INTRODUCTION TO DATA SCIENCE

KEYS TO MOCKUP MIDTERM EXAM

(This is a 90-minute exam)

STUDENT ID:	FULLNAME:	SIGNATURE:

Q.1) a) Explain the difference between integer-based indexing and label-based indexing in pandas DataFrames.

They differ in how they reference the position of the data. iloc relies on the position of the data within the DataFrame. It uses zero-based counting. It works well when the DataFrame index consists of sequential numbers, but may be less intuitive when custom labels are assigned to rows or columns. Loc uses the actual labels assigned to rows and columns. The labels are used to refer directly to specific datapoints. This makes your code easier to read, and it maintains consistency even if the order of the DataFrame is shuffled around.

b) What does the following code print to the console?

```
import pandas as pd

data = {'Age': [25, 30, 22, 38],
        'City': ['New York', 'Los Angeles', 'Chicago', 'Miami']}
df = pd.DataFrame(data, index=['Alice', 'Bob', 'Charlie', 'David'])

x = df.iloc[1, :]
print(x)
```

```
Age         30
City    Los Angeles
Name: Bob, dtype: object
```

```
y = df.loc['Bob', 'Age']
print(y)
```

```
30
```

```
u = df.iloc[:, :2]
v = df.iloc[-2:]
print(u)
```

```
         Age      City
Alice    25      New York
Bob      30      Los Angeles
Charlie  22      Chicago
David    38      Miami
```

```
print(v)
```

```
         Age   City
Charlie  22    Chicago
David    38    Miami
```

```
z = df.loc[['Alice', 'Charlie'], :]
print(z)
```

```
         Age City
Alice    25  New York
Charlie  22  Chicago
```

INTRODUCTION TO DATA SCIENCE

KEYS TO MOCKUP MIDTERM EXAM

(This is a 90-minute exam)

Q.2) Briefly explain the key steps involved in the data science workflow, from problem definition to model deployment.

The data science workflow is a structured process for handling data science projects.

Problem Definition: This is the identification of the business question or problem that you'll be solving with data science. Clearly defining the goals and desired outcomes is important.

Data Acquisition: Once you have a problem, you need data. This may involve multiple sources, databases, or APIs.

Data Cleaning and Preparation: Raw data is rarely perfect. This stage involves cleaning the data by addressing missing values, inconsistencies, and errors. You might also need to transform the data into a format suitable for analysis.

Exploratory Data Analysis: Here, you get familiar with the data. You use various techniques to understand patterns, trends, and relationships within the data. This helps identify potential features for modeling.

Model Building: This is where you create a model based on the prepared data. You choose appropriate algorithms and train the model on a portion of the data.

Model Evaluation: After training, you assess the model's performance on unseen data. This helps determine if the model is effective at solving the problem you defined.

Deployment: If the model performs well, you might deploy it into production. This involves integrating the model into an application or system where it can be used to make predictions or generate insights.

Q.3) Data types like temperature (Celsius or Fahrenheit) and time (seconds, minutes) are often used for measurement. Can you classify these as both interval and ratio data types? Explain why or why not.

The difference between 4°C and 8°C is the same as the difference between 12°C and 16°C (4°C), 0°C (or 0°F) doesn't represent absolute zero temperature. You can't say that 16°C is "four times hotter" than 4°C, because zero is not a true absence of temperature. Therefore, temperature data is classified as interval data.

The difference between one minute and two minutes is the same as the difference between two minutes and three minutes. zero seconds (or minutes) represents a true absence of time. There is no such thing as "negative time". Time data is classified as ratio data because both the difference in time intervals and the zero point have meaning.