Data Science 3 Captone Proposal Directions

As we are approaching the end of the bootcamp, it is time to start thinking about what type or project you would like to do for your capstone. This is an opportunity to choose a direction and delve deeply into an aspect of data science that interests you. If you want to learn more about machine learning, this would be a great opportunity to do so, but note that having machine learning is *not* a requirement for the capstone project. This is also an opportunity to show off to potential employers, as this will be what you showcase on demo day.

For your capstone, you have your choice as to which technology to use. You may wish to use R and create a Shiny app like you did for your midcourse project, or you may want to learn more Python. You may even find it useful to use some SQL, but it is not required.

To choose your topic, you should think about an interesting question you would like to analyze and try to answer. You must either find at least two datasets to combine or choose a single dataset, which requires extensive cleaning and preparation. Kaggle datasets may not be used for the capstone without approval. I have included a list of websites to begin looking for some datasets that interest you.

Keep in mind that the goal of the project is to answer a question or to build a predictive model for some task. As you do you exploratory analysis, think about the overall question you are trying to answer.

The requirements for this project are to clean and prepare your data, perform exploratory analysis on it, and to create a presentation to communicate your findings. Similar to the midterm, should should create a PowerPoint or similar type of presentations, but be ready to show either a (well-polished) Jupyter notebook or R Shiny app. In creating this presentation, focus on good clear communication and storytelling.

For your proposal, complete the project_proposal_templete.docx file and submit as a pdf.

Due Date for Proposal: Sunday, April 26 **Capstone Roundtable:** Wednesday, May 6

Potential Dataset Sources:

- awesomedata: Awesome Public Datasets: **https://github.com/awesomedata/awesome-public-datasets**
- Datasetlist: Machine learning datasets: **https://www.datasetlist.com/**
- UCI Machine Learning Repository: **https://archive.ics.uci.edu/ml/index.php**
- 18 Places to Find Free Datasets: **https://www.dataquest.io/blog/free-datasets-for-projects/**
- Data is Plural Newsletter: **https://tinyletter.com/data-is-plural**
- Google Dataset Search: **https://datasetsearch.research.google.com/**
- Centers for Disease Control and Prevention: **https://data.cdc.gov/**
- United States Census: **https://data.census.gov/cedsci/**
- Zillow Data: **https://www.zillow.com/research/data/**
- Austin R. Benson's Datasets (including several graph datasets): **https://www.cs.cornell.edu/~arb/data/index.html**
- Registry of Open Data on AWS: **https://registry.opendata.aws/**
- Google BigQuery Public Datasets: **https://cloud.google.com/bigquery/public-data/**
- Grouplens: **https://cloud.google.com/bigquery/public-data/**
- IMDb: **https://www.imdb.com/interfaces/**
- Amazon Reviews: **http://jmcauley.ucsd.edu/data/amazon/**