Table of Contents

|  |  |
| --- | --- |
| Content | Page |
| Introduction |  |
| Abstract |  |
| Workflow and Conclusion |  |
| References |  |
| Appendix |  |

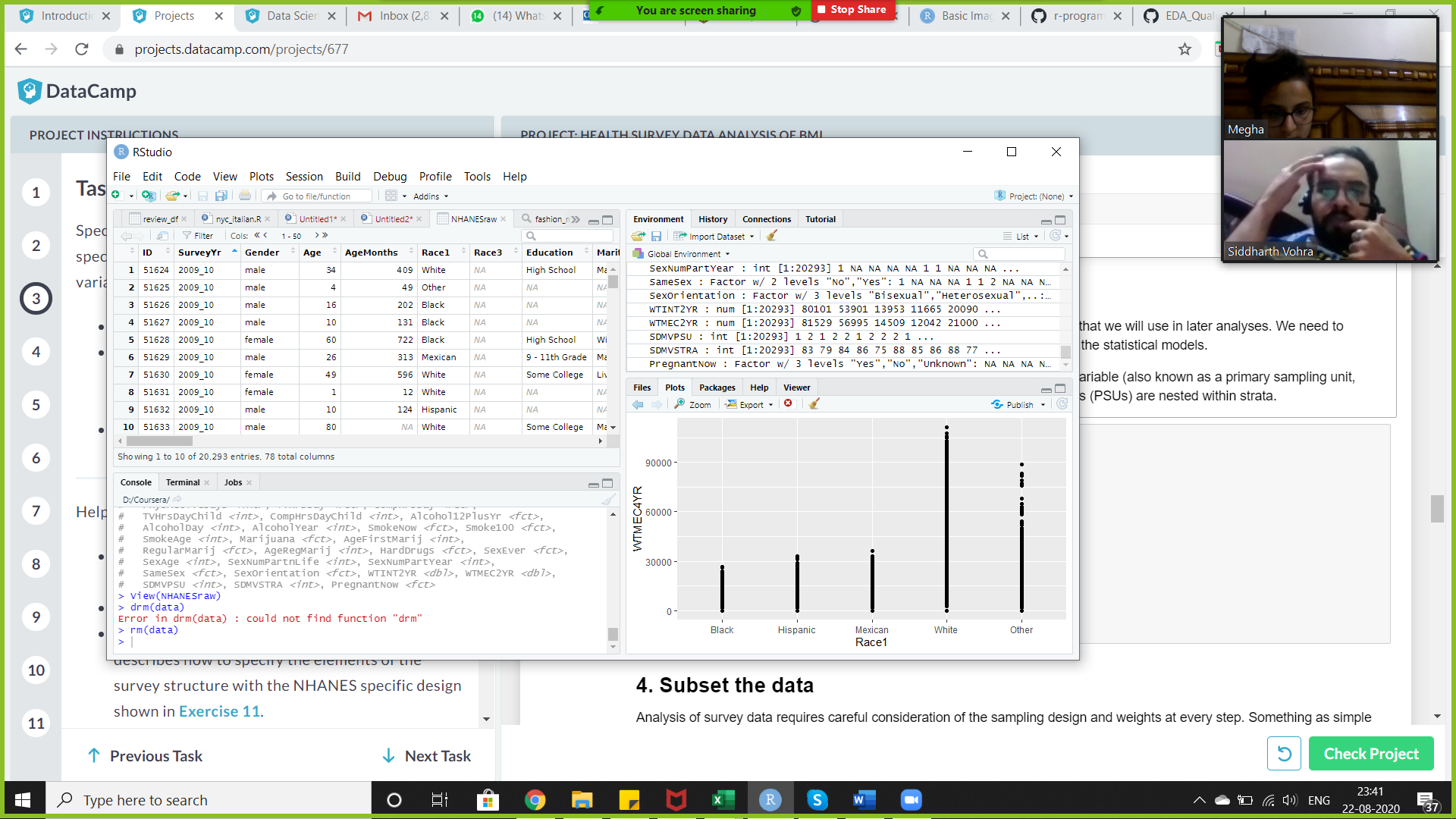
Introduction

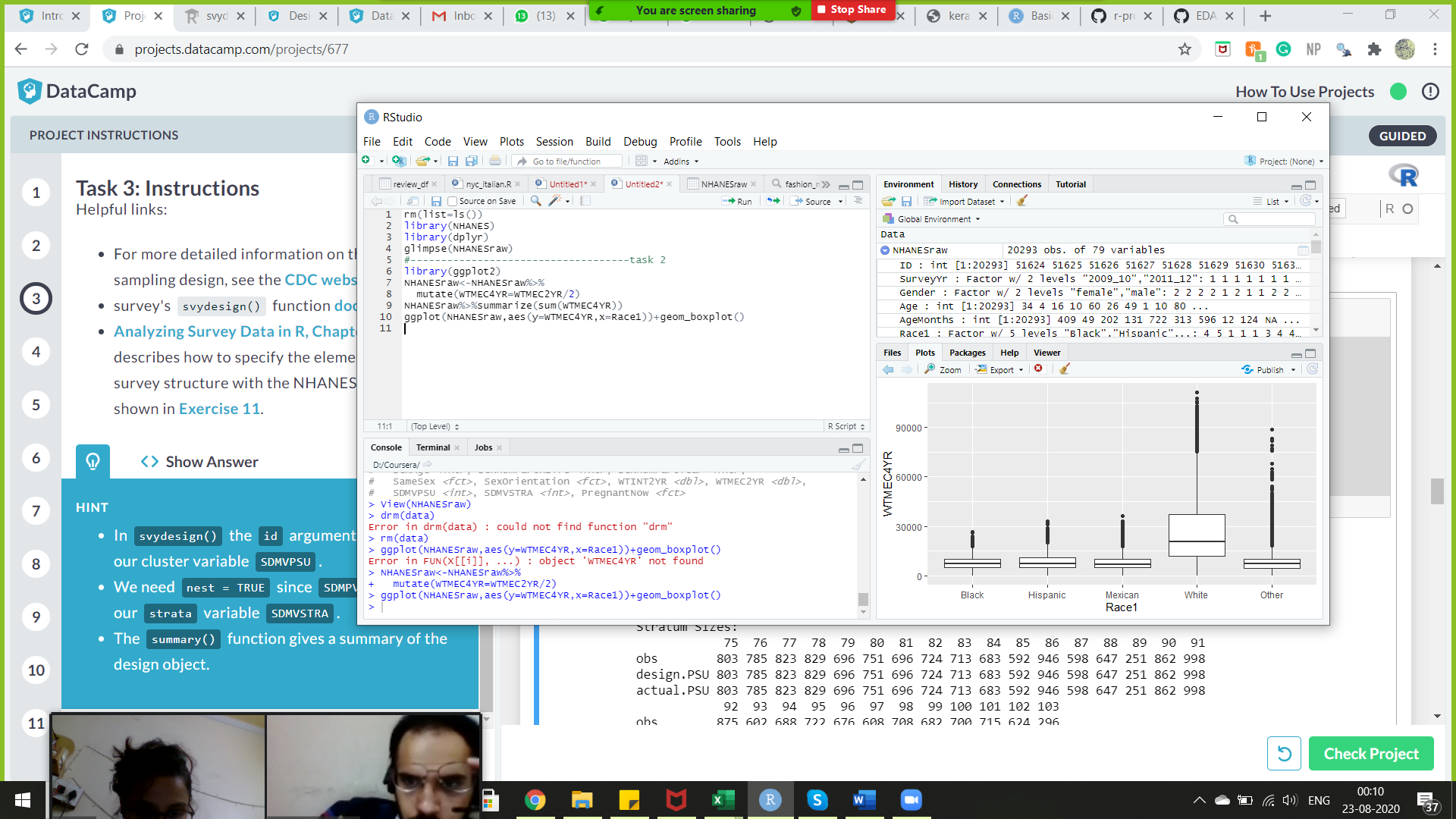
Abstract

Workflow and Conclusion

Since NHANESraw data spans 4 years (2009–2012) and the sampling weights are based on 2 years of data, we first need to create a weight variable that scales the sample across the full 4 years. Currently the weights sum to 2 times the US population number, so we need to divide the 2-year weight in half so that in total, the sum of the weights is equal to the US population.

The NHANES data has oversampled some geographic regions and specific minority groups. By examining the distribution of sampling weights for each race, we can see that Whites are undersampled and have higher weights while oversampled Black, Mexican, Hispanic people have lower weights since each sampled person in these minority groups represents fewer US people.





We will now use the survey package to specify the complex survey design that we will use in later analyses. We need to specify the design so the sampling weights and design are used properly in the statistical models.

The NHANESraw data contains a strata variable SDMVSTRA, and a cluster id variable (also known as a primary sampling unit, PSU), SDMVPSU, that accounts for design effects of clustering. These clusters (PSUs) are nested within strata.

**Tasks 1,2 and 3 involved importing libraries, reading csv files and pre-processing the data.**

#---task 4

**4. Analysis of survey data**

Analysis of survey data requires careful consideration of the sampling design and weights at every step. Something as simple as filtering the data becomes complicated when weights are involved.

When we wish to examine a subset of the data (i.e. the subpopulation of adult Hispanics with diabetes, or pregnant women), we must explicitly specify this in the design. We cannot simply remove that subset of the data through filtering the raw data because the survey weights will no longer be correct and will not add up to the full US population.

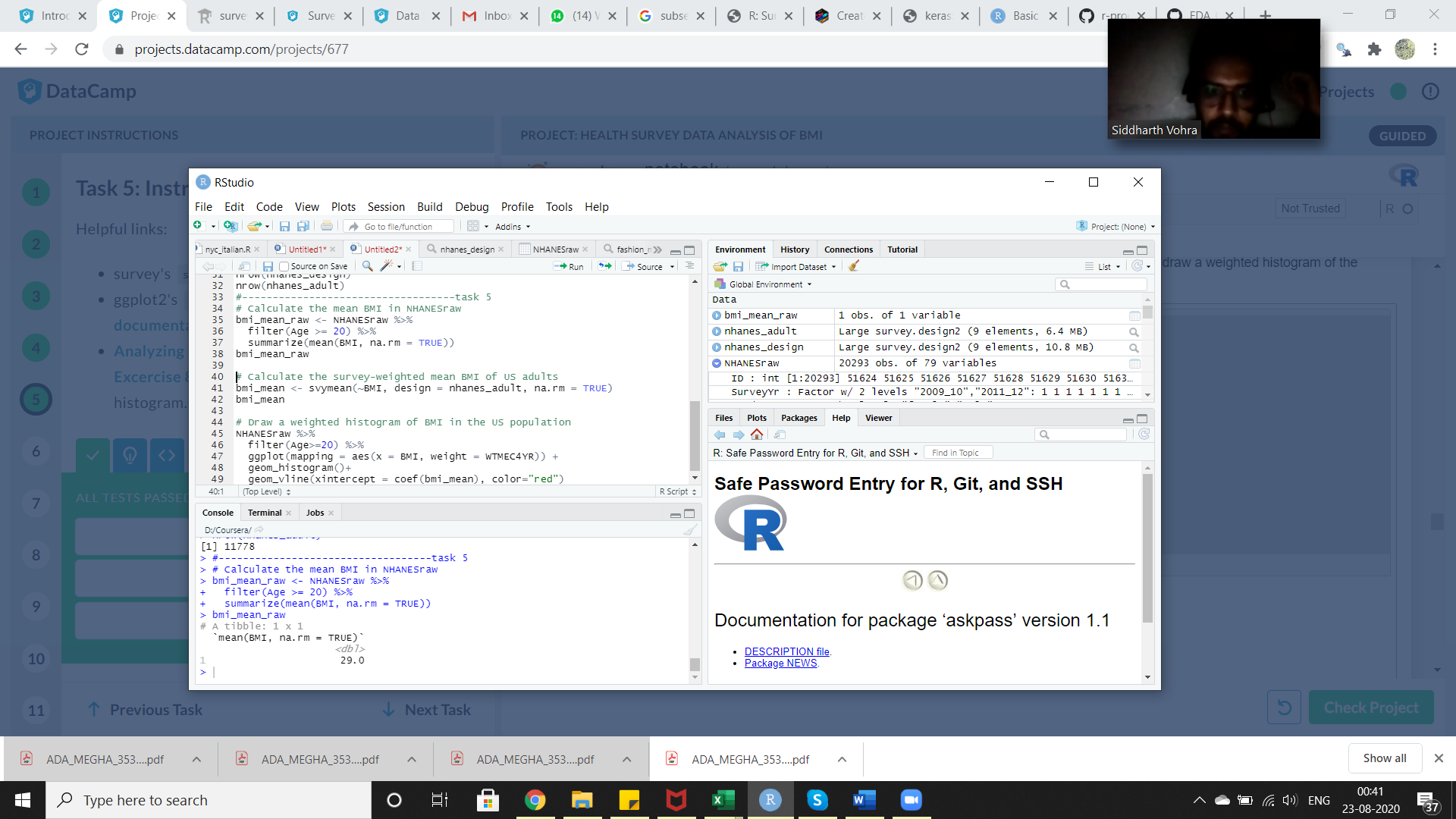
BMI categories are different for children and young adults younger than 20 so we will subset the data to only analyze adults of at least 20 years of age.

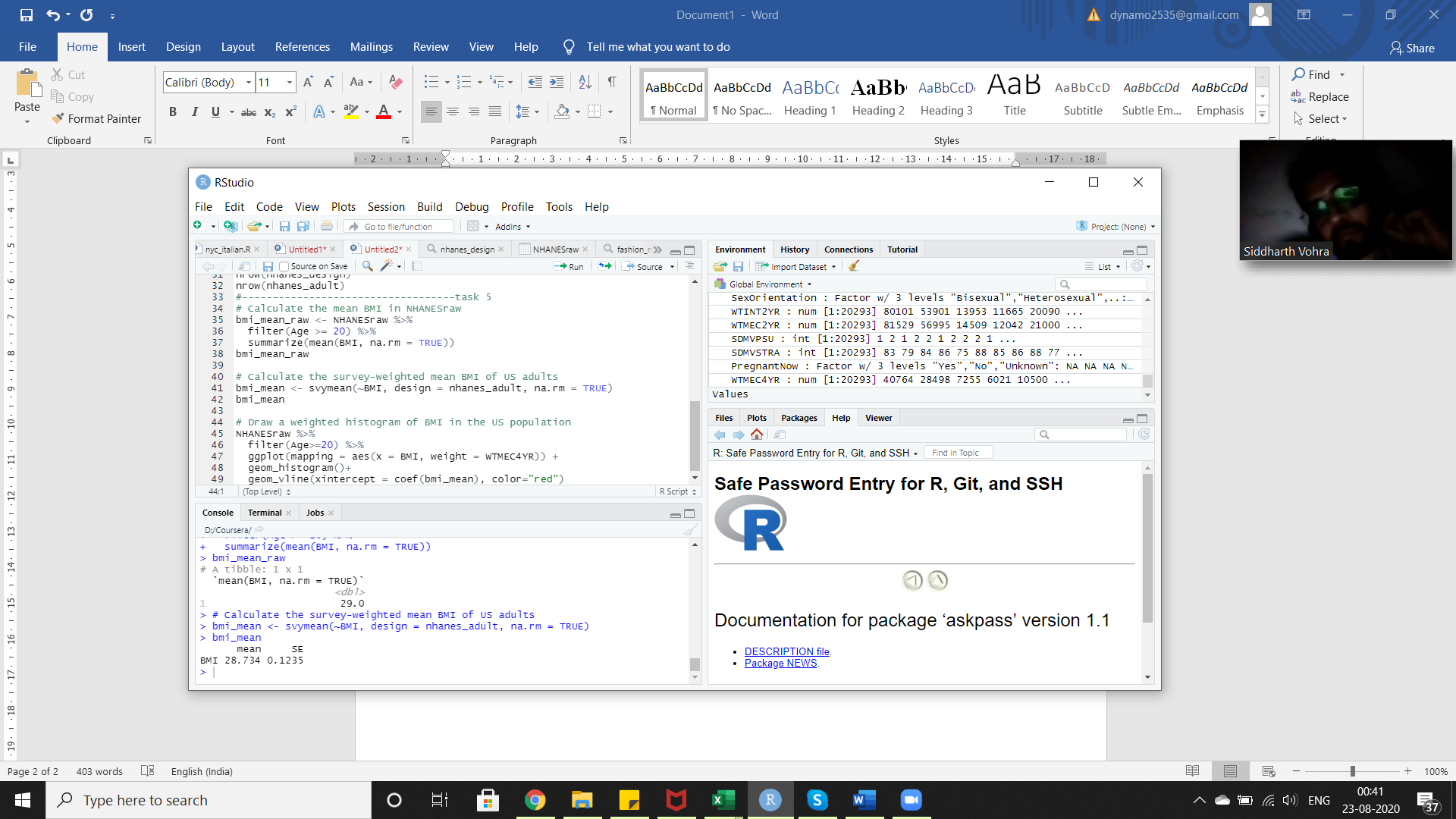
#---task 5

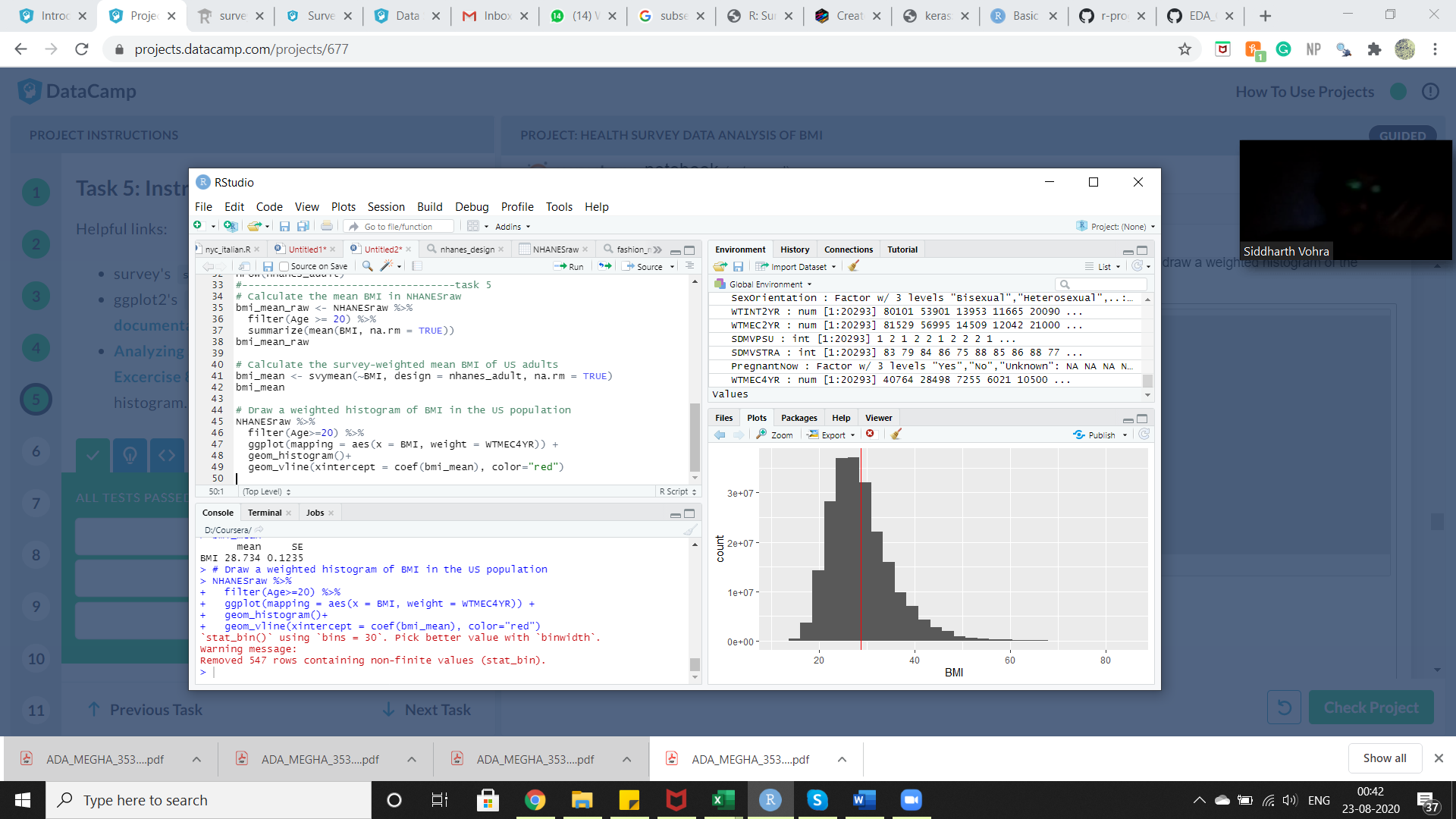
**5. Visualizing BMI**

We let svydesign() do its magic, but how does this help us learn about the full US population? With survey methods, we can use the sampling weights to estimate the true distributions of measurements within the entire population. This works for many statistics such as means, proportions, and standard deviations.

We'll use survey methods to estimate average BMI in the US adult population and also to draw a weighted histogram of the distribution.



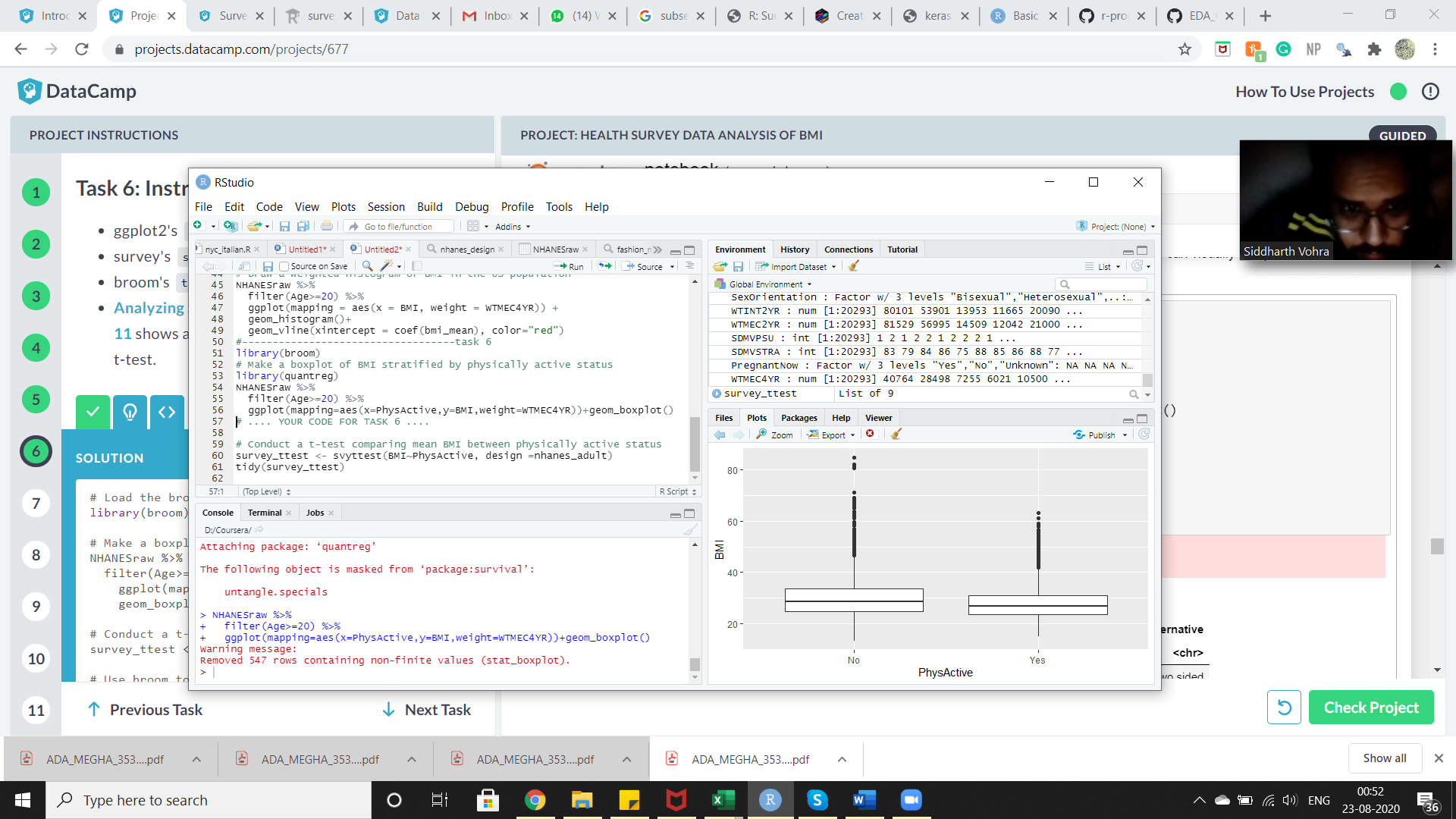


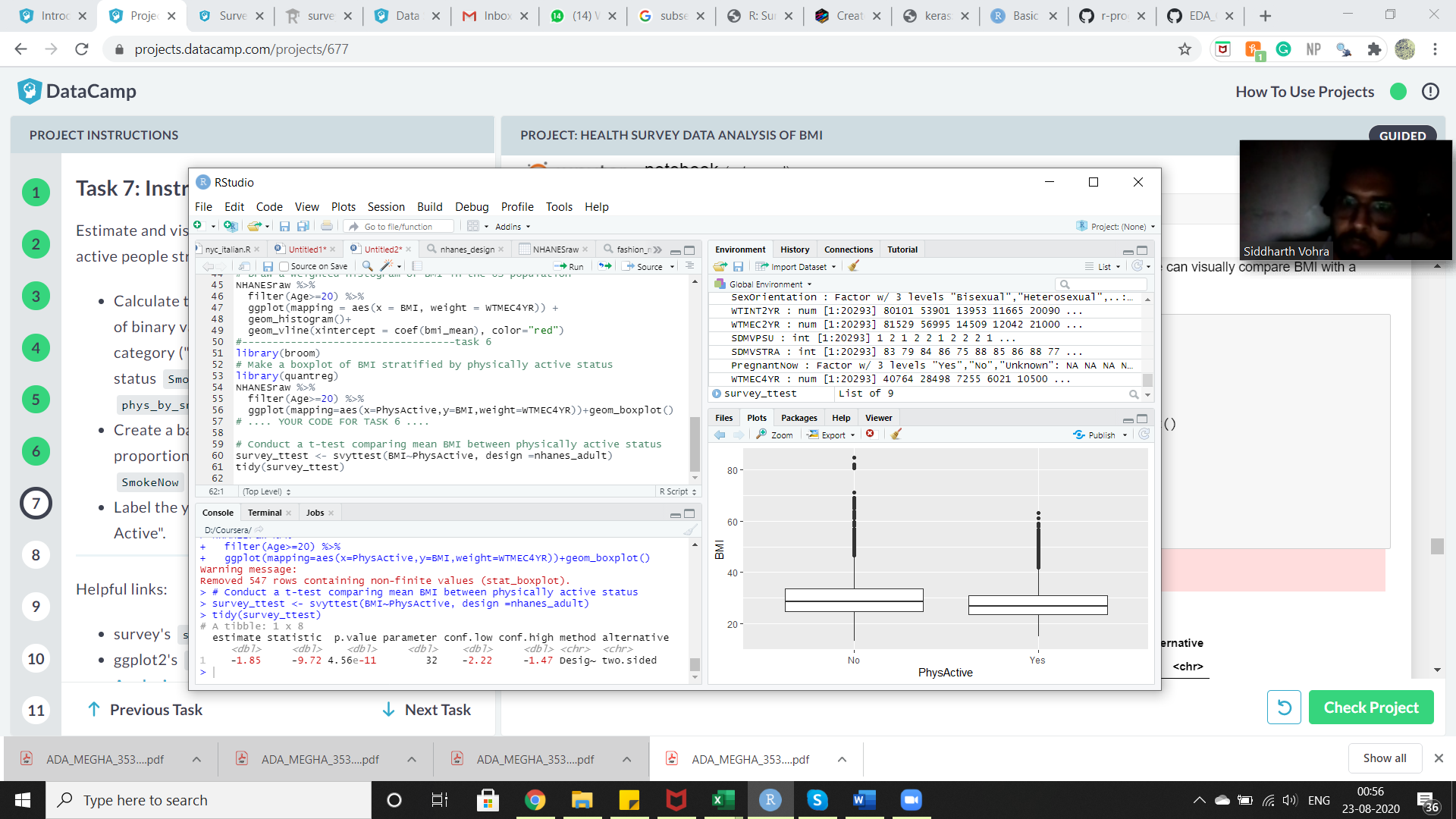


#---task 6

## 6. Is BMI lower in physically active people?

The distribution of BMI looks to be about what we might expect with most people under 40 kg/m2 and a slight positive skewness because a few people have much higher BMI. Now to the question of interest: does the distribution of BMI differ between people who are physically active versus those who are not physically active? We can visually compare BMI with a boxplot as well as formally test for a difference in mean BMI.

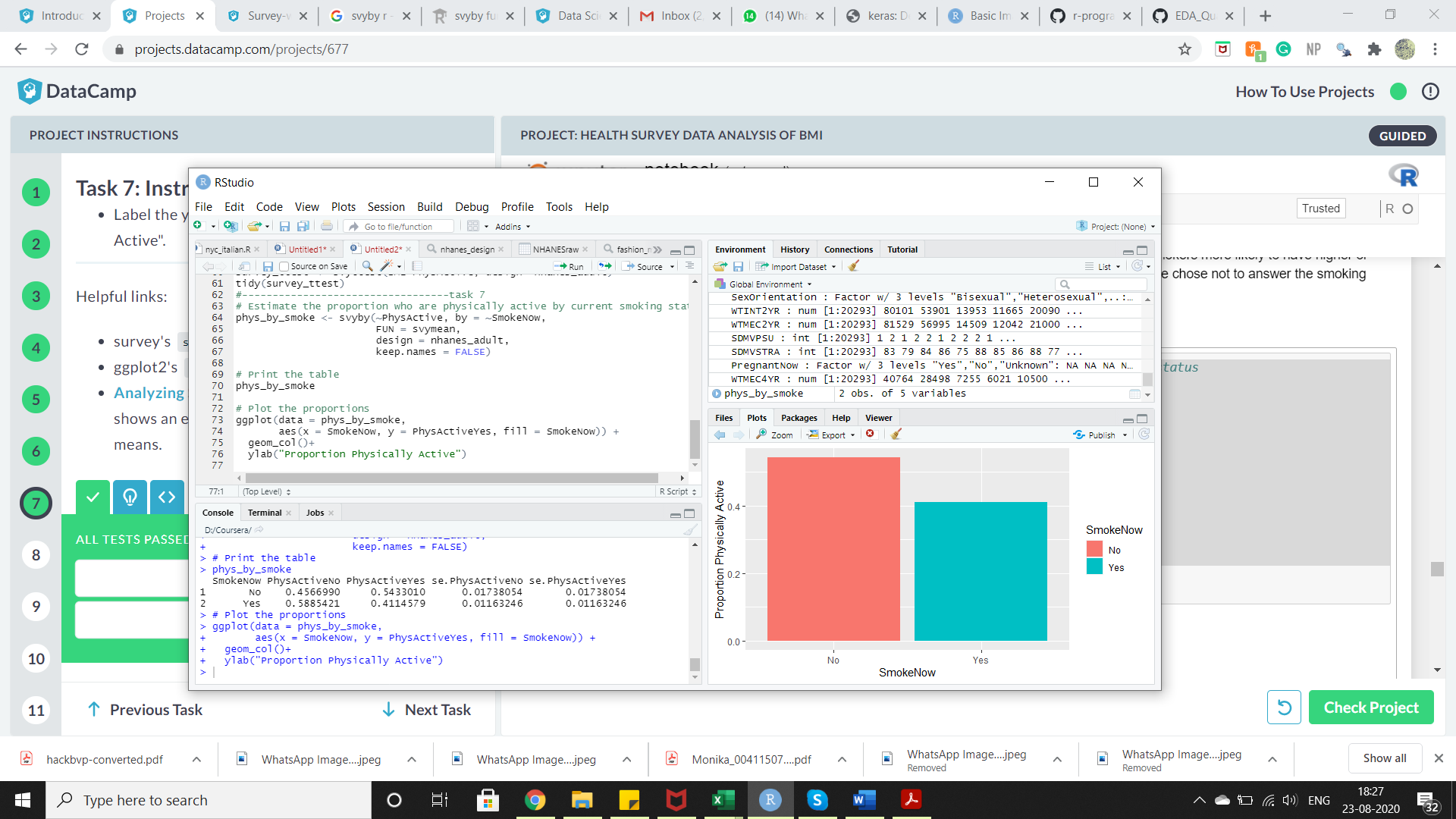




## 7. Could there be confounding by smoking? (part 1)

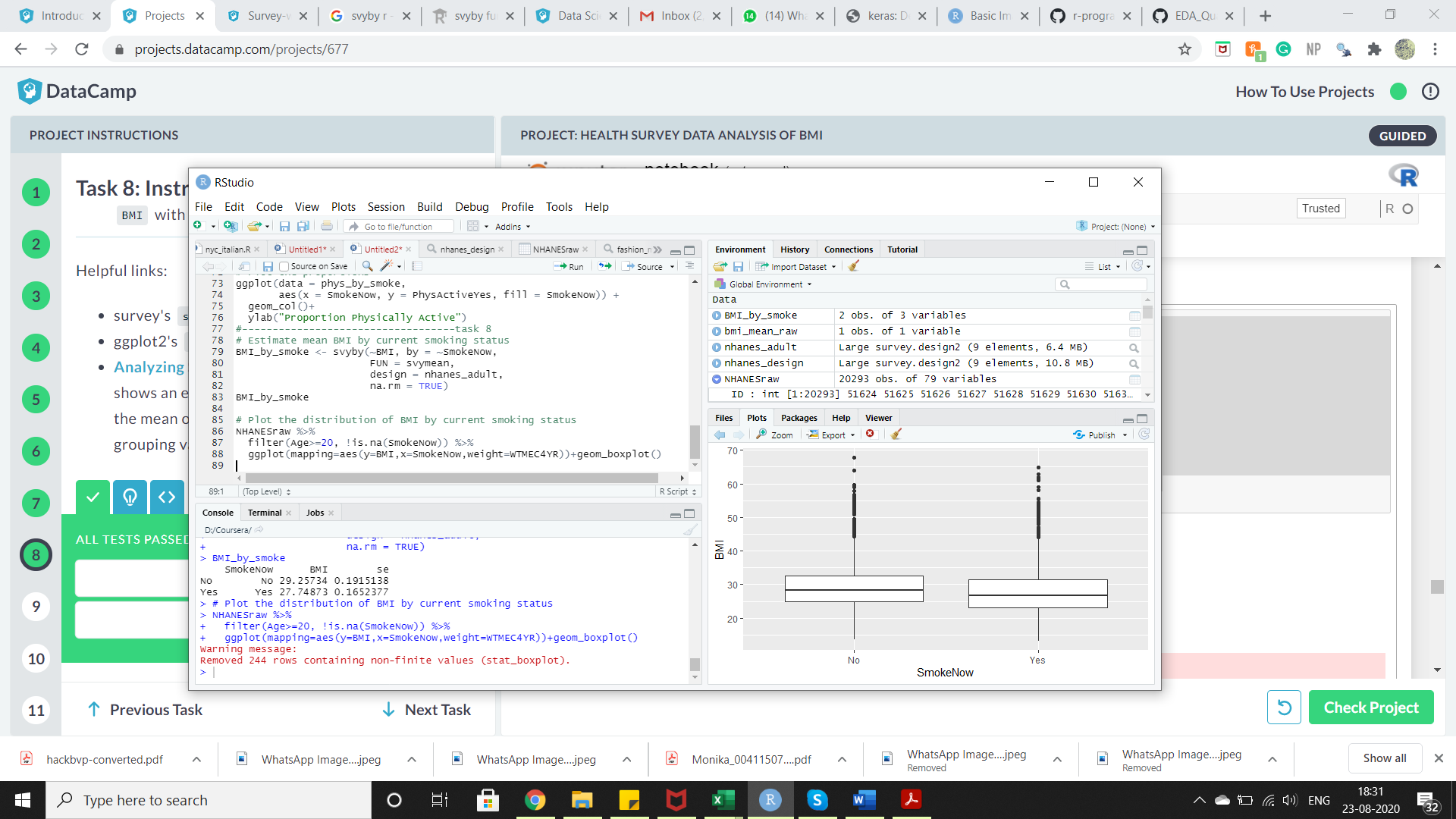
The relationship between physical activity and BMI is likely not so simple as "if you exercise you will lower your BMI." In fact, many other lifestyle or demographic variables could be confounding this relationship. One such variable could be smoking status. If someone smokes, is he or she more or less likely to be physically active? Are smokers more likely to have higher or lower BMI? We can examine these relationships in the survey data. Note that many people chose not to answer the smoking question, so we reduce our sample size when looking at this data.

First, let's look at the relationship between smoking and physical activity.



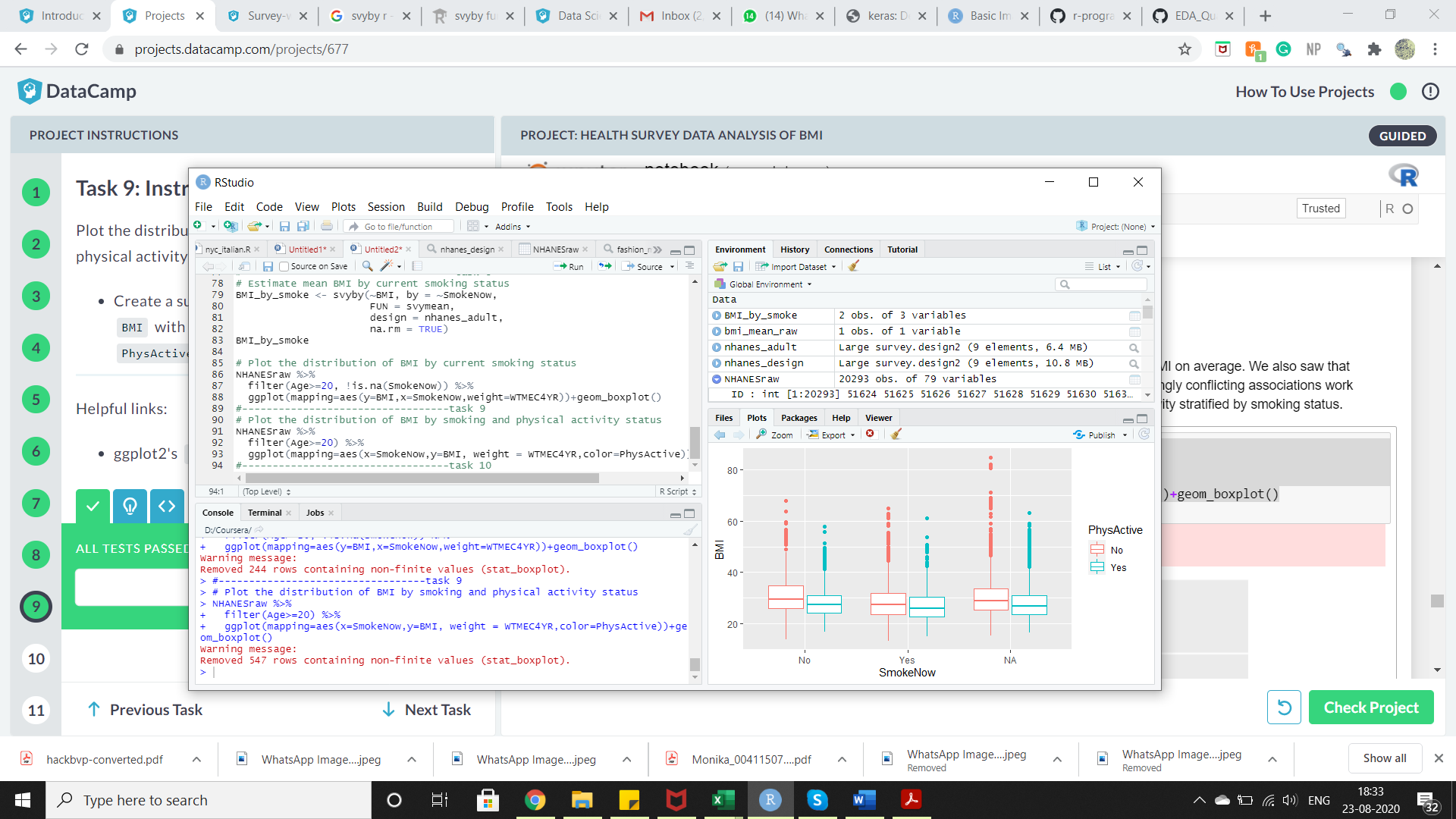
## 8. Could there be confounding by smoking? (part 2)

Now let's examine the relationship between smoking with BMI.



## 9. Add smoking in the mix

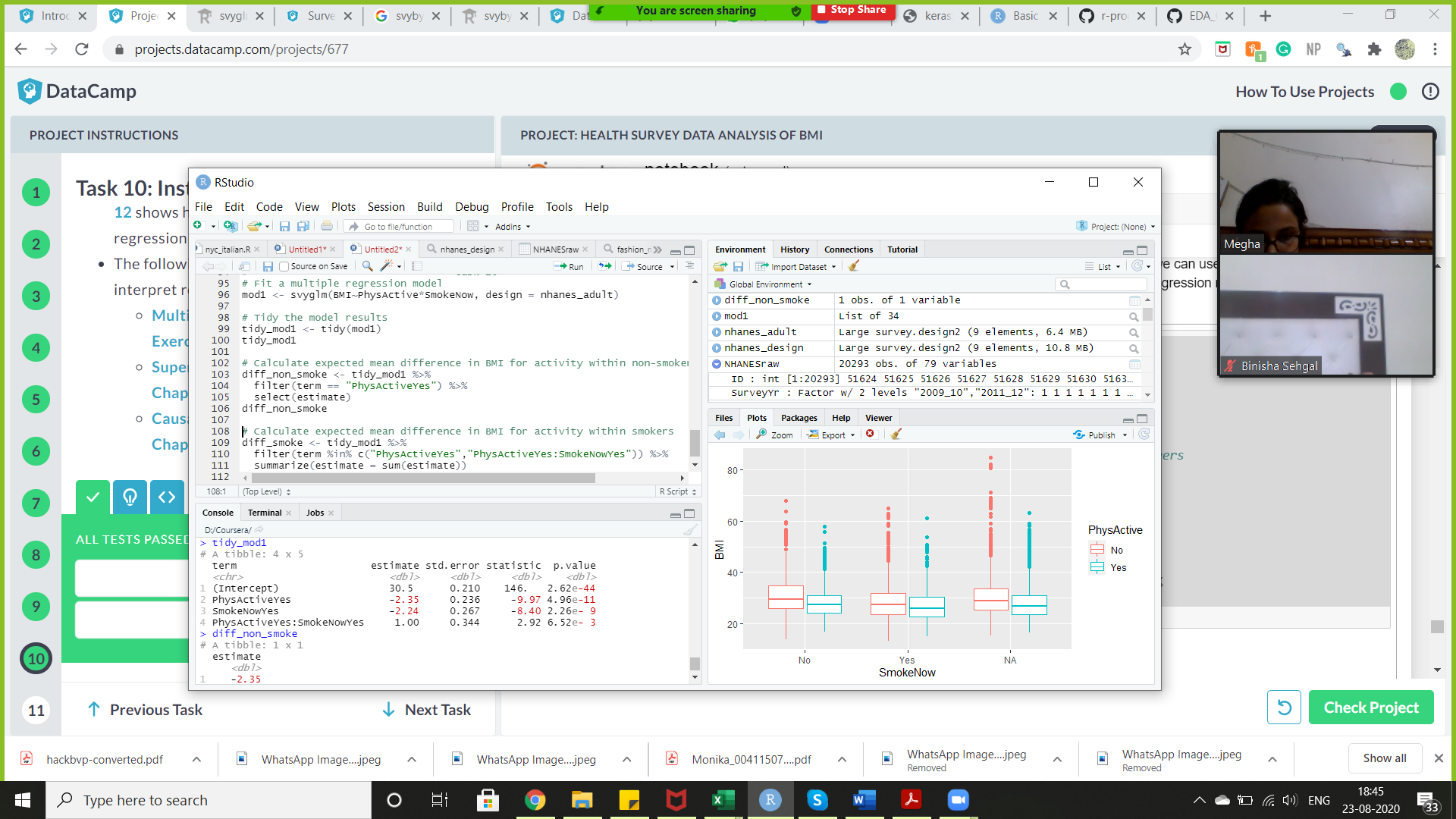
We saw that people who smoke are less likely to be physically active and have a lower BMI on average. We also saw that people who are not physically active have a higher BMI on average. How do these seemingly conflicting associations work together? To get a better sense of what's going on, we can compare BMI by physical activity stratified by smoking status.



## 10. Incorporate possible confounding in the model

In the above plot, we see that people who are physically active tend to have lower BMI no matter their smoking status, and this is true even if they didn't answer the question. However, we also see that smokers have lower BMI in general. Also, looking closely we see the difference in BMI comparing physically active people to non-physically active people is slightly smaller in smokers than in non-smokers.

Previously, we used a simple t-test to compare mean BMI in physically active people and non-physically active people. In order to adjust for smoking status, as well as other possible confounders or predictors of BMI, we can use a linear regression model with multiple independent variables. When using survey data, we use a weighted linear regression method which is a special case of generalized linear models (GLMs).

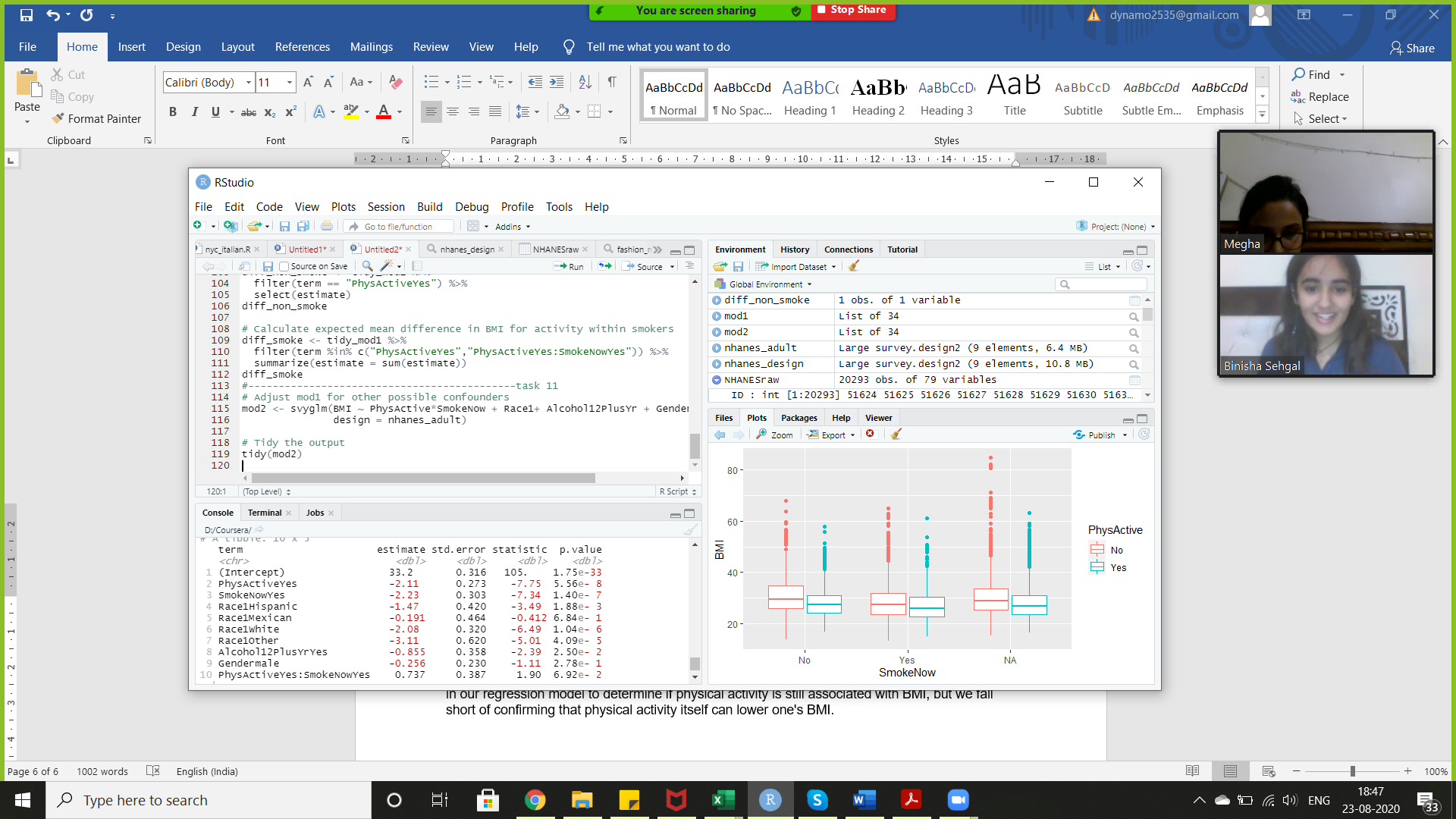


## 11. What does it all mean?

We fit a linear regression model where the association of physical activity with BMI could vary by smoking status. The interaction between physical activity and smoking has a small p-value, which suggests the association does vary by smoking status. The difference between physically active and non-physically active people is larger in magnitude in the non-smoker population.

We should check the [model fit and technical assumptions](https://campus.datacamp.com/courses/inference-for-linear-regression/technical-conditions-in-linear-regression?ex=1) of our regression model. Then, we can conclude that physically active people tend to have lower BMI, as do smokers. Although they have similar effect sizes, we probably wouldn't want to recommend smoking along with exercise!

In order to determine whether physical activity causes lower BMI, we would need to use causal inference methods or a randomized control study. We can adjust for other possible confounders in our regression model to determine if physical activity is still associated with BMI, but we fall short of confirming that physical activity itself can lower one's BMI.



References

Appendix

Task 1:

rm(list=ls())

library(NHANES)

library(dplyr)

glimpse(NHANESraw)

Task 2:

library(ggplot2)

NHANESraw<-NHANESraw%>%

mutate(WTMEC4YR=WTMEC2YR/2)

NHANESraw%>%summarize(sum(WTMEC4YR))

ggplot(NHANESraw,aes(y=WTMEC4YR,x=Race1))+geom\_boxplot()

Task 3:

library(survey)

# Specify the survey design

nhanes\_design <- svydesign(

data = NHANESraw,

strata = ~SDMVSTRA,

id = ~SDMVPSU,

nest = TRUE,

weights = ~WTMEC4YR)

#nested=true because id is nested in strata variable

# Print a summary of this design

summary(nhanes\_design)

Task 4:

# Select adults of Age >= 20 with subset

nhanes\_adult <- subset(nhanes\_design,Age>=20 )

# Print a summary of this subset

# .... YOUR CODE FOR TASK 4 ....

summary(nhanes\_adult)

# Compare the number of observations in the full data to the adult data

nrow(nhanes\_design)

nrow(nhanes\_adult)

Task 5:

# Calculate the mean BMI in NHANESraw

bmi\_mean\_raw <- NHANESraw %>%

filter(Age >= 20) %>%

summarize(mean(BMI, na.rm = TRUE))

bmi\_mean\_raw

# Calculate the survey-weighted mean BMI of US adults

bmi\_mean <- svymean(~BMI, design = nhanes\_adult, na.rm = TRUE)

bmi\_mean

# Draw a weighted histogram of BMI in the US population

NHANESraw %>%

filter(Age>=20) %>%

ggplot(mapping = aes(x = BMI, weight = WTMEC4YR)) +

geom\_histogram()+

geom\_vline(xintercept = coef(bmi\_mean), color="red")

Task 6:

library(broom)

# Make a boxplot of BMI stratified by physically active status

library(quantreg)

NHANESraw %>%

filter(Age>=20) %>%

ggplot(mapping=aes(x=PhysActive,y=BMI,weight=WTMEC4YR))+geom\_boxplot()

# .... YOUR CODE FOR TASK 6 ....

# Conduct a t-test comparing mean BMI between physically active status

survey\_ttest <- svyttest(BMI~PhysActive, design =nhanes\_adult)

tidy(survey\_ttest)

Task 7:

# Estimate the proportion who are physically active by current smoking status

phys\_by\_smoke <- svyby(~PhysActive, by = ~SmokeNow,

FUN = svymean,

design = nhanes\_adult,

keep.names = FALSE)

# Print the table

phys\_by\_smoke

# Plot the proportions

ggplot(data = phys\_by\_smoke,

aes(x = SmokeNow, y = PhysActiveYes, fill = SmokeNow)) +

geom\_col()+

ylab("Proportion Physically Active")

Task 8:

# Estimate mean BMI by current smoking status

BMI\_by\_smoke <- svyby(~BMI, by = ~SmokeNow,

FUN = svymean,

design = nhanes\_adult,

na.rm = TRUE)

BMI\_by\_smoke

# Plot the distribution of BMI by current smoking status

NHANESraw %>%

filter(Age>=20, !is.na(SmokeNow)) %>%

ggplot(mapping=aes(y=BMI,x=SmokeNow,weight=WTMEC4YR))+geom\_boxplot()

Task 9:

Plot the distribution of BMI by smoking and physical activity status

NHANESraw %>%

filter(Age>=20) %>%

ggplot(mapping=aes(x=SmokeNow,y=BMI, weight = WTMEC4YR,color=PhysActive))+geom\_boxplot()

Task 10:

# Fit a multiple regression model

mod1 <- svyglm(BMI~PhysActive\*SmokeNow, design = nhanes\_adult)

# Tidy the model results

tidy\_mod1 <- tidy(mod1)

tidy\_mod1

# Calculate expected mean difference in BMI for activity within non-smokers

diff\_non\_smoke <- tidy\_mod1 %>%

filter(term == "PhysActiveYes") %>%

select(estimate)

diff\_non\_smoke

# Calculate expected mean difference in BMI for activity within smokers

diff\_smoke <- tidy\_mod1 %>%

filter(term %in% c("PhysActiveYes","PhysActiveYes:SmokeNowYes")) %>%

summarize(estimate = sum(estimate))

diff\_smoke

Task 11:

# Adjust mod1 for other possible confounders

mod2 <- svyglm(BMI ~ PhysActive\*SmokeNow + Race1+ Alcohol12PlusYr + Gender,

design = nhanes\_adult)

# Tidy the output

tidy(mod2)