**Module 1: Exploratory analysis of air pollution and health data**

In this module we will explore the daily time time series data of air pollution and mortality and we will visually inspect their long-term, seasonal, and short-term variation. We will also calculate the associations between air pollution and mortality at these different scales of variation.

Data: The data frames for each of the cities have all the same variable names. The primary variables we will need from each data frame are

- `death`, daily mortality counts from all-cause non-accidental mortality
- `pm10`, daily $PM_{10}$ values
- `date`, the date
- `tmpd`, daily temperature

Questions to consider (Part 1)

1. What are the main characteristics of the time series data for mortality and air pollution?
2. What are the long-term, seasonal, and short-term variations in air pollution in U.S. cities? Are there differences between cities?
3. What are the long-term, seasonal and short-term variations in mortality in U.S. cities? Are there differences between cities? Are there differences between age categories?

Questions to consider (Part 2)

1. How do the long-term, seasonal, and short-term variations in $PM_{10}$ and mortality relate to each other? How do they relate on different timescales?
2. Is there any evidence of an association between $PM_{10}$ and mortality in these cities? Which timescale is more suitable for drawing inferences?

Overarching questions:

1. Based on the correlations between long-term trends, seasonal, and short-term variations, what is the evidence about the association between $PM_{10}$ and mortality?
2. How should we weigh the evidence from the different timescales? What evidence is more important? What evidence should be discounted and why?

**Part 1: Descriptive analyses**

Pollution ($PM_{10}$) data

1. Load data for Chicago (`chic`), New York (`ny`) and Los Angeles (`la`) into R.
2. For each city, plot the $PM_{10}$ data versus date. Try plotting $PM_{10}$ both with and without the trend added in. Try plotting the data in smaller windows of time to see more detail.
3. Using the `tsdecomp()` function, decompose the Chicago $PM_{10}$ data into 3 timescales: long-term variation, *seasonal* variation *short-term* variation. This is the default setting for `tsdecomp`. Plot your results.

Mortality data

1. Load mortality data for Chicago (`chic.mortality`), New York (`ny.mortality`) and Los Angeles (`la.mortality`) into R.
2. For each of the 3 cities, plot the all-cause non-accidental mortality data versus date separately for each of the 3 age categories: `under65`, `65to74`, and `75p`. Use the age category specific mortality (`*.mortality.rda`) datasets for this.
3. Using the `tsdecomp()` function, decompose the mortality data into 3 timescales (as before). Use the standard datasets for this part.

**Part 2: Looking at correlations**

Bringing the mortality and pollution data together

NOTE: Use the standard datasets for this part.

1. Revisit the timescale decompositions for both $PM_{10}$ and mortality in Chicago. Visually compare the long-term trend for mortality with the long-term trend for $PM_{10}$. Do the same comparison for the seasonal and short-term components.
2. Compute the correlation coefficient between the long-term trends for mortality and $PM_{10}$. Compute the correlation coefficients for both the seasonal and short-term components of mortality and $PM_{10}$. If there are any missing data, set `use = "complete"` in the call to `cor` when computing the correlation.
3. Try the same timescale/correlation analysis with the city of Seattle, WA (`seat`). Do you get the same correlations?
4. Try the same timescale/correlation analysis Pittsburgh, PA (`pitt`).
5. Fill in the following table with the correlations between $PM_{10}$ and mortality computed at different timescales in the previous steps:

| | Long-term | Seasonal | Short-term |
|---|---|---|---|
| Chicago | | | |
| Seattle | | | |
| Pittsburgh | | | |

Unfortunately, the timescale analysis using `tsdecomp()` can only be done with cities that have relatively complete data on $PM_{10}$. Later we will use other methods to get around this limitation.