Elements of Machine Learning and Data Science

Part I: Data Science — Exam Notes (Living Document)

Emir Pisirici

January 29, 2026

Exam likelihood: High (overall Data Science part)

This document is structured to match the lecture topics exactly and is designed for adding **exam-style notes**, **common traps**, and **visual summaries**.

Contents

1 Introduction to Data Science

1.1 Introduction

1.2 Tabular Data

1.3 Data Science Process

Exam likelihood: High

Framework questions are easy to grade and strongly test "big picture" understanding.

Examiner favorite (what they love to ask)

Typical asks: **ETL vs ELT**, **CRISP-DM phases**, and mapping a scenario to the correct phase. Also: where data leakage/bias lives (data understanding + evaluation).

1.3.1 ETL vs ELT (Definitions + Differences)

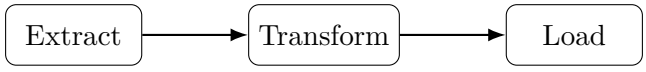
Cheat sheet / must-memorize

ETL: Extract → Transform → Load (transform before target).
ELT: Extract → Load → Transform (transform inside target platform).
Key contrast: where transformations happen; governance vs flexibility; raw history availability.

Common pitfall

People confuse "ELT = no cleaning". Wrong. It means cleaning happens *after loading*, often in warehouse/lakehouse layers (staging → curated).

Visual



ETL

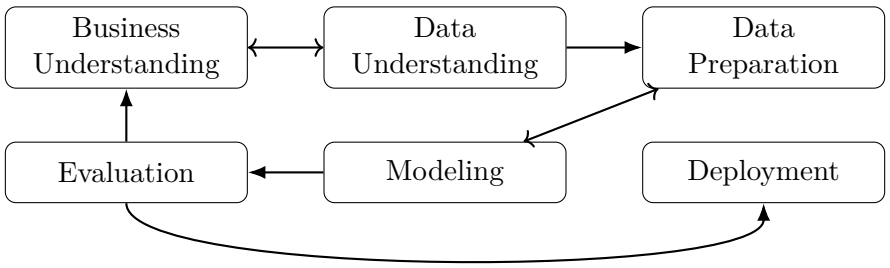
ELT

1.3.2 CRISP-DM

Cheat sheet / must-memorize

CRISP-DM: Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment (iterative loops).

Visual



1.3.3 PDCA

Cheat sheet / must-memorize

PDCA: Plan → Do → Check → Act (continuous improvement loop).

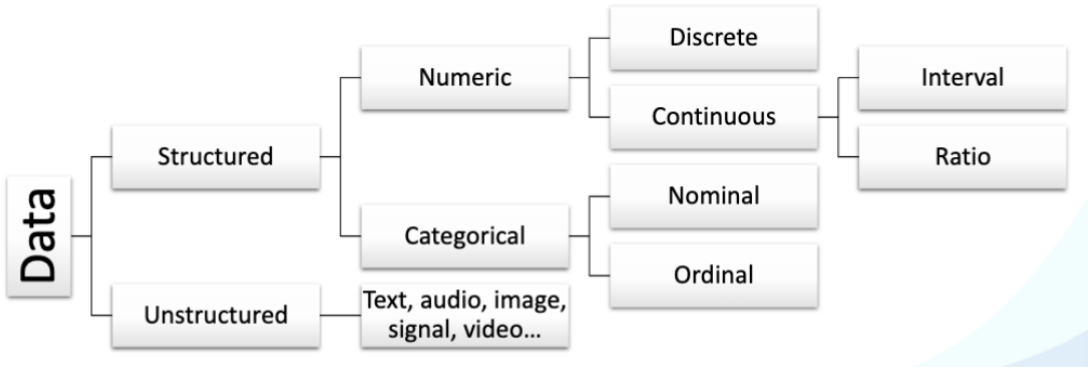
1.3.4 DMAIC

Cheat sheet / must-memorize

DMAIC: Define → Measure → Analyze → Improve → Control. Often used for process/quality improvement + monitoring and part of the Six Sigma methodology.

1.4 Data Types

Visual



1.5 Descriptive Statistics

Exam likelihood: High

Frequent: compute variance/STD/covariance/correlation by hand; read a correlation matrix.

Examiner favorite (what they love to ask)

Explain why covariance depends on units, and why correlation is normalized in $[-1, 1]$.

Why (motivation): Quantify spread and association between variables.
What (definition): Variance/STD measure spread; covariance/correlation measure linear association.
How (procedure/usage): Compute formulas, then interpret sign/magnitude and check the correlation matrix.

Cheat sheet / must-memorize

- **Variance (sample):** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Std dev:** $s = \sqrt{s^2}$
- **Covariance (sample):** $\text{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- **Correlation:** $r = \frac{\text{cov}(X,Y)}{s_X s_Y}$ (unitless, -1 to 1)
- **Correlation matrix:** table of pairwise correlations; symmetric with 1s on the diagonal.

Common pitfall

Correlation $\neq$ causation; a strong correlation can be driven by a confounder or Simpson's paradox.

Visual

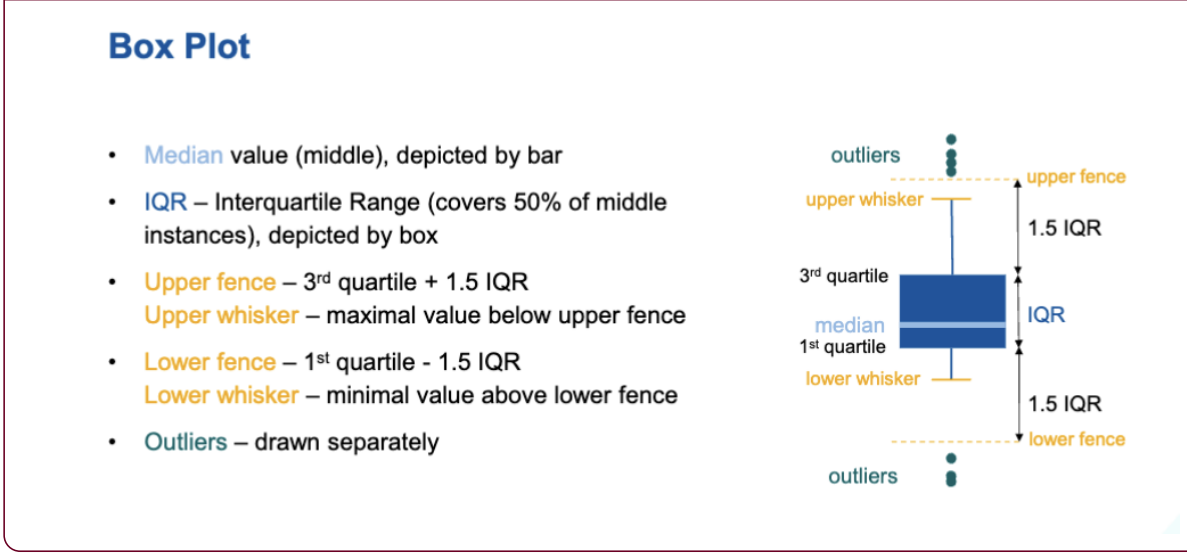
1	r_{12}	r_{13}
r_{21}	1	r_{23}
r_{31}	r_{32}	1

Correlation matrix

Key takeaways: Know formulas + interpretations; correlation matrix is symmetric with 1s on the diagonal.

1.6 Basic Visualizations

Visual



1.7 Feature Transformations

Exam likelihood: High

Typical: pick the right transform (scale, log, encode) and explain why.

Examiner favorite (what they love to ask)

Identify data leakage in preprocessing; name the correct order for train/test transformations.

Why (motivation): Turn raw categorical/continuous variables into model-ready features.
What (definition): Encoding or discretizing features without changing the target meaning.
How (procedure/usage): Choose encoding by category type; choose binning by distribution.

Cheat sheet / must-memorize

- **One-hot encoding:** create a 0/1 column per category (nominal).
- **Binary encoding:** represent categories as binary digits (compact one-hot).
- **Ordinal encoding:** map ordered categories to ranks (only if order is real).
- **Binning:** convert continuous to categories.
- **Equal-width binning:** fixed interval sizes across the range.
- **Equal-frequency binning:** same number of samples per bin.

Common pitfall

Fitting transforms on the full dataset (leakage). Always fit on training data, then apply to validation/test.

Key takeaways: Use one-hot for nominal, ordinal for ordered labels, and binning for simplification.

1.8 "How to lie with statistics"

2 Decision Trees

2.1 Introduction to Decision Trees

2.2 Entropy and Information Gain

Exam likelihood: Very High

Almost guaranteed: compute entropy / information gain on a small dataset.

2.3 ID3 Algorithm

2.4 Pruning

2.5 Continuous Data (Threshold splits)

2.6 Ensembles (Bagging/Random Forest/Boosting)

3 Clustering

3.1 Introduction to Unsupervised Learning

3.2 Introduction to Clustering

3.3 Similarity and Dissimilarity

3.4 K-means and K-medoids

3.5 Agglomerative Clustering

3.6 DBSCAN

3.7 Closing

4 Frequent Itemsets

4.1 Introduction

4.2 Properties of Frequent Itemsets

4.3 Apriori Algorithm

4.4 FP-Growth Algorithm

5 Association Rules

5.1 Introduction

5.2 Generating Association Rules

5.3 Evaluation (support, confidence, lift, conviction)

5.4 Applications

5.5 Simpson's Paradox

6 Time Series

6.1 Temporal Data

6.2 Introduction to Time Series

6.3 Analysis

6.4 Forecasting