post01-marc-mansour

Visualizing Basketball Data with Violin Plots in ggplot2

By Marc Mansour

Introduction

Background

In 2005, after years of research, Hadley Wickham released a package within the R language that embodied the principles of Leland Wilkinson's *Grammar of Graphics*. ggplot2 enabled R users to create cleaner and more sophisticated data visualizations than what could be produced with the base R functions (e.g. plot(), hist(), barplot()). In addition, ggplot2 follows a specific 'grammar', like a sentence, that allows users to easily add new elements to each visualization. These elements include anything from basic labels to various faceting features. Throughout this piece, we will expand upon our knowledge of the ggplot2 package and its capabilities by working with a new kind of visualization: violin plots.

Violin plots are very similar to boxplots in that they give viewers a better idea of the distribution of a given variable. However, boxplots leave us unsure about how densely distributed a variable is at various points. In order to solve this problem, violin plots essentially combine the elements of a boxplot and density plot to give us a more holistic understanding of the variables in question. Violin plots enable data analysts to more clearly understand how the distribution of a given variable changes at different points.

In this piece we will create a series of violin plots to help us better understand the different positions on a basketball team and their unique roles on the court. By analyzing the statistical contributions of each position across a standard box score (i.e. points, rebounds, assits, steals, blocks), we will bring clarity to a) the actual roles of each of the five positions in basketball and b) the degree to which these roles vary in each position. We will begin by creating a simple boxplot to better illustrate the enhanced capabilities of the violin plots we will create after. There will be five violin plots included in this piece, one for each of the major statistical categories in a standard box score, as previously mentioned. Each of these violin plots will treat positions as a factor to help us isolate the different roles of each position. In addition, violin plots will allow us to figure out the extent to which certain players are outliers in their position groups.

Installing ggplot2

```r
# Install the ggplot2 package, if you don't already have it
# install.packages('ggplot2')

# Load the ggplot2 package to use in your current R session
library(ggplot2)
```

The Data

As an avid fan of the NBA, I really enjoy analyzing and visualizing basketball data. As a result, I will use a dataset from the course's Github repository; however, I have made some changes to the dataset for my specific purposes. These changes are outlined in an R Script file titled 'make-per-game-data.R' which can be found in the 'code' folder of 'post01'. In this script file, I create a variables for total points ('PTS') and total rebounds ('REB'). Then I adjust all statistics to a 'per game' scale by dividing each statistic included by the number of games each player participated in. Finally, I isolated just the five statistics I wanted as well as some variables I deemed necessary, such as the player's names, team, and position. Please reference my script file to see how I made these specific changes. I have also included a data dictionary titled 'nba2017-per-game-dicitonary.Rmd' and a text file with the summary of the per game dataset for additional information on each variable. Below I have inserted some code to load the csv I created in my script file to this current R session.

```r
# Load per game NBA player statistics from 'nba2017-per-game.csv'
nba_per_game <- read.csv("../data/nba2017-per-game.csv",
                         stringsAsFactors = FALSE)
```

Creating Violin Plots with ggplot2

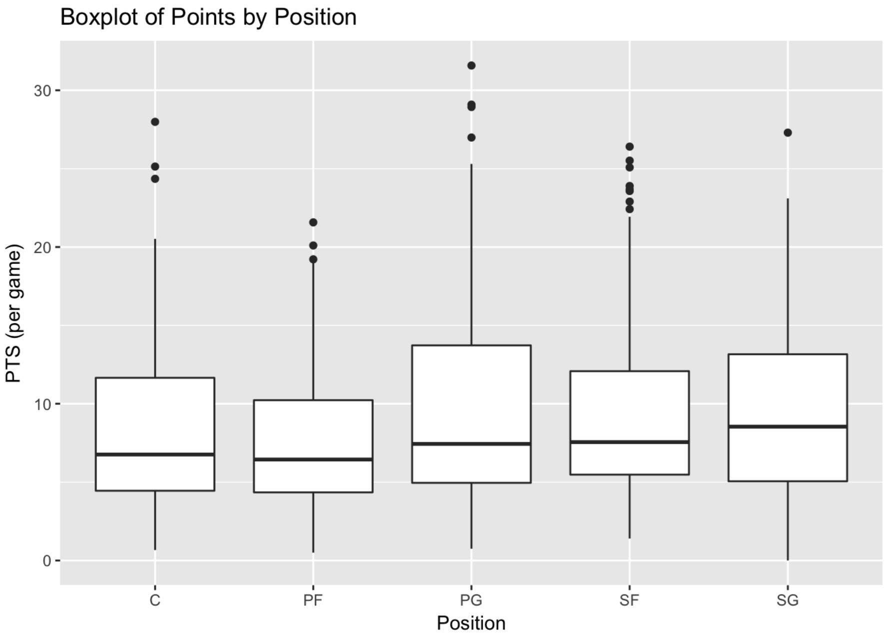
An Introductory Boxplot

We will begin with a very simple boxplot just to get us started. In this boxplot, we will investigate the differences in points per game by position. This should allow us to better understand which positions are most heavily burdened with scoring responsibilities. Ultimately, however, we will use this boxplot to identify the limitations of boxplots.

```
# Implement the base ggplot function and specify the desired variables
ggplot(nba_per_game, aes(factor(Position), PTS)) +

# Specify that we want a boxplot
  geom_boxplot() +

# Add a title and label the axes
  ggtitle('Boxplot of Points by Position') +
  xlab('Position') +
  ylab('PTS (per game)')
```



From the above plot, we initially notice that there seemingly isn’t too much of a difference among the points per game scored across the five positions. Point guards seem to have a slight edge over the rest of the positions as their maximum values are the highest among all positions. However, points guards also seem to have the most variance in scoring as they also have the largest interquartile range. This prompts us to wonder how values within the IQR are distributed. In addition, we might wonder how many point guards score over 20 point per game. Unfortunately, boxplots cannot properly give us an accurate representation of this distribution. Consequently, we will move on to our series of violin plots which will give us a more holistic representation of the distribution of each variable identified in the basic box score by position.

Violin Plots

Violin plots are a more effective means for visualizing continuous distributions. As previously mentioned, violin plots give us a better idea of the density of a given variables’ distribution within a dataset. This provides viewers with a more holistic visualization of the data. We will use five violin plots here in order to identify the specific roles of certain positions and extract some more specific insights on how statistical accomplishments vary among players of the same position to more accurarely interpret outliers.

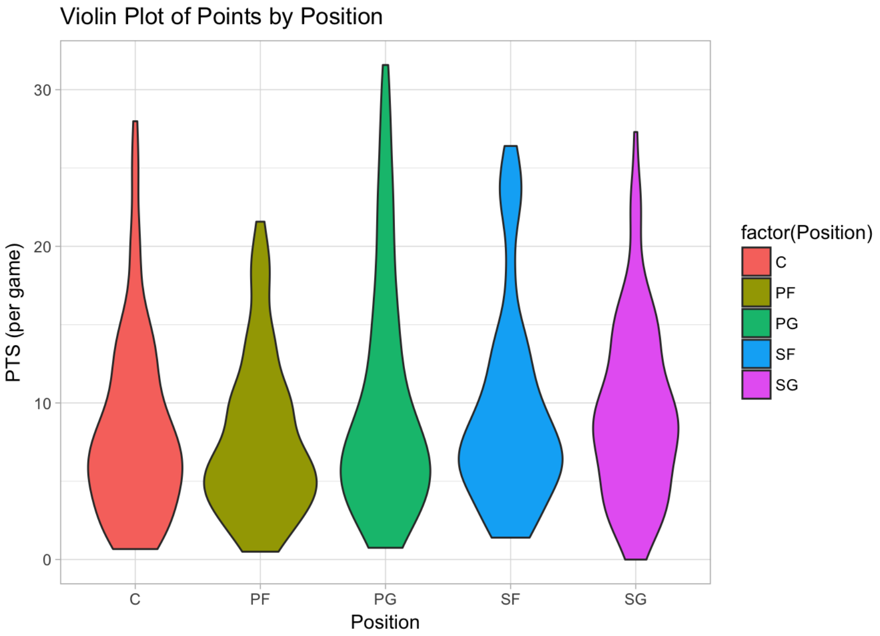
Points Violin Plot

```
# Implement the base function ggplot()
ggplot(nba_per_game, aes(factor(Position), PTS)) +

# Specify that we want a violin plot
  geom_violin(aes(fill = factor(Position))) +

# Add a title and labels
  ggtitle('Violin Plot of Points by Position') +
  xlab('Position') +
  ylab('PTS (per game)') +

# Add a theme for aesthetic purposes
  theme_light()
```



From the above violin plot, we get a much clearer representation of the offensive roles of each position, specificaly with regards to scoring. Once again, we note that point guards seem to score the most points among NBA players. This is likely due to two factors. First, point guards are the ones who are tasked with running the offense. As a result, point guards generally spend the most time with the ball in their hands. Therefore, point guards have the most opportunities to score. Additionally, the NBA's talent pool is currently filled with an abundance of highly talented point guards. This is why we notice the long tail at the top of the PG distribution which represents NBA All Stars such as Russell Westbrook, James Harden, and Stephen Curry, among others. These outliers make up for the large collection of point guards who score just about 5 points per game. This violin plot also shows us that there is a group of small forwards who score at an elite level. This subset of small forwards is represented by the cluster of players at the top of the violin plot. These small forwards likely include All Stars like LeBron James, Kevin Durant, Kawhi Leonard, and Giannis Antetokounmpo. Power forwards, on the other hand, are the least involved when it comes to scoring. We immediately notice that there are very few power forwards who score more than 20 points per game, which is benchmark for elite scorers in the NBA. In addition, power forwards are most densely distributed at the 5 points per game mark, which is pretty lackluster. Moreover, there are also a few centers that score the basketball at an elite level; however, shooting guards seem to be more effective at scoring overall as we notice that a larger number of shooting guards score over 10 points per game.

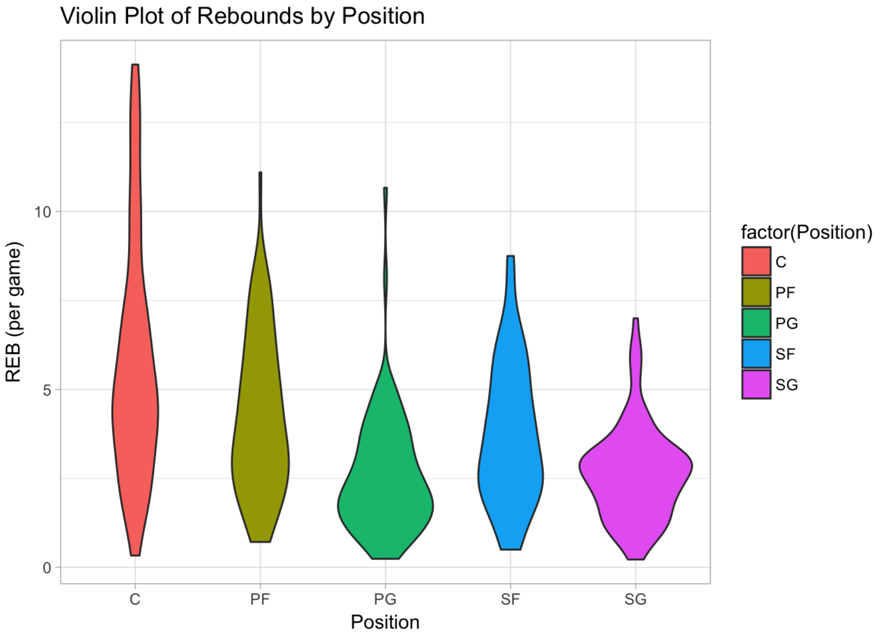
Rebounds Violin Plot

```r
# Implement the base function ggplot()
ggplot(nba_per_game, aes(factor(Position), REB)) +

# Specify that we want a violin plot
  geom_violin(aes(fill = factor(Position))) +

# Add a title and labels
  ggtitle('Violin Plot of Rebounds by Position') +
  xlab('Position') +
  ylab('REB (per game)') +

# Add a theme for aesthetic purposes
  theme_light()
```



This violin plot of rebounds by position gives us a nice visual representation of the hierarchy of rebounders by position. Centers clearly grab the most rebounds, followed by power forwards, small forwards, shooting guards, and points guards in that order. The reasoning behind this ordering is very straightforward as the number of rebounds collected clearly correlates to the typical height of each position. Centers are the usually the tallest players on each team, followed by power forwards, small forwards, shooting guards, and point guards. However, this violin plot illustrates a notable exception to this rule as we can clearly identify a small subset of point guards who average approximately 10 rebounds per game, which is considered exceptional at any position. This tail represents two point guards in particular: Russell Westbrook and James Harden. Although the largest collection of point guards averaged just about 2 rebounds per game, these point guards were able to register more rebounds because their teams specifically used their big men (i.e. centers and power forwards) to box out opposing players and clear space for Westbrook and Harden to get rebounds. This is an emerging trend within the NBA that enables point guards to get the ball and start running the offense immediately. This recent trend and its' clear deviation from the normal distribution of rebounds for point guards would not be as clear in a boxplot. The above violin plot highlights how rare this scheme is with the large cluster of point guards averaging approximately 2 rebounds a game.

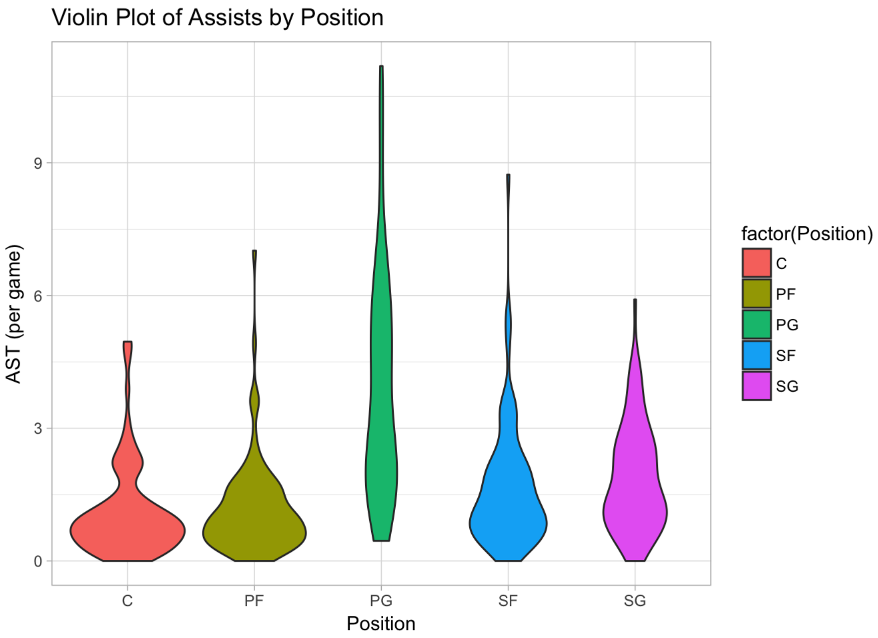
Assists Violin Plot

```
# Implement the base function ggplot()
ggplot(nba_per_game, aes(factor(Position), AST)) +

# Specify that we want a violin plot
  geom_violin(aes(fill = factor(Position))) +

# Add a title and labels
  ggtitle('Violin Plot of Assists by Position') +
  xlab('Position') +
  ylab('AST (per game)') +

# Add a theme for aesthetic purposes
  theme_light()
```



Next, we will analyze the passing roles of each position as shown by the number of assists recorded per game. As previously alluded to, point guards are tasked with running the offense and therefore have a clear lead over all other positions in terms of assists. A very limited number of players at the remaining positions record even just 6 assists per game whereas there are a handful of point guards who average over 10 assists per game (i.e. Russell Westbrook, John Wall, and James Harden). On average, shooting guards tend to be the next most involved passers; however, the density of this distribution shows us that there are not many elite passers at the shooting guard spot as none of them average over 6 assists per game. Nevertheless, there are a good number of shooting guards who average right around 3 assists per game or more, which is a significant contribution. On the other hand, there seem to be a handful of small forwards who pass the ball at an elite level. This subset of small forwards represents players who play the 'point forward' role. These players are small forwards who are often tasked with bringing the ball up for their team as if they were the point guard. Examples of point forwards include players like LeBron James and Giannis Antetokounmpo. Once again, violin plots help illustrate how rare it is for a small forward to pass at such a high level, which further emphasizes the great offensive impact these players have. Finally, centers and power forwards are very rarely tasked with handling the ball and when they do get the ball, it's likely to initiate offense and score right away instead of generating scoring opportunities for other players like point guards do.

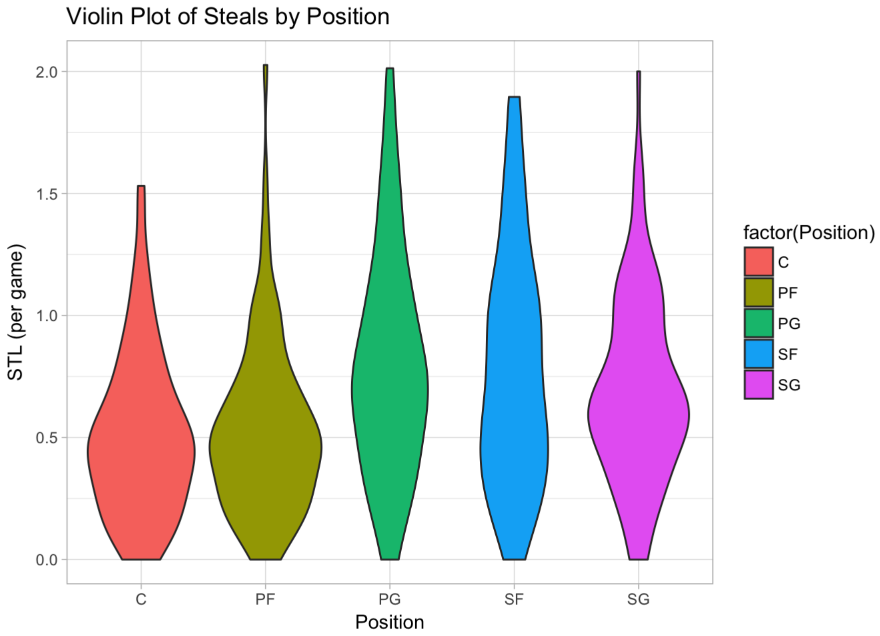
Steals Violin Plot

```
# Implement the base function ggplot()
ggplot(nba_per_game, aes(factor(Position), STL)) +

# Specify that we want a violin plot
  geom_violin(aes(fill = factor(Position))) +

# Add a title and labels
  ggtitle('Violin Plot of Steals by Position') +
  xlab('Position') +
  ylab('STL (per game)') +

# Add a theme for aesthetic purposes
  theme_light()
```



Now we will transition to the purely defensive roles of each position, starting with steals. A steal occurs when a player takes possession of the ball from an opposing player. Once again, we note that positions that are typically tasked with handling the ball more (i.e. point guards, shooting guards, and small forwards) are the ones who generate the most steals. This intuitively makes sense since more time with the ball enables players to generate more steals from opposing players. On the other hand, big men (i.e. centers and power forwards) record the least number of steals because, once again, they handle the ball the least. However, this violin plot uncovers yet another recent trend among NBA players in the small tail of power forwards who generate a comparable number of steals as other positions. This tail represents the growing group of power forwards who are tasked with guarding ball handlers. This group of power forwards are led by last year's Defensive Player of the year, Draymond Green of the Golden State Warriors, who actually led the league in steals per game. The uniqueness of this trend is further highlighted by the large cluster of power forwards who only generate approximately 0.5 steals per game, whereas Green averaged 2.0.

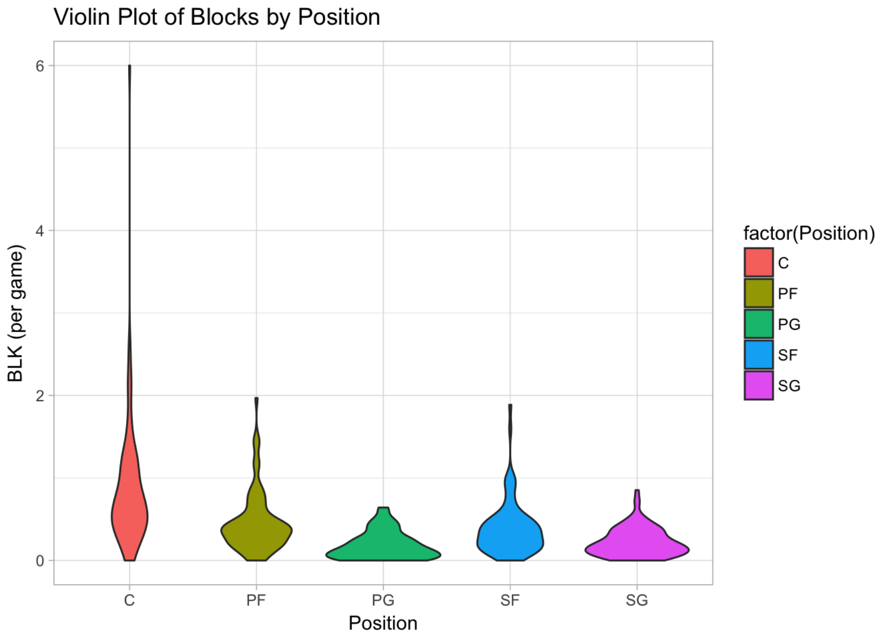
Blocks Violin Plot

```
# Implement the base function ggplot()
ggplot(nba_per_game, aes(factor(Position), BLK)) +

# Specify that we want a violin plot
  geom_violin(aes(fill = factor(Position))) +

# Add a title and labels
  ggtitle('Violin Plot of Blocks by Position') +
  xlab('Position') +
  ylab('BLK (per game)') +

# Add a theme for aesthetic purposes
  theme_light()
```



Finally, we will look at the shot blocking roles of each position. Once again, we find that centers are the most involved in this regard. The reasoning for this is very similar to the reasoning behind why centers are such active rebounders: height. Centers are the tallest players on the court and therefore are the most physically enabled players to block shots. Centers also are specifically tasked with defending the rim as most defensive schemes place centers right in front of the basket to fulfill this role. Point guards and shooting guards, on the other hand, are simply not active shot blockers since they typically defend opponents on the perimeter of the court. In addition, guards are traditionally the shortest players on the court. Power forwards and small forwards are the next most active, respectively. Although power forwards are typically the more productive shot blockers, there are two exceptionally gifted shot blockers at the small forward position: Kevin Durant and Giannis Antetokounmpo. Both of these players are gifted with elite length, standing approximately 7 feet tall. Due to this length, Durant and Antetokounmpo are able to block significantly more shots than their shorter counterparts at the small forward position and even some shorter power forwards.

Looking Forward

All in all, the use of violin plots sorted by position gave us a clearer look at the roles of each position in basketball. As the NBA continues to encourage more versatility, we will see more players who are able to contribute in more ways than other players at the same position. Violin plots give us a better visual representation of this trend towards "positionless basketball" in the NBA. By including density, violin plots allow us to better understand how a given variable is distributed. In this case, displaying density enabled us to comprehend the full extent to which players deviate from the standard capabilities of their peers. Violin plots can of course be extended to other datasets as well, giving data analysts a clearer visual representation of a variables' distribution and allowing us to more accurately understand them.

References

- *ggplot2: Elegant Graphics for Data Analysis* by Hadley Wickham
- ggplot2 Cheat Sheet
- Violin Plots in ggplot2
- "Violin Plots: A Box Plot-Density Trace Synergism" by Hintze & Nelson
- Video on Boxplots and Violin Plots in ggplot2
- An alternative 'vioplot' package
- Wikipedia page explaining the roles of each position in basketball
- NBA 2017 Player Statistics from Stat133 Github Repository

Additional Examples

- STHDA Tutorial on violin plots in ggplot2
- Another tutorial for violin plots in ggplot2
- The R Graph Gallery #95: Violin Plots with ggplot2