EXPLORATORY DATA ANALYSIS ON A DATASET

Objective:

The main goal of this assignment is to conduct a thorough exploratory analysis of the "cardiographic.csv" dataset to uncover insights, identify patterns, and understand the dataset's underlying structure. You will use statistical summaries, visualizations, and data manipulation techniques to explore the dataset comprehensively.

Dataset:

1. **LB** - Likely stands for "Baseline Fetal Heart Rate (FHR)" which represents the average fetal heart rate over a period.
2. **AC** - Could represent "Accelerations" in the FHR. Accelerations are usually a sign of fetal well-being.
3. **FM** - May indicate "Fetal Movements" detected by the monitor.
4. **UC** - Likely denotes "Uterine Contractions", which can impact the FHR pattern.
5. **DL** - Could stand for "Decelerations Late" with respect to uterine contractions, which can be a sign of fetal distress.
6. **DS** - May represent "Decelerations Short" or decelerations of brief duration.
7. **DP** - Could indicate "Decelerations Prolonged", or long-lasting decelerations.
8. **ASTV** - Might refer to "Percentage of Time with Abnormal Short Term Variability" in the FHR.
9. **MSTV** - Likely stands for "Mean Value of Short Term Variability" in the FHR.
10. **ALTV** - Could represent "Percentage of Time with Abnormal Long Term Variability" in the FHR.
11. **MLTV** - Might indicate "Mean Value of Long Term Variability" in the FHR.

Tools and Libraries:

- Python or R programming language

- Data manipulation libraries
- Data visualization libraries (Matplotlib and Seaborn in Python)
- Jupyter Notebook for documenting your analysis

Tasks:

1. **Data Cleaning and Preparation:**
    - Load the dataset into a DataFrame or equivalent data structure.
    - Handle missing values appropriately (e.g., imputation, deletion).
    - Identify and correct any inconsistencies in data types (e.g., numerical values stored as strings).
    - Detect and treat outliers if necessary.

Answer : **1. Data Cleaning and Preparation**

The dataset was loaded into a Pandas DataFrame. A preliminary inspection was conducted to check the shape, data types, and presence of duplicates.

- **Missing values:** The dataset contained very few missing values. Where missingness was minimal, rows were dropped; otherwise, imputation was carried out using the median (for skewed numerical variables) or the mode (for categorical variables).
- **Data types:** All features that were mistakenly stored as strings were converted to numeric, ensuring consistency.
- **Outliers:** Outlier detection was performed using the Interquartile Range (IQR) method. Several variables, such as Fetal Movements (FM) and Uterine Contractions (UC), displayed outliers. These may represent genuine physiological extremes rather than errors, so they were flagged for further domain-specific review rather than removed automatically.

Overall, the dataset was cleaned, structured, and made suitable for statistical analysis and visualization.

2. **Statistical Summary:**
    - Provide a statistical summary for each variable in the dataset, including measures of central tendency (mean, median) and dispersion (standard deviation, interquartile range).

- Highlight any interesting findings from this summary.

Answer :

For each variable, descriptive statistics were computed, including measures of central tendency (mean, median) and measures of dispersion (standard deviation, interquartile range).

- **Baseline FHR (LB):** Showed a relatively stable mean and median, indicating consistency across samples.
- **Fetal Movements (FM) and Uterine Contractions (UC):** Both exhibited high skewness and large standard deviations, suggesting most values are clustered at the lower end but with some extreme high values.
- **Variability metrics (ASTV, MSTV, ALTV, MLTV):** These measures revealed notable variation in fetal heart rate patterns. High interquartile ranges in ASTV and ALTV suggest irregularities in short- and long-term variability for subsets of patients.

The summary provided a quantitative foundation, highlighting variables with potential clinical importance due to variability or skewness.

3. **Data Visualization:**
    - Create histograms or boxplots to visualize the distributions of various numerical variables.
    - Use bar charts or pie charts to display the frequency of categories for categorical variables.
    - Generate scatter plots or correlation heatmaps to explore relationships between pairs of variables.
    - Employ advanced visualization techniques like pair plots, or violin plots for deeper insights.

Answer:

Visual exploration provided deeper insights into distributions and relationships:

- **Histograms & Boxplots:** Most physiological measures displayed non-normal distributions, with FM and UC showing clear right skewness. Boxplots confirmed the presence of outliers.
- **Correlation Heatmap:** Strong correlations were observed between measures of variability (e.g., ASTV and ALTV). This suggests these metrics may capture overlapping aspects of fetal heart rate irregularities.
- **Scatter Plots:** Relationships between uterine activity (UC) and deceleration patterns (DL, DS, DP) indicated possible clinical associations worth investigating.

- **Advanced plots:** Violin plots revealed distribution density with quartiles, while pair plots helped identify clusters and multicollinearity among certain features.

Visualizations helped validate statistical findings and provided intuitive understanding of data structure and anomalies.

4. **Pattern Recognition and Insights:**
    - Identify any correlations between variables and discuss their potential implications.
    - Look for trends or patterns over time if temporal data is available.

Answer :

From the analysis:

- **Correlations:** Variables related to heart rate variability (ASTV, MSTV, ALTV, MLTV) were moderately correlated, suggesting redundancy. This should be considered in feature selection for modeling.
- **Physiological events:** High fetal movements often coincided with accelerations (AC), reflecting normal physiological responses. Conversely, prolonged decelerations (DP) aligned more with higher uterine contraction counts, which may indicate fetal stress.
- **Outliers:** The extreme values in FM and UC highlight that while most pregnancies exhibit stable readings, a small subset of patients experience significantly higher activity, potentially signaling risk conditions.

These insights can guide medical decision-making, such as focusing on combinations of contractions and decelerations for risk monitoring.

5. **Conclusion:**
    - Summarize the key insights and patterns discovered through your exploratory analysis.
    - Discuss how these findings could impact decision-making or further analyses.

Answer:

The exploratory analysis of the Cardiotocographic dataset provided the following key findings:

1. The dataset was clean, with minimal missingness, though several features displayed significant outliers.
2. Statistical summaries revealed skewed distributions in FM and UC, indicating variability across patients.

3. Visualization confirmed these findings and exposed strong relationships among heart rate variability measures.
4. Correlation analysis suggested redundancy among some variables, and highlighted clinically relevant interactions such as UC with DP.

Implications: These findings support better feature engineering for predictive modeling in fetal monitoring, particularly focusing on contractions and deceleration events. They also underscore the need to consult medical expertise to interpret whether outliers are clinically meaningful or artifacts.

Title: Exploratory Data Analysis — *Cardiotocographic Dataset*

Objective:
Perform a detailed exploratory data analysis (EDA) to understand fetal heart monitoring variables, detect anomalies, and surface relationships that inform downstream modeling or clinical interpretation.

Data & Variables (summary):
Variables include baseline fetal heart rate (LB), accelerations (AC), fetal movements (FM), uterine contractions (UC), decelerations (DL, DS, DP), and variability measures (ASTV, MSTV, ALTV, MLTV). These metrics collectively describe fetal heart behavior and uterine activity.

Data cleaning & preparation:

- Loaded data into a pandas DataFrame.
- Converted string-encoded numerics to numeric types; inspected and removed exact duplicates.
- Handled missing values: rows with very small missingness were dropped; columns with more missingness were imputed by median (numerical) or mode (categorical).
- Outliers were detected using the IQR rule and *flagged* (not automatically deleted), since extreme values could be clinically meaningful.

Statistical summary (key metrics):

- Computed count, mean, median, std, min, max, IQR, skewness, kurtosis for all numeric features.
- Observations: LB (baseline FHR) is relatively stable; FM and UC show strong right-skew and high IQRs; variability measures (ASTV, ALTV) show moderate dispersion, indicating heterogeneity across observations.

Visualizations & findings:

- **Histograms & boxplots** confirm skewness and outliers for FM and UC.
- **Correlation heatmap** shows moderate-to-strong correlations among variability measures (ASTV, MSTV, ALTV, MLTV), suggesting partial redundancy.
- **Scatter plots** indicate associations between contractions (UC) and prolonged decelerations (DP), which may signal periods of fetal stress.
- **Pairwise plots / violin plots** show distribution shapes and confirm multimodal or heavy-tailed distributions in some features.

Pattern recognition & insights:

- Variability metrics cluster together and may be combined or reduced (PCA, feature selection) to avoid multicollinearity in predictive models.

- High fetal movement often coincides with accelerations — a physiologically normal pattern.
- A minority of records show unusually high UC or DP values — these cases should be prioritized for clinical review.

Recommendations:

1. Retain flagged outliers for domain expert review before removal — they could be clinically important.
2. Consider transformations (log or sqrt) for skewed variables before modeling.
3. Use feature selection or dimensionality reduction for correlated variability metrics (ASTV, MSTV, ALTV, MLTV).
4. If building predictive models (e.g., fetal distress detection), include interaction terms for UC × DP and test their predictive power.
5. Collect more metadata if available (maternal age, gestational week) to control for confounders.

Concluding remark:
This EDA reveals that while most fetal heart metrics are stable, a subset of observations shows extreme values and correlated variability features that warrant targeted clinical attention and careful feature engineering for predictive tasks.