Prof. Eren Bilen **DATA 101 – Study Guide for Exam I**

Overview:
The first exam focuses on understanding data, interpreting relationships, and reasoning about evidence. Topics combine conceptual reasoning with general data exploration using Python. *The list below is for guidance only and may not cover all material on the exam.*

Data and Variables

Define what data represent and what a dataset contains.
Identify the unit of observation (what each row corresponds to).
Distinguish between categorical and numerical variables.
Distinguish between discrete and continuous numerical data.
Recognize derived or computed variables (e.g., rates, ratios, averages).
Identify which type of variable fits a given visualization.
Understand why variable type matters for analysis and plotting.
Keywords: dataset, observation, variable, categorical, numerical, discrete, continuous

---

Descriptive Visualizations

Match visualization types to data types: bar chart (categorical), histogram (numerical), scatterplot (two quantitative).
Describe the direction, strength, and shape of a relationship in a scatterplot.
Identify skewness and outliers in histograms.
Add a categorical dimension to scatterplots using color or shape.
Interpret what visual patterns imply about relationships.
Recognize misleading graphs and explain why they mislead.
Keywords: bar chart, histogram, scatterplot, trend, skewness, outlier, pattern

---

Data Sources and Sampling

Identify data sources.
Distinguish between population and sample.
Recognize sampling bias and measurement bias.
Explain how sampling affects generalizability.
Understand why random sampling matters.
Keywords: population, sample, sampling bias, measurement bias, representativeness

---

Experiments and Causality

Distinguish between experimental and observational data.
Define treatment and control groups.
Explain why random assignment enables causal inference.
Identify threats to validity such as selection bias or confounding.
Define outcome variables and describe how they can be measured.
Explain why voluntary participation or choosing treatment introduces bias.
Keywords: experiment, treatment, control, randomization, bias, validity, confounder

---

Difference-in-Differences (DiD)

Compute changes within treatment and control groups.
Calculate and interpret a DiD estimate conceptually.
Explain the parallel-trends assumption.
Interpret the meaning and direction of an effect.
Keywords: before–after, treatment–control, parallel trends, policy effect

---

Regression Discontinuity (RD)

Understand what a cutoff or threshold represents.
Explain the logic of comparing observations just above and below the cutoff.
Interpret what the discontinuity represents in causal terms.
Distinguish RD from randomized experiments and DiD.
Keywords: cutoff, threshold, continuity, local comparison

Instrumental Variables (IV)

Understand when and why an instrument is used in causal analysis.
Define what makes a variable a valid instrument (relevance and exogeneity).
Explain how an instrument isolates variation in the treatment that is as-if random.
Keywords: instrument, exogeneity

Python and DataFrames

Understand how to inspect, summarize, and filter data in a DataFrame.
Use `pandas` to read, explore, and manipulate tabular data.
Write commands to view rows, select columns, compute averages, check data types, and filter based on a condition.
General syntax examples:
`df.head()` → preview the first few rows
`df.shape` → check dataset dimensions
`df.describe()` → view summary statistics
`df.dtypes` → inspect column types
`df["col"]` or `df.col` → select a specific column
`df["col"].mean()` → compute a column's average
`df[df["col"] == "value"]` → subset data matching a condition
`df.sort_values("col")` → order rows by a variable
Distinguish between selecting columns and filtering rows.
Recognize how DataFrames represent tabular data (rows = observations, columns = variables).
Keywords: pandas, DataFrame, column, row, filtering, selecting, sorting, summary, describe

---

Data Interpretation and Ethics

Recognize that data reflect both measurement and human decisions.
Explain how framing, sampling, or language can introduce bias.

Identify one insight and one limitation from a dataset or visualization.
Evaluate what a graph truly communicates versus what it implies.
Keywords: interpretation, bias, limitation, measurement, context