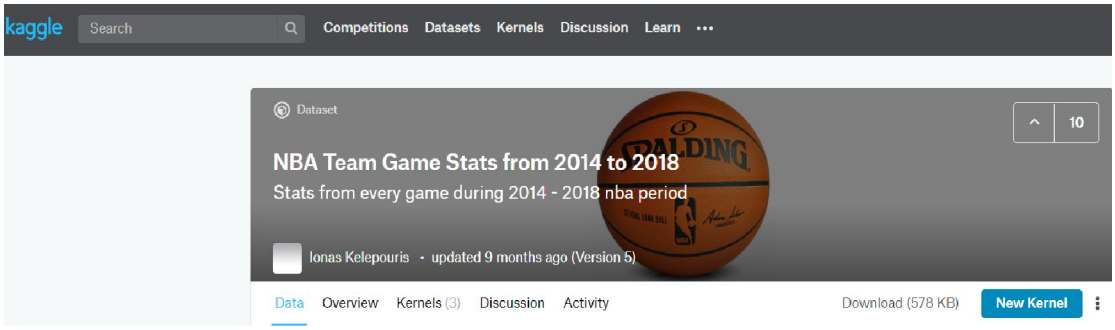
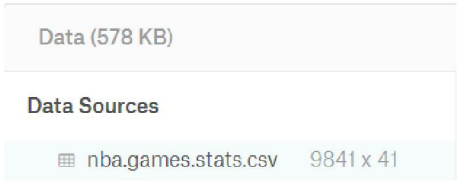
NBA 資料分析

資料介紹:





本次資料分析使用的資料是從 Kaggle 上找到的 Dataset，記載 NBA 2014-2015 到 2017-2018 四個球季共 4920 場例行賽的數據，每場共有 41 項數據，其中最重要的是勝/負，也有包含自己及對手的攻守數據，將無法分析的數據刪除後(例如比賽日期、對手球隊名稱)，整理得到變數如下:

	Game	TeamPoints	OpponentPoints	FieldGoals	FieldGoalsAttempted	FieldGoals.
1	9	97	100	43	86	0.500
2	8	114	103	42	75	0.560
3	10	109	114	41	85	0.482
4	2	102	92	35	69	0.507
5	11	99	89	38	80	0.475

X3PointShots	X3PointShotsAttempted	X3PointShots.	FreeThrows	FreeThrowsAttempted	FreeThrows.
5	23	0.217	6	12	0.500
11	28	0.393	19	23	0.826
9	27	0.333	18	23	0.783
7	20	0.350	25	33	0.758
5	24	0.208	18	26	0.692

OffRebounds	TotalRebounds	Assists	Steals	Blocks	Turnovers	TotalFouls
8	30	28	12	8	11	17
3	36	33	10	5	13	20
13	38	22	7	3	10	17
3	37	26	10	6	12	20
8	46	20	7	5	9	18

Opp.FieldGoals	Opp.FieldGoalsAttempted	Opp.FieldGoals.	Opp.3PointShots	Opp.3PointShotsAttempted
39	76	0.513	9	20
35	74	0.473	10	21
47	87	0.540	6	17
31	81	0.383	12	32
31	83	0.373	4	21

Opp.3PointShots.	Opp.FreeThrows	Opp.FreeThrowsAttempted	Opp.FreeThrows.	Opp.OffRebounds
0.450	13	18	0.722	13
0.476	23	25	0.920	5
0.353	14	22	0.636	13
0.375	18	21	0.857	11
0.190	23	24	0.958	13

Opp.TotalRebounds	Opp.Assists	Opp.Steals	Opp.Blocks	Opp.Turnovers	Opp.TotalFouls	WIN	HOME
46	23	8	4	18	12	0	0
32	27	10	3	14	20	1	1
44	24	7	0	11	24	0	1
44	25	5	5	18	26	1	1
45	12	6	3	12	20	1	1

Game : 場次，這場比賽是球隊在球季中的第幾場比賽 (1~82)

TeamPoints : 球隊得分　OpponentPoints : 對手得分

FieldGoals : 進球數 FieldGoalsAttempted :出手數 FieldGoals . : 出手命中率

X3PointShots : 三分球進球數　X3PointShotsAttempted : 三分球出手數

X3PointShots . : 三分球命中率 FreeThrows : 罰球進球數

FreeThrowsAttempted : 罰球出手數 FreeThrows . : 罰球命中率

OffRebounds : 進攻籃板數 TotalRebounds : 總籃板數 Assists : 助攻數

Steals : 抄截數 Blocks : 阻攻數 Turnovers : 失誤數 TotalFouls : 犯規數

Opp.FieldGoals : 對手進球數 Opp.FieldGoalsAttempted : 對手出手數

Opp.FieldGoals . : 對手出手命中率 Opp.X3PointShots : 對手三分球進球數

Opp.X3PointShotsAttempted : 對手三分球出手數

Opp.X3PointShots . : 對手三分球命中率 Opp.FreeThrows : 對手罰球進球數

Opp.FreeThrowsAttempted : 對手罰球出手數

Opp.FreeThrows . : 對手罰球命中率 Opp.OffRebounds : 對手進攻籃板數

Opp.TotalRebounds : 對手總籃板數 Opp.Assists : 對手助攻數

Opp.Steals : 對手抄截數 Opp.Blocks : 對手阻攻數

Opp.Turnovers : 對手失誤數 Opp.TotalFouls : 對手犯規數

WIN : 勝/負，勝為 1、敗為 0 HOME : 主客場，主場為 1，客場為 0

命中率相關(字尾有”.”)數據皆會是 0~1 的數，而其他數據則會是非負整數。

建立模型 :

在籃球比賽中，資料分析最大的功用就是探討哪些數據對於勝負有較大的影響力，或是影響方向與直觀印象相反(最後面會提到，進攻籃板在最終的模型中即是屬於這樣的例子)。而本次資料分析不作勝/負預測，因為樣本的所有自變數除了”Game”(場次)及”HOME”(主客場)之外，都是要在「比賽完成」後才能收集完成的數據，意即當收集完一個樣本的所有自變數時，比賽已經結束，預測勝負這件事就失去意義，因此本次資料分析以解釋資料為主，使用的模型是「羅吉斯迴歸模型」，將 勝/負 設為應變數，並分為三步建立模型

Step1 : Lasso Regression

```
> coef(best.lasso.model)
37 x 1 sparse Matrix of class "dgCMatrix"
                                    s0
(Intercept)                -9.556439e-01
Game                       -1.953614e-04
TeamPoints                  3.928019e+00
OpponentPoints             -3.928949e+00
FieldGoals                  .
FieldGoalsAttempted         .
FieldGoals.                 2.961322e+00
X3PointShots                .
X3PointShotsAttempted      -5.889369e-06
X3PointShots.               4.949323e-01
FreeThrows                  .
FreeThrowsAttempted         .
FreeThrows.                 .
OffRebounds                 2.218936e-02
TotalRebounds               .
Assists                     4.718554e-03
Steals                      .
Blocks                      3.232884e-02
Turnovers                   .
TotalFouls                 -6.103010e-03
Opp.FieldGoals              .
Opp.FieldGoalsAttempted     .
Opp.FieldGoals.            -2.758960e+00
Opp.3PointShots             .
Opp.3PointShotsAttempted    .
Opp.3PointShots.            .
Opp.FreeThrows             -3.397523e-03
Opp.FreeThrowsAttempted     .
Opp.FreeThrows.            -1.607138e-01
Opp.OffRebounds             .
Opp.TotalRebounds           .
Opp.Assists                 .
Opp.Steals                 -1.982884e-02
Opp.Blocks                  .
Opp.Turnovers               .
Opp.TotalFouls              3.616088e-02
HOME                        .
```

挑選自變數 – Lasso Regression

由上圖可知，Lasso 挑選出變數為:Game、TeamPoints、OpponentPoints、FieldGoals.、X3PointShotsAttempted、 X3PointShots.、OffRebounds、Assists、Blocks、TotalFouls、Opp.FieldGoals.、Opp.FreeThrows、Opp.FreeThrows.、Opp.Steals、Opp.TotalFouls 共 15 個變數

再將這 15 個變數跑一次羅吉斯迴歸，得到報表如下

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.036e+00  1.348e+04   0.000    1.000
Game                    -1.962e-03  3.036e+01   0.000    1.000
TeamPoints               1.929e+01  7.103e+02   0.027    0.978
OpponentPoints          -1.930e+01  7.063e+02  -0.027    0.978
FieldGoals               2.312e-02  3.325e+02   0.000    1.000
FieldGoals.              1.050e+00  2.257e+04   0.000    1.000
FreeThrows.              1.646e-01  7.820e+03   0.000    1.000
Blocks                   4.965e-02  3.119e+02   0.000    1.000
TotalFouls              -8.112e-03  2.014e+02   0.000    1.000
Opp.FieldGoals.         -1.766e+00  2.841e+04   0.000    1.000
Opp.3PointShotsAttempted 1.367e-03  1.231e+02   0.000    1.000
Opp.3PointShots.         5.267e-01  9.560e+03   0.000    1.000
Opp.OffRebounds          7.895e-03  2.355e+02   0.000    1.000
Opp.Assists              4.159e-03  1.837e+02   0.000    1.000
Opp.Steals              -2.723e-02  2.627e+02   0.000    1.000
Opp.TotalFouls           4.455e-02  2.133e+02   0.000    1.000
```

Step2：解決模型無法收斂的問題

由圖可知模型並沒有收斂，懷疑是因兩個解釋力極強且方向相反的自變數所導致，因此嘗試性的鎖定 TeamPoints & OpponentPoints 這兩個估計係數較大的變數，單獨跑一次羅吉斯迴歸

```
> model_OnlyPoint  <- glm(formula=WIN~.,family = "binomial",data=Data_OnlyPo
int)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(model_OnlyPoint )

Call:
glm(formula = WIN ~ ., family = "binomial", data = Data_OnlyPoint)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-9.222e-05  -2.100e-08   2.100e-08   2.100e-08   9.409e-05

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.4473  6970.7114   0.000    1.000
TeamPoints      19.3412   697.8233   0.028    0.978
OpponentPoints -19.3371   697.3773  -0.028    0.978
```

發現模型光是只有這兩個變數就無法收斂，假設成立

並且可以發現，應該要先建立一個有效的模型，再去跑 Lasso 比較洽當

因此將此二變數移除，並將其餘所有包括被 Lasso 刪減的自變數放入模型

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.757e+00	1.302e+05	0.000	1.000
Game	-9.573e-04	3.188e+01	0.000	1.000
FieldGoals	3.865e+01	2.664e+03	0.015	0.988
FieldGoalsAttempted	-8.616e-02	1.116e+03	0.000	1.000
FieldGoals.	-8.880e+00	1.949e+05	0.000	1.000
X3PointShots	1.927e+01	1.249e+03	0.015	0.988
X3PointShotsAttempted	-2.424e-02	4.030e+02	0.000	1.000
X3PointShots.	3.502e-02	2.431e+04	0.000	1.000
FreeThrows	1.927e+01	1.308e+03	0.015	0.988
FreeThrowsAttempted	-3.150e-02	8.348e+02	0.000	1.000
FreeThrows.	1.095e+00	2.409e+04	0.000	1.000
OffRebounds	2.050e-01	4.867e+02	0.000	1.000
TotalRebounds	-1.153e-01	3.431e+02	0.000	1.000
Assists	3.322e-02	2.188e+02	0.000	1.000
Steals	3.428e-02	4.223e+02	0.000	1.000
Blocks	4.400e-02	3.362e+02	0.000	1.000
Turnovers	4.462e-02	4.735e+02	0.000	1.000
TotalFouls	-2.173e-02	3.499e+02	0.000	1.000

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
Opp.FieldGoals	-3.862e+01	1.459e+03	-0.026	0.979
Opp.FieldGoalsAttempted	6.821e-02	3.708e+02	0.000	1.000
Opp.3PointShots	-1.909e+01	1.459e+03	-0.013	0.990
Opp.3PointShotsAttempted	-4.874e-02	4.675e+02	0.000	1.000
Opp.3PointShots.	-2.015e+00	3.245e+04	0.000	1.000
Opp.FreeThrows	-1.935e+01	1.198e+03	-0.016	0.987
Opp.FreeThrowsAttempted	7.800e-02	7.841e+02	0.000	1.000
Opp.FreeThrows.	9.200e-01	2.156e+04	0.000	1.000
Opp.OffRebounds	-1.251e-01	4.567e+02	0.000	1.000
Opp.TotalRebounds	7.788e-02	3.685e+02	0.000	1.000
Opp.Assists	8.643e-03	2.076e+02	0.000	1.000
Opp.Steals	-2.680e-02	3.965e+02	0.000	1.000
Opp.Blocks	-2.381e-02	3.425e+02	0.000	1.000
Opp.Turnovers	-9.733e-02	4.819e+02	0.000	1.000
Opp.TotalFouls	7.599e-02	3.214e+02	0.000	1.000
HOME	-1.416e-01	1.567e+03	0.000	1.000

模型依然無法收斂，依照上一步的邏輯繼續找可能使模型無法收斂的變數

這次鎖定 FieldGoals & Opp.FieldGoals，將此二變數移除，得到模型如下

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.176e+00	1.050e+01	0.112	0.9108
Game	-9.363e-03	9.559e-03	-0.979	0.3273
FieldGoalsAttempted	2.176e+00	2.374e-01	9.165	< 2e-16
FieldGoals.	4.387e+02	4.546e+01	9.650	< 2e-16
X3PointShots	2.559e+00	3.780e-01	6.770	1.29e-11
X3PointShotsAttempted	-4.223e-03	1.168e-01	-0.036	0.9712
X3PointShots.	-9.236e+00	7.001e+00	-1.319	0.1871
FreeThrows	2.327e+00	3.853e-01	6.040	1.54e-09
FreeThrowsAttempted	7.322e-02	2.420e-01	0.303	0.7622
FreeThrows.	3.741e+00	6.864e+00	0.545	0.5858
OffRebounds	1.526e-01	1.419e-01	1.076	0.2820
TotalRebounds	-9.508e-02	8.778e-02	-1.083	0.2787
Assists	6.965e-02	6.131e-02	1.136	0.2559
Steals	-3.187e-02	1.257e-01	-0.254	0.7999
Blocks	-1.485e-01	9.830e-02	-1.511	0.1308
Turnovers	4.833e-03	1.241e-01	0.039	0.9689
TotalFouls	-2.262e-01	9.447e-02	-2.394	0.0167

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
Opp.FieldGoalsAttempted	-2.123e+00	2.270e-01	-9.353	< 2e-16
Opp.FieldGoals.	-4.452e+02	4.605e+01	-9.668	< 2e-16
Opp.3PointShots	-2.526e+00	4.186e-01	-6.035	1.59e-09
Opp.3PointShotsAttempted	2.630e-02	1.228e-01	0.214	0.8305
Opp.3PointShots.	5.627e+00	8.259e+00	0.681	0.4956
Opp.FreeThrows	-2.582e+00	3.434e-01	-7.521	5.45e-14
Opp.FreeThrowsAttempted	1.947e-01	2.009e-01	0.969	0.3325
Opp.FreeThrows.	-1.013e+00	5.975e+00	-0.169	0.8654
Opp.OffRebounds	5.010e-03	1.278e-01	0.039	0.9687
Opp.TotalRebounds	-1.642e-02	1.023e-01	-0.161	0.8725
Opp.Assists	5.966e-03	6.212e-02	0.096	0.9235
Opp.Steals	-1.886e-01	1.154e-01	-1.634	0.1022
Opp.Blocks	1.365e-02	9.338e-02	0.146	0.8838
Opp.Turnovers	8.131e-02	1.293e-01	0.629	0.5294
Opp.TotalFouls	7.029e-02	9.363e-02	0.751	0.4528
HOME	2.349e-01	4.645e-01	0.506	0.6130

到這步模型終於收斂，並且經過剛剛的步驟可發現 ： 在做籃球勝負的資料分析時，並不一定要使用 Lasso 刪除高度相關的變數，也不需要用逐步迴歸刪除解釋力較小的變數，因為我們的模型是要去探討變數的影響力而非預測，再來是可以發現某些解釋力很強的變數本身並沒探討的意義，例如 ： 害模型無法收斂的 TeamPoints & OpponentPoints，兩個解釋力強到直接決定勝負的變數(籃球比賽只要得分比對手得分多，則肯定會贏球)，本身並無法成為球隊建立方針的指標，因為大家都知道得分多、失分少就能贏球，卻很難做到，所以可以發現解釋力越強的變數，就越難針對去做改善，在這邊改變我們挑選變數的方針，改為針對解釋力強且難以改善的變數進行刪減

Step3： 移除解釋力強且難以改善的變數

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.639868	2.237113	0.733	0.46354
Game	-0.006232	0.002520	-2.473	0.01341
FieldGoalsAttempted	0.120034	0.023517	5.104	3.32e-07
X3PointShots	0.682584	0.085885	7.948	1.90e-15
X3PointShotsAttempted	-0.177517	0.031292	-5.673	1.40e-08
X3PointShots.	-2.650881	1.991180	-1.331	0.18309
FreeThrows	0.055733	0.077502	0.719	0.47207
FreeThrowsAttempted	0.078549	0.061331	1.281	0.20029
FreeThrows.	0.817464	1.738592	0.470	0.63822
OffRebounds	-0.606311	0.037087	-16.348	< 2e-16
TotalRebounds	0.537521	0.022797	23.579	< 2e-16
Assists	0.141238	0.015656	9.021	< 2e-16
Steals	0.074350	0.031297	2.376	0.01752
Blocks	0.136102	0.024376	5.583	2.36e-08
Turnovers	-0.448838	0.034889	-12.865	< 2e-16
TotalFouls	-0.133133	0.024262	-5.487	4.08e-08

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
Opp.FieldGoalsAttempted	-0.132350	0.024095	-5.493	3.96e-08
Opp.3PointShots	-0.470126	0.092083	-5.105	3.30e-07
Opp.3PointShotsAttempted	0.097067	0.033462	2.901	0.00372
Opp.3PointShots.	-1.128245	2.202470	-0.512	0.60847
Opp.FreeThrows	-0.118922	0.072999	-1.629	0.10329
Opp.FreeThrowsAttempted	-0.024322	0.058666	-0.415	0.67845
Opp.FreeThrows.	-0.760448	1.641356	-0.463	0.64315
Opp.OffRebounds	0.645986	0.037831	17.075	< 2e-16
Opp.TotalRebounds	-0.513526	0.022760	-22.563	< 2e-16
Opp.Assists	-0.135266	0.014897	-9.080	< 2e-16
Opp.Steals	-0.058729	0.030823	-1.905	0.05673
Opp.Blocks	-0.181020	0.024789	-7.302	2.82e-13
Opp.Turnovers	0.442613	0.034614	12.787	< 2e-16
Opp.TotalFouls	0.109847	0.023402	4.694	2.68e-06
HOME	0.050883	0.116054	0.438	0.66107

這次刪除的是 FieldGoals . & Opp.FieldGoals . ，自己以及對手的投籃命中

率，刪完這兩個變數後發現模型的各自變數解釋力相互對稱(自己/對手)

故刪除對手相關的變數，只留下自己的數據以及 Opp.FieldGoalsAttempted

(對手的出手數，要探討節奏對勝負的影響，故保留)做為自變數

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.435054   1.219224  -1.997   0.0458 *
Game                    -0.008595   0.001704  -5.044 4.56e-07 ***
FieldGoalsAttempted     -0.132143   0.010062 -13.133  < 2e-16 ***
X3PointShots             0.378078   0.059109   6.396 1.59e-10 ***
X3PointShotsAttempted   -0.131755   0.021742  -6.060 1.36e-09 ***
X3PointShots.            0.396121   1.412454   0.280   0.7791
FreeThrows               0.034987   0.052375   0.668   0.5041
FreeThrowsAttempted      0.019387   0.040468   0.479   0.6319
FreeThrows.              2.782027   1.161242   2.396   0.0166 *
OffRebounds             -0.122750   0.016098  -7.625 2.44e-14 ***
TotalRebounds            0.314498   0.011130  28.256  < 2e-16 ***
Assists                  0.165460   0.010023  16.507  < 2e-16 ***
Steals                   0.301573   0.016721  18.035  < 2e-16 ***
Blocks                   0.124082   0.016603   7.473 7.81e-14 ***
Turnovers               -0.242928   0.014248 -17.049  < 2e-16 ***
TotalFouls              -0.078270   0.010653  -7.347 2.02e-13 ***
Opp.FieldGoalsAttempted -0.043769   0.007682  -5.697 1.22e-08 ***
HOME                     0.352227   0.079551   4.428 9.53e-06 ***
```

結論

最後得到的模型如上圖，共剩下 17 個變數，其中 TotalRebounds(籃板)對勝負正向的解釋力最強，但有趣的是 OffRebounds(進攻籃板)卻對勝負具負向影響，初步推測的原因是爛隊可能在比賽前期就已大幅落後，比賽早早進入垃圾時間，命中率低加上對手疏於鞏固籃板，導致抓很多進攻籃板但球隊卻輸球的狀況，但實際狀況如何還需要跟專業執教團隊討論才能夠下結論。

而”Game”(場次)具負向影響，因為當兩隊開打前的場次不同時，場次較高代表之前進行了較多比賽，必會對體能造成較大的負擔。

再來探討一下出手數，自己的出手數 FieldGoalsAttempted 及對方出手數 Opp.FieldGoalsAttempted 皆是對勝/負有負向的影響，我們可以把出手數的增加視為節奏的提升，雖然 NBA 各隊近年來有將節奏加快的趨勢，但數據顯示對大部分的球隊而言，加快節奏不一定對勝率有正向的影響，這點在三分球出手數 X3PointShotsAttempted 上也看的出來，雖然目前聯盟最強的金洲勇士隊是個節奏超快的球隊(出手數及三分出手數皆是聯盟頂尖)，但在這份數據分析的報告中，一昧地模仿勇士隊並無法複製勇士的致勝方程式，而是像傳統球隊那樣著重防守、助攻、放慢節奏並減少失誤，才是對勝利有幫助的。

參考資料

https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018