Data Analytics Midterm Exam

Q1) You are hired by a retail company that sells both online and offline. They provide you with the following data sources:

- An SQL database containing customer demographics
- CSV files with monthly sales transactions
- A REST API providing real-time product inventory
- Social media posts scraped from public Twitter accounts about the brand

Design a complete end-to-end data analytics pipeline for this scenario. Your answer must include:

1. Identification and justification of the type of each data source (structured / semi-structured / unstructured; internal vs external; streaming vs batch).
2. A detailed data ingestion plan, specifying tools/technologies (any are allowed).
3. A full preprocessing workflow for each data source.
4. A unified data model (diagram or explanation) showing how data from all sources will integrate.
5. A discussion of data quality risks and how you would mitigate them.

Q2) The numeric dataset **Dataset_for_Q2.csv** contains the following attributes: Age, Income, Spending Score, Number of Purchases, Loyalty Level (categorical). The raw dataset includes noise, skewed distributions, and missing values. You must prepare the dataset for cluster analysis. For each of the following preprocessing steps, choose *one specific method*, justify it, and explain why the alternatives are inferior for this dataset:

1. Missing value treatment
2. Outlier handling
3. Normalization / scaling
4. Encoding of the categorical attribute

Q3) The dataset **Dataset_for_Q3.csv** contains real-world data quality issues commonly found in IoT environments, including inconsistent timestamp formats, missing intervals, mixed temperature units, corrupted metadata, duplicate sensor readings, and isolated anomalies. Your task is to transform the dataset into a consistent and reliable time-series suitable for analysis.

1. Identify all inconsistent data formats and correct them.
2. Inspect the time sequence and list at least five hourly timestamps that are missing from the expected timeline.
3. Identify mixed units and provide them in a consistent format.
4. Explain why mixing temperature units can produce false anomaly signals in downstream analysis.
5. Identify one Fahrenheit value from the dataset that could be mistakenly interpreted as an extreme outlier if not converted.
6. By inspecting the temperature values, identify three single-timestamp spikes that do not align with normal sensor behavior.

7. Propose two different methods for handling such anomalies and explain when each method is appropriate.
8. For each spike you identified, state whether you would keep, correct, or remove the value, and justify your decision in the context of sensor monitoring.
9. The *StatusText* column contains corrupted labels (e.g., inconsistent spelling, symbols, accented characters). Create a mapping table that standardizes all corrupted entries into correct categories (e.g., "OK", "Running", "Active").
10. Briefly explain one potential risk of applying incorrect fuzzy text matching in an industrial monitoring or safety system.
11. Identify at least two duplicate airflow readings that appear across different timestamps.
12. Decide whether each repeated value is likely a true repeated measurement or an accidental duplication, and justify your reasoning based on airflow behavior in systems.

Write a short evaluation (5–7 sentences) describing:

- How the identified data-quality issues could lead to misleading analytical results
- Which issues you consider most harmful for time-series modeling
- Your recommendations for improving future data collection and validation procedures

Q4) Compare the following data sources for a national health-care analytics project:

1. Government open-data portals
2. Hospital EHR (Electronic Health Record) systems
3. Patient-reported survey data
4. Wearable device sensor streams (daily step count, heart rate, etc.)

For each data source:

1. Evaluate accuracy, completeness, timeliness, consistency, and potential bias.
2. Identify one major risk that could affect data reliability.
3. Propose a preprocessing technique that specifically mitigates that risk (e.g., truth cross-validation, sampling correction, deduplication, time alignment, noise filtering).

Then, choose one analytics task (e.g., predicting disease risk) and explain how weaknesses in these data sources could distort model performance—even after preprocessing.