Data Analysis Report for  
Canadian Light Source Coding Interview

Presented by: Khoa Nguyen

# Executive Summary

# Introduction

**Overview**: The dataset consists of roughly 1.5 million entries (rows), 6 features (columns).

The features are:

|  |  |
| --- | --- |
| **Features** | **Possible Meaning** |
| Stretch | This could refer to the elongation or compression of an object, perhaps related to deformation. In some contexts, it could relate to the stretching of a wavelength, or DNAs, RNAs, but without additional context, this is speculative. |
| Theta | Theta (θ) represents an angle, commonly used in polar coordinates. In the context of synchrotron experiments, it might relate to the angle of incidence of X-ray beams or any other beams onto a sample, possibly in a Tomography. |
| Twist | This might refer to a rotational or torsional deformation of an object. In some experimental setups, the "twist" might represent the angular orientation or rotation of a sample, possibly DNAs, or a detector. |
| Avg\_stretch | The average value of the stretch for a given set of measurements or over a certain period. |
| Avg\_twist | The average value of the twist for a given set of measurements or over a certain period. |
| Distance | Could refer to a variety of distances - from the distance between a detector and a sample, the distance over which a particular deformation occurs, or even a spatial resolution. The specific context in which "distance" is used would provide more clarity. |

A computer screen shot of a missing number

Description automatically generated**Objective**: This analysis aims to uncover patterns, anomalies, and relationships within the data, providing valuable insights

Figure . Data Quality Examination

# Data Quality

* Missing values: 0
* Duplicated rows: 0

Note: There are rows with same values,  
however, they have different IDs. Therefore,  
it was decided that there is no duplication.

# Univariate Analysis

## Statistical Summary



Figure . Statistical Summary of the Dataset

**Distribution of Data**

* Columns like "theta" and "avg\_twist" have relatively high mean values compared to others.
* Columns "twist" and "distance" have high standard deviations, suggesting varied data distribution.

**Presence of Potential Outliers**

* For columns like “theta”, "twist", "avg\_twist", and "distance", there are significant gaps between the 3rd quartile and the max value, and between the min value and the 1st quartile. This might suggest the presence of outliers.

## Histograms

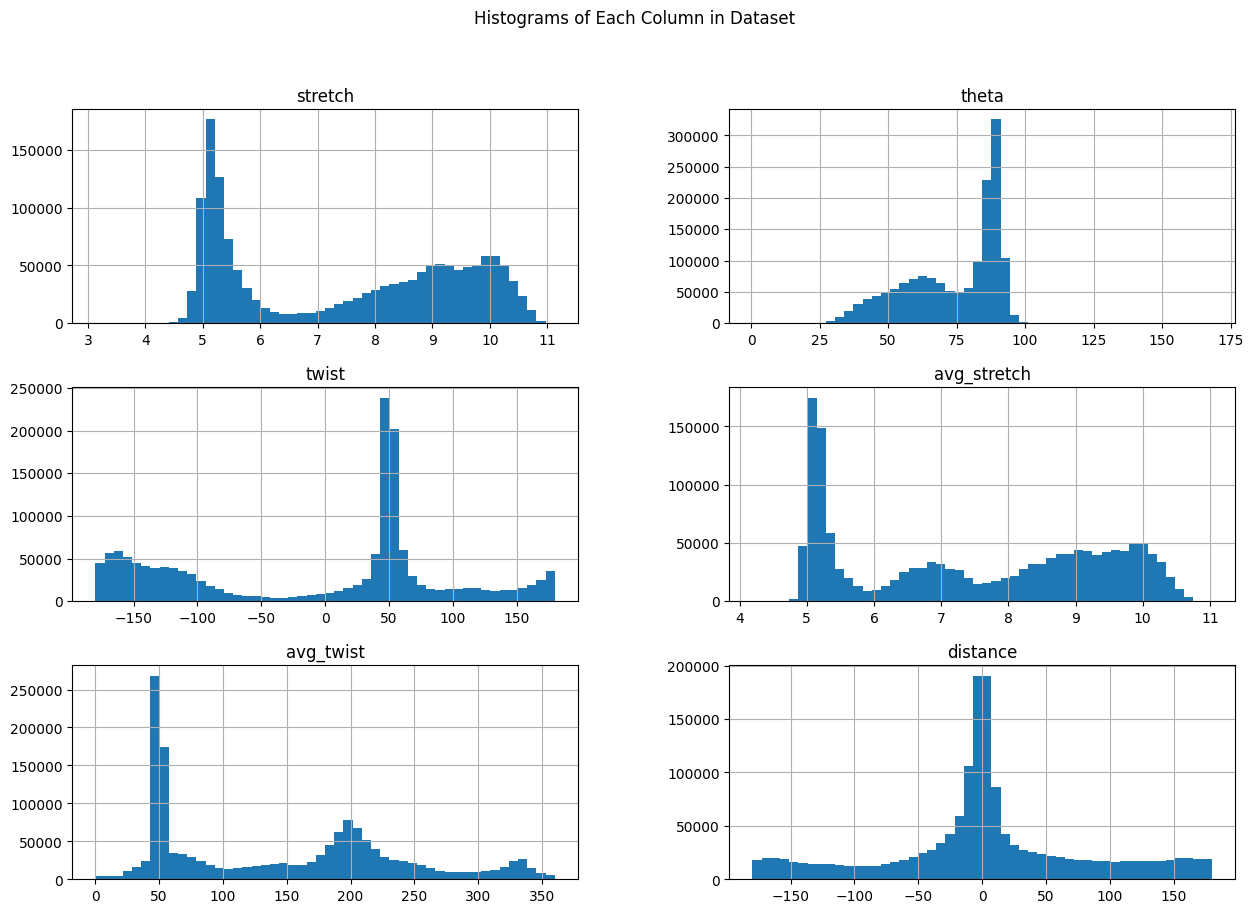


Figure . Histograms of each feature

|  |  |
| --- | --- |
| Features | Possible Interpretations |
| stretch | * The bimodal nature suggests there might be two distinct groups or types of observations in the dataset. * Gaps between peaks suggest that transition states or values between the major groups are less frequent or perhaps less desirable. |
| twist | * While many of these have a twist value close to 50, there are also significant occurrences where the twist is near -100 or 100. * The symmetry around the central peak suggests balanced deviations in the negative and positive directions, possibly indicating equally common clockwise and counterclockwise twists |
| avg\_twist | * The dominant mode in the 25-50 range suggests that many items or occurrences have a twist value within this bracket. * The secondary peak around 200-225 might represent another common twisting behavior or pattern that's distinct from the primary mode. |
| theta | * The dominant mode in the 80-100 range suggests that this is a common orientation or rotation for many observations. * The secondary peak at 50-75 indicates another typical angular orientation or rotation * The frequencies reduce significantly after the 100 mark. |
| avg\_stretch | * The histogram suggests two distinct groups or behaviors in the data: One that typically has an average stretch value around 5 and another that is more varied, with values ranging from 7 to 11. |
| distance | * The significant concentration around 0 might indicate that a lot of the measurements or events being studied occurred close to a reference point * The symmetry of the histogram suggests that deviations from this central value occur with roughly the same frequency on both sides, implying that factors causing these deviations might be random or unbiased in nature. |

## Boxplots

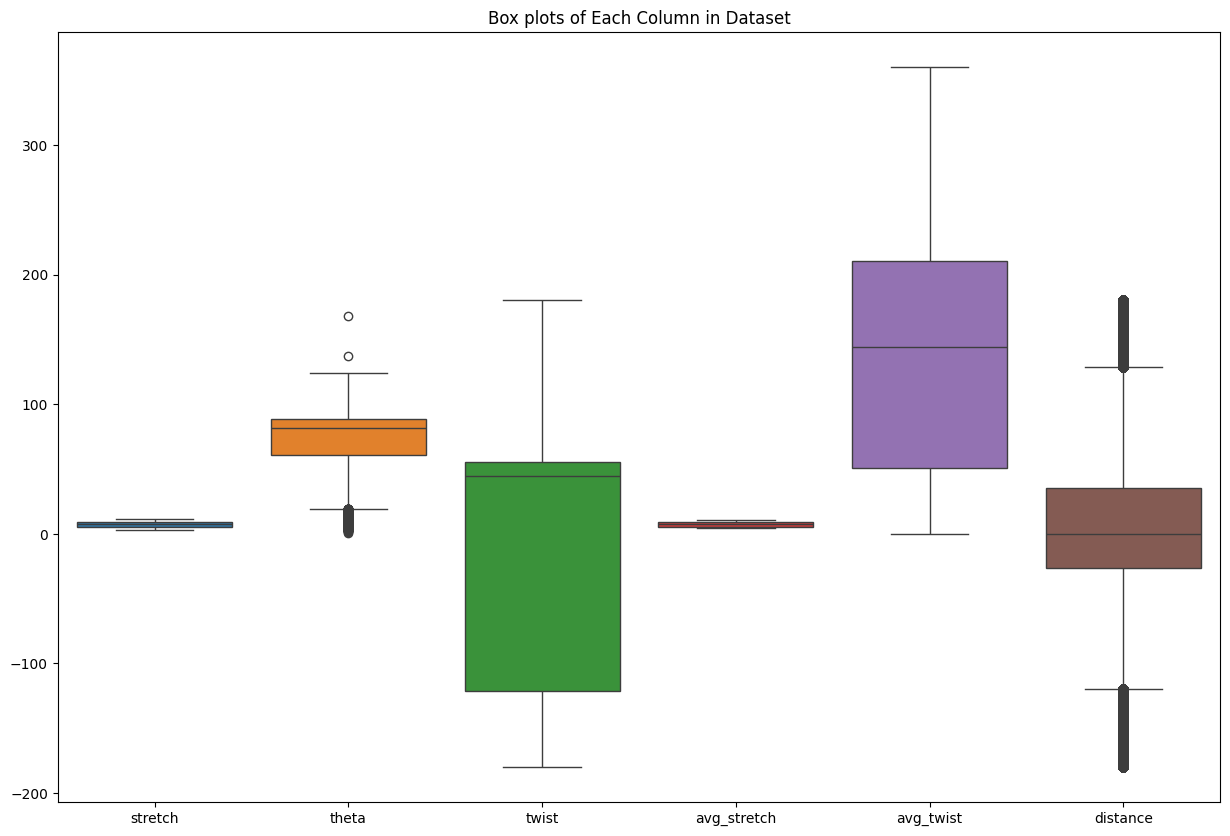


Figure . Boxplots of each feature

* Stretch and avg\_stretch have data values closely centered around zero.
* Theta and Distance seem to have multiple outliers on both ends.
* Twist displays a widespread, with the central 50% of values spanning a large range.
* Avg\_twist showcases a symmetric distribution with moderate variability.
* This could be an indicator that before any modeling or further analysis, some sort of normalization or standardization might be necessary.
* Given the presence of outliers in the theta and distance variables, further investigation into these data points is crucial.
* Some transformations, like log or square root, might be beneficial, especially for theta and twist variables that seem skewed or have wide ranges