University of California Santa Barbara

Department of Probability and Statistics

PSTAT 274 Final Project

Analysis of Alcohol Sales in the U.S.

Authors:

Luis Aragon
Perm Number: 3898269
Daniel Fields
Perm Number: 9565052
Max Gershman
Perm Number: 9458200
Nicole Grazda
Perm Number: 3745692

Professor:

Sudeep Bapat

Table of Contents

Table of Contents **2**

Abstract **3**

Introduction **3**

Exploratory Data Analysis **5**
- Data Exploration 5

Data Transformation **6**
- Box-Cox Transformation for Variance Stabilization 7
- Removing Seasonality and Trend 8

Model Selection **11**
- Preliminary Model Identification 11
- Narrowing Down Our Model 12

Model Estimation **13**

Model Diagnostics **14**
- 1. Normality of Errors 14
- 2. Serial Correlation Detection 16
- 3. Homoscedasticity 18

Forecasting **19**

Spectral Analysis **21**
- Periodogram 21
- Periodicities in Stationary Model 23
- Fisher Test 24
- Kolmogorov-Smirnov Test 24

Conclusion **25**

References **27**

Appendix **27**
- Appendix: Figures 27
- Appendix: Code 29

Abstract

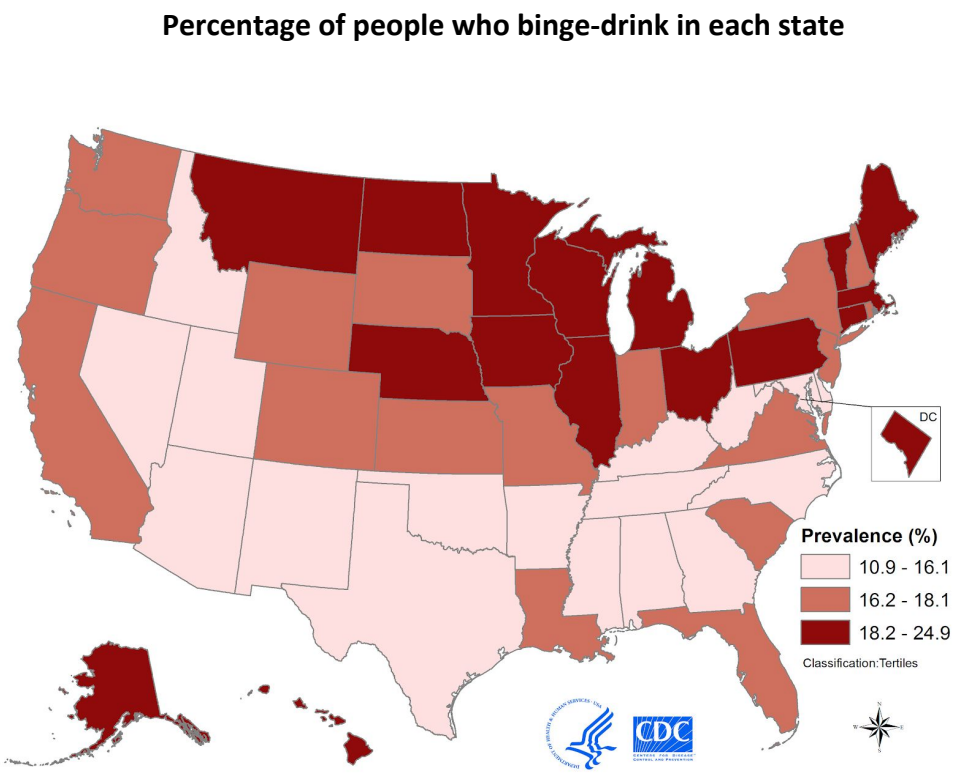
In the most recent decades, alcohol sales have increased in the United States. The goal of this project is to predict monthly American alcohol sales using methods of time series analysis and R software. Initially, we construct a proper SARIMA model based on the stationary data, which we obtain by differencing and transforming the original data. We then perform diagnostic checks to ensure our model's feasibility. Our forecast predicts the retail sales of booze between December 2017 and November 2018, which lies between the 95% confidence interval. In addition, we performed spectral analysis to gain a deeper understanding of our final model.

Introduction

Beer, wine, and spirits have shaped American history and have left an indelible mark on this nation's culture. Alcohol has a significant presence in this country and is glorified by the media. Thus, booze sales have increased steadily in the past century and it is worth our attention. It is therefore important that we examine its trend, and forecast its growth to be able to recognize how American physical and mental health will be affected. Additionally, analyzing alcohol sales provides deeper insight into the dynamic United States economy and the general well-being of the country.

Our dataset measures the monthly retail sales of beer, wine and liquor stores from January 1992 to November 2018, in millions of dollars. This data set is important to forecast to better understand the United State's future and attitude towards drinking. We have chosen this data not only for its relevance to society but for its large sample size that would yield accurate results. From the original time series plot, we notice an upward trend and slight heteroscedasticity. Additionally, we notice a seasonal pattern at uniform intervals, which is reasonable since our data is recorded monthly by the *U.S. Bureau of the Census*. We perform a box-cox transformation with λ = -0.22 and difference at lag 12 to obtain a stationary

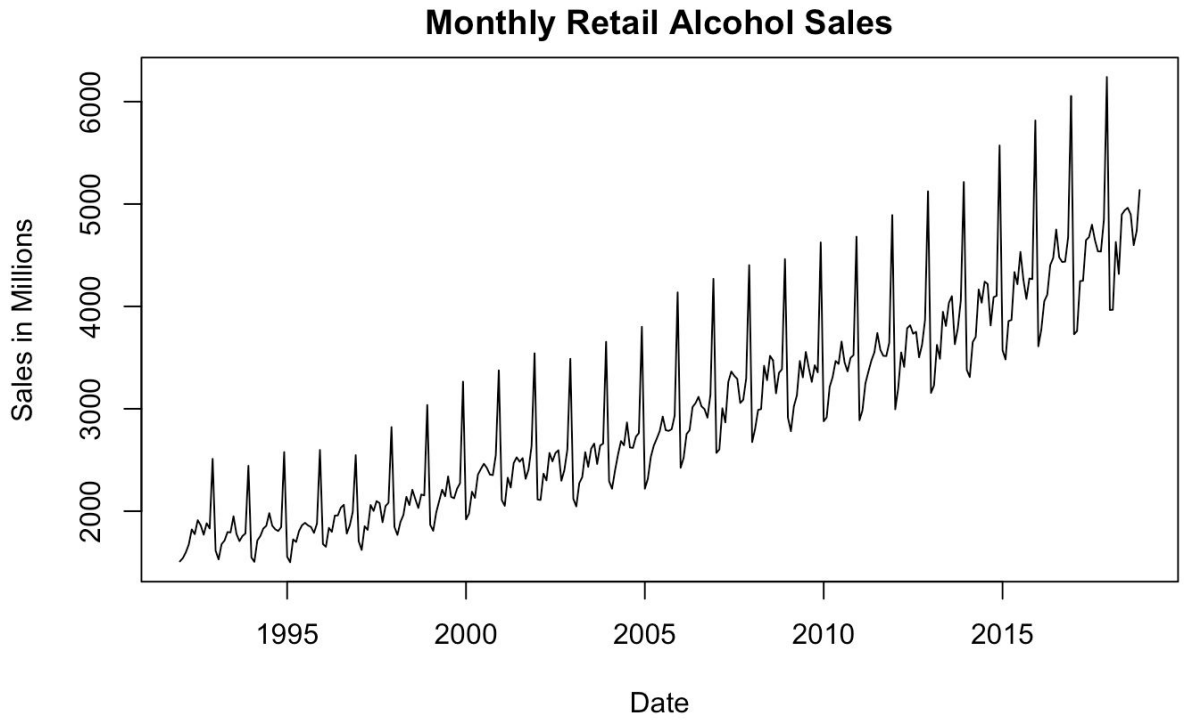
series. Since the variance is minimized after conducting the differencing and the transformation, we recognize that the data is ready to be identified as a certain model with its estimated orders. After analyzing our ACF and PACF plots and considering the AIC, AICc and BIC values, we narrow down our final model to be SARIMA(3,0,0)x(1,1,2) 12. We forecast from December 2017 to November 2018 and observe that our prediction lies in the 95% confidence interval and that our predicted values are approximately close to the true values in the original dataset. Furthermore, we conduct Spectral Analysis to examine its periodic behavior and conclude by the Fisher and Kolmogorov-Smirnov test that the residuals of the data are White Noise. Overall, our model proves to be feasible.



Exploratory Data Analysis

Data Exploration

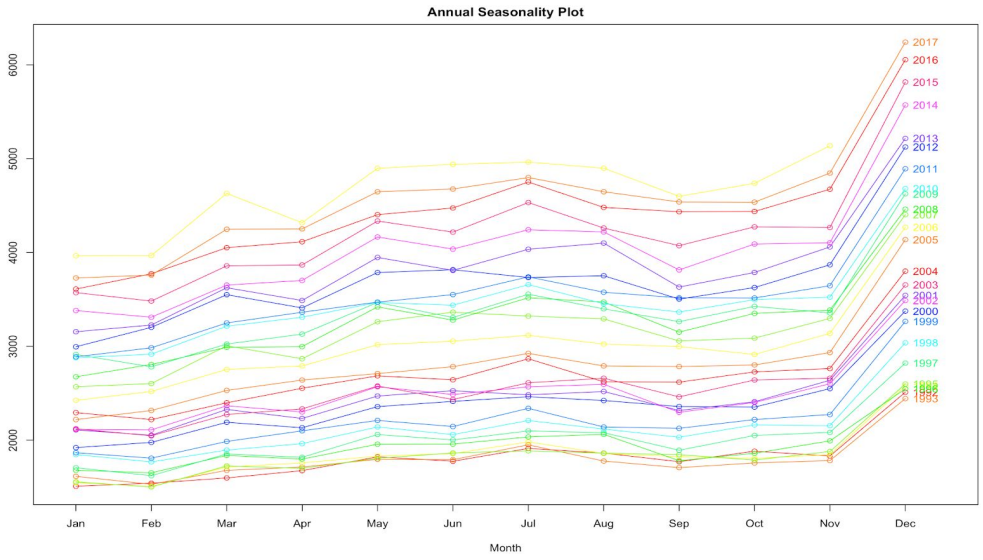
For this project, we used a dataset that included two variables: monthly dates from January 1992 to November 2018, and the monthly total retail sales of beer, wine, and liquor in the United States measured in millions of dollars. This dataset contains 323 observations. Below is a glimpse of the data we are working with.



We derive from this plot that the time series data has a very strong seasonal component of period d=12, which is understandable since the data is recorded monthly. Logically, the data shows an increasing trend, a result of both an increase in consumption over the years as well as monetary inflation. The range of values that the series can assume is clearly not constant across all time, concluding that the variance does vary with time. A transformation will be needed on this data to assist with normalization. Taking a closer

look at the seasonal plot we observe a large increase in sales in December and a large decrease in sales

in January. These considerable peaks and valleys are most likely a result of the holiday season. Many people drink on Thanksgiving, Christmas and New Years, yet conversely, many Americans begin their “New Year's Resolution” in January in an attempt to be healthier and reduce intake of alcoholic beverages. Another element to note is alcohol sales begin to spike in the summer, the time of the year when many will take a vacation from work and school.



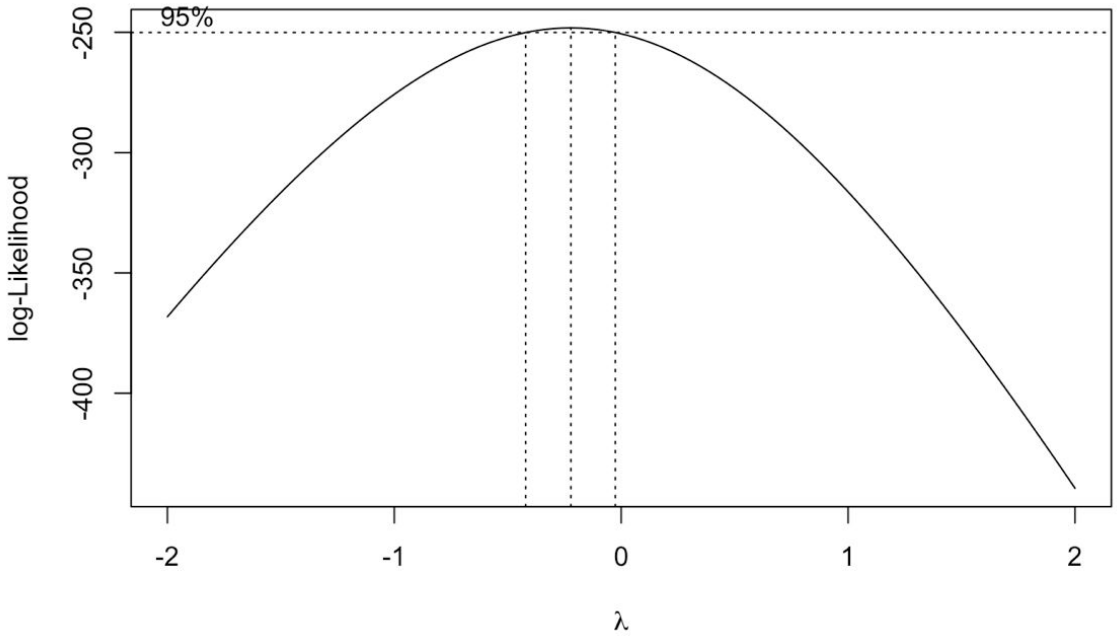
Data Transformation

As stated above the variance does vary with time, therefore the dataset is non-stationary. To make this data stationary we need to first stabilize the variance. Following this transformation, we will detrend and deseasonalize the data.

Box-Cox Transformation for Variance Stabilization

Due to heteroscedasticity, our original time series violated our constant error of variance assumption. A Box-Cox transformation is a way to transform non-normal dependent

variables into a normal shape. Using the Box-Cox package in R we graph the log-Likelihood in relation to λ defined by $f(B_t)_\lambda = \frac{(B_t\lambda - 1)}{\lambda}$ where $\lambda \neq 0$.



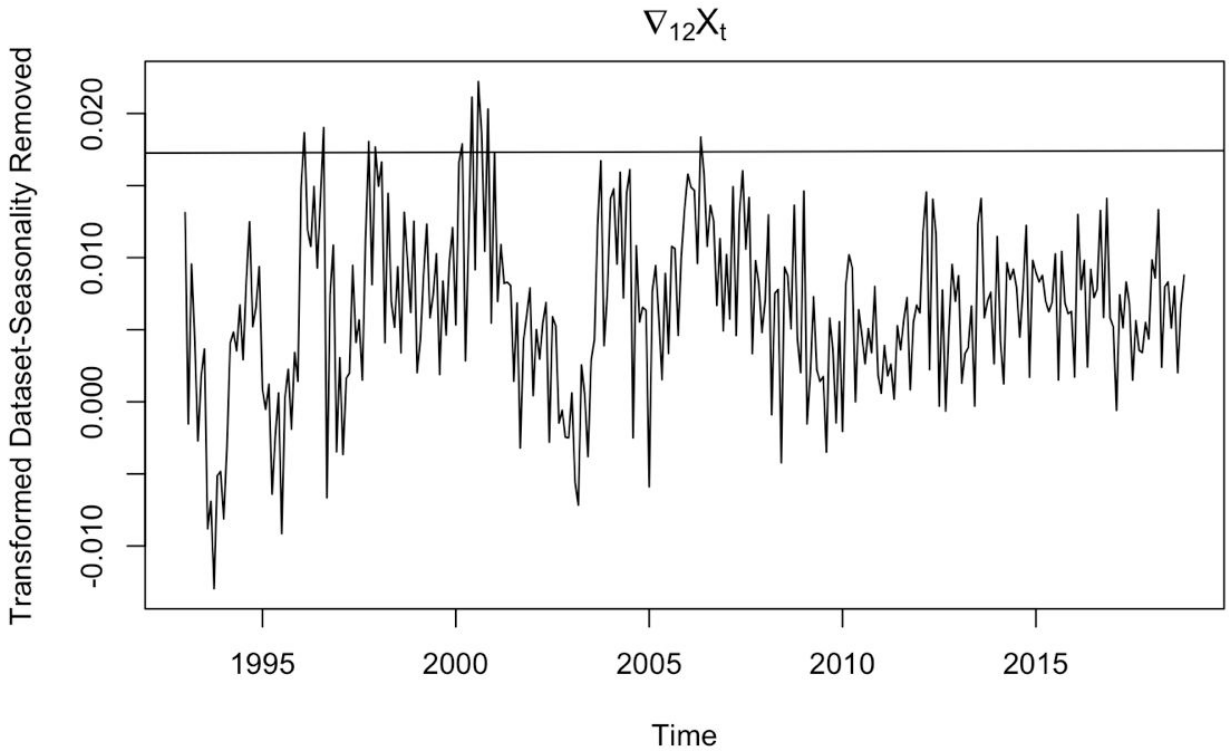
Thus calculating our λ = -.02222. We transform our data points according to the Box-Cox transform and raise them to the power of λ resulting in a new time series:

$$X_t = B_t^{-0.2222}$$

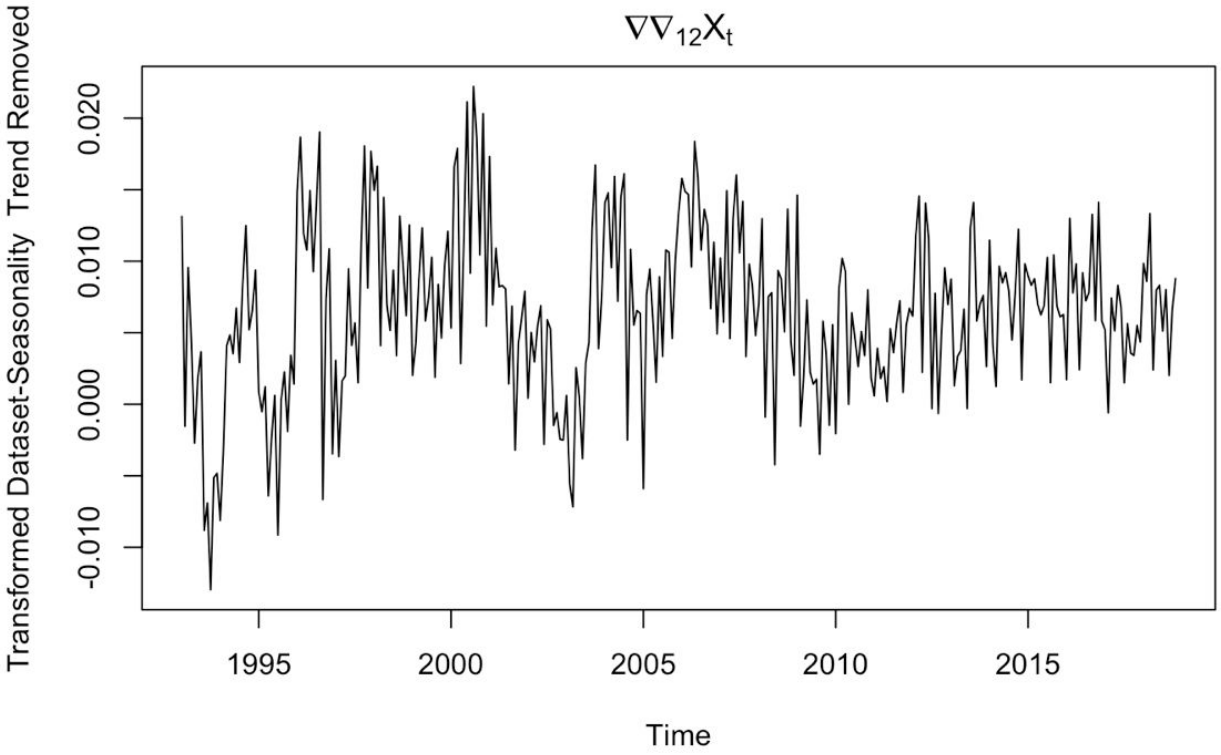
Though the data is now more normally distributed, it is not yet stationary. We must remove seasonality and trend. The resulting variance after this transformation is $4.759559e^{-05}$.

Removing Seasonality and Trend

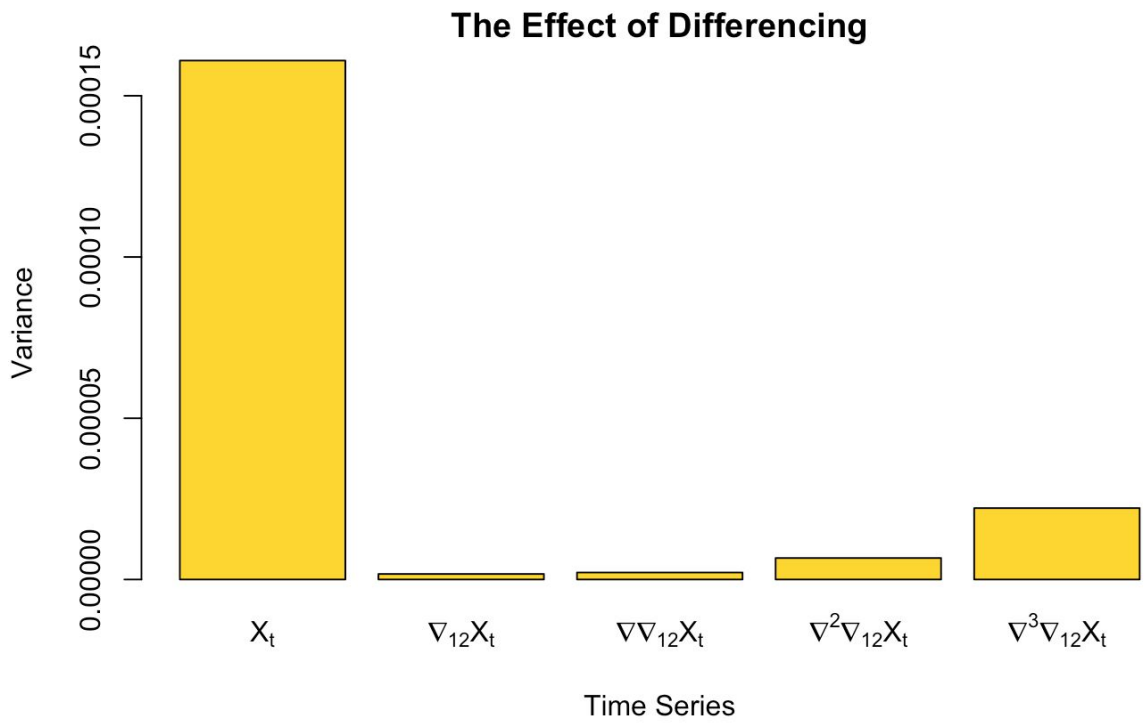
Differencing the data to remove the seasonality and trend will be our next goal on our path to developing a working model. Previously, it was stated that there seemed to be a strong monthly seasonal component, therefore, lagging the difference by 12 seemed appropriate.



The plot above is the dataset differenced at lag 12. The resulting variance after the differencing is $3.633156e^{-07}$. Next, we must remove the trend using the transformed dataset with the seasonality removed. Once again we will difference the dataset, yet this time using a lag of 1.



The plot of the transformed data deseasonalized and detrended is above. Then, to determine which time series we want to build a model on, we want to use the time series that has the lowest variance. To do this, we differenced the model at lag 12 to get rid of seasonality, differenced this the lag-12 differenced series once more to remove any trend we observed in the original data, and then repeated this 1-difference twice more to examine a trend of overfitting, which resulted in the following plot.



In regards to the transformation we applied to the data, we observe a reduction on orders of magnitude about 75 times from the Box-Cox transformed data to the final transformation. Due to this, the original variance is not included in the above plot so that the subtle evidence of overdifferencing will be visible due to the scale of the changes in the differenced transformed data's variances. The resulting variance of taking an additional difference *after* differencing at lag 12 is slightly larger than that of the deseasonalized data set with a resulting increase in variance of $5.767335e^{-07}$. Since the variance increases, we consider removing the trend to be overdifferencing. Nevertheless, we differenced the data again to check how the variance reacts to another lag. As expected, the variance increases substantially to $1.872324e^{-06}$ therefore, differencing at lag 1 after removing seasonality would be considered over-differencing. Our final transformation of the original data is then

$$Y_t = \nabla_{12} X_t = \nabla_{12} B_t^{-0.2222}$$

Using our final transformation Y_t an augmented Dickey-Fuller test was performed. The result was a p-value of less than 0.01, and thus a rejection of the null hypothesis that our

time series is non-stationary. Thus, Y_t is, in fact, a stationary dataset. At this point, we can move on to model identification and estimation.

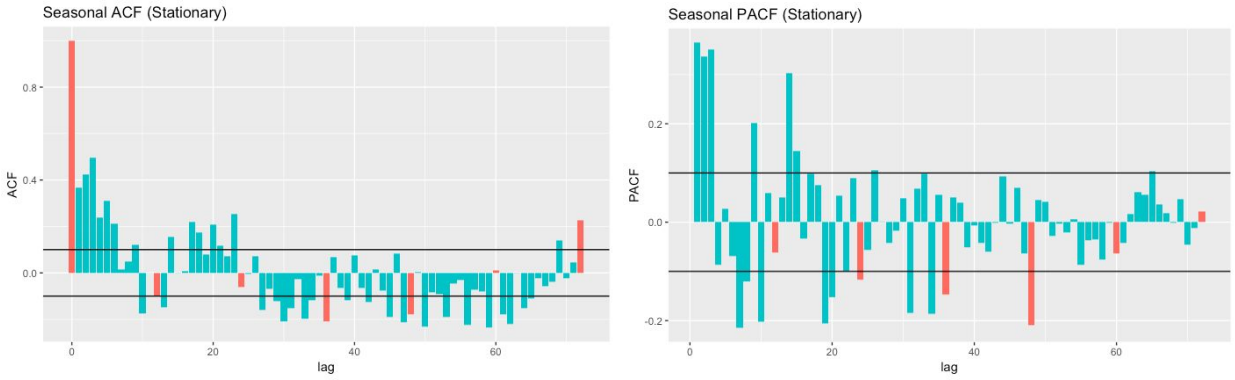
Model Selection

Preliminary Model Identification

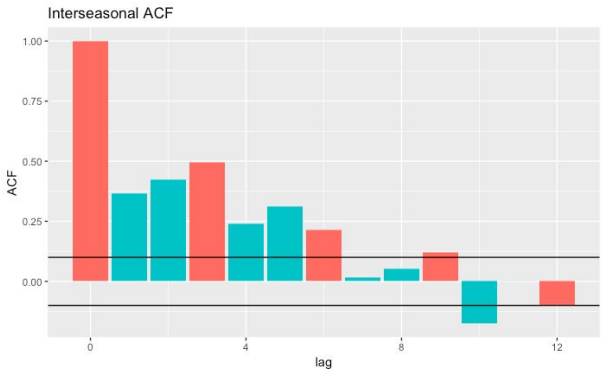
For our sales data, we will identify a seasonal ARIMA model in the form of:

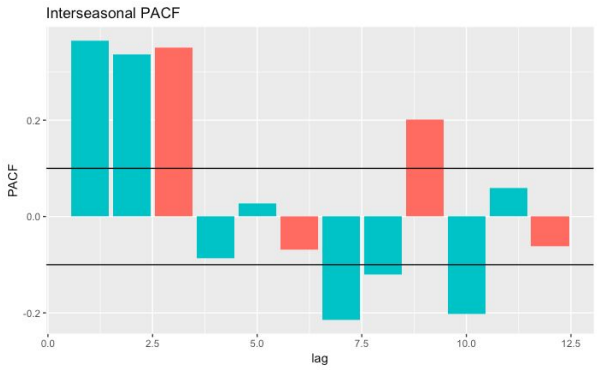
$$SARIMA(p,d,q) \times (P,D,Q)_s$$

In our data, we have identified the season to be 12, so we can conclude that s = 12. Next, we will identify possible seasonal components using the ACF and PACF plots and looking at lags that are a multiple of 12.



We notice that the ACF trails off at around lag 24, therefore, Q = 2. Also, by looking at the PACF above, we can see that there are large spikes around lag 12 before cutting off, so since we have a seasonal period of s = 12, then this cut off at lag 12 suggests that P = 1. As a side note, we notice a spike in the PACF at lag 48, which indicates that a possible value for P can be seasonal P = 4. We will eliminate the model where seasonal P is large because it makes the model very complicated and may cause overfitting. Lastly, because we differenced once at lag 12 (our season) to get our model, $\nabla_{12} X_t$, so we will set D=1. Now, we will identify the non-seasonal components of our model by looking at lags between 1 and 12 in our ACF and PACF plots.





For the non-seasonal part of our SARIMA, we notice that the trend was removed after differencing at lag 12, so we did not have to difference again at lag 1. Therefore, d = 0. If we look closely at the PACF (between lags 0 and 12), we notice that the PACF cuts off after 3. This indicates that p = 3. The ACF trails off after lag 0 and lag 3, but there does not seem to be a significant cut off point. The trail off after lag 3 in the ACF is a characteristic of an AR(3) model. Thus, we conclude that q = 0, q = 1, q= 2, or q = 3 . Therefore, we think that our time series model can be either:

$(i)\ \ SARIMA(3,0,0)\times(1,1,2)_{12}$

$(ii)\ SARIMA(3,0,1)\times(1,1,2)_{12}$

$(iii)\ SARIMA(3,0,2)\times(1,1,2)_{12}$

$(iv)\ SARIMA(3,0,3)\times(1,1,2)_{12}$

Narrowing Down Our Model

We will now compare the AIC, AICc and BIC values for each model in the table below:

Model	AIC	BIC	AICc
$(i)\ \ SARIMA(3,0,0)\times(1,1,2)_{12}$	-12.92723	-13.84537	-12.91962
$(ii)\ SARIMA(3,0,1)\times(1,1,2)_{12}$	-12.92709	-13.83353	-12.91912
$(iii)\ SARIMA(3,0,2)\times(1,1,2)_{12}$	-12.92309	-13.81783	-12.91472
$(iv)\ SARIMA(3,0,3)\times(1,1,2)_{12}$	-13.10557	-13.98862	-13.09675

We can see that model (iv) has the lowest AIC, BIC, and AICc followed by model (i). These models are very close, so using the idea of parsimony, I will eliminate model (iv) because it is the most complex model. Therefore, we will choose model (i):

$$(i)\ SARIMA(3,0,0)\times(1,1,2)_{12}$$

Model Estimation

Now we will fit this model using the Maximum Likelihood Estimator method:

Model	ar1	ar2	ar3	sar1	sma1	sma2
$SARIMA(3,0,0)\times(1,1,2)_{12}$	0.1275	0.2808	0.4208	0.3403	-1.9865	0.9998

Our full model can be written as:

$$(1-0.1275B-0.2808B^2-0.428B^3)(1-0.3403B^{12})(1-B^{12})X_t = (1-1.9865B^{12}+0.9998B^{24})Z_t$$

where Zt ~N(0, $1.627e^{-5}$)

By solving for the roots from the equation and plotting them (Appendix Figure 1), we can see that they lie outside the unit circle. This means that the model is both stationary and invertible. Additionally, the absolute value for the coefficient of the seasonal AR part is less than 1, so it is stationary. Now that we have our model, we can move forward with the model diagnostics. If this model does not pass diagnostic checks, then we will go back and try our second larger model, which is $SARIMA(3,0,3)\times(1,1,2)_{12}$

Model Diagnostics

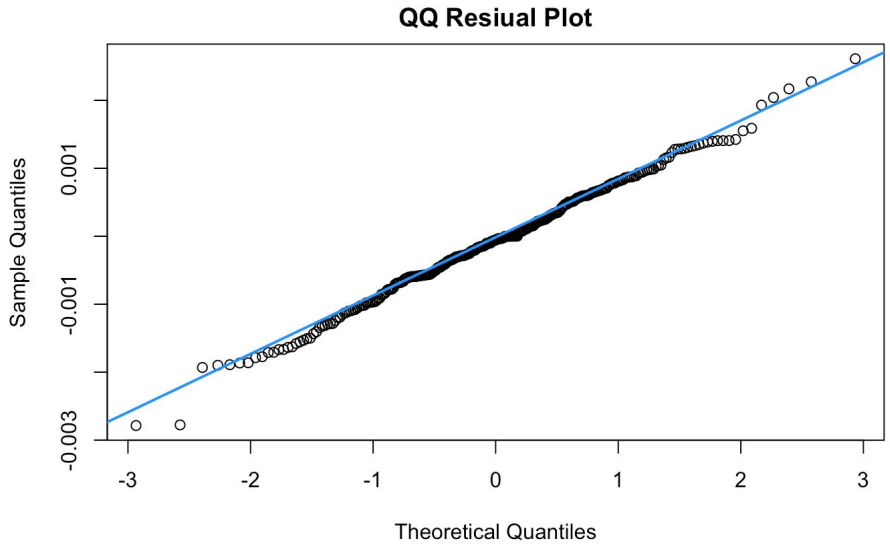
Now that we have fit the model, we need to check diagnostics to ensure our assumptions that the model is estimated upon are valid. These assumptions are the Normality of the

errors, there does not exist a serial correlation within the model, and the model is homoskedastic. These assumptions in more clear language are that the errors follow a normal distribution, the data is not serially correlated to itself, and that the data show constant variance across time.

1. Normality of Errors

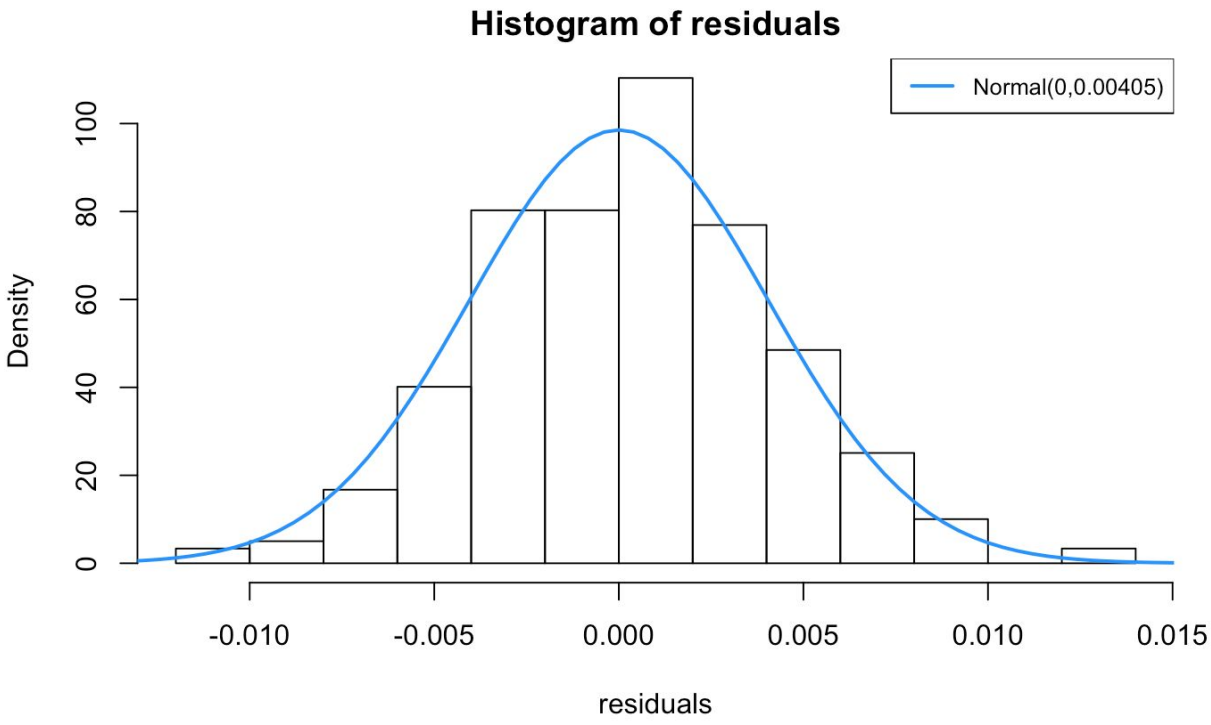
First, to assess the assumption that the errors are normally distributed. We want this to be the case in our model, as it shows that our model's errors fall equally above and below true values which is evidence of unbiased estimators which indicates a good model fit. To check the normality of errors we want to look at the QQ Plot of the errors around the mean of the time series, a histogram of the errors, and use the Shapiro Wilk Test to test if the residuals are approximately IID Gaussian.

For the QQ Plot, we want to see that the errors hug close to the 45° line shown in blue in the plot.



This is exactly what we see in our QQ Plot. This is good evidence that the errors are normally distributed.

Then, to further our intuition that the errors are indeed Gaussian, we can also examine a histogram of the values the errors themselves take. We have superimposed a normal distribution on top of the errors' histogram, with a mean of zero and a variance estimated by the fact that 99.9% of data in a normal distribution falls within 3 standard deviations of its mean. This histogram is shown below.

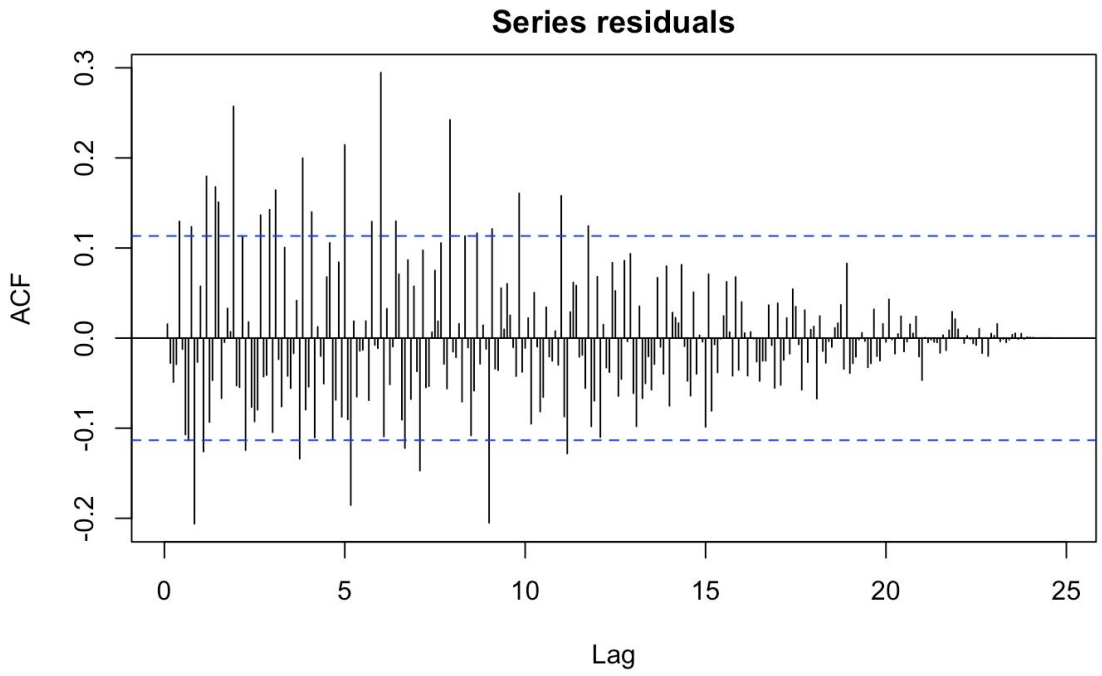


This is further evidence that our errors are normally distributed.

Finally, we employed a Shapiro-Wilkes Test. Specifically, we are testing the null hypothesis that the errors follow a normal distribution, versus the alternative hypothesis that the errors are not normally distributed. After using this test in R on our errors we receive a p-value of 0.6754. Using the rejection rule at $\alpha = 0.05$, if P-value > 0.05, we would reject the null hypothesis. Here, our null hypothesis is that the residuals are approximately IID Gaussian. Since, 0.6754 is *not* less than 0.05, we fail to reject the null hypothesis. Thus, the assumption that the residuals are approximately IID Gaussian is valid.

2. Serial Correlation Detection

We want to ensure that our data does not exhibit patterns of serial correlation because if there is serial correlation then that means that OLS is no longer an efficient linear estimator, the reported standard errors are likely incorrect and usually overstated, and the OLS estimates are biased and inconsistent due to the lagged dependent variable being used as a regressor. Thus, to ensure our model is not serially correlated is of high importance. To check we can examine a few things. First, we can examine the autocorrelation plot of the errors to visually examine if we will observe serial correlation which will be shown in the ACF by significant valued correlations. This is due to the fact that if we have serial correlation, then we would expect that errors have significant autocorrelation amounts at different lags. In our model, we get the following ACF for our errors.



So, we observe that at most of the lags in the errors, we have borderline significant values for the autocorrelation. To determine quantitatively if we have serial correlation present, we can use a use the Ljung-Box (Modified Box-Pierce) Lack-of-Fit Test. With this, we are

testing H_0 : $\rho_1 = \rho_2 = \ldots = \rho_k = 0$ for all k (where ρ_i = is the autocorrelation of the residuals at lag i) verus the alternative hypothesis H_A : There exists j in {1,..,k} such that $\rho_j \neq 0$.
After employing this test in R, we obtain a p-value of 0.7875. So, using the rejection rule at $\alpha = 0.05$, if P-value > 0.05, then we reject the null hypothesis. Here, our null hypothesis is that the residuals are uncorrelated. Since, 0.7875 is not less than 0.05, we fail to reject the assumption that the residuals are uncorrelated. Thus, we can say that no serial correlation is present, and therefore assumption 2 is upheld.

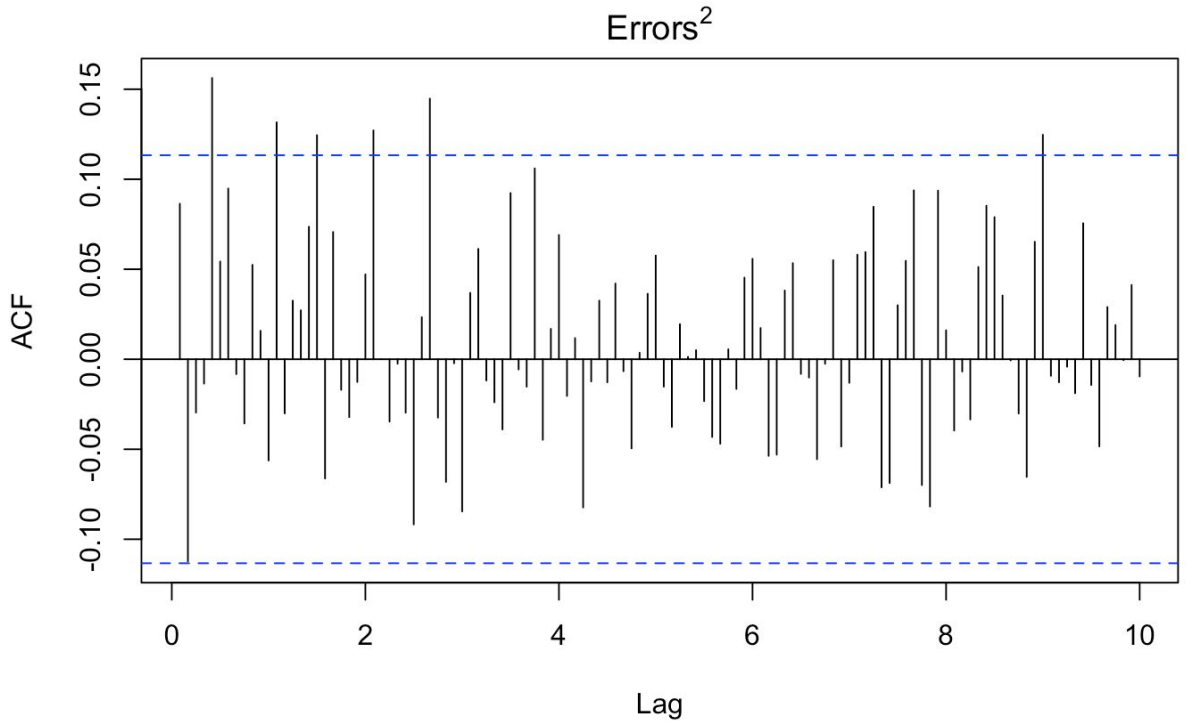
3. Homoscedasticity

Homoscedasticity means that the variance shown is approximately constant throughout time. Or in other words, that we do not observe periods where in the variance is significantly larger or smaller than other periods. Heteroskedasticity, when the variance of the time series *does* change over time periods, is a problem in a similar fashion to serial correlation. If we have heteroskedasticity, then the OLS estimates are no longer the BLUE (Best Linear Unbiased Estimators) because they are no longer efficient, so the regression predictions will be inefficient too. Additionally, due to non-constant variance the estimate for the covariances will be inaccurate and as a result the resulting hypothesis tests (t-test, F-test) become invalid to use when heteroskedasticity is present. To check for heteroskedasticity, we can examine a few things. First, we can examine the autocorrelation and partial autocorrelation plots of the squared residuals. The intuition behind this, is that the errors should be normally distributed with mean zero and variance σ^2_Z. By definition of variance, the variance of Z_t, the white noise component of the time series, is defined by:

$$\sigma_Z^2 = Var(Z_t) = E[(Z_t - \mu)^2] = E[(Z_t - 0)^2] = E[(Z_t)^2]$$

So, examining the autocorrelation of the squared errors should tell us about the variance. Thus, if there is not heteroskedasticity in the data, then we should observe that in the ACF and PACF of the errors, there will be values present that lack significance. Or in other words, the ACF and PACF should all be within acceptable bounds for simple white noise.

We observed the following plots of ACF and PACF for our errors.



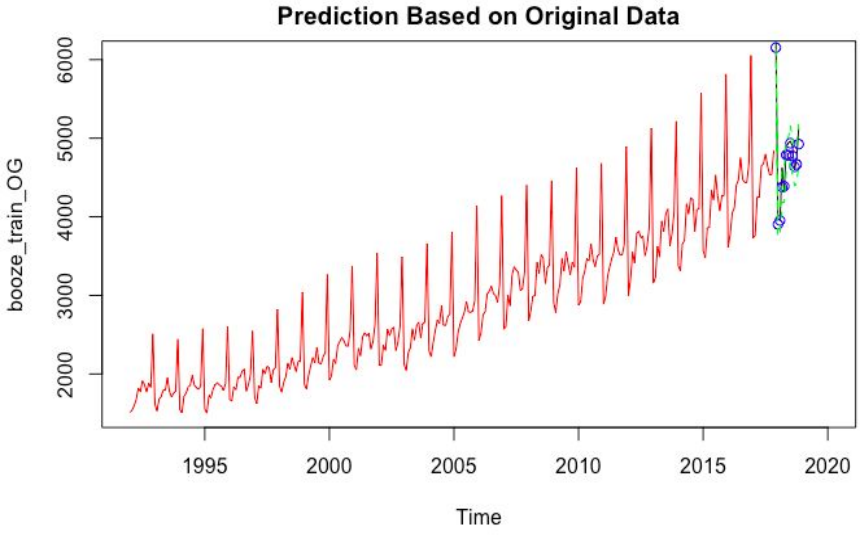
So, based on these it would appear from the ACF and PACF of the squared residuals that our errors are homoskedastic. Additionally, the Ljung-Box test, also quantitatively validates the assumption that our errors are homoskedastic.

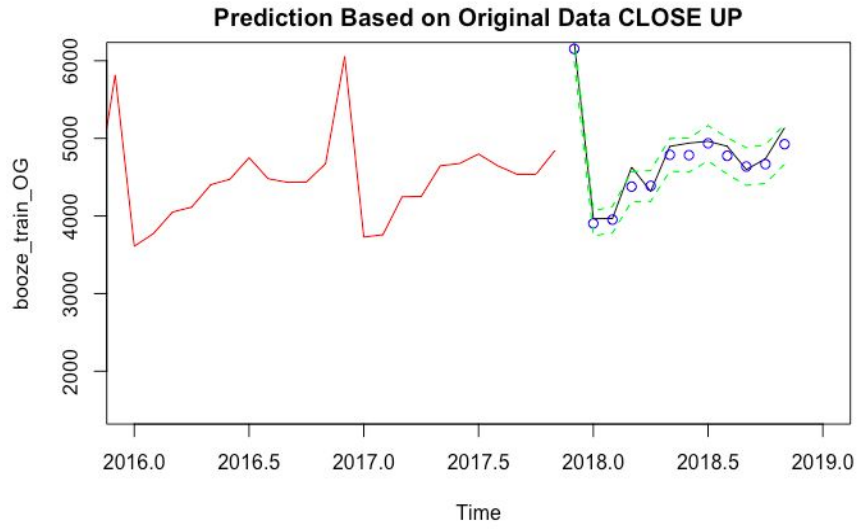
Our model passes all of the diagnostic checks, and thus, our model is valid since it upholds of all of its assumptions.

Forecasting

Now that we found a suitable model for our data, we can forecast future retail alcohol sales for the next 12 months (1 year). Below is a plot forecasting our original data (You can find a plot forecasting our transformed data in the Appendix). We can clearly see in the graph below that the predicted values follow nearly an identical seasonal pattern as the current data. There is a predicted spike in sales, followed by a drop. Additionally, we see that the

prediction follows the same upward trend we noticed from the original data, all while the variance increases. From a general perspective, we can predict that future sales for next year will follow a close pattern as previous years. We can see the predicted value (blue points) follow very closely with the actual values (black line). This is an indication that the model we chose works very well with the data. It recognizes the trend, seasonality, and increasing variance of the original sales data.





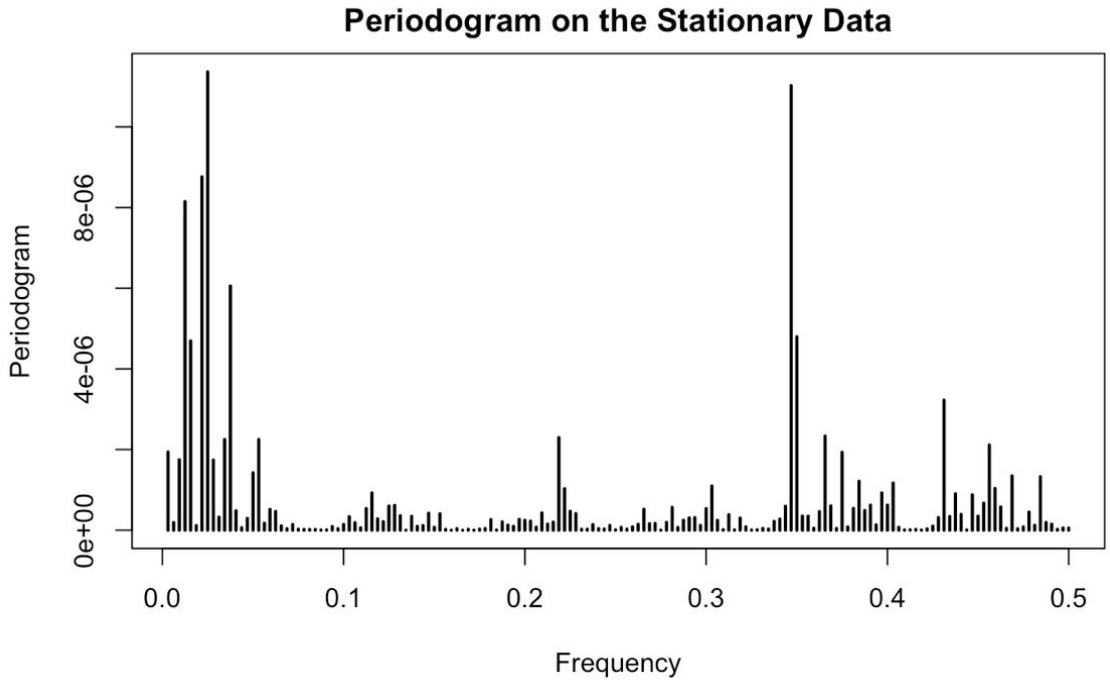
Spectral Analysis

The idea of spectral analysis is to decompose a stationary time series {Xt} into a combination of sinusoids, with random (and uncorrelated) coefficients. Instead of the time domain, spectral analysis uses the frequency domain, an approach that considers regressions on a sinusoid. The model will look like a Fourier Series.

$$X_t = \mu + \sum_{j=1}^{k}(A_j \cos 2\pi v_j t + B_j \sin 2\pi v_j t), \text{ where } v = \text{frequency}$$

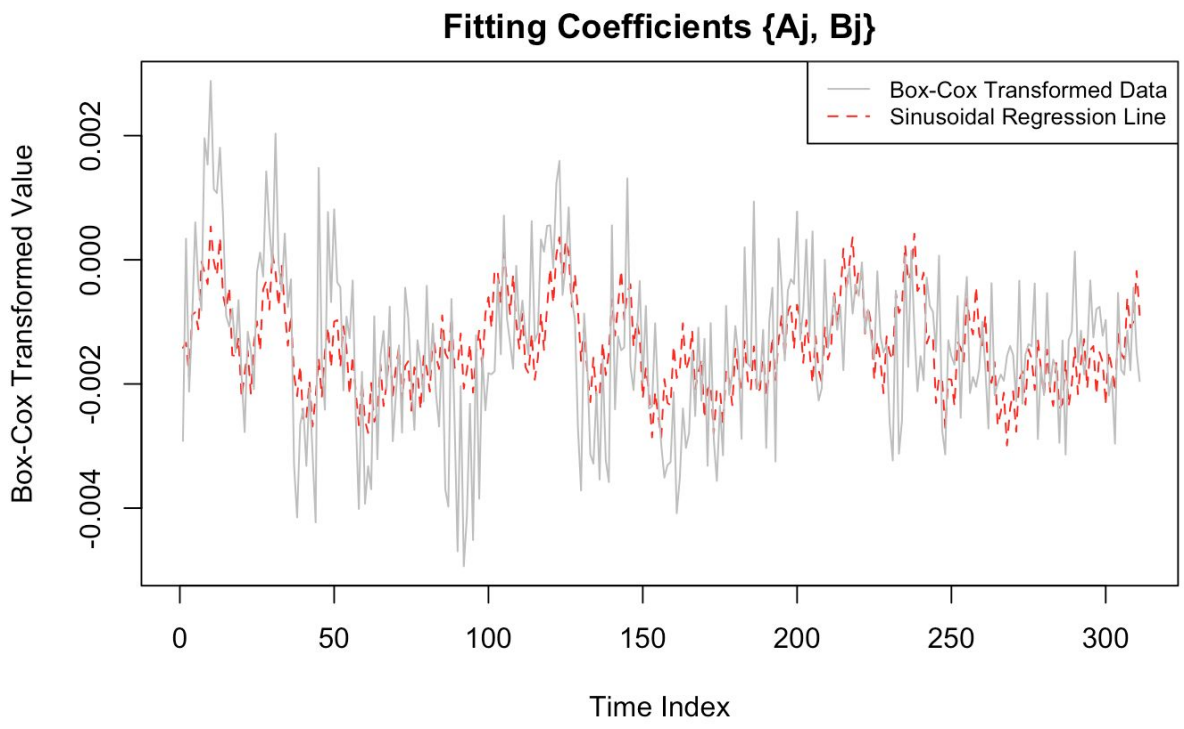
Periodogram

A periodogram graphs a measure of the relative importance of possible frequency values that might explain the cyclical pattern of our data.



The periodogram pictured identifies the eight most dominant frequencies in our model. These frequencies can then determine the coefficients A_j and B_j. After plotting our stationary data our eight most common frequencies observed are 0.025000, 0.346875,

0.021875, 0.012500, 0.037500 0.350000, 0.015625, 0.431250. Now using a regression the corresponding A_j and B_j can be calculated.

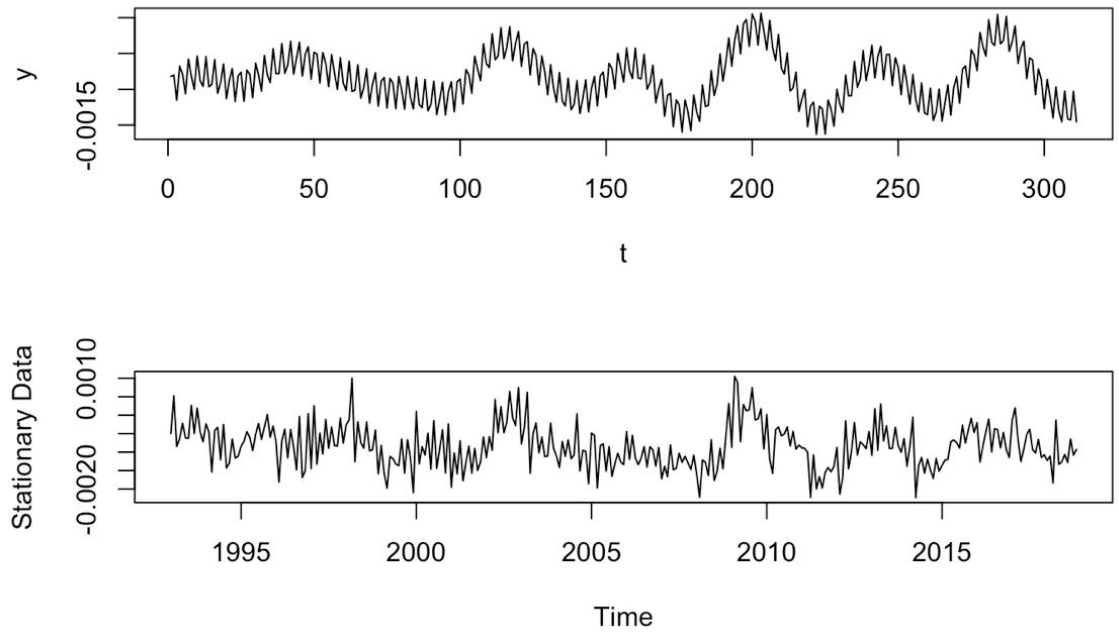


After fitting a linear model onto the sinusoidal representation of our time series, our final spectral model can be represented as

$$X_t \approx -8.47e^{-04} + 2.439e^{-04}\cos(2\pi * 0.025t) + 7.320e^{-05}\sin(2\pi * 0.025t) -$$
$$2.617e^{-04}\cos(2\pi * 0.346875t) - 2.835e^{-05}\sin(2\pi * 0346875t)$$
$$-1.498e^{-04}\cos(2\pi * 0.021875t) + 1.995e^{-04}\sin(2\pi * 0.021875t)$$
$$-3.1302e^{-04}\cos(2\pi * 0.0125t) - 1.511e^{-05}\sin(2\pi * 0.0125t)$$
$$-6.0589e^{-05}\cos(2\pi * 0.0375t) + 2.003e^{-04}\sin(2\pi * 0.0375t)$$
$$+ 1.458e^{-04}\cos(2\pi * 0.35t) - 8.726e^{-05}\sin(2\pi * 0.35t)$$
$$+ 1.477e^{-04}\cos(2\pi * 0.015625t)$$
$$+ 4.355e^{-05}\sin(2\pi * 0.43125t) + 1.25e^{-04}\sin(2\pi * 0.43125t)$$

It should be noted that the data used for this spectral analysis must be stationary, therefore the transformed and deseasonalized data was used to develop this model. Below is a plot of

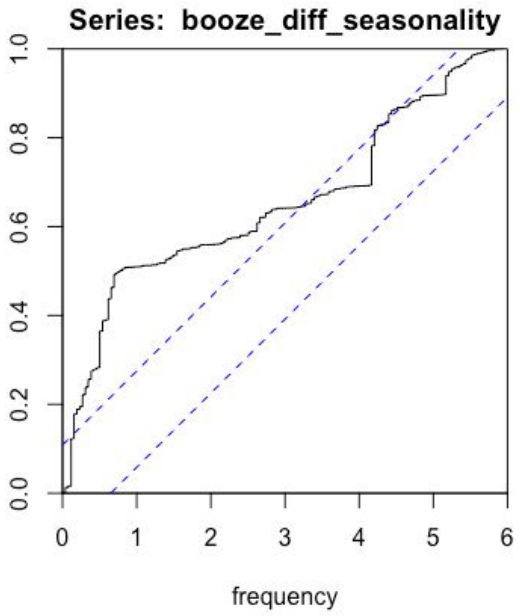
a comparison between the plot of the stationary data and the data approximated using spectral analysis.



From the plot above it seems that spectral techniques represent our data fairly well, nevertheless our adjusted R^2 value is only 0.4858. Therefore, about 48.58% of the variance in the data can be explained by the spectral analysis, which shows that our data very much resembles a sinusoid.

Periodicities in Stationary Model

We will use the Fisher test and Kolmogrov-Smirnov test on the stationary data to see if there are any periodicities. If there are, this means we can represent the data with a sinusoidal function. The null hypothesis is that there are no periodicities. Below, we see that the null is rejected for the Kolmogrov-Smirnov. In addition, the Fisher test gives us a value of $5.495556e^{-08}$. Therefore, there are periodicities and we can use a sinusoidal function. Now we will test the residuals.

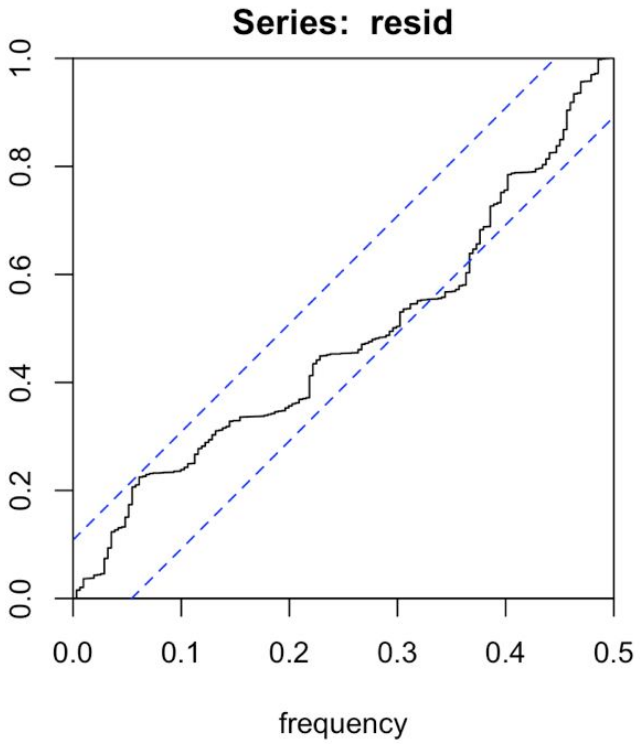


Fisher Test

The Fisher test enables one to test the data for the presence of hidden periodicities with unspecified frequency(Lec 10 pg.18). This test is applied to the residuals and a value of 0.2824072 is returned therefore the Fisher Test passed, and we can confirm that the simulation can be represented as Gaussian White Noise.

Kolmogorov-Smirnov Test

The Kolmogorov- Smirnov test can also be used to test whether the residuals follow Gaussian White Noise.



The cumulative periodogram never exits the boundary, therefore according to this test our residuals are Gaussian white noise. Since both tests pass we conclude the residuals of our final spectral model are Gaussian White Noise. We accept the null that there are no hidden periodicities.

Conclusion

Our goal is to construct a time series model that can explain the alcohol sales in the United States from the past 25 years and to predict its future sales 12 months ahead. We have observed that the data exhibited an upward trend with heteroscedasticity. Furthermore, we have observed a pattern and seasonality in the data that shows that the sales reaches its maximum per year in each December. This may potentially be the result of social drinking in the Christmas holiday season and the desire to consume copious amounts of alcohol before giving it up as a New Years Resolution. After making our data stationary, we narrow down our model by the model selection process and conduct diagnostic checks. Our final model is displayed below.

Let Xt be the transformed and differenced data:

$$Y_t = \nabla_{12} X_t = \nabla_{12} B_t^{-0.2222}$$

SARIMA(3,0,0)x(1,1,2) 12

$$(1 - 0.1275B - 0.2808B^2 - 0.428B^3)(1 - 0.3403B^{12})(1 - B^{12})X_t = (1 - 1.9865B^{12} + 0.9998B^{24})Z_t$$

where Zt ~N(0, $1.627e^{-5}$)

Our forecast indicates that between December 2017 and November 2018, the retail sales of alcohol will continue to increase. Since our forecast lies in the 95% confidence range, our model proves to be feasible. Finally, we performed spectral analysis to approximate the model into a combination of sinusoids using a periodogram and tests.

The goal of our project was to predict future monthly American alcohol sales using methods of time series analysis. Overall, we have accomplished this goal and gained a deeper insight into the booze industry in the United States.

We cordially thank Professor Bapat for teaching us over the course of the past 10 weeks. He has been super helpful and supportive in the learning process of such a difficult class. Additionally, we would like to thank our TAs for helping us along the way and contributing their time to allow us to master the material. Thank you for everything!

References

1) *Alcohol and Public Health* https://www.cdc.gov/alcohol/data-stats.htm

2) Underage Drinking https://pubs.niaaa.nih.gov/publications/UnderageDrinking/Underage_Fact.pdf

3) *History of Alcohol in America* https://axisresidentialtreatment.com/alcohol- addiction/ history-in-america/

Appendix

Appendix: Figures

Figure 1

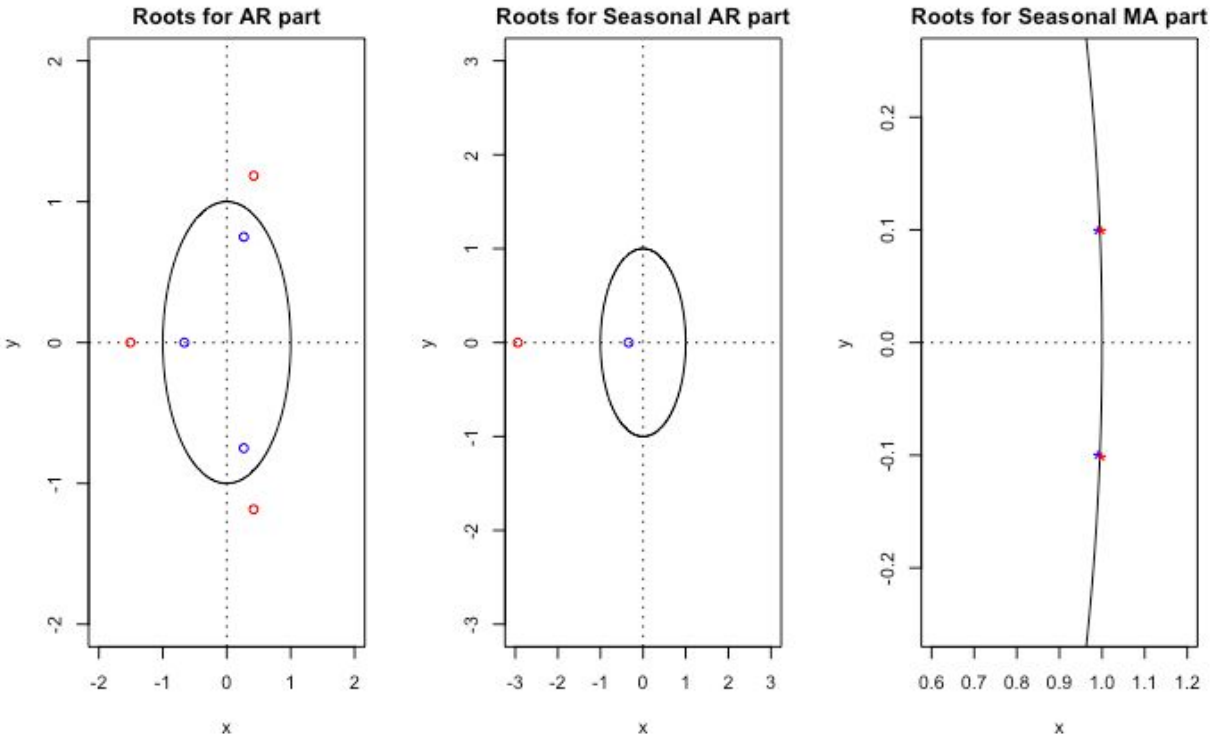


Figure 2

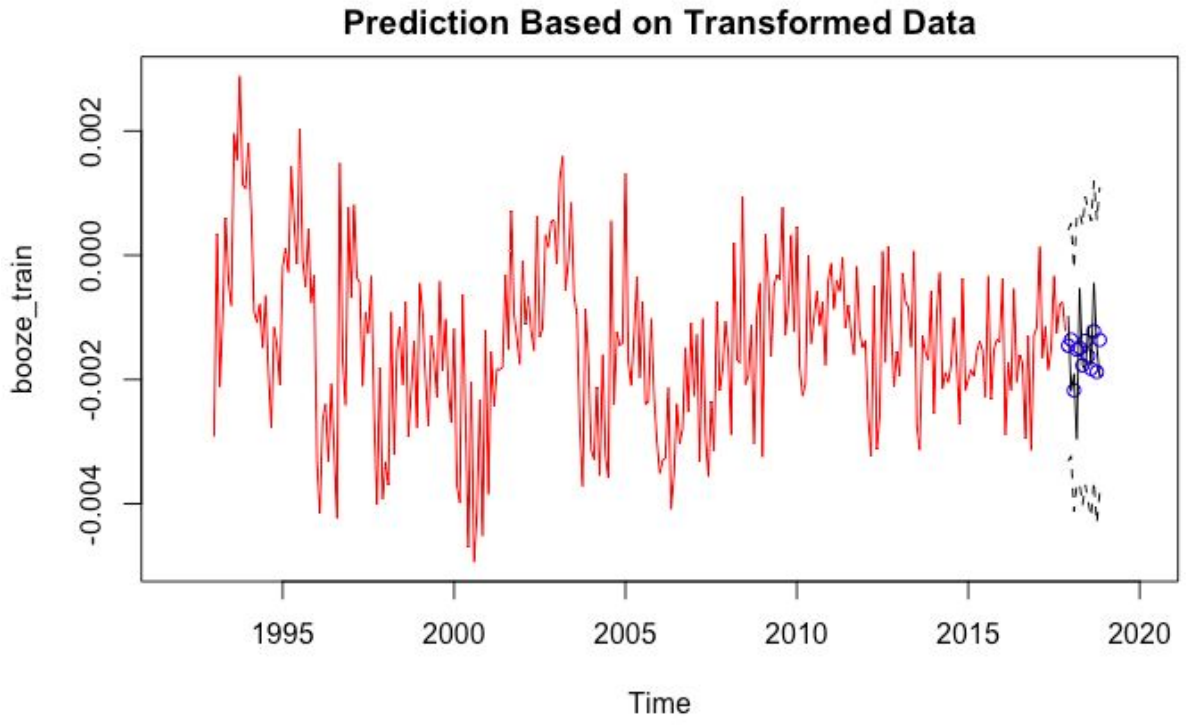
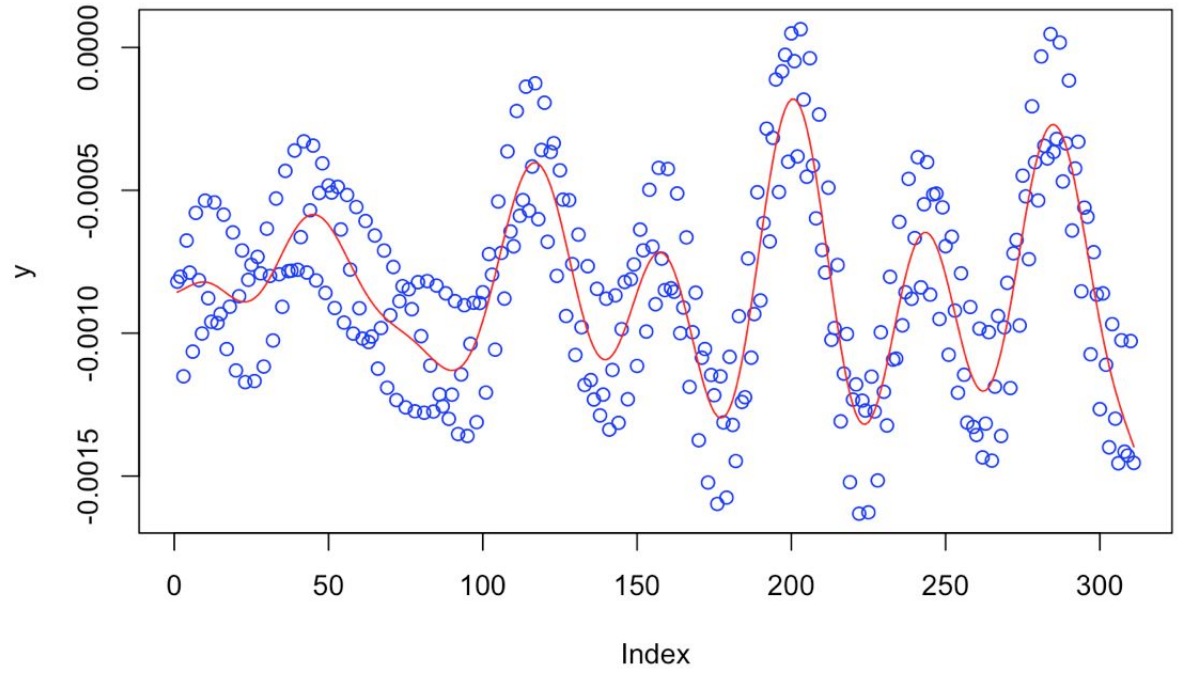


Figure 3



Appendix: Code

```
knitr::opts_chunk$set(echo = T, warning=F, message=F, results='hide')
library(tseries)
library(tidyverse)
library(forecast)
library(plotrix)
library(MASS)
library(astsa)
library(GeneCycle)
library(TSA)
library(car)
library(lattice)

booze <- read.csv("booze.csv", header = TRUE)
str(booze)
head(booze)

#Convert Data to time series
booze <- ts(booze[,2],frequency = 12, start = c(1992,1)) #Monthly data points
var_og <- var(booze)

#Stabilize Variance
#Plot raw time series
ts.plot(booze, main = "Monthly Retail Sales: Beer, Wine, and Liquor Stores" )

time <- 1:length(booze)
lin.model <- lm(booze ~ time)

predicted <- predict(lin.model)
residuals <- residuals(lin.model)
xy <- data.frame(time,booze,predicted,residuals)

ggplot(xy, aes(x = time, y = booze)) +
 geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
 geom_segment(aes(x = time,y = booze,xend = time, yend = predicted), alpha = .5) +
 geom_point(aes(x = time, y = booze, fill = "Actual"), size = 0.4)+
 ggtitle("Linear Model Intuition for Slight Heteroscedacity")

qqnorm(residuals, pch = 1, frame = FALSE) #QQ plots
qqline(residuals, col = "steelblue", lwd = 2)

#Do Box-Cox
transforms <- boxcox(booze~c(1:length(booze)))
which_transform <- data.frame(transforms)%>%
 filter(y == max(y))

which_transform <- which_transform$x
BoxCox <- function(data, lambda){
 data_transformed <- data^lambda
 return(data_transformed)
}
```

```
#Transform the time series
booze2 <- BoxCox(booze, which_transform)

#Examine the effects
ts.plot(booze2, main = "Monthly Retail Sales: Beer, Wine, and Liquor Stores After Box-Cox
Transform" )

time <- 1:length(booze2)
lin.model2 <- lm(booze2 ~ time)

predicted2 <- predict(lin.model2)
residuals2 <- residuals(lin.model2)
xy2 <- data.frame(time,booze2,predicted2,residuals2)

ggplot(xy2, aes(x = time, y = booze2)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  geom_segment(aes(x = time,y = booze2,xend = time, yend = predicted2), alpha = .5) +
  geom_point(aes(x = time, y = booze2, fill = "Actual"), size = 0.4)+
  ggtitle("Linear Model After Box-Cox")+
  ylab("Price Transformed")


var_boxcox <- var(booze2)

qqnorm(residuals2, pch = 1, frame = FALSE)
qqline(residuals2, col = "steelblue", lwd = 2)


decomposed_model <- stats::decompose(booze2, type = "additive")

autoplot(decomposed_model, main = "Additive Decomposition Plot")+
  xlab("Time in Months")



#Examine the data for seasonality
seasonplot(booze2, 12, col = rainbow(12),year.labels = TRUE, main = "Annual Seasonality Plot")
head(booze2)
var(booze2)
ts.plot(booze2, main = "Box-Cox Transformed Data")


#Remove Seasonality
booze_diff_seasonality <- diff(booze2, lag = 12)
ts.plot(booze_diff_seasonality, main = expression(paste(nabla[12],X[t])), ylab = "Seasonality
Removed")
var_season_removed <- var(booze_diff_seasonality)
```

```
#Remove trend
booze_stationary <- diff(booze_diff_seasonality, lag = 1)
#ts.plot(booze_diff_seasonality, main = expression(paste(nabla,nabla[12],X[t])))

var_stationary_diff1 <- var(booze_stationary)
#Check if more differencing helps
once_more <- diff(booze_stationary, lag = 1)
var_stationary_diff2 <- var(once_more)
two_more <- diff(once_more, lag = 1)
var_stationary_diff3 <- var(two_more)
variances <- c(var_boxcox,var_season_removed,var_stationary_diff1,var_stationary_diff2,
var_stationary_diff3)
labels1 <- c( expression(X[t]), expression(paste(nabla[12], X[t])), expression(paste(nabla,
nabla[12], X[t])), expression(paste(nabla^2, nabla[12], X[t])), expression(paste(nabla^3, nabla[12],
X[t])))

#check difference in variance
barplot(variances,names.arg = labels1, col = "gold", xlab = "Time Series", ylab = "Variance", main =
"The Effect of Differencing")

#dickey fuller test
adf.test(booze2)

#Make Train and Test Sets
par(cex.main = 0.8)
booze_data <- ts(booze_diff_seasonality,frequency = 12, start = c(1992,1)) #Monthly data points
#Split data into train and test sets.
booze_train <- head(booze_diff_seasonality, length(booze_diff_seasonality) - 12)
booze_test <- tail(booze_diff_seasonality,12)

# Define Acf and Pacf
acf_seasonality <- Acf(booze_diff_seasonality, lag.max = 72, main = "Sample ACF", plot = FALSE)
pacf_seasonality <- Pacf(booze_diff_seasonality,lag.max = 72, main = "Sample PACF", plot= FALSE)

acf_seasonality <- data.frame(ACF = acf_seasonality$acf, lag = acf_seasonality$lag)
pacf_seasonality<- data.frame(PACF = pacf_seasonality$acf, lag = pacf_seasonality$lag)

# Plot ACF
ggplot(acf_seasonality, aes(x = lag, y = ACF, fill = if_else(lag%%12 == 0, "blue", "red")))+
  geom_col()+
  geom_hline(yintercept = 0.1)+
  geom_hline(yintercept = -0.1)+
  theme(legend.position="none")+
  ggtitle("Seasonal ACF (Stationary)")
```

```
# Plot PACF
ggplot(pacf_seasonality, aes(x = lag, y = PACF, fill = if_else(lag%%12 == 0, "green", "red")))+
  geom_col()+
  geom_hline(yintercept = 0.1)+
  geom_hline(yintercept = -0.1)+
  theme(legend.position="none")+
  ggtitle("Seasonal PACF (Stationary)")

# Set up for smaller graph
acf_stationary2 <- Acf(booze_diff_seasonality, lag.max = 12, main = "Sample ACF", plot = FALSE)
pacf_stationary2 <- Pacf(booze_diff_seasonality,lag.max = 12, main = "Sample PACF", plot= FALSE)

acf_stationary2 <- data.frame(ACF = acf_stationary$acf, lag = acf_stationary$lag)
pacf_stationary2 <- data.frame(PACF = pacf_stationary$acf, lag = pacf_stationary$lag)

# Plot ACF Close Up
ggplot(acf_stationary2, aes(x = lag, y = ACF, fill = if_else(lag%%3 == 0, "blue", "red")))+
  geom_col()+
  geom_hline(yintercept = 0.1)+
  geom_hline(yintercept = -0.1)+
  theme(legend.position="none")+
  ggtitle("Interseasonal ACF")

# Plot Pacf Close Up
ggplot(pacf_stationary2, aes(x = lag, y = PACF, fill = if_else(lag%%3 == 0, "blue", "red")))+
  geom_col()+
  geom_hline(yintercept = 0.1)+
  geom_hline(yintercept = -0.1)+
  theme(legend.position="none")+
  ggtitle("Interseasonal PACF")


decomposed_model_train <- stats::decompose(booze_train, type =  "additive")

autoplot(decomposed_model_train, main = "Additive Decomposition Plot")+
  xlab("Time in Months")

# Model Selection
sarima(booze2, 3,0,0,1,1,2,12,details = FALSE, Model = FALSE)$AIC
sarima(booze2, 3,0,1,1,1,2,12,details = FALSE, Model = FALSE)$AIC
sarima(booze2, 3,0,2,1,1,2,12,details = FALSE, Model = FALSE)$AIC
sarima(booze2, 3,0,3,1,1,2,12,details = FALSE, Model = FALSE)$AIC

sarima(booze2, 3,0,0,1,1,2,12,details = FALSE, Model = FALSE)$AICc
sarima(booze2, 3,0,1,1,1,2,12,details = FALSE, Model = FALSE)$AICc
sarima(booze2, 3,0,2,1,1,2,12,details = FALSE, Model = FALSE)$AICc
sarima(booze2, 3,0,3,1,1,2,12,details = FALSE, Model = FALSE)$AICc
```

```
sarima(booze2, 3,0,0,1,1,2,12,details = FALSE, Model = FALSE)$BIC
sarima(booze2, 3,0,1,1,1,2,12,details = FALSE, Model = FALSE)$BIC
sarima(booze2, 3,0,2,1,1,2,12,details = FALSE, Model = FALSE)$BIC
sarima(booze2, 3,0,3,1,1,2,12,details = FALSE, Model = FALSE)$BIC

# Check for causality/invertibility
source("plot.root.R")
par(mfrow=c(1,3))
plot.roots(ma.roots = NULL, polyroot(c(1, 0.13, 0.28, 0.42)), main="Roots for AR part")
plot.roots(ma.roots = NULL, polyroot(c(1, 0.34)), main="Roots for Seasonal AR part", size = 3)
plot.roots(ar.roots = NULL, polyroot(c(1, -1.98, 0.99)), main="Roots for Seasonal MA part", size = 1)

# Fit the Model using MLE
booze_fit <- arima(booze_train, order = c(3,0,0), seasonal = list(order = c(1,1,2), period = 12),
method = "ML")

booze_fit


#Now that we have fit the model, we need to check diagnostics to ensure our #assumptions on the
model are valid. These assumptions are:
    #1.  Normality of Errors
    #2.  No Serial Correlation
    #3.  Homoskedasticity

    #Get residuals
    residuals <- booze_fit$residuals

    #Plot QQ
    qqnorm(residuals)
    qqline(residuals, col = "dodgerblue", lwd = 2)

    #estimate standard deviation using the fact that 99.9% of data between 3 standard devations
    sigma_hat = (max(residuals) - min(residuals))/6

    #Plot a histogram of the residuals
    hist(residuals, probability = TRUE)

    #overlay the estimated normal distribution onto the residuals
    curve(dnorm(x,0,sigma_hat), xlim = c(-0.02,0.02),add=TRUE, yaxt="n", lwd = 2, col =
        "dodgerblue")
    legend("topright", legend=c("Normal(0,0.00405)"),
        col=c("dodgerblue"), lty=c(1),lwd = c(2), cex=0.8)

    #2 check for serial correlation
    acf(residuals, lag.max = 300)
```

```
#Perform Ljung-Box test
Box.test(residuals, type = "Ljung-Box")

#3 check Homoskedacity
pacf(residuals2, lag.max = 120, main = "")
title(expression(Errors^2), line = 0.6)

#predictions on transformed data
pred_Tr= predict(booze_fit, n.ahead = 12)

#upper and lower bounds by point estimate +- 2* standard error since 2-approx 1.96 Z(0.025)
Up_tr = pred_Tr$pred+2*pred_Tr$se
Do_tr = pred_Tr$pred-2*pred_Tr$se
preds = pred_Tr$pred

#plot series and  predictions
ts.plot(booze_train, col = "red",  main= "Prediction Based on Transformed Data", xlim =
      c(1992,2020))
points( preds,col='blue')
lines(Up_tr, lty="dashed")
lines(Do_tr, lty="dashed")
lines(booze_test)


#FORECASTING
#create train/test untransformed data
booze_train_OG <- head(booze, length(booze)  - 12)
booze_test_OG <- tail(booze, 12)
#refit model to original non-transformed data
test_OG <- arima(booze_train_OG, order = c(3,0,0), seasonal = list(order = c(1,1,2), period =
      12), method = "ML")

preds_nonstationary <- predict(test_OG, n.ahead = 12)

#Set up Confidence Bounds
upper <- preds_nonstationary$pred+2*preds_nonstationary$se
lower <- preds_nonstationary$pred-2*preds_nonstationary$se

#predictions
preds = preds_nonstationary$pred

# Plot of forecasting original data
ts.plot(booze_train_OG, col = "red",  main= "Prediction Based on Original Data", xlim =
      c(1992,2020))
points( preds, col='blue')
lines(upper, lty="dashed", col = "green")
lines(lower, lty="dashed", col = "green")
```

```
lines(booze_test_OG)

# Plot of close up  of forecasting original data
ts.plot(booze_train_OG, col = "red",  main= "Prediction Based on Original Data CLOSE UP", xlim
      = c(2016,2019))
lines(booze_test_OG, col="black")
points(preds,col='blue')
lines(upper, lty="dashed", col = "green")
lines(lower, lty="dashed", col = "green")


#SPECTRAL ANALYSIS

periodogram(booze_diff_seasonality, main= 'Periodogram on the Stationary Data') #find the
      frequencies
abline=0
ch = periodogram(booze_diff_seasonality)
# grabs the top 8 frequencies since after checking this had the most 'significant'
FREQ = ch$freq[order(ch$spec, decreasing=T)][1:8] FREQ #top 8 are FREQ

t = 1:311 #size of dataset
w=2*pi*t #w is always this
x1 = cos(w*FREQ[1])
x2= sin(w*FREQ[1])
x3 = cos(w*FREQ[2])
x4= sin(w*FREQ[2])
x5 = cos(w*FREQ[3])
x6= sin(w*FREQ[3])
x7 = cos(w*FREQ[4])
x8= sin(w*FREQ[4])
x9 = cos(w*FREQ[5])
x10= sin(w*FREQ[5])
x11= cos(w*FREQ[6])
x12= sin(w*FREQ[6])
x13 = cos(w*FREQ[7])
x14= sin(w*FREQ[7])
x15= cos(w*FREQ[8])
x16= sin(w*FREQ[8]) #these are the x's for the different frequencies
z =
      lm(booze_diff_seasonality~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x1
      5+x16) #z is our fit
summary(z)
#to get our adjusted R squared value
coeffs=lm(booze_diff_seasonality~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x
      15+x16)$coeff #get the coefficients
resid=lm(booze_diff_seasonality~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x
      15+x16)$resid #grabbing the residuals to be tested later on fisher and kolmogorov test
```

```
y =
      coeffs[1]+coeffs[2]*x1+coeffs[3]*x2+coeffs[4]*x3+coeffs[5]*x4+coeffs[6]*x5+coeffs[7]*x
      6+coeffs[8]*x7+coeffs[9]*x8+coeffs[10]*x9+coeffs[11]*x10+coeffs[12]*x11+coeffs[13]*x
      12+coeffs[14]*x13+coeffs[15]*x14+coeffs[16]*x15+coeffs[17]*x16
#this is the final spectral model

op = par(mfrow=c(2,1))


plot(t,y, type='l')
plot(booze_diff_seasonality, ylab= "Stationary Data") #plotting the two models
par(op)
spline <- smooth.spline(c(1:length(y)), y) # you choose lambda
plot(y, col = "blue")
lines(1:length(y), predict(spline, x = 1:length(y))$y, col = 'red')
#graphs of our stationary data vs the spectral data

# CHECK for periodicity. Test on stationary model
cpgram(booze_diff_seasonality) #Kolmogorov-Smirnov test
fisher.g.test(booze_diff_seasonality) #fisher test

# CHECK residuals to make sure they are Gaussian White Noise. Same test on residuals
cpgram(resid) #Kolmogorov-Smirnov test
#fisher test
fisher.g.test(resid) #this is greater than .05 so this passes
```