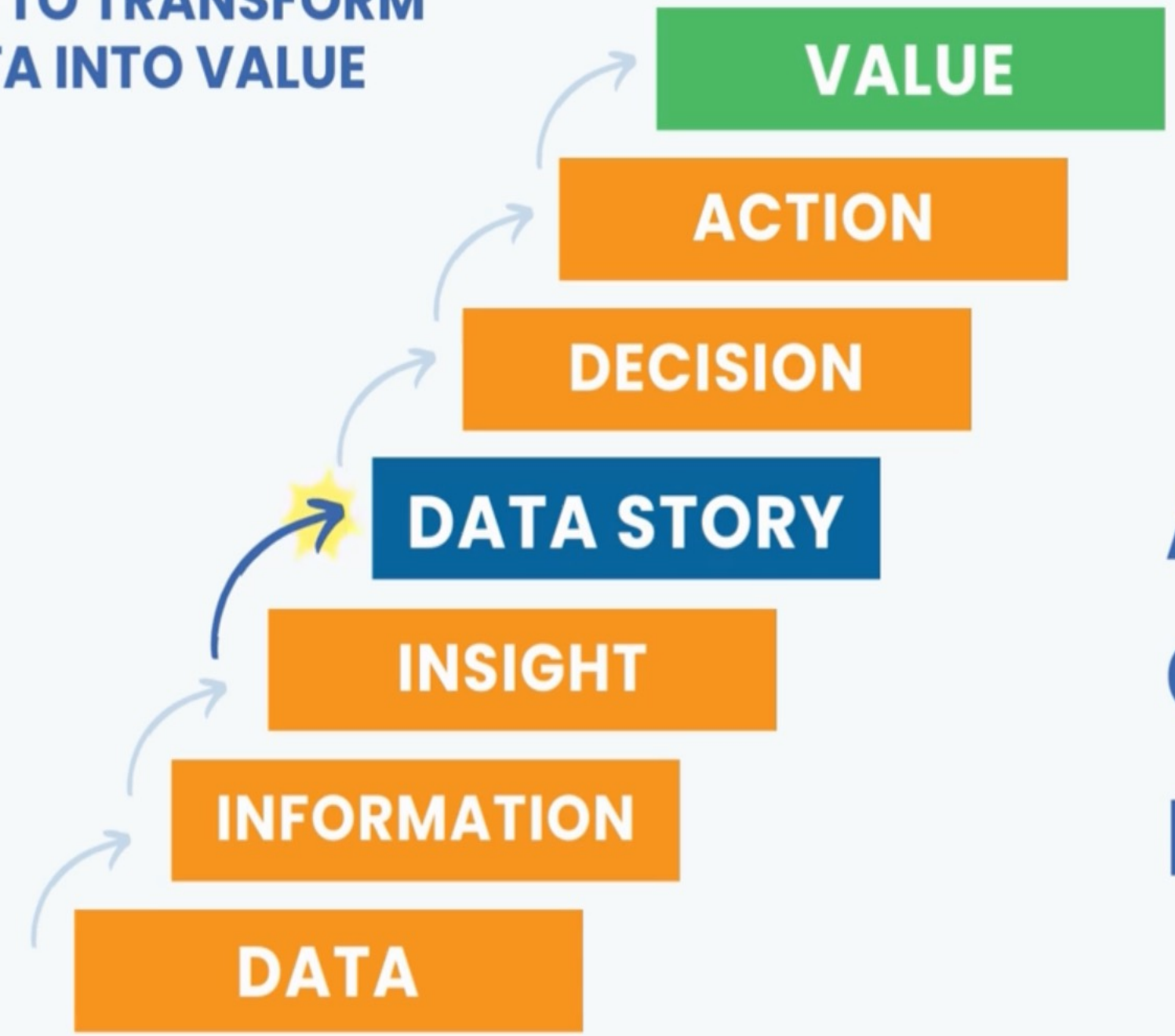
Quantitative Data Reasoning

"The ability to take data—to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a **hugely important skill in the next decades.**"

Hal Varian
Chief Economist
Google

STEPS TO TRANSFORM DATA INTO VALUE



Action.

Change.

Results.

Data-Driven Decision Making

1

Ask Questions to your Data !

2

Understand your Numbers !

3

Data Modelling
Mingle your Data!

4

Visualize your Data!

5

Data Storytelling Connect Dots!

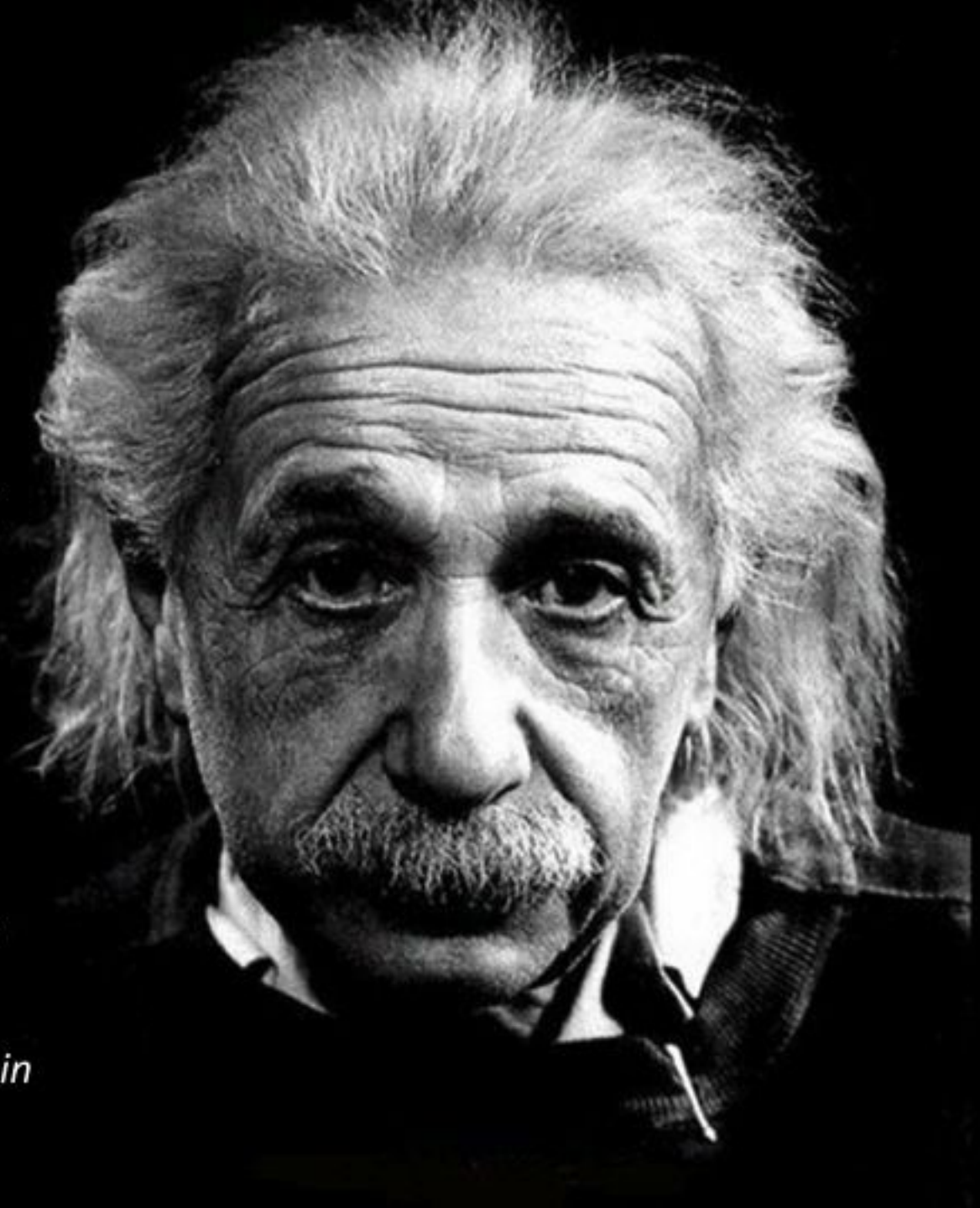
Do you prefer Chocolate or Vanilla?

1

Ask Questions to your Data !

"If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than 5 minutes."

- Albert Einstein



How to Ask AI Smart Questions for Better Insights

S . M . A . R . T

Specific

Measurable

Action-oriented

Relevant

Time-bound

Specific

Measurable

Action-oriented

Relevant

Time-bound

“Are kids getting enough exercise these days?”

Specific

Measurable

Action-oriented

Relevant

Time-bound

~~"Are kids getting enough exercise these days?"~~

"What percentage of kids achieve the recommended 60 minutes of physical activity at least five days a week?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

“Why did our recent video go viral?”

Specific

Measurable

Action-oriented

Relevant

Time-bound

~~"Why did our recent video go viral?"~~

"How many times was our video shared on social channels the first week it was posted?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

"How can we get customers to recycle our product packaging?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

~~"How can we get customers to recycle our product packaging?"~~

"What design features will make our packaging easier to recycle?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

"Why does it matter that Pine Barrens tree frogs started disappearing?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

~~"Why does it matter that Pine Barrens tree frogs started disappearing?"~~

"What environmental factors changed in Durham, North Carolina, between 1983 and 2004 that could cause Pine Barrens tree frogs to disappear from the Sandhills Regions?"

Specific

Measurable

Action-oriented

Relevant

Time-bound

"What environmental factors changed in Durham, North Carolina, between 1983 and 2004 that could cause Pine Barrens tree frogs to disappear from the Sandhills Regions?"

2

Understand your Numbers !

Understand
Numbers with

Statistics

“

Statistics without business interpretation is just math.

Business interpretation without statistical evidence is just opinion.

Together, they create powerful **data-driven decision making**.

Learning Objective

1. Understand what statistics is and why it matters in banking
2. Identify different types of data and variables
3. Learn data collection methods and their applications

Our Dataset

Our Dataset: Portuguese Bank Marketing Campaign

Source: UCI Machine Learning Repository

Domain: Direct marketing campaigns for term deposits

Size: 41,188 customers with 20 features

Time Period: May 2008 - November 2010

What is Statistics

The Science of Data-Driven Decision Making

In Banking & Finance, Statistics Helps Us

📊 Assess credit risk and predict loan defaults

🎯 Optimize marketing campaigns and customer targeting

💰 Detect fraudulent transactions and unusual patterns

📈 Understand customer behaviour and preferences

🏦 Make data-driven business decisions and strategy

Types of Data

Continuous Data 📈

Can take any value within a range (including decimals)

From Our Dataset:

age: Customer age (25, 35, 47.5, 62, etc.)

balance: Average yearly balance in euros (€1,234.56, €15,678.90)

duration: Last contact duration in seconds (180.5, 642.3, 1,205.7)

Discrete Data 🔢

Countable whole numbers only

From Our Dataset:

campaign: Number of contacts in current campaign (1, 2, 3, 4, 5...)

previous: Number of contacts in previous campaigns (0, 1, 2, 3...)

day: Day of month for last contact (1, 2, 3, ..., 31)

Quantitative

Types of Data

Nominal Data

Categories without meaningful order

From Our Dataset:

job: Type of occupation
admin., blue-collar, entrepreneur, housemaid, management

marital: Marital status (*divorced, married, single, unknown*)

contact: Communication type (*cellular, telephone, unknown*)

Ordinal Data

Categories with meaningful order/ranking

From Our Dataset:

education: Education level (ranked by years of education)
basic.4y → basic.6y → basic.9y → high.school → professional.course → university.degree

poutcome: Previous campaign outcome (ranked by success)
failure → unknown → other → success

Qualitative

Sampling vs Census

Census 🌍

Collecting data from entire population

When to Use:

- Small population size
- High accuracy required
- Regulatory compliance needed

Banking Example:

- Analyzing all transactions for fraud detection
- Reviewing all loan applications in a small branch

Sampling 🎲

Collecting data from subset of population

When to Use:

- Large population size
- Cost/time constraints
- Destructive testing

Our Dataset Example:

- 41,188 customers sampled from total Portuguese bank customers
- Represents broader customer base for marketing insights

Sampling vs Census

💡 **Sampling Benefits:** Cost-effective, faster analysis, still provides accurate insights when done properly.

Learning Objective

Descriptive Statistics - Making **Sense** of Numbers

1. Calculate and interpret measures of central tendency (mean, median, mode)
2. Understand measures of dispersion (range, variance, standard deviation)
3. Analyze data shape (skewness and kurtosis)
4. Apply all concepts to solve real banking business problems
5. Make business decisions using descriptive statistics

👀 **Our Focus:** Transform raw numbers into meaningful business insights using the Portuguese Bank Marketing dataset.

💡 **Key Question:** *"How can we summarize 41,188 customers' data into actionable business intelligence?"*

Descriptive Statistics

Your Data Storytelling Toolkit

Three Key Questions Descriptive Statistics Answer:

1. **What's Typical?** (Central Tendency) - *"What's our average customer like?"*
2. **How Much Variation?** (Dispersion) - *"How different are our customers?"*
3. **What's the Pattern?** (Shape) - *"Are most customers similar or diverse?"*

Business Value:

📊 Understand customer characteristics

🎯 Identify target segments

💰 Assess risk levels

📈 Measure campaign performance

🏆 Benchmark against competitors

Measures of Central Tendency

Finding the "Typical" Customer

Three Ways to Describe "Average":

1. Mean (Arithmetic Average) 📊
Formula: Sum of all values ÷ Number of values
Best for: Normal distributions, no extreme outliers
Business use: Budget planning, expected values

2. Median (Middle Value) 📍
Formula: Middle value when data is sorted
Best for: Skewed data, presence of outliers
Business use: Typical customer profile, realistic expectations

3. Mode (Most Common) 🎯
Formula: Most frequently occurring value
Best for: Categorical data, understanding preferences
Business use: Popular products, common characteristics

💡 **Key Insight:** Each measure tells a different story about your customers!

Choosing the Right Central Tendency

Situation	Best Measure	Why	Business Example
Normal Distribution	Mean	Mathematically precise	Average loan amount for budgeting
Skewed Data	Median	Not affected by outliers	Typical customer balance
Categorical Data	Mode	Shows most common	Most popular account type
Outliers Present	Median	Robust to extremes	Representative customer age
Small Sample	Median	Less sensitive to errors	Branch performance comparison

Real Banking Example: Executive asks: *"What's our typical customer's account balance?"*

Analysis 📊

- **Mean:** €4,500 (affected by millionaire customers)
- **Median:** €1,800 (true middle customer)
- **Recommendation:** Use median (€1,800) for realistic planning

Measures of Dispersion

Understanding Customer Diversity

Why Dispersion Matters: Knowing the average isn't enough - you need to understand the spread!

Example: Two branches both have average customer age of 40 years:

Branch A: Ages range 38-42 (very similar customers)

Branch B: Ages range 20-60 (very diverse customers)

Strategy: Branch A needs focused approach, Branch B needs diverse offerings

Four Key Measures:

Range - Simple spread

Variance - Average squared deviation

Standard Deviation - Average deviation in original units

Coefficient of Variation - Relative variability

💡 **Business Application:** Risk assessment, customer segmentation, product customization

Standard Deviation - The Most Useful Spread Measure

Your Go-To Measure for Understanding Variation

Formula: Standard Deviation = √Variance

Example Continued - Customer Age: Column Used: age

Variance = 250 years² Standard Deviation = √250 = 15.8 years

Business Interpretation:

- *"Customer ages typically vary by ±15.8 years from the average"*
- *"If average age is 45, most customers are between 29-61 years old"*
- *"Age diversity is moderate - need varied but not extreme approaches"*

💡 **Business Application:** Setting realistic ranges, identifying outliers, risk assessment

Coefficient of Variation – Comparing Different Variables

The Relative Variation Champion

Formula: CV = (Standard Deviation ÷ Mean) × 100%

Why CV is Powerful: Allows comparison between variables with different units!

Example from Our Dataset: Columns Used: age and balance

```
Customer Age: Mean = 40 years, SD = 10 years
CV = (10 ÷ 40) × 100% = 25%
Account Balance: Mean = €1,500, SD = €2,000
CV = (2,000 ÷ 1,500) × 100% = 133%
```

Business Interpretation:

- *"Account balances are much more variable (133%) than ages (25%)"*
- *"Customer ages are relatively consistent across our base"*
- *"Account balance-based strategies need more flexibility"*
- *"Age-based segmentation will be more stable than balance-based"*

CV Guidelines:

- **CV < 30%:** Low variation
- **CV 30-60%:** Moderate variation
- **CV > 60%:** High variation

Measures of Shape

Understanding Data Distribution Pattern

Two Key Measures:

Skewness 📈

Measures if data leans to one side

Positive Skew: Long tail to the right (few very high values)

Negative Skew: Long tail to the left (few very low values)

Symmetric: Balanced distribution

Kurtosis ⛰️

Measures how peaked or flat the distribution is

High Kurtosis: Sharp peak, heavy tails (more extreme values)

Low Kurtosis: Flat distribution (fewer extreme values)

💡 **Why Shape Matters:** The shape of your data tells you about customer behavior patterns and guides your analysis approach.

Portuguese Bank Campaign Analysis Challenge

Business Context: You're the Data Analyst for the Portuguese Bank's Marketing Department. The Marketing Director needs insights for next quarter's campaign strategy.

The Challenge: *"We spent €2.5M on phone campaigns last year with mixed results. Before planning next quarter's €800K campaign, I need to understand our customer base better. Help me make data-driven decisions."*

Available Data: 41,188 customers from previous campaigns

Timeline: 1 week to present findings

Stakeholders: Marketing Director, Branch Managers, Executive Team

👀 **Your Task:** Use descriptive statistics to provide actionable insights for campaign optimization.

Business Questions to Answer

Translating Business Needs into Statistical Analysis

Marketing Director's Key Questions:

Q1: Customer Demographics *"Who are our typical customers? What age should we target?"*

Q2: Financial Profiles *"What account balances do our customers have? Should we focus on wealthy or regular customers?"*

Q3: Contact Strategy *"How many calls does it typically take to convert a customer?"*

Q4: Market Segmentation *"What job categories should we prioritize?"*

Q5: Campaign Efficiency *"Are our campaigns consistent or highly variable in performance?"*

Business Questions to Answer

Q1: "Who are our typical customers?"

Column Analyzed: age

Statistical Analysis:

```
Age Statistics:
Mean = 40.9 years
Median = 39.0 years
Mode = 32 years
Standard Deviation = 10.6 years
Range = 18 to 95 years
Coefficient of Variation = 25.9%
```

Business Interpretation:

🎯 **Target Age:** 39 years (median gives best "typical" customer)

📊 **Age Spread:** Most customers between 29-51 years (±1 SD)

📈 **Distribution:** Slightly positive skew (some older customers)

🔄 **Consistency:** Moderate variation (CV = 25.9%)

Actionable Recommendations:

Primary Target: 30-50 age group (core market)

Marketing Message: Mid-career financial planning focus

Channel Strategy: Digital + traditional mix for this age group

Business Questions to Answer

Q2: "What account balances do our customers have?"

Column Analyzed: balance

Statistical Analysis:

```
Balance Statistics:
Mean = €1,362
Median = €448
Mode = €0
Standard Deviation = €3,044
Range = €-8,019 to €102,127
Coefficient of Variation = 223%
```

Business Interpretation:

💰 **Typical Customer:** €448 balance (median more realistic than mean)

📊 **Wealth Distribution:** Highly skewed - few wealthy customers

🔄 **High Variability:** CV = 223% indicates diverse financial situations

⚠️ **Outliers Present:** Some very wealthy and some in debt

Strategic Segmentation:

- **Mass Market (70%):** €0 - €1,000 balance
- **Affluent (25%):** €1,000 - €10,000 balance
- **Premium (5%):** €10,000+ balance

Campaign Budget Allocation:

- **60% budget:** Mass market (high volume, low margin)
- **30% budget:** Affluent (moderate volume, good margin)
- **10% budget:** Premium (low volume, high margin)

Business Questions to Answer

Q3: "How many calls does it take to convert customers?"

Columns Analyzed:
campaign (number of contacts) and
y (success)

Statistical Analysis:

Campaign Contacts for Successful Customers:

```
Mean = 2.8 contacts
Median = 2.0 contacts
Standard Deviation = 2.4 contacts
```

Success Rate by Contact Number:

```
1 contact: 15.2% success
2 contacts: 12.1% success
3 contacts: 8.9% success
4+ contacts: 5.3% success
```

Business Interpretation:

📞 **Optimal Strategy:** Most conversions happen in first 2 calls

📉 **Diminishing Returns:** Success rate drops significantly after 3rd call

💰 **Cost Efficiency:** Early contacts are most cost-effective

⏱️ **Quick Decision:** Customers who will convert decide early

Strategic Recommendations:

- **Focus Resources:** Prioritize high-quality first and second contacts
- **Call Limit:** Set maximum of 3 contacts per customer
- **Early Qualification:** Identify interested customers quickly
- **Resource Reallocation:** Redirect budget from repeated calls to new prospects

Business Questions to Answer

Q5: "Are our campaigns consistent or highly variable?"

Multiple Columns Analyzed: Key performance metrics

Variability Analysis:

```
Performance Consistency:
Success Rate by Month: CV = 34%
(moderate variation)
Success Rate by Age Group: CV = 28%
(low variation)
Success Rate by Balance Group: CV = 67%
(high variation)
Success Rate by Job Category: CV = 45%
(moderate-high variation)
```

Business Interpretation:

📅 **Seasonal Patterns:** Moderate monthly variation suggests seasonal effects

👥 **Age Consistency:** Similar response across age groups

💰 **Balance Sensitivity:** Highly variable by wealth level

🏢 **Job Variation:** Significant differences between job categories

Risk Assessment:

Low Risk: Age-based campaigns (consistent performance)

Medium Risk: Time-based campaigns (seasonal variation)

High Risk: Balance-based campaigns (unpredictable results)

Comprehensive Business Recommendations

Target Customer Profile 🎯

- **Primary Age:** 30-50 years (median 39 years)
- **Financial Profile:** €448 typical balance (focus on mass market)
- **Job Priority:** Students > Retired > Management > Technician

Contact Strategy 📞

- **Optimal Contacts:** Maximum 2-3 calls per customer
- **Success Timeline:** 72% of conversions happen in first 2 contacts
- **Resource Allocation:** Focus on quality, not quantity of contacts

Segmentation Strategy 📊

Segment 1: High-Value Niche (30% of budget)
- Students and Retired customers
- Expected conversion: 25-31%
- Lower volume, higher success

Segment 2: Volume Play (60% of budget)
- Management and Technician professionals
- Expected conversion: 12-14%
- High volume, moderate success

Segment 3: Avoid or Redesign (10% of budget)
- Services sector customers
- Current conversion: <10%
- Needs different approach

Expected ROI Improvement:

- **Current Campaign:** 11.7% overall success rate
- **Optimized Campaign:** 18-22% projected success rate
- **Budget Efficiency:** 40-60% improvement expected

Practice Problems

Problem 1: Credit Risk Assessment

The bank wants to assess risk for loan approvals.

Using the balance column:

- Calculate mean, median, and standard deviation
- Interpret what these values mean for loan risk
- Recommend risk thresholds

Problem 2: Branch Performance Compare campaign success rates between different education levels (education and y columns):

- Calculate success rate for each education level
- Determine which group to prioritize
- Explain your business reasoning

Problem 3: Customer Contact Optimization Analyze the duration column (call length in seconds):

- Find the optimal call duration range
- Calculate the coefficient of variation
- Recommend call center training guideline

Problem 4: Seasonal Analysis Using the month column:

- Identify the mode (most common contact month)
- Calculate success rate variability across months
- Suggest seasonal campaign strategies