	0.91	1.99

Table: Dissemilarity matrix for example data after step 1 using the average method.

Example

	1	2	3	4/6	5	7/9	8
2	1.24						
3	1.15	2.16					
4/6	1.71	1.00	2.14				
5	1.32	1.53	1.17	1.09			
7/9	2.51	2.95	1.65	2.31	1.43		
8	2.79	2.02	3.00	1.12	1.83	2.57	
10	3.12	2.65	2.98	1.66	1.89	2.06	0.91

Table: Dissemilarity matrix for example data after step 2 using the average method.

Example

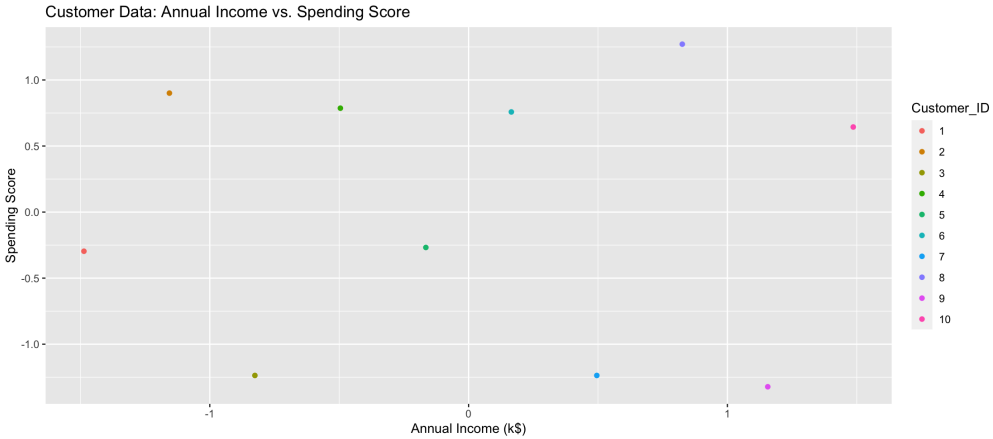
	1	2	3	4/6	5	7/9
2	1.24					
3	1.15	2.16				
4/6	1.71	1.00	2.14			
5	1.32	1.53	1.17	1.09		
7/9	2.51	2.95	1.65	2.31	1.43	
8/10	2.96	2.34	2.99	1.39	1.86	2.32

Table: Dissemilarity matrix for example data after step 3 using the average method.

Example

	1	2/4/6	3	5	7/9
2/4/6	1.48				
3	1.15	2.15			
5	1.32	1.31	1.17		
7/9	2.51	2.63	1.65	1.43	
8/10	2.96	1.86	2.99	1.86	2.32

Table: Dissemilarity matrix for example data after step 4 using the average method.


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Example

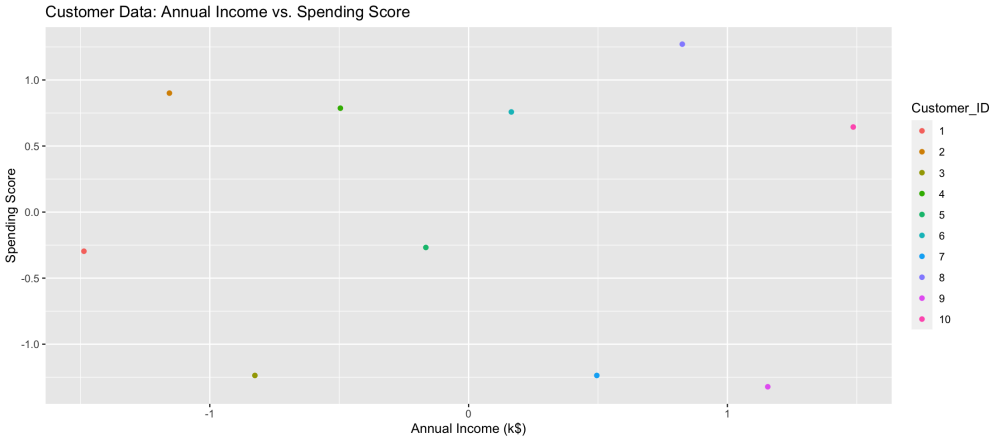
	1/3	2/4/6	5	7/9
2/4/6	1.81			
5	1.25	1.31		
7/9	2.08	2.63	1.43	
8/10	2.97	1.86	1.86	2.32

Table: Dissemilarity matrix for example data after step 5 using the average method.

Example

	1/3/5	2/4/6	7/9
2/4/6	1.56		
7/9	1.76	2.63	
8/10	2.42	1.86	2.32

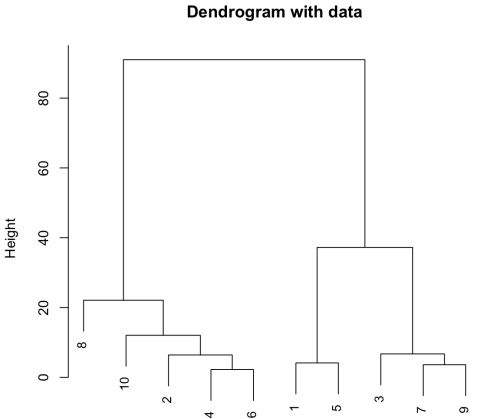
Table: Dissemilarity matrix for example data after step 6 using the average method.

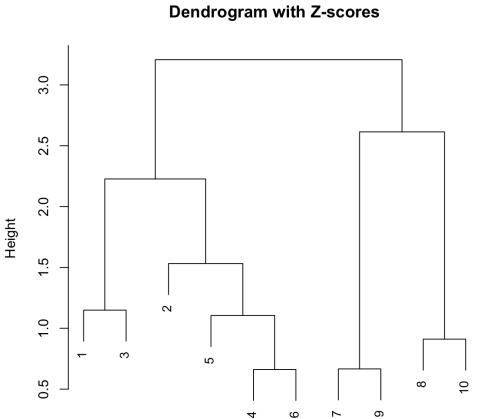

Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

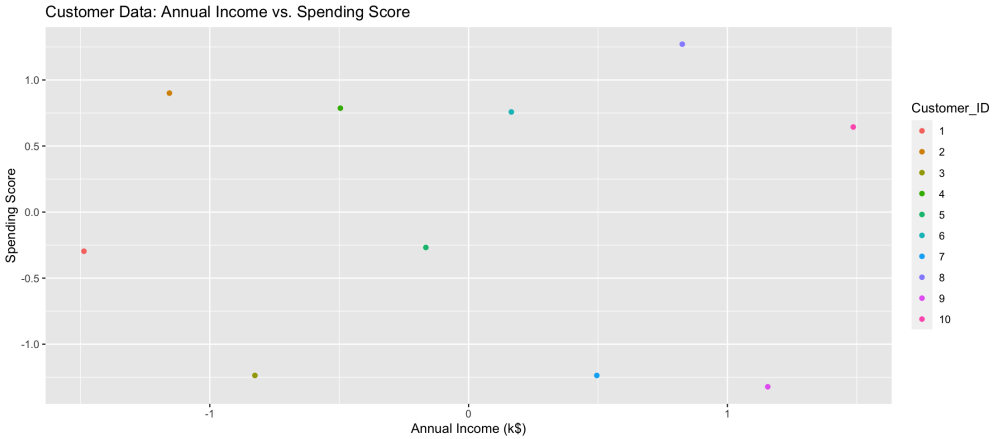
Example

	1/2/3/4/5/6	7/9
7/9	2.20	
8/10	2.14	2.32

Table: Dissemilarity matrix for example data after step 7 using the average method.


Dendrogram with data
Height
0
20
40
60
80
8
10
2
4
6
1
5
3
7
9


Dendrogram with Z-scores
Height
0.5
1.0
1.5
2.0
2.5
3.0
1
3
2
5
4
6
7
9
8
10

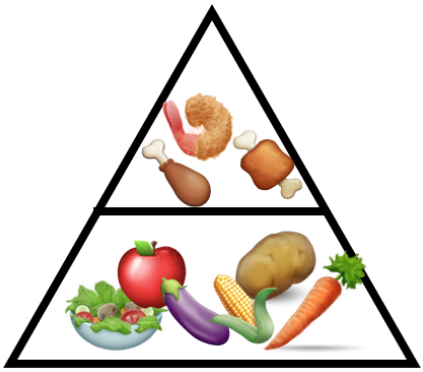

Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Principal Components Introduction

- Principal components explain the variance-covariance structure of a set of variables through some linear combinations.
- Knowing the variables that best differentiate your items has several uses:
  1. Visualization. Using the right variables to plot items will give more insights.
  2. Uncovering Clusters. With good visualizations, hidden categories or clusters could be identified.

Introduction

- Consider data from the United States Department of Agriculture.
- This data contains the nutritional content of one serving of several food items (Parsley, Kale, Broccoli, Corn, Chick, Beef, etc.).
- Four nutrition variables were analyzed: Vitamin C, Fiber, Fat and Protein.
- We want to consider making one variable that fully characterizes the food.




Vitamin C
Parsley
Kale
Broccoli
Cauliflower
Soybeans
Yam
Guinea Hen
(Vitamin C) - (Fat)
Parsley
Kale
Broccoli
Cauliflower
Cabbage
Spinach
Yam
Sweet Corn
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb
(Vitamin C + Fiber) - (Fat)
Parsley
Peas
Lotus Root
Chives
Cauliflower
Soybeans
Eggplant
Sweet Corn
Mushrooms
Haddock
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb

Introduction

- Principal components explain the variance-covariance structure of a set of variables through some linear combinations.
- $\boldsymbol{X}$: variables which we want to explain the variance-covariance matrix
- $\boldsymbol{Z}$: a linear combination of the X's (this is not the response)

$$
\begin{aligned}
\boldsymbol{Z}_1 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{11}\boldsymbol{X}_1 + v_{12}\boldsymbol{X}_2 + \ldots + v_{1p}\boldsymbol{X}_p \\
\boldsymbol{Z}_2 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{21}\boldsymbol{X}_1 + v_{22}\boldsymbol{X}_2 + \ldots + v_{2p}\boldsymbol{X}_p \\
\vdots \quad & \quad \vdots \quad \vdots \\
\boldsymbol{Z}_p &= \boldsymbol{X}\boldsymbol{v}_1 = v_{p1}\boldsymbol{X}_1 + v_{p2}\boldsymbol{X}_2 + \ldots + v_{pp}\boldsymbol{X}_p
\end{aligned}
$$

PC goal

Main idea of Principal Components

- Set $\boldsymbol{Z}_1 = \boldsymbol{X}\boldsymbol{v}_1$ such that the variance of $\boldsymbol{Z}_1$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$.
- Set $\boldsymbol{Z}_2 = \boldsymbol{X}\boldsymbol{v}_2$ such that the variance of $\boldsymbol{Z}_2$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = 0$.

⋮

- Set $\boldsymbol{Z}_i = \boldsymbol{X}\boldsymbol{v}_i$ such that the variance of $\boldsymbol{Z}_i$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_k, \boldsymbol{Z}_i) = 0$ for all $k = 1, 2, \ldots, i = 1$.

⋮

These $\boldsymbol{v}$ values can be found using some linear algebra techniques.

	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

Figure: Factor loadings for the food example.

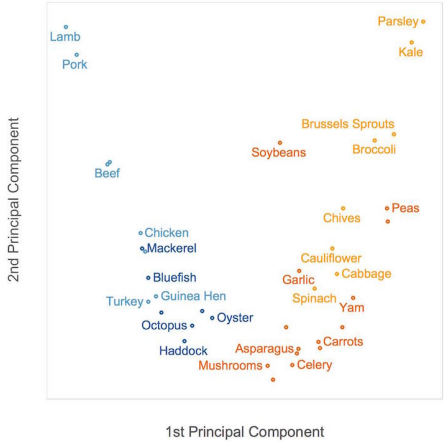


Figure: First and second PCs for the food example.

Properties

- Let d_k denote the variance of Z_k
- The total proportion variability that can be attributed to Z_k is

$$\frac{d_k}{\sum_{i=1}^{p} d_i}$$

- Suppose

$$\frac{d_1 + d_2 + d_3}{\sum_{i=1}^{p} d_i} = 0.97.$$

what would this imply about $Z_4, \ldots, Z_p$?

How many PC's to use

- If $\frac{\sum_{i=1}^{r} d_i}{\sum_{i=1}^{p} d_i}$ is "close to one," then only the first r PC's are needed.
- Scree plot: a plot of d_i used for selecting r
- *Kaiser's Rule* for the PC of a correlation matrix cutoff at 1.
- PC's are not invariant to scaling

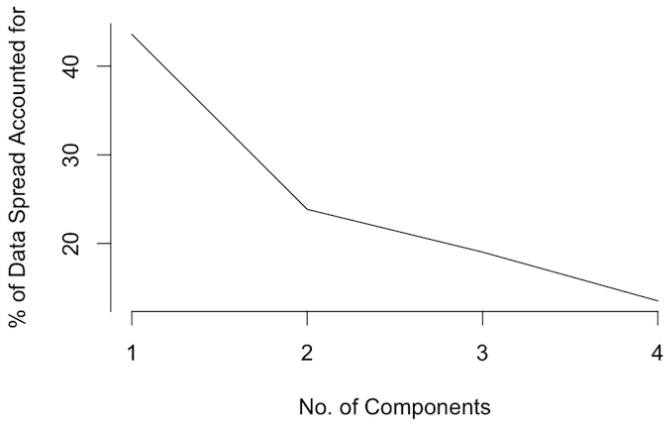


Figure: The total proportion variability attributed to the PCs.

Principal components regression

- Principal components regression (PCR) forms the linear combination $\boldsymbol{z}_m = \boldsymbol{X}\boldsymbol{v}_m$ and then regresses $\boldsymbol{y}$ on $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M$ for some $M \leq p$.

Principal components clustering

- Principal components clustering is simply clustering the data based on $M \leq p$ principal components.
- For example, we could cluster the food data based on the first two PCs