Time Series Analysis for Candy Production

Jiang Liu
May 2023

Summary

For this project, we have used data from FRED ECONOMIC DATA to track the industrial production index of U.S. candy production every month from March 1973 to March 2023. Our primary aim of this project is to construct a reliable time series forecasting model for the future 12-month candy production in the U.S.

After exploring the time series data set, we found this candy production data set consists of trend, monthly seasonality and level components. Consequently, we started our model fitting with models that can incorporate these components. We started with the following eight models:

Model 1: Regression model with linear trend and seasonality

Model 2: Regression model with quadratic trend and seasonality

Model 3: Two level model with linear trend and seasonality and trailing moving average for residuals

Model 4: Two level model with quadratic trend and seasonality and trailing moving average for residuals

Model 5: Two level model with linear trend and seasonality and AR(1) for residuals

Model 6: Two level model with quadratic trend and seasonality and AR(1) for residuals

Model 7: Auto ARIMA model

Model 8: Holt-Winter's model with auto selection of parameters

Based on the performance accuracy on the validation data set, we chose three models to fit our entire data set. These models are Model 5, Model 7 and Model 8. After fitting them on the entire data set, we found that the best model is the Auto ARIMA model($y_t = 0.0622+0.4804(y_{t-1})+0.4118(y_{t-2})+0.2140\varepsilon_{t-1}-0.2206\varepsilon_{t-2}+0.2013(y_{t-1}- y_{t-13}) - 0.0035(y_{t-2}- y_{t-14})-0.8113\rho_{t-1}$). Notably, this model showcased a significant improvement in accuracy compared to the seasonal naive model, making it our preferred choice for this project's requirements.

This project underscored the importance of tailored model selection in time series forecasting. The careful evaluation and comparison of different models enabled us to find the most suitable one, thereby underlining the significance of a data-driven approach in achieving reliable predictions. As we move forward, we will continue to monitor the model's performance and make necessary adjustments to ensure its ongoing effectiveness.

Introduction

Regardless of the occasion - be it Halloween, New Year's, or any other holiday, it's unanimous: we all relish candy and the sweet surge of energy it offers.

Given the notable growth in the candy industry in 2022, with record sales reaching $42.6 billion, it's evident that the sector holds significant potential. The National Confectioners Association's data highlights the extensive consumer expenditure on various confections, including candy, chocolate, gum, and mints.

However, the industry faces challenges. One key factor is the rising health awareness among consumers which, while driving innovation in product development, has also led to a slowdown in revenue growth. This necessitates the industry to be more responsive and adaptive, creating healthier confectionery options that align with the shifting consumer preferences.

Additionally, inflation poses another hurdle, impacting prices across commodities and goods, and potentially affecting the profit margins of candy manufacturers. As such, a deeper analysis of the candy production data is essential to navigate these challenges and to ensure the industry's sustainable growth.

Primarily, the focus of this project lies in forecasting future candy production volumes via a rigorous data analysis process. This forecasting element serves as an invaluable tool for strategic planning, empowering manufacturers to navigate with confidence through fluctuating market conditions, irrespective of whether they present challenges or opportunities. This initiative strengthens their resilience and adaptability within a swiftly evolving marketplace.

To supplement this primary focus, our project also emphasizes the crucial role of data visualization. By converting time-series data into visually compelling graphical representations, we aim to unravel intricate trends and patterns. This transformation of data into a more accessible and comprehensible format enhances its value for industry stakeholders.

With this two-pronged approach, our project offers strategic foresight while also improving our understanding of industry trends. This makes it an all-encompassing and invaluable resource for candy manufacturers operating in today's dynamic market environment.

Eight steps of Forecasting

1. **Define Goal:**
2. **Get the Data:**
3. **Explore and visualize the series**
4. **Data preprocessing**
5. **Partition Series**
6. **Apply Forecasting Methods**
7. **Evaluate and Compare Performance**
8. **Implement Forecasts/System**

Step 1: Define Goal

The goal of this project is to create numeric forecasts of the Candy production industry for the next twelve upcoming months. The objective is to create a predictive model which will properly consider all the time series components of the historical data and effectively forecast the desired future 12 months. Naturally, the model with the highest accuracy on the entire data set will be considered the model of choice. The resulting forecasts will be used to monitor Candy production in the USA. The forecasting models developed for this project were done via the R language.

Step 2: Get data

In this report, we will analyze the monthly Industrial Production Index of candies from the year March 1973 to March 2023 with baseline year of 2017.

The manufacturing industries in the United States are measured monthly by the Industrial Production Index (IPI) which is a key economic indicator. This provides insightful information about the state and future of the economy in the United States and the performance of various sectors within it. This data is sourced from Federal Resource Economic Data (FRED) provided by the Federal Reserve Bank of St. Louis.

The dataset consists of 601 observations and two variables: Date and Production. The date variable represents the month and year for which the Industrial Production Data is calculated. Production is the variable that represents the Industrial Production Data for the manufacturing sector. It is seasonally adjusted and measured in terms of percentage change in actual output from a base year.

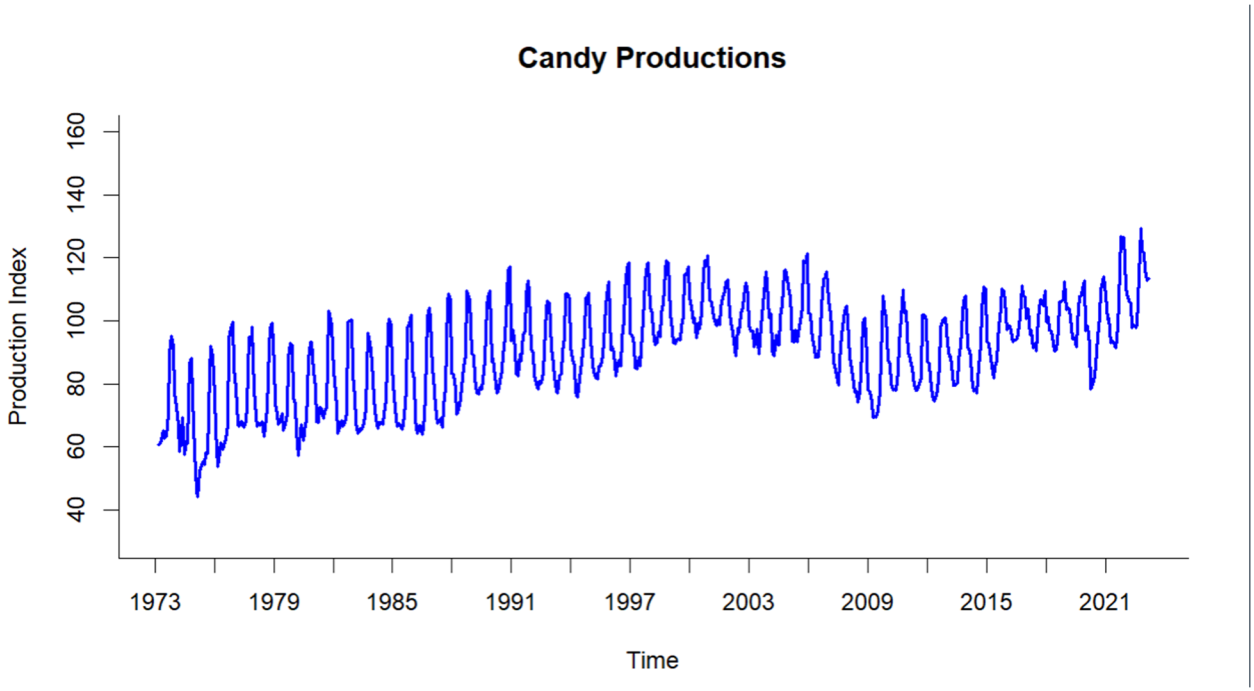
Step 3: Explore and Visualize Series

Initiating our data exploration, we scrutinized the dataset for null and missing values, with reassuringly positive results. The dataset exhibited no instances of null or missing values, thereby green-lighting our progression with the subsequent steps.

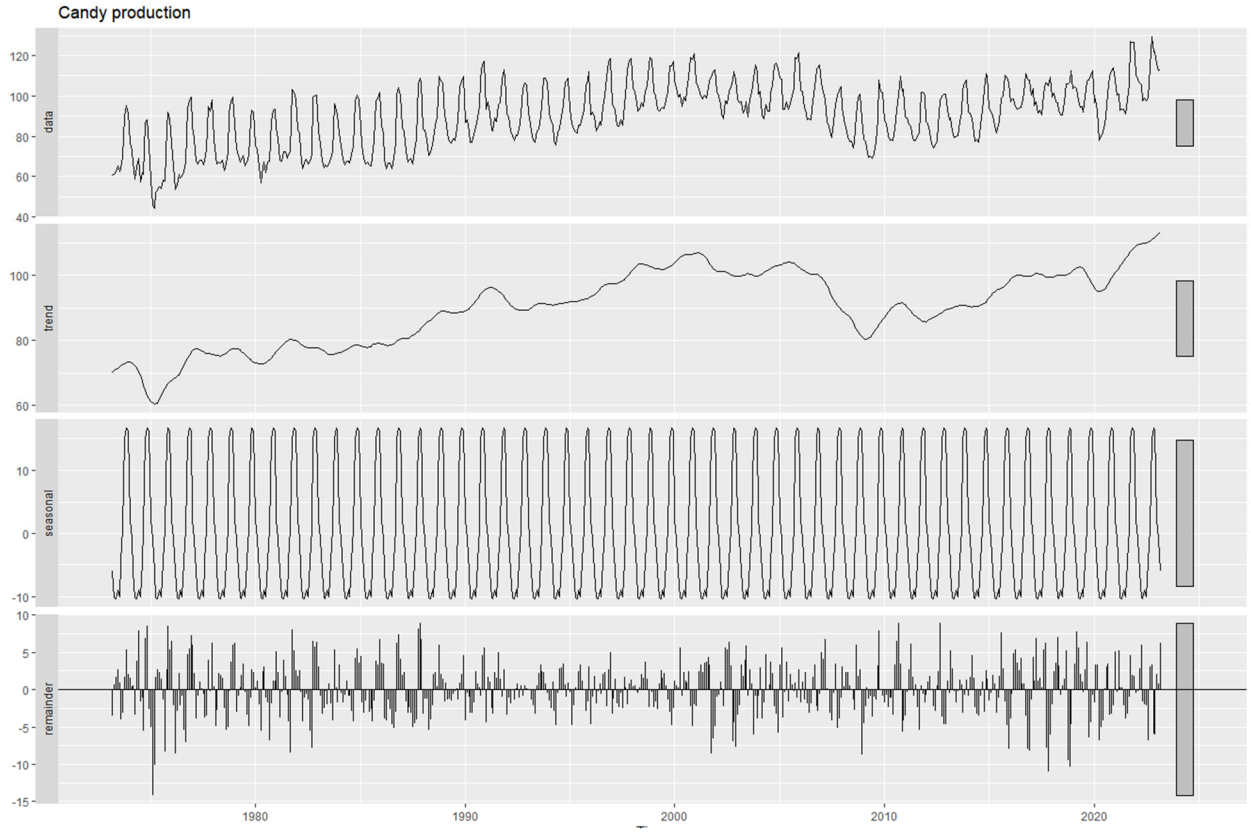
Summary statistics of the production index

Min	Max	range	Median	Mean	Standard deviation
44.21	129.27	85.06	91.73	89.93	15.86

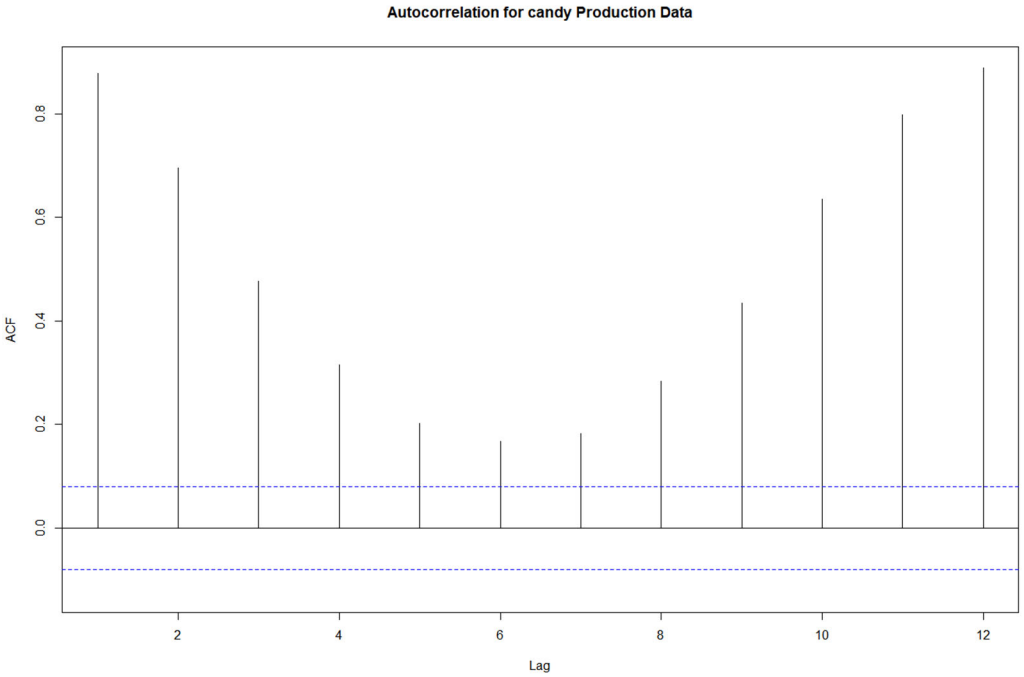
First, we looked at the summary statistics of the production index. There are 601 observations starting from March 1973 to March 2023. The production index ranges from 44.21 to 129.27, with the median of 91.73. The mean of this data set is 89.93 with a standard deviation of 15.86.



By looking at the line plot of the original data, we noticed the seasonality in the data set. The maximum occurred at time points of roughly equal lengths. The overall production index increased slowly before the year 2020 and then decreased from 2000 to 2010, then slightly increased afterwards, but we see a decrease from 2016 to 2019 relatively but also there is a rapid decrease in 2021 which might be an impact of covid as well and the sales slowly started increasing from 2022. We could also observe that the amplitude of cycles did not change much throughout the years.



Using the ACF function in R we decomposed the data set. From the above plot we can notice that in general there is an upward trend and monthly seasonality in the data set.



The autocorrelation chart of the data shows strong positive autocorrelation of all the 12 lags in the production index. A positive autocorrelation coefficient in lag 1 is substantially higher than the horizontal threshold, which is indicative of an upward trend component. The strong autocorrelation coefficient in lag 12, which is statistically significant, points to the monthly seasonality component being present in the data.

Step 4: Data Preprocessing

The data set we downloaded is a relatively clean one. There is no null value in the dataset and ready to be used.

Before model development, we would like to check the time series dataset predictability to ensure the dataset can be predictable by using these two approaches:

Approach 1: AR(1) Model

The output of AR(1) model summary is presented above. ARIMA(1,0,0) is an autoregressive(AR) model with order1, no differencing, no move average.The model intercept is 89.8516, the coefficient of the ar1(Yt-1) variable,β1 is 0.8845.

The model equation: $Yt = 89.8515 + 0.8845Yt-1$

Then, we conducted a Z-test to see if the time series data is predictable or random walk:

Null hypothesis Ho: $\beta1 = 1$

Alternative hypothesis H1: $\beta1 \neq 1$

ar1=0.8845

s.e.=0.0192

null_mean=1

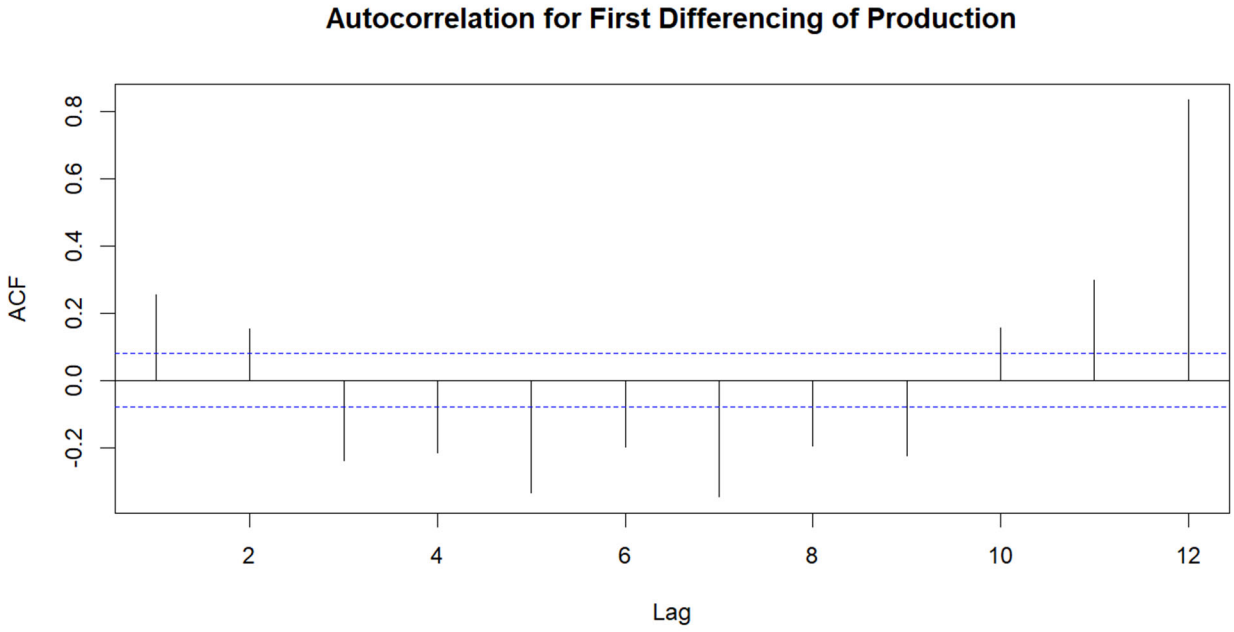
alpha =0.05

z.stat = (ar1-null_mean)/s.e= -6.205128

p.value=pnorm(z.stat)= 2.732617e-10

We can see that p-value is smaller than z.stat. We reject the null hypothesis. Therefore, the time series data for Candy Production can be predicted, not a random walk.

Approach 2:First differencing (lag1) of the historical data and Acf() function



Based on the autocorrelation graph, we can see that all of the autocorrelation coefficients that are not within the horizontal thresholds of the first difference are statistically significant. The results align with approach one.We can infer that the Candy Production is predictable.

Step 5: Partition Series

With a total of 621 records, we set a validation data set of 120 records, thus there are 401 records in the training dataset from March 1973 to January 2013. Below is a screenshot of the validation period data.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013		88.4444	88.9596	82.6390	79.5933	79.7108	80.3504	87.3886	92.0476	102.5614	106.4893	108.0843
2014	91.8221	91.3606	88.8589	81.7398	77.6142	78.3120	76.9918	85.4195	92.6979	103.4177	110.9391	110.0463
2015	93.7183	92.5262	90.2869	85.5492	81.9169	85.4599	87.4176	99.0636	99.8239	110.2747	109.3003	106.3806
2016	97.0845	97.8823	98.3514	95.0180	93.3329	93.7790	94.3197	96.2598	101.3600	111.2936	108.2221	107.7016
2017	100.6168	103.7016	95.3707	97.1333	91.3770	93.0478	90.4570	100.6918	106.5645	106.6768	104.8813	109.4814
2018	99.5552	101.3342	97.1693	96.7356	92.9164	90.4160	90.6298	96.9854	106.1213	105.9470	106.7965	112.5134
2019	103.5330	104.5199	103.8798	97.9744	94.3307	94.5633	91.8082	99.4010	107.5340	107.8665	110.2605	112.5942
2020	96.6461	98.3549	92.4525	78.1803	79.7221	82.4073	85.8352	94.9170	101.2075	109.7146	112.3875	114.0195
2021	108.0691	100.3232	101.0842	92.9450	93.4992	93.2982	91.2649	98.0119	106.2358	126.8703	126.4170	126.5746
2022	110.8825	108.4339	106.4680	105.5225	97.6997	98.8312	97.6889	98.7723	114.3283	129.2725	121.9423	121.8197
2023	116.3841											

Step 6 & 7: Apply Forecasting & Comparing Performance

Model 1: Regression model with linear trend and seasonality

For the model development, we first started with linear trend and seasonality, since we can identify the trend and seasonality from our previous analysis. Therefore, the model takes care of the trend and seasonality would be the better option.

Summary output for Regression model with linear trend and seasonality:

```
Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max
-27.585  -5.275   1.102   5.869  19.026

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.222160   1.522511  49.407  < 2e-16 ***
trend         0.060600   0.002833  21.393  < 2e-16 ***
season2      -3.116522   1.928743  -1.616 0.106808
season3      -9.400973   1.916654  -4.905 1.29e-06 ***
season4     -13.029306   1.916615  -6.798 3.26e-11 ***
season5     -12.508336   1.916579  -6.526 1.76e-10 ***
season6     -10.882561   1.916548  -5.678 2.40e-08 ***
season7     -12.164677   1.916520  -6.347 5.20e-10 ***
season8      -7.021969   1.916497  -3.664 0.000277 ***
season9      -2.214662   1.916479  -1.156 0.248441
season10     13.555210   1.916464   7.073 5.59e-12 ***
season11     15.413113   1.916453   8.043 7.34e-15 ***
season12     14.694693   1.916447   7.668 1.03e-13 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.571 on 466 degrees of freedom
Multiple R-squared:   0.72,     Adjusted R-squared:  0.7128
F-statistic: 99.84 on 12 and 466 DF,  p-value: < 2.2e-16
```

The regression model with linear trend and seasonality contains 1 trend index (t) and 11 seasonal dummy variables from season 2 to season 12.

The equation of model is:

$y_t = 75.22+0.06t -3.12D_2-9.40D_3+...+14.69D_{12}$

The regression model with linear trend and seasonality summary reveals that the coefficient of intercept, trend and majority seasons have p_values below 0.05 or 0.01, indicating statistically significant. However, the p-value of season 9 is higher than 0.05 and 0.01, therefore, they are statistically significant. The relatively high values of R-squared and Adjusted R-squared,suggest that the model fits the training partition data set well. Overall, these findings suggest that the model may also be considered for time series forecasting.

Forecasting Result for the training period from regression model with linear trend and seasonality :

	Point Forecast	Lo 0	Hi 0
Feb 2013	101.19372	101.19372	101.19372
Mar 2013	94.96987	94.96987	94.96987
Apr 2013	91.40213	91.40213	91.40213
May 2013	91.98370	91.98370	91.98370
Jun 2013	93.67008	93.67008	93.67008
Jul 2013	92.44856	92.44856	92.44856
Aug 2013	97.65187	97.65187	97.65187
Sep 2013	102.51978	102.51978	102.51978
Oct 2013	118.35025	118.35025	118.35025
Nov 2013	120.26875	120.26875	120.26875
Dec 2013	119.61093	119.61093	119.61093
Jan 2014	104.97684	104.97684	104.97684
Feb 2014	101.92092	101.92092	101.92092

Model 2: Regression model with quadratic trend and seasonality

In order to better capture the components of the dataset, the second model we develop still follows the rule that it needs to take care of the trend and seasonality.

Summary output of regression model with quadratic trend and seasonality:

```
Call:
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
     Min       1Q   Median       3Q      Max
-22.1124  -4.5893  -0.4383   5.0489  15.3859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.339e+01  1.376e+00  46.071  < 2e-16 ***
trend        2.086e-01  8.884e-03  23.478  < 2e-16 ***
I(trend^2)  -3.083e-04  1.792e-05 -17.201  < 2e-16 ***
season2     -3.416e+00  1.510e+00  -2.263   0.0241 *
season3     -9.401e+00  1.500e+00  -6.267 8.38e-10 ***
season4     -1.303e+01  1.500e+00  -8.688  < 2e-16 ***
season5     -1.251e+01  1.500e+00  -8.343 8.33e-16 ***
season6     -1.089e+01  1.500e+00  -7.260 1.64e-12 ***
season7     -1.217e+01  1.500e+00  -8.116 4.36e-15 ***
season8     -7.030e+00  1.500e+00  -4.687 3.65e-06 ***
season9     -2.222e+00  1.500e+00  -1.482   0.1391
season10     1.355e+01  1.500e+00   9.034  < 2e-16 ***
season11     1.541e+01  1.500e+00  10.273  < 2e-16 ***
season12     1.469e+01  1.500e+00   9.796  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.707 on 465 degrees of freedom
Multiple R-squared:  0.8289,    Adjusted R-squared:  0.8241
F-statistic: 173.2 on 13 and 465 DF,  p-value: < 2.2e-16
```

The regression model with quadratic trend and seasonality contains 14 independent variables, they are trend index(t), squared trend(t2)and 11 seasonal dummy variables from season 2 to season 11,the equation of model is:

$yt = 63.39+0.21t -0.0003t^2-3.42D_2-9.40D_3+...+14.69D_{12}$

The regression model with quadratic trend and seasonality summary reveals that the coefficient of intercept, trend ,and majority seasons have small p-values, indicating statistically significant, besides season 9 is larger than 0.05, therefore, it's statistically insignificant. The relatively high R-squared and Adjusted R-squared which suggest that the model fits the training partition data set quite well. Overall, these findings suggest that the model can be considered for time series forecasting.

Forecasting result of regression model with quadratic trend and seasonality on validation data:

	Point Forecast	Lo 0	Hi 0
Feb 2013	89.05949	89.05949	89.05949
Mar 2013	82.98731	82.98731	82.98731
Apr 2013	79.26790	79.26790	79.26790
May 2013	79.69780	79.69780	79.69780
Jun 2013	81.23249	81.23249	81.23249
Jul 2013	79.85930	79.85930	79.85930
Aug 2013	84.91093	84.91093	84.91093
Sep 2013	89.62716	89.62716	89.62716
Oct 2013	105.30595	105.30595	105.30595
Nov 2013	107.07278	107.07278	107.07278
Dec 2013	106.26328	106.26328	106.26328
Jan 2014	91.47751	91.47751	91.47751
Feb 2014	87.96655	87.96655	87.96655

Model 3: Regression model with linear trend and seasonality + Trailing MA

This model is a two-level model, the first level is a regression model with linear trend and seasonality. And the second level is Trailing MA to forecast the regression model's residuals(errors). The total forecast used in predictions is a combination of the regression model and trailing MA forecasts.

Summary Output of Regression Model with Linear Trend and Seasonality:

```
Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max
-27.585  -5.275   1.102   5.869  19.026

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  75.222160   1.522511  49.407 < 0.0000000000000002 ***
trend         0.060600   0.002833  21.393 < 0.0000000000000002 ***
season2      -3.116522   1.928743  -1.616              0.106808
season3      -9.400973   1.916654  -4.905  0.00000129378768600 ***
season4     -13.029306   1.916615  -6.798  0.00000000003256088 ***
season5     -12.508336   1.916579  -6.526  0.00000000017615006 ***
season6     -10.882561   1.916548  -5.678  0.00000002397824404 ***
season7     -12.164677   1.916520  -6.347  0.00000000052033292 ***
season8      -7.021969   1.916497  -3.664              0.000277 ***
season9      -2.214662   1.916479  -1.156              0.248441
season10     13.555210   1.916464   7.073  0.00000000000558577 ***
season11     15.413113   1.916453   8.043  0.00000000000000734 ***
season12     14.694693   1.916447   7.668  0.00000000000010291 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The regression model has 12 predictors including the trend and 11 seasonal dummy variables from February(season2/D2) to December(season12/D12):

The regression equation(coefficients rounded up to two decimal places) is:

yt=75.22+0.06t-3.12D2-9.40D2-13.03D3+…+14.694D12

trend(t) = time period index, e.g: t=1 for Mar 1973.

For example, if we want to calculate Feb 2013(t=480):

y_{480}= 75.22 +0.06*480-3.12=101.19

Example output of Forecast monthly production in the validation period:

	Point Forecast	Lo 0	Hi 0
Feb 2013	101.19372	101.19372	101.19372
Mar 2013	94.96987	94.96987	94.96987
Apr 2013	91.40213	91.40213	91.40213
May 2013	91.98370	91.98370	91.98370
Jun 2013	93.67008	93.67008	93.67008
Jul 2013	92.44856	92.44856	92.44856
Aug 2013	97.65187	97.65187	97.65187
Sep 2013	102.51978	102.51978	102.51978
Oct 2013	118.35025	118.35025	118.35025
Nov 2013	120.26875	120.26875	120.26875
Dec 2013	119.61093	119.61093	119.61093
Jan 2014	104.97684	104.97684	104.97684
Feb 2014	101.92092	101.92092	101.92092

Next, we developed a trailing MA(K=3) forecast for residual in the validation period without confidence interval, the sample of summary output is attached:

	Point Forecast	Lo 0	Hi 0
Feb 2013	-16.08332	-16.08332	-16.08332
Mar 2013	-16.08332	-16.08332	-16.08332
Apr 2013	-16.08332	-16.08332	-16.08332
May 2013	-16.08332	-16.08332	-16.08332
Jun 2013	-16.08332	-16.08332	-16.08332
Jul 2013	-16.08332	-16.08332	-16.08332
Aug 2013	-16.08332	-16.08332	-16.08332
Sep 2013	-16.08332	-16.08332	-16.08332
Oct 2013	-16.08332	-16.08332	-16.08332
Nov 2013	-16.08332	-16.08332	-16.08332
Dec 2013	-16.08332	-16.08332	-16.08332
Jan 2014	-16.08332	-16.08332	-16.08332
Feb 2014	-16.08332	-16.08332	-16.08332

Lastly, we combined these two models to forecast the validation part and generated a forecasting result:

	Jan	Feb	Mar	Apr	May	Jun
2013		85.11040	78.88655	75.31881	75.90038	77.58676
2014	88.89352	85.83760	79.61375	76.04602	76.62759	78.31396
2015	89.62072	86.56480	80.34095	76.77322	77.35479	79.04116
2016	90.34793	87.29200	81.06815	77.50042	78.08199	79.76837
2017	91.07513	88.01921	81.79535	78.22762	78.80919	80.49557
2018	91.80233	88.74641	82.52256	78.95482	79.53639	81.22277
2019	92.52953	89.47361	83.24976	79.68203	80.26360	81.94997
2020	93.25673	90.20081	83.97696	80.40923	80.99080	82.67717
2021	93.98394	90.92801	84.70416	81.13643	81.71800	83.40438
2022	94.71114	91.65522	85.43136	81.86363	82.44520	84.13158
2023	95.43834					
	Jul	Aug	Sep	Oct	Nov	Dec
2013	76.36524	81.56855	86.43646	102.26693	104.18543	103.52761
2014	77.09245	82.29575	87.16366	102.99413	104.91264	104.25482
2015	77.81965	83.02296	87.89086	103.72134	105.63984	104.98202
2016	78.54685	83.75016	88.61807	104.44854	106.36704	105.70922
2017	79.27405	84.47736	89.34527	105.17574	107.09424	106.43642
2018	80.00125	85.20456	90.07247	105.90294	107.82144	107.16362

Model 4: Two-level model - Quadratic Trend and Seasonality with Trailing MA

The next model we developed is another two-level model which combines Quadratic Trend and Seasonality with Trailing MA, the first level is a regression model with quadratic trend and seasonality. And the second level is Trailing MA to forecast the regression model's residuals(errors). The total forecast used in predictions is a combination of the regression model and trailing MA forecasts.

Summary Output of Quadratic Trend and Seasonality:

```
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
     Min       1Q   Median       3Q      Max
-22.1124  -4.5893  -0.4383   5.0489  15.3859

Coefficients:
              Estimate   Std. Error t value             Pr(>|t|)
(Intercept)  63.38789408   1.37586896  46.071 < 0.0000000000000002 ***
trend         0.20857858   0.00888386  23.478 < 0.0000000000000002 ***
I(trend^2)   -0.00030829   0.00001792 -17.201 < 0.0000000000000002 ***
season2      -3.41648683   1.50951797  -2.263               0.0241 *
season3      -9.40097335   1.49995693  -6.267 0.000000000837847995 ***
season4     -13.03208061   1.49992581  -8.688 < 0.0000000000000002 ***
season5     -12.51326879   1.49989798  -8.343 0.000000000000000833 ***
season6     -10.88903540   1.49987342  -7.260 0.000000000001638623 ***
season7     -12.17207543   1.49985214  -8.116 0.000000000000004360 ***
season8      -7.02967638   1.49983412  -4.687 0.000003647735094422 ***
season9      -2.22206076   1.49981937  -1.482               0.1391
season10     13.54873644   1.49980789   9.034 < 0.0000000000000002 ***
season11     15.40818022   1.49979968  10.273 < 0.0000000000000002 ***
season12     14.69191807   1.49979474   9.796 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.707 on 465 degrees of freedom
Multiple R-squared:  0.8289,    Adjusted R-squared:  0.8241
F-statistic: 173.2 on 13 and 465 DF,  p-value: < 0.00000000000000022
```

The regression model with quadratic trend and seasonality contains 5 independent variables, they are trend index(t), squared trend(t2), and 11 seasonal dummy variables: from February(season2/D2) to December(season12/D12),

the equation of the model(coefficients rounded up to two decimal places) is:

$y_t = 63.39+0.21\mathbf{t}-0.0003\mathbf{t^2}-3.42\mathbf{D_2}-9.40\mathbf{D_3}+...+14.69\mathbf{D_{12}}$

The regression model with quadratic trend and seasonality summary reveals that the coefficient of intercept, trend, and all the seasonals except seasonal 9 all have small p-values, indicating statistical significance. Additionally, the pretty high values of R-squared and Adjusted R-squared which are quite close to 1 suggest that the model fits the training partition on the data set pretty well. These findings suggest that the model can be considered forecast series forecasting.

Sample output of Forecasting Result for the training period from a regression model with quadratic trend and seasonality:

	Point Forecast	Lo 0	Hi 0
Feb 2013	89.05949	89.05949	89.05949
Mar 2013	82.98731	82.98731	82.98731
Apr 2013	79.26790	79.26790	79.26790
May 2013	79.69780	79.69780	79.69780
Jun 2013	81.23249	81.23249	81.23249
Jul 2013	79.85930	79.85930	79.85930
Aug 2013	84.91093	84.91093	84.91093
Sep 2013	89.62716	89.62716	89.62716
Oct 2013	105.30595	105.30595	105.30595
Nov 2013	107.07278	107.07278	107.07278
Dec 2013	106.26328	106.26328	106.26328
Jan 2014	91.47751	91.47751	91.47751
Feb 2014	87.96655	87.96655	87.96655

Next, we developed a trailing MA(K=3) forecast for residual in the validation period without confidence interval, the sample of summary output is attached:

	Point Forecast	Lo 0	Hi 0
Feb 2013	-2.7667281	-2.7667281	-2.7667281
Mar 2013	-1.3341663	-1.3341663	-1.3341663
Apr 2013	-0.1881168	-0.1881168	-0.1881168
May 2013	0.7287229	0.7287229	0.7287229
Jun 2013	1.4621946	1.4621946	1.4621946
Jul 2013	2.0489721	2.0489721	2.0489721
Aug 2013	2.5183940	2.5183940	2.5183940
Sep 2013	2.8939316	2.8939316	2.8939316
Oct 2013	3.1943617	3.1943617	3.1943617
Nov 2013	3.4347058	3.4347058	3.4347058
Dec 2013	3.6269811	3.6269811	3.6269811
Jan 2014	3.7808013	3.7808013	3.7808013
Feb 2014	3.9038575	3.9038575	3.9038575

Lastly, we combined these two models to forecast the validation part and generated a forecast result:

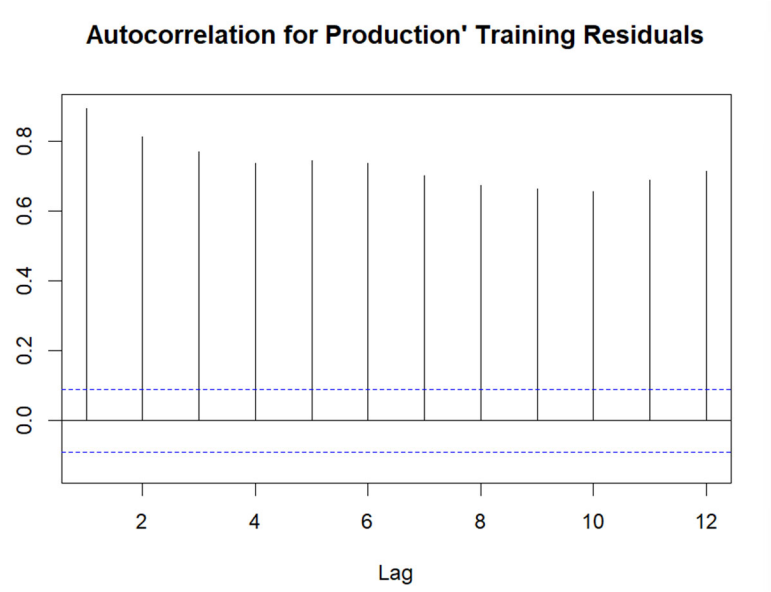
	Jan	Feb	Mar	Apr	May	Jun	Jul
2013		86.29276	81.65315	79.07979	80.42652	82.69469	81.90827
2014	95.25831	91.87041	85.88928	82.24123	82.72673	84.30443	82.96416
2015	94.65699	91.14709	85.06689	81.33809	81.75751	83.28088	81.89566
2016	93.43326	89.90809	83.81418	80.07295	80.48094	81.99368	80.59848
2017	92.08407	88.55096	82.44922	78.70024	79.10056	80.60567	79.20290
2018	90.64357	87.10302	80.99386	77.23746	77.63036	79.12806	77.71787
2019	89.11412	85.56617	79.44960	75.68580	76.07130	77.56160	76.14401
2020	87.49586	83.94051	77.81655	74.04534	74.42344	75.90635	74.48136
2021	85.78882	82.22607	76.09471	72.31611	72.68681	74.16231	72.72993
2022	83.99299	80.42284	74.28408	70.49808	70.86138	72.32949	70.88970
2023	82.10837						

	Aug	Sep	Oct	Nov	Dec
2013	87.42932	92.52109	108.50032	110.50748	109.89026
2014	88.04065	92.77529	108.46733	110.24327	109.43959
2015	86.93471	91.63791	107.30333	109.05649	108.23310
2016	85.62806	90.32222	105.97891	107.72362	106.89199
2017	84.22494	88.91159	104.56079	106.29803	105.45894
2018	82.73251	87.41174	103.05354	104.78337	103.93688
2019	81.15125	85.82309	101.45749	103.17992	102.32603

Model 5: Two-level model: Linear trend and seasonality + AR(1) for residuals

The first level is a regression model with linear trend and seasonality. And the second level is the Autoregressive AR(1) Model to forecast the regression model's residuals(errors). The total forecast used in predictions is a combination of the regression model and Autoregressive AR(1) Model forecasts.

Since we developed the first level model at model 2, we just moved forward to the second level model. We first take a look at the autocorrelation first level of residual:



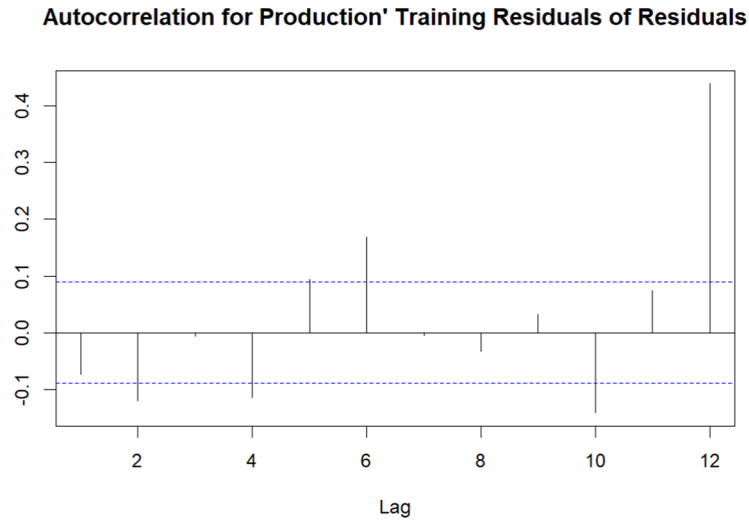
Based on the autocorrelation graph, we can see that all of the autocorrelation coefficients that are not within the horizontal thresholds and are statically significant. This indicates that these autocorrelations (relationships) between residuals are not incorporated into the regression model. Therefore, it will be a good idea to add to our forecast an AR model for residuals. This approach can potentially enhance the forecasting performance of the model.

Summary Output of AR(1) model:

```
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.8955  -0.2844
s.e.  0.0202   1.6187

sigma^2 = 14.24:  log likelihood = -1315.63
AIC=2637.27   AICc=2637.32   BIC=2649.78

Training set error measures:
                     ME     RMSE      MAE      MPE     MAPE
Training set 0.01471913 3.765942 2.863278 207.3927 312.1948
                   MASE        ACF1
Training set 0.6329507 -0.07325739
```

ARIMA(1,0,0) is an autoregressive(AR) model with order 1, no differencing, and no moving average. The model intercept is -0.2844, and the coefficient of the ar1(et-1) variable,β1 is 0.8955.

The model equation: $e_t = -0.2844 + 0.8955\ e_{t-1}$



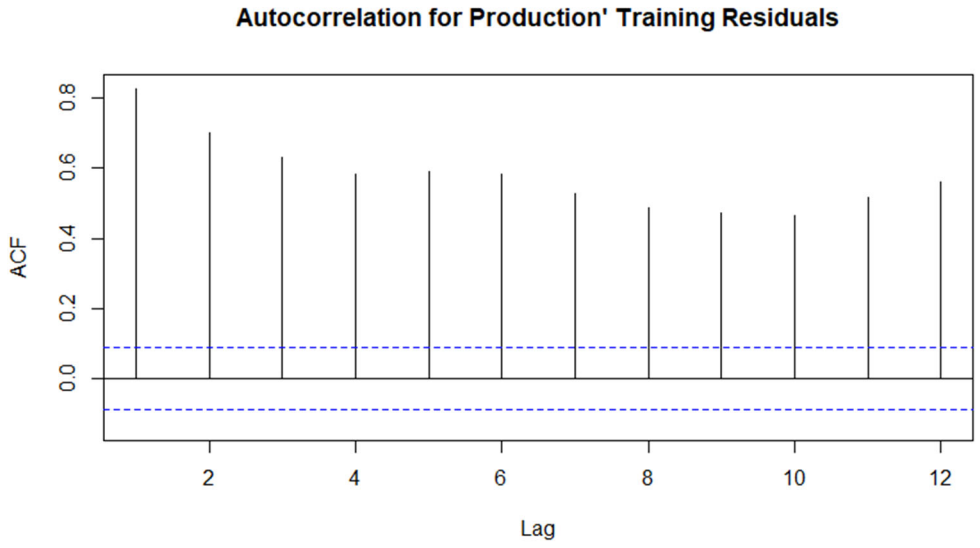
Based on the graph shown above, it is evident that many of the autocorrelations of the AR(1) model's residuals are statistically insignificant, as they lie within the horizontal thresholds. This indicates that the AR(1)residual model has captured the significant autocorrelation present in most of the lags. As a result,we can combine this model with our first-level model to enhance the model's overall forecasting performance.

The graph displayed below represents the example of validation data. In this graph, the first column shows the values of the validation data, while the second column presents the forecasted values generated by the regression model within the validation of the dataset. The third column displays the AR(1) model's forecasts of the regression residuals in the validation data, and the final column represents the sum of the second and third columns.

	Production	Reg.Forecast	AR(1)Forecast	Combined.Forecast
1	88.4444	101.19372	-10.2388836	90.95483
2	88.9596	94.96987	-9.1984628	85.77140
3	82.6390	91.40213	-8.2667844	83.13535
4	79.5933	91.98370	-7.4324827	84.55122
5	79.7108	93.67008	-6.6853803	86.98470
6	80.3504	92.44856	-6.0163632	86.43220
7	87.3886	97.65187	-5.4172702	92.23460
8	92.0476	102.51978	-4.8807930	97.63899
9	102.5614	118.35025	-4.4003871	113.94986
10	106.4893	120.26875	-3.9701920	116.29856
11	108.0843	119.61093	-3.5849600	116.02597
12	91.8221	104.97684	-3.2399914	101.73685

Model 6: Two-level model: Quadratic trend and seasonality + AR(1) for residuals

The first level is a regression model with quadratic trend and seasonality. And the second level is the Autoregressive AR(1) Model to forecast the regression model's residuals(errors). The total forecast used in predictions is a combination of the regression model and Autoregressive AR(1) Model forecasts. The quadratic regression model has been explained previously, so we will skip that part in this model's explanation.

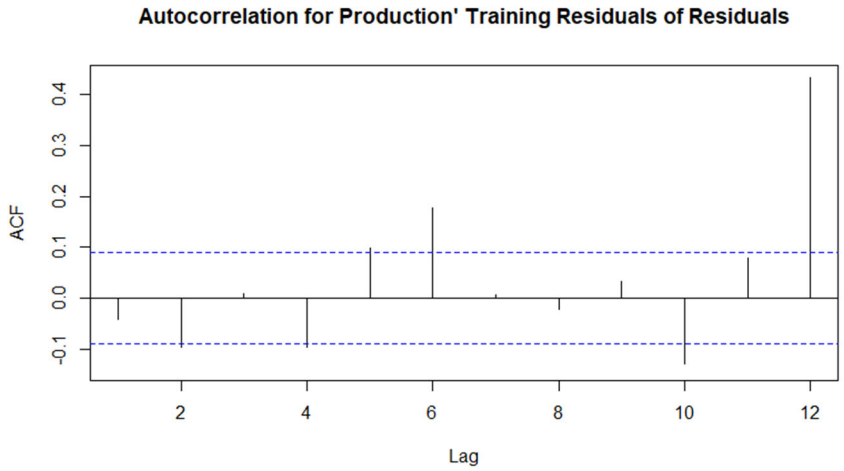


We can see from the above autocorrelation chart that the autocorrelation of the training residuals of the quadratic regression model is all statistically significant. This indicates that the first-level quadratic model failed to incorporate autocorrelation and an AR(1) model is needed for the second level.

```
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1    mean
      0.8281  0.0660
s.e.  0.0254  0.9726

sigma^2 = 13.72:  log likelihood = -1306.43
AIC=2618.85   AICc=2618.9   BIC=2631.37
```

ARIMA(1,0,0) is an autoregressive(AR) model with order 1, no differencing, and no moving average. The model intercept is 0.0660, and the coefficient of the ar1(et-1) variable,β1 is 0.8281.

The model equation: $e_t = 0.066 + 0.8281\ e_{t-1}$



From the above autocorrelation chart we can see, after applying the second level AR(1) model, the majority of autocorrelation coefficients dropped within the horizontal threshold. This implies that our AR(1) model successfully incorporates the autocorrelation relationship.

Model 7: Auto ARIMA model

The summary of the Auto ARIMA model is shown below:

```
ARIMA(2,0,2)(2,1,1)[12]

Coefficients:
         ar1     ar2     ma1      ma2    sar1     sar2     sma1
      0.3561  0.5446  0.3369  -0.2666  0.1194  -0.0521  -0.7330
s.e.  0.2771  0.2509  0.2732   0.0643  0.0701   0.0628   0.0556

sigma^2 = 10.79:  log likelihood = -1218.9
AIC=2453.81   AICc=2454.12   BIC=2486.98
```

This is a seasonal ARIMA model, ARIMA(p,d,q)(P, D,Q)[m] where:

p = 2, order 2 autoregressive model AR(2)

d = 0, no differencing

q = 2, order 2 moving average MA(2) for error lags

P = 2, order 2 autoregressive model AR(2) for seasonality

D= 1, order 1 differencing to remove linear trend

Q = 1, order 1 moving average MA(1) for the seasonal error lags

m =12, for monthly seasonality

The model equation is:

$y_t = 0.3561y_{t-1}+0.5446y_{t-2}+0.3369\varepsilon_{t-1} -0.2666\varepsilon_{t-2}+0.1194(y_{t-1} - y_{t-13})-0.0521(y_{t-2} - y_{t-14})- 0.733\rho_{t-1}$

Model 8: HW.ZZZ model

This model is the Holt-Winter's(HW) model with the automated selection of model options and automated selection of the smoothing parameters for the training period. The summary is presented below:

```
ETS(A,N,A)

Call:
 ets(y = train.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.5647
    gamma = 0.2446

  Initial states:
    l = 73.8263
    s = -3.5771 5.3768 20.4881 21.5874 19.923 -4.346
           -7.4779 -10.8329 -6.2561 -10.031 -12.5744 -12.28

  sigma:  3.4247

     AIC     AICc      BIC
4151.334 4152.371 4213.909
```

The HW model has the (A, N, A) options, which means additive error, no trend and additive seasonality. The optimal value for exponential smoothing constant alpha is 0.5647, no smoothing constant for trend estimate (beta = 0), and the smoothing constant for seasonal estimate gamma is 0.2446. The large value of alpha indicates that the model is more responsive to variation and the model's level component tends to be more local. The relatively small value of gamma means that the seasonal component is changing relatively slowly over time.

Accuracy performance on the validation period of the 7 models:

Model Number	Model Name	RMSE	MAPE
1	Linear trend and seasonality	9.071	7.777
2	Quadratic Trend and Seasonality	18.803	15.432
3	Regression model with linear trend and seasonality + Trailing MA	11.278	9.535
4	Quadratic Trend and Seasonality + Trailing MA	15.369	12.034
5	Linear trend and seasonality + AR(1) for residuals (first rank)	8.094	6.756
6	Quadratic trend and seasonality + AR(1) for residuals	18.748	15.375
7	Auto ARIMA model(third rank)	11.326	9.167
8	HW.ZZZ model (second rank)	10.514	8.416

After meticulously developing a variety of models using the training and validation datasets, we must now select the top three models that exhibit superior accuracy. These models will be used to forecast future periods by applying them to the entire dataset. Evaluating the performance metrics, Model 5 emerges as the best, showcasing a Mean Absolute Percentage Error (MAPE) of 6.756% and a Root Mean Square Error (RMSE) of 8.094. Following closely is Model 8, denoted as HW.zzz, with a MAPE of 8.416% and RMSE of 10.514. Lastly, the third most effective model is Model 7, featuring a MAPE of 9.167% and RMSE of 11.326.

Forecasting on the Entire Dataset:

Model 5 Linear trend and seasonality + AR(1) for residuals on the entire dataset:

Summary output:

```
Call:
tslm(formula = prod.ts ~ trend + season)

Residuals:
     Min       1Q   Median       3Q      Max
-29.2287  -6.3727   0.3148   5.8767  19.9812

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  76.875048   1.329238  57.834 < 0.0000000000000002 ***
trend         0.050972   0.001971  25.858 < 0.0000000000000002 ***
season2      -2.894440   1.676551  -1.726               0.0848 .
season3      -8.323777   1.668330  -4.989 0.000000799180183381 ***
season4     -12.498507   1.676644  -7.454 0.000000000000324635 ***
season5     -12.712918   1.676624  -7.582 0.000000000000132951 ***
season6     -11.258328   1.676607  -6.715 0.000000000044424223 ***
season7     -12.347752   1.676592  -7.365 0.000000000000602541 ***
season8      -6.833131   1.676579  -4.076 0.000052195771207842 ***
season9      -1.569579   1.676569  -0.936               0.3496
season10     12.763313   1.676560   7.613 0.000000000000107395 ***
season11     14.321959   1.676555   8.542 < 0.0000000000000002 ***
season12     13.976324   1.676551   8.336 0.000000000000000544 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.383 on 588 degrees of freedom
Multiple R-squared:  0.7261,    Adjusted R-squared:  0.7205
F-statistic: 129.9 on 12 and 588 DF,  p-value: < 0.00000000000000022
```

The regression model with linear trend and seasonality contains 1 trend index (t) and 11 seasonal dummy variables from February(season2/D2) to December(season12/D12), the equation of the model is:

$$y_t = 76.88+0.05t -2.89D_2-8.32D_3+...+13.98D_{12}$$

The regression model with linear trend and seasonality summary reveals that the coefficient of intercept, trend and most of the seasons all have p-values below 0.05 or 0.01, indicating statistically significant. However, the p-value of season 2 and season 9 are higher than 0.05., therefore, they are statistically insignificant. The relatively high values of R-squared and Adjusted R-squared suggest that Overall, this model is a good fit and can be used for forecasting candy production in 2023 and 2024.

Summary Output of AR(1) model on the entire dataset:

```
Series: trend.season$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1    mean
      0.8841  0.0748
s.e.  0.0191  1.3537

sigma^2 = 15.22:  log likelihood = -1670.77
AIC=3347.55   AICc=3347.59   BIC=3360.74

Training set error measures:
                     ME     RMSE      MAE    MPE     MAPE      MASE         ACF1
Training set 0.01901515 3.895358 2.94182 56.319 226.8498 0.6508605 -0.04484622
>
```

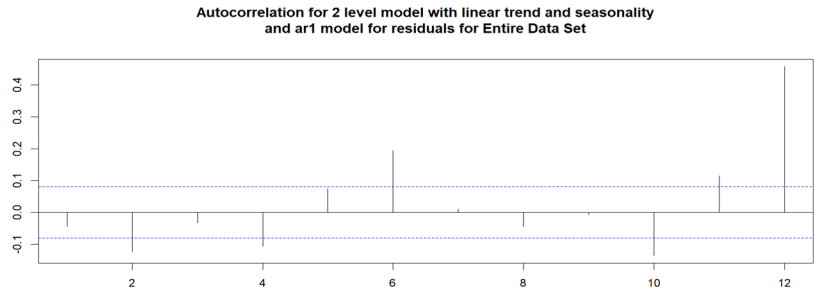
ARIMA(1,0,0) is an autoregressive(AR) model with order 1, no differencing, and no moving average. The model intercept is 0.0748, and the coefficient of the ar1(tt-1) variable,β1 is 0.8841.

The model equation: $e_t = 0.0748 + 0.8841\ e_{t-1}$

After combining these two models, the forecast for the 12 future periods is presented below:

	Reg.Forecast	AR(1)Forecast	Combined.Forecast
1	95.06185	12.416200	107.4781
2	94.89842	10.985592	105.8840
3	96.40398	9.720819	106.1248
4	95.36553	8.602657	103.9682
5	100.93112	7.614112	108.5452
6	106.24564	6.740159	112.9858
7	120.62951	5.967513	126.5970
8	122.23913	5.284433	127.5236
9	121.94446	4.680534	126.6250
10	108.01911	4.146639	112.1658
11	105.17564	3.674633	108.8503
12	99.79728	3.257342	103.0546

In order to validate that the model has effectively accounted for autocorrelation in the residuals, we can refer to the graph provided below. Upon examination, it's apparent that five lags remain statistically significant - specifically, the lags at positions 2, 4, 6, 10, and 12. Among these, the lag at position 12 is the most statistically significant. However, it's reassuring to note that the model has successfully addressed the autocorrelation in the remaining lags, hence ensuring its overall efficiency and accuracy.



Model 8 HW.ZZZ model on the entire dataset:

This model is Holt-Winter's(HW) model with the automated selection of model options and automated selection of the smoothing parameters for the training period. The summary is presented below:

```
ETS(A,N,A)

Call:
 ets(y = prod.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.5929
    gamma = 0.218

  Initial states:
    l = 77.4076
    s = -3.3594 4.8847 21.634 20.1123 18.2958 -5.4731
           -6.0482 -11.0616 -6.8725 -8.9349 -12.1054 -11.0719

  sigma:  3.533

     AIC     AICc      BIC
5378.506 5379.327 5444.485
```

The HW model has the (A, N, A) options, which means additive error, no trend and additive seasonality. The optimal value for exponential smoothing constant alpha is 0.5929, no smoothing constant for trend estimate (beta = 0), and the smoothing constant for seasonal estimate gamma is 0.218. The large value of alpha indicates that the model is more responsive to variation and the model's level component tends to be more local. The relatively small value of gamma means that the seasonal component is changing relatively slowly over time.

Model 7 Auto ARMIA on the entire dataset:

Summary output for Auto ARIMA model on the entire dataset:

```
Series: prod.ts
ARIMA(2,0,2)(2,1,1)[12] with drift

Coefficients:
         ar1     ar2     ma1      ma2    sar1     sar2     sma1
      0.4804  0.4118  0.2140  -0.2206  0.2013  -0.0035  -0.8113
s.e.  0.3637  0.3254  0.3604   0.0742  0.0586   0.0538   0.0414
       drift
      0.0622
s.e.  0.0266

sigma^2 = 11.29:  log likelihood = -1550.57
AIC=3119.14   AICc=3119.45   BIC=3158.55

Training set error measures:
                     ME   RMSE      MAE        MPE     MAPE
Training set -0.009261294 3.3039 2.489657 -0.1123761 2.829186
                   MASE         ACF1
Training set 0.5434558 0.0003639371
```

The ARIMA(2,0,2)(2,1,1)[12] is a seasonal ARIMA model (ARIMA(p,d,q)(P,D,Q)[m]):

p = 2, order 2 autoregressive model AR(2)

d = 0, order 0 differencing to remove linear trend

q = 2, order 2 moving average MA(0) for error lags

P = 2, order 2 autoregressive model AR(1) for seasonality

D = 1, order 1 differencing to remove linear trend

Q = 1, order 1 moving average MA(2) for error lags

M = 12, for monthly seasonality

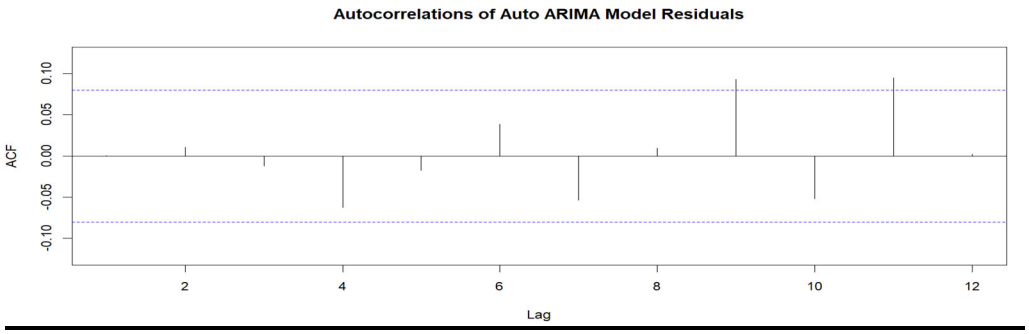
The model equation is:

$y_t = 0.0622+0.4804(y_{t-1})+0.4118(y_{t-2})+0.2140\varepsilon t-1-0.2206\varepsilon t-2+0.2013(y_{t-1}- y_{t-13}) - 0.0035(y_{t-2}- y_{t-14})-0.8113\rho_{t-1}$

The ARIMA(2,0,2)(2,1,1)[12] model Forecasstng result for the future 12 periods:

	Point Forecast	Lo 0	Hi 0
Apr 2023	107.7467	107.7467	107.7467
May 2023	102.2878	102.2878	102.2878
Jun 2023	102.7853	102.7853	102.7853
Jul 2023	101.5615	101.5615	101.5615
Aug 2023	106.0593	106.0593	106.0593
Sep 2023	115.8528	115.8528	115.8528
Oct 2023	126.6361	126.6361	126.6361
Nov 2023	123.8031	123.8031	123.8031
Dec 2023	124.1211	124.1211	124.1211
Jan 2024	114.6024	114.6024	114.6024
Feb 2024	112.0399	112.0399	112.0399
Mar 2024	110.3046	110.3046	110.3046

To ensure that the autocorrelation in the residuals is adequately accounted for by the model, we can examine the graph below. It becomes apparent that the model has effectively captured most of the lag. Although two significant lags persist, the randomness of the remaining lags suggests that the model has sufficiently addressed the issue of autocorrelation. Consequently, we can confidently conclude that any concerns regarding autocorrelation have been duly dealt with by the model.

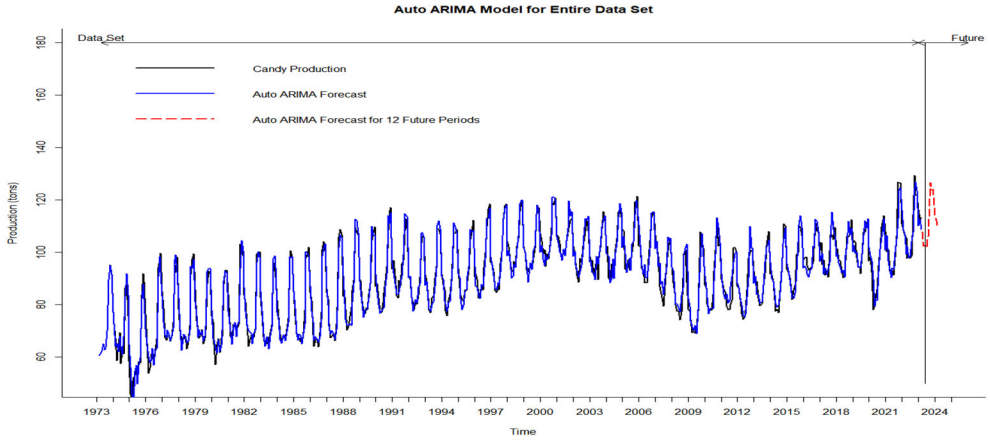


Conclusion:

Performance accuracy on the entire data set:

Model Number	Model Name	RMSE	MAPE
5	**Linear trend and seasonality + AR(1)**	**3.9**	**3.37**
7	**HW.ZZZ**	**3.49**	**3.06**
8	**Auto.ARMIA**	**3.3**	**2.83**
	Seasonal Naive	**6.06**	**5.36**

The error metric table provides a clear conclusion regarding the most effective model for forecasting candy production for the period 2023 - 2024. The Auto ARIMA model stands out as the premier choice, demonstrating the highest efficiency and accuracy. Closely following in terms of performance is Holt's Winter (ZZZ) model, which serves as the second most optimal selection for our forecasting needs. It is noticeable that these three models all performed better than the seasonal naive model.



Appendix

- Data resource link: https://fred.stlouisfed.org/series/IPG3113N

- 2022 candy industry report: https://www.fooddive.com/news/candy-sales-grew-2022-inflation-state-of-treating/644236/