

Assessing Drivers of Anxiety and Depression in Widows: A Data-Driven Analysis

General Dataset Information

File Name: 2023_Mental_health_survey_HI Analysis clean.xls

Description: Mental Health Analysis and Baseline Survey of Widows

Dataset Details: Raw data: 624 rows / 49 columns; Baseline data: 625 rows / 99 columns

Size: 1,754 KB (1.8 MB)

Source: Global Fund for Widows (GFW)

Method of Collection: Survey data collected from widows (anxiety, depression, and baseline factors such as income, hunger, dependents).

Data Profile

Initial Data Import and Setup:

- Imported dataset into Excel containing 2 sheets:
 - **Raw Data:** Mental health analysis of 624, focusing on anxiety and depression.
 - **Baseline Data:** Baseline survey of 625 widows including factors like income, dependents, food security, and more.

Data Profiling Steps in Excel:

- **Structure and Formatting Check:** Verified the initial structure and formatting of both sheets and ensured consistency in column names and data types.
- **Key Metric Calculation:** Calculated key mental health metrics, including the minimum, maximum, and average for anxiety and depression scores, to assess the data's range and typical values.
- **Integrity Checks:**
 - Checked for 0 or negative values in critical columns such as age, number of dependents, income, and mental health scores (anxiety and depression).
 - Checked for missing or null values across columns.
 - Reviewed distinct values across key columns to identify inconsistencies or errors.
 - Ensured consistency in anxiety and depression scores between different survey instruments (e.g., GAD-7 and PHQ-9 scores).
 - Conducted mental health distribution analyses across different variables.

Observations and Inconsistencies:

- The dataset includes data from over 600 widows with detailed mental health scores (anxiety, depression) and baseline factors such as income, dependents, and food security.
- **Anxiety and Depression Scores:**
 - **Severe Anxiety:** Only 45 widows report severe anxiety.
 - **Severe Depression:** Only 25 widows report severe depression.

- **Duplicates and Data Inconsistencies:**
 - Detected 14 duplicates (28 total entries) in the raw data.
 - Identified 17 widows in the baseline data that were not found in the raw data.
 - Noticed 137 mismatched age entries between the raw and baseline data for specific widows.
- **Null Values:** Several key columns, such as income, contained null values. However, it is important to note that all widows with null values in the income column reported having some form of income generation. This suggests that the null values may not represent a lack of income but rather incomplete or inconsistent reporting in the dataset.
- **Data Formatting Issues:**
 - Many columns are formatted as text instead of numerical or categorical variables. This will need to be addressed for correlation analysis.
 - Binary responses (e.g., "Yes"/"No" or "1"/"0") are prevalent, which may require conversion to proper categorical variables.
- **Suicidal Thoughts:** A high number of 579 respondents (out of 624) answered "0" to the question about suicidal thoughts, which may make it difficult to draw significant correlations in this area.
- **Limited Severe Mental Health Data:** The low count of severe anxiety and depression cases (especially severe depression) could limit the statistical power of any analysis focused on those conditions when testing the entire dataset. However, by filtering for specific groups or conditions, we may be able to gain more meaningful insights into the factors contributing to severe mental health outcomes.

Data Wrangling

Standardization of Text Data:

- Text functions such as **TRIM**, **UPPER**, and **PROPER** were applied to standardize categorical variables, ensuring consistency across the dataset (e.g., converting all text data to consistent case formats for easier analysis).

Binary Column Standardization:

- **IF** functions were used to standardize binary columns like "Yes"/"No" responses, converting them into numerical values (**1 for "Yes" and 0 for "No"**) to facilitate correlation analysis.
- **Missing values** in binary columns were replaced with **0 (indicating "No")** to ensure completeness and readiness for analysis.

Determination of Anxiety and Depression Levels:

- **Anxiety levels** were determined based on the **GAD-7 score (range: 0-21)**, using the following classification:
 - **Minimal Anxiety:** Scores from 0 to 4
 - **Mild Anxiety:** Scores from 5 to 9
 - **Moderate Anxiety:** Scores from 10 to 14
 - **Severe Anxiety:** Scores from 15 to 21

- **Depression levels** were determined based on the **PHQ-9 score (range: 0-27)**, using the following classification:
 - **Minimal Depression:** Scores from 0 to 4
 - **Mild Depression:** Scores from 5 to 9
 - **Moderate Depression:** Scores from 10 to 14
 - **Moderately Severe Depression:** Scores from 15 to 19
 - **Severe Depression:** Scores from 20 to 27

Handling Missing Values:

- **Missing income** values were replaced with the **average income** from the dataset to allow for correlation analysis without introducing bias due to missing data.

Duplicate Removal:

- **14 duplicate rows** (28 entries) were identified and removed. **Conditional formatting** was used to highlight duplicates based on **Widow ID**, ensuring no duplication of unique data.

Age Consistency Checks:

- **Age mismatches** between the **Raw Data** and **Baseline Data** were identified using the **VLOOKUP**, **INDEX/MATCH**, and **IF** functions. Any records where the **age** in the **Raw Data** did not match the age in the **Baseline Data** were **flagged and removed**.
- **137 rows** were deleted due to discrepancies in age between the two datasets. These rows were removed to ensure that the dataset maintained accurate and consistent demographic information.

Merging Datasets:

- **VLOOKUP** and **INDEX/MATCH** functions were used to **merge the Raw Data and Baseline Data** based on **Widow ID**, aligning important columns such as income, dependents, and other baseline factors into a single dataset.

Correcting Data Types:

- **Data types** for columns were corrected to ensure consistency and compatibility with analytical processes. Numeric data stored as text was converted to the appropriate **numeric format**.

Filtering Data for Validity:

- Only records with matching Widow ID and age from both the Raw and Baseline datasets were retained.

Final Cleanup:

- Unnecessary columns were removed to streamline the dataset for analysis.
- Columns necessary for further analysis were preserved and correctly formatted.

Final Dataset Overview:

- **After merging, cleaning, and filtering**, the final dataset included **457 unique records** and **72 columns** with all relevant information retained for analysis.
- The dataset was prepared for in-depth analysis of mental health outcomes (anxiety and depression) in relation to baseline factors such as income, dependents, and food security.