CPSC – 375 – 02

Linh Nguyen

**PROJECT 1 REPORT**

1. **Data preparation:**
2. 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.

1. Vaccine\_doses data frame:
2. 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.

1. 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.

1. 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.

1. Hospital\_bed data frame:

Keep rows that have hospital beds for the most year for each country, then remove the column Year.

1. Demographics data frame:
2. Discard unnecessary data:

Remove `Country Code`, `Series Names` columns.

1. Tidy up table:

Mess type 2: one observation with multiple rows.

Solution: pivot\_wider. There will be columns of population for different age ranges.

1. 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.

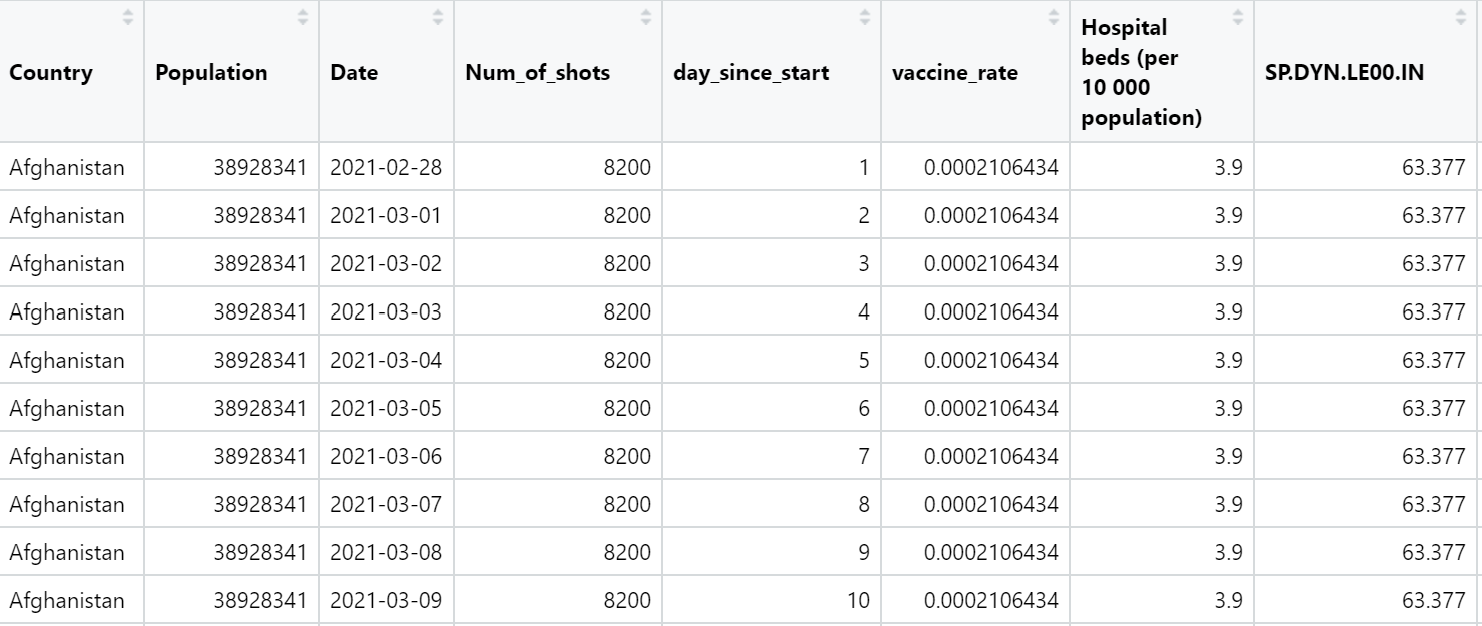
1. Merge data:

* 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.

The resulting data frame, called “my\_data”, looks like this:



1. **Linear modeling the Covid vaccine rate:**
2. Independent variables that are available for modeling:

* Population: indicates the population size. A country with large population may have some constraints in disseminating vaccines.
* ‘Hospital beds (per 10 000 population)’: indicates the infrastructure of healthcare system of a country. The lower number of hospital bed per 10,000 population may demonstrate that the country does not have enough capacity to provide health care services to its people, therefore affecting the vaccination rate.
* SP.DYN.LE00.IN (life expectancy at birth of each country): this statistic assesses the population health. For instance, Japan has this value of 83.79, meaning a child newly born in Japan is expected to live 83.79 years, hence their people may have better health foundation.
* SP.DYN.AMRT (Mortality rate of each country): this statistic may partly reflect the capability of a healthcare system, hence it can be combined with ‘Hospital beds (per 10 000 population)’to predict the vaccination rate.
* SP.URB.TOTL (Urban population), SP.POP.80UP (population ages 80 and above), SP.POP.1564.IN (population ages 15-64), SP.POP.0014.IN (population ages 0-14), SP.POP.65UP.IN (population ages 65 and above): since each country has different population size, we should examine the proportions of each kind per total population rather than the absolute parameters.

1. Transformed variables as predictors for modeling:

* Proportion\_65UP (proportion of population age 65 and above per total population): the higher rate indicates an aging population. This is also a population at high risk of death from Covid; hence the country may have to accelerate the vaccination of this population in order to reduce the number of deaths.
* Proportion\_1564 (proportion of population age 15-64 per total population): This is the main working population that creates wealth for society.
* Proportion\_urban: the proportion of urban population per total population may have some effect on the dissemination of vaccine over a country.

1. Scatterplot of the most recent vaccination rate for every country and the number of days since first vaccination:

Chart, scatter chart

Description automatically generated

1. Attempts to linearly model the vaccination rate:
2. Model1: vaccine\_rate ~ Population

R2 = 0.004795 : too low

1. Model2: vaccine\_rate ~ ‘Hospital beds (per 10 000 population)’

R2 = 0.05712 : too low

1. Model3: vaccine\_rate ~ SP.DYN.AMRT

R2 = 0.5006 : acceptable

1. Model4: vaccine\_rate ~ SP.DYN.LE00.IN

R2 = 0.5903: acceptable

1. Model5: vaccine\_rate ~ Proportion\_65UP

R2 = 0.2154 : low

1. Model6: vaccine\_rate ~ Proportion\_1564

R2 = 0.2862 : low

1. Model7: vaccine\_rate ~ Proportion\_urban

R2 = 0.2889 : low

1. Model8: vaccine\_rate ~ SP.DYN.LE00.IN + Proportion\_65UP

R2 = 0.5956: acceptable

1. Model9: vaccine\_rate ~ SP.DYN.LE00.IN + Proportion\_65UP + SP.DYN.AMRT

R2 = 0.5842 : acceptable

1. Model10: vaccine\_rate ~ SP.DYN.LE00.IN + Proportion\_65UP + ‘Hospital beds (per 10 000 population)’

R2 = 0.5968 : acceptable

Summary of R2 of 10 models:

Chart, bar chart

Description automatically generated

1. Conclusion:

Model1, model2, model5, model6, model7 have low values of R2 demonstrating that the population of a country, the ratio of hospital bed per 10,000 population, the proportion of population ages 65 and above, the proportion of population ages 15-64, the proportion of urban population do not affect the model. Therefore, population size, population composition, as well as infrastructure of the health care system are not necessarily the reasons for high or low Covid vaccination rate.

Model3, model4, model8, model9, model10 have R2 above 0.5 showing that the indices of life expectancy at birth and mortality rate are more likely to affect the model. This reflects that life quality or other certain aspects that contribute to high average life expectancy of civilians are more likely to affect the Covid vaccination rate.