Time Series Analysis on Amazon’s Quarterly Revenues

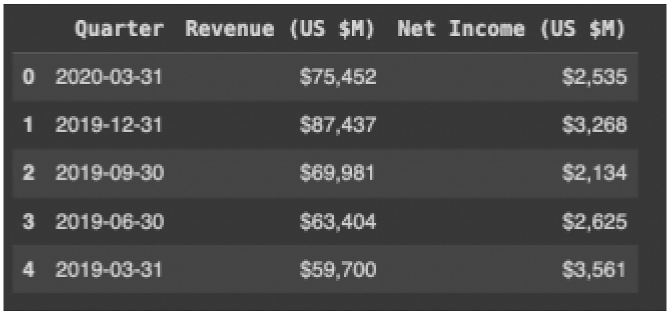
Sambit stita siri sai pradhan(21IM10029)

Abstract

This project undertakes a thorough examination of time series forecasting techniques applied to Amazon’s quarterly revenue dataset spanning the years 2006 to 2020. Employing a diverse set of methodologies including Autoregressive Integrated Moving Average (ARIMA), Moving Average (MA), Exponentially Weighted Moving Average (EWMA), Holt’s Linear Trend Method, Holt-Winters Method, and Seasonal ARIMA (SARIMA), the study aims to provide a comprehensive evaluation of forecasting accuracy and statistical significance. Each model is meticulously implemented, with attention to parameter tuning and validation techniques, to extract meaningful insights into Amazon’s revenue trends. The report not only presents the empirical findings of these forecasting models but also offers critical analyses of their strengths, limitations, and applicability in capturing the complex dynamics inherent in Amazon’s financial data. By synthesizing these methodologies and findings, this study contributes to a deeper understanding of the dataset dynamics, enabling stakeholders to make informed decisions and anticipate future revenue trends with greater confidence. The code and dataset are available in the following repo: time-series-analysis-rtsm.

1. Dataset and Preprocessing

The dataset has been sourced from data.world, and it is open access. The initial dataset view is provided in 1.



	Quarter	Revenue (US $M)	Net Income (US $M)
0	2020-03-31	$75,452	$2,535
1	2019-12-31	$87,437	$3,268
2	2019-09-30	$69,981	$2,134
3	2019-06-30	$63,404	$2,625
4	2019-03-31	$59,700	$3,561

Figure 1: Amazon’s Quarterly Revenue+Profits Dataset

1.1 Preprocessing

For the preprocessing of the dataset, nothing additional had to be done. Since the dataset consisted of both profits and revenues, we removed the profits(a choice which is discussed in following subsection) and considered the revenues alongside the quarters. Since the revenues were in string form, we changed them to integer format for the implementation. Additionally, a null value analysis yielded **0 null values** which made the preprocessing easier.

1.2 Revenues vs Profits

We considered focussing on revenues only instead of profits is because the profits dip below 0 in certain cases, which made prediction visualizations difficult to show as there remained a gap due to some values present below 0. This is the reason why we choose to focus on a time series analysis of the revenues rather than profits. A visualization of the quarterly revenues is provided in 2

2 Dataset Decomposition

Before moving onto implementing the various techniques of forecasting to the revenue dataset, we performed a decomposition of the data present into the main components i.e **trend**,**seasonality** and **residuals**. Following the decomposed view of the components of the dataset, we also implemented an **Augmented Dicky-Fuller Test** to perform a hypothesis testing related to the dataset being stationary or non-stationary.

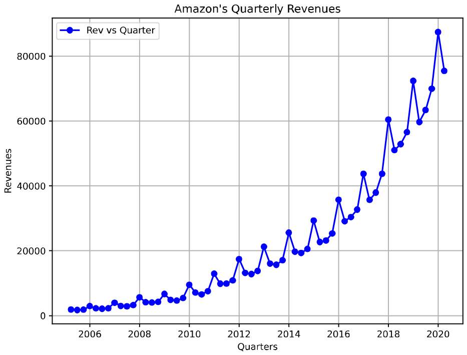


Figure 2: Revenues vs Quarters

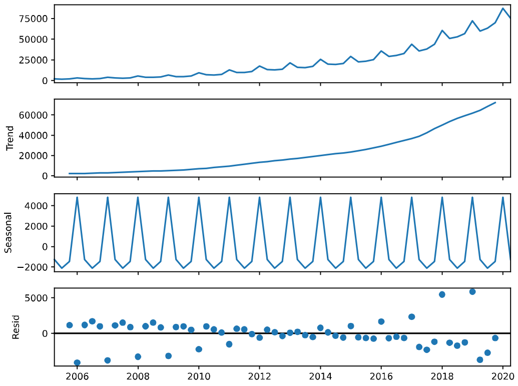


Figure 3: Trend+Seasonality+Residuals

2.1 Decomposed View

We use the `from statsmodels.tsa.seasonal import seasonal_decompose` to provide a decomposed view of the data as discussed before. The visualization is depicted in 3. Here we can see that the range of trend and residual is nominal, or we can say that trend is having variation between 0 to 80000, and most of the time residual is having the variation around 0. But for the seasonality, we can see that it varies between -2000 to 5000.

2.2 ADF Test

For the test of stationarity of the data, we use the augmented dicky-fuller test, wherein setting a confidence interval of **0.95**, we aim to test whether our null hypothesis can be considered to be true or not. Additionally In this context our hypothesis design is such:

$$H_0 : \text{time series has a unit root, indicating it is non-stationary}$$
$$H_1 : \text{time series has no unit root, indicating it is stationary}$$

The important values yielded from the ADF test are listed in 1.

ADF Metrics	**Values**	**Conclusion**
ADF	-2.445	-
P-Value	0.129	**Fail to reject** H_0

Table 1: ADF Test metrics

3 Forecasting Methods

For an in-depth time series analysis of the revenue dataset, we implemented multiple statistical methods which include: Auto-Regressive Integrated Moving Average (**ARIMA**), Moving Average (**MA**), Exponentially Weighted Moving Average (**EWMA)**, Holt's method as well as Holt & Winters method and Seasonal Auto-Regressive Integrated Moving Average (**SARIMA**). Alongside the implementation of the methods, a residual analysis was also performed to test the statistical significance of the methods. Additionally, we also performed a parameter variation tests to identify the locally best parameters for predictions. For all the methods, **RMSE** was used as the accuracy metric.

3.1 Moving Average

The equation for the moving average method is provided here. In this context, MA(t) is the value of the moving average at time t. w is the number of time periods over which the average is calculated and x_{t-i} is the value of the time series at time $t - i$ where $i = 1, 2, 3....w - 1$.

$$\begin{aligned} \text{MA}(t) &= \frac{1}{w} \sum_{i=0}^{w-1} x_{t-i} \\ &= \frac{1}{w} (x_t + x_{t-1} + \cdots + x_{t-w+1}) \end{aligned}$$

In the MA method, the tunable parameter is w, hence we vary the window size within two different ranges, considering one as a short-span variation in the range of $(3, 12)$ and another as a long-span variation in the range of $(6, 12)$. Using RMSE as the metric, we obtain $W_s = 3$ as the optimal short span and $W_l = 6$ as the optimal long span. The prediction visualizations against the original values are given in 4.

3.2 Exponentially Weighted Moving Average

The EWMA method works differently from the MA method, by not providing equal weightage to previous observations, but rather using a smoothing constant $\alpha \in (0, 1]$ which is meant to provide less weightage to old observations and more to recent observations. In the equation, EWMA(t) is the value of the EWMA at time t and x_t is the value of the time series at time t.

$$\text{EWMA}(t) = \alpha \cdot x_t + (1 - \alpha) \cdot \text{EWMA}(t - 1)$$

In the EWMA method, the tunable parameter is w and α, hence we vary the window size within two different ranges, considering one as a short span and another as a long span. Additionally, we vary α between a range of $[0.1, 1)$ with a step of 0.05. Using RMSE as the metric, we obtain $W_s = 3$ as the optimal short span and $W_l = 6$ as the optimal long span due to a positive linear nature of the RMSE vs window-size plot. For α, the optimal value comes to be 0.95, due to a exponentially decreasing nature of the RMSE vs alpha plot. The prediction visualizations against the original values are given in 5.

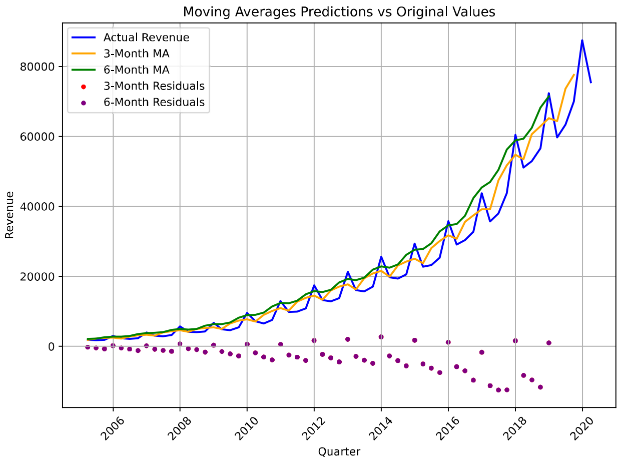


Figure 4: Moving Average for $W_s = 3$ and $W_l = 6$

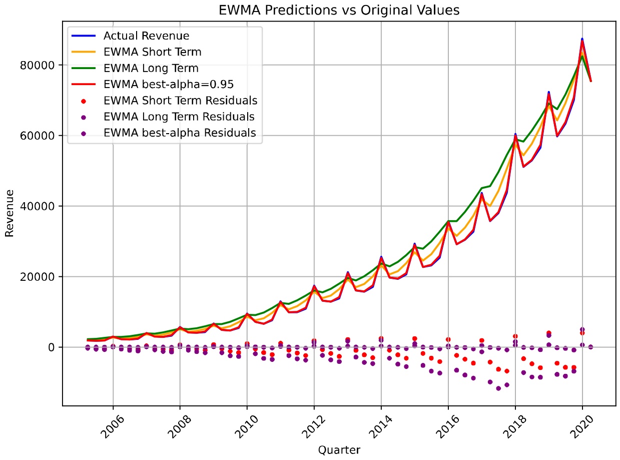


Figure 5: EWMA for $W_s = 3$, $W_l = 6$ and $\alpha = 0.95$

3.3 Holt's Method

Holt's method, also known as Holt's linear trend method or double exponential smoothing, is an extension of simple exponential smoothing that incorporates both a level component and a trend component. The equation for Holt's method (with additive trend) consists of two equations: one for the level (L_t) and one for the trend (T_t), both of which are updated at time t.

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

Following this update, $\hat{Y}_{t+h}$ is updated in the following way, which is an update as an addition of the trend and level components.

$$\hat{Y}_{t+h} = L_t + h \cdot T_t$$

We do not perform parameter tuning for this method and implement the default algorithm which sets $\alpha = 0.5$ and $\beta = 0.5$. The plot is shown in 6.

3.4 Holt and Winter's Method

The Holt-Winters method, also known as triple exponential smoothing, extends Holt's method by including seasonality. It comprises three equations: one for the level (L_t), one for the trend (T_t), and

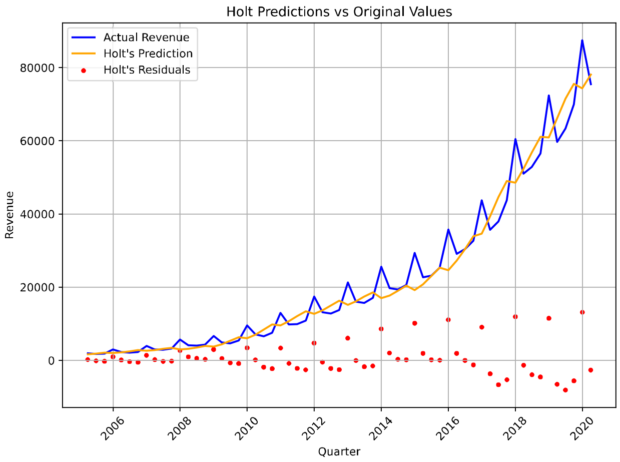


Figure 6: Holt's Method predictions

one for the seasonal component (S_t). The forecast $\hat{Y}_{t+h}$ at time t plus h periods ahead is obtained by combining these components.

The equations for Holt-Winters method (additive seasonality) are as follows:

$$L_t = \alpha(Y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$S_t = \gamma(Y_t - L_{t-1} - T_{t-1}) + (1-\gamma)S_{t-m}$$

Where L_t represents the level component at time t, T_t represents the trend component at time t, and S_t represents the seasonal component at time t. Y_t is the observed value at time t. α, β, and γ are smoothing parameters ($0 < \alpha, \beta, \gamma < 1$). m represents the number of periods in a season, and h represents the forecast horizon.

$$\hat{Y}_{t+h} = L_t + h \cdot T_t + S_{t+h-m\cdot(h//m)}$$

The last equation for $\hat{Y}_{t+h}$ includes an adjustment to handle the seasonality for forecasts beyond the current season. For this method as well, we do not perform individual parameter tuning, but just use the default algorithm for predictions shown in 7.

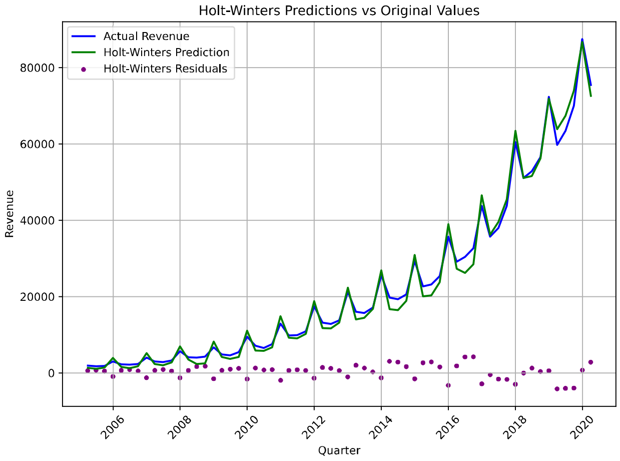


Figure 7: Holt and Winters Method predictions

3.5 Auto-Regressive Integrated Moving Average

The Autoregressive Integrated Moving Average (ARIMA) model, denoted as ARIMA(p, d, q), combines autoregression, differencing, and moving average components. The ARIMA equation is:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Where Y_t is the value of the time series at time t, c is a constant, $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive parameters, ϵ_t is the white noise error term at time t, $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average parameters, and $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ are lagged values of the time series.

For this method, we firstly stationarize the data by implemented a logarithmic difference function using a lag $l = 2$, following this we individually calculate the optimal values of p for the AR method as well as q for the MA method using RMSE as a metric. The optimal values are $p_0 = 2$ and $q_0 = 1$. The prediction plots of the stationarized data through AR(2) and MA(1) are given in 8, as well the final predictions of ARIMA on the normal data is present in 9.

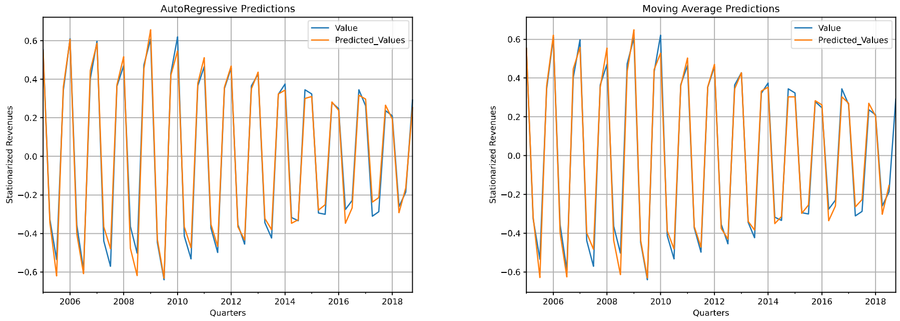


Figure 8: Stationarized Data Predictions (i) AR(2) (ii) MA(1)

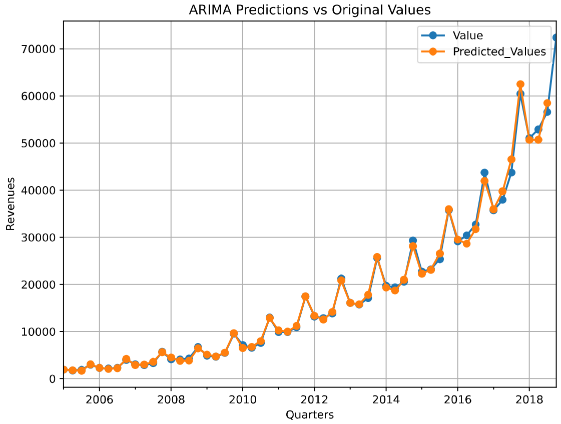


Figure 9: ARIMA with $p_0 = 2$ and $q_0 = 1$

3.6 Seasonal Auto-Regressive Integrated Moving Average

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA model to include seasonal components. The SARIMA equation is:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{P} \Phi_j Y_{t-jm} + \sum_{k=1}^{q} \theta_k \epsilon_{t-k} + \sum_{l=1}^{Q} \Theta_l \epsilon_{t-lm} + \epsilon_t$$

Where Y_t is the value of the time series at time t, c is a constant, ϕ_i and θ_k are the autoregressive and moving average parameters, respectively, Φ_j and Θ_l are the seasonal autoregressive and moving average parameters, respectively, m is the seasonal period, and ϵ_t is the white noise error term at time t.

For the implementation of this method, we do not need to stationarize the data. We recalculate p_0 and q_0 using RMSE with the optimal values being $p_0 = 5$ and $q_0 = 5$. We also set $m = 4$ due 4 quarters being present in a year. The other parameters are set as 0. The predictions are depicted in 10.

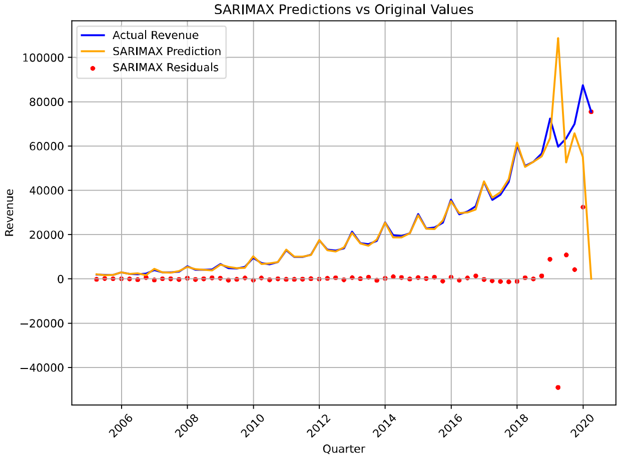


Figure 10: SARIMA with $p_0 = 5$, $q_0 = 5$, $m = 4$ and $P, Q = 0$

4 Residual Analysis and Results

Having implemented all the methods, we move forward to a residual analysis wherein we try and visualize if the residuals resemble a normal distribution. For this purpose, we use density plots as well as quantile-quantile plots.

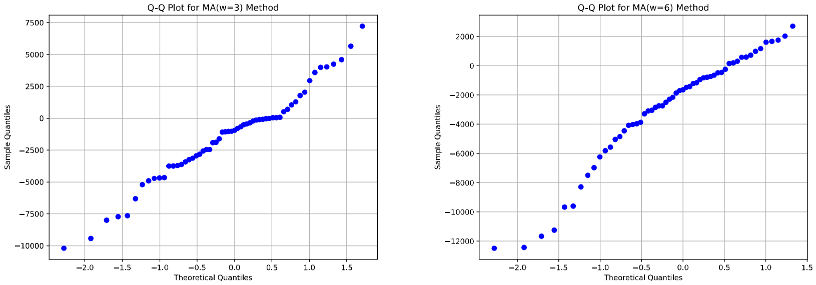


Figure 11: q-q plot MA (i)$W_s = 3$ (ii) $W_l = 6$

From all the quantile-quantile plots, it is clearly observable that ARIMA makes the best predictions from all the methods we had implemented. Better parameter choices to SARIMA could yield improved results, however in the current parameters, it is unable to yield good results. Now for a more concrete analysis of the methods, we also calculated the RMSE values of the optimal predictions made by each method against the given values. The results are shown in 2.

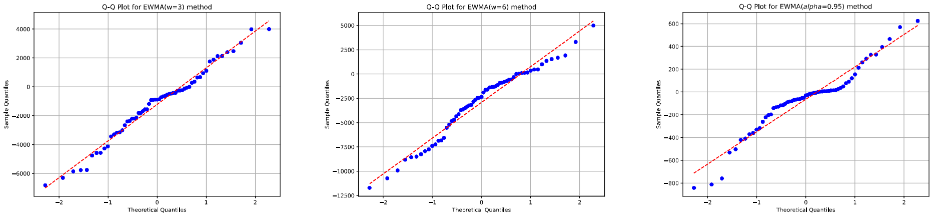


Figure 12: q-q plot EWMA (i)$W_s = 3$ (ii) $W_l = 6$ (iii) $\alpha = 0.95$

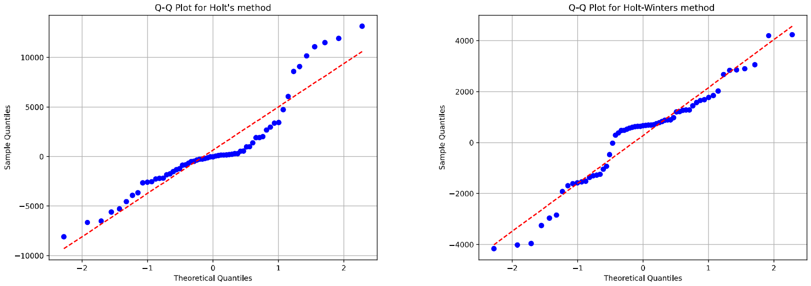


Figure 13: q-q plot (i) Holt (ii) Holt-Winters

Methods	Parameters	RMSE
MA	$W_s = 3$	3817.144
MA	$W_l = 6$	4783.103
EWMA	$W_s = 3$	2763.375
EWMA	$W_l = 6$	4652.107
EWMA	$\alpha = 0.95$	291.988
Holt	-	4534.357
Holt-Winters	-	1887.920
ARIMA	$p_0 = 2$,$q_0 = 1$	828.303
SARIMA	$p_0 = 5$,$q_0 = 5$,$P = 0$,$Q = 0$	12542.062

Table 2: Root Mean Square Values

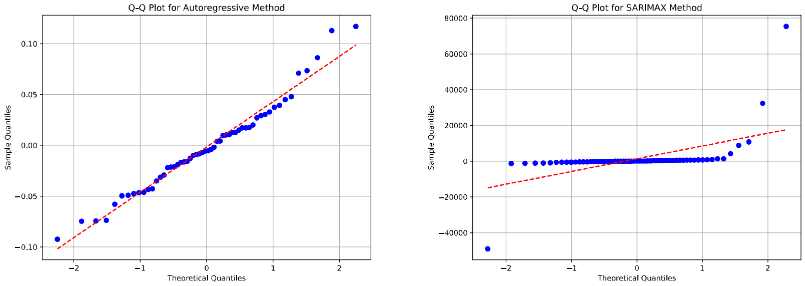


Figure 14: q-q plot (i) ARIMA (ii) SARIMA

5 Conclusion

The report goes into an analysis of the dataset which is based on Amazon's Quarterly Revenues, delving into the decomposition of the dataset into the trend, seasonality and residual components. Following this, we implement various statistical methods such as MA, EWMA, Holt's Method, Holt-Winter's Method, ARIMA and SARIMA alongside parameter tuning as well as residual analysis

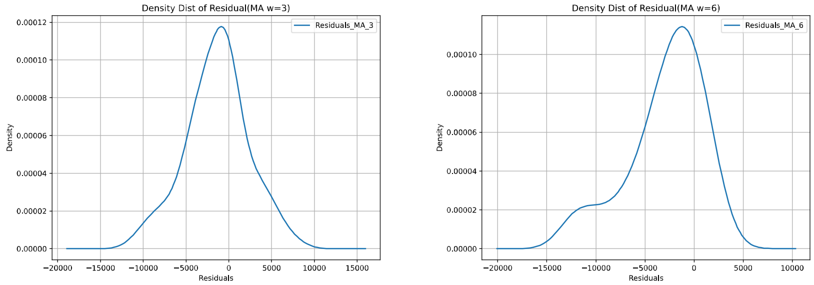


Figure 15: density plot MA (i)$W_s = 3$ (ii) $W_l = 6$

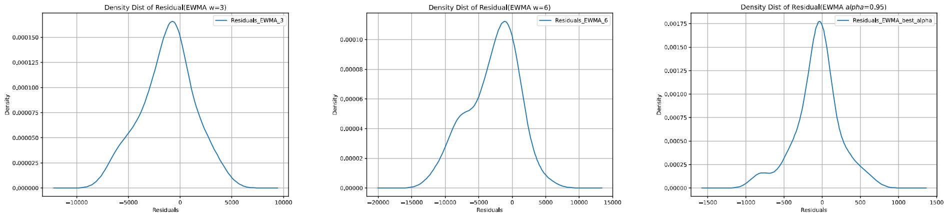


Figure 16: density plot EWMA (i)$W_s = 3$ (ii) $W_l = 6$ (iii) $\alpha = 0.95$

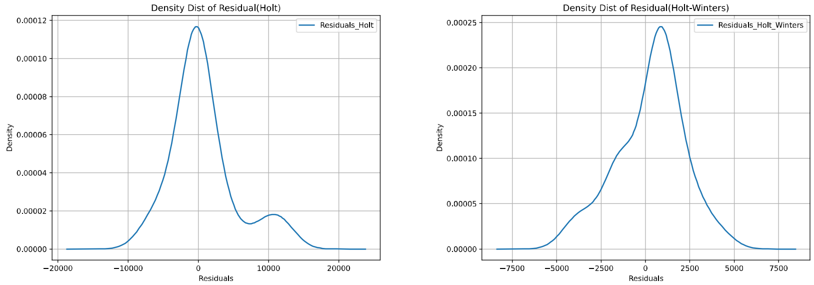


Figure 17: density plot (i) Holt (ii) Holt-Winters

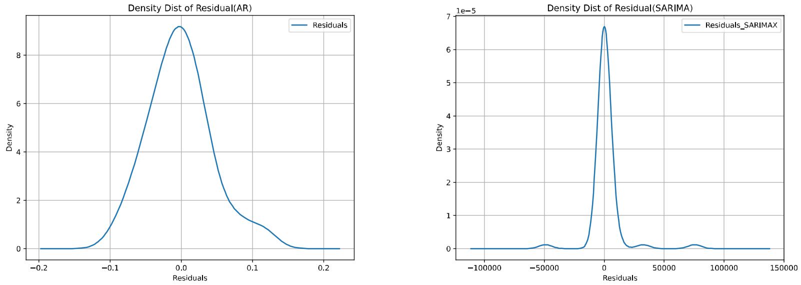


Figure 18: density plot (i) ARIMA (ii) SARIMA

and RMSE value comparison to provide an overall time-series analysis based on concepts covered in Regression and Time Series Models(MA60280). We humbly thank our professor Dr. Buddhananda Bannerjee for guiding us and helping us understand the complex concepts in time series models as well as regression.

6 Appendix

6.1 Auto-correlation Test

For the implementation of ARIMA, after having stationarized the data we perform autocorrelation tests to test for correlation as well as partial correlations. The visualizations are provided in 19

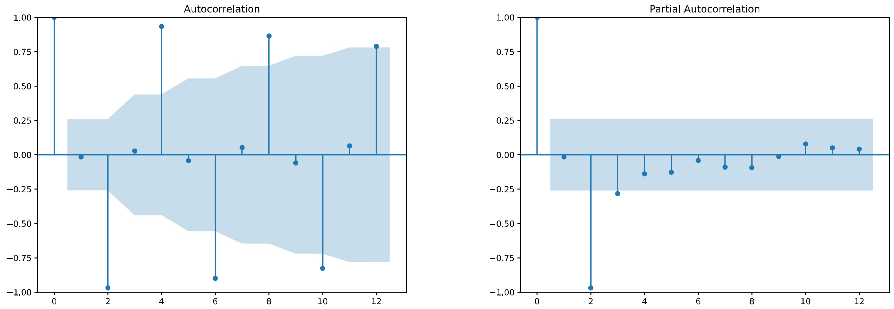


Figure 19: ACF Tests with lag $l = 12$

6.2 Parameter tuning

The various RMSE vs parameter variation plots are provided in this section under the subheadings of the different methods.

6.2.1 Moving Average

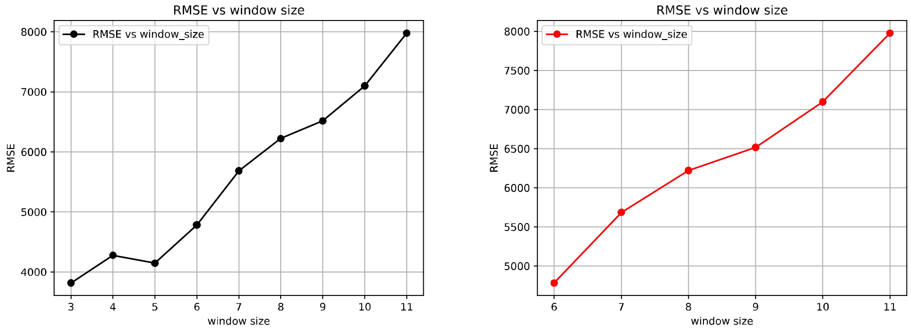


Figure 20: RMSE vs Window Size (i) W_s (ii) W_l

6.2.2 Exponentially Weighted Moving Average

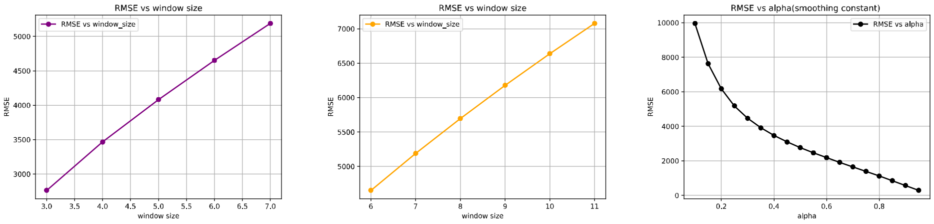


Figure 21: RMSE vs (i) W_s (ii) W_l (iii) α

6.2.3 ARIMA

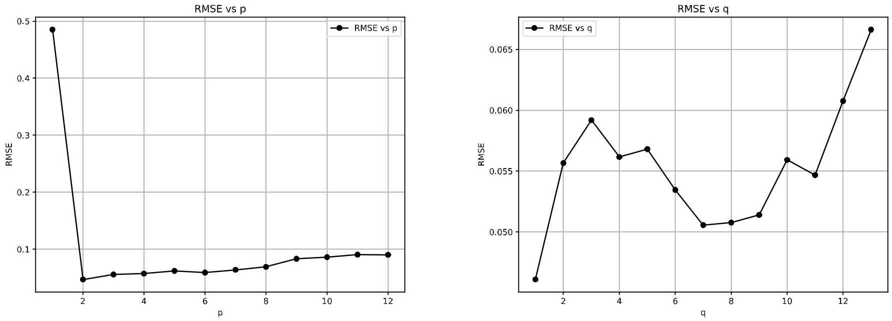


Figure 22: RMSE vs (i) p (ii) q

6.2.4 SARIMA

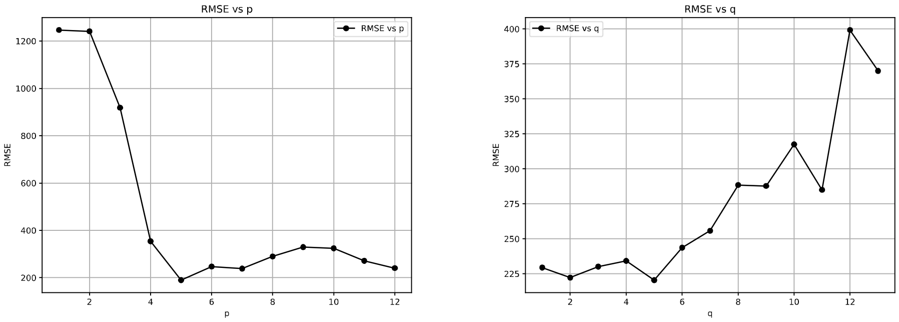


Figure 23: RMSE vs (i) p (ii) q