CPSC – 375 – 02

Linh Nguyen

PROJECT 1 REPORT

A. Data preparation:

1. Read data files:
   Store dataset that that contains the number of vaccine doses given everyday in different countries as “vaccine_doses” data frame.
   Store dataset that that contains hospital beds as “hospital_beds” data frame.
   Store demographics dataset as “demographics” data frame.

2. Vaccine_doses data frame:
   a. Discard unnecessary data:
      - Remove rows under countries that have data of provinces or states:
        Rows which are data of province/states are not NA. Rows having data of countries will be NA in column Province_State. Then remove rows which Province_State is not NA.
      - Only keep columns that have country name, population, number of shots each day.
        Remove other columns: UID, iso2, iso3, code3, FIPS, Admin2, Province_States, Lat, Long_, Combined_Key.

   b. Tidy up table:
      Mess type 1: Column headers are values, not variable names.
      Solution: pivot_longer. There will be a column of date and a column of number of vaccine shots.
      Remove rows that number of vaccine shots is NA or 0.
   c. Calculate required variable:
      - Vaccination rate:

      Add a new column to the table named “vaccination_rate”, its value is calculated by dividing the number of vaccine shots by population.
      - Day since vaccination starts:

      Add a new column called “day_since_start”, its value starts with 1 as the first day of vaccination in that country, increasing until the last day of the dataset.

3. Hospital_bed data frame:
   Keep rows that have hospital beds for the most recent 5 years for each country.

4. Demographics data frame:
   a. Discard unnecessary data:
      Remove `Country Code`, `Series Names` columns.

   b. Tidy up table:
      Mess type 2: one observation with multiple rows.
      Solution: pivot_wider. There will be columns of population for different age ranges.

c. Add up the male/female population:
Add new columns that store the sum of male and female population of each age range. Remove the old columns of male/female population.

5. Merge data:
   - Match country names between tables:
     - "Korea, Rep." (demographics) replaced by "South Korea"
     - "Iran, Islamic Rep." (demographics) replaced by "Iran"
     - "Iran (Islamic Republic of)" (hospital_beds) replaced by "Iran"
     - "United Kingdom of Great Britain and Northern Ireland" (hospital_bed) replaced by "Unied Kingdom"
     - "Republic of Korea" (hospital_beds) replaced by "South Korea"
     - "Korea, South" (vaccine_doses) replaced by "South Korea"
   - Join tables:
     Match column name to join: Rename "Country_Region" to "Country", "Country Name" to "Country"
     Full join vaccine_doses table and hospital_beds table by "Country", then full join the output table with demographics table by "Country", saved as "my_data" data frame.

Country	Population	Date	Num_of_shots	day_since_start	vaccine_rate	Year	Hospital beds (per 10 000 population)	SP.DYN.LE00.IN	SP.URB.TOTL	SP.POP.TOTL	SP.POP.80UP
Afghanistan	38928341	2021-02-28	8200	1	0.0002106434	2017	3.9	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-02-28	8200	1	0.0002106434	2016	5.0	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-02-28	8200	1	0.0002106434	2015	5.0	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-02-28	8200	1	0.0002106434	2014	5.0	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-02-28	8200	1	0.0002106434	2013	5.3	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-03-01	8200	2	0.0002106434	2017	3.9	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-03-01	8200	2	0.0002106434	2016	5.0	63.377	8535606	34413603	8555
Afghanistan	38928341	2021-03-01	8200	2	0.0002106434	2015	5.0	63.377	8535606	34413603	8555

B. Linear modeling the Covid vaccine rate: