Lab 8 (7/26/2019) on Statistical Computing/Open Data/Data Science/Data Exploration/Data Mining

Data Science, Statistical Modeling, and Machine Learning are important, current topics in Computer Science. There is a lot of open source software and open data available that are helping

The best way to learn data science is to do data science. In this lab we will give you some level of practice, but to learn it well, you will need to do more. This site recommends five projects to try in order to learn data science. Another place to look at is Kaggle. Kaggle holds competitions on machine learning/data science. A couple of years ago, Diogo, a student in RCOS, won first place in the cause-effect pairs competition and his code can be found here. If you are interested in data science you can participate (and maybe even win!).

For this lab, please do the following - your Lab report should be in your github page

1. Read Chapters 3 and 5 of https://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf on plotting and regression. The chapters work through some simple plotting and regression using datasets built into R. **You do not need to show anything from step 1 in your lab report.** This is just a learning step. You may have problems:
    2. With the **rgl** library if you are on a Mac. Feel free to ignore that specific plot, or, if you'd like, download XQuartz from http://xquartz.org.
    3. As the last part of the tutorial you will be asked to execute ***data("bodyfat", package="mboost")*** execute ***data("bodyfat", package = "TH.data")*** instead.
2. Now go to DataCamp https://www.datacamp.com/home and create an account. We will leverage the free lessons that DataCamp provides for the rest of the data anlysis portion of the lab.
3. Do the introductory lesson of "Data Visualization with ggplot2 (Part 1)". Take screen shots along and put them in your Lab Notebook for Lab 7.
4. Now do the Parallel Slopes lesson of "Multiple and Logistic Regression". Again, take screen shots along the way and put them in your Lab Notebook.
5. ***The rest of this lab applies to your project.*** If you haven't already done so, create an Observatory http://rcos.io page and a repo page for your project. Create a slack channel. Please choose a license for your repo. (If you are joining an existing project that has a communication channel, your group can just join that. Be sure to tell me this, and to tell me how to find it.) Write your first blog as a paragraph description of the status of your project - What did you do last week on your project? You only have 3 more weeks to finish. Add a pointer ***in your lab notebook*** to your page on Observatory and make sure we can get to the project page, the repo page and the blog page from it.
6. Submit a ***text file*** with a link to your Lab 7 notebook on github. Make

sure your lab notebook has been pushed to github. Your notebook should have:

- Screen shots from the Data Visualization with ggplot2 (Part 1) introduction on DataCamp
- Screen shots from the Multiple and Logistic Regression intrduction on DataCamp
- A pointer to your open source project on Observatory

7. (optional) Read the kaggle R tutorial on Machine Learning (Random forest is also discussed in chapter 4 of Zhao's book above.) (You can login to Kaggle with your facebook or google plus account or register for a new account.)