UNIVERSITY OF CALIFORNIA
SANTA BARBARA

DEPARTMENT OF PROBABILITY AND STATISTICS

PSTAT 174 FINAL PROJECT

TIME SERIES ANALYSIS:
MONTHLY U.S. CANDY PRODUCTION

Author:

Clayton Van Hovel
Derek Mahn
Dorsa Jenab
Kaitlyn Boyle
Nicolle Yaranga

Supervisor:

Sudeep Bapat

June 6, 2018

US Candy Production

TABLE OF CONTENTS

United States Candy Production

PSTAT 174 I BAPAT I SPRING 2018

Kaitlyn Boyle, Dorsa Jenab, Clayton Van Hovel, Derek Mahn, Nicolle Yaranga

ABSTRACT

Candy production can give us a glimpse of other aspects of our world such as economic or health effects and fulfill the curiosity of the candy connoisseur. Our groups goal for this project is to forecast monthly candy production in the United States based off of our data aging back to 1972 until 2017 (starting from January 1, 1972). We attempt a transformation and use differencing to remove the trend and seasonality within our original data. We chose a SARIMA(1,1,2) x $(0,1,1)_{12}$ model to represent our candy production data and use it to forecast a year of candy production, September, 2017 to August 1, 2018.

1 INTRODUCTION

For many Americans candy is an integral part of their diet, for others they would be unaffected by its disappearance. The amount of candy consumption is tightly related to candy production. The mass amounts of candy production in the United States can affect our society in many different ways such as economic and health trends. Lots of candy produced potentially means many new jobs but an increase in production could lead to an increase in consumption, causing declines in health throughout the nation. Being able to draw conclusions from this data and forecast future candy production will help us determine its effect on other factors that affect us more directly.

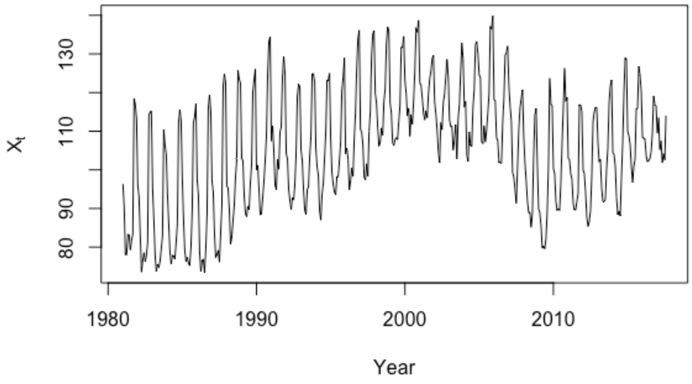
Our goal for this project is to create the most effective model possible to predict candy production for future times. The data spans from January of 1972 until August of 2017 with 441 observations and monthly candy production being the variable of interest. After exploring the initial data plots and observing immediate deviations, we decided to truncate the data to start from January 1st, 1981 to August 2017. We attempt to stabilize the variance and make the time-series stationary using a Box-Cox transformation. Next, we difference the data to remove seasonality and trend. At this point, we are able to make initial conjectures on the orders of models based off of the ACF and PACF plots of the data. We used the Akaike Information Criterion (AIC) to select a few models to further explore in diagnostic testing. Keeping the Principle of Parsimony in mind, we concluded a final model of SARIMA(1,1,2) x$(0,1,1)_{12}$ as it has fewer parameters. We used this model to forecast a year of candy production from September 1st, 2017 to August 1st, 2018 .

2 EXPLORATORY DATA ANALYSIS

2.1 Data Exploration

Our dataset consists of monthly Industrial Production Index (IPI) values from January 1981 to August 2017. The Industrial Production Index measures monthly real production output of candy manufacturers relative to the base year. We observed an average IPI of 105.1258, a minimum IPI of 73.4034 and a maximum IPI of 139.9153. On the side you can see a plot of all the observations.

Figure 1: Candy Production Time Series



There is an evident seasonality component as the data was collected monthly. There is also a trend present, gradually increasing up until around 2005 where it begins to fall but slowly picks up again around 2011. Before applying any type of transformation, note that the variance of our data is 245.1008.

Since our data has a seasonality, a trend and large variance, we can deduce that the time series data needs to be transformed and differenced.

Figure 2: ACF and PACF of Candy Production

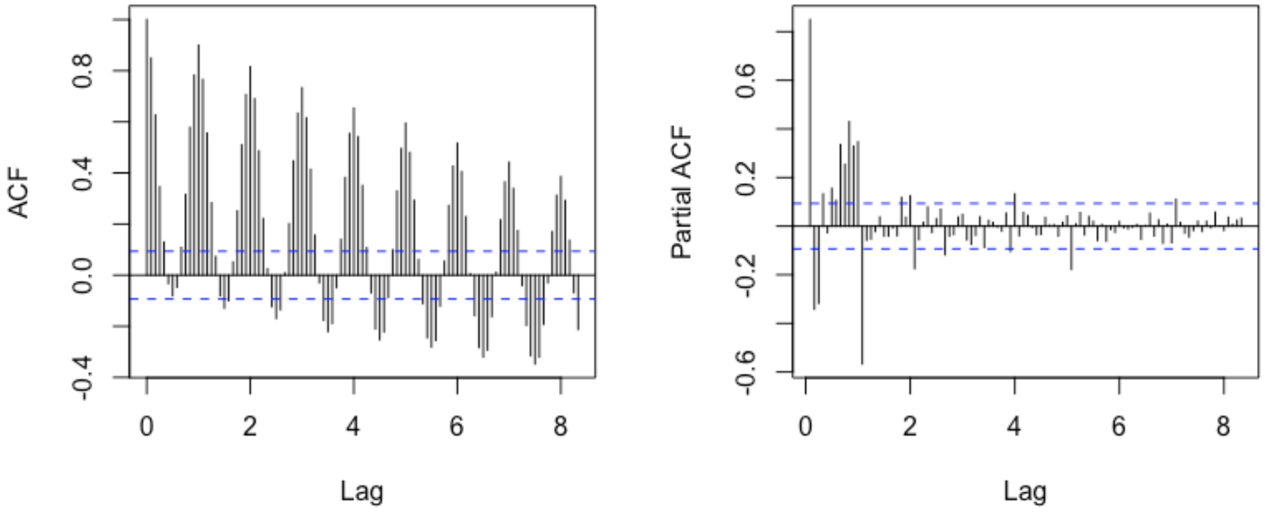


Figure 2 shows the ACF is cyclically decreasing geometrically, indicating that there is a seasonal component present in our data. The ACF tails off, implying an AR component, while the PACF shows a sharp spike at lag 1, further implying this AR component.

2.2 Box-Cox Transformation

In attempt to make the data stationary and stabilize the variance, we used a Box-Cox transformation. We performed the following calculation to the original data. X represents the original data:

$$y = \begin{cases} (x^{\lambda}) - 1/\lambda & when\ \lambda \neq 0 \\ \log(x) & when\ \lambda = 0 \end{cases}$$

Figure 3: Log-Likelihood of Box Cox transformation

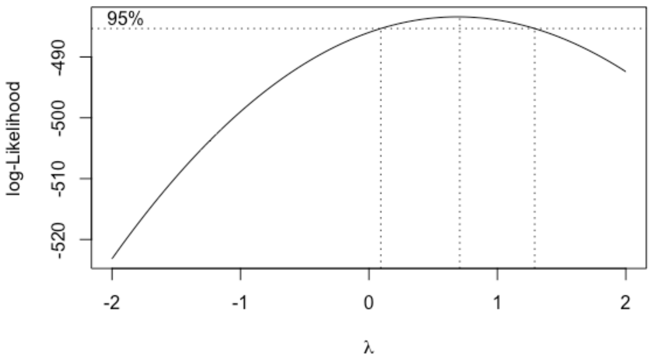
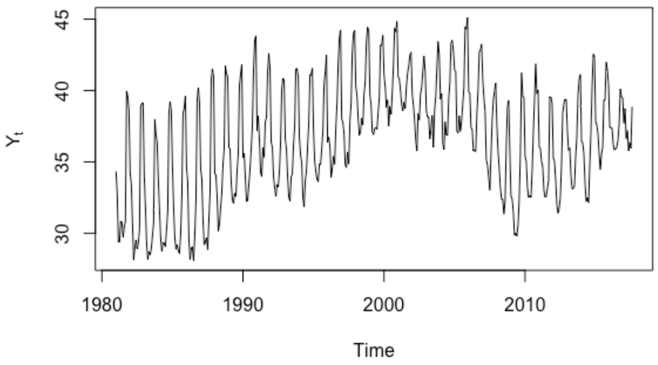
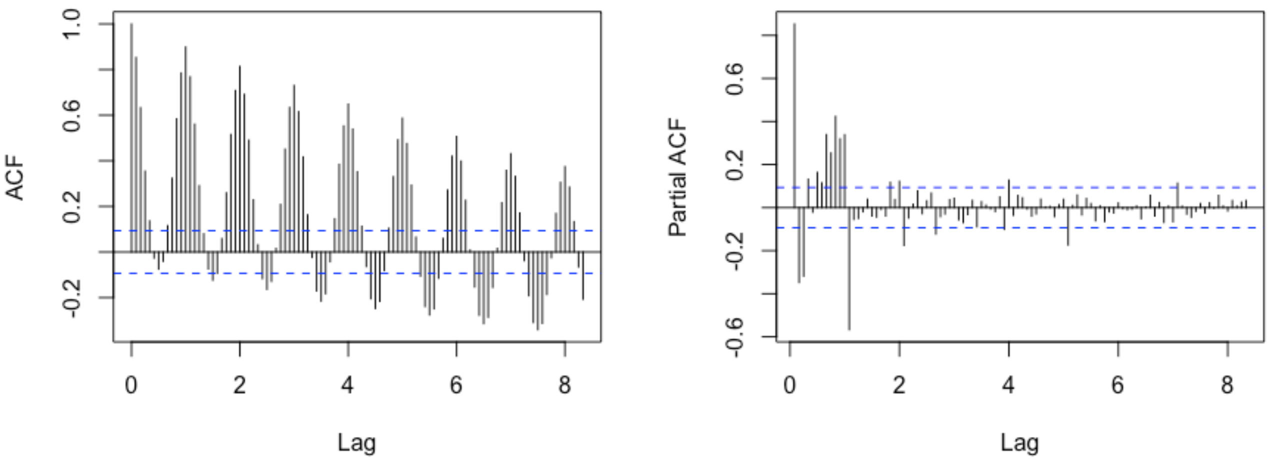


Figure 3: Box-Cox tranformed data



Y is the transformed data with λ as a parameter that is determined by the max log-likelihood. The optimal λ we found is 0.707071 however, $\lambda = 1$ is still included in the 95 percentile. Although the variance was reduced to 16.18151, a Box-Cox transformation is harder to interpret and its ACF and PACF are almost identical to that of the original , thus we chose to stick with the original data. (Reference to Figure 2 and 4).

Figure 4: ACF and PACF of Box-Cox Transformed Data



The slow decay of the ACF of Y reveals that Y is non stationary (reference Figure 3).

2.3 Differencing

2.3.1 De-Trending

We continue to perform differencing to remove the trend and seasonality. Figure 1 displays an obvious trend, so we difference the data at lag 1 in efforts to remove that trend. The variance was reduced from 245.1008 to 73.52247. We differenced once more but the variance increased to 111.0935, implying the data was over-differenced.

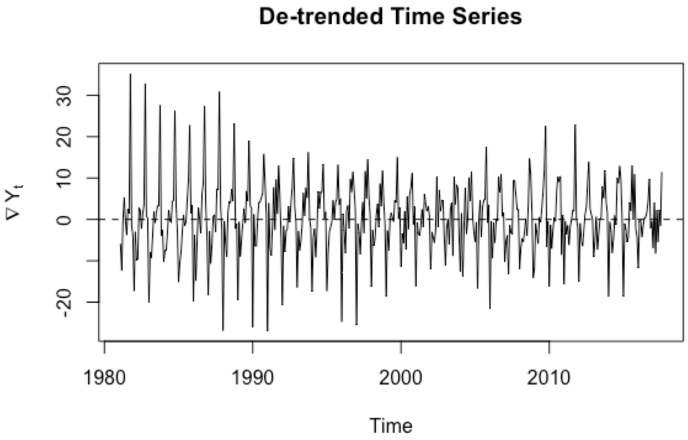
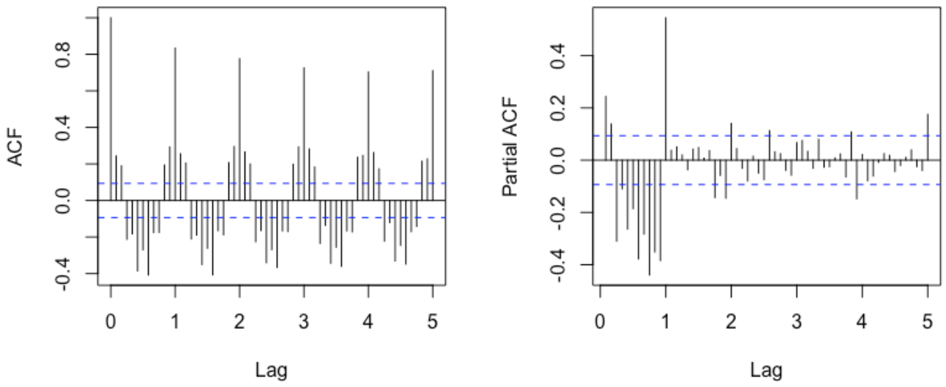


Figure 5: ACF and PACF of De-Trended Data



De-trending the data once did not change the ACF plot much. Figure 5 shows the ACF showing a geometrically decreasing pattern towards zero, but must less than it showed in Figure 3.

2.3.2 De-Seasonalizing

Since the ACF still has a cyclical pattern, we need to remove seasonality. Knowing the data was collected with a 12-month period, we differenced the original data at lag 12 to remove the seasonality. The variance decreased from 73.52247 to 19.87893.

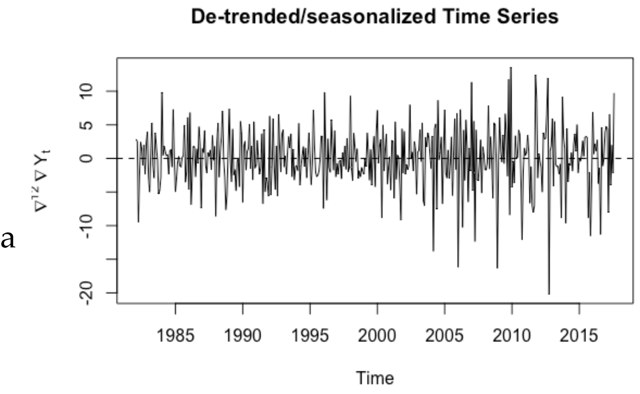
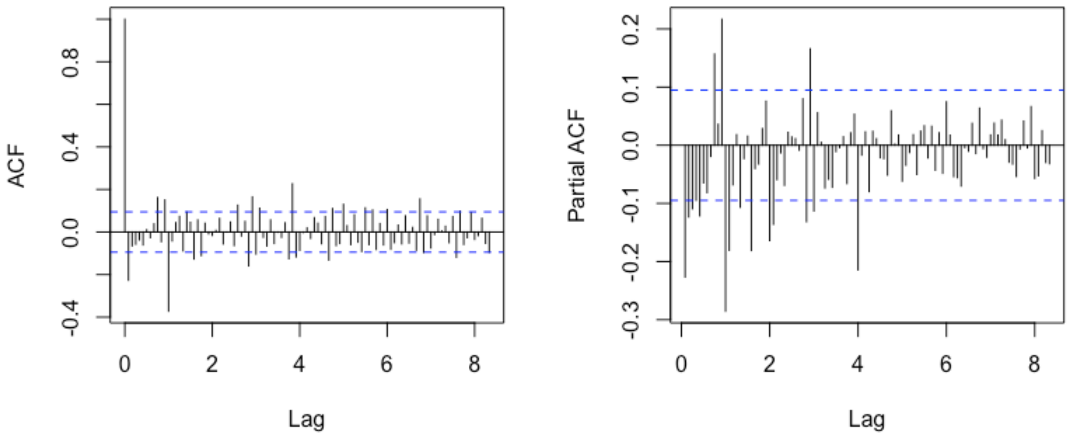


Figure 6: De-trended/seasonalized Time Series



Overall, we only differenced the data at lag 1 and again at lag 12.

$$Y_{12}= \nabla_{12}\nabla_{1}y_{t}$$

To confirm whether the transformed and differenced model is stationary, we applied the Augmented Dickey-Fuller Test where H_0 is that the data contains a unit root and H_A is the data is stationary. The test resulted in a significant p-value of less than 0.01. Thus, we rejected the null hypothesis and confirm with 95% confidence that the data is now stationary.

3 MODEL BUILDING

After removing the trend and seasonality to produce a stationary series, we can fit the data into a SARIMA model. A SARIMA model is illustrated by SARIMA(p,d,q)×(P,D,Q)$_s$ where p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and s = the length of the season. As the data was collected monthly and we de-seasonalized the data by differencing at lag 12, S = 12 and D =1. Since we differenced the de-seasonalized data only once before the variance began to increase, d =1. To find the order of the p,q ,and P, Q we use preliminary conjectures based off of the ACF and PACF of the data, followed by AIC comparisons to arrive to a final model.

3.1 Analyzing ACF and PACF

To determine the order of seasonal components (P and Q) we observe Figure 6 at lags 12, 24, 36, and so on. The ACF shows a prominent peak at lag 12 while the PACF tails off at lag 12. As a result, we are led to believe the MA order is 1 (Q =1) and the AR order is 0 (P=0). To determine the order of the non-seasonal components (p and q) we observe the plot at the lags within each season (1,2,...,11). The ACF seems to cut off after lag 1 while the PACF shows peaks at lags 1, 2, and 3. These graphs suggests several candidates for p and q. So we consider models for p = {0,1} and q = {1,2,3} which we will further explore through AIC comparisons.

3.2 Model Selection

Of these possible models, we compare the AIC, looking for the model with the lowest, most negative AIC.

	MA(0)	MA(1)	MA(2)	MA(3)
AR(0)	2491.367	2462.807	2456.272	2453.570
AR(1)	2470.596	2439.919	2435.851	2437.489
AR(2)	2465.978	2436.372	2436.976	2428.665
AR(3)	2462.804	2437.895	2440.406	2440.865

The lowest two AIC values were produced by the models ARMA(2,3) and ARMA(1,2). Because we have already de-trended and de-seasonalized, these AIC values are for the non-seasonal p, q. SARIMA(1,1,2) x (0,1,1)$_{12}$ as well as SARIMA(2,1,3) x (0,1,1)$_{12}$ produce AIC values that are lower than all other AIC values nearby them in the chart. The Principle of Parsimony led us to choose the SARIMA(1,1,2) x (0,1,1)$_{12}$ model as it has less parameters.

3.3 Model Fitting

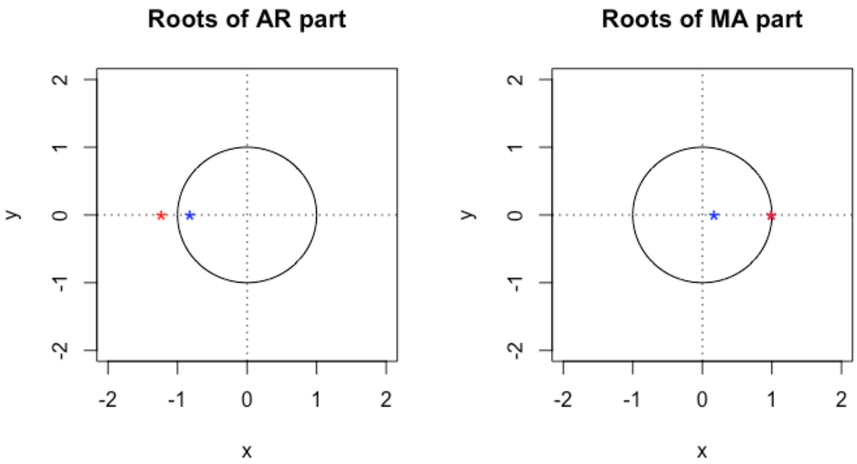
Let Model I be SARIMA(1,1,2) x (0,1,1)$_{12}$. The following table shows the coefficients of Model I.

	Model I
AR(1)	0.8146
AR(2)	—
MA(1)	-1.1693
MA(2)	-0.344

Our final model is illustrated by the following formula :

$$Y_t - 0.8146Y_{t-1} = Z_t - 1.1693Z_{t-1} + 0.1693Z_{t-2}$$

In order to confirm if the AR part is stationary and MA part is invertible, we plot the roots of both the AR and MA parts. The roots of the AR part lie outside of the unit circle, implying that the model is stationary. However, the root of the MA part lies exactly on 1, therefore our model is not invertible.

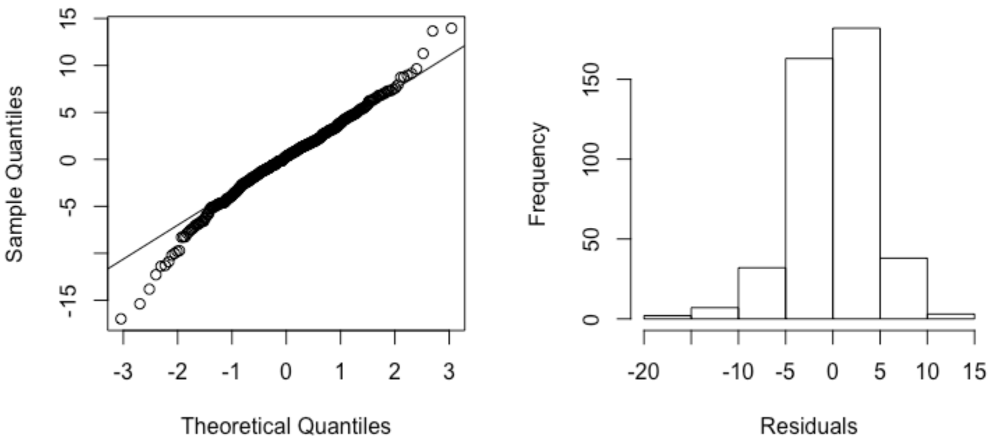


3.4 Diagnostics

3.4.1 Normality

We performed a series of diagnostic tests to check whether any of the assumptions of our SARIMA model are violated. A SARIMA model assumes the residuals are Gaussian White Noise with a mean of zero and a constant variance. A Q-Q plot can show us if the residuals follow a normal distribution. Our Q-Q plot exhibits a heavy-tailed pattern, meaning that towards the ends of the graph, points curve off towards the extremities. This implies that our data has a larger number of outliers than expected, therefore our residuals cannot be captured in a normal distribution. The histogram plot of our model also verifies our previous observation.

Figure 7: QQ Plot and Histogram



The Shapiro-Wilk Test resulted in a p-value of 4.873e-05, thus we reject the null hypothesis that the residuals are identically and independently normally distributed. Our data fails to pass the Shapiro-Wilk test.

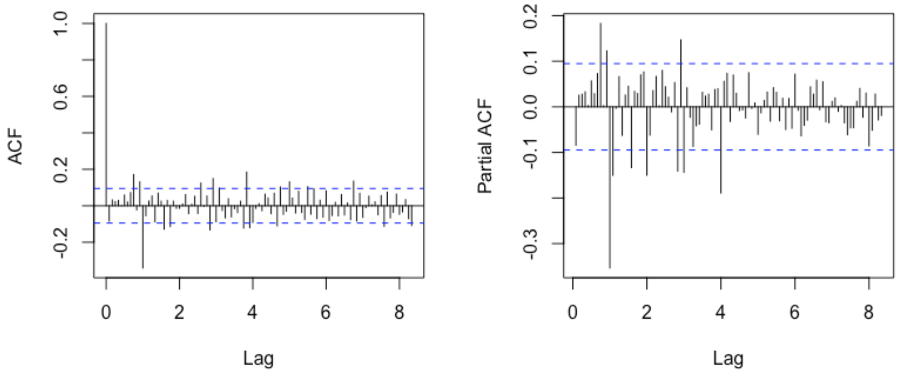
3.4.2 Independence

To check the independence of residuals, we performed Box-Pierce and Ljung-Box Test. Both the Box-Pierce and Ljung-Box test resulted in a significant p-value of 0.9386 and 0.9384 respectively, thus we reject the null hypothesis that there exists serial correlation, proving the residuals are independent and uncorrelated.

3.4.3 Constant Variance

To check for the homoscedasticity assumption, we plot the ACF and PACF of the residuals squared. Both the ACF and PACF plots fall within the 95% WN limits, except for a peak at lag 1. For convenience, we conclude that there is no heteroscedasticity present and the residuals have a constant variance.

Figure 8: ACF and PACF of Residuals



Collectively, we cannot confirm that our model's residuals follow an i.i.d. Gaussian distribution with a constant variance. Our model failed to pass the Shapiro-Wilk test, and the Q-Q plot & histogram showed signs of heavy-tailing. These are both indicative of non-normal residuals. The independence of the residuals was confirmed through the Box-Pierce and Ljung-Box. However, our ACF and PACF plots of the residuals did not completely fall within the 95% confidence intervals, therefore we cannot confidently say our model fulfills all assumptions.

4 FORECASTING

As the last point in our data set is 8/1/17, we forecasted the US candy production for the proceeding 12 months, 9/1/17-8/1/18. We can observe the 12 forecasted points in blue in figure 10, along with the 95% confidence intervals for these forecasted points. Because the uncertainty of the predictions increases with time, we see that the confidence interval widens towards the end of the forecasted points. This can be caused by the nonstationary aspect of our model. A stationary model with residuals that pass all diagnostic checks would produce a narrower, and therefore more precise confidence interval.

Figure 9: US Candy Production, Forecasted

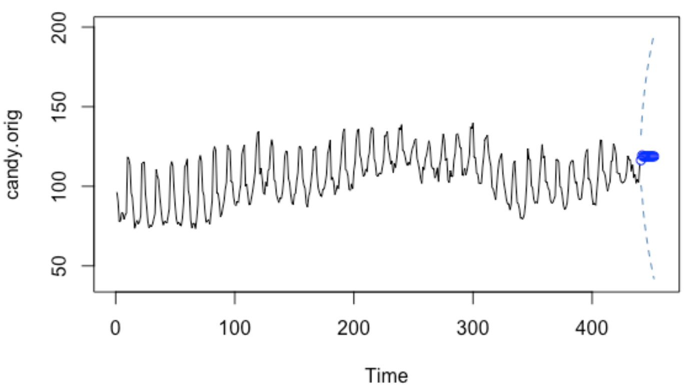
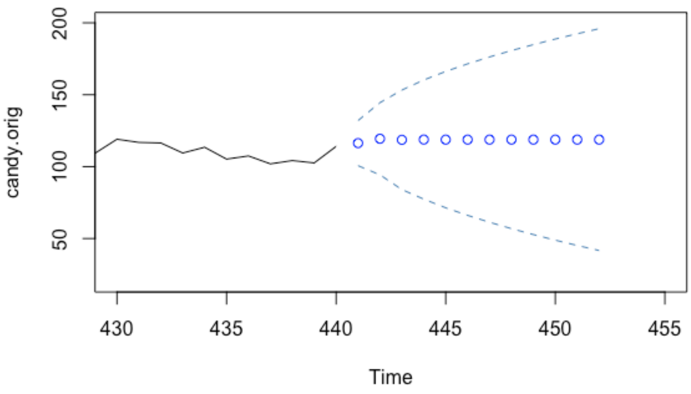


Figure 10: Twelve month Forecast of US Candy Production



5 CONCLUSION

Our first step in the analysis of candy production was to remove years 1972 - 1980 due to the heavy presence of outliers during that time period. This could have been due to many factors of the time period. In efforts to stabilize the variance we performed the Box-Cox transformation which we deemed inconclusive. Next, we differenced our data set in attempts to remove trend and seasonality. We differenced at lag 1 which reduced the variance of our data, however once we differenced an additional time the variance rose, so we limited ourselves to one difference at lag 1. To remove seasonality we differenced at lag 12 and concluded that the optimal amount of differencing for out data was once at lag 1 and once at lag 12.

The next step in our process was to build a model that accurately represents our data in hopes of forecasting future candy production. By analyzing the ACF and PACF of the data we brought our decision down to several models. By checking the AIC values of each model we selected a SARIMA(1,1,2) x $(0,1,1)_{12}$. By using the plot root function we found the coefficients for our model and determined that it is stationary but not invertible. By testing our data using several methods we concluded that it cannot be represented by a normal distribution and the residuals do not follow a gaussian white noise distribution which implies that our data would require an ARCH or GARCH model to optimally represent it. Lastly, we created a 95% confidence interval for candy production in 2018, however, because of the non-stationary aspect of our model our range is wider than a completely stationary model with residuals that pass all diagnostics checks would be.

6 REFERENCES

[1] "4.1 Seasonal ARIMA Models." 11.5 - Identifying Influential Data Points | STAT 501, newonlinecourses.science.psu.edu/stat510/node/67/.
We used this reference to learn about SARIMA models and model building for a SARIMA model.
[2] "ATSA." Function | R Documentation, www.rdocumentation.org/packages/aTSA/versions/3.1.2/topics/adf.test.
We learned about the Augmented Dickey Fuller test for stationarity and used this as a reference for our R code.

[3] Board of Governors of the Federal Reserve System (US). (2017). US Candy Production by Month. Retrieved from https://www.kaggle.com/rtatman/us-candy-production-by-month.
We researched Kaggle and decided on the dataset US Candy Production by Month because the dataset had sufficient observations to create a solid time series analysis.
[4] "Industry - Industrial Production - OECD Data." TheOECD, data.oecd.org/industry/industrial-production.htm.
We used this resource to learn about Industrial Production Index, since that is how our data values were measured.

7 APPENDEX

~~~
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
options(warn=-1)
```

```{r time series plot, echo=FALSE}
candy <- read.table("~/Desktop/candy_production.csv", sep = ",", header = FALSE,
skip = 1)
#head(candy)
candyts <- ts(candy[,2], start = c(1981, 1), frequency = 12)
ts.plot(candyts, xlab="Year", ylab=expression(X[t]), main="Figure 1: Candy
Production Time Series")
```

```{r acf and pacf of time series, echo=FALSE}
op = par(mfrow = c(1,2))
acf(candyts, lag.max=100, main="")
pacf(candyts, lag.max=100, main="")
title("Figure 2: ACF and PACF of Candy Production Data", line = -1, outer=TRUE)
par(op)
```

```{r boxcox transformation, echo=FALSE}
library(MASS)
t = 1:length(candyts)
fit = lm(candyts ~ t)
bcTransform = boxcox(candyts ~ t, plotit = TRUE)
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
candy.bc = (1/lambda)*(candyts^lambda-1)
title("Figure 3: Log-Likelihood of Box Cox transformation")
```

```{r BC Transformed data include=FALSE}
ts.plot(candy.bc,main = "Figure 3: Box-Cox tranformed data", ylab =
expression(Y[t]))
```

```{r acf and pacf of box-cox transformation, echo=FALSE}
var(candyts)
var(candy.bc)
op = par(mfrow = c(1,2))
acf(candy.bc, lag.max = 100, main="")
pacf(candy.bc, lag.max = 100, main="")
title("Figure 4: ACF and PACF of Box-Cox Transformed Data", line = -1, outer=TRUE)
par(op)
```

```{r differencing, echo=FALSE}
y1 = diff(candyts, 1)
plot(y1, main = "De-trended Time Series", ylab = expression(nabla~Y[t]))
abline(h = 0, lty = 2)
y12 = diff(y1, 12)
ts.plot(y12,main = "De-trended/seasonalized Time Series",ylab =
expression(nabla^{12}~nabla~Y[t]))
abline(h = 0,lty = 2)
```
```{r include=FALSE}
var(candyts)
var(y1)
~~~

~~~~
```

```{r differenced at lag 1 acf pacf, echo=FALSE}
#install.packages("tseries")
#library(tseries)
op = par(mfrow = c(1,2))
acf(y1,lag.max = 60,main = "")
pacf(y1,lag.max = 60,main = "")
title("Figure 5: ACF and PACF of De-Trended Data", line = -1, outer=TRUE)
par(op)
```

```{r variance analysis include=FALSE}
var(candyts)
var(y1)
y2 = diff(y1,1)
var(y2) # shows that we only need to difference once
var(y12)
```

```{r differenced at lag 12 acf pacf, echo=FALSE}
op = par(mfrow = c(1,2))
acf(y12,lag.max = 100,main = "")
pacf(y12,lag.max = 100,main = "")
title("Figure 6: De-trended/seasonalized Time Series",line = -1, outer=TRUE)
par(op)
```

```{r Augmented Dickey-Fuller echo=FALSE}
library(tseries)
adf.test(y12, k =12)
```

```{r model estimation, echo=FALSE}
fit_arma11 = arima(y12, order = c(1,0,1), method = "ML")
fit_arma12 = arima(y12, order = c(1,0,2), method = "ML")
fit_arma21 = arima(y12, order = c(2,0,1), method = "ML")
fit_arma23 = arima(y12, order = c(2,0,3), method = "ML")

library(Matrix)
library(robustbase)
library(rgl)
library(minpack.lm)
library(MASS)
library(qpcR)

aiccs <- matrix(NA, nr = 4, nc = 4)
for(p in 0:3)
{
  for(q in 0:3)
  {
    aiccs[p+1,q+1] = AICc(arima(y12, order = c(p,0,q), method="ML"))
  }
}

colnames(aiccs)<- c("MA(0)", "MA(1)", "MA(2)", "MA(3)")
rownames(aiccs)<- c("AR(0)", "AR(1)", "AR(2)", "AR(3)")
aiccs
min(aiccs)
```

```{r diagnostic checking,eval=FALSE, include=FALSE }
~~~~

~~~
plot(residuals(fit_arma11))
Box.test(residuals(fit_arma11), type = "Ljung")
Box.test(residuals(fit_arma11), type ="Box-Pierce")
shapiro.test(residuals(fit_arma11))
op = par(mfrow = c(1,2))
qqnorm(residuals(fit_arma11), main = "")
qqline(residuals(fit_arma11))
hist(residuals(fit_arma11),main="")
title("Figure 7: QQ Plot and Histogram",line = -1, outer=TRUE)
par(op)
```

```{r diagnostic checks echo=FALSE}
plot(residuals(fit_arma12))
Box.test(residuals(fit_arma12), type = "Ljung")
Box.test(residuals(fit_arma12), type ="Box-Pierce")
shapiro.test(residuals(fit_arma12))
op = par(mfrow = c(1,2))
qqnorm(residuals(fit_arma12), main = "")
qqline(residuals(fit_arma12))
hist(residuals(fit_arma12),main="")
title("Figure 7: QQ Plot and Histogram",line = -1, outer=TRUE)
par(op)
```

```{r diagnostic checks echo=FALSE}
plot(residuals(fit_arma23))
Box.test(residuals(fit_arma23), type = "Ljung")
Box.test(residuals(fit_arma23), type ="Box-Pierce")
shapiro.test(residuals(fit_arma23))
op = par(mfrow = c(1,2))
qqnorm(residuals(fit_arma23),main="")
qqline(residuals(fit_arma23))
hist(residuals(fit_arma23),main="")
title("Figure 8: QQ Plot and Histogram",line = -1, outer=TRUE)
par(op)
```

```{r  acf and pacf of residuals eval=FALSE, include=FALSE}
acf(residuals(fit_arma11),lag.max=100)
pacf(residuals(fit_arma11),lag.max=100)
```

```{r eval=FALSE, include=FALSE}
fit_arma11
```
```{r checking for invertibility/causality eval=FALSE, include=FALSE}
source('plot.roots.R.txt')
op = par(mfrow = c(1,2))
plot.roots(NULL,polyroot(c(.7387)), main="Roots of AR part")
plot.roots(NULL,polyroot(c(1)), main="Roots of MA part")
par(op)
```

```{r echo=FALSE}
fit_arma12
```
```{r checking for invertibility/causality echo=FALSE}
source('plot.roots.R.txt')
op = par(mfrow = c(1,2))
plot.roots(NULL,polyroot(c(1,.8146)), main="Roots of AR part")
plot.roots(NULL,polyroot(c(1,-1.1693,.1693)), main="Roots of MA part")
~~~

~~~
par(op)
```

```{r checking for invertibility/causality echo=FALSE}
fit_arma23
```
```{r echo=FALSE}
source('plot.roots.R.txt')
op = par(mfrow = c(1,2))
plot.roots(NULL,polyroot(c(1,-0.0958,0.7602)), main="Roots of AR part")
plot.roots(NULL,polyroot(c(1,-0.2247,-.9885,.2132)), main="Roots of MA part")
par(op)
```

```{r forecasting, echo=FALSE}
fit12 = arima(candyts, order = c(1,1,2), method = "ML")
mypred <- predict(fit12, n.ahead=12)
candy.orig <- ts(candyts)
ts.plot(candy.orig, xlim=c(0,455), ylim=c(40,200), main = "Figure 9: US Candy
Production, Forecasted")
points(441:452,mypred$pred, col="blue")
lines(441:452,mypred$pred+1.96*mypred$se,lty=2,col="steelblue")
lines(441:452,mypred$pred-1.96*mypred$se,lty=2,col="steelblue")
```

```{r zoomed in forecasting echo=FALSE}
ts.plot(candy.orig, xlim=c(430,455), ylim=c(20,200), main = "Figure 10: Twelve
month Forecast of US Candy Production")
points(441:452,mypred$pred, col="blue")
lines(441:452,mypred$pred+1.96*mypred$se,lty=2,col="steelblue")
lines(441:452,mypred$pred-1.96*mypred$se,lty=2,col="steelblue")
```
~~~