Influenza Time Series and Forecasting

Influenza viruses pose a significant global public health threat, impacting people of all ages. Each year, they cause hundreds of thousands of deaths worldwide. The historical pandemics of influenza have had far-reaching social consequences and placed substantial burdens on society (WHO, 2018).

Time series analysis is an effective tool for predicting disease incidence (Majumder, 2017). In this study, I employ time series analysis to examine the spread and incidence of positive Influenza virus cases from 2010 to more recent data in 2021. By constructing a time series model for year prediction, this analysis aims to enhance our understanding of Influenza outbreaks and contribute to the prevention efforts focused on public health.

Thus, I am going to try answer the following research question:

“What will be the monthly reports of Influenza be for 2022, based on historical data?"

To answer this question comprehensively, I will consider these further questions:

- *What’s the trend of positive Influenza report prevalence?*
- *Is there coherent seasonality?*
- *Are there unexpected sharp spikes or drops, if so, why?*
- *Can we forecast accurate entries in near future to help support public planning and resource allocation?*

Dataset and the Forecast Importance

Controlling and preventing the spread of the influenza virus is crucial to avoid another pandemic. Accurate forecasting plays a vital role in predicting the timing and severity of influenza outbreaks. Early detection allows for effective intervention strategies, such as resource allocation, public awareness campaigns, and prompt medical response. By minimizing the spread of the virus, accurate forecast models can reduce illness and mortality, particularly among vulnerable populations like the elderly, pregnant women, and young children.

To ensure the reliability of the data used in this analysis, the Influence dataset was sourced from the reputable website of the U.S. Centers for Disease Control and Prevention (CDC, 2019). The CDC's stringent protocols and standardized procedures followed by their specialized epidemiologists ensure the accuracy and reliability of the data. Relying on such trustworthy organizations facilitates a better understanding and prediction of influenza outbreaks.

Understanding Influenza Dataset

I conducted EDA to gain insights into the influenza data (**Table 1**), which helped me make decisions about how to pre-process and transform the data. No missing data was present.

Table 1. Summary of the datasets columns

Column Name	Key Information
Date	Weekly information from 2011-2021, presented as date type format.
Total A	The total number of positive tests tested for strain A of the influenza virus
Total B	The total number of positive tests tested for strain A of the influenza virus
Total # Tested	The total number of people tested for either strain of the influenza virus
% Positive	Percentage of results of patients tested positively for either strain A/B of influenza

To forecast the time dependent variable 'Total # Tests', I dropped other unnecessary columns. To understand how the total number of Influenza tests relates to time, I plotted the raw time series (**Figure 1**) this helped me visualise the overall pattern, trend, and seasonality over time.

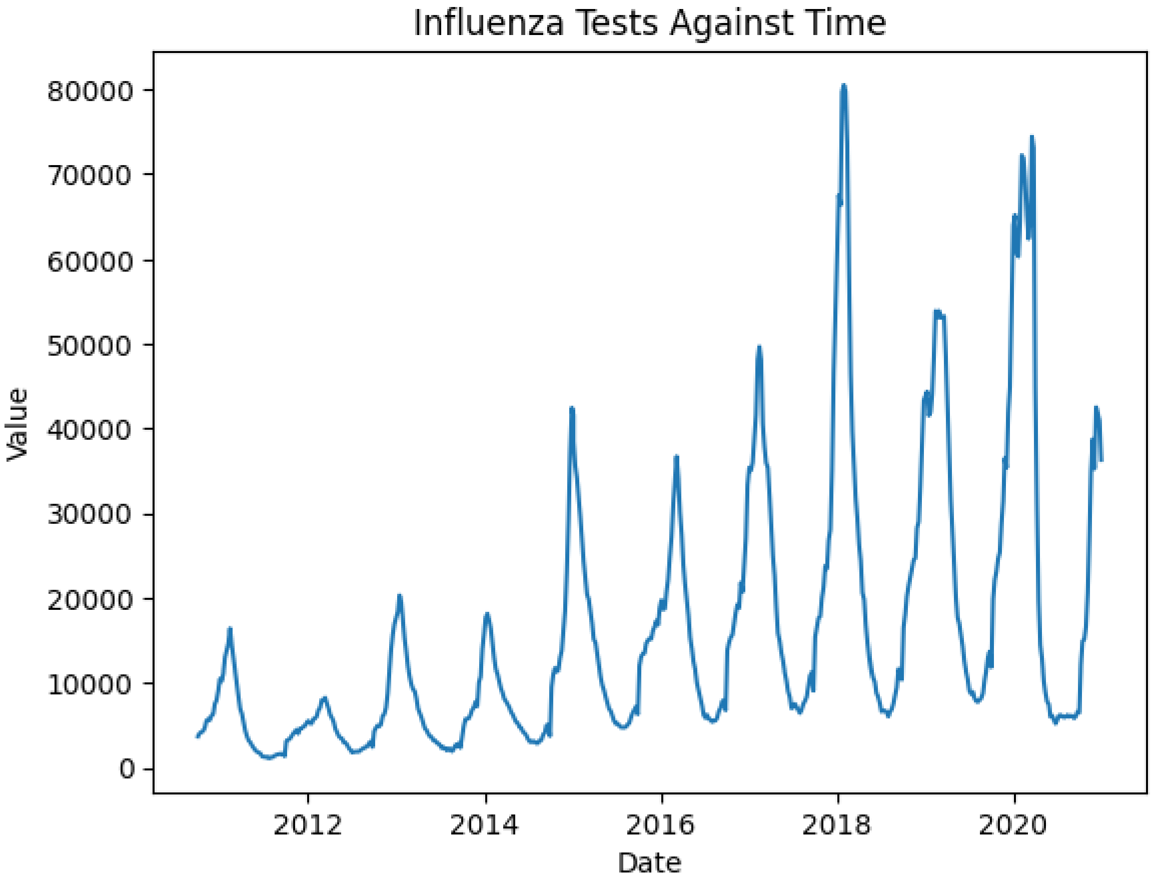

Figure 1. Time series plot showing the percentage of positive influenza reports from 2011-2021.

It is evident that there is seasonality in the Influenza test dataset with a repeating nature of seasonal patterns, representing a multiplicative model with increasing width and height of the peaks. The influenza incidence peaks correspond to winter months, including seasonal exchange (prominent in years 2015, 2018 and 2020). Thus, we can infer from a visual inspection that the dataset is not stationary.

To produce an interpretable quantity for the influenza data, I used the sum as an aggregation function to combine my original data into the new time intervals to identify the most meaningful way of analysing the data (**Figure 2**).

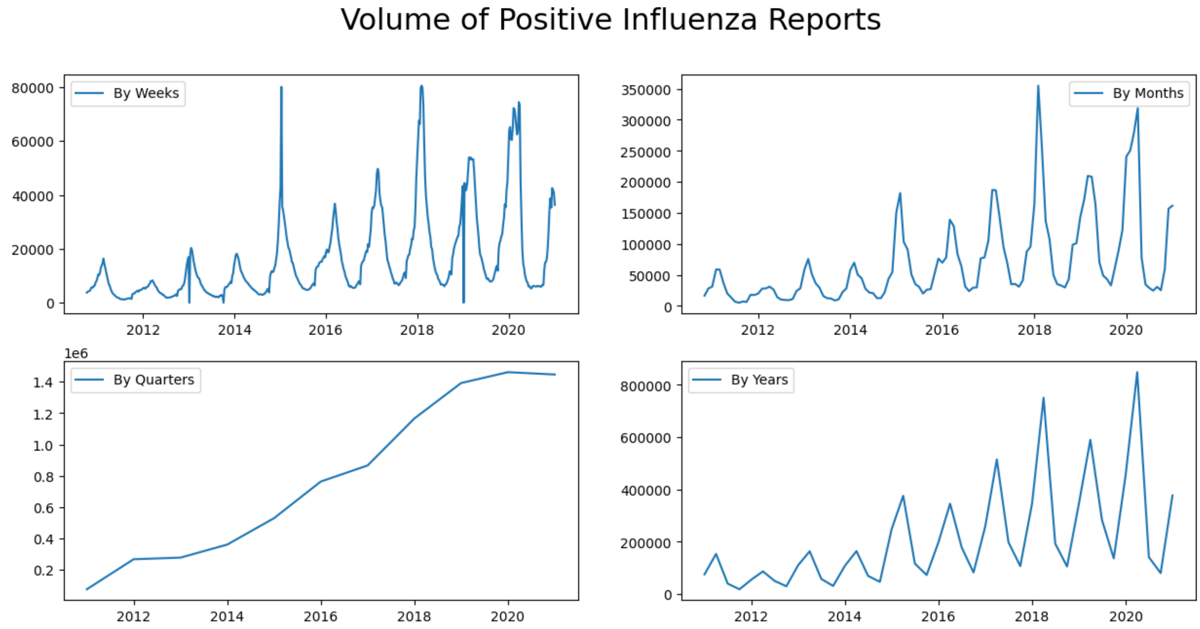

Figure 2. The volume of influenza reports, by different aggregations.

The figure above shows how the data varies with different time aggregations. Whilst majority of these show clear seasonality, I will be using monthly influenza data. As observed, the monthly data has reduced noise and short-term fluctuations in comparison to weekly data, making it easier to identify long-term patterns and trends, providing a more stable representation of underlying behaviour of the time series. Further, it considers seasonal effects, facilitating a better comparison with other monthly data set.

To provide a quantifiable overview of central tendency, variability, and spread of Influenza data, I calculated summary statistics, including the mean, standard deviation, and variance (**Table 2**).

Table 2. Summary statistics of original time series compared to boxcox transformed data.

	Original Time Series	**Transformed Time Series**
Mean	69951.89	0.0064
Standard Deviation	69720.86	0.62
Variance	258979309.73	0.38

The extremely large standard deviation, variance and mean of the data suggests heteroscedasticity with major abnormality within the data. However, this is just a snapshot of the overall statistics displayed by the series. To obtain visual information on how this varies with time, and to gain insights into data's underlying distribution, I plotted a boxplot, histogram and moving average alongside standard deviation (**Figure 3**).

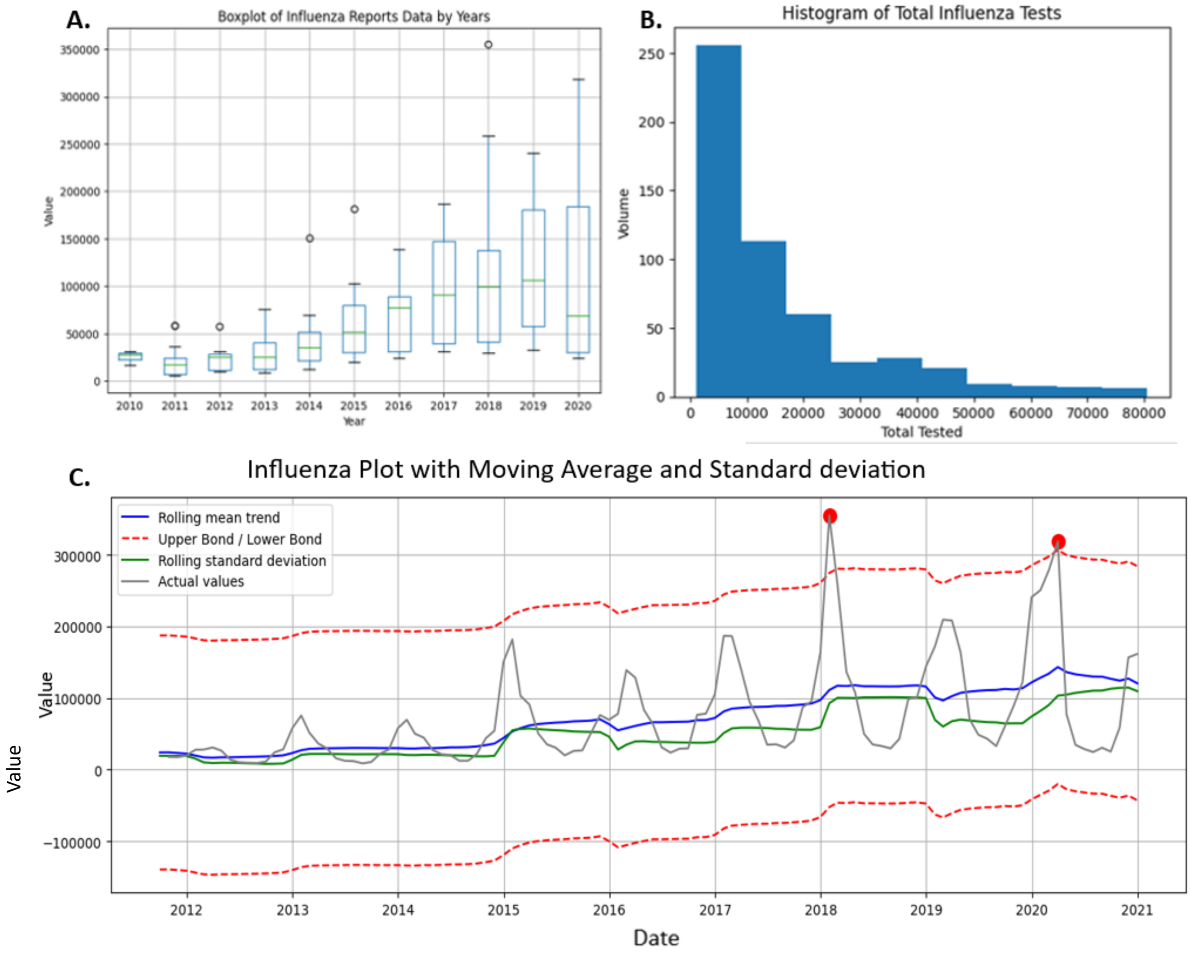

Figure 3. Graphs showing Influenza dataset distribution. A. Boxplot showing the spread of total Influenza tests annually. B. Histogram showing the spread of total influenza tests. C. Rolling mean plot, moving average window size = 12.

As observed, the boxplot (**Figure 3a**) alongside the moving average with anomaly detection (**Figure 3c**) helped reveal underlying trends and identify outliers within the dataset. The fluctuations in distribution for different years exhibit some major shifts in the quartiles, suggesting a change in trends. As the years increased, as did the distribution in the data, suggesting that more people are getting tested for influenza annually, maybe due to individuals being more educated on the major effects of the virus. However, this didn't have large effects on outlier count, suggesting that the data is rather compact and doesn't deviate much.

The histogram (**Figure 3b.**) visualises the positive skew and asymmetry within the data, proving that the variability and central tendency of the data is abnormal. The rolling mean (**Figure 2c.** blue line) and the variation in standard deviation (green line) evidently differ with time, graphically demonstrating heteroscedasticity though crests-troughs. Further demonstrating changes in the central tendency and variability over time.

<u>Decomposition</u>

I performed decomposition via the robust method of Seasonal Decomposition of Time Series by Loess (STL) (www.statsmodels.org, n.d.) (**Figure 4**) to separate the influenza time series into its different components and help me understand how each unique aspects contribute to the data's overall pattern identified previously, thereby aid with forecasting.

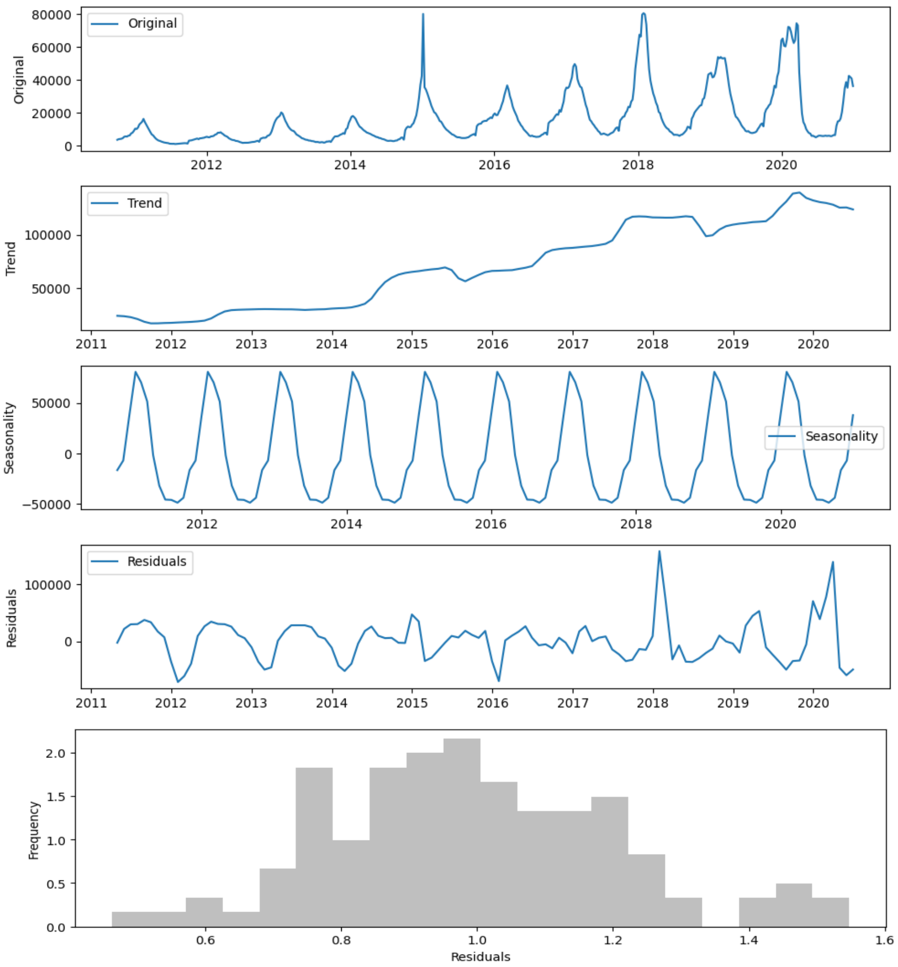

Figure 4. Decomposition of the Influenza Data into Trend, Seasonality, Residuals and Residuals distribution via a Histogram.

Evidently, the data demonstrates a non-linear uptrend with a clear-cut annual seasonality (lag = 12). The residuals distribution displayed by the histogram (**Figure 4**) follow a non-gaussian distribution. The residuals were further analysed by the Ljung-Box test (Stephanie, 2018), which assessed the presence of autocorrelation. The resulting p-value of **0.04** (critical value of 0.05), signifies that there is enough evidence to reject the null hypothesis and conclude that the residuals are not independently distributed.

Test Stationarity

To test whether the influenza dataset has a unit root, and further quantify evidence of non-stationarity, I employed the augmented Dickey-Fuller (ADF) Test (Verma, 2021).

Null hypothesis: *The time series contains a unit root and is non-stationary*

Results: ADF Statistic: **-0.83** p-value: **0.81**

Since the p-value of **0.81** is greater than the critical value of **0.05**, we do not have sufficient evidence to reject the null hypothesis, thus can statistically conclude that the series is non-stationary. This is further supported by a significant p-value of 0.02 in Kwiatkowski–Phillips–Schmidt–Shin (KPSS) (Stephanie, 2016), providing sufficient evidence to accept alternate hypothesis that the process is non-stationary.

Transform dataset to Stationarity

Now that non-stationary behaviour has been identified as present in the dataset, I will remove them from the series to transform into stationary via seasonal differencing to attempt to nullify the effects of statistical parameters on the forecast.

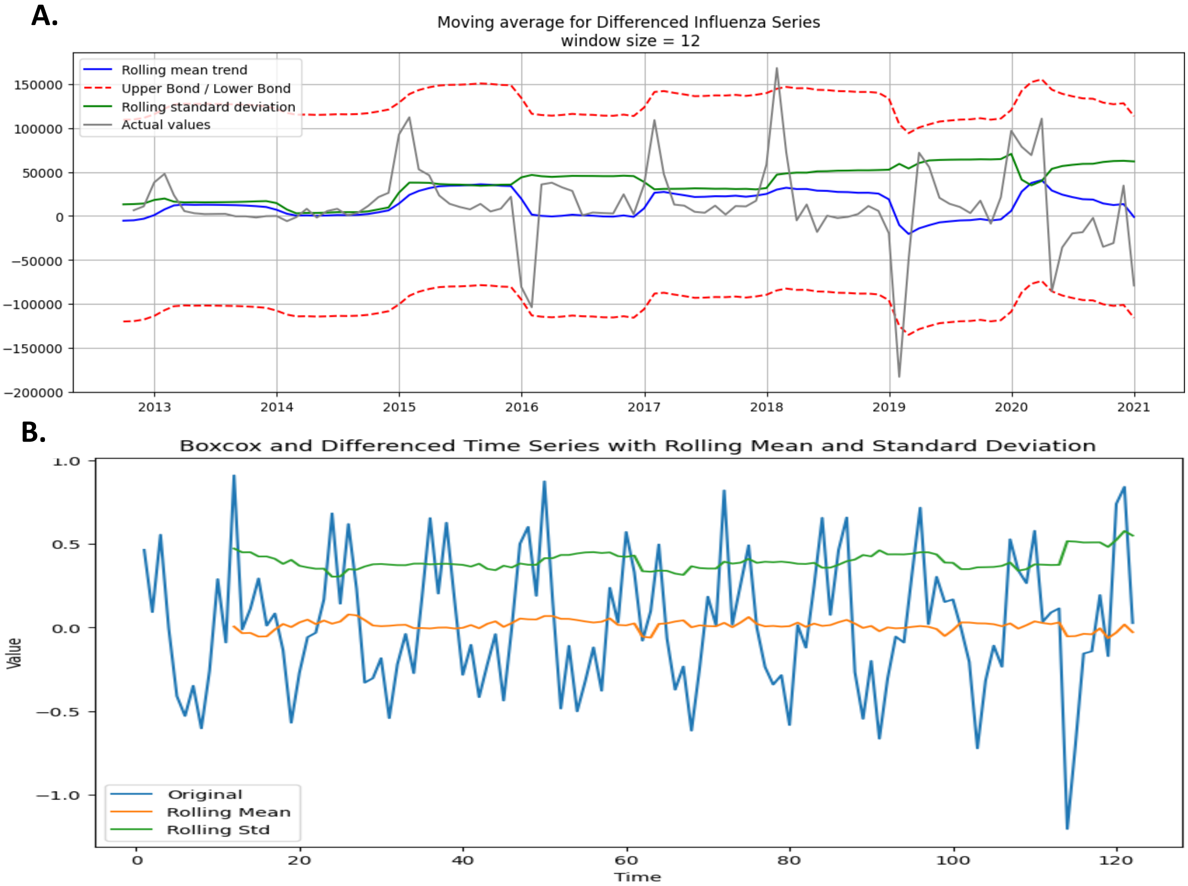

Figure 5. Moving Average for Differenced Influenza Series (A) vs Boxcox Transformed and Differenced Influenza Data (B).

However, there is still evidence that the data exhibits heteroscedasticity (**Figure 5a**). The dynamic standard deviation violates one of the assumptions of many statistical models. To stabilise the mean and variance I transformed the data.

I identified box-cox as the best transformation, yielding the smallest p-value of **2.55e-11** in the ADF test, indicating strong evidence against the presence of a unit root and stationarity. Thus, there is sufficient evidence to reject the null hypothesis and accept that the time series is stationary.

With differenced and transformed data, the rolling mean is very close to 0 (**Figure 5b**). The rolling standard deviation is close to 0.05, with some crests-and-troughs present. Overall, this suggests that the time series is stationary. This coupled with a quantifiable reduction in standard deviation and variation (**Table 2**) shows that seasonal adjustment and boxcox transformation has been effective.

Model Selection

Now that the Influenza series is stationary, homoscedastic and it fits many statistical models assumptions of stationarity, where the mean, variance, and autocorrelation structure do not change over time. This includes Auto Regressive Integrated Moving Average (ARIMA) model (Malik, 2018) and seasonal ARIMA (SARIMA) (Asamoah-Boaheng, 2014). Thus, I will estimate its parameters via autocorrelation analysis (**Figure 6**) and maximum likelihood estimation (MLE).

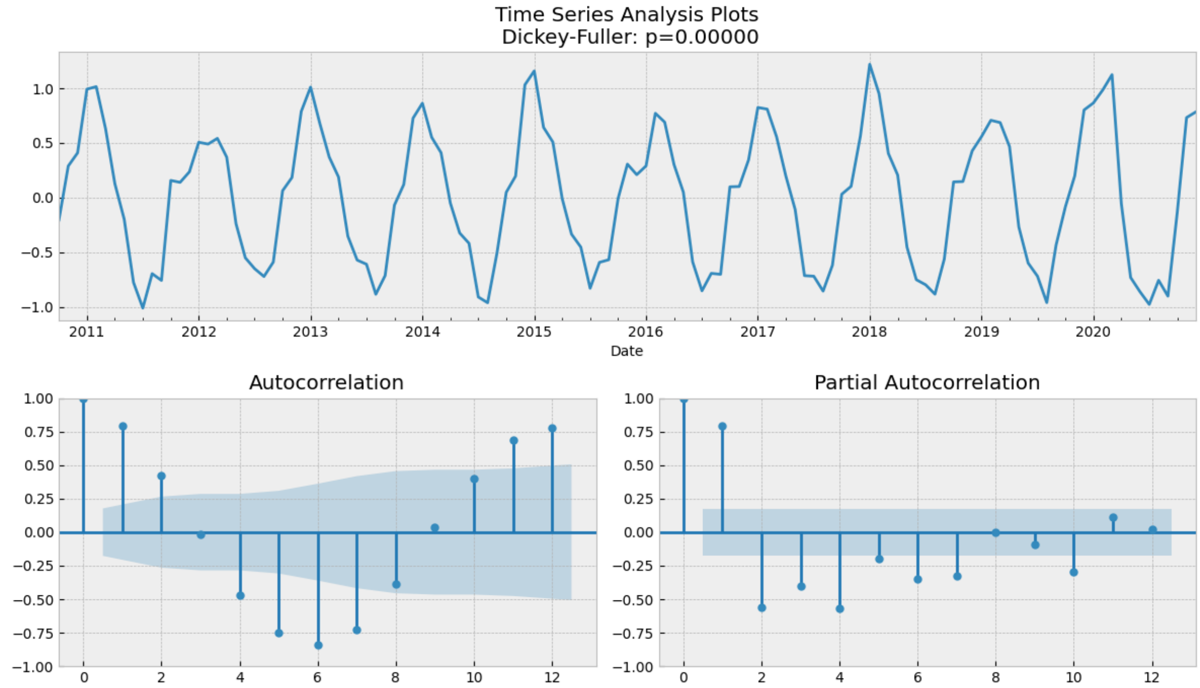

Figure 6. Time series plot alongside autocorrelation plots, ACF and PACF.

From the PACF plot the order of the AR component (p) was identified as 2, as that's where the graph cuts off for the first instance. The series underwent seasonal differencing once for stationarity, thus to prevent increasing noise in the data by further differencing, the 'd' parameter is 0. The MA component(q) to remove autocorrelation was identified as 2. This is because the first lag in ACF plot is significantly out of the limit and the second one is also out of the limit, but it is not that far and there is significant decrease below the significance threshold (blue solid) to 0 at lag 3. Parameter (2, 0, 2) yielded the best goodness of fit via lowest AIC (Akaike Information Criterion) and BIC (Bayesian information criterion) (**Table 3**).

When transitioning from an ARIMA to a SARIMA model, I started with the same non-seasonal parameters of (2, 0, 2). After considering model diagnostics, I determined the most appropriate parameter combination for SARIMA model were non-seasonal orders of (1, 1, 2) and seasonal orders of (0, 1, [1], 12) (**Table 3**).

Table 3. Comparison of performance metrics between ARIMA (2, 0, 2) and SARIMA (1, 1, 2) (0, 1, [1], 12) model.

Model	Performance Metrics	Score
ARIMA	AIC	-28.22
	BIC	-14.715
	HQIC	-22.74
	Ljung-Box	0.04
	Jarque-Bera	1.21
	Mean Squared Error	0.12
	Mean Absolute Error	0.22
	Root Mean Squared Error	0.45
SARIMA	AIC	-37.45
	BIC	-20.58
	HQIC	-30.60
	Ljung-Box	2.44
	Jarque-Bera	0.52
	Mean Squared Error	0.06
	Mean Absolute Error	0.19
	Root Mean Squared Error	0.26

The comparison between the two models (**Table 3**), showed that the SARIMA model outperformed the ARIMA model. This could be due to SARIMA considering seasonality patterns displayed by the series. The Ljung-Box test p-value of **2.44** for SARIMA (>0.05 critical value) caused me to reject the null hypothesis and conclude that the residuals are independently distributed. Confirming the SARIMA assumption of residual independence.

Forecast

To further evaluate how well the model fits the observed data and assess its predictive performance, providing insights into models to capture the data's underlying patterns, trends, and seasonality, I performed in-sample and out-of-sample forecasting for SARIMA model (**Figure 7**).

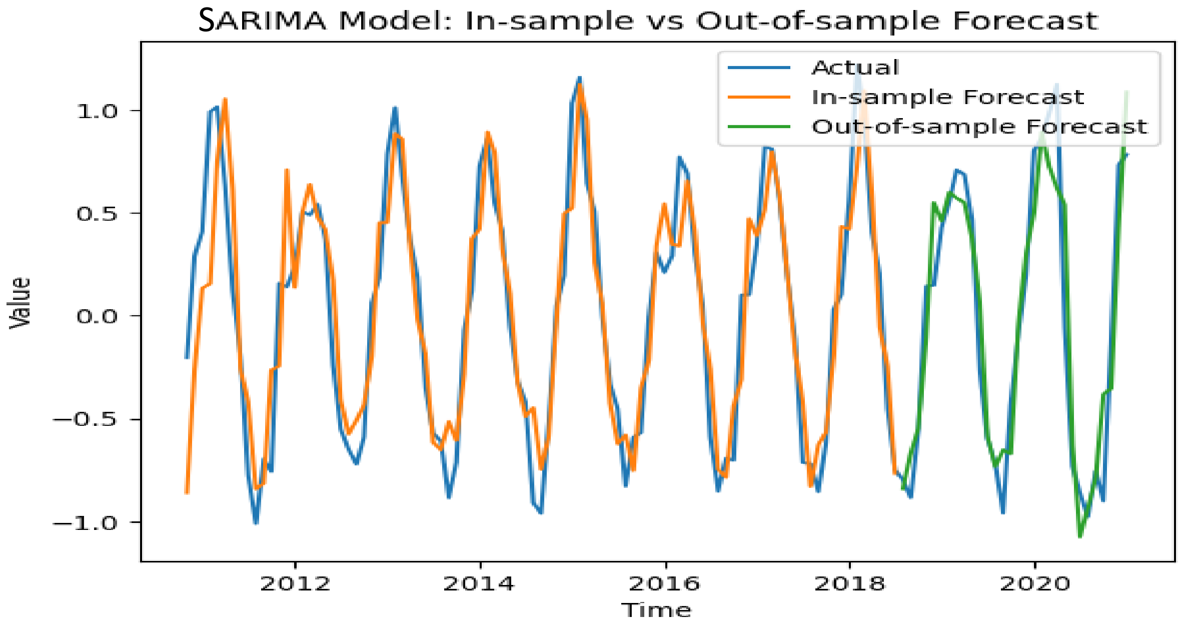

Figure 7. In-Sample and Out-of-Sample Forecast of Influenza Reports

Forecast Accuracy

Both the in-sample and out-of-sample forecast provided visual validation that the models goodness of fit has capability to make accurate predictions. To assess the accuracy of the forecasts generated by each model, I compared the forecasted values against the actual values and use metrics of mean squared error, mean absolute error, root mean squared error to quantify the forecast accuracy (**Table 3**). Overall, SARIMA performed more accurate forecasts, thus I will be using this model for future predictions (**Figure 9**).

Analysis of residues

To help identify any remaining patterns not captured by the model, I analysed the difference between the actual values and the predicted values. Re-running the ADF test for residuals validated the stationarity.

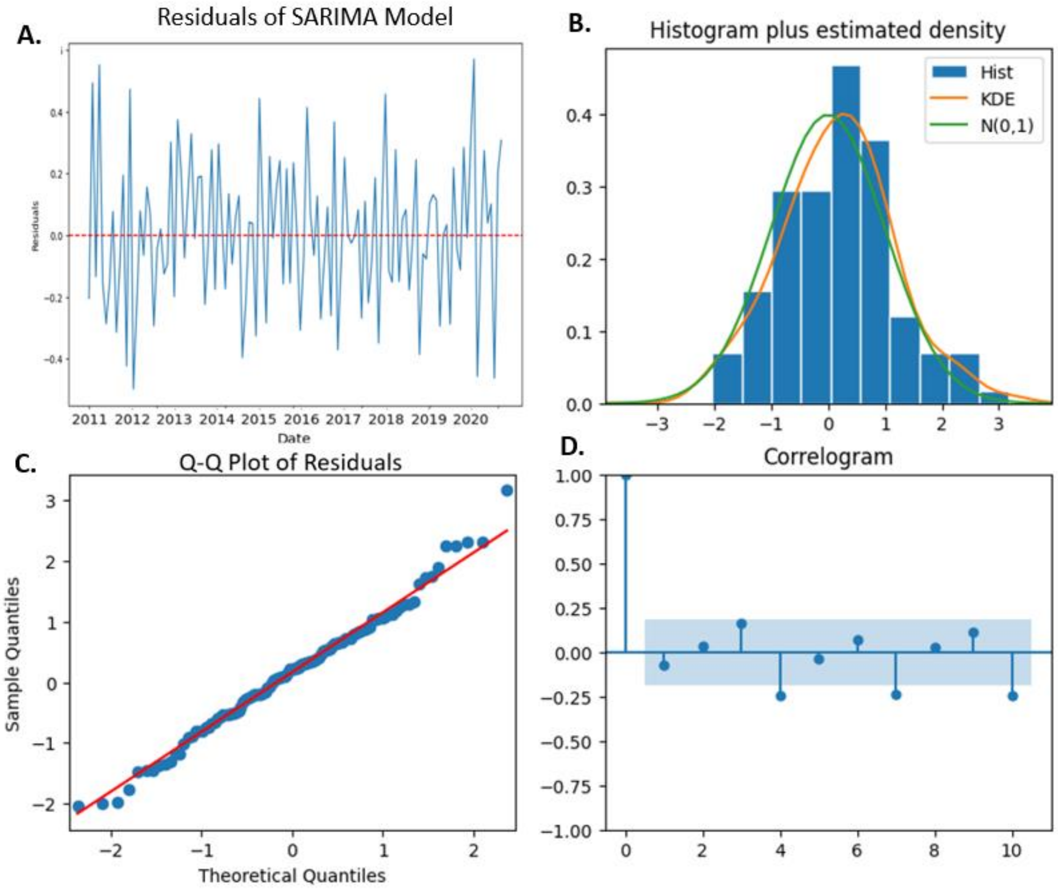

Figure 8. Residual plots for ARIMA residual analysis.

The residual errors seem to fluctuate around a mean of zero and have a uniform variance (**Figure 8a**). Thus, suggesting the model's predictions are unbiased and do not consistently overestimate or underestimate the actual values. Calculated residual standard deviation of **0.28** shows the spread of residuals is tightly clustered around the mean. Overall, a low residual mean, and a small residual standard deviation indicate that the SARIMA model is providing a good fit to the Influenza data, capturing the patterns and variations present in the time series.

The Q-Q plot evidently shows that the residuals have a normal distribution (**Figure 8c**), where the quantiles don't deviate from the normal distribution line (red). The histogram (**Figure 8b**) mostly shows Gaussian distribution, however there is a slight shift towards right, as supported by value of **0.12**. To quantify whether this slight deviation is significant, I performed the Shapiro-Wilk test on the residuals. The p-value of **0.075** fails to reject the null hypothesis of null hypothesis, thus validating that the residuals are normally distributed. the density plot shows that mean is shifted towards right.

Further, the residuals exhibit no autocorrelation (**Figure 8d**), as supported by Ljung-Box test p-value of **2.44** (**Table 3**), providing strong evidence to reject the null hypothesis and conclude that the residuals are independently distributed.

Overall, the residual analysis indicated that the model has captured the underlying patterns in the data and should be competent at predicting future values.

Future Prediction

To avoid any data leakage from the testing set into the training set and to evaluate the performance of the time series models, I utilised the rolling window cross validation (Shrivastava, 2020). Considering the temporal nature of the data, I chose this method as it provides the most effective method at simulating realistic forecasts by reducing the impact of data variability and captures the unique iterations within the data.

To employ future predictions from my SARIMA model on Influenza data, I inversed the previously used boxcox transformation function, to reverse the effects of the transformation and obtain the data in its original units (**Figure 9**).

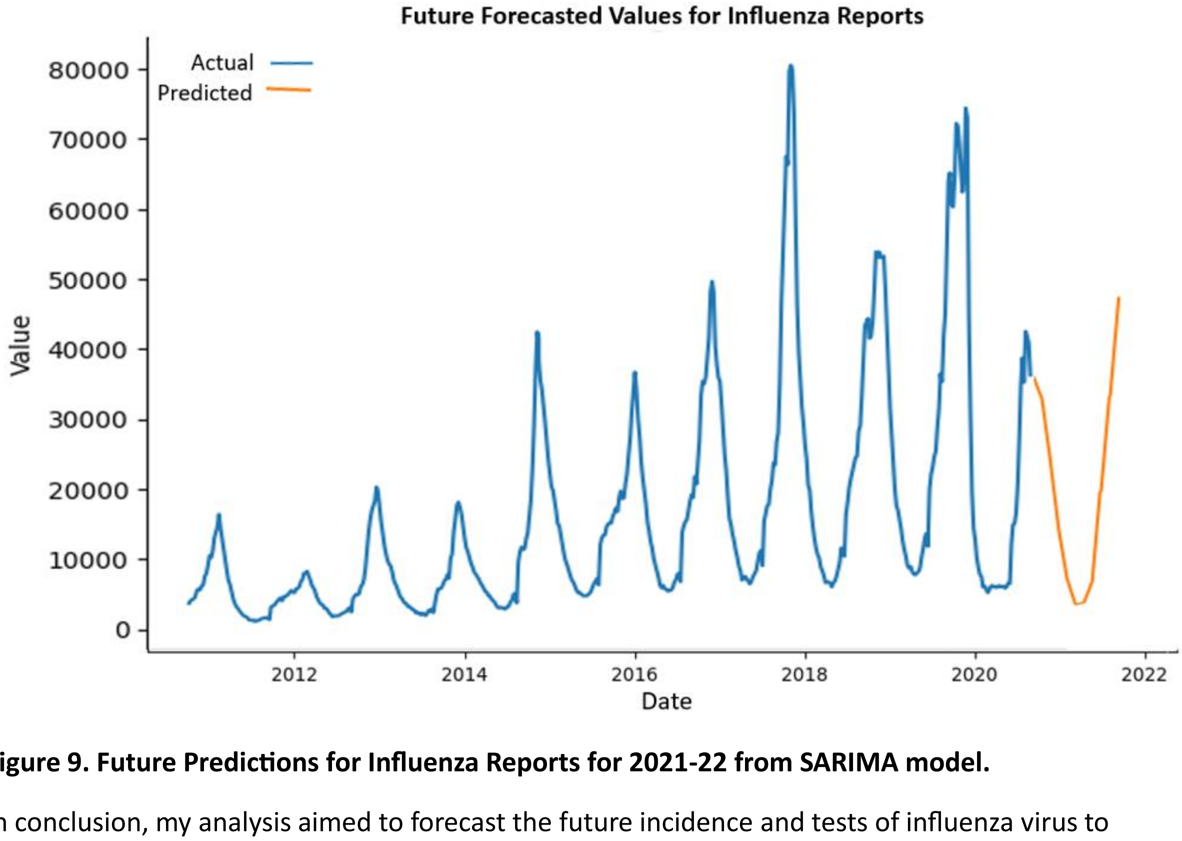

Figure 9. Future Predictions for Influenza Reports for 2021-22 from SARIMA model.

In conclusion, my analysis aimed to forecast the future incidence and tests of influenza virus to support public health planning and resource allocation. The SARIMA model captured and predicted the seasonal and trend components of the complex dynamics of influenza transmission, with high accuracy, as evidenced by low error metrics and close alignment of forecasted values with the observed data (**Table 3**).

SARIMA successfully captured the seasonality of influenza, with peaks occurring during the typical flu season. Thus, the forecast can provide valuable information for public health authorities to allocate resources effectively, such as increasing vaccine production and distributing antiviral medications during high-risk periods. The model's accuracy aids in early detection of potential outbreaks, enabling timely implementation of preventive measures and public health interventions.

Future considerations for enhancing the forecasting accuracy include incorporating external factors such as weather patterns and population demographics, alongside monitoring and model updating using real-time data. Further, the process of parameter estimation during model selection could be automated via a grid search, his would remove manual work and improve efficiency of the model.

References

Asamoah-Boaheng, M. (2014). Using SARIMA to Forecast Monthly Mean Surface Air Temperature in the Ashanti Region of Ghana. *International Journal of Statistics and Applications*, [online] 4(6), pp.292–298. Available at: http://article.sapub.org/10.5923.j.statistics.20140406.06.html [Accessed 29 May 2023].

Malik, F. (2018). *Understanding Auto Regressive Moving Average Model — ARIMA*. [online] Medium. Available at: https://medium.com/fintechexplained/understanding-auto-regressive-model-arima-4bd463b7a1bb.

Shrivastava, S. (2020). *Cross Validation in Time Series*. [online] Medium. Available at: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4.

Stephanie (2016). *KPSS Test: Definition and Interpretation*. [online] Statistics How To. Available at: https://www.statisticshowto.com/kpss-test/.

Stephanie (2018). *Ljung Box Test: Definition*. [online] Statistics How To. Available at: https://www.statisticshowto.com/ljung-box-test/.

Verma, Y. (2021). *Complete Guide To Dickey-Fuller Test In Time-Series Analysis*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/.

WHO (2018). *Influenza (Seasonal)*. [online] Who.int. Available at: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal).