




Feature Engineering (I)

SGA07_DATASCI

30th January 2020



Module Overview

- Split-Apply-Combine Strategy
- Data cleaning: Missing Values
- Data cleaning: Smooth out Noise



Book Keeping

- Direct any technical questions to TA
- Spend some time to build programming skills
- Expect 2 mini group-based projects
- Catch up on Tasks/Practice Labs so far



Outcome

After this Module, you will;

- Get an overview of Split-Apply-Combine strategy for data analysis
- Introductory to data cleaning as a preprocessing task



Split-Apply-Combine Strategy

- Used in data preprocessing, modelling and visualisation.
- Similar to Map-Reduce strategy popularised by Google

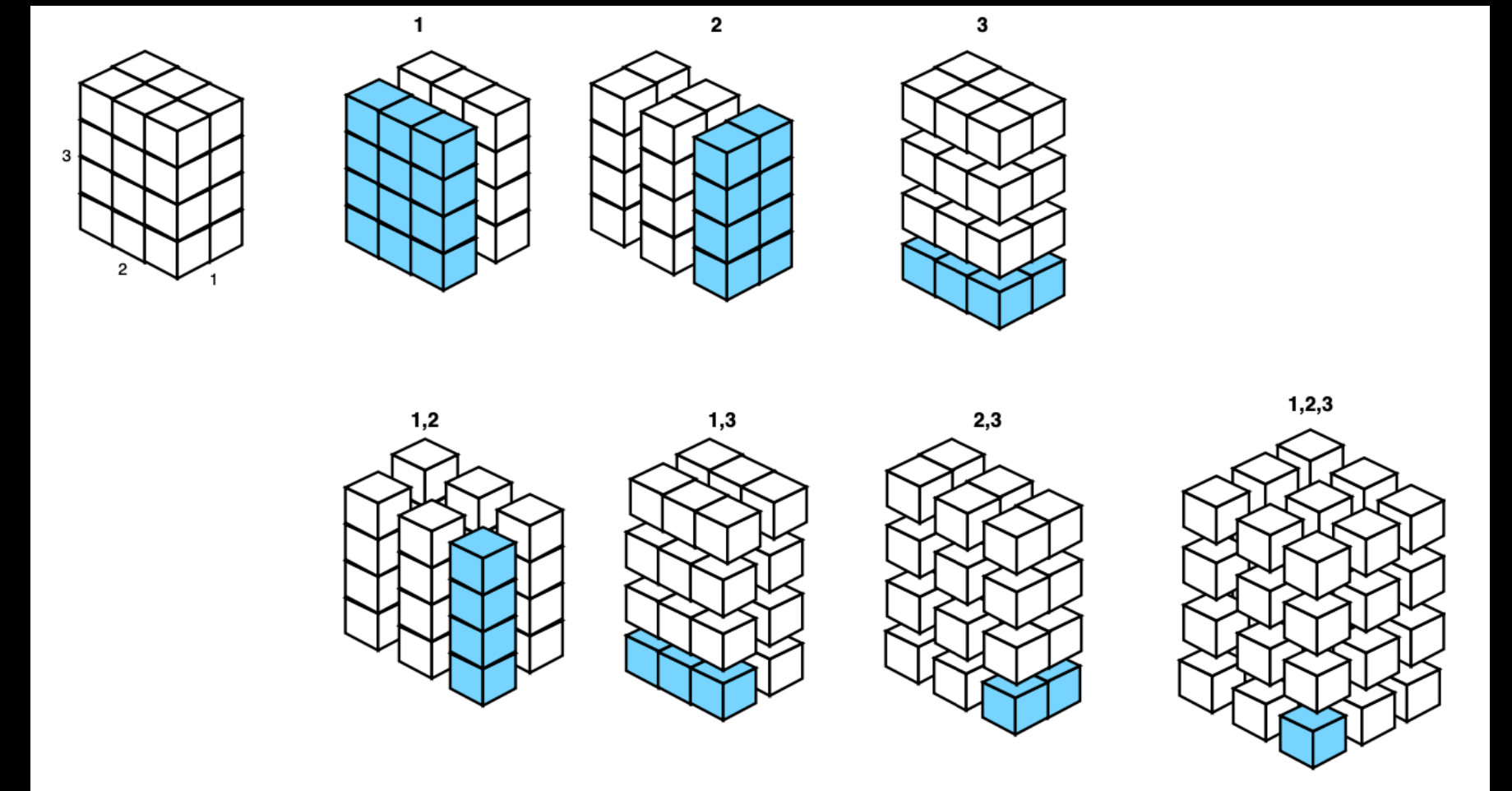
“

Break up a big problem into manageable components, operates on each pieces independently and then put all the pieces together

”

Split-Apply-Strategy

- Keep Data Structure consistent
- Reduce cognitive efforts
- Reduce book-keeping efforts



Plyr Package

- Package in R provides an intuitive way to use the split-apply-combine strategy
- Each type of input has different rules for how to split it up, and these rules are described in detail in the following sections. In short:
 - Arrays are sliced by dimension in to lower-d pieces: `a*ply()`.
 - Data frames are subsetted by combinations of variables: `d*ply()`.
 - Each element in a list is a piece: `l*ply()`.
- The output type defines how the pieces will be joined back together and how they will be labelled.

	Array	DF	List	Discard
Array	aapply	adply	alply	a_ply
DF	dapply	ddply	dlply	d_ply
List	lapply	ldply	llply	l_ply



Practice Lab

Implement split-apply-combine strategy to get summary statistics

Use the following Instructions:

- Create a new script `sac,r` in your root directory
- Create a data frame of years and random counts different years
- Use loop method to calculate the mean of each year count
- Use `plyr` method to calculate the mean of each year



Data Cleaning

- Fill out missing values
- Smooth out noise
 - Outlier analysis
- Data inconsistencies

“

Data cleaning tasks help to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

”



Fill our Missing Values “

- Ignore the observations with missing values
- Fill in the values manually
- Use a global constant to fill the missing values
- Use a statistic (measure of central tendency - mean, median or mode)
- Use a statistic for all samples belonging to the same class as the given observation
- Use the most probable value to fill in the missing value

Real-world data tend to be incomplete. It is not unusual to note that attributes of a dataset have no recorded values.

”

Smooth out Noise

- **Binning:** Binning methods smooth a sorted data value by consulting its “neighbourhood,” that is, the values around it.
- **Regression:** Data smoothing can also be done by regression, a technique that conforms data values to a function.
- **Outlier analysis:** Outliers may be detected by clustering, for example, where similar values are organised into groups, or “clusters.” Intuitively, values that fall outside of the set of clusters may be considered outliers.

“

Real-world data tend to be noisy. Noise is a random value that is recorded for a measured variable that do not comply with the general behaviour of the variable.

”



Data Inconsistencies

- As a result of human or system errors
- Leverage on existing knowledge
 - Expert systems
 - Exploratory analysis
- Consistent use of code
- Consistent use of data representation

“

Data quality assurance is to remove bias and use historical data to create actionable recommendations and predictions for the future. But this only works if the data is of high quality to begin with.

”



Practice Lab

Implement different method to fill in missing values

Use the following Instructions:

- Create a new script `data_cleaning.r` in your root directory
- Create a data frame that applies 3 methods of filling in missing values
- Join data frame and compare your statistics



Recap/Summary

At the end of this Module, you should understand;

- How to apply strategy to breakdown work into manageable components, apply your functions and combine your result for presentation
- The importance of data cleaning as a preprocessing task either to fill in missing values, smooth out noise or generally improve the quality of your data



Suggested Material (Programming)

- <https://www.edx.org/course/cs50s-introduction-to-computer-science>
- <https://www.hackerearth.com/practice/python/functional-programming/functional-programming-1/tutorial/>
- <https://www.coursera.org/learn/python>



Suggested Material

- Data Mining Concepts and Techniques (3rd Edition) by Jiawei Han, Micheline Kamper and Jian Pei: Chapter 3 - Data Preprocessing