post01-yowsean-li

Yowsean Li

10/30/2017

Using Color in Data Visualization

Introduction

In data visualization, color is a powerful tool that can not only improve aesthetics, but also make plots more concise and easier to understand. Color allows for more information to be communicated in the same amount of space, allowing for plots to be more versatile while having little to no drawbacks. However it is important that color is used properly since it can also make plots distracting and harder to read. In this post, we will explore a few examples of how and when to incorporate color into plots and also go over some general rules to follow when choosing colors to use.

Discussion and Examples

In the following discussion and examples, we will be using NBA player statistics from the 2016-2017 season.

Setup

```
# Import packages
library(dplyr)
```

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

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

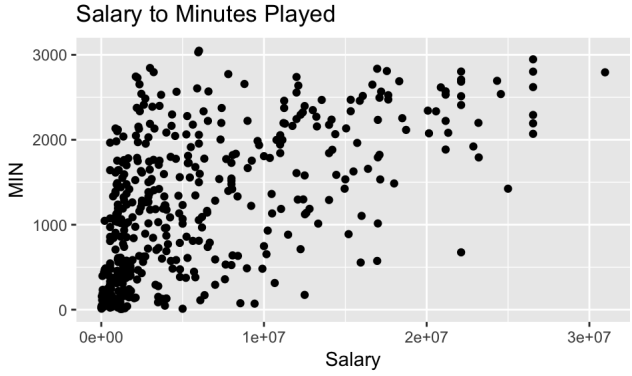
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# Import Data
dat <- read.csv('../data/nba2017-player-statistics.csv', colClasses=c("character", "character", "factor", "character", "double", rep("integer", 19)))
dat2 <- read.csv('../data/nba2017-roster.csv', colClasses=c("character", "character", "factor", rep("integer", 4), "double"))
# Data Cleaning - replace experience 'R' with '0'
dat$Experience[dat$Experience=="R"] <- "0"
dat$Experience <- as.integer(dat$Experience)
# Add RPM Column
dat <- mutate(dat, RPM=(OREB+DREB)/MIN)
```

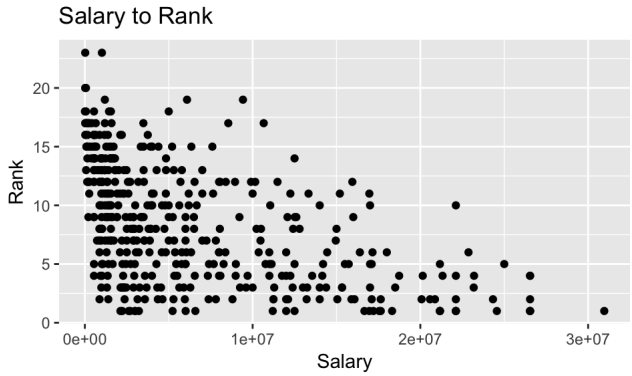
Plotting Three Variables

Say you want to visualize the relationship between three continuous variables. The most naive way of going about this would be to plot two separate plots, each comparing two of the three variables. In the example below, we will plot the relationship between salary, minutes played, and rank (lower is better). The two plots use the naive approach of plotting two separate scatter plots to visualize the relationship between two variables at a time.

```
ggplot(data=dat, aes(x=Salary, y=MIN)) +
  geom_point() +
  ggtitle("Salary to Minutes Played")
```

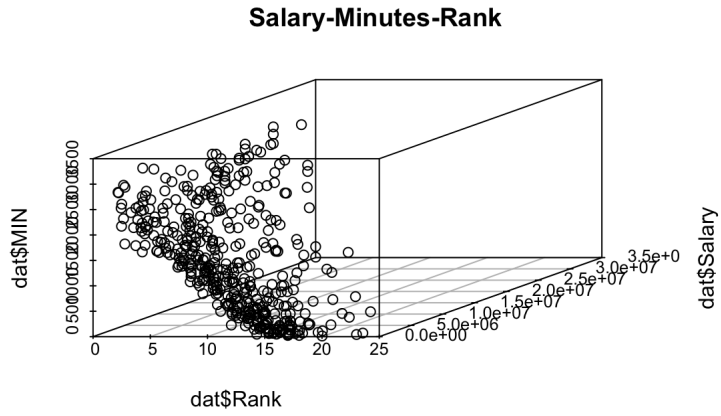



```
ggplot(data=dat, aes(x=Salary, y=Rank)) +
  geom_point() +
  ggtitle("Salary to Rank")
```



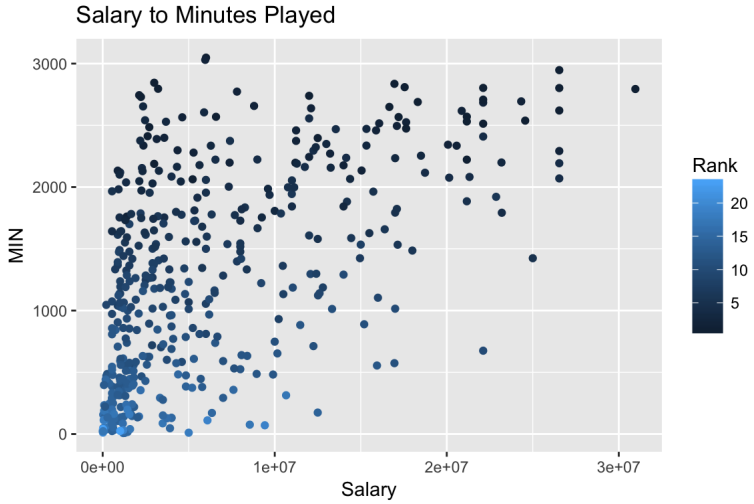
Another method would be to plot the three variables in a three-dimensional scatter plot. However, plotting a three-dimensional graph onto a two-dimensional plane is not always viable since we can only show one specific angle which can make the plot hard to understand. As you can see in the plot below, its hard to see any clear relationships between the variables due to the limited viewing angle.

```
library(scatterplot3d)
scatterplot3d(dat$Rank, dat$Salary, dat$MIN)
title('Salary-Minutes-Rank')
```



With the use of color, we add another dimension to the plot in a more intuitive way. With colors, we can use a continuous gradient of color to represent the values which works well in representing quantitative variables. In this case, information about the rank is incorporated into the color of the points. This is much easier to visualize than the previous examples of plotting two separate plots or using a three-dimensional plot. With this plot, we can easily see the positive relationships between rank, minutes played, and salary, all at once.

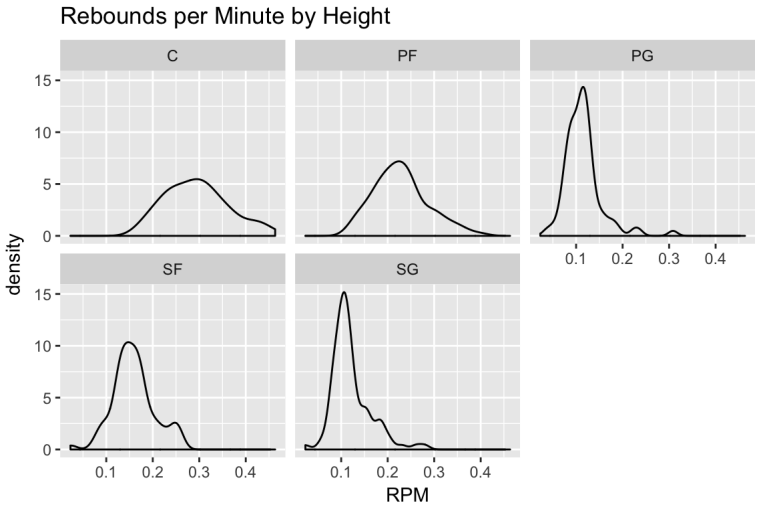
```
ggplot(data=dat, aes(x=Salary, y=MIN, color=Rank)) +
  geom_point() +
  ggtitle("Salary to Minutes Played")
```



Plotting Quantitative and Categorical Data

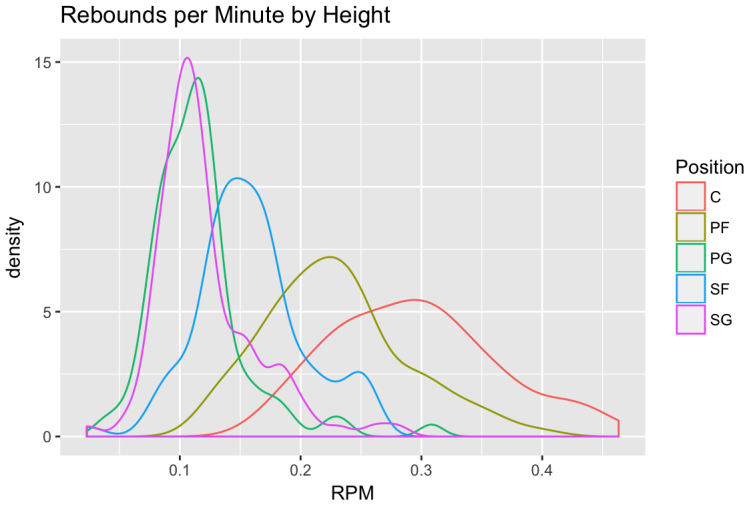
In the next example, we will compare the distribution of rebounds per minute between different positions. As simple way of doing this would be to facet by position and plot each distribution separately. However, with this plot, it is not particularly easy to compare the distributions. We can see the more prominent differences such as the distributions between PG and C. However, the PF and C or PG and SG distributions are more similar and are harder to compare.

```
ggplot(data=dat, aes(x=RPM)) +
  geom_density() +
  ggtitle("Rebounds per Minute by Height") +
  facet_wrap(~ Position)
```



Using color is an effective way of mixing quantitative data with categorical data into plots. By assigning each position a different color, we can plot all five plots from above into one plot, making the data far easier to visualize. With this plot we can clearly see that the PF distribution has a lower mean than the C distribution and we can see detailed differences between the PG and SG distributions.

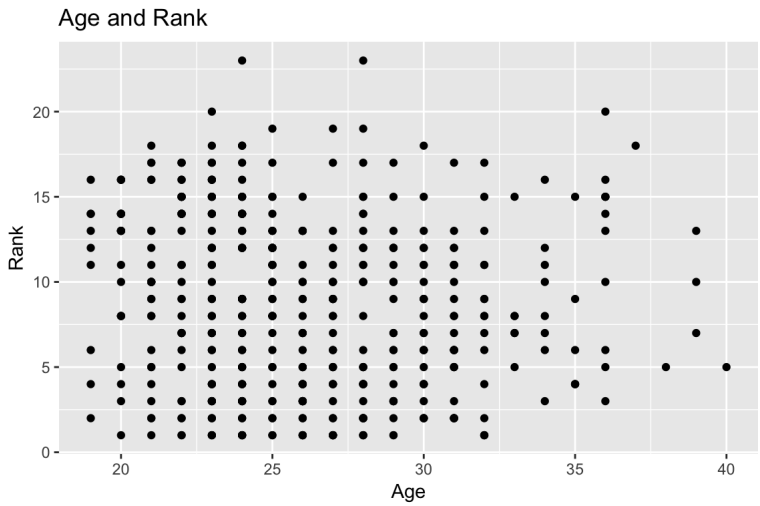
```
ggplot(data=dat, aes(x=RPM, color=Position)) +
  geom_density() +
  ggtitle("Rebounds per Minute by Height")
```



Overplotting

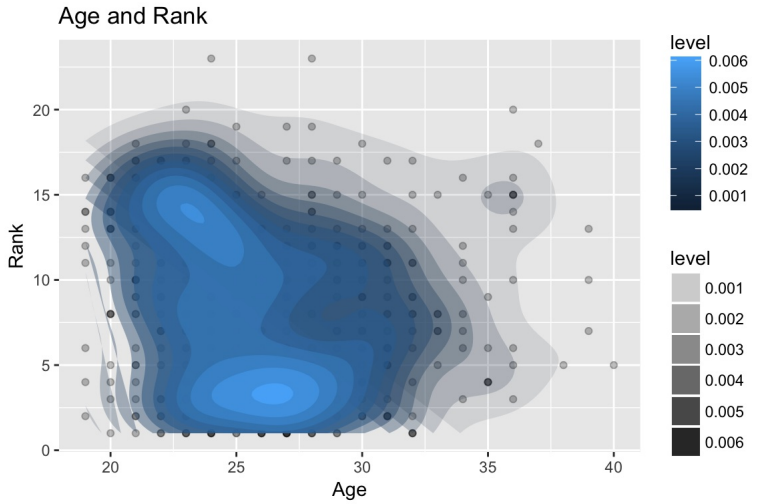
Another case where color can be effective is when plots are overplotted, or when points in a scatterplot overlap. For instance, in the scatterplot below, where we are comparing age to rank, many of the points are overlapped. As a result, the plot is not representative of the data. From the plot it appears that the data is uniformly spread out, but in reality some points should be weighted more than others since some points may be overlapped more than others.

```
ggplot(data=dat, aes(x=Age, y=Rank)) +
  geom_point() +
  ggtitle("Age and Rank")
```



One way of resolving this problem is to use a contour plot, which uses color to represent the density of the points. This gives us additional information that could not be portrayed in the normal scatterplot. In the plot below, we can see the distribution of players based on age and rank. Specifically, we can see that players ranked lower (lower is better) are between ages 25 and 28 and players ranked higher are around 22. This plot gives us a far better understanding of the data than the previous plot.

```
ggplot(data=dat, aes(x=Age, y=Rank)) +
  geom_point(alpha=0.25) +
  ggtitle("Age and Rank") +
  # Transparency and fill determined by level
  stat_density_2d(aes(alpha=..level.., fill=..level..), geom="polygon")
```



Choosing Colors

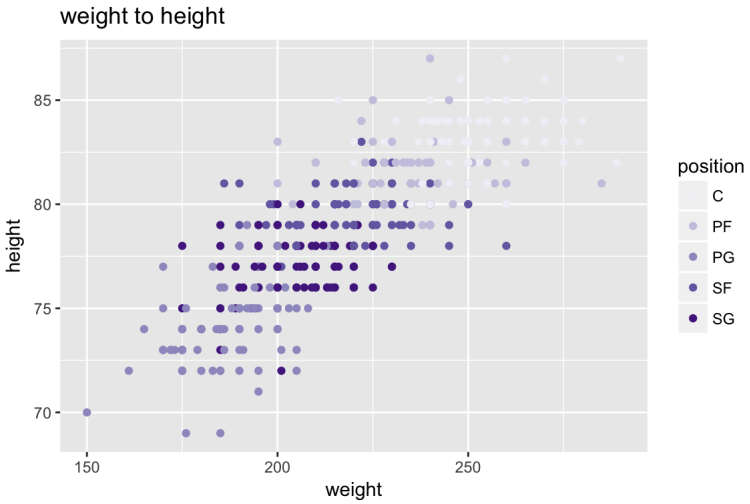
We have seen how color can be used to make plots easier to understand, but how exactly do we choose what colors to use? There are a few general rules to follow when using colors, but for the most part it depends on the context of the data.

In general, it is important to ensure that the colors are visible and not hard to see. For example, make sure not to use light colors on a light background and dark colors on a dark background. Furthermore, it is a good idea to avoid using colors that are distracting, such as bright and obnoxious colors.

In terms of context, it is important to know what type of data is being plotted. For continuous, or quantitative data, it is generally better to use gradients which can be either one color with gradual saturation, or a range across multiple colors. This allows the color to show progression of values. Unlike with quantitative data, with categorical data we want to see clear distinction between categories, so contrasting colors should be used to clearly differentiate the categories.

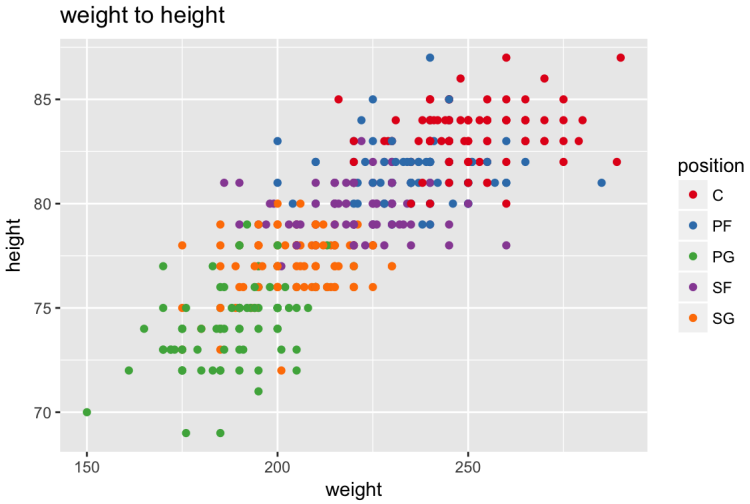
The following is an example of how choosing colors is important in making sure the plot is readable. In both plots, colorbrewer palettes were used. In the plot below, since the categorical data is represented by one color on a gradient, it is hard to differentiate between the distinct categories. Furthermore, the colors for C and PF are too light to see on the light gray background.

```
ggplot(data=dat2, aes(x=weight, y=height, color=position)) +
  geom_point() +
  ggtitle("weight to height") +
  scale_color_brewer(palette="Purples")
```



The plot below solves the problems from the previous plot, as the colors darker making it very easy to see on the light background. Furthermore, the colors for the different colors are all contrasting, making it very easy to see where the points on the scatterplot are grouped.

```
ggplot(data=dat2, aes(x=weight, y=height, color=position)) +
  geom_point() +
  ggtitle("weight to height") +
  scale_color_brewer(palette="Set1")
```



Conclusion

The main takeaway from this post is that color can make data visualization much more effective if done correctly. We have seen several cases in which colors allow for more data and information to be communicated in a plot. Furthermore, we have covered some general rules of thumb in terms of choosing colors that are easy to view and make intuitive sense.

References

- https://infogram.com/blog/color-theory-dos-and-donts-for-data-visualization/
- https://lisacharlotterost.github.io/2016/04/22/Colors-for-DataVis/
- https://stats.stackexchange.com/questions/31726/scatterplot-with-contour-heat-overlay
- https://www.r-statistics.com/2016/07/using-2d-contour-plots-within-ggplot2-to-visualize-relationships-between-three-variables/
- http://ggplot2.tidyverse.org/reference/geom_point.html
- http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually#gradient-colors-for-scatter-plots
- https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf