Exercise 10

Nicole Navarro

November 14, 2016

Crime data

```
#import data
filenames <- list.files("~/School/Fall 16/EDA/data/FDLE",
                        pattern="*.xls", full.names=TRUE)

ldf <- lapply(filenames, function(x)
  read.xlsx(x, sheetName ="Part II (1)", startRow=3))

CNames <- names(as.data.frame(ldf[1]))
Alachua_label <- lapply(ldf, function(x)
  which((x[,1] == "Alachua County") | (x[,1] =="Alachua") ))
Washington_label <- lapply(ldf, function(x)
  which((x[,1] == "Washington County") | (x[,1] == "Washington") ))

#check completeness
unlist(Washington_label)-unlist(Alachua_label)
```

```
## [1] 66 66 66 66 66 66 66
```

```
#get rid of col names (because they don't all match)
NULL_Name <- function(x) {
  names(x) <- NULL
  return(x)
}
ldf <- lapply(ldf, NULL_Name)
```

```
#combine all years into one data frame
L<- length(as.data.frame(ldf[1])[1,])
FL_Crime_0612 <- as.data.frame(ldf[1])[Alachua_label[[1]]:
                                       Washington_label[[1]],]
for (i in 2:length(ldf)) {
  FL_Crime_0612 <- rbind(FL_Crime_0612,
                         as.data.frame(ldf[i])[Alachua_label[[i]]:
                                       Washington_label[[i]],1:L])
}

FL_Crime_0612 <- as.data.frame(FL_Crime_0612)
names(FL_Crime_0612) <-CNames
str(FL_Crime_0612)
```

```
## 'data.frame':    469 obs. of  11 variables:
##  $ Agency.County       : Factor w/ 142 levels "Alachua County",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Manslaughter        : num  1 1 3 0 6 8 0 0 1 0 ...
```

```
##  $ Kidnap..Abduction   : num  17 0 12 1 10 22 0 1 3 3 ...
##  $ Arson               : num  8 0 5 1 24 39 0 7 8 7 ...
##  $ Simple.Assault      : num  1427 62 1394 189 2815 ...
##  $ Drug.Arrest         : num  2697 163 2020 217 5282 ...
##  $ Bribery             : num  1 0 1 0 0 1 0 1 1 4 ...
##  $ Embezzlement        : num  27 0 63 6 48 233 0 34 1 0 ...
##  $ Fraud               : num  1133 0 377 14 400 ...
##  $ Counterfeit..Forgery: num  52 7 93 18 74 310 8 28 15 121 ...
##  $ Extortion..Blackmail: num  2 0 2 0 7 12 0 0 1 2 ...
```

```
#clean up county names
FL_Crime_0612[,1] <- as.factor(gsub("\\ County", "", FL_Crime_0612[,1]))
FL_Crime_0612[,1] <- as.factor(gsub("Desoto", "DeSoto", FL_Crime_0612[,1]))

FL_Crime_0612[,2:length(FL_Crime_0612[1,])] <- sapply(FL_Crime_0612[,2:length(FL_Crime_0612[1,])], as.in

start <- rep(2006,67)
FL_Crime_0612$Year <- c(start,start+1,start+2,start+3,start+4,start+5,start+6)
colnames(FL_Crime_0612)[1] <- "County"
str(FL_Crime_0612)
```

```
## 'data.frame':    469 obs. of  12 variables:
##  $ County              : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Manslaughter        : int  1 1 3 0 6 8 0 0 1 0 ...
##  $ Kidnap..Abduction   : int  17 0 12 1 10 22 0 1 3 3 ...
##  $ Arson               : int  8 0 5 1 24 39 0 7 8 7 ...
##  $ Simple.Assault      : int  1427 62 1394 189 2815 5367 76 599 618 795 ...
##  $ Drug.Arrest         : int  2697 163 2020 217 5282 16466 154 872 741 1244 ...
##  $ Bribery             : int  1 0 1 0 0 1 0 1 1 4 ...
##  $ Embezzlement        : int  27 0 63 6 48 233 0 34 1 0 ...
##  $ Fraud               : int  1133 0 377 14 400 724 0 80 201 565 ...
##  $ Counterfeit..Forgery: int  52 7 93 18 74 310 8 28 15 121 ...
##  $ Extortion..Blackmail: int  2 0 2 0 7 12 0 0 1 2 ...
##  $ Year                : num  2006 2006 2006 2006 2006 ...
```

```
#save cleaned up version to csv
write.csv(FL_Crime_0612, "~/School/Fall 16/EDA/data/FDLE/FL_Crime_0612.csv")
```

Educational Attainment Data

```
#import data
filenames2 <- list.files("~/School/Fall 16/EDA/data/FLedu", pattern="*.csv", full.names=TRUE)

ldf2 <- lapply(filenames2, function(x)
  read.csv(x, skip=1))


#combine into one frame
Cnames <- c("County","pop25up","HS25up","bachelors25up")
FL_edu_0912 <- as.data.frame(ldf2[1])[c(3,34,82,88)]
```

```
names(FL_edu_0912)<-Cnames
for (i in 2:length(ldf2)) {
  data <- as.data.frame(ldf2[i])[c(3,34,82,88)]
  names(data) <- Cnames
  FL_edu_0912 <- rbind(FL_edu_0912, data)
}

FL_edu_0912 <- as.data.frame(FL_edu_0912)
str(FL_edu_0912)
```

```
## 'data.frame':    268 obs. of  4 variables:
##  $ County       : Factor w/ 67 levels "Alachua County, Florida",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ pop25up      : int  127014 16214 110534 19960 381023 1204588 9359 123895 107696 115948 ...
##  $ HS25up       : num  89.1 78.6 85.8 78.7 89.9 87 72 87.9 83.9 89.8 ...
##  $ bachelors25up: num  39.3 6.7 20.2 10.2 26.4 29.3 10.9 20.8 16.7 22.9 ...
```

```
#clean up county names
FL_edu_0912[,1] <- as.factor(gsub("\\ County, Florida", "", FL_edu_0912[,1]))
FL_edu_0912[,1] <- as.factor(gsub("Miami-Dade", "Miami Dade", FL_edu_0912[,1]))

FL_edu_0912[,2:length(FL_edu_0912[1,])] <- sapply(FL_edu_0912[,2:length(FL_edu_0912[1,])], as.numeric)


start <- rep(2009,67)
FL_edu_0912$Year <- c(start,start+1,start+2,start+3)
str(FL_edu_0912)
```

```
## 'data.frame':    268 obs. of  5 variables:
##  $ County       : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ pop25up      : num  127014 16214 110534 19960 381023 ...
##  $ HS25up       : num  89.1 78.6 85.8 78.7 89.9 87 72 87.9 83.9 89.8 ...
##  $ bachelors25up: num  39.3 6.7 20.2 10.2 26.4 29.3 10.9 20.8 16.7 22.9 ...
##  $ Year         : num  2009 2009 2009 2009 2009 ...
```

Unemployment data

```
#import data
filenames <- list.files("~/School/Fall 16/EDA/data/FLlabor", pattern="*.xls", full.names=TRUE)

#Read in only the data we want
ldf <- lapply(filenames, function(x) read.xlsx2(x, sheetIndex =1))
ldf <- lapply(ldf, function(x) x <- x[,c(4,5,10)])

CNames <- c("County","Year", "Unemployment_Rate") ##Future column names

Alachua_label <- lapply(ldf, function(x) grep("Alachua", x[,1])) ##starting lines in each frame
Washington_label <- lapply(ldf, function(x) grep("Washington County, FL", x[,1])) ##starting lines in ea

ldf <- lapply(ldf, NULL_Name)  ##Naming tricks
```

```
#create data frame
FL_Unemployment_0612 <- as.data.frame(ldf[1])[Alachua_label[[1]]:
                                        (Washington_label[[1]]),]

for (i in 2:length(ldf)) {
  FL_Unemployment_0612 <- rbind(FL_Unemployment_0612,
                          as.data.frame(ldf[i])[Alachua_label[[i]]:
                                        (Washington_label[[i]]),])
}
FL_Unemployment_0612 <- as.data.frame(FL_Unemployment_0612)
names(FL_Unemployment_0612) <-CNames
str(FL_Unemployment_0612)
```

```
## 'data.frame':    469 obs. of  3 variables:
##  $ County           : Factor w/ 3224 levels "","Abbeville County, SC",..: 30 126 162 269 281 296 364
##  $ Year             : Factor w/ 9 levels "","2006","Year",..: 2 2 2 2 2 2 2 2 2 ...
##  $ Unemployment_Rate: Factor w/ 247 levels "","(%)","1.7",..: 63 64 68 65 70 68 69 71 75 67 ...
```

```
#clean up data frames

FL_Unemployment_0612[,1] <- as.factor(gsub("\\ County, FL", "", FL_Unemployment_0612[,1]))
FL_Unemployment_0612[,2] <- as.factor(gsub("\\ Year", "", FL_Unemployment_0612[,2]))
FL_Unemployment_0612[,1] <- as.factor(gsub("Miami-Dade", "Miami Dade", FL_Unemployment_0612[,1]))
FL_Unemployment_0612[,2] <- as.numeric(as.character(FL_Unemployment_0612[,2]))
FL_Unemployment_0612[,3] <- as.numeric(as.character(FL_Unemployment_0612[,3]))

rownames(FL_Unemployment_0612) <- NULL
str(FL_Unemployment_0612)
```

```
## 'data.frame':    469 obs. of  3 variables:
##  $ County           : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Year             : num  2006 2006 2006 2006 2006 ...
##  $ Unemployment_Rate: num  2.7 2.8 3.1 2.9 3.3 3.1 3.2 3.4 3.8 3 ...
```

Population Data

```
#2010-2015
url <- "https://www.census.gov/popest/data/counties/totals/2015/files/CO-EST2015-alldata.csv"

popdata1015 <- read.csv(url, header=TRUE)
popFL1015 <- popdata1015 %>% select(c(6,7,10:15)) %>% filter(STNAME=="Florida")

#clean
popFL1015 <- popFL1015 %>% filter(CTYNAME != "Florida") %>% select(2:length(popFL1015[1,]))
head(popFL1015)
```

```
##           CTYNAME POPESTIMATE2010 POPESTIMATE2011 POPESTIMATE2012
## 1  Alachua County          247625          249688          251669
## 2    Baker County           27076           27076           27052
```

```
## 3      Bay County          169247          169647          171920
## 4 Bradford County           28539           28477           27133
## 5  Brevard County          543966          544323          547495
## 6  Broward County         1753263         1787582         1818491
##   POPESTIMATE2013 POPESTIMATE2014 POPESTIMATE2015
## 1          253252          256518          259964
## 2           26991           27135           27420
## 3          174859          178703          181635
## 4           26895           26681           26928
## 5          551148          556902          568088
## 6         1843583         1869679         1896425
```

```
#change to numeric
popFL1015[,2:length(popFL1015[1,])] <- sapply(popFL1015[,2:length(popFL1015[1,])], as.character)
popFL1015[,2:length(popFL1015[1,])] <- sapply(popFL1015[,2:length(popFL1015[1,])], as.numeric)

#reshape (wide form to long form)
popFL1015_long <- melt(popFL1015, id.vars= "CTYNAME", measure.vars= c("POPESTIMATE2010", "POPESTIMATE20
head(popFL1015_long)
```

```
##           CTYNAME        variable   value
## 1  Alachua County POPESTIMATE2010  247625
## 2    Baker County POPESTIMATE2010   27076
## 3      Bay County POPESTIMATE2010  169247
## 4 Bradford County POPESTIMATE2010   28539
## 5  Brevard County POPESTIMATE2010  543966
## 6  Broward County POPESTIMATE2010 1753263
```

```
#cleaning county names
names(popFL1015_long) <- c("County", "Year", "population")

popFL1015_long[,1] <- as.factor(gsub("\\ County", "", popFL1015_long[,1]))
popFL1015_long[,2] <- as.factor(gsub("[A-Z]", "", popFL1015_long[,2]))
popFL1015_long[,1] <- as.factor(gsub("Miami-Dade", "Miami Dade", popFL1015_long[,1]))

str(popFL1015_long)
```

```
## 'data.frame':    402 obs. of  3 variables:
##  $ County    : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Year      : Factor w/ 6 levels "2010","2011",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ population: num  247625 27076 169247 28539 543966 ...
```

```
#2000-2009
url <- "https://www.census.gov/popest/data/intercensal/county/files/CO-EST00INT-TOT.csv"

popdata0009 <- read.csv(url, header=TRUE)
popFL0009 <- popdata0009 %>% select(c(6,7,9:18)) %>% filter(STNAME=="Florida")
popFL0009 <- popFL0009 %>% filter(CTYNAME != "Florida") %>% select(2:length(popFL0009[1,]))

popFL0009[,2:length(popFL0009[1,])] <- sapply(popFL0009[,2:length(popFL0009[1,])], as.character)
popFL0009[,2:length(popFL0009[1,])] <-sapply(popFL0009[,2:length(popFL0009[1,])], as.numeric)
```

```
popFL0009_long <- melt(popFL0009, id.vars= "CTYNAME", measure.vars= c("POPESTIMATE2000", "POPESTIMATE20

names(popFL0009_long) <- c("County", "Year", "population")

#clean up county names
popFL0009_long[,1] <- as.factor(gsub("\\ County", "", popFL0009_long[,1]))
popFL0009_long[,1] <- as.factor(gsub("Miami-Dade", "Miami Dade", popFL0009_long[,1]))
popFL0009_long[,2] <- as.factor(gsub("[A-Z]", "", popFL0009_long[,2]))
popFL0015 <- rbind(popFL0009_long, popFL1015_long)
popFL0015$Year <- as.integer(as.character(popFL0015$Year))

str(popFL0015)
```

```
## 'data.frame':    1072 obs. of  3 variables:
##  $ County    : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Year      : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ population: num  218611 22374 148393 26064 477819 ...
```

```
write.csv(popFL0015,"~/School/Fall 16/EDA/data/FLpopulation.csv")
```

```
#crime rates
FCdata <- inner_join(FL_Crime_0612, popFL0015)

FCdata <- FCdata %>% mutate(Assault_Rate = 10000*Simple.Assault/population) %>% mutate(Manslaughter_Rat

str(FCdata)
```

```
## 'data.frame':    469 obs. of  18 variables:
##  $ County              : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Manslaughter        : int  1 1 3 0 6 8 0 0 1 0 ...
##  $ Kidnap..Abduction   : int  17 0 12 1 10 22 0 1 3 3 ...
##  $ Arson               : int  8 0 5 1 24 39 0 7 8 7 ...
##  $ Simple.Assault      : int  1427 62 1394 189 2815 5367 76 599 618 795 ...
##  $ Drug.Arrest         : int  2697 163 2020 217 5282 16466 154 872 741 1244 ...
##  $ Bribery             : int  1 0 1 0 0 1 0 1 1 4 ...
##  $ Embezzlement        : int  27 0 63 6 48 233 0 34 1 0 ...
##  $ Fraud               : int  1133 0 377 14 400 724 0 80 201 565 ...
##  $ Counterfeit..Forgery: int  52 7 93 18 74 310 8 28 15 121 ...
##  $ Extortion..Blackmail: int  2 0 2 0 7 12 0 0 1 2 ...
##  $ Year                : num  2006 2006 2006 2006 2006 ...
##  $ population          : num  239506 25571 165644 28506 535138 ...
##  $ Assault_Rate        : num  59.6 24.2 84.2 66.3 52.6 ...
##  $ Manslaughter_Rate   : num  0.0418 0.3911 0.1811 0 0.1121 ...
##  $ Drug_Rate           : num  112.6 63.7 121.9 76.1 98.7 ...
##  $ Embezzlement_Rate   : num  1.127 0 3.803 2.105 0.897 ...
##  $ Fraud_Rate          : num  47.31 0 22.76 4.91 7.47 ...
```

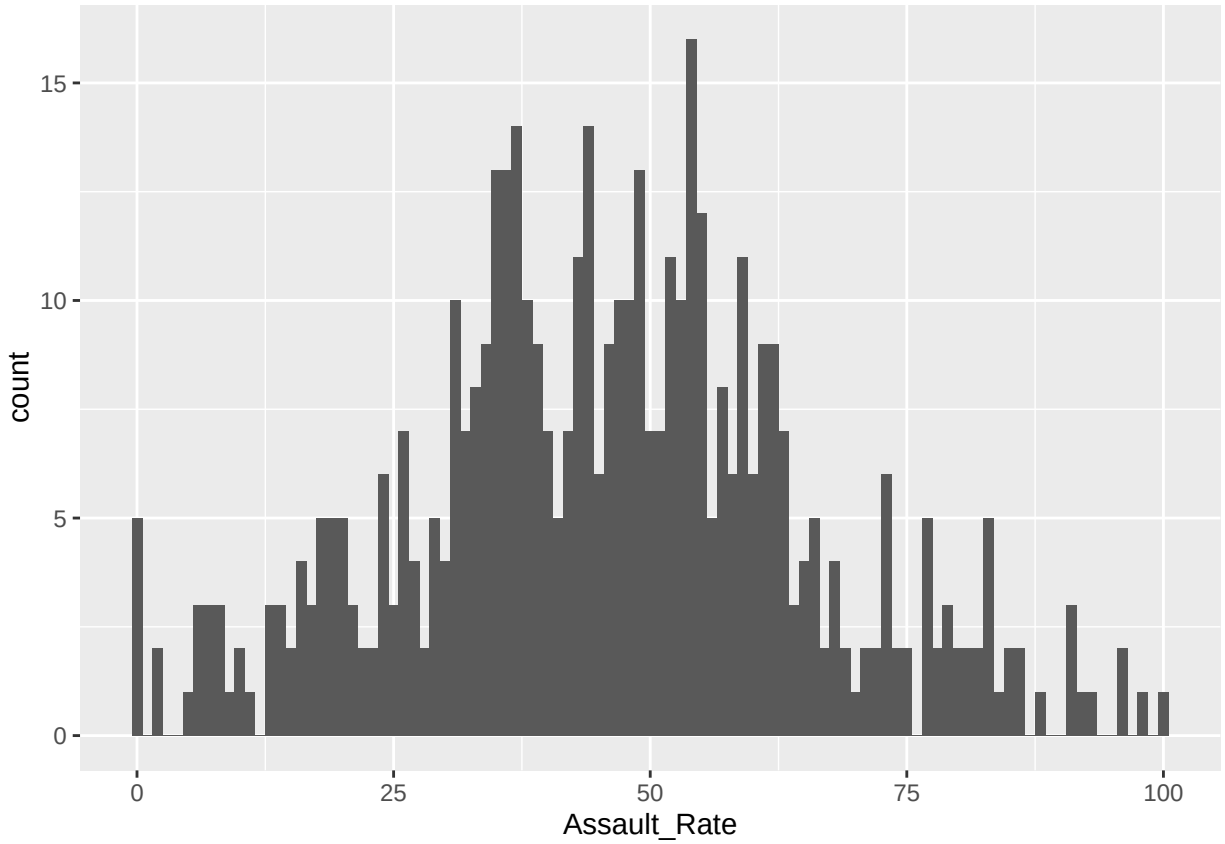
```
#join crime, edu, and unemployment dfs
big_frame <- full_join(FCdata,FL_edu_0912, by = c("County","Year"))
big_frame <- full_join(big_frame, FL_Unemployment_0612, by = c("County","Year"))
str(big_frame)
```

```
## 'data.frame':    469 obs. of  22 variables:
##  $ County              : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Manslaughter        : int  1 1 3 0 6 8 0 0 1 0 ...
##  $ Kidnap..Abduction   : int  17 0 12 1 10 22 0 1 3 3 ...
##  $ Arson               : int  8 0 5 1 24 39 0 7 8 7 ...
##  $ Simple.Assault      : int  1427 62 1394 189 2815 5367 76 599 618 795 ...
##  $ Drug.Arrest         : int  2697 163 2020 217 5282 16466 154 872 741 1244 ...
##  $ Bribery             : int  1 0 1 0 0 1 0 1 1 4 ...
##  $ Embezzlement        : int  27 0 63 6 48 233 0 34 1 0 ...
##  $ Fraud               : int  1133 0 377 14 400 724 0 80 201 565 ...
##  $ Counterfeit..Forgery: int  52 7 93 18 74 310 8 28 15 121 ...
##  $ Extortion..Blackmail: int  2 0 2 0 7 12 0 0 1 2 ...
##  $ Year                : num  2006 2006 2006 2006 2006 ...
##  $ population          : num  239506 25571 165644 28506 535138 ...
##  $ Assault_Rate        : num  59.6 24.2 84.2 66.3 52.6 ...
##  $ Manslaughter_Rate   : num  0.0418 0.3911 0.1811 0 0.1121 ...
##  $ Drug_Rate           : num  112.6 63.7 121.9 76.1 98.7 ...
##  $ Embezzlement_Rate   : num  1.127 0 3.803 2.105 0.897 ...
##  $ Fraud_Rate          : num  47.31 0 22.76 4.91 7.47 ...
##  $ pop25up             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ HS25up              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ bachelors25up       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Unemployment_Rate   : num  2.7 2.8 3.1 2.9 3.3 3.1 3.2 3.4 3.8 3 ...
```

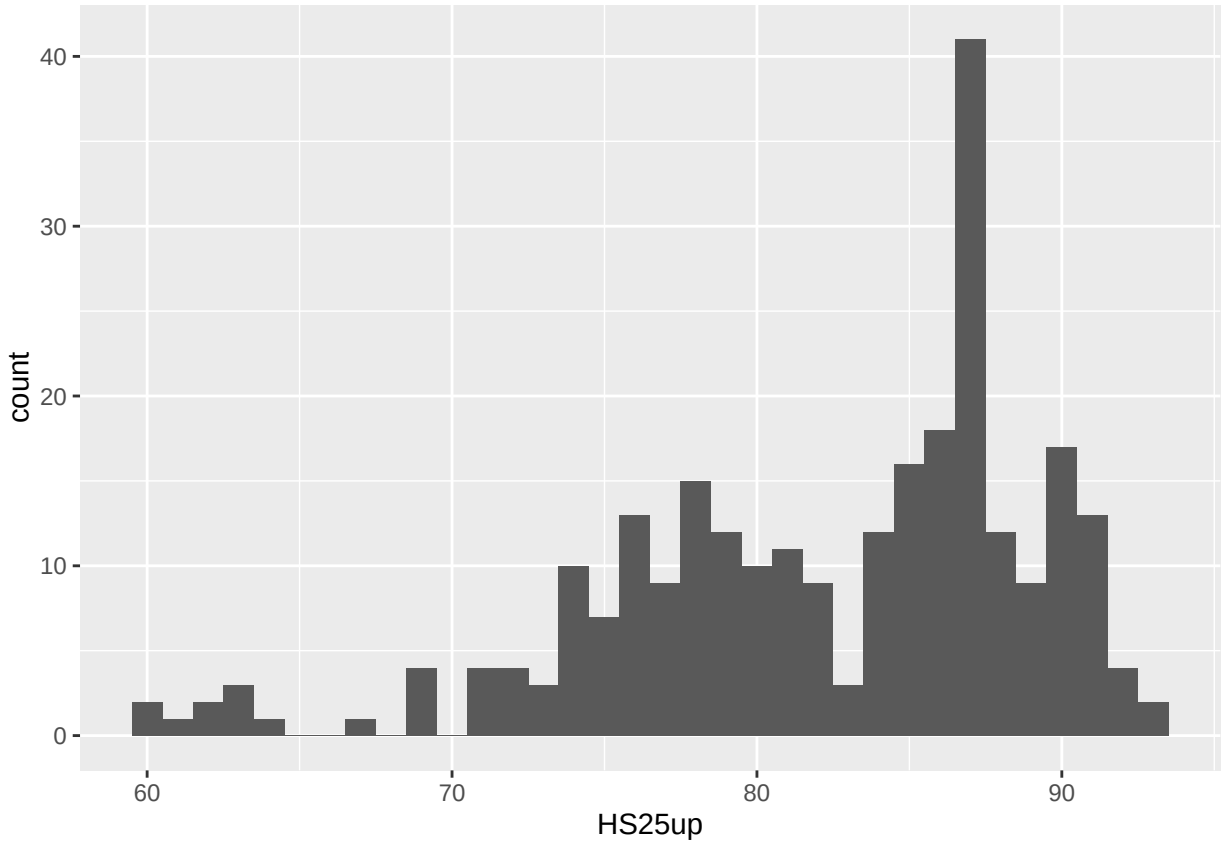
now for the fun stuff!

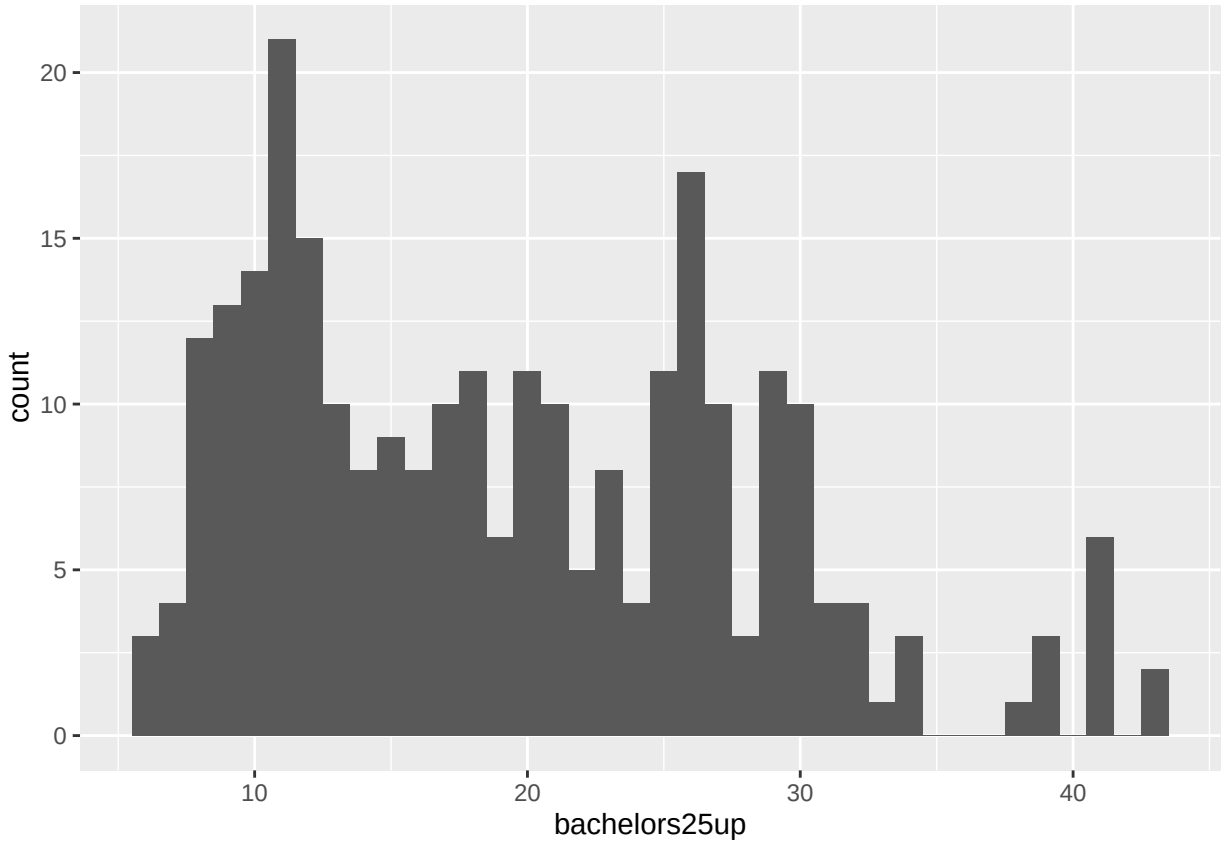
Distributions

Assault Rate:

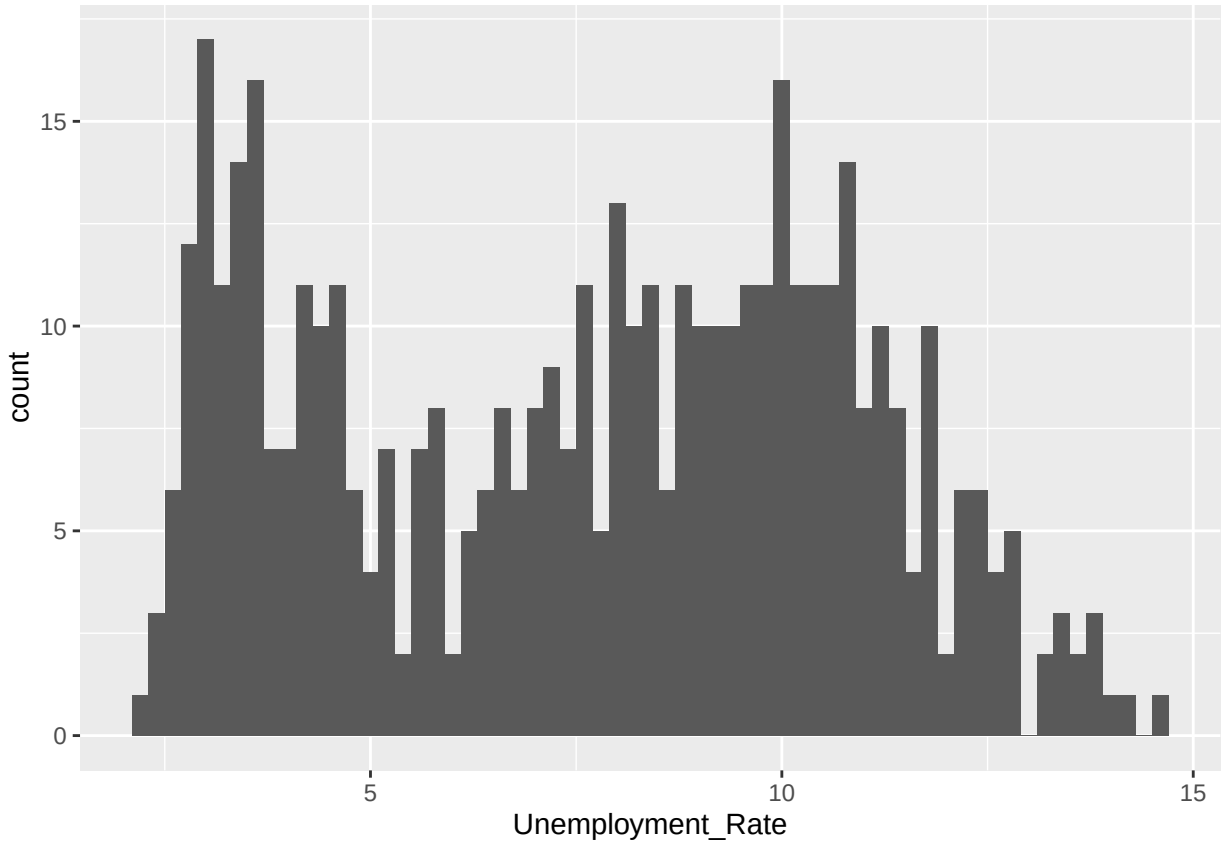


Education:


40
30
20
10
0
count
60
70
80
90
HS25up

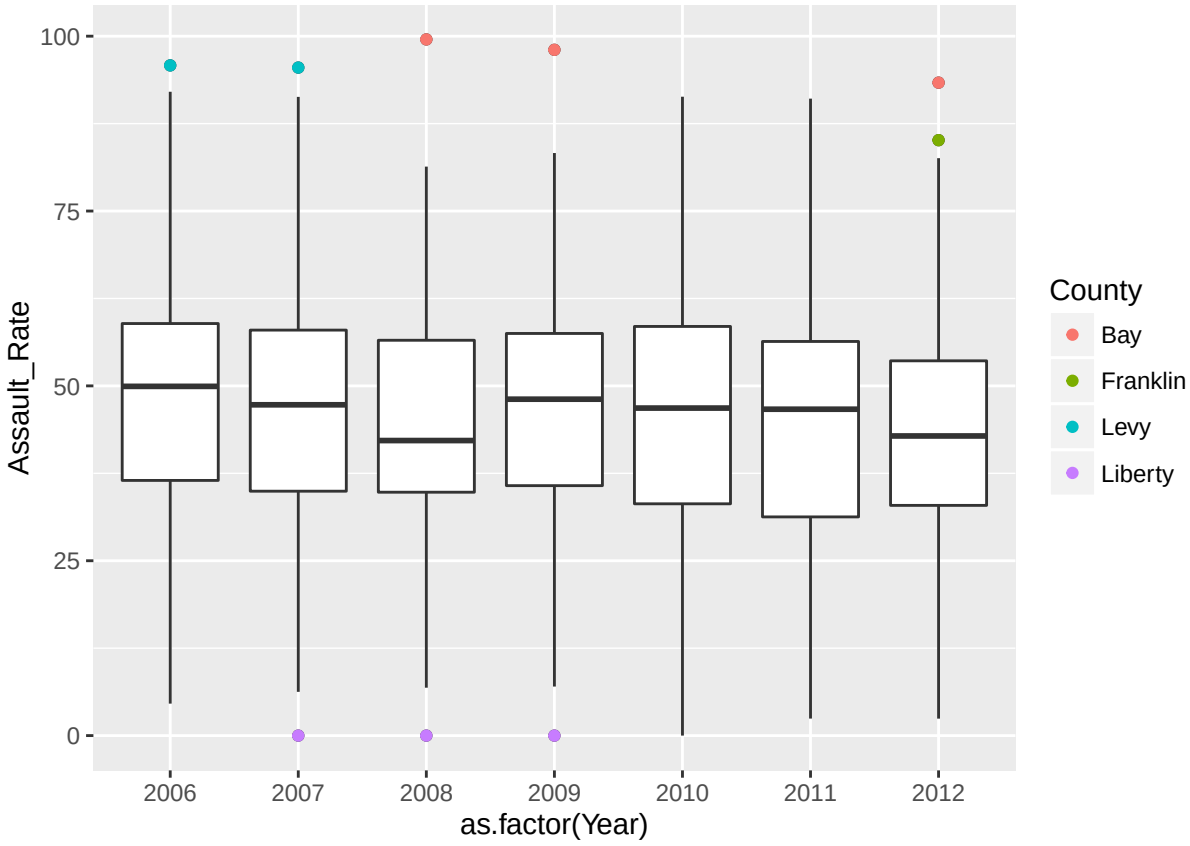


Unemployment:

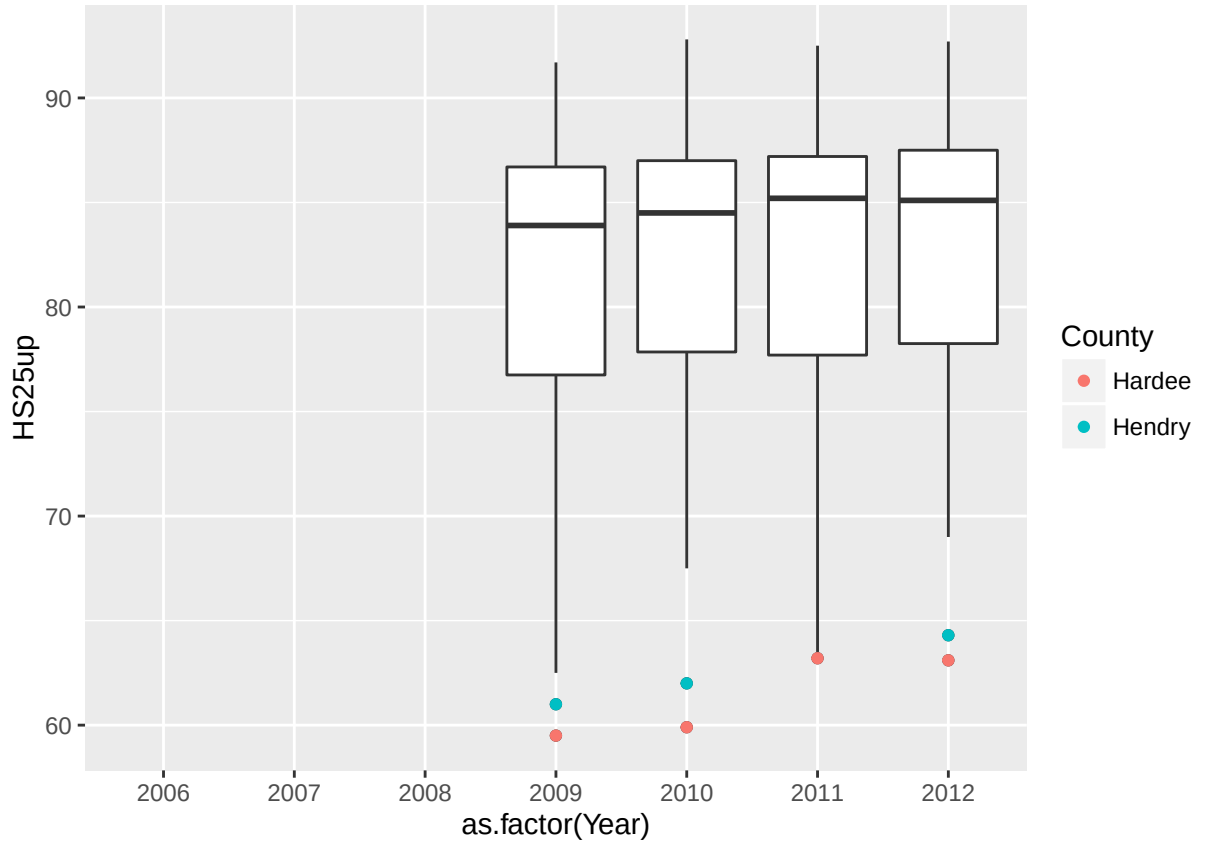


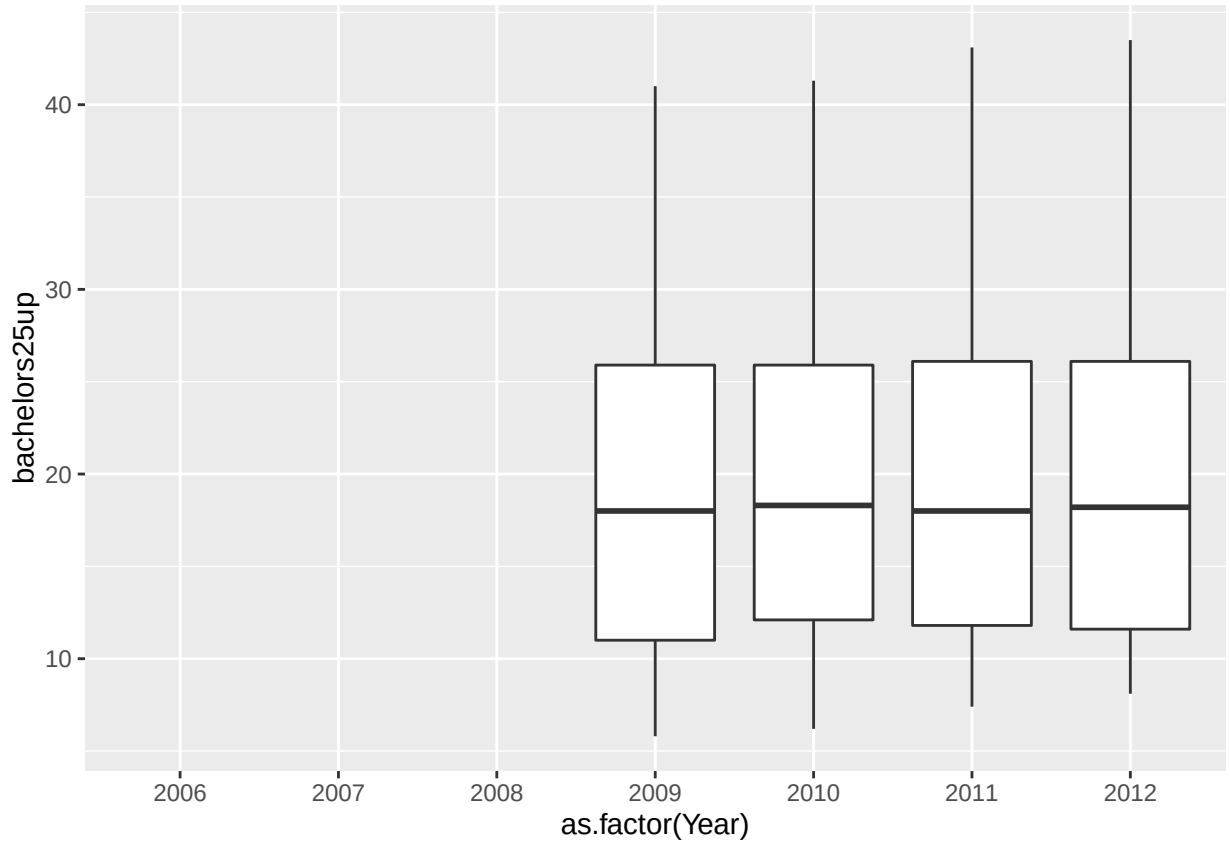
Boxplots

Assault Rates

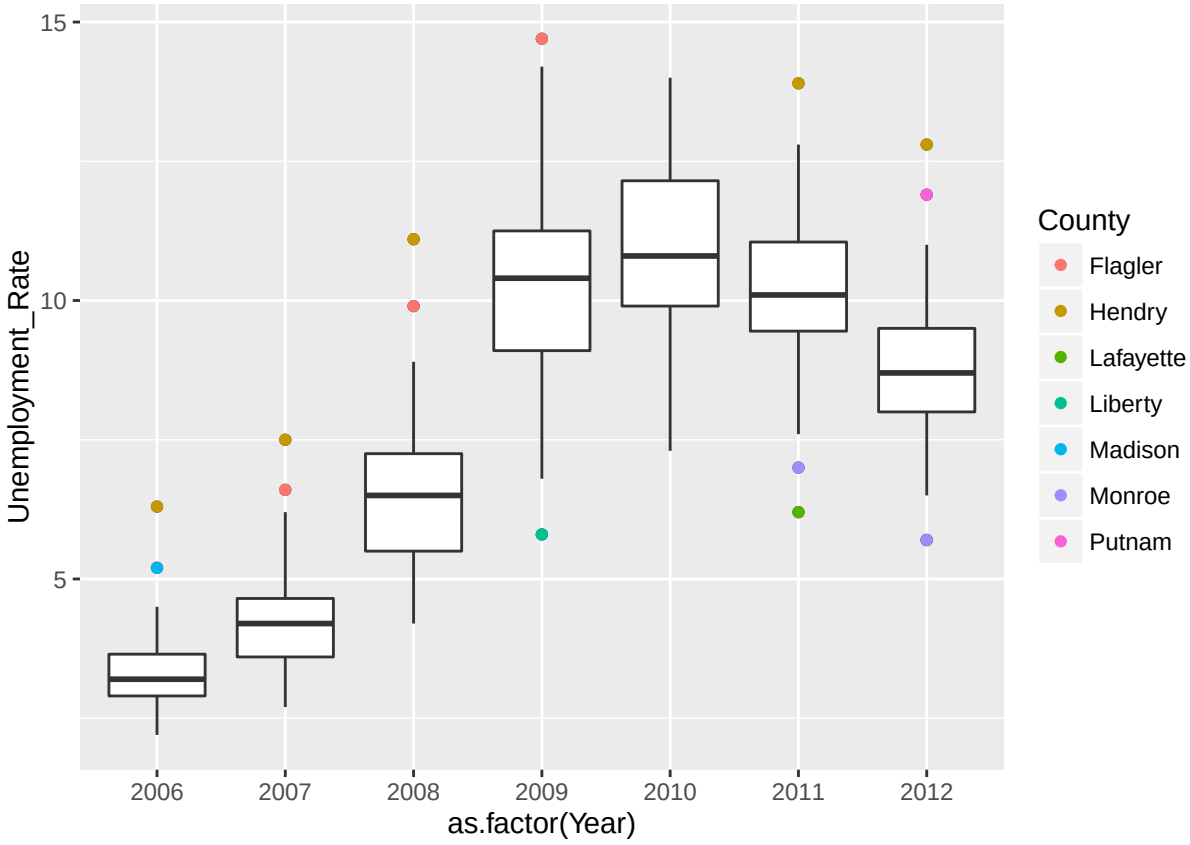


Educational Attainment

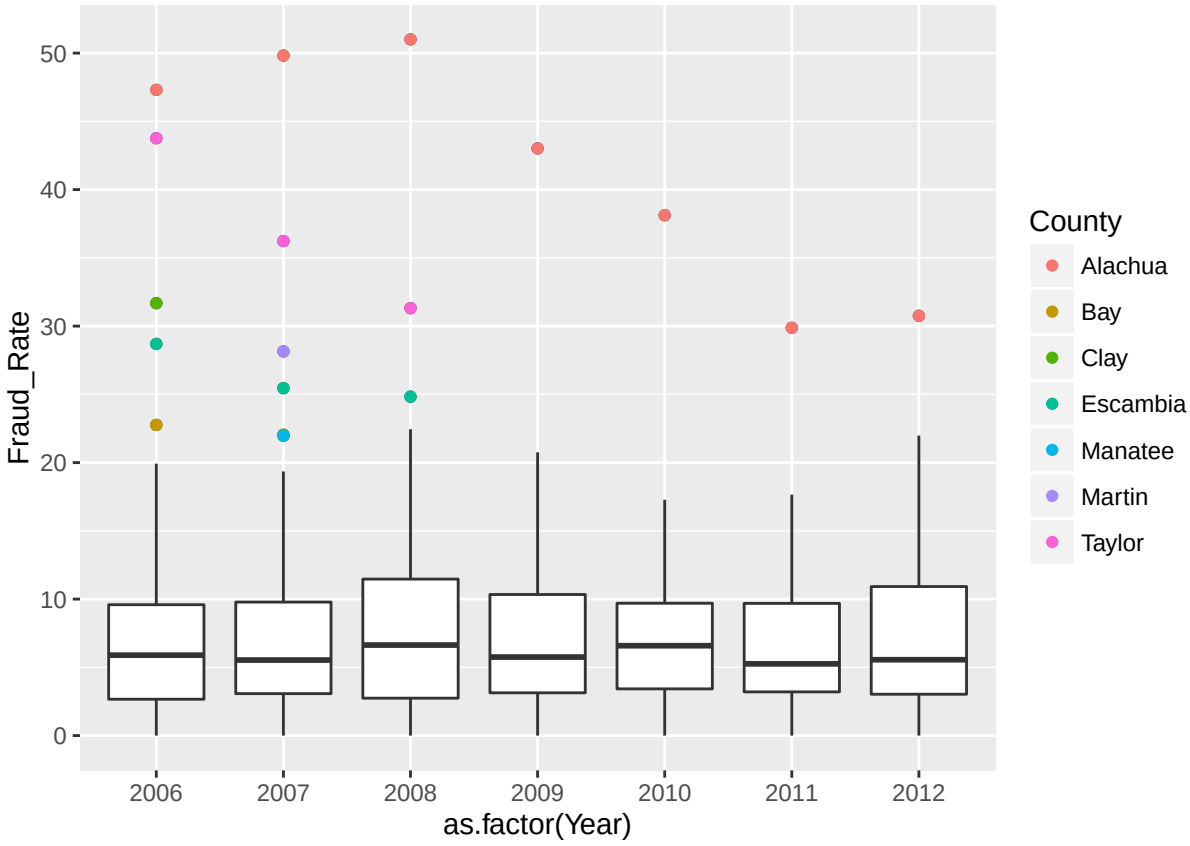

HS25up
90
80
70
60
2006
2007
2008
2009
2010
2011
2012
as.factor(Year)
County
Hardee
Hendry


bachelors25up
40
30
20
10
2006
2007
2008
2009
2010
2011
2012
as.factor(Year)

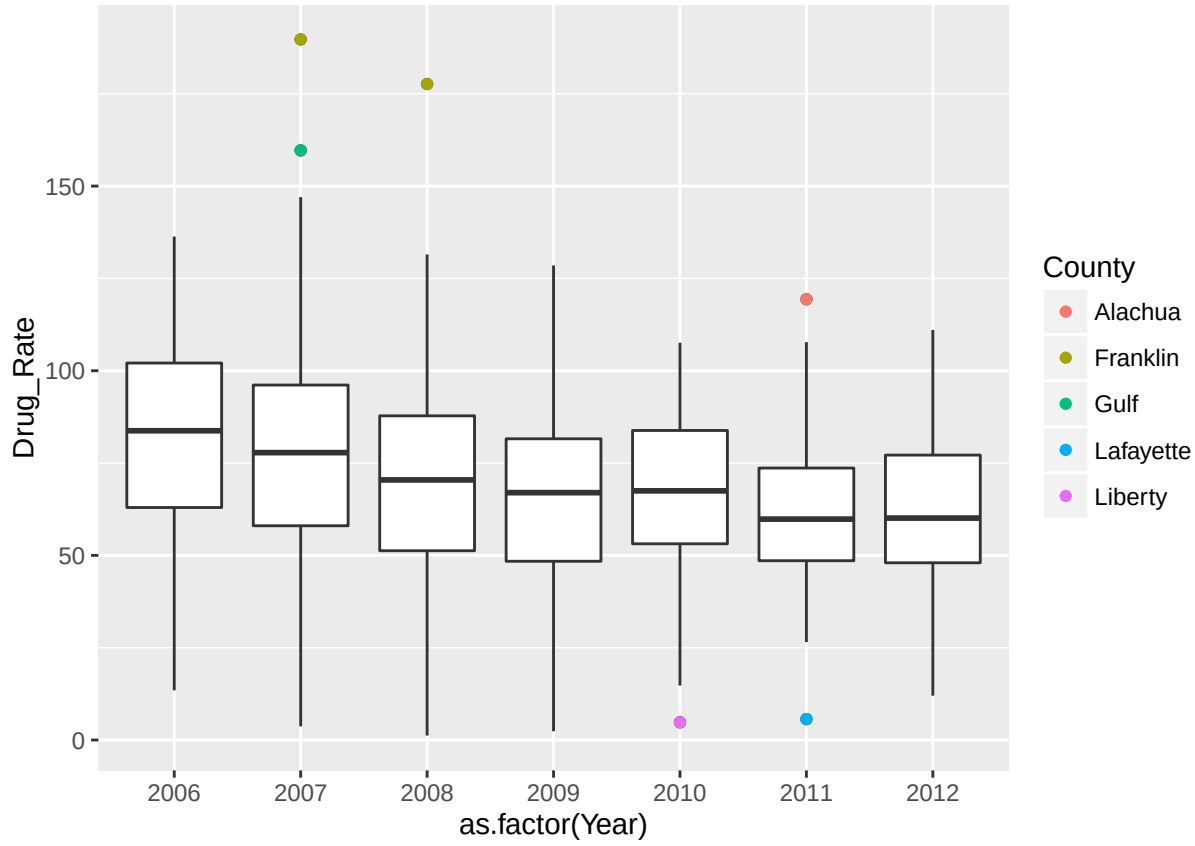
Unemployment rates



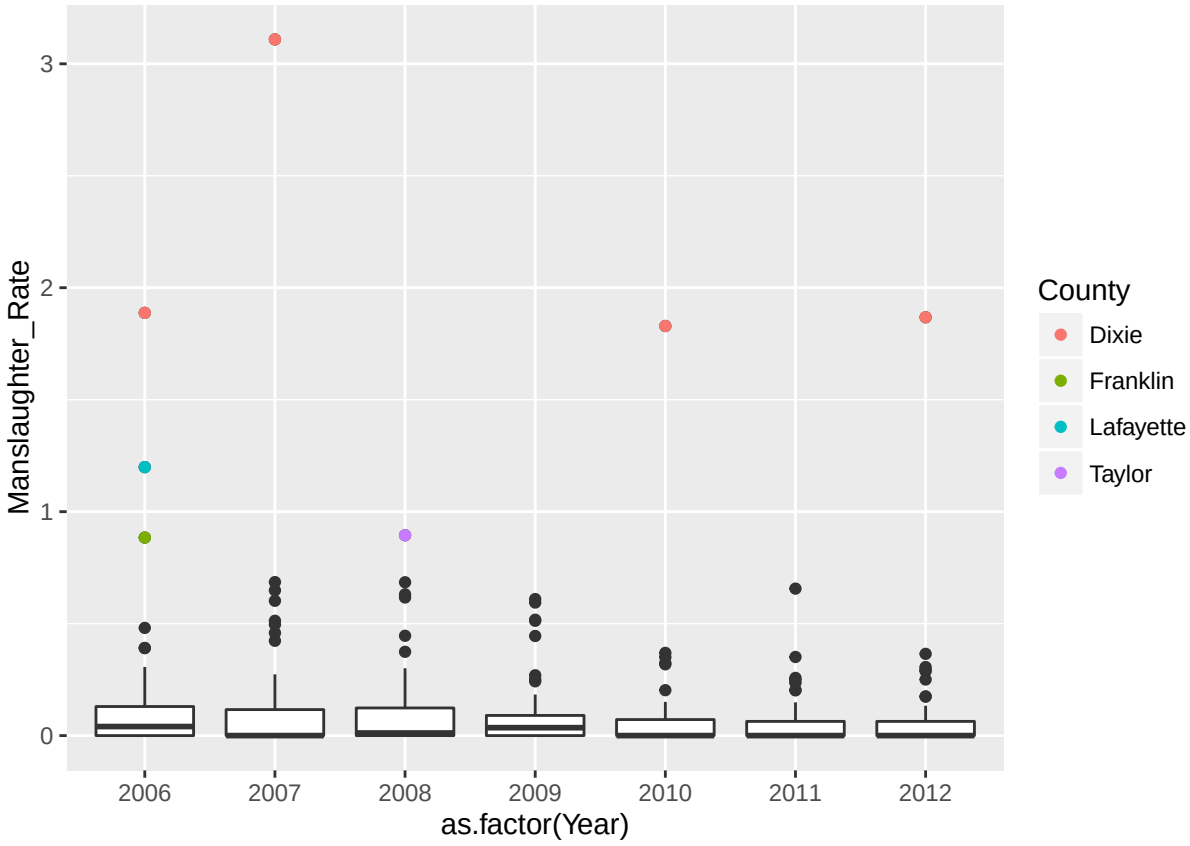
Fraud rates



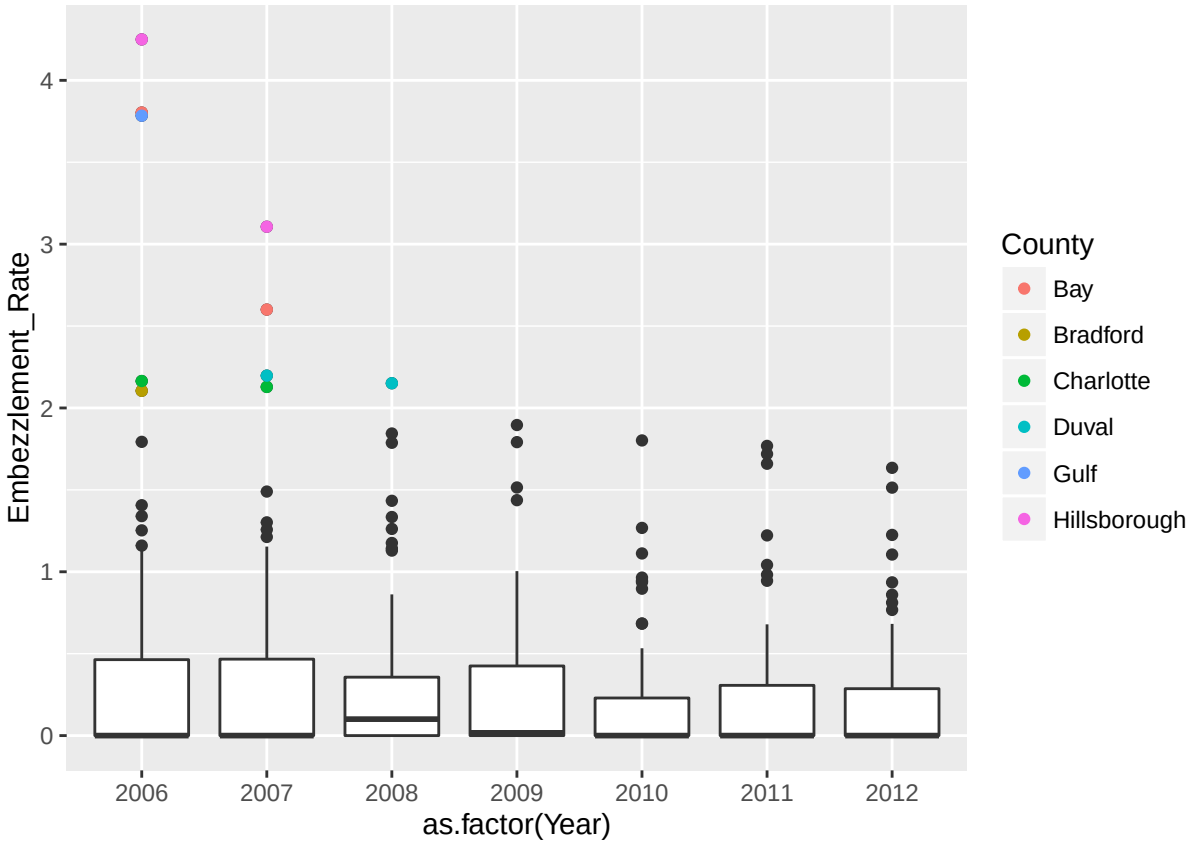
Drug arrest rates



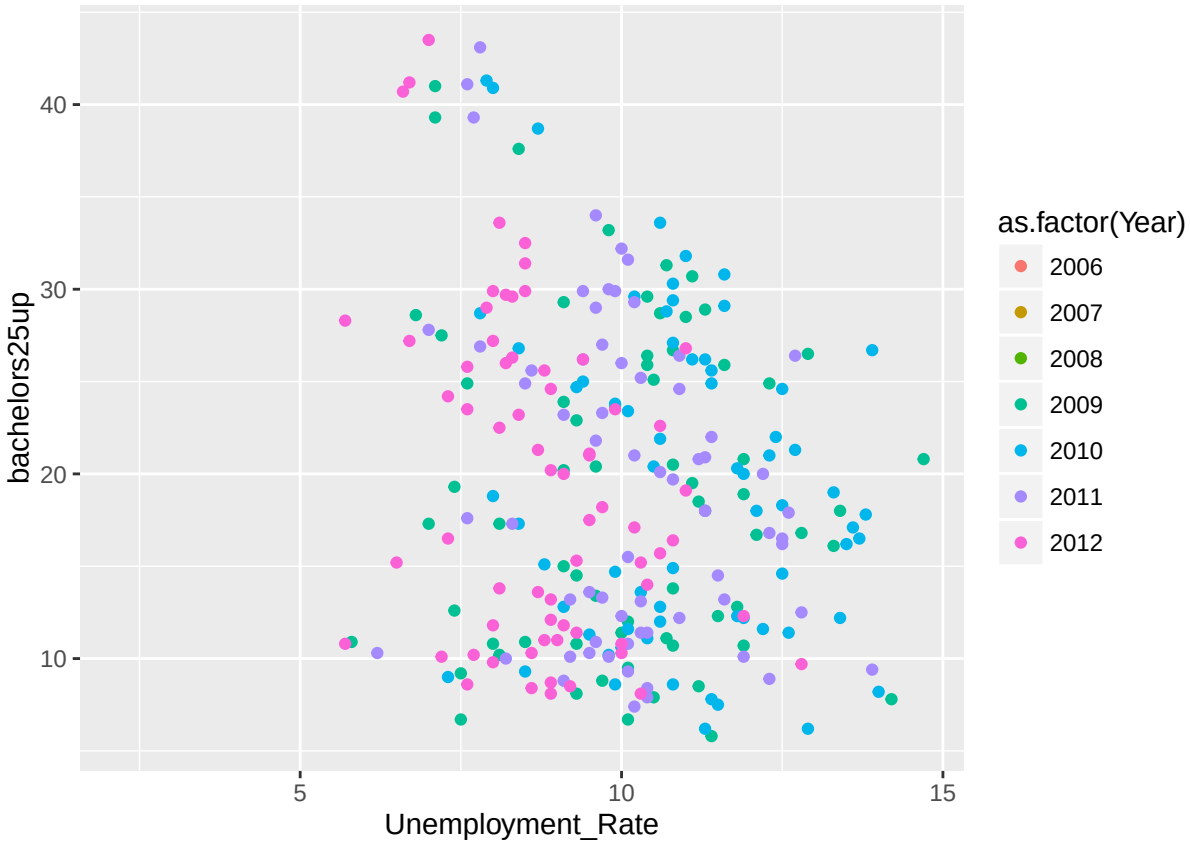
Manslaughter rates

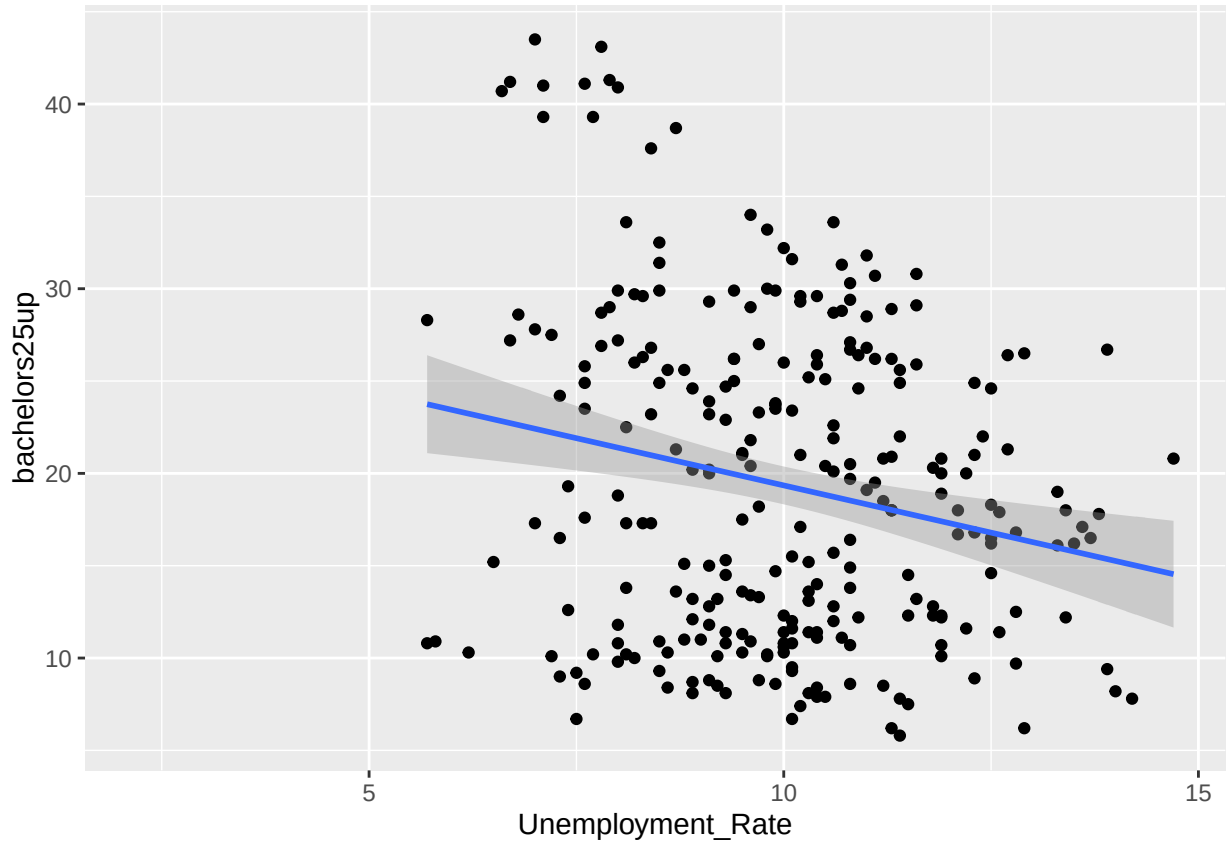


Embezzlement rates


Embezzlement_Rate
4
3
2
1
0
2006
2007
2008
2009
2010
2011
2012
as.factor(Year)
County
Bay
Bradford
Charlotte
Duval
Gulf
Hillsborough

unemployment vs education

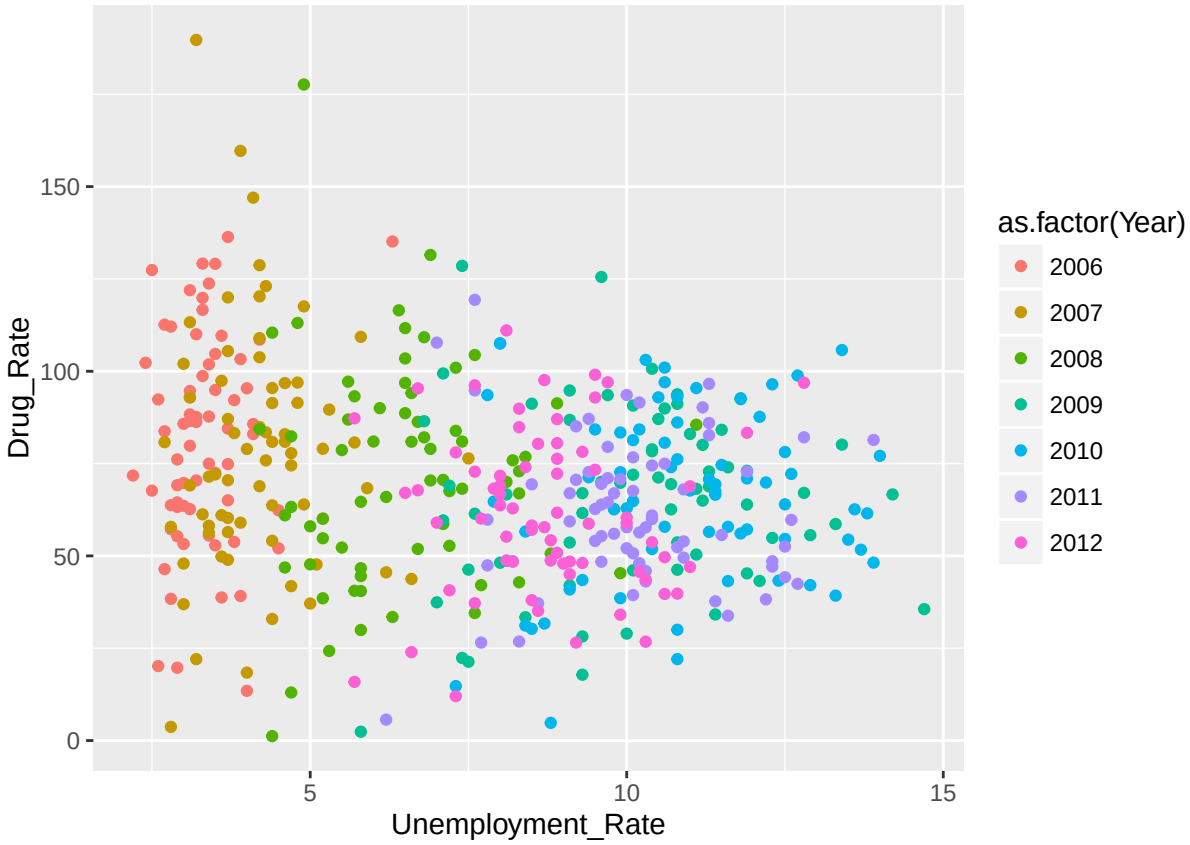


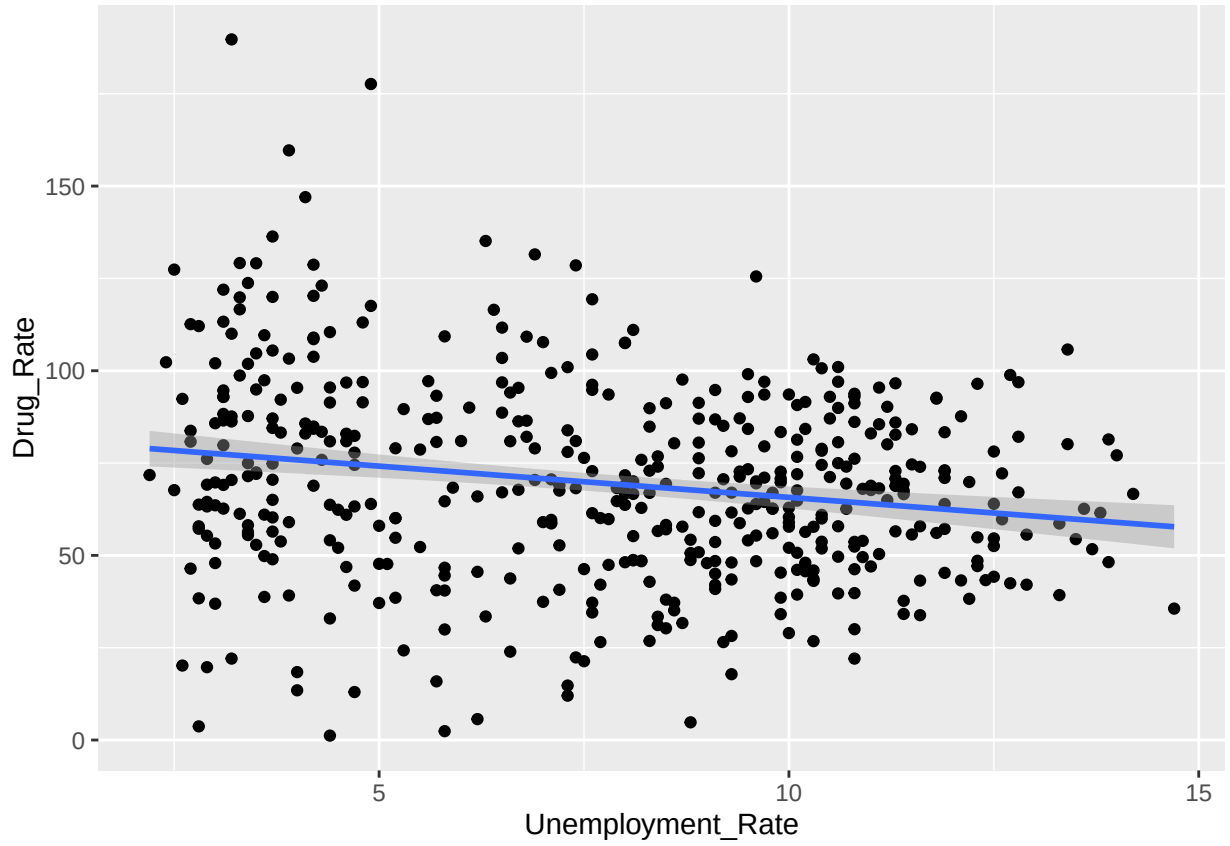



```
cor(big_frame$Unemployment_Rate,big_frame$bachelors25up,use="complete")
```

```
## [1] -0.2110115
```

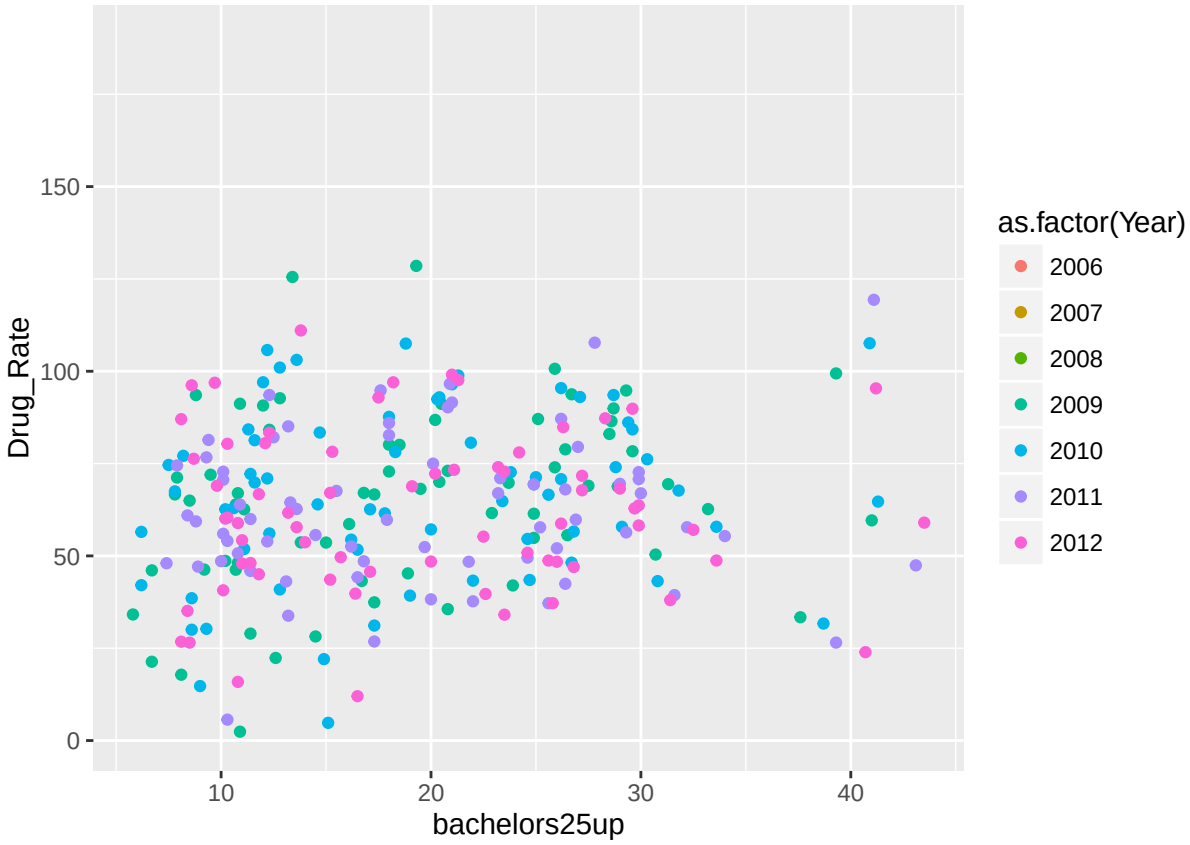
Unemployment and drug arrests

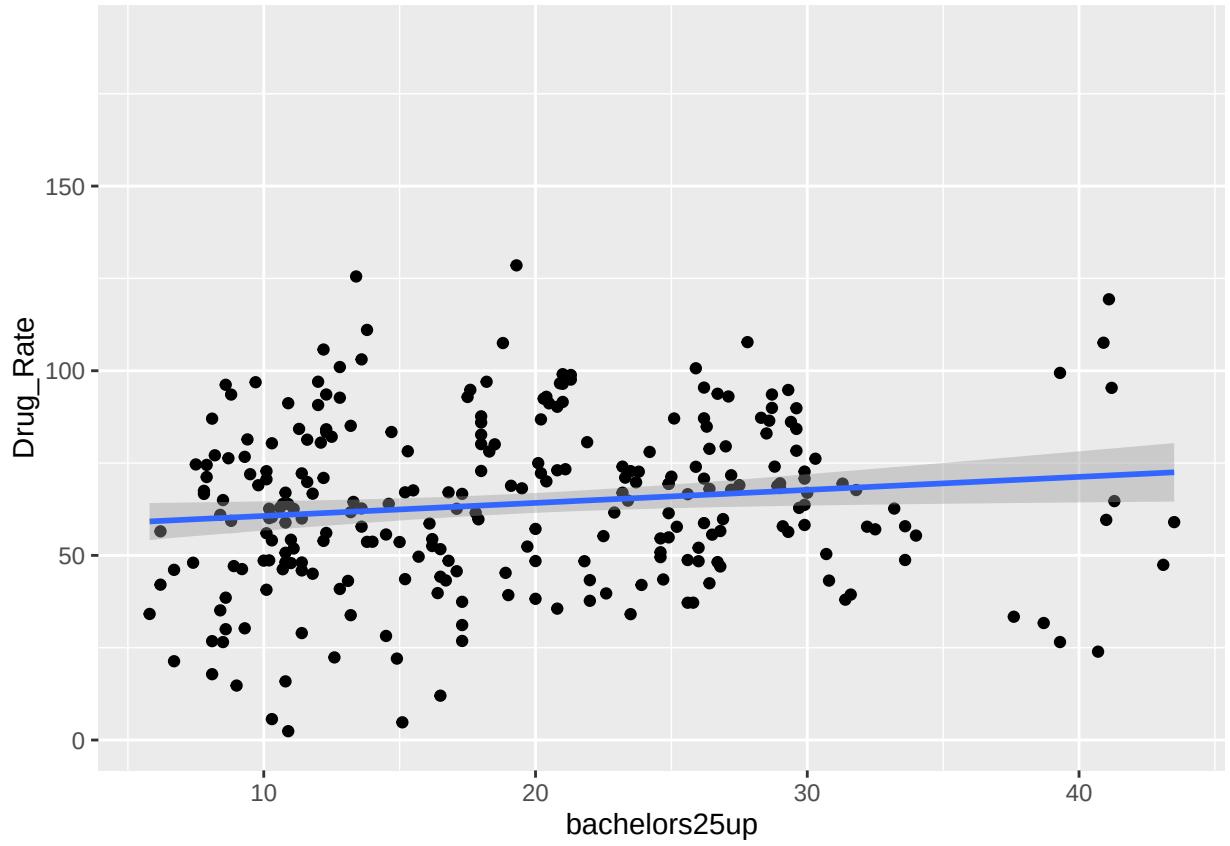





```
## [1] -0.1983548
```

Education and drug arrests

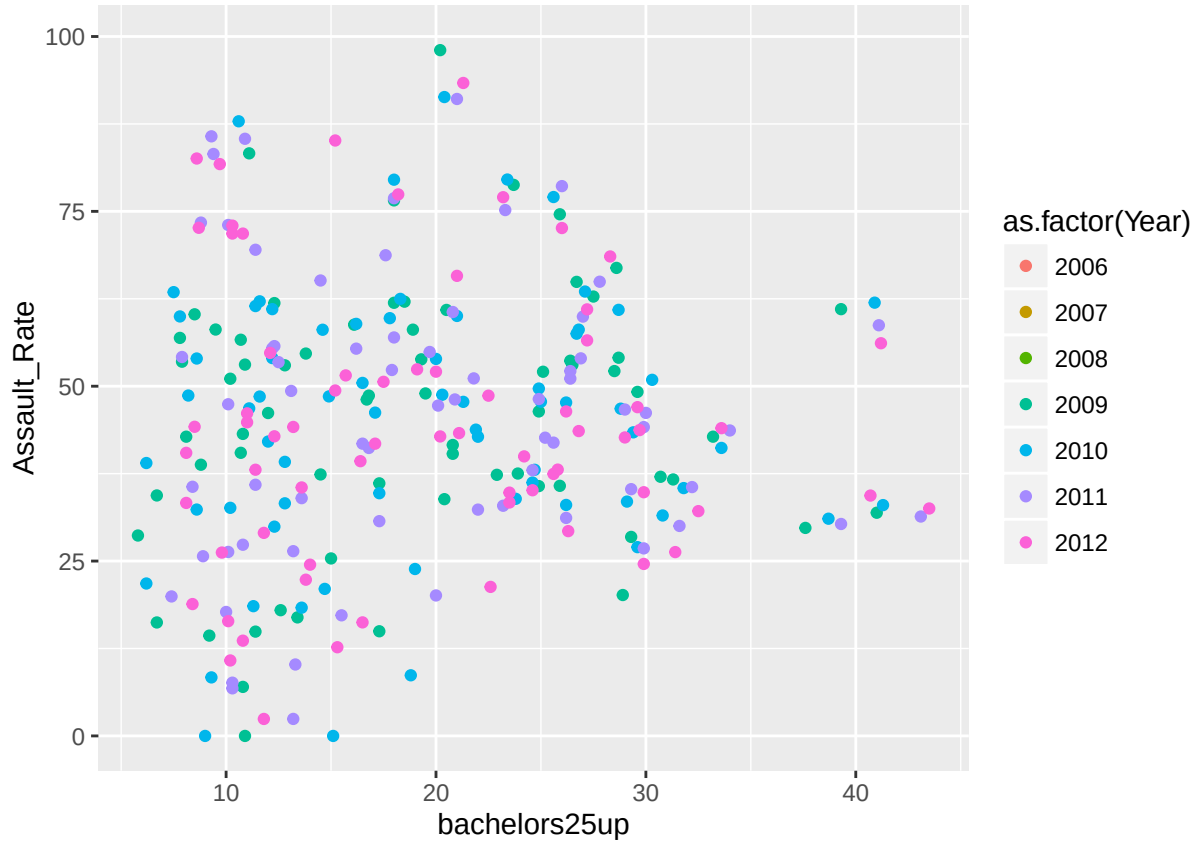


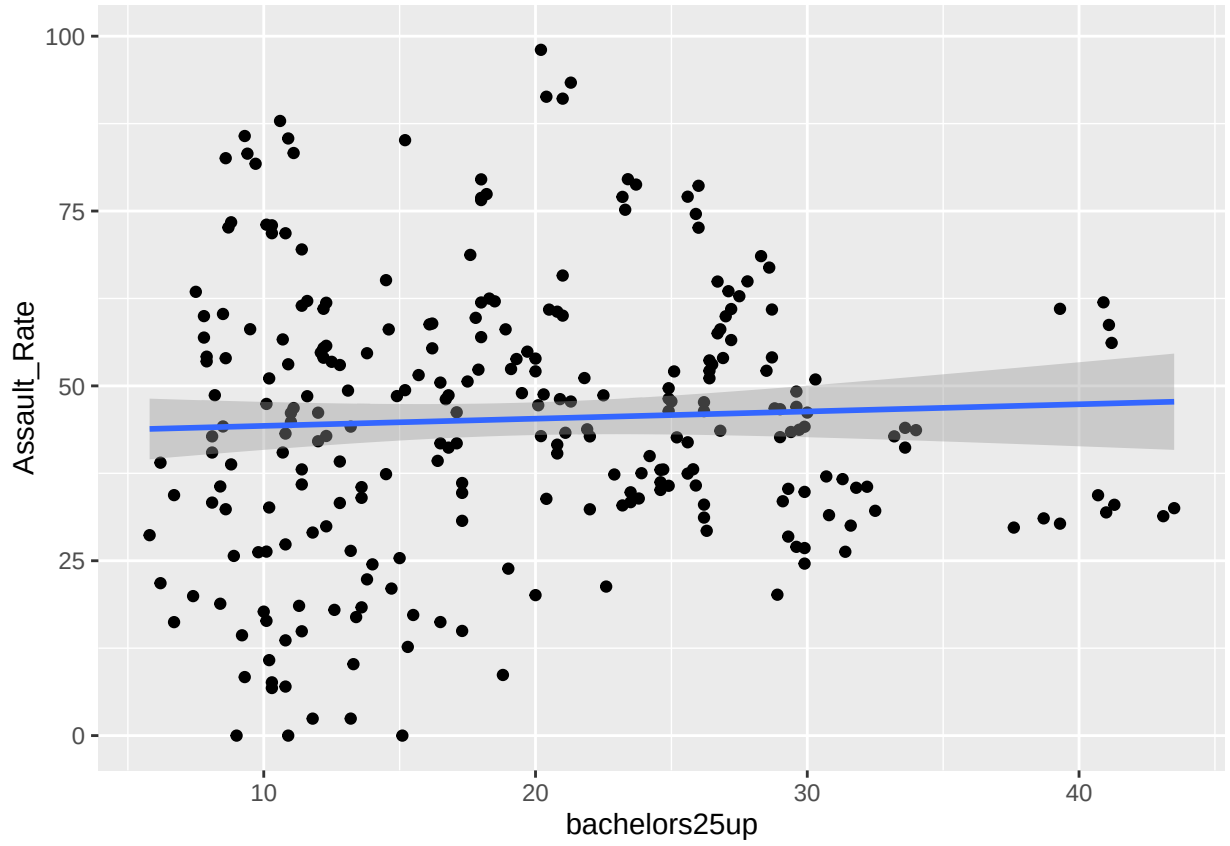



```
cor(big_frame$bachelors25up,big_frame$Drug_Rate,use="complete")
```

```
## [1] 0.1364718
```

education and assault

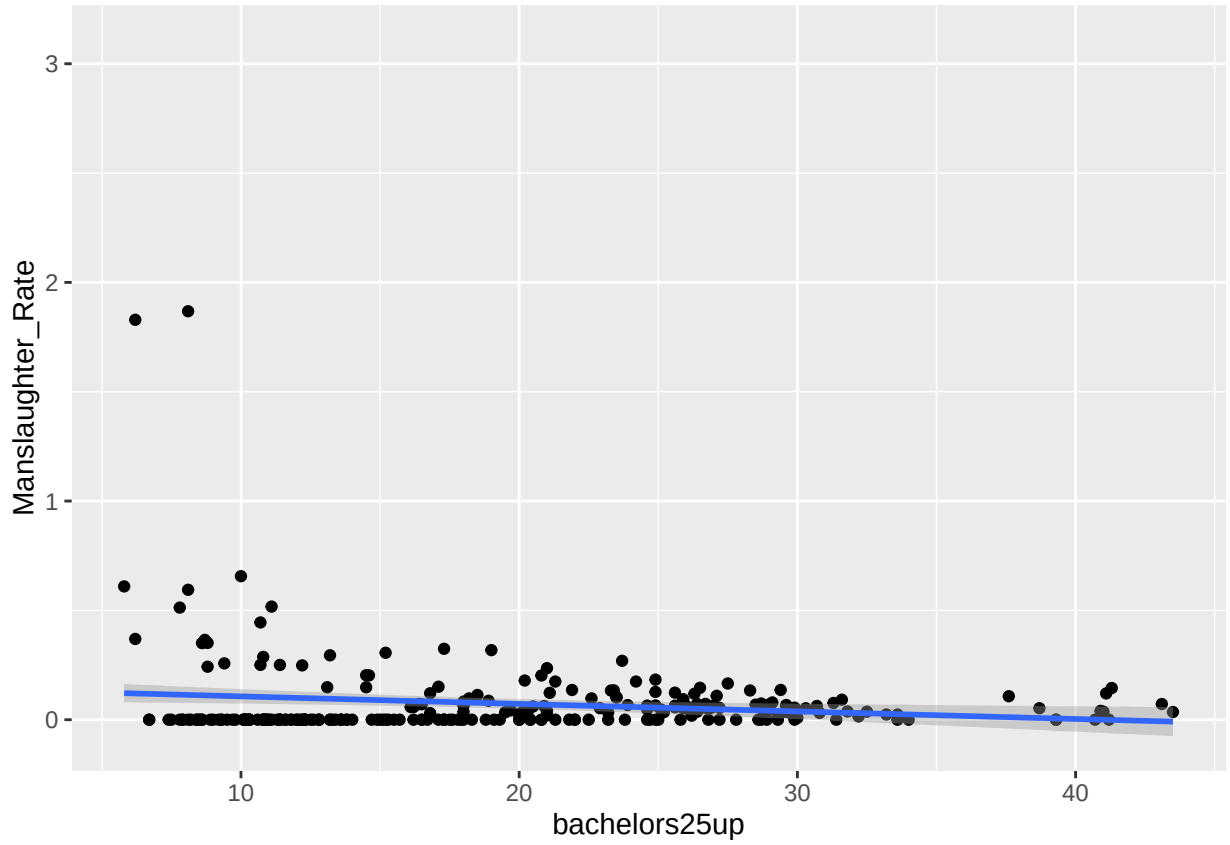





```
cor(big_frame$bachelors25up,big_frame$Assault_Rate,use="complete")
```

```
## [1] 0.04629918
```

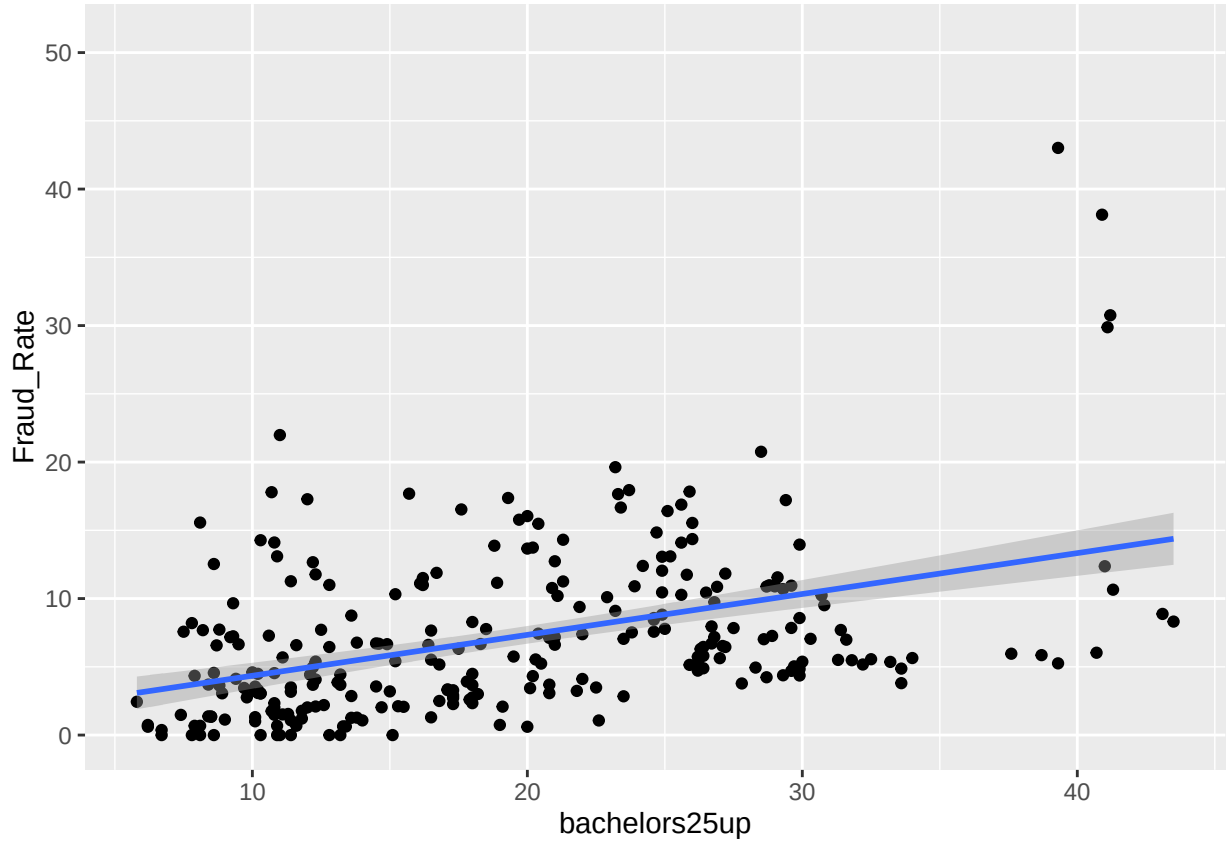
education and manslaughter




```
cor(big_frame$bachelors25up,big_frame$Manslaughter_Rate,use="complete")
```

```
## [1] -0.1589908
```

education and fraud




```
cor(big_frame$bachelors25up,big_frame$Fraud_Rate,use="complete")
```

```
## [1] 0.437218
```