NBA Scoring Correlation To Different Statistics Case Study

by Maxym Voytovych

Date and Location: September 5 2022, Lviv, Ukraine

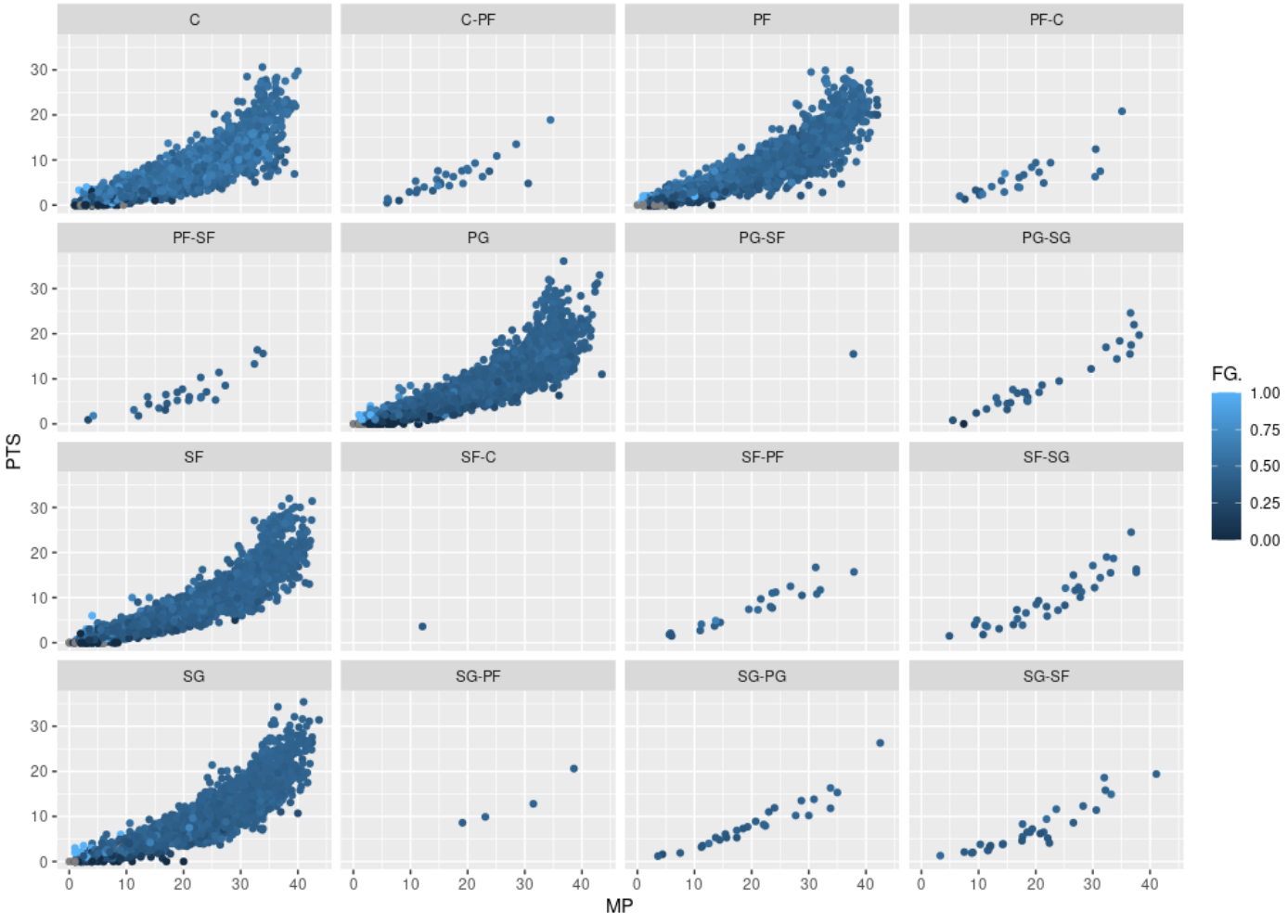
Main objective: Finding a correlation between different basketball statistics and points scored per game.
Data: Data is taken from free public source, you can check it yourself after opening this link [NBA Player Stats - dataset by etocco | data.world]
Analysis tactics: I will be completing a correlation analysis on this data set and also will present my findings with some dot plots to visually see results. This case study will be divided into a few hypotheses.

First hypothesis

Firstly, let's use some common sense. For a player to score a decent amount of points he has to play a decent amount of minutes per game. Sounds pretty obvious if you ask me, but it is too soon to make such a conclusion without doing some research. So, I decided to find a correlation coefficient (a coefficient that shows how much two variables relate to each other) on columns MP(minutes played) and PTS(points scored per game). The outcome is this:

```
> print(cor.test(df$MP, df$PTS))

        Pearson's product-moment correlation

data:  df$MP and df$PTS
t = 239.04, df = 14571, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8892915 0.8958910
sample estimates:
      cor
0.8926391
```

Correlation coefficient in this case equals 0.8926391, which shows us that there definitely is quite a strong correlation between those two variables. Lets see some visuals to better understand this correlation.



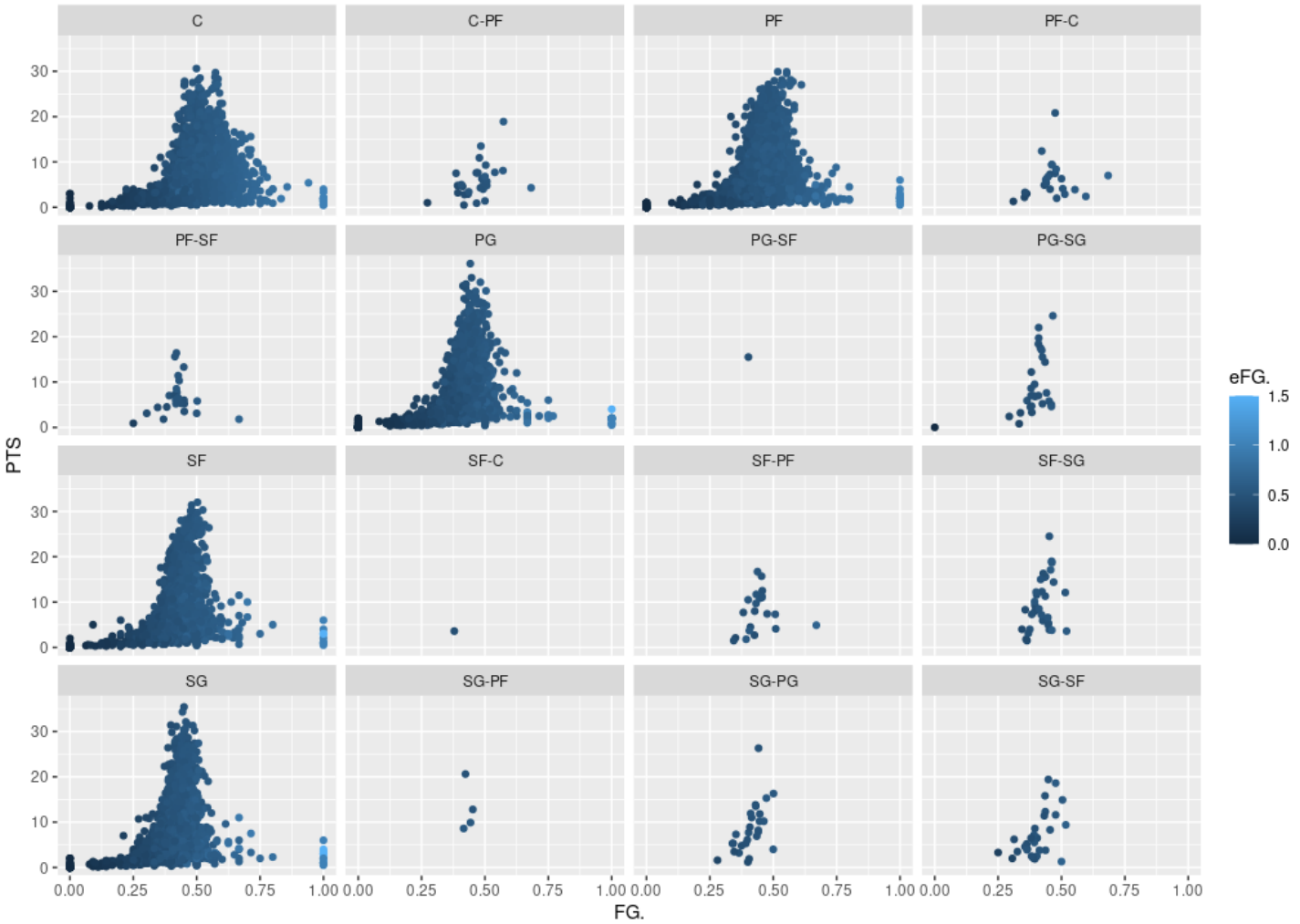
This plot is divided into more subplots by position played. We can now clearly see that the more minutes a player plays - the more points he can score, despite what position this average player plays.

Second hypothesis

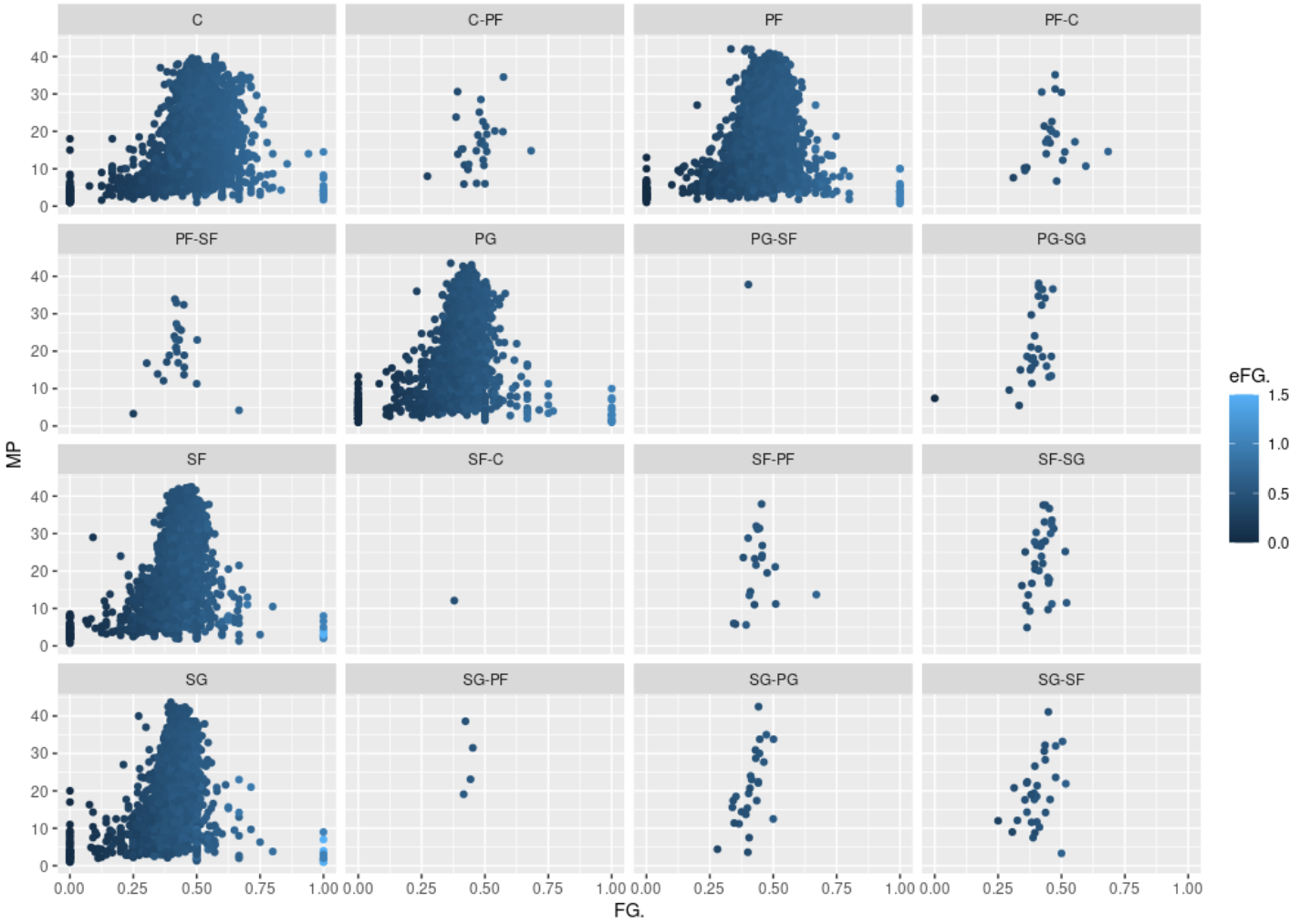
So, the second theory is that to score a lot of points - a player has to shoot the ball efficiently. That, like our first hypothesis, sounds pretty obvious, let's check for correlation.

```
> print(cor.test(df$FG., df$PTS))

        Pearson's product-moment correlation

data:  df$FG. and df$PTS
t = 32.448, df = 14483, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2450778 0.2754416
sample estimates:
      cor
0.2603241
```

As we see, the correlation is weak, let's look at some visuals to see a bigger picture.



We can see, that players with FG.(field goal percentage) that is close to 0.50 tend to score more than players who shoot with better field goal percentage. That seems a bit strange, why players that have FG. greater than 0.60 score less than their colleges with FG. 0.50 or close to that. Let's now look at a plot that may make things more clear to us. This plot will show us the relation between field goal percentage and minutes played.



If we look closely, we can see that players who have FG. greater than 0.60 mostly play less minutes, compared to those who have FG. closer to 0.50. And as we know from our first hypothesis, players score more points when they play more minutes. That is why a higher field goal percentage does not guarantee a lot of points scored per player.

Conclusions

So, after checking our first hypothesis we now know that for a player to score a lot of points per game he needs to play a lot of minutes. We also know, after checking our second hypothesis, that having a high field goal percentage does not guarantee that a player will score a lot of points, because such players tend to not play as much. There can be some explanations for such phenomena, but to know the true answer we will have to consult with a specialist in this field. But for now, thank you for your attention, hope that those insights were interesting for you!