**Final project Data Analysis guidelines**

**Example used (Oct 2013, Pew Research Center, Higher Education, Gender and Work)**

**Step 1: Construct the data set based on survey results**

**Step 2: Import the constructed data set into R.**

**Step 3: Data Analysis**

**1. Bar Plot, Hypothesis Testing for One Sample Proportion**

**2. Hypothesis Testing for Two Sample Proportions**

**3. Hypothesis Testing for One Sample Mean and Two Sample Means**

**4. Regression**

**Step 1: Construct the data set based on survey results**

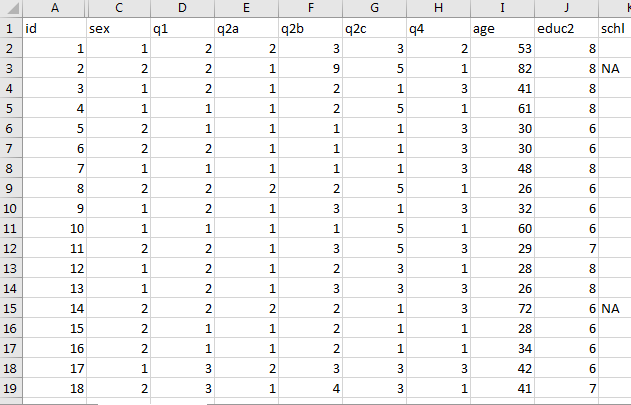
* Label the options for each question in your survey using numbers. For example, for sex variable, we label male as 1 and female as 2.

**SEX [ENTER RESPONDENT'S SEX:]**

**1 Male**

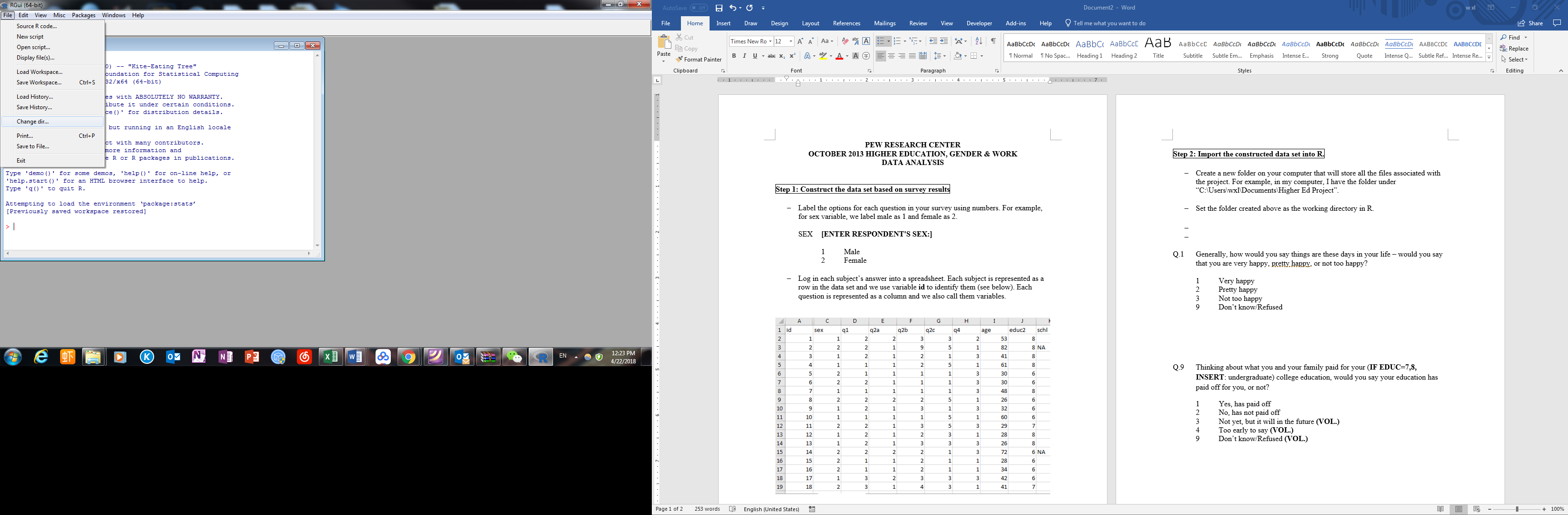
**2 Female**

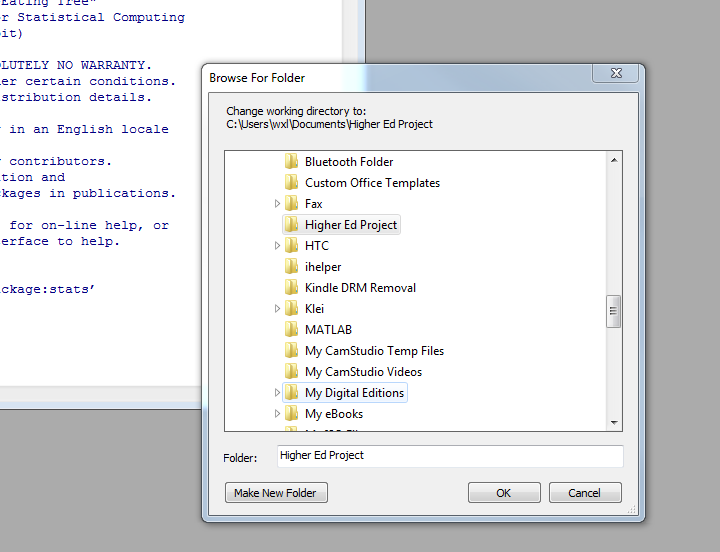
* Log in each subject’s answer into a spreadsheet. Each subject is represented as a row in the data set and we use variable **id** to identify them (see below). Each question is represented as a column and we also call them variables.



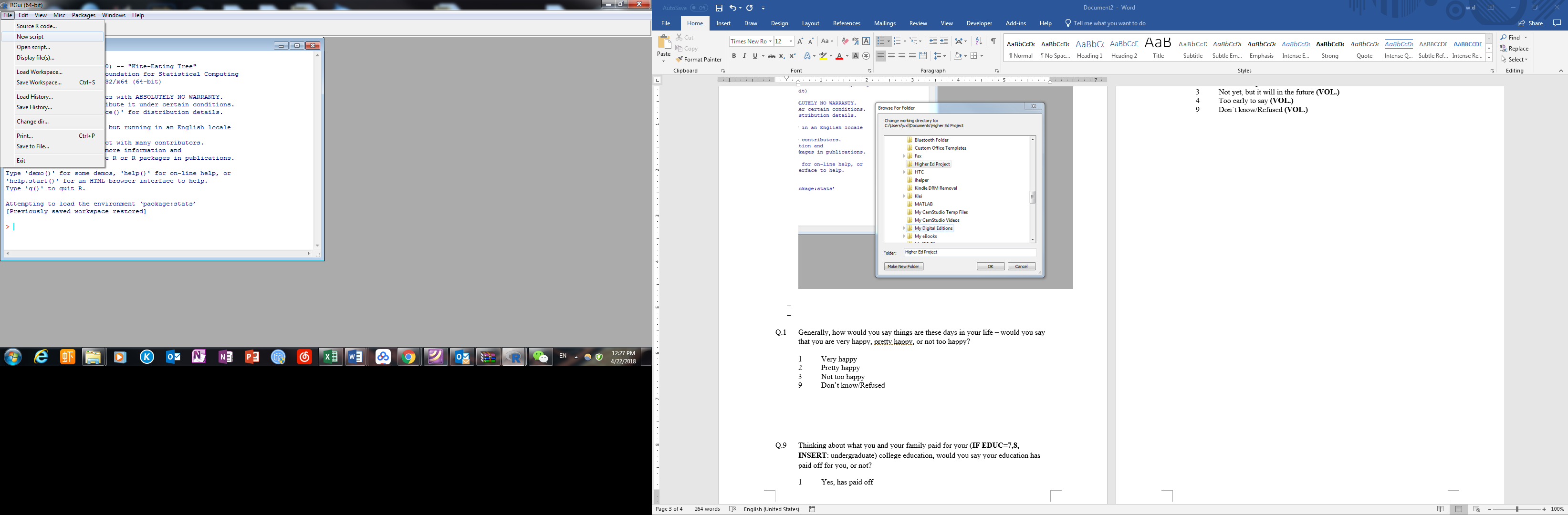
**Step 2: Import the constructed data set into R.**

* Create a new folder on your computer that will store all the files associated with the project. For example, in my computer, I have the folder under “C:\Users\wxl\Documents\Higher Ed Project”.
* Set the folder created above as the working directory in R. In Windows, click “File” in R and then choose “Change dir”.



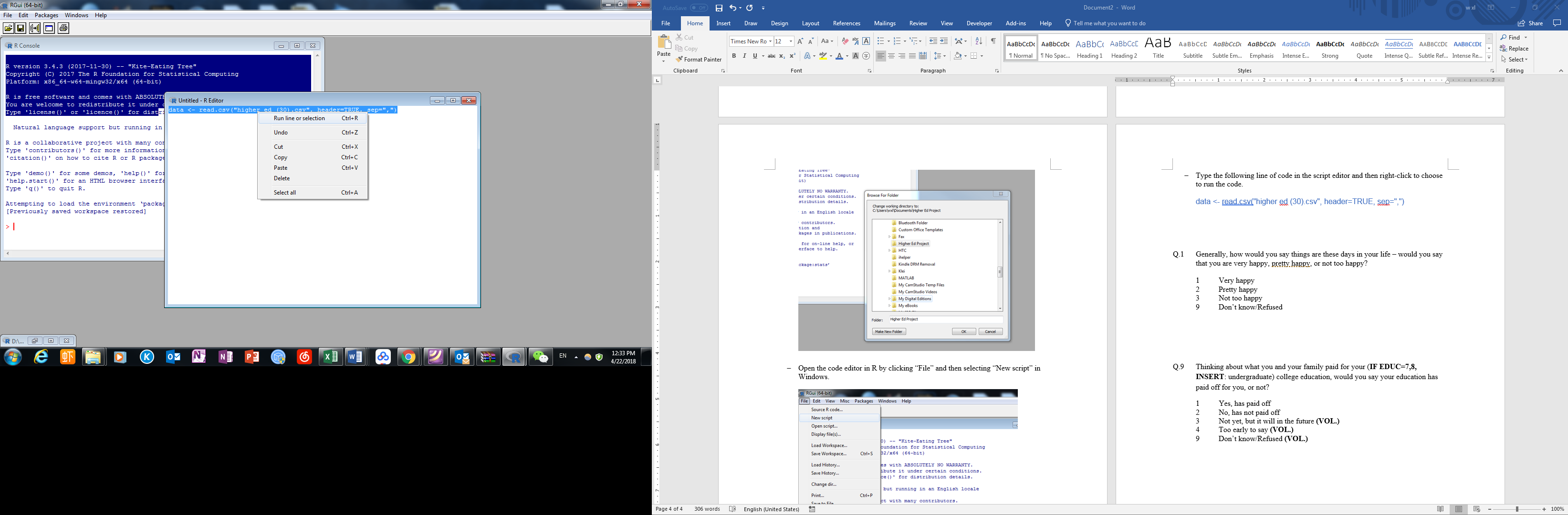


* Open the code editor in R by clicking “File” and then selecting “New script” in Windows.

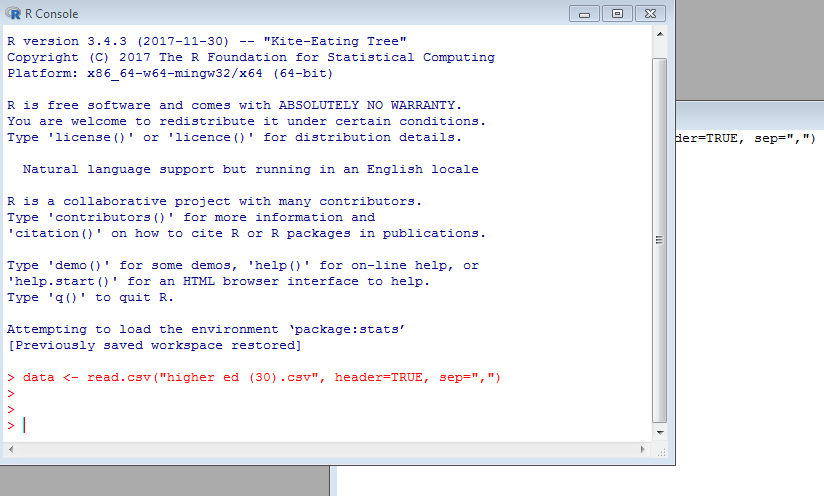


* Type the following line of code in the script editor and then right-click to choose to run the code. Here we name out dataset as “data”.

data <- read.csv("higher ed (30).csv", header=TRUE, sep=",")



* Then data should be imported if no error message come in the R console.



**Step 3: Data Analysis**

In the following, we will go through a few examples to see how to analyze data.

**1. Bar Plot, Hypothesis Testing for One Sample Proportion**

**Q.1 Generally, how would you say things are these days in your life – would you say that you are very happy, pretty happy, or not too happy?**

**1 Very happy**

**2 Pretty happy**

**3 Not too happy**

**9 Don’t know/Refused**

Given this happiness question, we want to know how many subjects choose each option. We use the following code.

happiness <- table(data$q1)

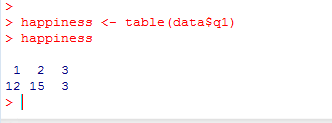
*#use the table function to find the frequency distribution of each option*

*#name the table as happiness.*

*#“data$q1” means that the frequency table is for variable “q1” in dataset “data”*

happiness

*#call the object "happiness and display in R*



12 out 30 people are very happy, 15 are pretty happy and 3 are not too happy.

We can also create a bar plot to visualize this.

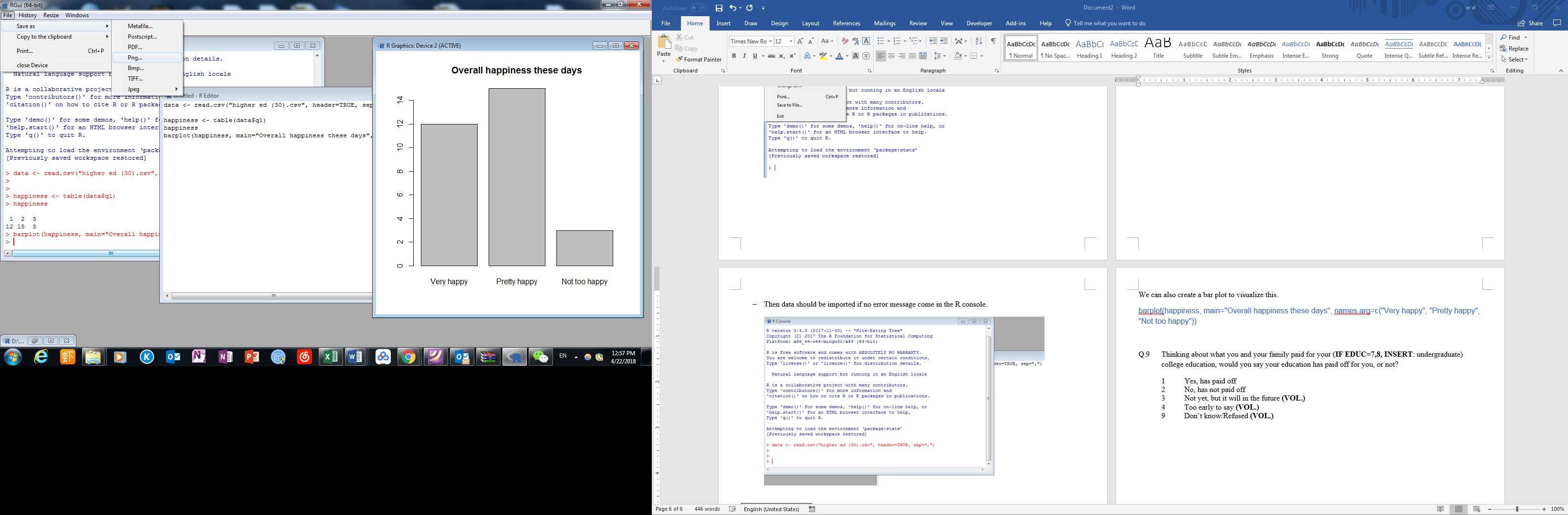
barplot(happiness, main="Overall happiness these days", names.arg=c("Very happy", "Pretty happy", "Not too happy"))

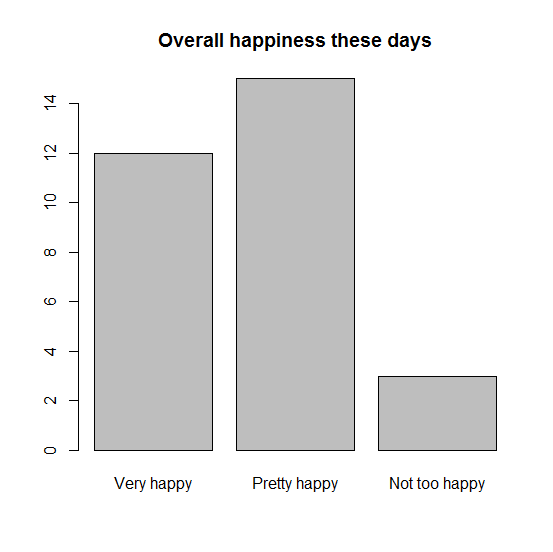
*#main specifies the title*

*#names.arg specifies the labels for each bar*

*#c means combining the elements as a vector*

After the graph is created, select the graph window and click on “File” to save the graph as a picture or copy to the clipboard to paste here.





We see that the sample proportion of very happy individuals is and we can use this to test the hypothesis that the proportion of very happy individuals in the population is . We use 95% confidence level, or in other words, the significance level .

prop.test(x=12,n=30,p=0.5,correct=F, conf.level=0.99, alternative="two.sided")

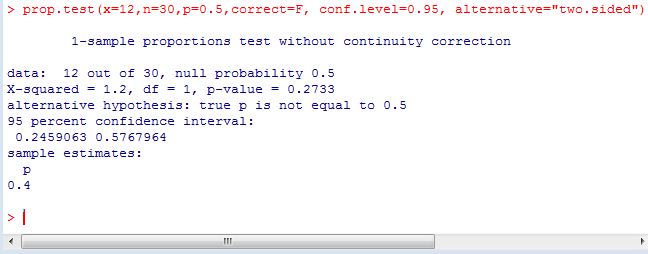
*#x is the number of individuals with the desired characristic*

*#n is the sample size*

*#correct=FALSE means to exclude the continuity correction*

*#conf.level specifies the confidence level*

*#alternative specifies the kinds of test, can be "two.tailed", "less" or "greater"*



Since the p-value is 0.2733, which is greater than 0.05, so we don’t reject the null hypothesis and the sample suggests that the population proportion is not significantly different from 0.5.

The result also shows that the confidence interval for the sample proportion is (0.2459, 0.5768). In other words, we are 95% confident that the proportion of very happy individuals in the population is from 25% to 58%.

Similarly, we can obtain graphs for other questions.

**Q.9 Thinking about what you and your family paid for your (IF EDUC=7,8, INSERT: undergraduate) college education, would you say your education has paid off for you, or not?**

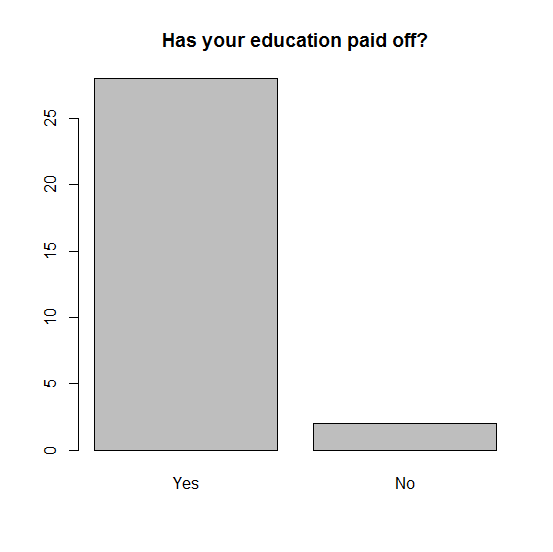
**1 Yes, has paid off**

**2 No, has not paid off**

**3 Not yet, but it will in the future**

**4 Too early to say**

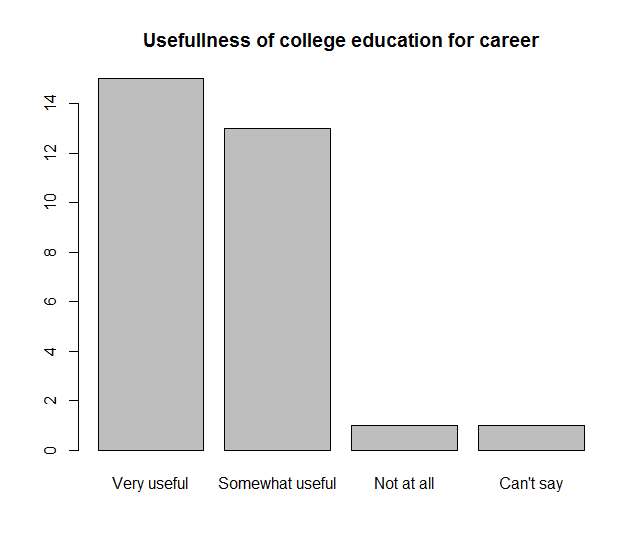
**9 Don’t know/Refused**



**Q.20 How useful was your college education in preparing you for a job or career? Would you say very useful, somewhat useful, not too useful, or not useful at all?**

1. **Very useful**
2. **Somewhat useful**
3. **Not too useful**
4. **Not at all useful**
5. **Can’t say, still in graduate school/college**

**9 Don’t know/Refused**



**Q.22 Still thinking back to when you were an undergraduate college student, do you think any of the following things would have better prepared you to get the kind of job you wanted, or not?**

**a. Choosing a different major**

**b. Gaining more work experience**

**c. Starting to look for work sooner**

**d. Studying harder**

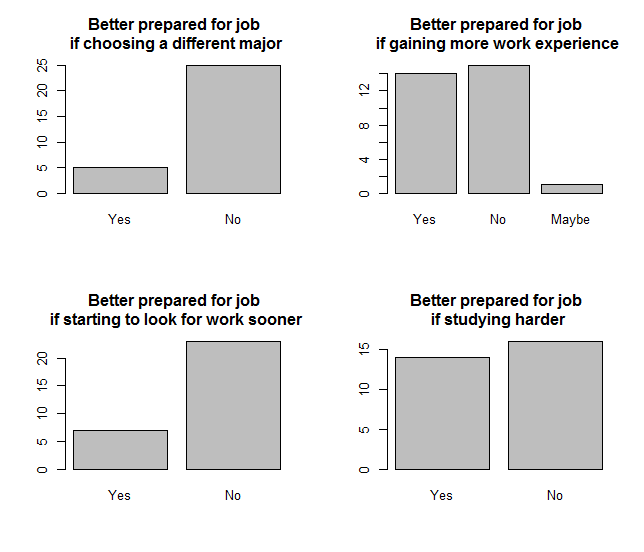
**RESPONSE CATEGORIES:**

**1 Yes**

**2 No**

**3 Maybe**

**9 Don’t know/Refused**



**2. Hypothesis Testing for Two Sample Proportions**

Besides, we can look at the relationship between two variables. First we look at gender and happiness by listing the distribution of happiness against gender.

**SEX**

**1 Male**

**2 Female**

**Q.1 Generally, how would you say things are these days in your life – would you say that you are very happy, pretty happy, or not too happy?**

**1 Very happy**

**2 Pretty happy**

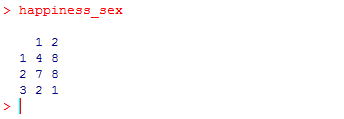
**3 Not too happy**

**9 Don’t know/Refused**

happiness\_sex <- table(data$q1, data$sex)

*#having both q1 and sex in the table function, we produce a cross table of two*

happiness\_sex



We can see that for male subjects, 4 are very happy, 7 pretty happy and 3 not too happy. For female, 8 are very happy, 8 pretty happy and 1 unhappy.

barplot(happiness\_sex, beside=T, main="Overall happiness these days", names.arg=c("Male", "Female"), legend.text=T, args.legend=list(x=8, y=-1,legend=c("Very happy", "Pretty happy", "Not too happy"),ncol=3))

*#beside=TRUE specifies that the bars for male and female are beside to each other*

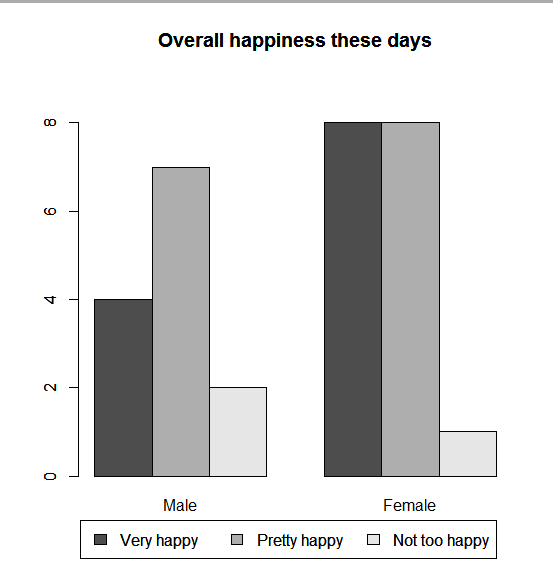
*#legend.text=TRUE means we add a legend to the graph*

*#args.legend specifies the parameters of the legend*

*#x= and y= indicate the position of the legend on the x-y plane*

*#legend= specifies the content of the legend*

*#ncol specifies the number of columns in the legend*



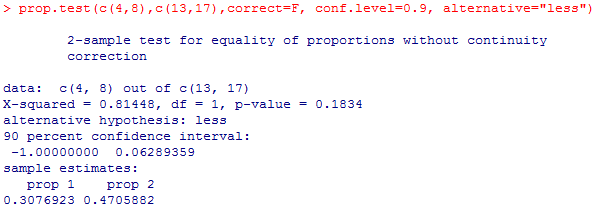
One question about gender difference is whether female is proportionally happier than male. We can conduct a two proportions hypothesis testing on this. We set confidence level to be 90%, or in other words, the significance level .

prop.test(c(4,8),c(13,17),correct=F, conf.level=0.9, alternative="less")

*#4,8 are x1 and x2, which are the number of individuals with the desired characteristc in each group*

*#13, 17 are n1 and n2, which are sample sizes for each group*

*#we are testing men has less proportion than women, so alternative="less"*



Here p value is 0.1834, which is greater than 0.1, so we cannot reject the null hypothesis, and the sample evidence does not suggest that female is proportionally happier than male. Probably this is because our sample size is so small that we cannot reject the null hypothesis.

We can also look at the effects of marriage on happiness, or, if married people are happier than singles.

MARITAL Are you currently married, living with a partner, divorced, separated, widowed, or have you never been married? **(IF R SAYS “SINGLE,” PROBE TO DETERMINE WHICH CATEGORY IS APPROPRIATE)**

1 Married

2 Living with a partner

3 Divorced

4 Separated

5 Widowed

6 Never been married

9 Don't know/Refused **(VOL.)**

Since the marital variable has many groups, we may simplify it first by creating a binary variable partner, that is, classifying married (1) and living with a partner (2) as “partnered” and the rest as “not partnered”.

data$partner <- ifelse(data$marital==1 | data$marital==2, 1, 2)

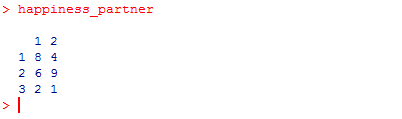
*#create a new variable named partner in the dataset*

*#partner = 1 when marital==1 or 2*

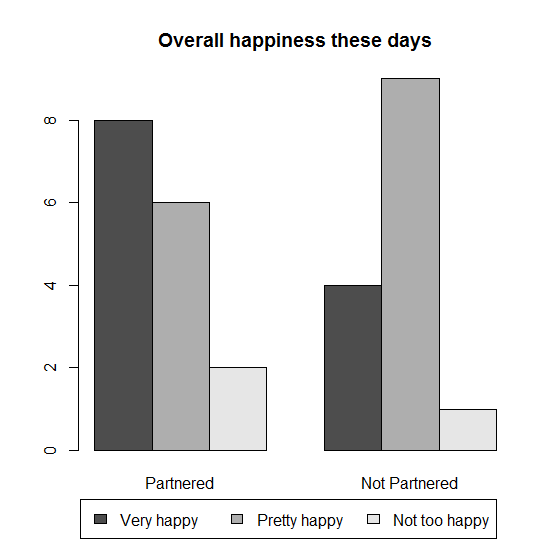
*#partner = 2 when marital== other value that is not 1 or 2*

*#"data$marital==1 | data$marital==2" is the condition, if it's true, we define partner as 1, if not true (else), partner as 2*

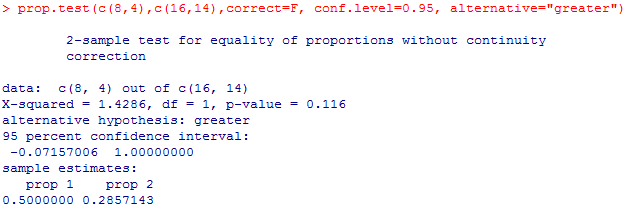
happiness\_partner <- table(data$q1, data$partner)



barplot(happiness\_partner, beside=T, main="Overall happiness these days", names.arg=c("Partnered", "Not Partnered"),legend=c("Very happy", "Pretty happy", "Not too happy"),legend.text=T, args.legend=list(x=8, y=-1.2 ,legend=c("Very happy", "Pretty happy", "Not too happy"),ncol=3))



prop.test(c(8,4),c(16,14),correct=F, conf.level=0.95, alternative="greater")



Similarly, hypothesis testing does suggest that there is a significance difference between partnered and not partnered in terms of happiness, but this might be because of small sample size.

**3. Hypothesis Testing for One Sample Mean and Two Sample Means**

**Q.30 There are many things people value in a job. How important is each of these things to you personally?**

**a. Having a high-paying job**

**b. Having job security**

**c. Being able to take time off for family or child care needs**

**d. Having a job you enjoy doing**

**e. Having a job that offers good benefits**

**f. Having opportunities for promotions or advancement**

**g. Having a job that helps society**

**RESPONSE CATEGORIES:**

**1 Extremely important**

**2 Very important**

**3 Somewhat important**

**4 Not too important**

**5 Not at all important**

**6 Does not apply**

**9 Don’t know/Refused**

In this question, people give their evaluation of the importance of a certain job characteristic and we look at how important it is for people to have a high-paying job first. Since the importance is on a 1 through 5 scale, we can calculate the average importance. The middle is 3, meaning that it’s somewhat important and somewhat unimportant.

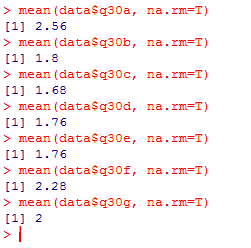
mean(data$q30a, na.rm=T)

# since we have missing values (shown as NA in the data), na.rm=TRUE means to remove them



We see that the mean is 2.56, so it seems important, but not very important.

But which one people think is the most important one?



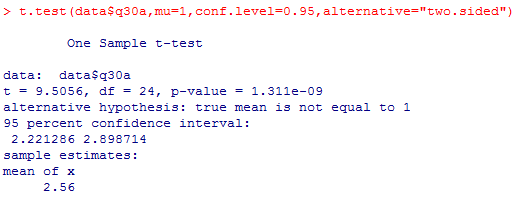
By comparing all the means, it seems that being able to take time off for family or child care needs, having a job you enjoy doing and having a job that offers good benefits rank the top 3 important things.

**Hypothesis Testing for One Sample Mean**

We can also test if any of the mean is significantly different from 1 which is “extremely important”. We use high-paying job as an example. The test is to see if the statement that “people in general think high-paying job is extremely important” is true or not.

t.test(data$q30a,mu=1,conf.level=0.95,alternative="two.sided")

#mu specifies that null hypothesis



Since p-value is very small, we can reject the null hypothesis and hence conclude that the sample indicates that people in general don’t think high-paying job is extremely important.

**Hypothesis Testing for Two Sample Means (Dependent Samples)**

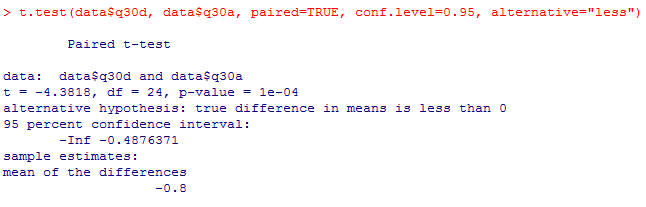
We can compare if any of the two means are the same or not. For example, we can test to see if people in general think a job they enjoy doing is more important than high-paying job.

(smaller number means more important)

Since each subject answer both questions, we have dependent samples here.

t.test(data$q30d, data$q30a, paired=TRUE, conf.level=0.95, alternative="less")

#paired=TRUE means dependent samples, paired=FALSE means independent samples



Because p-value is very small, we can reject the null hypothesis and hence conclude that the sample indicates that people in general think a job they enjoy doing is more important than high-paying job.

**Hypothesis Testing for Two Sample Means (Independent Samples)**

Another test that can be carried out is to test if there is any difference between male and female in terms of the importance they perceive to have a high-paying job.

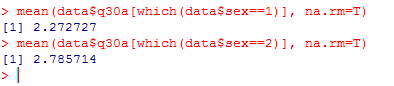
We first calculate the sample mean importance of having a high-paying job for both genders.

mean(data$q30a[which(data$sex==1)], na.rm=T)

#the brackets [ ] are used to subset the data vector

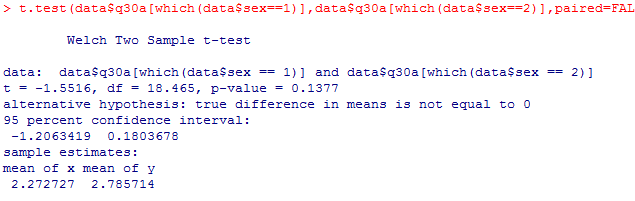
#which() returns the values that satisfy the condition specified in the parenthesis

mean(data$q30a[which(data$sex==2)], na.rm=T)



It seems that male individuals value high-paying job more than female since the male group has a smaller mean, and we test to see if this can be generalized to the population.

t.test(data$q30a[which(data$sex==1)],data$q30a[which(data$sex==2)],paired=FALSE,conf.level=0.95,alternative="two.sided")



The p-value is greater than 0.05, so we cannot reject the null at 95% confidence level and hence there is not statistical evidence suggesting that male individuals value high-paying job more than female.

**4. Regression**

We can use regression analysis to see if there is any relation between two variables, say, the importance of a high-paying job and age. In other words, do elder people value less about a high-paying job? If true, we may expect a positive relation since the higher the age the less important is a high-paying job (larger numbers in terms of the numeric importance value).

First, we obtain the correlation coefficient.



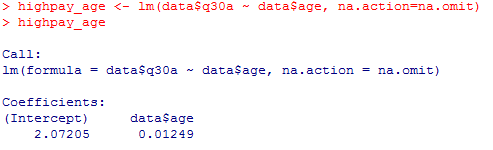
The correlation coefficient is quite small, so we doubt if there is a strong correlation between the two variables.

We carry out the regression analysis using the following code.

highpay\_age <- lm(data$q30a ~ data$age, na.action=na.omit)

#we create a regression model called highpay\_age

#lm means linear model in which we specify the dependent variable first, followed by ~ and independent variables



The slope is 0.01249, in fact, a quite small number. So we conclude that there not strong relation between age and the importance of a high-paying job.