How to Choose Graphs and Parameters in ggplot2 to Strengthen Our Answer to a Research Question

Jack Ji

October 25, 2017

Introdution:

When we do data analysis, we always start with a research question, which perhaps hypothesizes about some special property, pattern, or relationship that exists within the data. Graphs can certainly help us corroborate our hypothesis about the research question and convince others of it. Although we have used some graphs in the homework of Stat 133, we never formally discuss when we should use which type of graphs and what their respective specialties are. In this post, we will peruse four major categories of graphs and their functions in the `ggplot2` package and see how to select the appropriate type of graphs that we should use for different types of answers to research questions and how to utilize some parameters of the corresponding "geom" function in order to strengthen our answer very more.

Body:

Before we delve into the graph types, we need to import the package of `ggplot2` (and `dplyr` for purposes of data manipulation) and read in the NBA Player datasets we used for [HW3][HW].

```
library(ggplot2)
library(dplyr)
stats <- read.csv("./nba2017-stats.csv")
teams <- read.csv("./nba2017-teams.csv")
roster <- read.csv("./nba2017-roster.csv")
```

- Reference for `ggplot2` : https://www.rdocumentation.org/packages/ggplot2/versions/2.2.1

Although there are myriad graphing functions that start with the word "geom" in the `ggplot2` package, they can be grouped into four major categories–they each emphasize correlation, ranking, composition and distribution of the data. Note that this post does not focus on simply how to create graphs but rather when to use them and how to use some function parameters to tailor the graph closer to our research question.

Ranking

Let's say we came up with the research question "Why is BOS a good NBA team?". To answer that, we might focus on the aspect of total points scored by this team and compare it to the total points of other teams–the reason why BOS is good is that it is ranked very high among all teams in terms of total points. This is the notion of "ranking"–comparing a specific variable (attribute) of an element within the dataset to that of its counterparts. We should use graphs in this category when we want to how an element in the data is ranked with respect to a specific variable.

Ordered bar graphs

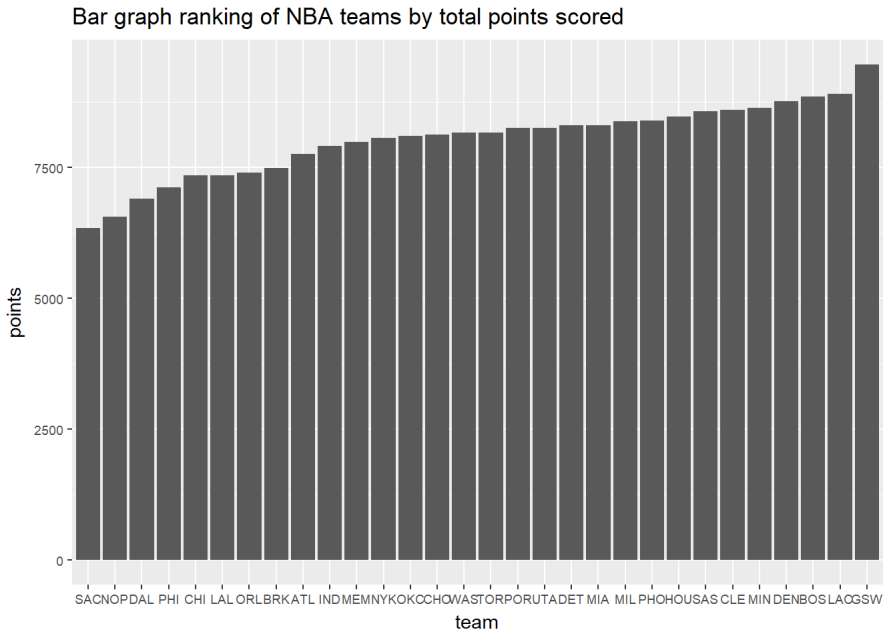
An ordered bar graph is ordered by its y-axis variable. It manifests where the total points of BOS is ranked among the total points of all teams.

To make a bar graph with `ggplot2` , one may call the function `geom_bar()` (as we have seen and used before in class, but we never touched upon why and when we want to use bar graphs). It is important to order our data by our variable of interest. The resulting graph shows that BOS is ranked third among all teams in terms of total points scored. BOS's high ranking provides an answer to our research question and elucidates why "BOS is a good NBA team" in the aspect of total points.

```
teams_bypoints <- teams[order(teams$points), ] # Rank teams in ascending order of points

# Maintain the order of increasing points in the graph (an advanced technique not covered in lecture)
teams_bypoints$team <- factor(teams_bypoints$team, levels = teams_bypoints$team)

# plot the bar graph with slight adjustment to the axis labels
ggplot(data = teams_bypoints, aes(x = team, y = points)) + geom_bar(stat = "identity") +
      theme(axis.text = element_text(size=7)) +
      ggtitle("Bar graph ranking of NBA teams by total points scored")
```



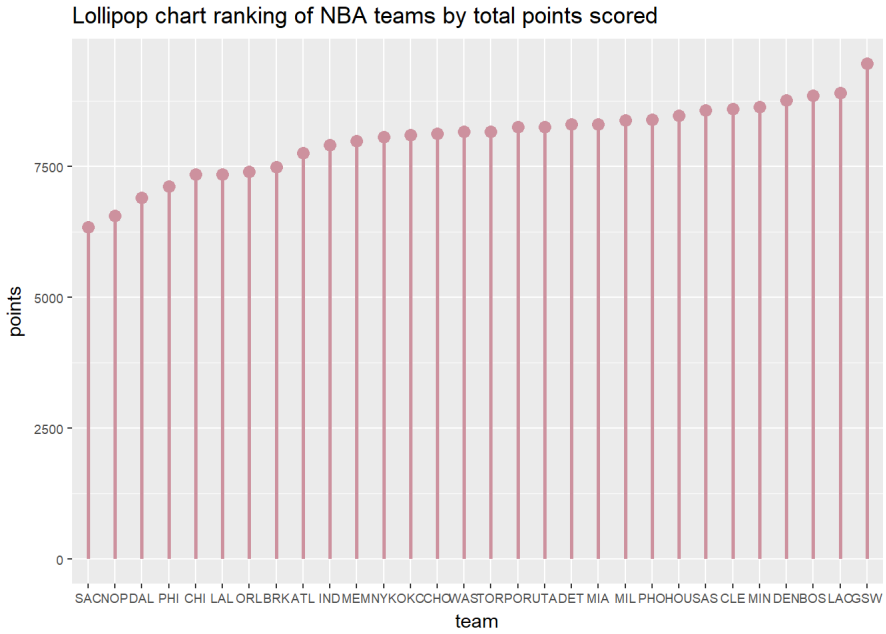
Lollipop charts

To answer the same research question and focus on BOS's ranking among all NBA teams, lollipop charts are as useful as a bar graph–largely because a lollipop chart is a variation of a bar graph. Similarly as a bar graph, a lollipop chart shows the ranking of BOS in terms of total points and puts more emphasis on the height of each bar (total points of each team).

To make a lollipop chart, one may call the function `geom_point()` to construct the points on the top of each "lollipop" first (or the apex of each bar above) and specify a large size of the points and then call the function `geom_segment()` to make a perpendicular line between each point and the x-axis.

- Reference: https://uc-r.github.io/lollipop

```
# use the data frame above to plot
ggplot(data = teams_bypoints, aes(x = team, y = points)) +
    geom_point(size = 3, col = "pink3") +
    geom_segment(aes(x = team, xend = team, y = 0, yend = points), col = "pink3", size = 1) +
    theme(axis.text = element_text(size=7)) +
    ggtitle("Lollipop chart ranking of NBA teams by total points scored")
```



Correlation

Suppose that we are now interested in another research question "what kind of people can become an NBA basketball player?". One might answer the question by examining some physical features of the current NBA players (e.g. height and weight). Focusing on a pair of variables like height and weight, we might wonder if it is fine for an NBA player to have arbitrary height and weight and if there exists a relationship between a player's height and weight. This is the notion of "correlation"–we use graphs in this category to demonstrate the existence and magnitude of variables' correlation.

Scatterplots

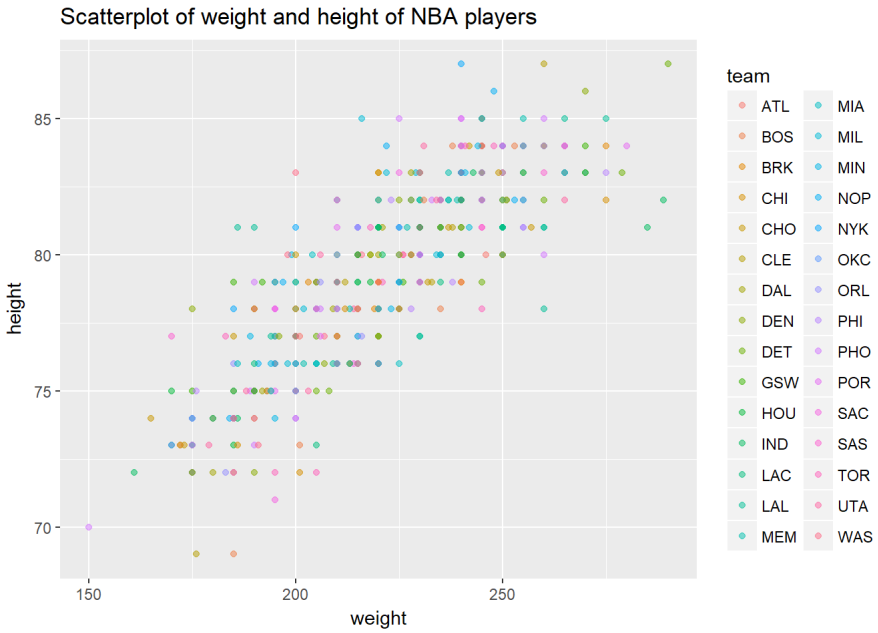
As bar graphs, we have seen and used these in class, but we did not discuss under what situation we should use them. We use a scatterplot

when we want to examine whether there exists a nexus between two variables and, if there exists one, how strong it is. In this case, we are interested in the correlation between the weight and the height of an NBA player.

We are familiar with the procedure of constructing a scatterplot, calling `geom_point()` from `ggplot2` and set `weight` as the x-value and `height` as the y-value. As we see in the plot, the cluster of points is almost in the shape of a line, indicating that there is a strong correlation between height and weight. That is to say, it is rare to see very tall and light or very short and heavy.

- Reference: http://www.cookbook-r.com/Graphs/Scatterplots_(ggplot2)/

```
# plot a scatterplot with player's weight and height; points are colored by team
ggplot(data = roster) + geom_point(aes(x = weight, y = height, col = team), alpha = 0.5) +
  ggtitle("Scatterplot of weight and height of NBA players")
```



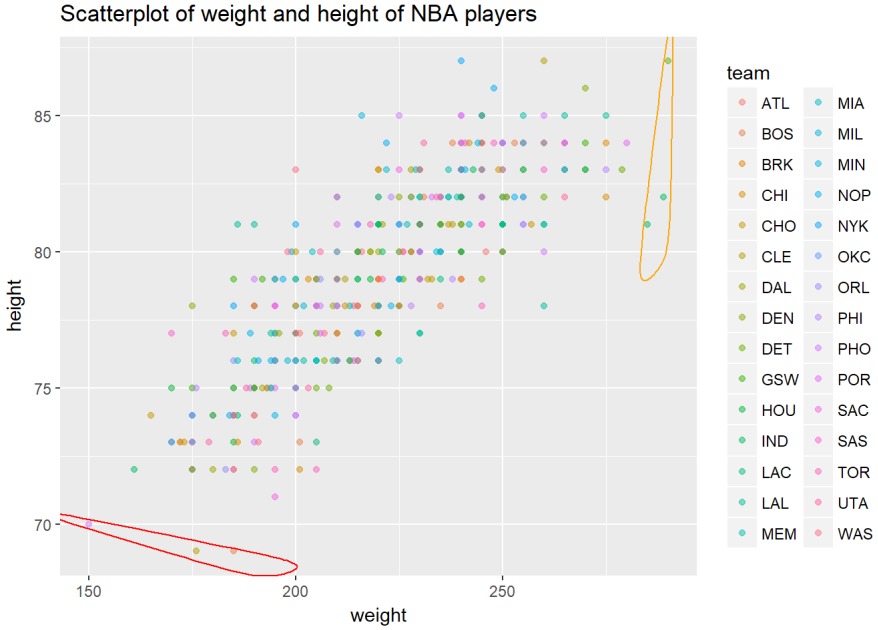
We can make use of a new package here named `ggalt`. It enables us to circle up some points in the scatterplot and call reader's attention to them. For instance, we can call `geom_encircle()` here to note some anomalies–the red circle includes players that are very short and the yellow one includes players that are very heavy. By doing this, we can tailor our graph more to our answer to the research question so that it becomes easier for us to convince the reader of our point.

```
library(ggalt)

short_players <- roster[roster$height <= 70, ]

heavy_players <- roster[roster$weight >= 285, ]

# plot a scatterplot with player's weight and height, colored by team
ggplot(data = roster) + geom_point(aes(x = weight, y = height, col = team), alpha = 0.5) +
    geom_encircle(aes(x = weight, y = height), data = short_players,
                  expand = 0.1, color = "red") +
    geom_encircle(aes(x = weight, y = height), data = heavy_players,
                  expand = 0.08, color = "orange") +
    ggtitle("Scatterplot of weight and height of NBA players")
```

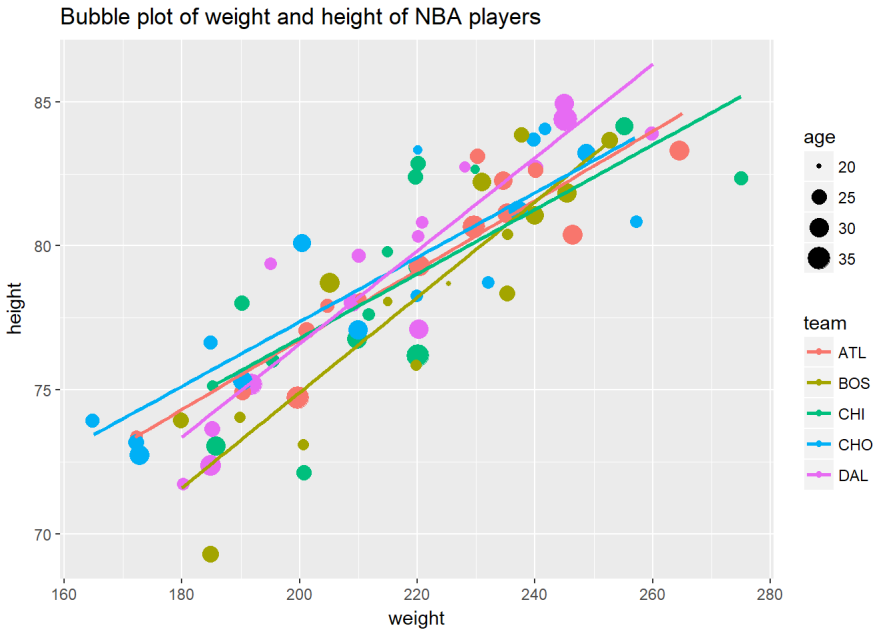


Bubble plots

To create a bubble plot, we use the function `geom_jitter()` in `ggplot2` to plot the points (this function is excellent at handling overlapping points as it randomly moves them slightly) and use `geom_smooth()` to add some lines of best fit to help us detect patterns. Bubble plots are good for any analysis of three variables–the size of the points adds information about the third variable, in this case `age`, to the original scatterplot. On top of our conclusion from the previous scatterplot, this bubble plot shows that age does not seem to have any correlation with height or weight as points of different sizes are uniformly distributed across the graph.

```
roster_reduced <- filter(roster, team == "ATL" | team == "BOS" | team == "CHI" | team == "CHO" | team == "DAL") # create a data frame with just the first five teams

ggplot(data = roster_reduced, aes(x = weight, y = height)) +
      geom_jitter(aes(col = team, size = age)) +
      geom_smooth(aes(col = team), method = "lm", se = FALSE) +
      ggtitle("Bubble plot of weight and height of NBA players")
```



Bubble plots with marginal histogram (a variation)

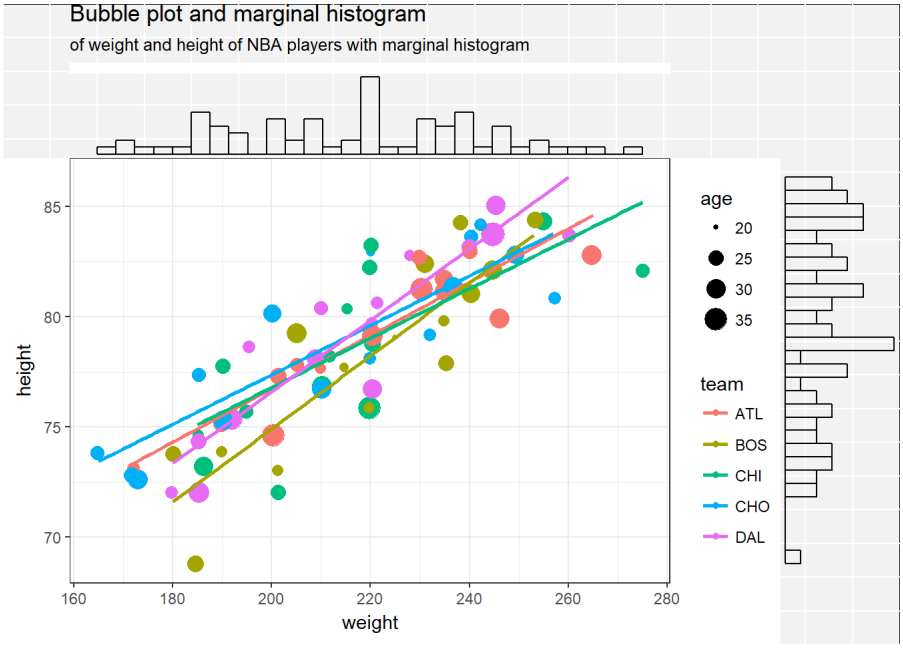
Using the package `ggExtra` and calling the function `ggMarginal()`, we can add two more graphs next to each axis. Such juxtaposition permits us to bring in other insights about the data. In this case, histograms show the distribution of height and weight among all players. With knowledge in both correlation and distribution, we may conclude that not only is there a linear correlation between height and weight but also that most players have medium height (~ 80 inches) with their weight skewed to the right.

Reference: https://cran.r-project.org/web/packages/ggExtra/README.html

```
library(ggExtra)
theme_set(theme_bw())

scatterplot_extra <- ggplot(data = roster_reduced, aes(x = weight, y = height)) +
                     geom_jitter(aes(col = team, size = age)) +
                     geom_smooth(aes(col = team), method = "lm", se = FALSE) +
          labs(title = "Bubble plot and marginal histogram",
               subtitle = "of weight and height of NBA players with marginal histogram")

plot(ggMarginal(scatterplot_extra, type = "histogram", fill = "transparent"))
```



Distribution

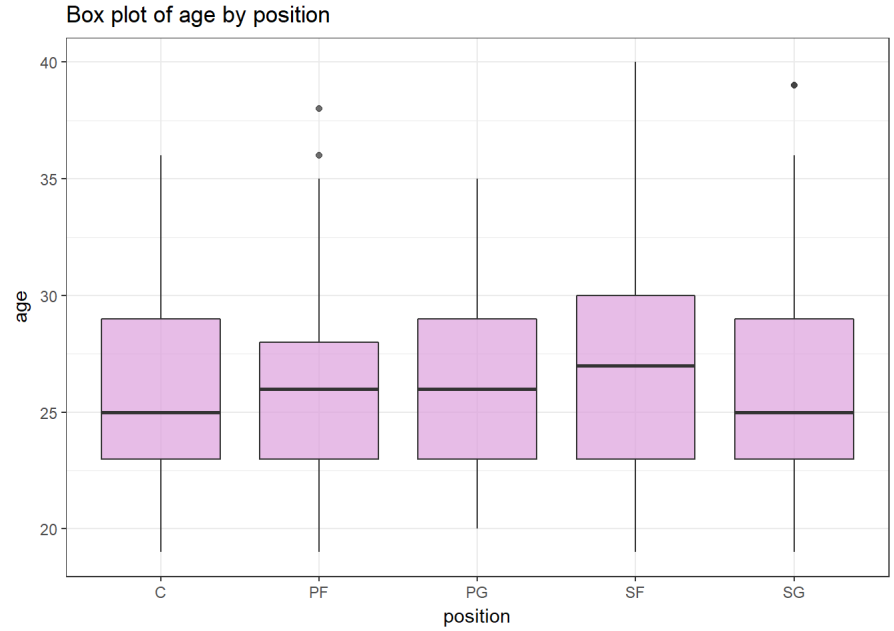
Finally, for research questions that, for example, ask about the age of an NBA player with respect to his position, we indeed care about "distribution" in lieu of "correlation". Since there are only five positions in basketball, it is very hard to plot a correlation graph with merely 5 values for one of the axes. It is more informative for us to analyze the "distribution" of the data–the range, median, quartiles of each group of data. We use graphs in this category to illustrate how many elements fall into which particular range of the entire span of a variable.

Box plots

To plot a box plot, call the function `geom_boxplot()` in the package `ggplot2`. A box plot displays the median and quartiles of players that play each position. From this box plot, we can conclude that most players, regardless of their positions, are around the age of 25. The variance is large, though, with most positions extending past 35 and SF to even 40.

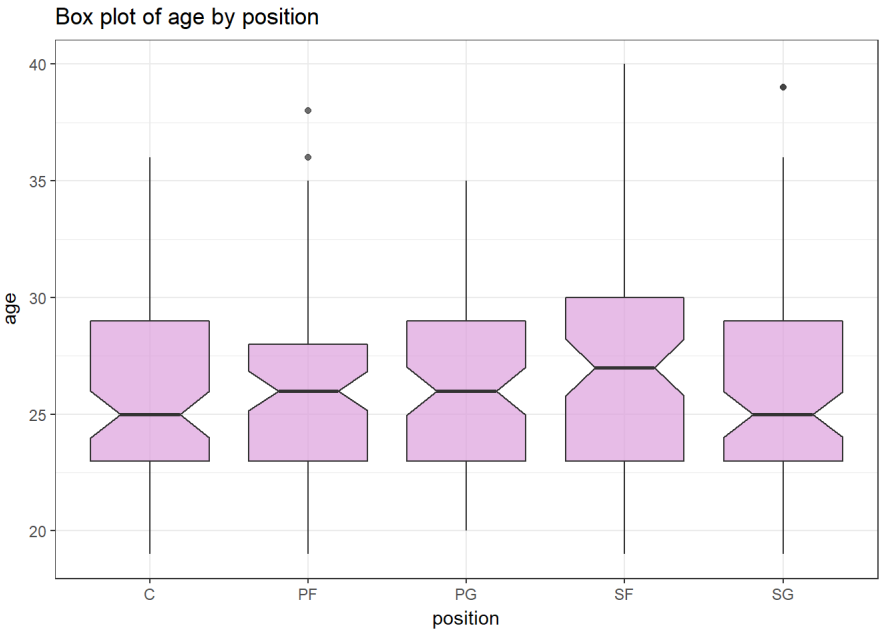
- Reference: http://t-redactyl.io/blog/2016/04/creating-plots-in-r-using-ggplot2-part-10-boxplots.html

```
# plot the graph
ggplot(data = roster, aes(x = position, y = age)) +
  geom_boxplot(fill = "plum", alpha = 0.7) +
  ggtitle("Box plot of age by position")
```



Note that we may turn on the `notch` parameter and get more information about the distribution, specifically quartiles.

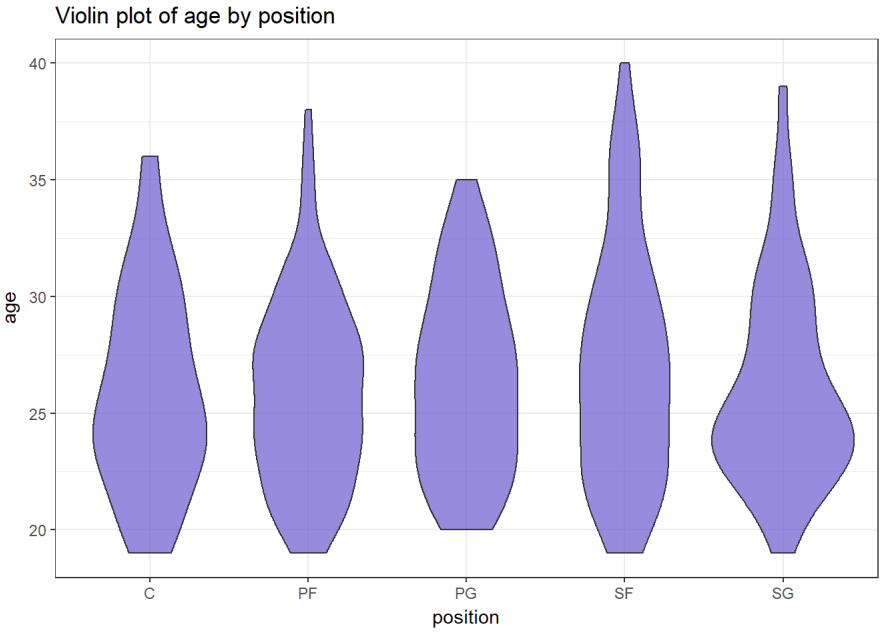
```
# plot the graph
ggplot(data = roster, aes(x = position, y = age)) +
  geom_boxplot(fill = "plum", alpha = 0.7, notch = TRUE) +
  ggtitle("Box plot of age by position")
```



Violin plots

To plot a violin plot, call the function `geom_violin()` in the package `ggplot2`. It is simply a variation of box plot. However, the curvature of each "violin" gives a continuous illustration of the distribution of the variable in each group–thinning implies that few elements take a particular value while thickening implies that many elements take the value. As above, we notice that each bar is long but most elements are centered around age 25. We can hence conclude that most players are around 25 years old.

```
# plot the graph
ggplot(data = roster, aes(x = position, y = age)) +
  geom_violin(fill = "slateblue", alpha = 0.7) +
  ggtitle("Violin plot of age by position")
```



Composition

Assume that we are now asking a research question like "How are the old players distributed among the NBA teams?". Be careful that this is not tantamount to asking about "distribution"–distribution focuses on the median, range, and other attributes within each group of data, whereas the question here inquires more about what percentage of old players each team has out of ALL old players. This invokes the notion of "composition". We use graphs in this category to illustrate how much an element makes up of the entire data.

Pie chart

In this particular example, we define "old players" as players with greater than or equal to 14 years of experience. In order to make a pie chart in `ggplot2`, we actually call the function `geom_bar()` and leave the `x` in `aes()` as blank and instead set `fill` to the variable that we would use for `x` in a bar graph (team in this case). Finally, we put it in polar coordinate. The larger the area of a team in the pie chart, the more old players from the team make up all the old players in the NBA. We can conclude that SAS has the most old players, but the number of old players is fairly close among the 11 teams that have old players.

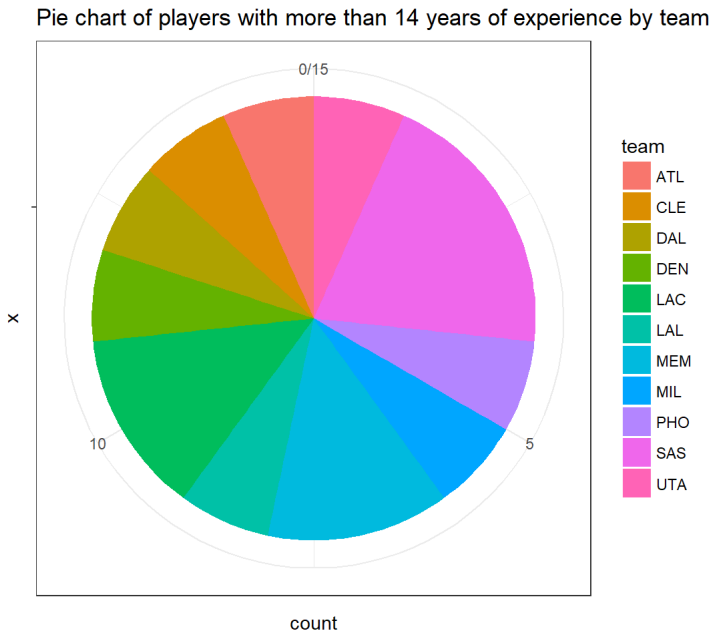
```r
# make a data frame of players with greater than or equal to 14 years of experience
exp_above14 <- roster[roster$experience >= 14, c("team", "experience")]

# group these players by team and count the number of such experienced players in each team
exp_above14 <- group_by(exp_above14, team) %>% summarise(count = n())

colnames(exp_above14) <- c("team", "count") # Set the column names

# plot the bar graph first
pie <- ggplot(data = exp_above14, aes(x = "", y = count, fill = team)) +
                    geom_bar(width = 1, stat = "identity") +
                ggtitle("Pie chart of players with more than 14 years of experience by team")

pie + coord_polar(theta = "y", start=0) # put the bar graph in polar coordinate
```



Take home message

We need to choose the type of graph that is appropriate to our answer to the research question. Specifically, if we want to note some correlation within the data, we should make use of graphs such as scatterplots and bubble plots (perhaps with marginal graphs); if we want to show the composition of a variable within the data, graphs like pie graphs would be a good choice; if we want to call attention to the ranking of an element with respect to some variable in the data, we should consider graphs such as ordered bar graphs and lollipop graphs; if we want to show the distribution of the data elements in terms of a variable, we need to use graphs like box plots and violin plots. Sometimes we can even fine-tune the graph and tailor it more to our argument by specifying some parameters (e.g. color, size of points, etc.) of the graphing function.

References

https://www.rdocumentation.org/packages/ggplot2/versions/2.2.1

https://blog.hubspot.com/marketing/data-visualization-choosing-chart

[correlation][http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#1.%20Correlation]

[ranking][http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#3.%20Ranking]

[distribution][http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#4.%20Distribution]

[composition][http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#5.%20Composition]

http://t-redactyl.io/blog/2016/04/creating-plots-in-r-using-ggplot2-part-10-boxplots.html

[HW][https://github.com/ucb-stat133/stat133-fall-2017/tree/master/data]

https://uc-r.github.io/lollipop

https://cran.r-project.org/web/packages/ggExtra/README.html

http://www.cookbook-r.com/Graphs/Scatterplots_(ggplot2)/