KimThan\_HW3

Kimberly Than

September 20, 2017

## Objective

The goal is to subset a certain class of job title (Athletic Trainer) into a new data frame, and generate boxplots showing the distribution for the different classes of Althletic Trainer (I, II, and Head).

## Setting the working directory

RStudio defaults to the .Rmd file's directory as the working directory, so it is not necessary to set your working directory when using RStudio go generate a Word or HTML document from this .Rmd file.

## Importing data

Next, I will import my data file of interest. Since the working directory used by RStudio (where this .Rmd file is located) is a sub-directory of my main directory 255E, I need to use the ../data/ syntax to get to my data folder.

CSU = read.csv("../data/california-state-university-2015.csv")

## Searching for the different classes

Since this data file contains numerous job titles, I want to find the different classes of Althletic Trainer. I will be using the ones under a 12 month contract.

jobs = levels(CSU$Job.Title)  
jobs[grep(pattern = "ATHLETIC TRAINER", x = jobs)]

## [1] "ATHLETIC TRAINER I - 12 MONTH" "ATHLETIC TRAINER I - AY"   
## [3] "ATHLETIC TRAINER II - 12 MONTH" "ATHLETIC TRAINER II - AY"   
## [5] "HEAD ATHLETIC TRAINER - 12 MONTH"

## Subsetting the different classes

Here I am subsetting out the three classes of althletic trainer.

AT1 = subset(CSU, Job.Title == "ATHLETIC TRAINER I - 12 MONTH")  
AT2 = subset(CSU, Job.Title == "ATHLETIC TRAINER II - 12 MONTH")  
AT3 = subset(CSU, Job.Title == "HEAD ATHLETIC TRAINER - 12 MONTH")

## Merging the subsets

Then, I am merging the three subsets into one data set.

ATAll = rbind(AT1, AT2, AT3)

## Organizing the levels

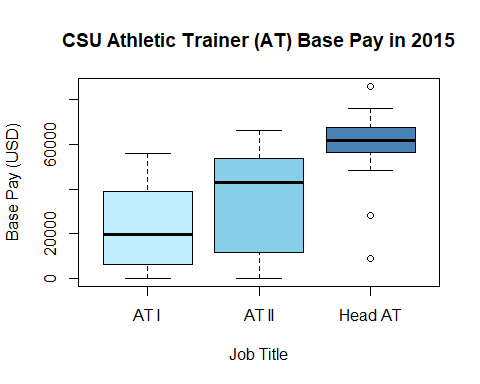
The data set now only has the job titles I am interesetd in, with three levels in the order that I want to have them displayed.

ATAll$Job.Title = factor(ATAll$Job.Title,   
 levels = c("ATHLETIC TRAINER I - 12 MONTH",   
 "ATHLETIC TRAINER II - 12 MONTH",  
 "HEAD ATHLETIC TRAINER - 12 MONTH" ))

## Including plots

Now we can make 3 boxplots, each representing the distribution of salary for each class. I abbreviated the name, specified the color for each class, and included axes titles.

boxplot(Base.Pay~Job.Title, data = ATAll, names = c("AT I", "AT II", "Head AT"),   
 main = "CSU Athletic Trainer (AT) Base Pay in 2015",  
 xlab = "Job Title",   
 ylab = "Base Pay (USD)",  
 col = c("light blue1", "skyblue", "steelblue"))



Boxplot of CSU salary distribution of three different classes of Athletic trainer in 2015.

## Including R results

We can use the aggregate function to calculate the mean base pay of each level of Athletic Trainer.

mymeans = aggregate(Base.Pay~Job.Title, data = ATAll, FUN = mean)  
mymeans

## Job.Title Base.Pay  
## 1 ATHLETIC TRAINER I - 12 MONTH 21924.99  
## 2 ATHLETIC TRAINER II - 12 MONTH 34482.87  
## 3 HEAD ATHLETIC TRAINER - 12 MONTH 58092.94

The function kable can also be used to make a table outside of the data frame.

kable(mymeans, caption = "Mean base pay for each job level.")

Mean base pay for each job level.

|  |  |
| --- | --- |
| Job.Title | Base.Pay |
| ATHLETIC TRAINER I - 12 MONTH | 21924.99 |
| ATHLETIC TRAINER II - 12 MONTH | 34482.87 |
| HEAD ATHLETIC TRAINER - 12 MONTH | 58092.94 |