Deciphering User Engagement: Insights from Social Media Platform Data Analysis

- By Anisha Bhanudas Mogal (am6196@rit.edu)

STAT 614

Professor: Carly Metcalfe

FALL 2023

TABLE OF CONTENTS

INTRODUCTION.

Social media has emerged as an indispensable hub for contemporary communication, creating a multifaceted arena where individuals can connect, share content, and interact within communities. The comprehension of user conduct in these spheres holds utmost significance for refining platforms, making informed strategic choices, and enriching user participation. The available dataset encompasses a wide array of user centric factors from a social media platform, furnishing valuable perspectives on engagement rates within various demographics and patterns of usage.

Objectives:

The analysis focuses on examining user engagement behavior through evaluating various critical factors, including location, content preference, and usage habits. Utilizing thorough statistical analysis, the goal is to address essential queries concerning user engagement on the platform.

Questions to be answered:

1. **Do urban and rural users significantly differ in their mean engagement rates?**

Test Method: Two-Sample T-Test

Approach: By comparing mean engagement rates between urban and rural users to check if there is a significant difference between the two groups.

2. **How does daily platform usage in hours influence the engagement rate?**

Test Method: Simple Linear Regression

Approach: By fitting a regression line to understand the relationship between daily platform usage (in hours) and engagement rate, assessing its significance in predicting engagement.

3. **Is there a significant difference in mean engagement rates before and after the feature change?**

Test Method: Paired T-Test

Approach: By comparing mean engagement rates before and after a feature change to identify if there's a statistically significant difference in engagement following the feature introduction.

4. **Does the data collected have the same proportion of users from each age group?**

Test Method: Chi-square Test for Goodness of Fit

Approach: By performing a chi-square test to assess the observed proportions of users across age groups determining if the data represents equal proportions for each age group.

METHODS.

Dataset Description:

The dataset captures user engagement and activity on a social media platform, comprising various user-related variables:

- **User ID:** Unique user identifier.
- **Age Group:** Categorical variable indicating user age intervals.
- **Location:** Categorical variable representing user locations (Rural, Urban).
- **Platform Usage (hours):** Continuous variable denoting average daily platform usage in hours.
- **Post Frequency:** Continuous variable reflecting the number of weekly user posts.
- **Follower Count:** Continuous variable signifying the total count of user followers.
- **Engagement Rate:** Continuous variable expressing the engagement rate (likes, comments, shares) relative to the user's account.
- **Preferred Content:** Categorical variable indicating the user's content type preference.
- **Engagement Rate Before Feature**: Continuous variable representing engagement rates before a new feature introduction.
- **Engagement Rate After Feature:** Continuous variable representing engagement rates after a new feature introduction.
- **Average Likes per Post:** Continuous variable denoting the average likes per user post.

Response Variable: Engagement Rate

Total Number of Observations: 500 observations.

Hypothesis For Testing

Two-Sample T-Test (μ_2 and μ_1 are the mean engagement rate of urban and rural users respectively.)

H_o: The mean engagement rate of urban users is equal to the mean engagement rate of rural users.

$$(\mu_2 = \mu_1)$$

H_1: The mean engagement rate of urban users is greater than the mean engagement rate of rural users.

$$(\mu_2 > \mu_1)$$

Simple Linear Regression (where β_1 is the coefficient of platform usage in the regression equation):

H_o: Daily platform usage hours do not significantly influence the engagement rate.

$$(\beta_1 = 0)$$

H_1: Daily platform usage hours have a significant influence on the engagement rate.

$$(\beta_1 \neq 0)$$

Paired T-Test (μ_D = Average difference in engagement rating before and after a feature change):

H_o: The mean engagement rate before the feature change is equal to the mean engagement rate after the feature change.

$$(\mu_D = 0)$$

H_1: The mean engagement rate before the feature change is not equal to the mean engagement rate after the feature change.

$$(\mu_D < 0)$$

Chi-square Test for Goodness of Fit (Π_i = proportion of users for i^{th} age group):

H_o: The observed proportions of users across different age groups are equal to the expected proportions.

(Π_i = 1.6667 for all i = 1)

H_1: The observed proportions of users across different age groups are not equal to the expected proportions.

(At least one $\Pi_i \neq 1.6667$)

RESULTS

Level of significance α **= 0.05 .**

For all the tests to be performed, **Two-Sample T-Test** has normality assumption which means data should be approximately from a normal distribution. Thus, by plotting the normal quantile plot of engagement rates, we get the diagram (f(a)).

Inference: From figure f(a) by fat pencil method it can be said that the given data approximately follows normal distribution as most of the data points are along the straight line and fall inside the confidence interval.

1. **Two-Sample T-Test (Equality of Means):**

Prior to conducting a T test for mean comparison, it's crucial to first ensure the equality of variances within the compared groups. This assessment is instrumental in deciding whether to employ a pooled estimator in the T test statistic computation.

Here,

$H_0 : \sigma_2 / \sigma_1 = 1$

$H_1 : \sigma_2 / \sigma_1 \neq 1$

We use F-test to test the two variances.

$F_0 = 1.111$

Here, α **= 0.05** and P-value **0.4281** (from f(1.a))

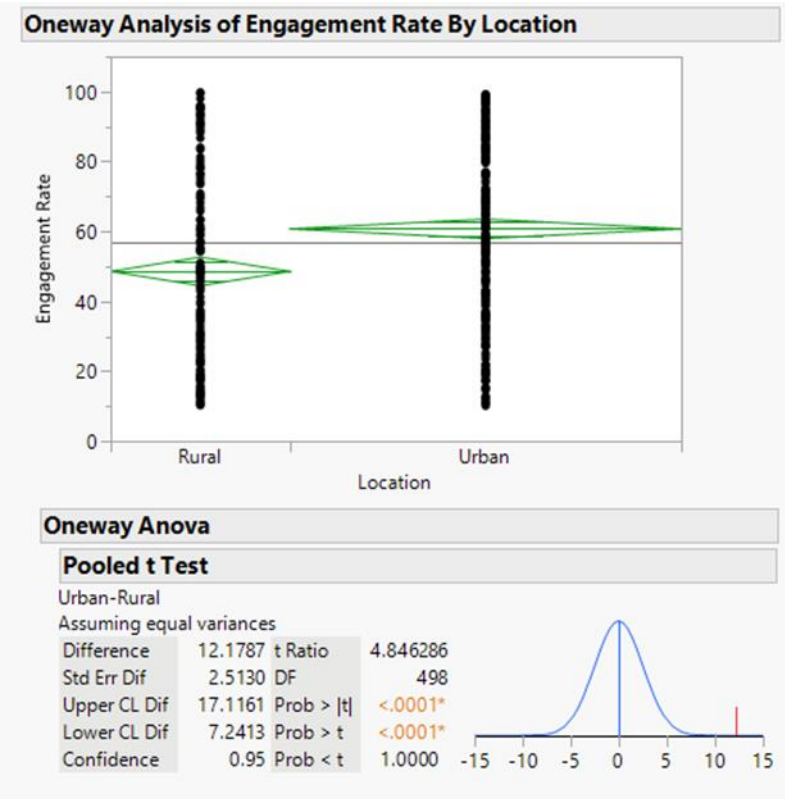
Decision:

We fail to reject null hypothesis as p-value > α.

We have insufficient evidence at 5% level of significance to claim $\sigma_2 / \sigma_1 \neq 1$.

Conclusion: Population variance can be assumed equal

We can use pooled estimator for the T-test.



f(1.b)

From f(1.b),

$T_0 = 4.8463$

Here, α **= 0.05** and P-value **< 0.0001**

The confidence interval for difference in means given in f(1.c) -> (7.2413, 17.1161) is greater than 0.

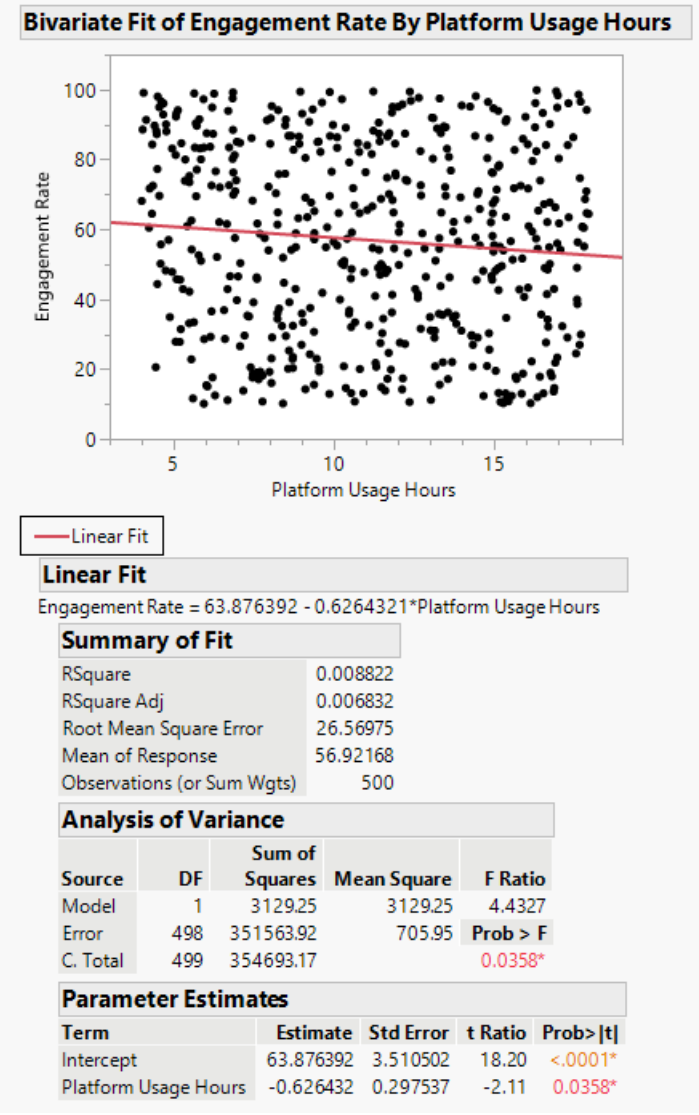
Thus, we can say that $\mu_2 > \mu_1$.

Inference:

We reject null hypothesis as p-value < α.

We have sufficient evidence at 5% level of significance to claim $\mu_2 > \mu_1$.

2. Simple Linear Regression



f(2.a)

From f(2.a),

The fitted line can be given by,

Engagement Rate =

63.8764 - 0.6264 * Platform Usage Hours

Where β_0 (intercept) = 63.8764 and β_1(coefficient of platform usage in the regression equation) = 0.6264

$T_0 = -2.11$

Here, α **= 0.05** and P-value **= 0.0358**

Inference:

We reject null hypothesis as p-value < α.

We have sufficient evidence at 5% level of significance to claim $\beta_1 \neq 0$.

From f(2.b) and f(2.c),

In the residual vs predicted plot, we can see a random scatter of points indicating no specific pattern. In the residual normal quantile plot by fat pencil method, it can be said that the residuals follow normal distribution. Thus, both the normality and constant variability assumption of the regression analysis is met.

3. Paired T-Test (Equality of Means):

D = Difference in engagement rating (after - before) feature.

From f(3.a),

$T_0 = 45.6369$

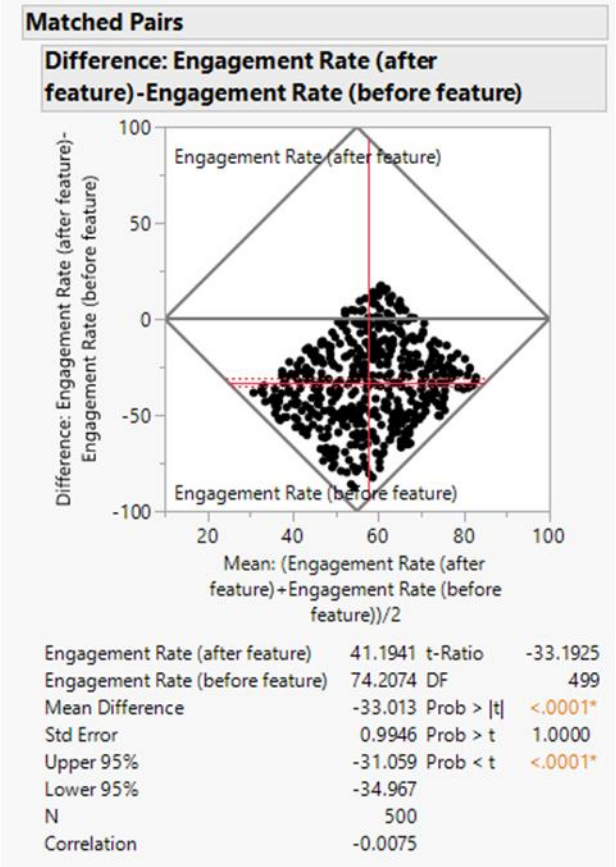
For one tailed test P ($T < T_0$),

Here, α **= 0.05** and P-value **< 0.0001**

The confidence interval for expected difference in f (3.b) -> (-34.967, -31.059) is less than 0. Thus, we can say that $\mu_D < 0$.

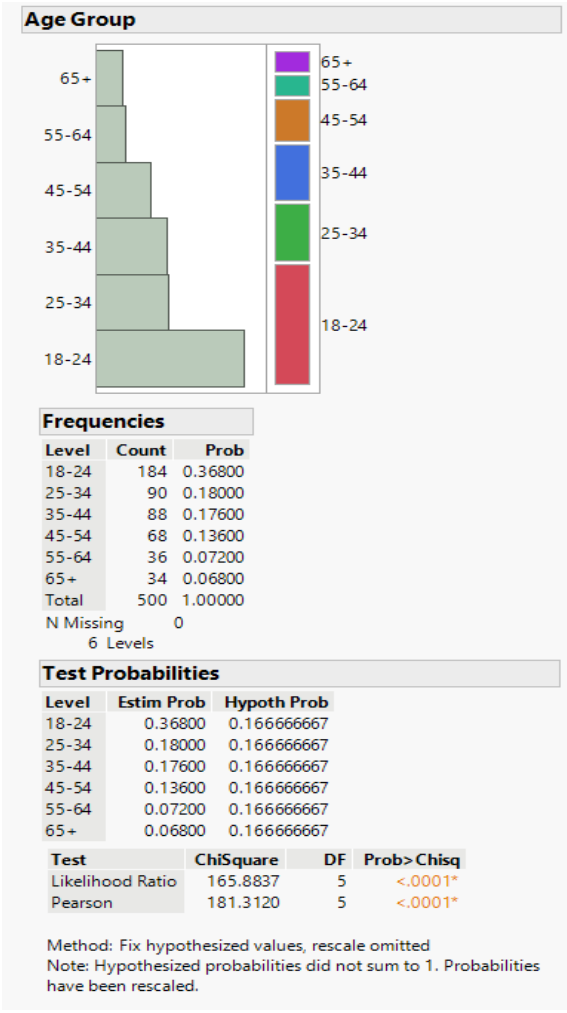
Inference:

We reject null hypothesis as p-value < α. We have sufficient evidence at 5% level of significance to claim $\mu_D < 0$.



f(3.a)

4. Chi-square Test for Goodness of Fit:



f(4.a)

From f(4.a),

$\chi_0 = 181.3120$

Here, α **= 0.05** and P-value **< 0.0001**

Inference:

We reject null hypothesis as p-value < α.

We have sufficient evidence at 5% level of significance to claim at least one $\Pi_i \neq 1.6667$.

The diagram also shows that the age group 18-24 has the highest proportion out of all the age groups while the age group 65+ has the lowest proportion.

Conclusions and Recommendations

Conclusions:

i. People living in urban areas have a mean engagement rate that is much higher than those in rural areas. It can be concluded that urban consumers interact at higher rates than rural users.

ii. Daily platform usage hours significantly influence the engagement rate. The regression analysis indicates that as platform usage hours increase, the engagement rate tends to decrease.

iii. There is a strong difference in mean engagement rates before and after the feature change. The engagement rate after the feature change is significantly lower than before the change.

iv. The observed proportion of users across different age groups is not equal. At least one age group's representation significantly deviates from the expected proportion.

Recommendations:

Engagement Strategies: To improve user engagement rates and close the engagement gap between urban and rural users, the platform should concentrate on strategies designed specifically for rural consumers.

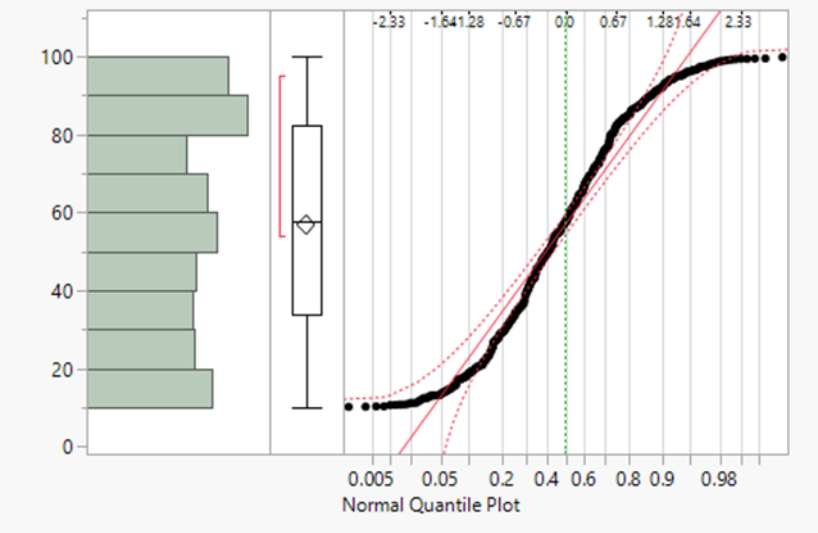
Usage Optimization: Implement measures to optimize user engagement concerning daily platform usage, considering that increased usage might lead to decreased engagement rates.

Feature Update Considerations: Review the recent feature change impact on user engagement. Implement modifications or additional features that can boost engagement to reverse the decrease in engagement post-feature change.

Age Group Targeting: Develop targeted strategies that appeal to age groups with significantly different representations to improve engagement and user experience within those specific areas.

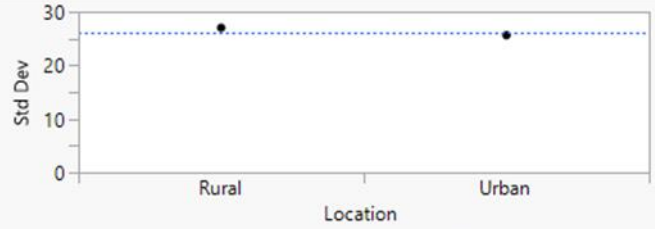
Through concentration on these analysis-derived recommendations, the platform can plan and customize its features, content, and engagement strategies to maximize user experience and increase overall engagement rates among various user groups.

APPENDIX



f(a)

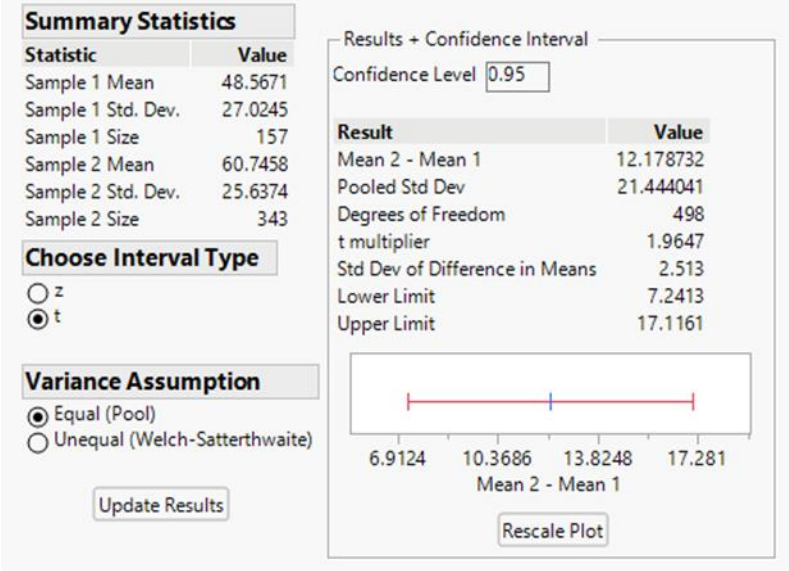
Tests that the Variances are Equal



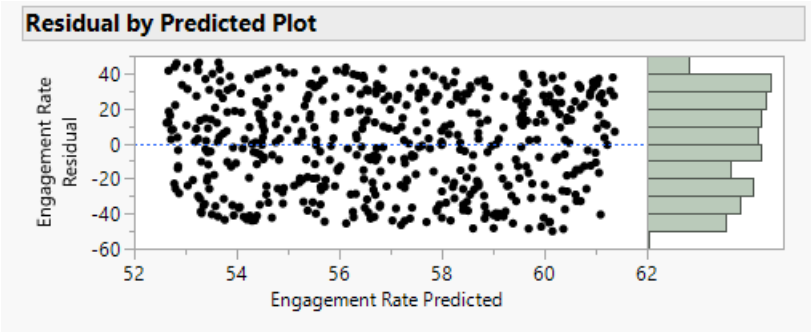
Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Rural	157	27.02453	22.87395	22.72121
Urban	343	25.63737	21.78610	21.68863

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	1.2934	1	498	0.2560
Brown-Forsythe	0.5692	1	498	0.4509
Levene	0.6769	1	498	0.4111
Bartlett	0.6011	1	.	0.4382
F Test 2-sided	1.1111	156	342	0.4281

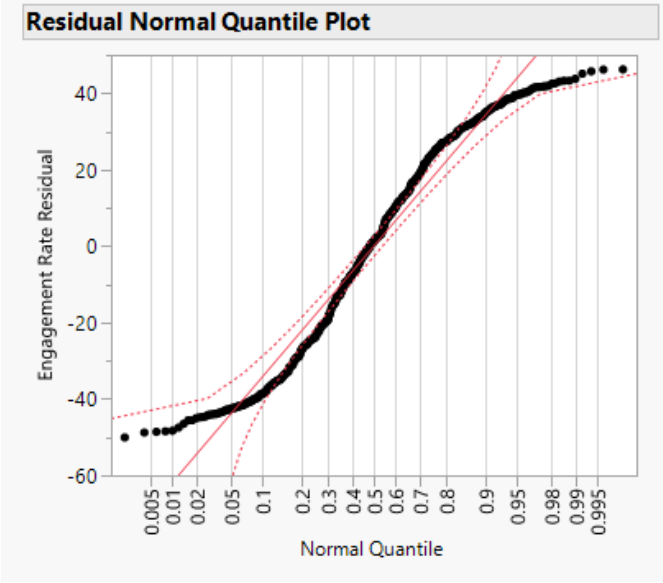
f(1.a)


Summary Statistics
Statistic
Value
Sample 1 Mean
48.5671
Sample 1 Std. Dev.
27.0245
Sample 1 Size
157
Sample 2 Mean
60.7458
Sample 2 Std. Dev.
25.6374
Sample 2 Size
343
Choose Interval Type
z
t
Variance Assumption
Equal (Pool)
Unequal (Welch-Satterthwaite)
Update Results
Results + Confidence Interval
Confidence Level
0.95
Result
Value
Mean 2 - Mean 1
12.178732
Pooled Std Dev
21.444041
Degrees of Freedom
498
t multiplier
1.9647
Std Dev of Difference in Means
2.513
Lower Limit
7.2413
Upper Limit
17.1161
6.9124
10.3686
13.8248
17.281
Mean 2 - Mean 1
Rescale Plot

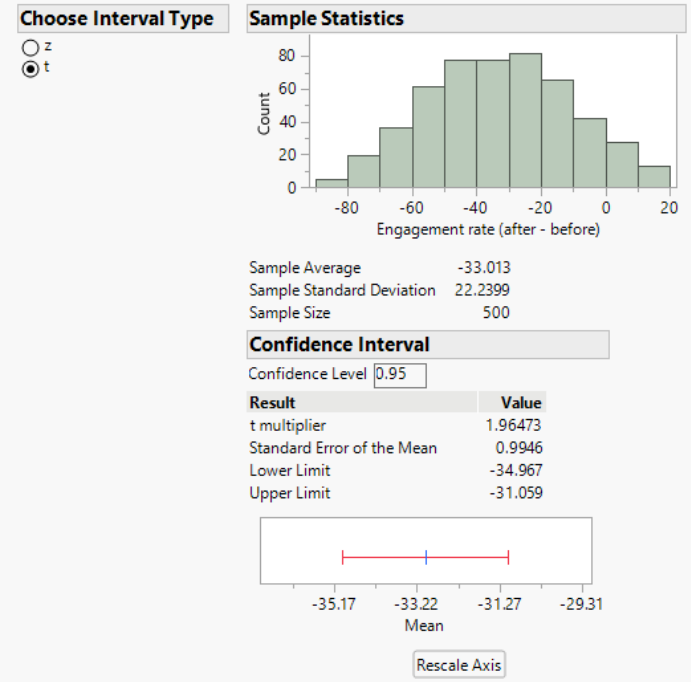
f(1.c)


Residual by Predicted Plot
Engagement Rate Residual
40
20
0
-20
-40
-60
52
54
56
58
60
62
Engagement Rate Predicted

f(2.b)


Residual Normal Quantile Plot
Engagement Rate Residual
40
20
0
-20
-40
-60
0.005
0.01
0.02
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.98
0.99
0.995
Normal Quantile

f(2.c)


Choose Interval Type
z
t
Sample Statistics
Count
80
60
40
20
0
-80
-60
-40
-20
0
20
Engagement rate (after - before)
Sample Average -33.013
Sample Standard Deviation 22.2399
Sample Size 500
Confidence Interval
Confidence Level 0.95
Result Value
t multiplier 1.96473
Standard Error of the Mean 0.9946
Lower Limit -34.967
Upper Limit -31.059
-35.17
-33.22
-31.27
-29.31
Mean
Rescale Axis

f(3.b)