Exploratory Data Analysis REPORT

**1. Findings and Insights**

**1. Dataset Overview**

* Columns in our dataset include LB, AC, FM, UC, DL, DS, DP, ASTV, MSTV, ALTV, MLTV, Width, Tendency, and NSP.
* There are continuous variables (e.g., LB, UC) and potentially categorical ones (e.g., NSP and Tendency).

**2. Data Quality Insights**

* Negative values were observed in some columns (**AC, FM, UC, ALTV, MLTV**) that are unlikely in the context of fetal measurements. We have replaced these with lower bounds (e.g., 0)
* These could result from data entry errors or measurement issues.
* Columns like NSP and Tendency (initially continuous) were recategorized into discrete classes (1, 2, 3) for analysis.

**3. Distribution and Trends**

* Many features exhibited skewed distributions and outliers, as expected for biological data.
* Histograms showed some multimodal distributions for several columns, indicating possible clustering or subgroup behavior in the dataset.
* **NSP** and **Tendency** displayed clear peaks, supporting their categorical or ordinal nature.

**4. Relationships Between Variables**

* Correlation analysis: Weak or no correlation between others, even for LB and UC.
* Bar charts analysis: High DS (Short Decelerations) and high FM (Fetal Movements) contributing to NSP = 3 (pathological) likely indicate fetal distress or overactivity due to stress.

**5. Key Findings**

* DL and DP: Almost no prolonged decelerations for Class 1 and 2. High DP and DL are more frequent in pathological cases.
* Negative or neutral tendencies in the **Tendency** column align with standardization or scoring methods. While peaks were evident, further analysis could clarify their meaning.

**2. Recommendations**

**1. Outlier Handling:** Confirm the outlier values and its significance with an expert to get a better understanding of how to deal with other outliers in the dataset

**2. Pattern Recognition:**

- Investigate clustering algorithms (e.g., k-means or hierarchical clustering) to identify potential subgroups in the dataset.

- Consider classification models to predict \*\*NSP\*\* or other key outcomes based on feature relationships.

**3. Domain Expert Validation:**

- Collaborate with cardiotocography experts to validate assumptions about data ranges, tendencies, and significance of identified patterns.

**4. Further Feature Engineering:** Derive additional features, such as variability ratios or normalized metrics, to enhance analysis.