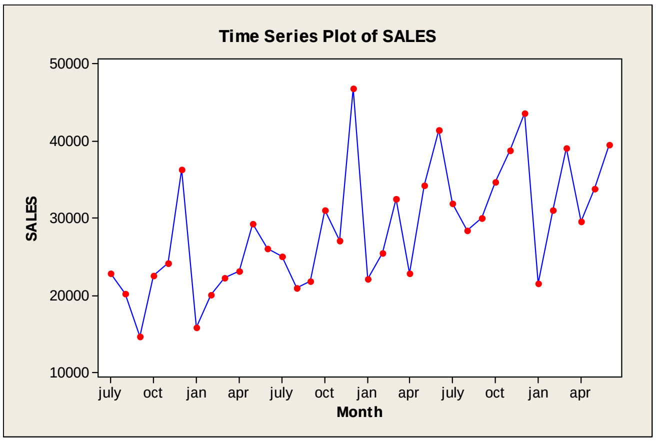
Vodka Sales - Solution

The time series plot of the sales series turned out as follows:



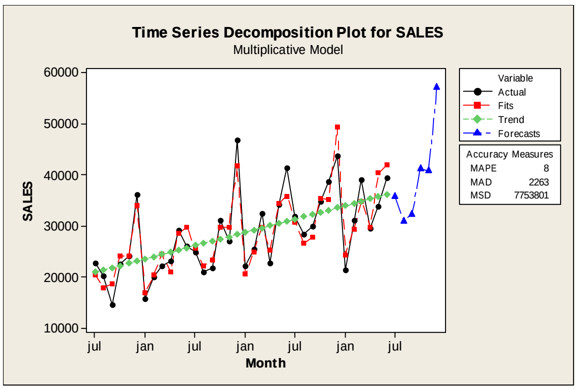
We see an increasing trend and seasonal peak in June and December.

Several modes of analysis may be possible, among them

- Trend/season decomposition: Simple and easy to communicate
- Forecasting (exponentially smoothing): Convenient and allows recursive updating
- Regression with seasonal dummies: Intermediate and flexible
- Time series modelling and prediction (ARIMA): Advanced not easily communicated

Simple trend/season decomposition does not pick up possible month to month autocorrelation as the other methods do. The last two methods provide probabilistic statements on prediction errors. Regression with trend and seasonal dummies has the added opportunity of lumping together months of similar sales level to give a model with few parameters, and with potentially better predictions. We will here limit ourselves to simple decomposition and to regression.

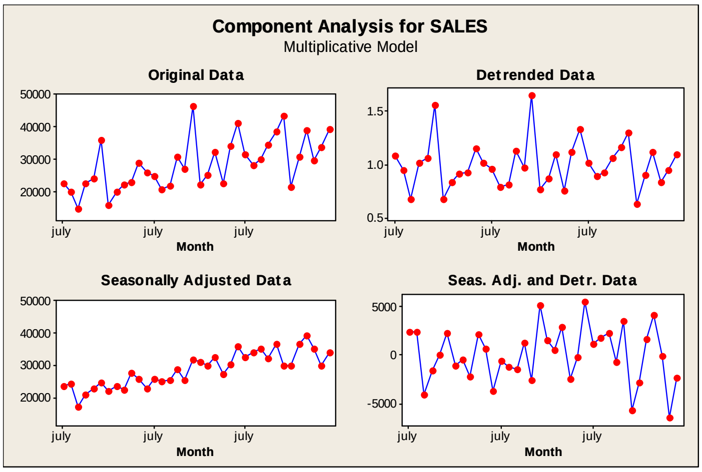
Here follows output for multiplicative decomposition with trend line, observations (true and "fitted ") and predictions 6 months ahead together with some accuracy measures.

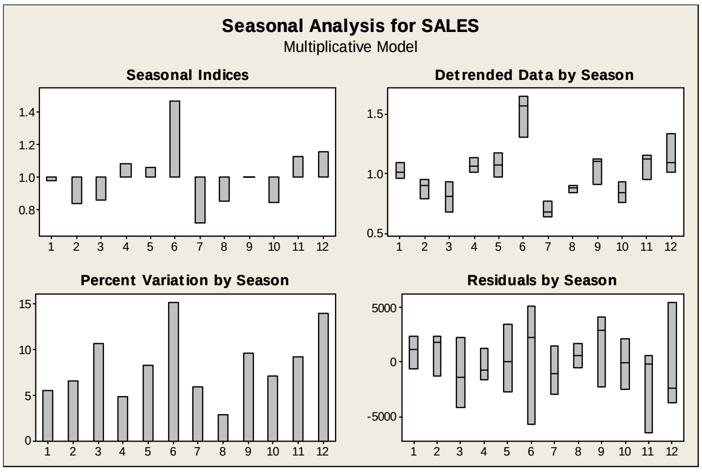


More details are given in the following (left) with formula for the trend line and seasonal indices for each month, accuracy measures and forecasts. Note that the periods are numbered from July onwards, so the peak month is December (period 6) and lowest month January (period 7). The first plot (right) shows the original data, detrended data, seasonal adjusted data and seasonal adjusted/detrended data. The second plot shows the seasonal indices and some other quantities by season.

Time Series Decomposition for SALES

```
Multiplicative Model

Data      SALES
Length    36

Fitted Trend Equation

Yt = 20499 + 435*t

Seasonal Indices

Period    Index
     1  0.97848
     2  0.83667
     3  0.86075
     4  1.08512
     5  1.06438
     6  1.47223
     7  0.71735
     8  0.85397
     9  1.00051
    10  0.84173
    11  1.13028
    12  1.15854

Accuracy Measures

MAPE         8
MAD       2263
MSD    7753801

Forecasts

Period  Forecast
jul      35819.7
aug      30992.4
sep      32259.3
okt      41140.6
nov      40817.9
des      57099.3
```





For the regression approach we start by generating a variable representing the time trend, here named TIMETREND, and generating indicator variables ("dummies") representing each month of the year. A regression analysis involving these variables, taking January as the basis month gave the following (leaving out some details from the regression output):

Regression Analysis (1)

```
The regression equation is

SALES = 11546 + 434 TIMETREND +
      5283 FEBRUARY + 10581 MARCH + 4061 APRIL +  10930 MAY +
      13679 JUNE + 9356 JULY + 5559 AUGUST + 4098 SEPTEMBER +
     10921 OCTOBER + 11053 NOVEMBER + 22884 DECEMBER

Predictor        Coef       StDev          T        P
Constant        11546        2065       5.59    0.000
TIMETREND      434.25       53.09       8.18    0.000
FEBRUARY         5283        2549       2.07    0.050
MARCH           10581        2551       4.15    0.000
APRIL            4061        2554       1.59    0.125
MAY             10930        2557       4.27    0.000
JUNE            13679        2562       5.34    0.000
JULY             9356        2568       3.64    0.001
AUGUST           5559        2562       2.17    0.041
SEPTEMBER        4098        2557       1.60    0.123
OCTOBER         10921        2554       4.28    0.000
NOVEMBER        11053        2551       4.33    0.000
DECEMBER        22884        2549       8.98    0.000

S = 3121        R-Sq = 89.7%     R-Sq(adj) = 84.3%

Unusual Observations
Obs       TID       SALES        Fit  StDev Fit   Residual    St Resid
 24      24.0      41353       35647       1802       5706       2.24R

R denotes an observation with a large standardized residual
```

As expected we see a positive time trend, and that all months have positive regression coefficients. This means that the basis month January is the lowest season. Furthermore we see that we have explained 89.7% of the variation in sales by TIMETREND and the monthly indicators (adjusted for the number of explanatory variables 84.3%). We note also the residual standard deviation of S=3121, which gives a rough idea of the prediction errors. Finally we see that one observation, no. 24 for June 2003, had higher sales than expected for its month of the year.

Classification of months from low sales to high sales, where months with sales of similar magnitude are grouped together gives:

(Jan) (Feb, Apr, Aug, Sept), (March, May, July, Oct, Nov) (June) (Dec).

In order to reduce the number of explanatory variables we may define indicators for each group, which may simply be done by adding the indicators for the members in the group.
For short we denote the two group indicators by LOWSEASON=(Feb, Apr, Aug, Sept) and MEDIUMSEASON=(March, May, July, Oct, Nov). We then get

Regression Analysis (2)

```
The regression equation is

SALES = 11489 + 437 TIMETREND + 4754 LOWSEASON + 1057 MEDIUMSEASON +
        13664 JUNE + 22887 DECEMBER

Predictor        Coef       StDev          T        P
Constant        11489        1834       6.26    0.000
TIMETREND      437.22       45.55       9.60    0.000
LOWSEASON        4754        1809       2.63    0.013
MEDIUMSEASON    10571        1772       5.96    0.000
JUNE            13664        2299       5.94    0.000
DECEMBER        22887        2288      10.00    0.000

S = 2801        R-Sq = 89.2%     R-Sq(adj) = 87.4%

Unusual Observations
Obs       TID       SALES       Fit  StDev Fit   Residual    St Resid
 18      18.0       46827     42246       1617       4581       2.00R
 24      24.0       41353     35647       1617       5706       2.49R

R denotes an observation with a large standardized residual
```

It turns out that the adjusted explanatory power measured by R-square is now increased from 84.3% to 87.4% and S is reduced from 3121 to 2801. For prediction purposes it is a virtue to have few predictor variables, but still high explanatory power. The adjusted R-square penalize for the number of explanatory variables.
If we had not lumped that many together the adjusted R-square becomes less, and this is so also if we add June to the medium group. Now S is increased as well.

Our preferred model for predicting the sales for the rest of the year will then be (2). For this we need to initiate the values of the predictor variables in the model for the subsequent months July to December. These are

TIMETREND:	37	38	39	40	41	42
LOWSEASON:	0	1	1	0	0	0
MEDIUMSEASON:	1	0	0	1	1	0
JUNE:	0	0	0	0	0	0
DECEMBER:	0	0	0	0	0	1

Regression Analysis (2b)

```
The regression equation is

SALES = 11489 + 437 TIMETREND + 4754 LOWSEASON + 10571 MEDIUMSEASON +
                             13664 JUNE + 22887 DECEMBER
```

...... the output as (2) above, with addition of the following predictions

```
 Predicted Values for New Observations
New
Obs    Fit  SE Fit      95% CI          95% PI
  1  38237    1128  (35934; 40541)  (32070; 44405)
  2  32858    1227  (30352; 35363)  (26612; 39103)
  3  33295    1261  (30719; 35871)  (27020; 39569)
  4  39549    1236  (37025; 42073)  (33296; 45802)
  5  39986    1273  (37386; 42586)  (33702; 46271)
  6  52739    1952  (48752; 56726)  (45766; 59713)
```

Here the predicted sales for each month of July to December are given in the leftmost column and prediction intervals with a 95% guarantee to the right. If we sum the six predictions we get a total 236665. By comparison simple trend/seasonal decomposition predicts the total sales for the rest of the year to be 238129.
It is not clear how to put a probability guarantee on any of the total predictions.

Predictions beyond this horizon may be risky, since there is no basis for knowing whether the time trend will continue.

In addition we could have made diagnostic plots to check the assumptions of the standard regression model. We could also have studied whether the found outlying observation(s) may effect the estimation and prediction, and perhaps do some corrections.

Further comments:

1. If predictions are made just month ahead it may be worthwhile to investigate the effect of adding the sales lagged one month to the regression equation. It turns out that this variable is not statistically significant and will not be helpful for predictions at all.
2. If we had further observations it would clearly be of interest to see which one of the two (or other) methods gave the best predictions. With a large dataset we may hold out the last portion and predict these observations and compare them with the actual values. By trying out different methods we could find the likely best one for the kind of data in question. For this comparison typically criteria used are Mean Squared Deviation (MSD) or Mean Absolute Deviation (MAD).