Source Code

August 30, 2023

[1]:
```
import numpy as pd
import pandas as pd
import matplotlib.pyplot as plt
pd.set_option('display.max_columns', 500)
import numpy as np
```

[2]:
```
train = pd.read_csv(r'Dataset/Train.csv')
test = pd.read_csv(r'Dataset/Test.csv')
test1 = pd.read_csv(r'Dataset/Test.csv')
sample = pd.read_csv(r'Dataset/Sample_Submission.csv')
```

[3]:
```
sample
```

[3]:
```
   Deal_title  Success_probability
0  TitleM5DZY                 48.6
1  TitleKIW18                 33.9
2  TitleFXSDN                 43.8
3  TitlePSK4Y                 39.5
4  Title904GV                 37.4
```

[4]:
```
def make_csv(predict, name):
    df = pd.DataFrame()
    df['Deal_title'] = test1['Deal_title']
    df['Success_probability'] = predict
    df.to_csv(name, index=False)
```

Ensure that the probability of success for the train is adjusted to a positive value during the process of data cleaning.

[5]:
```
train.Success_probability = train.Success_probability.abs()
```

[6]:
```
train
```

[6]:
```
      Deal_title                        Lead_name              Industry  \
0     TitleM5DZY    Davis, Perkins and Bishop Inc           Restaurants
1     TitleKIW18                  Bender PLC LLC  Construction Services
2     TitleFXSDN           Carter-Henry and Sons      Hospitals/Clinics
3     TitlePSK4Y                  Garcia Ltd Ltd           Real Estate
```

```
4     Title904GV                  Lee and Sons PLC     Financial Services
...          ...                               ...                    ...
7002  TitleJ7TDY                 Flowers-Adams PLC                 Banks
7003  TitleO1IIN  Kelly, Smith and Benton and Sons      Hospitals/Clinics
7004  TitleF1FWY                Romero-Juarez PLC         Semiconductors
7005  TitleXVBNJ  Sullivan, Mooney and Elliott LLC                 Banks
7006  TitleKXU3H               Jones LLC and Sons     Financial Services

     Deal_value Weighted_amount Date_of_creation      Pitch  \
0       320506$      2067263.7$       2020-03-29  Product_2
1        39488$       240876.8$       2019-07-10  Product_2
2       359392$      2407926.4$       2019-07-27  Product_1
3        76774$       468321.4$       2021-01-30  Product_2
4       483896$             NaN       2019-05-22  Product_2
...         ...             ...              ...        ...
7002    192800$      1195360.0$       2020-12-07  Product_1
7003    220208$      1453372.8$       2020-03-13  Product_2
7004    253608$             NaN       2020-03-10  Product_1
7005    118615$       794720.5$       2019-12-26  Product_1
7006    258627$     1642281.45$       2020-04-27  Product_2

               Contact_no             Lead_revenue Fund_category Geography  \
0             607.447.7883        50 - 100 Million    Category 2       USA
1             892-938-9493  500 Million - 1 Billion    Category 4     India
2             538.748.2271  500 Million - 1 Billion    Category 4       USA
3     (692)052-1389x75188  500 Million - 1 Billion    Category 3       USA
4     001-878-814-6134x015        50 - 100 Million    Category 3     India
...                    ...                     ...           ...       ...
7002            1734434912       100 - 500 Million    Category 4       NaN
7003  (002)106-0243x03346       100 - 500 Million    Category 1       NaN
7004     777-231-4109x712       100 - 500 Million    Category 2       USA
7005  001-212-631-5467x377  500 Million - 1 Billion    Category 2       USA
7006      +1-240-958-7728  500 Million - 1 Billion    Category 2       USA

                            Location          POC_name  \
0                  Killeen-Temple, TX   Charlene Werner
1                             Ratlam             rakhi
2          Albany-Schenectady-Troy, NY    Ariel Hamilton
3           Mount Vernon-Anacortes, WA       Erin Wilson
4                             Shimoga             kavita
...                               ...               ...
7002                       Kagaznagar    smt. chanchala
7003                        Proddatur     geeta @ komal
7004                  Rocky Mount, NC   Nicholas Duncan
7005  Atlanta-Sandy Springs-Roswell, GA    Wayne Williams
7006                    Brunswick, GA  Carol Strickland
```

```
                                      Designation  \
0                            Executive Vice President
1                              Chairman/CEO/President
2                                SVP/General Counsel
3                             CEO/Co-Founder/Chairman
4                            Executive Vice President
...                                               ...
7002                          CEO/Co-Founder/Chairman
7003                                              CEO
7004                             SVP/General Counsel
7005                         Executive Vice President
7006  Vice President / GM (04-present) : VP Sales an...

                     Lead_POC_email               Hiring_candidate_role  \
0            charlenewerner@davis.com               Community pharmacist
1               terrylogan@bender.com             Recruitment consultant
2         arielhamilton@carterhenry.com             Health service manager
3                 erinwilson@garcia.com     Therapist, speech and language
4               mr.christopher@lee.com                      Media planner
...                                ...                                ...
7002  clairewilliams@flowersadams.com               Call centre manager
7003        vanessaanderson@kelly.com             Financial risk analyst
7004  nicholasduncan@romerojuarez.com        Nature conservation officer
7005       waynewilliams@sullivan.com                  Designer, textile
7006        carolstrickland@jones.com  Government social research officer

          Lead_source Level_of_meeting                Last_lead_update  \
0              Website          Level 3                        No track
1               Others          Level 1  Did not hear back after Level 1
2     Marketing Event          Level 1                               ?
3       Contact Email          Level 2  Did not hear back after Level 1
4              Website          Level 2                      Up-to-date
...                ...              ...                             ...
7002    Contact Email          Level 1           More than a week back
7003  Marketing Event          Level 2                               ?
7004  Marketing Event          Level 3                             NaN
7005          Website          Level 3                             NaN
7006    Contact Email          Level 3           More than a week back

                  Internal_POC                     Resource  \
0              Davis,Sharrice A                          NaN
1                Brown,Maxine A                           No
2     Georgakopoulos,Vasilios T                           No
3                Brown,Maxine A  We have all the requirements
4                 Thomas,Lori E                           No
...                         ...                          ...
7002            Jones,Eyvette W  We have all the requirements
```

```
7003             Brown,Maxine A  We have all the requirements
7004              Logan,Kevin N                           NaN
7005              Cashin,Marc C  We have all the requirements
7006           Massiah,Gerard F  We have all the requirements

      Internal_rating  Success_probability
0                   3                73.60
1                   5                58.90
2                   4                68.80
3                   1                64.50
4                   4                62.40
...               ...                  ...
7002                4               107.34
7003                3                26.35
7004                1                70.60
7005                3                 5.00
7006                1                68.70

[7007 rows x 23 columns]
```

We will perform data cleaning on the numerical data first.

[7]:
```
train.describe()
```

[7]:
```
       Internal_rating  Success_probability
count      7007.000000          7007.000000
mean          3.009562            64.845034
std           1.418666            17.566890
min           1.000000             5.000000
25%           2.000000            60.600000
50%           3.000000            65.300000
75%           4.000000            69.600000
max           5.000000           107.340000
```

[8]:
```
test.describe()
```

[8]:
```
       Internal_rating
count      2093.000000
mean          3.143354
std           4.510451
min          -1.000000
25%           2.000000
50%           3.000000
75%           4.000000
max          82.340000
```

There appears to be a problem with the Internal_Rating feature in the test dataset.It should only contain ratings in the range of 1 to 5.

[9]:
```
test.Internal_rating.value_counts()
```

[9]:
```
 3.00     421
 5.00     417
 2.00     411
 1.00     399
 4.00     391
-1.00      48
 82.34      6
Name: Internal_rating, dtype: int64
```

[10]:
```
min_one_test = test[(test.Internal_rating == -1) | (test.Internal_rating == 1)]
```

[11]:
```
min_one_test.head()
```

[11]:

	Deal_title	Lead_name	Industry	Deal_value	Weighted_amount	Date_of_creation	Pitch	Contact_no	Lead_revenue	Fund_category	Geography	Location	POC_name	Designation	Lead_POC_email	Hiring_candidate_role	Lead_source
0	TitleAD16O	Bonilla Ltd Inc	Investment Bank/Brokerage	200988$	NaN	2020-04-15	Product_1	167.332.2751x989	100 - 500 Million	Category 4	India	Bhubaneshwar	sonia	Chairman/CEO/President	maureenthomas@bonilla.com	Designer, fashion/clothing	Marketing Event
1	TitleOW6CR	Williams, Rogers and Roach PLC	Electronics	409961$	2541758.2$	2021-01-23	Product_1	001-486-903-0711x7831	100 - 500 Million	Category 3	USA	Coeur d'Alene, ID	Daniel Bell	CEO/Co-Founder/Chairman	danielbell@williams.com	Horticultural consultant	Marketing Event
6	Title8CF62	Chen LLC PLC	Insurance	300288$	1711641.6$	2019-10-17	Product_1	299.856.4096x1588	100 - 500 Million	Category 2	USA	Pensacola-Ferry Pass-Brent, FL	Jamie Allen	Chairman/CEO/President	jamieallen@chen.com	Quality manager	Contact Email
9	TitleXK47X	Short-Simpson and Sons	Technology Consulting	452127$	3029250.9$	2019-03-07	Product_1	(805)537-7827x3795	500 Million - 1 Billion	Category 1	USA	Pittsfield, MA	Tyler Lucas	CEO			
11	TitleR9W92	Rogers, Alexander and Smith Inc	Semiconductors	300968$	1775711.2$	2020-09-20	Product_1	712-632-7468x301	50 - 100 Million	Category 2	USA	Gulfport-Biloxi-Pascagoula, MS	Eric Nielsen	CEO			

```
9   tylerlucas@shortsimpson.com              Engineer, site  Marketing Event
11       ericnielsen@rogers.com    Technical sales engineer          Website

   Level_of_meeting                     Last_lead_update      Internal_POC  \
0           Level 1                    more than a month  Massiah,Gerard F
1           Level 2                           Up-to-date    Smith,Keenan H
6           Level 1  Following up but lead not responding  Salyers,Daniel L
9           Level 2                                    ?   Rocks,Michael J
11          Level 2  Following up but lead not responding  Davis,Sharrice A

   Resource  Internal_rating
0        No            -1.0
1       Yes             1.0
6       Yes             1.0
9        No             1.0
11      Yes             1.0
```

[12]:
```
test.Internal_rating = test.Internal_rating.abs()
```

[13]:
```
test.Internal_rating.value_counts()
```

[13]:
```
1.00     447
3.00     421
5.00     417
2.00     411
4.00     391
82.34      6
Name: Internal_rating, dtype: int64
```

[14]:
```
weird_test = test[test.Internal_rating == 82.34]
```

[15]:
```
weird_test
```

[15]:
```
      Deal_title                          Lead_name           Industry  \
433   Title4KKQU         Johnson-Gutierrez and Sons  Peripherals/Comp.
1044  Title92JGX                 Hurley-Russell Inc               ISPs
1329  TitleSN96Y     Sutton, Walls and Williams Ltd           Services
1399  TitleVEJGJ  Oconnor, Graham and Sanders Group         Publishing
1475  TitleH5282                 Ware-Williams LLC              Banks
1931  TitleU5YAN              Evans-Burton and Sons        Wood/Timber

     Deal_value Weighted_amount Date_of_creation      Pitch  \
433      89249$       580118.5$       2019-05-28  Product_1
1044    250074$      1788029.1$       2020-06-14  Product_2
1329    304166$      2098745.4$       2020-10-18  Product_2
1399    160504$      1099452.4$       2020-01-08  Product_2
1475    347583$      2224531.2$       2021-01-12  Product_1
```

```
1931    364375$     2277343.75$       2019-07-20  Product_1

               Contact_no             Lead_revenue Fund_category Geography  \
433       642.371.3952x574  500 Million - 1 Billion    Category 1     India
1044      648.197.4666x470        100 - 500 Million    Category 1     India
1329  001-237-888-3906x081        100 - 500 Million    Category 1     India
1399            3931431891  500 Million - 1 Billion    Category 4     India
1475      027-607-4686x230         50 - 100 Million    Category 4     India
1931   (857)928-7428x39346        100 - 500 Million    Category 2     India

        Location       POC_name  \
433        Surat      baby  bai
1044     Kolkata         praveen
1329  Rajahmundry         pushpa
1399    Bamanpuri  aakanksha d/o
1475    Bhavnagar        santosh
1931    Brahmapur          pinki

                                        Designation  \
433                            SVP/General Counsel
1044                          CEO/Chairman/President
1329                                  CEO/President
1399                         Executive Vice President
1475  Vice President / GM (04-present) : VP Sales an...
1931                          CEO/Co-Founder/Chairman

                     Lead_POC_email                  Hiring_candidate_role  \
433   joeharrell@johnsongutierrez.com                     Personal assistant
1044  gracejennings@hurleyrussell.com  Engineer, control and instrumentation
1329        jeffreywilson@sutton.com                          Stage manager
1399        walterrivera@oconnor.com                   Clinical biochemist
1475   carriephelps@warewilliams.com                Secretary/administrator
1931      markpalmer@evansburton.com               Insurance claims handler

         Lead_source Level_of_meeting                     Last_lead_update  \
433     Contact Email          Level 2                    More than 2 weeks
1044  Marketing Event          Level 3                                  NaN
1329          Website          Level 3  Following up but lead not responding
1399          Website          Level 2                    more than a month
1475           Others          Level 3                More than a week back
1931           Others          Level 3                              Pending

                 Internal_POC        Resource  Internal_rating
433           Massiah,Gerard F      Not enough            82.34
1044            Hameier,Kurt E             Yes            82.34
1329  Abdul-Hamid,Saud Muhamad  Cannot deliver            82.34
1399               Ali,Mohamed     Deliverable            82.34
```

```
1475         Sutton,Michelle R  Cannot deliver            82.34
1931            Booker,David L             Yes            82.34
```

I intend to determine the Internal Rating value according to the Level_of_Meeting.

[16]:
```
weird_id = weird_test.index
```

[17]:
```
for x in weird_id:
    if test.loc[x,'Level_of_meeting'] == 'Level 2':
        test.loc[x,'Internal_rating'] = 2
    else:
        test.loc[x,'Internal_rating'] = 3
```

[18]:
```
test.Internal_rating.value_counts()
```

[18]:
```
1.0    447
3.0    425
5.0    417
2.0    413
4.0    391
Name: Internal_rating, dtype: int64
```

[19]:
```
train.shape
```

[19]:
```
(7007, 23)
```

There also appears to be an problem with the Success_Probability feature in the train dataset. If it's meant to be a percentage, it should only contain values within the range of 1 to 100.

[20]:
```
train.Success_probability.value_counts().head()
```

[20]:
```
107.34    515
26.35     447
65.60      74
62.40      73
5.00       70
Name: Success_probability, dtype: int64
```

[21]:
```
train[train.Success_probability == 100].shape
```

[21]:
```
(0, 23)
```

[22]:
```
train.Success_probability = train.apply(lambda x: 100 if␣
↪x['Success_probability'] == 107.34 else x['Success_probability'], axis=1)
```

[23]:
```
train.Success_probability.value_counts()
```

[23]:
```
100.00    515
26.35     447
```

```
65.60      74
62.40      73
5.00       70
          ...
50.90       1
78.80       1
51.90       1
54.30       1
52.20       1
Name: Success_probability, Length: 248, dtype: int64
```

[24]:
```
num_type = ['int64', 'float64']
```

Initially, we will save the Success Probability as the target variable.

[25]:
```
train_y = train.Success_probability
```

[26]:
```
train_x = train.drop('Success_probability', axis=1).copy()
```

[27]:
```
train_x
```

[27]:
```
      Deal_title                         Lead_name              Industry  \
0     TitleM5DZY     Davis, Perkins and Bishop Inc           Restaurants
1     TitleKIW18                 Bender PLC LLC  Construction Services
2     TitleFXSDN           Carter-Henry and Sons     Hospitals/Clinics
3     TitlePSK4Y                 Garcia Ltd Ltd           Real Estate
4     Title904GV                Lee and Sons PLC     Financial Services
...          ...                               ...                   ...
7002  TitleJ7TDY               Flowers-Adams PLC                 Banks
7003  TitleO1IIN  Kelly, Smith and Benton and Sons     Hospitals/Clinics
7004  TitleF1FWY               Romero-Juarez PLC        Semiconductors
7005  TitleXVBNJ  Sullivan, Mooney and Elliott LLC                 Banks
7006  TitleKXU3H              Jones LLC and Sons     Financial Services

     Deal_value Weighted_amount Date_of_creation      Pitch  \
0       320506$      2067263.7$       2020-03-29  Product_2
1        39488$       240876.8$       2019-07-10  Product_2
2       359392$      2407926.4$       2019-07-27  Product_1
3        76774$       468321.4$       2021-01-30  Product_2
4       483896$             NaN       2019-05-22  Product_2
...         ...             ...              ...        ...
7002    192800$      1195360.0$       2020-12-07  Product_1
7003    220208$      1453372.8$       2020-03-13  Product_2
7004    253608$             NaN       2020-03-10  Product_1
7005    118615$       794720.5$       2019-12-26  Product_1
7006    258627$     1642281.45$       2020-04-27  Product_2

              Contact_no            Lead_revenue Fund_category Geography  \
```

```
0             607.447.7883         50 - 100 Million    Category 2       USA
1             892-938-9493  500 Million - 1 Billion    Category 4     India
2             538.748.2271  500 Million - 1 Billion    Category 4       USA
3      (692)052-1389x75188  500 Million - 1 Billion    Category 3       USA
4     001-878-814-6134x015         50 - 100 Million    Category 3     India
...                    ...                      ...           ...       ...
7002            1734434912        100 - 500 Million    Category 4       NaN
7003   (002)106-0243x03346        100 - 500 Million    Category 1       NaN
7004      777-231-4109x712        100 - 500 Million    Category 2       USA
7005  001-212-631-5467x377  500 Million - 1 Billion    Category 2       USA
7006       +1-240-958-7728  500 Million - 1 Billion    Category 2       USA

                              Location          POC_name  \
0                    Killeen-Temple, TX   Charlene Werner
1                                 Ratlam            rakhi
2            Albany-Schenectady-Troy, NY    Ariel Hamilton
3             Mount Vernon-Anacortes, WA       Erin Wilson
4                               Shimoga            kavita
...                                 ...               ...
7002                          Kagaznagar    smt. chanchala
7003                           Proddatur     geeta @ komal
7004                       Rocky Mount, NC   Nicholas Duncan
7005  Atlanta-Sandy Springs-Roswell, GA    Wayne Williams
7006                        Brunswick, GA  Carol Strickland

                                          Designation  \
0                              Executive Vice President
1                                Chairman/CEO/President
2                                  SVP/General Counsel
3                               CEO/Co-Founder/Chairman
4                              Executive Vice President
...                                                 ...
7002                            CEO/Co-Founder/Chairman
7003                                               CEO
7004                               SVP/General Counsel
7005                           Executive Vice President
7006  Vice President / GM (04-present) : VP Sales an...

                       Lead_POC_email               Hiring_candidate_role  \
0             charlenewerner@davis.com                Community pharmacist
1                terrylogan@bender.com              Recruitment consultant
2         arielhamilton@carterhenry.com              Health service manager
3                 erinwilson@garcia.com      Therapist, speech and language
4               mr.christopher@lee.com                       Media planner
...                                ...                                 ...
7002  clairewilliams@flowersadams.com                 Call centre manager
7003        vanessaanderson@kelly.com              Financial risk analyst
```

```
7004  nicholasduncan@romerojuarez.com         Nature conservation officer
7005       waynewilliams@sullivan.com                   Designer, textile
7006        carolstrickland@jones.com  Government social research officer

          Lead_source Level_of_meeting                Last_lead_update  \
0             Website          Level 3                         No track
1              Others          Level 1  Did not hear back after Level 1
2     Marketing Event          Level 1                               ?
3       Contact Email          Level 2  Did not hear back after Level 1
4             Website          Level 2                      Up-to-date
...               ...              ...                             ...
7002    Contact Email          Level 1           More than a week back
7003  Marketing Event          Level 2                               ?
7004  Marketing Event          Level 3                             NaN
7005          Website          Level 3                             NaN
7006    Contact Email          Level 3           More than a week back

                  Internal_POC                       Resource  Internal_rating
0              Davis,Sharrice A                            NaN               3
1                Brown,Maxine A                             No               5
2     Georgakopoulos,Vasilios T                             No               4
3                Brown,Maxine A  We have all the requirements               1
4                 Thomas,Lori E                             No               4
...                         ...                            ...             ...
7002            Jones,Eyvette W  We have all the requirements               4
7003             Brown,Maxine A  We have all the requirements               3
7004              Logan,Kevin N                            NaN               1
7005              Cashin,Marc C  We have all the requirements               3
7006           Massiah,Gerard F  We have all the requirements               1

[7007 rows x 22 columns]
```

[28]:
```
train_x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7007 entries, 0 to 7006
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Deal_title            7007 non-null   object
 1   Lead_name             7007 non-null   object
 2   Industry              7006 non-null   object
 3   Deal_value            6956 non-null   object
 4   Weighted_amount       6482 non-null   object
 5   Date_of_creation      7007 non-null   object
 6   Pitch                 7007 non-null   object
 7   Contact_no            7007 non-null   object
 8   Lead_revenue          7007 non-null   object
```

```
 9   Fund_category          7007 non-null   object
 10  Geography              6035 non-null   object
 11  Location               6996 non-null   object
 12  POC_name               6999 non-null   object
 13  Designation            7007 non-null   object
 14  Lead_POC_email         7007 non-null   object
 15  Hiring_candidate_role  7007 non-null   object
 16  Lead_source            7007 non-null   object
 17  Level_of_meeting       7007 non-null   object
 18  Last_lead_update       6374 non-null   object
 19  Internal_POC           7007 non-null   object
 20  Resource               6858 non-null   object
 21  Internal_rating        7007 non-null   int64
dtypes: int64(1), object(21)
memory usage: 1.2+ MB
```

[29]:
```
num_type = ['int64', 'float64']
```

I believe I have completed the numeric data processing; now, let's proceed to address the categorical data.

[30]:
```
num_col = [x for x in train_x.columns if train_x[x].dtype in num_type]
```

[31]:
```
num_col
```

[31]:
```
['Internal_rating']
```

[32]:
```
cat_col = [x for x in train_x.columns if train_x[x].dtypes == 'O']
```

[33]:
```
len(cat_col)
```

[33]:
```
21
```

[34]:
```
train_x[cat_col].nunique()
```

[34]:
```
Deal_title              7007
Lead_name               7007
Industry                 171
Deal_value              6907
Weighted_amount         6480
Date_of_creation         777
Pitch                      2
Contact_no              7007
Lead_revenue               3
Fund_category              4
Geography                  2
Location                 597
POC_name                5261
```

```
Designation                10
Lead_POC_email           7007
Hiring_candidate_role     639
Lead_source                 4
Level_of_meeting            3
Last_lead_update           11
Internal_POC               60
Resource                    6
dtype: int64
```

[35]:
```
train_x
```

[35]:
```
      Deal_title                          Lead_name               Industry  \
0     TitleM5DZY     Davis, Perkins and Bishop Inc            Restaurants
1     TitleKIW18                     Bender PLC LLC  Construction Services
2     TitleFXSDN              Carter-Henry and Sons      Hospitals/Clinics
3     TitlePSK4Y                     Garcia Ltd Ltd            Real Estate
4     Title904GV                  Lee and Sons PLC     Financial Services
...          ...                                ...                    ...
7002  TitleJ7TDY                 Flowers-Adams PLC                  Banks
7003  TitleO1IIN  Kelly, Smith and Benton and Sons      Hospitals/Clinics
7004  TitleF1FWY                 Romero-Juarez PLC         Semiconductors
7005  TitleXVBNJ  Sullivan, Mooney and Elliott LLC                  Banks
7006  TitleKXU3H                Jones LLC and Sons     Financial Services

     Deal_value Weighted_amount Date_of_creation      Pitch  \
0       320506$      2067263.7$       2020-03-29  Product_2
1        39488$       240876.8$       2019-07-10  Product_2
2       359392$      2407926.4$       2019-07-27  Product_1
3        76774$       468321.4$       2021-01-30  Product_2
4       483896$             NaN       2019-05-22  Product_2
...         ...             ...              ...        ...
7002    192800$      1195360.0$       2020-12-07  Product_1
7003    220208$      1453372.8$       2020-03-13  Product_2
7004    253608$             NaN       2020-03-10  Product_1
7005    118615$       794720.5$       2019-12-26  Product_1
7006    258627$     1642281.45$       2020-04-27  Product_2

                Contact_no             Lead_revenue Fund_category Geography  \
0            607.447.7883        50 - 100 Million    Category 2       USA
1            892-938-9493  500 Million - 1 Billion    Category 4     India
2            538.748.2271  500 Million - 1 Billion    Category 4       USA
3     (692)052-1389x75188  500 Million - 1 Billion    Category 3       USA
4    001-878-814-6134x015        50 - 100 Million    Category 3     India
...                   ...                     ...           ...       ...
7002           1734434912       100 - 500 Million    Category 4       NaN
7003  (002)106-0243x03346       100 - 500 Million    Category 1       NaN
```

```
7004      777-231-4109x712        100 - 500 Million    Category 2       USA
7005  001-212-631-5467x377  500 Million - 1 Billion    Category 2       USA
7006       +1-240-958-7728  500 Million - 1 Billion    Category 2       USA

                            Location          POC_name  \
0                  Killeen-Temple, TX   Charlene Werner
1                               Ratlam            rakhi
2           Albany-Schenectady-Troy, NY    Ariel Hamilton
3            Mount Vernon-Anacortes, WA       Erin Wilson
4                               Shimoga            kavita
...                                 ...               ...
7002                         Kagaznagar    smt. chanchala
7003                          Proddatur     geeta @ komal
7004                  Rocky Mount, NC   Nicholas Duncan
7005  Atlanta-Sandy Springs-Roswell, GA    Wayne Williams
7006                      Brunswick, GA  Carol Strickland

                                       Designation  \
0                            Executive Vice President
1                              Chairman/CEO/President
2                                SVP/General Counsel
3                             CEO/Co-Founder/Chairman
4                            Executive Vice President
...                                               ...
7002                          CEO/Co-Founder/Chairman
7003                                              CEO
7004                             SVP/General Counsel
7005                         Executive Vice President
7006  Vice President / GM (04-present) : VP Sales an...

                     Lead_POC_email               Hiring_candidate_role  \
0            charlenewerner@davis.com                Community pharmacist
1               terrylogan@bender.com              Recruitment consultant
2       arielhamilton@carterhenry.com              Health service manager
3               erinwilson@garcia.com      Therapist, speech and language
4             mr.christopher@lee.com                       Media planner
...                              ...                                 ...
7002  clairewilliams@flowersadams.com                Call centre manager
7003       vanessaanderson@kelly.com              Financial risk analyst
7004  nicholasduncan@romerojuarez.com         Nature conservation officer
7005       waynewilliams@sullivan.com                   Designer, textile
7006        carolstrickland@jones.com  Government social research officer

         Lead_source Level_of_meeting                Last_lead_update  \
0             Website          Level 3                        No track
1              Others          Level 1  Did not hear back after Level 1
2     Marketing Event          Level 1                               ?
```

```
3       Contact Email          Level 2  Did not hear back after Level 1
4             Website          Level 2                      Up-to-date
...               ...              ...                             ...
7002    Contact Email          Level 1           More than a week back
7003  Marketing Event          Level 2                               ?
7004  Marketing Event          Level 3                             NaN
7005          Website          Level 3                             NaN
7006    Contact Email          Level 3           More than a week back

                  Internal_POC                      Resource  Internal_rating
0              Davis,Sharrice A                           NaN               3
1                Brown,Maxine A                            No               5
2     Georgakopoulos,Vasilios T                            No               4
3                Brown,Maxine A  We have all the requirements               1
4                 Thomas,Lori E                            No               4
...                         ...                           ...             ...
7002            Jones,Eyvette W  We have all the requirements               4
7003             Brown,Maxine A  We have all the requirements               3
7004              Logan,Kevin N                           NaN               1
7005              Cashin,Marc C  We have all the requirements               3
7006           Massiah,Gerard F  We have all the requirements               1

[7007 rows x 22 columns]
```

Removing attributes where every single value is distinct.

[36]:
```
elim_col = [x for x in train_x.columns if train_x[x].nunique() == 7007]
```

[37]:
```
for x in elim_col:
    train_x.drop(x, axis=1, inplace=True)
    test.drop(x, axis=1, inplace=True)
```

[38]:
```
cat_col = [x for x in train_x.columns if train_x[x].dtypes == 'O']
```

[39]:
```
len(cat_col)
```

[39]:
```
17
```

First, we'll focus on addressing the issue of missing values.

[40]:
```
train_x.isna().sum()
```

[40]:
```
Industry                1
Deal_value             51
Weighted_amount       525
Date_of_creation        0
Pitch                   0
Lead_revenue            0
```

```
Fund_category              0
Geography                972
Location                  11
POC_name                   8
Designation                0
Hiring_candidate_role      0
Lead_source                0
Level_of_meeting           0
Last_lead_update         633
Internal_POC               0
Resource                 149
Internal_rating            0
dtype: int64
```

[41]:
```
train_x.Industry.unique()
```

[41]:
```
array(['Restaurants', 'Construction Services', 'Hospitals/Clinics',
       'Real Estate', 'Financial Services', 'Banks',
       'Architecture/Engineering', 'Education/Training', 'REIT',
       'Healthcare Facilities/Services', 'Hotels/Motels',
       'Biotech/Healthcare', 'Services', 'Other Investment Firms',
       'Software', 'Other Biz Services', 'Materials/Manufacturing',
       'Other', 'Trucking', 'Automotive/Transportation',
       'Constr - Supplies', 'Casinos/Gaming', 'Food Processing',
       'BioTech/Drugs', 'Investment Bank/Brokerage',
       'Technology Consulting', 'Grocery', 'Online Banking', 'Staffing',
       'Associations', 'Consumer Products', 'Professional Services',
       'Insurance', 'Beverages (Non-Alcoholic)', 'Non-Profit',
       'Internet Software', 'Software Consulting', 'Energy',
       'Furniture/Fixtures', 'Recreational Products',
       'Chemical Manufacturing', 'Personal Services', 'Training',
       'Health/Accident', 'Marketing/Advertising', 'Telecom Hardware',
       'Gold/Silver', 'Semiconductors', 'Other Consulting',
       'Oil, Gas, Coal', 'Aerospace/Defense', 'Social Services',
       'Information Services', 'Telecom Consulting', 'Machine Tools',
       'Iron/Steel', 'Industrial Equip', 'Wireless Consulting',
       'Service Providers', 'Lending/Mortgage', 'Fabricated Products',
       'Advertising/PR', 'Wood/Paper', 'Networking/Comm.',
       'Electric Utilities', 'eCommerce', 'Office Supplies',
       'Photography', 'eBusiness', 'Conglomerates', 'Electronics',
       'Apparel Retail', 'Metals/Mining', 'Finance', 'Life', 'Specialty',
       'Alternative', 'Apparel Products', 'Water Utilities',
       'Recreational Services', 'MSPs (Mgmt)', 'Medical Equipment',
       'Catalog/MailOrder', 'Retail', 'Department/Discount',
       'Entertainment/Media', 'B2B eCommerce', 'Plumbing/HVAC',
       'Property/Casualty', 'Venture Capital', 'WSPs (Wireless)',
       'Sales/Marketing Services', 'Auto Manufacturers', 'Plastic/Rubber',
```

```
       'Auto Parts/Services', 'Legal', 'Religious Groups', 'Web',
       'Peripherals/Comp.', 'Finance Software', 'BSPs (Broadband)',
       'Outsourcing', 'Shipping', 'Personal/Household Products',
       'Home Improvement', 'Waste/Recycling', 'Government', 'Periodicals',
       'Scientific', 'Computer Hardware', 'Constr/Agric Machinery',
       'SaaS', 'Staffing/Outsourcing/HR', 'Television/Cable',
       'Warehousing/Logistics', 'Network Infrastructure', 'Schools',
       'Rentals/Leasing', 'Wood/Timber', 'Movies', 'Testing',
       'Human Resources', 'Packaging/Containers', 'Publishing',
       'Printing Services', 'Vitamins/Nutritionals/Other', 'Utilities',
       'Wireless Hardware', 'Supply Chain', 'Beverages (Alcoholic)',
       'Security Software', 'OSPs (Optical)', 'Technology Retail',
       'Hardware Consulting', 'Concrete/Cement', 'Music',
       'Internet Consulting', 'Telecom Services/Telephone Companies',
       'Airlines/Air Couriers', 'Satellite/RF/Micro',
       'Construction/Agriculture', 'Aircraft Parts',
       'Leisure/Hospitality', 'Servers/Storage', 'Trucks/Buses/RVs',
       'Security Hardware', 'Accounting', 'Natural Gas',
       'Multimedia Software', 'Security Services', 'Railroads',
       'Other Vehicles', 'Wireless Software Networking',
       'Sales/Marketing Software', 'Database', 'Office Equipment',
       'Enterprise', 'ISPs', 'Internet', 'eMail/Messaging',
       'Speech Recognition', 'Radio', 'Appliances/Tools',
       'Telecom Software', 'Book', nan, 'Other Biz Products', 'ERP',
       'Newspapers', 'Auto Dealers', 'Multimedia Hardware',
       'Web Development'], dtype=object)
```

[42]:
```
train_x[train_x.Industry.isna()]
```

[42]:
```
     Industry Deal_value Weighted_amount Date_of_creation      Pitch  \
4653      NaN    209418$      1266978.9$       2020-12-15  Product_2

          Lead_revenue Fund_category Geography  Location POC_name  \
4653  50 - 100 Million    Category 4     India  Pilibhit  roshani

               Designation Hiring_candidate_role    Lead_source  \
4653  SVP/General Counsel    Landscape architect  Contact Email

     Level_of_meeting Last_lead_update    Internal_POC Resource  \
4653          Level 1      2 days back  Hameier,Kurt E       No

      Internal_rating
4653                3
```

[43]:
```
# Since the role being hired for is "architect," I will assign an industry␣
↪related to architecture.
```

```
train_x.loc[4653,'Industry'] = 'Architecture/Engineering'
```

[44]:
```
test[test.Industry.isna()]
```

[44]:
```
     Industry Deal_value Weighted_amount Date_of_creation      Pitch  \
373       NaN    140313$      778737.15$       2020-04-27  Product_1

          Lead_revenue Fund_category Geography   Location POC_name  \
373  50 - 100 Million    Category 2     India  Shillong    kavita

                Designation        Hiring_candidate_role      Lead_source  \
373  CEO/Chairman/President  Lecturer, further education  Marketing Event

    Level_of_meeting Last_lead_update   Internal_POC  \
373          Level 1      2 days back  Pappas,Mark S

                         Resource  Internal_rating
373  We have all the requirements             2.0
```

[45]:
```
# Given that the hiring position is for a lecturer, I will allocate an industry␣
↪associated with education.

test.loc[373,'Industry'] = 'Education/Training'
```

Our task involves removing the dollar sign from both Deal_Value and Weighted_Amount, and subsequently converting them into integers.

[46]:
```
train_x.Deal_value = train_x.loc[:,'Deal_value'].str.replace('$','')
test.Deal_value = test.loc[:,'Deal_value'].str.replace('$','')

train_x.Weighted_amount = train_x.loc[:,'Weighted_amount'].str.replace('$','')
test.Weighted_amount = test.loc[:,'Weighted_amount'].str.replace('$','')
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\3225085736.py:1: FutureWarning:
The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as
literal strings when regex=True.
  train_x.Deal_value = train_x.loc[:,'Deal_value'].str.replace('$','')
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\3225085736.py:2: FutureWarning:
The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as
literal strings when regex=True.
  test.Deal_value = test.loc[:,'Deal_value'].str.replace('$','')
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\3225085736.py:4: FutureWarning:
The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as
literal strings when regex=True.
  train_x.Weighted_amount = train_x.loc[:,'Weighted_amount'].str.replace('$','')
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\3225085736.py:5: FutureWarning:
The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as
literal strings when regex=True.
  test.Weighted_amount = test.loc[:,'Weighted_amount'].str.replace('$','')
```

[47]:
```
int_col = ['Deal_value', 'Weighted_amount']
for x in int_col:
    train_x[x].fillna('0', inplace=True)
    test[x].fillna('0', inplace=True)
```

[48]:
```
train_x.head()
```

[48]:

	Industry	Deal_value	Weighted_amount	Date_of_creation	Pitch	Lead_revenue	Fund_category	Geography	Location	POC_name	Designation	Hiring_candidate_role	Lead_source	Level_of_meeting	Last_lead_update	Internal_POC
0	Restaurants	320506	2067263.7	2020-03-29	Product_2	50 - 100 Million	Category 2	USA	Killeen-Temple, TX	Charlene Werner	Executive Vice President	Community pharmacist	Website	Level 3	No track	Davis,Sharrice A
1	Construction Services	39488	240876.8	2019-07-10	Product_2	500 Million - 1 Billion	Category 4	India	Ratlam	rakhi	Chairman/CEO/President	Recruitment consultant	Others	Level 1	Did not hear back after Level 1	Brown,Maxine A
2	Hospitals/Clinics	359392	2407926.4	2019-07-27	Product_1	500 Million - 1 Billion	Category 4	USA	Albany-Schenectady-Troy, NY	Ariel Hamilton	SVP/General Counsel	Health service manager	Marketing Event	Level 1	?	Georgakopoulos,Vasilios T
3	Real Estate	76774	468321.4	2021-01-30	Product_2	500 Million - 1 Billion	Category 3	USA	Mount Vernon-Anacortes, WA	Erin Wilson	CEO/Co-Founder/Chairman	Therapist, speech and language	Contact Email	Level 2	Did not hear back after Level 1	Brown,Maxine A
4	Financial Services	483896	0	2019-05-22	Product_2	50 - 100 Million	Category 3	India	Shimoga	kavita	Executive Vice President	Media planner	Website	Level 2	Up-to-date	Thomas,Lori E

```
                      Resource  Internal_rating
0                          NaN                3
1                           No                5
2                           No                4
3  We have all the requirements                1
4                           No                4
```

[49]:
```
test.head()
```

[49]:
```
                   Industry Deal_value Weighted_amount Date_of_creation  \
0  Investment Bank/Brokerage     200988               0       2020-04-15
1                Electronics     409961       2541758.2       2021-01-23
2                      Banks     434433       3041031.0       2020-07-19
3                      Music     218952       1521716.4       2020-02-27
4                Real Estate     392835      2455218.75       2020-10-25

       Pitch             Lead_revenue Fund_category Geography  \
0  Product_1        100 - 500 Million    Category 4     India
1  Product_1        100 - 500 Million    Category 3       USA
2  Product_1        100 - 500 Million    Category 1       USA
3  Product_2        100 - 500 Million    Category 1     India
4  Product_1  500 Million - 1 Billion    Category 3       USA

                     Location           POC_name  \
0                Bhubaneshwar              sonia
1           Coeur d'Alene, ID        Daniel Bell
2  Portland-South Portland, ME       Andrew Davis
3                    Bareilly             shital
4                  Trenton, NJ  Shelly Stephenson

                        Designation            Hiring_candidate_role  \
0             Chairman/CEO/President       Designer, fashion/clothing
1           CEO/Co-Founder/Chairman         Horticultural consultant
2  Chairman/Chief Innovation Officer              Information officer
3             CEO/Chairman/President  Commercial/residential surveyor
4           CEO/Co-Founder/Chairman               Wellsite geologist

       Lead_source Level_of_meeting                 Last_lead_update  \
0  Marketing Event          Level 1                more than a month
1  Marketing Event          Level 2                       Up-to-date
2  Marketing Event          Level 2  Did not hear back after Level 1
3    Contact Email          Level 3                more than a month
4           Others          Level 3                More than 2 weeks

       Internal_POC     Resource  Internal_rating
0  Massiah,Gerard F           No              1.0
```

```
1    Smith,Keenan H          Yes              1.0
2     Gilley,Janine  Deliverable              5.0
3      Morsy,Omar A           No              5.0
4      Morsy,Omar A  Deliverable              2.0
```

[50]:
```
for x in int_col:
    train_x[x] = train_x[x].astype('float64')
    test[x] = test[x].astype('float64')
```

Dataset containing non-null values for deal value and weighted amount.

[51]:
```
train_x['W_div_D'] = train_x.Weighted_amount / train_x.Deal_value
test['W_div_D'] = test.Weighted_amount / train_x.Deal_value
```

The average of the x factor between deal value and weighted amount.

[52]:
```
train_x[train_x.W_div_D != np.inf].describe()
```

[52]:

	Deal_value	Weighted_amount	Internal_rating	W_div_D
count	6960.000000	6.960000e+03	6960.000000	6956.000000
mean	249512.544971	1.451080e+06	3.010345	5.805865
std	144840.900466	9.786007e+05	1.418842	1.700877
min	0.000000	0.000000e+00	1.000000	0.000000
25%	122106.500000	5.835328e+05	2.000000	5.900000
50%	246902.000000	1.420423e+06	3.000000	6.250000
75%	376802.000000	2.293298e+06	4.000000	6.550000
max	500000.000000	3.601416e+06	5.000000	7.450000

[53]:
```
test[test.W_div_D != np.inf].describe()
```

[53]:

	Deal_value	Weighted_amount	Internal_rating	W_div_D
count	2076.000000	2.076000e+03	2076.000000	2076.000000
mean	247720.996628	1.513542e+06	2.957611	16.694741
std	142317.028166	9.244484e+05	1.426474	61.922842
min	0.000000	0.000000e+00	1.000000	0.000000
25%	125825.500000	7.103544e+05	2.000000	2.923683
50%	250921.000000	1.517774e+06	3.000000	6.186514
75%	372179.500000	2.312600e+06	4.000000	12.421288
max	499392.000000	3.409163e+06	5.000000	1672.323698

[54]:
```
train_fac = train_x[train_x.W_div_D != np.inf].W_div_D.mean()
test_fac = test[test.W_div_D != np.inf].W_div_D.mean()
```

[55]:
```
test_fac
```

[55]:
```
16.694740557899745
```

[56]:
```
test.shape
```

[56]: (2093, 19)

[57]:
```
both_null = train_x[(train_x.Weighted_amount == 0) & (train_x.Deal_value == 0)]
```

[58]:
```
both_null_id = both_null.index
```

[59]:
```
for x in both_null_id:
        train_x.loc[x, 'Deal_value'] = 249512.5
for x in both_null_id:
        train_x.loc[x, 'Weighted_amount'] = 1451080.3
```

[60]:
```
null_tr_wa = train_x[(train_x.Weighted_amount == 0) & (train_x.Deal_value != 0)]
```

[61]:
```
null_tr_wa_id = null_tr_wa.index
```

[62]:
```
for x in null_tr_wa_id:
    train_x.loc[x, 'Weighted_amount'] = train_x.loc[x, 'Deal_value'] * train_fac
```

[63]:
```
train_x.loc[null_tr_wa_id, :]
```

[63]:

	Industry	Deal_value	Weighted_amount	Date_of_creation	\
4	Financial Services	483896.0	2.809435e+06	2019-05-22	
13	Banks	487351.0	2.829494e+06	2020-04-25	
17	Software	140949.0	8.183309e+05	2019-05-19	
23	REIT	207335.0	1.203759e+06	2020-10-27	
33	Automotive/Transportation	335694.0	1.948994e+06	2019-10-04	
...	...	...	...	...	
6971	Hospitals/Clinics	101621.0	5.899979e+05	2019-05-14	
6975	Financial Services	456307.0	2.649257e+06	2020-04-03	
6983	Professional Services	89949.0	5.222318e+05	2019-07-07	
6990	Insurance	240326.0	1.395300e+06	2020-11-23	
7004	Semiconductors	253608.0	1.472414e+06	2020-03-10	

	Pitch	Lead_revenue	Fund_category	Geography	\
4	Product_2	50 - 100 Million	Category 3	India	
13	Product_1	500 Million - 1 Billion	Category 3	NaN	
17	Product_1	100 - 500 Million	Category 3	USA	
23	Product_1	100 - 500 Million	Category 3	India	
33	Product_2	50 - 100 Million	Category 3	USA	
...	...	...	...	...	
6971	Product_1	50 - 100 Million	Category 1	USA	
6975	Product_2	500 Million - 1 Billion	Category 2	India	
6983	Product_1	100 - 500 Million	Category 2	NaN	
6990	Product_2	100 - 500 Million	Category 4	India	
7004	Product_1	100 - 500 Million	Category 2	USA	

	Location	POC_name	\

```
4                Shimoga              kavita
13              Pilibhit             saaniya
17          Columbus, IN  Christopher Walter
23            Ahmadnagar                rani
33             Chico, CA             Alex Le
...                  ...                 ...
6971         Madison, WI       Jeffrey Smith
6975             Gangtok               durga
6983  Grand Forks, ND-MN       Mary Cummings
6990       Muzaffarnagar              girija
7004     Rocky Mount, NC     Nicholas Duncan

                           Designation          Hiring_candidate_role  \
4              Executive Vice President                  Media planner
13    Chairman/Chief Innovation Officer               Hospital doctor
17             Executive Vice President                Phytotherapist
23               CEO/Chairman/President                 Warden/ranger
33                         CEO/President  Operational investment banker
...                                  ...                           ...
6971                               CEO            Clinical biochemist
6975             CEO/Chairman/President          Travel agency manager
6983               SVP/General Counsel          Volunteer coordinator
6990             Executive Vice President     Lecturer, higher education
7004                 SVP/General Counsel     Nature conservation officer

         Lead_source Level_of_meeting                     Last_lead_update  \
4             Website          Level 2                          Up-to-date
13    Marketing Event          Level 3                More than a week back
17    Marketing Event          Level 2                More than a week back
23            Website          Level 2                More than a week back
33      Contact Email          Level 1       Did not hear back after Level 1
...               ...              ...                                  ...
6971  Marketing Event          Level 3                                  NaN
6975  Marketing Event          Level 2                            5 days back
6983          Website          Level 1                                  NaN
6990    Contact Email          Level 1  Following up but lead not responding
7004  Marketing Event          Level 3                                  NaN

               Internal_POC                      Resource  Internal_rating  \
4             Thomas,Lori E                            No                4
13         Shelton,Sidney P  We have all the requirements                2
17           Mabrey,Kevin C                           Yes                1
23         Irizarry,Yolanda                           Yes                4
33             Morsy,Omar A                           Yes                3
...                     ...                           ...              ...
6971       Irizarry,Yolanda  We have all the requirements                4
6975            Ali,Mohamed                   Deliverable                4
```

```
6983          Charles,Caleb                    Deliverable                3
6990  Featherstone,Adrian R                             No                1
7004          Logan,Kevin N                            NaN                1

      W_div_D
4         0.0
13        0.0
17        0.0
23        0.0
33        0.0
...       ...
6971      0.0
6975      0.0
6983      0.0
6990      0.0
7004      0.0

[521 rows x 19 columns]
```

[64]:
```
test.Weighted_amount.replace(0, np.nan, inplace=True)
```

[65]:
```
test_wa_mean = test.Weighted_amount.mean()
```

[66]:
```
test.Weighted_amount.fillna(test_wa_mean, inplace=True)
```

The weighted amount is now resolved.

I've observed a connection between deal_value and weighted_amount. The average weighted_amount is approximately 6.01 to 6.9 times the deal value. I intend to populate the missing deal_value entries by dividing the weighted_amount by either train_fac or test_fac, depending on the context.

Managing deal value.

[67]:
```
train_x[train_x.Deal_value == 0].shape
```

[67]:
```
(47, 19)
```

[68]:
```
test[test.Deal_value == 0].shape
```

[68]:
```
(5, 19)
```

[69]:
```
train_x.Deal_value.replace(0, np.nan, inplace=True)
test.Deal_value.replace(0, np.nan, inplace=True)
```

[70]:
```
train_x[train_x.Deal_value.isna()]
```

[70]:
```
                          Industry  Deal_value  Weighted_amount  \
169                          Banks         NaN       1023175.80
```

381	Iron/Steel	NaN	668538.00
469	Software	NaN	1167666.85
519	Banks	NaN	2839513.60
1016	Network Infrastructure	NaN	2929018.90
1148	Furniture/Fixtures	NaN	2432947.95
1246	Staffing	NaN	1195450.00
1328	Insurance	NaN	2607096.45
1587	Music	NaN	408901.35
1599	Education/Training	NaN	224107.90
1626	Apparel Retail	NaN	1059562.40
1636	Banks	NaN	2748051.20
1651	Real Estate	NaN	66601.60
1849	Other	NaN	1455981.20
1943	Electronics	NaN	1666770.60
2039	Banks	NaN	3295175.40
2165	Banks	NaN	1936548.25
2241	Banks	NaN	1190311.85
2469	Financial Services	NaN	2139806.85
2487	Recreational Services	NaN	1469188.05
2869	Movies	NaN	1977882.70
2953	Medical Equipment	NaN	1482643.25
3139	Insurance	NaN	2055754.40
3272	Staffing	NaN	2263868.60
3327	Hospitals/Clinics	NaN	2498082.00
3586	Restaurants	NaN	2694762.50
3632	Oil, Gas, Coal	NaN	1851598.35
3919	Other	NaN	2485949.55
4269	Banks	NaN	3175824.40
4300	Software	NaN	1865694.60
4745	Food Processing	NaN	232377.00
4926	Non-Profit	NaN	1744578.00
5276	Other Biz Services	NaN	1339298.30
5401	Lending/Mortgage	NaN	1711770.80
5620	Banks	NaN	96495.75
5889	Financial Services	NaN	799571.20
6001	Retail	NaN	1701722.25
6100	Banks	NaN	150485.40
6115	Trucking	NaN	2210982.40
6213	Investment Bank/Brokerage	NaN	448921.35
6423	Sales/Marketing Services	NaN	823636.65
6501	Oil, Gas, Coal	NaN	2227864.80
6567	Insurance	NaN	2259299.65
6611	Security Services	NaN	532115.65
6653	Healthcare Facilities/Services	NaN	1106557.20
6878	Real Estate	NaN	2063642.90
6885	Banks	NaN	2141010.45

	Date_of_creation	Pitch	Lead_revenue	Fund_category	\
169	2020-06-12	Product_2	50 - 100 Million	Category 2	
381	2019-08-24	Product_2	100 - 500 Million	Category 1	
469	2020-05-19	Product_1	500 Million - 1 Billion	Category 1	
519	2020-07-25	Product_1	100 - 500 Million	Category 4	
1016	2019-05-26	Product_1	500 Million - 1 Billion	Category 1	
1148	2020-10-27	Product_2	500 Million - 1 Billion	Category 4	
1246	2019-08-19	Product_1	50 - 100 Million	Category 4	
1328	2020-01-31	Product_2	50 - 100 Million	Category 4	
1587	2020-03-22	Product_1	50 - 100 Million	Category 2	
1599	2019-09-27	Product_2	500 Million - 1 Billion	Category 4	
1626	2019-02-21	Product_2	100 - 500 Million	Category 4	
1636	2020-08-19	Product_1	500 Million - 1 Billion	Category 2	
1651	2020-08-09	Product_1	500 Million - 1 Billion	Category 2	
1849	2021-02-06	Product_1	500 Million - 1 Billion	Category 1	
1943	2019-08-23	Product_1	100 - 500 Million	Category 1	
2039	2020-09-09	Product_1	50 - 100 Million	Category 1	
2165	2020-12-19	Product_2	500 Million - 1 Billion	Category 2	
2241	2019-03-16	Product_1	100 - 500 Million	Category 1	
2469	2019-01-28	Product_1	50 - 100 Million	Category 4	
2487	2019-09-26	Product_1	50 - 100 Million	Category 3	
2869	2020-10-20	Product_2	500 Million - 1 Billion	Category 4	
2953	2019-06-22	Product_2	100 - 500 Million	Category 2	
3139	2020-09-27	Product_1	500 Million - 1 Billion	Category 1	
3272	2020-12-29	Product_1	100 - 500 Million	Category 2	
3327	2019-04-07	Product_1	50 - 100 Million	Category 1	
3586	2020-07-07	Product_2	100 - 500 Million	Category 2	
3632	2019-09-13	Product_2	500 Million - 1 Billion	Category 2	
3919	2020-02-03	Product_2	500 Million - 1 Billion	Category 3	
4269	2020-08-12	Product_1	500 Million - 1 Billion	Category 4	
4300	2020-08-07	Product_1	500 Million - 1 Billion	Category 2	
4745	2020-10-08	Product_2	100 - 500 Million	Category 4	
4926	2019-06-07	Product_2	100 - 500 Million	Category 4	
5276	2020-09-10	Product_2	500 Million - 1 Billion	Category 4	
5401	2019-08-20	Product_2	50 - 100 Million	Category 1	
5620	2019-04-07	Product_2	500 Million - 1 Billion	Category 3	
5889	2019-01-13	Product_1	100 - 500 Million	Category 1	
6001	2019-10-15	Product_2	100 - 500 Million	Category 1	
6100	2019-07-06	Product_1	100 - 500 Million	Category 2	
6115	2019-06-17	Product_2	50 - 100 Million	Category 3	
6213	2019-02-23	Product_2	500 Million - 1 Billion	Category 4	
6423	2019-03-05	Product_2	50 - 100 Million	Category 2	
6501	2019-01-06	Product_2	500 Million - 1 Billion	Category 2	
6567	2021-01-05	Product_2	50 - 100 Million	Category 2	
6611	2020-01-02	Product_2	500 Million - 1 Billion	Category 3	
6653	2019-09-06	Product_1	50 - 100 Million	Category 2	
6878	2019-09-07	Product_2	500 Million - 1 Billion	Category 3	

```
6885       2020-06-05  Product_2        100 - 500 Million    Category 4

     Geography                                       Location  \
169        USA                                 Panama City, FL
381      India                                           Akola
469      India                                       Bhagalpur
519      India                                          Haora
1016     India                                       Amritsar
1148       USA                                  Bellingham, WA
1246     India                                         Rampura
1328     India                                       Bharauri
1587       USA                                   Ann Arbor, MI
1599     India                                       Jabalpur
1626       USA                  Santa Maria-Santa Barbara, CA
1636       USA                                  Bellingham, WA
1651     India                                     Krishnanagar
1849       USA                                Salisbury, MD-DE
1943     India                                      Saharanpur
2039       USA                    Florence-Muscle Shoals, AL
2165     India                                   Rajapalaiyam
2241       USA                                Jacksonville, NC
2469       USA                   Weirton-Steubenville, WV-OH
2487       USA                                    Midland, TX
2869       USA                                    Jackson, TN
2953     India                                            Puri
3139     India                                         Mathura
3272     India                                       Mangalore
3327       NaN                                         Silchar
3586       NaN                                          Rohtak
3632     India                                      Aurangabad
3919       USA                                    Amarillo, TX
4269       USA                                  Greenville, NC
4300       USA                                   Corvallis, OR
4745     India                                           Alwar
4926       USA                                   Muskegon, MI
5276       NaN                                  Rochester, MN
5401       USA                                  Pine Bluff, AR
5620     India                                          Silchar
5889       USA                                 Springfield, MO
6001       USA  San Luis Obispo-Paso Robles-Arroyo Grande, CA
6100       NaN                                          Ranchi
6115       USA                                   Missoula, MT
6213       USA                                  Burlington, NC
6423       USA                                  Tuscaloosa, AL
6501       NaN                                    Savannah, GA
6567     India                                         Asansol
6611     India                                       Bengaluru
```

6653	NaN	Bloomington, IL
6878	India	Imphal
6885	USA	Elkhart-Goshen, IN

	POC_name	Designation	\
169	Jason Smith	Chairman/Chief Innovation Officer	
381	kamakshya @ mona	Executive Vice President	
469	gajala	Chairman/CEO/President	
519	kritika	Executive Vice President	
1016	miss. santra	SVP/General Counsel	
1148	Scott Johnson	SVP/General Counsel	
1246	sonia	Executive Vice President	
1328	tanvir fatma	CEO/President	
1587	Carol Clark	CEO/President	
1599	sushree sagar	CEO/President	
1626	Jacqueline Hull	Chairman/CEO/President	
1636	Brandi Miller	Executive Vice President	
1651	mamta	Vice President / GM (04-present) : VP Sales an...	
1849	Mark Garza	SVP/General Counsel	
1943	smt sunita	CEO/Chairman/President	
2039	Katherine Walker	SVP/General Counsel	
2165	soni kumari	Chairman/CEO/President	
2241	Wendy Valdez	Chairman/CEO/President	
2469	Robyn Braun PhD	Chairman/Chief Innovation Officer	
2487	Wendy Ball	SVP/General Counsel	
2869	Carrie Hawkins	CEO/President	
2953	anita	SVP/General Counsel	
3139	ranju sharma	Executive Vice President	
3272	neelam	SVP/General Counsel	
3327	karishama	CEO/Co-Founder/Chairman	
3586	unknown	Chairman/CEO/President	
3632	rizwana	Chairman/Chief Innovation Officer	
3919	Anthony Fields	Executive Vice President	
4269	James Cameron	Chairman/CEO/President	
4300	Allison Caldwell	CEO/President	
4745	shimran d/o	CEO/Co-Founder/Chairman	
4926	Joshua Duran	CEO/Chairman/President	
5276	Jeffrey Henderson	CEO/Chairman/President	
5401	Taylor Gomez	Executive Vice President	
5620	rukhsana	Chairman/Chief Innovation Officer	
5889	Corey Lewis	Executive Vice President	
6001	Tiffany Wang	CEO/Co-Founder/Chairman	
6100	simren	SVP/General Counsel	
6115	Stephen Long	Chief Executive Officer	
6213	Kimberly Hull	Chairman/CEO/President	
6423	Danny Hunter	Chairman/CEO/President	
6501	Desiree Kim	Chairman/Chief Innovation Officer	

6567	suman	CEO/President
6611	rani	Executive Vice President
6653	David Larson	Chairman/CEO/President
6878	seema	CEO/President
6885	Amanda Smith	CEO/Co-Founder/Chairman

	Hiring_candidate_role	Lead_source
169	Information officer	Others
381	Insurance account manager	Others
469	Minerals surveyor	Contact Email
519	Restaurant manager	Website
1016	Civil engineer, consulting	Marketing Event
1148	Community pharmacist	Contact Email
1246	Primary school teacher	Marketing Event
1328	Bonds trader	Others
1587	Designer, textile	Marketing Event
1599	Sports administrator	Marketing Event
1626	Health promotion specialist	Marketing Event
1636	Associate Professor	Marketing Event
1651	Printmaker	Others
1849	Chief Financial Officer	Others
1943	Nurse, adult	Marketing Event
2039	Interpreter	Contact Email
2165	Insurance claims handler	Website
2241	Restaurant manager	Others
2469	Embryologist, clinical	Others
2487	Clinical cytogeneticist	Contact Email
2869	Engineer, maintenance (IT)	Contact Email
2953	Fast food restaurant manager	Others
3139	Surveyor, planning and development	Contact Email
3272	Professor Emeritus	Marketing Event
3327	Surveyor, commercial/residential	Marketing Event
3586	Podiatrist	Website
3632	Designer, blown glass/stained glass	Others
3919	Interior and spatial designer	Others
4269	Animal nutritionist	Contact Email
4300	Charity officer	Marketing Event
4745	Recruitment consultant	Others
4926	Magazine features editor	Contact Email
5276	Hotel manager	Others
5401	Librarian, academic	Contact Email
5620	Teacher, secondary school	Others
5889	Doctor, hospital	Marketing Event
6001	Health service manager	Contact Email
6100	Proofreader	Marketing Event
6115	English as a foreign language teacher	Others
6213	Higher education careers adviser	Marketing Event

```
6423                                Media planner           Others
6501  Programme researcher, broadcasting/film/video           Others
6567                              Physiotherapist           Others
6611                         Biochemist, clinical          Website
6653                Medical sales representative    Contact Email
6878                           Technical author          Website
6885                         Biomedical engineer          Website

     Level_of_meeting                      Last_lead_update  \
169           Level 1                            Up-to-date
381           Level 2                              No track
469           Level 3                              No track
519           Level 3                           5 days back
1016          Level 2                                     ?
1148          Level 3  Following up but lead not responding
1246          Level 2                            Up-to-date
1328          Level 1                              No track
1587          Level 2                  More than a week back
1599          Level 1  Following up but lead not responding
1626          Level 1                               Pending
1636          Level 1       Did not hear back after Level 1
1651          Level 1       Did not hear back after Level 1
1849          Level 3                     more than a month
1943          Level 2                     more than a month
2039          Level 3                              No track
2165          Level 1                                     ?
2241          Level 3                                   NaN
2469          Level 2                                     ?
2487          Level 1                     more than a month
2869          Level 3                            Up-to-date
2953          Level 3                  More than a week back
3139          Level 3                                     ?
3272          Level 2                              No track
3327          Level 3                                   NaN
3586          Level 2                            Up-to-date
3632          Level 3                           5 days back
3919          Level 3                                   NaN
4269          Level 3                     more than a month
4300          Level 2                                 Pending
4745          Level 1                           5 days back
4926          Level 2       Did not hear back after Level 1
5276          Level 3                     More than 2 weeks
5401          Level 3                            5 days back
5620          Level 3                           2 days back
5889          Level 2       Did not hear back after Level 1
6001          Level 3  Following up but lead not responding
6100          Level 2                               Pending
```

6115	Level 3	2 days back	
6213	Level 3	No track	
6423	Level 1	Did not hear back after Level 1	
6501	Level 1	More than a week back	
6567	Level 1	Up-to-date	
6611	Level 1	?	
6653	Level 3	NaN	
6878	Level 2	More than 2 weeks	
6885	Level 1	Following up but lead not responding	

	Internal_POC	Resource	\
169	Cashin,Marc C	Cannot deliver	
381	Dunaway,Antoine	Yes	
469	Moran,Natalie A	We have all the requirements	
519	Houston,Arnold E	Deliverable	
1016	Bannister,Joan	Deliverable	
1148	Hanyok,John J	We have all the requirements	
1246	Brown,Maxine A	Deliverable	
1328	Van Arter,Derrick	Not enough	
1587	Vickers Jr.,Henry J	Not enough	
1599	Brown,Maxine A	No	
1626	Maine,John P	Not enough	
1636	Hebron,Artenia D	We have all the requirements	
1651	Himes,Maurice C	Yes	
1849	Davis,Brian R	Yes	
1943	Cash,Tyrone J	Yes	
2039	Pappas,Mark S	We have all the requirements	
2165	Tondeur,Keith D	Not enough	
2241	Massiah,Gerard F	Not enough	
2469	Jones,Eyvette W	Cannot deliver	
2487	Vickers Jr.,Henry J	Cannot deliver	
2869	Pappas,Mark S	Yes	
2953	McKenstry,Loretta A	No	
3139	Georgakopoulos,Vasilios T	Deliverable	
3272	Ullrich,Rose Anne	Cannot deliver	
3327	Abdul-Hamid,Saud Muhamad	Cannot deliver	
3586	Dyson,William A	Cannot deliver	
3632	Himes,Maurice C	Deliverable	
3919	Gaskins Jr,Franklin D	We have all the requirements	
4269	Hameier,Kurt E	We have all the requirements	
4300	Ryker,David	NaN	
4745	Hebron,Artenia D	Deliverable	
4926	Gilley,Janine	Cannot deliver	
5276	Hameier,Kurt E	Yes	
5401	Green,Candy	Cannot deliver	
5620	Green,Ann E	We have all the requirements	
5889	Murray,Younetta	We have all the requirements	

6001	Booker,David	L	No
6100	Himes,Maurice	C	Deliverable
6115	Himes,Maurice	C	Cannot deliver
6213	Houston,Arnold	E	Deliverable
6423	Mabrey,Kevin	C	No
6501	Green,Candy		Not enough
6567	Houston,Arnold	E	We have all the requirements
6611	Gould,Lisa	D	We have all the requirements
6653	Meli,Teresa	V	We have all the requirements
6878	Himes,Maurice	C	Cannot deliver
6885	Moran,Natalie	A	We have all the requirements

	Internal_rating	W_div_D
169	4	inf
381	1	inf
469	2	inf
519	1	inf
1016	4	inf
1148	4	inf
1246	2	inf
1328	2	inf
1587	4	inf
1599	1	inf
1626	5	inf
1636	5	inf
1651	2	inf
1849	3	inf
1943	4	inf
2039	1	inf
2165	2	inf
2241	1	inf
2469	5	inf
2487	5	inf
2869	1	inf
2953	4	inf
3139	2	inf
3272	3	inf
3327	3	inf
3586	5	inf
3632	1	inf
3919	5	inf
4269	1	inf
4300	5	inf
4745	3	inf
4926	3	inf
5276	4	inf
5401	3	inf

```
5620                2      inf
5889                3      inf
6001                1      inf
6100                4      inf
6115                5      inf
6213                1      inf
6423                3      inf
6501                2      inf
6567                4      inf
6611                3      inf
6653                3      inf
6878                2      inf
6885                2      inf
```

[71]:
```
null_trdv_id = train_x[train_x.Deal_value.isna()].index
```

[72]:
```
train_x.loc[null_trdv_id, :]
```

[72]:
```
                          Industry  Deal_value  Weighted_amount  \
169                          Banks         NaN       1023175.80
381                     Iron/Steel         NaN        668538.00
469                       Software         NaN       1167666.85
519                          Banks         NaN       2839513.60
1016  Network Infrastructure         NaN       2929018.90
1148      Furniture/Fixtures         NaN       2432947.95
1246                    Staffing         NaN       1195450.00
1328                   Insurance         NaN       2607096.45
1587                       Music         NaN        408901.35
1599      Education/Training         NaN        224107.90
1626          Apparel Retail         NaN       1059562.40
1636                         Banks         NaN       2748051.20
1651                  Real Estate         NaN         66601.60
1849                        Other         NaN       1455981.20
1943                  Electronics         NaN       1666770.60
2039                         Banks         NaN       3295175.40
2165                         Banks         NaN       1936548.25
2241                         Banks         NaN       1190311.85
2469      Financial Services         NaN       2139806.85
2487   Recreational Services         NaN       1469188.05
2869                      Movies         NaN       1977882.70
2953       Medical Equipment         NaN       1482643.25
3139                   Insurance         NaN       2055754.40
3272                    Staffing         NaN       2263868.60
3327       Hospitals/Clinics         NaN       2498082.00
3586                 Restaurants         NaN       2694762.50
3632          Oil, Gas, Coal         NaN       1851598.35
3919                        Other         NaN       2485949.55
```

4269	Banks	NaN	3175824.40
4300	Software	NaN	1865694.60
4745	Food Processing	NaN	232377.00
4926	Non-Profit	NaN	1744578.00
5276	Other Biz Services	NaN	1339298.30
5401	Lending/Mortgage	NaN	1711770.80
5620	Banks	NaN	96495.75
5889	Financial Services	NaN	799571.20
6001	Retail	NaN	1701722.25
6100	Banks	NaN	150485.40
6115	Trucking	NaN	2210982.40
6213	Investment Bank/Brokerage	NaN	448921.35
6423	Sales/Marketing Services	NaN	823636.65
6501	Oil, Gas, Coal	NaN	2227864.80
6567	Insurance	NaN	2259299.65
6611	Security Services	NaN	532115.65
6653	Healthcare Facilities/Services	NaN	1106557.20
6878	Real Estate	NaN	2063642.90
6885	Banks	NaN	2141010.45

	Date_of_creation	Pitch	Lead_revenue	Fund_category	\
169	2020-06-12	Product_2	50 - 100 Million	Category 2	
381	2019-08-24	Product_2	100 - 500 Million	Category 1	
469	2020-05-19	Product_1	500 Million - 1 Billion	Category 1	
519	2020-07-25	Product_1	100 - 500 Million	Category 4	
1016	2019-05-26	Product_1	500 Million - 1 Billion	Category 1	
1148	2020-10-27	Product_2	500 Million - 1 Billion	Category 4	
1246	2019-08-19	Product_1	50 - 100 Million	Category 4	
1328	2020-01-31	Product_2	50 - 100 Million	Category 4	
1587	2020-03-22	Product_1	50 - 100 Million	Category 2	
1599	2019-09-27	Product_2	500 Million - 1 Billion	Category 4	
1626	2019-02-21	Product_2	100 - 500 Million	Category 4	
1636	2020-08-19	Product_1	500 Million - 1 Billion	Category 2	
1651	2020-08-09	Product_1	500 Million - 1 Billion	Category 2	
1849	2021-02-06	Product_1	500 Million - 1 Billion	Category 1	
1943	2019-08-23	Product_1	100 - 500 Million	Category 1	
2039	2020-09-09	Product_1	50 - 100 Million	Category 1	
2165	2020-12-19	Product_2	500 Million - 1 Billion	Category 2	
2241	2019-03-16	Product_1	100 - 500 Million	Category 1	
2469	2019-01-28	Product_1	50 - 100 Million	Category 4	
2487	2019-09-26	Product_1	50 - 100 Million	Category 3	
2869	2020-10-20	Product_2	500 Million - 1 Billion	Category 4	
2953	2019-06-22	Product_2	100 - 500 Million	Category 2	
3139	2020-09-27	Product_1	500 Million - 1 Billion	Category 1	
3272	2020-12-29	Product_1	100 - 500 Million	Category 2	
3327	2019-04-07	Product_1	50 - 100 Million	Category 1	
3586	2020-07-07	Product_2	100 - 500 Million	Category 2	

3632	2019-09-13	Product_2	500 Million - 1 Billion	Category 2
3919	2020-02-03	Product_2	500 Million - 1 Billion	Category 3
4269	2020-08-12	Product_1	500 Million - 1 Billion	Category 4
4300	2020-08-07	Product_1	500 Million - 1 Billion	Category 2
4745	2020-10-08	Product_2	100 - 500 Million	Category 4
4926	2019-06-07	Product_2	100 - 500 Million	Category 4
5276	2020-09-10	Product_2	500 Million - 1 Billion	Category 4
5401	2019-08-20	Product_2	50 - 100 Million	Category 1
5620	2019-04-07	Product_2	500 Million - 1 Billion	Category 3
5889	2019-01-13	Product_1	100 - 500 Million	Category 1
6001	2019-10-15	Product_2	100 - 500 Million	Category 1
6100	2019-07-06	Product_1	100 - 500 Million	Category 2
6115	2019-06-17	Product_2	50 - 100 Million	Category 3
6213	2019-02-23	Product_2	500 Million - 1 Billion	Category 4
6423	2019-03-05	Product_2	50 - 100 Million	Category 2
6501	2019-01-06	Product_2	500 Million - 1 Billion	Category 2
6567	2021-01-05	Product_2	50 - 100 Million	Category 2
6611	2020-01-02	Product_2	500 Million - 1 Billion	Category 3
6653	2019-09-06	Product_1	50 - 100 Million	Category 2
6878	2019-09-07	Product_2	500 Million - 1 Billion	Category 3
6885	2020-06-05	Product_2	100 - 500 Million	Category 4

	Geography	Location
169	USA	Panama City, FL
381	India	Akola
469	India	Bhagalpur
519	India	Haora
1016	India	Amritsar
1148	USA	Bellingham, WA
1246	India	Rampura
1328	India	Bharauri
1587	USA	Ann Arbor, MI
1599	India	Jabalpur
1626	USA	Santa Maria-Santa Barbara, CA
1636	USA	Bellingham, WA
1651	India	Krishnanagar
1849	USA	Salisbury, MD-DE
1943	India	Saharanpur
2039	USA	Florence-Muscle Shoals, AL
2165	India	Rajapalaiyam
2241	USA	Jacksonville, NC
2469	USA	Weirton-Steubenville, WV-OH
2487	USA	Midland, TX
2869	USA	Jackson, TN
2953	India	Puri
3139	India	Mathura
3272	India	Mangalore

3327	NaN	Silchar
3586	NaN	Rohtak
3632	India	Aurangabad
3919	USA	Amarillo, TX
4269	USA	Greenville, NC
4300	USA	Corvallis, OR
4745	India	Alwar
4926	USA	Muskegon, MI
5276	NaN	Rochester, MN
5401	USA	Pine Bluff, AR
5620	India	Silchar
5889	USA	Springfield, MO
6001	USA	San Luis Obispo-Paso Robles-Arroyo Grande, CA
6100	NaN	Ranchi
6115	USA	Missoula, MT
6213	USA	Burlington, NC
6423	USA	Tuscaloosa, AL
6501	NaN	Savannah, GA
6567	India	Asansol
6611	India	Bengaluru
6653	NaN	Bloomington, IL
6878	India	Imphal
6885	USA	Elkhart-Goshen, IN

	POC_name	Designation	\
169	Jason Smith	Chairman/Chief Innovation Officer	
381	kamakshya @ mona	Executive Vice President	
469	gajala	Chairman/CEO/President	
519	kritika	Executive Vice President	
1016	miss. santra	SVP/General Counsel	
1148	Scott Johnson	SVP/General Counsel	
1246	sonia	Executive Vice President	
1328	tanvir fatma	CEO/President	
1587	Carol Clark	CEO/President	
1599	sushree sagar	CEO/President	
1626	Jacqueline Hull	Chairman/CEO/President	
1636	Brandi Miller	Executive Vice President	
1651	mamta	Vice President / GM (04-present) : VP Sales an...	
1849	Mark Garza	SVP/General Counsel	
1943	smt sunita	CEO/Chairman/President	
2039	Katherine Walker	SVP/General Counsel	
2165	soni kumari	Chairman/CEO/President	
2241	Wendy Valdez	Chairman/CEO/President	
2469	Robyn Braun PhD	Chairman/Chief Innovation Officer	
2487	Wendy Ball	SVP/General Counsel	
2869	Carrie Hawkins	CEO/President	
2953	anita	SVP/General Counsel	

3139	ranju sharma	Executive Vice President
3272	neelam	SVP/General Counsel
3327	karishama	CEO/Co-Founder/Chairman
3586	unknown	Chairman/CEO/President
3632	rizwana	Chairman/Chief Innovation Officer
3919	Anthony Fields	Executive Vice President
4269	James Cameron	Chairman/CEO/President
4300	Allison Caldwell	CEO/President
4745	shimran d/o	CEO/Co-Founder/Chairman
4926	Joshua Duran	CEO/Chairman/President
5276	Jeffrey Henderson	CEO/Chairman/President
5401	Taylor Gomez	Executive Vice President
5620	rukhsana	Chairman/Chief Innovation Officer
5889	Corey Lewis	Executive Vice President
6001	Tiffany Wang	CEO/Co-Founder/Chairman
6100	simren	SVP/General Counsel
6115	Stephen Long	Chief Executive Officer
6213	Kimberly Hull	Chairman/CEO/President
6423	Danny Hunter	Chairman/CEO/President
6501	Desiree Kim	Chairman/Chief Innovation Officer
6567	suman	CEO/President
6611	rani	Executive Vice President
6653	David Larson	Chairman/CEO/President
6878	seema	CEO/President
6885	Amanda Smith	CEO/Co-Founder/Chairman

	Hiring_candidate_role	Lead_source	\
169	Information officer	Others	
381	Insurance account manager	Others	
469	Minerals surveyor	Contact Email	
519	Restaurant manager	Website	
1016	Civil engineer, consulting	Marketing Event	
1148	Community pharmacist	Contact Email	
1246	Primary school teacher	Marketing Event	
1328	Bonds trader	Others	
1587	Designer, textile	Marketing Event	
1599	Sports administrator	Marketing Event	
1626	Health promotion specialist	Marketing Event	
1636	Associate Professor	Marketing Event	
1651	Printmaker	Others	
1849	Chief Financial Officer	Others	
1943	Nurse, adult	Marketing Event	
2039	Interpreter	Contact Email	
2165	Insurance claims handler	Website	
2241	Restaurant manager	Others	
2469	Embryologist, clinical	Others	
2487	Clinical cytogeneticist	Contact Email	

```
2869                    Engineer, maintenance (IT)    Contact Email
2953                  Fast food restaurant manager           Others
3139             Surveyor, planning and development    Contact Email
3272                              Professor Emeritus  Marketing Event
3327                Surveyor, commercial/residential  Marketing Event
3586                                    Podiatrist          Website
3632             Designer, blown glass/stained glass           Others
3919                   Interior and spatial designer           Others
4269                              Animal nutritionist    Contact Email
4300                                  Charity officer  Marketing Event
4745                           Recruitment consultant           Others
4926                         Magazine features editor    Contact Email
5276                                  Hotel manager           Others
5401                              Librarian, academic    Contact Email
5620                         Teacher, secondary school           Others
5889                                  Doctor, hospital  Marketing Event
6001                           Health service manager    Contact Email
6100                                      Proofreader  Marketing Event
6115           English as a foreign language teacher           Others
6213                Higher education careers adviser  Marketing Event
6423                                  Media planner           Others
6501  Programme researcher, broadcasting/film/video           Others
6567                                  Physiotherapist           Others
6611                             Biochemist, clinical          Website
6653                    Medical sales representative    Contact Email
6878                                Technical author          Website
6885                             Biomedical engineer          Website

     Level_of_meeting                       Last_lead_update  \
169           Level 1                             Up-to-date
381           Level 2                               No track
469           Level 3                               No track
519           Level 3                            5 days back
1016          Level 2                                      ?
1148          Level 3  Following up but lead not responding
1246          Level 2                             Up-to-date
1328          Level 1                               No track
1587          Level 2                   More than a week back
1599          Level 1  Following up but lead not responding
1626          Level 1                                Pending
1636          Level 1       Did not hear back after Level 1
1651          Level 1       Did not hear back after Level 1
1849          Level 3                     more than a month
1943          Level 2                     more than a month
2039          Level 3                               No track
2165          Level 1                                      ?
2241          Level 3                                    NaN
```

2469	Level 2	?	
2487	Level 1	more than a month	
2869	Level 3	Up-to-date	
2953	Level 3	More than a week back	
3139	Level 3	?	
3272	Level 2	No track	
3327	Level 3	NaN	
3586	Level 2	Up-to-date	
3632	Level 3	5 days back	
3919	Level 3	NaN	
4269	Level 3	more than a month	
4300	Level 2	Pending	
4745	Level 1	5 days back	
4926	Level 2	Did not hear back after Level 1	
5276	Level 3	More than 2 weeks	
5401	Level 3	5 days back	
5620	Level 3	2 days back	
5889	Level 2	Did not hear back after Level 1	
6001	Level 3	Following up but lead not responding	
6100	Level 2	Pending	
6115	Level 3	2 days back	
6213	Level 3	No track	
6423	Level 1	Did not hear back after Level 1	
6501	Level 1	More than a week back	
6567	Level 1	Up-to-date	
6611	Level 1	?	
6653	Level 3	NaN	
6878	Level 2	More than 2 weeks	
6885	Level 1	Following up but lead not responding	

	Internal_POC	Resource	\
169	Cashin,Marc C	Cannot deliver	
381	Dunaway,Antoine	Yes	
469	Moran,Natalie A	We have all the requirements	
519	Houston,Arnold E	Deliverable	
1016	Bannister,Joan	Deliverable	
1148	Hanyok,John J	We have all the requirements	
1246	Brown,Maxine A	Deliverable	
1328	Van Arter,Derrick	Not enough	
1587	Vickers Jr.,Henry J	Not enough	
1599	Brown,Maxine A	No	
1626	Maine,John P	Not enough	
1636	Hebron,Artenia D	We have all the requirements	
1651	Himes,Maurice C	Yes	
1849	Davis,Brian R	Yes	
1943	Cash,Tyrone J	Yes	
2039	Pappas,Mark S	We have all the requirements	

2165	Tondeur,Keith	D	Not enough
2241	Massiah,Gerard	F	Not enough
2469	Jones,Eyvette	W	Cannot deliver
2487	Vickers Jr.,Henry	J	Cannot deliver
2869	Pappas,Mark	S	Yes
2953	McKenstry,Loretta	A	No
3139	Georgakopoulos,Vasilios	T	Deliverable
3272	Ullrich,Rose Anne		Cannot deliver
3327	Abdul-Hamid,Saud Muhamad		Cannot deliver
3586	Dyson,William	A	Cannot deliver
3632	Himes,Maurice	C	Deliverable
3919	Gaskins Jr,Franklin	D	We have all the requirements
4269	Hameier,Kurt	E	We have all the requirements
4300	Ryker,David		NaN
4745	Hebron,Artenia	D	Deliverable
4926	Gilley,Janine		Cannot deliver
5276	Hameier,Kurt	E	Yes
5401	Green,Candy		Cannot deliver
5620	Green,Ann	E	We have all the requirements
5889	Murray,Younetta		We have all the requirements
6001	Booker,David	L	No
6100	Himes,Maurice	C	Deliverable
6115	Himes,Maurice	C	Cannot deliver
6213	Houston,Arnold	E	Deliverable
6423	Mabrey,Kevin	C	No
6501	Green,Candy		Not enough
6567	Houston,Arnold	E	We have all the requirements
6611	Gould,Lisa	D	We have all the requirements
6653	Meli,Teresa	V	We have all the requirements
6878	Himes,Maurice	C	Cannot deliver
6885	Moran,Natalie	A	We have all the requirements

	Internal_rating	W_div_D
169	4	inf
381	1	inf
469	2	inf
519	1	inf
1016	4	inf
1148	4	inf
1246	2	inf
1328	2	inf
1587	4	inf
1599	1	inf
1626	5	inf
1636	5	inf
1651	2	inf
1849	3	inf

```
1943                4       inf
2039                1       inf
2165                2       inf
2241                1       inf
2469                5       inf
2487                5       inf
2869                1       inf
2953                4       inf
3139                2       inf
3272                3       inf
3327                3       inf
3586                5       inf
3632                1       inf
3919                5       inf
4269                1       inf
4300                5       inf
4745                3       inf
4926                3       inf
5276                4       inf
5401                3       inf
5620                2       inf
5889                3       inf
6001                1       inf
6100                4       inf
6115                5       inf
6213                1       inf
6423                3       inf
6501                2       inf
6567                4       inf
6611                3       inf
6653                3       inf
6878                2       inf
6885                2       inf
```

[73]:
```
for x in null_trdv_id:
    train_x.loc[x, 'Deal_value'] = int(train_x.loc[x, 'Weighted_amount'] /␣
 ↪train_fac)
```

[74]:
```
null_test_dv_id = test[test.Deal_value.isna()].index
```

[75]:
```
for x in null_test_dv_id:
    test.loc[x, 'Deal_value'] = int(test.loc[x, 'Weighted_amount'] / test_fac)
```

[76]:
```
train_x.drop('W_div_D', axis=1, inplace=True)
test.drop('W_div_D', axis=1, inplace=True)
```

[77]:
```
train_x[(train_x.Geography.isna()) & (train_x.Location.isna())]
```

[77]:
```
                      Industry  Deal_value  Weighted_amount Date_of_creation  \
3466  Architecture/Engineering    485381.0        3252052.7       2021-01-18

          Pitch             Lead_revenue Fund_category Geography Location  \
3466  Product_2  500 Million - 1 Billion    Category 4       NaN      NaN

         POC_name             Designation            Hiring_candidate_role  \
3466  Sarah Lloyd  Chairman/CEO/President  Television production assistant

        Lead_source Level_of_meeting Last_lead_update   Internal_POC Resource  \
3466  Contact Email          Level 2       Up-to-date  Cashin,Marc C      Yes

      Internal_rating
3466                3
```

[78]:
```
train_x.loc[3466,'Geography'] = 'USA'
```

[79]:
```
tr_nul_geo = train_x[train_x.Geography.isna()]
```

[80]:
```
tr_nul_geo
```

[80]:
```
                           Industry  Deal_value  Weighted_amount  \
6                             Banks    384356.0     2.709710e+06
10    Healthcare Facilities/Services    311125.0     1.851194e+06
13                            Banks    487351.0     2.829494e+06
16            Other Investment Firms    390309.0     2.458947e+06
27                         Trucking    380817.0     2.513392e+06
...                              ...         ...              ...
6991               Water Utilities    242256.0     1.707905e+06
6999             Financial Services     31429.0     2.200030e+05
7000          Beverages (Alcoholic)    152908.0     9.709658e+05
7002                          Banks    192800.0     1.195360e+06
7003              Hospitals/Clinics    220208.0     1.453373e+06

     Date_of_creation      Pitch             Lead_revenue Fund_category  \
6          2019-11-20  Product_2  500 Million - 1 Billion    Category 1
10         2020-01-30  Product_1  500 Million - 1 Billion    Category 2
13         2020-04-25  Product_1  500 Million - 1 Billion    Category 3
16         2019-10-13  Product_2        50 - 100 Million    Category 2
27         2019-05-31  Product_2        50 - 100 Million    Category 1
...               ...        ...                     ...           ...
6991       2019-01-16  Product_2  500 Million - 1 Billion    Category 1
6999       2020-07-04  Product_2  500 Million - 1 Billion    Category 4
7000       2020-12-08  Product_1       100 - 500 Million    Category 1
7002       2020-12-07  Product_1       100 - 500 Million    Category 4
7003       2020-03-13  Product_2       100 - 500 Million    Category 1
```

```
     Geography                                  Location         POC_name  \
6          NaN                        Salisbury, MD-DE       Sara Dixon
10         NaN          Oxnard-Thousand Oaks-Ventura, CA   Tamara Sanchez
13         NaN                                 Pilibhit          saaniya
16         NaN  Miami-Fort Lauderdale-West Palm Beach, FL    Joshua Wright
27         NaN              Chambersburg-Waynesboro, PA   Jennifer Davis
...        ...                                      ...              ...
6991       NaN              Shreveport-Bossier City, LA  Nicholas Kaiser
6999       NaN                           Knoxville, TN   Valerie Wilson
7000       NaN                                  Kohima      sita kumari
7002       NaN                               Kagaznagar   smt. chanchala
7003       NaN                                Proddatur    geeta @ komal

                         Designation             Hiring_candidate_role  \
6               CEO/Co-Founder/Chairman                       Cartographer
10              CEO/Co-Founder/Chairman                    Hydrogeologist
13    Chairman/Chief Innovation Officer                   Hospital doctor
16               Chairman/CEO/President              Broadcast journalist
27                       CEO/President  Armed forces operational officer
...                                ...                               ...
6991               SVP/General Counsel               Engineer, automotive
6999            CEO/Co-Founder/Chairman            Applications developer
7000                     CEO/President          Counselling psychologist
7002            CEO/Co-Founder/Chairman               Call centre manager
7003                                CEO            Financial risk analyst

         Lead_source Level_of_meeting                    Last_lead_update  \
6       Contact Email          Level 3                  More than 2 weeks
10              Website         Level 1  Following up but lead not responding
13    Marketing Event          Level 3                More than a week back
16      Contact Email          Level 3                                   ?
27      Contact Email          Level 3       Did not hear back after Level 1
...               ...              ...                                 ...
6991  Marketing Event          Level 3                             Pending
6999          Website          Level 3                          Up-to-date
7000          Website          Level 2                        5 days back
7002    Contact Email          Level 1                More than a week back
7003  Marketing Event          Level 2                                   ?

          Internal_POC                     Resource  Internal_rating
6       Booker,David L                          NaN               4
10    Clavey,Therese A                Cannot deliver               2
13    Shelton,Sidney P  We have all the requirements               2
16    Dimattia,Frank D                   Deliverable               4
27        Leu,Darren L                            No               1
...                ...                          ...             ...
6991   Robinson,John C                          NaN               1
```

```
6999   Young,Valerie K                   Deliverable                4
7000      Gould,Lisa D                            No                1
7002   Jones,Eyvette W  We have all the requirements                4
7003    Brown,Maxine A  We have all the requirements                3

[971 rows x 18 columns]
```

[81]:
```
tr_nul_geo['Region'] = tr_nul_geo.Location.str.split(',')
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\4160386935.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  tr_nul_geo['Region'] = tr_nul_geo.Location.str.split(',')
```

[82]:
```
tr_nul_geo
```

[82]:
```
                            Industry  Deal_value  Weighted_amount  \
6                              Banks    384356.0     2.709710e+06
10    Healthcare Facilities/Services    311125.0     1.851194e+06
13                             Banks    487351.0     2.829494e+06
16          Other Investment Firms    390309.0     2.458947e+06
27                          Trucking    380817.0     2.513392e+06
...                              ...         ...              ...
6991                 Water Utilities    242256.0     1.707905e+06
6999              Financial Services     31429.0     2.200030e+05
7000           Beverages (Alcoholic)    152908.0     9.709658e+05
7002                           Banks    192800.0     1.195360e+06
7003               Hospitals/Clinics    220208.0     1.453373e+06

     Date_of_creation      Pitch          Lead_revenue Fund_category  \
6          2019-11-20  Product_2  500 Million - 1 Billion    Category 1
10         2020-01-30  Product_1  500 Million - 1 Billion    Category 2
13         2020-04-25  Product_1  500 Million - 1 Billion    Category 3
16         2019-10-13  Product_2         50 - 100 Million    Category 2
27         2019-05-31  Product_2         50 - 100 Million    Category 1
...               ...        ...                      ...           ...
6991       2019-01-16  Product_2  500 Million - 1 Billion    Category 1
6999       2020-07-04  Product_2  500 Million - 1 Billion    Category 4
7000       2020-12-08  Product_1        100 - 500 Million    Category 1
7002       2020-12-07  Product_1        100 - 500 Million    Category 4
7003       2020-03-13  Product_2        100 - 500 Million    Category 1

     Geography                              Location        POC_name  \
6          NaN                         Salisbury, MD-DE      Sara Dixon
```

```
10         NaN           Oxnard-Thousand Oaks-Ventura, CA   Tamara Sanchez
13         NaN                                   Pilibhit          saaniya
16         NaN  Miami-Fort Lauderdale-West Palm Beach, FL    Joshua Wright
27         NaN                Chambersburg-Waynesboro, PA   Jennifer Davis
...        ...                                        ...             ...
6991       NaN                Shreveport-Bossier City, LA  Nicholas Kaiser
6999       NaN                             Knoxville, TN   Valerie Wilson
7000       NaN                                    Kohima      sita kumari
7002       NaN                                 Kagaznagar   smt. chanchala
7003       NaN                                  Proddatur    geeta @ komal

                        Designation            Hiring_candidate_role  \
6              CEO/Co-Founder/Chairman                     Cartographer
10             CEO/Co-Founder/Chairman                   Hydrogeologist
13   Chairman/Chief Innovation Officer                  Hospital doctor
16              Chairman/CEO/President              Broadcast journalist
27                      CEO/President  Armed forces operational officer
...                               ...                             ...
6991              SVP/General Counsel              Engineer, automotive
6999           CEO/Co-Founder/Chairman            Applications developer
7000                    CEO/President          Counselling psychologist
7002           CEO/Co-Founder/Chairman               Call centre manager
7003                               CEO             Financial risk analyst

        Lead_source Level_of_meeting                    Last_lead_update  \
6      Contact Email          Level 3                   More than 2 weeks
10           Website          Level 1  Following up but lead not responding
13   Marketing Event          Level 3                More than a week back
16     Contact Email          Level 3                                   ?
27     Contact Email          Level 3       Did not hear back after Level 1
...              ...              ...                                 ...
6991 Marketing Event          Level 3                             Pending
6999         Website          Level 3                          Up-to-date
7000         Website          Level 2                        5 days back
7002   Contact Email          Level 1                More than a week back
7003 Marketing Event          Level 2                                   ?

         Internal_POC                    Resource  Internal_rating  \
6       Booker,David L                         NaN                4
10    Clavey,Therese A              Cannot deliver                2
13    Shelton,Sidney P  We have all the requirements                2
16    Dimattia,Frank D                 Deliverable                4
27       Leu,Darren L                          No                1
...               ...                         ...              ...
6991   Robinson,John C                         NaN                1
6999   Young,Valerie K                 Deliverable                4
7000      Gould,Lisa D                          No                1
```

```
7002   Jones,Eyvette W  We have all the requirements                4
7003    Brown,Maxine A  We have all the requirements                3

                                                 Region
6                                [Salisbury,  MD-DE]
10               [Oxnard-Thousand Oaks-Ventura,  CA]
13                                        [Pilibhit]
16      [Miami-Fort Lauderdale-West Palm Beach,  FL]
27                    [Chambersburg-Waynesboro,  PA]
...                                             ...
6991                  [Shreveport-Bossier City,  LA]
6999                               [Knoxville,  TN]
7000                                       [Kohima]
7002                                   [Kagaznagar]
7003                                    [Proddatur]

[971 rows x 19 columns]
```

[83]:
```
tr_null_geo_id = tr_nul_geo.index
```

[84]:
```
for x in tr_null_geo_id:
    if len(tr_nul_geo.loc[x, 'Region']) == 1:
        train_x.loc[x, 'Geography'] = 'India'
    else:
        train_x.loc[x, 'Geography'] = 'USA'
```

[85]:
```
train_x.loc[tr_null_geo_id]
```

[85]:
```
                            Industry  Deal_value  Weighted_amount  \
6                              Banks    384356.0     2.709710e+06
10    Healthcare Facilities/Services    311125.0     1.851194e+06
13                             Banks    487351.0     2.829494e+06
16            Other Investment Firms    390309.0     2.458947e+06
27                          Trucking    380817.0     2.513392e+06
...                              ...         ...              ...
6991                Water Utilities    242256.0     1.707905e+06
6999             Financial Services     31429.0     2.200030e+05
7000          Beverages (Alcoholic)    152908.0     9.709658e+05
7002                           Banks    192800.0     1.195360e+06
7003             Hospitals/Clinics    220208.0     1.453373e+06

     Date_of_creation      Pitch             Lead_revenue Fund_category  \
6          2019-11-20  Product_2  500 Million - 1 Billion    Category 1
10         2020-01-30  Product_1  500 Million - 1 Billion    Category 2
13         2020-04-25  Product_1  500 Million - 1 Billion    Category 3
16         2019-10-13  Product_2        50 - 100 Million    Category 2
27         2019-05-31  Product_2        50 - 100 Million    Category 1
```

```
...              ...        ...                        ...          ...
6991       2019-01-16  Product_2  500 Million - 1 Billion    Category 1
6999       2020-07-04  Product_2  500 Million - 1 Billion    Category 4
7000       2020-12-08  Product_1        100 - 500 Million    Category 1
7002       2020-12-07  Product_1        100 - 500 Million    Category 4
7003       2020-03-13  Product_2        100 - 500 Million    Category 1

     Geography                                  Location         POC_name  \
6          USA                          Salisbury, MD-DE       Sara Dixon
10         USA          Oxnard-Thousand Oaks-Ventura, CA   Tamara Sanchez
13       India                                  Pilibhit          saaniya
16         USA  Miami-Fort Lauderdale-West Palm Beach, FL    Joshua Wright
27         USA               Chambersburg-Waynesboro, PA   Jennifer Davis
...        ...                                       ...              ...
6991       USA               Shreveport-Bossier City, LA  Nicholas Kaiser
6999       USA                             Knoxville, TN   Valerie Wilson
7000     India                                    Kohima      sita kumari
7002     India                                 Kagaznagar   smt. chanchala
7003     India                                  Proddatur    geeta @ komal

                         Designation             Hiring_candidate_role  \
6               CEO/Co-Founder/Chairman                       Cartographer
10              CEO/Co-Founder/Chairman                    Hydrogeologist
13    Chairman/Chief Innovation Officer                   Hospital doctor
16               Chairman/CEO/President              Broadcast journalist
27                       CEO/President  Armed forces operational officer
...                                ...                               ...
6991               SVP/General Counsel               Engineer, automotive
6999            CEO/Co-Founder/Chairman            Applications developer
7000                     CEO/President          Counselling psychologist
7002            CEO/Co-Founder/Chairman               Call centre manager
7003                                CEO            Financial risk analyst

          Lead_source Level_of_meeting                    Last_lead_update  \
6       Contact Email          Level 3                   More than 2 weeks
10            Website          Level 1  Following up but lead not responding
13    Marketing Event          Level 3                More than a week back
16      Contact Email          Level 3                                   ?
27      Contact Email          Level 3       Did not hear back after Level 1
...               ...              ...                                 ...
6991  Marketing Event          Level 3                             Pending
6999          Website          Level 3                          Up-to-date
7000          Website          Level 2                           5 days back
7002    Contact Email          Level 1                More than a week back
7003  Marketing Event          Level 2                                   ?

          Internal_POC                    Resource  Internal_rating
```

```
6       Booker,David L                          NaN                4
10    Clavey,Therese A              Cannot deliver                 2
13    Shelton,Sidney P  We have all the requirements               2
16    Dimattia,Frank D                  Deliverable                4
27       Leu,Darren L                           No                 1
...               ...                         ...               ...
6991   Robinson,John C                        NaN                 1
6999   Young,Valerie K                 Deliverable                4
7000      Gould,Lisa D                         No                 1
7002   Jones,Eyvette W  We have all the requirements               4
7003    Brown,Maxine A  We have all the requirements               3

[971 rows x 18 columns]
```

[86]:
```
test[(test.Geography.isna()) & (test.Location.isna())]
```

[86]:
```
                    Industry  Deal_value  Weighted_amount Date_of_creation  \
1514  Constr/Agric Machinery    264533.0     1.556917e+06       2019-06-23

          Pitch      Lead_revenue Fund_category Geography Location  \
1514  Product_1  50 - 100 Million    Category 1       NaN      NaN

         POC_name           Designation Hiring_candidate_role Lead_source  \
1514  Randy Ramos  SVP/General Counsel     Arts administrator     Website

     Level_of_meeting                      Last_lead_update      Internal_POC  \
1514          Level 2  Following up but lead not responding  Massiah,Gerard F

            Resource  Internal_rating
1514  Cannot deliver              1.0
```

[87]:
```
test.loc[1514,'Geography'] = 'USA'
```

[88]:
```
test_nul_geo = test[test.Geography.isna()]
```

[89]:
```
test_nul_geo['Region'] = test_nul_geo.Location.str.split(',')
```

```
C:\Users\dell\AppData\Local\Temp\ipykernel_23844\4177573635.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  test_nul_geo['Region'] = test_nul_geo.Location.str.split(',')
```

[90]:
```
test_nul_geo
```

[90]:

	Industry	Deal_value	Weighted_amount	Date_of_creation
44	Oil, Gas, Coal	460575.0	1.556917e+06	2019-03-02
92	Associations	49910.0	2.994600e+05	2020-03-31
120	Insurance	334274.0	2.206208e+06	2020-12-07
163	Oil, Gas, Coal	421478.0	1.556917e+06	2020-11-25
230	Banks	20465.0	1.330225e+05	2020-04-09
...	...	...	...	...
1906	REIT	392073.0	2.587682e+06	2019-09-30
1962	Banks	137830.0	9.234610e+05	2020-08-25
2020	Financial Services	114371.0	7.205373e+05	2020-08-09
2079	Education/Training	490432.0	2.844506e+06	2021-01-31
2082	Restaurants	5082.0	3.430350e+04	2019-09-17

	Pitch	Lead_revenue	Fund_category	Geography
44	Product_1	100 - 500 Million	Category 2	NaN
92	Product_2	500 Million - 1 Billion	Category 2	NaN
120	Product_2	50 - 100 Million	Category 4	NaN
163	Product_1	100 - 500 Million	Category 4	NaN
230	Product_2	500 Million - 1 Billion	Category 4	NaN
...	...	...	...	...
1906	Product_2	50 - 100 Million	Category 3	NaN
1962	Product_2	500 Million - 1 Billion	Category 2	NaN
2020	Product_1	500 Million - 1 Billion	Category 1	NaN
2079	Product_2	50 - 100 Million	Category 3	NaN
2082	Product_1	100 - 500 Million	Category 4	NaN

	Location	POC_name	Designation
44	Jefferson City, MO	Anthony Garcia	Chairman/CEO/President
92	Morgantown, WV	Christian Kelly	Executive Vice President
120	Odessa, TX	Kevin Wells	Vice President / GM (04-present) : VP Sales an...
163	Rajapalaiyam	dipika ahuja	CEO
230	Cape Girardeau, MO-IL	Richard Clark	Chief Executive Officer
...	...	...	...
1906	Raleigh, NC	Samuel Osborne	CEO/President
1962	Vizianagaram	smts. manju	
2020	Ghandinagar	aasu	
2079	Stockton-Lodi, CA	Steven Herrera	
2082	Memphis, TN-MS-AR	Ralph Lopez	

```
1962                            Chairman/CEO/President
2020                            CEO/Chairman/President
2079                                    CEO/President
2082                            CEO/Chairman/President

              Hiring_candidate_role      Lead_source Level_of_meeting  \
44                  Tourism officer          Website          Level 2
92                Scientist, forensic  Marketing Event          Level 2
120              Engineer, production          Website          Level 2
163                       Press sub          Website          Level 1
230       Museum/gallery conservator    Contact Email          Level 2
...                               ...              ...              ...
1906     Careers information officer  Marketing Event          Level 3
1962  Commercial/residential surveyor          Website          Level 3
2020           Amenity horticulturist    Contact Email          Level 1
2079            Engineering geologist          Website          Level 1
2082    Garment/textile technologist  Marketing Event          Level 3

                        Last_lead_update         Internal_POC  \
44                                     ?  Anthony,Katherine D
92                            Up-to-date      Dunaway,Antoine
120                          2 days back        Thomas,Lori E
163   Following up but lead not responding     Clavey,Therese A
230                                    ?         Morsy,Omar A
...                                  ...                  ...
1906                          Up-to-date        Cashin,Marc C
1962                                   ?     Clavey,Therese A
2020                          Up-to-date    Sutton,Michelle R
2079                  More than 2 weeks          Ryker,David
2082                           No track    Ullrich,Rose Anne

                         Resource  Internal_rating                   Region
44                     Not enough             4.0     [Jefferson City,  MO]
92                     Not enough             1.0         [Morgantown,  WV]
120                    Not enough             3.0             [Odessa,  TX]
163                           NaN             4.0           [Rajapalaiyam]
230                Cannot deliver             1.0  [Cape Girardeau,  MO-IL]
...                           ...             ...                       ...
1906                          Yes             4.0            [Raleigh,  NC]
1962                  Deliverable             5.0           [Vizianagaram]
2020  We have all the requirements             4.0            [Ghandinagar]
2079  We have all the requirements             2.0      [Stockton-Lodi,  CA]
2082                            No             2.0      [Memphis,  TN-MS-AR]

[78 rows x 19 columns]
```

[91]:
```
test_nul_geo_id = test_nul_geo.index
```

[92]:
```
for x in test_nul_geo_id:
    if len(test_nul_geo.loc[x, 'Region']) == 1:
        test.loc[x, 'Geography'] = 'India'
    else:
        test.loc[x, 'Geography'] = 'USA'
```

[93]:
```
train_x[train_x.Location.isna()]
```

[93]:

	Industry	Deal_value	Weighted_amount	Date_of_creation	\
545	Investment Bank/Brokerage	321315.0	2056416.00	2019-06-30	
810	REIT	457184.0	3108851.20	2019-03-30	
1597	Insurance	134465.0	793343.50	2020-10-07	
1623	Banks	380753.0	2684308.65	2020-08-18	
1698	Telecom Consulting	383929.0	2687503.00	2020-01-14	
1901	Investment Bank/Brokerage	37105.0	222630.00	2019-09-27	
3466	Architecture/Engineering	485381.0	3252052.70	2021-01-18	
4201	Other Biz Services	15638.0	103992.70	2020-05-21	
5222	Staffing	449933.0	2924564.50	2020-12-04	
5675	Financial Services	164299.0	1026868.75	2021-01-17	
6862	Banks	63427.0	396418.75	2019-01-12	

	Pitch	Lead_revenue	Fund_category	Geography	Location	\
545	Product_1	100 - 500 Million	Category 1	USA	NaN	
810	Product_1	50 - 100 Million	Category 1	USA	NaN	
1597	Product_1	500 Million - 1 Billion	Category 3	USA	NaN	
1623	Product_1	500 Million - 1 Billion	Category 3	USA	NaN	
1698	Product_2	50 - 100 Million	Category 4	USA	NaN	
1901	Product_2	500 Million - 1 Billion	Category 2	USA	NaN	
3466	Product_2	500 Million - 1 Billion	Category 4	USA	NaN	
4201	Product_2	50 - 100 Million	Category 4	USA	NaN	
5222	Product_2	500 Million - 1 Billion	Category 4	USA	NaN	
5675	Product_1	100 - 500 Million	Category 2	USA	NaN	
6862	Product_2	100 - 500 Million	Category 3	USA	NaN	

	POC_name	Designation	\
545	Jennifer Thomas	SVP/General Counsel	
810	Bianca Kelly	CEO/Co-Founder/Chairman	
1597	Colin Camacho	Chairman/CEO/President	
1623	David Williams	CEO/Chairman/President	
1698	Brendan Reid	Vice President / GM (04-present) : VP Sales an...	
1901	Virginia Medina	Chairman/CEO/President	
3466	Sarah Lloyd	Chairman/CEO/President	
4201	Joseph Brown	Chairman/Chief Innovation Officer	
5222	Daniel Delacruz	CEO/Co-Founder/Chairman	
5675	Charles Robinson	Executive Vice President	
6862	Elizabeth House	CEO/Chairman/President	

```
                Hiring_candidate_role      Lead_source Level_of_meeting  \
545         Nurse, learning disability    Contact Email          Level 2
810                      Pharmacologist           Others          Level 3
1597         Radio broadcast assistant           Others          Level 2
1623          Arts development officer  Marketing Event          Level 3
1698   Dance movement psychotherapist    Contact Email          Level 3
1901           General practice doctor  Marketing Event          Level 2
3466  Television production assistant    Contact Email          Level 2
4201      Conservator, museum/gallery  Marketing Event          Level 2
5222             Multimedia specialist    Contact Email          Level 1
5675    Engineer, civil (contracting)           Others          Level 3
6862                Information officer         Website          Level 2

                         Last_lead_update           Internal_POC  \
545                      More than 2 weeks    McKenstry,Loretta A
810                              Up-to-date       Salyers,Daniel L
1597  Following up but lead not responding         Mabrey,Kevin C
1623                             Up-to-date        Jones,Michael L
1698                             Up-to-date            Ali,Mohamed
1901  Following up but lead not responding         Bannister,Joan
3466                             Up-to-date          Cashin,Marc C
4201                           5 days back      Van Arter,Derrick
5222  Following up but lead not responding  Featherstone,Adrian R
5675                               No track          Pappas,Mark S
6862                  More than a week back       Davis,Sharrice A

            Resource  Internal_rating
545       Not enough               1
810   Cannot deliver               5
1597            Yes                1
1623      Not enough               5
1698  Cannot deliver               5
1901              No               1
3466             Yes               3
4201      Not enough               5
5222             NaN               4
5675  Cannot deliver               4
6862     Deliverable               3
```

[94]:
```
train_x[train_x.Geography == 'USA'].Location.value_counts()
```

[94]:
```
Panama City, FL                      18
Ann Arbor, MI                        17
Phoenix-Mesa-Scottsdale, AZ          17
Los Angeles-Long Beach-Anaheim, CA   16
Milwaukee-Waukesha-West Allis, WI    16
                                     ..
```

```
Decatur, IL                              3
Farmington, NM                           3
Beckley, WV                              3
California-Lexington Park, MD            3
Jackson, MI                              3
Name: Location, Length: 385, dtype: int64
```

I'm going to replace any missing Location values in the train_x dataset with "Panama City, FL".

[95]:
```
train_x.Location.fillna('Panama City, FL', inplace=True)
```

[96]:
```
test[test.Location.isna()]
```

[96]:
```
                    Industry  Deal_value  Weighted_amount Date_of_creation  \
1000       Hospitals/Clinics    293714.0     1.835712e+06       2020-03-31
1514  Constr/Agric Machinery    264533.0     1.556917e+06       2019-06-23
1690                   Banks    120602.0     7.236120e+05       2019-09-10

          Pitch             Lead_revenue Fund_category Geography Location  \
1000  Product_2         50 - 100 Million    Category 1       USA      NaN
1514  Product_1         50 - 100 Million    Category 1       USA      NaN
1690  Product_2  500 Million - 1 Billion    Category 2       USA      NaN

         POC_name              Designation            Hiring_candidate_role  \
1000   James Dean  CEO/Co-Founder/Chairman               Academic librarian
1514  Randy Ramos     SVP/General Counsel                Arts administrator
1690  Logan Smith     SVP/General Counsel   Psychotherapist, dance movement

        Lead_source Level_of_meeting                      Last_lead_update  \
1000  Contact Email          Level 2                           5 days back
1514        Website          Level 2  Following up but lead not responding
1690         Others          Level 2                                     ?

          Internal_POC                      Resource  Internal_rating
1000    Mabrey,Kevin C  We have all the requirements             2.0
1514  Massiah,Gerard F                Cannot deliver             1.0
1690   Moran,Natalie A                            No             4.0
```

[97]:
```
test[(test.Geography == 'USA')].Location.value_counts().head(10)
```

[97]:
```
Riverside-San Bernardino-Ontario, CA     7
Racine, WI                               7
Farmington, NM                           7
Chattanooga, TN-GA                       7
Montgomery, AL                           7
Tucson, AZ                               7
Omaha-Council Bluffs, NE-IA              7
Coeur d'Alene, ID                        6
```

```
Fayetteville-Springdale-Rogers, AR-MO    6
Augusta-Richmond County, GA-SC           6
Name: Location, dtype: int64
```

I'm planning to replace any empty Location values in the test dataset with "Farmington, NM".

[98]:
```
test.Location.fillna('Farmington, NM', inplace=True)
```

[99]:
```
train_x[train_x.POC_name.isna()]
```

[99]:

	Industry	Deal_value	Weighted_amount	Date_of_creation	\
489	Insurance	209984.0	1186409.60	2019-06-07	
599	Insurance	326899.0	2043118.75	2019-07-04	
1862	Banks	457421.0	2813139.15	2019-06-04	
2008	Servers/Storage	118302.0	727557.30	2020-03-10	
2053	Banks	90309.0	546369.45	2021-02-07	
4142	Construction Services	208962.0	1211979.60	2019-11-08	
5040	Health/Accident	162684.0	1081848.60	2020-11-30	
6468	Banks	73637.0	449185.70	2019-04-30	

	Pitch	Lead_revenue	Fund_category	Geography	Location	\
489	Product_2	50 - 100 Million	Category 3	India	Alipurduar	
599	Product_1	50 - 100 Million	Category 1	India	Sirsa	
1862	Product_2	500 Million - 1 Billion	Category 2	India	Raurkela	
2008	Product_2	500 Million - 1 Billion	Category 2	India	Daman	
2053	Product_2	50 - 100 Million	Category 4	India	Thanjavur	
4142	Product_2	500 Million - 1 Billion	Category 2	India	Bareilly	
5040	Product_1	500 Million - 1 Billion	Category 4	India	Indore	
6468	Product_1	100 - 500 Million	Category 3	India	Ludhiana	

	POC_name	Designation	Hiring_candidate_role	\
489	NaN	Executive Vice President	Interior and spatial designer	
599	NaN	Chief Executive Officer	Teacher, secondary school	
1862	NaN	SVP/General Counsel	Statistician	
2008	NaN	Chief Executive Officer	Commissioning editor	
2053	NaN	CEO/Co-Founder/Chairman	Early years teacher	
4142	NaN	CEO/Chairman/President	Environmental health practitioner	
5040	NaN	Executive Vice President	Designer, jewellery	
6468	NaN	CEO/President	Horticulturist, commercial	

	Lead_source	Level_of_meeting	Last_lead_update	\
489	Contact Email	Level 2	2 days back	
599	Contact Email	Level 2	more than a month	
1862	Contact Email	Level 3	No track	
2008	Contact Email	Level 3	Pending	
2053	Marketing Event	Level 1	5 days back	
4142	Contact Email	Level 2	Up-to-date	
5040	Others	Level 3	?	

```
6468  Marketing Event          Level 1  More than a week back

            Internal_POC                     Resource  Internal_rating
489        Booker,David L                   Not enough                2
599      Davis,Sharrice A                  Deliverable                3
1862     Dimattia,Frank D                   Not enough                1
2008        Gilley,Janine  We have all the requirements                1
2053          Ross,Eric L                  Deliverable                1
4142        Davis,Brian R                   Not enough                2
5040  Vickers Jr.,Henry J                   Not enough                3
6468        Thomas,Lori E                          Yes                5
```

[100]:
```
tr_pocnm_id = train_x[train_x.POC_name.isna()].index
```

[101]:
```
train.loc[tr_pocnm_id]
```

[101]:
```
      Deal_title                          Lead_name               Industry  \
489   TitleS2D23  Martinez, Gutierrez and Knapp PLC              Insurance
599   TitleST3S0      Anderson, Buckley and Lee LLC              Insurance
1862  TitleDQJEL                  Armstrong Inc LLC                  Banks
2008  Title612IP      King, Davenport and James LLC        Servers/Storage
2053  TitleEPQ2Q                   Lozano-Clark Inc                  Banks
4142  TitleU3Y7W                    Lee-Alvarez Ltd  Construction Services
5040  TitleDBQIT       Hobbs, Reyes and Schmidt LLC        Health/Accident
6468  TitleQI25X                Forbes Ltd and Sons                  Banks

     Deal_value Weighted_amount Date_of_creation      Pitch  \
489     209984$      1186409.6$       2019-06-07  Product_2
599     326899$     2043118.75$       2019-07-04  Product_1
1862    457421$     2813139.15$       2019-06-04  Product_2
2008    118302$       727557.3$       2020-03-10  Product_2
2053     90309$      546369.45$       2021-02-07  Product_2
4142    208962$      1211979.6$       2019-11-08  Product_2
5040    162684$      1081848.6$       2020-11-30  Product_1
6468     73637$       449185.7$       2019-04-30  Product_1

              Contact_no             Lead_revenue Fund_category Geography  \
489          358-063-2204        50 - 100 Million    Category 3     India
599            9576256452        50 - 100 Million    Category 1     India
1862  +1-816-342-8073x681  500 Million - 1 Billion    Category 2     India
2008           6521077903  500 Million - 1 Billion    Category 2     India
2053   (032)148-1914x4137        50 - 100 Million    Category 4     India
4142    094-918-7390x0378  500 Million - 1 Billion    Category 2     India
5040         287.728.4362  500 Million - 1 Billion    Category 4     India
6468  +1-058-252-2637x094       100 - 500 Million    Category 3     India

        Location POC_name              Designation  \
```

```
489   Alipurduar      NaN  Executive Vice President
599        Sirsa      NaN   Chief Executive Officer
1862    Raurkela      NaN      SVP/General Counsel
2008       Daman      NaN   Chief Executive Officer
2053   Thanjavur      NaN   CEO/Co-Founder/Chairman
4142    Bareilly      NaN    CEO/Chairman/President
5040      Indore      NaN  Executive Vice President
6468    Ludhiana      NaN             CEO/President

                  Lead_POC_email              Hiring_candidate_role  \
489    chelseajohnson@martinez.com      Interior and spatial designer
599       williamcole@anderson.com          Teacher, secondary school
1862   jamesmarshall@armstrong.com                       Statistician
2008         wendygilbert@king.com               Commissioning editor
2053  lisacaldwell@lozanoclark.com                Early years teacher
4142  waynepeterson@leealvarez.com  Environmental health practitioner
5040       rebeccabrewer@hobbs.com                Designer, jewellery
6468   krystalschroeder@forbes.com         Horticulturist, commercial

         Lead_source Level_of_meeting       Last_lead_update  \
489     Contact Email          Level 2            2 days back
599     Contact Email          Level 2     more than a month
1862    Contact Email          Level 3              No track
2008    Contact Email          Level 3               Pending
2053  Marketing Event          Level 1            5 days back
4142    Contact Email          Level 2             Up-to-date
5040           Others          Level 3                      ?
6468  Marketing Event          Level 1  More than a week back

            Internal_POC                     Resource  Internal_rating  \
489        Booker,David L                   Not enough                2
599      Davis,Sharrice A                   Deliverable               3
1862     Dimattia,Frank D                   Not enough                1
2008        Gilley,Janine  We have all the requirements               1
2053          Ross,Eric L                   Deliverable               1
4142        Davis,Brian R                   Not enough                2
5040  Vickers Jr.,Henry J                   Not enough                3
6468        Thomas,Lori E                          Yes                5

      Success_probability
489                 55.60
599                 61.80
1862                61.70
2008                26.35
2053                56.70
4142                26.35
5040                64.20
```

```
6468                61.00
```

I will create a list based on the email addresses found in the train dataset.

[102]:
```
poc_dict = {0:'Chelsea Johnson', 1:'William Cole', 2:'James Marshall', 3:'Wendy␣
↪Gilbert', 4:'Lisa Caldwell', 5:'Wayne Peterson', 6:'Rebecca Brewer', 7:
↪'Krystal Schroeder'}
```

Populate the empty POC name fields using the information from the poc_list.

[103]:
```
poc_dict.values()
```

[103]:
```
dict_values(['Chelsea Johnson', 'William Cole', 'James Marshall', 'Wendy
Gilbert', 'Lisa Caldwell', 'Wayne Peterson', 'Rebecca Brewer', 'Krystal
Schroeder'])
```

[104]:
```
poc_dict_df = pd.DataFrame(poc_dict.values(), columns=['A'])
```

[105]:
```
poc_dict_df
```

[105]:
```
                  A
0    Chelsea Johnson
1       William Cole
2     James Marshall
3      Wendy Gilbert
4      Lisa Caldwell
5     Wayne Peterson
6     Rebecca Brewer
7  Krystal Schroeder
```

[106]:
```
poc_list = list(poc_dict.values())
```

[107]:
```
n = 0
for x in tr_pocnm_id:
    train_x.loc[x,"POC_name"] = poc_list[n]
    n += 1
```

[108]:
```
train_x.loc[tr_pocnm_id]
```

[108]:
```
                   Industry  Deal_value  Weighted_amount Date_of_creation  \
489                Insurance    209984.0       1186409.60       2019-06-07
599                Insurance    326899.0       2043118.75       2019-07-04
1862                   Banks    457421.0       2813139.15       2019-06-04
2008        Servers/Storage     118302.0        727557.30       2020-03-10
2053                   Banks     90309.0        546369.45       2021-02-07
4142  Construction Services    208962.0       1211979.60       2019-11-08
5040        Health/Accident     162684.0       1081848.60       2020-11-30
6468                   Banks     73637.0        449185.70       2019-04-30
```

```
         Pitch             Lead_revenue Fund_category Geography    Location  \
489   Product_2         50 - 100 Million    Category 3     India  Alipurduar
599   Product_1         50 - 100 Million    Category 1     India       Sirsa
1862  Product_2  500 Million - 1 Billion    Category 2     India    Raurkela
2008  Product_2  500 Million - 1 Billion    Category 2     India       Daman
2053  Product_2         50 - 100 Million    Category 4     India   Thanjavur
4142  Product_2  500 Million - 1 Billion    Category 2     India    Bareilly
5040  Product_1  500 Million - 1 Billion    Category 4     India      Indore
6468  Product_1        100 - 500 Million    Category 3     India    Ludhiana

              POC_name               Designation  \
489     Chelsea Johnson  Executive Vice President
599        William Cole   Chief Executive Officer
1862     James Marshall      SVP/General Counsel
2008      Wendy Gilbert   Chief Executive Officer
2053      Lisa Caldwell   CEO/Co-Founder/Chairman
4142     Wayne Peterson    CEO/Chairman/President
5040     Rebecca Brewer  Executive Vice President
6468  Krystal Schroeder             CEO/President

                 Hiring_candidate_role      Lead_source Level_of_meeting  \
489       Interior and spatial designer    Contact Email          Level 2
599           Teacher, secondary school    Contact Email          Level 2
1862                      Statistician    Contact Email          Level 3
2008              Commissioning editor    Contact Email          Level 3
2053               Early years teacher  Marketing Event          Level 1
4142  Environmental health practitioner    Contact Email          Level 2
5040               Designer, jewellery           Others          Level 3
6468        Horticulturist, commercial  Marketing Event          Level 1

          Last_lead_update         Internal_POC  \
489            2 days back       Booker,David L
599      more than a month     Davis,Sharrice A
1862              No track     Dimattia,Frank D
2008              Pending        Gilley,Janine
2053           5 days back          Ross,Eric L
4142            Up-to-date        Davis,Brian R
5040                     ?  Vickers Jr.,Henry J
6468  More than a week back        Thomas,Lori E

                       Resource  Internal_rating
489                    Not enough               2
599                   Deliverable               3
1862                   Not enough               1
2008  We have all the requirements               1
2053                  Deliverable               1
4142                   Not enough               2
```

```
5040                    Not enough               3
6468                           Yes               5
```

[109]:
```
test[test.POC_name.isna()]
```

[109]:
```
                          Industry  Deal_value  Weighted_amount  \
744   Healthcare Facilities/Services    287570.0        1725420.0
1095                     Real Estate    132712.0         855992.4

     Date_of_creation      Pitch             Lead_revenue Fund_category  \
744        2019-03-04  Product_1  500 Million - 1 Billion    Category 2
1095       2019-02-21  Product_1         50 - 100 Million    Category 4

     Geography    Location POC_name             Designation  \
744      India   Hyderabad      NaN           CEO/President
1095     India  Ahmadnagar      NaN  CEO/Chairman/President

           Hiring_candidate_role      Lead_source Level_of_meeting  \
744   Building services engineer  Marketing Event          Level 2
1095       Geneticist, molecular  Marketing Event          Level 2

           Last_lead_update     Internal_POC     Resource  Internal_rating
744             2 days back  Jones,Michael L  Deliverable              1.0
1095  More than a week back  Kiepea,Prince A  Deliverable              4.0
```

[110]:
```
test_ponm_id = test[test.POC_name.isna()].index
```

Complete the POC name entries in the test dataset with the value 'pooja'.

[111]:
```
test.POC_name.fillna('pooja', inplace=True)
```

[112]:
```
test.loc[test_ponm_id]
```

[112]:
```
                          Industry  Deal_value  Weighted_amount  \
744   Healthcare Facilities/Services    287570.0        1725420.0
1095                     Real Estate    132712.0         855992.4

     Date_of_creation      Pitch             Lead_revenue Fund_category  \
744        2019-03-04  Product_1  500 Million - 1 Billion    Category 2
1095       2019-02-21  Product_1         50 - 100 Million    Category 4

     Geography    Location POC_name             Designation  \
744      India   Hyderabad    pooja           CEO/President
1095     India  Ahmadnagar    pooja  CEO/Chairman/President

           Hiring_candidate_role      Lead_source Level_of_meeting  \
744   Building services engineer  Marketing Event          Level 2
1095       Geneticist, molecular  Marketing Event          Level 2
```

```
          Last_lead_update     Internal_POC     Resource  Internal_rating
744            2 days back  Jones,Michael L  Deliverable              1.0
1095  More than a week back  Kiepea,Prince A  Deliverable              4.0
```

[113]:
```
train_x.Last_lead_update.value_counts(dropna=False)
```

[113]:
```
NaN                                   633
Following up but lead not responding  627
Up-to-date                            623
more than a month                     605
No track                              601
?                                     585
5 days back                           564
2 days back                           559
More than 2 weeks                     556
More than a week back                 554
Did not hear back after Level 1       553
Pending                               547
Name: Last_lead_update, dtype: int64
```

[114]:
```
test.Last_lead_update.value_counts(dropna=False)
```

[114]:
```
2 days back                           204
?                                     190
More than 2 weeks                     186
5 days back                           186
more than a month                     180
Up-to-date                            176
Did not hear back after Level 1       173
NaN                                   168
No track                              165
Pending                               156
Following up but lead not responding  155
More than a week back                 154
Name: Last_lead_update, dtype: int64
```

[115]:
```
train_x[train_x.Last_lead_update.isna()].head(10)
```

[115]:
```
                    Industry  Deal_value  Weighted_amount Date_of_creation  \
12         Biotech/Healthcare    222331.0     1.400685e+06       2020-03-27
35          Constr - Supplies    216892.0     1.344730e+06       2019-07-30
48                      Banks    376485.0     2.560098e+06       2020-05-02
51               Associations    457274.0     2.972281e+06       2019-10-25
65                  Insurance     58778.0     4.055682e+05       2020-02-17
82                Restaurants     58529.0     3.804385e+05       2020-07-04
115  Architecture/Engineering    150122.0     9.757930e+05       2020-02-25
```

```
121        Financial Services     39425.0     2.288962e+05       2019-08-21
123          Telecom Hardware    214941.0     1.311140e+06       2021-01-19
125                     Banks    258822.0     1.734107e+06       2019-04-26

        Pitch             Lead_revenue Fund_category Geography  \
12   Product_2        100 - 500 Million    Category 2       USA
35   Product_1        100 - 500 Million    Category 3       USA
48   Product_2         50 - 100 Million    Category 1       USA
51   Product_2  500 Million - 1 Billion    Category 4     India
65   Product_2  500 Million - 1 Billion    Category 4     India
82   Product_2         50 - 100 Million    Category 1     India
115  Product_2  500 Million - 1 Billion    Category 4       USA
121  Product_2         50 - 100 Million    Category 4     India
123  Product_2  500 Million - 1 Billion    Category 1     India
125  Product_2  500 Million - 1 Billion    Category 1       USA

                            Location           POC_name  \
12                       Danville, IL          Frank Ali
35                         Monroe, MI  Anthony Armstrong
48   Kingsport-Bristol-Bristol, TN-VA         Amy Stuart
51                         Saharanpur       puja chauhan
65                 Tiruchchirappalli             kushum
82                       Muzaffarpur              mamta
115                Oklahoma City, OK    Deanna Schaefer
121                           Panaji             sakshi
123                         Itanagar      kajal@ poonam
125                         Ocala, FL     Patrick Jacobs

                                        Designation  \
12   Vice President / GM (04-present) : VP Sales an...
35                             CEO/Chairman/President
48   Vice President / GM (04-present) : VP Sales an...
51                           Executive Vice President
65                             CEO/Chairman/President
82                            CEO/Co-Founder/Chairman
115                                    CEO/President
121                              SVP/General Counsel
123                            CEO/Chairman/President
125                                    CEO/President

                    Hiring_candidate_role      Lead_source Level_of_meeting  \
12           Speech and language therapist  Marketing Event          Level 1
35                 Housing manager/officer           Others          Level 2
48                Secondary school teacher    Contact Email          Level 3
51              Psychologist, occupational  Marketing Event          Level 2
65                   Accountant, chartered          Website          Level 3
82   English as a foreign language teacher    Contact Email          Level 3
```

```
115            Industrial/product designer  Marketing Event          Level 1
121            Engineer, building services  Marketing Event          Level 2
123                   Scientist, forensic           Others          Level 2
125                 Agricultural engineer    Contact Email          Level 2

    Last_lead_update         Internal_POC        Resource  Internal_rating
12               NaN         Maine,John P  Cannot deliver                5
35               NaN        Davis,Brian R  Cannot deliver                5
48               NaN  McKenstry,Loretta A              No                3
51               NaN       Smith,Keenan H             Yes                2
65               NaN       Knox,Antonio D             Yes                4
82               NaN     Massiah,Gerard F              No                3
115              NaN     Houston,Arnold E              No                1
121              NaN  Anthony,Katherine D             Yes                4
123              NaN        Gilley,Janine             Yes                1
125              NaN     Salyers,Daniel L      Not enough                4
```

[116]:
```
tr_lv1 = train_x[train_x.Level_of_meeting == 'Level 1']
```

[117]:
```
tr_lv1.Last_lead_update.value_counts(dropna=False)
```

[117]:
```
Following up but lead not responding    222
?                                       219
No track                                217
NaN                                     213
Up-to-date                              206
5 days back                             206
More than 2 weeks                       200
more than a month                       199
Pending                                 189
2 days back                             182
Did not hear back after Level 1         179
More than a week back                   177
Name: Last_lead_update, dtype: int64
```

Due to a certain level of uncertainty, I will attempt to replace the missing values in the Last lead update column with 'unknown'.

[118]:
```
train_x.Last_lead_update.fillna('Unknown', inplace=True)
test.Last_lead_update.fillna('Unknown', inplace=True)
```

[119]:
```
train_x.Resource.value_counts(dropna=False)
```

[119]:
```
No                            1179
We have all the requirements  1160
Not enough                    1145
Deliverable                   1131
Yes                           1130
```

```
Cannot deliver                  1113
NaN                              149
Name: Resource, dtype: int64
```

[120]:
```
tr_nrs_id = train_x[train_x.Resource.isna()].index
```

[121]:
```
test.Resource.value_counts(dropna=False)
```

[121]:
```
Cannot deliver                  384
No                              365
Not enough                      346
We have all the requirements    345
Yes                             332
Deliverable                     307
NaN                              14
Name: Resource, dtype: int64
```

I don't observe any apparent connection between the Resource feature and other attributes, so I will simply fill it using backward fill (bfill) method.

[122]:
```
train_x.Resource.fillna(method='bfill', inplace=True)
test.Resource.fillna(method='bfill', inplace=True)
```

[123]:
```
train_x.isna().sum()
```

[123]:
```
Industry                0
Deal_value              0
Weighted_amount         0
Date_of_creation        0
Pitch                   0
Lead_revenue            0
Fund_category           0
Geography               0
Location                0
POC_name                0
Designation             0
Hiring_candidate_role   0
Lead_source             0
Level_of_meeting        0
Last_lead_update        0
Internal_POC            0
Resource                0
Internal_rating         0
dtype: int64
```

[124]:
```
test.isna().sum()
```

[124]:
```
Industry                0
Deal_value              0
Weighted_amount         0
Date_of_creation        0
Pitch                   0
Lead_revenue            0
Fund_category           0
Geography               0
Location                0
POC_name                0
Designation             0
Hiring_candidate_role   0
Lead_source             0
Level_of_meeting        0
Last_lead_update        0
Internal_POC            0
Resource                0
Internal_rating         0
dtype: int64
```

[125]:
```
train_x.nunique()
```

[125]:
```
Industry                171
Deal_value             6955
Weighted_amount        7002
Date_of_creation        777
Pitch                     2
Lead_revenue              3
Fund_category             4
Geography                 2
Location                597
POC_name               5268
Designation              10
Hiring_candidate_role   639
Lead_source               4
Level_of_meeting          3
Last_lead_update         12
Internal_POC             60
Resource                  6
Internal_rating           5
dtype: int64
```

[126]:
```
test.nunique()
```

[126]:
```
Industry                138
Deal_value             2088
Weighted_amount        2034
```

```
Date_of_creation          720
Pitch                       2
Lead_revenue                3
Fund_category               4
Geography                   2
Location                  565
POC_name                 1745
Designation                10
Hiring_candidate_role     618
Lead_source                 4
Level_of_meeting            3
Last_lead_update           12
Internal_POC               60
Resource                    6
Internal_rating             5
dtype: int64
```

[127]:
```
train_x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7007 entries, 0 to 7006
Data columns (total 18 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Industry               7007 non-null   object
 1   Deal_value             7007 non-null   float64
 2   Weighted_amount        7007 non-null   float64
 3   Date_of_creation       7007 non-null   object
 4   Pitch                  7007 non-null   object
 5   Lead_revenue           7007 non-null   object
 6   Fund_category          7007 non-null   object
 7   Geography              7007 non-null   object
 8   Location               7007 non-null   object
 9   POC_name               7007 non-null   object
 10  Designation            7007 non-null   object
 11  Hiring_candidate_role  7007 non-null   object
 12  Lead_source            7007 non-null   object
 13  Level_of_meeting       7007 non-null   object
 14  Last_lead_update       7007 non-null   object
 15  Internal_POC           7007 non-null   object
 16  Resource               7007 non-null   object
 17  Internal_rating        7007 non-null   int64
dtypes: float64(2), int64(1), object(15)
memory usage: 985.5+ KB
```

[128]:
```
train_x.head(10)
```

[128]:

	Industry	Deal_value	Weighted_amount	Date_of_creation
0	Restaurants	320506.0	2.067264e+06	2020-03-29
1	Construction Services	39488.0	2.408768e+05	2019-07-10
2	Hospitals/Clinics	359392.0	2.407926e+06	2019-07-27
3	Real Estate	76774.0	4.683214e+05	2021-01-30
4	Financial Services	483896.0	2.809435e+06	2019-05-22
5	Banks	418674.0	2.637646e+06	2019-06-30
6	Banks	384356.0	2.709710e+06	2019-11-20
7	Architecture/Engineering	245205.0	1.642874e+06	2020-12-22
8	Education/Training	343280.0	1.991024e+06	2019-01-08
9	REIT	293611.0	1.791027e+06	2020-07-04

	Pitch	Lead_revenue	Fund_category	Geography
0	Product_2	50 - 100 Million	Category 2	USA
1	Product_2	500 Million - 1 Billion	Category 4	India
2	Product_1	500 Million - 1 Billion	Category 4	USA
3	Product_2	500 Million - 1 Billion	Category 3	USA
4	Product_2	50 - 100 Million	Category 3	India
5	Product_1	50 - 100 Million	Category 2	India
6	Product_2	500 Million - 1 Billion	Category 1	USA
7	Product_2	500 Million - 1 Billion	Category 1	USA
8	Product_1	100 - 500 Million	Category 3	USA
9	Product_1	500 Million - 1 Billion	Category 3	USA

	Location	POC_name
0	Killeen-Temple, TX	Charlene Werner
1	Ratlam	rakhi
2	Albany-Schenectady-Troy, NY	Ariel Hamilton
3	Mount Vernon-Anacortes, WA	Erin Wilson
4	Shimoga	kavita
5	Bulandshahr	kamla devi
6	Salisbury, MD-DE	Sara Dixon
7	Jacksonville, FL	Christian Stanley
8	Seattle-Tacoma-Bellevue, WA	Joseph Thompson
9	Clarksville, TN-KY	William Grimes

	Designation
0	Executive Vice President
1	Chairman/CEO/President
2	SVP/General Counsel
3	CEO/Co-Founder/Chairman
4	Executive Vice President
5	CEO/Co-Founder/Chairman
6	CEO/Co-Founder/Chairman
7	Vice President / GM (04-present) : VP Sales an...
8	Chairman/Chief Innovation Officer
9	CEO

```
                  Hiring_candidate_role      Lead_source Level_of_meeting  \
0                   Community pharmacist          Website          Level 3
1                 Recruitment consultant           Others          Level 1
2                 Health service manager  Marketing Event          Level 1
3         Therapist, speech and language    Contact Email          Level 2
4                           Media planner          Website          Level 2
5                         Microbiologist  Marketing Event          Level 3
6                           Cartographer    Contact Email          Level 3
7              Engineer, maintenance (IT)           Others          Level 2
8  Product/process development scientist    Contact Email          Level 1
9                  Engineering geologist    Contact Email          Level 1

                 Last_lead_update               Internal_POC  \
0                        No track           Davis,Sharrice A
1  Did not hear back after Level 1             Brown,Maxine A
2                                ?  Georgakopoulos,Vasilios T
3  Did not hear back after Level 1             Brown,Maxine A
4                      Up-to-date              Thomas,Lori E
5                     2 days back      Featherstone,Adrian R
6                More than 2 weeks             Booker,David L
7                     5 days back              Cashin,Marc C
8                        No track          Van Arter,Derrick
9                more than a month              Hanyok,John J

                      Resource  Internal_rating
0                           No               3
1                           No               5
2                           No               4
3  We have all the requirements               1
4                           No               4
5                   Deliverable               3
6                           No               4
7                           No               5
8  We have all the requirements               2
9                   Deliverable               4
```

[129]:
```
date_col = ['Creation_date']
num_col = [x for x in train_x.columns if (train_x[x].dtype == 'float64') |␣
↪(train_x[x].dtype == 'int64')]
cat_col = [x for x in train_x.columns if (train_x[x].dtype == 'O')]
```

[130]:
```
from sklearn.preprocessing import LabelEncoder
lbl = LabelEncoder()
lbl2 = LabelEncoder()
train_x.Lead_revenue = lbl.fit_transform(train_x.Lead_revenue)
train_x.Level_of_meeting = lbl2.fit_transform(train_x.Level_of_meeting)
```

[131]:
```
test.Lead_revenue = lbl.transform(test.Lead_revenue)
test.Level_of_meeting = lbl2.transform(test.Level_of_meeting)
```

[132]:
```
num_col = [x for x in train_x.columns if (train_x[x].dtype == 'float64') |␣
↪(train_x[x].dtype == 'int32')]
```

[133]:
```
num_col
```

[133]:
```
['Deal_value', 'Weighted_amount', 'Lead_revenue', 'Level_of_meeting']
```

[134]:
```
cat_col = [x for x in train_x.columns if train_x[x].dtype == 'O']
```

[135]:
```
cat_col
```

[135]:
```
['Industry',
 'Date_of_creation',
 'Pitch',
 'Fund_category',
 'Geography',
 'Location',
 'POC_name',
 'Designation',
 'Hiring_candidate_role',
 'Lead_source',
 'Last_lead_update',
 'Internal_POC',
 'Resource']
```

[136]:
```
low_car = [x for x in cat_col if train_x[x].nunique() <= 12]
```

[137]:
```
low_car
```

[137]:
```
['Pitch',
 'Fund_category',
 'Geography',
 'Designation',
 'Lead_source',
 'Last_lead_update',
 'Resource']
```

[138]:
```
from sklearn.preprocessing import OneHotEncoder, StandardScaler, LabelEncoder
from sklearn.compose import make_column_transformer, ColumnTransformer
from sklearn.pipeline import Pipeline
num_transformer = Pipeline(steps=[
    ('scaler', StandardScaler(with_mean=False))])
cat_transformer = Pipeline(steps=[
    ('ohe', OneHotEncoder(handle_unknown='ignore')),
    ('scaler', StandardScaler(with_mean=False))
```

```
])


preprocessor = ColumnTransformer(
    transformers=[
        ('num', num_transformer, num_col),
        ('cat', cat_transformer, low_car)])
```

First Model

[139]:
```
from sklearn.linear_model import LinearRegression
model1 = Pipeline(steps=[('pre', preprocessor),
                         ('reg', LinearRegression())])
```

[140]:
```
train_x1 = train_x[num_col + low_car]
test_1 = test[num_col + low_car]
```

[141]:
```
train_x1
```

[141]:

	Deal_value	Weighted_amount	Lead_revenue	Level_of_meeting	Pitch	\
0	320506.0	2.067264e+06	1	2	Product_2	
1	39488.0	2.408768e+05	2	0	Product_2	
2	359392.0	2.407926e+06	2	0	Product_1	
3	76774.0	4.683214e+05	2	1	Product_2	
4	483896.0	2.809435e+06	1	1	Product_2	
...	...	...	...	...	...	
7002	192800.0	1.195360e+06	0	0	Product_1	
7003	220208.0	1.453373e+06	0	1	Product_2	
7004	253608.0	1.472414e+06	0	2	Product_1	
7005	118615.0	7.947205e+05	2	2	Product_1	
7006	258627.0	1.642281e+06	2	2	Product_2	

	Fund_category	Geography	\
0	Category 2	USA	
1	Category 4	India	
2	Category 4	USA	
3	Category 3	USA	
4	Category 3	India	
...	...	...	
7002	Category 4	India	
7003	Category 1	India	
7004	Category 2	USA	
7005	Category 2	USA	
7006	Category 2	USA	

	Designation	Lead_source	\
0	Executive Vice President	Website	
1	Chairman/CEO/President	Others	

```
2                                SVP/General Counsel   Marketing Event
3                              CEO/Co-Founder/Chairman    Contact Email
4                             Executive Vice President          Website
...                                                ...              ...
7002                           CEO/Co-Founder/Chairman    Contact Email
7003                                               CEO  Marketing Event
7004                             SVP/General Counsel    Marketing Event
7005                          Executive Vice President          Website
7006  Vice President / GM (04-present) : VP Sales an...    Contact Email

                    Last_lead_update                       Resource
0                           No track                             No
1     Did not hear back after Level 1                            No
2                                   ?                            No
3     Did not hear back after Level 1  We have all the requirements
4                          Up-to-date                            No
...                               ...                          ...
7002           More than a week back  We have all the requirements
7003                               ?  We have all the requirements
7004                         Unknown  We have all the requirements
7005                         Unknown  We have all the requirements
7006           More than a week back  We have all the requirements

[7007 rows x 11 columns]
```

[142]:
```
model1.fit(train_x1, train_y)
```

[142]:
```
Pipeline(steps=[('pre',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Deal_value',
                                                   'Weighted_amount',
                                                   'Lead_revenue',
                                                   'Level_of_meeting']),
                                                 ('cat',
                                                  Pipeline(steps=[('ohe',
OneHotEncoder(handle_unknown='ignore')),
                                                                  ('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Pitch', 'Fund_category',
                                                   'Geography', 'Designation',
                                                   'Lead_source',
                                                   'Last_lead_update',
                                                   'Resource'])])),
                ('reg', LinearRegression())])
```

[143]:
```
tr_pred = model1.predict(train_x1)
```

[144]:
```
tr_pred = [ "{:0.1f}".format(x) for x in tr_pred ]
```

[145]:
```
tr_y = train.Success_probability.values
```

[146]:
```
from sklearn.metrics import r2_score
import numpy as np

train_y_numeric = np.array(train_y, dtype=float)
tr_pred_numeric = np.array(tr_pred, dtype=float)
r2 = r2_score(train_y_numeric, tr_pred_numeric)
print(r2)
```

```
0.06345011538112089
```

[147]:
```
from sklearn.metrics import mean_absolute_error
import numpy as np

train_y_numeric = np.array(train_y, dtype=float)
tr_pred_numeric = np.array(tr_pred, dtype=float)
mae = mean_absolute_error(train_y_numeric, tr_pred_numeric)
print(mae)
```

```
7.7442771514200075
```

[148]:
```
pred_1 = model1.predict(test_1)
```

[149]:
```
pred_1 = [ "{:0.1f}".format(x) for x in pred_1 ]
```

[150]:
```
make_csv(pred_1, 'Output_1.csv')
```

[151]:
```
pd.read_csv('Output_1.csv')
```

[151]:

	Deal_title	Success_probability
0	TitleAD16O	60.6
1	TitleOW6CR	67.2
2	TitleVVJQ5	71.2
3	TitleUS8NA	63.0
4	Title5VGWW	68.7
...	...	...
2088	Title2R8VU	70.1
2089	Title7HCNJ	67.5
2090	TitleCD5YZ	58.9
2091	Title8OKXL	61.6
2092	TitleHFQT8	60.6

[2093 rows x 2 columns]

[152]:
```
sample
```

[152]:

	Deal_title	Success_probability
0	TitleM5DZY	48.6
1	TitleKIW18	33.9
2	TitleFXSDN	43.8
3	TitlePSK4Y	39.5
4	Title904GV	37.4

Second Model

[153]:
```
from sklearn.ensemble import RandomForestRegressor
model2 = Pipeline(steps=[('pre', preprocessor),
                         ('rfreg', RandomForestRegressor(n_estimators=100,␣
 ↪random_state=42))])
```

[154]:
```
model2.fit(train_x1, train_y)
```

[154]:
```
Pipeline(steps=[('pre',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Deal_value',
                                                   'Weighted_amount',
                                                   'Lead_revenue',
                                                   'Level_of_meeting']),
                                                 ('cat',
                                                  Pipeline(steps=[('ohe',
OneHotEncoder(handle_unknown='ignore')),
                                                                  ('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Pitch', 'Fund_category',
                                                   'Geography', 'Designation',
                                                   'Lead_source',
                                                   'Last_lead_update',
                                                   'Resource'])])),
                ('rfreg', RandomForestRegressor(random_state=42))])
```

[155]:
```
import joblib as jb
mod2_name = 'Model_2.sav'
jb.dump(model2, mod2_name)
model2 = jb.load(mod2_name)
```

[156]:
```
pred_2 = model2.predict(test_1)
```

[157]:
```
pred_2 = [ "{:0.1f}".format(x) for x in pred_2 ]
```

[158]:
```
make_csv(pred_2, 'Output_2.csv')
```

[159]:
```
pd.read_csv('Output_2.csv')
```

[159]:

	Deal_title	Success_probability
0	TitleAD16O	63.8
1	TitleOW6CR	65.7
2	TitleVVJQ5	70.5
3	TitleUS8NA	67.1
4	Title5VGWW	69.2
...	...	...
2088	Title2R8VU	61.2
2089	Title7HCNJ	70.1
2090	TitleCD5YZ	56.4
2091	Title8OKXL	58.6
2092	TitleHFQT8	56.2

[2093 rows x 2 columns]

Third Model

[160]:
```
model3 = Pipeline(steps=[('pre', preprocessor),
                         ('rfreg', RandomForestRegressor(max_depth=10,␣
↪n_estimators=50, random_state=42))])
```

[161]:
```
model3.fit(train_x1, train_y)
```

[161]:
```
Pipeline(steps=[('pre',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Deal_value',
                                                   'Weighted_amount',
                                                   'Lead_revenue',
                                                   'Level_of_meeting']),
                                                 ('cat',
                                                  Pipeline(steps=[('ohe',
OneHotEncoder(handle_unknown='ignore')),
                                                                  ('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Pitch', 'Fund_category',
                                                   'Geography', 'Designation',
                                                   'Lead_source',
                                                   'Last_lead_update',
                                                   'Resource'])])),
                ('rfreg',
                 RandomForestRegressor(max_depth=10, n_estimators=50,
                                       random_state=42))])
```

[162]:
```
mod3_name = 'Model_3.sav'
jb.dump(model3, mod3_name)
model3 = jb.load(mod3_name)
```

[163]:
```
pred_3 = model3.predict(test_1)
pred_3 = [ "{:0.1f}".format(x) for x in pred_3 ]
```

[164]:
```
make_csv(pred_3, 'Output_3.csv')
```

[165]:
```
pd.read_csv('Output_3.csv')
```

[165]:

	Deal_title	Success_probability
0	TitleAD16O	59.6
1	TitleOW6CR	67.6
2	TitleVVJQ5	71.7
3	TitleUS8NA	65.2
4	Title5VGWW	67.9
...	...	...
2088	Title2R8VU	62.0
2089	Title7HCNJ	71.6
2090	TitleCD5YZ	58.0
2091	Title8OKXL	58.4
2092	TitleHFQT8	59.5

[2093 rows x 2 columns]

Fourth Model

[166]:
```
from sklearn.svm import LinearSVR

model4 = Pipeline(steps=[('pre', preprocessor),
                         ('l_svr', LinearSVR(random_state=0))])
```

[167]:
```
model4.fit(train_x1, train_y)
```

[167]:
```
Pipeline(steps=[('pre',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('scaler',
StandardScaler(with_mean=False))]),
                                                  ['Deal_value',
                                                   'Weighted_amount',
                                                   'Lead_revenue',
                                                   'Level_of_meeting']),
                                                 ('cat',
                                                  Pipeline(steps=[('ohe',
OneHotEncoder(handle_unknown='ignore')),
                                                                  ('scaler',
StandardScaler(with_mean=False))]),
```

```
                                                    ['Pitch', 'Fund_category',
                                                     'Geography', 'Designation',
                                                     'Lead_source',
                                                     'Last_lead_update',
                                                     'Resource'])])),
                    ('l_svr', LinearSVR(random_state=0))])
```

[168]:
```
pred_4 = model4.predict(test_1)
```

[169]:
```
pred_4 = [ "{:0.1f}".format(x) for x in pred_4 ]
```

[170]:
```
make_csv(pred_4, 'Output_4.csv')
```

[171]:
```
pd.read_csv('Output_4.csv')
```

[171]:

	Deal_title	Success_probability
0	TitleAD16O	62.5
1	TitleOW6CR	68.3
2	TitleVVJQ5	74.8
3	TitleUS8NA	63.9
4	Title5VGWW	70.2
...	...	...
2088	Title2R8VU	71.6
2089	Title7HCNJ	67.8
2090	TitleCD5YZ	60.5
2091	Title8OKXL	62.6
2092	TitleHFQT8	60.6

[2093 rows x 2 columns]