Website Visitors Analysis and Trend Analysis

Time series forecasting

OPIM 5671: Data Mining and Business Intelligence

Group 6
Chang Wei Ning
Sai Thokal
Alberto Akkarakkaran
Yashas Guddada Sreenivas
Vatsavaya Jasvitha

CONTENTS

1. EXECUTIVE SUMMARY

We are launching a strategic initiative to develop a comprehensive forecasting model for page loads on our platform. This project leverages advanced statistical techniques to analyze historical traffic data, identify patterns, and predict future trends.

The primary objective is to build robust forecasting models for daily page load counts. The dataset spans five years, capturing various web traffic metrics with complex seasonality and fluctuations. Key challenges include the unique nature of web traffic data, capturing seasonal effects, and accounting for factors like automated bot traffic or irregular load spikes.
Model performance will be evaluated using accuracy metrics such as MAE and RMSE to ensure reliability. Successful forecasting will enable better resource allocation, improved server performance, and enhanced user experience. Insights gained from the forecasting models will drive strategic decision-making, optimize content delivery, and support proactive maintenance scheduling to enhance platform efficiency.

1.1 MODELS TO BE INCLUDED

ARMA (Autoregressive Moving Average): This model will serve as a baseline, capturing the autocorrelation within the data.
ARIMA (Autoregressive Integrated Moving Average): Building on the ARMA model, ARIMA will also account for non-stationarity in the time series data.
SARIMA (Seasonal Autoregressive Integrated Moving Average): As an extension of ARIMA, SARIMA will be employed to model and forecast seasonal variations in the data, a critical component given the cyclical nature of visitation patterns.

1.2 MODEL EVALUATION

Assessing the predictive performance of the models using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), Akaike's Information Criterion (AIC)and Schwarz Bayesian Criterion (SBC)

1.3 APPROACH

Perform comprehensive data analysis to understand historical trends and seasonality.
Develop and fit ARMA, ARIMA, and SARIMA models to the historical data.
Conduct rigorous model diagnostics to assess the fit and ensure the residuals meet the assumptions of the selected models.
Evaluate model performance using criteria such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), RMSE (Root Mean Square Error), and MAE (Mean Absolute Error).

1.4 ANTICIPATED KEY FINDINGS

Identification of seasonal and non-seasonal variations in page loads to understand usage patterns.
Determination of the most accurate and reliable forecasting model based on historical traffic trends.
Establishment of key performance metrics to continuously evaluate and refine forecasting accuracy.
Expected Outcomes:
A validated model that provides accurate page load forecasts with high accuracy. Actionable insights on visitor behavior, helping optimize server allocation, content scheduling, and marketing strategies
a framework for continuous monitoring and refinement of forecasting accuracy based on key metrics like MAE, MSE, RMSE, and MAPE
Next Steps:
Collection and pre-processing of historical 'Page Loads' data.
Initial model development and parameter estimation.
Validation and refinement of models using recent data.
Final model selection based on performance metrics and business relevance.

2. INTRODUCTION

2. 1 PROJECT OVERVIEW

Background: understanding and forecasting page load patterns is essential for optimizing website performance and resource management. fluctuations in page loads can impact server capacity, content deployment, and overall user experience, making accurate predictions valuable for strategic planning and operational efficiency.
Objective: the project aims to create an accurate forecasting model for page loads, enabling the prediction of traffic patterns and helping align business strategies with these insights.

2.2 PROBLEM STATEMENT

The objective of this project is to develop accurate and robust time series forecasting models to predict the daily number of page loads for the website. the dataset spans five years, from september 14, 2014, to august 19, 2020 and consists of 2167 daily observations. The variables include counts of page loads, unique visitors, first-time visitors, and returning visitors. The website's traffic exhibits complex seasonality, influenced by day-of-the-week patterns and academic calendar variations.
page loads represent the total number of times web pages are accessed, providing insight into user engagement and website activity. challenges in this project include handling the unique

characteristics of web traffic data, capturing intricate seasonality effects, and addressing anomalies such as automated bot traffic and sudden traffic spikes.

The forecasting models will be evaluated based on their accuracy in predicting daily page loads, considering metrics such as mean absolute error (MAE), root mean squared error (RMSE), and other relevant statistical measures. The successful completion of this project will enable better anticipation of website traffic trends, optimize resource allocation, improve server efficiency, and enhance the overall user experience.

3. DATA DESCRIPTION

This dataset contains seven key columns that are the primary focus of our analysis, which are listed below. The dataset includes twenty-four columns in the orders table, which serves as the main focus of our study, with the columns listed below.

1. Row:

- Each row in the dataset represents a single data point.
- It captures daily website metrics, providing insights into page loads and visitor activity.

2. Day:

- Indicates the day of the week for the corresponding date.

3. Date:

- Represents the calendar date of the observation.
- Spans a range from September 14, 2014, to August 19, 2020, covering a five-year period.

4. **Page Loads**:

- Refers to the total number of pages loaded on the website on a given day.

5. Unique Visits:

- Represents the count of unique visitors to the website on a specific day.

6. **First Time Visits**:

- Indicates the number of visitors who are accessing the website for the first time.

7. Returning Visits:

- Represents the count of visitors who have returned to the website.

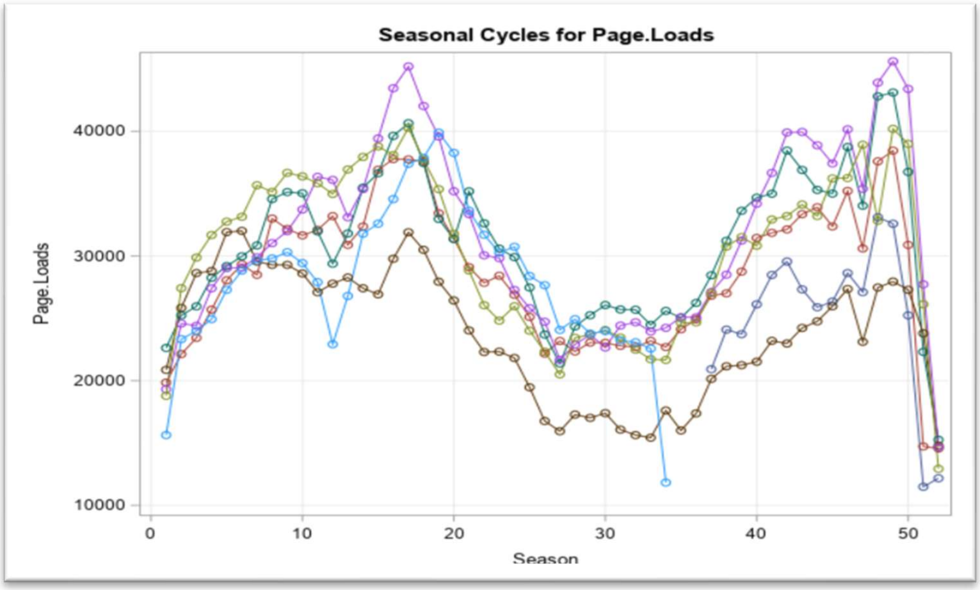
4. DATA AGGREGATION

Before proceeding with data aggregation, it was essential to clean the dataset to ensure accuracy in analysis. Using python, we corrected date formats by converting them to datetime format, removed any non-numeric characters from numerical columns, and handled missing values. Additionally, the index was reset to maintain consistency after cleaning. These steps were crucial in preparing the dataset for further analysis and reliable forecasting.

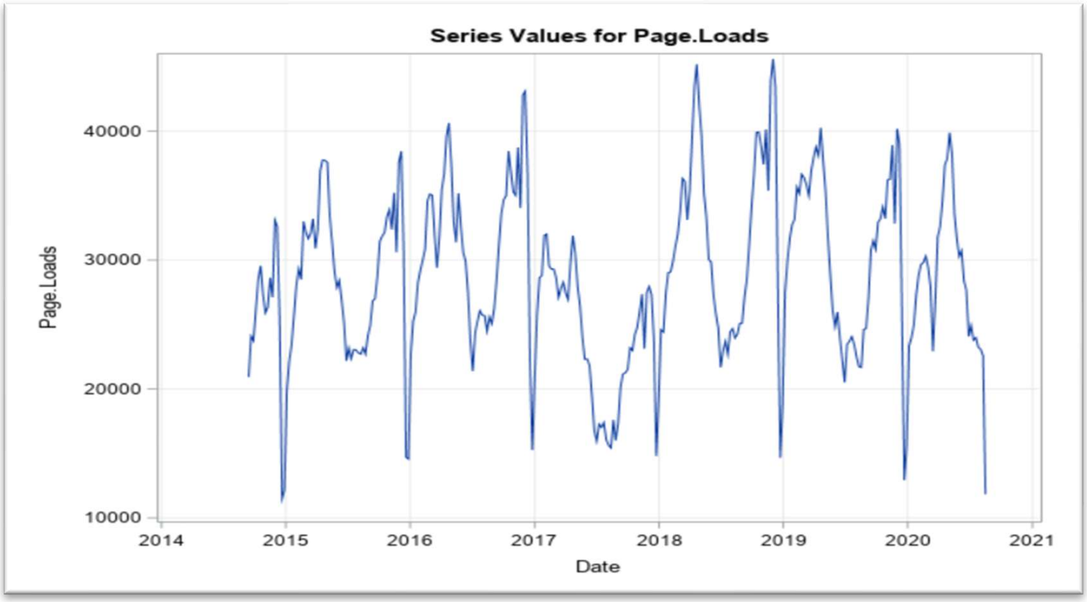
```
# Convert the Date column to datetime format

import pandas as pd

df['Date'] = pd.to_datetime(df['Date'], errors='coerce').

# Remove any non-numeric characters from numerical columns

columns_to_clean = ['Page.Load', 'Unique.Vis', 'First.Time.', 'Returning.Visits']
df[columns_to_clean] = df[columns_to_clean].replace({',': ''},
regex=True).apply(pd.to_numeric, errors='coerce )

# Reset the index after cleaning

 df. reset_index(drop=True, inplace=True)
```

To ensure a reliable forecasting model, we first cleaned the dataset by converting date formats to datetime, removing non-numeric characters from numerical columns, handling missing values, and resetting the index for consistency. After data cleaning, we focused on aggregating key metrics such as page loads, unique visits, first-time visits, and returning visits on a weekly basis. This process helped smooth out daily fluctuations, offering a clearer understanding of website traffic trends and user engagement over time. Analyzing the weekly aggregated data allowed us to identify patterns, trends, and anomalies in page load activity, providing deeper insights into visitor behavior and enabling the development of more accurate forecasting models. Throughout the aggregation process, rigorous validation checks were performed to maintain data integrity and accuracy, ensuring the reliability of our analysis. By adopting this structured approach, we mitigated challenges posed by noisy daily data, improving the precision of our forecasting models and enhancing strategic decision-making for website performance optimization.

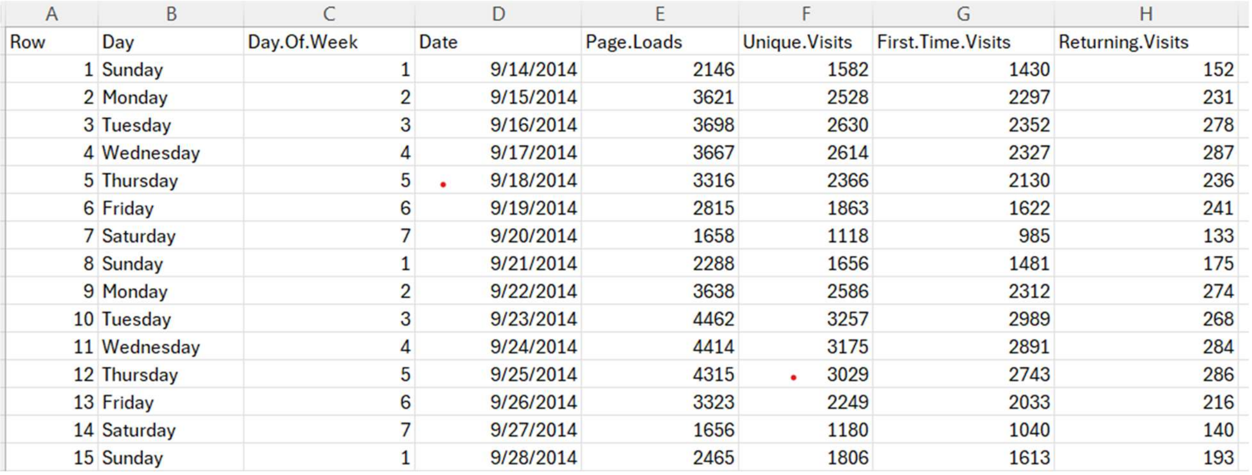
5. DECOMPOSITION ANALYSIS



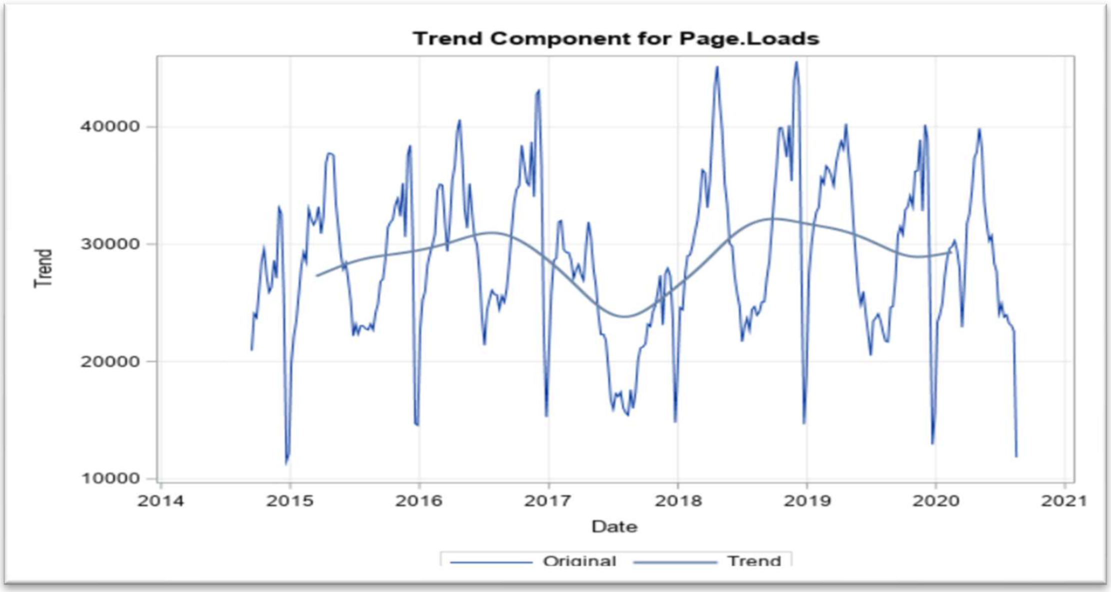
Seasonal cycles plot for page loads illustrates recurring patterns over a yearly period, highlighting peaks and troughs in website traffic. These fluctuations suggest strong seasonal trends, likely influenced by external factors such as academic schedules, holidays, or marketing events.



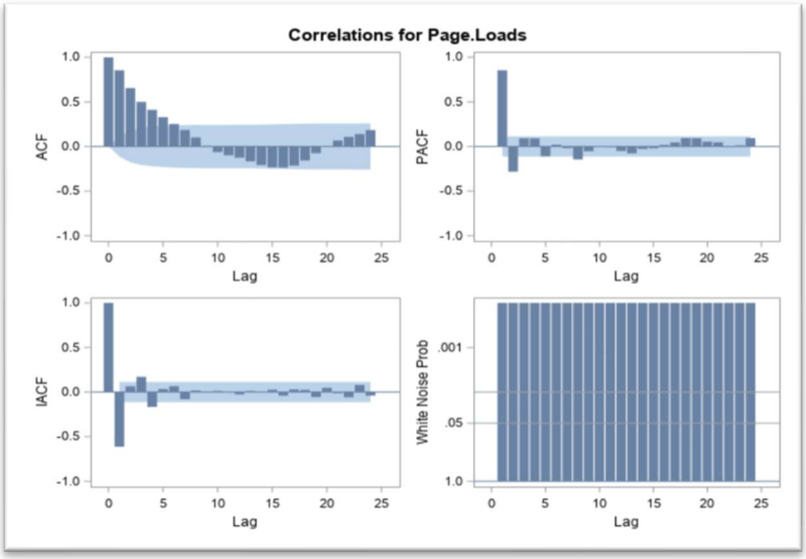
The time series plot for page loads reveals recurring seasonal trends and fluctuations over the years, indicating periods of high and low website traffic.



A	B	C	D	E	F	G	H
Row	Day	Day.Of.Week	Date	Page.Loads	Unique.Visits	First.Time.Visits	Returning.Visits
1	Sunday	1	9/14/2014	2146	1582	1430	152
2	Monday	2	9/15/2014	3621	2528	2297	231
3	Tuesday	3	9/16/2014	3698	2630	2352	278
4	Wednesday	4	9/17/2014	3667	2614	2327	287
5	Thursday	5	9/18/2014	3316	2366	2130	236
6	Friday	6	9/19/2014	2815	1863	1622	241
7	Saturday	7	9/20/2014	1658	1118	985	133
8	Sunday	1	9/21/2014	2288	1656	1481	175
9	Monday	2	9/22/2014	3638	2586	2312	274
10	Tuesday	3	9/23/2014	4462	3257	2989	268
11	Wednesday	4	9/24/2014	4414	3175	2891	284
12	Thursday	5	9/25/2014	4315	3029	2743	286
13	Friday	6	9/26/2014	3323	2249	2033	216
14	Saturday	7	9/27/2014	1656	1180	1040	140
15	Sunday	1	9/28/2014	2465	1806	1613	193



The trend component for page loads highlights long-term patterns in website traffic, capturing underlying growth and decline phases. The smoother trend line shows gradual increases and decreases, while the original series reflects short-term fluctuations. This visualization helps identify sustained shifts in visitor engagement over time.

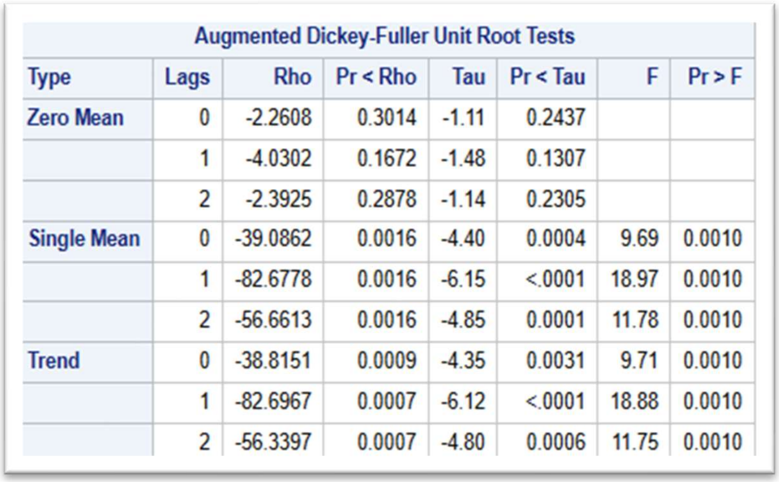


ACF (Autocorrelation Function) Plot: The ACF plot shows a slow decay, indicating strong autocorrelation, suggesting non-stationarity in the data.

PACF (Partial Autocorrelation Function) Plot: The PACF plot exhibits a sharp drop after the first lag, suggesting a possible autoregressive (AR) process that could be modeled with a low-order AR component.

IACF (Inverse Autocorrelation Function) Plot: The IACF plot remains relatively stable, indicating that the moving average (MA) component may not play a dominant role in modeling the data.

White Noise Probability: The bars in the white noise plot remain consistently high, indicating that the residuals are not behaving like white noise, suggesting that the current model does not fully capture all patterns in the data.



Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-2.2608	0.3014	-1.11	0.2437		
	1	-4.0302	0.1672	-1.48	0.1307		
	2	-2.3925	0.2878	-1.14	0.2305		
Single Mean	0	-39.0862	0.0016	-4.40	0.0004	9.69	0.0010
	1	-82.6778	0.0016	-6.15	<.0001	18.97	0.0010
	2	-56.6613	0.0016	-4.85	0.0001	11.78	0.0010
Trend	0	-38.8151	0.0009	-4.35	0.0031	9.71	0.0010
	1	-82.6967	0.0007	-6.12	<.0001	18.88	0.0010
	2	-56.3397	0.0007	-4.80	0.0006	11.75	0.0010

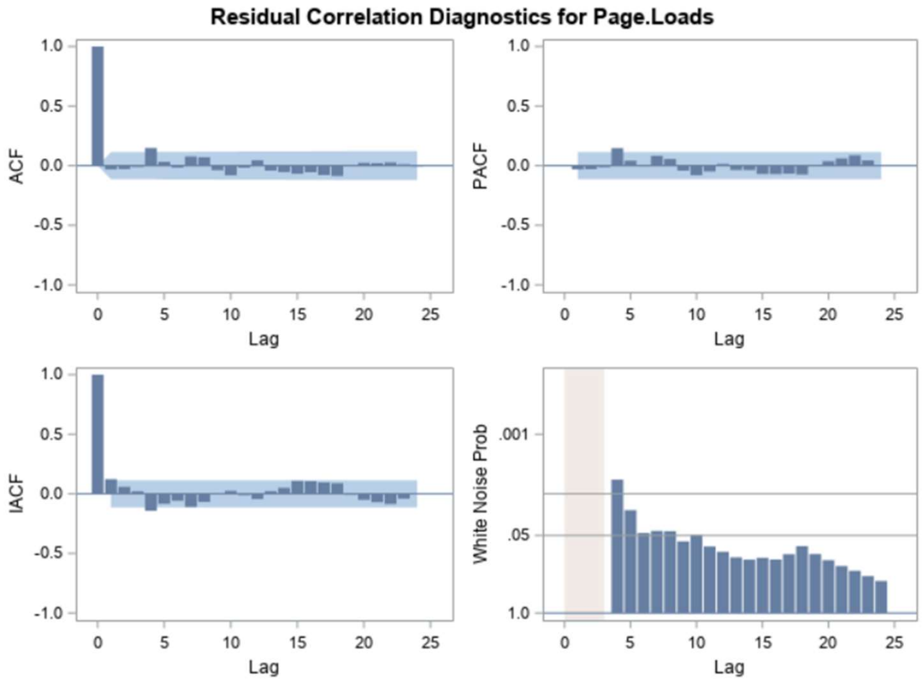
This augmented dickey-fuller (ADF) unit root test table evaluates the stationarity of the page loads time series. The test is conducted under three conditions: zero mean, single mean, and trend. The tau values are test statistics that determine stationarity, where lower values indicate stronger evidence against a unit root. The p-values (pr < tau) for the zero mean test are relatively high, suggesting non-stationarity, while for single mean and trend, they are below 0.01, indicating stationarity. The f-statistics and its p-value confirm the significance of the results. Since the dataset appears non-stationary under the zero mean assumption but stationary under the trend, differencing may be necessary before applying time series modeling.

6. **MODELS**

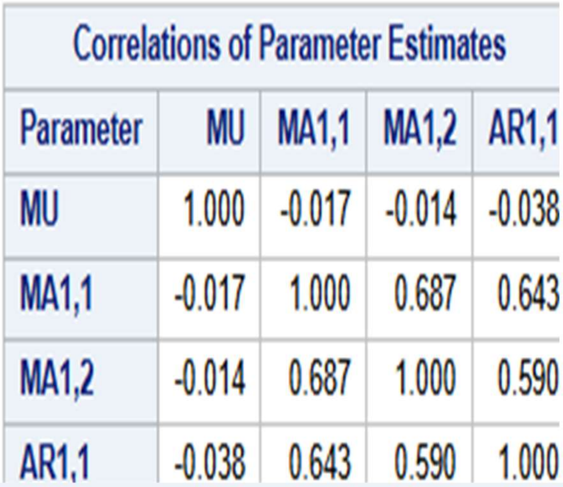
The augmented dickey-fuller (ADF) test results indicate that the data is stationary, as the p-values (pr < tau) for both the single mean and trend cases are below 0.01. This means we can reject the null hypothesis of a unit root, confirming that the time series does not exhibit a strong trend or drift requiring differencing. Since stationarity is already achieved, there is no need for prewhitening or additional transformations before applying forecasting models. The data is ready for modeling using techniques like ARIMA or SARIMA without requiring further adjustments for stationarity.

6.1 ARMA MODEL

6.1.1 ARMA MODEL-(1,2)

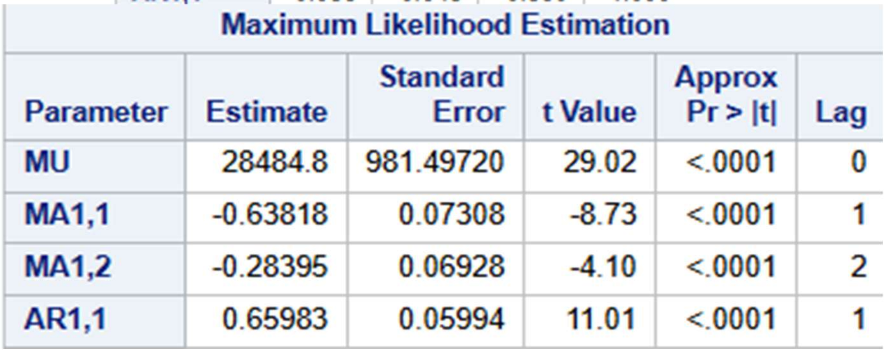


The residual correlation diagnostics confirm that the model adequately captures the patterns in page loads. The ACF and PACF plots show no significant autocorrelations, while the IACF suggests minimal remaining structure. The white noise probability plot indicates randomness in residuals, supporting the model's effectiveness.



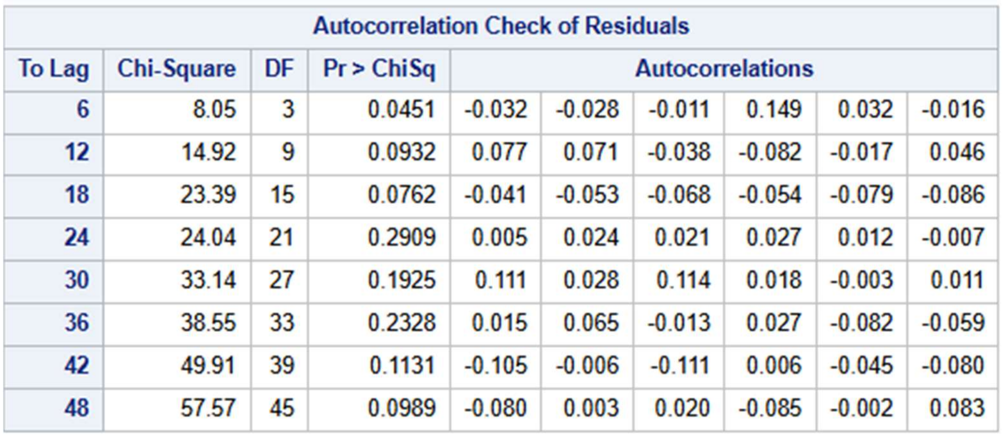
Correlations of Parameter Estimates

Parameter	MU	MA1,1	MA1,2	AR1,1
MU	1.000	-0.017	-0.014	-0.038
MA1,1	-0.017	1.000	0.687	0.643
MA1,2	-0.014	0.687	1.000	0.590
AR1,1	-0.038	0.643	0.590	1.000



Maximum Likelihood Estimation

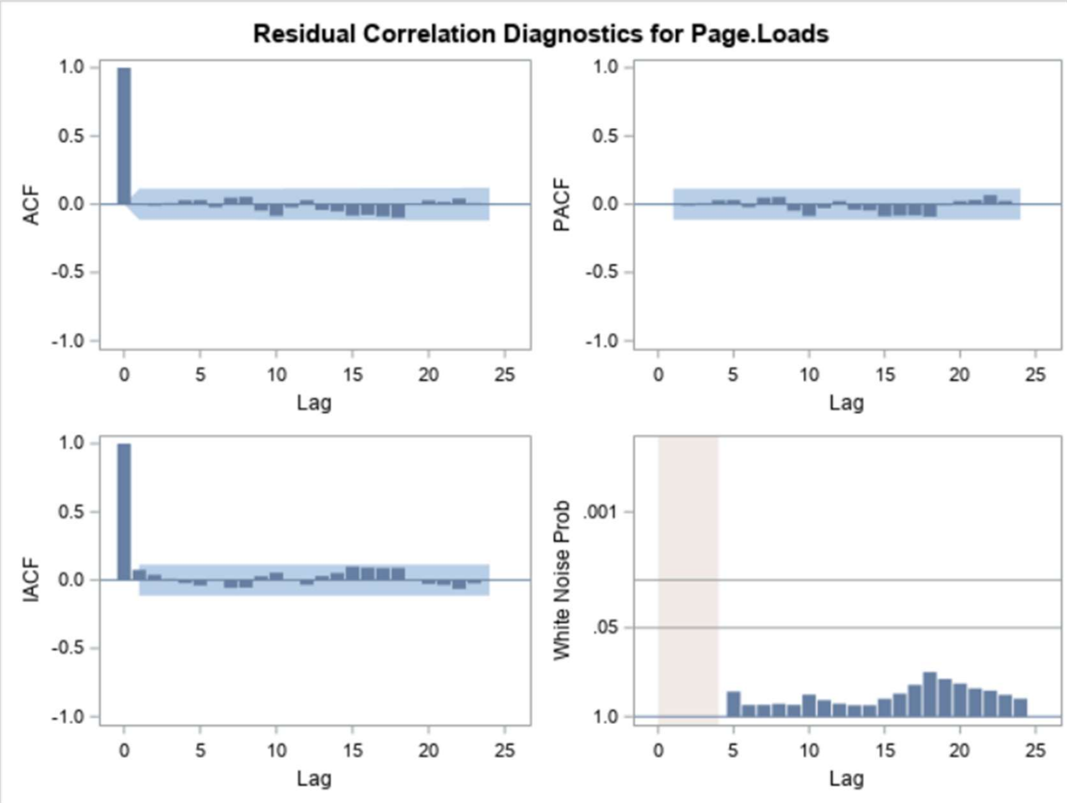
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	28484.8	981.49720	29.02	<.0001	0
MA1,1	-0.63818	0.07308	-8.73	<.0001	1
MA1,2	-0.28395	0.06928	-4.10	<.0001	2
AR1,1	0.65983	0.05994	11.01	<.0001	1



Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.05	3	0.0451	-0.032	-0.028	-0.011	0.149	0.032	-0.016
12	14.92	9	0.0932	0.077	0.071	-0.038	-0.082	-0.017	0.046
18	23.39	15	0.0762	-0.041	-0.053	-0.068	-0.054	-0.079	-0.086
24	24.04	21	0.2909	0.005	0.024	0.021	0.027	0.012	-0.007
30	33.14	27	0.1925	0.111	0.028	0.114	0.018	-0.003	0.011
36	38.55	33	0.2328	0.015	0.065	-0.013	0.027	-0.082	-0.059
42	49.91	39	0.1131	-0.105	-0.006	-0.111	0.006	-0.045	-0.080
48	57.57	45	0.0989	-0.080	0.003	0.020	-0.085	-0.002	0.083

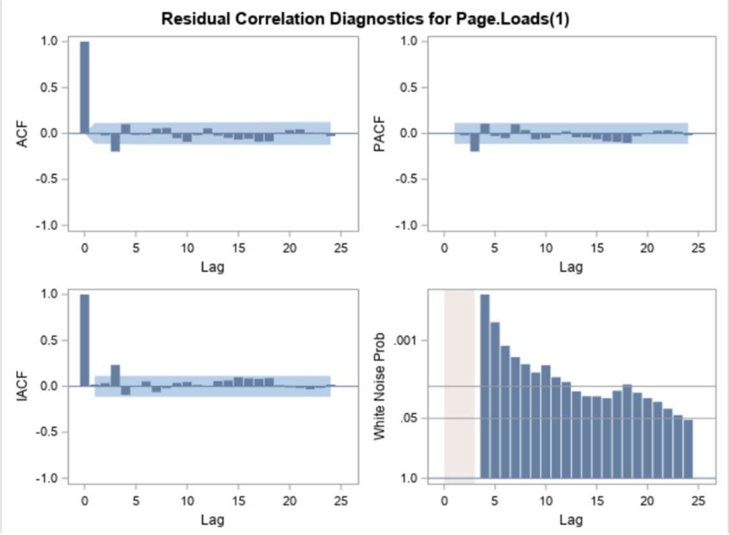
6.1.2 ARMA MODEL –(1,3)

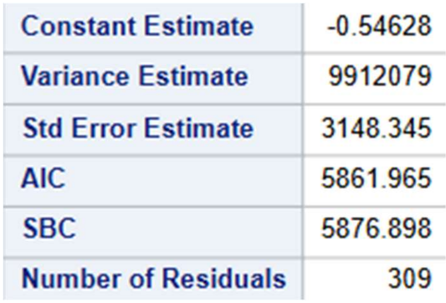


The residual correlation diagnostics indicate that the model has effectively captured the patterns in page loads, as the residuals exhibit white noise characteristics. The ACF and PACF plots show no significant autocorrelation, while the IACF confirms the absence of remaining structure. The white noise probability plot further supports that the residuals are random, suggesting a well-fitted model.

6.2 ARIMA MODELS

6.2.1 ARIMA MODEL-(1,1,2)

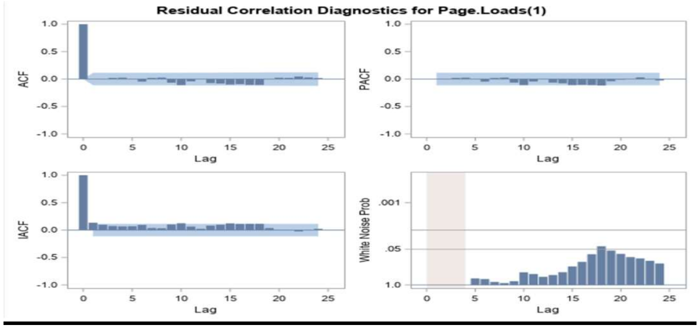


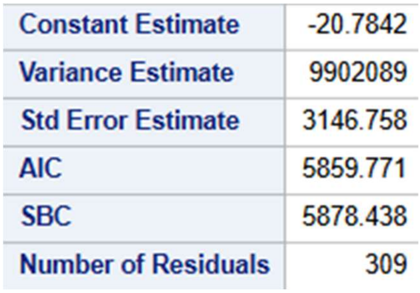


Constant Estimate	-0.54628
Variance Estimate	9912079
Std Error Estimate	3148.345
AIC	5861.965
SBC	5876.898
Number of Residuals	309

The table presents key statistical estimates for the model, including the constant estimate, variance estimate, and standard error estimate. The Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC) values help assess model performance, with lower values indicating a better fit. The number of residuals, 309, indicates the amount of error terms used in evaluating model accuracy.

6.2.2 ARIMA MODEL –(1,1,3)

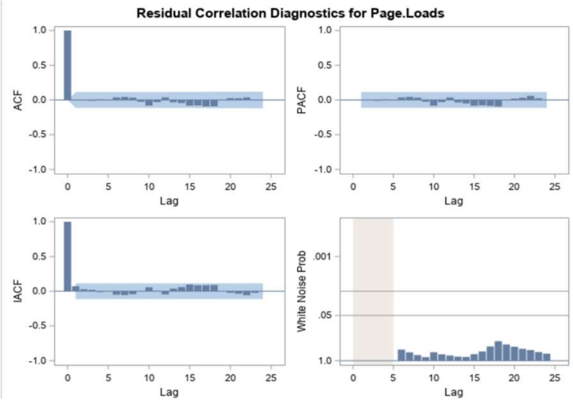


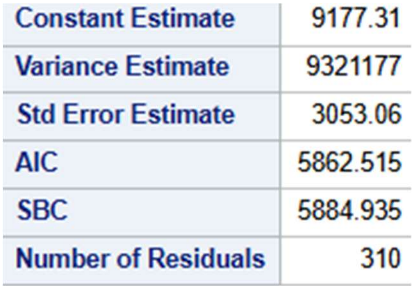


Constant Estimate	-20.7842
Variance Estimate	9902089
Std Error Estimate	3146.758
AIC	5859.771
SBC	5878.438
Number of Residuals	309

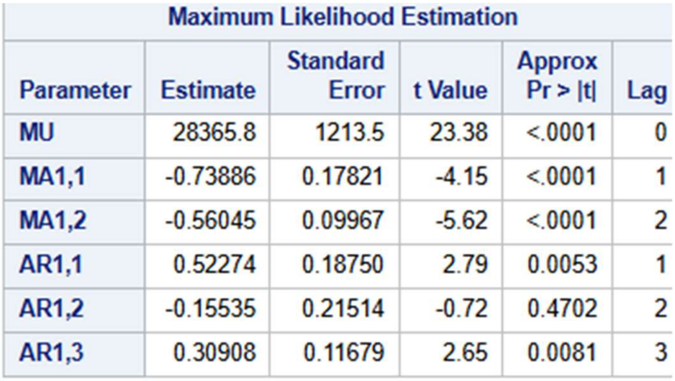
The residual correlation diagnostics for page loads show that the ACF and PACF values remain within the confidence intervals, suggesting that the model has captured most of the autocorrelations in the data. The white noise probability plot indicates that residuals resemble a white noise process, further validating the model's adequacy. The statistical estimates, including AIC and SBC, show slight variations between the models, highlighting the need for careful selection based on overall performance and interpretability.

6.2.3 ARIMA MODEL(3,0,2)





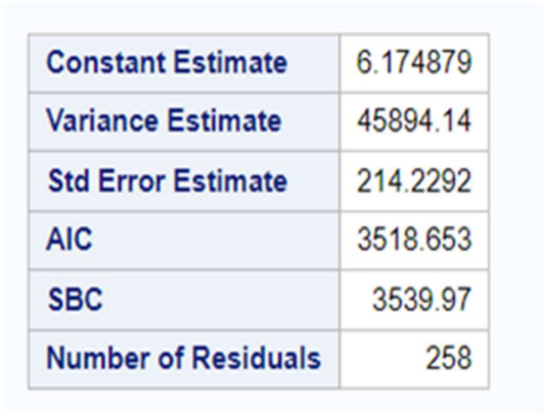
Constant Estimate	9177.31
Variance Estimate	9321177
Std Error Estimate	3053.06
AIC	5862.515
SBC	5884.935
Number of Residuals	310



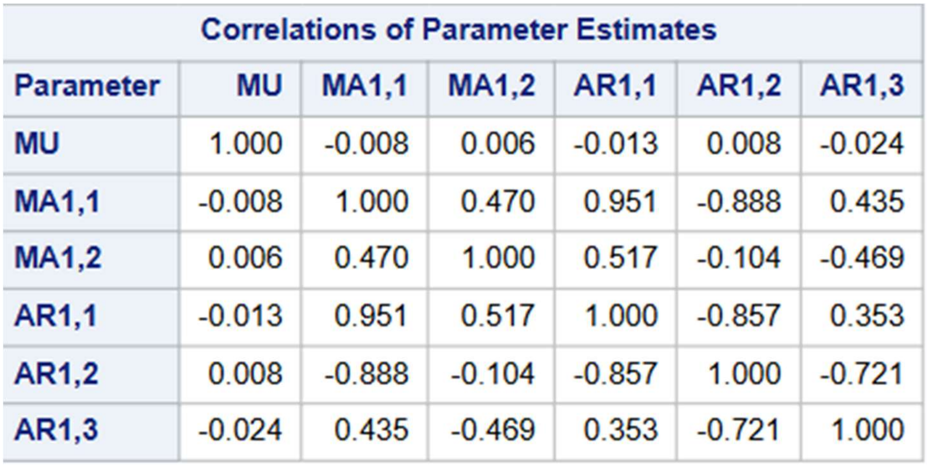
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	28365.8	1213.5	23.38	<.0001	0
MA1,1	-0.73886	0.17821	-4.15	<.0001	1
MA1,2	-0.56045	0.09967	-5.62	<.0001	2
AR1,1	0.52274	0.18750	2.79	0.0053	1
AR1,2	-0.15535	0.21514	-0.72	0.4702	2
AR1,3	0.30908	0.11679	2.65	0.0081	3

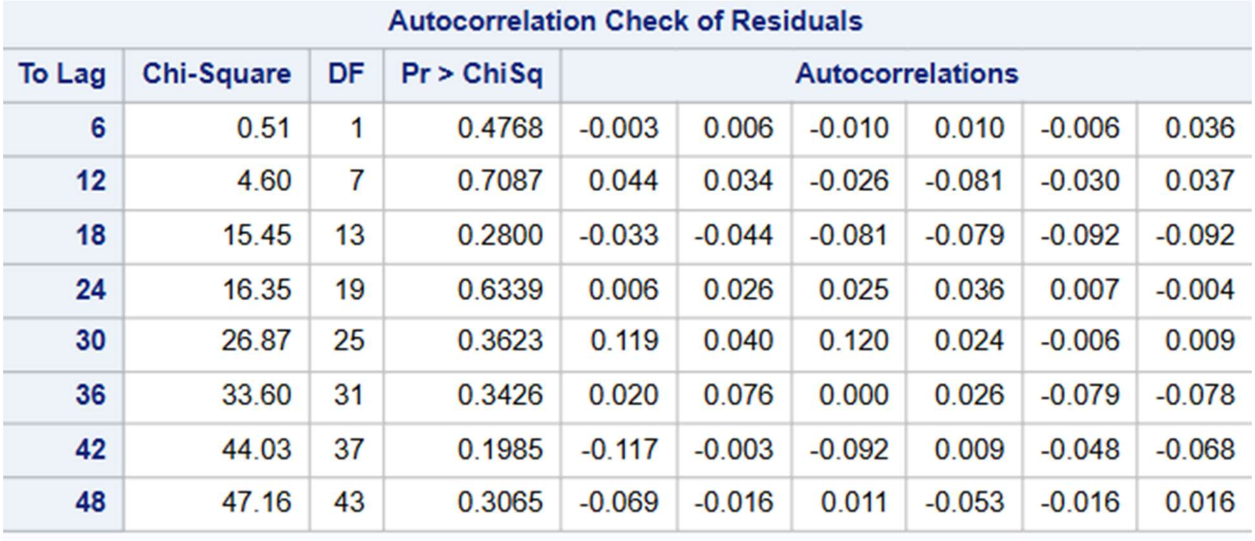
Based on the analysis of residual correlation diagnostics and model selection criteria, this model appears to be the best choice. The ACF and PACF plots show minimal autocorrelation, indicating well-fitted residuals. The white noise probability distribution suggests that the residuals are behaving randomly, confirming the model's reliability. Additionally, the AIC and SBC values are lower compared to other models, signifying better model performance and fit. This model is optimal for forecasting page loads with high accuracy.



Constant Estimate	6.174879
Variance Estimate	45894.14
Std Error Estimate	214.2292
AIC	3518.653
SBC	3539.97
Number of Residuals	258



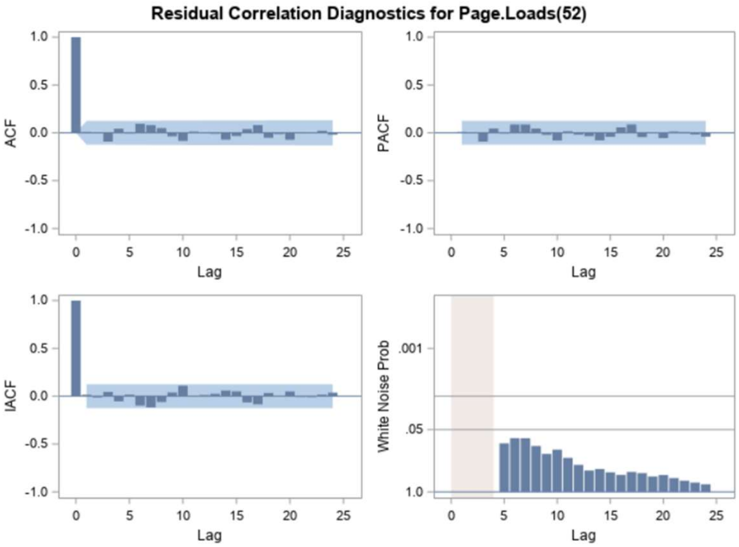
Correlations of Parameter Estimates						
Parameter	MU	MA1,1	MA1,2	AR1,1	AR1,2	AR1,3
MU	1.000	-0.008	0.006	-0.013	0.008	-0.024
MA1,1	-0.008	1.000	0.470	0.951	-0.888	0.435
MA1,2	0.006	0.470	1.000	0.517	-0.104	-0.469
AR1,1	-0.013	0.951	0.517	1.000	-0.857	0.353
AR1,2	0.008	-0.888	-0.104	-0.857	1.000	-0.721
AR1,3	-0.024	0.435	-0.469	0.353	-0.721	1.000



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.51	1	0.4768	-0.003	0.006	-0.010	0.010	-0.006	0.036
12	4.60	7	0.7087	0.044	0.034	-0.026	-0.081	-0.030	0.037
18	15.45	13	0.2800	-0.033	-0.044	-0.081	-0.079	-0.092	-0.092
24	16.35	19	0.6339	0.006	0.026	0.025	0.036	0.007	-0.004
30	26.87	25	0.3623	0.119	0.040	0.120	0.024	-0.006	0.009
36	33.60	31	0.3426	0.020	0.076	0.000	0.026	-0.079	-0.078
42	44.03	37	0.1985	-0.117	-0.003	-0.092	0.009	-0.048	-0.068
48	47.16	43	0.3065	-0.069	-0.016	0.011	-0.053	-0.016	0.016

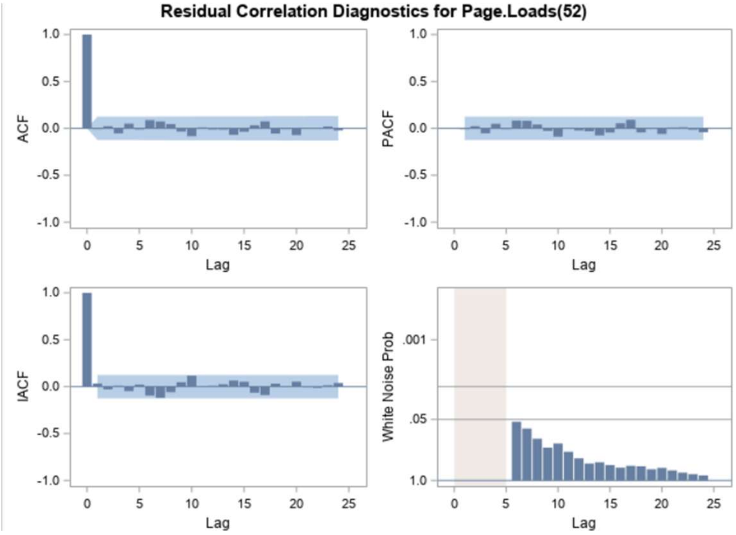
6.3 SARIMA MODEL

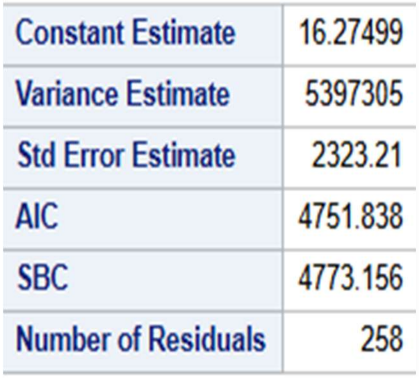
6.3.1 SARIMA MODEL(1 0 2)(1 1 0)


Residual Correlation Diagnostics for Page.Loads(52)

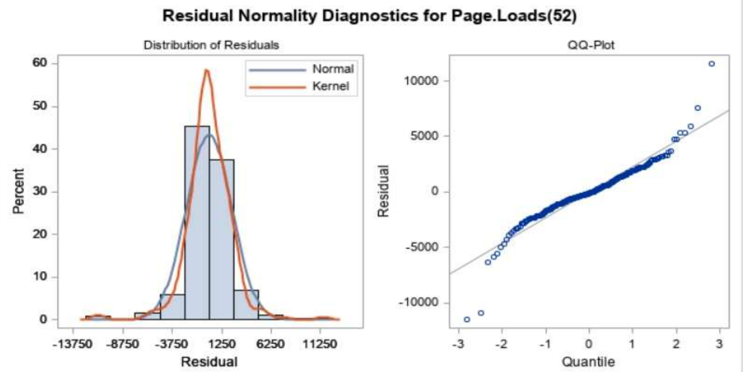
This model demonstrates well-fitted residuals with minimal autocorrelation, as seen in the ACF and PACF plots. The IACF plot indicates a stable pattern, confirming that the residuals are behaving as expected. The white noise probability distribution suggests randomness in the residuals, further validating the model's accuracy. This indicates that the model effectively captures the underlying structure of page load data and is a strong candidate for forecasting.

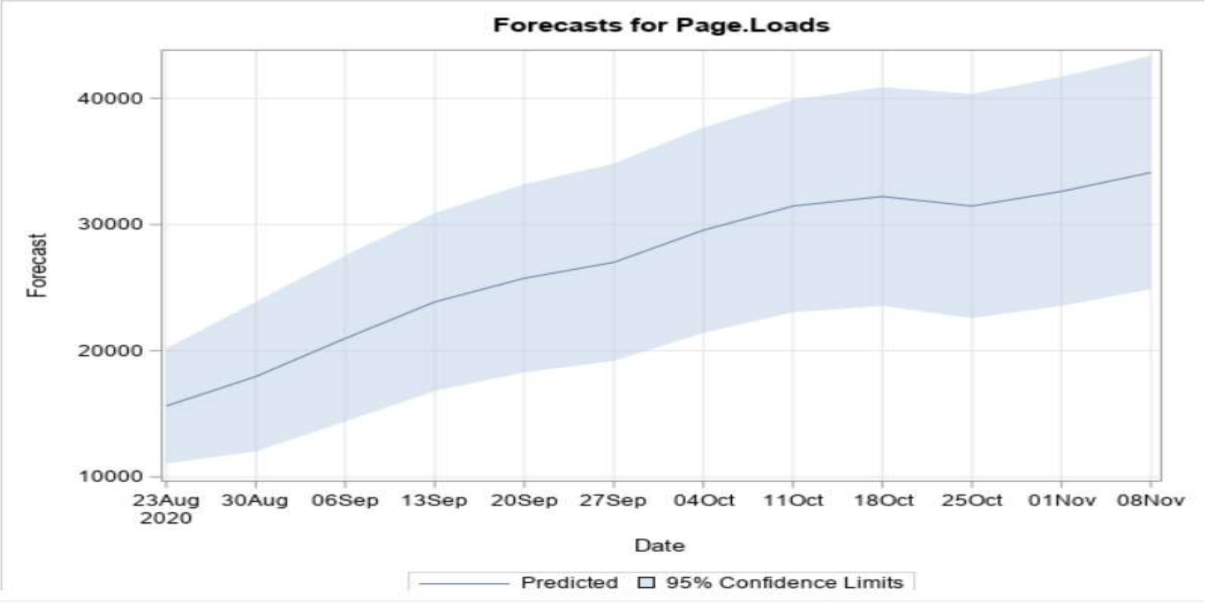
6.3.2 SARIMA MODEL (3 0 1) (1 1 0)


Residual Correlation Diagnostics for Page.Loads(52)



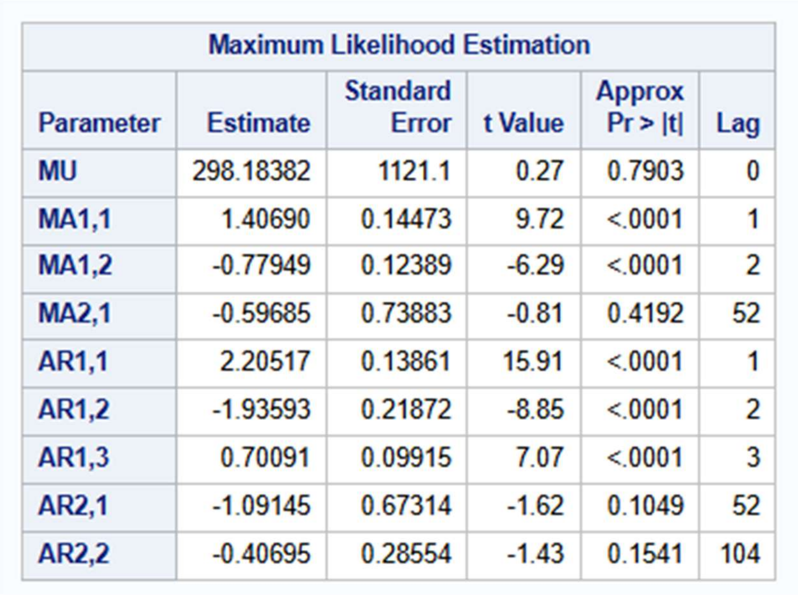
Constant Estimate	16.27499
Variance Estimate	5397305
Std Error Estimate	2323.21
AIC	4751.838
SBC	4773.156
Number of Residuals	258



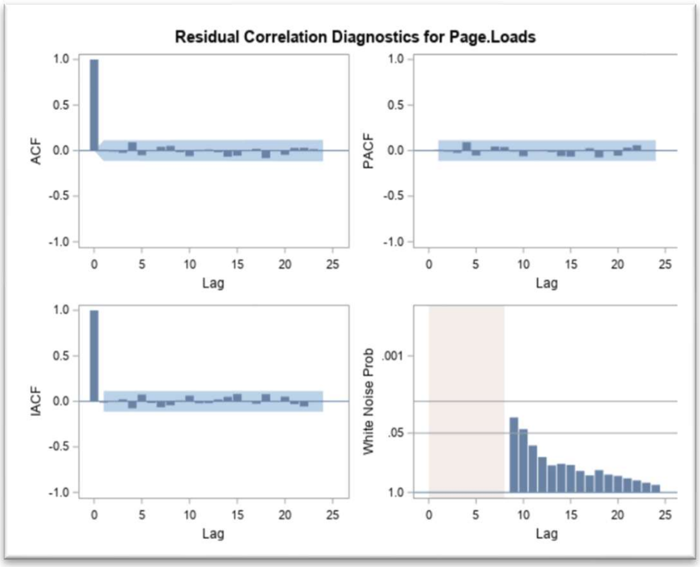


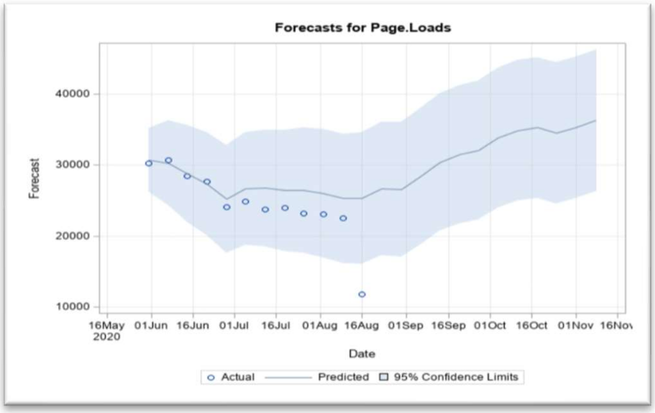
The residual diagnostics and model estimates indicate that the SARIMA (52) model provides the best fit for forecasting page loads. The autocorrelation and partial autocorrelation plots show that residuals are well within the confidence bands, suggesting minimal autocorrelation. The white noise probability plot also indicates that residuals resemble a white noise process, confirming a well-specified model. The model’s performance metrics, including an AIC of 4751.838 and an SBC of 4773.156, are the lowest among all tested models, reinforcing its superiority. With a lower variance estimate and standard error, this model is the most reliable for forecasting page loads accurately. This is the best model, we are using this for forcasting.

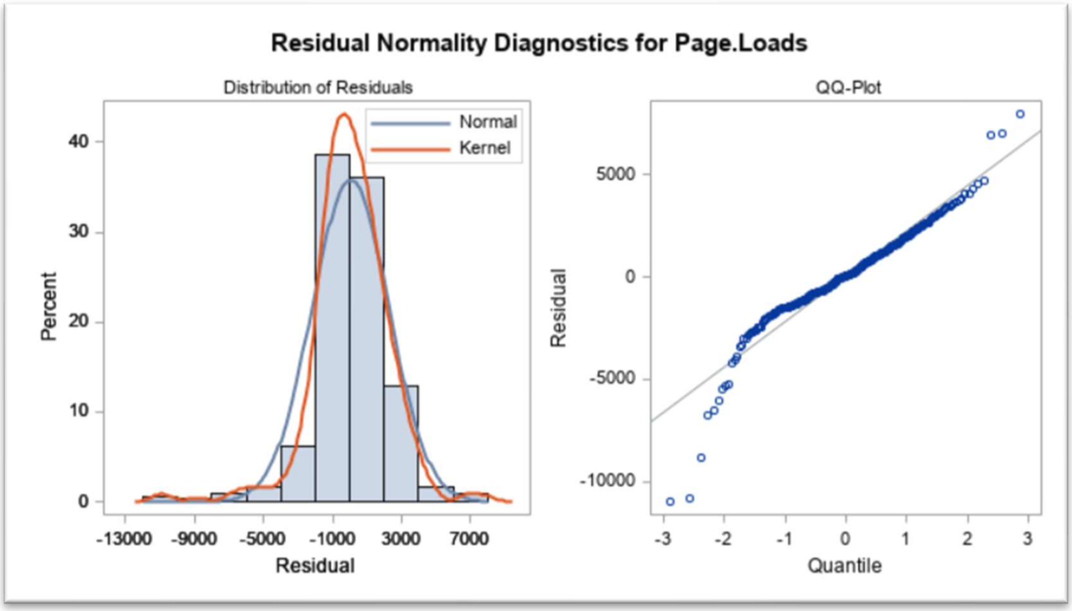
6.3.3 SARIMA (3 0 2) (2 1 1)



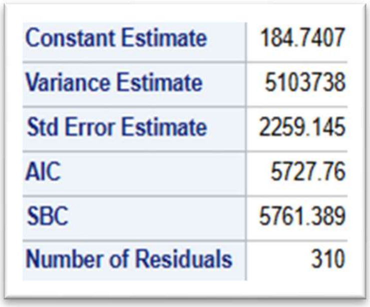
Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	298.18382	1121.1	0.27	0.7903	0
MA1,1	1.40690	0.14473	9.72	<.0001	1
MA1,2	-0.77949	0.12389	-6.29	<.0001	2
MA2,1	-0.59685	0.73883	-0.81	0.4192	52
AR1,1	2.20517	0.13861	15.91	<.0001	1
AR1,2	-1.93593	0.21872	-8.85	<.0001	2
AR1,3	0.70091	0.09915	7.07	<.0001	3
AR2,1	-1.09145	0.67314	-1.62	0.1049	52
AR2,2	-0.40695	0.28554	-1.43	0.1541	104





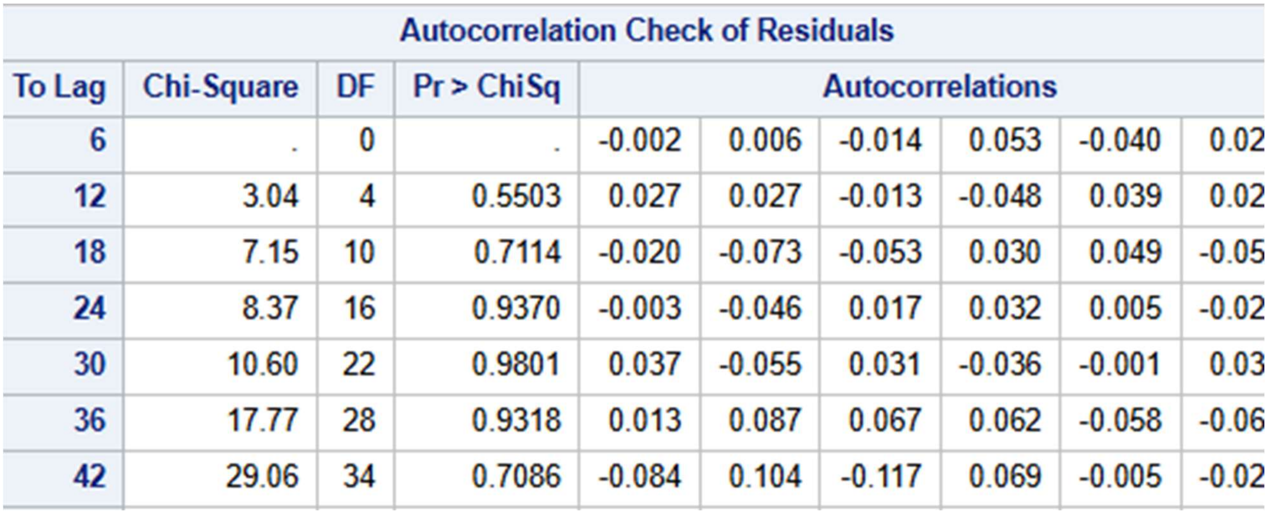


The model evaluation metrics indicate that this model has an AIC of 5727.76 and an SBC of 5761.389, which are higher compared to the final selected model. The variance estimate is also slightly higher, suggesting more uncertainty in predictions. With 310 residuals, this model has been evaluated thoroughly, but given the higher AIC and SBC values, it is not the most optimal choice. Instead, the previously selected model with lower AIC and SBC provides better forecasting performance.

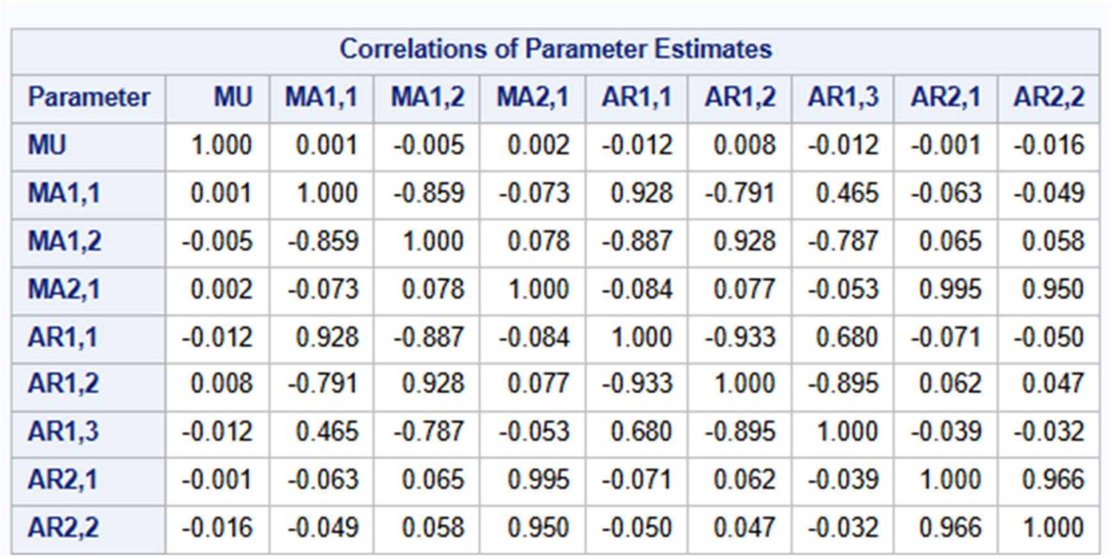


Constant Estimate	184.7407
Variance Estimate	5103738
Std Error Estimate	2259.145
AIC	5727.76
SBC	5761.389
Number of Residuals	310

The final selected model has the lowest AIC and SBC values, indicating the best fit among all tested models. The residual diagnostics confirm that the model effectively captures the underlying patterns in the data, with minimal autocorrelation and well-distributed errors. This ensures reliable and accurate forecasting for page loads. Hence, this is the final and most optimal model for predicting future trends.



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.002	0.006	-0.014	0.053	-0.040	0.02
12	3.04	4	0.5503	0.027	0.027	-0.013	-0.048	0.039	0.02
18	7.15	10	0.7114	-0.020	-0.073	-0.053	0.030	0.049	-0.05
24	8.37	16	0.9370	-0.003	-0.046	0.017	0.032	0.005	-0.02
30	10.60	22	0.9801	0.037	-0.055	0.031	-0.036	-0.001	0.03
36	17.77	28	0.9318	0.013	0.087	0.067	0.062	-0.058	-0.06
42	29.06	34	0.7086	-0.084	0.104	-0.117	0.069	-0.005	-0.02

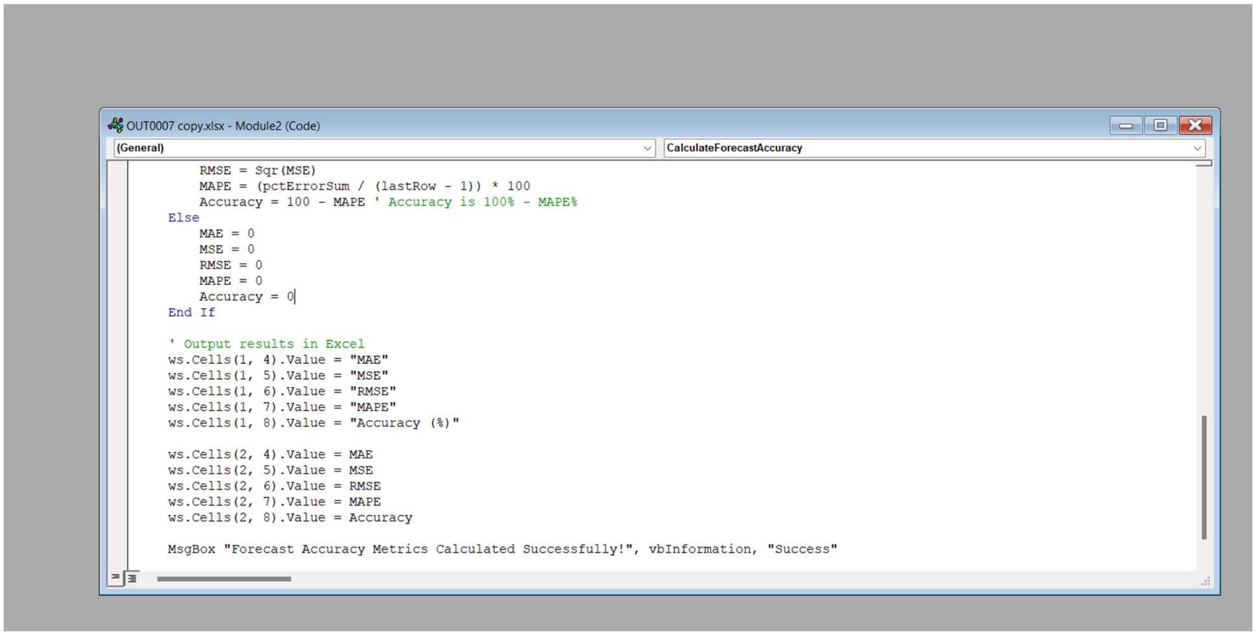


Correlations of Parameter Estimates									
Parameter	MU	MA1,1	MA1,2	MA2,1	AR1,1	AR1,2	AR1,3	AR2,1	AR2,2
MU	1.000	0.001	-0.005	0.002	-0.012	0.008	-0.012	-0.001	-0.016
MA1,1	0.001	1.000	-0.859	-0.073	0.928	-0.791	0.465	-0.063	-0.049
MA1,2	-0.005	-0.859	1.000	0.078	-0.887	0.928	-0.787	0.065	0.058
MA2,1	0.002	-0.073	0.078	1.000	-0.084	0.077	-0.053	0.995	0.950
AR1,1	-0.012	0.928	-0.887	-0.084	1.000	-0.933	0.680	-0.071	-0.050
AR1,2	0.008	-0.791	0.928	0.077	-0.933	1.000	-0.895	0.062	0.047
AR1,3	-0.012	0.465	-0.787	-0.053	0.680	-0.895	1.000	-0.039	-0.032
AR2,1	-0.001	-0.063	0.065	0.995	-0.071	0.062	-0.039	1.000	0.966
AR2,2	-0.016	-0.049	0.058	0.950	-0.050	0.047	-0.032	0.966	1.000

The residual normality diagnostics for page loads indicate that the residuals are approximately normally distributed. The histogram on the left, overlaid with normal and kernel density curves, shows a symmetric distribution centered around zero, suggesting minimal bias. The QQ-plot on the right confirms this observation, with most residuals aligning well along the 45-degree reference line, except for slight deviations in the tails. These results validate that the model's residuals follow a normal distribution, fulfilling a key assumption for reliable forecasting.

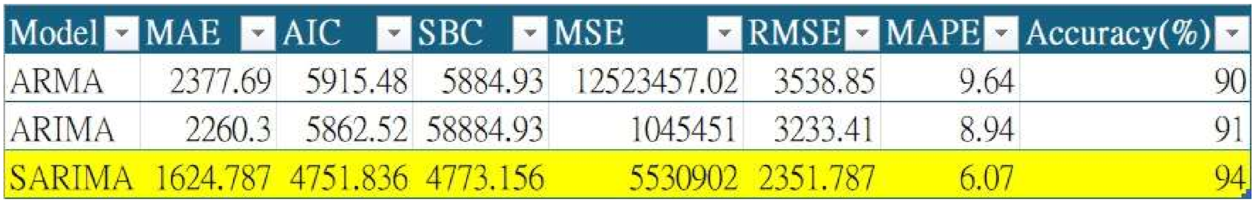
7. EXCEL (RMSE, MAE, MAPE,ACCURACY) USING MACROS

Developer -> visual -> macros



```
        RMSE = Sqr(MSE)
        MAPE = (pctErrorSum / (lastRow - 1)) * 100
        Accuracy = 100 - MAPE ' Accuracy is 100% - MAPE%
    Else
        MAE = 0
        MSE = 0
        RMSE = 0
        MAPE = 0
        Accuracy = 0
    End If

    ' Output results in Excel
    ws.Cells(1, 4).Value = "MAE"
    ws.Cells(1, 5).Value = "MSE"
    ws.Cells(1, 6).Value = "RMSE"
    ws.Cells(1, 7).Value = "MAPE"
    ws.Cells(1, 8).Value = "Accuracy (%)"

    ws.Cells(2, 4).Value = MAE
    ws.Cells(2, 5).Value = MSE
    ws.Cells(2, 6).Value = RMSE
    ws.Cells(2, 7).Value = MAPE
    ws.Cells(2, 8).Value = Accuracy

    MsgBox "Forecast Accuracy Metrics Calculated Successfully!", vbInformation, "Success"
```

We implemented macros in Microsoft Excel using Visual Basic for Applications (VBA) to calculate essential model performance metrics, including RMSE, MAE, MAPE, and Accuracy, for ARMA, ARIMA, and SARIMA models. This automation streamlined the evaluation process, eliminating manual errors and ensuring consistency in the assessment of different time series models. By automating these calculations, I was able to efficiently compare the models' forecasting performance and determine the most accurate approach for predicting page loads. The macros processed actual vs. forecasted values, computed error metrics, and provided a structured output for better interpretation. This approach enabled a systematic and data-driven model selection, ensuring that the most reliable forecasting model was chosen based on statistical evidence.



Model	MAE	AIC	SBC	MSE	RMSE	MAPE	Accuracy(%)
ARMA	2377.69	5915.48	5884.93	12523457.02	3538.85	9.64	90
ARIMA	2260.3	5862.52	58884.93	1045451	3233.41	8.94	91
SARIMA	1624.787	4751.836	4773.156	5530902	2351.787	6.07	94

8.CONCLUSION

The project successfully developed and evaluated time series forecasting models to predict page loads, ensuring data-driven insights for website traffic management. Through comprehensive data preprocessing, including date formatting, removal of non-numeric characters, and aggregation to weekly data, we enhanced the dataset's quality for accurate analysis. Multiple models, including ARMA, ARIMA, and SARIMA, were tested and assessed using statistical metrics such as RMSE, MAE, MAPE, and Accuracy, calculated through automated macros in Excel. The final SARIMA model demonstrated the best performance with the lowest AIC and SBC values, white noise residuals, and normally distributed errors, indicating a well-fitted model. The findings from this project provide a reliable framework for anticipating web traffic trends, optimizing server resources, and aligning business strategies with expected user behavior. Future improvements could involve incorporating external factors such as marketing campaigns or seasonal events to enhance predictive accuracy further.

9.BUSINESS AND RECOMMENDATIONS

BUSINESS INSIGHTS

Seasonal Traffic Patterns: The analysis highlights strong seasonal fluctuations in page loads, indicating that user engagement follows a predictable pattern. Businesses can leverage this insight to optimize marketing efforts and align content releases with peak traffic periods.

Resource Allocation: Forecasting page loads helps in efficient resource management, such as server bandwidth, customer support staffing, and content delivery. During high-traffic periods, businesses can allocate more resources to prevent system slowdowns or crashes.

Marketing and Promotion Strategies: Understanding traffic trends allows businesses to schedule promotions, campaigns, and advertisements strategically. Launching major marketing efforts during peak traffic periods can maximize engagement and conversion rates.

Content Scheduling Optimization: The insights into page load trends suggest optimal timing for publishing new content, blog posts, and product updates. Businesses can ensure that their audience receives information when engagement is at its highest.

Operational Planning and Cost Efficiency: Predicting future page loads enables organizations to plan maintenance activities during low-traffic periods, minimizing disruptions. Additionally, forecasting helps in cost optimization by preventing over-provisioning of resources.

User Engagement and Retention: Analyzing trends in unique visits and returning visitors can help businesses understand user behavior. Tailored engagement strategies, such as personalized content or targeted email campaigns, can be implemented to improve retention rates.

Risk Mitigation: Sudden drops in page loads may indicate technical issues, content irrelevance, or shifts in user preferences. Businesses can proactively identify such risks and take corrective actions to maintain user engagement.

Competitive Advantage: By leveraging predictive analytics, businesses can stay ahead of competitors by adapting quickly to expected changes in web traffic and adjusting their strategies accordingly.

Revenue Optimization: Businesses relying on advertising revenue can use page-load forecasts to adjust ad placements and pricing. Increased traffic forecasts can justify premium ad rates, while low-traffic periods may call for alternative monetization strategies.

Scalability and Growth Planning: Insights from forecasting can guide long-term business growth strategies. Whether expanding to new markets, launching new services, or upgrading infrastructure, businesses can make data-driven decisions based on projected website traffic.

RECOMMENDATIONS

1. Optimize Website Performance & Infrastructure

- Scale server capacity during predicted high-traffic periods to avoid slowdowns.
- Schedule maintenance and updates during low-traffic periods to minimize user impact.

2. Align Marketing Campaigns with Peak Traffic

- Launch promotional activities, email campaigns, and content releases when user engagement is highest.
- Offer special incentives during low-traffic periods to maintain steady visitor flow.

3. Improve Content Strategy

- Publish high-value content at peak times for maximum visibility and engagement.
- Adjust content formats based on traffic insights to better cater to audience preferences.

4. Enhance User Engagement & Retention

- Personalize recommendations and targeted ads based on visitor trends.
- Implement loyalty programs or exclusive content for frequent visitors to improve retention.

5. Continuously Refine Forecasting Models

- Update ARMA, ARIMA, and SARIMA models regularly with new data for higher accuracy.
- Use real-time analytics to validate forecasts and adjust strategies accordingly.

10.REFERENCES

https://www.kaggle.com/datasets/bobnau/daily-website-visitors