Report for TaskD

Group 37

September 25, 2016

Top 15 Companies (TaskD4)

Here, we list the top 15 companies on the basis of their standardized scores.

Rank	Company Index	Score
1	525	72.37672
2	477	72.07414
3	253	72.06298
4	101	71.84182
5	358	71.81531
6	207	71.3849
7	334	70.85849
8	179	70.61823
9	13	70.49138
10	519	70.48215
11	24	70.38025
12	145	70.3692
13	103	70.35524
14	497	70.25587
15	576	70.13045

Limitations in this task

There are potential limits to the previous analysis. One of them can be that new companies may be simply out of competition since they do not have more data than the older players in the market. One more case can be that some companies suffered bad stocks and we would want to consider the n-best scores and rank the companies based on that. This way we can limit out the luck factor that may have affected the company just once or twice. There could be many other problems but we will focus on a few ones.
Hence, we will consider the best-n scores and rank on basis of that.

Task D5 — Top companies based on n-best scores

We choose $X = 15$ and $n = 15$ to compute the best n-scores. After running the script, we came up with the following data (top 5 shown for simplicity).

Rank	Company Index	Score
1	253	22.23603
2	101	22.00681
3	78	21.27863
4	317	21.15653
5	145	21.04884

As we can see, the ranks have considerably changed now, bringing with it some number of inversions. The number of inversions between the previous ranking (task D4) and **this** task is — **53285**.

Task D5 — Top companies based on n-best scores

We choose $X = 15$ and $n = 10$ to compute the best n-scores. After running the script, we came up with the following data (top 5 shown for simplicity).

Rank	Company Index	Score
1	78	16.77166
2	358	16.41489
3	293	16.0625
4	253	15.64493
5	480	15.53922

As we can see, the ranks have considerably changed now, bringing with it some number of inversions. The number of inversions between the previous ranking (task D4) and **this** task is — **50919**.

Combined stocks?

Till now, we have only assumed that each column is independent of the other. But that may not always be true. Sometimes stocks may get mixed up and we may want to normalize all the related content under one *bin*.

In taskD6, we standardize based on bins of size 4. Then, we repeat the procedure of task D5.

Task D6 — Bin-standardized scores

After standardizing on basis of bins, we choose $X = 15$ and $n = 15$ to compute the best n-scores. After running the script, we came up with the following data (top 5 shown for simplicity).

Rank	Company Index	Score
1	101	1142.634
2	253	1139.968
3	317	1121.875
4	404	1119.689
5	384	1109.452

This change is also visible, since now we apply different scaling factors to different bins instead of different columns. Here, the inversion between the data from task D5 (n=15) and data from task D6 (n=15) is — **5930**

Task D6 — Bin-standardized scores

After standardizing on basis of bins, we choose $X = 15$ and $n = 10$ to compute the best n-scores. After running the script, we came up with the following data (top 5 shown for simplicity).

Rank	Company Index	Score
1	101	773.0868
2	253	771.4785
3	320	766.2707
4	78	762.7466
5	293	758.5553

This change is also visible, since now we apply different scaling factors to different bins instead of different columns. Here, the inversion between the data from task D5 (n=10) and data from task D6 (n=10) is — **9121**

Normalized data?

The data we normalized was based on the assumption that the data is already a Gaussian distribution. But it does not have to be that way. In this case, we would use the Shapiro Wilk test and if the data isn't Gaussian, we convert it into a near-Gaussian distribution using a power distribution.
We want to perform this on the bins as well, because of the mixing of data that we mentioned before. In the next slide, we have some values of λ for applying the power transformation.

Values of λ and p.values

Index	Value of λ	Max. p-value
1	NA	NA
2	NA	NA
3	2.7708	0.3410764
4	2.271372	0.1479099
5	NA	NA
6	NA	NA
7	2.315637	0.694019
8	NA	NA
9	NA	NA
10	NA	NA
11	2.685704	0.523623
12	1.969344	0.05513378
13	NA	NA
14	2.482261	0.3465166
15	3.369244	0.3385539

Values of λ and p.values

Index	Value of λ	Max. p-value
16	2.452356	0.4104091
17	NA	NA
18	2.467869	0.9154822
19	2.690514	0.1700652
20	NA	NA
21	NA	NA
22	2.250627	0.8854992
23	2.712406	0.7315938
24	2.001719	0.6904559
25	NA	NA
26	2.320246	0.7682865
27	2.549874	0.02111368
28	3.215067	0.4237685
29	NA	NA
30	2.667507	0.4414825

Values of λ and p.values

Index	Value of λ	Max. p-value
31	2.354934	0.4079119
32	NA	NA
33	NA	NA
34	NA	NA
35	2.431053	0.6273555
36	2.158383	0.3359656
37	NA	NA
38	NA	NA
39	2.603305	0.4420314
40	NA	NA
41	NA	NA
42	NA	NA
43	2.170817	0.07388404
44	2.248474	0.835606
45	NA	NA

Values of λ and p.values

Index	Value of λ	Max. p-value
46	NA	NA
47	NA	NA
48	2.504228	0.7604583
49	2.083446	0.1579835
50	NA	0.01960388
51	NA	NA
52	2.325627	0.438944
53	NA	NA
54	4.284068	0.3450772
55	5.656239	0.07720072
56	NA	0.01546341
57	5.220352	0.1030435
58	4.32525	0.07520227
59	6.68118	0.3770296
60	6.213394	0.02504522

Values of λ and p.values

Index	Value of λ	Max. p-value
61	NA	NA
62	4.942675	0.5196723
63	6.202803	0.3341738
64	5.10602	0.02576021
65	5.332308	0.4642536
66	5.53205	0.06834371
67	6.849416	0.5106512
68	6.45511	0.1220777
69	3.890508	0.4069139
70	5.197715	0.2263008

Values of λ and p.values

Index	Value of λ	Max. p-value
71	7.082642	0.03759176
72	NA	0.0147447
73	4.29295	0.9593105
74	4.974628	0.1018002
75	7.105998	0.07151621
76	6.533667	0.2779635
77	NA	NA
78	5.46435	0.06011275
79	NA	0.003800226
80	NA	0.009763912

Converting to Gaussian and analyzing

After applying the power transformation and standardizing the values, we repeat the procedure for task D4, D5 and D6. Here is the data for D4 (top 5 shown for simplicity).

Rank	Company Index	Score
1	24	74.3676
2	519	74.21533
3	253	74.03203
4	477	73.87042
5	13	73.56091

We can see that the values are different from the previous task D4.

Task D5 after power transformation

Next, we take the task D5 and repeat the same ($X = 15$, $n = 15$). Only first 5 are shown for simplicity.

Rank	Company Index	Score
1	253	1165.855
2	101	1157.329
3	497	1147.976
4	145	1147.585
5	52	1147.16

We can see that the values are different from the previous task D5. Now, we find the number of inversions with respect to the **previous (D4)** task. The number comes out to be — **51395**.

Task D5 after power transformation

Next, we take the task D5 and repeat the same ($X = 15$, $n = 10$). Only first 5 are shown for simplicity.

Rank	Company Index	Score
1	78	791.567
2	497	788.1285
3	293	787.4321
4	253	785.318
5	101	785.2796

We can see that the values are different from the previous task D5. Now, we find the number of inversions with respect to the **previous (D4)** task. The number comes out to be — **48712**.

Converting bins to Gaussian and analyzing

Here, we will have to standardize and classify based on the bins. Here are the λ and p-values of the data.

Index	Value of λ	Max. p-value
1	2.310283	0.07221841
2	1.716053	0.4372291
3	2.170939	0.02576221
4	2.549081	0.02040344
5	2.219386	0.05863462
6	2.150485	0.2607789
7	NA	0.006666137
8	2.130212	0.0550808
9	1.887119	0.1191538
10	2.019534	0.04810784

Converting bins to Gaussian and analyzing

Index	Value of λ	Max. p-value
11	2.027465	0.2840496
12	1.919642	0.3044805
13	NA	0.005074563
14	NA	0.0003121151
15	NA	8.115146e-05
16	NA	0.001025854
17	NA	0.001559148
18	NA	0.0001549022
19	NA	0.0003705385
20	NA	8.88055e-06

Task D6 after power transformation

Now, we find out the best-n rankings using $X = 15, n = 15$. Here is a sample.

1	101	1181.003
2	253	1178.565
3	317	1149.224
4	404	1146.918
5	384	1132.584

The number of inversions with respect to the **previous** task (D5) is — **6285**.

Task D6 after power transformation

Now, we find out the best-n rankings using $X = 15, n = 10$. Here is a sample.

1	101	804.9609
2	253	803.7876
3	78	789.8092
4	320	787.2204
5	293	781.5668

The number of inversions with respect to the **previous** task (D5) is — **9636**.

Variables to access

If you want to see individual values, run this in RStudio, and search the following variables.

- Data for taskD — *firstScores*
- Data for taskE ($n = 15$) — *secondScores*
- Data for taskE ($n = 10$) — *secondScoresX_10*
- Data for taskF ($n = 15$) — *thirdScores*
- Data for taskF ($n = 10$) — *thirdScoresX_10*
- Inversions between taskD and taskE($n = 15$) — *inv12_X15*
- Inversions between taskD and taskE($n = 10$) — *inv12_X10*
- Inversions between taskE($n = 15$) and taskF($n = 15$) — *inv23_X15*
- Inversions between taskE($n = 10$) and taskF($n = 10$) — *inv23_X10*

To get the variables after doing the Shapiro Wilk test, simply prepend the name with '*new*' to it and capitalize the first letter of the old word e.g. *firstScores* — *newFirstScores*