**Point Guard Statistics Report**

**Introduction:**

The dataset I will be looking at is the per-game career average statistics of Point Guards in the NBA (usually always the shortest players on the court). My goal in investigating their career statistics was to find any underlying patterns that occur in that specific position. With these patterns, I can make some solid statements and assumptions on playing that role in basketball. I picked this dataset because I have played and watched basketball ever since I came to Canada.

**Data:**

Here is the data I was working with:

All these statistics were found on [**http://www.basketball-reference.com/**](http://www.basketball-reference.com/). This data contains the names of 49 different Point Guards in the NBA (retired and active) all of whom have played during and after 1979-1980 (when the 3 point line was introduced). Each of these players has a total of 30 variables (categorical and numerical). Here is what each statistic means:

* **Season:** The per-game statistics of a player in that year
  + Our column is filled with “Career” because we will be looking at career per-game averages
* **Age:** The per-game average statistics of a player during a certain age
  + Since we are looking at career numbers, age has no significance
* **Tm:** The per-game average statistics of a player when he played for a certain team
  + Since we are looking at a player’s career as a whole, we will not be looking at what team a player played best on
* **Lg:** The per-game average statistics of a player who played in a certain league
  + As I stated before, all the players are from the NBA, hence this column is filled with “NBA”
* **Pos:** The per-game average statistics of a player who played in a certain position on a team
  + As previously stated, I have picked a dataset of point guards so this column is empty
* **G:** The amount of total games a player has played throughout his career
* **GS:** The amount of games a player “started” throughout his career
* **MP:** The average amount of minutes played per game (maximum of 48 minutes)
* **FG:** The average amount of field goals made per game
  + A field goal made is when the player scores on the opposing basket (excluding free throws)
* **FGA:** The average amount of field goals attempted per game
* **FG.**: The percentage ratio of FG vs. FGA
* **X3P:** The average amount of 3 pointers made per game
* **X3PA:** The average amount of 3 pointers attempted per game
* **X3P.**: The percentage ratio of X3P vs. X3PA
* **X2P:** The average amount of shots made inside the 3 point line per game
* **X2PA:** The average amount of shots attempted inside the 3 point line per game
* **X2P.**: The percentage ratio of X2P vs. X2PA
* **eFG.:** This is just like the FG. accept it adjusts for the fact that 3-point field goals are worth 50% more than 2-point field goals, so it adds an additional parameter
  + **Equation:** eFG. = (FG + (0.5 x 3PT)) / FGA
* **FT:** The average amount of free throws made per game
  + A free throw is an unopposed attempt to score points from a restricted area on the court (the **free throw**), and are generally awarded after a foul on the shooter by the opposing team. Each successful **free throw** is worth one point.
* **FTA:** The average amount of free throws attempted per game
* **FT.:** The percentage ratio of FT vs FTA
* **ORB:** The average amount of offensive rebounds collected per game
  + An offensive rebound is when a player collects the ball after a missed shot from his/her team on the opposing basket
* **DRB:** The average amount of defensive rebounds collected per game
  + A defensive rebound is when a player collects the ball after a missed shot from the opposing team on your basket
* **TRB:** The average amount of total rebounds collected per game
  + ORB + DRB
* **AST:** The average amount of assists performed by a player per game
  + An assist in basketball is when a player passes the ball to a teammate in a way that leads to a score by field goal
* **STL:** The average amount of steals performed by a player per game
  + A steal in basketball occurs when a defensive player legally causes a turnover by his positive, aggressive action(s)
* **BLK:** The average amount of blocks performed by a player per game
  + A block in basketball occurs when a defensive player legally deflects a field goal attempt from an offensive player
* **TOV:** The average amount of turnovers committed by a player per game
  + A turnover in basketball occurs when a player loses possession of the ball to the opposing team before a field goal is attempted
* **PF:** The average amount of personal foul committed by a player per game
  + A personal foul in basketball is a breach of the rules that concerns illegal personal contact with an opponent
* **PTS:** The average amount of points a player scores per game

**Methodology and Application:**

**Multiple Linear Regression:**

For my first statistical analysis, I fit a multiple linear regression model with PTS as my response variable and using all other variables as predictors other than Season, Age, Tm, Lg, and Pos. Here is the summary of the R output:

Call:

lm(formula = career$PTS ~ career$G + career$GS + career$MP +

career$FG + career$FGA + career$FG. + career$X3P + career$X3PA +

career$X3P. + career$X2P + career$X2PA + career$X2P. + career$eFG. +

career$FT + career$FTA + career$FT. + career$ORB + career$DRB +

career$TRB + career$AST + career$STL + career$BLK + career$TOV +

career$PF)

Residuals:

Min 1Q Median 3Q Max

-0.135166 -0.028990 0.002495 0.027440 0.139579

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.458e+00 1.661e+00 -0.878 0.388677

career$G -1.082e-04 1.901e-04 -0.569 0.574573

career$GS 4.716e-05 2.021e-04 0.233 0.817491

career$MP 3.776e-03 1.057e-02 0.357 0.723964

career$FG 1.694e+00 3.745e-01 4.524 0.000139 \*\*\*

career$FGA 9.085e-02 3.760e-01 0.242 0.811128

career$FG. -4.452e+00 5.218e+00 -0.853 0.401967

career$X3P 6.782e-01 3.185e-01 2.130 0.043666 \*

career$X3PA 5.812e-02 3.200e-01 0.182 0.857412

career$X3P. -1.724e+00 8.619e-01 -2.000 0.056948 .

career$X2P 2.400e-01 3.116e-01 0.770 0.448667

career$X2PA -2.678e-02 3.616e-01 -0.074 0.941583

career$X2P. -7.114e+00 3.725e+00 -1.910 0.068157 .

career$eFG. 1.416e+01 4.289e+00 3.302 0.003000 \*\*

career$FT 9.455e-01 2.198e-01 4.301 0.000246 \*\*\*

career$FTA 8.135e-03 1.794e-01 0.045 0.964209

career$FT. 4.810e-01 8.176e-01 0.588 0.561824

career$ORB -4.660e-01 4.742e-01 -0.983 0.335494

career$DRB -4.940e-01 4.816e-01 -1.026 0.315294

career$TRB 4.813e-01 4.797e-01 1.003 0.325670

career$AST -4.510e-03 1.758e-02 -0.257 0.799736

career$STL 4.176e-02 4.415e-02 0.946 0.353651

career$BLK -9.000e-02 1.225e-01 -0.735 0.469755

career$TOV -4.876e-02 6.048e-02 -0.806 0.428067

career$PF 9.948e-02 3.838e-02 2.592 0.015981 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07276 on 24 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9996

F-statistic: 5020 on 24 and 24 DF, p-value: < 2.2e-16

**Variable Selection:**

With the summary of the multiple linear regression model, I manually performed backwards selection with a threshold of a = 0.05. This is my R summary output I ended up with:

Call:

lm(formula = career$PTS ~ career$FG + career$X3P + career$X2P. +

career$eFG. + career$FT + career$ORB + career$DRB + career$TRB +

career$TOV + career$PF)

Residuals:

Min 1Q Median 3Q Max

-0.149071 -0.038708 -0.004267 0.036796 0.155202

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.01907 0.20648 0.092 0.926903

career$FG 2.04386 0.01480 138.127 < 2e-16 \*\*\*

career$X3P 0.89090 0.04026 22.127 < 2e-16 \*\*\*

career$X2P. -5.05681 1.32204 -3.825 0.000473 \*\*\*

career$eFG. 4.62651 1.33905 3.455 0.001368 \*\*

career$FT 0.98601 0.01249 78.966 < 2e-16 \*\*\*

career$ORB -0.74932 0.28117 -2.665 0.011239 \*

career$DRB -0.79555 0.28618 -2.780 0.008409 \*\*

career$TRB 0.78397 0.28561 2.745 0.009191 \*\*

career$TOV -0.06571 0.02910 -2.258 0.029778 \*

career$PF 0.09325 0.02841 3.283 0.002211 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06706 on 38 degrees of freedom

Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997

F-statistic: 1.418e+04 on 10 and 38 DF, p-value: < 2.2e-16

From this backward selection, we see that the resulting model uses FG, X3P, X2P., eFG., FT, ORB, DRB, TRB, TOV, and PF to predict PTS. This is the resulting estimated equation of a line:

PTS = 2.04386(FG) + 0.89090(X3P) – 5.05681(X2P) + 4.62651(eFG.) + 0.98601(FT) – 0.74932(ORB) – 0.79555(DRB) + 0.78397(TRB) – 0.06571(TOV) + 0.09325(PF)

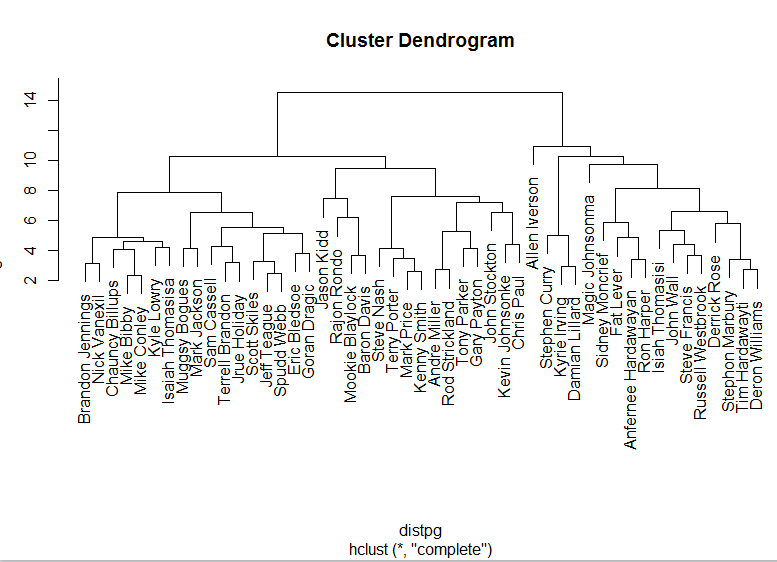
From this equation we see that FG, X3P, eFG., FT, and TRB have a positive coefficient when predicting PTS. This makes sense because the higher the average field goals made, 3-pointers made, average effective field goal percentage, free throws made, and total rebounds collected (all per game averages), the higher your average points per game will be.

We also see that TOV has a negative coefficient. This makes sense because this means the more you give away the ball to the other team, the lower your average points per game will be. We also see 2P. having a very large negative coefficient which implies that the higher your average percentage of shots made inside the 3-point line, the lower your PTS statistic. This means point guards who attempt more shots inside the 3-point line tend to have a higher PTS. The more shots a point guard attempts, the lower his 2P. statistic will be.

There are a few coefficients here that contradict the game of basketball or need further explanation. Most notably, PF having a positive coefficient. This is implying that the more you foul a person the higher your PTS statistic will be. This is not true because there is a limit on how many PFs you can commit. Once you commit your 6th PF, you do not get to play for the rest of the game. Also, ORB and DRB have a negative coefficient but TRB (a summation of ORB and DRB) has a positive coefficient. This does not seem to make sense since TRB is a summation of ORB and DRB.

**Hierarchical Clustering:**

I have performed single, average, and complete hierarchical clustering on the career dataset. Scaling was required to calculate the distances for each player in order to perform the clustering algorithm. Here are the resulting plots:



Out of the three cluster dendrograms, the cluster dendrogram for complete linkage gives the most clear-ish result since you can tell there are about 2 or 3 major groups. One thing that jumps out at me the most is the fact that Allen Iverson and Magic Johnson are the highest on this dendrogram. They connect with the other groups not only as a leaf, but at a much higher distance than the other players. This observation shows that they are substantially better point guards than anyone else, which holds true to their real-life status. They are hall-of-famers in the NBA and always show up on sports analyst’s, reporter’s, and legend’s top ten list of best basketball players of all time.

The first group that seems the most prominent to me is the group that expands to the right of the dendrogram (contains players like Allen Iverson, Magic Johnson, Stephen Curry, etc.). These players who are all grouped together are known, or have been known, as elite point guards for their incredible scoring ability. All these players do exceptionally well in arguably the most important aspect of basketball, scoring. This is evident since their PTS statistic is well above the average for this dataset.

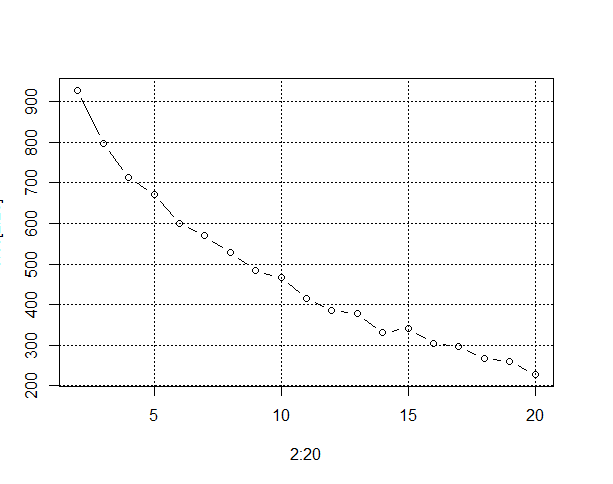
The second group that seems prominent ranges from Jason Kidd to Chris Paul (left to right). These players are known as the “true point guards”. A point guard’s role in basketball is to set up his teammates and find the open man. This can be shown through the AST statistic, which is very high (well above the average) for all of these point guards.

The last group on the left is considered the “average joes”. These point guards have all-around average numbers in every statistic.

**K-Means Clustering:**

The hierarchical clustering data gave us a good look at how many groups there might be within our dataset through a dendrogram. But if we want R to give us a specific amount of clusters, the K-means clustering algorithm is a better method. K-means clustering finds the “best” way to group all the observations in a certain amount of clusters.

After scaling all my numerical data, I performed a for loop calculation which collected a quantity called “total within-cluster sum of squares” for each number of clusters from 2 to 20. The scree plot below is a result of that calculation.



As you can see, the scree plot is not very helpful at giving us a good idea on how many clusters we should pick for our k value. Although, we do see an elbow at 5 (sort of) since all the points after seem to be on a very steady, linear decrease. So I picked my k value to be 5. After picking the k value, a cluster, centres, withinss, and size calculation were performed. I also got to see which players were in which cluster. Here were the reults:

$`1`

[1] "Mookie Blaylock" "Baron Davis" "Anfernee Hardawayan"

[4] "Ron Harper" "Jason Kidd" "Fat Lever"

[7] "Rajon Rondo"

$`2`

[1] "Steve Francis" "Allen Iverson" "Magic Johnsonma" "Isiah Thomasisi"

[5] "John Wall" "Russell Westbrook"

$`3`

[1] "Mike Bibby" "Chauncy Billups" "Eric Bledsoe" "Muggsy Bogues"

[5] "Terrell Brandon" "Mike Conley" "Goran Dragic" "Jrue Holiday"

[9] "Mark Jackson" "Kyle Lowry" "Steve Nash" "Terry Porter"

[13] "Mark Price" "Scott Skiles" "Kenny Smith" "Jeff Teague"

[17] "Isaiah Thomasisa" "Nick Vanexil" "Spudd Webb"

$`4`

[1] "Sam Cassell" "Kevin Johnsonke" "Andre Miller" "Sidney Moncrief"

[5] "Tony Parker" "Gary Payton" "John Stockton" "Rod Strickland"

$`5`

[1] "Stephen Curry" "Tim Hardawayti" "Kyrie Irving" "Brandon Jennings"

[5] "Damian Lillard" "Stephon Marbury" "Chris Paul" "Derrick Rose"

[9] "Deron Williams"

The one cluster that stands out to me is cluster 2. This cluster has elite-level point guards of the past and the present. These players have very high averages in PTS and AST statistics. There is an argument to be made for Stephen Curry not being on this list since in the current season he is playing like the best player in the world. But this could be due to him being in cluster 5 (the “scorers” or “high-volume shooters” of the league) which makes sense because that is one thing he excels in. Cluster 4 could be classified as “true” point guards. These point guards play within their role of setting teammates up and passing the ball first. Cluster 1 could be classified as the assisting masters. These players always average a high amount of assists. Although, I would replace Ron Harper with Steve Nash because Ron Harper is a very average point guard and belongs in the average point guards cluster (cluster 3).

**Clustering Conclusion:**

Both Hierarchical and K-means clustering work relatively well when grouping data. Although, I do prefer the Hierarchical clustering more. First of all, it requires less code and no need for standardizing the data beforehand. Secondly, it provides a much more interesting representation of groups in the form of a dendrogram. Lastly, although both methods do have some misclassifications, In my experience with my data, I found Hierarchical Clustering to have less misclassifications.

**Principal Component Analysis:**

After running a principal component analysis, here were the results of the summary, the biplot, an the loadings of the first 2 principal components:

**Summary:**

Importance of com: PC1 PC2 PC3 PC4 PC5 PC6 PC7

Standard dev: 2.90581 2.075306 1.817184 1.532399 1.359524 0.977152 0.9701234

Prop. of Var.: 0.33775 0.172280 0.132090 0.093930 0.073930 0.038190 0.0376500

Sum Of Prop.: 0.33775 0.510030 0.642110 0.736040 0.809970 0.848170 0.8858100

PC8 PC9 PC10 PC11 PC12 PC13

Standard dev: 0.894691 0.753745 0.7230171 0.5650194 0.4768577 0.4029087

Prop. of Var.: 0.032020 0.022730 0.0209100 0.0127700 0.0091000 0.0064900

Sum of Prop.: 0.917830 0.940560 0.9614700 0.9742400 0.9833300 0.9898300

PC14 PC15 PC16 PC17 PC18 PC19

Standard dev: 0.3564915 0.2462923 0.1730881 0.1632772 0.06596116 0.05107726

Prop. of Var.: 0.0050800 0.0024300 0.0012000 0.0010700 0.00017000 0.00010000

Sum of Prop.: 0.9949100 0.9973400 0.9985300 0.9996000 0.99977000 0.99988000

PC20 PC21 PC22 PC23 PC24

Standard dev: 0.03764327 0.02834823 0.01955087 0.01636758 0.010884

Prop. Of Var.: 0.00006000 0.00003000 0.00002000 0.00001000 0.000000

Sum of Prop.: 0.99994000 0.99997000 0.99998000 0.99999000 1.000000

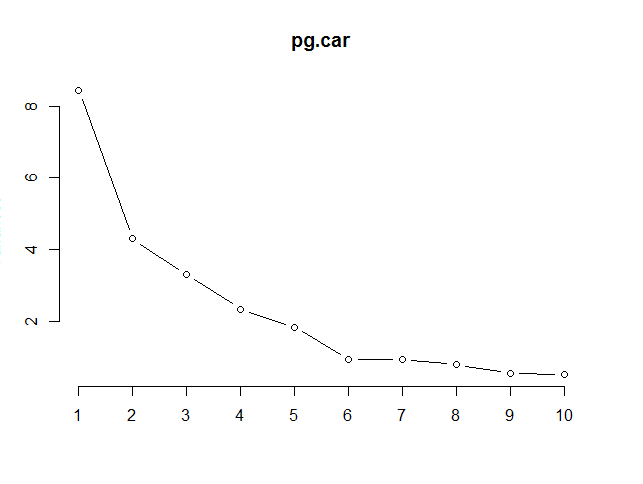
PC25

Standard deviation 0.006938237

Proportion of Variance 0.000000000

Cumulative Proportion 1.000000000

**Biplot:**



**Loadings:**

PC1 PC2

G -0.05 0.24

GS 0.03 0.19

MP 0.29 -0.07

FG 0.30 -0.16

FGA 0.30 -0.20

FG. -0.01 0.17

X3P 0.02 -0.42

X3PA 0.04 -0.41

X3P. -0.10 -0.34

X2P 0.30 0.05

X2PA 0.31 0.02

X2P. -0.02 0.11

eFG. -0.08 -0.13

FT 0.26 -0.09

FTA 0.28 -0.04

FT. -0.05 -0.25

ORB 0.17 0.29

DRB 0.22 0.13

TRB 0.22 0.20

AST 0.12 0.12

STL 0.18 0.18

BLK 0.15 0.08

TOV 0.28 -0.01

PF 0.15 0.05

PTS 0.30 -0.21

As we can see from the loadings, the X3P. statistic has the lowest score in the first principal components and the X2PA statistic has the highest score. This means that your 3-point percentage has the biggest negative coefficient in determining your point guard score but the amount of shots attempted inside the 3-point line has the biggest positive coefficient. This is due to the fact of a point guard’s primary scoring objective of driving to the basket. Point guards are the quickest and fastest players on the court, so their best bet on scoring would be to blow past their defender and get to the basket to get an easy layup. It is not only the easiest shot in basketball, but you could also absorb enough contact from the defender for the referee to call a foul and earn some free throws.

Although 3-pointers are worth more points, they are also a much harder shot to take due to the distance of the shot. Therefore, all the point guards X3P. statistic will be quite low (with the exception of players like Steve Nash and Stephen Curry) which is a good thing. We have previously stated that the more shots you take the lower your percentage gets but the higher amount of points you receive. So the lower your X3P. statistic, the better your point guard score is.

Kaiser Criterion:

I would only be able to keep PC1, PC2, PC3, PC4, and PC5 through the Kaiser criterion because they are the only PCs in my summary plot whose square root of the deviation (which is the variance and ultimately the eigenvalue) is greater than or equal to 1.

Retaining 90% of Variance:

I would only be able to keep PC1, PC2, PC3, PC4, PC5, PC6, and PC7 because their cumulative proportions retain at least 90% of the variance in the data.

Scree Plot:

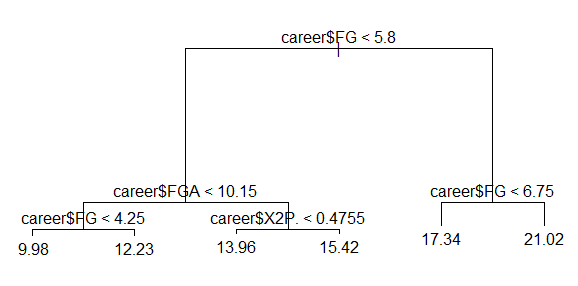
Based on the scree plot, I would keep PC1, PC2, PC3, PC4, PC5, and PC6 since the elbow appears to be at PC6.

**Neural Networks:**

**Random Forrests/Bagging/Regression Trees:**

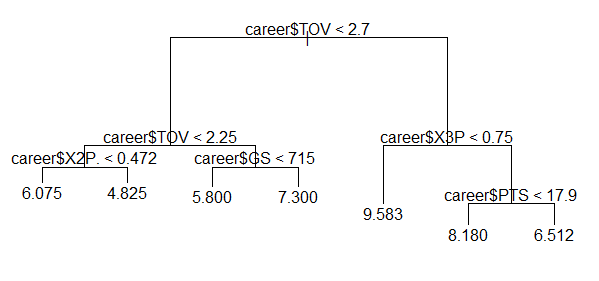
Regression Tree:

I created a regression tree for the point guards using the PTS statistic. Here were the results:



This decision tree brings obvious observations. It implies the more field goals you make, the more points you get. It also implies the more shots you take, the more points you get. Also, if your percentage of shots made inside the 3-point line is high, you will have more points.

I also created a decision tree for AST. Here were the results:



This decision tree had much more interesting results. First of all, this tree implies that the greater amount of turnovers you have, the better your chances are for getting more assists. This could be a result of the point guard’s attempts to set up teammates. An assist is usually when a point guard passes the ball and the player scores right away, so this could mean the point guard attempts more passes which in turn leads to more turnovers because of the opposing team’s defense or because the pass was not on target. This could also occur because the point guard controls the ball the most. The more he holds on to the ball, the more susceptible he is to turning the ball over to the other team, but the higher the chance is to get the assist because the he dictates where the ball goes the majority of the time.

Another interesting observation is that the more 3-pointers and total points you have, the lower amount of assists you collect. This is because the point guard is too focused on scoring during the game and not focused on setting up his teammates. In fact, this is the case with many scoring point guards in the NBA. They become a “shoot-first” point guard instead of a “pass-first” point guard.

Random Forests/Bagging:

I also ran random forests and bagging for AST from my career dataset, here were my results for the summary and the resulting plot:

**Summary:**

Call:

randomForest(formula = AST ~ G + GS + MP + FG + FGA + FG. + X3P + X3PA + X3P. + X2P + X2PA + X2P. + eFG. + FT + FTA + FT. + ORB + DRB + TRB + PTS + STL + BLK + TOV + PF, data = career, importance = TRUE)

Type of random forest: regression

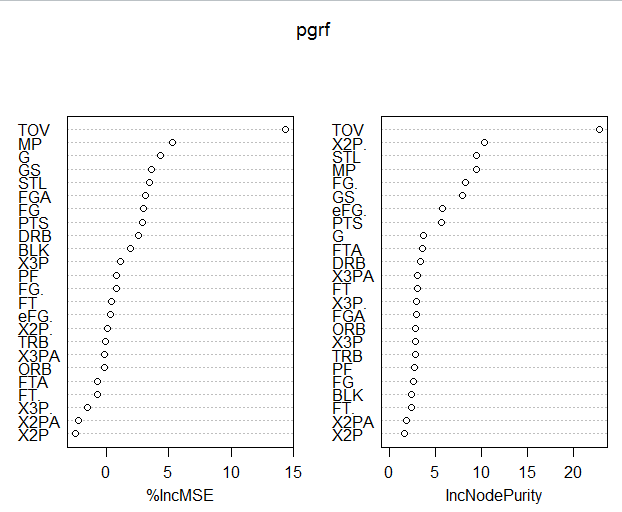
Number of trees: 500

No. of variables tried at each split: 8

Mean of squared residuals: 2.1301177

% Var explained: 21.7

**Plot:**



As we can see from the plot, the most important variable for predicting AST is TOV. This observation solidifies our previous claim on how important turnovers are when analyzing assists: The more turnovers you have, the more assists you have.