# The Steaming Mug

Adya Sarin, CEO of 'The Steaming Mug', a hot-beverage chain, had just finished the weekly Monday morning meeting with all the VPs of the company, including the heads of sales, marketing, operations and finance. The meeting proceeded as usual with no major surprises or shocks. The high level numbers looked fairly good. The company seemed to be meeting their targets and as such nobody had any major issues to report. The meeting was dispersed with "continue business as usual". Yet, after the meeting, Adya was left with some discomfort. Were they being complacent? Were they doing enough to understanding their strengths and weaknesses? In a recently attended business expo she felt that the entire focus was on using data to make decisions and organizations that were data-driven. Adya wondered if "The Steaming Mug" was doing enough to leverage their data to take their business to the next level.

Consequently, she set up a meeting with the head of the analytics team Vir Narula, a trained statistician, to follow up on her intuition to explore their data better.

**Adya** : Good Morning Vir. I hope I didnot interrupt any important task.

**Vir**: Good morning Ma'm. I was compiling reports for the sales team. They wanted some numbers collated on the recent fall-campaign, but it's not very urgent.

**Adya**: Good. Listen, I know this is unusual, but I was wondering if we have had a good look at all the comprehensive data we have been collecting as part of our Data-Warehouse initiative. As you know, our Data-warehouse project was initiated almost two and a half years back and it has enabled us to collate detailed information on all our stores and products.

**Vir**: Hmm, well. Yes of-course. Our Data Warehouse project has been a great success in the quality and quantity of fine-grained information we have been able to collect. We do use some of that information to prepare our weekly and monthly reports, both for internal and external purposes.

**Adya**: That's good. However, these reports have standard formats that we have been using for a long time. They do not really allow us to dig deeper and look for any anomalies or exceptions at a fine-grained level. Do you think, your team would be able to explore this data further?

**Vir**: I guess so, but what exactly are you looking for?

**Adya**: This may sound crazy, but I don't know what we need to look for ! I just know that we need to study the data better.

Prepared by Professor Kavitha Ranganathan, Indian Institute of Management Ahmedabad.

Vir is now perplexed and is speechless. Adya realises that perhaps the task she is proposing is beyond the ability of the in-house analytics team and decides to bring in an external expert, Dr. Avital, a trained "Data Scientist" who could work with their analytics team and help them "look at the data better".

**About 'The Steaming Mug'**

The Steaming Mug is a US based hot beverage chain which started as a local café in San Jose, CA in 1985. Owning to the immense popularity of their specialty coffee, they soon started opening many additional branches and by 1989 had 12 stores spread across California and Washington. In time, they extended their core coffee offering to a range of coffee products. In 1999 they expanded their offerings to include various varieties of tea, which they continue to experiment with. The Steaming Mug currently has 156 stores spread across 20 states in the United States (exhibit 1) which offer 12 different types of brewed hot beverages, though not all products are offered at all stores.

As part of their data consolidation initiative, the company has collated monthly sales data across all their stores for the last two years. They record product-wise monthly sales, COGS , total expenses and inventory levels at each of their 156 stores across the country. Budgeted sales and COGS are also recorded (refer to Exhibit 2 for the detailed data model).

**The Potential of Exploratory Data Analysis**

Vir set up an initial meeting with Avital to tell him about the company and to jumpstart their new project with the aim of gaining new insights for their company.

**Avital**: Good morning Vir. I hear your team has been doing some good work around reporting and visualization.

**Vir**: Thanks! We have been creating a lot of summary reports and also doing some hypothesis testing when a particular department wants some questions answered. However, Adya's request has left us stumped as we do not know where to start and what to look for.

**Avital**: Well, there are actually three types of data that any organization has to deal with. Firstly, data that you know. This includes the kind of data available in the weekly or monthly reports that you prepare. The second type is data that you know you need to know. For example, if you see some exception in the weekly report, you want additional data on why that exception has occurred. The third type of data is data that you don't know you need to know. For example, if a particular relationship among data is not captured in the reports, unless there is ad-hoc and/or spontaneous exploration of the data, this new insight might be hidden forever.

When you get into a data-set without any pre-conceived notions and are open to discovering new insights, this particular type of analysis is called exploratory data analysis or EDA. There are of-course various techniques and tools one can use to get better at EDA, but some of it does come with practice.

**Vir**: Well, this sounds fascinating. It would help immensely to know a bit more about EDA. Could you elaborate on the techniques of EDA?

**Avital**: That could fill a book but I'll try to be brief. EDA as described by John Tukey, who proposed it, is a philosophy of "how to look at data to see what it seems to say". There are

two guiding principles to EDA: (1) to make the description of the data simpler so that one can understand the data better and (2) to look deeper into it than just at the existing level so as to gain additional insights.

**Vir**: Ah, sounds like we need a Sherlock Holmes here to look at the clues and solve the mystery!

**Avital** (smiling): That's exactly the description of the job. Tukey called it "graphical detective work", since EDA relies heavily on visualizing data to gain insights and spot trends, patterns or exceptions.

Let me give you an example. There is a famous data quartet called Anscombe's quartet (Figure 1) and all four pairs of values have the same summary statistics: the same mean, the same standard deviation, even their correlation co-efficient is identical. If we were looking for a relationship between x and y, no matter which data is examined the conclusion would be the same, a strong correlation with a correlation coefficient of 0.816

**Figure 1 : Anscombe's quartet**

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

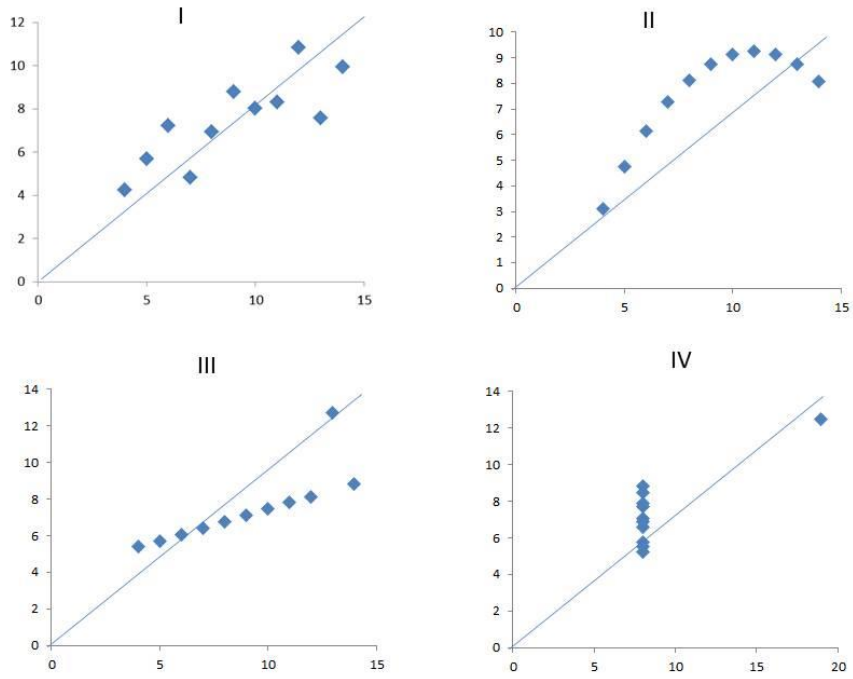
Now if we plot this data (

Without this kind of visual inspection, if we only looked at summary statistics, we might actually come to misleading conclusions about our data. There-in lies the strength of EDA.

Figure 2), we can see four very different scatterplots. The first exhibits a simple linearity between x and y and follows normality. The second exhibits a non-linear relationship. The third dataset has a strong positive correlation which is nearly 1 but its correlation coefficient is brought down because of one outlier. The fourth dataset contains no relationship between x and y, but for the one exception which produces a high correlation coefficient for the data.

Without this kind of visual inspection, if we only looked at summary statistics, we might actually come to misleading conclusions about our data. There-in lies the strength of EDA.

**Figure 2: Scatterplots representing Anscombe's quartet (tool : MS Excel)**




**Vir**: I am beginning to understand the significance of EDA now. Can you explain some more EDA techniques for analyzing data?

**Avital**: There are a whole lot of visual constructs (also called graphs or charts) that are recommended for EDA. Let me explain how to use some of them – the most useful ones. But before that, let us try to identify some common relationships in data that we normally explore. We came across one just now. This shows a correlation between two quantitative values. Can you think of some others – based on the reports you normally make or the questions the different departments ask you to answer.

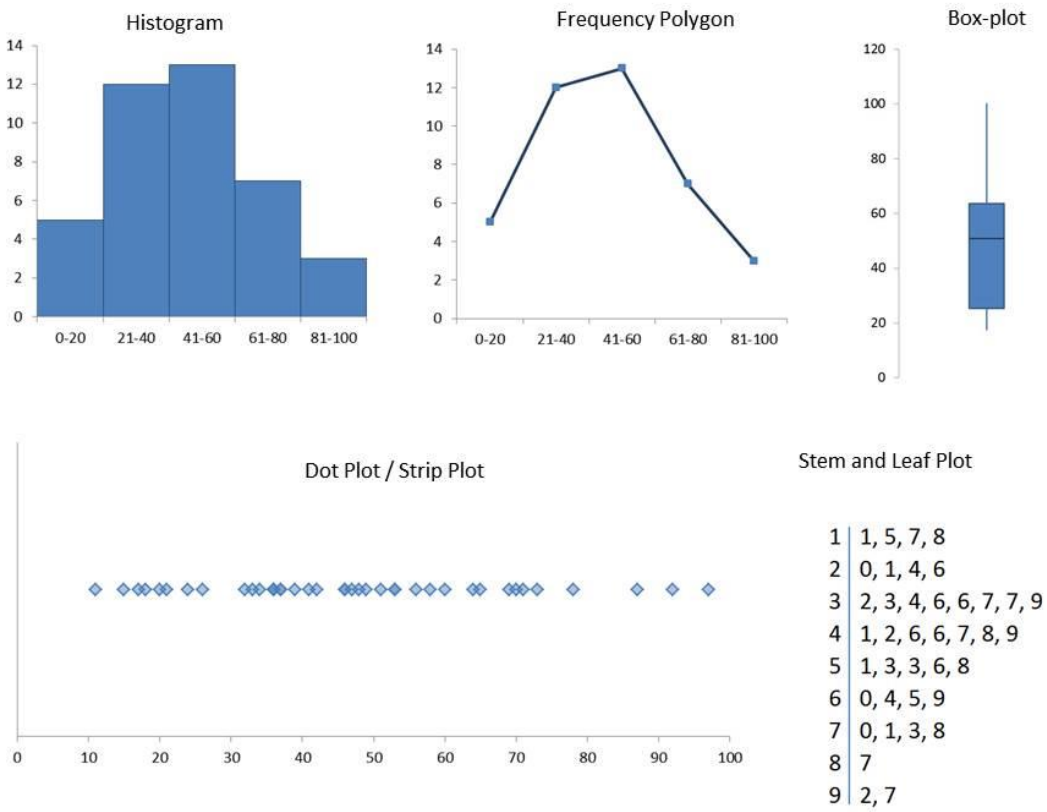
**Vir**: Let's see, apart from correlations which are of high interest to many people we also get a lot of requests for trends on a timeline. Understanding the frequency distribution of a particular variable is also popular.

**Avital**: Yes, both these are frequently used. So we have correlations, time-series or trends and distributions. Let me add multivariate analysis as a special case to this list.

You see, there are specific graphical constructs that are recommended for analyses depending on which relationship in the data you want to explore. For depicting **time series** it is a good idea to use trend lines and scatterplots are ideal for investigating possible correlations between two quantitative variables.

If you want to understand the **distribution** of a data-set there are a few different constructs that could be used, depending on the situation. Apart from the histogram that is the most well-known, we also have the frequency polygon, box-plot, strip plot and stem-and-leaf plot. The same dataset consisting of 40 numbers in the range 0-100 is represented by the five different constructs in Figure 3.

**Figure 3: Distribution constructs (tool : MS Excel)**




The histogram is most commonly used to find out how the values in a data-set are distributed within a range. Identifying the correct bin size for a particular histogram is a crucial design element in understanding your data and this is often a process of trial and error. Note that the bars in a histogram do not have any space between them. This is intentional – the continuity in the bars denotes the continuity across the bin ranges.

Both the histogram and the frequency polygon represent the data at a lower level of granularity as data is bunched together and represented. However, if you are more interested in the overall shape of the distribution rather than the individual bin size, then the frequency polygon works better than the histogram. Moreover, if you want to compare across multiple distributions, the frequency polygon comes in handy, with multiple lines of different colors to represent different data-sets in the same chart.

**Vir**: What if there are many distributions to compare? Won't all the lines overlap and make it difficult to interpret what is going on?

**Avital**: Yes, that is definitely an issue. The frequency polygon might be able to accommodate 4 or 5 different distributions after which it might start to look like a spaghetti bowl (laughs loudly).

**Vir** (smiling) : Is there any solution for that?

**Avital**: The box-plot which is also known as the box and whisker plot is the ideal solution for this situation. Invented by John Tukey, it condenses your data to 5 key points: the

highest, lowest, median, 25th and 75th percentiles. This gives you a quick sense of the distribution: effectively 4 bins for your data. But since it is so compact, a whole bunch of box-plots can be accommodated in a small amount of space and compared across. Box-plots are also handy if you want to use the x-axis in your graph for the time-line and analyses how a distribution has changed over time.

**Vir**: Box-plots sound pretty nifty! I have seen them a couple of times but never knew how to interpret them!

**Avital**: That is a good point to keep in mind if you plan to use box-plots to present your data. Your audience may find them daunting, so budget in a couple of minutes to explain how to read a box-plot to your audience.

**Vir**: So when are the other two constructs, the strip plot and the stem-and-leaf plot useful?

**Avital**: The strip plot uses one dot for each data point in the data-set. For example, if the data-set represented marks scored for a particular subject in a particular class, each dot represents a student. There is only one axis in this graph. Hence, a strip-plot is useful when you want to look at the data at a fine granularity. To compensate for over plotting (for example: multiple students getting the same marks), the dots are either made transparent or some amount of jitter is used to make the dots sit on top of each other vertically.

And finally, we have the stem and leaf plot, which is handy when you have a small data-set and want to look at the distribution within each bin as well. In the example shown here, the stem is represented by the tens of the number and the leaves by the units. You can see that five students are in the top thirties, just missing the passing cut-off of forty. Stem-and-leaf plots however, are usually used for back of the envelope calculations and are drawn by hand. This usually does not impose a problem since these are effective only if your data-set is small.

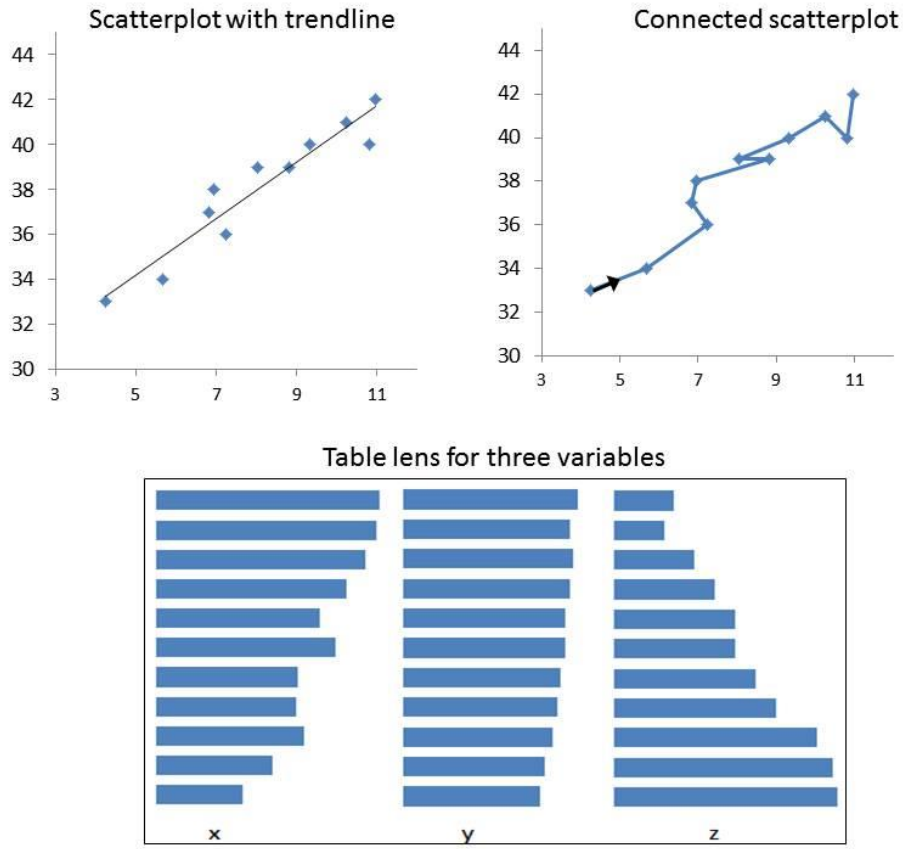
**Vir**: Is the scatter plot the only option available for investigating **correlations**?

**Avital**: Scatter plots are usually the default construct used to investigate a possible correlation. If you want to see how a particular correlation has played out over time, you can use what is called a connected scatter-plot. In a connected scatter-plot, the dots are joined by a line in chronological order. If the number of data points you have are relatively small you could also use what is called a table-lens, especially if you want to look at possible correlations between more than two variables. The various correlation constructs are depicted in

Figure 4.

Look at the example of the scatter plot. Let us assume that the x-axis denotes ice-cream sales and the y-axis denotes temperature in centigrade scale. Can you see any relationship between the two?

**Figure 4: Correlation constructs (tool : MS Excel)**




**Vir**: It looks like the ice-cream sales are greater when the temperature is higher. The rising heat causes more people to buy ice-creams.

**Avital**: Well, the scatter plot does depict a strong positive correlation between temperature and ice-cream sales. However, we need to take a step back in our analysis because we can't jump to the conclusion that the rise in temperature is causing ice-cream sales. Remember that a scatter-plot only confirms a correlation, and not causation. These could be a reverse-causation or a third hidden variable effecting both ice-cream sales and temperature or it could be a mere coincidence. Additional analysis is needed to rule out all these possibilities before we can assume a causal relationship between these two variables.

Now let us look at the table lens. Can you see any relationships here between x, y and z ? Let's assume these are monthly ice-cream sales, temperature and coffee sales respectively, sorted in descending order of ice-cream sales.

**Vir**: From a visual inspection, it looks like ice-cream sales and temperature are positively correlated to each other. Coffee sales seem negatively correlated to both ice-cream sales and temperature.

**Avital** (smiles): Wonderful, I noticed how you stayed away from any causal inferences this time. We've looked at time-series, distributions and correlations. Now, if you could show

me the data that you have as part of the data-warehousing effort I will try to work with that to illustrate the other constructs.

Vir shows Avital their data construct and explains the different fields in it and they decide to take it forward later.

Vir and Avital meet the next day to look over multivariate examples that Avital has created from The Steaming Mug's data.

**Avital**: We have already discussed the table lens for looking at possible correlations between multiple variables. As the number of variables we want to simultaneously analyze increase in number, the constructs get more complicated. Heat-maps and small-multiples are two effective options available for multivariate analysis.

Refer to the heat-map for product and state-wise sales, where the darker the color the higher the quantitative value. We can choose to retain the actual values (Figure 6) or remove them (Figure 5), depending on what you want to focus on. This heat map can be used to quickly identify top-selling products across states (Columbian coffee and Cinnamon tea for example) and strong markets across products ( California, Illinois, New York).

**Figure 5: Heat map of sales for different products across different states (tool : Tableau)**


**Figure 6: Heatmap (with values) of sales for different products across different states (tool : Tableau)**

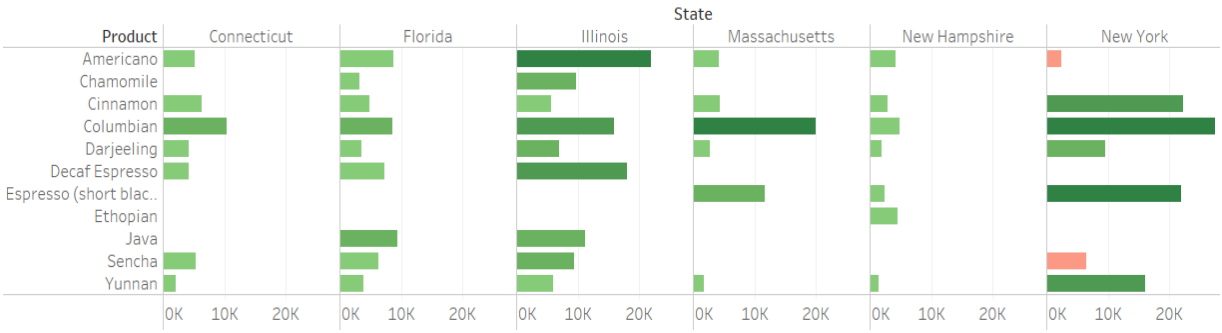
Product

| State | Americano | Caffe Latte | Chamomile | Cinnamon | Columbian | Darjeeling | Decaf Espr.. | Espresso (s.. | Ethopian | Java | Sencha | Yunnan |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| California | 5,465 | 8,424 | 4,374 | 8,948 | 12,819 | 4,619 | 10,268 | | 1,842 | 2,663 | 2,597 | 6,113 |
| Colorado | 3,321 | | 5,409 | 2,001 | 2,597 | 2,009 | 2,792 | | 4,619 | 4,374 | 2,145 | 4,377 |
| Connecti.. | 2,385 | | | 2,954 | 4,797 | 1,865 | 1,956 | | | | 2,490 | 1,073 |
| Florida | 4,193 | | 1,494 | 2,247 | 4,017 | 1,536 | 3,440 | | | 4,455 | 2,979 | 2,075 |
| Illinois | 10,268 | | 4,619 | 2,597 | 7,538 | 3,321 | 8,424 | | | 5,307 | 4,374 | 2,792 |
| Iowa | 891 | | 10,268 | 5,307 | 1,184 | 8,424 | 741 | | 849 | 707 | | 10,155 |
| Louisiana | 2,370 | 2,385 | 2,634 | 2,954 | 1,805 | | 1,905 | | | 2,075 | | |
| Massach.. | 1,940 | | | 2,009 | 9,345 | 1,226 | | 5,409 | | | | 933 |
| Missouri | 1,865 | | 1,986 | 2,226 | 2,954 | 1,562 | 1,574 | | | 2,490 | | 2,541 |
| Nevada | 849 | 741 | 5,465 | 7,538 | 1,383 | 10,268 | 891 | | | 1,184 | 5,307 | 8,873 |
| New Ha.. | 1,986 | | | 1,383 | 2,250 | 849 | | 1,158 | 2,135 | | | 741 |
| New Mex.. | 1,986 | 1,158 | 849 | 1,184 | 2,250 | | 1,562 | | | 2,226 | | |
| New York | 1,163 | | | 10,545 | 12,819 | 4,374 | | 10,350 | | | 2,882 | 7,940 |
| Ohio | 5,409 | | 933 | 1,226 | 1,797 | 4,017 | 2,009 | | 2,145 | 2,001 | | 4,455 |
| Oklahoma | 933 | 4,455 | 2,247 | 4,193 | 2,316 | | 4,017 | | | 1,226 | | |
| Oregon | 2,145 | 2,009 | 1,221 | 2,316 | 2,133 | 933 | 5,409 | | 2,001 | 1,797 | | 8,472 |
| Texas | 4,619 | 2,792 | 2,145 | 1,797 | 8,948 | | 3,321 | | | 2,597 | | |
| Utah | 2,634 | 1,574 | 2,135 | 2,250 | 1,956 | 1,986 | 1,865 | | 2,490 | 2,954 | 2,226 | 2,720 |
| Washing.. | 2,247 | 1,494 | 4,797 | 1,805 | 2,975 | 2,370 | 2,979 | | | 4,193 | | 4,290 |
| Wisconsin | 2,979 | | 2,370 | 2,075 | 4,193 | 1,905 | 1,494 | | 2,247 | 3,440 | | 2,385 |

Small multiples are also handy when we want to look at multiple dimensions. For example, in the example below ( Figure 7) we want to compare sales and profits across products and

are using the construct of small multiples. Small multiples allow us to juxtapose multiple graphs together, to make cross comparisons easy. However, to enable cross comparisons, the layout of the graphs – the range of axis's, category labels, color scheme used, labeling etc. should be uniform. For instance, this graph shows us that Americano is unusually strong in Illinois but is unprofitable in New York.

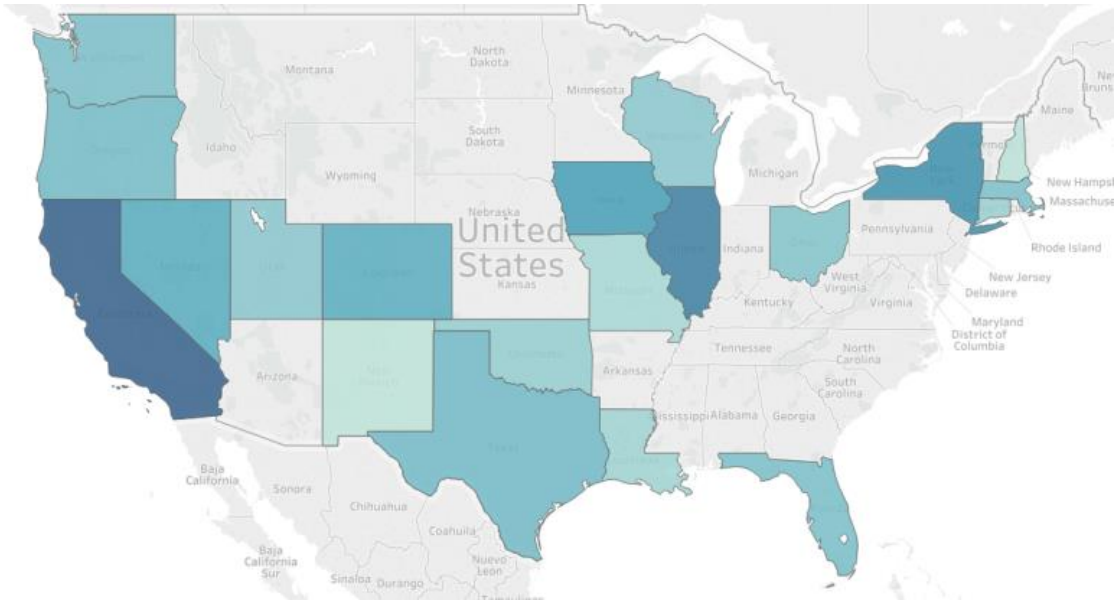
**Figure 7: Small multiples. Sales (bar) and Profits (color) across states and products ( tool : Tableau)**




Finally, constructs like chloropeth maps (Figure 8 ) and proportional symbol maps (Figure 9) can be used to investigate geographical or spatial relationships.

In the chloropeth map, entire areas get different colors. The color scale chosen typically represents a range of quantitative values, with darker intensities generally representing larger values. If the quantitative values are on a divergent scale (for example: profits, which can take on both positive and negative values), then the color scheme chosen should also be divergent, a common choice being increasing intensities of reds for negatives and increasing intensity of greens for positives.

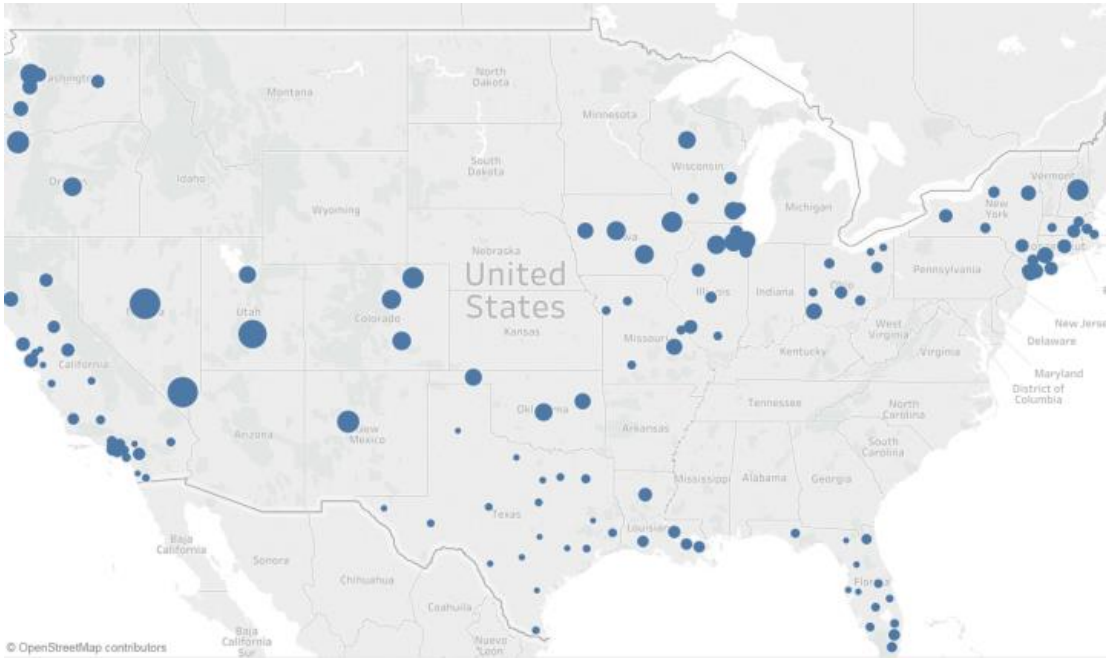
**Figure 8: Chloropeth map depicting profits across states (tool : Tableau)**




An alternative to using color to represent a quantitative value on a map is to change the size of the dot. This leads to a propotional symbol map, with the area of the dot propotional to the quantitative value represented. This construct would not serve variables which are on a

diverging scale, for obvious reasons. An effective combination often used in quantitative maps is to use the size of the dot to represent a sequential variable ( say revenue), and the color of the dot to represent a divergent variable (say profits).

**Figure 9: Proportional Symbol map depicting sales across stores (tool : Tableau)**




**Exploratory Analysis Tools**

**Vir:** Thanks Avital, now I have a handle on the various techniques that can been used for exploring the data. However, would you be able to recommend a tool for the same?

**Avital:** That is a good question but a tricky one to answer. The market is full of various visual BI tools that can be used for exploratory analysis but as usual there are tradeoffs to consider. A good place to start is Gartner's classification and evaluation of the various analysis tools each year (Exhibit 3). While Tableau, Microsoft BI and Qlik usually lead the pack in their reviews, some of these proprietary tools can be quite expensive. There are open-source alternatives like R with its ggplot libraries and Mondrian but their learning curve is steeper. With R for example, one has to be willing to learn the various commands. Of-course, those comfortable with programming might prefer the complete flexibility offered by languages like Python which come with a host of visualization libraries. Finally, if you feel you cannot invest either time or money in buying or learning a new tool, there is a lot you can do with MS Excel itself! Now that you know the basics of EDA, let us start our investigations. Are you ready Sherlock?

**Not All Those Who Wander Are Lost**

Vir's team has met and gone over some of the essential techniques of EDA. They decided that they would try to conduct investigations under four major heads: budget planning, marketing & sales, new markets and inventory management. Vir is supposed to present their findings and possible recommendations to Adya and the rest of the top management in two weeks.

Imagine you are in Vir's position.

Choose a topic from budget planning, marketing & sales, new markets or inventory management for your investigation.

Choose the BI tool you want to use. Sometimes multiple tools come in handy.

Conduct an Exploratory Data Analysis of the data-set available for The Steaming Mug.

Prepare a 10 minute presentation with your key-findings, which you will present to the top management of The Steaming Mug.

**Exhibit 1: Store locations of The Steaming Mug in the US (tool : Tableau)**

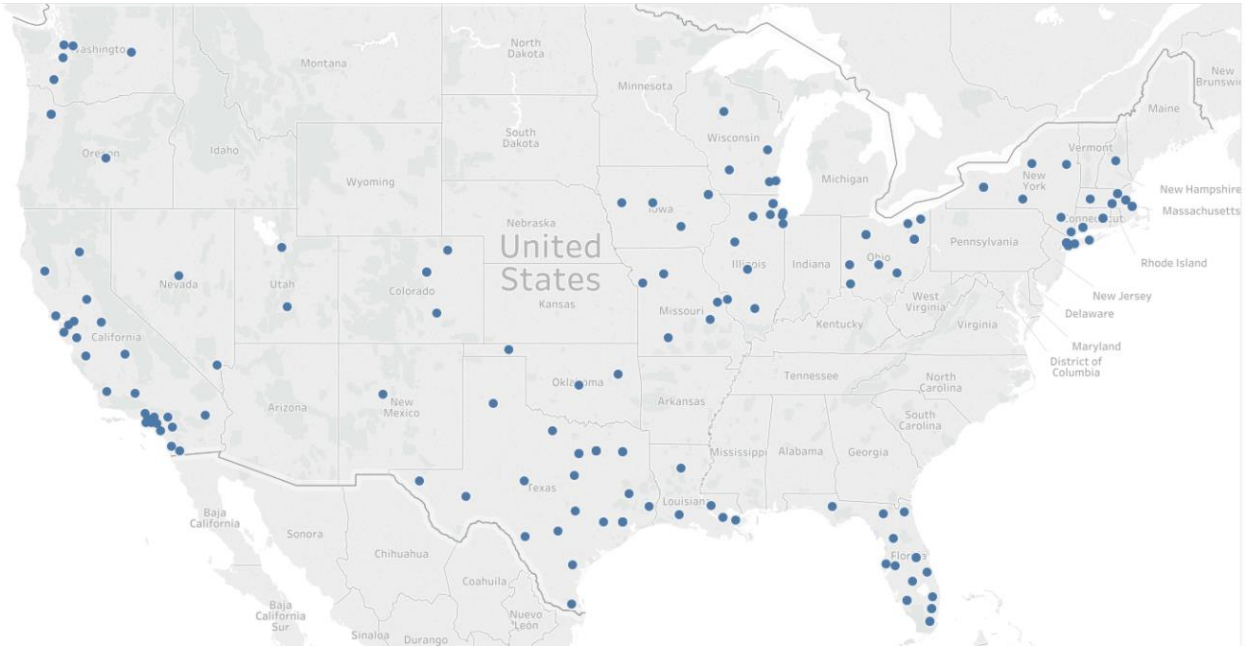



**Exhibit 2: Data Model for The Steaming Mug**

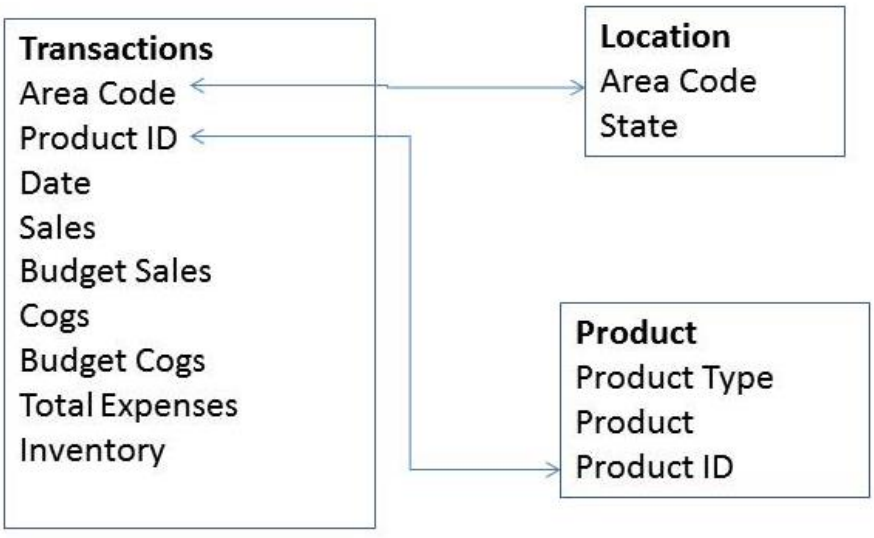


**Exhibit 3: Magic Quadrant for Analysis and Business Intelligence Platforms (source Gartner 2018)**