R project #1

Contents

Orange juice data mining

This data set gives:

- Weekly sales data
- Refrigerated 64-ounce Orange juice containers
- 83 stores in the Chicago area for 121 weeks
- 3 brands of juices Dominicks, MinuteMaid, and Tropicana

- There are many stores throughout the city, many time periods, and also three different brands
- The data are arranged in rows, with each row giving the recorded store sales as well as brand, price, presence/absence of feature advertisement, and the various demographic characteristics of the store
- There are 28,947 rows (records) and 18 columns (attributes or variables) in this data set
- The data is taken from P. Rossi's bayesm package for R, and it has been used earlier in Montgomery (1987)
- Values such as sales are in logarithmic form i.e. 'logmove' variable since data set has been *normalized*

Packages used

1. tigerstats
2. lattice
3. latticeExtra
4. ggplot2
5. stats
6. stats4
7. labeling
8. formatR

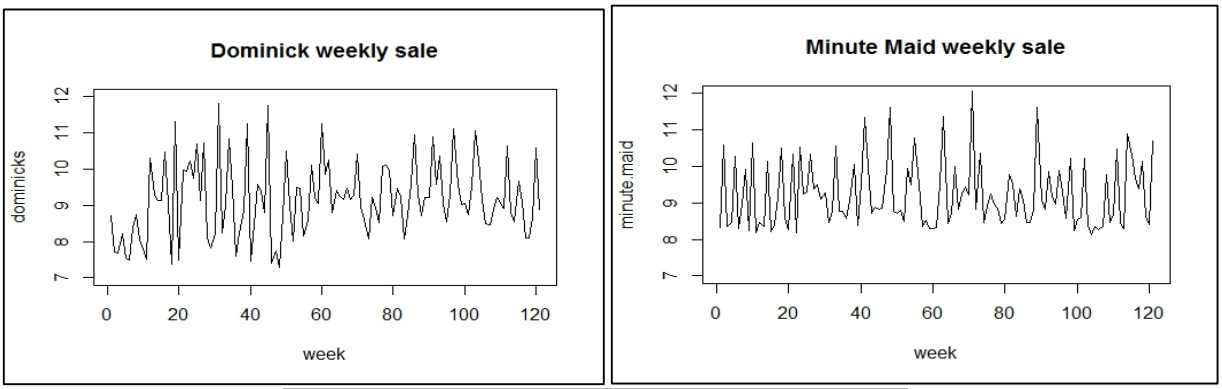
Objective

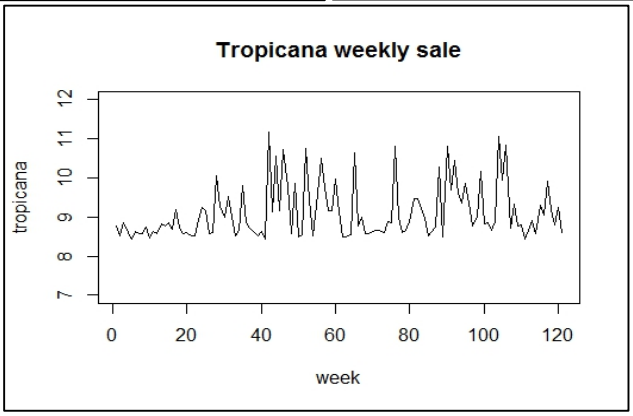
The objective of this project is to analyze the effect of the variables on the sale of the 3 brands of Juices and suggest recommendations as per the insights obtained from the analysis.

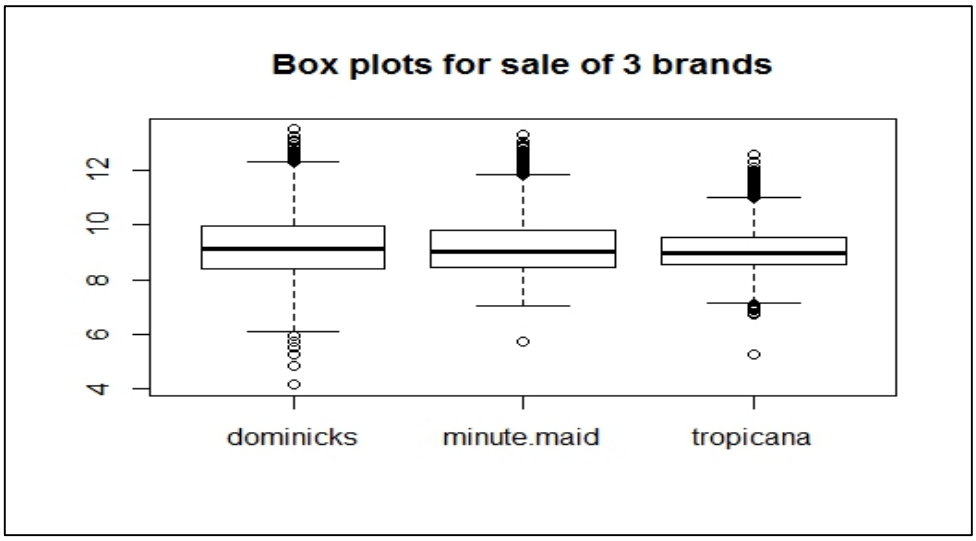
Data analysis and key findings

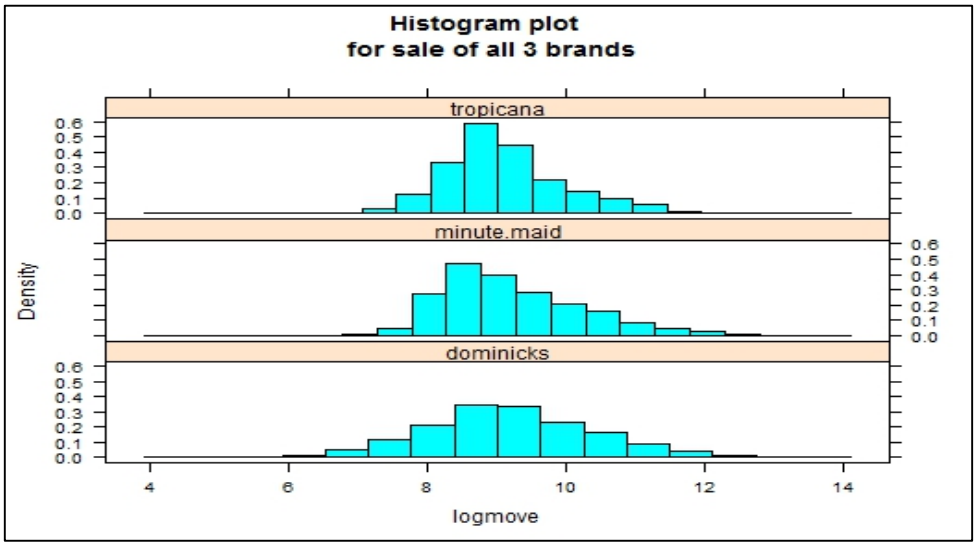
- Time sequence plots of weekly sales, averaged over all 83 stores are shown for the three brands. These plots are created by first obtaining the average sales for a given week and brand (averaged over the 83 stores). For this, R function 'tapply' is used
- Time sequence plots of the averages are then graphed for each brand, and the plots are arranged on the same scale for easy comparison. An equivalent display, as 3 panels on the same plotting page, is produced through the xyplot function of the 'lattice' package. When all 3 plots are displayed are together it is easier to analyze the data
- Then box plots, histograms, and smoothed density plots for sales, stratified for all 3 brands are drawn. These plots average the information across the 83 stores and the 121 weeks given in the data set

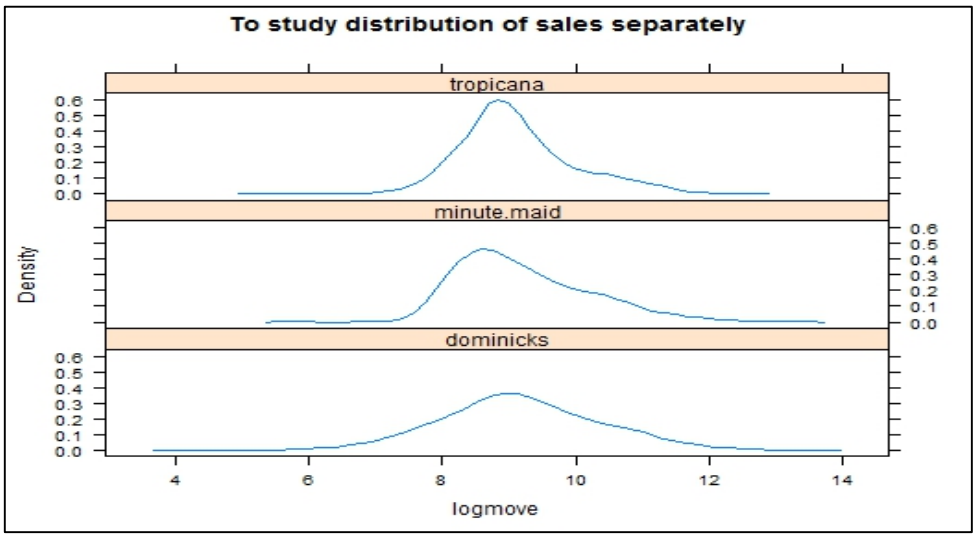
Plots with presence of feature advertisements – NOT taken into consideration

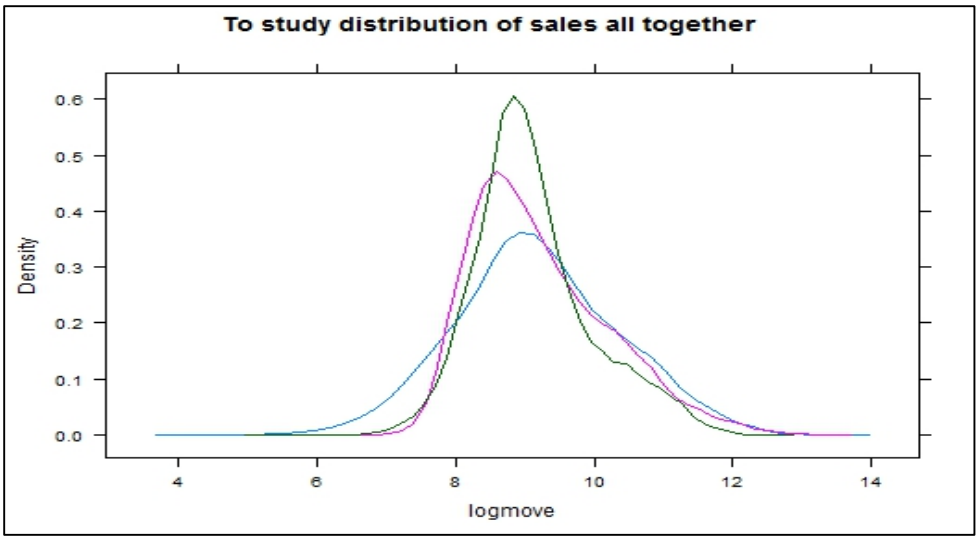




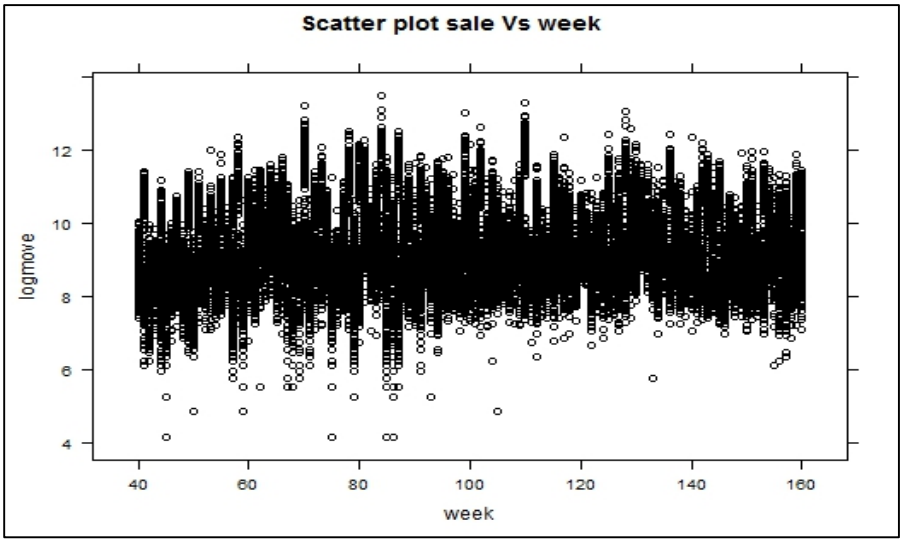

Box plots for sale of 3 brands
4
6
8
10
12
dominicks
minute.maid
tropicana

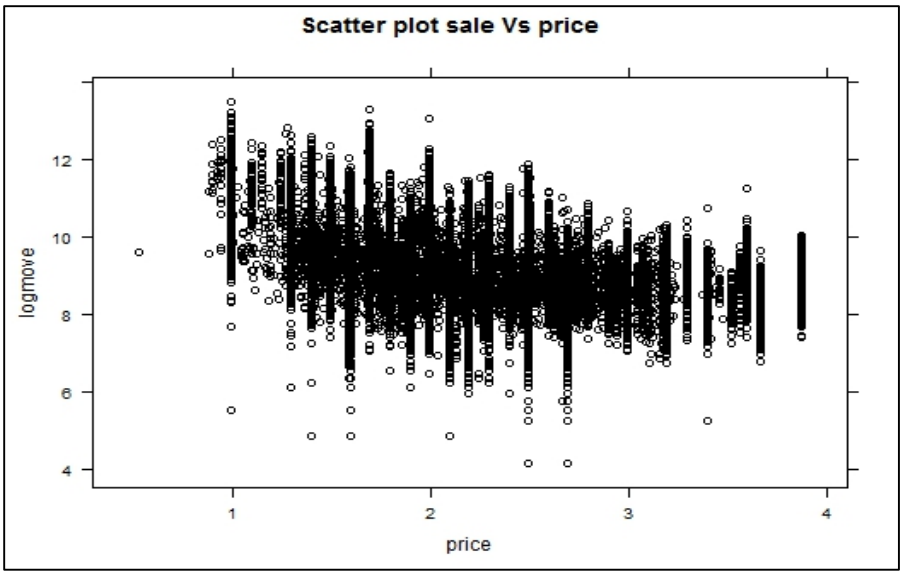

Histogram plot
for sale of all 3 brands
tropicana
minute.maid
dominicks
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4
6
8
10
12
14
logmove

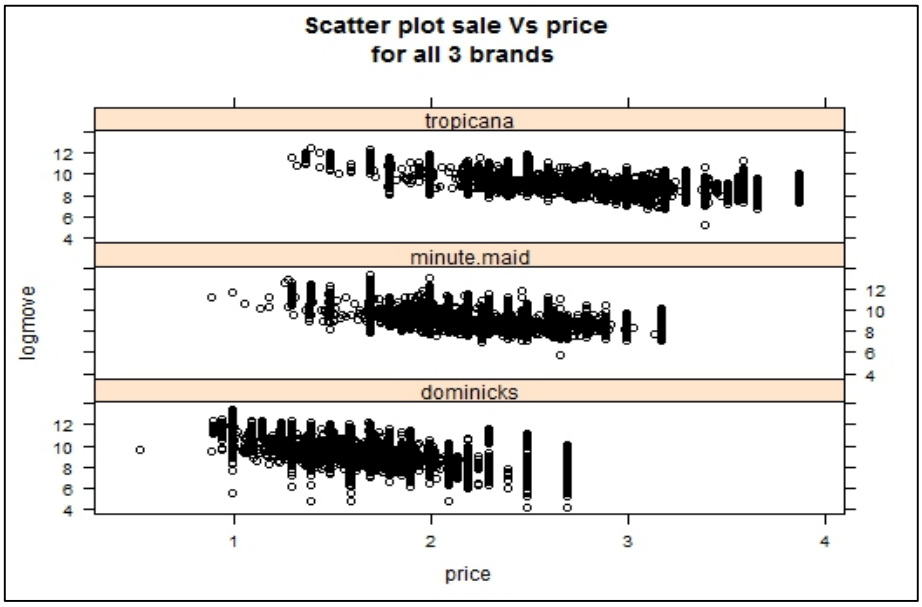

To study distribution of sales separately
tropicana
minute.maid
dominicks
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4
6
8
10
12
14
logmove

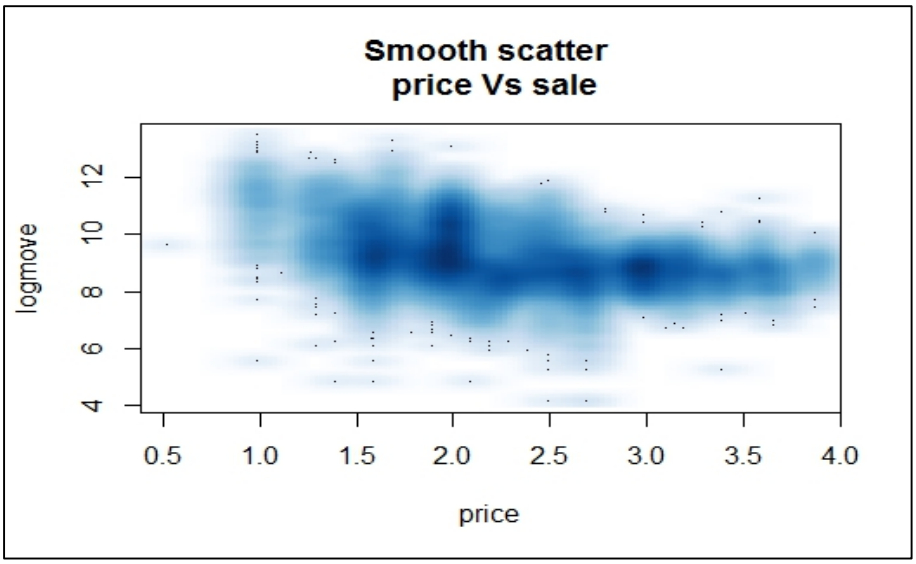


- Graphs of sales against price for each brand separately but aggregating over weeks and stores
- The graph shows that sales decrease with increasing price



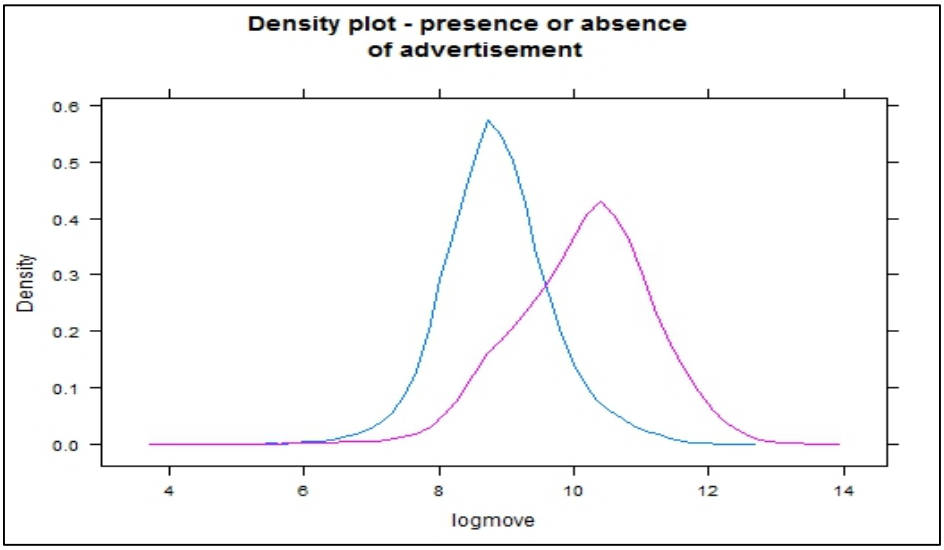

Scatter plot sale Vs price
logmove
price
12
10
8
6
4
1
2
3
4

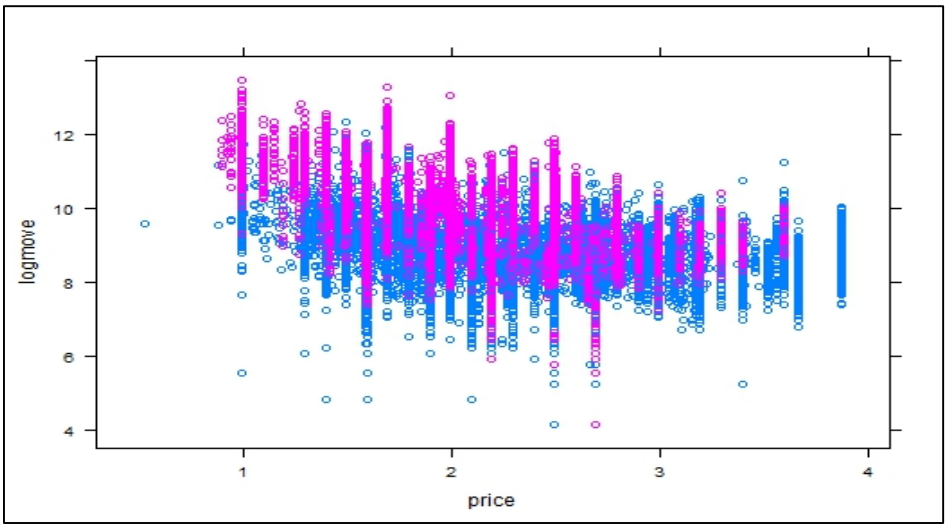

Scatter plot sale Vs price
for all 3 brands
tropicana
minute.maid
dominicks
logmove
price
12
10
8
6
4
1
2
3
4



Price with presence/absence of feature advertisements – taken into consideration

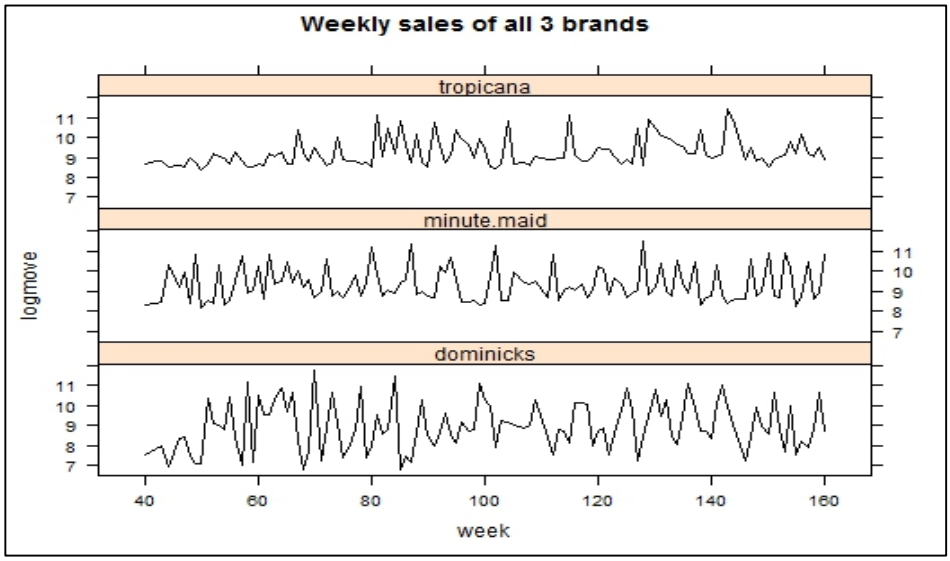
- A density plot of sales for weeks with and without feature advertisement
- A scatter plot of sales against price with the presence of feature advertisement indicated by different colors
- Both plots show positive effect of feature advertisement on sales

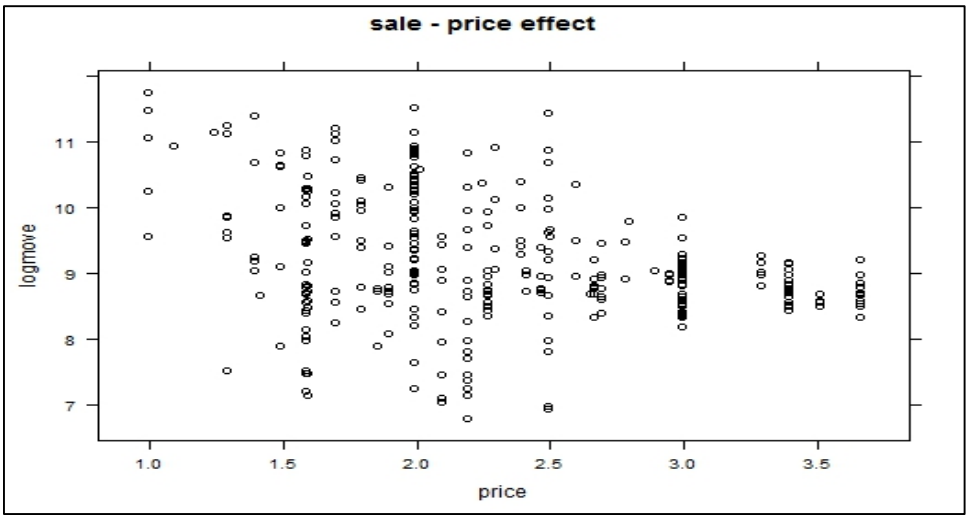


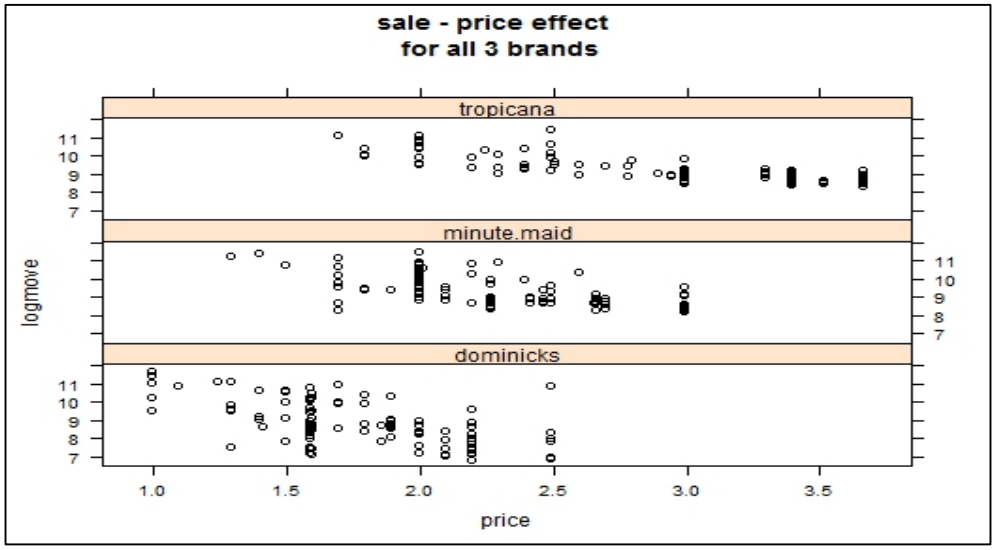


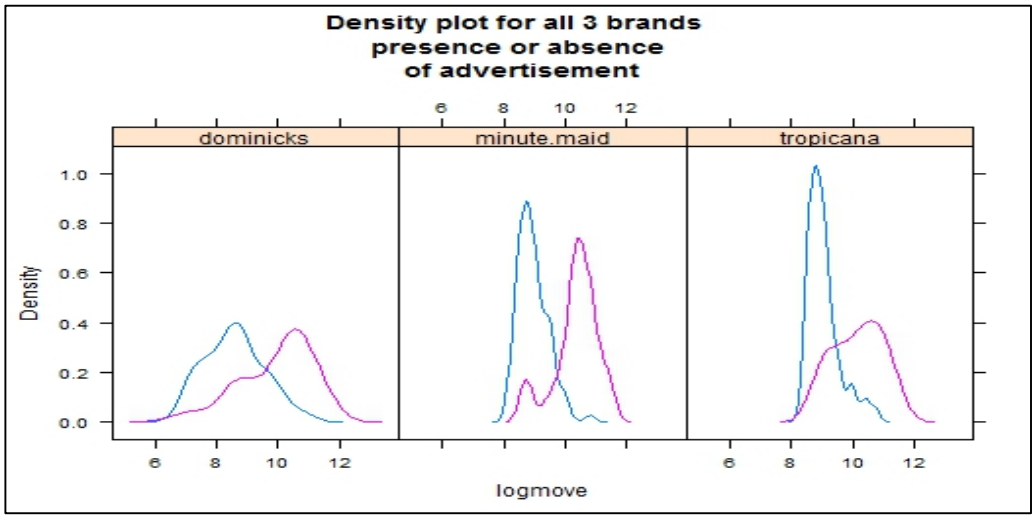
Considering one particular store – store 5

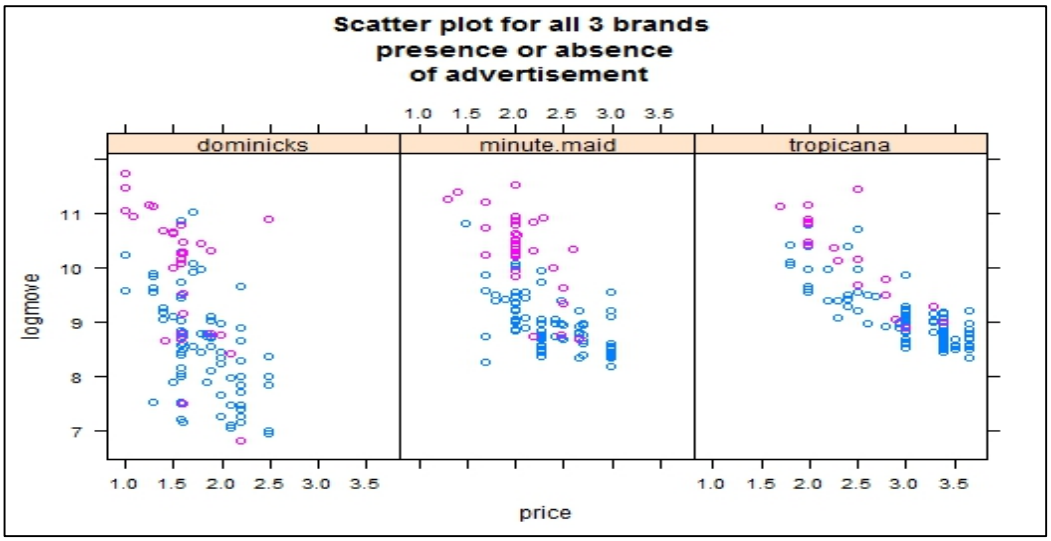
- Time sequence plots of the sales of 'store 5' are shown for all 3 brands
- Scatter plots of sales against price, separately for the 3 brands
- Sales decrease with increasing price
- Density histograms of sales and scatter plots of sales against price, with weeks with and without feature advertisement coded in different colors, are shown for each of the 3 brands
- Again, these graphs show very clearly that feature advertisement increases the sales



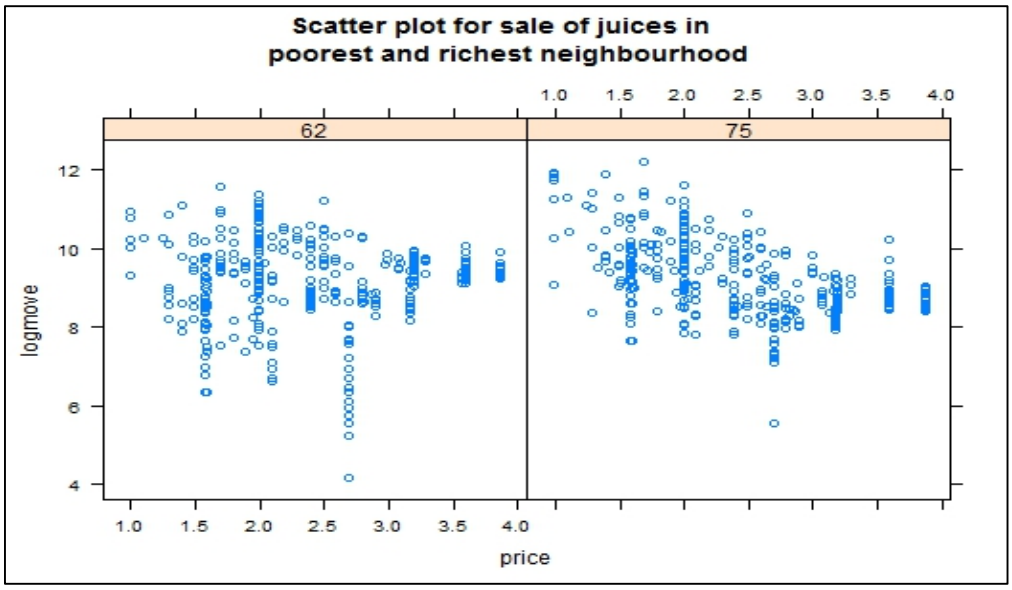

sale - price effect
logmove
price
1.0
1.5
2.0
2.5
3.0
3.5
7
8
9
10
11

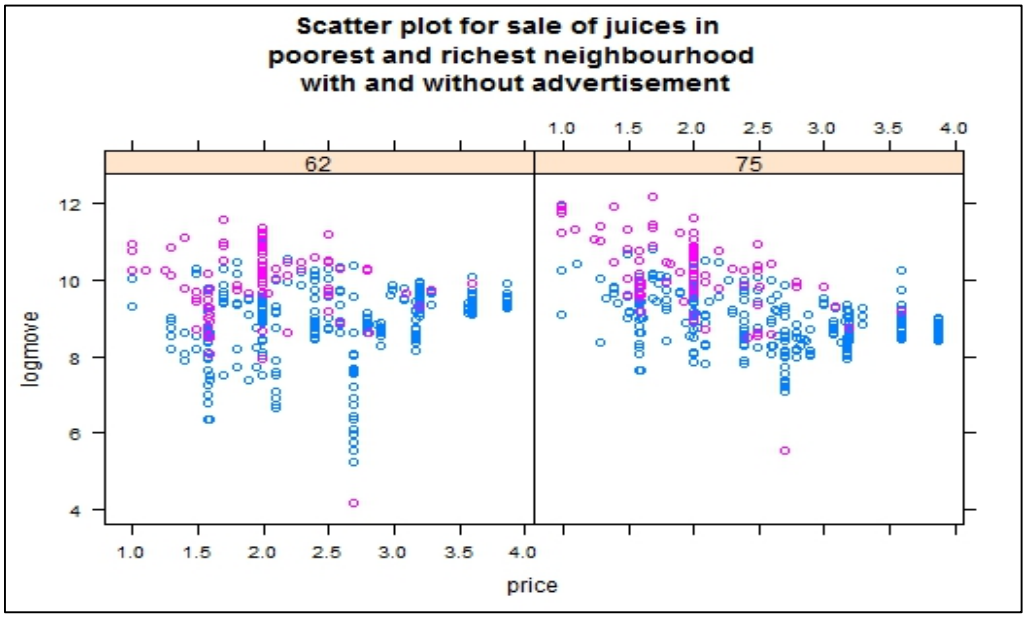

sale - price effect
for all 3 brands
tropicana
minute.maid
dominicks
logmove
price

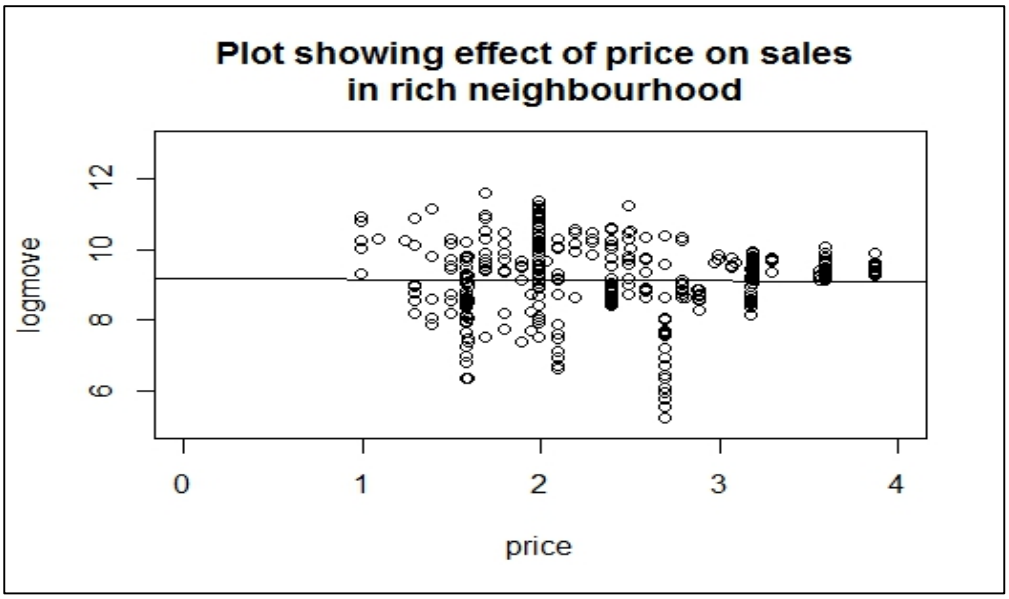

Density plot for all 3 brands
presence or absence
of advertisement
dominicks
minute.maid
tropicana
Density
logmove

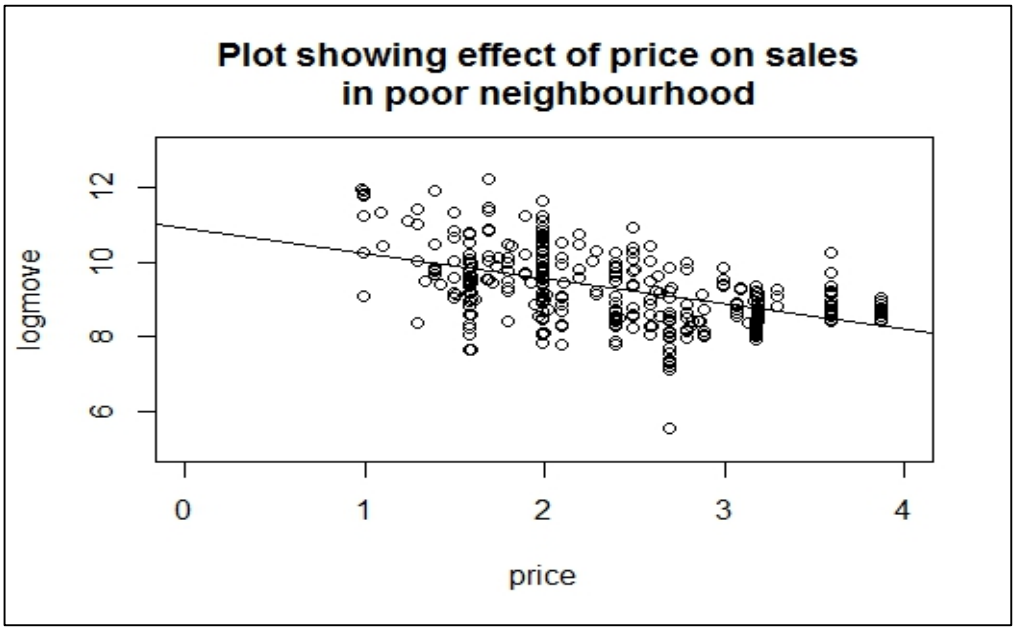


- The volume of the sales of a given store certainly depends on the price that is being charged and on the feature advertisement that is being run
- In addition, sales of a store may depend on the characteristics of the store such as the income, age, and educational composition of its neighborhood
- Whether the sensitivity (elasticity) of the sales to changes in the price depends on the income of the customers who live in the store's neighborhood, can also be assessed
- Price elasticity can be largest in poorer neighborhoods as poorer customers have to watch their spending budgets more closely. To follow up on this hypothesis, analysis of stores which are in the wealthiest and the poorest neighborhoods can be performed
- On analysis, it is found the store 62 is in the wealthiest area, while store 75 is in the poorest one
- Lattice scatter plots of sales versus price, are plotted, on separate panels for these two stores, with and without the presence of feature advertisements
- In order to get a better idea about the effect of price on sales, the best fitting (least squares) line to the graph is added
- The slope of the fitted line is more negative for the poorest store, indicating that its customers are much more sensitive to changes in the price


Scatter plot for sale of juices in
poorest and richest neighbourhood
62
75
logmove
price


Scatter plot for sale of juices in
poorest and richest neighbourhood
with and without advertisement
62
75
logmove
price





Inferences/Recommendations

- Feature advertisements are effective for this product. It should be considered to continue this practice whenever possible to boost the sale
- Pricing is important for this product. As price increases, sale decreases
- Income, age and educational composition also affect the sale. It can inferred that in rich neighborhoods, increase in price does decrease the sale significantly; whereas in poorer neighborhoods, sale decreases with increase in price. Hence, the price of the product could be mitigated accordingly
- Psychological pricing could be used i.e. to set the price in such a way that it ends with the digit 9
- Special seasoning pricing, promotional offers and deals could be used for the same

- It is important to note that the study is limited to the data provided and may not give a comprehensive picture of all the factors driving the sales of all the brands of orange juices. For example, we don't have data about pricing and promotion activity of competitor retailers that could possibly affect sales.

Future Scope

This data set could be used to investigate clustering. We may want to learn whether it is possible to reduce the 83 stores to a smaller number of homogeneous clusters. Furthermore, we may want to explain sales as a function of explanatory variables such as price, feature advertisements, and the characteristics of the store neighborhood.
In particular, we may want to study whether the effects of price changes and feature advertisements depend on demographic characteristics of the store neighborhood.
