

# **Project Guide for the Blue Hen Data Analytics Club**

## **1. Define Your Data Analysis Question**

**Identify an Area of Interest:** Start by thinking about topics or areas you are passionate about or curious to explore. Consider fields such as sports, entertainment, health, finance, social media, marketing, or any other domain that excites you.

**Formulate a Data Analysis Question:** Based on your area of interest, formulate a specific question that you want to answer through data analysis. A good data analysis question should be:

- **Clear and Specific:** Clearly defines what you want to know or understand.
- **Measurable:** Can be answered using data.
- **Relevant:** Addresses a significant aspect of the topic or field you are interested in.

**Determine the Data Requirements:** Think about what data you need to answer your primary question and sub-questions. Consider the following:

- **Type of Data:** Will you need numerical data, categorical data, time-series data, or a combination?
- **Variables:** What are the key variables or metrics you need? (e.g., player statistics, match outcomes, weather conditions).
- **Scope and Time Frame:** Do you need data for a specific period, geographical region, or group of individuals?

Write down the data requirements clearly to guide your search for a suitable dataset.

**Ensure Feasibility:** Review your question and data requirements to ensure they are feasible within the scope of this project. Consider the following:

- **Data Availability:** Can you find or access data that meets your requirements?
- **Complexity:** Is the question suitable for your current skill level in data analytics? Start with a question that is challenging but achievable.
- **Time Constraints:** Ensure that the question can be answered within the time allocated for this project.

## **Break Down the Question into Sub-Questions:**

- Decompose your primary question into smaller, more manageable sub-questions. These will help guide your data exploration and analysis.
- For example, if your main question is about factors influencing soccer team performance, sub-questions could include:
  - “How does player experience affect match outcomes?”
  - “What is the impact of home vs. away games on team performance?”

- “How does weather influence the number of goals scored?”
- 

## 2. Data Preparation and Analysis in Python

**Select Your Dataset for Analysis:** Find a dataset that aligns with your interests (e.g., career interest, personal pursuits). Here are some sources where you can find free datasets:

- [Kaggle Datasets](#)
- [UCI Machine Learning Repository](#)
- [Google Dataset Search](#)
- [Data.gov](#)

The dataset should have at least 10-15 columns and 1000+ rows to ensure sufficient data for analysis. You can also organize your data across multiple sheets.

### Data Cleaning in Python:

- **Remove Duplicates:** Identify and remove duplicate records using the `drop_duplicates()` method.
- **Handle Missing Values:** Use methods such as removing records (`dropna()`), filling with the mean/median/mode (`fillna()`), or other appropriate techniques.
- **Convert Data Types:** Ensure data is in the correct format (e.g., using `to_datetime()` for dates or `to_numeric()` for numeric types).
- **Data Validation:** Ensure data accuracy by applying validation rules (e.g., using conditions to filter data within a specific range).

### Data Manipulation in Python:

- **Data Aggregation:** Use functions like `groupby()` and aggregation methods (`sum()`, `mean()`, etc.) to calculate metrics (e.g., total sales by category, average customer rating by region).
- **Lookup Functions:** Use `merge()` to:
  - Combine data from multiple DataFrames based on a common key (e.g., customer IDs).
  - Extract specific information from a DataFrame.
- **Sorting and Filtering:** Apply sorting with `sort_values()` to organize data (e.g., ascending/descending order) and use filtering techniques (e.g., Boolean indexing) to focus on specific subsets (e.g., by category or date range).

## Data Enrichment in Python:

- **Create New Calculated Fields:** Use mathematical operations or functions to create new fields that add value to the dataset (e.g., adding two columns together).
  - **Conditional Fields:** Use the `apply()` method with custom functions or `np.where()` to create new categories (e.g., categorize sales as "High," "Medium," or "Low" based on sales amount).
  - **Mathematical Calculations:** Perform calculations like growth rates, profit margins, percentage changes, etc., using arithmetic operations.
  - **Date Calculations:** Extract components from date fields (e.g., month, year, day) using `dt` accessor functions (`dt.month`, `dt.year`, `dt.day`).
- 

## 3. Creating Visualizations in Tableau

**Open Tableau Desktop:** Launch Tableau Desktop on your computer.

### Connect to Data Source:

- Click on the “Connect” pane on the left side of the start page.
- Choose the appropriate file type that matches your cleaned dataset:
  - For Excel: Click on “Microsoft Excel” and browse to select your file.
  - For CSV: Click on “Text File” and select your CSV file.
  - For other formats (e.g., database files), choose the appropriate connector and follow the instructions.

### Load Data:

- Once the data file is selected, Tableau will display the data source page.
- Drag and drop the worksheet(s) from the left pane to the “Drag sheets here” area.
- Review the data preview to ensure the data is correctly imported. Check for any errors in data types or formatting.

### Perform Initial Data Checks:

- Check the “Data Grid” at the bottom of the Data Source page to verify that all data types (e.g., strings, numbers, dates) are correctly identified by Tableau.
- Make any necessary adjustments, such as changing the data type by clicking on the data type icon next to each column.

**Save Data Source:** Click on the “Sheet 1” tab at the bottom to move to the worksheet area. This will save your data source and allow you to start building visualizations.

**Create Visualizations:** Transform raw data into meaningful insights through visual representations. Here are some examples of visualizations:

- **Bar Chart:** Visualize categorical data to compare values across different categories (e.g., Sales by Region, Product Category vs. Profit).
- **Line Chart:** Visualize trends over time (e.g., Monthly Sales Trends, Stock Prices over Time).
- **Scatter Plot:** Visualize the relationship between two numerical variables (e.g., Sales vs. Profit, Customer Age vs. Purchase Amount).
- **Map Visualization:** Visualize geographical data (e.g., Sales by Country or City, Store Locations).
- **Heatmap:** Visualize data intensity and patterns (e.g., Sales by Product and Month, Customer Engagement by Day and Hour).

**Other Visualization Types:**

- **Pie Charts/Donut Charts, Treemaps, etc.**

**Create Filters and Parameters:**

- **Add Filters:** Allow users to filter data in the visualizations to enhance interactivity.
- **Create Parameters:** Enable users to change the parameters of the visualizations for advanced control.

**Use Calculated Fields to Enhance Insights:**

- **Basic Calculated Fields:** Add new calculated fields to the dataset to derive additional insights (e.g., profit margin, growth rate).
- **Advanced Calculated Fields:** Create more complex calculated fields to perform advanced calculations (e.g., cumulative totals, year-over-year growth).

## **4. Building Dynamic Dashboards in Tableau**

**Create a New Dashboard:**

- **Open a New Dashboard:** Click on the “New Dashboard” icon at the bottom of the Tableau window. This will open a blank dashboard canvas where you will assemble your visualizations.
- **Set the Dashboard Size:** In the “Dashboard” pane on the left, choose an appropriate size for your dashboard based on the device and platform (desktop, tablet, mobile) where it will be viewed most frequently.

**Organize the Layout:**

- **Determine the Dashboard Structure:** Before adding visualizations, decide on a structure for your dashboard:
  - **Top Section:** Title and overview metrics (KPIs).
  - **Middle Section:** Main visualizations (e.g., trends, comparisons).
  - **Bottom Section:** Filters, parameters, and additional details.
- Use a logical flow that guides the user through the data story, typically from left to right or top to bottom. Utilize multiple sheets to enhance the organization and presentation of your dashboard.

#### **Use Containers to Arrange Elements:**

- Use horizontal and vertical containers to group elements together. Horizontal containers place elements side by side, while vertical containers stack elements on top of each other.

#### **Add a Dashboard Title:**

- Click on the "Title" checkbox at the top of the dashboard pane to add a dashboard title.

#### **Add Text Boxes for Context:**

- Drag a "Text" object from the "Objects" pane to the desired location on the dashboard.

#### **Use Annotations for Key Insights:**

- Add annotations directly to visualizations by right-clicking on a specific data point and selecting "Annotate." Customize the annotation text, style, and placement to make it visually distinct and clear.

#### **Consistent Visual Style:**

- Ensure all visualizations have a consistent font style, size, and color scheme. Use the same color palette for similar data categories to avoid confusion.
- Align all elements properly to create a neat, organized appearance. Use Tableau's grid lines and alignment tools to help with positioning.