Post-01

Christopher Lau

10/31/2017

Data Visualization Labels and Conditional Formatting

It's important to add conditional formatting and labeling in your data visualization. It makes the data stand out better and highlights the points you are trying to make in your data analysis.

In this tutorial, I will be using the package ggplot to show examples.

```
library('ggplot2')
```

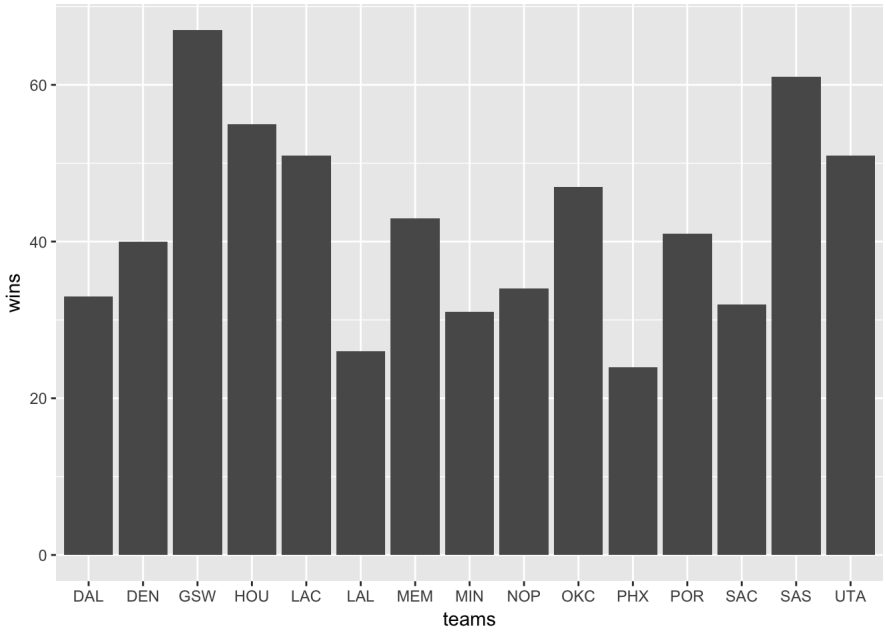
The data I will use is data from the NBA about teams, # of all previous all stars on each team on the 2017 roster, total salary per team, and how many wins each team had by the end of that regular season.

```
teams <- c('LAL', 'GSW', 'HOU', 'MEM', 'LAC', 'OKC', 'DAL', 'PHX', 'POR', 'SAC', 'SAS', 'UTA', 'MIN', 'DEN', 'NOP'
)

allStars <- c(0, 4, 1, 1, 3, 1, 0, 0, 1, 0, 3, 1, 0, 0, 2)

salary <- c(101846550, 99746910, 100052341, 112035324, 1114740032, 96074548, 111991221, 80900983, 113698084, 97345
391, 109221228, 82361096, 83971308, 78785722, 108166365)

wins <- c(26, 67, 55, 43, 51, 47, 33, 24, 41, 32, 61, 51, 31, 40, 34)

df <- data.frame(teams, allStars, salary, wins)

df
```

```
##    teams allStars     salary wins
## 1    LAL        0  101846550   26
## 2    GSW        4   99746910   67
## 3    HOU        1  100052341   55
## 4    MEM        1  112035324   43
## 5    LAC        3 1114740032   51
## 6    OKC        1   96074548   47
## 7    DAL        0  111991221   33
## 8    PHX        0   80900983   24
## 9    POR        1  113698084   41
## 10   SAC        0   97345391   32
## 11   SAS        3  109221228   61
## 12   UTA        1   82361096   51
## 13   MIN        0   83971308   31
## 14   DEN        0   78785722   40
## 15   NOP        2  108166365   34
```

I will make a standard bar graph of wins with ggplot here.

```
ggplot(df, aes(x = teams, y = wins)) + geom_bar(stat='identity')
```



The bar graph looks fine, but it's hard to analyze the data. In statistics, you generally want to compare the data to the mean, and the standard deviation.

```
winsMean <- mean(wins)
winsMean
```

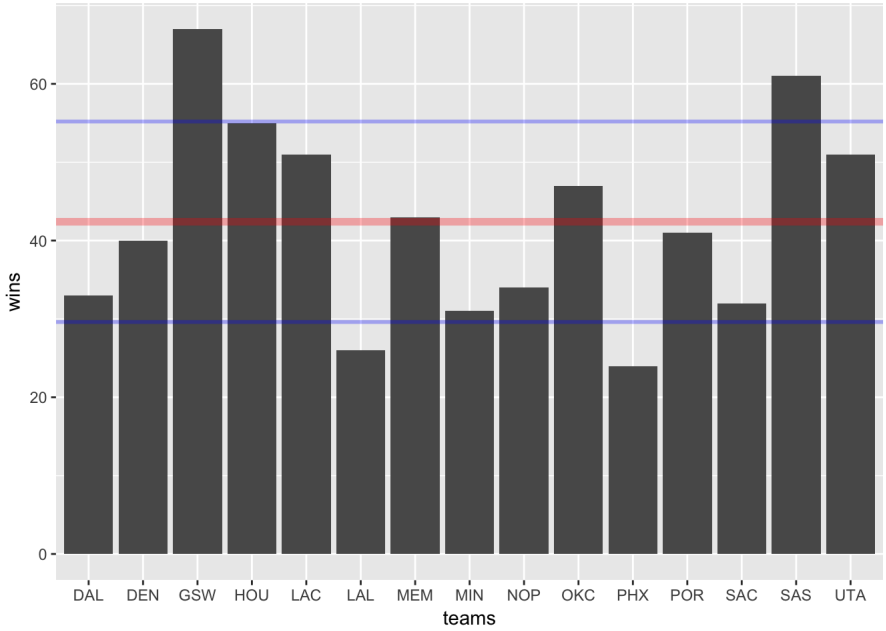
```
## [1] 42.4
```

```
winsSD <- sd(wins)
winsSD
```

```
## [1] 12.79397
```

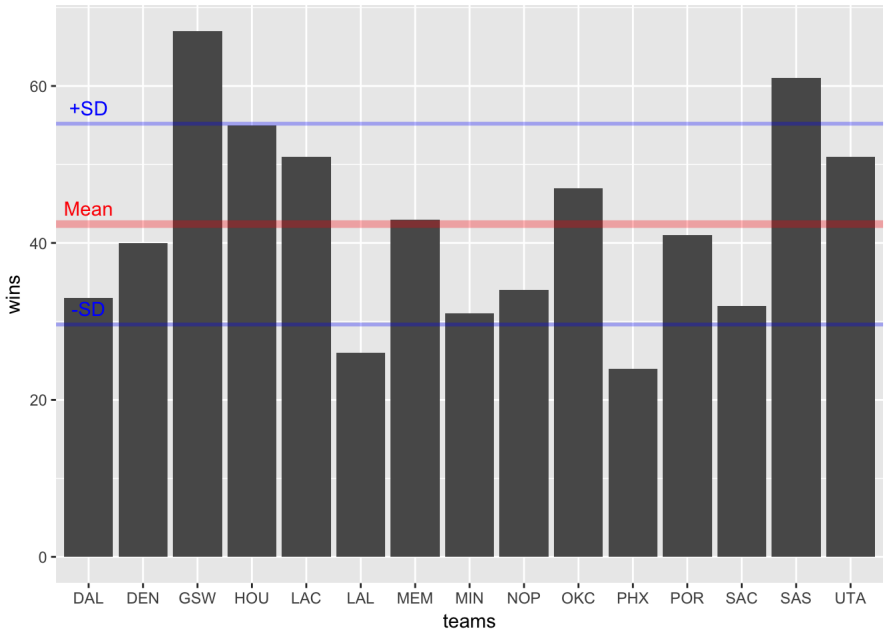
Now lets add the lines to the graph using geom_hline. We want to make the lines transparent so you can see the bars underneath.

```
ggplot(df, aes(x = teams, y = wins)) + geom_bar(stat='identity') + geom_hline(size = 2, color = 'red', alpha = 0.3
, aes(yintercept = winsMean)) + geom_hline(size = 1, color = 'blue', alpha = 0.3, aes(yintercept = winsMean + wins
SD)) + geom_hline(size = 1, color = 'blue', alpha = 0.3, aes(yintercept = winsMean - winsSD))
```



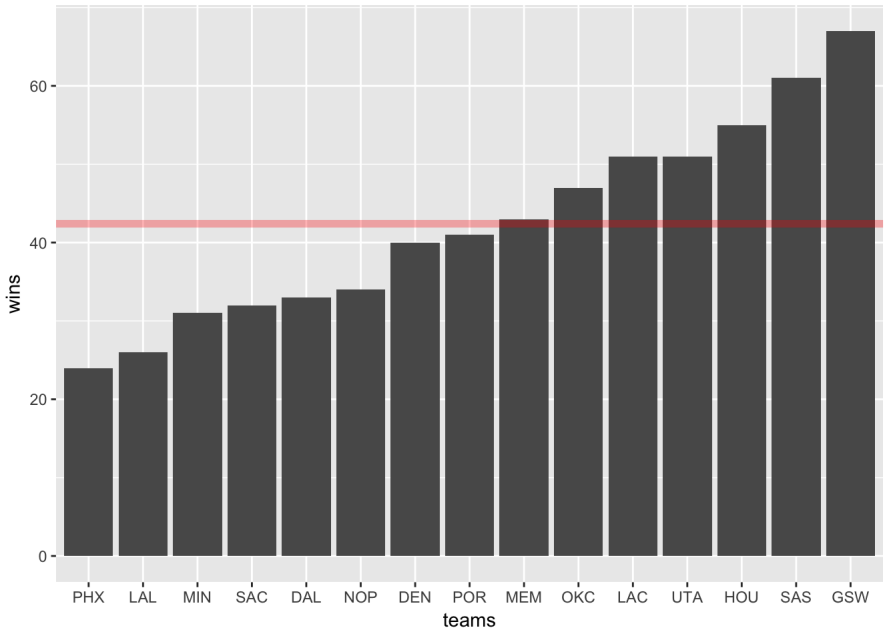
We can also use text to add labels to the lines so we know what the lines mean.

```
ggplot(df, aes(x = teams, y = wins)) + geom_bar(stat='identity') + geom_hline(size = 2, color = 'red', alpha = 0.3
, aes(yintercept = winsMean)) + geom_hline(size = 1, color = 'blue', alpha = 0.3, aes(yintercept = winsMean + wins
SD)) + geom_hline(size = 1, color = 'blue', alpha = 0.3, aes(yintercept = winsMean - winsSD)) + annotate("text", x
= 1, y = winsMean + 2, label = 'Mean', color = 'red') + annotate("text", x = 1, y = winsMean + winsSD + 2, label =
'+SD', color = 'blue') + annotate("text", x = 1, y = winsMean - winsSD + 2, label = '-SD', color = 'blue')
```



In statistics, median is an important number as well. We could find that mathematically or we could display it by sorting the teams by wins.

```
ggplot(df, aes(x = reorder(teams, wins), y = wins)) + geom_bar(stat='identity') + geom_hline(size = 2, color = 're
d', alpha = 0.3, aes(yintercept = winsMean)) + xlab('teams')
```

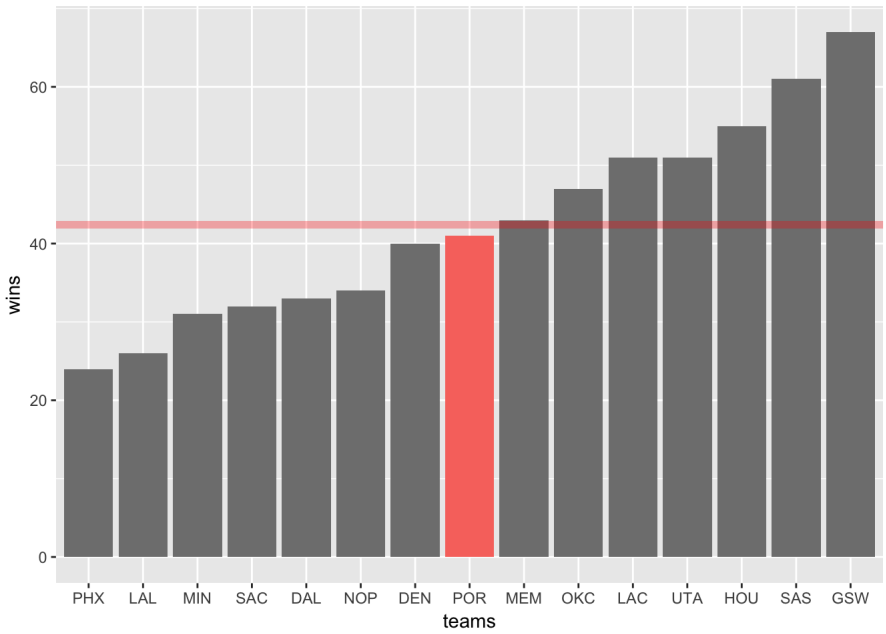


We can compare the mean to the median, but if only we could tell which team was the median straight off the bat. Here we use conditional formatting to highlight which team is the median, using fill in the aes of ggplot.

```
df['winsColor'] <- ifelse(df$teams == 'POR', 'red', NA)
df
```

```
##    teams allStars     salary wins winsColor
## 1    LAL        0  101846550   26      <NA>
## 2    GSW        4   99746910   67      <NA>
## 3    HOU        1  100052341   55      <NA>
## 4    MEM        1  112035324   43      <NA>
## 5    LAC        3 1114740032   51      <NA>
## 6    OKC        1   96074548   47      <NA>
## 7    DAL        0  111991221   33      <NA>
## 8    PHX        0   80900983   24      <NA>
## 9    POR        1  113698084   41       red
## 10   SAC        0   97345391   32      <NA>
## 11   SAS        3  109221228   61      <NA>
## 12   UTA        1   82361096   51      <NA>
## 13   MIN        0   83971308   31      <NA>
## 14   DEN        0   78785722   40      <NA>
## 15   NOP        2  108166365   34      <NA>
```

```
ggplot(df, aes(x = reorder(teams, wins), y = wins, fill = winsColor)) + geom_bar(stat='identity') + geom_hline(siz
e = 2, color = 'red', alpha = 0.3, aes(yintercept = winsMean)) + xlab('teams') + guides(fill=FALSE)
```

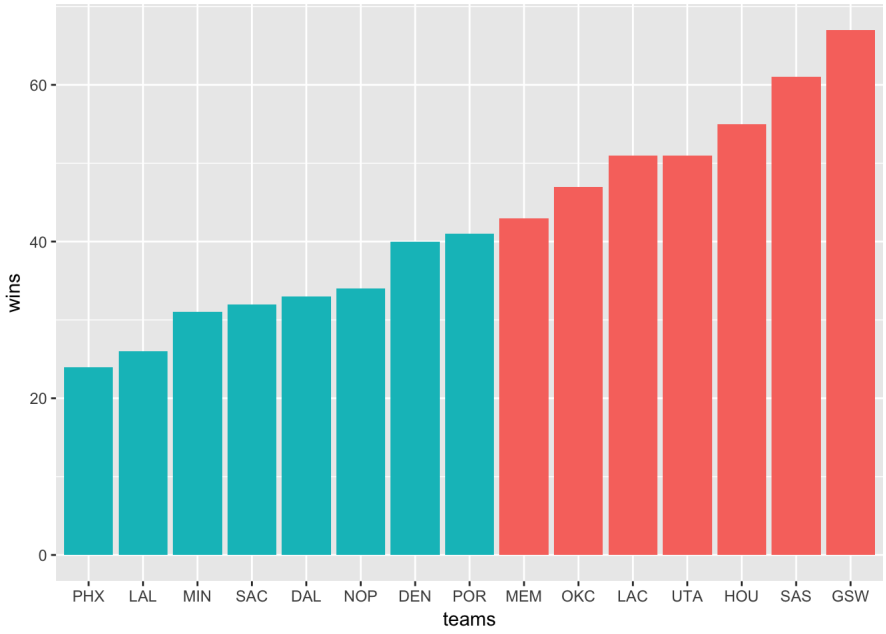


Now we can see which team is the median. All we had to do was add a new column to the data frame with a conditional statement, so R knows which team is the median.

Let's try adding the color blue to every single team that has wins over the mean. Let's color every team below the mean red. Now we don't need

a mean line.

```
df['overMeanColor'] <- ifelse(df$wins > winsMean, 'blue', 'red')
ggplot(df, aes(x = reorder(teams, wins), y = wins, fill = overMeanColor)) + geom_bar(stat='identity') + xlab('team
s') + guides(fill=FALSE)
```



Key tip: use + guides(fill=FALSE) to remove the legend guide that appears when we use fill in ggplot.