*Time Series Analysis and Forecasting*

- Fundamental concepts
- Model-based time series analysis
- Stationarity and auto-correlation
- Forecasting
- Multi-horizon forecasting
- Multi-variate time series

# MIS772

## Predictive Analytics

## Time Series Analysis and Forecasting

**Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.**

DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

# So far nm the unit…

- Predictive models on **cross-sectional** data
  - A set of predictors or independent variables
  - To predict the class or dependent variable (different from the predictors)
  - Regardless of the time dimension

# What if...

- There is a time dimension (data are time-stamped)
- Aim: To predict the value/s of an attribute that is/are changing over time


Image source: https://aboveintelligent.com/the-a-i-gold-mine-predicting-stock-market-success-19082ec87ef5

Image source: https://desertscreenwritersgroup.com/2015/09/21/screenplay-sequence/

THE LANCET
Planetary Health

*"…ambulance callouts, emergency department presentations, and data on hospital admissions for* Nov 21, 2016*, as well as leading up to and following the event were collected…"*

*"…At 1800 AEDT, a gust front crossed Melbourne, plunging temperatures 10°C, raising humidity above 70%, and concentrating particulate matter. Within 30 h, there were 3365 (672%) excess respiratory-related presentations to emergency departments, and 476 (992%) excess asthma-related admissions to hospital"*

*Source: Thien, F. et al. (2018), Vol 2, Issue 6, E255-E263*

DEAKIN BUSINESS SCHOOL

# Mini case study:
# Demand forecasting

- An emergency hospital wants to estimate their staff requirements (doctors, specialists, nurses, support staff)
  - Considerations…
- Weather seasons: spike in demand because of climatic conditions, different activities people engage in
- Recent years: overall demand has grown as a result of growth in region
- These all mean: there is **seasonality** and **trend** in the demand for the emergency services
  - Aim: The hospital would like to…
- forecast the demand for different staff requirements (weekly, monthly, annually)
- plan for required resources (rostering, on call)

the general **trend** shows growth in demand over time

every few months, the demand increases and then decreases (**seasonality**)

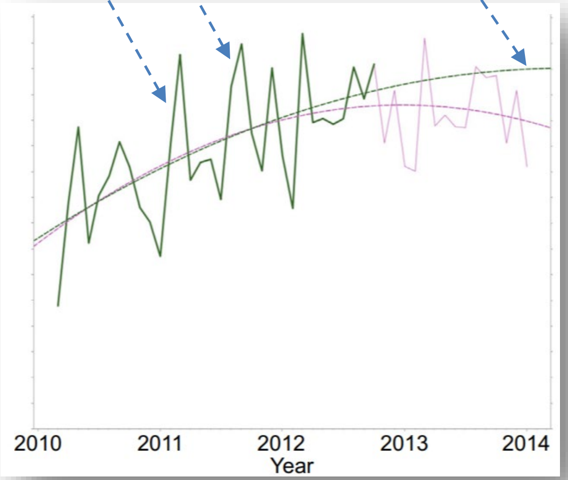



Image source: Page 307, KD Ch10

# Terminology

- Time periods
  - Refers to any time interval that is of interest
  - E.g., seconds, minutes, weeks, months, years, etc.
- Horizon
  - The time period for which forecasting is done
  - E.g., next week, three weeks from now, or two years from now
- Forecast error
  - Is the difference between the predicted value of the attribute and its actual value at any time
  - Calculated as: $e_t = y_t - F_t$

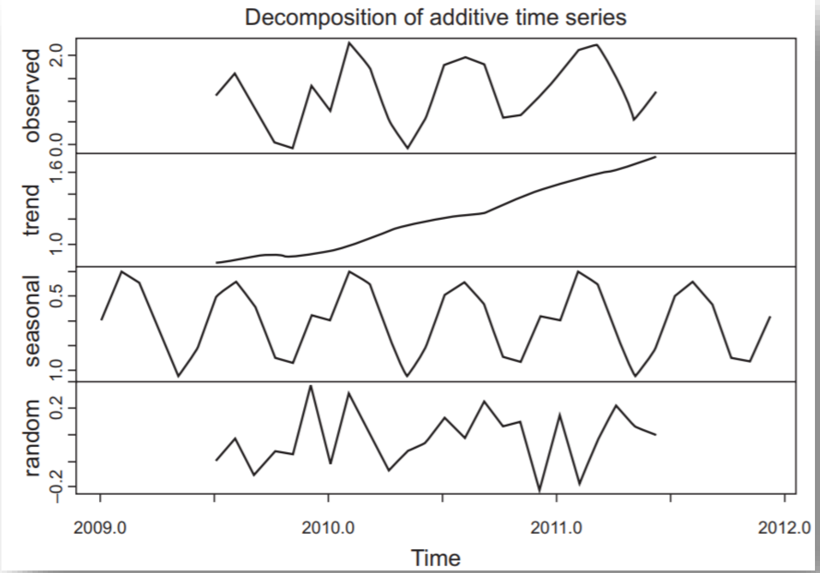
the actual value
at time t

forecast
at time t

# Data components

- Time series data are called **non-stationary** if the data exhibits…
  - **Seasonality:** The data have regular and predictable changes that recur over time, with a **(relatively) fixed** period. This is called **cyclicality** when the data have regular changes that recur over time, but with a **non-fixed** period.
  - **Trend:** The data have a pattern of gradual change over time.

Image source: Page 307, KD Ch10




  - **Random noise:** The data have some normal fluctuations even after the seasonality and trend components have been taken out.
  - The above characteristics will result in different **means** and **variances** over time.

# Data: Ames Real Estate

***A real-estate agency in Ames (USA) would like to determine if the housing market is going to increase in volume in the next quarter and if the prices are likely to go up or down. They are also interested in the time periods that are best for selling or buying property and if the auctions give best results on weekends or weekdays.***

ExampleSet (2930 examples, 2 special attributes, 79 regular attributes) Filter (2,930 / 2,930

| Row No. | PID | SalePrice | MS_SubClass | MS_Zoning | Lot_Frontage | Lot_Area | Street | Alley | Lot_Shape | Land_Contour | Utilities | Lot_Config | Land_Slope | Neighborhood | Condition_1 | Condition_2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 526301100 | 215000 | 20 | RL | 141 | 31770 | Pave | NA | IR1 | Lvl | AllPub | Corner | Gtl | NAmes | Norm | Norm |
| 2 | 526350040 | 105000 | 20 | RH | 80 | 11622 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | NAmes | Feedr | Norm |
| 3 | 526351010 | 172000 | 20 | RL | 81 | 14267 | Pave | NA | IR1 | Lvl | AllPub | Corner | Gtl | NAmes | Norm | Norm |
| 4 | 526353030 | 244000 | 20 | RL | 93 | 11160 | Pave | NA | Reg | Lvl | AllPub | Corner | Gtl | NAmes | Norm | Norm |
| 5 | 527105010 | 189900 | 60 | RL | 74 | 13830 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 6 | 527105030 | 195500 | 60 | RL | 78 | 9978 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 7 | 527127150 | 213500 | 120 | RL | 41 | 4920 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | StoneBr | Norm | Norm |
| 8 | 527145080 | 191500 | 120 | RL | 43 | 5005 | Pave | NA | IR1 | HLS | AllPub | Inside | Gtl | StoneBr | Norm | Norm |
| 9 | 527146030 | 236500 | 120 | RL | 39 | 5389 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | StoneBr | Norm | Norm |
| 10 | 527162130 | 189000 | 60 | RL | 60 | 7500 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 11 | 527163010 | 175900 | 60 | RL | 75 | 10000 | Pave | NA | IR1 | Lvl | AllPub | Corner | Gtl | Gilbert | Norm | Norm |
| 12 | 527165230 | 185000 | 20 | RL | ? | 7980 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 13 | 527166040 | 180400 | 60 | RL | 63 | 8402 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 14 | 527180040 | 171500 | 20 | RL | 85 | 10176 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | Gilbert | Norm | Norm |
| 15 | 527182190 | 212000 | 120 | RL | ? | 6820 | Pave | NA | IR1 | Lvl | AllPub | Corner | Gtl | StoneBr | Norm | Norm |
| 16 | 527216070 | 538000 | 60 | RL | 47 | 53504 | Pave | NA | IR2 | HLS | AllPub | CulDSac | Mod | StoneBr | Norm | Norm |
| 17 | 527225035 | 164000 | 50 | RL | 152 | 12134 | Pave | NA | IR1 | Bnk | AllPub | Inside | Mod | Gilbert | Norm | Norm |
| 18 | 527258010 | 394432 | 20 | RL | 88 | 11394 | Pave | NA | Reg | Lvl | AllPub | Corner | Gtl | StoneBr | Norm | Norm |
| 19 | 527276150 | 141000 | 20 | RL | 140 | 19138 | Pave | NA | Reg | Lvl | AllPub | Corner | Gtl | Gilbert | Norm | Norm |
| 20 | 527302110 | 210000 | 20 | RL | 85 | 13175 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | NWAmes | Norm | Norm |
| 21 | 527358140 | 190000 | 20 | RL | 105 | 11751 | Pave | NA | IR1 | Lvl | AllPub | Inside | Gtl | NWAmes | Norm | Norm |
| 22 | 527358200 | 170000 | 85 | RL | 85 | 10625 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | NWAmes | Norm | Norm |
| 23 | 527368020 | 216000 | 60 | FV | ? | 7500 | Pave | NA | Reg | Lvl | AllPub | Inside | Gtl | Somerst | Norm | Norm |
| 24 | 527402200 | 149000 | 20 | RL | ? | 11241 | Pave | NA | IR1 | Lvl | AllPub | CulDSac | Gtl | NAmes | Norm | Norm |

*https://www.kaggle.com/c/house-prices-advanced-regression-techniques*

***To answer these questions they need to conduct time series analysis and forecasting.***

DEAKIN BUSINESS SCHOOL

# Time Series: Data Preparation

*Create an "index" for the time series*

| PID | Time_Month_No | Yr_Sold | Mo_Sold | SalePrice ↓ | Lot… | |
|---|---|---|---|---|---|---|
| 528351010 | 133 | 2017 | 1 | 755000 | 215… | [illegible] |
| 528320050 | 139 | 2017 | 7 | 745000 | 156… | [illegible] |
| 528320060 | 127 | 2016 | 7 | 625000 | 357… | [illegible] |
| 528164060 | 146 | 2018 | 2 | 615000 | 127… | [illegible] |
| 528110020 | 159 | 2019 | 3 | 610000 | 136… | [illegible] |
| 527216080 | 138 | 2017 | 6 | 591587 | 519… | [illegible] |
| 528360050 | 125 | 2016 | 5 | 584500 | 172… | [illegible] |
| 528110090 | 157 | 2019 | 1 | 582933 | 138… | [illegible] |
| 527212030 | 127 | 2016 | 7 | 556581 | 160… | [illegible] |
| 528106020 | 160 | 2019 | 4 | 555000 | 15431 | 105 |
| 528176030 | 158 | 2019 | 2 | 552000 | 14836 | 100 |
| 527210040 | 129 | 2016 | 9 | 545224 | 18062 | 60 |

*Order time series by index*

| Row No. | Time_Month_No | median(SalePrice) | count(PID) |
|---|---|---|---|
| 1 | 121 | 176700 | 18 |
| 2 | 122 | 188050 | 24 |
| 3 | 123 | 149500 | 51 |
| 4 | 124 | 141576 | 48 |
| 5 | 125 | 1570… | [illegible] |
| 6 | 126 | 1550… | [illegible] |
| 7 | 127 | 1567… | [illegible] |
| 8 | 128 | 1877… | [illegible] |
| 9 | 129 | 1970… | [illegible] |
| 10 | 130 | 1425… | [illegible] |
| 11 | 131 | 2090… | [illegible] |
| 12 | 132 | 1610… | [illegible] |

*There are many specialist TS pre-processing operations, e.g. replacement of missing values or aggregation*

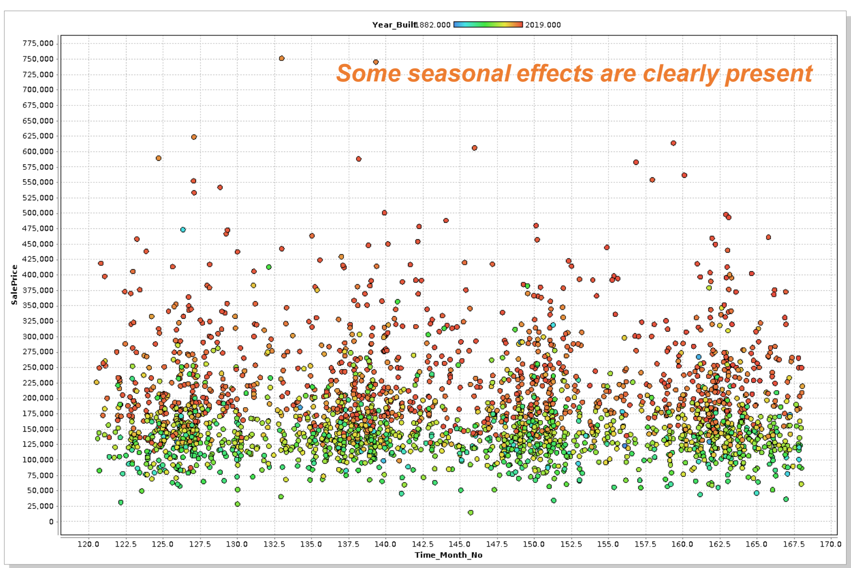
| Row No. | Time_Month_No | median(SalePrice) | count(PID) |
|---|---|---|---|
| 1 | ? | ? | ? |
| 2 | ? | ? | ? |
| 3 | 123.000 | 162565.200 | 43.200 |
| 4 | 124 | 158225.200 | 59 |
| 5 | 125.000 | 151965.200 | 78.600 |
| 6 | 126 | 159615.200 | 77.400 |
| 7 | 127.000 | 170700 | 76.000 |
| 8 | 128 | 167800 | 70.800 |
| 9 | 129 | 178600 | 57.600 |
| 10 | 130 | 179450 | 38 |
| 11 | 131 | 174400 | 35.200 |
| 12 | 132.000 | 166800 | 32.600 |

*Smooth a time series (if required) for its exploration*

- **As with any data set, time series needs some pre-processing for its effective use.**
- **Time series must have an "index", which can be date, time, time-stamp, or some quantity to be considered as sequencing your data set.**
- **A series can be defined over non-time entities and events, e.g. vehicle manoeuvres along the length of the road, shape of a landscape in a photograph, or a DNA sequence.**
- **Time series needs to be sorted by its index, usually in an increasing order.**
- **We then need to decide on the time dependent attributes that will be modelled and their values forecast.**
- **Ensure that an index if of an appropriate granularity, e.g. years, months, weeks, days, hours, etc.**
- **The selected time series attributes may need to be aggregated to the required index granularity.**
- **In an univariate case only a single attribute is analysed.**
- **In a multivariate case, several time series attributes are being modelled.**

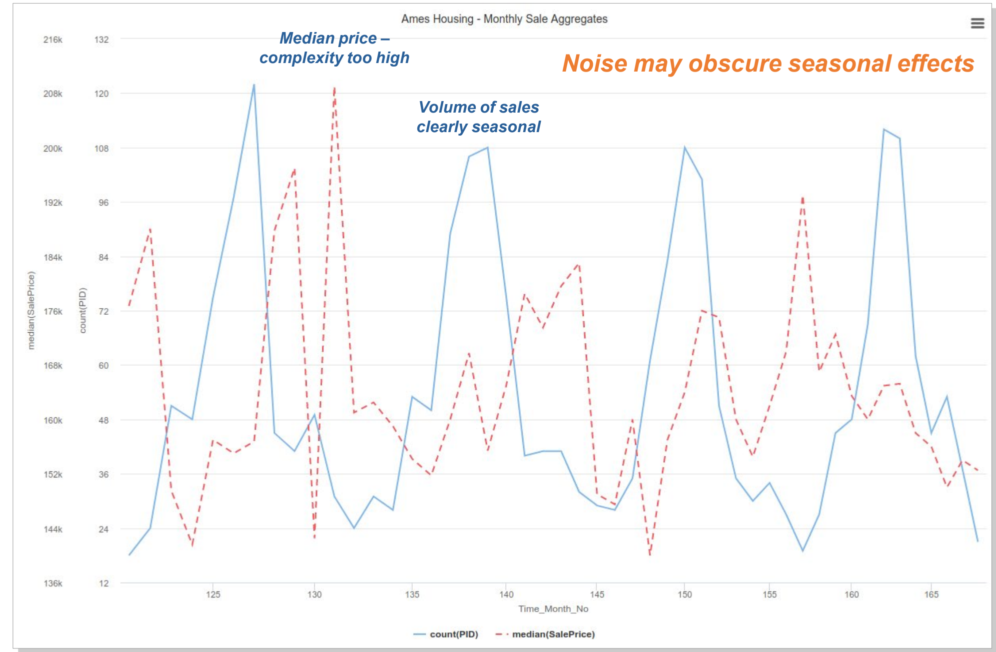
# The Initial Exploration: Seasonality

- A time dimension often implies seasonality in data
- Seasonality may not be obvious unless correct aggregation and smoothing of data is applied first
- Once a seasonal component is identified it is also possible to see the trend in the time series
- A seasonal component is common in business datasets involving time, e.g. sales, supply/demand forecasts
- The distribution of a Stationary time series does not change when shifted in time

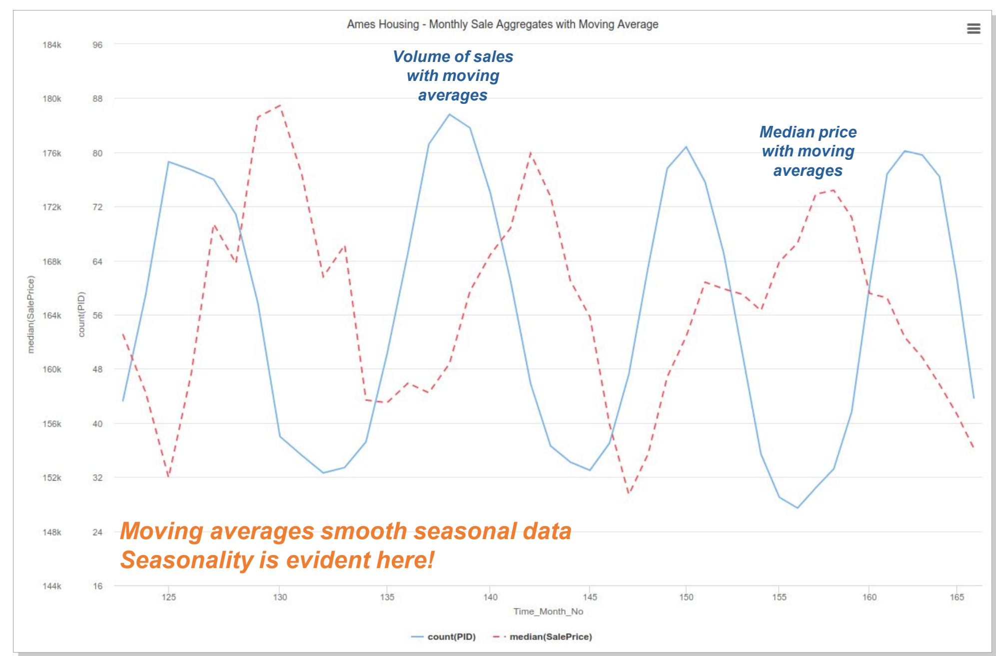



## The elusive seasons

*Data aggregation sometimes reveals seasonal effect – if they exist*

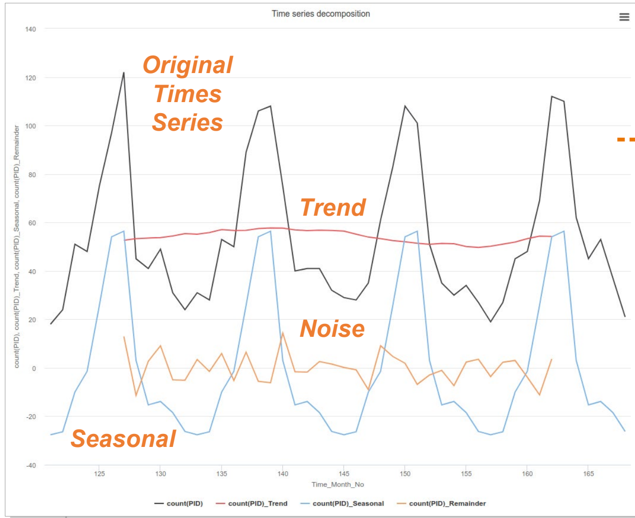


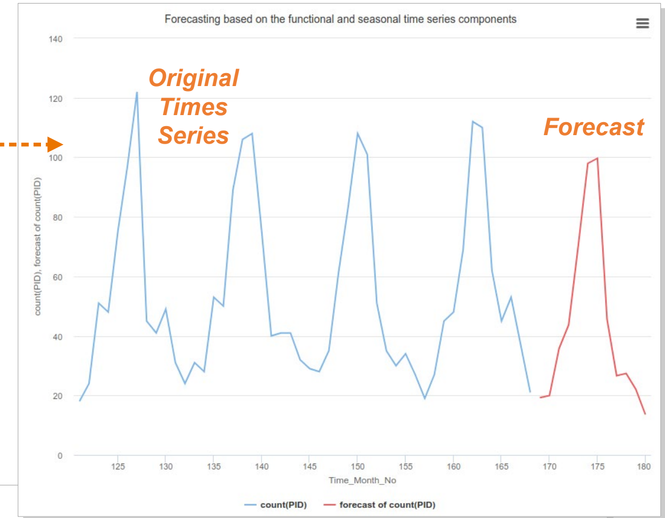

*Smoothing by moving (rolling) averages Highlights seasonal features*




DEAKIN BUSINESS SCHOOL

# Time Series Decomposition: Trend by smoothing or LOESS

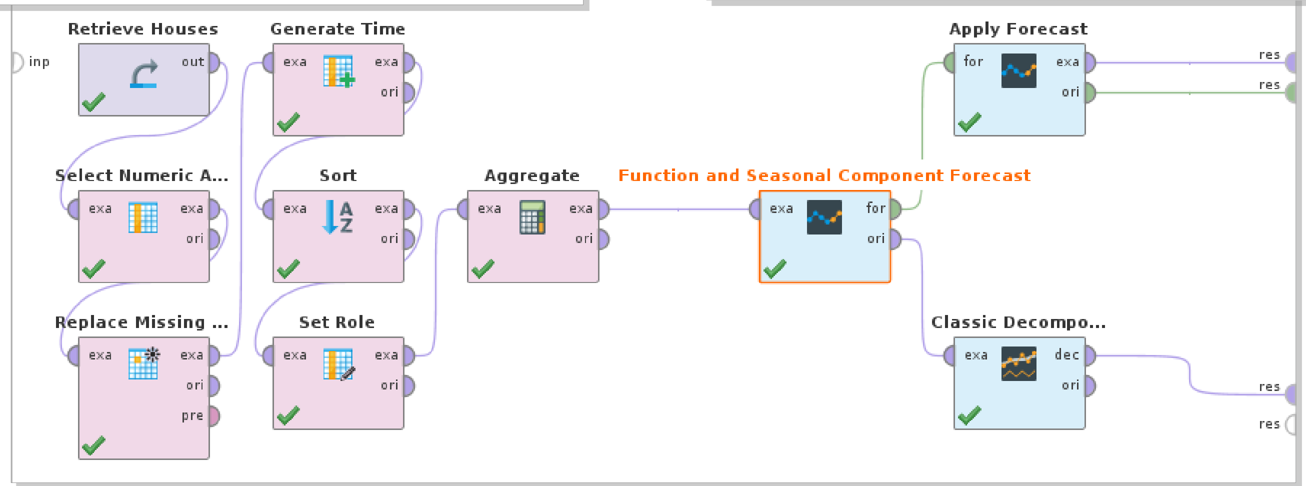



*Forecasting on decomposed time series is simple: forecast values can be calculated from the trend function and adding a seasonal component.*

*Non-stationary time-series can be transformed to make them stationary, e.g. by using a Log transformation*

*Alternatively, you can apply seasonal differencing, i.e. subtract from a given value, the value from the previous season*




*There are two ways of decomposing the time series, i.e. Classical which finds a trend by smoothing, or LTS which uses non-parametric LOESS interpolation*

- *Forecasting* (normally) makes it is possible to decompose the series into its components, I.e. *trend* (by smoothing), *seasonal* (by removing the trend and averaging), and *noise* (remainder obtained by subtracting the trend and seasonal components).
- Two modes of operation are possible, i.e. *additive* ts = trend + seasonal + noise, when seasonal and noise components do not depend on trend, and *multiplicative* ts = trend x seasonal x noise, otherwise.
- The additive model is preferred when the seasonal variation is relatively constant over time. The multiplicative model is better when the seasonal variation increases over time.

Also see KD 12.1-2 on Time-Series Decomposition and Smoothing Methods

DEAKIN BUSINESS SCHOOL

# Autoregressive forecasting

- Makes use of the autocorrelation concept; the values of one attribute are correlated with the values of the same attribute in the past
  - The variable is regressed on its **own lagged** (prior) values; because of the correlation
  - The lag can be any number of time periods (depending on the data)
  - Can be determined from experience/intuition
  - Can be determined through application of several lag sizes

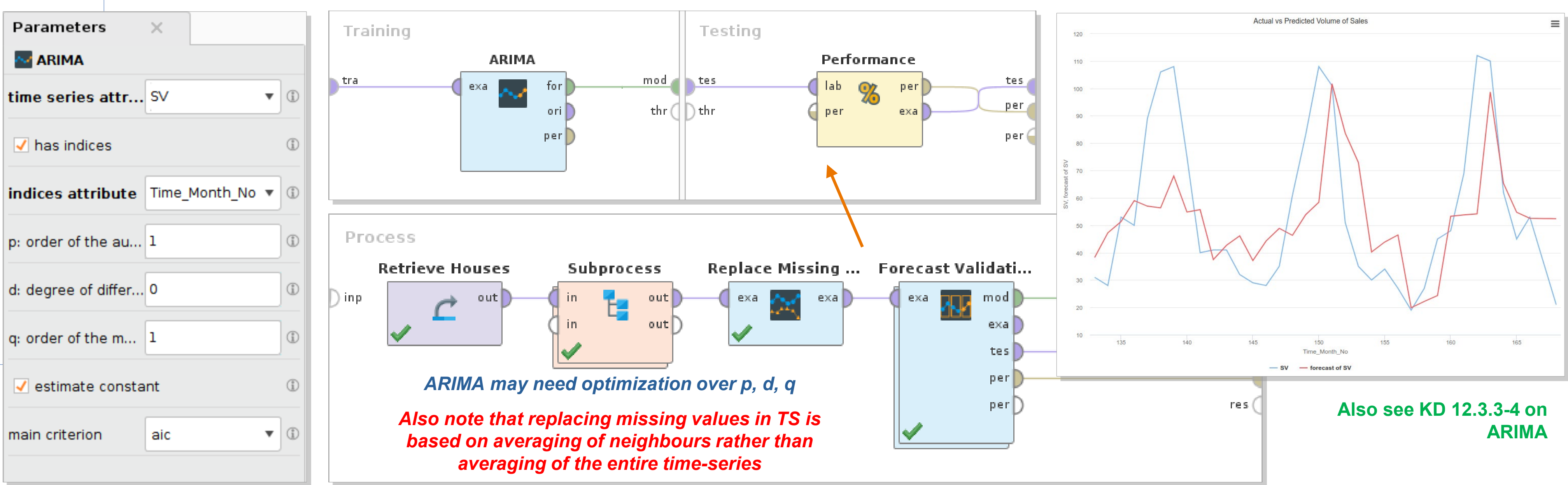
lag=6 months

the values are autocorrelated (values in the different columns are strongly correlated)

DEAKIN BUSINESS SCHOOL

# ARIMA

- ***Autoregressive*** models assume that future time-series values are a linear combination of the past *p* values.
- Non-stationary time-series can be made stationary by differencing, i.e. subtracting the previous TS value from the current.
- ***First-order differencing*** subtracts the consecutive data points.
- If the time-series is still non-stationary, this process can be repeated leading to ***second-order differencing***, third-order, …, *d*-order, etc.
- ***Seasonal differencing*** can also be applied, by subtracting values from the same period of the consecutive seasons.
- Autoregressive models are trained over the time-series. At each step they generate some error. The error itself can also be expressed as a linear combination of the past errors, viewed as a ***moving average of past q errors***.
- ***ARIMA (Autoregressive Integrated Moving Average) combines all of the above elements.***
- ***ARIMA(p,d,q)***: ***AR (p)*** indicates that the time-series is regressed on its past *p* data points. ***MA (q)*** suggests that the forecast error is a linear combination of its past *q* errors. ***I (d)*** shows that the time-series can be made stationary by *d*-degree differencing.



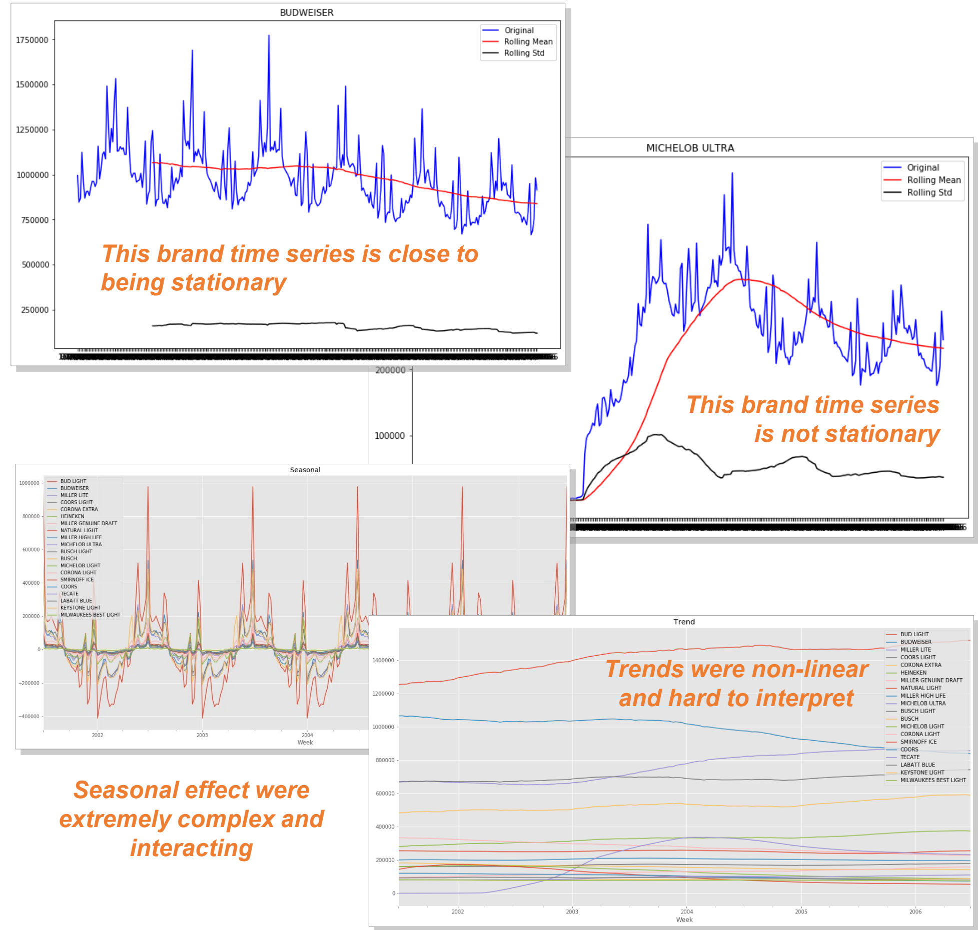

*ARIMA may need optimization over p, d, q*

*Also note that replacing missing values in TS is based on averaging of neighbours rather than averaging of the entire time-series*

**Also see KD 12.3.3-4 on ARIMA**

# Multi-Variate Forecasting

- **Traditional statistical methods of forecasting from multi-variate data are extremely complex**
- **They are based on linear models and require that all time-series are tested for stationarity and co-integration**
- **This could not always be achieved**
- **However, we deployed deep LSTM (Long-Short Term Memory) neural network, which can be used to analyse time-series with non-linear features**
- **Deep learning model proved very simple, was easy to train and its results were superior to those obtained from the traditional models**

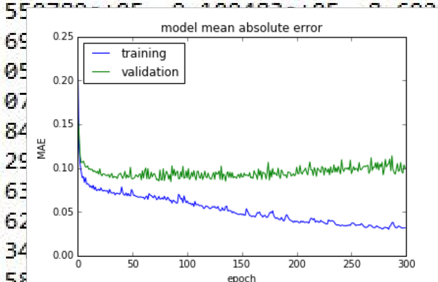



*Multi-variate methods required special encoding of sales data*

| | INPUT | | | | | | | | OUTPUT | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Brand1(t-4) | Brand2(t-4) | Brand1(t-3) | Brand2(t-3) | Brand1(t-2) | Brand2(t-2) | Brand1(t-1) | Brand2(t-1) | Brand1(t) | Brand2(t) |
| Week | | | | | | | | | | |
| 1118.0 | 1.109784e+06 | 9.953423e+05 | 8.734214e+05 | 8.475431e+05 | 8.943546e+05 | 8.690308e+05 | 1.213969e+06 | 1.123630e+06 | 1.011556e+06 | 9.550789e+05 |
| 1119.0 | 8.734214e+05 | 8.475431e+05 | 8.943546e+05 | 8.690308e+05 | 1.213969e+06 | 1.123630e+06 | 1.011556e+06 | 9.550789e+05 | 9.109483e+05 | 8.692257e+05 |
| 1120.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 9.650034e+05 | 9.058158e+05 |
| 1121.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 9.653520e+05 | 9.079342e+05 |
| 1122.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 9.294550e+05 | 8.843647e+05 |
| 1123.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 9.852230e+05 | 9.299287e+05 |
| 1124.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.044415e+06 | 9.634913e+05 |
| 1125.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.051271e+06 | 9.622128e+05 |
| 1126.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.025763e+06 | 9.343008e+05 |
| 1127.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.066392e+06 | 9.585089e+05 |
| 1128.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.302172e+06 | 1.158150e+06 |
| 1129.0 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1.145184e+06 | 9.759201e+05 |
| 1130.0 | 1.025763e+06 | 9.343008e+05 | 1.066392e+06 | 9.585089e+05 | 1.302172e+06 | 1.158150e+06 | 1.145184e+06 | 9.759201e+05 | 1.212033e+06 | 1.030224e+06 |
| 1131.0 | 1.066392e+06 | 9.585089e+05 | 1.302172e+06 | 1.158150e+06 | 1.145184e+06 | 9.759201e+05 | 1.212033e+06 | 1.030224e+06 | 1.275100e+06 | 1.113123e+06 |

```
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 200)               184800
_________________________________________________________________
dense_3 (Dense)              (None, 30)                6030
=================================================================
Total params: 190,830
Trainable params: 190,830
Non-trainable params: 0
```

*A simple LSTM deep learning model*



# Summary / Review

- Provide recent examples of time series data used in media
- Provide examples of forecasting used in media
- What data pre-processing is needed for time-series analysis?
- What is a time series index?
- Can time series analysis be applied to non-time data?
- What is the difference between univariate and multivariate time series?
- What is time series smoothing, what is it used for, and how can it be done?
- What is time series decomposition?
- What is a time series trend?
- What is seasonality?
- How can time series decomposition be used in forecasting?
- What is a window and horizon in autoregressive forecasting?
- How is time series forecasting different from multi-variate forecasting?
- How can anomalies be detected in time series?

DEAKIN BUSINESS SCHOOL