# VISUALISING INFECTIOUS DISEASE SURVEILLANCE DATA FOR PUBLIC HEALTH

DATA7903 Capstone Project Report

Sem 1, 2022

School of Information Technology and Electrical Engineering
The University of Queensland

Lalith Veerabhadrappa Badiger
Student ID: 46557829

# Executive Summary

Surveillance is essential to improve public health as the threat of infectious diseases continues to rise. In this project, I demonstrate a robust reporting system for public health using Influenza disease surveillance data sourced from the National Notifiable Disease Surveillance System (NNDSS) website, as explained below.

Firstly, I have been building an interactive dashboard using Microsoft PowerBI that visualises every attribute of the notified cases. As per my findings, children under the age of 15 years are the most affected age group accounting for over 30% of the cases. Moreover, the Influenza-A virus is the dominant strain in Australia, and NSW is the most infected state. The flu has a yearly seasonality that peaks every winter, i.e., June to September. Secondly, I have been constructing a Phylogenetic tree that visually illustrates the evolutionary relationship between different virus types (Ludmir & Enquist, 2009). Here, I observed that H1N1 and H3N2 are closely related, and Influenza type B share a common ancestor with influenza type C. Thirdly, I have been forecasting Influenza cases for 52 weeks by experimenting with the Holt-Winters method, LSTM (Long short-term memory), Facebook Prophet, and SARIMA (Seasonal Autoregressive Integrated Moving Average), and SARIMAX (SARIMA with exogenous factors) model. After evaluating based on root mean squared error, I found the SARIMA to be the most accurate model. It shows yearly seasonality and a significant peak during winter. Finally, I have been simulating the influenza epidemic based on a deterministic SEIR (Susceptible-Exposed-Infectious-Recovered) model. It visualises the impact of preventive measures like wearing a mask and incorporating social distancing to help reduce the spread of infection (He, Peng, & Sun, 2020). However, this project has one shortcoming, as explained below.

The only limitation of this project is that the dashboard uses historical data. Therefore, for surveillance to be effective, real-time data analysis is essential. For example, the Australian government could create an API that streams national public health surveillance data accessible by common people. Additionally, using Apache Beam and Python, real-time analysis can be achieved by constructing data pipelines to extract, manipulate, and analyse streaming data.

To conclude, this project is an effective public health reporting system that enables quick and informed decision-making by the authorities, thereby preventing yet another pandemic.

# Table of Contents

## Introduction

Surveillance is critical to improving public health as the threat of infectious diseases escalates. Insights obtained from public health monitoring systems enable authorities to come to a timely resolution and act accordingly. The consequences of an uncontrolled spread of infections can disrupt a country's economy and social system. It can result in a significant loss of human life, posing an unprecedented threat to people's livelihood, public health, and food systems (Chriscaden, 2020). Although the Australian government established the National Notifiable Disease Surveillance System (NNDSS) to coordinate the national surveillance of an agreed list of infectious diseases, it has limited reporting functionality, such as the lack of interactive visualisations or forecasting methods. As explained below, I demonstrated a robust reporting system for the influenza virus to address this problem.

Firstly, I created an interactive dashboard for public health using Influenza disease surveillance data sourced from the NNDSS website. Additionally, I investigated time series plots to discover the trend and seasonality of the flu using the ETS (Error-Trend-Seasonality) Model. Secondly, I constructed a Phylogenetic tree to visualise different strains of the Influenza virus evolving in a hierarchical nature. Thirdly, I forecasted Influenza cases by experimenting with the Holt-Winters method, LSTM (Long short-term memory), Facebook Prophet, SARIMA (Seasonal Autoregressive Integrated Moving Average), and SARIMAX (SARIMA with exogenous factors) model. Finally, I simulated the influenza epidemic based on a deterministic SEIR (Susceptible-Exposed-Infectious-Recovered) model that visualises the impact of preventive measures like wearing a mask and incorporating social distancing to help reduce the impact of the spread of infection.

I followed a typical data science process to build this reporting system, i.e., data gathering, pre-processing, exploration, visualisation, and modelling. Below, I have discussed the significance and reasons behind choosing influenza disease for my project.

## Background

Influenza (or flu) is a highly contagious disease that affects people's nose, lungs, and throat. The possible symptoms include cold, cough, fever, sore throat, muscular pains, and tiredness. In rare cases, flu also causes death. Australian Bureau of Statistics reported influenza deaths to be 1181, 148, and 902 in 2017, 2018, and 2019, respectively (Attwooll, 2021), as shown in fig 1.

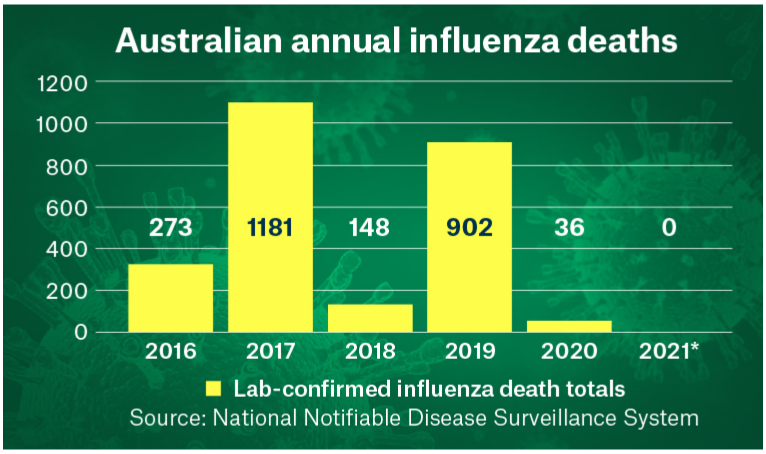


Fig 1. Australian annual influenza deaths

There are four types of influenza viruses: A, B, C, and D viruses. Influenza type A and B viruses cause seasonal epidemics almost every winter. The virus's genetic material provides the instructions for making additional copies of the same virus. Projecting from the virus's outer envelope are the H and N spikes of protein molecules, as shown in fig 2. The flu virus utilises its H spikes as a key to access the cells, while N spikes allow copies of the virus to break off from infected cells and infect more cells. Scientists use 18 different types of H spikes and 11 different N spikes to designate the subtypes of flu viruses like H1N1, H3N2, etc. ("Types of influenza viruses," 2019).

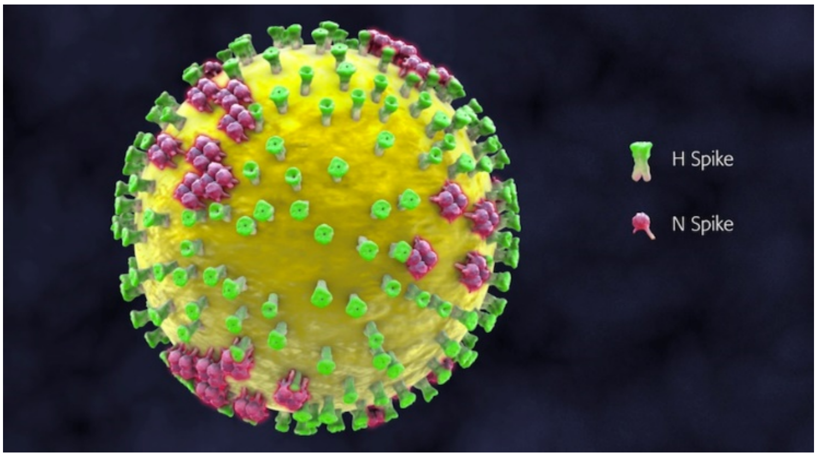


Fig 2. H and N spikes of protein molecules in Influenza virus

People can get infected by touching a contaminated object or contacting bodily fluids from infected humans or animals. For example, a person can catch the flu when infected individual talks, coughs, or sneezes. Droplets containing the influenza virus may land in their mouth or nose and then move into their lungs. The influenza virus enters the body, establishes contact with cells, and replicates to infect more cells (Gopinath et al., 2014). I have discussed the established methods to build a reporting system for public health surveillance below.

## Related works

There are two established methods for infectious disease surveillance:

1. **Search Query Analysis:** This method analyses search query data from various search engines like Google and Yahoo. The research shows a strong correlation between the number of individuals searching for influenza-related subjects and those with influenza symptoms (Carneiro & Mylonakis, 2009).
2. **Natural Language Processing:** This method automatically classifies and visualises Internet media reports such as news media and expert-curated ProMED Mails in a user-friendly interface to find critical disease outbreak information. This approach primarily uses text processing algorithms to categorise alerts by location and illness, then overlay them on an interactive world map called the "HealthMap" (Freifeld, Mandl, Reis, & Brownstein, 2008).

However, in this project, I obtained the data from Australia's National Notifiable Disease Surveillance System (NNDSS), a trusted official source. Hence, it resolves the ambiguities and inaccuracies found in the results of the above two methods. I have expanded upon this limitation of the above two methods below.

## Problem identification

Public health authorities find it very difficult to gain any actionable insights quickly from the NNDSS website. The primary reason for this problem is that the data is in a raw tabular format without any interactive visualisations, trends, or predictions. Although the Australian government established NNDSS to coordinate the national surveillance of an agreed list of infectious diseases, it has limited reporting functionality.

The problem with utilising search query data or news media data is that influenza recurs in predictable cycles each season. However, each epidemic's geographic location, timing, and magnitude differ, making accurate and timely assessments of influenza activity very problematic. This ambiguity will not be a problem since this project takes advantage of reasonably clean and accurate historical data obtained from a government-recognised body.

Hence, this is an excellent opportunity to build an effective and intuitive reporting system for infectious disease surveillance.

## Project objectives

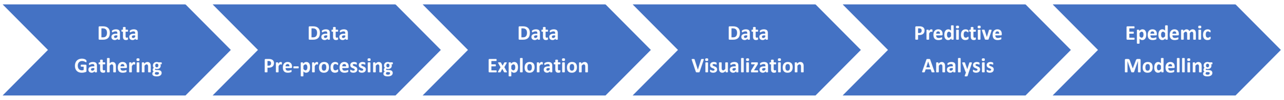
This project aims to build an effective reporting system for the flu that enables quick and informed decision-making by the public health authorities. This reporting system is achieved by:

- Developing an interactive dashboard with various visualisations and filters.
- Forecasting the number of flu infections using time-series data modelling.
- Using auxiliary data like weather data to discover any trends in the spread of infections.
- Visualising the impact of preventive measures intended to “Flatten the curve”.

I have delivered the above objectives using the steps and methods discussed in the next section.

## Methodology

To build a robust reporting system for influenza, I followed the six steps below:



### Data Gathering

This section focuses on data collection, data specifications and data caveats. Firstly, I downloaded the public health surveillance data from Australia’s National Notifiable Disease Surveillance System (NNDSS) website for “Influenza (laboratory confirmed)” disease. This data is recorded from 2008 to 2019 throughout Australia. Moreover, it is in Microsoft Excel format with a file size of around 26MB.

Secondly, each row of the dataset contains information about an individual infected with the flu characterised by the following attributes:

1. **Week Ending (Friday):** The date of the Friday following the day the notified case was diagnosed with influenza.
2. **State:** The Australian State or Territory that reports the influenza diagnosis to NNDSS. The included states are NSW, NT, Qld, SA, Tas, Vic, and WA.
3. **Age group:** The age group with 5-year bins.
4. **Sex:** The gender of the infected individual, i.e., “Male”, “Female”, “X”, or “Unknown”.
5. **Indigenous status:** It is “Indigenous” (Aboriginal or Torres Strait Islander origin), “Non-Indigenous” (Not of Aboriginal or Torres Strait Islander origin), or “Unknown” (Not stated, unknown or blank).
6. **Type/Subtype:** The type and subtype of the Influenza virus, namely “A(H3N2)”, “A(H1N1)pmd09”, “A(H1N1)”, “A(unsubtyped)”, “B”, “C”, “A and B”, and “Untyped”.

Lastly, there are three data caveats to be accounted for:

1. The dataset does not contain the Australia Capital Territory (ACT) information.
2. The recorded data is only for the cases that health care professionals diagnose. Hence, there will be a degree of under-representation compared with the actual number of cases.
3. It is crucial to note that variations in the number of cases over time do not always represent disease prevalence. Other factors like changes in testing regulations, screening programmes, and public awareness may impact the number of notified cases received each year.

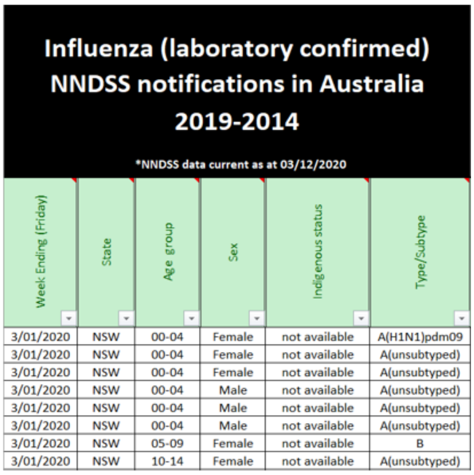
After gathering the required data, I pre-processed it as detailed below.

## Data Pre-processing

This section focuses on data cleaning, where I pre-processed the influenza dataset to a correct format using the “Pandas” library in Python.

Fig 3 shows the dataset before pre-processing. I followed the below three steps to pre-process the data:

1. I deleted the first row containing the heading and removed all the formatting and comments in the dataset.
2. This Excel file has two worksheets. One for data from 2019 to 2014 and the other from 2013 to 2008. I separated these two worksheets into two Excel files and then concatenated them to have a single Excel file with data from 2008 to 2019.
3. For predictive analysis, I reduced the Influenza dataset to only two columns. Since each row contains information about an individual infected with the flu, I grouped the entire dataset by the “Week Ending (Friday)” column. The second column is the counts of the incidents for that week in Australia. Then I parsed the “Week Ending (Friday)” column to the DateTime datatype and converted it to index.



Influenza (laboratory confirmed) NNDSS notifications in Australia 2019-2014

*NNDSS data current as at 03/12/2020

| Week Ending (Friday) | State | Age group | Sex | Indigenous status | Type/Subtype |
|---|---|---|---|---|---|
| 3/01/2020 | NSW | 00-04 | Female | not available | A(H1N1)pdm09 |
| 3/01/2020 | NSW | 00-04 | Female | not available | A(unsubtyped) |
| 3/01/2020 | NSW | 00-04 | Female | not available | A(unsubtyped) |
| 3/01/2020 | NSW | 00-04 | Male | not available | A(unsubtyped) |
| 3/01/2020 | NSW | 00-04 | Male | not available | A(unsubtyped) |
| 3/01/2020 | NSW | 00-04 | Male | not available | A(unsubtyped) |
| 3/01/2020 | NSW | 05-09 | Female | not available | B |
| 3/01/2020 | NSW | 10-14 | Female | not available | A(unsubtyped) |

Fig 3. Before data pre-processing

The following section discusses the exploratory data analysis of the pre-processed data.

## Data Exploration

In this step, I performed exploratory data analysis to determine if there are any patterns, traits, or areas of interest. Firstly, I plotted a time-series line plot to determine the trend of the influenza cases in Australia, as shown in fig 4. As a result, I observed an increasing trend in infections, with the highest number of cases in the third week of August 2017, i.e., 24161 cases.

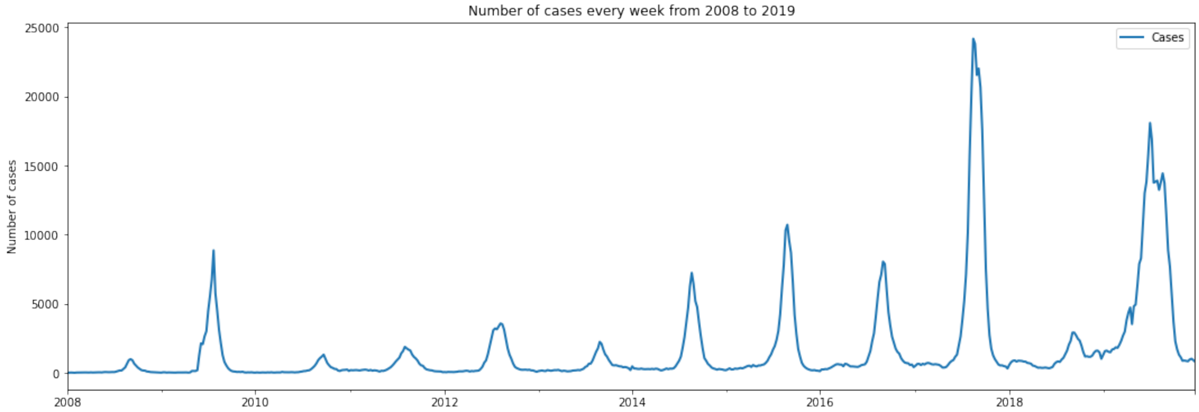


Fig 4. Number of influenza cases every week from 2008 to 2019

In addition, I plotted a bar chart that shows the total number of cases each year from 2008 to 2019 in Australia, as in fig 5. Although 2017 has the highest peak in the number of cases weekly, I observed that 2019 is one of the worst-hit flu seasons. The number of flu cases for 2019 has surpassed 300,000, increasing 81% over 2018 and 20% over 2017.

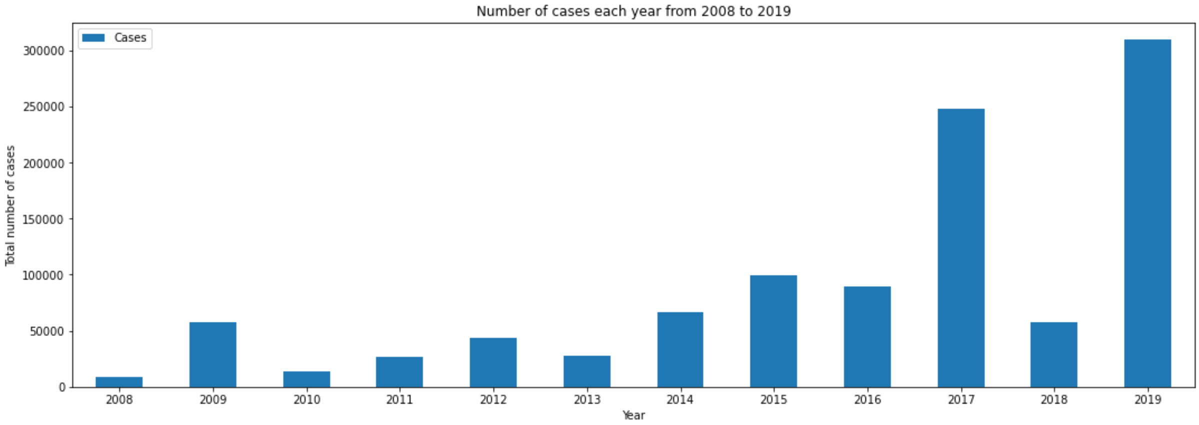


Fig 5. Number of cases every week from 2008 to 2019

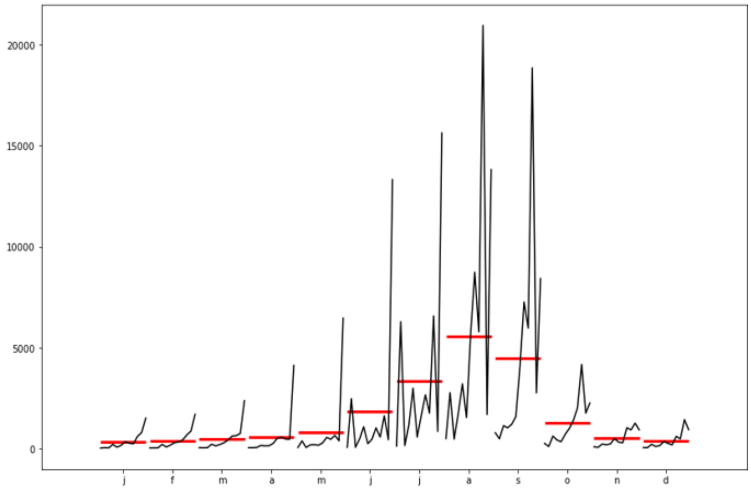


Fig 6. Seasonality plot

Secondly, I determined the seasonality of the model to be yearly seasonality that peaks every winter, i.e., June to September, as observed from the seasonality plot in fig 6. This plot shows the average number of cases every month, indicated by the red bars. The black line represents the trend of the total number of cases every year from 2008 to 2019, grouped by month. For example, the number of infections always peaks in August, with the highest total number of cases being around 24000 in 2017 and the average number of cases being just over 5000.

In addition, I decomposed the time-series data into Trend, Seasonal and Error components using the ETS (Error-Trend-Seasonality) model, as shown in fig 7. From the "Trend" plot, I observed an upward trend in the number of infections from 2008 to 2019. Additionally, I observed yearly periodicity from the "Seasonal Plot". The "Residual" plot is the difference of original time-series data from the sum of the trend and seasonal plot.

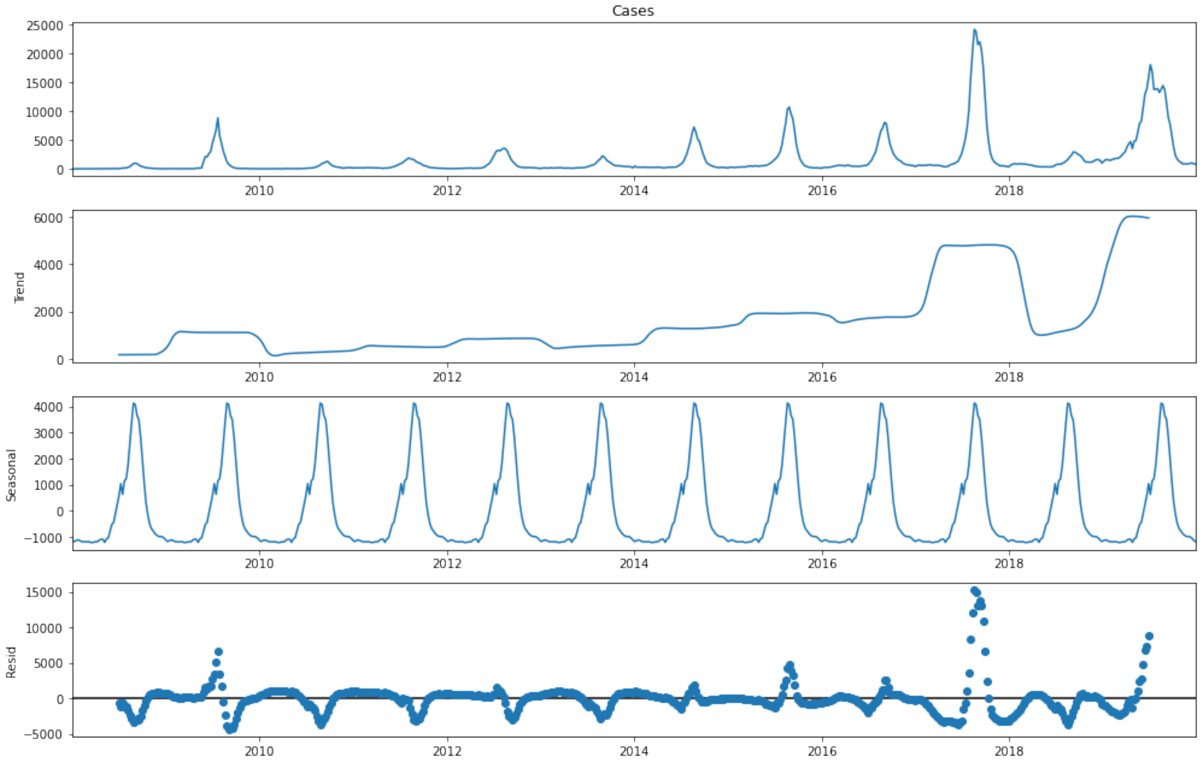


Fig 7. Error-Trend-Seasonality Plot

Thirdly, I used auxiliary data like temperature and rainfall to determine the relationship between these variables and the influenza cases. I obtained the daily minimum temperature (in degree Celsius) and rainfall (in millimetres) data for Brisbane from Australia's Bureau of Meteorology, the national weather, climate, and water agency. Then, I performed an Inner Join of this auxiliary time-series data with the "Week Ending (Friday)" column of the influenza dataset. From fig 8, I observed that the influenza cases generally increase as the temperature decreases during winter. However, from fig 9, I observed no concrete relationship between influenza cases and the rainfall amount.

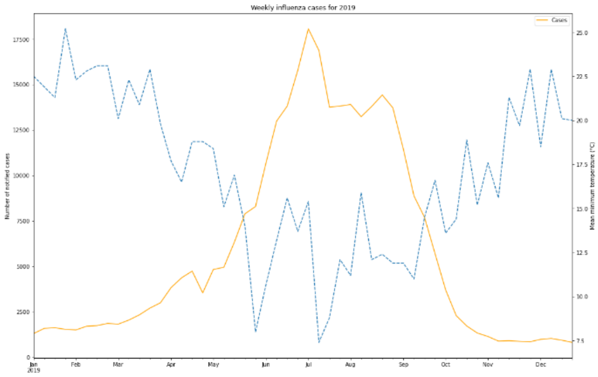


Fig 8. Influenza cases vs. temperature

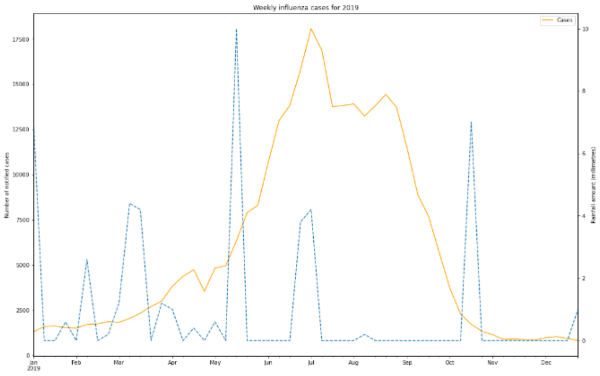


Fig 9. Influenza cases vs. rainfall

Finally, I plotted a Phylogenetic tree to visualise different strains of the influenza virus evolving in a hierarchical nature from an evolutionary standpoint. I used Nucleotide sequences to plot phylogenetic trees. The virus sequences separated by shorter evolutionary distances are expected to be more similar than those separated over longer evolutionary distances (Poon et al., 2013). Therefore, the Phylogenetic trees are simple and easy to understand by common people. I plotted a Phylogenetic tree that will represent only the types or subtypes of viruses found in Australia, as shown in fig 10. I observed that Influenza Type-A (H1N1) and Type-A (H3N2) are closely related, and Influenza Type-B share a common ancestor with Influenza Type-C.

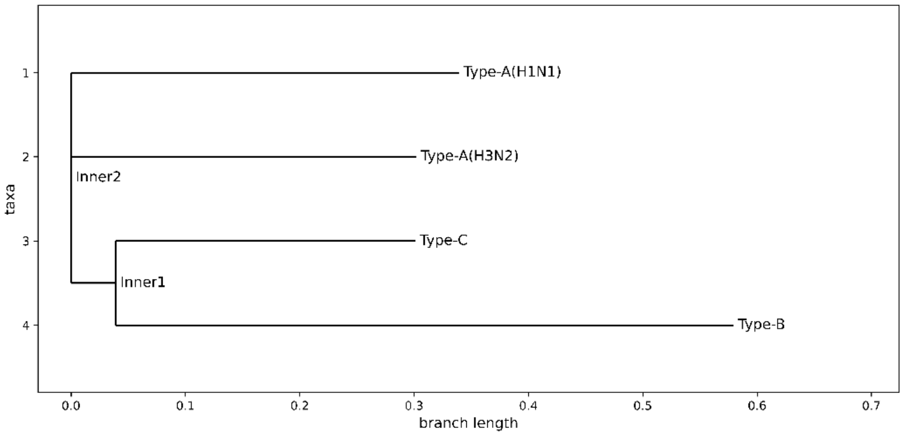


Fig 10. Phylogenetic tree of Influenza virus

I built a Phylogenetic tree following the steps below (Talevich, Invergo, Cock, & Chapman, 2012):

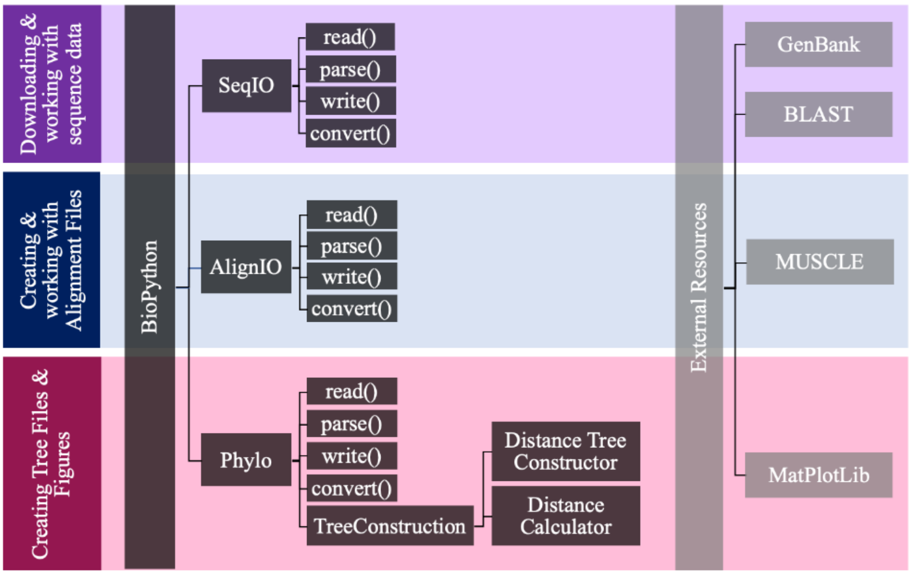
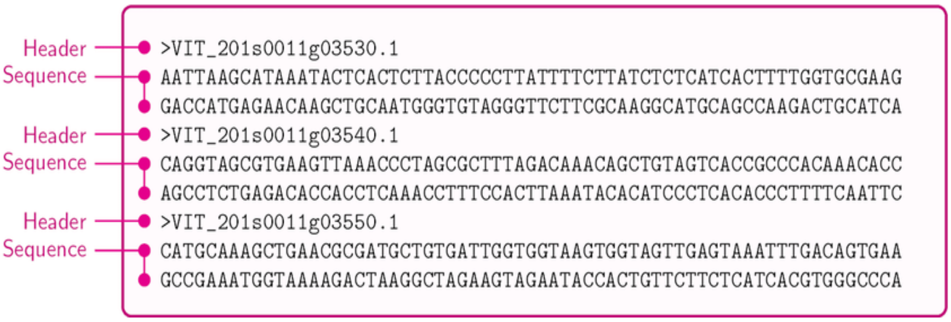


Fig 11. Overview of the steps to construct a Phylogenetic tree

1. I downloaded nucleotide sequence data in FASTA format, as shown in fig 12, from GenBank. GenBank is a federally funded database containing information on various virus strains. Then, I imported this data using the SeqIO module from BioPython, which reads and writes sequence files.


Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

Fig 12. A sample FASTA file

2. I aligned the sequences and removed the headers using the AlignIO module from BioPython and an external website called MUSCLE. Sequences are aligned such that the equivalent nucleotide in each sequence lines up to form a column, making comparing sequences easier.
3. I constructed the Phylogenetic tree using the distance matrix calculated using the “Phylo” module from BioPython. Lastly, I plotted the tree using Python’s “matplotlib” library.

After exploring the data, I constructed interactive visualisations as detailed below.

## Data Visualization

I constructed an interactive dashboard using Microsoft Power BI to visualise the data, as shown in fig 13.

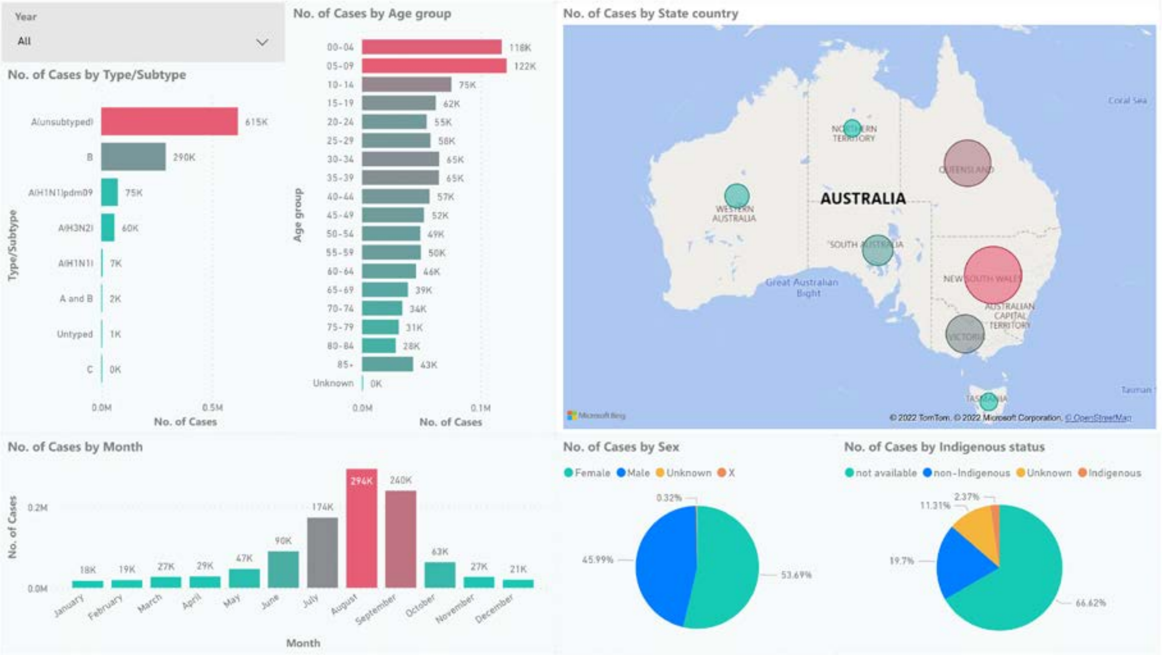


Fig 13. Influenza Dashboard for Australia (2008 to 2019)

The main features of the dashboard are:

- **Interactivity:** The dashboard is interactive, i.e., clicking on any visualisation affects the entire dashboard. So, for instance, when the bubble over Queensland is selected, the whole dashboard reflects the change, as shown in fig 14.

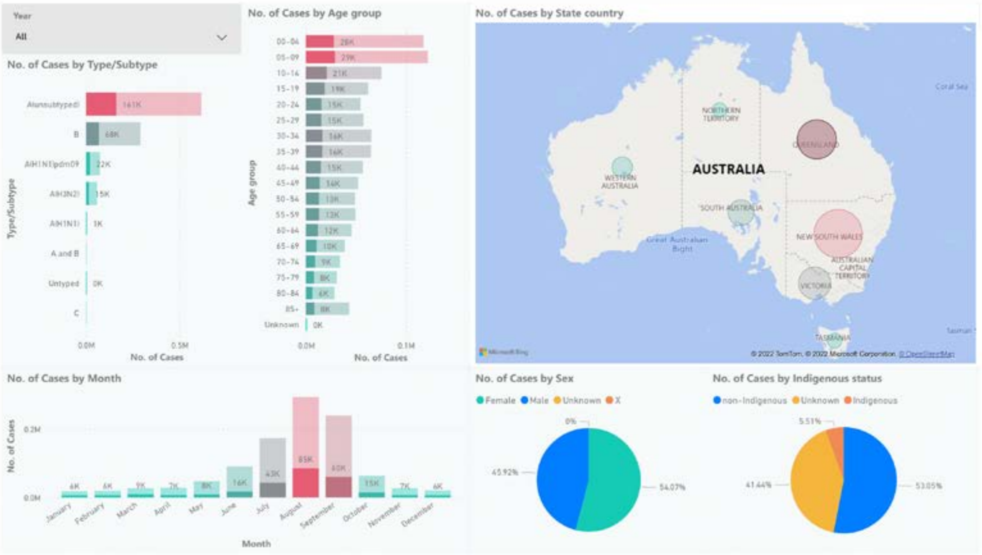


Fig 14. Demo of dashboard interactivity

- **Filtering:** Filters the entire dashboard by single or multiple years from 2008 to 2019.
- **Conditional Formatting:** As the number of cases increases, the colour of the visualisation changes from green to red. For instance, as shown in fig 15, the bar gradually turns red going towards the winter season.

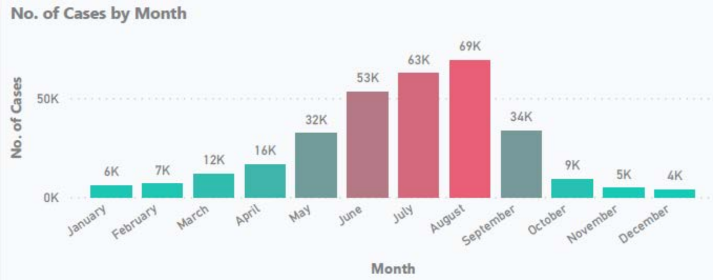


Fig 15. Demo of dashboard conditional formatting

I obtained the following insights from the dashboard:

- Influenza Type-A virus is the dominant strain in Australia, accounting for over 70% of all the cases, followed by Type-B virus, accounting for 27.3%.
- Children under the age of 10 years are the most vulnerable, accounting for over 22% of cases.
- The infections peak during winter from July to September. The highest number of infections occurred in August, accounting for 28% of all the cases. It is followed by September at 22.9%.
- From the Australian map in the dashboard, New South Wales records the highest cases, almost 35%, followed by Queensland and Victoria at 25.5% and 18.6%, respectively. The high infection rate in NSW is expected, considering its high population density making it easy for the virus to spread.
- The distribution of infections among males and females is almost equal.
- Over 75% of the data for Indigenous status is either “Unknown” or “Not available”. However, among the available data, more than 90% of the infected individuals are non-indigenous.

In the next section, I have described various forecasting methods used to predict the number of influenza cases, given historical data.

## Predictive Analysis

The ability to predict the dynamics of influenza cases is crucial for public health planning of efficient health care allocation and monitoring of the effects of policy interventions (Zhao et al., 2021). Therefore, I predicted Influenza cases for 52 weeks by experimenting with various forecasting methods like the Holt-Winters method, LSTM (Long short-term memory), Facebook Prophet, SARIMA (Seasonal Autoregressive Integrated Moving Average), and SARIMAX (SARIMA with exogenous factors) model.

Before applying predictive models to the influenza dataset, this time-series data is split into train and test data, as shown in fig 16. I split the data to evaluate the forecasting models and obtain the best performing model. Train data is the number of infections from 2008 to 2018, represented by blue colour, and the test data is the data only from 2019, represented by orange colour.

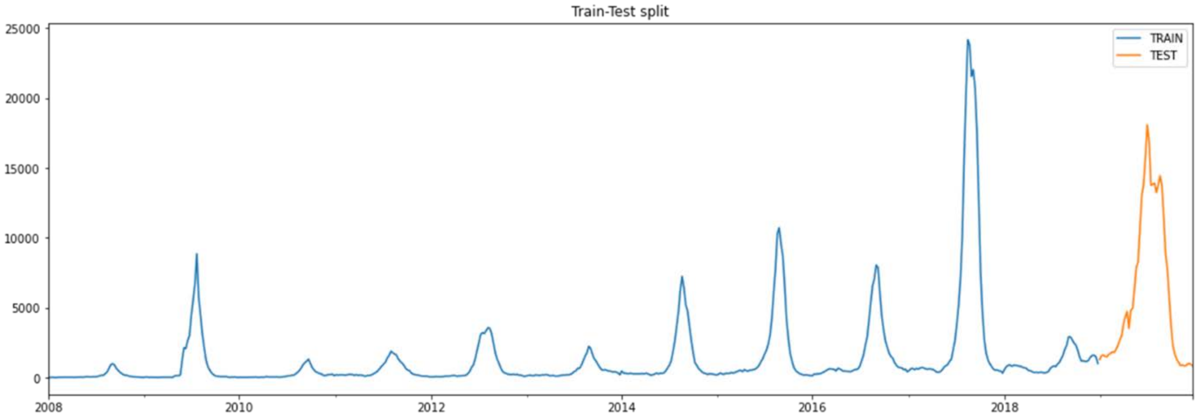


Fig 16. Train and test split

Using the predictive models below, I forecasted the influenza cases for 52 weeks (i.e., for 2019).

### *Holt-Winters Method*

The first method I tried was the Holt-Winters method since it is a simple model that is computationally inexpensive. This method is a seasonal decomposition tool that uses triple exponential smoothing to decompose the time-series data into three components, i.e., Level $l_t$, Trend $b_t$ and Seasonal $s_t$ (Holt, 2004). The corresponding smoothing parameters are α, β and γ. All these coefficients range from 0 to 1. Additionally, the seasonal period is denoted by $m$. Since the influenza dataset has weekly data and the flu has yearly seasonality, $m$ is equal to 52 weeks.

The seasonal component can be either multiplicative or additive. The multiplicative approach is preferable for influenza since the seasonal periodicity changes proportionately to the 'Level' of the time series. In contrast, the additive approach is desired when seasonal variations are reasonably stable throughout the series. The parameter α controls how much the Level component of the time series is smoothed. A low number indicates that older values in the x-direction are highly weighted. When the value is close to one, the most recent values are given greater weight.

Hence, the Holt-Winter's multiplicative model is given by,

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\
\ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}
\end{aligned}
$$

However, Holt-Winter's method performs poorly on the influenza dataset, whose predictions are almost equal to zero, as shown in fig 17. The orange line represents true labels in the test set, and the green line represents the predictions using Holt-Winter's method. This model obtained an RMSE value of 8002, and due to this poor performance, I used a more powerful deep-learning-based model, as detailed below.

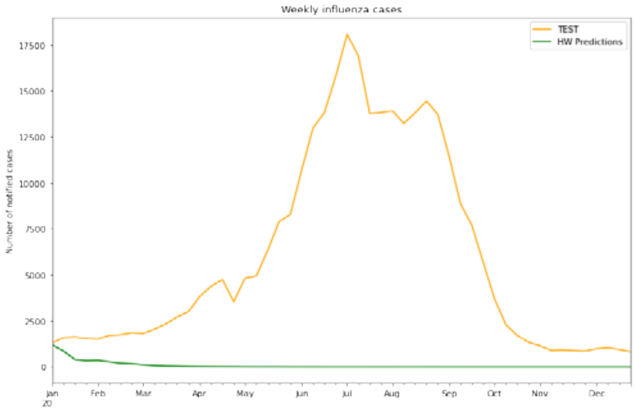


Fig 17. Holt-Winter's model Predictions

### *Long short-term memory (LSTM)*

Long Short-Term Memory (LSTM) is an artificial neural network with feedback connections. It is a type of Recurrent Neural Network (or RNN) that can learn long-term dependencies in sequential data. Hence, LSTM is suitable for language translation, time-series forecasting and speech recognition.

LSTM was proposed to solve the vanishing gradient problem (i.e., gradients tending to zero) of vanilla RNNs. LSTM solves this problem by enabling gradients to flow unmodified. However, the expanding gradient problem (i.e., gradients tending to infinity) can still affect LSTM networks.

A typical LSTM unit contains a cell, an input gate, an output gate, and a forget gate, as shown in fig 18. The three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time periods (Hochreiter & Schmidhuber, 1997).

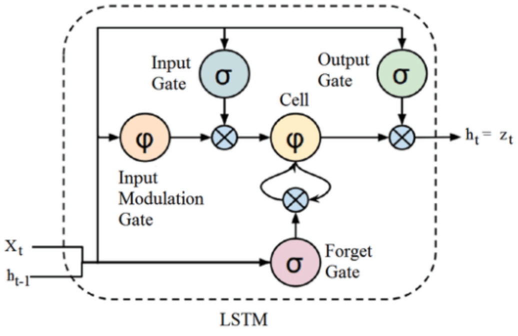


Fig 18. LSTM unit

I used Keras to build an LSTM model for influenza. I chose the following hyper-parameters:

1. ReLU (Rectified Linear Unit) is the activation function. It is given by *R(z) = max(0, z)*, as shown in the fig 19.
2. Adam is the optimiser, a blend of two gradient descent methodologies, i.e., Momentum and the Root Mean Square Propagation (RMSP).
3. Mean Squared Error (MSE) is the loss function. It is the average of the squared difference between the estimated values and the actual value.
4. The output layer is a dense layer with one neuron.

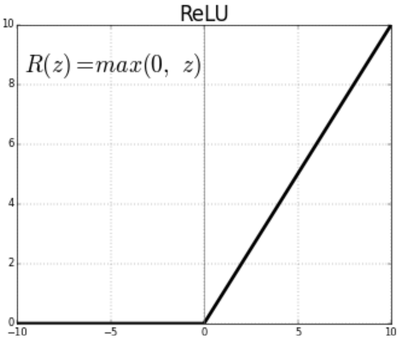


Fig 19. ReLU activation

However, LSTM performs better than Holt-Winter's method, but it still performs poorly, with an RMSE value of 6489.7 on the influenza dataset, as shown in fig 20. To improve model performance further, I used Prophet, as discussed below.

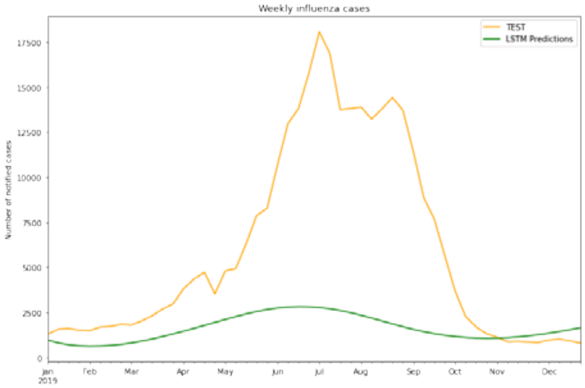


Fig 20. LSTM model predictions

### *Prophet*

Prophet is an open-source time-series forecasting method developed by Facebook. It is based on an additive model that fits non-linear trends with annual, weekly, or daily seasonality (Taylor & Letham, 2017). It also considers the impacts of public holidays. Moreover, Prophet is forgiving of missing data and trend changes, and it usually handles outliers well (Tseng & Shih, 2019). Prophet expects the input data to be in a specific format. Hence, I pre-processed the data to comprise two columns named *ds* and *y* containing the dates and the influenza cases, respectively.

Prophet decomposes the time-series data into three components, that is:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- *g(t):* piecewise linear or logistic growth curve
- *s(t):* periodic changes (daily, weekly or yearly seasonality)
- *h(t):* effects of holidays with irregular schedules
- $\epsilon_t$: error term

From fig 21, Facebook Prophet performs better than the previous two forecasting models with the RMSE value of 5342. Nevertheless, the peak of the predicted values during winter is not as prominent as the true values. To address this problem, I tried a more sophisticated forecasting method called SARIMA, which I discussed below.

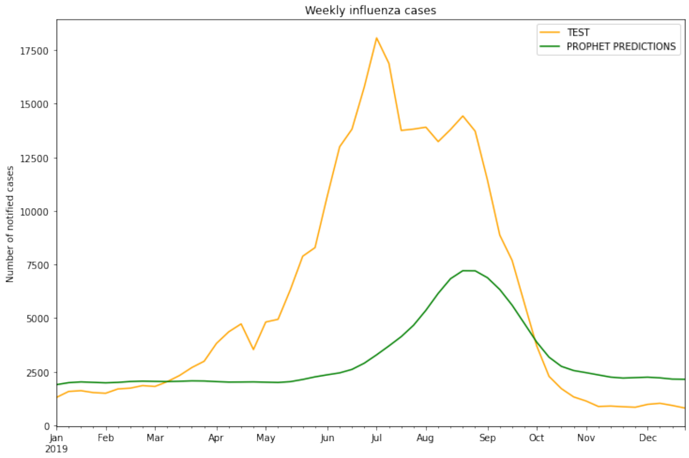


Fig 21. Facebook Prophet predictions

### *Seasonal Auto-Regressive Integrated Moving Average (SARIMA)*

SARIMA is a variation of ARIMA that enables univariate time series data with a seasonal component. The Auto-Regressive (AR) component performs regression on the variable of interest with its own lagged values. The Moving Average (MA) component is a model that exploits the relationship between the observation and residual error. This residual error is obtained from applying the moving average model to lagged data. Moreover, it allows non-stationary time series data thanks to the Differencing (I) component (Permanasari, Hidayah, & Bustoni, 2013).

Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the time series, i.e.,

$$SARIMA\underbrace{(p,d,q)}_{non-seasonal}\underbrace{(P,D,Q)_m}_{seasonal}$$

Hence, SARIMA has seven hyper-parameters to configure. There are three parameters for AR, I, and MA order for each trend (p, d, q) and seasonal (P, D, Q) components of the time series. The last parameter is ‘m’, which is the seasonal periodicity. It is set to 52 weeks for this influenza dataset.

I used the auto-ARIMA function in Python from the ‘pmdarima’ package to determine these hyper-parameters. This function provides the most optimal SARIMA model when the seasonal component is enabled. Hence, after running the dataset through this function, the optimal parameters obtained are:

$$SARIMA(p,d,q)(P,D,Q)_m = SARIMA(2,1,4)(2,0,0)_{52}$$

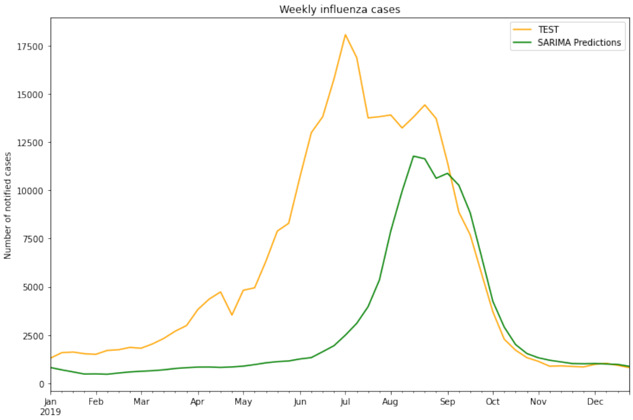


Fig 22. SARIMA model predictions

From fig 22, I observed that this is the best performing forecasting model so far, with the lowest RMSE value of 5303. The predicted values show a peak during winter that roughly follows the true values. However, since I observed a high negative correlation between influenza cases and the temperature in the data exploration stage, I tried incorporating this variable in the model as explained below.

*Seasonal Auto-Regressive Integrated Moving Average with exogenous factors (SARIMAX)*

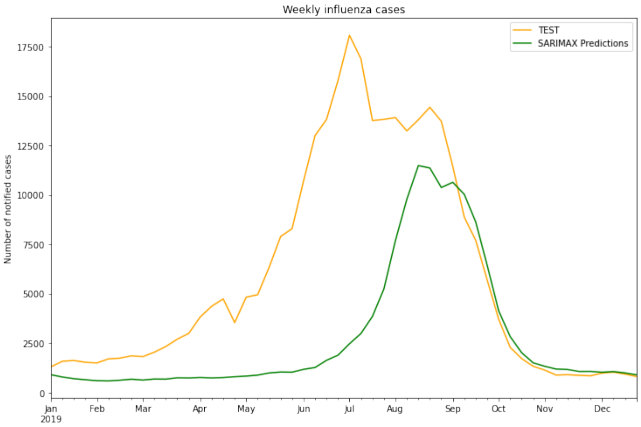


Fig 23. SARIMAX model predictions

SARIMAX is an extension of the SARIMA model but with the capability to handle exogenous variables. In this case, I chose the weekly minimum temperature in Brisbane, obtained from Australia’s Bureau of Meteorology, as my exogenous element. I used the same hyper-parameters of the SARIMA model and got the predictions as shown in fig 23. The chart looks very similar to the SARIMA model predictions with an RMSE value of 5350. I considered various factors to find the best performing forecasting model, as detailed below.

**Model Selection**

Root Mean Squared Error (RMSE) is my preferred evaluation metric to determine the best forecasting model. The lower the RMSE value, the better the model performance.

RMSE is given by,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

where,

- *Predicted$_i$* is the predicted value for the ith observation.
- *Actual$_i$* is the observed (or true) value for the ith observation.
- *N* is the total number of observations.

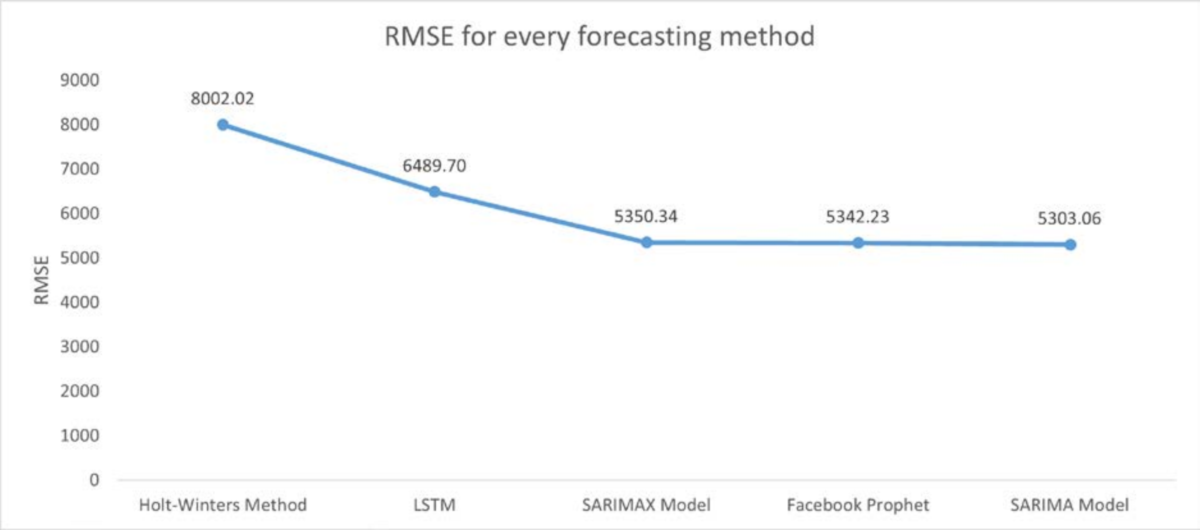


Fig 24. RMSE for every forecasting method

However, from fig 24, I observed that the SARIMAX, Prophet, and SARIMA models have almost the same RMSE value.

Finally, I selected SARIMA as the preferred model because:

- SARIMA has the lowest RMSE value of 5303.
- In Prophet, the peak of the predicted values during winter is not as prominent as the true values.
- SARIMAX requires an additional variable to enable prediction, i.e., temperature as its exogenous variable. Unfortunately, the actual (or true) temperature values for the future cannot be known before. Therefore, the only way to use SARIMAX is to use predicted temperature values that may not be accurate.

## Epidemic Modelling

Epidemics describe the process by which diseases spread. This process consists of a pathogen, a population of hosts, and a spreading mechanism. One of the two methodologies listed below can be utilised to study epidemics:

1. **Using Contact Network:**
    - This technique examines how hosts communicate with one another and develops strategies to describe the cause of the epidemics.
    - A contact network is a graph in which nodes represent the hosts, and edges represent the interactions between these hosts. For instance, hosts (nodes) that breathe the same air in an influenza contact network are connected.
2. **Fully mixed method:**
    - Analyse solely the rates at which hosts become infected, recover, etc. and avoid considering network information.

I implemented a Fully mixed epidemic model that assumes that there is no contact network information and the process by which hosts get infected is unknown.

Initially, I considered the following Fully mixed epidemic models:

1. **SI:** It has susceptible and infected individuals. Once infected, they cannot be cured.
2. **SIR:** SI + individuals can get recovered. Once recovered, they cannot be infected again.
3. **SIS:** SI + individuals can get recovered. And recovered individuals can be infected again.
4. **SIRS:** SI + individuals can get recovered. For some time, recovered individuals will not be infected again. However, after that time, they can be infected again.
5. **SEIR:** SI + individuals can get recovered + individuals can be exposed to the virus but cannot spread it.

Finally, I chose the SEIR epidemic model since it accounts for all the characteristics of the influenza epidemic. SEIR stands for Susceptible-Exposed-Infectious-Recovered, which means:

- Susceptible (S): People in this group are not infected yet. Here, it is the entire Australian population since every individual of all age groups and gender is susceptible to the flu. As per the Australian Bureau of Statistics, the Australian population is 25,750,198 ("National, state and territory population," 2022).
- Exposed (E): The individuals in this group are infected but not yet infectious like people in the incubation period. Initially, it is the reciprocal of the number of susceptible people.
- Infectious (I): The individuals who infect other people. Here, the initial value is zero.
- Recovered (R): People in this group have recovered from the flu. Here, the initial value is zero.

SEIR model is built based on the below equations:

$$\frac{dS}{dt} = -(1-u)\,\beta SI$$

$$\frac{dE}{dt} = (1-u)\,\beta SI - \alpha E$$

$$\frac{dI}{dt} = \alpha E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

where,

- $\alpha$ is the reciprocal of the incubation period. The time between being exposed to an illness and experiencing the first symptoms is known as the incubation period. Symptoms of the flu generally present 1 to 3 days after infection. Even before they show symptoms, people might be contagious to others ("Influenza (the flu)," 2022). Hence, I have taken $\alpha = 1 / 2 = 0.5$ assuming incubation period for influenza to be 2 days.
- $\gamma$ is the reciprocal of the infective period. Patients can shed influenza virus for up to 24 hours (1 day) before symptoms appear and up to 7 days after symptoms appear. In adults, viral shedding peaks in the first 1 to 2 days following symptom onset. Fever is not present in all cases of influenza infection, but when a person has a fever, it is linked to viral shedding ("Control Guidelines - Influenza control guideline," 2018). Therefore, I have taken $\gamma = 1 / 3 = 0.33$ assuming the infective period for influenza to be three days.
- $\beta$ is the product of $\gamma$ and reproduction rate $R_0$. The virus's transmissibility between people will impact the virus's spread around the globe and in Australia. The reproduction rate ($R_0$) of the influenza virus is typically between 1.2 and 2.5 ("Australian Health Management Plan for Pandemic Influenza," 2019). Hence, I have taken $\beta = 0.33 * 1.85 = 0.61$ assuming $R_0$ to be 1.85.

I simulated the influenza epidemic that visualises the impact of preventive measures like wearing a mask and incorporating social distancing to help reduce the spread of infection. This simulation is done by changing the value of 'u' accordingly.

For instance:

1. Let u = 0 when people follow no preventive measures, resulting in fig 25. In 2017–18, Australia had 3.9 beds per 1,000 population in public and private hospitals, equating to 0.39% of the total population (“Hospital resources 2017–18: Australian Hospital Statistics, hospitals and average available beds,” 2019). A blue horizontal straight line represents the healthcare system’s capacity. As per the Centres for Disease Control and Prevention (CDC), 7.14% of the infected people get hospitalised (“Frequently asked questions about estimated flu burden,” 2021). As I observed from the plot below, if people take no preventive measures, the healthcare system will be overwhelmed, and the people who need critical care might not get the proper treatment they deserve. Hence, it is essential to flatten the curve, as shown in fig 26.

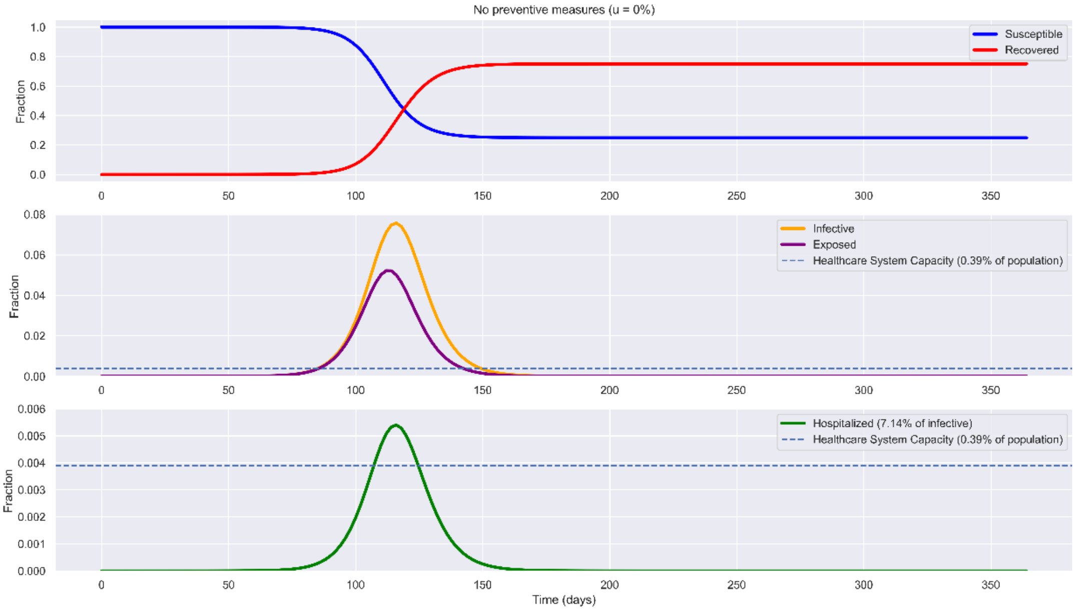
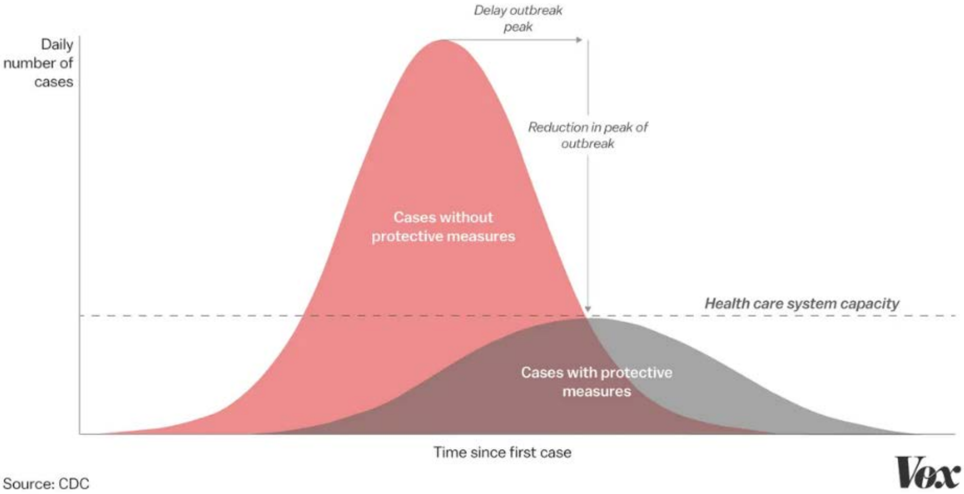


Fig 25. No preventive measures (u = 0)



Source: CDC

Christina Animashaun/Vox

Fig 26. Flatten the curve

2. Let u = 0.1 (or 10%) while wearing a mask. Consequently, from fig 27, I observed that the rate of hospitalisation decreases and is almost close to filling up the healthcare capacity. Hence, although wearing a mask is effective against the spread of infections, it is not enough.

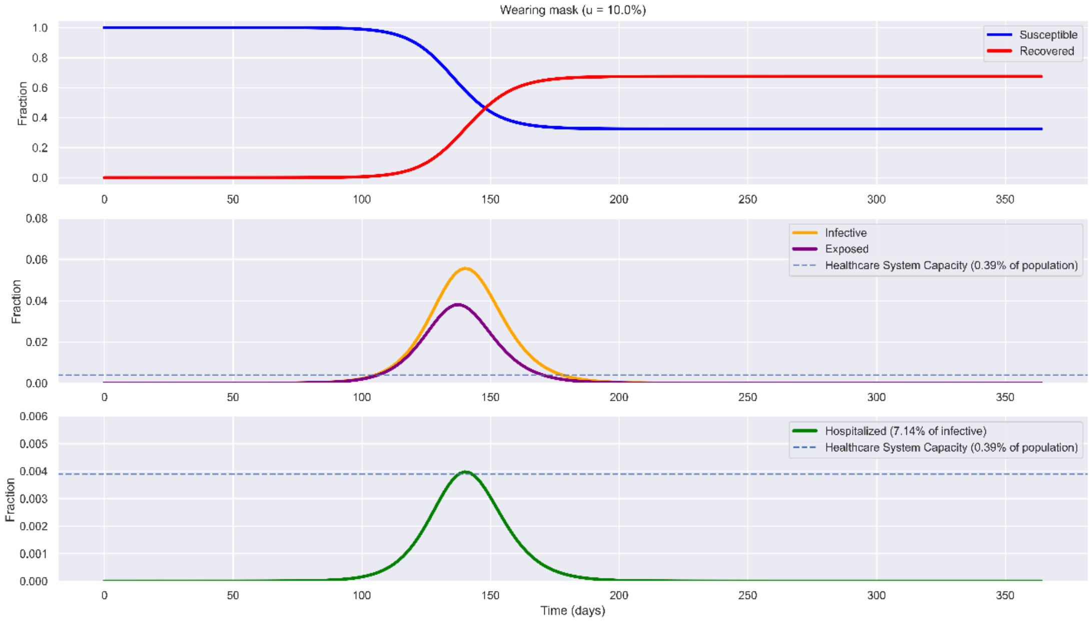


Fig 27. Wearing a mask (u = 0.1)

3. Let u = 0.2 (or 20%) while wearing a mask and maintaining social distancing. As a result, from fig 28, I observed that the number of hospitalisations drastically reduces and goes below the healthcare capacity line.

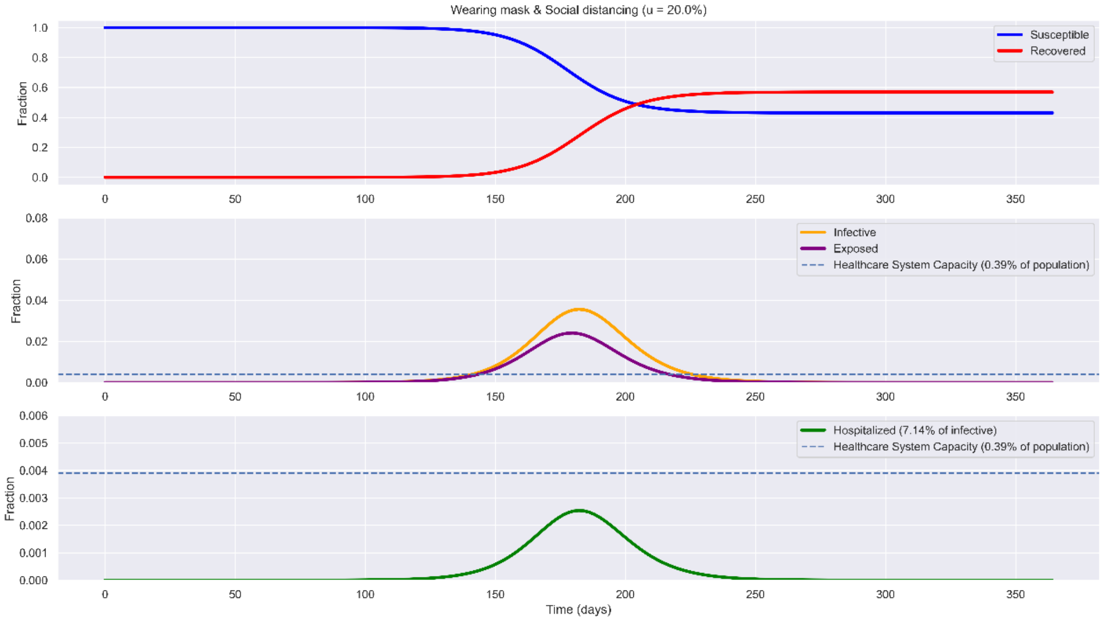


Fig 28. Wearing a mask and maintaining social distance (u = 0.2)

Therefore, if nobody follows preventive measures like wearing a mask and maintaining social distance, getting infected with the flu is inevitable. However, suppose the number of individuals who get sick is spread out. In that case, patients will be able to receive the treatment they require over time since hospitals and other medical resources will not be overwhelmed.

## Limitation

The only limitation of this project is that the dashboard uses historical data. Therefore, for surveillance to be effective, real-time data analysis is essential. For example, the Australian government could create an API that streams national public health surveillance data accessible by common people. Additionally, using Apache Beam and Python, real-time analysis can be achieved by constructing data pipelines to extract, manipulate, and analyse streaming data.

## Future Work

This project could be extended by:

1. Influenza has been linked to many subsequent bacterial infections like Streptococcus pneumoniae, the most common infection that causes secondary bacterial pneumonia. Influenza has a role in developing Invasive Pneumococcal Disease (IPD) (Grabowska, Högberg, Penttinen, Svensson, & Ekdahl, 2006). Therefore, efforts should be made to build a similar reporting system to track IPD.
2. The naive Ordinary Least Squares (OLS) model can be used to determine the relationship between the number of weekly infections, weekly temperature, and weekly rainfall.
3. The dashboard created could be hosted on a live website available to view publicly. The website can be built using HTML, CSS, JavaScript, and Flask.
4. Genetic sequences I used to build a Phylogenetics tree can also be used to detect genetic alterations that impact the virus's characteristics. For instance, determining the modifications linked to influenza viruses spreading more quickly, causing more severe illness, or acquiring antiviral medication resistance.
5. Data from external sources can be used to determine the influenza vaccine's effectiveness against getting infected or reinfected. This data can also be used to track the vaccination status of Australian citizens in real-time.
6. Isolate and research novel influenza A virus to assess the risks and develop tests to detect it, determine whether existing influenza antiviral medications would be effective against it, and take early measures to build vaccinations to respond to the next pandemic.

## Conclusion

Australia and the rest of the world must remain vigilant to protect themselves against seasonal influenza and from novel influenza A viruses, which account for over 70% of Australia’s flu cases. Epidemiologic and virologic influenza surveillance is the cornerstone of influenza preparedness and response to influenza viruses. Unfortunately, the federal government has put less than ideal effort into creating a robust reporting system to monitor public health. The consequences of not taking effective actions during the initial stages of an epidemic will be disastrous. It will burden the health care system, have adverse economic impacts and the most horrifying instance of all is the pain of losing a loved one.

This project discussed how to build a sophisticated yet simple reporting system for the influenza epidemic by accomplishing the following:

1. I built an interactive dashboard that enables public health authorities to determine the intensity of influenza activity throughout Australia, categorised by attributes like age, gender, indigenous status, etc. With the insights obtained from this dashboard, authorities can take necessary action, like declaring a curfew in influenza hotspots.
2. I constructed a Phylogenetic tree that helps medical researchers organise biological information and communicate hypothesised evolutionary links between influenza virus strains circulating in Australia. Moreover, it gives a visual overview of what influenza viruses are circulating.
3. I built a sophisticated epidemic model, SEIR, that measures influenza’s impact on hospitalisations. This model will enable authorities to determine if a mandate is necessary to make people follow certain preventive measures like wearing a mask.
4. I forecasted influenza cases that will enable public health planning of efficient health care allocation and monitoring effects of policy intervention, like increasing the number of hospital beds or intensive care units during winter, determining when to administer flu shots, etc.

Moreover, this project also gives a general framework to build reporting systems for any infectious diseases using its surveillance data. Prevention is better than cure and hence preventing an epidemic from becoming a pandemic is imperative in saving people’s life and livelihood.

# References

Attwooll, J. (2021, August 16). Flu-zero: More than a year since Australia's last flu death. Retrieved June 08, 2022, from https://www1.racgp.org.au/newsgp/clinical/australia-records-zero-flu-deaths-over-past-12-mon

Australian Health Management Plan for Pandemic Influenza. (2019, August). Retrieved June 7, 2022, from https://www1.health.gov.au/internet/main/publishing.nsf/Content/519F9392797E2DDCCA257D47001B9948/$File/w-AHMPPI-2019.PDF

Carneiro, H., & Mylonakis, E. (2009). Google trends: A Web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557-1564. doi:10.1086/630200

Chriscaden, K. (2020, October 13). Impact of covid-19 on people's livelihoods, their health and our Food Systems. Retrieved May 02, 2022, from https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems

Control Guidelines - Influenza control guideline. (2018, July 2). Retrieved June 7, 2022, from https://www.health.nsw.gov.au/Infectious/controlguideline/Pages/influenza.aspx

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HEALTHMAP: Global infectious disease monitoring through automated classification and visualisation of internet media reports. *Journal of the American Medical Informatics Association, 15*(2), 150-157. doi:10.1197/jamia.m2544

Frequently asked questions about estimated flu burden. (2021, October 21). Retrieved June 7, 2022, from https://www.cdc.gov/flu/about/burden/faq.htm

Gopinath, S. C., Tang, T., Chen, Y., Citartan, M., Tominaga, J., & Lakshmipriya, T. (2014). Sensing strategies for influenza surveillance. *Biosensors and Bioelectronics, 61*, 357-369. doi:10.1016/j.bios.2014.05.024

Grabowska, K., Högberg, L., Penttinen, P., Svensson, Å, & Ekdahl, K. (2006). Occurrence of invasive pneumococcal disease and number of excess cases due to influenza. *BMC Infectious Diseases, 6*(1). doi:10.1186/1471-2334-6-58

He, S., Peng, Y., & Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics, 101*(3), 1667-1680. doi:10.1007/s11071-020-05743-y

Hochreiter, S., & Schmidhuber, J. (1997, November 15). Long short-term memory. Retrieved June 6, 2022, from https://doi.org/10.1162/neco.1997.9.8.1735

Holt, C. (2004, January 28). Forecasting seasonals and trends by exponentially weighted moving averages. Retrieved June 6, 2022, from https://doi.org/10.1016/j.ijforecast.2003.09.015

Hospital resources 2017–18: Australian Hospital Statistics, hospitals and average available beds. (2019, June 26). Retrieved June 7, 2022, from https://www.aihw.gov.au/reports/hospitals/hospital-resources-2017-18-ahs/contents/hospitals-and-average-available-beds

Influenza (the flu). (2022, May 27). Retrieved June 7, 2022, from http://conditions.health.qld.gov.au/HealthCondition/condition/14/217/82/influenza-the-flu

Ludmir, E. B., & Enquist, L. W. (2009). Viral genomes are part of the phylogenetic tree of life. *Nature Reviews Microbiology, 7*(8), 615-615. doi:10.1038/nrmicro2108-c4

National, state and territory population. (2022, March 17). Retrieved June 7, 2022, from https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release

Permanasari, A. E., Hidayah, I., & Bustoni, I. A. (2013, October). SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. Retrieved June 7, 2022, from https://ieeexplore.ieee.org/document/6676239/

Poon, A. F., Walker, L. W., Murray, H., McCloskey, R. M., Harrigan, P. R., & Liang, R. H. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS ONE, 8*(11). doi:10.1371/journal.pone.0078122

Talevich, E., Invergo, B. M., Cock, P. J., & Chapman, B. A. (2012). Bio.phylo: A unified toolkit for processing, analysing and visualising phylogenetic trees in Biopython. *BMC Bioinformatics, 13*(1). doi:10.1186/1471-2105-13-209

Taylor, S., & Letham, B. (2017, September 27). Forecasting at scale. Retrieved June 6, 2022, from https://doi.org/10.7287/peerj.preprints.3190v2

Tseng, Y., & Shih, Y. (2019). Developing epidemic forecasting models to assist disease surveillance for influenza with Electronic Health Records. *International Journal of Computers and Applications, 42*(6), 616-621. doi:10.1080/1206212x.2019.1633762

Types of influenza viruses. (2021, November 02). Retrieved June 8, 2022, from https://www.cdc.gov/flu/about/viruses/types.htm

Zhao, H., Merchant, N., McNulty, A., Radcliff, T., Cote, M., Fischer, R., . . . Ory, M. (2021, April 14). Covid-19: Short term prediction model using daily incidence data. Retrieved June 6, 2022, from https://doi.org/10.1371/journal.pone.0250110