Boxplot with ggplot2

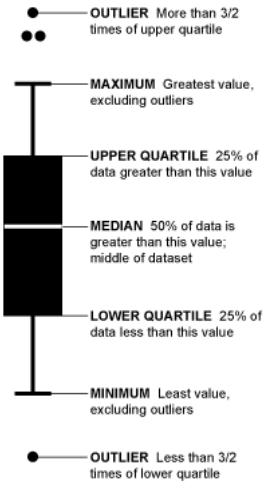
Zhenya Xiao

Introduction of Boxplot

I am going to talk about making boxplot with ggplot2 in this post. The **boxplot** is a standard method to display the distribution of a data set and it is based on the five number summary. Boxplot allows people to visually estimate various features of the data set and it is an efficient tool for data analysis. The name of the boxplot comes from the box in the middle of the graph. (see Boxplot for more information.)

- `minimun` : Least value in the data, excluding outliers.
- `lower quartile` : 25% of the data less than this value.
- `median` : 50% of the data is greater than this value; middle of the data set.
- `upper quartile` : 25% of the data greater than this value.
- `maximum` : Greatest value in the data, excluding outliers.
- `outliers` : Either more than or less than 3/2 times of upper quartile or lower quartile respectively.

Here is a picture to help understand what the different parts of the boxplot mean:



Boxplot Example

Some Examples of Boxplot with ggplot2

In this post, I am going to introduce several types of boxplot using ggplot2 in R. The data I choose to work with is `nba2017-player-statistics.csv` and the main source of the data (see source for more details about the data).

```
library(ggplot2)
setwd("~/stat133/stat133-hws-fall17/post01")
dat <- read.csv("data/nba2017-player-statistics.csv")
dat$Salary <- dat$Salary / 1000000
```

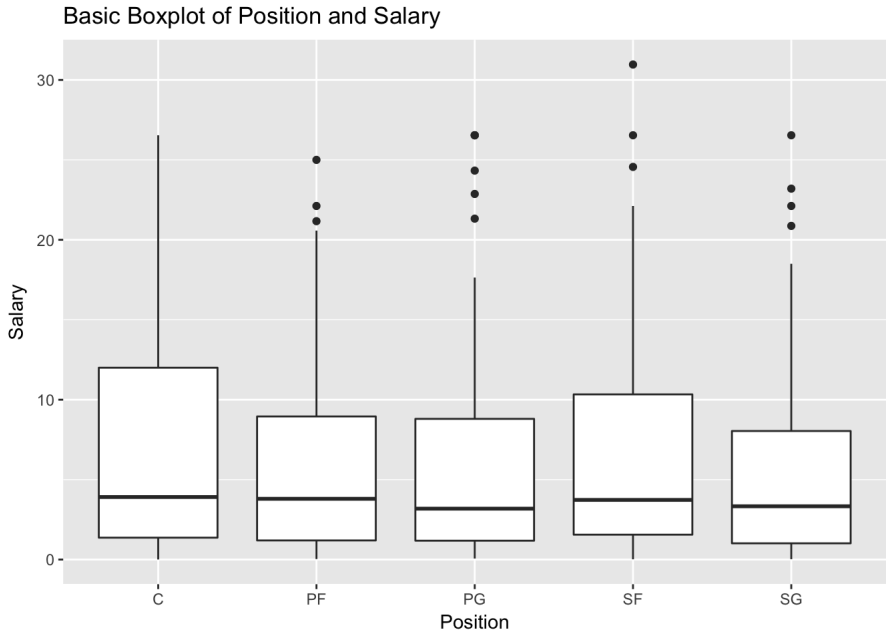
Basic Boxplot with ggplot2

We can make basic boxplot with ggplot2 by using the geom function `geom_boxplot()`. This function works with two variables which are *discrete X* and *continuous Y*.

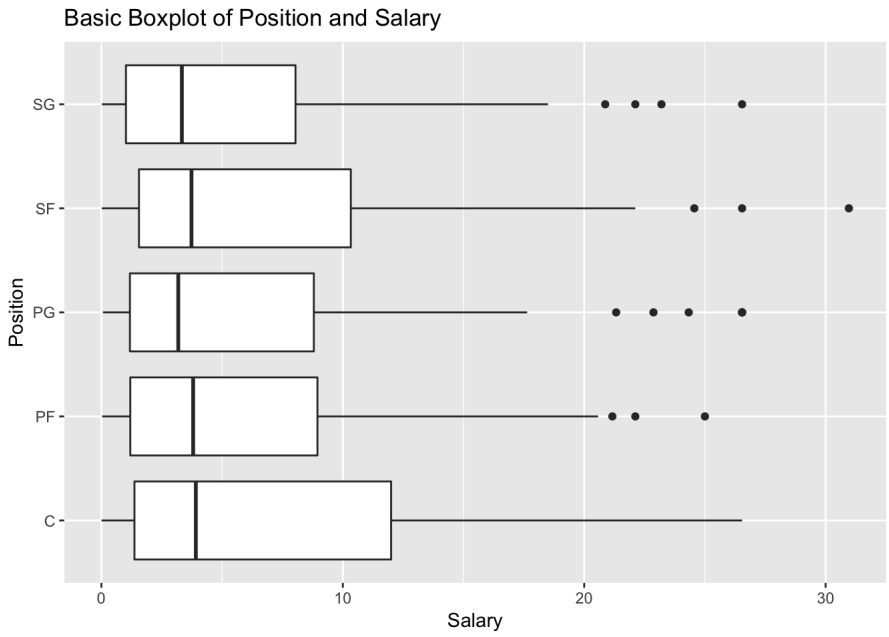
Here is more detail information about `geom_boxplot` function: see geom_boxplot in R documentation website.

Boxplot can be drawn either vertically or horizontally:

```
# a basic boxplot of Position and Salary.
ggplot(data = dat, aes(x=Position, y=Salary)) +
  geom_boxplot() +
  ggtitle("Basic Boxplot of Position and Salary")
```



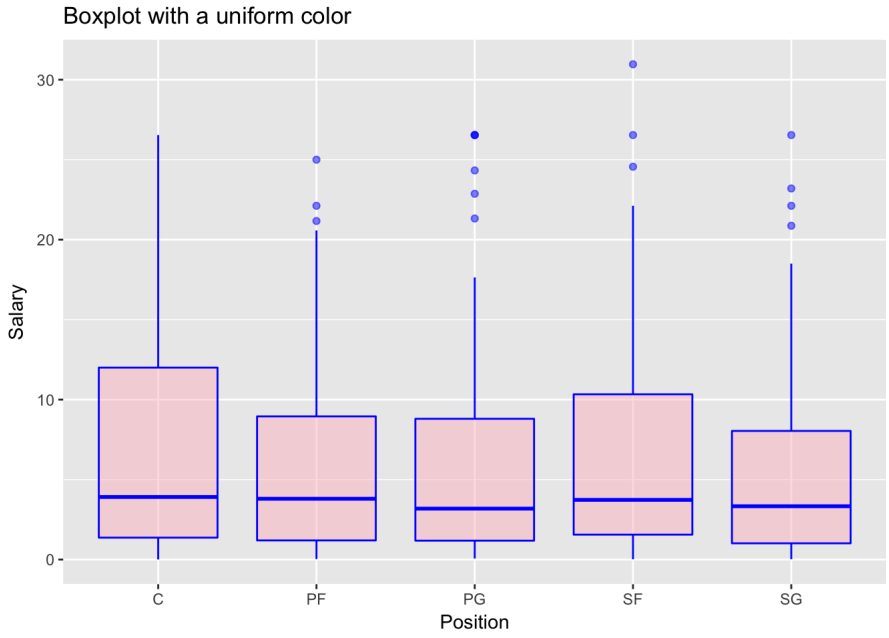

```
# horizontal boxplot
ggplot(data = dat, aes(x=Position, y=Salary)) +
  geom_boxplot() +
  coord_flip() +
  ggtitle("Basic Boxplot of Position and Salary")
```



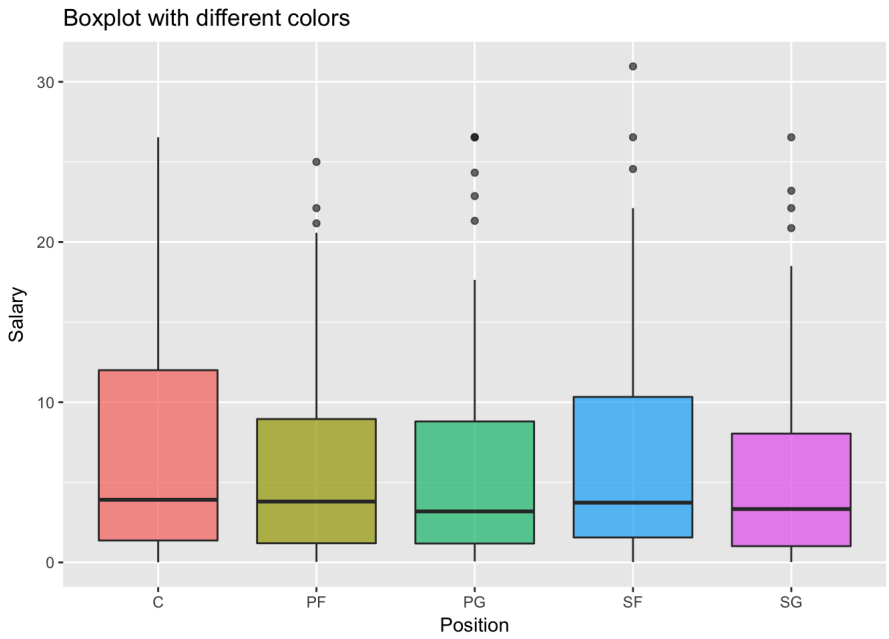
Change Colors of Boxplot with ggplot2

With the arguments `color` and `fill` in the geom function `geom_boxplot()`, we can control boxplot colors and make amazing plots. We can set the same color for the whole boxplot or set different colors for each group. In addition, the degree of transparency in the box fill area could be specified by using the argument `alpha` in `geom_boxplot`. It ranges from 0 to 1.

```
# same color for thw whole boxplot
ggplot(data = dat, aes(x=Position, y=Salary)) +
  geom_boxplot(color = "blue", fill = "pink", alpha = 0.5) +
  ggtitle("Boxplot with a uniform color")
```



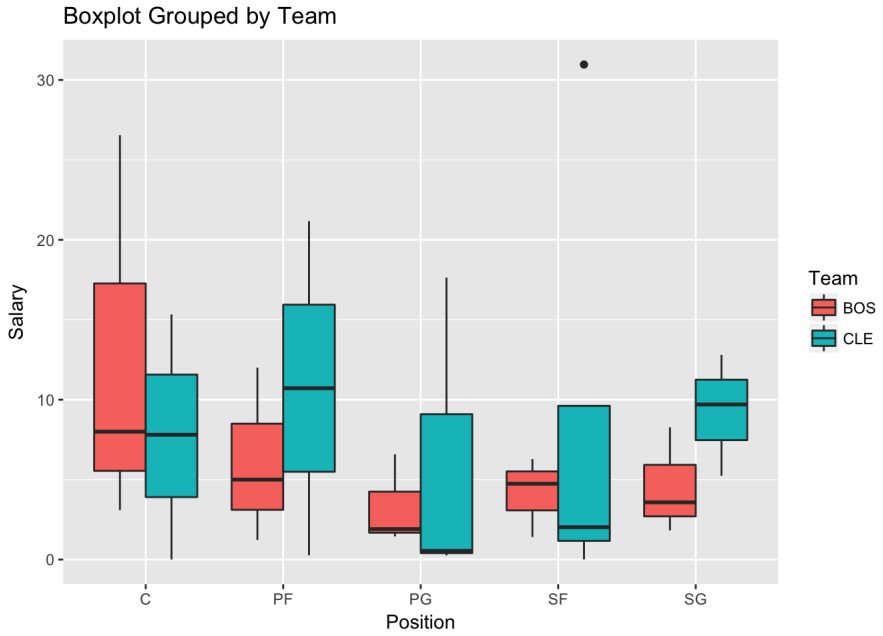

```
# different colors for each group
ggplot(data = dat, aes(x=Position, y=Salary)) +
  geom_boxplot(aes(fill = Position), alpha = 0.7) +
  theme(legend.position = "none") +
  ggtitle("Boxplot with different colors")
```



Grouped Boxplot with ggplot2

Grouped boxplot allow people to divide every category into several groups and plot them together. We just need to change the argument `fill` in `geom_boxplot` to group the data.

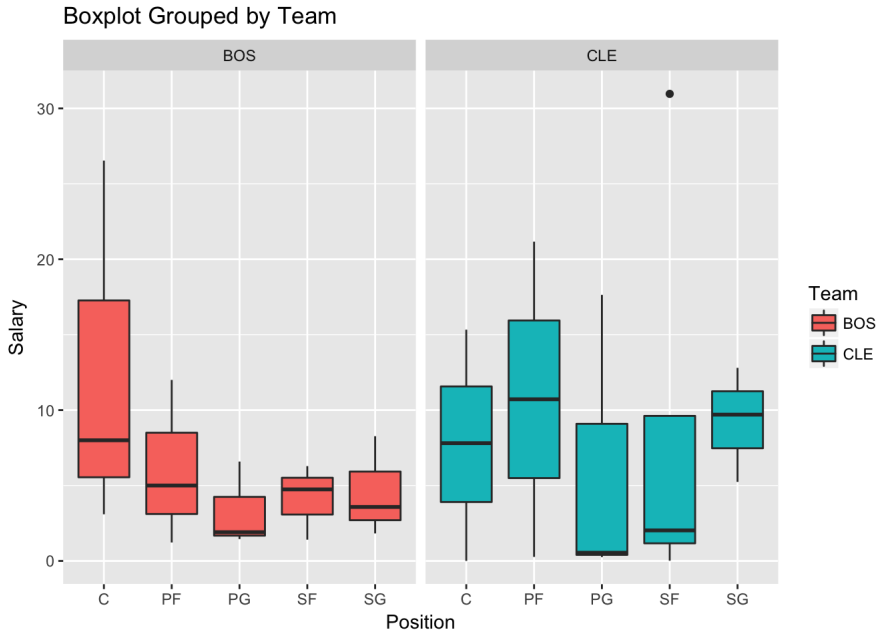
```
# choose teams BOS and CLE for example
dat1 <- dat[dat$Team == "BOS" | dat$Team == "CLE", ]

# grouped boxplot for teams BOS and CLE
ggplot(data = dat1, aes(x=Position, y=Salary)) +
  geom_boxplot(aes(fill = Team)) +
  ggtitle("Boxplot Grouped by Team")
```



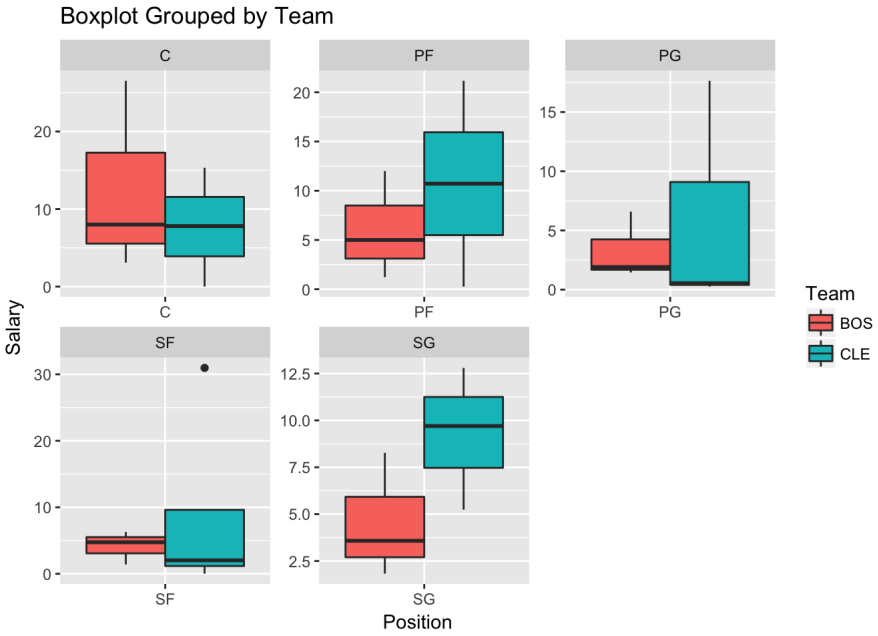
To group data, we can use `faceting` to divide plots into subplots as a good alternative to grouped boxplot. Function `facet_wrap()` will be used.

There are two ways to display grouped boxplots:

```
# one subplot per team
ggplot(data = dat1, aes(x=Position, y=Salary)) +
  geom_boxplot(aes(fill = Team)) +
  facet_wrap(~ Team) +
  ggtitle("Boxplot Grouped by Team")
```




```
# one subplot per position
ggplot(data = dat1, aes(x=Position, y=Salary)) +
  geom_boxplot(aes(fill = Team)) +
  facet_wrap(~Position, scales = "free") +
  ggtitle("Boxplot Grouped by Team")
```

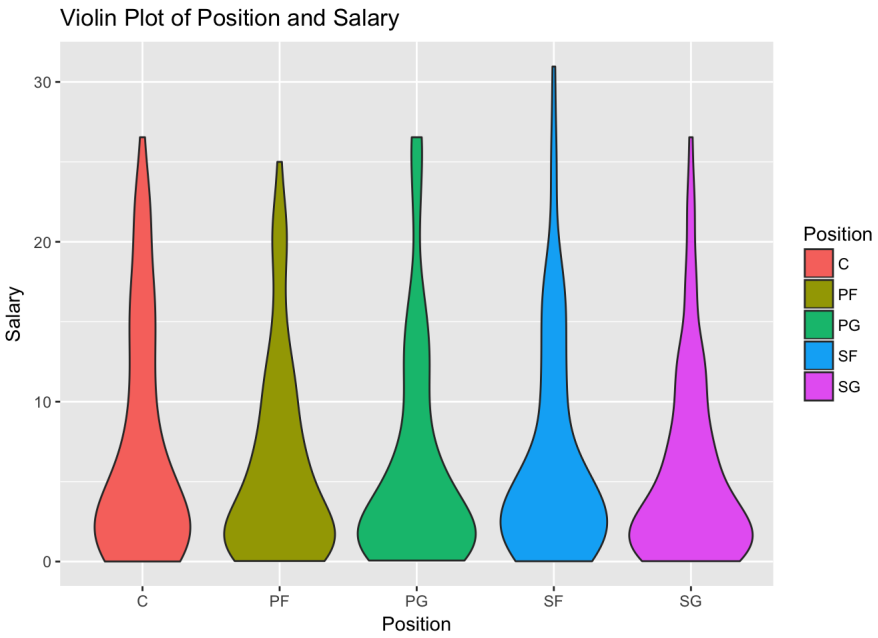


Violin Plot with ggplot2

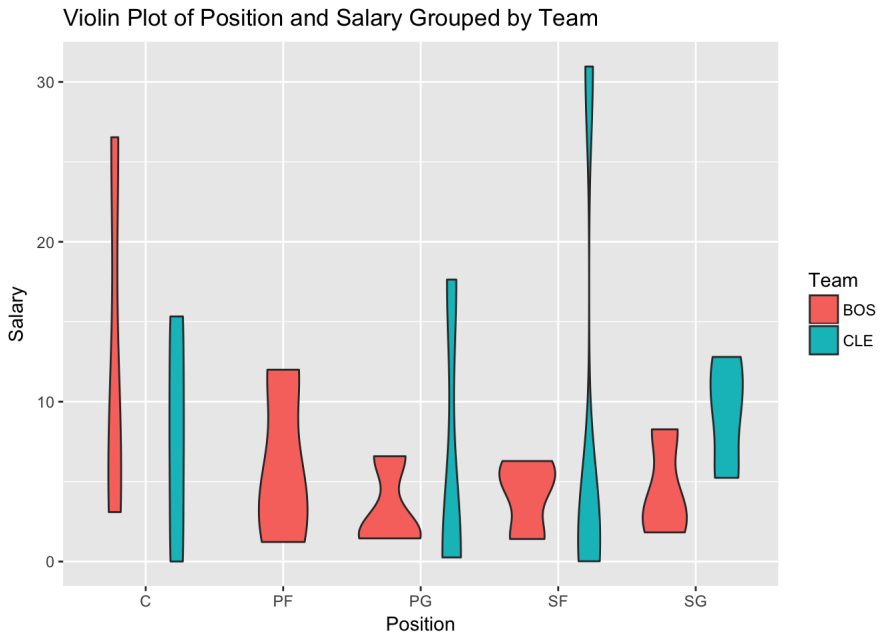
Violin plot is a boxplot with a rotated kernel density plot on each side. It displays the distribution shape of the data and the probability density. The violin plot is a combination of *box plot* and *density plot*. For more information about violin plot, see violin plot.

We use geom function `geom_violin()` to make the violin plot and see geom_violin in R documentation website for detail information about the usage of `geom_violin`.

```
# violin plot
ggplot(data = dat, aes(x = Position, y = Salary)) +
  geom_violin(aes(fill = Position)) +
  ggtitle("Violin Plot of Position and Salary")
```




```
# violin plot grouped by team
ggplot(data = dat1, aes(x = Position, y = Salary)) +
  geom_violin(aes(fill = Team)) +
  ggtitle("Violin Plot of Position and Salary Grouped by Team")
```



Conclusion

Boxplots are excellent and important tool for displaying location and variation changes between different groups of data. We can use the boxplot to check the significance of a factor, and it is also an effective tool for summarizing large quantities of information.

In R program, we can also use R base function `boxplot` to make boxplot besides `geom_boxplot` in ggplot2. Check the website of geom_boxplot in ggplot2 for more examples of boxplot with ggplot2. It is fun to make boxplots with ggplot2 and I love these colorful plots. Come and try to make some amazing boxplots.