Research Documentation

Introduction:

This section outlines the overall context and objectives of the research.

Define the Population of Interest:

Population of Interest:

The population of interest in this study consists of all text messages sent by me over the course of the last two months. This period is chosen to reflect my recent communication patterns. In order to extrapolate to a larger population (say the entire year), I would need to sample my messages from a larger time frame to account for seasonal variations and changes in behavior.

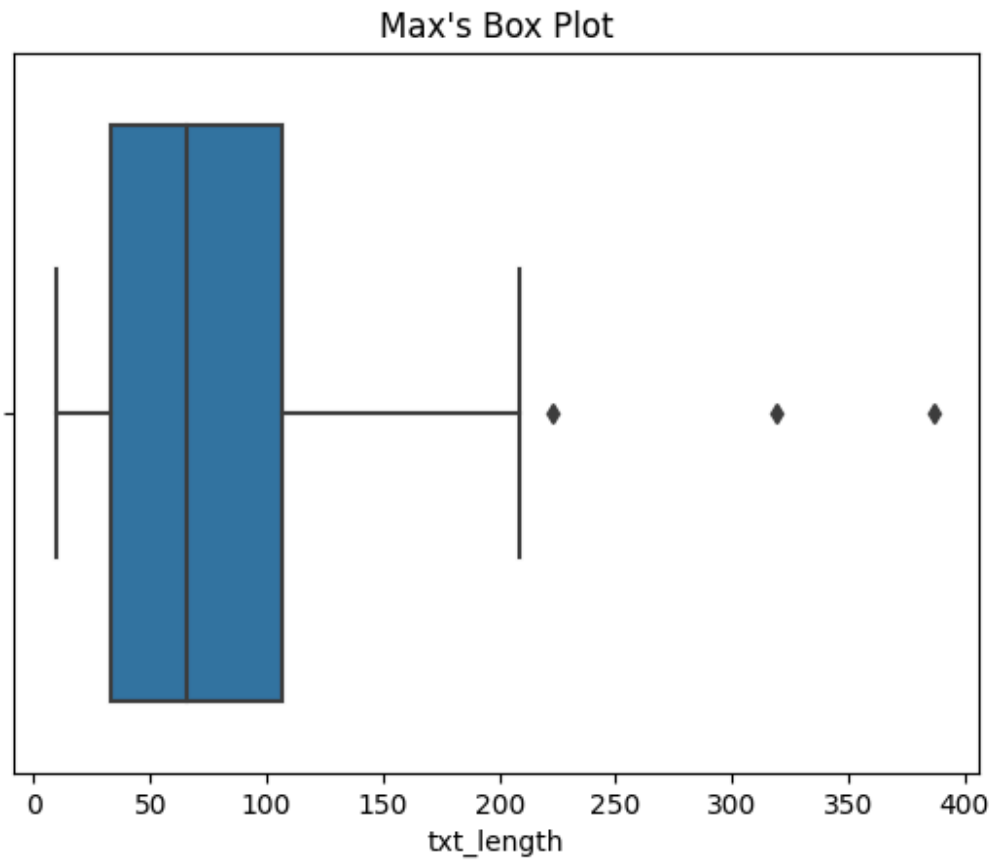
Estimation of the Population Size (N)

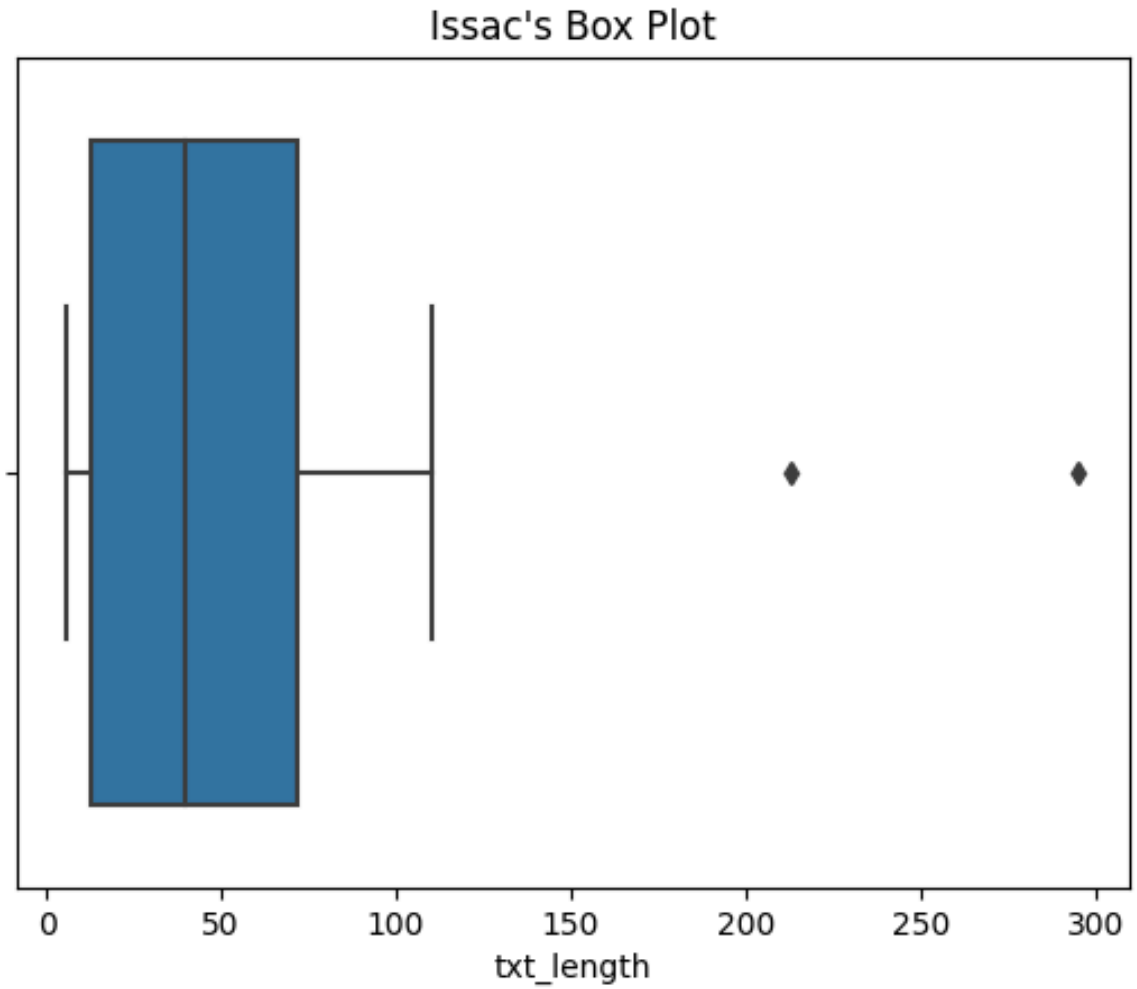
The population size (N) is the total number of text messages I have sent in these last two months. This estimation is based on a comprehensive review of my messaging history during this specific period.

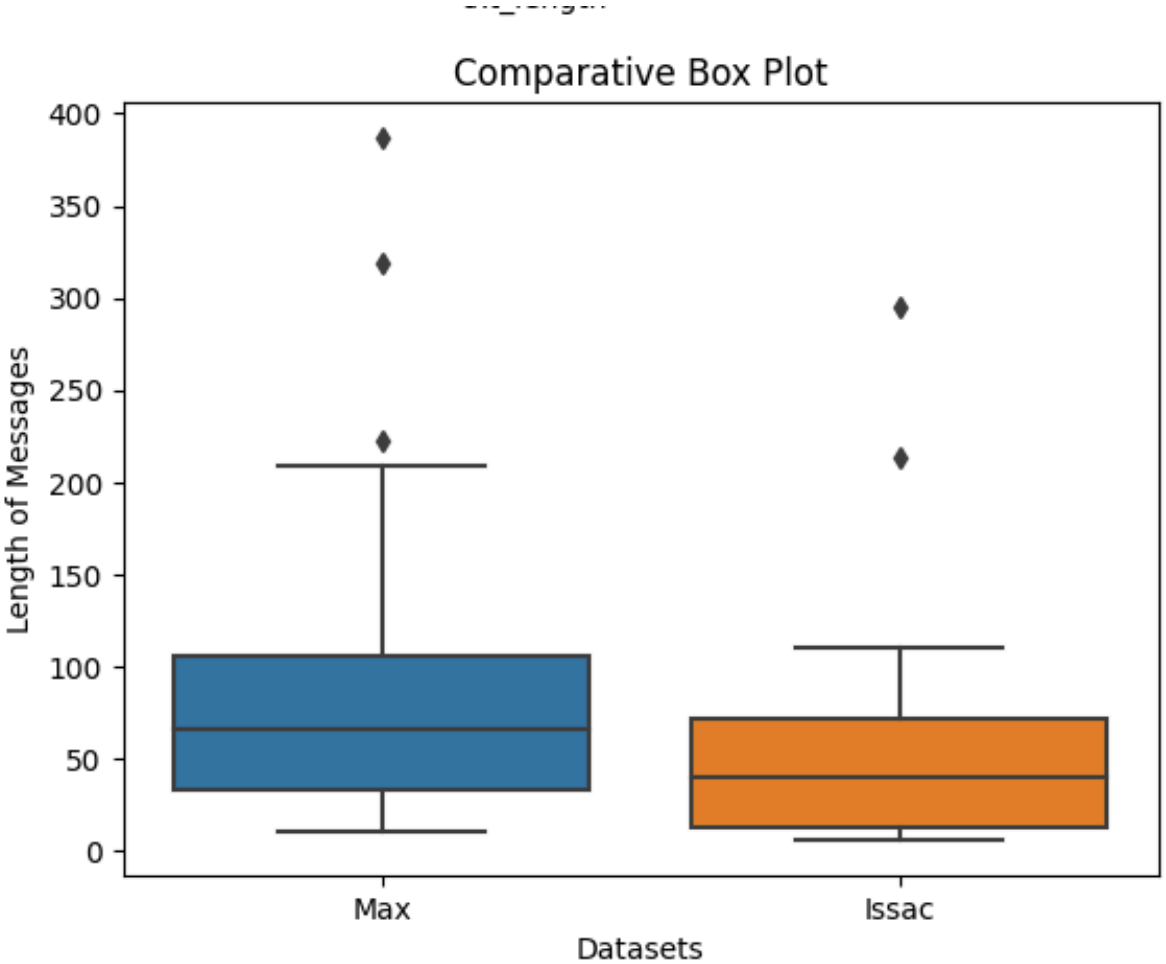
Define Sampling Method

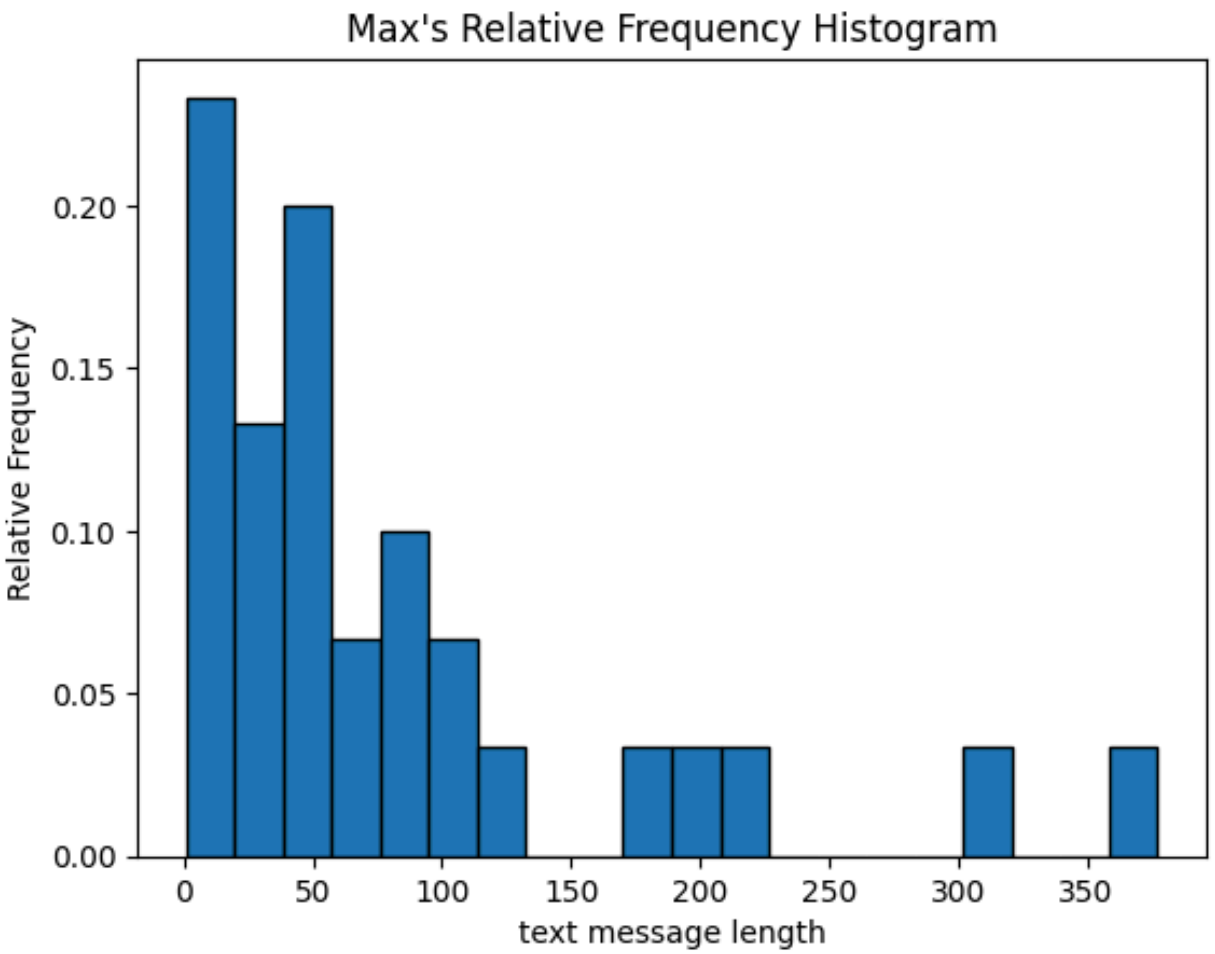
- Choose random text messages from the last two month period. (I collected about 140)
- We will use a random number generator and a seed value to select the n = 30 sample for our analysis
- As mentioned earlier, potential bias may include:
    - changes in behavior, (as result of weather or other causes)
    - changes in weather, (last two months we went from warm to cold)
    - changes in work and school (finals week, sprint reviews, big exams)

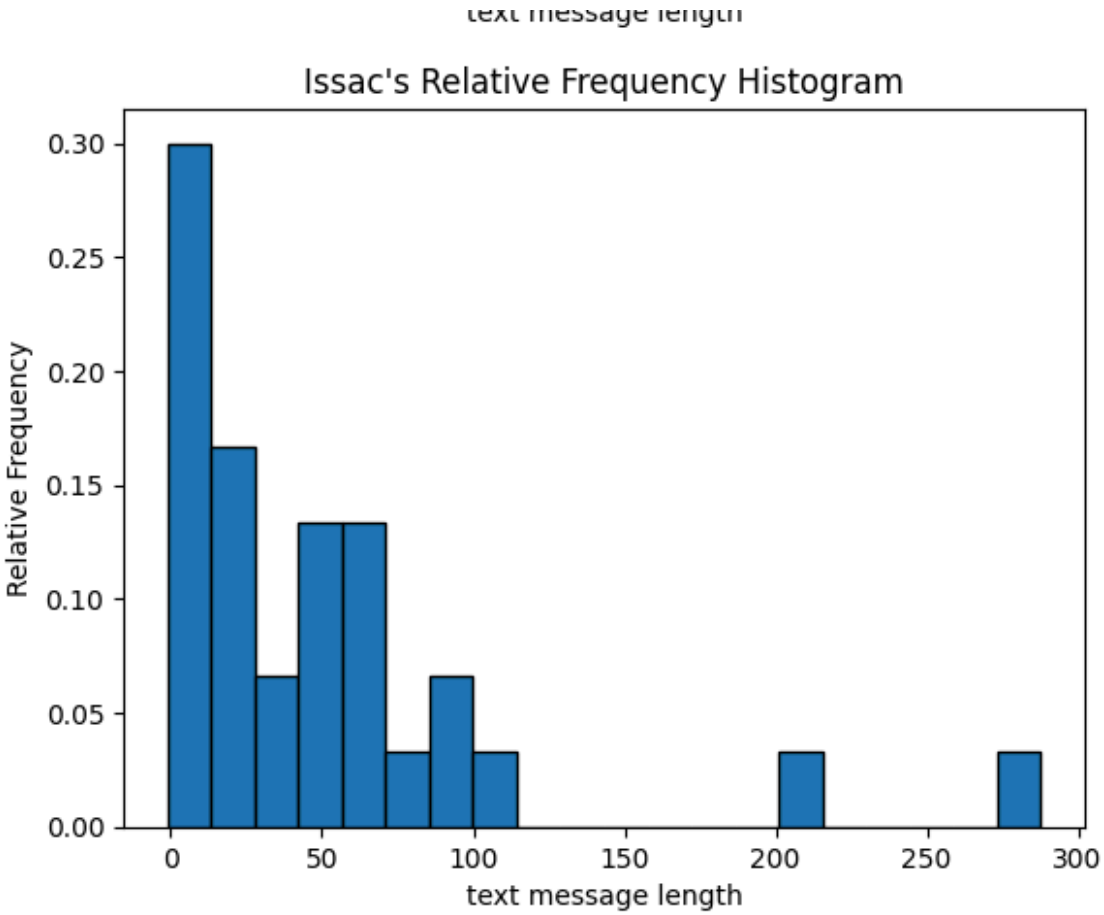
```
Max's Five Number Summary:
 count     30.000000
mean      92.700000
std       89.927615
min       10.000000
25%       33.000000
50%       65.500000
75%      106.250000
max      387.000000
Name: txt_length, dtype: float64
Issac's Five Number Summary:
 count     30.000000
mean      57.600000
std       62.986917
min        6.000000
25%       12.750000
50%       40.000000
75%       72.000000
max      295.000000
Name: txt_length, dtype: float64
```




Issac's Box Plot
0
50
100
150
200
250
300
txt_length


Comparative Box Plot
Length of Messages
400
350
300
250
200
150
100
50
0
Max
Issac
Datasets


Max's Relative Frequency Histogram
Relative Frequency
0.00
0.05
0.10
0.15
0.20
0
50
100
150
200
250
300
350
text message length


text message length
Issac's Relative Frequency Histogram
Relative Frequency
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
text message length

```
================================
Max's Relative Frequency Table:
================================
```

Text Length	Frequency	Relative Frequency
(9.999, 28.85]	7	0.233333
(47.7, 66.55]	6	0.200000
(28.85, 47.7]	4	0.133333
(85.4, 104.25]	3	0.100000
(66.55, 85.4]	2	0.066667
(104.25, 123.1]	2	0.066667
(179.65, 198.5]	1	0.033333
(311.6, 330.45]	1	0.033333
(217.35, 236.2]	1	0.033333
(198.5, 217.35]	1	0.033333
(123.1, 141.95]	1	0.033333
(368.15, 387.0]	1	0.033333
(160.8, 179.65]	0	0.000000
(141.95, 160.8]	0	0.000000
(236.2, 255.05]	0	0.000000
(255.05, 273.9]	0	0.000000
(273.9, 292.75]	0	0.000000
(292.75, 311.6]	0	0.000000
(330.45, 349.3]	0	0.000000
(349.3, 368.15]	0	0.000000

```
================================
Issac's Relative Frequency Table:
================================
    Text Length  Frequency  Relative Frequency
 (5.999, 20.45]          9            0.300000
  (20.45, 34.9]          5            0.166667
  (49.35, 63.8]          4            0.133333
  (63.8, 78.25]          4            0.133333
  (34.9, 49.35]          2            0.066667
 (92.7, 107.15]          2            0.066667
(208.3, 222.75]          1            0.033333
(280.55, 295.0]          1            0.033333
(107.15, 121.6]          1            0.033333
  (78.25, 92.7]          1            0.033333
(121.6, 136.05]          0            0.000000
(136.05, 150.5]          0            0.000000
(164.95, 179.4]          0            0.000000
(179.4, 193.85]          0            0.000000
(193.85, 208.3]          0            0.000000
(222.75, 237.2]          0            0.000000
(237.2, 251.65]          0            0.000000
(251.65, 266.1]          0            0.000000
(266.1, 280.55]          0            0.000000
(150.5, 164.95]          0            0.000000
```

```
Max's 95% Confidence Interval: (59.12047685592172, 126.27952314407828)

Issac's 95% Confidence Interval: (34.080298624201106, 81.1197013757989)

Max's std deviation range:
{'1_std_dev': 83.33333333333334, '2_std_dev': 93.33333333333333, '3_std_dev': 96.66666666666667}

Issac's std deviation range:
{'1_std_dev': 93.33333333333333, '2_std_dev': 93.33333333333333, '3_std_dev': 96.66666666666667}



Two sample t-test:
T-statistic: 1.7510407337841827, P-value: 0.08522603113442435
```

Comments and Analysis:

Response to Frequency Table and Histogram

- The sample appears to follow an exponential curve for both Issac and my sample datasets.
    - This is likely because we tend to use text messages to get the point accross quickly, and if we need to communicate more than a few words, we typically call people.
    - Text messages are typically a precursor to longer forms communication, so it would make sense that there are exponentially more shorter messages in the dataset.

Analysis of Confidence Interval:

Output from program:

```
Max's 95% Confidence Interval: (59.12047685592172, 126.27952314407828)
```

```
Issac's 95% Confidence Interval: (34.080298624201106, 81.1197013757989)
```

- I am 95% confident that the true average number of words per text message for Max and Issac falls within their respective intervals.
- This seems to state that it is more likely that I tend to write longer messages on average than Issac based on the intervals.

Analysis of Hypothesis Test:

Output from program:

```
Two sample t-test: T-statistic: 1.7510407337841827, P-value: 0.08522603113442435
```

Two-Sample T-Test Results

Test Details:

- **Significance Level**: 0.05 (95% Confidence Level)
- **T-Test Type**: Two-Sample T-Test

Test Results:

- **T-Statistic**: 1.751
- **P-Value**: 0.085

Interpretation:

Step 1: Declare Significance Level

The alpha (α) level is set at 0.05, indicating a 95% confidence level.

Step 2: Conduct T-Test

A two-sample t-test was performed, yielding a t-statistic of 1.751.

Step 3: Evaluate P-Value

The calculated p-value from the t-test is 0.085.

The Significance Level is 0.05.

[(α) < P value]: => : 0.05 < 0.085

Step 4: Reject or Accept the Null Hypothesis

- **Decision**: Do not reject the null hypothesis.
- **Rationale**: Since the p-value (0.085) is greater than the significance level (0.05), there is not sufficient evidence to conclude a statistically significant difference between the means of the two samples.
- **Explanation**: This means that the inferred population mean of Issac and My texting habits do not differ enough to reject the null hypothesis at a 95% Confidence Level.

Final Conclusions:

Conclusions from confidence intervals: The confidence intervals seem to suggest that I write longer messages on average as my minimum and maximum values are both greater than Issac's.

Conclusions from Hypothesis Testing: From our findings with both the confidence interval and the hypothesis tests we can conclude that the true population mean of our texting habits does not differ at a statistically significant level.

Conclusiosn from both CI and Hypothesis Tests: Interestingly, the confidence intervals for Max and Issac suggest different texting styles, with one potentially favoring shorter or longer texts than the other. However, these differences aren't significant enough to impact the overall average text lengths when considering the broader population.

Final Thoughts: Moreover, the absence of a significant difference in average texting lengths, as demonstrated by the hypothesis test, points to a broader pattern in texting behavior. Our initial observation of an exponential curve in text message lengths supports the idea that texting behavior is more likely to follow an exponential distribution rather than a uniform one. This implies that shorter texts are more common, aligning with the general tendency to use text messages for brief communications and opting for longer communication methods for more in-depth conversations.