DS 201 Final Project

The goals of this project are:

1. Apply the steps of the Data Science Process
2. Explain how to apply the steps of the Data Science Process
3. Perform EDA on a large dataset
4. Train a Machine Learning model
5. Work with GitHub
6. Make a professional presentation

Description

The final project will be a group project (up to 5 students) where students will turn in a tutorial that will walk users through the entire data science process: Business Understanding>Data Understanding> Data Preparation> Modeling+ Evaluating (EDA + Machine Learning Model Training)>Deployment.

You will need to create material for a tutorial that includes: Code, Text to explain the steps and the code, and a Video taking users through the entire process. The tutorial should be self-contained (e.g., code can download the data, do all the data preprocessing, perform all the EDA and ML model training and evaluation) a mix of Markdown text and Python or R code. Your code should have a link that allows users to run it in either Kaggle R or Python Kernels or Colab for Python. These files should be self-contained (automatically download the dataset and run all the analysis) so the user should just click and run (for example see this Jupiter notebook). You will host all this material on a GitHub statically hosted Page. You will also create a video taking users through the entire process, which has to be hosted on YouTube (as unlisted) but you will have the link on your GitHub page. Finally, you will need to create a video presentation of the work summarizing it in a professional manner.
This video should not be more than 5mins long, and it should present to decision-makers the importance of the problem you are trying to tackle, the insight you have found from your analysis, and the action that can be taken based on this insight. This video presentation has to be hosted on YouTube (as unlisted) but you will have the link on your GitHub page as well.

Students may choose an application area and dataset(s) of a **problem** that is of interest to them; please feel free to be **creative about this**! (the objective is to gain actionable insights about a problem that your team is interested in). Think about this final project as a mixture of Lab 3-5, all in one, but with the problem of your choice.

Submission

You will only need to submit the GitHub link where you are hosting the project. In this link user should find:

1. README file with Title of the project and team names, links of the code, video, and .io page. The file should also have a quick intro to the project, and the findings (with some images, as this)
2. Markdown text and Python or R code (i.e., R Markdown file with text &code, and/or Jupiter Notebook with text & code).
3. Link to run the code in the cloud (e.g., via Kaggle or Colab)
4. Link of YouTube Tutorial and Presentation Video

Grading

• Procedure (55%) – How well did you follow all the steps of the Data Science process and clearly show what you did in each step?

• Documentation (20%) –how well commented is the program? (e.g., GitHub README File, code comments, etc – the README file should look something like this, quick intro to the problem, links, few images, and findings)

•Tutorial (15%) - how well the process and the insights are presented?

• Presentation Video (10%) - how well did your team present your approach and results?

- There are a vast amount of freely accessible databases online (remember that for a unique problem you can mix multiple datasets), like the one in:

  - Google Dataset Search
  - US Government data
  - UCI Machine Learning Repository
  - Open Access Directory
  - Kaggle datasets
  - AWS Open Data Registry
  - Others….(public, federal, research datasets…)