Data Appendix

The purpose of this appendix is to provide a detailed description of the datasets used in the Exploratory Data Analysis (EDA) and our Time-Series Analysis. This document details information regarding the variables of the original and cleaned datasets, descriptive statistics, and general trends and observations of all datasets used in this project.

Appendix A. Original Dataset (original_dat.csv)

Our dataset was sourced from the United Nations Global Sustainable Development Goals (SDG) Indicator Database. Within the database, this dataset is referred to as “Indicator 3.5.2: Alcohol consumption per capita (aged 15 years and older) within a calendar year (litres of pure alcohol)” and is saved as “original_dat.csv” within the DATA folder on this github repository. Each unit of observation represents the average alcohol intake within a calendar year for females, males or both sexes at specific time points for a specific country and geographic region. A key aspect of this dataset is that alcohol consumption data is recorded at approximately five-year intervals, including the following years: 2000, 2005, 2010, 2015, and 2019.

Descriptive Statistics:

- 624 observations from 188 countries total
- Timepoints: 2000, 2005, 2010, 2015, 2019
- Regions with the most alcohol consumption data:
    - Asia– 165 Total Observations
    - Africa– 156 Total Observations
    - Europe– 120 Total Observations

The following table showcases the number of observations for each geographical region within this dataset. For our time-series analysis, we analyzed average alcohol consumption for Africa, Asia, and Europe because regions from these continents had the most observations overall, as indicated below.

```
> region_count

             Australia and New Zealand                              Caribbean
                                     6                                     39
                       Central America                           Central Asia
                                    24                                     15
                        Eastern Africa                           Eastern Asia
                                    51                                     15
                        Eastern Europe                          Middle Africa
                                    30                                     27
                       Northern Africa                       Northern America
                                    15                                      6
                       Northern Europe Oceania (exc. Australia and New Zealand)
                                    30                                     72
                         South America                     South-Eastern Asia
                                    36                                     33
                       Southern Africa                          Southern Asia
                                    15                                      3
        Southern Asia (excluding India)                        Southern Europe
                                    48                                     39
                        Western Africa                           Western Asia
                                    48                                     51
                        Western Europe
                                    21
```

Table 1: Number of Observations Per Region

Cleaned Data Dictionary:
The following data dictionary contains information about variables deemed relevant for our analysis and providing proper contextualization. Within our EDA.R file under the SCRIPTS folder on this github repository, we provide the code to remove irrelevant variables.

Column Name	Description
Value 2000-2019	**Yearly alcohol consumption values for each region.**
Latest Value	**Most recent recorded alcohol consumption value.**
Geographic Area Code	**Numeric identifier for geographic regions.**
Geographic Area Name	**Name of the geographic region.**
Geographic Area Level	**Level of geographic aggregation (e.g., country, region).**
Parent Geographic Area Code	**Numeric code of the parent geographic region.**
Parent Geographic Area Name	**Name of the parent geographic region.**
ISO Code	**ISO country code for international standardization.**
X	**Longitude coordinate for geographic mapping.**
Y	**Latitude coordinate for geographic mapping.**
x2	**Additional longitude coordinate (if available).**
y2	**Additional latitude coordinate (if available).**
Number of Years Available	**Count of available years of data for each region.**
Earliest Year Available	**First year for which alcohol consumption data is available.**
Latest Year Available	**Most recent year with available alcohol consumption data.**
Available Years	**List of years with available alcohol consumption data.**
Goal Code	**Numeric identifier for the sustainable development goal (SDG).**
Target Code	**Numeric code representing the target under the SDG.**
Indicator Code	**Code identifying the alcohol consumption indicator.**
Indicator Description	**Explanation of what the alcohol consumption indicator measures.**
Series Code	**Unique identifier for the dataset series.**
Series Description	**Description of the dataset series.**

Table 2: Data Dictionary

Appendix B. Final Dataset (long_dat.csv)

This data file contains the dataset used for our time-series analysis using the ARIMA model, which is located in the DATA folder under "long_dat.csv" within this github repository. Each observation in this dataset represents the average alcohol consumption (L) within a calendar year for people within a specific continental region (e.g., Europe, Africa, Asia) for a specific year (e.g., 2000, 2005, 2010, 2015, 2019). This is a long version of the final dataset, in which all years are repeated for each continent and alcohol intake is an actual variable. There are 15 total entries that have corresponding information across three variables that are described below.

Final Dataset Variables:

Year: character variable

- Indicates the year that the average alcohol consumption data was recorded in the database.
    - Values: E.g., 2000

Continent: character variable

- Indicates the continental region that the average alcohol consumption data was taken from.
    - Values: E.g., Asia

Alc: numeric variable

- Indicates the average alcohol intake per capita (litres of pure alcohol) for people aged 15 years or older within a calendar year in a specific continental region.
    - Values: E.g., 2.449354

Visualizations:

The following visualizations display a line plot in Figure 1 and a bar plot in Figure 2 showcasing the same information regarding alcohol consumption trends by continental region from 2000-2019 over five-year intervals. From these visualizations, we can see that each continental region has a distinct average alcohol consumption that does not vary much over time. However, we can see that countries in Europe slightly declined in alcohol consumption when comparing 2000 to 2019, regions within Asia slightly increased within that time period, and African countries stayed relatively the same over time. These visualizations provide interesting insights into what we can expect our ARIMA models to predict. We hypothesize that alcohol consumption per capita will increase in Europe over time, but will stagnate or decrease in Africa and Asia.

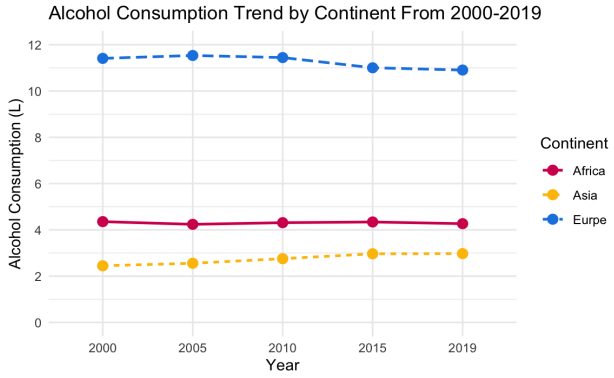

Figure 1: Line Plot of Alc. Intake Over Time Per Region

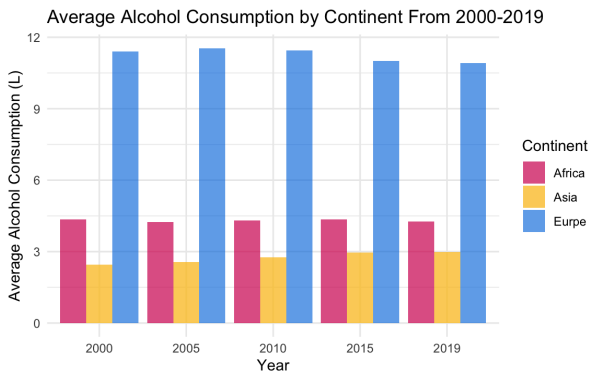

Figure 2: Line Plot of Alc. Intake Over Time Per Region