ECON 613 Aassignment 1

Nond Prueksiri

Importing Data

```
rm(list = ls())

datstu <- read.csv("C:/Users/nprue/Desktop/econ613/datstu.csv")
datsss <- read.csv("C:/Users/nprue/Desktop/econ613/datsss.csv")
datjss <- read.csv("C:/Users/nprue/Desktop/econ613/datjss.csv")
```

Excercise 1 Missing Data

- Number of Students

```
# Assuming each obervation represents each students
  length(unique(datstu$X))
```

```
## [1] 340823
```

- Number of Schools

```
# Eliminate missing data
  datsssA <- datsss[!(is.na(datsss$schoolname)
                       |is.na(datsss$schoolcode)
                       |is.na(datsss$sssdistrict)
                       |is.na(datsss$ssslong)
                       |is.na(datsss$ssslat)),]
# Eliminate duplicated schoolcode
  datsssB <- datsssA[!(duplicated(datsssA$schoolcode)),]
# Report the number of schools
  length(datsssB$schoolcode)
```

```
## [1] 689
```

- Number of programs

```
# Clean out missing programs
  datstuA <- datstu[!(is.na(datstu[11])
                      |is.na(datstu[12])
                      |is.na(datstu[13])
                      |is.na(datstu[14])
                      |is.na(datstu[15])
                      |is.na(datstu[16])),]
# Report the number of programs
  length(unique(unlist(datstuA[,11:16])))
```

```
## [1] 33
```

- Number of Choices (School, Program)

```
# Clean out missing school choices
  datstuB <- datstuA[!(is.na(datstuA[5])
                      |is.na(datstuA[6])
                      |is.na(datstuA[7])
                      |is.na(datstuA[8])
                      |is.na(datstuA[9])
                      |is.na(datstuA[10])),]

# Combine school and program choices, keep only unique ones
  datstuB$choice1 <- paste(datstuB[,5],"-" , datstuB[,11])
  datstuB$choice2 <- paste(datstuB[,6],"-" , datstuB[,12])
  datstuB$choice3 <- paste(datstuB[,7],"-" , datstuB[,13])
  datstuB$choice4 <- paste(datstuB[,8],"-" , datstuB[,14])
  datstuB$choice5 <- paste(datstuB[,9],"-" , datstuB[,15])
  datstuB$choice6 <- paste(datstuB[,10],"-" , datstuB[,16])

# Report the number of program choices (school, program)
  length(unique(unlist(datstuB[,19:24])))
```

```
## [1] 3068
```

- Missing Test Score

```
# Create missing score vector
missing <- datstu[is.na(datstu$score), ]
length(missing$X)
```

```
## [1] 179887
```

- Number of students applying to the same school

Who has the repeated schools with in her choices. . . .

```
# Create a logical column (TRUE = at least has one repeat)
datstu$sameschool <-  (datstu$schoolcode1 == datstu$schoolcode2
                    | datstu$schoolcode1 == datstu$schoolcode3
                    | datstu$schoolcode1 == datstu$schoolcode4
                    | datstu$schoolcode1 == datstu$schoolcode5
                    | datstu$schoolcode2 == datstu$schoolcode3
                    | datstu$schoolcode2 == datstu$schoolcode4
                    | datstu$schoolcode2 == datstu$schoolcode5
                    | datstu$schoolcode3 == datstu$schoolcode4
                    | datstu$schoolcode3 == datstu$schoolcode5
                    | datstu$schoolcode4 == datstu$schoolcode5)

# Return the number
nrow(datstu[datstu$sameschool == 'TRUE',])
```

```
## [1] 113586
```

Who has applied to only one school. . . .

```
# Create a logical column (TRUE = only one school)
datstu$oneschool <- (is.na(datstu$schoolcode2)
                    | datstu$schoolcode1 == datstu$schoolcode2) & (is.na(datstu$schoolcode3)
                    | datstu$schoolcode1 == datstu$schoolcode3) & (is.na(datstu$schoolcode4)
                    | datstu$schoolcode1 == datstu$schoolcode4) & (is.na(datstu$schoolcode5)
                    | datstu$schoolcode1 == datstu$schoolcode5) & (is.na(datstu$schoolcode6)
                    | datstu$schoolcode1 == datstu$schoolcode6)

# Return the number
nrow(datstu[datstu$oneschool == 'TRUE',])
```

```
## [1] 764
```

- Number of students applying less than 6 choices

```
# Create a logical column (TRUE = below 6 choices)
datstu$belowSix <- (is.na(datstu$schoolcode1) | is.na(datstu$schoolcode2) | is.na(datstu$schoolcode3)
                   |is.na(datstu$schoolcode4) | is.na(datstu$schoolcode5) | is.na(datstu$schoolcode6))

# Return the number
nrow(datstu[datstu$belowSix == 'TRUE',])
```

```
## [1] 17734
```

Excercise 2 Data

```
  library(data.table)
# Remove unused data
rm(datstuA, datstuB, datsssA, datsssB, missing)
datadmit <- datstu[ , 1:18]
# Remove invalid rankplace, i.e., NA and 99
datadmit <- datadmit[!(is.na(datadmit$rankplace) | datadmit$rankplace == 99), ]

# Create variable "schoolcode" = school that a student is placed into
datadmit$schoolcode <- NA
datadmit$schoolcode[which(datadmit$rankplace == 1)] <- datadmit$schoolcode1[which(datadmit$rankplace =
datadmit$schoolcode[which(datadmit$rankplace == 2)] <- datadmit$schoolcode2[which(datadmit$rankplace =
datadmit$schoolcode[which(datadmit$rankplace == 3)] <- datadmit$schoolcode3[which(datadmit$rankplace =
datadmit$schoolcode[which(datadmit$rankplace == 4)] <- datadmit$schoolcode4[which(datadmit$rankplace =
datadmit$schoolcode[which(datadmit$rankplace == 5)] <- datadmit$schoolcode5[which(datadmit$rankplace =
datadmit$schoolcode[which(datadmit$rankplace == 6)] <- datadmit$schoolcode6[which(datadmit$rankplace =

# Create variable "adprog" = program that a student is placed into
  datadmit$adprog <- NA
  datadmit$adprog[which(datadmit$rankplace == 1)] <- as.character(datadmit$choicepgm1[which(datadmit$
  datadmit$adprog[which(datadmit$rankplace == 2)] <- as.character(datadmit$choicepgm2[which(datadmit$
  datadmit$adprog[which(datadmit$rankplace == 3)] <- as.character(datadmit$choicepgm3[which(datadmit$
  datadmit$adprog[which(datadmit$rankplace == 4)] <- as.character(datadmit$choicepgm4[which(datadmit$
  datadmit$adprog[which(datadmit$rankplace == 5)] <- as.character(datadmit$choicepgm5[which(datadmit$
  datadmit$adprog[which(datadmit$rankplace == 6)] <- as.character(datadmit$choicepgm6[which(datadmit$


# Eliminate missing data
  datsss <- datsss[!(is.na(datsss$schoolname)
                    |is.na(datsss$schoolcode)
                    |is.na(datsss$sssdistrict)
                    |is.na(datsss$ssslong)
                    |is.na(datsss$ssslat)),]

# Eliminate duplicated schoolcode
  datsss <- datsss[!(duplicated(datsss$schoolcode)),]

# Merge with datsss for sssdistrict, ssslong, ssslat by schoolcode
 ssdat <- merge(datadmit,datsss, by = "schoolcode", all.x = TRUE)

# Create variable size, quality, size
size <- rep(1,nrow(ssdat))
ssdat <- data.table(ssdat)
ssdat <- ssdat[, list(quality=mean(score), cutoff=min(score), size = sum(size)), by=c("schoolcode","ad

# Here is the data requied by the exercise
summary(ssdat)
```

```
##    schoolcode         adprog
##  Min.   :  10101   Length:2300
##  1st Qu.:  30107   Class :character
##  Median :  50606   Mode  :character
##  Mean   : 665894
##  3rd Qu.:  70602
##  Max.   :9100101
##
##                               schoolname
##  KUMASI TECH. INST., KUMASI        :  13
##  BOLGATANGA TECH. INST., BOLGATANGA:  11
##  CAPE COAST TECH. INST., CAPE COAST:  11
##  KPANDO TECH. INST., KPANDO        :  11
##  ANLO TECH. INST., ANLOGA          :   9
##  ASUANSI TECH. INST., ASUANSI      :   9
##  (Other)                           :2236
##                                sssdistrict      ssslong
##  Accra Metropolitan                  : 106   Min.   :-2.9267
##  Kumasi Metro                        : 100   1st Qu.:-1.5972
##  Ho Municipal                        :  69   Median :-0.9692
##  Shama/Ahanta/East (Sekondi/Takoradi):  63   Mean   :-0.9183
##  Cape Coast Municipal                :  59   3rd Qu.:-0.1971
##  Keta                                :  56   Max.   : 1.0327
##  (Other)                             :1847
##      ssslat          quality          cutoff           size
##  Min.   : 4.835   Min.   :209.0   Min.   :158.0   Min.   :  1.00
##  1st Qu.: 5.786   1st Qu.:248.7   1st Qu.:215.0   1st Qu.: 28.00
##  Median : 6.415   Median :268.5   Median :240.0   Median : 48.00
##  Mean   : 6.772   Mean   :282.8   Mean   :255.5   Mean   : 60.53
##  3rd Qu.: 7.184   3rd Qu.:308.8   3rd Qu.:286.0   3rd Qu.: 85.50
##  Max.   :11.036   Max.   :445.0   Max.   :433.0   Max.   :360.00
##
```

Excercise 3 Distance

```
# Prepare location of junior high schools
jssloc <- datjss[,2:4]
colnames(jssloc) <- c("jssdistrict","jsslong", "jsslat")

# Merge JSS locations with choices(school,program) in the cleaned datstu data and drop 'NA' district
datadmit <- merge(datadmit,jssloc, by = "jssdistrict", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$jsslat),]

# Prepare location of high schools, collapse each school using unique()
sssloc <- unique(ssdat[,c("schoolcode","ssslong", "ssslat")])

# Merge location of high school,then, calculate the distance from given formular.
# Repeat them by choice, i.e., choice1, ... , choice6
colnames(sssloc) <- c("schoolcode1", "ssslong1", "ssslat1")
datadmit <- merge(datadmit,sssloc, by ="schoolcode1", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat1),]
dist1 <- sqrt(((69.172*(datadmit$ssslong1-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist1)

colnames(sssloc) <- c("schoolcode2", "ssslong2", "ssslat2")
datadmit <- merge(datadmit,sssloc, by ="schoolcode2", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat2),]
dist2 <- sqrt(((69.172*(datadmit$ssslong2-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist2)

colnames(sssloc) <- c("schoolcode3", "ssslong3", "ssslat3")
datadmit <- merge(datadmit,sssloc, by ="schoolcode3", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat3),]
dist3 <- sqrt(((69.172*(datadmit$ssslong3-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist3)

colnames(sssloc) <- c("schoolcode4", "ssslong4", "ssslat4")
datadmit <- merge(datadmit,sssloc, by ="schoolcode4", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat4),]
dist4 <- sqrt(((69.172*(datadmit$ssslong4-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist4)

colnames(sssloc) <- c("schoolcode5", "ssslong5", "ssslat5")
datadmit <- merge(datadmit,sssloc, by ="schoolcode5", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat5),]
dist5 <- sqrt(((69.172*(datadmit$ssslong5-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist5)
```

```r
colnames(sssloc) <- c("schoolcode6", "ssslong6", "ssslat6")
datadmit <- merge(datadmit,sssloc, by ="schoolcode6", all.x = TRUE)
datadmit <- datadmit[!is.na(datadmit$ssslat6),]
dist6 <- sqrt(((69.172*(datadmit$ssslong6-datadmit$jsslong))*cos(datadmit$jsslat/57.3))^2 + (69.172*(
datadmit <- cbind(datadmit, dist6)

# The Summary of required data
summary(datadmit)
```

```
##   schoolcode6       schoolcode5       schoolcode4       schoolcode3
##  Min.   :  10102   Min.   :  10102   Min.   :  10101   Min.   :  10101
##  1st Qu.:  21010   1st Qu.:  21007   1st Qu.:  21302   1st Qu.:  21303
##  Median :  50204   Median :  50204   Median :  50139   Median :  50109
##  Mean   :  45346   Mean   :  45168   Mean   : 212080   Mean   : 171292
##  3rd Qu.:  60303   3rd Qu.:  60303   3rd Qu.:  60701   3rd Qu.:  60604
##  Max.   :9090401   Max.   :9090401   Max.   :9100101   Max.   :9100101
##
##   schoolcode2       schoolcode1
##  Min.   :  10101   Min.   :  10101
##  1st Qu.:  21303   1st Qu.:  21303
##  Median :  50105   Median :  50105
##  Mean   : 154187   Mean   : 143791
##  3rd Qu.:  60601   3rd Qu.:  60304
##  Max.   :9100101   Max.   :9100101
##
##                                  jssdistrict           X
##  Accra Metropolitan                  :13770   Min.   :179888
##  Kumasi Metro                        :11205   1st Qu.:219391
##  Tema                                : 5653   Median :260475
##  Ga West (Amasaman)                  : 3635   Mean   :260497
##  Shama/Ahanta/East (Sekondi/Takoradi): 3279   3rd Qu.:300831
##  Ga East (Abokobi)                   : 3101   Max.   :340823
##  (Other)                             :90120
##      score            agey            male                 choicepgm1
##  Min.   :185.0   Min.   : 9.00   Min.   :0.000   General Arts   :52055
##  1st Qu.:256.0   1st Qu.:15.00   1st Qu.:0.000   Business       :26204
##  Median :288.0   Median :16.00   Median :1.000   General Science:18994
##  Mean   :295.4   Mean   :16.66   Mean   :0.596   Home Economics :12076
##  3rd Qu.:329.0   3rd Qu.:18.00   3rd Qu.:1.000   Visual Arts    : 7953
##  Max.   :469.0   Max.   :54.00   Max.   :1.000   Agriculture    : 7945
##                  NA's   :172                     (Other)        : 5536
##           choicepgm2               choicepgm3               choicepgm4
##  General Arts   :50638   General Arts   :50907   General Arts   :50005
##  Business       :28599   Business       :27119   Business       :25227
##  General Science:14137   Home Economics :13657   Home Economics :14273
##  Home Economics :12915   General Science:12070   Agriculture    :13127
##  Agriculture    : 9782   Agriculture    :11054   General Science:10531
##  Visual Arts    : 8675   Visual Arts    : 9433   Visual Arts    : 9990
##  (Other)        : 6017   (Other)        : 6523   (Other)        : 7610
##           choicepgm5               choicepgm6       rankplace
##  General Arts   :54833   General Arts   :55364   Min.   :1.000
##  Business       :26452   Business       :25100   1st Qu.:1.000
##  Home Economics :12703   Home Economics :13578   Median :2.000
##  Agriculture    :11557   Agriculture    :12318   Mean   :2.434
##  General Science:11360   General Science: 9641   3rd Qu.:3.000
##  Visual Arts    : 7832   Visual Arts    : 8471   Max.   :6.000
##  (Other)        : 6026   (Other)        : 6291
##    schoolcode          adprog              jsslong             jsslat
##  Min.   :  10101   Length:130763      Min.   :-3.0435   Min.   : 4.835
##  1st Qu.:  21501   Class :character   1st Qu.:-1.6237   1st Qu.: 5.665
##  Median :  50113   Mode  :character   Median :-1.0217   Median : 6.258
##  Mean   : 228266                      Mean   :-1.0452   Mean   : 6.627
##  3rd Qu.:  60901                      3rd Qu.:-0.1971   3rd Qu.: 7.002
##  Max.   :9100101                      Max.   : 1.0327   Max.   :11.036
```

```
##
##     ssslong1          ssslat1            dist1           ssslong2
##  Min.   :-2.9267   Min.   : 4.835   Min.   :  0.00   Min.   :-2.9267
##  1st Qu.:-1.5972   1st Qu.: 5.690   1st Qu.:  0.00   1st Qu.:-1.5972
##  Median :-1.1801   Median : 6.436   Median : 20.29   Median :-1.0180
##  Mean   :-1.0390   Mean   : 6.674   Mean   : 34.38   Mean   :-1.0274
##  3rd Qu.:-0.2975   3rd Qu.: 6.901   3rd Qu.: 48.31   3rd Qu.:-0.2975
##  Max.   : 1.0327   Max.   :11.036   Max.   :418.47   Max.   : 1.0327
##
##     ssslat2           dist2            ssslong3          ssslat3
##  Min.   : 4.835   Min.   :  0.00   Min.   :-2.9267   Min.   : 4.835
##  1st Qu.: 5.690   1st Qu.:  0.00   1st Qu.:-1.6237   1st Qu.: 5.726
##  Median : 6.415   Median : 20.94   Median :-1.0171   Median : 6.415
##  Mean   : 6.695   Mean   : 33.17   Mean   :-1.0272   Mean   : 6.703
##  3rd Qu.: 7.002   3rd Qu.: 45.85   3rd Qu.:-0.2975   3rd Qu.: 7.028
##  Max.   :11.036   Max.   :450.35   Max.   : 1.0327   Max.   :11.036
##
##      dist3           ssslong4          ssslat4            dist4
##  Min.   :  0.00   Min.   :-2.9267   Min.   : 4.835   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:-1.6237   1st Qu.: 5.726   1st Qu.:  0.00
##  Median : 18.74   Median :-1.0180   Median : 6.383   Median : 14.43
##  Mean   : 30.87   Mean   :-1.0367   Mean   : 6.706   Mean   : 26.42
##  3rd Qu.: 41.74   3rd Qu.:-0.2682   3rd Qu.: 7.028   3rd Qu.: 35.61
##  Max.   :433.23   Max.   : 1.0327   Max.   :11.036   Max.   :412.51
##
##     ssslong5          ssslat5            dist5            ssslong6
##  Min.   :-2.9267   Min.   : 4.835   Min.   :  0.000   Min.   :-2.9267
##  1st Qu.:-1.6237   1st Qu.: 5.778   1st Qu.:  8.813   1st Qu.:-1.5628
##  Median :-0.9692   Median : 6.436   Median : 23.765   Median :-1.0054
##  Mean   :-1.0342   Mean   : 6.681   Mean   : 30.448   Mean   :-1.0392
##  3rd Qu.:-0.3561   3rd Qu.: 7.184   3rd Qu.: 47.591   3rd Qu.:-0.3561
##  Max.   : 1.0327   Max.   :11.036   Max.   :368.827   Max.   : 1.0327
##
##     ssslat6           dist6
##  Min.   : 4.835   Min.   :  0.00
##  1st Qu.: 5.786   1st Qu.:  9.44
##  Median : 6.436   Median : 24.12
##  Mean   : 6.687   Mean   : 31.01
##  3rd Qu.: 7.031   3rd Qu.: 48.31
##  Max.   :11.036   Max.   :373.97
##
```

Exercise 4 Descriptive Characteristics

```r
# Remove unused values
rm(dist1, dist2, dist3, dist4, dist5, dist6)

# merge variable "cutoff" and "quality" to original data by choices [1:6]
ssdat <- ssdat[, c("schoolcode","adprog","cutoff","quality")]
colnames(ssdat) <- c("schoolcode1","choicepgm1","cutoff1","quality1")
datadmit <- merge(datadmit,ssdat, by = c("schoolcode1","choicepgm1"), all.x = TRUE)

colnames(ssdat) <- c("schoolcode2","choicepgm2","cutoff2","quality2")
datadmit <- merge(datadmit,ssdat, by = c("schoolcode2","choicepgm2"), all.x = TRUE)

colnames(ssdat) <- c("schoolcode3","choicepgm3","cutoff3","quality3")
datadmit <- merge(datadmit,ssdat, by = c("schoolcode3","choicepgm3"), all.x = TRUE)

colnames(ssdat) <- c("schoolcode4","choicepgm4","cutoff4","quality4")
datadmit <- merge(datadmit,ssdat, by = c("schoolcode4","choicepgm4"), all.x = TRUE)

colnames(ssdat) <- c("schoolcode5","choicepgm5","cutoff5","quality5")
datadmit <- merge(datadmit,ssdat, by = c("schoolcode5","choicepgm5"), all.x = TRUE)
```

```
  colnames(ssdat) <- c("schoolcode6","choicepgm6","cutoff6","quality6")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode6","choicepgm6"), all.x = TRUE)

# Calculate mean and sd of "Cutoff", "Quality" and "Distance"
result <- data.frame("Choice1", "Choice2", "Choice3", "Choice4", "Choice5", "Choice6")

xcutoff <- c(mean(as.numeric(datadmit$cutoff1), na.rm=TRUE),
             mean(as.numeric(datadmit$cutoff2), na.rm=TRUE),
             mean(as.numeric(datadmit$cutoff3), na.rm=TRUE),
             mean(as.numeric(datadmit$cutoff4), na.rm=TRUE),
             mean(as.numeric(datadmit$cutoff5), na.rm=TRUE),
             mean(as.numeric(datadmit$cutoff6), na.rm=TRUE)
             )

sdcutoff <- c(sd(as.numeric(datadmit$cutoff1), na.rm=TRUE),
             sd(as.numeric(datadmit$cutoff2), na.rm=TRUE),
             sd(as.numeric(datadmit$cutoff3), na.rm=TRUE),
             sd(as.numeric(datadmit$cutoff4), na.rm=TRUE),
             sd(as.numeric(datadmit$cutoff5), na.rm=TRUE),
             sd(as.numeric(datadmit$cutoff6), na.rm=TRUE)
             )

xquality <- c(mean(as.numeric(datadmit$quality1), na.rm=TRUE),
             mean(as.numeric(datadmit$quality2), na.rm=TRUE),
             mean(as.numeric(datadmit$quality3), na.rm=TRUE),
             mean(as.numeric(datadmit$quality4), na.rm=TRUE),
             mean(as.numeric(datadmit$quality5), na.rm=TRUE),
             mean(as.numeric(datadmit$quality6), na.rm=TRUE)
             )

sdquality  <- c(sd(as.numeric(datadmit$quality1), na.rm=TRUE),
             sd(as.numeric(datadmit$quality2), na.rm=TRUE),
             sd(as.numeric(datadmit$quality3), na.rm=TRUE),
             sd(as.numeric(datadmit$quality4), na.rm=TRUE),
             sd(as.numeric(datadmit$quality5), na.rm=TRUE),
             sd(as.numeric(datadmit$quality6), na.rm=TRUE)
             )

xdistance <- c(mean(as.numeric(datadmit$dist1), na.rm=TRUE),
             mean(as.numeric(datadmit$dist2), na.rm=TRUE),
             mean(as.numeric(datadmit$dist3), na.rm=TRUE),
             mean(as.numeric(datadmit$dist4), na.rm=TRUE),
             mean(as.numeric(datadmit$dist5), na.rm=TRUE),
             mean(as.numeric(datadmit$dist6), na.rm=TRUE)
             )

sddistance  <- c(sd(as.numeric(datadmit$dist1), na.rm=TRUE),
             sd(as.numeric(datadmit$dist2), na.rm=TRUE),
             sd(as.numeric(datadmit$dist3), na.rm=TRUE),
             sd(as.numeric(datadmit$dist4), na.rm=TRUE),
             sd(as.numeric(datadmit$dist5), na.rm=TRUE),
             sd(as.numeric(datadmit$dist6), na.rm=TRUE)
             )
result <- rbind(xcutoff, sdcutoff, xquality, sdquality, xdistance, sddistance)
colnames(result) <- c("Choice1", "Choice2", "Choice3", "Choice4", "Choice5", "Choice6")
result
```

```
##              Choice1   Choice2   Choice3   Choice4   Choice5   Choice6
## xcutoff    315.38556 297.25446 284.05394 269.80623 255.18089 250.13240
## sdcutoff    53.41517  49.93377  47.90471  46.08840  32.45614  31.95417
## xquality   336.56375 319.38690 307.48907 295.24666 283.20489 278.66488
## sdquality   48.05315  44.04184  41.73266  39.67875  26.20330  25.96806
## xdistance   34.38466  33.17020  30.86659  26.42078  30.44816  31.00968
## sddistance  47.99236  46.08994  44.07567  41.75364  28.53081  28.59082
```

```r
# Divide student into quartile according to her score
summary(datadmit$score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   185.0   256.0   288.0   295.4   329.0   469.0
```

```r
datadmit$stQr <- NA
datadmit$stQr[which(datadmit$score < 256)] <- 1
datadmit$stQr[which(datadmit$score >= 256 & datadmit$score < 288)] <- 2
datadmit$stQr[which(datadmit$score >= 288 & datadmit$score < 329)] <- 3
datadmit$stQr[which(datadmit$score >= 329)] <- 4

# Calculate descriptive statistics
rm(quartile, qdata)
```

```
## Warning in rm(quartile, qdata): object 'quartile' not found
```

```
## Warning in rm(quartile, qdata): object 'qdata' not found
```

```r
quartile <- data.frame("Choice1", "Choice2", "Choice3", "Choice4", "Choice5", "Choice6")
frow <- c(0,0,0,0,0,0)
quartile <- rbind(frow)

for (i in 1:4) {
  qdata <- datadmit[which(datadmit$stQr == i), ]

  qxcutoff <- c(mean(as.numeric(qdata$cutoff1), na.rm=TRUE),
             mean(as.numeric(qdata$cutoff2), na.rm=TRUE),
             mean(as.numeric(qdata$cutoff3), na.rm=TRUE),
             mean(as.numeric(qdata$cutoff4), na.rm=TRUE),
             mean(as.numeric(qdata$cutoff5), na.rm=TRUE),
             mean(as.numeric(qdata$cutoff6), na.rm=TRUE)
             )

  qsdcutoff <- c(sd(as.numeric(qdata$cutoff1), na.rm=TRUE),
             sd(as.numeric(qdata$cutoff2), na.rm=TRUE),
             sd(as.numeric(qdata$cutoff3), na.rm=TRUE),
             sd(as.numeric(qdata$cutoff4), na.rm=TRUE),
             sd(as.numeric(qdata$cutoff5), na.rm=TRUE),
             sd(as.numeric(qdata$cutoff6), na.rm=TRUE)
             )

  qxquality <- c(mean(as.numeric(qdata$quality1), na.rm=TRUE),
             mean(as.numeric(qdata$quality2), na.rm=TRUE),
             mean(as.numeric(qdata$quality3), na.rm=TRUE),
             mean(as.numeric(qdata$quality4), na.rm=TRUE),
             mean(as.numeric(qdata$quality5), na.rm=TRUE),
             mean(as.numeric(qdata$quality6), na.rm=TRUE)
             )

  qsdquality  <- c(sd(as.numeric(qdata$quality1), na.rm=TRUE),
             sd(as.numeric(qdata$quality2), na.rm=TRUE),
             sd(as.numeric(qdata$quality3), na.rm=TRUE),
             sd(as.numeric(qdata$quality4), na.rm=TRUE),
             sd(as.numeric(qdata$quality5), na.rm=TRUE),
             sd(as.numeric(qdata$quality6), na.rm=TRUE)
             )

  qxdistance <- c(mean(as.numeric(qdata$dist1), na.rm=TRUE),
             mean(as.numeric(qdata$dist2), na.rm=TRUE),
             mean(as.numeric(qdata$dist3), na.rm=TRUE),
             mean(as.numeric(qdata$dist4), na.rm=TRUE),
             mean(as.numeric(qdata$dist5), na.rm=TRUE),
             mean(as.numeric(datadmit$dist6), na.rm=TRUE)
             )
  qsddistance  <- c(sd(as.numeric(qdata$dist1), na.rm=TRUE),
             sd(as.numeric(qdata$dist2), na.rm=TRUE),
             sd(as.numeric(qdata$dist3), na.rm=TRUE),
             sd(as.numeric(qdata$dist4), na.rm=TRUE),
             sd(as.numeric(qdata$dist5), na.rm=TRUE),
             sd(as.numeric(qdata$dist6), na.rm=TRUE)
             )

 quartile <- rbind(quartile, qxcutoff, qsdcutoff, qxquality, qsdquality, qxdistance, qsddistance)
 rm(qxcutoff, qsdcutoff, qxquality, qsdquality, qxdistance, qsddistance)
}

quartile <- quartile[-1,]
row.names(quartile)<- c("mean.cut.q1", "sd.cut.q1", "mean.qual.q1", "sd.qual.q1","mean.dist.q1","sd.dis
                          "mean.cut.q2", "sd.cut.q2", "mean.qual.q2", "sd.qual.q2","mean.dist.q2","sd.di
                          "mean.cut.q3", "sd.cut.q3", "mean.qual.q3", "sd.qual.q3","mean.dist.q3","sd.di
                          "mean.cut.q4", "sd.cut.q4", "mean.qual.q4", "sd.qual.q4","mean.dist.q4","sd.di
                          )
colnames(quartile) <- c("Choice1", "Choice2", "Choice3", "Choice4", "Choice5", "Choice6")
#Print result
quartile
```

```
##                Choice1   Choice2   Choice3   Choice4   Choice5   Choice6
## mean.cut.q1  276.66997 262.80348 253.46973 242.50353 242.63734 238.38356
## sd.cut.q1     44.03093  40.24884  38.97435  37.32502  31.27866  30.26387
## mean.qual.q1 300.94764 288.24508 280.12242 271.05266 271.00272 267.15122
## sd.qual.q1    38.18512  34.60469  33.42321  31.94334  25.70278  25.14254
## mean.dist.q1  28.51427  29.10862  28.23327  25.46411  29.87025  31.00968
## sd.dist.q1    45.25706  44.10377  42.94375  41.17081  29.14634  29.19500
## mean.cut.q2  296.57800 279.97978 267.98428 255.10190 251.37324 246.81518
## sd.cut.q2     44.51153  41.28993  39.77415  38.38146  31.79764  31.28516
## mean.qual.q2 318.94408 303.60229 293.00618 282.16605 279.24106 275.19133
## sd.qual.q2    38.60791  35.54368  33.92918  32.37772  25.55595  25.37030
## mean.dist.q2  32.23710  31.65921  30.13329  26.36467  30.03666  31.00968
## sd.dist.q2    49.12941  47.70704  45.91059  43.50364  28.73384  28.78936
## mean.cut.q3  323.27131 303.36958 288.59678 273.37908 259.83559 254.46094
## sd.cut.q3     43.01964  41.80838  40.67447  39.50504  31.43416  31.27641
## mean.qual.q3 343.18864 324.67777 311.37958 298.25629 287.36526 282.54808
## sd.qual.q3    37.84537  36.02182  34.75239  33.33981  24.57416  24.74428
## mean.dist.q3  34.56131  33.42151  31.17684  26.61729  30.84777  31.00968
## sd.dist.q3    48.77901  46.63911  44.74205  42.31957  28.15696  28.28514
## mean.cut.q4  363.58389 341.54074 324.99672 307.20833 266.56297 260.62880
## sd.cut.q4     37.22376  37.36432  38.89449  40.92360  30.21728  30.54142
## mean.qual.q4 381.83139 359.81604 344.39695 328.59358 294.89173 289.52279
## sd.qual.q4    34.10365  33.16425  33.71793  34.96533  22.61092  23.06570
## mean.dist.q4  42.02907  38.35563  33.84586  27.21802  31.01489  31.00968
## sd.dist.q4    47.69991  45.37073  42.49419  39.98462  28.07316  28.08144
```

Exercise 5 Diversification

```
  # Create deciles for choices(school,program) by "cutoff"
  quantile(ssdat$cutoff6, c(.1, .2, .3, .4, .5, .6, .7, .8, .9))
```

```
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
## 207 212 218 226 240 256 275 298 335
```

```
  ssdat$deci <- NA
  ssdat$deci[which(ssdat$cutoff6 < 207)] <- 1
  ssdat$deci[which(ssdat$cutoff6 >= 207 & ssdat$cutoff6 < 212)] <- 2
  ssdat$deci[which(ssdat$cutoff6 >= 212 & ssdat$cutoff6 < 218)] <- 3
  ssdat$deci[which(ssdat$cutoff6 >= 218 & ssdat$cutoff6 < 226)] <- 4
  ssdat$deci[which(ssdat$cutoff6 >= 226 & ssdat$cutoff6 < 240)] <- 5
  ssdat$deci[which(ssdat$cutoff6 >= 240 & ssdat$cutoff6 < 256)] <- 6
  ssdat$deci[which(ssdat$cutoff6 >= 256 & ssdat$cutoff6 < 275)] <- 7
  ssdat$deci[which(ssdat$cutoff6 >= 275 & ssdat$cutoff6 < 298)] <- 8
  ssdat$deci[which(ssdat$cutoff6 >= 298 & ssdat$cutoff6 < 335)] <- 9
  ssdat$deci[which(ssdat$cutoff6 >= 335)] <- 10

  # Assign value of decile to each choice
  ssdat <- ssdat[,-3:-4]
  colnames(ssdat) <- c("schoolcode1","choicepgm1","deci1")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode1","choicepgm1"), all.x = TRUE)

  colnames(ssdat) <- c("schoolcode2","choicepgm2","deci2")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode2","choicepgm2"), all.x = TRUE)

  colnames(ssdat) <- c("schoolcode3","choicepgm3","deci3")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode3","choicepgm3"), all.x = TRUE)

  colnames(ssdat) <- c("schoolcode4","choicepgm4","deci4")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode4","choicepgm4"), all.x = TRUE)

  colnames(ssdat) <- c("schoolcode5","choicepgm5","deci5")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode5","choicepgm5"), all.x = TRUE)

  colnames(ssdat) <- c("schoolcode6","choicepgm6","deci6")
  datadmit <- merge(datadmit,ssdat, by = c("schoolcode6","choicepgm6"), all.x = TRUE)

  # Calculate the number of unique group within the application
totgroup <- datadmit[,54:59]
totgroup$ngroup <- apply(totgroup, 1, function(x)length(unique(x)) )
datadmit <- merge(datadmit,totgroup, by =c("deci1", "deci2", "deci3", "deci4" , "deci5" , "deci6"), all
summary(datadmit$ngroup)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.179   4.000   6.000
```

```
  # Calculate the number of unique group within the application by student's score quartile
ngroup1 <- summary(datadmit$ngroup[which(datadmit$stQr ==1)])
ngroup2 <- summary(datadmit$ngroup[which(datadmit$stQr ==2)])
ngroup3 <- summary(datadmit$ngroup[which(datadmit$stQr ==3)])
ngroup4 <- summary(datadmit$ngroup[which(datadmit$stQr ==4)])
sumngroup <- rbind(ngroup1, ngroup2, ngroup3, ngroup4)
sumngroup
```

```
##         Min. 1st Qu. Median     Mean 3rd Qu. Max.
## ngroup1    1       4      4 4.275826       5    6
## ngroup2    1       3      4 3.957598       4    6
## ngroup3    1       3      3 3.496592       4    6
## ngroup4    1       3      3 3.069786       3    6
```