```
> setwd("~/Desktop")
>
> datjss = read.csv("datjss.csv")
> datsss = read.csv("datsss.csv")
> datstu = read.csv("datstu.csv")


# Ex1
> dim(datstu)
[1] 340823     18
>
> dim(datsss)
[1] 6165     6
>
> # exercise 1
>
> # number of students
> length(unique(datstu$X))
[1] 340823
>
> # number of schools
> length(unique(datsss$schoolname))
[1] 842
>
> # number of programs
> pgm = datstu[!duplicated(datstu$choicepgm1), ]
> dim(pgm)
[1] 32 18
>
> # number of choices
> datstu$choice1 <- paste(datstu$schoolcode1, "-", datstu$choicepgm1)
> datstu$choice2 <- paste(datstu$schoolcode2, "-", datstu$choicepgm2)
> datstu$choice3 <- paste(datstu$schoolcode3, "-", datstu$choicepgm3)
> datstu$choice4 <- paste(datstu$schoolcode4, "-", datstu$choicepgm4)
> datstu$choice5 <- paste(datstu$schoolcode5, "-", datstu$choicepgm5)
> datstu$choice6 <- paste(datstu$schoolcode6, "-", datstu$choicepgm6)
> choice = cbind(datstu$choice1, datstu$choice2, datstu$choice3, datstu$choice4, datstu$choice5, datstu$choice6)
> choice %>%
+   pivot_longer(
+     cols = starts_with("choice"),
+     names_to = "list",
+     names_prefix = "choice",
+     values_to = "choice",
+         values_drop_na = TRUE
+     )
> choice = choice[choice!=""]
> length(unique(choice))
[1] 3086
>
> # missing test score
> sum(is.na(datstu$score))
[1] 179887

# Ex3
> library("dplyr")
> library("tidyr")
> schchoice <- datstu %>%
+     select(-c(5:16)) %>%
+     pivot_longer(
+         cols = starts_with("choice"),
+         names_to = "list",
+         names_prefix = "choice",
+         values_to = "choice",
+         values_drop_na = TRUE
+     ) %>%
+     filter(rankplace == list) %>%
+     separate(choice, c("schoolcode", "program"), sep = " - ")
>
> sss <- datsss %>%
+     select(c(3:6))
> unique_school = sss[!duplicated(sss$schoolcode), ]
>
> schcho_sss <- merge(x = schchoice, y = sss, by = "schoolcode", all.x = TRUE)
> schcho_sss1 = group_by(schcho_sss, schoolcode)
> dfschcho = data_frame(summarise(schcho_sss1, cutoff = min(score), quality = mean(score),
size=n()))
>
> school_level = merge(schcho_sss1, dfschcho, by = 'schoolcode', all.x = T, all.y = T)
> school_level1 = subset(school_level, select = -c(X, agey, male, jssdistrict, rankplace) )
> school_level2 = school_level1[!duplicated(school_level1[c("schoolcode", "program")]), ]
>
> school_levelFin = subset(school_level2, select = -c(score) )
> school_levelFin[1:20, ]
     schoolcode list          program   sssdistrict    ssslong     ssslat cutoff
1        100101    4        Technical Wa Municipal -2.285030 10.03062    198
4        100101    3    General Arts Wa Municipal -2.285030 10.03062    198
37       100101    3  Home Economics Wa Municipal -2.285030 10.03062    198
```

```
505          100102       1    Home Economics Wa Municipal -2.285030 10.03062       250
511          100102       3       General Arts Wa Municipal -2.285030 10.03062       250
517          100102       2             Business Wa Municipal -2.285030 10.03062       250
529          100102       2 General Science Wa Municipal -2.285030 10.03062       250
553          100102       1         Agriculture Wa Municipal -2.285030 10.03062       250
559          100102       2         Visual Arts Wa Municipal -2.285030 10.03062       250
3205         100104       1       General Arts          Wa West             NA            NA       282
3223         100104       1    Home Economics           Wa West             NA            NA       282
3253         100104       1 General Science          Wa West             NA            NA       282
4015         100105       3    Home Economics Wa Municipal -2.285030 10.03062       242
4018         100105       2       General Arts Wa Municipal -2.285030 10.03062       242
4021         100105       1             Business Wa Municipal -2.285030 10.03062       242
4735         100106       2             Business Wa Municipal -2.285030 10.03062       223
4738         100106       3       General Arts Wa Municipal -2.285030 10.03062       223
4741         100106       2         Agriculture Wa Municipal -2.285030 10.03062       223
5095         100201       1       General Arts               Lawra -2.800941 10.54640       288
5098         100201       1 General Science              Lawra -2.800941 10.54640       288
           quality size
1       238.1250    504
4       238.1250    504
37      238.1250    504
505    296.4956 2700
511    296.4956 2700
517    296.4956 2700
529    296.4956 2700
553    296.4956 2700
559    296.4956 2700
3205 326.9333    810
3223 326.9333    810
3253 326.9333    810
4015 266.9708    720
4018 266.9708    720
4021 266.9708    720
4735 254.3667    360
4738 254.3667    360
4741 254.3667    360
5095 335.9600    600
5098 335.9600    600
```

```
> # Ex3
> school_mapping <- merge(school_levelFin,datjss,by=c("jssdistrict"))
>
> # replace with zeros
> school_mapping[is.na(school_mapping)] <- 0#4
>
> # element wise thing
>        school_mapping$distance        =        sqrt(69.172*(school_mapping$ssslong-
school_mapping$point_x)*cos(school_mapping$point_y/57.3)^2+(69.172*(school_mapping$
ssslat-school_mapping$point_y))^2)
Warning message:
In sqrt(69.172 * (school_mapping$ssslong - school_mapping$point_x) *  :
  产生了 NaNs
> school_mapping$distance
   [1]   6.128696 354.968125 354.968125 354.968125 354.968125 354.968125
   [7]   6.128696 354.968125 387.892180   0.000000 387.892180 387.892180
  [13] 387.892180 387.892180   0.000000 387.892180   0.000000 387.892180
  [19] 387.892180  77.117486 387.892180   0.000000  39.284770   0.000000
  [25]  25.568202   0.000000   6.555235  98.676089  73.688660   0.000000
  [31]  35.036790 387.892180 387.892180 387.892180 387.892180  57.226039
  [37] 387.892180 387.892180  39.284770 387.892180 387.892180  30.150849
  [43] 387.892180 387.892180 387.892180 387.892180        NaN  77.117486
  [49] 387.892180 387.892180 387.892180   0.000000   7.654021  34.848325
  [55] 387.892180 387.892180  65.132379 387.892180 387.892180 387.892180
  [61] 387.892180  77.117486   0.000000   1.405689  30.150849  77.117486
  [67] 387.892180   0.000000  89.372460  69.775455   0.000000 387.892180
  [73] 387.892180  89.372460 387.892180  27.069135   1.405689  57.226039
  [79]  30.150849 387.892180 387.892180        NaN  34.848325 387.892180
  [85]   0.000000   0.000000 387.892180   6.555235  34.848325 387.892180
  [91]  30.150849 387.892180   0.000000 387.892180 387.892180  22.422627
  [97]  12.679963 387.892180 387.892180   0.000000   0.000000   0.000000
 [103] 387.892180  39.284770   1.405689 387.892180 387.892180   0.000000
 [109]  69.775455  22.422627  25.568202 387.892180   0.000000  22.907403
 [115] 387.892180 387.892180 387.892180 387.892180 387.892180 387.892180
```

#Ex4

```
> jss = datjss %>%
+     select(c(2:4))
> schcho_sss2 <- merge(x = school_level, y = jss, by = "jssdistrict", all.x = TRUE)
> schcho_sss3 = group_by(schcho_sss2, schoolcode)
> cutoff_mean <- mean(schcho_sss3$cutoff)
> cutoff_mean
[1] 247.2349
> quality_mean <- mean(schcho_sss3$quality)
> quality_mean
[1] 295.0693
> distance <- mean(school_mapping$distance, na.rm=TRUE)
> distance
[1] 217.8308
> cutoff_stdev <- sd(schcho_sss3$cutoff)
> cutoff_stdev
[1] 48.39953
> quality_stdev <- sd(schcho_sss3$quality)
> quality_stdev
[1] 43.16776
> distance_stdev <- sd(school_mapping$distance, na.rm=TRUE)
> distance_stdev
[1] 232.9432
```