Homework 2

Talal Khodr

7/15/2019

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# Part 1  
  
library(readr)  
  
bobble = read\_csv("/Users/tkhodr/Desktop/Predictive Analytics/Week 2-Files/HW2/bobbleheads.csv")

## Parsed with column specification:  
## cols(  
## year = col\_double(),  
## home\_team = col\_character(),  
## month = col\_character(),  
## day = col\_double(),  
## attend = col\_double(),  
## day\_of\_week = col\_character(),  
## opponent = col\_character(),  
## temp = col\_double(),  
## skies = col\_character(),  
## day\_night = col\_character(),  
## cap = col\_character(),  
## shirt = col\_character(),  
## fireworks = col\_character(),  
## bobblehead = col\_character()  
## )

dim(bobble)

## [1] 2421 14

head(bobble)

## # A tibble: 6 x 14  
## year home\_team month day attend day\_of\_week opponent temp skies  
## <dbl> <chr> <chr> <dbl> <dbl> <chr> <chr> <dbl> <chr>  
## 1 2012 Baltimor… APR 6 46773 Friday Minneso… 59 Clear  
## 2 2012 Baltimor… APR 7 31532 Saturday Minneso… 63 Clear  
## 3 2012 Baltimor… APR 8 14738 Sunday Minneso… 68 Clear  
## 4 2012 Baltimor… APR 9 25478 Monday New Yor… 65 Clou…  
## 5 2012 Baltimor… APR 10 24659 Tuesday New Yor… 62 Clear  
## 6 2012 Baltimor… APR 11 22919 Wednesday New Yor… 53 Clou…  
## # … with 5 more variables: day\_night <chr>, cap <chr>, shirt <chr>,  
## # fireworks <chr>, bobblehead <chr>

bobble$ordered\_day\_of\_week

## Warning: Unknown or uninitialised column: 'ordered\_day\_of\_week'.

## NULL

#define an ordered day-of-week variable   
bobble$ordered\_day\_of\_week <- with(data=bobble,  
ifelse ((day\_of\_week == "Monday"),1,  
ifelse ((day\_of\_week == "Tuesday"),2,  
ifelse ((day\_of\_week == "Wednesday"),3,  
ifelse ((day\_of\_week == "Thursday"),4,  
ifelse ((day\_of\_week == "Friday"),5,  
ifelse ((day\_of\_week == "Saturday"),6,7)))))))  
  
bobble$ordered\_day\_of\_week <- factor(bobble$ordered\_day\_of\_week, levels=1:7,  
labels=c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun"))  
  
# The difference is how its ordered, Friday just appears first.  
# However in the second one it is given significance levels and it understands the   
# Relationship the string has to the factor level provided.  
  
table(bobble$day\_of\_week)

##   
## Friday Monday Saturday Sunday Thursday Tuesday Wednesday   
## 379 284 392 388 235 366 377

table(bobble$ordered\_day\_of\_week)

##   
## Mon Tue Wed Thur Fri Sat Sun   
## 284 366 377 235 379 392 388

sapply(bobble,class)

## year home\_team month   
## "numeric" "character" "character"   
## day attend day\_of\_week   
## "numeric" "numeric" "character"   
## opponent temp skies   
## "character" "numeric" "character"   
## day\_night cap shirt   
## "character" "character" "character"   
## fireworks bobblehead ordered\_day\_of\_week   
## "character" "character" "factor"

team\_stats <- as.data.frame(table(bobble$home\_team))  
summary(team\_stats)

## Var1 Freq   
## Arizona Diamondbacks: 1 Min. :79.0   
## Atlanta Braves : 1 1st Qu.:81.0   
## Baltimore Orioles : 1 Median :81.0   
## Boston Red Sox : 1 Mean :80.7   
## Chicago Cubs : 1 3rd Qu.:81.0   
## Chicago White Sox : 1 Max. :81.0   
## (Other) :24

# 30 teams  
# Median :81.0  
  
#Part 2  
  
load("/Users/tkhodr/Desktop/Predictive Analytics/Week 2-Files/HW2/dodgers\_bad.Rda")  
  
dim(dodgers\_bad)

## [1] 97 17

unique\_dodger <- unique(dodgers\_bad)  
  
dim(unique\_dodger)

## [1] 81 17

library(pastecs)  
options(scipen = 1000)  
options(digits=2)  
stat.desc(unique\_dodger$attend)

## nbr.val nbr.null nbr.na min max   
## 81.00 0.00 0.00 24312.00 100000.00   
## range sum median mean SE.mean   
## 75688.00 3622209.00 40619.00 44718.63 1807.69   
## CI.mean.0.95 var std.dev coef.var   
## 3597.42 264687899.81 16269.23 0.36

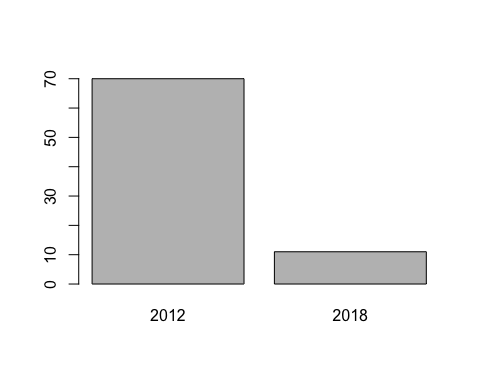
stat.desc(unique\_dodger$temp)

## nbr.val nbr.null nbr.na min max   
## 81.00 0.00 0.00 -30.00 95.00   
## range sum median mean SE.mean   
## 125.00 4883.00 71.00 60.28 3.88   
## CI.mean.0.95 var std.dev coef.var   
## 7.72 1218.31 34.90 0.58

## Temp stands out s it has a 1218 variation  
## Max is 100K for attend, more than capacity.  
  
table(unique\_dodger$year)

##   
## 2012 2018   
## 70 11

barplot(table(unique\_dodger$year))



as.data.frame(table(unique\_dodger$opponent))

## Var1 Freq  
## 1 Arizona Diamondbacks 9  
## 2 Atlanta Braves 3  
## 3 Baltimore Orioles 0  
## 4 Boston Red Sox 0  
## 5 Chicago Cubs 3  
## 6 Chicago White Sox 3  
## 7 Cincinnati Reds 3  
## 8 Cleveland Indians 0  
## 9 Colorado Rockies 9  
## 10 Detroit Tigers 0  
## 11 Houston Astros 3  
## 12 Kansas City Royals 0  
## 13 Los Angeles Angels 3  
## 14 Los Angeles Dodgers 0  
## 15 Miami Marlins 3  
## 16 Milwaukee Brewers 4  
## 17 Minnesota Twins 0  
## 18 New York Mets 4  
## 19 New York Yankees 0  
## 20 Oakland Athletics 0  
## 21 Philadelphia Phillies 3  
## 22 Pittsburgh Pirates 3  
## 23 San Diego Padres 18  
## 24 San Francisco Giants 0  
## 25 Seattle Mariners 0  
## 26 St. Louis Cardinals 7  
## 27 Tampa Bay Rays 0  
## 28 Texas Rangers 0  
## 29 Toronto Blue Jays 0  
## 30 Washington Nationals 3

#San Diego Padres 18 Stands out as they were played 18 times.  
  
ids <- subset(unique\_dodger,   
 date =='2012-05-07' |   
 date =='2012-05-08' |  
 date =='2012-05-09' |  
 date =='2012-08-20' |  
 date =='2012-08-21' |  
 date =='2012-08-22' |  
 date =='2012-10-01' |  
 date =='2012-10-02' |  
 date =='2012-10-03'   
 ,  
 select=c(gameid, date, opponent)  
)  
ids

## gameid date opponent  
## 1628 13 2012-05-07 San Diego Padres  
## 1629 14 2012-05-08 San Diego Padres  
## 1630 15 2012-05-09 San Diego Padres  
## 1674 59 2012-08-20 San Diego Padres  
## 1675 60 2012-08-21 San Diego Padres  
## 1676 61 2012-08-22 San Diego Padres  
## 1694 79 2012-10-01 San Diego Padres  
## 1695 80 2012-10-02 San Diego Padres  
## 1696 81 2012-10-03 San Diego Padres

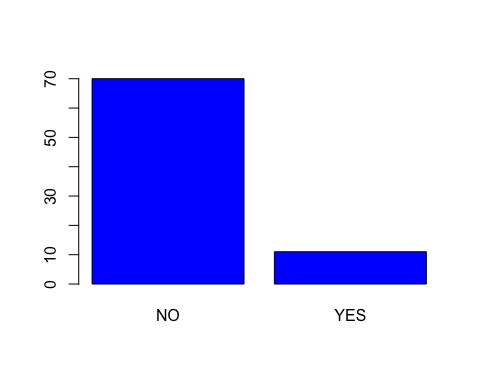
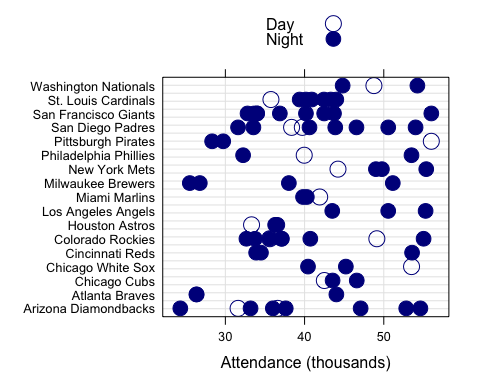
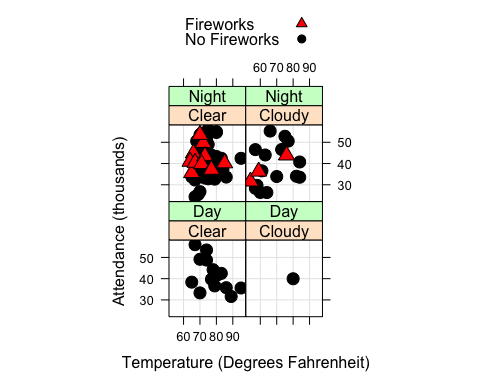
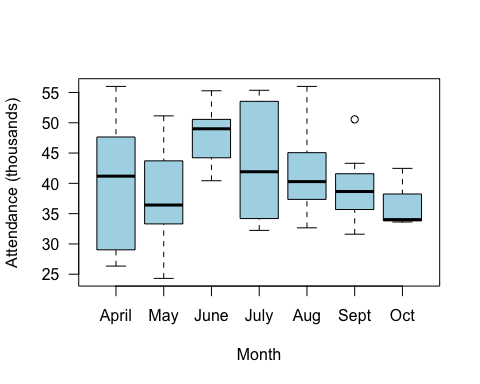
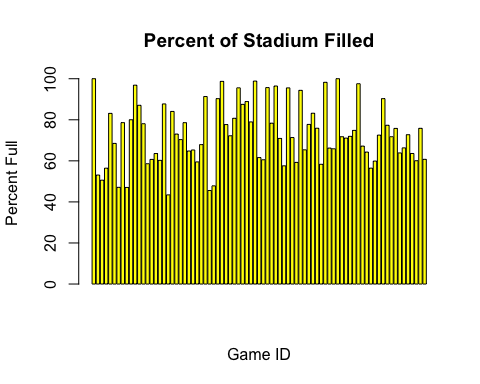
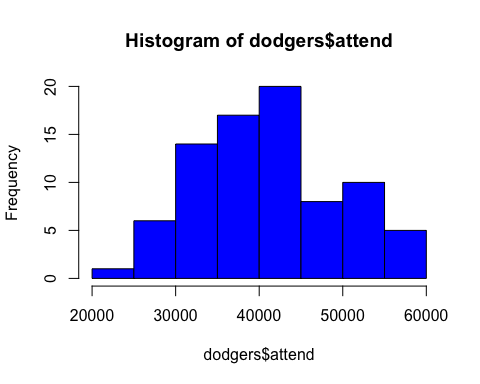
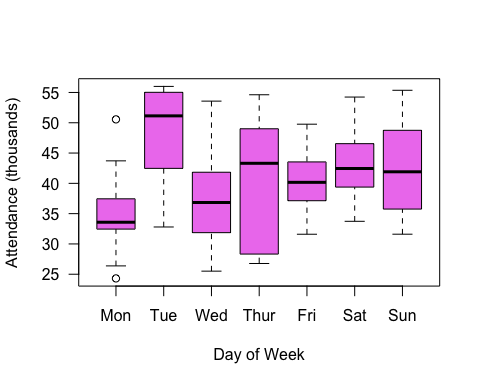
unique\_dodger[complete.cases(unique\_dodger),]

## year home\_team month day attend day\_of\_week opponent  
## 1658 2012 Los Angeles Dodgers JUL 4 53570 Wednesday Cincinnati Reds  
## temp skies day\_night cap shirt fireworks bobblehead  
## 1658 70 Clear Night NO NO YES NO  
## ordered\_day\_of\_week gameid date  
## 1658 Wed 43 2012-07-04

# Just surpirised so many attended for a non booble head game.

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.