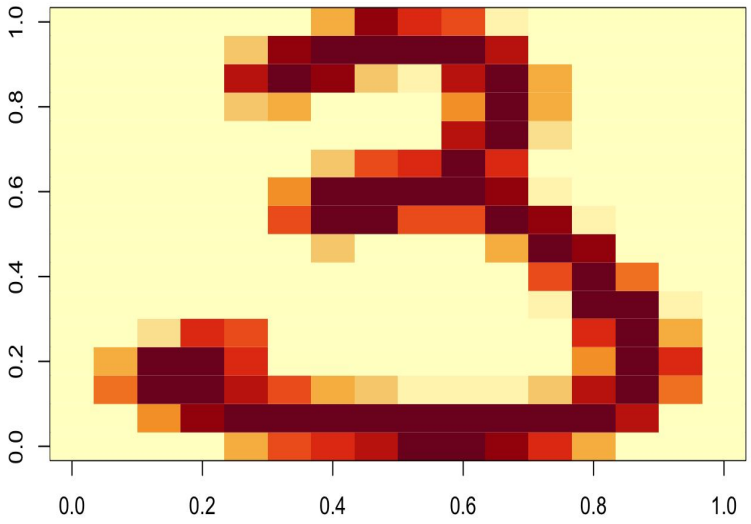
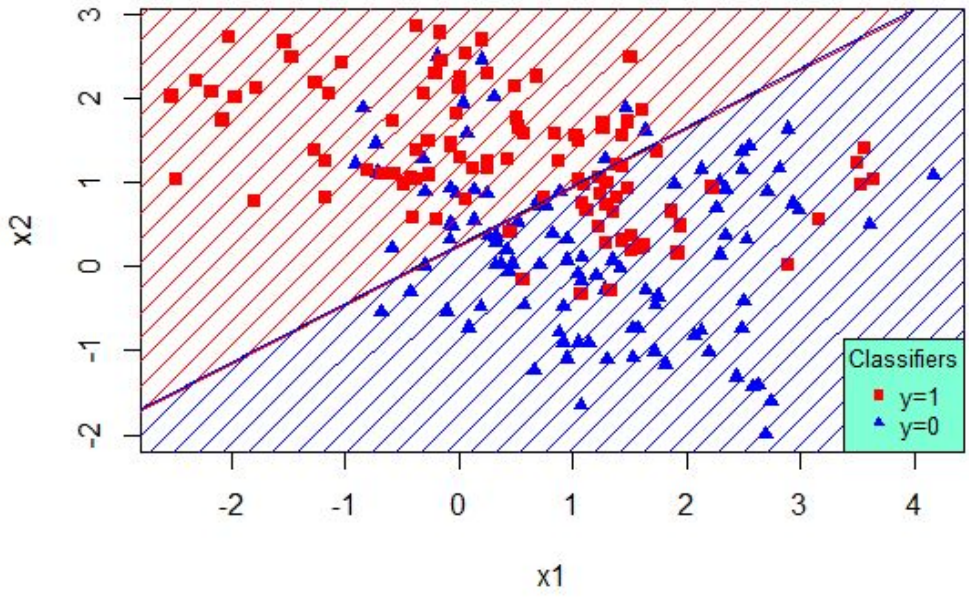
STAT 3280 - Section 002
Homework 1
Nancy Zhou (nz2ay), Philip Lee (phl6cw), Zinnia Lin (zl3rk)

Question 1. (3 pts)

The resulting digit is the number “3”

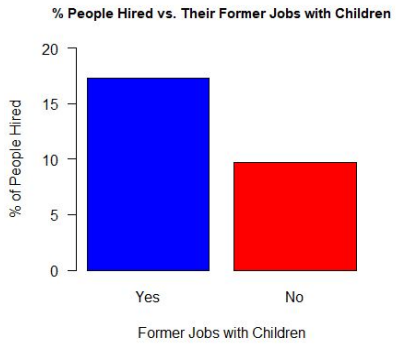


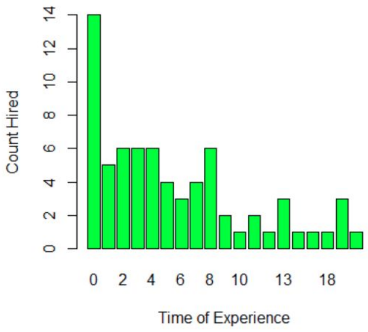
Question 2. (5 pts)

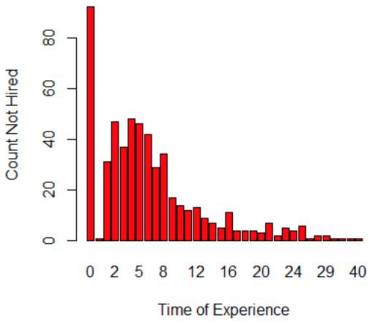


The red square boxes indicate class label y of 1 while the blue triangles indicate class label y of 2. The red region is the region that will predict classifiers as 1 while the blue region is the region that will predict classifiers as 0.

Question 3 (6 pts)

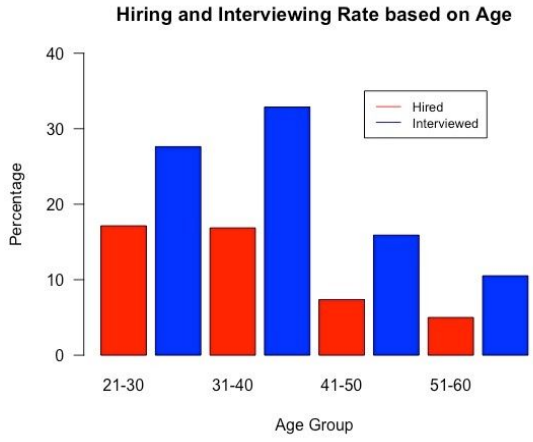


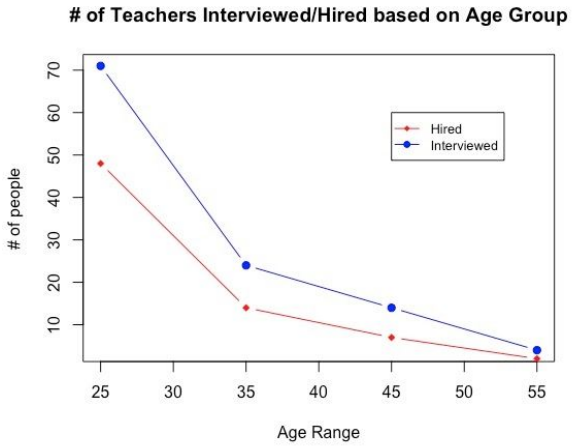




The graph on the upper left corner demonstrates that 17% of candidates who had experience working with kids were hired while 10% of those who did not have experience working with kids were hired. This conclusion is reasonable, as having experience working with kids is a very relevant factor in applying for teaching roles. However, having the experience of working with kids may not be the most important factor, as almost 83% of the candidates who **had** experience with children were not hired.

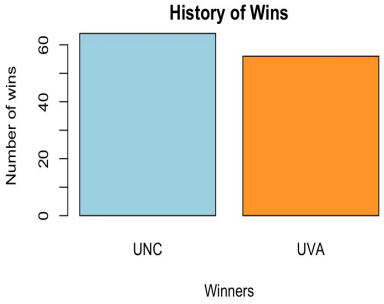
This visual on the upper right corner demonstrates that having more years of experience does not produce a higher chance of obtaining employment. This supports the idea that many schools prefer to hire candidates with less experience due to limited budgets.

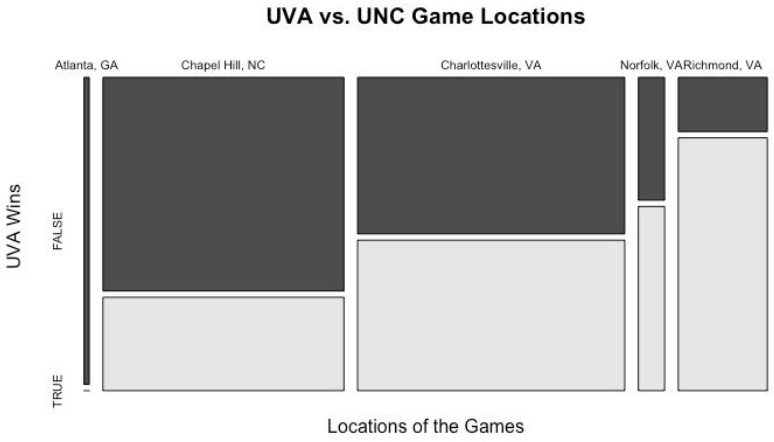




The graphs above depicts 4 age groups and their respective interview and hiring rates. In both the interview and hiring process, it is apparent that there is an decrease in the hiring rates from the youngest to oldest age groups. The interviewing rate follows a similar trend with the exception that candidates between 31-40 years old have (slightly) higher interview rates than candidates in the 21-30 age group. Based on this information, one can deduce that there may exist age discrimination with the older candidates less attractive to schools. In applying this data to the real world, this conclusion sounds reasonable as older candidates may have more experience teaching and therefore can command higher salaries than younger candidates who have no/less experience.

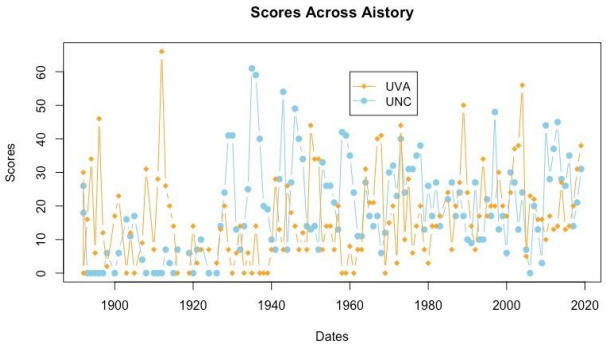
Question 4 (6 pts)

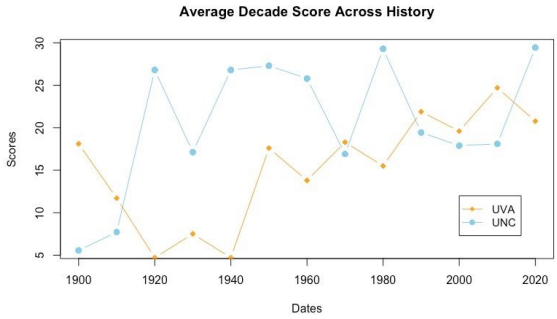




Note: ties are excluded

Looking at the graphs above, in the entirety of their rivalry, UNC has won slightly more games than UVA. For UNC's part, it scored more wins on its home turf and less wins at away games, suggesting that location may play a factor in their games. UVA, on the other hand, has done exceptionally well in its away games versus UNC at Richmond, VA around the late 19th century to early 20th century, with many consecutive wins over UNC in the first 22 years of their rivalry.





In its earliest years, UVA consistently scored above UNC, resulting in two outperforming decades. However, UNC scored higher than UVA when we look at the average scores per decade. Nonetheless, UVA's average score is positively trending from the 1940s to 2019. This may indicate that UVA has been investing more and more in its football team, such as the players, the coach, the resources available, and/or the environment.

Appendix

Question 1 code:

```
a<-read.table("HandWritten.txt")
b<-matrix(as.numeric(a),nrow=16,ncol=16,byrow=T)
image(b)
rotate <- function(x) t(apply(x, 2, rev))
image(rotate(b))
```

Question 2 code:

```
aa<-read.csv("mixture.csv",header=T)
bb<-aa[,2:4]
bb$color<-"black"
bb$color[bb$y==1]<-"red"
bb$color[bb$y==0]<-"blue"
qq<-c(17,15)[factor(bb$color)]     ## numbers inside c() dictates the plot symbols
plot(bb$x1,bb$x2,col=bb$color,xlab="x1",ylab="x2",pch=qq)
abline(a=.134*bb$x1/1.398,b=.978/1.398)
polygon(x=c(-2.9,3.92,3),y=c(-1.8,2.97,999),col="red",density=10)
polygon(x=c(-3.9,4.5,5),y=c(-2.46,-2.97,3.75),col="blue",density=10)
legend("bottomright",col=c("red","blue"),legend=c("y=1","y=0"),pch =
c(15,17),bg="aquamarine",title="Classifiers",cex=0.8)
```

Question 3 code:

```
hires=read.csv("TeacherHires.csv")
#head(hires)
hires=hires[,-c(4,6,9,11,13)] #takes out all the x, x.1 columns with N/A
#hires[,-c("X","X.1","X.2","X.3","X.4")]
head(hires)
row.has.na <- apply(hires, 1, function(x){any(is.na(x))})
sum(row.has.na)
final.filtered <- hires[!row.has.na,]

na.omit(hires)
#delete rows with NA varialbes
```

```
library(YaleToolkit)
x <- read.csv("TeacherHires.csv")
dim(x)
summary(x)
whatis(x)

head(x)

wi <- whatis(x)
class(wi)

names(wi)

wi$missing

#There are some columns which are empty.

which(wi$missing==nrow(x))

x <- x[, -which(whatis(x)$missing==nrow(x))]

dim(x)

whatis(x)


names(x)

#It is also helpful to improve variable names; try to use short-yet-descriptive words, without awkward capitalization. I use the type.truncate option to narrow the whatis display.

names(x) <- c("interviewed", "hired", "appdate",
              "age", "sex", "residence", "GPA.u",
              "GPA.g", "MA", "substitute",
              "teaching", "experience", "workkids",
              "volunteer")


whatis(x, type.truncate = 4)


# Clean up sex: Why are there 5 distinct values for this variable?


x$sex <- as.character(x$sex)
```

```
unique(x$sex)
levels(x$sex)
x$sex[x$sex==""] <- NA
x$sex[x$sex=="M"] <- "Male"
x$sex[x$sex=="F"] <- "Female"
x$sex <- factor(x$sex)
levels(x$sex)
table(x$sex)

# Clean up age: age was listed as a mixed factor
levels(x$age)
table(x$age)

sum(table(x$age))
sum(is.na(x$age))

### convert it to numeric

### but you cannot do this by directly using as.numeric

summary(as.numeric(x$age))

## this is because age is a factor now

as.numeric("N/A")
as.numeric("")
as.numeric("22")

summary(as.numeric(as.character(x$age)))
## this is what we want

x$age <- as.numeric(as.character(x$age))

x$agegroup <- factor(x$age<=39, levels=c(FALSE,TRUE),
                     labels=c("older", "younger"))


#omit all N/A's in age:
x<-x[is.na(x$age)==F,]

#changing yes* to yes for hired column
x$hired<-as.character(x$hired)
x$hired[x$hired=="yes "] <-"yes"
x$hired[x$hired=="yes*"] <-"yes"
x$hired<-factor(x$hired)
```

```
#changing “no “ and “” to “no” and NA for workkids column
x$workkids<-as.character(x$workkids)
x$workkids[x$workkids=="no "] <-"no"
x$workkids[x$workkids==""] <-NA
x$workkids<-factor(x$workkids)

#changing yes* to yes for interviewed column
x$interviewed<-as.character(x$interviewed)
x$interviewed[x$interviewed=="yes "] <-"yes"
x$interviewed<-factor(x$interviewed)

#makes a barplot for percentage of people who got hired or not based on their former jobs with
#children column
yeskids<-table(x$hired[x$workkids=='yes'])
nokids<-table(x$hired[x$workkids=='no'])
perc1<-yeskids[[2]]/yeskids[[1]]*100
perc2<-nokids[[2]]/nokids[[1]]*100
barplot(height=c(perc1,perc2),ylab="% of People Hired",xlab= "Former Jobs with
Children",names.arg = c("Yes","No"),main="% People Hired vs. Their Former Jobs with
Children",cex.main=.95,col=c("Blue","red"),las=1,ylim=c(0,20))

#creating new category: AgeRange:
Agerange<-ifelse(x$age>=21 & x$age<=30,x$AgeRange<-"21-30", ifelse(x$age>=31 &
x$age<=40, x$AgeRange<-"31-40", ifelse(x$age>=41 & x$age<=50,x$AgeRange<-"41-50",
x$AgeRange<-"51-60")))

x$AgeRange<-Agerange


table(x$hired,x$AgeRange)
table(x$interviewed,x$AgeRange)

#Create barplot for hire and interview % based on age group
twenty<- table(x$hired[x$AgeRange == "21-30"])
twentyInt<- table(x$interviewed[x$AgeRange == "21-30"])
thirty <- table(x$hired[x$AgeRange == "31-40"])
thirtyInt <- table(x$interviewed[x$AgeRange == "31-40"])
forty <- table(x$hired[x$AgeRange == "41-50"])
fortyInt <- table(x$interviewed[x$AgeRange == "41-50"])
fifty <- table(x$hired[x$AgeRange == "51-60"])
fiftyInt <- table(x$interviewed[x$AgeRange == "51-60"])

perc20 <- twenty[[2]]/twenty[[1]]*100
perc20Int <- twentyInt[[2]]/twentyInt[[1]]*100
```

```
perc30 <- thirty[[2]]/thirty[[1]]*100
perc30Int <- thirtyInt[[2]]/thirtyInt[[1]]*100
perc40 <- forty[[2]]/forty[[1]]*100
perc40Int <- fortyInt[[2]]/fortyInt[[1]]*100
perc50 <- fifty[[2]]/fifty[[1]]*100
perc50Int <- fiftyInt[[2]]/fiftyInt[[1]]*100

barplot(height = c(perc20, perc20Int, perc30, perc30Int, perc40, perc40Int, perc50,
perc50Int),names.arg = c("21-30", "21- 30", "31-40", "31-40", "41-50","41-50", "51-60","51-60"),
xlab = "Age Group", ylab = "Percentage", main = "Hiring and Interviewing Rate based on Age",
col = c("Red", "Blue"), las=1, ylim=c(0,40))
legend(6, 35, legend = c("Hired", "Interviewed"), col = c("Red", "Blue"), lty = 1:1, cex = 0.8)

#Code to generate a line graph of teachers hired and interviewed based on age group
y20 = 0
for(i in x$AgeRange[x$hired == "yes"]){
  print(i)
  if (i =="21-30")
    {y20 = y20+1}
  }
y20
y30 = 0
for(i in x$AgeRange[x$hired == "yes"]){
  print(i)
  if (i =="31-40")
  {y30 = y30+1}
}
y30
y40 = 0
for(i in x$AgeRange[x$hired == "yes"]){
  print(i)
  if (i =="41-50")
  {y40 = y40+1}
}
y40

y50 = 0
for(i in x$AgeRange[x$hired == "yes"]){
  print(i)
  if (i =="51-60")
  {y50 = y50+1}
}
y50

y20Int = 0
```

```
for(i in x$AgeRange[x$interviewed == "yes"]){
  print(i)
  if (i =="21-30")
  {y20Int = y20Int+1}
}
y20Int
y30Int = 0
for(i in x$AgeRange[x$interviewed == "yes"]){
  print(i)
  if (i =="31-40")
  {y30Int = y30Int+1}
}
y30Int
y40Int = 0
for(i in x$AgeRange[x$interviewed == "yes"]){
  print(i)
  if (i =="41-50")
  {y40Int = y40Int+1}
}
y40Int

y50Int = 0
for(i in x$AgeRange[x$interviewed == "yes"]){
  print(i)
  if (i =="51-60")
  {y50Int = y50Int+1}
}
y50Int


yhired <- c(y20, y30, y40, y50)
y2 <- c(y20Int, y30Int, y40Int, y50Int)

xlabs <- c(25, 35, 45, 55)

x1 <- xlabs

plot(x1, y2, type = "b", pch = 19, col = "blue", xlab = "Age Range", ylab = "# of people", main =
"# of Teachers Interviewed/Hired based on Age Group", cex.lab = 1)
lines(x1, yhired, type = "b", pch = 18, col = "red")
legend(45,60, legend = c("Hired", "Interviewed"), col = c("Red", "Blue"), pch = 18:19,lty = 1:1,
cex = 0.8)


##makes a barplot of the number of people hired vs not hired
```

```
exp<-as.character(x$experience)
exp[exp=="N/A"] <- NA    ##changes N/A into NA
rownumbers<-which(grepl(" ",exp)==T)  ##gets the row numbers that has weird unit of time for
                                                    experience
x$experience<-as.numeric(exp)
newx<-x[-rownumbers,c(2,12)]     ##makes a new matrix without the weird unit of time
noNAnewx<-newx[which(is.na(newx$experience)==F),]
test<-noNAnewx[noNAnewx$hired=="yes",]
test2<-noNAnewx[noNAnewx$hired=="no",]
par(mfrow=c(1,2))
barplot(table(test),xlab="Time of Experience",ylab="Count Hired",col="green")
barplot(table(test2),xlab="Time of Experience",ylab="Count Not Hired",col="red")
```

Question 4 code:

```
##data cleaning:
rival<-read.csv("rivalriesdata.csv")
rival$Score<-as.character(rival$Score)
noTies<-rival[-which(rival$Winner=="Tie"),]
yesTies<-rival[which(rival$Winner=="Tie"),]
VArowWin<-which(noTies$Winner=="Virginia")
NCrowWin<-which(noTies$Winner=="North Carolina")
VArowLose<-NCrowWin
NCrowLose<-VArowWin
numbers<-as.numeric(sapply(strsplit(noTies$Score,split='_'), function(x) x[1]))
numbers2<-as.numeric(sapply(strsplit(noTies$Score,split='_'), function(x) x[2]))
noTies$VAscores[VArowWin]<-numbers[VArowWin]
noTies$VAscores[VArowLose]<-numbers2[VArowLose]
noTies$NCscores[NCrowLose]<-numbers2[NCrowLose]
noTies$NCscores[NCrowWin]<-numbers[NCrowWin]

#complete history of wins:
barplot(height = c(length(NCrowWin),length(VArowWin)),names.arg = c("UNC","UVA"), xlab =
"Winners", ylab = "Number of wins", main = "History of Wins",col = c("light blue", "Orange"))


#Detailed line graph of wins according to dates
plot(noTies$Date, noTies$NCscores, type = "b", pch = 19, col="sky
blue",ylab="Scores",xlab="Dates", main = "Average Scores across history", ylim=
c(min(noTies$NCscores), max(noTies$VAscores)))
lines(noTies$Date, noTies$VAscores, type = "b", pch = 18, col = "orange")
legend(1960,60, legend = c( "UVA","UNC"), col = c( "ORANGE", "skyblue"), pch = 18:19,lty =
1:1, cex = 0.95)
```

```
#Add in separate categories in rival for va score and nc scores
score1 <- as.numeric(sapply(strsplit(rival$Score,split='_'), function(x) x[1]))
score2 <- as.numeric(sapply(strsplit(rival$Score,split='_'), function(x) x[2]))
rival$VAScore <- score1
rival$NCScore <- score2
length(rival$Date)
rival

#sorting the dates into groups
noTies$DateGroup <- ifelse(noTies$Date<=1900,noTies$DateGroup<-1900,
ifelse(noTies$Date>1900 & noTies$Date<=1910, noTies$DateGroup<-1910,
ifelse(noTies$Date>1910 & noTies$Date<=1920,noTies$DateGroup<-1920,
ifelse(noTies$Date>1920 & noTies$Date<=1930, noTies$DateGroup <- 1930,
ifelse(noTies$Date>1930 & noTies$Date <=1940, noTies$DateGroup <- 1940,
ifelse(noTies$Date > 1940 & noTies$Date <=1950, noTies$DateGroup <-
1950,ifelse(noTies$Date>1950 & noTies$Date<=1960, noTies$DateGroup<-1960,
ifelse(noTies$Date>1960 & noTies$Date<=1970,
noTies$DateGroup<-1970,ifelse(noTies$Date>1970 & noTies$Date<=1980,
noTies$DateGroup<-1980,ifelse(noTies$Date>1980 & noTies$Date<=1990,
noTies$DateGroup<-1990,ifelse(noTies$Date>1990 & noTies$Date<=2000,
noTies$DateGroup<-2000,ifelse(noTies$Date>2000 & noTies$Date<=2010,
noTies$DateGroup<-2010, noTies$DateGroup <- 2020)))))))))))


score1900 <- mean(noTies$NCscores[noTies$DateGroup == 1900])
score1910 <- mean(noTies$NCscores[noTies$DateGroup == 1910])
score1920 <- mean(noTies$NCscores[noTies$DateGroup == 1920])
score1930 <- mean(noTies$NCscores[noTies$DateGroup == 1930])
score1940 <- mean(noTies$NCscores[noTies$DateGroup == 1940])
score1950 <- mean(noTies$NCscores[noTies$DateGroup == 1950])
score1960 <- mean(noTies$NCscores[noTies$DateGroup == 1960])
score1970 <- mean(noTies$NCscores[noTies$DateGroup == 1970])
score1980 <- mean(noTies$NCscores[noTies$DateGroup == 1980])
score1990 <- mean(noTies$NCscores[noTies$DateGroup == 1990])
score2000 <- mean(noTies$NCscores[noTies$DateGroup == 2000])
score2010 <- mean(noTies$NCscores[noTies$DateGroup == 2010])
score2020 <- mean(noTies$NCscores[noTies$DateGroup == 2020])

score1900_2 <- mean(noTies$VAscores[noTies$DateGroup == 1900])
score1910_2 <- mean(noTies$VAscores[noTies$DateGroup == 1910])
score1920_2 <- mean(noTies$VAscores[noTies$DateGroup == 1920])
score1930_2 <- mean(noTies$VAscores[noTies$DateGroup == 1930])
score1940_2 <- mean(noTies$VAscores[noTies$DateGroup == 1940])
```

```
score1950_2 <- mean(noTies$VAscores[noTies$DateGroup == 1950])
score1960_2 <- mean(noTies$VAscores[noTies$DateGroup == 1960])
score1970_2 <- mean(noTies$VAscores[noTies$DateGroup == 1970])
score1980_2 <- mean(noTies$VAscores[noTies$DateGroup == 1980])
score1990_2 <- mean(noTies$VAscores[noTies$DateGroup == 1990])
score2000_2 <- mean(noTies$VAscores[noTies$DateGroup == 2000])
score2010_2 <- mean(noTies$VAscores[noTies$DateGroup == 2010])
score2020_2 <- mean(noTies$VAscores[noTies$DateGroup == 2020])

#computing averages for each 2 decades
noTies$VAAvg <- ifelse(noTies$Date<=1900,noTies$VAAvg<-score1900_2,
ifelse(noTies$Date<=1910,noTies$VAAvg<-score1910_2,ifelse(noTies$DateGroup==1920,
noTies$VAAvg<-score1940_2, ifelse(noTies$Date<=1930,noTies$VAAvg<-score1930_2,
ifelse(noTies$DateGroup==1940,noTies$VAAvg<-score1940_2,
ifelse(noTies$Date<=1950,noTies$VAAvg<-score1950_2,ifelse(noTies$DateGroup==1960,
noTies$VAAvg <- score1960_2,
ifelse(noTies$Date<=1970,noTies$VAAvg<-score1970_2,ifelse(noTies$DateGroup==1980,
noTies$VAAvg <- score1980_2,
ifelse(noTies$Date<=1990,noTies$VAAvg<-score1990_2,ifelse(noTies$DateGroup==2000,
noTies$VAAvg <-score2000_2,ifelse(noTies$Date<=2010,noTies$VAAvg<-score2010_2,
noTies$VAAvg <- score2020_2))))))))))
noTies$NCAvg <- ifelse(noTies$Date<=1900,noTies$NCAvg<-score1900,
ifelse(noTies$Date<=1910,noTies$NCAvg<-score1910,ifelse(noTies$DateGroup==1920,
noTies$NCAvg<-score1940, ifelse(noTies$Date<=1930,noTies$NCAvg<-score1930,
ifelse(noTies$DateGroup==1940,noTies$NCAvg<-score1940,
ifelse(noTies$Date<=1950,noTies$NCAvg<-score1950,ifelse(noTies$DateGroup==1960,
noTies$NCAvg <- score1960,
ifelse(noTies$Date<=1970,noTies$NCAvg<-score1970,ifelse(noTies$DateGroup==1980,
noTies$NCAvg <- score1980,
ifelse(noTies$Date<=1990,noTies$NCAvg<-score1990,ifelse(noTies$DateGroup==2000,
noTies$NCAvg <-score2000,ifelse(noTies$Date<=2010,noTies$NCAvg<-score2010,
noTies$NCAvg <- score2020))))))))))

#plots based on dates and scores
par(mfrow = c(1,1))
plot(noTies$DateGroup, noTies$NCAvg, type = "b", pch = 19, col="sky
blue",ylab="Scores",xlab="Dates", main = "Average Decade Score Across History", ylim=
c(min(noTies$NCAvg), max(noTies$NCAvg)))
lines(noTies$DateGroup, noTies$VAAvg, type = "b", pch = 18, col = "orange")
legend(2000,12, legend = c( "UVA","UNC"), col = c( "ORANGE", "skyblue"), pch = 18:19,lty =
1:1, cex = 0.95)
```

```
#plots based on dates and average scores
par(mfrow = c(1,1))
plot(noTies$DateGroup, noTies$NCAvg, type = "b", pch = 19, col="sky
blue",ylab="Scores",xlab="Dates", main = "Average Scores across history", ylim=
c(min(noTies$NCAvg), max(noTies$NCAvg)))
lines(noTies$DateGroup, noTies$VAAvg, type = "b", pch = 18, col = "orange")
legend(1900,25, legend = c( "UVA","UNC"), col = c( "ORANGE", "skyblue"), pch = 18:19,lty =
1:1, cex = 0.95)

#mosaicplot for location of UVA wins
mosaicplot(noTies$Location~(noTies$VAscores>noTies$NCscores), ylab = "UVA Wins", xlab =
"Locations of the Games",main= "UVA vs. UNC Game Locations", color = TRUE)
```