Industrial Engineering Department
IE 360- Statistical Forecasting and Time Series
Spring 2021

Project Report

Hatice Pınar YILDIRIM- 2016402015
Ahmet Atakan DEMİRKAN- 2016402144

Table of Contents

1. Introduction

Everyday, millions of people purchase many different items via online shopping websites. As Internet evolves throughout years, online shopping became one of the most important sectors around world. Trendyol.com is the largest mobile commerce company in Turkey and the MENA region selling more than 70 million items per year. Considering the wide product spectrum and strong sales of Trendyol, the correct decisions should be made with great precision in order to continue the operation successfully. Deciding how many products to sell in the future is one of the most important of these decisions. Forecasting the sales quantity of a product for an online retail is a challenging task. Forecasting heavily relies on the historical data and modeling skills. In this project, we were expected to forecast the sales of eight different products in Trendyol. Groups were free to choose their own approach to predict the sales. All approaches are explained in Approach part.

The data, covers daily sales of nine products (Table.1) of Trendyol and includes product, category level, brand and site level details. There were 4,2,3 products in textile, electronics and cosmetics categories correspondingly.

Product Id	Top Hier.	Bottom Hier.	Brand
48740784	Dış Giyim	Mont	ALTINYILDIZ CLASSICS
73318567	Plaj Giyim	Bikini Üstü	TRENDYOLMİLLA
32737302	Plaj Giyim	Bikini Üstü	TRENDYOLMİLLA
31515569	Alt Giyim	Tayt	TRENDYOLMİLLA
6676673	Elektronik Aksesuarlar	Bluetooth Kulaklık	Xiaomi
7061886	Elektrikli Ev Aletleri	Dik Süpürge	Fakir
85004	Cilt Bakım	Yüz Temizleyici	La Roche Posay
4066298	Anne ve Bebek Bakım	Bebek Islak Mendil	Sleepy
32939029	Ağız Bakım	Şarj Edebilir Diş Fırçası	Oral-B

Table 1: Product characteristics

1. Related Literature

The main resource that is used throughout the project was DataCamp and lecture notes of Prof Baydoğan. The assignments that were conducted as a lecture enhancement was quite helpful in finding the best model that fits the data. Also, “Forecasting in R” course of Hyndman is a very helpful resource about plotting the data, training and test data concepts, exponential smoothing, ARIMA models and other advanced methods. Lastly, “Introduction to Time Series analysis and Forecasting” book helped us to gain an in-depth theoretical knowledge.

3. Approach

The first step while approaching the problem was analyzing the data. Time series are plotted as seen in Figure 1 to observe them visually.

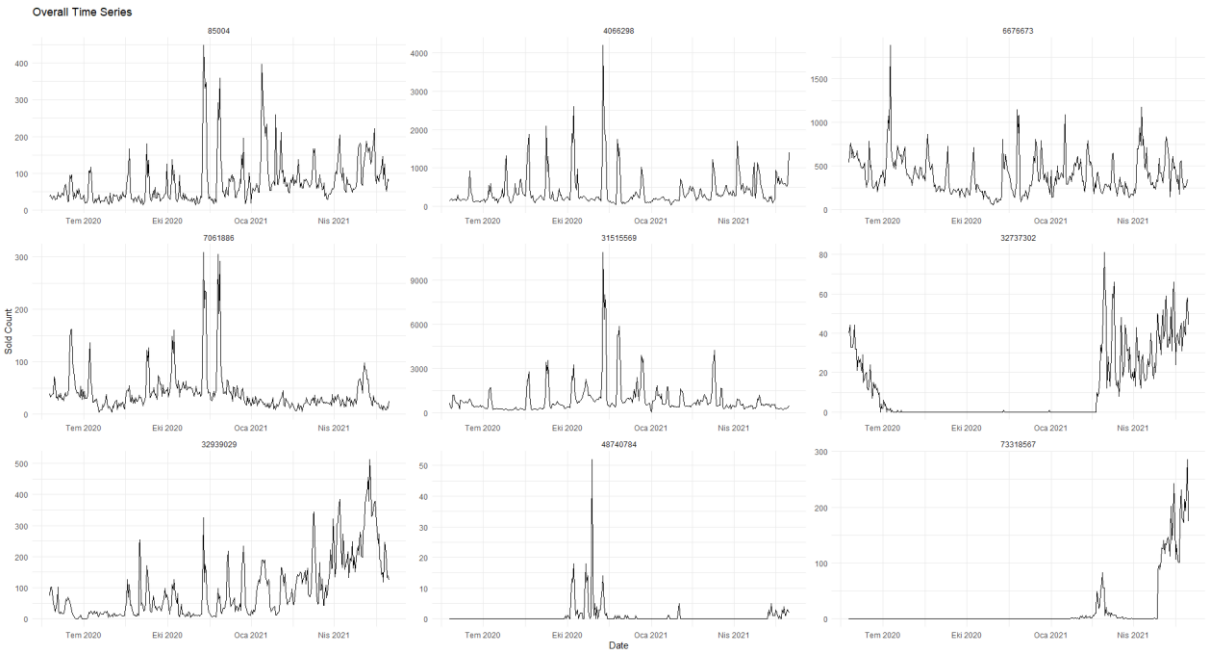

Figure 1. Time Series Plots

To have a better understanding time series are plotted for 3-month intervals in Figure 2, Figure 3, Figure 4 and Figure 5.

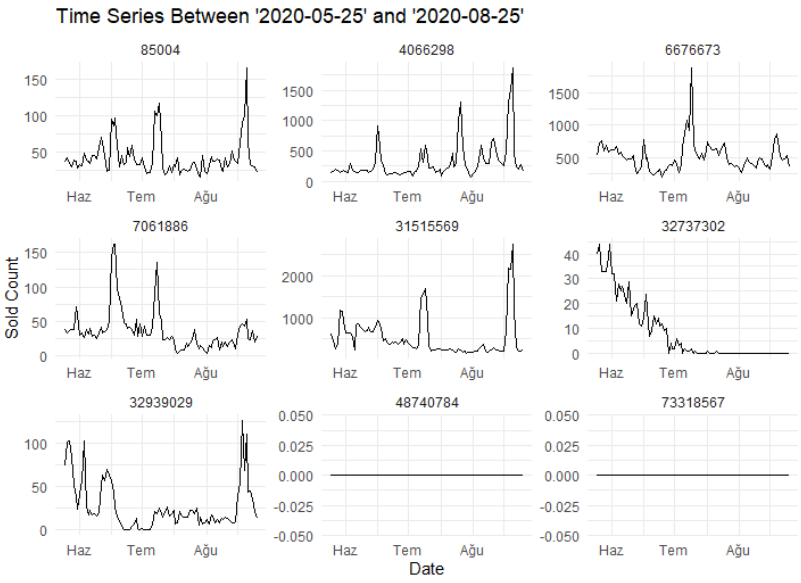

Figure 2. Data between '2020-05-25' and '2020-08-25'

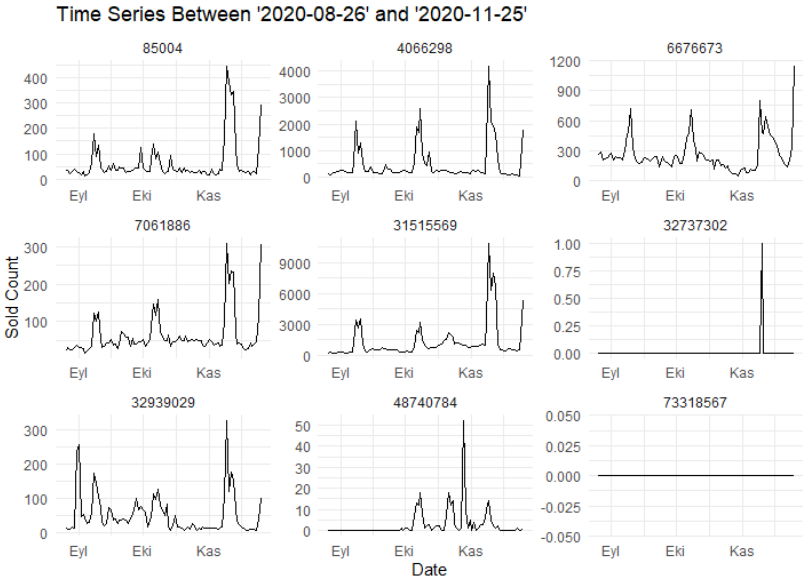

Figure 3. Data between '2020-08-26' and '2020-11-25'

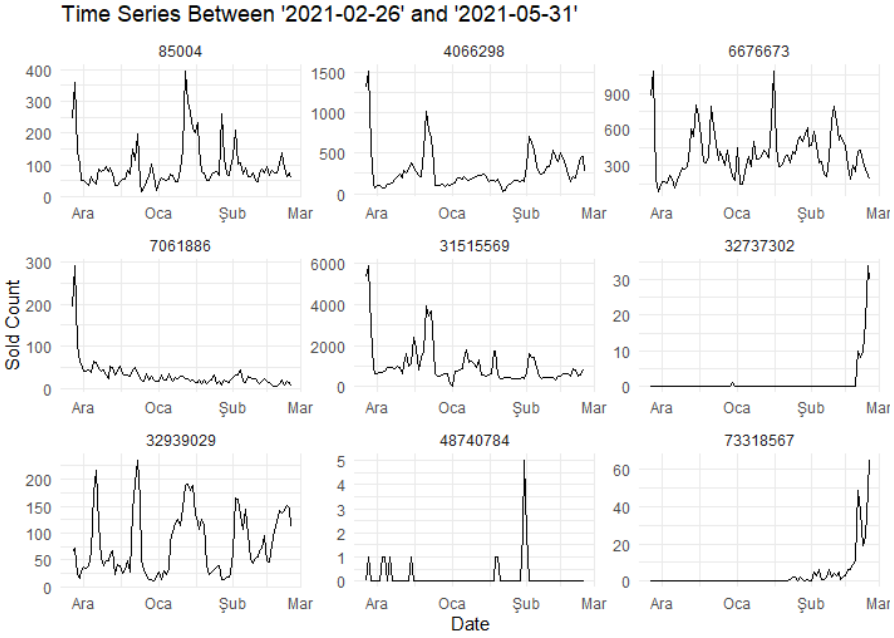

Figure 4. Data between '2021-02-26' and '2021-05-31'

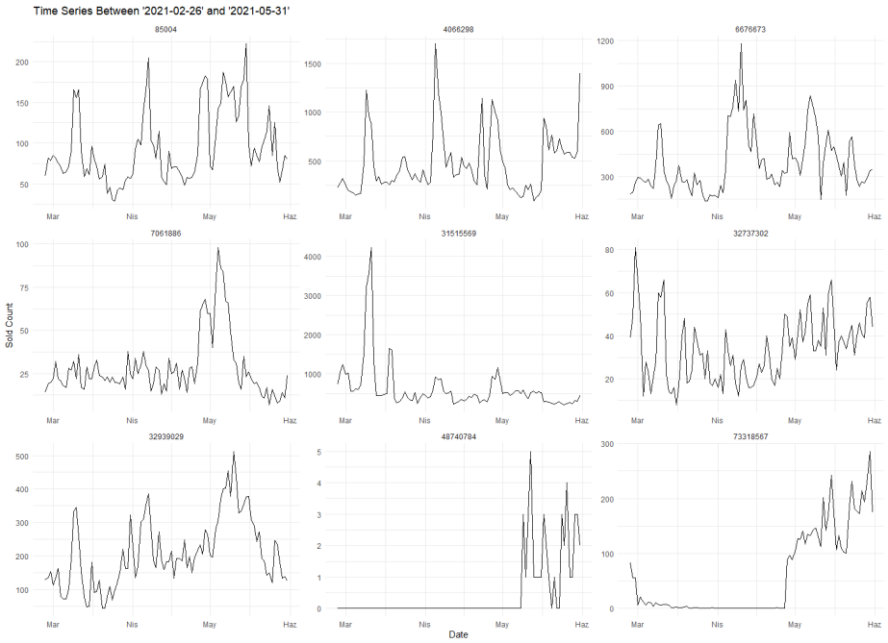

Figure 5. Data between '2021-02-26' and '2021-05-31'

After understanding the characteristics of the given data set the second step is cleaning and manipulating the data to perform a healthy analysis.

3.1. Analysis of Products

Data consists of observations for a total of 9 distinct products. Given data set is read as a data table and divided into 9 separate data tables, one for each product and each product is analyzed separately

.

3.1.1. Product 1

Sold count of Product 1 is plotted as in Figure 6. It is observed that there are several but short periods that have high sales.

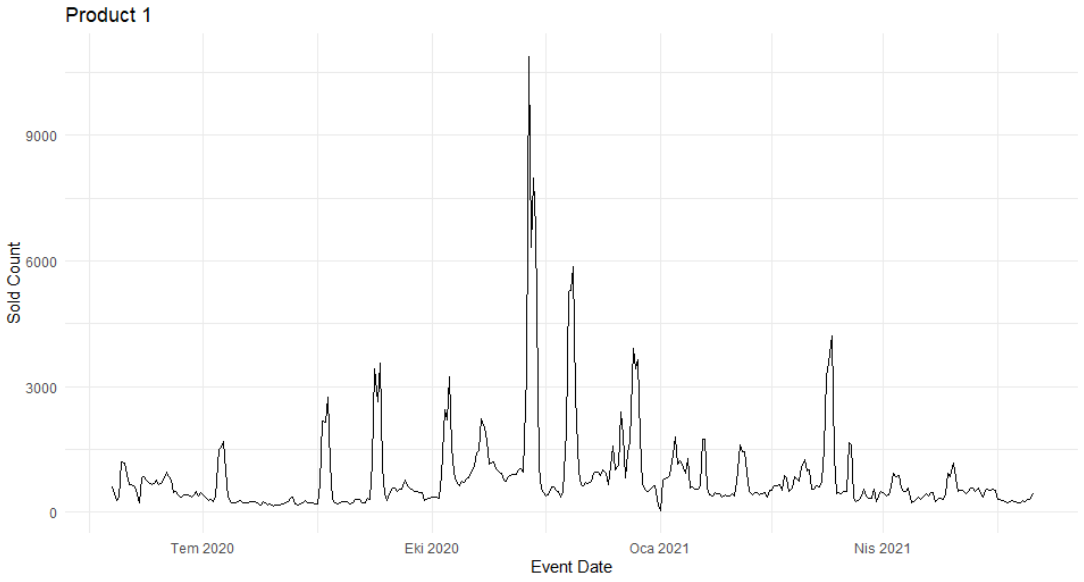

Figure 6. Sold Count of Product 1

Histogram and Normal Q-Q plot of Product 1 are given in Figure 7 and Figure 8 relatively. It is observed that data is not normally distributed.

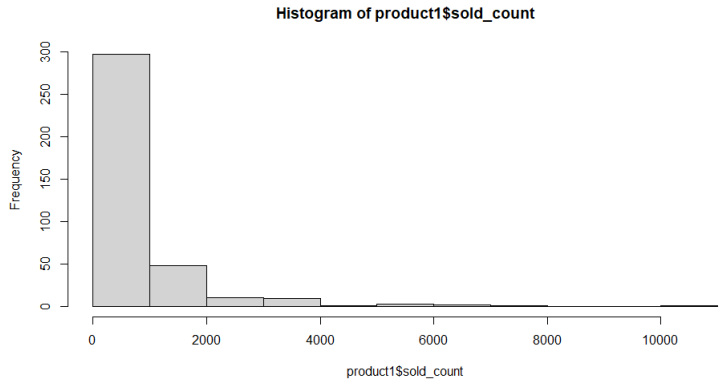

Figure 7. Histogram of Sold Count of Product 1

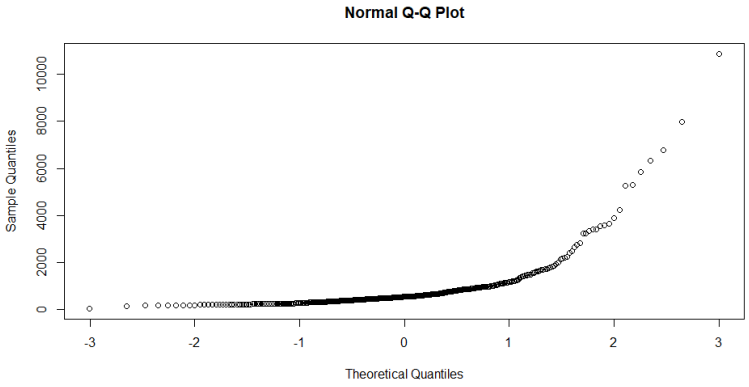

Figure 8. Normal Q-Q plot of Sold Count of Product 1

3.1.2. Product 2

Sold count of Product 2 is plotted as in Figure 9. It is observed that sales are small in winter months and high in the summer. Sales are highest in the beginning of summer.

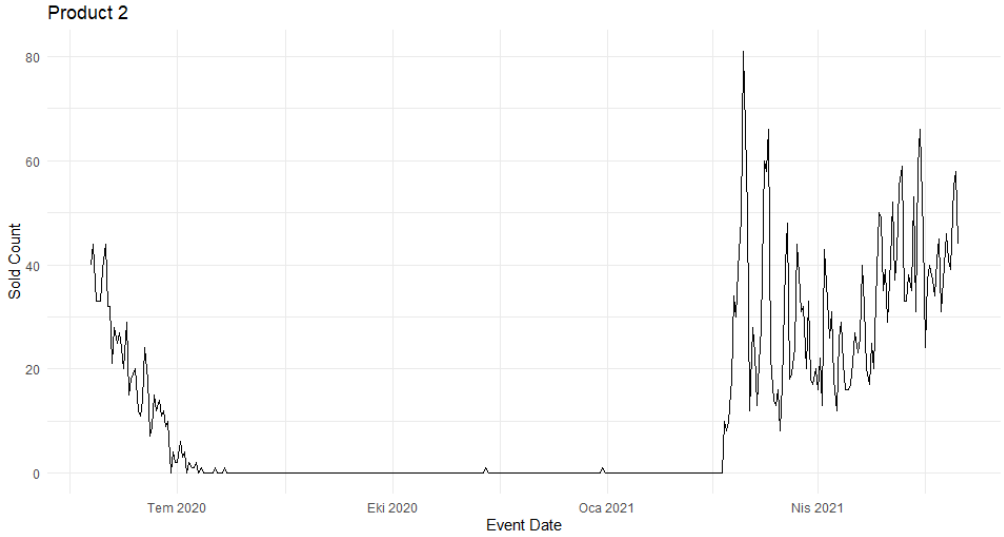

Figure 9. Sold Count of Product 2

Histogram and Normal Q-Q plot of Product 2 are given in Figure 10 and Figure 11 relatively. It is observed that data is not normally distributed.

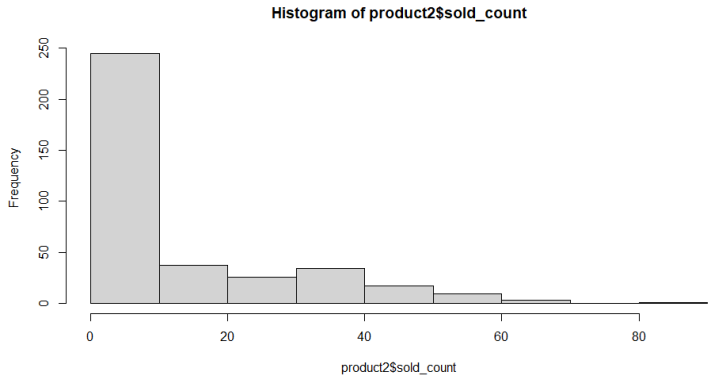

Figure 10. Histogram of Sold Count of Product 2

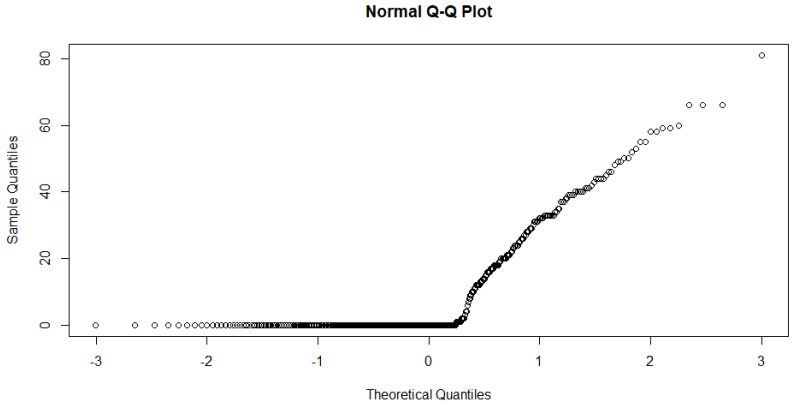

Figure 11. Normal Q-Q plot of Sold Count of Product 2

3.1.3. Product 3

Sold count of Product 3 is plotted as in Figure 12. It is observed that there is a mean shift starting from the spring of 2021. Variance of the series is not constant.

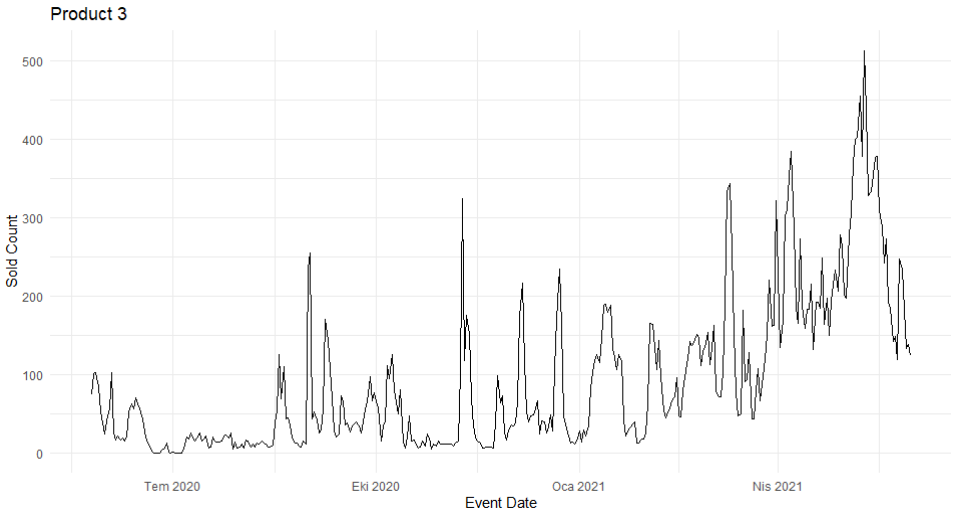

Figure 12. Sold Count of Product 3

Histogram and Normal Q-Q plot of Product 3 are given in Figure 13 and Figure 14 relatively. It is observed that data is not normally distributed.

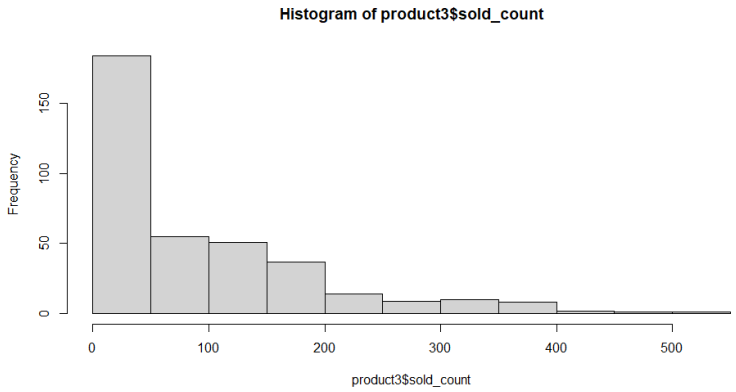

Figure 13. Histogram of Sold Count of Product 3

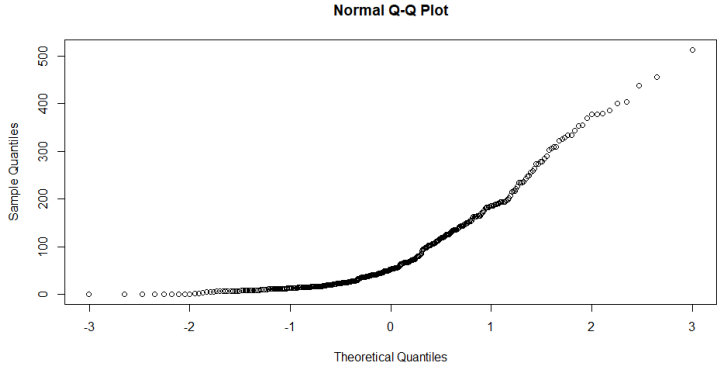

Figure 14. Normal Q-Q plot of Sold Count of Product 3

3.1.4. Product 4

Sold count of Product 4 is plotted as in Figure 15. Seasonality is observed in 2020, however there is not any seasonality in 2021.Variance of the series is not constant. There are several short periods with high sales.

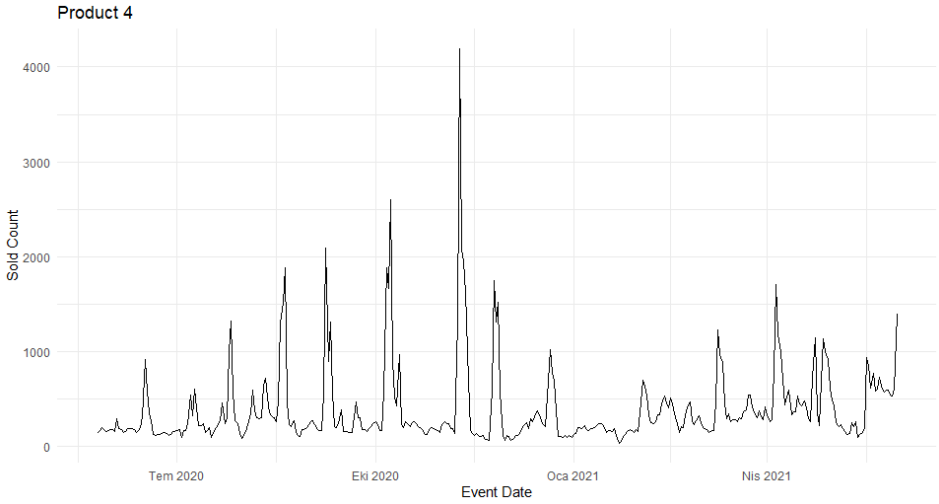

Figure 15. Sold Count of Product 4

Histogram and Normal Q-Q plot of Product 4 are given in Figure 16 and Figure 17 relatively. It is observed that data is not normally distributed.

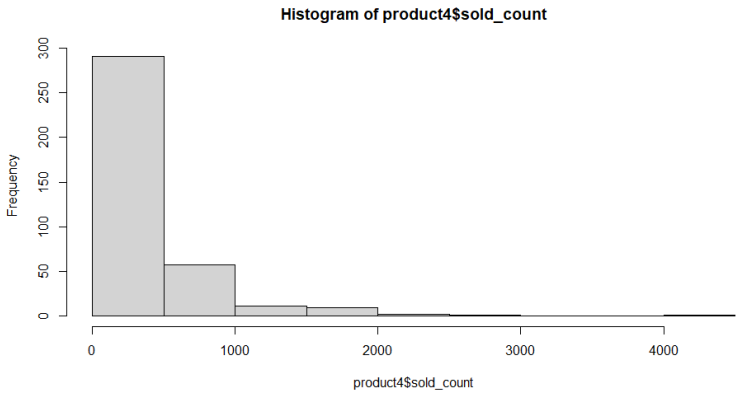

Figure 16. Histogram of Sold Count of Product 4

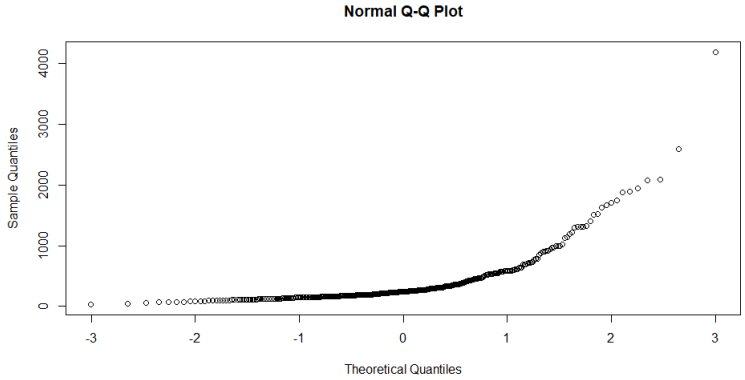

Figure 17. Normal Q-Q plot of Sold Count of Product 4

3.1.5. Product 5

Sold count of Product 5 is plotted as in Figure 18. This product is mostly sold in the winter of 2020. There are a few periods with small sold counts. Recently there is a small increase in the sales possibly due to summer discounts.

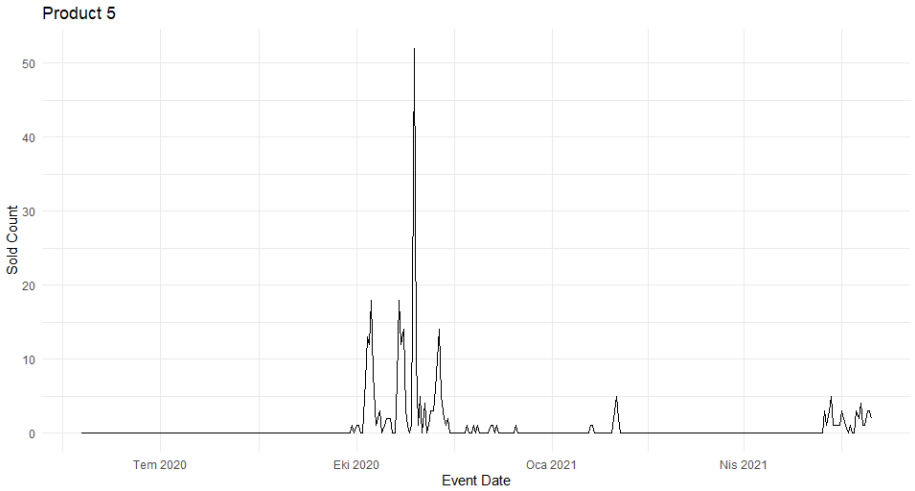

Figure 18. Sold Count of Product 5

Histogram and Normal Q-Q plot of Product 5 are given in Figure 19 and Figure 10 relatively. It is observed that data is not normally distributed.

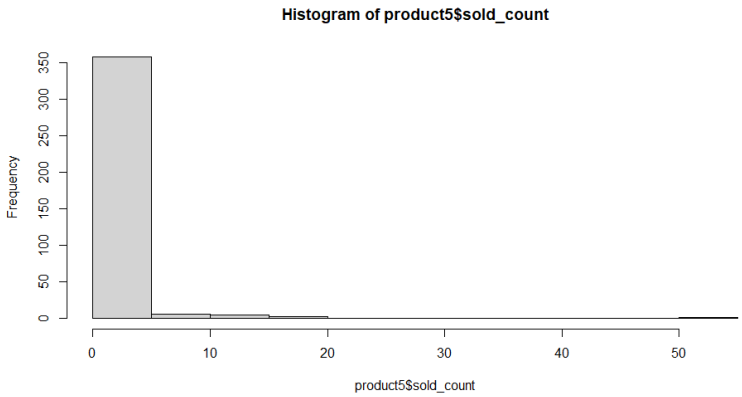

Figure 19. Histogram of Sold Count of Product 5

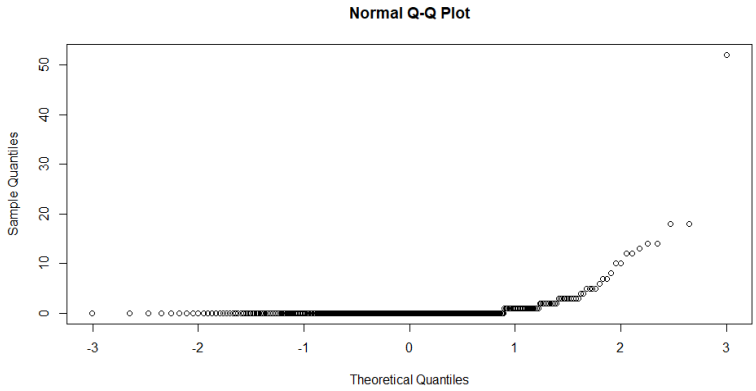

Figure 20. Normal Q-Q plot of Sold Count of Product 5

3.1.6. Product 6

Sold count of Product 6 is plotted as in Figure 21. There are several mean-shifts and the variance of the series is not constant.

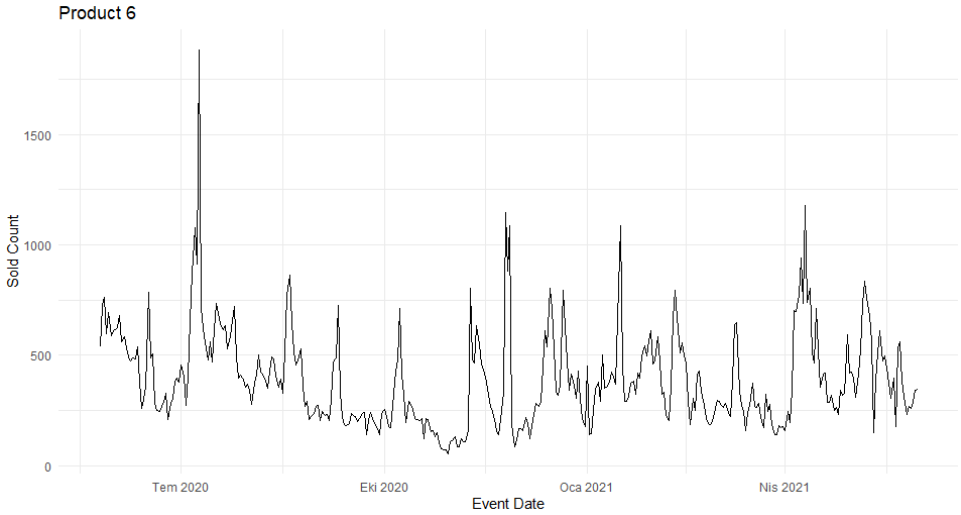

Figure 21. Sold Count of Product 6

Histogram and Normal Q-Q plot of Product 6 are given in Figure 22 and Figure 23 relatively. Although distribution of this product is more similar to a normal distribution than the other products. It is not possible to say that data is normally distributed.

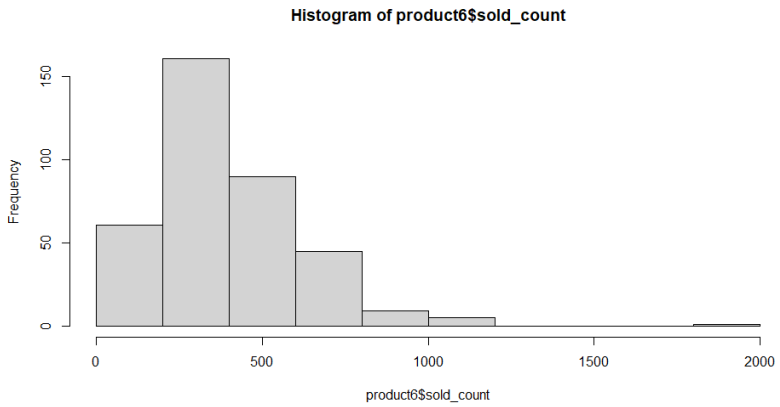

Figure 22. Histogram of Sold Count of Product 6

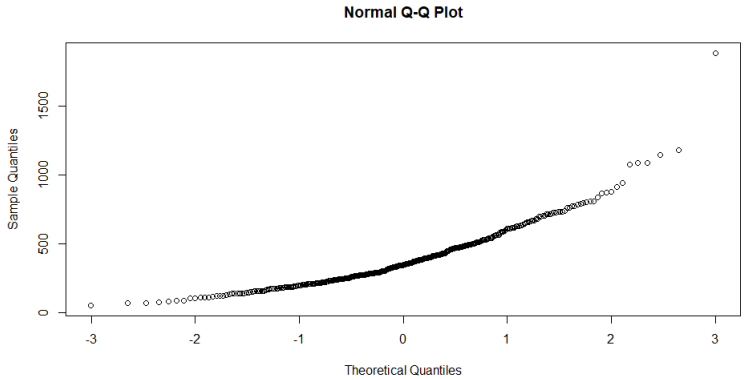

Figure 23. Normal Q-Q plot of Sold Count of Product 6

3.1.7. Product 7

Sold count of Product 7 is plotted as in Figure 24. Mean of the data is not constant. There are several high-sale periods therefore variance is not constant.

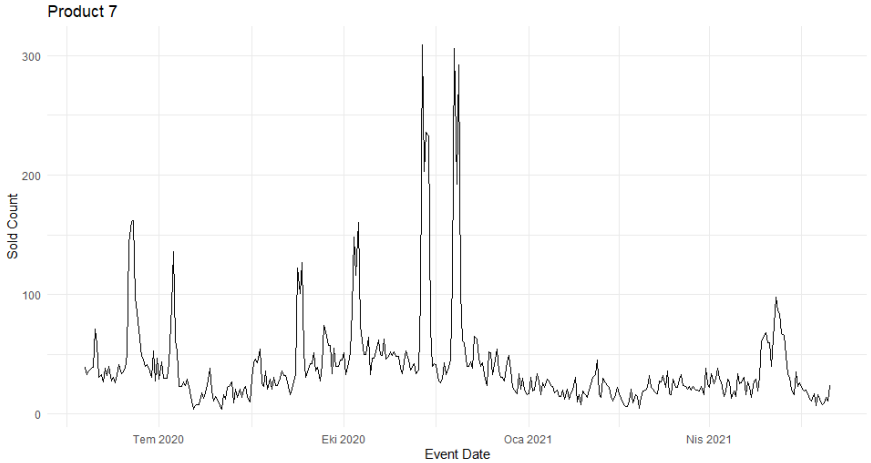

Figure 24. Sold Count of Product 7

Histogram and Normal Q-Q plot of Product 7 are given in Figure 25 and Figure 26 relatively. It is not possible to say that data is normally distributed.

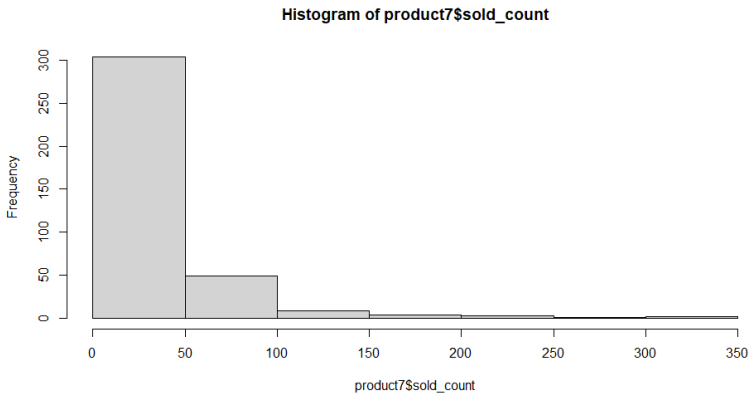

Figure 25. Histogram of Sold Count of Product 7

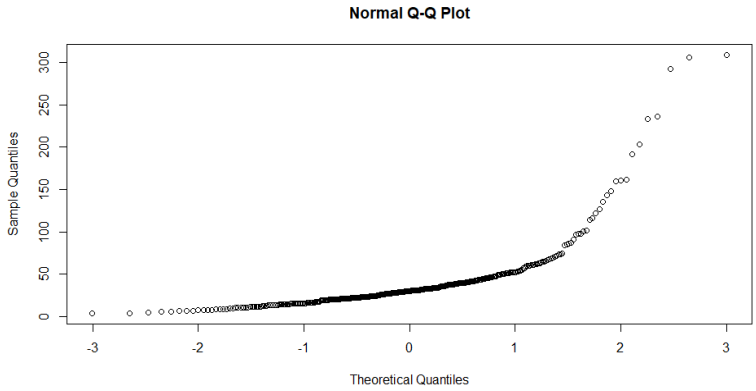

Figure 26. Normal Q-Q plot of Sold Count of Product 7

3.1.8. Product 8

Sold count of Product 8 is plotted as in Figure 27. There are not any sales before winter of 2021, this period is removed in modeling. During the winter sales counts are varying however there is an increasing trend in summer. Variance of the sales during the summer and winter are not constant.

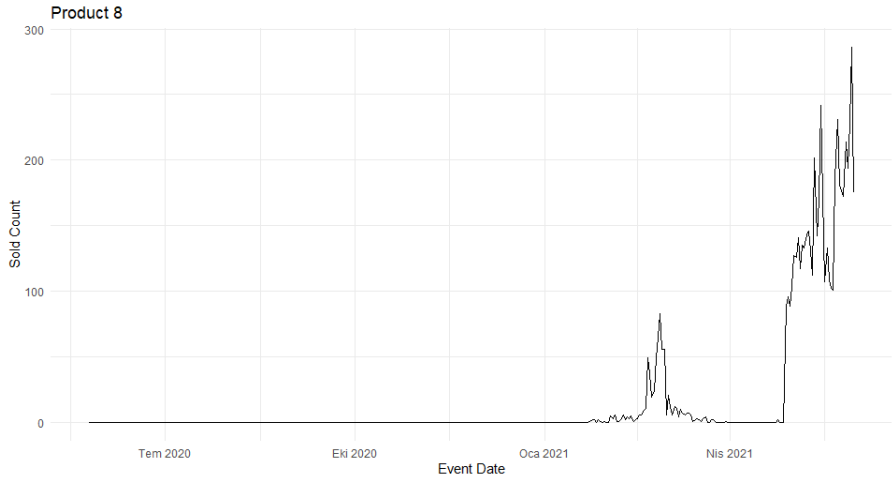

Figure 27. Sold Count of Product 8

Histogram and Normal Q-Q plot of Product 8 are given in Figure 28 and Figure 29 relatively. It is not possible to say that data is normally distributed.

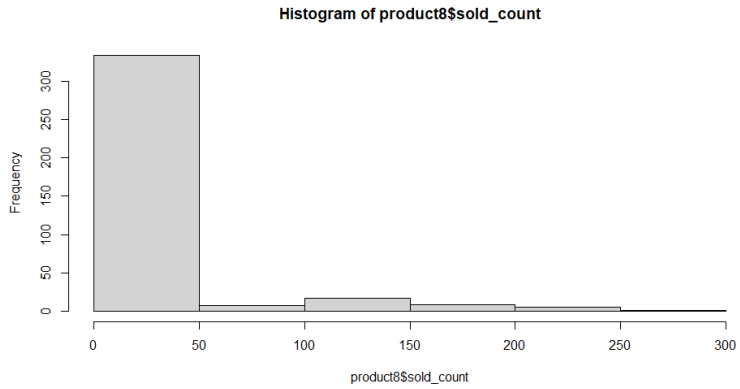

Figure 28. Histogram of Sold Count of Product 8

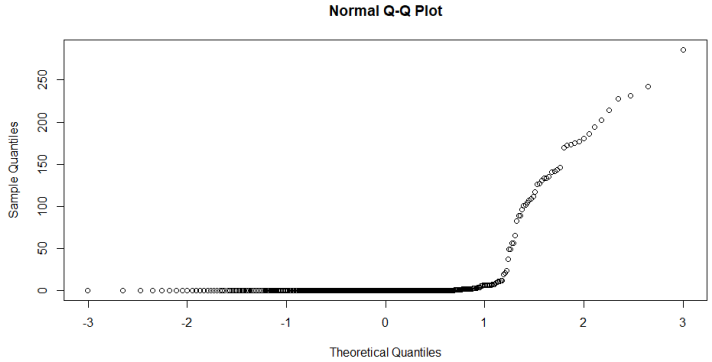

Figure 29. Normal Q-Q plot of Sold Count of Product 8

3.1.9. Product 9

Sold count of Product 9 is plotted as in Figure 30. Sales have different means in 2020 and 2021. Variance is not constant in both years.

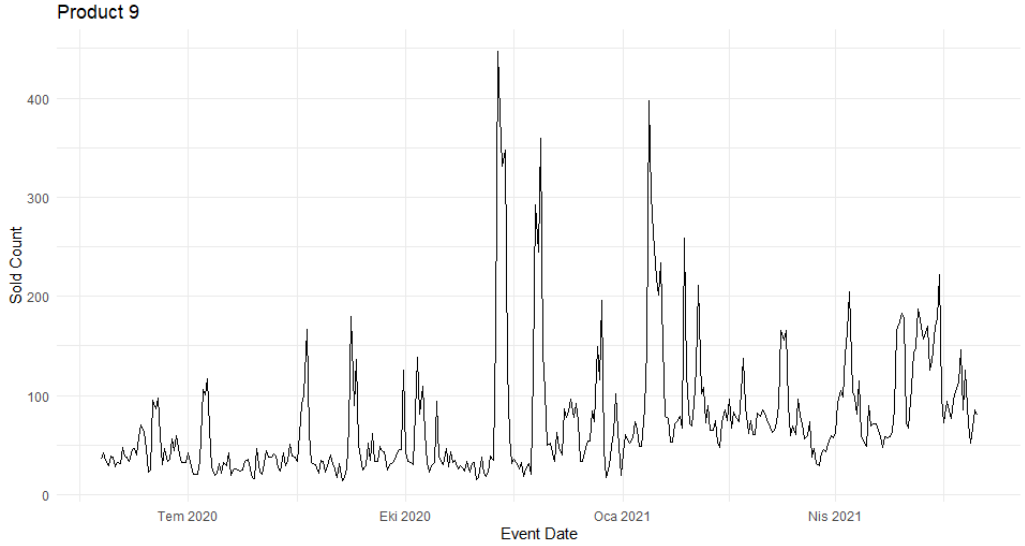

Figure 30. Sold Count of Product 9

Histogram and Normal Q-Q plot of Product 9 are given in Figure 31 and Figure 32 relatively. It is not possible to say that data is normally distributed.

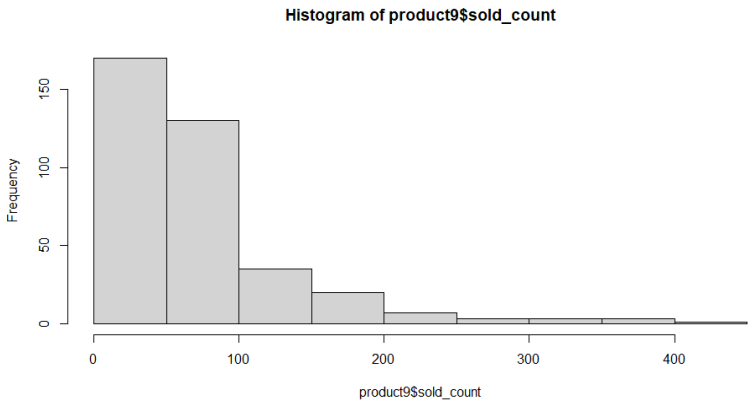

Figure 31. Histogram of Sold Count of Product 9

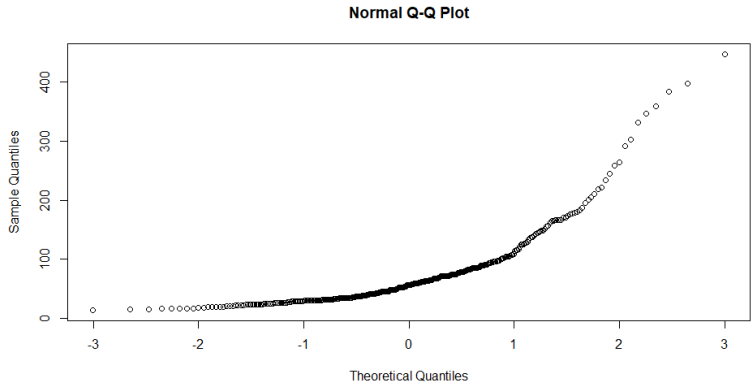

Figure 32. Normal Q-Q plot of Sold Count of Product 9

3.2 Time Series Analysis Methods

Time series related to each product are decomposed on monthly and weekly levels and with multiplicative and additive methods. All 4 decompositions are inspected and the one which has the most stationary random term is chosen for each product. ACF and PACF plots of random terms are checked. Then *auto.arima* function is used to model random terms. Results of *auto.arima* function were satisfactory for most of the products. However, for Product 2, Product 5 and Product 8 results were not satisfactory and external regressors are used. For Product 2 correlation plot is given in Figure 33.

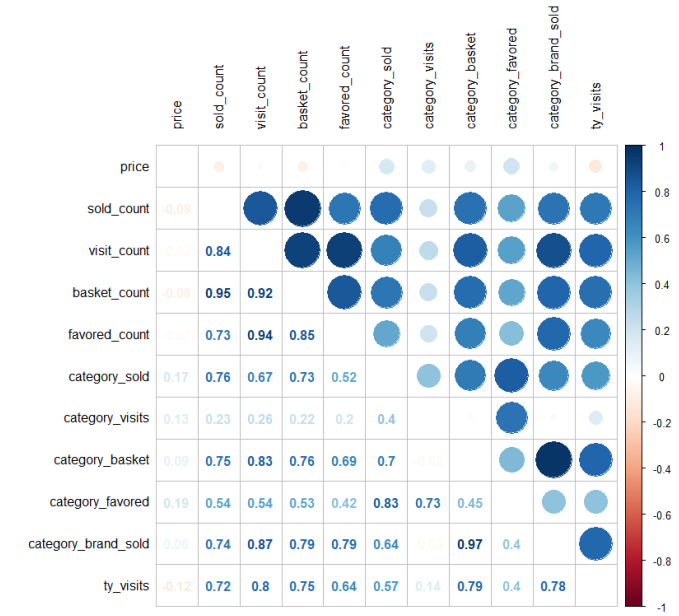

Figure 33. Correlation plot for Product 2

In addition to CCF functions are also checked and basket_count, visit_count, category_sold, category_favored, category_basket, category_brand and ty_visits at lag 1, as well as the fitted values of ARIMA function are used as inputs to a linear regression model for sold count. After removing the non-significant features final inputs to the linear model are fitted values, visit_count, category_sold, category_favored and category_basket.

For Product 5 correlation plot is given in Figure 34.

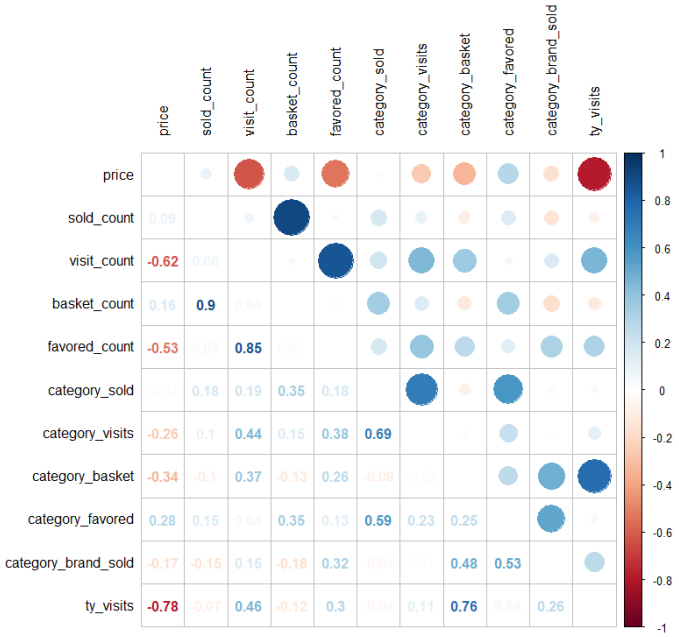

Figure 34. Correlation plot for Product 5

After checking the CCF functions only basket_count at lag 1 and fitted values of ARIMA function are used as inputs to a linear regression model for sold count.

For Product 8 correlation plot is given in Figure 35.

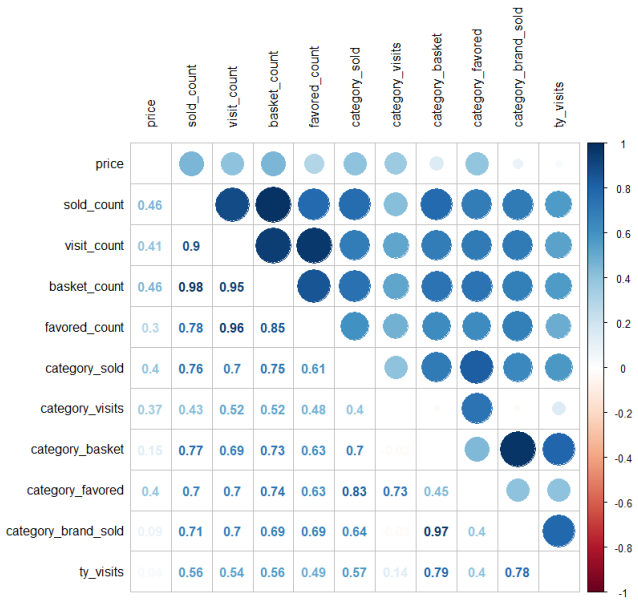

Figure 35. Correlation plot for Product 8

CCF functions are also checked and basket_count, visit_count, category_sold, category_favored, favored_count and category_baske at lag 1, as well as the fitted values of ARIMA function are used as inputs to a linear regression model for sold count. After removing the non-significant features final inputs to the linear model are fitted values, basket_count, category_favored and favored_count.

4. Results

Plots of fitted values are given in the following figures:

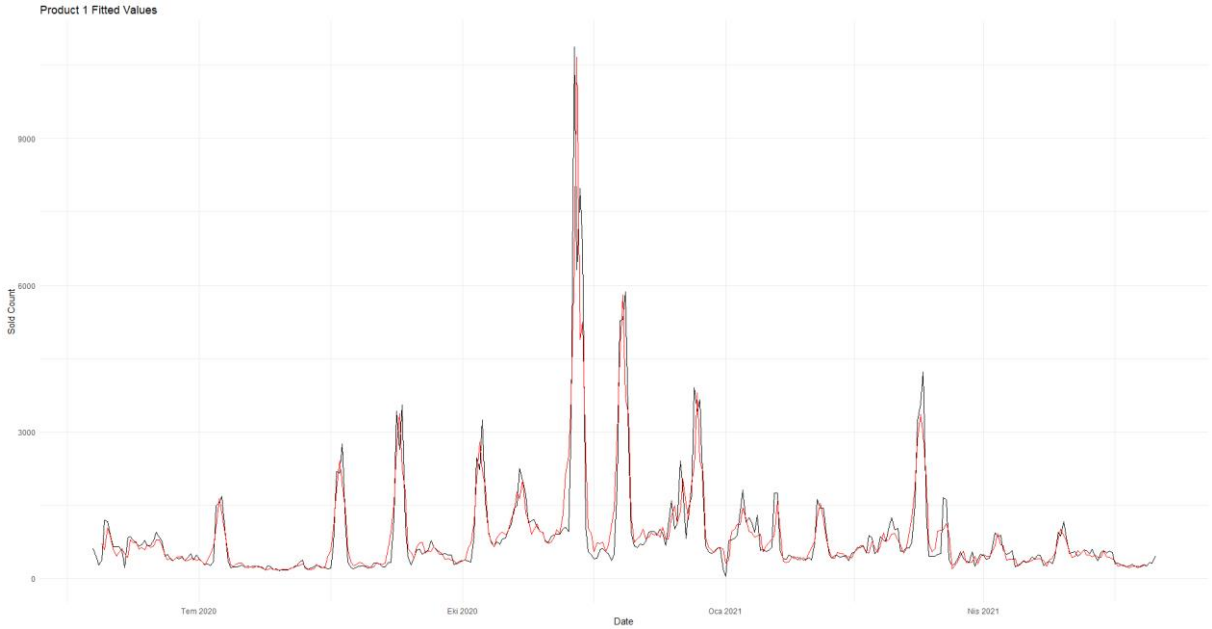

Figure 36. Predictions for Product 1

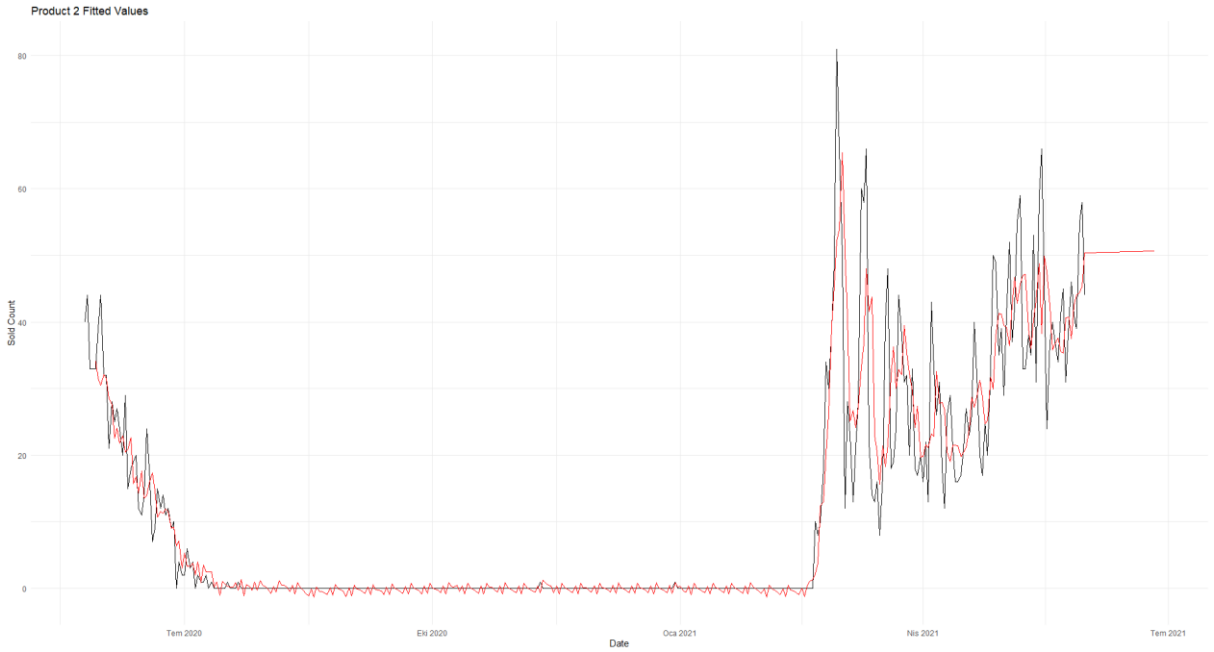

Figure 37. Predictions for Product 2

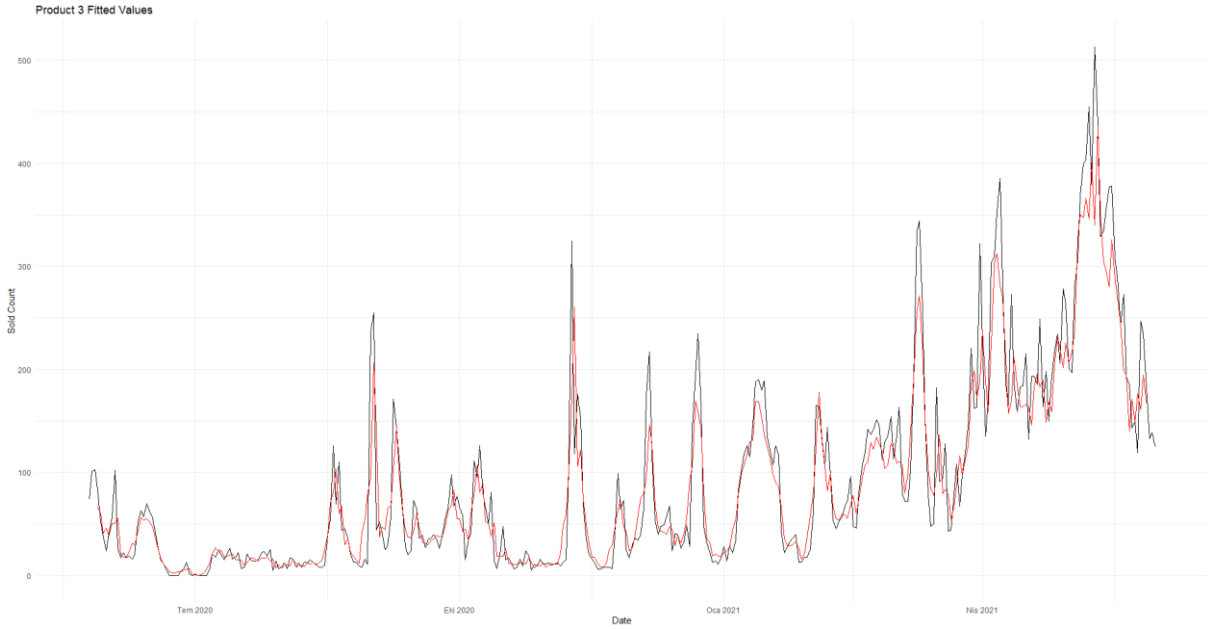

Figure 38. Predictions for Product 3

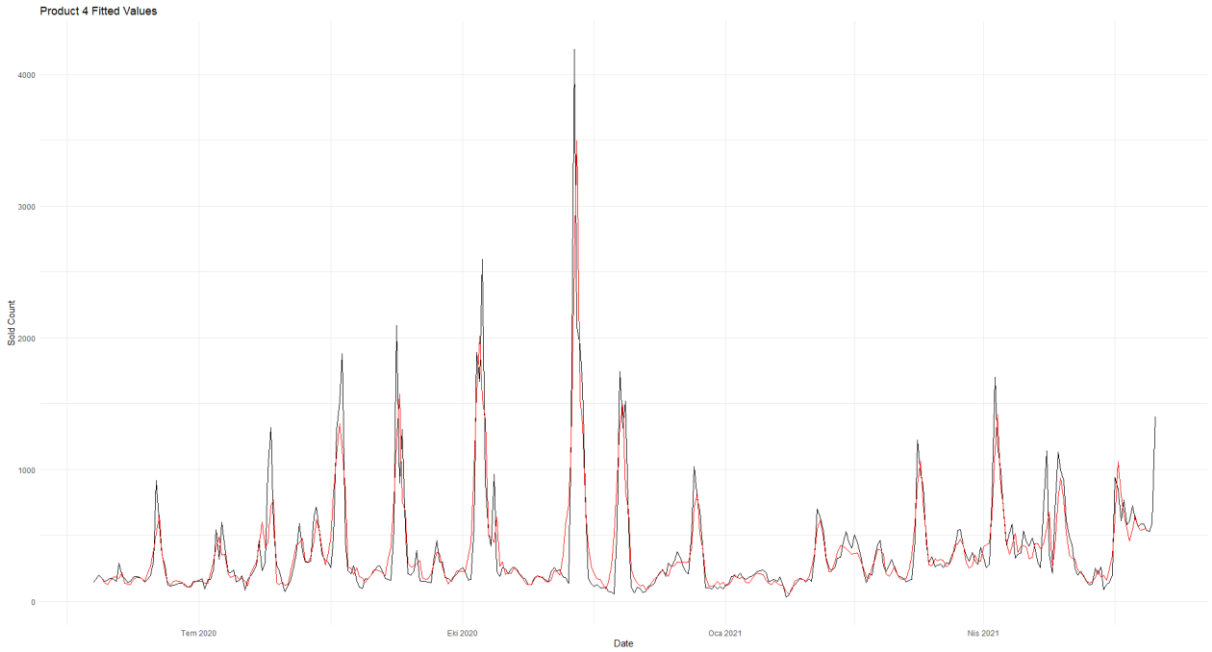

Figure 39. Predictions for Product 4

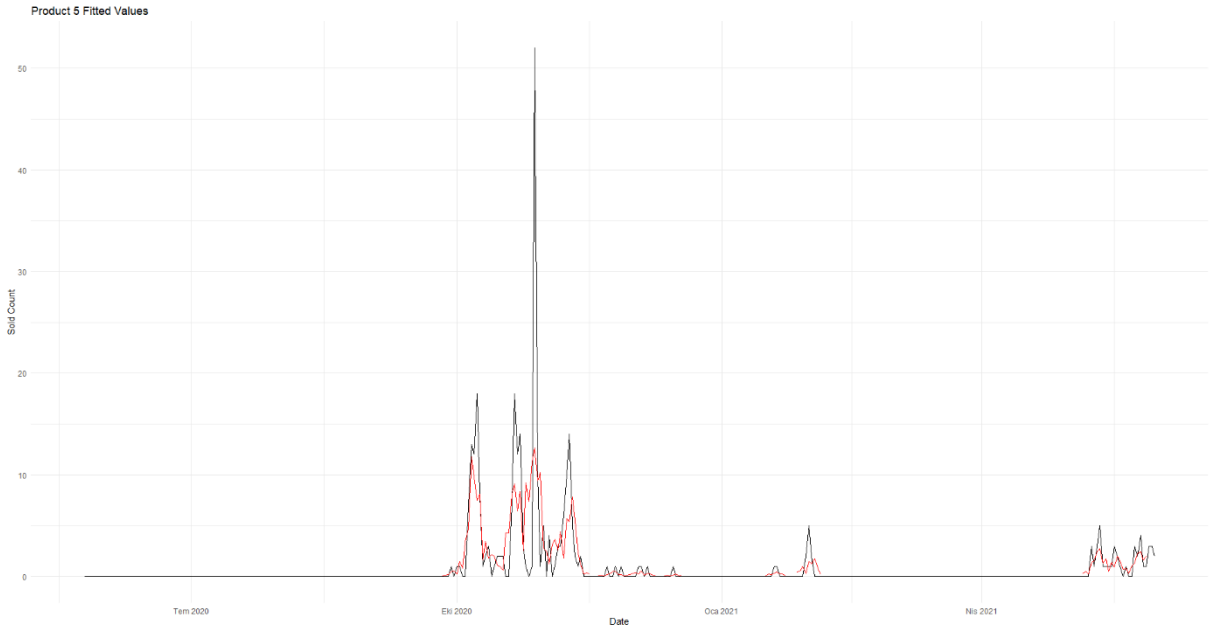

Figure 40. Predictions for Product 5

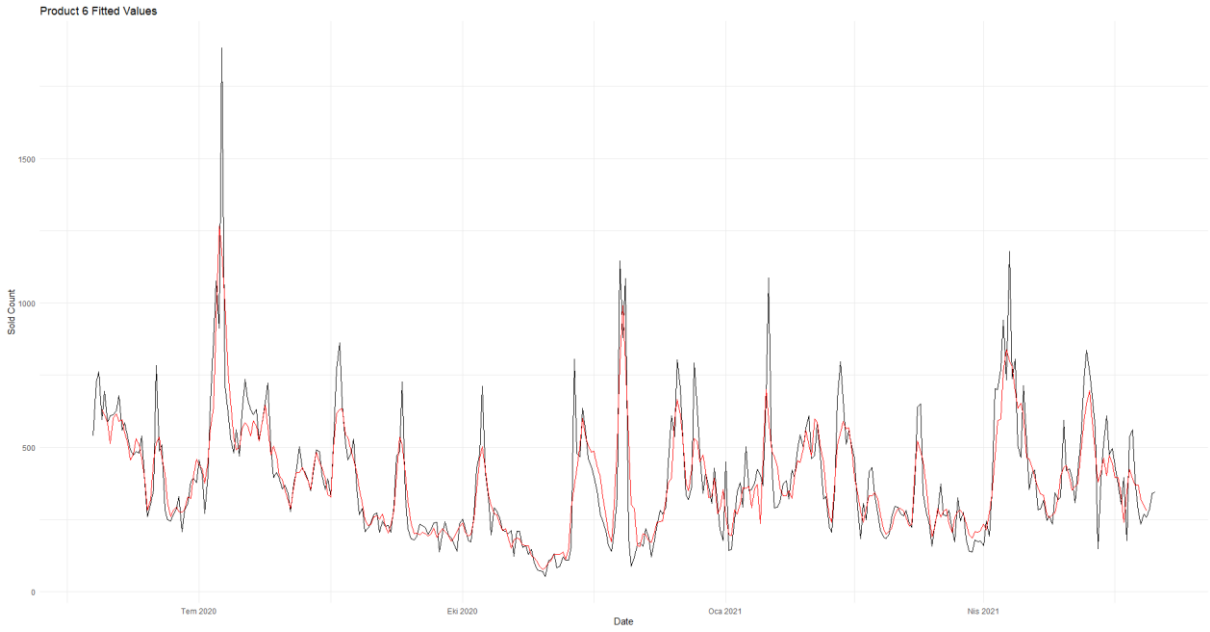

Figure 41. Predictions for Product 6

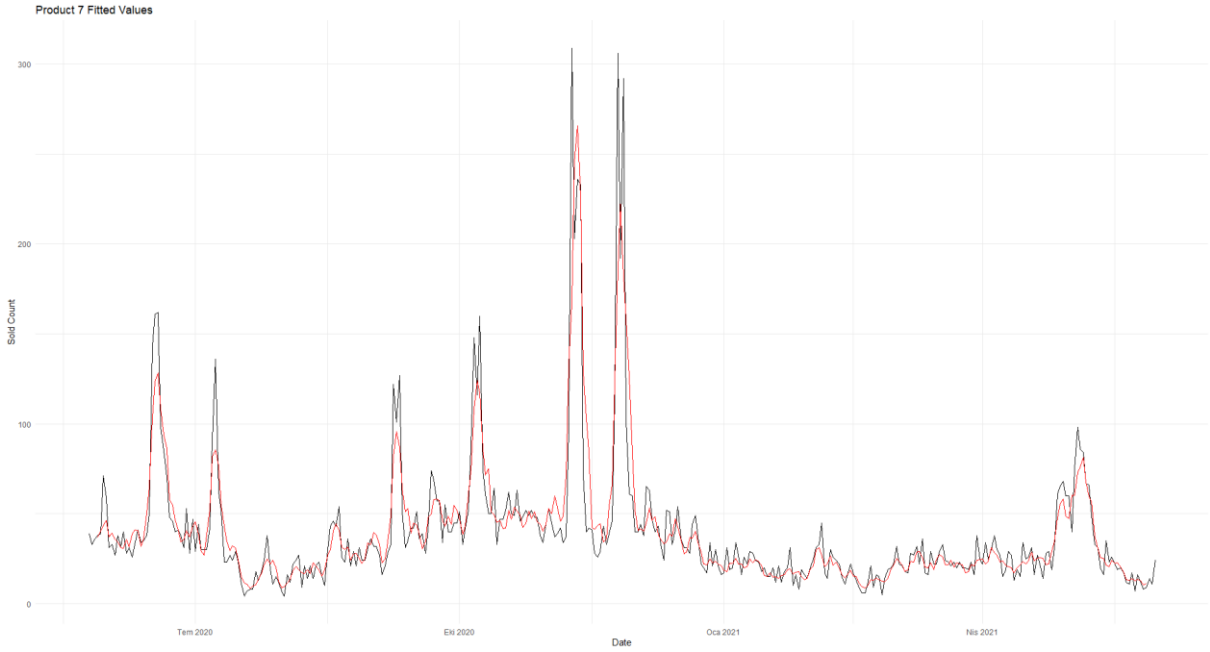

Figure 42. Predictions for Product 7

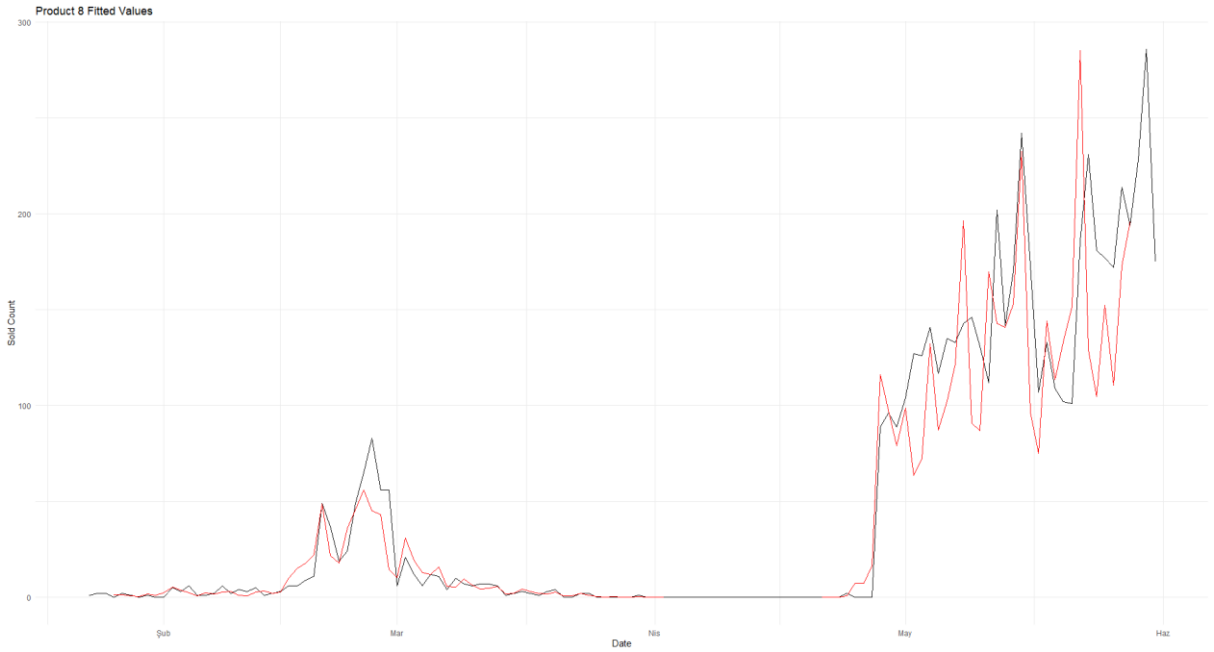

Figure 43. Predictions for Product 8 before adding regressors

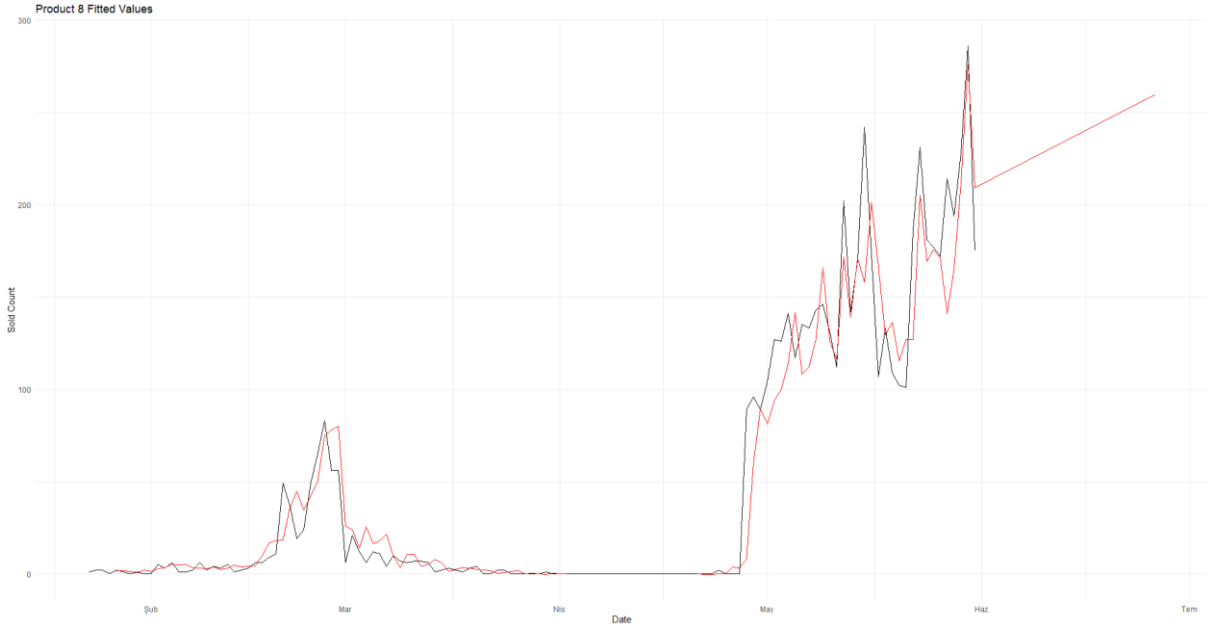

Figure 44. Predictions for Product 8 after adding regressors

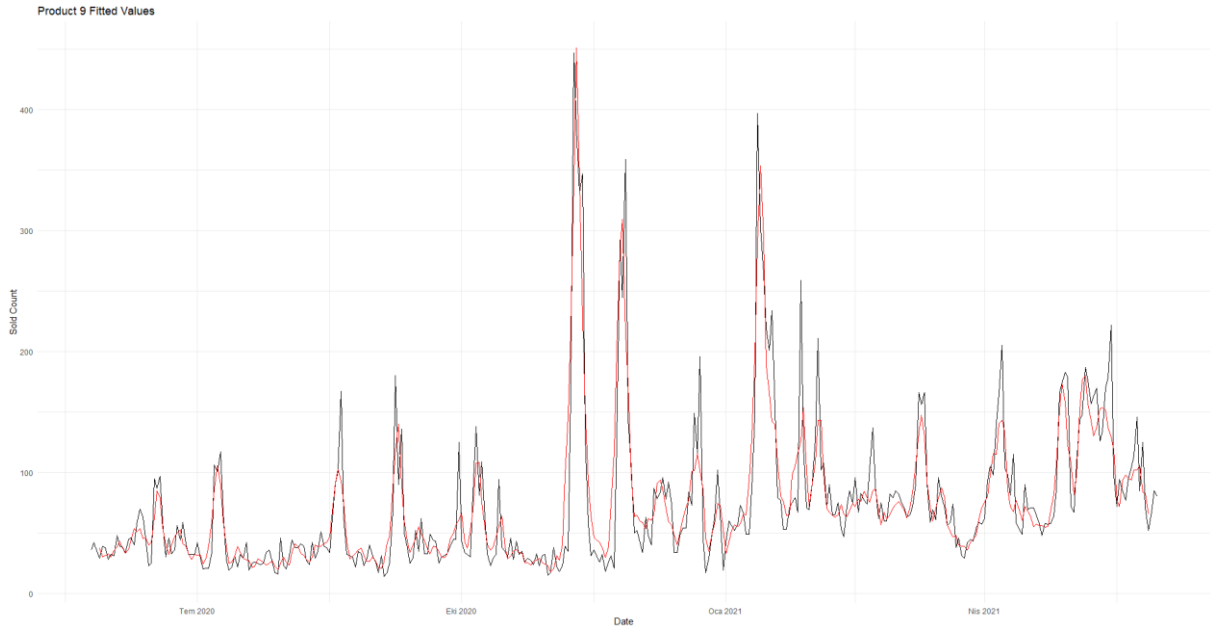

Figure 45. Predictions for Product 9

Performance measures for Product 1:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	858.207	0.001146426	0.002160051	0.9919027	0.983871	0.001146426	0.001146426

Performance measures for Product 2:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	11.01882	0.08929007	Inf	0.9919027	0.983871	0.08929007	0.08929007

Performance measures for Product 3:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	92.20968	0.01066993	Inf	0.9919027	0.983871	0.01066993	0.01066993

Performance measures for Product 4:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	385.1452	0.002554546	0.004564582	0.9919027	0.983871	0.002554546	0.002554546

Performance measures for Product 5:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
245	1.261224	0.368932	NaN	0.6821335	0.4653061	0.368932	0.368932

Performance measures for Product 6:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	392.0323	0.002509668	0.003424822	0.9919027	0.983871	0.002509668	0.002509668

Performance measures for Product 7:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	39.72581	0.02476654	0.04023471	0.9919027	0.983871	0.02476654	0.02476654

Performance measures for Product 8:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
129	45.54264	0.01787234	NaN	0.9021937	0.8139535	0.01787234	0.01787234

Performance measures for Product 9:

n	mean	FBias	MAPE	RMSE	MAD	MADP	WMAPE
372	74.19624	0.01326039	0.02099574	0.9919027	0.983871	0.01326039	0.01326039