Driving Ad Performance

Insights from Exploratory Data Analysis and Time Series Forecasting

Date: 25 May 2023

Presented by:

Sakshi Shende

https://github.com/sakshi-shende/Time-Series-Analysis

What are you looking for?

Introduction

Dataset Exploration

Exploratory Data Analysis

Time Series Forecasting Techniques

Model Evaluation

Conclusion and Future Work

Introduction

PROBLEM STATEMENT

- Accurately Predicting Wikipedia Page Views for Effective Ad Placement

ANALYSIS OBJECTIVES

- Optimize ad placement for clients based on data-driven insights
- Forecast page views to predict ad performance and allocate resources effectively
- Provide actionable recommendations for maximizing clicks and minimizing costs

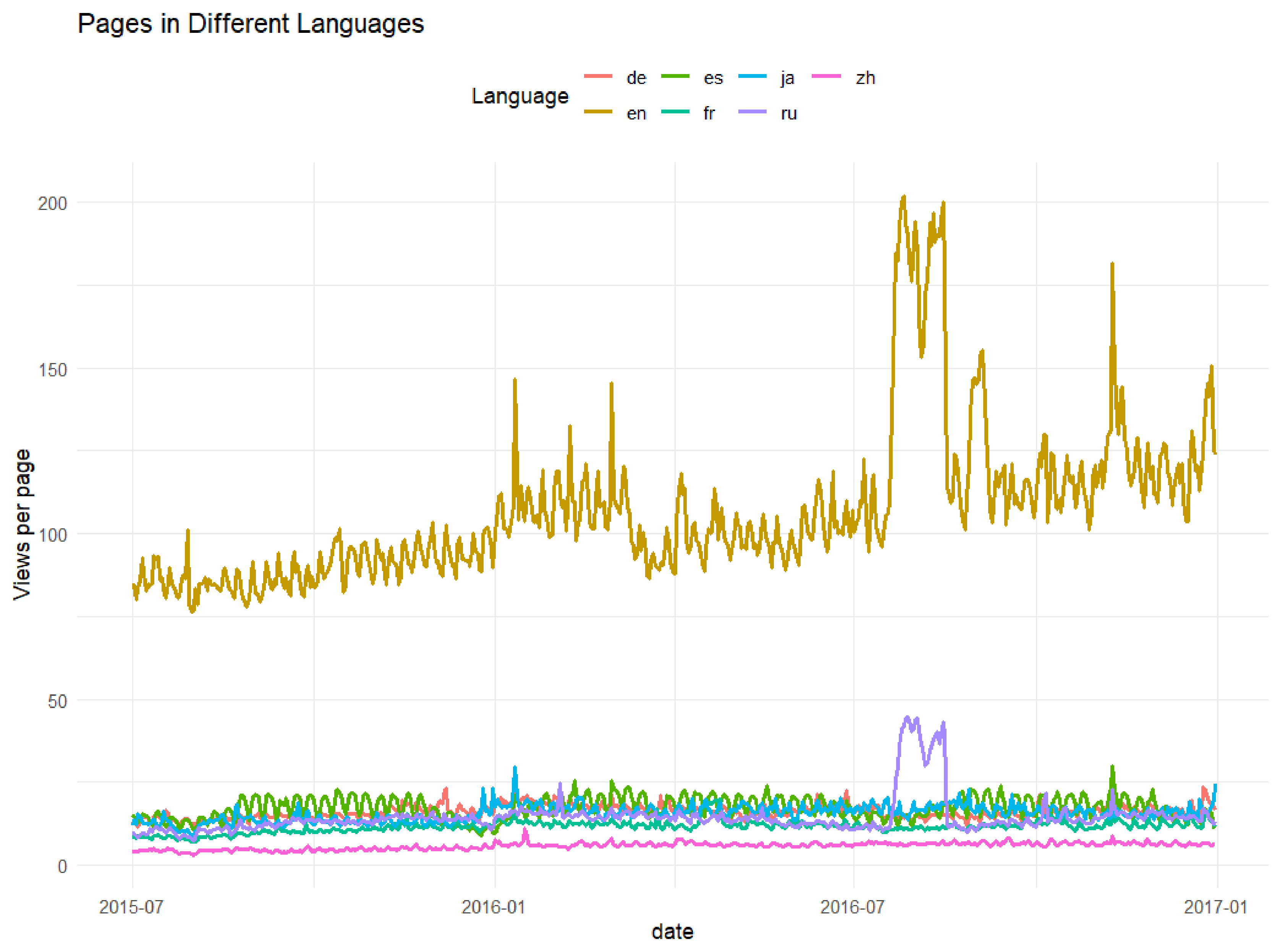
BENEFITS OF ANALYSIS RESULTS

- Targeted advertising based on language and region-specific insights
- Efficient allocation of resources by forecasting page views
- Minimization of ad spending while maximizing clicks and conversions

ASSUMPTIONS

- The dataset is representative and provides sufficient variability.
- The time series data exhibit temporal dependencies and trends.
- The selected models and metrics are appropriate for forecasting and evaluation.

Dataset Exploration



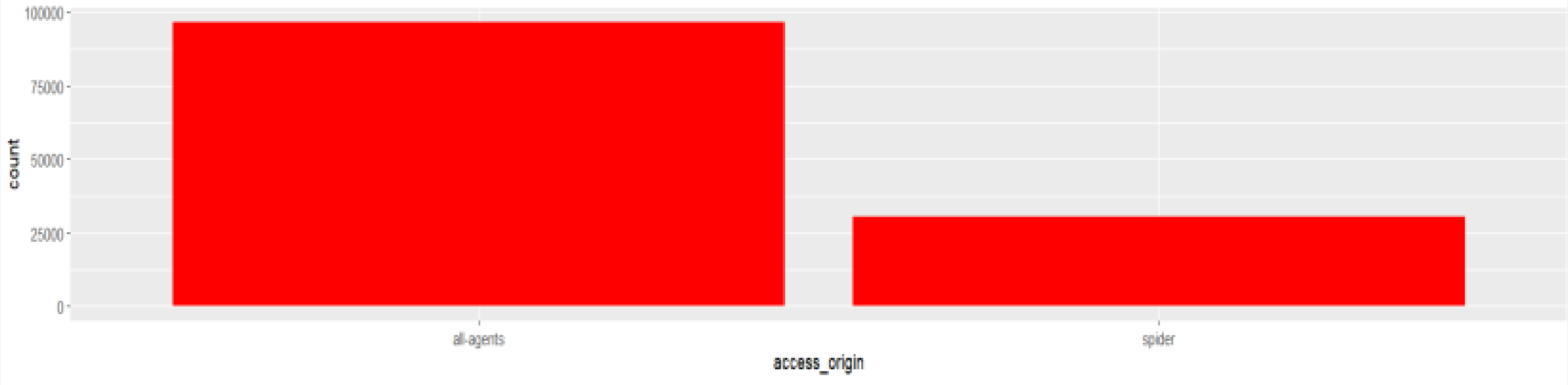
145K Wikipedia Pages

Frequency: Daily

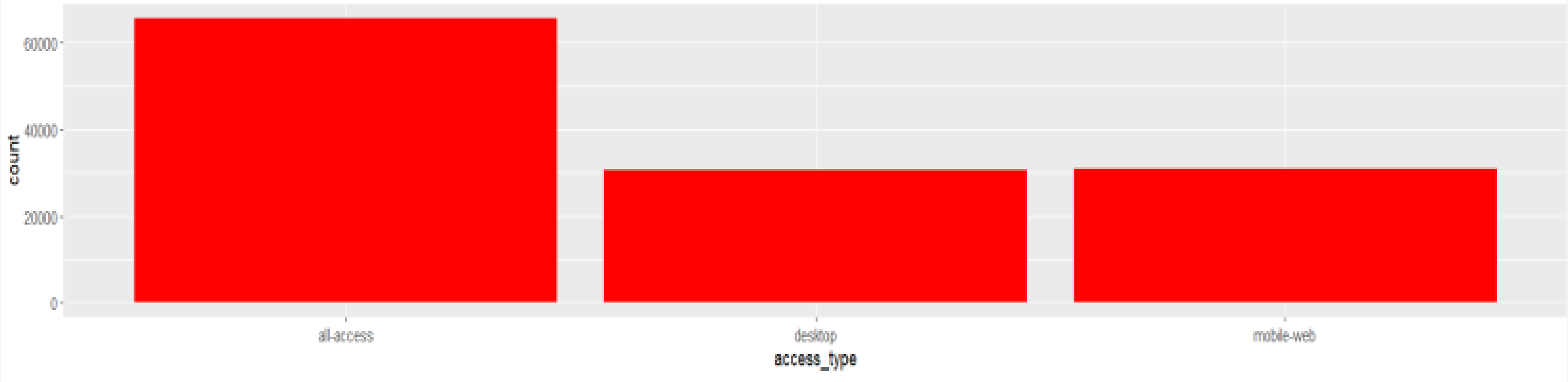
Period: July 1st 2015 – December 31st 2016

Language: 7

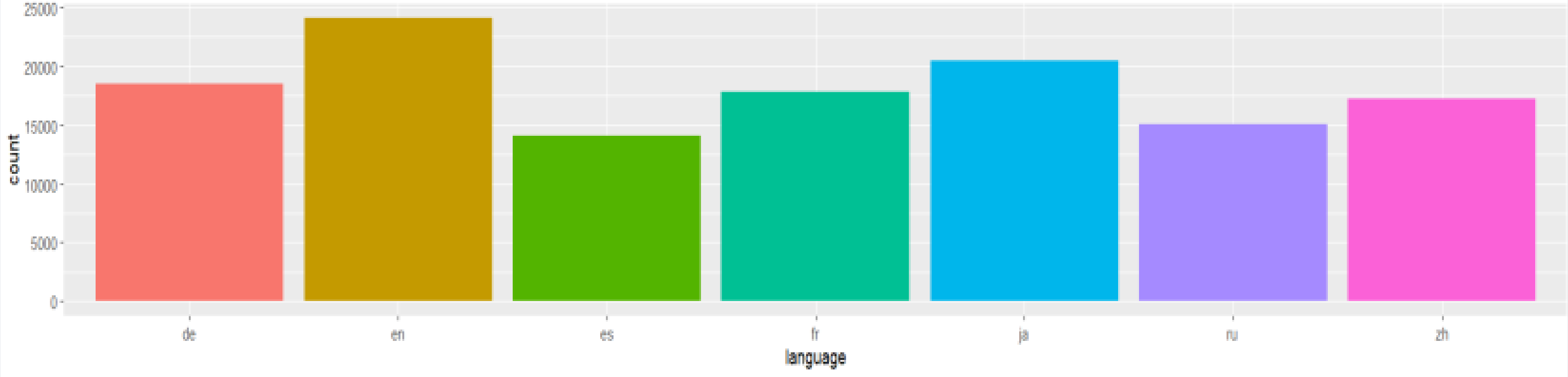
Dataset Exploration



- Plot display distribution of page views across access origin
- Observed that the count of "all-agents" is higher compared to "spider“
- Focus on optimizing ad placement for real users, to increase the likelihood of higher click-through rates and conversions, helping to achieve the goal of maximizing clicks at minimum cost.



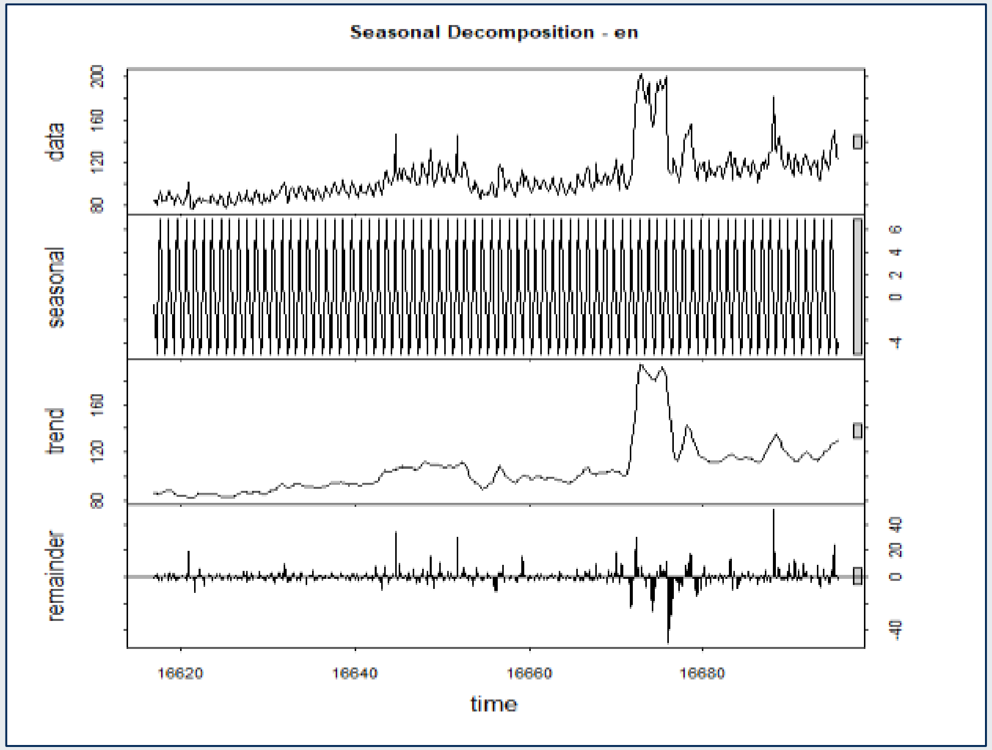
- Plot displays the distribution of page views across access types
- Majority of page views are coming from various access types combined, indicating a diverse user base across different platforms
- Optimize the ad spending by strategically tailoring the campaigns to the most effective access types



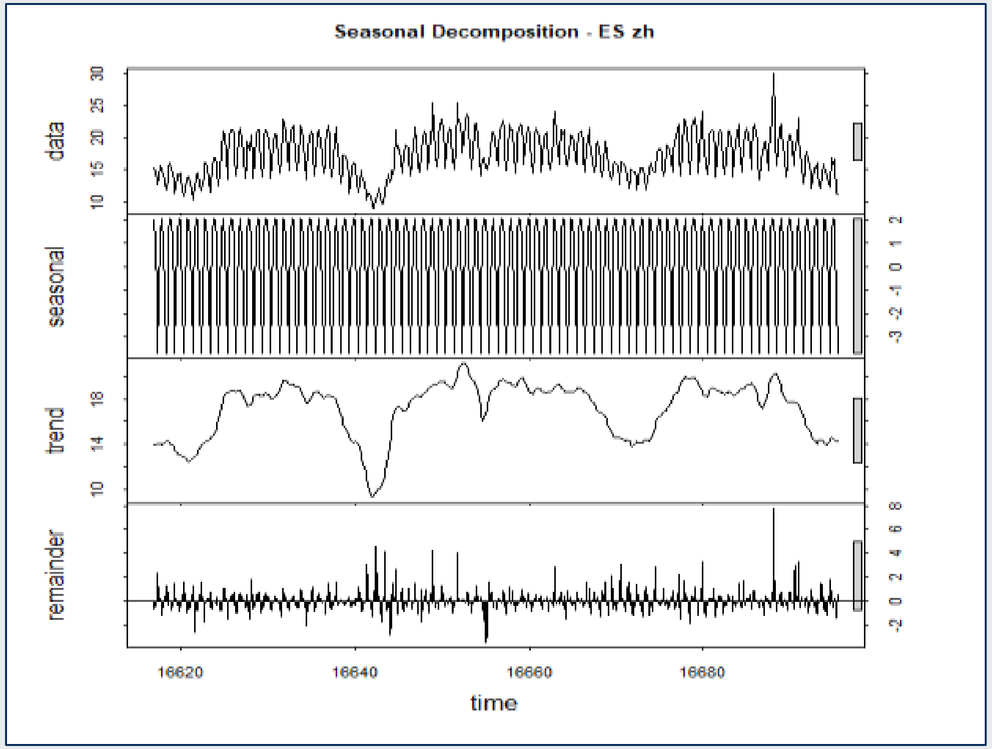
- Plot displays the distribution of page views across different languages
- English has the highest number of views, indicating a larger user base and potential reach for ad campaigns.
- Tailor the advertising campaigns to specific language-speaking audiences, ensuring targeted messaging and higher engagement.

Exploratory Data Analysis

Applied Seasonal Decomposition to separate the time series into three components: trend, seasonal, and residual to the underlying patterns and behaviours of the data.



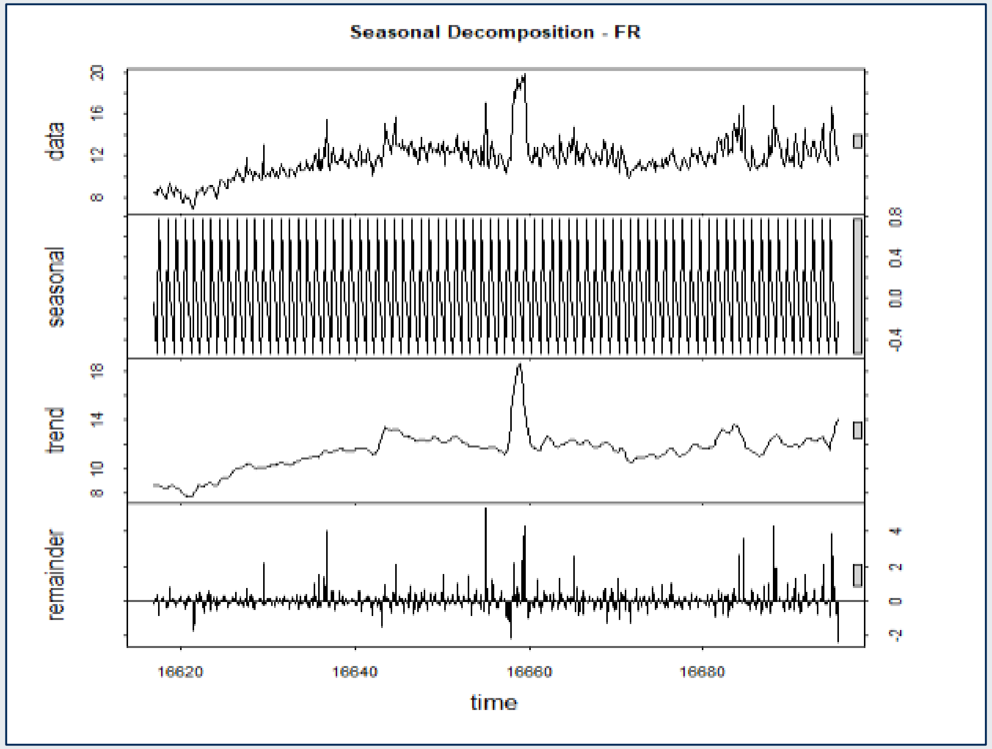
The above graph shows the seasonal decomposition of the English Wikipedia page. We can observe a sharp increase in views trend in August 2016 and drop drastically.



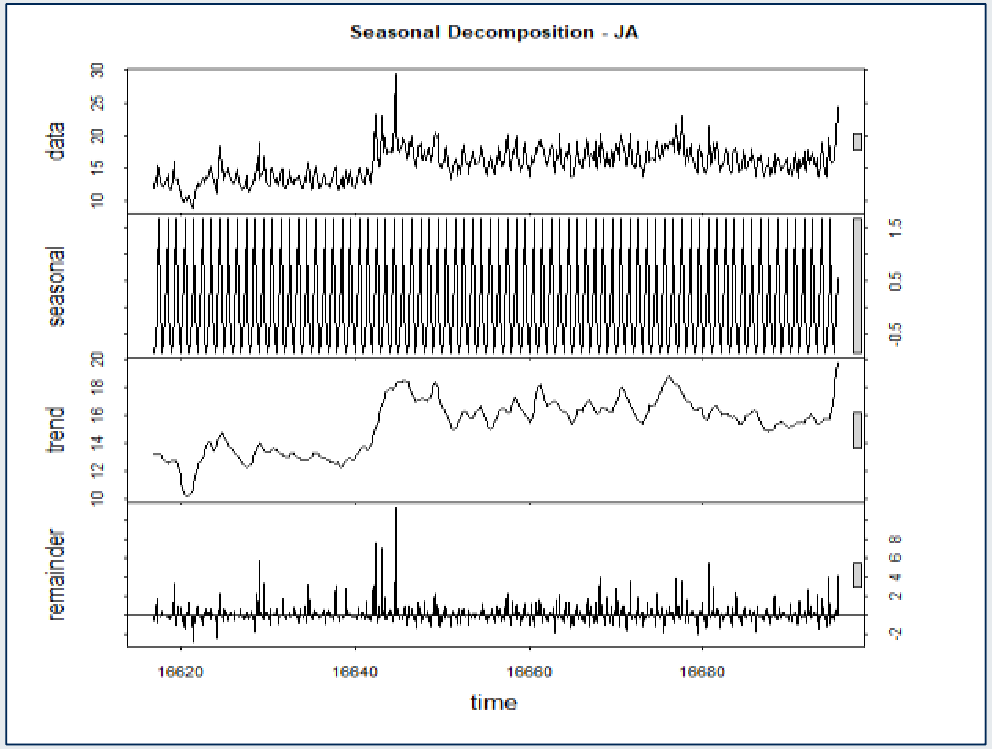
The above graph shows the seasonal decomposition of the Spanish Wikipedia page. We can observe a sharp decrease in the views trend from the end of December 2015.

Exploratory Data Analysis

Applied Seasonal Decomposition to separate the time series into three components: trend, seasonal, and residual to the underlying patterns and behaviors of the data.



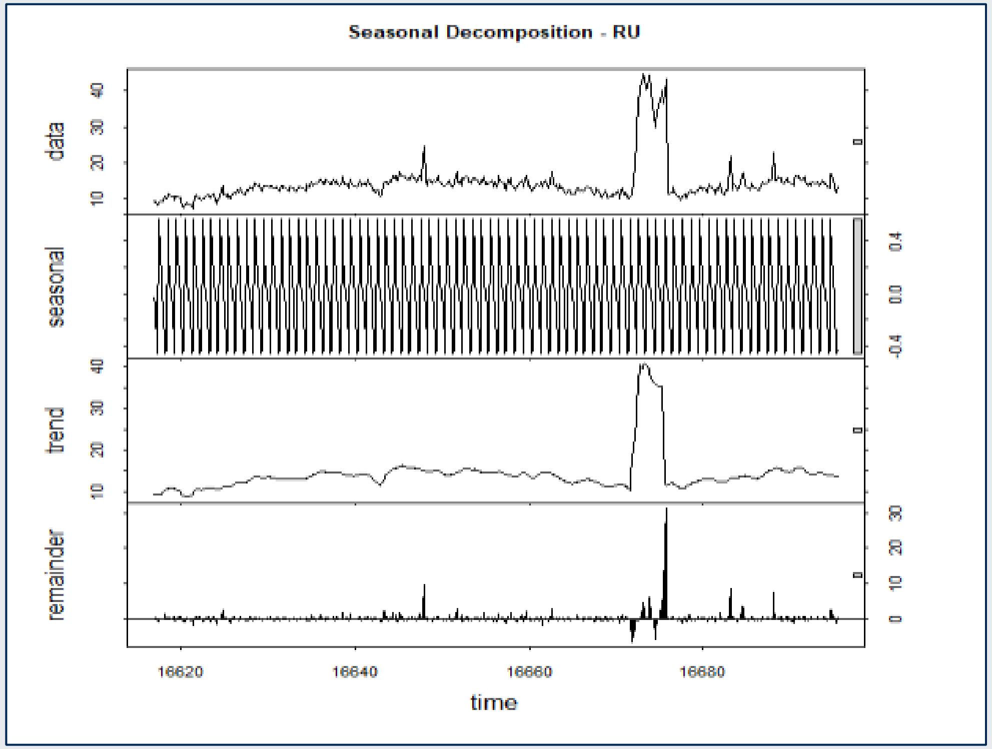
The above graph shows the seasonal decomposition of the French Wikipedia page. We can observe a sharp increase in views trend in between April 2016 and May 2016.



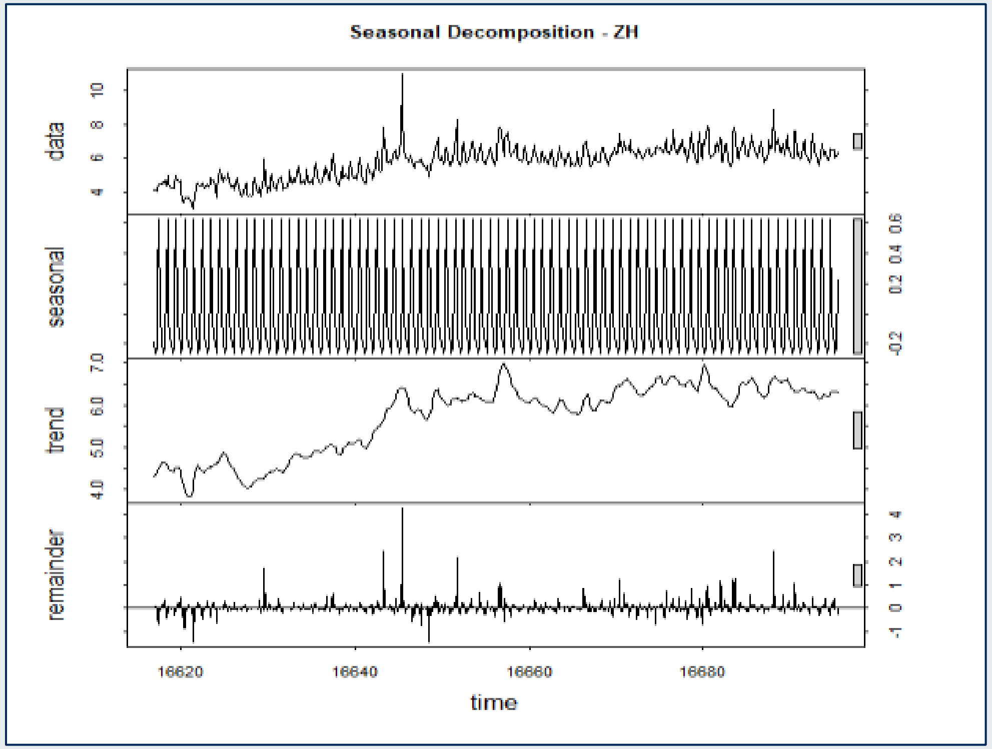
The above graph shows the seasonal decomposition of the Japanese Wikipedia page. We can observe increasing page views trends from January 2016.

Exploratory Data Analysis

Applied Seasonal Decomposition to separate the time series into three components: trend, seasonal, and residual to the underlying patterns and behaviors of the data.



The above graph shows the seasonal decomposition of the Russian Wikipedia page. We can observe a sharp increase in views trend in July 2016 to September 2016 and drop drastically.

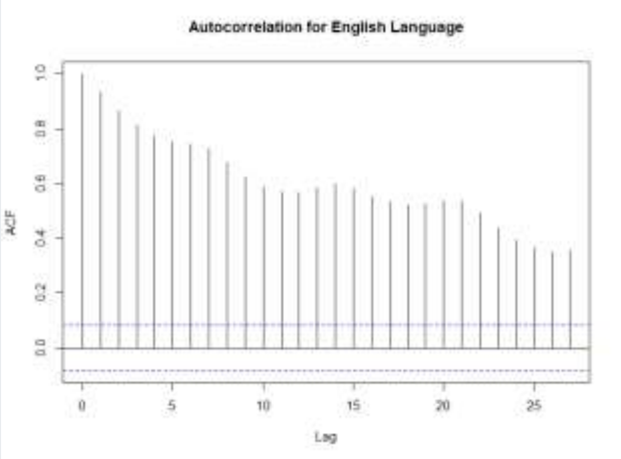


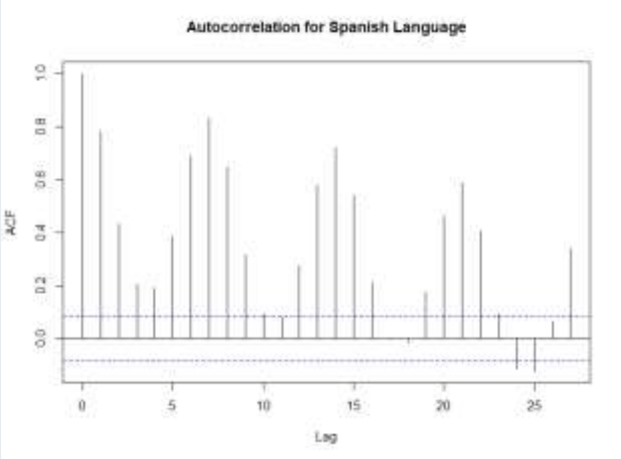
The above graph shows the seasonal decomposition of the Chinese Wikipedia page. We can observe a gradual increase in page view trend.

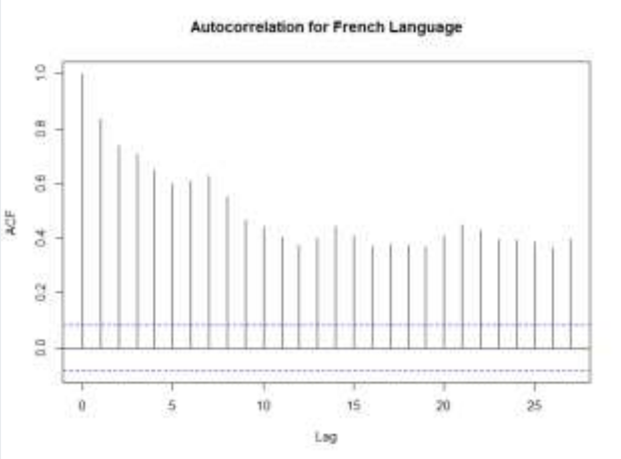
Data - Stationarity

Utilized ACF plot to evaluate the stationarity of page view time series data for different languages.

Observed non-stationarity in the time series

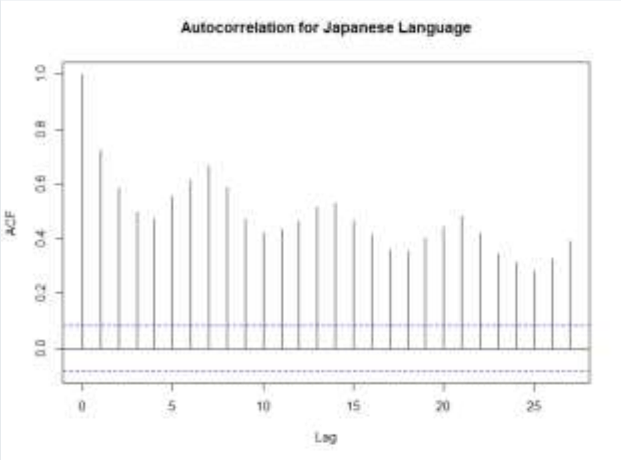


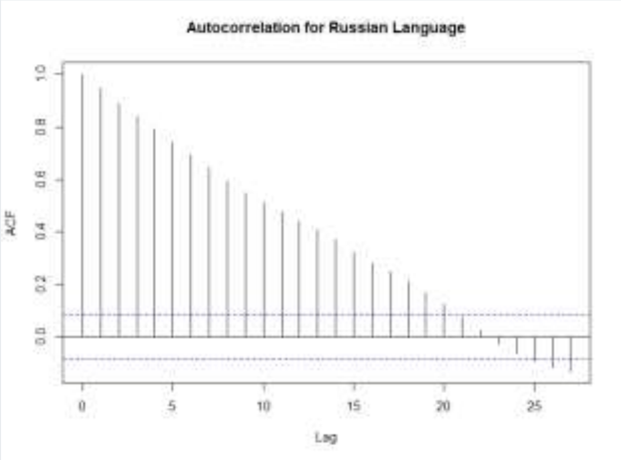


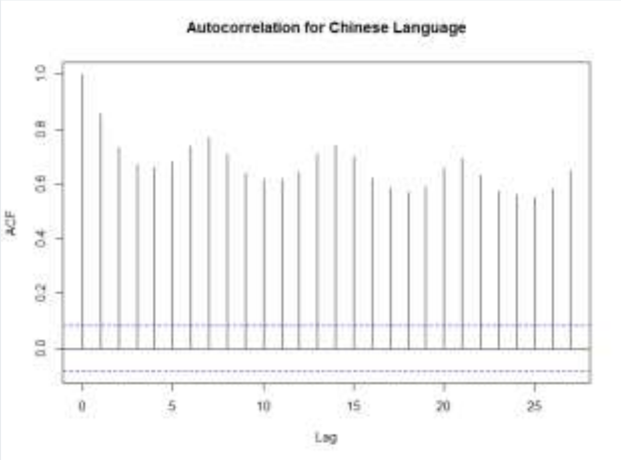


Observation

- ACF plots demonstrate non-stationarity for all language-specific page view time series
- Indicates the presence of trends, seasonality, or other factors affecting stationarity







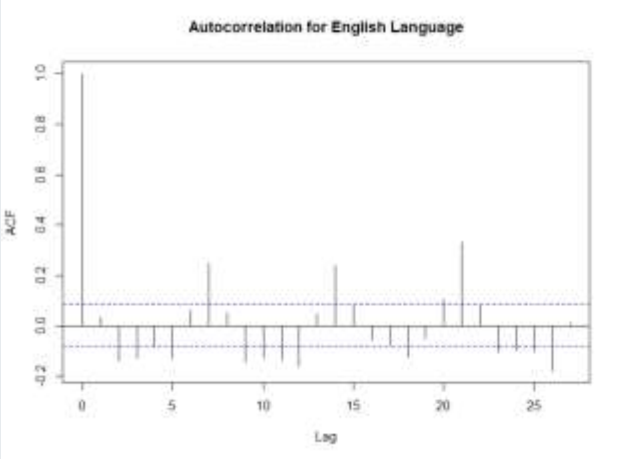
Implication

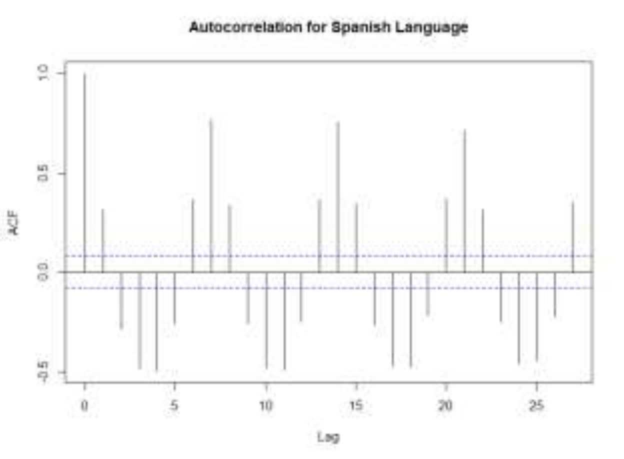
- Non-stationary time series require appropriate transformations for accurate forecasting
- Methods such as differencing or decomposition can be employed to achieve stationarity

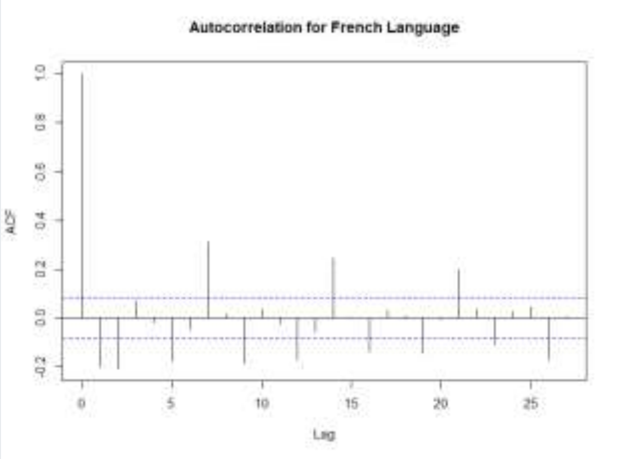
Note: The slide only shows the ACF plots of 6 languages. The 7th language, German, was also found to be non-stationary. Please see the reference for more details.

Data - Differencing

Applied first-order differencing technique to the language-specific page view time series to convert non-stationary data into a stationary form

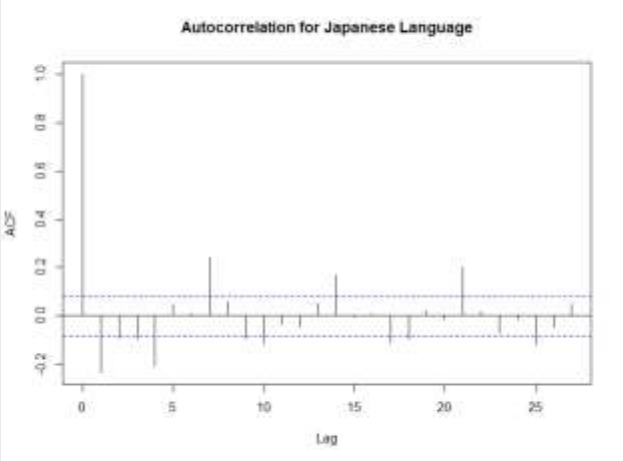

Autocorrelation for English Language

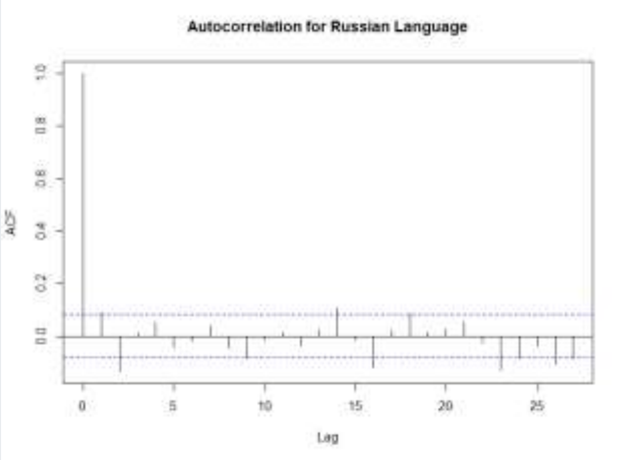

Autocorrelation for Spanish Language

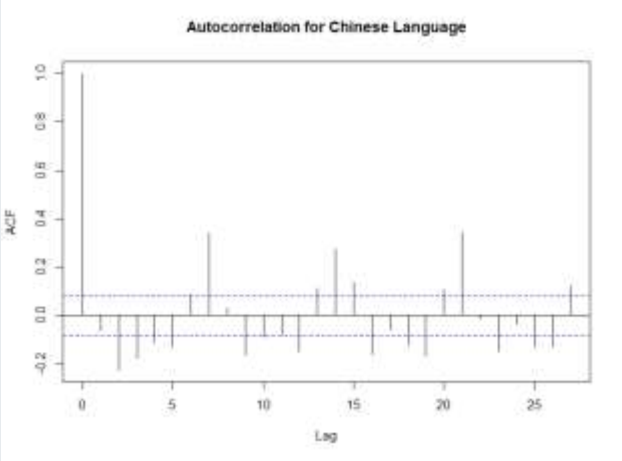

Autocorrelation for French Language

Methodology

- First Order Differencing: Calculation of the difference between consecutive observations in the time series
- Removes trends and seasonality to achieve stationarity


Autocorrelation for Japanese Language


Autocorrelation for Russian Language


Autocorrelation for Chinese Language

Impact on Stationarity:

- First-order differencing effectively eliminated the non-stationarity of the language-specific page view time series
- Resulted in a stationary time series suitable for further analysis and forecasting

Note: The 7th language, German, is also converted to non-stationary by applying first-order differencing. Please see the reference for more details.

Time Series Forecasting Techniques

Split the pre-processed data into training and testing subsets

Train Data: Up to December 16, 2016, | Test Data: From December 17, 2016, onwards.

ARIMA
AutoRegressive
Integrated Moving Average

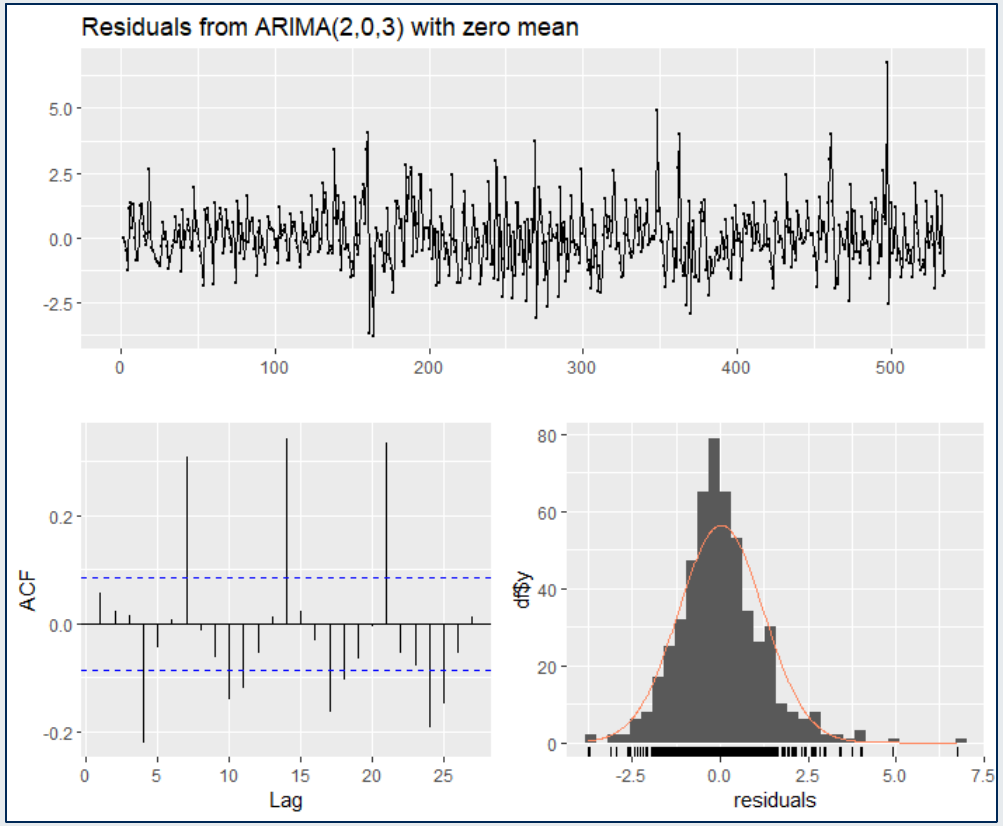
Justification

- ARIMA model suitable for forecasting Wikipedia page views with stationary properties
- Enables understanding and predicting the overall trend and seasonality of the page view time series

Trade-Offs

- Assumes linear relationships and may not capture complex non-linear patterns
- Requires stationarity transformation and appropriate selection of model parameters (p, d, q)

The auto.arima() function is used to fit the ARIMA model on the data we got after applying the first-order differencing. To select the best ARIMA model based on the provided data, I have applied an exhaustive search.



The plot shows the residuals for the ARIMA model with order (2, 0, 3). We can observe that the residuals are normally distributed for the page views of German Language.

Note: Applied auto_arima() with same configuration for other 6 languages. Please see the appendix for more details.

Time Series Forecasting Techniques

Split the pre-processed data into training and testing subsets

Train Data: Up to December 16, 2016, | Test Data: From December 17, 2016, onwards.

SARIMA

Seasonal AutoRegressive Integrated Moving Average

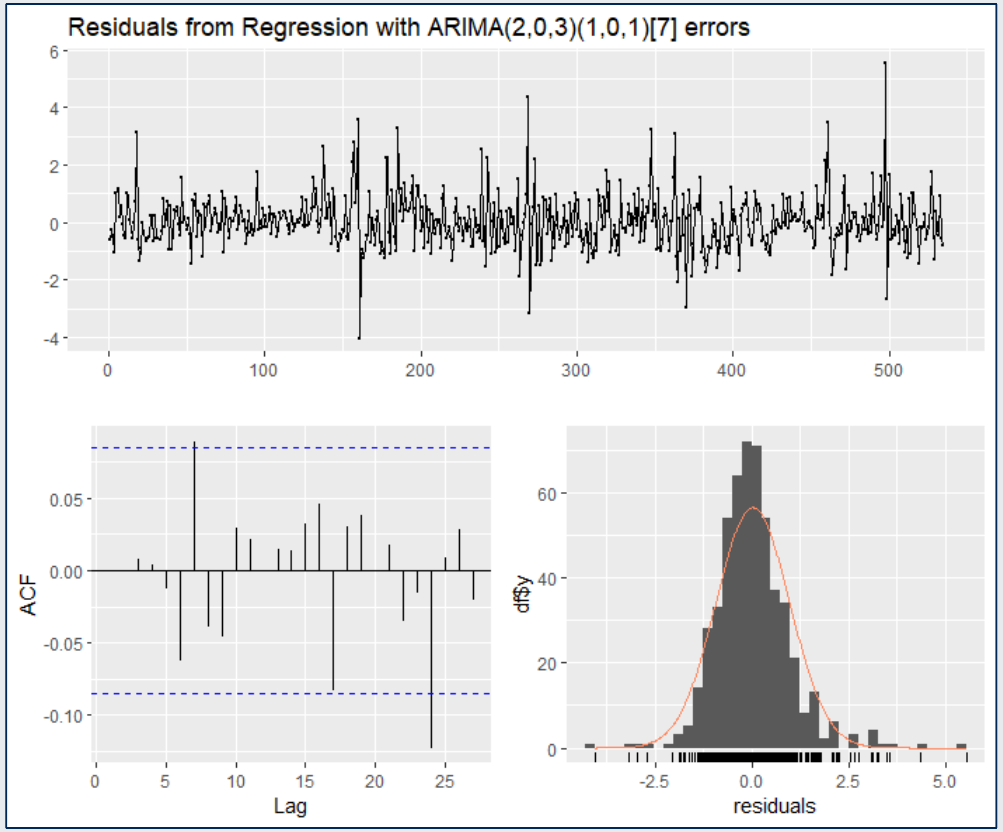
The SARIMA model was trained using the Arima() function with an order obtained from the auto_arima() function, ensuring consistency in the modeling approach.

Justification

- SARIMA model extends ARIMA by incorporating seasonal patterns and external factors
- Useful for modeling and forecasting page views considering seasonal variations and external influences

Trade-Offs

- Increased complexity due to the inclusion of seasonal components
- Requires careful selection of model parameters (p, d, q, P, D, Q) to accurately capture seasonality



The plot shows the residuals for the SARIMA model. We can observe that the residuals are normally distributed for the page views of German Language.

Note: Applied Arima() with same configuration for other 6 languages. Please see the appendix for more details.

Time Series Forecasting Techniques

Split the pre-processed data into training and testing subsets

Train Data: Up to December 16, 2016, | Test Data: From December 17, 2016, onwards.

SARIMAX

Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables

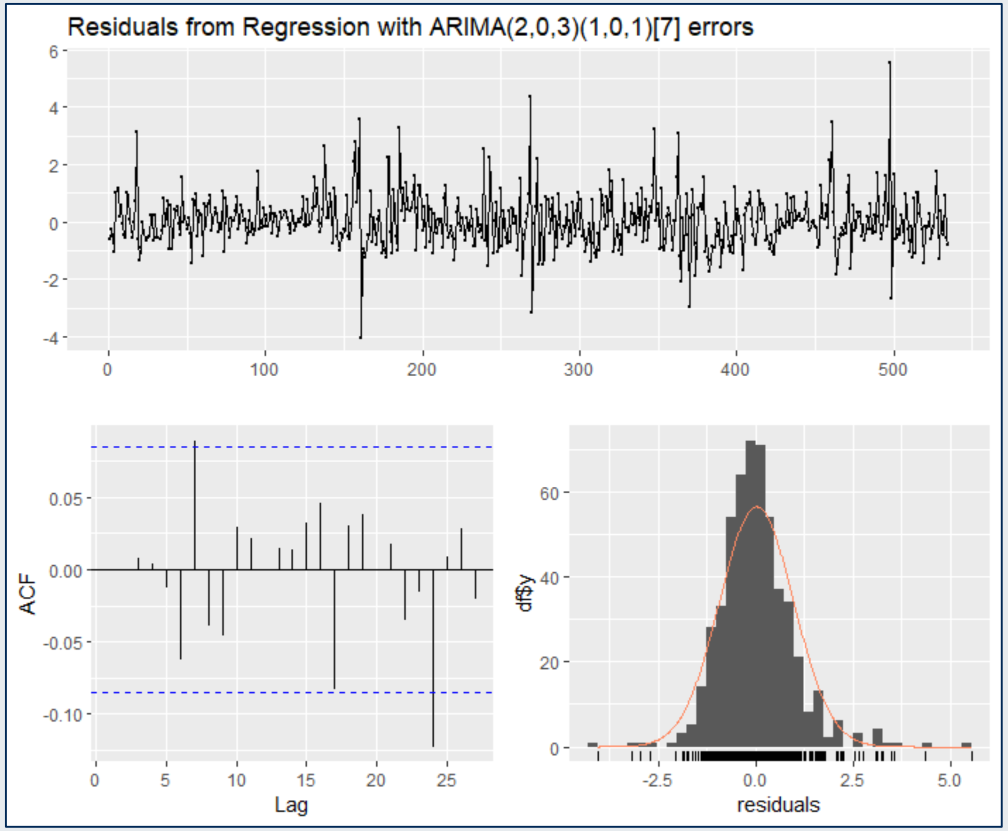
Justification

- SARIMA model further enhanced by incorporating additional exogenous variables
- Enables capturing the impact of external factors on the page view forecast

Trade-Offs

- Increased complexity due to the inclusion of exogenous variables and potential multicollinearity issues
- Requires thorough analysis and selection of relevant exogenous variables to avoid overfitting

The SARIMA model with exogenous variables was trained using an order of determined by auto_arima(), seasonal order of (1, 0, 1) with a period of 7.

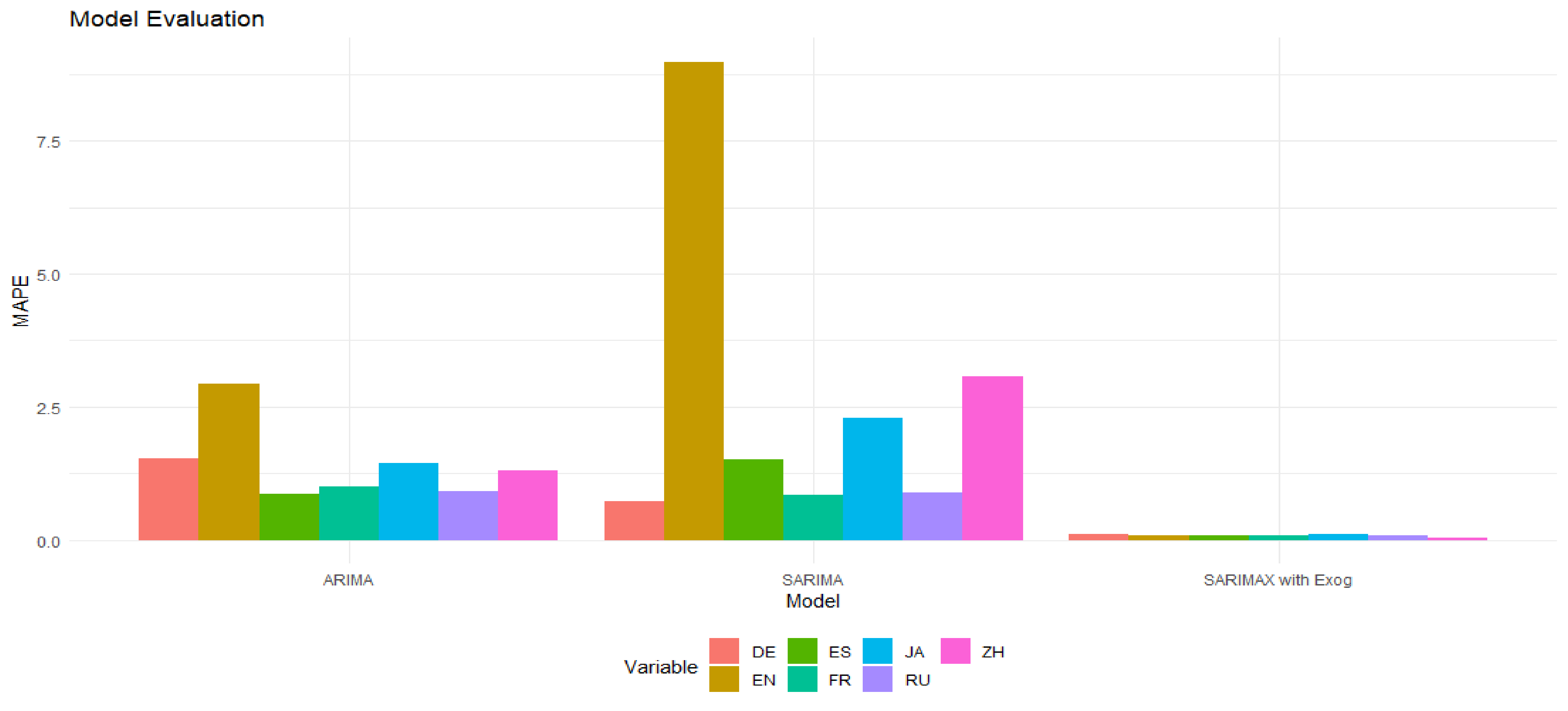


The plot shows the residuals for the SARIMAX model. We can observe that the residuals are normally distributed for the page views of the German Language.

Note: Applied Arima() with same configuration for other 6 languages. Please see the appendix for more details.

Model Evaluation

Identified **SARIMA with Exogenous Variables** as the best model for accurate forecasting of Wikipedia page views across different languages, using MAPE as the evaluation metric.



Conclusion

- Performed extensive exploratory analysis, including checking the dataset's structure and characteristics, handling null values, and extracting relevant information from the page names.
- Applied various time series forecasting techniques, including ARIMA, SARIMA, and SARIMA with exogenous variables.
- SARIMAX consistently achieved the lowest MAPE, indicating its superior accuracy in forecasting Wikipedia page views.

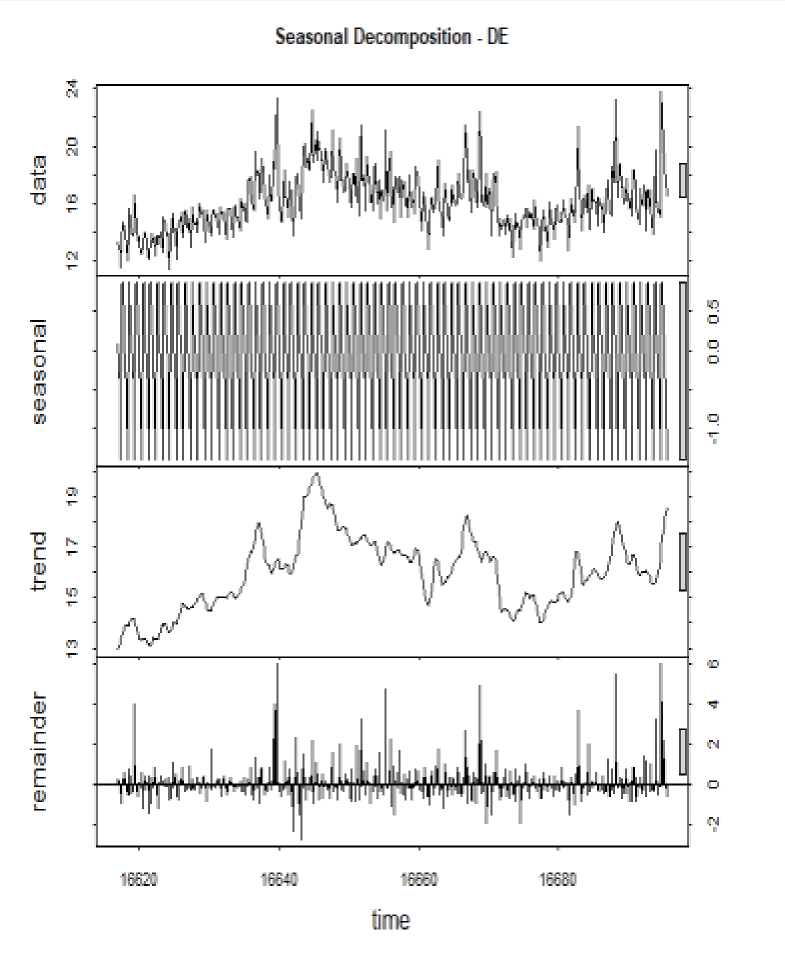
Future Work

- Explore more sophisticated time series forecasting models, such as Long Short-Term Memory (LSTM) networks or hybrid models combining deep learning and traditional forecasting techniques.
- Perform a more granular analysis by considering different segments or categories within each language.
- Integrating external data sources, such as social media trends, news events, or public holidays, can provide additional contextual information that may impact page views and user engagement.

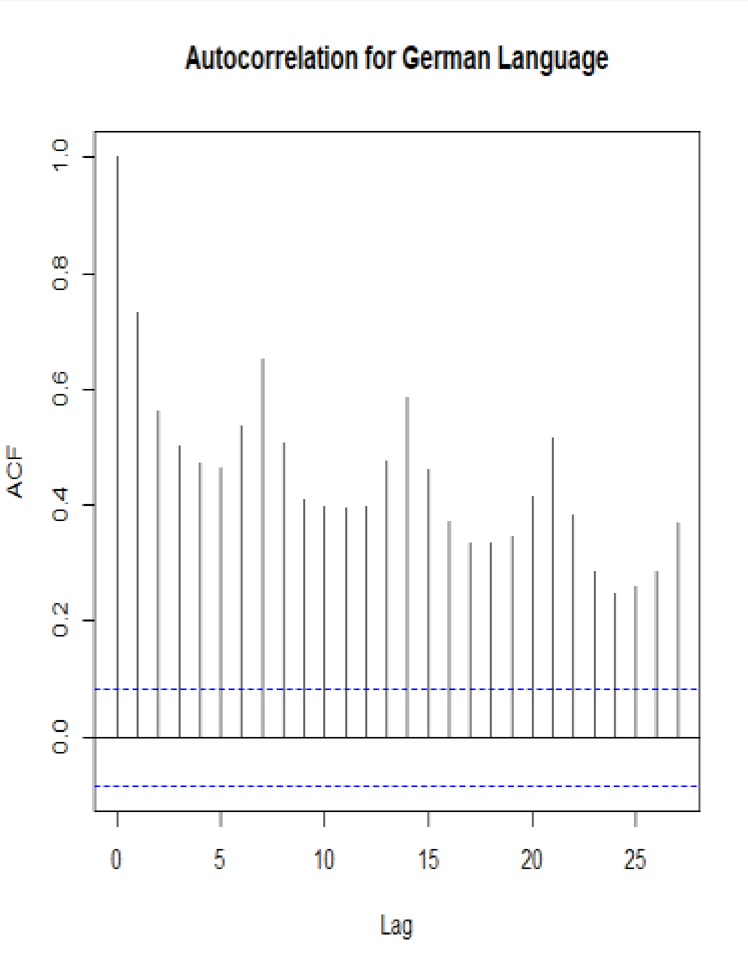
THANK YOU!

APPENDIX

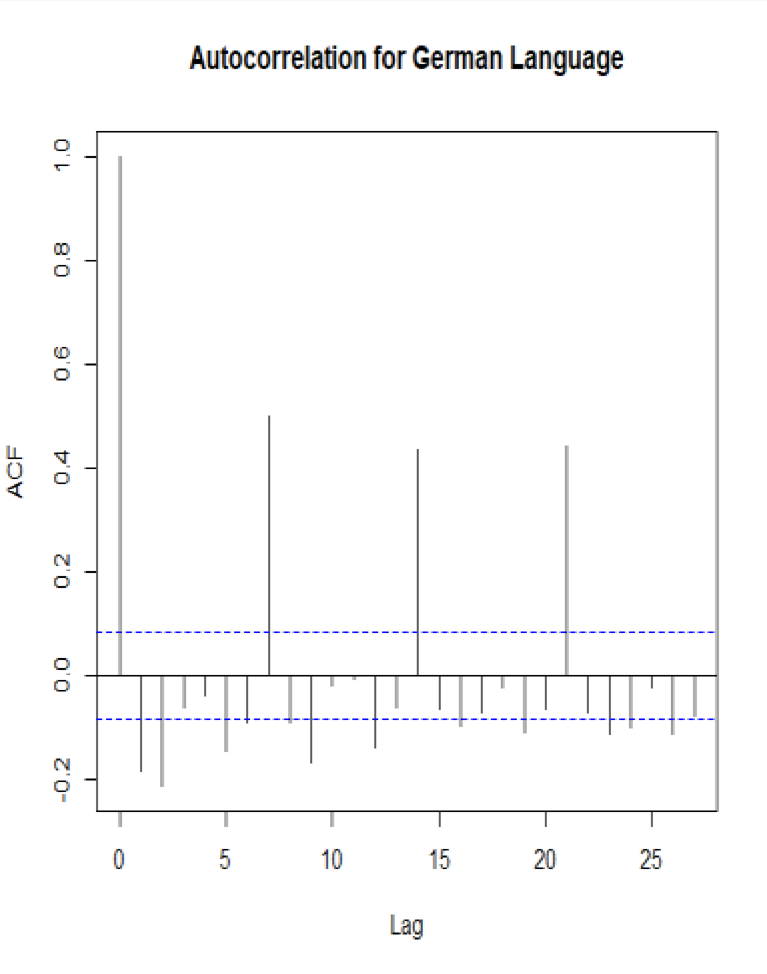
Seasonal Decomposition and ACF



The above graph shows the seasonal decomposition of the English Wikipedia page. We can observe a sharp increase in views trend in August 2016 and drop drastically.

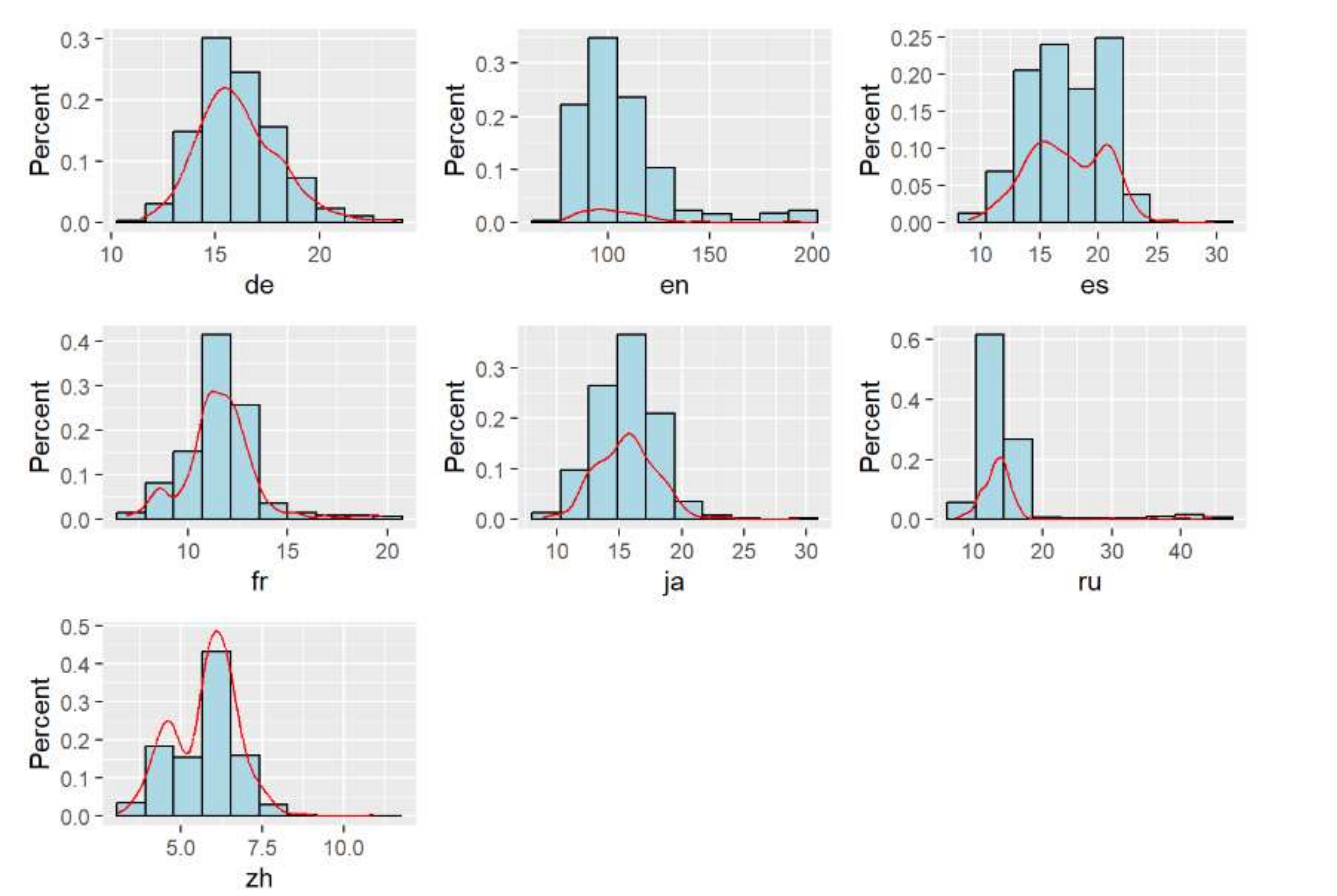


The graph show the ACF plot of German language. We can observe that it shows non-stationarity.

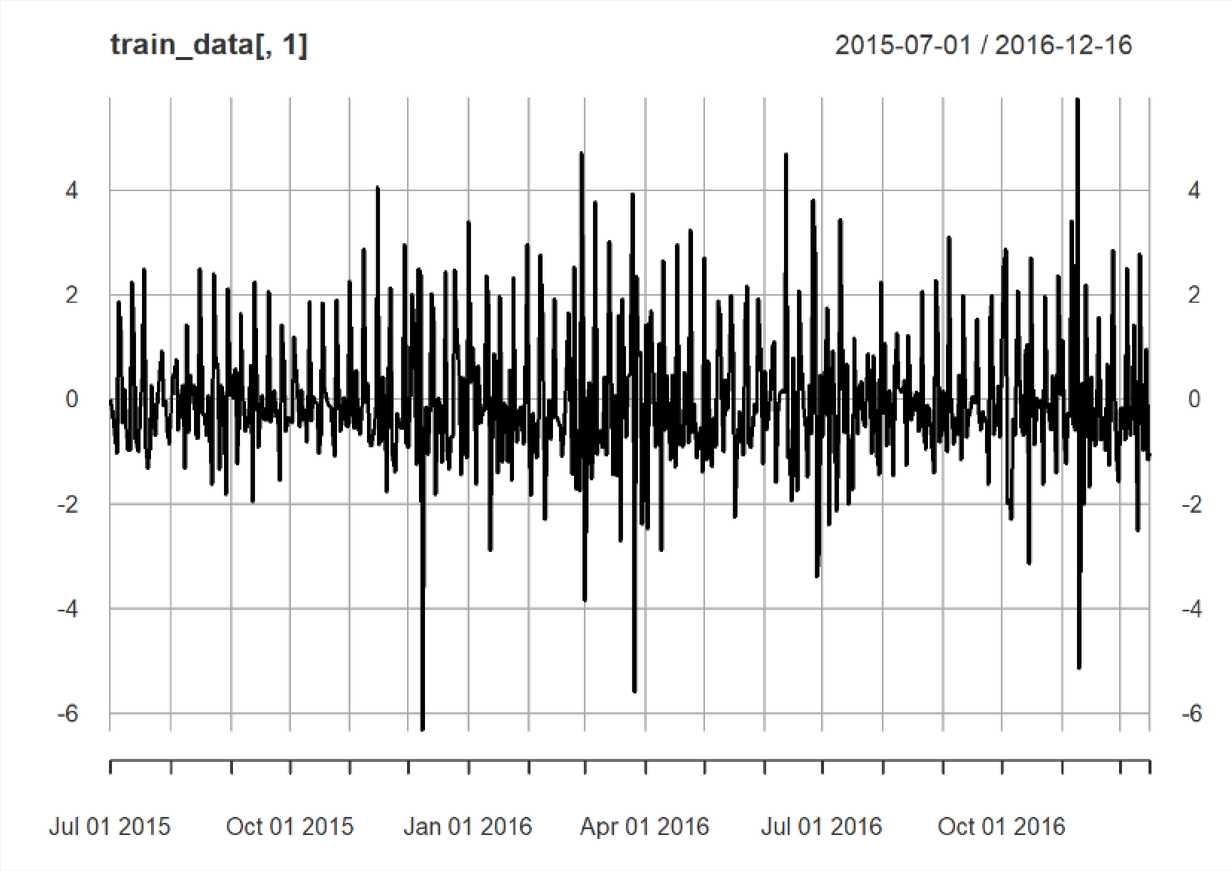


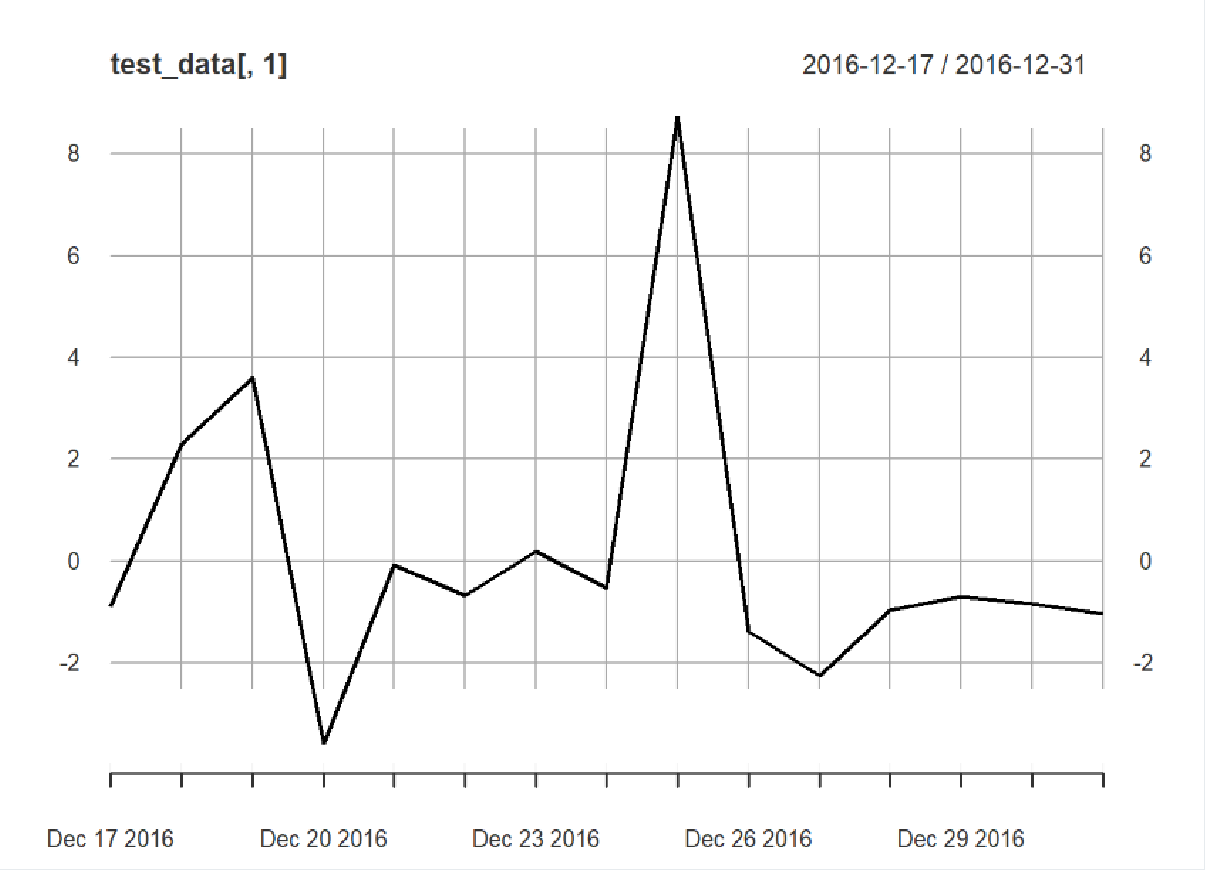
The graph shows the ACF plot of German language after applying first-order differencing. We can observe that it shows stationarity

Page View Distribution by Language

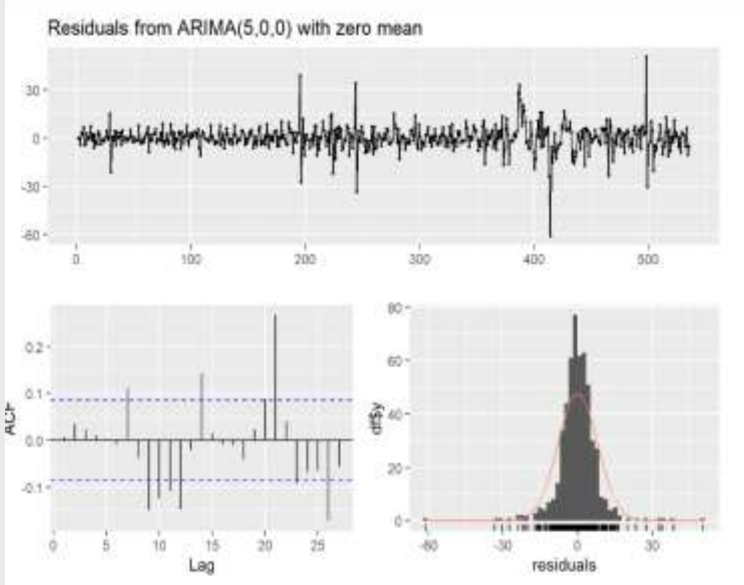


Train and Test Dataset

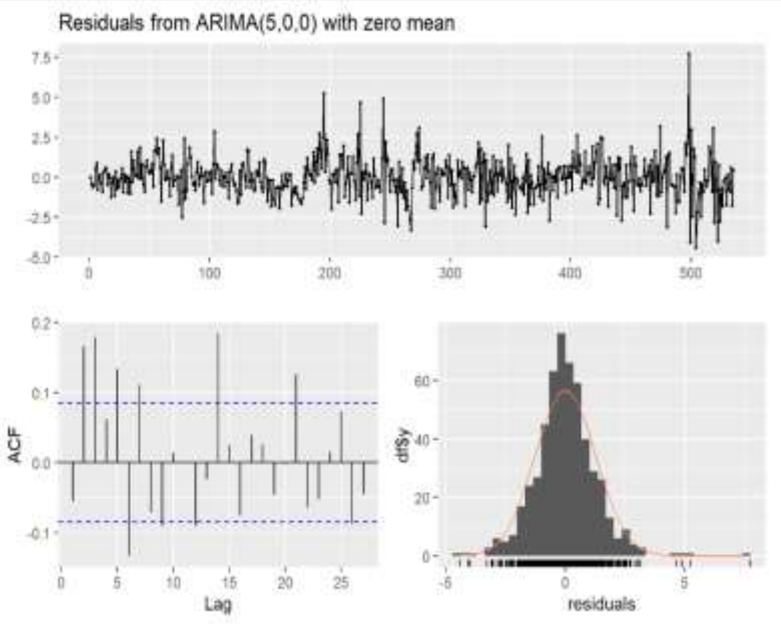




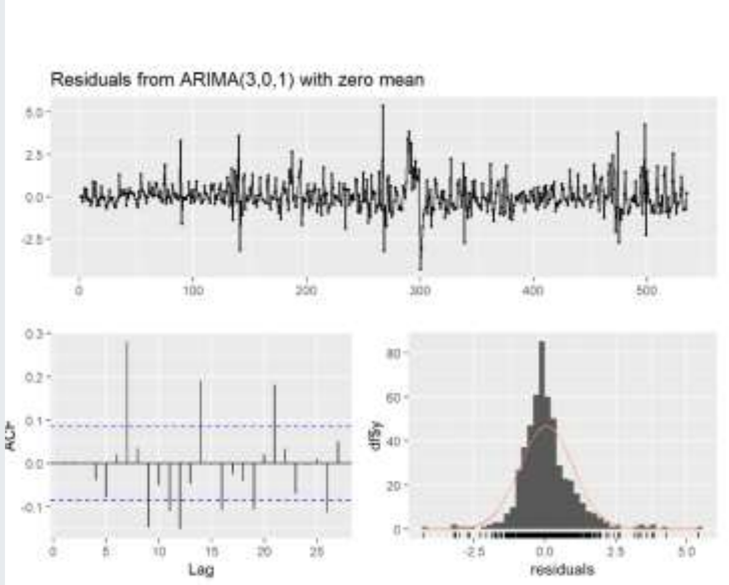
ARIMA Residuals



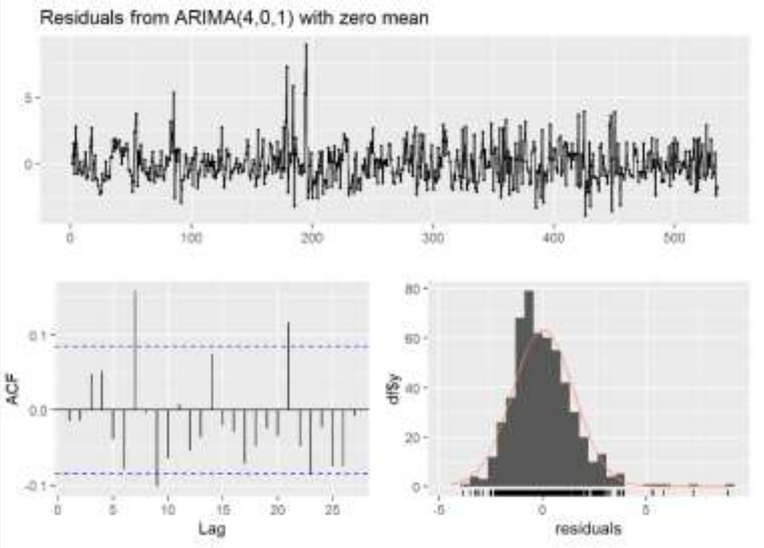
Residuals from ARIMA(5,0,0) with zero mean



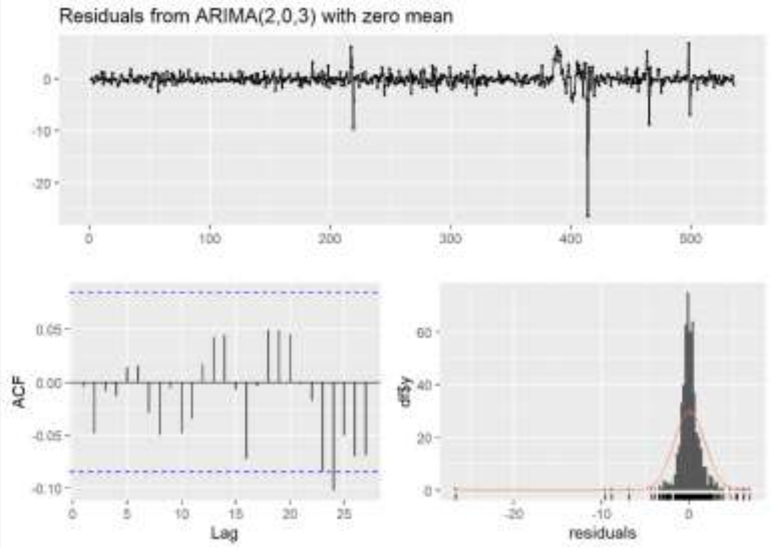
Residuals from ARIMA(5,0,0) with zero mean



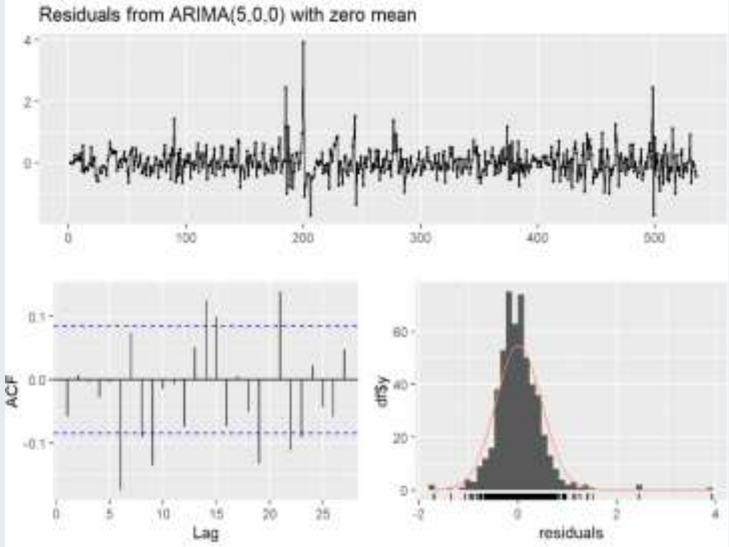
Residuals from ARIMA(3,0,1) with zero mean



Residuals from ARIMA(4,0,1) with zero mean

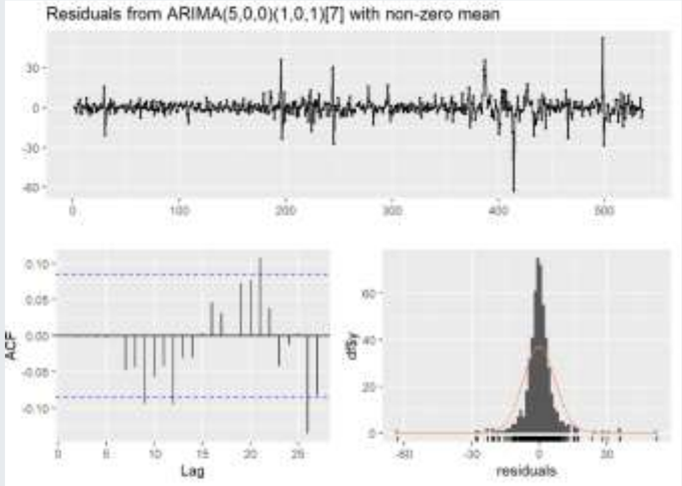


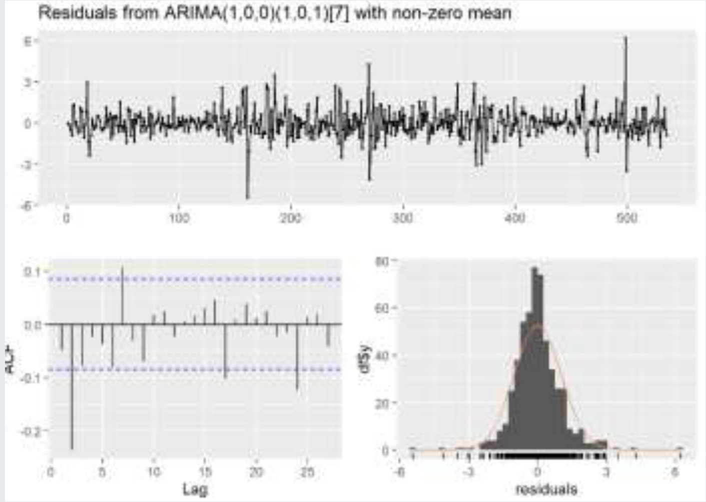
Residuals from ARIMA(2,0,3) with zero mean

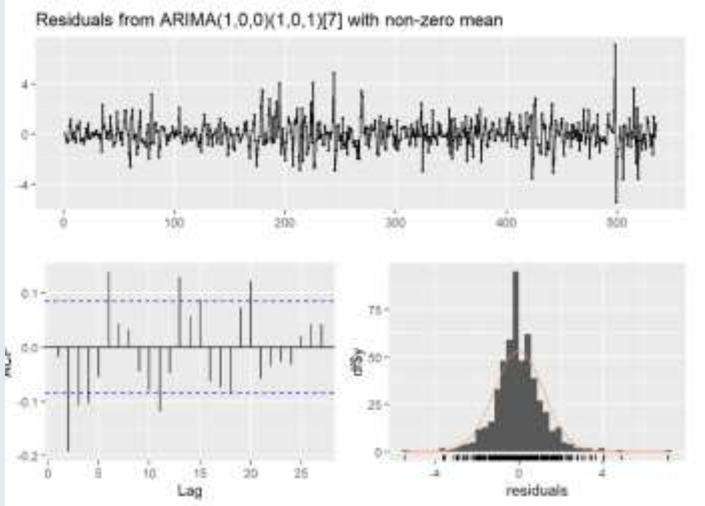


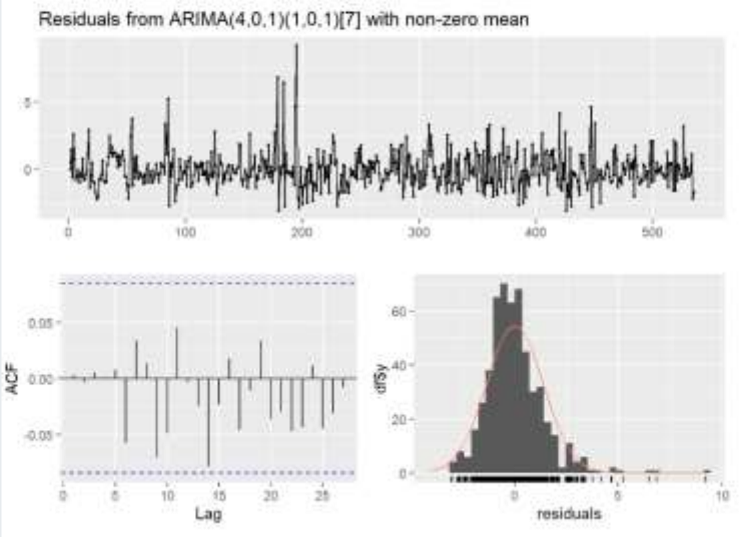
Residuals from ARIMA(5,0,0) with zero mean

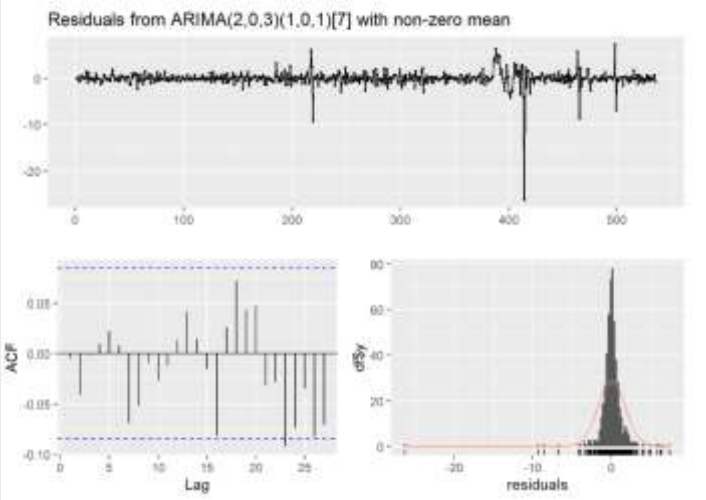
SARIMA Residuals

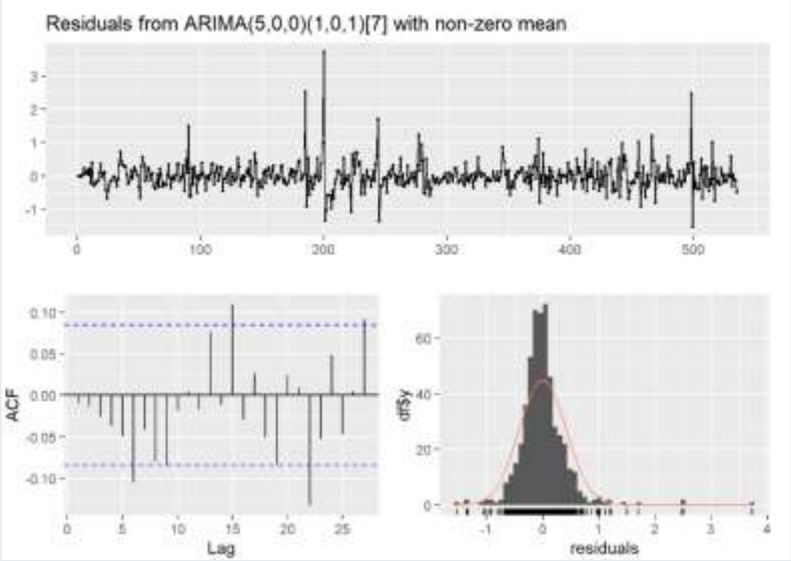

Residuals from ARIMA(5,0,0)(1,0,1)[7] with non-zero mean


Residuals from ARIMA(1,0,0)(1,0,1)[7] with non-zero mean

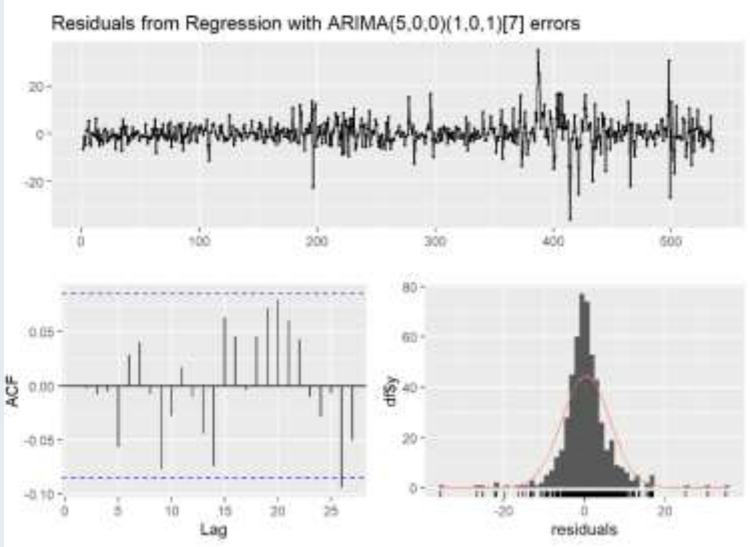

Residuals from ARIMA(1,0,0)(1,0,1)[7] with non-zero mean


Residuals from ARIMA(4,0,1)(1,0,1)[7] with non-zero mean

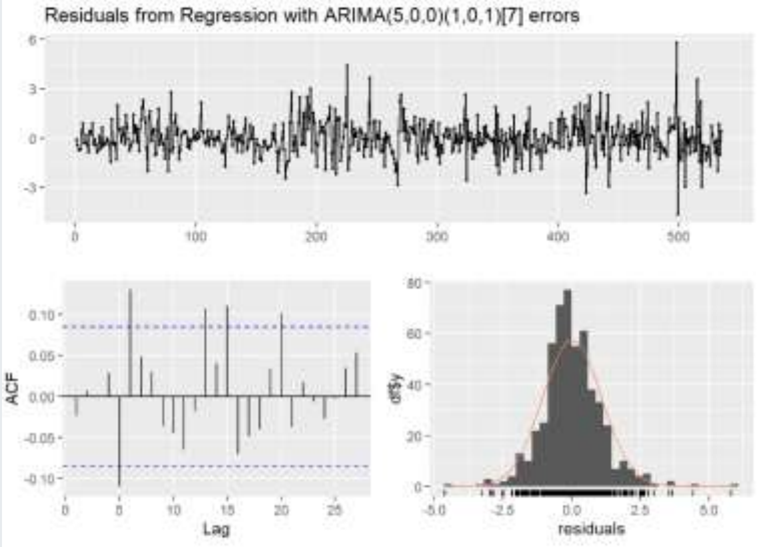

Residuals from ARIMA(2,0,3)(1,0,1)[7] with non-zero mean


Residuals from ARIMA(5,0,0)(1,0,1)[7] with non-zero mean

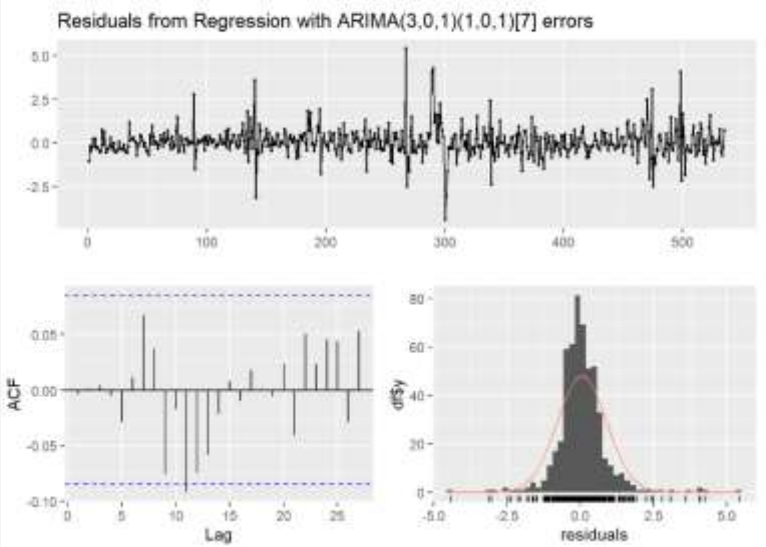
SARIMAX Residuals



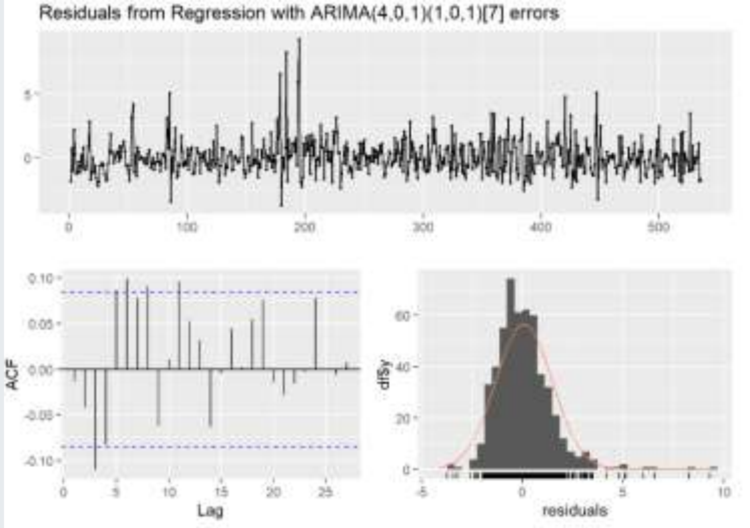
Residuals from Regression with ARIMA(5,0,0)(1,0,1)[7] errors



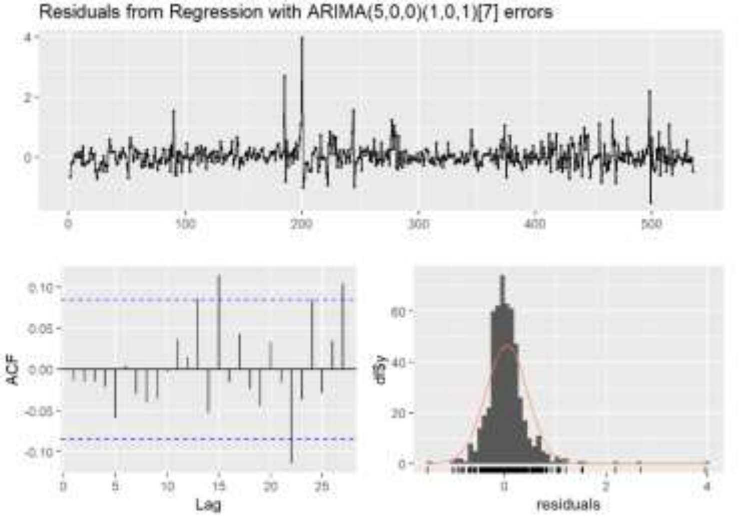
Residuals from Regression with ARIMA(5,0,0)(1,0,1)[7] errors



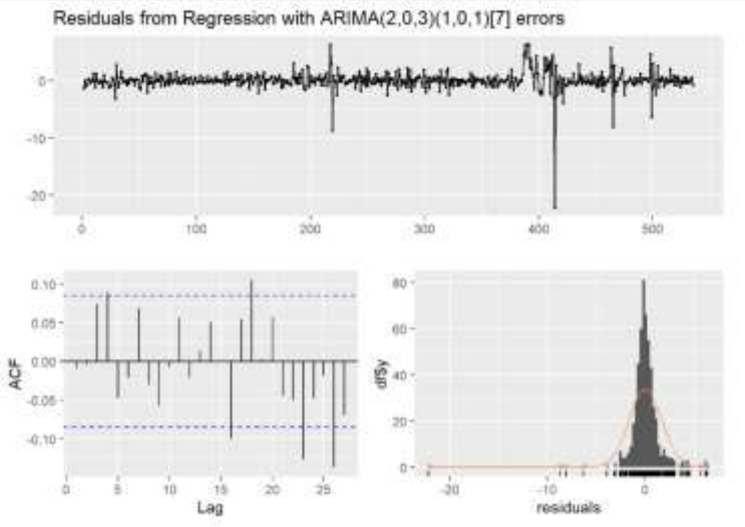
Residuals from Regression with ARIMA(3,0,1)(1,0,1)[7] errors



Residuals from Regression with ARIMA(4,0,1)(1,0,1)[7] errors



Residuals from Regression with ARIMA(5,0,0)(1,0,1)[7] errors



Residuals from Regression with ARIMA(2,0,3)(1,0,1)[7] errors