INTRODUCTION & DATASET:

This paper presents an analysis of the yearly export data of the CAF from 1960 to 2017, employing an Autoregressive Integrated Moving Average (ARIMA) model to understand the underlying patterns and forecast future trends.

Seen below are various variables, and to provide insight into its structure, the first row of data is displayed below. While some columns may exhibit missing values, our project primarily centers on the "Exports" variable, which notably does not contain any missing data.

Country	Code	Year	GDP	Growth	CPI	Imports	Exports	Population
Central African Republic	CAF	1960	1.121556e+08	NaN	NaN	34.181812	23.272724	1503508

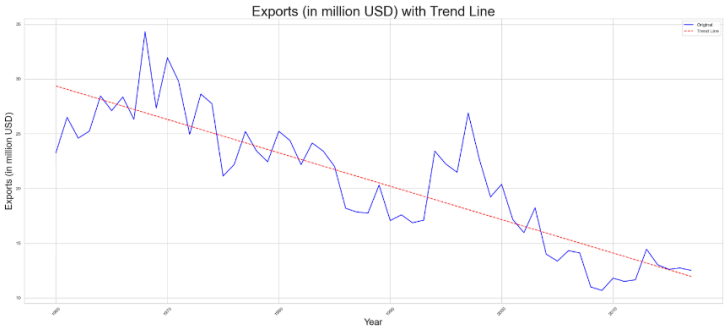


Figure 1

```
engle_test = het_arch(df['Exports'])
print('Test statistic: ', engle_test[0])
print('P-value: ', engle_test[1])
```

```
Test statistic:  39.967911653709294
P-value:  1.7166628139057225e-05
```

Figure 2

Figure 1 depicts the time series plot illustrating the CAF's exports spanning from 1960 to 2017. The blue line captures fluctuations in export values and significant variation, reflecting distinct periods of growth and decline.

NON-STATIONARITY:

A time series achieves stationarity when its statistical attributes, including mean, variance, and autocorrelation, remain constant across time. Figure 1reveals an overall downward trend and periods where fluctuations exhibit varying degrees of intensity, suggesting the presence of non-constant mean and variance. To quantify heteroscedasticity, I employed Engle's Autoregressive Conditional Heteroscedasticity (ARCH) test, a method which contrasts the null hypothesis of constant variance against its alternative. As depicted in Figure 2, the resulting p-value was 0.0000171, leading us to reject the hypothesis of constant variance.

Next, we examined Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, alongside conducting the Augmented Dickey-Fuller (ADF) test. The ACF plot (Figure 3)

displayed a gradual decay, indicating that current values in the series have a prolonged correlation with past values, which violates the conditions for stationarity.

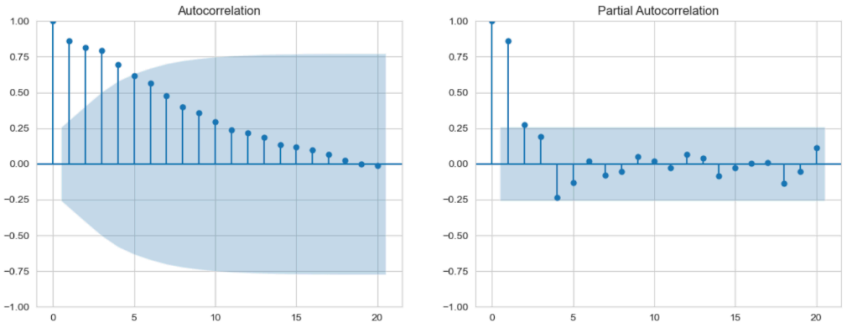


Figure 3

```
In [10]: adf_test = adfuller(df['Exports'])

         print('Test statistic: ' , adf_test[0])
         print('p-value: '  ,adf_test[1])
         print('Critical Values:' ,adf_test[4])
```

```
Test statistic:  -0.808241033526914
p-value:  0.816689332299944
```

Figure 4

Finally, we conducted the Augmented Dickey-Fuller (ADF) test, which evaluates the presence of a unit root in a time series sample, indicating non-stationarity. The test results, depicted in Figure 4, revealed a p-value of 0.8167, reinforcing our earlier observations of non-stationarity from trend analysis, examination of non-constant variance, and autocorrelation plots.

TREATING NON-CONSTANT VARIANCE:

Based on our findings, the subsequent steps were aimed at addressing non-constant variance, mean, and autocorrelation. While techniques like differencing could be immediately applied, we initially explored various data transformations to help stabilize variance and mitigate trend effects.

Our first task was to identify the most suitable transformation among the square root, logarithmic, and Box-Cox transformations. These were chosen due to their recognized effectiveness in stabilizing variance, reducing skewness, and normalizing distributions. Upon plotting histograms and calculating skewness for each transformation, as depicted in Figure 5, we observed that the original dataset and the

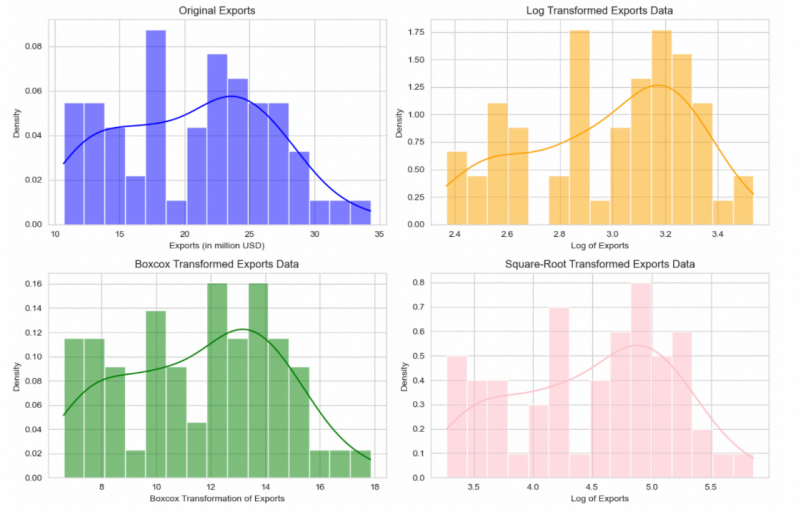


Figure 5

```
{'Original Skewness': 0.03397663478316858,
 'Log Skewness': -0.4174993355111485,
 'Square Root Skewness': -0.1983603589685804,
 'Box-Cox Skewness': -0.08091612000497714}
```

Figure 6

```
Shapiro-Wilk Test P-value (Original): 0.1145
Shapiro-Wilk Test P-value (Log Transformed): 0.0143
Shapiro-Wilk Test P-value (Sqrt Original): 0.0620
Shapiro-Wilk Test P-value (BoxCox Transformed): 0.0954
```

Figure 7

Box-Cox transformation exhibited the lowest skewness values, indicating their effectiveness in symmetrizing the distribution. The results of this analysis are presented in Figure 6.

Further examination utilizing Quantile-Quantile (QQ) plots and the Shapiro-Wilk normality test reinforced our findings. Both the original and the Box-Cox transformed data produced p-values that upheld the null hypothesis of normality, with these p-values illustrated in Figure 7. Interestingly, the original data's slightly higher p-value compared to the Box-Cox transformed data implied that a transformation might not be imperative prior to differencing. Nonetheless, Engle's Test was applied on the transformed data and the results still pointed towards non-constant variance and non-stationarity. This indicates that while transformation helps in symmetrizing the distribution, further steps are required to address the data's underlying non-stationarity.

DIFFERENCING:

Our next step was to apply differencing: a technique aimed at achieving stationarity by focusing on the change between consecutive observations to eliminate trends and stabilize mean and variance. After applying first-order differencing to the original, logarithmically transformed, and Box-Cox transformed data, they all resulted in series exhibiting stationary behavior with no discernible trend. Given that the plots for all three differed series appeared identical, we chose to present just a single plot of the differenced data in Figure 8. The trend lines fitted to this data were nearly flat, showcasing a minimal change in mean over time.

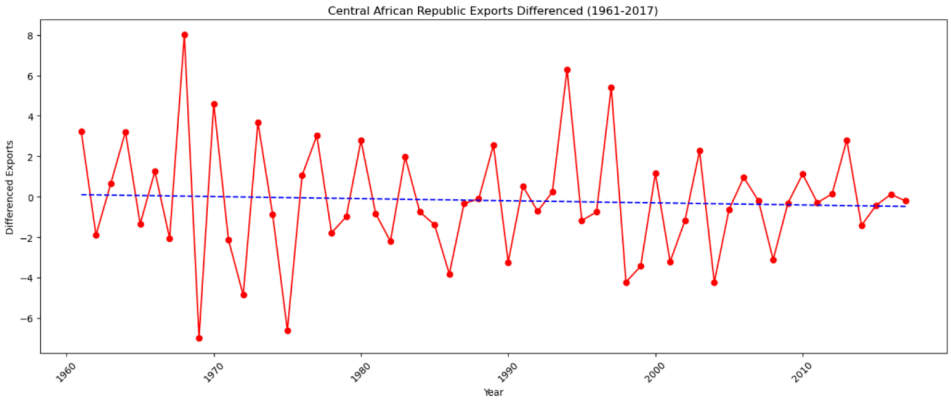


Figure 8

To formally assess the stationarity of our differenced data, we recalculated the ADF statistic for all three datasets, yielding the corresponding p-values:

```
{'Original First-Difference p-value': {'p-value': 0.00026339432031339437},
 'Log Difference p-value': {'p-value': 0.0006794976404052878},
 'BoxCox Difference p-value': {'p-value': 0.0003430359582321000 6}}
```

While all three forms of data demonstrated stationarity, the original non-transformed data reaffirmed our earlier findings, exhibiting the lowest p-value among the three. Furthermore, conducting Engle's Test on the original differenced data provided a p-value of 0.114, finally affirming constant variance.

The analysis done so far suggests that the data, without transformation, satisfies the critical assumptions for effective time series modeling and outperforms the transformed data in diagnostics. However, as this report unfolds, you will observe that I opted for the Box-Cox transformed data. This decision emerged from a critical insight gained during the subsequent model selection phase. Specifically, as shown in Figure 9, models derived from transformed data consistently showed lower AIC values, indicating better performance by revealing more intricate patterns in the data not captured in the original dataset models. Despite initial analyses suggesting transformations were unnecessary, the lower AIC values highlighted the benefits of using transformed data for optimizing model performance. This illustrates that statistical tests, while vital, are not the sole factors in preprocessing decisions.

	Combination	AIC: Transformed-Difference Data
11	(2, 0, 3)	188.045987
14	(3, 0, 2)	188.268137
15	(3, 0, 3)	188.282259
12	(3, 0, 0)	188.335023
10	(2, 0, 2)	188.604349
8	(2, 0, 0)	188.652457
3	(0, 0, 3)	189.212058
9	(2, 0, 1)	189.551677
13	(3, 0, 1)	189.927285
1	(0, 0, 1)	190.125980
2	(0, 0, 2)	190.784703
6	(1, 0, 2)	191.796861
5	(1, 0, 1)	191.829665
7	(1, 0, 3)	191.899193
4	(1, 0, 0)	192.285194

	ombination	AIC: Non-Transformed Difference Data
	(2, 0, 3)	274.942599
	(3, 0, 2)	275.117325
	(3, 0, 0)	275.179853
	(2, 0, 0)	275.253517
	(3, 0, 3)	275.297676
	(2, 0, 2)	275.377847
	(0, 0, 3)	275.796723
	(2, 0, 1)	276.284892
	(3, 0, 1)	276.710022
	(0, 0, 1)	276.820924
	(0, 0, 2)	276.927502
	(1, 0, 2)	277.980342
	(1, 0, 3)	278.000789
	(1, 0, 1)	278.361020
	(1, 0, 0)	278.944140

Figure 9

MODEL SELECTION

ARIMA(p,d,q) is a statistical model used for forecasting time series data, combining AutoRegressive (AR) and Moving Average (MA) components, with 'p' and 'q' indicating their orders, respectively. AR models leverage past values for predictions, while MA uses past errors. The 'I' stands for 'Integrated', signified by 'd', which involves differencing data to achieve stationarity. Analysis of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots is essential for selecting initial model parameters. As highlighted in Figure 10, the ACF plot's rapidly decreasing autocorrelations and the PACF plot's spike at the first lag suggest a stationary series, guiding the preliminary selection of ARIMA models, specifically ARIMA(2,0,1) and ARIMA(2,0,3).

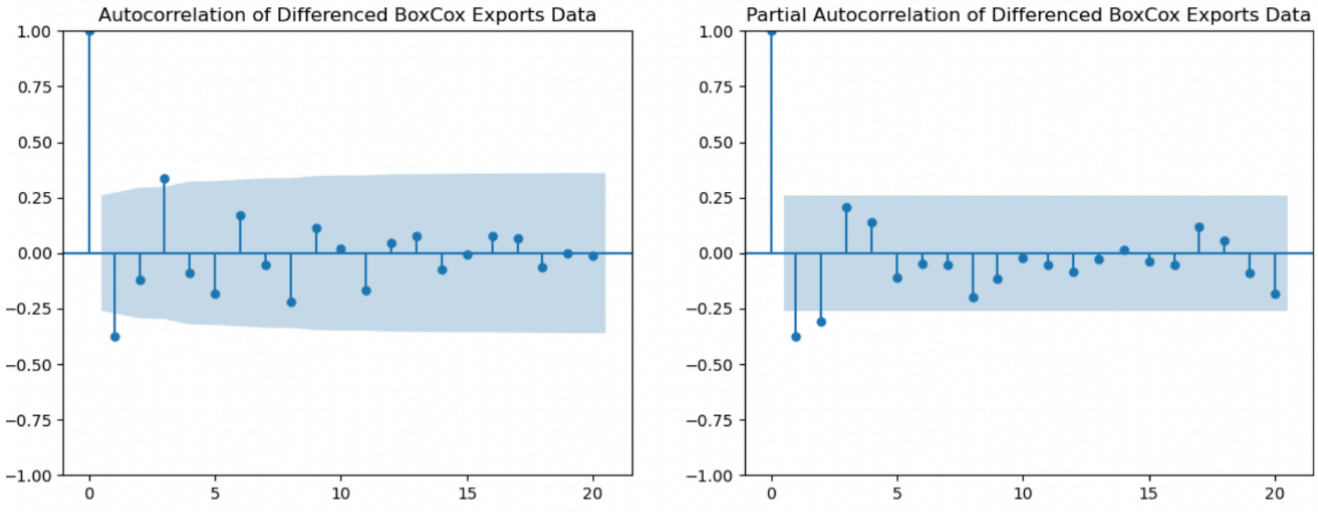


Figure 10

To refine our model selection process, we expanded our search beyond the insights gained solely from ACF and PACF plots. We conducted extensive testing of ARIMA model combinations, exploring different values for parameters 'p' and 'q', as illustrated earlier in Figure 9. AIC was used to identify the most suitable model, which, in line with initial insights from the ACF and PACF, was ARIMA(2,0,3). This model's adequacy was confirmed by plotting the ACF of residuals and performing the Ljung-Box test, which indicated that the residuals resembled white noise, as shown in Figure 11. Yet, a closer look at parameter significance revealed that all MA parameters were statistically insignificant, prompting the exploration of alternative models. Ultimately, ARIMA(2,0,0) emerged as an alternative choice, characterized as the model that had the lowest AIC with statistically significant parameters and whose residuals that looked like white noise.

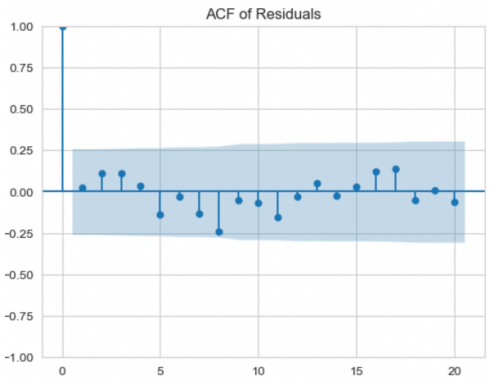


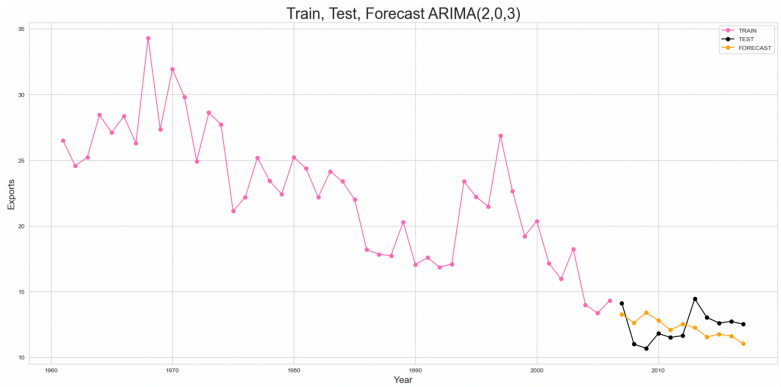
Figure 11

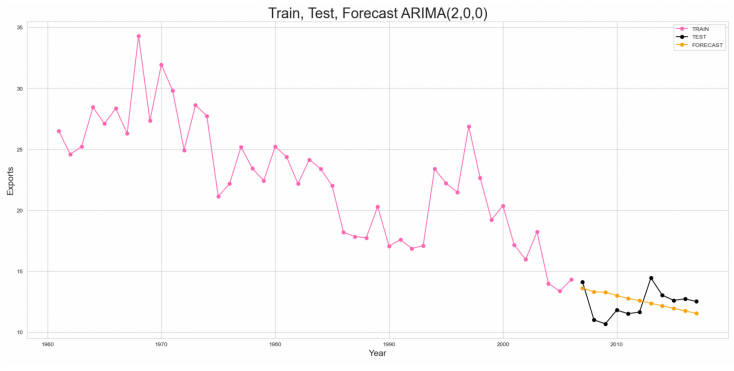
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1064	0.113	-0.943	0.346	-0.328	0.115
ar.L1	-0.6633	0.068	-9.685	0.000	-0.797	-0.529
ar.L2	-0.9757	0.066	-14.736	0.000	-1.106	-0.846
ma.L1	0.2768	1.183	0.234	0.815	-2.043	2.596
ma.L2	0.8376	3.858	0.217	0.828	-6.724	8.399
ma.L3	-0.2836	1.257	-0.226	0.822	-2.748	2.180
sigma2	1.1485	4.728	0.243	0.808	-8.118	10.415

Figure 12

FORECASTING:

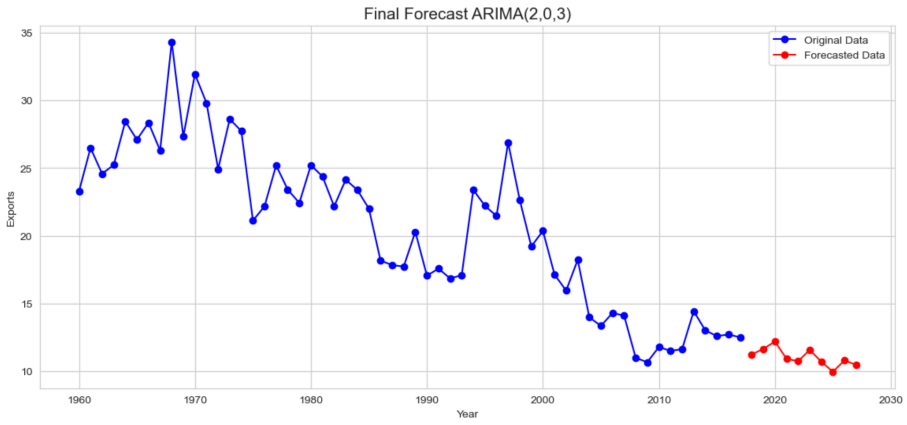
Upon selecting ARIMA(2,0,0) and ARIMA(2,0,3) as our candidate models, I adopted an 80-20 split for training and testing. The models were fitted on the training set to gauge their performance, using RMSE for accuracy assessment. The ARIMA(2,0,0) model yielded an RMSE of 0.743, while the ARIMA(2,0,3) model slightly trailed with an RMSE of 0.76. Post-training, I reversed any differencing and transformations to align the forecasts with the original data scale, a critical step for accurate comparison. The comparison between the model forecasts and the actual test set was visually represented through the following plots that visualized the training set, test set, and the forecasts from both ARIMA models:





FINAL FORECAST

Motivated by these results, I applied the ARIMA(2,0,3) model to the entire dataset, projecting ten years into the future. This involved a similar procedure of reversing differencing and transformations for the final forecasts. The final forecasts were summarized in subsequent plots and analyses:



Dep. Variable:	Boxcox_Diff	**No. Observations:**	57
Model:	ARIMA(2, 0, 3)	**Log Likelihood**	-87.023
Date:	Sat, 16 Mar 2024	**AIC**	188.046
Time:	21:21:53	**BIC**	202.347
Sample:	0	**HQIC**	193.604
	- 57		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1064	0.113	-0.943	0.346	-0.328	0.115
ar.L1	-0.6633	0.068	-9.685	0.000	-0.797	-0.529
ar.L2	-0.9757	0.066	-14.736	0.000	-1.106	-0.846
ma.L1	0.2768	1.183	0.234	0.815	-2.043	2.596
ma.L2	0.8376	3.858	0.217	0.828	-6.724	8.399
ma.L3	-0.2836	1.257	-0.226	0.822	-2.748	2.180
sigma2	1.1485	4.728	0.243	0.808	-8.118	10.415

Ljung-Box (L1) (Q):	0.02	**Jarque-Bera (JB):**	1.81
Prob(Q):	0.89	**Prob(JB):**	0.41
Heteroskedasticity (H):	0.40	**Skew:**	0.26
Prob(H) (two-sided):	0.05	**Kurtosis:**	3.70

This study embarked on a comprehensive exploration of time series forecasting, meticulously analyzing the statistical properties of our dataset to ascertain the most suitable model for predicting future values. Through a detailed examination involving visualizations, statistical tests, transformations, differencing, and model selection, we successfully navigated the complexities of achieving stationarity and optimizing model parameters. Our rigorous approach led us to the ARIMA(2,0,3) model, which, despite initial hesitations regarding the significance of its parameters, demonstrated superior predictive accuracy. This finding not only highlights the importance of thorough statistical analysis but also underscores the nuanced decision-making required in effective time series forecasting.