The purpose of this document is to provide guidance for navigating the execution of your capstone project. It is divided into three sections: a project summary, a checklist, and an example process outlining how one might approach the execution of their capstone project. The summary section provides an overview of the requirements you need to fulfill for successful project completion. The checklist section organizes these requirements in a checklist format, allowing you to assess your progress toward completing the project successfully. Lastly, the example process section outlines an approach for completing a capstone project. This section includes examples based on previous students' projects.

# Project Summary

The capstone project is an opportunity for you to showcase the skills you learned throughout the bootcamp by designing, executing, and presenting an individually crafted project. The challenge of completing a project that has not been pre-defined enables you to stretch your analytical, technical, and organization skills. The end result of this effort will be a dashboard or presentation that enables you to communicate the findings of your analysis. This project will serve as the keystone of your portfolio, demonstrating your ability to work with multiple technologies.

## Data Selection and Preparation

* Determine an area of personal or professional interest that you would like to explore with data.
* The data you work with should enable you to demonstrate your proficiency in cleaning, merging, and preparing data for analysis.
* The dataset should also enable you to answer three to four data questions so you can showcase your analytical abilities.
* Ideally, the project should involve either multiple datasets or one sufficiently complex dataset that requires extensive cleaning, integration, data preparation, or involves the use of APIs or web scraping techniques.
* The primary dataset cannot be sourced from Kaggle or used in previous class projects.
* While proprietary data may be used with prior instructor approval, your data cannot contain Personally Identifiable Information (PII) visible in the final deliverables.

## Data Questions

* Formulate appropriately scoped and sufficiently complex data question(s) to guide your project.
* Maintaining a clear focus on the data question(s) throughout the project, will ensure all data collection, preparation, analysis, and visualizations are aligned with answering them effectively.
* The questions should be distinct from those explored in previous class projects.

## Technologies

* Your project must utilize at least two of the technologies you learned in the class - Excel, SQL, Power BI, Tableau, and Python.

## GitHub Requirements

* You will need to maintain an organized and well-documented codebase on GitHub.
* You should make at least one commit per class session to show regular progress.
* As you are wrapping up your project, you should include a demonstration of the analysis process, showcasing the steps taken to arrive at the final results.
* You should include an informative and representative README file that provides an overview of the project, instructions for running the code, and details on the chosen technologies and techniques.
* Ensure there is no visible PII in the code or final deliverables.
* The exception to the aforementioned requirements if projects using proprietary data. In this case, communicate with your instructor about how to make your code assemble for review.

## Capstone Presentation

* As your capstone presentation is used to share your project on Demo Day, the presentation must effectively communicate your project's findings using relevant visualizations that meaningfully contribute to your audience's understanding of your findings. Your presentation must be contained in PowerPoint, an interactive dashboard, or another appropriate and approved medium.
* In order to participate in demo day and graduate, you must demonstrate your completed capstone project to the instructor and obtain approval by the end of your final class session.

## Capstone Support

* Seek regular support and troubleshooting from your instructor throughout the project, especially in areas related to data acquisition, analysis, visualization, and presentation.
* Maintain regular check-ins and feedback sessions with the instructor to receive guidance and ensure progress aligns with the Minimum Viable Product (MVP) defined by the student.

# 

# Capstone Checklist

**Data Selection and Preparation**

* Select an Area of Interest: Determine a personal or professional topic that you want to explore using data.
* Data Source Selection: Choose a dataset that meets the following criteria:
  + Enables you to demonstrate proficiency in data cleaning, merging, and data preparation for analysis.
  + Allows you to answer three to four data questions, showcasing your analytical abilities.
  + Involves either multiple datasets or one complex dataset that requires extensive cleaning, integration, data preparation, or incorporates APIs or web scraping techniques.
  + Is not sourced from Kaggle or used in previous class projects.
* Data Privacy Consideration: Ensure the data does not contain Personally Identifiable Information (PII) visible in the final deliverables. If using proprietary data, obtain prior approval from the instructor.

**Data Questions**

* Define Clear and Focused Data Question(s): Formulate appropriately scoped and specific data question(s) that guide your project's objectives.
* Ensure Complexity and Relevance: The data question(s) should be sufficiently complex and relevant to demonstrate your analytical abilities and address real-world challenges.
* Avoid Repetition: Ensure the data question(s) are distinct from those explored in previous class projects or assignments to showcase your ability to explore new topics.
* Verify Data Availability: Ensure that the necessary data to answer the formulated question(s) is available and accessible.
* Data Question Alignment: Throughout the project, maintain a clear focus on the data question(s) to ensure all data collection, preparation, analysis, and visualizations are aligned with effectively addressing them.

**Technologies**

* Project Technology Selection: Ensure that your project incorporates at least two of the technologies you learned in the class such as -
  + Excel
  + SQL
  + Power BI
  + Tableau
  + Python

**GitHub Requirements**

* GitHub Repository Setup: Create a GitHub repository to maintain an organized and well-documented codebase for your data analysis project.
* Regular Commits: Make at least one commit per class session to demonstrate regular progress throughout the project.
* Demonstration of Analysis Process: As you wrap up the project, include a clear demonstration of the analysis process, showcasing the steps taken to arrive at the final results.
* Informative README File: Include an informative and representative README file in the repository. The README should contain the following:
  + Overview of the project and its objectives.
  + Details on the chosen technologies and techniques used in the project.
* Protect Personally Identifiable Information (PII): Ensure that there is no visible Personally Identifiable Information (PII) in the code or final deliverables. This is crucial to protect privacy and comply with data protection regulations.
* Handling Proprietary Data: If your project uses proprietary data, communicate with your instructor about how to make your code accessible for review while respecting any confidentiality agreements.
* Code Documentation: Provide clear and concise documentation within the codebase to explain the purpose and functionality of major code sections and functions.
* Presentation or Dashboard: Include a pdf copy of your presentation slides in your repository or a link to your published dashboard in your README.

**Capstone Presentation**

* Project Completion: Ensure that your capstone project is fully completed, including data analysis, visualization, and documentation.
* Approval Demonstration: Schedule a session with the instructor to demonstrate your completed capstone project and obtain approval before the end of your final class session.
* Effective Communication: Create a presentation that effectively communicates your project's findings to the audience. Use relevant visualizations that enhance the audience's understanding of your project's results.
* Relevant Visualizations: Select visualizations that meaningfully contribute to the audience's understanding of your findings and highlight key insights from your data analysis.
* Presentation Medium: Choose an appropriate and approved medium for your capstone presentation. This could be in the form of an interactive dashboard, a PowerPoint presentation, or any other suitable format.
* Demo Day Preparation: Prepare your presentation for Demo Day, ensuring it is engaging, clear, and effectively showcases the value of your project to potential stakeholders.
* Check for Errors: Review your presentation for any spelling, grammar, or formatting errors, and ensure that all visuals are accurate and appropriately labeled.
* Incorporate Feedback: If you received feedback during the project's development, incorporate any necessary changes or improvements into your final presentation.
* Clarity and Conciseness: Make sure your presentation is clear, concise, and avoids unnecessary jargon to ensure that all audience members can follow along.
* Storytelling: Organize your presentation in a coherent and logical manner, telling a compelling data-driven story from start to finish.
* Stay Within Time Limit: Practice delivering your presentation within the designated time limit to ensure it fits into the allocated time on Demo Day.
* Practice and Rehearsal: Practice your capstone presentation multiple times to ensure you are confident and well-prepared to present on Demo Day.
* Final Review: Conduct a final review of your capstone project, presentation, and documentation to ensure everything aligns with the project's objectives and requirements.

# 

# Capstone Project Example

The following process illustrates how one could approach a capstone project. Please note, this is just an example, and the process you utilize to complete your capstone does not need to follow this outline.

1. **Determine an area of interest for your project.** You can pick an area that is of personal or professional interest to you like sports, healthcare, finance, etc. At this point you do not need to define anything specific about your project, just its general topic.
   * For example, Sam was a data analytics student interested in working in healthcare, so she decided to create her capstone project using COVID-19 data.
2. **Determine 3-4 general questions you would like to answer about your domain of choice.** At this step, your goal is to generally define the direction of your project. These questions do not need to be measurable, but they should help inform what data needs to be collected.
   * For example, Sam was interested in the following questions, which helped her determine what data she should collect for her analysis:
     1. How were rates of COVID-19 cases impacted by personal income?
     2. What demographic factors impacted COVID-19 cases rates?
     3. What events drove spikes in COVID-19 case counts?
3. **Collect data related to your chosen domain.** Ensure your data has a blend of numeric and categorical variables that you can use for answering questions. The data you collect should enable you to measure specific phenomena related to your general data questions.
   * For example, after reviewing several data sources, Sam decided to focus her analysis on the COVID-19 cases in Tennessee. She went to the Tennessee Department of Health’s website and pulled the following files:
     1. County New: This file provided details of daily COVID-19 cases by each county in Tennessee.
     2. Age By County: This file listed case counts by age groups by day for each county in Tennessee.
     3. Covid Vaccine County Demographics: This file provided case counts by demographic category by day for each county in Tennessee.
   * Because she was interested in the relationship between COVID-19 case rates and individuals’ income, she then went to the CensusReporter.org website and pulled the following datasets:
     1. Household Income in the Past 12 Months: This dataset provided household income level for each county in Tennessee.
     2. County Population: This dataset provided the population count for each county in Tennessee.
4. **Perform an initial pass of exploratory data analysis.** At this step in the process, you want to familiarize yourself with the dataset(s), create a list of data cleansing tasks that need to be performed, and understand what questions the data you have collected can and cannot answer.
   * For example, Sam loaded all of the datasets she collected into Pandas dataframes, and performed basic exploratory analysis on each dataset. She noted that she needed to perform the following data cleaning tasks:
     1. Remove duplicate values found in a couple of the datasets.
     2. Address outlier values found in a couple of the datasets.
     3. Standardize date formats and data types.
     4. Standardize key fields used across datasets.
     5. Create dataframes aggregating certain values to the county level for analysis.
   * As Sam performed her review, she took notes on what numeric and categorical values were available in the data she collected. This helped her in the next step of the process when she started to revise her data questions.
5. **Refine your data questions into specific measurements.** Based on your initial analysis of the collected data, refine your data questions into specific measurable units of analysis. At this point, you want to refine the scope of your data questions into specifications for your project.
   * For example, in step two of this process, Sam identified three broad questions she was interested in exploring through her analysis. After exploring the datasets she collected, she refined her questions into four specific measurements:
     1. Among all Tennesseans, what was the per capita rate of COVID-19 cases, hospitalizations, and deaths over the course of the pandemic?
     2. Among all Tennesseans, what is the per capita rate of COVID-19 cases, hospitalizations, and deaths by age category?
     3. Among all vaccinated Tennesseans, what is the per capita rate of vaccination by race, ethnicity, and gender.
     4. For each county in Tennessee, what is the correlation between median income and the per capita rate of COVID-19 cases, hospitalizations, and deaths?
6. **Prepare the data you have collected for analysis.** At this point you want to clean and enhance your data for analysis. Use the list of data cleaning tasks you created during your initial analysis of the dataset as a starting point for cleaning. At this point if you know you need to create any calculated fields, go ahead and do that work as well.
   * For example, Sam created a list of data cleaning tasks in step four of this process. At this point in the process, she completed those tasks. In Sam’s case, she merged the data she collected from the CensusReporter with the Tennessee Department of Health data in order to calculate the per capita rates of cases and vaccinations. She did not, however, need to create any calculated fields, so her data was prepared and ready to be loaded into Power BI.
7. **Visualize measures that answer your specific data questions.** For each of your data questions, visualize one or two primary measures and when appropriate two to four secondary measures. A primary measure is the metric that answers the data question most directly. A good choice for visualizing a primary measure is a chart that displays the measure over time. This enables your audience to better understand the historical performance of the metric when it is explored in secondary metrics. Secondary measures should support and clarify the primary measure(s) and may or may not be derived from the same measurement as the primary measure(s).
   * For example, Sam created the following visualizations for her data questions:
     1. How did the rate of COVID cases, hospitalizations, and deaths in Tennessee progress over the course of the pandemic?
        1. Primary Measure(s)
           1. A line chart displaying the per capita rate of cases in Tennessee
           2. A line chart displaying the per capita rate of hospitalizations in Tennessee
           3. A line chart displaying the per capita rate of deaths in Tennessee
        2. Secondary Measure(s)
           1. A line chart displaying the count of cases in Tennessee
           2. A line chart displaying the count of hospitalizations in Tennessee
           3. A line chart displaying the count of deaths in Tennessee
           4. A card reporting the total number of cases in Tennessee across the reporting period.
           5. A card reporting the total number of hospitalizations in Tennessee across the reporting period.
           6. A card reporting the total number of deaths in Tennessee across the reporting period.
     2. Among all Tennesseans, what is the per capita rate of COVID-19 cases, hospitalizations, and deaths by age category?
        1. Primary Measure(s)
           1. A clustered bar chart showing the per capita rate of cases, hospitalizations, and deaths by age category.
     3. Among all vaccinated Tennesseans, what is the per capita rate of vaccination by race, ethnicity, and gender?
        1. Primary Measure(s)
           1. A line chart displaying the per capita rate of vaccinated Tennesseans by race
           2. A line chart displaying the per capita rate of vaccinated Tennesseans by ethnicity
           3. A line chart displaying the per capita rate of vaccinated Tennesseans by gender
        2. Secondary Measure(s)
           1. A card reporting the total number of vaccinated Tennesseans across the reporting period.
           2. A bar chart showing the total number of vaccinated Tennesseans by race
           3. A bar chart showing the total number of vaccinated Tennesseans by ethnicity
           4. A bar chart showing the total number of vaccinated Tennesseans by gender
     4. For each county in Tennessee, what is the correlation between median income and the per capita rate of COVID-19 cases, hospitalizations, and deaths?
        1. Primary Measure(s)
           1. A scatterplot correlating the per capita rate of cases and median income by county
           2. A scatterplot correlating the per capita rate of hospitalizations and median income by county
           3. A scatterplot correlating the per capita rate of deaths and median income by county
        2. Secondary Measures(s)
           1. A choropleth map of the median household income by county
           2. A choropleth map of the per capita rate of cases by county
           3. A choropleth map of the per capita rate of hospitalizations by county
           4. A choropleth map of the per capita rate of deaths by county
8. **Organize your dashboard or presentation.** After you have visualized your measurements, take time to get familiar with them. What is significant or notable about each primary and secondary measurement? As you review your findings, what narrative arises? How can you best assemble your findings into a coherent and engaging story? After answering these questions, create a dashboard or presentation that conveys this information to your audience.
   * For example, Sam decided to structure her dashboard so that each of her data questions were answered on a single page. This enabled her to speak to the interrelated nature of the primary and secondary measures for each data question. Next, she organized the ordering of each page in the dashboard so that the first page offered the broadest context of her analysis for her audience and the subsequent pages showed narrower, more specific analyses. After assembling her dashboard, she then spent time with each measure noting any insights that she wanted to share during her presentation.
9. **Prepare to present your capstone.** After you have completed your dashboard or presentation, practice delivering it. There are a few things that are useful to focus on when preparing. First, be able to quickly orient your audience to each chart. You want to succinctly explain what is being measured, the x-axis, and the y-axis. In the world of remote meetings, don’t assume your audience can see your presentation clearly enough to decipher what each chart is reporting. Second, be able to inform your audience of one or two notable takeaways from each primary and secondary measure. Whatever is significant about each chart, should be clearly noted to your audience. Finally, after you have walked through your presentation a couple of times, create a compelling introduction and conclusion that summarize the key findings from your analysis. When preparing your introduction and conclusion, use the Aristotelian triptych - your introduction should consist of you telling your audience what you are about to tell them and your conclusion should tell them what you just told them.
   * For example, to ensure her presentation was a success, Sam practiced delivering it multiple times. Through her practice, she developed short explanations of each chart and refined the key takeaways she noted for each measure in step eight. After a few rounds of practice, she crafted a compelling introduction and conclusion for her presentation using the Aristotelian triptych approach. In her introduction, she briefly previewed what she would be presenting, setting clear expectations for her audience. In the conclusion, she effectively summarized the key findings from her analysis, reiterating the important points she had discussed. By following this approach, she ensured that her audience would leave the presentation with a clear understanding of her work and the conclusions drawn from the data. Once Sam had her introduction and conclusion drafted, she continued practicing her presentation, eventually enlisting her friends, family, and classmates to serve as her audience.