ALY6000 Introduction to Analytics

Project Report 2

Assignment: Exploratory Data Analysis (EDA) of Two Data Sets

Submission Date: 31st January 2024, Wednesday

Submitted by

Shraddha P Gupte

Executive Summary

The objective of this report is to give analysis and observation on the data base of the 1986 Major League Baseball season game and its batting statistics. The aim is to derive the eligible players for the end of season awards and Most Valued Player of the season. The data set provided gives wide information on the number of players, games played, number of Home Runs scored by the players and other wide rage of information.

Note- As mentioned in the instruction file the report is only required to be created for the baseball data analysis hence the report is started from question 12.

Preparing the data for analysis –

Question 12

The baseball dataset is add in the R.Script to initiate the functions and statistical findings

Console | Terminal | Background Jobs

R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/

```
> #PART2
> baseball <- read.csv("/Users/shraddha/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/
Project2.R/baseball.csv")
> baseball
          Last  First Age   G  PA  AB   R   H X2B X3B HR RBI SB CS  BB  SO
1        Acker    Jim  27  21  28  28   1   3   1   0  0   0  0  0   0  21
2       Adduci    Jim  26   3  13  11   2   1   1   0  0   0  0  0   1   2
3       Aguayo   Luis  27  62 146 133  17  28   6   1  4  13  1  1   8  26
4     Aguilera   Rick  24  32  57  51   4   8   0   0  2   6  0  0   3  12
5    Akerfelds Darrel  24   1   0   0   0   0   0   0  0   0  0  0   0   0
6      Aldrete   Mike  25  84 256 216  27  54  18   3  2  25  1  3  33  34
7    Alexander  Doyle  35  18  45  38   2   8   1   0  0   5  0  0   0   8
8     Allanson   Andy  24 101 324 293  30  66   7   3  1  29 10  1  14  36
9        Almon   Bill  33 102 230 196  29  43   7   2  7  27 11  4  30  38
10     Amelung     Ed  27   8  11  11   0   1   0   0  0   0  0  0   0   4
11    Andersen  Larry  33  48   7   6   0   0   0   0  0   0  0  0   0   3
12    Anderson   Dave  25  92 241 216  31  53   9   0  1  15  5  1  22  39
13    Anderson   Rick  29  15  12  11   1   1   0   0  0   0  0  0   0   4
14    Anderson  Allan  22   1   0   0   0   0   0   0  0   0  0  0   0   0
15       Armas   Tony  32 121 453 425  40 112  21   4 11  58  0  3  24  77
16     Asadoor  Randy  23  15  60  55   9  20   5   0  0   7  1  2   3  13
17       Ashby   Alan  34 120 361 315  24  81  15   0  7  38  1  0  39  56
```

Question 13

The data is then loaded correctly, cleaned up and visualized in another configuration by using the ‘head’, ‘names’ and ‘glimpse’ functions.

Gupte_Project.R | baseball | data_2015

Filter

	Last	First	Age	G	PA	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO
1	Acker	Jim	27	21	28	28	1	3	1	0	0	0	0	0	0	21
2	Adduci	Jim	26	3	13	11	2	1	1	0	0	0	0	0	1	2
3	Aguayo	Luis	27	62	146	133	17	28	6	1	4	13	1	1	8	26
4	Aguilera	Rick	24	32	57	51	4	8	0	0	2	6	0	0	3	12
5	Akerfelds	Darrel	24	1	0	0	0	0	0	0	0	0	0	0	0	0
6	Aldrete	Mike	25	84	256	216	27	54	18	3	2	25	1	3	33	34
7	Alexander	Doyle	35	18	45	38	2	8	1	0	0	5	0	0	0	8
8	Allanson	Andy	24	101	324	293	30	66	7	3	1	29	10	1	14	36
9	Almon	Bill	33	102	230	196	29	43	7	2	7	27	11	4	30	38
10	Amelung	Ed	27	8	11	11	0	1	0	0	0	0	0	0	0	4

Showing 1 to 11 of 771 entries, 16 total columns

Console | Terminal | Background Jobs

R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/

```
[14] "CS"    "BB"    "SO"
> view(baseball)
> glimpse(baseball)
Rows: 771
Columns: 16
$ Last  <chr> "Acker", "Adduci", "Aguayo", "Aguilera", "Akerfelds", "Aldrete", "Alexander", "Allanson", "Almon"…
$ First <chr> "Jim", "Jim", "Luis", "Rick", "Darrel", "Mike", "Doyle", "Andy", "Bill", "Ed", "Larry", "Dave", "…
$ Age   <int> 27, 26, 27, 24, 24, 25, 35, 24, 33, 27, 33, 25, 29, 22, 32, 23, 34, 25, 27, 26, 24, 24, 27, 36, 3…
$ G     <int> 21, 3, 62, 32, 1, 84, 18, 101, 102, 8, 48, 92, 15, 1, 121, 15, 120, 61, 3, 124, 1, 57, 145, 1, 83…
$ PA    <int> 28, 13, 146, 57, 0, 256, 45, 324, 230, 11, 7, 241, 12, 0, 453, 60, 361, 8, 0, 440, 0, 182, 618, 0…
$ AB    <int> 28, 11, 133, 51, 0, 216, 38, 293, 196, 11, 6, 216, 11, 0, 425, 55, 315, 6, 0, 387, 0, 153, 570, 0…
$ R     <int> 1, 2, 17, 4, 0, 27, 2, 30, 29, 0, 0, 31, 1, 0, 40, 9, 24, 0, 0, 67, 0, 9, 72, 0, 25, 1, 54, 0, 28…
$ H     <int> 3, 1, 28, 8, 0, 54, 8, 66, 43, 1, 0, 53, 1, 0, 112, 20, 81, 0, 0, 124, 0, 27, 169, 0, 58, 3, 117,…
$ X2B   <int> 1, 1, 6, 0, 0, 18, 1, 7, 7, 0, 0, 9, 0, 0, 21, 5, 15, 0, 0, 18, 0, 5, 29, 0, 8, 1, 25, 0, 9, 35, …
$ X3B   <int> 0, 0, 1, 0, 0, 3, 0, 3, 2, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 1, 0, 0, 2, 0, 4, 5…
$ HR    <int> 0, 0, 4, 2, 0, 2, 0, 1, 7, 0, 0, 1, 0, 0, 11, 0, 7, 0, 0, 1, 0, 4, 21, 0, 4, 0, 29, 0, 2, 40, 0, …
$ RBI   <int> 0, 0, 13, 6, 0, 25, 5, 29, 27, 0, 0, 15, 0, 0, 58, 7, 38, 0, 0, 27, 0, 15, 88, 0, 19, 0, 88, 0, 2…
$ SB    <int> 0, 0, 1, 0, 0, 1, 0, 10, 11, 0, 0, 5, 0, 0, 0, 0, 1, 1, 0, 0, 13, 0, 1, 2, 0, 0, 0, 0, 0, 8, 0, 1…
$ CS    <int> 0, 0, 1, 0, 0, 3, 0, 1, 4, 0, 0, 1, 0, 0, 3, 2, 0, 0, 0, 7, 0, 1, 1, 0, 1, 0, 0, 0, 1, 8, 0, 7, 1…
$ BB    <int> 0, 1, 8, 3, 0, 33, 0, 14, 30, 0, 0, 22, 0, 0, 24, 3, 39, 2, 0, 36, 0, 28, 38, 0, 27, 2, 43, 0, 22…
$ SO    <int> 21, 2, 26, 12, 0, 34, 8, 36, 38, 4, 3, 39, 4, 0, 77, 13, 56, 3, 0, 32, 0, 45, 89, 0, 37, 7, 146, …
>
```

Filtering the data

Question 14

The data is then filtered to remove any unnecessary data points to give the consized report to work the analysis on. This is done by using the filter function to remove all players that have 0 at bats (AB).

Console | Terminal | Background Jobs

R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/

```
> baseball <- baseball %>% filter(AB != 0)
> baseball
```

	Last	First	Age	G	PA	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO
1	Acker	Jim	27	21	28	28	1	3	1	0	0	0	0	0	0	21
2	Adduci	Jim	26	3	13	11	2	1	1	0	0	0	0	0	1	2
3	Aguayo	Luis	27	62	146	133	17	28	6	1	4	13	1	1	8	26
4	Aguilera	Rick	24	32	57	51	4	8	0	0	2	6	0	0	3	12
5	Aldrete	Mike	25	84	256	216	27	54	18	3	2	25	1	3	33	34
6	Alexander	Doyle	35	18	45	38	2	8	1	0	0	5	0	0	0	8
7	Allanson	Andy	24	101	324	293	30	66	7	3	1	29	10	1	14	36
8	Almon	Bill	33	102	230	196	29	43	7	2	7	27	11	4	30	38
9	Amelung	Ed	27	8	11	11	0	1	0	0	0	0	0	0	0	4
10	Andersen	Larry	33	48	7	6	0	0	0	0	0	0	0	0	0	3
11	Anderson	Dave	25	92	241	216	31	53	9	0	1	15	5	1	22	39
12	Anderson	Rick	29	15	12	11	1	1	0	0	0	0	0	0	0	4
13	Armas	Tony	32	121	453	425	40	112	21	4	11	58	0	3	24	77
14	Asadoor	Randy	23	15	60	55	9	20	5	0	0	7	1	2	3	13
15	Ashby	Alan	34	120	361	315	24	81	15	0	7	38	1	0	39	56
16	Assenmacher	Paul	25	61	8	6	0	0	0	0	0	0	0	0	2	3
17	Backman	Wally	26	124	440	387	67	124	18	2	1	27	13	7	36	32
18	Bailey	Mark	24	57	182	153	9	27	5	0	4	15	1	1	28	45
19	Baines	Harold	27	145	618	570	72	169	29	2	21	88	2	1	38	89
20	Baker	Dusty	37	83	271	242	25	58	8	0	4	19	0	1	27	37
21	Baker	Doug	25	13	30	24	1	3	1	0	0	0	0	0	2	7
22	Balboni	Steve	29	138	562	512	54	117	25	1	29	88	0	0	43	146
23	Baller	Jay	25	36	6	5	0	0	0	0	0	0	0	0	0	1
24	Bando	Chris	30	92	290	254	28	68	9	0	2	26	0	1	22	49
25	Barfield	Jesse	26	158	671	589	107	170	35	2	40	108	8	8	69	146
26	Bargar	Greg	27	22	2	2	0	0	0	0	0	0	0	0	0	2
27	Barrett	Marty	28	158	713	625	94	179	39	4	4	60	15	7	65	31
28	Bass	Kevin	27	157	640	591	83	184	33	5	20	79	22	13	38	72
29	Bathe	Bill	25	39	112	103	9	19	3	0	5	11	0	0	2	20
30	Baylor	Don	37	160	687	585	93	139	23	1	31	94	3	5	62	111
31	Beane	Billy	24	80	194	183	20	39	6	0	3	15	2	3	11	54
32	Bedrosian	Steve	28	68	6	5	0	1	0	0	0	0	0	0	1	1
33	Bell	Buddy	34	155	655	568	89	158	29	3	20	75	2	8	73	49

Initiating the analysis

Question 15

Most optimized analysis would be derived by calculating the Batting Average by division of the number of Hits (H) by number the number of at Bats (AB).

Console | Terminal | Background Jobs

R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/

```
> baseball$BA <- baseball$H / baseball$AB
> baseball
          Last  First Age   G  PA  AB   R   H X2B X3B HR RBI SB CS  BB  SO         BA
1        Acker    Jim  27  21  28  28   1   3   1   0  0   0  0  0   0  21 0.10714286
2       Adduci    Jim  26   3  13  11   2   1   1   0  0   0  0  0   1   2 0.09090909
3       Aguayo   Luis  27  62 146 133  17  28   6   1  4  13  1  1   8  26 0.21052632
4     Aguilera   Rick  24  32  57  51   4   8   0   0  2   6  0  0   3  12 0.15686275
5      Aldrete   Mike  25  84 256 216  27  54  18   3  2  25  1  3  33  34 0.25000000
6    Alexander  Doyle  35  18  45  38   2   8   1   0  0   5  0  0   0   8 0.21052632
7     Allanson   Andy  24 101 324 293  30  66   7   3  1  29 10  1  14  36 0.22525597
8        Almon   Bill  33 102 230 196  29  43   7   2  7  27 11  4  30  38 0.21938776
9      Amelung     Ed  27   8  11  11   0   1   0   0  0   0  0  0   0   4 0.09090909
10    Andersen  Larry  33  48   7   6   0   0   0   0  0   0  0  0   0   3 0.00000000
11    Anderson   Dave  25  92 241 216  31  53   9   0  1  15  5  1  22  39 0.24537037
12    Anderson   Rick  29  15  12  11   1   1   0   0  0   0  0  0   0   4 0.09090909
13       Armas   Tony  32 121 453 425  40 112  21   4 11  58  0  3  24  77 0.26352941
14     Asadoor  Randy  23  15  60  55   9  20   5   0  0   7  1  2   3  13 0.36363636
15       Ashby   Alan  34 120 361 315  24  81  15   0  7  38  1  0  39  56 0.25714286
16 Assenmacher   Paul  25  61   8   6   0   0   0   0  0   0  0  0   2   3 0.00000000
17     Backman  Wally  26 124 440 387  67 124  18   2  1  27 13  7  36  32 0.32041344
18      Bailey   Mark  24  57 182 153   9  27   5   0  4  15  1  1  28  45 0.17647059
19      Baines Harold  27 145 618 570  72 169  29   2 21  88  2  1  38  89 0.29649123
20       Baker  Dusty  37  83 271 242  25  58   8   0  4  19  0  1  27  37 0.23966942
21       Baker   Doug  25  13  30  24   1   3   1   0  0   0  0  0   2   7 0.12500000
22     Balboni  Steve  29 138 562 512  54 117  25   1 29  88  0  0  43 146 0.22851562
23      Baller    Jay  25  36   6   5   0   0   0   0  0   0  0  0   0   1 0.00000000
```

Question 16

The on base percentage (OBP)would be comparatively beneficial for this analysis as it includes computation of base on balls with the Batting Average. This would derive the percentage analysis of the data.

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> baseball$OBP <- (baseball$H + baseball$BB) / (baseball$AB + baseball$BB)
> baseball
          Last  First Age   G  PA  AB   R   H X2B X3B HR RBI SB CS  BB  SO         BA        OBP
1        Acker    Jim  27  21  28  28   1   3   1   0  0   0  0  0   0  21 0.10714286 0.10714286
2       Adduci    Jim  26   3  13  11   2   1   1   0  0   0  0  0   1   2 0.09090909 0.16666667
3       Aguayo   Luis  27  62 146 133  17  28   6   1  4  13  1  1   8  26 0.21052632 0.25531915
4     Aguilera   Rick  24  32  57  51   4   8   0   0  2   6  0  0   3  12 0.15686275 0.20370370
5      Aldrete   Mike  25  84 256 216  27  54  18   3  2  25  1  3  33  34 0.25000000 0.34939759
6    Alexander  Doyle  35  18  45  38   2   8   1   0  0   5  0  0   0   8 0.21052632 0.21052632
7     Allanson   Andy  24 101 324 293  30  66   7   3  1  29 10  1  14  36 0.22525597 0.26058632
8        Almon   Bill  33 102 230 196  29  43   7   2  7  27 11  4  30  38 0.21938776 0.32300885
9      Amelung     Ed  27   8  11  11   0   1   0   0  0   0  0  0   0   4 0.09090909 0.09090909
10    Andersen  Larry  33  48   7   6   0   0   0   0  0   0  0  0   0   3 0.00000000 0.00000000
11    Anderson   Dave  25  92 241 216  31  53   9   0  1  15  5  1  22  39 0.24537037 0.31512605
12    Anderson   Rick  29  15  12  11   1   1   0   0  0   0  0  0   0   4 0.09090909 0.09090909
13       Armas   Tony  32 121 453 425  40 112  21   4 11  58  0  3  24  77 0.26352941 0.30289532
14     Asadoor  Randy  23  15  60  55   9  20   5   0  0   7  1  2   3  13 0.36363636 0.39655172
15       Ashby   Alan  34 120 361 315  24  81  15   0  7  38  1  0  39  56 0.25714286 0.33898305
16 Assenmacher   Paul  25  61   8   6   0   0   0   0  0   0  0  0   2   3 0.00000000 0.25000000
17     Backman  Wally  26 124 440 387  67 124  18   2  1  27 13  7  36  32 0.32041344 0.37825059
18      Bailey   Mark  24  57 182 153   9  27   5   0  4  15  1  1  28  45 0.17647059 0.30386740
19      Baines Harold  27 145 618 570  72 169  29   2 21  88  2  1  38  89 0.29649123 0.34046053
20       Baker  Dusty  37  83 271 242  25  58   8   0  4  19  0  1  27  37 0.23966942 0.31598513
21       Baker   Doug  25  13  30  24   1   3   1   0  0   0  0  0   2   7 0.12500000 0.19230769
22     Balboni  Steve  29 138 562 512  54 117  25   1 29  88  0  0  43 146 0.22851562 0.28828829
23      Baller    Jay  25  36   6   5   0   0   0   0  0   0  0  0   0   1 0.00000000 0.00000000
```

Question 17

In order to find and analyze the data further, the dataset should be filtered and arranged from the highest value to lowest value by the strikeout rate. The resulting data would give the list of top ten players that provide the most value in that season. Here the top 10 players are calculated by using the Head and Arrange function.

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> strikeout_artist <- head(arrange(baseball, desc(SO)), 10)
> strikeout_artist
         Last  First Age   G  PA  AB   R   H X2B X3B HR RBI SB CS BB  SO        BA       OBP
1  Incaviglia   Pete  22 153 606 540  82 135  21   2 30  88  3  2 55 185 0.2500000 0.3193277
2        Deer    Rob  25 134 546 466  75 108  17   3 33  86  5  2 72 179 0.2317597 0.3345725
3     Canseco   Jose  21 157 682 600  85 144  29   1 33 117 15  7 65 175 0.2400000 0.3142857
4     Presley    Jim  24 155 660 616  83 163  33   4 27 107  0  4 32 172 0.2646104 0.3009259
5   Tartabull  Danny  23 137 578 511  76 138  25   6 25  96  4  8 61 157 0.2700587 0.3479021
6     Balboni  Steve  29 138 562 512  54 117  25   1 29  88  0  0 43 146 0.2285156 0.2882883
7    Barfield  Jesse  26 158 671 589 107 170  35   2 40 108  8  8 69 146 0.2886248 0.3632219
8      Samuel   Juan  25 145 633 591  90 157  36  12 16  78 42 14 26 142 0.2656514 0.2965964
9      Murphy   Dale  30 160 692 614  89 163  29   7 29  83  7  7 75 141 0.2654723 0.3454282
10 Strawberry Darryl  24 136 562 475  76 123  27   5 27  93 28 12 72 141 0.2589474 0.3564899
>
```

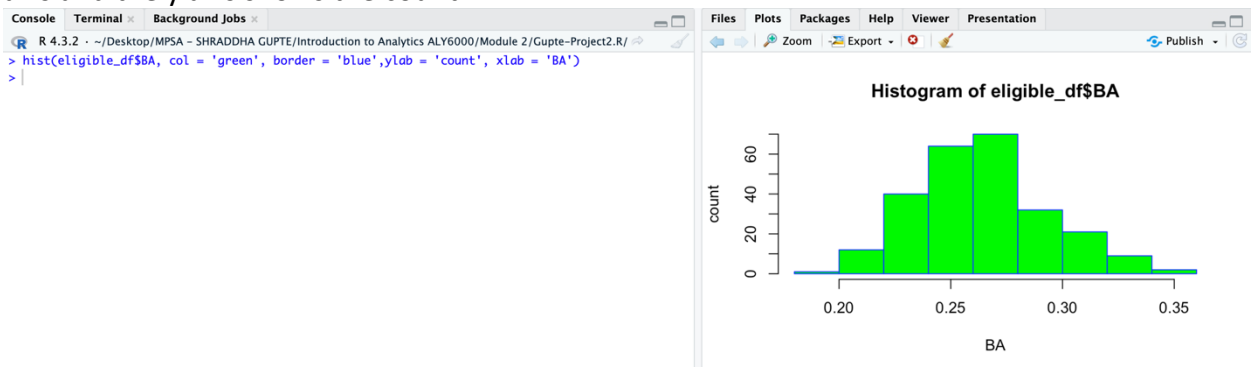
Question 18

The defined eligibility criteria for the end of season awards is the player having at least 300 at bats and more than 100 games. This result can be obtained by performing the 'filter' function on at bats and number of games.

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA - SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> eligible_df <- filter(baseball, AB >=300 | G>=100)
> eligible_df
         Last   First Age   G  PA  AB   R   H X2B X3B HR RBI  SB CS  BB  SO        BA       OBP
1    Allanson    Andy  24 101 324 293  30  66   7   3  1  29  10  1  14  36 0.2252560 0.2605863
2       Almon    Bill  33 102 230 196  29  43   7   2  7  27  11  4  30  38 0.2193878 0.3230088
3       Armas    Tony  32 121 453 425  40 112  21   4 11  58   0  3  24  77 0.2635294 0.3028953
4       Ashby    Alan  34 120 361 315  24  81  15   0  7  38   1  0  39  56 0.2571429 0.3389831
5     Backman   Wally  26 124 440 387  67 124  18   2  1  27  13  7  36  32 0.3204134 0.3782506
6      Baines  Harold  27 145 618 570  72 169  29   2 21  88   2  1  38  89 0.2964912 0.3404605
7     Balboni   Steve  29 138 562 512  54 117  25   1 29  88   0  0  43 146 0.2285156 0.2882883
8    Barfield   Jesse  26 158 671 589 107 170  35   2 40 108   8  8  69 146 0.2886248 0.3632219
9     Barrett   Marty  28 158 713 625  94 179  39   4  4  60  15  7  65  31 0.2864000 0.3536232
10       Bass   Kevin  27 157 640 591  83 184  33   5 20  79  22 13  38  72 0.3113367 0.3529412
11     Baylor     Don  37 160 687 585  93 139  23   1 31  94   3  5  62 111 0.2376068 0.3106646
12       Bell   Buddy  34 155 655 568  89 158  29   3 20  75   2  8  73  49 0.2781690 0.3603744
13       Bell  George  26 159 690 641 101 198  38   6 31 108   7  8  41  62 0.3088924 0.3504399
14   Belliard  Rafael  24 117 350 309  33  72   5   2  0  31  12  2  26  54 0.2330097 0.2925373
15   Beniquez    Juan  36 113 395 343  48 103  15   0  6  36   2  3  40  49 0.3002915 0.3733681
16  Bernazard    Tony  29 146 636 562  88 169  28   4 17  73  17  8  53  77 0.3007117 0.3609756
```

Question 19

The below graphical representation shows the visualization for the Batting Average of the eligible players for the end of season awards. According to the below data the BA is on the x axis and the y axis shows the count.



Question 20

To identify the Most Valuable Player (MVP) there are three key indicators, On-Base Percentage (OBP), Number of Home Runs Scored(HR) and the number of runs batted in (RBI).

- Step 1 – Creating a data frame with only important columns including the first and last name and the three indicators OPB, HR and RBI and storing it in the variable MVP_df.

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> # selecting only important stats columns from baseball and storing in MVP_df.
> MVP_df <- baseball[c("Last", "First", "OBP", "HR", "RBI")]
> MVP_df
           Last   First        OBP HR RBI
1         Acker     Jim 0.10714286  0   0
2        Adduci     Jim 0.16666667  0   0
3        Aguayo    Luis 0.25531915  4  13
4      Aguilera    Rick 0.20370370  2   6
5       Aldrete    Mike 0.34939759  2  25
6     Alexander   Doyle 0.21052632  0   5
7      Allanson    Andy 0.26058632  1  29
8         Almon    Bill 0.32300885  7  27
9       Amelung      Ed 0.09090909  0   0
10     Andersen   Larry 0.00000000  0   0
11     Anderson    Dave 0.31512605  1  15
12     Anderson    Rick 0.09090909  0   0
13        Armas    Tony 0.30289532 11  58
14      Asadoor   Randy 0.39655172  0   7
```

- Step 2 – Getting a mean score for each important statistics of OPB, HR and RBI; rounding the operations for easier application and storing it in variables 'MeanOBP', 'MeanHR' and 'MeanRBI' respectively

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> # Getting Mean score for each imp stat.
> meanOBP <- round(mean(MVP_df$OBP), 2)
> meanHR <- round(mean(MVP_df$HR),0)
> meanRBI <- round(mean(MVP_df$RBI),0)
> meanOBP
[1] 0.27
> meanHR
[1] 5
> meanRBI
[1] 24
>
```

- Step 3 - Filtering players above mean score of each indicator to get list of above avg players and storing the data back in MVP_df

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> # Filtering players above mean score to get list of above avg players.
> MVP_df <- filter(MVP_df, OBP>0.25 & HR>9 & RBI>39)
> MVP_df
          Last    First       OBP HR RBI
1        Armas     Tony 0.3028953 11  58
2       Baines   Harold 0.3404605 21  88
3      Balboni    Steve 0.2882883 29  88
4     Barfield    Jesse 0.3632219 40 108
5         Bass    Kevin 0.3529412 20  79
6       Baylor      Don 0.3106646 31  94
7         Bell    Buddy 0.3603744 20  75
8         Bell   George 0.3504399 31 108
9    Bernazard     Tony 0.3609756 17  73
10       Bonds    Barry 0.3284519 16  48
11     Bradley     Phil 0.3980100 12  50
12       Bream      Sid 0.3436426 16  77
13      Brenly      Bob 0.3479853 16  62
14       Brett   George 0.3992322 16  73
```

- Step 4 – As the scales of the three indicators are not consistent it is necessary to bring them under one common score. Hence, three columns are added ('OPB','HR','RBI') having each stat computed into a percentage score by using the 'max' function and rounded off to 0 decimal by using 'round' function.

```
Console   Terminal ×   Background Jobs ×
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> # Computing % score for each stat individually since scales are not consistent.
> MVP_df$OBP <- round((((max(MVP_df$OBP) - MVP_df$OBP) / max(MVP_df$OBP) )*100),0)
> MVP_df$HR <- round((((max(MVP_df$HR) - MVP_df$HR) / max(MVP_df$HR) )*100),0)
> MVP_df$RBI <- round((((max(MVP_df$RBI) - MVP_df$RBI) / max(MVP_df$RBI) )*100),0)
> MVP_df
          Last    First OBP HR RBI
1        Armas     Tony  26 72  52
2       Baines   Harold  17 48  27
3      Balboni    Steve  30 28  27
4     Barfield    Jesse  12  0  11
5         Bass    Kevin  14 50  35
6       Baylor      Don  25 22  22
7         Bell    Buddy  12 50  38
8         Bell   George  15 22  11
9    Bernazard     Tony  12 57  40
10       Bonds    Barry  20 60  60
11     Bradley     Phil   3 70  59
12       Bream      Sid  17 60  36
```

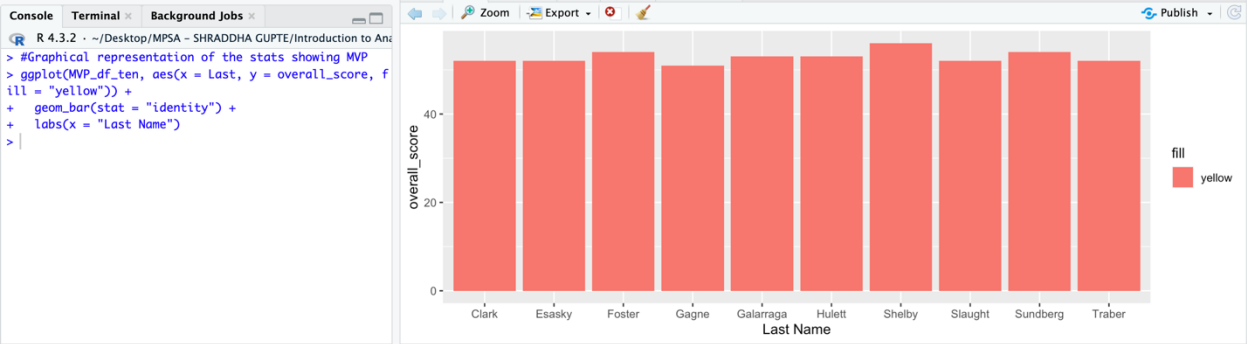
- Step 5 – As all the scores are now in %, these scores can be computed to get an overall score of all indicators which is the average of % of each stat for every player and adding a column on 'overall_score'

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module 2/Gupte-Project2.R/
> # Calculating overall_score which is the avg of % of each stat for every player.
> MVP_df$overall_score <- round((MVP_df$OBP + MVP_df$HR + MVP_df$RBI) / 3, 0)
> MVP_df
          Last    First OBP HR RBI overall_score
1        Armas     Tony  26 72  52            50
2       Baines   Harold  17 48  27            31
3      Balboni    Steve  30 28  27            28
4     Barfield    Jesse  12  0  11             8
5         Bass    Kevin  14 50  35            33
6       Baylor      Don  25 22  22            23
7         Bell    Buddy  12 50  38            33
8         Bell   George  15 22  11            16
9    Bernazard     Tony  12 57  40            36
10       Bonds    Barry  20 60  60            47
11     Bradley     Phil   3 70  59            44
12       Bream      Sid  17 60  36            38
13      Brenly      Bob  15 60  49            41
14       Brett   George   3 60  40            34
```

- Step 6 – These scores are then arranged from highest to lowest by using the arrange and head function to get the top 10 players eligible for the Most Valuable Player award (MVP) and storing it in 'MVP_df_ten'

```
Console  Terminal  Background Jobs
R 4.3.2 · ~/Desktop/MPSA – SHRADDHA GUPTE/Introduction to Analytics ALY6000/Module
> # Selecting top 10 high score value players.
> MVP_df_ten <- head(arrange(MVP_df, desc(overall_score)), 10)
> MVP_df_ten
         Last  First OBP HR RBI overall_score
1      Shelby   John  37 72  60            56
2      Foster George  31 65  65            54
3    Sundberg    Jim  26 70  65            54
4  Galarraga Andres  19 75  65            53
5      Hulett    Tim  37 57  64            53
6       Clark   Will  17 72  66            52
7      Esasky   Nick  21 70  66            52
8     Slaught    Don  27 68  62            52
9      Traber    Jim  24 68  64            52
10      Gagne   Greg  28 70  55            51
>
```

- Step 7 – Presenting the graphical representation of the stats to identify the MVP.



Key Findings

After analyzing the data, several players stand out based on their performance across multiple metrics:

- **On-Base Percentage (OBP)** is a crucial metric indicating a player's ability to contribute to scoring runs for their team.
- **Home Runs (HR)** is an important offensive statistic that demonstrates a player's ability to hit for power and drive in runs.
- **Runs Batted-In (RBI)** indicates a player's ability to produce runs by successfully driving in base runners.

However, these high performance in these individual statistics would not demonstrate the most valuable player as the player having high performance in one indicator may have below average performance in other indicator. Hence it is necessary to identify the player that has a consistent performance amongst all three indicators.

Overall score is a combination of OBP, HR, and RBI, that identifies players who consistently perform well across various aspects of the game.

Conclusion and Recommendations

From the above operations and graphical representations, it is conclusively derived that the player named John Shelby is the deserving player to be awarded as the Most Valuable Player award for the season.

Shelby John has demonstrated exceptional and consistent performance in key areas such as OBP, HR, and RBI, contributing significantly to their team's achievements. The recommendation is supported by robust statistical analysis and provides valuable insights into the player's impact on the game.

Citation:

1. ***Book-***
   - *Kabacoff, R.I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd edition). Manning Publications.*
   - *Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.*
2. ***Proffessor Notes*** *–*
   - *Thomas Goulding (2024). Introduction to Analytics Notes [ALY6000 Course Notes]Canvas :https://northeastern.instructure.com/courses/164773/pages/module-2-%7C-resources?module_item_id=9797122*

3. ***Website –***
   - *R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.R-project.org/*