Project Assignment 5

by Alisiya Balayan, Bishoy Abdelmalik, Arian Dehghani, Ariel Kohanim, Sergio Ramirez, Natalie Weingart

1. Dataset sources

a. Dataset A: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

b. Dataset B: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

c. Dataset C: https://www.kaggle.com/eng0mohamed0nabil/population-by-country-2020

2. Datasets Attributes

a. Dataset a

i. **Numeric:**

1. Total vaccinations per date and country (only U.S.)
2. Total vaccinations per hundred
3. Number of people vaccinated
4. Number of people vaccinated per hundred
5. Number of people fully vaccinated
6. Number of people fully vaccinated
7. Daily vaccinations
8. Daily vaccinations per million

ii. **Time-series:**

1. Date

b. Dataset b

i. Numeric:

1. Cumulative number of confirmed cases
2. Cumulative number of confirmed deaths
3. Cumulative number of recovered cases

ii. Categorical:

1. Province/State

iii. Time-series:

1. Observation dates

c. Dataset c

i. Numeric:

1. Population
2. Population density
3. Median age
4. Fert. Rate

d. Dataset d

i. **Categorical:**
   1. Entity
   2. Country Code

ii. **Numeric:**
   1. Daily Vaccinations

iii. **Time-series:**
   1. Dates

3. Feature Selection

a. Dataset a

- i. **Numeric:**
    1. Total Vaccinations
    2. Number of people vaccinated
    3. Number of people fully vaccinated
- ii. **Time-series:**
    1. Date

b. Dataset b

- **i.** **Time-series:**
    1. Observation date
- ii. **Numeric:**
    1. Cumulative number of confirmed cases
    2. Cumulative number of confirmed deaths
    3. Cumulative number of recovered cases

c. Dataset c

- **i.** **Numeric:**
    1. Population
    2. Population density
    3. Median age

d. Dataset d

- **i.** **Categorical:**
    1. Entity
- **ii.** **Time-series:**
    1. Day
- **iii.** **Numeric:**
    1. Daily Vaccinations

We selected our attributes to improve the accuracy of our results so we selected the data relevant to the spread of COVID-19 and the increase of herd immunity. We selected the total number of vaccinations per date and the total vaccinations per hundred because of its significance in immunization. Similarly, we selected total number of vaccinated people and people vaccinated per hundred, as well as, total number of fully vaccinated and fully vaccinated people per hundred. Additionally, the total daily vaccinations and vaccinations per million were also selected as the increase of vaccines given is relevant to the increase of immunization.

Attributes regarding the spread of COVID-19 and population were also selected, like the cumulative numbers of cases, deaths, and recoveries to improve the accuracy of immunization predictions in regards to the continuing spread of COVID-19. Population attributes were selected like population, density and median age as these factors greatly influence the chances of catching, spreading, dying or recovering from COVID-19.

4. Approach

a. Numerical inputs

   i. Outputs will be either numeric or categorical. Therefore, we will be using Pearson algorithm (numeric output) and ANOVA regression (categorical output). For example, our vaccination data contains both numeric and time-series data so we will first analyze the numeric using Pearson algorithm, then we will proceed to analyze data points for different time-series using cross correlation function.

b. Time-series inputs

   i. Since we have time-sequence data which we discretized into monthly periods of time for evaluation, we will be using cross correlation function which is a signal processing technique to measure the similarity between different time periods.