How to do Data Analysis (Step by step)

Complete steps for a successful Business case

By SHAILESH SHAKYA @BEGINNERSBLOG.ORG

MINDSET:

What a real business case actually is

A business case is not “EDA + charts.”

It is:

1. A clear business problem.
2. A clean, well-understood dataset.
3. A structured analysis (EDA → tests → models).
4. A small set of sharp decisions with quantified impact.

Everything you do with data must line up with decisions the company can actually take.

1.CLARIFY THE BUSINESS PROBLEM FIRST :

Do this before opening a notebook.

1.1 Define the business objective (1–2 lines, no jargon)

Examples for e-commerce:

- "Increase monthly revenue by 15% in the next 12 months."
- "Reduce customer churn from 20% to 15% in 6 months."
- "Increase conversion rate on product page from 2.5% to 3.5%."

1.2 DEFINE PRIMARY & SECONDARY KPIS

Typical e-commerce KPIs: revenue, AOV, conversion rate, repeat purchase rate, CLTV, churn, margin, inventory turnover

Example

- Primary KPI: Conversion ra
- Secondary: AOV, revenue per visitor, bounce_rate:.

1.3 TURN THIS INTO 3–5 CONCRETE QUESTIONS

Example:

- Which product categories and customer segments drive most revenue?
- When are sales highest/lowest (month, weekday, season)?
- Where do we lose customers in the funnel (traffic → add-to-cart → checkout → purchase)?
- Do repeat buyers behave differently from one-time buyers?

These questions will drive all your EDA, tests, and models.

2.

UNDERSTAND AND AUDIT THE DATA PROPERLY

Now open the data and behave like a QA engineer, not a model builder.

2.1 INVENTORY DATA SOURCES

LIST CLEARLY:

- Transactions: order_id, customer_id, order_date, product_id, quantity, price, discount, etc.
- Product catalog: product_id, category, subcategory, brand, cost price.
- Customers: customer_id, signup_date, country, segment, marketing channel.
- Web / app analytics: sessions, pageviews, add_to_cart events, device type, source/medium.
- Marketing: campaigns, spend, channel, impressions, clicks.

You will rarely have all of these, but you must know what exists and what is missing.

2.2 BASIC STRUCTURE CHECKS

In pandas:

- df.shape → number of rows and columns (data volume)
- df.info() → dtypes, non-null counts
- df.head() → sanity-check columns and example records

Ask yourself:

- Do columns match business logic? (E.g., “Order Date” is datetime, “Quantity” is integer, “Sales” is numeric.)
- Are key IDs present and non-null? (order_id, customer_i d, product_id)

2.3 MISSING VALUES

Steps:

- Count missing per column: df.isnull().sum()
- Calculate percentage missing: df.isnull().mean() * 100

Classify:

- Operationally critical fields: order_id, date, quantity, price → usually must not be missing.
- Optional fields: phone number, secondary address, coupon_code → can sometimes be missing w ithout harm.

2.3 continue)

HANDLING STRATEGY (SIMPLE RULE-SET):

- If a column has > 40–50% missing and isn't critical → often drop the column or treat as "unknown" category.
- If a small % of rows (< 5%) has missing in critical columns → drop those rows.
- For numeric features (e.g., discount, rating) with moderate missing → impute:
    - Median for skewed data (e.g. sales, prices)
    - Mean for roughly symmetric distributions
    - More advanced: KNN / model-based imputation when necessary
- For categorical features (e.g., segment, city) → use mode or a separate category "Missing".

Always document what you did and why.

2.4 DUPLICATES

- df.duplicated().sum() → check full row duplicates
- Also check duplicate order_id or customer_id + date combinations if that should be unique.

If duplicates are true data errors (same order repeated), drop them. If they represent multiple items in one order, that's fine.

2.5 OUTLIERS AND OBVIOUS ERRORS

Examples from real e-commerce datasets: negative quantity, zero or negative price, insane prices like 999999, timestamps in the future.

Steps:

- Use df.describe() to see min/max, mean vs median.
- Visualize with boxplots / histograms for key numeric columns (sales, quantity, AOV).

Rules of thumb:

- Impossible values → fix or remove (negative price, negative quantity)
- Extreme but possible values (very high AOV) → flag them as outliers but don't blindly drop; investigate first (B2B ord ers, wholesale)

3.

FEATURE ENGINEERING FOR USEFUL BUSINESS VIEWS

You rarely get the features you actually need directly. Build the m.

3.1 TIME-BASED FEATURES

From order_date:

- order_year, order_month, order_week, order_dayofweek (0 = Monday)
- is_weekend = dayofweek in {5, 6}
- order_hour if you have timestamps

Use these to see seasonality, weekday patterns, holiday spikes.

3.2 REVENUE AND MARGIN

Create:

- sales = quantity * unit_price
- gross_margin = sales - cost (if cost available)
- discount_rate = discount / (unit_price * quantity)

These support profit-focused decisions, not just revenue.

3.3 CUSTOMER METRICS (RFM-STYLE

Per customer_id:

- recency = days since last purchase
- frequency = number of orders
- monetary = total revenue from that customer

Use RFM to segment into VIP, regular, new, at- risk, etc.

3.4 PRODUCT AND CATEGORY METRICS

Per product or category:

- Total sales
- Number of orders
- Average discount
- Return rate (if you have returns)
- Profit per unit

This tells you which products or categories really drive the business.

4.

EXPLORATORY DATA ANALYSIS (EDA) WITH A BUSINESS LENS

EDA is not random charting. Each plot should answer a specific questio n.

4.1 DESCRIPTIVE STATISTICS

Use df.describe() on key numeric fields (sales, quantity, AOV, sessions, conversion_rate).

Look for:

- Mean vs median differences → skewness.
- Very wide ranges → potential outliers.
- High std → unstable metric.

Example interpretation:

- "Average sale is 4× the median" → only a few very large orders; consider treating enterprise / wholesale separately.

4.2 CORE VISUALIZATIONS YOU SHOULD ALWAYS DO

Sales over time (line chart)

- Total monthly sales, AOV, orders over tim
- Look for growth trend, seasonality (e.g., Q4 spikes), drops (stockouts, tracking issues)

Sales by category / segment (bar char)

- Revenue by product category, subcategory, and customer segmen.
- Identify top categories (e.g., Technology 36.4% of revenue) and weak ones

Order value distribution (histogram)

- Histogram of AOV or order_sale
- Look at right-skew; decide whether to use median, log-transform, or treat big orders separately.

4.2 CONTINUE

1. **Customer behavior (RFM plots)**
    - Scatter of frequency vs monetary; recency vs monetary.
    - Quickly see VIPs (high frequency, high monetary) vs at-risk (high monetary, high recency).
2. **Correlation matrix + scatter plots**
    - Correlations among numeric features: orders, sessions, conversion_rate, AOV etc.
    - Example: Fulfilled Orders vs Delivered Orders with strong linear correlation → features are redundant.

Each visual should end with 1–2 sentences of business meaning, not just “this is skewed.”

5. FORMULATE HYPOTHESES FROM EDA

Take patterns from EDA and convert them into explicit statements you can test.

Examples:

- "Customers acquired via paid ads have a higher AOV than customers from organic search."
- "The new checkout design increases conversion rate compared to the old design."
- "High-frequency buyers have significantly higher CLTV than low-frequency buyers."

For each:

- Define null hypothesis (H_0): no difference.
- Define alternative hypothesis (H_1): there is a difference (or a dire ctional effect).

6.

HYPOTHESIS TESTING / A/B TESTING

You now evaluate if patterns are statistically real or random noise.

6.1 STANDARD 4-STEP PROCESS

1. State H_0 and H_1
2. Choose significance level (α = 0.05 is standard)
3. Choose a test & compute statistic / p-value
4. Decision: if $p < \alpha$, reject H_0, else fail to reject H_0

6.2 CHOOSING THE RIGHT TEST (SIMPLE RULES)

- Compare means, large sample, roughly normal → t-test
- Compare medians, non-normal / skewed → Wilcoxon / Mann–Whitney
- Compare proportions (e.g., conversion rates) → z-test or proportion test
- Association between two categorical variables (e.g., device type vs conversion) → chi-square test

Always:

- Check normality (histogram + QQ-plot, Shapiro–Wilk) before using parametric tests.
- Use non-parametric tests when normalit y is violated.

6.3 EXAMPLE

Check normality (histogram + QQ-plot, Shapiro–Wilk) before using parametric tests.
Use non-parametric tests

Claim: “High-order-count customers have higher revenue than average.”

- H_0: median revenue_high = median revenue_all
- H_1: median revenue_high > median revenue_all
- Data is skewed → use Wilcoxon / Mann–Whitney instead of t-test.
- If p < 0.05 → evidence supports claim; treat high-order customers as VIP segment with targ eted offers.

7.

ADVANCED ANALYTICS AND MODELING

Use only after EDA and tests are sound. Goal: predict or segment, not just describe.

7.1 FORECASTING (REGRESSION / TIME SERIES)

Use historical data to predict:

- Future sales
- Future demand per category
- Impact of traffic on sales

Example: linear regression "next_month_sales = 0.02 × this_month_traffic + 983".

Use cases:

- Inventory planning (avoid stockouts/overstock).
- Budget planning for marketing.

For more advanced, use ARIMA, Prophet, or ML models (XGBoost, LSTM). Validate with train/test splits and cross-validation.

7.2 CUSTOMER SEGMENTATION

Methods:

- RFM segmentation → simple, business friendly.
- K-means or other clustering on RFM or behavior features (pages visited, categories viewed).

Output:

- Named segments: “High-value loyalists,” “Discount hunters,” “New low-value,” “At-risk high-value.”
- Action per segment: loyalty perks, win-back campaigns, product recommendations.

7.3 CLASSIFICATION MODELS

Predict:

- Probability a visitor buys within a session (lead scoring).
- Probability a customer churns in the next 30 days.

Use logistic regression, decision trees, random forests, gradient boosting. Evaluate with accuracy, AUC, precision/recall depending on the problem

7.4 RECOMMENDATION SYSTEMS

- Recommend products to increase basket size:
- Collaborative filtering: “users like you bought…”
- Content-based: similar products by category/attributes.

Tie results back to uplift in AOV and c ross-sell.

8.

HANDLING LACK OF DATA AND DATA QUALITY ISSUES

Real datasets are im
documents correctly

By SHAILESH SHAKYA

KEY STRATEGIES:

Gather more data

- Extend time window.
- Add external / open datasets (macro trends, demographics).

Use synthetic data carefully

- Generate realistic but fake transactions to augment small samples (GANs, bootstrapping, etc.).
- Good for model training when privacy or volume is a problem, but always flag that you used syn thetic data.

KEY STRATEGIES:

Simplify models

- With little data, avoid complex deep models.
- Prefer simpler, interpretable models with regularization.

Be explicit about limitations in the business case

- "Results are based on 3 months of data, which may not capture seasonality."
- "Synthetic data was used to augment minority segments; actual behavior may differ."

9. MAP EVERYTHING BACK TO KPIS

Every finding must connect back to KPIs and business levers.

Examples:

- "If we lift conversion rate from 2.5% to 3.0% at current traffic, monthly revenue increases by X."
- "Reducing churn by 5 percentage points in the VIP segment increases CLTV by Y and annual revenue by Z."
- "Bundling underperforming items with top sellers is expected to increase AOV by 10%, adding $A per month."

This is the "language" executives understand.

10.

HOW TO WRITE THE BUSINESS CASE DOCUMENT / DECK

Structure your final output like this.

By SHAILESH SHAKYA

POWERED BY:
BEGINNERSBLOG.ORG

10.1 EXECUTIVE SUMMARY

3 bullets: problem, key insight, recommended actions + expected impact.

Example:

- "Conversion rate is healthy, but 70% of revenue comes from 15% of customers."
- "VIP customers are not targeted; churn in this segment is high."
- "Introduce VIP program + targeted win-back; expected +8–12% revenue uplift."

10.2 Problem & objectives

Business context: what the company does, current challenges.

Concrete objectives and KPIs defined in section 1.

10.3 Data & methodology

- Data sources (transactions, customers, web, marketing).
- Time period used.
- Key cleaning steps: missing values, duplicate removal, outlier treatment.
- Methods used: EDA, visualizations, hypothesis tests, models.

10.4 Key EDA findings (with charts)

- Sales over time (trend & seasonality).
- Revenue by category/segment.
- Distribution of order values, customer spend.
- Main correlations (e.g., traffic vs sales, discount vs conversion).

Each chart must have:

- Clear title that states the insight, not the chart type.
- 1–2 bullet insights under it.

10.5 Hypothesis test results

For each important test:

- Hypothesis, metric, groups compared.
- Test used (t-test, Wilcoxon, etc.), p-value, conclusion.

Example:

- "New checkout vs old: conversion increased from 2.5% to 3.0%, p = 0.01 → statistically significant improvement."

10.6 Modeling & projections (if used)

- Short explanation of model: target, features, model type.
- Quality metrics (R^2, AUC, etc. in simple terms).
- Forecasts: likely range of outcomes (not just a single number).

10.7 Recommendations

Each recommendation should have:

1. What to do.
2. Why (link to findings).
3. Expected impact (numbers).
4. Implementation notes / risks.

Example:

- Action: Launch VIP loyalty program for top 10% customers by monetary value.
- Why: This segment contributes 45% of revenue but has 18% churn.
- Impact: Reducing churn to 12% increases annual revenue by ~$X.
- Notes: Requires coupon system + email automation.

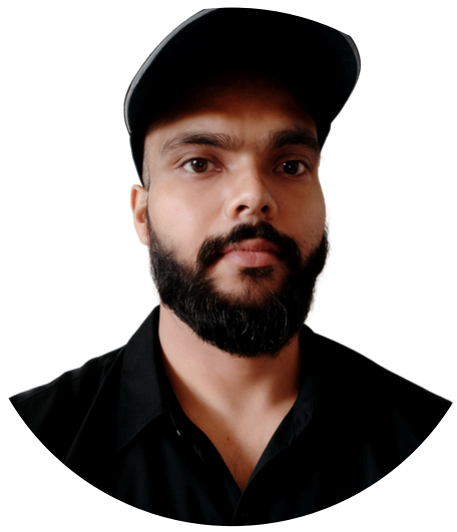
10.8 Risks, limitations, and next steps

- Data limitations, experimental caveats, assumptions
- Next steps: refine models, run more A/B tests, improve tracking.

11. Simple mental template you can reuse on any e-commerce case

Use this as a checklist:

- What is the business problem and KPI?
- What data do I have, and is it clean (shape, types, missing, duplicates, outliers)?
- What features do I need to answer the questions (time, revenue, RFM, categories)?
- What do the distributions and trends say (EDA + visuals)?
- What hypotheses can I test, and what tests will I use?
- Do I need forecasting, segmentation, or prediction to support decisions?
- What data limitations exist, and how did I handle them?
- What 3–5 actions should the business take, and what is the expect ed impact on KPIs?



I hope you have found this information helpful

Join **OpenAILearning** to get more educational stuff Similar to this you finished reading⤵️⤵️

Telegram: **OpenAILearning**

WhatsApp: **OpenAILearning**

Thank You!

Swipe to Next Slide

By SHAILESH SHAKYA

POWERED BY: BEGINNERSBLOG.ORG

Created by Shailesh Shakya

@BEGINNERSBLOG.ORG

Did you find this post helpful? Please...