Data Warehousing and Data Mining

Assignment - 02 – CLO – 02

You are tasked with preparing a **centralized data warehouse** for a **smart city traffic management system** that collects data from over **15,000 IoT sensors** deployed across highways, intersections, parking zones, and public transport hubs. These sensors include **traffic cameras**, **induction loops**, **smart parking meters**, and **air quality monitors**. The data arrives in **heterogeneous formats and intervals**, often with **misaligned timestamps**. The dataset is **messy**, containing **missing values**, **sensor glitches causing extreme outliers**, **redundant measurements from overlapping sensors**, and **subtle inconsistencies**.

Your task is to **clean, integrate, transform, and visualize the data** so city planners can:

- Detect congestion patterns
- Optimize parking space usage
- Monitor air quality
- Plan infrastructure improvements

You must apply **critical thinking** at each step: some extreme values are **true traffic/pollution events**, while others are sensor errors. Decisions made in early cleaning steps affect later transformations, correlations, and visualizations.

Sensor_ID	Type	Timestamp	Vehicle_Count	Avg_Speed_kmh	PM2.5	Parking_ Available
T1	Camera	2025-10-15 08:00	45	NA	NA	NA
T2	Loop	2025-10-15 08:00	NA	65	NA	NA
T3	Air_Quality	2025-10-15 08:00	NA	NA	78	NA
T4	Parking	2025-10-15 08:00	NA	NA	NA	12
T1	Camera	2025-10-15 08:01	48	NA	NA	NA
T2	Loop	2025-10-15 08:01	NA	500	NA	NA
T3	Air_Quality	2025-10-15 08:01	NA	NA	80	NA
T4	Parking	2025-10-15 08:01	NA	NA	NA	11
T1	Camera	2025-10-15 08:02	50	NA	NA	NA
T2	Loop	2025-10-15 08:02	NA	62	NA	NA

Part 1 – Identify Data Issues (3 Marks)

1. Identify missing values, extreme outliers, and redundancies in the dataset.
2. Explain why each issue is a problem for data warehouse analytics.

Part 2 – Handling Missing Values (2 Marks)

1. Propose an appropriate imputation method for each column with missing values.
2. Fill in missing values using your chosen method and justify your reasoning.

Part 3 – Noise Handling & Outlier Treatment (3 Marks)

1. Detect extreme outliers and explain their impact on analysis.
2. Apply binning (equal-frequency, 3 bins) to smooth Avg_Speed_kmh readings.

3. Discuss how smoothing affects analysis.

Part 4 – Redundancy & Correlation Analysis (3 Marks)

1. Calculate the correlation between Vehicle_Count and Avg_Speed_kmh after cleaning.
2. Decide if one attribute is redundant and justify your decision.

Part 5 – Data Transformation & Discretization (3 Marks)

1. Apply min-max normalization to Vehicle_Count and Avg_Speed_kmh.
2. Apply equal-width discretization with 3 bins.
3. Explain why these transformations are critical for analytics.

Part 6 – Visualization & Interpretation (6 Marks)

Using Python:

1. Draw a histogram of Vehicle_Count before and after noise handling.
2. Draw a boxplot for Avg_Speed_kmh to highlight outliers.
3. Draw a scatter plot of Vehicle_Count vs Avg_Speed_kmh to visualize correlation.
4. Write 3–4 sentences interpreting the plots, discussing the effects of noise handling, binning, normalization, and correlation.

Deliverables

- **Part 6:** Submit a **Jupyter Notebook (.ipynb file)** with proper code, comments, and visualizations.
- **Part 1-5:** Submit a **handwritten report (scanned PDF or hard copy)**.
    - No typed answers, no AI-generated content, and no plagiarism.
    - **Zero marks** if AI tools or copy-paste work is detected.

Rubrics:

Part	**Marks**
Part 1 – Identify Issues	3
Part 2 – Missing Values	2
Part 3 – Noise & Binning	3
Part 4 – Correlation & Redundancy	3
Part 5 – Normalization & Discretization	3
Part 6 – Visualization & Interpretation	6
Total	20

Part	Deliverable	Marks	Rubrics / Criteria
06	Jupyter Notebook (Practical Implementation)	3	Correct code (1), Proper output (1), Clarity & comments (1)
1-5	Handwritten Assignment (Concepts & Explanation)	17	Accuracy (10), Depth of explanation (4), Neatness (3)
Total		**20**	-
Penalty	Late Submission	-2	Applied for each late day

Submission Rules

- **Deadline: 25 October 2025**
- **Late Submission Penalty:** -2 marks per day
- **Plagiarism / AI Detection: 0 marks** (strictly enforced)
- **Total Marks:** 20