# Assignment 7:  Final – PDA Report – Airlines Model: PDA

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

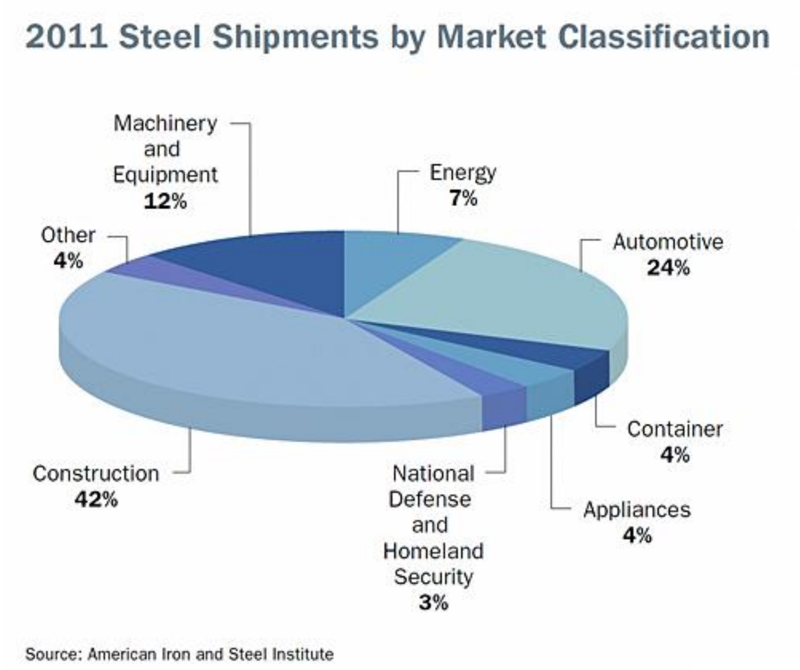
## Executive Summary

For much of the 20th century, the US was the paragon for the iron and steel industries. In 1910, it was the largest producer of steel and iron in the world with over 24 million tons produced, and reached its peak in 1969 with over 141 million tons produced. (History, John Gordon). This exploratory data analysis (EDA) studies that trend of US iron and steel exports over 44 years starting in 1937 and ending in 1980. Upon initial analysis, it was found that the time series needed was not stationary and needed one differencing in order to transform the data into a stationary time series. After analyzing the autocorrelation factors (ACF) and partial correlation factors (PACF) in conjunction with the sum square errors, it was found that the moving average (MA) model was the most parsimonious model to fit the data. From this EDA, management can utilize the model to predict future steel and iron export trends to improve their bottom line.

## Introduction

Since 2000 BC, humans have been using iron and steel to defend, capture, and build the known world (Anslem.edu). In the US, the 19th century saw the rapid expanse of steel and iron through the industrial revolution. This was propelled as a result of new low-cost advances in the refinement process. By the time the twentieth century rolled around, the US was the largest producer of steel and iron in the world with over 24 million tons produced in 1910. It is estimated the steel production reached its peak in 1969 with over 141 million tons produced (History, John Gordon).

Iron and steel are often associated with one another as a result of the refinement process. Out of all the elements on earth iron is the fourth most abundant (Anselm College). Steel is a hardened refinement of iron with carbon added at the melting stage. Thus, steel and iron are synonymous with each other in modern production.


2011 Steel Shipments by Market Classification

Source: American Iron and Steel Institute

The use of iron and steel in modern USA is inundated throughout every industry. To the left, the pie chart from AISI visually shows which industries utilize steel throughout the United States. Automotive and construction represent 66% of steel shipments throughout the US. I would expect the steel and iron industry to be very susceptible to economic growth and recession as a result of the major industries that comprise its major shipments.

The objective of this report is to explore time series analysis methods by conducting an EDA on the Steel Export dataset. I expect that as time increases so too will the exportation of US steel. This is based on the assumption that at the time this data was recorded US steel mining and manufacturing techniques advanced such that costs dropped and worldwide demand increased. This is the first time I will be working with time series data, which is a reason why this data set is rather small and lacks many different variables.

## **Analysis**

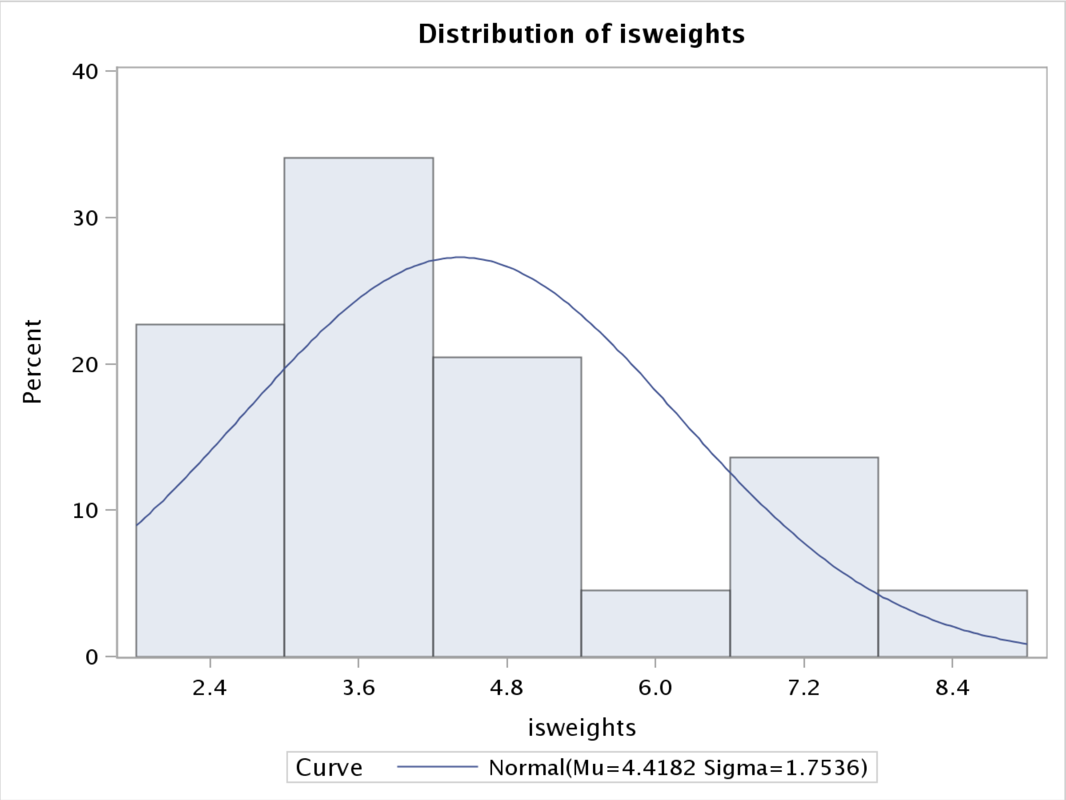
In order to meet the objective of exploring time series analysis methods by utilizing this dataset, an exploratory data analysis (EDA) must be conducted. This EDA will start with a basic Time Series Plot of I_S_Weights versus Year. From that plt, analysis will be drawn and an appropriate model will be 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*:

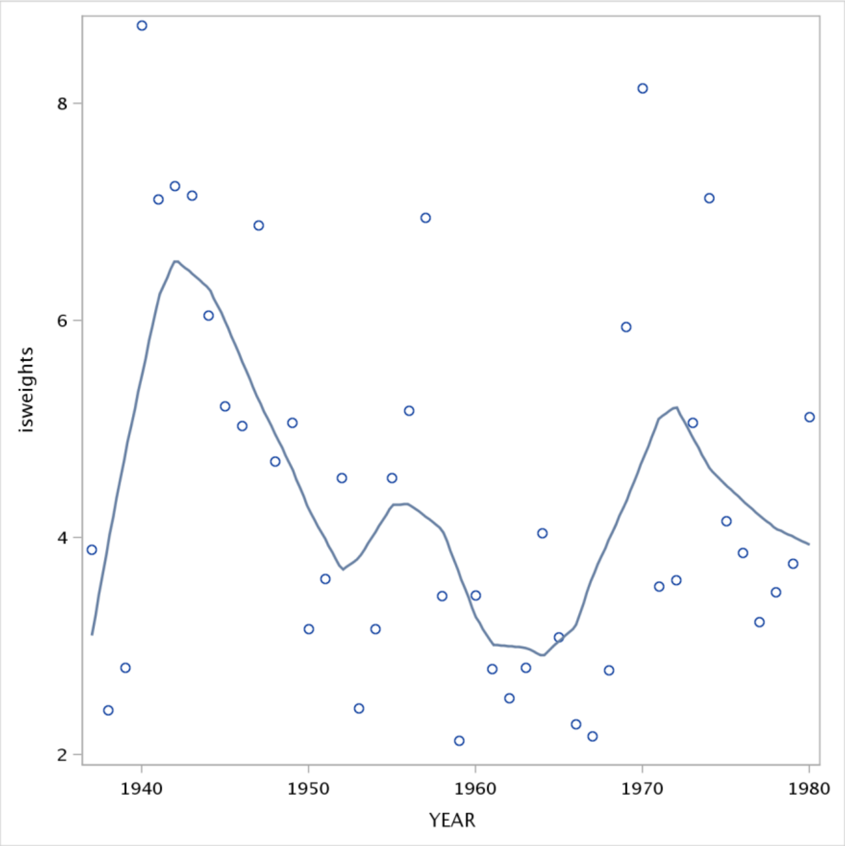
> Problem/Objective: Explore the relationship between I_S_Weights and year.
>
> Data: The data has been aggregated and has been supplied from management. There are no missing values.
>
> Analysis: I will describe the data via simple descriptive statistics at first. After the initial analysis, I will 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 with a NLAG of 10 in order to identify the appropriate model. In addition to the ARIMA, the estimate statement will also be used to fit an appropriate model to the data.
>
> 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.

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.

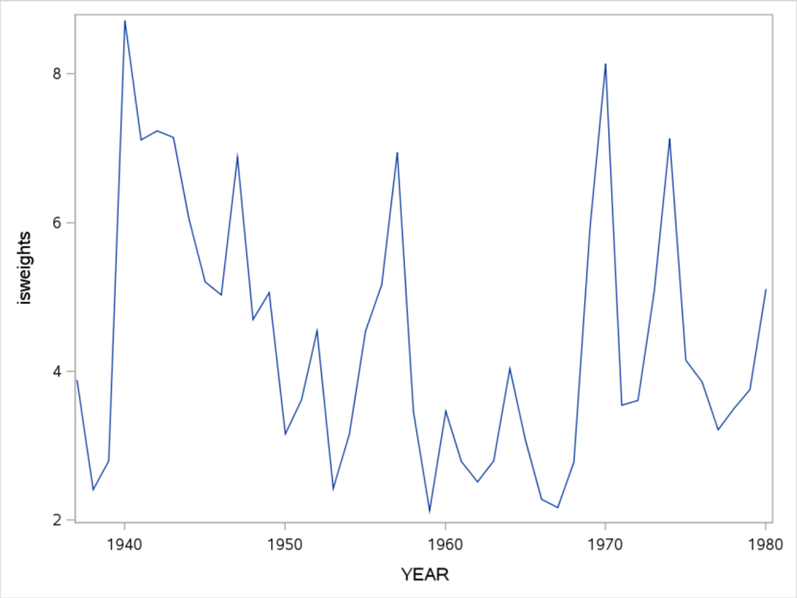
## **Data**

There are a total of 44 observations with 0 completely missing values per row. The data is comprised of one variable, I_S_Weights, and  represents the iron and steel (including scraps) exports (in million tons) for the period 1937 to 1980. The average value for exports per year is 4.418 million tons, and a median of 3.875. Standard deviation for this data set is 1.733 with a range of 6.590, which demonstrates a positive skewness of .762. Overall, the values represent a wide spectrum, and the data as a whole is slightly skewed. The distribution of steel exportation can be visually seen from the histogram. Over 30% of the export values were around 3.6 million tons. The distribution shows that 70% of the exports were less than 5million tons, which highlights the positive skewness.





To the left, one can see the scatter plot of exportation 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. 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 steel exports.



## Results

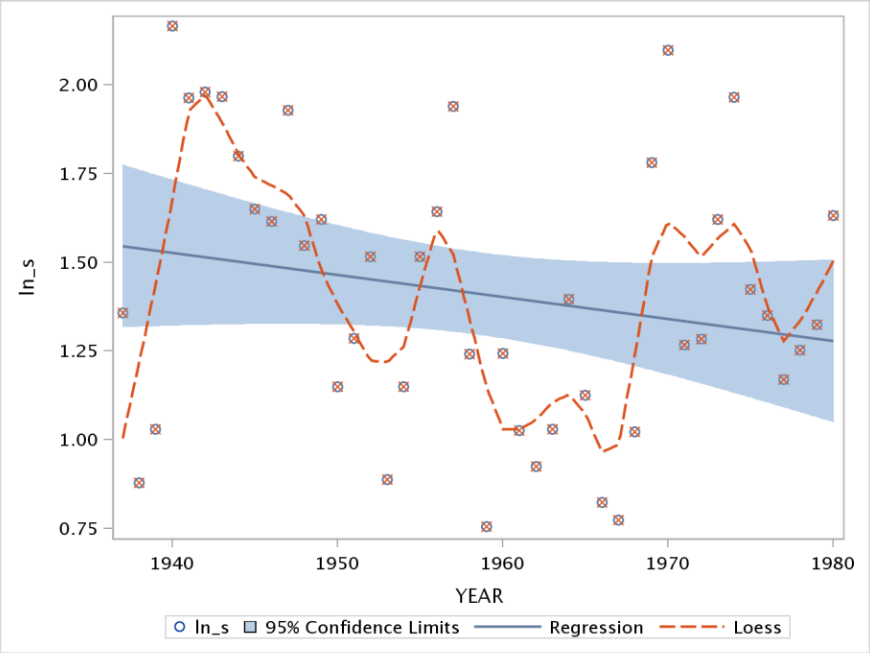
The time series plot to the right is the same scatter plot as above with the exception that a line connects each point. Notice how the overall trends between both scatter plots are the same.

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 modes. the one major assumption of ARIMA is that the data has been transformed into a stationary time series. As seen above, the data is not stationary and there is a downward 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 negative 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 is absolutely 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.**

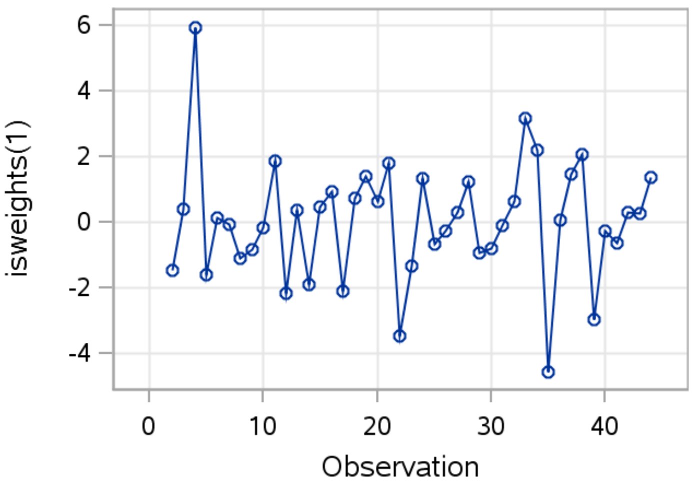
**Rule 2: If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and pattern less, then the series does not need a higher order of differencing. If the lag-1 autocorrelation is -0.5 or more negative, the series may be over differenced. BEWARE OF OVERDIFFERENCING!!**

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

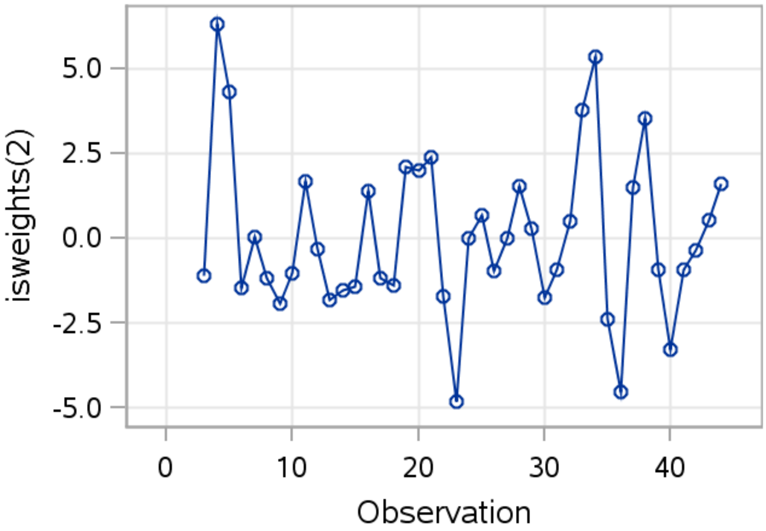
**Rule 4: A model with no orders of differencing assumes that the original series is stationary (mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth). A model with two orders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

Once the time series data has the correct differences and is stationary, selecting the optimal model to fit the data is next.

From the helpful tips from above, I will iterate the data to a stationary form. It has already been established that the original time series is not stationary. After the first difference, the scatter plot shows the new time series distribution (please refer to SAS Ouput for the full output). Initially, the time series has a stationary outcome.

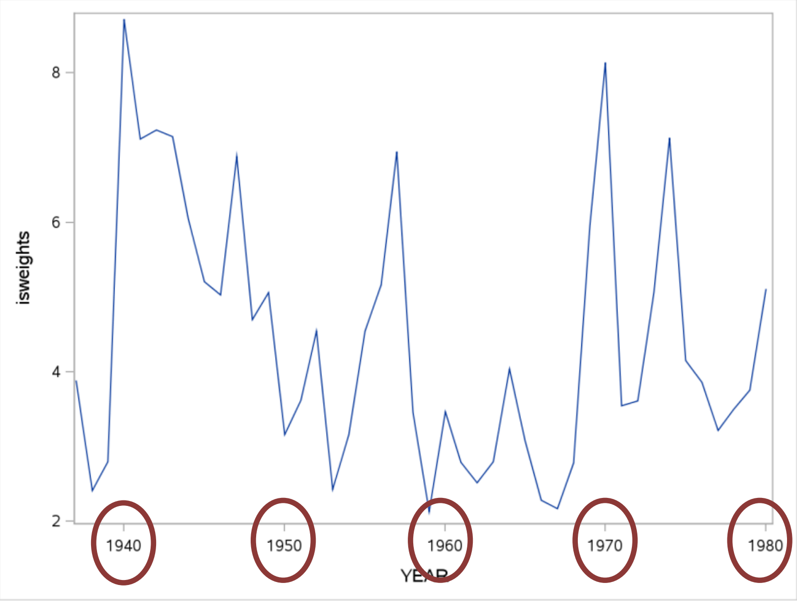


The standard deviation for this difference is 1.797 and the check for white noise is okay. For discernment sake, analyzing the second difference is helpful in settling on a stationary model. The second difference time series can be seen below. Notice how the weights are not as far apart.



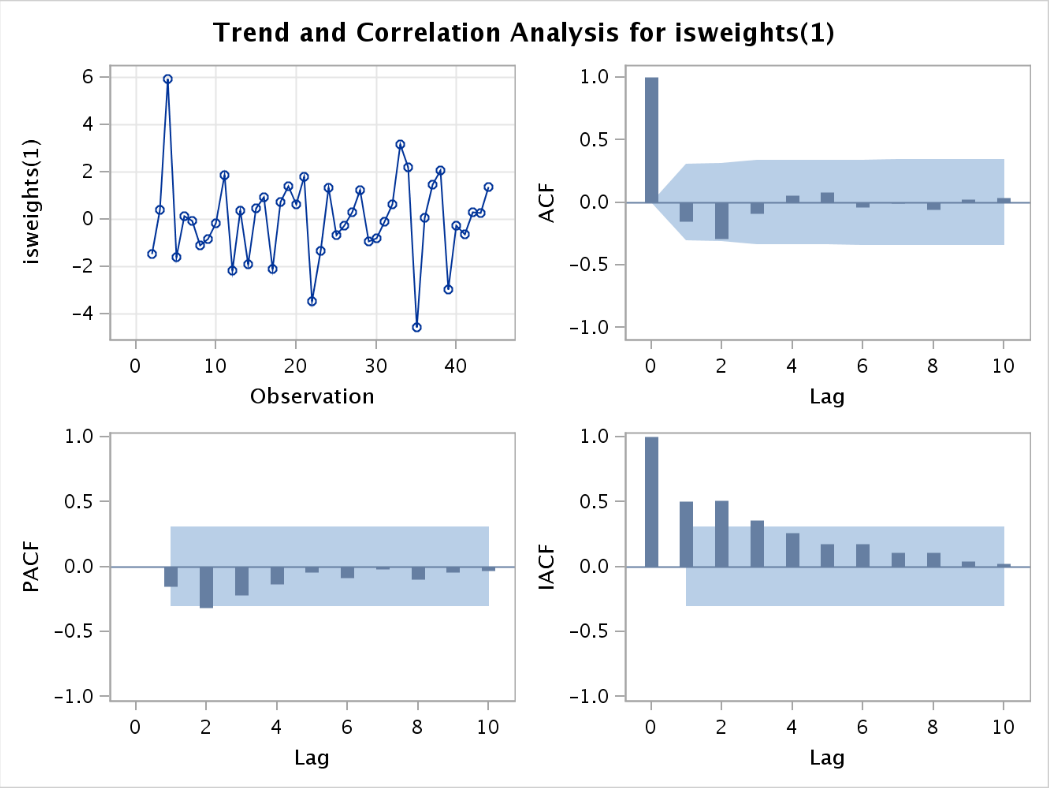
One has to guard against over-differencing, and in this case the first model is sufficient. In addition, the standard deviation for the second difference is 2.347, which is less desirable than the first difference. The first model difference is better and will be the time series.

Now that a stationary time series has been established, moving to model selection is appropriate. The autoregressive model (AR) is the simplest model and is very similar to OLS. The only difference is the dependent variable is the present value of the variable and  the independent variable(s)  are past values of the dependent variables except at different time intervals. For example, the scatter plot below shows the exportation of steel over the past 44 years. This is demonstrated via 10 year delineating segments, circled in red. One could arbitrarily create 5 independent variables based on these years. The moving average (MA) model is more complex distribution used to fit a model.



than (AR) in that it calculates a moving average of the fixed forecast errors. In addition, the average does not add up to one, and more recent forecasts carry more weight. 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 for the AR(1) lag model in regard to the ACF and PACF plots, either an AR or MA model will be used. F

or the steel EDA, I will fit both 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. After the time series is fit with an AR model, the variance (sum square errors) is 3.307. The variance for the MA model is 2.637, which is desirable compared to the AR model. When fitted with the ARMA model, the variance is 2.643. In conclusion, the stationarized time series was with fit a Moving Average model based on the smallest sum square of errors and the ACF and PACF output.



## Future Work

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

- Time permitting; it would be great to analyze this dataset updated for the most recent years.
- It would be exciting analyze other companies from different countries and compare their steel and iron export data output during this time period. Countries of interest would include Japan, Germany, Australia, New Zealand, and Switzerland.
- Economists often compare the Great Recession to The Great Depression, and one could analyze the same steel and iron export data for these different time periods and look for similar trends.

Through this initial EDA, coupled with the future work recommendations, economists will gather pertinent information in regard to predicting the ebb and flow of iron and steel exports from the US.

## References

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

AISI. "About AISI: Market Application in Steel | AISI - American Iron & Steel Institute." *AISI: American Iron and Steel Institute | Steel Industry News, Public Policy, Statistical & Production Resources*. N.p., n.d. Web. 9 Feb. 2013. <http://www.steel.org/About%20AISI/Statistics/Market%20Applications%20in%20Steel.aspx>.

Gordon, John. "Iron and Steel Industry â€” History.com Articles, Video, Pictures and Facts."

*History.com â€” History Made Every Day â€" American & World History*. N.p., n.d. Web. 9 Feb. 2013. <http://www.history.com/topics/iron-and-steel-industry>.

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

Mariam Webster. "Export - Definition and More from the Free Merriam-Webster Dictionary." *Dictionary and Thesaurus - Merriam-Webster Online*. N.p., n.d. Web. 9 Feb. 2013. <http://www.merriam-webster.com/dictionary/export>.

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.

Spoerl, Joseph. "A Brief History of Iron and Steel Manufacture." *Saint Anselm College : Saint Anselm College*. N.p., n.d. Web. 9 Feb. 2013. <http://www.anselm.edu/homepage/dbanach/h-carnegie-steel.htm>.

"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>.