Statistics.com

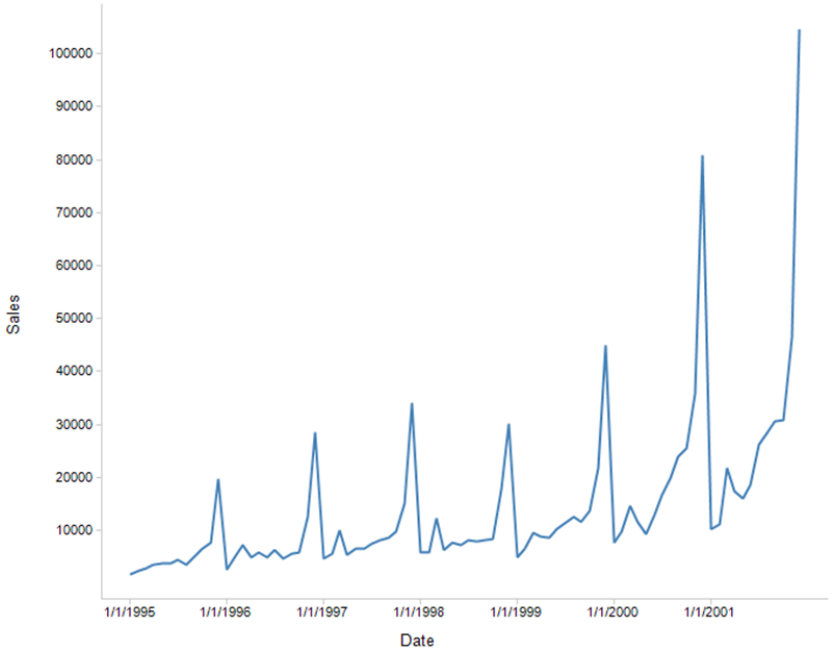
Forecasting: Assignment #1 Solutions

Chapter 2, Problem #5

5. *Souvenir Sales:* The file *SouvenirSales.xls* contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995 and 2001. [Source: R. J. Hyndman, Time Series Data Library, `www.robjhyndman.com/TSDL`; accessed on Dec 1, 2011.]

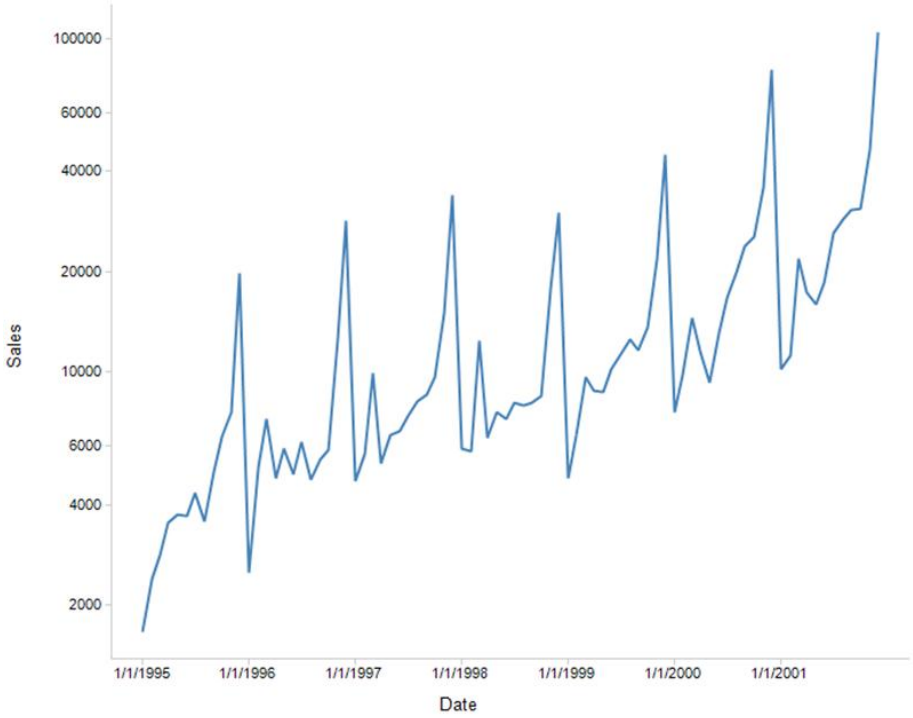
Back in 2001, the store wanted to use the data to forecast sales for the next 12 months (year 2002). They hired an analyst to generate forecasts. The analyst first partitioned the data into training and validation periods, with the validation period containing the last 12 months of data (year 2001). She then fit a regression model to sales, using the training period.

(a) Create a well-formatted time plot of the data.



(b) Change the scale of the x axis, of the y axis, or of both to logarithmic scale in order to achieve a linear relationship. Select the time plot that seems most linear.

The most linear trend is obtained by using a logarithmic scale on the y-axis.



(c) Comparing the two time plots, what can be said about the type of trend in the data?

The trend can be approximated by an exponential function.

Chapter 3, Problem #1

1. *Souvenir Sales:* The file *SouvenirSales.xls* contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995 and 2001. [Source: R. J. Hyndman, Time Series Data Library, `www.robjhyndman.com/TSDL`; accessed on Dec 1, 2011.]

   Back in 2001, the store wanted to use the data to forecast sales for the next 12 months (year 2002). They hired an analyst to generate forecasts. The analyst first partitioned the data into training and validation periods, with the validation period containing the last 12 months of data (year 2001). She then fit a forecasting model to sales, using the training period.

   Partition the data into the training and validation periods as explained above.

(a) Why were the data partitioned?

The validation partition provides a benchmark against which to test predictions, which is important because the goal of this effort is to forecast future sales.

(b) Why did the analyst choose 12 months for the validation period?

The forecast horizon is monthly forecasts for 1-12 months ahead. Choosing 12 months for the validation partition allows evaluating the prediction accuracy of 12-month ahead forecasts. A choice of a longer validation period might have been avoided to include recent data in the training period.

(c) What is the naive forecast for the validation period? (assume that you must provide forecasts for 12 months ahead)

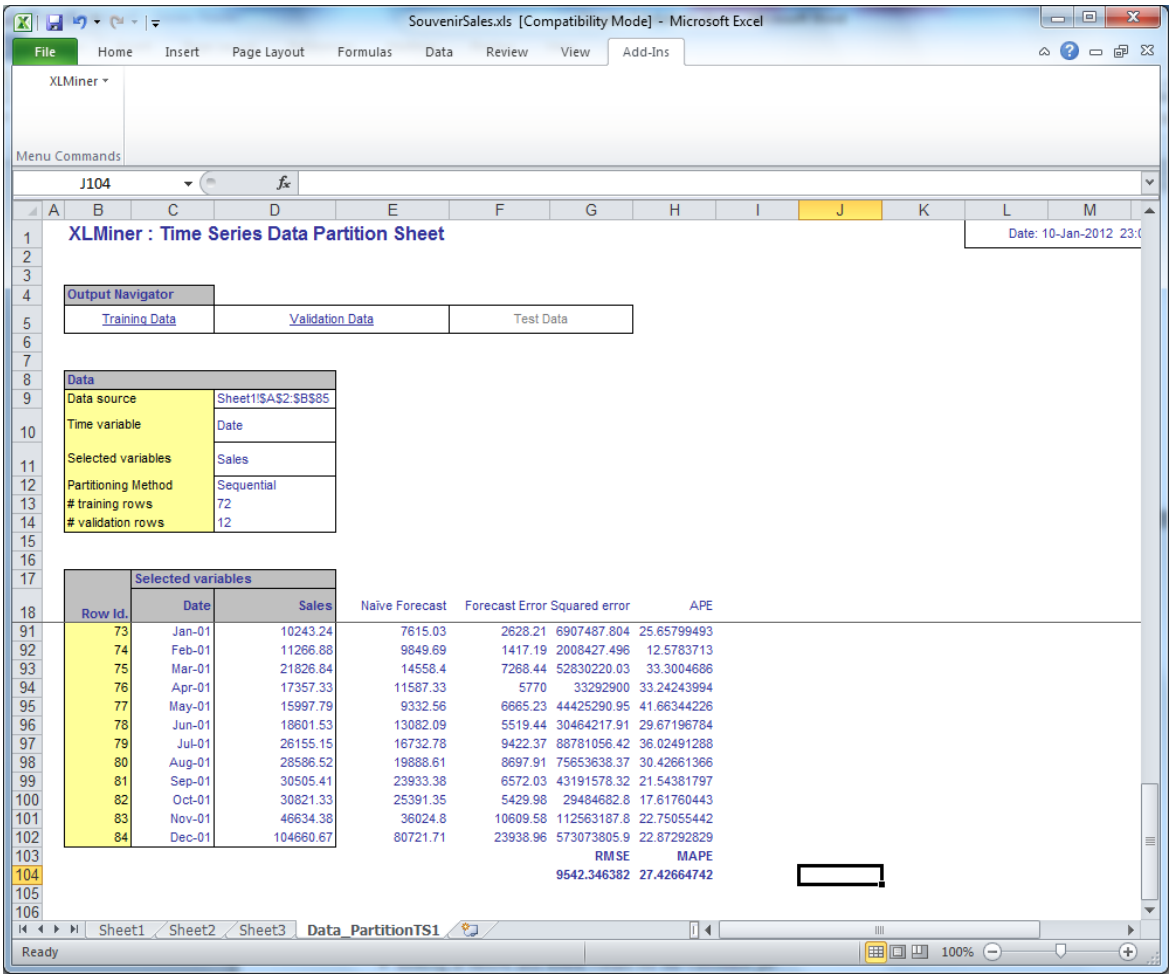
Because there is annual seasonality in the series, the naïve forecast for each of the months in the validation period is equal to sales in the most recent similar month. For example, the forecast for Jan 2001 being equal to sales in Jan 2000.

(d) Compute the RMSE and MAPE for the naive forecasts.

RMSE = 9542

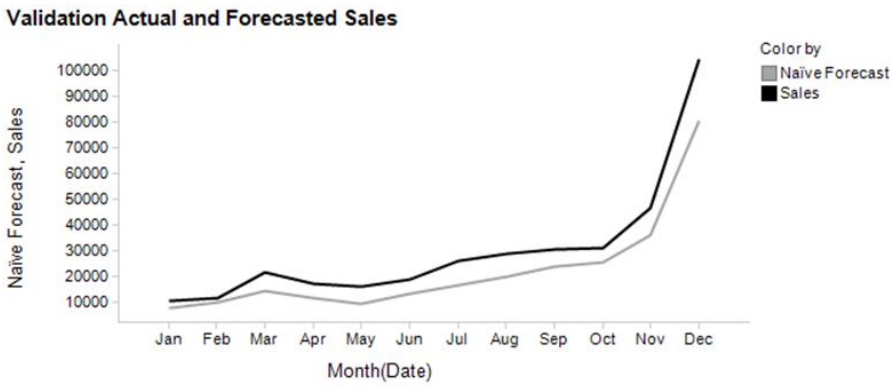
MAPE = 27.4%

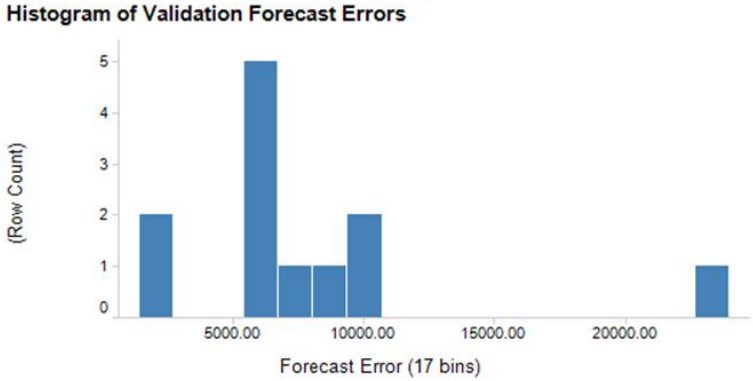
See spreadsheet for calculations:



(e) Plot a histogram of the forecast errors that result from the naive forecasts (for the validation period). Plot also a time plot for the naive forecasts and the actual sales numbers in the validation period. What can you say about the behavior of the naive forecasts?

The naïve forecasts are under-predictions of actual sales. This can be seen in both plots.



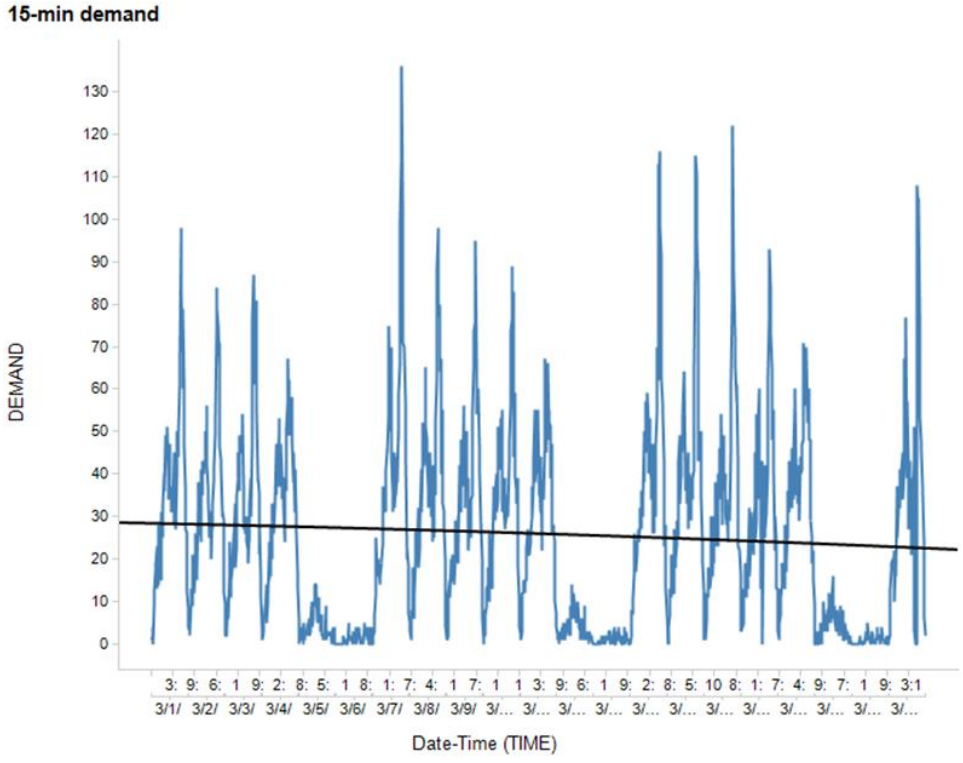


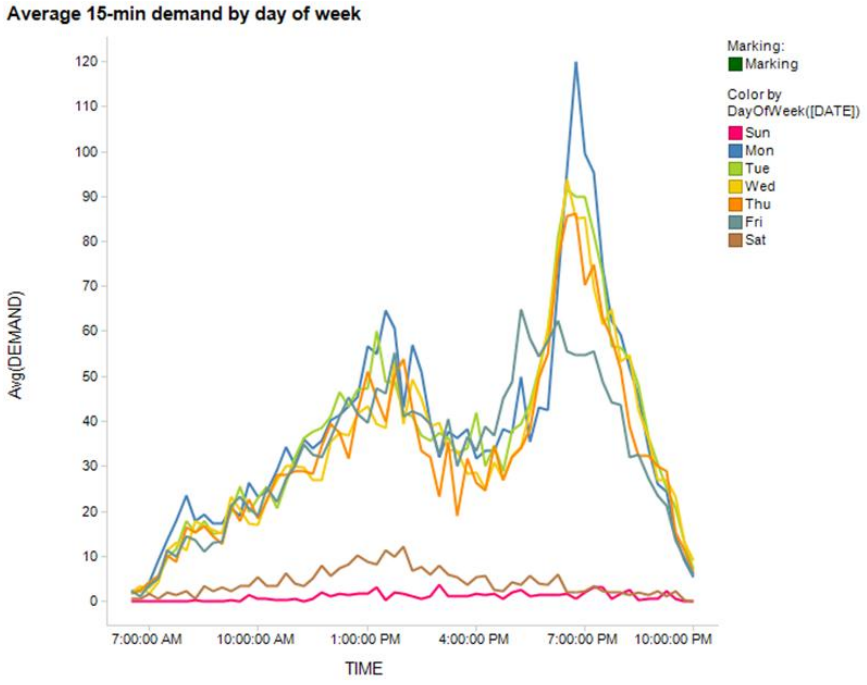
(f) The analyst found a forecasting model that gives satisfactory performance on the validation set. What must she do to use the forecasting model for generating forecasts for year 2002?

The training and validation periods must be re-combined, and then the forecasting model should be applied to the complete series before producing future forecasts.

Chapter 9, Case 9.1, Tips and Suggested Steps #1-4

1. *Use exploratory analysis to identify the components of this time series. Is there a trend? Is there seasonality? If so, how many "seasons" are there? Are there any other visible patterns? Are the patterns global (the same throughout the series) or local?*





The time plot with an overlaid line (polynomial) shows that there is barely any trend. Seasonality exists in two types of cycles: within-day and by day of week. Saturdays and Sundays look different from weekdays. Within weekdays, there are two peak demand periods. All these patterns appear to be global, occurring equally throughout the 3 week period.

2. *Consider the frequency of the data from a practical and technical point of view. What are some options?*

   The data are given in 15-min intervals. One option is to model this frequency. Another is to aggregate to hourly data. A third option is to aggregate into bins with similar demand (for example, five periods during weekdays which capture the two demand peaks; two-three periods during Saturday and Sunday to capture the single peak)