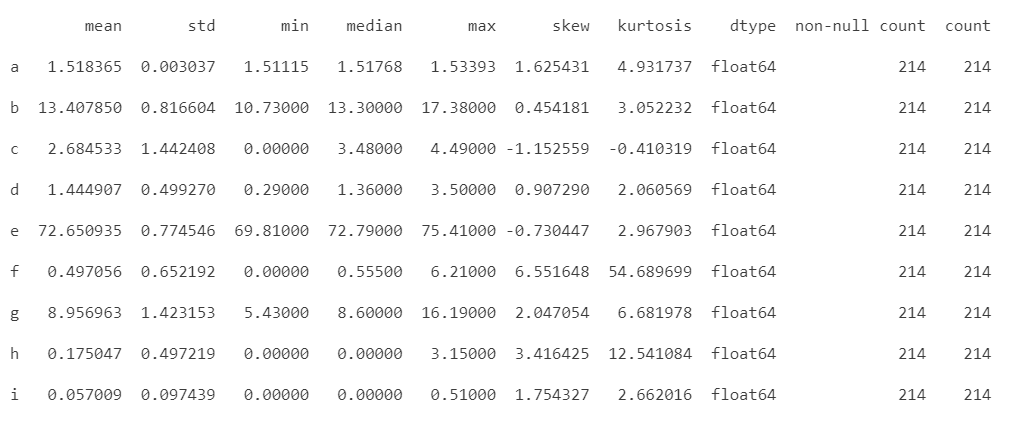
**Question 1:**

**Objective:**

The primary objective of this analysis was to determine if different petrol formulations, differentiated by varying levels of additives, exhibit statistically significant differences. This analysis supports a third-party certification organization’s requirement to validate the distinctness of these formulations.

**1.0 Descriptive Statistics Overview**

The dataset comprises several variables, each showing unique statistical characteristics, which are summarized as follows:



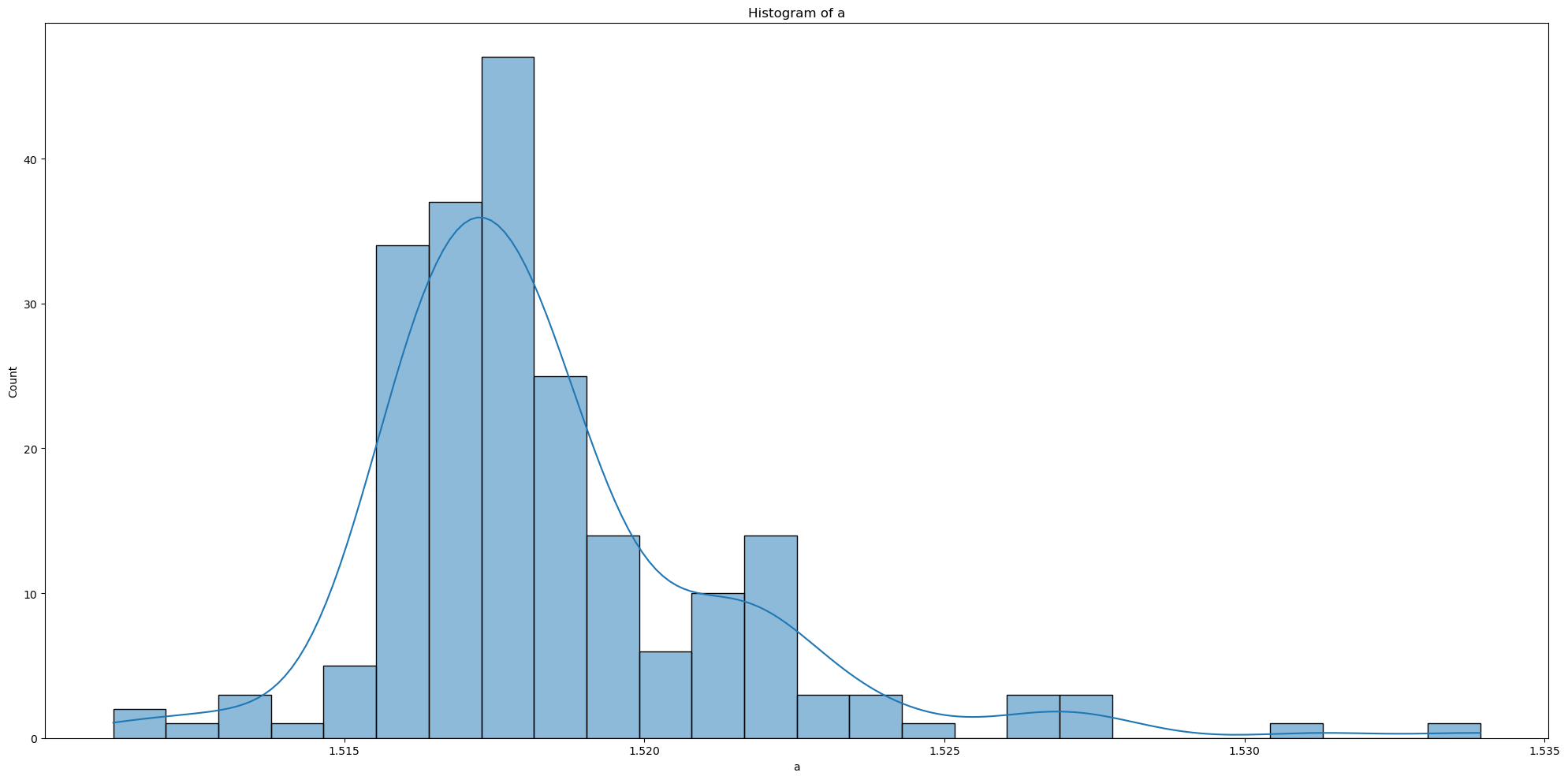
* **Central Tendency and Variability:** Variables exhibit a wide range of means and standard deviations, indicating **diverse scales and distributions**. For instance, variable b has a mean of 13.48 with a standard deviation of 0.816, suggesting moderate variability around a higher mean value. In contrast, variable **i** shows minimal average and variability, with a mean of only 0.057 and a standard deviation of 0.097.
* **Skewness and Kurtosis:** The skewness and kurtosis values suggest **varying degrees of asymmetry and tailedness in the data distributions**. **Variable f is highly skewed** (skewness = 6.55) and exhibits extreme kurtosis (54.69), indicating a distribution with a pronounced tail and potential outliers. This contrasts with **variable c, which shows negative skewness and kurtosis**, suggesting a distribution that leans towards lower values with fewer and less extreme outliers.
* **Data Integrity**: All variables show complete data integrity with **no null values** across the board (214 non-null counts for each), ensuring robust statistical analysis and modeling.

**2.0 Graphical Analysis**

**2.1 Histogram**

In this section, we will look into each and single variable distribution with its corresponding kernel density.

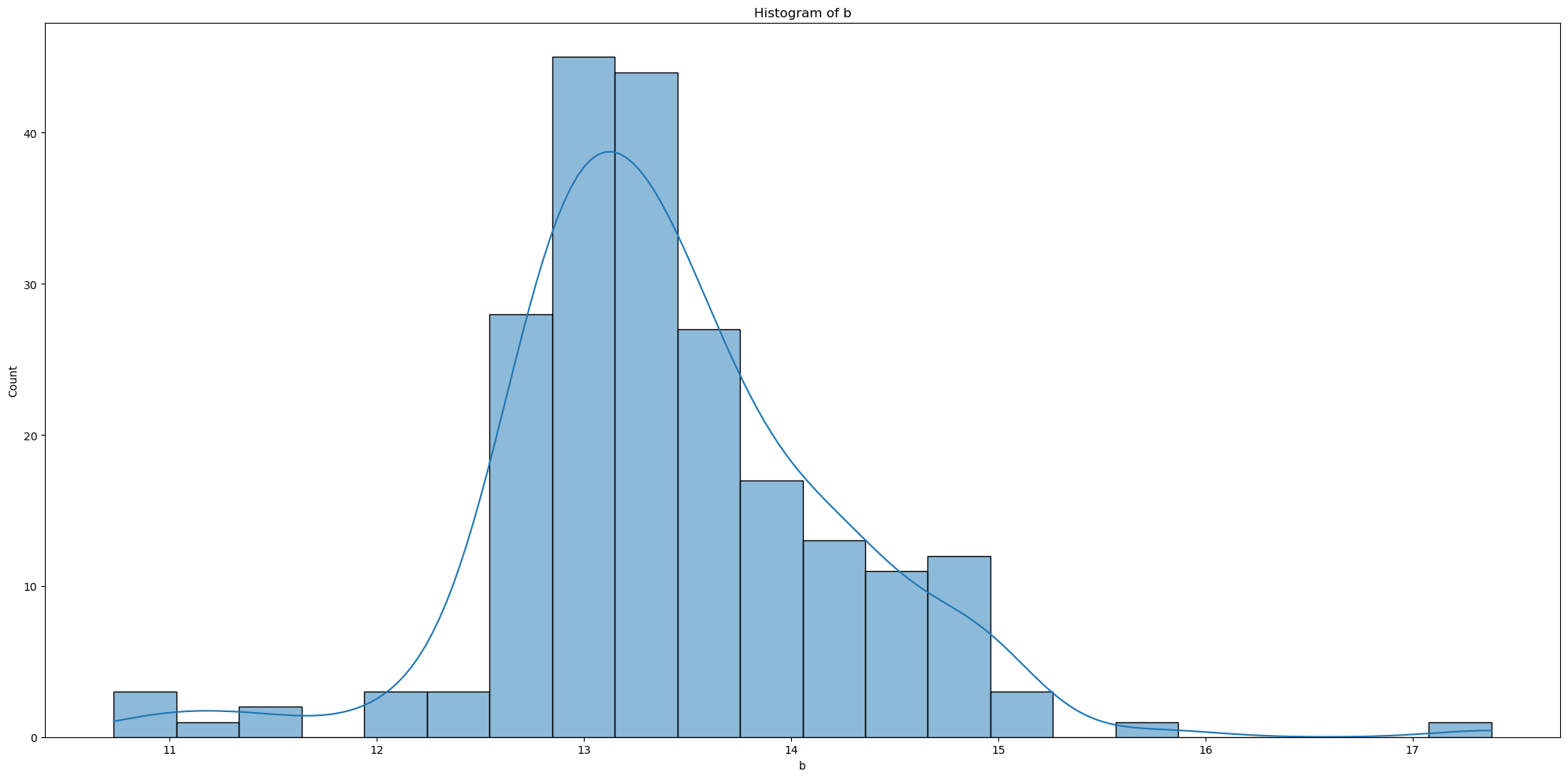
**2.1.1 Histogram of variable ‘a’**



The histogram shows a single-variable [a] distribution with its corresponding kernel density estimate overlaid.

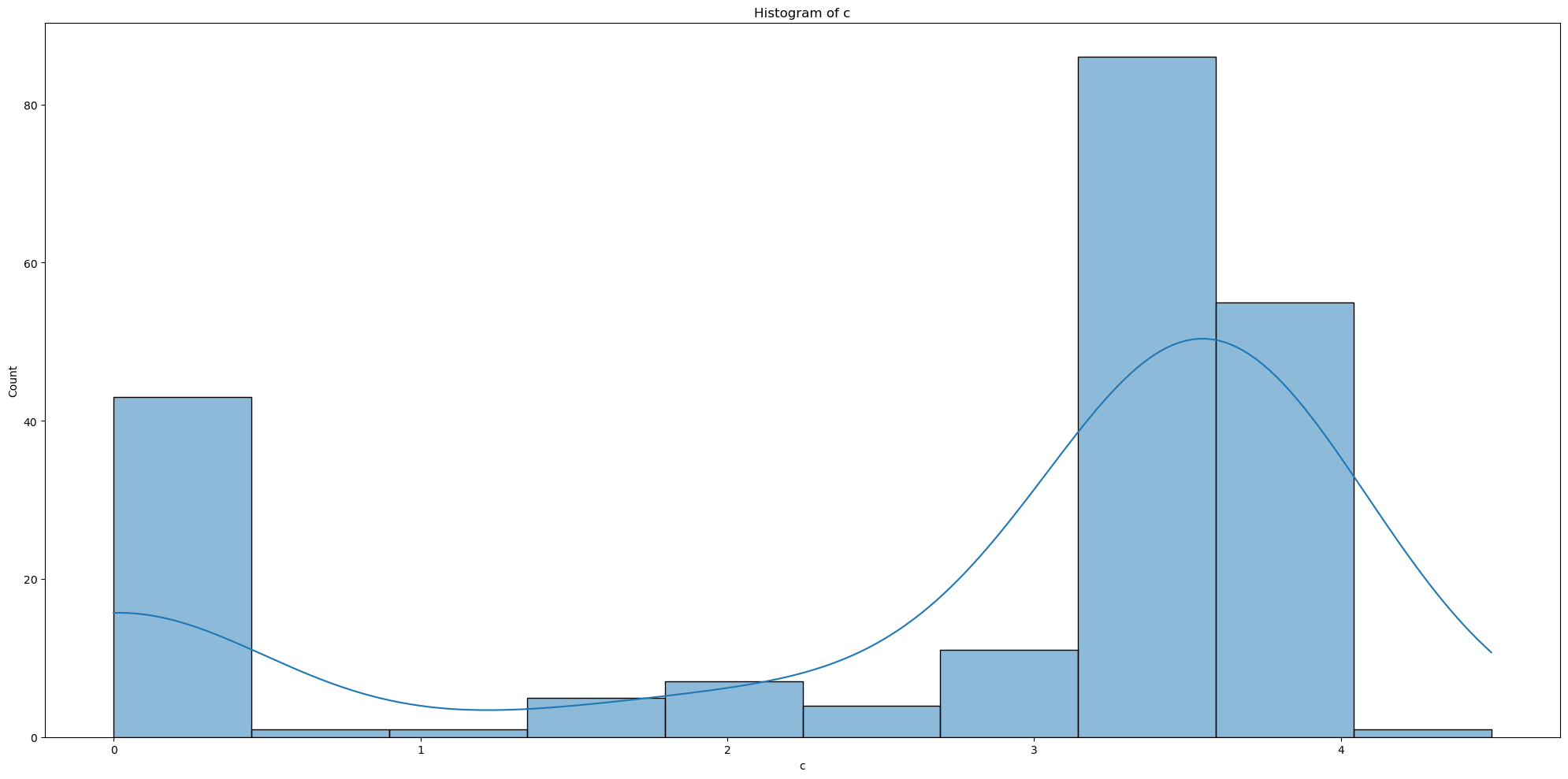
* **Distribution Shape**: Slightly **right-skewed** .
* **Data Concentration**: Most values cluster around a central peak slightly above 1.52.
* **Tail**: Visible right tail suggests the presence of outliers or extreme values.
* **Statistical Properties:** Indications of **non-normality** confirmed by skewness and kurtosis values, important for models assuming normal distributions.

**2.1.2 Histogram of variable ‘b’**



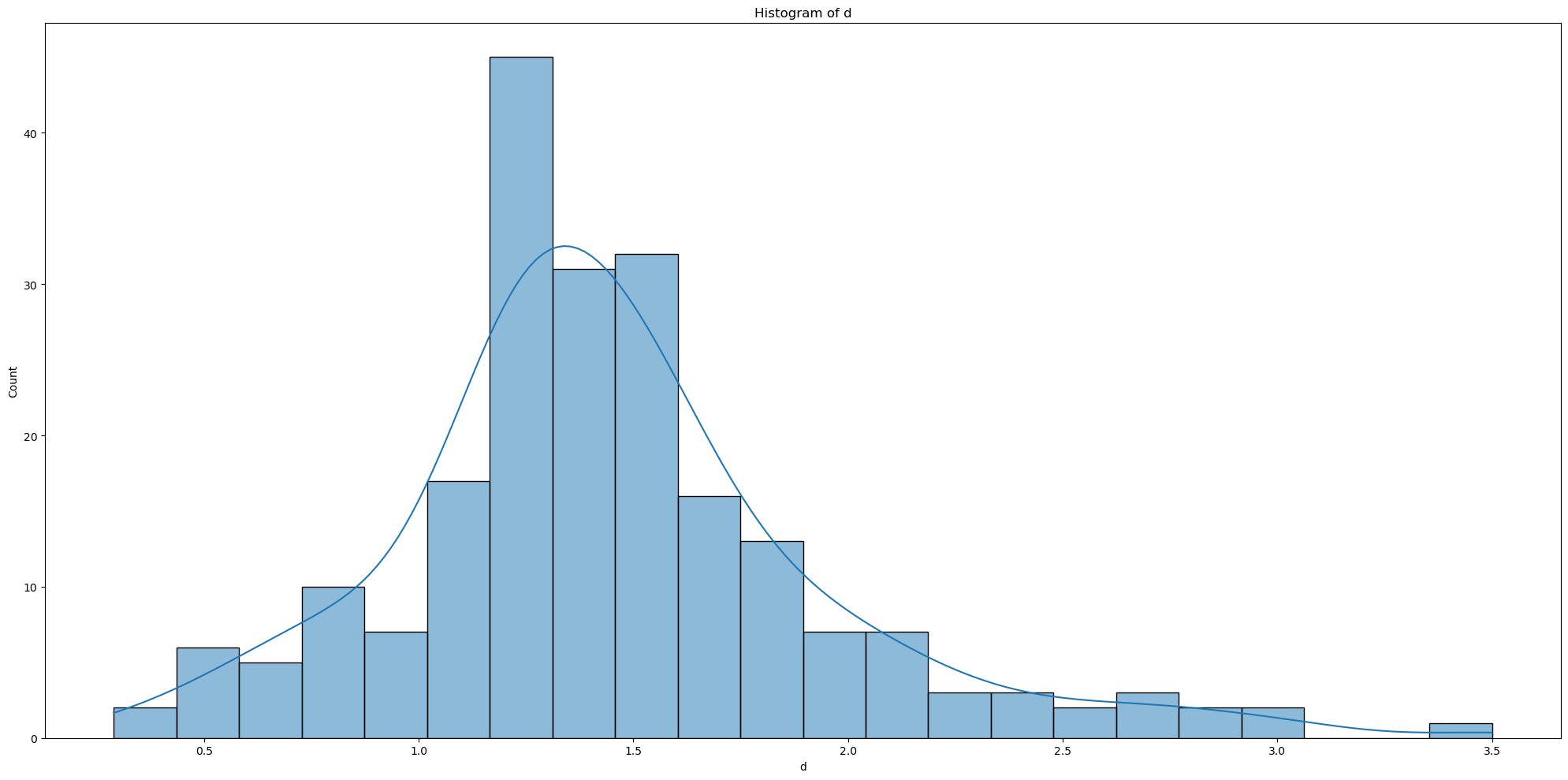
* **Central Tendency:** Data peaks around the central value near 13, indicating the mode.
* **Skewness (0.45)**: Distribution is slight **right-skewed**, with a longer tail extending towards higher values.
* **Symmetry**: Fairly symmetrical around the peak despite right-skewed.
* **Outliers**: Potential outliers or rare high values are indicated by the tail on the right side.
* **Normality**: The distribution **does not perfectly fit a normal distribution** due to the skewness to the right. Addition, kurtosis (3.05) indicates more data is present in the tails compared to normal distribution.

**2.1.3 Histogram of variable ‘c’**



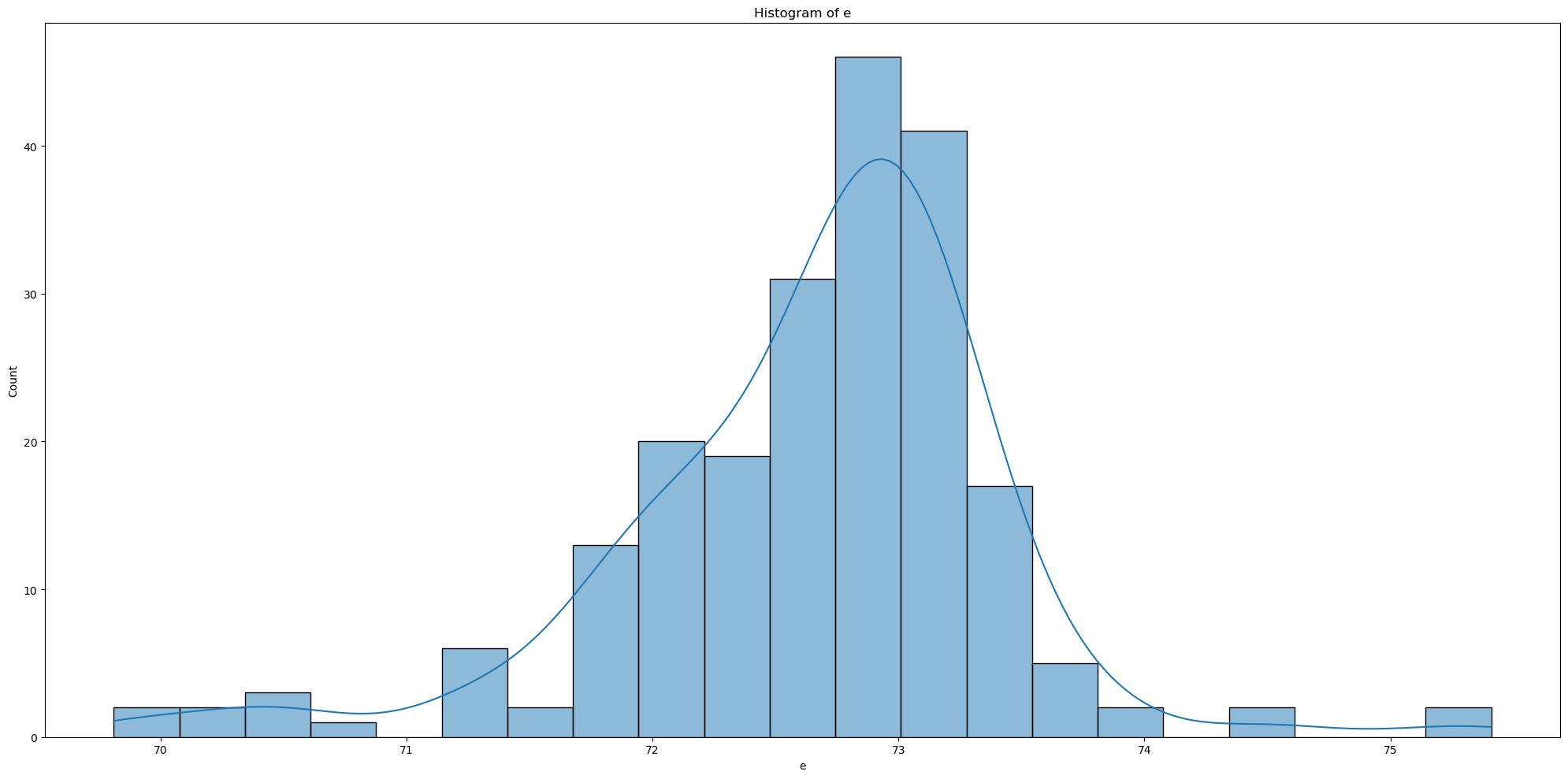
* **Bimodal Distribution**: The histogram shows two distinct peaks, suggesting a bimodal distribution.
* **Skewness** (-1.14): The distribution is **left-skewed**, indicating a longer tail extending towards lower values.
* **Normality**: The negative skewness and visible separation between modes indicate the **distribution deviates from normality**.
* **Kurtosis** (3.05): The kurtosis value is close to that of a normal distribution, implying the tails are neither particularly heavy nor light, but the bimodality affects the overall shape.

**2.1.4 Histogram of variable ‘d’**



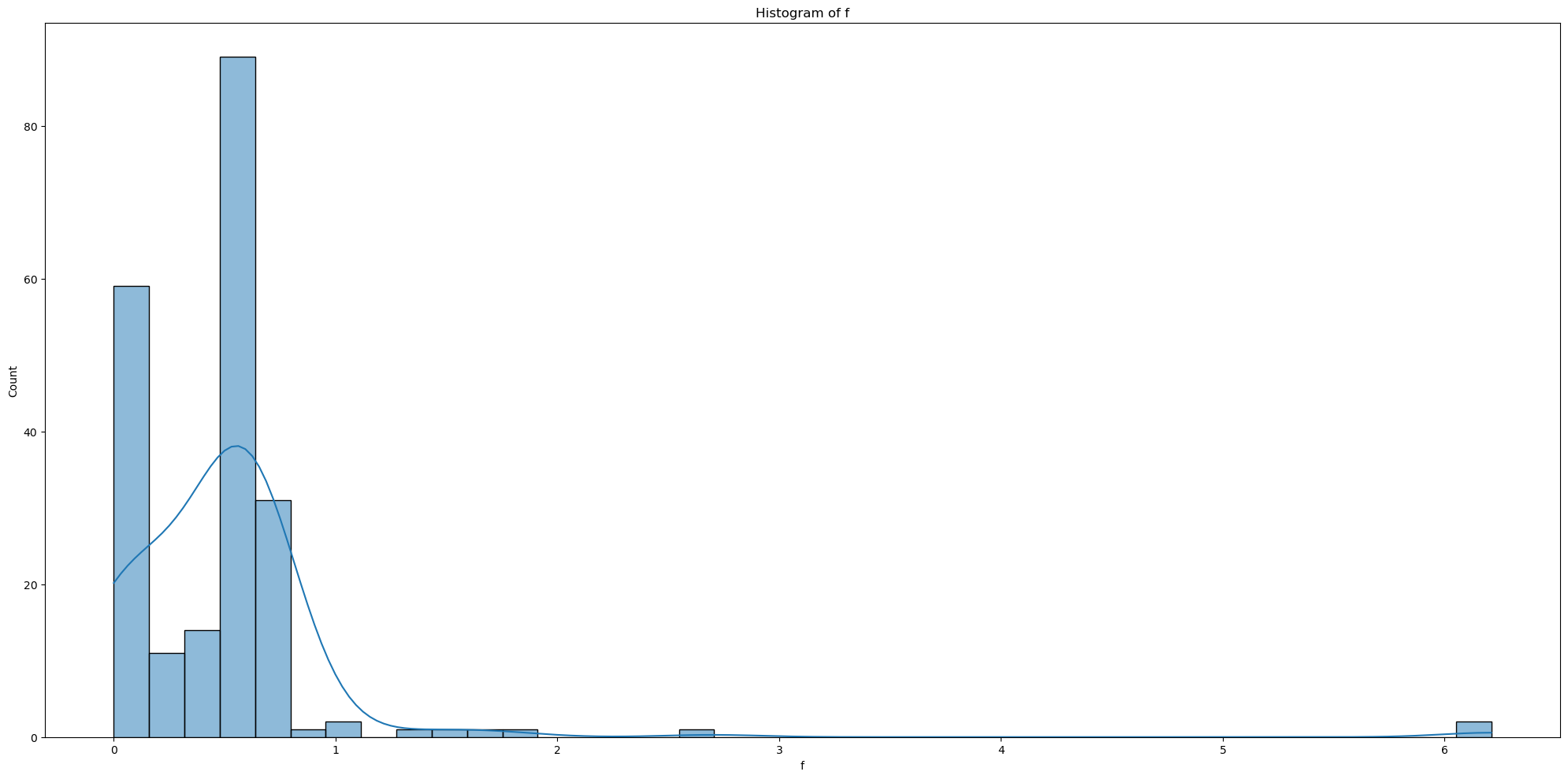
* **Central Tendency:** Data peaks around the central value near 1.0, indicating the mode.
* **Skewness (0.907)**: Distribution is **right-skewed**, with a longer tail extending towards higher values.
* **Symmetry**: Fairly symmetrical around the peak despite right-skewed.
* **Outliers**: Potential outliers or rare high values are indicated by the tail on the right side.
* **Normality**: The distribution **does not perfectly fit a normal distribution** due to the skewness to the right. Addition, kurtosis (2.06) indicates more data is present in the tails compared to normal distribution.

**2.1.5 Histogram of variable ‘d’**



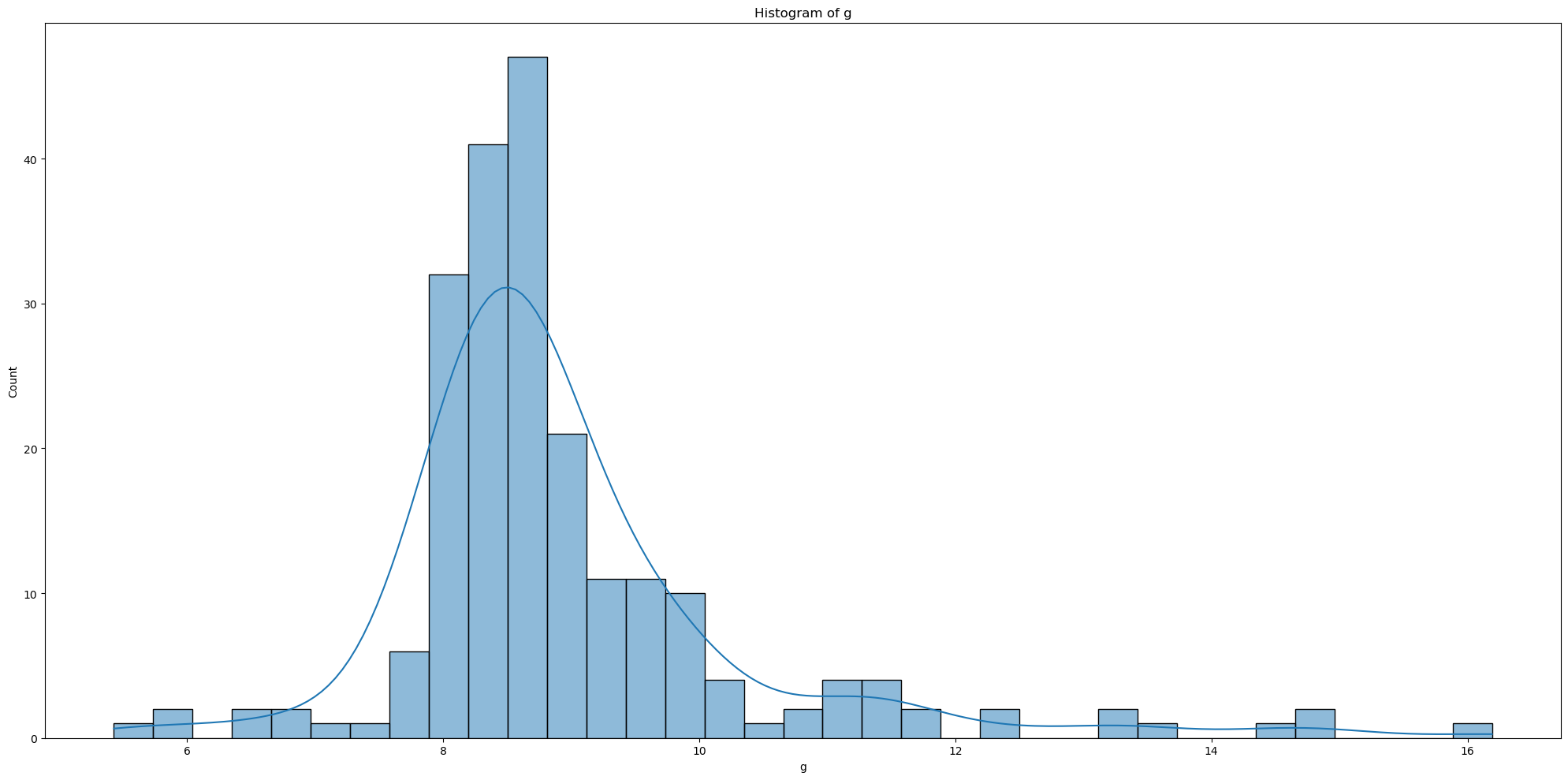
* **Central Tendency**: The data clusters around 72.5 to 73.
* **Skewness**: (0.73): The distribution is **left-skewed**, indicating a longer tail extending towards lower values.
* **Normality**: Kurtosis is close to 3 suggests a distribution similar to **a normal curve but with a slight flatness**, indicating fewer extreme outliers than a normal distribution would have.
* **Outliers**: Likely fewer and less severe outliers due to the kurtosis value.

**2.1.6 Histogram of variable ‘f’**



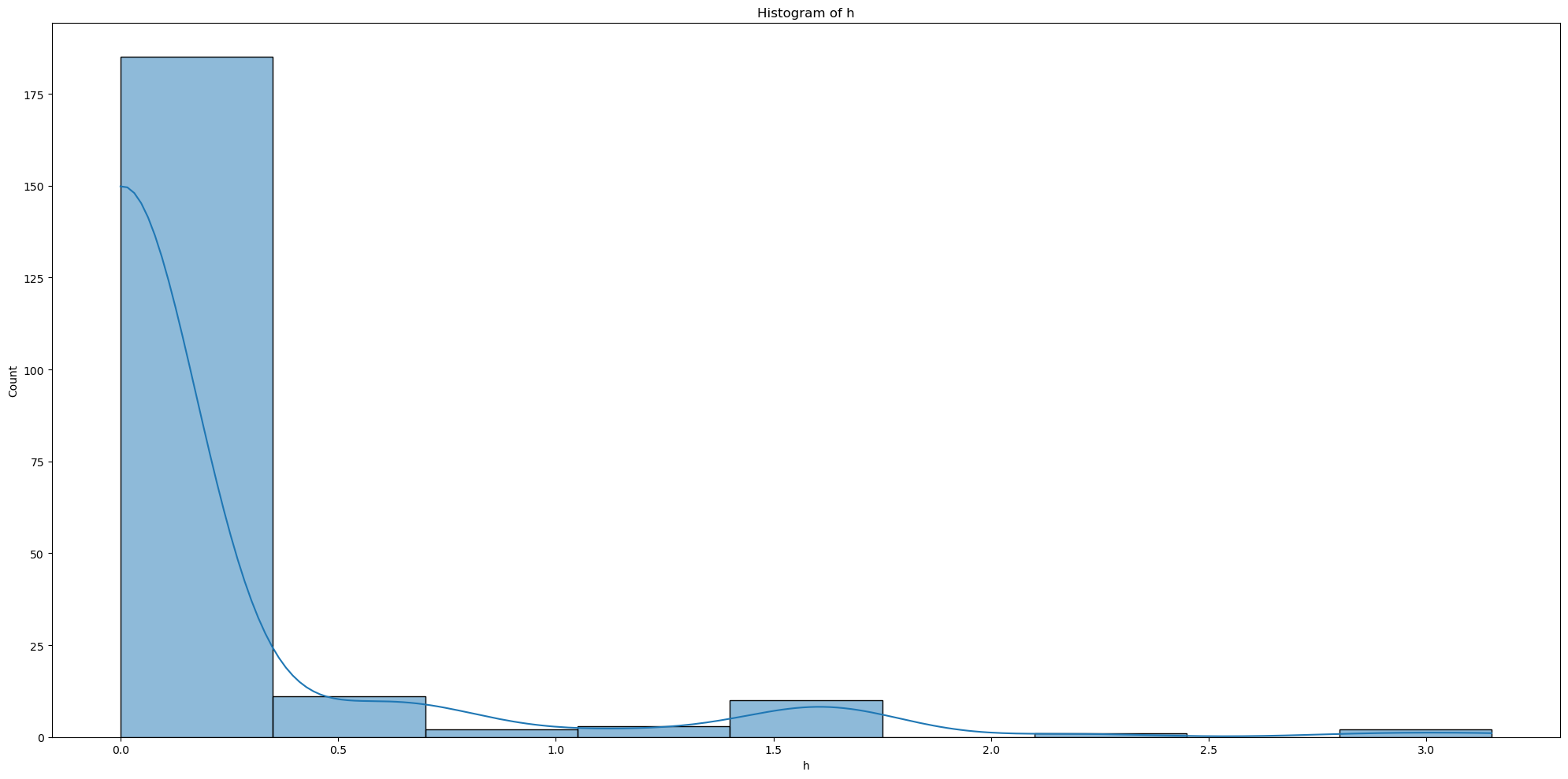
* **Central Tendency**: The data heavily concentrates near the lower values, primarily around 0 to 1.
* **Skewness**: **Extremely high positive skewness** indicates a significant tail stretching toward higher values, which is less common in the data.
* **Normality**: High kurtosis (54.68) indicates a very peaked distribution with heavy tails, suggesting a **significant presence of outliers**, particularly towards the higher end of the scale.
* **Outliers**: The presence of extreme values is pronounced, affecting the overall distribution shape.

**2.1.7 Histogram of variable ‘g’**



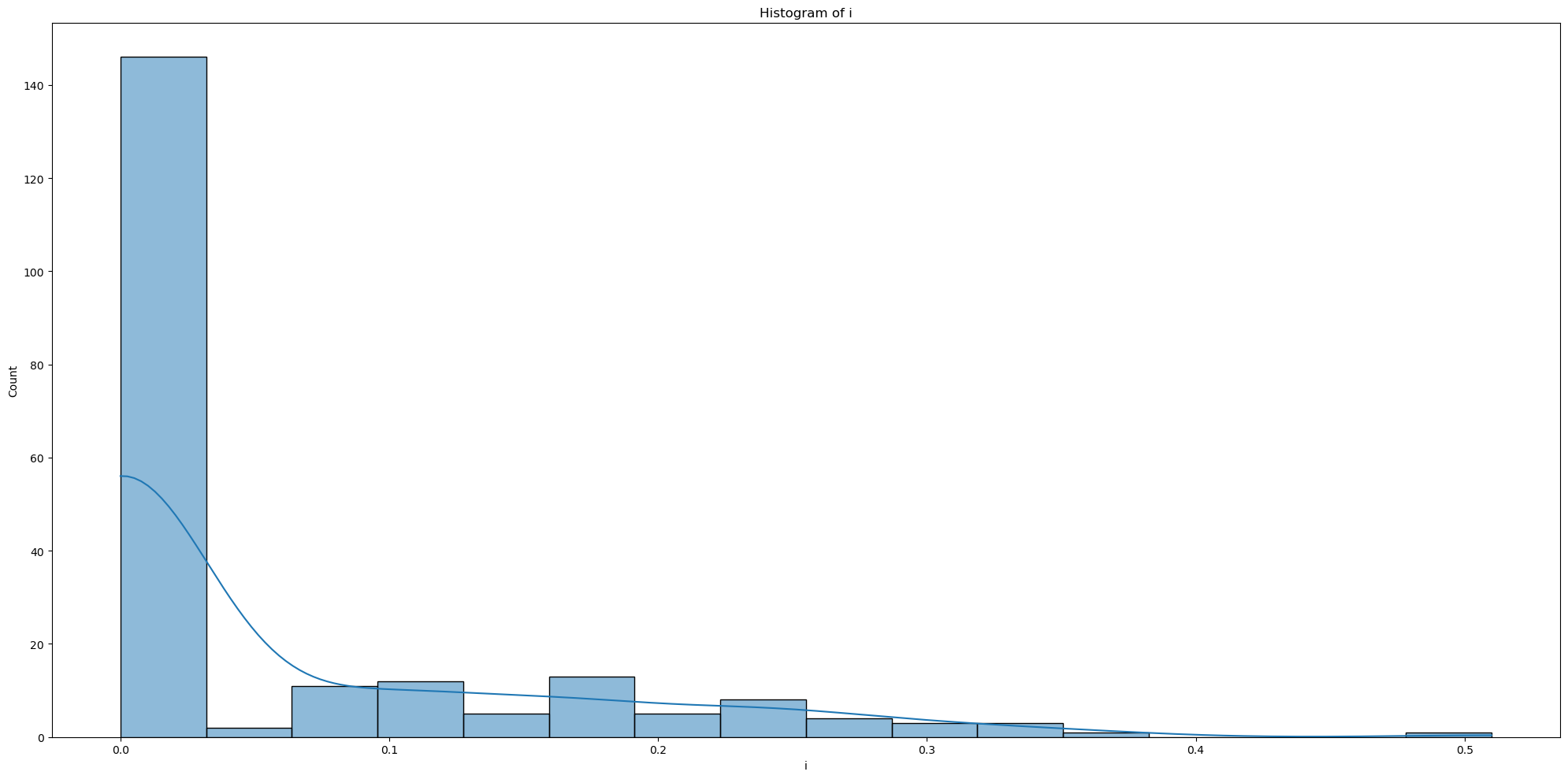
* **Central Tendency**: Most data points cluster around the range of 8 to 10.
* **Skewness**: Noticeable **right skew**, indicated by a tail extending towards higher values beyond 10.
* **Normality**: The distribution exhibits a pronounced peak and fat tails, **deviating significantly from normality**.
* **Outliers**: Likely presence of outliers, particularly on the higher end of the distribution spectrum due to the extended right tail.

**2.1.8 Histogram of variable ‘h’**



* **Central Tendency**: The data is heavily concentrated near 0.
* **Skewness**: The distribution has a very **high right skew**, with a long tail extending towards higher values.
* **Normality**: The pronounced peak and extreme fat tail indicate a **significant deviation from normal distribution**.
* **Outliers**: The extended right tail suggests the presence of outliers, with most data points clustering at the lower end.

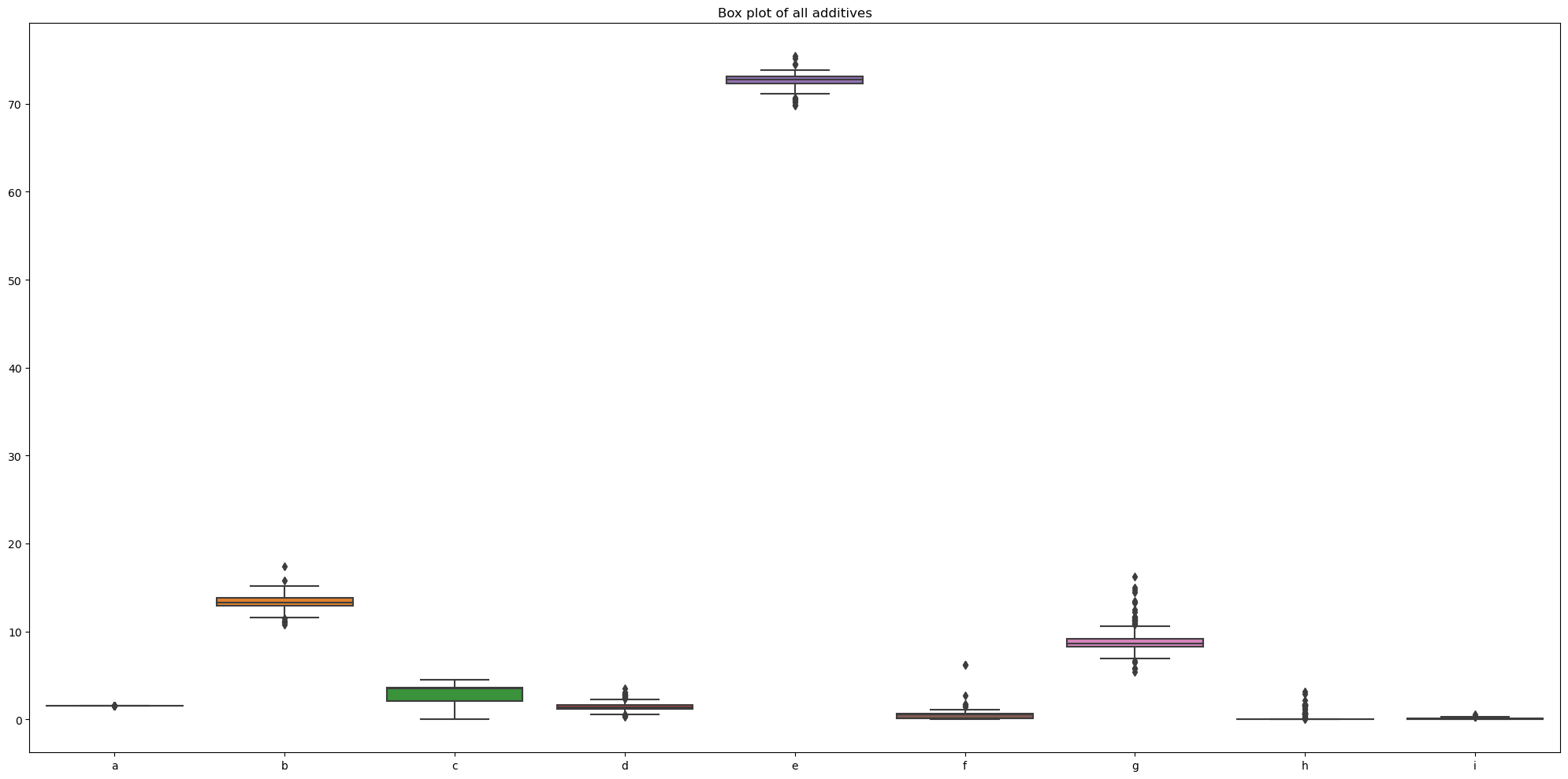
**2.1.9 Histogram of variable ‘i’**



* **Central Tendency:** The data predominantly clusters near 0.
* **Skewness**: The distribution has a very **high right skew**, with a long tail extending towards higher values.
* **Normality**: Kurtosis value of 2.66 is less than 3, which typically suggests that the distribution has lighter tail than a normal distribution. However, the visual impression of the graph and the high skewness contradict this somewhat, showing that the **tail is long but not heavy**.
* **Outliers**: The right tail shows outliers at higher values, but the moderate kurtosis of 2.66 indicates these are not overly pronounced or numerous.

**2.2 Box plot**

**2.2.1 Box plot for all additives**



In this box plot, we can see a comparison of different variables (from 'a' to 'i') based on their values. Here are some key observations:

* Variable 'e' has the highest median value among all, as shown by the line in the middle of the box, and it also has higher overall data points spread (the highest point near 70).
* Variable 'a' and 'i' have very low median values and fewer data points, indicated by the short horizontal line and very few points.
* Variable C is more consistent, and variable (g, d, and f) have more extreme value, and variable such as (a, h, and i) have more uniform data.

**2.2.2 Box plot for ‘a’ additive**

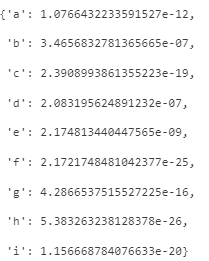
|  |  |
| --- | --- |
|  | **Median**: Centred around 1.52.  **Spread**: Narrow IQR from approximately 1.515 to 1.520, indicating **low variability**.  **Outliers**: Multiple outliers present both above and below the main data cluster, with more significant values towards the upper range.  **Skewness**: Slightly **right-skewed** due to higher outliers. |
|  | **Additive B:**  Most values cluster around 13, but there are outliers as high as 17.38. The box plot shows a wide spread, reflecting **moderate variability**. Its skewness suggests a slight tilt towards higher values. |
|  | **Additive C**: Values are more uniform, ranging only from 0 to 4.49, with a median around 3.48. The box plot indicates **low variability** and is **slightly skewed left**, suggesting more values are concentrated at the higher end. |
|  | **Additive D:** The values range from 0.29 to 3.5, with a median at 1.36. Outliers suggest some extreme values. The distribution is **right-skewed**, indicating a concentration of lower values with some high extremes. |
|  | **Additive E:** This has a narrow spread from about 69.81 to 75.41, indicating consistent data with a few outliers. The box plot shows minimal skew, meaning the data is **fairly symmetrical** around the median. |
|  | **Additive F:** It displays significant skewness and kurtosis, indicating a **heavy tail of outliers**, despite most data clustering near 0.555. This could imply occasional but extreme deviations from typical values. |
|  | **Additive G:** This additive's values generally range between 5.43 and 16.19, with a dense central region indicated by the box plot around 8 to 10. The plot shows several outliers above 10, suggesting occasional values far from the norm. The data is **right-skewed**, hinting at more frequent higher values. |
|  | **Additive H:** Most of the data for this additive is zero, as indicated by the concentration at the base of the plot, but there are several outliers up to a maximum of 3.15. The data is heavily **right-skewed** due to these high outliers. |
|  | **Additive I:** Values are tightly packed near zero, ranging up to 0.51 with a median close to zero. The box plot shows few outliers, indicating a uniform distribution with **right skewness**. |

**2.3 Q-Q Plot**

|  |  |
| --- | --- |
|  | **Variable a:** The plot shows that the points form a curve rather than a straight line, indicating that the data is **not normally distributed**. Specifically, it appears to curve upwards, suggesting the presence of right-skewness or heavy tails on the right side. |
|  | **Variable b**: The points on this plot deviate from the line at both ends but follow it more closely in the center. This pattern might suggest some **light-tailed distribution or a distribution with fewer extreme values** than a normal distribution would have. However, the deviation is not as pronounced as in variable a. |
|  | **Variable c:** This plot reveals a significant deviation from the red line, especially in the center, where there is a plateau, and at the ends. This is indicative of a **distribution that is not normal**. The stepped nature of the central points suggests that the data might be discrete or binned, and the heavy clustering at the lower end indicates a strong skew or a distribution with a threshold effect. |
|  | **Variable d:** The plot indicates that the data starts to **deviate from normality as the values get higher**. The curve bending upwards away from the red line at the higher end suggests that the tail is heavier than that of a normal distribution (right-skewed). |
|  | **Variable d**: The plot indicates that the data starts to **deviate from normality as the values get higher**. The curve bending upwards away from the red line at the higher end suggests that the tail is heavier than that of a normal distribution (right-skewed). |
|  | **Variable f**: There is a **clear deviation from the normality** in the central portion of the data, where the points form a step-like pattern, which is quite characteristic of discrete data or data with a lot of repeated values. Additionally, there are outliers at the higher end, as seen by the points that sit far from the red line. |
|  | **Variable g**: The points deviate significantly from the red line, especially at the higher quantiles. The upward curve at the end indicates a **right-skewed distribution** with a long tail on the right side. This suggests that variable g is **not normally distributed** and has **more high-value outliers** than would be expected in a normal distribution. |
|  | **Variable h**: The plot has a distinctive S-shape with a plateau at the center, suggesting that the values in the middle of the dataset are clustered together more than expected for a normal distribution. This might indicate a **bimodal distribution** or some other **non-normal characteristic**. It's also showing some extreme values at the higher end, deviating from normality. |
|  | **Variable i:** This Q-Q plot shows a **stepped pattern**, particularly towards the extremes, which suggests discrete data or data that have been rounded or binned. The steps at the lower end of the data are notably far from the line, indicating that there are more lower values than expected in a normal distribution. This is not a pattern typical of continuous data, and thus variable i is **not normally distributed.** |

**2.4 Normality Test (Shapiro-Wilk test)**

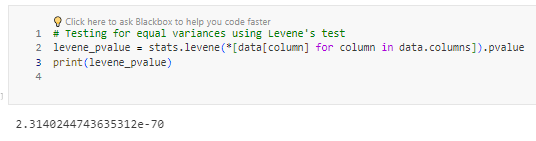
In this section, we will use Shapiro-Wilk test to check the data for each additive if is normal distributed.



The Shapiro-Wilk test results show **very low p-values** for all additives. A low p-value suggests that we **reject the null hypothesis**, which stated that the data is normal distributed. Therefore, the data for each additive is **not normally distributed**.

**2.5 Levene’s test**

The Levene’s test is used to access the equality of variances across groups.

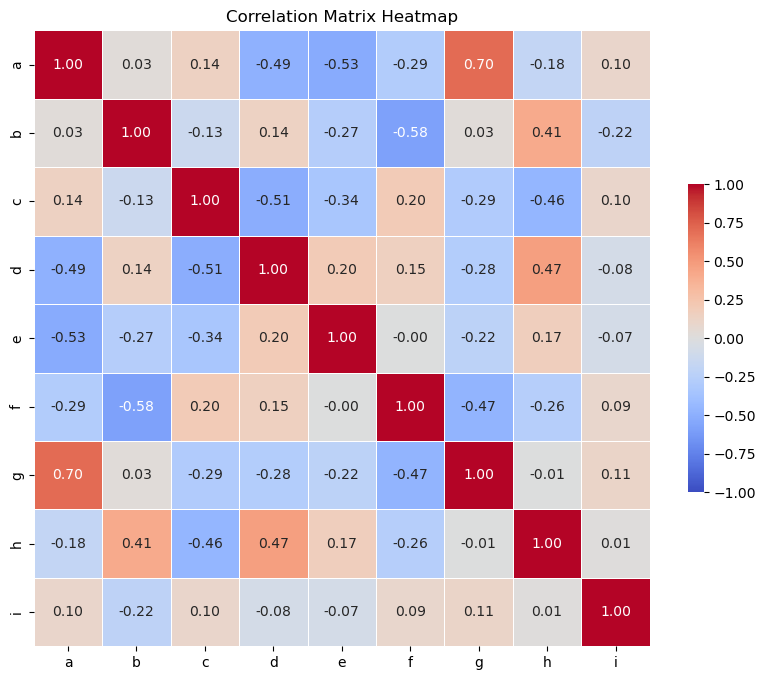


A very small p-value from the Levene's test indicate that the assumption of equal variances is also violated. It indicates that there are **significant differences in the variances across the groups**.

Given that both assumptions for ANOVA are not met, proceeding with a standard ANOVA would not be appropriate because ANOVA results could be unreliable under these conditions.

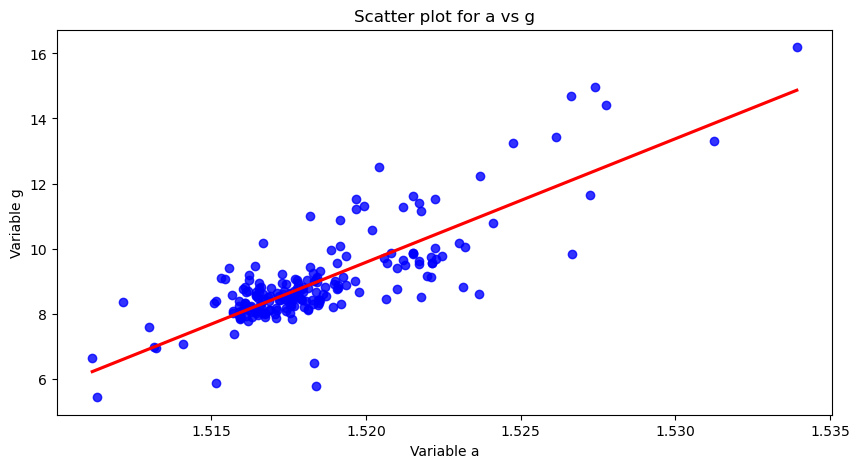
**3.0 Correlation analysis**

In this step will calculate the correlation between each pair of additives. This helps to understand the relationships between different additives. Since we found the data to be **non-parametric in previous step**, we will proceed with correlation analysis to understand the relationship between different additives. We will use **Spearman's rank correlation coefficient**, which is suitable for non-parametric data.



* **Strong Positive Correlation:** 
  + a and g (0.70): Suggest a strong positive relationship; as ‘a’ increase, so does ‘g’.
* **Moderate Negative Correlation:** When one variable increases, the other variable tends to decrease:
  + a and e (-0.53)
  + a and d (-0.49)
  + b and f (-0.58)
  + f and g (-0.47)
  + c and d (-0.51)
  + c and h (-0.46)
* Moderate Correlation: When one variable increase, the other variable tends to increase as well:
  + b and f (0.41)
  + d and h (0.47)
* **Weak Correlations:**
  + Most other pairs show weak or no correlation (both positive and negative), indicating little to no linear relationship.
* **Independence**:
  + ‘i’ shows almost no strong correlations with any other variables, suggesting it acts independently of others in this dataset.

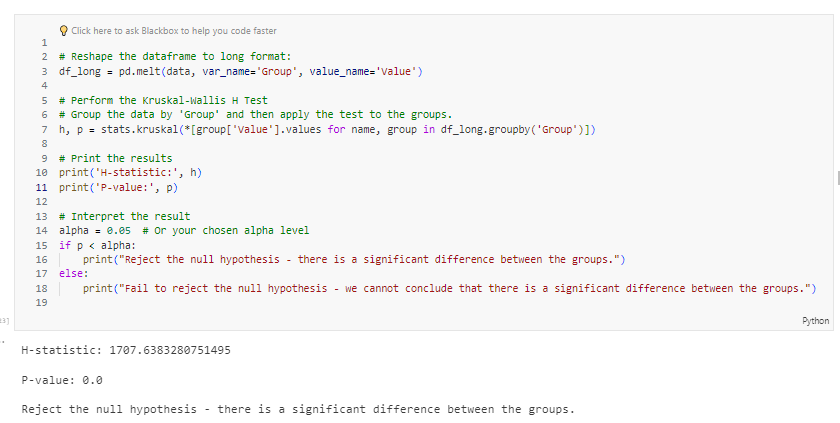
**3.1 Scatter Plot**

****

Scatter plot for variables **a vs g**: As discussed previously, the Spearman rank correlation coefficient of 0.70 indicates a strong positive correlation between these two variables. As variable a increases, variable g tends to increase as well, which is visually supported by the upward trend shown by the red regression line in the scatter plot

**4.0 Kruskal-Wallis H Test**

Given the violation of both normality and homogeneity of variances, a non-parametric alternative to ANOVA should be used. Non-parametric tests are more robust against violations of these assumptions. Hence, Kruskal-Wallis H-test is used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. It does not require the assumption of normal distribution or equal variances.



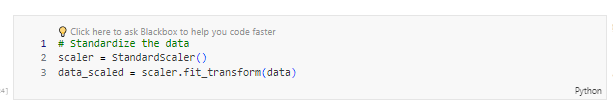
Since the p-value is much less than the alpha level of 0.05, we **reject the null hypothesis**. This suggests that there is a **statistically significant difference** between the median ranks of at least one pair of groups in the dataset. In other words, it supports the conclusion that there are significant differences between the levels of additives.

**5.0 Clustering**

Cluster Analysis helps to identify distinct groups of formulations based on the concentration of additives. This method can help us understand if there are naturally occurring clusters within the data, which might suggest different formulation group with unique characteristics.

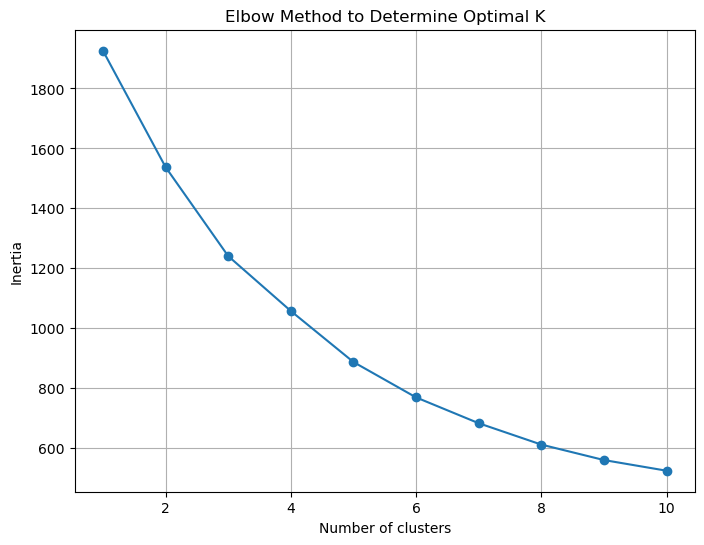
**5.1 Standardize the data**

First, we'll standardize the data because clustering algorithms like K-Means are sensitive to the scales of the data points.



**5.2 Elbow Method**

Elbow method is used to determine the optimal number of clusters.



From the plot, it seems that the elbow could be **around k = 3 or k =4**, as the rate decrease in inertia slows down notably after these points. This indicates that having three or four clusters might effectively capture the inherent groupings within the data.

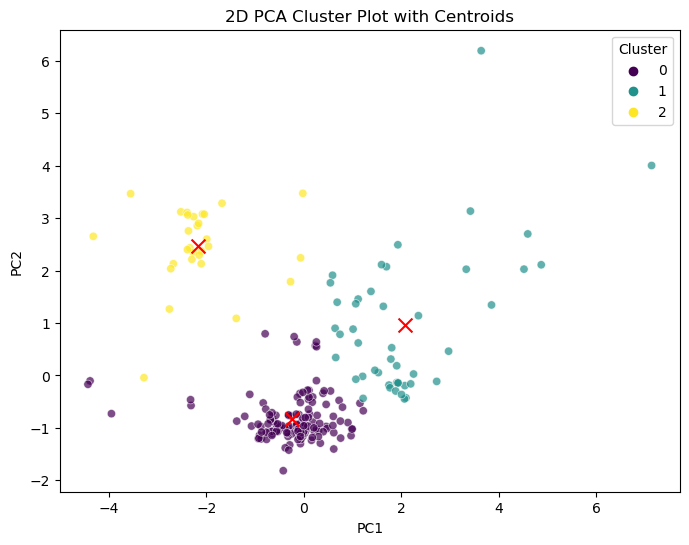
**5.2.1 Silhouette Score**

As there are no true labels available for the data, we will use the silhouette score to assess the quality of our clustering. The silhouette score measures how well-defined the clusters are, helping us evaluate the effectiveness of our clustering algorithm without relying on external labels.



With a small difference in silhouette scores between the two options (0.37 for 3 clusters and 0.39 for 4 clusters), both solutions perform similarly. I've chosen to go with **3 clusters** for simplicity.

**5.2.2 PCA Cluster Plot with Centroids**



The scatter plot above shows the result of clustering the formulations into three groups, visualized in a reduced two-dimensional space using Principal Component Analysis (PCA). Each point represents a formulation, and the colour indicate the different cluster.

**Observation:**

1. **Clusters and Spread**: The data points are grouped into three clusters (0, 1, and 2), which are indicated by different colours. The spread along both principal components suggests that PCA was effective in capturing the variance and separating the clusters to some extent.

2. **Centroids**: The red 'X' markers represent the centroids of each cluster. The centroid is the mean position of all the points in a cluster, essentially the "centre" of that cluster in the reduced PCA space.

3. **Cluster Separation**: Cluster 0 (in purple) and Cluster 1 (in aqua) are distinct and separate from each other, with their centroids located within the dense areas of their respective clusters. Cluster 2 (in yellow) is spread out and somewhat overlaps with Cluster 0, suggesting that the distinction between these two clusters is less clear.

4. **Relative Sizes and Densities**: Cluster 0 appears to be denser and possibly has more points than Cluster 1, which is a bit more spread out. Cluster 2 has the largest spread, indicating a wider range of variation within this cluster.

5. **Potential Outliers**: There are a few points in Cluster 1 that are distant from the rest, which could be potential outliers or points that don't fit as neatly within the defined clusters.

The clusters seem to be distinctly separated, which suggests that there are indeed different groups of formulations with unique characteristics based on their addictive concentrations.

This separation could imply that different clusters might have different burning patterns or other performance characteristics, as influenced by their specific combinations of addictive. The overlap between Clusters 1 and 2 might warrant further examination to understand the characteristics that cause the points to be grouped together.

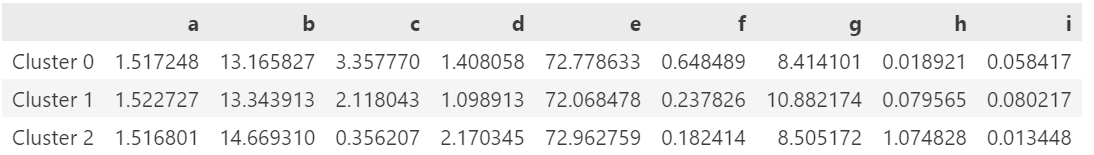
**5.2.3 Cluster Centres Analysis**

This section performs a detailed analysis by examine the mean values of additives at the cluster centres to identify what characterizes each cluster.

|  |  |
| --- | --- |
|  | **Variable a:**  Cluster 1 has a higher median and wider interquartile range compared to clusters 0 and 2, indicating more variability and generally higher values of "a" in this cluster. Cluster 0 and 2 have similar median values but different ranges and fewer outliers compared to cluster 1. |
|  | **Variable b:**  The plot suggests that the value of 'b' tends to be highest in cluster 2 and lowest in cluster 0. Cluster 1 and 2 show higher variability in their values compared to cluster 0. |
|  | **Variable c:**  The plot suggests that while clusters 0 and 1 have relatively more concentrated distributions of 'c' with less variability, cluster 2 shows a much wider spread of values indicating higher variability. |
|  | **Variable d:**  **Cluster 0** shows a tighter interquartile range, with fewer outliers, suggesting moderate variability.  **Cluster 1** has wider IQR but a slightly lower median than Cluster 0.  **Cluster 2** displays the highest median and variability with some high outliers. |
|  | **Variable e:**  **Cluster 0** shows a slightly wider range with some lower outliers.  **Cluster 1** shows a narrower range, centralized distribution, but with several outliers both low and high.  **Cluster 2**'s has the highest median value and a compact interquartile range, with a few outliers. |
|  | **Variable f:**  **Cluster 0's** shows a very narrow range close to zero, with multiple outliers.  **Cluster 1** shows a slightly wider range  **Cluster 2** tightly grouped around a median slightly above zero with higher outlier. |
|  | **Variable g:**  **Cluster 0's** distribution is concentrated with a lower median.  **Cluster 1** displays a much higher median and wider spread, indicating greater variability.  **Cluster 2** has a lower median similar to Cluster 0 but less spread. |
|  | **Variable h:**  **Cluster 0 and 1** show minimal to no values, with Cluster 0 presenting a couple of outliers.  **Cluster 2** shows a significant increase in median and range, indicating this cluster predominantly contains higher values of h. |
|  | **Variable i:**  **Cluster 0** has a lower median with some variability.  **Cluster 1** shows a higher median and slightly wider spread.  **Cluster 2** shows a very narrow distribution close to zero with few outliers. |

**5.3 Cluster Centre Analysis**

The cluster canters are shown in the table below, providing average values for each additive across the three identified clusters:



**Cluster 0:**

* Additive 'c': Has the highest average concentration at 3.357770.
* Additive 'f': Shows the highest value among the clusters at 0.648489.
* Additive 'h': Has the lowest average concentration at 0.018921, indicating lesser emphasis on the properties that 'h' influences compared to the other clusters.

**Cluster 1:**

* Additive 'g': The highest average concentration at 10.882174, possibly indicative of a focus on the characteristics enhanced by this additive.
* Additive 'i': Also shows a relatively higher average value at 0.080217.
* Additives 'c', 'f', and 'h': While not the lowest, they have moderate average values compared to the other clusters (2.118043 for 'c', 0.237826 for 'f', and 0.079565 for 'h'), which may suggest these characteristics are considered but not the primary focus.

**Cluster 3:**

* Additive 'h': The average value is significantly higher at 1.074828, suggesting a specialized focus for this cluster.
* Additive 'c': Has the lowest average at 0.356207, which might mean the properties affected by 'c' are not prioritized in these formulations.
* Additive 'i': With the lowest average value at 0.013448, it suggests minimal emphasis on the characteristics associated with 'i'.

These interpretations consider the relative average concentrations of additives within each cluster and are based on the provided data. Each additive's average concentration helps us infer which characteristics might be emphasized in the formulations of that cluster.

**6.0 Conclusion**

The statistical analysis verified that the formulations are distinct based on the concentrations of the additives. The cluster analysis not only grouped the formulations into statistically significant categories but also highlighted the unique characteristics of each cluster. These findings can effectively assist the third-party certification organization in recognizing and certifying the formulations based on their specific properties.