Data Mining Assignment – Tips

Created by Dave Eargle  
Last Updated 4/5/2016

Data Source: R, reshape2::tips

Download: https://vincentarelbundock.github.io/Rdatasets/csv/reshape2/tips.csv

Description from the reshape2 documentation:

**Description**

One waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables:

**Format**

A data frame with 244 rows and 7 variables

**Details**

* tip in dollars,
* bill in dollars,
* sex of the bill payer,
* whether there were smokers in the party,
* day of the week,
* time of day,
* size of the party.

In all he recorded 244 tips. The data was reported in a collection of case studies for business statistics (Bryant & Smith 1995).

**References**

Bryant, P. G. and Smith, M (1995) *Practical Data Analysis: Case Studies in Business Statistics*. Homewood, IL: Richard D. Irwin Publishing:

# Learning Objectives

Learn how to do the following data mining and analysis tasks in RapidMiner:

* Correlation
* Plotting: Scatterplots
* Attribute Generation
* Linear Regression
  + Model Performance, including squared correlation (*R*2)

# Before you Start

* Complete all of the RapidMiner tutorials
* Complete the following chapters’ exercises “Data Mining for the Masses”, Dr. North, 1st edition. Available to purchase on Amazon or for free PDF download [here](http://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf).
  + Chapter 4: Correlation
  + Chapter 8: Linear Regression

# Questions

* What is the association between day and time. Are customers really more likely to come in at lunch during the week, and dinner on the weekends?
  + Show the correlation between day and time
  + Plot the relationship between day and time. Use “size” as the column color. Remember to use a “Jitter.”
* Calculate a tip ratio (tip / total\_bill)
  + What is the average tip ratio? Highest? Lowest? Plot the distribution of tip ratios.
  + Are females better tippers? Do bigger groups tip with better ratios?
  + Effect of time of day?
  + Does bill size impact tip ratio? (it shouldn’t, right?)
* Let’s look at the tips.
  + Show a scatterplot. Plot total\_bill against tip, and use your tip\_ratio calculated value as the color column.
  + What is the average tip ratio, approximately?
    - On the raw dataset, it’s 16.1%
  + Let’s see who was the most giving, tip-ratio wise.
    - What is the highest tip ratio? What was the total\_bill for the highest tip ratio, and what was the tip amount?
      * Highest was 71% tip ratio.
      * Describe the observation with the highest tip ratio amount.
        + Hint: You can quickly get the row id for a point by changing one of the axes to “att1” (the row id), removing any jitter, then hovering over the point until you see its coordinates. One of the coordinates will be “att1”, the row id. Or, just sort the data by your tip\_ratio column.
      * Row id 173, Male tipper, smoker, Sunday, dinner time, two in party
  + What was the highest tip amount? Describe this observation
    - Row id 171. Tip $10. Total bill over $50. Male tipper, smoker, Saturday, dinnertime. 3 in party.
* Perform a linear regression predicting tip\_ratio.
  + Hint: In RapidMiner, you will need to convert all polynomial attributes over to be numerical in order for the regression to run. An easy way to do this is to use the “Nominal to Numerical” operator.
  + Hint: By default, RapidMiner, will select a subset of your attributes to fit the best model. For this assignment, set the “feature selection” parameter of the Linear Regression operator to “None”
  + Interpret the relationship between total\_bill and the tip ratio
    - Answer: for every dollar increase in total\_bill, the tip ratio is predicted to be .8% lower.
  + What is the squared correlation (*R*2) for this linear regression?
    - Answers may vary depending on the split, but it should be about 87.5%
  + What impacted the tip ratio: Was there a difference between genders? Smokers? Day of the week?
    - Answer: Compared to non-smokers, smokers are predicted to tip 2% more.
    - Answer: Sunday was the highest-tipping day.
    - Answer: No difference between Lunch and Dinner time on tip ratio.
    - Answer: Group size didn’t predict tip ratio.
    - Answer: higher tips were associated with higher tip ratios. 5% increase for every additional dollar tipped. So higher tips aren’t entirely explained by higher total bills!