Swiss UMEF
UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics

05: In-Memory Analytics with Pandas. Exploratory Data Analysis

Instructor: Oleh Tymchuk

#05: Agenda

- Introduction to EDA
- Summary statistics
- Practical cases
- Useful Links

Introduction to EDA

What is EDA?

- Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their key characteristics
- It helps in understanding the structure, distribution, and relationships within the data
- EDA allows us to identify patterns, anomalies, missing values, and outliers before applying machine learning models

Why is EDA important?

- Understand data structure
- Identify relationships between variables
- Prepare data for modeling

Types of EDA techniques

- Univariate analysis (examining single variables, e.g., histograms, box plots)
- Bivariate analysis (exploring relationships between two variables, e.g., scatter plots, correlation analysis)
- Multivariate analysis (analyzing more than two variables, e.g., heatmaps, pair plots)

Tools for EDA

- Pandas: Data manipulation and analysis
- NumPy: Numerical computations
- Matplotlib/Seaborn/Plotly: Data visualization

Steps in EDA

[x] Understanding the dataset

[x] Handling missing values

[x] Checking data types and conversions

[-] Summary statistics

[-] Data visualization

[-] Identifying outliers and anomalies

Scales

Nominal Scale

Definition: categorical data **without order.**

Used to classify objects into distinct groups.

Examples:

- Colors (red, blue, green)
- Product types (smartphones, laptops)
- Countries (Ukraine, Germany, Japan)
- Gender (male, female)

Key Features:

- No mathematical meaning in values
- Only **frequency** or **mode** (most frequent category) can be calculated
- Visualization: pie charts, bar plots

Ordinal Scale

Definition: categorical data **with order**, but intervals between values are not equal or measurable.

Examples:

- Education level (primary < secondary < tertiary)
- Product ratings (1 ★ < 2 ★ < 5 ★)
- Disease stages (mild < moderate < severe)
- Income levels (low, medium, high)

Key Features:

- Order matters, but differences between values are not quantified.
- **Median** and **ranks** are appropriate statistics.
- Visualization: ordered bar plots, Likert scales.

Quantitative Scale

Definition: numerical data **with mathematical meaning**. Divided into two subtypes:

- Discrete (integers): number of products, children in a family.
- Continuous (decimal numbers): weight, height, temperature.

Examples:

- Age (25 years, 30.5 years)
- Salary ($50,000)
- Delivery time (2.5 hours)
- Website views (1,000,000)

Key Features:

- All mathematical operations (+, −, ×, ÷) apply.
- Use **mean**, **standard deviation**, **variance**.
- Visualization: histograms, box plots, scatter plots.

Comparison of Scales

Criterion	Nominal	Ordinal	Quantitative
Order	✘ No	✔ Yes	✔ Yes
Equal Intervals	✘ No	✘ No	✔ Yes
Math Operations	✘ Not meaningful	✘ Limited (on ranks only)	✔ All arithmetic operations allowed
Statistics	Mode, frequency	Median, mode, rank order	Mean, median, mode, variance, standard deviation
Example	Gender, colors, country names	Product ratings, education levels	Weight, height, temperature, age

Important Notes:

- Common Mistake: Calculating the mean for ordinal data (e.g., "average rating 3.8" is technically incorrect).
- Rule: Statistical methods and visualizations depend on the scale type. Always validate assumptions before analysis.

Summary statistics

Summary Statistics

Concept:

- Summary statistics are a subset of descriptive statistics that provide a concise overview of the data
- They summarize key characteristics of the dataset using numerical metrics

Why is it important?

- Helps us understand the overall structure of the data
- Identifies patterns, trends, and potential issues (e.g., outliers, missing data)
- Provides a foundation for further analysis or modeling

Central Tendency. Mean

- **Calculation**:
  ages = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]
  $$\text{Mean} = \frac{\sum x_i}{n} = \frac{23 + 29 + \cdots + 65}{10} = \frac{447}{10} = 44.7$$
  Interpretation: The average age is 44.7 years.
- **Meaning**: Balances all values equally; "center of gravity"
- **Use when**: Data is symmetric, continuous, no outliers
- **Not for**: Categorical data or ordinal where intervals aren’t equal
- **Good use**: Mean income, mean height
- **Bad use**: Mean customer satisfaction (on 1–5 scale) — misleading due to ordinal nature

Central Tendency. Mean

- **Monthly Salaries**: $3000, $3200, $2800, $3100, $2950

  Mean: Yes/No?

- **Student Exam Scores**: 72, 85, 90, 65, 78

  Mean: Yes/No?

- **Product Ratings**: 3, 4, 2, 5

  Mean: Yes/No?

- **Zip Codes**: 90210, 10001, 30301

  Mean: Yes/No?

Central Tendency. Mean

How to interpret:

- High mean → overall tendency toward larger values
- Low mean → most observations are relatively small

Example insight:

- The average income in the region is $48,000 — most people earn around this amount

Warning: Sensitive to outliers!

- One billionaire can completely skew the result

Central Tendency. Median

- **Calculation**:
  ages = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]
  Sort data: [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]
  Median = (42 + 45) / 2 = 43.5
  Interpretation: 50% of individuals are younger than 43.5 years, and 50% are older.
- **Meaning:** Middle-ranked value; resistant to outliers
- **Use when:** Skewed distributions, ordinal data
- **Not for:** Nominal data (no order)
- **Good use:** Median household income
- **Bad use:** Median country name

Central Tendency. Median

- **Apartment Prices**: $200k, $220k, $180k

  Median: Yes/No?

- **Test Scores**: 40, 60, 90, 95, 100

  Median: Yes/No?

- **Customer Satisfaction**: 1, 2, 2, 4, 5

  Median: Yes/No?

- **Cities**: Tokyo, Paris, London, Berlin

  Median: Yes/No?

Central Tendency. Median

How to interpret

- Median > Mean → right-skewed distribution (some large outliers)
- Median < Mean → left-skewed distribution (some small outliers)

Example insight

- The median income is $35,000, which is lower than the mean — the wealthy pull the average up. Most people earn less than the average.

Central Tendency. Mode

- **Calculation**: Most frequent value(s)
  ages = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]
  Here, 42 occurs twice
  Interpretation: The most common age is 42 years
- **Meaning:** Most frequent observation
- **Use when:** You care about frequency
- **Applies to:** all scales
- **Good use:** Most common customer complaint type
- **Less useful:** Continuous data (e.g., weight with all unique values)

Central Tendency. Mode

- **Shoe Sizes:** 38, 38, 39, 40, 38

  Mode: Yes/No
- **Car Colors:** Red, Blue, Blue, Black

  Mode: Yes/No
- **Temperatures (°C):** 22, 23, 22, 21

  Mode: Yes/No
- **Product Codes:** A123, B321, A123, A123

  Mode: Yes/No

Central Tendency. Mode

How to interpret

- Especially useful for categorical data (nominal or ordinal)
- Tells you what's most common, not what's "central"

Example insight

- The most popular coffee type is “Latte” — we should consider promoting it more

Central Tendency. Practical cases

Measures of Spread. Range

- **Calculation:**
  ages = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]
  $\text{Range} = \text{Max} - \text{Min} = 65 - 23 = 42$
  Interpretation: The ages span 42 years.
- **Meaning:** Difference between extremes (max - min)
- **Use when:** Quick sense of spread
- **Not for:** Categorical data, ordinal scales with unclear spacing
- **Good use:** Range of temperatures
- **Bad use:** Range of product satisfaction ratings (1 to 5) may ignore distribution shape

Measures of Spread. Range

- **Lifespan (years)**: 70, 85, 90, 95

  Range: Yes/No?

- **Temperature (°F):** 32, 45, 60, 55

  Range: Yes/No?

- **Star Ratings:** 1⭐, 2⭐, 4⭐, 5⭐

  Range: Yes/No?

- **Country Names:** USA, France, Germany

  Range: Yes/No?

Measures of Spread. Range

How to interpret

- A simple measure of total spread; shows how far apart the smallest and largest values are.

Example insight

- The age range in the group is 22 to 65 — a diverse age group.

Warnings

- Very sensitive to outliers
- Doesn’t reflect variability in the middle of the data

Measures of Spread. Variance / Standard deviation

Calculation. Variance

Situation	Formula	Denominator	Reason
Population	$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$	N	You have all data — no estimation
Sample	$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	$n - 1$	Corrects bias in small samples

Calculation. Standard deviation

$$s = \sqrt{s^2} \quad \text{or} \quad \sigma = \sqrt{\sigma^2}$$

Measures of Spread. Variance / Standard deviation

Let's say we have a sample of **monthly sales in $1000:**

[5, 7, 3, 7, 10]

Step 1: Calculate the mean

$$\bar{x} = \frac{5+7+3+7+10}{5} = \frac{32}{5} = 6.4$$

Step 2: Subtract the mean and square the result

$$(5-6.4)^2 = 1.96 \quad (7-6.4)^2 = 0.36 \quad (3-6.4)^2 = 11.56 \quad (7-6.4)^2 = 0.36 \quad (10-6.4)^2 = 12.96$$

Step 3: Add the squared deviations

$$\sum (x_i - \bar{x})^2 = 1.96 + 0.36 + 11.56 + 0.36 + 12.96 = 27.2$$

Step 4: Divide by $n-1$ (sample size – 1)

$$s^2 = \frac{27.2}{4} = 6.8 \quad \text{(Variance)}$$

Step 5: Take the square root

$$s = \sqrt{6.8} \approx 2.61 \quad \text{(Standard Deviation)}$$

Measures of Spread. Variance / Standard deviation

- **Calculations:** IQR = Q3 - Q1,
  where: Q1 = 25th percentile (lower quartile) ; Q3 = 75th percentile (upper quartile)

  Data: [3, 7, 8, 5, 12, 14, 21, 13, 18] -> [3, 5, 7, 8, 12, 13, 14, 18, 21]
  Median = 12
  Q1 = median of lower half: 5
  Q3 = median of upper half: 14
  IQR = 14 – 5 = 9

- **Meaning:** Spread around the mean (variance is squared, std is same units as data)
- **Use when:** Data is numerical, especially if symmetric
- **Not for:** Categorical or ordinal without equal intervals
- **Good use:** Std deviation of monthly sales
- **Bad use:** Variance of education levels coded as 1–5

Measures of Spread. Standard deviation

- **Product Weights (kg)**: 2.3, 2.5, 2.1, 2.4

  Std Dev: Yes/No?

- **Blood Pressure (mmHg)**: 120, 130, 125

  Std Dev: Yes/No?

- **Satisfaction Scores**: 3, 4, 5, 2

  Std Dev: Yes/No?

- **ID Numbers**: 102, 105, 110

  Std Dev: Yes/No?

Measures of Spread. Standard deviation

How to interpret

- A low standard deviation → data points are close to the mean
- A high standard deviation → data is more spread out

Example insight

- The standard deviation of monthly sales is $1500 — sales vary moderately around the average.

Measures of Spread. Interquartile range

- **Meaning: Middle 50% spread (Q3 − Q1)**
- **Use when: Resistant to outliers; non-normal data**
- **Not for: Nominal**
- **Good use: IQR of salaries, delivery times**
- **Bad use: IQR of product names**

Measures of Spread. Interquartile range

- **Daily Steps:** 5000, 6000, 7000, 8000, 9000

  IQR: Yes/No?

- **Exam Scores:** 55, 60, 65, 90, 95

  IQR: Yes/No?

- **Survey Ratings:** 2, 3, 3, 4, 5

  IQR: Yes/No?

- **Phone Numbers**: 12345, 23456, 34567

  IQR: Yes/No?

Measures of Spread. Interquartile range

How to interpret

- Describes where the bulk of values lie, ignoring extremes.

Example insight

- The IQR of exam scores is 20 — most students scored within a 20-point range.

Warnings

- Doesn’t show the full range of variability
- May not reflect multimodal distributions

Measures of Spread (Dispersion). Practical cases

Measures of Shape. Skewness / Kurtosis

- **Meaning:** Shape of distribution — asymmetry and tailedness
- **Use when:** You want to assess normality or detect outliers
- **Only for:** Quantitative data
- **Good use:** Distribution of investment returns
- **Bad use:** Shape of nominal variables (e.g., brand names)

Measures of Shape. Skewness

How to interpret

- Positive skew: long tail to the right
- Negative skew: long tail to the left
- Skew ≈ 0: fairly symmetric

Example insight

- Income data shows strong positive skew — a few individuals earn much more than the rest.

Warnings

- Sensitive to outliers
- Not useful on very small samples
- Skewed data may affect mean-based statistics

Measures of Shape. Kurtosis

How to interpret

- High kurtosis: heavy tails, more outliers
- Low kurtosis: light tails, fewer outliers
- Normal distribution has kurtosis ≈ 3 (excess kurtosis = 0)

Example insight

- Sales data shows high kurtosis — frequent extreme changes month to month

Warnings

- Often misunderstood as "peakness" (but it's about tails)
- Easily distorted by a few outliers
- Use with other statistics for a full picture

Measures of Shape. Z-scores

- **Meaning:** How far a point is from the mean in std units
- **Use when:** You need to compare across variables or detect outliers
- **Only for:** Quantitative
- **Good use:** Compare student scores across tests with different scales
- **Bad use:** Z-score of phone brands

Measures of Shape. Z-scores

How to interpret

- Tells how unusual a value is in the context of the dataset

Example insight

- A z-score of 2.1 for this month's sales means sales were significantly higher than usual

Warnings

- Assumes a normal (or roughly symmetric) distribution
- Not meaningful for categorical or skewed data
- Outliers will have very large/small z-scores

Measures of Shape. Practical cases

What Can Summary Statistics Tell Us?

Data Distribution

- Is the data symmetric, skewed, or uniform?
- Are there outliers or extreme values?

Data Quality

- How much missing data is there?
- Are there unexpected values (e.g., negative values in a positive-only dataset)?

Insights for Modeling

- Do we need to normalize or scale the data?
- Should we handle outliers or missing values before modeling?

Business Insights

- What are the typical values for key metrics?
- How much variability exists in the data?

Practical cases

Useful Links

[Exploratory Data Analysis with Pandas]

[Mastering Exploratory Data Analysis (EDA): A Comprehensive Python (Pandas) Guide for Data Insights and Storytelling]