**Week 1 Quiz** [**Help Center**](https://accounts.coursera.org/i/zendesk/courserahelp?return_to=https://learner.coursera.help/hc/articles/201523125-Quizzes)

**Warning:** The hard deadline has passed. You can attempt it, but **you will not get credit for it**. You are welcome to try it as a learning exercise.

Haut du formulaire

In accordance with the Coursera Honor Code, I (Debernardi Nicolas) certify that the answers here are my own work.

**Question 1**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>   
  
and load the data into R. The code book, describing the variable names is here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf>  
  
How many properties are worth $1,000,000 or more?

25

47

2076

53

**Question 2**

Use the data you loaded from Question 1. Consider the variable FES in the code book. Which of the "tidy data" principles does this variable violate?

Tidy data has one variable per column.

Numeric values in tidy data can not represent categories.

Each variable in a tidy data set has been transformed to be interpretable.

Each tidy data table contains information about only one type of observation.

**Question 3**

Download the Excel spreadsheet on Natural Gas Aquisition Program here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx>  
  
Read rows 18-23 and columns 7-15 into R and assign the result to a variable called:

dat

What is the value of:

sum(dat$Zip\*dat$Ext,na.rm=T)

(original data source: <http://catalog.data.gov/dataset/natural-gas-acquisition-program>)

184585

36534720

NA

33544718

**Question 4**

Read the XML data on Baltimore restaurants from here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml>  
  
How many restaurants have zipcode 21231?

28

181

127

100

**Question 5**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv>   
  
using the fread() command load the data into an R object

DT

Which of the following is the fastest way to calculate the average value of the variable

pwgtp15

broken down by sex using the data.table package?

tapply(DT$pwgtp15,DT$SEX,mean)

mean(DT[DT$SEX==1,]$pwgtp15); mean(DT[DT$SEX==2,]$pwgtp15)

DT[,mean(pwgtp15),by=SEX]

rowMeans(DT)[DT$SEX==1]; rowMeans(DT)[DT$SEX==2]

sapply(split(DT$pwgtp15,DT$SEX),mean)

mean(DT$pwgtp15,by=DT$SEX)

In accordance with the Coursera Honor Code, I (Debernardi Nicolas) certify that the answers here are my own work.

Bas du formulaire

**Week 2 Quiz** [**Help Center**](https://accounts.coursera.org/i/zendesk/courserahelp?return_to=https://learner.coursera.help/hc/articles/201523125-Quizzes)

**Warning:** The hard deadline has passed. You can attempt it, but **you will not get credit for it**. You are welcome to try it as a learning exercise.

Haut du formulaire

In accordance with the Coursera Honor Code, I (Debernardi Nicolas) certify that the answers here are my own work.

**Question 1**

Register an application with the Github API here https://github.com/settings/applications. Access the API to get information on your instructors repositories (hint: this is the url you want "https://api.github.com/users/jtleek/repos"). Use this data to find the time that the datasharing repo was created. What time was it created? This tutorial may be useful (https://github.com/hadley/httr/blob/master/demo/oauth2-github.r). You may also need to run the code in the base R package and not R studio.

2013-11-07T13:25:07Z

2014-03-05T16:11:46Z

2012-06-20T18:39:06Z

2014-02-06T16:13:11Z

**Question 2**

The sqldf package allows for execution of SQL commands on R data frames. We will use the sqldf package to practice the queries we might send with the dbSendQuery command in RMySQL. Download the American Community Survey data and load it into an R object called

acs

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv>   
  
Which of the following commands will select only the data for the probability weights pwgtp1 with ages less than 50?

sqldf("select \* from acs")

sqldf("select \* from acs where AGEP < 50")

sqldf("select pwgtp1 from acs where AGEP < 50")

sqldf("select \* from acs where AGEP < 50 and pwgtp1")

**Question 3**

Using the same data frame you created in the previous problem, what is the equivalent function to unique(acs$AGEP)

sqldf("select distinct AGEP from acs")

sqldf("select unique AGEP from acs")

sqldf("select AGEP where unique from acs")

sqldf("select distinct pwgtp1 from acs")

**Question 4**

How many characters are in the 10th, 20th, 30th and 100th lines of HTML from this page:   
  
http://biostat.jhsph.edu/~jleek/contact.html   
  
(Hint: the nchar() function in R may be helpful)

45 31 7 31

43 99 7 25

45 31 7 25

45 31 2 25

43 99 8 6

45 92 7 2

45 0 2 2

**Question 5**

Read this data set into R and report the sum of the numbers in the fourth of the nine columns.   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110.for>   
  
Original source of the data: <http://www.cpc.ncep.noaa.gov/data/indices/wksst8110.for>   
  
(Hint this is a fixed width file format)

28893.3

222243.1

32426.7

35824.9

36.5

101.83

**Week 3 Quiz** [**Help Center**](https://accounts.coursera.org/i/zendesk/courserahelp?return_to=https://learner.coursera.help/hc/articles/201523125-Quizzes)

**Warning:** The hard deadline has passed. You can attempt it, but **you will not get credit for it**. You are welcome to try it as a learning exercise.

Haut du formulaire

In accordance with the Coursera Honor Code, I (Debernardi Nicolas) certify that the answers here are my own work.

**Question 1**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>   
  
and load the data into R. The code book, describing the variable names is here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf>  
  
Create a logical vector that identifies the households on greater than 10 acres who sold more than $10,000 worth of agriculture products. Assign that logical vector to the variable agricultureLogical. Apply the which() function like this to identify the rows of the data frame where the logical vector is TRUE. which(agricultureLogical) What are the first 3 values that result?

25, 36, 45

403, 756, 798

125, 238,262

153 ,236, 388

**Question 2**

Using the jpeg package read in the following picture of your instructor into R   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fjeff.jpg>  
  
Use the parameter native=TRUE. What are the 30th and 80th quantiles of the resulting data? (some Linux systems may produce an answer 638 different for the 30th quantile)

-14191406 -10904118

-15259150 -10575416

-16776430 -15390165

10904118 -594524

**Question 3**

Load the Gross Domestic Product data for the 190 ranked countries in this data set:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv>   
  
Load the educational data from this data set:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv>   
  
Match the data based on the country shortcode. How many of the IDs match? Sort the data frame in descending order by GDP rank (so United States is last). What is the 13th country in the resulting data frame?   
  
Original data sources:   
<http://data.worldbank.org/data-catalog/GDP-ranking-table>   
<http://data.worldbank.org/data-catalog/ed-stats>

189 matches, 13th country is Spain

190 matches, 13th country is St. Kitts and Nevis

190 matches, 13th country is Spain

234 matches, 13th country is St. Kitts and Nevis

234 matches, 13th country is Spain

189 matches, 13th country is St. Kitts and Nevis

**Question 4**

What is the average GDP ranking for the "High income: OECD" and "High income: nonOECD" group?

30, 37

23, 45

32.96667, 91.91304

133.72973, 32.96667

23, 30

23.966667, 30.91304

**Question 5**

Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group. How many countries are Lower middle income but among the 38 nations with highest GDP?

1

18

5

0

**Week 4 Quiz \*\*Please Note: No Grace Period\*\*** [**Help Center**](https://accounts.coursera.org/i/zendesk/courserahelp?return_to=https://learner.coursera.help/hc/articles/201523125-Quizzes)

**Warning:** The hard deadline has passed. You can attempt it, but **you will not get credit for it**. You are welcome to try it as a learning exercise.

Haut du formulaire

In accordance with the Coursera Honor Code, I (Debernardi Nicolas) certify that the answers here are my own work.

**Question 1**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>   
  
and load the data into R. The code book, describing the variable names is here:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf>  
  
Apply strsplit() to split all the names of the data frame on the characters "wgtp". What is the value of the 123 element of the resulting list?

"wgtp" "15"

"wgt" "15"

"" "15"

"wgtp"

**Question 2**

Load the Gross Domestic Product data for the 190 ranked countries in this data set:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv>   
  
Remove the commas from the GDP numbers in millions of dollars and average them. What is the average?   
  
Original data sources: http://data.worldbank.org/data-catalog/GDP-ranking-table

293700.3

381615.4

381668.9

377652.4

**Question 3**

In the data set from Question 2 what is a regular expression that would allow you to count the number of countries whose name begins with "United"? Assume that the variable with the country names in it is named countryNames. How many countries begin with United?

grep("\*United",countryNames), 5

grep("^United",countryNames), 3

grep("^United",countryNames), 4

grep("\*United",countryNames), 2

**Question 4**

Load the Gross Domestic Product data for the 190 ranked countries in this data set:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv>   
  
Load the educational data from this data set:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv>   
  
Match the data based on the country shortcode. Of the countries for which the end of the fiscal year is available, how many end in June?   
  
Original data sources:   
<http://data.worldbank.org/data-catalog/GDP-ranking-table>   
<http://data.worldbank.org/data-catalog/ed-stats>

15

8

13

31

**Question 5**

You can use the quantmod (<http://www.quantmod.com/>) package to get historical stock prices for publicly traded companies on the NASDAQ and NYSE. Use the following code to download data on Amazon's stock price and get the times the data was sampled.

library(quantmod)

amzn = getSymbols("AMZN",auto.assign=FALSE)

sampleTimes = index(amzn)

How many values were collected in 2012? How many values were collected on Mondays in 2012?

365, 52

251, 47

250, 47

252, 47

* [Peer Assessments](https://class.coursera.org/getdata-015/human_grading/)
* /Getting and Cleaning Data Course Project

[Help Center](https://accounts.coursera.org/i/zendesk/courserahelp?return_to=https://learner.coursera.help/hc)

* closed 1wk 1d ago

Submission Phase

* + [1.Do assignment](https://class.coursera.org/getdata-015/human_grading/view/courses/973502/assessments/3/submissions)
* closed 9hr 43m ago

Evaluation Phase

* + [2.Evaluate peers](https://class.coursera.org/getdata-015/human_grading/view/courses/973502/assessments/3/peerGradingSets)
* Results Phase
  + [3.See results](https://class.coursera.org/getdata-015/human_grading/view/courses/973502/assessments/3/results/mine)

No work was submitted before the submission deadline. You will not be able to evaluate the work of your peers or receive an evaluation.

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit: 1) a tidy data set as described below, 2) a link to a Github repository with your script for performing the analysis, and 3) a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.    
  
One of the most exciting areas in all of data science right now is wearable computing - see for example [this article](http://www.insideactivitytracking.com/data-science-activity-tracking-and-the-battle-for-the-worlds-top-sports-brand/) . Companies like Fitbit, Nike, and Jawbone Up are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained:   
  
<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>   
  
Here are the data for the project:   
  
<https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>   
  
 You should create one R script called run\_analysis.R that does the following.

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.
3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names.
5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

Good luck!

Please upload the tidy data set created in step 5 of the instructions. Please upload your data set as a txt file created with write.table() using row.name=FALSE (do not cut and paste a dataset directly into the text box, as this may cause errors saving your submission).

### Evaluation/feedback on the above work

**Note**: this section can only be filled out during the evaluation phase.

Has the student submitted a tidy data set? Either a wide or a long form of the data is acceptable if it meets the tidy data principles of week 1 (Each variable you measure should be in one column, Each different observation of that variable should be in a different row).

Please submit a link to a Github repo with the code for performing your analysis. The code should have a file run\_analysis.R in the main directory that can be run as long as the Samsung data is in your working directory. The output should be the tidy data set you submitted for part 1. You should include a README.md in the repo describing how the script works and the code book describing the variables.

### Evaluation/feedback on the above work

**Note**: this section can only be filled out during the evaluation phase.

Did the student submit a Github repo with the required scripts?

Was code book submitted to GitHub that modifies and updates the codebooks available to you with the data to indicate all the variables and summaries you calculated, along with units, and any other relevant information?

I was able to follow the README in the directory that explained what the analysis files did.

### Overall evaluation/feedback

**Note**: this section can only be filled out during the evaluation phase.

As far as you can determine, does it appear that the work submitted for this project is the work of the student who submitted it?

Please use the space below to provide constructive feedback to the student who submitted the work. Point out the submission's strengths as well as areas in need of improvement. You may also use this space to explain your grading decisions.



You've written 0 words

Bas du formulaire

Bas du formulaire

Bas du formulaire