module_04

April 12, 2023

[1]:
```
config_path = 'config.yaml'
```

1 packages

[2]:
```
import pandas as pd
import plotly.express as px
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import yaml
from IPython.display import HTML
import matplotlib.pyplot as plt
from ipywidgets import Combobox, HBox, Checkbox, interactive_output
import pdb
```

2 loading the config

[3]:
```
config = yaml.safe_load(open(config_path, 'rb'))
```

[4]:
```
fit_intercept = config['fit_intercept']
alpha = config['chart_alpha']
```

3 preparing column names

[5]:
```
lines = open('data/spambase.names').readlines()[33:]
colnames = [x.split(':')[0] for x in lines]
colnames.append('spam_class')
```

[6]:
```
colnames
```

[6]:
```
['word_freq_make',
 'word_freq_address',
 'word_freq_all',
 'word_freq_3d',
```

```
'word_freq_our',
'word_freq_over',
'word_freq_remove',
'word_freq_internet',
'word_freq_order',
'word_freq_mail',
'word_freq_receive',
'word_freq_will',
'word_freq_people',
'word_freq_report',
'word_freq_addresses',
'word_freq_free',
'word_freq_business',
'word_freq_email',
'word_freq_you',
'word_freq_credit',
'word_freq_your',
'word_freq_font',
'word_freq_000',
'word_freq_money',
'word_freq_hp',
'word_freq_hpl',
'word_freq_george',
'word_freq_650',
'word_freq_lab',
'word_freq_labs',
'word_freq_telnet',
'word_freq_857',
'word_freq_data',
'word_freq_415',
'word_freq_85',
'word_freq_technology',
'word_freq_1999',
'word_freq_parts',
'word_freq_pm',
'word_freq_direct',
'word_freq_cs',
'word_freq_meeting',
'word_freq_original',
'word_freq_project',
'word_freq_re',
'word_freq_edu',
'word_freq_table',
'word_freq_conference',
'char_freq_;',
'char_freq_(',
'char_freq_[',
```

```
 'char_freq_!',
 'char_freq_$',
 'char_freq_#',
 'capital_run_length_average',
 'capital_run_length_longest',
 'capital_run_length_total',
 'spam_class']
```

4 dataset analysis

4.1 reading the dataset

[7]:
```
df = pd.read_csv('data/spambase.data', header = None)
df.columns = colnames
```

[8]:
```
df.head()
```

[8]:

	word_freq_make	word_freq_address	word_freq_all	word_freq_3d	\
0	0.00	0.64	0.64	0.0	
1	0.21	0.28	0.50	0.0	
2	0.06	0.00	0.71	0.0	
3	0.00	0.00	0.00	0.0	
4	0.00	0.00	0.00	0.0	

	word_freq_our	word_freq_over	word_freq_remove	word_freq_internet	\
0	0.32	0.00	0.00	0.00	
1	0.14	0.28	0.21	0.07	
2	1.23	0.19	0.19	0.12	
3	0.63	0.00	0.31	0.63	
4	0.63	0.00	0.31	0.63	

	word_freq_order	word_freq_mail	…	char_freq_;	char_freq_(	\
0	0.00	0.00	…	0.00	0.000	
1	0.00	0.94	…	0.00	0.132	
2	0.64	0.25	…	0.01	0.143	
3	0.31	0.63	…	0.00	0.137	
4	0.31	0.63	…	0.00	0.135	

	char_freq_[	char_freq_!	char_freq_$	char_freq_#	\
0	0.0	0.778	0.000	0.000	
1	0.0	0.372	0.180	0.048	
2	0.0	0.276	0.184	0.010	
3	0.0	0.137	0.000	0.000	
4	0.0	0.135	0.000	0.000	

	capital_run_length_average	capital_run_length_longest	\

```
0                        3.756                          61
1                        5.114                         101
2                        9.821                         485
3                        3.537                          40
4                        3.537                          40

   capital_run_length_total  spam_class
0                       278           1
1                      1028           1
2                      2259           1
3                       191           1
4                       191           1

[5 rows x 58 columns]
```

[9]:
```
df.shape
```

[9]:
```
(4601, 58)
```

4.2 correlation matrix

[10]:
```
df_corr = df.corr()
df_corr
```

[10]:
```
                          word_freq_make  word_freq_address  word_freq_all  \
word_freq_make                  1.000000          -0.016759       0.065627
word_freq_address              -0.016759           1.000000      -0.033526
word_freq_all                   0.065627          -0.033526       1.000000
word_freq_3d                    0.013273          -0.006923      -0.020246
word_freq_our                   0.023119          -0.023760       0.077734
word_freq_over                  0.059674          -0.024840       0.087564
word_freq_remove                0.007669           0.003918       0.036677
word_freq_internet             -0.003950          -0.016280       0.012003
word_freq_order                 0.106263          -0.003826       0.093786
word_freq_mail                  0.041198           0.032962       0.032075
word_freq_receive               0.188459          -0.006864       0.048254
word_freq_will                  0.105801          -0.040398       0.083210
word_freq_people                0.066438          -0.018858       0.047593
word_freq_report                0.036780          -0.009206       0.008552
word_freq_addresses             0.028439           0.005330       0.122113
word_freq_free                  0.059386          -0.009117       0.063906
word_freq_business              0.081928          -0.018370       0.036262
word_freq_email                 0.053324           0.033500       0.121923
word_freq_you                   0.128243          -0.055476       0.139329
word_freq_credit                0.021295          -0.015806       0.031111
word_freq_your                  0.197049          -0.018191       0.156651
word_freq_font                 -0.024349          -0.008850      -0.035681
```

word_freq_000	0.134072	-0.020502	0.123671
word_freq_money	0.188155	0.001984	0.041145
word_freq_hp	-0.072504	-0.043483	-0.087924
word_freq_hpl	-0.061686	-0.038211	-0.062459
word_freq_george	-0.066424	-0.030307	-0.108886
word_freq_650	-0.048680	-0.029221	-0.050648
word_freq_lab	-0.041251	-0.021940	-0.057726
word_freq_labs	-0.052799	-0.027508	-0.032547
word_freq_telnet	-0.039066	-0.018097	-0.038927
word_freq_857	-0.032058	-0.003326	-0.061870
word_freq_data	-0.041014	-0.024903	-0.054759
word_freq_415	-0.027690	-0.004303	-0.061706
word_freq_85	-0.044954	-0.024058	-0.048335
word_freq_technology	-0.054673	-0.028198	-0.046504
word_freq_1999	-0.057312	-0.024013	-0.067015
word_freq_parts	-0.007960	-0.008922	0.032407
word_freq_pm	-0.011134	-0.019124	-0.014809
word_freq_direct	-0.036095	-0.014821	-0.047066
word_freq_cs	-0.009703	-0.015420	-0.030956
word_freq_meeting	-0.026070	-0.025177	-0.005811
word_freq_original	-0.024292	-0.002370	-0.044325
word_freq_project	-0.022116	-0.019739	-0.053464
word_freq_re	-0.037105	-0.016418	-0.050664
word_freq_edu	-0.034056	-0.023858	-0.056655
word_freq_table	-0.000953	-0.009818	0.029339
word_freq_conference	-0.017755	-0.015747	-0.026344
char_freq_;	-0.026505	-0.007282	-0.033213
char_freq_(	-0.021196	-0.049837	-0.016495
char_freq_[	-0.033301	-0.018527	-0.033120
char_freq_!	0.058292	-0.014461	0.108140
char_freq_$	0.117419	-0.009605	0.087618
char_freq_#	-0.008844	0.001946	-0.003336
capital_run_length_average	0.044491	0.002083	0.097398
capital_run_length_longest	0.061382	0.000271	0.107463
capital_run_length_total	0.089165	-0.022680	0.070114
spam_class	0.126208	-0.030224	0.196988

	word_freq_3d	word_freq_our	word_freq_over	\
word_freq_make	0.013273	0.023119	0.059674	
word_freq_address	-0.006923	-0.023760	-0.024840	
word_freq_all	-0.020246	0.077734	0.087564	
word_freq_3d	1.000000	0.003238	-0.010014	
word_freq_our	0.003238	1.000000	0.054054	
word_freq_over	-0.010014	0.054054	1.000000	
word_freq_remove	0.019784	0.147336	0.061163	
word_freq_internet	0.010268	0.029598	0.079561	
word_freq_order	-0.002454	0.020823	0.117438	

word_freq_mail	-0.004947	0.034495	0.013897
word_freq_receive	-0.012976	0.068382	0.053900
word_freq_will	-0.019221	0.066788	0.009264
word_freq_people	-0.013199	0.031126	0.077631
word_freq_report	0.012008	0.003445	0.009673
word_freq_addresses	0.002707	0.056177	0.173066
word_freq_free	0.007432	0.083024	0.019865
word_freq_business	0.003470	0.143443	0.064137
word_freq_email	0.019391	0.062344	0.078350
word_freq_you	-0.010834	0.098510	0.095505
word_freq_credit	-0.005381	0.031526	0.058979
word_freq_your	0.008176	0.136605	0.106833
word_freq_font	0.028102	-0.020207	0.007956
word_freq_000	0.011368	0.070037	0.211455
word_freq_money	0.035360	0.000039	0.059329
word_freq_hp	-0.015181	-0.072502	-0.084402
word_freq_hpl	-0.013708	-0.075456	-0.087271
word_freq_george	-0.010684	-0.088011	-0.069051
word_freq_650	-0.010368	-0.061501	-0.066223
word_freq_lab	-0.007798	0.032048	-0.048673
word_freq_labs	-0.010476	-0.052066	-0.048127
word_freq_telnet	-0.007529	-0.042535	-0.046383
word_freq_857	-0.006717	-0.026748	-0.036835
word_freq_data	-0.008075	-0.031998	-0.034164
word_freq_415	-0.006729	-0.026960	-0.037315
word_freq_85	-0.006122	-0.049732	-0.054315
word_freq_technology	-0.006515	-0.048844	-0.052819
word_freq_1999	-0.007761	-0.072599	-0.057465
word_freq_parts	-0.002669	0.130812	-0.017918
word_freq_pm	-0.004602	-0.042044	-0.047619
word_freq_direct	-0.007643	-0.021442	-0.029866
word_freq_cs	-0.005670	-0.047505	-0.029457
word_freq_meeting	-0.008095	0.115041	-0.054812
word_freq_original	-0.009268	-0.048879	-0.030616
word_freq_project	-0.005933	0.015234	-0.028826
word_freq_re	-0.012957	-0.042336	-0.053637
word_freq_edu	-0.009181	-0.077986	-0.033046
word_freq_table	-0.003348	-0.026900	-0.014343
word_freq_conference	-0.001924	-0.032005	-0.031693
char_freq_;	-0.000591	-0.032759	-0.019119
char_freq_(	-0.012370	-0.046361	-0.008705
char_freq_[	-0.007148	-0.026390	-0.015133
char_freq_!	-0.003138	0.025509	0.065043
char_freq_$	0.010862	0.041582	0.105692
char_freq_#	-0.000298	0.002016	0.019894
capital_run_length_average	0.005260	0.052662	-0.010278
capital_run_length_longest	0.022081	0.052290	0.090172

```
capital_run_length_total        0.021369       0.002492        0.082089
spam_class                      0.057371       0.241920        0.232604

                          word_freq_remove  word_freq_internet  \
word_freq_make                    0.007669           -0.003950
word_freq_address                 0.003918           -0.016280
word_freq_all                     0.036677            0.012003
word_freq_3d                      0.019784            0.010268
word_freq_our                     0.147336            0.029598
word_freq_over                    0.061163            0.079561
word_freq_remove                  1.000000            0.044545
word_freq_internet                0.044545            1.000000
word_freq_order                   0.050786            0.105302
word_freq_mail                    0.056809            0.083129
word_freq_receive                 0.159578            0.128495
word_freq_will                   -0.001461           -0.002973
word_freq_people                  0.013295            0.026274
word_freq_report                 -0.022723            0.012426
word_freq_addresses               0.042904            0.072782
word_freq_free                    0.128436            0.051115
word_freq_business                0.187981            0.216422
word_freq_email                   0.122011            0.037738
word_freq_you                     0.111792            0.020641
word_freq_credit                  0.046134            0.109163
word_freq_your                    0.130794            0.156905
word_freq_font                   -0.002093           -0.016192
word_freq_000                     0.064795            0.089226
word_freq_money                   0.030575            0.034127
word_freq_hp                     -0.089494           -0.053038
word_freq_hpl                    -0.080330           -0.041450
word_freq_george                 -0.065893           -0.057189
word_freq_650                    -0.066947           -0.049988
word_freq_lab                    -0.048482           -0.037047
word_freq_labs                   -0.058101           -0.043405
word_freq_telnet                 -0.046280           -0.035816
word_freq_857                    -0.040538           -0.034276
word_freq_data                   -0.041372           -0.039220
word_freq_415                    -0.040910           -0.034811
word_freq_85                     -0.053202           -0.035174
word_freq_technology             -0.053978           -0.033747
word_freq_1999                   -0.052035           -0.017466
word_freq_parts                  -0.014781           -0.012119
word_freq_pm                     -0.046978           -0.030392
word_freq_direct                 -0.022121           -0.005988
word_freq_cs                     -0.033120           -0.003884
word_freq_meeting                -0.049664           -0.043626
word_freq_original               -0.049079           -0.004542
```

word_freq_project	-0.034461	-0.030134
word_freq_re	-0.050811	-0.002423
word_freq_edu	-0.056166	-0.037916
word_freq_table	-0.017512	-0.006397
word_freq_conference	-0.031408	-0.021224
char_freq_;	-0.033089	-0.027432
char_freq_(	-0.051885	-0.032494
char_freq_[	-0.027653	-0.019548
char_freq_!	0.053706	0.031454
char_freq_$	0.070127	0.057910
char_freq_#	0.046612	-0.008012
capital_run_length_average	0.041565	0.011254
capital_run_length_longest	0.059677	0.037575
capital_run_length_total	-0.008344	0.040252
spam_class	0.332117	0.206808

	word_freq_order	word_freq_mail	…	char_freq_;	\
word_freq_make	0.106263	0.041198	…	-0.026505	
word_freq_address	-0.003826	0.032962	…	-0.007282	
word_freq_all	0.093786	0.032075	…	-0.033213	
word_freq_3d	-0.002454	-0.004947	…	-0.000591	
word_freq_our	0.020823	0.034495	…	-0.032759	
word_freq_over	0.117438	0.013897	…	-0.019119	
word_freq_remove	0.050786	0.056809	…	-0.033089	
word_freq_internet	0.105302	0.083129	…	-0.027432	
word_freq_order	1.000000	0.130624	…	-0.014646	
word_freq_mail	0.130624	1.000000	…	0.011945	
word_freq_receive	0.137760	0.125319	…	-0.032410	
word_freq_will	0.030344	0.071157	…	-0.027711	
word_freq_people	0.034738	0.045737	…	-0.023445	
word_freq_report	0.066840	0.017901	…	-0.019045	
word_freq_addresses	0.238436	0.160543	…	-0.018277	
word_freq_free	0.008269	0.025601	…	-0.026841	
word_freq_business	0.158390	0.081363	…	-0.031542	
word_freq_email	0.098804	0.035977	…	-0.039519	
word_freq_you	0.039017	0.093509	…	-0.044314	
word_freq_credit	0.123217	0.030859	…	-0.020851	
word_freq_your	0.159112	0.098072	…	-0.058660	
word_freq_font	-0.019648	0.008200	…	0.416608	
word_freq_000	0.126800	0.096809	…	-0.027362	
word_freq_money	0.099461	0.052129	…	-0.019139	
word_freq_hp	-0.069931	-0.033534	…	0.029181	
word_freq_hpl	-0.049775	-0.013045	…	0.013558	
word_freq_george	-0.064608	-0.067817	…	-0.022724	
word_freq_650	-0.056764	0.019356	…	-0.025020	
word_freq_lab	-0.044840	-0.026903	…	-0.018502	
word_freq_labs	-0.043643	0.008677	…	-0.019845	

word_freq_telnet	-0.040158	-0.024423	…	-0.016280
word_freq_857	-0.033984	-0.015137	…	-0.008853
word_freq_data	-0.014403	-0.035366	…	-0.005691
word_freq_415	-0.033601	-0.014434	…	-0.009290
word_freq_85	-0.041847	-0.020092	…	-0.021592
word_freq_technology	-0.056270	-0.016955	…	-0.018947
word_freq_1999	-0.033244	-0.004944	…	0.052138
word_freq_parts	-0.002216	-0.017950	…	0.007886
word_freq_pm	-0.040844	-0.016091	…	0.034492
word_freq_direct	-0.009867	0.004163	…	-0.018693
word_freq_cs	-0.035177	-0.025084	…	0.053034
word_freq_meeting	-0.048223	-0.054467	…	-0.007817
word_freq_original	-0.034190	0.023200	…	0.015385
word_freq_project	-0.035159	-0.026654	…	-0.007257
word_freq_re	-0.075558	-0.032065	…	-0.024698
word_freq_edu	-0.056817	-0.030326	…	0.015382
word_freq_table	0.007521	-0.015546	…	0.000995
word_freq_conference	-0.026017	-0.016842	…	-0.002290
char_freq_;	-0.014646	0.011945	…	1.000000
char_freq_(	-0.031003	0.003936	…	0.049124
char_freq_[	0.013601	0.007357	…	0.009070
char_freq_!	0.043639	0.036737	…	0.020539
char_freq_$	0.149365	0.075786	…	0.006392
char_freq_#	-0.000522	0.044830	…	0.055057
capital_run_length_average	0.111308	0.073677	…	0.003443
capital_run_length_longest	0.189247	0.103308	…	0.040829
capital_run_length_total	0.248724	0.087273	…	0.055298
spam_class	0.231551	0.138962	…	-0.059630

	char_freq_(	char_freq_[	char_freq_!	\
word_freq_make	-0.021196	-0.033301	0.058292	
word_freq_address	-0.049837	-0.018527	-0.014461	
word_freq_all	-0.016495	-0.033120	0.108140	
word_freq_3d	-0.012370	-0.007148	-0.003138	
word_freq_our	-0.046361	-0.026390	0.025509	
word_freq_over	-0.008705	-0.015133	0.065043	
word_freq_remove	-0.051885	-0.027653	0.053706	
word_freq_internet	-0.032494	-0.019548	0.031454	
word_freq_order	-0.031003	0.013601	0.043639	
word_freq_mail	0.003936	0.007357	0.036737	
word_freq_receive	-0.055089	-0.025183	0.024992	
word_freq_will	-0.030940	-0.044966	0.013369	
word_freq_people	-0.051151	-0.028283	0.040737	
word_freq_report	-0.005804	-0.014349	-0.008499	
word_freq_addresses	-0.002551	-0.003111	0.018607	
word_freq_free	-0.046578	-0.029560	0.104261	
word_freq_business	-0.035897	-0.036691	0.077049	

```
word_freq_email               -0.035897    -0.017439     0.039350
word_freq_you                 -0.128882    -0.063826     0.153381
word_freq_credit              -0.021431    -0.012071     0.048350
word_freq_your                -0.085181    -0.045469     0.084017
word_freq_font                -0.046244    -0.001137    -0.004838
word_freq_000                 -0.033174    -0.000467     0.070103
word_freq_money               -0.033113    -0.020798     0.051076
word_freq_hp                   0.136979     0.039723    -0.090862
word_freq_hpl                  0.144771     0.064349    -0.078367
word_freq_george              -0.028748    -0.017676    -0.067500
word_freq_650                  0.313835     0.031979    -0.063495
word_freq_lab                  0.158593     0.006575    -0.042330
word_freq_labs                 0.224192     0.004667    -0.061694
word_freq_telnet               0.233392     0.010718    -0.045273
word_freq_857                  0.304679     0.013805    -0.041529
word_freq_data                 0.028655     0.113105    -0.048493
word_freq_415                  0.303606     0.013687    -0.038626
word_freq_85                   0.200713     0.034435    -0.048822
word_freq_technology           0.245454     0.001017    -0.060379
word_freq_1999                 0.107674     0.073010    -0.054578
word_freq_parts               -0.010430     0.001767    -0.015126
word_freq_pm                   0.107928     0.038474    -0.024846
word_freq_direct               0.268701     0.014065    -0.032509
word_freq_cs                   0.017584     0.034408    -0.025911
word_freq_meeting             -0.013082     0.011091    -0.038094
word_freq_original             0.060568     0.115548    -0.049362
word_freq_project             -0.003203    -0.010733    -0.033837
word_freq_re                   0.001413     0.008838     0.067569
word_freq_edu                  0.014763    -0.003168    -0.028845
word_freq_table               -0.003085    -0.004592    -0.017679
word_freq_conference          -0.012795    -0.006310    -0.026576
char_freq_;                    0.049124     0.009070     0.020539
char_freq_(                    1.000000     0.022316    -0.030354
char_freq_[                    0.022316     1.000000    -0.031769
char_freq_!                   -0.030354    -0.031769     1.000000
char_freq_$                    0.044722    -0.026400     0.142913
char_freq_#                    0.023322    -0.006863     0.020924
capital_run_length_average     0.034365    -0.008180     0.054308
capital_run_length_longest     0.370963    -0.013994     0.077392
capital_run_length_total       0.112209     0.006016     0.036321
spam_class                    -0.089672    -0.064709     0.241888

                           char_freq_$  char_freq_#  \
word_freq_make                0.117419    -0.008844
word_freq_address            -0.009605     0.001946
word_freq_all                 0.087618    -0.003336
word_freq_3d                  0.010862    -0.000298
```

word_freq_our	0.041582	0.002016
word_freq_over	0.105692	0.019894
word_freq_remove	0.070127	0.046612
word_freq_internet	0.057910	-0.008012
word_freq_order	0.149365	-0.000522
word_freq_mail	0.075786	0.044830
word_freq_receive	0.070227	0.001126
word_freq_will	0.016723	-0.030445
word_freq_people	0.205905	-0.014195
word_freq_report	0.080953	0.006545
word_freq_addresses	0.123854	-0.005446
word_freq_free	0.049953	0.035534
word_freq_business	0.098323	-0.000466
word_freq_email	0.063872	0.020978
word_freq_you	0.091470	-0.002434
word_freq_credit	0.034948	0.007214
word_freq_your	0.141649	-0.004355
word_freq_font	-0.011036	0.184428
word_freq_000	0.310971	0.020140
word_freq_money	0.104691	0.000703
word_freq_hp	-0.086634	0.058780
word_freq_hpl	-0.081198	-0.020691
word_freq_george	-0.068728	-0.020561
word_freq_650	-0.061441	-0.011438
word_freq_lab	-0.050231	0.002076
word_freq_labs	-0.065475	0.082593
word_freq_telnet	-0.047475	0.000225
word_freq_857	-0.043484	-0.010735
word_freq_data	-0.048101	-0.009928
word_freq_415	-0.039844	-0.010635
word_freq_85	-0.048947	-0.009650
word_freq_technology	-0.057933	0.006452
word_freq_1999	-0.063895	-0.022637
word_freq_parts	-0.012909	-0.003627
word_freq_pm	-0.044513	-0.011326
word_freq_direct	-0.016724	-0.010661
word_freq_cs	-0.036610	-0.011755
word_freq_meeting	-0.043653	-0.003873
word_freq_original	-0.054698	-0.013925
word_freq_project	-0.036241	0.001167
word_freq_re	-0.049367	-0.023878
word_freq_edu	-0.050109	-0.015040
word_freq_table	-0.018549	0.000308
word_freq_conference	-0.030751	-0.008575
char_freq_;	0.006392	0.055057
char_freq_(	0.044722	0.023322
char_freq_[	-0.026400	-0.006863

char_freq_!	0.142913	0.020924
char_freq_$	1.000000	0.012613
char_freq_#	0.012613	1.000000
capital_run_length_average	0.079998	0.013497
capital_run_length_longest	0.183144	0.061657
capital_run_length_total	0.201948	0.042568
spam_class	0.323629	0.065067

	capital_run_length_average \
word_freq_make	0.044491
word_freq_address	0.002083
word_freq_all	0.097398
word_freq_3d	0.005260
word_freq_our	0.052662
word_freq_over	-0.010278
word_freq_remove	0.041565
word_freq_internet	0.011254
word_freq_order	0.111308
word_freq_mail	0.073677
word_freq_receive	0.029258
word_freq_will	-0.010002
word_freq_people	-0.013446
word_freq_report	0.003023
word_freq_addresses	0.017383
word_freq_free	0.015036
word_freq_business	0.038126
word_freq_email	-0.007979
word_freq_you	-0.030592
word_freq_credit	0.067140
word_freq_your	0.041066
word_freq_font	0.021497
word_freq_000	0.008372
word_freq_money	0.007681
word_freq_hp	-0.017285
word_freq_hpl	-0.024234
word_freq_george	-0.025504
word_freq_650	-0.013757
word_freq_lab	-0.014936
word_freq_labs	-0.016599
word_freq_telnet	-0.010897
word_freq_857	-0.010498
word_freq_data	-0.015509
word_freq_415	-0.002404
word_freq_85	-0.013707
word_freq_technology	-0.019185
word_freq_1999	-0.014423
word_freq_parts	-0.006012

```
word_freq_pm                                -0.014032
word_freq_direct                            -0.003945
word_freq_cs                                -0.008895
word_freq_meeting                           -0.017899
word_freq_original                          -0.017681
word_freq_project                           -0.013157
word_freq_re                                -0.026979
word_freq_edu                               -0.017408
word_freq_table                             -0.006465
word_freq_conference                        -0.008114
char_freq_;                                  0.003443
char_freq_(                                  0.034365
char_freq_[                                 -0.008180
char_freq_!                                  0.054308
char_freq_$                                  0.079998
char_freq_#                                  0.013497
capital_run_length_average                   1.000000
capital_run_length_longest                   0.492638
capital_run_length_total                     0.162314
spam_class                                   0.109999

                           capital_run_length_longest  \
word_freq_make                               0.061382
word_freq_address                            0.000271
word_freq_all                                0.107463
word_freq_3d                                 0.022081
word_freq_our                                0.052290
word_freq_over                               0.090172
word_freq_remove                             0.059677
word_freq_internet                           0.037575
word_freq_order                              0.189247
word_freq_mail                               0.103308
word_freq_receive                            0.086791
word_freq_will                               0.021774
word_freq_people                             0.041962
word_freq_report                             0.060993
word_freq_addresses                          0.213992
word_freq_free                               0.026528
word_freq_business                           0.062672
word_freq_email                              0.075122
word_freq_you                                0.006530
word_freq_credit                             0.099463
word_freq_your                               0.085321
word_freq_font                               0.027775
word_freq_000                                0.123036
word_freq_money                              0.044870
word_freq_hp                                -0.051206
```

word_freq_hpl	-0.051806
word_freq_george	-0.054400
word_freq_650	-0.038772
word_freq_lab	-0.034733
word_freq_labs	-0.039001
word_freq_telnet	-0.027449
word_freq_857	-0.027732
word_freq_data	-0.025919
word_freq_415	-0.024532
word_freq_85	-0.030236
word_freq_technology	-0.038100
word_freq_1999	-0.033204
word_freq_parts	-0.009487
word_freq_pm	-0.029229
word_freq_direct	-0.004835
word_freq_cs	-0.023658
word_freq_meeting	-0.034585
word_freq_original	-0.017279
word_freq_project	-0.025918
word_freq_re	-0.051858
word_freq_edu	-0.033365
word_freq_table	-0.010154
word_freq_conference	-0.016894
char_freq_;	0.040829
char_freq_(	0.370963
char_freq_[	-0.013994
char_freq_!	0.077392
char_freq_$	0.183144
char_freq_#	0.061657
capital_run_length_average	0.492638
capital_run_length_longest	1.000000
capital_run_length_total	0.475486
spam_class	0.216097

	capital_run_length_total	spam_class
word_freq_make	0.089165	0.126208
word_freq_address	-0.022680	-0.030224
word_freq_all	0.070114	0.196988
word_freq_3d	0.021369	0.057371
word_freq_our	0.002492	0.241920
word_freq_over	0.082089	0.232604
word_freq_remove	-0.008344	0.332117
word_freq_internet	0.040252	0.206808
word_freq_order	0.248724	0.231551
word_freq_mail	0.087273	0.138962
word_freq_receive	0.115055	0.234529
word_freq_will	0.020076	0.007741

word_freq_people	0.105150	0.132927
word_freq_report	0.169257	0.060027
word_freq_addresses	0.151626	0.195902
word_freq_free	0.003007	0.263215
word_freq_business	0.064261	0.263204
word_freq_email	0.046364	0.204208
word_freq_you	-0.007307	0.273651
word_freq_credit	0.075751	0.189761
word_freq_your	0.051797	0.383234
word_freq_font	0.103954	0.091860
word_freq_000	0.165977	0.334787
word_freq_money	0.080993	0.216111
word_freq_hp	-0.043267	-0.256723
word_freq_hpl	-0.059601	-0.232968
word_freq_george	-0.096548	-0.183404
word_freq_650	-0.067596	-0.158800
word_freq_lab	-0.056628	-0.133523
word_freq_labs	-0.064115	-0.171095
word_freq_telnet	-0.045923	-0.126912
word_freq_857	-0.046796	-0.114214
word_freq_data	0.006919	-0.119931
word_freq_415	-0.044529	-0.112754
word_freq_85	-0.045963	-0.149225
word_freq_technology	-0.045792	-0.136134
word_freq_1999	-0.003490	-0.178045
word_freq_parts	-0.013897	-0.031035
word_freq_pm	-0.049256	-0.122831
word_freq_direct	-0.028806	-0.064801
word_freq_cs	-0.026373	-0.097375
word_freq_meeting	-0.056511	-0.136615
word_freq_original	-0.036529	-0.135664
word_freq_project	-0.040661	-0.094594
word_freq_re	-0.095444	-0.140408
word_freq_edu	-0.046371	-0.146138
word_freq_table	0.005158	-0.044679
word_freq_conference	-0.010033	-0.084020
char_freq_;	0.055298	-0.059630
char_freq_(	0.112209	-0.089672
char_freq_[	0.006016	-0.064709
char_freq_!	0.036321	0.241888
char_freq_$	0.201948	0.323629
char_freq_#	0.042568	0.065067
capital_run_length_average	0.162314	0.109999
capital_run_length_longest	0.475486	0.216097
capital_run_length_total	1.000000	0.249164
spam_class	0.249164	1.000000

```
[58 rows x 58 columns]
```

4.3 correlation matrix heatmap

[11]:
```
fig = px.imshow(df_corr)
fig.show()
```

5 other markdown examples

this text is **bold**
this text is *italic*
example of **ordered list** 1. item 1 2. item 2 3. item 3

example of **unordered list** - item 1 - item 2 - item 3

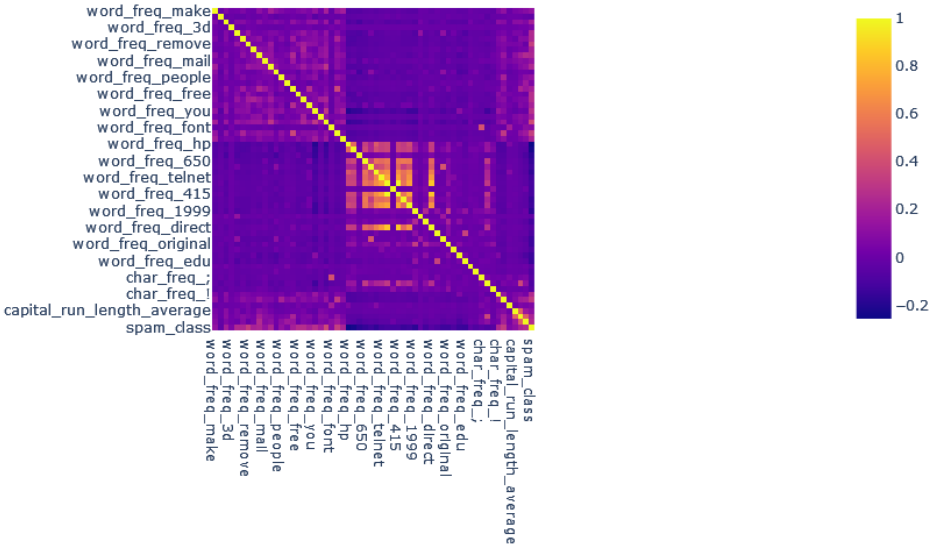
this is the way to write R**code**
`import pandas as pd` or **block of code**

```
import pandas as pd
df = pd.read_csv('data/spambase.data', header = None)
```

we can create a **table** as well:

Measure	Value
MSE	0.5
R^2	0.33

this is a link to codered learning platform
and lastly, lets show a **picture**



6 linear regression

Given a dataset of variables (X_i, Y_i) where X_i is the explanatory variable and Y_i is the dependent variable that varies as X_i does, the simplest model that could be applied for the relation between two of them is a linear one. Simple linear regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

ϵ_i - is the random component of the regression handling the residue, i.e. the lag between the estimation and actual value of the dependent parameter.
β_0 - constant term or the intercept
β_1 - coefficient term or slope of the intercept line

Loss function
Mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

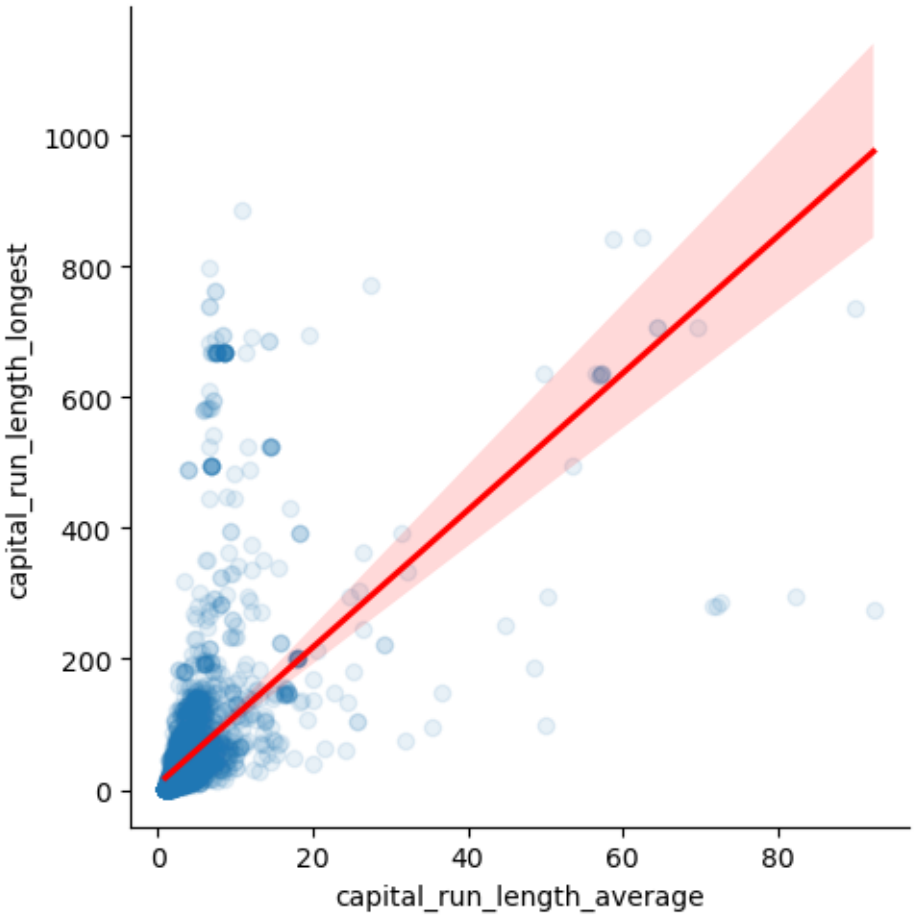
n - number of observations
$\hat{Y}_i$ - predicted value

Y_i - observed value

6.1 chart

[12]:
```
condition = (df['capital_run_length_average'] < 100) &␣
 ↪(df['capital_run_length_longest'] < 1000)
sns.lmplot(x='capital_run_length_average',y='capital_run_length_longest',
           data=df[condition],
           fit_reg=True,
          scatter_kws={'alpha':alpha},
           line_kws={'color': 'red'})
```

[12]:
```
<seaborn.axisgrid.FacetGrid at 0x7efe210cd8a0>
```



[13]:
```
df_corr['capital_run_length_average']['capital_run_length_longest']
```

[13]: 0.49263829723867375

6.2 fitting sklearn model

[14]:
```
#pdb.set_trace()
x = df['capital_run_length_average'].to_frame()
y = df['capital_run_length_longest']
model = LinearRegression(fit_intercept = fit_intercept)
model.fit(x, y)
y_pred = model.predict(x)
```

6.3 model evaluation

Coefficient of determination R^2

[15]:
```
model.score(x,y)
```

[15]:
```
0.24269249190622155
```

Mean squared error MSE

[16]:
```
mean_squared_error(x,y)
```

[16]:
```
35096.723299171266
```

6.4 model intercept and coefficient

Coefficient β_1

[17]:
```
model.coef_
```

[17]:
```
array([3.02592471])
```

Intercept β_0

[18]:
```
model.intercept_
```

[18]:
```
36.46365470176502
```

[19]:
```
import time
time.sleep(5)
```

[20]:
```
import sys
print(sys.executable)
```

```
/opt/conda/bin/python
```

[21]:
```
%reload_ext rpy2.ipython
```

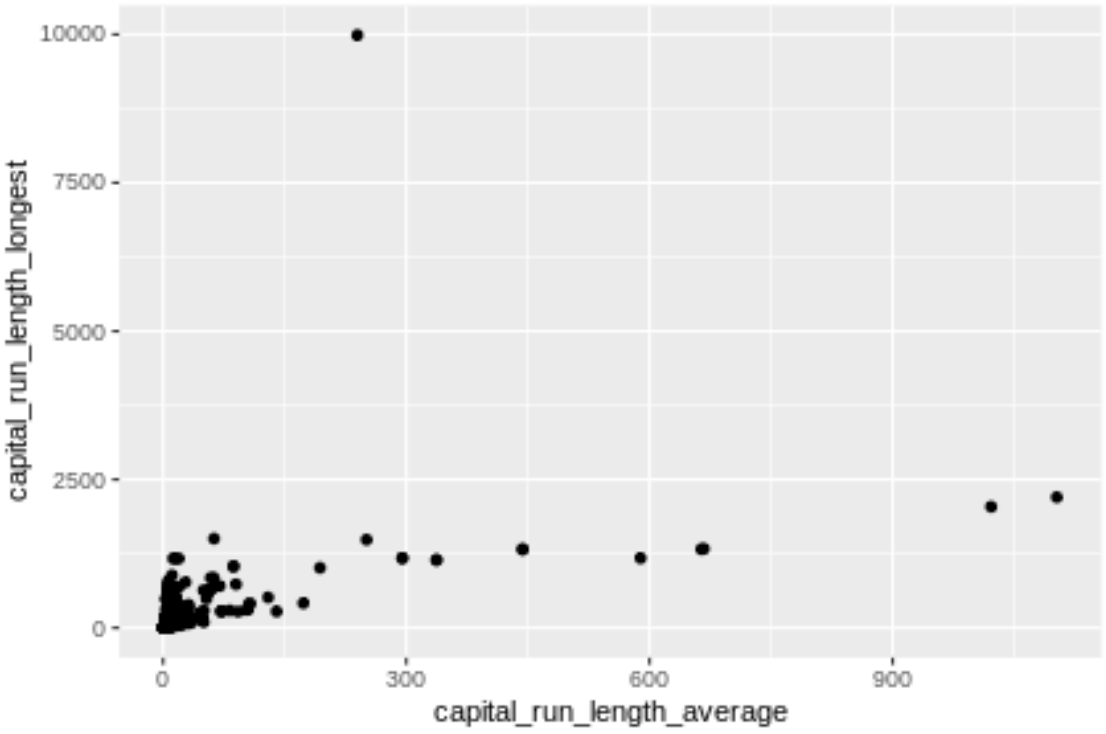
6.5 chart in R

[22]:
```
%%R -i df -w 15 -h 10 --units cm

if (!require('ggplot2')) install.packages('ggplot2',repos='https://cloud.
 ↪r-project.org/');
library('ggplot2')

ggplot(df,
       aes(x = capital_run_length_average, y = capital_run_length_longest)) +
geom_point()
```

```
R[write to console]: Loading required package: ggplot2
```



6.6 model in R

[23]:
```
%%R -i df -o intercept_r -o coef_r -o y_pred_r

model <- lm(capital_run_length_longest ~ capital_run_length_average, data = df)
print(model)

intercept_r <- coef(model)['(Intercept)']
coef_r <- coef(model)['capital_run_length_average']
y_pred_r <- predict(model)
```

```
Call:
lm(formula = capital_run_length_longest ~ capital_run_length_average,
    data = df)

Coefficients:
               (Intercept)  capital_run_length_average
                    36.464                       3.026
```

[24]:
```
intercept_r, coef_r, type(intercept_r)
```

[24]:
```
(array([36.4636547]), array([3.02592471]), numpy.ndarray)
```

[25]:
```
y_pred_r, type(y_pred_r)
```

[25]:
```
(array([47.8290279 , 51.93823365, 66.18126125, …, 40.71205299,
        39.93439034, 40.24606059]),
 numpy.ndarray)
```

7 embedding HTML

[26]:
```
HTML('<h1>It seems embedding HTML works!!!</h1>')
```

[26]:
```
<IPython.core.display.HTML object>
```

8 ipywidgets

[30]:
```
#%debug
v1 = Combobox(
    value = 'word_freq_make',
    placeholder='Choose variable 1',
    options=colnames,
    description='X variable',
    ensure_option=True)
v2 = Combobox(
    value = 'word_freq_all',
    placeholder='Choose variable 2',
    options=colnames,
    description='Y variable',
    ensure_option=True,
)
reg = Checkbox(
    value=False,
    description='Fit regresson',
    indent=False
)

def plot_reg(x_var, y_var, reg):
    sns.lmplot(x=x_var,
               y=y_var,
               data=df,
               fit_reg=reg,
               scatter_kws={'alpha':alpha},
               line_kws={'color': 'red'})
    plt.show()
    if reg:
```

```
        x = df[x_var].to_frame()
        y = df[y_var]
        model = LinearRegression(fit_intercept = fit_intercept)
        model.fit(x, y)
        y_pred = model.predict(x)
        r2 = model.score(x,y)
        print(f'Coeficient of determination R^2 = {r2}')

# An HBox lays out its children horizontally
ui = HBox([v1, v2, reg])

# we link the plotting function, its params to our widgets
out = interactive_output(plot_reg, {'x_var': v1, 'y_var': v2, 'reg':reg})

display(out, ui)
```

```
Output()
```

```
HBox(children=(Combobox(value='word_freq_make', description='X variable',␣
 ↪ensure_option=True, options=('word_f…
```

9 debugging

[31]:
```
def fit_predict_regression(df, var_explanatory, var_dependent, fit_intercept):
    #pdb.set_trace()
    x = df[var_explanatory].to_frame()
    y = df[var_dependent]
    model = LinearRegression(fit_intercept = fit_intercept)
    model.fit(x, y)
    y_pred = model.predict(y)
    return model, y_pred
```

[32]:
```
model, y_pred = fit_predict_regression(df, 'capital_run_length_average',␣
 ↪'capital_run_length_longest', fit_intercept)
```

```
/opt/conda/lib/python3.10/site-packages/sklearn/base.py:420: UserWarning:

X does not have valid feature names, but LinearRegression was fitted with
feature names
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[32], line 1
----> 1 model, y_pred =␣
 ↪fit_predict_regression(df, 'capital_run_length_average', 'capital_run_length_longest', fit_
```

```
Cell In[31], line 7, in fit_predict_regression(df, var_explanatory,␣
 ↪var_dependent, fit_intercept)
      5 model = LinearRegression(fit_intercept = fit_intercept)
      6 model.fit(x, y)
----> 7 y_pred = model.predict(y)
      8 return model, y_pred

File /opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_base.py:354,
 ↪in LinearModel.predict(self, X)
    340 def predict(self, X):
    341     """
    342     Predict using the linear model.
    343
   (…)
    352         Returns predicted values.
    353     """
--> 354     return self._decision_function(X)

File /opt/conda/lib/python3.10/site-packages/sklearn/linear_model/_base.py:337,
 ↪in LinearModel._decision_function(self, X)
    334 def _decision_function(self, X):
    335     check_is_fitted(self)
--> 337     X =␣
 ↪self._validate_data(X, accept_sparse=["csr", "csc", "coo"], reset=False)
    338     return safe_sparse_dot(X, self.coef_.T, dense_output=True) + self.
 ↪intercept_

File /opt/conda/lib/python3.10/site-packages/sklearn/base.py:546, in␣
 ↪BaseEstimator._validate_data(self, X, y, reset, validate_separately,␣
 ↪**check_params)
    544     raise ValueError("Validation should be done on X, y or both.")
    545 elif not no_val_X and no_val_y:
--> 546     X = check_array(X, input_name="X", **check_params)
    547     out = X
    548 elif no_val_X and not no_val_y:

File /opt/conda/lib/python3.10/site-packages/sklearn/utils/validation.py:902, i␣
 ↪check_array(array, accept_sparse, accept_large_sparse, dtype, order, copy,␣
 ↪force_all_finite, ensure_2d, allow_nd, ensure_min_samples,␣
 ↪ensure_min_features, estimator, input_name)
    900     # If input is 1D raise error
    901     if array.ndim == 1:
--> 902         raise ValueError(
    903             "Expected 2D array, got 1D array instead:\narray={}.\n"
    904             "Reshape your data either using array.reshape(-1, 1) if "
    905             "your data has a single feature or array.reshape(1, -1) "
    906             "if it contains a single sample.".format(array)
    907         )
```

```
    909 if dtype_numeric and array.dtype.kind in "USV":
    910     raise ValueError(
    911         "dtype='numeric' is not compatible with arrays of bytes/strings "
    912         "Convert your data to numeric values explicitly instead."
    913     )

ValueError: Expected 2D array, got 1D array instead:
array=[ 61 101 485 …   6   5   5].
Reshape your data either using array.reshape(-1, 1) if your data has a single␣
 ↪feature or array.reshape(1, -1) if it contains a single sample.
```

!!! In case you encounter latex error when converting this file to PDF via Latex, clear above cells output (the output of debugger) !!!

10 exercises

1. write latex equation for mean absolute error
2. create a table with Markdown with evaluation metrics of our regression model and their values
3. select 2 variables from our dataset you think might correlate and plot their scatterplot with seaborn
4. create a simple regression model of these variables and evaluate it with R^2 and MSE
5. parameterize variables used in regression model: add target_variable and input_variable to config file and adapt accordingly the notebook

11 references

1. Markdown cheatsheet
2. Latex
3. Latex machine learning equations
4. Simple linear regression
5. rpy2 package
6. R magic
7. IPython magic
8. Nbparametrise
9. Ipywidgets
10. Custom jupyter widgets
11. Voila
12. Python debugger
13. Nbviewer
14. Binder

[ ]: