post01

Albert Lee

10/25/2017

Title: Using `ggplot2` to code and analyze different plots of **correlation** to verify if correlation exists between experience and points within the dataset `nba2017-players.csv`

Introduction

Hi, there! Before I start with this introduction, I just want to say thank you for spending time reading my post. You may or may not find it interesting, but the whole purpose of this post as well as all of our posts is to expand our knowledge in R and to have some fun learning and analyzing data. I hope you enjoy my post and give me a critical feedback at the end!

A few weeks ago, we had the opportunity to encounter a very sophisticated package that is embedded in RStudio, `ggplot2`. We also had a hands-on experience, manipulating data and displaying datasets on plots using `ggplot2` and other supportive packages. The purpose of this post is to further dive into the world of `ggplot2` and explore some more fascinating functions that we might not have been able to cover in our labs and homework.

Throughout this post, you will be able to read and learn codes that will allow you to explore the visual and analytical tools of `ggplot2`. I have decided to use the dataset of `nba2017-players` not only because we already have familiarity with using this dataset but also as a huge basketball fan, I want to take this opportunity to solidly verify whether there is any correlation between experience of a player and number of points a player scored in NBA season 2017.

To sum up the introduction, I will be using `ggplot2` to code and analyze different plots that can be used to show interesting correlations that will eventually sum up to verify that whether the correlation exists between the experience and the points scored. This post will serve largely as an educational tool for readers to enjoy learning and thinking through the analytical aspects of NBA dataset.

My biggest inspiration came from two amazingly designed and formatted R websites

- Reference for my inspiration behind this post
  - The main source: R-bloggers
  - Secondary source: r-statistics

Preparations

Remember that it is important to load within RMarkdown not in console

```
# loading necessary package ggplot2
library(ggplot2)

# setting up dataframe for nba2017-players.csv file
github <- "https://github.com/ucb-stat133/stat133-fall-2017/raw/master/"
csv <- "data/nba2017-players.csv"
download.file(url = paste0(github, csv), destfile = 'nba2017-players.csv')
dat <- read.csv('nba2017-players.csv', stringsAsFactors = FALSE)
```

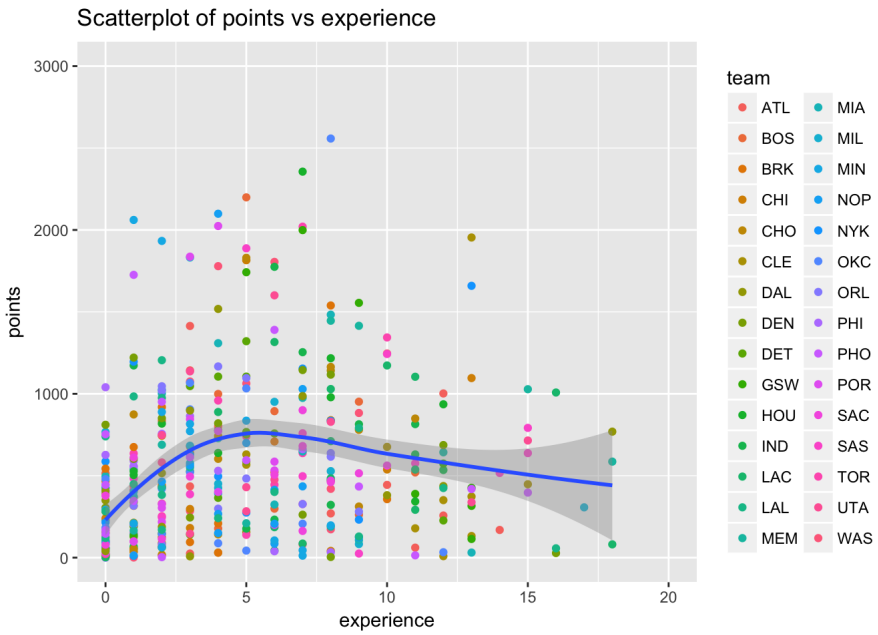
Correlations

Finding correlation between two variables is important in the world of statistics as it can help us understand to what extent two variables are related. In R, there is a function `plot()`, but `ggplot()` is much easier to deploy and perform on csv datasets.

1. Scatterplot

The most frequently used plot for data analysis is scatterplot. Scatterplot analyzes the nature of relationship between two variables. As we learned it before, we can use `geom_point()` to plot the points of all players. Additionally, `geom_smooth()`, which draws a loess smooth line can be altered to make it to the line of best fit.

```
# loading necessary package
library(ggplot2)

# To get rid of scientific notation (e.g 1.4e+24)
options(scipen=999)

# scatterplot of points vs experience
gg_1 <- ggplot(data = dat, aes(x = experience, y = points)) +
  geom_point(aes(x = experience, y = points, col = team)) +
  geom_smooth(method = 'loess') + xlim(c(0, 20)) + ylim(c(0, 3000)) +
  ggtitle('Scatterplot of points vs experience')

plot(gg_1)
```



Analysis: The above scatterplot is do-able with a good understanding of lab materials we have covered in this class. However, what I am interested is to indicate that there is a set of points(i.e players) which deviate from the smooth lowess line (also known as anomalies).*But how can you show that on the graph without describing in words?* That is the next task.

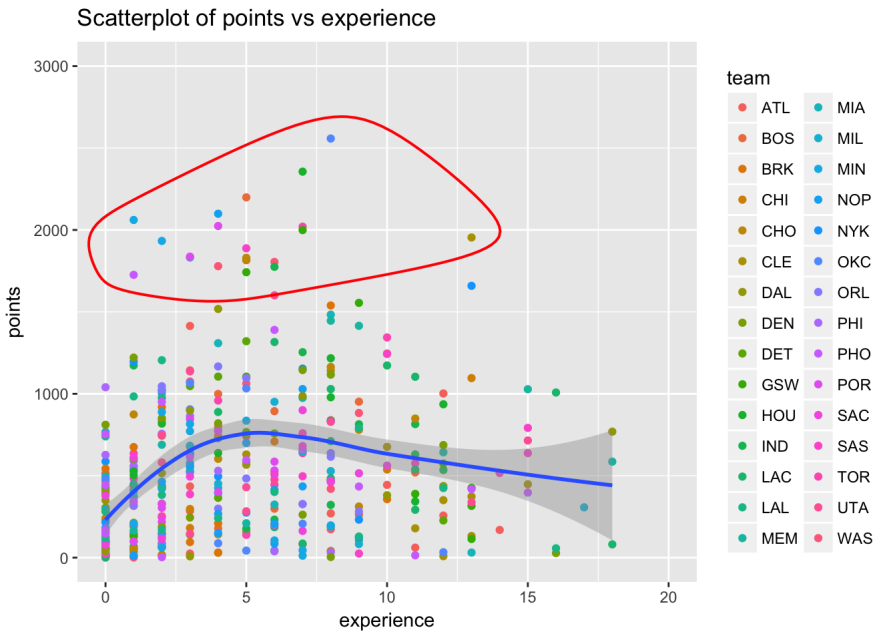
2. Scatterplot with encircling

From my research, I figured that there is a function `geom_encircle()`, which can help me achieve what I mentioned in the previous paragraph. The `geom_encircle()` function is within the `ggalt` package, so you need to download the `ggalt` package before you can use the function.

Within `geom_encircle()`, you need to first set your original data, `dat`, to a new dataframe that includes only the points of your interest. To expand this encircle of certain data points, simply increase the range of both variables(x and y axis) of your encircle.

- Reference for learning
  - The main source: ggalt examples

```
# loading necessary package ggalt
library(ggalt)

# Creating a new dataframe of the encircling points of interest
dat_encircle <- dat[dat$experience >= 0 & dat$experience <= 15 &
                      dat$points > 1700 & dat$points < 2700, ]

# scatterplot of points vs experience
gg_2 <- ggplot(data = dat, aes(x = experience, y = points)) +
  geom_point(aes(x = experience, y = points, col = team)) +
  geom_smooth(method = 'loess') + xlim(c(0, 20)) + ylim(c(0, 3000)) +
  ggtitle('Scatterplot of points vs experience') +

# and now add geom_encircle() function
  geom_encircle(data = dat_encircle, aes(x = experience, y = points), color = 'red', size = 2)

plot(gg_2)
```



Analysis: `geom_encircling()` function is a convenient tool to draw a boundary around the set of anomalies that are off the lowess smooth line. From both the lowess smooth line and the encircling, we can see that players who are in between 4-6 years of experience achieve the highest number of points. This is probably true as those players are now entering what is known as 'peak' performance (after 4 years of NBA rookie season) and are fully equipped with set of skills that have been developed in first 4 years. As experience accumulates beyond 8 years, the lowess smooth line has a negative slope but not too steep. This is probably due to the fact that there are much less number of players who have played in a competitive NBA league for more than 10 years. Thus, the range of highest score and lowest score in players of experience less than 10 is much greater than that in players of experience greater than 10.

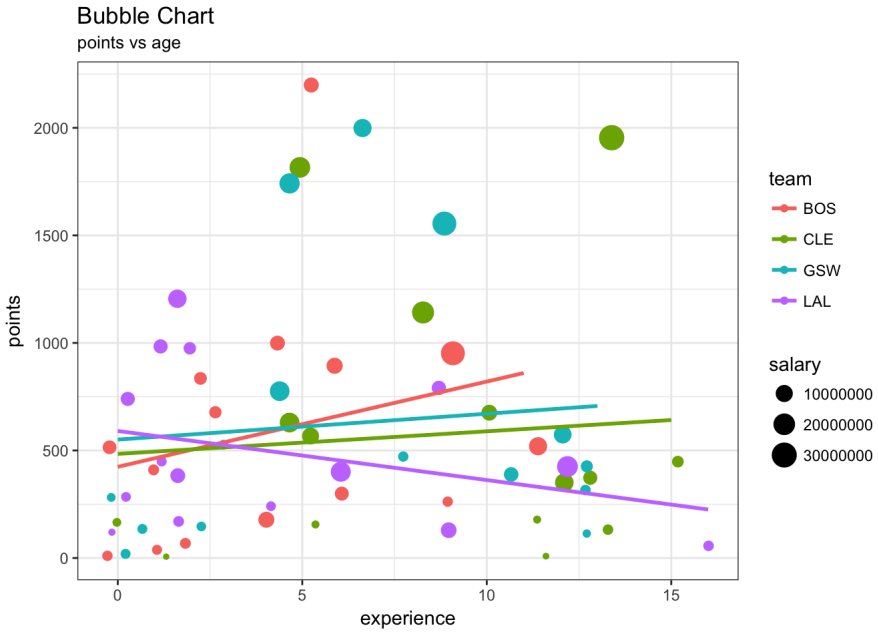
3. Bubble plot

The scatterplot and the scatterplot with encircling largely served to analyze the correlation between two numeric variables (points scored vs experience). *What if you want to have more variables displayed within the same scatterplot with attrative visuals?* Bubble plot is one tool you can use within the `ggplot2` package. Bubble plot allows you to add two more variables: one categorial which is indicated by its color and another continuous numeric variable which is indicated by the size of the bubble as will be shown below.

In furtherance to the two previous variables which I analyzed in the scatterplots above, I decided to factor in `salary` and `team` to illustrate an example of a bubble plot. For `team`, I decided to pick four of my favorite NBA teams `CLE`, `GSW`, `LAL`, `BOS`, for a concise and clear image.

- Reference for learning
  - The main source: Bubble Charts in R

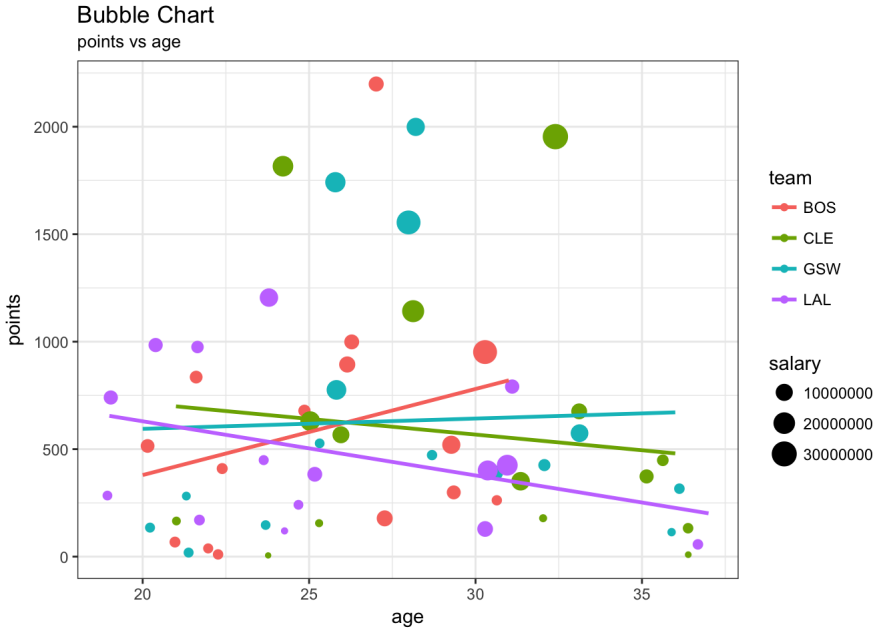
```
# loading necessary packages
library(ggplot2)

# data preparation
dat_bubble <- dat[dat$team %in% c("CLE", "GSW", "LAL", "BOS"), ]
theme_set(theme_bw()) # pre-set the background theme of the plot

# using ggplot2
gg_bubble <- ggplot(dat_bubble, aes(x = experience, y = points)) + geom_jitter(aes(col = team, size = salary)) + g
eom_smooth(aes(col = team), method = 'lm', se = F) + labs(title = "Bubble Chart", subtitle = "points vs age")

plot(gg_bubble)
```



Analysis: It can be shown that BOS, CLE, and GSW have fairly positive slope while LAL has a higher degree of negative slope. This could be fundamentally interpreted as that the experienced players in BOS, CLE, and GSW are contributing more in points for the team than those in LAL. However, this basic analysis is subject to misinterpretation as there are many deviations of big bubbles (indicating highly paid players) from the smooth lines.

I decided to see if there is any difference betweeen NBA experience and age. Below is another bubble chart that factors in age rather than experience as an independent variable.

```
# loading necessary packages
library(ggplot2)

# data preparation
dat_bubble <- dat[dat$team %in% c("CLE", "GSW", "LAL", "BOS"), ]
theme_set(theme_bw()) # pre-setting the background theme

# ggplot
gg_bubble <- ggplot(dat_bubble, aes(x = age, y = points)) + geom_jitter(aes(col = team, size = salary)) + geom_smo
oth(aes(col = team), method = 'lm', se = F) + labs(title = "Bubble Chart", subtitle = "points vs age")

plot(gg_bubble)
```



Analysis: Interestingly, the slope of the smooth line changed for the team `CLE` from positive to negative when I changed the x-axis from experience to age. What this means is that there are playeres in team `CLE` who are relatively younger in NBA expeience but older in real age (scaled). This could be due to the fact that they had an early debut (probably right after their fresman year at college). The slope for other teams look roughly the same.

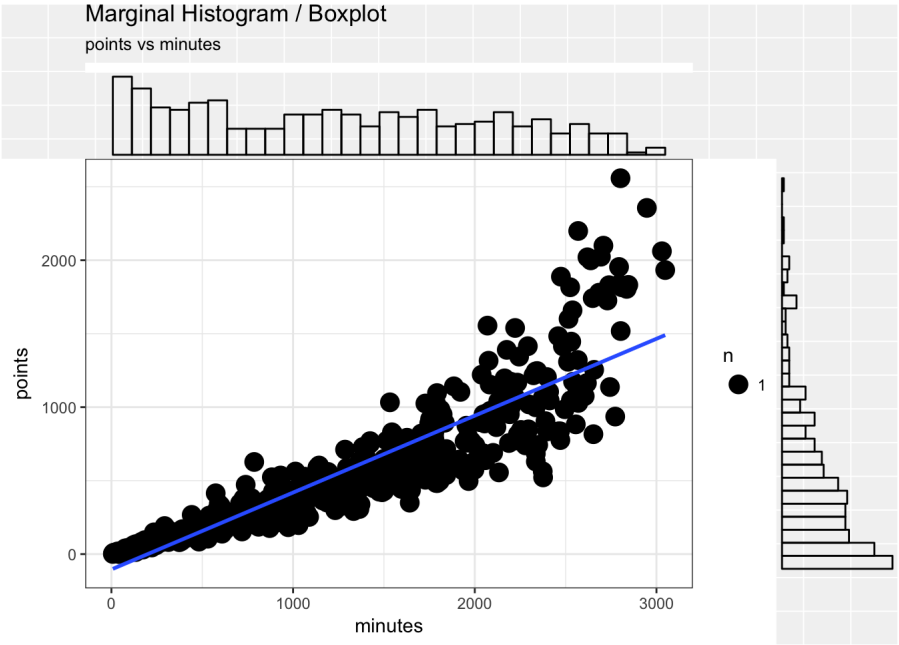
4. Marginal Histogram / Boxplot

Marginal Histogram and Boxplot is a little more advanced than the previous tools we used to analyze. It requires a little more data manipulation and knowledge of how to use the functions in different package. The motivation behind this tool is I wanted to have the distribution and the scatterplot/boxplot displayed on the same page in both axes. I did this using the function called `ggMarginal` from the package `ggExtra`

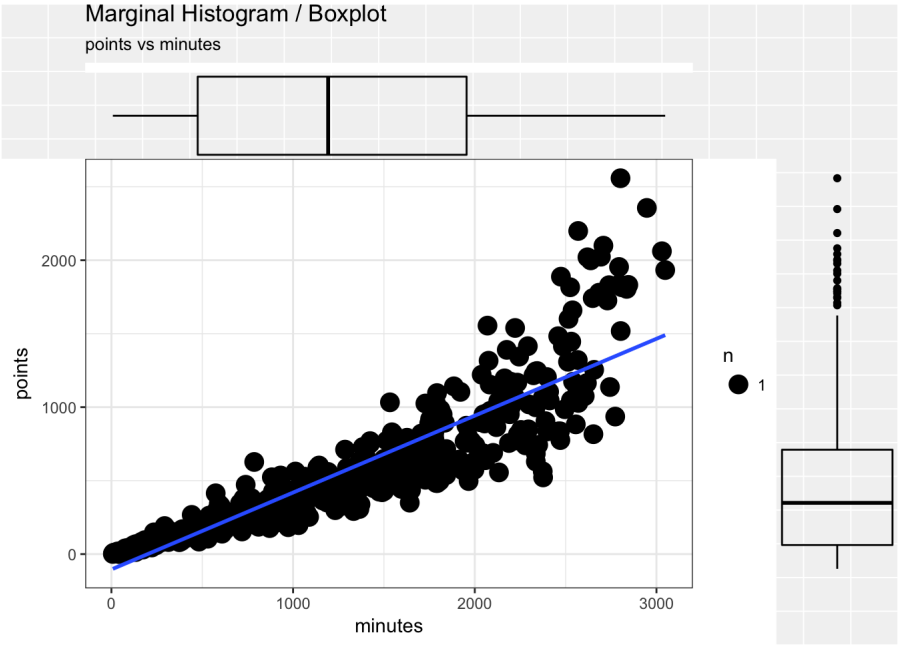
To make an effective use of `ggpMarginal` function, I decided to analyze the correlation between minutes of game played and the points, a correlation that is very obvious since you would expect more points to be scored if the player played more minutes of game.

- Reference for learning
  - The main source: ggExtra
  - Secondary source: CRAN ggExtra

```
# loading necessary packages
library(ggplot2)
library(ggExtra)

theme_set(theme_bw()) # pre-setting the background theme

# data preparation
dat_marginal <- dat[dat$minutes >= 0 & dat$points >= 0, ]
gg_marginal <- ggplot(dat, aes(x = minutes, y = points)) +
  geom_count() +
  geom_smooth(method="lm", se=F) + labs(title = "Marginal Histogram / Boxplot", subtitle = "points vs minutes")

# plotting Marginal histogram
plot(ggMarginal(gg_marginal, type = "histogram", fill="transparent"))
```

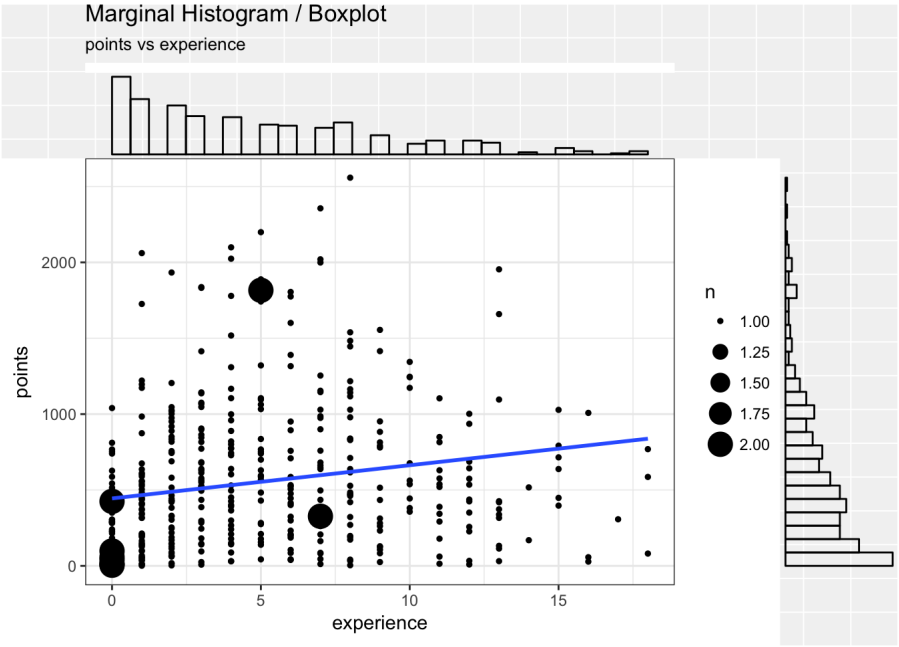



```
# plotting Marginal boxplot
plot(ggMarginal(gg_marginal, type = "boxplot", fill="transparent"))
```

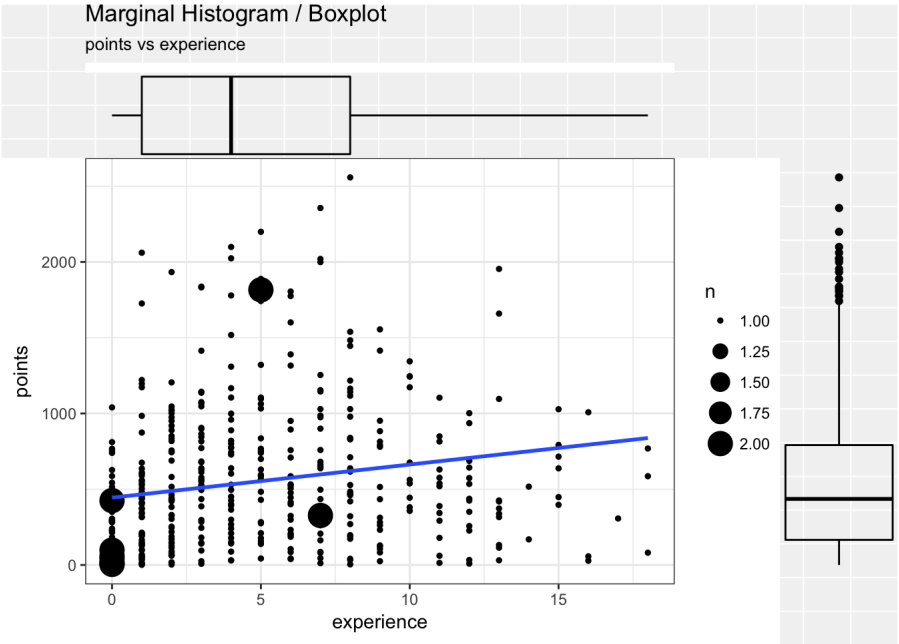


Should we check out what would happen if we use our previous x and y variables, points vs experience?

```r
# Marginal histogram and boxplot points vs experience

# loading necessary packages
library(ggplot2)
library(ggExtra)

theme_set(theme_bw()) # pre-setting the background theme

# data preparation
dat_marginal <- dat[dat$experience >= 0 & dat$points >= 0, ]
gg_marginal <- ggplot(dat, aes(x = experience, y = points)) +
  geom_count() +
  geom_smooth(method="lm", se=F) + labs(title = "Marginal Histogram / Boxplot", subtitle = "points vs experience")

# plotting Marginal histogram
plot(ggMarginal(gg_marginal, type = "histogram", fill="transparent"))
```




```r
# plotting Marginal boxplot
plot(ggMarginal(gg_marginal, type = "boxplot", fill="transparent"))
```



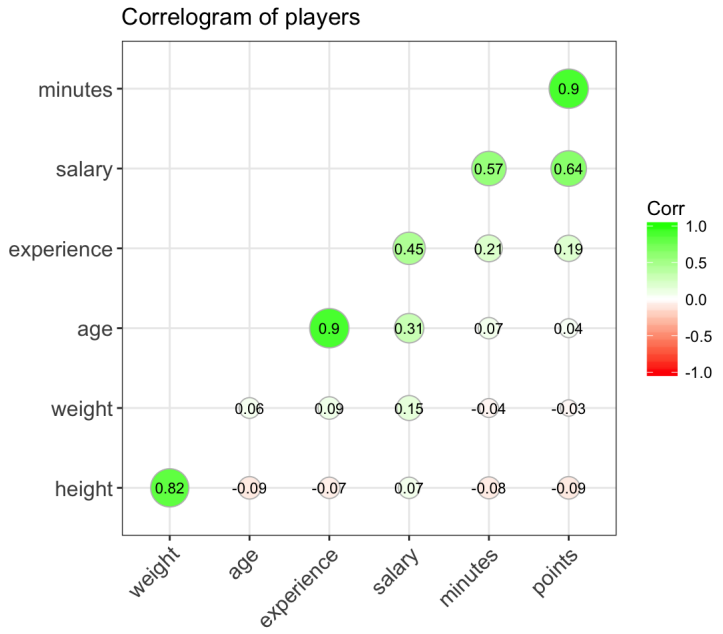
Analysis: It is now getting pretty obvious that the correlation between experience and points is weak to support.

5. Correlogram

The last tool I will be using to analyze the correlation between multiple continuous variables is correlogram. Correlogram is an advanced technique/function that also needs to be summoned from a new package `ggcorrplot`. This will give correlation coeffcients for many numeric variables within the dataset `nba2017-players.csv`

- Reference for learning
  - The main source: Correlogram

```
# loading necessary packages
library(ggplot2)
library(ggcorrplot)

# Data preparation
dat2 <- data.frame('height' = dat$height, 'weight' = dat$weight, 'age' = dat$age, 'experience' = dat$experience, '
salary' = dat$salary, 'minutes' = dat$minutes, 'points' = dat$points)
cor_dat2 <- round(cor(dat2), 2)

# ggplot
ggcorrplot(cor_dat2, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
           colors = c("red", "white", "green"),
           title= "Correlogram of players",
           ggtheme = theme_bw)
```



Analysis: Although we previously predicted that the correlation between points and experience is very weak, it surprises me that the correlation between these two variables is recorded to be 0.64. Although correlation does not necessarily imply causation, taking into account that many other variables revolve around 0 correlation coefficient, it might be feasible to say there is more to the correlation between points and experience than what meets the eye!

Evaluation

Throughout this post, we were able to learn and analyze 5 different correlation charts. We started off with the basic scatterplot that hopefully served a good revision, and we incorporated encircling property to indicate there were players who deviated from the lowess smooth line by a lot. Then we were able to move on to an intermediate level bubble plot, a new plot that confused us with many dispersed dataset. The marginal histogram / boxplot gave us some new insights into how we can utilize `ggplot2` and other corresponding packages to display both the distribution and the correlation on the same chart. The correlogram, however, surprised me as it showed a significantly higher correlation coefficient between experience and points than what I have previously predicted. I have effectively come to a conclusion that there is a undeniable correlation between experience of a NBA player and points they scored in NBA season 2017.

I hope you enjoyed my post and thank you for reading!