#1. Data wrangling

```
#input the data, change height units to inches
height.summary <- read.csv("~/Desktop/2016-heights.csv", header = TRUE, stringsAsFactors = TRUE)
height.inches <- 12*height.summary$Feet + height.summary$Inches
height.inches <- as.data.frame(height.inches)
height.summary$Feet <- NULL
height.summary$Inches <- NULL
height.summary <- data.frame(height.summary, height.inches)

#remove one suspicious answer, clean answers about ethnicity
height.summary <- height.summary[-c(43), ]
rownames(height.summary) <- c(1:66)

height.summary$Self.reported.ethnicity <- gsub("South Asian", "Asian",
height.summary$Self.reported.ethnicity)
height.summary$Self.reported.ethnicity <- gsub("Black", "African", height.summary$Self.reported.ethnicity)
height.summary$Self.reported.ethnicity <- gsub("Polish", "Others", height.summary$Self.reported.ethnicity)
height.summary$Self.reported.ethnicity <- gsub("Romanian", "Others",
height.summary$Self.reported.ethnicity)
height.summary$Self.reported.ethnicity <- gsub("Latin American", "Others",
height.summary$Self.reported.ethnicity)
```

#2. Plotting

```
library(ggplot2)
#1) distribution of heights
height.summary$Self.reported.ethnicity <- factor(height.summary$Self.reported.ethnicity, levels = c("White",
"Asian", "Hispanic", "African", "Others"), ordered = TRUE)
p1.1 <- ggplot(height.summary, aes(x = Self.reported.ethnicity, y = height.inches)) + geom_boxplot() +
geom_jitter() + labs(title = "Fig1.1 height vs ethnicity", x = "Ethnicity", y = "height(inches)")
print(p1.1)
p1.2 <- ggplot(height.summary, aes(x = Self.reported.gender, y = height.inches)) + geom_boxplot() +
geom_jitter() + labs(title = "Fig1.2 height vs gender", x = "Gender", y = "height(inches)")
print(p1.2)
p1.3 <- ggplot(height.summary, aes(x = First.letter.of.last.name, y = height.inches)) + geom_boxplot() +
geom_jitter() + labs(title = "Fig1.3 height vs letter", x = "First letter of last name", y = "height(inches)")
print(p1.3)
# I chose to plot height against gender, ethnicity and age, because all of them are factors which could
potentially affect height.

#2) distribution of ages
p2.1 <- ggplot(height.summary, aes(x = Self.reported.ethnicity, y = Age)) + geom_boxplot() + geom_jitter() +
labs(title = "Fig2.1 age vs ethnicity", x = "Ethnicity", y = "age(years)")
print(p2.1)
p2.2 <- ggplot(height.summary, aes(x = Self.reported.gender, y = Age)) + geom_boxplot() + geom_jitter() +
labs(title = "Fig2.2 age vs gender", x = "Gender", y = "age(years)")
print(p2.2)
p2.3 <- ggplot(height.summary, aes(x = First.letter.of.last.name, y = Age)) + geom_boxplot() + geom_jitter() +
labs(title = "Fig2.3 age vs letter", x = "First letter of last name", y = "age(years)")
print(p2.3)
# I chose to plot age against gender, ethnicity and age, because all of them are factors which could potentially
affect age.

#3) age vs height
p3 <- ggplot(height.summary, aes(x = Age, y = height.inches)) + geom_point(aes(color = Self.reported.gender),
size = 4) + labs(title = "Fig3 age vs height")
print(p3)

#4) age vs height (separate figures)
p4 <- ggplot(height.summary, aes(x = Age, y = height.inches)) + geom_point(aes(color = Self.reported.gender),
size = 4) + labs(title = "Fig4 age vs height") + facet_grid(Self.reported.gender ~.)
print(p4)

#5) barchart
p5 <- ggplot(height.summary, aes(x = First.letter.of.last.name)) + geom_bar(aes(fill = Self.reported.ethnicity))
+ scale_x_discrete(limits = LETTERS, drop=FALSE) + scale_fill_discrete(drop=FALSE) + labs(title = "Fig5", x =
"First letter of last name", y = "Height(inches)")
print(p5)

pdf(file = 'proj2_figures.pdf')
print(p1.1);print(p1.2);print(p1.3);print(p2.1);print(p2.2);print(p2.3);print(p3);print(p4);print(p5)
dev.off() # The figures are attached in a separate file in the email.
```

#3.Statistics
#1) heights: man vs women
Males <- subset(height.summary, Self.reported.gender == 'Male', select = height.inches)
Females <- subset(height.summary, Self.reported.gender == 'Female', select = height.inches)
t.test(Males, Females)

```
        Welch Two Sample t-test

data:  Males and Females
t = 5.5385, df = 61.078, p-value = 6.84e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.374356 7.187549
sample estimates:
mean of x mean of y
 70.16667  64.88571
```

<u># The average of male heights is significantly higher than that of the female heights.</u>

#2) gender distribution across sections
height.summary <- height.summary[-c(59),]
height.summary$Self.reported.gender <- factor(height.summary$Self.reported.gender, levels = c('Male', 'Female'), ordered = TRUE)

gender.section <- table(height.summary$Group, height.summary$Self.reported.gender)
gender.section

```
  Male Female
1    6     10
2   14      7
3    5      8
4    5      9
```

prop.test(table(height.summary$Group, height.summary$Self.reported.gender), correct=FALSE)

```
        4-sample test for equality of proportions without continuity correction

data:  table(height.summary$Group, height.summary$Self.reported.gender)
X-squared = 4.9378, df = 3, p-value = 0.1764
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.3750000 0.6666667 0.3846154 0.3571429
```

<u>When checking the male/female ratio in individual groups, more women were sitting in each group except for group2, where more men were sitting there.</u>

# change absolute number of people into percentage.
total <- colSums(gender.section)
gender.section[,1] <- sapply(gender.section[,1], function(x) x/total[1])
gender.section[,2] <- sapply(gender.section[,2], function(x) x/total[2])
gender.section

	Male	Female
1	0.2000000	0.2941176
2	0.4666667	0.2058824
3	0.1666667	0.2352941
4	0.1666667	0.2647059

```
left.right.M <- t.test(c(gender.section[1,1],gender.section[2,1]), c(gender.section[3,1], gender.section[4,1]))
left.right.M
```

```
        Welch Two Sample t-test

data:  c(gender.section[1, 1], gender.section[2, 1]) and c(gender.section[3, 1], gender.section[4, 1])
t = 1.25, df = 1, p-value = 0.4296
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.527494  1.860827
sample estimates:
mean of x mean of y
0.3333333 0.1666667
```

```
left.right.F <- t.test(c(gender.section[1,2],gender.section[2,2]), c(gender.section[3,2], gender.section[4,2]))
left.right.F
```

```
        Welch Two Sample t-test

data:  c(gender.section[1, 2], gender.section[2, 2]) and c(gender.section[3, 2], gender.section[4, 2])
t = 0, df = 1.2195, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3903561  0.3903561
sample estimates:
mean of x mean of y
     0.25      0.25
```

```
middle.side.M<- t.test(c(gender.section[2,1],gender.section[3,1]), c(gender.section[1,1], gender.section[4,1]))
middle.side.M
```

```
        Welch Two Sample t-test

data:  c(gender.section[2, 1], gender.section[3, 1]) and c(gender.section[1, 1], gender.section[4, 1])
t = 0.88345, df = 1.0247, p-value = 0.5365
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.678519  1.945186
sample estimates:
mean of x mean of y
0.3166667 0.1833333
```

middle.side.F <- t.test(c(gender.section[2,2],gender.section[3,2]), c(gender.section[1,2], gender.section[4,2]))
middle.side.F

```
        Welch Two Sample t-test

data:  c(gender.section[2, 2], gender.section[3, 2]) and c(gender.section[1, 2], gender.section[4, 2])
t = -2.8284, df = 2, p-value = 0.1056
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.14830691  0.03065985
sample estimates:
mean of x mean of y
0.2205882 0.2794118
```

When comparing male/female ratio sitting on left vs right, center vs sides, there's no significant differences of distribution across sections.