Final Project/Tutorial

DATA 602 -Principle of Data Science
Fall Semester 2025
Section - PSC2
Instructor: Dr. Fardina Alam

Summary

For your final semester-long group project, you will develop a comprehensive tutorial that guides users through the entire data science pipeline. This includes acquiring and curating data, parsing it into a queryable format, performing exploratory data analysis, conducting hypothesis testing and machine learning, and explaining results through both written descriptions and visualizations. You will also compile a detailed report and a short presentation on your project. This project aims to showcase your data science expertise and enhance your professional portfolio.

Deadlines

- **Project posted:** Friday, September 19, 2025
- **First deliverable:** Friday, October 3, 2025
- **Second deliverable:** Tuesday, October 28, 2025
- **Final deliverable (Project Report + Tutorial + PPT):** Tuesday, December 16, 2025
- **Presentation Day:** December 17, 2025 1:30PM to 4:30PM https://app.testudo.umd.edu/soc/exam/search?courseId=&sectionId=&termId=202508

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

Rubrics

The final project is worth 150 points, accounting for 30% of your total course grade. All deliverables must be submitted as **a group on Gradescope**, with **only one** submission per group. Ensure that the submitting member **includes the names of all group members**, as missing names will result in point deductions. Although individual work is permitted with prior instructor approval, we recommend **forming groups of three people**. However, if you choose to create a group of four to five people, ensure that your group's contribution reflects this size accordingly.

IMPORTANT: Any form of plagiarism, including the use of AI tools like ChatGPT or code from previous semesters, will result in a grade of 0 for the final tutorial. To properly cite the Python libraries you use, simply include the import code block at the top of your tutorial.

Checkpoint 1 (5 points): Due Oct 3, 2025

Submit a **pdf** on **Gradescope** with:

- **(1 point)** A **link** to your Github repository(include your Project name, topic, and group member name there).
- **(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): Due Oct 28, 2025

Submit a **jupyter notebook file (.ipynb)** on **Gradescope**, which should include the following components:

- **(10 points) Data preprocessing and Data Cleaning:** You are free to choose any dataset you want. (a) import your chosen dataset, (b) parse and transform as needed (e.g., convert data types, extract features, format conversion), (c) organize into appropriate data structures (e.g., set up a database, pandas DataFrame, or other suitable format), (d) clean the data (address missing values, correct inconsistencies, remove duplicates, and handle data quality issues specific to your dataset to ensure it is ready for analysis).

- **(15 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 used, include at least one **visually appealing plot** to support your findings.

Checkpoint 3 - Final deliverables (55+50+15 = 120 points):

Due Dec 16, 2025

Submit a **pdf** of your project report including a *single URL* pointing to your final tutorial in the first page of your report, and your **powerpoint presentation** on **Gradescope.**

The final deliverable consists of **three** components:

1. **Your Final Data Science Tutorial/Project on GitHub:** This should include the complete tutorial and code for your data science project.
2. **Project Report:** A detailed report documenting your methodology, analysis, and findings.
3. **Presentation:** Submit a PowerPoint presentation summarizing your project.

1.Data Science Tutorial/Project on GitHub (55 points):

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.

- **(10 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.
- **(25 points) Primary/Machine learning analysis** that will help you answer the questions you posed in the introduction. Specifically, your analysis should include a comparison of the current day's data with your model's predictions, which are based on data from previous days. This comparison should clearly illustrate how well your model performs by showing both the predictions and the actual data.
- **(10 points) Visualizatio**n based on your results. Ensure that all elements of the plot are labeled and explained, including a legend.
- **(10 points) Insights and conclusions**. Assess whether an uninformed reader would gain a clear understanding of the topic, and whether a knowledgeable reader would learn something new.

Remember that, the final tutorial should include all components from the previous checkpoints. Your third checkpoint submission should build on and continue from the work done in the earlier checkpoints. If you have made any improvements or adjustments based on earlier feedback, clearly describe these changes and provide justifications in your tutorial.

2. Project Report (50 points):

Create a detailed project report using **Overleaf**. The report should be approximately **8 pages in length, formatted in 11pt font with one-inch margins all around, excluding the first header page and the second page detailing contributions**. It should include a general background on your approach to handling the data, a detailed description of your algorithm implementation, and a comprehensive presentation of your findings, supported by **diagrams (your model diagram), graphs, and other relevant visuals**. For more details, follow the format given at the end of this section. Use section headers.

The format of the final project report/github tutorial should be as follows:

1. **First Page:**
   In the first page of your project report, include the following information in this order:
   - Course Code and Name: [Insert course code and name here]
   - Semester and Year
   - Instructor Name: [Insert instructor's name here]
   - Group Project Name: [Insert the name of your group project here]
   - URL to Final Tutorial: [Insert the single URL pointing to your final tutorial on GitHub here]
   - Group Members: [List the names of all group members here]
   - Date of Submission: [Insert the date of submission here]

2. **Second Page:**
   Include a header titled Contributions.For each group member, specify their contributions by listing the relevant sections and summarizing their work in 2-3 sentences.
   Be specific about their roles and contributions!
   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)

3. **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?
4. **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.
5. **Exploratory data analysis.** (See checkpoint 2.)
6. **Primary/Machine Learning analysis.** Based on the results of your data exploration, select a machine learning technique (e.g., classification, regression, clustering, or neural networks, etc.) that will effectively address the questions posed in the introduction. Provide a clear explanation for your choice, detailing how this technique aligns with your analysis objectives and how it will help answer the key questions.
7. **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!).
8. **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?
9. **Data Science Ethics:** Address any ethical considerations related to your project, such as potential biases in data collection or analysis. Discuss how you mitigated these concerns and ensured that your analysis is fair and transparent.

If needed, you may add subsections. You also need to add **relevant screenshots** in the project report.

2. Presentation Format and Instructions (15 points):

You will prepare and submit a PowerPoint presentation that highlights the main aspects of your project. The presentation should be clear, well-structured, and concise, with a maximum duration of **6 minutes,** followed by **4 minutes of Q&A.**

- **(1 point) Group Introduction:** Briefly introduce your group members and describe the role each member portrayed in the project .
- **(2 points) Problem Summary:** Provide a clear and concise summary of the problem you selected. Ensure it is understandable to both technical and non-technical audiences.
- **(5 points) Model & Rationale:** Explain the model you chose, why you selected it, and how it addresses the problem. Keep explanations simple, with just enough detail to show your reasoning.
- **(5 points) Results Visualization:** Present your results using clear and effective visualizations (charts, plots, or tables). All visual elements should be labeled, explained, and easy to interpret.
- **(2 points) Conclusion & Future Work:** Summarize the main takeaways from your project. Suggest possible improvements, extensions, or future directions.

The presentation should demonstrate that your group can communicate your project effectively to both informed and general audiences. Slides should be visually engaging and emphasize clarity, use visuals and concise bullet points rather than text-heavy content. You must include a summary contribution slide at the end of your presentation that lists each team member's name and their contributions to the project. This slide does not need to be presented during the 6-minute presentation time. Simply show it at the end of your presentation after concluding your main content.

IMPORTANT POINTS:

Your GitHub tutorial should include all specified sections with precise writing, detailed project content, and a focus on code and visualizations, with minimal text.

In contrast, your project report should provide a detailed narrative that thoroughly explains your methodology, analysis, and findings. The report should include screenshots of your analysis, visualizations, training details, and any other relevant elements to offer a comprehensive understanding of your work.

While both the GitHub project and the report should follow the same format, the GitHub tutorial will emphasize concise coding and visualizations, whereas the report will provide an in-depth written explanation.

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//. The deliverable to the DATA 602 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 well-executed final tutorials from the undergraduate version of this course:

- 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/

These examples can provide initial ideas, but remember that this is a graduate course, so your work should meet the higher expectations accordingly.

DO NOT COPY THEIR PROJECTS OR CODE.