Data Analysis Report for
Canadian Light Source Coding Interview

Presented by: Khoa Nguyen

1. Executive Summary

- Most features do not follow a normal distribution. As such, techniques that don't presuppose a normal distribution are favored. For instance, normalization is preferred over standardization for scaling purposes.
- The distribution of many features is bimodal, suggesting the presence of two or more distinct groups or clusters within the observations.
- The features "stretch" and "twist" can be omitted to simplify the dataset. They exhibit a high correlation, mutual information with the "avg_stretch" and "avg_twist" features, and the latter two even demonstrate a more robust relationship with other attributes.
- Should "theta" be considered as the target variable, it's advisable to exclude "distance" from the dataset due to their lack of interdependency.

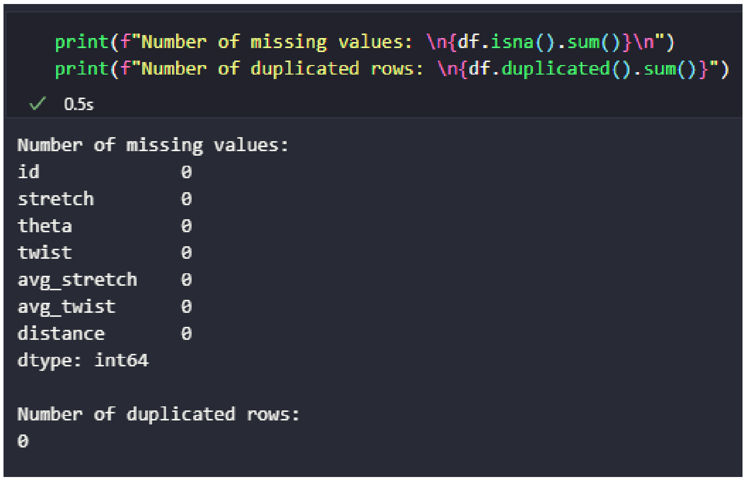
2. Introduction

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.

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



```
print(f"Number of missing values: \n{df.isna().sum()}\n")
print(f"Number of duplicated rows: \n{df.duplicated().sum()}")
```

```
Number of missing values:
id             0
stretch        0
theta          0
twist          0
avg_stretch    0
avg_twist      0
distance       0
dtype: int64

Number of duplicated rows:
0
```

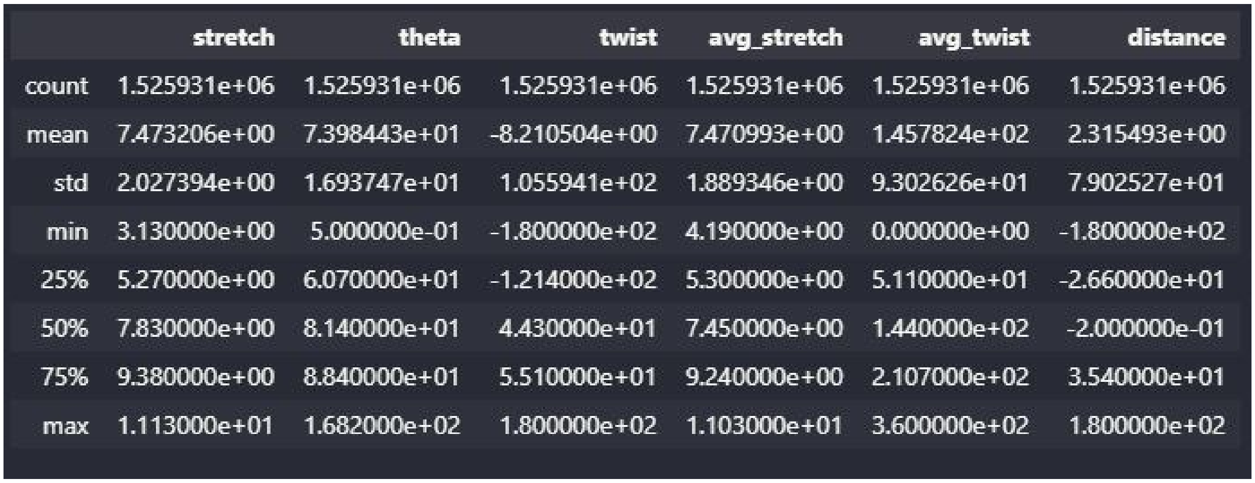
3. Data Quality

- 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

3.1. Statistical Summary



	stretch	theta	twist	avg_stretch	avg_twist	distance
count	1.525931e+06	1.525931e+06	1.525931e+06	1.525931e+06	1.525931e+06	1.525931e+06
mean	7.473206e+00	7.398443e+01	-8.210504e+00	7.470993e+00	1.457824e+02	2.315493e+00
std	2.027394e+00	1.693747e+01	1.055941e+02	1.889346e+00	9.302626e+01	7.902527e+01
min	3.130000e+00	5.000000e-01	-1.800000e+02	4.190000e+00	0.000000e+00	-1.800000e+02
25%	5.270000e+00	6.070000e+01	-1.214000e+02	5.300000e+00	5.110000e+01	-2.660000e+01
50%	7.830000e+00	8.140000e+01	4.430000e+01	7.450000e+00	1.440000e+02	-2.000000e-01
75%	9.380000e+00	8.840000e+01	5.510000e+01	9.240000e+00	2.107000e+02	3.540000e+01
max	1.113000e+01	1.682000e+02	1.800000e+02	1.103000e+01	3.600000e+02	1.800000e+02

Figure 2. 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.

3.2. Histograms

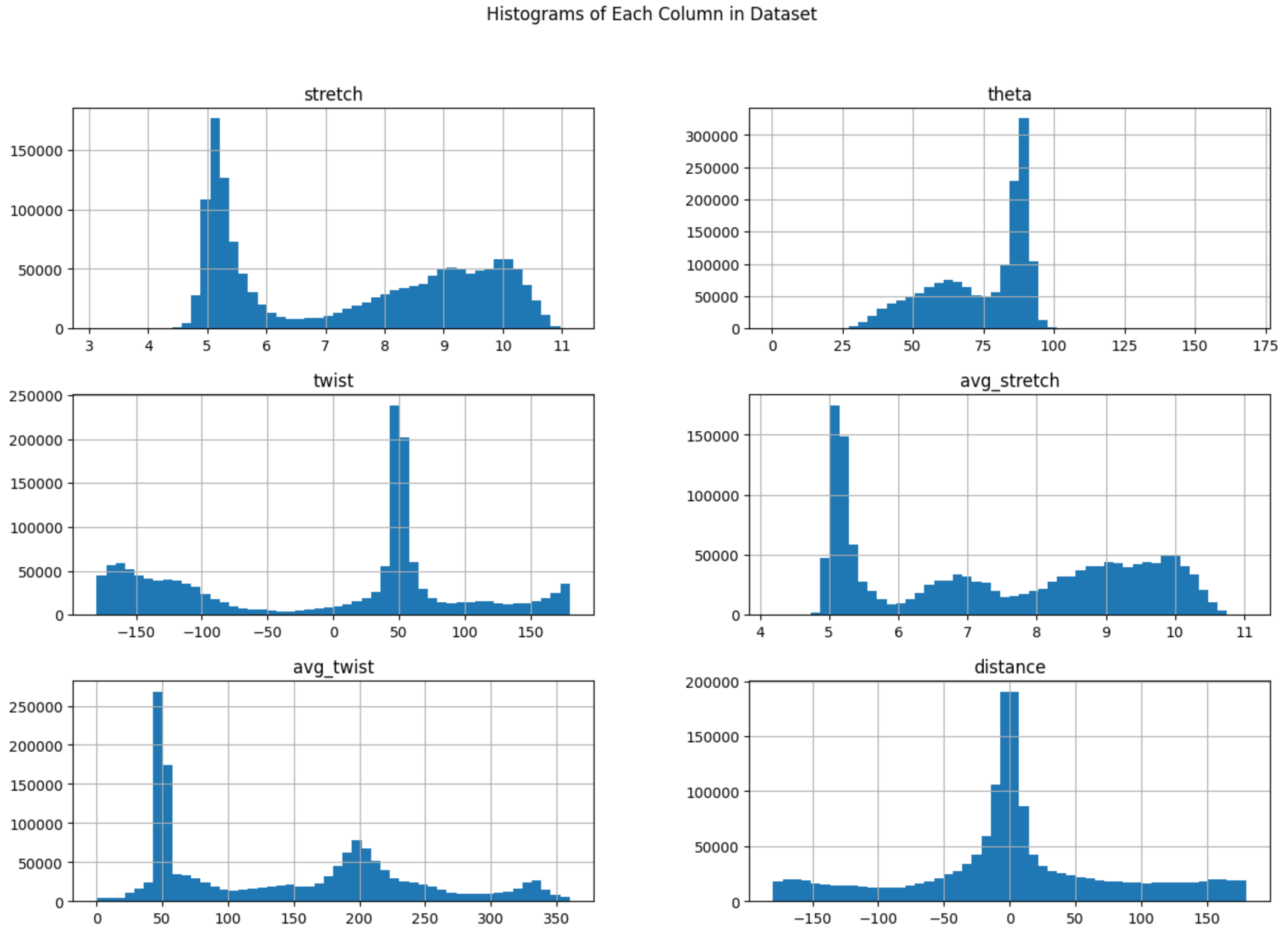


Figure 3. Histograms of each feature

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.

KEY TAKEAWAYS:

- **There is no normal distribution, except for “distance”**
- **Most distributions are bimodal, implying there might be two or more potential groups, clusters.**

3.3. Boxplots

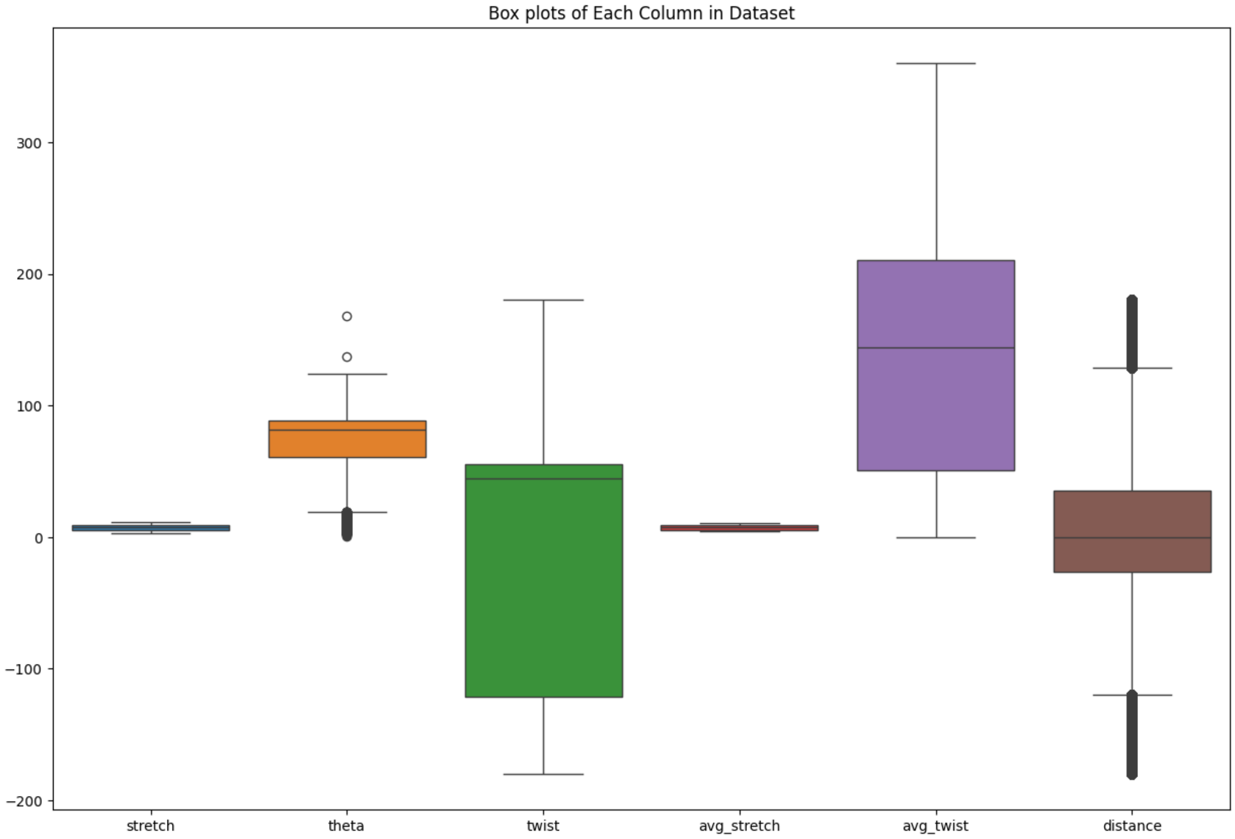


Figure 4. Boxplots of each feature

- 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.

KEY TAKEWAYS:

- **Normalization might be necessary (Standardization is not suitable since most distributions are not normal).**
- **Given the presence of outliers in the theta and distance variables, further investigation into these data points is crucial.**
- **Even though some transformations, might be beneficial, especially for theta and twist variables that seem skewed or have wide ranges, we don't have enough domain-knowledge to ensure there would be no loss of information.**

4. Data Processing and Transformation

Before proceeding with further analysis, it's beneficial to eliminate outliers and normalize the data for two primary reasons:

- Firstly, by comparing the analysis results of both the original and processed data, we can ensure consistency. If the results align, it's prudent to continue with the processed data.
- Secondly, cleansed and normalized data often enhances processing speed in subsequent steps.

The IQR technique was used to remove outliers from the dataset. Then, the dataset was normalized with Min-Max Scaling.

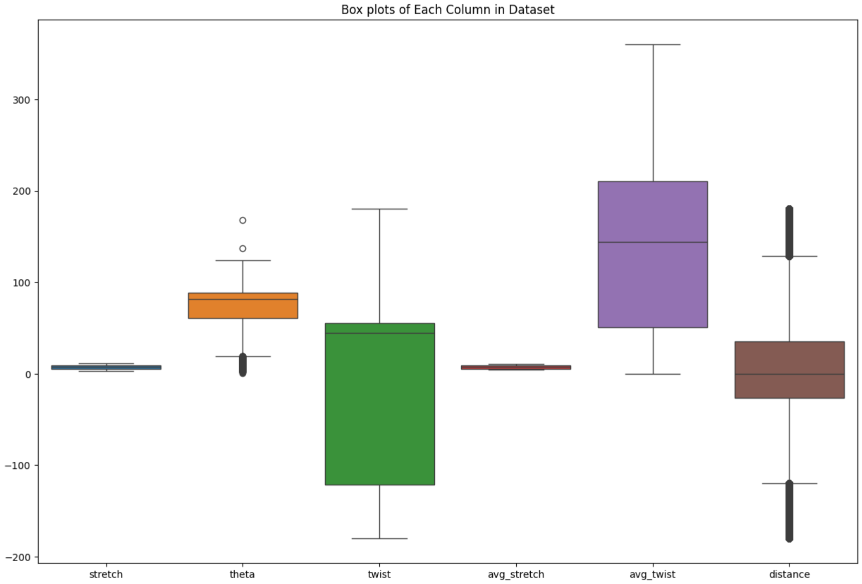


Figure 5. Boxplots of each column in original dataset.

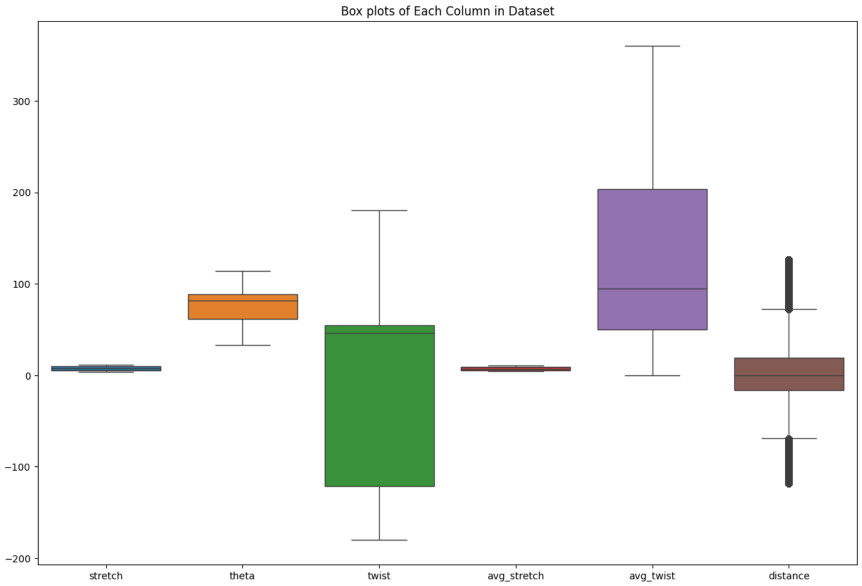


Figure 6. Boxplots of each column in dataset with outliers removed.

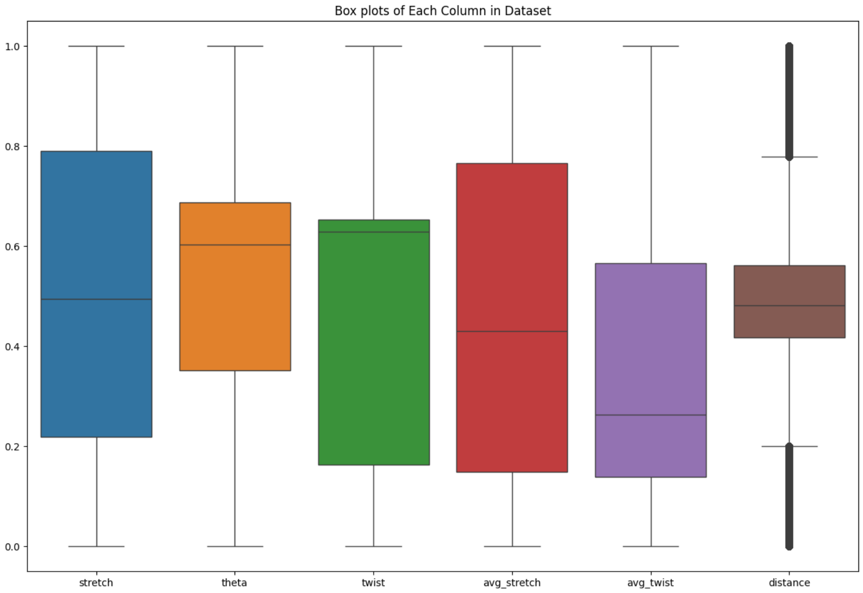


Figure 7. Boxplots of each column in normalized dataset with outliers removed.

5. Bivariate/Multivariate Analysis

Correlation analysis was performed by plotting matrices of Pearson's Coefficient, Spearman's Rank Correlation Coefficient, and Kendall's Tau. In addition, a matrix showing mutual information between features was also calculated.

5.1. Original vs Processed Dataset

Comparing results of correlation analysis between original and processed datasets yields similar results. The correlation analysis even produced slightly better results. Therefore, it was decided to go with the processed dataset for downstream experiments.

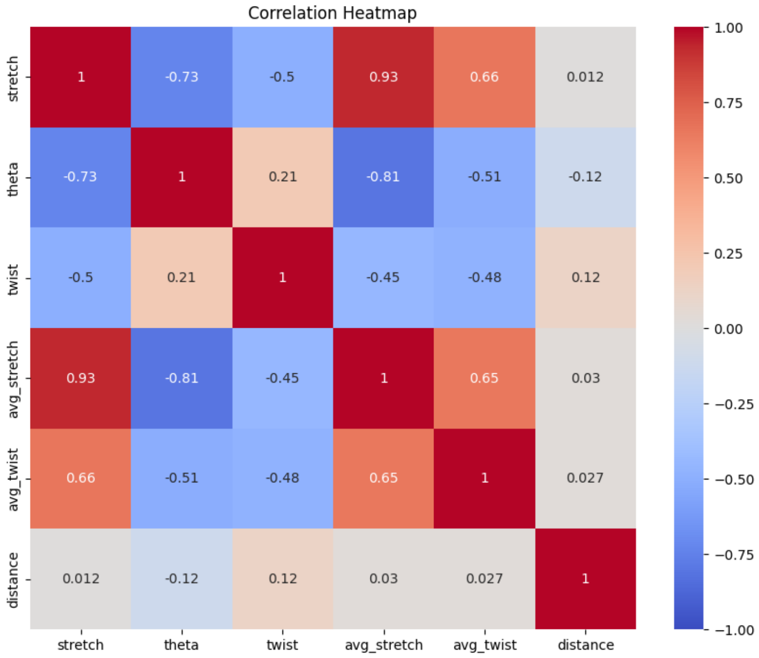


Figure 8. Pearson's Correlation Heatmap (Orginal)

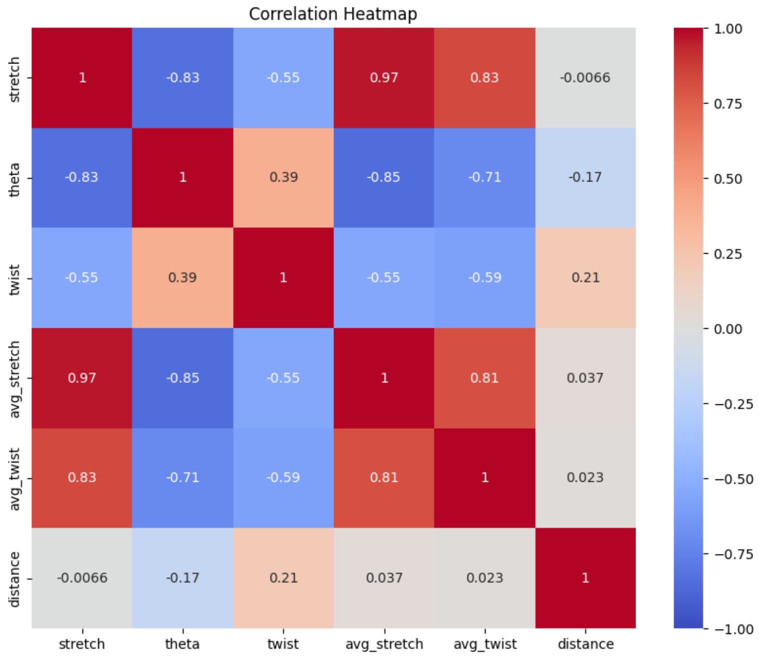


Figure 9. Pearson's Correlation Heatmap (Processed)

5.2. Discussion

- The three correlations matrices demonstrate similar results in the sense that there are fairly strong relationships between “stretch” vs. “theta” (strong negative linear relationship and strong negative monotonic relationship), average relationships between “stretch” vs. “twist” (negative linear relationship and negative monotonic relationship), and between “theta” vs. “twist” (positive linear relationship and positive monotonic relationship).
- The three correlations matrices also show similar, even stronger, relationships between “avg_stretch”, “avg_twist”, and “theta” to those of “stretch” and “twist”.
- The mutual information heatmap shows there are extremely high mutual information between “stretch” vs. “avg_stretch”, and “twist” vs. “avg_twist”.
- The feature “distance” does not seem to have any kind of relationship with other features. However, there is still some dependency between “distance” and other features.

KEY TAKEAWAYS:

- **Both features “stretch” and “twist” can be removed to reduce the complexity of the dataset**
- **If “theta” is the target variable, then “distance” can be removed as well since they don’t have any kind of relationship, dependency.**

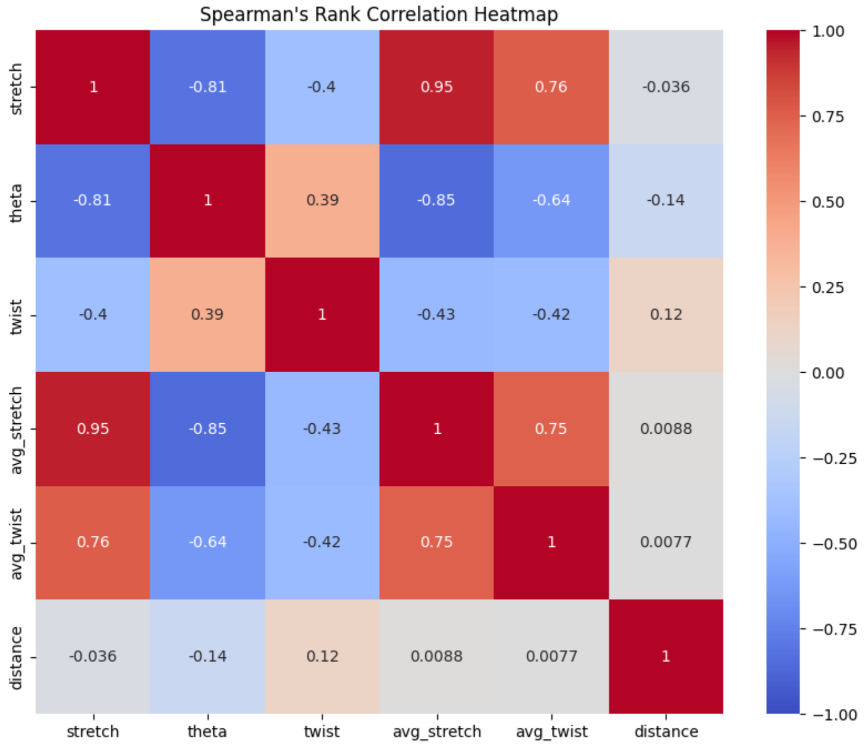


Figure 10. Spearman's Rank Correlation Heatmap (Processed).

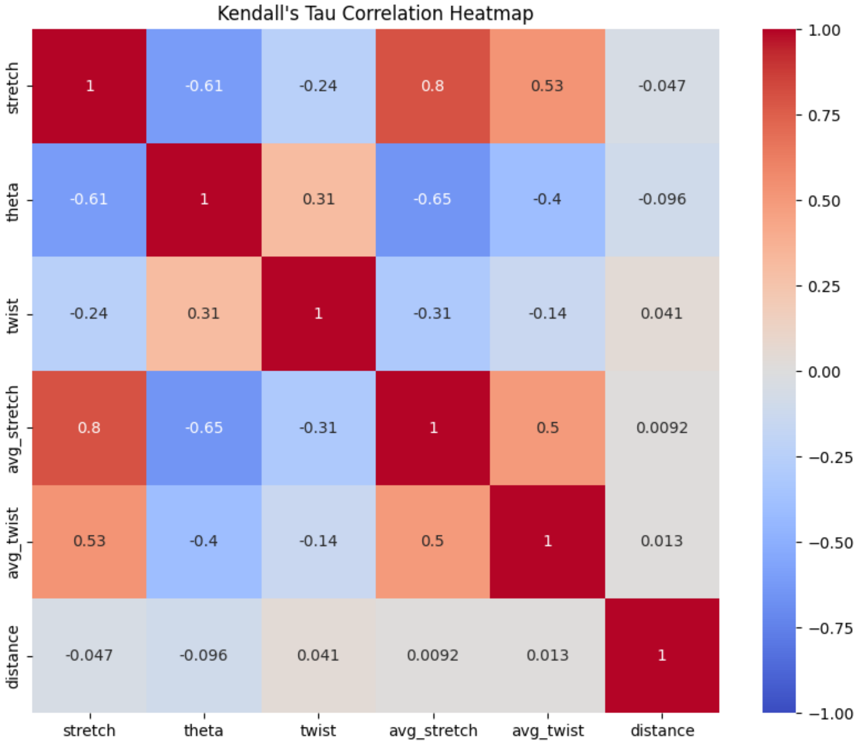


Figure 11. Kendall's Tau Heatmap (Processed).

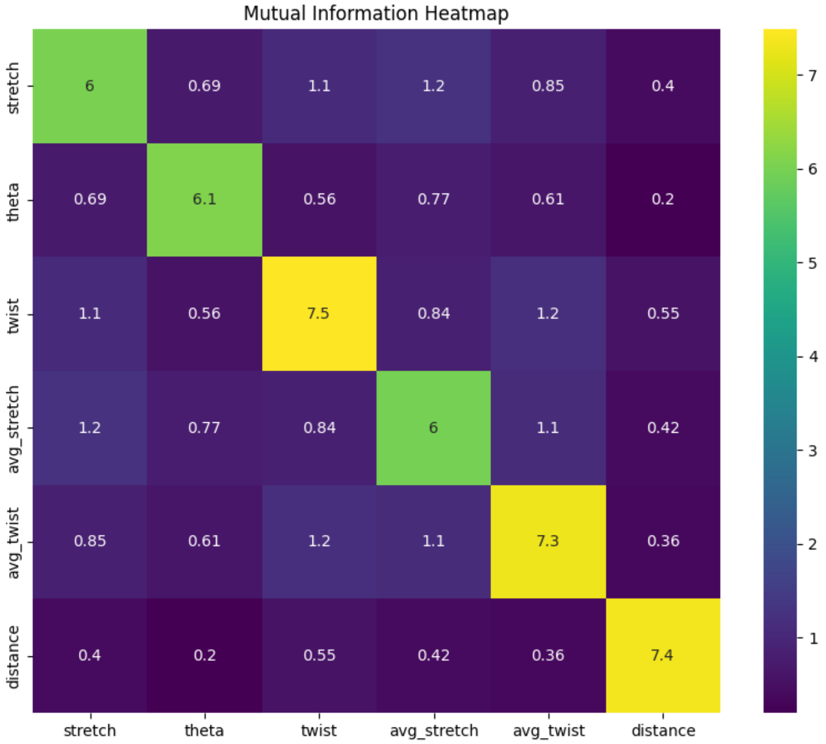


Figure 12. Mutual Information Heatmap (Processed)