SpringFinalProject

May 21, 2023

#

Spring Final Project

[1]:
```
# Load the CourseKata library
suppressPackageStartupMessages({
    library(coursekata)
})
```

0.1 Pick a Data Set

Go here to get the list of data sets to choose from. Take a look at them, choose one that interests you, and answer the questions below.

[2]:
```
draft <- subset(nba_draft_2015, draft_year == 2015)
draft$proj_spm <- round(draft$projected_spm, 5)
draft$superstar <- round(draft$superstar, 5)
draft$starter <- round(draft$starter, 5)
draft$role_player <- round(draft$role_player, 5)
draft$bust <- round(draft$bust, 5)
```

For this data analysis report, I used the `nba_draft_2015` data frame because I like watching basketball. The data was collected by FiveThirtyEight, which is a website that focuses on statistical analysis of politics, economics, and sports in the US. They collected this data to analyze it and make it understandable for their readers.

The data frame contains data on the top projected college players from the 2001 to 2015 NBA drafts. I chose to analyze only those from the 2015 draft class, of which there are 77 top projected players. The variables in the data frame are the players' names, positions, draft year (2015), projected statistical plus/minus (SPM)*, and probability of becoming either a superstar, starter, role player, or bust in the NBA**.

- * statistical plus/minus (SPM): an estimate of the player's contribution to the team's point differential per 100 possessions, using his box score stats as inputs
- projected SPM: projected statistical plus/minus per 100 possessions by his team
- ** categories are defined as follows:
- superstar: the best players in the league (about 1 per class, SPM >= +3.3)
- starter: players that are good enough to start on a team (about 10 per class, SPM >= +0.5)

- role player: players that are good enough to be see regular play time (about 25 per class, SPM >= -1.4)
- bust: everyone else (SPM < -1.4)

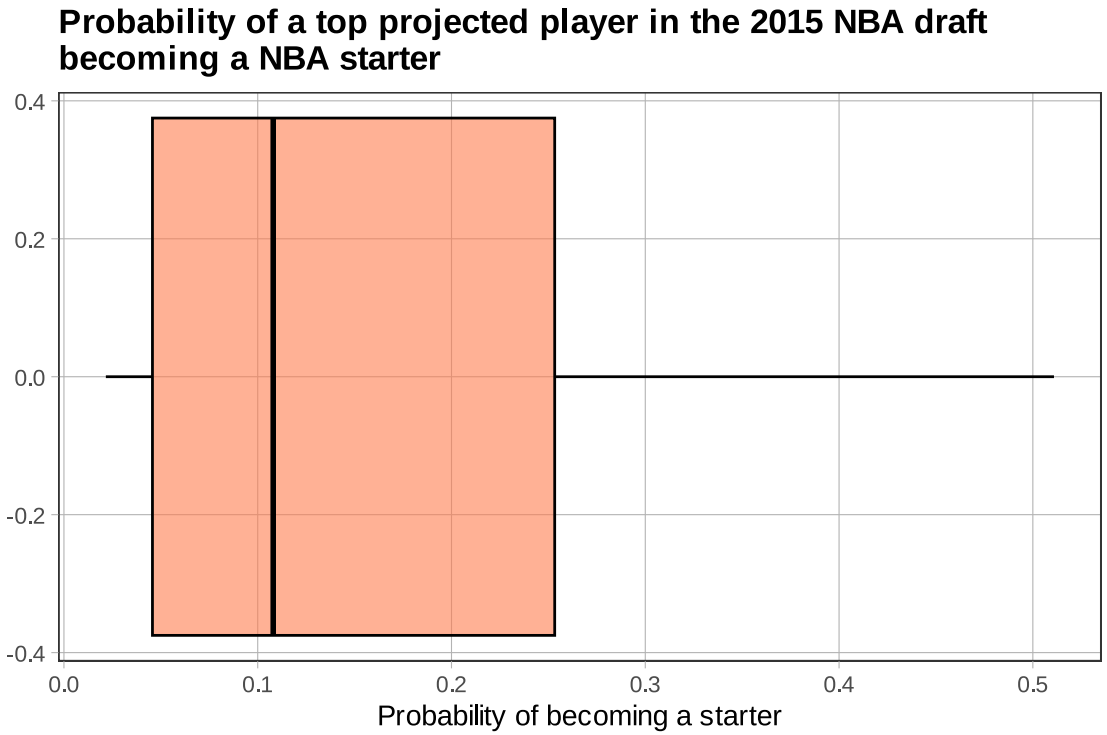
[3]:
```
head(select(draft, player, position, proj_spm, superstar, starter, role_player,␣
↪bust), 10)
```

A tibble: 10 × 7

player <chr>	position <chr>	proj_spm <dbl>	superstar <dbl>	starter <dbl>	role_player <dbl>	bust <dbl>
Karl-Anthony Towns	C	1.03061	0.13477	0.42718	0.16308	0.27497
Justise Winslow	SF	0.87533	0.08353	0.51090	0.17677	0.22880
Stanley Johnson	SF	0.67949	0.06780	0.42373	0.27850	0.22997
Jahlil Okafor	C	0.52166	0.05872	0.40990	0.23553	0.29585
D'Angelo Russell	PG	0.51197	0.15203	0.34228	0.09658	0.40910
Dakari Johnson	C	0.49179	0.02134	0.36754	0.41757	0.19354
Devin Booker	SG	0.47258	0.07337	0.32447	0.39017	0.21200
Willie Cauley-Stein	C	0.35117	0.04711	0.40599	0.24319	0.30371
Rondae Hollis-Jefferson	SF	0.31191	0.01459	0.36853	0.39248	0.22440
Trey Lyles	PF	0.26751	0.02239	0.35133	0.40305	0.22323

In my report, I explored the variable `starter`, which contains the probability of becoming a starting-caliber player in the NBA for each player in the data frame. The distribution of `starter` has been visualized in the boxplot and histogram below:

[4]:
```
gf_boxplot(~starter, data = draft, fill = "coral")%>%
gf_labs(title = "Probability of a top projected player in the 2015 NBA␣
↪draft\nbecoming a NBA starter",
       x = "Probability of becoming a starter")
```



The distribution of `starter` is skewed to the right with a median of around 0.11, meaning the typical player in the 2015 NBA draft had about an 11% chance of becoming a starting-caliber player in the NBA. The distribution has a range of about 0.5 and an interquartile range of about 0.2. The highest chance any player had of becoming a starter was just above 50%, while the lowest was around 2.5%. There are no outlier players at either end of the distribution.

In this report, I explored the relationship between a player's position and their probability of becoming a starting-caliber player in the NBA. Since some positions are more necessary to a team (center) than other, more flexible positions (shooting guard, small forward), I predict that a player's position affects their probability of becoming a starter in the NBA.

My hypothesis can be represented by the word equation: **`starter` = `position` + other stuff**

[5]:
```
summary(subset(draft, position == "C")$starter)
summary(subset(draft, position == "PF")$starter)
summary(subset(draft, position == "PG")$starter)
summary(subset(draft, position == "SF")$starter)
summary(subset(draft, position == "SG")$starter)

gf_boxplot(~starter, data = draft, fill = "coral")%>%
gf_facet_grid(position ~ .)%>%
```

```
gf_labs(title = "Probability of a top projected player in the 2015 NBA␣
  ↪draft\nbecoming a NBA starter by position",
         x = "Probability of becoming a starter")
```

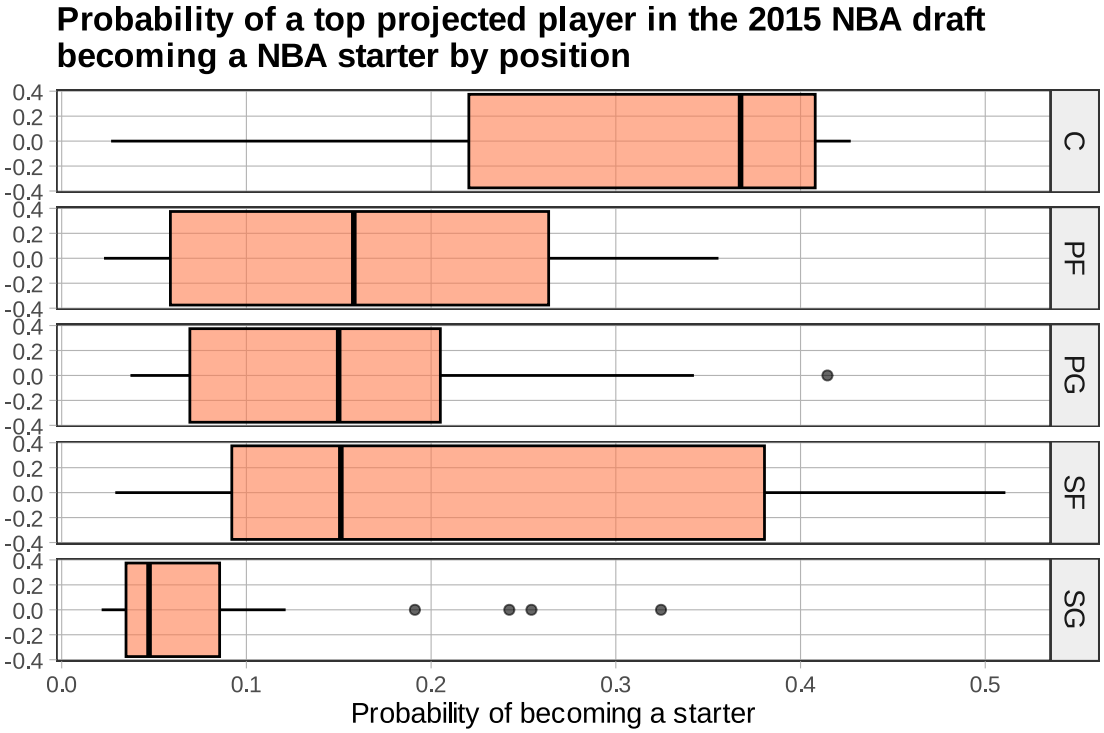
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02676 0.22043 0.36754 0.29689 0.40795 0.42718

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02292 0.05885 0.15815 0.17034 0.26364 0.35557

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03722 0.06939 0.14996 0.15968 0.20500 0.41456

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02902 0.09211 0.15113 0.22848 0.38049 0.51090

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02163 0.03483 0.04730 0.08344 0.08551 0.32447
```



- The distribution of `starter` for centers is skewed left with a median of 0.36754, a mean of 0.29689, and a range of about 0.027 to 0.427. There are no outliers.
- The distribution of `starter` for power forwards is roughly symmetrical, with a median of 0.15815, a mean of 0.17034, and a range of about 0.023 to 0.356. There are no outliers.
- The distribution of `starter` for point guards is roughly symmetrical, with a median of 0.14996, a mean of 0.15969, and a range of about 0.037 to 0.342, excluding the

upper outlier at about 0.415.

- The distribution of `starter` for small forwards is skewed right with a median of 0.15113, a mean of 0.22848, and a range of about 0.029 to 0.511. There are no outliers.
- The distribution of `starter` for shooting guards is skewed right with a median of 0.04730, a mean of 0.08344, and a range of about 0.022 to 0.121, excluding the upper outliers at about 0.191, 0.242, 0.254, and 0.324.

The distributions of `starter` for power forwards, point guards, and small forwards are relatively similar, with their centers around the same point. The overall distribution of `starter` for centers is much higher than the other distributions. The overall distribution of `starter` for shooting guards is much lower than the other distributions.

The differences between these distributions suggest that centers have a relatively high probability of becoming an NBA starter while shooting guards have a relatively low probability of becoming an NBA starter. Overall, these observations support my hypothesis that a player's position affects their likelihood of becoming a starting-caliber player in the NBA.

[6]:
```
position.model <- lm(starter ~ position, data = draft)
position.model
```

```
Call:
lm(formula = starter ~ position, data = draft)

Coefficients:
(Intercept)   positionPF   positionPG   positionSF   positionSG
    0.29689     -0.12655     -0.13721     -0.06841     -0.21345
```

The best-fitting model I decided to use is the `position` model. The GLM notation for this model is:

$$Y_i = 0.29689 - 0.12655X_{1i} - 0.13721X_{2i} - 0.06841X_{3i} - 0.21346X_{4i} + e_i$$

- b_0: The average probability of becoming an NBA starter for a center is 0.29689
- b_1: The average increment in probability of becoming an NBA starter for a power forward compared to a center is -0.12655
- b_2: The average increment in probability of becoming an NBA starter for a point guard compared to a center is -0.13721
- b_3: The average increment in probability of becoming an NBA starter for a small forward compared to a center is -0.06841
- b_4: The average increment in probability of becoming an NBA starter for a shooting guard compared to a center is -0.21346

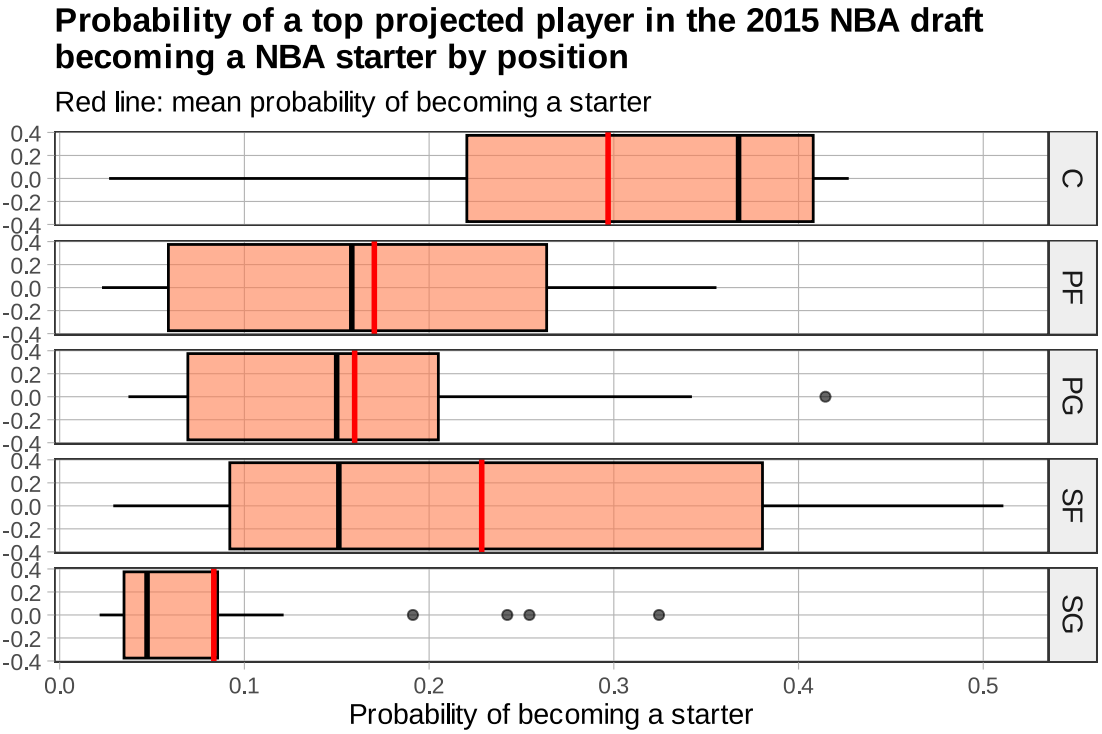
- X_{1i}: Whether or not the player is a power forward or not (0 = not a PF, 1 = PF)
- X_{2i}: Whether or not the player is a point guard or not (0 = not a PG, 1 = PG)
- X_{3i}: Whether or not the player is a small forward or not (0 = not a SF, 1 = SF)
- X_{4i}: Whether or not the player is a point guard or not (0 = not a SG, 1 = SG)

- e_i: error (actual probability - predicted probability)

Below is the model overlayed on the boxplot of `starter` by `position`:

[7]:
```
dat <- draft%>%
group_by(position)%>%
summarize(mean = mean(starter))

gf_boxplot(~starter, data = draft, fill = "coral")%>%
gf_facet_grid(position ~ .)%>%
gf_vline(xintercept = ~mean, data = dat, color = "red")%>%
gf_labs(title = "Probability of a top projected player in the 2015 NBA draft\nbecoming a NBA starter by position",
        subtitle = "Red line: mean probability of becoming a starter",
        x = "Probability of becoming a starter")
```



[8]:
```
supernova(position.model)
```

```
Analysis of Variance Table (Type III SS)
Model: starter ~ position

                          SS df    MS     F   PRE     p
----- --------------- | ----- -- ----- ----- ----- -----
Model (error reduced) | 0.330  4 0.083 5.492 .2338 .0006
```

```
Error (from model)    | 1.082 72 0.015
----- --------------- | ----- -- ----- ----- ----- -----
Total (empty model)   | 1.412 76 0.019
```

- **PRE = 0.2338:** The `position` model reduces 23.38% of the error in the empty model
- **F-ratio = 5.492:** The variance reduced by the `position` model is 5.492 times greater than the variance left unexplained
- **p-value = 0.0006:** There is a 0.06% chance of having an observed F-ratio greater than our observed sample F-ratio of 5.492 given that the empty model is true

Compared to the empty model, the `position` model reduces more error, as shown by the values of the PRE and F-ratio. Since the p-value of 0.0006 is less than the critical value $\alpha = 0.05$, we have sufficient evidence to reject the empty model and conclude there is a difference in the probability of becoming an NBA starter depending on the player's position.

[9]:
```
position.model
confint(position.model)
```

```
Call:
lm(formula = starter ~ position, data = draft)

Coefficients:
(Intercept)   positionPF   positionPG   positionSF   positionSG
    0.29689     -0.12655     -0.13721     -0.06841     -0.21345
```

A matrix: 5 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	0.2045432	0.38923682
positionPF	-0.2338428	-0.01924819
positionPG	-0.2503063	-0.02410370
positionSF	-0.1846105	0.04779054
positionSG	-0.3184069	-0.10849977

However, while a conclusion can be made, random chance can still never be entirely ruled out as a possible cause for the patterns seen in the sample data. This is because whenever we reject the empty model, there is always a possibility of making a Type I Error, where you reject the empty model when it is in fact true in the DGP. The probability of making such an error is equal to the value of our critical value $\alpha = 0.05$.

The ANOVA table doesn't tell us which specific values of `position` cause a difference in `starter` in the DGP. To find that, you have to use pairwise comparisons, which compare each possible pair of `position` values to determine which pairs are correlated with a difference in `starter` in the DGP.

[10]:
```
pwc <- pairwise(position.model)
pwc
plot(pwc)
```

Tukey's Honestly Significant Differences

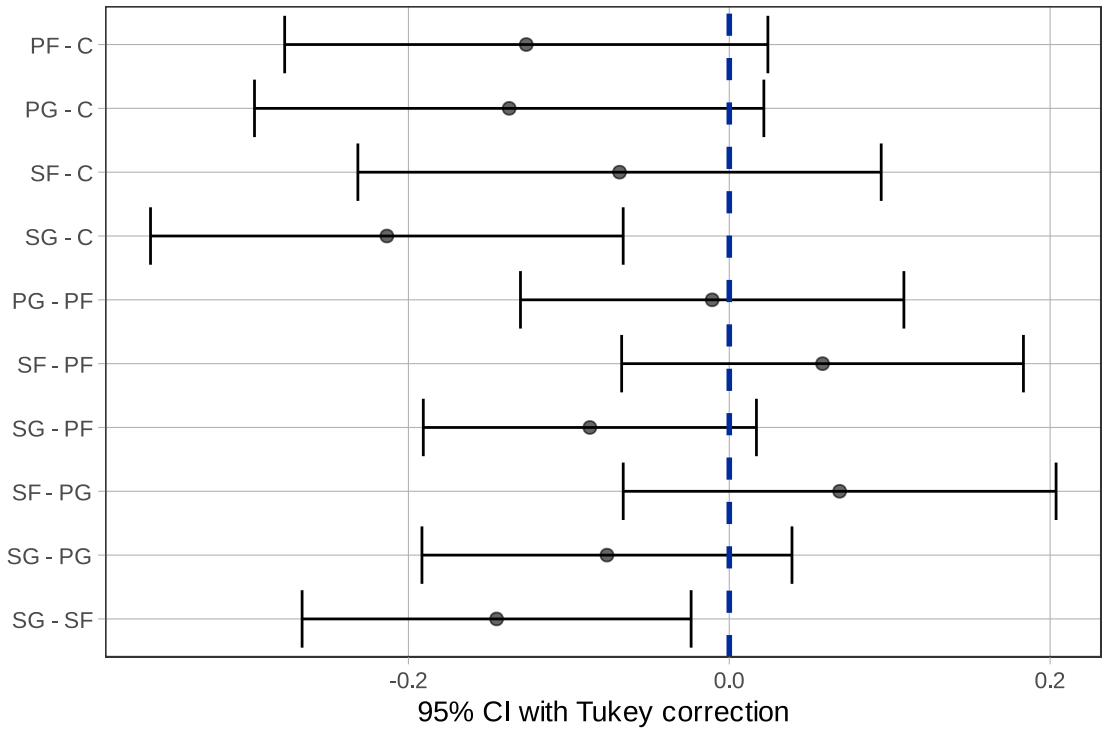
Model: starter ~ position

position

Levels: 5

Family-wise error-rate: 0.05

	group_1	group_2	diff	pooled_se	q	df	lower	upper	p_adj
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	PF	C	-0.127	0.038	-3.325	72	-0.277	0.024	.1410
2	PG	C	-0.137	0.040	-3.420	72	-0.296	0.022	.1222
3	SF	C	-0.068	0.041	-1.660	72	-0.232	0.095	.7663
4	SG	C	-0.213	0.037	-5.734	72	-0.361	-0.066	.0012
5	PG	PF	-0.011	0.030	-0.353	72	-0.130	0.109	.9991
6	SF	PF	0.058	0.032	1.837	72	-0.067	0.183	.6927
7	SG	PF	-0.087	0.026	-3.312	72	-0.191	0.017	.1437
8	SF	PG	0.069	0.034	2.018	72	-0.066	0.204	.6126
9	SG	PG	-0.076	0.029	-2.616	72	-0.192	0.039	.3537
10	SG	SF	-0.145	0.031	-4.734	72	-0.266	-0.024	.0110



The pairwise comparisons table and plot display the 95% confidence intervals for the true mean differences in `starter` for each pair of `position` values. The limits to the confidence intervals are given in the `lower` and `upper` columns of the table. - The 95% confidence intervals for **PF-C (-0.227, 0.024)**, **PG-C (-0.296, 0.022)**, **SF-C (-0.232, 0.095)**, **PG-PF (-0.130, 0.109)**, **SF-PF (-0.067, 0.183)**, **SG-PF (-0.191, 0.017)**, **SF-PG (-0.066, 0.204)**, and **SG-PG (-0.192, 0.039)** all contain 0, meaning that it is possible that the true mean difference in probability of becoming an NBA starter between each of those pairs is 0 in the DGP. - The 95% confidence intervals for **SG-C (-0.361, -0.066)** and **SG-SF (-0.266, -0.024)** do not contain 0, meaning that we are 95% confident that there is a true mean difference in probability of becoming an NBA starter between each of those pairs in the DGP.

The `p_adj` column of the pairwise comparisons table contains the p-values for the mean difference in `starter` for each pair of `position` values. SG-C and SG-SF are the only pairs with a p-value less than the critical value $\alpha = 0.05$. - **SG-C p-value = 0.0012:** There is a 0.12% chance of getting a mean difference in `starter` between shooting guards and centers as extreme as or more extreme than the mean difference seen in our sample data (-0.21346) - **SG-SF p-value = 0.011:** There is a 1.1% chance of getting a mean difference in `starter` between shooting guards and small forwards as extreme as or more extreme than the mean difference seen in our sample data (-0.14504)

In other words, if you are comparing a pair of players whose respective positions are shooting guard and center or shooting guard and small forward, the position the players play (`position`) affects their probability of becoming a starting-caliber player in the NBA (`starter`). However, if you are comparing two players whose respective positions

are not either of those two possible pairs, then the mean difference in `starter` is not extreme enough to conclude that `position` significantly affects `starter`.

In conclusion, my hypothesis of **`starter` = `position` + other stuff** was only partially correct. A player's position significantly affects their chance of becoming an NBA starter only when comparing specific pairs of positions. Still, I opted to reject the empty model in favor of the `position` model, which explains at least some of the total error. The results of my study imply that centers are in high demand in the NBA while shooting guards have the lowest demand, which would impact how prospects present themselves to NBA teams during the drafting process.

[11]:
```
table(draft$position)
table(draft$position)/77
```

```
 C PF PG SF SG
 7 20 14 12 24
```

```
         C         PF         PG         SF         SG
0.09090909 0.25974026 0.18181818 0.15584416 0.31168831
```

However, I didn't consider how many players of each position were in the 2015 NBA draft. As seen in the tables above, there were the least amount of centers (9.1%) and small forwards (15.6%) and the highest amount of shooting guards (31.2%). I would have to analyze more draft years to see if my conclusions are consistent.

0.2 1.0 - Intro/ Overview of the Problem or Question

The goal of this section is to provide an overview of the **context, situation,** or **problem.** Below are some questions for you to consider while writing the report.

Introduce the **context** - Which data frame are you using and why? - Who has provided the data (as much as you know)? In general, why do you think they collected this data? - What are the cases and variables in the data frame? Showcase your data frame as evidence. - What outcome variable do you want to explore? *Showcase and describe that distribution.* (Remember that statistics is the study of *variation* – so be sure to pick a variable that has a lot of variation!)

Introduce the beginnings of your **situation** or **problem**: - What research question do you want to explore with the data? In other words, what outcome will you try to explain and what variable will you use to predict it? (You will explore the variation of the two variables in the next section. This section is to explain, in words, what you are *wanting* to explore.) - What are your initial predictions about this situation? - What word equation might represent your hypothesis?

0.3 2.0 - Exploring Variation

The goal of this section is to explore variation in your explanatory and outcome variables. That exploration will almost certainly include visual displays of your data. Below are some questions for you to consider.

Start your exploration:

- What visualizations would best convey the variability you're trying to explain? (You can include multiple visualizations if you think they will be helpful to your readers.)
- If needed, would it help to create new variables based on the existing ones to better see any differences?
- Are there any mistakes in the data that you found that need to be fixed? For example, maybe there are some cases that need to be excluded? Or maybe you had some missing data that you needed to handle?

Start your analysis:

- What do you see in your visualization?
- So far, just based on these visualizations, what do you think about your research question or hypothesis?

0.4 3.0 - Model Variation

The goal of this section is to create a model or models of your hypothesis to explain some of the variation in your outcome variable.

Fit a model: - What is the best-fitting model that you decided to use? Add your model to your visualization. - What is the GLM notation for your model? - What do all parts of your GLM mean in terms of your context?

0.5 4.0 Evaluate Model(s)

This goal of this section is to evaluate how good your model is in explaining the variation in your outcome variable.

Evaluate your model: - How does your model compare to the empty model? - What does the ANOVA table tell you about the how well the model fits the data and, if applicable, how does it compare the fit of alternative models? - Can you rule out randomness/ chance as a possibility? - *What does the confidence interval tell you about your research question? - Using p-value or confidence intervals, how does it hold up against the empty model? What is the rationale for which model you opt to keep?*

0.6 5.0 Conclusion

The goal of this section is to help your audience understand what can be learned from your data analysis. You are reminding them of how all this work relates back to your initial word equation.

Questions to consider: - Why did you choose to pursue this research question and what were you predicting? - How would you summarize all that you did, what you found, and how it relates to the motivating question? - What are the implications of the results and what it means for the audience or the world? - How well do you think your model represents the DGP? Are there any other factors that did not previously consider that you are now curious about?

0.7 Submit your rough draft here

- Due by Monday 05/15/2023
- You can submit earlier too
- The sooner you submit your rough draft, the sooner you can get feedback and start on your final draft.

0.8 Submit your final draft here

- Due on the day of the final, Tuesday 05/23/2023

0.9 Present your report on the day of the final

- Find expectations for your poster here