Data Science and Data Analytics (WS 2025/26)

International Business Management (B. A.)

© Benjamin Gross

September 3, 2025

This document provides the course material for Data Science and Data Analytics (B. A. – International Business Management). Upon successful completion of the course, students will be able to: recognize important technological and methodological advancements in data science and distinguish between descriptive, predictive, and prescriptive analytics; demonstrate proficiency in classifying data and variables, collecting and managing data, and conducting comprehensive data evaluations; utilize R for effective data manipulation, cleaning, visualization, outlier detection, and dimensionality reduction; conduct sophisticated data exploration and mining techniques (including PCA, Factor Analysis, and Regression Analysis) to discover underlying patterns and inform decision-making; analyze and interpret causal relationships in data using regression analysis; evaluate and organize the implementation of a data analysis project in a business environment; and communicate the results and effects of a data analysis project in a structured way.

Table of contents

1 Scope and Nature of Data Science

xy

1.1 Defining Data Science as an Academic Discipline

Data science draws from and interacts with multiple foundational disciplines: * Informatics / Information Systems * Computer Science (algorithms, data structures, systems design) * Mathematics (linear algebra, calculus, optimization) * Statistics & Econometrics (inference, modeling, causal analysis) * Social Science & Behavioral Sciences (contextual interpretation, experimental design)

1.2 Significance of Business Data Analysis for Decision-Making

- Supports evidence-based strategic, tactical, and operational decisions.
- Reduces uncertainty in forecasting, pricing, resource allocation, and risk management.
- Enables performance measurement and continuous improvement.
- Facilitates customer understanding, personalization, and retention strategies.

1.3 Emerging Trends

Key technological and methodological developments shaping the data landscape: * Evolution of computing and data processing architectures. * Digitalization of processes and platforms. * Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). * Big Data ecosystems (volume, velocity, variety, veracity, value). * Internet of Things (IoT) and sensor-driven data generation. * Cloud computing and elastic infrastructure. * Blockchain for distributed trust and data integrity. * Industry 4.0: cyber-physical systems and automation. * Remote and hybrid working environments: collaboration, distributed analytics, governance.

1.4 Types of Analytics

- Descriptive Analytics: What happened?
- Predictive Analytics: What is likely to happen?
- Prescriptive Analytics: What should we do?

2 Data Analytic Competencies

xy

2.1 Types of Data

- Cross-sectional data
- Panel (longitudinal) data
- Time-series data
- Geo-referenced / spatial data
- (Potentially) streaming / real-time data

2.2 Types of Variables

- Continuous (interval/ratio)
- Count
- Ordinal
- Categorical (nominal / binary)
- (Possibly) compositional or hierarchical structures

2.3 Conceptual Framework: Knowledge & Understanding of Data

- Clarify analytical purpose and domain context.
- Define entities, observational units, and identifiers.
- Align business concepts with data structures.

2.4 Data Collection

- Identify internal and external sources.
- Acquire via APIs, databases, surveys, sensors, or third-party vendors.
- Assess provenance, licensing, and ethical considerations.

2.5 Data Management

- Organize: schema design, naming conventions.
- Clean: resolve duplicates, inconsistencies, missingness.
- Convert: type casting, normalization, encoding.
- Curate: maintain lineage, documentation, metadata.
- Preserve: backups, versioning, retention policies.

2.6 Data Evaluation

- Plan analyses aligned with objectives and stakeholders.
- Conduct exploratory, inferential, and predictive procedures appropriately.
- Evaluate robustness, reliability, and validity.
- Assess limitations, bias, and ethical impact.

3 Applications in the Programming Language R

xy

3.1 Core tidyverse Tooling

Explore fundamental packages: * `dplyr` for data manipulation (filter, mutate, summarise, joins). * `tidyr` for data reshaping (pivoting, nesting, separating, unnesting). * `ggplot2` for layered grammar-based visualization. * (Optionally) `readr`, `purrr`, `stringr`, `forcats` for ingestion, functional iteration, text, and factor handling.

3.2 Data Visualization Principles

- Choose encodings appropriate to variable types.
- Emphasize clarity: reduce chart junk; apply perceptual best practices.
- Support comparison, trend detection, and anomaly spotting.

3.3 Detecting Outliers and Anomalies

- Rule-based methods (IQR, z-scores).
- Robust statistics (median, MAD).
- Model-based or multivariate detection (e.g., Mahalanobis distance, clustering residuals).
- Distinguish errors vs. novel but valid observations.

3.4 Dimensionality Reduction

- Motivation: mitigate multicollinearity, noise, and curse of dimensionality.
- Techniques: Principal Component Analysis (PCA), Factor Analysis, (optionally) t-SNE / UMAP (for exploration).
- Interpretability vs. compression trade-offs.

3.5 Data Exploration and Mining

- Structured EDA workflow: question $\rightarrow$ visualize $\rightarrow$ quantify $\rightarrow$ refine.
- PCA for variance structure.
- Factor Analysis for latent constructs.
- Regression Analysis for relationships and predictive structure.
- Clustering (k-means, hierarchical) for pattern discovery (if included).

3.6 Causal Inference with Regression Analysis

- Distinguish association vs. causation.
- Model specification and confounding control.
- Assumptions: linearity, independence, homoskedasticity, exogeneity.
- Interpretation of coefficients and marginal effects.
- Sensitivity and robustness checks.

4 Literature

References