Pharma Drug Sales – Analysis and Forecasting

Time Series and Forecasting 2024/2025

Group 18

Athos Freitas

Luís Du

Master in Artificial Intelligence

Professor: Maria Eduarda Silva

10th January, 2025

1. Introduction

This project aims to analyze and forecast pharmaceutical drug sales using time series data collected from the Point-of-Sale system of a single pharmacy over six years (2024–2019). The primary objective is to uncover underlying patterns in the data, such as trends and seasonality, and to develop models capable of generating accurate and reliable sales forecasts.

Trend and seasonality patterns of the time series will be explored, along with the evaluation of different models. Various forecasting models and strategies will be evaluated, with an emphasis on identifying the most effective and computationally efficient approaches. Accuracy measures will be applied to assess the quality of the results. Finally, the study will conclude with a discussion of the benefits and limitations of the chosen approaches.

2. Methodology

The analysis involves several key steps to ensure a comprehensive understanding and accurate forecasting of pharmaceutical drug sales.

Firstly, **Exploratory Data Analysis (EDA)** will be conducted to visually inspect the time series and detect patterns such as trends and seasonality. Variance will be analyzed to identify any instability, and different transformations will be applied as needed to stabilize it. Statistical tests, including the Augmented Dickey-Fuller (ADF) test, will be used to confirm stationarity. If the time series is not stationary, transformations or differencing techniques will be applied to stabilize its mean and variance, ensuring that it meets the assumptions for time series modeling.

The next step is **Data Partitioning**, where the dataset will be divided into training and test sets. This separation is crucial for model development and evaluation, enabling the assessment of the models' performance on unseen data.

During **Model Development**, ARIMA and SARIMA models will be constructed to capture the underlying patterns in the data. The process will involve diagnostic checks such as autocorrelation and partial autocorrelation analysis, unit-root testing, and the Ljung-Box test to ensure the models' adequacy. Models will be shortlisted based on their Akaike Information Criterion (AIC) scores. Additionally, an alternative modeling approach using STL decomposition will be applied to each time series component – trend, seasonality, and residuals – to enhance forecasting accuracy.

In the **Forecasting and Evaluation** phase, the selected models will be employed to generate forecasts for the test set. Prediction accuracy will be assessed using cross-validation and comparisons with actual values.

Both rolling forecasts and long-term step-ahead forecasting strategies will be explored to evaluate the models' performance in different scenarios. To quantify the uncertainty associated with the forecasts, 95% prediction intervals will be calculated.

The **Quality Measures** phase will involve applying accuracy metrics to evaluate the reliability and precision of the forecasting results, ensuring a robust assessment of the models' effectiveness.

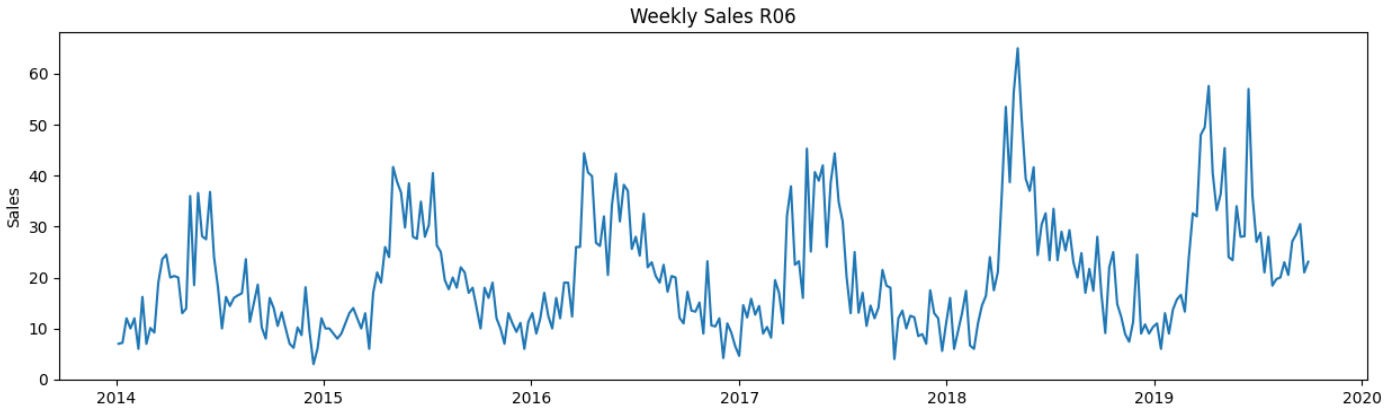
Finally, the study will conclude with a **Discussion** of the strengths and limitations of the ARIMA and SARIMA models in forecasting pharmaceutical sales. This section will provide insights into the models' suitability for this domain, highlight areas for improvement, and offer recommendations for future research.

3. Exploratory Data Analysis

3.1. Time series plot

The chosen dataset for this study is derived from a Point-of-Sale (POS) system of a single pharmacy, covering a period of six years. The research underlying this project considers eight distinct time series, each summarizing the sales of a specific group of pharmaceutical products. These series exhibit varying statistical features, offering a diverse foundation for analysis and forecasting.

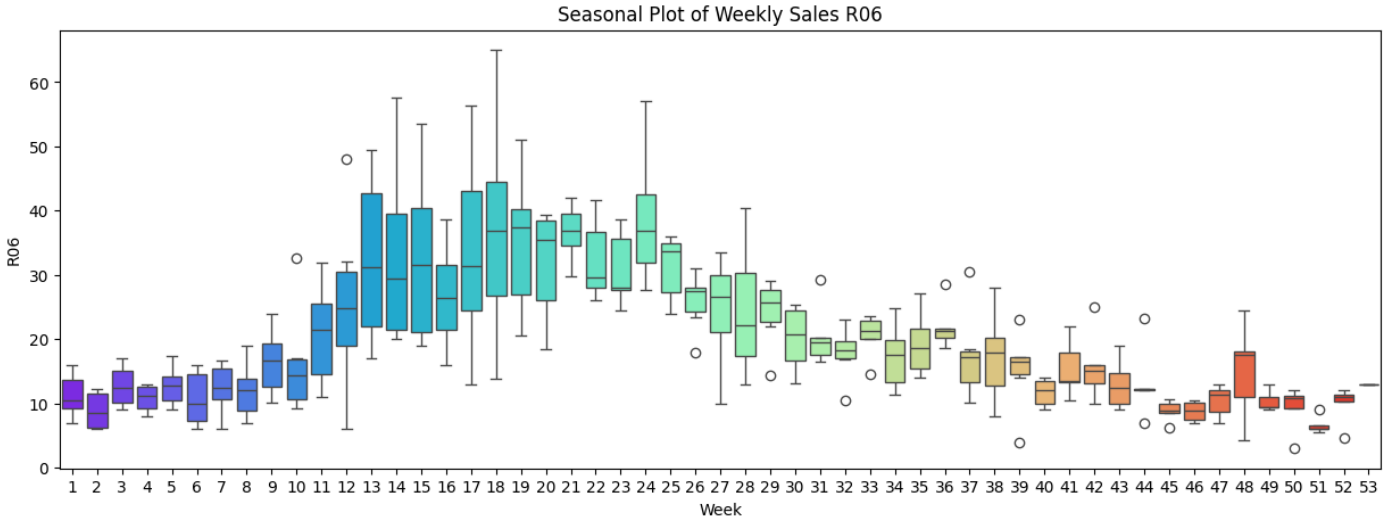
In this study, we will focus on the analysis of the **R06 category**, which includes **antihistamines for systemic use.** This category is particularly relevant due to its consistent demand and significance within the pharmaceutical sector, making it an ideal candidate for exploring time series patterns.



By visualizing the time series data, we observe several key characteristics. First, there is no clear evidence of a trend over time, indicating that the data does not exhibit a consistent upward or downward movement. Second, the data displays pronounced seasonality, with recurring patterns that suggest regular fluctuations at specific intervals. Lastly, the variance is relatively stable throughout the series, except for a noticeable spike during the early weeks of 2018.

3.2. Seasonal plot

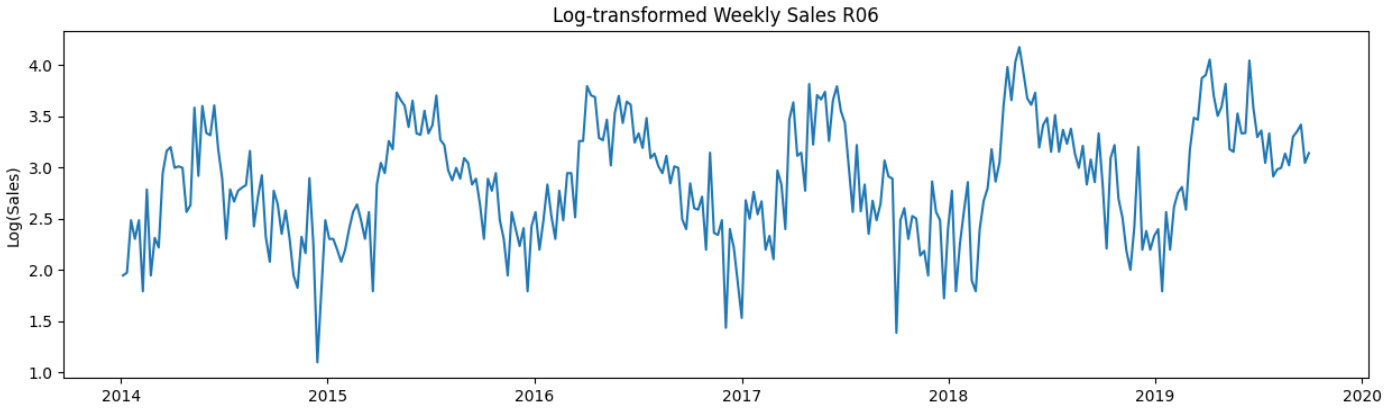
For a clearer representation of the seasonal patterns, we may use the seasonal plots.



The seasonal plot reveals that sales consistently peak during the spring, followed by a gradual decline throughout the remainder of the year. This pattern aligns with expectations, as it reflects the increased demand for antihistamines during allergy seasons, when symptoms are most prevalent.

3.3. Variance stabilization

To stabilize the variability over the series, the Box-Cox transformations can be applied. One specific case of this transformation is to take the **logarithm of the data**.



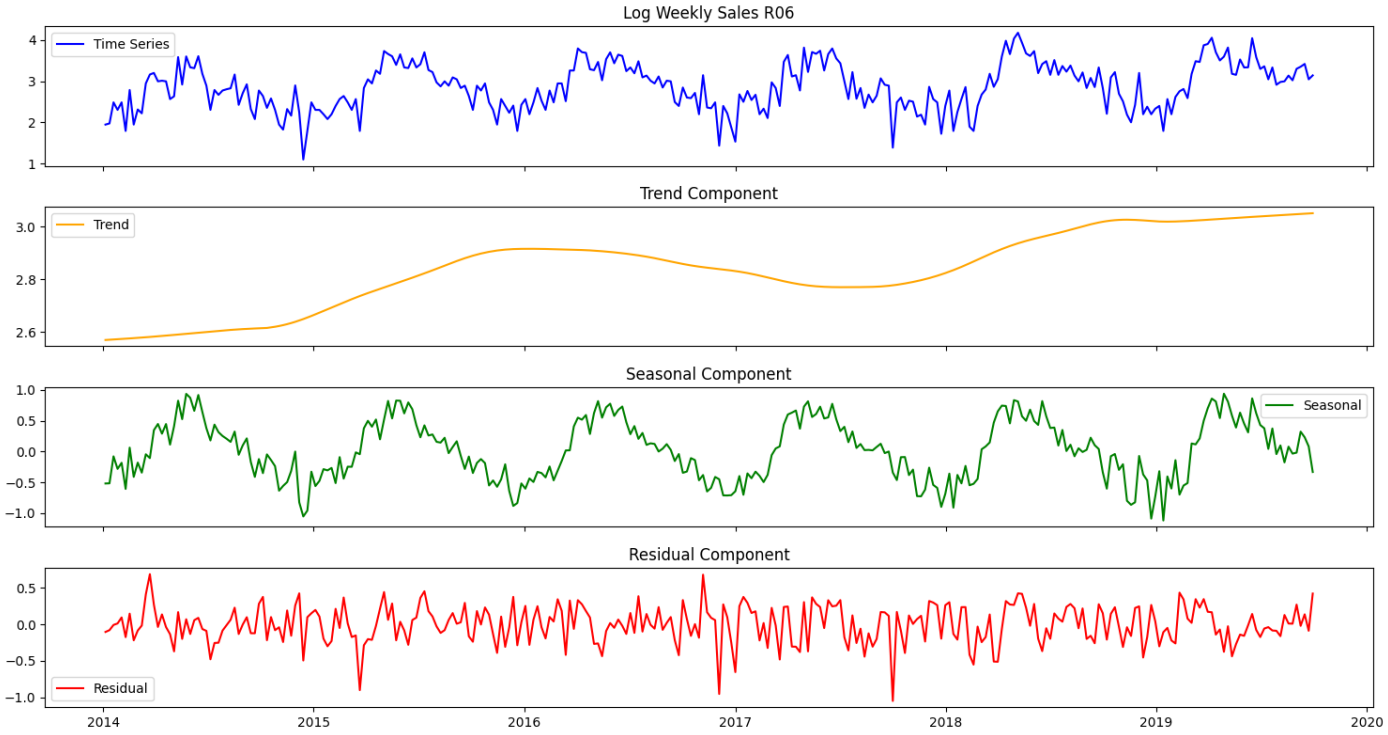
After applying a log transformation to the time series, the variance appears significantly reduced, with the previously noticeable spike in early 2018 no longer evident. This indicates that the transformation effectively stabilized the data, making it more suitable for modeling.

The best **Box-Cox transformation** was also explored by identifying the lambda value that minimizes variance (λ=0.10). Since this value is close to zero, the result of the Box-Cox transformation is very similar

to the log transformation. To maintain simplicity and enhance interpretability, we will use the log-transformed time series in the subsequent sections.

3.4. Seasonal-Trend decomposition using Loess (STL)

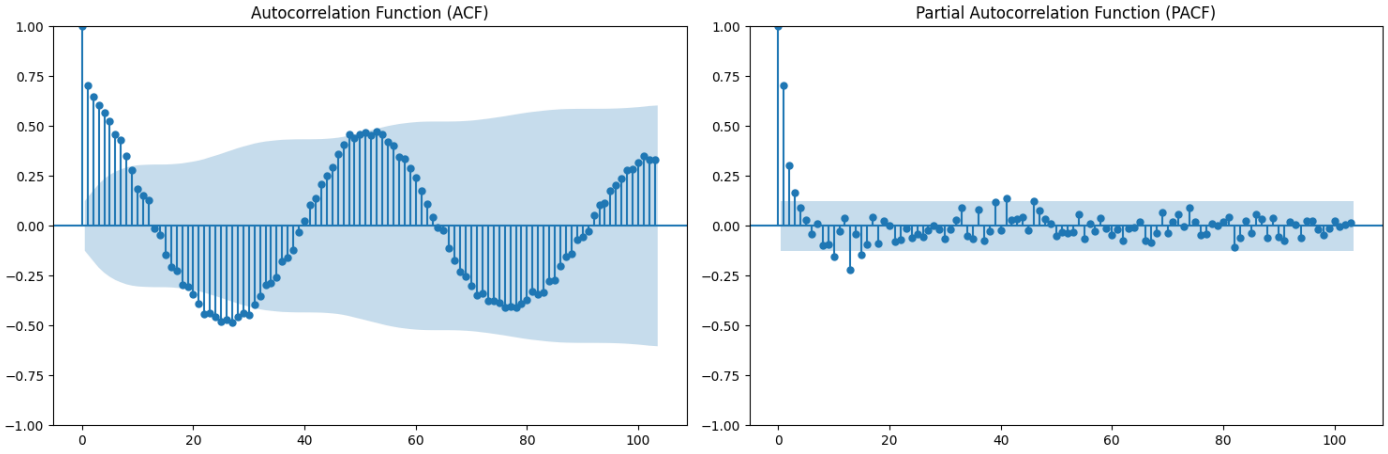
The Seasonal Decomposition of Time Series by Loess decomposes a time series into seasonal, trend and irregular components using Loess.



From the decomposition, we can confirm some of the earlier observations about the time series. Firstly, there is a lack of an evident trend, as the trend component does not exhibit a clear directional movement over time. Secondly, the seasonal component reveals a recurring pattern that aligns with the previously observed seasonality, reinforcing the periodic nature of the data.

3.5. Autocorrelation and Partial Autocorrelation Functions

By examining the Autocorrelation and Partial Autocorrelation functions of the original time series, we observe several key patterns. The **ACF** clearly exhibits cyclic patterns, gradually approaching zero, which indicates seasonality and diminishing correlations at higher lags. On the other hand, the **PACF** shows distinct spikes at lags 1, 2, and 3, suggesting strong direct correlations at these lags and highlighting their potential relevance in modeling autoregressive components of the time series.



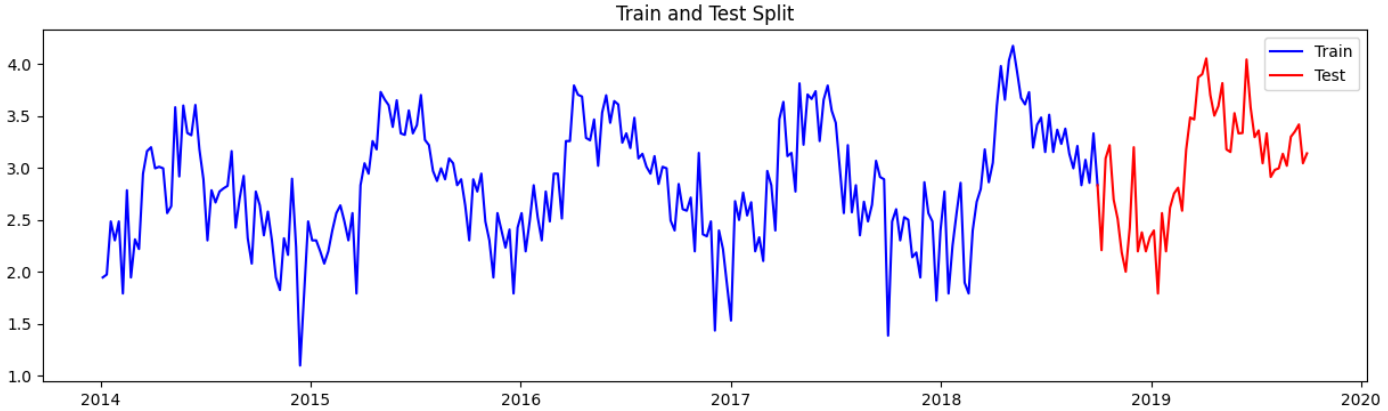
3.6. Stationarity tests

To assess the stationarity of the time series, both the **Augmented Dickey-Fuller (ADF)** and **Kwiatkowski-Phillips-Schmidt-Shin (KPSS)** tests were applied. The ADF test examines the null hypothesis that a unit root is present in the time series, indicating non-stationarity. In this case, the ADF statistic is lower than the critical values, and the p-value is less than 0.05, confirming that **the log-transformed time series is stationary.**

The KPSS test, on the other hand, evaluates the null hypothesis that the time series is stationary around a deterministic trend. The KPSS test indicates that 296 differences would be required to achieve a relatively stationary time series. However, applying such a high number of differences is impractical in real-world scenarios, suggesting that alternative approaches should be considered for addressing stationarity.

4. Data partitioning

The most recent 52 weeks of data will be designated as the testing set, providing a realistic forecast horizon for evaluating model performance and prediction accuracy.



5. Modelling

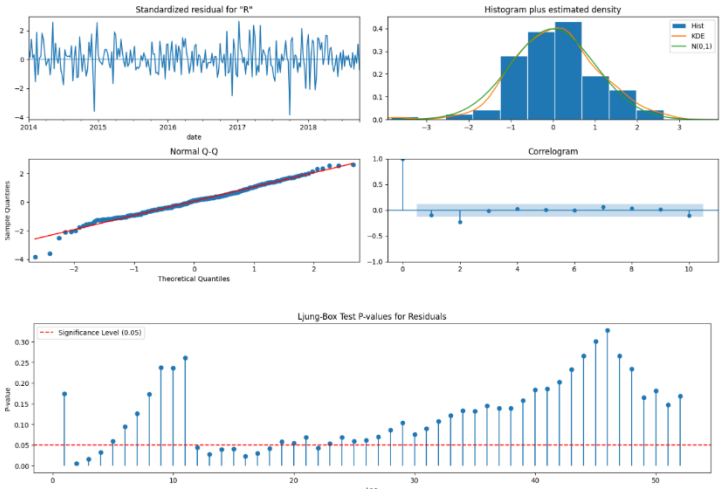
5.1. Autoregressive models

Based on the observed ACF and PACF during the data analysis phase, which suggest the presence of autoregressive patterns, we will explore **AR(2)** and **AR(3)** models to capture the temporal dependencies in the data.

- **AR(2)**

SARIMAX Results

Dep. Variable:		R06		No. Observations:		248
Model:		SARIMAX(2, 0, 0)		Log Likelihood		-119.475
Date:		Thu, 09 Jan 2025		AIC		244.950
Time:		09:26:59		BIC		255.490
Sample:		01-05-2014		HQIC		249.193
		- 09-30-2018				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5835	0.054	10.873	0.000	0.478	0.689
ar.L2	0.4093	0.053	7.713	0.000	0.305	0.513
sigma2	0.1509	0.011	13.443	0.000	0.129	0.173
Ljung-Box (L1) (Q):		2.28	Jarque-Bera (JB):			11.92
Prob(Q):		0.13	Prob(JB):			0.00
Heteroskedasticity (H):		1.17	Skew:			-0.19
Prob(H) (two-sided):		0.47	Kurtosis:			4.01

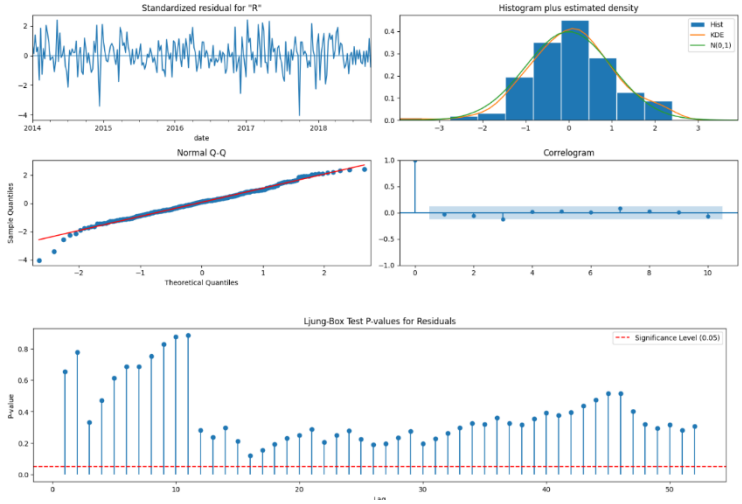


For the **AR(2)** model, all the **parameters are statistically significant**; however, there is still **some correlation between the residuals.** This suggests that the model does not fully capture all the underlying patterns in the data, indicating the need for a more complex model to better explain the observed behavior.

- **AR(3)**

SARIMAX Results

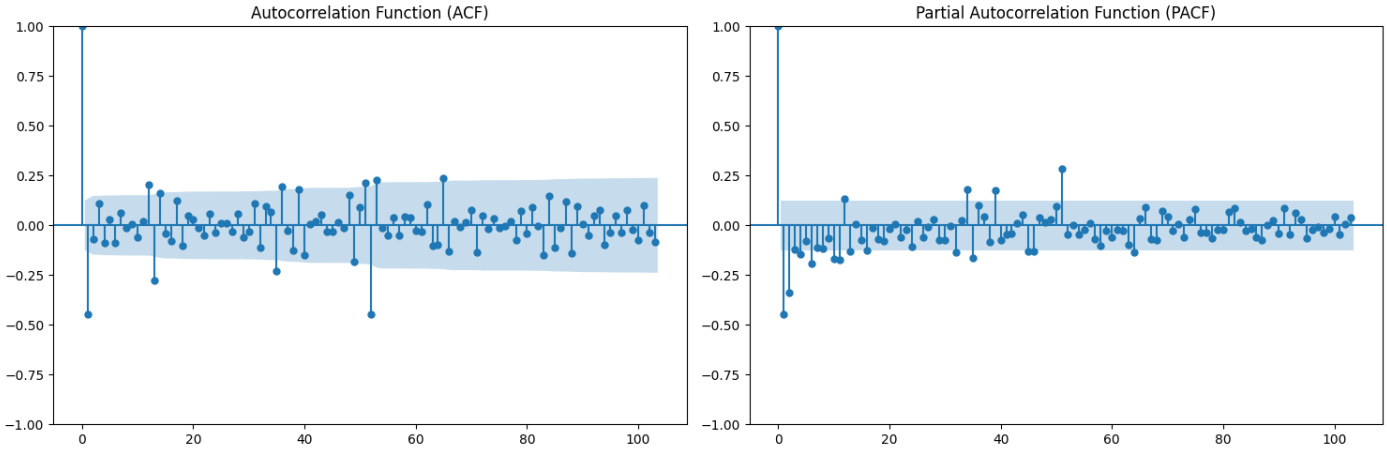
Dep. Variable:		R06		No. Observations:		248
Model:		SARIMAX(3, 0, 0)		Log Likelihood		-113.195
Date:		Thu, 09 Jan 2025		AIC		234.389
Time:		09:27:00		BIC		248.443
Sample:		01-05-2014		HQIC		240.047
		- 09-30-2018				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4924	0.062	7.960	0.000	0.371	0.614
ar.L2	0.2785	0.064	4.331	0.000	0.152	0.405
ar.L3	0.2233	0.063	3.571	0.000	0.101	0.346
sigma2	0.1434	0.011	13.471	0.000	0.123	0.164
Ljung-Box (L1) (Q):		0.27	Jarque-Bera (JB):			14.82
Prob(Q):		0.60	Prob(JB):			0.00
Heteroskedasticity (H):		1.15	Skew:			-0.28
Prob(H) (two-sided):		0.54	Kurtosis:			4.06



Using the AR(3) model, we observe that all the **parameters remain statistically significant**, and the **residuals are now independent**. This indicates that the AR(3) model provides a better fit to the data compared to the AR(2) model, effectively capturing the underlying patterns and reducing residual autocorrelation.

5.2. SARIMA models

To address the seasonality, we have decided to apply a **52-step difference** to the time series, which corresponds to the seasonal cycle observed over the course of a year. The **ADF test** was then applied to assess the stationarity of the differenced series. Based on the results of the ADF test, we conclude that only **one difference (d=1)** is sufficient to achieve stationarity in the time series, ensuring that the data is appropriately transformed for modeling.

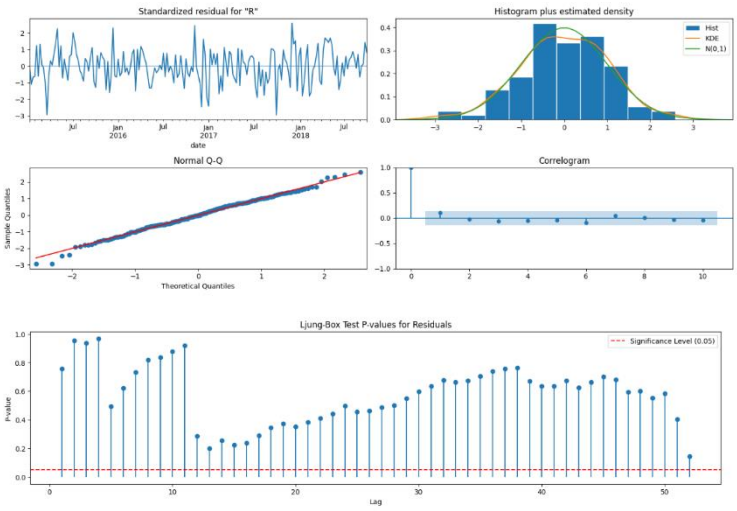


Looking at the ACF and PACF, we observe clear spikes at both **lag 1** and **lag 52** in the plots. This indicates strong correlations at these specific lags, which likely correspond to **immediate** and **seasonal dependencies** in the time series, suggesting that both short-term and yearly cyclical patterns are important factors in modeling the data.

- **SARIMA$(0,1,1)(1,1,1)_{52}$**

SARIMAX Results

Dep. Variable:		R06		No. Observations:		248
Model:	SARIMAX(0, 1, 1)x(1, 1, 1, 52)			Log Likelihood		-105.063
Date:		Thu, 09 Jan 2025		AIC		218.127
Time:		09:27:36		BIC		231.219
Sample:		01-05-2014		HQIC		223.427
		- 09-30-2018				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9254	0.032	-29.141	0.000	-0.988	-0.863
ar.S.L52	-0.3076	0.164	-1.873	0.061	-0.629	0.014
ma.S.L52	-0.4197	0.195	-2.150	0.032	-0.802	-0.037
sigma2	0.1476	0.016	9.017	0.000	0.116	0.180
Ljung-Box (L1) (Q):	2.32		Jarque-Bera (JB):		0.97	
Prob(Q):	0.13		Prob(JB):		0.61	
Heteroskedasticity (H):	1.32		Skew:		-0.17	
Prob(H) (two-sided):	0.26		Kurtosis:		3.09	

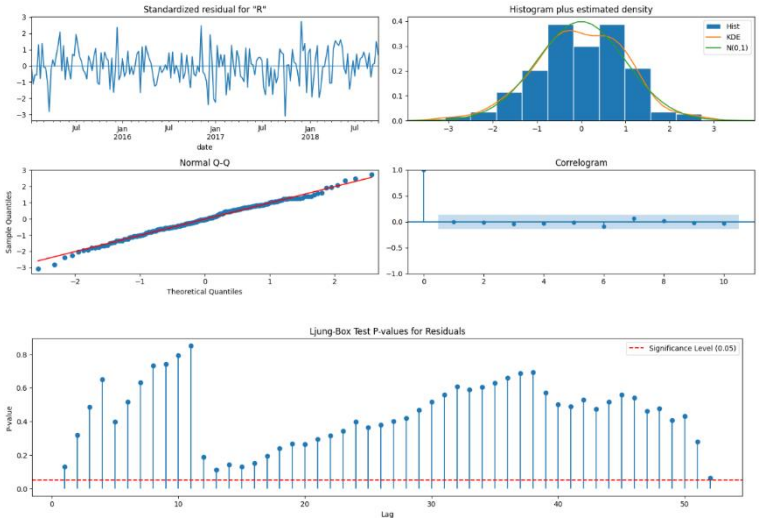


In this SARIMA model, **all the parameters are statistically significant, and the residuals are uncorrelated**, indicating a good fit. Let's now check the model with p=1 to see if it further improves the performance.

- **SARIMA(1,1,1)(1,1,1)$_{52}$**

SARIMAX Results

Dep. Variable:			R06	No. Observations:		248
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 52)			Log Likelihood		-103.633
Date:			Thu, 09 Jan 2025	AIC		217.266
Time:			09:28:02	BIC		233.631
Sample:			01-05-2014	HQIC		223.892
			- 09-30-2018			
Covariance Type:			opg			
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1293	0.080	1.614	0.106	-0.028	0.286
ma.L1	-0.9395	0.033	-28.153	0.000	-1.005	-0.874
ar.S.L52	-0.3455	0.164	-2.101	0.036	-0.668	-0.023
ma.S.L52	-0.3694	0.202	-1.830	0.067	-0.765	0.026
sigma2	0.1463	0.016	9.196	0.000	0.115	0.178
Ljung-Box (L1) (Q):		0.00	Jarque-Bera (JB):		1.02	
Prob(Q):		0.98	Prob(JB):		0.60	
Heteroskedasticity (H):		1.30	Skew:		-0.17	
Prob(H) (two-sided):		0.29	Kurtosis:		3.12	



In this model, both requirements are also fulfilled, with **all parameters being statistically significant and the residuals uncorrelated**, making it also a strong candidate for forecasting.