Time interval (minuets) between past and present Geyser eruptions.

1. Dataset heads.
   - Past Eruptions

```
past_df data

   times
0     97
1     94
2     66
3     86
4     73
```

   - Recent Eruptions

```
recent_df data

   times
0     60
1     73
2     54
3     65
4     67
```

   - The problem is investigating potential changes in the time intervals between past and recent eruptions. We will make the assumption that the mean time interval between recent eruptions is represented by $\mu 1$, while the mean time interval between past eruptions is represented by μ_2.

2. Summary Statistics

Statistic	Recent	Past
Mean	69.58	77.33
Median	66.00	73.00
StDev	15.25	10.90
Q1	60.75	70.75
Q3	74.50	85.25
Max	107.00	97.00
Min	52.00	63.00

   - In both the recent and past distributions, the mean surpasses the median, implying the presence of extreme values exerting an upward influence on the mean. The recent eruption time intervals exhibit a greater spread, as evidenced by their larger standard deviation (15.25), wider range (55.00), and more extreme minimum (52.00) and maximum (107.00) values. Conversely, the past eruption time intervals are more tightly clustered, indicated by the smaller standard deviation (10.90), narrower range (34.00), and less extreme minimum (63.00) and maximum (97.00) values.

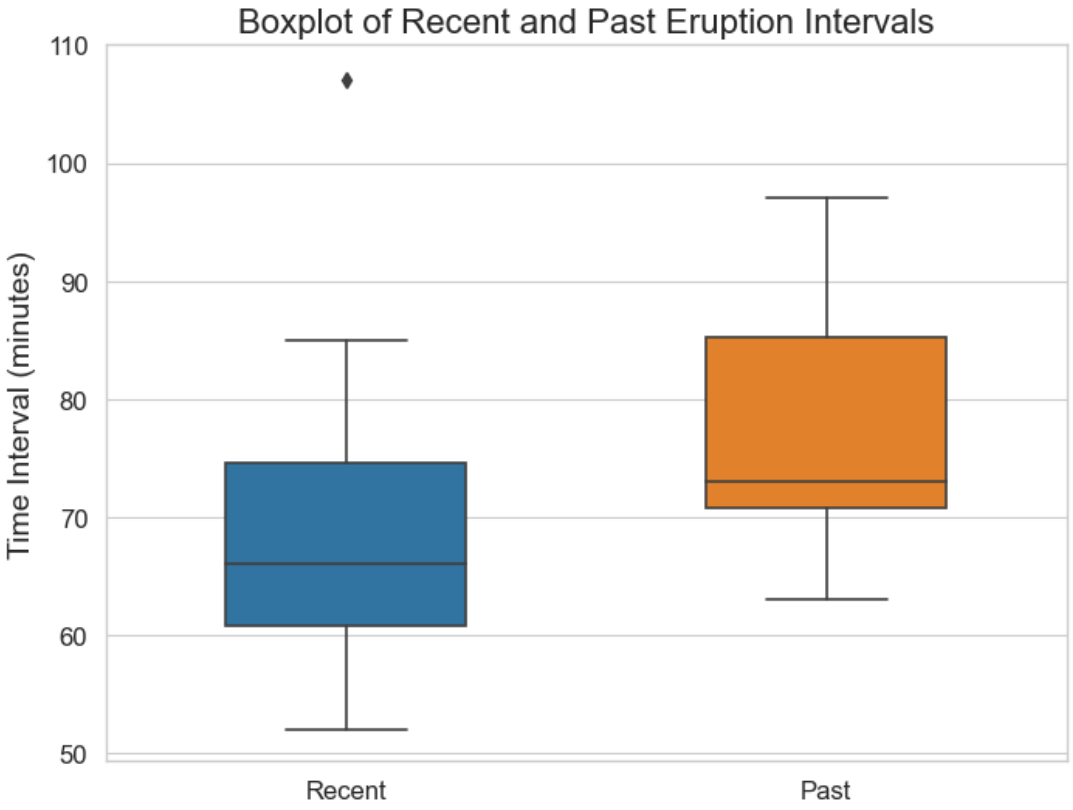
     However, it is crucial to highlight that the recent eruption distribution contains a single outlier with a value of 107.00. It is advisable to validate the legitimacy of this outlier to

ensure it is not an erroneous data point, as its presence significantly impacts the overall dataset.

3. Data Visualization

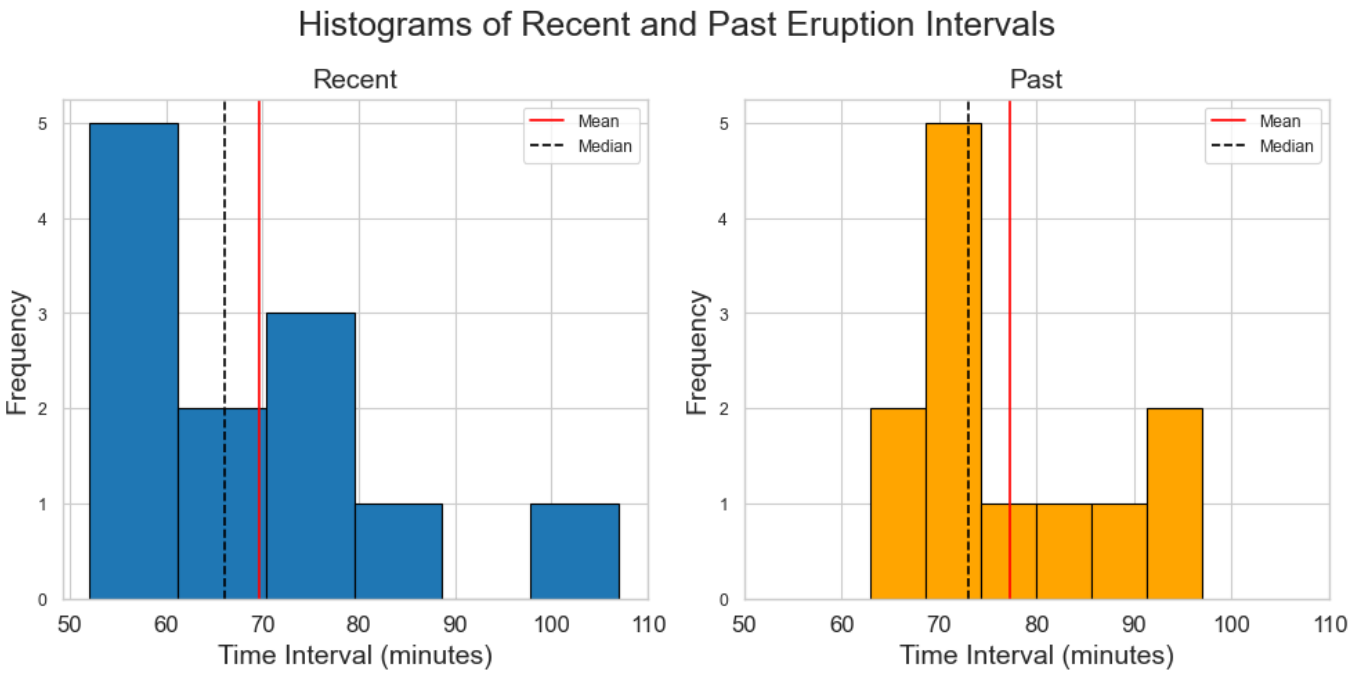
- The provided boxplot illustrates the datasets of time intervals between recent and past eruptions. In the boxplot representing recent eruption time intervals, a notable outlier is observed at a value of 107. Consequently, for the upper whisker, the outlier threshold was employed as the endpoint, while the minimum value was retained as the lower whisker endpoint.

  On the other hand, the dataset depicting time intervals between past eruptions does not exhibit any outliers. As a result, the whisker endpoints for this dataset were determined using the minimum and maximum values.



- The provided histogram provides a more detailed visualization of the distribution of time intervals between recent and past eruptions. Notably, the histogram portraying recent eruption time intervals exhibits a broader spread, while the one representing past eruption time intervals appears narrower in comparison.

  Furthermore, both histograms suggest a slight right skewness in the distribution of time intervals for recent and past eruptions. To quantify this skewness, Pearson's Second Coefficient formula was employed: (3 * (mean - median)) / standard deviation. This calculation yielded a skewness value of 1.216 for recent eruption intervals and 0.578 for past eruption intervals. Consequently, it can be inferred that the data for recent eruption time intervals is right-skewed, while the data for past eruption time intervals exhibits a mild right skewness.



- Both distributions share common characteristics, notably a right-skewed shape where the mean surpasses the median. Furthermore, their interquartile ranges (IQR) exhibit a slight disparity of merely 0.75 units. Specifically, the IQR for recent eruption time intervals stands at 13.75 units, while that for past eruption time intervals measures 14.5 units.

  However, a notable distinction arises when examining the range of these distributions. The range for recent eruption time intervals spans 55 units, whereas the range for past eruption time intervals is notably narrower, spanning only 34 units. This discrepancy amounts to a substantial difference of 21 units between the two datasets.

- It is important to highlight that these datasets do not conform to the assumptions of a normal distribution, nor do they meet the criteria of having a sample size exceeding 30. In practical terms, opting for parametric testing in such a real-world scenario would not be advisable or appropriate.

4. Hypothesis Testing

Test 1: Alpha 0.05

Claim: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

Null Hypothesis: The mean time interval of recent eruptions is the same as the mean time interval of past eruptions.

Alternative Hypothesis: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

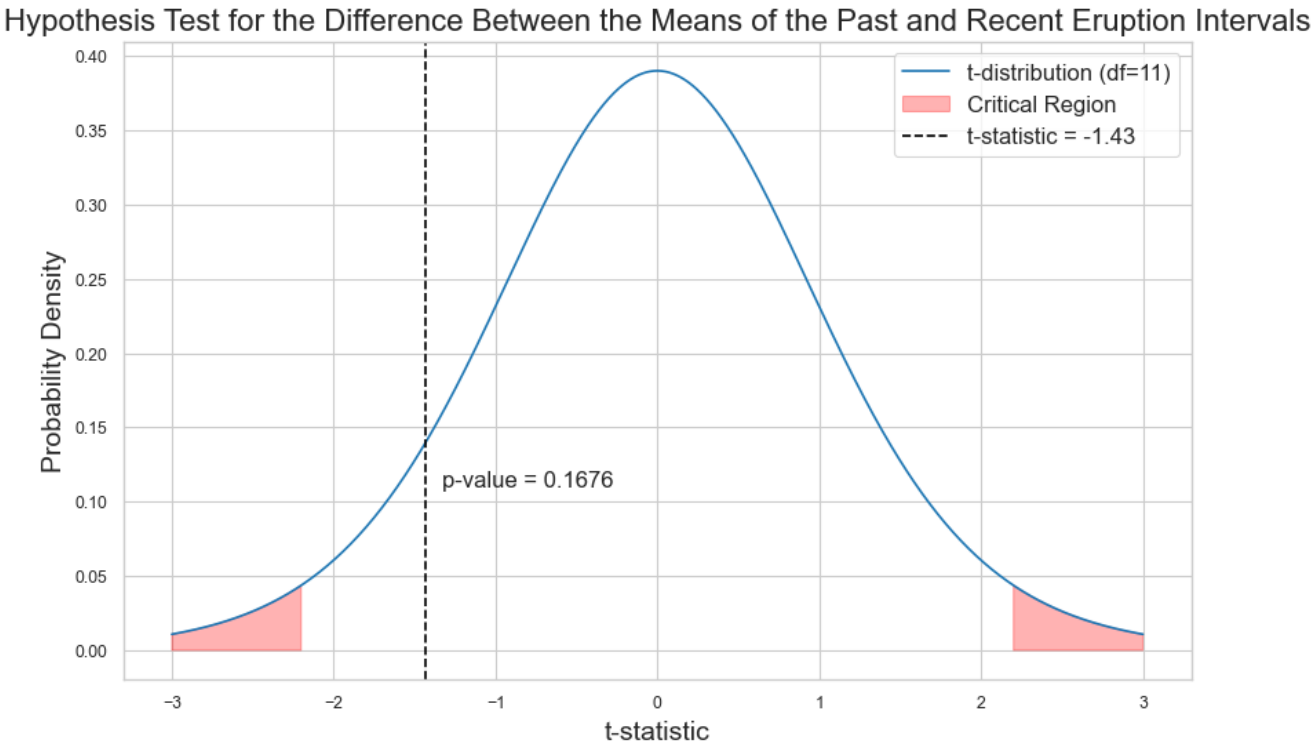
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

Hypothesis Test	
significance level:	$\alpha = 0.05$
t-statistic:	-1.4321
p-value:	0.1676

Conclusion: Fail to reject the null hypothesis.

There is not sufficient evidence to support the claim that the mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

- This conclusion is supported by the 95% confidence interval for the difference between the means of the past and recent eruption intervals. (-19.661, 4.161)



Test 2: Alpha 0.01

Claim: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

Null Hypothesis: The mean time interval of recent eruptions is the same as the mean time interval of past eruptions.

Alternative Hypothesis: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

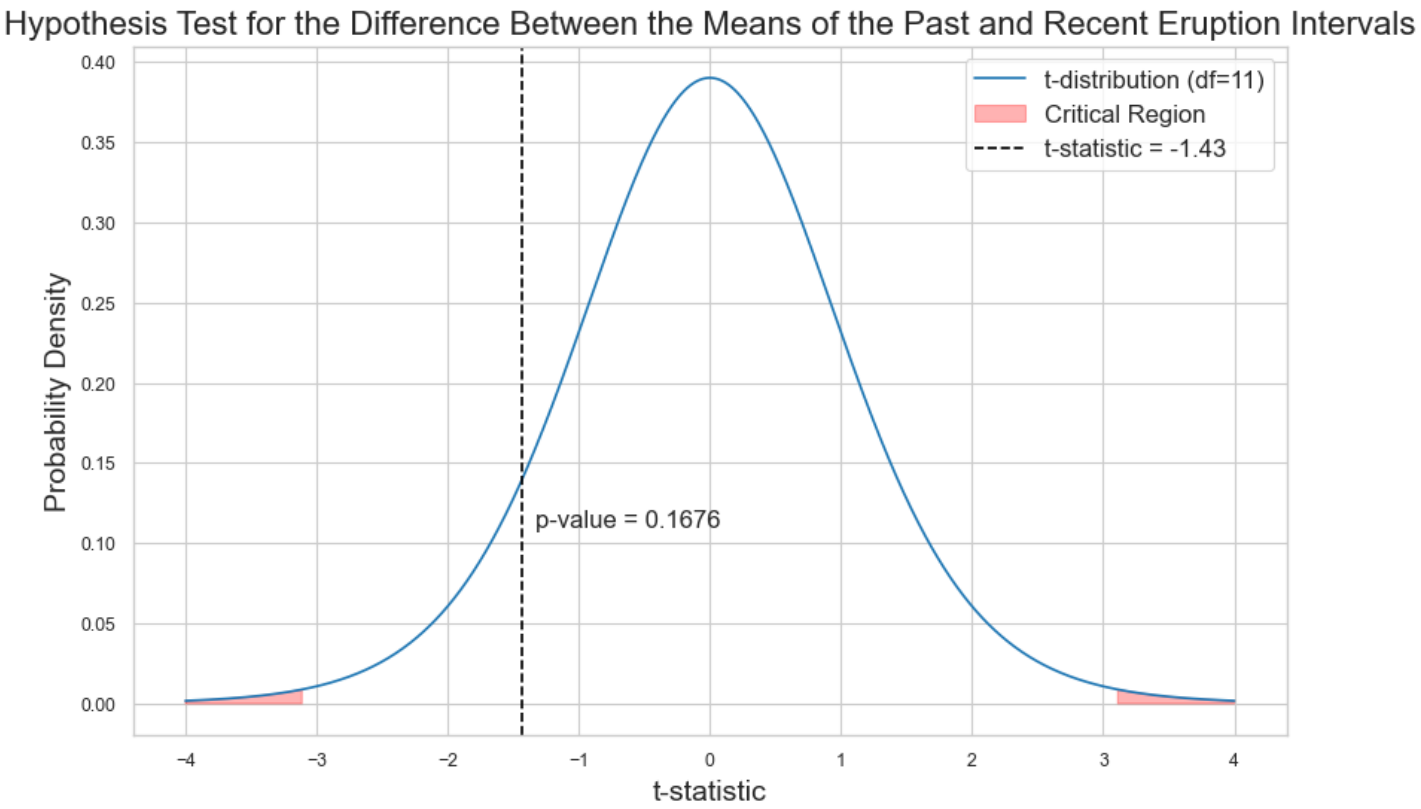
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

Hypothesis Test	
significance level:	$\alpha = 0.01$
t-statistic:	-1.4321
p-value:	0.1676

Conclusion: Fail to reject the null hypothesis.

There is not sufficient evidence to support the claim that the mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

- This conclusion is supported by the 99% confidence interval for the difference between the means of the past and recent eruption intervals. (-24.557, 9.057)



- In this instance, it's important to note that altering the significance level does not have an impact on the statistical conclusion. The significance level itself does not exert a direct influence on either the t-statistic or the p-value. Instead, it plays a pivotal role in establishing the threshold for determining statistical significance within the critical region.

  Given that the p-value resided beyond the critical region at a significance level of 0.05, reducing the significance level to 0.01 did not yield any alteration in the outcome of the hypothesis test.

Crucial Consideration: It is of vital importance to ascertain the validity of the outlier present within the recent eruption time intervals dataset, discerning whether it is a genuine data point or an error. The exclusion of this outlier exerts a profound influence on the overall analysis and the resulting outcomes of statistical testing.

Upon removal of the outlier, the standard deviation stands at 10.16, aligning much more closely with the standard deviation of 10.90 observed in the historical eruption time intervals.

As illustrated below, if it is established that the outlier is indeed erroneous and subsequently eliminated, the hypothesis test conducted at a significance level of 0.05 would yield significantly significant results.

Conversely, it is noteworthy that reducing the significance level to 0.01 does bear consequences in this particular scenario. This reduction in significance level leads to a narrowing of the critical region, causing the t-statistic to fall outside of said critical region. Consequently, this change shifts the outcome from rejecting the null hypothesis to failing to reject it.

Test 3: Outlier removed alpha 0.05

Claim: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

Null Hypothesis: The mean time interval of recent eruptions is the same as the mean time interval of past eruptions.

Alternative Hypothesis: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

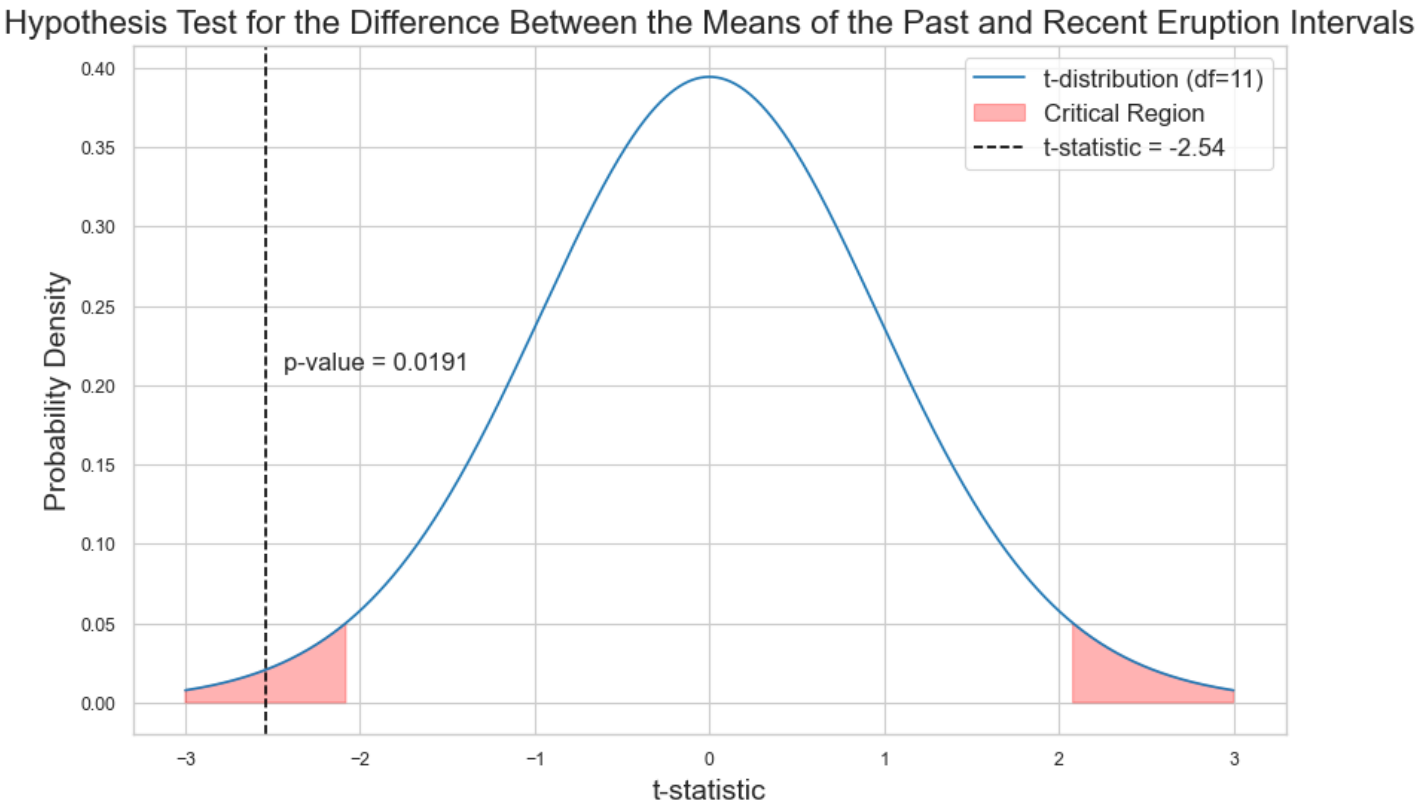
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

Hypothesis Test	
significance level:	$\alpha = 0.05$
t-statistic:	-2.54
p-value:	0.0191

Conclusion: Reject the null hypothesis.

There is sufficient evidence to support the claim that the mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

- This conclusion is supported by the 95% confidence interval for the difference between the means of the past and recent eruption intervals. (-20.934, -1.369)



Test 4: Outlier removed alpha 0.01

Claim: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

Null Hypothesis: The mean time interval of recent eruptions is the same as the mean time interval of past eruptions.

Alternative Hypothesis: The mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

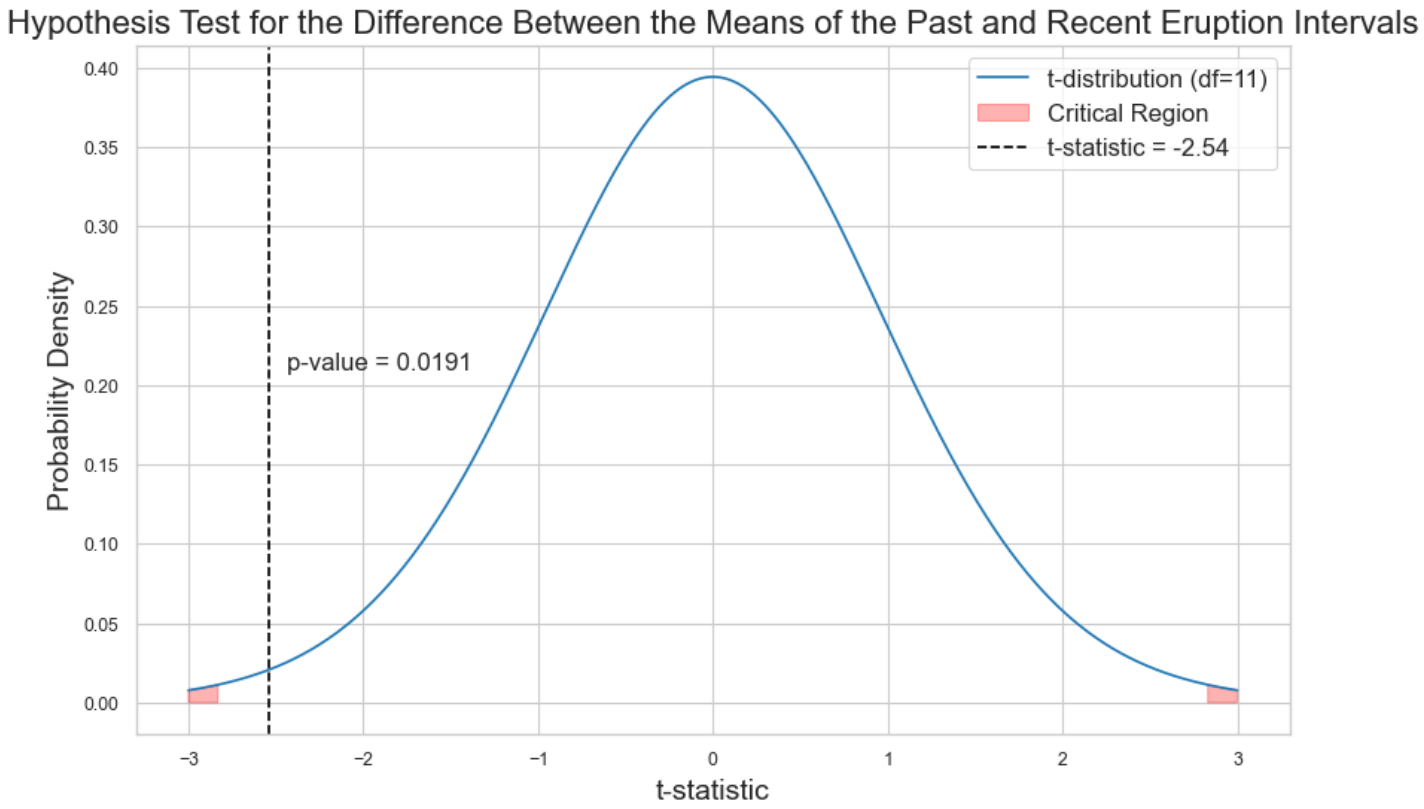
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

Hypothesis Test	
significance level:	$\alpha = 0.01$
t-statistic:	-2.54
p-value:	0.0191

Conclusion: Fail to reject the null hypothesis.

There is not sufficient evidence to support the claim that the mean time interval of recent eruptions is not the same as the mean time interval of past eruptions.

- This conclusion is supported by the 95% confidence interval for the difference between the means of the past and recent eruption intervals. (-22.067, 2.764)



In conclusion, to ensure accurate and dependable results, it is imperative to conduct testing with a significantly larger sample size. The findings obtained from the limited sample size are inconclusive, as explained below.

The outcomes of hypothesis testing conducted at significance levels of 0.05 and 0.01 using the original datasets have led to the consistent finding of failing to reject the null hypothesis. This indicates that there is insufficient evidence to

suggest a difference between the mean time intervals of recent and past eruptions. The construction of both 95% and 99% confidence intervals further bolsters this conclusion.

Conversely, when hypothesis testing was performed on datasets with outliers removed, a contrasting set of conclusions emerged. At a significance level of 0.05, the analysis led to the rejection of the null hypothesis, supporting the alternative hypothesis that a difference exists in the mean time intervals of recent and past eruptions. A subsequent construction of a 95% confidence interval aligned with this result.

However, at the more stringent significance level of 0.01, the hypothesis testing on outlier-excluded datasets reverted to the outcome of failing to reject the null hypothesis, signifying a lack of evidence for a difference in the mean time intervals of recent and past eruptions. This conclusion was further substantiated by the construction of a 99% confidence interval.