Data 102 Final Project, Spring 2024

Due Date: Monday May 6, 2024 at 11:59 PM

In this project, you will complete a guided analysis for a dataset of your choice. We have curated a list of suggested datasets, but you are welcome to select an external dataset. Your analysis should include the following steps:

1. **Data Overview** Describe and discuss your dataset.
2. **EDA** Perform exploratory data analysis (EDA) and describe key features of your dataset. You may want to complete this step before deciding on a research question below.
3. **Research Questions** List **two** research questions that you will explore in this project. Between the two research questions, you should use at least two of the following four techniques that you've learned this semester. You may use more than two techniques, and you may use more than one technique for any particular question (see Section 1 for examples).
   - Binary decision-making and hypothesis testing
   - Bayesian hierarchical modeling
   - Comparing generalized linear models (GLMs) to nonparametric methods for prediction
   - Causal inference

   At least one of your techniques should be either **Bayesian hierarchical modeling** or **causal inference**. You should use the same dataset to answer both questions. Please see Section 1 for examples and clarification.
4. **Inference and Decisions** Apply the two techniques you chose above to answer your research questions, explaining your choices.
5. **Conclusion** Highlight key findings, identify potential next steps, and assess the strengths and limitations of your analysis.

You must work in groups of four, and you must fill out the project group form by Wednesday, March 20. If you don't have a group, you must fill out the same form the same form, and you will randomly be assigned a group. If you do have a group, **one person must fill out the same form** by the deadline to declare your group. In very special circumstances (e.g., extenuating personal circumstances or ongoing personal project such as a senior thesis), we will allow students to work alone. If you believe you qualify for this exception, please email data102@berkeley.edu ASAP with relevant information/documentation. Do **not** assume the exception has been granted until you receive a confirmation email. Please note that you will be evaluating your group members at the end of the project: we strongly encourage you to discuss and resolve any conflicts with your group members sooner rather than later.

Detailed guidelines are provided below. Please read through this entire document before you begin working!

Contents

1 Research Question Examples

Here are some examples of research questions on hypothetical datasets. Note that all of them use either Bayesian hierarchical modeling or causal inference, all of them use at least two of the techniques described, and all of them answer at least two research questions using the same dataset.

- If you were looking at a dataset of Data 102 students, you might choose as your research questions (1) does attending office hours cause an improvement in homework grades (causal inference), and (2) Can we fit a Bayesian hierarchical model to the distributions of assignment grades by student year (Bayesian hierarchical modeling).
- If you were looking at a dataset involving jellybean consumption, acne, and other demographics, you might choose as your research questions (1) does consuming different colors of jellybeans cause acne (causal inference *and* multiple hypothesis testing), and (2) predicting jelly bean consumption from personal demographics, using negative binomial regression and random forests (prediction with GLMs and nonparametric methods).
- If you were looking at a dataset involving characters in a TV show (lines of dialogue, gender, age, etc.), you might choose as your research questions (1) how well does character demographic information predict lines of dialogue for each season, comparing GLMs to nonparametric methods (prediction with GLMs and nonparametrics); and (2) for each season of the show, is there a significant association between gender and lines spoken (multiple hypothesis testing).

2 Section Guidelines

Your report should include each of the following sections, and address the listed questions at minimum. You should include additional, relevant discussion to each section that is specific to the features of your dataset.

Depending on your research questions, you should choose at least two of the corresponding sections for options A through D.

2.1 Data Overview

- How were your data generated? Is it a sample or census?
- If you chose to use your own data, describe the data source and download process.
- If you chose to add additional data sources, explain why.
- If your data represents a sample:
  - Compare the distribution of one of your variables to what is expected in the population. For example, if your data has an age variable, compare it to the age structure of the population.
    - Do you notice any differences?
    - How does this affect the generalizability of your results?
- If your data represents a census:
  - Are there any groups that were systematically excluded from your data?

- To what extent were participants aware of the collection/use of this data?
- What is the granularity of your data? What does each row represent? How will that impact the interpretation of your findings?
- Are any of the following concerns relevant in the context of your data?
  - Selection bias
  - Measurement error
  - Convenience sampling
- Was your dataset modified for differential privacy? If so, explain how.
- Are there important features/columns that you wish you had, but are unavailable? What are they and what questions would they help you answer?
- Are there columns with missing data? What do the missing entries mean? How will you deal with this in your analysis?
- What cleaning or pre-processing did you apply to the data and why? How will these decisions impact your models and inferences?

2.2 Research Questions

- Your research questions should involve using the methods mentioned above (i.e., the ones you've learned in Data 102) to answer them. For each of your two research questions, describe:
  - What is the research question? What real-world decision(s) could be made by answering it?
  - Explain why the method you will use is a good fit for the question (for example, if you choose causal inference, you should explain why causal inference is a good fit for answering your research question).
  - What are the limitations of the method you chose? Under what circumstances might it not do well, and what could go wrong under those circumstances?

2.3 EDA

- Visualize at least two quantitative variables and two categorical variables. Your visualizations must be relevant to your research questions!
- Describe any trends you observe, and any relationships you may want to follow up on.
- Explain how your visualizations should be relevant to your research questions: either by motivating the question, or suggesting a potential answer. You must explain why they are relevant.

2.4 Option A: Multiple hypothesis testing / decision making

Test at least six different hypotheses, correctly calculating p-values for each one, and use two different multiple hypothesis testing correction techniques to control error rates.

- **Methods**

  - Describe the hypotheses that you'll be testing using your dataset, and explain why it makes sense to test many hypotheses instead of just one to answer your question.
  - For at least one of the tests, describe a specific alternative hypothesis, and compute the power of the test you plan to use.
  - Describe how you'll be testing each hypothesis (A/B test, correlation/association, etc.), and justify your choice.
  - Describe at least two different ways you'll correct for multiple hypothesis tests, and explain the error rates being controlled.
  - Compare and contrast FWER control and FDR control for your tests: specifically, use one method that controls FWER and one that controls FDR, and compare the number of discoveries made by each. Explain which is more appropriate for your research question.
- **Results**
  - Summarize and interpret the results from the hypothesis tests themselves.
  - For the two correction methods you chose, clearly explain what kind of error rate is being controlled by each one.
- **Discussion**
  - After applying your correction procedures, which discoveries remained significant? If none did, explain why.
  - What decisions can or should be made from the individual tests? What about from the results in aggregate?
  - Discuss any limitations in your analysis, and if relevant, how you avoided p-hacking.
  - What additional tests would you conduct if you had more data?

2.5 Option B: Bayesian Hierarchical Modeling

Formulate a hierarchical Bayesian model, similar to the models you saw in lecture, lab, and homework (e.g., kidney cancer data from Lecture 7, school funding model from HW2, COVID transmission model from Lab 4). Use this model estimate parameters of interest from your dataset. You must implement your model in PyMC and present the results.

- **Methods**
  - Draw a graphical model, clearly indicating which variables are observed. Provide descriptions of any hidden variables you're trying to estimate.
  - Clearly describe the groups in your dataset, and explain why a hierarchical model is a good choice for modeling variability across the groups.
  - Justify and explain your choice of each prior and conditional distribution in the model (Gaussian, Beta, etc.).
- **Results**
  - Explain how you chose the hyperparameters for your prior distribution: did you use empirical Bayes? A full hierarchical model? Why?

- Summarize and interpret your results. Are there any counterintuitive findings or surprises?
- Provide a credible interval for at least one hidden variable in the model, and provide clear quantitative statements of the uncertainty in plain English.

- **Discussion**
  - Elaborate on the limitations of your methods.
  - If your inference procedure had trouble converging, can you explain why?
  - Did you try other formulations / graphical models? If so, what worked or didn't work about each one?
  - What additional data would be useful for answering this question, and why? How would you add it to the graphical model?

2.6 Option C: Prediction with GLMs and nonparametric methods

Set up a prediction problem, and use **both a generalized linear model (GLM) and a nonparametric model** to carry out your prediction. You should identify the best choice of link function/likelihood model for your GLM, and present results from either the frequentist or the Bayesian perspective depending on your preference. Similarly, you should also choose one or two nonparametric methods (nearest neighbors, decision trees, random forests and neural networks) for prediction, and compare the results.

- **Methods**
  - Describe what you're trying to predict, and what features you're using. Justify your choices. If there are additional features that are relevant to the prediction and are readily available in public datasets, you must include at least one such feature.
  - Describe the GLM you'll be using, justifying your choice. Describe any assumptions being made by your modeling choice. If you are using a prior for the coefficients of the Bayesian GLM, explain why you chose the prior you did.
  - Describe the nonparametric method(s) you'll be using, justifying your choice. Describe any assumptions being made by your modeling choice.
  - How will you evaluate each model's performance?

- **Results**
  - Summarize and interpret the results from your models.
  - Estimate any uncertainty in your GLM predictions, providing clear quantitative statements of the uncertainty in plain English.

- **Discussion**
  - Which model performed better, and why? How confident are you in applying this to future datasets?
  - Discuss how well each model fits the data.

- Interpret the results from each model: based on the parameters of the model(s), can you make any general statements about the relationships between the outcome and the features? You may choose to not provide interpretations, but you must justify this choice.
- Elaborate on the limitations of each model.
- What additional data would be useful for improving your models?
- Explain the uncertainty in your results: is it qualitatively high or low? Explain why this might be (e.g., noisy data, dataset size, variance in the estimation, etc.).

Note that you will not be graded on model accuracy: if your accuracy is high, you should explain why the prediction problem is relatively easy, and if it's low, you should explain why the problem is relatively hard.

2.7 Option D: Causal Inference

Formulate a causal question, clearly defining the treatment, control, and units (people, states, months, etc.). Use one of the techniques you learned in class to answer the question, clearly stating and justifying any and all assumptions you make.

- **Methods**
  - Describe which variables correspond to treatment and outcome.
  - Describe which variables (if any) are confounders. If the unconfoundedness assumption holds, make a convincing argument for why.
  - What methods will you use to adjust for confounders?
  - Are there any colliders in the dataset? If so, what are they?
  - Draw the causal DAG for your variables.
- **Results**
  - Summarize and interpret your results, providing a clear statement about causality (or a lack thereof) including any assumptions necessary. In addition to statistical significance, discuss the magnitude of any effect you find.
  - Where possible, discuss the uncertainty in your estimate and/or the evidence against the hypotheses you are investigating.
- **Discussion**
  - Elaborate on the limitations of your methods.
  - What additional data would be useful for answering this causal question, and why?
  - How confident are you that there is (or isn't) a causal relationship between your chosen treatment and outcome? Why?

2.8 Conclusions

- Summarize your key findings.
- How generalizable are your results? How broad or narrow are your findings?

- Based on your results, suggest a call to action. What interventions, policies, real-world decisions, or action should be taken in light of your findings?
- Did you merge different data sources? What were the benefits and/or consequences of combining different sources?
- What limitations are there in the data that you could not account for in your analysis?
- What future studies could build on your work?
- What did you learn during the project?

3 Project Deliverables

You must work in groups of four. Each group will submit one set of the following deliverables, submitted to Gradescope as a zip file.

3.1 Project Proposal (Due Friday, April 5)

By **Friday, April 5 at 11:59PM**, please fill out the Project Proposal Form (more info will be provided on Ed) to indicate the dataset and research questions you'll be working on.

If you plan to use any methods beyond what you learned in Data 102, please include this in your proposal (or let us know on Ed if you decide to do so after the proposal is due): note that course staff will be less equipped to help you if you choose to do so.

Your research questions should be well-defined. Course staff will respond to your proposal by Wednesday, April 10 with feedback: you must incorporate this feedback in order to receive full credit on the project!

The proposal (and both checkpoints) will be graded on a credit/no-credit basis.

3.2 Checkpoint 1: EDA (Due Friday, April 18)

By **Friday, April 18 at 11:59PM**, you must submit a draft of the EDA section of your project. If (and only if) you address all the criteria in the EDA section above, you will receive full credit on the checkpoint.

You are free to change your EDA section or add (or remove) content between the checkpoint and your final submission. Course staff will not provide any feedback on the EDA checkpoint.

3.3 Checkpoint 2: Research Question Results (Due Friday, April 25)

By **Friday, April 25 at 11:59PM**, you must submit a draft of your results for at least one research question (we recommend trying to have a draft of both done by this time). If (and only if) you address all the criteria in the corresponding Methods and Results section above, you will receive full credit on the checkpoint. Note that you do not need to complete the Discussion section for this checkpoint.

You are free to change your results section, or add (or remove) content between the checkpoint and your final submission. Course staff will not provide any feedback on the research question checkpoint.

3.4 Written Report

You must submit a typed PDF document that contains each of the sections described in "Section Guidelines". Your report should be between 3000 and 5000 words of text, in addition to tables, figures, and references. All mathematical equations must be rendered properly in LaTeX, Equation Editor, or similar. We won't be strictly enforcing this limit, but reports that are much longer than this are subject to a penalty (reports that are much shorter are probably missing important discussion).

Your report should be a proper written document: you cannot just submit a printed Jupyter notebook (including data sources, code, outputs, etc.) We highly recommend using Overleaf or Google Docs. For your convenience, we will post a list of LaTeX resources on Ed.

If relevant, include a reference page with citations of all outside sources used.

All figures and tables should be included in your written report. Clearly label all figures and include informative captions. These labels should be used to reference figures and tables in your written report. Refer to this guide for instructions on inserting images into a LaTeX file.

3.5 Jupyter Notebook

You must submit a single notebook that contains all the code run for the project. Your code should be clear and well-documented. Please label each section of code in markdown.

Your results should be completely reproducible from the code you submit. For all random processes, we recommend that you set a seed or random state to ensure that your results are consistent when your code is rerun. If you use a nonstandard library (in other words, any library that hasn't been used in the course so far), please include installation code.

3.6 Dataset (for external datasets)

If you choose to use your own data set, please submit the files. Make sure that the file names and files paths correspond to how you load the data in your Jupyter notebooks.

3.7 Team Member Assessment

In the middle of the project and at the time of submission, you will be required to fill out a form to summarize the contributions that each team member (including yourself) made to the project. If you have issues working with your group members, we encourage you to work together to resolve them earlier rather than later!

4 Datasets

We highly recommend selecting from the listed datasets; course staff has ensured that these datasets are sufficient in satisfying the project requirements. You can and should supplement the suggested data with publicly available secondary datasets (e.g., US Census, American Community Survey). If you use additional data, please reference the source in your write-up and submit the files in your ZIP file. (See Section 3.6).

If you choose to use an outside dataset, please refer to the suggested guidelines. Please note that staff will be less equipped to provide assistance with data outside the recommended list.

Note that some of the supplemental readings include already-answered research questions: make sure your research questions and approaches are different from theirs! Please contact course staff if you aren't sure about this requirement.

4.1 Dataset 1: Chronic Disease and Air Quality

Data Description:

The Center for Disease Control and Prevention (CDC) maintains the U.S. Chronic Disease Indicators dataset, containing state-specific data of chronic illness prevalence as well as relevant policies and regulations.

The CDC also publishes daily air quality data through the National Environmental Public Health Tracking Network to monitor environmental exposures.

Some of these datasets are large: we recommend subsetting the data to a particular geographic region and/or chronic illness to make the data easier to work with. See the Large Datasets section below for some additional tips.

Potential Directions:

1. What has been the impact of substance regulation on chronic disease onset?
2. How do levels of particulate matter and ozone affect the onset of chronic illnesses, such as asthma?
3. Are there any geographical trends in air pollution and/or chronic illness?

Relevant datasets:

1. CDC: Annual State-Level U.S. Chronic Disease Indicators
   (a) Filtered for COPD
   (b) Filtered for Asthma
   (c) Filtered for Cardiovascular Disease
   (d) Filtered for Tobacco
2. CDC: Daily Census-Tract PM2.5 Concentrations
3. CDC: Daily Census-Tract Ozone Concentrations

Supplemental Readings:

1. Public health impact of global heating due to climate change: potential effects on chronic non-communicable diseases (2009)
2. Past Racist "Redlining" Practices Increased Climate Burden on Minority Neighborhoods (2020)

4.2 Dataset 2: Primary Election Endorsements and Financing

Data Description:

FiveThirtyEight compiled a dataset with information on numerous primary elections (primary elections determine the candidates that each political party nominates for the general election) in 2022. The data contains information about each candidate, including their political leanings, endorsements, and gender, race, and veteran identities. The dataset also reports outcomes for each election.

The Federal Election Commission (FEC) publishes campaign financing data, including the amount of data raised and spent by each candidate. Public data includes donor names, contribution amounts, and dates of contribution.

Potential Directions:

1. Is there a relationship between the number of unique donations a candidate received and the proportion of the vote they received in the primary?
2. What type of candidates did Joe Biden, Donald Trump, or Bernie Sanders endorse? What candidates were popular among different special interest groups?
3. What are the characteristics of the most contentious elections?

Relevant datasets:

1. FiveThirtyEight: 2018 Primary Candidate Endorsements
2. Federal Election Commission: 2022 Campaign Financing Data

Supplemental Readings:

1. The Persuasion Effects of Political Endorsements (2016)
2. How Money Affects Elections (2018)

4.3 Dataset 3: Electricity and Carbon Emissions Forecasting

Data Description:

Given the rapidity and intensity of global warming, much attention has been devoted to understanding how electricity, emissions, and climate have been closely intertwined. In particular, the U.S government (along with companies such as ElectricityMaps) have open-sourced the data they've collected on emissions, climate/weather patterns, and electricity generation.

Potential Directions:

1. How can we predict electricity demand from weather/temperature data? How can we utilize point estimates of temperature to estimate electricity demand across entire states?
2. How are emissions, electricity demand, and electricity production related? Are there particular events that cause spikes in emissions? Can we use this to forecast emission rates?

3. How has electrification (i.e the adoption of EVs, adoption of electricity/battery powered appliances over fossil fuel powered ones) impacted electricity demand on the grid?

Relevant Datasets:

1. US Energy Information Administration: has an API that provides hourly forecasted and actual electricity demand for various states/power-authorities across the U.S. We encourage you to explore the various data available, but here are a few examples of datasets available through the EIA:
    - Historical state data shows generation, consumption, and emissions broken down by state, year, and type of energy from different year ranges spanning 1960-2022.
    - Power plant EIA-923 data has data on power plants across the US, showing electricity and data generated.
2. EPA eGrid: contains information about electric power generated in the US, including electricity generation, carbon dioxide emissions, and more.
3. National Oceanic and Atmospheric Administration: provides short-term and longer-term forecasts and reports on weather, temperature, and climate. For an API to access historical data, see the Climate Data Online page.
4. ElectricityMaps: provides free, open-source data on international emissions at various spatial and temporal granularities

Supplemental Readings

1. Two-thirds of the U.S. is at risk of power outages this summer—but it's not stopping Americans from electrifying everything in their homes
2. How Electrifying Everything Became a Key Climate Solution
3. U.N. Report Card Shows World Is Far From Meeting Climate Goals

4.4 Useful Additional Datasets

Depending on your research questions, you may find it helpful to join the following datasets with one of the above categories. These will likely be useful as confounding variables in causal inference, groups for hierarchical Bayesian modeling or A/B tests, features for prediction, or anything else you may want to use them for.

- The Census and American Community Survey are a valuable source for demographic data (race, socioeconomic status, housing, etc.). We have provided a short tutorial video that explains how to use their website.
- You are welcome (and encouraged) to mix and match individual data files across the datasets if it helps answer your research questions.
- The Opportunity Insights Data Library contains data on social mobility and economic opportunity, and could be a valuable addition to any analysis. Note that some of these datasets are large: see the Large Datasets section below for some tips.)
- The Bureau of Transportation Statistics (under the U.S. Department of Transportation) publishes monthly data on transportation utilization and spending. The dataset includes infor-

mation on airline traffic, transit ridership, transportation employment, construction spending, and transborder movement.

These are simply suggestions: you're welcome to add any publicly available dataset that you find, subject to the guidelines below.

4.5 Guidelines for External Data

If you choose not to use one of the three suggested datasets, your dataset must meet the following guidelines:

1. Data cannot contain sensitive and/or identifying data. If you do not have permission to share data with course staff, you are not allowed to use it for this project.
2. The data source(s) must be known. Why were the data collected? Who conducted/funded data collection? When were the data measured or recorded? How were data collected?
3. At minimum, the combined data should include 2 numerical and 2 categorical variables (it's okay if any individual source has less than this).
4. Your dataset should have a sufficient number of observations. This is a fairly subjective measure: please check with staff if you are concerned that your dataset is too small.

4.6 External Data Sources

Listed below are some suggested data sources for either an external dataset or supplementary data for one of our provided datasets. This is not a comprehensive list, nor can we guarantee that data found on these sites will meet the guidelines for this project.

1. Humanitarian Data Exchange
2. Gapminder
3. World Bank
4. WHO
5. UNICEF
6. UC Irvine Machine Learning Repository
7. Google Data Repository
8. AWS Data Repository
9. FiveThirtyEight
10. Kaggle
11. Dataverse
12. Meta's Data for Good
13. The General Social Survey

4.7 Large Datasets

When loading large datasets, you may find it helpful to use the `chunksize` or `nrows` parameters of `pd.read_csv`: see the documentation for more information. This Stack Overflow answer describing how to read large files in pandas might also be helpful.

5 Grading

The final project is worth 20% of your overall grade. Each section of the project will be weighted as follows:

1. Project proposal (5%)
2. Checkpoint 1: EDA (5%)
3. Checkpoint 2: Research Questions (5%)
4. Data Overview (7%)
5. EDA (10%)
6. Research Question 1 (25%)
7. Research Question 2 (25%)
8. Conclusion (15%)
9. Group Member Evaluation (3%; additional adjustments may be made based on individual contributions)

These values are subject to change (by small amounts). See the rubric (linked on Ed) for more detailed information about how each section will be graded.

6 Working in Groups

Here are some tips for working in your project groups. If you have any concerns with your group members, the best way to resolve them is usually proactive, direct, and considerate communication! Most of these are based on the collective wisdom of course staff and advice from past semesters' Data 102 students.

- Start a group chat and make sure everyone knows to check it regularly.
- Consider meeting in person rather than online: students have told us that this tends to be more productive and efficient, and more fun too.
- Don't be afraid to reach out to course staff and come to office hours early if your group is struggling to refine your research questions. As usual, you'll likely be able to get more focused help at instructor office hours earlier in the week.
- Even if you divide up the work, it's often helpful to have all group members contribute to brainstorming, reviewing results, and discussing how best to write up the report.
- If you have concerns with your group members, communicate them earlier: don't wait until the last minute! It's much easier to resolve issues and improve collaboration early than the night before the deadline.
- For the remainder of the semester, all vitamins will contain a very short survey on how well your project is going. While we won't share your comments directly with anyone else in your group, course staff will reach out to any groups that raise issues, so that we can help you resolve them and work productively with your group.