Dataset Analysis Report

Uploaded Dataset

August 13, 2025

Introduction

This report provides an initial analysis of a dataset designed for assessing fetal health, likely derived from Cardiotocography (CTG) measurements. The dataset encompasses a range of physiological parameters and histogram-derived features. The primary objective of this preliminary review is to understand the dataset's structural characteristics, data types, completeness regarding missing values, and to gain initial insights into the distributions of its various features.

Methods

The analysis was conducted by thoroughly examining the provided dataset summary. This involved reviewing the dataset's overall shape (number of rows and columns), identifying all column names, inspecting the assigned data types for each feature, checking for the presence of missing values across all columns, and analyzing the basic descriptive statistics (count, mean, standard deviation, min, quartiles, and max) for all numerical features. This systematic approach allowed for a comprehensive, high-level understanding of the data's integrity and preliminary statistical properties.

Executive Summary

- The dataset consists of 2106 observations and 22 distinct features.
- All features in the dataset are uniformly represented as `float64` data types.
- A significant finding is the complete absence of missing values across all 22 columns, indicating a highly complete dataset.
- Several features exhibit no variance and are constant across all observations: `severe_decelerations`, `prolongued_decelerations`, and `histogram_number_of_zeroes` all have a mean and standard deviation of 0, implying their values are uniformly zero.
- Critically, the `fetal_health` target variable also appears constant, with a mean, min, max, and all quartiles at 1.0, and a standard deviation of 0. This suggests that all records in this specific dataset summary are labeled as '1.0' (presumably 'healthy'), which is a major limitation for any classification task.
- Other features, such as `baseline value` (mean ~133 bpm), `accelerations` (mean ~0.003), `abnormal_short_term_variability` (mean ~47), and various histogram-derived parameters, show meaningful distributions and variability, suggesting their potential utility if the target variable issue is resolved.

Data Cleaning Notes

Based on the provided dataset summary:

* **Missing Values:** No missing values were detected in any of the columns. The dataset appears entirely complete, eliminating the need for imputation or row removal due to missing data.

* **Data Types:** All columns are consistently `float64`, which is an appropriate data type for numerical analysis and direct input into most machine learning algorithms. No type conversions are required.

* **Constant Features:** Three features (`severe_decelerations`, `prolongued_decelerations`, `histogram_number_of_zeroes`) exhibit zero variance, meaning they hold the same value (0.0) for every entry. These features offer no discriminatory information for modeling and are strong candidates for removal.

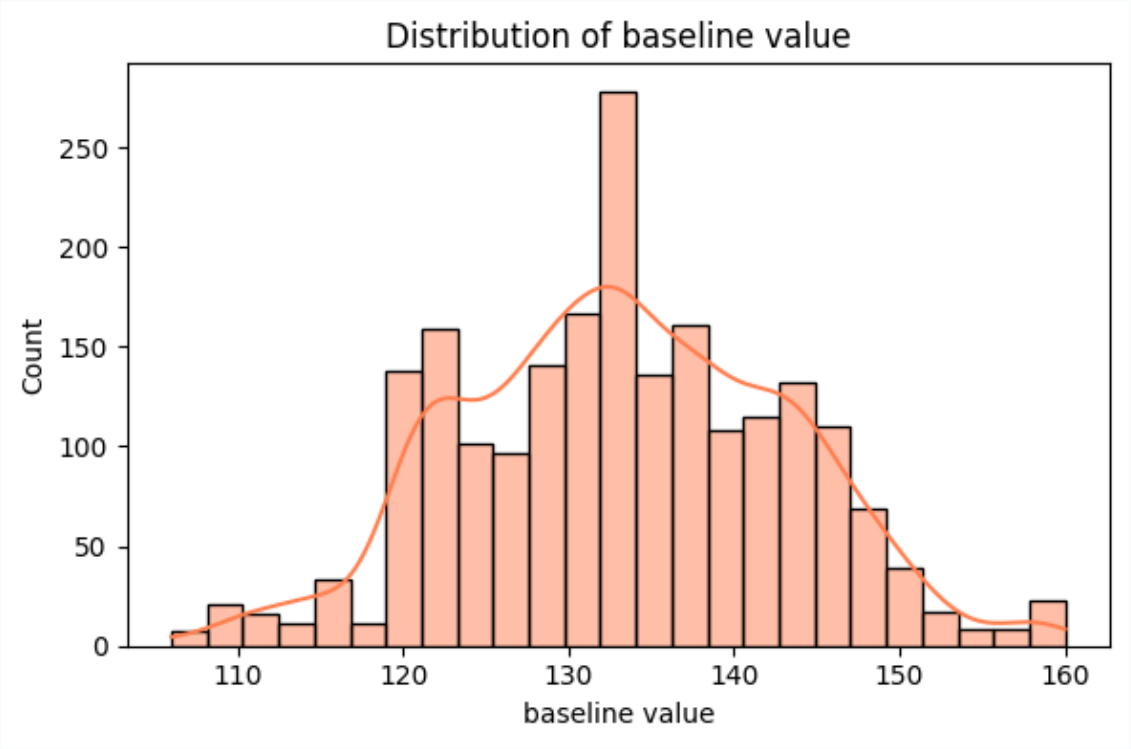
* **Target Variable Uniformity:** The `fetal_health` column has zero variance, with all observed values being 1.0. This indicates that, within the scope of this summary, the dataset contains only observations categorized under a single `fetal_health` outcome. If the goal is to build a predictive model to classify different fetal health states, this dataset subset is fundamentally unsuitable as it lacks the necessary variability in the target variable.

Statistical Summary

Variable	Unweighted_mean	Unweighted_s
baseline value	133.3	0.21
accelerations	0.0	0.0
fetal_movement	0.0	0.0
uterine_contractions	0.0	0.0
light_dec	0.0	0.0
severe_decelerations	0.0	0.0
prolongued_decelerations	0.0	0.0
abnormal_short_term_variability	46.94	0.37
mean_v_o_s_t_variability	1.3	0.02
percentage_of_time_with_abnormal_long_term_variability	6.59	0.22
mean_value_of_long_term_variability	7.99	0.11
histogram_width	70.65	0.85
histogram_min	93.5	0.64
his_max	163.96	0.38
histogram_number_of_peaks	4.07	0.06
histogram_number_of_zeroes	0.0	0.0
histogram_mode	137.92	0.32
histogram_mean	134.79	0.33
histogram_median	138.2	0.31
histogram_variance	15.68	0.4
histogram_tendency	0.32	0.01
fetal_health	1.0	0.0

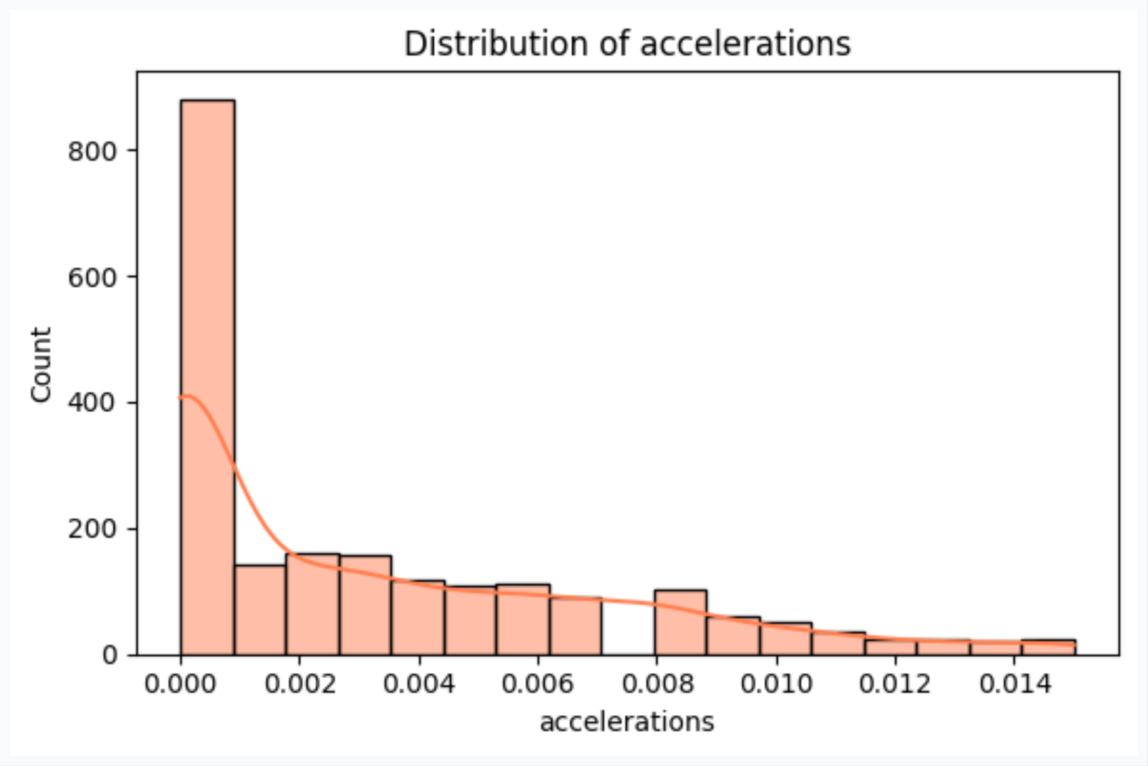
Results & Insights

Distribution: baseline value



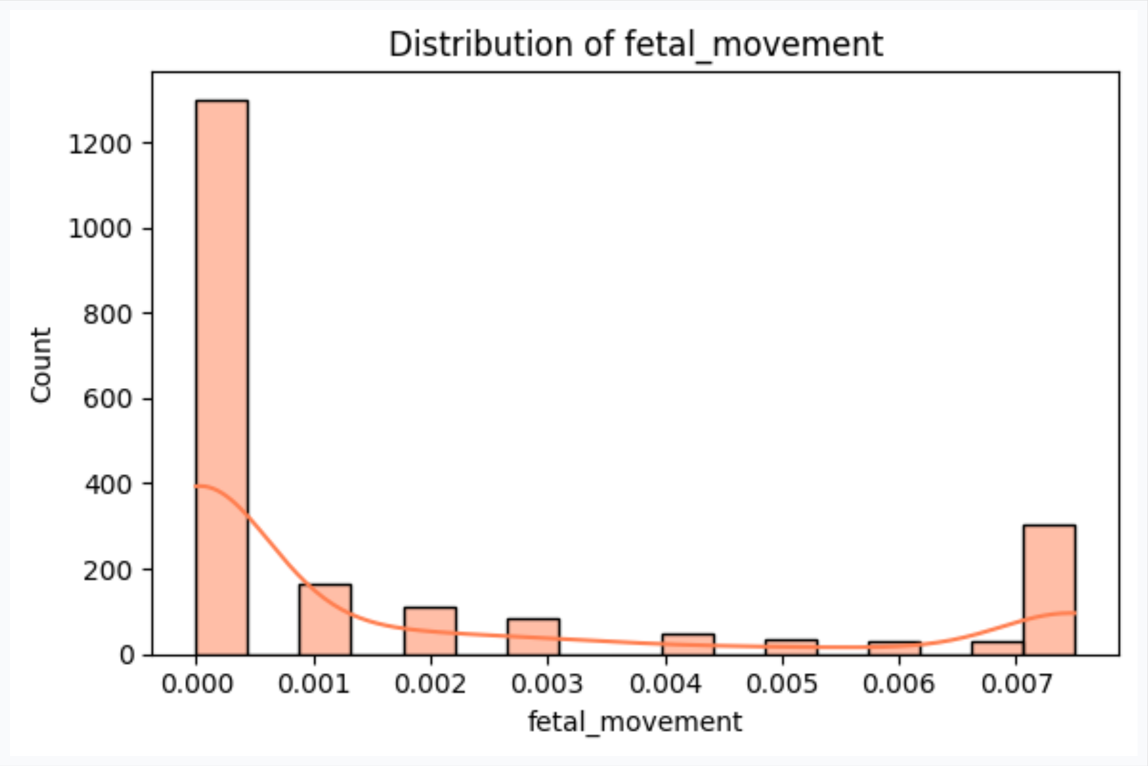
The average baseline value is 133.30 with a spread of 9.84.

Distribution: accelerations



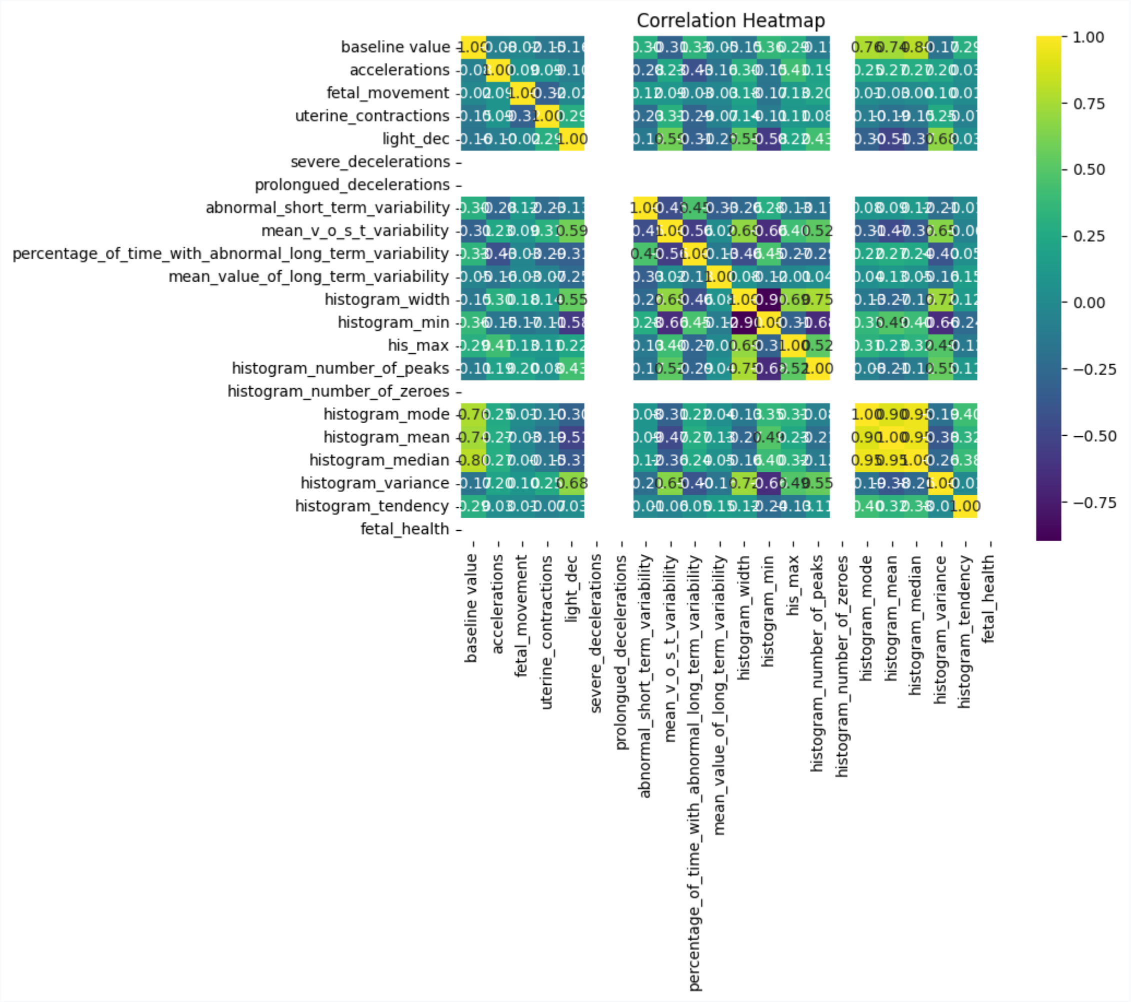
The average accelerations is 0.00 with a spread of 0.00.

Distribution: fetal_movement

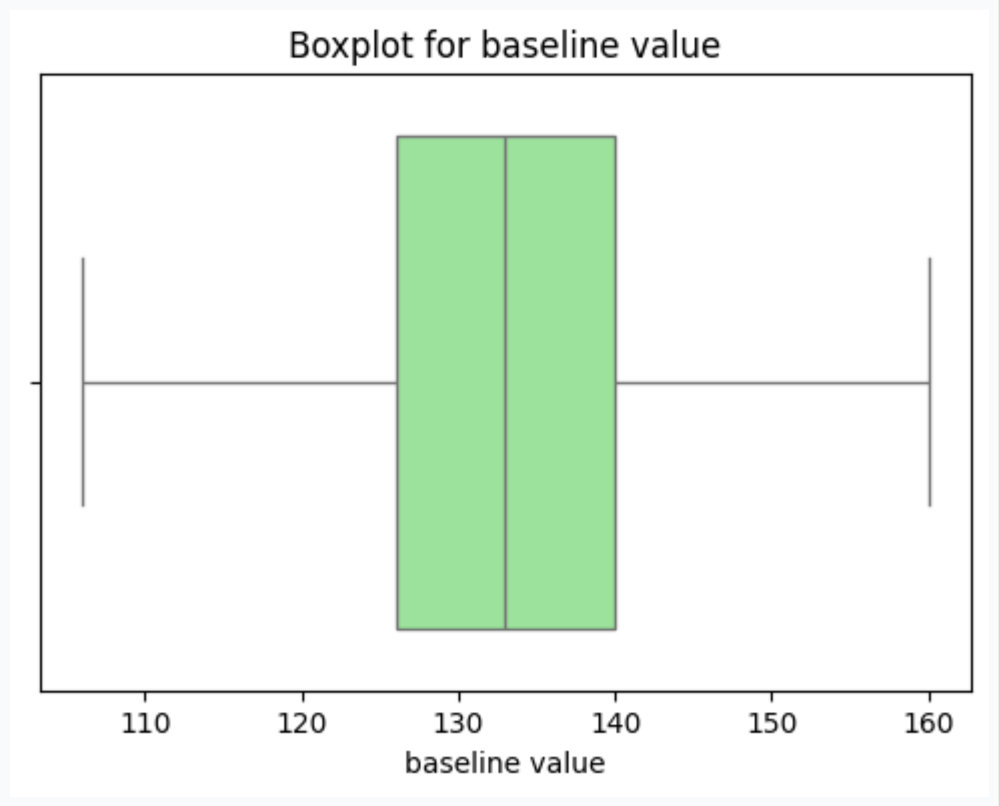


The average fetal_movement is 0.00 with a spread of 0.00.

Correlation Heatmap

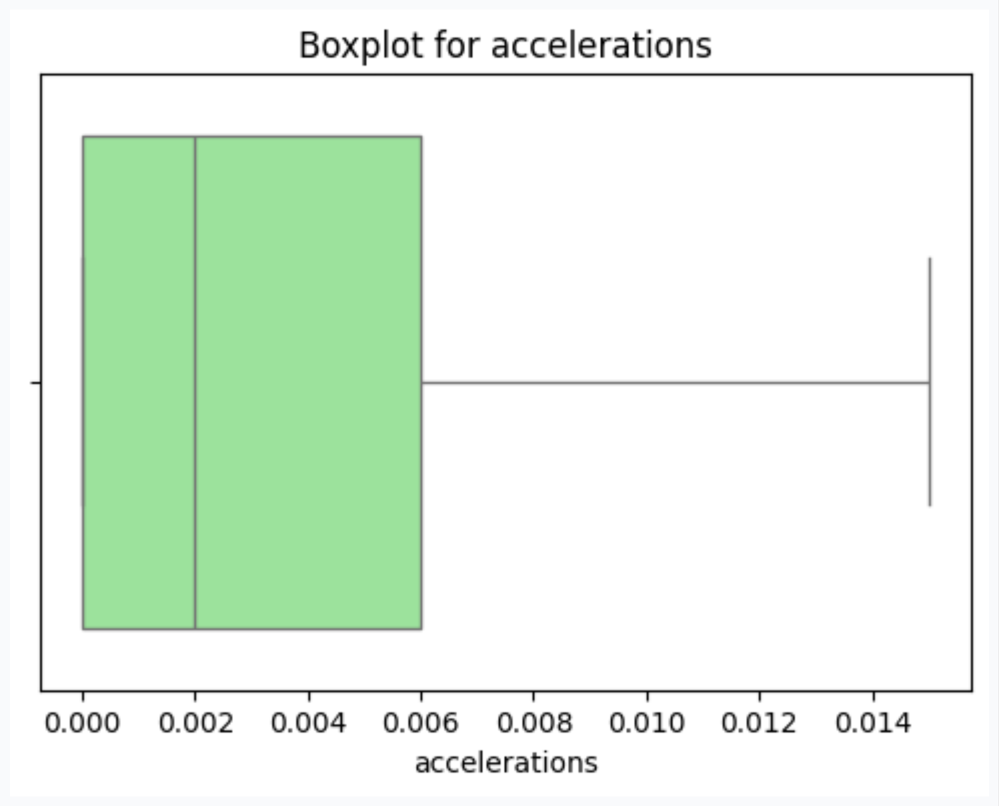


Boxplot: baseline value



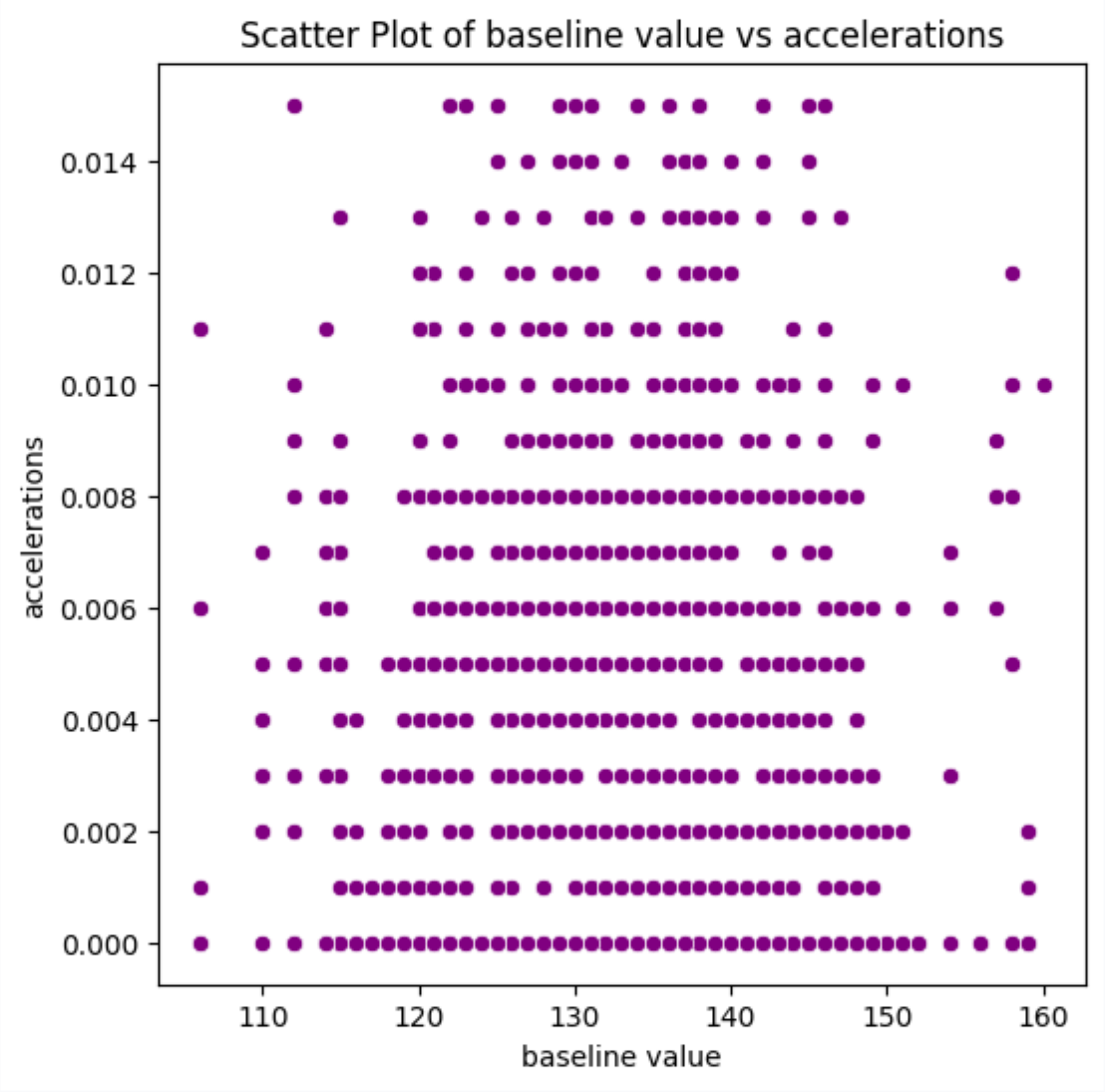
The average baseline value is 133.30 with a spread of 9.84.

Boxplot: accelerations



The average accelerations is 0.00 with a spread of 0.00.

Scatter Plot: baseline value vs accelerations



Conclusion & Recommendations

['The dataset, while commendably clean in terms of missing values and consistent data types, presents significant limitations for typical machine learning applications, particularly classification tasks. The complete absence of missing data is a notable positive aspect. However, the presence of multiple constant features, and most critically, the `fetal_health` target variable being uniformly 1.0, severely restricts its current utility for building a predictive model aimed at differentiating between various health outcomes. If the objective is classification, the current dataset implies a mono-class scenario for the target, rendering it ineffective for training without access to a more diverse dataset or a corrected subset where the target variable exhibits true variability.']

- **Verify and Diversify `fetal_health` variable:** It is paramount to confirm if the `fetal_health` column is indeed constant across the entire original dataset. If so, a different or augmented dataset containing observations with varying `fetal_health` outcomes (e.g., normal, suspect, pathological) is essential for any meaningful classification task.
- **Remove Constant Features:** Eliminate `severe_decelerations`, `prolongued_decelerations`, and `histogram_number_of_zeroes` from the dataset. These features provide no predictive power due to their lack of variance.
- **Conduct Comprehensive Exploratory Data Analysis (EDA):** Perform in-depth visualization and statistical analysis on the remaining, variable features to understand their distributions, identify potential outliers, and assess correlations among them, as well as with the target variable (once variability is established).
- **Engage Domain Expertise:** Consult with medical professionals or domain experts to gain a deeper understanding of the clinical significance of each feature and the implications of their statistical distributions, particularly for features with very low mean values (e.g., accelerations, fetal_movement, light_dec).
- **Prepare for Modeling (Post-Target Rectification):** Once a suitable dataset with a truly variable `fetal_health` target is acquired, proceed with standard machine learning preparatory steps, including feature scaling, potential dimensionality reduction, and the selection and application of appropriate classification algorithms.