

Term Project:

During the course, you will be working on a term project that takes you through an exploratory data analysis project. You are going to pose a question and then evaluate a dataset to answer the question, in the same way the author is attempting to answer the question, “Do first babies tend to arrive late?”. While you will be gaining experience in the course with the datasets the author of our book selected, you will also be required to select a dataset, form a question or hypothesis of that dataset and then using the various techniques we learn in the course work towards proving or disproving your hypothesis. What will be required of you is a summary of your results along the way, along with a written and visual presentation of your findings. The term project will be graded at the end of the course, however each week, you should follow-along with the book and apply the learned methods to your own dataset.

The first step is selecting a dataset and forming your question/hypothesis.

Some helpful places to find datasets include:

- <https://community.tableau.com/docs/DOC-10635>
- <https://www.kaggle.com/datasets>
- <http://www.data.gov>
- <http://www.science.gov>
- <http://data.gov.uk>
- <http://gss.norc.oregon.edu/>
- <http://www.europeansocialsurvey.org>

There are no restrictions on what dataset you use, other than you cannot use the specific datasets used primarily in the book which are from the following sources: National Survey of Family Growth and the Behavioral Risk Factor Surveillance System.

You will turn your entire term project in during the final week of class, however, here are some of the milestones to help you gauge where you should be at during each week of class.

Milestone 1 (Week 1) – Evaluate datasets, start thinking of statistical questions

Milestone 2 (Weeks 2-3) – Select a dataset, solidify your statistical question, begin describing the single variables in your dataset to determine which variables are relevant to your question (Distributions, PMFs, CDFs). You should know your statistical question you are trying to answer by no later than Week 2.

Milestone 3 (Weeks 4-5) – Start identifying relationships between the variables you have identified vs looking at just one variable at a time.

Milestone 4 (Weeks 6-8) – Start evaluating if the results you are seeing in a sample would happen in the large population and start testing out the results and hypothesis you have made up to this point.

Milestone 5 (Weeks 9-10) – Wrap up your PowerPoint presentation and summarization of analysis.

The following is due submitted to your GitHub repository. Submit a link to your repository to the assignment link during the final week of class:

- Your dataset

- A PowerPoint presentation outlining your statistical question/hypothesis
 - A minimum of 5 variables in your dataset used during your analysis (for help with selecting, the author made his selection on page 6 of your book). Consider what you think could have an impact on your question – remember this is never perfect, so don't be worried if you miss one (Chapter 1).
 - Describe what the 5 variables mean in the dataset (Chapter 1).
 - Include a histogram of each of the 5 variables – in your summary and analysis, identify any outliers and explain the reasoning for them being outliers and how you believe they should be handled (Chapter 2).
 - Include the other descriptive characteristics about the variables: Mean, Mode, Spread, and Tails (Chapter 2).
 - Using pg. 29 of your text as an example, compare two scenarios in your data using a PMF. Reminder, this isn't comparing two variables against each other – it is the same variable, but a different scenario. Almost like a filter. The example in the book is first babies compared to all other babies, it is still the same variable, but breaking the data out based on criteria we are exploring (Chapter 3).
 - Create 1 CDF with one of your variables, using page 41-44 as your guide, what does this tell you about your variable and how does it address the question you are trying to answer (Chapter 4).
 - Plot 1 analytical distribution and provide your analysis on how it applies to the dataset you have chosen (Chapter 5).
 - Create two scatter plots comparing two variables and provide your analysis on correlation and causation. Remember, covariance, Pearson's correlation, and Non-Linear Relationships should also be considered during your analysis (Chapter 7).
 - Conduct a test on your hypothesis using one of the methods covered in Chapter 9.
 - For this project, conduct a regression analysis on either one dependent and one explanatory variable, or multiple explanatory variables (Chapter 10 & 11).
- Using Python, submit your results via your notebook or export your code and submit via the assignment link. You must show your code and work for full credit.
- A 250-500-word paper summarizing the following:
 - Statistical/Hypothetical Question
 - Outcome of your EDA
 - What do you feel was missed during the analysis?
 - Were there any variables you felt could have helped in the analysis?
 - Were there any assumptions made you felt were incorrect?
 - What challenges did you face, what did you not fully understand?