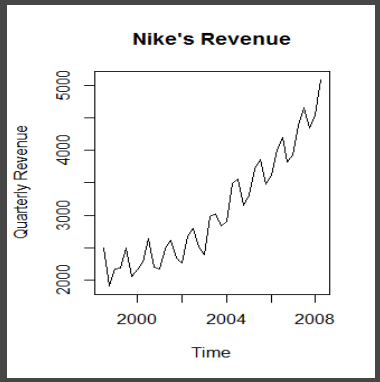
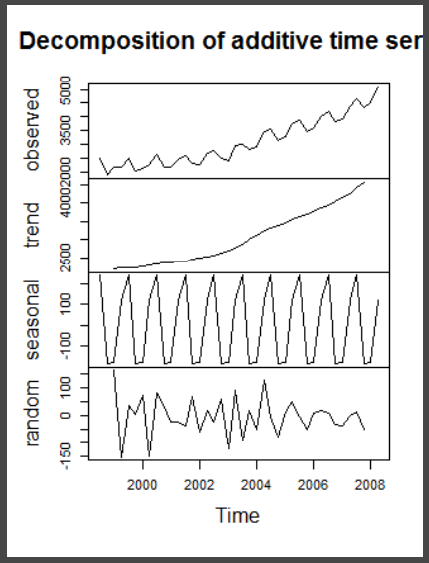
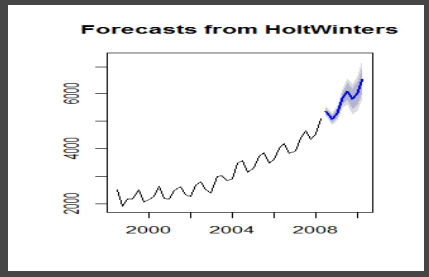
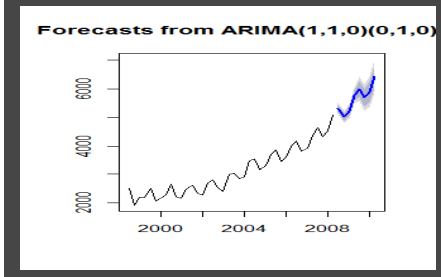
DATA SCIENCE PROJECT

Data Diggers
Let's Uncomplicate Business









Ali Syed
Chief Data Scientist

MACHINE LEARNING

Nike Revenue Forecast

Domain Retail

Topic Nike Revenue Forecast

Objective:

Build forecasting models to forecast Nike's revenue for 2010. Prepare a report to summarize approach(es) and findings.

Project's Scope

Project is required to cover following broad tasks:

- Plot the data. Which time series components seem to be present in this series? Interpret the chart in practical terms.
- Part I: Regression
  - Build a regression candidate model(s) and use that model(s) to forecast Nike's revenue for the validation set.
  - Do the forecasts seem reasonable? Briefly discuss
  - What is/are the value(s) of RMSE for the training set? What is/are the value(s) of RMSE for the validation set?

Project's Scope (cont.…)

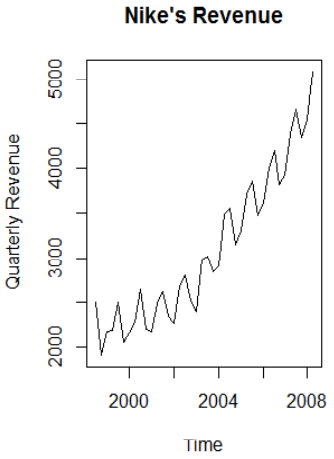
- Part II: Smoothing methods
  - Identify an appropriate smoothing model(s) you should use for Nike's revenue forecasting and discuss why you selected this/these model(s).
- Part III: Classical time series decomposition
  - Perform time series decomposition on Nike sales revenue.
- Part IV: ARIMA models
  - Is the data stationary? How do you know? Is there a way to make non-stationary data stationary? How? Apply these ideas to Nike's revenue data.

Dataset Description

- Data has been collected quarterly on Nike's revenue for the fiscal years 1998/99 through 2008/99; for instance, data for fiscal year 1999 refers to the time period from June 1, 1998 through May 31, 1999. For validation set, 2009 data has been provided.
- Both training (1999 – 2008) and validation (1999) data are in one csv file
- Data is organized as cross-sectional data i.e. columns contains quarterly data while rows represents years.
- Note revenue values given are in Million $s.

Summary Report

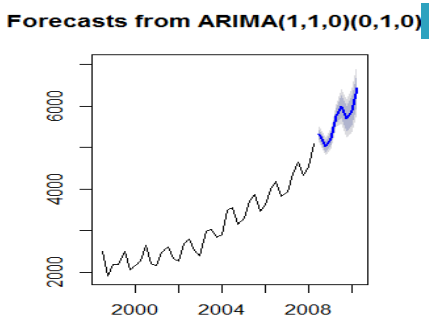
- *Following Summary Report sums up the inferences and the process of selecting the best fit model(s) in a jargon-free manner for senior management to make evidence-based decision:*
  - *Nike's Revenue data collected for fiscal years 1998/99 thru 2008/09 show that there is gradual rise in trend and as well seasonal repetitive pattern periodically over each financial year.*
  - *Minimum revenue was recorded at 1913 during 2nd Quarter of 1999 Fiscal Year, while Maximum revenue was recorded at 5088 4th Quarter of 2008 Fiscal Year. Average revenues for the period is 3094.*
  - *A simple Regression Analysis was not satisfactory due to Seasonal and Trend Variations; plus a multiple linear regression (with trend and seasonal effect incorporated) left significant information in residuals;*

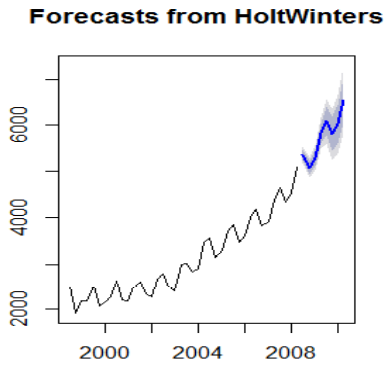


Summary Report (cont....)

- *After Detailed analysis of various models and subsequently their residuals analysis leads us to couple of best fit models i.e. HoltWinters Triple Exponential Smoothing Method and ARIMA(1,1,0)(0,1,0)s Method that probably cannot be improved upon by extrapolative TS forecasting since no information left in the residuals* (rather cause & effect forecasting method may bring some improvement) ;*
- *We suggest a forecasting approach : Prediction by taking average of forecasts generated by both HoltWinters Triple Exponential Smoothing Method and ARIMA(1,1,0)(0,1,0)s Method*

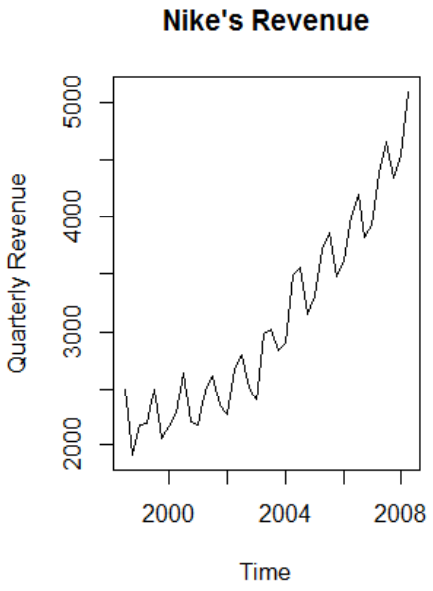
* *For complete analysis details, please review complete technical report following this summary*





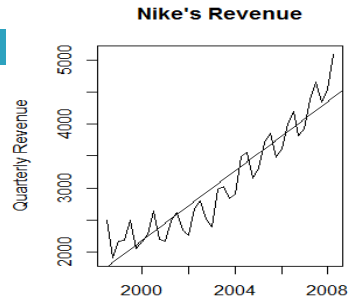
Data Preparation and Visualization

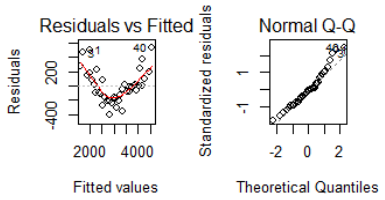
- First and foremost critical task was to bring the data in a format supporting time series analysis and modeling; that is done using R code.
- Nike Revenue data clearly exhibits trend (increasing) seasonal variations plus randomness; i.e.
  - Seasonality – notice repetition pattern periodically over each fiscal year with high revenues in the beginning of fiscal year and then revenues fall in the next two quarter and rises up until 1st quarter of the next fiscal year.
  - Trend - notice a gradual rise upward in revenues
  - Random - notice significant traces of Randomness
- Decomposition of TS validates our visual findings Later in the report

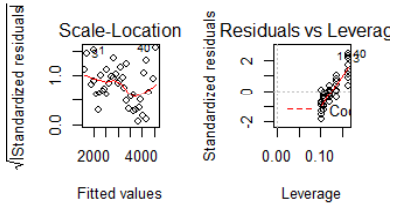


Part I: Regression

- Notwithstanding our visual observations of presence of trend, seasonality, lets check statistically how linear regression model fits to Nike Revenue data.
- Multiple linear regression model that incorporate seasonal component along with trend component give better fit compared to fitting straight line.

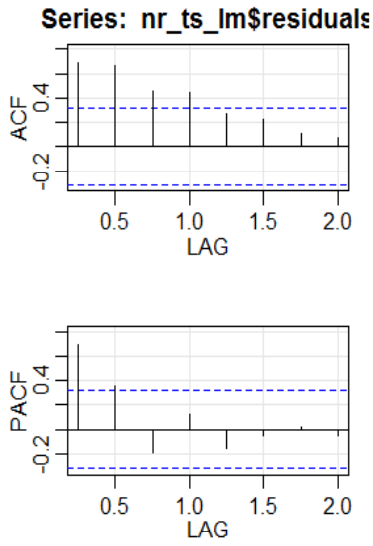






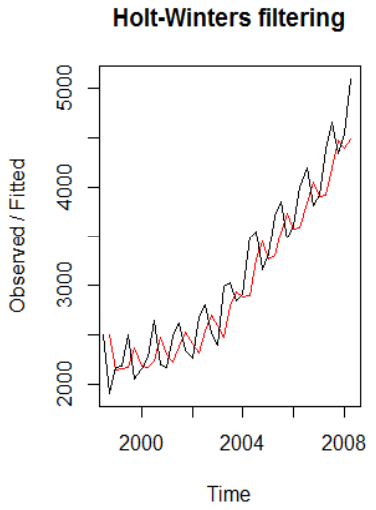
Part I: Regression (cont…)

- Root Mean Square Error (RMSE) of Training dataset is much lower compared to validation dataset, highlights unsatisfactory predictive power of linear regression model even for next year.
- RMSE (Training)=222 < RMSE (Validation)=369
- Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) of Residuals (Actual – Fitted) made clear that there is significant evidence of non-zero correlations at various lags.
- Implies that a better fit model is needed to incorporate information left in the residuals.



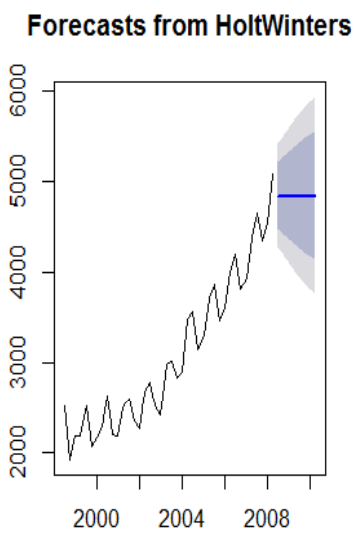
Part II: Smoothing methods (SES)

- While noticing both trend and seasonality in the Nike's revenue data implies triple exponential smoothing (TES) may be most apt model; let's validate our observations by first fitting simple exponential smoothing (SES).
- Note we estimated smoothing parameter alpha using the given data.
- Estimated smoothing parameter (0.6), gives more weight to recent observations.



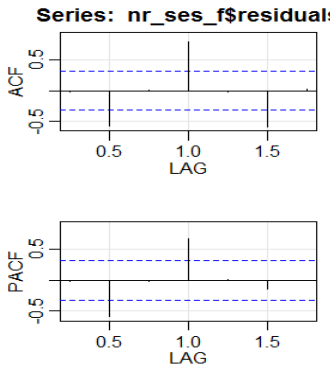
Part II: Smoothing methods (SES) (cont…)

- Estimated Sum of Square Errors (SSE) for SES= 3652701;
- Note from SES forecast for next 8 quarters (blue line) that no trend and seasonality incorporated in SES forecast
- Note dark shaded area denotes 80% prediction interval and light shaded area extended to 95% of prediction interval



Part II: Smoothing methods (SES) (cont…)

- Residuals analysis of SES:
  - from the correlogram it is quite evident that there is significant evidence of non-zero correlations at various lags;
  - Notice results of Ljung-Box test (p-value much lower then 0.01) further confirms evidence of non-zero autocorrelations at various legs:
  - It is abundantly clear from acf/pacf functions and Ljung-box tex that residuals have lots of information left.



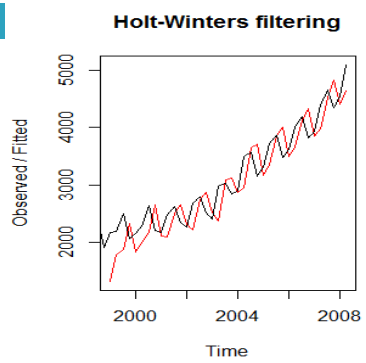
Box.test(ses$residuals, lag=16, type="Ljung-Box")

Box-Ljung test data: nr_ses_f$residuals X-squared = 148.79,

df = 16, p-value < 2.2e-16

Part II: Smoothing methods (DES)

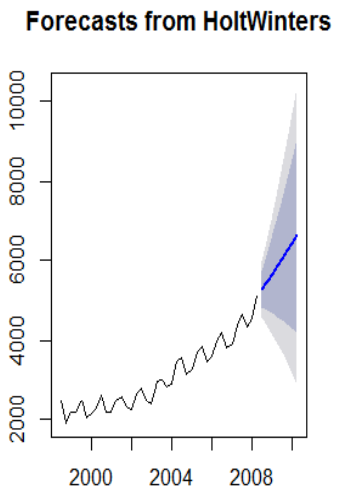
- Double Exponential Smoothing (DES), estimated value of parameters alpha (1) implies that forecast level is dependent on most recent values of time series;
- however small value of beta (0.22) implies that slop of the trend component is less dependent on recent values rather on historical values ;



Smoothing parameters:
alpha: 1
beta : 0.2249913
gamma: FALSE
Coefficients: [,1]
a 5088.0000
b 192.3685

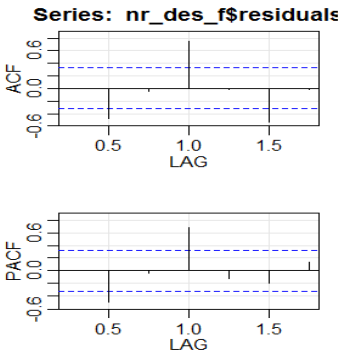
Part II: Smoothing methods (DES) (cont…)

- Note from DES forecast for next 8 quarters (blue line) that trend is clearly incorporated in forecast however seasonality remains not modeled in DES forecast;
- Estimated Sum of Square Errors (SSE) for DES= 5137920;



Part II: Smoothing methods (DES) (cont…)

- Residuals analysis of DES:
  - from the correlogram it is quite evident that there is still significant evidence of non-zero correlations at various lags;
  - Notice results of Ljung-Box test (p-value much lower then 0.01) further confirms evidence of non-zero autocorrelations at various legs:
  - It is abundantly clear from acf/pacf functions and Ljung-box tex that residuals have lots of information left.



Box-Ljung test data: des$residuals

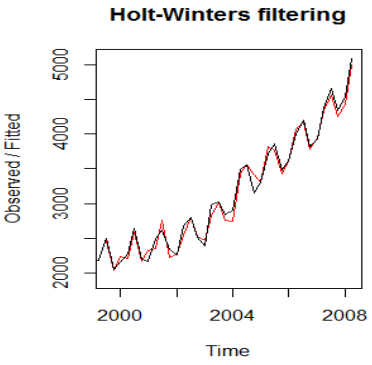
X-squared = 116.01,

df = 16,

p-value < 2.2e-16

Part II: Smoothing methods (TES)

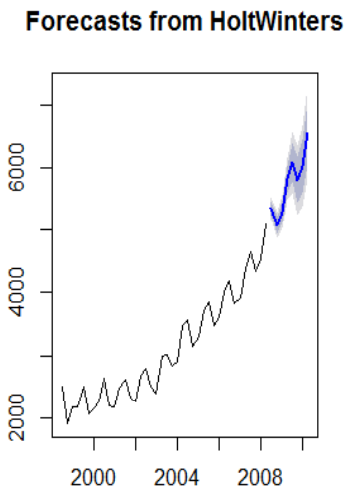
- Triple Exponential Smoothing (TES), as observed earlier, fits the best to the given Nike revenue TS compared to other SES and DES (notice how close fitted values are to actuals.
- The estimated values of alpha (0.38) implies the foreast level at the current time point is based upon both recent observations and more distant past values, however past values have more weight;
- beta (0.6) implies that the slope b of the trend component, are based largely upon very recent observations in the time series;
- while the value of gamma (1) indicating that the estimate of the seasonal component at the current time point is entirly based upon very recent observations



Smoothing parameters:
alpha: 0.3775294 beta : 0.6180435 gamma: 1
Coefficients: [,1]
a 4816.81956
b 181.93644
s1 358.38880
s2 -106.02232
s3 -77.69796
s4 271.18044

Part II: Smoothing methods (TES) (cont…)

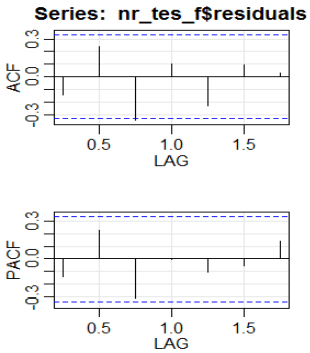
- Note, as was anticipated, from TES forecast for next 8 quarters (blue line) that both trend and seasonality are clearly incorporated in the forecast;
- Estimated Sum of Square Errors (SSE) for TES= 310893, which is much lower than that SSE of both SES and DES; i.e.



SSE of TES =310893 < SSE of SES=3652701 < SSE of DES=5137920

Part II: Smoothing methods (TES) (cont…)

- Residuals analysis of TES:
  - from the correlogram it is quite evident that there is no significant evidence of non-zero correlations at various lags (note 3rd lag slightly exceeds the significance bound however that can be entirly by chance);
  - Notice results of Ljung-Box test (p-value higher then 0.05) further confirms evidence of no non-zero autocorrelations at various legs:
  - It is abundantly clear from acf/pacf functions and Ljung-box tex that residuals have no more information left and TES enabled us to model all the information available in actual data.



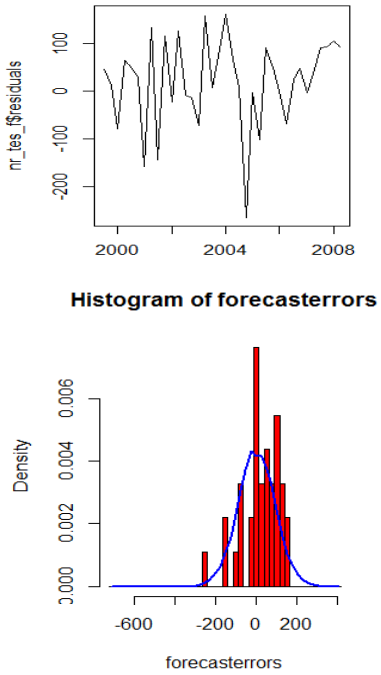
Box-Ljung test data: tes$residuals

X-squared = 25.725, df = 16,

p-value = 0.05802

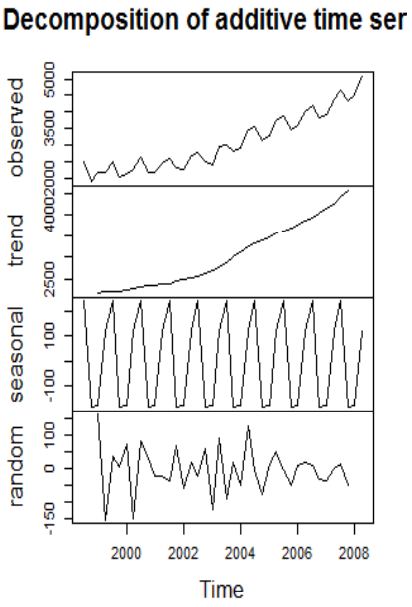
Part II: Smoothing methods (TES) (cont…)

- Residuals analysis of TES: (cont…)
  - by examining the residuals charts, from the time plot, it appears plausible that the forecast errors have constant variance over time except the abnormal drop in 2006
  - From the histogram of forecast errors, it seems plausible that the forecast errors are largely normally distributed with mean zero and constant variance.
  - From ACF/PACF plots of residuals little evidence of autocorrelation at various lags for the forecast errors,
  - With all above findings, we conclude that Holt-Winters exponential smoothing provides an adequate predictive model of the Nike revenue, which probably cannot be improved upon.



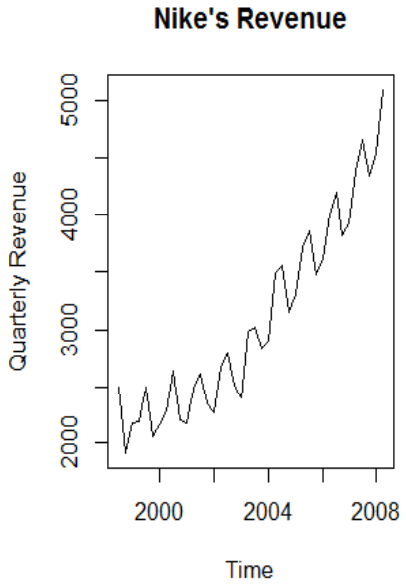
Part III: Classical time series decomposition

- From the plots, we notice that the time series has Seasonality, Trend and Randomness.
- Also we notice that the random fluctuations are roughly constant in size over time. So this suggests data can be describes as Additive Model.
  - Seasonality - We notice a repetition pattern periodically over each fiscal year with high revenues in the beginning of fiscal year and then revenues fall in the next two quarter and rises up until 1st quarter of the next fiscal year.
  - Trend - We notice a constant pattern of gradual rise an upward pattern in revenues
  - Random - We notice traces of Randomness with significant high and low spikes along the time period



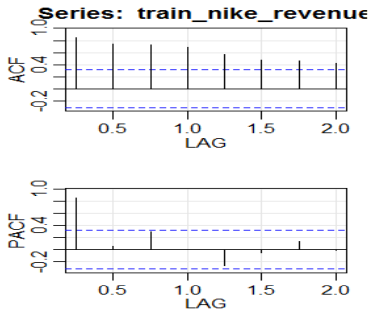
Part IV: ARIMA models

- Note ARIMA models require time series (TS) must be stationary.
- For a TS to be stationary; mean, variance and covariance shell not be a function of time i.e. should not change over time
- Generally following are steps for ARIMA modeling:
  - plot and visualize the time series
  - Stationarize the time series
  - find Optimal Parameters of ARIMA model generally using ACF/PACF measures
  - Fit ARIMA model using optimal parameters
  - Forecast values using Selected/fitted ARIMA models
- From the TS plot, it seems that time series is in not stationary; will further check if it is.



Part IV: ARIMA models (cont....)

- ACF plot of Nike Revenue data clearly gradually decreasing implies that TS in non-stationary.
- dicky-fuller test further confirms that Nike revenue is a non-stationary time series



Augmented Dickey-Fuller
Test data: nike_revenue
Dickey-Fuller = -0.1661,
Lag order = 3,

p-value = 0.99

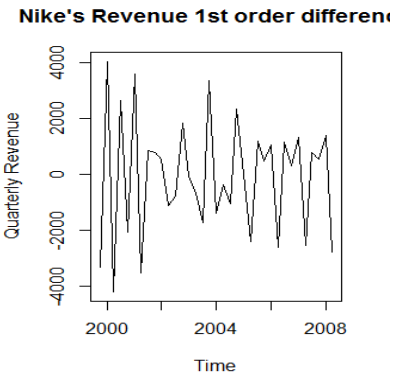
alternative hypothesis:
stationary

Part IV: ARIMA models (cont....)

- To make a non-stationary time series stationary:
  - Two most common ways to make a non-stationary time series curve stationary are:
    - Differencing (Most commenly used technique to make TS stationary; can be of 1st order or 2nd order or 3rd order)
    - Transforming (most common transformation is log transformation; However, it is normally suggested that you use transformation only in case differencing is not working)

Part IV: ARIMA models (cont....)

- We took 1st order difference and it appears that
- Examining plot of nike revenues TS's 1st order differences (both normal and seasonal differencing) appears to be somewhat stationary i.e. stationary in mean and in variance (as the level of the series stays roughly constant over time, and the variance of the series appears roughly constant over time.);
- Dicky-fuller test further confirms that the Nike revenue TS of 1st order difference (both normal and seasonal differencing) is stationary.



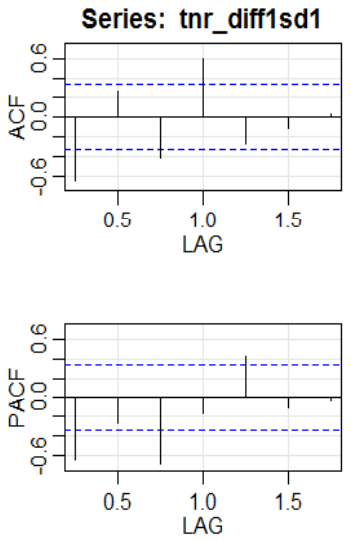
Augmented Dickey-Fuller
Test data: tnr_diff1sd1
Dickey-Fuller = -18.801,
Lag order = 3,

p-value = 0.01

alternative hypothesis:
stationary

Part IV: ARIMA models (cont....)

- by examining ACF and PACF of first order differenced series (for both simple and seasonal) limited or no possibility of any auto regressive (i.e. p=0 or p=1) model plus no possibility of moving average of any order arima(0,1,0)(0,1,0)s or sarima(1,1,0)(0,1,0)s likely be candidate model



Part IV: ARIMA models (cont.…)

- ARIMA (1,1,0)(0,1,0)s appears better fit model by comparing both candidate models vs ARIMA (0,1,0)(0,1,0)s
- Further confirmed as auto-arima function of R, selected the same model that we suggested i.e. ARIMA(1,1,0)(0,1,0)s

arima(x = train_nike_revenue, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 0), period = 4))
sigma^2 estimated as 12021:

log likelihood = -214.07,

aic = 430.13 >

arima(x = train_nike_revenue, order = c(1, 1, 0), seasonal = list(order = c(0, 1, 0), period = 4))

Coefficients: ar1 -0.5067

s.e. 0.1480

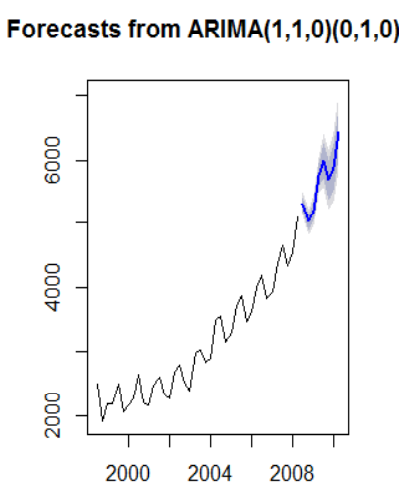
sigma^2 estimated as 8982:

log likelihood = -209.11,

aic = 422.23

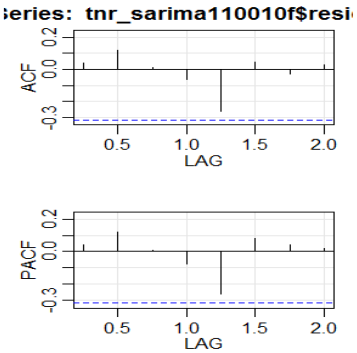
Part IV: ARIMA models (cont.…)

- Note, ARIMA forecast for next 8 quarters (blue line) that both trend and seasonality are clearly incorporated in the forecast;
- Further confirmed as auto-arima function of R, selected the same model that we suggested i.e. ARIMA(1,1,0)(0,1,0)s



Part IV: ARIMA models (cont....)

- Residuals analysis of ARIMA(1,1,0)(0,1,0)s:
  - from the correlogram it is quite evident that there is no significant evidence of non-zero correlations at various lags;
  - Notice results of Ljung-Box test (p-value higher then 0.05) further confirms evidence of no non-zero autocorrelations at various legs:
  - It is abundantly clear from acf/pacf functions and from Ljung-box test that residuals have no more information left and ARIMA(1,1,0)(0,1,0)s enabled us to model all the information available in actual data.



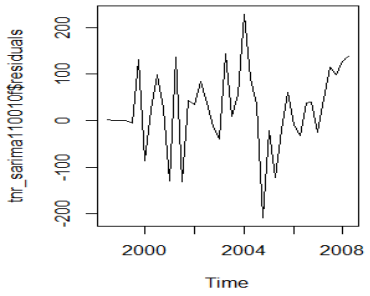
Box-Ljung test data: sarima110010f$residuals

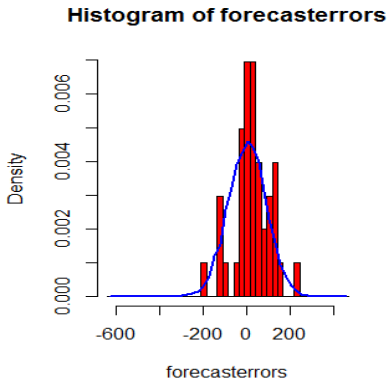
X-squared = 16.646, df = 16,

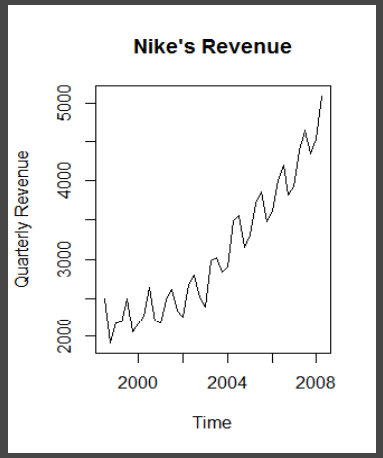
p-value = 0.4088

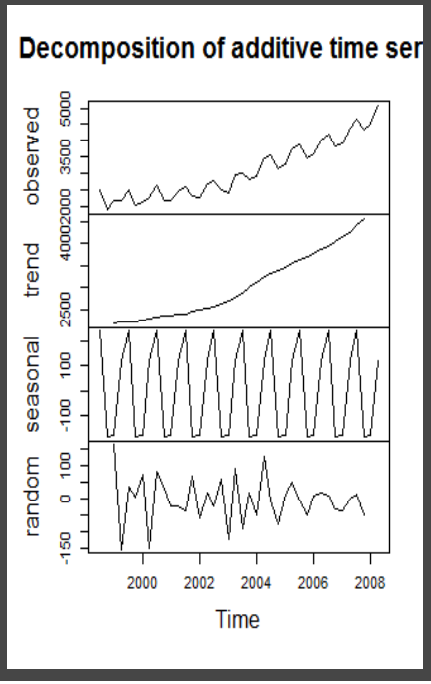
Part IV: ARIMA models (cont.…)

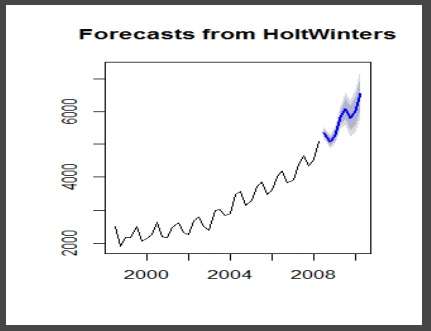
- Residuals analysis of ARIMA(1,1,0)(0,1,0): (cont…)
  - by examining the residuals charts, from the time plot, it appears plausible that the forecast errors have constant variance over time except the abnormal drop in 2006
  - From the histogram of forecast errors, it seems plausible that the forecast errors are largely normally distributed with mean zero and constant variance.
  - From ACF/PACF plots of residuals little evidence of autocorrelation at various lags for the forecast errors,
  - With all above findings, we conclude that ARIMA(1,1,0)(0,1,0)s provides an adequate predictive model of the Nike revenue, which probably cannot be improved upon.

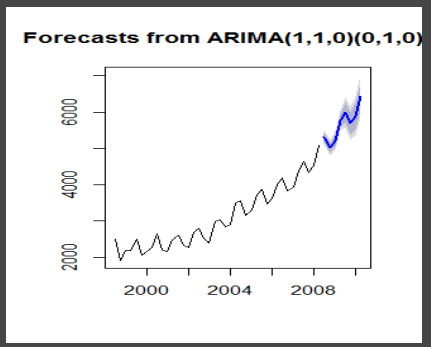












Data *is the new oil* & Analytics *is the new combustion engine*

Data Diggers
Let's Uncomplicate Business

THANK YOU