The dataset consists of 13 variables that have 54809 observations from different employees.

**Features:**

* employee\_id: Unique ID for employee
* department: Department of employee
* region: Region of employment (unordered)
* education: Education Level
* gender: Gender of Employee
* recruitment\_channel: Channel of recruitment for employee
* no\_ of\_ trainings: no of other trainings completed in previous year on soft skills, technical skills etc.
* age: Age of Employee
* previous\_ year\_ rating: Employee Rating for the previous year
* length\_ of\_ service: Length of service in years
* awards\_ won?: if awards won during previous year then 1 else 0
* avg\_ training\_ score: Average score in current training evaluations
* is\_promoted: (Target) Recommended for promotion

Practical Question: Should every company really need to invest more in programs where they recruit new employees to contribute to their success?

Statistical Question: Is there a significant difference in the Avg. Scores between the 3 groups from which they came from?

I chose the Recruitment Channel as the Independent variable, then the Average Training Score won for the dependent variable.

* **Missing Values**

In this dataset, fortunately, there are no missing values.

* **Descriptive Analysis & Visualization**

1. MEAN

* The **mean** represents the **average training score** for the entire dataset. In practical terms, the average training score for all employees across all recruitment channels is about **63.26**.
* This gives an overall idea of the general performance level of employees on their training assessments.

1. STANDARD ERROR

* The **standard error** measures the accuracy of the sample mean estimate of the population mean. A low standard error (0.0875) indicates that the sample mean is a precise estimate of the population mean.
* This suggests that the calculated mean (63.26) is a reliable estimate of the true mean training score.

1. MEDIAN

* The **median** is the middle value when all training scores are ordered from lowest to highest. Here, it is **60**, indicating that half of the training scores are below 60 and half are above.
* The median being lower than the mean suggests that there may be some **higher values** pulling the mean upward, indicating a slight right skew.

1. MODE

* The **mode** is the most frequently occurring value in the dataset, which is **50**. This means that **50** is the most common training score among employees.
* Knowing the mode helps in understanding the most typical value and could indicate a performance threshold that a large number of employees are hitting.

1. STANDARD DEVIATION

* The **standard deviation** measures the spread of the training scores around the mean. A standard deviation of **13.41** suggests that there is a **moderate spread** in the training scores.
* In practical terms, most employees' training scores fall within approximately **13 points above or below** the mean (63.26), indicating some variation in performance levels.

1. SAMPLE VARIANCE

* The **sample variance** is the square of the standard deviation, which is **179.87**. It represents the spread of the data points. Higher variance indicates greater diversity in the training scores.
* In practical terms, this tells us that there is some variability in training performance among employees, but it is not extremely high.

1. KURTOSIS

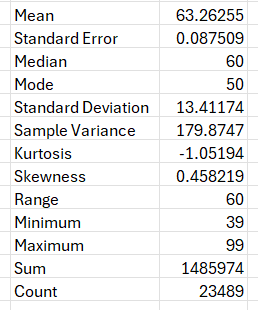
* **Kurtosis** measures the "tailedness" of the distribution. A kurtosis of **-1.05** suggests a **platykurtic distribution**, meaning the data has **lighter tails** and is **flatter** than a normal distribution.
* In practical terms, there are **fewer extreme values** (outliers) in the training scores than would be expected in a normal distribution.

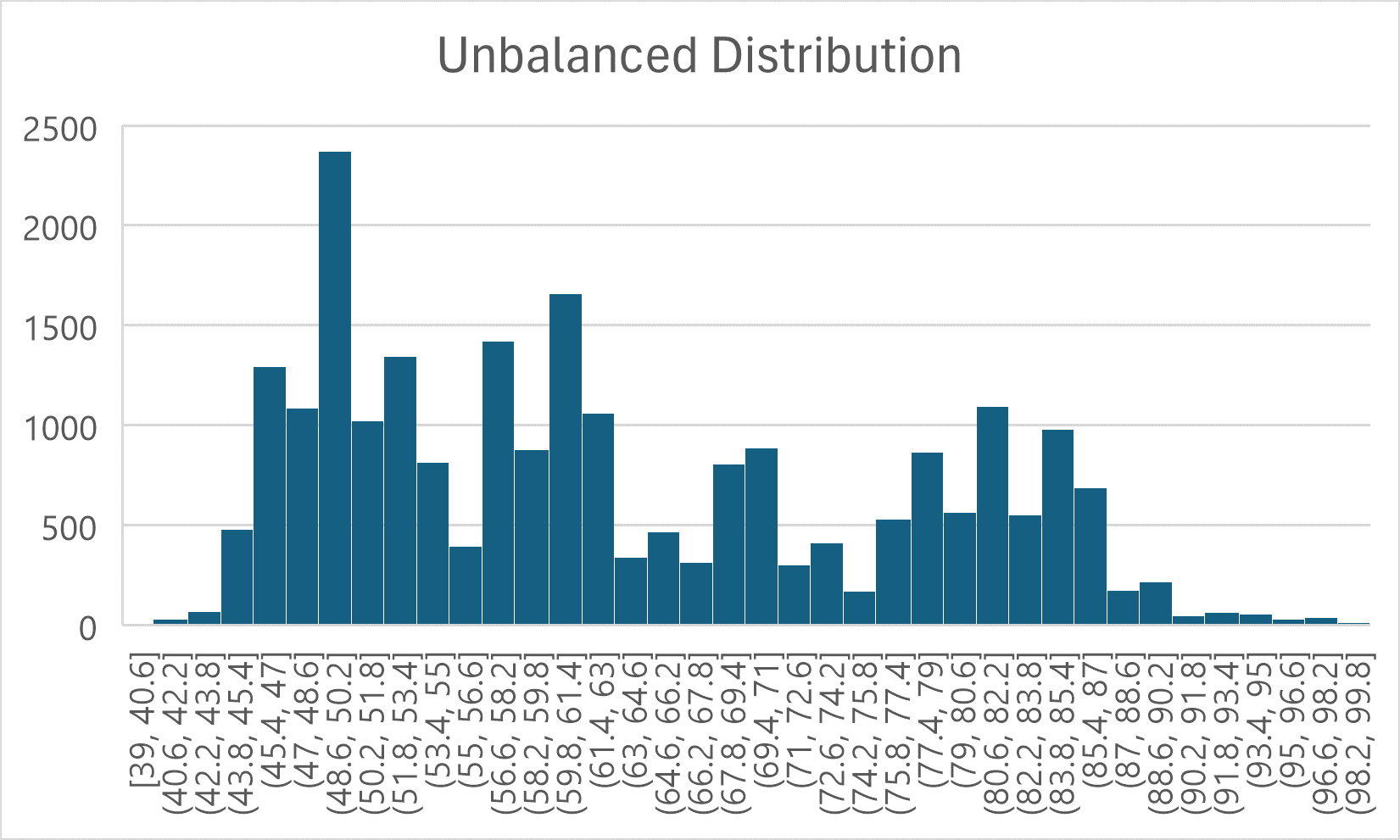
1. SKEWNESS

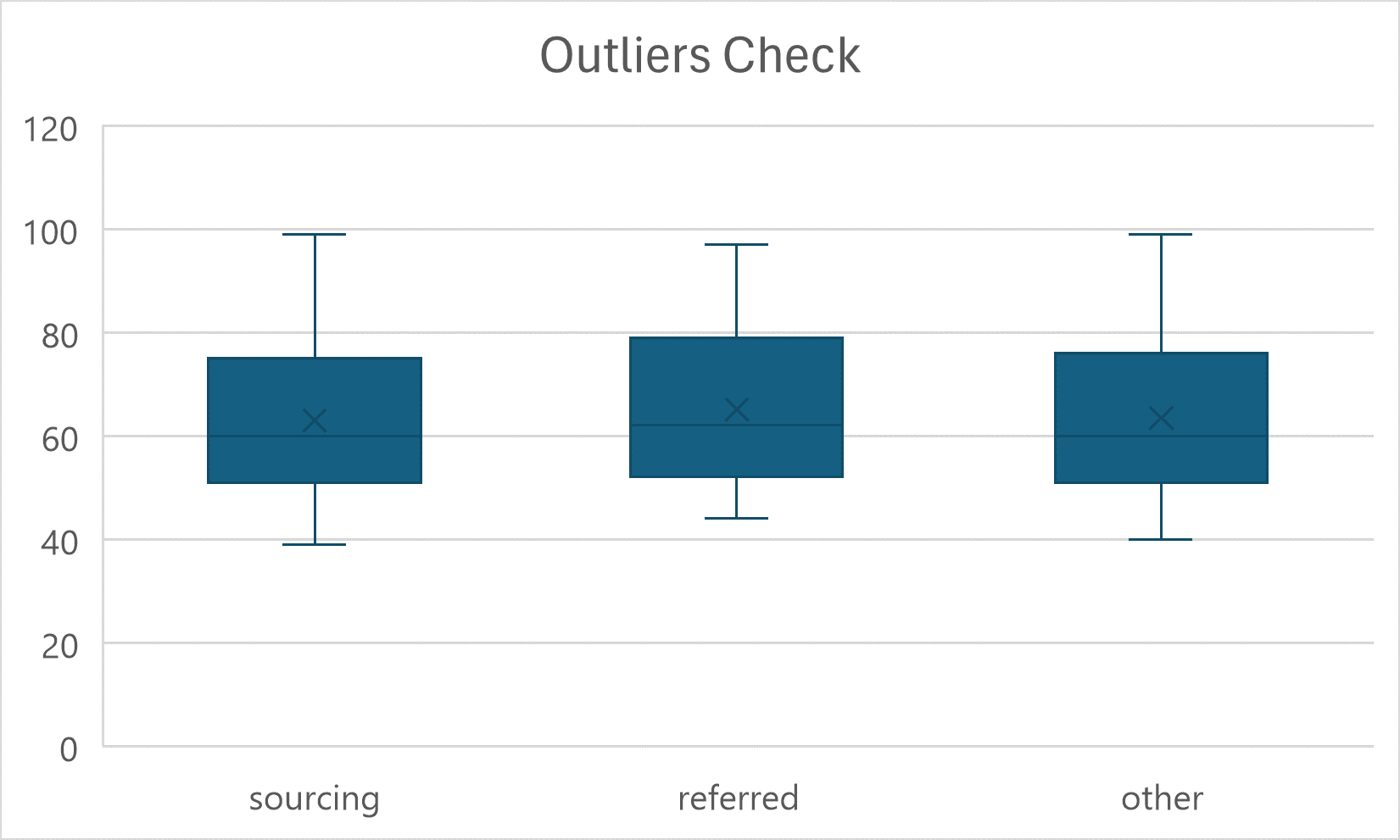
* **Skewness** measures the asymmetry of the data distribution. A skewness of **0.46** indicates a **slight positive skew**, meaning the tail on the right side is a bit longer than the left.
* In practical terms, this suggests that there are slightly more employees with **higher training scores** above the mean than below.

1. RANGE

* The **range** is the difference between the maximum and minimum values in the dataset (99 - 39 = 60).
* A range of **60** suggests a wide spread between the lowest and highest training scores, indicating variability in employee performance.







**Employee Performance Overview**: The mean score (63.26) and standard deviation (13.41) suggest that while most employees perform around the average, there is still a notable spread in training scores, with some employees excelling and others lagging.

**Training Program Effectiveness**: The relatively high variability in training scores could suggest differences in training delivery, comprehension, or individual abilities. Organizations may want to investigate further to understand the reasons behind these differences.

**Identifying Focus Areas**: Since the mode is **50** and the median is **60**, there might be a need to provide additional support or training resources for employees who are not performing as well.

**Skewness and Kurtosis Insights**: The slight positive skewness and negative kurtosis suggest that most training scores are close to the mean, but there are more high scores than low scores. This could indicate an overall positive outcome of the training programs but might also highlight areas for improvement to bring lower performers up to the mean.

**Further Analysis**: The insights from these descriptive statistics provide a basis for more targeted analysis, such as looking into specific groups, training methods, or employee characteristics that could influence these scores.

**Statistically Significant Difference Between Groups**:

* The **P-value (0.000229)** is much **smaller** than the common significance level of **0.05**.
* This indicates that there is a **statistically significant difference** between the average avg\_training\_score of at least one of the groups (sourcing, referred, other).
* In practical terms, this means that the **recruitment channel** has a significant impact on the **training performance** of employees.

**Decision to Reject the Null Hypothesis**:

* The **null hypothesis (H0)** in ANOVA states that all group means are equal.
* Given the **P-value < 0.05**, we **reject the null hypothesis**, suggesting that at least one group's mean training score is significantly different from the others.
* This conclusion can lead to further investigation to understand which specific groups differ from each other.

**Implications for Recruitment and Training Strategies**:

* The mean training scores differ slightly between groups:
  + **Referred** candidates have a slightly higher average training score (**65.09**) compared to **sourcing** (**62.94**) and **other** (**63.45**).
* This suggests that candidates who come through **referral** programs might perform better in training, indicating that referral programs could be more effective in finding suitable candidates.
* Organizations might want to **invest more** in **referral programs** or analyze why referred candidates perform better to replicate similar strategies for other recruitment channels.

**Analyzing Variance Within Groups**:

* The variances within each group (ranging from **178.62** to **184.46**) are similar, indicating a consistent spread of training scores within each group.
* This shows that, although the mean scores differ slightly, the variability within each group is quite uniform, suggesting that other factors beyond recruitment channels could influence training performance.

**Potential Action Points**:

* **Review Referral Programs**: Consider what aspects of the referral process might contribute to higher training performance and whether these aspects can be implemented in other recruitment methods.
* **Tailor Training Programs**: If significant differences between groups are confirmed, it may be beneficial to tailor training programs according to the recruitment channel to address the specific needs of different types of recruits.
* **Recruitment Strategy Adjustments**: Given the differences in performance, the organization might want to adjust its recruitment strategy to focus more on channels yielding higher training outcomes.