GEA1000 QUANTITATIVE REASONING WITH DATA
Tutorial 3

Please work on the problems before coming to class. In class, you will engage in group work.

Question 1: Exploring Relationships in Chesapeake Bay Dataset

Context

The dataset ("*sample_with_seasons.csv"*) we are working with comes from the Chesapeake Bay region and is accessible at Dryad. This dataset contains information about water properties such as temperature (in degree Celsius) and salinity collected over several years. Salinity is the amount of salt dissolved in a body of water. It is measured in grams of salt per kilogram of water.

As part of our analysis, the GEA instructors performed **feature engineering** by creating a new column for **season** based on the date of sample collection. Feature engineering is choosing and modifying data in a smart way to help a regression model make better predictions.

Feature engineering is important because it helps extract meaningful patterns from raw data and enhances the dataset by providing a new perspective—grouping data into seasons allows us to uncover seasonal trends in temperature and salinity that may not be immediately evident.

The goal of this question is to conduct **exploratory data analysis (EDA)**, calculate **individual-level** and **ecological correlations**, and understand when we will be committing **ecological fallacy**.

Exploratory Data Analysis (EDA)

a) Perform EDA on Salinity by doing the following:
    i. Calculate suitable summary statistics (5-number summary, mean, SD and IQR).

Mean	
Median	
Minimum	
Maximum	
SD	
Q1	
Q3	
IQR	

ii. Construct a boxplot.

iii. Find the number of outliers in the boxplot constructed in (ii).
Hint: You can use filters or consider creating a new categorical variable that indicates whether a value is an outlier.

b) Using a suitable plot, compare the distribution of Temperature between the four seasons. Based on the plot, are your observations as expected? Which seasons have the smallest and largest temperature ranges?

c) Now, compare the distribution of Salinity between the seasons using suitable plots. How similar/different are the salinity levels between the seasons?

Correlation Analysis

d) Calculate the correlation coefficient between Salinity and Temperature using the original data (without removing any outliers). Plot a scatter plot and regression line of Salinity against Temperature. Observe and interpret the strength and direction of the correlation between these two variables.

e) Seeing what you have done in part (d), one of the GEA instructors decided to calculate the ecological correlation between the average values of salinity and temperature for each season. Find these two sets of averages. Plot a scatter plot and regression line of average Temperature against average Salinity. Calculate the ecological correlation and compare it with the individual-level correlation found in (d). State your observations.

Fallacy

f) With the calculated ecological correlation in (e), the instructor immediately claimed that the correlation calculated from the individual data points will have the same value and direction as the ecological correlation. Has the instructor committed a fallacy? If so, what fallacy has been committed?

g) Reflect on how the difference between individual correlation and ecological correlation might impact the interpretation of the relationship between salinity and temperature.

Question 2: Analysing Global Demographic Trends

As part of an ongoing project to analyse global demographic trends, researchers have compiled a simulated dataset, *population_age.csv*, generated by an intern. This dataset focuses on key indicators such as median age and total population. While this data serves as a valuable resource for applications like public policy planning, economic strategies, and projections for future healthcare and educational needs, it is important to note that the dataset was simulated for the purpose of this tutorial and may contain inaccuracies or errors.

Median_Age serves as a pivotal measure, showing the age that divides the population in two parts of equal size - there are as many persons in that region with ages above the median as there are ages below[1] the median. It is an indicator used to understand societal aging, potential workforce demographics, and consumer behavior patterns.

Population quantifies the total number of individuals residing within each region, influencing numerous aspects of national planning, including infrastructure development, resource management, and environmental strategies.

However, preliminary reviews of the data set have revealed some inconsistencies, particularly concerning median age values.

Objective: The primary goal of this analysis is to scrutinise the distribution of Median_Age and Population, identifying logical inconsistencies and significant outliers that could impact the validity of subsequent analyses. By addressing these data integrity issues, the research aims to refine the data set, ensuring that further demographic studies and policy recommendations are based on accurate and reliable data.

Approach: The analysis will begin by charting the distributions of both Median_Age and Population to visually identify and statistically confirm any aberrant or anomalous values. This step is crucial in maintaining the integrity of demographic data, which must be free from significant errors to support effective decision-making and accurate forecasting.

a) Describe the distributions of Median_Age in the data set using
    i. suitable summary statistics (5-number summary, SD and IQR)
    ii. visualisation tools (histogram and boxplot).

Are there any noticeable trends or outliers that could affect the analysis?

[1] People with age = Median_Age are not included in the top or bottom half.

b) Before attempting this part, remove data points with Median_age below 15[2].
   i. Construct a scatter plot of Population vs Median_Age. What do you observe about the relationship between these two variables?
   ii. Now construct another scatter plot, this time for ln(Population) vs Median_Age. (Note: ln() is the natural logarithm. You could also try using logarithm of base 10, that is, log10(Population)). What do you notice about the relationship shown in this scatter plot?
   iii. Calculate and compare the correlation coefficients in parts b(i) and b(ii).
   iv. Share a brief observation based on parts b(i), (ii) and (iii).
   v. Discuss why taking the natural logarithm of the population might be beneficial for this type of analysis. How does transforming the data affect the interpretation of the results?

c) Fit a linear regression model using Median_Age as the independent variable and the natural log of Population as the dependent variable. What is the equation of the fitted line?

d) Interpret the slope and intercept of the regression line you obtained for the previous part. What do these coefficients/values imply about how Median_Age is related to Population?

e) Using the regression model, predict the natural log of the Population for the given Median_Age of 35 years. Convert this logarithmic prediction back into the actual population estimate.

[2] Countries do not typically have a median age below 15 years.
https://ourworldindata.org/age-structure