Just Do it

General Work flow

1. Clean data
2. Fill data
    2.1 Fill 0 into missing data
    2.2 Calculate “duration” and add more columns of date
3. Create Sentiment Analysis variables
    3.1 Calculate “polarity”* and “subjectivity”* of message1-message5
    3.2 Calculate “similarity” *of message1-message5
    3.3 Calculate “distance”* between message1-message5
4. Show statistics information of useful variables
5. Linear regression

As for the detail, I will show you the codes and explanation.

** “Polarity”* item is the positiveness of the text, which is a floating point number in the range of [-1.0, 1.0]
** “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
** “Similarity” item 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.
** “Distance” 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

Question & Problems

1. In the email I notice that “NPC status **(this one ignore first, something wrong with this entry.)”** So I add it in variable. I tried to add it in independent variables I found this variable is helpless to improve the R-Square. You can see the result in codes.
2. I used the “Leon Dataset 3 Nov 2022.xlsx” as data set. I think this data set have more information but there is no “Number of causes the campaign identifies with” in this data set. Could you tell me how to add these information into it or change data set?

1. Descriptive statistics

Table1 Descriptive statistics

Descriptive statistics						
statistics	Number	Median	Min	Max	Variance	Average
Actual_Donation_Amount	15957	1300.0	0	3431670	4788325250	9813
Donation_per_donor	15957	84.3	0	163050	2492432	199
Campaign_Goal	15957	5000	100	5000000	23847804493	44800
Campaign_Video	15957	0	0	1	0.2409	0.4046
Campaign_Image_num	15957	3	0	5	2.524	2.868
duration_day	15957	60	0	630	12089	107
Msg1_category	15957	1	0	2	0.7603	0.8325
Msg2_category	15957	1	0	2	0.7772	0.8326
Msg3_category	15957	1	0	2	0.7762	0.8247
Msg4_category	15957	1	0	2	0.7444	0.7708
Msg5_category	15957	1	0	2	0.6741	0.7680

If need more info about this please tell directly.

2. Correlation matrix

2.1 Correlation between important variables and actual amount donation.

Table2 Correlations Matrix

Correlations Matrix													
	0	1	2	3	4	5	6	7	8	9	10	11	12
0.Actual Donation													
1.Campaign goal	0.47												
2.Duration	0.31	0.43											
3.Images Number	0.19	0.079	0.11										
4.Video	-0.14	-0.04	-0.05	-0.08									
5.impact message1	0.18	0.5	0.25	-0.05	-0.08								
6.impact message2	0.19	0.51	0.26	-0.055	-0.08	0.97							
7.impact message3	0.19	0.5	0.25	-0.05	-0.08	0.95	0.98						
8.impact message4	0.17	0.48	0.23	-0.034	-0.06	0.89	0.92	0.93					
9.impact message5	0.17	0.51	0.23	-0.032	-0.06	0.93	0.92	0.91	0.88				
10.Total_polarity	0.12	0.22	0.21	-0.065	-0.06	0.44	0.45	0.44	0.38	0.42			
11.Total_subjectivity	0.24	0.4	-0.22	0.15	-0.14	0.78	0.79	0.78	0.72	0.76	0.57		
12.Total_similarity	-0.14	-0.45	0.23	0.024	0.034	-0.85	-0.83	-0.82	-0.73	-0.8	-0.45	-0.66	
13.Total_distance	0.16	0.46	0.23	0.0004	-0.04	0.87	0.21	0.86	0.79	0.84	0.48	0.71	-0.97

The video correlation turned out to be negative, which is unexpected totally. As for others, I think they are reasonable.
The same, pls tell me if need more.

2.2 Other correlations

In this picture I add some other control variables into it. You can take it as reference.

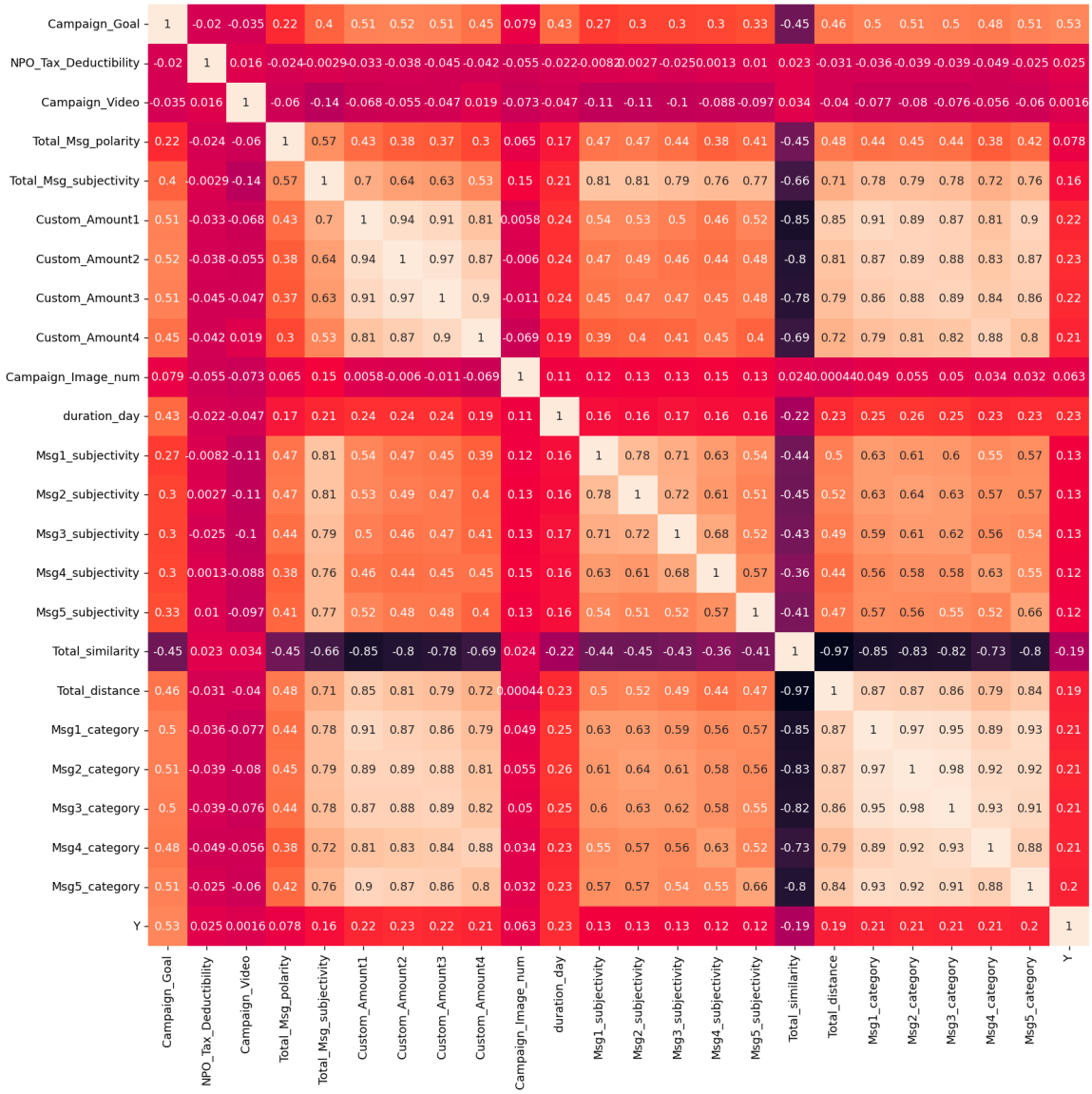

Figure1 Correlations Matrix

3. Linear Regression Result

3.1 The Result of selected independent variables

OLS Regression Results

Dep. Variable:	Actual_Donation_Amount	R-squared:	0.304
Model:	OLS	Adj. R-squared:	0.304
Method:	Least Squares	F-statistic:	775.2
Date:	Sun, 13 Nov 2022	Prob (F-statistic):	0.00
Time:	02:05:02	Log-Likelihood:	-1.9633e+05
No. Observations:	15977	AIC:	3.927e+05
Df Residuals:	15967	BIC:	3.928e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1484.3545	1054.573	1.408	0.159	-582.726	3551.435
Campaign_Goal	0.2282	0.003	81.991	0.000	0.223	0.234
Campaign_Video	1470.4701	852.906	1.724	0.085	-201.322	3142.262
Campaign_Image_num	108.4456	264.359	0.410	0.682	-409.729	626.620
duration_day	-10.4460	3.944	-2.649	0.008	-18.177	-2.715
Msg1_category	-2277.4151	1859.604	-1.225	0.221	-5922.449	1367.619
Msg2_category	1023.2697	2549.303	0.401	0.688	-3973.652	6020.191
Msg3_category	-865.9064	2384.108	-0.363	0.716	-5539.027	3807.214
Msg4_category	-414.7742	1296.364	-0.320	0.749	-2955.794	2126.245
Msg5_category	520.4973	1232.742	0.422	0.673	-1895.816	2936.811

Omnibus:	35657.054	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	500752276.907
Skew:	20.491	Prob(JB):	0.00
Kurtosis:	869.331	Cond. No.	1.24e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.24e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

3.2 The Result of selected + other control variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Actual_Donation_Amount   R-squared:                       0.314
Model:                                OLS   Adj. R-squared:                  0.313
Method:                     Least Squares   F-statistic:                     331.6
Date:                    Sun, 13 Nov 2022   Prob (F-statistic):              0.00
Time:                            02:05:11   Log-Likelihood:            -1.9622e+05
No. Observations:                   15977   AIC:                             3.925e+05
Df Residuals:                       15954   BIC:                             3.927e+05
Df Model:                              22
Covariance Type:                nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    -1619.3993   6606.661     -0.245      0.806   -1.46e+04    1.13e+04
Campaign_Goal                    0.2329      0.003     83.219      0.000       0.227       0.238
NPO_Tax_Deductibility         2244.7877   1735.079      1.294      0.196   -1156.162    5645.737
Campaign_Video                1455.8683    850.688      1.711      0.087    -211.576    3123.313
Total_Msg_polarity           -2424.6686   1079.932     -2.245      0.025   -4541.456    -307.881
Total_Msg_subjectivity        1943.4403    708.242      2.744      0.006     555.207    3331.674
Custom_Amount1                  -1.4295      0.240     -5.955      0.000      -1.900      -0.959
Custom_Amount2                  -0.1177      0.361     -0.326      0.745      -0.826       0.590
Custom_Amount3                  -0.0744      0.330     -0.225      0.822      -0.722       0.573
Custom_Amount4                  -0.1029      0.152     -0.676      0.499      -0.401       0.196
Campaign_Image_num              40.5187    266.697      0.152      0.879    -482.238     563.276
duration_day                   -10.4315      3.927     -2.656      0.008     -18.130      -2.733
Msg1_subjectivity             1.067e+04   3176.407      3.359      0.001    4444.410    1.69e+04
Msg2_subjectivity            -1.124e+04   3233.922     -3.476      0.001   -1.76e+04   -4902.012
Msg3_subjectivity              2.28e+04   3360.905      6.783      0.000    1.62e+04    2.94e+04
Msg4_subjectivity            -1.066e+04   3043.884     -3.504      0.000   -1.66e+04   -4698.320
Msg5_subjectivity            -9617.7378   2506.873     -3.837      0.000   -1.45e+04   -4703.984
Total_similarity               327.5966   1600.996      0.205      0.838   -2810.535    3465.728
Total_distance                   7.2033      6.477      1.112      0.266      -5.492      19.899
Msg1_category                -3406.3712   1939.860     -1.756      0.079   -7208.716     395.974
Msg2_category                 1026.4132   2619.997      0.392      0.695   -4109.075    6161.902
Msg3_category                -4061.3118   2423.154     -1.676      0.094   -8810.966     688.342
Msg4_category                 1008.1028   1627.788      0.619      0.536   -2182.544    4198.750
Msg5_category                 2326.4253   1321.522      1.760      0.078    -263.907    4916.758
==============================================================================
Omnibus:                    35371.049   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        477276588.419
Skew:                          20.099   Prob(JB):                         0.00
Kurtosis:                     848.772   Cond. No.                     3.63e+19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.13e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Actually, the additional variables improve the result a little.

3.3 The regression result of donation of per donor

The R-squared is very low, I want to make sure “donation of per donor = Amount donation/ Distinct Donors” right? If not please tell me.

OLS Regression Results

Dep. Variable:	Donation_per_donor	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	4.780
Date:	Sun, 13 Nov 2022	Prob (F-statistic):	8.99e-13
Time:	02:05:34	Log-Likelihood:	-1.4028e+05
No. Observations:	15977	AIC:	2.806e+05
Df Residuals:	15954	BIC:	2.808e+05
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.2901	199.259	-0.328	0.743	-455.860	325.280
Campaign_Goal	0.0005	8.44e-05	5.550	0.000	0.000	0.001
NPO_Tax_Deductibility	-29.5756	52.331	-0.565	0.572	-132.149	72.998
Campaign_Video	14.3132	25.657	0.558	0.577	-35.977	64.604
Total_Msg_polarity	42.1417	32.571	1.294	0.196	-21.701	105.985
Total_Msg_subjectivity	-30.5281	21.361	-1.429	0.153	-72.398	11.342
Custom_Amount1	0.0073	0.007	1.014	0.311	-0.007	0.022
Custom_Amount2	0.0139	0.011	1.278	0.201	-0.007	0.035
Custom_Amount3	0.0104	0.010	1.049	0.294	-0.009	0.030
Custom_Amount4	-0.0091	0.005	-1.971	0.049	-0.018	-5.1e-05
Campaign_Image_num	22.3573	8.044	2.779	0.005	6.591	38.124
duration_day	0.2883	0.118	2.434	0.015	0.056	0.520
Msg1_subjectivity	-41.4160	95.801	-0.432	0.666	-229.198	146.366
Msg2_subjectivity	25.4180	97.536	0.261	0.794	-165.764	216.600
Msg3_subjectivity	107.0225	101.366	1.056	0.291	-91.666	305.711
Msg4_subjectivity	-39.9107	91.805	-0.435	0.664	-219.858	140.036
Msg5_subjectivity	-81.6419	75.608	-1.080	0.280	-229.842	66.559
Total_similarity	30.4887	48.287	0.631	0.528	-64.158	125.136
Total_distance	-0.2056	0.195	-1.053	0.293	-0.589	0.177
Msg1_category	67.0142	58.507	1.145	0.252	-47.666	181.694
Msg2_category	-17.6539	79.020	-0.223	0.823	-172.542	137.234
Msg3_category	-30.7882	73.083	-0.421	0.674	-174.039	112.463
Msg4_category	-5.4594	49.095	-0.111	0.911	-101.690	90.772
Msg5_category	97.5549	39.858	2.448	0.014	19.430	175.680

Omnibus:	58192.128	Durbin-Watson:	1.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36028162900.975
Skew:	76.283	Prob(JB):	0.00
Kurtosis:	7358.050	Cond. No.	3.63e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.13e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.