DS Web Application Project

Multiple rolling deadlines, see below.
Final deliverables are due by Noon, 12/07/20

For the final project, you and your team members will develop a database-backed website that provides the results of a data science project of your choosing. Fill in the team signup form (one submission per team) before Noon, Sunday, 11/08 and we will send you a Google Folder to store some of your submissions.

Students that have taken a standard Data Science course will be familiar with the tasks associated with a data science project. Most projects typically involve the following:

1. Gathering data
2. Processing data
3. Performing exploratory analysis
4. Exploring various machine learning and statistical approaches
5. Reporting of final results and process documentation

For the DATA 1050 project, teams will create a live data-science web application. This application will allow users to interactively view aspects of the project's EDA and do something creative and useful.

Minimal Requirements and Suggested Enhancements

Requirements are divided into **two aspects**: Data Science and Data Engineering.

Data Science **requirements** for the project are as follows:

1. Dataset: You are required to make use of an online data source that is either being incrementally updated (for example, weather data or sports data) or so large that user-specific queries need to be applied against it (e.g., Google Search APIs, Genomic Databases).
2. EDA, ETL: You should develop an **EDA_ETL.ipynb** notebook or notebooks that document your data exploration and data cleaning process. It should include a section on the raw data and another section on the final data.
3. Your project should include Dash/plot.ly based interactive visualization of the data. You should prototype your visualization using a notebook named **Visualization.ipynb**
4. Your project should also do *something useful* with the data. For example, it could try and make a prediction based on the present. Alternatively, it could try and summarize the data in a new and useful way. Prototype your useful enhancement to the data in a notebook called **Enhancement.ipynb**

Suggested Data Science specific **enhancements**:

1. Datasets: Many interesting data science projects make use of *at least two* data sets. For example, joining a primary dataset with relevant geospatial weather, census or economic data can provide valuable insights.
2. Predictions: If your project includes predictions we suggest the following enhancements:
    a. Create a simple baseline model and use it to evaluate your efforts as you create more complex models. Include the results of your model and the baseline model.
    b. Include a summary of the results of past predictions. If you are updating your model incrementally, be sure to avoid violating causality as you do this.

Data Engineering **requirements** for the project are as follows:

1. At least one dataset is live (i.e., additional data is becoming available or a user query is used to pull additional data)
2. Results from web API calls or data-scraping are cached in **raw form** on disk (**or database**) as much as possible
3. Appropriate ETL is used on **raw form** results is used to populate and update a database
4. Creation of at least one Interactive Dashboard that showcases the project's data and additional functionality. Generation of Interactive Dashboard(s)should be done using appropriate queries to the database.
5. Data updates will be retrieved incrementally and also applied incrementally to the backing database(s).

The caching and incremental update requirements are typical for live data science applications. The use of a database as a data store is typical in environments where the data is being used by multiple users or processes.

Specifically, caching raw form data will speed up the development process because the application will not need to re-request data each time the ETL component is run. It will also limit the number of web data requests made by the application. This is considered good practice since each data request requires a certain amount of system resources from the data provider. Many websites will also block IP addresses and or API users that make too many requests. Appropriate use of the database will allow incremental updates to occur without end-users potentially seeing inconsistent results.

Suggested Data Engineering specific ***enhancements***:

1. GCP, AWS or another service like databricks.com or heroku:
    a. To retrieve or store your data
    b. To retrieve, store and process your data
2. Static Website:
    a. Design your application so that each time it updates, it creates a new static version of the website
    b. Push this site to a free hosting platform like github.io
3. Dynamic Website:
    a. Host your website on GCP or other cloud provider (AWS, Azure, Heroku)

Additional Resources

The course staff is releasing some Example Project components that teams are free to reuse.

BPP Energy Application Example

Students in DATA 1030 are also free to reuse materials they have developed for their projects in that course. A post on additional tips and additional resources will be maintained on Piazza. (Students are encouraged to contribute suggestions and updates to this post.)

Project Team Setup, due not later than Sunday, 11/08, Noon

One member of each team should use this form to provide the email address for each member of their team, and their team's name. Teams of 4 students are recommended, but teams of 3 students are allowed.

Project Outline, due no later than Sunday, 11/15, Noon

Submit your proposal by creating a Google Document called **Project Outline** in your team's Google Folder. The course staff will try and provide feedback (in the form of document comments) within 48 hours of when you *submit* your proposal. ***If you submit your proposal outline early, you will get feedback early.**
Send an email to Professor and the HTA when you have done so.

The purpose of the outline is to identify data sources and an outline of what your team wants to do with them. Your direction may change based on what you discover as you proceed with your project and also on feedback on your proposal - **both are okay.**

Specifically, in your proposal, we are looking for outlines on how you plan to do the following:

- **The name of your project**
- **The name of your team, and the name of each team member**
- **Location of your team's git/gitpod repo:**
    - You are expected to actively use Github as you develop your project.
- **Vision/Summary**
    - Provide an Executive Summary
    - What is your "big idea"?
    - What might you find at the end of your project?
- **Data**
    - What datasets do you plan on using?
    - How big are they?
    - How do you plan to collect the initial dataset?
    - How will incremental updates be collected?
- **Previous Work**
    - Provide an annotated bibliography to previous work by others on your problem and or data (try and find at least three paper references or websites)
- **Methodology**
    - What do you plan on doing with your data?
    - What techniques do you think you will use to analyze the data?
    - How might you visualize your results?
    - Identify at least one base-line model.
- **Visualization idea:** Your system must include ways to meaningfully summarize its inputs and (and possibly results) in a graphical manner using plot.ly and Dash.
- **Enhancement idea:** Describe how you will use the data in a creative way.
- **Next Steps:** Include a clear description of your team's next steps. How and who will gather the initial data? Who will research plot.ly for use with Python and in Notebooks? Who will design your database schema and work on the ETL component? Who will research Dash and look for interesting examples that you can pattern your final website after?

Project Check, due no later than Noon, Wednesday, 11/25/20

Data sets should have been collected and cleaned, using your ETL_EDA notebook. Your Prediction or enhancement notebook should also be mostly complete. Create a Google Document called '**Project Check'** in your Google Folder. At the beginning of your document please include a section called **Current Status** that explains clearing where your team is in terms of creating your web application. This section should include links to viewable versions of your **ETL_EDA.ipynb**, **Visualization.ipynb** and **Enhancement.ipynb** notebooks. If you are using colab to work with these notebooks please provide a link to them there.

Your Project Check document should then include an optional **Help Request** section that includes a description of all areas your team would like help with from the course support staff.

The remainder of the Project Check document should follow the general format of your team's Project Outline but updated with everything using your current dataset, progress, ambitions, etc. It should finish with a detailed Next Steps section that explains what each team member will be doing in order to complete the project.

***The sooner you submit your Project Check the sooner the course staff will try to assist with your Help Request(s).**

Website, due Noon, Sunday, 12/06/20

Your website will act as the platform for your final report. In addition to **one or more interactive dashboard pages**, your site should have an About page which provides the following details on your efforts:

About

- Project & Executive Summary
- Names of all team members
- Possible next steps
- References to related work

Your site should also have an additional page or page(s) providing full project detail divided up into the following sections:

Project Details

- Datasets used
- Development Process and Final Technology Stack
    - Explain how you created the site, and the final technology stack used
- Data Acquisition, Caching, ETL Processing, Database Design
    - Describe how the data is accessed, cached and ETL processing steps
    - Describe the Database used (include a schema diagram if appropriate)
- Link to a static version of your ETL_EDA.ipynb notebook, or equivalent web page
- Link to a static version of your Enhancement.ipynb notebook, or equivalent web page

The About page and Additional Project Details page should allow interested users to fully understand how your project was developed and how it currently works. You are encouraged to use content from your Project Check document and notebooks when creating these sections of your website.

Screencast, due Noon, Monday, 12/07/20

- Your team should jointly create a screencast (all members should speak) that begins with a review of the first few sections of the About page and then gives an interactive demonstration of the site's dashboard functionality
- The screencast should then cover the remaining content in the About page and Project Details in whatever order you think leads to the best narrative
- Your screencast should be between 8 and 15 minutes long
- You should create one powerpoint (or equivalent) slide that will be displayed at the beginning of your presentation; it should include a bit.ly link to your project that will allow interested viewers (and graders) to follow along with your presentation. Additional slides are fine to create but are not required.
- Audience: You may assume your audience is familiar with data science basics but be sure to include clear definitions for any application-specific terms.

Grading Rubric

Project Outline	10 %
Project Check	10 %
Website and final codebase	60 %
Screencast & QA	20 %