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

Objective of the Dashboard

The purpose of this dashboard is to explore potential relationships between global mental health trends, specifically major depressive disorder (MDD), and patterns in music streaming behavior over time. This analysis investigates whether certain attributes of popular music—such as its emotional tone or "valence"—might reflect broader societal shifts in mental health. By examining these patterns in music alongside mental health data, the dashboard seeks to provide insights that could support early detection of societal mood shifts, contributing to a broader understanding of public well-being. Importantly, this analysis does not aim to establish causation, but rather to observe potential correlations that may provide unique insights into public sentiment.

SDG Indicator Focus

This project aligns with Sustainable Development Goal (SDG) 3: Good Health and Well-being, with a specific focus on Indicator 3.4.2: Suicide mortality rate. While the IHME dataset primarily provides data on major depressive disorder rather than direct suicide mortality rates, MDD is widely recognized as a significant risk factor for suicide. Therefore, examining trends in MDD offers valuable insights into mental health challenges relevant to

SDG 3.4.2. This alignment supports the SDG goal of reducing premature mortality from non-communicable diseases through prevention, treatment, and improved mental health awareness.

Insights and Purpose of the Dashboard

The dashboard is designed to provide the following insights:

1. Global Trends in Mental Health: By visualizing the prevalence of MDD over time across various demographics and regions, the dashboard highlights global patterns and key shifts in mental health.

2. Music as a Proxy Indicator of Societal Mood: By analyzing characteristics of music—particularly valence (a measure of positivity)—this dashboard explores whether trends in music preferences reflect shifts in mental health indicators. This analysis could suggest music as a potential non-traditional, proxy indicator of societal mood, but without implying any causal relationship.

3. Cross-Cultural and Temporal Analysis: The dashboard facilitates comparisons across regions and time periods, enabling an exploration of whether societal mood, as expressed through music attributes, has any observable relationship with mental health indicators like MDD.

4. These insights could support public health officials in identifying early indicators of shifts in societal mood, potentially aiding in proactive mental health campaigns. For policymakers, understanding these correlations could inform broader social programs that address mental well-being.

Through this approach, the dashboard contributes to SDG 3 by offering an innovative way to examine mental health trends, leveraging music data as a complementary indicator. While correlations observed may yield valuable public health insights, they are not indicative of causation. Insights drawn from the dashboard should be used with caution and in conjunction with other established mental health indicators.

2. Data Understanding

2.1 Data Collection

The data for this project comes from two primary sources:

- **Mental Health Data:**

  The Global Burden of Disease (GBD) study from the Institute for Health Metrics and Evaluation (IHME) provides global estimates for major depressive disorder (MDD) by region, gender, and age group, spanning several years. This dataset is structured to allow for comparisons across demographics and time, making it suitable for observing trends in mental health. Originally spread across multiple CSV files, this dataset requires consolidation for effective analysis.

- **Spotify Streaming Data:**

  This dataset includes music track data from Spotify, containing various attributes such as valence, energy, tempo, and loudness, which are indicative of each track's emotional and acoustic characteristics. The data also includes each track's release year, allowing for time-based analysis. However, the popularity score is a snapshot at the time of collection (likely 2021), which restricts its use for historical trend analysis. This data will primarily be used to analyze the emotional tone of music over time.

2.2 Initial Data Exploration

Overview of Variables

The following are the main variables in each dataset:

- **Mental Health Data Variables:**
    - **location_name**: Represent countries/regions globally.
    - **age_name:** Denote different age groups.
    - **sex_name:** Indicate gender (1 for male, 2 for female).
    - **year**: Shows the observation year.
    - **metric_name**: Describes the type of metric (e.g., Number, Rate per 100,000, Percent).
- **Spotify Data Variables:**
    - **year**: Represents the release year of each track.
    - **energy**, **valence, tempo, loudness, danceability, and acousticness**: Describe the emotional characteristics of each track.
    - **popularity:** Represents the track's popularity at the time the dataset was collected.

2.3 Data Quality Assessment

Several data quality issues were identified during the initial inspection:

1. Redundant Columns in IHME Data: Some columns in the mental health dataset (e.g., measure_id, measure_name, cause_id) contained only one value across all rows, offering no analytical value and suggesting redundancy.

2. Structure and Consolidation Needs: The IHME dataset was spread across multiple CSV files, creating an inconsistency that required consolidation into a single dataset for effective analysis.

3. Mixed Metrics in a Single Column: The metric_name column in the IHME data combined multiple metric types (Number, Rate, Percent), complicating direct metric-specific analysis.

4. Unscaled Numeric Values: Some IHME metrics displayed extremely high values, likely due to large population counts. Additionally, Spotify attributes like energy, valence, and tempo required scaling adjustments to standardize ranges across variables.

5. Missing Values in Spotify Data: Certain Spotify features (e.g., loudness, tempo) had missing values, potentially due to variations in genre or recording quality. These missing values were initially noted as needing handling or imputation if necessary.

2.4 Insights

Initial data exploration provided the following insights:

- **Observations on Mental Health Trends**:

  The IHME data reveals rising trends in MDD prevalence over time, with differences across age groups and gender. These initial trends highlight the relevance of focusing on younger populations and females, who often show higher prevalence rates. This insight shapes the demographic focus of the analysis.

- **Shifts in Musical Characteristics Over Time**:

  Initial analysis of Spotify data suggests a decrease in valence (positivity) over recent

decades, indicating that popular music may be trending toward a more reflective or melancholic tone. This trend is relevant to the study as it aligns with the observed rise in mental health challenges.

- **Limitations in Temporal Popularity Trends**:

  The popularity metric in Spotify data is a snapshot from a specific point in time, which restricts its usefulness for trend analysis. Instead, characteristics like valence, energy, and tempo will serve as primary indicators for observing changes in musical tone over time.

- **Refinement of Location Data**: Upon review, it was noted that the location data included several non-country names (e.g., regions or undefined areas). A Python algorithm was applied to filter out rows without valid country names, greatly increasing optimization and decreasing file size.

3. Data Preparation

This section outlines the key steps taken to clean, transform, and organize the datasets to prepare them for analysis. The process involved merging multiple files restructuring the data, and creating a fact table along with supporting

```python
import pandas as pd
import glob
import os

# Path where all CSV files are located
file_path = "C:\\2024-25A-FAI1-ADSAI--FlorisLokhorst236918\\IMHE data"

csv_files = glob.glob(os.path.join(file_path, "*.csv"))

combined_df = pd.concat((pd.read_csv(file) for file in csv_files), ignore_index=True)

# new CSV file
combined_df.to_csv("fact_table.csv", index=False)

print("CSV files combined successfully into 'combined_output.csv'!")
input("press enter to exit")
```

figure 3.1

dimension tables to streamline data modeling and analysis in Power BI.

3.1 Data Cleaning and Transformation

- Mental Health Data (IHME):
    1. Removing Redundant Columns:
        - Certain columns (measure_id, measure_name, and cause_id) in the mental health data contained a single, value across all entries. These were removed to reduce file size and improve performance.
    2. Merging Multiple CSV Files:
        - Initially spread across 16 CSV files, the IHME data was consolidated into a single dataset using a Python script. This ensured consistent formatting and enabled smoother integration into Power BI. (see figure 3.1)
    3. Pivoting the metric_name Column:
        - The metric_name column contained multiple metric types (Number, Rate, and Percent) within a single field. To simplify analysis, a pivot transformation was applied to separate these metrics into individual columns, resulting in distinct fields for each metric.
    4. Scaling Numeric Values:
        - Some metrics in the IHME data displayed excessively high values (e.g., population counts in billions), requiring scaling adjustments. These values were divided by different multiples of 10 to make them align with reported values from the IHME and WHO websites. Calculation of Correlation

Coefficient: A custom DAX formula was implemented to calculate the correlation coefficient between music valence and MDD prevalence. This calculation is essential for identifying potential associations within the data and is displayed directly on the dashboard to support immediate interpretation by viewers.

5. **Removal of Age Slicer:**

   - An age slicer was initially included to allow filtering by age groups. However, functionality issues were discovered that affected data consistency when different ages were selected. Due to time constraints, the slicer was removed to maintain dashboard integrity and prevent misleading data interpretations.

   - Partial Resolution of Location-Based Issues: Certain location data inconsistencies were identified but were left partially unresolved due to project scope constraints. These remaining issues may affect some location-based insights but were deemed minor enough not to impact the overall analysis.

   - It was discovered that certain male or female values for certain years were still not correct. The following DAX formula was used to correct this:

```
Adjusted_Number =
IF(
    [sex_id] = 2 && [Number] < 50000000,
    [Number] * 10,
    IF(
        [sex_id] = 1 && [Number] < 1000000,
        [Number] * 100,
        IF(
            [sex_id] = 1 && [Number] < 10000000,
            [Number] * 10,
            [Number]
        )
    )
)
```

- Spotify Streaming Data:
-

1. **Removing Unnecessary Columns:**
    - Columns like ID, duration, and release_date were dropped from the dataset, as they did not contribute meaningfully to the analysis. The year field was retained to represent each track's release date and serve as the primary time variable.

2. **Normalizing Audio Features:**
    - Spotify attributes like energy, valence, tempo, and acousticness required scaling adjustments to standardize their ranges. These were divided by 1,000 to align them with comparable units and simplify analysis.

3.2 Data Modeling: Creating Dimension and Fact Tables

To facilitate efficient analysis in Power BI, the data was structured into a fact table supported by three dimension tables:

- **Fact Table:**
    - The main dataset for analysis containing key metrics and audio features from both the IHME and Spotify datasets.

- **Dimension Tables:**
    - dim_demographics: Includes demographic information, such as age groups and gender, to allow segmentation by these variables.

    - dim_location: Contains location information, including country and region, enabling geographical analysis.

- dim_time: Holds time-related information, primarily year, to support time-based analysis of trends.

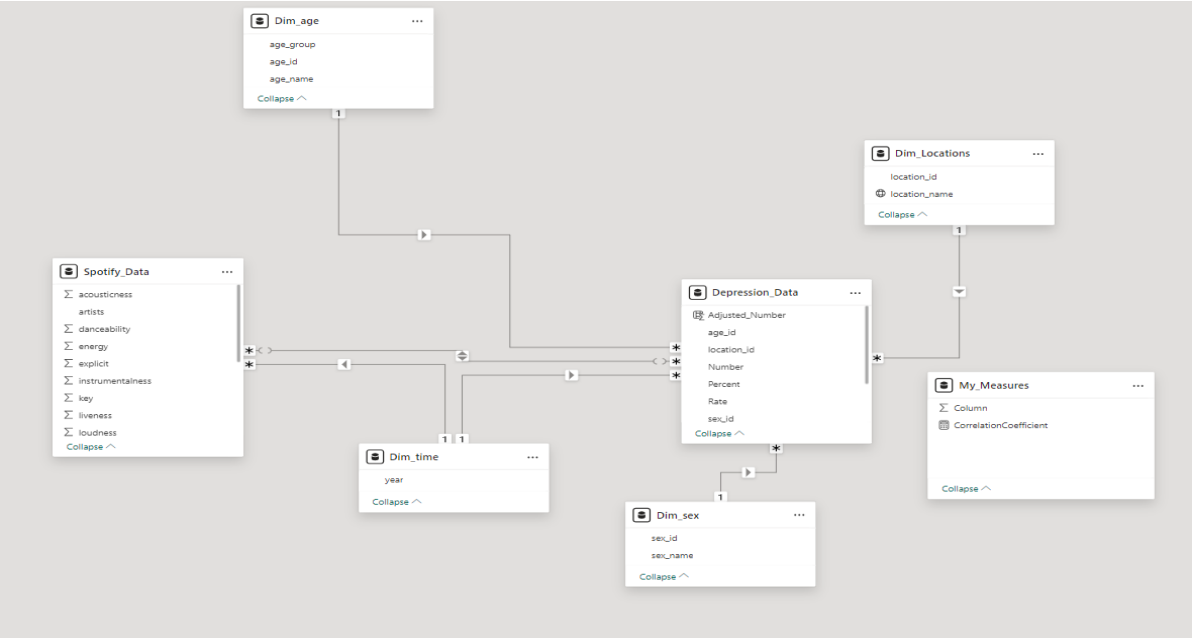

Figure 3.2: data structure

This star schema optimizes data retrieval and simplifies the analytical process, allowing more flexibility in slicing and filtering by demographic, location, and time

Additional Features:

- **Navigation Enhancements**: Interactive buttons were added to improve dashboard navigation. A home button was implemented to quickly return to the main page, and additional buttons allow users to switch between pages, enhancing user experience.

- **Valence/Energy Toggle**: A toggle button was added on the fourth page, allowing users to switch between valence and energy metrics. This feature provides flexibility in exploring the emotional tone of music, offering an additional dimension to the analysis.

.

4. Modeling

Since this project is focused on visualization rather than predictive modeling, the main objective in this phase was to structure the data effectively for analysis and to add analytical elements that enhance the dashboard's interpretability. Key efforts included calculating correlations and optimizing data organization to support the visual exploration of mental health trends and music characteristics over time.

Data Structuring for Visualization:

The datasets were structured using a star schema, consisting of a primary fact table and dimension tables for demographics, location, and time. This approach allows efficient filtering and slicing of data by age group, gender, region, and year, optimizing performance and enabling flexible data exploration.

Calculation of Correlation Coefficient

A custom DAX formula was used to calculate the correlation coefficient between music valence and MDD cases, providing a quantitative measure of association between these variables. This coefficient, displayed on the fourth page of the dashboard, offers a quick insight into the strength and direction of the relationship between music's emotional tone and mental health prevalence. While the correlation does not imply causation, it serves as a valuable indicator of association for further exploration.

Visualization Rationale:

The visualizations were chosen to intuitively represent different aspects of the analysis. Line charts illustrate trends over time, making it easier for users to observe changes in mental health prevalence or music valence. Scatter plots effectively show correlations, such as the association between valence and MDD, and are ideal for conveying the non-causal nature of these relationships.

Ensuring Accuracy of Calculations

To ensure the reliability of the calculated correlation coefficient and other data metrics:

- **Data Validation Checks**: Initial data cleaning processes included removing redundant or irrelevant entries (such as non-country locations) to maintain data consistency.
- **Scaling Adjustments**: Numeric values in the mental health data and Spotify attributes were scaled appropriately (e.g., dividing large population counts) to standardize ranges across variables. This alignment minimized the risk of skewed correlations due to scale mismatches.
- **Testing of DAX Calculations:** The custom DAX formula for the correlation coefficient was tested with smaller sample datasets to confirm accuracy. By verifying that the formula produced expected results on test data, confidence in its accuracy for the full dataset was ensured.
- **Valence and Energy Metric Toggle**

  A toggle button was added on the fourth page to allow users to switch between the valence and energy metrics, providing flexibility in analyzing different emotional characteristics of music. This feature enables a more comprehensive

view of music's role as a proxy for societal mood, allowing the user to observe potential patterns in both positivity (valence) and activity levels (energy).

Interactive Navigation

To enhance user experience, interactive navigation buttons were incorporated, including a home button to return to the main page and buttons to switch between pages. This design choice supports a seamless user journey across different aspects of the dashboard, making it easier to explore global mental health trends, demographic variations, and the relationship between music and mood.

Justification of Visual Structure

The dashboard was designed with a mix of line charts, scatter plots, and interactive elements to present data in an intuitive manner. Line charts illustrate trends over time, scatter plots display correlations, and interactive elements allow users to customize the analysis by switching between metrics or focusing on specific time periods. This visual structure was chosen to facilitate quick insights while maintaining a clear and engaging user interface.

5. Evaluation

This section evaluates whether the dashboard meets the initial objectives outlined in the Business Understanding phase. It also discusses the limitations encountered during the project and potential improvements that could enhance the dashboard's functionality and accuracy.

Meeting Business Objectives

The primary objective of this dashboard was to explore potential correlations

between mental health trends (specifically Major Depressive Disorder, MDD) and music streaming behavior, using emotional characteristics of music as a proxy for societal mood. The dashboard successfully presents these trends through:

- **Visualization of Global MDD Trends:** The dashboard effectively visualizes MDD prevalence over time and across demographic groups, allowing users to observe global and demographic-specific mental health patterns.

- **Analysis of Music Attributes:** By incorporating Spotify data on valence (positivity) and energy, the dashboard provides a visual representation of changes in musical tone over time. This supports the exploration of music as a potential indicator of societal mood.

- **Correlation Insight:** The inclusion of a calculated correlation coefficient between valence and MDD prevalence helps quantify the association between these variables. While not indicative of causation, this metric provides users with a basis for further investigation into societal mood trends. The negative correlation coefficient suggests a potential inverse relationship, where a lower valence in music correlates with an increase in MDD cases. This pattern could indicate that societal mood, as reflected in popular music, aligns with broader mental health trends

These elements collectively align with the project's goal of examining SDG 3 (Good Health and Well-being) through an innovative lens, leveraging music data to gain insights into mental health trends.

Limitations

Despite its strengths, the dashboard has some limitations that may impact its accuracy and the scope of insights:

- **Partial Location Data Issues:** Certain inconsistencies in the location data were only partially addressed, meaning some geographic insights may be affected. This limitation may reduce the reliability of certain location-specific interpretations.

- **Age-Based Analysis Constraints:** Initially, the dashboard included an age-based slicer, but this feature was removed due to functional issues. As a result, the dashboard cannot currently filter insights by age group, limiting the ability to perform detailed demographic comparisons.

- **Correlational Analysis Only:** This analysis is purely correlational and does not imply causation. While associations between musical characteristics and mental health trends can be observed, other factors such as socioeconomic or cultural influences may also play a role. Therefore, these correlations should be interpreted cautiously and in conjunction with other well-established mental health indicators.

Potential Improvements

Several improvements could enhance the dashboard's analytical depth and usability:

- **Expand Age-Based Filtering:** Resolving the issues with age-based filtering would add significant value, allowing users to view mental health trends and musical preferences for specific age groups.

- **Enhance Location Data Consistency:** Cleaning up the remaining location inconsistencies would improve the accuracy of geographic insights, providing a more reliable view of regional variations in mental health and music trends**.**

- **Add Other Health Indicators:** Incorporating additional health metrics (e.g., anxiety prevalence, suicide mortality rate) would enrich the analysis, aligning the dashboard more closely with SDG Indicator 3.4.2 and providing a broader view of mental health challenges.

- **Incorporate Predictive Analysis:** In future iterations, adding predictive analysis capabilities could enable the dashboard to anticipate trends in mental health based on shifts in music characteristics. For example, using time series forecasting on valence or energy levels in popular music might reveal patterns that precede changes in mental health trends. This would make the dashboard a more proactive tool for early detection of shifts in societal mood.

Overall, the dashboard effectively meets its primary goal, shedding light on the potential connections between mental health trends and shifts in music preferences. While it achieves this by highlighting patterns and correlations, there's room to make it even more impactful. With enhancements like age-based filtering, better location consistency, and predictive analysis, the dashboard could evolve into a real-time proxy indicator of societal mental well-being. These improvements would enable it to not only reflect current trends but also anticipate changes in public sentiment, offering a proactive tool for mental health monitoring and early intervention. By deepening its analytical capabilities, this dashboard could provide a more nuanced and timely view of how music might mirror and signal shifts

in collective mental health.

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?
- - Who is the target audience for the dashboard?
- - How will you ensure the dashboard stays up to date

Deployment Strategy

The dashboard will be published using Power BI's cloud service, making it accessible via a secure web link or Power BI workspace. This allows stakeholders to interact with the dashboard in real-time, exploring trends and insights at their convenience.

Target Audience

The primary users of this dashboard include mental health researchers, public health organizations, policymakers, and music industry analysts. Each of these groups may find unique value in the dashboard:

- **Mental Health and Public Health Organizations**: These stakeholders can use the dashboard to gain insights into societal mental well-being trends, potentially informing public health initiatives and awareness campaigns.

- **Policymakers**: By examining correlations between cultural trends and mental health, policymakers may gain a broader understanding of population sentiment, aiding in the design of preventative and supportive programs.

- **Music Industry Analysts**: For those in the music industry, this dashboard provides a novel way to examine how music consumption patterns may reflect or respond to societal mood shifts.

Long-Term Vision for Real-Time Updates

In the future, as real-time data sources become available for both music and mental health indicators, the dashboard could be adapted to support live updates. This would enhance its role as a potential real-time proxy indicator of societal mental well-being, allowing stakeholders to monitor and respond to emerging trends more quickly. For now, the dashboard will be manually updated regularly as well as enhanced functionality.

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?
- How might the research impact different stakeholders or communities?

- This section explores potential next steps to address limitations, expand the dashboard's analytical capabilities, and deepen its relevance as a tool for public health insights.
- **1. Incorporate Additional Mental Health Indicators**

  To enhance the depth of mental health analysis, future research could integrate additional indicators, such as anxiety prevalence, suicide mortality rate, and substance abuse data. These additional metrics would align with Sustainable Development Goal (SDG) Indicator 3.4.2 and offer a more comprehensive view of mental health trends. By examining a broader range of health indicators, the dashboard could provide a more complete picture of societal mental well-being.

- **2. Integrate Predictive Analytics**

  Introducing predictive analytics would enable the dashboard to forecast potential shifts in mental health trends based on patterns in music characteristics and historical mental health data. Time-series forecasting or machine learning models could predict how societal mood might evolve, adding a proactive element to the dashboard. This capability would transform the dashboard into a tool not only for observing past and present trends but also for anticipating future changes in mental well-being.

- **3. Improve Geographic and Demographic Granularity**

  Currently, the dashboard’s geographic and demographic filtering is limited by partial data inconsistencies and the removal of the age-based slicer. Future research could focus on refining these data points to allow for more detailed regional and age-specific analyses. This would enhance the dashboard’s usefulness for stakeholders interested in understanding mental health trends in specific communities or age groups.

- **4. Explore Emerging Techniques in Sentiment Analysis**

  Future iterations could incorporate sentiment analysis on song lyrics as a complementary measure to audio features like valence and energy. By analyzing the sentiment expressed in popular lyrics, the dashboard could provide a richer understanding of societal mood, potentially revealing insights that go beyond the musical tone. This could be especially valuable for exploring how lyrical themes in popular music might correlate with mental health trends.

- **5. Collaborate with Real-Time Data Sources**

  For the dashboard to function as a real-time indicator of societal mental well-being, collaborations with data providers would be essential. Partnering with platforms like Spotify or public health organizations could enable access to live-streaming or frequently updated data. Such collaborations would allow the dashboard to provide more immediate insights, supporting early intervention and timely responses to emerging mental health trends.

- **6. Broader Impact on Stakeholders and Communities**

  The insights generated by this dashboard have the potential to impact diverse stakeholders and communities. Public health officials and mental health organizations could leverage these insights to identify early indicators of societal mood shifts and target mental health resources accordingly. For policymakers, understanding correlations between cultural trends and mental health could inform more holistic social policies. Furthermore, music industry professionals could use these insights to gauge how societal moods influence musical preferences, potentially guiding more responsive artistic and marketing strategies.

By building on these future research directions, this dashboard could evolve into a powerful tool for understanding and responding to societal mental well-being, supporting informed decision-making across public health, policy, and industry.

References

Institute for Health Metrics and Evaluation (IHME). (2021). *Global Burden of Disease Study 2021 (GBD 2021) Results*. Seattle, WA: IHME, University of Washington. Retrieved from http://ghdx.healthdata.org/gbd-2021

Ay, Y. E. (2021). *Spotify Data* [Data set]. Kaggle. Retrieved from https://github.com/gabminamedez/spotify-data/blob/master/data.csv

ChatGPT was used for grammar control and formatting in this document. Prompt used: "please correct the grammar of this document and format it according to university conventions"