**CASE STUDY:**

**Time Series Analysis of monthly sales of petroleum in the United States from Jan 1971 – Dec 1991**

**Introduction:**

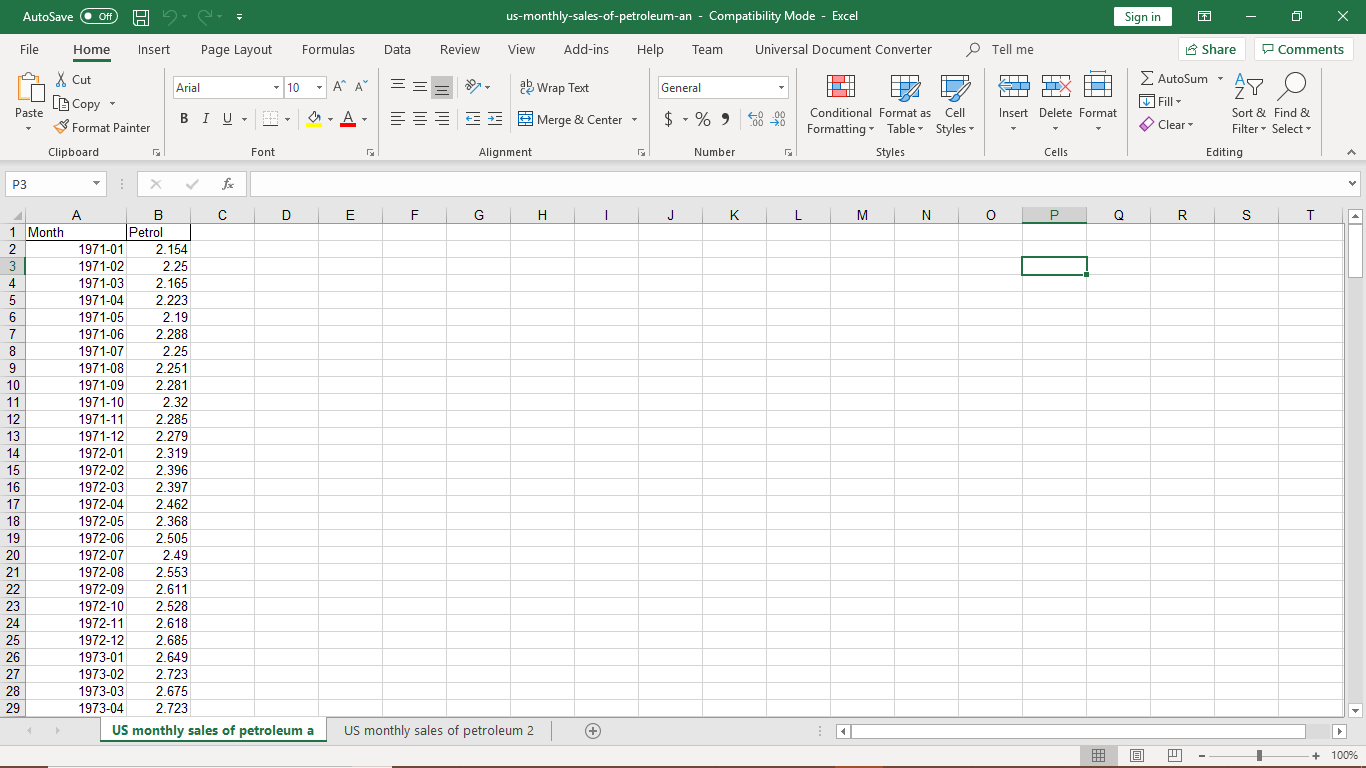
Time series analysis is a [statistical technique](http://www.statisticssolutions.com/directory-of-statistical-analyses/) that deals with time series data, or trend analysis. Time series data means that data is in a series of time periods or intervals. Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of the data. This case study is to analyze a time series dataset which has the monthly sales of petroleum in the United States from Jan 1971 till Dec 1991 with 250 observations. We wished to observe whether there were any patterns in the sales of petrol or if it differed from month to month.

**Methodology:**

The aim of this paper is to obtain a descriptive or statistical measure of a time series by plotting the data. After the data is plotted, we fit a model for the data and check the adequacy of the model using Residual analysis. The observed variation of a time series can be used to explain the variation of a related time series. Also, the observed values will be used to predict the future values of a time series by using various forecasting techniques in R.

**Dataset:**

The dataset used for the study is monthly sales of petroleum in the United States from Jan 1971 till Dec 1991 with 250 observations.



Below are the first 10 values of the dataset.

|  |  |
| --- | --- |
| Month | Petrol |
| 1971-01 | 2.154 |
| 1971-02 | 2.25 |
| 1971-03 | 2.165 |
| 1971-04 | 2.223 |
| 1971-05 | 2.19 |
| 1971-06 | 2.288 |
| 1971-07 | 2.25 |
| 1971-08 | 2.251 |
| 1971-09 | 2.281 |
| 1971-10 | 2.32 |

**Analysis and Interpretations:**

1. **Load the dataset and plot a time series plot:**

**library(tseries)**

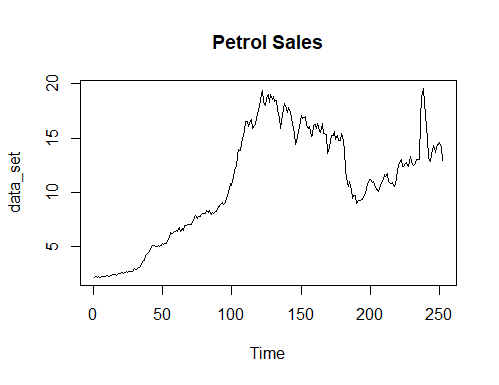
**library(astsa)**

**attach(us\_monthly\_sales\_of\_petroleum\_an)**

**class(us\_monthly\_sales\_of\_petroleum\_an)**

**data\_set=ts(Petrol)**

**ts.plot(data\_set,main="Petrol Sales")**

****

**> x<-decompose(us\_monthly\_sales\_of\_petroleum\_an)**

**Error in decompose(us\_monthly\_sales\_of\_petroleum\_an) :**

**time series has no or less than 2 periods**

**Interpretation:**

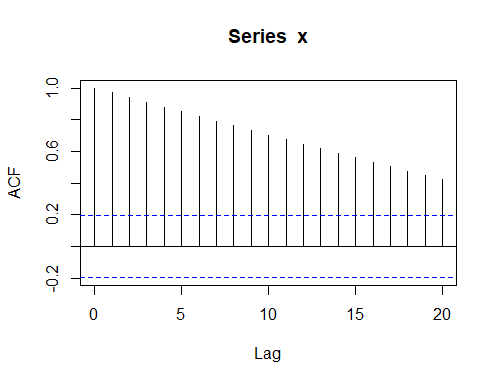
Time series plot of the data is done, and we observe that there exists only trend component.

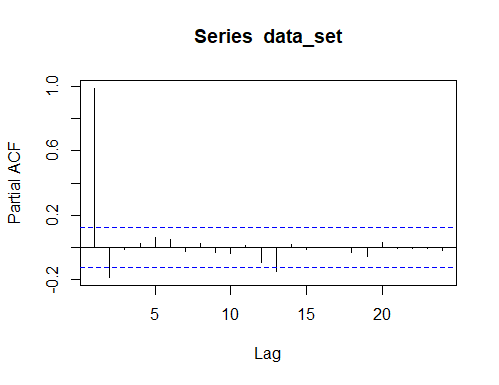
On decomposing it gives us a message stating,” The error time series has no seasonal cycles or less than 2 seasonal cycles.” This may indicate that the data are not seasonal.

1. **ACF and PACF plots:**

**acf(data\_set)**

**pacf(data\_set)**

****

****

**Interpretations:**

From the observed ACF and PACF plot we can see that the data set is not random and there is a strong correlation between the variables. The ACF plots suggest the presence of non-stationarity in the data. The PACF plot suggest that observations at nearby observations.

1. **Augmented Dickey – Fuller test.**

Since we observe non-stationarity in the ACF plot we can test the non-stationarity with the

help of Augmented Dickey – Fuller test. ADF test- used to test presence of non-stationarity

in data

The hypothesis is given by:

#NH: data is non-stationary

#AH: data is stationary

If we want to make the data set stationary, then we can create a differencing model to obtain a stationary set.

**adf.test(data\_set)**

**#null=not stationary, if p>0.05 accept it**

**Augmented Dickey-Fuller Test**

**data: data\_set**

**Dickey-Fuller = -1.1517, Lag order = 6, p-value = 0.9126**

**alternative hypothesis: stationary**

**data1=diff(data\_set)**

**adf.test(data1)**

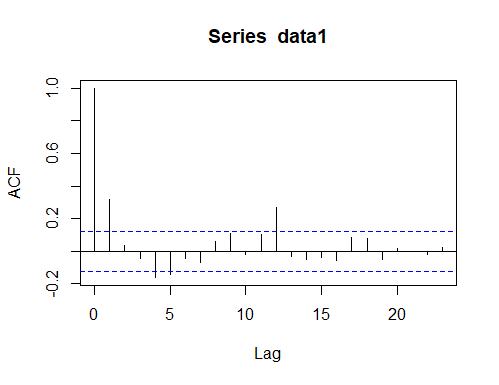
**Augmented Dickey-Fuller Test**

**data: data1**

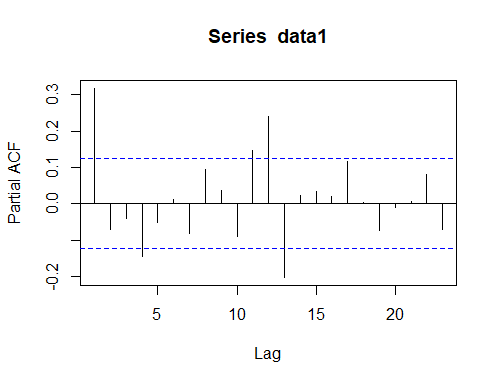
**Dickey-Fuller = -6.9136, Lag order = 6, p-value = 0.01**

**alternative hypothesis: stationary**

**acf(data1)**

****

**pacf(data1)**

****

**Interpretations:**

From the adf.test it is seen that since the p-value is 0.9126, which is greater than 0.05, we accept the null hypothesis i.e., the dataset is non-stationary. Therefore, we need to difference the data. After differencing the dataset, we can see that the p-value is 0.01 which is less than 0.05 and hence we reject the null hypothesis and thus we can now conclude that the data is stationary.

We know that When PACF plot is exponentially decreasing or oscillatory and while the ACF plot cuts at lag p, we conclude that we have an MA model of order p. So from the above ACF and PACF plots, it suggest a MA model as PACF plot is oscillatory and ACF is cutting off at lag 1.

In order to test this theory we can use auto.arima which suggest the suitable model for our

dataset.

1. **ARIMA modelling:**

**fit=auto.arima(Petrol, seasonal=FALSE)**

**fit**

**Series: Petrol**

**ARIMA(0,1,1)**

**Coefficients:**

**ma1**

**0.3333**

**s.e. 0.0577**

**sigma^2 estimated as 0.296: log likelihood=-202.94**

**AIC=409.88 AICc=409.93 BIC=416.93**

**fit1=arima(data\_set,order=c(0,1,2))**

**fit1**

**Call:**

**arima(x = data\_set, order = c(0, 1, 2))**

**Coefficients:**

**ma1     ma2**

**0.3462  0.0477**

**s.e.  0.0628  0.0638**

**sigma^2 estimated as 0.2942:  log likelihood = -202.66,  aic = 411.33**

**Interpretation:**

ARIMA - Autoregressive integrated moving average auto.arima gives the best model that

can be fitted to a given data in the form ARIMA(p,d,q).

Here, p = order of AR model, q = order of MA model, d = level of differencing applied

We have obtained ARIMA(0,1,1). This implies that the best model that can be fitted to the

data is ARMA(0,1) with 1 differencing or we can say it is MA model of order 1, i.e. MA(1).

We can check whether the above model is appropriate or not by choosing another model with some other order for example (0,1,2) and compare the AIC of both the models. Whichever model has the least Akaike information criterion that model should be considered for the dataset.

And here it is observe that that the first model i.e. ARIMA(0,1,1) has the lowest value when compared with the new model. Therefore, the observed model is the appropriate model.

1. **Holt-Winters exponential smoothing with trend and without seasonal component.**

**HoltWinters(x = data\_set, gamma = FALSE)**

**Call:**

**HoltWinters(x = data\_set, gamma = FALSE)**

**Smoothing parameters:**

**alpha: 0.5649418**

**beta : 0.5446736**

**gamma: FALSE**

**Coefficients:**

**[,1]**

**a 2.59544181**

**b 0.04712395**

**> forecasts$fitted**

**Time Series:**

**Start = 1973**

**End = 1991**

**Frequency = 1**

**xhat level trend**

**1973 2.346000 2.250000 0.096000000**

**1974 2.284050 2.243746 0.040304687**

**1975 2.271079 2.249560 0.021518992**

**1976 2.221844 2.225274 -0.003429858**

**1977 2.276145 2.259218 0.016926812**

**1978 2.270256 2.261375 0.008881678**

**1979 2.262334 2.259378 0.002956315**

**1980 2.281579 2.272879 0.008700020**

**1981 2.323807 2.303285 0.020522439**

**1982 2.310465 2.301883 0.008581130**

**1983 2.291588 2.292689 -0.001100778**

**1984 2.314408 2.307074 0.007334105**

**1985 2.392943 2.360503 0.032440586**

**1986 2.428924 2.395235 0.033688817**

**1987 2.491477 2.447610 0.043866600**

**1988 2.427591 2.421720 0.005871747**

**1989 2.501014 2.471323 0.029691107**

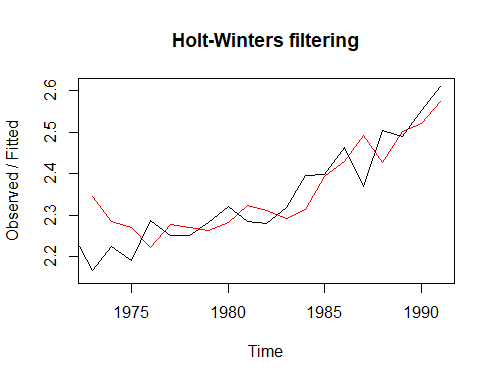
**1990 2.521094 2.494792 0.026302063**

**1991 2.575239 2.539119 0.036119913**

**> forecasts$SSE**

**[1] 0.08498939**

**> plot(forecasts)**

****

**Interpretations:**

The estimated value of alpha is 0.56, and of beta is 0.54. These are both high, telling us that both the estimate of the current value of the level, and of the slope b of the trend component, are based mostly upon very recent observations in the time series. This makes good intuitive sense, since the level and the slope of the time series both change quite a lot over time. The value of the sum-of-squared-errors for the in-sample forecast errors is 0.0849.

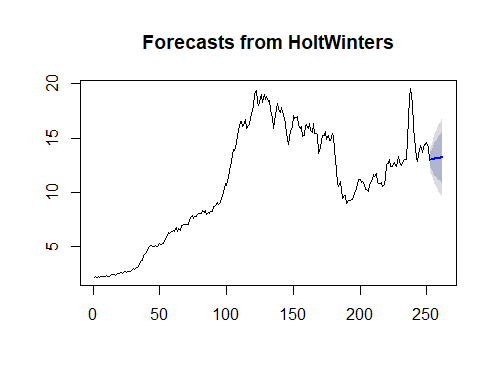
The plot of the observed values versus the fitted values showed that the in-sample forecasts agree well with the observed values, although they tend to lag behind the observed values a little bit.

1. **Forecasting:**

**library(forecast)**

**forecast\_data<-forecast(dataforecasters,h=5)**

**plot(forecast\_data)**

****

|  |
| --- |
| **Interpretations:** |
| After this using the library named “library( forecast)” we can forecast some future values and plot it. Here we have to predict the next 5 values and when gets plotted we can observe the next 10 observations as blue in color. |

1. **Residual Analysis:**

|  |
| --- |
| When a model has been fitted to a time series it is advisable to check that the model really  does provide an adequate description of the data. This is usually done by looking at the residuals which are defined by Residual = observed value -fitted value.  Here we check whether the residuals from the fitted model satisfy the assumptions of Time series.  Assumptions of time series:  1) Residual should be uncorrelated or independent of each other  2) Mean of residuals = 0, Variance of residuals should be constant.  3) Residuals are normally distributed random variables  We check for these assumptions by deriving the residuals from the fitted model and  performing the necessary tests.  **d=residuals(fit1) mean(d) 0.03107945  ts.plot(d)**    **acf(d)   pacf(d)**    **Interpretation:**  From the acf plot, we observe that the correlation values lie within the range -0.1 to 0.1 for  all the lags, hence they are negligible. Thus, the residual values are uncorrelated or independent of each other. Also, Mean = 0, Variance = constant   1. **Box-Ljung Test**   The ACF and the PACF plots clearly indicate that the residues are uncorrelated. Even though the ACF/PACF plots show the absence of dependence in residues, we confirm it by using the Box test. This can be done in R using the “Box.test()”, function.  In Box test:  H0: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).  H1: The data are not independently distributed; they exhibit serial correlation.  If the obtained p value is less than 0.05, we reject the null hypothesis and otherwise accept.  **Box.test(d,type="Ljung") Box-Ljung test  data: d X-squared = 0.04272, df = 1, p-value = 0.8363**  **Interpretations:**  Here we can see that the p-value is 0.8363 which is greater 0.05 therefore the data are independently distributed and serially correlated.   1. **Shapiro-Wilk Test**   We check for normality using the statistical test: Shapiro-Wilk Test. We consider the  following hypothesis:  H0: Residuals are Normally Distributed  H1: Residuals are not Normally Distributed  **shapiro.test(d) Shapiro-Wilk normality test  data: d W = 0.90611, p-value = 1.864e-11**  **qqline(d)** |
|  |

**Interpretation:**

Usually to check whether the data is normally independent or not we use Q-Q plot or Shapiro Wilk test. In Shapiro Wilk test if p>0.05, accept null hypothesis-residuals which suggests that the residuals are normally distributed and if residuals are not normal, transform data and run all steps again.

From the Shapiro-Wilk normality test we observe that the p-value, 1.864e-11, is greater than 0.05 so we accept the null hypothesis. Therefore. the data is normal.

**Conclusion:**

From our analysis of the dataset containing monthly sales of petroleum in the United States: we made the following conclusions:

i) The collected data exhibited a trend; that is, at certain points there was either an increasing or decreasing trend that implied a growth or decline in sales of petrol.

ii) After differencing , we were able to remove the trend component in order to work with stationary data.

iii) Using Holt Winter’s exponential smoothing, we were able to forecast future values or future sales of petroleum.

iv) Fitting an ARIMA(0,1,1) model to the data, we were able to check whether the model was appropriate by performing residual analysis.