Project – Part 1: Reflection Statement

# Raw Data Schema

The common analysis research question needed the following datasets:

* The **RAW\_us\_confirmed\_cases.csv** file from [the Kaggle repository of John Hopkins University COVID-19 data](https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university). This data is updated daily. You can use any revision of this dataset posted after October 1, 2022.
* The [CDC dataset of masking mandates by county](https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i). Note that the CDC stopped collecting this policy information in September 2021.
* The [New York Times mask compliance survey data](https://github.com/nytimes/covid-19-data/tree/master/mask-use).

## Schema of Raw US confirmed cases

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data type** | **Description** |
| Province\_State | String | State Name (Example, “New Jersey”) |
| Admin2 | String | County Name (Example, “Essex”) |
| FIPS | Int | FIPS code for the county (34013 == Essex County, NJ) |
| 1/22/20 | Int | Count of confirmed cases for 1/22/20 |
| … | … | … |
| 10/31/22 | Int | Count of confirmed cases for 10/31/22 |

## Schema of Mask mandates by county

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data type** | **Description** |
| State\_Tribe\_Territory | String | State Name (Example, “NJ”) |
| County\_Name | String | Example, “Essex County” |
| FIPS\_State | Int | FIPS code for state. 34 == NJ |
| FIPS\_County | Int | FIPS code for county. 13 == Essex County |
| Date | DateTime |  |
| Face\_Masks\_Required\_in\_Public | Boolean | “Yes” if mask mandate is in effect; “No” otherwise |

## Schema of Mask compliance survey

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data type** | **Description** |
| COUNTYFP | String | FIPS code for the county. ‘34013’ == Essex County, NJ |
| NEVER | Float | Percentage of population that never complies with mask mandates |
| RARELY | Float | Percentage of population that rarely complies with mask mandates |
| SOMETIMES | Float | Percentage of population that sometimes complies with mask mandates |
| FREQUENTLY | Float | Percentage of population that frequently complies with mask mandates |
| ALWAYS | Float | Percentage of population that always complies with mask mandates |

# Data cleaning

There were 3 different datasets that were obtained from multiple sources with different schemas. So, the first step toward performing data analysis was to join these datasets and obtain a single combined dataset that can be analyzed.

From the above 3 datasets, we obtain a combined infection rates dataset that has the following schema:

* date: The date from Feb 2020 to Oct 2021.
* infected: This is the effective infected population on any given date. The raw dataset has cumulative confirmed cases, but the model assumes a 14-day recovery window and computes effective infected population considering the recoveries.
* rate: This is the infection rate, which is computed as (today's infected population / yesterday's infected population).
* mandate: 1 if there was a mask mandate in effect; 0 otherwise.
* usage: The mask usage percentage, which is 68.2% for Essex County, NJ.

The resulting cleaned dataset for **Essex County, NJ** is shown below:

Table

Description automatically generated with medium confidence

# Learnings

The key learnings from this part of the project were about how to combine the raw datasets with disparate schemas into a single clean dataset for visualization and analysis.

Firstly, the raw confirmed cases dataset had dates as columns, so the cleaning required to pivot the dates as rows in order to join it with the mask mandate dataset.

Further, the raw confirmed cases contained a cumulative total of COVID-19 cases across the date range. As part of the cleaning, a recovery window of 14 days was assumed, so that we eliminate the recovered population to get an effective infected population count for each day.