Here's a **summary** of your lecture note **"Sprint 1 – Introductory Data Analysis" (Prof. Dr. Matteo Marouf)**:

🧠 Summary: Sprint 1 – Introductory Data Analysis

1. Course Overview

- Focus: Understanding **data analysis using Python**.
- Tools: `pandas`, `numpy`, `matplotlib`, `seaborn`, `scikit-learn`.
- Structure: 12 learning blocks from Python basics → statistical inference → storytelling with data.
- Requirements:
  - 80% attendance in exercises for exam eligibility.
  - Rehearsal exam required.
  - No AI assistance in assignments; collaboration is encouraged.

---

2. Data Analysis Process

Goal: Convert raw data → meaningful insights.

Stages:

1. **Collect Data** (sampling)
2. **Summarize Data** (descriptive statistics)
3. **Interpret and Analyze** (inferential statistics)

---

3. Statistical Investigation Cycle

1. **Problem Understanding** – define research question.
2. **Produce Data** – collect a representative sample.
3. **Exploratory Data Analysis (EDA)** – summarize, visualize.
4. **Modeling & Analysis** – apply statistical tools.
5. **Inference** – generalize findings to population.
6. **Communication** – storytelling and recommendations.

---

4. Data Representation

- Convert various data (text, image, signal) into **numerical form** for analysis.
  Examples:
  - **Text:** TF-IDF, Word2Vec.
  - **Image:** pixel intensity values or spectrograms.
  - **Graph:** Node2Vec embeddings.

---

5. Data Types

Type	Description	Examples
Qualitative (Categorical)	Non-numeric categories	Gender, Color
Quantitative (Numeric)	Measurable values	Age, Salary
Structured	Tables, databases	Customer records
Unstructured	Free text, media	Tweets, Videos
Semi-structured	JSON, XML	Web logs
Signal Data	Sequential numeric data	ECG, audio waveform

6. Sampling Concepts

- **Population**: Entire group of interest.
- **Sample**: Subset studied due to practical limits.
- **Good Sampling Practices:**
  - Clear definition
  - Representativeness
  - Random selection
  - Adequate sample size
  - Bias avoidance (selection/nonresponse bias)

7. Descriptive Statistics

Measures of Center:

Measure	Description	Use Case
Mean	Average value	Normal data
Median	Middle value	Skewed data
Mode	Most frequent	Categorical data

Measures of Variability:

- **Range**: max - min
- **Variance (s^2)** and **Standard Deviation (s)**: how spread out the data is.

Percentiles:

- Indicate relative standing (e.g., 85th percentile = better than 85% of observations).

Five-number summary:

- **Min, Q1, Median, Q3, Max** → forms **Boxplot** to show spread and outliers.

8. Measures of Shape

Measure	Meaning	Interpretation
Skewness	Asymmetry	Right-skew → long right tail
Kurtosis	Tail heaviness	High → frequent extreme values

9. Visualizations

Type	Purpose
Histogram	Distribution of numeric data
Bar Chart / Pie Chart	Categorical distribution
Box Plot	Spread, outliers
Scatter Plot	Relationship between two variables
Time Chart	Change over time

10. Common Statistical Pitfalls

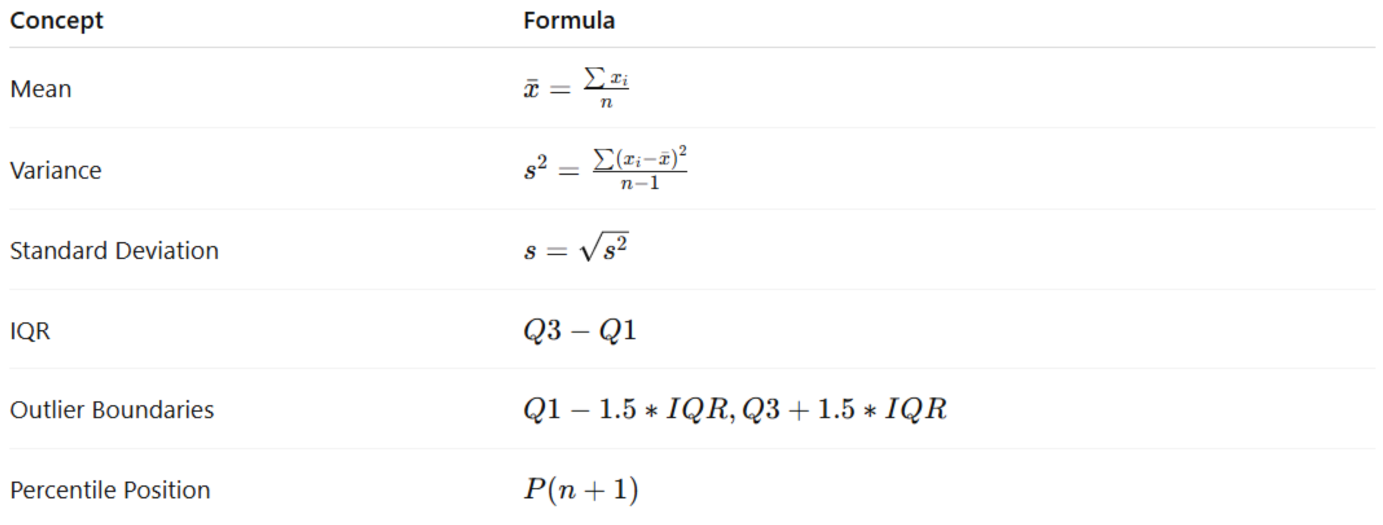
- **Simpson’s Paradox** – Aggregated data can mislead.
- Always check **subgroups** and **denominators**.

11. Communication & Storytelling

- Combine **data**, **visualization**, and **narrative**.
- Outcome: clear insights for decision-making.

🧾 Exam Cheat Sheet

Key Formulas



Concept	Formula
Mean	$\bar{x} = \frac{\sum x_i}{n}$
Variance	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Standard Deviation	$s = \sqrt{s^2}$
IQR	$Q3 - Q1$
Outlier Boundaries	$Q1 - 1.5 * IQR, Q3 + 1.5 * IQR$
Percentile Position	$P(n+1)$

Data Analysis Steps

1. Define question
2. Collect data (sampling)
3. Clean & visualize
4. Summarize (mean, median, mode, SD)

5. Analyze (correlation, distributions)
6. Infer and communicate

Visualization Shortcuts (Python)

```
# Histogram
df['age'].hist()

# Boxplot
sns.boxplot(x='sex', y='age', data=df)

# Scatterplot
sns.scatterplot(x='age', y='fare', data=df)

# Correlation heatmap
sns.heatmap(df.corr(), annot=True)
```

🧾 DATA ANALYSIS EXAM CHEAT SHEET

🔹 1. DATA ANALYSIS PROCESS

Goal: Turn raw data → meaningful insights.

Steps:

1. Define problem
2. Collect sample data
3. Clean and summarize
4. Visualize patterns
5. Analyze (statistical tools)
6. Draw conclusions & communicate

🔹 2. TYPES OF DATA

Type	Description	Example
Qualitative (Categorical)	Names or labels	Gender, Color
Quantitative (Numeric)	Measurable values	Age, Salary
Structured	Tabular format	SQL tables
Unstructured	Free text/media	Tweets, Images
Semi-structured	JSON/XML	Web logs
Signal Data	Sequential numeric	Audio, ECG

🔹 3. SAMPLING CONCEPTS

- **Population:** All individuals of interest.
- **Sample:** Subset used for analysis.
- **Good Sampling:** Random, representative, adequate size, unbiased.
- **Common Biases:** Selection bias, nonresponse bias.

🔹 4. DESCRIPTIVE STATISTICS

Measures of Center

Measure	Description	Best For
Mean	Arithmetic average	Symmetrical data
Median	Middle value	Skewed data
Mode	Most frequent value	Categorical data

Measures of Spread

Measure	Formula	Meaning
Range	max – min	Total spread
Variance (s^2)	$\frac{\sum(x_i-\bar{x})^2}{n-1}$	Avg. squared deviation
Std. Dev. (s)	$\sqrt{s^2}$	Typical distance from mean
IQR	Q3 – Q1	Central 50% spread

Outliers:

- Lower bound = Q1 - 1.5×IQR
- Upper bound = Q3 + 1.5×IQR

🔹 5. SHAPE OF DISTRIBUTION

Measure	Meaning	Interpretation
Skewness	Asymmetry	Right-skew = long right tail
Kurtosis	Tail heaviness	High = more extreme values

🔹 6. VISUALIZATION TYPES

Chart	Purpose
Histogram	Show distribution of numeric data
Bar Chart / Pie Chart	Compare categories
Box Plot	Show median, quartiles, outliers
Scatter Plot	Relationship between two variables
Line / Time Plot	Trends over time
Heatmap	Correlation matrix visualization

🔹 7. COMMON STATISTICAL FALLACIES

- **Simpson's Paradox:** Trends change when groups are combined.
  👉 Always check subgroup data.
- **Misleading Averages:** Mean distorted by outliers.
- **Ignoring Variability:** Always pair mean with SD/IQR.

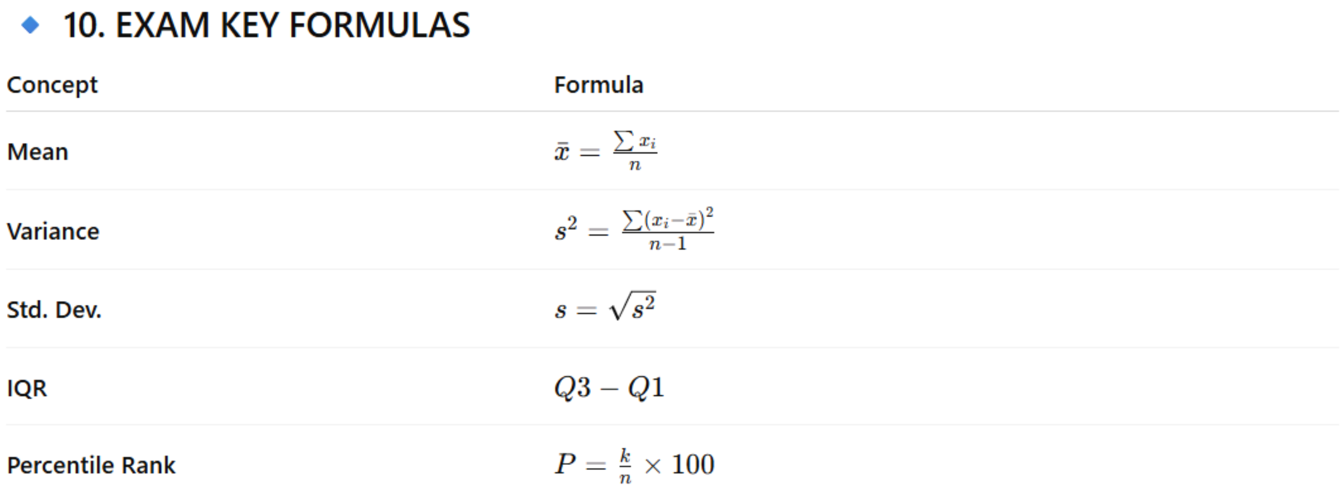
🔹 8. PYTHON COMMANDS (Pandas & Seaborn)

Task	Command
Import pandas	`import pandas as pd`
Load dataset	`df = pd.read_csv('file.csv')`
View first rows	`df.head()`

Shape (rows, cols)	`df.shape`
Missing values	`df.isnull().sum()`
Summary stats	`df.describe()`
Select rows	`df.loc[0]` / `df.iloc[0]`
Filter	`df[df['age'] < 18]`
Group mean	`df.groupby('sex')['age'].mean()`
New column	`df['fare_per_age'] = df['fare']/df['age']`
Correlation	`df.corr()`

◆ 9. VISUALIZATION COMMANDS

Plot	Python Command
Histogram	`df['age'].hist()`
Box Plot	`sns.boxplot(x='sex', y='age', data=df)`
Scatter Plot	`sns.scatterplot(x='age', y='fare', data=df)`
Bar Plot	`sns.barplot(x='sex', y='survived', data=df)`
Pie Chart	`df['survived'].value_counts().plot.pie()`
Heatmap	`sns.heatmap(df.corr(), annot=True, cmap='coolwarm')`



◆ 10. EXAM KEY FORMULAS

Concept	Formula
Mean	$\bar{x} = \frac{\sum x_i}{n}$
Variance	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
Std. Dev.	$s = \sqrt{s^2}$
IQR	$Q3 - Q1$
Percentile Rank	$P = \frac{k}{n} \times 100$

◆ 11. DATA STORYTELLING TIPS

- Use **simple visuals** (avoid clutter).
- Combine **data + visuals + narrative**.
- Always answer:
  “What does this mean and why does it matter?”