Analysis on 2022 Major League Baseball Data

Indo Badial, RJ Burson, and

Melissa Gomez

California State University, Long Beach

STAT 550 - Multivariate Statistical Analysis

15 December 2022

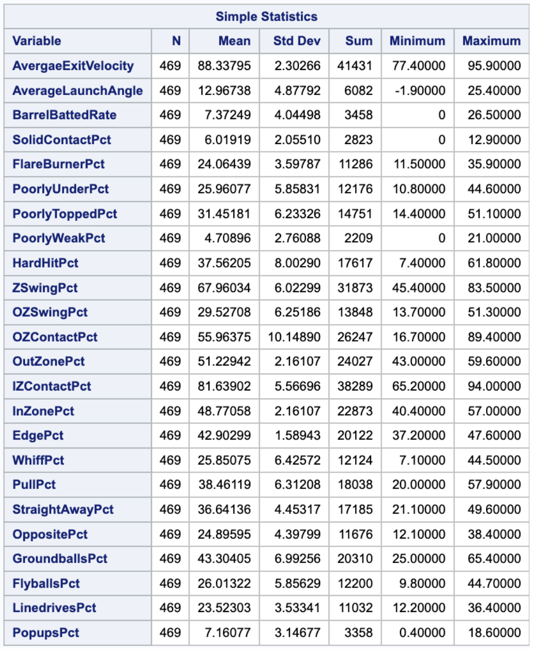
Introduction

The purpose of our project is to analyze 2022 Major League Baseball hitting statistics in order to predict a players on base plus slugging percentage over the course of a 162 game season. In order to do so, we use techniques such as principal component analysis to reduce the dimension of our dataset, inferences on the mean vector, factor analysis to explain any covariance/correlation structures amongst variables, and cluster analysis to possibly reduce the number of observations by grouping similar individuals together. Our data was collected from Baseball Savant, a site dedicated to providing Statcast metrics, and advanced statistics. Our dataset includes metrics such as On-base Plus Slugging (OPS), Exit Velocity (EV), and Launch Angle (LA). The dataset consists of 469 observations (players) and 25 observable variables.

Methods and Analysis

The first step in the process was to perform exploratory data analysis to learn more about the nature of the dataset. To accomplish this, a correlation matrix between all 25 variables was produced and analyzed to see if any variables were highly correlated with each other. This would mean that several variables essentially give us the same information about OPS, meaning data reduction techniques should be used for simplicity.

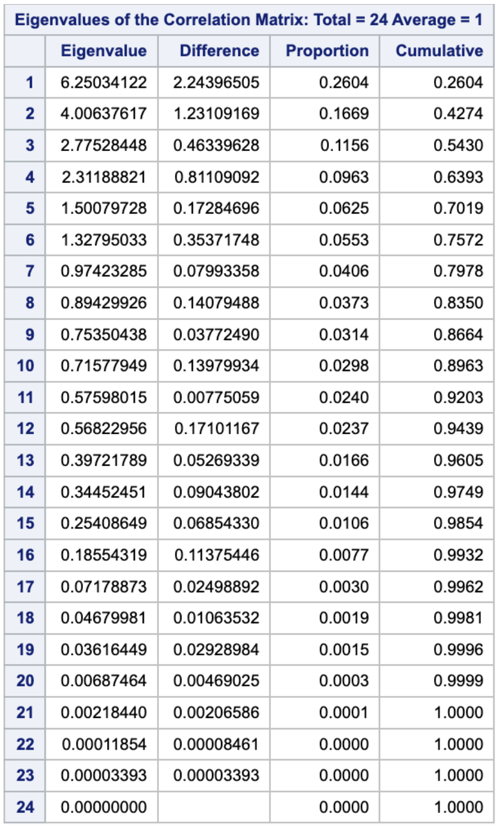
The simple statistics of the original full model are given in the table to the right. The correlation matrix is very large (25x25 size matrix) so only the code for it is included in the appendix. However, there were several variables with high correlation (>.80) which justify the use of methods such as Principal Component Analysis and Factor Analysis used later.



Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
AvergaeExitVelocity	469	88.33795	2.30266	41431	77.40000	95.90000
AverageLaunchAngle	469	12.96738	4.87792	6082	-1.90000	25.40000
BarrelBattedRate	469	7.37249	4.04498	3458	0	26.50000
SolidContactPct	469	6.01919	2.05510	2823	0	12.90000
FlareBurnerPct	469	24.06439	3.59787	11286	11.50000	35.90000
PoorlyUnderPct	469	25.96077	5.85831	12176	10.80000	44.60000
PoorlyToppedPct	469	31.45181	6.23326	14751	14.40000	51.10000
PoorlyWeakPct	469	4.70896	2.76088	2209	0	21.00000
HardHitPct	469	37.56205	8.00290	17617	7.40000	61.80000
ZSwingPct	469	67.96034	6.02299	31873	45.40000	83.50000
OZSwingPct	469	29.52708	6.25186	13848	13.70000	51.30000
OZContactPct	469	55.96375	10.14890	26247	16.70000	89.40000
OutZonePct	469	51.22942	2.16107	24027	43.00000	59.60000
IZContactPct	469	81.63902	5.56696	38289	65.20000	94.00000
InZonePct	469	48.77058	2.16107	22873	40.40000	57.00000
EdgePct	469	42.90299	1.58943	20122	37.20000	47.60000
WhiffPct	469	25.85075	6.42572	12124	7.10000	44.50000
PullPct	469	38.46119	6.31208	18038	20.00000	57.90000
StraightAwayPct	469	36.64136	4.45317	17185	21.10000	49.60000
OppositePct	469	24.89595	4.39799	11676	12.10000	38.40000
GroundballsPct	469	43.30405	6.99256	20310	25.00000	65.40000
FlyballsPct	469	26.01322	5.85629	12200	9.80000	44.70000
LinedrivesPct	469	23.52303	3.53341	11032	12.20000	36.40000
PopupsPct	469	7.16077	3.14677	3358	0.40000	18.60000

Principal component analysis allows us to reduce our large datasets with many variables into a smaller dataset with few variables. Thus, for the purpose of data reduction, we use principal analysis on our dataset. Since On-base Plus Slugging (OPS) is our dependent variable we exclude it from our analysis. The first step of principal component analysis is to analyze the eigenvalues of the correlation matrix; the larger eigenvalues are extracted first and will range from 1-24 because there are 24 other observable variables aside from OPS, as shown in Figure 1. In the figure, the 6 largest eigenvalues are 6.25, 4.00, 2.78, 2.31, 1.50, and 1.33 which together account for 75.72% of the standardized variance. Only the first 6 components are used on the criterion of the eigenvalues-greater-than-one rule; in Figure 1 we see that the 7th eigenvalue is just below 1 at 0.97. A useful visual aid that helps determine an appropriate amount of principal components to use is the scree plot. We look for an elbow/bend in the scree plot to help determine the appropriate amount of principal components. In Figure 2, we see that there are a couple "elbow" that occur at about i=3 and i=5 but based on what we have discussed, 6 principal components provides a sufficient summary of our data.



Eigenvalues of the Correlation Matrix: Total = 24 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	6.25034122	2.24396505	0.2604	0.2604
2	4.00637617	1.23109169	0.1669	0.4274
3	2.77528448	0.46339628	0.1156	0.5430
4	2.31188821	0.81109092	0.0963	0.6393
5	1.50079728	0.17284696	0.0625	0.7019
6	1.32795033	0.35371748	0.0553	0.7572
7	0.97423285	0.07993358	0.0406	0.7978
8	0.89429926	0.14079488	0.0373	0.8350
9	0.75350438	0.03772490	0.0314	0.8664
10	0.71577949	0.13979934	0.0298	0.8963
11	0.57598015	0.00775059	0.0240	0.9203
12	0.56822956	0.17101167	0.0237	0.9439
13	0.39721789	0.05269339	0.0166	0.9605
14	0.34452451	0.09043802	0.0144	0.9749
15	0.25408649	0.06854330	0.0106	0.9854
16	0.18554319	0.11375446	0.0077	0.9932
17	0.07178873	0.02498892	0.0030	0.9962
18	0.04679981	0.01063532	0.0019	0.9981
19	0.03616449	0.02928984	0.0015	0.9996
20	0.00687464	0.00469025	0.0003	0.9999
21	0.00218440	0.00206586	0.0001	1.0000
22	0.00011854	0.00008461	0.0000	1.0000
23	0.00003393	0.00003393	0.0000	1.0000
24	0.00000000		0.0000	1.0000

Figure 1. Eigenvalues of the Correlation Matrix

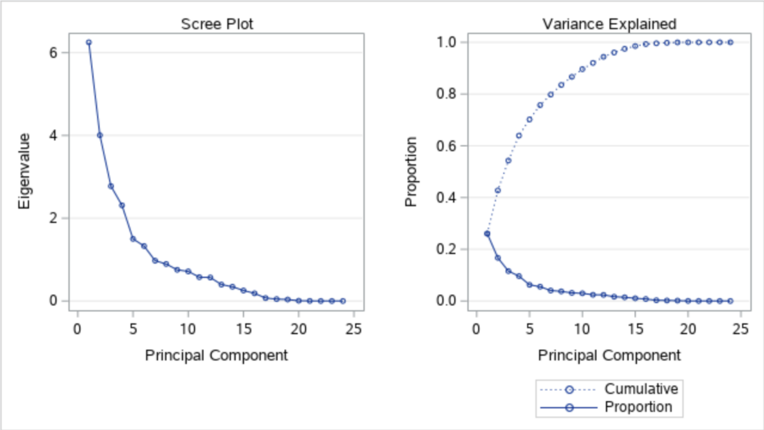
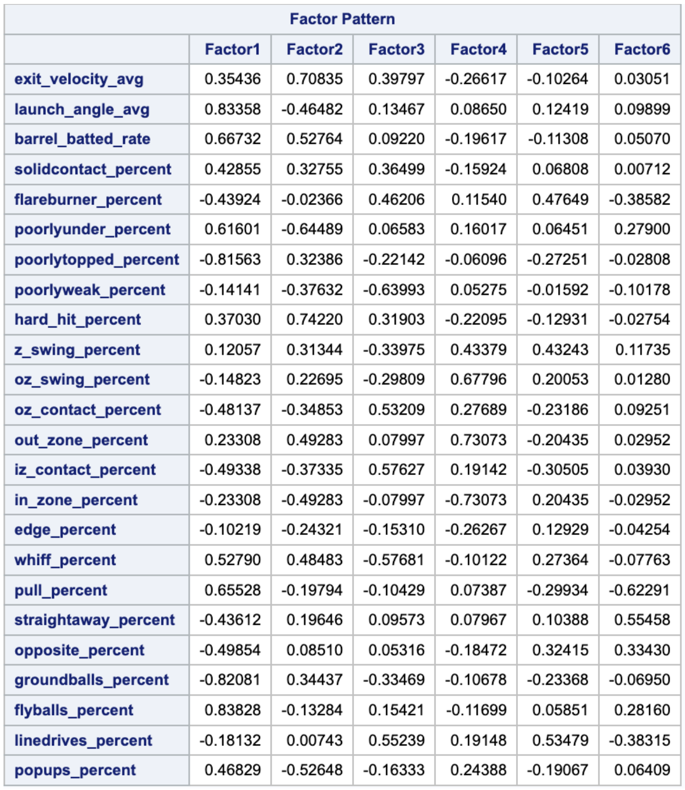


Figure 2. Scree Plot and Variance Explained Plot

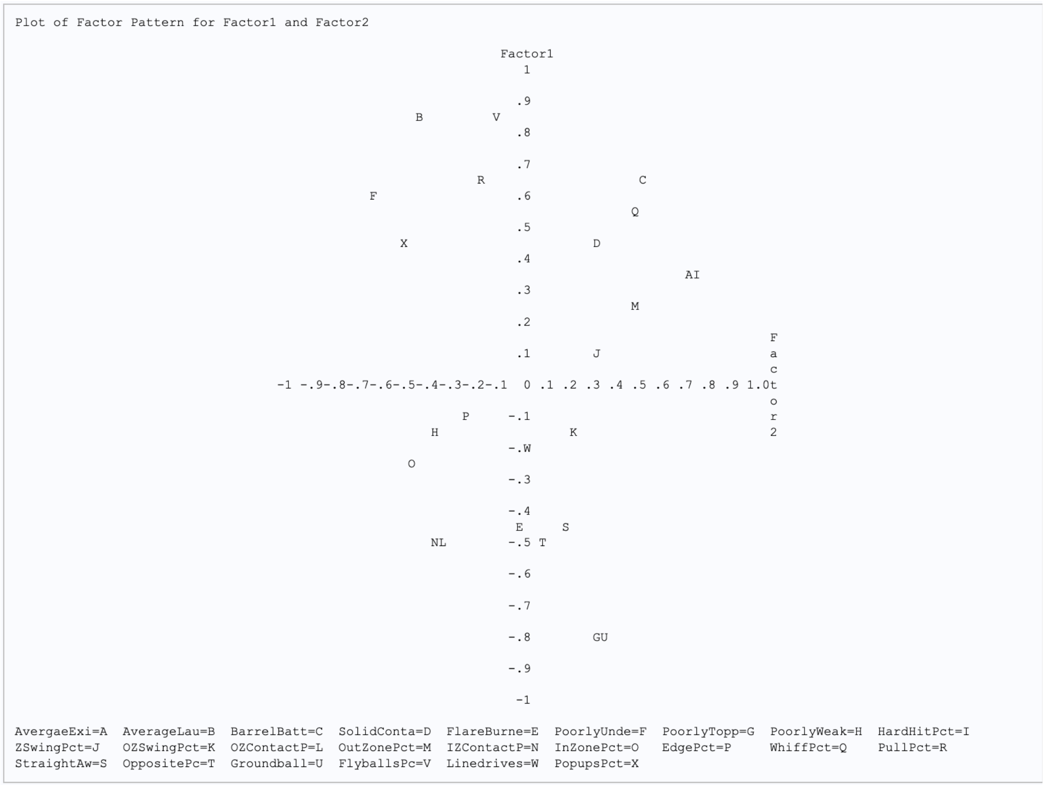
We extend our analysis to Factor Analysis, to help us determine the pattern of the relation among the variables. In Figure 3, we have the factor pattern also known as the factor loading matrix where each element in the matrix is called factor loadings. The factor loadings tell us the strength of the relation between the variables and the components. We can see that Factor1 and Factor2 have a significant correlation to many variables and together account for 42.74% of the variance. Next, this reduced model, that preserves much of the data's variability, is used for cluster analysis and grouping certain hitters together based on similar characteristics.



Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
exit_velocity_avg	0.35436	0.70835	0.39797	-0.26617	-0.10264	0.03051
launch_angle_avg	0.83358	-0.46482	0.13467	0.08650	0.12419	0.09899
barrel_batted_rate	0.66732	0.52764	0.09220	-0.19617	-0.11308	0.05070
solidcontact_percent	0.42855	0.32755	0.36499	-0.15924	0.06808	0.00712
flareburner_percent	-0.43924	-0.02366	0.46206	0.11540	0.47649	-0.38582
poorlyunder_percent	0.61601	-0.64489	0.06583	0.16017	0.06451	0.27900
poorlytopped_percent	-0.81563	0.32386	-0.22142	-0.06096	-0.27251	-0.02808
poorlyweak_percent	-0.14141	-0.37632	-0.63993	0.05275	-0.01592	-0.10178
hard_hit_percent	0.37030	0.74220	0.31903	-0.22095	-0.12931	-0.02754
z_swing_percent	0.12057	0.31344	-0.33975	0.43379	0.43243	0.11735
oz_swing_percent	-0.14823	0.22695	-0.29809	0.67796	0.20053	0.01280
oz_contact_percent	-0.48137	-0.34853	0.53209	0.27689	-0.23186	0.09251
out_zone_percent	0.23308	0.49283	0.07997	0.73073	-0.20435	0.02952
iz_contact_percent	-0.49338	-0.37335	0.57627	0.19142	-0.30505	0.03930
in_zone_percent	-0.23308	-0.49283	-0.07997	-0.73073	0.20435	-0.02952
edge_percent	-0.10219	-0.24321	-0.15310	-0.26267	0.12929	-0.04254
whiff_percent	0.52790	0.48483	-0.57681	-0.10122	0.27364	-0.07763
pull_percent	0.65528	-0.19794	-0.10429	0.07387	-0.29934	-0.62291
straightaway_percent	-0.43612	0.19646	0.09573	0.07967	0.10388	0.55458
opposite_percent	-0.49854	0.08510	0.05316	-0.18472	0.32415	0.33430
groundballs_percent	-0.82081	0.34437	-0.33469	-0.10678	-0.23368	-0.06950
flyballs_percent	0.83828	-0.13284	0.15421	-0.11699	0.05851	0.28160
linedrives_percent	-0.18132	0.00743	0.55239	0.19148	0.53479	-0.38315
popups_percent	0.46829	-0.52648	-0.16333	0.24388	-0.19067	0.06409

Figure 3. Factor Loading Matrix



```
AvergaeExi=A   AverageLau=B   BarrelBatt=C   SolidConta=D   FlareBurne=E   PoorlyUnde=F   PoorlyTopp=G   PoorlyWeak=H   HardHitPct=I
ZSwingPct=J    OZSwingPct=K   OZContactP=L   OutZonePct=M   IZContactP=N   InZonePct=O    EdgePct=P      WhiffPct=Q     PullPct=R
StraightAw=S   OppositePc=T   Groundball=U   FlyballsPc=V   Linedrives=W   PopupsPct=X
```

Figure 4. Plot of Factor1 and Factor2

After analyzing the factor pattern of the factor analysis, a couple of the factors are very similar to each other and are not needed in the model. Factor 1 and Factor 6 are very similar to each other because a player with a high value in each of these factors is a player who hits the ball in the air a lot. A player with a low value in Factor 1 or 6 is a player who hits the ball on the ground a lot. Factor 2 and Factor 3 were also similar to each other. Players with a high value in Factors 2 and 3 were patient hitters who made a lot of good contact. Players with a low value in these factors were hitters who swung the bat alot and didnt make very good contact when doing so.

The cluster analysis moving forward was done with just using Factors 1, 2, 4 and 5. If a player had a high value in Factor 4 it meant that the batter was thrown a lot of pitches out of the strike zone and swung at a lot of pitches out of the strike zone. If a batter had a low Factor 4 value it is because they saw more pitches in the zone and didn't swing at the pitches out of the zone. Factor 5 is contrasting how the batters spray the ball around the field. A high value in

Factor 5 indicates that the batter pulls the ball more often whereas a player with a low value is someone who hits the ball to all fields.

When using proc cluster in SAS, Ward's minimum variance method was applied and the tree plot in Figure 5 helped determine how many clusters should be used.

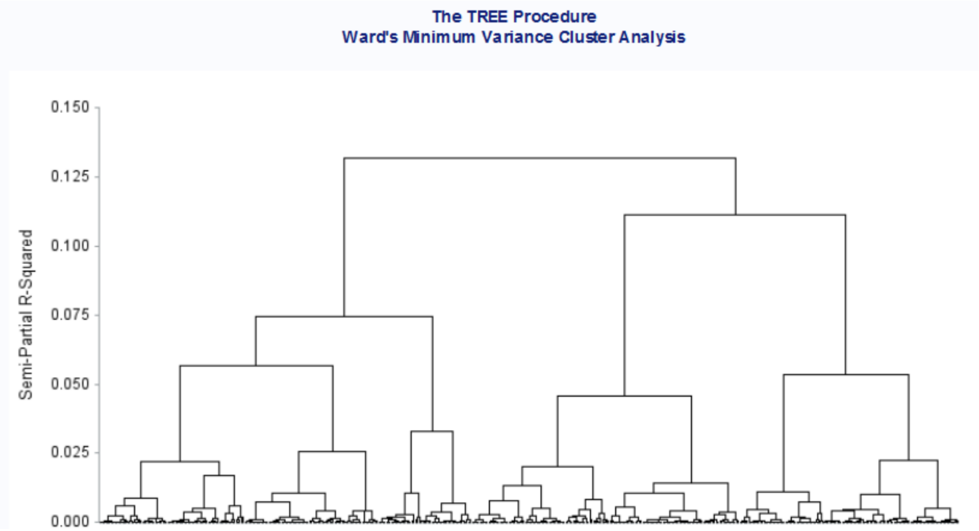


Figure 5. Tree Plot

From this tree plot it was determined to use a number of 3 clusters because the more clusters that were added the less interpretable the clusters became. They all became intermixed with one another and it was hard to differentiate one cluster from another. The next step was to plot these clusters onto a plot with Factor1 on the x-axis and Factor2 on the y-axis that can be seen in Figure 6. Another plot in Figure 7 shows the clusters on a plot with Factor1 on the x-axis and Factor4 on the y-axis.

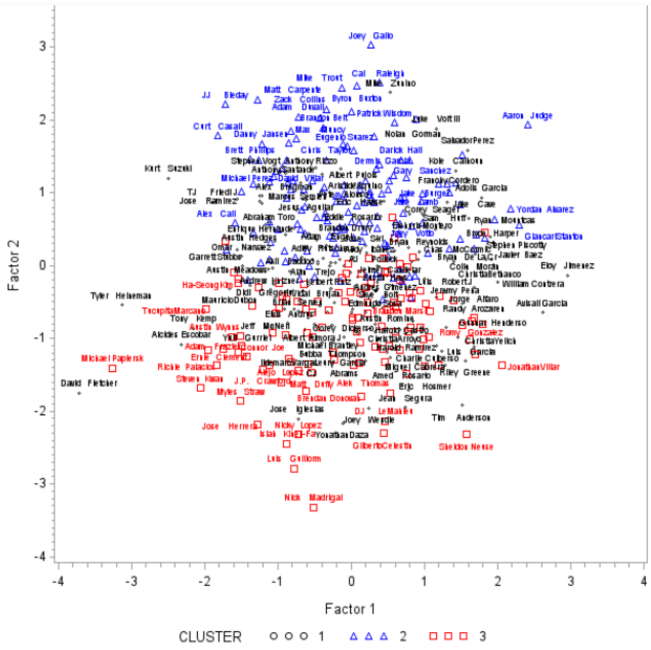


Figure 6. Factor1*Factor2

In Figure 6 all three of the clusters seem to have hitters spread all across the Factor 1 values. Meaning there are a variety of different types of hitters in each cluster(heavy ground ball hitter, heavy fly ball hitter, and in between). The clusters looked grouped by tiers when it comes to Factor 2. Which means they were grouped by how patient they were at the plate. Cluster 3(red) is grouped toward the bottom of Factor 2 values being the least patient, cluster 1(black) is grouped near the middle values of Factor 2, and cluster 2(blue) is towards the top of the graph with the highest Factor 2 values being the most patient. This is backed up by the fact that Joey Gallo at the top of the graph saw 4.2 pitches per plate appearance in 2022 and Nick Madrigal at the bottom of the graph saw 3.81 pitches per plate appearance in 2022. The league average was around 3.95.

Conclusion

As shown in the analysis, the issue of creating a sufficient prediction model for the on-base plus slugging of MLB batters requires extensive data reduction as a result of high correlation amongst variables. This was done through the use of principal component analysis which retained 75.72% of the standardized variance of the data. In addition to this, the factor analysis yielded a final model that consisted of only 6 factors as opposed to 25 original variables. This factor model was then used for more efficient cluster analysis used to group similar players together. As seen in Figure 6, there are different types of successful hitters and different types of bad hitters in baseball. So it isn't easy to group by a hitter's tendency to hit the ball in the air or on the ground because there are good and bad hitters with each tendency. It is easier to group by how patient the hitters are at the plate however.

References

“Factor Analysis.” *Statistics Solutions*, 10 Aug. 2021, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/.

Johnson, Richard Arnold, and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007.

Code

FULL CODE

```
FILENAME REFFILE "C:/Users/014497819/Downloads/stats (7).csv";

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=BASEBALL;
        GETNAMES=YES;
RUN;

DATA BASEBALL; *renaming vars for simplicity;
RENAME exit_velocity_avg=AvergaeExitVelocity
launch_angle_avg=AverageLaunchAngle
barrel_batted_rate=BarrelBattedRate
solidcontact_percent=SolidContactPct
flareburner_percent=FlareBurnerPct
poorlyunder_percent=PoorlyUnderPct
poorlytopped_percent=PoorlyToppedPct
poorlyweak_percent=PoorlyWeakPct
hard_hit_percent=HardHitPct
z_swing_percent=ZSwingPct
oz_swing_percent=OZSwingPct
oz_contact_percent=OZContactPct
out_zone_percent=OutZonePct
iz_contact_percent=IZContactPct
in_zone_percent=InZonePct
edge_percent=EdgePct
whiff_percent=WhiffPct
pull_percent=PullPct
straightaway_percent=StraightAwayPct
opposite_percent=OppositePct
groundballs_percent=GroundballsPct
flyballs_percent=FlyballsPct
linedrives_percent=LinedrivesPct
popups_percent=PopupsPct;
SET BASEBALL;
full_name=cat(' ',_first_name , last_name);
RUN;


*Cleaning up data set;
DATA BASEBALL;
        set BASEBALL (drop = last_name _first_name player_id  year VAR30);
RUN;


/******** Correlation Matrix *******/
proc corr data=BASEBALL;
run;


*Factor Analysis to determine how many factors to use;
PROC FACTOR DATA=BASEBALL SIMPLE CORR;
        VAR AvergaeExitVelocity
AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct
PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct
IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct
FlyballsPct LinedrivesPct PopupsPct;
RUN;

PROC PRINCOMP DATA=BASEBALL OUT=PRINCOMP;
        VAR AvergaeExitVelocity
```

```
AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct
PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct
IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct
FlyballsPct LinedrivesPct PopupsPct;
RUN;

*factor analysis without OPS var;
*Determined to use 6 factors;
Proc Factor DATA=BASEBALL METHOD=prin NFACT=6 out=baseballfact
    ROTATE=NONE PREPLOT PLOT; *NO ROTATION;
                  var AvergaeExitVelocity
AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct
PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct
IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct
FlyballsPct LinedrivesPct PopupsPct;
run;



/**** Decided to use only Factors 1,2,4&5 for clustering because after interpreting the factors Factor3 was similar to
                                   factor 2 and factor 6 was similar to factor 1*/

Proc Cluster Data=baseballfact Method=ward OutTree=baseballtree pseudo;
                                                   Var Factor1 Factor2 Factor4 Factor5; *all of the variables to be used in cluster analysis;
 Id full_name;
Run;
* Method options: single, complete, average, centroid, ward, etc. ;

GOptions Reset=Symbol Reset=Axis;
Proc GPlot Data=baseballtree;
 Plot _HEIGHT_*_NCL_=1 / VAxis=Axis1;
 Axis1 Label=(A=90);
 Symbol1 C=Black V=Dot I=SplineS;
Run;
Quit;

Proc Gplot;
 plot _PSF_*_NCL_=1;
Axis1 Label=(A=90);
 Symbol1 C=Black V=Dot I=SplineS H=.34;
run;

Proc Tree Data=baseballtree NCL=3 out=clusters VAxis=Axis1;
 Id full_name;
 Axis1 Label=(A=90);
Run;


*Merge the datasets with clusters and the dataset with factors;
Proc Sort data=baseballfact;
By full_name; run;
Proc sort data=clusters;
by full_name; run;
data baseballclust;
merge baseballfact clusters;
by full_name;
drop ClusName; run;




/***************** Plot FACTOR1 and FACTOR2 *********/

Proc Gplot data=baseballclust;
plot Factor1*Factor2=cluster / VAxis=Axis1 HAxis=Axis2;
 Axis1 Label=(A=90 "Factor 2")
      Order=(-4 To 4 By 1)
      Length=5.75in;
 Axis2 Label=("Factor 1")
      Order=(-4 To 4 By 1)
      Length=5.75in;
 Symbol1 C=Black V=Circle    I=None Pointlabel=(C=Black H=0.75 "#full_name");
 Symbol2 C=Blue  V=Triangle  I=None Pointlabel=(C=Blue  H=0.75 "#full_name");
 Symbol3 C=Red   V=Square    I=None Pointlabel=(C=Red   H=0.75 "#full_name");
Run;
Quit;
```