統計學一下期末報告

NBA 球員薪水與相關數據分析

陳禹翰、陳承妤、周子馨、宋凱翔、陳詣斌

Outline

01 研究動機、主題

02 資料簡介

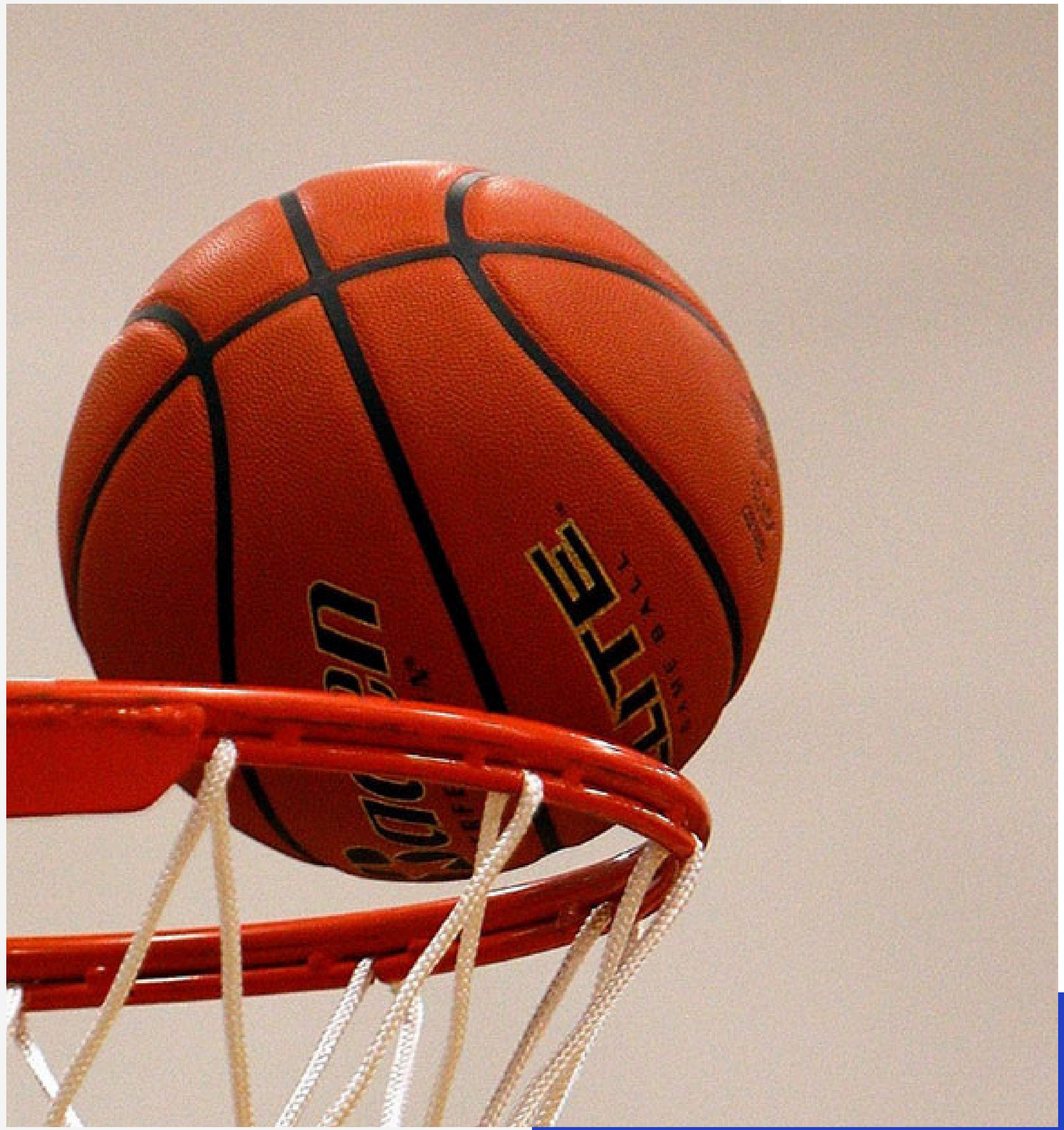
03 敘述統計–球員數據

04 球隊薪資總額分析與模型

05 球員薪水分析與模型預測

06 結論

問題



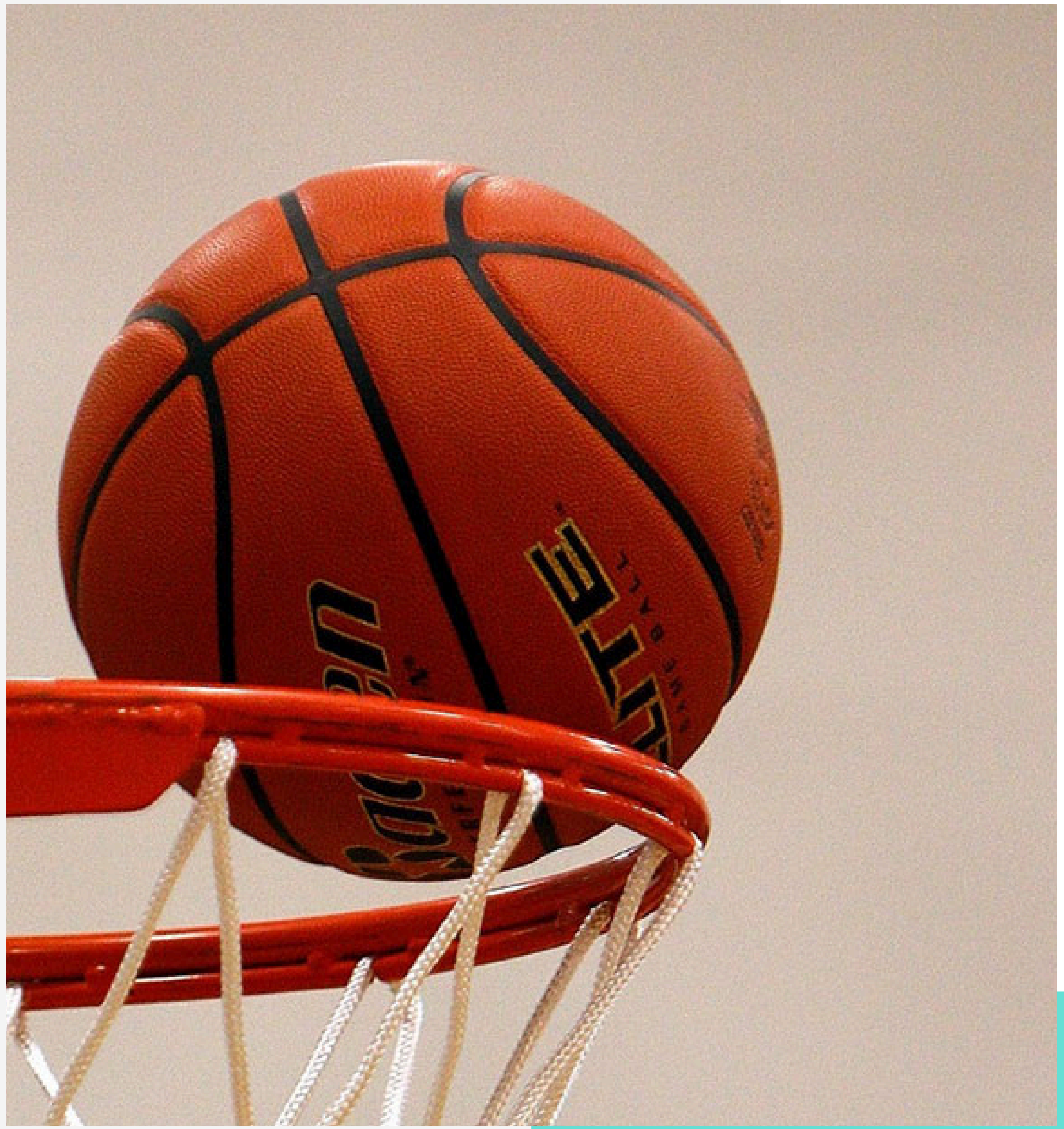
研究動機、主題

研究動機

一位球員的薪資會由多重因素所構成，不單單取決於其個人能力和表現，還可能受到其他因素的影響， 例如球員上場時間、位置、球隊所在地區經濟水平等。我們想要了解一位球員乃至於一個球隊的薪資可能會和哪些指標有所關聯，分析其背後的原因。

研究主題

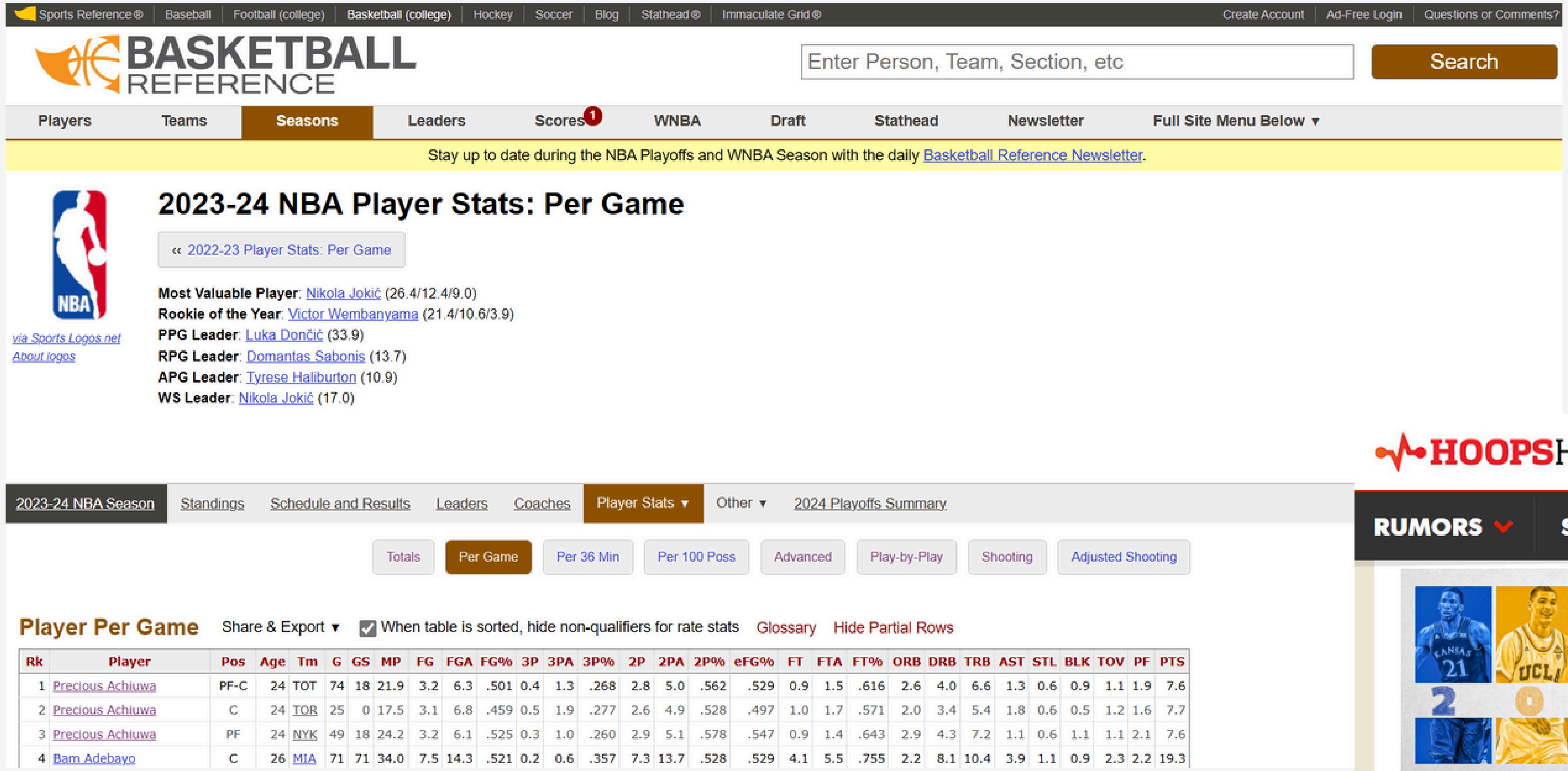
- 分析 NBA 球員各項指標對薪資的影響，建立回歸模型解釋之
- 驗證 NBA 各位置的球員薪資是否有所不同
- 建立 NBA 各位置球員的薪資回歸模型，觀察自變數是否有所不同
- 分析NBA 球隊所在地區經濟水平與球隊相關數據對球隊薪資總額的影響



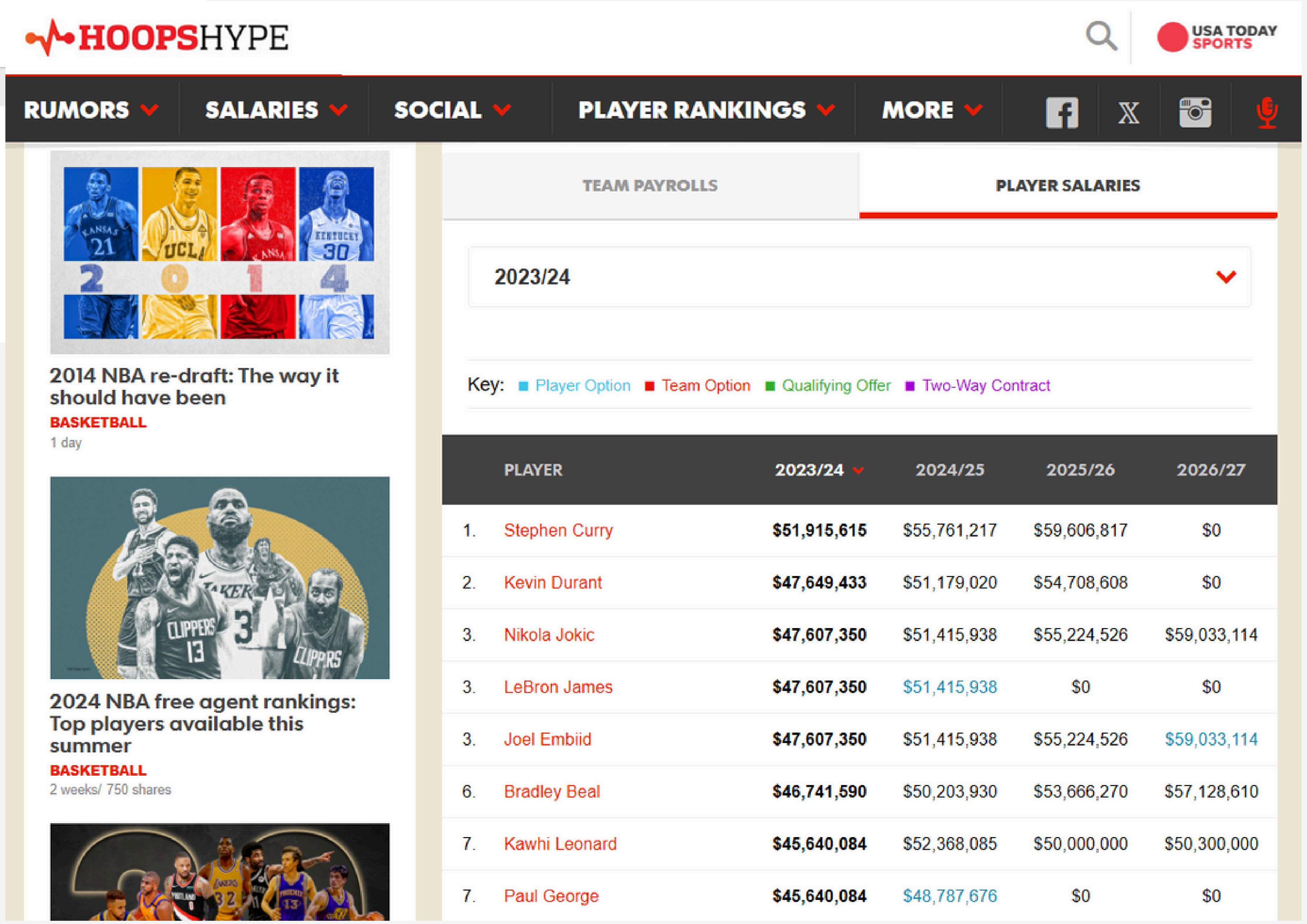
資料簡介

資料簡介 – nba_player_stats

- 資料範圍：NBA 2023-24 賽季有簽署合約且有上場比賽的球員（共494位）
- 資料來源
    - basketball-reference.com
    - hoopshype.com
- 資料前處理
    - 合併球員轉隊前後的資料
    - 合併傳統數據、進階數據與薪水至同一檔案
    - 計算自定義變數（城市經濟水平、球員位置劃分、dummy variables 等）



basketball-reference.com



hoopshype.com

傳統數據　　薪水

	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	...	VORP	2023-24_Salary	StateLevel	Level1	Level2	PosCtg	Gaurd	Forward	gs_rate	Core
0	Stephen Curry	PG	35	GSW	74	74	32.7	8.8	19.5	0.450	...	4.4	51915615	1	1	0	G	1	0	1.0	Main
1	Kevin Durant	PF	35	PHO	75	75	37.2	10.0	19.1	0.523	...	4.3	47649433	2	0	1	F	0	1	1.0	Main
2	Joel Embiid	C	29	PHI	39	39	33.6	11.5	21.8	0.529	...	4.5	47607350	1	1	0	C	0	0	1.0	Main
3	LeBron James	PF	39	LAL	71	71	35.3	9.6	17.9	0.540	...	5.4	47607350	1	1	0	F	0	1	1.0	Main
4	Bradley Beal	SG	30	PHO	53	53	33.3	7.1	13.9	0.513	...	1.1	46741590	2	0	1	G	1	0	1.0	Main
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
489	Jalen Crutcher	PG	24	NOP	1	0	3.0	0.0	1.0	0.000	...	0.0	64343	3	0	0	G	1	0	0.0	Other
490	Izaiah Brockington	PG	24	NOP	1	0	3.0	2.0	5.0	0.400	...	0.0	64343	3	0	0	G	1	0	0.0	Other
491	Timmy Allen	SF	24	MEM	5	0	25.0	1.2	4.6	0.261	...	-0.3	64343	2	0	1	F	0	1	0.0	Other
492	Kaiser Gates	SF	27	NOP	1	0	7.0	0.0	4.0	0.000	...	-0.1	35389	3	0	0	F	0	1	0.0	Other
493	Dmytro Skapintsev	C	25	NYK	2	0	1.0	0.0	0.5	0.000	...	0.0	26322	1	1	0	C	0	0	0.0	Other

494 rows × 58 columns

進階數據　　自定義參數

自定義參數

- StateLevel: 球員所屬的球隊所在城市的經濟發展水平，分成四個 level
- PosCtg: 將球員位置劃分成 C (中鋒), F (前鋒), G (後衛)
- gs_rate: GS / G
- Core: 將 gs_rate >= 0.9以及出場數 >= 10 和 出場時間 >= 5

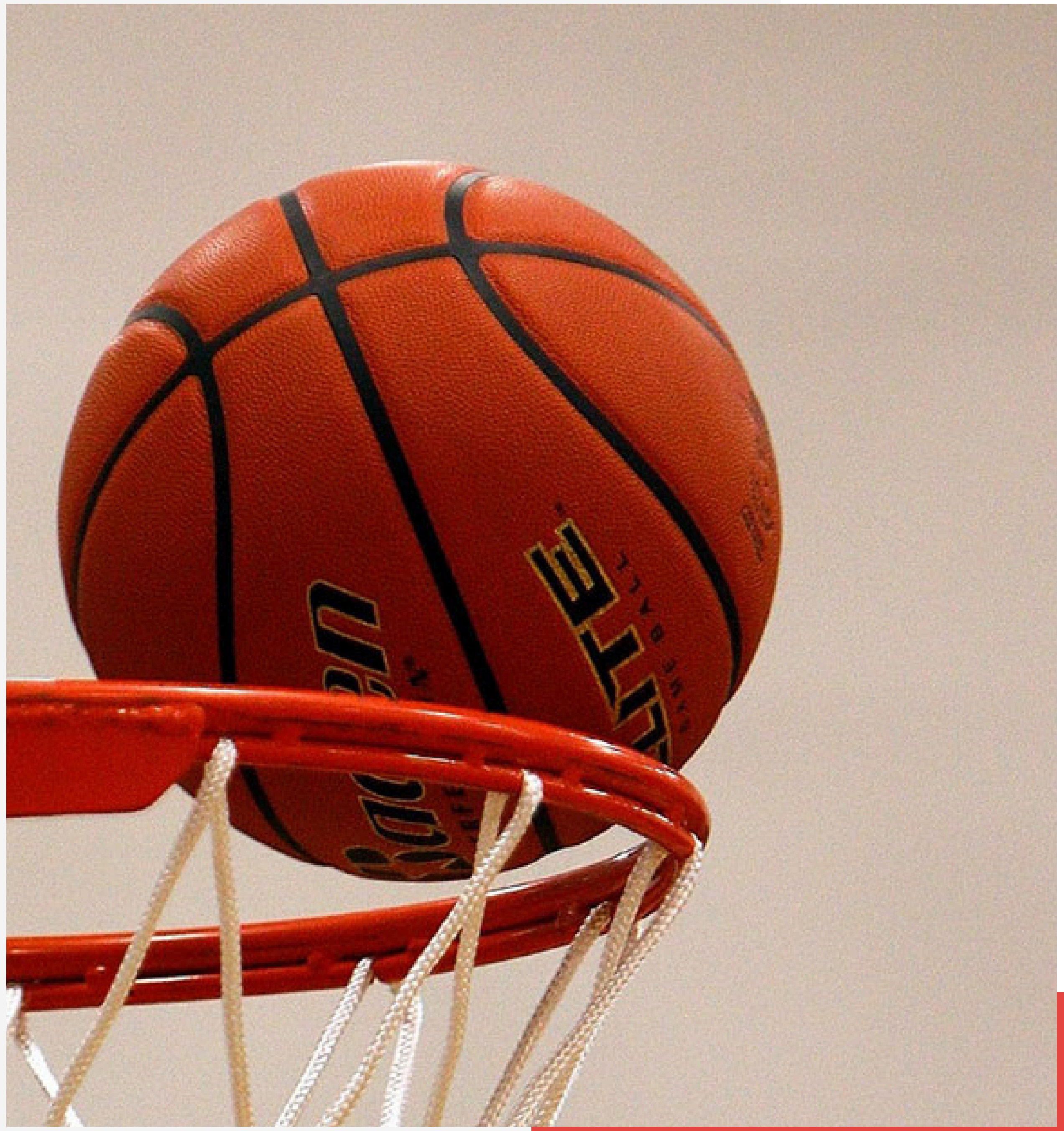
資料簡介 - city_team_salary

- 資料範圍：28 座城市之經濟表現與所在球隊的相關數據
- 資料來源
  - american-growth-project-january-01042023r.pdf (unc.edu)
  - 2023 Fall Statement | In Brief (ontario.ca)
  - Sportico's NBA Valuations
- 資料前處理
  - 如該城市擁有兩支球隊，則取其平均

城市數據　　球隊薪資　　球隊數據

	City	GDP (billion)	population (million)	GDP per person	Team Salaries (milllions)	Win	Revenue (million)	Net Worth (m)
0	Boston	842	8.5	9906	185.6823	64	407.0	500
1	New York	2538	23.6	10754	160.9242	41	455.0	4950
2	Philadelphia	601	7.4	8122	166.5278	47	377.0	5900
3	Ontario (Toronto)	800	15.4	5195	164.4861	25	347.0	2100
4	Chicago	875	10.0	8750	165.6304	39	380.0	1850
5	Cleveland	256	3.6	7111	166.8743	48	342.0	1570
6	Detroit	373	5.4	6907	138.3696	14	282.0	6200
7	Indiana	205	2.5	8200	148.4723	47	305.0	3300
8	Milwaukee	147	2.1	7000	187.3467	49	325.0	1800
9	Atlanta	587	6.9	8507	159.1534	36	373.0	6300

圖表

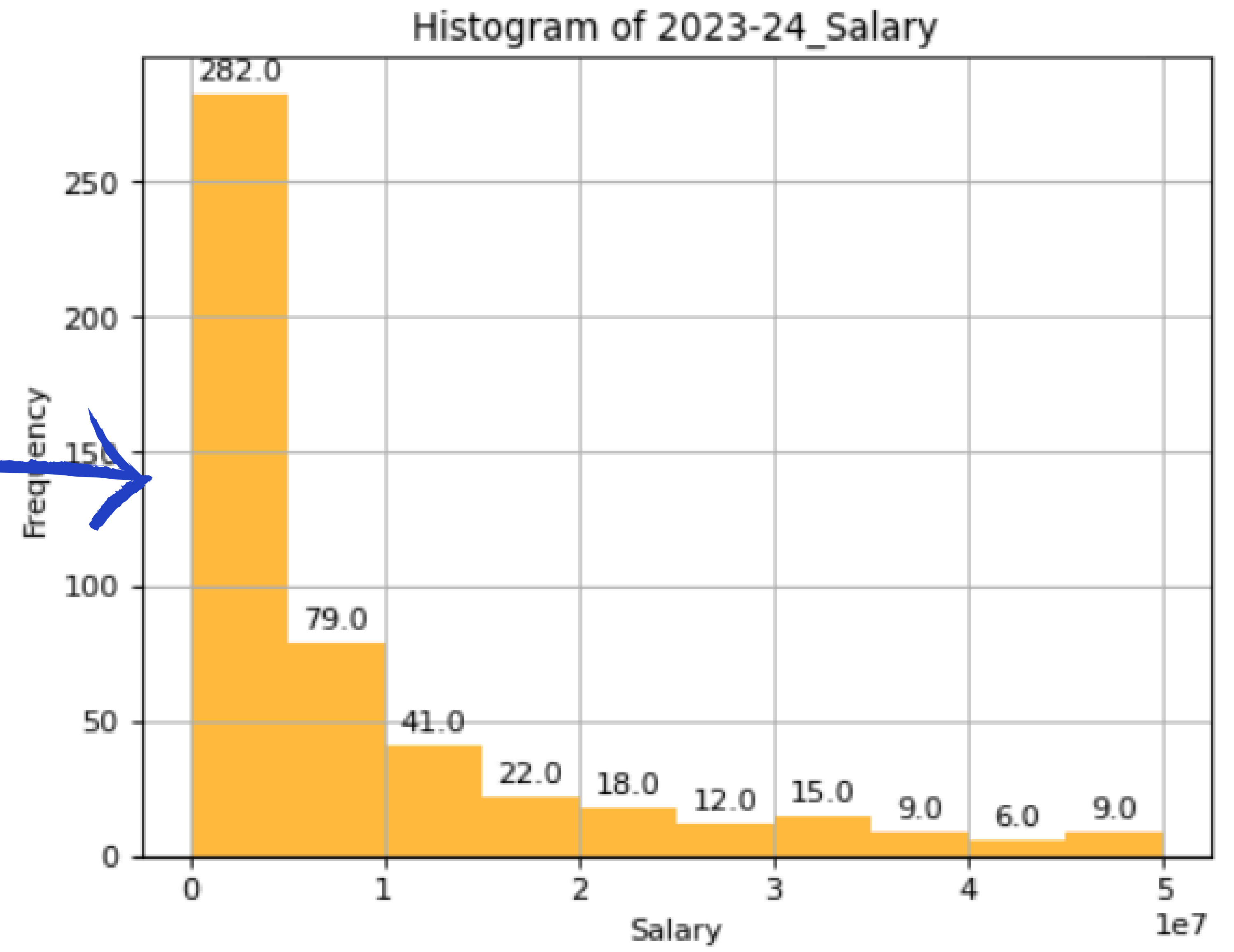


敘述統計

敘述統計

低薪球員比例高

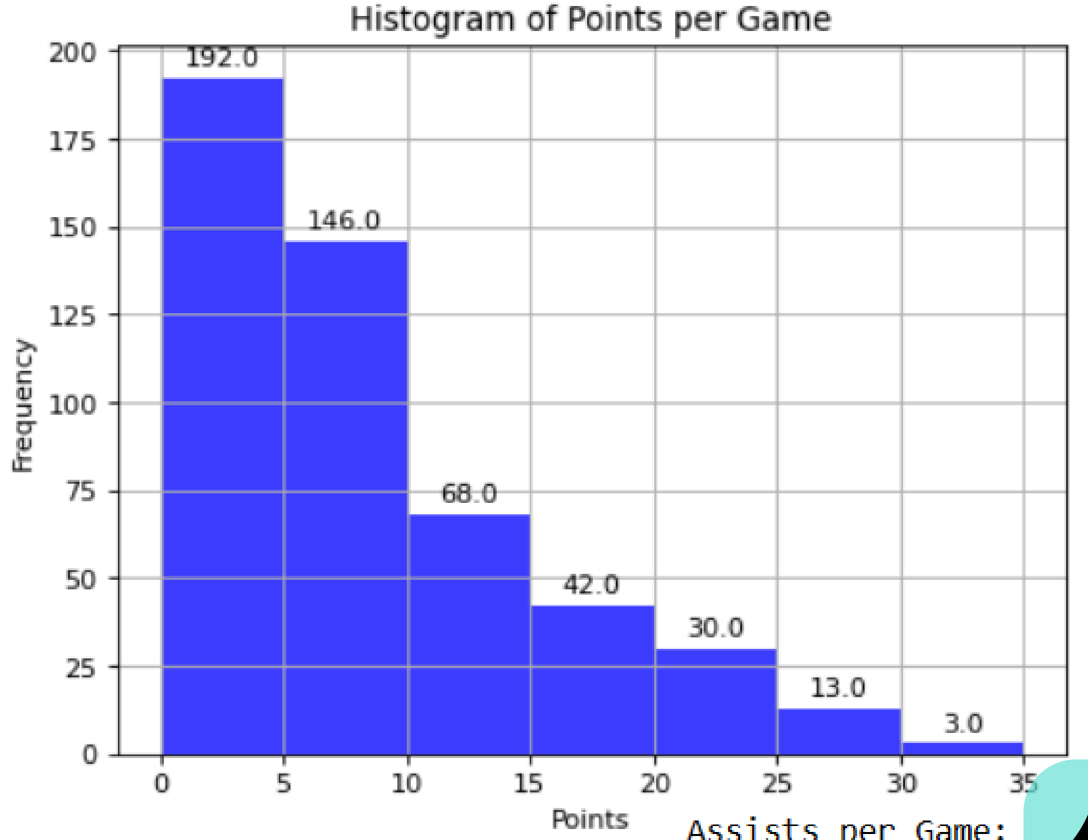
```
2023-24_Salary:
average:  8593873.4048583
standar deviation:  11067602.72773844
skewness:  1.8500508210381228
```



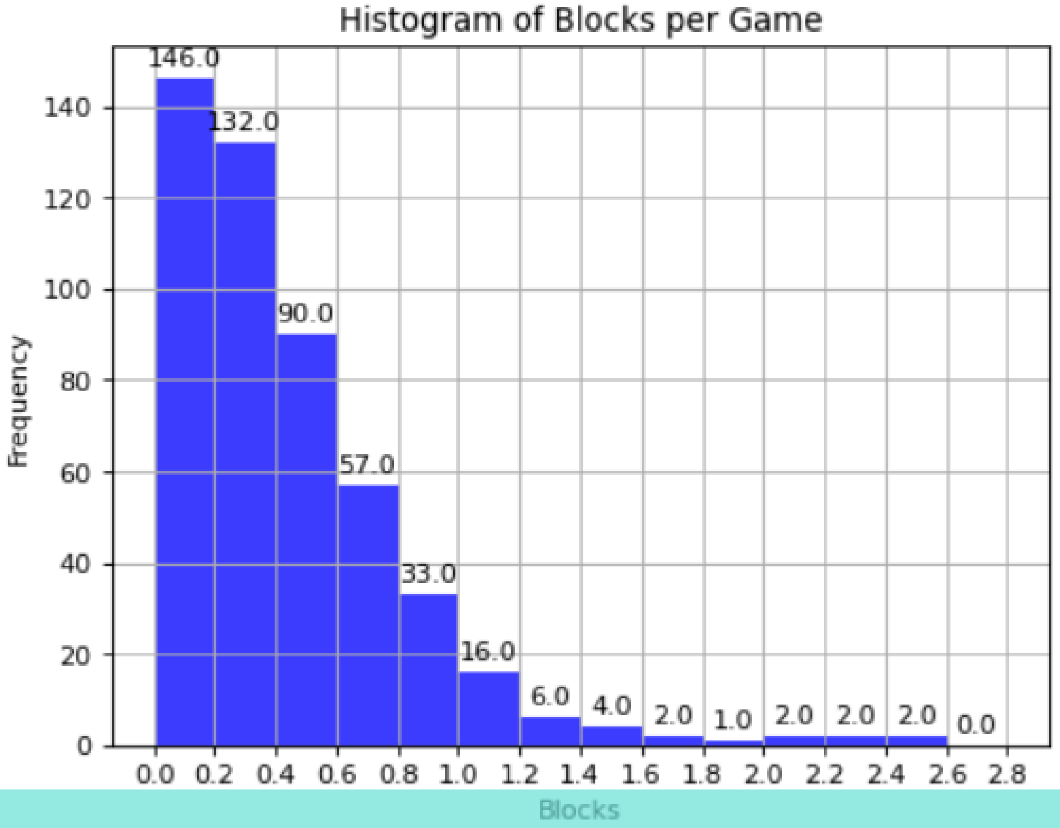
圖：2023-34 賽季球員薪水直方圖

傳統累積數據之直方圖右偏情況明顯

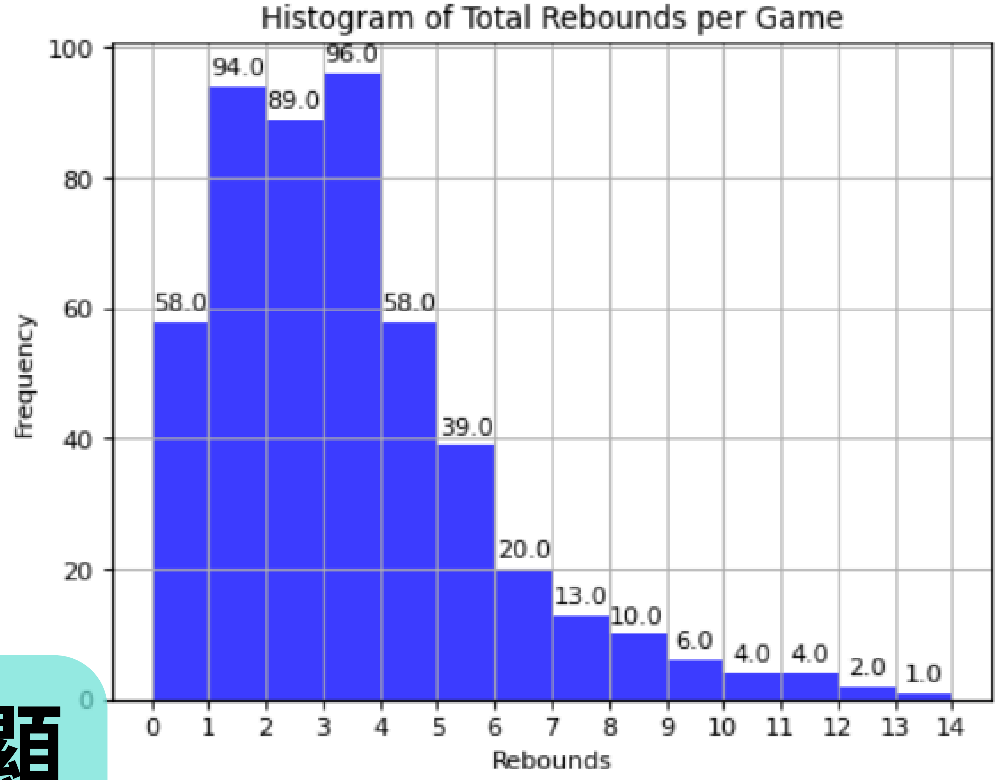
Points per Game:
average: 8.47975708502024
standar deviation: 6.846323161499208
skewness: 1.1736429118868292



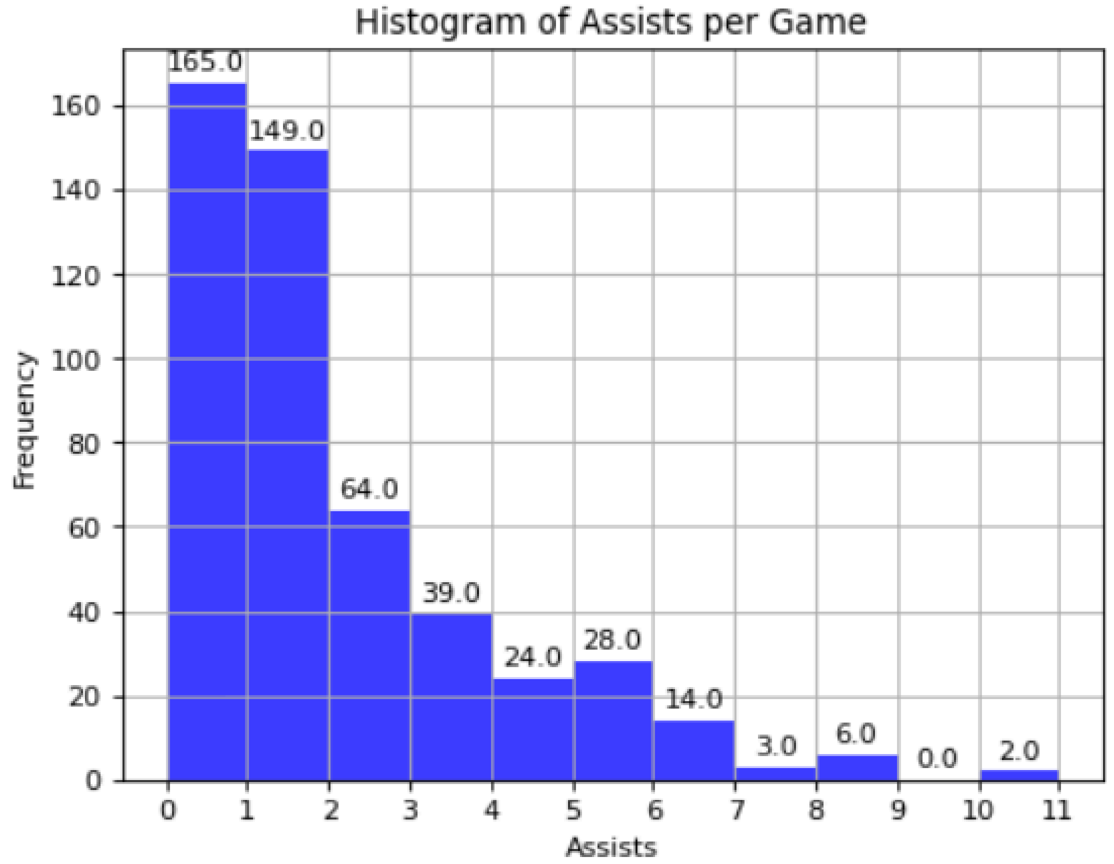
Blocks per Game:
average: 0.4022267206477724
standar deviation: 0.4084818109699562
skewness: 2.6501916837811534



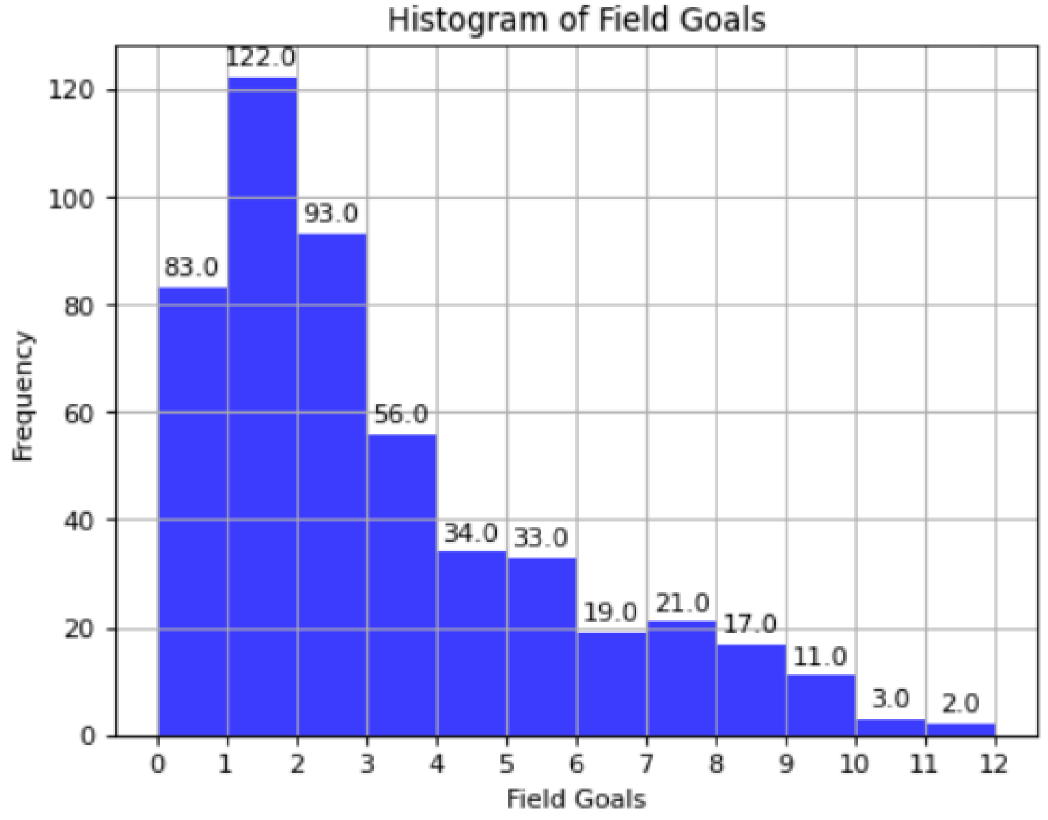
Total Rebounds per Game:
average: 3.363360323886639
standar deviation: 2.35611308518218
skewness: 1.299311192798905



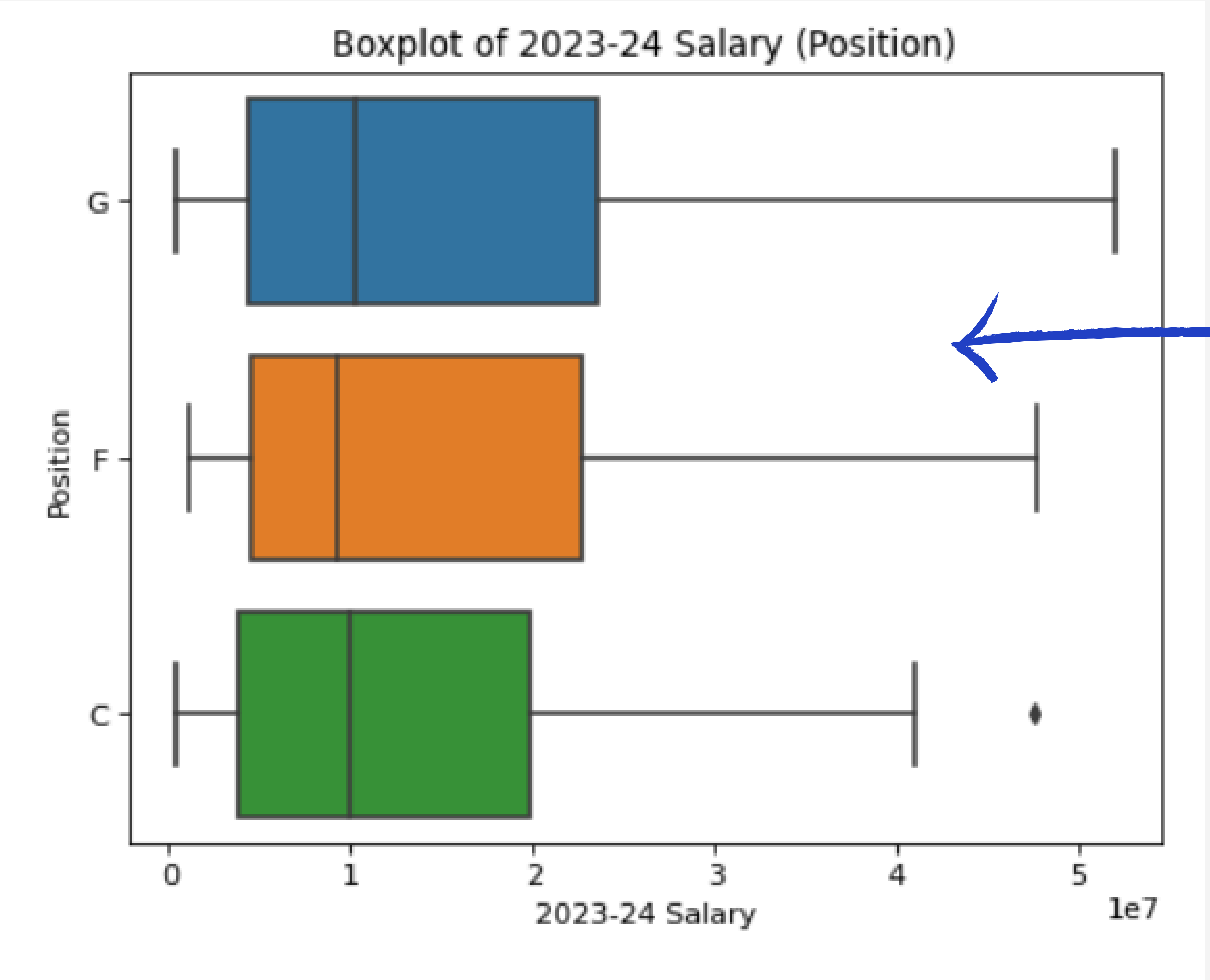
Assists per Game:
average: 2.043522267
standar deviation: 1.8844850622674214
skewness: 1.540596094519236



Field Goals:
average: 3.140283400809716
standar deviation: 2.4751511467052287
skewness: 1.0874964868490835

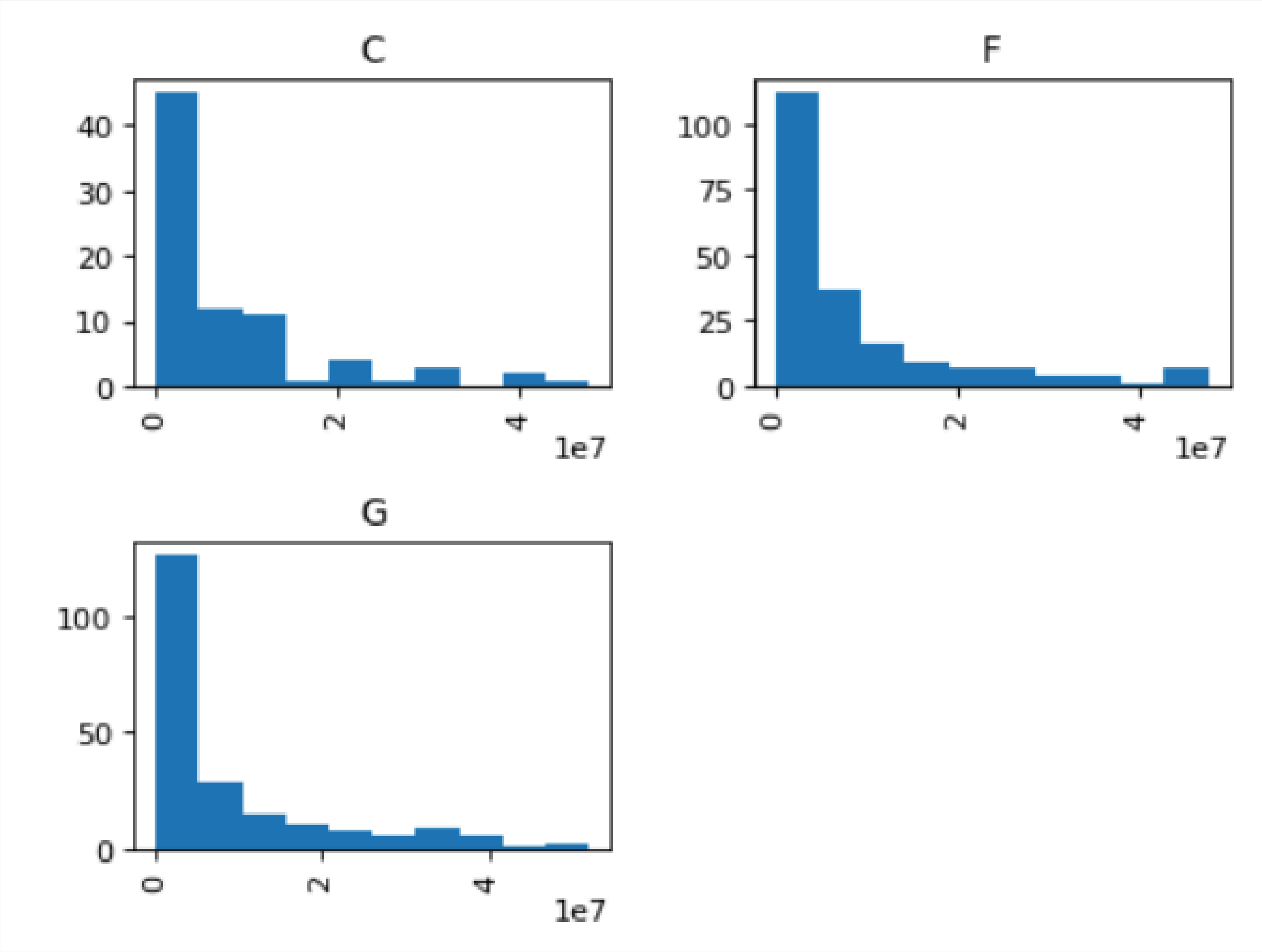


敘述統計



沒有顯著差異？

圖：2023–34 賽季各位置球員薪水盒鬚圖




```
For Pos = C
Shapiro statistic = 0.721105 and p_value = 0.000000
For Pos = F
Shapiro statistic = 0.727757 and p_value = 0.000000
For Pos = G
Shapiro statistic = 0.738136 and p_value = 0.000000
```

各位置薪水分布不為常態，故使用 Kruskal Wallis Test 作檢定

H0: 所有母體的位置皆相同
H1: 至少有兩個母體位置不同

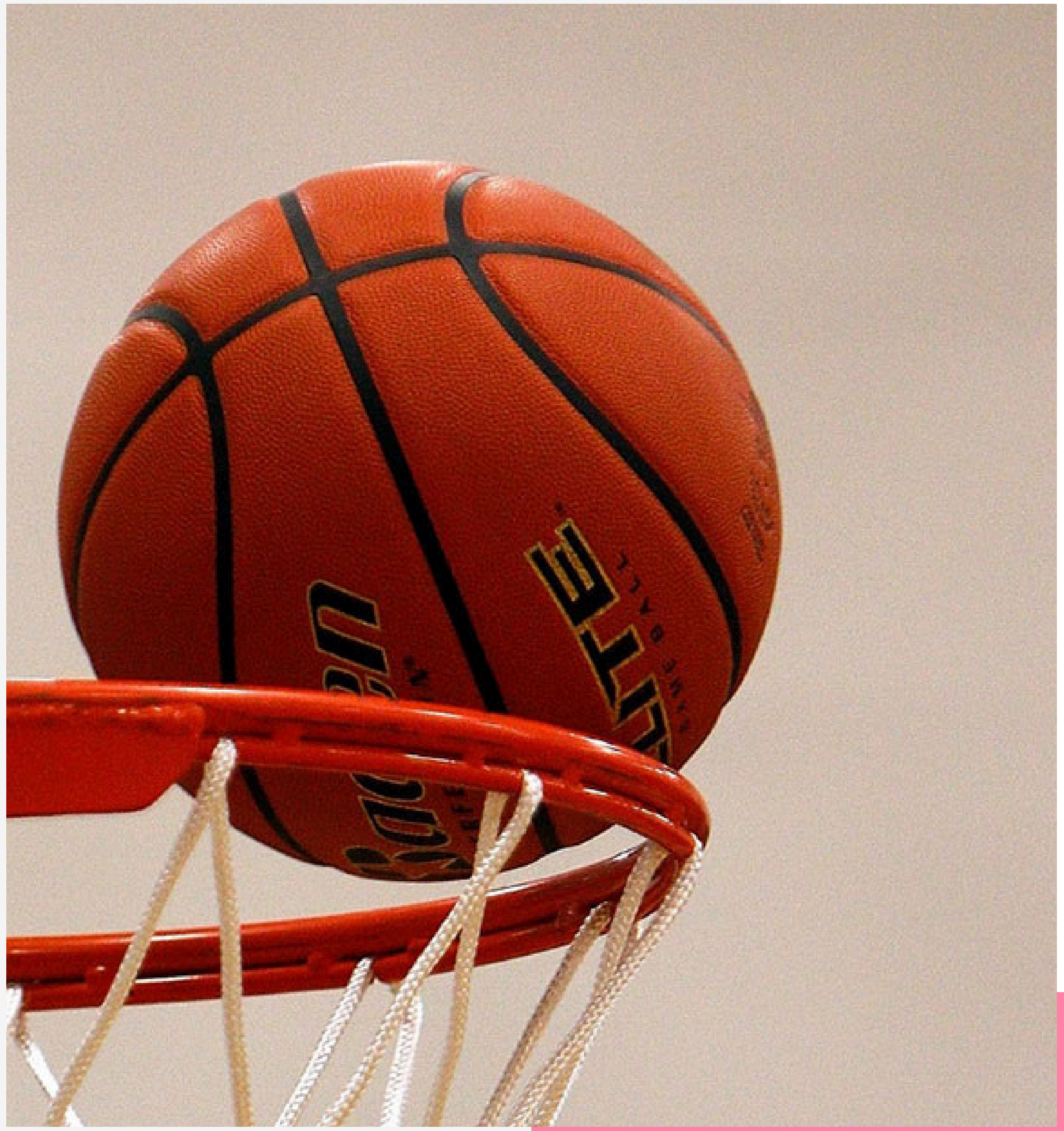
- 計算結果 p-value = 0.6245，不拒絕虛無假設，故沒有足夠證據顯示母體位置有不相同的情況。

	C	F	G
0	47607350.0	47649433.0	51915615
1	41000000.0	47607350.0	46741590
2	40600080.0	45640084.0	45640084
3	32600060.0	45640084.0	40806300
4	32459438.0	45640084.0	40064220
...	...	...	...
205	NaN	NaN	134863
206	NaN	NaN	120250
207	NaN	NaN	70687
208	NaN	NaN	64343
209	NaN	NaN	64343

210 rows × 3 columns

```
T1 = 20628.500000
T2 = 51044.000000
T3 = 50592.500000
H = 0.941534
pvalue = 0.624523
```

回歸模型



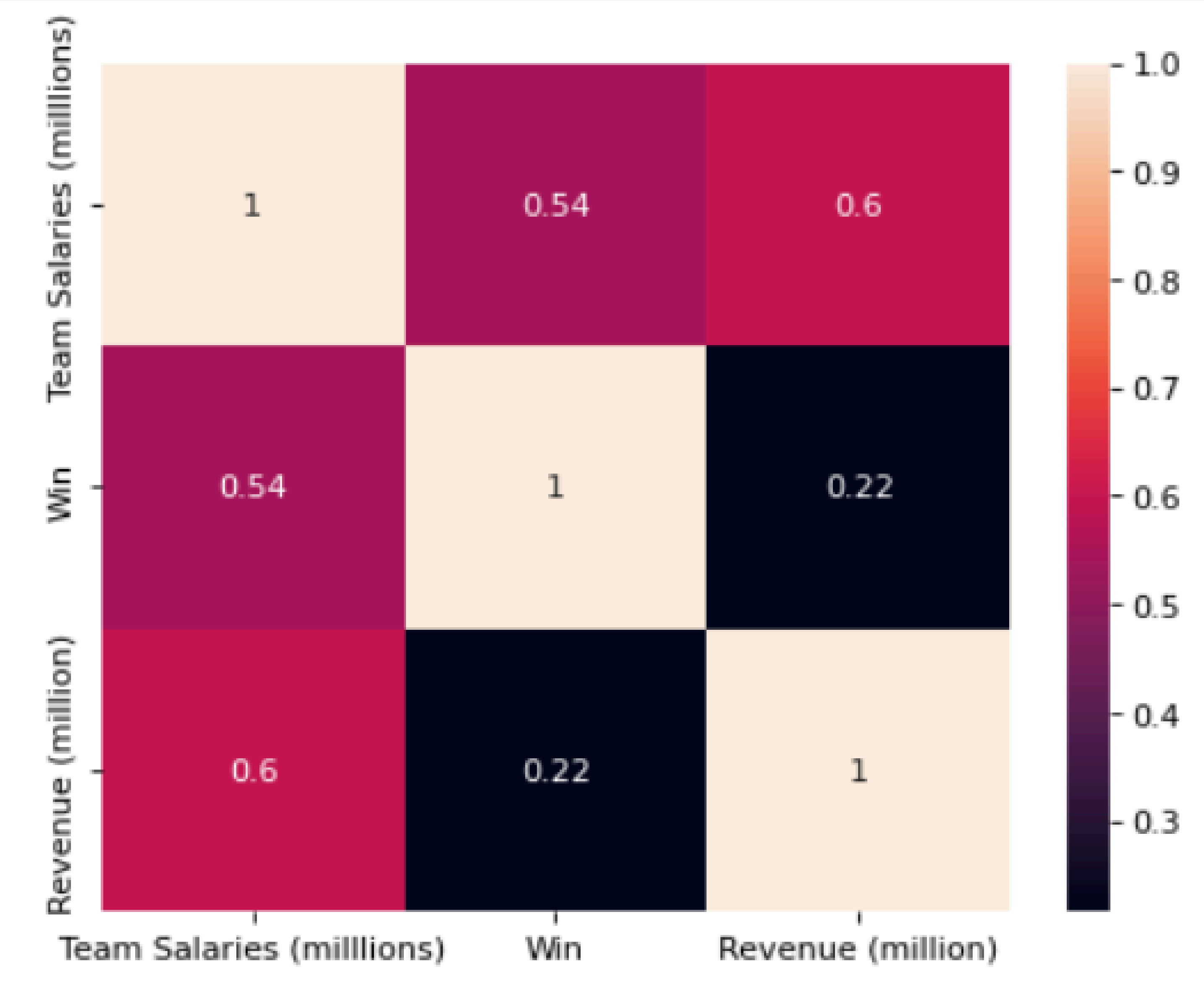
球隊薪資總額分析與模型

Foward Stepwise

注：此模型通過殘差檢定，
無法驗證是否有自回歸

```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     Team Salaries (milllions)   R-squared:                       0.533
Model:                                   OLS   Adj. R-squared:                  0.495
Method:                        Least Squares   F-statistic:                     14.24
Date:                       Wed, 29 May 2024   Prob (F-statistic):           7.45e-05
Time:                               00:04:25   Log-Likelihood:                -110.61
No. Observations:                         28   AIC:                             227.2
Df Residuals:                             25   BIC:                             231.2
Df Model:                                  2
Covariance Type:                   nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               103.2381     11.650      8.862      0.000      79.245     127.231
Win                   0.5959      0.193      3.086      0.005       0.198       0.994
Revenue (million)     0.1016      0.028      3.583      0.001       0.043       0.160
==============================================================================
Omnibus:                        0.209   Durbin-Watson:                   2.646
Prob(Omnibus):                  0.901   Jarque-Bera (JB):                0.021
Skew:                           0.057   Prob(JB):                        0.990
Kurtosis:                       2.931   Cond. No.                     1.71e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



Best subset

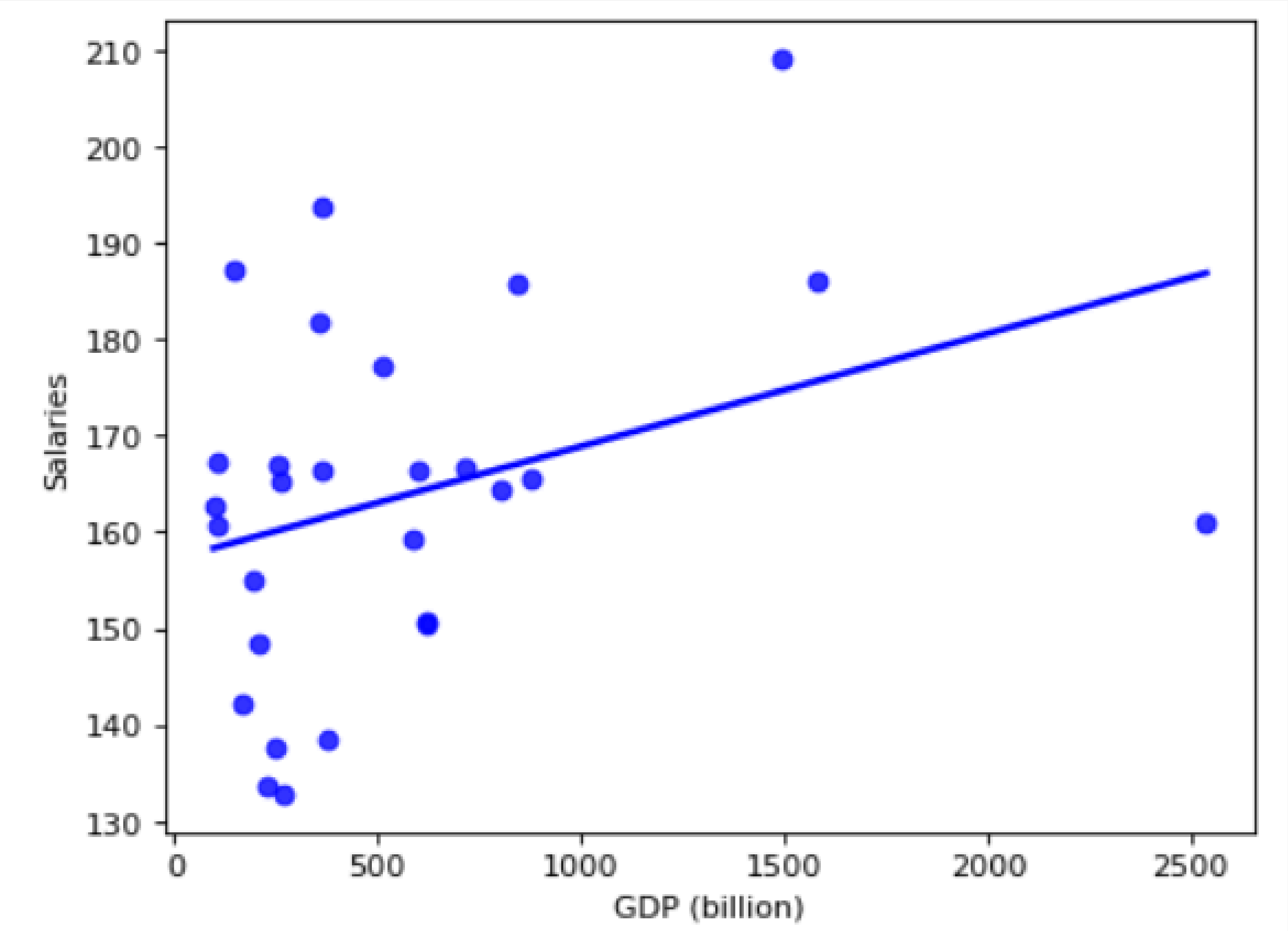
注：此模型通過殘差檢定，無法驗證是否有自回歸

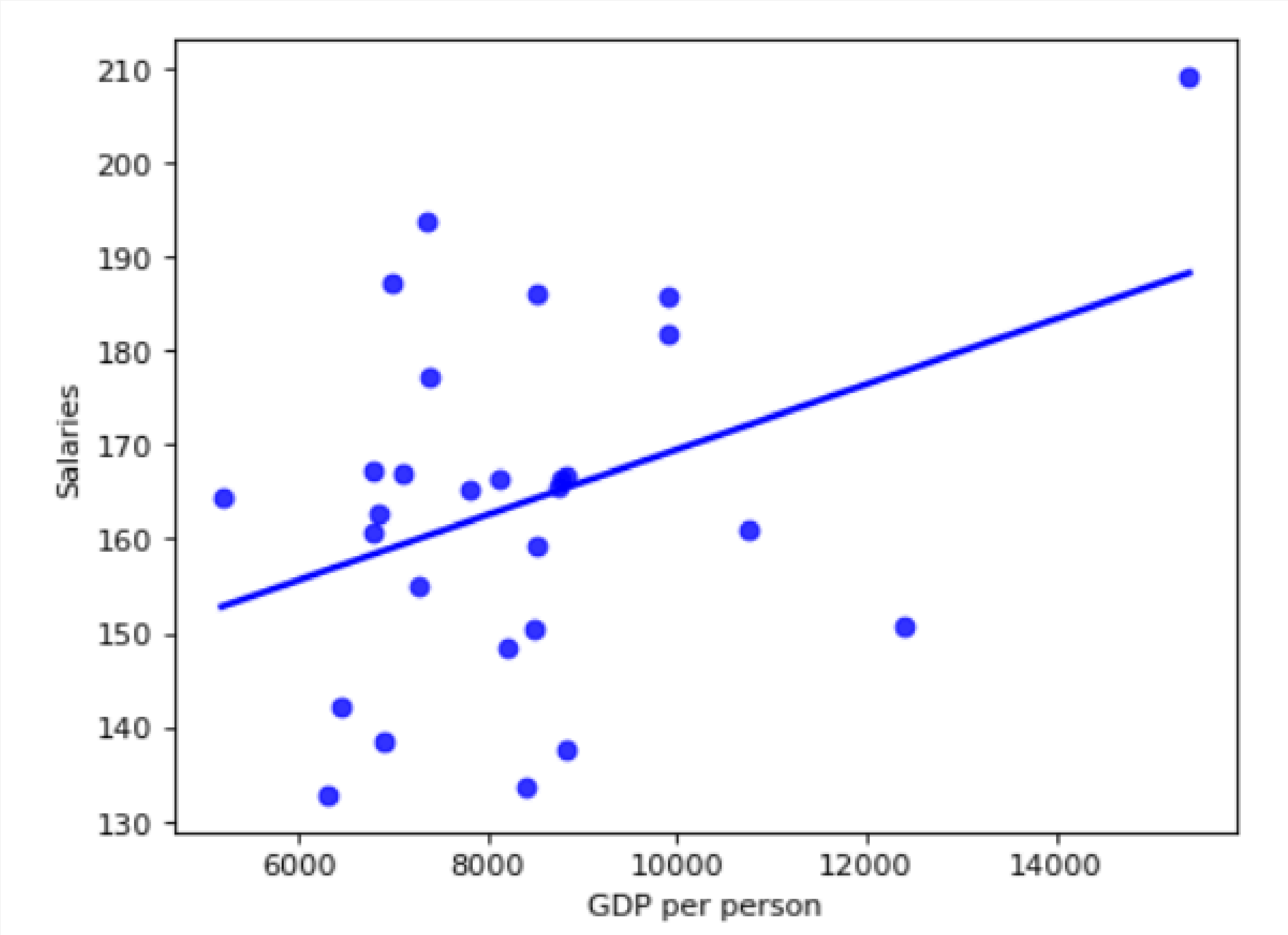
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Team Salaries (milllions)   R-squared:                       0.575
Model:                            OLS   Adj. R-squared:                  0.522
Method:                 Least Squares   F-statistic:                     10.82
Date:                Wed, 29 May 2024   Prob (F-statistic):           0.000110
Time:                        00:11:05   Log-Likelihood:                -109.28
No. Observations:                  28   AIC:                             226.6
Df Residuals:                      24   BIC:                             231.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             100.9394     11.436      8.826      0.000      77.337     124.542
Win                 0.6180      0.188      3.279      0.003       0.229       1.007
Revenue (million)   0.0973      0.028      3.511      0.002       0.040       0.155
Net Worth (m)       0.0005      0.000      1.546      0.135      -0.000       0.001
==============================================================================
Omnibus:                        0.722   Durbin-Watson:                   2.847
Prob(Omnibus):                  0.697   Jarque-Bera (JB):                0.087
Skew:                          -0.053   Prob(JB):                        0.957
Kurtosis:                       3.252   Cond. No.                     4.64e+04
==============================================================================
```

Variables

variable	interpretation
Win	在球員投資較多金額，球隊較有機會拿到較高的勝場數。
Revenue	球隊收入越高，越有意願投入更多資金在球員薪水。
Net Worth	球隊老闆身價越高，可能會願意花費較多金錢與球員簽約。

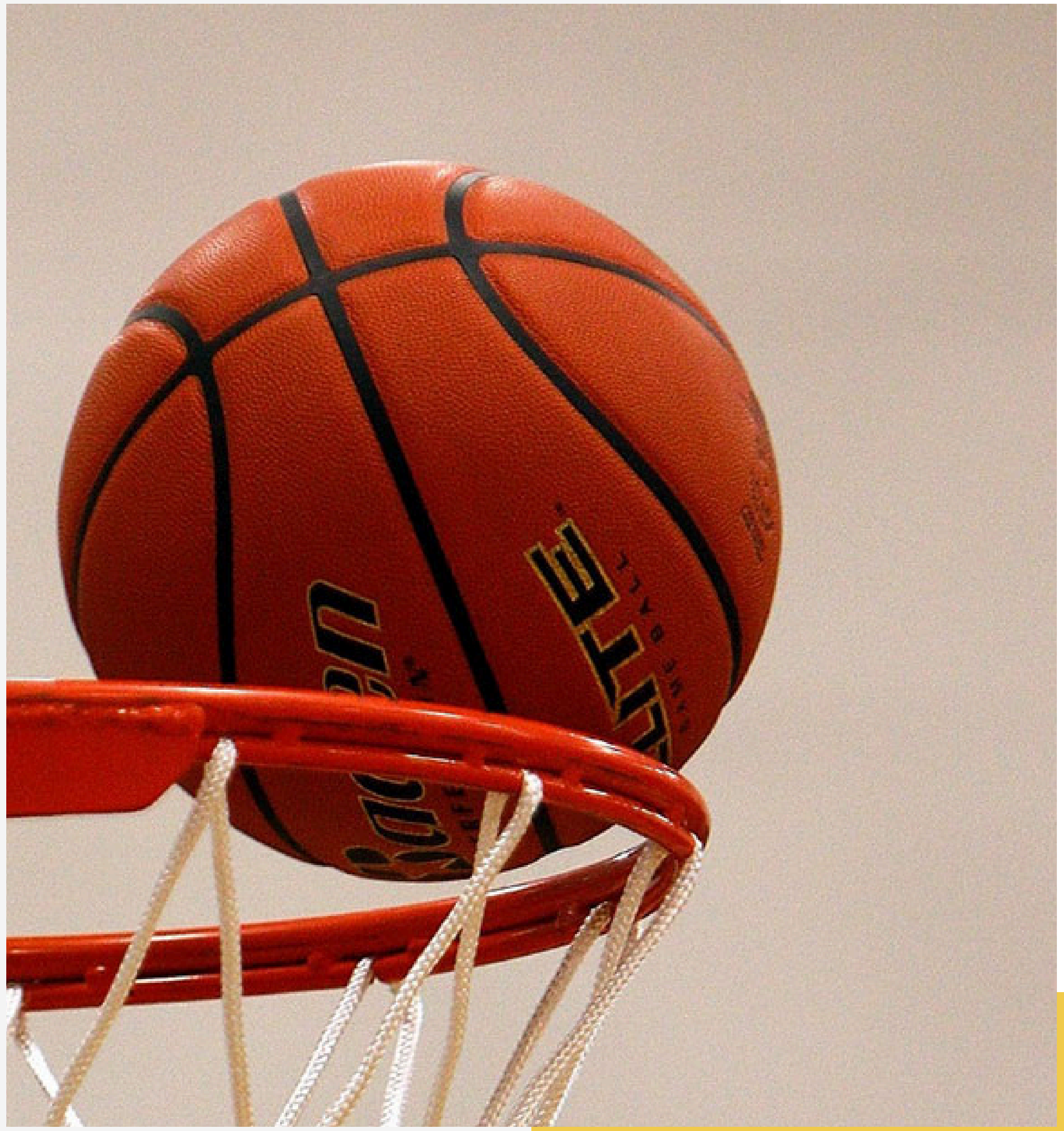
What about the Economics of City?





無明顯正相關，其模型解釋力也很低

回歸模型



球員薪水分析與模型預測

Model 1

NBA 球員年齡、上場時間、使用率、得分數對薪資的影響

猜想

- NBA 球員年齡、上場時間、使用率、得分數越高可能造成更高的薪資水準

- 使用率公式：
  - 球員出手次數+0.44*球員罰球次數+球員失誤次數)*(球隊所有球員上場時間÷5)/[球員上場時間*(球隊所有總球員出手次數+0.44*球隊所有球員罰球次數+球隊所有球員失誤次數)]

實證結果

	Dependent Variable	Independent Variables			
	log(Salary)	Intercept	Age	MP	USG%
Model 1-1	$R^2 = 0.596$	10.9139	0.0824	0.1078	
	Adj. $R^2 = 0.594$	(0.2750)	(0.0110)	(0.0040)	
	p-value	0.000	0.000	0.000	
Model 1-2	$R^2 = 0.193$	10.5124	0.1299		0.0664
	Adj. $R^2 = 0.189$	(0.433)	(0.015)		(0.010)
	p-value	0.000	0.000		0.000

Table 1: MP and USG% seperated

- **Model 1-1: Regress Salary on Age and MP**
  - **年齡跟上場時間都對薪資有很顯著且正向的影響，符合猜想。**
- **Model 1-2: Regress Salary on Age and USG%**
  - **使用率也對薪資有很顯著且正向的影響，符合猜想。**

實證結果

	Dependent Variable	Independent Variables				
	log(Salary)	Intercept	Age	MP	USG%	USG%:MP
Model 1-3	$R^2 = 0.596$	10.7876	0.0836	0.1062	0.0069	
	Adj. $R^2 = 0.594$	(0.307)	(0.011)	(0.005)	(0.007)	
	p-value	0.000	0.000	0.000	0.356	
Model 1-4	$R^2 = 0.604$	11.3814	0.0834	0.0692	-0.0221	0.0017
	Adj. $R^2 = 0.600$	(0.364)	(0.011)	(0.013)	(0.012)	(0.001)
	p-value	0.000	0.000	0.000	0.072	0.003
Model 1-5	$R^2 = 0.541$	11.1932	0.0994			0.0036
	Adj. $R^2 = 0.539$	(0.291)	(0.011)			(0.000)
	p-value	0.000	0.000			0.000

Table 2: MP and USG% mixed

- **Model 1-3: Regress Salary on Age, MP, and USG%**
  - **把年齡、使用率、上場時間同時放進模型，三者都有非常顯著的正向影響，符合猜想。**
- **Model 1-4: Regress Salary on Age, MP, USG%, and USG%*MP**
  - **USG%*MP 是該球員的總終結進攻次數佔球隊總終結進攻次數的比率。**

實證結果

	Dependent Variable	Independent Variables				
	log(Salary)	Intercept	Age	MP	USG%	USG%:MP
Model 1-3	$R^2 = 0.596$	10.7876	0.0836	0.1062	0.0069	
	Adj. $R^2 = 0.594$	(0.307)	(0.011)	(0.005)	(0.007)	
	p-value	0.000	0.000	0.000	0.356	
Model 1-4	$R^2 = 0.604$	11.3814	0.0834	0.0692	-0.0221	0.0017
	Adj. $R^2 = 0.600$	(0.364)	(0.011)	(0.013)	(0.012)	(0.001)
	p-value	0.000	0.000	0.000	0.072	0.003
Model 1-5	$R^2 = 0.541$	11.1932	0.0994			0.0036
	Adj. $R^2 = 0.539$	(0.291)	(0.011)			(0.000)
	p-value	0.000	0.000			0.000

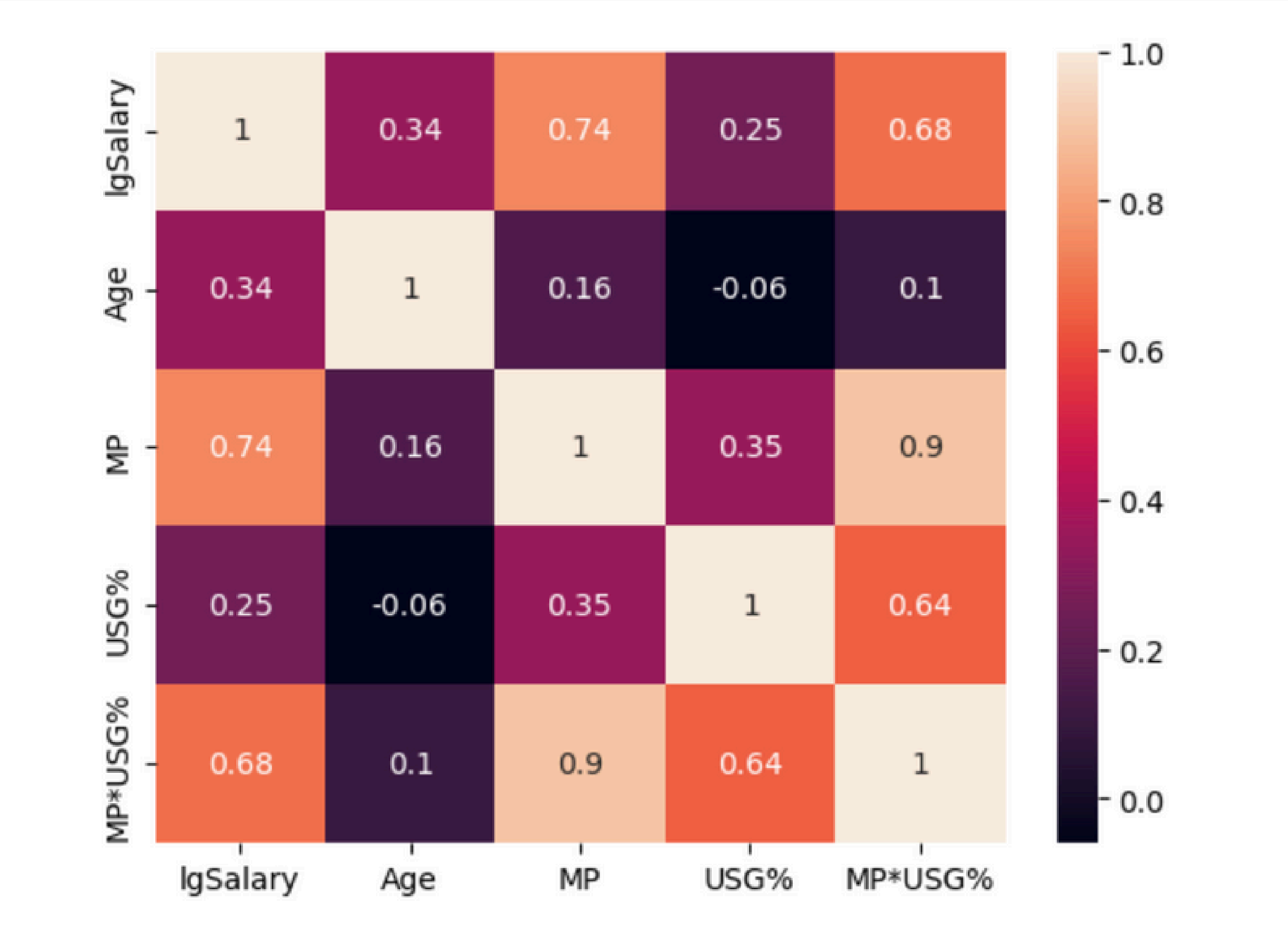
Table 2: MP and USG% mixed

- **Model 1-5: Regress Salary on Age and USG%*MP**
  - **USG%*MP意謂整場比賽中球員終結進攻次數佔球隊總終結進攻次數的比率，而終結進攻的情況只有三種，分別是出手、被犯規後罰球與失誤轉換球權，故USG%*MP越高其實可能代表此球員的得分數也越高，因為出手次數越多，得分數也很容易跟著多。**

實證結果

	Dependent Variable	Independent Variables				
	log(Salary)	Intercept	Age	MP	USG%	USG%:MP
Model 1-3	$R^2 = 0.596$	10.7876	0.0836	0.1062	0.0069	
	Adj. $R^2 = 0.594$	(0.307)	(0.011)	(0.005)	(0.007)	
	p-value	0.000	0.000	0.000	0.356	
Model 1-4	$R^2 = 0.604$	11.3814	0.0834	0.0692	-0.0221	0.0017
	Adj. $R^2 = 0.600$	(0.364)	(0.011)	(0.013)	(0.012)	(0.001)
	p-value	0.000	0.000	0.000	0.072	0.003
Model 1-5	$R^2 = 0.541$	11.1932	0.0994			0.0036
	Adj. $R^2 = 0.539$	(0.291)	(0.011)			(0.000)
	p-value	0.000	0.000			0.000

Table 2: MP and USG% mixed



- **Model 1–5: Regress Salary on Age and USG%*MP**
  - **USG%*MP意謂整場比賽中球員終結進攻次數佔球隊總終結進攻次數的比率，而終結進攻的情況只有三種，分別是出手、被犯規後罰球與失誤轉換球權，故USG%*MP越高其實可能代表此球員的得分數也越高，因為出手次數越多，得分數也很容易跟著多。**

實證結果

	Dependent Variable		Independent Variables		
	log(Salary)	Intercept	Age	USG%:MP	PTS
Model 1-5	R^2= 0.541	11.1932	0.0994	0.0036	
	Adj. R^2 = 0.539	(0.291)	(0.011)	(0.000)	
	p-value	0.000	0.000	0.000	
Model 1-6	R^2= 0.557	11.3688	0.09445		0.1490
	Adj. R^2 = 0.555	(0.285)	(0.011)		(0.007)
	p-value	0.000	0.000		0.000

Table 3: MP:USG% and PTS

- **Model 1-5: Regress Salary on Age and USG%*MP**
- **Model 1-6: Regress Salary on Age and PTS**
  - 只看年齡跟得分數的話，兩者都有很顯著的正向影響，符合猜想。
  - USG%*MP跟PTS的相關係數高達0.99 → 我們於1-5的推論被驗證，USG%*MP越高可能代表此球員的得分數也越高，因為出手次數越多，得分數也很容易跟著多。

小結

- 在這部分討論的每個模型中，年齡都對薪資呈現非常顯著且正向的影響，年齡越大代表球員越資深，薪水也跟著上漲。
- 在相同年齡的情況下，球員的上場時間越久代表球隊越器重此球員，薪資也跟著上漲。
- 在相同年齡的情況下，球員的使用率越高代表球員的進攻能力越高，薪資也跟著上漲。
- USG%*MP越高還可能代表此球員的得分數越高，因為出手次數越多，得分數也很容易跟著多。
- 第一部分的模型中，Model 1-5(Regress Salary on Age and USG%*MP)和 Model 1-6(Regress Salary on Age and PTS)有著最好的解釋效力。

Model 2

NBA 球員位置對薪資的影響

猜想

- 打不同位置的球員薪資水準可能有根本的差異
- 得分數、籃板數、助攻數、失誤數分別跟位置會有不一樣的交互影響，並可以藉此觀察到不同位置的球員較看重哪些數值

實證結果

	Dependent Variable	Independent Variables						
	log(Salary)	Intercept	Age	PTS	Guard	Forward	PTS:Guard	PTS:Forward
Model 2-1	$R^2 = 0.565$	11.5710	0.0927	0.1506	-0.3303	-0.0778		
	Adj. $R^2 = 0.562$	(0.307)	(0.011)	(0.007)	(0.133)	(8.134)		
	p-value	0.000	0.000	0.000	0.014	0.561		
Model 2-1-1	$R^2 = 0.566$	11.5149	0.0937	0.1545	-0.3436	0.0152	0.0010	-0.0115
	Adj. $R^2 = 0.561$	(0.346)	(0.011)	(0.018)	(0.214)	(0.216)	(0.021)	(0.021)
	p-value	0.000	0.000	0.000	0.110	0.944	0.961	0.591

Table 4: PTS and Position

- **Model 2-1: Regress Salary on Age, PTS, and Pos**
  - **相比年齡跟得分數，位置對薪資的影響似乎不太重要**
- **Model 2-1-1: Regress Salary on Age, PTS, Pos, and PTS*Pos**
  - **加入位置跟得分數的交乘項，但所有跟位置相關的項依然都不顯著**
  - **在年齡跟得分數的討論下，加入位置 dummy variables 沒有提高解釋效力**
- **因為發現把得分數跟位置放在一起討論看時，對位置沒有比較好的解釋意義，所以接下來我們改看幾個更細緻的數據與位置對薪資的影響**

實證結果

Dependent Variable		Independent Variables						
	log(Salary)	Intercept	Age	TRB	Guard	Forward	TRB:Guard	TRB:Forward
Model 2-2	$R^2 = 0.425$	10.1825	0.1094	0.3954	0.9979	0.7486		
	Adj. $R^2 = 0.420$	(0.370)	(0.012)	(0.025)	(0.168)	(0.161)		
	p-value	0.000	0.000	0.000	0.000	0.000		
Model 2-2-1	$R^2 = 0.457$	10.7060	0.1078	0.3060	-0.0817	0.4384	0.3277	0.0399
	Adj. $R^2 = 0.450$	(0.412)	(0.012)	(0.040)	(0.292)	(0.292)	(0.065)	(0.054)
	p-value	0.000	0.000	0.000	0.780	0.134	0.000	0.464

Table 5: TRB and Position

- **Model 2-2: Regress Salary on Age, TRB, and Pos**
  - **籃板數越高，薪資也越高，但不能直接說薪資主要受籃板數影響**
  - **籃板數跟得分數的相關係數有0.65，得分數高的球員通常有較高的籃板數**
  - **相同年齡跟籃板數的情況下，薪資水平：後衛>前鋒>中鋒，因為籃板對中鋒來說很輕鬆就撿到，然後前鋒距離籃筐更遠一些、個子也更矮一些，所以較難撿到籃板，後衛則更難撿到籃板**

實證結果

	Dependent Variable	Independent Variables						
	log(Salary)	Intercept	Age	TRB	Guard	Forward	TRB:Guard	TRB:Forward
Model 2-2	$R^2 = 0.425$	10.1825	0.1094	0.3954	0.9979	0.7486		
	Adj. $R^2 = 0.420$	(0.370)	(0.012)	(0.025)	(0.168)	(0.161)		
	p-value	0.000	0.000	0.000	0.000	0.000		
Model 2-2-1	$R^2 = 0.457$	10.7060	0.1078	0.3060	-0.0817	0.4384	0.3277	0.0399
	Adj. $R^2 = 0.450$	(0.412)	(0.012)	(0.040)	(0.292)	(0.292)	(0.065)	(0.054)
	p-value	0.000	0.000	0.000	0.780	0.134	0.000	0.464

Table 5: TRB and Position

- **Model 2-2-1: Regress Salary on Age, TRB, Pos, and TRB*Pos**
  - **後衛的籃板數對薪資的影響較前鋒跟中鋒大，推測可能是因為後衛較難取得籃板，通常只有明星球員才做得到**

實證結果

	Dependent Variable	Independent Variables						
	log(Salary)	Intercept	Age	AST	Guard	Forward	AST:Guard	AST:Forward
Model 2-3	$R^2 = 0.465$	12.3802	0.0796	0.5134	-0.8438	-0.1346		
	Adj. $R^2 = 0.461$	(0.341)	(0.012)	(0.029)	(0.153)	(0.148)		
	p-value	0.000	0.000	0.000	0.000	0.365		
Model 2-3-1	$R^2 = 0.470$	12.3622	0.0771	0.5734	-0.6499	-0.1815	-0.1012	0.0206
	Adj. $R^2 = 0.463$	(0.371)	(0.012)	(0.094)	(0.220)	(0.216)	(0.100)	(0.109)
	p-value	0.000	0.000	0.000	0.003	0.402	0.313	0.850

Table 6: AST and Position

- **Model 2-3: Regress Salary on Age, AST, and Pos**
  - 助攻數越高，薪資也越高，但不能直接說薪資主要受助攻數影響
  - 助攻數跟得分數的相關係數高達0.79，得分數高的球員通常有較高的助攻數
  - 相同年齡與助攻數下，後衛的薪資水平比前鋒跟中鋒都低，這可能是因為後衛比較好拿助攻，換句話說，相同薪資水平下，後衛的助攻數理應較高
- **Model 2-3-1: Regress Salary on Age, AST, Pos, and AST*Pos**
  - 相同年齡與助攻數下，後衛的薪資水平較前鋒跟中鋒低，可能是因為後衛比較好拿助攻，結論同 Model 2-3

實證結果

	Dependent Variable	Independent Variables						
	log(Salary)	Intercept	Age	TOV	Guard	Forward	TOV:Guard	TOV:Forward
Model 2-4	$R^2 = 0.505$	11.2544	0.1053	1.1897	-0.2693	0.0761		
	Adj. $R^2 = 0.501$	(0.329)	(0.012)	(0.061)	(0.142)	(0.143)		
	p-value	0.000	0.000	0.000	0.059	0.594		
Model 2-4-1	$R^2 = 0.505$	11.2529	0.1057	1.1811	-0.2991	0.0917	0.0285	-0.0178
	Adj. $R^2 = 0.499$	(0.374)	(0.012)	(0.163)	(0.233)	(0.235)	(0.184)	(0.194)
	p-value	0.000	0.000	0.000	0.200	0.696	0.877	0.927

Table 7: TOV and Position

- **Model 2-4: Regress Salary on Age, TOV, Pos**
  - **不能直接說薪資主要受失誤數影響，推論是因為當一個球員在場上越為重要時，持球時間也越長，因此更有機會造成失誤。**
- **Model 2-4-1: Regress Salary on Age, TOV, Pos, and TOV*Pos**
  - **有位置 dummy variables 的項都不顯著 → 相同失誤數的情況下不同位置似乎對薪資沒什麼影響**

小結

- 相比年齡跟得分數，位置對薪資的影響似乎不太重要
- 相同年齡跟籃板數的情況下，薪資水平：後衛>前鋒>中鋒，因為籃板對中鋒來說很輕鬆就撿到，然後前鋒距離籃筐更遠一些、個子也更矮一些，所以較難撿到籃板，後衛則更難撿到籃板，而難度越高的情況下又能達到一樣的籃板數，所以薪資才會較高
- 相同年齡跟助攻數的情況下，後衛的薪資水平比前鋒跟中鋒都低，這可能是因為後衛比較好拿助攻，換句話說，相同薪資水平下，後衛的助攻數應該要比較高才合理
- 相同年齡跟失誤數的情況下，不同位置對薪資並無顯著影響

Model 3

NBA 球員球隊所在地區經濟水平對薪資的影響

猜想

- 球隊所在地區的經濟水平越高，球員的薪資也會跟著越高

實證結果

	Dependent Variable	Independent Variables					
	log(Salary)	Intercept	Age	PTS	USG%:MP	Level1	Level2
	$R^2 = 0.120$	11.8417	0.1240			0.1007	-0.1084
Model 3-1	Adj. $R^2 = 0.115$	(0.410)	(0.015)			(0.177)	(0.199)
	p-value	0.000	0.000			0.569	0.586
	$R^2 = 0.545$	11.1210	0.0984		0.0036	0.1787	-0.0093
Model 3-2	Adj. $R^2 = 0.541$	(0.297)	(0.011)		(0.000)	(0.127)	(0.143)
	p-value	0.000	0.000		0.000	0.161	0.948
	$R^2 = 0.5611$	11.3020	0.0936	0.1491		0.1701	-0.0253
Model 3-3	Adj. $R^2 = 0.557$	(0.291)	(0.011)	(0.007)		(0.125)	(0.141)
	p-value	0.000	0.000	0.000		0.174	0.857

Table 8: GDP Level

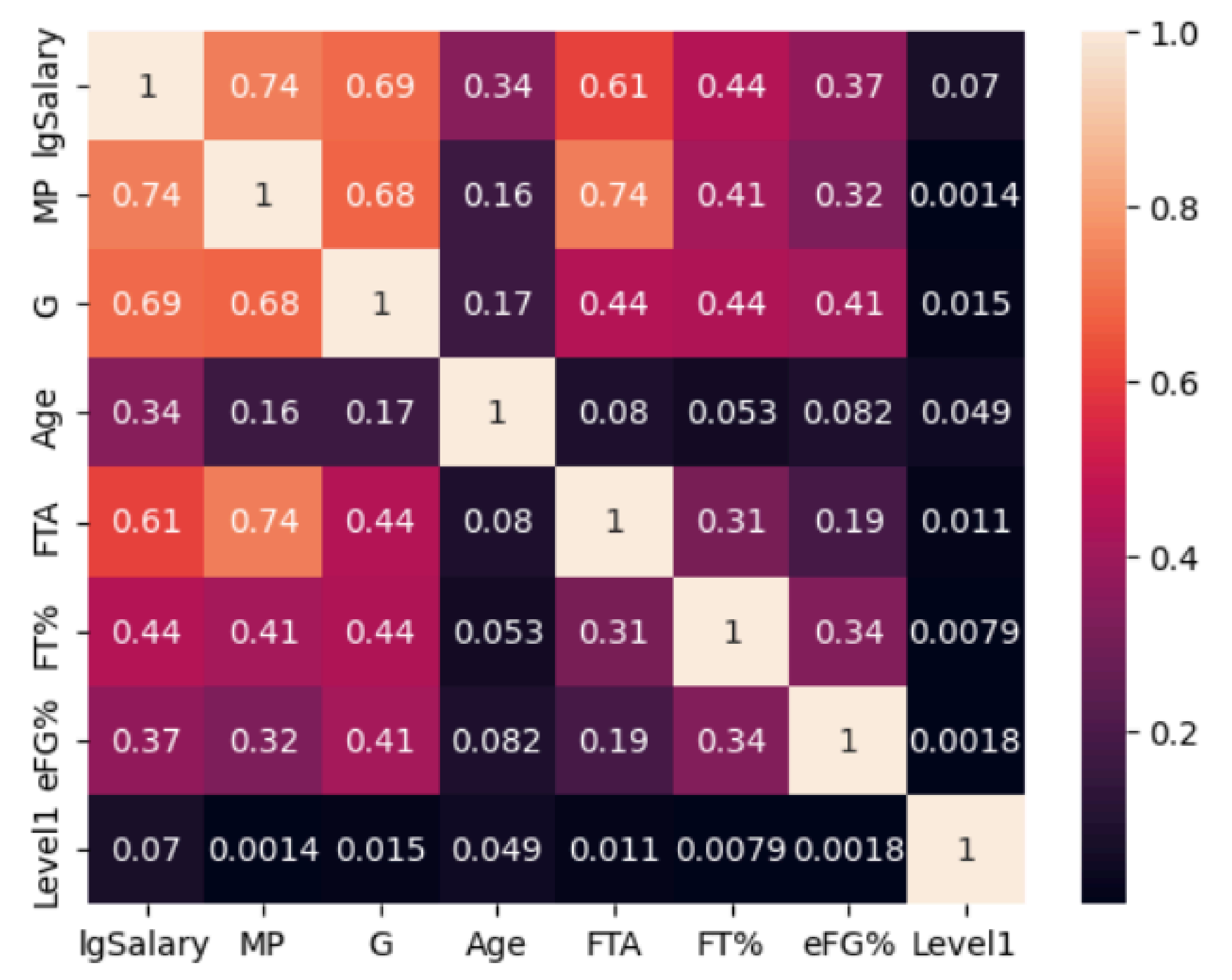
- **Model 3-1: Regress Salary on Age and Level**
- **Model 3-2: Regress Salary on Age, USG%*MP, and Level**
- **Model 3-3: Regress Salary on Age, PTS, and Level**
  - **以上三個 Model 都看不出 Level 有什麼顯著的影響**

小結

- 球隊所在地區的經濟水平高低對球員薪資並沒有顯著的影響，其可能的原因是用 GDP 看的區域經濟水平囊括的面向太廣了，地區經濟興盛不代表此地區的球隊薪資水平就比較高

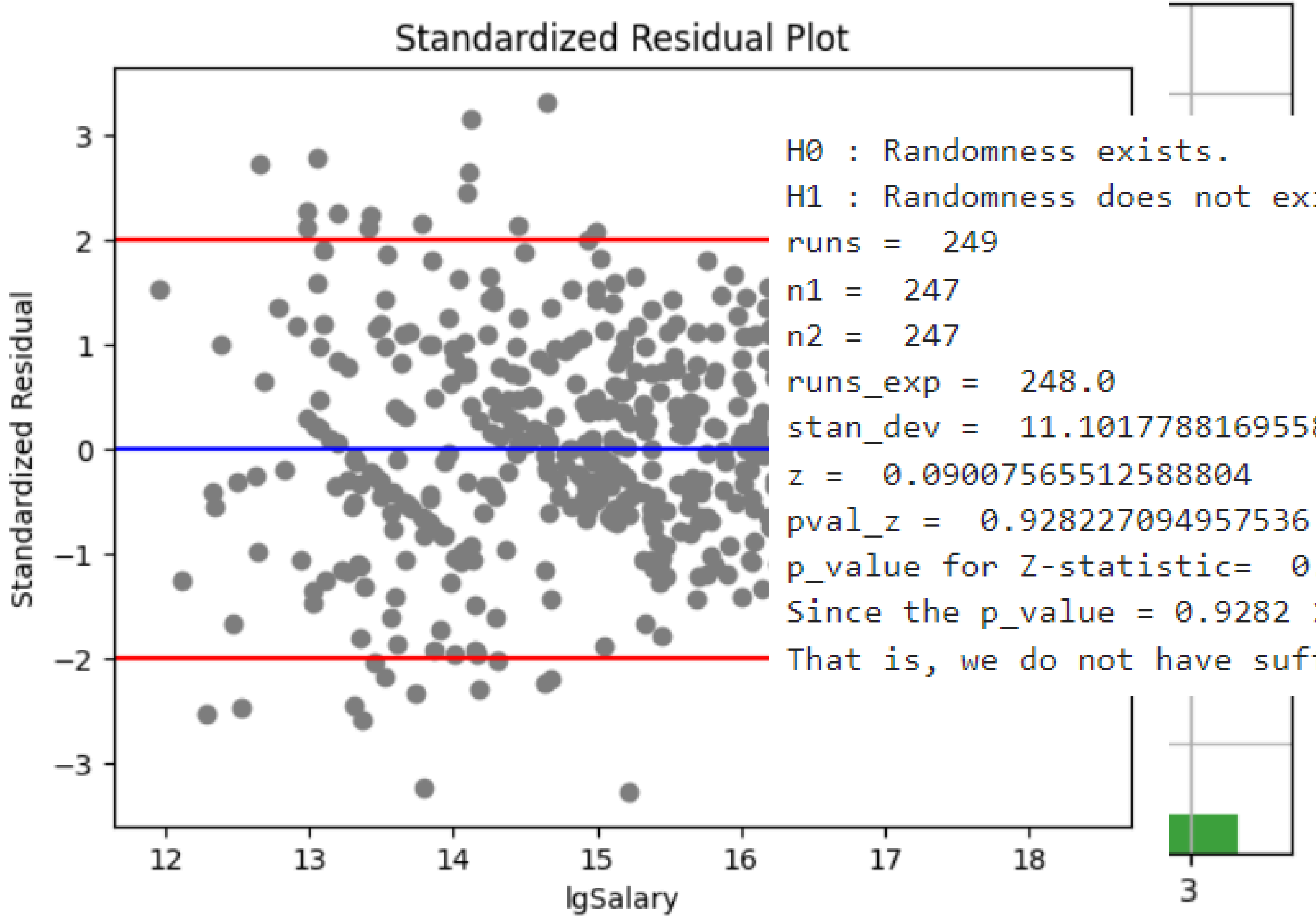
Forward-All Players

```
Multicollinearity exists between [('lgSalary', 'MP'), ('MP', 'FTA')].
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.690
Model:                            OLS   Adj. R-squared:                  0.685
Method:                 Least Squares   F-statistic:                     154.3
Date:                Tue, 28 May 2024   Prob (F-statistic):          3.68e-119
Time:                        22:04:57   Log-Likelihood:                -620.93
No. Observations:                 494   AIC:                             1258.
Df Residuals:                     486   BIC:                             1291.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0972      0.297     33.943      0.000       9.513      10.682
MP             0.0444      0.007      6.212      0.000       0.030       0.058
G              0.0181      0.002      8.178      0.000       0.014       0.022
Age            0.0779      0.009      8.333      0.000       0.059       0.096
FTA            0.1797      0.034      5.327      0.000       0.113       0.246
FT%            0.6342      0.191      3.315      0.001       0.258       1.010
eFG%           0.8920      0.375      2.379      0.018       0.155       1.629
Level1         0.1595      0.078      2.051      0.041       0.007       0.312
==============================================================================
Omnibus:                        4.000   Durbin-Watson:                   2.027
Prob(Omnibus):                  0.135   Jarque-Bera (JB):                4.860
Skew:                           0.021   Prob(JB):                       0.0880
Kurtosis:                       3.484   Cond. No.                         662.
==============================================================================
```



Forward–All Players

H0: Homoskedasticity
H1: Heteroskedasticity

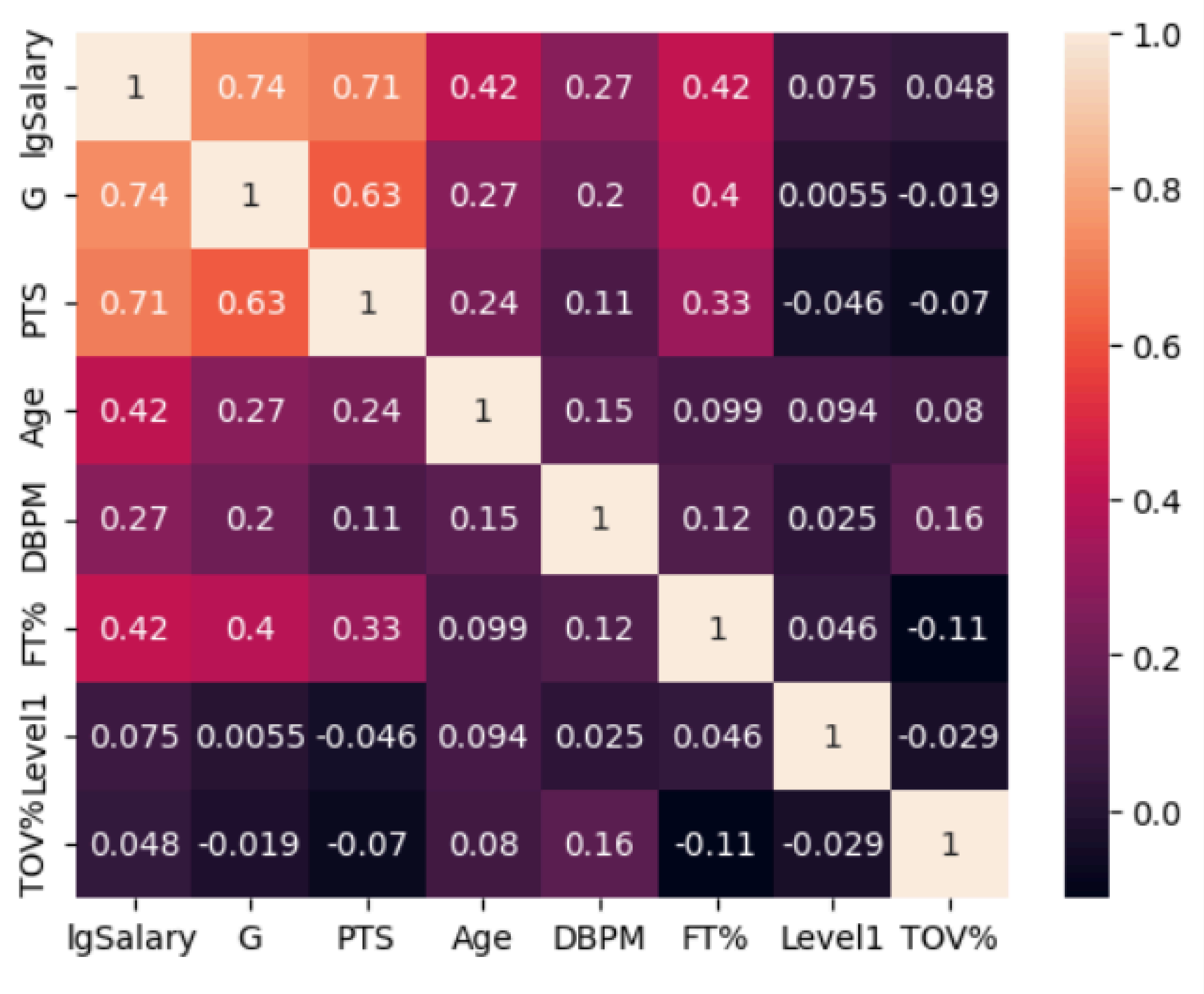



```
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  249
n1 =  247
n2 =  247
runs_exp =  248.0
stan_dev =  11.1017788169558
z =  0.09007565512588804
pval_z =  0.928227094957536
p_value for Z-statistic=  0.928227094957536
Since the p_value = 0.9282 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Do not rejected H0. Heteroscedasticity does not appear to be a problem.

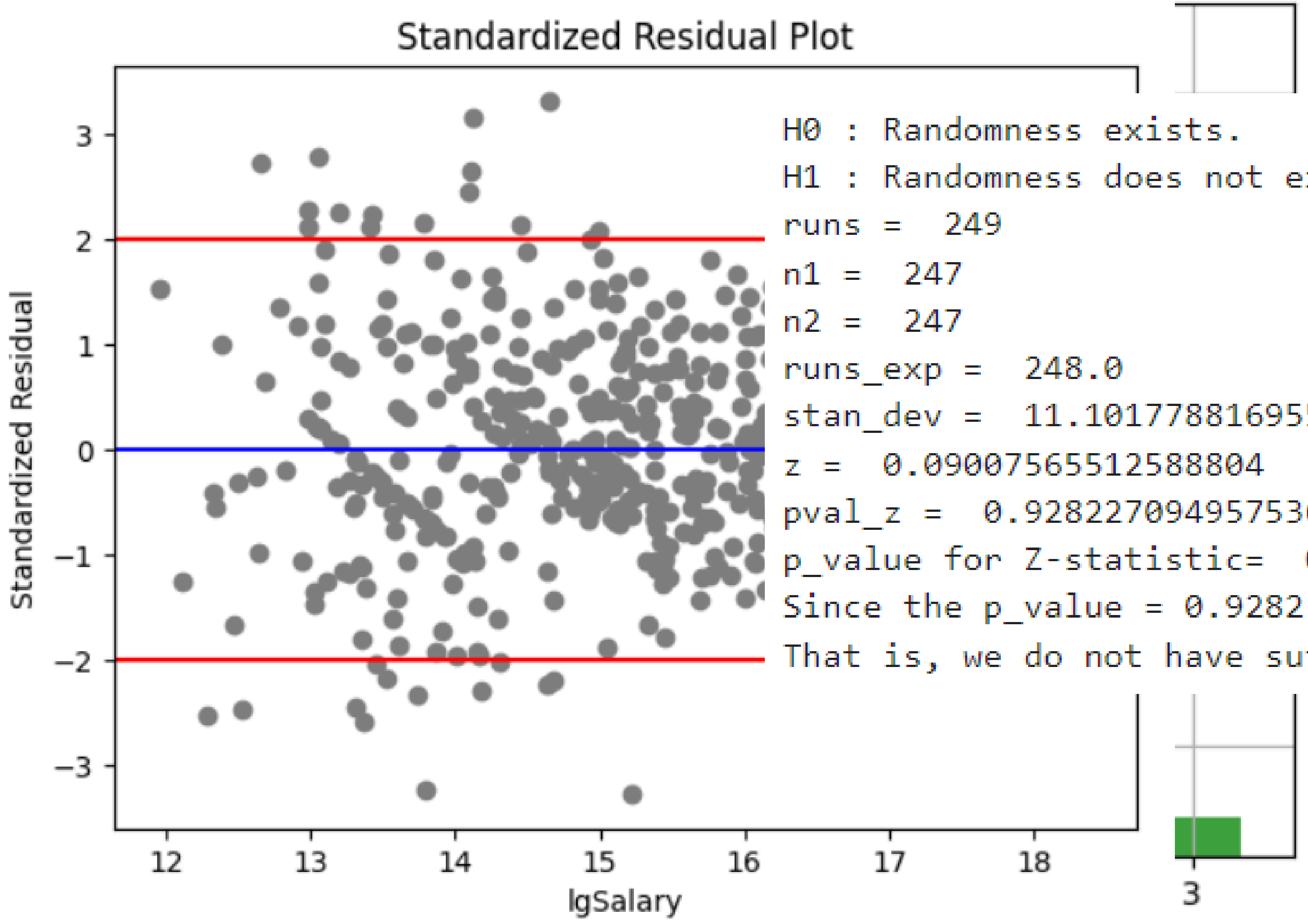
Forward–Forward

```
Multicollinearity exists between [('lgSalary', 'G'), ('lgSalary', 'PTS')].
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.705
Method:                 Least Squares   F-statistic:                     70.47
Date:                Tue, 28 May 2024   Prob (F-statistic):           3.94e-50
Time:                        22:11:52   Log-Likelihood:                -240.69
No. Observations:                 204   AIC:                             497.4
Df Residuals:                     196   BIC:                             523.9
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.8691      0.398     27.301      0.000      10.084      11.654
G              0.0225      0.003      7.292      0.000       0.016       0.029
PTS            0.0858      0.011      7.766      0.000       0.064       0.108
Age            0.0640      0.014      4.710      0.000       0.037       0.091
DBPM           0.0925      0.037      2.472      0.014       0.019       0.166
FT%            0.7570      0.267      2.830      0.005       0.229       1.285
Level1         0.1999      0.115      1.743      0.083      -0.026       0.426
TOV%           0.0205      0.012      1.692      0.092      -0.003       0.044
==============================================================================
Omnibus:                        3.197   Durbin-Watson:                   2.023
Prob(Omnibus):                  0.202   Jarque-Bera (JB):                3.174
Skew:                          -0.142   Prob(JB):                        0.205
Kurtosis:                       3.541   Cond. No.                         445.
==============================================================================
```



Forward–Forward

H0: Homoskedasticity
H1: Heteroskedasticity

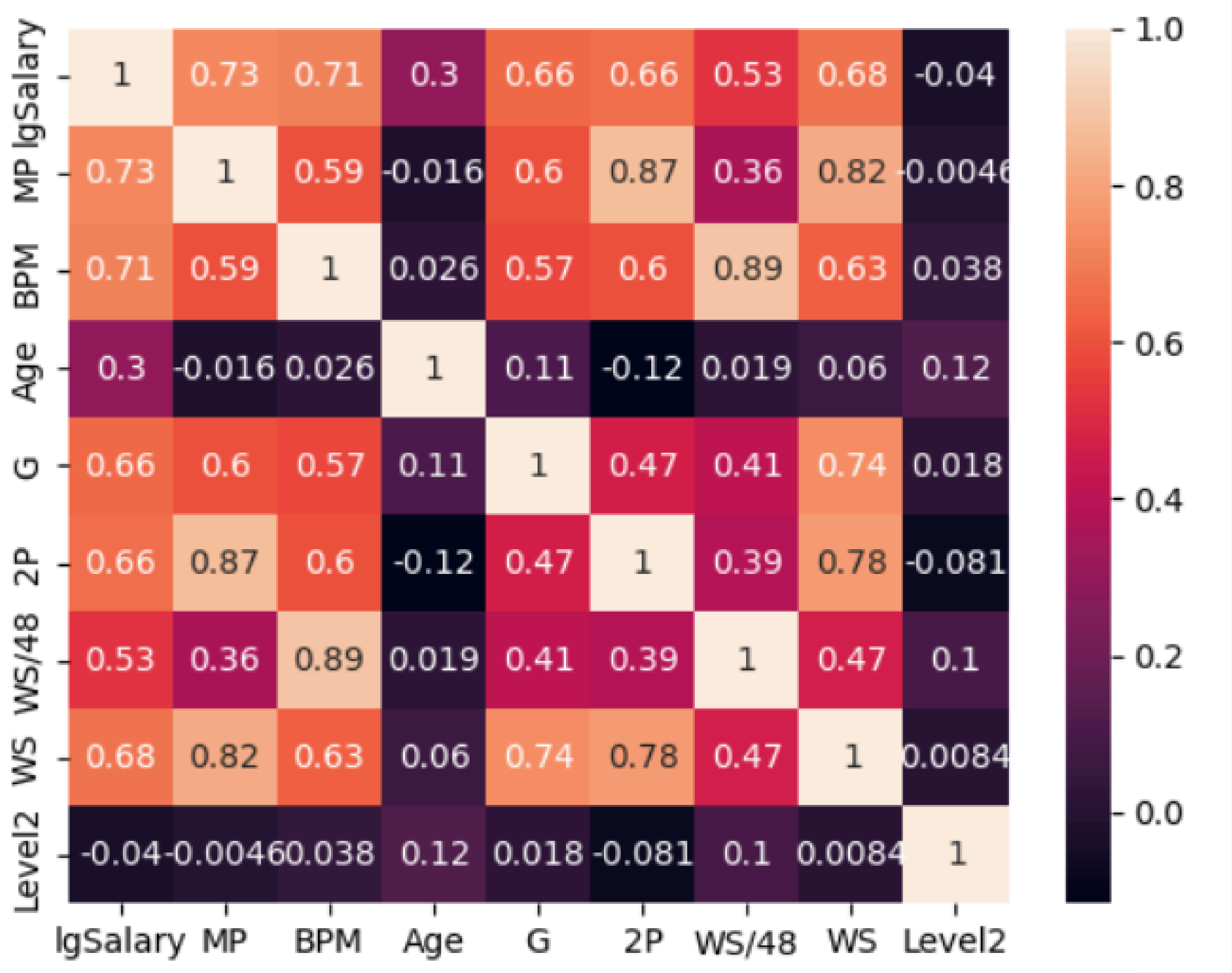


H0 : Randomness exists.
H1 : Randomness does not exist.
runs = 249
n1 = 247
n2 = 247
runs_exp = 248.0
stan_dev = 11.1017788169558
z = 0.09007565512588804
pval_z = 0.928227094957536
p_value for Z-statistic= 0.928227094957536
Since the p_value = 0.9282 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.

Do not rejected H0. Heteroscedasticity does not appear to be a problem.

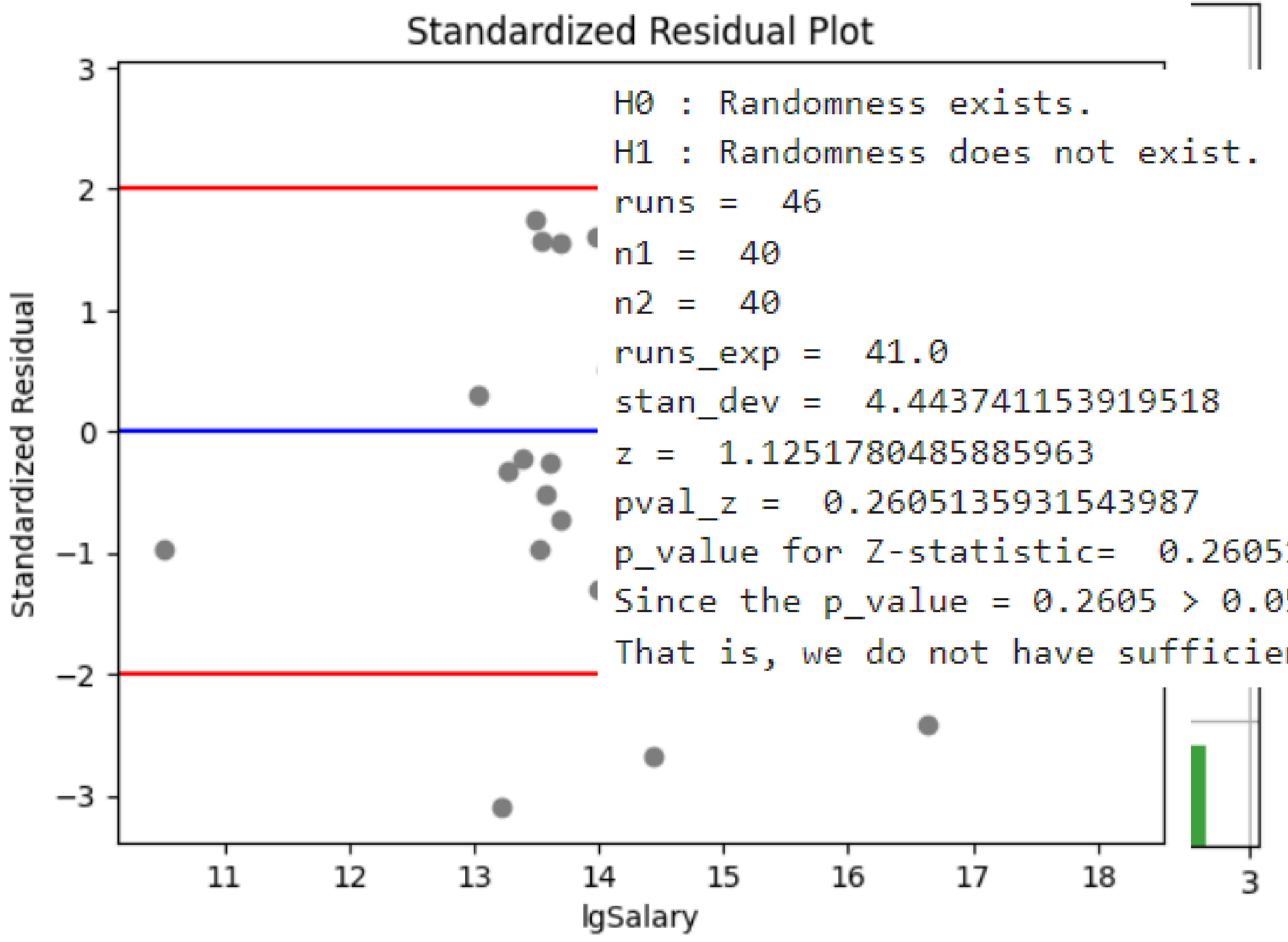
Forward-Center

```
Multicollinearity exists between [('lgSalary', 'MP'), ('lgSalary', 'BPM'), ('MP', '2P'), ('MP', 'WS'), ('BPM', 'WS/48'), ('G', 'WS'), ('2P', 'WS')].
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.777
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     30.89
Date:                Tue, 28 May 2024   Prob (F-statistic):           3.18e-20
Time:                        22:47:10   Log-Likelihood:                -79.142
No. Observations:                  80   AIC:                             176.3
Df Residuals:                      71   BIC:                             197.7
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.7020      0.791     13.538      0.000       9.126      12.278
MP             0.0615      0.021      2.907      0.005       0.019       0.104
BPM            0.1145      0.054      2.119      0.038       0.007       0.222
Age            0.1054      0.020      5.198      0.000       0.065       0.146
G              0.0156      0.006      2.705      0.009       0.004       0.027
2P             0.1122      0.086      1.299      0.198      -0.060       0.284
WS/48         -0.1217      2.190     -0.056      0.956      -4.488       4.245
WS            -0.0975      0.060     -1.626      0.108      -0.217       0.022
Level2        -0.2461      0.184     -1.340      0.185      -0.612       0.120
==============================================================================
Omnibus:                        3.853   Durbin-Watson:                   2.176
Prob(Omnibus):                  0.146   Jarque-Bera (JB):                3.836
Skew:                          -0.199   Prob(JB):                        0.147
Kurtosis:                       3.996   Cond. No.                     1.84e+03
==============================================================================
```



Forward–Center

H0: Homoskedasticity
H1: Heteroskedasticity

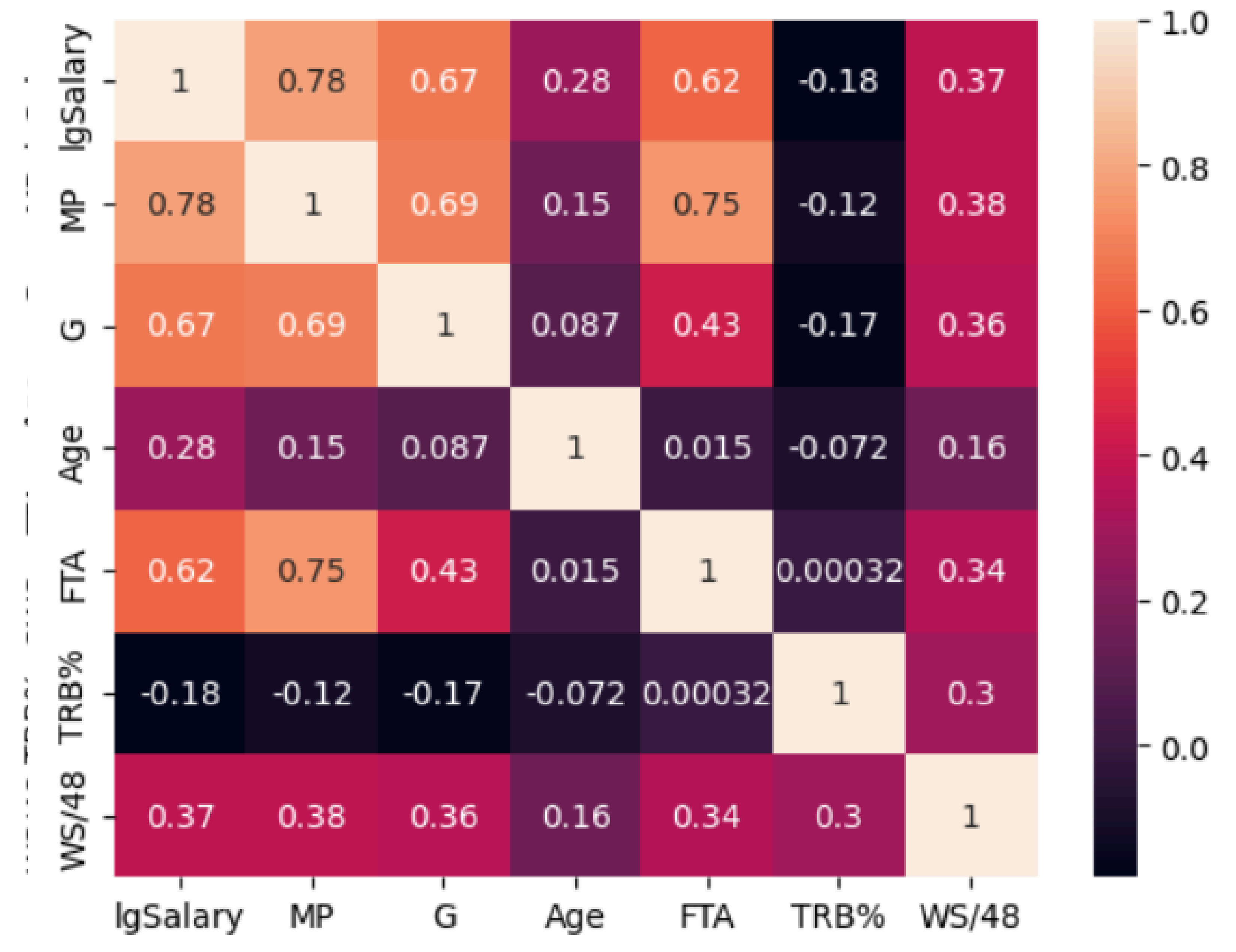


H0 : Randomness exists.
H1 : Randomness does not exist.
runs = 46
n1 = 40
n2 = 40
runs_exp = 41.0
stan_dev = 4.443741153919518
z = 1.1251780485885963
pval_z = 0.2605135931543987
p_value for Z-statistic= 0.2605135931543987
Since the p_value = 0.2605 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.

Do not rejected H0. Heteroscedasticity does not appear to be a problem.

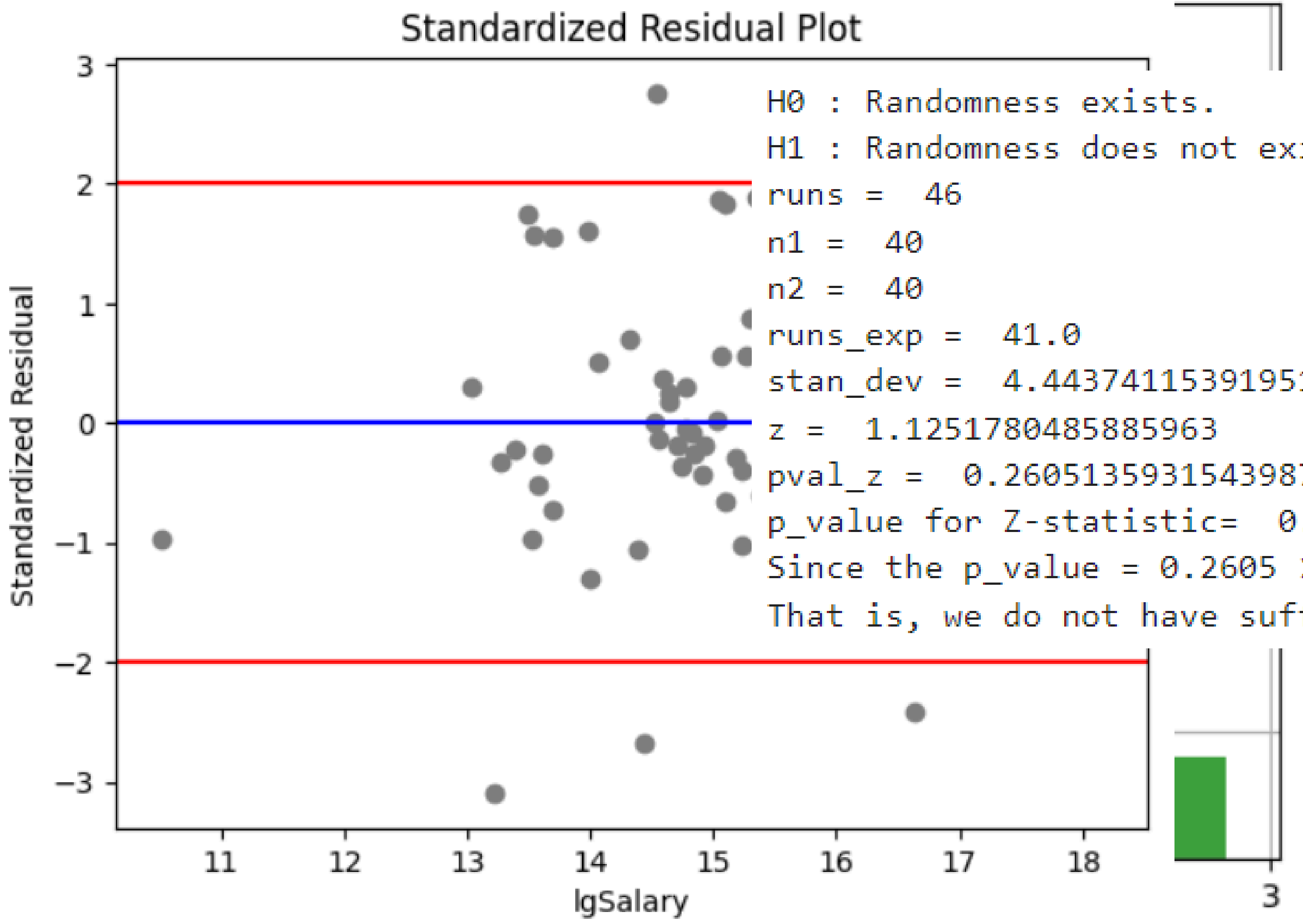
Forward–Guard

```
Multicollinearity exists between [('lgSalary', 'MP'), ('MP', 'FTA')].
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.695
Model:                            OLS   Adj. R-squared:                  0.686
Method:                 Least Squares   F-statistic:                     77.07
Date:                Tue, 28 May 2024   Prob (F-statistic):           1.20e-49
Time:                        23:58:17   Log-Likelihood:                -274.70
No. Observations:                 210   AIC:                             563.4
Df Residuals:                     203   BIC:                             586.8
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.2112      0.452     24.790      0.000      10.319      12.103
MP             0.0632      0.012      5.230      0.000       0.039       0.087
G              0.0154      0.003      4.542      0.000       0.009       0.022
Age            0.0703      0.016      4.414      0.000       0.039       0.102
FTA            0.1788      0.060      2.995      0.003       0.061       0.296
TRB%          -0.0399      0.017     -2.338      0.020      -0.073      -0.006
WS/48          1.1399      0.785      1.451      0.148      -0.409       2.688
==============================================================================
Omnibus:                        3.153   Durbin-Watson:                   2.109
Prob(Omnibus):                  0.207   Jarque-Bera (JB):                2.776
Skew:                           0.222   Prob(JB):                        0.250
Kurtosis:                       3.348   Cond. No.                         780.
==============================================================================
```



Forward–Guard

H0: Homoskedasticity
H1: Heteroskedasticity




```
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  46
n1 =  40
n2 =  40
runs_exp =  41.0
stan_dev =  4.443741153919518
z =  1.1251780485885963
pval_z =  0.2605135931543987
p_value for Z-statistic=  0.2605135931543987
Since the p_value = 0.2605 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Do not rejected H0. Heteroscedasticity does not appear to be a problem.

Forward

position	variable	r squared
all players	MP, G, Age, FTA, FT%, FG%, OWS, Level1	0.690
forward	G, PTS, Age, DBPM, FT%, Level1, TOV%	0.716
center	MP, BPM, Age, G, VORP, 2P, WS/48, WS, Level2	0.777
guard	MP, G, Age, FTA, OWS, TRB%, WS/48	0.695

Forward

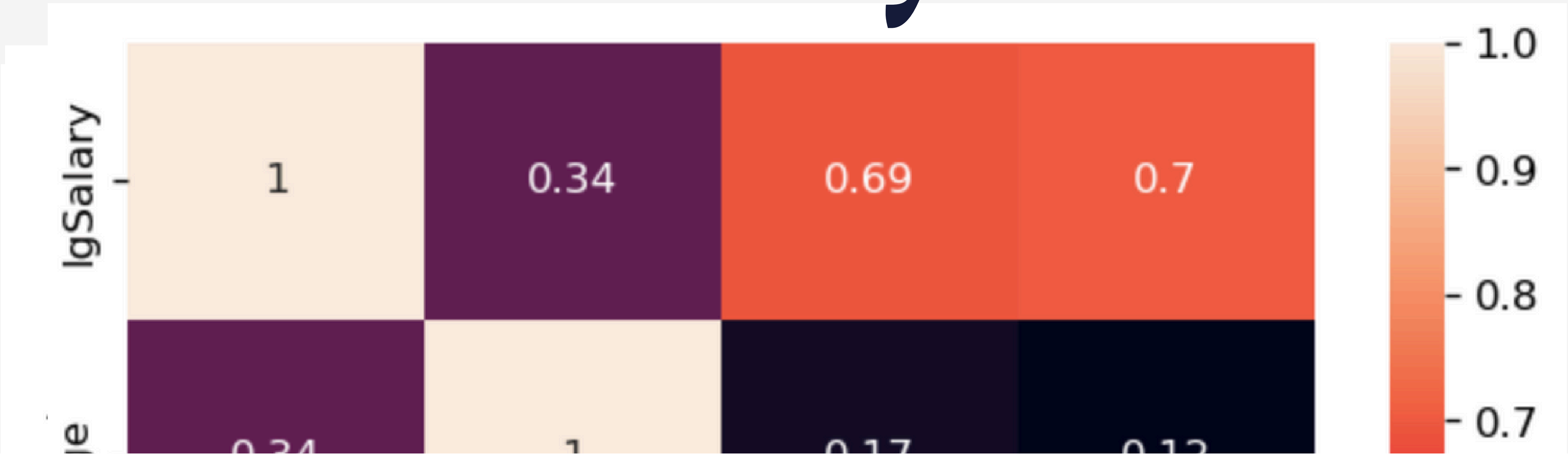
- Age 跟 G 在所有回歸模型中皆有被選中

- 在 All players 跟 Forward 中，Level 1 是顯著 variable
- 在 Center 中，Level 2 是顯著 variable
- 在 Guard 中，兩者皆沒被選中

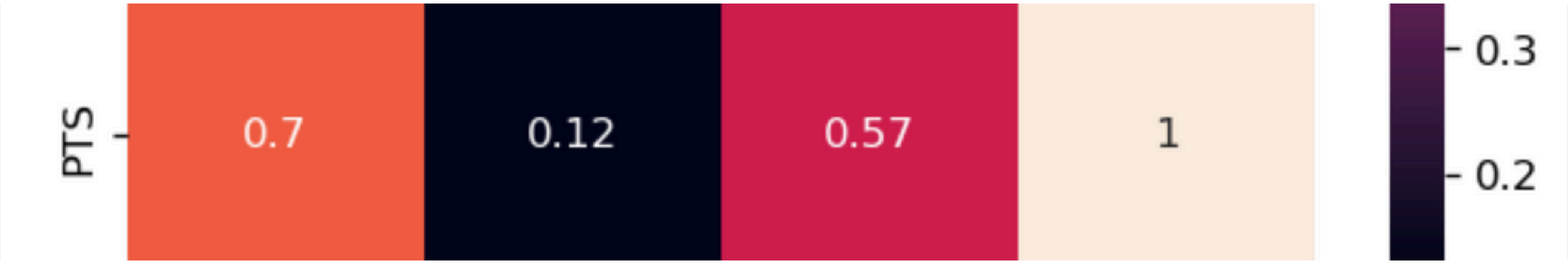
Best subset–All Players

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                     323.2
Date:                Tue, 28 May 2024   Prob (F-statistic):          1.06e-115
Time:                        21:56:13   Log-Likelihood:                -640.35
No. Observations:                 494   AIC:                             1289.
Df Residuals:                     490   BIC:                             1306.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.0488      0.250     44.212      0.000      10.558      11.540
Age            0.0798      0.010      8.287      0.000       0.061       0.099
G              0.0241      0.002     12.509      0.000       0.020       0.028
PTS            0.0986      0.007     13.822      0.000       0.085       0.113
==============================================================================
Omnibus:                        5.097   Durbin-Watson:                   2.016
Prob(Omnibus):                  0.078   Jarque-Bera (JB):                6.735
Skew:                           0.013   Prob(JB):                       0.0345
Kurtosis:                       3.571   Cond. No.                         368.
==============================================================================
```

Best subset–All Players




```
runs =  247
n1 =  247
n2 =  247
runs_exp =  248.0
stan_dev =  11.1017788169558
z =  -0.09007565512588804
pval_z =  0.928227094957536
p_value for Z-statistic=  0.928227094957536
Since the p_value = 0.9282 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

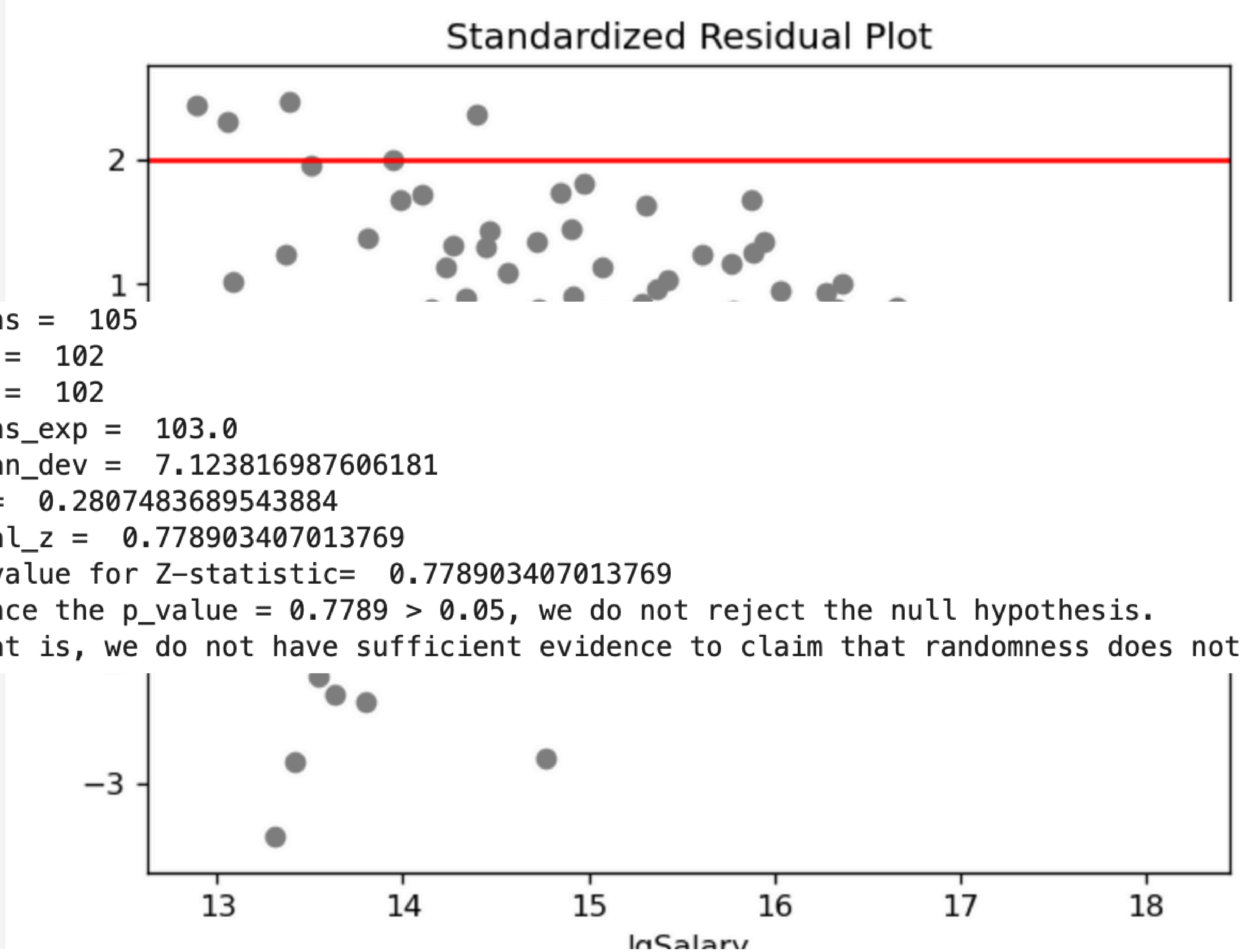



```
Shapiro statistic = 0.995608 and p_value = 0.181016
Since the p_value = 0.1810 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Best subset–Forward

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.684
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     144.1
Date:                Tue, 28 May 2024   Prob (F-statistic):           9.43e-50
Time:                        22:07:23   Log-Likelihood:                -251.53
No. Observations:                 204   AIC:                             511.1
Df Residuals:                     200   BIC:                             524.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.3517      0.352     32.281      0.000      10.658      12.045
Age            0.0714      0.014      5.108      0.000       0.044       0.099
G              0.0262      0.003      8.530      0.000       0.020       0.032
PTS            0.0853      0.011      7.491      0.000       0.063       0.108
==============================================================================
Omnibus:                        5.193   Durbin-Watson:                   2.129
Prob(Omnibus):                  0.075   Jarque-Bera (JB):                5.156
Skew:                          -0.274   Prob(JB):                       0.0759
Kurtosis:                       3.553   Cond. No.                         352.
==============================================================================
```

Best subset–Forward




```
runs =  105
n1 =  102
n2 =  102
runs_exp =  103.0
stan_dev =  7.123816987606181
z =  0.2807483689543884
pval_z =  0.778903407013769
p_value for Z-statistic=  0.778903407013769
Since the p_value = 0.7789 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

```
Shapiro statistic = 0.990032 and p_value = 0.170157
Since the p_value = 0.1702 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Best subset–Center

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.743
Model:                            OLS   Adj. R-squared:                  0.733
Method:                 Least Squares   F-statistic:                     73.40
Date:                Tue, 28 May 2024   Prob (F-statistic):           2.16e-22
Time:                        22:17:34   Log-Likelihood:                -84.718
No. Observations:                  80   AIC:                             177.4
Df Residuals:                      76   BIC:                             187.0
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.2706      0.582     19.373      0.000      10.112      12.429
Age            0.1007      0.020      5.024      0.000       0.061       0.141
MP             0.0770      0.011      6.815      0.000       0.054       0.100
BPM            0.1299      0.023      5.735      0.000       0.085       0.175
==============================================================================
Omnibus:                        8.399   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.015   Jarque-Bera (JB):                9.402
Skew:                          -0.531   Prob(JB):                      0.00909
Kurtosis:                       4.302   Cond. No.                         235.
==============================================================================
```

Best subset–Center

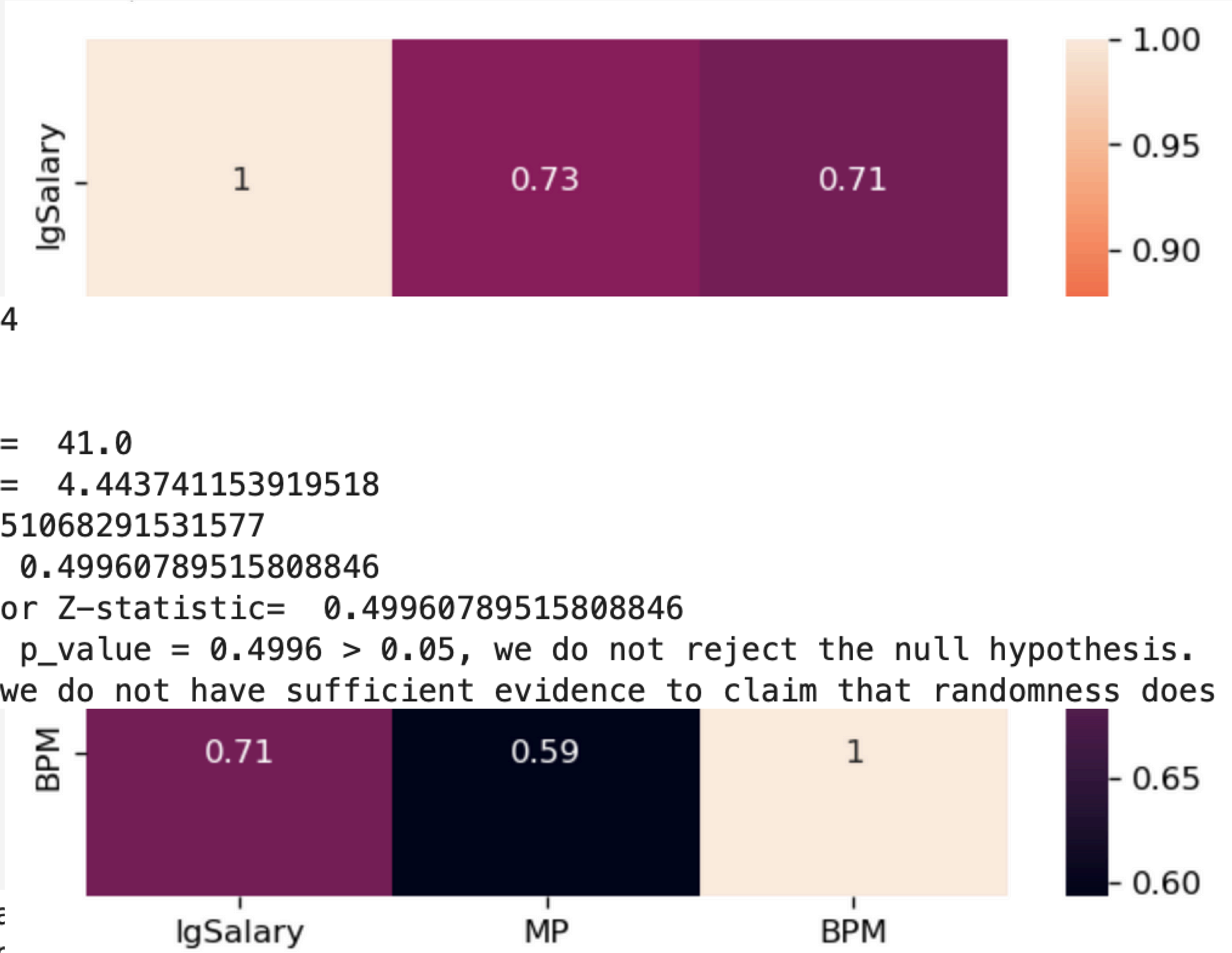

```
runs =  50
n1 =  40
n2 =  40
runs_exp =  41.0
stan_dev =  4.443741153919518
z =  2.025320487459473
pval_z =  0.04283446613171164
p_value for Z-statistic=  0.04283446613171164
Since the p_value = 0.0428 < 0.05, we reject the null hypothesis.
That is, we have sufficient evidence to claim that randomness does not exist.
```

```
Shapiro statistic = 0.968677 and p_value = 0.046751
Since the p_value = 0.0468 < 0.05, we reject the null hypothesis.
That is, we have sufficient evidence to claim that the distribution is not normal.
```

Best subset-Center

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.658
Model:                            OLS   Adj. R-squared:                  0.649
Method:                 Least Squares   F-statistic:                     74.15
Date:                Tue, 28 May 2024   Prob (F-statistic):           1.12e-18
Time:                        23:13:51   Log-Likelihood:                -96.188
No. Observations:                  80   AIC:                             198.4
Df Residuals:                      77   BIC:                             205.5
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.9678      0.257     54.357      0.000      13.456      14.479
MP             0.0748      0.013      5.776      0.000       0.049       0.101
BPM            0.1349      0.026      5.200      0.000       0.083       0.187
==============================================================================
Omnibus:                        4.479   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.107   Jarque-Bera (JB):                3.706
Skew:                          -0.489   Prob(JB):                        0.157
Kurtosis:                       3.394   Cond. No.                         55.5
==============================================================================
```

Best subset–Center



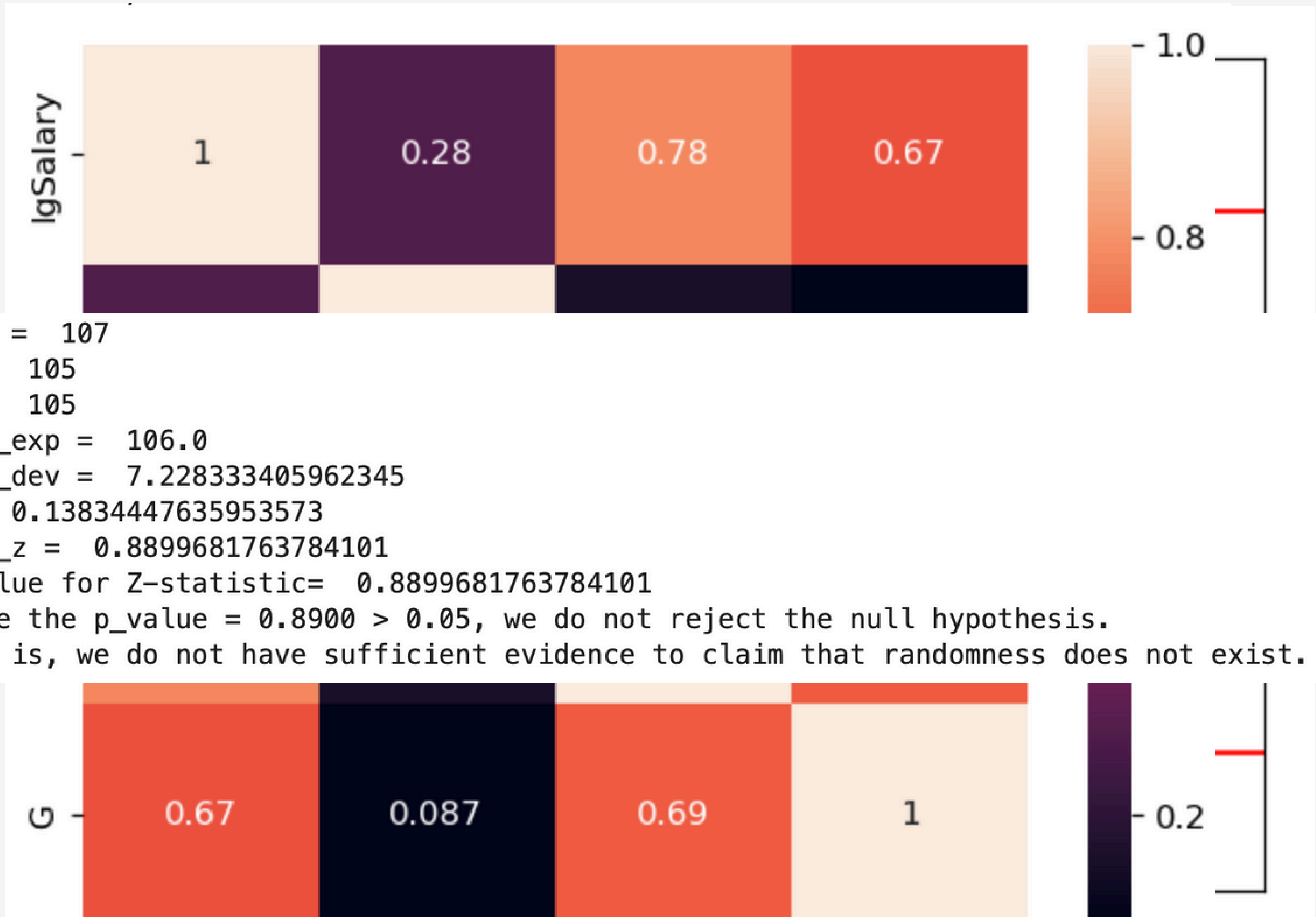

```
runs =  44
n1 =  40
n2 =  40
runs_exp =  41.0
stan_dev =  4.443741153919518
z =  0.6751068291531577
pval_z =  0.49960789515808846
p_value for Z-statistic=  0.49960789515808846
Since the p_value = 0.4996 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

```
Shapiro sta
Since the p_
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Best subset–Guard

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lgSalary   R-squared:                       0.673
Model:                            OLS   Adj. R-squared:                  0.668
Method:                 Least Squares   F-statistic:                     141.2
Date:                Tue, 28 May 2024   Prob (F-statistic):           1.02e-49
Time:                        22:22:24   Log-Likelihood:                -282.07
No. Observations:                 210   AIC:                             572.1
Df Residuals:                     206   BIC:                             585.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.7920      0.414     26.047      0.000       9.975      11.609
Age            0.0678      0.016      4.266      0.000       0.036       0.099
G              0.0152      0.003      4.614      0.000       0.009       0.022
MP             0.0914      0.009     10.492      0.000       0.074       0.109
==============================================================================
Omnibus:                        0.307   Durbin-Watson:                   2.153
Prob(Omnibus):                  0.858   Jarque-Bera (JB):                0.105
Skew:                          -0.012   Prob(JB):                        0.949
Kurtosis:                       3.107   Cond. No.                         390.
==============================================================================
```

Best subset–Guard




```
runs =  107
n1 =  105
n2 =  105
runs_exp =  106.0
stan_dev =  7.228333405962345
z =  0.13834447635953573
pval_z =  0.8899681763784101
p_value for Z-statistic=  0.8899681763784101
Since the p_value = 0.8900 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

```
Shapiro statistic = 0.996230 and p_value = 0.890476
Since the p_value = 0.8905 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Best subset

position	variable	r squared
all players	Age, G, PTS	0.664
forward	Age, G, PTS	0.684
center	MP, BPM	0.658
guard	Age, G, MP	0.673

Best subset

variable	interpretation
Age	年紀小的球員可能會為了進入nba而先接受較低的薪水，年紀很大的球員雖會因為表現不如年輕人，簽約薪水較低，但他們可能會選擇退休、不繼續打球。2~30歲的球員正值體力高峰，年紀越大者經驗越多，簽約薪水也越高。
PTS	得分多的球員價值高，簽約薪水也高

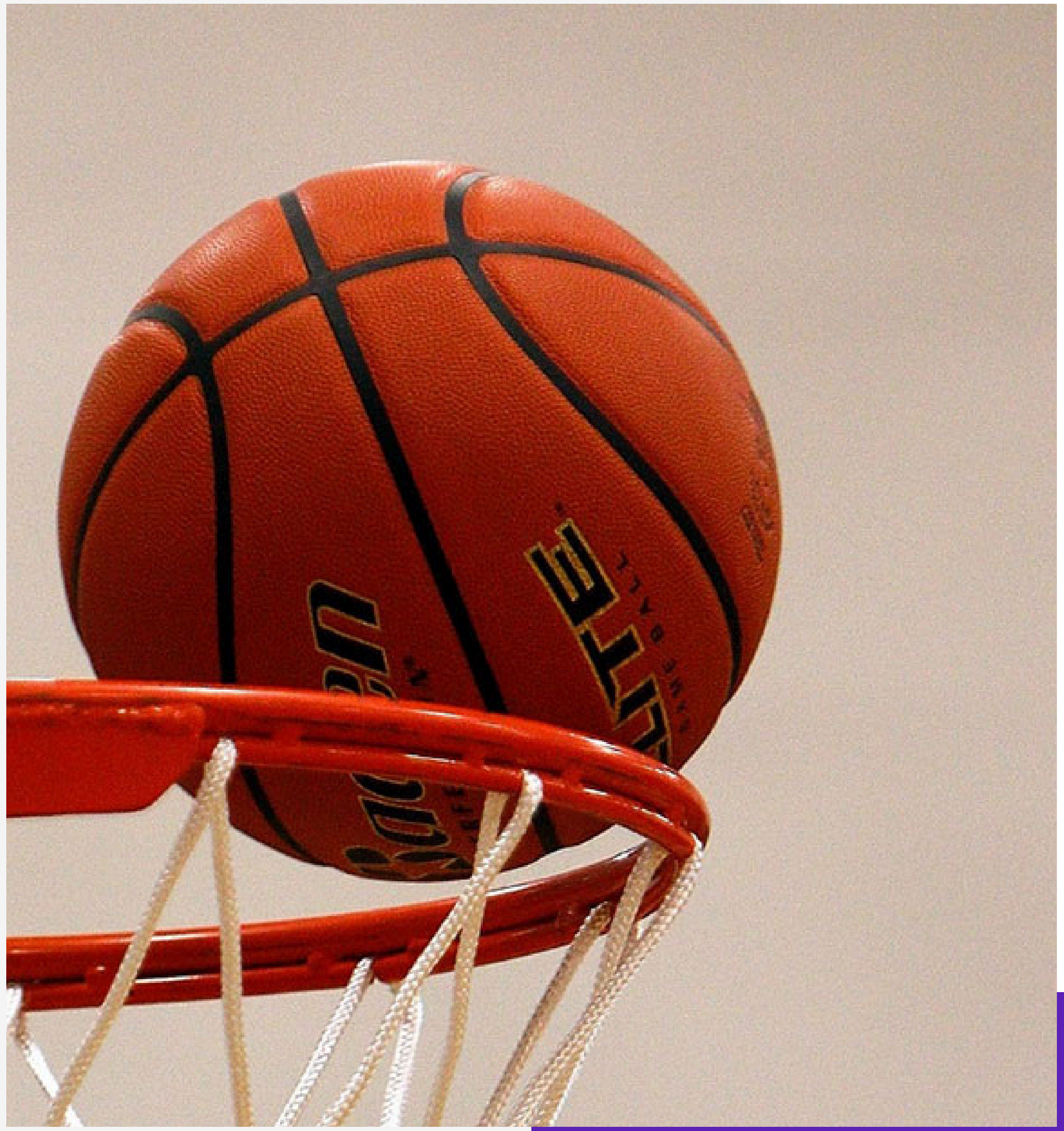
Best subset

variable	interpretation
G	出場次數較多的球員應能力較好、更被看重，薪水也因此較高
MP	在場上時間久的選手大多能力好，雇主會願意給高額薪水
BPM	BPM 計算包含許多進階數據，能看出球員的綜合表現、為球隊帶來多少淨勝分，數值越高表現越好，更有可能領到較高薪水

Forward regression v.s. Best subset

position		variable	r squared
all players	F	MP, G, Age, FTA, FT%, FG%, OWS, Level1	0.690
	B	Age, G, PTS	0.664
forward	F	G, PTS, Age, DBPM, FT%, Level1, TOV%	0.716
	B	Age, G, PTS	0.684
center	F	MP, BPM, Age, G, VORP, 2P, WS/48, WS, Level2	0.777
	B	MP, BPM	0.658
guard	F	MP, G, Age, FTA, OWS, TRB%, WS/48	0.695
	B	Age, G, MP	0.673

結論



結論

結論

- NBA 球員的年齡、上場時間、使用率、得分數對薪資的影響較顯著
- NBA 球員位置的差異會使部分數據有些不同，進而影響薪資的高低；影響不同位置球員的變數也有所差別
- NBA 球隊所在地區經濟水平發展與球員薪水和球隊薪資總額並無顯著影響
- NBA 球隊薪資總額與球隊收入、勝場數與球隊擁有者自身財力較有關連性

好想趕快放暑假

Thank you for listening!