Homework 2 - Data Scraping, APIs and Advanced Regression

Due Date

Friday February 15, 2026 by 11:59pm.

Learning Goals

- Use an API, scrape, wrangle, and get familiar different datasets.
- Make visualizations
- Conduct GAM regression modeling and compare to linear regression

Assignment Description

For this assignment, we will be analyzing data on life expectancy and alcohol consumption. The learning objectives are to conduct data wrangling and visualization keeping key questions in mind. We will also do a regression analysis.

The primary questions of interest are:

- Is there an association between life expectancy and alcohol consumption?
- How have life expectancy and alcohol consumption changed over time?
- Do other country specific factors such as per capita gross domestic product or population play a role in the relationship between alcohol consumption and life expectancy?

Data Acquisition and Wrangling

1. (15 points) Obtain two datasets: life expectancy (via API) and alcohol consumption (via web scraping).

For life expectancy, population, and GDP: Use the World Bank API to fetch the following indicators:

- `SP.DYN.LE00.IN` - Life expectancy at birth, total (years)
- `SP.POP.TOTL` - Population, total
- `NY.GDP.PCAP.CD` - GDP per capita (current US$)

You will need to:

- Parse the nested JSON response into a clean DataFrame for each indicator
- Handle missing values and filter to relevant years (1996, 2016, 2019)
- Note: The API returns data for regions/aggregates as well as countries; you may want to filter these out
- Merge the three indicators together by country and year

See the World Bank API documentation for more details on available endpoints and filtering options.

For alcohol consumption: Use web scraping to extract the **historical consumption table (with data for 1996, 2016, and 2019)** from the Wikipedia page on alcohol consumption per capita.

Before merging, prepare both datasets as follows:

a. **World Bank API data:** Parse the JSON responses from the World Bank API into DataFrames. The API response contains nested dictionaries with fields like `country`, `date`, and `value`. Extract the relevant fields into clean DataFrames for each indicator (life expectancy, population, and GDP per capita), then merge them together by country and year.
b. **Alcohol (scraped data):** Identify which Wikipedia table contains the historical data (1996, 2016, 2019) and extract it. Clean the scraped data: remove any footnote markers (e.g., `[1]`), convert columns to appropriate data types, and handle any missing values.
c. Reshape the alcohol data from wide to long format, creating a "Year" column with values 1996, 2016, and 2019. You can use `pd.melt()` in Python.
d. Filter the life expectancy data to include only the years that match the alcohol data (1996, 2016, 2019).

Merge these datasets by country name and year. Note: You may need to standardize country names between datasets (e.g., "United States" vs "United States of America").

Conduct basic EDA on the merged dataset: check dimensions, missing values, summary statistics of key variables to look for outliers. Which countries have the lowest and highest life expectancies and alcohol consumption?

Write a paragraph describing your scraping and cleaning steps. Include a written summary of the variables and the number of observations before and after merging. Document any countries that didn't match and how you handled them. Also summarize your findings from the basic EDA.

2. (4 points) Create a summary table of the merged dataset showing the mean and sd of life expectancy, alcohol consumption, population, and GDP per capita by year. Briefly summarize.

3. (4 points) Create a new categorical variable named "gdp_level" using the GDP per capita variable. Calculate the quartiles of GDP per capita. Categorize GDP level as low (0-q1), medium (q1-q3), and high (q3+). To make sure the variable is correctly coded, create a summary table that contains the minimum GDP per capita, maximum GDP per capita, and number of observations for each category.

Visualization

4. (15 points) Create the following figures with `plotnine` and interpret them. Be sure to include easily understandable axes, titles, and legends.

a. Two line plots: one of life expectancy by year and the other of alcohol consumption by year with lines for Canada and 3 other contrasting countries.
b. Facet plot by year for 1996, 2016, and 2019 showing scatterplots with linear and lowess regression lines of life expectancy and alcohol consumption.
c. Facet plot by year for 1996, 2016, and 2019 showing boxplots of alcohol consumption by GDP level.
d. A barplot showing life expectancy in 1996 and 2019 for the 10 countries with the largest change in life expectancy between 1996 and 2019.
e. Come up with your own plot that will help to understand the role of GDP in the association between alcohol and life expectancy.

Advanced Regression

5. (8 points) Construct a multiple linear regression model to examine the association between life expectancy and alcohol consumption adjusted for GDP per capita, and year. Note you should scale GDP by population to get GDP per capita since the values can be very large. Then fit a gam model where you put a smooth on GDP per capita and alcohol consumption.

Provide the following in your analyses:

a. Summaries of your models, including overall model fit and interpretation of the parameter estimates (linear and non-linear).
b. Plot of the smoothed variables from the gam model and interpret.