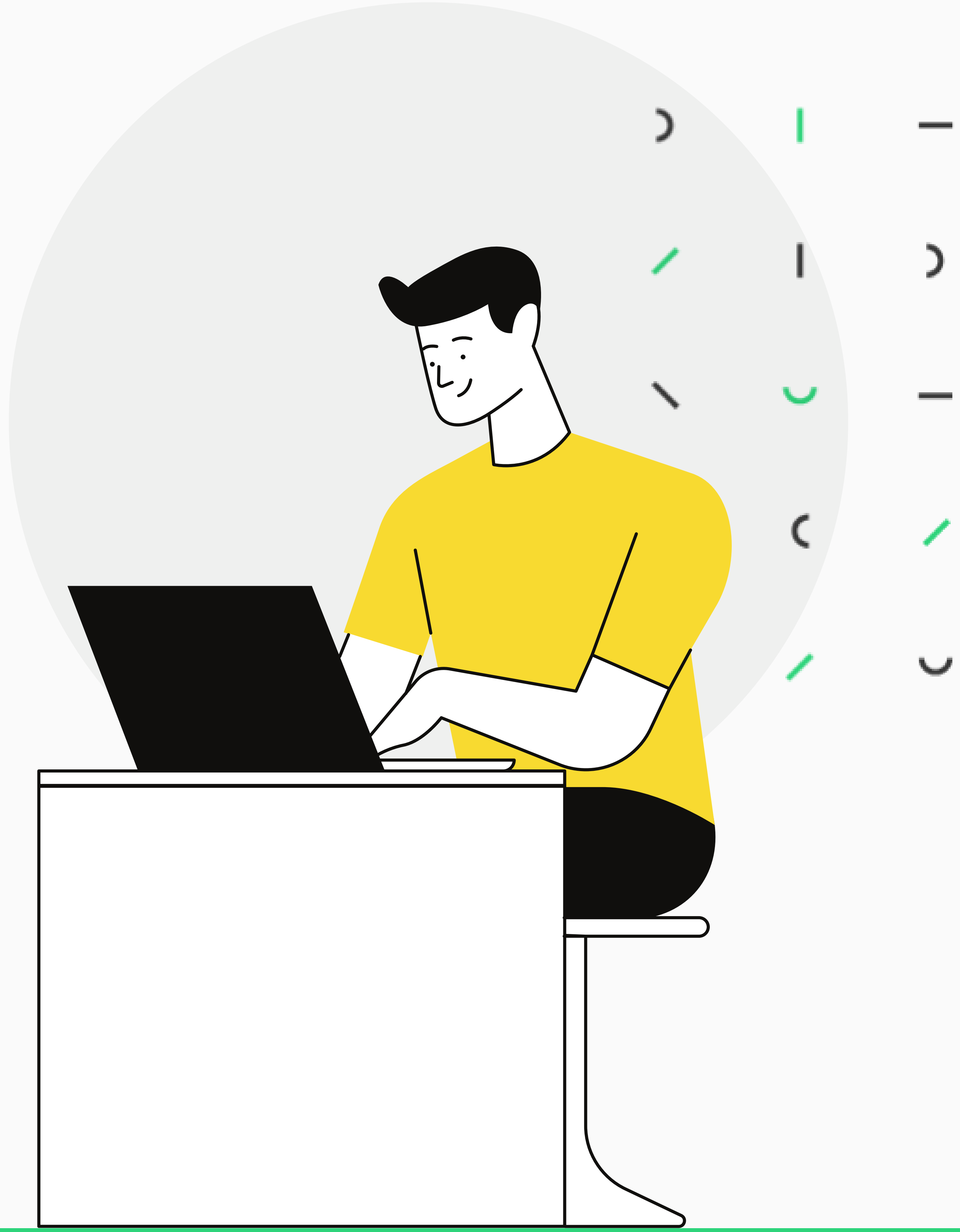
Data Mining with Python

CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING: A DATA MINING APPROACH WITH PYTHON

01

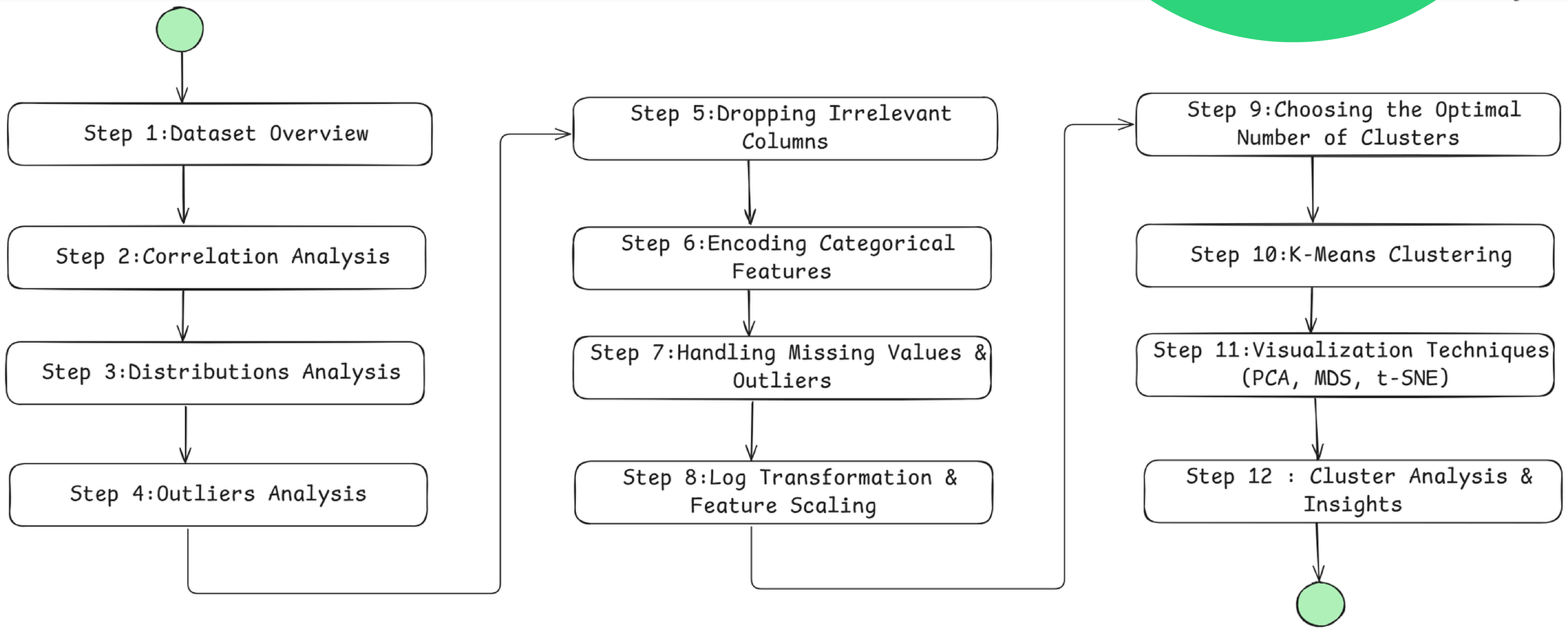


MOHAMED EL AMRAOUI , ERASMUS WORKSHOP, FEBRUARY 2025

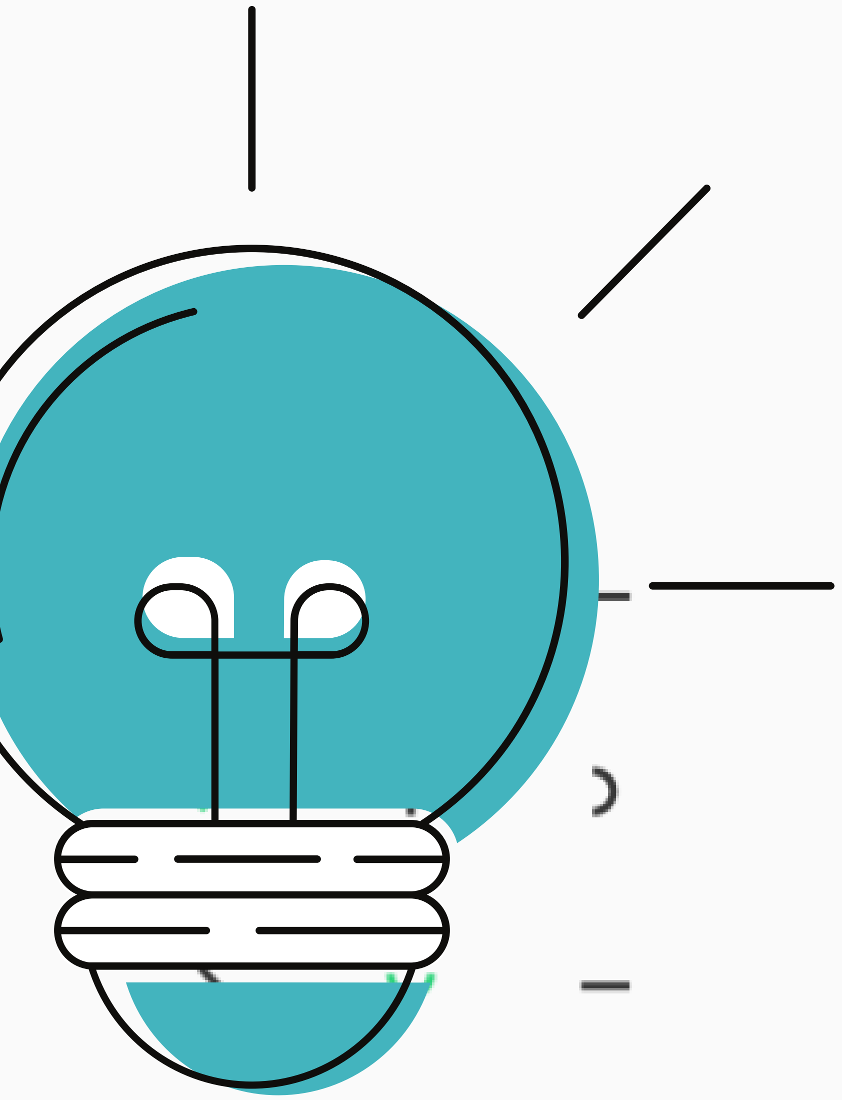
Presentation Plan

- Part 1: Exploratory Data Analysis (EDA)
- Part 2: Data Preparation
- Part 3: Clustering & Dimensionality Reduction
- Part 4: Cluster Analysis & Insights
- Conclusion & Next Steps

Data Preprocessing



Part 1: Exploratory Data Analysis (EDA)



04

Introduction to the Dataset

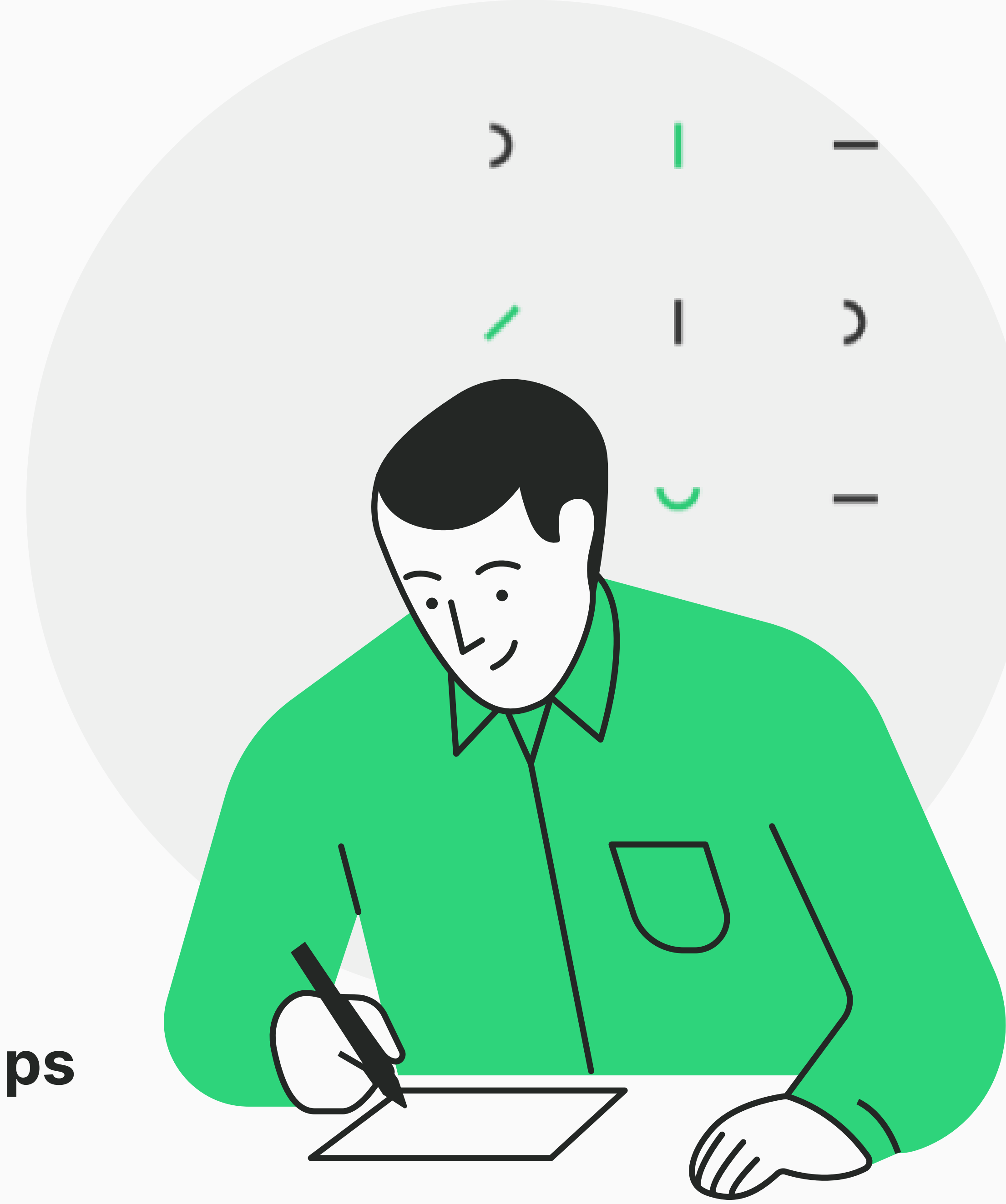
DATASET OVERVIEW

- Contains **bank client** data and previous marketing campaign details.
- Goal: **Cluster customers** based on features and compare groups with the target variable ("**subscribed**").

MAIN DATA CATEGORIES:

🔹 **Client Info:** Age, Job, Marital Status, Education
🔹 **Financial:** Balance, Loans, Credit Default
🔹 **Last Contact:** Contact Method, Call Duration, Last Contact Date
🔹 **Campaign Data:** Previous Contacts, Campaign Outcome
🔹 **Target Variable:** Subscribed (Yes/No)

✅ **17 Columns** | ✅ **K-Means Clustering** | ✅ **Goal: Identify behavioral groups**



Dataset Summary

KEY INSIGHTS FROM DF.INFO()

- **Total Rows:** 2,000
- **Total Columns:** 17
- **Data Types:**
    - ◆ **Numerical (7):** age, balance, day, duration, campaign, pdays, previous
    - ◆ **Categorical (10):** job, marital, education, default, housing, loan, contact, month, poutcome, subscribed
- **Missing Values:**
    - ◆ **age** (12 missing)
    - ◆ **job** (10 missing)
    - ◆ **education** (104 missing)
    - ◆ **contact** (191 missing)
    - ◆ **poutcome** (454 missing)

Correlation Heatmap Analysis

📌 **KEY INSIGHTS:**

- No strong correlations between variables.
- Weak positive correlations:

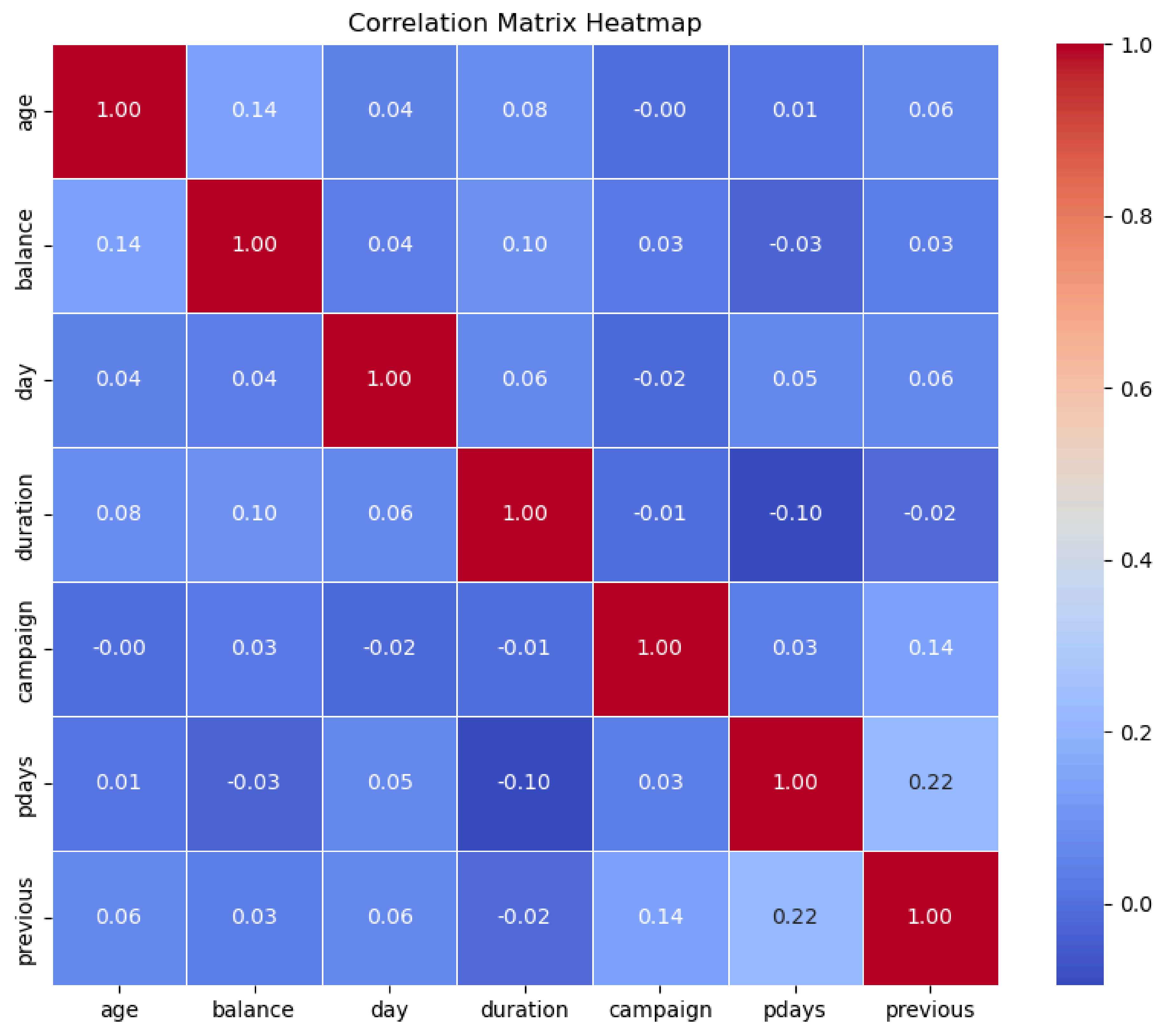
  🔹 **age & balance** (0.14) → Older clients have slightly higher balances.

  🔹 **campaign & previous** (0.14) → Clients contacted before are slightly more likely to be re-contacted.
- Weak negative correlations:

  🔹 **duration & pdays** (-0.10) → Longer gaps between contacts may result in shorter conversations.

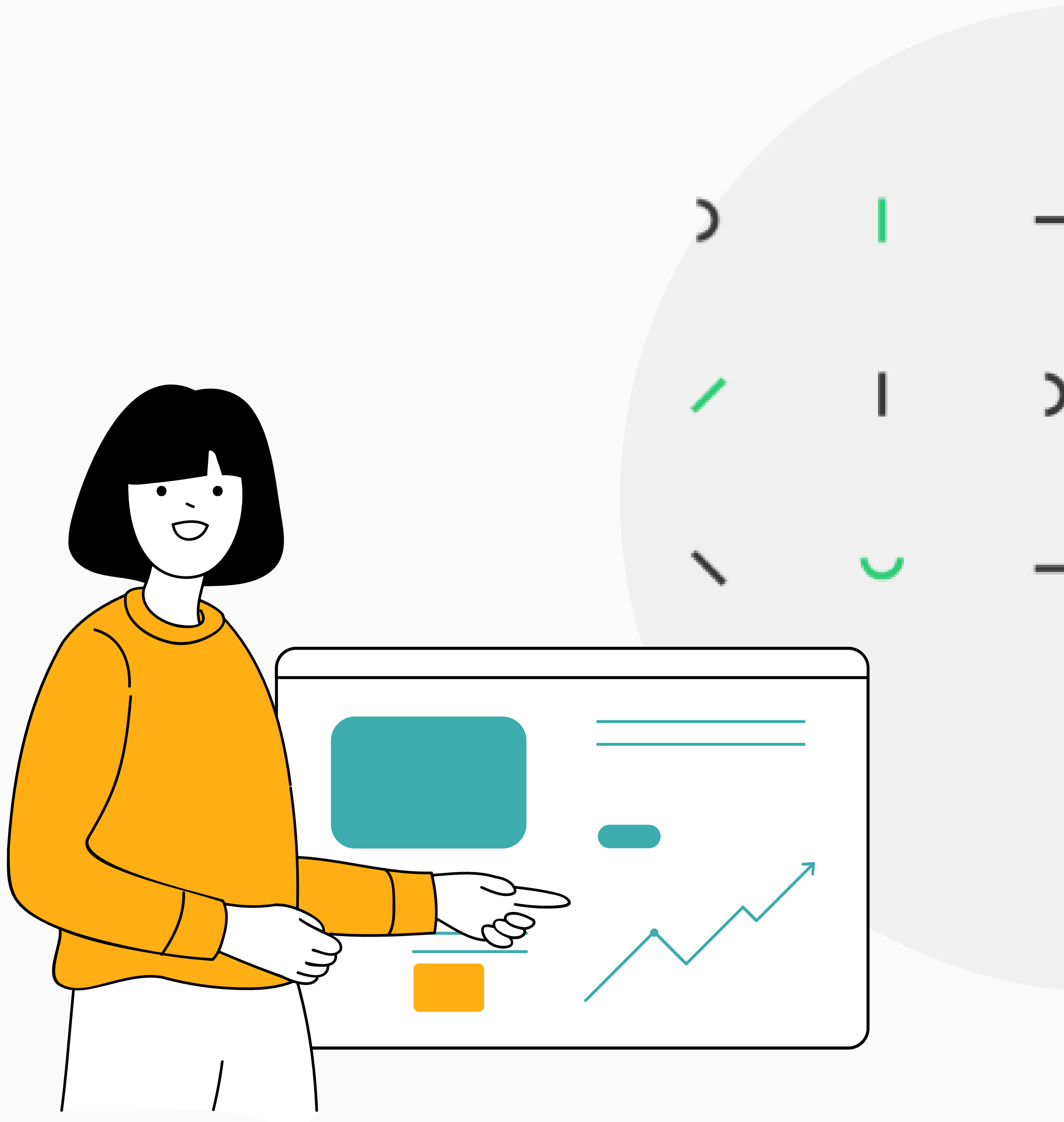
🎯 **CONCLUSION:**

- Variables are mostly independent → Suitable for clustering analysis.

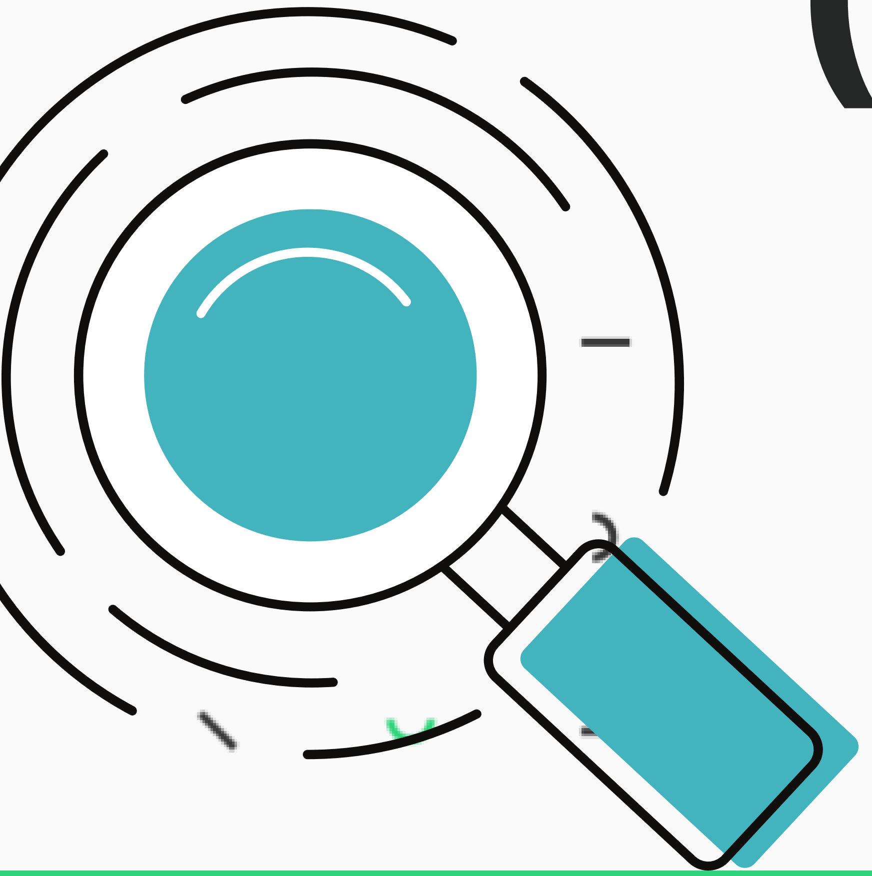


Interpretation of the Pair Plot

- **Distributions:** Most variables are right-skewed, indicating a concentration of lower values.
- **Relationships:** Weak correlations between most variables, with some clustering patterns.
- **Outliers:** Some extreme values exist, particularly in balance and duration.
- **Insights:** Data suggests distinct subgroups, requiring further analysis for meaningful patterns.

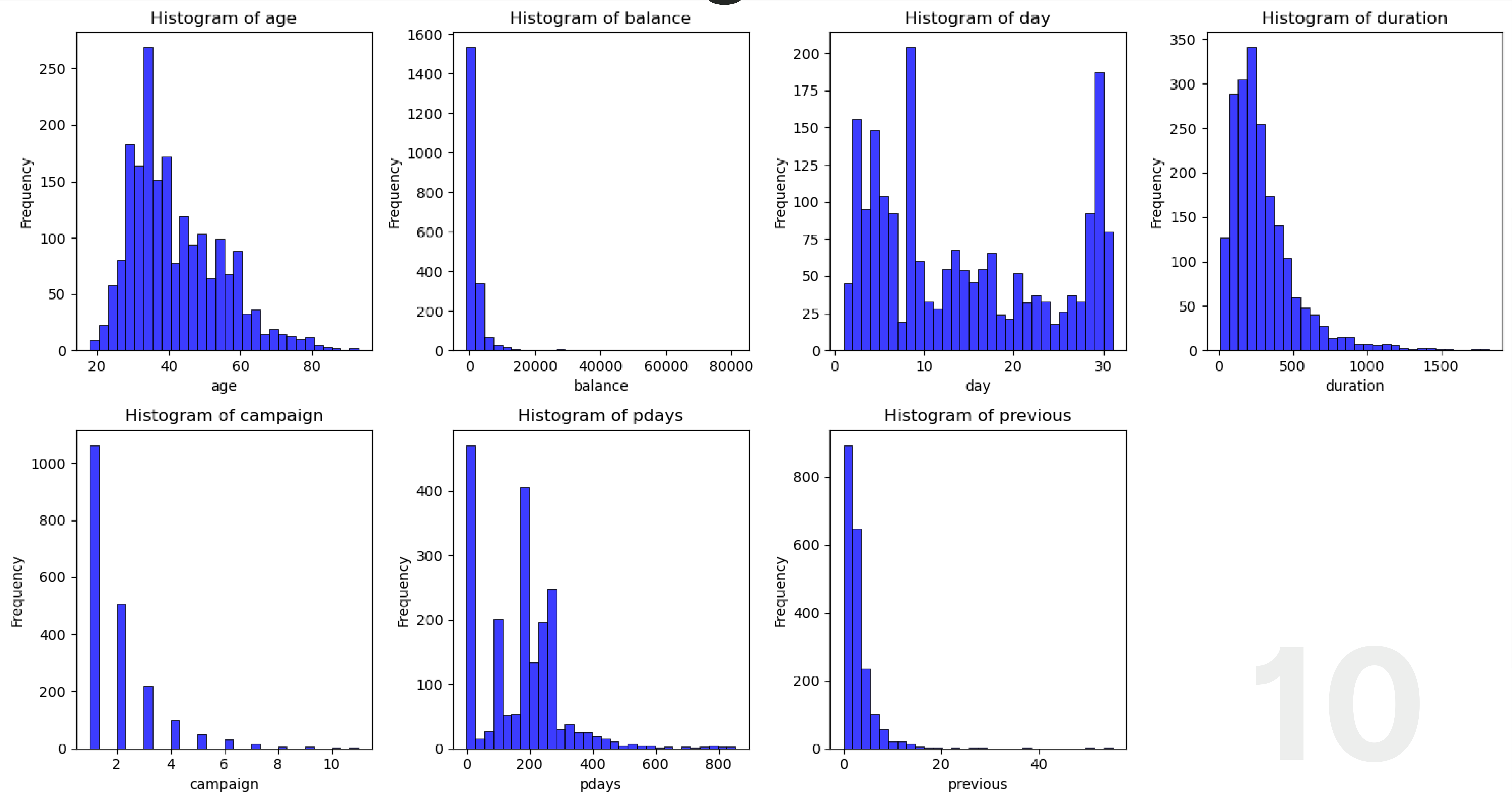


Distribution Analysis (Numerical Features)



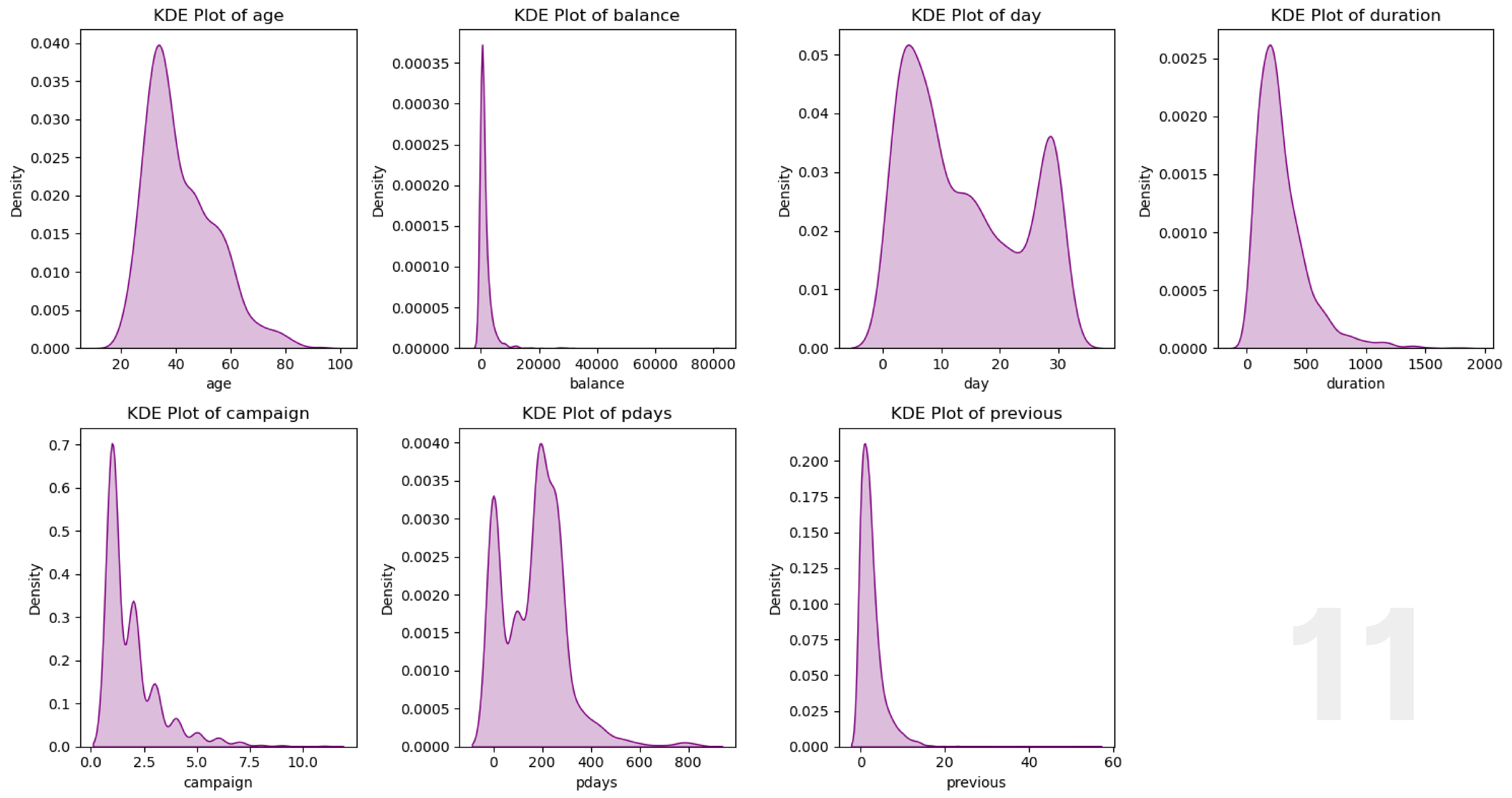
09

Histograms



Histogram of age
Frequency
age
Histogram of balance
Frequency
balance
Histogram of day
Frequency
day
Histogram of duration
Frequency
duration
Histogram of campaign
Frequency
campaign
Histogram of pdays
Frequency
pdays
Histogram of previous
Frequency
previous

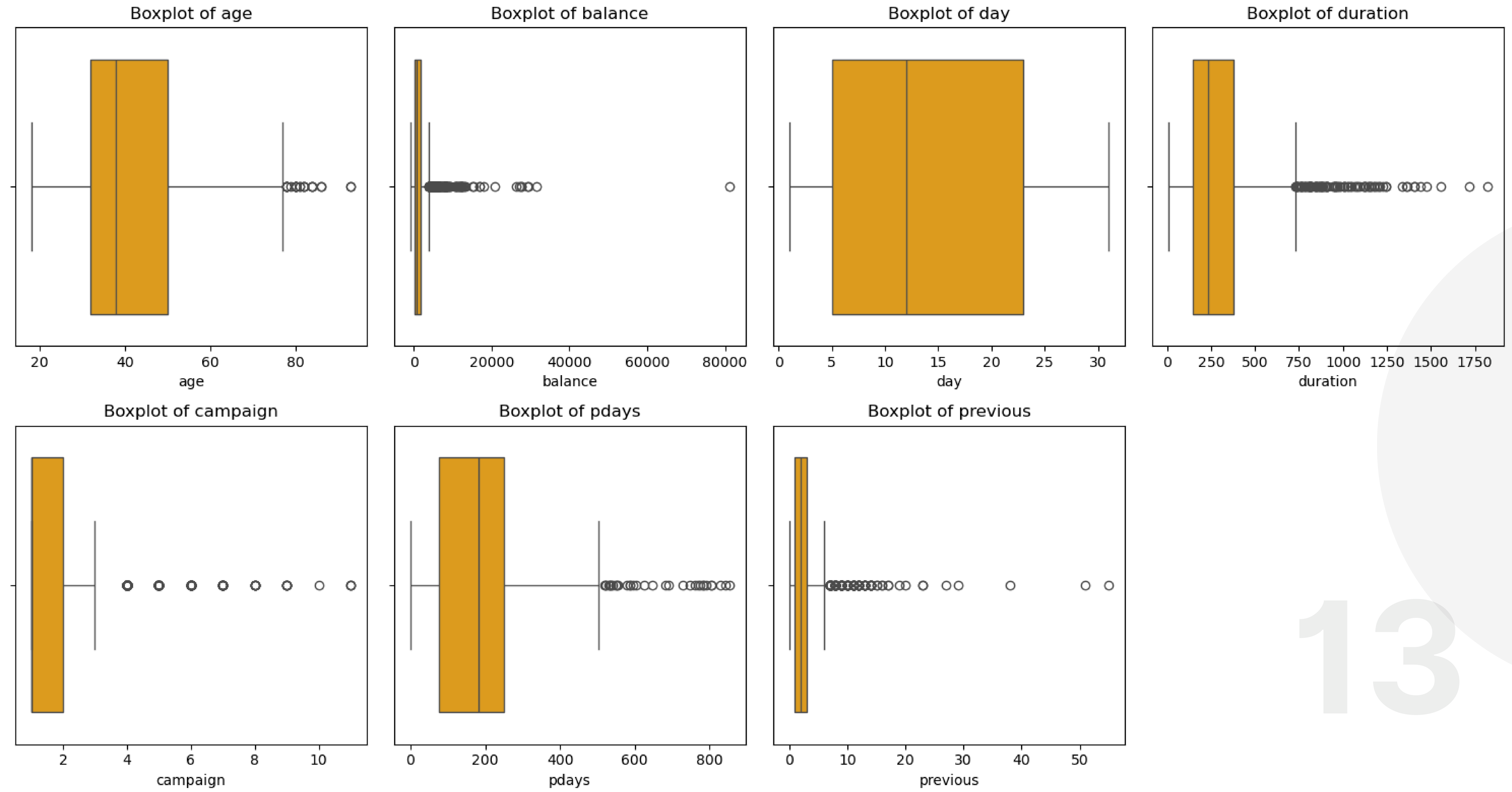
KDE Plots


KDE Plot of age
Density
age
KDE Plot of balance
Density
balance
KDE Plot of day
Density
day
KDE Plot of duration
Density
duration
KDE Plot of campaign
Density
campaign
KDE Plot of pdays
Density
pdays
KDE Plot of previous
Density
previous

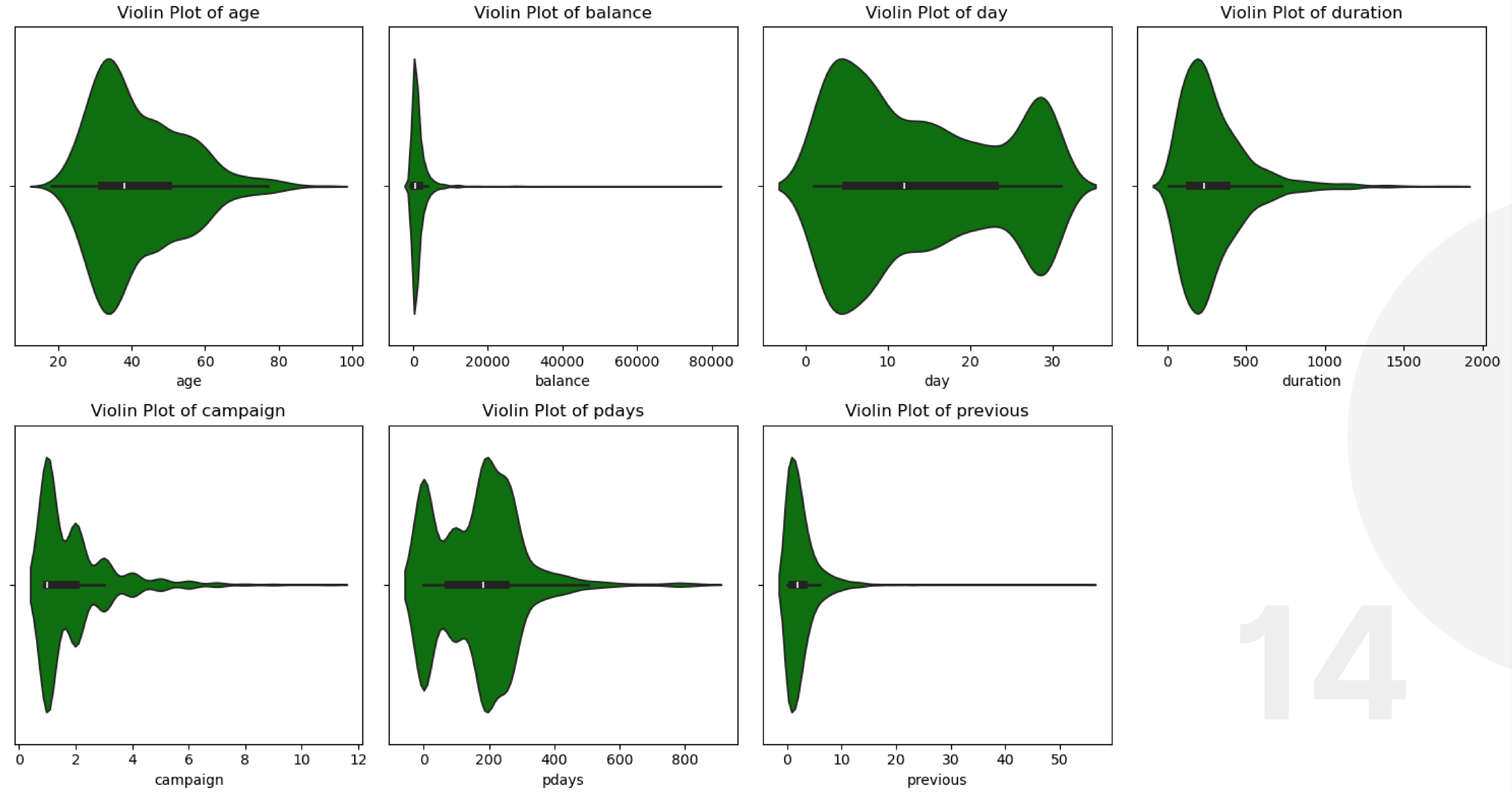
Interpretation of Histograms

- **Age:** Most clients are between 30 and 60 years old, with fewer younger and older clients.
- **Balance:** Highly skewed, meaning most clients have low balances, while a few have very high ones.
- **Day:** Calls were mostly made at the beginning and end of the month.
- Duration: Right-skewed, indicating that most calls are short, but some last significantly longer.
- **Campaign:** Most clients received only a few contacts, with rare cases of high-frequency contacts.
- **Pdays:** Peaks at specific intervals, possibly due to structured follow-up schedules.
- **Previous:** Most clients had very few prior contacts, suggesting many are new leads.

Boxplots


Boxplot of age
20
40
60
80
age
Boxplot of balance
0
20000
40000
60000
80000
balance
Boxplot of day
0
5
10
15
20
25
30
day
Boxplot of duration
0
250
500
750
1000
1250
1500
1750
duration
Boxplot of campaign
2
4
6
8
10
campaign
Boxplot of pdays
0
200
400
600
800
pdays
Boxplot of previous
0
10
20
30
40
50
previous

Violin Plots



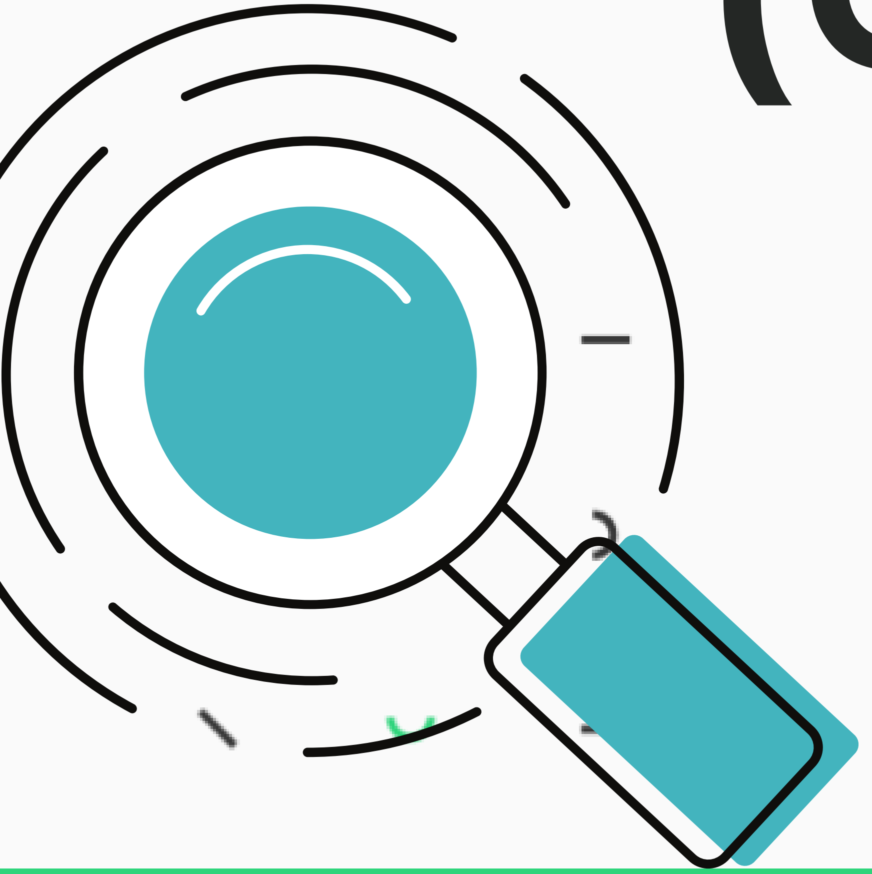
Interpretation of Boxplots & Violin Plots

- **Age:** Mostly 30-40 years, few at extremes.
- **Balance:** Highly skewed with many low values, some extreme outliers.
- **Day:** Evenly spread contacts across the month.
- **Duration:** Right-skewed; mostly short calls, some very long with many outliers .
- **Campaign:** Few contacts for most; some had many.
- **Pdays:** Peaks at lower values, indicating long gaps before contact.
- **Previous:** Most had few past contacts, some over 50 times.

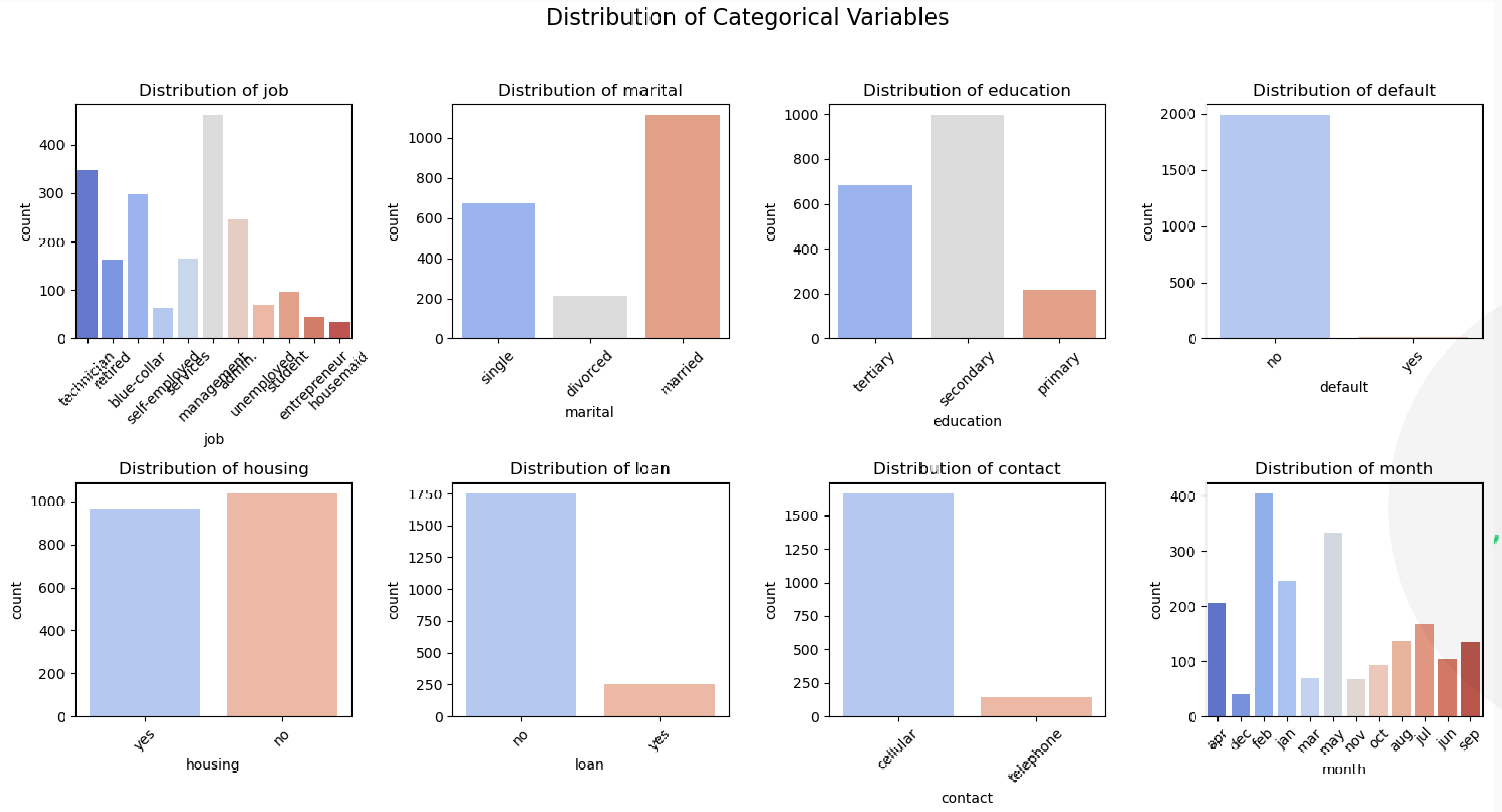
KEY INSIGHTS:

Many variables are right-skewed with outliers (balance, duration, campaign). Most clients had little or no prior contact. Call duration and past interactions may predict subscriptions.

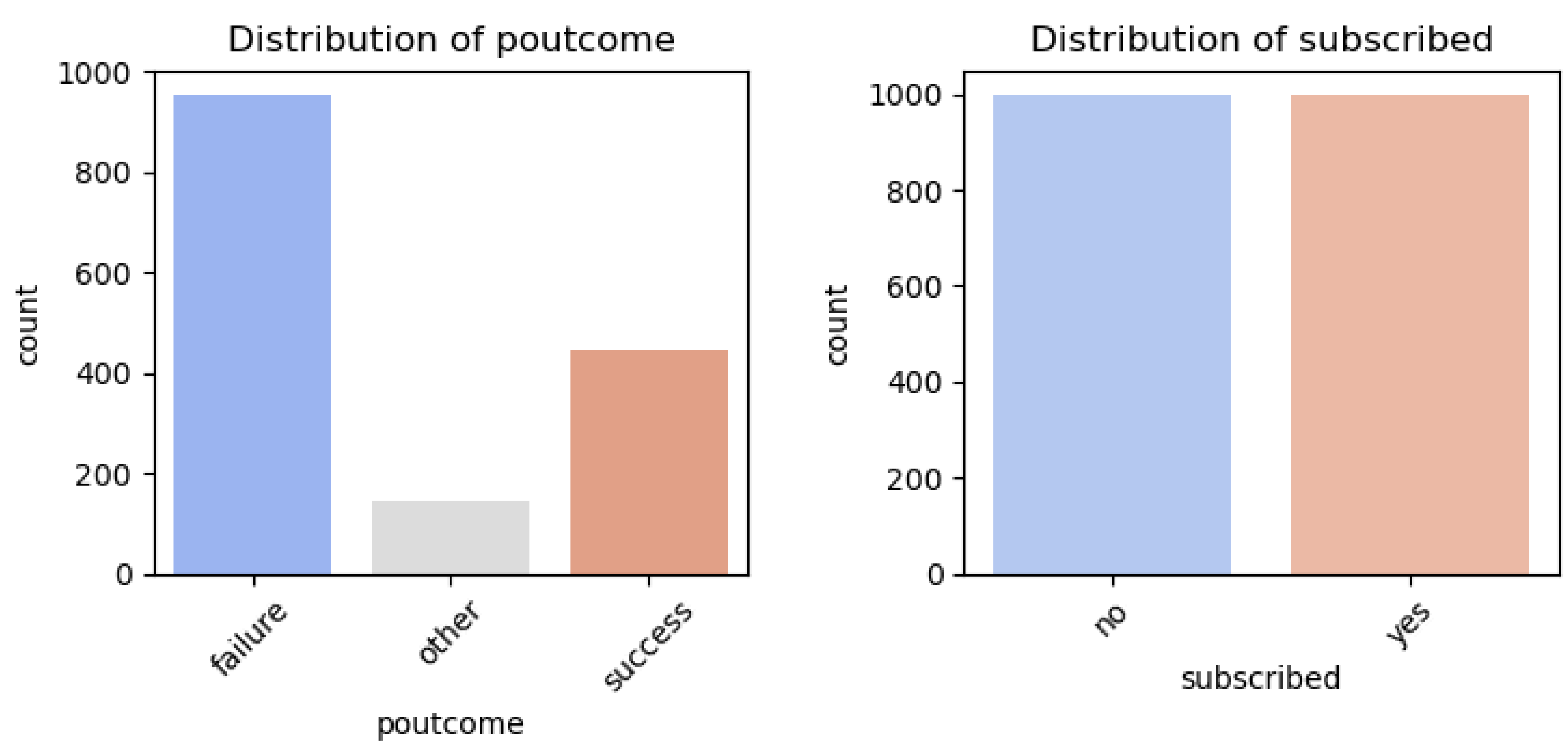
Distribution Analysis (Categorical Features)



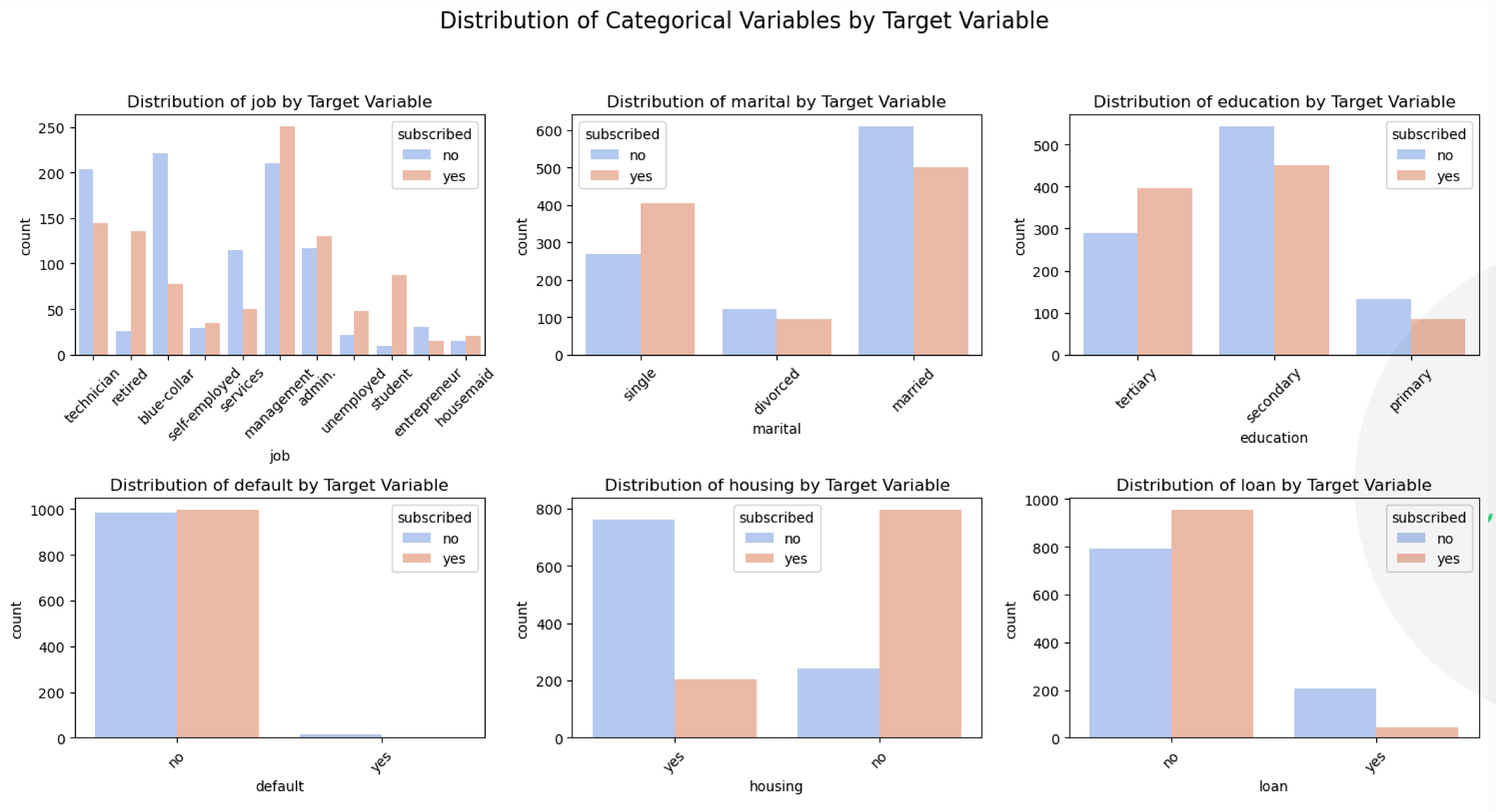
16

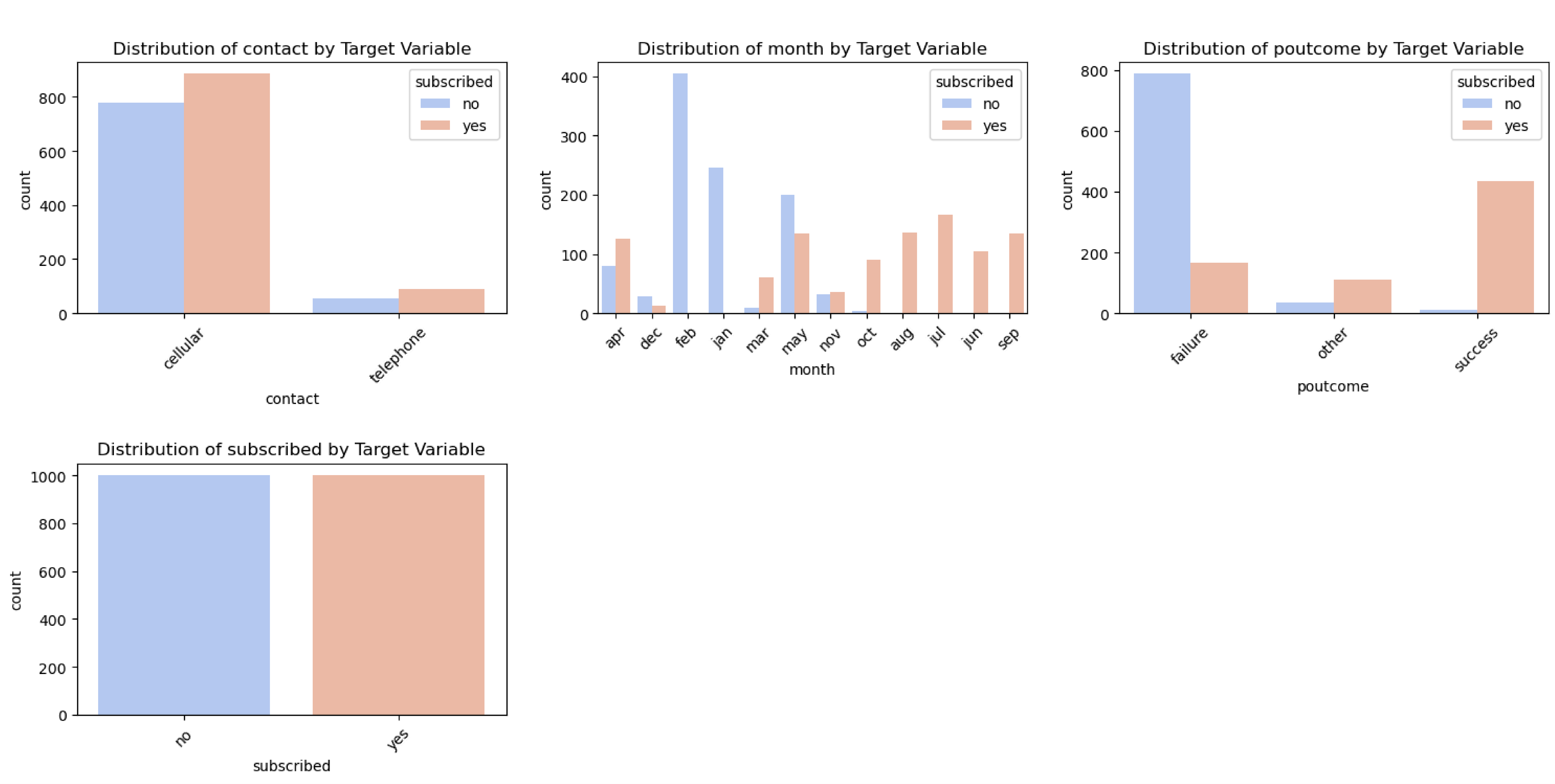


Distribution of Categorical Variables
Distribution of job
Distribution of marital
Distribution of education
Distribution of default
Distribution of housing
Distribution of loan
Distribution of contact
Distribution of month
count
job
marital
education
default
housing
loan
contact
month
technician
retired
blue-collar
self-employed
services
management
admin.
unemployed
student
entrepreneur
housemaid
single
divorced
married
tertiary
secondary
primary
no
yes
cellular
telephone
apr
dec
feb
jan
mar
may
nov
oct
aug
jul
jun
sep



- **Job:** Most clients are in management, followed by technicians and blue-collar workers.
- **Marital Status:** Majority are married, followed by single and divorced.
- **Education:** Most have secondary or tertiary education; few have primary.
- **Default:** Very few clients have credit defaults.
- **Housing & Loan:** Homeownership is balanced, but most don’t have personal loans.
- **Contact Method:** Most contacts were via cellular rather than telephone.
- **Campaign Month:** Contacts peak in February, May, and jan.
- **Previous Outcome:** More failures than successes in past campaigns.
- **Subscription:** Most clients did not subscribe, but a significant portion did.

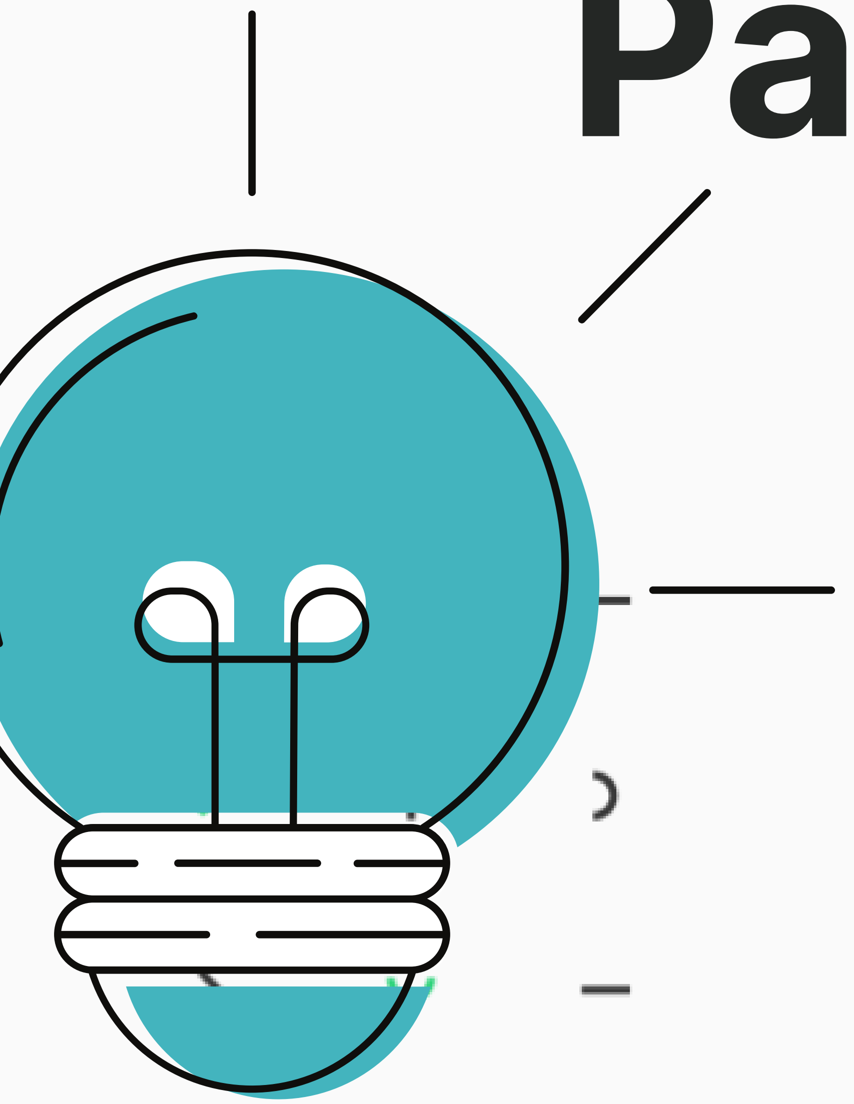

Distribution of Categorical Variables by Target Variable
Distribution of job by Target Variable
Distribution of marital by Target Variable
Distribution of education by Target Variable
Distribution of default by Target Variable
Distribution of housing by Target Variable
Distribution of loan by Target Variable
subscribed
no
yes
count
job
technician
retired
blue-collar
self-employed
services
management
admin.
unemployed
student
entrepreneur
housemaid
marital
single
divorced
married
education
tertiary
secondary
primary
default
housing
loan


Distribution of contact by Target Variable
subscribed
no
yes
count
0
200
400
600
800
cellular
telephone
contact
Distribution of month by Target Variable
subscribed
no
yes
count
0
100
200
300
400
apr
dec
feb
jan
mar
may
nov
oct
aug
jul
jun
sep
month
Distribution of poutcome by Target Variable
subscribed
no
yes
count
0
200
400
600
800
failure
other
success
poutcome
Distribution of subscribed by Target Variable
count
0
200
400
600
800
1000
no
yes
subscribed

Distribution of Categorical Variables by Subscription Status

- **Job:** Management & technician roles have higher subscriptions, blue-collar lower.
- **Marital:** Single clients subscribe slightly more than married ones.
- **Education:** Higher education levels show better subscription rates.
- **Default:** Clients with credit defaults rarely subscribe.
- **Housing & Loan:** No housing loan → higher subscriptions; personal loan → lower.
- **Contact:** Cellular contacts perform better than telephone.
- **Month:** Higher subscriptions in May & March; lower in December & April.
- **Previous Outcome:** Prior campaign success boosts subscription chances.
- **Overall:** Most didn't subscribe, but education, loans & past success matter.

Part 2: Data Preparation



Dropping Irrelevant Columns

Why Were These Columns Removed?

1️⃣Target Variable (subscribed) – Clustering is unsupervised; keeping it would turn it into classification.

2️⃣Day & Month (day, month) – Exact contact dates don't help in meaningful segmentation.

3️⃣ Call Duration (duration) – Strongly linked to subscription, making it unsuitable for clustering.

4️⃣ Previous Outcome (poutcome) – Too many missing values, reducing reliability.

5️⃣ Days Since Last Contact (pdays) – Many -1 values distort clustering patterns.

📌 Removing these ensures the model captures real customer behavior!

Encoding Ordinal Categorical Features

Why Encode Ordinal Attributes?

✓ ML models need numerical input.
✓ Some categories have a meaningful order (e.g., education level).
✓ Encoding preserves ranking relationships.

Ordinal Attributes & Encoding:

- Education: **"primary" < "secondary" < "tertiary"** → Encoded as 1, 2, 3
- Default, Housing, Loan: Binary → **"yes" = 1, "no" = 0**

Why Not Encode Other Features as Ordinal?

- ❌ **Marital Status, Job, Contact** → No natural ranking → One-Hot Encoding preferred.

One-Hot Encoding for Categorical Columns

Why One-Hot Encoding?

✓ K-Means requires numerical data.
✓ Prevents false ordinal relationships.
✓ Converts categories into binary indicators.

How We Encoded the Data

- **pd.get_dummies()** → Converts categories to binary variables.
- **dummy_na=True** → Handles missing values as a category.
- **drop_first=True** → Prevents the dummy variable trap.

✅ All categorical columns to ensure proper clustering.

Handling Missing Values & Outliers

Handling Missing Values in Numerical Columns

Why Handle Missing Values?

✓ Prevents bias and data loss.

✓ Improves clustering accuracy.

Imputation Strategy:

- **Age** (12 missing values)→ Mean Imputation

  ✓ Normally distributed → Mean is a good representative.

  ✓ Preserves data structure without skewing results.

- **Education** (104 missing values) → Mode Imputation

  ✓ Categorical → Mode (most frequent value) is best.

  ✓ Avoids introducing artificial values.

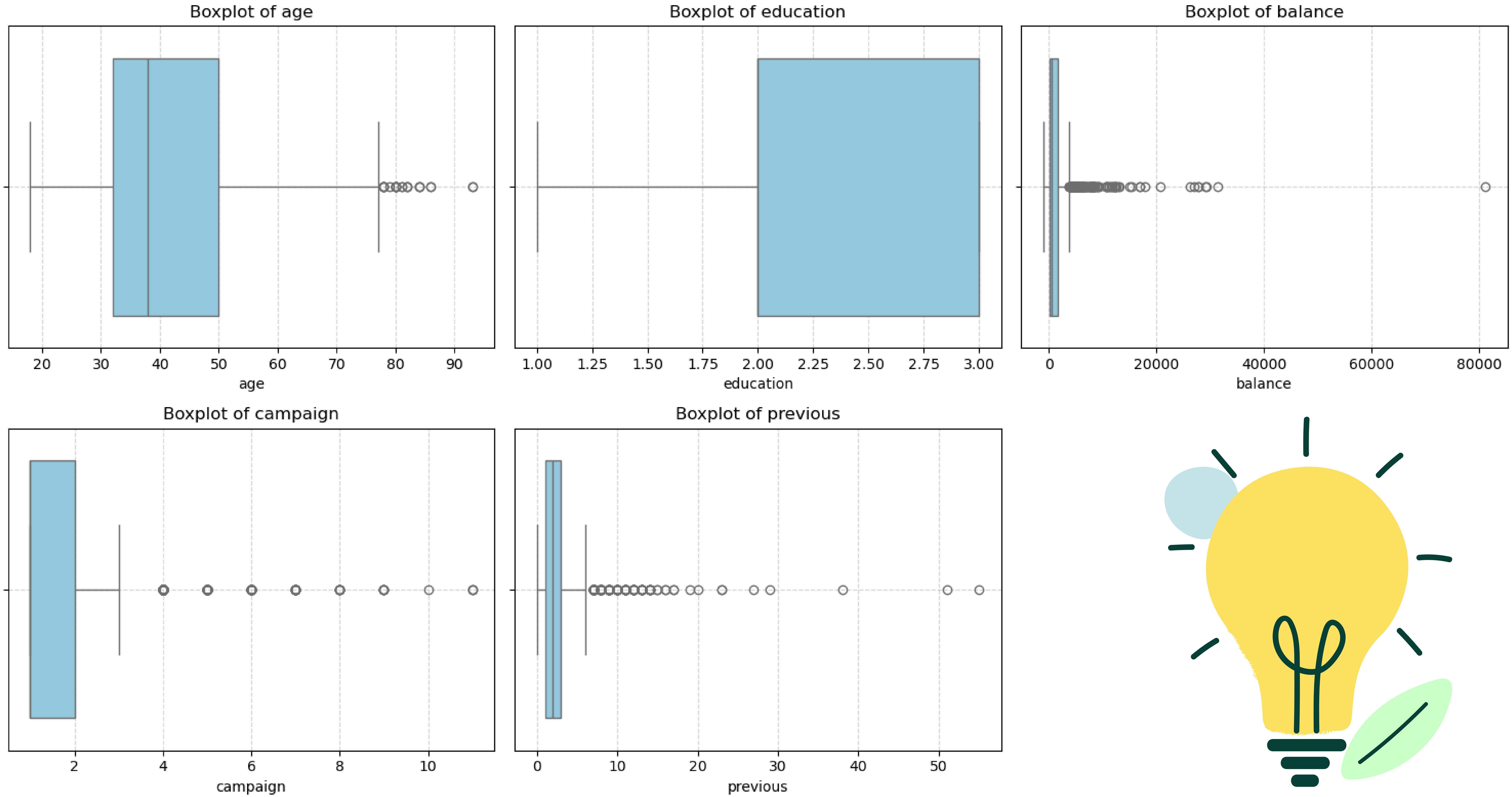
Handling Outliers in Numerical Features

Why Detect and Handle Outliers?

✔ Outliers can distort clustering results.
✔ Extreme values affect distance-based algorithms like K-Means.
✔ Proper handling improves model robustness.

Detected Outliers:

- Age → 🟠 24 outliers
- Education → ✅ 0 outliers (No extreme values)
- Balance → 🔴 158 outliers
- Campaign → 🔴 212 outliers
- Previous → 🔴 169 outliers


Boxplot of age
20
30
40
50
60
70
80
90
age
Boxplot of education
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
education
Boxplot of balance
0
20000
40000
60000
80000
balance
Boxplot of campaign
2
4
6
8
10
campaign
Boxplot of previous
0
10
20
30
40
50
previous

Logarithmic Transformation for Skewed Data

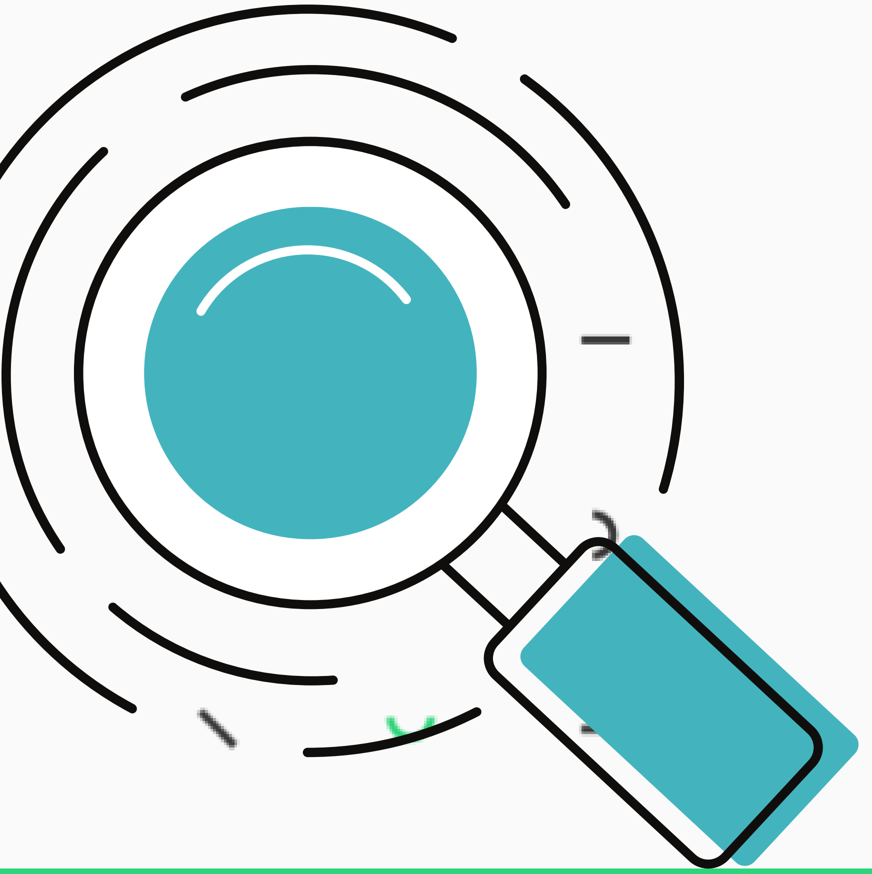
Why Consider Skewness?

✓ Skewness measures asymmetry in data distribution.
✓ Highly skewed features can distort clustering results.
✓ Log transformation reduces the impact of extreme values.

When to Apply Log Transformation?

- **Skewness > 1** → Right-skewed (Apply log(x + 1))
- **Skewness < -1** → Left-skewed (Consider Box-Cox or sqrt)
- **Skewness ≈ 0** → No transformation needed

Feature Scaling & Normalization



Feature Scaling & Normalization

Why Scale Features?

✓ K-Means clustering is sensitive to feature magnitudes.
✓ Scaling ensures equal contribution of all numerical features.
✓ Helps improve convergence and cluster stability.

Types of Scaling Used:

1️⃣ **StandardScaler:** Scales to mean = 0, variance = 1 (for normal distributions).

2️⃣ **MinMaxScaler:** Scales between 0 and 1 (for bounded data).

3️⃣ **RobustScaler:** Uses median and IQR, robust to outliers.

RobustScaler: Why Use It?

RobustScaler is a data scaling technique that is resistant to outliers. Unlike StandardScaler, it uses median and interquartile range (IQR) instead of mean and standard deviation.

When to Use?

✓ When the dataset contains outliers.

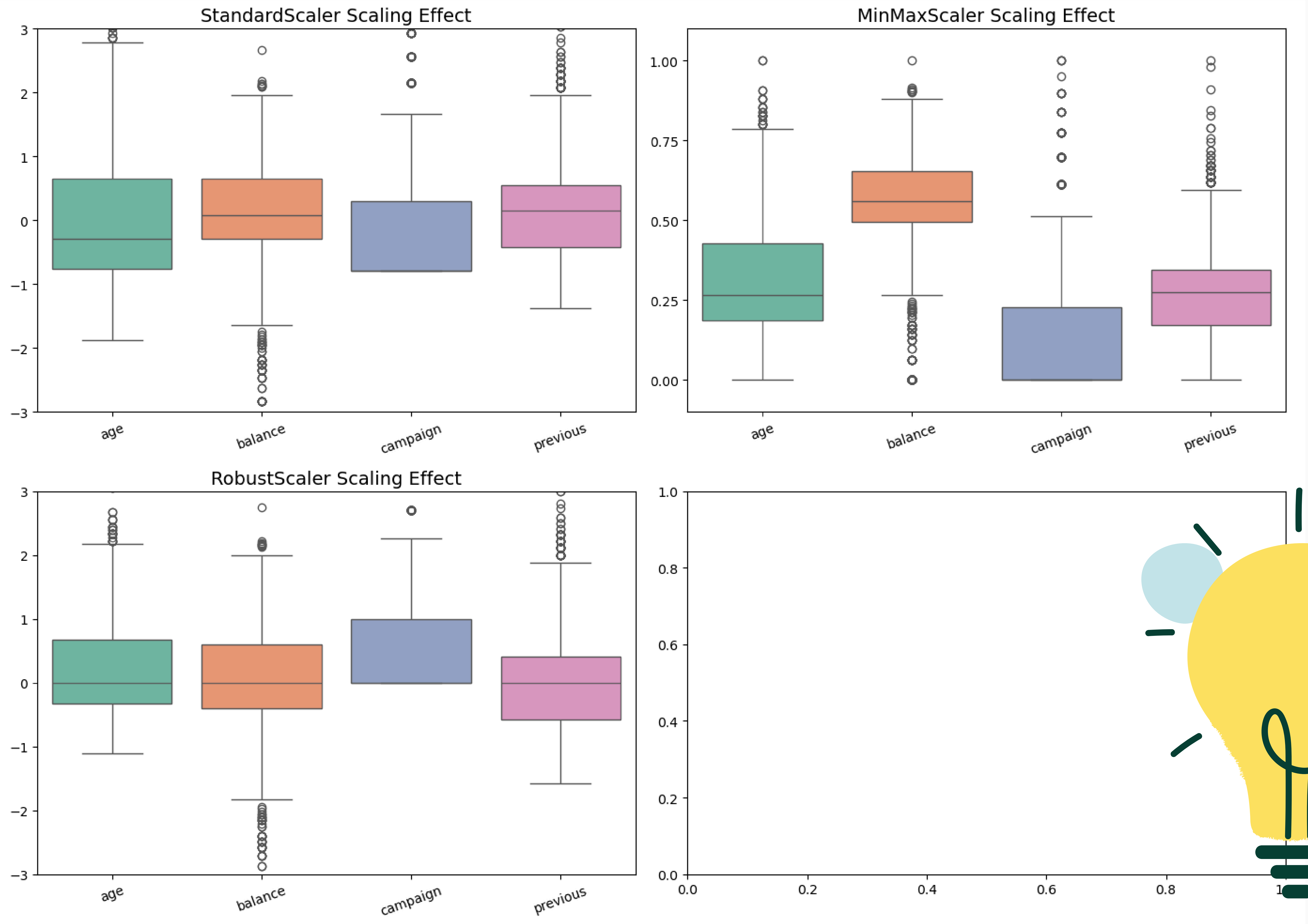
✓When other scalers (like StandardScaler) get distorted by extreme values.

Advantages

✓ More stable scaling for skewed distributions.

✓ Not affected by large outliers.

🔹 Ensures a robust feature transformation, making clustering results more reliable!


StandardScaler Scaling Effect
MinMaxScaler Scaling Effect
RobustScaler Scaling Effect
age
balance
campaign
previous
3
2
1
0
−1
−2
−3
1.00
0.75
0.50
0.25
0.00
1.0
0.8
0.6
0.4
0.2
0.0

Part 3: Clustering & Dimensionality Reduction

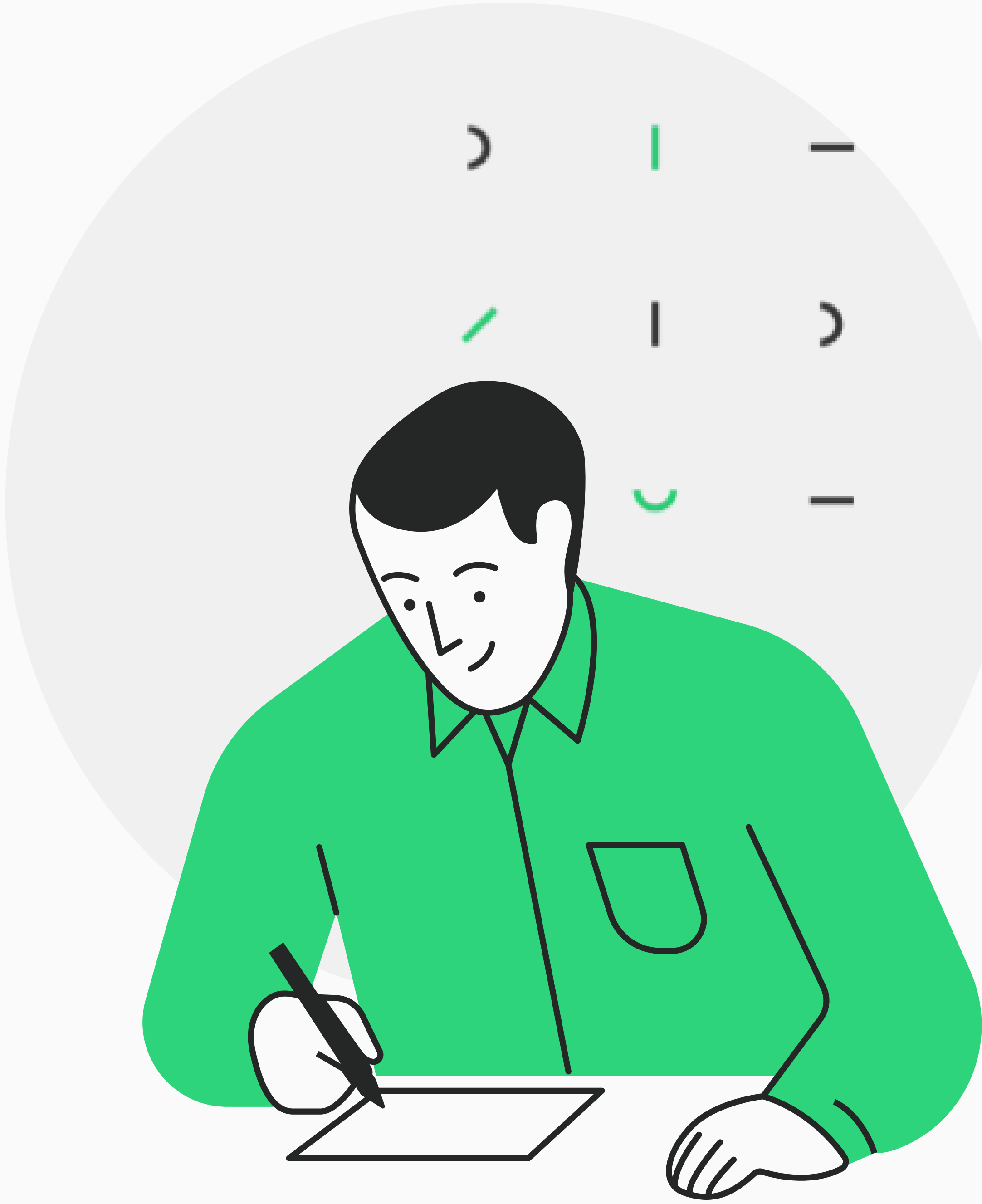
Introduction to K-Means Clustering

WHY K-MEANS?

✓ Simple & Efficient: Fast and easy to implement.
✓ Scalability: Works well with large datasets.
✓ Partitioning Method: Groups similar data points into clusters.

CHALLENGES IN K-MEANS

⚠ Choosing the Right K: Selecting the optimal number of clusters is not straightforward.
⚠ Sensitivity to Initialization: Different initial centroids can lead to different results.
⚠ Interpretability: Clusters may not always have a clear meaning.
⚠ Assumption of Spherical Clusters: Struggles with complex cluster shapes.



Choosing the Optimal Number of Clusters (K)

ELBOW METHOD

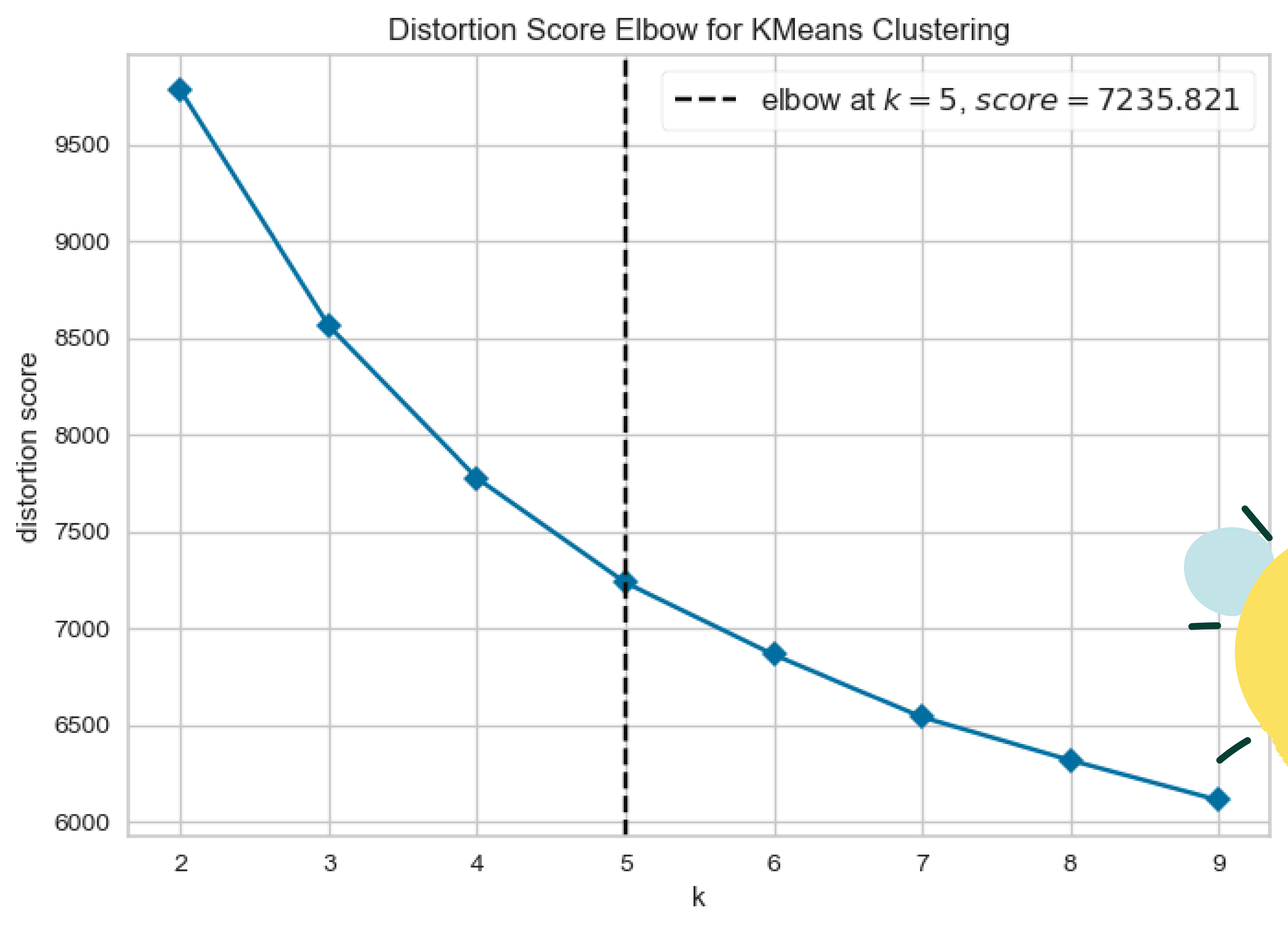
- Shows K = 5 as the elbow point, balancing variance and simplicity.

SILHOUETTE SCORE

- Measures cluster separation (higher = better).
- Best score at **K = 3** → Strongest cluster separation.

  - 🔹 K = 2 → 0.2731 (Too simple)
  - 🔹 K = 3 → 0.2732 ✅ (Best choice)
  - 🔹 K = 4 → 0.2641 (Weaker separation)
  - 🔹 K = 5 → 0.2289 (Poor separation)

✅ Final Choice: **K = 3**


Distortion Score Elbow for KMeans Clustering
elbow at $k = 5$, $score = 7235.821$
distortion score
9500
9000
8500
8000
7500
7000
6500
6000
2
3
4
5
6
7
8
9
k

K-Means Clustering (K=3) & Visualization

CLUSTER INSIGHTS

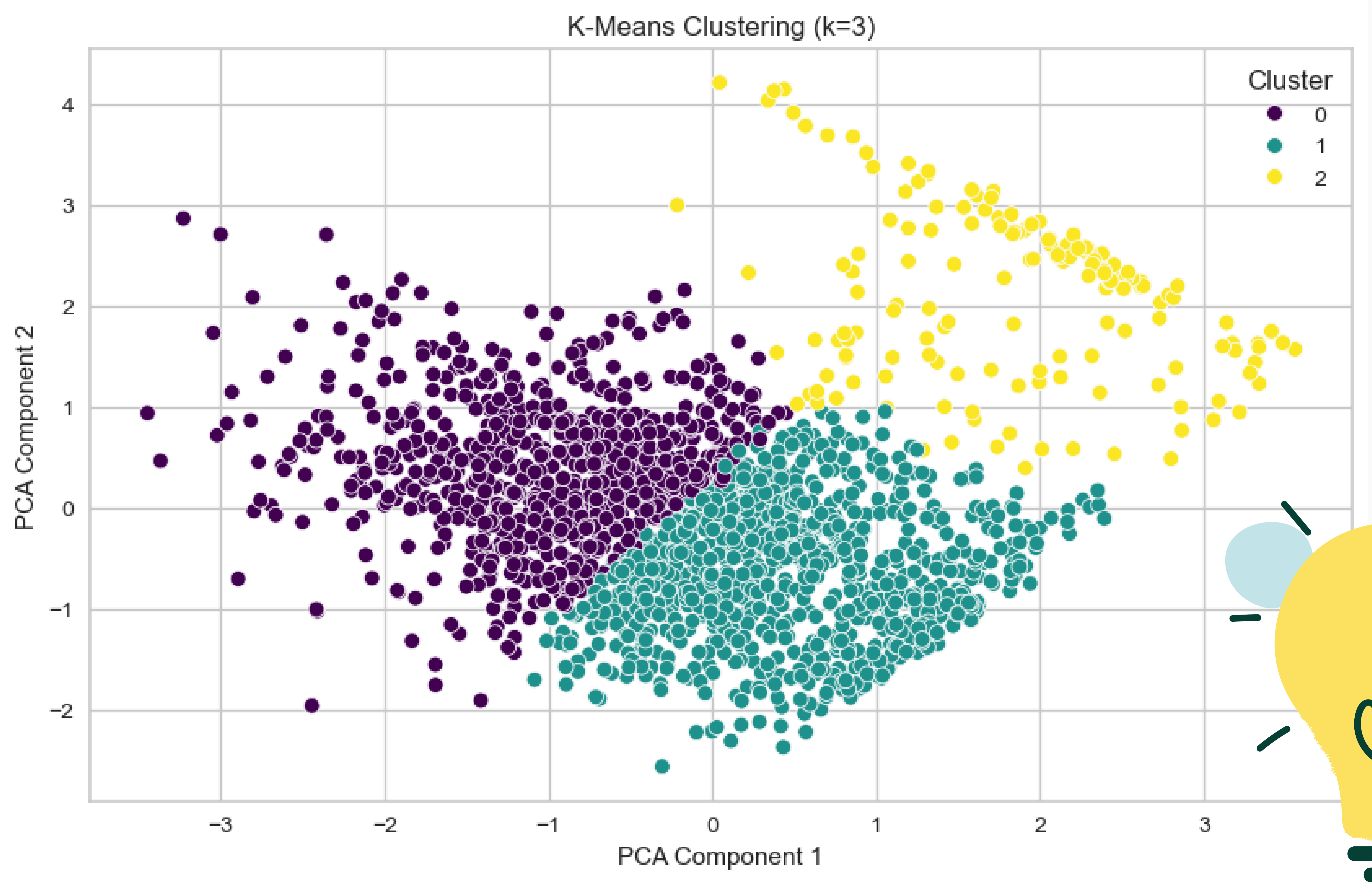
- Cluster 0 (🔵): Younger individuals with low balance.
- Cluster 1 (🟠): Moderate balance & campaign frequency.
- Cluster 2 (🟢): Older individuals with higher balance.

PCA VISUALIZATION

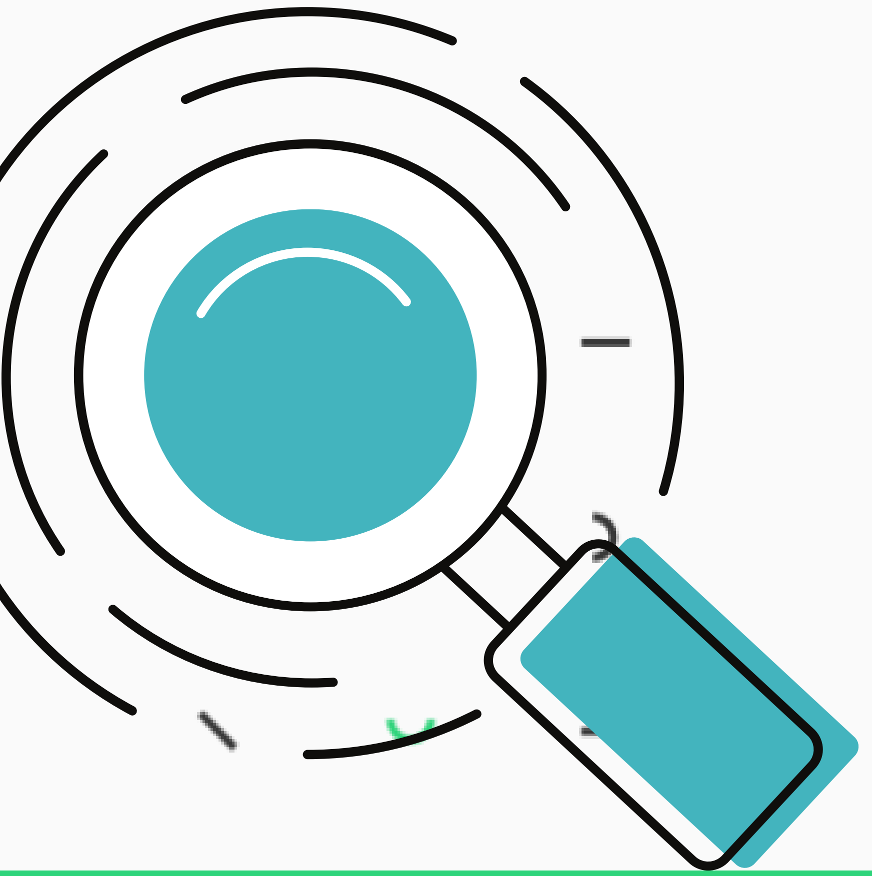
- X-Axis (PCA Component 1) → Primary variance direction.
- Y-Axis (PCA Component 2) → Secondary variance.
- Cluster 2 is more distinct, while Clusters 0 & 1 overlap.

Key Takeaways

✅ Cluster 2 stands out with high balance.
✅ Clusters 0 & 1 share similarities, needing further refinement.
✅ Outliers exist, mainly in Cluster 2.


K-Means Clustering (k=3)
Cluster
0
1
2
PCA Component 2
PCA Component 1
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

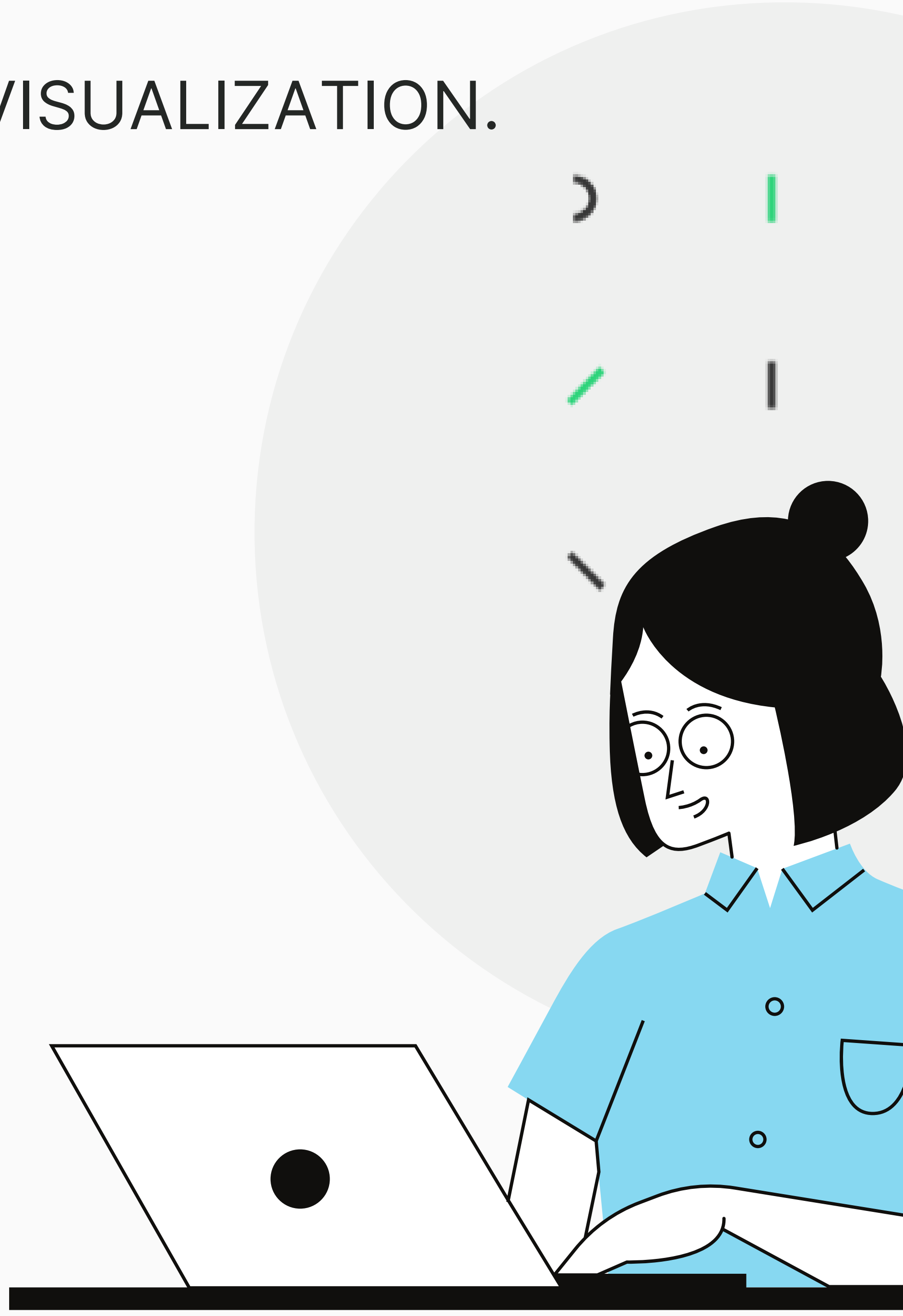
Dimensionality Reduction

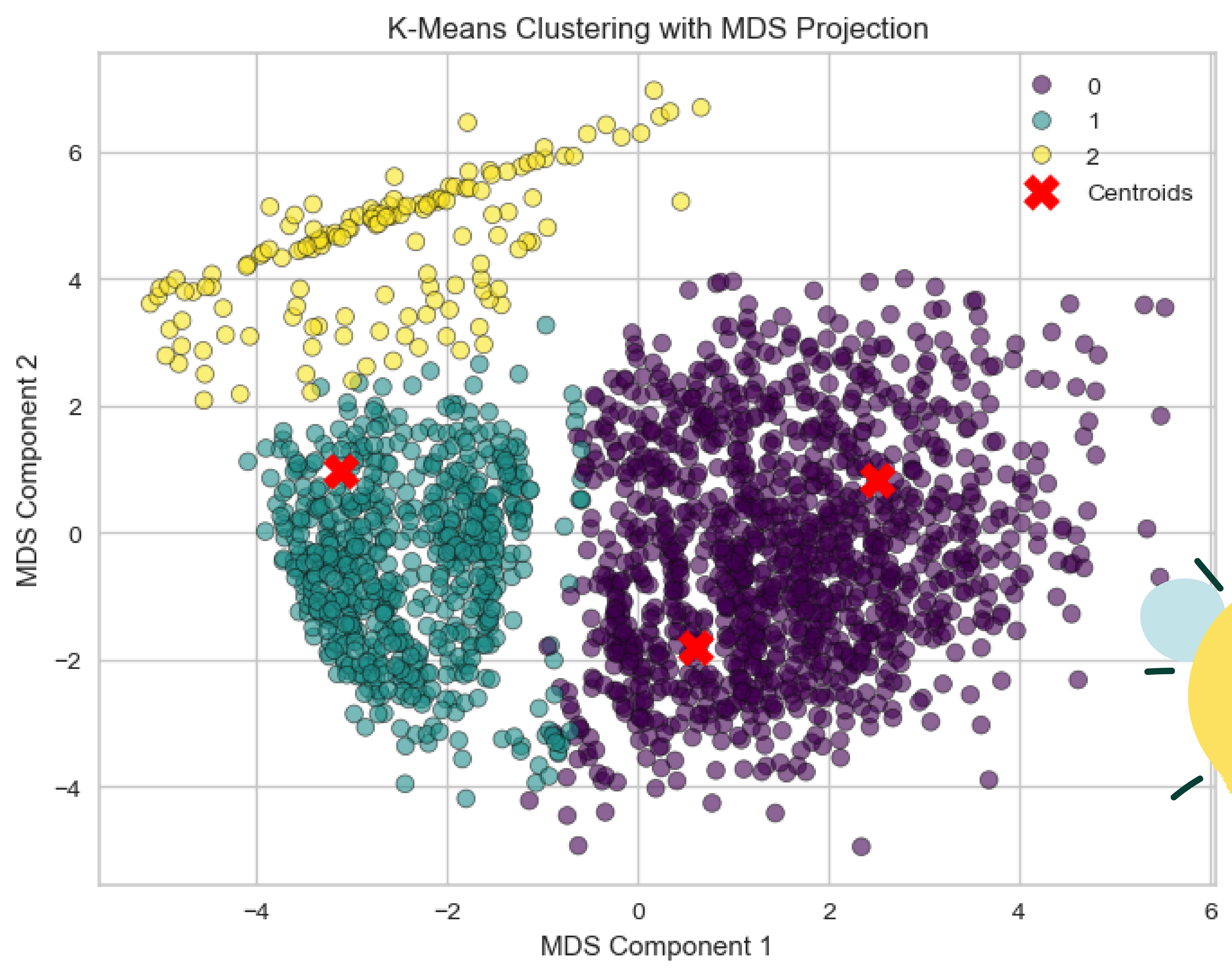


K-Means Clustering with MDS Projection

- **MDS (MULTI-DIMENSIONAL SCALING)** REDUCES DATA TO 2D FOR VISUALIZATION.
- **THREE CLUSTERS (0, 1, 2) ARE IDENTIFIED:**
  - Cluster 0 (purple): Largest & compact.
  - Cluster 1 (yellow): More spread-out, possible outliers.
  - Cluster 2 (green): Well-defined, slight overlap with Cluster 0.

- **Red "X"** marks centroids, showing cluster centers.
- Some overlap exists, suggesting potential improvements with alternative clustering methods (e.g., DBSCAN) or better feature selection

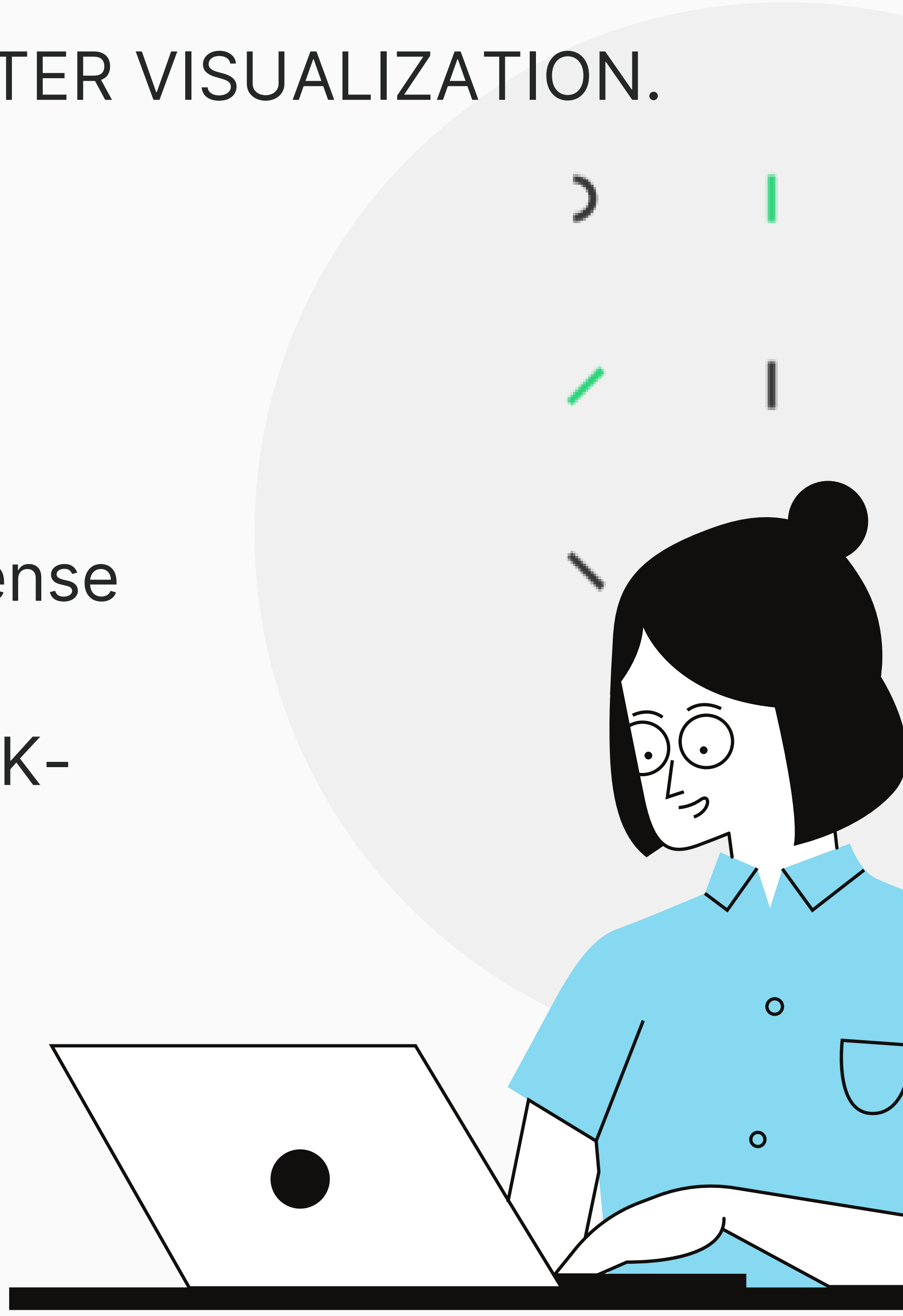


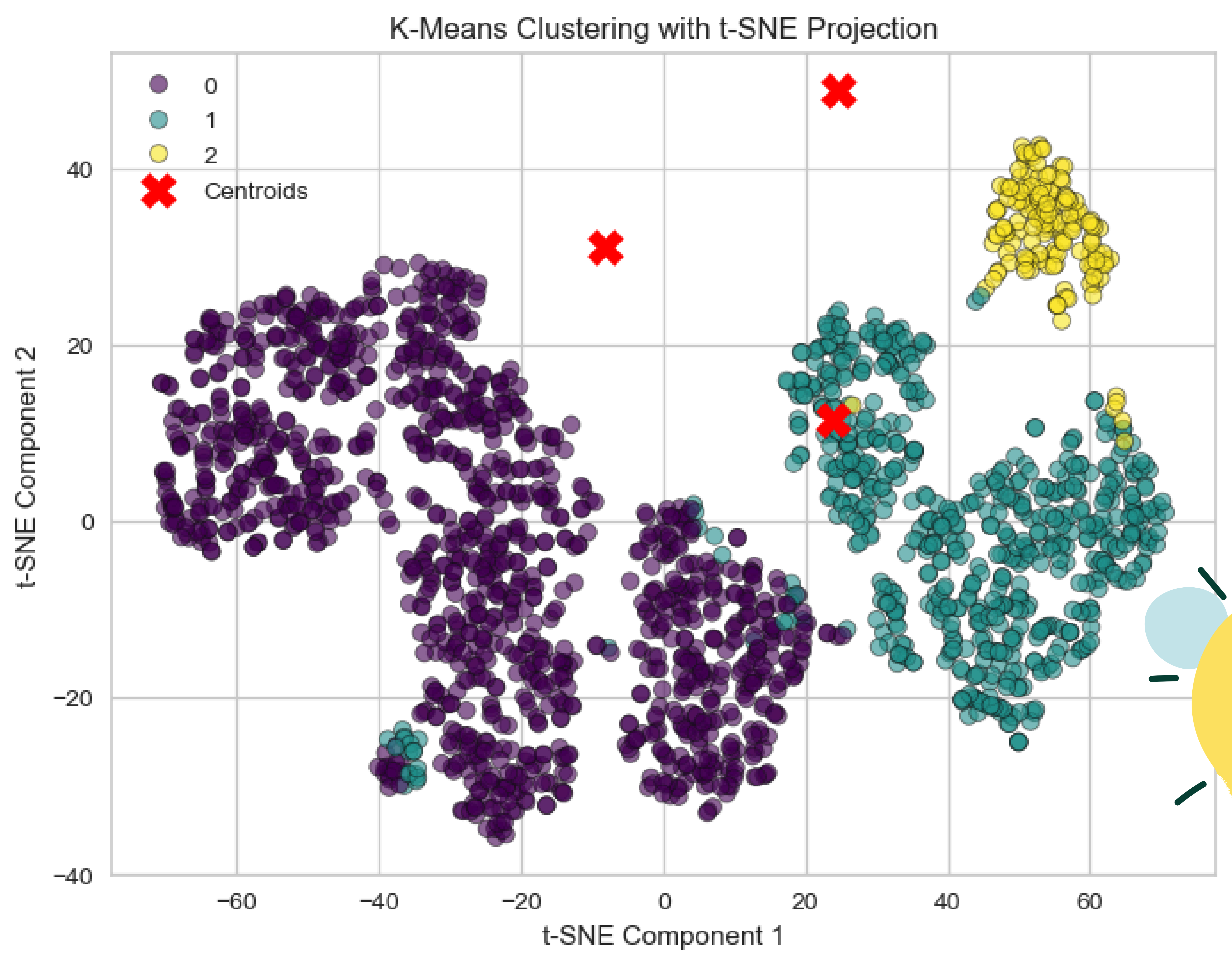

K-Means Clustering with MDS Projection
MDS Component 1
MDS Component 2
0
1
2
Centroids
−4
−2
0
2
4
6

t-SNE Clustering Visualization

- **T-SNE** PROJECTS **K-MEANS CLUSTERS** INTO A **2D SPACE** FOR BETTER VISUALIZATION.
- **CLUSTERS (0, 1, 2) ARE WELL-SEPARATED, WITH:**
  - Cluster 0 (purple): Widely spread.
  - Cluster 2 (yellow): Compact & distinct.

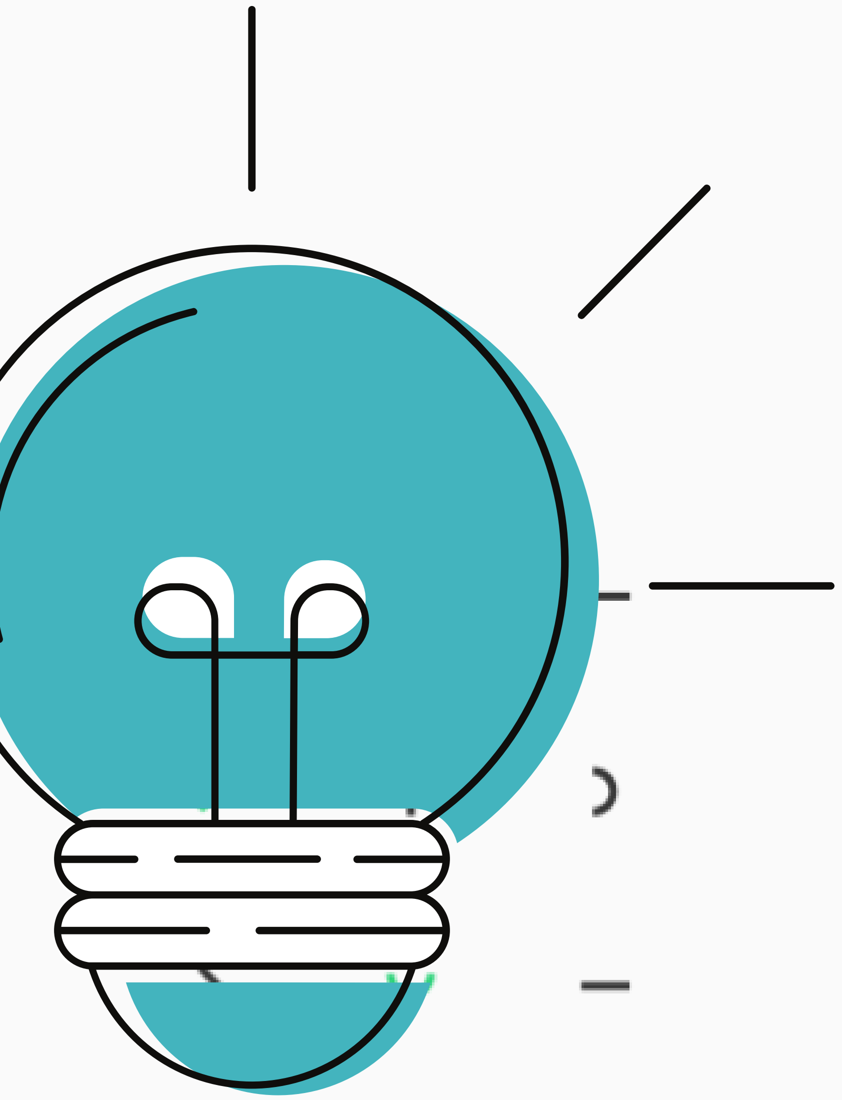
🔹 Red "X" centroids show cluster centers but seem far from dense areas, indicating possible clustering refinement.
🔹 Conclusion: t-SNE reveals non-linear structures, suggesting K-Means may not fully capture true cluster shapes.





K-Means Clustering with t-SNE Projection
0
1
2
Centroids
40
20
0
−20
−40
t-SNE Component 2
−60
−40
−20
0
20
40
60
t-SNE Component 1

Part 4: Cluster Analysis & Insights



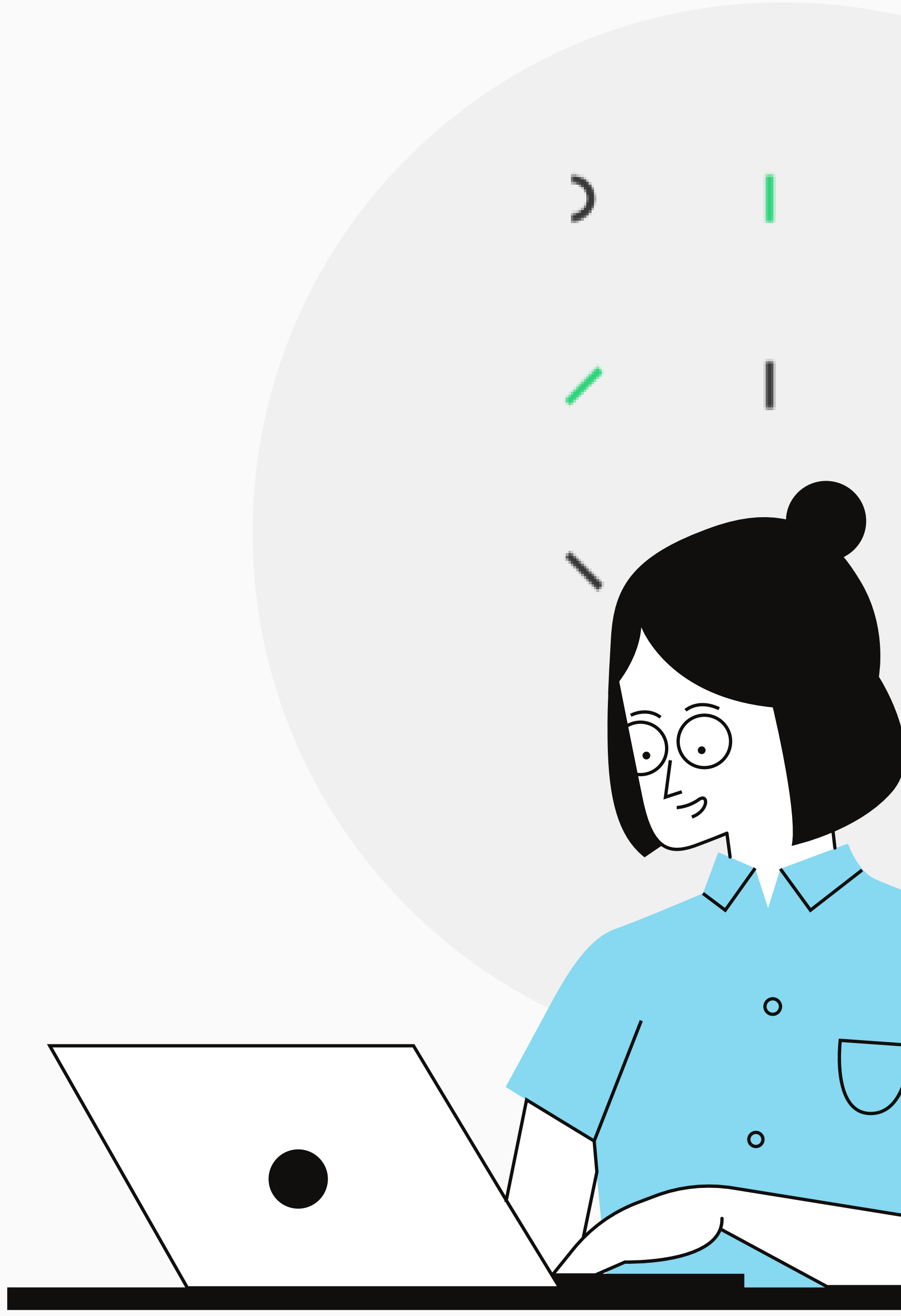
Cluster Analysis & Insights

CLUSTER OVERVIEW:

- Cluster 0: Moderate balance, fewer loans, stable job distribution.
- Cluster 1: More single individuals, higher unknown contact methods, slightly higher default rate.
- Cluster 2: Negative balance, higher loan dependency, distinct job patterns.

🔑 **KEY INSIGHT:**

Cluster 0 is financially stable, Cluster 1 has mixed financial status, and Cluster 2 struggles financially.

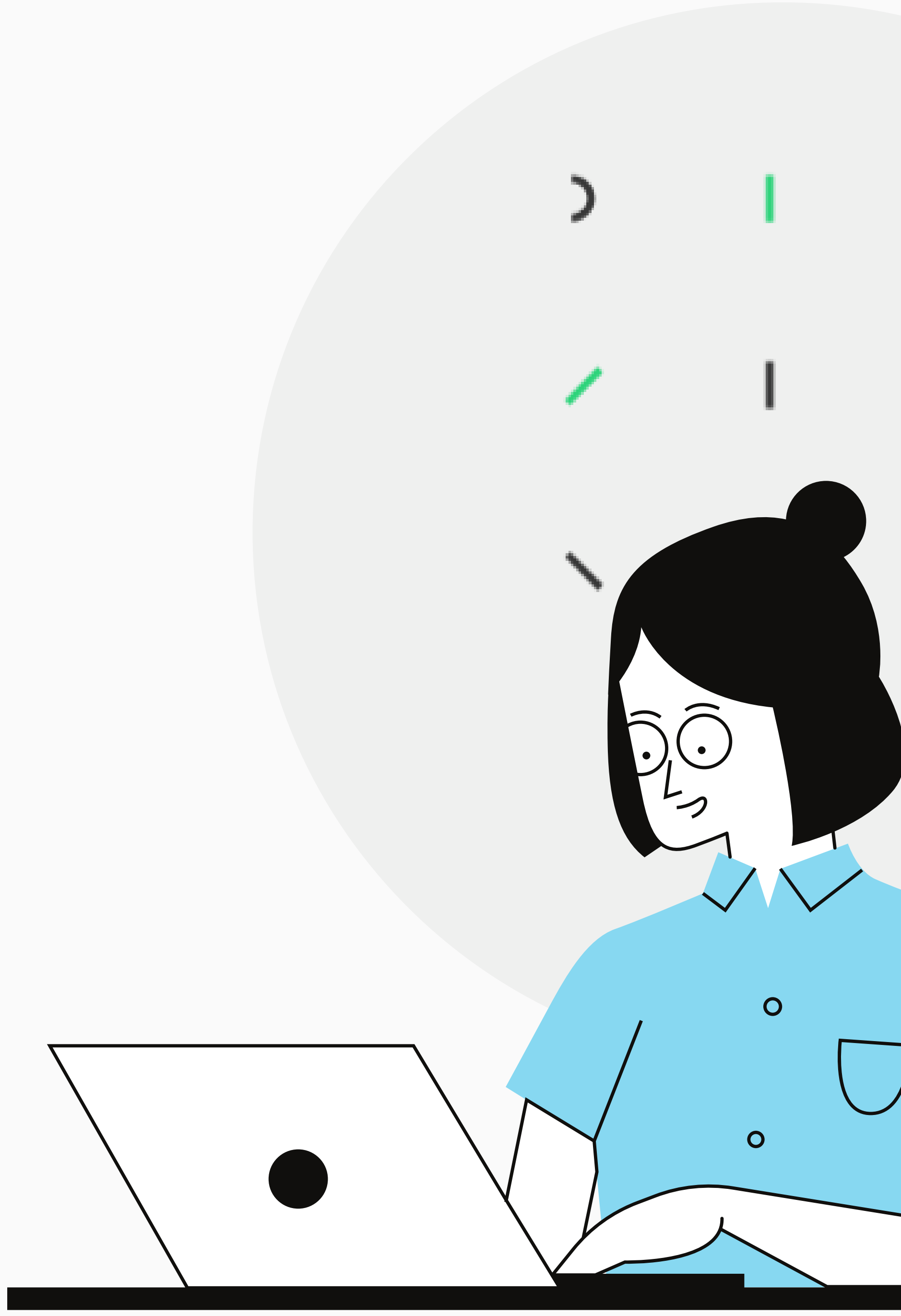


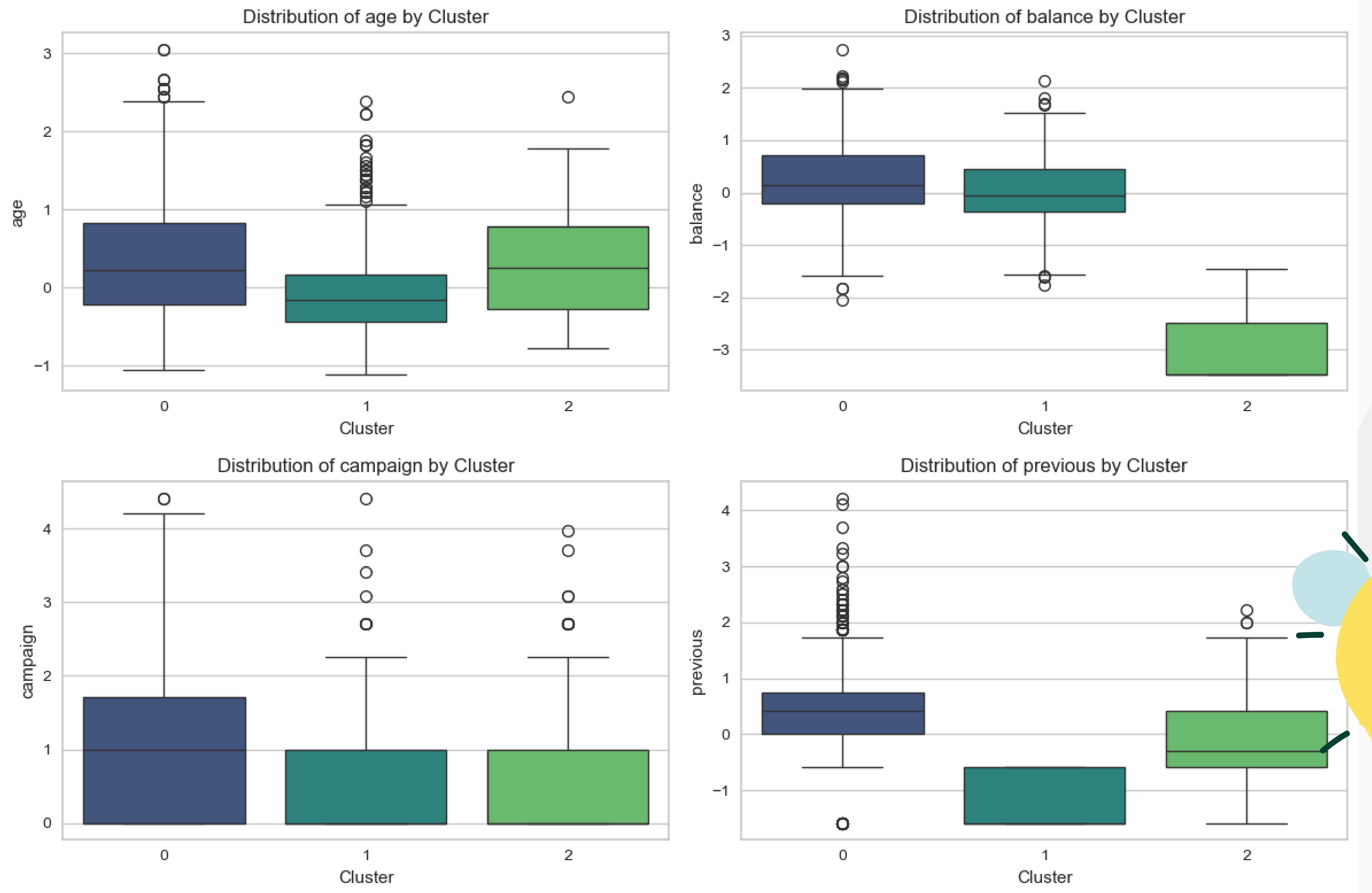
Cluster-wise Feature Distribution

- Age: Cluster 1 skews younger; Cluster 0 & 2 have broader distributions.
- Balance: Cluster 2 has significantly lower balances.
- Campaign: Cluster 0 gets more marketing attempts.
- Previous Contacts: Cluster 2 has more past interactions, Cluster 1 the least.

🔑 **KEY INSIGHT:**

Cluster 2 faces financial struggles but has higher past engagement.




Distribution of age by Cluster
age
Cluster
Distribution of balance by Cluster
balance
Cluster
Distribution of campaign by Cluster
campaign
Cluster
Distribution of previous by Cluster
previous
Cluster

Evaluating Clustering Performance

📏 SILHOUETTE SCORE = 0.2732

- Indicates moderate clustering, with some overlapping data points.
- Clusters are somewhat distinct but not well-separated.

🔍 Possible Improvements:

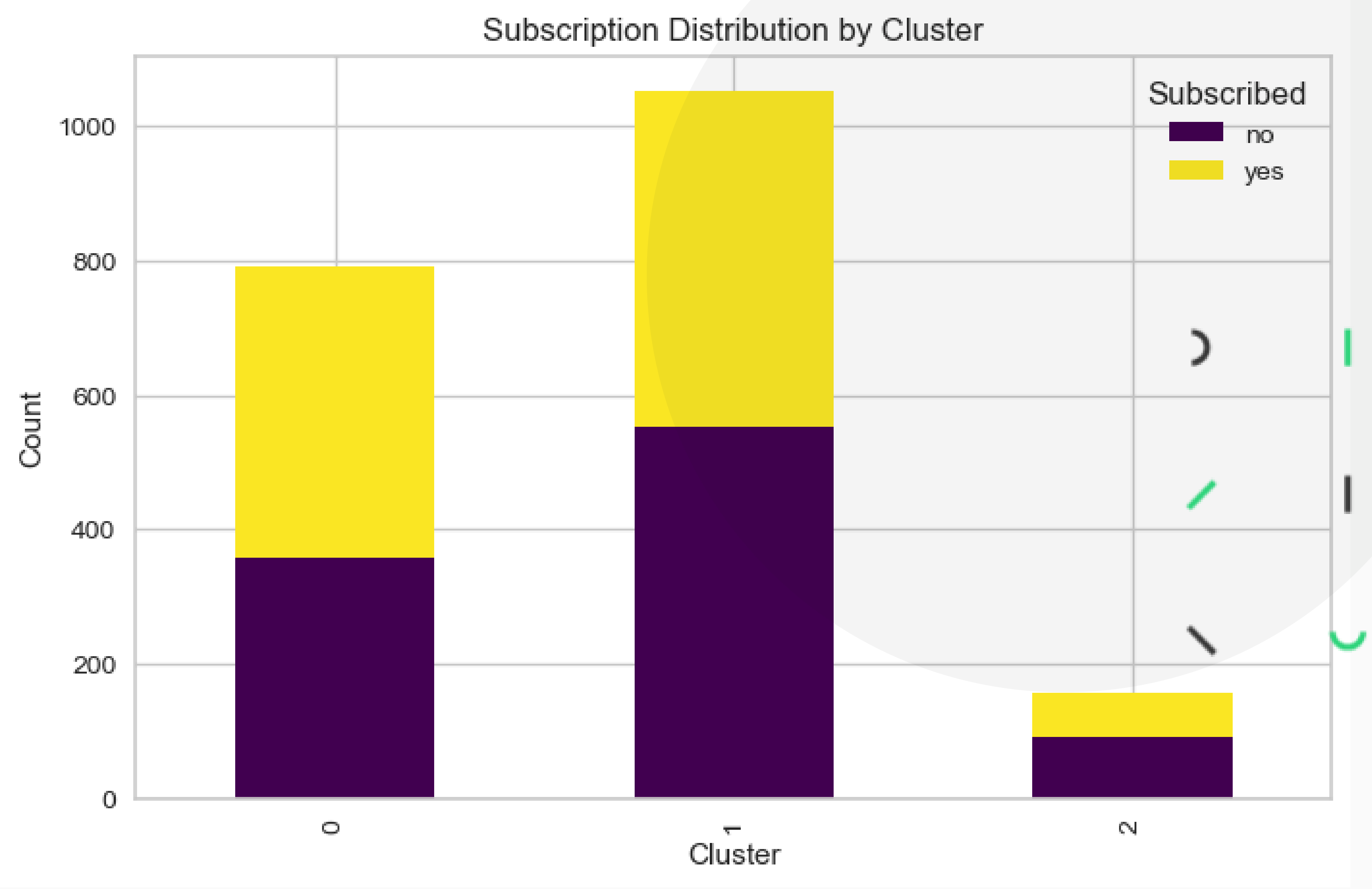
- Try different clustering methods (e.g., DBSCAN, Agglomerative).
- Adjust the number of clusters.
- Use feature selection or transformation (e.g., PCA).

Comparing Clusters with Target Variable

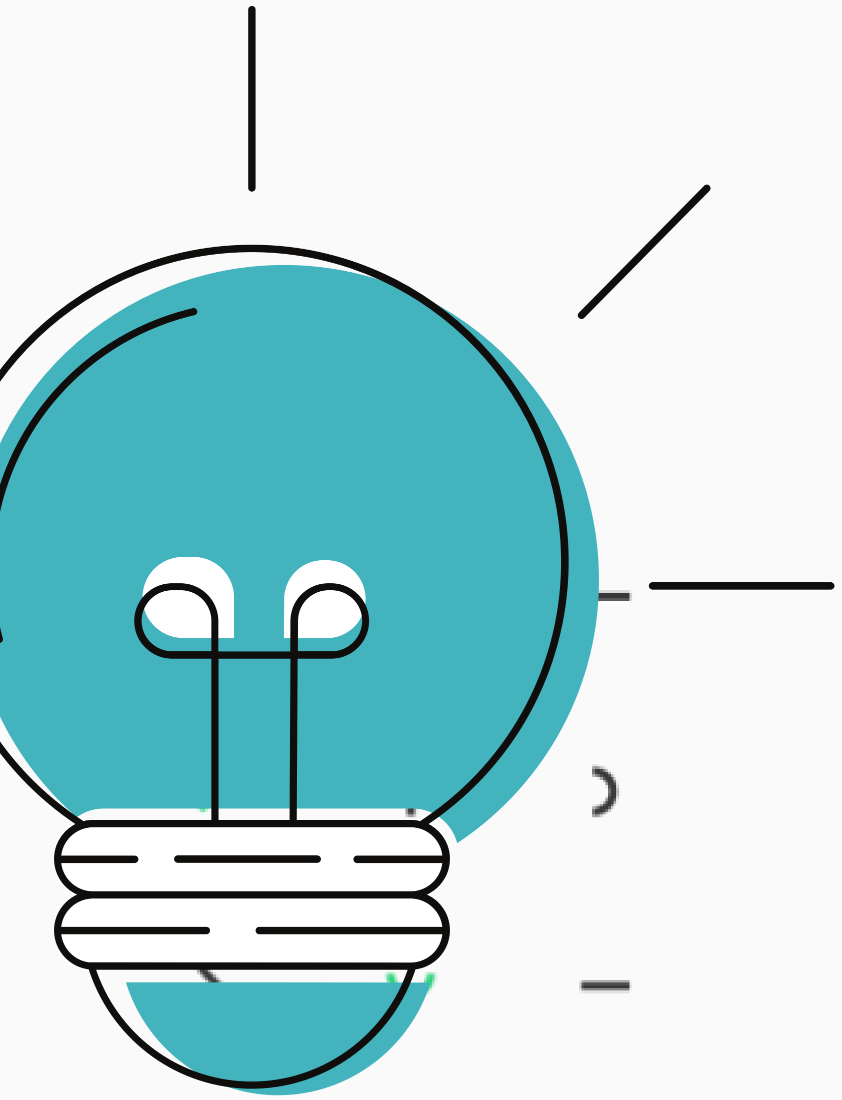
- 🔹 Cluster 0: Balanced between subscribers (435) and non-subscribers (357).
- 🔹 Cluster 1: Largest group, with more non-subscribers (552) but also a high number of subscribers (499).
- 🔹 Cluster 2: Smallest group, with fewer subscribers (66) than non-subscribers (91).

🔍 **KEY INSIGHT:**

Clusters show different likelihoods of subscription, useful for targeted marketing strategies.



Conclusion & Next Steps



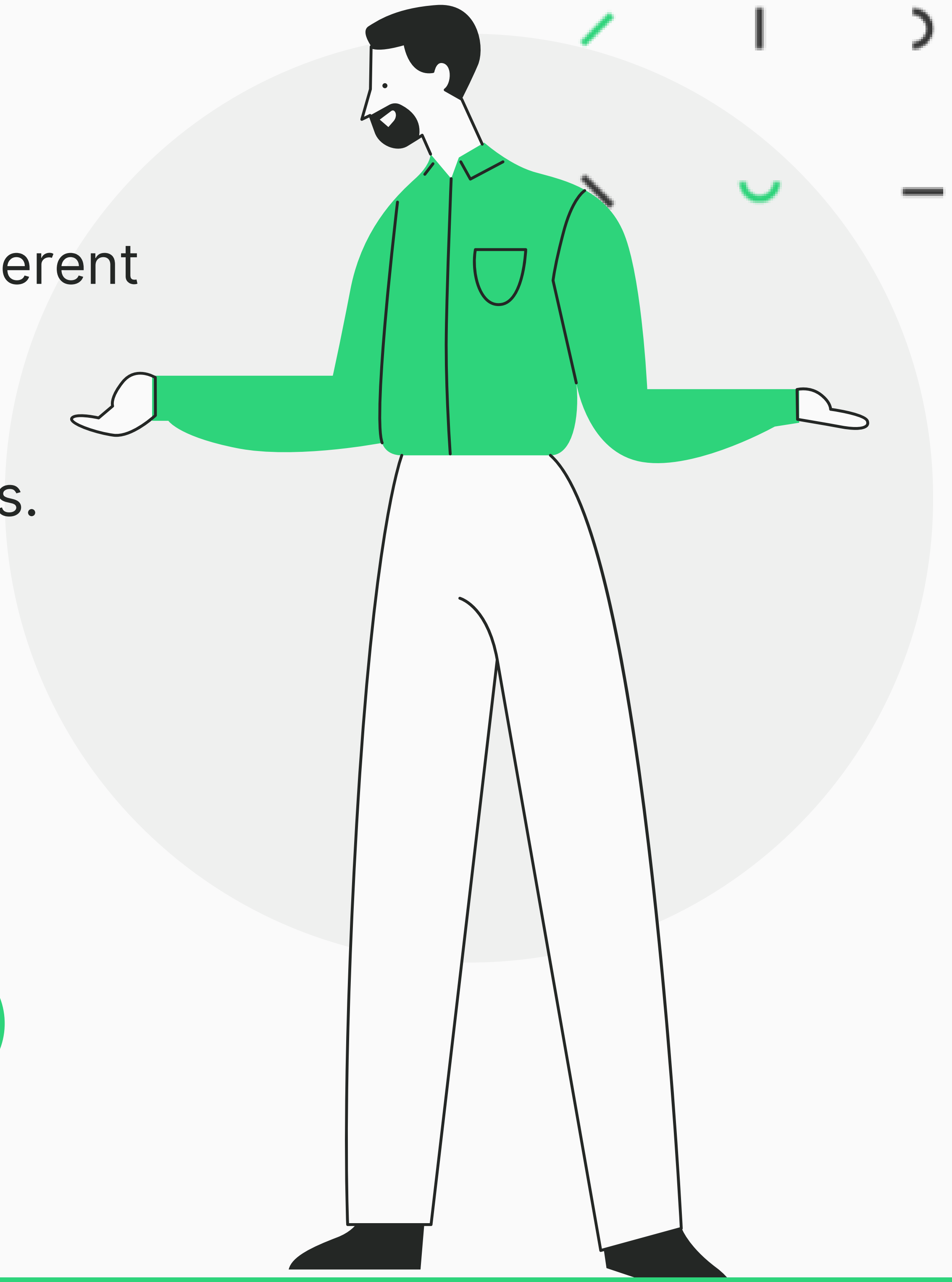
Conclusion & Key Insights

📌 SUMMARY OF FINDINGS

- Clustering identified three customer segments with different financial behaviors.
- **Cluster 0:** Financially stable, engaged in marketing.
- **Cluster 1:** Largest group, mix of subscribers & non-subscribers.
- **Cluster 2**: Financially weaker, least engaged.

📌 Business Implications

- Helps target marketing strategies more effectively.
- Optimizing outreach can improve subscription rates.
- Enables personalized financial recommendations.



Challenges & Future Work

📌 CHALLENGES

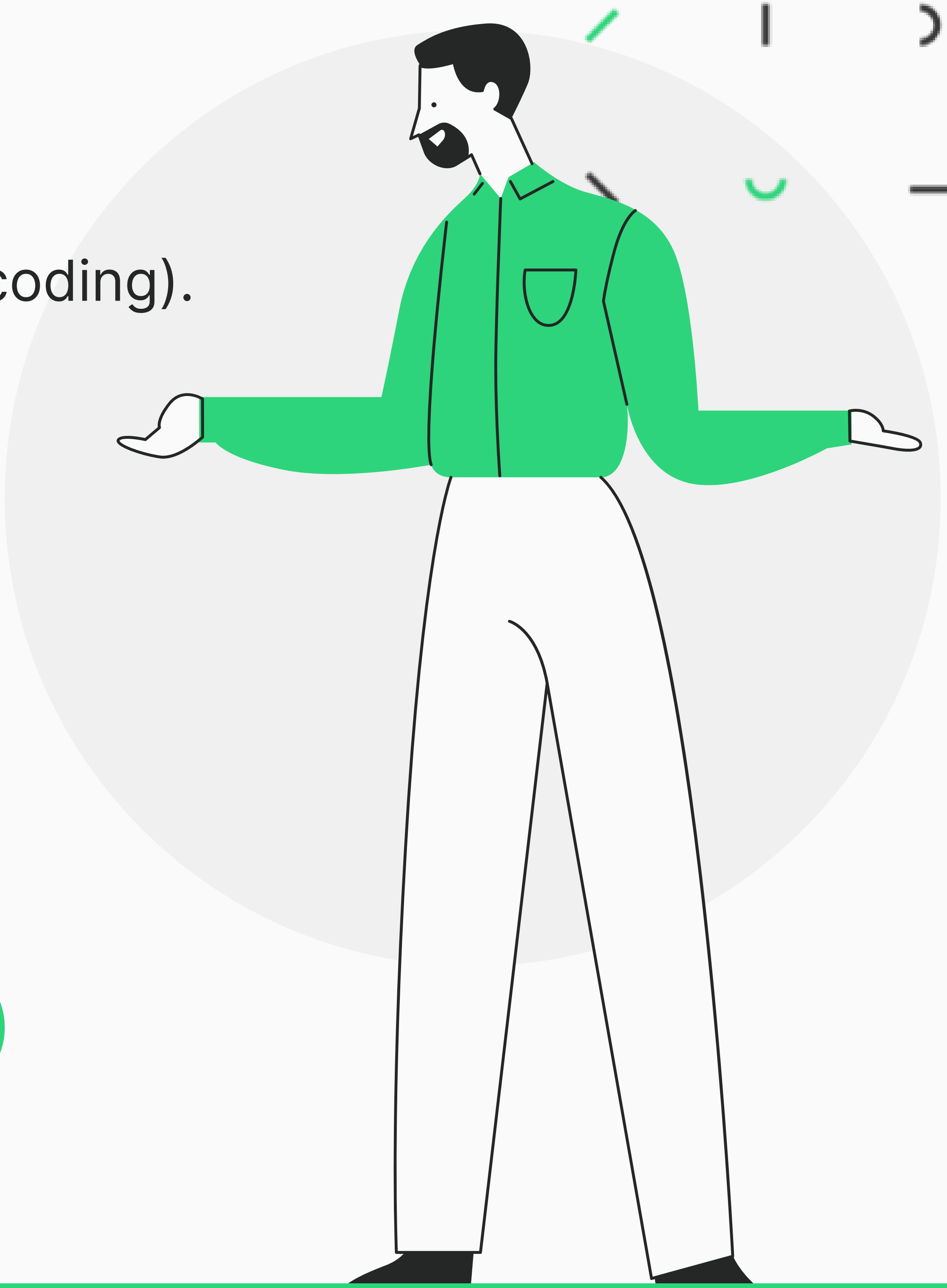
- Clusters have **some overlap** (Silhouette Score = 0.27).
- **Data preprocessing** issues (missing values, categorical encoding).
- **Scalability** concerns for larger datasets.

📌 Next Steps

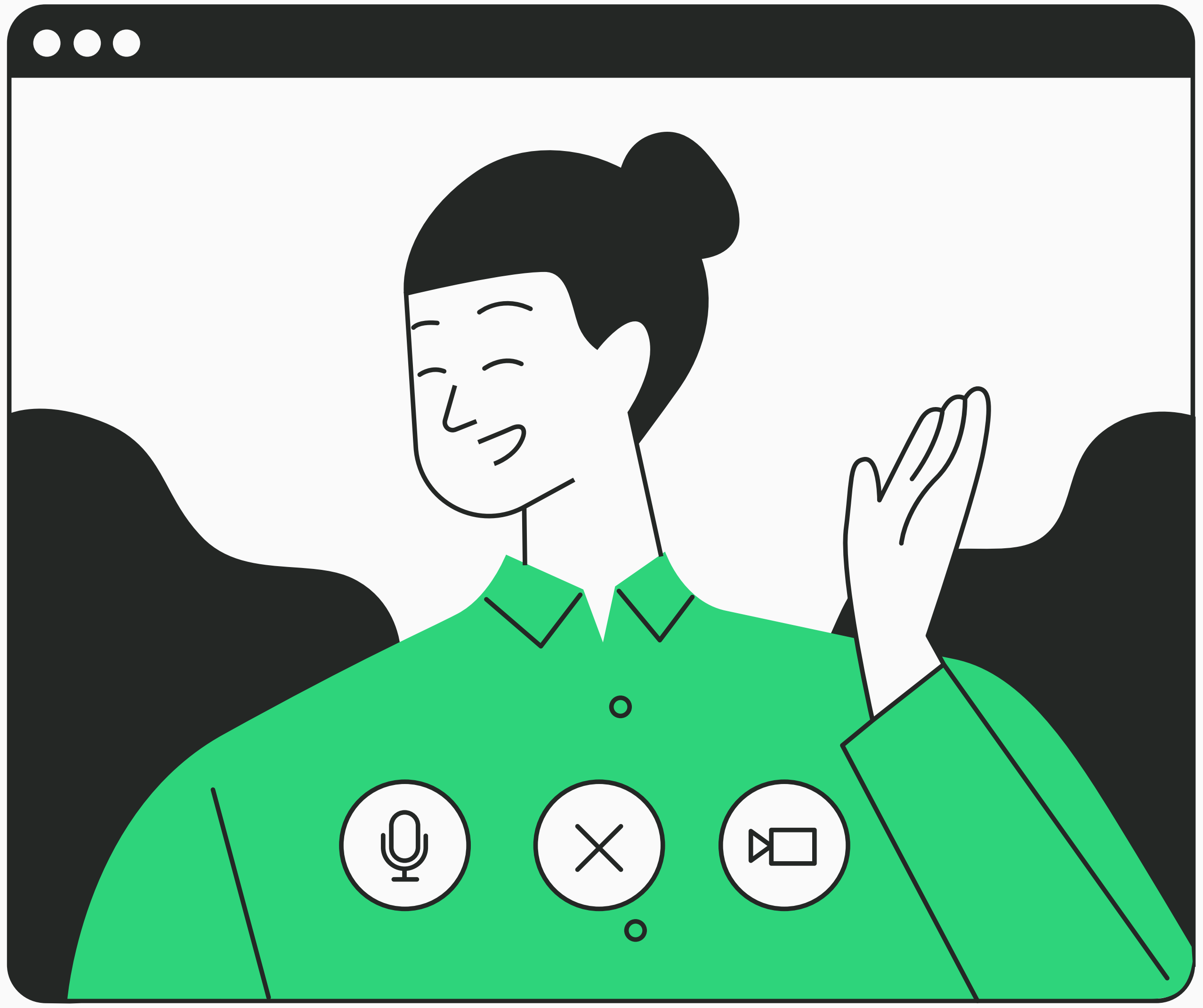
- Test other clustering methods (DBSCAN, Hierarchical).
- Use predictive modeling to improve conversion rates.
- Enhance feature engineering for better cluster separation.

🔍 Final Thought:

Data-driven segmentation improves marketing strategy and customer experience.



Contact me



EMAIL

mohamedelamrawi@yahoo.com

WEBSITE

https://linktr.ee/el_amraoui_mohamed

PHONE

+212712542184