Report: Milestone Report

I .Problem statement: Why it’s a useful question to answer and for whom (get this from your proposal)

This first capstone project will involve creating a model to score intro calls. Specifically we want to understand how valuable a prospect is based on certain characteristics about that prospect and how likely an intro call will be qualified.

My client for the project is my current company, a B2B Saas company interested in scaling sales operations. Currently we use a third-party vendor for modeling and lead scoring. A significant amount of effort and financial investment has gone into the third-party tool and it would be great if we could benchmark the value that tool provides.

Lead scoring is a classic candidate for machine learning. We want to classify intro call outcomes based on meta-data about the lead (person, lead source, campaign, associated company, etc). We have 4 years of consistent data spanning thousands of accounts and interactions with customers and prospects.

II. Description of the dataset, how you obtained, cleaned, and wrangled it (get this from your data wrangling report)

Description of Data

I have two tables of data from the data warehouse split into Leads and Intro Calls.

Leads table	Intro Calls table
• RangeIndex: 493767 entries, 0 to 493766 • Columns: 108 entries id 493767 isdeleted 493767 masterrecordid 32 lastname 491233 firstname 381490 salutation 6181 name 493519	• RangeIndex: 23241 entries, 0 to 23240 • Columns: 115 entries id 23241 ownerid 23241 isdeleted 23241 name 23241 currencyisocode 23241 recordtypeid 18201

title	292855	createddate	23241
company	492978	createdbyid	23241
street	67935	lastmodifieddate	23241
city	427036	lastmodifiedbyid	23241
state	250352	systemmodstamp	23241
postalcode	61654	lastactivitydate	1312
country	445275	lastvieweddate	0
email	490989	lastreferenceddate	0
leadsource	493371	actual_start_date_time__c	0
status	493767	application_type__c	17761
industry	40737	assigned_to__c	23229
ownerid	493767	lead_additional_phone__c	6287
hasoptedoutofemail	493767	lead_company__c	22676
isconverted	493767	lead_email__c	22600
converteddate	20192	lead_name__c	22676
convertedaccountid	20076	lead_phone__c	22576
convertedcontactid	20062	meeting_comments__c	18885
convertedopportunityid	13814	meeting_status__c	23241
createddate	493767	new_existing_customer__c	21424
createdbyid	493767	no_show_other_reason__c	176
lastmodifieddate	493767	no_show_reason__c	1395
lastmodifiedbyid	493767	no_of_days_from_schedule_to_meeting__c	0
...		opportunity__c	23241
sbc_rejected_reason__c	2494	...	
sbc_rejected_note__c	599	double_dipper__c	9
sf_edit_lead__c	493767	infer_score__c	20225
sbc_demo__c	493767	assigned_to_manager__c	23068
walkme_mobile__c	493767	created_by_manager__c	22678
scheduled_se_meeting__c	3259	lead_status__c	22676
se_meeting_time__c	424	lead_owner_id__c	22676
se_lead__c	493767	opportunity_presales_stage__c	3558
outbound_lead__c	493767	lead_company_size__c	6877
se_lead_progress__c	47890	se_rep__c	32
se_lead_rejected__c	10587	is_qualified__c	23241
of_employees__c	5260	se_reviewed__c	23241
jaco_lead__c	493767	introduced_walkme_se__c	932
ad_text__c	52279	of_employees__c	9016
marketing_camp_id__c	72861	time_zone__c	21311
marketing_channel_campaign_id__c	340467	owner_sales_team__c	22834
marketing_channel_ad_id__c	268150	end_date_formula__c	23237
marketing_channel_ad_name__c	170494	lead_customer_type__c	22564
marketing_channel_campaign_name__c	461031	implementing_partner__c	272
marketing_channel_ad_group_name__c	92004	copy_country_from_lead__c	23241
account_gclid__c	62161	lead_country_text__c	12157
ods_update_date	493767	assigned_to_role__c	23224
of_employees_category__c	155699	outbound_lead__c	23241

market_segment__c 155699 market_segment_code__c 493767 account__c 66842 owner_sales_team__c 312608 mintigo_score__c 109311 mintigo_rank__c 109310 clearbit_employees__c 490248	implementing_partner_details__c 129 decision_maker_picklist__c 9725 contact__c 538 attributed__c 23241 bdr_intro_call__c 23241 intro_call_source_marketing_outbound__c 13657 attribution_date__c 3 product_s__c 5978

Given the Leads and Intro Call data sets also contained in progress prospects, I needed to focus on Intro Calls that had been closed out. Of 22.9K records, 12.8K (56%) were qualified vs 10K (44%) disqualified.

Data Wrangling

In order to collect the data I first needed to access my company's internal Redshift data warehouse. Using the python libraries sqlalchemy and psycopg2 (a postgreSQL driver) I queried five main tables representing Salesforce objects (Leads, Opportunities, Demos, Accounts, and Products). Demos are our main objects of interest with Leads, Accounts, Opportunities and Products providing further clarity into customer demographics and demo outcomes. Providing my user credentials and database information, I opened a connection to pass queries (which is later shutdown after the queries are completed".

An important concept to understand in analyzing sales data is the business process from Lead to Opportunity and how the different entities are represented. The stage of the sales cycle we are interested in examining will determine how the different tables are joined. The typical sales process is depicted below:

[Lead] → MQL→ Sales Accepted Lead → Sales Qualified Lead → [Opportunity] → [Customer]

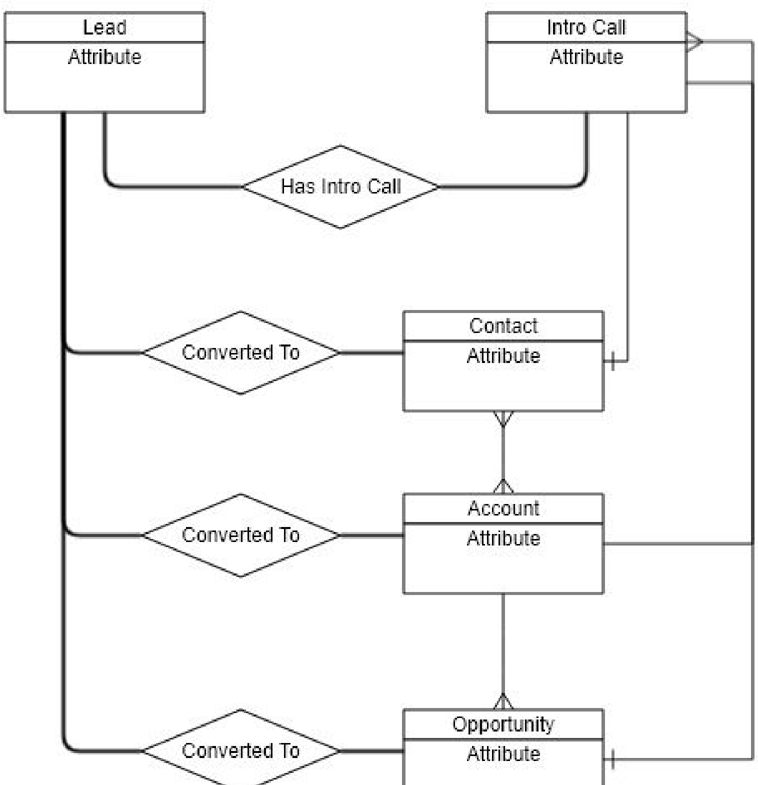
Along the way from Prospect to Customer, at least 3+ entities can be created when the individual enters the system as a Prospect, engages with our sales team, is converted to an Opportunity which is connected to an Account and the individual (now represented as a Contact). Each object will have a number of standard and custom editable fields that can be easily created to enrich a company's insight into the individual, deal, or company.

Original API Name	New Name	Type	select	new_names	Transformation	Transformation 2	Transformation 3	Use for model?	Use for EDA?	Data Issues	Type
email	email	Lead_PersonalInformation	'email',	'email___Lead_PersonalInformation',	Create new column - filled in non gmail/hotmail/etc email domain		String	N	Y	Can be fake "xyz","123" | Can be n	Lead: Personal Information
firstname	firstname	Lead_PersonalInformation	'firstname',	'firstname___Lead_PersonalInformation',	Create new column - Filled in First Name		String	N	Y	Can be fake "xyz","123" | Can be n	Lead: Personal Information
lastname	lastname	Lead_PersonalInformation	'lastname',	'lastname___Lead_PersonalInformation',	Create new column - Filled in last name		String	N	Y	Can be fake "xyz","123"	Lead: Personal Information
title	title	Lead_PersonalInformation	'title',	'title___Lead_PersonalInformation',	Create new column - filled in title	Create new column - title contains "manager/director/etc"	String	Y	Y	Not always filled	Lead: Personal Information
customer_type__c	customerType	Lead_LeadCompanyInformation	'customer_type__c',	'customerType___Lead_LeadCompanyInformation',			String	Y	Y	Can be null	Lead: Lead Company Information
company	company	Lead_LeadCompanyInformation	'company',	'company___Lead_LeadCompanyInformation',	Create new column - filled in company	Create new column - filled in non-dummy	String	N	Y	Not filled or fake	Lead: Lead Company Information
street	street	Lead_LeadCompanyInformation	'street',	'street___Lead_LeadCompanyInformation',			Address	N	N		Lead: Lead Company Information
city	city	Lead_LeadCompanyInformation	'city',	'city___Lead_LeadCompanyInformation',			Address	N	Y		Lead: Lead Company Information
state	state	Lead_LeadCompanyInformation	'state',	'state___Lead_LeadCompanyInformation',			Address	N	Y		Lead: Lead Company Information
country	country	Lead_LeadCompanyInformation	'country',	'country___Lead_LeadCompanyInformation',			Address	Y	Y		Lead: Lead Company Information
linkedin_page__c	linkedinPage	Lead_MarketingInformation	'linkedin_page__c',	'linkedinPage___Lead_MarketingInformation',			String	Y	Y		Lead: Marketing Information
traffic_channels__c	trafficChannel	Lead_MarketingInformation	'traffic_channels__c',	'trafficChannel___Lead_MarketingInformation',			String	Y	Y	Picklist value, can be null	Lead: Marketing Information
marketing_channel_camp	mktChannelcampaign	Lead_MarketingInformation	'marketing_channel_campa	'mktChannelcampaign___Lead_MarketingInformation',			String	Y	Y	Can be null	Lead: Marketing Information
landing_page__c	landingPage	Lead_MarketingInformation	'landing_page__c',	'landingPage___Lead_MarketingInformation',	Create new column - Has a linkedin page		String	Y	Y	Can be null & not always lined up	Lead: Marketing Information
landing_page_url__c	landingPageUrl	Lead_MarketingInformation	'landing_page_url__c',	'landingPageUrl___Lead_MarketingInformation',			String	Y	Y	Can be null & not always lined up	Lead: Marketing Information
google_campaign__c	googleCampaign	Lead_MarketingInformation	'google_campaign__c',	'googleCampaign___Lead_MarketingInformation',			String	Y	Y	Optional	Lead: Marketing Information
leadsource	leadsource	Lead_MarketingInformation	'leadsource',	'leadsource___Lead_MarketingInformation',			String	Y	Y	Can have nulls	Lead: Marketing Information
converteddate	converteddate	Lead_ConversionInformation	'converteddate',	'converteddate___Lead_ConversionInformation',	Standardize date		Date	N	Y	3/19/2015	Lead: Conversion Information
status_reason__c	statusReason	Lead_ConversionInformation	'status_reason__c',	'statusReason___Lead_ConversionInformation',	Filter out duplicates		Text	N	Y	Can have duplicates – but is there	Lead: Conversion Information
status	status	Lead_ConversionInformation	'status',	'status___Lead_ConversionInformation',			String	Y - Target	Y	Can have leads that are open	Lead: Conversion Information
id	PK_LeadID	Lead_ImportantJoinKey	'id',	'PK_LeadID___Lead_ImportantJoinKey',			String	N	Y		Lead: Important ID Info
convertedaccountid	FK_LeadtoAccount	Lead_ImportantJoinKey	'convertedaccountid',	'FK_LeadtoAccount___Lead_ImportantJoinKey',			String	N	Y		Lead: Important Join Key
convertedcontactid	FK_LeadtoContact	Lead_ImportantJoinKey	'convertedcontactid',	'FK_LeadtoContact___Lead_ImportantJoinKey',			String	N	Y		Lead: Important Join Key
convertedopportunityid	FK_LeadtoOpportunity	Lead_ImportantJoinKey	'convertedopportunityid',	'FK_LeadtoOpportunity___Lead_ImportantJoinKey',			String	N	Y		Lead: Important Join Key
ownerid	FK_LeadtoUser	Lead_ImportantJoinKey	'ownerid',	'FK_LeadtoUser___Lead_ImportantJoinKey',			String	N			Lead: Important Join Key
createdbyid	createdbyid	Lead_ImportantSystemInfo	'createdbyid',	'createdbyid___Lead_ImportantSystemInfo',			String	N			Lead: Important System Info
createddate	createddate	Lead_ImportantSystemInfo	'createddate',	'createddate___Lead_ImportantSystemInfo',	Need to Standardize to date		Date	Y			Lead: Important System Info
duplicate_lead__c	duplicateLead	Lead_ImportantSystemInfo	'duplicate_lead__c',	'duplicateLead___Lead_ImportantSystemInfo',	Need to filter out duplicate leads		Boolean	N		Useless field	Lead: Important System Info
isconverted	isconverted	Lead_ImportantSystemInfo	'isconverted',	'isconverted___Lead_ImportantSystemInfo',			Boolean	N - Target Check			Lead: Important System Info
isdeleted	isdeleted	Lead_ImportantSystemInfo	'isdeleted',	'isdeleted___Lead_ImportantSystemInfo',	Filter out		Boolean	N			Lead: Important System Info

Lead | Intro Call | Oppty | Oppty Line Items | Account | InitialSelect_Lists | Rename MasterDataSet

Given how frequently the metadata and schemas in Salesforce change in the start-up, I needed to query for all the relevant columns and fields and export csv samples for a visual inspection of the available data names, types and quality. Using the summaries, I manually constructed a data catalog showing the objects, related fields, the data warehouse names, the new names, necessary data transformations as well as possible data quality issues. After completing the first round of checks and evaluations and labeling fields which could be used for prediction or labeling, I created strings of candidate fields to subset the queried tables.

Once the data frames were subset, the columns were renamed, indicating the attribute, originating object, and type of attribute.

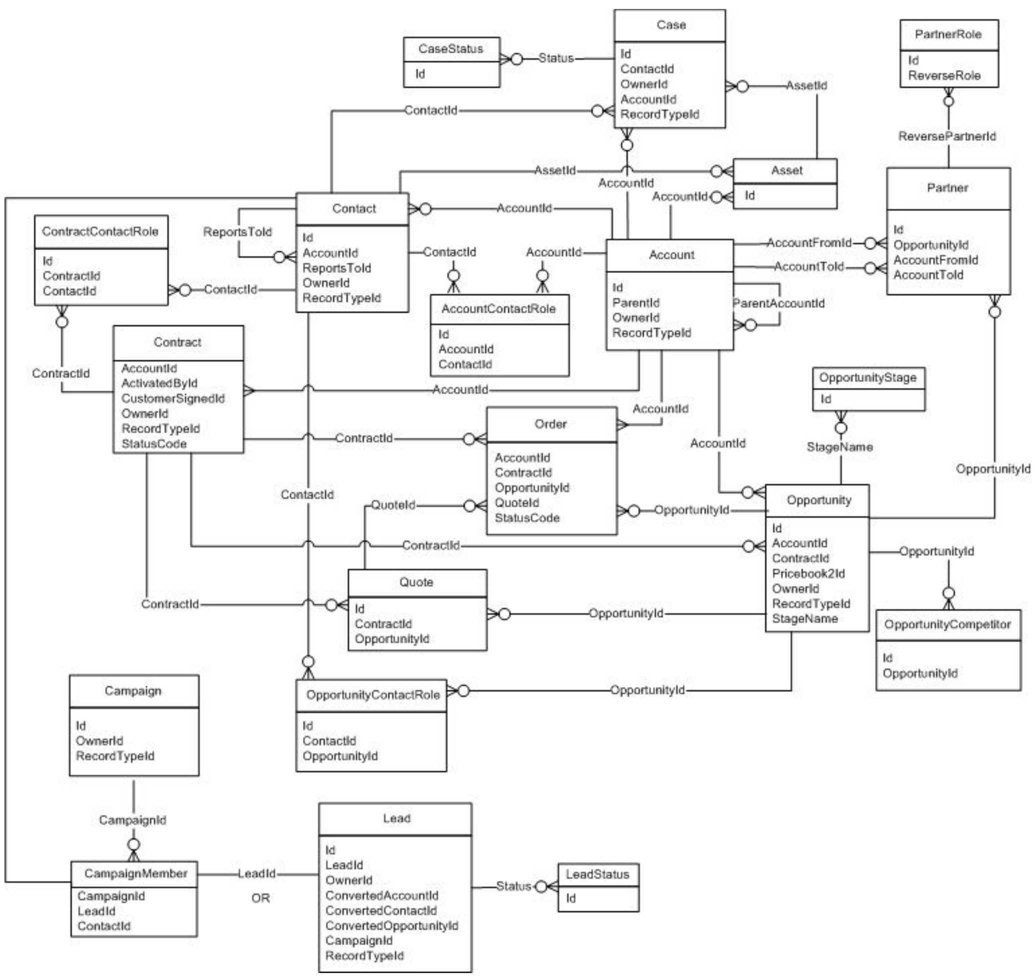


For example: “email___Lead_PersonalInformation” indicates the field describes Personal Information about a lead (in this case the email address) and came from the Lead dataframe (which was derived from the Lead table in the data warehouse). Another example: the field ‘PK_OpptyID___Oppty_ImportantJoinKey' indicates that the field is the Opportunity ID, is meant to be used in a join, and is the primary key that describes an opportunity instance.

The naming convention and detailed data catalog is crucial for a few reasons: (1) similarly named fields could be duplicated across Salesforce objects without necessarily being related or exact - especially in the case of field mismatches; (2) real-time feedback was being given back to the data engineering team about data quality issues and solutions; (3) multiple data sets for exploration were being created and ideally we’d need clarity for the order of joins for 1:M and M:M relationships.

The subsetted data frames were then joined via the table/object keys identified during the data cataloging using merge. Two master data sets were created, masterDataSet where each demo call is a unique row (left joined by Leads, Opportunities, Accounts) and a masterDataSet_product where each product line item from the opportunity was left joined and enriched by the masterDataSet. The rationale for creating split data sets was to be able to accurately classify demo call outcomes but have the product data set available in order to facilitate exploratory analysis around product purchases, SLA’s, and support add-ons.

In sales analytics and sales operations, time stamping is crucial to: (1) estimating velocity of opportunity pipeline, (2) triaging individual opportunities for attention, (3) ensuring sales teams are meeting the agreed upon standard of performance.



https://developer.salesforce.com/docs/atlas.en-us.api.meta/api/sforce_api_erd_majors.htm

One complication was that all the time fields were different data types and had different patterns. For example: ‘2018-11-08T20:12:05.000Z’, ‘Q1-2015’, ‘10/17/2014 17:09’, ‘10/28/2014’ are just four examples of the 10+ data columns that needed to be parsed and converted to a datetime object. I wrote a function that would take a dataframe, the target column to be parsed, name of a new column, and the date time pattern to pass in. The function clean_dates takes the specified columns, parses the timedate string, returns a datetime object as the new column and deletes the old column.

The next import step was re-grouping the categorical features. After inspecting the different grouped columns and values, I remapped the values using dictionaries containing the new group values and deleted the old columns.

	A	B	E	F	J	K	O	P
1	Target___IntroCall_Outcome		rejectedReason___IntroCall_Outcome		status___Lead_ConversionInformation		trafficChannel___Lead_MarketingInformation	
2	**Original Value**	**New Label**	**Original Value**	**New Label**	***Original Value***	***New Value***	***Original Value***	***New Value***
3	Attributed	Qualified	Company Too Small	Wrong_Demographic	Cancelled	Not Qualified	Affiliate	Affiliate
4	Cancelled	Not Qualified	Does Not See Benefit of WalkMe	Not_Interested	Conference Rejuvenated	Open	Banner	Other
5	No Show	Not Qualified	Existing Opportunity	Duplicate	Contacted	Open	Bing	Bing
6	Qualified	Qualified	No Budget/Price Too High	Price_Too_High	Converted	Qualified	Biz Dev	Other
7	Rejected	Rejected	No Commercial influence	Not_Right_Person	Engaged	Open	BizoLi	Other
8	Rescheduling	Open	Not A Use Case Fit	Not_Interested	Finished Sequence	Open	Brand	Brand
9	Scheduled	Open	Not Decision Maker	Not_Right_Person	Junk	Not Qualified	Bulk upload - R&D's bug	Other
0			Other (please specify)	Other	Moved to SE	Open	Conference Emails	Email
1			Project Fully Outsourced	Other	No Show	Not Qualified	Conferences	Event/Conference
2			Startup - Too Expensive	Price_Too_High	No longer with company	Not Qualified	Conferences Lead Swaps	Event/Conference
3			Too Few Users	Wrong_Demographic	Not Relevant	Not Qualified	Customer Engagement Event	Event/Conference
4	"Attributed":"Qualified",		Wrong Source Version	Other	Nurture	Open	Email Nurturing	Email
5	"Cancelled":"Not Qualified",		Wrong Timing	Not_Interested	Nurture (Outbound)	Open	Email Nurturing Conferences	Email
6	"No Show":"Not Qualified",		__	Other	Open	Open	EmailMarketing	Email
7	"Qualified":"Qualified",				Prospecting	Open	EmailNurturing	Email

Fields with long text data were also dropped due to inexperience with NLP techniques.
Additional columns were created in order to measure the duration between various stages of leads, demos, and opportunities.
After creating all the necessary columns for exploratory analysis, additional id and demographic columns were deleted.

III. Initial findings from exploratory analysis

Even before starting the analysis and data exploration, I had some assumptions about what factors could be the biggest contributors to Intro Call qualification.

Below is a table of features I assumed would impact Qualification Status of Intro Calls and the results of analysis.

Variable	Summary of Visual Inspection	Summary of Statistical Inspection
• Intro Call Creation Date (Month, Year)	No clear pattern	
• Marketing Channel	Difficult to discern but could be a driver of qualification	
• Customer Type	Could be a driver of qualification	
• Country	Difficult to discern, would need to perform goodness-of-fit test potentially	
• Landing Page	Difficult to discern, would need to perform goodness-of-fit test potentially	
• Total Calls & Emails		• Permutation Test: ○ P-val: 0.0000 • Bootstrap Test: ○ P-val: 0.0000 • Mann-Whitney: ○ P-val: 6.468e-58 • Welch’s T-Test: ○ P-val: 5.839e-91
• Lead Score		• Permutation Test: ○ P-val: 0.0000

		• Bootstrap Test: ○ P-val: 0.0000 • Mann-Whitney: ○ P-val: 1.24e-107 • Welch's T-Test: ○ P-val: 3.89e-99
• Intro Call Creation Delta		• Permutation Test: ○ P-val: 0.1170 • Bootstrap Test: ○ P-val: 0.0588 • Mann-Whitney: ○ P-val: 2.43857e-05 • Welch's T-Test: ○ P-val: 0.116

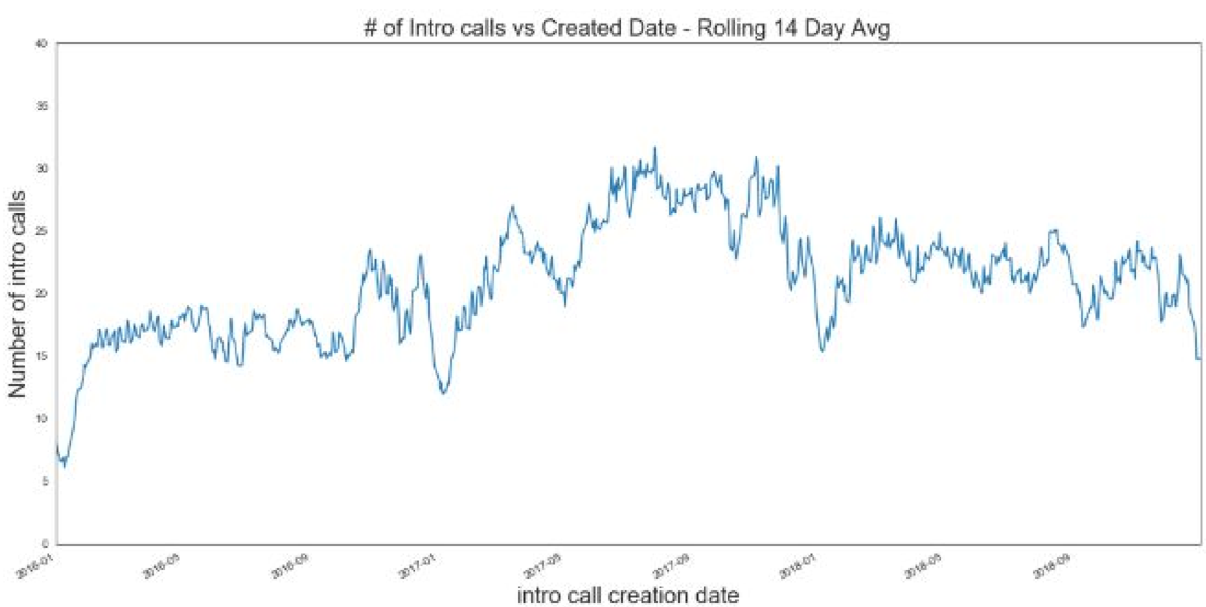
For deeper statistical analysis I focused on Lead Score, Intro Call/Lead Creation Delta, and Total Calls/Emails to determine strength of association with Qualification Status.

Prior to performing statistical analysis I visually inspected the list of variables to identify initial trends.

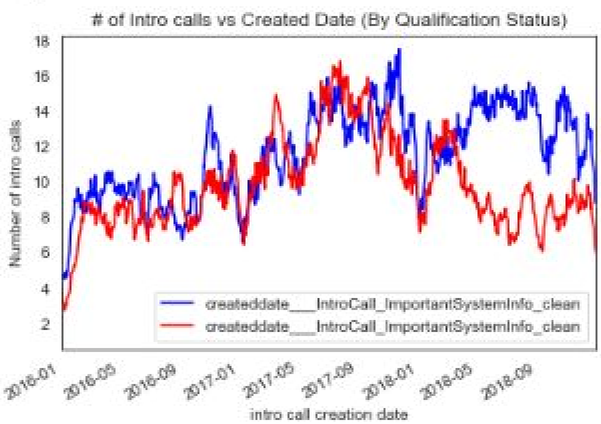
The first question I had was how many records I had in my dataset and the number of qualified vs. disqualified demo calls. Of 22.9K records, 12.8K (56%) were qualified vs 10K (44%) disqualified. An interesting extension to this question is the volume of leads necessary to get to these qualification rates.

Intro Call Volume:

Next I wanted to understand how the volume of intro calls has changed over time, both overall numbers and by qualification status (qualified vs disqualified).



But what we're really interested in is understanding the drivers of qualified calls, so it makes sense to view the volume by time and qualification status.



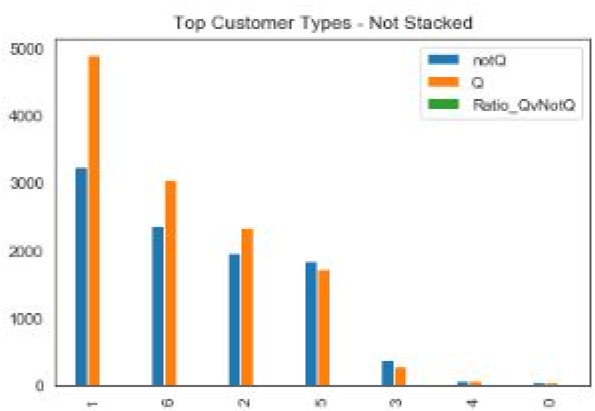
Red line is "unqualified", blue line is "qualified". It looks like volume has remained high in the last year but disqualifieds are making up a smaller proportion (possibly reinforcing the marketing team's assertion that they're providing higher quality leads).

Next I want to understand the sources of the intro calls (like marketing channels, landing pages, business type, etc).

Marketing Channels:

What's interesting is that even though some marketing channels have produced a large volume of qualifieds, they aren't necessarily the same channels responsible for producing a high ratio of qualifieds to disqualifieds.

For example, lead source 3 & 0 (corresponding to "Brand" and "Affiliate") have a ratio of 1.9 but are in sixth place and up, with additional intro call sources in between having a ratio of around 0.8~1.6. This is the first time I've seen the marketing funnel from the perspective of the intro calls (even within the company) so it's fascinating to see the different levels of quality.

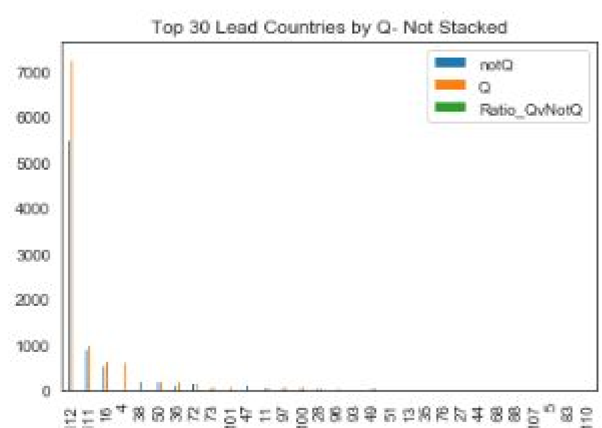


Customer Type:

The next question I was trying to answer was whether the customer type could be a driver of qualified intro calls.

After generating the following charts, it seems that customer type could be a driver (as well as an indicator of the company's strategic focus on the enterprise space).

1 corresponds to 'Enterprise' (2 is 'Unknown', which doesn't exactly bode the best in terms of our data quality) and 4 corresponds to 'Nonprofits' (which makes sense, the company primarily markets to companies willing to invest significant resources in onboarding and digital adoption).

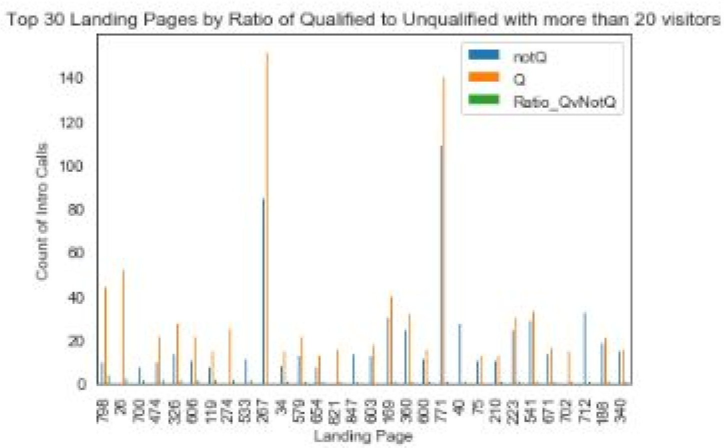


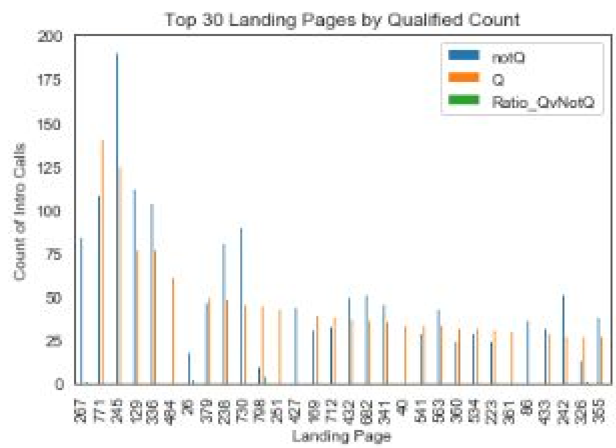
Countries:

Countries is a little surprising as we have some EMEA and ANZ/APAC countries listed as the top producers of qualified intro calls. The company started in Israel and has major presence in AMER but it’s interesting to see the UK (#111), Australia (#4), and Germany (#38) up in the top 8.

Landing Pages:

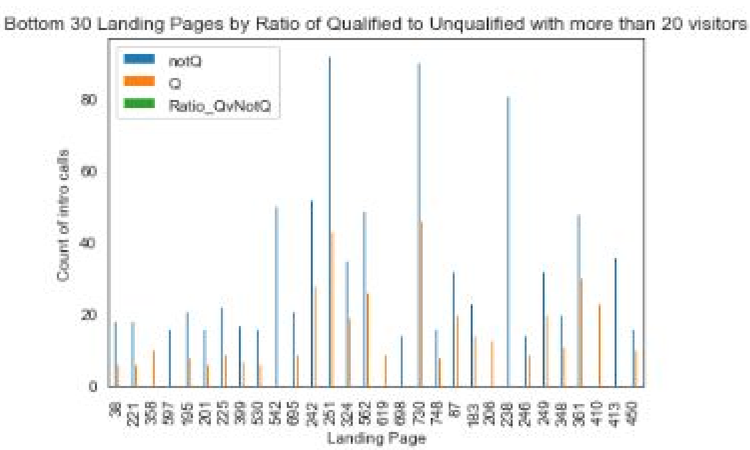
When we look at landing pages and try to create top 30 charts, we see some interesting trends where the top 30 best landing pages by qualifieds count aren’t the same as the top 30 landing pages by ratio of qualified to disqualified intro

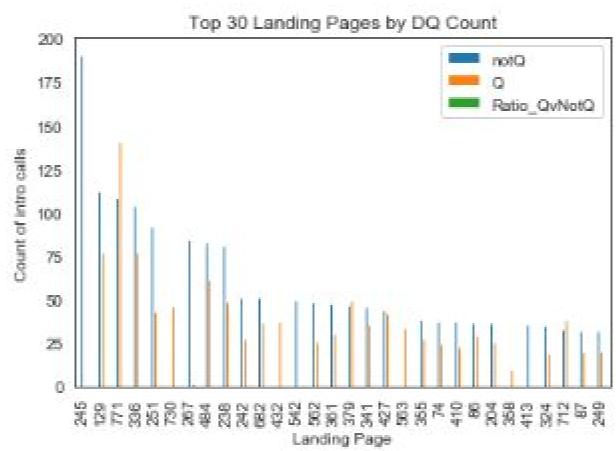




calls.

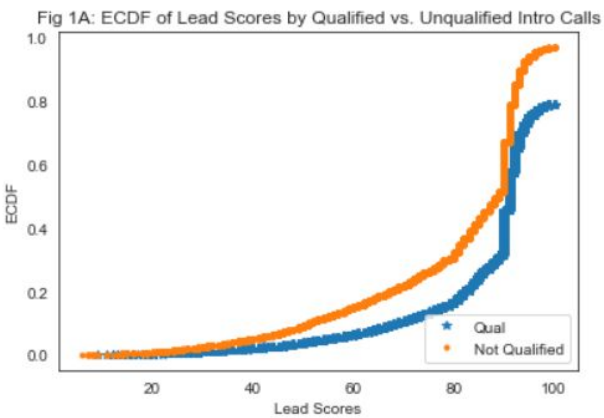
The first set of charts below show “Top 30 Landing Pages by Qualified Count” and “Top 30 Landing Pages by Ratio (Qualified/Disqualified)”. Landing pages had to have more than 20 visitors in order to weed out cases where only 5 people visited a landing page and all converted (or none converted). Notice how the top 5 landing pages are completely different depending on the particular cut.
This next set of charts focuses on landing page rankings of disqualified intro calls.

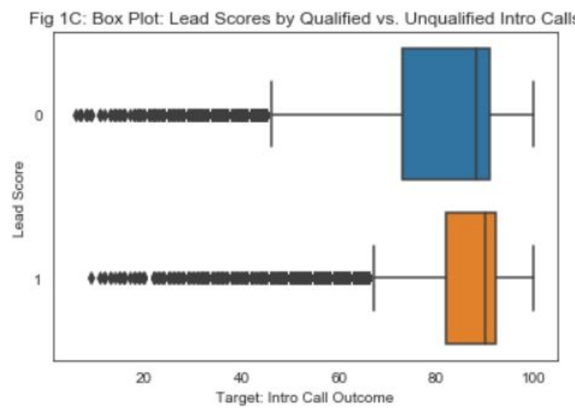




Lead Score:

- 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).





- 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

- 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

Fig 1E: Histogram of Shifted Arrays of Qualified & Disqualified Intro Calls for Boostrap Hypothesis

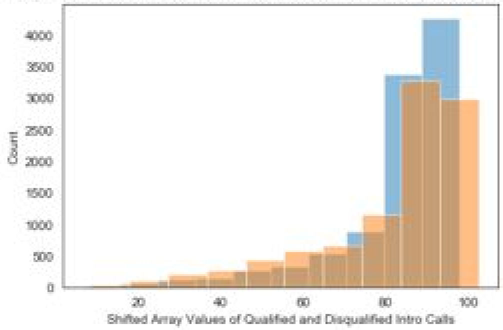
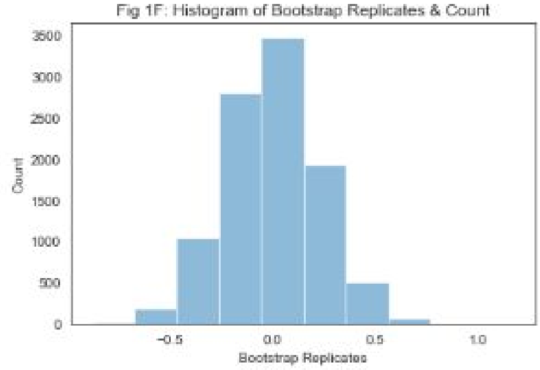


Fig 1F: Histogram of Bootstrap Replicates & Count



- Third & 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.

Total Calls & Emails:

- 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%)]).]

Fig 2A: ECDF of Total Calls & Emails by Qualified vs. Unqualified Intro Calls

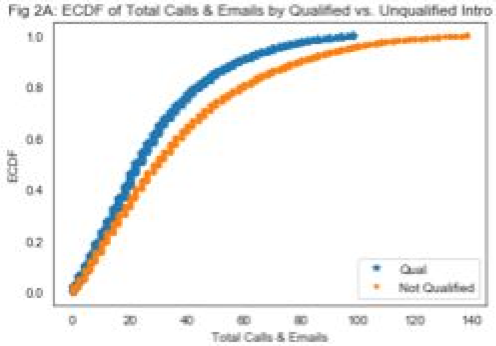
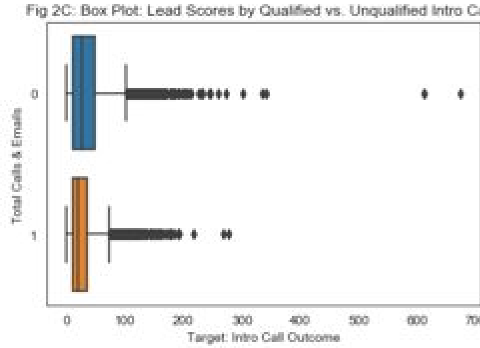


Fig 2C: Box Plot: Lead Scores by Qualified vs. Unqualified Intro Calls



- 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

- 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

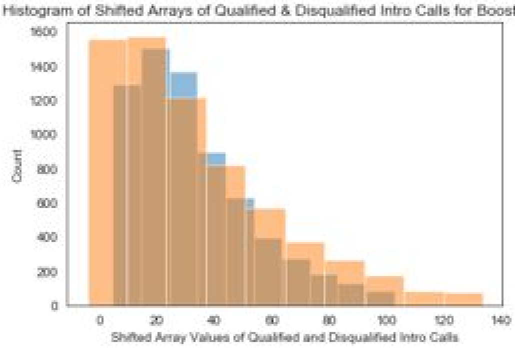


Fig 2E: Histogram of Shifted Arrays of Qualified & Disqualified Intro Calls for Boostrap Hypothesis

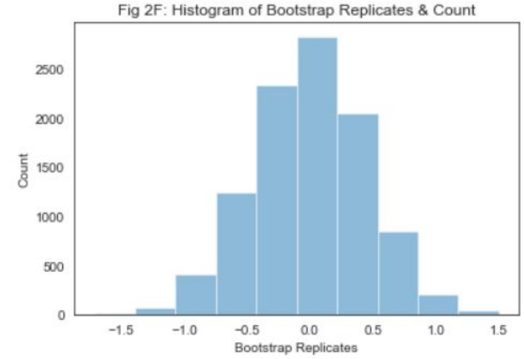


Fig 2F: Histogram of Bootstrap Replicates & Count

- Third & 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.

Lead - Intro Call Created Delta:

- 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).

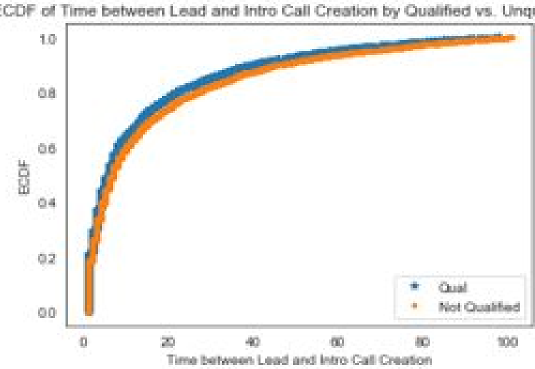


Fig 3A: ECDF of Time between Lead and Intro Call Creation by Qualified vs. Unqualified Intro Calls

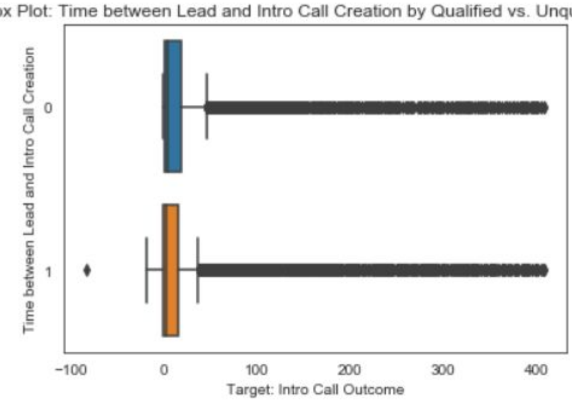
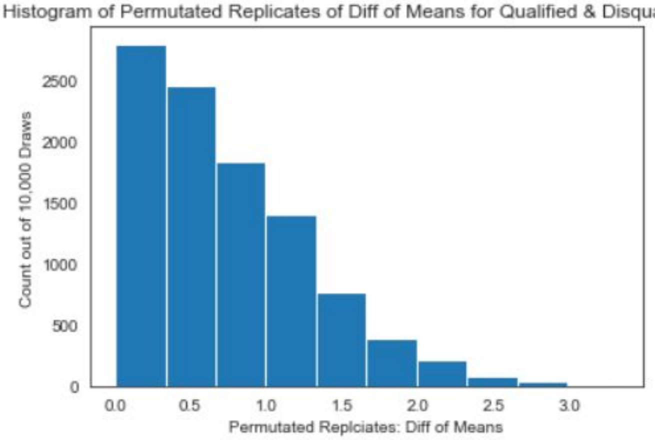


Fig 3C: Box Plot: Time between Lead and Intro Call Creation by Qualified vs. Unqualified Intro Calls

- 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

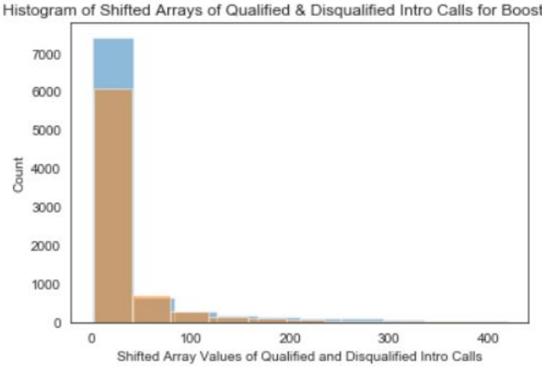
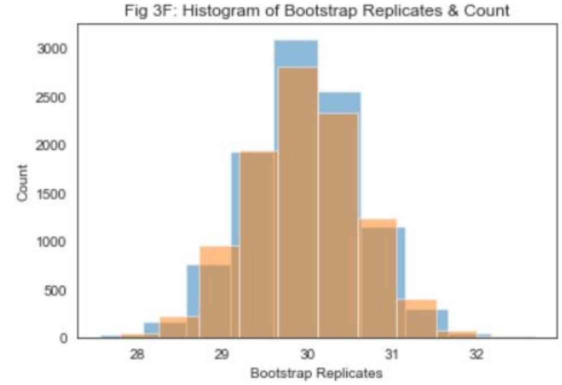


Fig 3F: Histogram of Bootstrap Replicates & Count



- Third & 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).