CRISP-DM Template for SDG Dashboard Design

This template will guide you through the CRISP-DM process while designing a dashboard focusing on one of the Sustainable Development Goals (SDG) indicators. Each section corresponds to a step in the CRISP-DM cycle. Provide detailed explanations and documentation of your work at each stage.

1. Business Understanding

In this section, define the objectives of your dashboard, which is the data-driven research question. Identify the specific SDG indicator you will focus on and explain the importance of this indicator to the broader SDG goals. Clearly state what you aim to achieve with the dashboard.

Key Questions to Address:

- - What is the purpose of the dashboard?

  The purpose of the dashboard is to highlight the number of people rendered homeless by natural disasters and explore the potential for predicting natural disasters amid climate change. It also aims to emphasize the broader social impact of natural disasters.

- - Which SDG indicator are you focusing on?

  The SDG indicator I'm focusing on is 13.1.1, under the "Climate Action" goal

- - What insights do you hope to provide with this dashboard?

  I aim to illustrate the impact of natural disasters on homelessness and underscore that, without action, the frequency of these events may increase, leading to more people losing their homes.

2. Data Understanding

Describe the data you are using to build the dashboard. Explain the sources of the data, the variables included, and any initial findings from exploring the data. Highlight any data quality issues or unusual patterns.

Key Questions to Address:

- - What are the sources of your data?

  My main source of data is UN Office for Disaster Risk Reduction from the SDG site.

  Sources:

Our World in Data based on EM-DAT, CRED / UCLouvain, Brussels, Belgium – www.emdat.be (D. Guha-Sapir) – processed by Our World in Data. "Number of deaths from disasters" [dataset]. Our World in Data based on EM-DAT, CRED / UCLouvain, Brussels, Belgium – www.emdat.be (D. Guha-Sapir) [original data].

EM-DAT, CRED / UCLouvain (2024) – with major processing by Our World in Data. "Number of recorded natural disaster events – EM-DAT" [dataset]. EM-DAT, CRED / UCLouvain, "Natural disasters" [original data]. Retrieved October 30, 2024 from https://ourworldindata.org/grapher/number-of-natural-disaster-events

Contains modified Copernicus Climate Change Service information (2019) – with major processing by Our World in Data. "Global temperature anomalies by month" [dataset]. Contains modified Copernicus Climate Change Service information, "ERA5 monthly averaged data on single levels from 1940 to present 2" [original data]. Retrieved October 30, 2024 from https://ourworldindata.org/grapher/global-temperature-anomalies-by-month

Contains modified Copernicus Climate Change Service information (2019) – with major processing by Our World in Data. "Temperature anomaly" [dataset]. Contains modified Copernicus Climate Change Service information, "ERA5 monthly averaged data on single levels from 1940 to present 2" [original data].

- - What variables are included in the dataset?

Natural disasters	Disaster type, Year, Count of disasters
Disaster impact	Country, Year, Number of Deaths, Injuries, Affected Population, Homelessness, Reconstruction costs, economic damages, death rates
Temperature	Month, Year, Temperature Anomaly, Country

- - Are there any data quality issues?

  In the country variable, entries included terms like "middle-income countries" or "world" in addition to specific country names. I created a key table of countries for filtering. The data had no blank or missing values.

- - What initial insights did you gain?

Floods and extreme weather were the most common disasters, peaking in 2021. Asia was the most affected continent, with India notably impacted in 2002. In 2005 and 2006, over 6 million people were left homeless due to natural disasters.

3. Data Preparation

Document the steps you took to clean, transform, and prepare the data for analysis. This may include removing duplicates, handling missing values, and creating new variables that will be useful for the dashboard.

Key Questions to Address:

- - What data cleaning processes did you perform?
  Firstly, I had to promote first rows to headers. Then I changed the type from texts to numbers because some numeric values were as text. Then I had to filter rows to only show data from the year 2000+. After that for the "Country_NDs_Info" I removed most of the columns because there were 224 columns but in reality, I was interested in 11 of them. Then in "death rates" column I replaced dot with comma because it was an issue for Power Bi and then I was able to change the type from text to number. Then I created a new column in this data set where I fixed the death rates, because some were over 100 which seemed unnecessary.
  For other data sets I also promoted first rows to headers, removed columns which had no values in them and sometimes renamed the columns for quality of life. I also changed the type of year because it was as text and not as number and I also changed the temperature anomaly into fixed decimal numbers.

- - How did you handle missing values?
  My data usually had 0 instead of missing values. That was difficult to determine whether they really had 0 deaths from natural disasters or if they just didn't bother to fill it in. So, I relied on the integrity of the data I worked with and left them at 0.
- - Did you create any new variables or features?
  I usually worked with all the variables I was given, but I did make a separate table for some variables which I wanted to observe individually. For example, I created a continent variable which helped me categorize countries by continents.

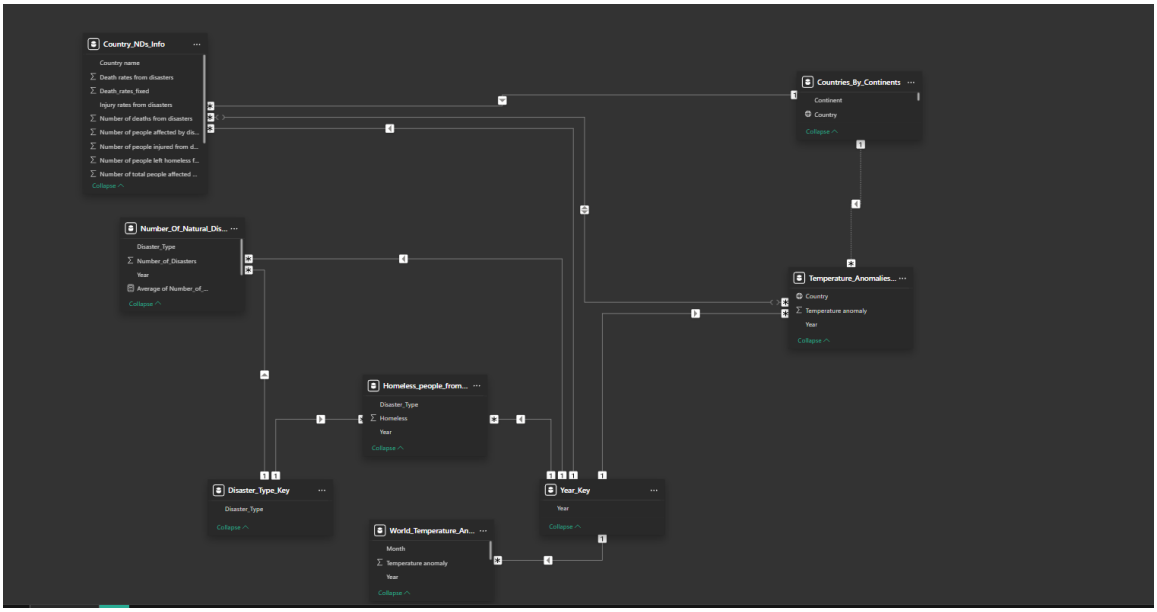
4. Modeling

Explain any models or algorithms you used to analyze the data. If your dashboard includes predictive elements, describe the models and how they were trained. If not, explain how you structured the data to create meaningful visualizations.

Key Questions to Address:

- - What models or algorithms did you use?

This is the model I used:



- - How did you ensure the model's accuracy?

  By defining exact links between important datasets and creating specific tables for essential keys like country, year, and disaster type, I was able to ensure the accuracy of the model. This method made sure that every data point could be properly checked against these reference tables and allowed me to maintain correct, relevant selections.

- - How did you structure the data for visualization?

  I structured the data in a snowflake schema, normalizing dimension tables to reduce redundancy and improve data integrity. This setup enabled a clear, organized data model where related attributes, such as country details or disaster classifications, are stored in separate, linked tables. This approach allowed for efficient filtering making it easier to visualize complex relationships and trends accurately.

5. Evaluation

Evaluate the results of your modeling or visualization efforts. Discuss whether the dashboard meets the original objectives set out in the business understanding phase. Identify any limitations and how they might impact the dashboard's effectiveness.

Key Questions to Address:

- - Does the dashboard meet the business objectives?

  Yes, the dashboard effectively displays the impact of natural disasters and provides quantifiable data on these effects. Although no strong correlations are yet present, the dashboard serves as a warning sign for future.

- - What are the limitations of your analysis?

    I feel like I could get more accurate results if some data values were more accurate in terms of what they mean, like if they are missing or if they are really a 0. I also think that to be more certain about my findings, this project needs to be further observed. I would say in the next 5 years with more precise data, it would harvest a better understanding whether there is really a correlation or not.

- - How could the dashboard be improved?

    Dashboard could be further improved by focusing on specific regions like Asia or Europe. I also think that finding more accurate and trustworthy data would also come a long way in improving this dashboard.

6. Deployment

Outline the steps required to deploy the dashboard for use. This may include publishing it to a platform like Power BI, sharing it with stakeholders, and ensuring it is regularly updated with new data.

Key Questions to Address:

- - How will you deploy the dashboard?

    I will deploy my dashboard using the standard Power Bi tools.

- - Who is the target audience for the dashboard?

    My target audience is people who are interested in climate change and natural disasters.

- - How will you ensure the dashboard stays up to date?

    The data I am using get updated once a year and the next update should be in April of 2025. After the update I will collect the data again, go through the same process of cleaning and implement them into my visualizations and calculate correlations or standard deviations.

7. Future Research

List down the steps that might address the shortcomings in your research. These could be availability of data, outdated research data.

Key Questions to Address:

- Are there emerging techniques or tools that could provide deeper insights?- Who is the target audience for the dashboard?

  With more data accuracy I could provide more meaningful conclusions and results, which would mean more for people who may not realize that they can also be impacted by natural disasters and end up homeless. Sadly, I was also not able to find data on what type of disaster occurred in a specific country. Maybe with that I would be able to provide more specific information. If there were a data about what types of people were affected (age, gender, nationality) it would be more impactful. There would be an improvement in impact if the data also contained which areas of economy were impacted by which natural disaster and when. Having more details about the data would definitely provide deeper insights.

- How might the research impact different stakeholders or communities?

  It could provide more awareness of this issue and shine light on people who end up homeless because of natural disasters. They might end up getting better care or even prevent it from happening. There is no simple way of preventing heavy rains, floods or extreme weathers and it can impact anyone. That's why this research should make people aware of the casualties it can bring.

This template is designed to help you systematically document your work at each stage of the CRISP-DM process, ensuring a comprehensive and well-structured approach to designing your SDG dashboard.