Time Series Analysis of Candy Production in the United States

Course: MA5781 - Time Series Analysis
Professor: Dr. Yeonwoo Rho

Kasia Krueger
Lorenzo Gordon

Table of Contents

1. Abstract

The aim of this study is to analyze U.S. candy production data from 1972 to 2017 in order to identify the best-fitting model for the candy production, and forecast for future production based on various seasons. The goal is to detrend the data, find an appropriate model, and forecast future candy production from this model. With more than 40 years of candy production data in the U.S., there is plenty of data to analyze trends, seasonality, and monthly mean production of candy in the U.S. The aim of this study is to find if there is any seasonal pattern in candy production data, and can it be predicted?

2. Overview

The U.S. candy production data dataset was retrieved from Kaggle and was chosen for analysis for it's interesting trends and seasonality. The dataset explores annual candy production by month from 1972-2017. The data is retrieved from the industrial production (IP) index, which measures the real output of all relevant establishments located in the United States, regardless of their ownership, but not those located in U.S. territories [3]. The total candy production is measured in capacity index (2012=100). The production index measures real output and is expressed as a percentage of real output in a base year, currently 2012. The capacity index, which is an estimate of sustainable potential output, is also expressed as a percentage of actual output in 2012 [2].

3. Introduction

Candy consumption is a very large industry in the United States, and a very big part of the U.S. holidays. Candy production is a $13 billion market in the United States and is expected to grow 0.8% per year [4]. For candy manufacturers, it is important to predict and forecast the amount of candy to produce before the heavy consumption during the U.S. holiday months (October-December). Furthermore, it is equally as important to not produce more than needed, especially during the "New Year's resolution" months and the summer months, when candy consumption is much lower than during the holidays. It is not as simple as producing 0.8% more candy each year; candy manufacturers must account for the seasonality and volatility of candy consumption, and therefore, produce according to the seasonality to match the demands of the U.S. market. A model can be built to estimate the production by using time series analysis techniques. Building the model will be approached in the following ways, as outlined in *Time Series Analysis with applications in R* [1]:

1. Plotting and detrending the time series data.
2. Choosing the model that best fits the data.
    a. Seasonal means: a least squares estimation comparing means for each month.
    b. Cosine trend: more parsimonious than seasonal means.
    c. Seasonal ARIMA using auto.arima()
    d. Seasonal ARIMA using information criteria
3. Checking residuals for each model.
    a. Zero mean and Homoscedasticity.
        i. Residual plots: random, even distribution over mean zero.
        ii. Residual plots: random, evenly distributed variance.
    b. Normality
        i. Q-Q plot: residuals are on a straight line without heavily- skewed tails or outliers.
        ii. Histogram: normal distribution in a 68-95-99.7 pattern.
        iii. Shapiro-Wilk test: formal test for normality.
    c. Independence
        i. Runs test: tests hypothesis that the elements of a data sequence are mutually independent.
4. Check for overdifferencing (stationarity).

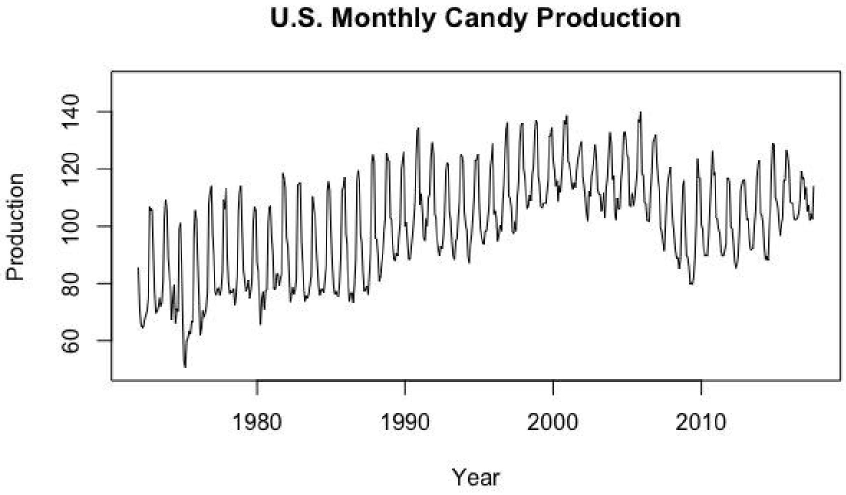
   a. Augmented Dickey-Fuller (ADF) test: null hypothesis that time series is nonstationary.
   b. Phillips-Perron (PP) test: null hypothesis that time series is nonstationary.
   c. KPSS test: null hypothesis that time series is stationary.
5. Choosing a model by minimizing information criteria.
   a. AIC, AICc - used for forecasting.
   b. BIC - used for finding a true model; more parsimonious than AIC, AICc.
6. Forecasting and modeling future production using chosen models.

4. Modeling and Analysis

4.1. Plotting the time series

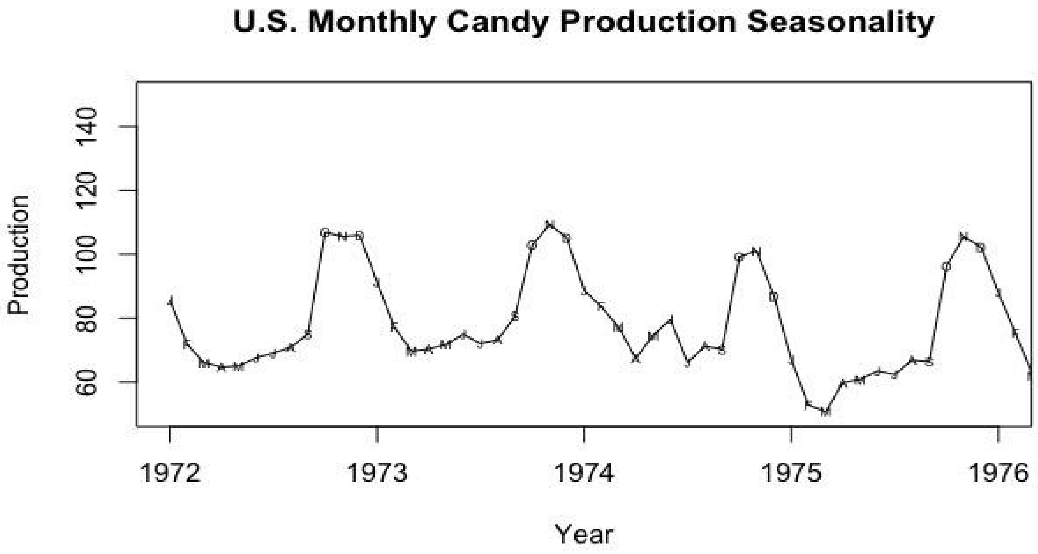
The candy production time series plot in *Figure 1* shows an upward trend and high volatility in production for each year. The strong volatility does not seem to change over time, so there is evidence of seasonality.

Figure 1. Time Series Plot of Monthly U.S. Candy Production



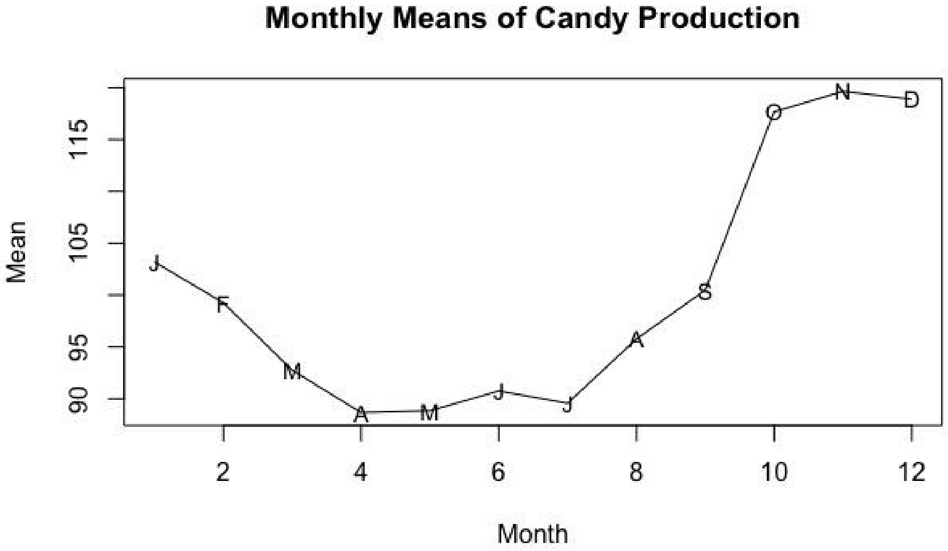
The first four years of the data were plotted to see the seasonality better. In *Figure 2*, the candy production is much higher during October, November, and December, while the lowest production occurs between March and August.

Figure 2. Seasonality of U.S. Monthly Candy Production



The monthly means plot in *Figure 3* shows an obvious uptick of production before the U.S. “holiday months” (August-December), then a slight drop during the “New Years Resolution” period (January-March), and the lowest production during the spring and early summer months (March-July), indicating roughly 6-month seasonality. Later, the study will explore how the 6-month seasonality affects the model selection.

Figure 3. Monthly Means of Candy Production



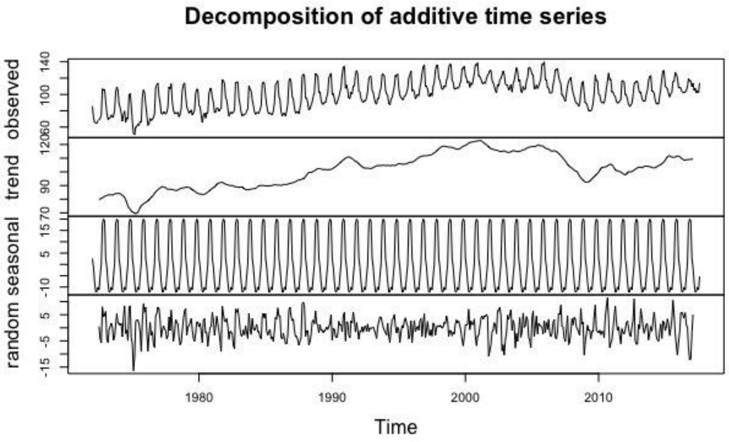
4.2 Decomposing the time series

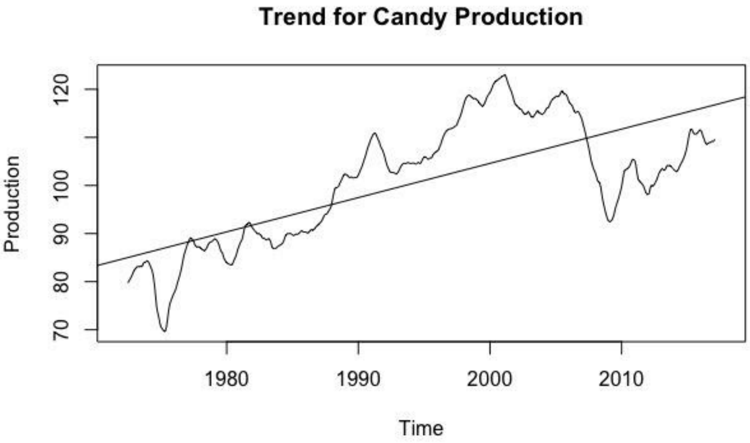
In *Figure 4* the time series decomposition shows there is an upward linear trend. The trend plot shows a possible stochastic trend. The random plot shows a steady random fluctuation around the mean, with no patterns present.

A closer look at the trend plot in *Figure 5* shows that fitting a line linear model to the trend series, an upward linear trend can be observed.

Figure 4. Decomposition of U.S. Monthly Candy Production

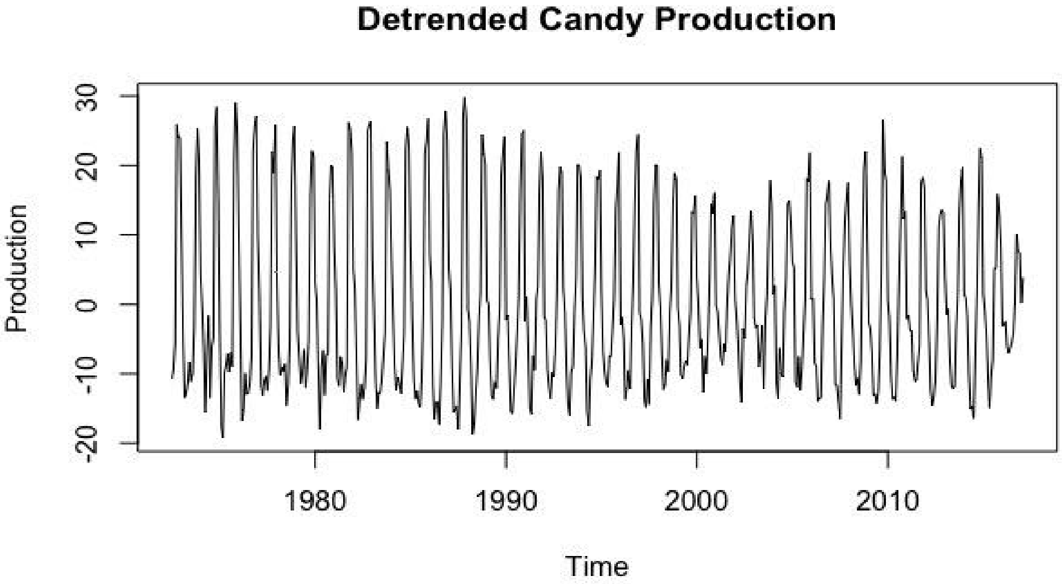
Figure 5. *Trend plot with Linear Model*





An Augmented Dickey-Fuller (ADF) test was on the detrended seasonal data and found the p-value to be 0.01, indicating that there is enough evidence to reject the null hypothesis and conclude that the detrended time series data is stationary. The stationary condition is met.

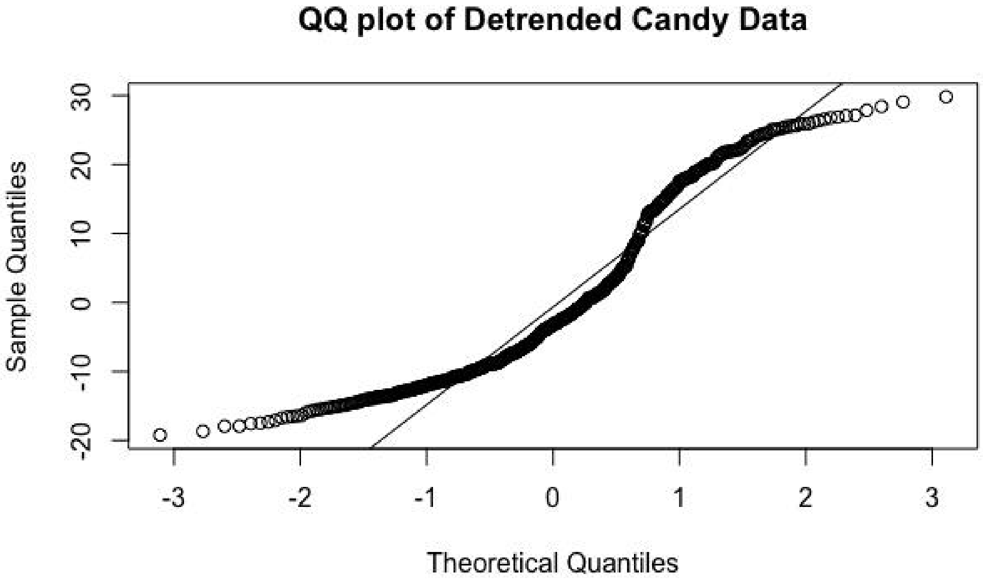
Figure 6. Detrended Candy Production plot



4.2.1 Residual Analysis

In *Figure 7* the Q-Q plot indicates the detrended data is not normal as indicated by the heavy tails in the Q-Q plot. The normality condition is not met.

Figure 7. Q-Q plot of Detrended U.S. Candy Production Data



4.2.2. Conclusion

A proper model ws not found by detrending the time series as the residuals were not normally distributed, leaving crucial information in the errors. When the data was differenced, the same residual analysis conclusion was found for the differenced data, and for the log of differenced data (see *Appendix 7.3.1*). Therefore the differenced log data is not included in this study since it did not offer any new information.

4.3 Deterministic trend - Seasonal means model

A seasonal means model was fit by setting the season function to the monthly data and fitting a linear model to the detrended seasonal data.

Table 1. Seasonal Means Statistics

Seasonal Means Statistics		
p-value	< 2.2e-16	Model is statistically significant
R^2	0.89	Explains 89% of Seasonality

The stationarity condition has already been met for the detrended data so the model can proceed with residual analysis. The residuals are tested to confirm whether the p-value found in *Table 1* for the seasonal means trend model is reliable or not. Supplementary R codes for the seasonal means deterministic model can be found in *Appendix 7.3.2*.

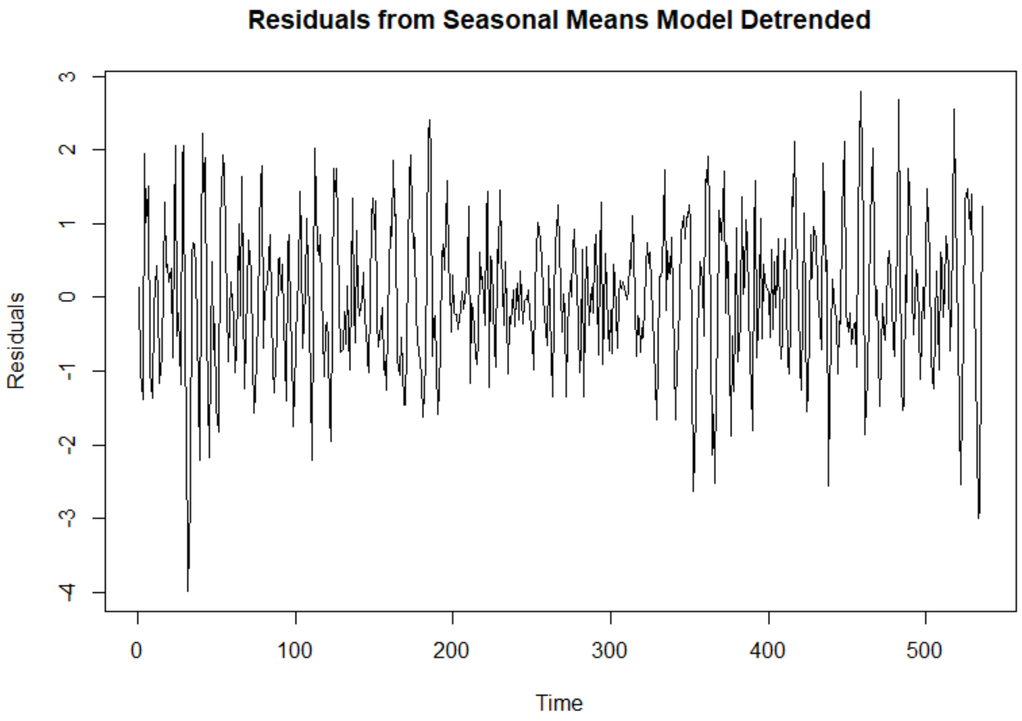
4.3.1 Residual Analysis

Residual analysis was conducted to observe whether or not the data satisfies the conditions of zero mean, homoscedasticity, independence, and normality,

4.3.2 Zero mean and homoscedasticity

The residual plot in *Figure 8* shows a random fluctuation around the mean zero with no pattern, however, the variance is slightly more volatile than should occur for a normal distribution, as the residuals often oscillate between 3 and -4, indicating possible outliers. The zero mean and homoscedasticity condition is met, but the formal normality test will be necessary.

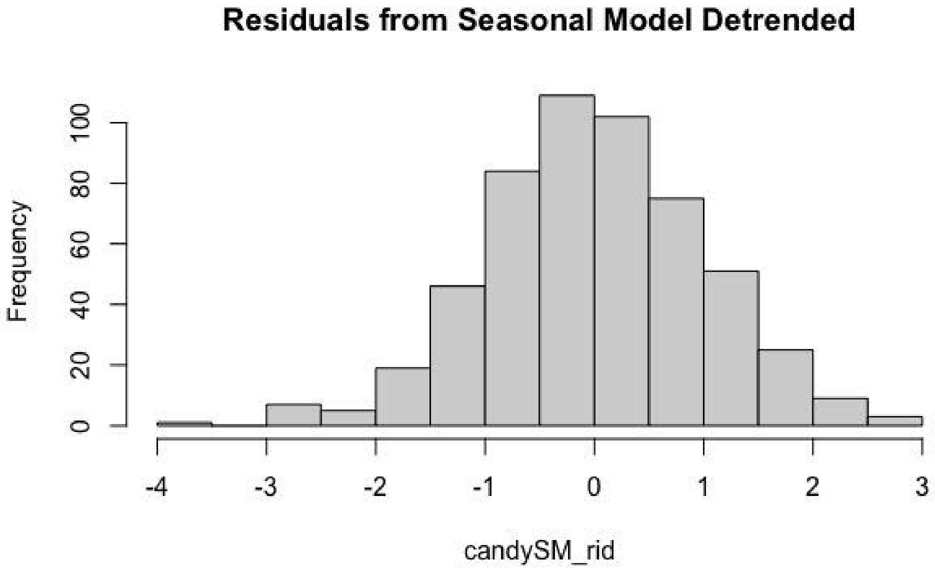
Figure 8. Residual Plot of Detrended Seasonal Means Model



4.3.3 Normality

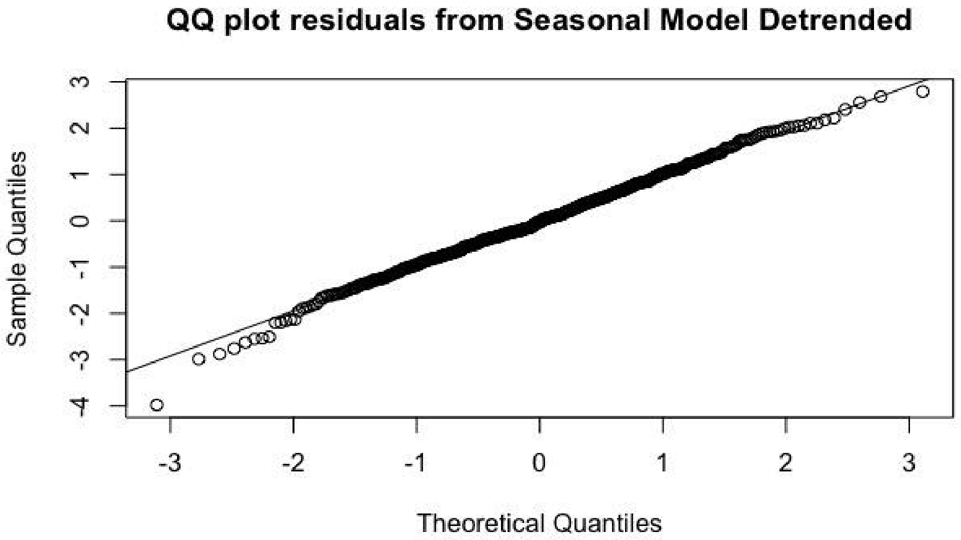
In *Figure 9* the histogram indicates the outliers that were observed in the residual plot, however, the histogram is otherwise a nice bell-shape, indicating possible normal distribution in the residuals.

Figure 9. Histogram of residuals of Detrended Seasonal Means Model



In *Figure 10* the Q-Q plot appears fairly straight, with some outliers in the tails, particularly in the lower tail.

Figure 10. Q-Q plot of residuals of Detrended Seasonal Means Model



As a final test for normality, the formal Shapiro-Wilk test showed a P-value of 0.2281, so the null hypothesis is retained that residuals are normally distributed. The normality condition is met.

4.3.4. Independence

A Runs test is performed to look for randomness (independence) in the data. In *Table 2* the observed runs are far lower than should occur for an independent set of data. The p-value confirms that there is statistically significant evidence to reject the null hypothesis that the data is independent. The independence condition is not met.

Table 2. Runs test of residuals of Detrended Seasonal Means Model

P-value	**Observed runs**	**Expected runs**
1.52e-11	191	268

4.3.5. Seasonal means model conclusion

The seasonal means model cannot be used since the residuals are not totally random. There is still data left in the residuals that needs to be removed. Therefore, the p-value that was observed in the model is not reliable. The analysis continues with another deterministic trend model.

4.4. Deterministic trend - Cosine trend model

A cosine model is fit by setting the harmonic function to the detrended monthly data and fitting a linear model to the harmonic data. Supplementary R codes for cosine trend deterministic model can be found in *Appendix 7.3.3.*

Table 3. Cosine model statistics

Cosine model statistics		
p-value	< 2.2e-16	Model is statistically significant
R^2	0.78	Explains 78% of Seasonality

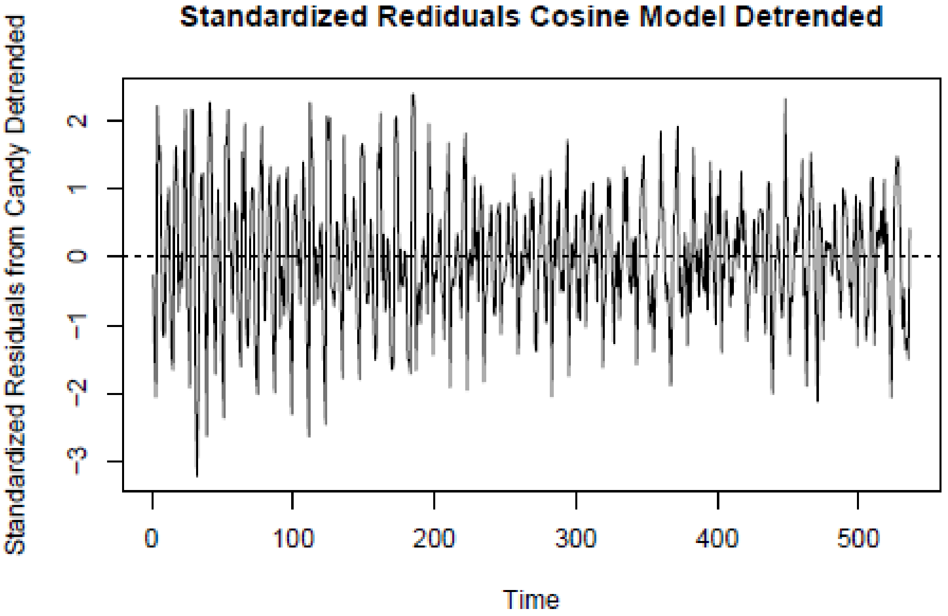
4.4.1 Residual analysis

A residual analysis was conducted to observe whether or not the data satisfies the conditions of zero mean, homoscedasticity, independence, and normality.

4.4.2. Zero mean and homoscedasticity

In *Figure 11* the residual plot has a random fluctuation around the mean zero, with no patterns visible. The residuals fluctuate between 2 and -3, with most of the errors between 2 and -2, which follows the normal distribution. The zero mean and homoscedasticity condition is met.

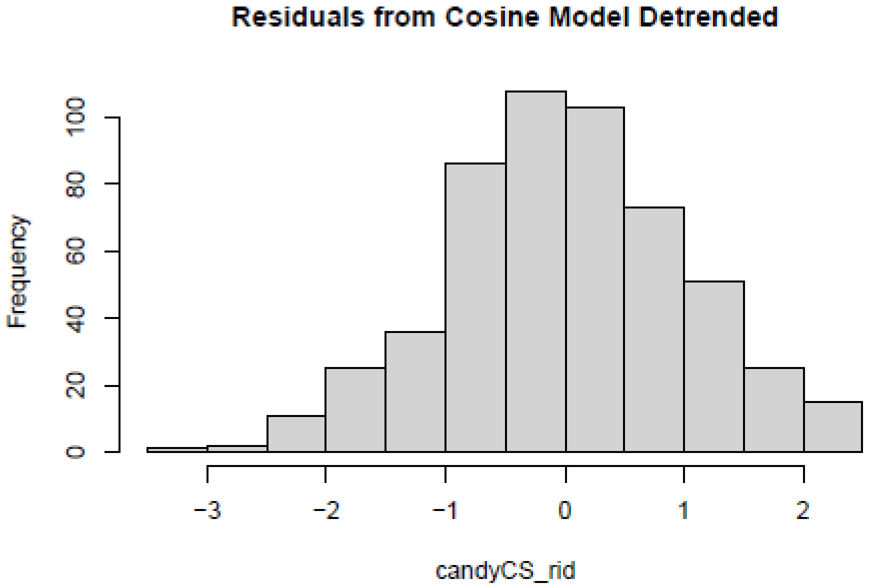
Figure 11. Residual plot of Detrended Cosine Model



4.4.3 Normality

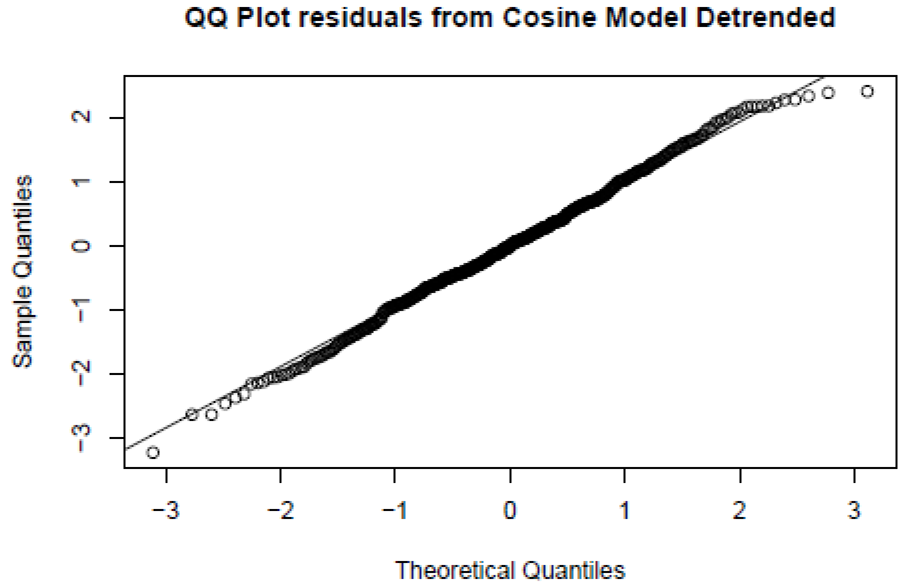
In *Figure 12* the histogram of residuals appear fairly normal, good bell-shaped distribution. Further residual analysis is needed by using the Shaprio-Wilk test for normality.

Figure 12. Histogram plot of Detrended Cosine Model



In *Figure 13* the Q-Q plot has a few outliers, but otherwise, has a straight line with no heavily skewed tails, indicating normality. The upper tail could potentially be an issue.

Figure 13. Q-Q plot of Detrended Cosine Model



The formal Shapiro-Wilk test returns a value of 0.2954, indicating the null hypothesis cannot be rejected, and that the residuals are normally distributed. The small issues observed in the plots were not significant enough to reject the null hypothesis for normality. Therefore, the normality condition is met.

4.4.4. Independence

In *Table 4* the Runs test for independence shows that there is significant evidence to reject the null hypothesis that the residuals are independent. The independence condition is not met.

Table 4. Residual Runs test results

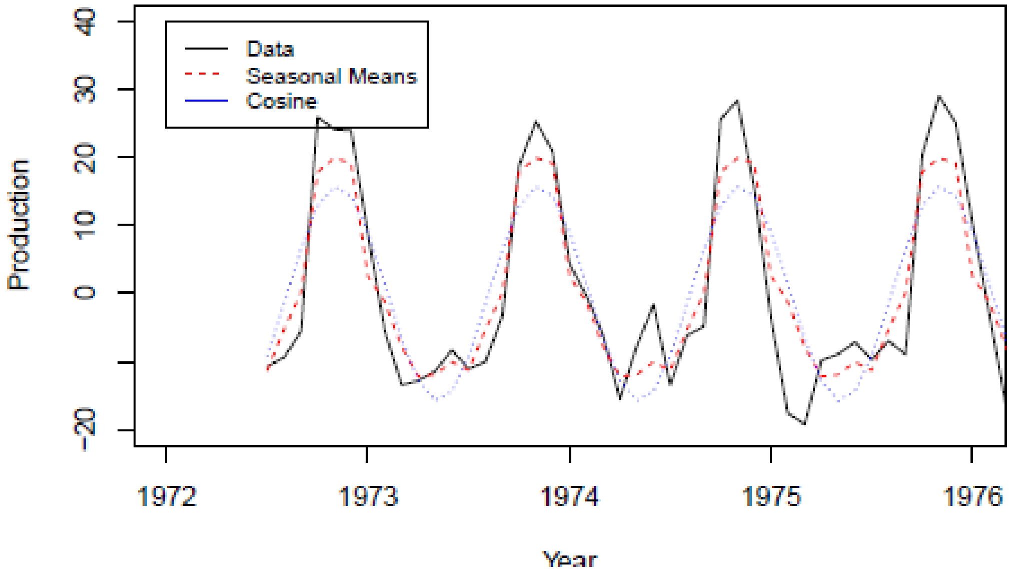
P-value	Observed runs	Expected runs
5.7e-08	206	268

4.4.5. Cosine trend model conclusion

There is still data left in the residuals, so the cosine model did not remove all the dependence from the residuals. While the residuals in the cosine model were slightly less dependent, it was still not the ideal model for the data. The comparison plot in *Figure 14, Detrended comparison plot: Seasonal Means model, Cosine Model, Real data* shows the cosine model was less seasonal than the seasonal means plot. In *Figure 5, Trend plot with Linear Model,* the plot appears to be a stochastic trend. Since deterministic and stochastic models did not fit the data, a stochastic

Seasonal ARIMA model may be more appropriate and will be tested next.

Figure 14. Detrended comparison plot: Seasonal Means model, Cosine Model, Real data



4.5. Choosing Seasonal ARIMA model using auto.arima()

4.5.1. Model determination

Using the auto.arima function on the detrended data, the auto.arima function returns a multiplicative seasonal ARIMA model, ARIMA(0,0,2)(1,1,2)[12] with drift. A seasonal ARIMA model is classified as an ARIMA(p,d,q)x(P,D,Q) model, where P=number of seasonal autoregressive terms, D=number of seasonal differences, Q=number of seasonal moving average terms. Supplementary R codes for the seasonal ARIMA model using auto.arima() can be found in *Appendix 7.3.4.*

4.5.2 Choosing order of integration

The auto.arima() function returns a seasonal ARIMA model with D=1 and d=0. Observing the ACFs of the detrended data in *Figure 15*, there is evidence of seasonality in the lags. The ACF supports the data that U.S. candy production has about a 6-month seasonality and a seasonal unit root, and supports the seasonal differencing, D=1. An Augmented Dickey Fuller test (ADF test), KPSS test, and Phillips-Perron test (PP test) can be used here to find evidence to support the stationarity of the ARIMA model.

Figure 15. Detrended Data ACF plot

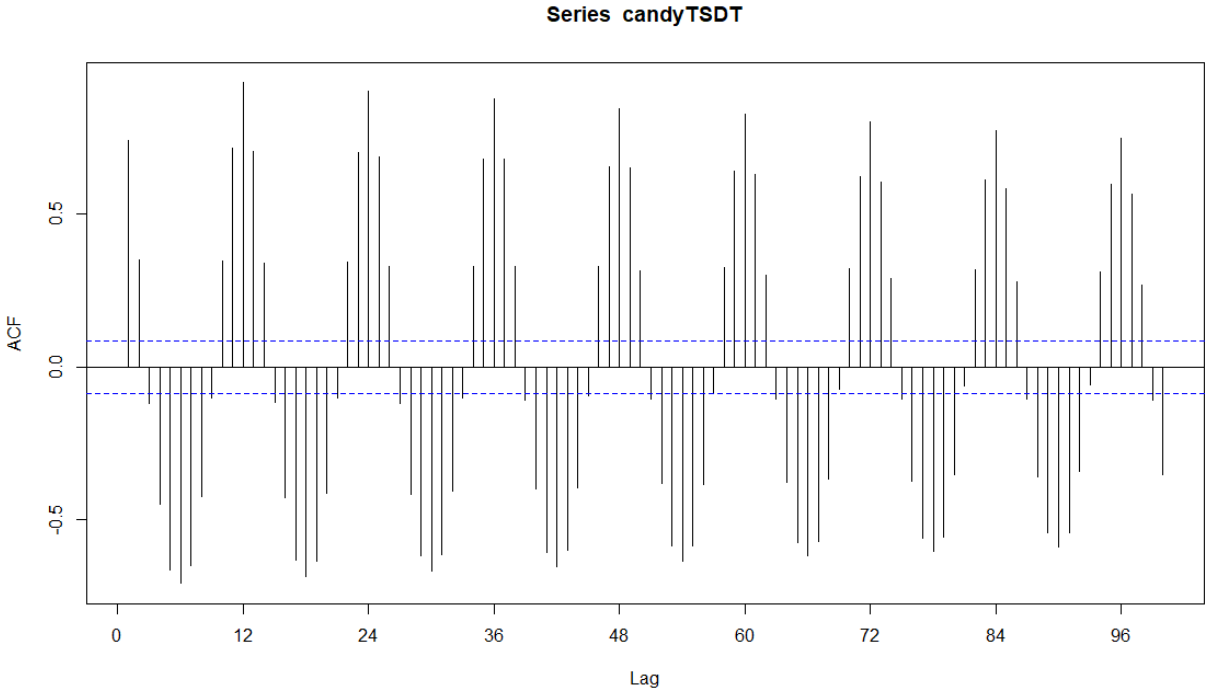


Table 5. P-values for ARIMA stationarity tests

ADF test	PP test	KPSS test
0.0164	-0.01	0.1

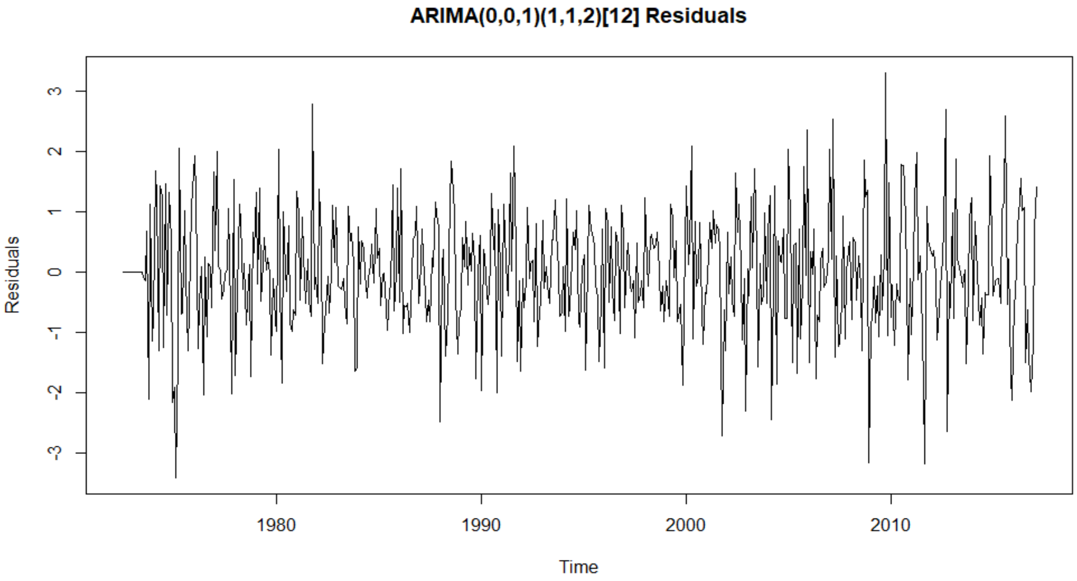
The Augmented Dickey-Fuller test (ADF test) on the time series data returns a value of 0.0164, the PP test returns a value of -0.01, and KPSS test returns 0.1. The null hypothesis for ADF and PP tests is that the data is nonstationary, so there is significant evidence to reject both hypotheses and conclude the data are stationary. The null hypothesis for the KPSS test is the time series data is stationary. Therefore, d=0 can be used as the order of integration for the ARIMA model.

4.5.3. Residual analysis

4.5.4. Zero mean and homoscedasticity

In *Figure 16* the residuals plot a random distribution and fluctuation around the zero mean and are evenly distributed between -3 and 3. There are no patterns or concerning volatility.

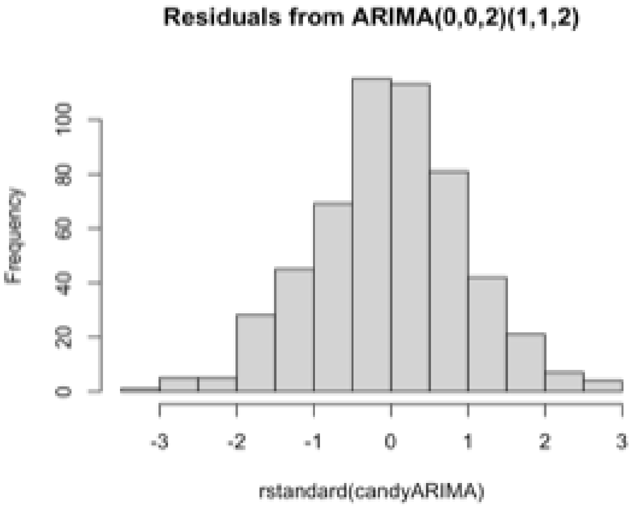
Figure 16. Residual plot of Seasonal ARIMA model



4.5.5. Normality

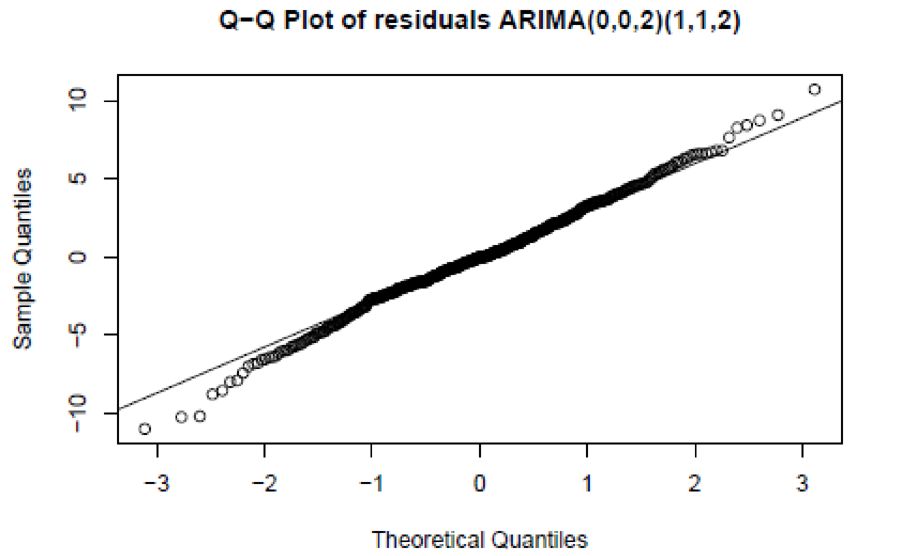
In *Figure 17* the histogram appears normal with outliers, but otherwise the residuals appear normally distributed.

Figure 17. Histogram plot of Seasonal ARIMA model



In *Figure 18* the Q-Q plot has a problem in the tails, indicating a possible problem with outliers in the data.

Figure 18. Q-Q plot of Seasonal ARIMA model



The formal test for normality, the Shapiro-wilk test, returns a value of 0.08244, indicating that there is not significant evidence to reject the null hypothesis, therefore, the residuals are normal. The normality condition is met.

4.5.6 Independence

The Runs test for this seasonal ARIMA model, however, returns a value of 0.0233, therefore, the residuals are not independent.

4.5.7. Seasonal ARIMA model conclusion

The auto.arima() function returned a good model but left time dependence in the model. The auto.arima() function provides a good basis to find an appropriate model for the data which will be explored in the next section.

4.6 Choosing Seasonal ARIMA model using information criteria

In order to find the best model, every ARIMA(p,d,q)*(P,D,Q) for p,d,q,P,D,Q in 0:2 was searched for a lower AIC than the previous model by creating a loop in R. The loop finds that an ARIMA (2, 0, 2) (1, 1, 2) [12] model is the best model for the data. *Table 6* shows the AIC comparison between the previous seasonal ARIMA model and the seasonal ARIMA found using a loop search in R; the AIC has been minimized from the previous seasonal ARIMA model.

Supplementary R codes for the seasonal ARIMA model using information criteria can be found in *Appendix 7.3.5.*

Table 6. Seasonal ARIMA Information Criteria

	AIC
ARIMA(0,0,2)(1,1,2)[12]	2738.47
ARIMA (2, 0, 2) (1, 1, 2)[12]	2629.62

Since the ACI is not a reliable model for D, the new seasonal ARIMA model will use d=0 and D=1, which is consistent with the previous seasonal ARIMA model using the auto.arima() function and consistent with the ACF plots. See *Section 4.5.2 Choosing order of integration* for more information.

The parameters found from the ARIMA (2, 0, 2) (1, 1, 2) [12] are consistent with the model chosen. The parameters in *Table 7* show that all the coefficients are significantly different from zero at a 95% confidence level, furthermore, the MA(q) and SMA(q) parameters are statistically significantly different from 1, therefore, the model does not appear to be overdifferenced.

Table 7. Seasonal ARIMA parameters

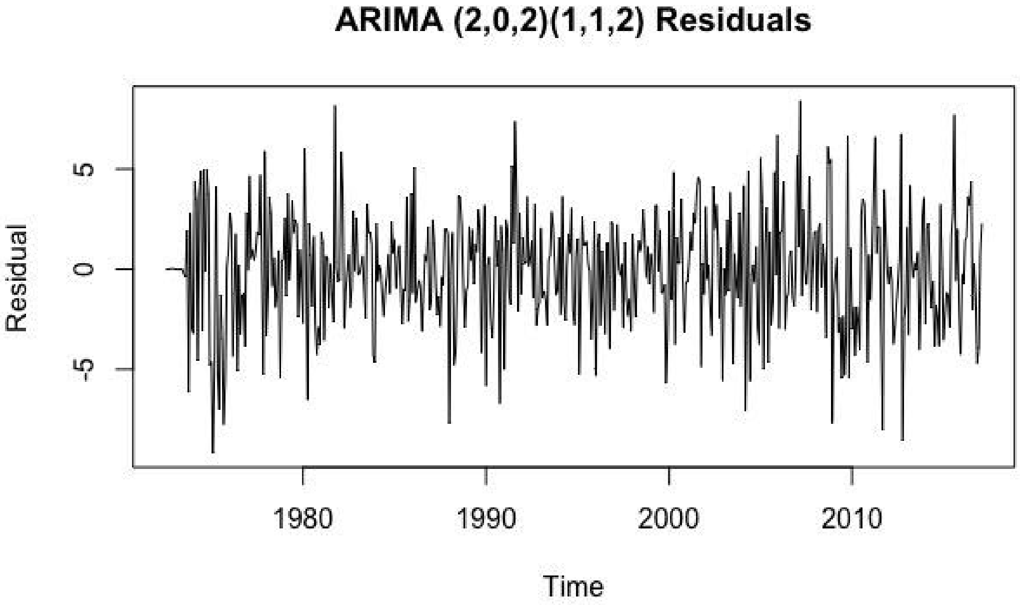
Coefficients						
	ar1	ar2	ma1	ma2	sma1	sma2
	1.4078	-0.6312	-1.2100	0.2100	-0.6622	-0.1009
s.e.	0.0653	0.0565	0.0829	0.0827	0.0455	0.0443

4.6.1 Residual Analysis

4.6.2 Zero mean and homoscedasticity

In *Figure 19* the residuals are evenly and randomly distributed around the mean zero. The zero mean and homoscedasticity condition is met.

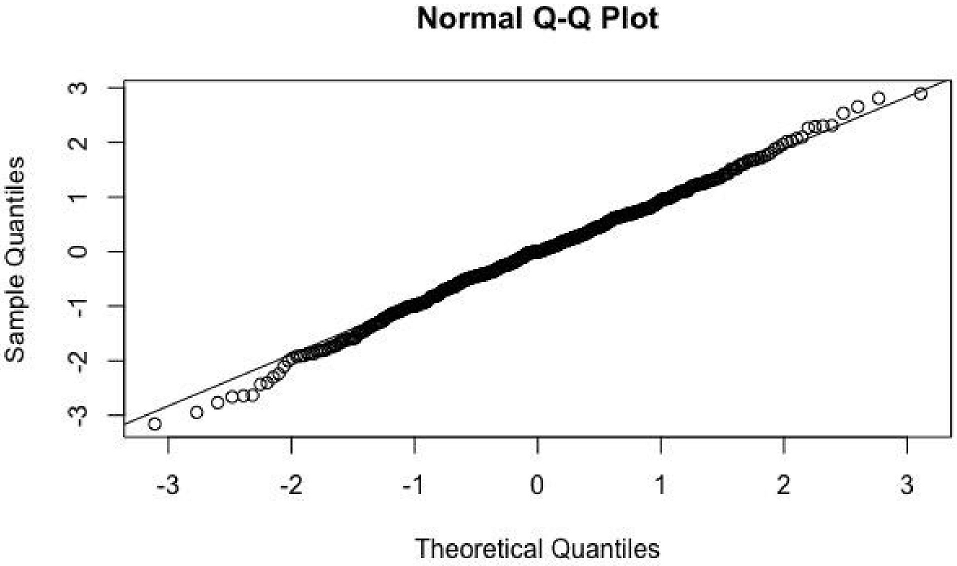
Figure 19. Residual plot of Seasonal ARIMA model



4.6.3 Normality

In *Figure 20* the Q-Q plot is fairly straight with no outstanding issues or outliers in the tails.

Figure 20. Q-Q plot of Seasonal ARIMA model



The Shapiro-wilk formal test for normality returns a p-value of 0.4989, indicating the residuals are normally distributed. The normality condition is met.

4.6.4 Independence

The runs test returns a value of 0.697, so the null hypothesis is retained; the residuals are independent. The independence condition is met.

4.6.5. Seasonal ARIMA model conclusion

The conditions for residuals are met using the ARIMA (2, 0, 2) (1, 1, 2) [12] model. This model can be used to predict future candy production without leaving valuable data in the residuals.

5. Conclusion

The aim of this study was to find if there is any seasonal pattern in candy production data, and can it be predicted? There is distinct seasonality present in the model, as evidenced by the ACF plots, which required a seasonal ARIMA model with d=0 and D=1, and a seasonal lag of 12. The seasonality is likely due to higher candy consumption during the U.S. winter holiday months (October- December), and lower consumption during the spring and summer months (April-September). Overlaying the seasonal ARIMA model on the original data shows that the ARIMA model found using the minimized information criteria functions is a very close and accurate model to the original data. See *Figure 21*. To see the accuracy closely, the last 10 years of U.S. candy production data is overlaid on the seasonal ARIMA model to show how close the model is to the original data (*Figure 22*). Because this model is a very good fit of the true model, the forecast should also be quite accurate. *Figure 23* shows the prediction of U.S. Candy Production for the following two years, 2018 and 2019, with a +/- 2 standard errors, or a 95% confidence interval. The seasonal pattern was accurately modeled using the seasonal ARIMA by minimizing the AIC, and was a sufficient model to predict future candy production.

Figure 21. Seasonal ARIMA model predicted vs. actual, 1972-2017

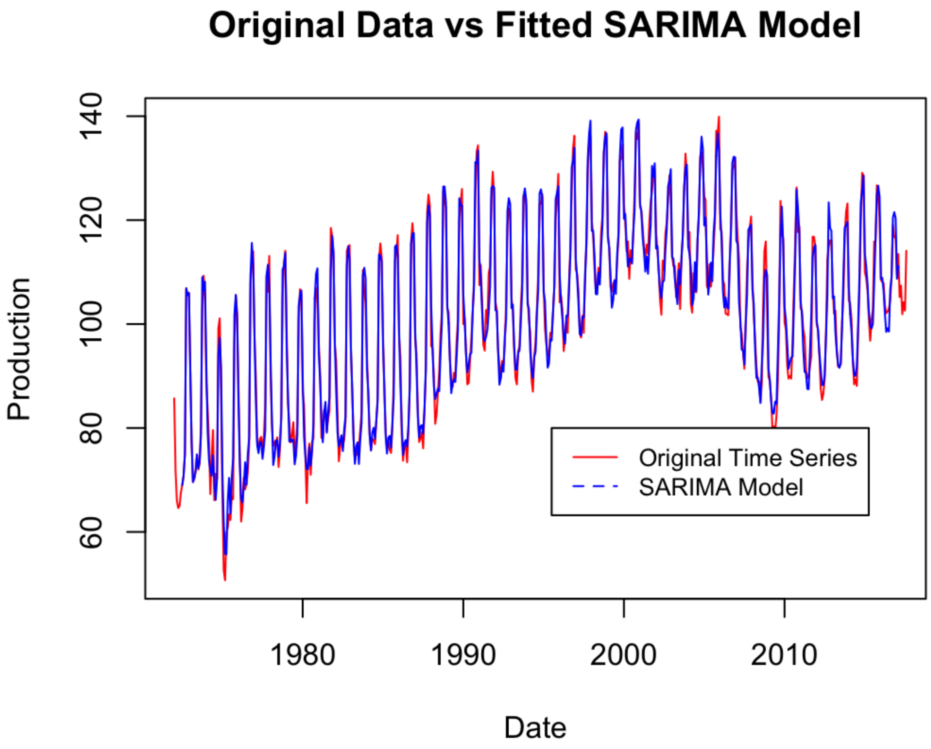


Figure 22. Seasonal ARIMA model predicted vs. actual, 2000-2010

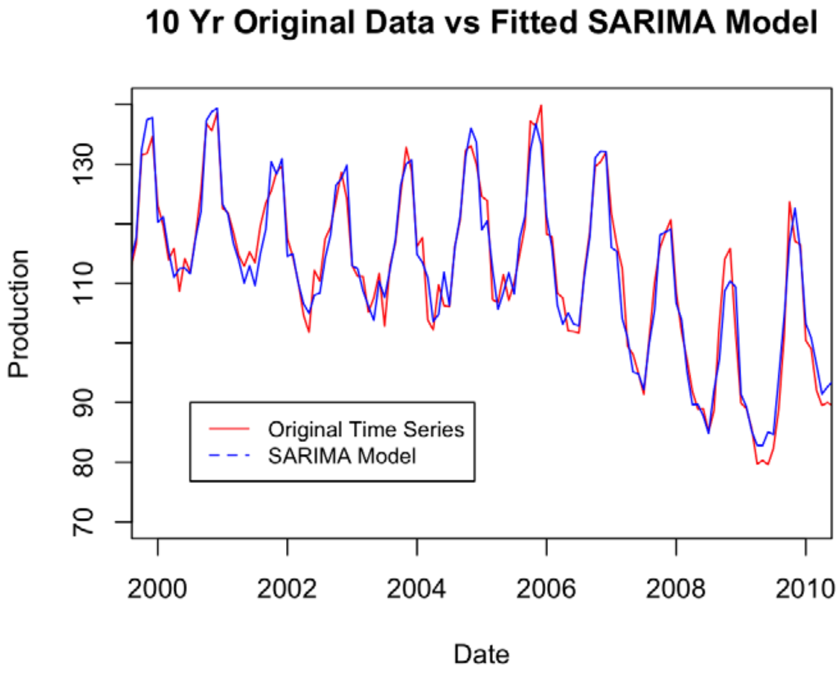
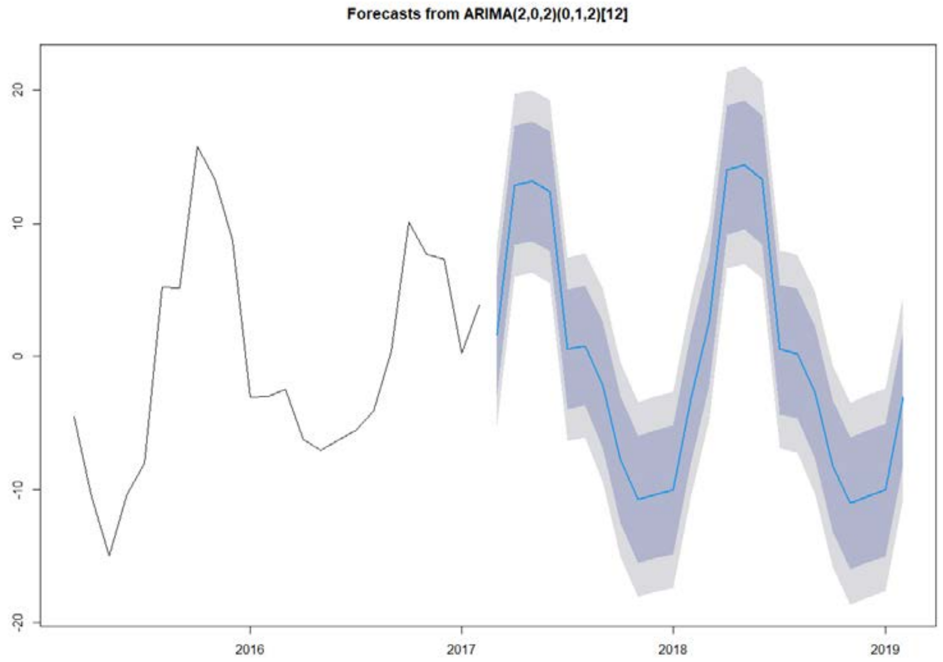


Figure 21. Forecast of Seasonal ARIMA model with 95% C.I., 2018-2019



6. References

[1] Cryer and Chan, "Time Series Analysis with Applications in R," 2nd Edition. 2011. Springer.

[2] Industrial Production and Capacity Utilization - G.17. January 20, 2021. Board of Governors of the Federal Reserve System. June 25, 2020. https://www.federalreserve.gov/releases/g17/about.htm

[3] Tatman, Rachael. 2017. January 20, 2021.US Candy Production by Month From January 1972 to August 2017. Version 1. Kaggle. https://www.kaggle.com/rtatman/us-candy-production-by-month/metadata

[4] U.S. Candy Market Size & Share: Industry Analysis Report, 2018-2025. January 20, 2021. U.S. Candy Market Size & Share | Industry Analysis Report, 2018-2025. Jan, 2018. https://www.grandviewresearch.com/industry-analysis/us-candy-market

7. Appendix

7.1 List of Figures

7.2 List of Tables

7.3 R codes

```
###############################################################
#
# MTU 5781 Time Series Analysis
# Spring 2021, Final Project
# Lorenzo Gordon & Kasia Krueger
#
###############################################################

###############################################################
#
# Monthly U.S. Candy Production 1972-2017
# https://www.kaggle.com/rtatman/us-candy-production-by-month
#
###############################################################

library(TSA)
library(sandwich)
library(lmtest)
library(forecast)
library(tseries)

# Import the data file
candy = read.csv(file.choose())

# Convert to time series
candyTS = ts(round(candy$IPG3113N, digits=1), freq=12, start=c(1972,1))

# 1. Plot the data. From 7.1 How to construct a time series model

# Plot candyTS (first four years to see seasonality)
months = c("J", "F", "M", "A", "M", "J", "J", "A", "S", "O", "N", "D")
plot(candyTS, type="l", xlim=c(1972,1976), ylim=c(50,150), xlab="Year",
     ylab="Production", main="U.S. Monthly Candy Production 
Seasonality")
points(candyTS, xlim=c(1972,1976), pch = months, col = 1, cex=0.5)

# Plot candyTS (1972-2017)
plot(candyTS, type="l", ylim=c(50,150), xlab="Year",
     ylab="Production", main="U.S. Monthly Candy Production")

# Add moving average to plot
candyTrend = ma(candyTS, order = 12)
plot(candyTS, type="l", ylim=c(50,150), xlab="Year",
     ylab="Production", main="U.S. Monthly Candy Production")
lines(candyTrend)

# decompose candy time series
candyTScomponents <- decompose(candyTS)
candyTScomponents
plot(candyTScomponents)
plot(candyTScomponents$trend, ylab="Production", main="Trend for Candy 
Production")
plot(candyTScomponents$seasonal, ylab="Production", ylim=c(-20, 30),
main="Seasonality for Candy Production")
plot(candyTScomponents$random, ylab="Production", ylim=c(-20, 20),
main="Random component of Candy Production")
plot(candyTScomponents$trend+candyTScomponents$seasonal+candyTScomponen
ts$random,
     ylab="Production", main="Candy Trend + Seasonal + Random")

#
# data appears to have both a seasonal and overall trend
#

# Remove the trend from original data for later testing
candyTSDT = candyTScomponents$seasonal + candyTScomponents$random
candyTSDT = na.omit(candyTSDT)
plot(candyTSDT, type='l', ylab="Production", main="Detrended Candy 
Production")

# Plot seasonal difference
candyTS %>% diff(lag=12) %>% ggtsdisplay()
```

```
# acf function
Acf(candyTSDT, lag.max=20, plot=TRUE)

# adf test Augmented Dickey Fuller Test, p-value < 0.05 indicates the
TS is stationary
adf.test(candyTSDT)
# p < 0.01, data is stationary

# QQ plot of data
qqnorm(candyTSDT, main="QQ plot of Detrended Candy Data")
qqline(candyTSDT)
# detrended candyTS looks non-normal
```

7.3.1. Non-normal differenced data

```
# look at the first difference of the detrended data
plot(diff(candyTSDT), type='l', ylab="Production", main="First
Difference Detrended Candy Production")
Acf(diff(candyTSDT), lag.max=20, plot=TRUE)
adf.test(diff(candyTSDT))
qqnorm(diff(candyTSDT), main="QQ Plot of First Difference Detrended
Candy Data")
qqline(diff(candyTSDT))
# still non-normal

# Plot monthly menas
candym = matrix(candyTS, ncol=12, byrow=TRUE)
month.means = apply(candym, 2, mean)
plot(month.means, type="l", main="Monthly Means of Candy Production",
xlab="Month", ylab="Mean")
points(month.means, pch = months, col = 1, cex=0.9)

# Fit linear model to trend line
candyTrendLM = lm(candyTScomponents$trend ~
time(candyTScomponents$trend))
summary(candyTrendLM)

# plot candy trend with linear model (may need for forecasting)
```

```
plot(candyTScomponents$trend, ylab="Production", main="Trend for Candy
Production")
abline(candyTrendLM)

# Deterministic models
```

7.3.2. Seasonal Means

```
# Seasonal Means model on detrended data
plot(candyTSDT, type="l", xlim=c(1972,2017), ylim=c(-30,40),
xlab="Year",
     ylab="Production", main="Detrended Candy Production")

month = season(candyTSDT)
candySMwoi =  lm(candyTSDT ~ month - 1)   # fit without intercept
summary(candySMwoi)
# R^2 = .89

## compare the result with intercept
candySMwi =  lm(candyTSDT ~ month)  # fit with intercept
summary(candySMwi)

# 2. Diagnostics of the Residuals

# Residual plot (zero mean and homoskedasticity)
candySM_rid = rstandard(candySMwoi)  #stardardized residual
plot(candySM_rid, ylab="Residuals", xlab="Time",
     type="l",main="Residuals from Seasonal Means Model Detrended")
abline(h=0,lty=2)
# residuals look mean zero, some concern with variance

# Histogram of residuals
hist(candySM_rid, main="Residuals from Seasonal Model Detrended")
# looks normal, some outliers in the tails

# QQ plot of residuals
qqnorm(candySM_rid, main="QQ plot residuals from Seasonal Model
Detrended")
qqline(candySM_rid)
# small problem in the lower tail
# Shapiro-Wilk test for normality
shapiro.test(candySM_rid)
# p = 0.2281, normal

# Runs test
runs(candySM_rid)
# pval=1.52e-11, observed=191, expected=268, reject independence

# Sample ACF plot
Acf(candySM_rid, main="ACF plot from residuals Seasonal Means
Detrended")

# adf test Augmented Dickey Fuller Test, p-value < 0.05 indicates the
TS is stationary
adf.test(candySM_rid)
# p < 0.01, stationary
```

7.3.3 Cosine trend model

```
# Cosine model of detrended data
har = harmonic(candyTSDT, 1)
candyCS = lm(candyTSDT ~ har)
summary(candyCS)
# R^2 = .78

plot(ts(fitted(candyCS), freq=12, start=c(1972,1)), type="l", ylim=c(-
20,20), xlab="Year",
     ylab="Production", main="Cosine Model Detrended Candy Production")

# 2. Diagnostics of the Residuals

# Residual plot (zero mean and homoskedasticity)
candyCS_rid = rstandard(candyCS)  #standardized residual
plot(candyCS_rid, ylab="Standardized Residuals from Candy Detrended",
xlab="Time",
     type="l",main="Standardized Residuals Cosine Model Detrended")
abline(h=0,lty=2)
# mean zero, but problem with variance

# Histogram of residuals
hist(candyCS_rid, main="Residuals from Cosine Model Detrended")
```

```
# looks normal, some outliers in the lower tail

# QQ plot of residuals
qqnorm(candyCS_rid, main="QQ Plot residuals from Cosine Model
Detrended")
qqline(candyCS_rid)
# small problem in the upper tail, normal

# Shapiro-Wilk test for normality
shapiro.test(candyCS_rid)
# p = 0.2954, normal

# Runs test
runs(candyCS_rid)
# pvalue=5.7e-08, observed=206, expected=268, reject independence

# Sample ACF plot
Acf(candyCS_rid, main="ACF plot of residuals from Cosine Detrended")
# looks like time series, dependent data

# Compare the fitted trends with Seasonal Means model and Cosine Trends
plot(candyTSDT, type="l", xlim=c(1972,1976), ylim=c(-20,40),
xlab="Year",
     ylab="Production", main="Seasonal Model VS Cosine Model")
candySM_fit = ts(fitted(candySMwoi), freq=12, start=c(1972.5,1))
lines(candySM_fit, lty=2, col="red")
candyCS_fit = ts(fitted(candyCS), freq=12, start=c(1972.5,1))
lines(candyCS_fit, lty=3, col="blue")
legend(1972, 40, legend=c("Data", "Seasonal Means", "Cosine"),
       col=c("black", "red", "blue"), lty=1:2, cex=0.8)
```

7.3.4. Seasonal ARIMA model using auto.arima()

```
# Find SARIMA model stochastic model

# 4. Detrend the data
# see line 62
# 5. Choose the order of integration 'd'

# Plot a section of the raw data
months = c("J", "F", "M", "A", "M", "J", "J", "A", "S", "O", "N", "D")
```

```
plot(candyTS, type="l", xlim=c(1972,1978), xlab="Year",
     ylab="Production", main="Monthly Candy Production")
points(candyTS, xlim=c(1972,1982), pch = months, col = 1, cex=0.8)

# adf test Augmented Dickey Fuller Test, p-value < 0.05 indicates the
TS is stationary
adf.test(candyTS)
# p = 0.0164, stationary

# Phillips-Perron Unit Root Test
pp.test(candyTS)
# p < .01, no unit root

# Select d=0
# undifferenced data passed the adf and pp test.

# see https://otexts.com/fpp2/seasonal-arima.html
# Check auto.arima on detrended data
# Warning, this function takes several minutes
candyARIMA = auto.arima(candyTSDT)
summary(candyARIMA)
# auto.arima returned a multiplicative seasonal SRIMA model
# ARIMA(0,0,2)(1,1,2)[12] with drift
# AIC=2738.47   AICc=2738.69   BIC=2768.3

# ARIMA residuals analysis for default (0,0,2)(1,1,2)
candyARIMA = arima(candyTSDT, order=c(0,0,2),
seasonal=list(order=c(1,1,2), period=12))
summary(candyARIMA)
# AIC=2736.49

plot.ts(candyARIMA$residuals, ylab="Residual", main="ARIMA
(0,0,2)(1,1,2) Residuals")

Acf(candyARIMA$residuals, lag.max=24, main="ACF of the ARIMA Model
Residuals")
Pacf(candyARIMA$residuals, lag.max=24, plot=TRUE)
signif(acf(candyARIMA$residuals, plot="F")$acf[1:12],2)

hist(candyARIMA$residuals, main="Residuals from ARIMA(0,0,2)(1,1,2)")
```

```
# normal with outliers

qqnorm(candyARIMA$residuals, main="Q-Q Plot of residuals
ARIMA(0,0,2)(1,1,2)")
qqline(candyARIMA$residuals)
# problem in the tails

# Shapiro-Wilk test for normality, p-value > .05 = normal
shapiro.test(candyARIMA$residuals)
# p-value = 0.08, normal

# Runs test
runs(candyARIMA$residuals)
# pvalue=0.0233, observed=242, expected=268

# adf test, p-value < 0.05 indicates the TS is stationary
adf.test(candyARIMA$residuals)
# pvalue < 0.01, stationary

# Plot 24 month forecast
plot(forecast(candyARIMA),24)

summary(candyARIMA)
confint(candyARIMA)
```

7.3.5 Seasonal ARIMA model using information criteria

```
#
# try all ARIMA(p,d,q)(P,D,Q) for p,d,q,P,D,Q in 0:2
# take a coffee break!
minAIC = 9999999
for(p in 0:2) {
   for(d in 0:2) {
      for(q in 0:2) {
         for(P in 0:2) {
            for(D in 0:2) {
               for(Q in 0:2) {
                  tryCatch( {model = arima(x=candyTSDT, order=c(p, d,
q),
                     seasonal=list(order=c(P,D,Q),period=12))},
                     error=function(e){}, warning=function(w){})
                  print(paste("(", p, ",", d, ",", q,
                     ")(", P, ",", D, ",", Q, ")", model$aic))
                  if (!is.null(model$aic) && model$aic < minAIC) {
                     minAIC = model$aic
                     minAICp = p
                     minAICd = d
                     minAICq = q
                     minAICP = P
                     minAICD = D
                     minAICQ = Q
                  }
               }
            }
         }
      }
   }
}
print(paste("Min AIC of ", minAIC, "at (", minAICp, ",", minAICd, ",",
minAICq,
            ")(", minAICP, ",", minAICD, ",", minAICQ, ")"))

# order=c(2,0,2),seasonal=c(0,1,2) has lowest AIC
# AIC = 2629.62

# ARIMA residuals analysis for default (2,0,2)(0,1,2)
candyARIMA = arima(candyTSDT, order=c(2,0,2),
seasonal=list(order=c(0,1,2), period=12))
summary(candyARIMA)
# AIC=2629.62
# note: Arima and arima differ

plot.ts(candyARIMA$residuals, ylab="Residual", main="ARIMA
(2,0,2)(0,1,2) Residuals")
#Box.test(candyARIMA$residuals,lag=20, type="Ljung-Box") # H0 Model
fits

Acf(candyARIMA$residuals, lag.max=24, main="ACF of the ARIMA Model
Residuals")
Pacf(candyARIMA$residuals, lag.max=24, plot=TRUE)
signif(acf(candyARIMA$residuals, plot="F")$acf[1:12],2)
```

```
hist(rstandard(candyARIMA), main="Residuals from ARIMA(2,0,2)(0,1,2)")
# normal with outlier in lower tail

qqnorm(rstandard(candyARIMA), main="Q-Q Plot of residuals
ARIMA(2,0,2)(0,1,2)")
qqline(rstandard(candyARIMA))
# small problem in lower tail

# Shapiro-Wilk test for normality, p-value > .05 = normal
shapiro.test(candyARIMA$residuals)
# p-value = 0.5149 normal

# Runs test
runs(candyARIMA$residuals)
# p-value = 0.697, independent

# adf test, p-value < 0.05 indicates the TS is stationary
adf.test(candyARIMA$residuals)
# P-value < 0.01, stationary

# 24 month forecast
candyARIMA = Arima(candyTSDT, order=c(2,0,2),
seasonal=list(order=c(0,1,2), period=12))
plot(forecast(candyARIMA),24)
```