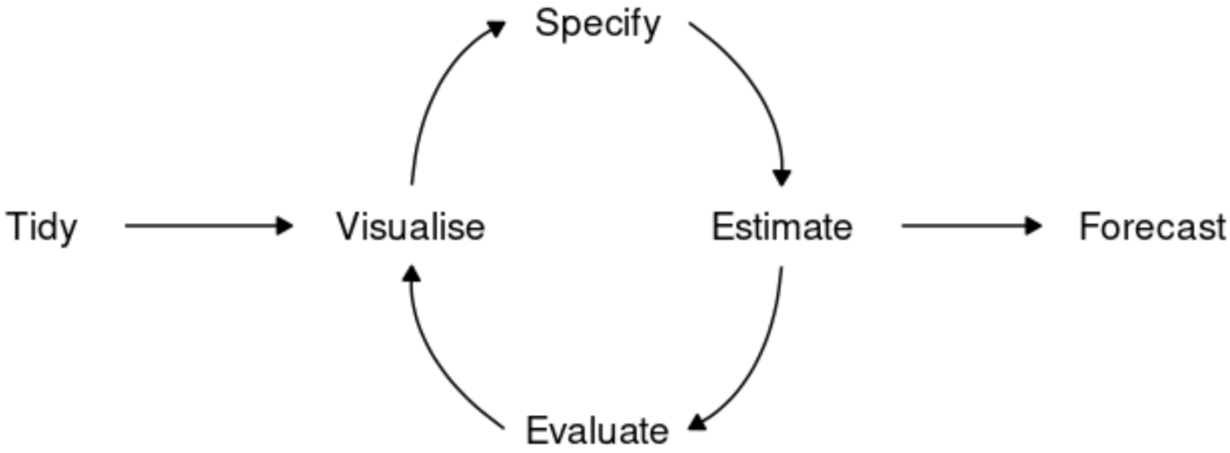
Week 8 - time series 2

Literature: FPP3 chapters 1-7

Today we will dive into chapters 4, 5, and 7, and start getting an idea of a time series analysis workflow.

Hyndman and Athanasoupolus suggest the following workflow.



As you are well aware, data science and model building are iterative processes of trying several different things out, testing assumptions, and optimizing models on training/validation data before making your final predictions/forecasts on the test data.

Last week you practiced the first step: getting your data into the right format for your analysis and perhaps you even got to make some baseline models to evaluate against. Today, we will learn to extract some features from time series as well as how to produce some simple regression models.

You will need to load the following datasets and packages

```
pacman::p_load(ggplot2, dplyr, tsibble, feasts, fpp3)
data("PBS")
data("global_economy")
data("hh_budget")
data("aus_retail")
```

Exercise 1 - Time Series Features

1. In the PBS data (data from pharmacies), calculate which series has the highest mean value of Cost
    a. Plot it.
    b. Should we do an additive or multiplicative decomposition of this series?
2. Calculate STL features of Costs.
    a. Which series shows strongest and weakest seasonality? Plot them.
    b. Which series are most and least trended? Plot them.

Exercise 2 - The Forecaster's Toolbox

1. Produce and plot forecasts for the following series using whichever of MEAN(y), NAIVE(y), SNAIVE(y) or RW(y ~ drift()) is more appropriate in each case. I.e. plot the time series and inspect it to see which method you believe is most appropriate:
    - Australian Population (global_economy)
    - Lambs in New South Wales (aus_livestock)
2. Perform residual diagnostic checks of your models.
    - Were you able to capture all information?
    - Do the residuals look like white noise; do they display autocorrelation?
3. Divide the hh_budget dataset into a training and test set, by withholding the last four years as test data.
    a. Fit all the appropriate benchmark models (those mentioned in 1) for household *Wealth* on the training data, and forecast on the test data.
    b. Calculate the performance of your forecasts. Which method performs best?
    c. Check the residuals of the best model for each country. Do they resemble white noise?
    d. Choose one of the countries (Australia/Canada/Japan/USA) and see if you can beat the benchmarks you just made using TSLM. Experiment with trend(), season() and knots.
4. Divide the Australian takeaway food turnover (aus_retail) into a training and test set by withholding the last four years as test data.
    a. Fit all the appropriate benchmark models as well as a TSLM on the training data using time series cross validation and evaluate their performance.
    b. Train the same models on the full training data and calculate the performance on the test data. How do the two methods compare in terms of performance measures?

Done? Work on the exercise from yesterday's lecture.

Tips and ideas for lecture exercise

The papers from the M4 and M5 competitions contain a lot of useful knowledge for time series analysis. Some of the main points are:

- (Seasonal) naive models tend to perform *really* well and should always be created as a baseline.
- Combining/ensembling multiple models tends to boost forecasting performance.
- Having a suitable cross-validation regime is imperative for obtaining trustworthy estimates of model performance.
- Hybrid approaches (between statistical/time series specific models and machine learning models) seem to perform best.

Remember that transforming or decomposing your data is often an essential part of time series analysis. Experiment with log/Box-Cox transformations, STL decompositions or other methods from your readings.