MIS40970 Assignment 1
Orange Juice Sales

Prof Michael O'Neill
School of Business
University College Dublin
Ireland
m.oneill@ucd.ie

1 Dataset

The dataset (`oj.csv`) is available from Blackboard, and contains weekly sales figures for three chilled orange juice brands from 83 shops based in Chicago. The data is based on Rossi's **bayesm** R package [1] and was used earlier by [2]. A list of the attributes follows:

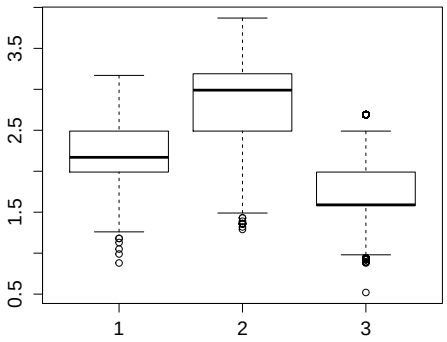
- `store`: store number
- `brand`: brand indicator
- `week`: week number
- `logmove`: log of the number of units sold
- `feat`: feature advertisement
- `price`: price of brand
- `STORE`: store number
- `AGE60`: percentage of the population that is aged 60 or older
- `EDUC`: percentage of the population that has a college degree
- `ETHNIC`: percent of the population that is black or Hispanic

- `INCOME`: median income
- `HHLARGE`: percentage of households with 5 or more persons
- `WORKWOM`: percentage of women with full-time jobs
- `HVAL150`: percentage of households worth more than $150,000
- `SSTRDIST`: distance to the nearest warehouse store
- `SSTRVOL`: ratio of sales of this store to the nearest warehouse store
- `CPDIST5`: average distance in miles to the nearest 5 supermarkets
- `CPWVOL5`: ratio of sales of this store to the average of the nearest five stores

2 Assignment

Use R to explore the orange juice dataset (`oj.csv`) to undertake the tasks detailed below. Please submit a report (pdf) through Blackboard by 17h00 Wednesday 10 February. This assignment is worth 10% of the total credit for MIS40970. **In your submission, please make sure that in each case you include the R code which was used to address each item below. Also, include the output (if any) generated.**

1. Load the oj.csv data into R.
2. How many records and how many attributes are in the orange juice dataset?
3. What is the mean, standard deviation and range of the price of orange juice?
4. What is the median of the "log of number of units sold" (`logmove`)?
5. What are the names of the 3 orange juice brands? (*Hint: brand is of type factor, and a factor is comprised of a number of ...<`insert function name here!`>...*)
6. Create a histogram of prices for each brand of orange juice.
7. Generate a boxplot, which includes a separate plot for the prices of each brand (*Hint: the boxplot should look something like the example below*).

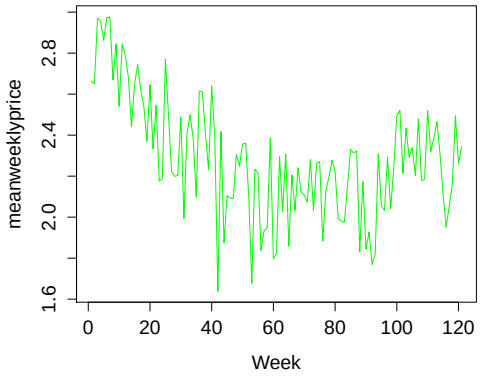


8. What does the boxplot tell us about the relative prices of each brand? (*Hint: answer this one by interpreting what you observe in the boxplot, i.e., no coding required!*).

9. Generate a scatterplot of the `logmove` compared to `price`, and color the points according to their brand (e.g., "red" for brand 1, "green" for brand 2 etc).

10. Based on what you observe in the scatterplot, what can we say about the price of each brand of chilled orange juice and the volume of sales?

11. Calculate the mean price of orange juice sold each week, and create a line plot of this timeseries. (*Hint: use* ***`tapply()`****, and an example plot is provided below.*)



12. Extract the mean weekly price of orange juice sold each week according to each brand (*Hint: use* ***`tapply()`*** *with a* ***`list`*** *for the* ***`INDEX`***)

13. Create a plot which compares the mean weekly price of orange juice for all brands versus each individual brand.

14. When there is an advertising campaign for orange juice does it impact on the number of units sold? (*Hint: use* ***`factor()`*** *to ensure* ***`feat`*** *is a factor! and use* ***`tapply`*** *as per the last question*).

15. Can you create a line plot of the mean weekly units sold without a promotion overlayed with the mean weekly units sold with a promotion? What is interesting about this plot?

16. Consider the demographic and competitive variables[1]. **Using descriptive analytics** are there patterns you can observe that might suggest the potential for profiling individual stores or customers, which might then be used for marketing purposes? Please

    (a) describe your approach (and provide code) to uncover potential patterns,

    (b) why the findings may (or may not) be of significance?, and

    (c) how they might be applied by the marketing team?

References

[1] Peter Rossi (2012). "bayesm: Bayesian Inference for Marketing/Microeconometrics". `http://cran.r-project.org/web/packages/bayesm/index.html`

[2] Alan L. Montgomery (1997). "Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data," *Marketing Science* 16(4) 315-337.

[1] `AGE60,EDUC,ETHIC,INCOME,HHLARGE,WORKWOM,HVAL150,SSTRDIST,SSTRVOL,CPDIST5,CPWVOL5`