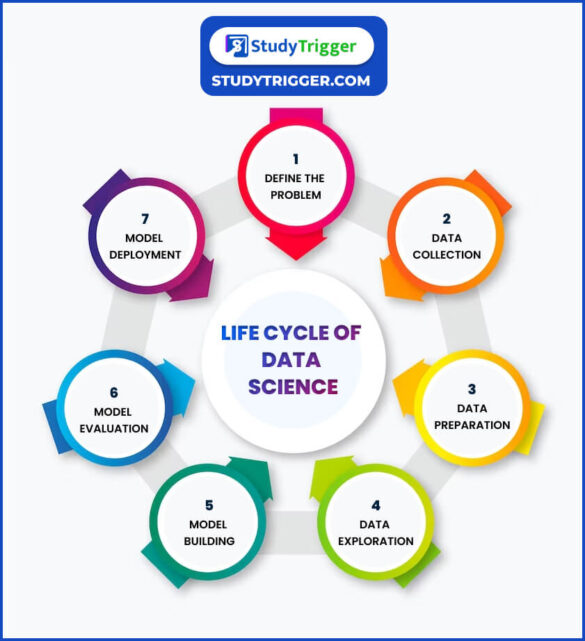
****

The group portion of the final project will include a group project [**proposal**](#_mytkdgoo3g1c), a group **write-up** in a Jupyter notebook (with visualizations and functioning code) and a group **poster** presentation.

Your final project group **write-up** must:

1. Discuss (but not necessarily implement) the 7 stages of the data science lifecycle as they relate to your topic of exploration. Stages 4, 5, and 6 will comprise the bulk of your project, but all the sections should be addressed. For example:
   1. If data is already collected, the data collection section could discuss concerns or important aspects of the methodology that should have been used when collecting the data.
   2. The data preparation section might describe the steps taken to ensure the data is in a useful form.
   3. The model deployment section might propose how the model could be deployed and/or potential concerns if it actually was deployed.
2. Investigate a hypothesis or a question.
3. Build and evaluate at least one model.
   1. The model doesn’t have to be predictive (for example, linear regression or decision tree that we cover in the last unit). A model just has to make some reasonable assumption on the data, and it doesn’t have to be complicated.
   2. For evaluation, you can use a technique appropriate for the model you have chosen, such as simulation, sampling, comparing distributions, or calculating accuracy. You might also evaluate the effects of outliers.
   3. You are also welcome to build multiple models and compare how they perform.
4. Investigate at least 3 variables and their potential associations.
5. Write at least one Jupyter Notebook file with at least 2 custom functions.
6. Create at least 3 original Python visualizations with appropriately labeled axes and titles.

Your final project group **poster** will be presented to the class during our scheduled finals time slot. It should include the most important elements from your write-up and should use visualizations effectively to “tell the story” of your data and your findings. All team members must participate in presenting the poster and should be prepared to explain any portion, including:

1. The problem being investigated.
2. The visualizations, why they were chosen, and how they were generated.
3. The model, why it was chosen, and how it was evaluated.

# 

## Project proposal (80 points, 8% of final course grade)

Please make a group submission on Gradescope: <https://guides.gradescope.com/hc/en-us/articles/21863861823373-Adding-Group-Members-to-a-Submission>

Make sure to include a project title(1 point) and a (cool) team name (1 point)!

Expected length: 1~2 pages, 400~500 words. A little longer or shorter is okay.

1. Define the problem (20 points)
   1. What is a question you have about the world that this project could potentially help answer? Or what is a problem or unmet need that this project can help with? Write down the “big picture” question and why you are interested in it.
2. Data collection (10 points)
   1. Even if you are working with a dataset we have provided, try to describe the origin of this dataset in your words. Who collected the data? Who is the population? Comment on potential limitations and biases. If you wish, you are welcome to collect your own data through, for example: surveys, counting the number of people who visit Meantime Coffee, or measuring the arrival times of Chapel Hill transit. If you intend to collect your own data, include a detailed plan for doing so.
3. Data preparation (10 points)
   1. Even if you are working with a dataset we have provided, it is likely that you would have to prepare them in some manner for your specific project. Refer to lecture 11 slides for examples. If this step is not relevant, say so.
4. Data exploration (20 points)
   1. Given what we have learned in this class so far, you already have the tools to calculate summary statistics and visualize the data. Include some of those preliminary results (a few numbers and/or a figure is fine) and comment on what you have learned about the data. No need to share your code. Refer to lecture content for examples. Your data exploration may help add specificity to part 1.
5. Model building (10 points)
   1. You’re not expected to have done any work or have a very detailed plan for model building at this point. But given what you have learned about the data from step 4, do you have any assumptions on data that you would like to explore more?
6. Risks and timeline (8 points)
   1. What are potential risks or failures you foresee in this project? How would you proceed if that happens? What is the planned timeline for this project? Remember that all assignments related to this course have to be submitted by the final exam date.

## Write-up and Code (80 points, 8% of final grade)

This is a **group** assignment. Please make a group submission on Gradescope: <https://guides.gradescope.com/hc/en-us/articles/21863861823373-Adding-Group-Members-to-a-Submission>

You are welcome to reuse anything you wrote for the project proposal. Just be aware that something that might have gotten full credit in the proposal might not get full credit in the final report if some of the details aren’t ironed out.

Your final report can also be very different from what you originally suggested in the proposal. That’s totally fine!

Here are things you need to have:

* Do you have at least 3 data visualizations? Include them in your write up.
  + Tables of summary statistics, correlation coefficients, linear regression coefficients all count as a visualization as long as they are well-formatted and have a title. Screenshots of raw data don’t count unless you collected your own data.
  + Examples: scatter plots, histograms, bar charts, line plots, visualization of your decision tree classifier
* Do you have at least 2 custom functions in your code? Don’t include your code in the write up, just submit the notebook file.
* Did you build and evaluate at least one model?
* Did you investigate a hypothesis or a question?
* Do you work with at least 3 variables?
  + Fitting a model with 3 features works!
  + Even if you only end up investigating 1~2 features for your model, if you did data exploration with 3+ features and mention why you narrowed the list down for further analysis, that’s ok.
* Did you discuss (but not necessarily implement) the 7 stages of the data science lifecycle? See below:

1. Define the problem
   1. This should include your “big picture problem” as well as a clear statement of the specific problem or hypothesis you chose to focus in on with your analysis.
2. Data collection
   1. Source(s) of the data, including if they come from multiple tables.
   2. Clearly state the number of samples and number of features.
   3. Mention the features (or categories of features if there are many of them).
   4. Mention potential sampling bias or limitations in the data collection process.
3. Data preparation
   1. If you dropped any data, describe why and how you did it.
   2. If you normalized any data, describe why and how you did it.
   3. If you converted any categorical data to numeric, describe why and how you did it.
   4. Describe any other data preparation steps you may have taken.
   5. If you didn’t need any data preparation, explain why.
4. Data exploration
   1. Include your data exploration from your project proposal.
   2. Include any new data exploration you may have done to help inform your model choice, but not required to receive full credit.
5. Model building
   1. This step and the Model Evaluation step should be the heart of the final write-up.
   2. Describe the model you chose to use and why.
   3. Describe the features or variables you used and why. Remember that you need to investigate at least 3 variables and their relationships. Clearly state which variables (out of all of them in your dataset) you chose to investigate.
   4. You are also welcome to build multiple models and compare how they perform, but not required to receive full credit.
6. Model evaluation
   1. This step and the Model Building step should be the heart of the final write-up.
   2. Describe the hypothesis you tested or question you investigated. What does your model say about your hypothesis or question?
   3. For evaluation, you can use a technique appropriate for the model you have chosen, such as simulation, sampling, comparing distributions, or calculating accuracy.
   4. If you’re looking for more analysis to do, you might also evaluate the effects of outliers, model complexity (e.g. tree depth, degree for polynomial regression), or effect of random train/test split. Not required for full credit.
   5. You are also welcome to build multiple models and compare how they perform, but not required to receive full credit.
7. Model deployment
   1. Briefly describe how you could envision this model being used or deployed.
   2. Briefly describe any issues, risks, and or challenges you see with deploying your model.
8. Code
   1. As part of your write-up, please submit your Python notebook.
   2. This should include at least 2 custom functions written by your team
   3. Code will not be graded on efficiency, complexity or style, just that it was used to generate the figures and/or tables in the write-up.

Rubrics for writing:

* Completeness – includes all the steps of the data science life cycle and the required elements described above
* Continuity – each described step of the data science life cycle logically follows from the previous steps
* Clarity – understandable and succinct
* Organization – well-structured
* Accuracy – truthful and not overstated
* Mechanics – correctness of grammar and spelling

## Poster (100 points, 10% of the final grade)

This is a **group** assignment.

Our exam session is on **Saturday, December 7, 4-7 pm** at **133 Rosenau**. This is different from our usual classroom! We booked this room so that you don’t have to print out the poster. I need to figure out the technology, but I’ll probably just need the pdf submission the night before or the morning of.

**Process**

Your final project group **poster** will be presented to the class during our scheduled finals time slot in the manner typical of academic conference Poster Sessions. The process will be that for 1 hour, 8 or 9 teams will simultaneously stand next to their poster and explain it or answer questions about it as the professors, TAs, and the rest of the class walk from poster to poster. At the end of the hour, we will switch, and the next 8 or 9 teams will simultaneously stand next to their poster and do the same. And the same for the final hour. Based on their interaction and questions, the students that are not currently presenting their poster will fill out their individual “Feedback to other teams” submission described below (to be completed before the end of the final period.) All individuals are expected to provide feedback for 10 other projects during the blocks they’re not presenting.

**Team Poster Composition**

A poster is intended to be an engaging summary of a more detailed written analysis (e.g. your team write-up). It should clearly convey the problem, approach, and results of the project. It should include the most important elements from your team write-up and should use visualizations (i.e. plots) effectively to “tell the story” of your data and your findings.

* Gillings has a very nice poster template that you can use (just swap out the logo): <https://sph.unc.edu/comm/academic-poster-templates/>
* Designing Effective Posters: <https://guides.lib.unc.edu/posters>
* Academic Posters: <https://gradschool.unc.edu/academics/resources/postertips.html>

All team members must participate in presenting the poster for 1 hour, and all team members should be prepared to explain any portion, including:

1. The problem your team investigated.
2. The data you chose to use for analysis, and why.
3. The hypothesis you tested or question you investigated
4. The model(s) you chose for analysis and why.
5. How you evaluated your model(s)
6. For each visualization, what it is showing, how it was generated, and why it is important to your project

You don’t need to prepare a scripted talk. It’ll be a short “tell me about your project” and then Q&A.

Poster rubrics:

* Completeness – includes problem, approach, and results
* Continuity – the entire poster feel cohesive and part of a whole story of the data
* Clarity – understandable
* Organization – well-structured
* Accuracy – truthful and not overstated
* Mechanics – correctness of grammar and spelling

Individual presentation rubrics:

* Knowledge – ability to describe the elements listed above
* Engagement – ability to respond to impromptu audience questions
* Delivery – appropriate pace, volume, enthusiasm/energy

## Reflection and teamwork (20 points, 2% of the grade)

This is an **individual** assignment.

It needs to contain these components:

1. One thing you learned through this project.
2. One challenge you encountered during this project and how you overcame it. (This exercise will be helpful in your job interview or grad school applications.)
3. List a contribution that each team member made to the project (including you).
4. [Optional] Anything else you would like for us to know about the project.
5. [Extra credit] Add this project to your CV, resume, LinkedIn page, personal website, or anything similar, and include a screenshot of it.

We’re expecting something short (1 page), but if you run over that’s ok.

## Feedback to other teams (20 points, 2% of the grade)

This is an **individual** assignment.

Everyone will be told ahead of time which teams they are reviewing during the poster session. We will give you a printed out rubric and/or send a short google form/canvas survey for you to fill out electronically. You need to submit them before the end of the final exam period.