ggplot in R

Minjeong Kim

11/27/2017

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

ggplot

Introduction

Welcome to the world of ggplot! We talked about vector at the middle of the semester. It is a philosophy of visualization in R. the basic concept we must master for further learnings in R. But, I thought that it is important to go over the basic steps to learn about more difficult concepts later on. You will explore more about ggplot in depth: definition, label, scales, zooming, and themes Sounds exciting? YEAH!!! There is a quiz created at the end, so read my post carefully!

What is ggplot?

ggplot2 is a widely used package for creating graphs : data exploration and visualisation package written in R. A function ggplot "ggplot()" is a function used in ggplot2. A data frame object is input for ggplot().

Label

Label is the text for the axis, plot title or caption below the plot. It is helpful for readers to understand the graph well. You can add labels with labs() function.

Let's take a look at several examples. We are using a data-set, "nba2017-player-statistcs" in R.

```
data <- read.csv('../hw02/nba2017-player-statistics.csv')
```

cleaning data

```
data$Experience[data$Experience == "R"] <- "0"
```

```
## Warning in `[<-.factor`(`*tmp*`, data$Experience == "R", value =
## structure(c(18L, : invalid factor level, NA generated
```

```
data$Experience <- as.integer(data$Experience)
```

```
data$Salary <- data$Salary/1000000
```

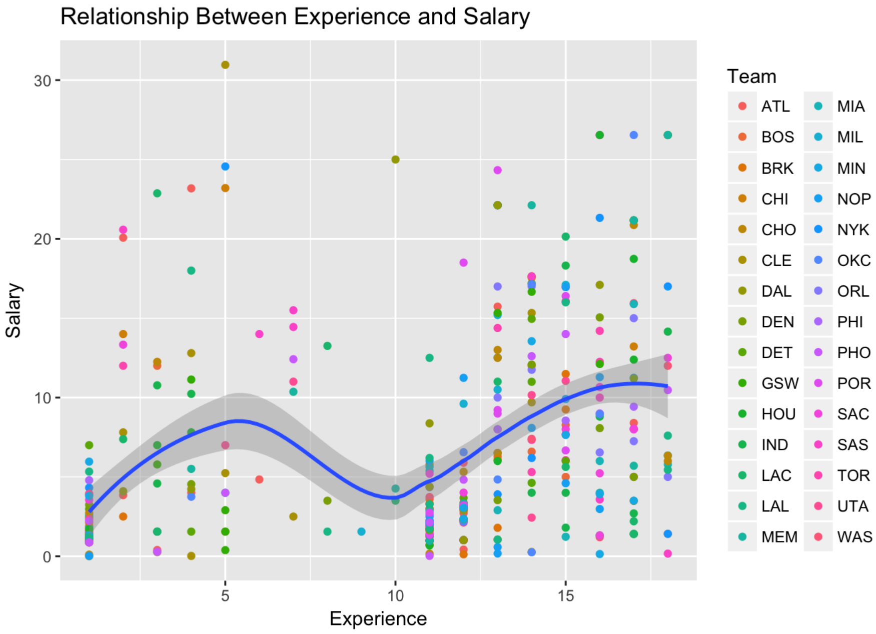
1.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point(aes(color = Team)) +
  geom_smooth() +
  labs(title = "Relationship Between Experience and Salary")
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



```
cor(as.numeric(data$Experience), as.numeric(data$Salary))
```

```
## [1] NA
```

The plot shows a relationship between experience and salary.

As the function cor "cor()" shows, correlation between two variables is 0.05925. Experience and salary have a weak relationship.

In this case, labs() is used to summarize the broad finding.

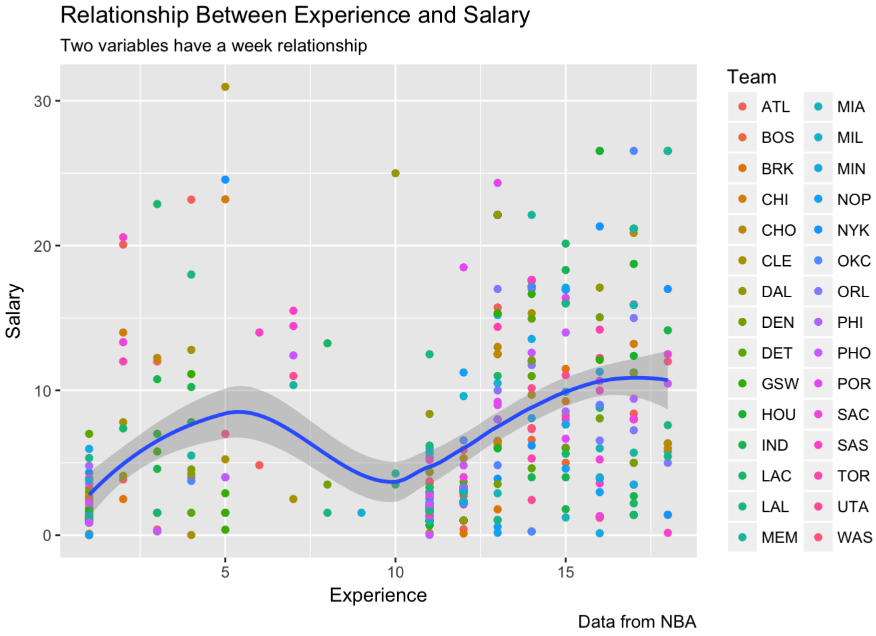
2.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point(aes(color = Team)) +
  geom_smooth() +
  labs(title = "Relationship Between Experience and Salary",
       subtitle = "Two variables have a week relationship",
       caption = "Data from NBA")
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



In this case, labs() is used to show the subtitle and caption.

"subtitle"" adds "additional detail in a smaller font beneath the title." "caption"" adds "text at the bottom right of the plot, often used to describe the source of the data." The caption here is from NBA data, seen from stat133 homeworks and labs.

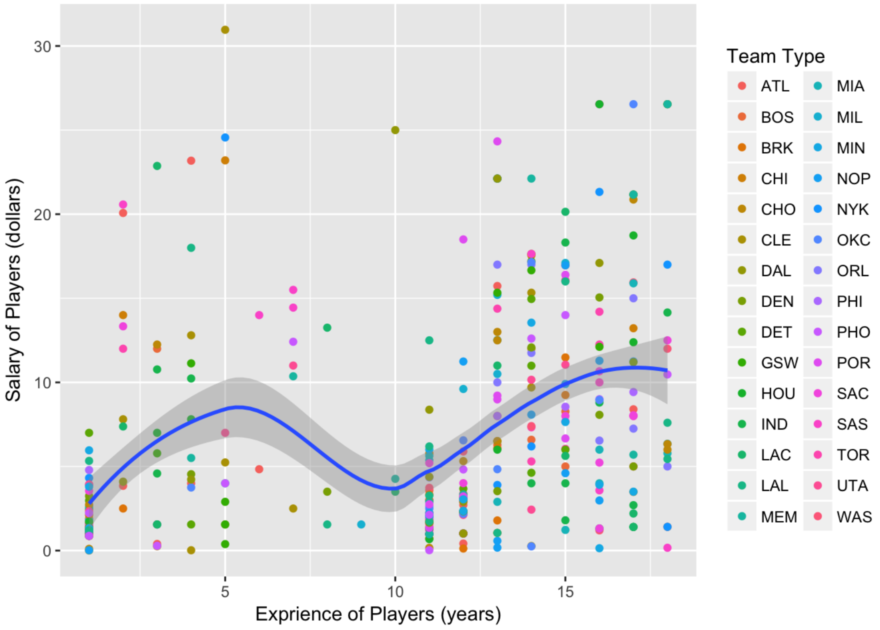
3.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point(aes(color = Team)) +
  geom_smooth() +
  labs(x = "Exprience of Players (years)",
       y = "Salary of Players (dollars)",
       colour = "Team Type")
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



In this case, labs() is used to show the x, y axes and leend titles.

“x” adds x axes with additional descriptions. “y” adds y axes with detailed units.

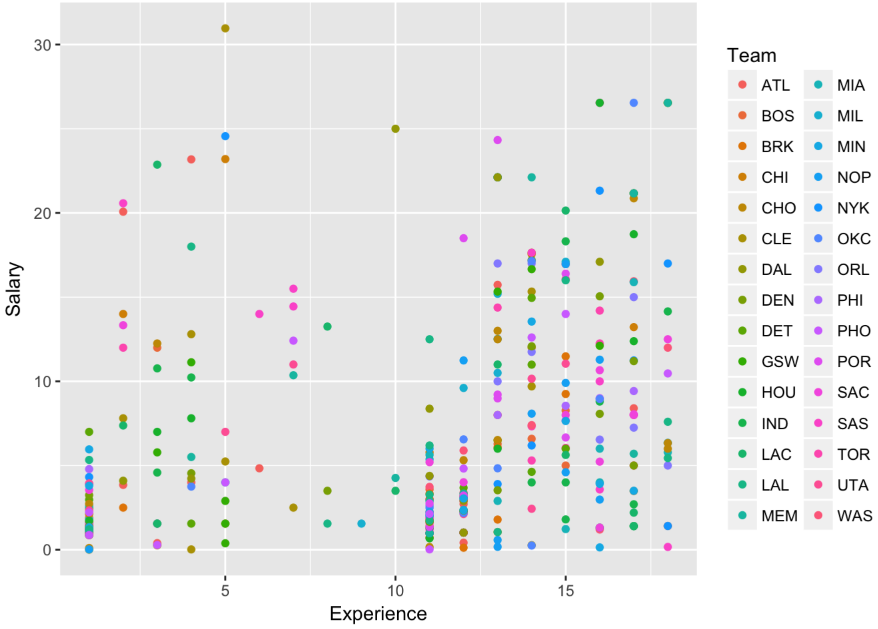
Scales

You can also adjust the scales for readers to look at plots better. ggplot2 automatically has default scales behind the scenes.

Let’s take a look at example.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point(aes(colour = Team))
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```

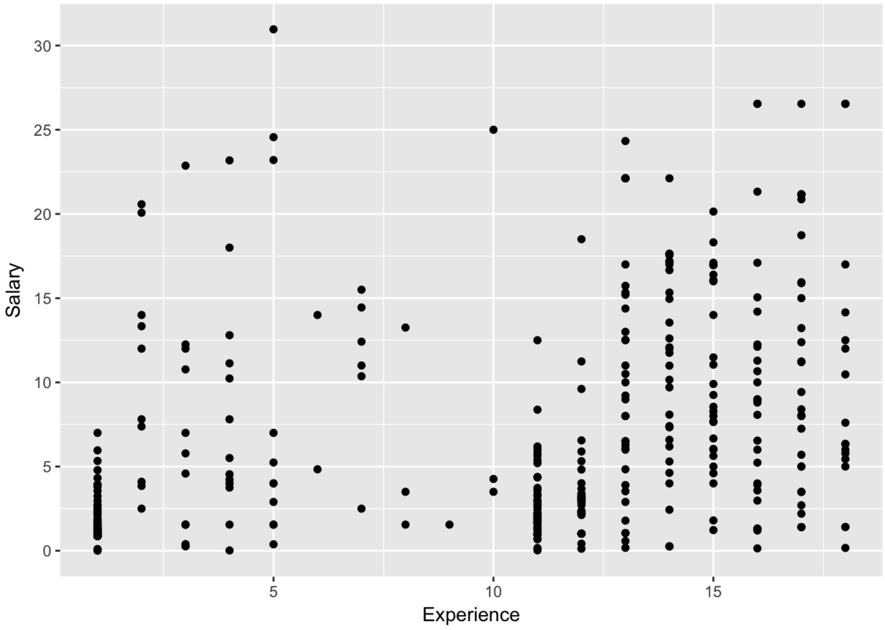


“breaks” is used to overrule the default. It sets the position of the dots, or the values related with the keys. Labels controls the text label related with each key or dot.

1.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point() +
  scale_y_continuous(breaks = seq(0, 40, by = 5))
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```

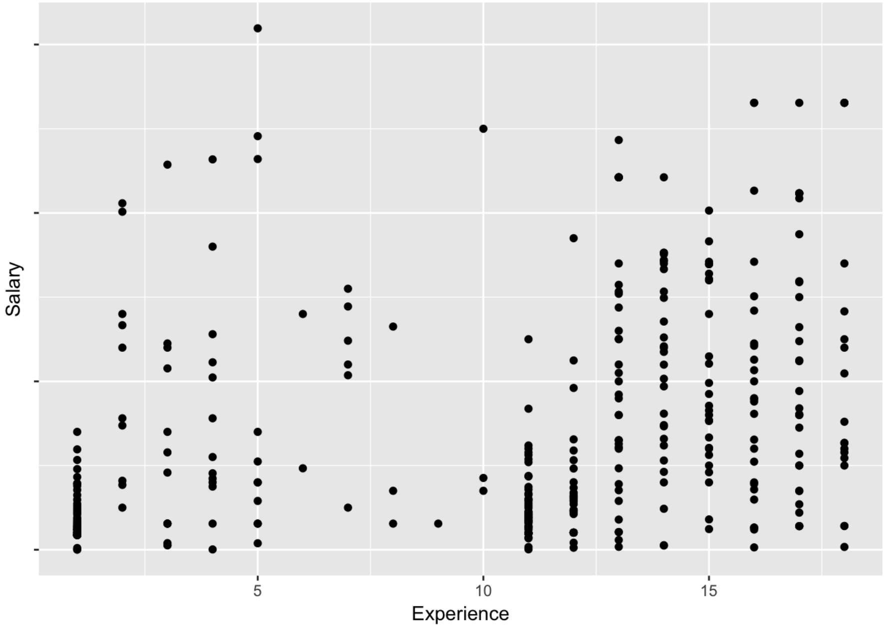


Just notice that a character vector has the same length as breaks; however, setting it to NULL suppresses the labels altogether. In this way, there is no absolute numbers.

2.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point() +
  scale_y_continuous(labels = NULL)
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



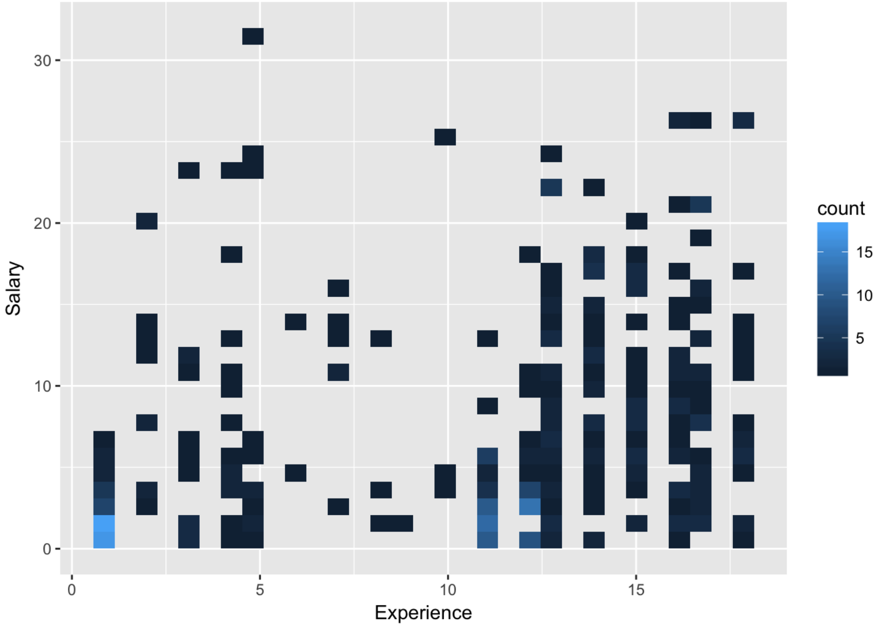
The only difference between two graphs is a white, bold gap in salary.

You can also replace the scales altogether. Let’s take a look at several examples.

3.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_bin2d()
```

```
## Warning: Removed 80 rows containing non-finite values (stat_bin2d).
```

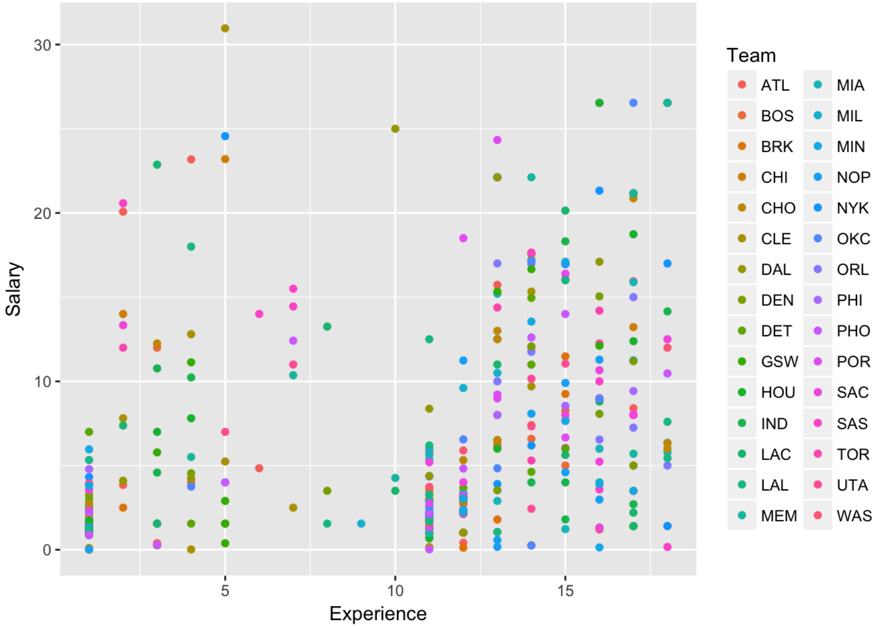


This shows one type of scales: continuous position scales.

4.

```
ggplot(data, aes(x = Experience, y = Salary)) +
  geom_point(aes(color = Team))
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



The colors of dots are differencated by team. This shows another type of scales: color scales.

Zooming

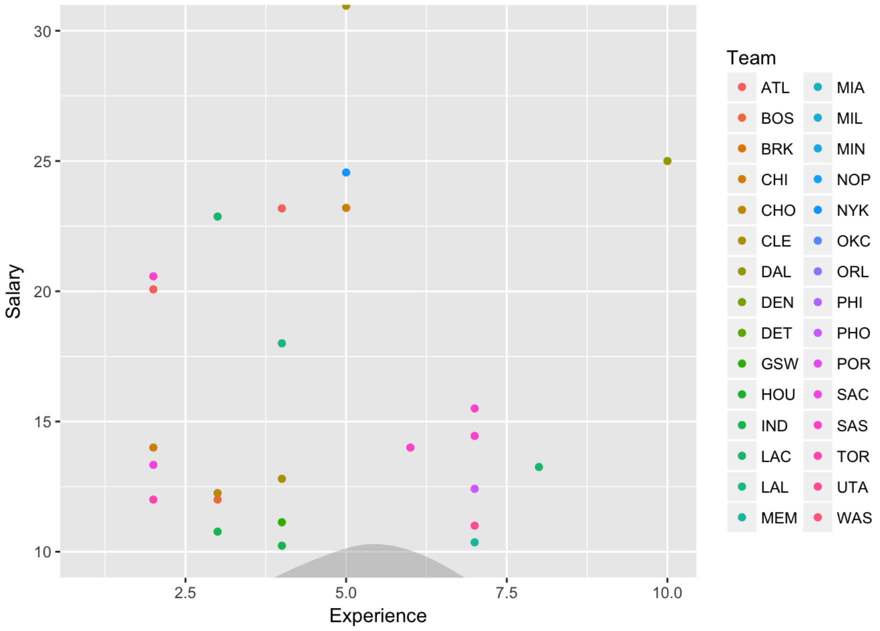
There are three ways of zooming graphs. " 1) adjusting what data are plotted
2) setting the limits in each scale
3) setting xlim and ylim in coord_caresian()
"

```
ggplot(data, mapping = aes(x= Experience, y = Salary)) +
  geom_point(aes(color = Team)) +
  geom_smooth() +
  coord_cartesian(xlim = c(1, 10), ylim = c(10, 30))
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

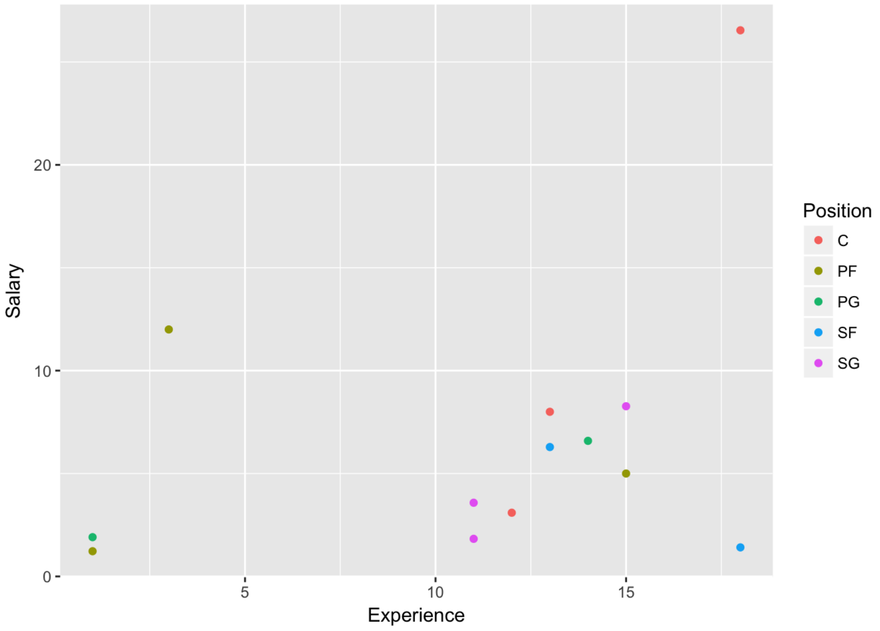
```
## Warning: Removed 80 rows containing missing values (geom_point).
```



In this case, you can set x-limit and y-limit. The disadvantage of this kind of zooming only shows several points. Therefore, you can't see the exact graph.

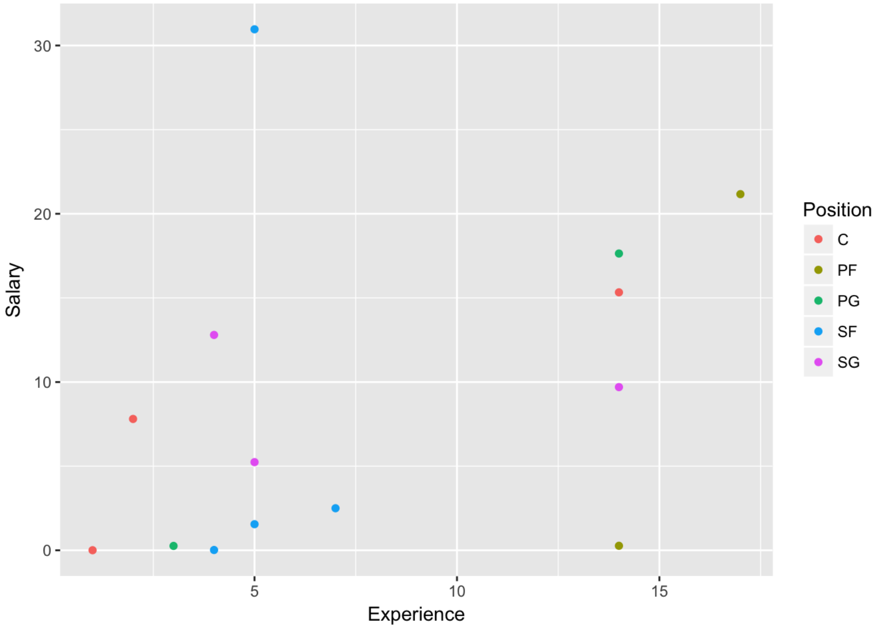
```
boston <- data %>% filter(Team == "BOS")
cleveland <- data %>% filter(Team == "CLE")

ggplot(boston, aes(x = Experience, y = Salary, colour = Position)) +
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



```
ggplot(cleveland, aes(x= Experience, y = Salary, colour = Position)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

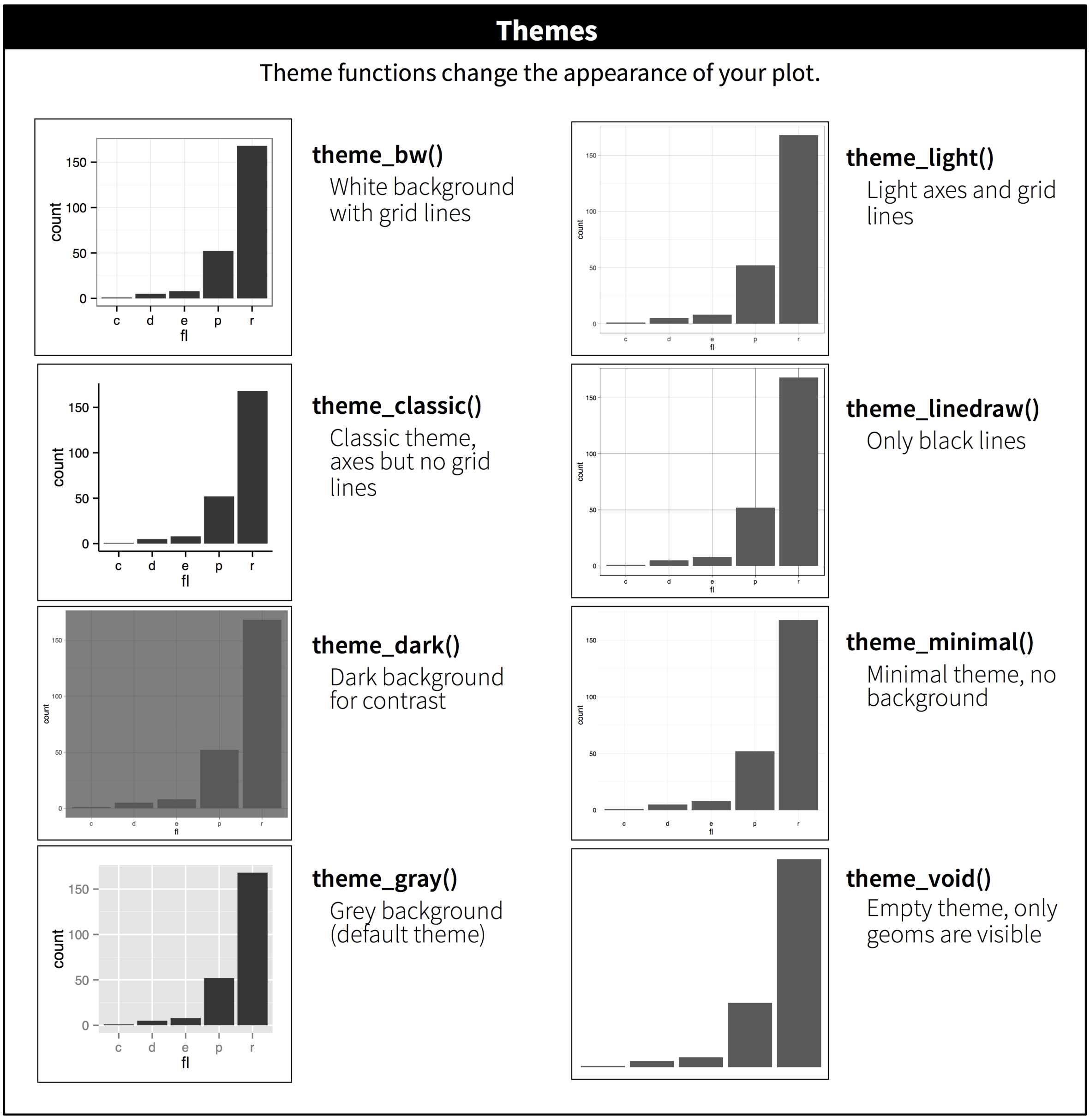


In this case, you can set limits on each scale, called boston and cleveland. You extract two teams of players and plot them separately. However, it is difficult to compare the plots because “all three scales (the x-axis, the y-axis, and the colour aesthetic) have different ranges.”

Themes

You can plot elements with a theme.

ggplot2 includes eight themes by default. Here is an image of theme functions.



Themes

Theme functions change the appearance of your plot.

theme_bw()
White background with grid lines

theme_light()
Light axes and grid lines

theme_classic()
Classic theme, axes but no grid lines

theme_linedraw()
Only black lines

theme_dark()
Dark background for contrast

theme_minimal()
Minimal theme, no background

theme_gray()
Grey background (default theme)

theme_void()
Empty theme, only geoms are visible

However, there are more theme functions in packages like ggthemes.

Do you wonder why the default theme has a grey background? It is because it puts the data forward, stimultaneously making the white grid lines detectable. The grey background "gives the plot a similar typographic colour to the text, ensuring that the graphics fit in with the flow of a document without jumping out with a bright white background."" Finally, the grey background creates "a continuous field of colour which ensures that the plot is perceived as a single visual entity.""

Quiz

Since you've mastered ggplot, you may deal with the quiz I made easily.
1) What is a package name including ggplot()?
2) What is a function name for labels?
3) What is one way of zooming graphs?
4) _____ is used to overrule the default.
5) What are two types of scales?

Answers for the Quiz

1. ggplot2
2. labs()
3.
    - adjusting what data are plotted
    - setting the limits in each scale
    - setting xlim and ylim in coord_caresian()
4. breaks
5. continuous position scales, color scales

Congratulations on completing the quiz! On my next post, you will explore more about ggplot. Take a look at the example below.

References

1. https://www.aridhia.com/technical-tutorials/the-fundamentals-of-ggplot-explained/
2. http://ggplot2.tidyverse.org/reference/labs.html
3. http://r4ds.had.co.nz/graphics-for-communication.html#scales
4. http://colorbrewer2.org/
5. http://ggplot2.tidyverse.org/reference/coord_cartesian.html
6. https://github.com/jrnold/ggthemes
7. http://r4ds.had.co.nz/images/visualization-themes.png

Conclusion

In conclusion, the purpose of this post was for you to master many parts about ggplot. Once you read the post and completed the quiz, I wish you clearly know all the information about ggplot On the next post, you will explore about modeling.