coursera

Discussion Forums

Week 4

Discuss this week's module: Week 4.

← Week 4

[TIP] The infamous and so useful post from David Hood for the final assignment .. and quizzes

Philippe Alcouffe Mentor Week 4 · 2 months ago · Edited

Hello,

Yes .. I know .. I already pinned that in the general discussion ! But I also know this final assignment was most challenging to me when I took the course and that this post really shed so much light on all the questions I had.

Definitively worth looking at it HERE.

Good luck with the assignment and keep on the good work !

Ph

12 Upvote · Follow 21 · Reply to Philippe Alcouffe

Earliest **Top** **Most Recent**

HJ

hamid junejo · 2 months ago

Good luck with assignments

0 Upvote · Reply

KT

Krishna C Thirumalasetty · 2 months ago · Edited

Help Center

Philippe - My Course Project did not get thru, because 2/3 of the peers believed my Dataset and run_analysis.R, did not produce the correct results.

I am absolutely positive, my code and tidyData.txt is correct, and to the required specifications. I think, people assume my tidyData is incorrect, because it only has Mean() and Std() values, whereas every other tidyData.txt I have looked at during the review process, they have also included MeanFreq() - I am positive, this measurement MeanFreq() should not be included in the tidyData.txt, as its not part of the requirements.

How do I challenge the Peer-Review, so I can resubmit my result and not have to waste another 4 weeks on this. Please advice.

0 Upvote · Hide 4 Replies

Philippe Alcouffe Mentor · 2 months ago

Hello,

Thank you for reaching to me on this concern. Peer reviewing might sometimes be frustrating (I myself lost points on that assignment as I chose the narrow form vs the wide form though both are explicitly stated to be ok ... :-( )

Your session is not over so you might still have time to resubmit within your session. If you were to run out of time, you can go on the next session and in order not to wait 4 weeks, post your submission link in the forum in a new post (not as an answer to this general pinned post). This usually encourages people that are in advance in the session to peer review so that you do not wait 4 weeks. As soon as it is peer reviewed you will get your final grade without waiting 4 weeks. If you resubmit in that session you can also reinforce with a separate post in the week 4 forum.

As for resubmitting, you have the choice to modify or not accordingly to the peer-review feedback. In my personal opinion, using or not MeanFreq should not be penalized (David Hood in the post mentions to explicit the choice made) ...

The grading is done through an average so usually, extreme peer reviews do not tilt you over board.

Wish you the best in this.

Ph

Help Center

1 Upvote

coursera

AP

Adur Pandya · 2 months ago

The project requirements definitely need to be better worded.

17 Upvote

KT

Krishna C Thirumalasetty · 2 months ago

And the Points need to be broken down. Giving 12 Points for the correct and half of that for partial seems a bit drastic. If a peer thinks all your answers are partially right, then you end up failing - because you get half the points and don't meet the 80% Pass requirement. Maybe break it down 2 -4 Point Questions.

4 Upvote

vivek singh · 2 months ago

Hey folks,

Kindly grade my assignment.

https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project/review/5QnACFOWEeapUhLW_ceEKQ

Thanks.

0 Upvote

Reply

Reply

AN

Asher Nig · 2 months ago

Hi Phillipe a reviewer failed me of no reason , I need to understand the reason for failure.I have submitted the script files and also the tidy dataset . Who should I contact to get it reviewed again.

0 Upvote · Hide 1 Reply

Philippe Alcouffe Mentor · 2 months ago

Hi Asher,

Help Center

Sorry that your peer reviewer did not elaborate enough on the comment part of his review ...

You can simply resubmit within your session (if there is still time) or go on to the next session. To encourage peers to review, you can then post your submission link in the week 4 forum.

Ph

0 Upvote

Reply

Reply

Chandramani Rangari · 2 months ago

I am not able to understand data set for a project . Can anybody point me some direction.

0 Upvote · Hide 8 Replies

Philippe Alcouffe Mentor · 2 months ago

Hello,

The link posted at the beginning of this thread HERE is usually a good start though it might not cover all of your questions. Can you elaborate on the questions it did not help with ?

Ph

4 Upvote

DP

Daniel Petruk · a month ago

Hi Phillipe! How long does the grading process take? I'm debating taking the next courses(but the anticipation is killing me!. In the mean time I've been grading as much as possible in hopes of helping other students get their grade.

Thank you!

Daniel

0 Upvote

Help Center

Philippe Alcouffe Mentor · a month ago

Hello Daniel,

From the classes I took with peer assignment, you usually get your grade 2-3 days at most after the deadline of the assignment.

But you can still enrol in the next courses even if you did not get your grade yet (and even re-submit an assignment ..)

And thank you for helping grading other students !

Ph

1 Upvote

DP

Daniel Petruk · a month ago

Thank you!

0 Upvote

SV

Sam Vennell · 24 days ago

Hi Phillippe,

I am unclear on how the x and y components of the testing and training data are structured within the dataset. I find that the "x" files each consist of several million numbers in scientific notation, and the "y" files contain integers separated by newline characters, but only a few thousand of these - no where near as many as in the "x" variable.

How should I interpret this data? Do the "y" values somehow shed light on the interpretation of the "x" values? Have I missed something in the ReadMe that will clarify?

Regards,

Sam

2 Upvote

Philippe Alcouffe Mentor · 24 days ago

Hello Sam,

The y_files are the id of the labels (this is clarified HERE) and you have as many lines as in the x_files (measurements). Labels (id + text) are in activity_labels.txt.

Help Center

When I did that assignment, I draw as lego blocks the different pieces of data (as advised by David Hood in his post HERE / section How do I put the data together).

This of course assume that you do not alter the sorting of the y_files as they are to be 'glued' to other piece of data.

Hope this helps

Ph

0 Upvote

SV

Sam Vennell · 22 days ago

Thank you Philippe for your prompt reply and kind advice :-) It's all making sense now, I think it was just getting a little late in the day before...

0 Upvote

CR

Chase Renick · 11 days ago

Philippe - or any other empathetic student ,

Can you review my assignment so I do not get stuck in Coursera academic purgatory? Thanks ;)

https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project/review/7aW_OnVDEeaAlg6zA9zuZQ

0 Upvote

Reply

Reply

HR

HARISH KUMAR RONGALA · 2 months ago

Hello,

Do we need to include the steps of downloading and unzipping the data set in "run_analysis.R" ? Or can I assume that the data set is already extracted in to the folder "UCI HAR Dataset" in the working directory ?

Help Center

Thanks

3 Upvote · Hide 4 Replies

KT

Krishna C Thirumalasetty · 2 months ago

Harish - I would include it, but do a Guardian check for, the folder before downloading the file.

0 Upvote

HR

HARISH KUMAR RONGALA · 2 months ago

Thank you Krishna, I will make my script check for the folder and accordingly download and extract the data set from web.

0 Upvote

vivek singh · 2 months ago

Hey folks,

Kindly grade my assignment.

https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project/review/5QnACFOWEeapUhLW_ceEKQ

Thanks.

1 Upvote

DP

Daniel Petruk · a month ago

just did!

0 Upvote

Reply

Reply

YZ

Yuxuan Zhao · a month ago

Hi, I read your guide, but still I don't understand the second step "Extract only the measurements on the mean and standard deviation for each measurements".

Help Center

Doe it mean to calculate the mean and standard deviation for each measurements and place them in a new row.

0 Upvote · Hide 6 Replies

SC

Sharmistha Chakrabarti · a month ago

Hello Zhao, once you merge the data set (let's call it myData), you would have several column names containing mean() and std(). You want to extract those columns of myData whose names contain mean() and std(). Make sure you also keep the column activity and subject.

I hope it helps. Good luck !

Sharmistha

1 Upvote

Apichart Thanomkiet · a month ago

Hello Yuxuan,

It took me a while to understand the whole assignment. Sharmistha told you the right thing. When you do the extraction of mean() column is a bit tricky because there are some variable that contain meanFreq. You may or may not want to extract these column.

Good luck!

Apichart

1 Upvote

DP

Daniel Petruk · a month ago

I agree but I can see how I would interpret meanFreq as Mean Frequency which can meet the requirement of being a measure of mean, for some students.

If during analysis it is not necessary we can simply omit it at time of analysis or remove it once we have actually tested and validated our data. Not to mention that not being able to speak to the person who is providing requirements has it's own set of challenges. Now, if there are other columns that should clearly not be there, that's an objective not met.

0 Upvote

Help Center

DZ

Dong Zheng · 6 days ago

HI Sharmistha, Apichart, Daniel,

i am looking at 'X_train.txt' and 'X_test.txt', there are no column headings. How does one figure out which columns are for what measurements? I guess i just don't understand how 'features.txt' (with 561 measured features) fit with the rest of the data files? Could you help?

Dong

0 Upvote

DZ

Dong Zheng · 6 days ago

hmm, i think i understand now, each of the main data files has 561 columns, so each column represents a measured feature.

0 Upvote

Apichart Thanomkiet · 6 days ago

it was a bit confusing at the beginning ;)

0 Upvote

Reply

Reply

Apichart Thanomkiet · a month ago · Edited

I would like to thank to Phillipe. Her guideline gave me an idea how to build the logic to finish this assignment. I recommend to anyone who is taking the assignment to read her articles and please be patient. Take your time and go through it. I am not good at this, so it took me sometime to finish this assignment. If I can do it, you guys can do it as well.

By the way my assignment is waiting for reviewing. Please grade me ;)

Apichart

1 Upvote · Reply

Philippe Alcouffe Mentor · a month ago

Hello Apichart,

Help Center

The credit goes to David Hood for his post (covering this assignment and also the quizzes). I only pinned a reference text as I found it so useful when I took the class.

... and i am a guy ;-) .. with trendy long hair (actually even longer now than on that pict)

Ph

0 Upvote · Hide 1 Reply

Apichart Thanomkiet · a month ago

Opps my mistake sorry!!

0 Upvote

Reply

Reply

Juan Bosco Mendoza Vega · a month ago

I'm not sure how descriptive need to be the variable names in our tidy data set.

is it enough to have names different to "V1", "V2", "x.1", and the like?

Or do we need to make expicit some abbreviations, like naming them "Accountant" instead of "Acct"?

0 Upvote · Hide 2 Replies

Apichart Thanomkiet · a month ago · Edited

Hello Juan,

The V1, V2, V3, and etc, are generated column names by R. I think you probably want to rename them according to the variable according to the name of the features which you can get some ideas from the linked given by Philipe from this post. Otherwise you might need to extract the mean and std columns manually and that will be a bit painful because you have 561 variables if I am not wrong.

Good luck.

Help Center

Apichart

0 Upvote

Juan Bosco Mendoza Vega · a month ago

Hi!

That makes sense. I guess it's time to figure out the appropiate layout for the code book, then.

Thanks for the reply!

0 Upvote

Reply

Reply

WS

wendy sarrett · a month ago · Edited

Thank you for the article. My question (and I posted in the week for section as well as well but people seem to be looking here) is the y_train.txt, y_test.txt and subject_test.txt and subject_train.txt appear to be gibberish. From the description I expected to see the lables that go with the data but all I see is gibberish..like it's corrupted? Please advise. Thanks!

I opened y_test.txt in a hex editor and it was a series of numbers with a period in between 5.5.4.4 etc.

1 Upvote · Hide 1 Reply

Monica Brisnehan · a month ago

Wendy - there are some good hints to figuring that out in the link in the first post of this thread. (Link: https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-the-assignment/)

1 Upvote

Reply

Help Center

coursera

Reply

RG

Robert Gerath · 21 days ago

Under "**what columns are measurements on the mean and standard deviation**," the sentence "Based on interpreting column names in the features is an open question as to is the the entries that include mean() and std() at the end, or does it include entries with mean in an earlier part of the name as well" doesn't make any sense to me. Would the author please revise this passage to make it clear what they're trying to say?

0 Upvote · Reply

Philippe Alcouffe Mentor · 20 days ago

Hello,

When looking at the column names, you will see (using regexp for example) that some names are ending with mean(), some names are with FreqMean.

Students regularly wonder whether they should consider everything that has mean (thus also the FreqMean kind of names) or only the ones that have mean() at the end.

What David Hood says is that it is actually an open question.

Hope this helps clarifying.

Ph

0 Upvote · Reply

AC

Alex Cheung · 15 days ago

Hi All,

I have successfully read all the data, but seemly I still have no idea what do the question intend me to do. First step, ask me to merge training and test set. I assume this means to use Y_train / Y_test as the header and merge the Y data with X_train/X_test?

1 Upvote · Hide 19 Replies

See earlier replies

BA

Beverly Andrews · 14 days ago

After reading this I changed everything around, that's what I was doing originally, but then after reading this I added the subject and activity data, so then I re-labeled those columns (two V1 columns causes a problem

Help Center

and when I put the labels on before I was getting a bunch of "duplicate columns" errors) and then [illegible] all of them together and pull the std, mean, and label columns. And I still don't get it. I've read the document referenced in this post forwards and backwards. It's really disheartening.

0 Upvote

Juan Bosco Mendoza Vega · 13 days ago

I think I have made this more confusing. Sorry about that.

Try importing each file in Train to R (x, y and subject_test), each to its own object. Then, use the functions str, dim adn/or summary to explore what each object contains.

Doing so, hopefully, will reveal that these objects share either the number of columns or the number of rows they have. With this information, you can decide what kind of merging to do (by rows, or by columns) so it results in a rectangular data set for Train.

Then, do the same with the files in Test, and this time you can check what traits the data set for Train has in common with the data set for Test, either the number of columns or the number of rows they have.

This way, you can now merge these two data set into one. Keep in mind that bot data sets have data for the same features and that there's a file called "features" that can help you in naming and selecting them.

2 Upvote

AC

Alex Cheung · 13 days ago

Hi Juan,

Thanks for the more explanation. It helps the task becomes more clear. But I still have a few questions and I hoped you dont mind to answer.

First, I noticed there are 561 variables in both x file and feature file. So I believe that feature mainly is for the header and it allows me to merge both train and test data because of the same header. However, I dont see any purpose of y and subject. My initial thought use their the value to tag back to the feature's unique id. But it doesn't make sense since there are so many duplicated in y and subject.

Does it make sense to you?

Help Center

1 Upvote

BA

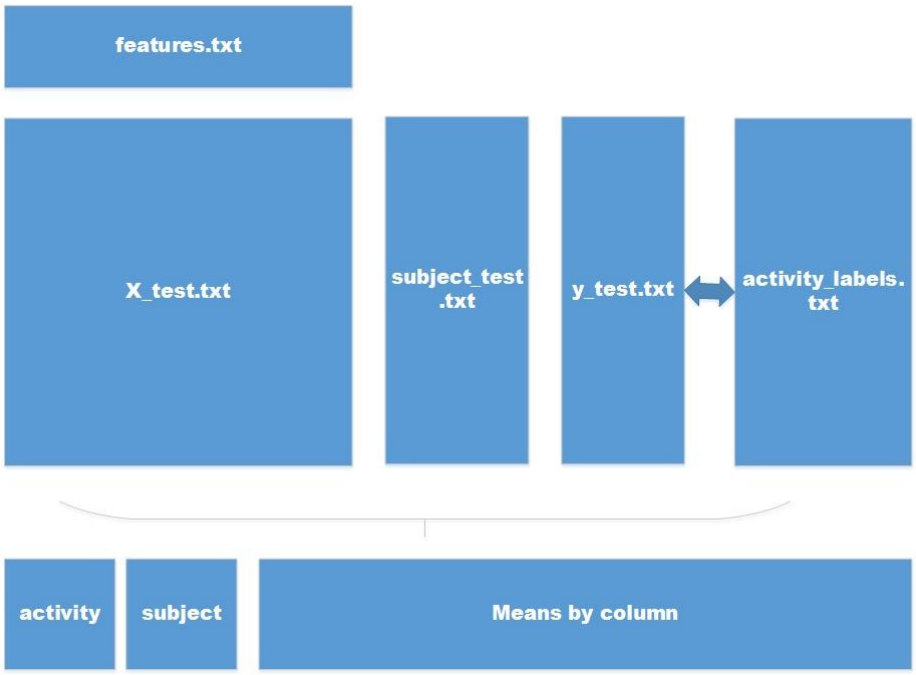
Beverly Andrews · 13 days ago

Hi Juan,

Thanks for the clarification. It has become abundantly clear that I was doing it wrong originally and you helped put me on the right track. Thank you.

Alex,

Remember that the final data set is going to be aggregated by subject and activity. So the files you'll need the three files in test, the three files in train, activity_labels and features.txt... This is what I did:



3 Upvote

AC

Alex Cheung · 7 days ago

Hi Beverly,

I hope you well! I found your diagram helps me a lot. I managed to apply feature as header and y tagged with activity as a row name. But I dont understand the need of subject.txt, does it for extra column that cbind with main x.txt or something familiar purpose of y.txt?

Help Center

0 Upvote

Ignas · 7 days ago

Thank you, this is very helpful, a graph like this should have been included in the assignment instructions.

How do you link it to the real datasets, though? I get it that the labels for 561 column are the feature.txt values as size is a match.

But how do you find out the correct labels for 128 columns of real data sets? There is no file / vector with 128 values in the zip, and filtering out irrelevant features from 561 still doesn't leave me anywhere close to 128.

Am I missing a file?

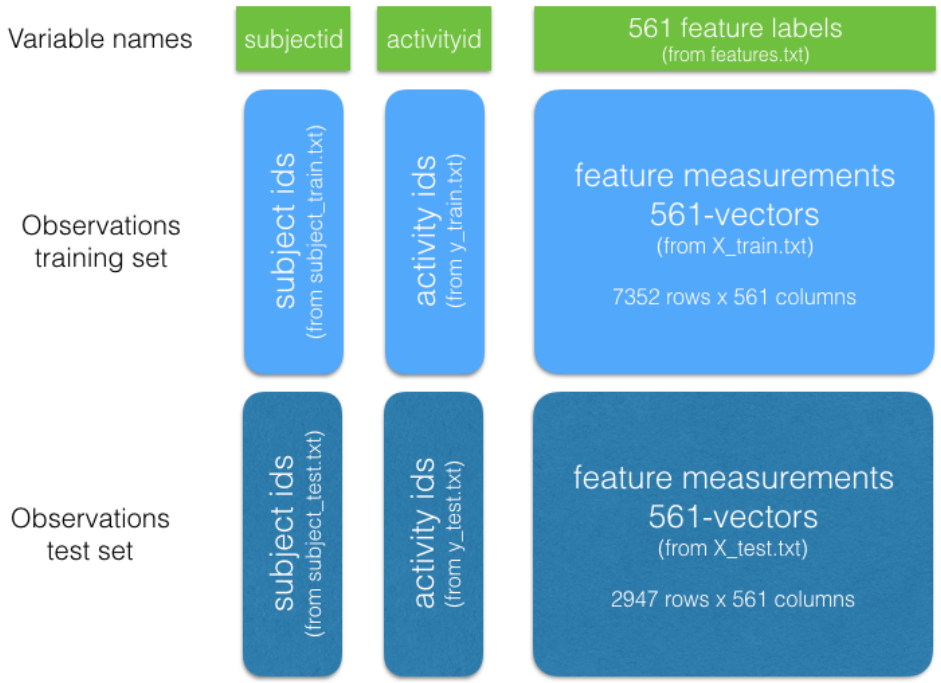
0 Upvote

Philippe Alcouffe Mentor · 7 days ago

Hello,

Here is the diagram I built on my side when I took the class. Feel free to use in your assignment.

Ph



0 Upvote

Help Center

Ignas · 7 days ago

Thanks Philippe, the diagram makes it easier to understand, however my question remains unanswered. I'll try to elaborate better.

Assuming we need to look into the 18 files in Inertial Signals folders (example below), they have 128 columns each (7352 or 2947 rows accordingly), with no headers. It's not written anywhere that these can be ignored, therefore I **assume** the requirement is to merge these into a single file (Part 1), like we do with the other files with 561 columns. While this can be done, I don't see how I can extract mean & std columns (Part 2) as without any labels I can't identify which columns from 128 to select and which to ignore.

Could you 1) confirm whether my assumptions are correct and if yes 2) advise from where to source the labels for 128 columns?

It's very confusing as the instructions on Coursera are so inexplicit, grateful for any clarification.

Ignas

From READ.ME:

- 'train/Inertial Signals/total_acc_x_train.txt': The acceleration signal from the smartphone accelerometer X axis in standard gravity units 'g'. Every row shows a 128 element vector. The same description applies for the 'total_acc_x_train.txt' and 'total_acc_z_train.txt' files for the Y and Z axis.

0 Upvote

BA

Beverly Andrews · 7 days ago

Subject.txt is the subject number that you'll need to summarize by.

0 Upvote

BA

Beverly Andrews · 7 days ago

@Ignas,

Don't use the Inertial Signals folders, use the file names listed in the diagram in test and train.

0 Upvote

AC

Alex Cheung · 6 days ago

Help Center

Thanks all the help, finally I am on last step of the assignment! One thing I dont get is the last question intend us to breakdown all the mean for every single activity, subject and finally each variables?

0 Upvote

Mark Yow · 6 days ago

Take subject #1 and walking for example. There are many measurements of him just walking. You will get the mean of those measurements for each of the variables. Follow this for all the subjects and all the activities. This was my interpretation.

0 Upvote

AC

Alex Cheung · 5 days ago

Finally got it done!! Thanks for everyone clear explanation and support. One more finally question: they need to store the final table by using write.table() function what does it mean?

0 Upvote

Juan Bosco Mendoza Vega · 4 days ago

This means you need to export your data set to a file in your computer.

Check the documentation for write.table() (call ?write.table inside R) for more details on how to go about this step.

0 Upvote

Peng Cheng Han · 6 hours ago

for X_train.txt, i still don't see the 561 columns. is it delimited by some characters?

0 Upvote

Reply

Reply

Help Center

David Carnahan · 13 days ago

I'm really having a hard time getting through the last step!!!

I have the merged dataset -- with subject, and activity, and the measures.

I was going to try to use the dplyr functions:

ds1 <- ds0 %>% group_by(subject) %>% aggregate ... but this is not working. Any advice you can give is appreciated.

1 Upvote · Hide 3 Replies

AC

Alex Cheung · 13 days ago

Hi David, may I ask you what data do you merge since I still not 100% sure I am following the correct path. Do you use the feature label as header and merge under the x and y data?

0 Upvote

David Carnahan · 13 days ago

Hi, Alex, thanks for reaching out with your question. I have done everything needed to get the final merged dataset -- it is the final step of creating the summary dataset with the mean of all the variables. I think I have figured it out ... I need to convert things to factors before I can do the summary stats. I hope this helps someone else.

0 Upvote

Francisco Jaramillo Aguilar · 11 days ago

Hello, Can you tell me what about body_acc_x_train, ... etc. those files has to be merged too?

0 Upvote

Reply

Reply

CR

Chase Renick · 11 days ago

Hi Phillipe,

Would you be willing to review my programming assignment? Thanks

Help Center

https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project/review/7aW_O[illegible]zuZQ

0 Upvote · Reply

1 2

Reply

Reply

Help Center