

FIT3152 Data analytics. Tutorial 04

1. Slide 45 lists 3 models for data analytics: KDD, SEMMA and CRISP-DM. Describe each of them and outline the origin, main similarities and differences between models. You can use these Wikipedia pages as a starting point.

http://en.wikipedia.org/wiki/Data_mining

<http://en.wikipedia.org/wiki/SEMMA>

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Tips

2. Briefly read: A Taxonomy of Dirty Data. (Springer link will work from University: <http://link.springer.com/article/10.1023%2FA%3A1021564703268>)

(a) Simplify the taxonomy by making groups of errors you think are closely related.

(b) Choose 10 specific error types and give an example of each.

Tips

3. Briefly read: Tidy Data <http://www.jstatsoft.org/v59/i10/paper> and summarize the main principles of tidy data.

Tips

4. Using lecture notes and/or Chapter 12 of R for Data Science <http://r4ds.had.co.nz/> as a guide, manually transform the table below to put it into tidy form. Write out the first 10 or so lines of the transformed table.

Student	English31	English32	Maths31	Maths32
Anna	50	-	77	69
Bobby	-	52	-	47
Carl	5	30	-	55
Duy	37	80	18	10
Enid	82	-	96	58
Fey	73	36	63	-
Geoff	95	72	13	90

Tips

5. The data file “Dunnhumby1-20.csv” is a cut down and modified set of test data from the Kaggle competition to predict when consumers would next visit a Dunnhumby supermarket and how much they would spend. See: <http://www.kaggle.com/c/dunnhumbychallenge> for more information. The current modified data set contains the customer ID, Date of visit, Date since last visit (visit_delta), and Spend for 20 customers from the test set.

- a. Using the customer spend data from the beginning of September 2010 to the end of March 2011 investigate whether there is any difference between the amount spent by customers having the most predictable relationship for time between visits *vs* spend, and the least. To do this first calculate the coefficient of determination (r^2) of visit_delta and visit_spend for each customer. Using r^2 as your criterion create two groups of 10 customers: those with the most predictable visit_spend *vs* visit_delta (highest r^2) and those with the lowest. By

comparing the average spend of customers in each group determine whether it is possible to see a difference between groups. Is this difference significant? At what level?

- b Over the same time period investigate whether customers who spend the most in total have a greater number of visits to the store than those who spend the least. To do this create two groups of 10: those having the highest spend in total and those with the lowest. You can now compare the number of visits made by each customer in each of the two groups. Hint: you might want to use the “length” function to count the number of visits made by each customer.

Tips

- 6 The data file “govhackerlectricitytimeofusedataset.csv” has been created from the .txt file originally available as part of the Australian Government’s data resources. See link at: <https://data.gov.au/dataset/sample-household-electricity-time-of-use-data>. The file contains the smart meter records for a number of households recorded at 30 minute intervals over varying periods of time. The first few rows of the csv file are below.

CUSTOMER_KEY	End Datetime	General Supply KWH	Off Peak KWH	Gross Generation KWH	Net Generation KWH
8170837	4/04/2013 11:59	0.137	0	0	0
8170837	4/04/2013 12:29	0.197	0	0	0
8170837	4/04/2013 12:59	0.296	0	0	0
8170837	4/04/2013 13:29	0.24	0	0	0
8170837	4/04/2013 13:59	0.253	0	0	0
8170837	4/04/2013 14:29	0.24	0	0	0
8170837	4/04/2013 14:59	0.238	0	0	0
8170837	4/04/2013 15:29	0.225	0	0	0
8170837	4/04/2013 15:59	0.246	0	0	0

The columns of interest are “Customer_Key” (meter), “End Datetime”, and “General SupplyKWH” (power used each 30 mins).

Using the 30 minute general supply, calculate the daily supply for each meter for every day there is data available. Because the number of records is unreliable you will also need to count the number of daily observations for each (day, meter). You should then discard any (day, meter) readings that do not have the complete number of observations.

Draw a boxplot of the daily consumption for each meter in January 2013 by meter.

Extension, now exclude those meters that do not have a complete set of readings over the month of January (that is 31 days).

Tips

7. Analyse the Anscombe data set (anscombe). This data set is part of the base R installation and consists of 4 pairs of x,y observations.
- (a) Using summary statistics and correlation describe the main similarities and differences between the pairs.
- (b) Now, using some visual analysis describe the similarities and differences between the pairs.