Lasata Tuladhar (Leader), Katherine Driebe, Ramya Tangirala
DS 4002
3/19/24

Data Appendix

Introduction

The purpose of this project is to determine any trends in national COVID data from 2020 to 2024 if certain factors (such as weekly deaths, weekly hospitalizations, etc.) are highlighted. This project is targeted toward hospital staffers and policymakers to understand how to best combat COVID-19 in the future.

Original Datasets Overview

- Entries: Each of the two CSV files contains 215 entries, with each entry representing a survey response.
- Variables:
    - The first dataset includes data on weekly COVID-19 hospital admissions, comprising a total of three variables: geography, date, and weekly COVID-19 hospital admissions.
    - The second dataset, which focuses on weekly deaths from COVID-19, contains four variables: geography, date, weekly deaths, and the date of the reported deaths.
- Time Frame: The datasets encompassing weekly hospital admissions and deaths in the United States from 2020 to 2024
- Dataset CSV: The data is derived from two distinct CSV files obtained from the CDC data tracker. These datasets are hosted on GitHub, enabling users to conveniently download them. Moreover, the script provides functionality for users to upload the CSV files directly from their computers.

DF Variable Descriptions:

Data Dictionary 1:

Variable Name	Variable Type	Description
Geography	pandas Series / String	The location, specifically country, of where the record was taken. In this dataset, the Geography is the United States for all entries.
Date	pandas Series / String	The week the data entries correspond to.
Weekly COVID-19 Hospital Admissions	pandas Series / Float64	The count of new hospital admissions due to COVID-19 that week.

Lasata Tuladhar (Leader), Katherine Driebe, Ramya Tangirala
DS 4002
3/19/24

Data Dictionary 2:

Variable Name	Variable Type	Description
Geography	pandas Series / String	The location, specifically country, of where the record was taken. In this dataset, the Geography is the United States for all entries.
Date	pandas Series / String	The week the data entries correspond to.
Weekly Deaths	pandas Series / Float64	The number of deaths that week.
Death Date As Of	pandas Series / String	The date the data was reported.

Missing Values and Data Integrity

- The datasets initially had missing values across various columns. Furthermore, although the data is updated weekly, the analysis conducted by our script will only reflect the most recent update from the time we last downloaded it from the CDC website.

Merged Dataset Overview

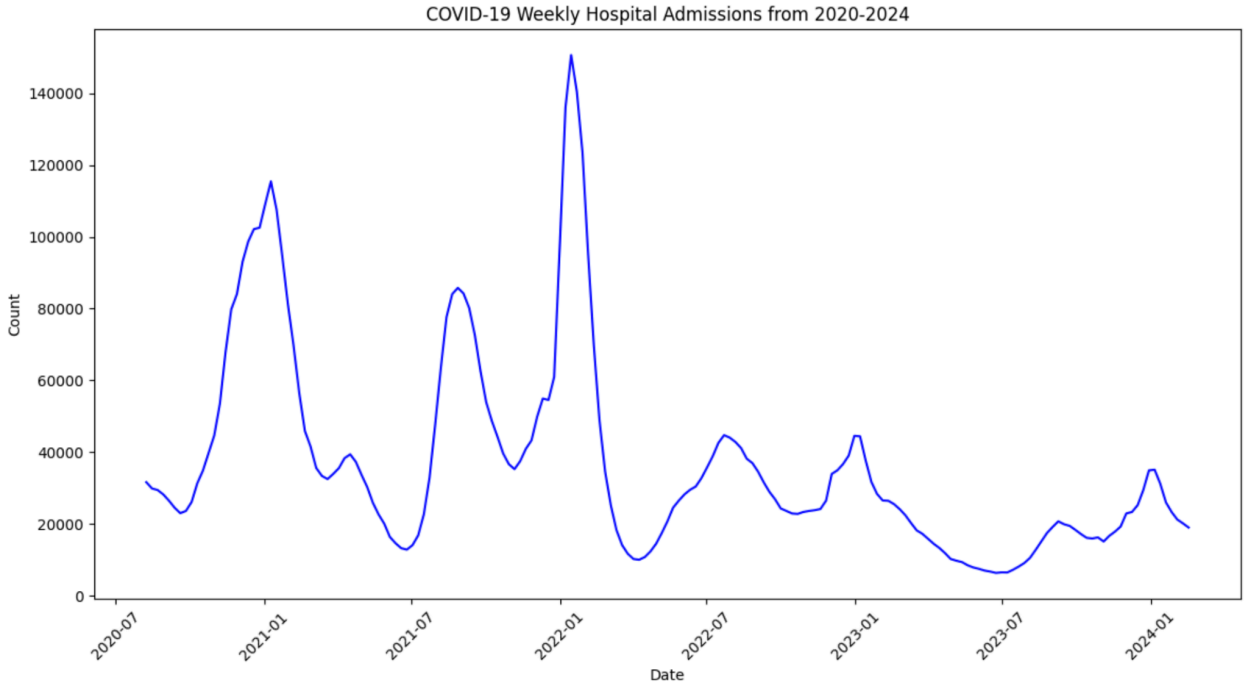
- Entries and Variables: The dataset contains 185 entries, each representing a survey response. There are five variables in total: geography, date, weekly COVID-19 hospital admissions, weekly deaths, and the date of the reported deaths.
- Changes in the Dataset: During the cleaning process duplicates and rows with missing variables were dropped. The `Date` variable in both DataFrames was converted from string format to Pandas datetime objects. The two datasets were merged on the `Date` and `Geography` columns, creating a single DataFrame (`df_merged`) that contains both weekly hospital admissions and deaths data.

Cleaned DF Variable Descriptions

Variable Name	Variable Type	Description
Geography	pandas Series / String	The location, specifically country, of where the record was taken. In this dataset, the Geography is the United States for all entries.

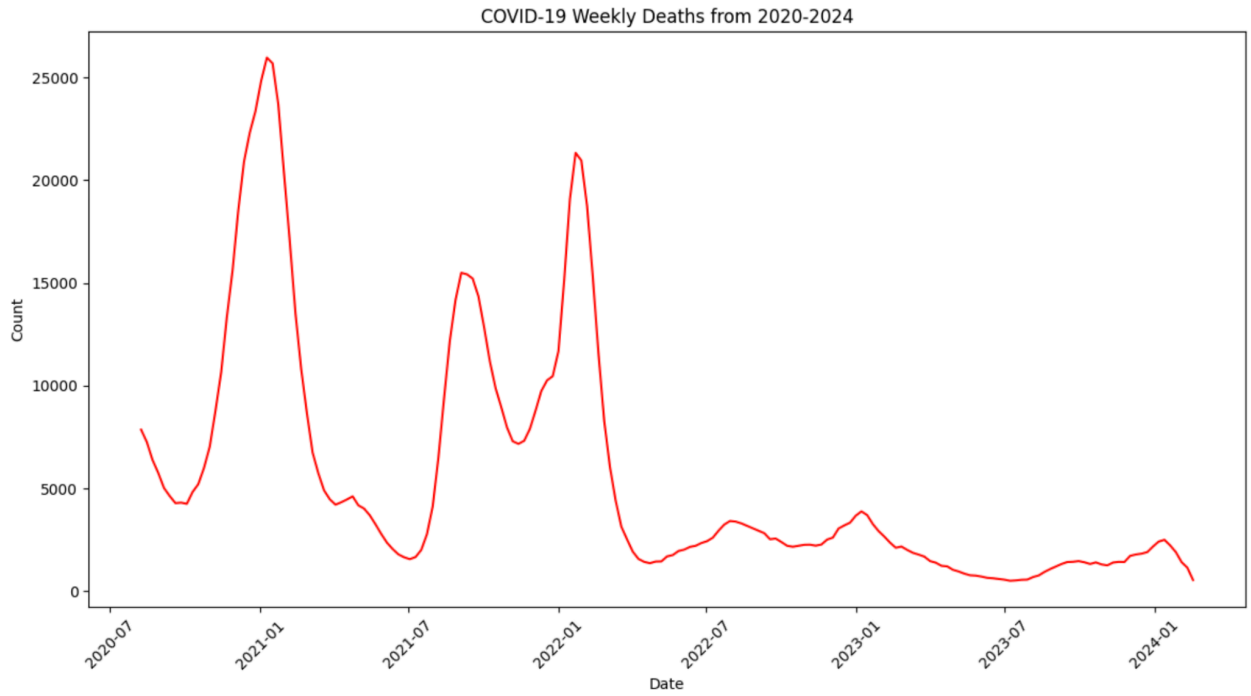
Date	pandas datetime / object	The week the data entries correspond to.
Weekly Deaths:	pandas Series / Float64	The number of deaths that week.
Death Data As Of:	pandas Series / String	The date the data was reported.
Weekly COVID-19 Hospital Admissions	pandas Series / Float64	The count of new hospital admissions due to COVID-19 that week.

Detailed Statistical Analysis



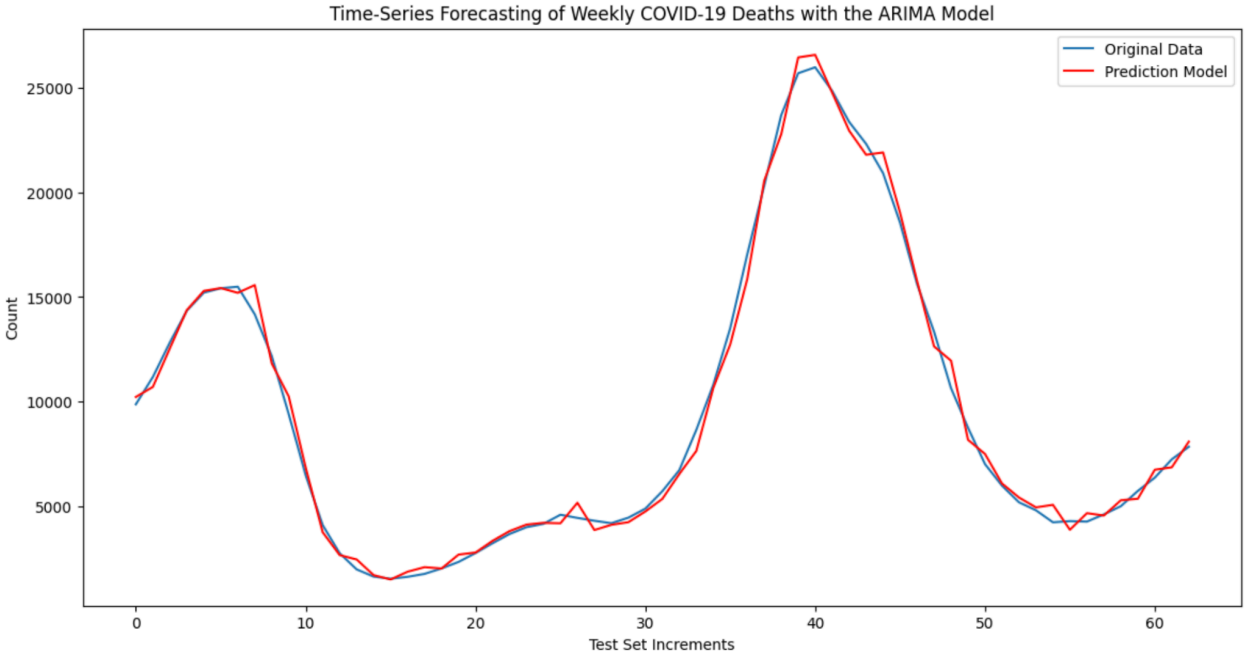
The graph depicts COVID-19 weekly hospital admissions from 2020 to early 2024. From the graph, we can observe several peaks, which suggest periods when the number of hospital admissions due to COVID-19 significantly increased. These peaks may correspond to surges or waves of COVID-19 infections. The valleys indicate periods when admissions were lower, possibly suggesting times when the infection rates were under better control or during lulls between surges. The trend appears to be cyclical, with the number of admissions rising and falling in a somewhat regular pattern, potentially correlating with seasonal changes or the effects of public health interventions and vaccination campaigns.

Lasata Tuladhar (Leader), Katherine Driebe, Ramya Tangirala
DS 4002
3/19/24

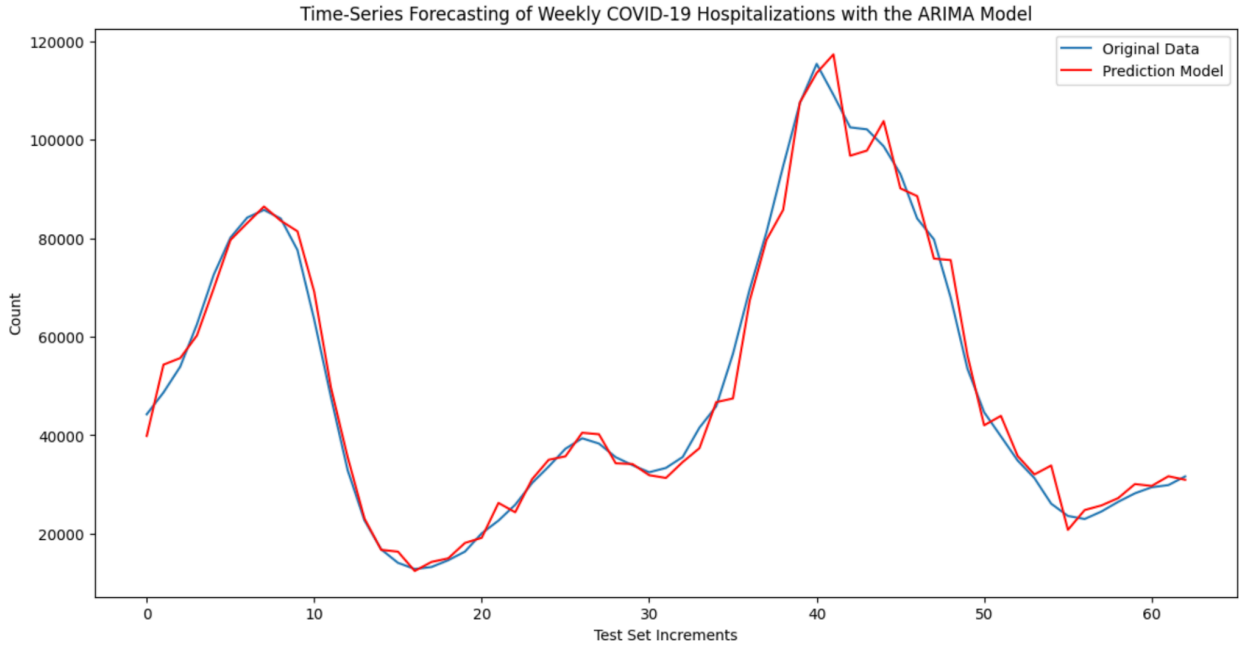


The graph depicts COVID-19 weekly deaths from 2020 to early 2024. From the graph, we can observe several peaks, which suggest periods when the number of deaths due to COVID-19 significantly increased. The valleys indicate periods when the number of deaths caused by COVID-19 was lower. The trend appears to be cyclical, with the number of deaths rising and falling in a somewhat regular pattern mirroring the peaks in valleys observed in the COVID-19 weekly hospital admissions. Additionally, an overall decline in deaths due to COVID-19 can be observed.

Lasata Tuladhar (Leader), Katherine Driebe, Ramya Tangirala
DS 4002
3/19/24



This graph shows the time-series forecasting of weekly COVID-19 deaths using the ARIMA model and plots the original data and the prediction data. The prediction model fits the original data well as can be seen by the graph above. Although there are discrepancies, they are minor and the model has an overall good fit. The future predictions cannot be seen but they cannot be verified as the actual data has not been released for comparison.



This graph shows the time-series forecasting of weekly COVID-19 hospitalizations using the ARIMA model and plots the original data and the prediction data. The prediction model fits the

original data well as can be seen by the graph above. The model has an overall good fit, but it can be observed that the prediction model was slightly off at the highest peak indicating that at that point, the original data does not match the predicted values. The future predictions cannot be seen but they cannot be verified as the actual data has not been released for comparison.

References [IEEE Format]:

[1] "CDC Museum COVID-19 Timeline" *https://www.cdc.gov/museum/timeline/covid19.html* https://www.cdc.gov/museum/timeline/covid19.html