Bazeley, Mikiko
Springboard 2018

Report: Exploratory Data Analysis

- *Are there variables that are particularly significant in terms of explaining the answer to your project question*?

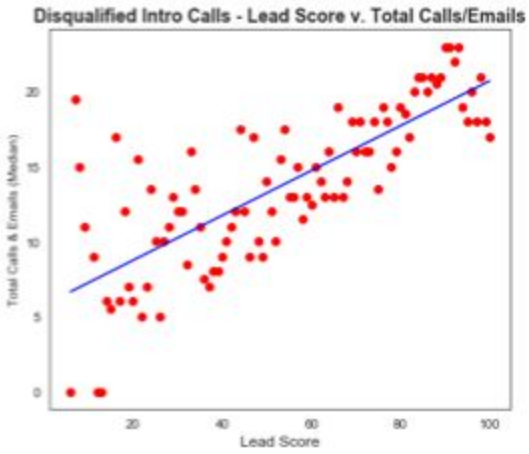
For this section of the capstone project I chose to analyse the relationship between the qualification status of intro calls (the Target variable) and the Lead Score, Total Calls & Emails (exchanged between the sales team and the prospect), and the duration between when the lead was created and when an intro call was created (or the time it took for someone to respond and qualify the prospect for a demo).

There are additional features that could potentially be valuable (Country, Landing Page, Marketing Traffic Channel, Customer Type) that won't be considered here but are potentially still valuable.

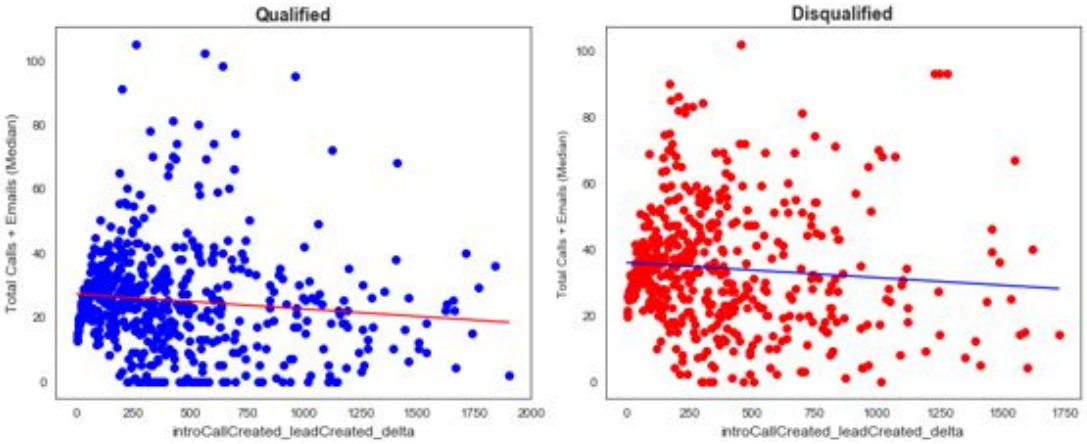
- *Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?*

From previous exploration with this data set we have the following intuition:

- **Lead Score vs. Total Calls/Emails**
    - Higher lead scores for disqualified intro calls were accompanied by more engagement (Total Calls + Emails), while Qualified Intro Calls showed a flatter relationship.



- **Time Delta Between Lead & Intro Call Creation vs. Target Outcome**
    - Qualified and Unqualified Intro Calls showed similar relationships between Time Delta and Total Calls & Emails.



From experience with the business context of the project, additional assumptions should be tested:

- **Lead Scores vs. Target Outcome** -
    - There should be no difference in average scores between two groups (and scores should be concentrated between 80-90, with no leads with scores below 60) because marketing claimed to automatically reject leads with a grade of D or below (which should correspond to a lead score of 60).

- **Total Calls & Emails vs. Target Outcome** -
    - We could see potentially two opposing trends at work, with regards to level of engagement with the prospect and their suitability for sales.
        - We could a positive relationship between amount of engagement for unqualified candidates (where sales reps need to exert additional effort to pull them further into the sales cycle).

    - We could see a negative relationship between level of engagement (via Total Calls & Emails) for qualified prospects (higher quality due to greater suitability or interest, so less effort needed to pull the prospect into the sales cycle).
    - We could also see a positive relationship where savvy, highly engaged shoppers are requiring more engagement from our sales teams through follow-up questions and due diligence.

- **Time Delta Between Lead & Intro Call Creation vs. Target Outcome**
    - Freshness is an important concept in sales and the common intuition is the longer the sales cycle takes (assuming same resulting payoffs), the less likely a prospect will stay engaged in the sales cycle.

By using both frequentist and randomized approaches, we can understand how significantly different the means are between the groups (by Target Outcome) and whether these differences could be explained through chance.

- *What are the most appropriate tests to use to analyse these relationships?*

For each variable I chose to do a permutation test, a bootstrap test, a Mann-Whitney test, and a Welch's t-test in order of restrictiveness of criteria. I chose to perform the permutation and bootstrap tests due to lack of assumptions around normality, distributions, and equal variances. I then performed a Mann-Whitney test and Welch's t-test to further validate the results of the randomization models.

Analysis 1: Lead Score vs Target Outcome

- We first want to understand the summary statistics of Qualified vs. Unqualified Intro Calls and whether the assertion that there is no difference (and lead scores should be 60+).

- From printing the summary statistics, we can already see that the assertion that the sales team doesn't interact with leads below 60 is false. Both samples of Qualified and Disqualified Intro Calls had a minimum below 60 (Qualified: 9, Disqualified: 6).
- However our Qualified sample is displaying an IQR of [82 (25%), 92 (75%)] and our Disqualified sample is displaying an IQR of [73 (25%), 91 (75%)], so it's possible the assertion that the majority of leads leading to demo calls should be around 70-90. We also observe a difference in means: Qualified (84), Unqualified (80).

Summary of Qualified Intro Calls:

count	10164.000000
mean	84.190279
std	13.998116
min	9.000000
25%	82.000000
50%	90.000000
75%	92.000000
max	100.000000

Summary of Not Qualified Intro Calls:

count	9813.000000
mean	79.379395
std	17.714944
min	6.000000
25%	73.000000
50%	88.000000
75%	91.000000
max	100.000000

Fig 1A: ECDF of Lead Scores by Qualified vs. Unqualified Intro Calls

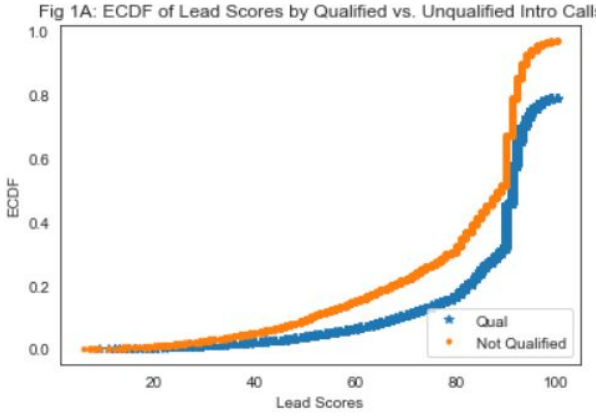


Fig 1B: Violin Plot: Lead Scores by Qualified vs. Unqualified Intro Calls

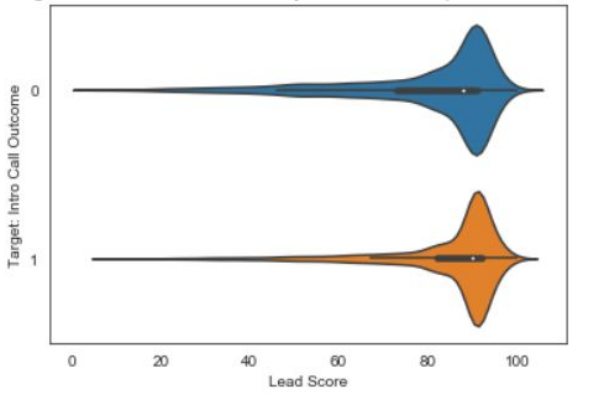
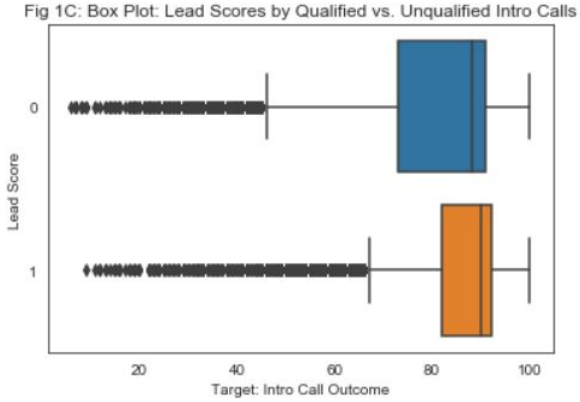


Fig 1C: Box Plot: Lead Scores by Qualified vs. Unqualified Intro Calls

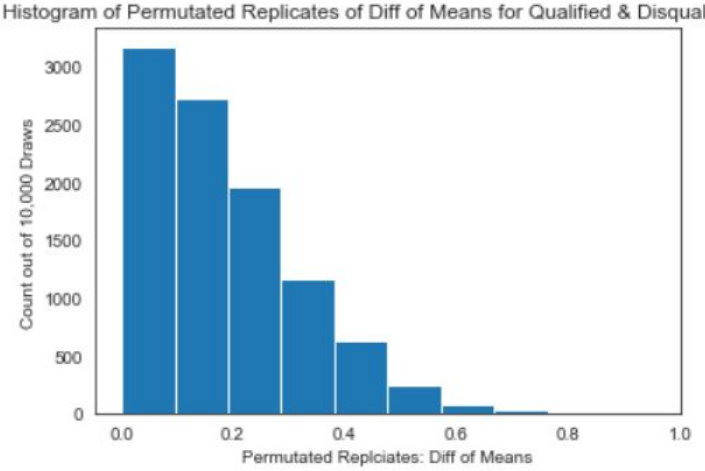


- I created charts showing the empirical cumulative distribution function (Fig 1A), the distribution of lead scores (Fig 1B), and a box-plot displaying the summary statistics (Fig 1C) in order to further understand the distributions of lead scores.
  - **[Fig 1A]** We can see from the ECDF that both groups experience a pick up at Lead Score ~ 80, where a relatively larger proportion of prospects exist. However our

sample of Disqualified Intro Calls shows a relatively higher proportion of leads between 40 and 90.
    - **[Fig 1B]** Each violin plot is scaled by count. We can see that the samples are roughly the same size, seem to have medians concentrated around ~90 (visually confirm the summary statistics we printed earlier), and both are left skewed. The Unqualified Lead Scores (Target Outcome = 0) sample also has a fatter peak and more observations in the 60 & below range (as shown by the fatter tail).
    - **[Fig 1C]** We can see that the Unqualified Lead Scores potentially display higher variance and have lower lead scores at the 25%.

- **First hypothesis: Permutation Test** - Simulating the null hypothesis that Qualified and Unqualified Lead Scores have identical distributions even while the means differ. Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value.
    - Results:
        - `Empirical Diff of Mean: 4.812948133518233`
        - `Proportion of replicates with value as great or greater than empirical diff of means`
          `p-value = 0.0000`
    - The histogram of permuted replicates below shows how extreme a value of 4.8 is relative to expectation if the populations had been the same (Fig 1D).

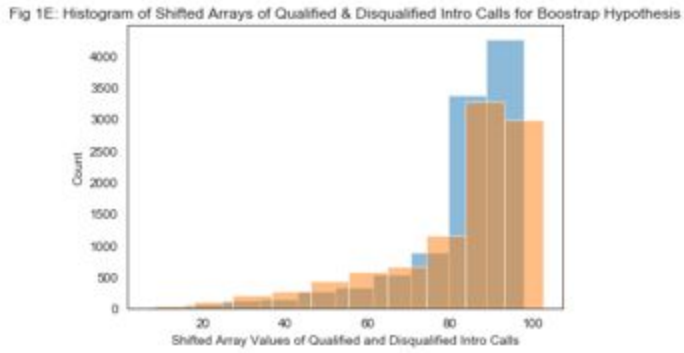
Fig 1D: Histogram of Permutated Replicates of Diff of Means for Qualified & Disqualified Intro Calls

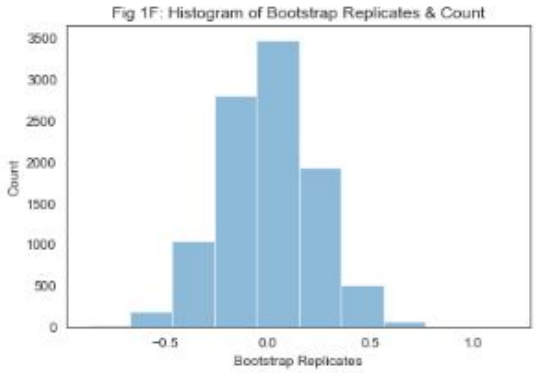


- **Second hypothesis: Bootstrap Test** - Simulating the null hypothesis that Qualified and Unqualified Lead Scores have identical means but come from different populations.

Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value given the shifted arrays (Fig 1E).

- Results:
  - `Mean Values of Concatenated Data:  81.82739465518966`
  - `Empirical Diff of Mean: 4.812948133518233y`
  - `Proportion of replicates with value as great or greater than empirical diff of means`
    `p-value = 0.0000`
- The histogram of bootstrap replicates below shows how extreme a value of 4.8 is relative to expectation if the populations had been the same (Fig 1F).





- **Third 7 Fourth hypothesis: Mann-Whitney vs Welch's T-Test** -
  - Results:
    - `MannwhitneyuResult(statistic=40960991.0, pvalue=1.2447515478111441e-107)`
    - `Ttest_indResult(statistic=array([21.26104234]), pvalue=array([3.89798532e-99]))`
  - Both the Mann-Whitney test and Welch's T-Test seem to also support rejecting the null hypothesis that the means are the same.

Analysis 2: Total Calls & Emails vs Target Outcome

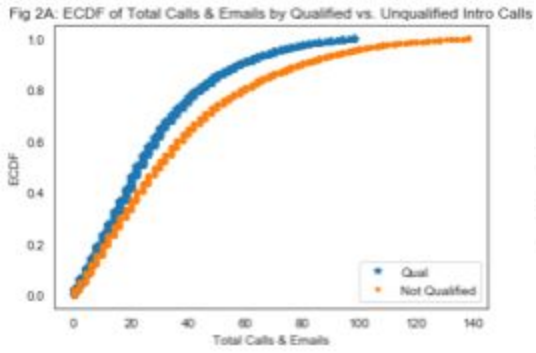
- The first step in analyzing the possible relationship between Total Calls & Emails on Target outcome is to examine the summary statistics and note differences in mean, median, min/max, In an ideal sales world, most sales managers would like sales reps to

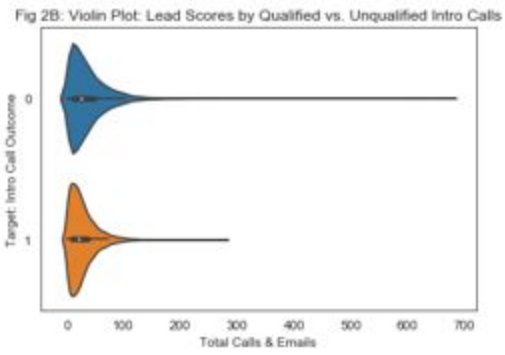
engage in the minimum amount of correspondence needed to: (1) qualify a prospect and (2) ensure good prospects are pulled into the sales process.

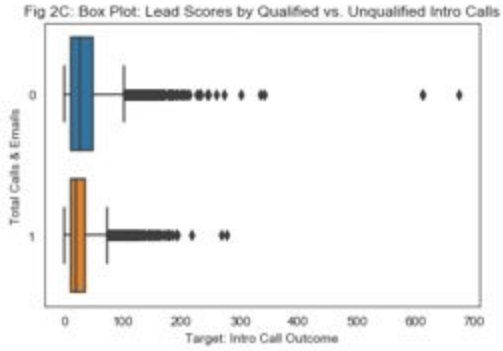
- From printing the summary statistics, we can already see that Disqualified Intro Calls were associated with a higher mean of Total Calls & Emails compared to Qualified Intro Calls (36.9 vs. 28.0).
- We can also see a difference in the IQR of Disqualified vs Qualified Intro Calls, indicating that prospects of Disqualified Intro Calls could be taking up more sales rep time (Qualified: [12 (25%), 40 (75%)], Disqualified: [14 (25%), 52 (75%)]).]

```
Qualified:        totalCallsEmails
count         6953.000000
mean            28.003452
std             21.391253
min              0.000000
25%             12.000000
50%             24.000000
75%             40.000000
max            100.000000
```

```
Not Qualified:        totalCallsEmails
count         6870.000000
mean            36.900728
std             29.253542
min              0.000000
25%             14.000000
50%             30.000000
75%             52.000000
max            138.000000
```



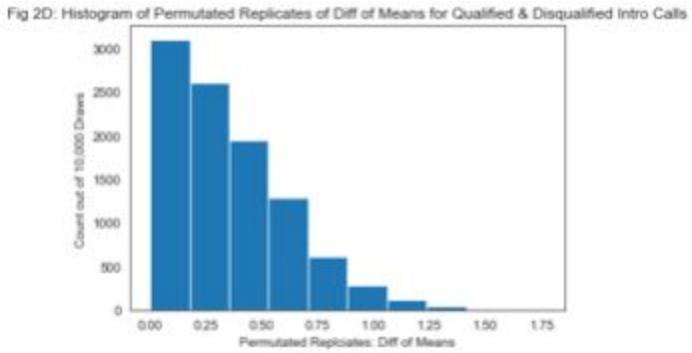




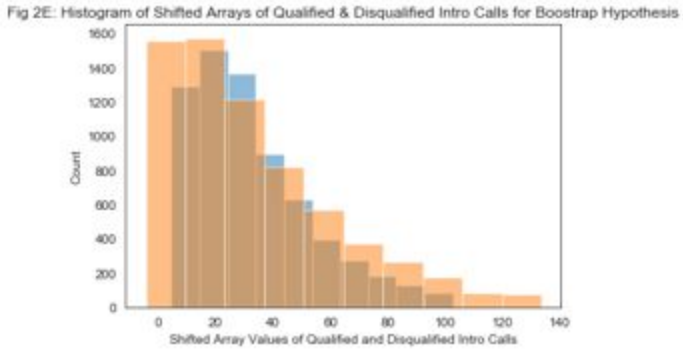
- Similarly to Analysis Part 1, I visually examine the sample distributions further.
  - **[Fig 2A]** Looking at the ECDF's we can see that 60% of Qualified Intro Calls were associated with =<~30 Total Calls & Emails while the same % of Disqualified Intro Calls were associated with =< ~40 Total Calls & Emails.
  - **[Fig 2B]** Another trend we can observe (from both the violin plot and the summary statistics printout) is the presence of significant outliers in the Disqualified sample of Total Calls & Emails.

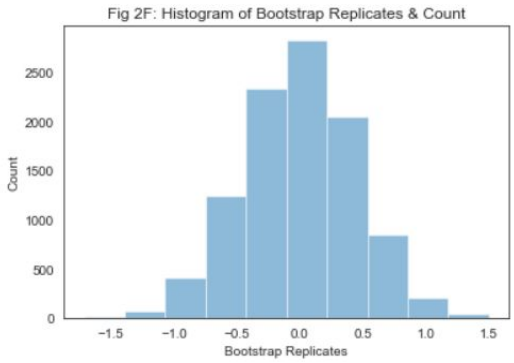
- **[Fig 2C]** We can also verify in the boxplot the wider IQR, mostly due to the 75 percentile of data being shifted to the right (i.e. towards higher Total Calls & Emails) with two incredibly extreme outliers (~600+ Emails & Calls!)

- **First hypothesis: Permutation Test** - Simulating the null hypothesis that Qualified and Unqualified Total Calls & Emails have identical distributions even while the means differ. Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value.
  - Results:
    - `Empirical Diff of Mean: 8.897276054590698`
    - `Proportion of replicates with value as great or greater than empirical diff of means`
      `p-value = 0.0000`
  - The histogram of permuted replicates below shows how extreme a value of 9.0 is relative to expectation if the populations had been the same (Fig 2D).



- **Second hypothesis: Bootstrap Test** - Simulating the null hypothesis that Qualified and Unqualified Total Calls/Emails have identical means but come from different populations. Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value given the shifted arrays (Fig 2E).
  - Results:
    - `Mean Values of Concatenated Data:  32.42537799319974`
    - `Empirical Diff of Mean: 8.897276054590698`
    - `Proportion of replicates with value as great or greater than empirical diff of means`
      `p-value = 0.0000`
  - The histogram of bootstrap replicates below shows how extreme a value of 9.0 is relative to expectation if the populations had been the same (Fig 2F).





- **Third 7 Fourth hypothesis: Mann-Whitney vs Welch's T-Test** -
  - Results:
    - `MannwhitneyuResult(statistic=20132210.0, pvalue=6.468704474691752e-58)`
    - `Ttest_indResult(statistic=array([-20.39150626]), pvalue=array([5.83924509e-91]))`
  - Both the Mann-Whitney test and Welch's T-Test seem to also support rejecting the null hypothesis that the means are the same.

Analysis 3: Lead-Intro Call Delta vs Target Outcome

- As written previously, lead freshness is an important concept in sales and we could expect to see Disqualified Intro Calls associated with higher Time Deltas.
  - On average however, Qualified Intro Calls have higher means (39 days) than Disqualified Intro Calls (30 days).

Qualified:	introCallCreated_leadCreated_delta
count	8938.000000
mean	30.622734
std	62.353438
min	1.000000
25%	2.000000
50%	7.000000
75%	25.000000
max	420.000000

Not Qualified:	introCallCreated_leadCreated_delta
count	7507.000000
mean	29.170108
std	56.099237
min	1.000000
25%	3.000000
50%	8.000000
75%	28.000000
max	390.000000

- We can also observe some interesting characteristics about the data with regards to the Time Delta of Lead Created to Intro Call Created.
    - **[Fig 3A]** The ECDF's are very similar giving the first indication that there might not be significant differences between Disqualified and Qualified Intro Calls with regards to the Time Delta.
    - **[Fig 3B]** We do see however that there are potentially some negative values that could be impacting the mean.
    - **[Fig 3C]** For both groups, we notice there are a number of outliers for both Qualified and Disqualified Intro Calls. Qualified Intro Calls also has some strangely negative values which may need to be excluded.

Fig 3A: ECDF of Time between Lead and Intro Call Creation by Qualified vs. Unqualified Intro Calls

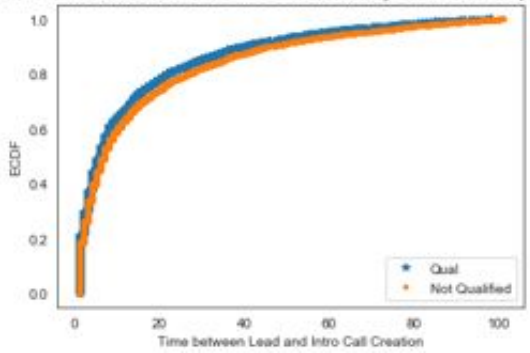


Fig 3B: Violin Plot: Time between Lead and Intro Call Creation by Qualified vs. Unqualified Intro Calls

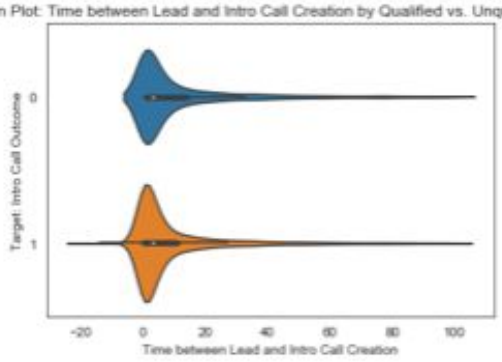
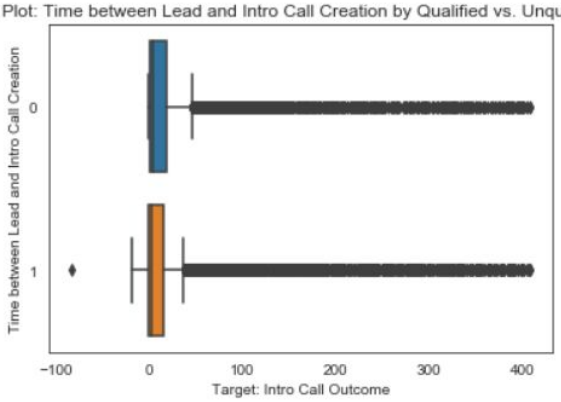
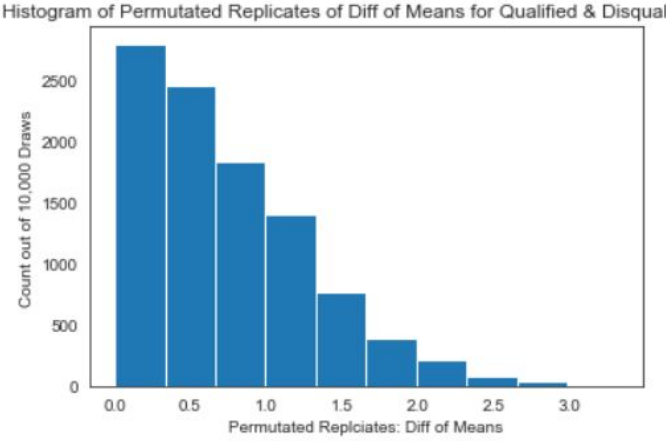


Fig 3C: Box Plot: Time between Lead and Intro Call Creation by Qualified vs. Unqualified Intro Calls



- **First hypothesis: Permutation Test** - Simulating the null hypothesis that Qualified and Unqualified Lead Scores have identical distributions even while the means differ. Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value.
    - Results:
        - `Empirical Diff of Mean: 1.452626493187548`
        - `Proportion of replicates with value as great or greater than empirical diff of means`
          `p-value = 0.1170`
    - From the histogram of permuted replicates we can visually see that the empirical mean of 1.5 isn't an extreme value with about 12% of the permuted values having a value as great or greater than the empirical difference of means. The permutation test result doesn't seem to provide evidence to reject the null hypothesis that Qualified and Disqualified Intro Calls are significantly different with regards to the Time Delta(Fig 3D).

Fig 3D: Histogram of Permutated Replicates of Diff of Means for Qualified & Disqualified Intro Calls



- **Second hypothesis: Bootstrap Test** - Simulating the null hypothesis that Qualified and Unqualified Lead Scores have identical means but come from different populations. Alpha = 5%. Our goal is to understand how likely we would have calculated a difference of means as great or greater than the current value given the shifted arrays (Fig 1E).
    - Results:
        - `Mean Values of Concatenated Data:  29.95962298570994`
        - `Empirical Diff of Mean: 1.452626493187548`

        - `Proportion of replicates with value as great or greater than empirical diff of means`
          `p-value = 0.0588`
    - Similarly the Bootstrap test isn't significant at the 5% level, with ~5.9% of the bootstrap replicates exhibiting a value equal to or greater than the empirical difference of means. (Fig 3F).

Fig 3E: Histogram of Shifted Arrays of Qualified & Disqualified Intro Calls for Boostrap Hypothesis

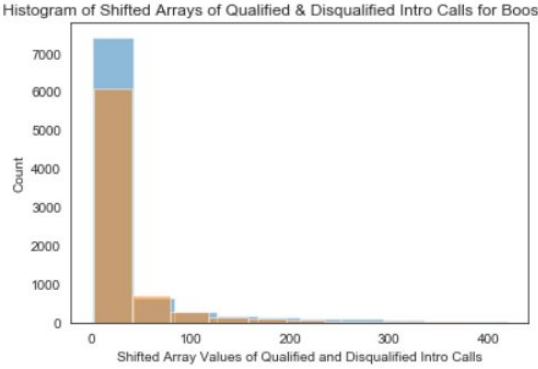
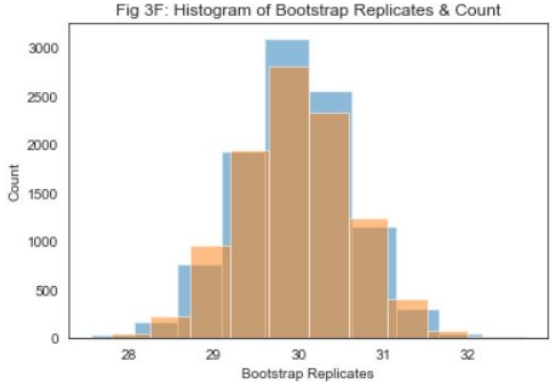


Fig 3F: Histogram of Bootstrap Replicates & Count



- **Third 7 Fourth hypothesis: Mann-Whitney vs Welch's T-Test** -
    - Results:
        - `MannwhitneyuResult(statistic=32321837.5, pvalue=2.4385713413183856e-05)`
        - `Ttest_indResult(statistic=array([1.5717012]), pvalue=array([0.1160392]))`
    - We are seeing conflicted results from the Mann-Whitney test (which seems to reject the null hypothesis that the populations are similar) and Welch's T-Test (which doesn't result in a statistically significant p-value).