Codebook

C. Andre

2024-12-07

Introduction

This document serves as a codebook for the analysis of global smoking prevalence data. It describes the datasets and key variables and libraries used.

Datasets

1. Raw Dataset

- **Dataset Name**: `rawdata`
- **Description**: This is the original dataset containing global smoking prevalence data from multiple sources.
- **Variables**:
  - `Entity`: Country or region name.
  - `Code`: ISO 3-letter country code.
  - `Year`: Year of the observation.
  - `Prevalence.of.current.tobacco.use....of.adults.`: Percentage of adults who currently use tobacco products.

---

2. Cleaned Dataset

- **Dataset Name**: `countries_data`
- **Description**: Cleaned dataset after removing regions and income categories, and focusing on relevant countries and years.
- **Key Changes**:
  - Removed entries for regions such as "World", "Sub-Saharan Africa", etc.
  - Filtered out years 2018 and 2019.
  - Renamed `Prevalence.of.current.tobacco.use....of.adults.` to `Prevalence`.

---

3. Merged Dataset

- **Dataset Name**: `map_data`
- **Description**: Merged dataset combining the world map data (`world`) with the cleaned smoking prevalence data (`countries_data`).
- **Key Variables**:
  - `iso_a3`: 3-letter ISO country codes.
  - `Prevalence`: Percentage of adults who smoke.
  - `geometry`: Spatial data for visualizing countries on a map.

---

4. Filtered Subsets

- **Dataset Name**: `subset_data` and `subset_data2`
- **Description**: These datasets are subsets of `plot_data` filtered by specific years.
-
  - `subset_data`: Includes data for years 2000 and 2005.
  - `subset_data2`: Includes data for years 2010, 2015, and 2020.
- **Purpose**: These subsets are combined into `combined_data` for visualization. My first interactive visualization worked for seperate years but not combined. Seperating the data is the only way I managed to make my interactive visualization work with all the data.

---

5. Combined Dataset

- **Dataset Name**: `combined_data`
- **Description**: Combines data from `subset_data` and `subset_data2` for visualizing smoking prevalence trends across multiple years (2000, 2005, 2010, 2015, and 2020).
- **Key Variables**:
  - `iso_a3`: 3-letter ISO country codes.
  - `Entity`: Country name.
  - `Prevalence`: Percentage of adults who smoke.
  - `Year`: Year of observation.

Variables

Below are key variables used across datasets:

1. ISO Codes

- **Name**: `iso_a3`
- **Description**: 3-letter ISO code representing each country.
- **Example Values**: `USA`, `FRA`, `IND`.

2. Prevalence

- **Name**: `Prevalence`
- **Description**: Percentage of adults who smoke in each country.
- **Example Values**: `0`, `15.6`, `68.5`.

3. Year

- **Name**: `Year`
- **Description**: Year of observation.
- **Example Values**: `2000`, `2005`, `2010`, `2015`, `2020`.

Libraries Used

The following R libraries were used in this project:

- **`tidyverse`**: Data manipulation and visualization tools (e.g., `dplyr` for filtering and summarizing, `tidyr` for reshaping data, and `ggplot2` for creating static visualizations).
- **`plotly`**: For interactive visualizations like choropleth maps.
- **`rnaturalearth` and `rnaturalearthdata`**: Provides geospatial data for world maps.
- **`sf` (Simple Features)**: Handles and manipulates spatial data.
- **`htmlwidgets`**: Exports interactive plots (e.g., `initial_interactive_plot.html`).