In [95]:
```
import warnings
warnings.filterwarnings('ignore')
from operator import itemgetter
import pandas as pd #dataframe
import numpy as np #mathematical computations
import matplotlib.pyplot as plt #visualization
import matplotlib
import seaborn as sns #visualization
import json #exporting columns
import pickle #saving the model
from sklearn.linear_model import LinearRegression #Linear Regression
from sklearn.linear_model import Lasso #Lasso Regression
from sklearn.tree import DecisionTreeRegressor #Decision Tree Regression
from sklearn.ensemble import RandomForestRegressor #Random Forest Regression
from sklearn.model_selection import train_test_split #Splitting the dataset
from sklearn.model_selection import ShuffleSplit #Random shuffling
from sklearn.model_selection import cross_val_score #Score cross validation
from sklearn.model_selection import GridSearchCV #Hyper parameter tuning
from warnings import simplefilter #Filtering warnings
import seaborn as sns
import missingno as msno
import statsmodels.api as sm
from datetime import datetime
from scipy import stats
```

Observe the data

Import the data set and show the title

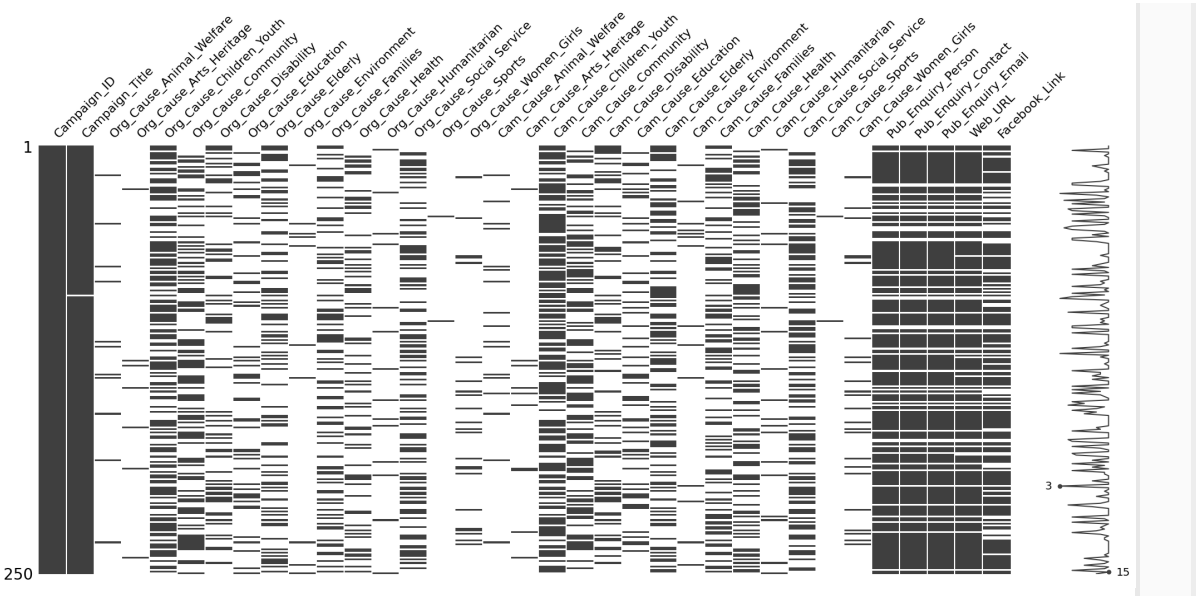
In [96]:
```
Orginal_data = pd.read_csv('./Combined.csv',encoding = "ISO-8859-1")
Causes_data = pd.read_csv('./Causes.csv',encoding = "ISO-8859-1")
```

In [97]:
```
Orginal_data.columns
```

Out[97]:
```
Index(['Campaign_ID', 'Campagin_Title ', 'Receiving_NPO_name ',
       'Receiving_NPO_Id', 'NPO_Status_orignal', 'NPO_Status',
       'Number_campaigns_NPO', 'Public_Campaign_Access', 'Creator_Type',
       'Creator_Id', 'Campaign_Status', 'Actual_Donation_Amount',
       'Distinct_Donors', 'Campaign_Goal', 'Campaign_Completion_Rate',
       'Days_Left_for_Campaign', 'Campaign_Start_Date', 'Campaign_End_Dat
e',
       'NPO_Tax_Deductibility', 'Campaign_Image1', 'Campaign_Image2',
       'Campaign_Image3', 'Campaign_Image4', 'Campaign_Image5',
       'Campaign_Video', 'Impact_Message1', 'Impact_Message2',
       'Impact_Message3', 'Impact_Message4', 'Impact_Message5',
       'Custom_Amount1', 'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount
4',
       'Description_Campaign', 'Description_NPO'],
      dtype='object')
```

In [98]:
```
msno.matrix(Causes_data.sample(250))
```

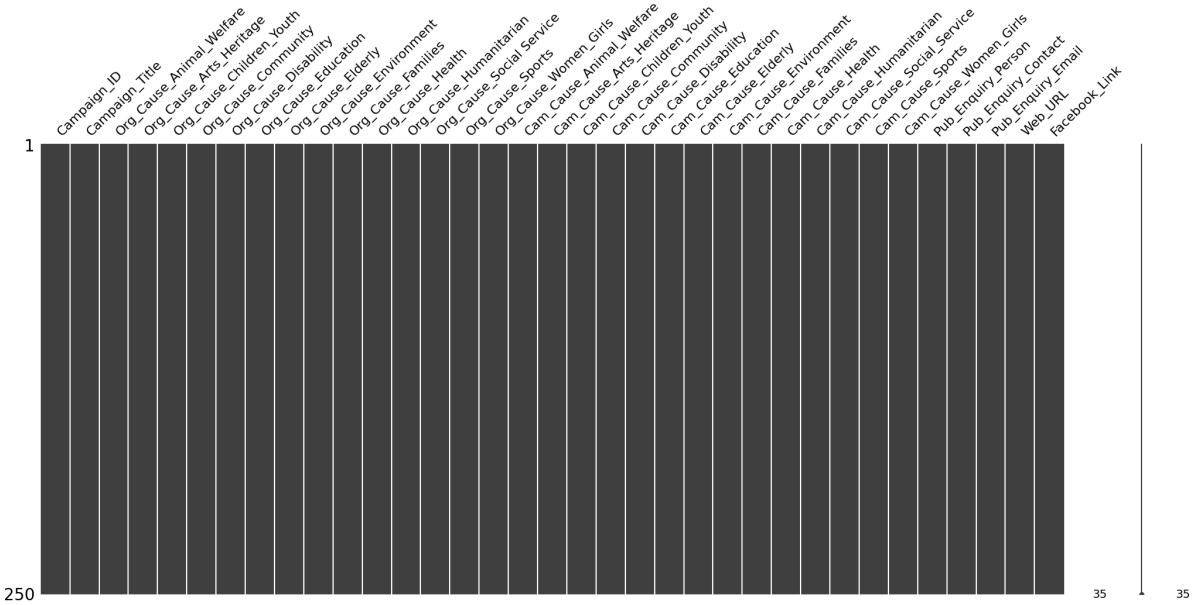
Out[98]:
```
<AxesSubplot:>
```



In [99]:
```
Causes_data= Causes_data.fillna(0)
```

In [100…
```
msno.matrix(Causes_data.sample(250))
```

Out[100]: <AxesSubplot:>



In [101…
```
combined_data = pd.merge(Orginal_data, Causes_data, how='left', on=['Campaig
```

In [102…
```
Total_Rows = combined_data.shape[0]
print(Total_Rows)
```

```
15979
```

In [103…
```
print(combined_data.columns)
```

```
Index(['Campaign_ID', 'Campagin_Title ', 'Receiving_NPO_name ',
       'Receiving_NPO_Id', 'NPO_Status_orignal', 'NPO_Status',
       'Number_campaigns_NPO', 'Public_Campaign_Access', 'Creator_Type',
       'Creator_Id', 'Campaign_Status', 'Actual_Donation_Amount',
       'Distinct_Donors', 'Campaign_Goal', 'Campaign_Completion_Rate',
       'Days_Left_for_Campaign', 'Campaign_Start_Date', 'Campaign_End_Dat
e',
       'NPO_Tax_Deductibility', 'Campaign_Image1', 'Campaign_Image2',
       'Campaign_Image3', 'Campaign_Image4', 'Campaign_Image5',
       'Campaign_Video', 'Impact_Message1', 'Impact_Message2',
       'Impact_Message3', 'Impact_Message4', 'Impact_Message5',
       'Custom_Amount1', 'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount
4',
       'Description_Campaign', 'Description_NPO', 'Campaign_Title',
       'Org_Cause_Animal_Welfare', 'Org_Cause_Arts_Heritage',
       'Org_Cause_Children_Youth', 'Org_Cause_Community',
       'Org_Cause_Disability', 'Org_Cause_Education', 'Org_Cause_Elderly',
       'Org_Cause_Environment', 'Org_Cause_Families', 'Org_Cause_Health',
       'Org_Cause_Humanitarian', 'Org_Cause_Social Service',
       'Org_Cause_Sports', 'Org_Cause_Women_Girls', 'Cam_Cause_Animal_Welfa
re',
       'Cam_Cause_Arts_Heritage', 'Cam_Cause_Children_Youth',
       'Cam_Cause_Community', 'Cam_Cause_Disability', 'Cam_Cause_Educatio
n',
       'Cam_Cause_Elderly', 'Cam_Cause_Environment', 'Cam_Cause_Families',
       'Cam_Cause_Health', 'Cam_Cause_Humanitarian',
       'Cam_Cause_Social_Service', 'Cam_Cause_Sports', 'Cam_Cause_Women_Gir
ls',
       'Pub_Enquiry_Person', 'Pub_Enquiry_Contact', 'Pub_Enquiry_Email',
       'Web_URL', 'Facebook_Link'],
      dtype='object')
```

I found there is no "Organizational Causes" and "Campaign Causes" in this data set.

Here are all variables I plan to operate, ignnore other columns temporarily

In [104…

```
Need_variable = ["Actual_Donation_Amount","NPO_Tax_Deductibility", "Distinct
"Campaign_Goal","Campaign_Start_Date", "Campaign_End_Date",
"Campaign_Image1", "Campaign_Image2","Campaign_Image3",
"Campaign_Image4","Campaign_Image5","Campaign_Video",
"Impact_Message1","Impact_Message2","Impact_Message3","Impact_Message4",
"Impact_Message5","Custom_Amount1","Custom_Amount2","Custom_Amount3",
"Custom_Amount4","Description_Campaign","Description_NPO",
'Org_Cause_Animal_Welfare', 'Org_Cause_Arts_Heritage',
'Org_Cause_Children_Youth', 'Org_Cause_Community',
'Org_Cause_Disability', 'Org_Cause_Education', 'Org_Cause_Elderly',
'Org_Cause_Environment', 'Org_Cause_Families', 'Org_Cause_Health',
'Org_Cause_Humanitarian', 'Org_Cause_Social Service',
'Org_Cause_Sports', 'Org_Cause_Women_Girls', 'Cam_Cause_Animal_Welfare',
'Cam_Cause_Arts_Heritage', 'Cam_Cause_Children_Youth',
'Cam_Cause_Community', 'Cam_Cause_Disability', 'Cam_Cause_Education',
'Cam_Cause_Elderly', 'Cam_Cause_Environment', 'Cam_Cause_Families',
'Cam_Cause_Health', 'Cam_Cause_Humanitarian',
'Cam_Cause_Social_Service', 'Cam_Cause_Sports', 'Cam_Cause_Women_Girls'
]
```

```
extract_data = combined_data[Need_variable]
extract_data
```

Out[104]:

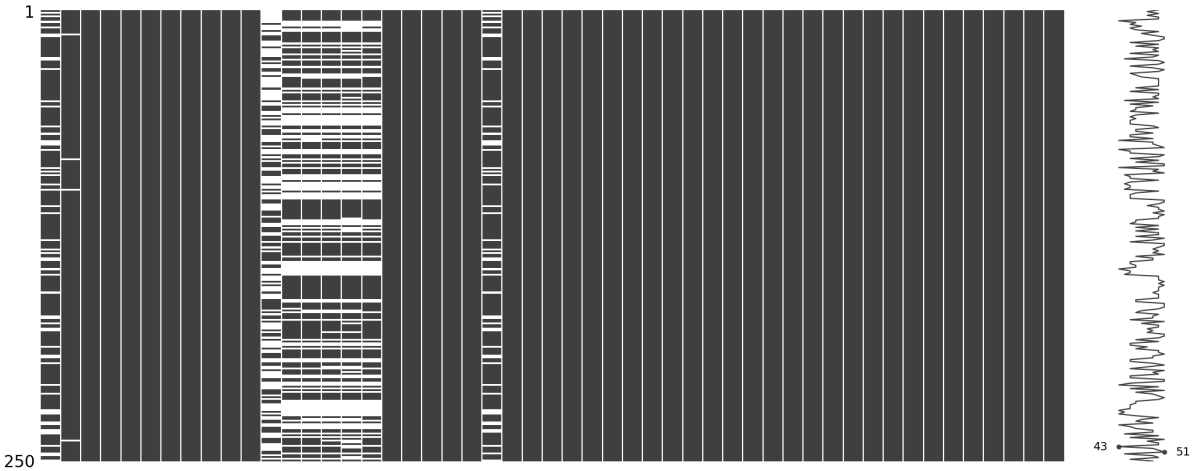
	Actual_Donation_Amount	NPO_Tax_Deductibility	Distinct_Donors	Campaign_Goal
0	5561.0	True	66	50000
1	2810.0	True	32	20000
2	1118.0	True	22	30000
3	2800.0	True	7	2000
4	2030.0	True	27	5000
...	...	...	...	...
15974	10.0	True	1	5000
15975	150.0	True	4	10000
15976	1000.0	True	10	1000
15977	120.0	True	2	3000
15978	120.0	True	2	40000

15979 rows × 51 columns

In [105…

```
msno.matrix(extract_data.sample(250))
```

Out[105]: <AxesSubplot:>



We can see that "Actual_Donation_Amount" "Campaign_Video" "Impact_Message1" "Impact_Message2" "Impact_Message3" "Impact_Message4" and "Impact_Message5" are many missing data, fill them first so that it's more convenient to operate. "NPO_Tax_Deductibility" has been ignore temporarily just like you said in email

In [106…
```
extract_data['NPO_Tax_Deductibility'] = extract_data['NPO_Tax_Deductibility'
extract_data['Actual_Donation_Amount'] = extract_data['Actual_Donation_Amoun
extract_data['Actual_Donation_Amount'] = pd.to_numeric( extract_data['Actual
extract_data['Distinct_Donors'] = extract_data['Distinct_Donors'].fillna('0'
extract_data['Distinct_Donors'] = pd.to_numeric( extract_data['Distinct_Donc
extract_data['Campaign_Video'] = extract_data['Campaign_Video'].fillna('0')
extract_data['Impact_Message1'] = extract_data['Impact_Message1'].fillna('0'
extract_data['Impact_Message2'] = extract_data['Impact_Message2'].fillna('0'
extract_data['Impact_Message3'] = extract_data['Impact_Message3'].fillna('0'
extract_data['Impact_Message4'] = extract_data['Impact_Message4'].fillna('0'
extract_data['Impact_Message5'] = extract_data['Impact_Message5'].fillna('0'
```

In [107…
```
extract_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15979 entries, 0 to 15978
Data columns (total 51 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Actual_Donation_Amount    15979 non-null  float64
 1   NPO_Tax_Deductibility     15979 non-null  object 
 2   Distinct_Donors           15979 non-null  int64  
 3   Campaign_Goal             15979 non-null  int64  
 4   Campaign_Start_Date       15979 non-null  object 
 5   Campaign_End_Date         15979 non-null  object 
 6   Campaign_Image1           15979 non-null  int64  
 7   Campaign_Image2           15979 non-null  int64  
 8   Campaign_Image3           15979 non-null  int64  
 9   Campaign_Image4           15979 non-null  int64  
 10  Campaign_Image5           15979 non-null  int64  
 11  Campaign_Video            15979 non-null  object 
 12  Impact_Message1           15979 non-null  object 
 13  Impact_Message2           15979 non-null  object 
 14  Impact_Message3           15979 non-null  object 
 15  Impact_Message4           15979 non-null  object 
 16  Impact_Message5           15979 non-null  object 
 17  Custom_Amount1            15979 non-null  int64  
 18  Custom_Amount2            15979 non-null  int64  
 19  Custom_Amount3            15979 non-null  int64  
 20  Custom_Amount4            15979 non-null  int64  
 21  Description_Campaign      15971 non-null  object 
 22  Description_NPO           13270 non-null  object 
 23  Org_Cause_Animal_Welfare  15979 non-null  object 
 24  Org_Cause_Arts_Heritage   15979 non-null  object 
 25  Org_Cause_Children_Youth  15979 non-null  object 
 26  Org_Cause_Community       15979 non-null  object 
 27  Org_Cause_Disability      15979 non-null  object 
 28  Org_Cause_Education       15979 non-null  object 
 29  Org_Cause_Elderly         15979 non-null  object 
 30  Org_Cause_Environment     15979 non-null  object 
 31  Org_Cause_Families        15979 non-null  object 
 32  Org_Cause_Health          15979 non-null  object 
 33  Org_Cause_Humanitarian    15979 non-null  object 
 34  Org_Cause_Social Service  15979 non-null  object 
 35  Org_Cause_Sports          15979 non-null  object 
 36  Org_Cause_Women_Girls     15979 non-null  object 
 37  Cam_Cause_Animal_Welfare  15979 non-null  object 
 38  Cam_Cause_Arts_Heritage   15979 non-null  object 
 39  Cam_Cause_Children_Youth  15979 non-null  object 
 40  Cam_Cause_Community       15979 non-null  object 
 41  Cam_Cause_Disability      15979 non-null  object 
 42  Cam_Cause_Education       15979 non-null  object 
 43  Cam_Cause_Elderly         15979 non-null  object 
 44  Cam_Cause_Environment     15979 non-null  object 
 45  Cam_Cause_Families        15979 non-null  object 
 46  Cam_Cause_Health          15979 non-null  object 
 47  Cam_Cause_Humanitarian    15979 non-null  object 
 48  Cam_Cause_Social_Service  15979 non-null  object 
 49  Cam_Cause_Sports          15979 non-null  object 
 50  Cam_Cause_Women_Girls     15979 non-null  object 
dtypes: float64(1), int64(11), object(39)
memory usage: 6.3+ MB
```

There is no donations per donor, So add a columns of

donations per donor

In [108…
```
extract_data.columns
```

Out[108]:
```
Index(['Actual_Donation_Amount', 'NPO_Tax_Deductibility', 'Distinct_Donor
s',
       'Campaign_Goal', 'Campaign_Start_Date', 'Campaign_End_Date',
       'Campaign_Image1', 'Campaign_Image2', 'Campaign_Image3',
       'Campaign_Image4', 'Campaign_Image5', 'Campaign_Video',
       'Impact_Message1', 'Impact_Message2', 'Impact_Message3',
       'Impact_Message4', 'Impact_Message5', 'Custom_Amount1',
       'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount4',
       'Description_Campaign', 'Description_NPO', 'Org_Cause_Animal_Welfar
e',
       'Org_Cause_Arts_Heritage', 'Org_Cause_Children_Youth',
       'Org_Cause_Community', 'Org_Cause_Disability', 'Org_Cause_Educatio
n',
       'Org_Cause_Elderly', 'Org_Cause_Environment', 'Org_Cause_Families',
       'Org_Cause_Health', 'Org_Cause_Humanitarian',
       'Org_Cause_Social Service', 'Org_Cause_Sports', 'Org_Cause_Women_Gi
rls',
       'Cam_Cause_Animal_Welfare', 'Cam_Cause_Arts_Heritage',
       'Cam_Cause_Children_Youth', 'Cam_Cause_Community',
       'Cam_Cause_Disability', 'Cam_Cause_Education', 'Cam_Cause_Elderly',
       'Cam_Cause_Environment', 'Cam_Cause_Families', 'Cam_Cause_Health',
       'Cam_Cause_Humanitarian', 'Cam_Cause_Social_Service',
       'Cam_Cause_Sports', 'Cam_Cause_Women_Girls'],
      dtype='object')
```

In [109…
```
extract_data['NPO_Tax_Deductibility'][0:20]
```

Out[109]:
```
0      True
1      True
2      True
3      True
4      True
5      True
6      True
7     False
8      True
9      True
10     True
11     True
12     True
13     True
14     True
15     True
16     True
17     True
18     True
19     True
Name: NPO_Tax_Deductibility, dtype: object
```

In [138…
```
# I am not sure Distinct_Donors is the total donors or not ?
extract_data['Donation_per_donor'] = 0
for j in range(len(extract_data["Actual_Donation_Amount"])):
    if extract_data["Distinct_Donors"].iloc[j] != 0:
        extract_data['Donation_per_donor'].iloc[j] = extract_data['Actual_Do
    else:
        extract_data['Donation_per_donor'].iloc[j] = 0
```

```
    if extract_data['NPO_Tax_Deductibility'].iloc[j] == True:
        extract_data.loc[j, 'NPO_Tax_Deductibility'] = 1
    else:
        extract_data.loc[j, 'NPO_Tax_Deductibility'] = 0
```

Sum the numbers of org_causes and camp_causes

In [111…

```
Org_causes = ['Org_Cause_Animal_Welfare', 'Org_Cause_Arts_Heritage',
       'Org_Cause_Children_Youth', 'Org_Cause_Community',
       'Org_Cause_Disability', 'Org_Cause_Education', 'Org_Cause_Elderly',
       'Org_Cause_Environment', 'Org_Cause_Families', 'Org_Cause_Health',
       'Org_Cause_Humanitarian', 'Org_Cause_Social Service',
       'Org_Cause_Sports', 'Org_Cause_Women_Girls', 'Cam_Cause_Animal_Welfar
       ]
Cam_causes = ['Cam_Cause_Arts_Heritage', 'Cam_Cause_Children_Youth',
       'Cam_Cause_Community', 'Cam_Cause_Disability', 'Cam_Cause_Education',
       'Cam_Cause_Elderly', 'Cam_Cause_Environment', 'Cam_Cause_Families',
       'Cam_Cause_Health', 'Cam_Cause_Humanitarian',
       'Cam_Cause_Social_Service', 'Cam_Cause_Sports',
       'Cam_Cause_Women_Girls']
Length_Org_causes = len(Org_causes)
Length_Cam_causes = len(Cam_causes)
extract_data['Org_causes'] = 0
extract_data['Cam_causes'] = 0

for j in range(Total_Rows):
    num_Org_causes = 0
    num_Cam_causes = 0
    for position1 in range(Length_Org_causes):
        num_Org_causes += 1 if extract_data[Org_causes[position1]].iloc[j] !
    extract_data['Org_causes'].iloc[j] = num_Org_causes
    for position2 in range(Length_Cam_causes):
        num_Cam_causes += 1 if extract_data[Cam_causes[position2]].iloc[j] !
    extract_data['Org_causes'].iloc[j] = num_Org_causes
    extract_data['Cam_causes'].iloc[j] = num_Cam_causes
```

In [112…

```
extract_data.iloc[0:10,20:50]
```

Out[112]:

	Custom_Amount4	Description_Campaign	Description_NPO	Org_Cause_Animal_Welfare
0	200	Suicide is often preventable. For those at ris...	Founded in 1969, Samaritans of Singapore (SOS...	0
1	200	Over the years at SPD, we saw how assistive te...	SPD is a local charity set up in 1964 to help ...	0
2	200	In 2007, SPD started its Charity Hongbao fundr...	SPD is a local charity set up in 1964 to help ...	0
3	0	Hi Everybody! \r\n\r\nWe are a group of 4 pers...	Habitat for Humanity Singapore is part of an i...	0
4	0	My name is Dhanyatha and I am turning 2 this m...	Children's Cancer Foundation (CCF) is a social...	0
5	200	Women On Mountains (WOM) originated from Ace A...	NaN	0
6	0	Hello everyone! Happy New Year!\r\n\r\nWith t...	The VIVA Foundation for Children with Cancer i...	0
7	200	Stray rescue in Singapore is a determined and ...	Oasis Second Chance Animal Shelter Ltd (OSCAS)...	Animal Welfare
8	0	Do you want to have a different 2017? \r\nSuppo...	Community Chest is the philanthropy and engage...	0
9	200	GIVE THE GIFT OF HOPE\r\n\r\nPersons with auti...	Out of passion to care for the physical, emoti...	0

10 rows × 30 columns

In [113…
```
print(extract_data['NPO_Tax_Deductibility'][0:10])
```

```
0     True
1     True
2     True
3     True
4     True
5     True
6     True
7    False
8     True
9     True
Name: NPO_Tax_Deductibility, dtype: object
```

Add a columns of numbers of images

In [114…
```
Add_Campaign_Image_num = lambda x0,x1,x2,x3,x4: (x0 != 0).astype(np.int) +(x
```

```
extract_data["Campaign_Image_num"] = Add_Campaign_Image_num(extract_data["Ca
```

Classfy video into “0” and ”1“ two categories

In [115…
```
Video_or_not = lambda x0: (x0 != '0').astype(np.int)
extract_data["Campaign_Video"] = Video_or_not(extract_data["Campaign_Video"]
extract_data
```

Out[115]:

	Actual_Donation_Amount	NPO_Tax_Deductibility	Distinct_Donors	Campaign_Goal
0	5561.0	True	66	50000
1	2810.0	True	32	20000
2	1118.0	True	22	30000
3	2800.0	True	7	2000
4	2030.0	True	27	5000
...	...	...	...	...
15974	10.0	True	1	5000
15975	150.0	True	4	10000
15976	1000.0	True	10	1000
15977	120.0	True	2	3000
15978	120.0	True	2	40000

15979 rows × 55 columns

The format of the date needs to be modified and the duration will be calculated below

In [116…
```
month_dictionary = {'Jan':'1',
 'Feb':'2',
 'Mar':'3',
 'Apr':'4',
 'May':'5',
 'Jun':'6',
 'Jul':'7',
 'Aug':'8',
 'Sep':'9',
 'Oct':'10',
 'Nov':'11',
 'Dec':'12'}
extract_data['Campaign_Start_Day'] = '0'
extract_data['Campaign_Start_Month'] = '0'
extract_data['Campaign_Start_Year'] = '0'
extract_data['Campaign_End_Day'] = '0'
extract_data['Campaign_End_Month'] = '0'
extract_data['Campaign_End_Year'] = '0'
extract_data['Campaign_Start'] = '0'
extract_data['Campaign_End'] = '0'
extract_data['duration_day'] = '0'
i = 0
```

```
for row in extract_data['Campaign_Start_Date']:
    extract_data.loc[i, 'Campaign_Start_Day'] = extract_data['Campaign_Start
    extract_data.loc[i, 'Campaign_Start_Month'] =month_dictionary[ extract_d
    extract_data.loc[i, 'Campaign_Start_Year'] = '20'+ extract_data['Campaig
    extract_data.loc[i, 'Campaign_End_Day'] = extract_data['Campaign_End_Dat
    extract_data.loc[i, 'Campaign_End_Month'] = month_dictionary[extract_dat
    extract_data.loc[i, 'Campaign_End_Year'] = '20' + extract_data['Campaign
    extract_data.loc[i, 'Campaign_Start'] = extract_data['Campaign_Start_Yea
    extract_data.loc[i, 'Campaign_End'] = extract_data['Campaign_End_Year'].
    extract_data.loc[i, 'duration_day'] = (datetime.strptime(extract_data['C
    if extract_data.loc[i, 'duration_day'] < 0:
        extract_data.loc[i, 'duration_day'] = 0
    i += 1

extract_data.iloc[:,20:]
```

Out[116]:

	Custom_Amount4	Description_Campaign	Description_NPO	Org_Cause_Animal_Wel
0	200	Suicide is often preventable. For those at ris...	Founded in 1969, Samaritans of Singapore (SOS...	
1	200	Over the years at SPD, we saw how assistive te...	SPD is a local charity set up in 1964 to help ...	
2	200	In 2007, SPD started its Charity Hongbao fundr...	SPD is a local charity set up in 1964 to help ...	
3	0	Hi Everybody! \r\n\r\nWe are a group of 4 pers...	Habitat for Humanity Singapore is part of an i...	
4	0	My name is Dhanyatha and I am turning 2 this m...	Children's Cancer Foundation (CCF) is a social...	
...	...	...	...	
15974	0	Endowus is an investing platform dedicated to ...	Gardens by the Bay is a national garden with c...	
15975	0	Diabetes is a major public health concern. Glo...	Family Medicine is a medical discipline dedica...	
15976	0	Diabetes is a major public health concern. Glo...	Family Medicine is a medical discipline dedica...	
15977	0	The Women's ERG group at Coinbase - WE@SG - is...	Empowering Women, Enabling Families. \r\n\r\nD...	
15978	200	This year, RLAF hosted our inaugural Rahmatan ...	The Rahmatan lil Alamin Foundation (RLAF) was ...	

15979 rows × 44 columns

In [117…

```
extract_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15979 entries, 0 to 15978
Data columns (total 64 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Actual_Donation_Amount    15979 non-null  float64
 1   NPO_Tax_Deductibility     15979 non-null  object 
 2   Distinct_Donors           15979 non-null  int64  
 3   Campaign_Goal             15979 non-null  int64  
 4   Campaign_Start_Date       15979 non-null  object 
 5   Campaign_End_Date         15979 non-null  object 
 6   Campaign_Image1           15979 non-null  int64  
 7   Campaign_Image2           15979 non-null  int64  
 8   Campaign_Image3           15979 non-null  int64  
 9   Campaign_Image4           15979 non-null  int64  
 10  Campaign_Image5           15979 non-null  int64  
 11  Campaign_Video            15979 non-null  int64  
 12  Impact_Message1           15979 non-null  object 
 13  Impact_Message2           15979 non-null  object 
 14  Impact_Message3           15979 non-null  object 
 15  Impact_Message4           15979 non-null  object 
 16  Impact_Message5           15979 non-null  object 
 17  Custom_Amount1            15979 non-null  int64  
 18  Custom_Amount2            15979 non-null  int64  
 19  Custom_Amount3            15979 non-null  int64  
 20  Custom_Amount4            15979 non-null  int64  
 21  Description_Campaign      15971 non-null  object 
 22  Description_NPO           13270 non-null  object 
 23  Org_Cause_Animal_Welfare  15979 non-null  object 
 24  Org_Cause_Arts_Heritage   15979 non-null  object 
 25  Org_Cause_Children_Youth  15979 non-null  object 
 26  Org_Cause_Community       15979 non-null  object 
 27  Org_Cause_Disability      15979 non-null  object 
 28  Org_Cause_Education       15979 non-null  object 
 29  Org_Cause_Elderly         15979 non-null  object 
 30  Org_Cause_Environment     15979 non-null  object 
 31  Org_Cause_Families        15979 non-null  object 
 32  Org_Cause_Health          15979 non-null  object 
 33  Org_Cause_Humanitarian    15979 non-null  object 
 34  Org_Cause_Social Service  15979 non-null  object 
 35  Org_Cause_Sports          15979 non-null  object 
 36  Org_Cause_Women_Girls     15979 non-null  object 
 37  Cam_Cause_Animal_Welfare  15979 non-null  object 
 38  Cam_Cause_Arts_Heritage   15979 non-null  object 
 39  Cam_Cause_Children_Youth  15979 non-null  object 
 40  Cam_Cause_Community       15979 non-null  object 
 41  Cam_Cause_Disability      15979 non-null  object 
 42  Cam_Cause_Education       15979 non-null  object 
 43  Cam_Cause_Elderly         15979 non-null  object 
 44  Cam_Cause_Environment     15979 non-null  object 
 45  Cam_Cause_Families        15979 non-null  object 
 46  Cam_Cause_Health          15979 non-null  object 
 47  Cam_Cause_Humanitarian    15979 non-null  object 
 48  Cam_Cause_Social_Service  15979 non-null  object 
 49  Cam_Cause_Sports          15979 non-null  object 
 50  Cam_Cause_Women_Girls     15979 non-null  object 
 51  Donation_per_donor        15979 non-null  float64
 52  Org_causes                15979 non-null  int64  
 53  Cam_causes                15979 non-null  int64  
 54  Campaign_Image_num        15979 non-null  int64  
 55  Campaign_Start_Day        15979 non-null  object 
```

```
 56  Campaign_Start_Month      15979 non-null  object
 57  Campaign_Start_Year       15979 non-null  object
 58  Campaign_End_Day          15979 non-null  object
 59  Campaign_End_Month        15979 non-null  object
 60  Campaign_End_Year         15979 non-null  object
 61  Campaign_Start            15979 non-null  object
 62  Campaign_End              15979 non-null  object
 63  duration_day              15979 non-null  object
dtypes: float64(2), int64(15), object(47)
memory usage: 8.5+ MB
```

See more infomation about every columns

Check whether there are missing data

In [118…
```
extract_data.isnull().sum()
```

Out[118]:
```
Actual_Donation_Amount    0
NPO_Tax_Deductibility     0
Distinct_Donors           0
Campaign_Goal             0
Campaign_Start_Date       0
                         ..
Campaign_End_Month        0
Campaign_End_Year         0
Campaign_Start            0
Campaign_End              0
duration_day              0
Length: 64, dtype: int64
```

Sentiment Analysis

In [119…
```
comm_data = pd.DataFrame()
extract_data['Msg1_polarity'] = 0
extract_data['Msg1_subjectivity'] = 0
extract_data['Msg2_polarity'] = 0
extract_data['Msg2_subjectivity'] = 0
extract_data['Msg3_polarity'] = 0
extract_data['Msg3_subjectivity'] = 0
extract_data['Msg4_polarity'] = 0
extract_data['Msg4_subjectivity'] = 0
extract_data['Msg5_polarity'] = 0
extract_data['Msg5_subjectivity'] = 0
```

In [120…
```
extract_data.columns
```

Out[120]:
```
Index(['Actual_Donation_Amount', 'NPO_Tax_Deductibility', 'Distinct_Donors',
       'Campaign_Goal', 'Campaign_Start_Date', 'Campaign_End_Date',
       'Campaign_Image1', 'Campaign_Image2', 'Campaign_Image3',
       'Campaign_Image4', 'Campaign_Image5', 'Campaign_Video',
       'Impact_Message1', 'Impact_Message2', 'Impact_Message3',
       'Impact_Message4', 'Impact_Message5', 'Custom_Amount1',
       'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount4',
       'Description_Campaign', 'Description_NPO', 'Org_Cause_Animal_Welfare',
       'Org_Cause_Arts_Heritage', 'Org_Cause_Children_Youth',
       'Org_Cause_Community', 'Org_Cause_Disability', 'Org_Cause_Education',
       'Org_Cause_Elderly', 'Org_Cause_Environment', 'Org_Cause_Families',
       'Org_Cause_Health', 'Org_Cause_Humanitarian',
       'Org_Cause_Social Service', 'Org_Cause_Sports', 'Org_Cause_Women_Girls',
       'Cam_Cause_Animal_Welfare', 'Cam_Cause_Arts_Heritage',
       'Cam_Cause_Children_Youth', 'Cam_Cause_Community',
       'Cam_Cause_Disability', 'Cam_Cause_Education', 'Cam_Cause_Elderly',
       'Cam_Cause_Environment', 'Cam_Cause_Families', 'Cam_Cause_Health',
       'Cam_Cause_Humanitarian', 'Cam_Cause_Social_Service',
       'Cam_Cause_Sports', 'Cam_Cause_Women_Girls', 'Donation_per_donor',
       'Org_causes', 'Cam_causes', 'Campaign_Image_num', 'Campaign_Start_Day',
       'Campaign_Start_Month', 'Campaign_Start_Year', 'Campaign_End_Day',
       'Campaign_End_Month', 'Campaign_End_Year', 'Campaign_Start',
       'Campaign_End', 'duration_day', 'Msg1_polarity', 'Msg1_subjectivity',
       'Msg2_polarity', 'Msg2_subjectivity', 'Msg3_polarity',
       'Msg3_subjectivity', 'Msg4_polarity', 'Msg4_subjectivity',
       'Msg5_polarity', 'Msg5_subjectivity'],
      dtype='object')
```

The polarity item is the positiveness of the text, which is a floating point number in the range of [-1.0, 1.0] The subjectivity item is a subjective score, which is a floating point number in the range of [0.0, 1.0], where 0.0 is very objective and 1.0 is very subjective

In [121…
```
from textblob import TextBlob
# polarity项为文本积极性，是在[-1.0，1.0]范围内的浮点数
# subjectivity项为主观评分，是在[0.0，1.0]范围内的浮点数，其中0.0是非常客观的，而1.0是
Impact_msg_list = ['Impact_Message1','Impact_Message2','Impact_Message3','Im
Msg_polarity_list = ['Msg1_polarity','Msg2_polarity','Msg3_polarity','Msg4_p
Msg1_subjectivity_list = ['Msg1_subjectivity','Msg2_subjectivity','Msg3_subj
for j in range(len(Impact_msg_list)):
    t=0
    for i in extract_data[Impact_msg_list[j]]:
        blob = TextBlob(i)
        sentiment = blob.sentiment
        extract_data[Msg_polarity_list[j]].iloc[t] = sentiment.polarity
        extract_data[Msg1_subjectivity_list[j]].iloc[t] = sentiment.subjecti
        t+=1
# sum the total five messages polarity and subjectivity
```

```
extract_data["Total_Msg_polarity"] = extract_data["Msg1_polarity"]+extract_d
extract_data["Total_Msg_subjectivity"] = extract_data["Msg1_subjectivity"]+e
extract_data.iloc[0:30,28:]
```

Out[121]:

	Org_Cause_Education	Org_Cause_Elderly	Org_Cause_Environment	Org_Cause_Familie
0	0	0	0	Familie
1	0	0	0	
2	0	0	0	
3	0	Elderly	Environment	
4	0	0	0	
5	0	0	0	
6	Education	0	0	Familie
7	Education	0	Environment	
8	0	Elderly	0	Familie
9	Education	Elderly	0	
10	0	Elderly	0	Familie
11	0	Elderly	0	
12	0	0	0	Familie
13	0	Elderly	Environment	
14	0	Elderly	0	
15	Education	0	0	
16	Education	0	0	Familie
17	0	0	0	
18	0	Elderly	0	
19	0	0	0	
20	0	0	0	
21	0	Elderly	0	Familie
22	Education	Elderly	0	
23	0	0	0	
24	0	0	0	
25	0	Elderly	0	
26	0	0	0	
27	0	0	0	Familie
28	0	0	0	Familie
29	Education	0	0	

30 rows × 48 columns

The method of judging the similarity uses the difflib library

It is a score, which in range of [0.0, 1.0]. 0 means this two sentences are totally different and 1 means there are the same.

In [122…
```
import difflib
def get_equal_rate_1(str1, str2):
   return difflib.SequenceMatcher(None, str1, str2).quick_ratio()
extract_data['Total_similarity'] = 0
Impact_msg_list = ['Impact_Message1','Impact_Message2','Impact_Message3','Im
for j in range(len(Impact_msg_list)-1):
    for i in range(extract_data[Impact_msg_list[j]].shape[0]):
        str1 = extract_data[Impact_msg_list[j]].iloc[i]
        str2 = extract_data[Impact_msg_list[j+1]].iloc[i]
        extract_data['Total_similarity'].iloc[i] += get_equal_rate_1(str1, s
extract_data.iloc[:,28:]
```

Out[122]:

	Org_Cause_Education	Org_Cause_Elderly	Org_Cause_Environment	Org_Cause_Fam
0	0	0	0	Fan
1	0	0	0	
2	0	0	0	
3	0	Elderly	Environment	
4	0	0	0	
...	...	...	...	
15974	Education	0	0	Fan
15975	0	0	0	
15976	0	0	0	
15977	0	0	0	
15978	0	0	0	

15979 rows × 49 columns

Between two strings, the minimum number of editing operations required to convert one into another, if the distance between them is greater, it means that they are more different

In [123…
```
import distance
extract_data['Total_distance'] = 0
def edit_distance(s1, s2):
    return distance.levenshtein(s1, s2)

for j in range(len(Impact_msg_list)-1):
```

```
    for i in range(extract_data[Impact_msg_list[j]].shape[0]):
        str1 = extract_data[Impact_msg_list[j]].iloc[i]
        str2 = extract_data[Impact_msg_list[j+1]].iloc[i]
        extract_data['Total_distance'].iloc[i] += edit_distance(str1, str2)

extract_data.iloc[0:30,28:]
```

Out[123]:

	Org_Cause_Education	Org_Cause_Elderly	Org_Cause_Environment	Org_Cause_Familie
0	0	0	0	Familie
1	0	0	0	
2	0	0	0	
3	0	Elderly	Environment	
4	0	0	0	
5	0	0	0	
6	Education	0	0	Familie
7	Education	0	Environment	
8	0	Elderly	0	Familie
9	Education	Elderly	0	
10	0	Elderly	0	Familie
11	0	Elderly	0	
12	0	0	0	Familie
13	0	Elderly	Environment	
14	0	Elderly	0	
15	Education	0	0	
16	Education	0	0	Familie
17	0	0	0	
18	0	Elderly	0	
19	0	0	0	
20	0	0	0	
21	0	Elderly	0	Familie
22	Education	Elderly	0	
23	0	0	0	
24	0	0	0	
25	0	Elderly	0	
26	0	0	0	
27	0	0	0	Familie
28	0	0	0	Familie
29	Education	0	0	

30 rows × 50 columns

Divide into four category

In [124…
```
extract_data['Msg1_category'] = 0
extract_data['Msg2_category'] = 0
extract_data['Msg3_category'] = 0
extract_data['Msg4_category'] = 0
extract_data['Msg5_category'] = 0
#
extract_data['Num_desc_cam'] = 0
extract_data['Num_desc_NPO'] = 0
Impact_msg_list = ['Impact_Message1','Impact_Message2','Impact_Message3','Im
Msg_category_list = ['Msg1_category','Msg2_category','Msg3_category','Msg4_c
def sentence_length(s):
    return len([i for i in s.split(' ') if i])

for j in range(len(Impact_msg_list)):
    cnt=0
    for s in extract_data[Impact_msg_list[j]]:
        extract_data[Msg_category_list[j]].iloc[cnt] = 0 if sentence_length(
        cnt += 1
```

In [125…
```
for r in range(Total_Rows):
    extract_data.loc[r, 'Num_desc_cam'] = sentence_length( str(extract_data[
    extract_data.loc[r, 'Num_desc_NPO'] = sentence_length( str(extract_data[
extract_data.iloc[0:30,34:]
```

Out[125]:

	Org_Cause_Social Service	Org_Cause_Sports	Org_Cause_Women_Girls	Cam_Cause_Animal_W
0	Social Service	0	0	
1	Social Service	0	0	
2	Social Service	0	0	
3	Social Service	0	0	
4	Social Service	0	0	
5	0	0	0	
6	0	0	0	
7	0	0	0	Animal W
8	0	0	0	
9	0	0	0	
10	0	0	0	
11	Social Service	0	0	
12	0	0	Women & Girls	
13	Social Service	0	0	
14	Social Service	0	0	
15	0	0	0	
16	0	0	0	
17	Social Service	0	0	
18	0	0	0	
19	0	0	0	
20	0	0	0	
21	Social Service	0	0	
22	0	0	0	
23	0	0	0	
24	Social Service	0	0	
25	0	0	0	
26	Social Service	0	0	
27	Social Service	0	0	Animal W
28	0	0	Women & Girls	
29	Social Service	0	0	Animal W

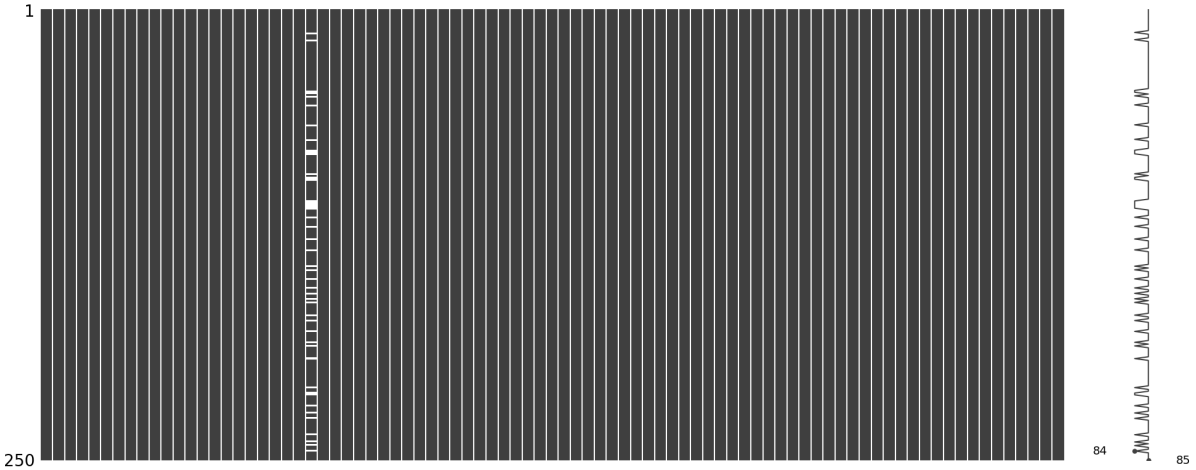
30 rows × 51 columns

Well done of data cleaning and feature structure

In [126…

```
msno.matrix(extract_data.sample(250))
```

Out[126]: <AxesSubplot:>



In [127…
```
extract_data.columns
```

Out[127]:
```
Index(['Actual_Donation_Amount', 'NPO_Tax_Deductibility', 'Distinct_Donor
s',
       'Campaign_Goal', 'Campaign_Start_Date', 'Campaign_End_Date',
       'Campaign_Image1', 'Campaign_Image2', 'Campaign_Image3',
       'Campaign_Image4', 'Campaign_Image5', 'Campaign_Video',
       'Impact_Message1', 'Impact_Message2', 'Impact_Message3',
       'Impact_Message4', 'Impact_Message5', 'Custom_Amount1',
       'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount4',
       'Description_Campaign', 'Description_NPO', 'Org_Cause_Animal_Welfar
e',
       'Org_Cause_Arts_Heritage', 'Org_Cause_Children_Youth',
       'Org_Cause_Community', 'Org_Cause_Disability', 'Org_Cause_Educatio
n',
       'Org_Cause_Elderly', 'Org_Cause_Environment', 'Org_Cause_Families',
       'Org_Cause_Health', 'Org_Cause_Humanitarian',
       'Org_Cause_Social Service', 'Org_Cause_Sports', 'Org_Cause_Women_Gi
rls',
       'Cam_Cause_Animal_Welfare', 'Cam_Cause_Arts_Heritage',
       'Cam_Cause_Children_Youth', 'Cam_Cause_Community',
       'Cam_Cause_Disability', 'Cam_Cause_Education', 'Cam_Cause_Elderly',
       'Cam_Cause_Environment', 'Cam_Cause_Families', 'Cam_Cause_Health',
       'Cam_Cause_Humanitarian', 'Cam_Cause_Social_Service',
       'Cam_Cause_Sports', 'Cam_Cause_Women_Girls', 'Donation_per_donor',
       'Org_causes', 'Cam_causes', 'Campaign_Image_num', 'Campaign_Start_D
ay',
       'Campaign_Start_Month', 'Campaign_Start_Year', 'Campaign_End_Day',
       'Campaign_End_Month', 'Campaign_End_Year', 'Campaign_Start',
       'Campaign_End', 'duration_day', 'Msg1_polarity', 'Msg1_subjectivit
y',
       'Msg2_polarity', 'Msg2_subjectivity', 'Msg3_polarity',
       'Msg3_subjectivity', 'Msg4_polarity', 'Msg4_subjectivity',
       'Msg5_polarity', 'Msg5_subjectivity', 'Total_Msg_polarity',
       'Total_Msg_subjectivity', 'Total_similarity', 'Total_distance',
       'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4_category',
       'Msg5_category', 'Num_desc_cam', 'Num_desc_NPO'],
      dtype='object')
```

Convert to numeric type

In [128…
```
extract_data['Total_Msg_subjectivity'] = pd.to_numeric( extract_data['Total_
extract_data['Total_Msg_polarity'] = pd.to_numeric( extract_data['Total_Msg_
```

```
extract_data['NPO_Tax_Deductibility'] = pd.to_numeric(extract_data['NPO_Tax_
extract_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15979 entries, 0 to 15978
Data columns (total 85 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Actual_Donation_Amount    15979 non-null  float64
 1   NPO_Tax_Deductibility     15979 non-null  int64  
 2   Distinct_Donors           15979 non-null  int64  
 3   Campaign_Goal             15979 non-null  int64  
 4   Campaign_Start_Date       15979 non-null  object 
 5   Campaign_End_Date         15979 non-null  object 
 6   Campaign_Image1           15979 non-null  int64  
 7   Campaign_Image2           15979 non-null  int64  
 8   Campaign_Image3           15979 non-null  int64  
 9   Campaign_Image4           15979 non-null  int64  
 10  Campaign_Image5           15979 non-null  int64  
 11  Campaign_Video            15979 non-null  int64  
 12  Impact_Message1           15979 non-null  object 
 13  Impact_Message2           15979 non-null  object 
 14  Impact_Message3           15979 non-null  object 
 15  Impact_Message4           15979 non-null  object 
 16  Impact_Message5           15979 non-null  object 
 17  Custom_Amount1            15979 non-null  int64  
 18  Custom_Amount2            15979 non-null  int64  
 19  Custom_Amount3            15979 non-null  int64  
 20  Custom_Amount4            15979 non-null  int64  
 21  Description_Campaign      15971 non-null  object 
 22  Description_NPO           13270 non-null  object 
 23  Org_Cause_Animal_Welfare  15979 non-null  object 
 24  Org_Cause_Arts_Heritage   15979 non-null  object 
 25  Org_Cause_Children_Youth  15979 non-null  object 
 26  Org_Cause_Community       15979 non-null  object 
 27  Org_Cause_Disability      15979 non-null  object 
 28  Org_Cause_Education       15979 non-null  object 
 29  Org_Cause_Elderly         15979 non-null  object 
 30  Org_Cause_Environment     15979 non-null  object 
 31  Org_Cause_Families        15979 non-null  object 
 32  Org_Cause_Health          15979 non-null  object 
 33  Org_Cause_Humanitarian    15979 non-null  object 
 34  Org_Cause_Social Service  15979 non-null  object 
 35  Org_Cause_Sports          15979 non-null  object 
 36  Org_Cause_Women_Girls     15979 non-null  object 
 37  Cam_Cause_Animal_Welfare  15979 non-null  object 
 38  Cam_Cause_Arts_Heritage   15979 non-null  object 
 39  Cam_Cause_Children_Youth  15979 non-null  object 
 40  Cam_Cause_Community       15979 non-null  object 
 41  Cam_Cause_Disability      15979 non-null  object 
 42  Cam_Cause_Education       15979 non-null  object 
 43  Cam_Cause_Elderly         15979 non-null  object 
 44  Cam_Cause_Environment     15979 non-null  object 
 45  Cam_Cause_Families        15979 non-null  object 
 46  Cam_Cause_Health          15979 non-null  object 
 47  Cam_Cause_Humanitarian    15979 non-null  object 
 48  Cam_Cause_Social_Service  15979 non-null  object 
 49  Cam_Cause_Sports          15979 non-null  object 
 50  Cam_Cause_Women_Girls     15979 non-null  object 
 51  Donation_per_donor        15979 non-null  float64
 52  Org_causes                15979 non-null  int64  
 53  Cam_causes                15979 non-null  int64  
 54  Campaign_Image_num        15979 non-null  int64  
 55  Campaign_Start_Day        15979 non-null  object 
```

```
 56  Campaign_Start_Month      15979 non-null  object
 57  Campaign_Start_Year       15979 non-null  object
 58  Campaign_End_Day          15979 non-null  object
 59  Campaign_End_Month        15979 non-null  object
 60  Campaign_End_Year         15979 non-null  object
 61  Campaign_Start            15979 non-null  object
 62  Campaign_End              15979 non-null  object
 63  duration_day              15979 non-null  object
 64  Msg1_polarity             15979 non-null  float64
 65  Msg1_subjectivity         15979 non-null  float64
 66  Msg2_polarity             15979 non-null  float64
 67  Msg2_subjectivity         15979 non-null  float64
 68  Msg3_polarity             15979 non-null  float64
 69  Msg3_subjectivity         15979 non-null  float64
 70  Msg4_polarity             15979 non-null  float64
 71  Msg4_subjectivity         15979 non-null  float64
 72  Msg5_polarity             15979 non-null  float64
 73  Msg5_subjectivity         15979 non-null  float64
 74  Total_Msg_polarity        15979 non-null  float64
 75  Total_Msg_subjectivity    15979 non-null  float64
 76  Total_similarity          15979 non-null  float64
 77  Total_distance            15979 non-null  int64
 78  Msg1_category             15979 non-null  int64
 79  Msg2_category             15979 non-null  int64
 80  Msg3_category             15979 non-null  int64
 81  Msg4_category             15979 non-null  int64
 82  Msg5_category             15979 non-null  int64
 83  Num_desc_cam              15979 non-null  int64
 84  Num_desc_NPO              15979 non-null  int64
dtypes: float64(15), int64(24), object(46)
memory usage: 11.1+ MB
```

In [129…

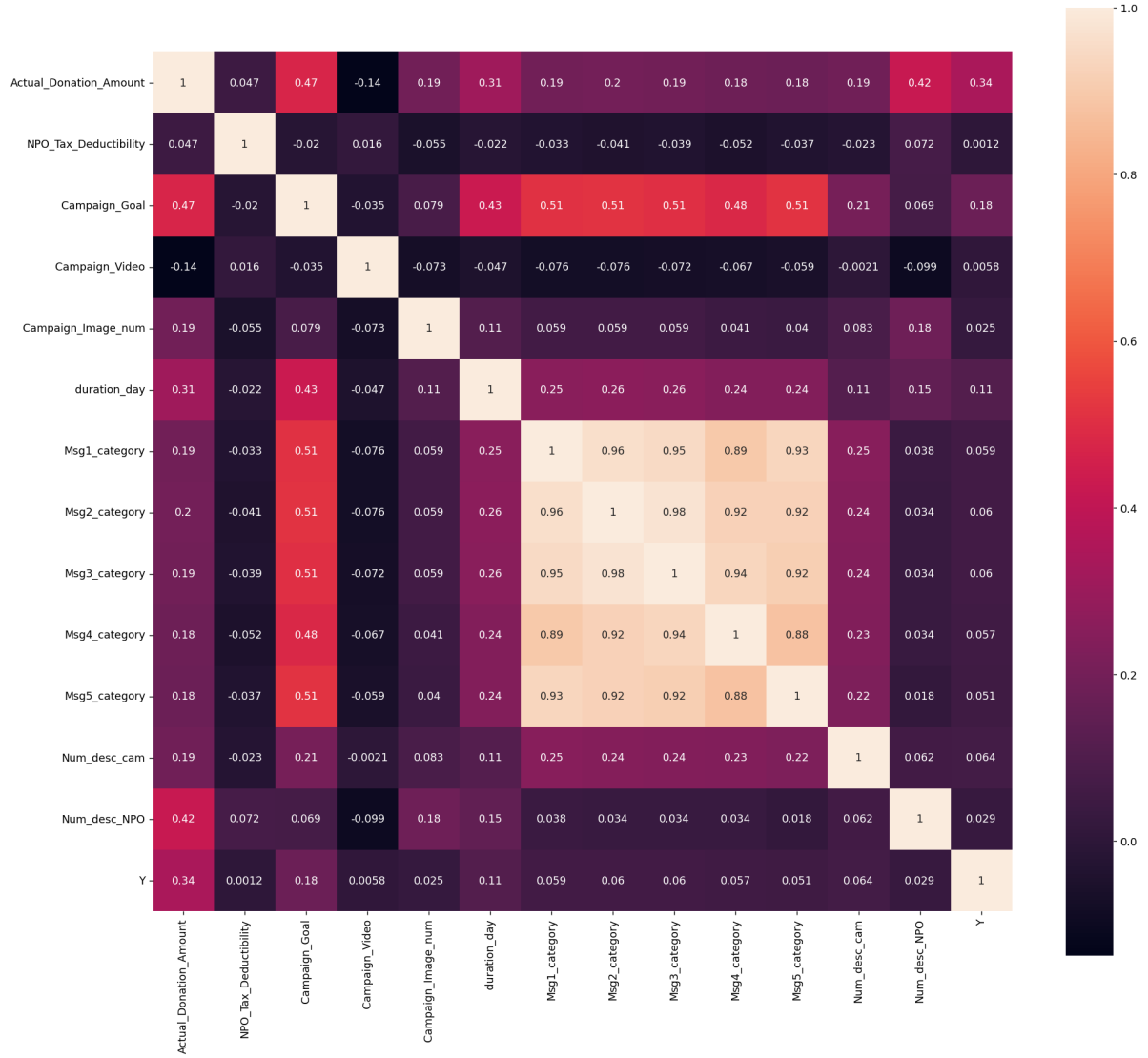
```
extract_data[20:].info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15959 entries, 20 to 15978
Data columns (total 85 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Actual_Donation_Amount    15959 non-null  float64
 1   NPO_Tax_Deductibility     15959 non-null  int64  
 2   Distinct_Donors           15959 non-null  int64  
 3   Campaign_Goal             15959 non-null  int64  
 4   Campaign_Start_Date       15959 non-null  object 
 5   Campaign_End_Date         15959 non-null  object 
 6   Campaign_Image1           15959 non-null  int64  
 7   Campaign_Image2           15959 non-null  int64  
 8   Campaign_Image3           15959 non-null  int64  
 9   Campaign_Image4           15959 non-null  int64  
 10  Campaign_Image5           15959 non-null  int64  
 11  Campaign_Video            15959 non-null  int64  
 12  Impact_Message1           15959 non-null  object 
 13  Impact_Message2           15959 non-null  object 
 14  Impact_Message3           15959 non-null  object 
 15  Impact_Message4           15959 non-null  object 
 16  Impact_Message5           15959 non-null  object 
 17  Custom_Amount1            15959 non-null  int64  
 18  Custom_Amount2            15959 non-null  int64  
 19  Custom_Amount3            15959 non-null  int64  
 20  Custom_Amount4            15959 non-null  int64  
 21  Description_Campaign      15951 non-null  object 
 22  Description_NPO           13251 non-null  object 
 23  Org_Cause_Animal_Welfare  15959 non-null  object 
 24  Org_Cause_Arts_Heritage   15959 non-null  object 
 25  Org_Cause_Children_Youth  15959 non-null  object 
 26  Org_Cause_Community       15959 non-null  object 
 27  Org_Cause_Disability      15959 non-null  object 
 28  Org_Cause_Education       15959 non-null  object 
 29  Org_Cause_Elderly         15959 non-null  object 
 30  Org_Cause_Environment     15959 non-null  object 
 31  Org_Cause_Families        15959 non-null  object 
 32  Org_Cause_Health          15959 non-null  object 
 33  Org_Cause_Humanitarian    15959 non-null  object 
 34  Org_Cause_Social Service  15959 non-null  object 
 35  Org_Cause_Sports          15959 non-null  object 
 36  Org_Cause_Women_Girls     15959 non-null  object 
 37  Cam_Cause_Animal_Welfare  15959 non-null  object 
 38  Cam_Cause_Arts_Heritage   15959 non-null  object 
 39  Cam_Cause_Children_Youth  15959 non-null  object 
 40  Cam_Cause_Community       15959 non-null  object 
 41  Cam_Cause_Disability      15959 non-null  object 
 42  Cam_Cause_Education       15959 non-null  object 
 43  Cam_Cause_Elderly         15959 non-null  object 
 44  Cam_Cause_Environment     15959 non-null  object 
 45  Cam_Cause_Families        15959 non-null  object 
 46  Cam_Cause_Health          15959 non-null  object 
 47  Cam_Cause_Humanitarian    15959 non-null  object 
 48  Cam_Cause_Social_Service  15959 non-null  object 
 49  Cam_Cause_Sports          15959 non-null  object 
 50  Cam_Cause_Women_Girls     15959 non-null  object 
 51  Donation_per_donor        15959 non-null  float64
 52  Org_causes                15959 non-null  int64  
 53  Cam_causes                15959 non-null  int64  
 54  Campaign_Image_num        15959 non-null  int64  
 55  Campaign_Start_Day        15959 non-null  object 
```

```
 56  Campaign_Start_Month      15959 non-null  object
 57  Campaign_Start_Year       15959 non-null  object
 58  Campaign_End_Day          15959 non-null  object
 59  Campaign_End_Month        15959 non-null  object
 60  Campaign_End_Year         15959 non-null  object
 61  Campaign_Start            15959 non-null  object
 62  Campaign_End              15959 non-null  object
 63  duration_day              15959 non-null  object
 64  Msg1_polarity             15959 non-null  float64
 65  Msg1_subjectivity         15959 non-null  float64
 66  Msg2_polarity             15959 non-null  float64
 67  Msg2_subjectivity         15959 non-null  float64
 68  Msg3_polarity             15959 non-null  float64
 69  Msg3_subjectivity         15959 non-null  float64
 70  Msg4_polarity             15959 non-null  float64
 71  Msg4_subjectivity         15959 non-null  float64
 72  Msg5_polarity             15959 non-null  float64
 73  Msg5_subjectivity         15959 non-null  float64
 74  Total_Msg_polarity        15959 non-null  float64
 75  Total_Msg_subjectivity    15959 non-null  float64
 76  Total_similarity          15959 non-null  float64
 77  Total_distance            15959 non-null  int64
 78  Msg1_category             15959 non-null  int64
 79  Msg2_category             15959 non-null  int64
 80  Msg3_category             15959 non-null  int64
 81  Msg4_category             15959 non-null  int64
 82  Msg5_category             15959 non-null  int64
 83  Num_desc_cam              15959 non-null  int64
 84  Num_desc_NPO              15959 non-null  int64
dtypes: float64(15), int64(24), object(46)
memory usage: 10.5+ MB
```

In [152…

```
#numeric_features Store the following variables that need to draw correlatic
numeric_feature = ['Actual_Donation_Amount', 'NPO_Tax_Deductibility', 'Campa
                   'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4
                   'Msg5_category', 'Num_desc_cam', 'Num_desc_NPO' ]
numeric_features1 = [ 'Campaign_Goal', 'NPO_Tax_Deductibility',
       'Campaign_Video', 'Total_Msg_polarity', 'Total_Msg_subjectivity',
       'Custom_Amount1', 'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount4
       'Campaign_Image_num', 'duration_day', 'Msg1_subjectivity',
       'Msg2_subjectivity', 'Msg3_subjectivity', 'Msg4_subjectivity',
       'Msg5_subjectivity', 'Total_similarity', 'Total_distance',
       'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4_category',
       'Msg5_category','Org_causes', 'Cam_causes']

numeric_features2 = ['Actual_Donation_Amount', 'Campaign_Goal', 'duration_da
       'Campaign_Video',
       'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4_category',
       'Msg5_category','Total_Msg_polarity', 'Total_Msg_subjectivity','Total

#Correlation analysis
price_numeric = extract_data[numeric_feature]
correlation = price_numeric.corr()
y_train = Orginal_data['Actual_Donation_Amount']
corr = plt.subplots(figsize = (18,16), dpi=128)
corr= sns.heatmap(price_numeric.assign(Y=y_train).corr(method='spearman'), a
```



In [ ]:

Modeling verfication

Model1 'Org_causes' 'Cam_causes' are ignored?

In [140…
```
variable_list1 = ['Actual_Donation_Amount', 'Campaign_Goal','duration_day' ,
                  'Campaign_Image_num', 'duration_day', 'Num_desc_cam', 'Num_d
variables_data1 = extract_data[variable_list1]
```

Model2

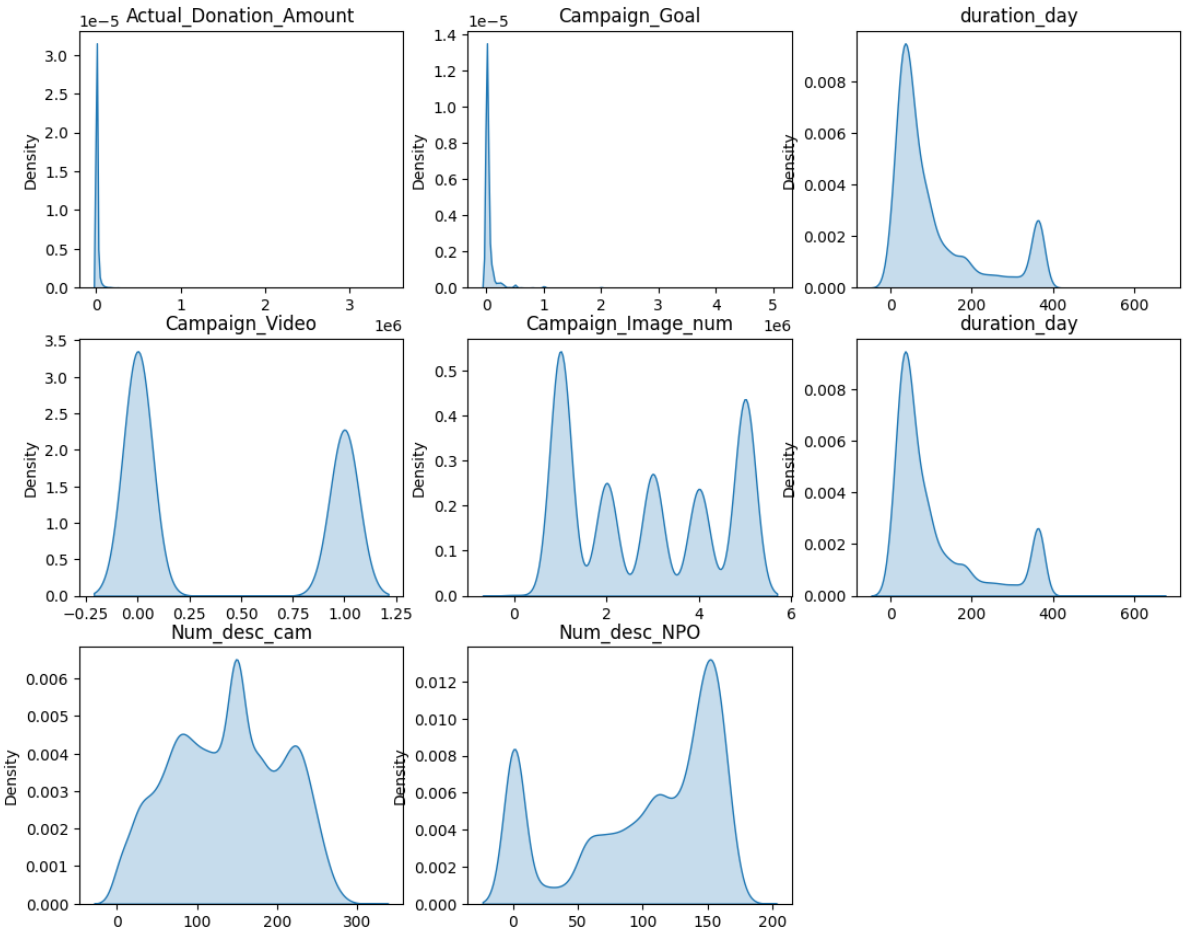
In [147…
```
variable_list2 = ['Actual_Donation_Amount', 'NPO_Tax_Deductibility', 'Campai
                    'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4
                    'Msg5_category', 'Num_desc_cam', 'Num_desc_NPO' ]
variables_data2 = extract_data[variable_list2]
```

Model3

In [148…
```
variable_list3 = ['Actual_Donation_Amount',  'Campaign_Goal', 'NPO_Tax_Deduc
                  'Campaign_Video', 'Msg1_category', 'Msg2_category', 'Msg
                  'Msg5_category', 'Num_desc_cam', 'Num_desc_NPO', 'Org_ca
variables_data3 = extract_data[variable_list3]
```

Variance, Average, Max, Min, Median calculation

In [134…
```
i = 0
plt.figure(figsize=(13, 14))
plt.xticks([])
for title in variable_list1:
    plt.subplot(4,3,i+1)
    plt.title(title)
    sns.kdeplot(extract_data[title], shade=True)
    plt.xlabel(" ")
    i += 1


#plt.hist(extract_data['Campaign_Goal'], bins=80, histtype="stepfilled", alp
```



In [153…
```
for title in variable_list2:

    extract_data[title] = pd.to_numeric( extract_data[title])
    print( title, "Average:",np.average(extract_data[title]))
    print( title, "Variance:" ,np.var(extract_data[title]))
    print( title, "Min:" ,np.min(extract_data[title]))
    print( title, "Max:" ,np.max(extract_data[title]))
    print( title, "Median:",np.median(extract_data[title]))
```

```
Actual_Donation_Amount Variance: 3966231020.7807913
Actual_Donation_Amount Max: 3431670.0
Actual_Donation_Amount Min: 0.0
Actual_Donation_Amount Median: 1300.0
Actual_Donation_Amount Average: 9813.046623693597
NPO_Tax_Deductibility Variance: 0.0576239683380983
NPO_Tax_Deductibility Max: 1
NPO_Tax_Deductibility Min: 0
NPO_Tax_Deductibility Median: 1.0
NPO_Tax_Deductibility Average: 0.9386069215845798
Campaign_Goal Variance: 23844896574.26401
Campaign_Goal Max: 5000000
Campaign_Goal Min: 100
Campaign_Goal Median: 5000.0
Campaign_Goal Average: 44797.3589085675
Campaign_Video Variance: 0.2408976052926278
Campaign_Video Max: 1
Campaign_Video Min: 0
Campaign_Video Median: 0.0
Campaign_Video Average: 0.40459352900682144
Campaign_Image_num Variance: 2.5242411674773044
Campaign_Image_num Max: 5
Campaign_Image_num Min: 0
Campaign_Image_num Median: 3.0
Campaign_Image_num Average: 2.8678265223105326
duration_day Variance: 12084.032362458203
duration_day Max: 630
duration_day Min: 0
duration_day Median: 60.0
duration_day Average: 107.74403905125477
Msg1_category Variance: 1.1590159280375665
Msg1_category Max: 3
Msg1_category Min: 0
Msg1_category Median: 1.0
Msg1_category Average: 0.9761562050190875
Msg2_category Variance: 1.1546736542365836
Msg2_category Max: 3
Msg2_category Min: 0
Msg2_category Median: 1.0
Msg2_category Average: 0.9678953626634959
Msg3_category Variance: 1.1734991716022736
Msg3_category Max: 3
Msg3_category Min: 0
Msg3_category Median: 1.0
Msg3_category Average: 0.9658301520745979
Msg4_category Variance: 1.1787740872940982
Msg4_category Max: 3
Msg4_category Min: 0
Msg4_category Median: 0.0
Msg4_category Average: 0.925214343826272
Msg5_category Variance: 1.0636654300489776
Msg5_category Max: 3
Msg5_category Min: 0
Msg5_category Median: 1.0
Msg5_category Average: 0.9041867451029476
Num_desc_cam Variance: 4579.946342115488
Num_desc_cam Max: 309
Num_desc_cam Min: 1
Num_desc_cam Median: 143.0
Num_desc_cam Average: 137.2523311846799
Num_desc_NPO Variance: 3181.0814422902467
```

```
Num_desc_NPO Max: 179
Num_desc_NPO Min: 1
Num_desc_NPO Median: 115.0
Num_desc_NPO Average: 100.5663683584705
```

In [136…
```
variable_list4 = ['Actual_Donation_Amount','Donation_per_donor','Campaign_Go
       'Campaign_Video', 'Total_Msg_polarity', 'Total_Msg_subjectivity',
       'Custom_Amount1', 'Custom_Amount2', 'Custom_Amount3', 'Custom_Amount4
       'Campaign_Image_num', 'duration_day', 'Msg1_subjectivity',
       'Msg2_subjectivity', 'Msg3_subjectivity', 'Msg4_subjectivity',
       'Msg5_subjectivity', 'Total_similarity', 'Total_distance',
       'Msg1_category', 'Msg2_category', 'Msg3_category', 'Msg4_category',
       'Msg5_category', 'Num_desc_cam', 'Num_desc_NPO', 'Org_causes', 'Cam_c
variables_data4 = extract_data[variable_list4]
```

The Linear regression of selected variables Model 1

In [142…
```
import statsmodels.formula.api as smf

model = smf.ols(formula = 'Actual_Donation_Amount ~ Campaign_Goal  + duratic
    Campaign_Image_num  + Campaign_Video + Num_desc_cam + Num_desc_NPO', dat

results1 = model.summary()
predicts = model._results
print(results1)
```

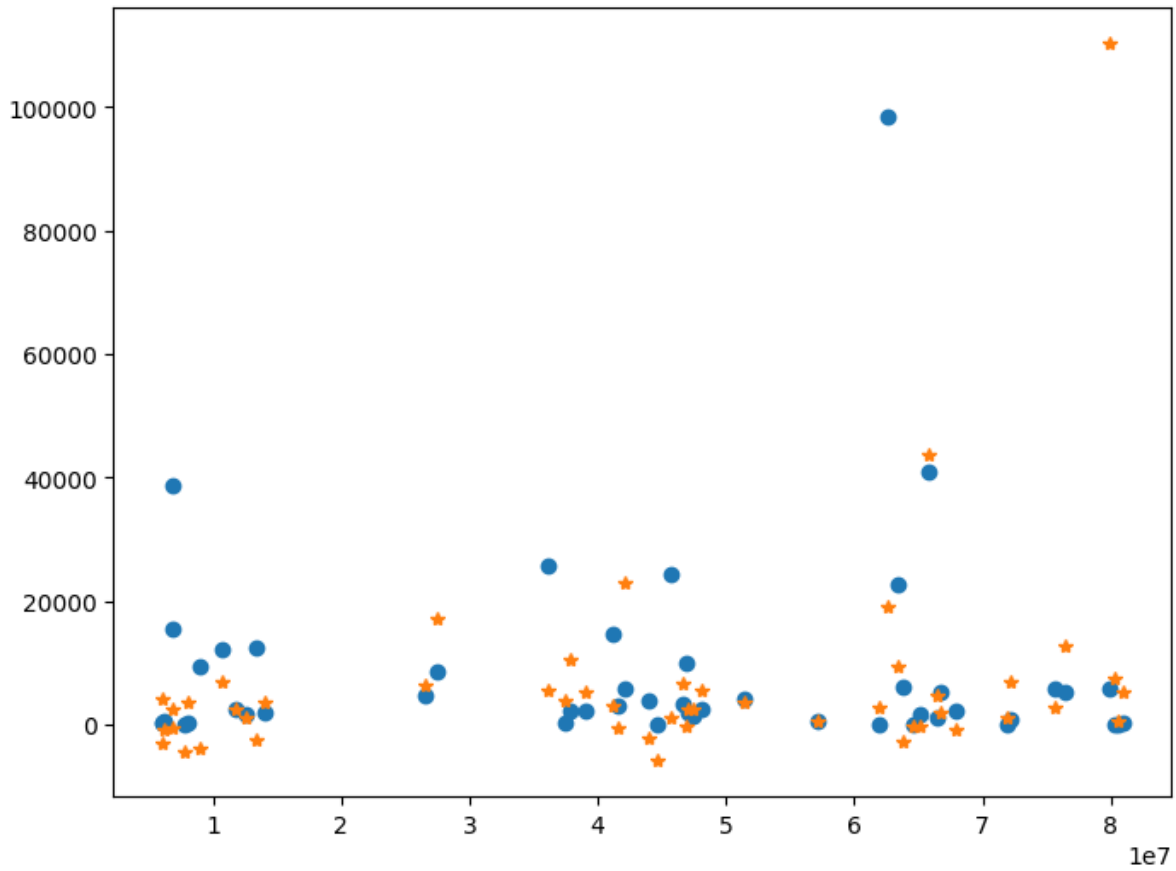
```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:     Actual_Donation_Amount   R-squared:
0.304
Model:                                OLS   Adj. R-squared:
0.304
Method:                     Least Squares   F-statistic:
1165.
Date:                    Tue, 15 Nov 2022   Prob (F-statistic):
0.00
Time:                            13:56:36   Log-Likelihood:            -1.9
635e+05
No. Observations:                   15979   AIC:                         3.
927e+05
Df Residuals:                       15972   BIC:                         3.
928e+05
Df Model:                               6
Covariance Type:                nonrobust
==============================================================================
===========
                         coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
Intercept          -4229.6166   1387.035     -3.049      0.002   -6948.361
-1510.873
Campaign_Goal          0.2259      0.003     82.288      0.000       0.220
0.231
duration_day[0]       -8.0055      1.955     -4.095      0.000     -11.837
-4.174
duration_day[1]       -8.0055      1.955     -4.095      0.000     -11.837
-4.174
Campaign_Image_num  -207.2837    269.294     -0.770      0.441    -735.131
320.564
Campaign_Video      2237.3319    856.557      2.612      0.009     558.384
3916.279
Num_desc_cam          13.3506      6.186      2.158      0.031       1.226
25.476
Num_desc_NPO          34.8610      7.683      4.538      0.000      19.802
49.920
==============================================================================
===
Omnibus:                    35795.486   Durbin-Watson:                   1.
943
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        510585570.
663
Skew:                          20.679   Prob(JB):
0.00
Kurtosis:                     877.743   Cond. No.                     5.79e
+18
==============================================================================
===

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corr
ectly specified.
[2] The smallest eigenvalue is 1.23e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Randomly choose 50 points of prediction and actual data to compare

Circle is actual donation star is regression result

In [143…
```
from random import sample
mysample = sample(range(0,Total_Rows), 50)
x = combined_data['Campaign_ID'][mysample]
y = extract_data['Actual_Donation_Amount'][mysample]
y_fitted = model.fittedvalues
fig, ax = plt.subplots(figsize=(8,6))
ax.plot(x, y, 'o', label='data')
ax.plot(x, y_fitted[mysample],'*',label='OLS')
```

Out[143]: `[<matplotlib.lines.Line2D at 0x7fb42702b4a8>]`



Test normality.

In [144…
```
import openturns as ot
from statsmodels.stats.diagnostic import lilliefors
model_resid = model.resid
result = lilliefors(list(model_resid))
print(result)
```

```
(0.3237611422221752, 0.0009999999999998899)
```

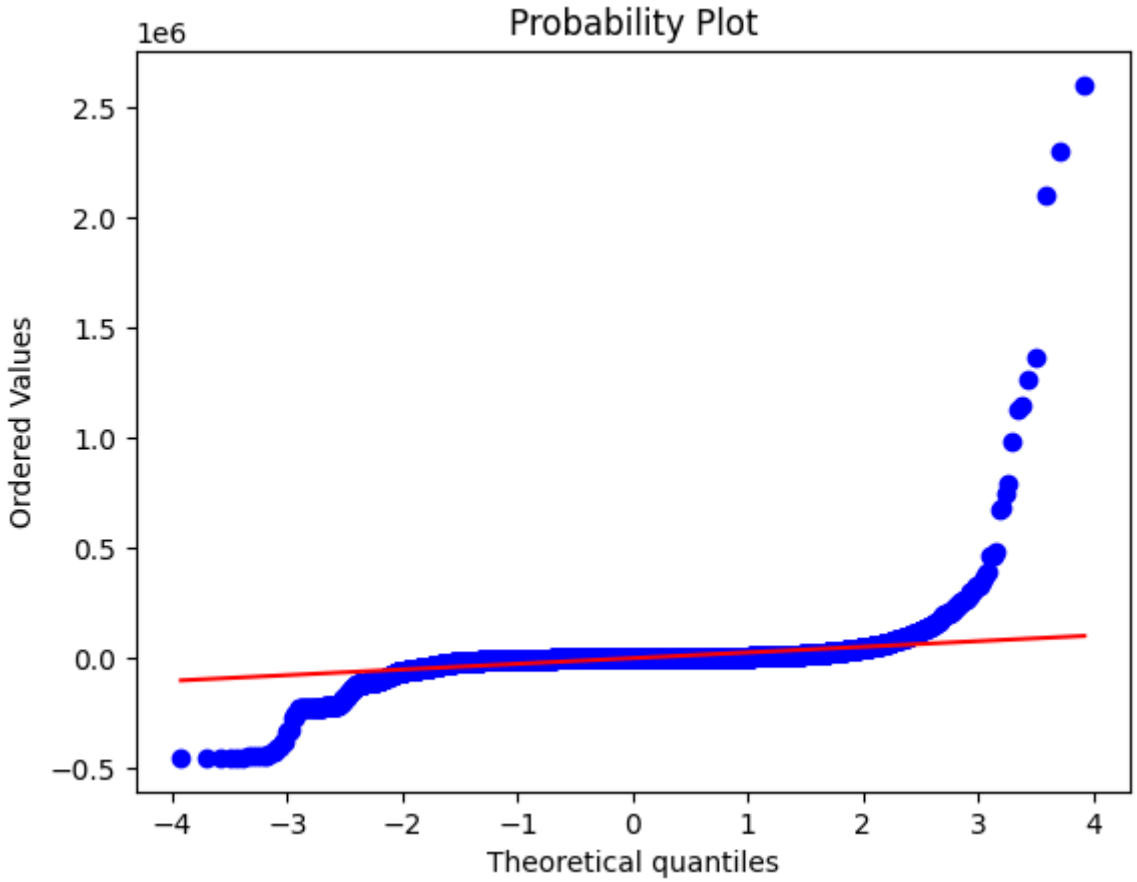
In [145…
```
# Example of the Anderson-Darling Normality Test
from scipy.stats import anderson
result = anderson(list(model_resid), dist='norm')
```

```
print('stat=%.3f' % (result.statistic))
print('significance_level:', (result.significance_level))
```

```
stat=3489.734
significance_level: [15.  10.   5.   2.5  1. ]
```

In [ ]:
```
#stats.probplot(sample, dist=stats.norm, plot=plt)
res = stats.probplot(list(model_resid),  dist=stats.norm, plot=plt)
```

Out[ ]:
```
((array([-3.92492883, -3.70596896, -3.58603376, ...,  3.58603376,
          3.70596896,  3.92492883]),
  array([-452382.65075766, -450608.46965487, -450211.41856847, ...,
         2107564.18449976, 2300161.2985986 , 2604998.28188512])),
 (25739.23427099197, -7.798737455211335e-10, 0.4899510880870473))
```



Model 2

In [149…
```
import statsmodels.formula.api as smf

model2 = smf.ols(formula = 'Actual_Donation_Amount ~ Campaign_Goal  + NPO_Ta
    Campaign_Image_num  + Campaign_Video +\
    Msg1_category + Msg2_category + Msg3_category +Msg4_category + Msg5_cate
     Num_desc_cam + Num_desc_NPO', data = variables_data2).fit()

results2 = model2.summary()
print(results2)
```

OLS Regression Results

Dep. Variable:	Actual_Donation_Amount	R-squared:	0.306
Model:	OLS	Adj. R-squared:	0.305
Method:	Least Squares	F-statistic:	586.4
Date:	Tue, 15 Nov 2022	Prob (F-statistic):	0.00
Time:	13:57:05	Log-Likelihood:	-1.9633e+05
No. Observations:	15979	AIC:	3.927e+05
Df Residuals:	15966	BIC:	3.928e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5727.7576	2160.079	-2.652	0.008	-9961.755	-1493.761
Campaign_Goal	0.2283	0.003	82.129	0.000	0.223	0.234
NPO_Tax_Deductibility	2273.6186	1743.392	1.304	0.192	-1143.626	5690.863
duration_day	-11.9256	3.977	-2.999	0.003	-19.721	-4.130
Campaign_Image_num	-133.5819	269.981	-0.495	0.621	-662.775	395.611
Campaign_Video	1874.0238	859.058	2.181	0.029	190.173	3557.875
Msg1_category	441.2637	1273.095	0.347	0.729	-2054.146	2936.673
Msg2_category	-3803.4040	1722.385	-2.208	0.027	-7179.472	-427.336
Msg3_category	1677.9746	1652.925	1.015	0.310	-1561.944	4917.894
Msg4_category	427.7441	1036.464	0.413	0.680	-1603.841	2459.330
Msg5_category	-996.6393	884.970	-1.126	0.260	-2731.280	738.002
Num_desc_cam	20.4618	6.357	3.219	0.001	8.001	32.922
Num_desc_NPO	33.9582	7.711	4.404	0.000	18.844	49.072

Omnibus:	35698.394	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	504354940.759
Skew:	20.542	Prob(JB):	0.00
Kurtosis:	872.390	Cond. No.	1.02e+06

```
==============================================================================
===

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corr
ectly specified.
[2] The condition number is large, 1.02e+06. This might indicate that there
are
strong multicollinearity or other numerical problems.
```

In [ ]:
```
model_resid2 = model2.resid
result = lilliefors(list(model_resid2))
print(result)
```

```
(0.3288416968646011, 0.0009999999999998899)
```

In [ ]:
```
variables_data3
```

Out[ ]:

	Actual_Donation_Amount	Campaign_Goal	NPO_Status	duration_day	Campaign_Ima
0	5561.0	50000	1.0	252	
1	2810.0	20000	1.0	89	
2	1118.0	30000	1.0	58	
3	2800.0	2000	1.0	88	
4	2030.0	5000	0.0	50	
...	...	...	...	...	
15974	10.0	5000	1.0	62	
15975	150.0	10000	1.0	30	
15976	1000.0	1000	1.0	30	
15977	120.0	3000	1.0	61	
15978	120.0	40000	1.0	117	

15979 rows × 15 columns

Model 3

In [150…
```
model3 = smf.ols(formula = 'Actual_Donation_Amount ~ Campaign_Goal  + NPO_Ta
    Campaign_Image_num  + Campaign_Video +\
    Msg1_category + Msg2_category + Msg3_category +Msg4_category + Msg5_cate
    Num_desc_cam + Num_desc_NPO+ Org_causes + Cam_causes', data = variables_

results3 = model3.summary()
print(results3)
```

OLS Regression Results

Dep. Variable:	Actual_Donation_Amount	R-squared:	0.307
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	506.2
Date:	Tue, 15 Nov 2022	Prob (F-statistic):	0.00
Time:	13:57:09	Log-Likelihood:	-1.9631e+05
No. Observations:	15979	AIC:	3.927e+05
Df Residuals:	15964	BIC:	3.928e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7082.6781	2568.904	-2.757	0.006	-1.21e+04	-2047.336
Campaign_Goal	0.2280	0.003	82.058	0.000	0.223	0.233
NPO_Tax_Deductibility	2543.8432	1749.430	1.454	0.146	-885.236	5972.922
duration_day	-12.6239	3.980	-3.172	0.002	-20.426	-4.822
Campaign_Image_num	-242.2811	271.282	-0.893	0.372	-774.023	289.461
Campaign_Video	2393.4330	866.222	2.763	0.006	695.541	4091.325
Msg1_category	625.4163	1272.373	0.492	0.623	-1868.578	3119.411
Msg2_category	-3815.8630	1720.965	-2.217	0.027	-7189.148	-442.578
Msg3_category	1608.6491	1651.428	0.974	0.330	-1628.335	4845.634
Msg4_category	479.3842	1035.441	0.463	0.643	-1550.196	2508.965
Msg5_category	-1020.4505	884.401	-1.154	0.249	-2753.977	713.076
Num_desc_cam	19.2951	6.353	3.037	0.002	6.842	31.749
Num_desc_NPO	-13.5626	11.289	-1.201	0.230	-35.691	8.566
Org_causes	2428.7210	420.132	5.781	0.000	1605.214	3252.228
Cam_causes	-354.9354	453.246	-0.783	0.434	-1243.348	533.478

Omnibus:	35711.055	Durbin-Watson:	1.946
Prob(Omnibus):	0.000	Jarque-Bera (JB):	505347367.924

```
Skew:                          20.559   Prob(JB):
0.00
Kurtosis:                     873.245   Cond. No.                         1.10e
+06
==============================================================================
===

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corr
ectly specified.
[2] The condition number is large, 1.1e+06. This might indicate that there
are
strong multicollinearity or other numerical problems.
```

In [ ]:
```
model.resid
```

Out[ ]:
```
0       -7370.000019
1        3405.415122
2       -5538.064057
3        2191.834759
4        8623.184623
            ...
15974    2642.841336
15975     388.687552
15976    3265.211928
15977   -4896.802085
15978   -2027.614579
Length: 15979, dtype: float64
```

The Linear regression with more variables

In [151…
```
model4 = smf.ols(formula = 'Actual_Donation_Amount ~ Campaign_Goal + NPO_Tax
       +Campaign_Image_num+duration_day+Msg1_subjectivity\
       +Msg2_subjectivity+Msg3_subjectivity+Msg4_subjectivity\
       +Msg5_subjectivity+Total_similarity+Total_distance\
       +Msg1_category+ Msg2_category + Msg3_category + Msg4_category +Msg5_c
       + Num_desc_cam + Num_desc_NPO+ Org_causes + Cam_causes', data = varia
results4 = model4.summary()
print(results4)
```

OLS Regression Results

Dep. Variable:	Actual_Donation_Amount	R-squared:	0.316
Model:	OLS	Adj. R-squared:	0.315
Method:	Least Squares	F-statistic:	284.0
Date:	Tue, 15 Nov 2022	Prob (F-statistic):	0.00
Time:	13:57:17	Log-Likelihood:	-1.9621e+05
No. Observations:	15979	AIC:	3.925e+05
Df Residuals:	15952	BIC:	3.927e+05
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8560.7673	6483.220	-1.320	0.187	-2.13e+04	4147.074
Campaign_Goal	0.2328	0.003	83.319	0.000	0.227	0.238
NPO_Tax_Deductibility	2032.4131	1745.649	1.164	0.244	-1389.256	5454.082
Campaign_Video	2373.2585	863.459	2.749	0.006	680.782	4065.735
Total_Msg_polarity	-2281.6259	1078.072	-2.116	0.034	-4394.768	-168.484
Total_Msg_subjectivity	2289.7276	721.070	3.175	0.001	876.349	3703.106
Custom_Amount1	-1.4122	0.239	-5.898	0.000	-1.881	-0.943
Custom_Amount2	-0.0466	0.360	-0.130	0.897	-0.752	0.659
Custom_Amount3	-0.1679	0.330	-0.508	0.611	-0.815	0.480
Custom_Amount4	-0.0746	0.152	-0.490	0.624	-0.373	0.224
Campaign_Image_num	-308.6069	272.193	-1.134	0.257	-842.136	224.923
duration_day	-12.4086	3.967	-3.128	0.002	-20.185	-4.632
Msg1_subjectivity	9050.9147	3210.036	2.820	0.005	2758.882	1.53e+04
Msg2_subjectivity	-1.088e+04	3229.534	-3.369	0.001	-1.72e+04	-4551.087
Msg3_subjectivity	2.16e+04	3382.728	6.384	0.000	1.5e+04	2.82e+04
Msg4_subjectivity	-9273.7575	2935.031	-3.160	0.002	-1.5e+04	-3520.765
Msg5_subjectivity	-8201.2299	2543.022	-3.225	0.001	-1.32e+04	-3216.620
Total_similarity	865.0656	1521.674	0.568	0.570	-2117.587	3847.719

```
Total_distance             6.2347      6.776      0.920      0.358      -7.
047      19.516
Msg1_category            303.0135   1366.038      0.222      0.824   -2374.
576    2980.603
Msg2_category          -3501.9019   1748.831     -2.002      0.045   -6929.
807     -73.996
Msg3_category           -320.8461   1704.144     -0.188      0.851   -3661.
161    3019.469
Msg4_category            152.3777   1221.328      0.125      0.901   -2241.
562    2546.318
Msg5_category            196.5563    950.672      0.207      0.836   -1666.
868    2059.980
Num_desc_cam              18.4832      6.348      2.912      0.004       6.
040      30.926
Num_desc_NPO             -19.5919     11.274     -1.738      0.082     -41.
690       2.506
Org_causes              2424.9798    418.639      5.793      0.000    1604.
401    3245.558
Cam_causes              -477.9566    452.158     -1.057      0.291   -1364.
238     408.325
==============================================================================
===
Omnibus:                    35420.474   Durbin-Watson:                   1.
947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        481501488.
441
Skew:                          20.160   Prob(JB):
0.00
Kurtosis:                     852.456   Cond. No.                     1.04e
+16
==============================================================================
===

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corr
ectly specified.
[2] The smallest eigenvalue is 3.8e-18. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Donation per donor Linear regression result by using selected variables

In [ ]:
```
model = smf.ols(formula = 'Donation_per_donor ~ Campaign_Goal + NPO_Tax_Dedu
       +Campaign_Image_num+duration_day+Msg1_subjectivity\
       +Msg2_subjectivity+Msg3_subjectivity+Msg4_subjectivity\
       +Msg5_subjectivity+Total_similarity+Total_distance\
       +Msg1_category+ Msg2_category + Msg3_category + Msg4_category +Msg5_c
results2 = model.summary()
print(results2)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/compat.py in call_and_wrap_exc(msg, origin, f, *args, **kwargs)
     35     try:
---> 36         return f(*args, **kwargs)
     37     except Exception as e:

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/eval.py in eval(self, expr, source_name, inner_namespace)
    165         return eval(code, {}, VarLookupDict([inner_namespace]
--> 166                                             + self._namespaces))
    167

<string> in <module>

NameError: name 'Total_distance' is not defined

The above exception was the direct cause of the following exception:

PatsyError                                Traceback (most recent call last)
/var/folders/vw/f8nhkr8d497gmh8ytfr1p9jr0000gn/T/ipykernel_5266/320398811.p
y in <module>
      3        +Msg2_subjectivity+Msg3_subjectivity+Msg4_subjectivity\
      4        +Msg5_subjectivity+Total_similarity+Total_distance\
----> 5        +Msg1_category+ Msg2_category + Msg3_category + Msg4_categor
y +Msg5_category', data = variables_data2).fit()
      6 results2 = model.summary()
      7 print(results2)

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/statsmodels/base/model.py in from_formula(cls, formula, data, subset, dr
op_cols, *args, **kwargs)
    199
    200         tmp = handle_formula_data(data, None, formula, depth=eval_e
nv,
--> 201                                   missing=missing)
    202         ((endog, exog), missing_idx, design_info) = tmp
    203         max_endog = cls._formula_max_endog

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/statsmodels/formula/formulatools.py in handle_formula_data(Y, X, formul
a, depth, missing)
     62         if data_util._is_using_pandas(Y, None):
     63             result = dmatrices(formula, Y, depth, return_type='data
frame',
---> 64                                NA_action=na_action)
     65         else:
     66             result = dmatrices(formula, Y, depth, return_type='data
frame',

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/highlevel.py in dmatrices(formula_like, data, eval_env, NA_action,
return_type)
    308     eval_env = EvalEnvironment.capture(eval_env, reference=1)
    309     (lhs, rhs) = _do_highlevel_design(formula_like, data, eval_env,
--> 310                                       NA_action, return_type)
    311     if lhs.shape[1] == 0:
    312         raise PatsyError("model is missing required outcome variabl
es")
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/highlevel.py in _do_highlevel_design(formula_like, data, eval_env,
NA_action, return_type)
    163         return iter([data])
    164     design_infos = _try_incr_builders(formula_like, data_iter_make
r, eval_env,
--> 165                                       NA_action)
    166     if design_infos is not None:
    167         return build_design_matrices(design_infos, data,

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/highlevel.py in _try_incr_builders(formula_like, data_iter_maker,
 eval_env, NA_action)
     68                                       data_iter_maker,
     69                                       eval_env,
---> 70                                       NA_action)
     71     else:
     72         return None

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/build.py in design_matrix_builders(termlists, data_iter_maker, eva
l_env, NA_action)
    694                                                 factor_states,
    695                                                 data_iter_maker,
--> 696                                                 NA_action)
    697     # Now we need the factor infos, which encapsulate the knowledge
of
    698     # how to turn any given factor into a chunk of data:

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/build.py in _examine_factor_types(factors, factor_states, data_ite
r_maker, NA_action)
    441     for data in data_iter_maker():
    442         for factor in list(examine_needed):
--> 443             value = factor.eval(factor_states[factor], data)
    444             if factor in cat_sniffers or guess_categorical(value):
    445                 if factor not in cat_sniffers:

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/eval.py in eval(self, memorize_state, data)
    564         return self._eval(memorize_state["eval_code"],
    565                           memorize_state,
--> 566                           data)
    567
    568     __getstate__ = no_pickling

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/eval.py in _eval(self, code, memorize_state, data)
    549                                     memorize_state["eval_env"].eval,
    550                                     code,
--> 551                                     inner_namespace=inner_namespace)
    552
    553     def memorize_chunk(self, state, which_pass, data):

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/compat.py in call_and_wrap_exc(msg, origin, f, *args, **kwargs)
     41                                     origin)
     42             # Use 'exec' to hide this syntax from the Python 2 pars
er:
---> 43             exec("raise new_exc from e")
     44         else:
```

```
     45             # In python 2, we just let the original exception escap
e -- better

/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packag
es/patsy/compat.py in <module>

PatsyError: Error evaluating factor: NameError: name 'Total_distance' is no
t defined
    Donation_per_donor ~ Campaign_Goal + NPO_Status+Campaign_Video+Total_Ms
g_polarity+Total_Msg_subjectivity+Custom_Amount1+Custom_Amount2+Custom_Amou
nt3+Custom_Amount4       +Campaign_Image_num+duration_day+Msg1_subjectivity
+Msg2_subjectivity+Msg3_subjectivity+Msg4_subjectivity       +Msg5_subjecti
vity+Total_similarity+Total_distance       +Msg1_category+ Msg2_category +
 Msg3_category + Msg4_category +Msg5_category

^^^^^^^^^^^^^^
```

test

In [ ]:
```
import nltk
# nltk.download('punkt')
from textblob import TextBlob
texts=["Thank you",'OK!']
for text in texts:
    blob=TextBlob(text)
    emotion=blob.sentiment
    print(emotion)
```

```
Sentiment(polarity=0.0, subjectivity=0.0)
Sentiment(polarity=0.625, subjectivity=0.5)
```