

Lab4: K-Means Clustering in KNIME

This lab demonstrates how to perform **customer segmentation** using **K-Means clustering** in KNIME. You will:

- normalize numeric features,
- search for the **best number of clusters (K)** using **Silhouette Coefficient**,
- run the final clustering using the selected K,
- visualize and interpret clusters.

Learning goals

By the end of this lab, you can:

1. Prepare numeric features for distance-based clustering.
2. Explain why **normalization** is required for k-means.
3. Use a loop to evaluate K candidates and choose best K using **Silhouette**.
4. Interpret cluster profiles using aggregated statistics and scatter plots.

Dataset Description

The **Mall Customer** dataset contains **200 customers**. Each row is one subject. The goal is to predict the **Spending Class (High or Low Spender)** of the customer.

Target Variable

High Spender: “yes” if spending score ≥ 60 ; else “no”.

Features:

CustomerID

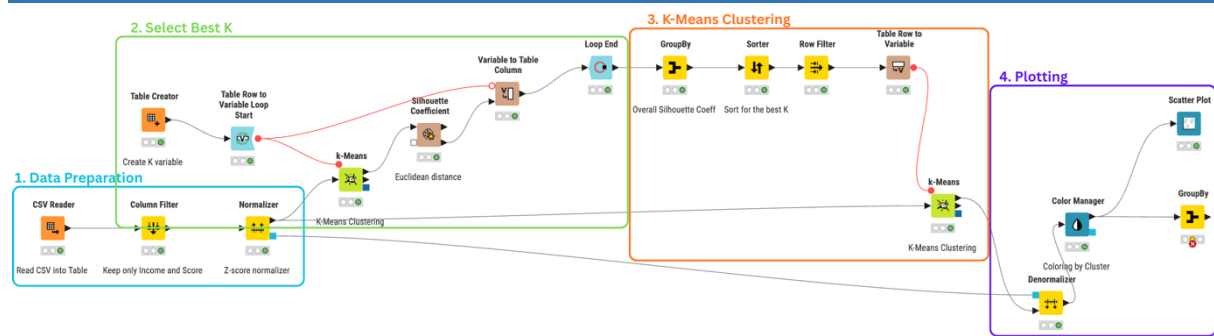
Gender: Male/Female

Age: (years)

Annual Income: (k\$)

Spending Score: Derived Metric (0-100)

KNIME Instructions



1) Data Preparation (Blue block)

1.1 CSV Reader

- Load your dataset (e.g., Mall_Customers.csv).
- Confirm the column types are correct.

1.2 Column Filter

- Keep only:
 - Age
 - Annual Income (k\$)

Column Filter

Column filter

Manual

Wildcard

Regex

Type

Q

Search

Aa

Excludes

123

CustomerID

T

Gender

123

Spending Score (1-1...

>

>>

<

<<

Any unknown column

Includes

123

Age

123

Annual Income (k\$)

1.3 Normalizer

Normalizer

Number columns

Manual Wildcard Regex Type

Search Aa

Excludes

No columns in this list.

Includes

Age

Annual Income (k\$)

Any unknown column

Normalization method

Min-max Z-score Decimal scaling

2) Select Best K (Green block)

2.1 Table Creator (Create K candidates)

Create a 1-column table named k containing:

- 2, 3, 4, 5, 6, 7, 8, 9, 10

2.2 Table Row to Variable Loop Start

- Loop over each row (each candidate K)
- Create flow variable for k

2.3 k-Means (inside loop)

- Number of clusters: controlled by flow variable k
- Input data: the normalized table from the blue block
- Seed: 2026

2.4 Silhouette Coefficient

- Distance: Euclidean
- Cluster column: the cluster output from k-means
- Output: Overall silhouette coefficient (higher is better)

2.5 Variable to Table Column → Loop End

- Convert silhouette metric into a table column

Variable to Table Column

Output as rows

Manual
Wildcard
Regex
Type

Search
Aa

Excludes

RowID

currentIteration

maxIterations

knime.workspace

Any unknown variable

Includes

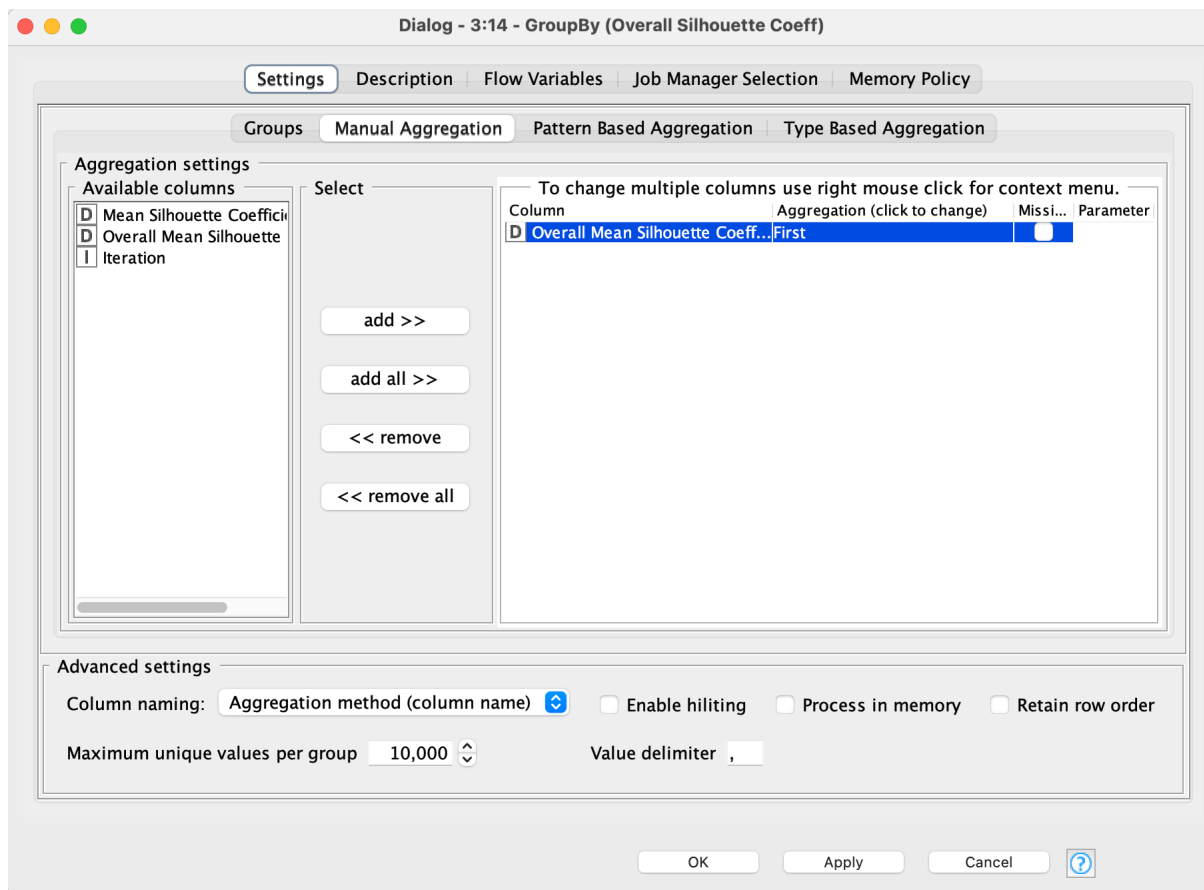
k

Overall Mean Silhou...

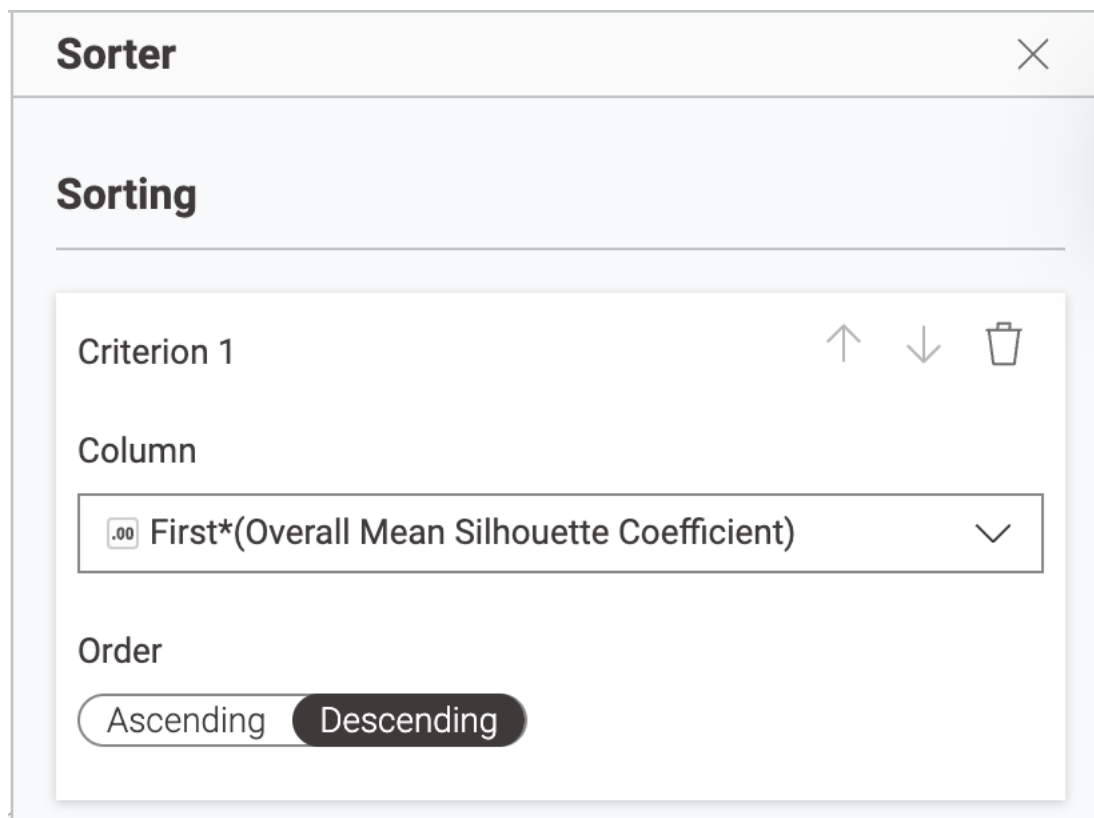
>
>>
<
<<

3) K-Means Clustering with Best K (Orange block)

3.1 GroupBy



3.2 Sorter



3.3 Row Filter

- Keep only **top 1 row** (best K)

Row Filter

Filter

Criterion 1

Filter column

Operator

Row number

Equals

Value

1

3.4 Table Row to Variable

- Convert best K row into a flow variable: best_k

Table Row to Variable

Output as variables

Excludes

No columns in this list.

Includes

123 k

.00 First*(Overall Mean ...

Any unknown column

>

>>

<

<<

3.5 k-Means (final run)

- Run k-means again using best_k
- Output: final cluster assignment per customer
- Seed: 2026

4) Plotting & Interpretation (Purple block)

4.1 Denormalizer (optional but recommended)

- Convert normalized values back to original scale for plotting.

4.2 Color Manager

- Color by cluster label.

4.3 Scatter Plot

- X-axis: Annual Income (k\$)
- Y-axis: Spending Score (1-100)
- Color: cluster

Scatter Plot

Data

Horizontal dimension

.00

Age

▼

Vertical dimension

.00

Annual Income (k\$)

▼

Color dimension

T

Cluster

▼