PSTAT 134 - Data Wrangling, Web Scraping & APIs

Assignment 1

PSTAT 134 (Spring 2025)
Due Date: May 4th, 11:59 PM

Part I - Case Study: New York Times Ad Impressions (Simulated)

There are 10 data sets in the `/data` sub directory named `nyt1.csv`, `nyt2.csv`, ..., `nyt10.csv`. Each file represents one day's worth of simulated data on ad impressions and clicks on the New York Times homepage. Each row represents a single user. There are five columns:

- `Age` (user's age)
- `Gender` (user's gender, coded as `0` = female, `1` = male)
- `Impressions` (number of ads displayed during the user's visit)
- `Clicks` (number of clicks made by the user)
- `Signed_In` (whether or not the user was signed in as a member)

1. Load **a single data file**. Write a single bash command that lists (or otherwise "loads") all ten CSV files: nyt1.csv through nyt10.csv; from the /data directory by matching their file names with a regular expression.

Hint: you might combine `ls /data` with `grep -E '^....$'` so that only `nyt1.csv`, `nyt2.csv`, ..., `nyt10.csv` are selected.

2. Create a new variable, `age_group`, that categorizes users into the following age groups: $<$ 18, 18-24, 25-34, 35-44, 45-54, 55-64, and 65+.

3. Plot the distributions of impressions and "click-through rate" for all 6 age categories. (*Note:* Click-through rate is defined as the number of clicks divided by the number of impressions; it represents the proportion of ads that generated clicks.)
4. Create a new variable to categorize users based on their click behavior. (The name and categories for this variable are up to you. Explain what decision[s] you make and why.)
5. Explore the data and make visual and quantitative comparisons across user segments/demographics to answer the following:
    - How do <18 year old males differ from <18 year old females in terms of clicks and impressions?
    - How does the distribution of click-through rate for users who are signed in differ from the distribution for those who are **not** signed in?
    - Are certain age groups more likely to be signed in than others? Which ones?
6. Calculate summary statistics for the click-through rate. These should include (1) quartiles, (2) mean, (3) median, (4) minimum and maximum, and (5) variance. Choose two user segments to compare these statistics across (for example, compare the mean, median, and quartiles for users 25-34 to those for users 65+).
7. Summarize your findings in a brief (1-2 paragraph) report intended for a New York Times (NYT) advertising team.

Part II: Analyzing the Weather

In this section, you will gain more practice working with public APIs, this time using a public weather API, WeatherAPI. The first thing you'll need to access the API is an API key. You can sign up for a key here: https://www.weatherapi.com/signup.aspx

Exercise 1

Use the http://api.weatherapi.com/v1/current.json URL to access the API and obtain real-time weather data. Note that you will want to specify three query parameters, at least – `key`, which should be set to your individual API key, `q`, which should equal the city name of a specified location – for example `q = "Isla Vista"` – and `aqi`, which indicates whether you want to obtain air quality data (`"yes"` or `"no"`).

Obtain current real-time weather data for **fifty cities**. I have saved a data file containing the names of fifty cities to `/data/cities.csv`. This ensures that you are all working with the same locations (although your results will still differ, depending on when you obtain the data).

Exercise 2

Write code in R or Python (your choice) to extract and store the following data for each location:

- City name
- Country
- Whether or not it is currently daytime there
- Temperature (in Fahrenheit)
- Humidity
- Weather description (`condition` text; for example, "Mist", "Clear", etc.)
- Wind speed (in miles per hour)
- Precipitation (in millimeters)
- US EPA air quality index (ranges from 1 to 6, representing the 6 categories of air quality: https://www.airnow.gov/aqi/aqi-basics/)

Exercise 3

Create a scatterplot of temperature vs. humidity. Add a linear regression line to the plot. What are the estimated intercept and slope values for this linear regression? Does there appear to be a significant relationship between temperature and humidity?

Exercise 4

Create a bar chart of the EPA air quality index values. What does the distribution of air quality look like? Identify the location(s) with the best air quality and the worst air quality.

Exercise 5

Create a bar chart of the current weather description. Which conditions are the most common? Which are the least?

Part III: Scraping Books

In this section, you'll practice your web scraping skills by experimenting with a fictional online bookstore located at https://books.toscrape.com/. Use the tools that we demonstrate in class to do the following, in either R or Python (your choice):

Exercise 6

Scrape the first 20 results from this site. Create a data frame (or tibble) that stores the following for each book:

- Title
- Price (excluding tax)
- Star rating
- Whether the book is in stock

Exercise 7

Create a histogram of prices for these 20 books. What is the average price?

Exercise 8

Create a bar chart of star rating for these 20 books. Find the book(s) with the highest and lowest star ratings.