# Explore & Summarize Data: For Baseball

## Summary

I decided for the Baseball data set including the basic data of 1157 baseball players. In the appendix you see further analysis of the data in R. I looked at the features, the distributions and correaltions to define a storyline. The two main additions/changes I made are:

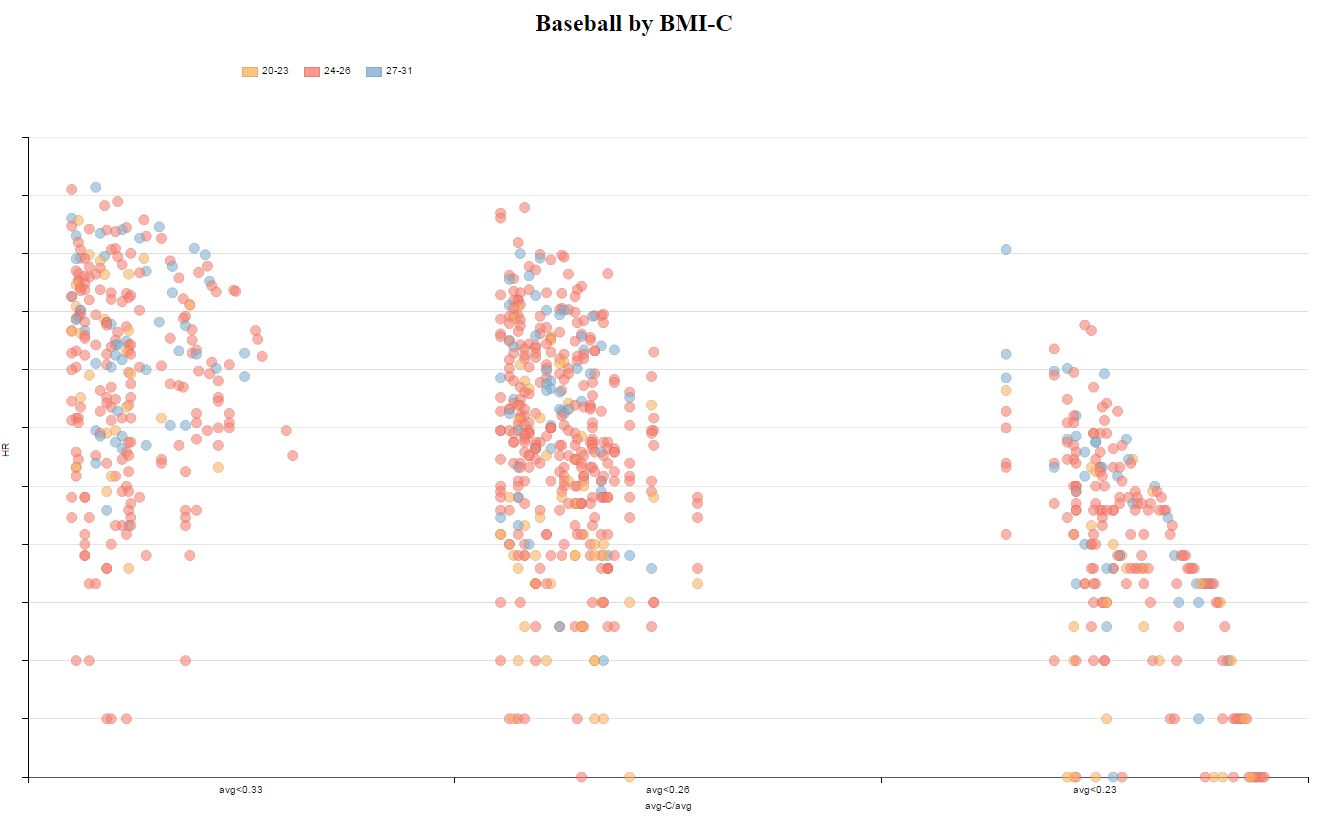
* Data cleaning - removed player with avg=0 and HR=0
* we had player with the same name, to distugish them I added the initial for preferred hand to the name
* introduce a new feature Body Mass Index (BMI - <https://en.wikipedia.org/wiki/Body_mass_index>), more infos under this link

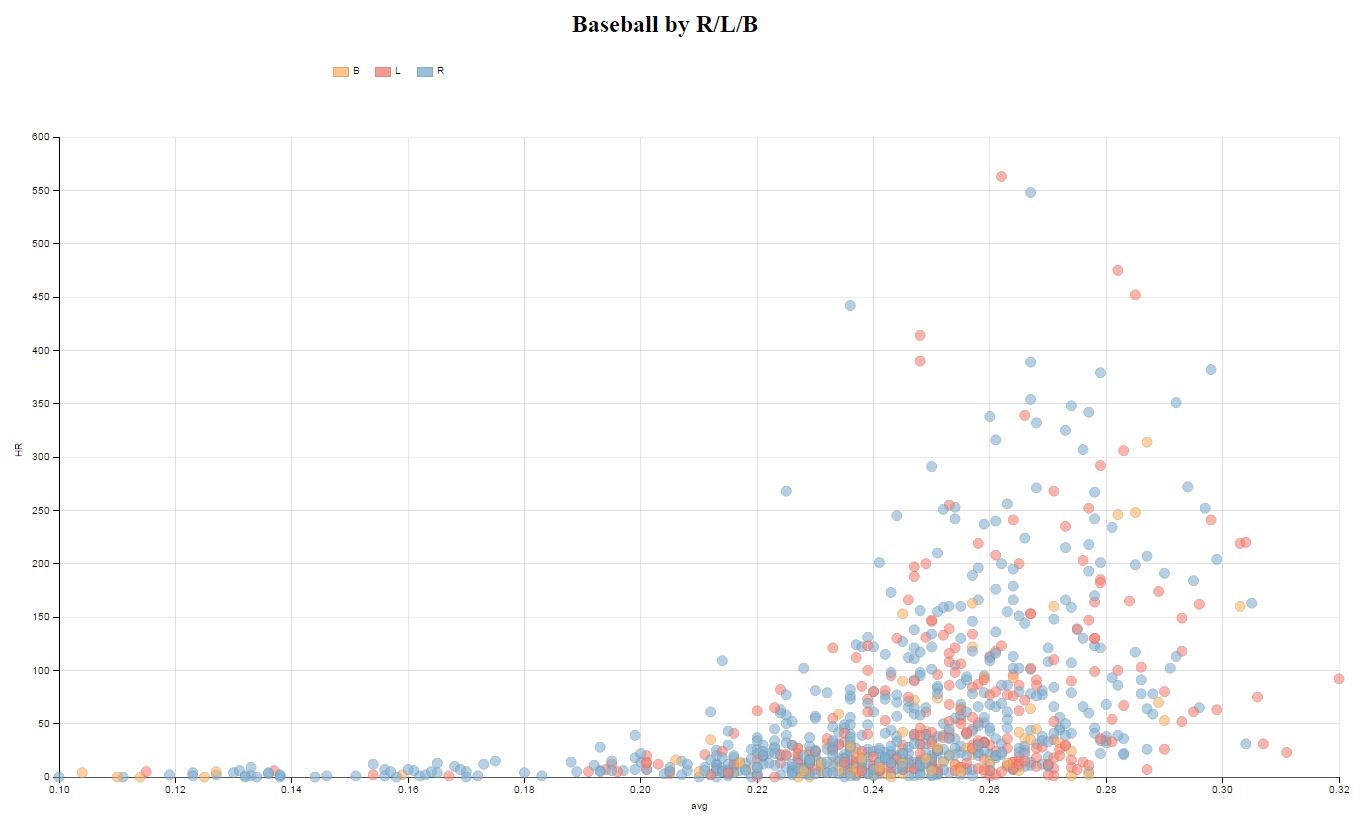
Storylines:

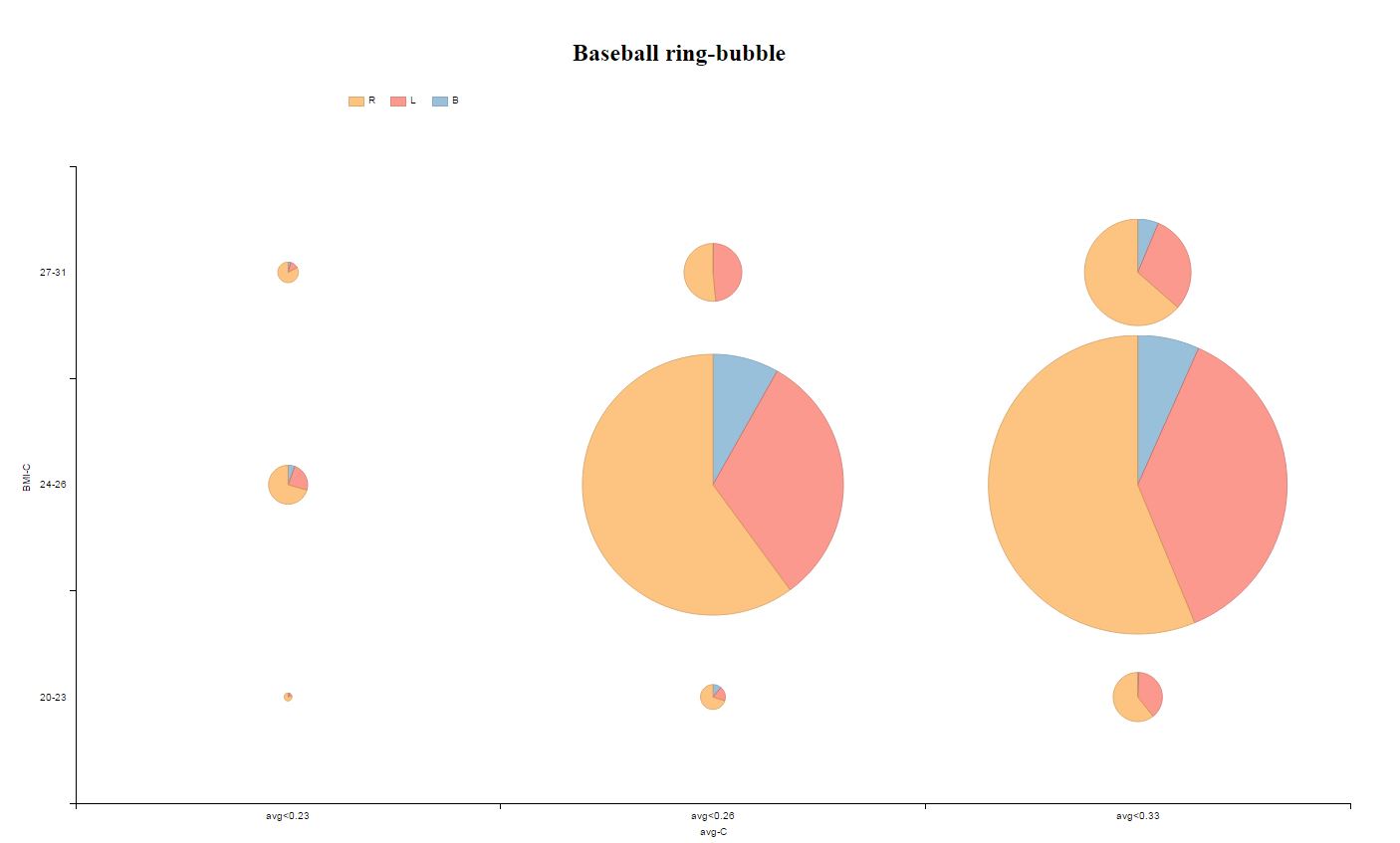
* The is a high correlation between weight and height and the performance of a baseball player. Performance as measured in Home Runs and Batting average
* In addition the handedness as an influnce too
* I wanted to use a naimation to expalin the audience the impact of BMI and Handedness on perfromacne metric, by showing the distribution of the players as a scatterplot
* in addition the audience can explore themselves by filtering the data nad find the best players and tehir indicators
* decided for the scatterplot since I was mainly intrested in the dsitribution of the baseball players by certain metric

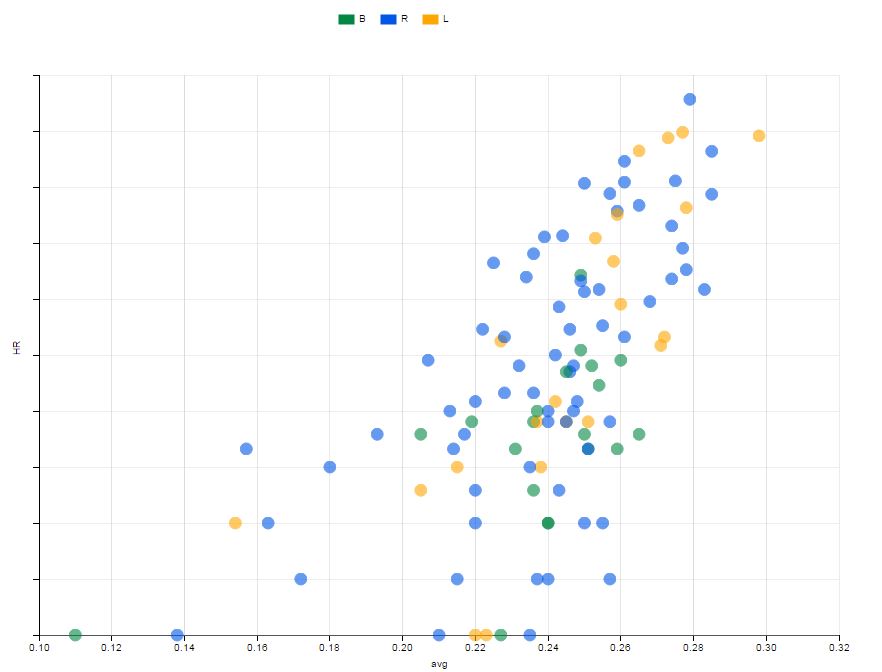
## Design

Attached my first d3.js charts to make the outlined story (see summarize) expalnatory.

Sketch 1: Introduce a category feature to explore HR by average by BMI-Index 

Sketch 2: Look at handedness by HR and avaerage 

Sketch 3: Explore bubble chart, the size represents the sum of HR by category BMI, average. In addition the color represents handedness 

Sketch 4: Use scatterplot to show players by BMI on a grid of HR and average, use color to show handedness 

SKetch 4 was the winner and will be used to further enhance the story.

Design choices:

* log 2 basis for HR - why
* change to Google colors and use opacity to show desnsity
* remove gridlines and reduce to minmum
* add animation and additiona chart to show HR distribution by BMI support expantory
* change text for axis and legend
* add filter options to provide eploartory capnbilities

See different version as .html file (v1\_xxxxxx.html - index.html)

explain any design choices you made including changes to the visualization after collecting feedback

## Feedback

Interviewer 1 (Marvin)

Interviewer 2 (Kirsta)

Interviewer 3 (JF)

include all feedback you received from others on your visualization from the first sketch to the final

## Visualization

## Final visualization

Resources - list any sources you consulted to create your visualization:

* <http://dimplejs.org/advanced_examples_viewer.html?id=advanced_storyboard_control>
* <http://dimplejs.org/>
* <http://d3js.org/>
* <https://en.wikipedia.org/wiki/Body_mass_index>
* Book: "Interactive Data Visualization for the Web", O'Reilly

## Appendix

Based on the insights I tested 2 apporaves Boxplot and scatter plot (see sketches)

### Analysing and understand the dataset

str(rw) #Compactly Display the Structure of an Arbitrary R Object

## 'data.frame': 891 obs. of 10 variables:  
## $ name : Factor w/ 891 levels "Adolfo Phillips",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ handedness: Factor w/ 3 levels "Both","Left",..: 3 2 3 3 3 3 2 2 1 2 ...  
## $ height : int 73 68 73 71 73 72 70 72 72 75 ...  
## $ weight : int 175 170 197 175 200 180 175 195 160 190 ...  
## $ avg : num 0.247 0.281 0.27 0.127 0.259 0.263 0.235 0.303 0.219 0.271 ...  
## $ HR : int 59 54 108 2 51 11 33 219 7 35 ...  
## $ BMI : int 24 26 27 25 27 25 26 27 22 24 ...  
## $ BMI.C : Factor w/ 3 levels "20-23","24-26",..: 2 2 3 2 3 2 2 3 1 2 ...  
## $ avg.C : Factor w/ 3 levels "avg<0.23","avg<0.26",..: 2 3 3 1 2 3 2 3 1 3 ...  
## $ size : int 2 2 2 2 2 2 2 2 2 2 ...

summary(rw) #Object SUmmaries

## name handedness height weight   
## Adolfo Phillips: 1 Both : 89 Min. :65.00 Min. :140.0   
## Al Bumbry : 1 Left :256 1st Qu.:71.00 1st Qu.:170.0   
## Al Cowens : 1 Right:546 Median :72.00 Median :180.0   
## Al Downing : 1 Mean :72.34 Mean :182.7   
## Al Ferrara : 1 3rd Qu.:74.00 3rd Qu.:193.0   
## Al Gallagher : 1 Max. :80.00 Max. :230.0   
## (Other) :885   
## avg HR BMI BMI.C   
## Min. :0.0660 Min. : 0.0 Min. :20.00 20-23:110   
## 1st Qu.:0.2300 1st Qu.: 10.0 1st Qu.:24.00 24-26:646   
## Median :0.2480 Median : 27.0 Median :25.00 27-31:135   
## Mean :0.2426 Mean : 58.9 Mean :25.11   
## 3rd Qu.:0.2630 3rd Qu.: 77.0 3rd Qu.:26.00   
## Max. :0.3280 Max. :563.0 Max. :31.00   
##   
## avg.C size   
## avg<0.23:218 Min. :2   
## avg<0.26:410 1st Qu.:2   
## avg<0.33:263 Median :2   
## Mean :2   
## 3rd Qu.:2   
## Max. :2   
##

Initial observations:

To first explore this data visually, I'll use the ggpair function and apply it to the data-set "rw". This will give me quick insights on the 5 variables. The intention here is to see a quick distribution of the values.

ggpairs summary as jpeg: 