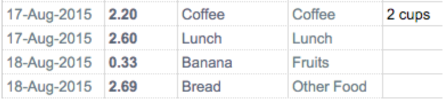
Kah-Vis : Visualizing the effect of Bulk Purchase of Coffee on consumption

Kunal Ghosh - 546247 - kunal.ghosh@aalto.fi

April 13, 2016

The Data

Our Chosen dataset comes from the purchase receipts we collected over the past 6 months. It is in the form of a table and a snippet of it is shown below. The dataset lists date of purchase, amount of money spent, item purchased, category and remarks. In this report we **focus only on the subset of the data which records the purchase of coffee**. We filtered the dataset and replaced the remarks field with the count of, cups of coffee. Note that:



17-Aug-2015	2.20	Coffee	Coffee	2 cups
17-Aug-2015	2.60	Lunch	Lunch	
18-Aug-2015	0.33	Banana	Fruits	
18-Aug-2015	2.69	Bread	Other Food	

Figure 1: Snippet of the Dataset.

- The number of cups for bulk purchases are taken from the of suggested servings on the pack.
- We had been informed apriori that this consumption pattern is only of a single person.

While performing exploratory analysis on the dataset, we observed that the amount of money being spent on coffee has gone down over the months.2

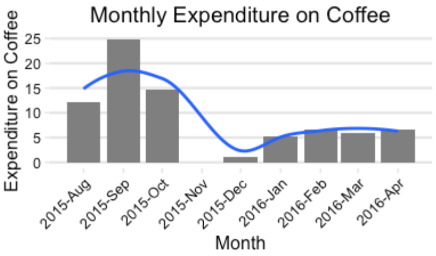


Figure 2: Monthly spend on Coffee

Although, at first sight this might appear to be a good sign, however this downward trend is because of subject started purchasing coffee in bulk and we **hypothesise that, bulk purchase of coffee has resulted in its increased consumption**.

Our visualization approach

Who is the Target audience of this visualization ?: This visualization is to help show the trend in coffee consumption to the person whose purchase dataset we are using. From the Visualization, the target audience might be interested to know about the number of Cups of Coffee consumed, the corresponding expenditure in Euros and overall trends, if they exist. To that effect we have aggregated, by month, the Expenditure on Coffee (in Euros) and the number of Cups of coffee purchased (or consumed). This grouping should also average out any daily or weekly fluctuations in the trend.

Munzner Model

The following Data 1 and Task 2 abstractions are based on the Munzer model described in the course slides.

Data Abstraction

- Data Type

  Attribute

  1. Date of Purchase
  2. Cups of Coffee Purchased
  3. Expenditure (Euros)

- Dataset Type : Table
- Dataset Availability : Static
- Attributes Types : Quantitative

Task Abstraction

Chained sequence of tasks or Action, Target Pairs.

- **PRODUCE : Derive - Distribution**
  - Group the Tabular data by Month.

- **QUERY : Identify - Trends**
  - Plot the trend lines

Visual Idiom

A visual idiom is a combination of marks and visual variables which are together called the visual encoding. We describe the visual encoding next.

Visual Encoding

Visual encodings used and their justification.

- Marks : Lines
  - Bars in the bar plot to encode **magnitude**
  - Interpolated lines which aid in the perception of overall trends 6
- Visual Variables
  - Position : **Horizontal Position of bars**. Aided by the month labels on the x axis, position of the bars can effectively portray ordinal data according to Mackinlay's ranking of visual variables 3
  - Size : **Length** of bars used to encode **magnitude** is the best available (Position already used to convey the ordinal month data) visual variable to encode Quantitative Data, according to **Mackinlay's ranking** 3.

Description of Visual Idioms used

- We have used **bar plots** to indicate the **magnitude** (of Cups of Coffee and Expenditure) as it aids in comparing values. 3
  - Also Bar plots are suitable because, the dataset has 1 key attribute (Date) and 1 quantitative attribute (either Expense or Cups of Coffee in their respective plots).
- We have also used a smoothed (cubic interpolated) **line plot** as it aids in perceiving **trends**. 4
  - In this case we used date as the ordered key attribute and the corresponding Expense and Cups of Coffee and the quantitative attribute for the two plots.

Tufte's Principles

- We have ensured that the bar plots are proportional to the measured quantities. This follows the principle laid down by Tufte which states that *The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.*

- We use only two visual variables to encode the dimensions of time and magnitude. Conforming with tufte's principle which states *The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.*

- **Data-ink ratio** is maximized by **removing the default gray background along with the grids** of the plots (made using ggplot in R). We have **added light gray horizontal guides** which helps in estimating the quantity represented by the barplots. Also the **facet bounding boxes are removed**. Although this might cause problems in distinguishing the two faceted plots, we have mitigated this problem by using the Gestalt principle of *connectedness* as described in the next section.

Gestalt Laws

- **Continuity** : The overall Trend is better perceived if the trend line is a smooth curve. 7 To that effect, we have smoothed the line plots using cubic interpolation.

- **Closure**: We have made use of facets to draw two graphs with comparable X axis. This is done because Cups of coffee consumed and Spend (in Euros) represent two separate quantities with different units and scales. Visualizing such quantities in the same plot can make the viewer draw false conclusions about the data, this has been discussed extensively in 10 and also in the "Planned Parenthood" example in lecture 2 9.

- **Connectedness** : Horizontal grid lines meet the corresponding facet labels indicating connectedness. This helps in perceiving the two sets of plots as two separate entities as it is a more powerful grouping principle than proximity.8

Methods used to analyze the data

- We first filtered the complete set of purchase records to get only the subset of data we were interested in analyzing.
- We then **order the data** by date to ensure that we get the true trend. **NOTE** : We could get "some" interesting trend and lie with the plots (by not showing the x axis labels) if the data was shown in a random order by date.
- We had also **grouped the data** by month, to get the monthly cumulative expense and cups of coffee consumed. This helped in averaging out small variabilities in the data and made it easier to spot the overall trends.
- Finally, instead of drawing a linear trend line we chose to use a **cubic polynomial** which we thought portrayed the overall trend well without being too jagged.

We mainly used the software environment **R** and its packages **lubridate** for date transformations, **sqldf** for data "juggling" and **ggplot2** for plotting and plot "touch ups", to increase the data-ink ratio.

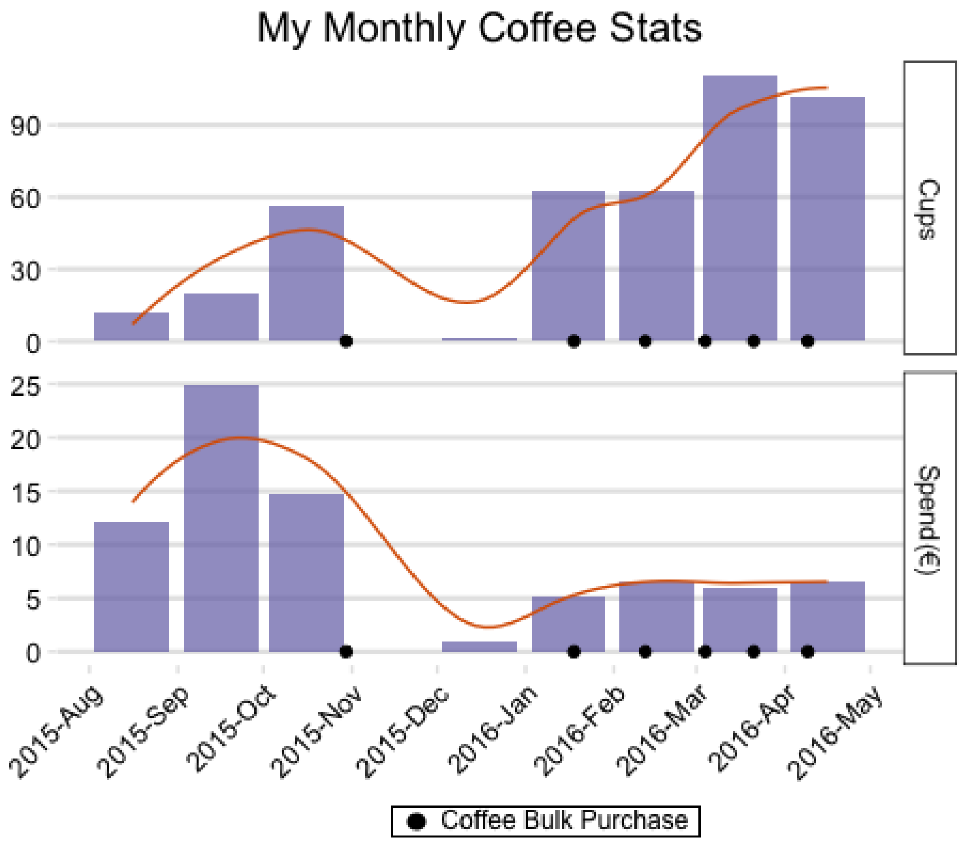


Figure 3: Note that the consumption is indicative only, if the coffee was purchased on the last day of the month, it would be consumed on the next month. However, it is grouped in the same month in the visualization. That explains why the number of cups don't increase in **November** after a bulk purchase in October.

Chosen Visualization Technique and Rationale

In this visualization, We have the challenge of visualizing two quantities (Monthly Expense and Cups of Coffee) with different units and scales. It is a known problem 10 that having two scales on the vertical axis leaves scope for manipulative visualizations. Which we have also previously discussed in the section on *Closure* above. We could have transformed the data to show only percentage change in the quantities as was done in 11 for the stock data for use in Index Charts. However in doing so, the plot would mislead the viewer into thinking the coffee consumption has gone down because, once the number of cups of coffee stays high but relatively constant the percent change would go down. To avoid the above caveat we make use of Facets to show the two quantities in separate graphs. **Facets** 5, as also discussed in the "How" to visualize section of **Munzner's Model**, allow us to juxtapose two graphs such that they have the same Horizontal axis which aids in making meaningful comparisons. Our choice of colors for the bars and line plot is from a color blind and print friendly color palette as suggested by the ColorBrewer2.org web-application.

Results

After analyzing the data and visualizing it 3, we observe that since the bulk purchase (lower coffee spends) have become more frequent, the consumption has risen steadily and shows a positive trend.

Discussion on Hypothesis after the results:

The graph quite strongly suggests that bulk purchase (Less monthly expenditure) of coffee has resulted in increased consumption. So we **can confirm our Hypothesis that the, bulk purchase of coffee has resulted in its increased consumption, based on the above visualization**. Given that the dataset records consumption pattern of only the subject. However it would advisable to run this experiment for a few more months and visualize the trends, to be certain this upward trend wasn't just a coincidence.

References

1. Lecture 7 slide 44 - T-61-5010 - Visualization analysis and design
2. Lecture 7 slide 52 - T-61-5010 - Visualization analysis and design
3. Lecture 7 slide 76 - T-61-5010 - Visualization analysis and design
4. Lecture 7 slide 79 - T-61-5010 - Visualization analysis and design
5. Lecture 7 slide 81 - T-61-5010 - Visualization analysis and design
6. Visualization Analysis and Design Page - (Munzner 2015) page 157
7. Lecture 6 slide 63 - Human perception part3
8. Lecture 6 slide 58 - Human perception part3
9. Lecture 2 T-61-5010 Visualization analysis and design Slide titled "Planned parenthood"
10. Stephen Few (2008) "Dual-Scaled Axes in Graphs Are They Ever the Best Solution?"
11. Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky (2010). "A Tour Through the Visualization Zoo." Communications of the ACM 53(6) : 59-67