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 on the training for their employees?

Statistical Question: Is there a significant difference in the average training scores between employees who have won awards and those who have not?

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

* **Missing Values**

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

* **Descriptive Analysis & Visualization**

1. MEAN

* **Practical Implication**: The average training score for employees who have not won awards is **63.39**. This serves as a baseline for understanding the overall training performance in this group.

1. Standard Error

* **Practical Implication**: The standard error is relatively small, suggesting that the sample mean (63.39) is a precise estimate of the true population mean for the training scores of employees without awards. The large sample size (54,807) contributes to this precision.

1. Median

* **Practical Implication**: The median score (60) is lower than the mean, suggesting a **right-skewed distribution** where more employees score below the mean, but there are some high scores pulling the mean up.

1. Mode

* **Practical Implication**: The mode, or most frequently occurring score, is **50**, which is lower than both the mean and median. This further supports the idea of a skewed distribution where a significant number of employees have lower training scores.

1. Standard Deviation

* **Practical Implication**: The standard deviation shows how spread out the training scores are around the mean. A standard deviation of **13.37** indicates moderate variability within this group.

1. Sample Variance

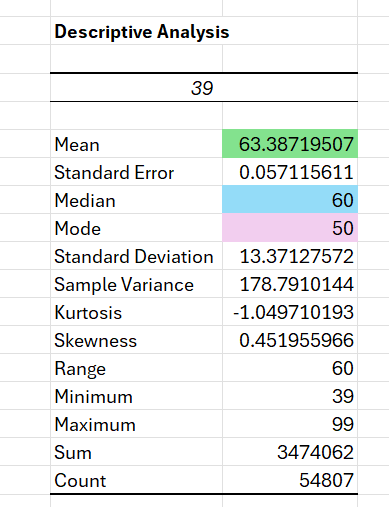
* **Practical Implication**: Variance measures the spread of the data points from the mean. A higher variance means more dispersion. This supports the idea that while the average score is **63.39**, scores can vary significantly.

1. Kurtosis

* **Practical Implication**: Negative kurtosis indicates a **platykurtic distribution**, meaning the distribution has lighter tails and a flatter peak compared to a normal distribution. This implies that extreme values (both low and high) are less likely than in a normal distribution.

1. Skewness

* **Practical Implication**: The positive skewness indicates that the distribution of training scores is **right-skewed**. This suggests that there are more low scores, but a few higher scores pull the mean to the right.

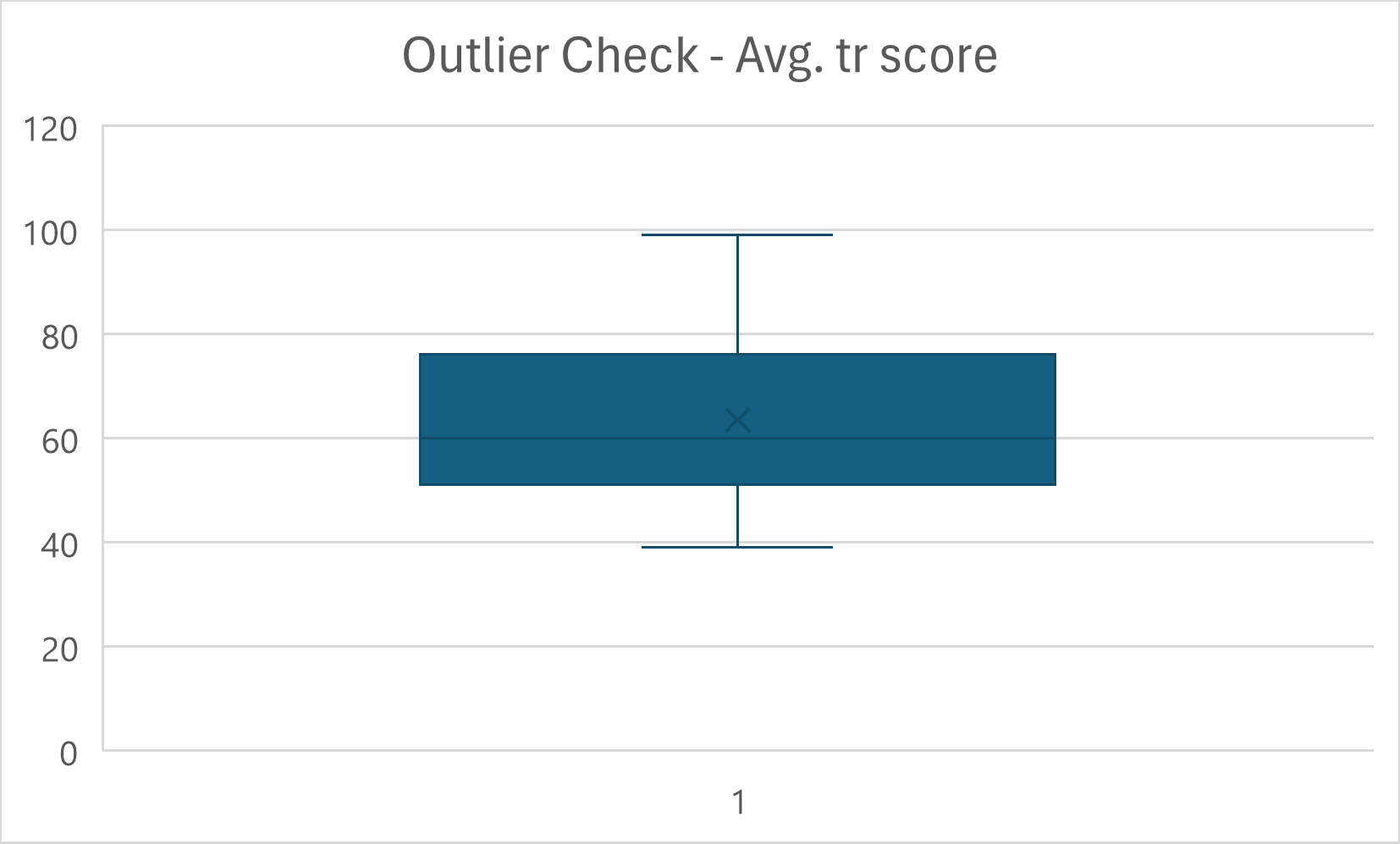


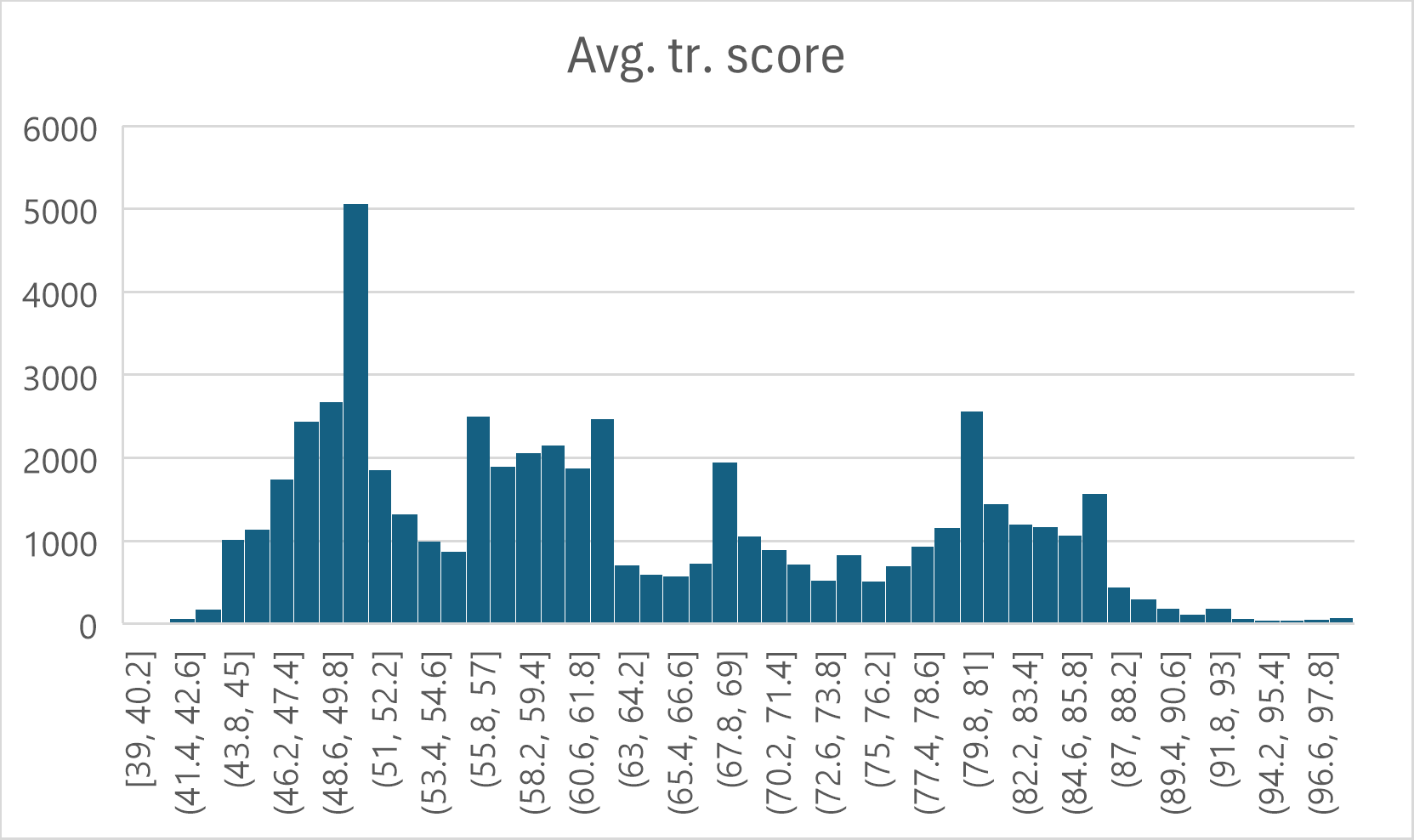
1. Skewness

* **Practical Implication**: The positive skewness indicates that the distribution of training scores is **right-skewed**. This suggests that there are more low scores, but a few higher scores pull the mean to the right.

The descriptive statistics show a skewed distribution of training scores among employees without awards, with a central tendency (mean) of around 63.39. The variability is moderate, and the distribution has a slightly skewed and flat shape, indicating that scores are not tightly clustered around the mean. These characteristics could suggest that only employees with substantially higher scores than the average are likely to win awards.

The descriptive statistics for employees without awards reveal a skewed distribution with variability, suggesting that while some employees have high training scores, a significant proportion have lower scores.



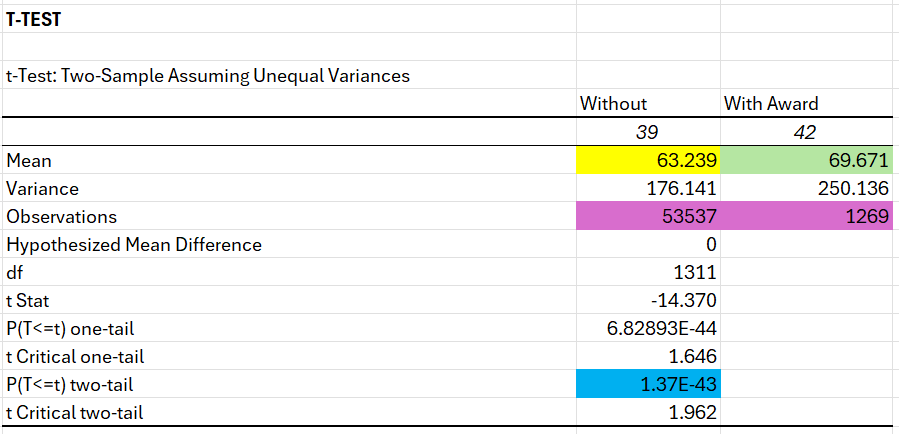


* T-Test Results: Two-Sample Assuming Unequal Variances

HYPOTHESIS:

1. NULL HYPOTHESIS which both variables are the same.
2. ALTERNATE HYPOTHESIS which are completely different to one another.

Alpha Value - 0.05 or 5%



1. MEAN

* **Practical Implication**: Employees who have won awards have a higher average training score (**69.671**) compared to those who have not (**63.239**). This suggests that, on average, award-winning employees perform better in training than non-award-winning employees.

1. VARIANCE

* **Practical Implication**: The higher variance among employees with awards (**250.136**) compared to those without (**176.141**) indicates that there is more variability in the training scores of award winners. This could mean that while some award winners perform exceptionally well in training, others do not perform as consistently, resulting in a wider spread of scores.

1. OBSERVATIONS

* **Practical Implication**: There is a significant difference in the sample sizes of the two groups, with a much larger number of employees who have not won awards (53,537) compared to those who have (1,269). This large disparity could influence the reliability and interpretation of the results, but the large sample sizes still provide robust statistical power for analysis.

1. Hypothesized Mean Difference

* **Practical Implication**: The null hypothesis assumes there is no difference in the mean training scores between the two groups (difference = 0).

1. DF

* **Explanation**: Degrees of freedom (df) refer to the number of independent values that can vary in the analysis.
* **Practical Implication**: Degrees of freedom are used to determine the critical values of the t-distribution needed to assess the statistical significance of the test result.

1. T-Stat

* **Explanation**: The **t-statistic** measures the size of the difference relative to the variation in the sample data. It tells us how many standard deviations the means are apart.
* **Practical Implication**: The large negative t-statistic value (**-14.370**) indicates that there is a substantial difference between the means of the two groups. The negative sign shows that the "Without Award" group has a significantly lower mean than the "With Award" group.

1. P-Value

* **Explanation**: The **p-value** represents the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true.
* **Practical Implication**: Both p-values (one-tail and two-tail) are extremely close to **0**, indicating that the difference in means is highly statistically significant. This means the likelihood of observing such a difference due to random chance is virtually zero. Therefore, we reject the null hypothesis and conclude that there is a significant difference in average training scores between employees with and without awards.

1. Critical Value

* **Explanation**: These are the cutoff points for determining statistical significance based on the t-distribution and degrees of freedom.
* **Practical Implication**: The critical t-values define the thresholds that the t-statistic must exceed to reject the null hypothesis at the chosen significance level (usually 0.05). Since the absolute value of the t-statistic (**14.370**) is much larger than both critical values, we can confidently reject the null hypothesis.

Conclusion:

NULL Hypothesis is rejected due to low P-Value, meaning that there are statistical differences between the two groups. Every company should heavily invest in training their employees for better performance.