Data Analysis utilizing Multiple Regression and Visualization of the Correlation(s) between National Values and Happiness

By: Jeanne Spaulding

1. Project Overview:

This project will look at the potential correlation between national values as assessed by the World Values Survey (WVS) and happiness according to the World Happiness Report (WHR) for the year 2018. The WVS 2018 raw dataset of individual responses will be condensed into underlying domains using factor analysis to achieve a value for each domain per country. The WHR 2018 data is grouped by country into values for 6 domains that comprise the report. The two datasets will be integrated based on countries participating in both surveys, WVS 2018 and WHR 2018. Multiple linear regression will be performed to determine strength of correlation among national value domains and happiness scores.

1. Technical Insights:
   1. Data Sources:
      1. World Values Survey (WVS) wave 7 (2017-2022): This dataset provides insights into societal beliefs, values, and cultural norms of people across the world. The survey is comprised of approximately 300 questions spanning topics of family, politics, education, economics, nationalism, health, and safety. Data are freely available in CSV format downloads via WVS website ([www.worldvaluessurvey.org](http://www.worldvaluessurvey.org)).
      2. World Happiness Report (WHR) 2020 (2005-2019): This dataset provides insights into happiness utilizing life evaluations from the Gallup World Poll. The WHR has indices covering social supports, freedom, government, and perceptions. Data are available for download in CSV format via WHR website (<https://worldhappiness.report>).
   2. Tools and Technologies:
      1. SQL (SQLite): SQL will be used for storing of the cleaned , processed datasets and merging. Tasks will include:
         1. Create tables for data storage
         2. ~~Querying data on year and country~~
         3. Merging datasets
         4. Exporting single analyzable dataset for Python analysis
      2. Python: will be used for data analysis and visualization of data. Libraries will include:
         1. Pandas: for data manipulation and working with dataframes
         2. Numpy: for numerical computations
         3. ~~SciPy: for numerical analysis using scientific and technical computing; expanding Numpy capabilities~~
         4. Scikit-Learn: for factor analysis and regression; builds on Numpy, SciPy, and Matplotlib
         5. Matplotlib and/or Seaborn: for creating visuals with plots, charts, regression models
         6. Statsmodel: for generating the regression model statistical summary table
      3. GitHub: will be used for repository storage, version control and sharing
         1. <https://github.com/spauldingj2/NationalValues_Happiness>
   3. Data Integration:
      1. Python pre-processing:
         1. WVS
            1. Load dataset into Python
            2. Filter dataframe for year 2018 and relevant variables for analysis
            3. Explore and drop missing values
            4. Import cleaned dataset into SQL table for data storage and merging
         2. WHR
            1. Load dataset into Python
            2. Filter dataframe for year 2018 and relevant variables
            3. Explore and drop missing values
            4. Import cleaned dataset into SQL table for data storage and merging
      2. SQL merging of datasets:
         1. Filtered WVS dataset uploaded to SQL table
         2. Primary Key identified for join, Country
         3. Perform Left Join to merge filtered WHR and filtered WVS
         4. Export CSV file of merged dataset for further Python analysis
      3. Python Regression Analysis:
         1. Define *y* – dependent variable (Happiness Score)
         2. Define *x* – independent variables (WVS factors)
         3. Split *x* and *y* into test and training sets; fit the regression model
         4. Evaluate the model
   4. Data Visualization
      1. Multiple Regression Analysis
         1. Scatter plot with regression line – show the relationship between *x* (independent variables) and *y* (dependent variable)

References:

World Values Survey website: [www.worldvaluessurvey.org](http://www.worldvaluessurvey.org)

World Happiness Report website: <https://worldhappiness.report>

Tabachnick, B and Fidell, L. (2007). *Using Multivariate Statistics*, Fifth Edition.

Shields, W. (2020). *SQL Quick Start Guide*.

Navlani, A., Fandango, A., and Idris, I. (2021). *Python Data Analysis*, Third Edition.

Cornell University. (2024). Python for Data Science e-course. Class Notes

Landup, D. (2021). *Data Visualization in Python*, First Edition.