Data Science

Short notes for B.Tech CSE / IT · BCA · MCA · M.Tech

Uncodemy Handwritten Notes · free for students

I. Introduction

- **Data Science** = getting useful knowledge and decisions from data using statistics, programming and domain knowledge.
- It mixes three areas: maths / statistics, computer science, and domain expertise.

➤ Data science life cycle

1. Define the business problem
2. Collect data
3. Clean and prepare data
4. Explore data (EDA)
5. Build models
6. Evaluate models
7. Deploy and communicate results

- A popular process model is **CRISP-DM**: business understanding, data understanding, data preparation, modelling, evaluation, deployment.

➤ Related fields

Field	Focus
Data analytics	Answer questions from existing data
Data science	Analytics + prediction + building data products
Machine learning	Algorithms that learn from data
Big data	Storing and processing very large data
Business intelligence	Reports and dashboards on past data

➤ Types of analytics

- **Descriptive**: what happened? **Diagnostic**: why did it happen?
- **Predictive**: what will happen? **Prescriptive**: what should we do?

2. Data and Data Types

- **Structured**: rows and columns (SQL tables). **Semi-structured**: JSON, XML. **Unstructured**: text, images, audio.

➤ Levels of measurement

Level	Property	Example
Nominal	Names only, no order	Gender, city
Ordinal	Order but unequal gaps	Rating: low / medium / high
Interval	Equal gaps, no true zero	Temperature in °C
Ratio	Equal gaps + true zero	Height, salary, age

- **Qualitative** (categorical) vs **quantitative** (numeric: discrete or continuous).
- Data sources: databases, files (CSV, Excel), APIs, web scraping, sensors, surveys.

3. Descriptive Statistics

➤ Central tendency

- **Mean** = sum / count. **Median** = middle value of sorted data. **Mode** = most frequent value.

EXAMPLE Data 2, 4, 4, 5, 7, 8. Mean = 30 / 6 = 5. Median = (4 + 5) / 2 = 4.5. Mode = 4.

➤ Spread

FORMULA Population variance $\sigma^2 = \Sigma(x - \mu)^2 / N$. Sample variance $s^2 = \Sigma(x - \bar{x})^2 / (n - 1)$. SD = square root of variance.

EXAMPLE Data 2, 4, 6. Mean = 4. Squared deviations 4, 0, 4, sum = 8. Population variance = 8/3 = 2.67. Sample variance = 8/2 = 4, sample SD = 2.

- **Range** = max − min. **IQR** = Q3 − Q1 (middle 50%).
- **Outlier rule**: below Q1 − 1.5·IQR or above Q3 + 1.5·IQR.

Shape

- **Skewness**: right (positive) skew → mean > median. Left (negative) skew → mean < median.
- **Kurtosis**: how heavy the tails are.

REMEMBER Median is robust to outliers; mean is not. Use median for skewed data like incomes.

4. Probability and Inferential Statistics

- **Probability** P(A) = favourable / total outcomes, between 0 and 1.
- P(A or B) = P(A) + P(B) − P(A and B). Independent events: P(A and B) = P(A)·P(B).
- Bayes: P(A|B) = P(B|A)·P(A) / P(B).

Distributions

Distribution	Type	Use
Bernoulli	Discrete	One yes / no trial
Binomial	Discrete	k successes in n trials
Poisson	Discrete	Count of events in a fixed interval
Uniform	Both	All outcomes equally likely
Normal	Continuous	Bell curve, many natural measures

FORMULA Binomial P(X = k) = nCk · p^k · (1 − p)^(n − k). Mean = np, variance = np(1 − p).

EXAMPLE A fair coin tossed 4 times, P(exactly 2 heads) = 4C2 × 0.5^2 × 0.5^2 = 6 × 0.0625 = 0.375.

- **Normal distribution**: 68–95–99.7 rule (within 1, 2 and 3 SD of the mean).

FORMULA z-score z = (x − μ) / σ.

- **Central Limit Theorem**: the mean of large samples is approximately normal, whatever the population shape.

Hypothesis testing

1. State **H0** (no effect) and **H1** (effect).
2. Choose significance level α (often 0.05).
3. Compute the test statistic and **p-value**.
4. If p-value < α, reject H0.

- **Type I error**: reject a true H0 (false positive). **Type II error**: fail to reject a false H0 (false negative).
- Tests: z-test, t-test (means), chi-square test (categories), ANOVA (3 or more means).

Correlation

- **Pearson r** from −1 to +1. r = 0 means no linear relation.
- Correlation does not mean causation.

5. Data Wrangling

- **Data cleaning**: handle missing values, remove duplicates, fix types, fix inconsistent labels, handle outliers.
- **Missing data**: drop, fill with mean / median / mode, forward fill, or model-based imputation.
- **Transformation**: scaling, log transform for skewed data, binning, encoding categories.
- **Data integration**: join data from many sources (inner, left, right, outer join).

Pandas essentials

```
import pandas as pd
df = pd.read_csv("sales.csv")
df.head(); df.info(); df.describe()
df = df.drop_duplicates()
df["price"] = df["price"].fillna(df["price"].median())
df[df["city"] == "Pune"]
df.groupby("city")["sales"].sum()
df.sort_values("sales", ascending=False)
```

NumPy essentials

- `np.array`, `arr.shape`, `arr.reshape()`, `np.mean`, `np.std`, slicing, broadcasting.
- NumPy arrays are faster than Python lists for maths because operations are vectorised.

6. Exploratory Data Analysis and Visualisation

- **EDA** = understanding data with summaries and plots before modelling.
- **Univariate**: one variable. **Bivariate**: two. **Multivariate**: many.

Chart	Best for
Histogram	Distribution of one numeric variable
Box plot	Median, IQR, outliers
Bar chart	Comparing categories
Line chart	Trend over time
Scatter plot	Relation between two numeric variables
Heatmap	Correlation matrix
Pie chart	Parts of a whole (few categories)

- Tools: Matplotlib, Seaborn, Plotly, Tableau, Power BI.

```
import matplotlib.pyplot as plt
plt.hist(df["age"], bins=10)
plt.xlabel("Age"); plt.ylabel("Count")
plt.show()
```

TIP Always label axes and add a title. A plot without labels loses marks in practicals.

7. Modelling Overview

- **Supervised**: regression (linear regression, random forest regressor), classification (logistic regression, decision tree, kNN, SVM).
- **Unsupervised**: clustering (k-means, hierarchical), dimensionality reduction (PCA).
- **Train-test split** (e.g. 80 / 20) and **cross-validation** to check generalisation.
- Metrics: accuracy, precision, recall, F1 (classification); MAE, RMSE, R^2 (regression).

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
Xtr, Xte, ytr, yte = train_test_split(X, y, test_size=0.2)
model = LinearRegression().fit(Xtr, ytr)
print(model.score(Xte, yte))   # R squared
```

FORMULA $R^2 = 1 - (SS_res / SS_tot)$. $R^2 = 1$ means a perfect fit.

8. Big Data and Tools

- **5 V's of big data**: Volume, Velocity, Variety, Veracity, Value.
- **Hadoop**: HDFS (distributed storage with replicated blocks) + MapReduce (batch processing) + YARN (resource manager).
- **Spark**: in-memory processing, much faster than MapReduce for repeated jobs. Has Spark SQL and MLlib.
- **NoSQL** types: key-value (Redis), document (MongoDB), column-family (Cassandra), graph (Neo4j).
- **SQL** is still core: SELECT, WHERE, GROUP BY, HAVING, JOIN, ORDER BY.

9. Ethics and Communication

- Data privacy, consent, anonymisation of personal data.
- **Bias** in data can lead to unfair models. Check data balance across groups.
- Tell a clear story: problem → method → result → recommendation.
- Reproducibility: keep code, versions and data steps documented (Jupyter notebooks, Git).

10. Most Asked Questions (practice)

1. Define data science. Explain the data science life cycle.
2. Descriptive vs predictive vs prescriptive analytics.
3. Explain nominal, ordinal, interval and ratio data.
4. Find mean, median, mode, variance and SD for given data.
5. Explain normal distribution and the Central Limit Theorem.
6. Explain hypothesis testing, p-value, Type I and Type II errors.
7. How do you handle missing values and outliers?
8. What is EDA? Which plots are used and when?
9. Explain Hadoop components. Hadoop vs Spark.
10. Write Pandas code to load, clean and group a dataset.

TIP Statistics numericals and Pandas / Matplotlib code snippets carry easy marks. Learn the life cycle and the 5 V's word for word.