# Covid Market Modeling

Pavan Chaitanya Penagamuri, Sz-Wei Charng, Artur Rocha, Joshua Reifel, Viswa Jayasankar, Krishnan Unni

PURDUE UNIVERSITY | The Data Mine

abbvie

## Introduction

This project aims to develop a forecasting model for COVID-19 cases across the USA to assist in AbbVie's understanding of the disease in its endemic state because of the large disruption to patients across the nation due to the pandemic.

### Objectives

- Estimate total and peak cases for future COVID-19 seasons as reflected in AbbVie data.
- Establish a clear understanding of case seasonality.

### Central Questions

- What variables may contribute to the spread of COVID-19?
- How may we best model the seasonality and covariances of COVID-19 spread?

## Data Exploration

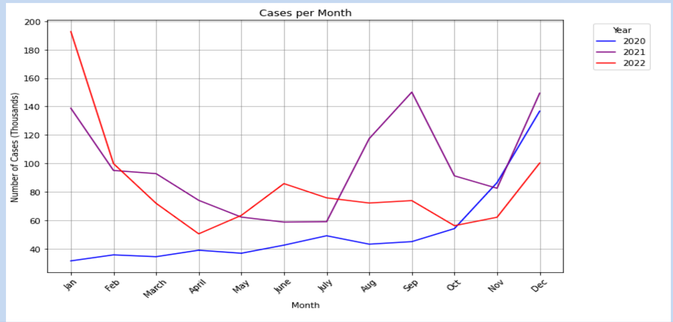
| Patient ID | Diagnosis Code | Diagnosis Description | Income Code | Date | State | Ethnicity | | Gender | Birth Year |
|---|---|---|---|---|---|---|---|---|---|

**Given Data:**

- 126 million cases of COVID

**What we did:**

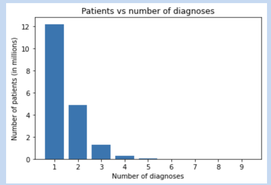
- Data cleaning (Removed all the undefined rows and unnecessary columns)
- Data Exploration (Graphed the data and identified patterns)
- Data Visualization (Made the visuals with the Data)



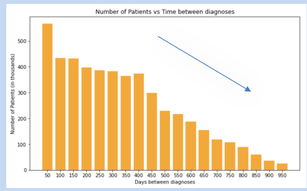

- Peaks in the data:
  - **September**
  - **End of December** and **Beginning of January**
- End of August is when **school starts**
- End of December is when people **travel for Christmas**
- Both lead to people coming in **close contact with each other**

## Exploratory Data Analysis

### Multiple diagnosis & Time Gap between diagnoses

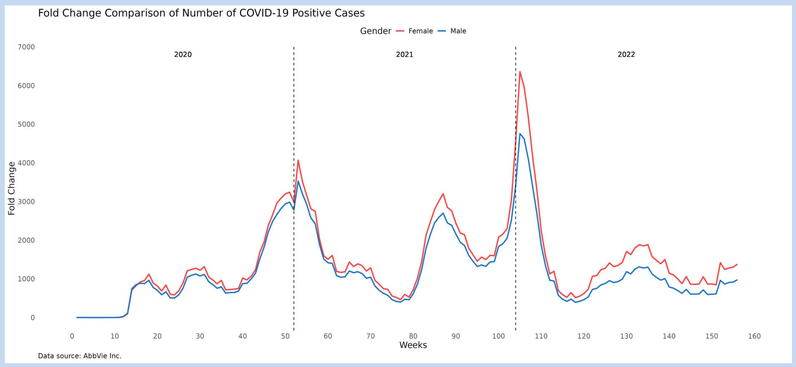



- Very few people have more than two diagnoses
- Time gap between diagnoses has decreasing trend




### Fold change analysis

- Considered the number of cases in the first week as the baseline
- There was **no significant difference between males and females** compared each month
- The **highest peak** in COVID-19 cases (January 2022) was **6,000 times greater** than in January 2020.




### Data Sampling

Data was sampled in a way which would minimize bias. Non-covid data was included in the dataset to ensure minimal bias after sampling.

**Combining fixed effects to create a new column.**

**Calculate frequency of unique combinations in the newly created column.**

**Grouping data by values in the new column.**

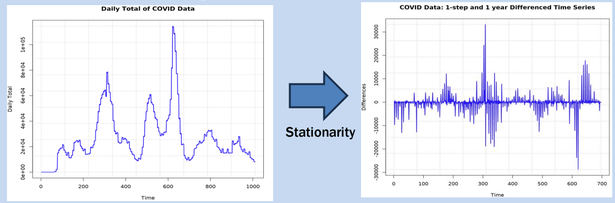
**Sampling is performed independently within each group.**

## Time series Analysis

### Seasonal Auto-regressive Integrated Moving-Average (SARIMA)

**Step 1: Check stationary of data and Converting Non-Sationary data into Sationary**

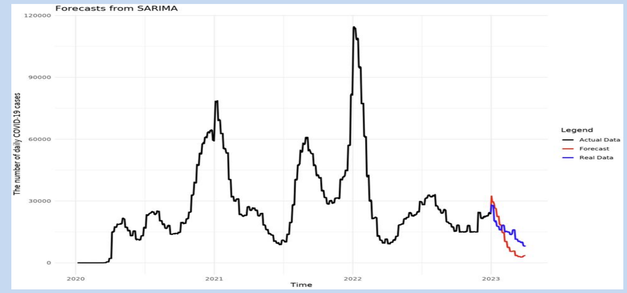
- Use ADF/KPSS tests check stationary
- First order difference + 312 steps seasonality difference




**Step 2: Select model by Auto-Correlation Function and Partial Auto-Correlation Function**

**Step 3: Fit SARIMA model and predict the future**

- Our SARIMA Model captures the the overall trend when forecasting the next three months




## Alternate Approaches

**Long short-term memory**

- Model too complex for the data
- Over-fits to the training data

**Generalized linear mixed models**

- Required too much memory
- More time for Bayesian methods

## Conclusion

Utilizing a massive dataset of nearly 120 million rows, we modeled COVID cases, identifying peak months and recurring patterns in patient diagnoses. Through meticulous data sampling and time-series analysis using SARIMA modeling, we forecasted COVID cases across the USA, emphasizing the critical role of data-driven insights in understanding and predicting pandemic trends.

## Future Work

- Collect more data features related to symptoms to develop a classification model for predicting covid diagnosis
- Perform geo-analysis of covid spread across US states

## Acknowledgements

Corporate Partner Mentors

- Tom Fenwick
- Ash Kharkar

TA: Chandni 'Maggie' Garg