Introduction and analysis:

Introduction:

The food retail turnover dataset estimated turnover value within the Australian food retail trade sector. Extracted from the monthly retail business survey, the dataset includes 700 large businesses and 2,700 small businesses. The focus of this analysis is specifically on the food retail industry, encompassing supermarket and grocery stores, non-petrol sales (convenience stores) in selected fuel retailing, liquor retailing, and other specialized food retailing. The latter category includes fresh meat, fish, and poultry retailing, fruit and vegetable retailing, among others. The dataset, "FoodRetailTurnover.csv," spans from April 1982 to August 2022, comprising monthly observations and turnover values measured in $Million. This initiative aims to develop a robust time series forecasting model for predicting turnover in the food retail sector for the upcoming 12 months, spanning from September 2022 to August 2023To achieve this, two baseline models, including random walk and last value, alongside the primary method SARIMA, are employed. The dataset contains two variables including the turnover variable serves as a critical attribute, and the date variable. The dataset does not have any missing values; however, the data type of the date variable is currently not in the correct format. Hence, the initial step involves changing the data type of the date variable, which is currently an object, to a date type. The figure 1 illustrates the presence of seasonal patterns in the data. The data is non-normality and non-stationarity, posing challenges for modelling. Various approaches to fix normality have been explored, but achieving complete normalization remains elusive. There are significant seasonal patterns in the turnover, I need to apply seasonal adjustment methods to mitigate bias caused by regular fluctuations. Mean Squared Error (MSE) is deployed as an evaluation metric which is common in time series forecasting.

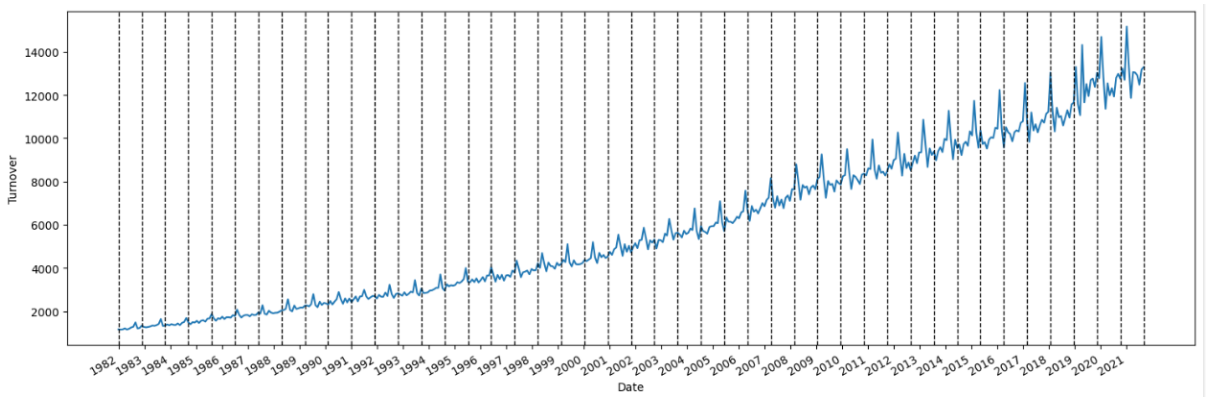


Figure 1: Seasonality behavior

The figure1 & 2 improve that there is an obvious trend and seasonality, so the data is not stationary as the mean and variance is changing over time.

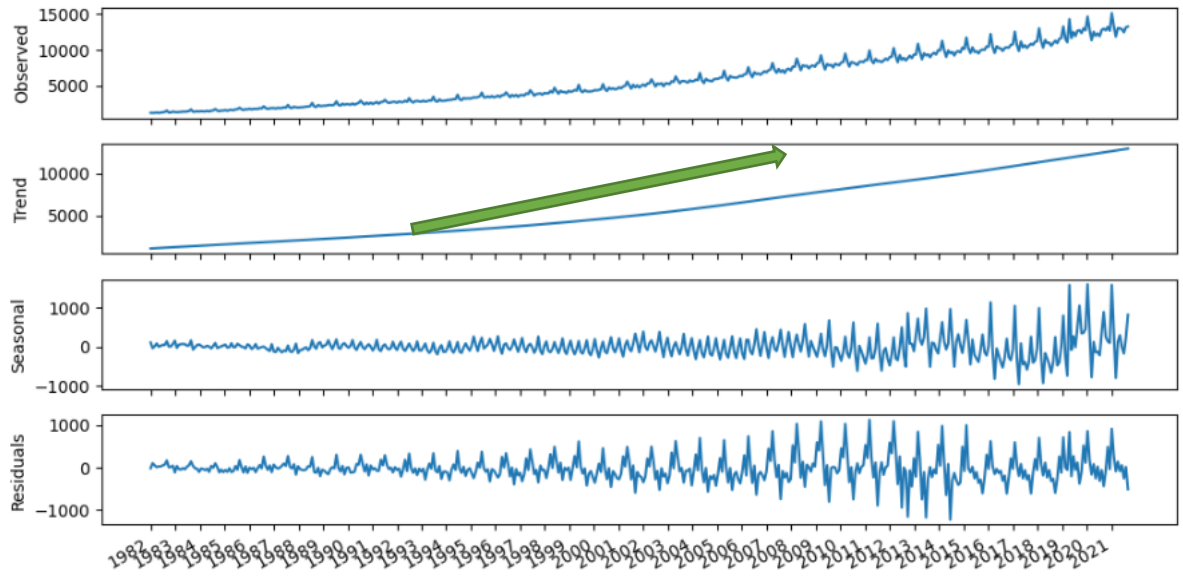


Figure 2: time series decomposition plot

Figure 2 illustrates the following observations:

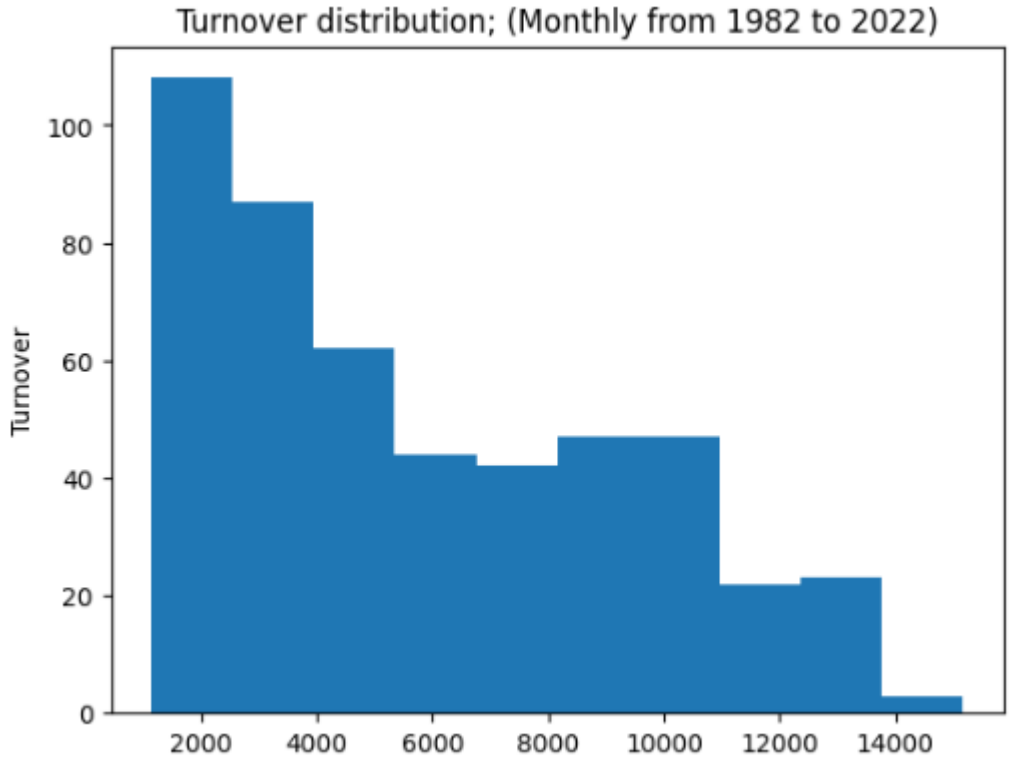
The dataset exhibits a positive trend, indicating a gradual increase over time.

Seasonal patterns are evident in the data, showcasing repetitive increases at specific intervals.

Residuals display white signals, incorporating seasonality and rapid fluctuations.

The normality of data is investigated by using bar plot and p-value.

The Shapiro test indicates that the dataset is not normally distributed. Additionally, the table below illustrates right skewness in the data.



Various techniques, such as Box-Cox, z-score, square root, and log transformation, were applied to address the non-normality of the data. However, none of these transformations succeeded in achieving normality. As a compromise, the square root transformation was chosen

to provide some stabilization to the data, as depicted in Figure 3.

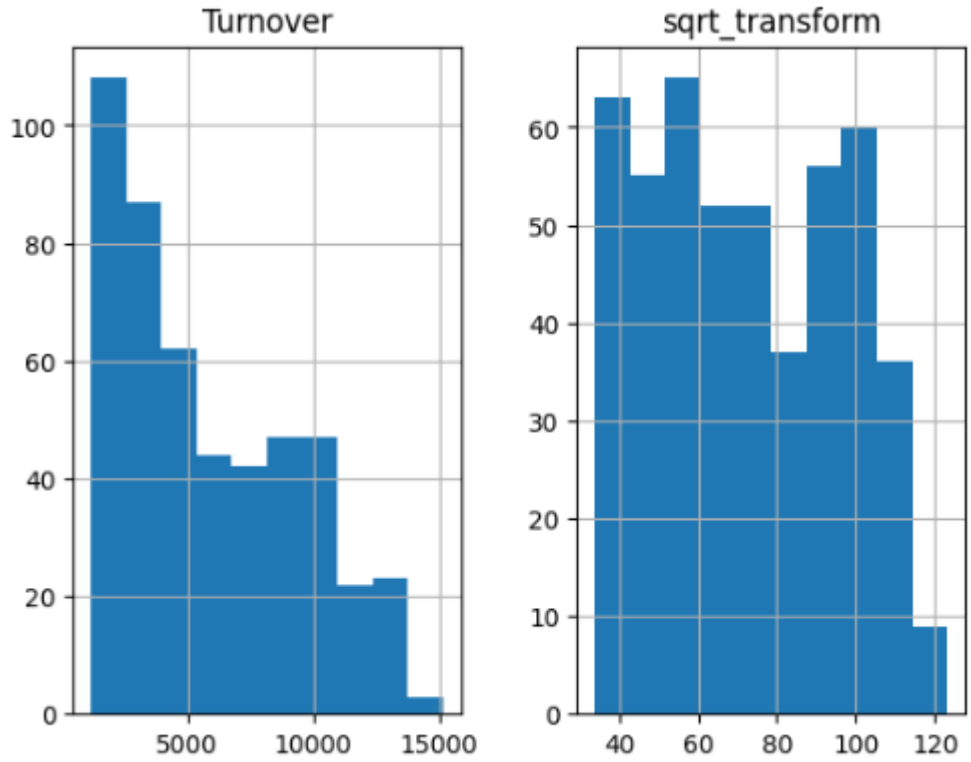


Figure 3: compare original and normalized data

Let's compare the original and transformed data.

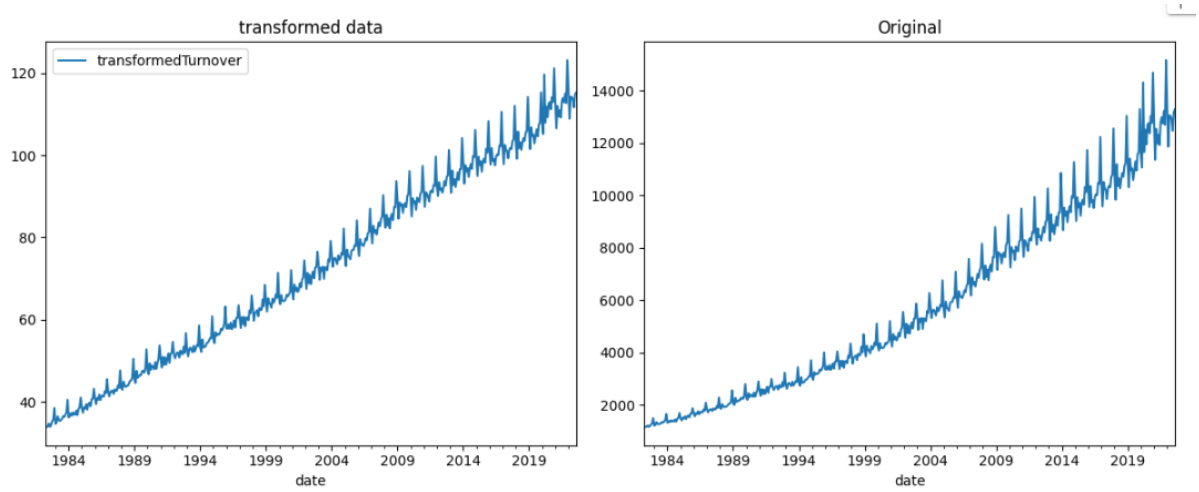


Figure 4: seasonality for original vs transformed data

The following graph shows the time series decomposition after applying normality technique.

The time series data represents the turnover of the food retail trade industry in Australia from April 1982 to August 2022. Preliminary analysis reveals a positive trend and seasonality in the data. The trend indicates a gradual increase in turnover over time, while the seasonal component shows repetitive patterns, suggesting potential periodic influences.

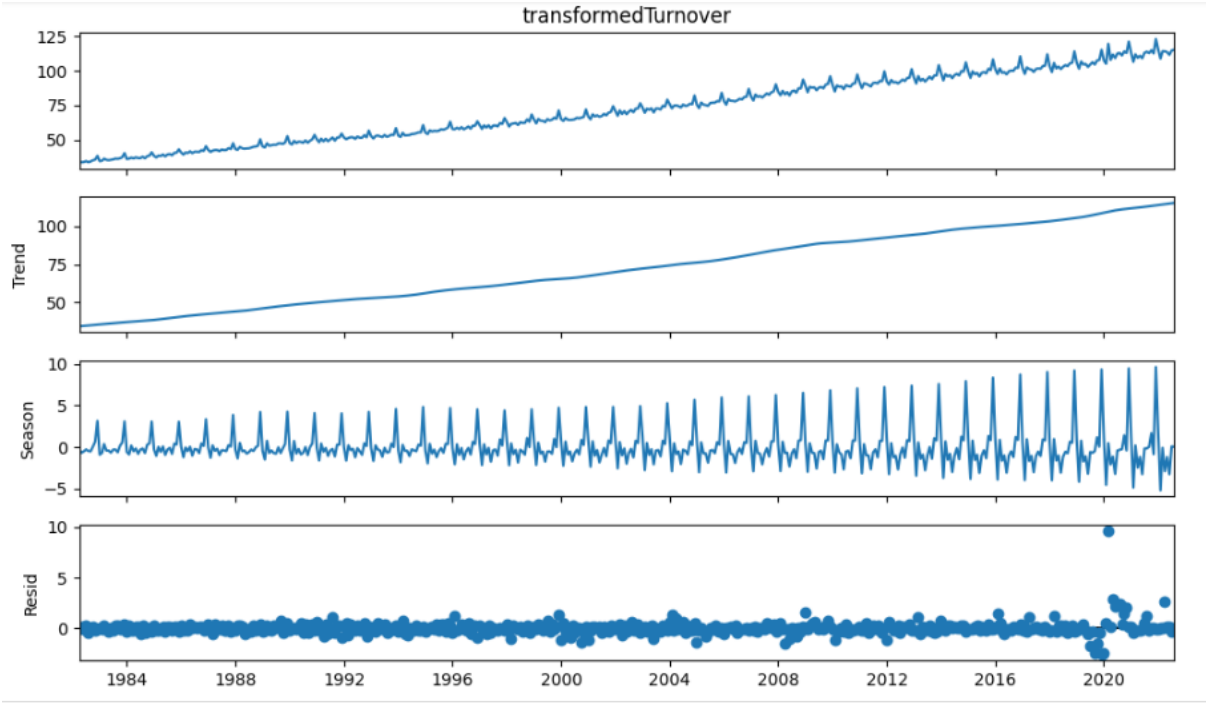


Figure 5: decomposition of transformed data(after applying normality)

The plots above indicate that the data exhibits non-stationary behaviour. To assess the stationarity, statistical and visualization techniques were employed on both the original turnover and transformed turnover (normalized turnover). The results reveal that neither dataset is stationary. Auto-correlation and partial correlation were also examined.

For the original turnover and normalized turnover, the Augmented Dickey-Fuller (ADF) statistic yielded values of 3.75 and 1.27, with corresponding p-values of 1.0 and 0.99, respectively. Given that the p-values exceed the significance level of 0.05, we cannot reject the null hypothesis, indicating non-stationarity in the time series. This suggests the presence of a trend in the data.

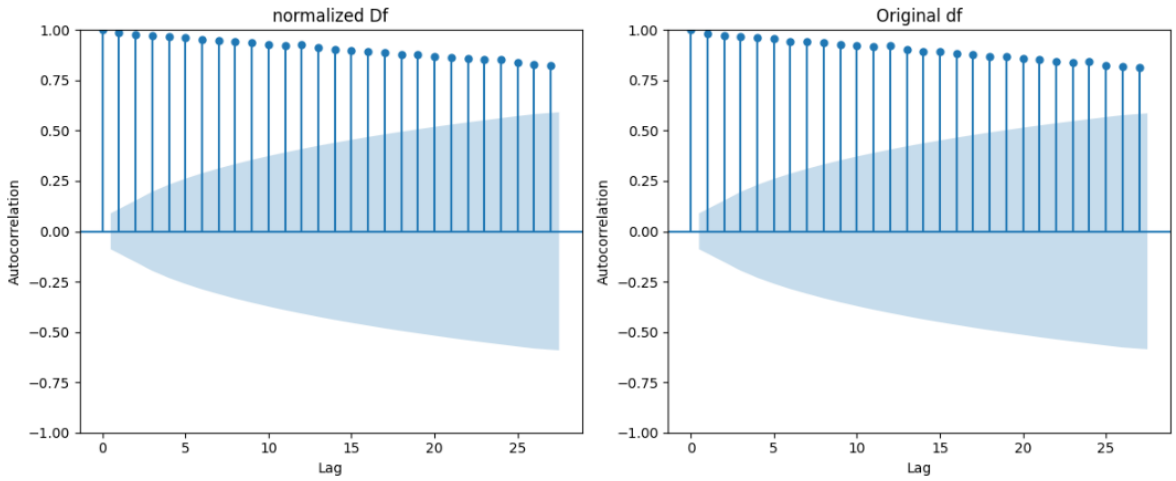


Figure 6: Autocorrelation

The autocorrelation coefficients slowly decrease as the lag increases (correlated) figure 6.

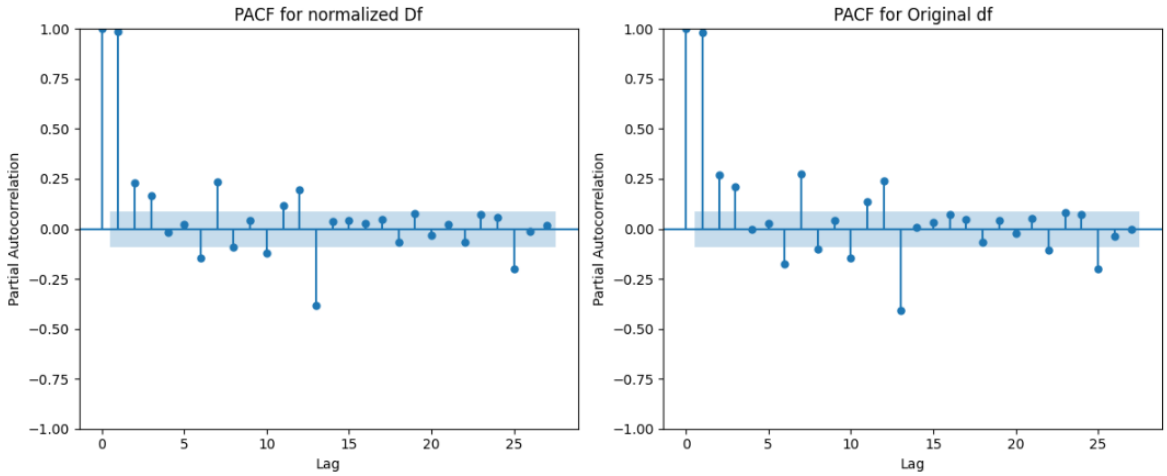
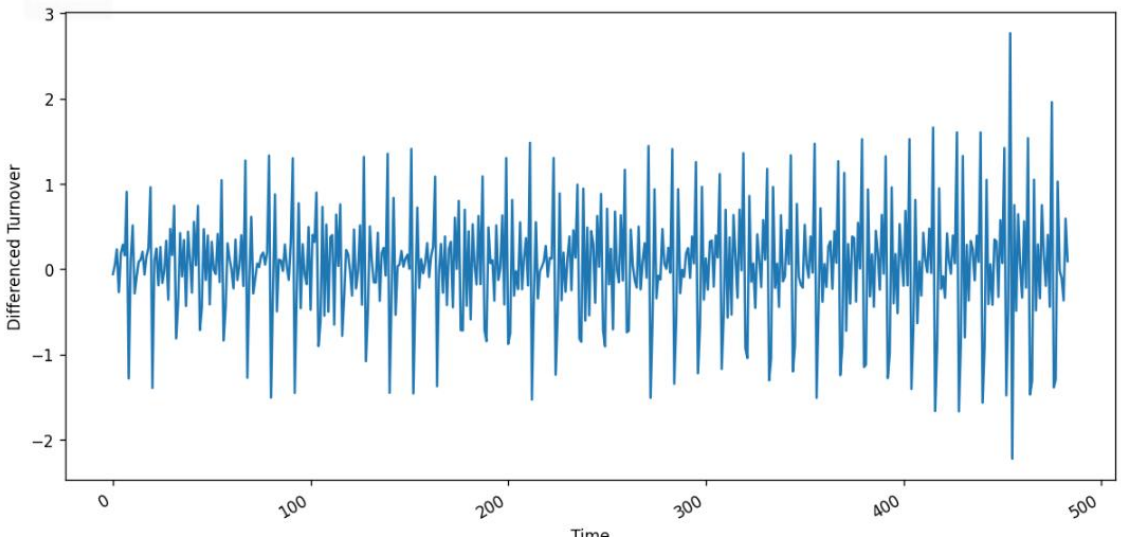


Figure 7: Partial Autocorrelation

As the above plots show the data is not stationary.

From this stage onward, our conclusions are based solely on the normalized data. To achieve stationarity, differencing was applied to the normalized data time series. The outcome indicates that the differenced data is now stationary, as evidenced by an ADF statistic of -6.11 and an extremely low p-value of 1.8e-08.



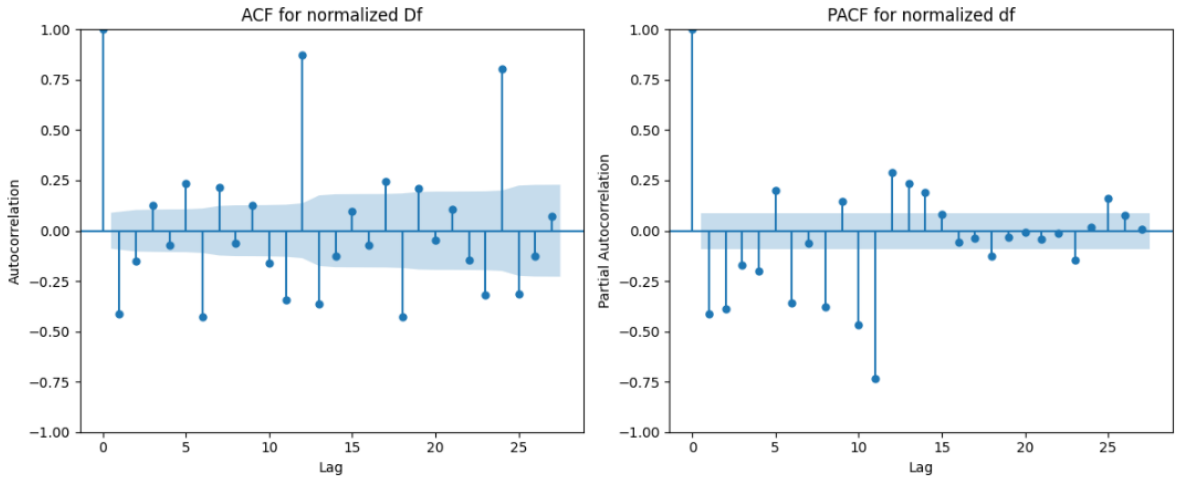
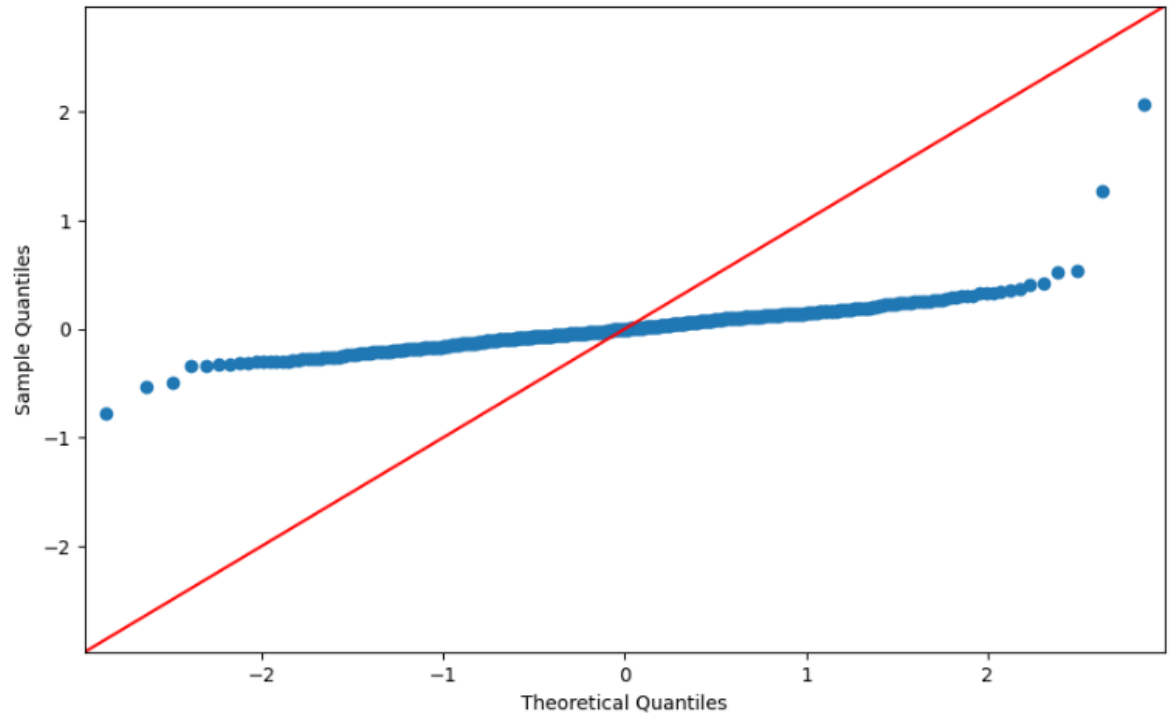
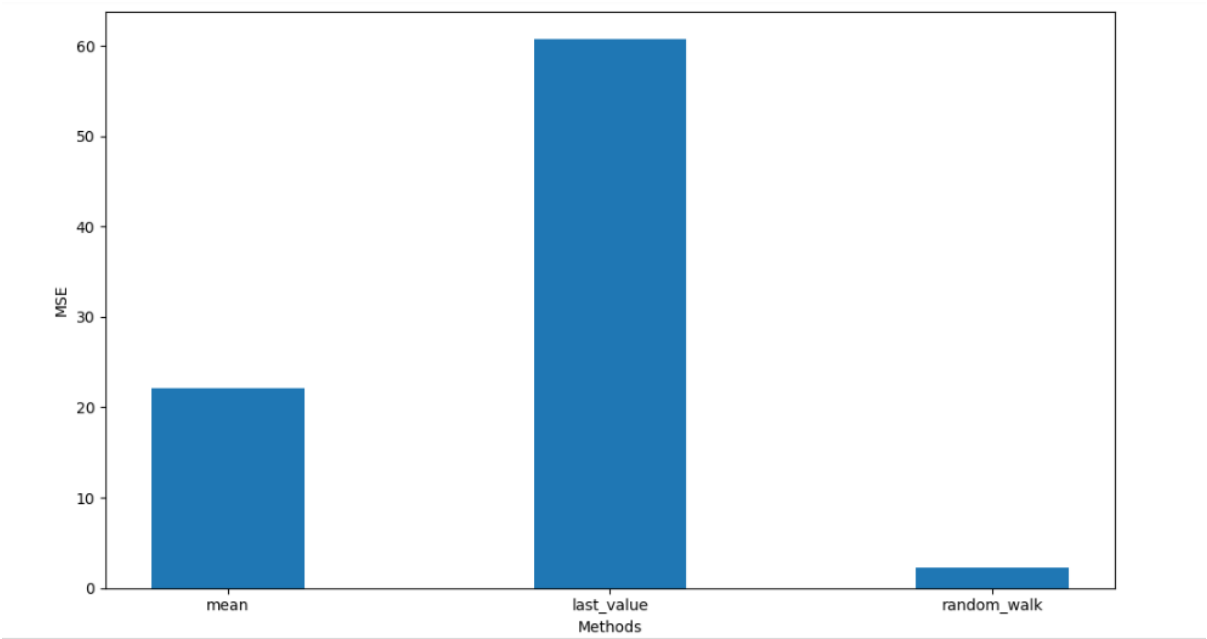


Figure 8: Stationary data



Let's compare the baseline methods. We use last known value and random walk models.

Mean Squared Error (Mean): 22.074976567515776
Mean Squared Error (Last Value): 60.736260798422535
Mean Squared Error (Random Walk): 2.2541553834128067



In summary, the random walk model outperforms the mean and last value models in terms of MSE, indicating its effectiveness in capturing patterns in the food retail trade turnover data.

Main Method SARIMA:

To determine appropriate parameters for SARIMA, the Auto ARIMA method is employed to attain the optimal fit. The outcome indicates that the SARIMA model with parameters (3, 0, 1) for the non-seasonal components and (2, 0, 1, 12) for the seasonal components, denoted as SARIMA(3, 0, 1)(2, 0, 1)[12], provides the best fit to the data.

SARIMAX Results

Dep. Variable:	y	No. Observations:	484
Model:	SARIMAX(3, 0, 1)x(2, 0, 1, 12)	Log Likelihood	118.538
Date:	Sun, 11 Feb 2024	AIC	-221.077
Time:	06:13:37	BIC	-187.620
Sample:	0	HQIC	-207.930
	- 484		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1154	0.069	1.684	0.092	-0.019	0.250
ar.L2	0.2513	0.054	4.625	0.000	0.145	0.358
ar.L3	0.4146	0.045	9.199	0.000	0.326	0.503
ma.L1	-0.9621	0.037	-25.817	0.000	-1.035	-0.889
ar.S.L12	1.0853	0.020	54.372	0.000	1.046	1.124
ar.S.L24	-0.0869	0.020	-4.364	0.000	-0.126	-0.048
ma.S.L12	-0.8255	0.025	-33.292	0.000	-0.874	-0.777
sigma2	0.0329	0.001	34.769	0.000	0.031	0.035

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	24160.72
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	2.53	Skew:	2.72
Prob(H) (two-sided):	0.00	Kurtosis:	37.18

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Figure 9 shows that SARIMA model has the better outcome compare to the random walk method.

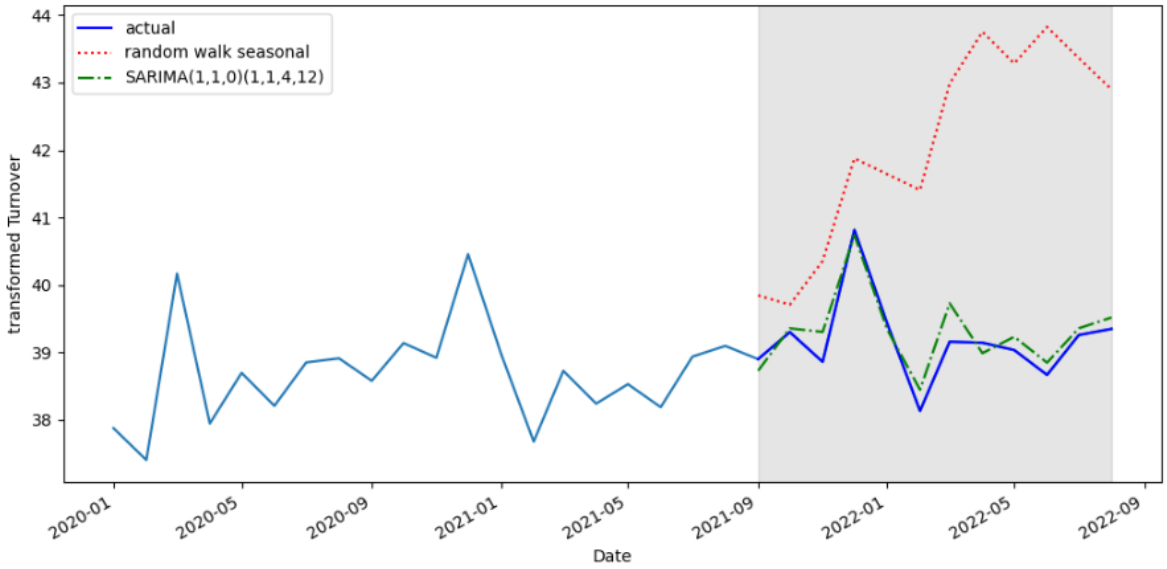


Figure 9: random walk Vs SARIMA

The SARIMA method demonstrates the lowest MSE, suggesting that it provides the most accurate predictions among the evaluated forecasting techniques (Figure 10).

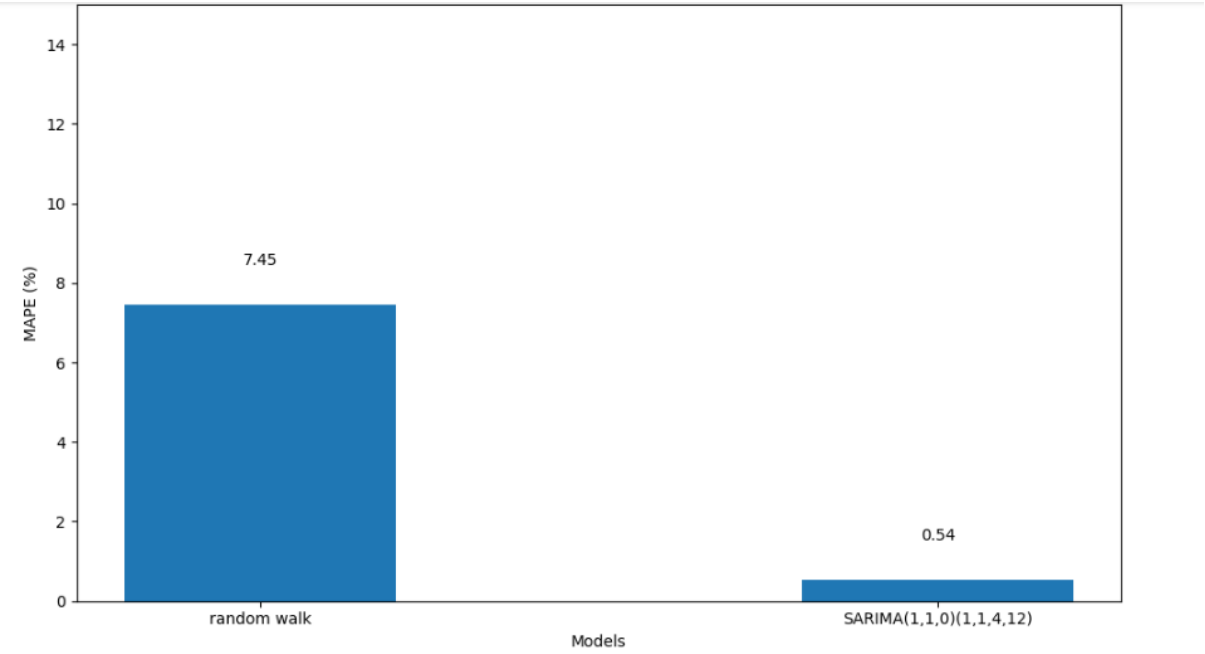
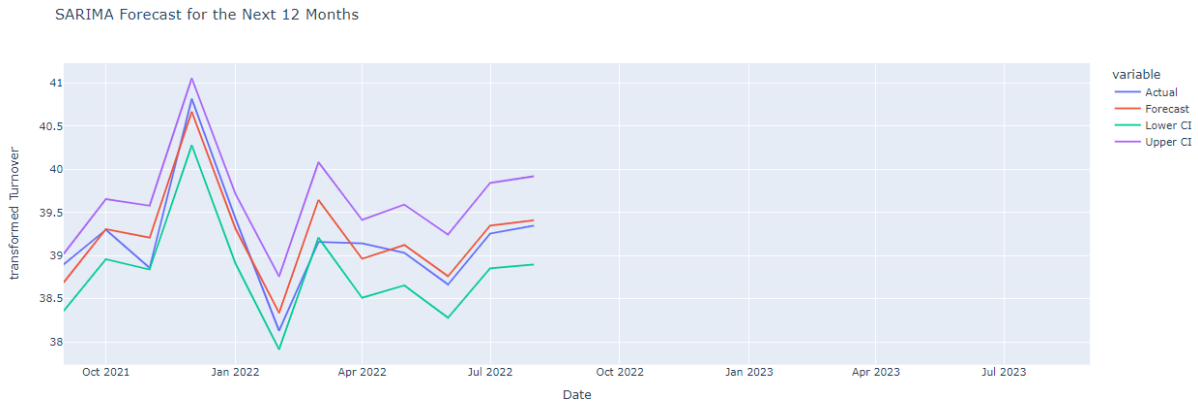


Figure 10: MAPE

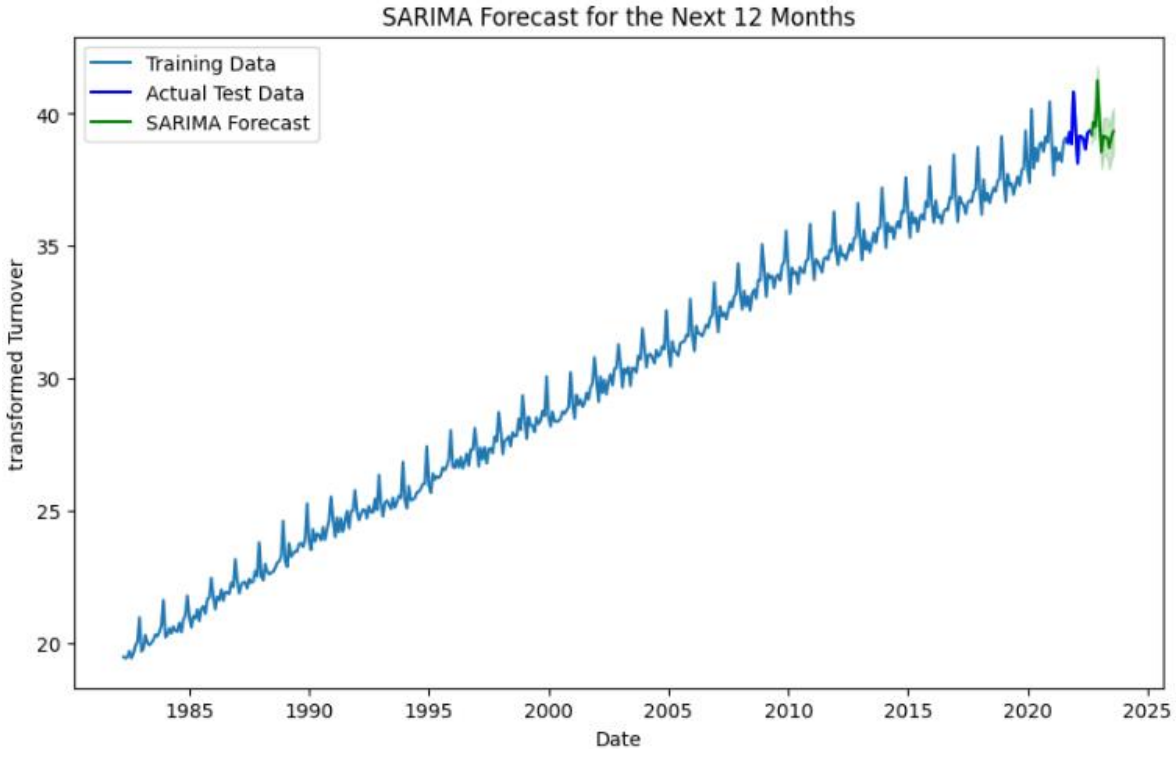
Predict the next 12 month without inversing:



	Date	Forecast	Lower CI	Upper CI
0	2021-09-01	38.689728	38.359054	39.020402
1	2021-10-01	39.306200	38.958066	39.654334
2	2021-11-01	39.208357	38.838897	39.577816
3	2021-12-01	40.665844	40.278280	41.053407
4	2022-01-01	39.311170	38.908329	39.714012
5	2022-02-01	38.334159	37.912443	38.755875
6	2022-03-01	39.643028	39.205954	40.080103
7	2022-04-01	38.962355	38.511308	39.413402
8	2022-05-01	39.122404	38.654293	39.590515
9	2022-06-01	38.760906	38.279398	39.242413
10	2022-07-01	39.348013	38.853361	39.842665
11	2022-08-01	39.408329	38.898095	39.918563

Second approach:

In the second approach, the Box-Cox transformation technique is applied to the data, followed by inversing the transformed data back to its original format. The outcome of this approach still improves SARIMA method has the best accuracy compare to the baseline models (MSE SARIMA: 0.05869050164999492).



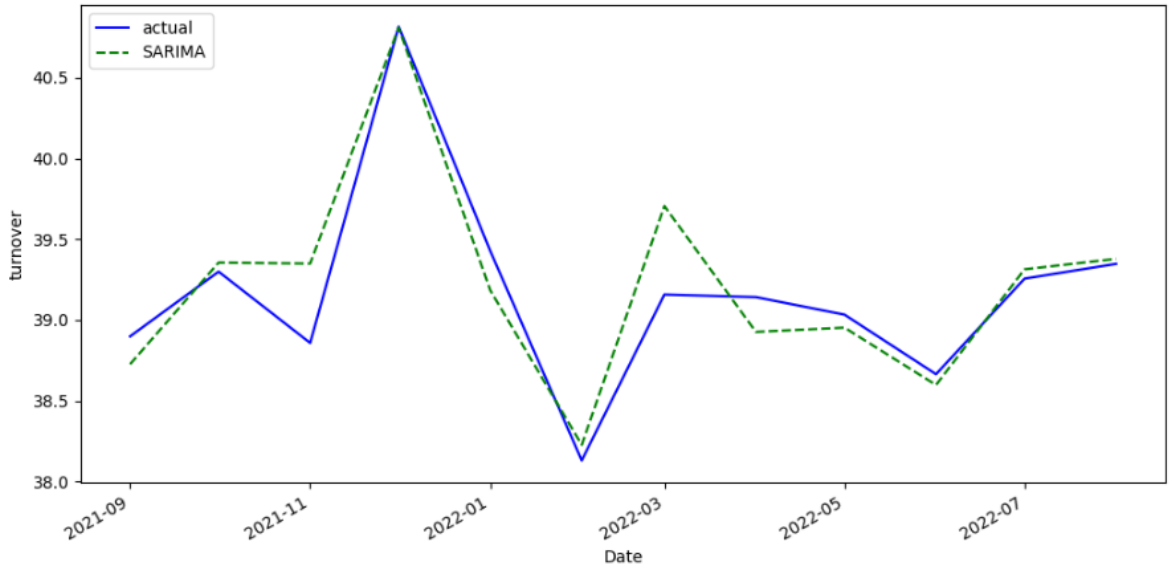


Figure 11: prediction using transformed data

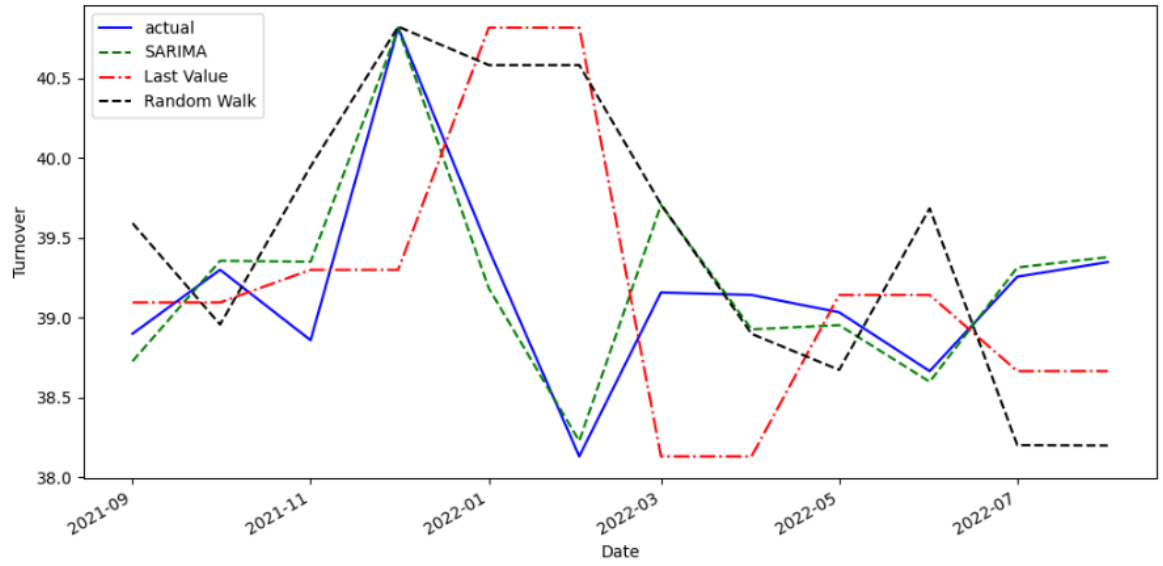


Figure 12: Comparison between three methods.

The SARIMA method demonstrates the lowest MSE, suggesting that it provides the most accurate predictions among the evaluated forecasting techniques (Figure 14).

Two baseline models, namely the Last Value and Random Walk, were employed for comparison. The Mean Squared Error (MSE) was used as a metric to evaluate their performance. The MSE for the Last Value method was found to be 1.24, for the Random Walk method was 1.09, and for SARIMA was 0.06. The lower MSE for the SARIMA method suggests this is the best model between other models.

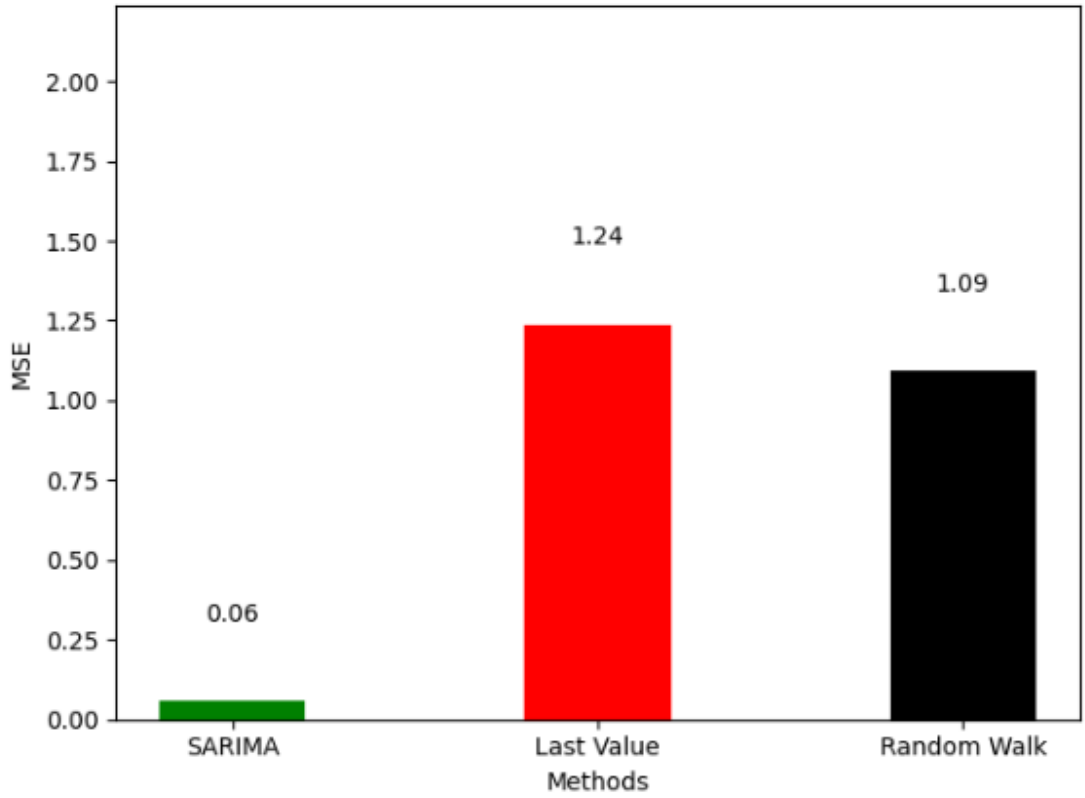


Figure 13: Average Square difference

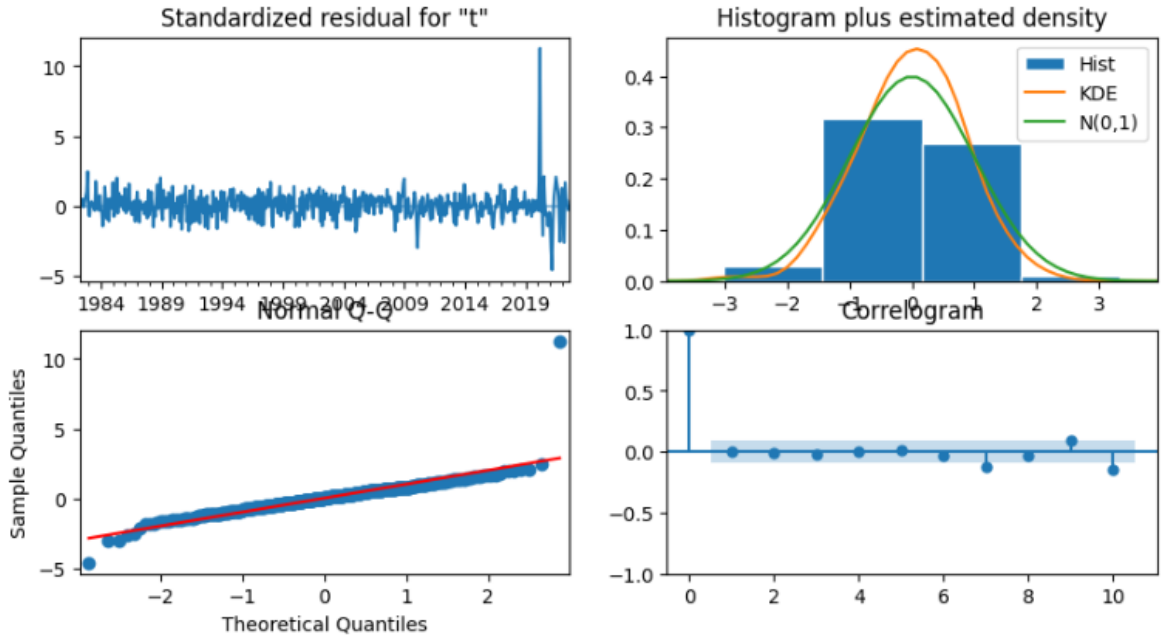


Figure 14: After transformation

	Date	Forecast	Lower CI	Upper CI
0	2022-09-01	39.163619	38.803615	39.523624
1	2022-10-01	39.653907	39.289376	40.018437
2	2022-11-01	39.506043	39.122832	39.889254
3	2022-12-01	41.091531	40.658082	41.524979
4	2023-01-01	39.683455	39.234284	40.132626
5	2023-02-01	38.705356	38.234363	39.176350
6	2023-03-01	39.898423	39.400840	40.396007
7	2023-04-01	39.356029	38.839868	39.872189
8	2023-05-01	39.505633	38.969335	40.041931
9	2023-06-01	39.130324	38.573596	39.687053
10	2023-07-01	39.649497	39.074654	40.224340
11	2023-08-01	39.793474	39.200408	40.386540

Figure 15: prediction inversed data

In summary, The SARIMA forecasting model was developed using Auto ARIMA to determine optimal parameters, resulting in SARIMA (3, 0, 1)(2, 0, 1)[12]. The model was trained on the normalized and differenced time series data. The forecast was then made for the next 12 months (September 2022 – August 2023). The results indicate that SARIMA is effective in capturing the intricate patterns within the data, providing a reliable basis for future predictions.

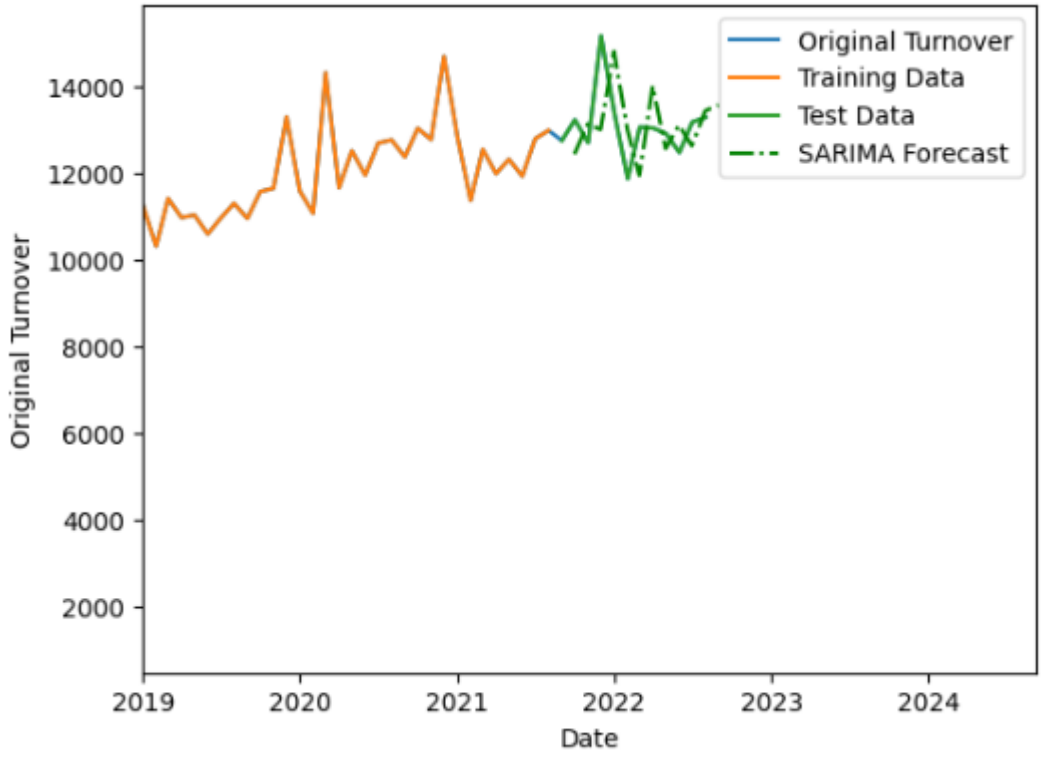


Figure 16: prediction real values

	Date	Forecast	Lower CI	Upper CI
0	2021-09-01	12430.427389	11948.895653	12926.336942
1	2021-10-01	13123.599861	12618.480551	13643.701546
2	2021-11-01	12999.208838	12459.162466	13556.585968
3	2021-12-01	14802.731583	14183.614526	15441.855102
4	2022-01-01	13003.110169	12416.920657	13609.768508
5	2022-02-01	11940.575054	11370.071112	12532.236663
6	2022-03-01	14007.558654	13341.306792	14698.403853
7	2022-04-01	12556.336732	11922.658881	13214.889423
8	2022-05-01	13100.076785	12425.810556	13801.358675
9	2022-06-01	12624.475724	11949.958677	13327.090064
10	2022-07-01	13429.508293	12703.802707	14185.802493
11	2022-08-01	13565.002390	12815.223669	14347.137913

Figure 17: prediction real values

Scope, Limitation, and Application of the Model:

The SARIMA forecasting model's scope may be limited in capturing abrupt, unforeseen events, such as economic downturns or pandemics, which can significantly impact turnover. The model's effectiveness relies on the assumption that future patterns mirror historical data.

Transparency and Accountability:

The inclusion of model validation metrics, such as Mean Squared Error (MSE), enhances accountability by quantifying prediction accuracy. To further address transparency, comprehensive documentation of data preprocessing, model selection criteria, and performance evaluation should be provided.

Conclusion and Recommendation:

In conclusion, the SARIMA model effectively captures the time series components, providing a robust forecasting tool for food retail trade turnover in Australia. Based on comparison with baseline models, SARIMA outperforms in terms of accuracy, showcasing its suitability for predicting future turnover trends. Recommendations include continuous monitoring, periodic model reassessment, and integrating external factors to enhance forecasting precision.