Announcements

Data Analysis Assignment One is Posted

Hi Everyone

I have posted the first data analysis assignment. If you go to the Data Analysis Assignments page and scroll to the bottom of the page you should be able to go to the assignment.

I need to bring a couple of things to your attention:

1. It was pointed out that there are potential security issues with having you run each other's code for the peer evaluations. Due to these issues, you will no longer be submitting R code to reproduce your analyses and the reproducibility criterion will no longer be part of the grading rubric. I still think reproducibility is incredibly important and I hope you will take the time to write code that reproduces your analyses, but you will not have to submit/evaluate it for the assignment.

2. I moved the due date to give you the entire weekend to submit the data analysis. It is now due Monday, February 18th, 2013 at 7:30AM UTC - 5:00 (Baltimore Time).

Upcoming Deadlines

Quizzes

Weekly Quiz 2

Sun 3 Feb 2013 7:30 PM AKST

New Lectures

Example Analysis Assignment (7:47)

Exploratory Graphs Part 1 (20:27)

Exploratory Graphs Part 2 (23:19)

Expository Graphs (15:46)

Hierarchical Clustering (17:40)

K-Means Clustering (8:36)

Dimension Reduction (21:24)

3. You will be submitting two components for the data analysis. The first is the main text and references (no more than 2000 words), submitted as pdf or input into the text box. The second is the figure uploaded as pdf, png, or jpg along with a figure caption (no more than 500 words). You can see the evaluation criteria for each component if you go to the data analysis assignment.

4. For the moment, I have only placed the data on the Coursera site. If you are unable to get them from there, let me know on the message boards and we will do our best to make sure you can access them.

5. I will try to post the lecture videos for statistical modeling basics by mid-late week this week. You should have two full weekends with the lecture videos from this week and next while you are working on this assignment.

Again, this is the first time trying this type of data analysis at this type of scale. I am going to be working hard to make sure everyone understands the assignment and can put their best analysis forward. Please post questions to the message boards if you have them.

Good luck!

Jeff

Sun 3 Feb 2013 7:04 PM AKST

Example Data Analysis Assignment

I've uploaded an example data analysis project that (I hope) will make expectations clearer. The example analysis is available as a zip file from here:

https://dl.dropbox.com/u/7710864/courseraPubli

or here:

https://spark-public.s3.amazonaws.com/dataanalysis/example

This folder has the question I was answering with the data analysis (prompt.pdf) and a folder with the components you will submit for your data analyses (the assignments folder). The components are:

The main text with references (pdf)
https://spark-public.s3.amazonaws.com/dataanalysis/example

The figure (pdf)

https://spark-public.s3.amazonaws.com/dataanalysis/finalfigu

The figure caption (pdf) - you may submit this as text

https://spark-public.s3.amazonaws.com/dataanalysis/figureCa

The .R or .Rmd file to reproduce your analyses

https://spark-

public.s3.amazonaws.com/dataanalysis/earthqu

The .rda file for the analysis

https://spark-public.s3.amazonaws.com/dataanalysis/quakesF

(Note, your final submission will not need to include the data set. But your .Rmd/.R file should run in the same directory as the .rda file)

The folder also includes all of the raw analysis files I created when performing the data analysis. The goal was to give you an idea of all the files involved in a data analysis.

I have also uploaded a video where I describe the contents of the exampleAnalysis folder under the Week 3 lectures.

If you are unable to get access to the files above or the zipped project folder, please post to the message boards. The links above are for Dropbox and for the Coursera data store. If there are other places that are more accessible to people in other parts of the world, it would be great if we could continue to help each other make the materials available.

I have been seeing a lot of questions about the data analysis on the message boards so I thought I'd take a minute to answer a few of them here.

Q. Can I/will I need to use my own data

set?

You will be provided a data set along with the question to answer about that data set. This uniformity is necessary to permit peer review at the scale of a class like this. We also don't want to put people without data sets to analyze at a disadvantage. That being said, if you perform analyses using the techniques of this class and your own data, I'd love to be able to highlight them, so please feel free to post them to the message boards.

Q. What format will the data be in?

I will provide the data both as a .csv file and a .rda file. I would suggest that your code that you submit uses the .rda file and one of the first commands you run is to load the data with the load() command. Loading the data in this way will reduce potential problems with people running your code on another platform.

Q. Do I need to submit R code or can I submit a different kind of code?

Yes, if you want to get the reproducibility points you should submit R code or a .Rmd file that can be run by one of your classmates. If you have a Mac you may want to test your code on Windows and vice versa. You should assume that the code will be run in a directory where the .rda file is located. Your code should clearly state all the packages that need to be installed/loaded to run your code.

Q. When will the data be made available?

The data analysis assignment and corresponding data will be made available

tomorrow, February 3rd at 11:30PM Baltimore (UTC - 5:00) time. It will be due on February 16th at 11:30PM Baltimore (UTC - 5:00) time.

Q. But we haven't learned all of the statistical methods we need yet, how do we do the analysis?

Since this course covers a complete data analysis, we needed to cover background/data cleaning material before statistical methods. We will be covering exploratory analysis methods this week. The videos will be made available on February 3rd at 11:30PM Baltimore (UTC - 5:00) time. I will also release the Week 4 videos midweek. These videos will cover basic inferential analysis. You should therefore, have two complete weekends to watch the exploratory and inferential videos and perform your analysis.

This simultaneously learning/performing of a data analysis is the best way to learn how to master these skills. It is also necessary given the short time frame of the class and the necessity to leave sufficient time for peer review.

Q. I have a ton more questions.

That's not a question! Just kidding. I have created a new forum for the first Data Analysis Assignment. I will be around both later this evening, over the course of the weekend, and over the next couple of weeks on the message board. I will do my best to answer questions there.

Q. This seems like a huge hassle. Why not more quizzes?

Great question! The most important component of data analysis training is practicing data analysis. I hope that these assignments will give you a chance to practice that. Admittedly this is the first time anyone has tried to do data analysis projects at this kind of scale with peer review. There will definitely be a few kinks. Your instructor is committed to spending the time it takes to help iron out those kinks. He really hopes you will get the same sense of adventure from this experiment that he does and be willing to be a little patient.

Sat 2 Feb 2013 1:40 PM AKST

Welcome to Week 2 of Data Analysis

Hi Everyone,

Welcome to the 2nd week of Data Analysis. This week we will be covering material on how to organize a data analysis, the structure of files in a data analysis, how to get data, and the basics of how to clean data. The course material will start getting more specific on how to use data in R, but there will still be plenty of room for conceptual discussions that are happening on the course message boards.

I would like to thank everyone for their active participation on the message boards. It is making the class way more fun and interactive than it would be if it were just a bunch of video lectures. I appreciate all of the help that folks have given me in making the source materials better and finding typos in the lectures/quizzes. I'd also like to extend particular gratitude to the community T.A.'s

who have been tireless about answering questions and keeping me informed of issues that need to be fixed.

A couple of notes on the upcoming data analysis assignment that will be assigned starting next week. First, the data set you will use for this analysis, along with the scientific question, will be defined by the assignment. You will not need to find your own data set and question to answer. Second, I will be posting an example analysis late this week so that you can see an example of how it should be formatted and what the different components of the analysis are.

Just a reminder, the quiz for this week is due on February 3, 11:30PM UTC - 5:00.

Thanks again for all of your hard work and have a great week!

Jeff

Mon 28 Jan 2013 9:44 AM AKST

Data Analysis | Update to Quiz Due Dates and Due Date Policy

Hi Everyone,

I'm glad to see that the first week of data analysis has led to so much enthusiasm. I'll put out a recap announcement later in the weekend, but I wanted to address some confusion about quiz due dates, late days, and to address a popular and reasonable request about the due dates of quizzes. I am sorry for any confusion about this, this is my

first Coursera class and I'm still learning the platform.

(1) To accomodate a very common request, I will make all lecture material/quizzes available Monday morning at 12am UTC-5:00 every week and have the quizzes/assignments due Sunday evenings at 11:30pm UTC-5:00. Many people who are taking the class while working have asked for the assignments to be due at the end of the weekend.

(2) There are soft and hard deadlines for the quizzes. The soft deadline will be Sunday at 11:30PM UTC-5:00 every week. The hard deadline will be Tuesday at 11:30PM UTC-5:00 every week.

(3) Quiz 1 now has a soft deadline of January 27 at 11:30PM UTC-5:00 and hard deadline of January 29 at 11:30PM UTC-5:00.

(4) You may submit the quiz after the soft deadline but there will be a 10% penalty for each day after the soft deadline.

(5) You may not submit the quiz after the hard deadline.

(6) You have up to 5 late days that you can use on the quizzes. You may use your late days on whichever quizzes you want. If you use a late day on a quiz, then you will not incur the 10% penalty for that day.

(7) You may not use late days after the hard deadline. So only 2 late days may be applied to each quiz.

I hope that makes things clearer. I will change the information on the Course

Logistics page to reflect this change.

Best,

Jeff

Sat 26 Jan 2013 7:58 AM AKST

Grammar in the data analysis peer assessments

Hi Everyone,

One of the most popular threads on the discussion board is about the component of the peer evaluation that asks, "Is the analysis written in grammatically correct English?" I do appreciate very much that there is a broad range of English language skills for this course and that English is not the first language of many people in the class. If you asked me to write a data analysis in a language that was not my first language, I would be nervous as well.

I do believe that communication is a critical, and often overlooked, component of data analysis. Therefore, I believe that including some evaluation of the clarity of writing is important. Since the course is being given in English, I used the term "grammatically correct English". I believe that grammar is important, but it is less important for the purposes of this course than clarity. I will therefore revise that criteria to be, "Is the analysis written in clear and understandable English?". I hope that people will evaluate this criteria seriously, but be understanding of the heterogeneity of English skills in our

community. I also hope that people who are not fully comfortable with their English skills try to have a friend who is more comfortable read their analysis before submitting.

Jeff

Thu 24 Jan 2013 3:47 AM AKST

Personal emails

I'm excited to have you all in class and looking forward to working with you on Data Analysis. I did want to point out that I have been getting a very large volume of personal emails related to the course. While I wish that I could answer them all, given the size of the class it is going to be impossible. So in the interest of fairness I won't be responding to any personal emails about the class.

I hope that you will instead consider posting to the message boards. This will allow your fellow students to see your questions/comments and get involved. I am also making an effort to be on the message boards as often as possible, so hopefully I can answer a lot of the questions/comments. We also have some wonderful and helpful Community TA's who can answer your questions if you post them to the board.

Thanks for your understanding and let's keep the energy going!

Jeff

Tue 22 Jan 2013 4:37 PM AKST

Data Analysis is now

open! Welcome!

We originally announced this class in July of 2012, along with 7 other MOOCs from the Johns Hopkins Bloomberg School of Public Health. Since then I've been eagerly awaiting the chance to open this class, take part in this grand experiment, and get to know all of you.

As part of that getting to know you process, I would really appreciate it if you could fill out the pre-course survey so we can understand a little more about your background and interests:

Launch Pre-Course Survey

All the anticipation for the course has me fired up and ready to go. I hope you are equally excited about Data Analysis. The goal is to start from scratch and zoom through all the key parts of analyzing data. This means that we will be covering a lot of material in 8 weeks - everything from exactly what we mean when we say data, to downloading data off the web, to linear models, to classification trees!

Since we will be covering a huge breadth of material, we will have to skim the surface of a lot of topics that I'm sure you'd be fascinated to know more about (I know I am!). I'll do my best to point to resources that will get you going on these topics; if you have any questions, make sure you post them to the message boards and I'll try to help you find answers.

It is also important to get into the "hacker" mentality when learning data analysis. Data analysis is currently at least as much art as it is science. It takes practice, searching out

answers to questions that aren't covered in lectures, asking lots of questions, and learning from your failures. Along the way, I hope that our interaction and the community of data analysts we are building can help you tackle the real challenges involved in working with messy and complicated real world data.

I've posted information about the course syllabus, course logistics, and data analysis projects. If you have any questions, please post them to the message boards and I'll do my best to answer as fast as I can.

Welcome again and see you on the message boards!

Jeff

Tue 22 Jan 2013 5:00 AM AKST