Milestone 2

NETIDs: el3418 ac8480

Methodology: Describe the methodology that you applied.

After much searching and web-scraping efforts, we decided to gather per game and advanced college player statistics from Sports Reference, and restricted it from 2000 to 2018 due to time constraints and tediousness of the manual process since there was no easy way to autonomously scrape the data efficiently due to pay wall and web scraping rate limits.

Some confounding variables mentioned in Milestone 1, such as game condition, defensive positioning distance, shot distance, etc. were too specific for public availability. Additionally, we noticed in our college stats dataset, there were many inconsistencies in stat recording among players. At first, we imputed using a forward fill since our data was already ordered by year. We chose to use the median since there wasn't that much variance in the dataset, and the number of years a college player usually plays is capped at 4 years, but rarely at 5. However, after closely looking at the data, we instead interpolated based on team conference for a more accurate representation taking into account their strength of schedule (SOS) within their own conference, and to prevent extra noise being added to the dataset through cross-conference interference. There were still some missing values even after this due to some conferences not collecting player data, thus we decided to use the geographically closest conference to impute using domain knowledge. If the entire conference data was missing for players, we imputed the remaining stats conditioned on player positions in our dataset. Finally, we standardized our variables to prevent over or under-weighting of certain features before we fed it into our model.

The first topic mentioned in the Methodology section of Milestone 1 was to examine correlation and causal effects. To do this, we first generated a correlation heatmap (Figure 7, 8). Our main focus was the relationships between our target variable, VORP, and the features. Due to the multicollinearity observed in our dataset from the correlation matrix, it made sense to apply PCA next. We do this first by examining what the data would look like when reduced to two dimensions as a baseline, then graph the explained variance against the number of components to select a good k, which was found to be k=3 (Figure 4). This would allow us to see which features are responsible for capturing the variance in our dataset and remove much of the noise, which would complement our next step: Applying a K-Means model to examine the separability of players based on their features. Utilizing the Elbow Method and checking against distortion and inertia, we iteratively checked the clusters starting at 2, 5, 7, until 10 clusters (Figure 9, 10). Prior domain knowledge would suggest the need to train separate classifiers based on player position, or at least some sort of player clustering, and these methods will be run in Milestone 3 in order to validate this claim.

After performing our EDA and pre-processing, we trained some predictive models. Looking at VORP, the target variable of our modeling procedure, out of 933 college players who now play in the NBA, 189 of them became all-stars (our benchmark for an exceptional player). Essentially, VORP is

the value of replacement of a player. The highest NBA all-star caliber player career VORP in our dataset belongs to Michael Jordan at 116.1. After merging the VORP with the college basketball stats of players, we input into a baseline linear regression model. Instead of using seasonal data, we used the full career college data to predict career NBA VORP. With this in mind, we are aware that some players are currently still active and their career VORP may be higher or lower than what the model currently predicts. One alternative we may explore in Milestone 3 is average seasonal VORP (career VORP divided by the number of seasons played).

Results: Describe your results in a succinct and effective way.

The first notable result was our correlation heatmap. Our main focus was the relationships between our target variable, VORP, and the features. Curiously, what was observed was that VORP tended to correlate much more with traditionally guard heavy stats than traditional big men stats. For example, assists (0.16) vs rebounds (0.08), or steals (0.26) vs blocks (-0.018). Even the more granular, contextual statistics showed this same relationship - AST% (0.18) vs TRB% (0.032).

To further examine this, PCA was applied on the dataset. Before performing PCA, our hypothesis was that the first and second principal directions would show the difference between guards and centers. This was confirmed in our results (Figure 3) - the first principal component represented the guard related stats, while the second principal component captured the center related stats. For example, in the first component some of the highest positive coefficients were AST (0.236885) and 3PA (0.281975), while some of the highest negative coefficients were BLK (-0.199675) and TRB% (-0.230915). The second principal component displayed much of the opposite - some of the highest positive coefficients were 2PA(0.310986) and TRB (0.287831), while some of the highest negative coefficients were 3PAr (-0.110406) and TOV% (-0.092503). When reduced to 3 dimensions, it appeared that the 3rd dimension tried to capture efficient '3-and-D' wings by prioritizing TS%, 3PAr, and low TOV% (Figure 5). Following that, the results of our K-Means clustering suggested that it would be difficult to separate clusters in a 2D space. However, we achieve decent separability in the 3D space, as shown in Figure 6.

Our linear regression using the PCA features gave us an average cross-validation score of 12.53 with scoring set to root mean squared error. The root mean square error (RSME) was 10.985 when ran the model on the test set. Because of the magnitude of high error, due to small size of our dataset and the large fluctuations caused by our loss metric, we decided to switch to Mean Absolute Error (MAE), where the 5-folds cross-validation came out to be 7.96, and the MAE on the test set was 7.53. This was worse overall compared to a linear regression directly using the features: FT%, FG%, 3P%, and AST, which are free throw, field goal, three-point %s, and assists respectively. The coefficients of this model were ~ 4.23, 52, 3.39, and .0067, compared to higher weights: ~ 0.439, 0.397, and -0.039 from a 3 component PCA transformation of the original dataset. It is interesting to see that FG% without the PCA transformation adds the most weight to the regression results, which makes sense because VORP is calculated through considering how many more points a replacement player can score. The 3rd principal component seems to take away from VORP score contribution if they mainly shoot from distance and play defense, thus they are more so role-players.

Analysis: What conclusion can you draw from your results?

After collecting the data, we performed a kernel density contour plot on various features. One example of this is shown in Figure 1, where we wanted to explore the relationship between free throw (FT) shooting % and field goal (FG) shooting %. After min max scaling the dataset so that the ranges are more compact, we see that our data captures that higher FT% moves with higher FG% with a majority of FG% around ~0.4. We can also observe it stretch lower for FT% as FG% increases due to big men getting easier shots in the paint area.

VORP's high correlation with guard heavy stats led to the assumption that perhaps there was an imbalance in the dataset, where guards were more prevalent as first round draft picks. However, as shown in Figure 2, this was not the case. Upon inspection, VORP estimates the points that a player contributes above a replacement level player. From that perspective, it can be easy to see how guards would be more correlated - they tend to be more ball dominant and score more as a result. As for the PCA and K-Means, we believe the results show good promise for separability and offer us avenues for feature engineering that will be expanded upon in the next section.

Our initial linear regression results are not good with a high MAE error, however, this offers a baseline before adding more features as building accurate models is an iterative process.

Plan for additional analysis: Based on your results, propose additional analysis

We are thinking of using thresholds for VORP based on percentiles to generate labels of player effectiveness on a team, such as high, medium, or low. The average career VORP among the all-stars in our dataset was ~28.84. This may be too high of a threshold for a simple binary classification among our college players, however we can simply use average seasonal VORP to make this objective more feasible. This will be useful when we run Random forest classifiers on our dataset.

This can be evaluated through macro-average precision scores, where the precision of each class will be collected and averaged. Precision would be used here because it is costly for a team if a player that isn't as effective, but falsely projected as "highly effective," is paid more using the franchise's budget.

Another avenue for exploration would be to create separate models for each player cluster - our results showed viable clustering when PCA reduced down to 3 dimensions, and from domain knowledge we know it makes intuitive sense to make the prediction task for our models easier through this approach. There can be multiple ways to accomplish this - we could subset our dataset based on player position or cluster label. We could also selectively use PCA - for example, it would make sense to use PCA to take in ORB, DRB, TRB, ORB%, DRB%, and TRB% and reduce it down to one dimension for use as a training feature.

Finally, in our next steps we will incorporate new features, through interaction terms among certain features, or also including pick standing of the college player as a dummy variable.

Appendix

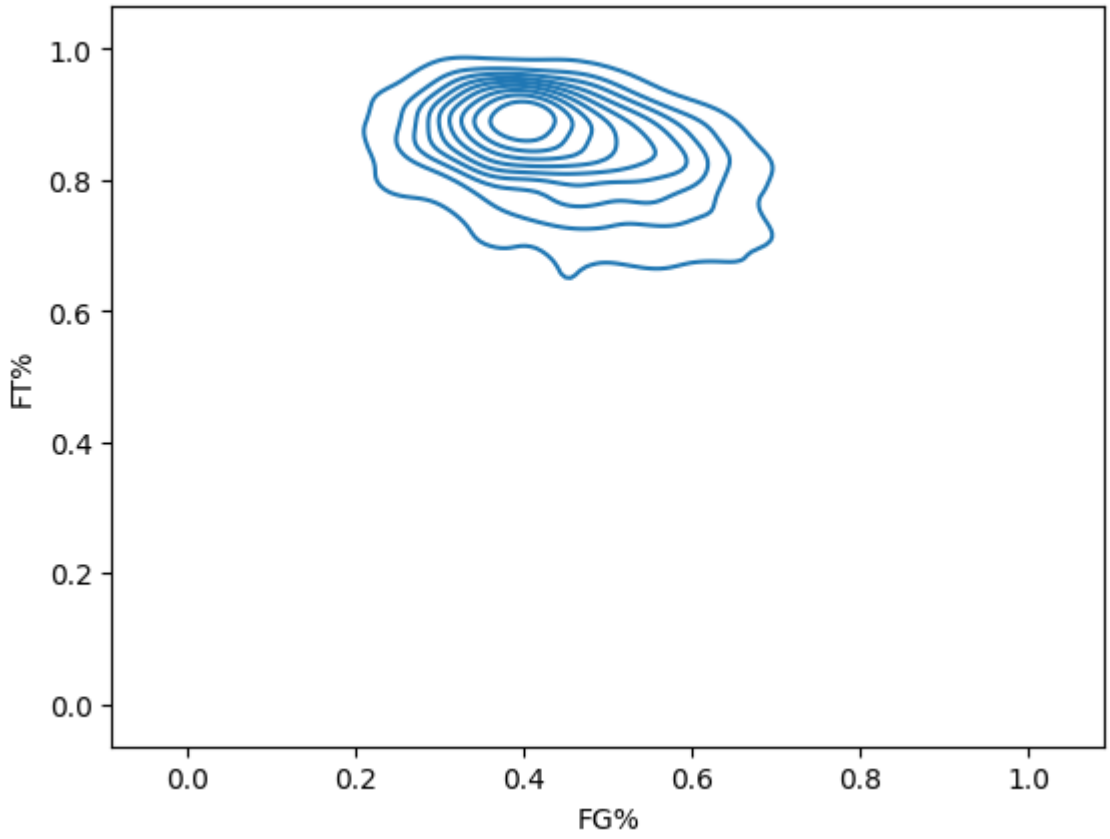

Figure 1

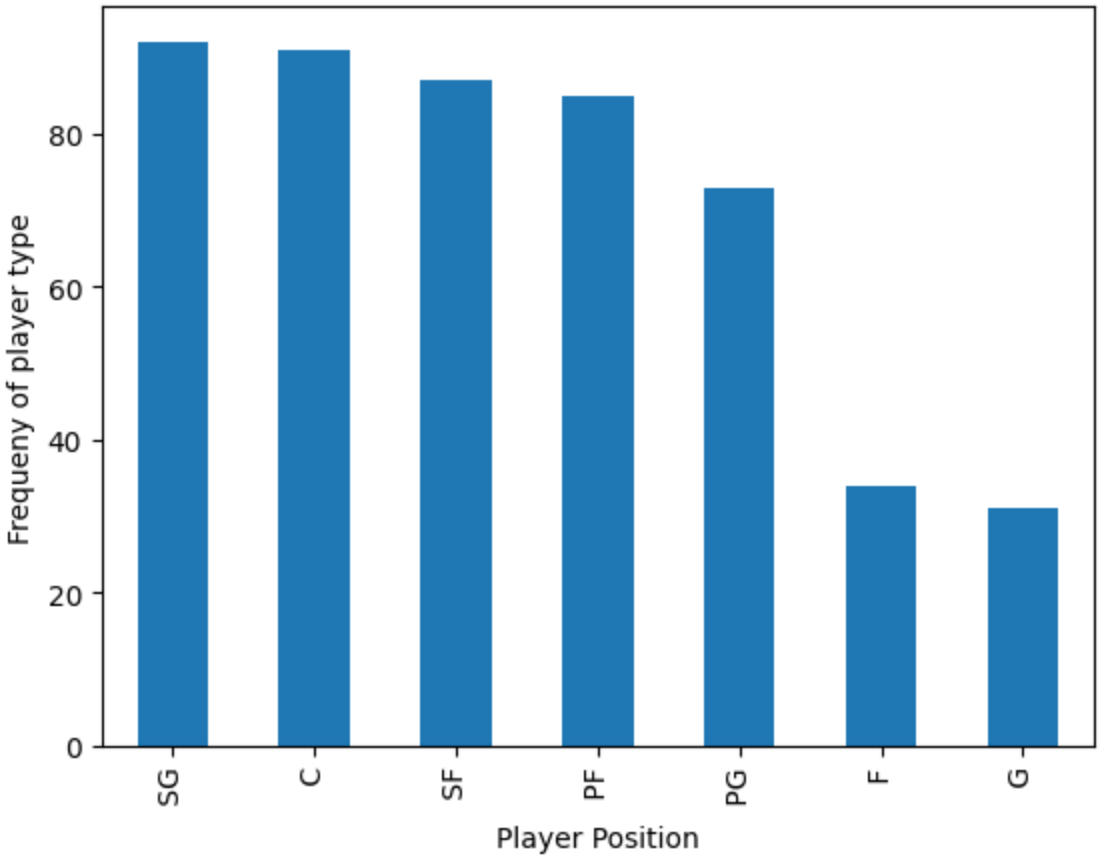
Frequeny of player type
0
20
40
60
80
SG
C
SF
PF
PG
F
G
Player Position

Figure 2

	PC-1	PC-2
FG	0.164363	0.284384
FGA	0.238717	0.221058
FG%	-0.222614	0.138285
2P	0.024332	0.327212
2PA	0.079699	0.310986
2P%	-0.179442	0.094900
3P	0.272930	-0.035835
3PA	0.281975	-0.037728
3P%	0.083536	-0.018152
FT	0.164585	0.250570
FTA	0.116598	0.277705
FT%	0.223407	0.026699
ORB	-0.180255	0.237101
DRB	-0.072568	0.274729
TRB	-0.115429	0.287831
AST	0.236885	-0.009382
STL	0.203111	0.064866
BLK	-0.199675	0.147341
TOV	0.191094	0.163539
PF	-0.050504	0.174570
PTS	0.209052	0.262018
TS%	0.045704	-0.008106
eFG%	-0.126926	0.089647
3PAr	0.199428	-0.110406
FTr	0.007919	0.023314
ORB%	-0.210401	0.092657
DRB%	-0.152779	0.127155
TRB%	-0.230915	0.175700
AST%	0.216071	-0.028443
STL%	0.092317	-0.016187
BLK%	-0.223135	0.067439
TOV%	-0.052484	-0.092503
USG%	0.133854	0.194017

Figure 3

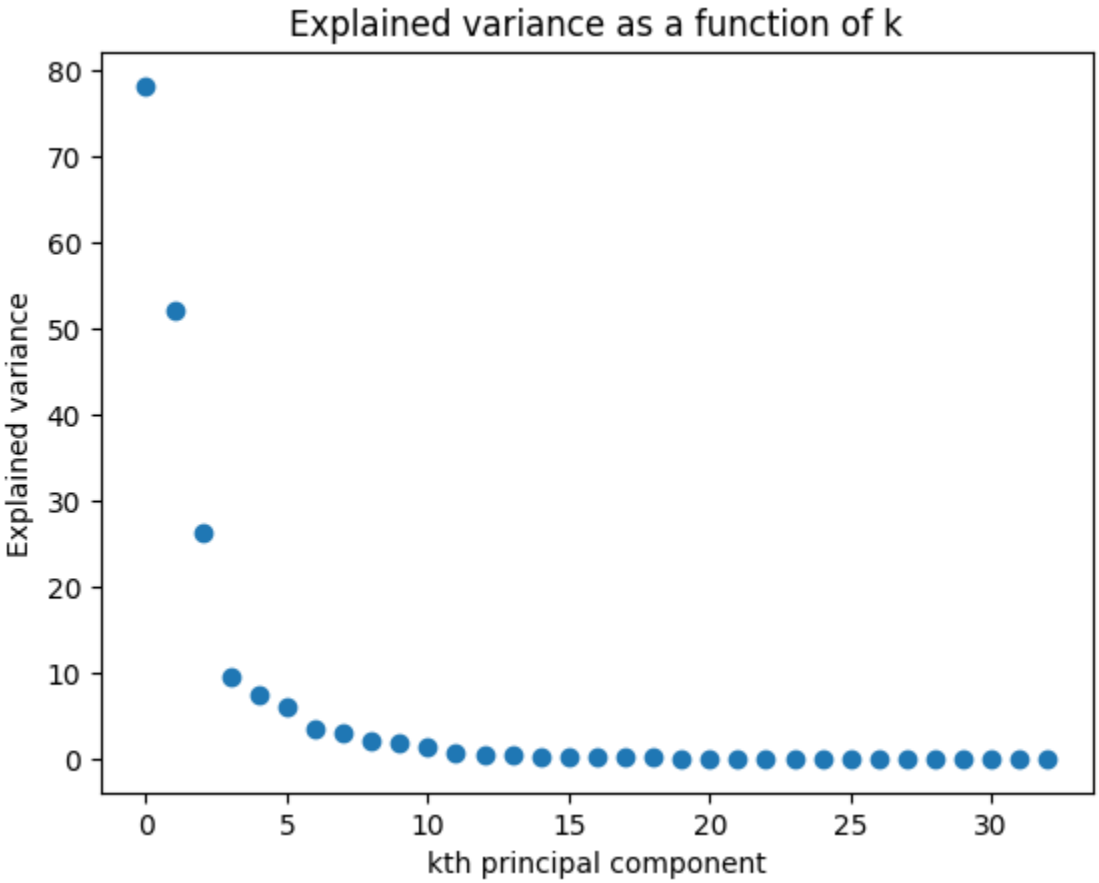
Explained variance as a function of k
Explained variance
kth principal component
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
30

Figure 4

	PC-1	PC-2	PC-3
FG	0.164363	0.284384	0.070597
FGA	0.238717	0.221058	0.049955
FG%	-0.222614	0.138285	0.069577
2P	0.024332	0.327212	-0.004597
2PA	0.079699	0.310986	-0.032427
2P%	-0.179442	0.094900	0.103965
3P	0.272930	-0.035835	0.143657
3PA	0.281975	-0.037728	0.119856
3P%	0.083536	-0.018152	0.108938
FT	0.164585	0.250570	-0.011223
FTA	0.116598	0.277705	-0.032845
FT%	0.223407	0.026699	0.066771
ORB	-0.180255	0.237101	0.033492
DRB	-0.072568	0.274729	0.032332
TRB	-0.115429	0.287831	0.035466
AST	0.236885	-0.009382	-0.193568
STL	0.203111	0.064866	-0.125015
BLK	-0.199675	0.147341	0.048386
TOV	0.191094	0.163539	-0.177583
PF	-0.050504	0.174570	-0.065311
PTS	0.209052	0.262018	0.069489
TS%	0.045704	-0.008106	0.497241
eFG%	-0.126926	0.089647	0.131043
3PAr	0.199428	-0.110406	0.400257
FTr	0.007919	0.023314	0.443442
ORB%	-0.210401	0.092657	-0.001247
DRB%	-0.152779	0.127155	-0.046114
TRB%	-0.230915	0.175700	-0.015859
AST%	0.216071	-0.028443	-0.244671
STL%	0.092317	-0.016187	-0.028587
BLK%	-0.223135	0.067439	0.047472
TOV%	-0.052484	-0.092503	-0.337135
USG%	0.133854	0.194017	-0.153396

Figure 5

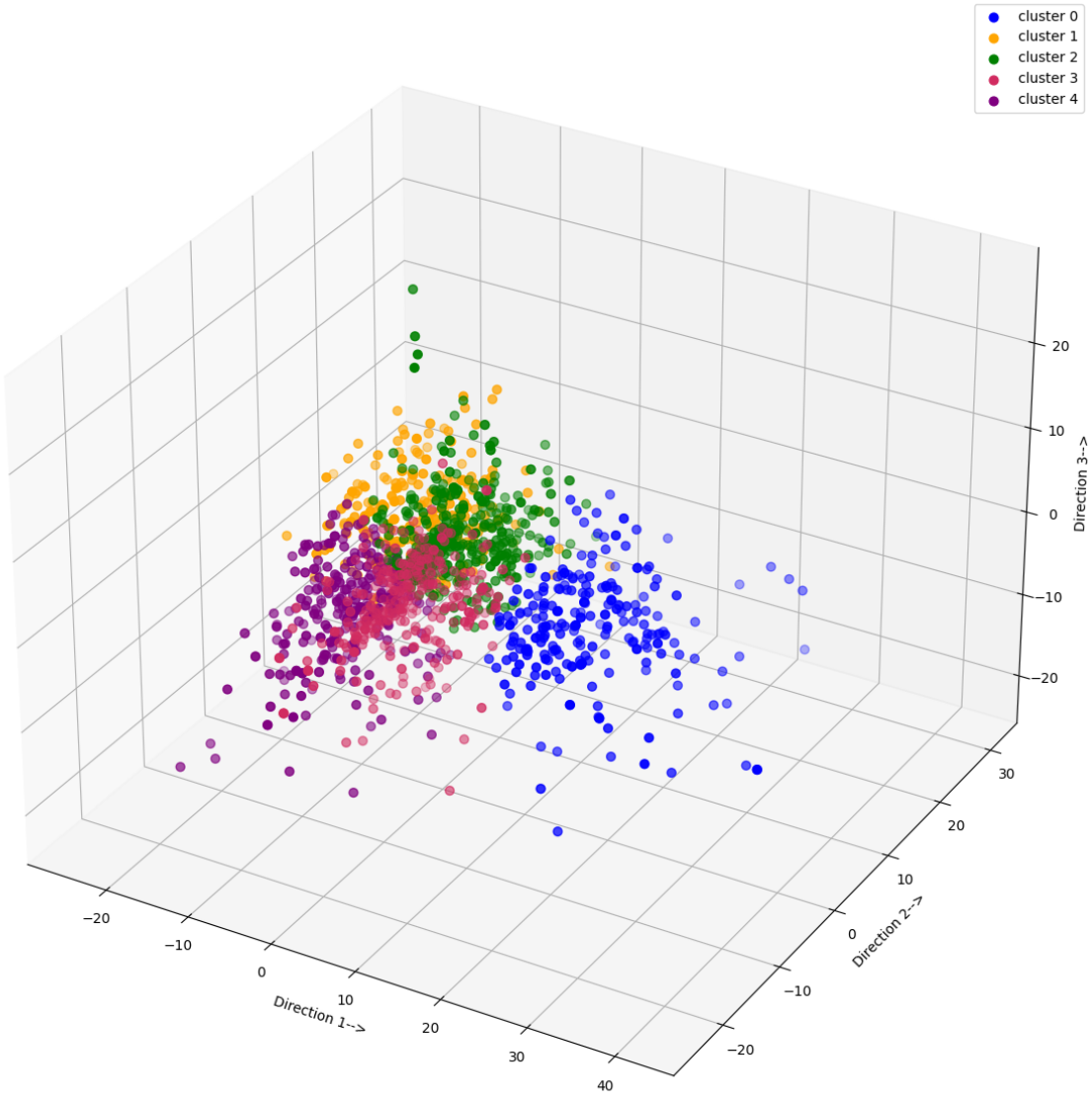
cluster 0
cluster 1
cluster 2
cluster 3
cluster 4
Direction 1-->
Direction 2-->
Direction 3-->

Figure 6

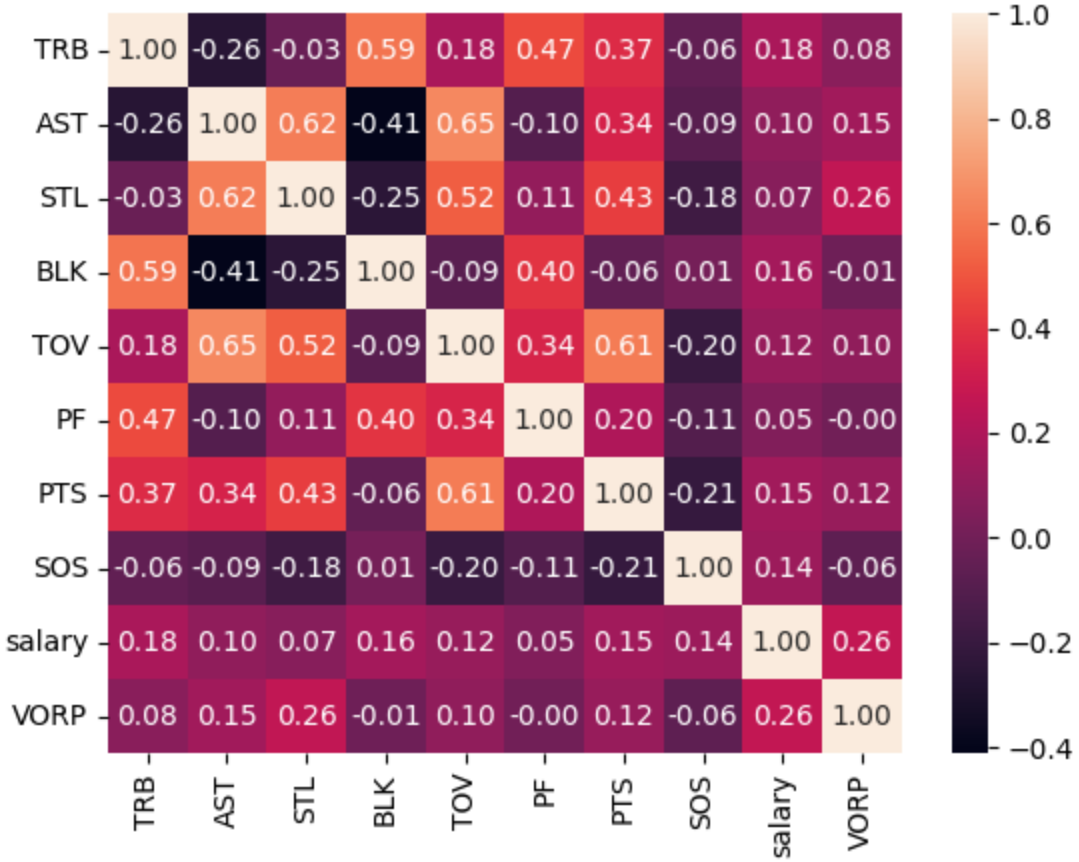
TRB 1.00 -0.26 -0.03 0.59 0.18 0.47 0.37 -0.06 0.18 0.08
AST -0.26 1.00 0.62 -0.41 0.65 -0.10 0.34 -0.09 0.10 0.15
STL -0.03 0.62 1.00 -0.25 0.52 0.11 0.43 -0.18 0.07 0.26
BLK 0.59 -0.41 -0.25 1.00 -0.09 0.40 -0.06 0.01 0.16 -0.01
TOV 0.18 0.65 0.52 -0.09 1.00 0.34 0.61 -0.20 0.12 0.10
PF 0.47 -0.10 0.11 0.40 0.34 1.00 0.20 -0.11 0.05 -0.00
PTS 0.37 0.34 0.43 -0.06 0.61 0.20 1.00 -0.21 0.15 0.12
SOS -0.06 -0.09 -0.18 0.01 -0.20 -0.11 -0.21 1.00 0.14 -0.06
salary 0.18 0.10 0.07 0.16 0.12 0.05 0.15 0.14 1.00 0.26
VORP 0.08 0.15 0.26 -0.01 0.10 -0.00 0.12 -0.06 0.26 1.00
TRB
AST
STL
BLK
TOV
PF
PTS
SOS
salary
VORP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Figure 7

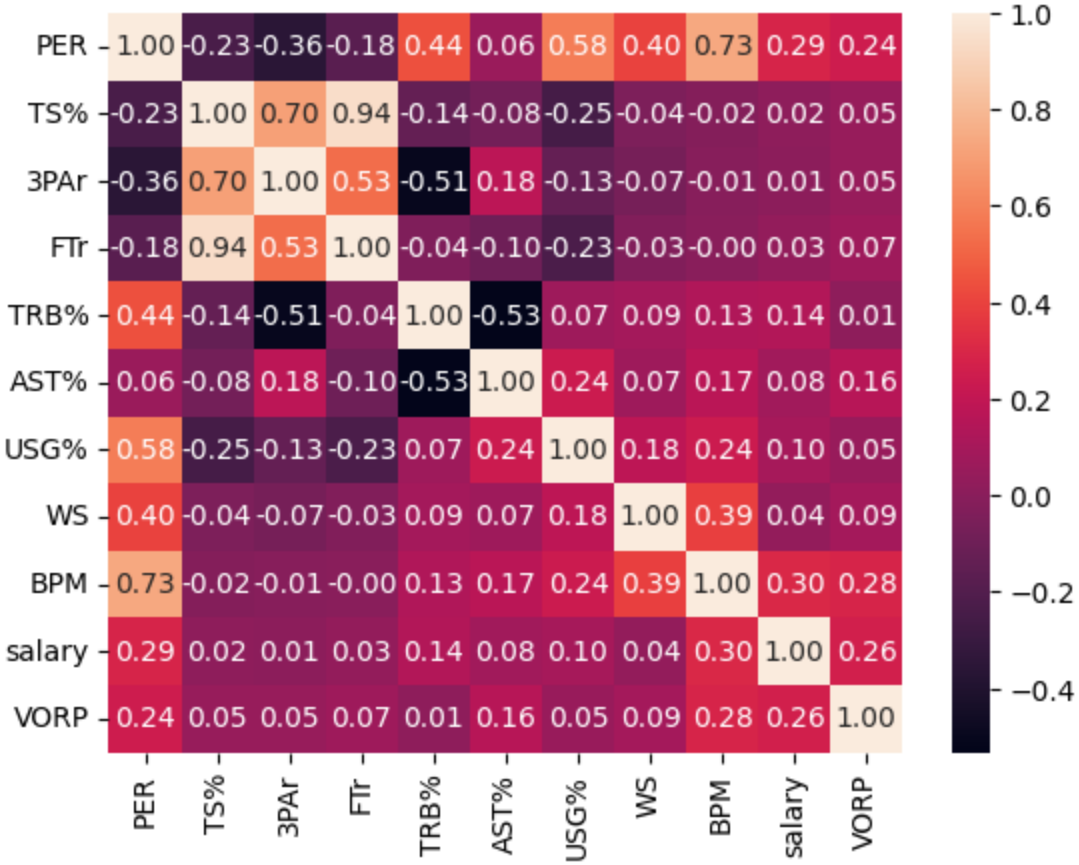
PER 1.00 -0.23 -0.36 -0.18 0.44 0.06 0.58 0.40 0.73 0.29 0.24
TS% -0.23 1.00 0.70 0.94 -0.14 -0.08 -0.25 -0.04 -0.02 0.02 0.05
3PAr -0.36 0.70 1.00 0.53 -0.51 0.18 -0.13 -0.07 -0.01 0.01 0.05
FTr -0.18 0.94 0.53 1.00 -0.04 -0.10 -0.23 -0.03 -0.00 0.03 0.07
TRB% 0.44 -0.14 -0.51 -0.04 1.00 -0.53 0.07 0.09 0.13 0.14 0.01
AST% 0.06 -0.08 0.18 -0.10 -0.53 1.00 0.24 0.07 0.17 0.08 0.16
USG% 0.58 -0.25 -0.13 -0.23 0.07 0.24 1.00 0.18 0.24 0.10 0.05
WS 0.40 -0.04 -0.07 -0.03 0.09 0.07 0.18 1.00 0.39 0.04 0.09
BPM 0.73 -0.02 -0.01 -0.00 0.13 0.17 0.24 0.39 1.00 0.30 0.28
salary 0.29 0.02 0.01 0.03 0.14 0.08 0.10 0.04 0.30 1.00 0.26
VORP 0.24 0.05 0.05 0.07 0.01 0.16 0.05 0.09 0.28 0.26 1.00
PER
TS%
3PAr
FTr
TRB%
AST%
USG%
WS
BPM
salary
VORP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Figure 8

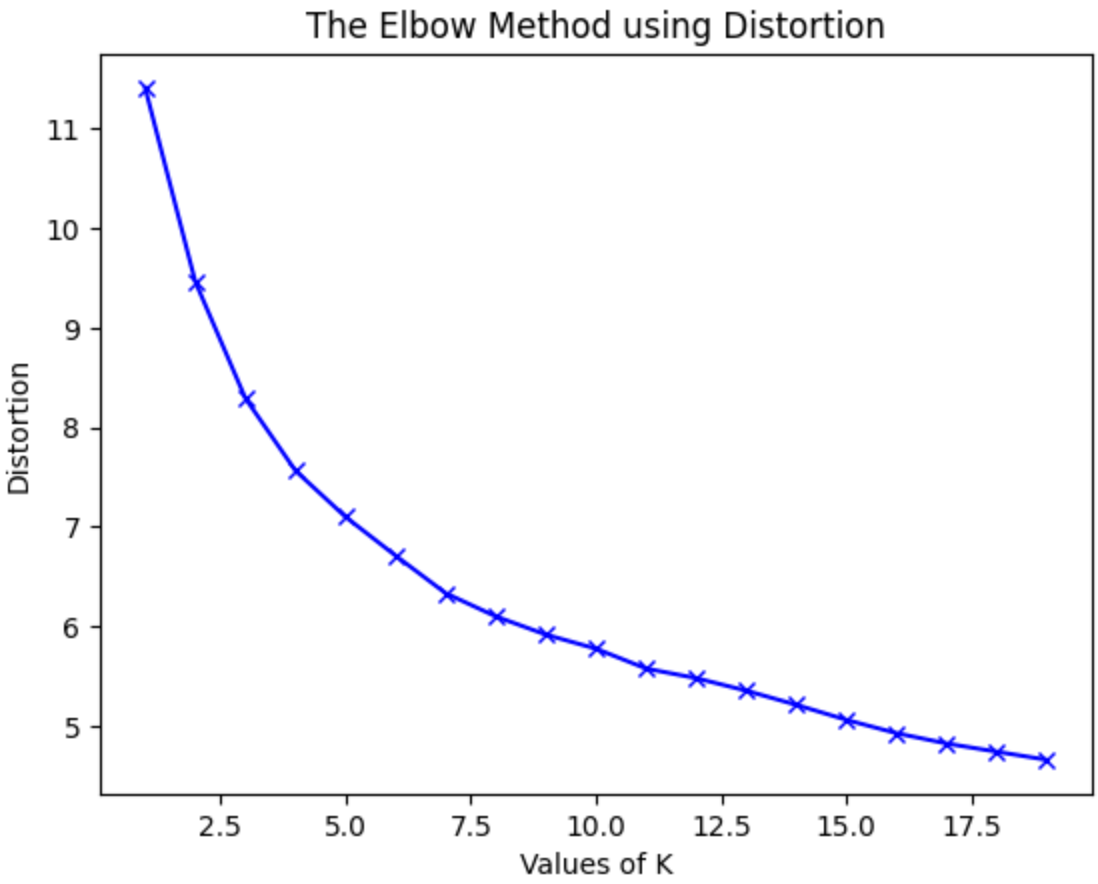
The Elbow Method using Distortion
Distortion
11
10
9
8
7
6
5
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Values of K

Figure 9

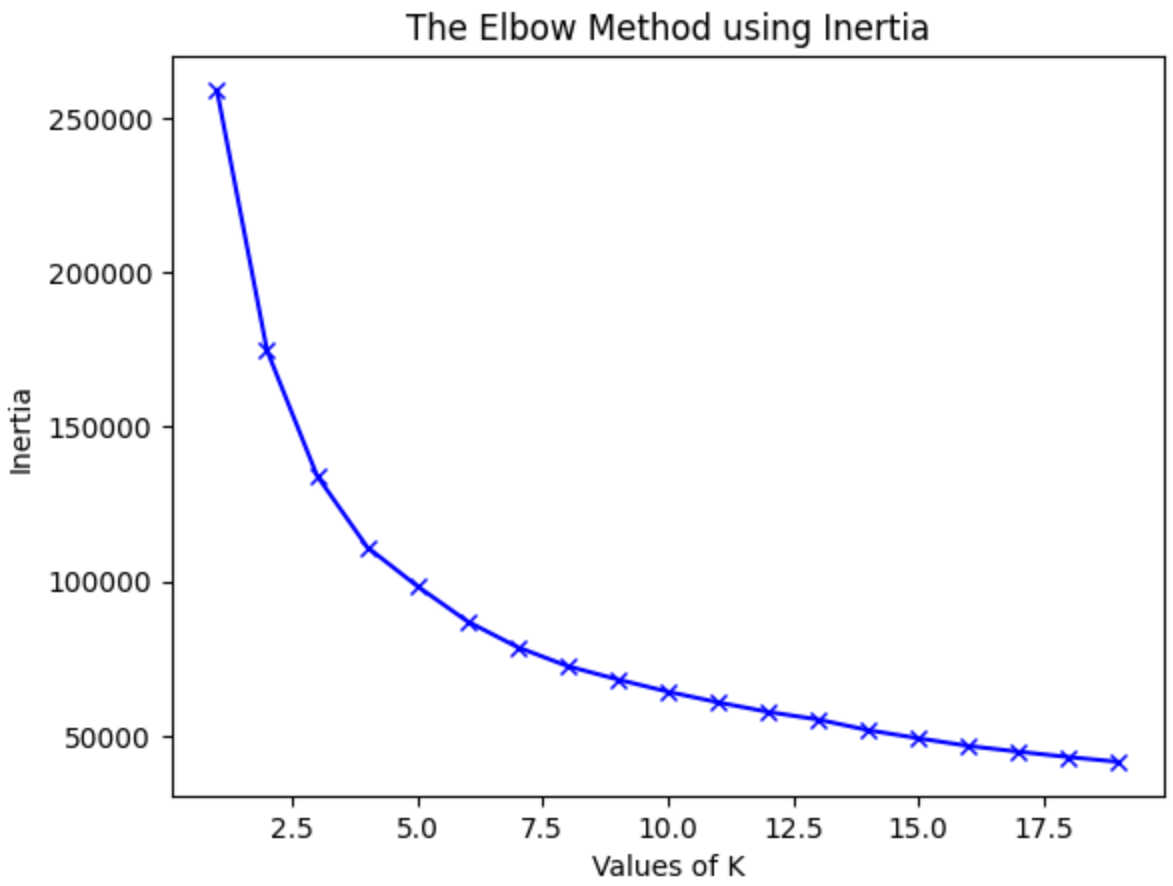
The Elbow Method using Inertia
250000
200000
150000
100000
50000
Inertia
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Values of K

Figure 10