140SL MIDTERM

JIASHU MIAO 804786709

2019/2/6

1.

```
# rm(list=ls())
pkg <- c("readr","readxl","dplyr","stringr","ggplot2","tidyr")
pkgload <- lapply(pkg, require, character.only = TRUE)
```

```
## Loading required package: readr

## Loading required package: readxl

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

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

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

## Loading required package: stringr

## Loading required package: ggplot2

## Loading required package: tidyr
```

1.

- Provide a demographic profile for both test components (AnionGap and SODIUM).

```
data1 <- read_excel("/Users/MichaelMiao/UCLA/uclatextbook/140sl/140\ review\ from\ Wilbur/demographicDa
```

```
## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.
```

```
load("/Users/MichaelMiao/UCLA/uclatextbook/140sl/140\ review\ from\ Wilbur/LAB4PM.RData")
data2 <- LAB4PM
summary(data1)
```

```
##    STUDY_ID            Gender              Race
##  Length:18721       Length:18721       Length:18721
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##   Ethnicity
##  Length:18721
##  Class :character
##  Mode  :character
```

```r
summary(data2)
```

```
##    STUDY_ID            AnionGap          Age             Year     
##  Length:14161       Min.   : 4.00   Min.   :30.00   Min.   :2016  
##  Class :character   1st Qu.:12.00   1st Qu.:32.00   1st Qu.:2016  
##  Mode  :character   Median :14.00   Median :34.00   Median :2016  
##                     Mean   :13.77   Mean   :34.45   Mean   :2016  
##                     3rd Qu.:15.00   3rd Qu.:37.00   3rd Qu.:2017  
##                     Max.   :35.00   Max.   :40.00   Max.   :2017  
##  Inpatient_Outpatient     SODIUM     
##  Length:14161         Min.   :123.0  
##  Class :character     1st Qu.:139.0  
##  Mode  :character     Median :140.0  
##                       Mean   :140.1  
##                       3rd Qu.:142.0  
##                       Max.   :151.0  
```

```r
names(data1)
```

```
## [1] "STUDY_ID"  "Gender"    "Race"      "Ethnicity"
```

```r
names(data2)
```

```
## [1] "STUDY_ID"             "AnionGap"             "Age"                 
## [4] "Year"                 "Inpatient_Outpatient" "SODIUM"
```

```r
datajoin <- inner_join(data1,data2,by = "STUDY_ID")
#View(datajoin)
datajoind <- distinct(datajoin)
attach(datajoin)
summary(datajoin)
```

```
##    STUDY_ID            Gender              Race          
##  Length:14161       Length:14161       Length:14161      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##   Ethnicity            AnionGap          Age             Year     
##  Length:14161       Min.   : 4.00   Min.   :30.00   Min.   :2016  
##  Class :character   1st Qu.:12.00   1st Qu.:32.00   1st Qu.:2016  
##  Mode  :character   Median :14.00   Median :34.00   Median :2016  
##                     Mean   :13.77   Mean   :34.45   Mean   :2016  
##                     3rd Qu.:15.00   3rd Qu.:37.00   3rd Qu.:2017  
##                     Max.   :35.00   Max.   :40.00   Max.   :2017  
##  Inpatient_Outpatient     SODIUM     
##  Length:14161         Min.   :123.0  
##  Class :character     1st Qu.:139.0  
##  Mode  :character     Median :140.0  
##                       Mean   :140.1  
##                       3rd Qu.:142.0  
##                       Max.   :151.0  
```

```r
tapply(datajoin$AnionGap,datajoin$Race,mean)
```

```
##          American Indian or Alaska Native
##                                  14.23913
##                                     Asian
##                                  13.83436
##                 Black or African American
##                                  13.15815
##                            Multiple Races
##                                  13.61905
##                                        NA
##                                  13.13636
## Native Hawaiian or Other Pacific Islander
##                                  14.45946
##                                     Other
##                                  13.80646
##                           Patient Refused
##                                  14.21692
##                                   Unknown
##                                  13.92087
##                        White or Caucasian
##                                  13.65992
```

```r
tapply(datajoin$AnionGap,datajoin$Gender,mean)
```

```
##   Female     Male  Unknown
## 13.67799 13.87348 16.00000
```

```r
tapply(datajoin$AnionGap,datajoin$Ethnicity,mean)
```

```
##                                Cuban                   Hispanic or Latino
##                             16.00000                             13.59548
##        Hispanic/Spanish origin Other Mexican, Mexican American, Chicano/a
##                             14.01170                             13.00000
##               Not Hispanic or Latino                      Patient Refused
##                             13.69819                             14.19592
##                         Puerto Rican                              Unknown
##                             12.50000                             13.93228
```

```r
tapply(datajoin$SODIUM,datajoin$Race,mean)
```

```
##          American Indian or Alaska Native
##                                  139.3261
##                                     Asian
##                                  140.1168
##                 Black or African American
##                                  140.0032
##                            Multiple Races
##                                  139.7738
##                                        NA
##                                  139.9545
## Native Hawaiian or Other Pacific Islander
##                                  139.7297
##                                     Other
##                                  140.1326
##                           Patient Refused
##                                  140.2988
##                                   Unknown
##                                  140.2171
##                         White or Caucasian
##                                   140.1332
```

```r
tapply(datajoin$SODIUM,datajoin$Gender,mean)
```

```
##   Female     Male  Unknown
## 139.7425 140.6421 143.0000
```

```r
tapply(datajoin$SODIUM,datajoin$Ethnicity,mean)
```

```
##                                 Cuban                   Hispanic or Latino
##                              141.2500                             139.9774
##        Hispanic/Spanish origin Other Mexican, Mexican American, Chicano/a
##                              140.2047                             139.7419
##               Not Hispanic or Latino                      Patient Refused
##                              140.1228                             140.3020
##                         Puerto Rican                              Unknown
##                              140.7500                             140.2659
```

```r
prop.table(table(datajoin$Gender))
```

```
##
##       Female         Male      Unknown
## 5.528564e-01 4.470729e-01 7.061648e-05
```

```r
prop.table(table(datajoin$Race))
```

```
##
##           American Indian or Alaska Native
##                                0.003248358
##                                      Asian
##                                0.102746981
##                  Black or African American
##                                0.044205918
##                             Multiple Races
##                                0.005931784
##                                         NA
##                                0.003107125
## Native Hawaiian or Other Pacific Islander
##                                0.002612810
##                                      Other
##                                0.170397571
##                            Patient Refused
##                                0.078454911
##                                    Unknown
##                                0.149918791
##                         White or Caucasian
##                                0.439375750
```

```r
table(Gender)
```

```
## Gender
##  Female    Male Unknown
##    7829    6331       1
```

```r
table(Race)
```

```
## Race
##           American Indian or Alaska Native
##                                       46
##                                    Asian
##                                     1455
##                Black or African American
##                                      626
##                           Multiple Races
##                                       84
##                                       NA
##                                       44
## Native Hawaiian or Other Pacific Islander
##                                       37
##                                    Other
##                                     2413
##                          Patient Refused
##                                     1111
##                                  Unknown
##                                     2123
##                       White or Caucasian
##                                     6222
```

```
table(Ethnicity)
```

```
## Ethnicity
##                                 Cuban                    Hispanic or Latino
##                                     4                                  1372
##         Hispanic/Spanish origin Other Mexican, Mexican American, Chicano/a
##                                   171                                   124
##                Not Hispanic or Latino                       Patient Refused
##                                  9042                                  1225
##                          Puerto Rican                               Unknown
##                                     8                                  2215
```

```
datajoinnew <- datajoin %>% mutate(.,Noinfo = replace(Race, Race=="Patient Refused","Unknown"))
table(datajoinnew$Noinfo)
```

```
## 
##         American Indian or Alaska Native
##                                       46
##                                    Asian
##                                     1455
##                Black or African American
##                                      626
##                           Multiple Races
##                                       84
##                                       NA
##                                       44
## Native Hawaiian or Other Pacific Islander
##                                       37
##                                    Other
##                                     2413
##                                  Unknown
##                                     3234
##                       White or Caucasian
##                                     6222
```

```
datajoinnew <- datajoinnew %>% select_at(.,vars(-c(Race)))
datajoinnew <- rename(datajoinnew, Race = Noinfo)
dim(datajoinnew)
```

```
## [1] 14161     9
```

```
datajoinnew <- datajoinnew %>% mutate(., NO = replace(Ethnicity, Ethnicity == "Patient Refused","Unknow
datajoinnew <- datajoinnew %>% select_at(.,vars(-c(Ethnicity)))
dim(datajoinnew)
```

```
## [1] 14161     9
```

```
datajoinnew <- rename(datajoinnew, Ethnicity = NO)
```

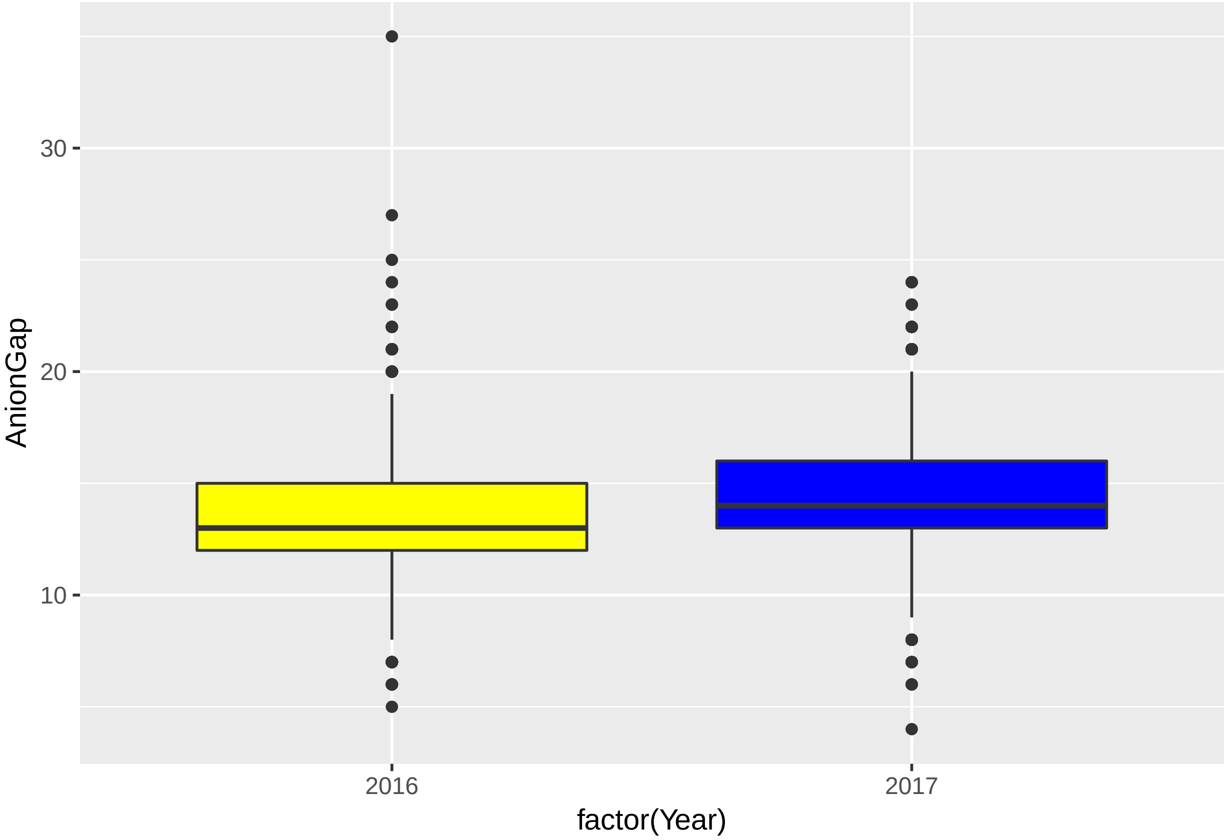
- There are about 55% females and 44% males enroll in this test. There are many unknows and I would like to make the patient refused as unknown for both race and ethicnicity. The two component seems to have no specific pattern but they distribute well.

2

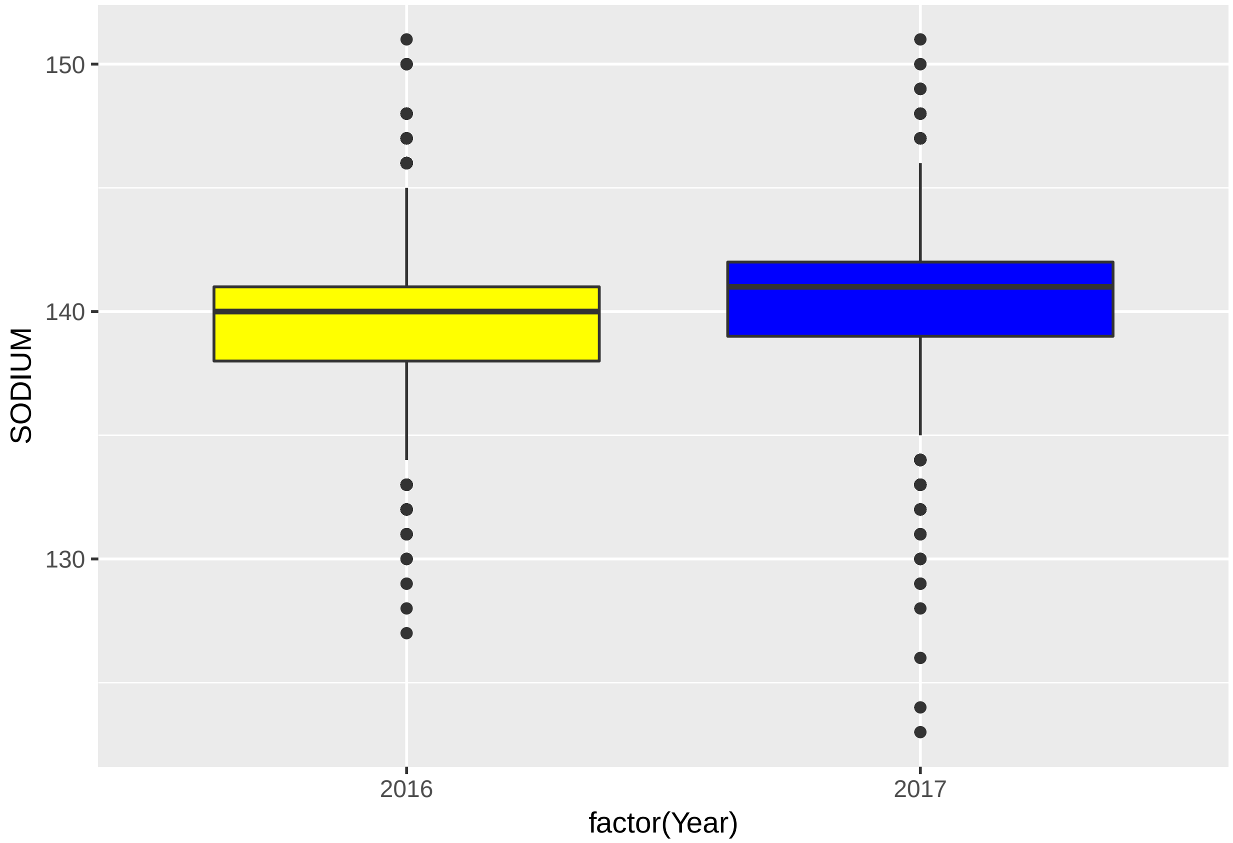
```
str(datajoinnew)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    14161 obs. of  9 variables:
##  $ STUDY_ID           : chr  "000EA425FFF3622052E55966D027AEC7" "0013E732CFD6284BBBDC47C8B3D6132A"
##  $ Gender             : chr  "Female" "Female" "Female" "Female" ...
##  $ AnionGap           : num  13 15 13 15 14 15 16 14 12 11 ...
##  $ Age                : int  33 35 38 39 33 33 32 31 35 32 ...
##  $ Year               : int  2016 2016 2016 2016 2016 2017 2016 2017 2017 2016 ...
##  $ Inpatient_Outpatient: chr  "IP" "OP" "OP" "OP" ...
##  $ SODIUM             : num  141 140 141 143 141 142 141 144 139 136 ...
##  $ Race               : chr  "Other" "Other" "White or Caucasian" "White or Caucasian" ...
##  $ Ethnicity          : chr  "Not Hispanic or Latino" "Unknown" "Not Hispanic or Latino" "Not Hispa
```

```
par(mfrow=c(1,2))
ggplot(data = datajoinnew,aes(factor(Year),AnionGap))+geom_boxplot(fill = c("yellow","blue"))
```



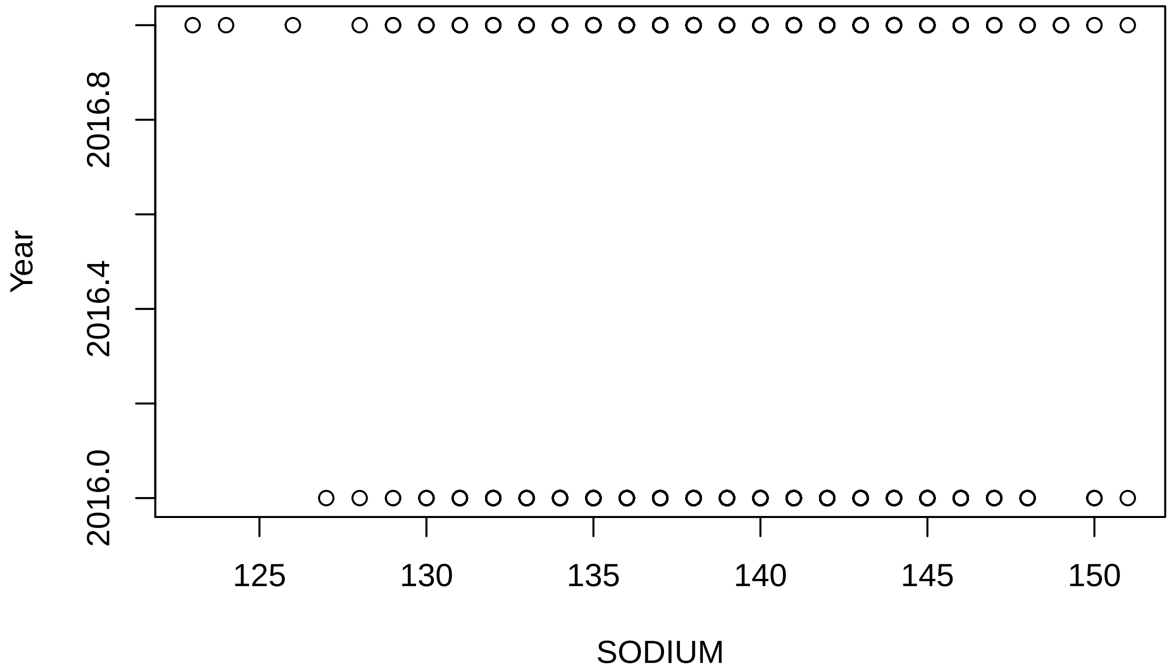
```r
ggplot(data = datajoinnew,aes(factor(Year),SODIUM))+geom_boxplot(fill = c("yellow","blue"))
```



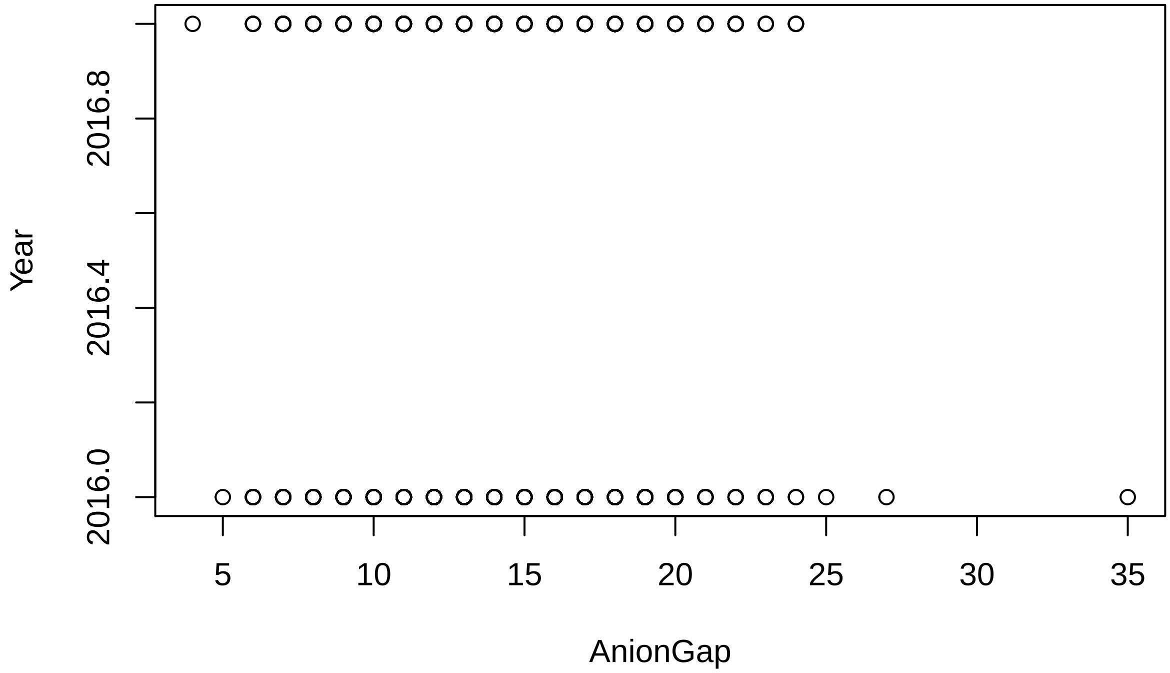
```
attach(datajoinnew)
```

```
## The following objects are masked from datajoin:
##
##     Age, AnionGap, Ethnicity, Gender, Inpatient_Outpatient, Race,
##     SODIUM, STUDY_ID, Year
```

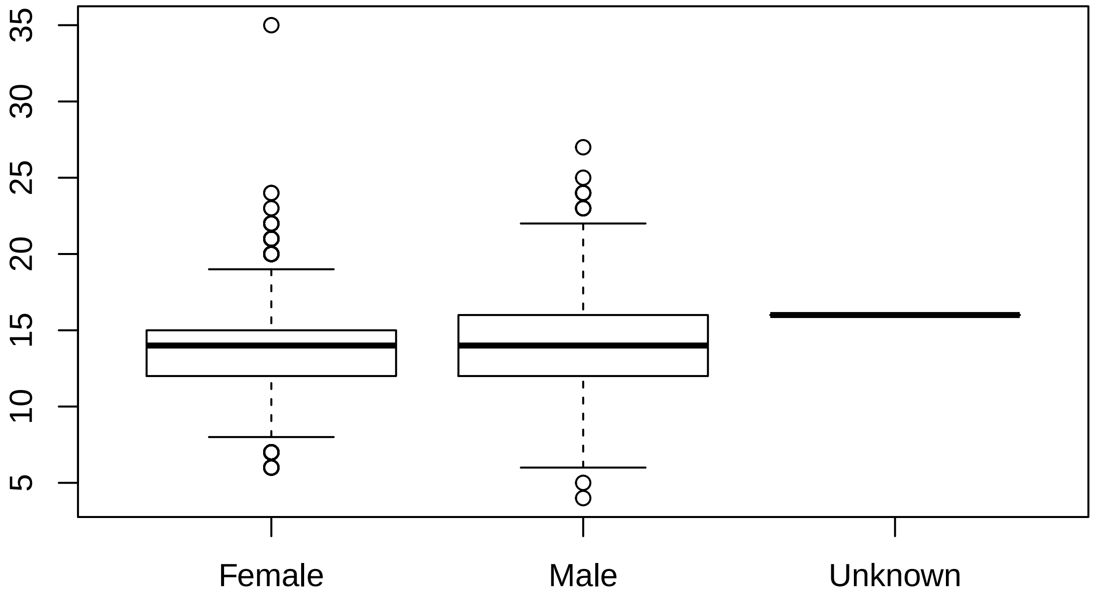
```
plot(SODIUM,Year)
```



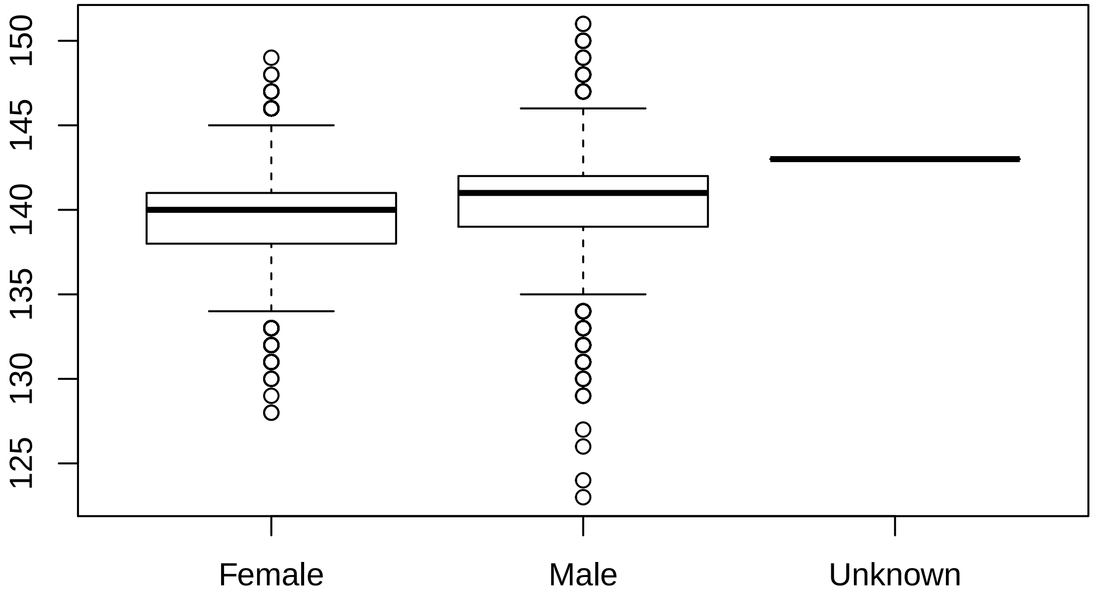
```r
plot(AnionGap,Year)
```



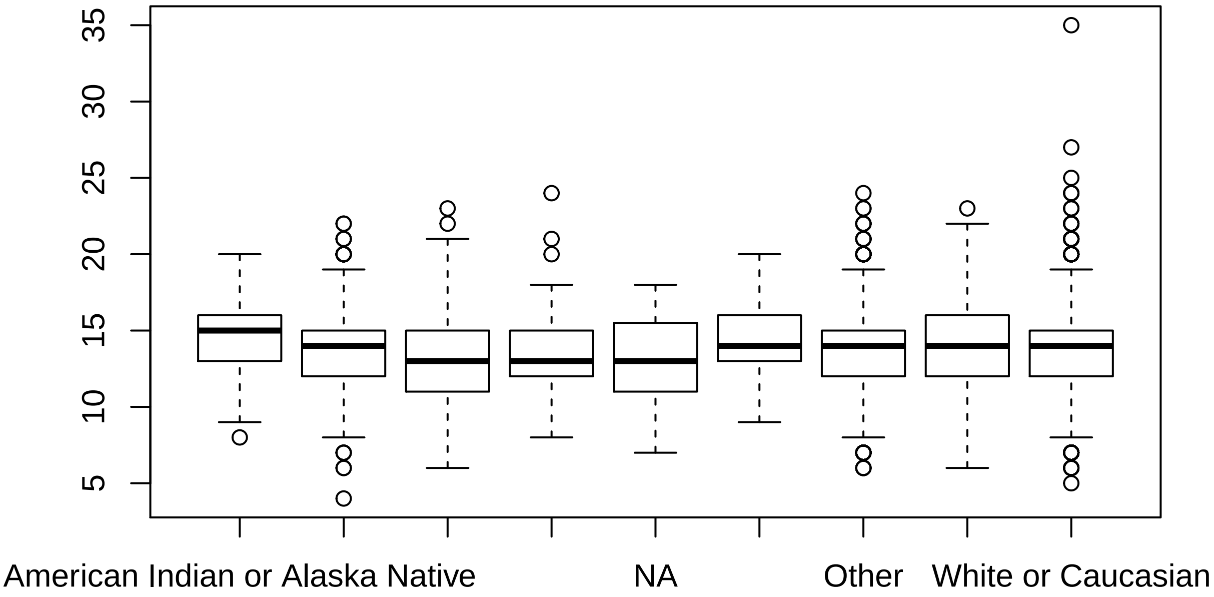
```r
plot(as.factor(Gender),AnionGap)
```



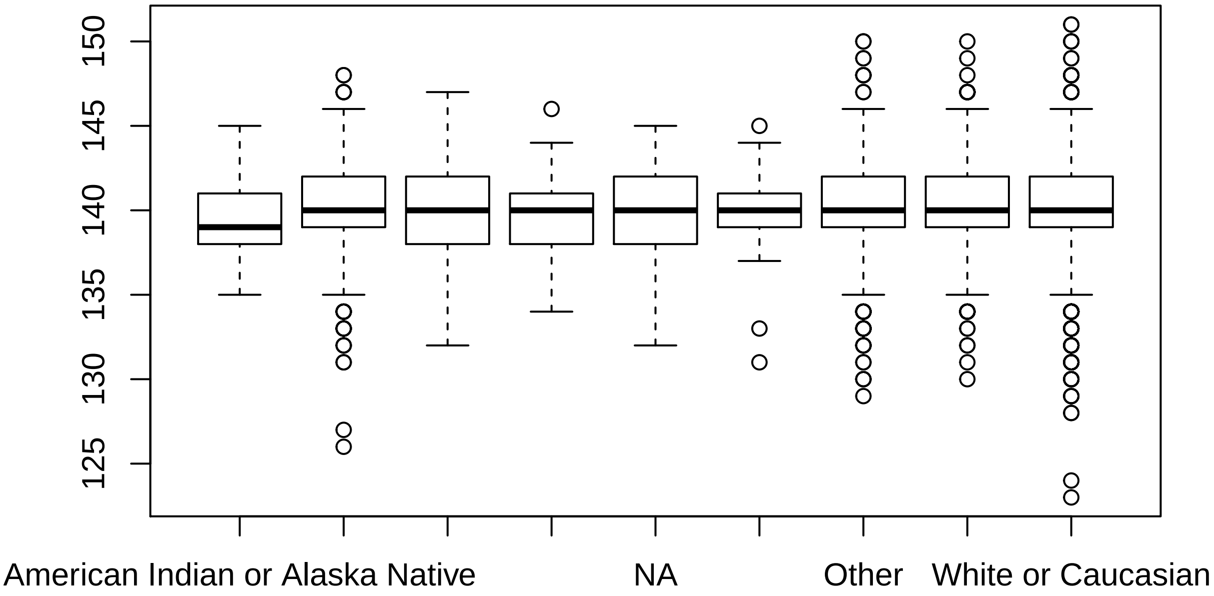
```
plot(as.factor(Gender),SODIUM)
```



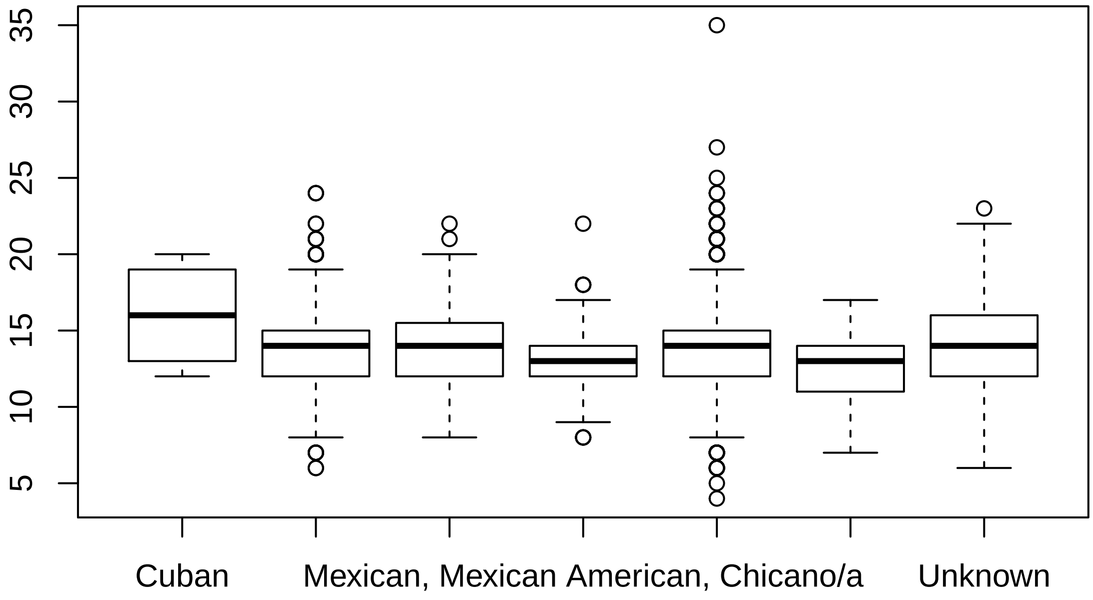
```
plot(as.factor(Race),AnionGap)
```



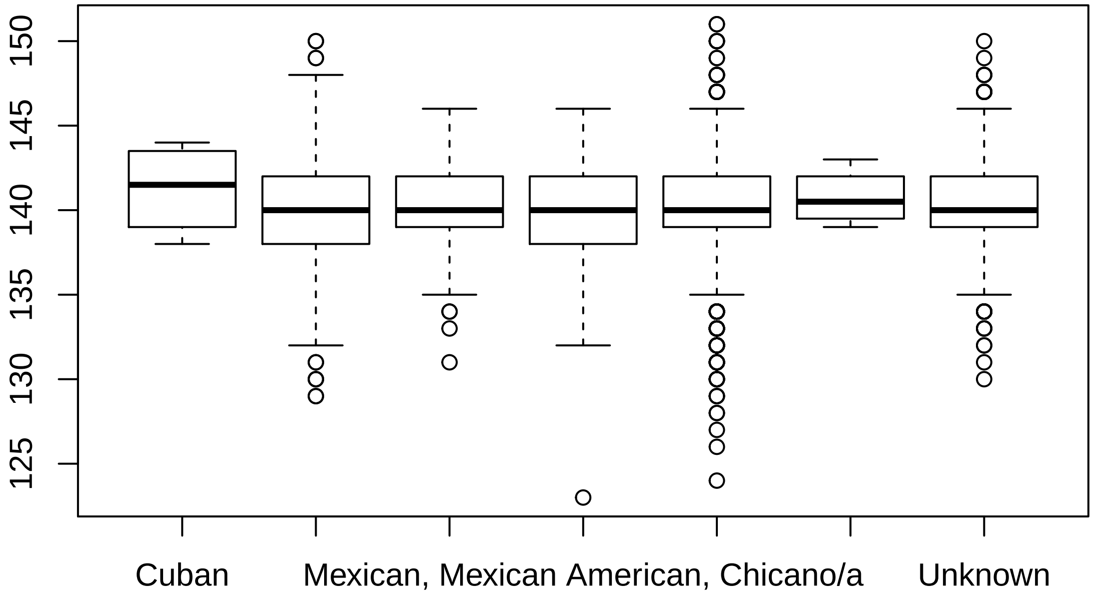
```r
plot(as.factor(Race),SODIUM)
```



```r
plot(as.factor(Ethnicity),AnionGap)
```



```
plot(as.factor(Ethnicity),SODIUM)
```



3

```
attach(datajoinnew)
```

```
## The following objects are masked from datajoinnew (pos = 3):
##
##     Age, AnionGap, Ethnicity, Gender, Inpatient_Outpatient, Race,
##     SODIUM, STUDY_ID, Year
```

```
## The following objects are masked from datajoin:
##
##     Age, AnionGap, Ethnicity, Gender, Inpatient_Outpatient, Race,
##     SODIUM, STUDY_ID, Year
```

```
model1 <- glm(data = datajoinnew, SODIUM~ factor(Ethnicity)+ factor(Race) + factor(Gender))
summary(model1)
```

```
## 
## Call:
## glm(formula = SODIUM ~ factor(Ethnicity) + factor(Race) + factor(Gender), 
##     data = datajoinnew)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -17.2885   -1.6203    0.2097    1.3797   10.3797  
## 
## Coefficients:
##                                                         Estimate Std. Error
## (Intercept)                                            140.12018    1.23701
## factor(Ethnicity)Hispanic or Latino                     -1.19461    1.18805
## factor(Ethnicity)Hispanic/Spanish origin Other          -0.99804    1.20011
## factor(Ethnicity)Mexican, Mexican American, Chicano/a  -1.39230    1.20539
## factor(Ethnicity)Not Hispanic or Latino                 -1.06050    1.18649
## factor(Ethnicity)Puerto Rican                           -0.49213    1.45296
## factor(Ethnicity)Unknown                                -0.88782    1.18953
## factor(Race)Asian                                        0.73789    0.35554
## factor(Race)Black or African American                    0.54781    0.36271
## factor(Race)Multiple Races                               0.41494    0.43520
## factor(Race)NA                                           0.25916    0.50791
## factor(Race)Native Hawaiian or Other Pacific Islander    0.27560    0.52417
## factor(Race)Other                                        0.73058    0.35342
## factor(Race)Unknown                                      0.59755    0.36162
## factor(Race)White or Caucasian                           0.66737    0.35121
## factor(Gender)Male                                       0.89330    0.04027
## factor(Gender)Unknown                                    3.34386    2.37335
##                                                        t value Pr(>|t|)    
## (Intercept)                                            113.273   <2e-16 ***
## factor(Ethnicity)Hispanic or Latino                     -1.006   0.3147    
## factor(Ethnicity)Hispanic/Spanish origin Other          -0.832   0.4056    
## factor(Ethnicity)Mexican, Mexican American, Chicano/a  -1.155   0.2481    
## factor(Ethnicity)Not Hispanic or Latino                 -0.894   0.3714    
## factor(Ethnicity)Puerto Rican                           -0.339   0.7348    
## factor(Ethnicity)Unknown                                -0.746   0.4555    
## factor(Race)Asian                                        2.075   0.0380 *  
## factor(Race)Black or African American                    1.510   0.1310    
## factor(Race)Multiple Races                               0.953   0.3404    
## factor(Race)NA                                           0.510   0.6099    
## factor(Race)Native Hawaiian or Other Pacific Islander    0.526   0.5990    
## factor(Race)Other                                        2.067   0.0387 *  
## factor(Race)Unknown                                      1.652   0.0985 .  
## factor(Race)White or Caucasian                           1.900   0.0574 .  
## factor(Gender)Male                                      22.185   <2e-16 ***
## factor(Gender)Unknown                                    1.409   0.1589    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 5.627735)
```

```
##
##     Null deviance: 82581  on 14160  degrees of freedom
## Residual deviance: 79599  on 14144  degrees of freedom
## AIC: 64672
##
## Number of Fisher Scoring iterations: 2
```

```
model2 <- glm(data = datajoinnew, AnionGap~ factor(Ethnicity)+ factor(Race) + factor(Gender))
summary(model2)
```

```
##
## Call:
## glm(formula = AnionGap ~ factor(Ethnicity) + factor(Race) + factor(Gender),
##     data = datajoinnew)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.9460  -1.7726   0.0458   1.4140  21.4140
##
## Coefficients:
##                                                        Estimate Std. Error
## (Intercept)                                            16.48070    1.32351
## factor(Ethnicity)Hispanic or Latino                    -2.46108    1.27112
## factor(Ethnicity)Hispanic/Spanish origin Other         -2.05224    1.28403
## factor(Ethnicity)Mexican, Mexican American, Chicano/a -3.06385    1.28968
## factor(Ethnicity)Not Hispanic or Latino                -2.26967    1.26945
## factor(Ethnicity)Puerto Rican                          -3.50846    1.55455
## factor(Ethnicity)Unknown                               -2.07179    1.27271
## factor(Race)Asian                                      -0.45157    0.38040
## factor(Race)Black or African American                  -1.13579    0.38807
## factor(Race)Multiple Races                             -0.60796    0.46563
## factor(Race)NA                                         -1.36555    0.54342
## factor(Race)Native Hawaiian or Other Pacific Islander  0.17267    0.56082
## factor(Race)Other                                      -0.42089    0.37814
## factor(Race)Unknown                                    -0.45472    0.38690
## factor(Race)White or Caucasian                         -0.62499    0.37577
## factor(Gender)Male                                      0.18653    0.04308
## factor(Gender)Unknown                                   2.40128    2.53931
##                                                        t value Pr(>|t|)
## (Intercept)                                             12.452  < 2e-16 ***
## factor(Ethnicity)Hispanic or Latino                     -1.936  0.05287 .
## factor(Ethnicity)Hispanic/Spanish origin Other          -1.598  0.11000
## factor(Ethnicity)Mexican, Mexican American, Chicano/a  -2.376  0.01753 *
## factor(Ethnicity)Not Hispanic or Latino                 -1.788  0.07381 .
## factor(Ethnicity)Puerto Rican                           -2.257  0.02403 *
## factor(Ethnicity)Unknown                                -1.628  0.10358
## factor(Race)Asian                                       -1.187  0.23521
## factor(Race)Black or African American                   -2.927  0.00343 **
## factor(Race)Multiple Races                              -1.306  0.19168
## factor(Race)NA                                          -2.513  0.01199 *
## factor(Race)Native Hawaiian or Other Pacific Islander   0.308  0.75818
## factor(Race)Other                                       -1.113  0.26570
## factor(Race)Unknown                                     -1.175  0.23991
## factor(Race)White or Caucasian                          -1.663  0.09629 .
## factor(Gender)Male                                       4.330  1.5e-05 ***
## factor(Gender)Unknown                                   0.946  0.34435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6.442289)
##
##     Null deviance: 92020  on 14160  degrees of freedom
## Residual deviance: 91120  on 14144  degrees of freedom
## AIC: 66586
##
## Number of Fisher Scoring iterations: 2
```

```
model3 <- glm(data = datajoinnew,SODIUM~AnionGap)
summary(model3)
```

```
##
## Call:
## glm(formula = SODIUM ~ AnionGap, data = datajoinnew)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -16.9951   -1.4079    0.0049    1.5921   10.3964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.45071    0.10905 1260.44   <2e-16 ***
## AnionGap      0.19572    0.00779   25.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.583425)
##
##     Null deviance: 82581  on 14160  degrees of freedom
## Residual deviance: 79056  on 14159  degrees of freedom
## AIC: 64545
##
## Number of Fisher Scoring iterations: 2
```

I think there is very tiny association of Sodium between gender(male) and race(asian) ; and some tiny association of Anigap between gender(male) and race(black). The association is very small overall.There is association between the Sodium and Aniongap. Also, the value goes up for the Sodium and Aniongap with year goes up.