edX

Course > Week 6… > Engage… > Mini-Pr…

Mini-Project Description

Mini-Project Description

This project allows you to use all that you've learned so far in this course to analyze one of the datasets you've already seen. We've allotted you one whole week to work on this project and you should expect to spend 5-10 hours total on the project. **Be sure to read all of the items below before starting the project**

- This project description

- The slides you will modify to present your work

- The rubric for evaluating your work (next reading)

There are a number of steps outlined below, but is critical that you do not view this as an entirely linear process. Remember that the "science" component in data science is the creation of a hypothesis based on exploration and testing of that hypothesis through analysis. You may need to go through many of these steps multiple times before you arrive at meaningful hypothesis or conclusions.

Deliverables: Although you should do your work in a Jupyter notebook, submitting the notebook itself is optional. The only piece you are required to submit is the PDF of your slides. Keep in mind you will be limited to a 5MB PDF so be judicious about including high-resolution images in slides. If both your slides and notebook both fit under 5MB, please feel free to include both in a single PDF file.

Steps

Step 1: Select a dataset we've already seen

A significant part of the data science process is finding, exploring, cleaning, and preparing the datasets. For that reason, we're having you select one of the datasets we've already used:

- Soccer Dataset

- IMDB Movie Dataset

- World Development Indicators Dataset

By using a dataset we've already explored, you can focus on the remaining preparation and cleaning, asking an appropriate research question, exploring the data more, and presenting your results. If you are aching to work with your own data or a dataset of interest, don't worry, you'll get a chance to do that in Week 9 and 10 when you start your Final Project.

Step 2: Continue to explore the dataset(s)

In this step, you'll continue to explore the dataset to see what kinds of questions it can answer. If you already have a good idea of what question you want to answer based on our exploration of the data together, feel free to advance to Step 3.

Step 3: Identify one research question

You likely had an idea of something you wanted to explore in the data when watching our videos. Go ahead and articulate your idea as a research question. The research question should be broad enough to be of interest to a reader but narrow enough that the question can be answered with the data. Some examples:

- **Too Narrow:** *What is the GDP of the U.S. for 2011?* This is just asking for a fact or a single data point.

- **Too Broad:** *What is the primary reason for global poverty?* This could be a Ph.D. thesis and would still be way too broad. What data will you use to answer this question? Even if a single dataset offered an answer, would it be defendable given the variety of datasets out there?

- **Good:** *Do science fiction movies tend to be rated more highly than other movie genres?* You can pull out the ratings by genre and see how they stack up to one another. You could also see if the distributions of ratings within genre are comparable across genres (e.g., maybe science fiction movies tend to be either highly or poorly rated, with little in between).

-

Step 4: Use appropriate methods to explore your data.

Based on your research question, you can now explore your data to answer the question. Use appropriate methods to answer that question. For example, if you are looking for a relationship between two items (say, CO_2 emissions and GDP), you may wish to use scatter plots and statistical correlation. Be sure to document, in your notebook, what you are exploring and why.

Step 5: Present your findings

In this step, you can begin to report your findings. What did you learn from the data and how do your findings help answer your research question? Use visualizations to present these findings.

Step 6: Present your work!

<u>Fill in this slide template to present your project</u>. Feel free to add more slides but aim to keep your entire presentation below 10 total slides. Optionally, you can attach a PDF of your Jupyter notebook to this presentation in one single PDF file to be uploaded.   As you are writing, please follow the guidelines below:

General Guidelines for Project Reports and Presentations:

1. This is academic writing. Keep it formal and coherent as a self-contained entity. Anyone reading your presentation should have a full understanding of the question, approach and the results.

2. Be professional. You should be comfortable giving this presentation to the general public, your boss, or your academic advisor.

3. Write for a diverse audience including:

   1. **General public:** Reads only the title and abstract looking for high-level point mainly for conversational purposes.

   2. **A company CEO:** Reads introduction, research question, findings and conclusions looking for business value and details related to that.

   3. **An academic advisor (or company CTO):** Reads the full presentation AND your Jupyter Notebook paying particular attention to technical coherence, academic value, and technical data science strengths.

Presentations should be for the above-mentioned three audiences. Think of the diversity of the audience. The whole point is to tell a story - so you should motivate a reader to care based on the question you are exploring, answer that question in a clear and concise manner, provide an honest appraisal of your

results, and give the reader valuable insights. Use charts whenever possible. Avoid slides with a lot of text and bullet points - break the slide into multiple slides when this happens. Be concise!