Part 1: Topic and Dataset Selection

Begin by choosing a topic that resonates with your interests or professional goals. This part involves selecting a suitable dataset for your project, which can be sourced online or generated using AI. The chosen topic and dataset should offer a rich ground for analysis and insights, laying the foundation for your project.

1A: Topic Selection

1A.1: What is your chosen topic for this project?
Briefly describe the topic you've selected and explain why it interests you or how it relates to your professional goals or personal passions.

I will be focusing on food swamps in NYC and how they correlate to income levels of various zip codes. This is a continuation of a project I began working on in the Python class. This is something I am very passionate about because I used to work for the nutrition department at my college over the summer and would deliver nutrition education to underfunded areas in Hartford. These communities were deeply affected by food instability which was caused by food deserts/swamps. This is also an issue in NYC and so I would like to explore its effects here.

1A.2: Why is this topic important or relevant to study?
Discuss the significance of your chosen topic in the current context. Is it related to emerging trends, societal issues, technological advancements, or professional practices?

Food swamps, areas where unhealthy food options are more accessible than healthier alternatives, have become a growing public health concern in urban environments like NYC. These areas often have a high density of fast food outlets compared to grocery stores, making it difficult for residents to access nutritious food. This imbalance has been linked to higher rates of obesity, diabetes, and heart disease, especially in low-income neighborhoods where healthier options are limited or expensive. In NYC, food swamps disproportionately affect lower-income communities and communities of color, further exacerbating health disparities. Lower-income areas tend to have a higher concentration of fast food restaurants and fewer grocery stores, making it challenging for residents to maintain a healthy diet. While city initiatives have tried to increase grocery store availability and promote fresh produce, these efforts have had mixed success, leaving many areas underserved.

My project's data on fast food restaurant locations, income levels by zip code, and the availability of assistance programs would like to explore these issues. Comparing the number of fast food restaurants to grocery store concentrations in different zip codes will help determine whether food swamps are more prevalent in lower-income neighborhoods. Additionally, analyzing the accessibility and effectiveness of government health aid programs, such as SNAP, in these areas can provide insight into how well these initiatives are addressing the problem.

1A.3: What specific questions or problems do you hope to address with your analysis? Identify 2-3 specific questions or challenges within your topic that you aim to explore through your data analysis. These should guide the focus of your project.

I would like to explore the correlation between the number of fast food restaurants broken down by zip code and combine that with a table showing income level to identify where potential food swamps exist. I would also like to explore the correlation between zip code and accessibility to large super markets.

1B: Dataset Selection

1B.1: Describe the dataset you will be using for your project.
Provide a detailed overview of your dataset including:

- The types of data it contains (numerical, categorical, textual, etc.).
- An estimate of the size of your dataset (number of tables, number of columns per table, and the approximate number of rows).
- Identify the most meaningful table(s) and column(s) within your dataset and explain their significance to your topic.

My main dataset is a list of all the restaurants in NYC. The data contains mostly categorical data, specifically geographic data. It has the information for almost 250,000 restaurants (rows) and contains information about locations and safety ratings. It has 26 columns. I am only concerned with the location of these restaurants, specifically their zip codes. I will be filtering this out by only having the dataset show the names and zip codes of the top 15 fast food restaurants in the USA. I also have a small table I created off of data on income I found online. It only has three columns, the zip code, the average income level of that zip code and the percent of high income households (income > 200k). Finally I have a dataset which shows the number of supermarkets in each zip code.

1B.2: Dataset Characteristics
Consider the following guidelines in your response for 1B.2.

- How is your dataset organized? Discuss whether the data is structured (e.g., in tables with defined relationships) or unstructured (e.g., text documents, images). ● What is the source of your dataset? Explain whether your data was sourced from a formal database (like government records, academic research, or industry reports), generated using AI, or collected through other means.
- Is your dataset ready for analysis or does it require cleaning and organization? Reflect on the initial state of your dataset and what preprocessing steps might be necessary before analysis.

The data is structured in tables and it comes from the NYC Health database. I was able to clean this data during the Python class so it is ready for analysis.

1B.3: Where did you obtain your dataset, and why did you choose this particular source? Detail the origin of your dataset, citing the source if it's publicly available or describing the process used to generate the data if it's artificially created.

The data comes from the NYC Health database and is publically available. I chose this because it was the most reliable source I could find for this data.

1B.4: Describe the organization and cleanliness of your dataset.
Is your dataset well-structured and clean, or does it contain inconsistencies, missing values, or other issues that will require attention before analysis?

I was able to remove all missing values and repeated rows during the Python class. It is also filtered to only show the top 15 fast food restaurants in the USA.

Part 2: Dataset Cleaning and Restoring

This part focuses on the essential process of cleaning and preparing your dataset for analysis. You'll be tasked with importing the dataset into PGAdmin, ensuring data integrity by cleaning, normalizing, and, if necessary, transforming the data to a suitable format for querying and analysis.

2A: Dataset Cleaning and Management

Dataset Cleaning and Restoring Instructions

In Part 2 of your final project, you'll focus on preparing your dataset for analysis by uploading it into PGAdmin and ensuring it's clean and organized. Proper data cleaning and restoration are crucial steps that significantly impact the quality of your analysis and findings.

Uploading and Restoring CSV Files into PGAdmin:

1. Prepare Your CSV File: Ensure your CSV files are structured correctly, with the first row containing column headers that match the column names in your PostgreSQL table.

2. Open PGAdmin and Connect to Your Database: Launch PGAdmin and connect to the database where you want to import your CSV file.

3. Create a Table: Before importing, you need a table with a structure that matches the CSV file. Use the SQL editor to create a table, specifying the appropriate data types for each column.

Ex.

```
CREATE TABLE your_table_name (
    column1 datatype,
    column2 datatype,
    column3 datatype,
    ...
);
```

4. Import the CSV File: Navigate to the table you created within PGAdmin's browser panel, right-click it, and select the Import/Export option. Choose to import, specify the path to your CSV file, and configure the import options (e.g., delimiter, quote character). Ensure the columns in the CSV align with the table's structure and initiate the import.

5. Verify the Data: After importing, run a simple SELECT query to verify the data has been imported correctly. For further assistance, please contact your instructor.

2A.1: Describe Your Strategy for Preparing Your Dataset:
Considering the diverse methods available for preparing and incorporating your dataset into PGAdmin, describe your comprehensive strategy for dataset restoration. Will your approach involve creating a new database or utilizing an existing one found online? If your data is in CSV format, how do you plan to integrate it into the database? Alternatively, if you have access to a pre-packaged database file (such as a `.tar` file) that aligns with your project topic, explain how you intend to restore it. Ensure to reflect on whether the structure of the chosen or created database accurately represents the needs of your topic/project.

The dataset already came as a CSV and was very easy to load into PGAdmin.

2A.2: Describe the organization and cleanliness of your dataset.
Is your dataset well-structured and clean, or does it contain inconsistencies, missing values, or other issues that will require attention before analysis? Do you have any strategies in place to ensure the organization or cleanliness of at least the most meaningful parts of your dataset?

I was able to remove all missing values and repeated rows during the Python class. It is also filtered to only show the top 15 fast food restaurants in the USA which will allow me to focus my data and look at larger patterns.

2A.3: If your dataset comprises tables from multiple sources (i.e. databases), describe how you plan to integrate these into a single database. What challenges do you anticipate in ensuring consistency across different data sources, and how will you address them? If your dataset does not include tables or data from multiple sources, then simply state **'Not Applicable'** in your response below.

I will have to do a few JOINS to combine the fast food table with the income table. There will also be a lot of GROUP BY and ORDER BY to find which zip codes have the highest rate of fast food restaurants.

2A.4: If creating CSV files is necessary for your dataset (e.g., for AI-generated data or data extraction from non-relational sources), outline the process you will use to prepare these files. Include how you'll structure the data, define column headers, and ensure the data types are compatible with your database schema.

N/A

Part 3: Query Development and Testing

With your dataset ready, begin designing and crafting five queries designed to explore and uncover insights related to your chosen topic. This part emphasizes the practical application of SQL concepts learned throughout the course, testing your ability to retrieve meaningful information from your data. With your dataset now restored and ready within PGAdmin, it's time to craft and execute queries that will bring you closer to answering the big question driving your project.

3A: Central Question

3A.1 What is the Central Question of Your Project?

Begin by clearly stating the overarching question or objective your project aims to address. This main question will guide the development of your scenarios and queries.

What zipcodes in NYC would be considered food deserts/swamps and is there a correlation with the income levels of those zipcodes?

3B: Developing Scenarios and Queries

For the next steps, consider various angles or sub-questions that, when combined, will help you answer the main question. Each scenario should represent a piece of the puzzle, contributing valuable insights toward your final analysis.

Guidelines for Success

1. **Relevance:** Ensure each scenario closely aligns with your main question and that, collectively, they cover the breadth of your topic.
2. **Clarity:** Each scenario should be clearly defined, with a direct link to how it aids in answering the overarching question.
3. **Feasibility:** The queries should be executable within your dataset's structure and limitations. Opt for clarity and efficiency in your SQL syntax.
4. **Insight:** Focus on crafting queries that not only retrieve data but also provide actionable insights, contributing to a well-rounded understanding of your topic.
5. **Reflection:** After developing each query, briefly reflect on what the expected outcome is

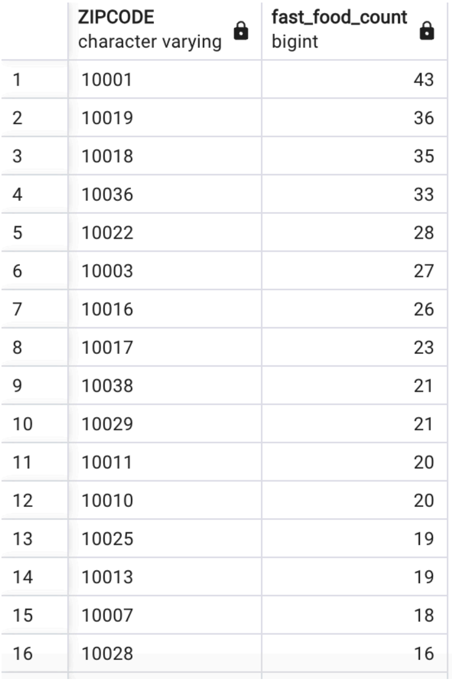
and how it will help in answering the main question.

This approach ensures that your querying process is methodical and purpose-driven, directly supporting your project's objectives and paving the way for a compelling analysis in your final presentation.

3B.1 Describe the first scenario that will help answer your main question. What specific aspect of your topic does it explore? Present the SQL query that you would use to investigate this scenario.

What number of fast food restaurants exist within each zip code?

```
SELECT "ZIPCODE", COUNT("DBA") AS
fast_food_count

FROM public."Restaurant_Data"

GROUP BY "ZIPCODE"

ORDER BY fast_food_count DESC
```



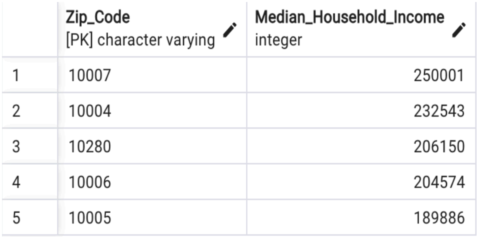
	ZIPCODE character varying	fast_food_count bigint
1	10001	43
2	10019	36
3	10018	35
4	10036	33
5	10022	28
6	10003	27
7	10016	26
8	10017	23
9	10038	21
10	10029	21
11	10011	20
12	10010	20
13	10025	19
14	10013	19
15	10007	18
16	10028	16

Before looking at any income levels of these zip codes, I just wanted to get a general understanding for where the highest concentration of fast food restaurants might be. The above query allowed me to count the number of fast food restaurants in each zip code and then sort from the highest concentration to the lowest concentration. My original theory was that we would see a high concentration of these restaurants in more uptown neighborhoods where income levels tend to be lower and where fast food companies would want to profit off of the convenience factor of fast food. However, the highest concreation seems to be in Midtown. This could be attributed to the fact that most businesses and companies work out of Midtown and therefore they probably get a large number of customers from employees during their lunch breaks. I will take a deeper look through my second scenario.

3B.2 Outline the second scenario. How does it build upon the previous scenario or introduce a new facet relevant to your main question? Include the SQL query designed to retrieve this information.

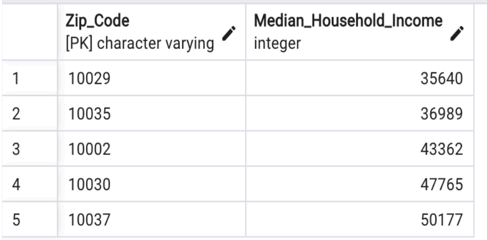
What zip codes are the wealthiest and poorest?

```
SELECT "Zip_Code", "Median_Household_Income"

FROM public."Income_By_Zipcode_NYC"

ORDER BY "Median_Household_Income" DESC

LIMIT 5;
```



	Zip_Code [PK] character varying	Median_Household_Income integer
1	10007	250001
2	10004	232543
3	10280	206150
4	10006	204574
5	10005	189886

```
SELECT "Zip_Code", "Median_Household_Income"

FROM public."Income_By_Zipcode_NYC"

ORDER BY "Median_Household_Income" ASC

LIMIT 5;
```



	Zip_Code [PK] character varying	Median_Household_Income integer
1	10029	35640
2	10035	36989
3	10002	43362
4	10030	47765
5	10037	50177

The above queries are used to get a better understanding of which neighborhoods would be considered high income vs low income. All I had to do was order by the median household income by descending and ascending and then limit by 5 to get the highest earning zip codes and lowest earning zip codes. As of right now I'm not seeing a lot of correlation between income and high fast food restaurant concentration. But my next query will allow me to dig into this further.

3B.3 Explain the third scenario, focusing on how it contributes to understanding your main question. Share the corresponding SQL query you've developed for this scenario.

What zip codes have the highest and lowest concentration of supermarkets?

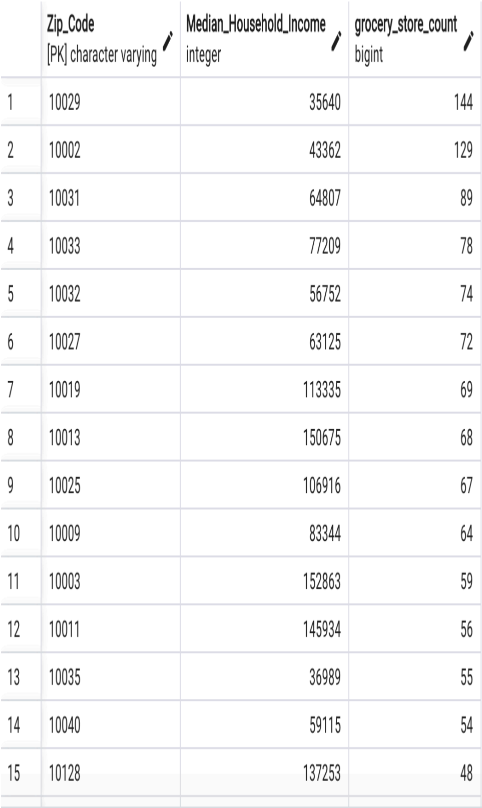
```
SELECT
    r."ZIPCODE",
    COUNT(r."DBA") AS fast_food_count,
    i."Median_Household_Income",
    i."High_Income_Households"
FROM
    public."Restaurant_Data" r
JOIN
    public."Income_By_Zipcode_NYC" i
ON
    r."ZIPCODE" = i."Zip_Code"
GROUP BY
    r."ZIPCODE", i."Median_Household_Income", i."High_Income_Households"
ORDER BY
    fast_food_count DESC;
```

The above query is used to relate the median household income of each zip code to the concentration of fast food restaurants within the zip codes. This is done by joining the income by zip code table with the restaurant table by the zip code. Once this is done I was able to group by the median household income and the percentage of high income households. I then ordered by the concentration of fast food restaurants in each zip code. As I had noticed with my first query, a lot of the highest concentrations of fast food restaurants are in Midtown. This makes me believe that something I had not considered when I originally decided to work on this project is the fact that NYC is one of the most important cities in the world in terms of employment. People from all over the Tristate area commute everyday to work in Manhattan and this creates a high demand for easily accessible food options for lunch. This makes it difficult to only consider income levels of these zip codes while looking to identify food swamps.

3B.4 Detail the fourth scenario, highlighting its importance in the context of your project's main question. Provide the SQL query that will help you explore this aspect of your topic.

Do the areas with the least supermarkets have lower median incomes or fewer high-income households?

```
SELECT
    i."Zip_Code",
    i."Median_Household_Income",
    COUNT(g."DBA Name") AS grocery_store_count
FROM
    public."Income_By_Zipcode_NYC" i
LEFT JOIN
    public."Grocery_Stores" g
ON
    i."Zip_Code" = g."Zip Code"
GROUP BY
    i."Zip_Code", i."Median_Household_Income"
ORDER BY
    grocery_store_count DESC;
```



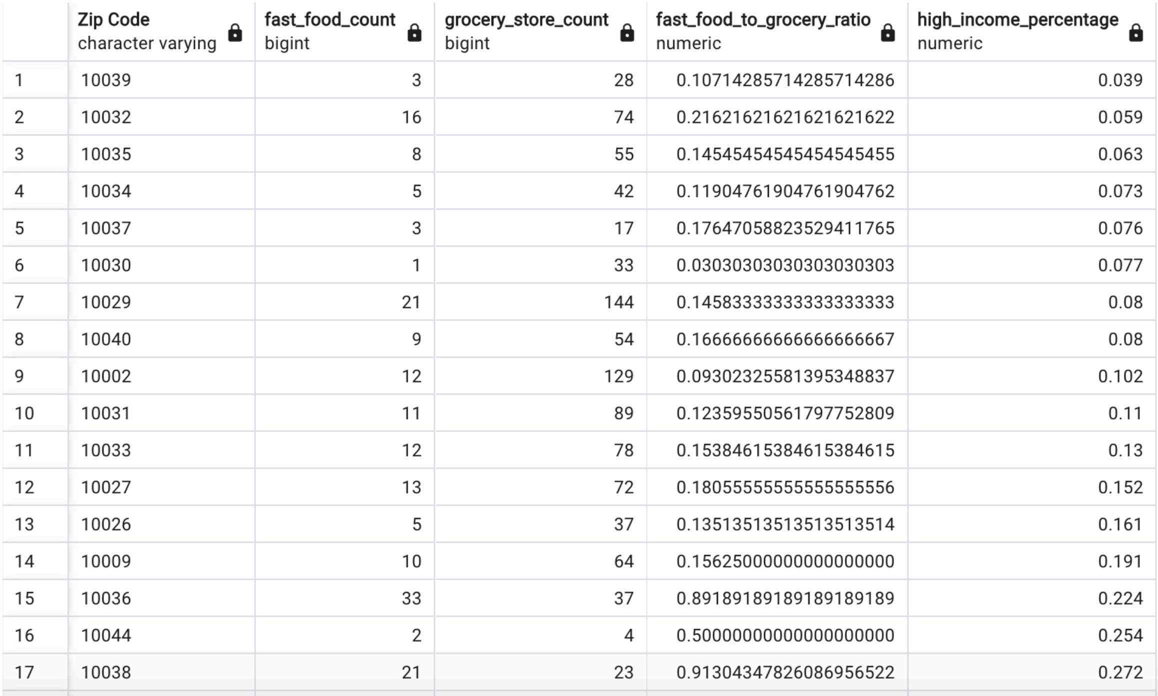
	Zip_Code [PK] character varying	Median_Household_Income integer	grocery_store_count bigint
1	10029	35640	144
2	10002	43362	129
3	10031	64807	89
4	10033	77209	78
5	10032	56752	74
6	10027	63125	72
7	10019	113335	69
8	10013	150675	68
9	10025	106916	67
10	10009	83344	64
11	10003	152863	59
12	10011	145934	56
13	10035	36989	55
14	10040	59115	54
15	10128	137253	48

The above query is used to relate the household median income to the concentration of grocery stores. This is done by joining the income by zip code table with the grocery store table by zip code count. Once this was done I could group by zip code and median household income and order by the grocery store count in descending order to see which zip codes had the highest concentrations of grocery stores. From the preliminary comparison it looks like NYC does a pretty good job of putting grocery stores in low income neighborhoods, which makes healthier food more readily available and decreases food insecurity.

3B.5 Describe the fifth and final scenario, indicating how it rounds out the insights needed to address your main question comprehensively. Present the associated SQL query.

How does the concentration of grocery stores compare to concentration of fast food restaurants?

```sql
SELECT

    g."Zip Code",

    COUNT(DISTINCT r."CAMIS") AS fast_food_count,

    COUNT(DISTINCT g."License Number") AS grocery_store_count,

    (COUNT(DISTINCT r."CAMIS")::decimal / NULLIF(COUNT(DISTINCT g."License Number"), 0)) AS fast_food_to_grocery_ratio,

    i."High_Income_Households" AS high_income_percentage

FROM

    public."Grocery_Stores" g

LEFT JOIN

    public."Restaurant_Data" r

ON

    g."Zip Code" = r."ZIPCODE"

LEFT JOIN

    public."Income_By_Zipcode_NYC" i

ON

    g."Zip Code" = i."Zip_Code"

GROUP BY

    g."Zip Code", i."High_Income_Households"

ORDER BY

    high_income_percentage ASC;
```



	Zip Code character varying	fast_food_count bigint	grocery_store_count bigint	fast_food_to_grocery_ratio numeric	high_income_percentage numeric
1	10039	3	28	0.10714285714285714286	0.039
2	10032	16	74	0.21621621621621621622	0.059
3	10035	8	55	0.14545454545454545455	0.063
4	10034	5	42	0.11904761904761904762	0.073
5	10037	3	17	0.17647058823529411765	0.076
6	10030	1	33	0.03030303030303030303	0.077
7	10029	21	144	0.14583333333333333333	0.08
8	10040	9	54	0.16666666666666666667	0.08
9	10002	12	129	0.09302325581395348837	0.102
10	10031	11	89	0.12359550561797752809	0.11
11	10033	12	78	0.15384615384615384615	0.13
12	10027	13	72	0.18055555555555555556	0.152
13	10026	5	37	0.13513513513513513514	0.161
14	10009	10	64	0.15625000000000000000	0.191
15	10036	33	37	0.89189189189189189189	0.224
16	10044	2	4	0.50000000000000000000	0.254
17	10038	21	23	0.91304347826086956522	0.272

The above query is used to analyze the relationship between the concentration of grocery stores and the percentage of high-income households in different zip codes in NYC. This is achieved by joining the grocery store data with the income data on the zip code, allowing for a comprehensive comparison. Once the data is joined, the query groups the results by zip code and the percentage of high-income households. This grouping enables the aggregation of the number of grocery stores and the calculation of the ratio of fast food restaurants to grocery stores within each zip code. The results are ordered by the count of fast food restaurants in descending order to identify areas with the highest concentration of fast food relative to grocery stores. From what I can understand of the data now, there doesn't seem to be much of a correlation between the ratio of fast food places to grocery stores in relation to income level. Further regression analysis will likely be needed to paint a full picture of the relationship between these three variables.

Part 4: Final Presentation

Prepare your final presentation of your project to be delivered by a short 5-15 minute recording, depending on your preference. This presentation should highlight the key insights discovered through your queries, and the relevance of your findings to your topic, and demonstrate your analytical prowess. Creativity and clarity in communicating your data story are crucial. Your presentation should be a concise yet informative reflection of your work, findings, and the relevance of SQL skills to your chosen field or interests. Do not feel pressured to conduct your project or analysis if you are still learning how to perform queries. Take the opportunity to instead discuss, in your presentation, the analytical (thinking) process you would experience if you were to conduct the project with a specific set of funding, data, and time. What type of analysis would you perform? What type of data would you want to source? What is the central problem-space that you are trying to solve, or understand? What type of questions, and queries would you ask to understand this problem?

Presentation Tips

- **Choose to present with 2 slides, or to consolidate all of the presentation requirements highlighted below into 1 slide.**
- Clarity and Conciseness: Aim to clearly articulate your points within the slide limit. Use bullet points or short paragraphs to maintain clarity.
- Visuals: Incorporate charts, graphs, or tables to visually represent key data points or findings. Visual aids can make your insights more accessible and engaging. ● Reflection: Your presentation should not only display what you did but also reflect on the learning process. Highlight any SQL skills you've developed and how they applied to solving real-world problems.
- Engagement: Consider your audience. Explain SQL concepts or project-specific jargon in layman's terms to ensure your presentation is engaging and understandable to everyone.

Examples:

Slide 1: Introduction and Motivation

- Personal Introduction: Begin by introducing yourself. Share your full name, where you are from, a professional picture of yourself on your slide (optional), and a brief overview of your background (e.g. personal interests and aspirations, current job/occupation, etc).
- Highlight how you envision SQL or other computing/programming skills in the certification enhancing your career or contributing to advancements within your industry area.
- Future Aspirations: Discuss what you hope to achieve with the knowledge and skills acquired upon completing the certification program. How do you plan to apply what you've learned in your current role, future career, or personal projects?
- Central Question: Present the central question that guided your project. Explain the rationale behind topic selection. Was it driven by personal passion, a gap in existing knowledge, or a specific industry challenge?

Slide 2: Understanding the Data

- Data Sourcing: Share where and how you sourced your dataset. Was it through online research, AI-generated, or another method? Discuss any challenges you encountered in this process.
- Dataset Overview: Provide an overview of your dataset's structure. Mention the number of tables and key columns, and why this dataset was apt for your project.

- Querying Process: Summarize how you approached querying the database. What were the key scenarios you explored? Highlight any SQL techniques or concepts that were particularly useful.
- Final Insights: Conclude with the main insights or answers revealed through your analysis. Reflect on how these findings contribute to understanding your central question and any implications or recommendations that emerge from your data.

4A: Complete your Presentation

Complete your presentation slides using Google Slides, PowerPoint, or any other software that you prefer. It is okay to have 1-8 slides, but don't feel pressured to have so many slides. Quality over quantity!

Part 5: Course Reflection and Professional Outlook

Conclude your project by reflecting on the learning journey throughout the course and the entire certification program. Discuss how the skills acquired, particularly in SQL and data analysis, align with your professional aspirations or career choices. This part is an opportunity to articulate the impact of your newly gained knowledge on your future endeavors in the data field.

5A.1 What has been your most significant learning milestone in the Data, Databases, & SQL course, and why? Reflect on the moment or concept that stood out to you as a pivotal point in your learning journey. How did it change your approach or understanding of data analysis?

One of the most significant milestones in this course was mastering the concept of JOINS. Before this, my understanding of SQL was somewhat limited; I could perform basic queries and manipulate single datasets, but I struggled to see the bigger picture. Learning about JOINS definitely honed my approach to data analysis. It helped me create a larger, more cohesive story than before and I could explore more complex questions that I previously thought were beyond my reach.

5A.2 How do the SQL and data analysis skills acquired during the course align with your current or future career aspirations? Discuss how the knowledge gained through this course fits into your professional goals. Are there specific roles, industries, or projects where you see these skills being particularly useful?

The SQL and data analysis skills I've acquired during this course are crucial as I pursue a career in data analytics. My passion for public health and my newfound skills helped me realize that data plays in shaping health policies and interventions. The ability to analyze and interpret data effectively will enable me to identify trends, assess health outcomes, and contribute to evidence-based decision-making in the public health sector. In addition to public health, I see these skills being highly applicable in the fields of politics which I am also very passionate about. In politics, data analytics can provide insights into voter behavior, public sentiment, and the impact of policy changes, allowing for more informed campaigning and governance. Overall, I'm hoping that these skills will allow for me to create meaningful change, one way or another, in my community.

5A.3 Can you identify a real-world problem or project where you can apply the SQL skills you've learned? Imagine a scenario in your current job, future career, or even a personal project where SQL could be used to solve a problem or provide insights. Describe what that would involve.

I think the current project I am working on about food insecurity shows that these skills in SQL are highly applicable to the real world. By the end of this project I'm hoping to compile data on all the boroughs in NYC so better understand where assistance for food insecurity may be needed. With SQL I can use multiple datasets to see which neighborhoods don't have good access to fresh food and others may be able to use those findings to implement community programs.

5A.4 What are the next steps in your data analysis or SQL learning journey? Consider what areas you wish to explore further or additional skills you want to acquire. Are there specific topics, technologies, or courses you plan to pursue next?

I would like to get better at doing more complex queries as I start adding more datasets into the mix for this project. I would also like to use these datasets in tangent with come data visualization tools like Tableau so that I can get a better sense of the relationships within the data. Finally, I think some ML techniques would also be useful to flesh out this project some more.

5A.5 How do you envision sharing or utilizing your SQL knowledge to benefit others or contribute to a community or organization? Think about ways you could use your skills to mentor others, contribute to community projects, or improve processes or decision-making within an organization.

I would like to share my findings about food insecurity in NYC with those working directly in the community so that they will be able to more easily target communities in need.