SOCIAL CONNECTION x PREVALENCE OF DEPRESSIVE & ANXIETY DISORDERS

Social connection (SC) is a measure of how connected a country's residents are, based on SC scores of Facebook users. Here, I explore the relationship between SC and the prevalence of depressive and anxiety disorders, using correlation coefficient heatmaps, scatter plots & choro maps. Lower depression, and anxiety disorders, in more connected countries.

TABITHA STUART | GITHUB | TABITH-STUART

This was done during the Tableau Foundations challenge and the data developed throughout the course. Thanks to Andy Kriebel & Jeff Drisgill for providing the challenge.

DATASET LIMITATIONS:

Before delving into the analysis, it is important to acknowledge the inherent limitations of the data, including:

- **Reporting prevalence of Depression and Anxiety Disorders is derived using model-estimated values**
  - The IHME GBD 2019 study utilized the broadest spectrum of epidemiological experience to generate prevalence and incidence rates for all covered disorders. The model is significant as it directly affects the comprehensiveness of the data, and the accuracy of the estimates. The model's ability to account for "hard-to-reach" populations is particularly significant, given the nature of the topic.
- **Subjective nature of diagnosing Depression and Anxiety Disorders**
  - Depression and anxiety disorders are subjective in nature, and are diagnosed through questionnaires or interviews, which are inherently biased, flawed and not standardized.

Datasets are compiled from various sources, and originally produced for different purposes

- The differing intentions and purposes across the datasets makes it challenging to combine them, and draw conclusions. Some datasets are more suited to the analysis than others, but this is not accounted for in the standardization and filtering.

DATASETS:

The project consisted of a combination of the following datasets:

- Prevalence of Depression & Anxiety Disorders (.csv)
- Continents (.csv)
- Facebook SC Scores (.csv), 250 per country (Original SC) (2020) and 1 per country (Updated SC) (2021)
- World Cities Database (.csv)
- World Bank Country Boundaries (.shp)
- Continent Polygons (.shp)
- Country Polygons (.shp)
- Country Centroids (.shp)

Facebook Social Connectedness and SC variables were developed carefully to understand the potential and limitations of each.

Click datasets for examples of the compiled datasets and their sources, and the calculated fields created to prepare the data.

DATA PREPARATION:

The initial cleaning of the datasets was performed in Excel, while exploratory/descriptive analysis was performed in Tableau. Tableau was used for the majority of the data preparation, because calculations and joins were quick and efficient with datasets of this size.

- As the datasets were sourced from different organizations, I faced a challenge of them having certain inaccuracies, and inconsistencies in how countries were listed, including groupings, and geo-spatial. To address this, a standardization of countries, regions and continents was performed, ensuring all datasets could be joined accurately.
- Columns not pertinent to the analysis were removed, and calculated fields were created to standardize and group variables, such as the Depression and Anxiety Disorders moving data for some countries. For a more complete and robust analysis, additional datasets could be joined, such as World Bank Country Income Levels.
- Some datasets included aggregate values, such as World, by income or continent, which were removed since all datasets
- Lastly, the data was reshaped, disaggregated, cleansed/corrected, and aggregated.
  - Pivot/unpivot using ASCII Tools
  - Unmerged/split or created columns, and overlapping/combined across datasets, and data joins, using Alteryx, Tableau, and Excel. Joins were performed using Alteryx, Tableau, or Excel depending on the dataset, which contributed part of the decision. Joins were performed using Alteryx, Tableau, or Excel depending on the dataset, which contributed part of the decision. Tableau was used for "center-based" intersect joins, and Alteryx was used for "overlap" joins, which contributed to each decision.

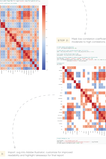
The preprocessing to increase data consistency and accuracy ensured subsequent steps changing the granularity of the data, and joining datasets, could be performed with less inaccuracies, and inconsistencies.

Once imported into Tableau:

- Geo-spatial data was created and standardized
  - Polygon, centroid and point files were created for geo-spatial analysis
  - Categorical data (such as "GDP groupings) was calculated and utilized from Japan to standardize groupings, and ensure accurate and consistent comparisons
  - Country names were altered to ensure mapping

DATA EXPLORATION:

Before diving into the main line, I made use of data exploration. Click on images for a detailed view and explanation.



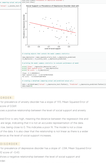
The exploratory data analysis was initiated by selecting a subset of variables from the compiled datasets, and creating a correlation matrix, to identify any potential relationships. Correlation coefficients, and p-values, were calculated using Alteryx, and visualized using a combination of choro/scatter plots. The insights gained from this stage (scatter) informed decisions on which variables to include in further analysis, and which to exclude from further analysis.

Correlation between each set of variables were assessed by creating a correlation coefficient heatmap using Alteryx.

SCAT 1: Depression & Anxiety Disorders, correlation by continent

SCAT 2: Depression & Anxiety Disorders, correlation by income

SCAT 3: Depression & Anxiety Disorders, correlation by continent, excluding Oceania



SCAT 4: Depression & Anxiety Disorders, correlation by income, excluding Oceania

Choro/Scatter plots were created to visualize the correlations between variables, and identify any potential outliers, or other patterns.

Prevalence ratios were performed to rank each Depression:

- The higher the prevalence ratio, the higher the proportion of anxiety disorder
- The higher the ratio, the higher the proportion of depression disorder

SCATTERPLOTS

Key KPIs were also examined and interpreted, based on the prevalence of each disorder, Social Connectedness, and the different ratios, using the Facebook's Updated Original Scores.

CHOROPLETHS

Choro/Scatter plots were created to visualize the correlations between variables, and identify any potential outliers, or other patterns.

PREVALENCE RATIOS

Prevalence ratios were performed to rank each Depression:

- The higher the prevalence ratio, the higher the proportion of anxiety disorder
- The higher the ratio, the higher the proportion of depression disorder

ANXIETY DISORDERS:

- Regression line (in percentage of anxiety disorder) has a slope of 170.78(see Calculation 2 in the Calculations section)
- Slope of 170.78 shows positive relationship between the level of Social Connectedness, and Anxiety Disorder
- Positive Correlation Coefficient is also high, meaning the relationship between the regression line and the data points is high, indicating that it is a more accurate representation of the data
- R-square value of 0.29, meaning that 29.0 % of the variance in the Anxiety Disorder data can be explained by the Social Connectedness. This is not a very strong correlation, but it is still a fairly high value, given the nature of the data, and the number of variables not included in the analysis

DEPRESSION DISORDERS:

- Regression line (in percentage of depression disorder) has a slope of -1.284 (see Calculation 2 in the Calculations section)
- Slope of -1.284 shows a negative relationship between the level of social support and Depression Disorder
- The Pearson Correlation Coefficient is also high, meaning the relationship between the regression line and the data points is high, indicating that it is an accurate representation of the data
- R-square value of 0.49, meaning that 49.0 % of the variance in the Depression Disorder data can be explained by the Social Connectedness. This is a fairly strong correlation, and is the most significant of the three

RATIO TOGETHER:

- Regression line (in percentage of the ratio) has a slope of 0.0000 (see Calculation 2 in the Calculations section)
- As a slope of approximately zero means that there is no relationship between the variables, the ratio of anxiety disorder to depression disorder is not impacted by the level of Social Connectedness, and is instead impacted by other factors
- Correlation Coefficient is fairly strong, 0.71, but R-square value of 0.50 indicates that only 50.0 % of the variance in the ratio can be explained by the Social Connectedness. This is the most significant of the three, and is the most representative of the data

PRESENTATION HIGHLIGHTS:

Correlations between key variables

Social support & Testing rates comparison

Continent Polygon Map

CHALLENGES:

Some of the challenges I faced were:

- Data limitations
  - Subjective nature of depression and anxiety
  - The inherent biased nature, coupled with overall underrepresented data, due to the nature of the topic, and the fact that it is more prevalent in some cultures than others, creates some issues.
  - In the case of the U Mapping Questionnaire, the small sample size significantly hampered the analysis, and I ended up not including it in the analysis.
- Technical challenges
  - Accounting for datasets, which contained a combination of variables and aggregate values, and ensuring that the calculations, and aggregations, were performed at the correct granularity, and level of detail, contributed to the need for complex workarounds, or more complex calculations, that I had to research extensively, in order to produce the outcomes I wanted.
  - Tableau Prep was a valuable tool, but it was not capable of handling the preparation required, so I had to use other tools, such as Alteryx, and Excel, to prepare the datasets.
  - Tableau's Spatial functions are very powerful, but they are not as robust as other tools, such as Alteryx, and Arcgis, when it comes to spatial joins, and other functions.

CONCLUSION:

Social connection is a measure of how connected a country's residents are, based on SC scores of Facebook users. Lower depression, and anxiety disorders, in more connected countries.

Key KPIs were also examined and interpreted, based on the prevalence of each disorder, Social Connectedness, and the different ratios, using the Facebook's Updated Original Scores.

Correlation coefficients, and p-values, were calculated using Alteryx, and visualized using a combination of choro/scatter plots. The insights gained from this stage informed decisions on which variables to include in further analysis, and which to exclude from further analysis.

Key KPIs were also examined and interpreted, based on the prevalence of each disorder, Social Connectedness, and the different ratios, using the Facebook's Updated Original Scores.

Choro/Scatter plots were created to visualize the correlations between variables, and identify any potential outliers, or other patterns.

Prevalence ratios were performed to rank each Depression:

- The higher the prevalence ratio, the higher the proportion of anxiety disorder
- The higher the ratio, the higher the proportion of depression disorder

SKILLS ACQUIRED:

- Data Preparation
  - Pivots/ Unpivots
  - Pivots/ Unpivots, Unmerge/Split
  - Data Cleansing, Aggregation, Data Visualization
  - Prep/ Alteryx geographic calculations
  - Data Exploration, Cleansing
- Tableau
  - Joins
  - Correlation coefficient heatmaps
  - geo-spatial analysis
  - Linear Regression analysis
  - Choro/Scatter plots
  - Data analysis, analysis and interpretation