Assignment 8: Final – BJA Report – Box-Jenkins Model: BJA

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

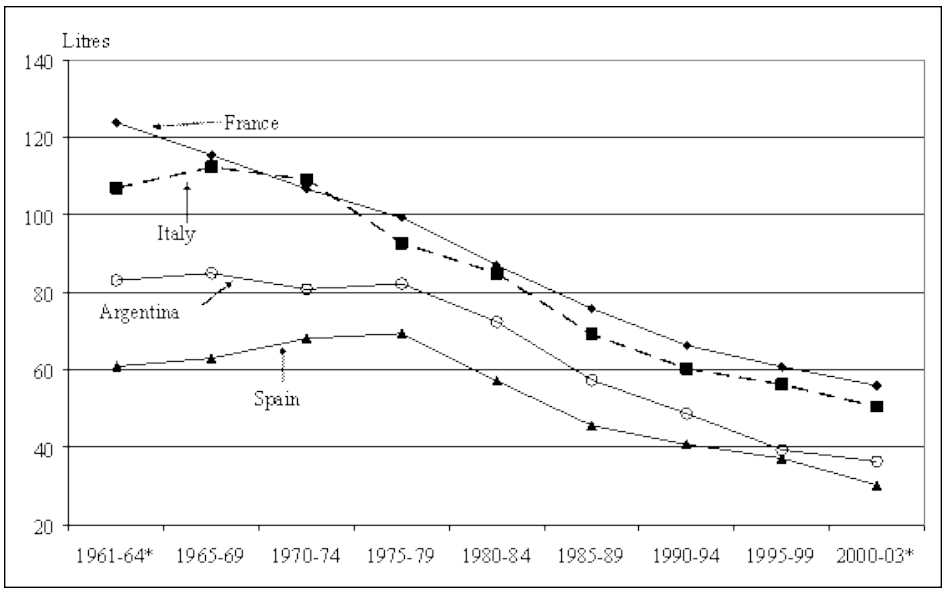
Eden Prairie, MN 55344

Executive Summary

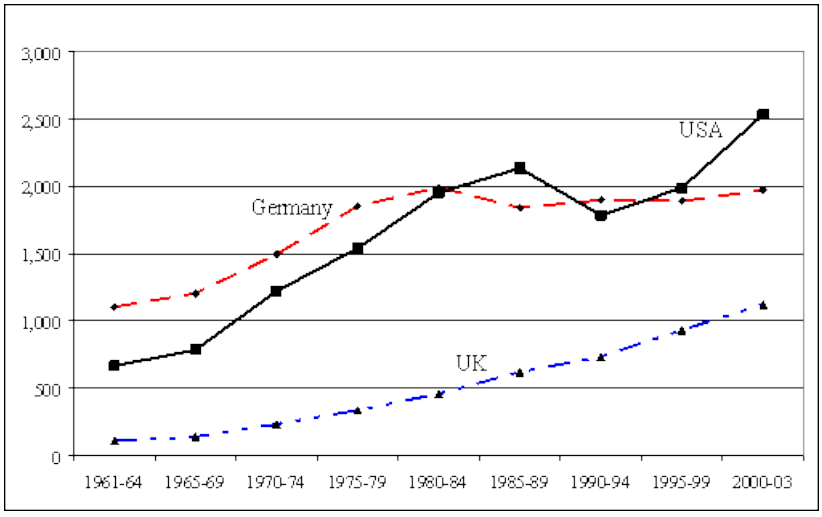
Over the past 20 years, Australia has dominated the wine export business over any other country. This is due in part to the fact that the US, Germany, and the UK have all increased their consumption of wine over the past 20 years. The objective of this exploratory data analysis (EDA) is to utilize the Box-Jenkins methodology to analyze the wine data set, fit an appropriate model, and forecast 10 periods ahead. In its original form, the data had a large variance and the log transformation was used to assuage this affect. The time series was not stationary and displayed distinct seasonality. Differencing transformed the time series such that both issues were placated, and involved differencing at the first lag as well as at the 12^{th} lag, in order to remove the seasonality. The time series data was still not ideal with a few of the autocorrelation function (ACF) plots outside of the confidence ban, but it was sufficient in order to move on to model selection. Given the statistical diagnostics as well as adherence to parsimony, it was found that a (0,1,1) X (0,1,1) – Moving Average with time lag one with differencing at the first and 12^{th} month model proved to be the best model. The forecast shows an initial dip in wine exportation, but a continual increase over time such that it exceeds its original point of entry.

Introduction

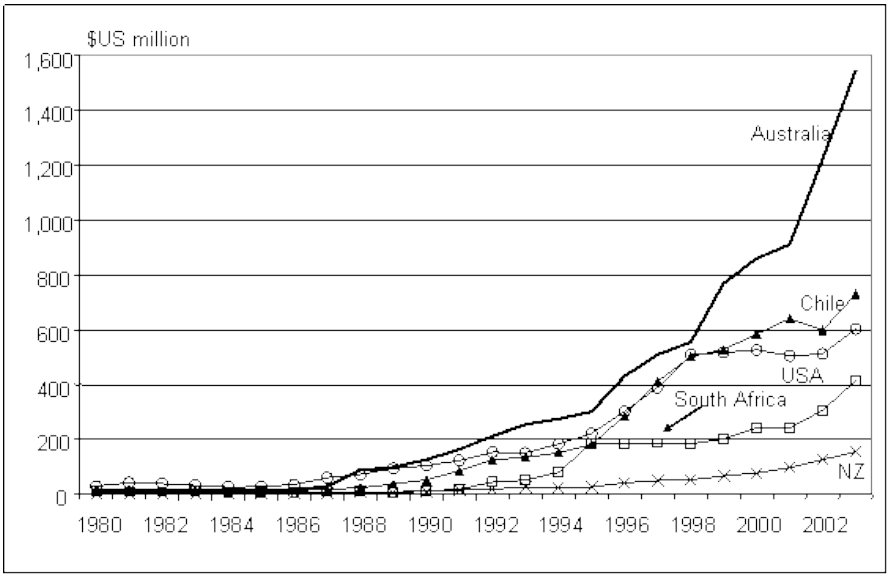
Global consumption for wine has steadily increases over the past ten years. The world's largest wine producing countries have experienced an overall decline in wine consumption.



These two factors have driven down the global price for grapes, as well as increased the exportation of wine from traditional wine consuming countries to new markets (Wittwer & Rothfield). The graph above shows the consumption rates of wine (liters) for the four largest wine producing countries. It can be seen that there has been a negative trend of wine consumption for the last 50 years in the traditional wine consumer countries. From the 1960's on, the world shifted to a more inclusive trading platform, which resulted in additional spirits in these countries. On the global stage, the USA, Germany, and the UK all increased their consumption of wine and importation of wine during the same time period. Essentially the two graphs show



that while wine consumption decreased in France, Italy, Spain, and Argentina, it increased in the US, UK and Germany. While new markets were expanded into over the past 40 years, it is vitally important to analyze which countries exported the most wine during this time period. The graph to the left shows which countries exportation of wine grew the largest over the past 20years. It can clearly be seen that Australia, Chile, and the US have been winning the export war of wine. Analyzing Australia's production and exportation of wine is a helpful exercise in understanding successful export strategies.



Given the preliminary information above, I expect that as time increases so too does wine consumption as well as exportation for the Australian wine industry. The dataset in this EDA records the monthly sales of Australian wine makers. Specifically, the objective for this EDA is to utilize the Box-Jenkins methodology to analyze the wine data set, fit an appropriate model, and forecast 10 periods ahead.

Analysis

In order to meet the objective of exploring and predicting the relationship between Australian red wine sales and time, an exploratory data analysis (EDA) must be conducted. . This EDA will start with a basic Time Series Plot of wine sales versus time. From that plot, analysis

will be drawn and an appropriate model will be fit to the data. As a data scientist in training, I am inculcating a paradigm of which to study data. While this paradigm is redundant report to report, it is training me to have the correct mindset. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

Problem/Objective: Explore the relationship between wine sale and time in months, and forecast 10 periods ahead.

Data: The data has been aggregated and has been supplied from management. There are no missing values.

Analysis: I will describe the data via simple descriptive statistics at first. After the initial analysis, I will analyze the data via time series plot. The goal of the analysis is to lead the EDA to an appropriate model.

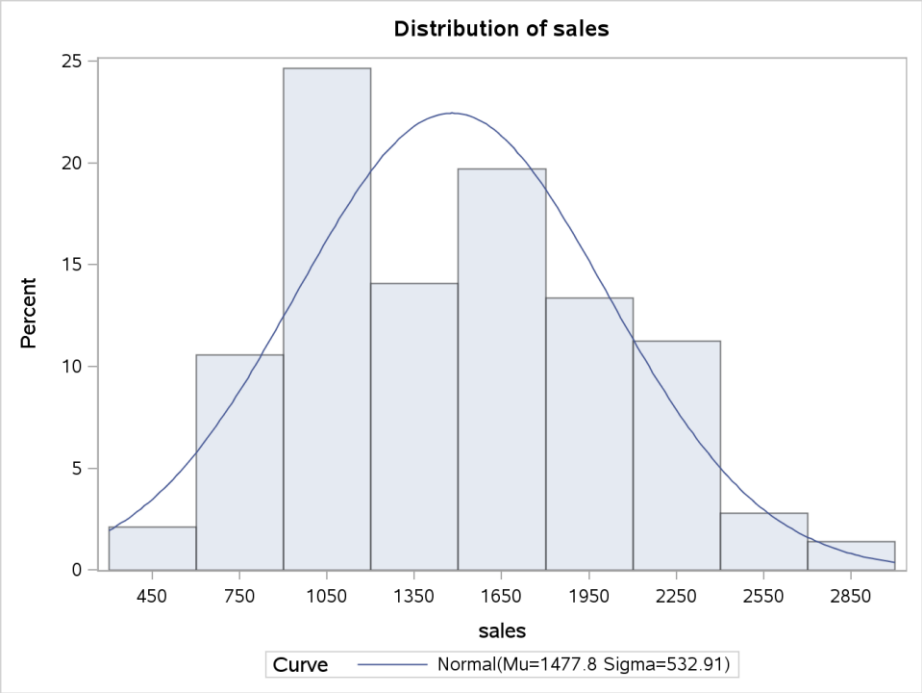
Model: Proc ARIMA in SAS will be used in order to identify the appropriate model..

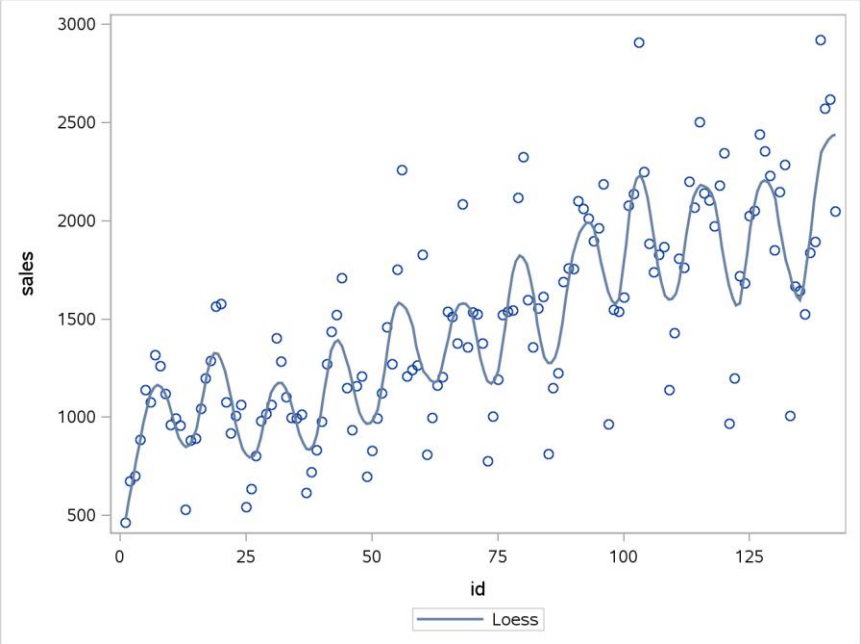
Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the model fits that data and the statistical backing of the model. The final step will be forecasting ten periods ahead.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best model, and the analyst's personal bias is mitigated.

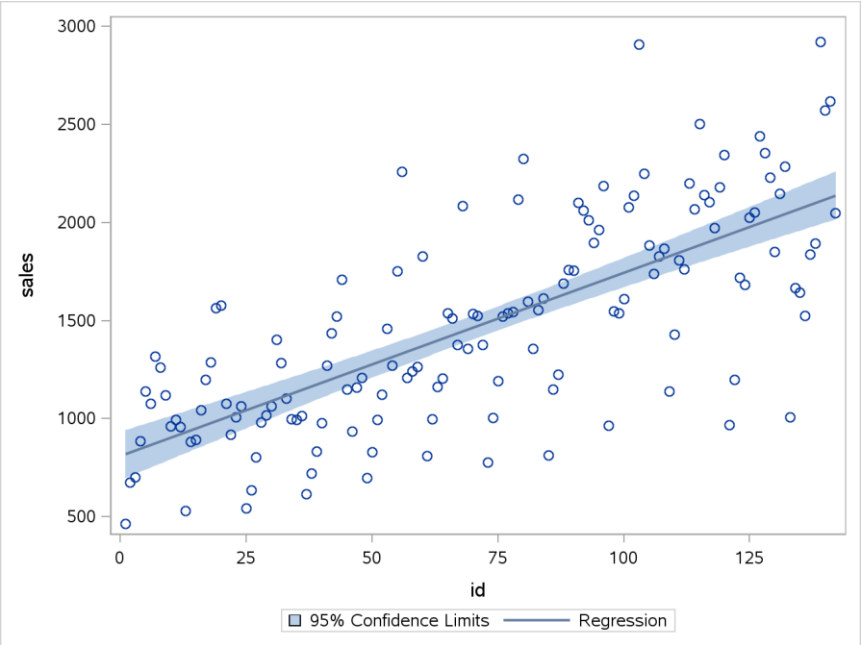
Data

There are a total of 142 observations with 0 completely missing values per row. The data is comprised of one variable, Kiloliters_in_1000_s, and represents the volume of Australian red wine sold for the time period between 1980 through 1991. Each year has 12 recorded observations at one month intervals starting in January first 1980. This is the first dataset where time has its own column, which directly correlates to the month and is expressed in the id column. The sales column equals the Kiloliters in 1000's column, and is easier to understand simply as sales. The average value for wine sales per month is 1,477.768 thousand kiloliters, and a median of 1433.5, which demonstrates a slight positive skew. Standard deviation for this data set is 532.914 with a range of 2456, and a positive skew of .3936. In my mind, it is easier to understand standard deviation as 66% of the sales fall between 945 and 2,009 thousand kiloliters. Overall, the sales are fairly normally distributed, and the data as a whole is slightly skewed. The distribution of wine sales can be visually seen from the histogram. Around 25% of the sale volume was around 1050 thousand kiloliters. The distribution shows a slight positive skew, but as a whole has a solid distribution. At this point, the data set is in fairly good shape and should not require any initial major transformations.

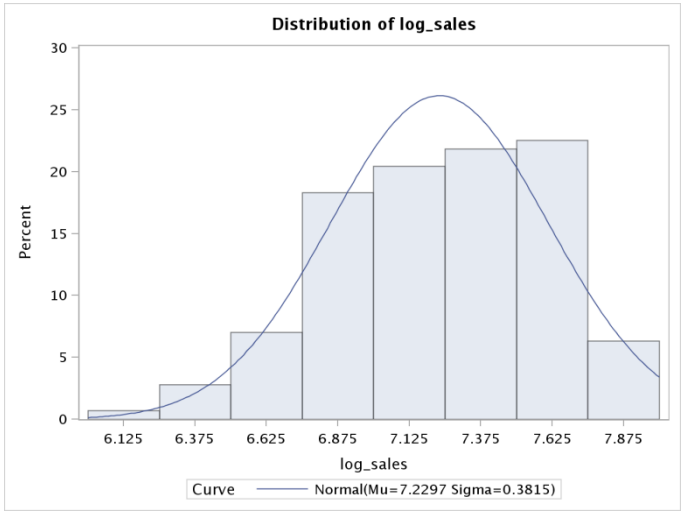


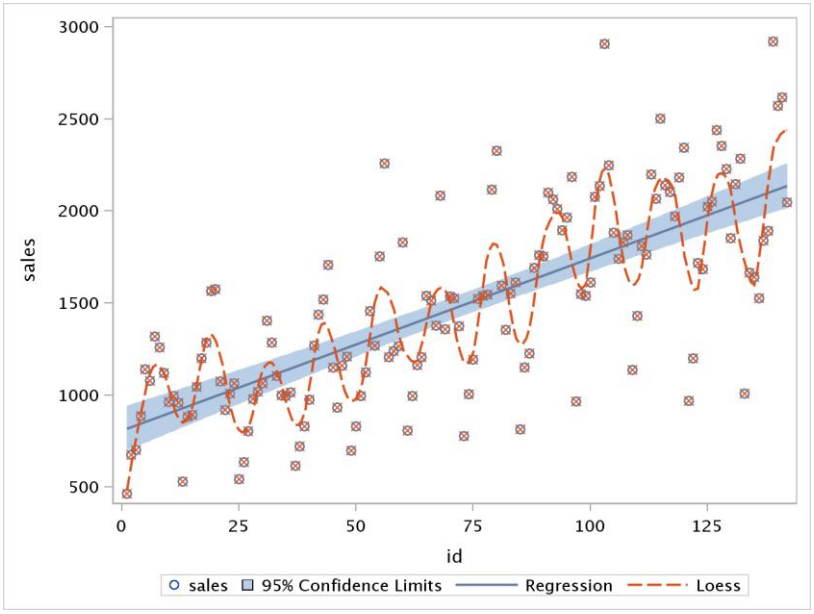


To the left, one can see the scatter plot of wine sales and time. The plots are hard to discern a pattern from, but the Loess smoother is helpful for identifying the overall trend. In regression analysis, the locally weighted scatterplot smoothing (Loess) technique combines least squares regression with localized subsets of data to delineate variation and produce a helpful trend line. It can be seen from the Loess smoother that the series is not stationary and exhibits seasonality by the wavy nature. The regression line, as seen below, also shows an upward trend. From the initial data analysis, the core descriptive statistics help frame the data, and the histogram and scatterplot visually demonstrated the distribution and initial interaction between time and wine sales. Before fitting a time series model, this data will have to be stationary and rid of the seasonality which is going to require a few data transformations.



The log transformation is a technique that helps the data follow a more normal distribution. After the log transformation is performed, the average value for wine sales per month is 7.230 thousand kiloliters, and a median of 7.230, which demonstrates a relatively balanced distribution. Standard deviation for this data set is .381 with a range of 1.840, and a slight negative skew of .372. Overall, the sales are fairly normally distributed, and the data as a whole is slightly skewed. The distribution of wine sales can be visually seen from the histogram. Around 80% of the sales volume is between 6.875 and 7.625 thousand kiloliters. The distribution shows a slight negative skew, but as a whole has a solid distribution. Below, one can see the scatter plot of wine sales and time.





It can be seen from the Loess smoother that the series is not stationary and exhibits seasonality by the wavy nature. The regression line, as seen below, also shows an upward trend. The log transformation appears to have calmed the precipitous variation and I will

use this data set in the EDA, but it first must be stationary and rid of the seasonality before fitting a model.

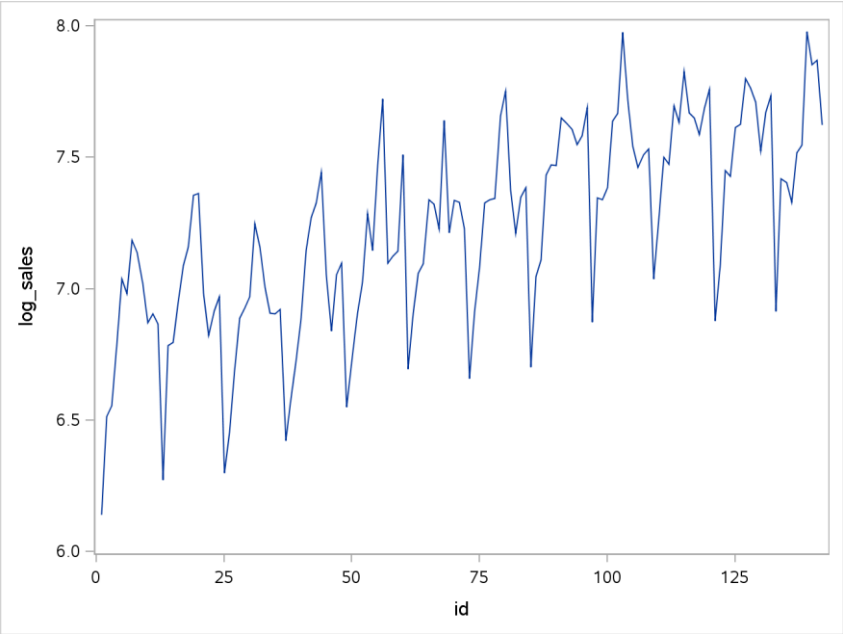
Results

Modeling time series data utilizing Box-Jenkins techniques follows four main steps:

1. Model Identification
2. Model Estimation
3. Diagnostic Checking
4. Forecasting

The last three steps are nothing new in regard to regression modeling techniques. Step 1 on the other hand is a new technique and will require background information before moving forward with the EDA.

Model Identification is centered on the acronym ARIMA which means Autoregressive Integrated Moving Average. Within this acronym, there are three models that can be used to fit time series data. Before delving into the models, the one major assumption of ARIMA is that the data has been transformed into a stationary time series, which means the data has a mean of zero and no trend overtime. As seen above, the data is not stationary and there is a seasonal upward trend. As a result of these findings, I have conducted a log transformation as seen to the left. Despite the log transformation, it can clearly be



seen there is still a positive seasonal trend, which violates the first assumption of utilizing Box-Jenkins modeling.

One might be inclined to skip transforming the data into a relatively stationary series, but there are serious ramifications for not stationarizing such as:

- The logic that the statistical properties in the past will be the same for the future is rendered invalid.
- Without stationarizing, an appropriate model cannot be chosen.
- The descriptive statistics cannot be compared to other variables, and furthermore without stationarization future forecasts are malarkey (Duke.edu).

In conclusion, making the time series data stationary is necessary for modeling with efficacy. As demonstrated above, the log transformation did not stationarize the data. Differencing is a technique that transforms the time series to a different time series. For example, X(t) equals the variable X over specific intervals. When differencing, x(t) is transformed to d(t) and d(t) equals the difference between consecutive values of x(t). This is considered the first difference. The second difference is d(1) (t) – d(1) (t-1), and can progress on to additional differences (statistics.com). Determining the order of differencing is the next step, and also one of the most important in defining the model. According to Duke University, the goal of differencing is to choose the lowest order coupled with a stationary mean where the time series fluctuates around zero and the ACF decline to zero quickly. Listed below are the rules of thumb taken from Duke University:

> Rule 1: If the series has positive autocorrelations out to a high number of lags, then it probably needs a higher order of differencing.
>
> Rule 2: If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and pattern less, then the series does not need a higher order of differencing. If the

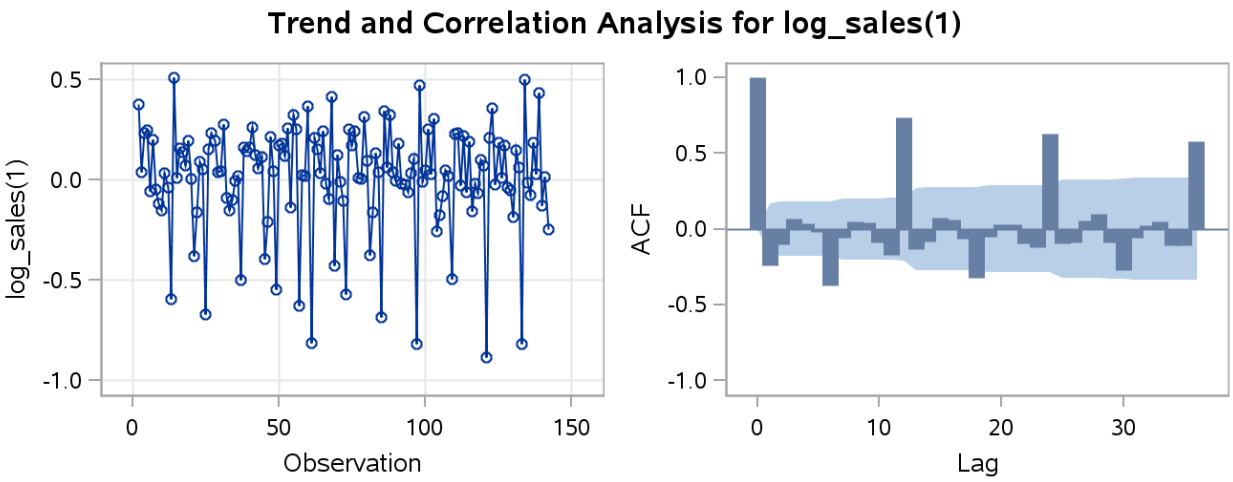
lag-1 autocorrelation is -0.5 or more negative, the series may be over differenced. BEWARE OF OVERDIFFERENCING!!

Rule 3: The optimal order of differencing is often the order of differencing at which the standard deviation is lowest.

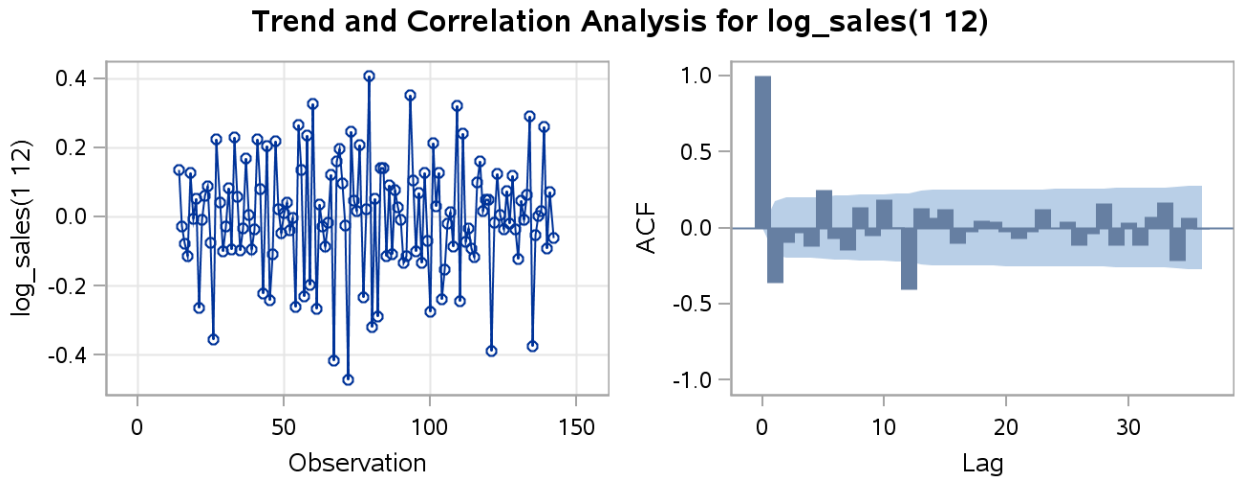
Rule 4: A model with no orders of differencing assumes that the original series is stationary (mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth). A model with two orders of total differencing assumes that the original series has a time-varying trend (e.g. a random trend or LES-type model).

An initial Proc Arima in SAS shows that the data is not stationary, the ACF plot is systematically seasonal and the white noise test shows autocorrelation, but it is pointless to draw a conclusion from the white noise test because the data is not stationary (JamesMadisonUniversity.edu). Once the time series is stationary, I will come back to the white noise test to validate that the data is not white noise, which tests whether there is actually any useful data to model.

After the first differencing, the model has the following scatter plot and ACF.



The good news with the first differencing is the mean hovers around zero, but there is still a trend in the data. I identify the seasonality from the scatter plot based on the observed minimums and maximums overtime that form repeating patterns. The seasonality appears to have a 12 month cycle, thus adding another differencing effect at the 12th lag might get rid of the seasonality. Bear in mind that there is no need to difference further in regard to the mean based on the principle of parsimony and the explanation above.

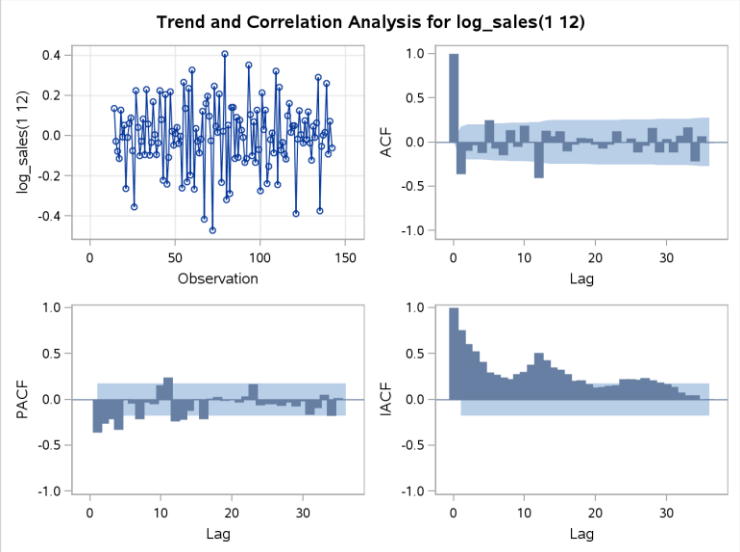


After differencing at the 1 and 12 periods, the scatterplot still looks seasonal, but the ACF plot has all but three values within the confidence ban.

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	29.93	6	<.0001	-0.362	-0.098	-0.034	-0.124	0.249	-0.072
12	64.97	12	<.0001	-0.149	0.137	-0.056	0.189	0.007	-0.406
18	72.45	18	<.0001	0.130	0.064	0.124	-0.105	-0.026	0.048
24	76.15	24	<.0001	0.043	-0.029	-0.071	-0.031	0.122	-0.005
30	85.91	30	<.0001	0.045	-0.119	-0.041	0.158	-0.119	0.038
36	103.27	36	<.0001	-0.114	0.074	0.166	-0.219	0.066	-0.008

The autocorrelation check for white noise statistically demonstrates that we can reject the null hypothesis that the autocorrelations in all the lags are jointly zero. While the stationarity is not ideal, for the purposes of building this model it will suffice. The next step is to fit the model.

When fitting a model to time series data, a specific nomenclature is followed to keep track of the specific model. The ARIMA model has three factors involved, (p,d,q), where p= the order of autoregressive components that are statistically significant, d= the number of differencing that has been done on the data, and q= the order of the moving average (JMU.edu). In the model selection, the nomenclature will follow a (p,d,q) list of the ideal model selected. Analyzing the inferential statistics AIC is an informal method to assess the model fit. This statistic can be used to compare different sets of variables. Higher values mean a worse fit to the data. As in Ordinary Least Squares, AIC is used to penalize models that have more variables. The ARMA model is comprised of terms from both AR and MA. Now the questions of how many terms get included as well as which model is used are answered. Depending on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) values, a specific model will be used. Both of these functions are similar to what the r-value represents in OLS. Visually the ACF and PACF are great ways to discern if an AR, MA ARMA model is appropriate. The similarity between AR and MA is the use of lagged terms, but the difference is the AR model uses the lags from the actual timer series where the MA model uses lags from the noise or residuals (colorado.edu). Similar to OLS, the goal of fitting the correct model is a minimized sum of square errors, and statistical validation. Depending on the graphical output and diagnostics, either an AR, MA, or ARMA model will be used. I will fit all the models and work through the output to discern the best model. The initial correlation analysis, to the left, points to an MA model. When the ACF has one



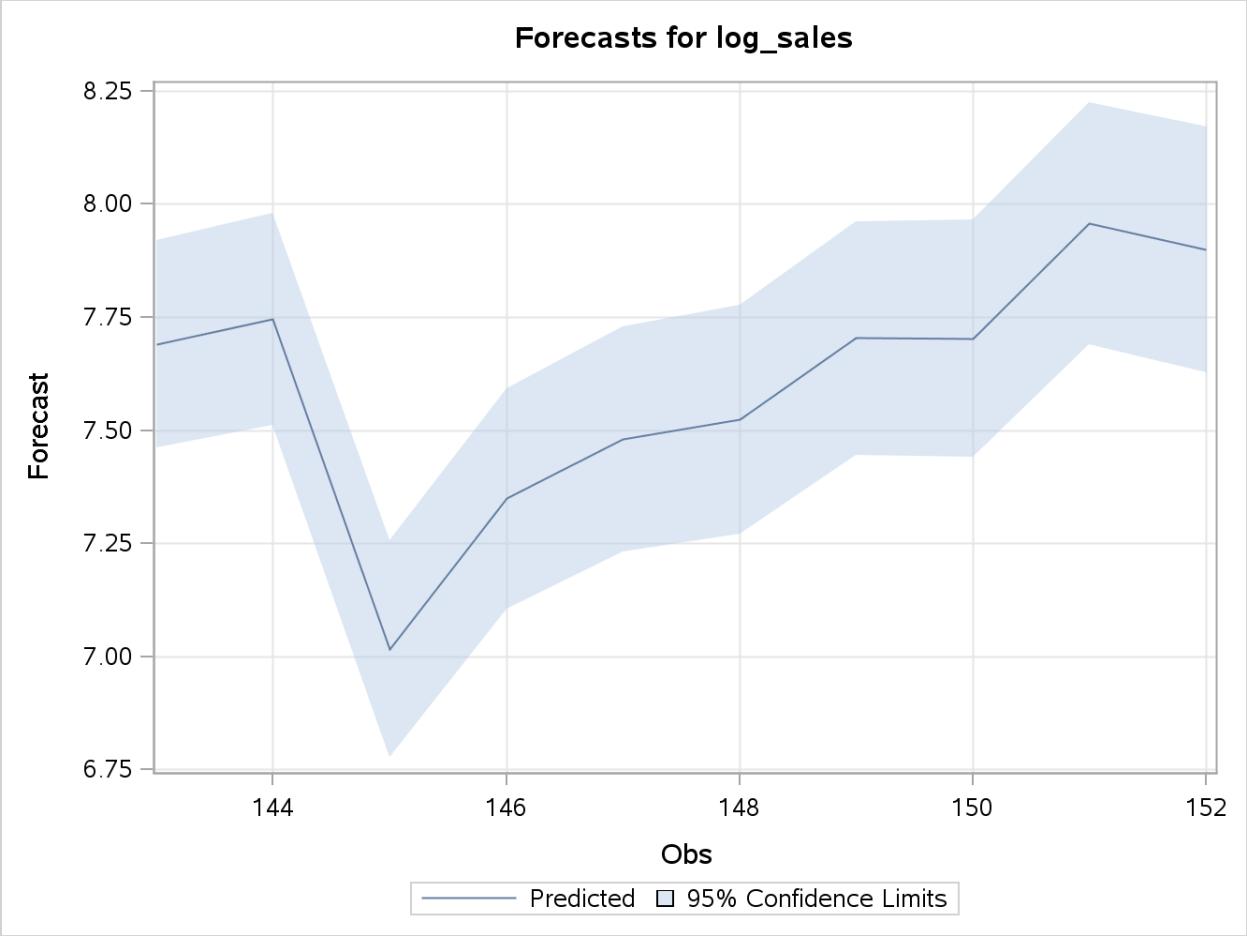
large spike, and the PACF diminishes over time, this points to utilizing the MA model. Bear in mind that through this modeling phase a small sum of squares errors is desired.

The first model will follow the nomenclature of a (1,1,0) X (1,1,0), which means that this is a 1 lag autoregressive model with one differencing period with another autoregressive lag term with another differencing component at the twelfth month. Appendix 1 shows the total output for this model and I will highlight the key components. The variance is .018 and the AIC is -147.262. In addition to these values, the model is statistically significant, but the autocorrelation check of the residuals show that important information has been left. As a result of the autocorrelations, this model can be improved upon.

An ARIMA model is comprised of (1,1,1) X (1,1,1) and Appendix 2 has the total output. Utilizing conditional least squares the autoregressive coefficients are not statistically significant. The variance is .013 and the AIC is -181.84, and the residuals pass for white noise. While these values are better than the AR model, the total ARIMA model is not statistically significant and a Moving Average (MA)_model will be fit.

Bingo, the MA model is the most parsimonious as well as statistically valid. Appendix 3 lists the full diagnostics. The model is statistically significant, variance is .013, and the AIC is -184.973. In addition, the autocorrelation check of residuals shows that the remaining values are white noise.

Now that the ideal model has been produced, simply forecasting is left. Utilizing the SAS command “forecast lead 10” produces the plot below. It can be seen that the blue line is the predicted value, and the light blue band is the 95% confidence limit.



The entire EDA was focused on creating the forecast plot above, but notice that preparing the data and selecting the correct model involved the most time and research.

Future Work

Further recommendations on how this study can be improved upon are the following:

- This study was very thorough in its nature, if time series was going to be a topic for the next couple of weeks delving into other models would be helpful beyond Box-Jenkins.
- It would be interested to compare the forecast plots for the different models and comment on the differences.
- Using time series to predict future stock trends is modern alchemy for analysts, and looking into different stock data sets would provide a useful experience.

Through this initial EDA, coupled with the future work recommendations, students would continue to gather relevant experience and exposure to time-series modeling.

References

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.

Cobb. "Time Series Analysis in a Nut Shell ." *James Madison University*. James Madison University, n.d. Web. 28 Feb. 2013. <cob.jmu.edu/doylejm/ec485_s04_timeseries.ppt>.

"Identifying the Order of Differencing." *Decision 411 Forecasting*. Duke University, n.d. Web. 21 Feb. 2013. <http://people.duke.edu/~rnau/411arim2.htm>.

Ratner, Bruce. *Statistical and machine-learning data mining techniques for better predictive modeling and analysis of big data*. 2nd ed. Boca Raton, FL: CRC Press, 2012. Print.

"Statistical Glossary." *Differencing (of Time Series):*. N.p., n.d. Web. 21 Feb. 2013. <http://www2.statistics.com/resources/glossary/t/tsdiff.php>.

"Time Series Analysis." *Stochastic Processes*. Colorado University, n.d. Web. 22 Feb. 2013. <http://www.colorado.edu/geography/class_homepages/geog_4023_s11/Lecture16_TS3.pdf>.

“Previously, reversing whatever mathematical transformations were. "Stationarity and Differencing." *Decision 411 Forecasting*. Duke University, n.d. Web. 21 Feb. 2013. <http://people.duke.edu/~rnau/411diff.htm>.

Wittwer, Glyn, and Jeremy Rothfield. "Projecting the world wine market from 2003 to 2010." *Australasian Agribusiness Review* 13.1 (2005): 21. *Centre of Policy Studies, Monash University*. Web. 18 Feb. 2013.