Final Project/Tutorial

CMSC 320 - Introduction to Data Science
Spring 2025
Instructor: Dr. Fardina Alam

Summary

For your final semester long group project, you will publish a tutorial that will walk users through the entire data science pipeline: acquiring and curating data, parsing the data into a queryable format, exploratory data analysis, hypothesis testing and machine learning, and explaining the results with words and visualizations. This project is designed to demonstrate your data science knowledge and contribute to your portfolio.

Deadlines

- **Project posted:** Thursday, February 6, 2025
- **First deliverable:** Tuesday, February 25, 2025
- **Second deliverable:** Tuesday, April 1, 2025
- **Final deliverable:** Tuesday, May 8, 2025

Late submissions will **not** be accepted for the final deliverable.

Rubrics

The final project is graded out of 100 points. Each deliverable must be submitted on **Gradescope as a group**. Working alone is permitted. We recommend four people per group. The maximum group size is 6 people.

IMPORTANT: If any kind of plagiarism is found, including the use of AI tools like ChatGPT or direct code from previous semesters, you will receive a grade of 0 for the final tutorial. To cite the python libraries you use, simply include the import code block at the top of the tutorial.

Checkpoint 1 (5 points)

Submit a **pdf** on Gradescope with:

- **(1 point)** A **link** to your Github repository.
- **(2 points)** **What** datasets are you choosing? Cite the source(s).
    - The dataset should be large enough and appropriate for making your analysis.
- **(2 points)** **Why** are you choosing this dataset?

Checkpoint 2 (25 points)

Submit a **jupyter notebook file (.ipynb)** on Gradescope with:

- **(5 points) Data preprocessing:** (a) import, (b) parse (e.g., convert strings to ints), (c) organize (e.g., set up a database or a pandas DataFrame).
- **(20 points) Basic data exploration and summary statistics**
  - You must present three conclusions using at least three different statistical methods including hypothesis testing.
    - For example: What are the main characteristics of your dataset? How many features and entries are there? Is a feature over-represented? Are features correlated? Are there outliers? Identify the attributes that will affect your choice of primary analysis technique. Etcetera.
  - For each method, you must have at least one **gorgeous** plot.

Checkpoint 3 - Final deliverable (70 points)

Submit a **single URL** pointing to your final tutorial to Gradescope. The tutorial should be self-contained, a mix of Markdown prose and Python code, and delivered as a GitHub statically-hosted Page (see the Publishing section below for instructions). The format of the final tutorial is given at the end of this section.

- **(15 points) Formatting and prose.**
  - Follow the format given at the end of this section. Use section headers.
  - For each section, include a clear explanation of what you are doing. We will be checking the entire tutorial, including the parts you submitted in previous checkpoints.
  - Write code that is documented, well-organized, and reproducible.
    - Does it help the reader understand the tutorial?
  - Cite your sources: Link to other resources that would: (a) give a lagging reader additional help on specific topics, and (b) give an advanced reader the opportunity to dive more deeply into a technique or idea.
- **(30 points) Machine learning analysis** that will help you answer the questions you posed in the introduction.
- **(10 points) Visualization** based on your results.
- **(15 points) Insights and conclusions**.

The format of the final tutorial (final deliverable) should be as follows:

1. **Header with contributions.**

   Title
   Spring 2025 Data Science Project
   Your name(s)

   Contributions:
   For each member, list which of the following sections they worked on, and summarize the contributions in 1-2 sentences. Be specific!
   A: Project idea
   B: Dataset Curation and Preprocessing
   C: Data Exploration and Summary Statistics
   D: ML Algorithm Design/Development
   E: ML Algorithm Training and Test Data Analysis
   F: Visualization, Result Analysis, Conclusion
   G: Final Tutorial Report Creation
   H: Additional (not listed above)

2. **Introduction.** The introduction should motivate your work: what is your topic? What question(s) are you trying to answer with your analysis? Why is answering those questions important?
3. **Data curation.** Cite the source(s) of your data. Explain what it is. Transform the data so that it is ready for analysis. For example, set up a database and use SQL to query for data, or organize a pandas DataFrame.
4. **Exploratory data analysis.** (See checkpoint 2.)
5. **Primary analysis.** Based on the results of your exploration, choose a machine learning technique (e.g., classification, regression, clustering, etc.) that will help you answer the questions you posed in the introduction. Explain your reasoning.
6. **Visualization.** Explain the results and insights of your primary analysis with at least one plot. Make sure that every element of the plots are labeled and explained (don't forget to include a legend!).
7. **Insights and Conclusions.** After reading through the project, does an uninformed reader feel informed about the topic? Would a reader who already knew about the topic feel like they learned more about it?

If needed, you may add subsections.

Publishing

GitHub provides a service called Pages (https://pages.github.com/) that provides website hosting functionality backed by a GitHub-based git repository. We would like you to host your final project on a GitHub Pages project site. To do this, you will need to:

1. Create a GitHub account (or use the one you already have) with username <username>.
2. Create a git repository titled username.github.io; make sure username is the same as whatever you chose for your global GitHub account.
3. Create a project within this repository. This is where you'll dump your iPython notebook file and an HTML export of that notebook file.

These instructions are also given on the front page of https://pages.github.com/ and Annie Zhou, a past student, created a tutorial: https://azhou4847.github.io/CMSC320ProjectPublication/. The deliverable to the CMSC320 staff will then be a single URL pointing to this publicly-hosted GitHub Pages-backed website.

Dataset Ideas

Choose an application area or dataset that is of interest to you. Please feel free to be creative! Remember that you can use API calls or scrape websites for data.

There are lots of options for large, open-access datasets that could yield multiple insights, especially from governmental agencies. Here are some examples:

- Canada has published some pretty interesting statistics – check the left hand side for filtering options: https://www.statcan.gc.ca/en/start. One of their datasets is the retail price of products over time: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810024501
- FEC campaign finance data: https://www.fec.gov/data/
- USGS earthquake data: https://earthquake.usgs.gov/earthquakes/search/

You could also curate a personalized dataset. Keep in mind that the dataset needs to be large and varied enough for machine learning analysis.

- Is your data being tracked by a website? For example, YouTube watch history or Goodreads book histories. Check whether you can export it. (Only do this if you're willing to discuss the trends within – not everyone is comfortable with that).
- What are your interests? What do you like to watch? Is there a subreddit for it that hosts a survey every year, or a fan-made spreadsheet?
- Generate your own dataset from audio or video data, e.g. https://www.reddit.com/r/educationalgifs/comments/64rjns/i_did_a_center_of_mass_analysis_of_a_triple/
    - Relatedly, what have you learned about in your other classes that you would want to demonstrate through data analysis?

Previous Examples

Here are some examples of good final tutorials from previous sections of CMSC 320:

- The Effect of Storms in the United States, https://shahsean.github.io/
- An Evaluation of American Presidential Elections, https://jcurran0499.github.io/
- Analysis of S&P 500 Companies, https://neo-zhao.github.io/
- Predicting Dementia and Alzheimer's, https://amygracecruz.github.io/
- Does the University of Maryland Computer Science Department need more faculty?, https://krixly.github.io/
- Analysis of Crime Data at UMD, College Park, https://andresgogo.github.io/

DO NOT COPY THEIR PROJECTS OR CODE.