152_project code

Lucy Wang, Michelle Seo

5/3/2020

##EDA

##Plots/EDA

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##filling in total_staff male/female calculation
##invalid data = 0 staff members
data15_18$TOTALSTAFFF <- data15_18$CORRSTAFFF+data15_18$OTHERSTAFFF
data15_18$TOTALSTAFFM <- data15_18$TOTALSTAFF - data15_18$TOTALSTAFFF
data15_18$STRATUM <- as.factor(data15_18$STRATUM)

summary(data15_18$TOTALSTAFF)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -9.0    32.0    78.0   181.2   190.5 11449.0
```

```
summary(data15_18$CORRSTAFFF)
```

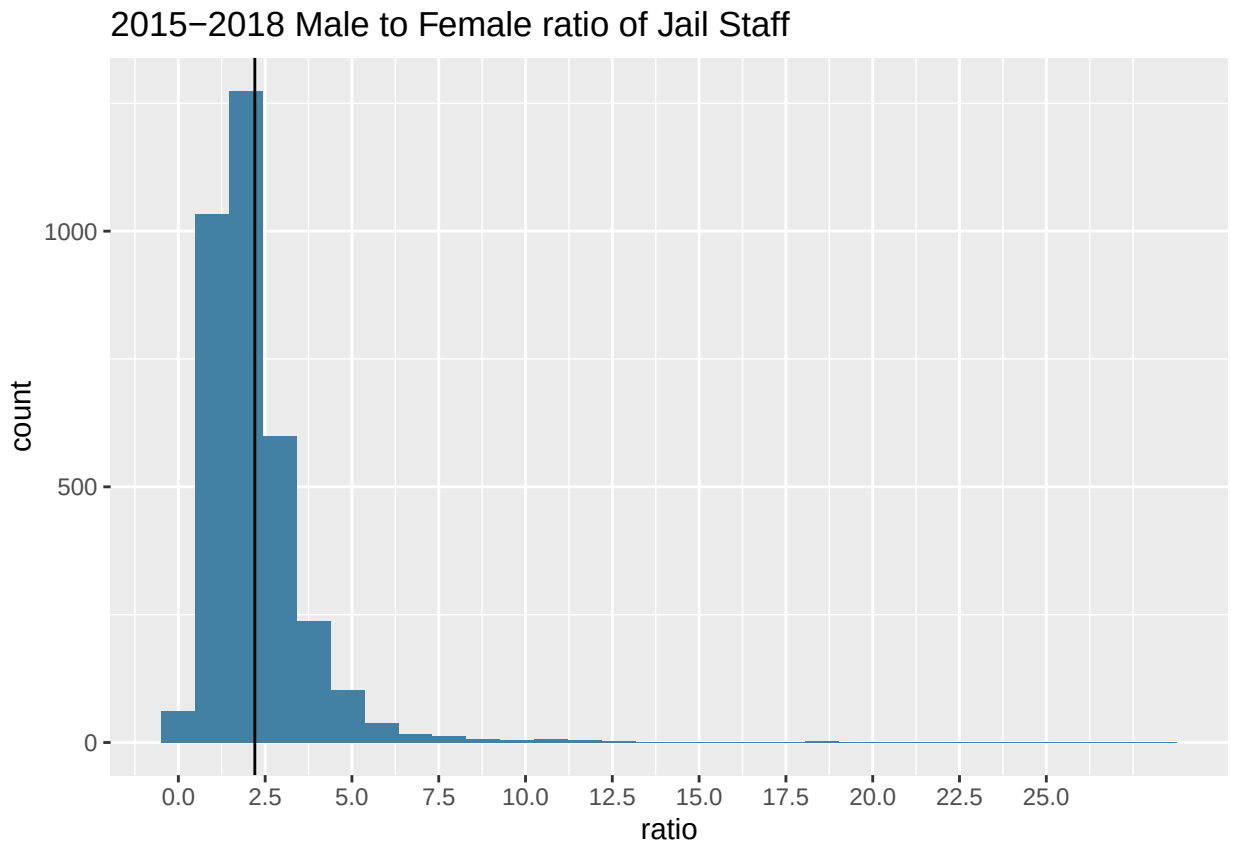
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -9.00    8.00   17.00   42.71   37.00 4370.00
```

```
summary(data15_18$OTHERSTAFFF)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -9.00    2.00    7.00   21.73   22.00  513.00
```

```
indices <- which(data15_18$TOTALSTAFF==0)

data15_18$ratio <- data15_18$TOTALSTAFFM/data15_18$TOTALSTAFFF
data15_18$ratio[data15_18$ratio<=0 | is.infinite(data15_18$ratio)] <- NA
```

```
ratio <- data15_18$ratio[is.na(data15_18$ratio)==F]
ratio <- ratio[is.infinite(ratio)==F]

ggplot()+
geom_histogram(aes(ratio), bins = 30, fill="#4281a4")+
  ggtitle("2015-2018 Male to Female ratio of Jail Staff")+
  geom_vline(aes(xintercept=mean(ratio)))+
  scale_x_continuous(breaks = seq(0,25, 2.5))
```



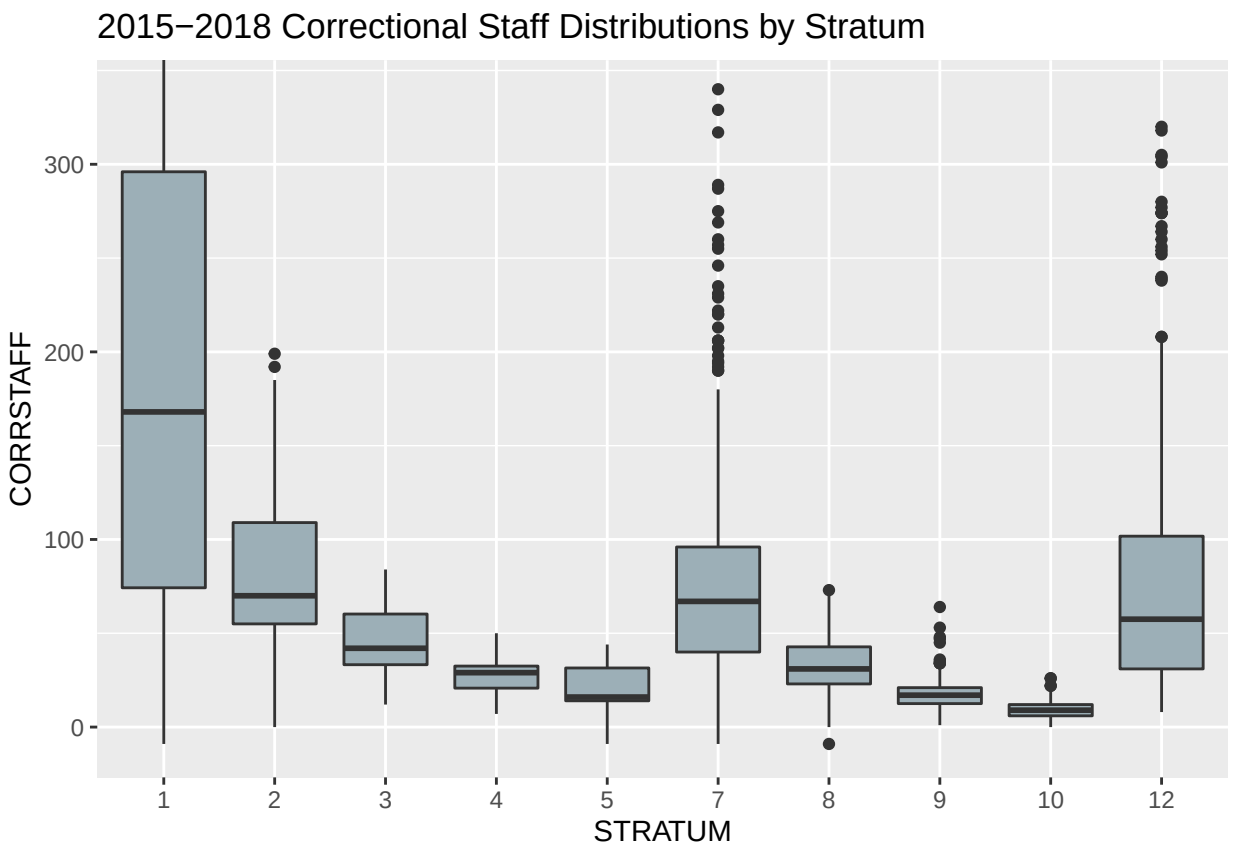

```
data15_18 %>% group_by(STRATUM) %>% summarise(ct = n(), avg=mean(ratio, na.rm = T))
```

```
## # A tibble: 10 x 3
##    STRATUM    ct   avg
##    <fct>   <int> <dbl>
##  1 1        1386  2.33
##  2 2         133  2.76
##  3 3          70  1.99
##  4 4          32  2.07
##  5 5          44  2.19
##  6 7         828  2.15
##  7 8         306  1.98
##  8 9         239  2.43
##  9 10        219  1.89
## 10 12        262  1.81
```

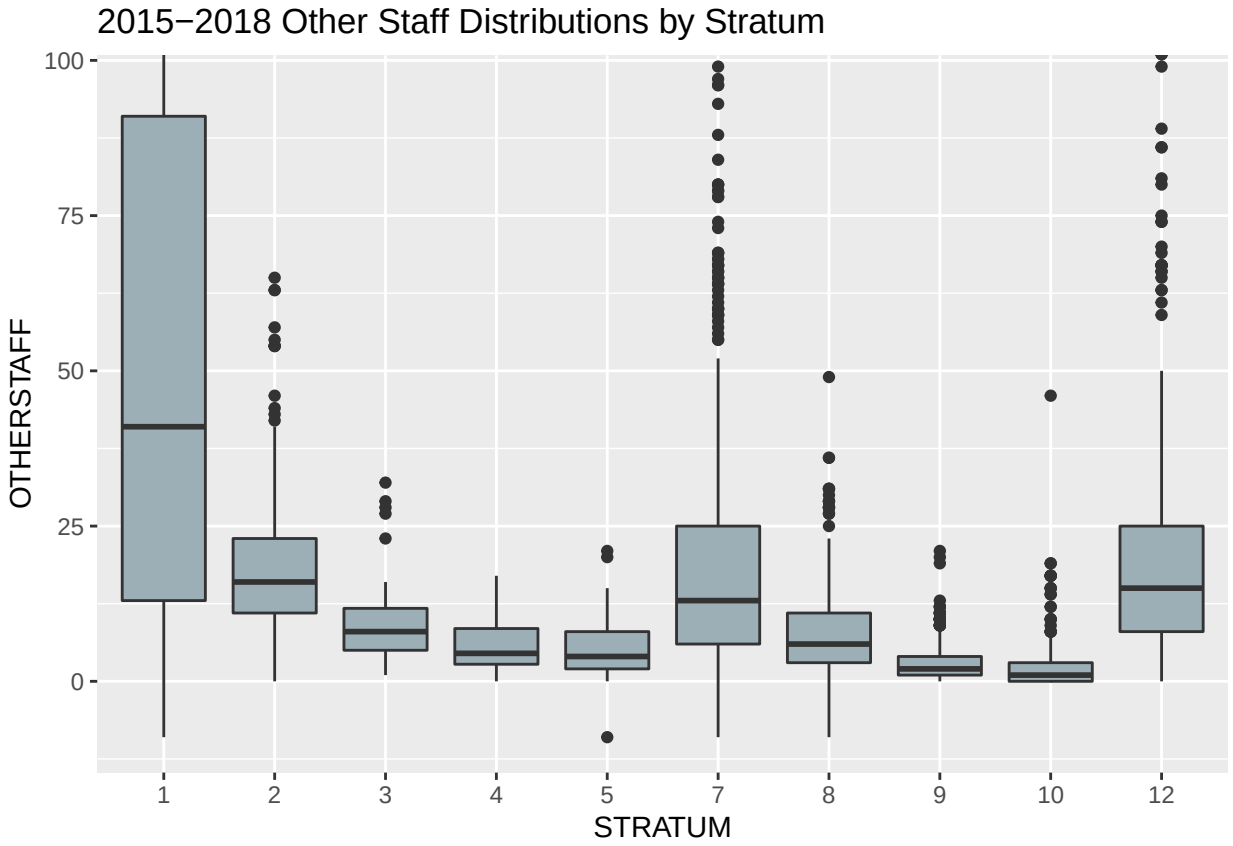
```
mean(ratio)
```

```
## [1] 2.202839
```

```
sts <- boxplot.stats(data15_18$CORRSTAFF)$stats
## Correctional Staff
ggplot(data15_18, aes(x=STRATUM, y=CORRSTAFF)) +
    geom_boxplot(aes(group=STRATUM),fill="#9cafb7")+
    scale_x_discrete(breaks = seq(1,12,1))+
    coord_cartesian(ylim = c(sts*1.05,sts/1.05))+
  ggtitle("2015-2018 Correctional Staff Distributions by Stratum")
```

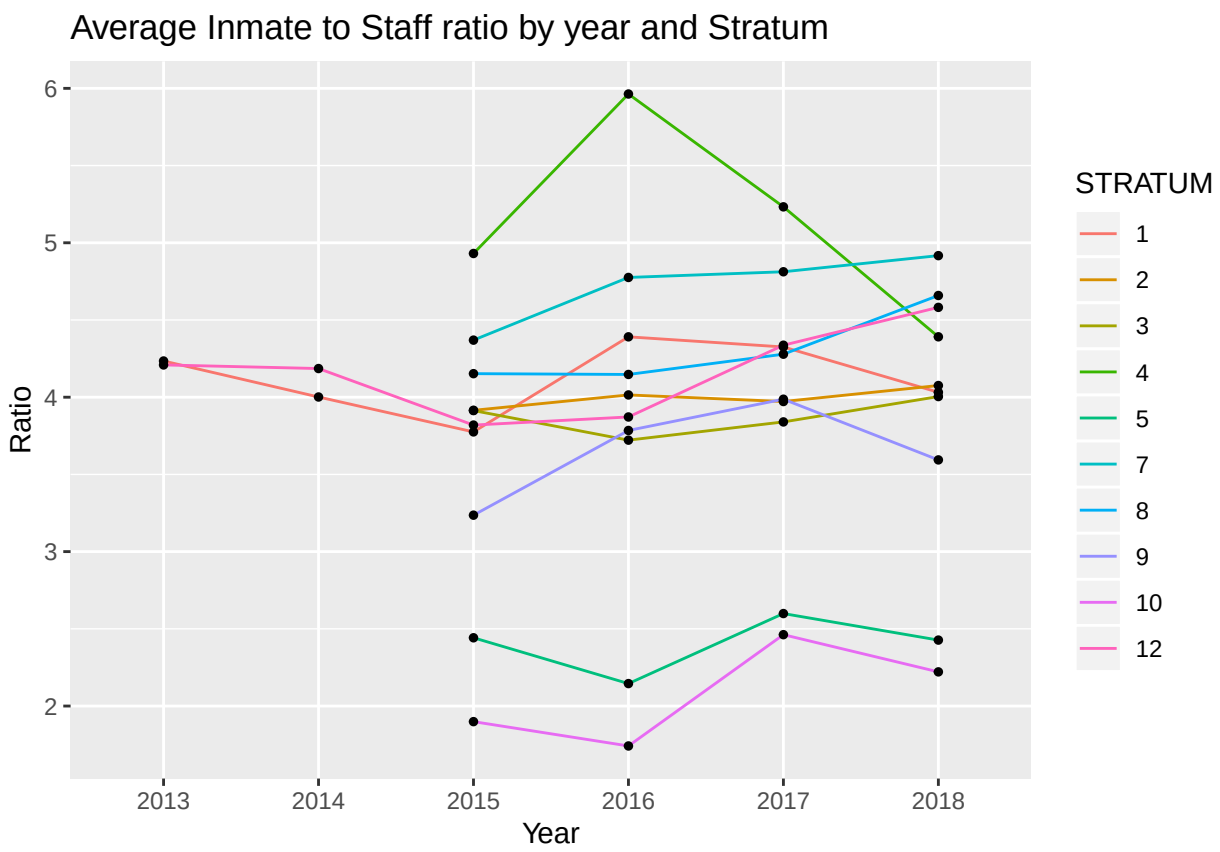



```
## Other Staff
sts <- boxplot.stats(data15_18$OTHERSTAFF)$stats
##Gender comparison per type
ggplot(data15_18, aes(x=STRATUM, y=OTHERSTAFF)) +
    geom_boxplot(aes(group=STRATUM), fill="#9cafb7")+
    scale_x_discrete(breaks = seq(1,12,1))+
    coord_cartesian(ylim = c(sts*1.05,sts/1.05))+
  ggtitle("2015-2018 Other Staff Distributions by Stratum")
```



##ADP/Staff Ratio

```
data13_18$STRATUM <- sapply(data13_18$STRATUM, round)
data13_18$year <- as.factor(data13_18$year)
data13_18$STRATUM <- as.factor(data13_18$STRATUM)

##inmate to staff ratio
data13_18$ratios2 <- data13_18$ADP/data13_18$TOTALSTAFF
data13_18$ratios2[is.infinite(data13_18$ratios2) | data13_18$ratios2 <=0] <- NA

grpRatio <- data13_18 %>% group_by(year, STRATUM) %>% summarise(avg = mean(ratios2, na.rm=T))
grpRatio <- grpRatio[is.na(grpRatio$avg)==F,]

ggplot(grpRatio, aes(year, avg))+
  geom_line(aes(group = STRATUM, color=STRATUM))+
  geom_point(size = 1)+
  ggtitle("Average Inmate to Staff ratio by year and Stratum")+
  xlab("Year")+
  ylab("Ratio")
```




```
grpRatio
```

```
## # A tibble: 44 x 3
## # Groups:   year [6]
##    year  STRATUM   avg
##    <fct> <fct>   <dbl>
##  1 2013  1        4.23
##  2 2013  12       4.21
##  3 2014  1        4.00
##  4 2014  12       4.19
##  5 2015  1        3.78
##  6 2015  2        3.92
##  7 2015  3        3.91
##  8 2015  4        4.93
##  9 2015  5        2.44
## 10 2015  7        4.37
## # ... with 34 more rows
```

```
data15_18$ratios3 <- data15_18$ADP/data15_18$CORRSTAFF
data15_18$ratios3[is.infinite(data15_18$ratios3) | data15_18$ratios3 <=0] <- NA

grpRatio <- data15_18 %>% group_by(year, STRATUM) %>% summarise(avg = mean(ratios3, na.rm=T))
grpRatio <- grpRatio[is.na(grpRatio$avg)==F,]
grpRatio
```

```
## # A tibble: 40 x 3
## # Groups:   year [4]
##    year  STRATUM   avg
```

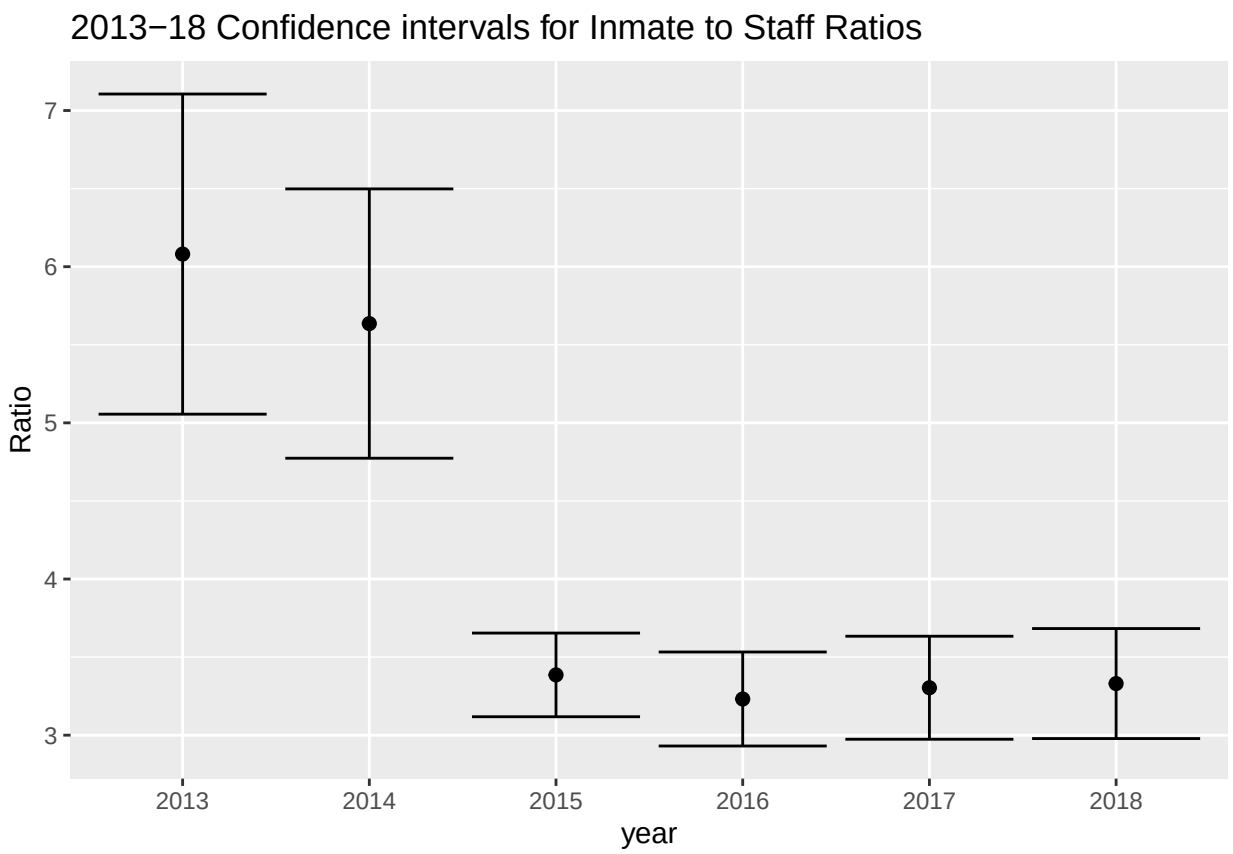
```
##    <chr> <fct>   <dbl>
##  1 2015  1        4.98
##  2 2015  2        4.74
##  3 2015  3        5.01
##  4 2015  4        5.68
##  5 2015  5        3.24
##  6 2015  7        5.61
##  7 2015  8        5.13
##  8 2015  9        3.80
##  9 2015  10       2.56
## 10 2015  12       5.00
## # ... with 30 more rows
```

##Inmate staff Ratio

```
##Inmate to staff
ratio_fn <- function(dat) {
  ##Returns Ratio and SE
strat_design<-svydesign(ids = ~1, strata = ~STRATUM, weights=~FINALWT, data = dat)
r <- svyratio(~ADP, ~TOTALSTAFF, strat_design, na.rm=T)
c(unname(r$ratio)[[1]], sqrt(unname(r$var)[[1]]))
}

yrs <- as.character(2013:2018)

estimates <- data.frame()

for (i in 1:6) {
  yr <- yrs[i]
  dat <- data13_18[data13_18$year==yr,]
  estimates <- rbind(estimates, ratio_fn(dat))
}

colnames(estimates) <- c("Ratio", "SE")
estimates$year <- yrs

estimates
```

```
##      Ratio        SE year
## 1 6.080715 0.5229804 2013
## 2 5.635782 0.4399577 2014
## 3 3.386003 0.1367137 2015
## 4 3.231654 0.1535714 2016
## 5 3.303769 0.1683169 2017
## 6 3.330533 0.1797114 2018
```

```
ggplot(estimates, aes(x = year, y = Ratio)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymax = Ratio+1.96*SE, ymin = Ratio-1.96*SE))+
  ggtitle("2013-18 Confidence intervals for Inmate to Staff Ratios")
```



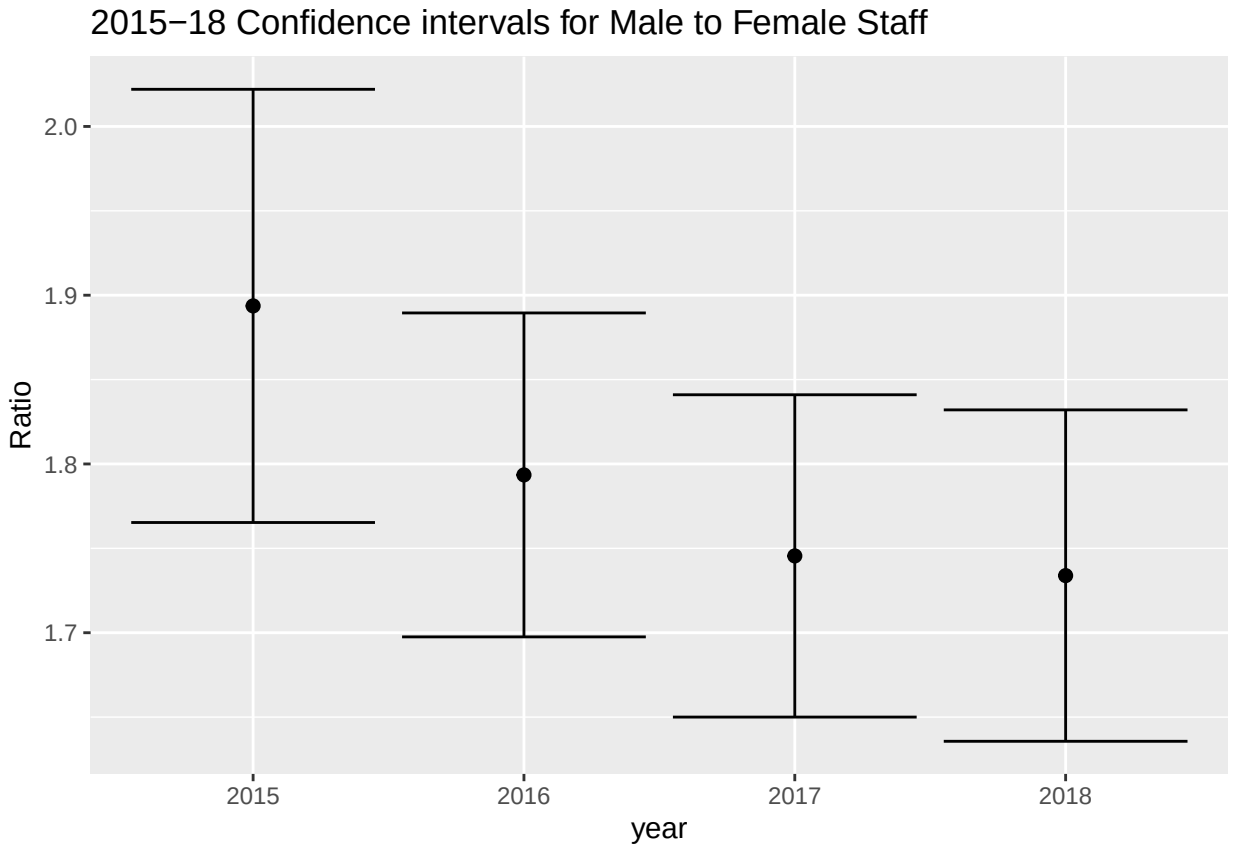

```
estimates <- estimates[c("year","Ratio","SE")]
estimates
```

```
##   year    Ratio        SE
## 1 2013 6.080715 0.5229804
## 2 2014 5.635782 0.4399577
## 3 2015 3.386003 0.1367137
## 4 2016 3.231654 0.1535714
## 5 2017 3.303769 0.1683169
## 6 2018 3.330533 0.1797114
```

##Functions for estimation

```
yrs <- as.character(2015:2018)
##Female to Male
estimates <- data.frame()

ratio_fn2 <- function(dat) {
  ##Returns Ratio and SE
strat_design<-svydesign(ids = ~1, strata = ~STRATUM, weights=~FINALWT, data = dat)
r <- svyratio(~TOTALSTAFFM, ~TOTALSTAFFF, strat_design, na.rm=T)
c(unname(r$ratio)[[1]], sqrt(unname(r$var)[[1]]))
}

for (i in 1:4) {
  yr <- yrs[i]
  dat <- data15_18[data15_18$year==yr,]
  estimates <- rbind(estimates, ratio_fn2(dat))
```

```
}

colnames(estimates) <- c("Ratio", "SE")
estimates$year <- yrs


ggplot(estimates, aes(x = year, y = Ratio)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymax = Ratio+1.96*SE, ymin = Ratio-1.96*SE))+
  ggtitle("2015-18 Confidence intervals for Male to Female Staff")
```




```
estimates <- estimates[c("year","Ratio","SE")]
data15_18$year <- as.factor(data15_18$year)
summary(aov(ratio~year, data15_18))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## year           3      9   2.846   1.145   0.33
## Residuals   3401   8456   2.486
## 114 observations deleted due to missingness
```

##t-test

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.0.0     v forcats 0.4.0
## v purrr   0.3.3
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x tidyr::expand()    masks Matrix::expand()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x tidyr::pack()      masks Matrix::pack()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
## x tidyr::unpack()    masks Matrix::unpack()
```

```
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:plyr':
##
##     mutate
```

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
```

```
## The following objects are masked from 'package:plyr':
##
##     desc, mutate
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
#Inmate to staff
dat <- data15_18[data15_18$STRATUM != 1 & data15_18$STRATUM != 2,]
summary(aov(ratio~STRATUM, dat))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## STRATUM        7     63   9.013   4.106 0.000173 ***
## Residuals   1933   4243   2.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 59 observations deleted due to missingness
```

```
inm_staff<-data13_18[data13_18$year != "2013" & data13_18$year != "2014",]


t_test <- inm_staff %>% pairwise_t_test(ratios2~STRATUM, pool.sd = F)
t_test <- t_test[t_test$p.adj < 0.05,]
t_test <- t_test[t_test$group1!=1 & t_test$group2 !=12,c("group1", "group2","p.adj")]
head(t_test)
```

```
## # A tibble: 6 x 3
```

```
##   group1 group2    p.adj
##   <chr>  <chr>     <dbl>
## 1 2      5      3.96e- 7
## 2 2      7      2.00e- 3
## 3 2      10     1.03e-11
## 4 3      5      1.55e- 5
## 5 3      7      2.00e- 3
## 6 3      10     6.26e- 9
```

Q1Q2

```
library(ggplot2)
library(dplyr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(survey)
```

#Loading data

```
fifteen_eighteen_data <- read.csv("data15_18.csv")
thirteen_fourteen_data <- read.csv("data13_14.csv")
total_data <- read.csv("data13_18.csv")
```

#Individual Years

```
thirteen_data <- total_data %>% filter(year==2013)
fourteen_data <- total_data %>% filter(year==2014)
fifteen_data <- total_data %>% filter(year==2015)
sixteen_data <- total_data %>% filter(year==2016)
seventeen_data <- total_data %>% filter(year==2017)
eighteen_data <- total_data %>% filter(year==2018)
```

#Total staff values using survey

```
survey_design13 <- svydesign(data=thirteen_data, ids=~1, weights=thirteen_data$FINALWT, strata=thirteen_
svytotal(thirteen_data$TOTALSTAFF,design=survey_design13)
```

```
##       total    SE
## [1,] 120274 13653
```

```
survey_design14 <- svydesign(data=fourteen_data, ids=~1, weights=fourteen_data$FINALWT, strata=fourteen_
svytotal(fourteen_data$TOTALSTAFF,design=survey_design14)
```

```
##       total    SE
## [1,] 131122 13688
```

```
survey_design15 <- svydesign(data=fifteen_data, ids=~1, weights=fifteen_data$FINALWT, strata=fifteen_da
svytotal(fifteen_data$TOTALSTAFF,design=survey_design15)
```

```
##       total    SE
## [1,] 212489 12634
```

```
survey_design16 <- svydesign(data=sixteen_data, ids=~1, weights=sixteen_data$FINALWT, strata=sixteen_da
svytotal(sixteen_data$TOTALSTAFF,design=survey_design16)
```

```
##       total    SE
## [1,] 226294 14067
```

```
survey_design17 <- svydesign(data=seventeen_data, ids=~1, weights=seventeen_data$FINALWT, strata=sevent
svytotal(seventeen_data$TOTALSTAFF,design=survey_design17)
```

```
##       total    SE
## [1,] 225690 14748
```

```
survey_design18 <- svydesign(data=eighteen_data, ids=~1, weights=eighteen_data$FINALWT, strata=eighteen
svytotal(eighteen_data$TOTALSTAFF,design=survey_design18)
```

```
##       total    SE
## [1,] 221564 14500
```

#Assault Analysis for years 13-14

```
thirteendata <- thirteen_fourteen_data %>% filter(year==2013)
fourteendata <- thirteen_fourteen_data %>% filter(year==2014)
thirteensurvey <- svydesign(data=thirteendata, ids=~1, weights=thirteendata$FINALWT, strata=thirteendata
svytotal(~ASSAULTCORR, design=thirteensurvey)
```

```
##              total     SE
## ASSAULTCORR 3000.9 443.15
```

```
svytotal(~ASSAULTOTRSTAFFF, design=thirteensurvey)
```

```
##                    total     SE
## ASSAULTOTRSTAFFF 159.18 36.256
```

```
fourteensurvey <- svydesign(data=fourteendata, ids=~1, weights=fourteendata$FINALWT, strata=fourteendata
svytotal(~ASSAULTCORR, design=fourteensurvey)
```

```
##              total     SE
## ASSAULTCORR 2687.1 411.78
```

```
svytotal(~ASSAULTOTRSTAFFF, design=fourteensurvey)
```

```
##                    total     SE
## ASSAULTOTRSTAFFF 93.392 26.478
```

#Chisq Test of Independence

```
sum(thirteendata$TOTALSTAFF-thirteendata$OTHERSTAFF)
```

```
## [1] 85808
```

```
sum(thirteendata$OTHERSTAFF)
```

```
## [1] 27551
```

```
sum(thirteendata$ASSAULTCORR)
```

```
## [1] 2854
```

```
sum(thirteendata$ASSAULTOTRSTAFFF)
```

```
## [1] 150
```

```r
correctional_staff <- c(2854, 85808)
other_staff <- c(150, 27551)
chisquared_table <- data.frame(rbind(correctional_staff, other_staff))
names(chisquared_table) <- c("assault", "no_assault")
chisquared_table
```

```
##                    assault no_assault
## correctional_staff    2854      85808
## other_staff            150      27551
```

```r
chisq.test(chisquared_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  chisquared_table
## X-squared = 600.58, df = 1, p-value < 2.2e-16
```

#Chisq Test of Independence

```r
emptyvector <- c()
emptyvector[thirteendata$ADP < 2000] <- "one"
emptyvector[thirteendata$ADP > 2001 & thirteendata$ADP < 4001] <- "two"
emptyvector[thirteendata$ADP > 4001] <- "three"
thirteendata$category <- emptyvector
gettingvals1 <- thirteendata %>% select(category, ASSAULTCORR, ASSAULTOTRSTAFFF) %>% group_by(category)
thirteendata$noassault <- thirteendata$TOTALSTAFF - thirteendata$ASSAULTCORR - thirteendata$ASSAULTOTRST
gettingvals2 <- thirteendata %>% select(noassault, category) %>% group_by(category) %>% summarise(sum=su

assault <- c(1470, 878, 598)
no_assault <- c(60027, 29406, 20415)

chisquared_table <- data.frame(rbind(assault, no_assault))
names(chisquared_table) <- c("one", "two", "three")
chisquared_table
```

```
##              one   two three
## assault     1470   878   598
## no_assault 60027 29406 20415
```

```r
chisq.test(chisquared_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisquared_table
## X-squared = 26.216, df = 2, p-value = 2.029e-06
```

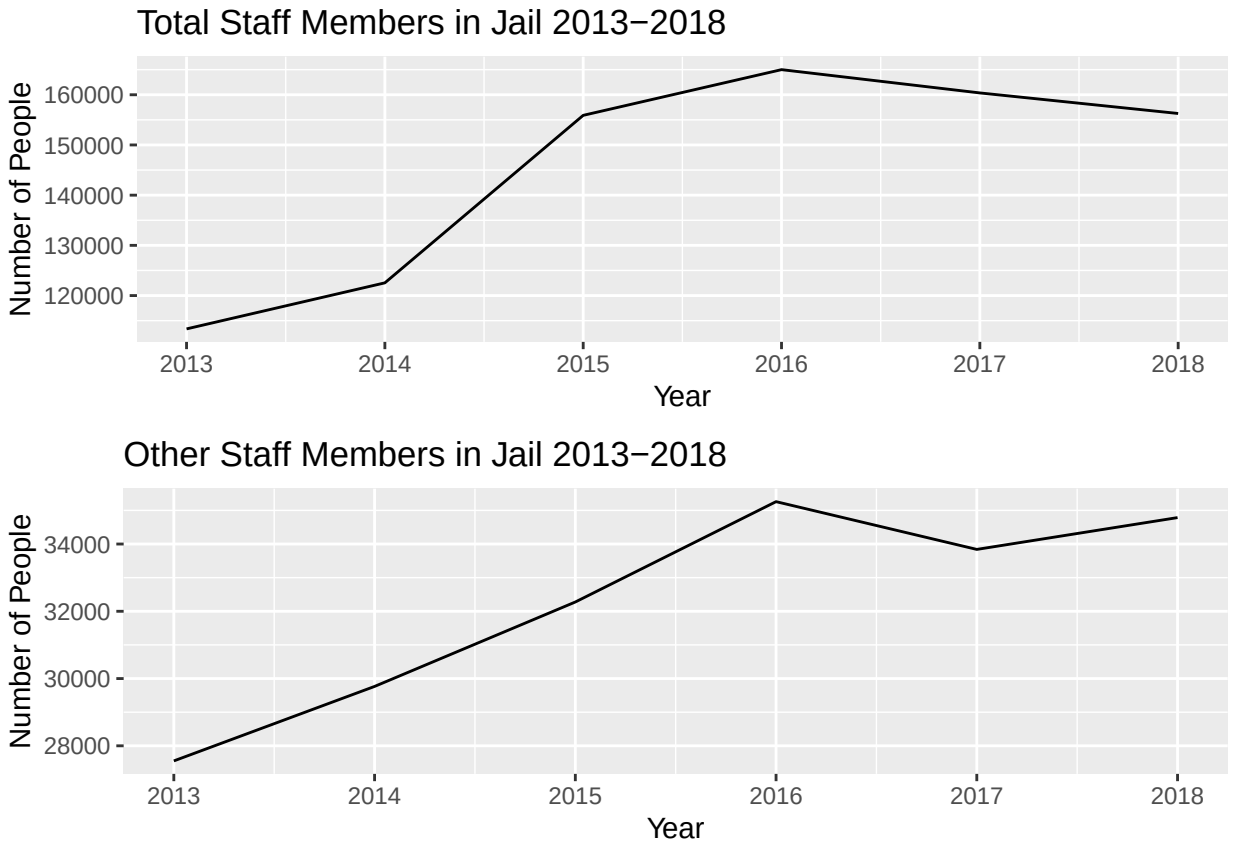
#Exploratory Analysis

```r
staff_data <- total_data %>% select(year, TOTALSTAFF, OTHERSTAFF) %>% group_by(year) %>% summarise_all(
staff_data
```

```
## # A tibble: 6 x 3
##    year TOTALSTAFF OTHERSTAFF
##   <int>      <int>      <int>
## 1  2013     113359      27551
## 2  2014     122533      29766
```

```
## 3  2015      155897       32276
## 4  2016      165000       35260
## 5  2017      160366       33840
## 6  2018      156268       34786
```

```
#reference https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2
a <- ggplot(data=staff_data)+geom_line(aes(x = year, y = TOTALSTAFF))+xlab("Year")+ylab("Number of Peopl

b <- ggplot(data=staff_data)+geom_line(aes(x = year, y = OTHERSTAFF))+xlab("Year")+ylab("Number of Peopl

grid.arrange(a, b)
```




```
assault_data <- thirteen_fourteen_data %>% select(ASSAULTCORR, ASSAULTOTRSTAFFF, DEATHSTAFFYN)
assault_data$DEATHSTAFFYN[assault_data$DEATHSTAFFYN==-9] <- 0
summary(assault_data)
```

```
##    ASSAULTCORR      ASSAULTOTRSTAFFF  DEATHSTAFFYN
##  Min.   :  0.000   Min.   :0.000    Min.   :0.0000
##  1st Qu.:  0.000   1st Qu.:0.000    1st Qu.:0.0000
##  Median :  0.000   Median :0.000    Median :0.0000
##  Mean   :  3.097   Mean   :0.137    Mean   :0.4819
##  3rd Qu.:  0.000   3rd Qu.:0.000    3rd Qu.:0.0000
##  Max.   :217.000   Max.   :9.000    Max.   :2.0000
```

```
assault_data <- thirteen_fourteen_data %>% select(ASSAULTCORR, ASSAULTOTRSTAFFF, DEATHSTAFFYN)
assault_data$DEATHSTAFFYN[assault_data$DEATHSTAFFYN==-9] <- 0
assault_data <- assault_data %>% select(ASSAULTCORR, ASSAULTOTRSTAFFF, DEATHSTAFFYN) %>% summarise_all(s
assault_count <- data.frame("Type"=c("Assault on Corr Officers", "Assault on Other", "Deaths by Assaults
assault_count
```

```
##                         Type Count
## 1 Assault on Corr Officers  2526
## 2         Assault on Other    88
## 3       Deaths by Assaults   836
```

```
ggplot(data=assault_count)+geom_col(aes(x=Type, y=Count))+xlab("Assault Type")+ylab("Number of Cases")+g
```