Time Series Analysis of Personal Consumption Expenditures

Betül Yıldırım
Department of Statistics
Middle East Technical University, Ankara, Turkey
yildirim.betul@metu.edu.tr

January 18, 2025

Abstract

This study forecasts monthly Personal Consumption Expenditures (PCE) utilizing multiple forecasting models, including ARIMA, ETS, TBATS, Neural Networks, and Prophet. The research predominantly employed Jupyter Notebook within Visual Studio Code, supplemented by R Studio for particular studies such as the GARCH model. Before predicting, the data was subjected to preprocessing procedures including outlier analysis, data cleansing, and stationarity assessments. Following the rectification of data anomalies the models were trained and assessed on both training and test datasets, with their performances compared utilizing measures such as RMSE and MAE.

1 Introduction

The principal objective of this study is to examine and model a time series dataset to identify patterns, trends, and potential anomalies. The dataset employed for this project encompasses monthly Personal Consumption Expenditures (PCE) from January 1959 to September 2024, offering a total of 789 observations across a comprehensive time frame. The data was obtained from the Federal Reserve Economic Data (FRED) website, particularly from the PCE series (https://fred.stlouisfed.org/series/PCE). FRED, operated by the Federal Reserve Bank of St. Louis, is a prominent platform providing complimentary access to economic data, financial trends, and macroeconomic indicators from credible sources. The extensive database guarantees the reliability and precision of the data utilized in this investigation.

The research commences with a comprehensive evaluation of the dataset to discern underlying trends, seasonal patterns, and possible anomalies. This research seeks to develop prediction models utilizing advanced statistical

approaches , aiming to accurately capture the data's fundamental structure and produce accurate predictions for future periods.

Multiple forecasting models, such as ARIMA, Prophet with hyperparameter optimization, GARCH, ETS, TBATS, and Neural Network-based models, were employed in the investigation. Their performances were evaluated utilizing evaluation criteria including Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). This study mostly utilized Jupyter Notebook within Visual Studio Code to do the analyses. Nonetheless, specific analyses, including the GARCH model, necessitated the utilization of functions accessible in R Studio. Both technologies were essential for the thorough assessment and precision of the models.

2 Data Description

The dataset consists of 789 monthly observations from January 1959 to September 2024. Key statistics are shown in Table 1.

Statistic	**PCE**
Count	789.000000
Mean	5759.638910
Standard Deviation	5252.222913
Minimum	306.100000
25%	1026.800000
50% (Median)	4003.600000
75%	9852.400000
Maximum	20024.300000

Table 1: Summary Statistics for Personal Consumption Expenditures (PCE)

Figure 1 The time series plot.

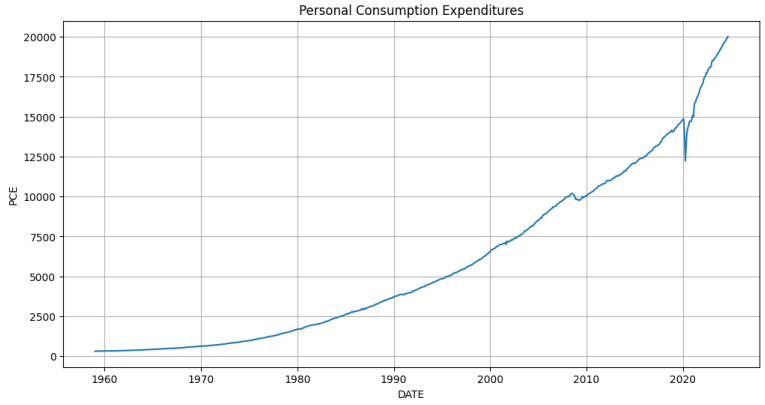


Figure 1: Personal Consumption Expenditures Over Time

As shown in Figure 1,The time series plot of the PCE data demonstrates an upward tendency, suggesting that the series is non-stationary and may display a stochastic trend. Furthermore, the series exhibits non-stationarity in its mean. The data exhibits no evidence of seasonal patterns. A more specific time period was shown for a detailed investigation of the fluctuation in the PCE data.

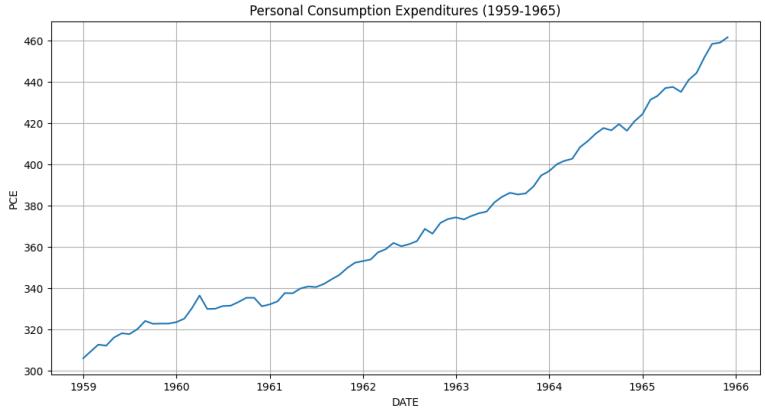


Figure 2: Personal Consumption Expenditures (1959-1965)

Figure 2 Concentrates on a defined temporal span from 1959 to 1965, facilitating a more detailed analysis of fluctuations in the data. Although the rising trend remains apparent, this deeper perspective highlights short-term variations in the PCE levels.

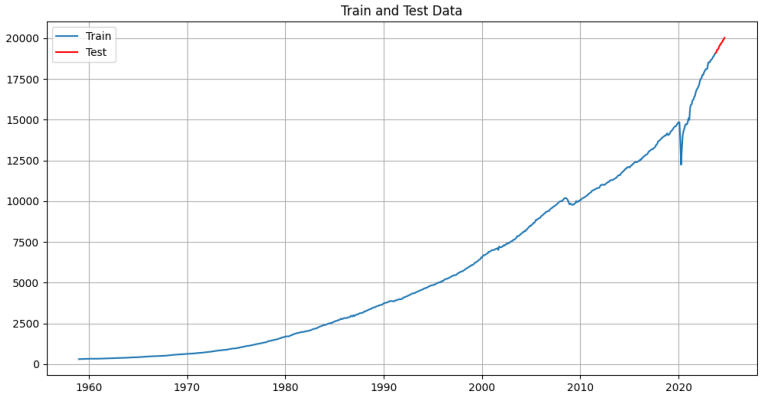


Figure 3: Train and Test Split of Personal Consumption Expenditures

Figure 3 The dataset was partitioned into training and testing subsets to assess the efficacy of predicting models. The training set consists of data from January 1959 to September 2023, encompassing 777 months, while the testing set comprises the final 12 months, extending from October 2023 to September 2024.

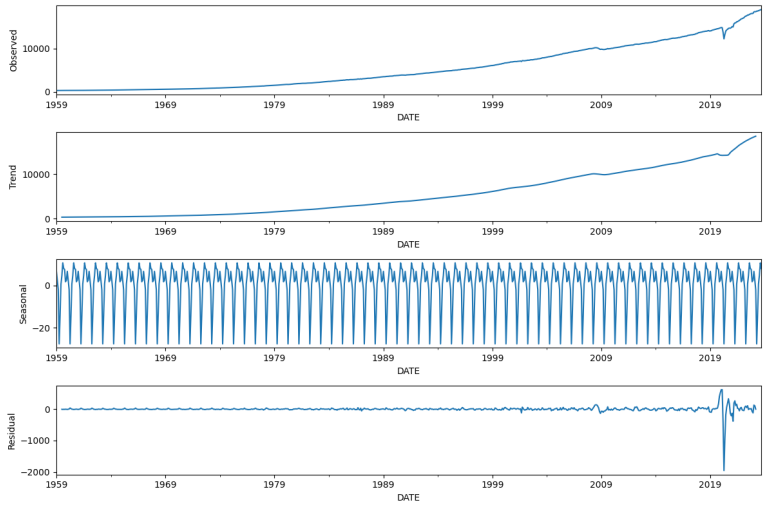
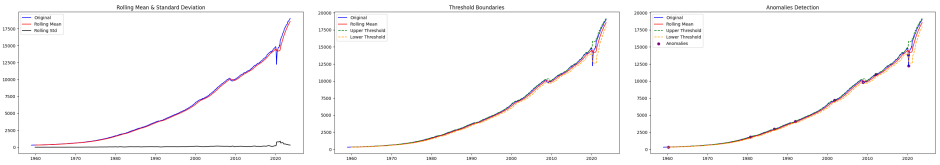


Figure 4: Decomposition of the Train Data

Figure 4 demonstrates the decomposition of the training data into its fundamental elements: trend, seasonality, and residuals. The trend component emphasizes the fundamental upward trend in the series. The residual component signifies the irregular fluctuations. The residuals exhibit increased fluctuation in recent years, potentially due to significant events such as economic crises or other disturbances.

3 Data Preprocessing

- **Handling Missing Values:** There is no missing values in the data.
- **Anomaly Detection:**



(a) Rolling Mean and sd (b) Threshold Boundaries (c) Anomalies Highlighted

Figure 5: Anomaly Detection in Personal Consumption Expenditures Data

Figure 5Demonstrates the procedure of anomaly detection inside the PCE dataset.The initial graph presents the computed moving statistics, encompassing the rolling mean and rolling standard deviation, which facilitate the identification of fluctuations and patterns within the time series data. The second panel presents threshold boundaries established by rolling statistics, including the rolling mean and standard deviation. The third panel emphasizes the identified anomalies that exceed the threshold limits. These anomalies represent substantial deviations from the anticipated trend and must be rectified to enhance the precision of forecasting models.

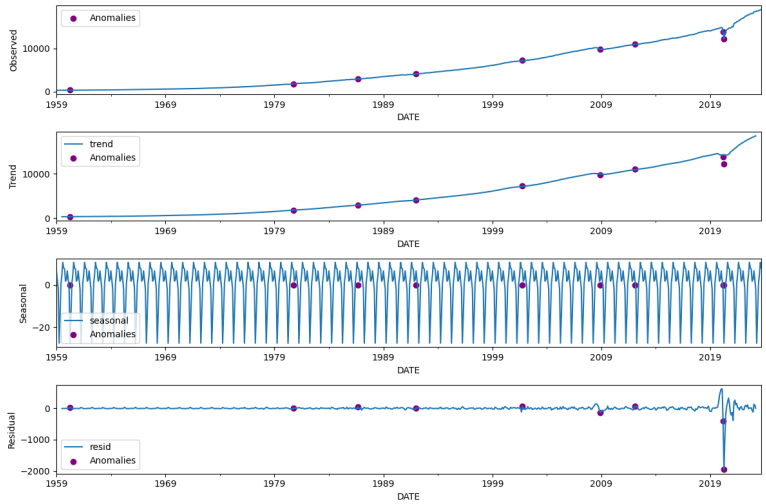


Figure 6: Anomaly Detection in Personal Consumption Expenditures Data

Figure 6 The graph breaks down the time series into its main parts: Observed, Trend, Seasonal, and Residuals, and points out any anoma-

lies in each section. The observed component shows important differences from what was expected. The Trend component shows a steady upward pattern, with unusual spikes that correspond to sudden changes related to economic events. The Seasonal component shows patterns that repeat, but there are some unusual cases that don't fit these expected cycles. The Residuals component shows unusual changes, where anomalies point to times of unexpected volatility. Some major anomalies are the huge drop during the COVID-19 pandemic, which was the biggest monthly decrease in the dataset, notable changes during the economic instability of the 1980s, and the steep decline in 2008 during the global financial crisis.

- **Transformations:** Applied transformations, which is the Box-Cox transformation, to stabilize variance.

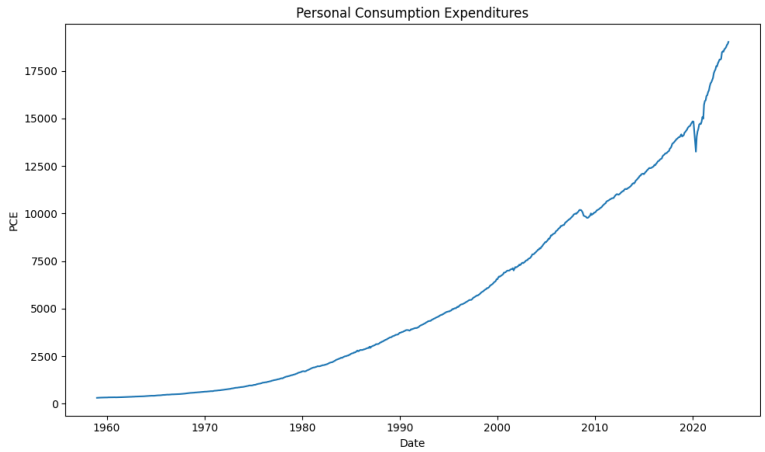


Figure 7: Transformed PCE Data after Removing Anomalies and Applying Box-Cox Transformation

Figure 7 represents the Personal Consumption Expenditures (PCE) data subsequent to the elimination of anomalies and the use of the Box-Cox transformation. The elimination of anomalies has reduced irregular fluctuations, while the Box-Cox transformation has stabilized the data's variance.

- **Stationarity Check:** Conducted unit root tests, such as the KPSS and ADF tests, to check the stationarity of the series.

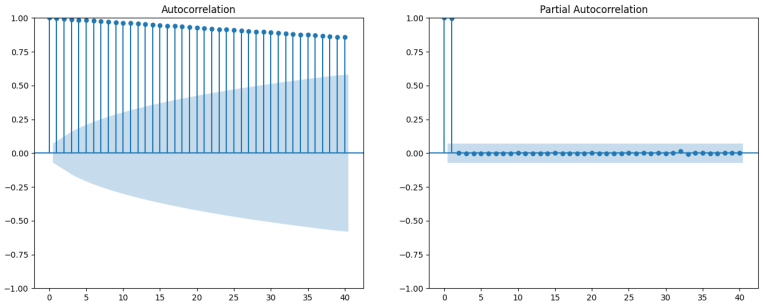


Figure 8: ACF and PACF Plots of the Transformed PCE Data

Figure 8 Displays the ACF and PACF graphs of the transformed PCE data. The ACF plot exhibits a gradual decline, signifying the existence of non-stationarity within the series.

Once variance stationarity was addressed, tests were used to guarantee that the dataset was completely stationary and to get stationarity in the mean. The data set stationarity has been examined using both the KPSS and ADF tests.

The initial level of the KPSS test demonstrated that the data is non-stationary ($p < 0.05$). The KPSS trend test indicated that the series exhibits a stochastic trend ($p < 0.05$). Non-stationarity was confirmed by the ADF test, which revealed that the data had a unit root ($p > 0.05$).

- **Differencing:**

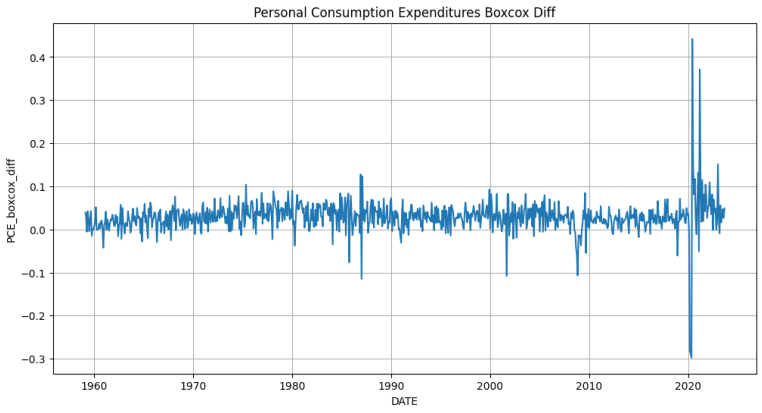


Figure 9: Differenced PCE Data

Figure 9Due to the non-stationarity of the data and its stochastic trend, first-order differencing of the time series was employed to stabilize the mean and eliminate the trend component.

The graph of the differenced data suggests that the process is stationary with a mean of 0. The persistent variations and lack of apparent pattern indicate that the differencing effectively mitigated the non-stationarity in the series.

- **Stationarity Check After Differencing:** Following the implementation of first-order differencing, the stationarity of the modified data was examined again utilizing both the KPSS and ADF tests. The KPSS test produced a p-value of 0.1, exceeding the significance threshold of 0.05. This outcome demonstrated that the differenced series is stationary and prevented the null hypothesis from being rejected. Correspondingly, the ADF test yielded a p-value of roughly 4.59×10^{-21}, which is substantially less than 0.05, so permitting the rejection of the null hypothesis. This demonstrates that the series is stationary and no longer has a unit root.

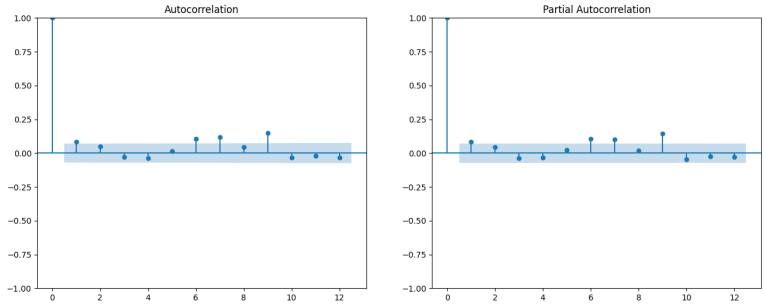


Figure 10: ACF and PACF Plots of the Differenced PCE Data

Figure 10 Similar to the ACF plot, the PACF plot also displays a considerable spike at lag 1. The results from the KPSS and ADF tests, in conjunction with the ACF and PACF plots, affirm that the converted series is stationary and suitable for modeling.

4 Modeling

4.1 Model Selection

The training dataset for ARIMA models was preprocessed by eliminating identified anomalies and implementing a Box-Cox transformation. The differenced series was not utilized directly, as the ARIMA model intrinsically incorporates differencing in the modeling procedure. According to the ACF and PACF plots, ARIMA(1,1,1) was initially proposed as an appropriate model. The AutoARIMA function was utilized to determine the optimal model by evaluating several ARIMA configurations through Maximum Like-

lihood Estimation (MLE) scores. The optimal model for the dataset, as identified by AutoARIMA, is ARIMA(1,1,0).

4.2 Compare Model

Metric	**ARIMA(1,1,0)**	**ARIMA(1,1,1)**
Log Likelihood	1411.961	1412.302
AIC	-2817.921	-2816.605
BIC	-2803.959	-2797.988
HQIC	-2812.550	-2809.443
Intercept (p-value)	0.000	0.000
AR.L1 (p-value)	0.000	0.014
MA.L1 (p-value)	-	0.085
Sigma2 (p-value)	0.000	0.000
Ljung-Box (Q) Prob	0.92	0.92
Jarque-Bera (JB)	52635.92	55457.16
Heteroskedasticity (H)	5.27	5.29
Skew	0.92	1.06
Kurtosis	43.31	44.37

Table 2: Comparison of ARIMA(1,1,0) and ARIMA(1,1,1) Models

The comparison of ARIMA(1,1,0) and ARIMA(1,1,1) models underscores variations in their performance according to essential criteria. The ARIMA(1,1,0) model demonstrates marginally superior AIC and BIC values, suggesting a more advantageous match based on these metrics. Moreover, all parameters in the ARIMA(1,1,0) model are statistically significant.

In summary, ARIMA(1,1,0) demonstrates superior efficiency for AIC, BIC, and model simplicity, rendering it the optimal selection.

5 Diagnostic Checking for ARIMA Model

5.1 Normality Check

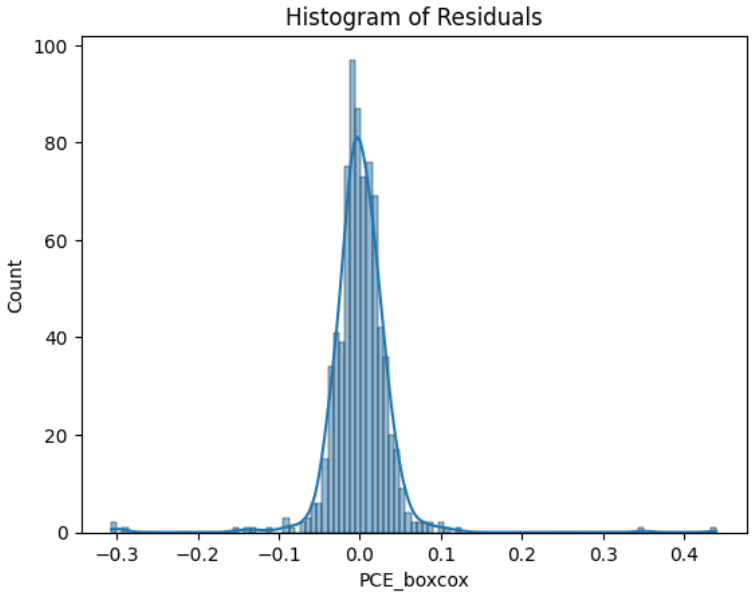


Figure 11: Histogram of Residuals for ARIMA(1,1,0)

Figure 18 The histogram of residuals indicates a nearly symmetric distribution, although there is some indication of slight skewness and heavy tails.

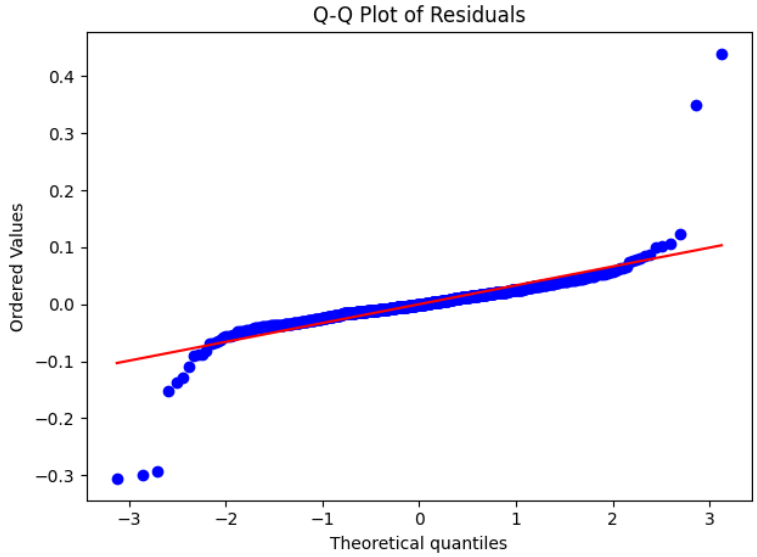


Figure 12: Q-Q Plot of Residuals for ARIMA(1,1,0)

Figure 20 The Q-Q plot exhibits departures from the normal line at the tails, signifying that the residuals are not perfectly normally distributed.

Test	**p-value**	**Conclusion**
Shapiro-Wilk Test	< 0.05	Residuals are not normally distributed
Jarque-Bera Test	0.00	Residuals are not normally distributed

Table 3: Normality Test Results for Residuals

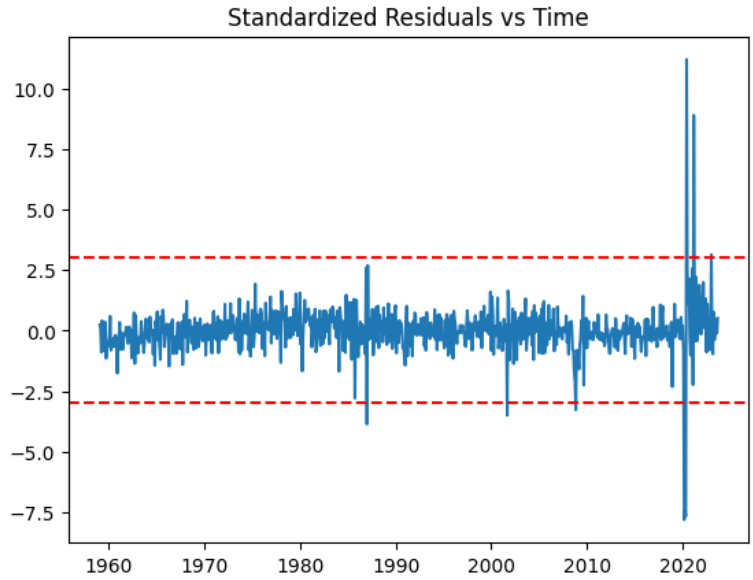


Figure 13: Standardized Residuals of ARIMA(1,1,0) Model

Figure 13 Displays the standardized residuals from the ARIMA(1,1,0) model across time. The residuals are mainly centered around zero, exhibiting no obvious trend.

5.2 Autocorrelation in Residuals

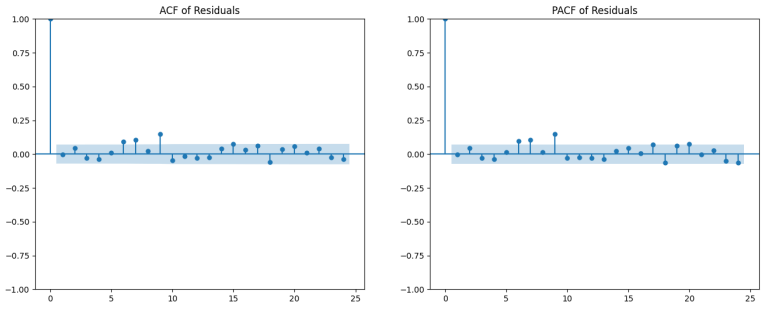


Figure 14: ACF and PACF Plots of Residuals for ARIMA(1,1,0)

Figure 14 The ACF and PACF plots of the residuals demonstrate that most of them of lags fall inside the white noise band, signifying the absence of substantial autocorrelation in the residuals.

The Breusch-Godfrey test was performed to assess serial correlation in the residuals. Given the high p-value in the test results, the null hypothesis—that there is no serial correlation—cannot be rejected. The results of

the Breusch-Godfrey test indicate that there is no significant autocorrelation in the residuals.

5.3 Heteroskedasticity

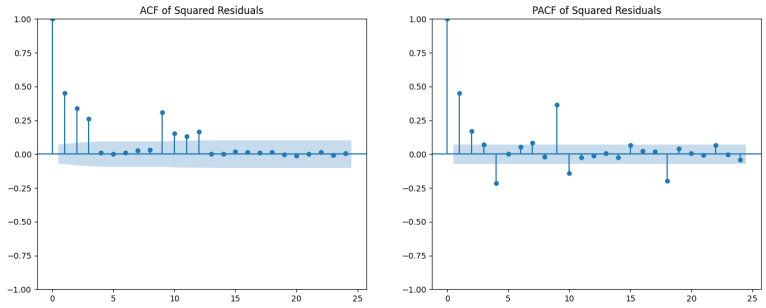


Figure 15: ACF and PACF Plots of Squared Residuals for ARIMA(1,1,0)

Figure 15 displays the squared residuals from the ARIMA(1,1,0) model's ACF and PACF plots. Heteroskedasticity is indicated by significant spikes in the graphs. Furthermore, the ARCH LM test findings demonstrate considerable heteroskedasticity in the residuals. Consequently, a GARCH model will be utilized to address the issue of non-constant variance in the residuals.

5.4 Summary

The ARIMA(1,1,0) model exhibits no substantial autocorrelation in the residuals, as verified by the Ljung-Box Q-test and the Breusch-Godfrey test. The residuals exhibit a non-normal distribution, as indicated by the Shapiro-Wilk and Jarque-Bera tests. The detection of heteroskedasticity, as shown by the ARCH LM test, necessitates the adoption of models capable of handling time-varying variance, such as GARCH.

6 GARCH Model

6.1 Model Fit and Parameters

Parameter	Estimate	Std. Error	t-value	p-value
μ	0.000398	0.000810	0.49161	0.622992
ar1	-0.482886	0.165564	-2.91661	0.003539
ma1	0.365575	0.178408	2.04909	0.040453
ω	0.000412	0.000077	5.34966	0.000000
α_1	0.612899	0.091494	6.69878	0.000000
β_1	0.123234	0.049740	2.47757	0.013228

Table 4: Optimal Parameters of GARCH(1,1) Model with ARFIMA(1,0,1) Mean

The GARCH(1,1) parameters, alpha1 and beta1 , are statistically significant ($p < 0.05$), demonstrating the appropriateness of the GARCH model for modeling heteroskedasticity. The omega parameter is significant, indicating that the model effectively represents the variance dynamics.

6.2 Model Fit Statistics

Metric	Value
Log-Likelihood	1644.094
AIC	-4.2219
BIC	-4.1859
Shibata	-4.2220
Hannan-Quinn	-4.2080

Table 5: GARCH(1,1) Model Fit Statistics

6.3 Residual Diagnostics

Test	Lag	p-value
Weighted Ljung-Box Test on Squared Residuals	1	0.9680
	5	0.9868
	9	0.9988
Weighted ARCH LM Test	3	0.7406
	5	0.8855
	7	0.9737

Table 6: Residual Diagnostics for GARCH(1,1) Model

Discussion: The Ljung-Box test on squared residuals reveals higher p-values across all lags, signifying the absence of significant autocorrelation in the squared residuals. This verifies that the GARCH(1,1) model effectively captures time-varying volatility. The Weighted ARCH LM test indicates no significant ARCH effects in the residuals, hence affirming the model's adequacy.

6.4 Sign Bias Test

Test	t-value	p-value	Significance
Sign Bias	0.1346703	0.8929077	Not Significant
Negative Sign Bias	0.7857006	0.4322843	Not Significant
Positive Sign Bias	0.7335744	0.4634312	Not Significant
Joint Effect	2.0047853	0.5714141	Not Significant

Table 7: Sign Bias Test Results for GARCH Model Residuals

Discussion: The sign bias, negative sign bias, and positive sign bias tests have high p-values , signifying the absence of systematic bias in the model .

6.5 GARCH Model Forecast

Metric	Value
ME (Mean Error)	19553.21
RMSE (Root Mean Squared Error)	19555.46
MAE (Mean Absolute Error)	19553.21
MPE (Mean Percentage Error)	100
MAPE (Mean Absolute Percentage Error)	100

Table 8: GARCH Model Error Metrics on Test Set

The error metrics indicate that the GARCH model exhibits a significant degree of prediction error, in both absolute and percentage terms. This indicates that the model is not equipped to precisely represent the dynamics of the test set.

7 Forecasting

7.1 Arima Pipeline

I check the diagnostics of ARIMA model in section 5. Since the ARIMA model uses BOX-COX transformation, need to reverse the transformation to get the actual values for forecasting. To do that, sktime Transformer

Pipeline is used. This pipeline applies the inverse Box-Cox transformation to the forecasted values. I have found this method in the official documentation of sktime. $https://www.sktime.net/en/latest/api_reference/auto_generated/sktime.forecasting.co$

7.2 Prophet

7.2.1 Prophet Hyperparameter Tuning

Parameter	Value
`changepoint_prior_scale`	0.01
`holidays_prior_scale`	0.01
`seasonality_mode`	Additive
`seasonality_prior_scale`	10.0

Table 9: Hyperparameters of the Final Tuned Prophet Model

7.2.2 Diagnostic Checking - Normality Check

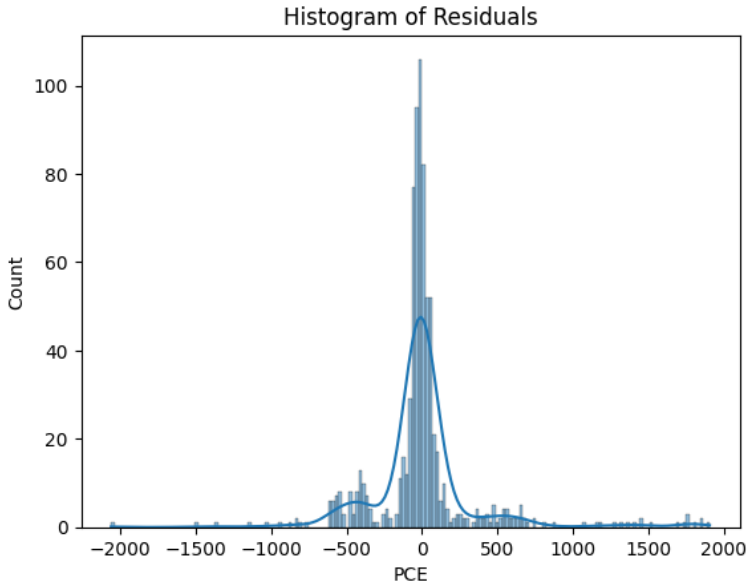


Figure 16: Histogram

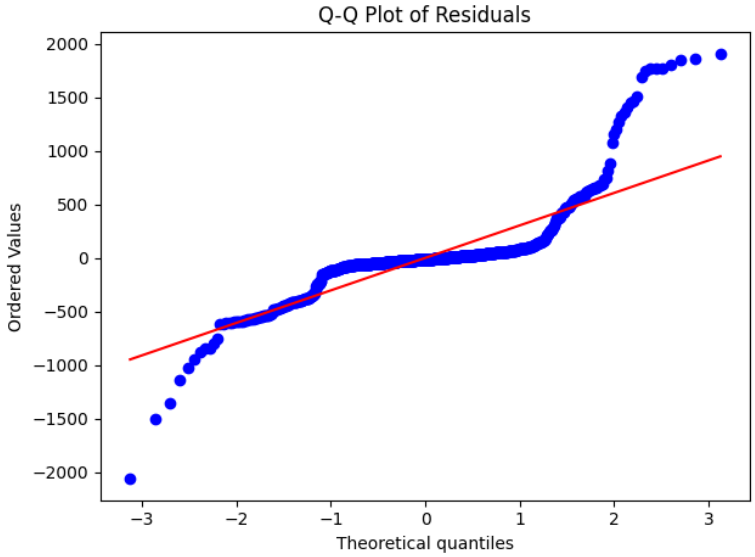


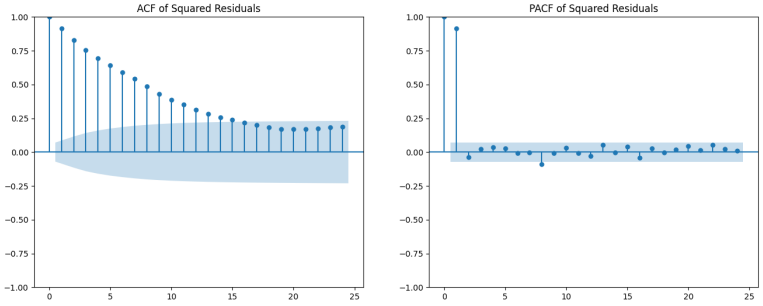
Figure 17: Q-Q Plot

Metric	Value
`Test Statistic`	0.7226651829190489
`p-value`	$5.754953666004976 \times 10^{-34}$

Table 10: Shapiro-Wilk Test Results for Residuals

The p value of the Shapiro-Wilk test is smaller than 0.05, so we can reject the H0 which is residuals is are distributed normally. Residuals are not distributed normally.

7.2.3 Diagnostic Checking -Heterockedasticity Check



Metric	Value
LM Statistic	697.0225797685475
P-value	$2.7360556607469145 \times 10^{-143}$
F Statistic	753.0272887484591
F-test P-value	0.0

Table 11: ARCH LM Test Results for Residuals

The ARCH LM test indicates significant heteroskedasticity in the residuals of the model.

7.3 ETS

7.3.1 Diagnostic Checking - Normality Check

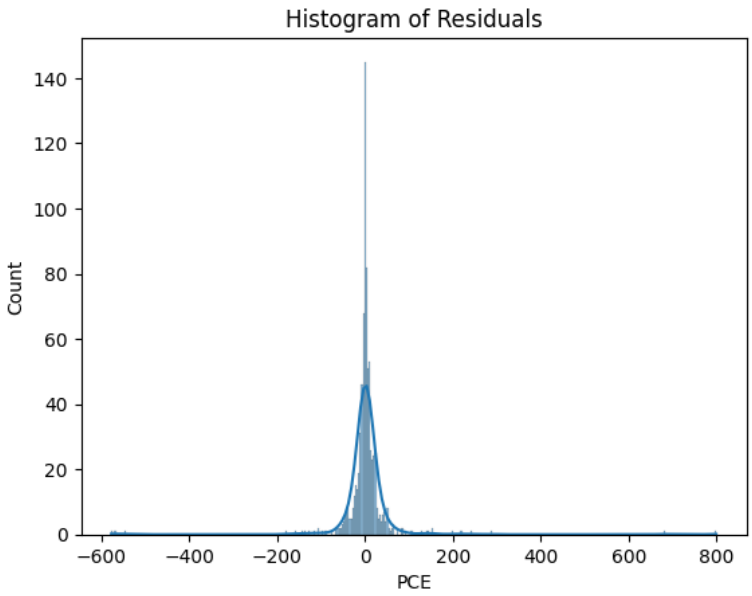


Figure 18: Histogram

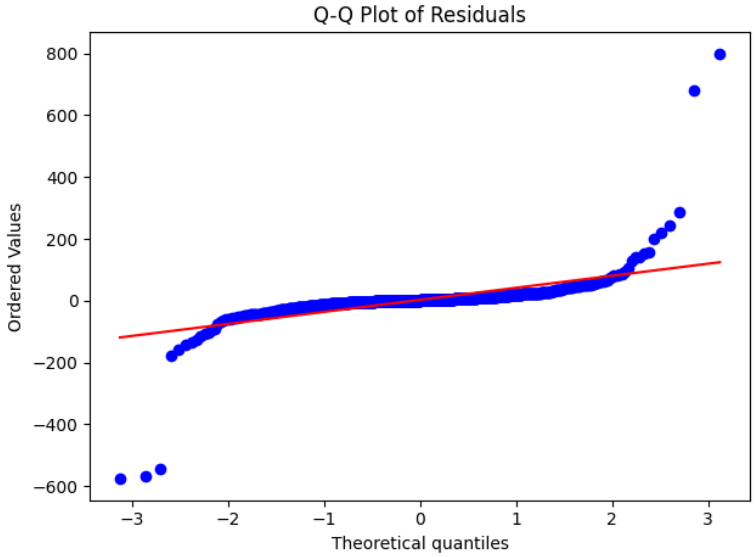


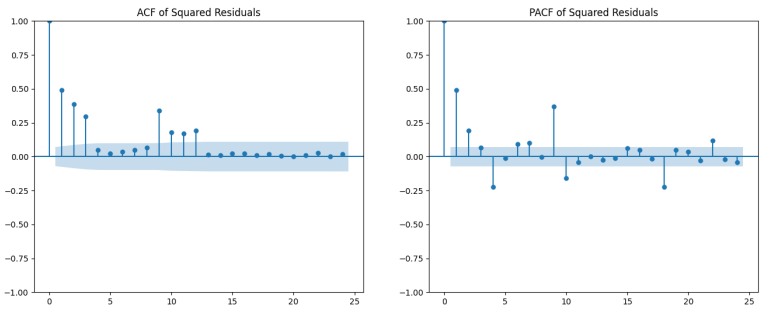
Figure 19: Q-Q Plot

Test	**Result**
Shapiro-Wilk Test Statistic	0.4135483062180507
Shapiro-Wilk p-value	$2.1383357701128927 \times 10^{-44}$
Jarque-Bera Test p-value	0.0

Table 12: Normality Test Results for Residuals

The p value of the Shapiro-Wilk test is smaller than 0.05, so we can reject the H0 which is residuals is are distributed normally. Residuals are not distributed normally. The p value of the Jarque-Bera test is smaller than 0.05, so we can reject the H0 which is residuals is distribudted normally. Residuals are not distributed normally.

7.3.2 Diagnostic Checking - Heterockedasticity Check



Metric	Value
LM Statistic	330.1356443653315
P-value	$6.50102213372923 \times 10^{-65}$
F Statistic	57.1304444291312
F-test P-value	$1.2057188078816973 \times 10^{-85}$

Table 13: ARCH LM Test Results for Residuals

The ARCH LM test results indicate that there is significant heteroskedasticity in the residuals.

7.4 TBATS

7.4.1 Diagnostic Checking - Normality Check

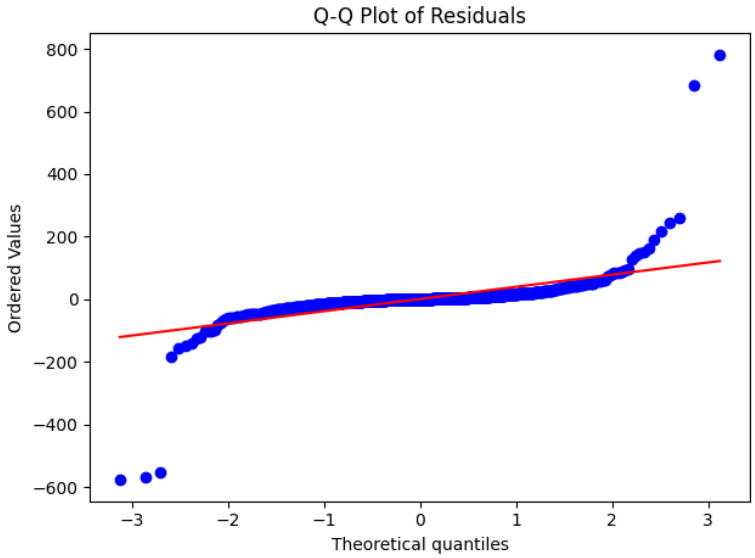


Figure 20: Q-Q Plot

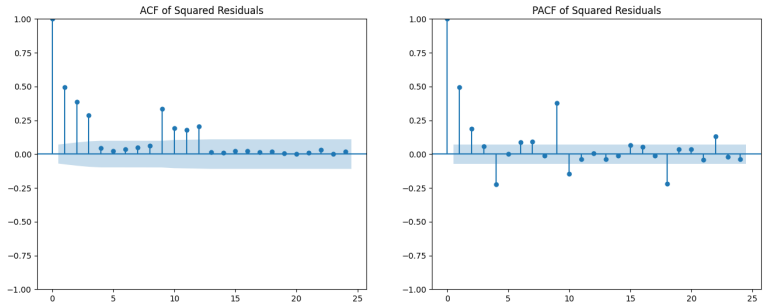
Test	Result
Shapiro-Wilk Test Statistic	0.4158152815368523
Shapiro-Wilk p-value	$2.442085899892703 \times 10^{-44}$
Jarque-Bera Test p-value	0.0

Table 14: Normality Test Results for Residuals

The p value of the Shapiro-Wilk test is smaller than 0.05, so we can reject the H0 which is residuals is are distributed normally. Residuals are not distributed normally. The p value of the Jarque-Bera test is smaller than

0.05, so we can reject the H0 which is residuals is distribudted normally. Residuals are not distributed normally.

7.4.2 Diagnostic Checking - Heterockedasticity Check



Metric	Value
LM Statistic	329.13137623956084
P-value	$1.061191602853258 \times 10^{-64}$
F Statistic	56.82602199266986
F-test P-value	$2.83748378193945 \times 10^{-85}$

Table 15: Updated ARCH LM Test Results for Residuals

The ARCH LM test results indicate that there is significant heteroskedasticity in the residuals.

7.5 FNN

For Machine Learning Model, I have tried to use sktime's NeuralForecastRNN and PytorchForecastingNBeats models Neither of them worked, they gave errors.

7.6 PLOT

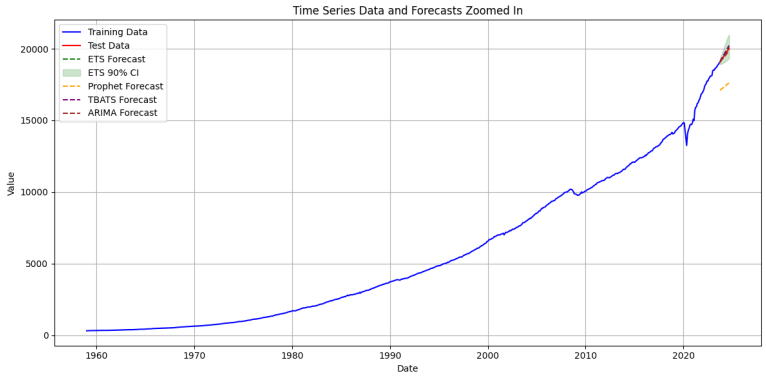


Figure 21: All Model Forecasts with Train Data included

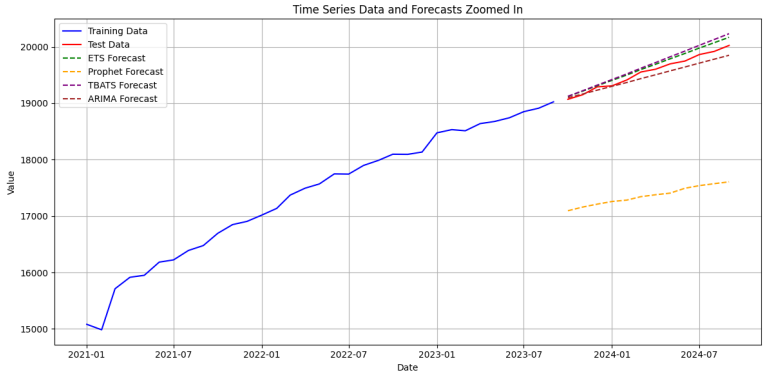


Figure 22: All Model Forecasts with Train Data included - Zoom In

7.7 Comparison

Metric	ETS	Prophet	TBATS	ARIMA
(MSE)	9842.4455	4818450.2461	17751.4722	10721.5941
(RMSE)	99.2091	2195.0969	133.2347	103.5451
(MAPE)	0.0046	0.1116	0.0061	0.0044

Table 16: Comparison of Error Metrics for Different Models

8 Conclusion

When looking at the different forecasting models like ETS, Prophet, TBATS, and ARIMA, it turns out that ETS and ARIMA do a better job than the others, showing much lower values for MSE, RMSE, and MAPE. ETS shows the

lowest RMSE at 99.21 and a competitive MAPE of 0.0046, which suggests it has strong accuracy in predicting the test data. ARIMA shows comparable performance, with a slightly higher RMSE of 103.55 and MAPE of 0.0044, positioning it as another solid option. TBATS performs is good, but it's not as accurate as ETS and ARIMA. Prophet doesn't do well, showing high error values in all metrics, which means it's not a good fit for this dataset.

References

- Hyndman, R. J.,
  "Forecasting: Principles and Practice."
- Box, G. E. P., Jenkins, G. M.,
  "Time Series Analysis: Forecasting and Control."
- Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, Franz Király (2019): "sktime: A Unified Interface for Machine Learning with Time Series"