# Prediction of Consumer Demand of Beer

# Summary

The IRI data set was analyzed in an attempt to predict various aspects of consumer demand related to beer. Key drivers were identified towards understanding sales by day of the week, value of promotions, and forecasting future sales. When preparing the data for these tasks, there was a severe contraction in the amount of data with each join due to the inconsistency in presence of data in the different fields.

For sale by day of the week, R^2 of 5 models that averaged ~0.84 was achieved. The dominant features were number of units, items sold, and vendor—promotions had a minor effect.

Value of promotions were examined by aggregating the data on a weekly basis; coupons or rebates did have a positive benefit, price reductions had a negative benefit towards revenue. Displays were not significant.

For prediction of future sales, models were created based on sales from 2008-10, including a 5 term linear model, and demonstrated high performance (r^2 > 0.95) for predicting the sales of 2011.

# Methodology

The CRISP-DM methodology was applied with some modifications:

* Deployment stage was not performed.
* Evaluation was not a separate step but part of each modelling iteration.

## Business Understanding

Our team took on the 1st option which was:

*Create a model of consumer demand to forecast future (on a daily bases) sales of particular product(s) for a subset of stores. Show how these might vary over different (types of) stores. Compare how this prediction might change under (deep) discount/promotion period.*

### Problem statements

The 3 items we chose to concentrate on were:

1. Discover the important features for predicting beer sales revenue (in dollars) on a daily basis. This meant splitting the data into sales for each day and modelling each separately.
2. Assessing whether and which promotions provide value ie lead to increased sales revenue.
3. Creating a model that could predict future sales.

## Data Understanding

Kevin and Conway

## Data Preparation

Kevin and Conway

### Aggregation by weeks

There was some realization that the way the data was arranged, by transaction would likely not lead to novel findings, as sales in dollars would likely be mostly of a function of number if items sold times and the item type (which would have a prescribed price), and there was risk that the promotions would only have a minor effect.

Therefore, a transformation was performed on the data to split it into weekly bins arranged in order by time, with aggregations of the features, either by sums or counts (for features that could not sum). This was done using excel on the “final” 66K line file that was generated in the python data preparation activities.

This led to some of the modelling that assessed promotions as well as forecasting future sales.

## Modelling

### Python

Kevin and Conway

#### Day by Day model activity

Data was initially split into the transactions that occurred on a daily basis (Sunday vs Monday vs Tuesday, etc).

#### Feature importances

As the tree based sklearn methods have the ability to show feature importances, random forest was used. Kevin found this.

Another method used was the selectKbest feature selection which uses F-tests to detect the most significant features.

#### Promotions – for 1 store

### Weka / JMP

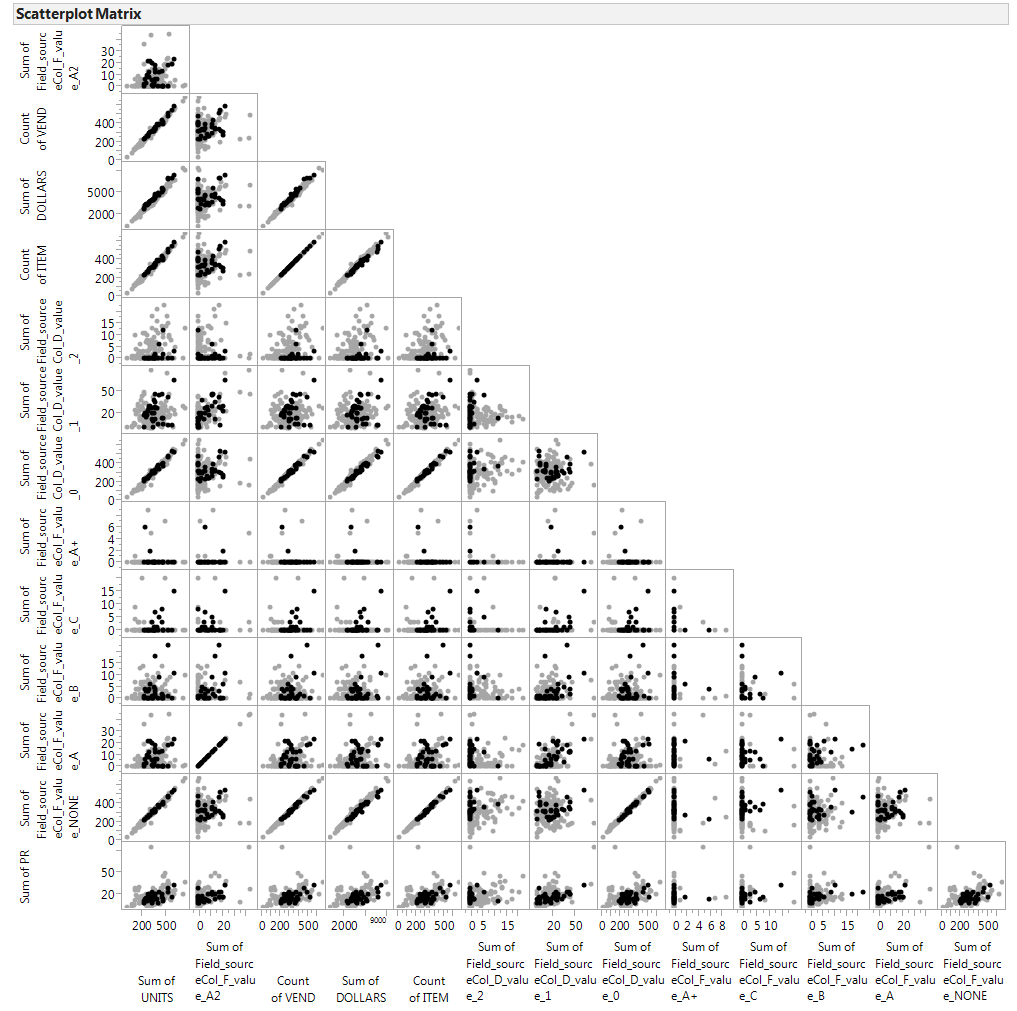
Beyond Python related model evaluation, Excel, Weka and JMP were used.

Excel was used to a) aggregate the data into weekly buckets and b) later, after the “best” model was chosen for sales forecasting, individual predicted values were calculated, as well as the r^2 model assessment.

#### Feature relationships

Since this data transformation was performed “after” the data understanding phase, some aspects of it were performed at this phase, particularly univariate comparisons of all the parameters. This was done in JMP for ease of display. See Figure X below and the appendix. It was evident that there were several highly correlated parameters:

* Count of Items with “no promotion” features, “no display” features, sum of units, sum of dollars, count of vendors
* Count of vendors with sum of units and “no promotion” features
* Sum of dollars with sum of units and “no promotion” features
* Sum of units with “no promotion” features
* “No display” features and “no promotion” features



#### Feature importance assessment

In weka, two types of attribute selection routines were run. The point of this was not to immediately eliminate the features that were indicated not to be important, but to couple that information with the correlations as a guide to which features could be eliminated. For Best First + CFSsubset Eval and then WrapperSubsetEval using M5Rules and GreedyStepwise search the following results were attained:

|  |  |  |
| --- | --- | --- |
| Best First + CFSsubset Eval | WrapperSubsetEval using M5Rules and GreedyStepwise search | ~Agree |
| 10(100 %) 1 Week number | 2( 20 %) 1 Week number |  |
| 10(100 %) 2 Sum of UNITS | 9( 90 %) 2 Sum of UNITS | Agree |
| 10(100 %) 3 Sum of Field\_sourceCol\_F\_value\_A2 | 4( 40 %) 3 Sum of Field\_sourceCol\_F\_value\_A2 |  |
| 10(100 %) 4 Count of VEND – correlated, plan to delete | 10(100 %) 4 Count of VEND | Agree |
| 10(100 %) 5 Count of ITEM | 2( 20 %) 5 Count of ITEM |  |
| 0( 0 %) 6 Sum of Field\_sourceCol\_D\_value\_2 | 1( 10 %) 6 Sum of Field\_sourceCol\_D\_value\_2 | Agree |
| 3( 30 %) 7 Sum of Field\_sourceCol\_D\_value\_1 | 1( 10 %) 7 Sum of Field\_sourceCol\_D\_value\_1 |  |
| 3( 30 %) 8 Sum of Field\_sourceCol\_D\_value\_0 | 1( 10 %) 8 Sum of Field\_sourceCol\_D\_value\_0 |  |
| 0( 0 %) 9 Sum of Field\_sourceCol\_F\_value\_A+ | 2( 20 %) 9 Sum of Field\_sourceCol\_F\_value\_A+ |  |
| 8( 80 %)10 Sum of Field\_sourceCol\_F\_value\_C | 3( 30 %)10 Sum of Field\_sourceCol\_F\_value\_C |  |
| 9( 90 %)11 Sum of Field\_sourceCol\_F\_value\_B | 0( 0 %)11 Sum of Field\_sourceCol\_F\_value\_B | Agree |
| 8( 80 %)12 Sum of Field\_sourceCol\_F\_value\_A | 3( 30 %)12 Sum of Field\_sourceCol\_F\_value\_A |  |
| 0( 0 %)13 Sum of Field\_sourceCol\_F\_value\_NONE – correlated plan to delete | 9( 90 %)13 Sum of Field\_sourceCol\_F\_value\_NONE |  |
| 0( 0 %)14 Sum of PR | 3( 30 %)14 Sum of PR |  |

#### Modeling and Evaluation with all Features

Models run and results <table> and appendix from powerpoint.

#### Modeling and Evaluation with some correlated features eliminated

Models run and results <table> and appendix from powerpoint.

#### Modeling of Years 8-10, with year 11 as test, and Evaluation

Models run and results <table> and appendix from powerpoint.

Show plot and methodology and R^2

# Conclusion

Relate back to original problem statement

## Areas for Future focus

# Appendix

