```
#setwd('C:/Users/Ekaterina K/Desktop/project_star')

projectStar <- readRDS("projectStar_noNA.Rds")
head(projectStar)

projectStar_female <- subset(projectStar, subset = ssex == "female")
projectStar_medExp <- subset(projectStar, subset = projectStar$totexpk > 10 &
projectStar$totexpk < 15)

# OLS in R
lm(formula = tcombssk ~ totexpk, data = projectStar)

# dummy variables
projectStar$female <- ifelse(projectStar$ssex == "female", yes = 1, no = 0)
projectStar$male <- ifelse(projectStar$ssex == "male", yes = 1, no = 0)
head(projectStar)

lm(formula = tcombssk ~ female, data = projectStar)
lm(formula = tcombssk ~ male, data = projectStar)

my_ols <- lm(formula = tcombssk ~ ssex, data = projectStar)
mm <- model.matrix(my_ols)
head(mm)

projectStar2 <- projectStar
table(projectStar2$ssex)
projectStar2$ssex <- relevel(x = projectStar2$ssex, ref = "female")
lm(formula = tcombssk ~ ssex, data = projectStar2)


# Part 1
# Randomization is important because there are other factors than just
# the class size that influence students' performance, for example, gender.
# As we will see in Part 3, girls on average get higher scores than boys. That
is why if the assignment
# was not randomized, the results could be not conclusive. For instance, if in
smaller
# classes 75% of students were girls and in regular classes only 25% were
girls,
# then the improved performance in smaller classes could come from the fact
that
# they were dominated by girls, and regular classes were dominated by boys.
# That is why it is important to make sure that the composition of samples is
the
# same for small classes as it is for regular classes.


# PART 2
# let's drop "regular + aide" category.
projectStar <- subset(projectStar, subset = cltypek != "regular + aide class")
lm(formula = tcombssk ~ cltypek, data = projectStar)

# PART 3

# First, we check if the data is balanced by comparing the numbers of students
# in small classes and regular classes.
```

```
sum(projectStar$cltypek == "regular class")
sum(projectStar$cltypek == "small class")

# As we can see, the numbers are similar, so the data is balanced.

# Summary
summary(projectStar)

projectStar_small <- subset(projectStar, subset = cltypek == "small class")
projectStar_regular <- subset(projectStar, subset = cltypek == "regular
class")

summary(projectStar_small)
summary(projectStar_regular)

# From the summary we can see that for both math and reading scores, the
students
# from smaller classes have higher scores across all the quantiles.
# What can be a bit surprising is that, in spite of that, the maximum scores
for
# math and reading are identical for small and regular classes.
# This single statistic, however, is not very important in this study.

ols_sex <- lm(formula = tcombssk ~ ssex, data = projectStar)
ols_sex

ols_race <- lm(formula = tcombssk ~ srace, data = projectStar)
ols_race


ols_ses <- lm(formula = tcombssk ~ sesk, data = projectStar)
ols_ses



# As we can see from the summaries above, the randomization of the assignment
of
# students to different classes was carried out well and we have very similar
samples
# for both small and regular classes with respect to factors such as gender,
# socio-economic status, race, etc, which we showed above using the linear
fits,
# have very significant influence over the scores. We further confirm it by
looking at the ratios
# of students belonging to different groups for both class sizes.
# The results are presented below. As we can see, the ratios are very similar,
# making it a good sample to carry out the study of the influence of the class
# size on student's performance.


sum(projectStar_small$srace == "black") / sum(projectStar_small$srace ==
"white")
sum(projectStar_regular$srace == "black") / sum(projectStar_regular$srace ==
"white")
```

```
sum(projectStar_small$sesk == "non-free lunch") / sum(projectStar_small$sesk
== "free lunch")
sum(projectStar_regular$sesk == "non-free lunch") /
sum(projectStar_regular$sesk == "free lunch")

sum(projectStar_small$ssex == "male") / sum(projectStar_small$ssex ==
"female")
sum(projectStar_regular$ssex == "male") / sum(projectStar_regular$ssex ==
"female")



#PART 4
ols_math <- lm(tmathssk ~ cltypek, data = projectStar)
ols_math
summary(ols_math)

ols_read <- lm(treadssk ~ cltypek, data = projectStar)
ols_read
summary(ols_read)

# As we can see from the p-values, the class size is a significant predictor
# for both the math and reading scores. Also, small class size is more
beneficial
# for math scores, since on average it is associated with around 8.1 points
increase
# in the score for math and only around 5.7 points increase in the score for
reading.
```