Calculating a Qualifying Offer from Web Scraped Data

Nick Cianci

October 24, 2017

#Load in relevant packages  
library('DataComputing')  
library('rvest')  
library('readr')  
library('scales')  
library('knitr')

#Read the url  
#Source for webscraping commands: 'https://rpubs.com/Radcliffe/superbowl'  
  
url <- 'https://questionnaire-148920.appspot.com/swe/'  
webpage <- read\_html(url)

#Creates the data table from web scraped data and returns the first 6 rows  
  
mlb\_table <- html\_nodes(webpage, 'table')  
plr\_sal <- html\_table(mlb\_table)[[1]]  
names(plr\_sal) <- c("Player Name", "Salary ($)")  
kable(head(plr\_sal))

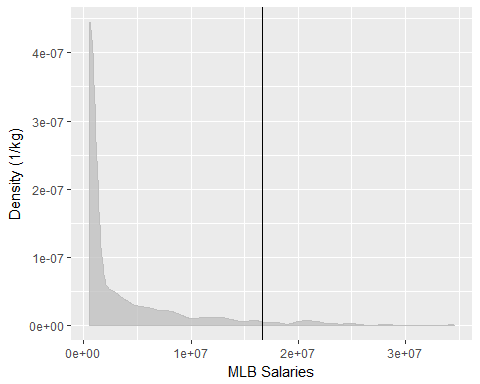
|  |  |
| --- | --- |
| Player Name | Salary ($) |
| Abreu, Jose | $11,666,667 |
| Acevedo, Andury | $507,500 |
| Ackley, Dustin | $3,200,000 |
| Adames, Cristhian | $507,500 |
| Adams, Austin | $507,500 |
| Adams, Matt | $1650000 |

#Data Cleaning: removed rows with no salary data   
  
clean\_sal<-(  
 plr\_sal %>%  
 filter(`Salary ($)`!= 'no salary data')  
)

#Strip Salaries of punctuation and sort them into descending order  
  
sorted <- sort(parse\_number(clean\_sal$`Salary ($)`), decreasing = TRUE)  
  
#Take the average of the top 125 salaries and format correctly  
#Source for dollar\_format()(): 'https://rdrr.io/cran/scales/man/dollar\_format.html'  
  
  
top125\_avg <- mean(sorted[0:124] )  
QO <- dollar\_format()(top125\_avg)  
  
#Prints what the Qualifying Offer will be  
  
paste('The monetary value of an upcoming Qualifying Offer is',QO)

## [1] "The monetary value of an upcoming Qualifying Offer is $16,703,018"

#Prepares clean\_sal for graphing  
clean\_sal$`Salary ($)`<-(sort(parse\_number(clean\_sal$`Salary ($)`)))  
  
#Creates a density plot of all mlb salaries with the vertical line being the Qualified Offer  
  
 clean\_sal %>%  
 ggplot(aes(x= `Salary ($)`))+  
 geom\_density(color = "gray", fill="gray", alpha = 0.75)+  
 xlab("MLB Salaries")+ylab("Density (1/kg)")+geom\_vline(xintercept = top125\_avg)



#Extracts the salary as a number and sorts them in descending order  
clean\_sal$`Salary ($)`<-(sort(parse\_number(clean\_sal$`Salary ($)`),decreasing = TRUE))  
  
#Takes the top half of the salaries in Major Leagues  
clean\_sal$`Salary ($)`<-clean\_sal$`Salary ($)`[0:(length(clean\_sal$`Salary ($)`)/2)]  
  
#Creates a density plot for the top half of the largest salaries and plots a vertical line where the Qualified Offer falls  
  
clean\_sal %>%  
 ggplot(aes(x= `Salary ($)`))+  
 geom\_density(color = "gray", fill="gray", alpha = 0.75)+  
 xlab("MLB Salaries")+ylab("Density (1/kg)")+geom\_vline(xintercept = top125\_avg)

