Exercise 6.1

Data Source

1. Dataset: Covid-19 pandemic dataset from "Our World in Data".
    a. Data Source Summary
        i. Dataset sourced from: https://ourworldindata.org/covid-vaccinations
        ii. Type of Data: External data
        iii. Owner: Sourced in from various public/government websites like WHO, Ministry of Health of various countries, vaccine trial from vaccine companies and so on.
        iv. Trustworthiness: The platform is run by researchers and data scientists, primarily from the University of Oxford. They provide thorough analysis and visualization of data, making complex data accessible and understandable. Many of the analysis is peer-reviewed and source of data is provided at the end of every analysis page.

    b. Data Collection Method
        i. Type of Data: Administrative data
        ii. Collection Method: The data is collected automatically.
        iii. Time Lag: There is no time lag as the data is updated regularly.

    c. Overview of Data Contents
        i. Variables

    Included for analysis:

        1. Iso_code: Country code
        2. Continent: Continent
        3. Location: Country
        4. Date: Date of record
        5. Month: Month extracted from date
        6. Year: Year extracted from date
        7. Total_cases: Total number of cases of corona (continuous increase which means the current number is the total number of cases)
        8. Total_deaths: Total number of deaths due to corona (continuous increase which means the current number is the total number of deaths)
        9. Reproduction_rate: The reproduction rate indicates the average number of secondary infections produced by a single infected individual in a fully susceptible population.
        10. Icu_patients: Patients admitted in ICU due to corona
        11. Hosp_patients: Patients who had to be hospitalized
        12. Total_tests: total testing done for corona
        13. People_vaccinated: Vaccinated persons
        14. People_fully_vaccinated: Fully vaccinated persons
        15. Aged_65_older: Population percentage of above 65 years olds
        16. GDP_per_capita (in usd): Gross Domestic Product per capita provides a per-person measure of standard of living.
        17. Cardiovasc_death_rate (per 100,000): Death due to CVD
        18. Diabetes_prevalence: % of population diagnosed with Diabetes
        19. Female_smokers: % of female Smokers
        20. Male_smokers: % of male Smokers
        21. Life_expectancy: Average life expectancy

22. Population: Total population as in 2024

d. Why this dataset was chosen?

I am particularly interested in pandemic because I was working for a pharmaceutical industry and was closely monitoring the IP generated in vaccines against Coronavirus. It is relevant for me to look at the pandemic data at a deeper level.

Data Cleaning

I have used both Python and Excel to clean the dataset. I have performed the following steps:

1. Used Python to select columns which were useful for my analysis. Identified missing data and duplicates (none). Found out what steps to be taken in Excel to generate a final dataset.
2. Used Excel to remove rows which were not needed.
3. Used Excel to pivot the information.
4. I started with 409882 rows and 67 columns.
5. The final dataset comprises 14 columns and 232 rows.

Limitation

Data in OWID is collected from various sources like directly from ministry/WHO (on later dates). These are the limitations stated on their websites.

1. Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.
2. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Bias & Ethics

1. Collection bias: Since information is being fed automatically from various websites, the accuracy is not checked personally. Some countries also lack any data and therefore, these countries were excluded and other sources were not looked for.
2. The data on pandemic is not personal, so it ethical and transparent. All the data sources are clearly mentioned.

Questions for analysis

a. Geographical Analysis:
    i. Countries that registered higher number of cases

b. Temporal Analysis:
    i. Identifying peak season

c. Predictive Modeling:
    i. Factors that could predict increased risk of spread of disease
        A. Population
        B. Vulnerable population (Immunocompromised population, smoking etc)
        C. GDP
        D. Vaccination

Please find the final dataset here.