RULES AND REGULATIONS FOR THE WRITTEN EXAMS

AT THE FACULTY OF ENGINEERING AND NATURAL SCIENCES

Please read this document carefully, sign the bottom part, and include this form with your solution sheet.

- The exam duration (start-end times) will be written on your exam sheet or will be announced, you should adhere to these times.
- You cannot leave the examination room in the **first 30 minutes** of the exam.
- You are **NOT allowed** to take the exam once a student has left one of the exam rooms.
- Talking with another student during the exam for any reason (asking for an explanation, for a calculator or eraser or any other object) is **NOT allowed**.
- If you **GIVE** to or **RECEIVE** from a student any item (eraser, calculator, pen, etc.) during the exam, this will be taken as a sign of cheating.
- **TURN OFF** your **CELLPHONES AND SMART DEVICES**; and PLACE THEM **away from you** such that those **CANNOT BE SEEN** from outside.
- **HAVING a CELLPHONE or a SMART DEVICE** that is VISIBLE FROM OUTSIDE during the exam is **NOT ALLOWED**. Interacting with a cellphone in any way (EVEN IF IT IS TURNED OFF) will be taken as a sign of cheating.

In case of a sign of cheating, or if you do not obey the above rules, there will be a disciplinary action taken towards you based on the STUDENT DISCIPLINARY REGULATIONS of YÖK Items 2.5, 2.7, and 2.8.

I understand the above-mentioned rules and regulations, and I hereby accept them, and any consequences following my actions during the exam.

Student Full Name:

Student Number:

Course Code and Title:

Signature:

Date:

INTRODUCTION TO DATA SCIENCE

KEYS TO MIDTERM EXAM

(This is a 90-minute exam)

STUDENT ID:	FULLNAME:	SIGNATURE:

Q.1:	Q.2:	Q.3:	Q.4:	TOTAL:

Q.1) (20) Define and provide examples for each of the following types of attributes: Nominal, Ordinal, Interval, and Ratio.

Nominal data categorizes items into different groups with no inherent order or ranking. Examples are blood type (A, B, AB, O) and Marital status (e.g., single, married, divorced).

Ordinal data puts the categories in a specific order. However, the difference between the values cannot be determined precisely. Examples are Shirt size (small, medium, large, extra-large) and Education level (high school, bachelor's degree, master's degree, Ph.D.).

Interval data has all the properties of ordinal data and adds the benefit of having equal intervals between each value. However, the zero point is arbitrary and doesn't necessarily represent an absence of the quantity being measured. Examples are Temperature measured in Celsius or Fahrenheit (e.g., 20°C, 30°C, 40°C) and Calendar dates (e.g., January 1st, February 15th, December 31st).

Ratio attributes are numerical variables that have a true zero point, where zero indicates the absence of the attribute being measured. Ratios between values are meaningful and consistent. Examples are Height (e.g., 171 cm), Weight (e.g., 85 kg), and Time (e.g., 42 minutes).

Q.2) (20) Which data types can be used to represent both the order of information (like a sequence of events) and the relationships between different pieces of data (like connections between webpages)?

Graphs are mathematical structures composed of nodes (vertices) and edges (connections) that represent relationships between pairs of nodes. They can be directed (edges have a specific direction) or undirected (edges have no direction). Graphs are versatile data structures that can represent various types of relationships, including sequential relationships and interconnected relationships. A directed graph can be used to represent a sequence of events where each node represents an event, and directed edges indicate the order of occurrence between events.

For example, in a timeline of historical events, each event can be a node, and the edges indicate the chronological order of events.

Sequences are ordered collections of elements where the order of elements is significant. Arrays and lists are common data structures used to represent sequences in computer science. An array or list can be used to represent a sequence of events where each element represents an event, and the position of the element in the array or list indicates its order in the sequence. For example, an array of timestamps representing the occurrence of events in chronological order. A list or array of hyperlinks can be used to represent connections between webpages where each element represents a hyperlink, and the order of hyperlinks reflects the navigational path or sequence of connections between webpages like history in any Web browser.

Q.3) (15) (a) Define data quality problems and provide examples of common issues encountered in real-world datasets.

Data quality problems refer to issues or anomalies in datasets that impact their reliability, accuracy, consistency, completeness, and overall usefulness for analysis or decision-making purposes. These problems can arise from various sources such as data collection processes, storage, transformation, or entry errors. Some common data quality problems:

Missing Values:

In a dataset containing customer information, some entries might have missing values for the "Phone Number" field.

Incorrect Values: A dataset of student ages might contain an entry with an age of 200, which is clearly incorrect and needs to be corrected.

Duplicate Records: A dataset of online transactions might have multiple identical entries for the same purchase.

Outliers: In a dataset of employee salaries, there might be a few entries with salaries significantly higher or lower than the rest of the data, which could be outliers.

Data Integrity Issues: In a relational database, foreign key constraints might not be enforced properly, leading to referential integrity issues where references to non-existent entities exist.

Data Consistency Problems: In a dataset of customer addresses, inconsistencies might arise due to variations in the representation of the same address (e.g., "123 Main St." vs. "123 Main Street").

Bias and Inaccuracy: In datasets used for machine learning models, bias might be introduced due to underrepresentation or misrepresentation of certain demographic groups, leading to inaccurate predictions or decisions.

(b) (15) Explain the significance of data quality in ensuring the reliability, accuracy, and usefulness of data for decision-making.

Reliability refers to the consistency and dependability of data. Reliable data can be trusted to be consistent and stable over time, regardless of changes in circumstances or conditions. Reliable data ensures that decision-makers can have confidence in the information they are using. It reduces the risk of making decisions based on inaccurate or inconsistent data, which could lead to unreliable outcomes or consequences.

Accuracy refers to the correctness and precision of data. Accurate data is free from errors, bias, or distortion and reflects the true values or characteristics of the entities it represents. Accurate data is essential for making informed decisions. Decision-makers rely on accurate data to understand the current situation, assess risks, identify opportunities, and forecast future outcomes. Inaccurate data can lead to flawed analyses, misguided decisions, and potential financial or reputational losses.

Usefulness refers to the relevance and applicability of data for achieving specific objectives or addressing particular needs. Useful data provides actionable insights and supports decision-making processes effectively. Useful data empowers decision-makers to derive meaningful insights, formulate strategies, and take appropriate actions. It enables organizations to optimize resource allocation, improve operational efficiency, enhance customer satisfaction, and gain competitive advantages in the marketplace.

Q.4) (30) Consider the following dataset stored in a Pandas DataFrame:

```
import pandas as pd

data = {'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Emily'],
        'Age': [25, 30, 35, 40, 45],
        'City': ['New York', 'Los Angeles', 'Chicago', 'Houston', 'Boston']}

df = pd.DataFrame(data)
```

a) (10) Write a Pandas code snippet to display the first 3 rows of the DataFrame.

```
print( df.head(3) )
```

b) (10) Write a Pandas code snippet to calculate the average age of the individuals in the DataFrame.

```
average_age = df['Age'].mean()
print("Average Age:", average_age)
```

c) (10) Write a Pandas code snippet to filter the DataFrame to only include individuals aged 35 or older.

```
filtered_df = df[ df[df['Age'] >= 35] ]
print(filtered_df)
```