Visualizing Big Data: Everything Old is New Again

Belinda A. Chiera and Małgorzata W. Korolkiewicz

Abstract Recent advances have led to increasingly more data being available, leading to the advent of Big Data. The volume of Big Data runs into petabytes of information, offering the promise of valuable insight. Visualization is key to unlocking these insights, however repeating analytical behaviors reserved for smaller data sets runs the risk of ignoring latent relationships in the data, which is at odds with the motivation for collecting Big Data. In this chapter, we focus on commonly used tools (SAS, R, Python) in aid of Big Data visualization, to drive the formulation of meaningful research questions. We present a case study of the public scanner database Dominick's Finer Foods, containing approximately 98 million observations. Using graph semiotics, we focus on visualization for decision-making and explorative analyses. We then demonstrate how to use these visualizations to formulate elementary-, intermediate- and overall-level analytical questions from the database.

1 Introduction

Recent advances in technology have led to more data being available than ever before, from sources such as climate sensors, transaction records, cell phone GPS signals, social media posts, digital images and videos, just to name a few. This phenomenon is referred to as 'Big Data', allowing governments, organizations and researchers to know much more about their operations, thus leading to decisions that are increasingly based on data and analysis, rather than experience and intuition [16].

Belinda A. Chiera
University of South Australia, Australia, e-mail: belinda.chiera@unisa.edu.au

Małgorzata W. Korolkiewicz
University of South Australia, Australia e-mail: malgorzata.korolkiewicz@unisa.edu.au

Big Data is typically defined in terms of its *Variety*, *Velocity* and *Volume*. Variety refers to expanding the concept of data to include unstructured sources such as text, audio, video or click streams. Velocity is the speed at which data arrives and how frequently it changes. Volume is the size of the data, which for Big Data typically means 'large', given how easily terabytes and now petabytes of information are amassed in today's market place.

One of the most valuable means through which to make sense of Big Data is visualization. If done well, a visual representation can uncover features, trends or patterns with the potential to produce actionable analysis and provide deeper insight [21]. However Big Data brings new challenges to visualization due to its speed, size and diversity, forcing organizations and researchers alike to move beyond well-trodden visualization paths in order to derive meaningful insights from data. The techniques employed need not be new — graphs and charts can effectively be those decision makers are accustomed to seeing — but a new way to look at the data will typically be required.

Additional issues with data volume arise when current software architecture becomes unable to process huge amounts of data in a timely manner. Variety of Big Data brings further challenges due to unstructured data requiring new visualization techniques. In this chapter however, we limit our attention to visualization of 'large' structured data sets.

There are many visualization tools available; some come from established analytics software companies (e.g. Tableau, SAS or IBM), while many others have emerged as open source applications[1]. For the purposes of visualization in this chapter, we focus on SAS, R and Python, which together with Hadoop, are considered to be the key tools for Data Science [20].

The use of visualization as a tool for data exploration and/or decision-making is not a new phenomenon. Data visualization has long been an important component of data analysis, whether the intent is that of data exploration or as part of a model building exercise. However the challenges underlying the visualization of Big Data are still relatively new; often the choice to visualize is between simple graphics using a palette of colors to distinguish information or to present overly-complicated but aesthetically pleasing graphics, which may obfuscate and distort key relationships between variables.

Three fundamental tenets underlie data visualization: (1) visualization for data exploration, to highlight underlying patterns and relationships; (2) visualization for decision making; and (3) visualization for communication. Here we focus predominantly on the first two tenets. In the case of the former, previous work in the literature suggests a tendency to approach Big Data by repeating analytical behaviors typically reserved for smaller, purpose-built data sets (e.g. [9, 22, 10]). There appears, however, to be less emphasis on the exploration of Big Data itself to formulate questions that drive analysis.

[1] http://www.tableau.com
http://thenextweb.com/dd/2015/04/21/the-14-best-data-visualization-tools/
http://opensource.com/life/15/6/eight-open-source-data-visualization-tools

In this chapter we propose to redress this imbalance. While we will lend weight to the use of visualization of Big Data in support of good decision-making processes, our main theme will be on visualization as a key component to harnessing the scope and potential of Big Data to drive the formulation of meaningful research questions. In particular, we will draw upon the seminal work of Bertin [1] in the use of graph semiotics to depict multiple characteristics of data. We also explore the extension of this work [17] to apply these semiotics according to data type and perceived accuracy of data representation and thus perception. Using the publicly available scanner database Dominick's Finer Foods, containing approximately 98 million observations, we demonstrate the application of these graph semiotics [1, 17] for data visualization. We then demonstrate how to use these visualizations to formulate elementary-, intermediate- and overall-level analytical questions from the database, before presenting our conclusions.

2 Case Study Data: Dominick's Finer Foods

To illustrate Big Data visualization, we will present a case study using a publicly available scanner database from Dominick's Finer Foods[2] (DFF), a supermarket chain in Chicago. The database has been widely used in the literature, ranging from consumer demand studies through to price point and rigidity analysis, as well as consumer-preferences studies. The emphasis in the literature has been on using the data to build analytical models and drive decision-making processes based on empirically driven insights [9, 22, 10, 8, 5, 6, 19])[3].

The DFF database contains approximately nine years of store-level data with over 3,500 items, all with Unique Product Codes (UPC). Data is sampled weekly from September 1989 through to May 1997, totaling 400 weeks of scanner data and yielding approximately 98 million observations [15]. The sample is inconsistent in that there is missing data and data that is non-homogeneous in time, for a selection of supermarket products and characteristics. The full database is split across 60 files, each of which can be categorized broadly as either:

1. **General files:** files containing information on store traffic such as coupon usage and store-level population demographics (cf. Table 1); and
2. **Category-specific files:** grocery items are broadly categorized into one of 29 categories (*e.g. Analgesics, Bath Soap, Beer*, and so forth) and each item category is associated with a pair of files. The first of the pair contains product description information such as the name and size of the product and UPC, for all brands of that specific category. The second file contains movement information for each UPC, pertaining to weekly sales data including *store, item price, units sold, profit margin, total dollar sales* and *coupons redeemed* (cf. Table 2)

[2] http://edit.chicagobooth.edu/research/kilts/marketing-databases/dominicks/dataset

[3] An expanded list of literature analyzing the Dominick's Finer Foods Database can be found at https://research.chicagobooth.edu/kilts/marketing-databases/dominicks/papers

Table 1 Sample of customer information recorded in the DFF database. Coupon information appeared only for those products offering coupon specials

Variable	Description	Type
DATE	Date of observation	Date (yymmdd)
Week	Week number	Quantitative
Store	Unique Store ID	Quantitative
BAKCOUP	Bakery Coupons Redeemed	Quantitative
BAKERY	Bakery Sales in Dollars	Quantitative
BEER	Beer Sales in Dollars	Quantitative
...	...	...
WINE	Wine Sales in Dollars	Quantitative

Table 2 Sample of demographic information recorded in the DFF database. A total of 510 unique variables comprise the demographic data. This brief excerpt gives a generalized overview

Variable	Description	Type
NAME	Store name	Qualitative
CITY	City in which store is located	Qualitative
ZONE	Geographic zone of store	Quantitative
STORE	Unique Store ID	Quantitative
age60	Percentage of population over 60	Quantitative
hizeavg	Average household size	Quantitative
...	...	...

In total, across the general and category-specific files there are 510 store-specific demographic variables and 29 item categories recorded as 60 separate data sets. A 524-page data manual and codebook accompanies the database. Given the amount of information recorded in the database, we are able to present a *breadth-and-depth* data overview. Specifically, we will demonstrate data visualization across a range of characteristics and/or demographic attributes of the supermarket product *beer*, to provide a summary of the breadth of the data set, as well as an in-depth analysis of beer products to demonstrate the ability of visualization to provide further insight into big databases.

3 Big Data: Pre-Processing and Management

Prior to visualization, the database needs to be checked for inconsistencies and, given the disparate nature of the recorded data, merged in a meaningful way for informative visualization. Unlike smaller databases however, any attempt to view the data in its raw state will be overwhelmed by the volume of information available, due to the prohibitive size of Big Data. What is ordinarily a rudimental step of any statisical analysis — checking data validity and cleaning — is now a difficult

exercise, fraught with multiple challenges. Thus alternative approaches need to be adopted to prepare the data for visualization.

The two main areas which need to be addressed at this stage are:

1. Data pre-processing; and
2. Data management.

Of the three software platforms considered here, the Python programming language provides tools which are both flexible and fast, to aid in both data pre-processing and manipulation, a process referred to as either *data munging* or *wrangling*.

Data munging encompasses the process of data manipulation from cleaning through to data aggregation and/or visualization. Key to the success of any data munging exercise is the flexibility provided by the software to manipulate data. To this end, Python contains the specialized *Pandas* library (PANel DAta Structures), which provides the *data frame* structure. A data frame allows for the creation of a data table, mirroring *e.g.* Tables 1 and 2 (page 3), in that variables of mixed type are stored within a single structure.

The advantage of using a Python data frame is that this structure allows for data cleaning, merging and/or concatenation, plus data summarization or aggregation, with a necessarily fast and simple implementation. It should be noted that R, and to an extent SAS, also offer a data frame structure which is equally easy to use and manipulate, however Python is preferred for Big Data as the Pandas data frame is implemented using a programming construct called *vectorization*, which allows for faster data processing over non-vectorized data frames [18].

3.1 Data Pre-Processing

We first addressed the general store files individually (Tables 1 and 2) to perform an initial investigation of the data. The two files were read into separate Python data frames, named *ccount* and *demo*, and were of dimension $324{,}133 \times 62$ and 108×510 respectively, with columns indicating unique variables and rows giving observations over those variables. A sample of each data frame was viewed to compare the database contents with the data manual, at which time it was determined that the store data was not perfectly mirrored in the manual. Any variable present in the database that did not appear in the manual was further investigated to resolve ambiguity around the information recorded, and if a resolution could not be achieved, the variable was removed from the analysis.

Rather than pre-process the two complete data frames, we elected to remove columns not suitable, or not of immediate interest, for visualization. Given an end goal was to merge disparate data frames to form a cohesive database for visualization, we identified common variables appearing in *ccount* and *demo* and used these variables for the merging procedures, as will be discussed in what follows. We removed missing values using Python's *drop.na()* function, which causes the listwise removal for any record containing at least one missing value. We opted for listwise

deletion since the validation of imputed data would be difficult due to inaccessibility of the original data records, and sample size was not of concern. Other operations performed included the removal of duplicate records and trailing whitespace characters in variable names, since statistical software could potentially treat these whitespaces as unique identifiers, and introduce erroneous qualitative categories.

In what follows, rather than attempt to read the database as a whole, category-specific files (cf. page 3) for a selection of products were processed and held in computer memory only as needed. All three software applications considered here — SAS, Python and R — are flexible and allow easy insertion of data, thus supporting the need to keep the data frames as small as possible, with efficient adaptation on-the-fly.

We focused on a single product item for visualization, in this case *Beer*, as given the scope of the data there was a plethora of information available for a single product and the data was more than sufficient for our purposes here. Each product was represented by two files, as was the case for Beer. The first captured information such as the Unique Product Code, name, size and item coding. The second file contained movement information indicating price, profit margins, sales amounts and codes, as well as identifying information such as the store ID and the week in which the data was recorded. We elected to use the movement data only, since: (1) the information contained therein was more suited to meaningful visualization; and (2) the information in the movement data overlapped with the *ccount* and *demo* data frames, namely the variables *Store* (a number signifying the store ID) and *week*, allowing for potential merging of the data frames. Finally, a map was made available on the DFF website, containing geographic information of each store (City, Zone, Zip code) as well as the ID and the *Price Tier* of each store, indicating the perceived socio-economic status of the area in which each store was located. In what follows, *Store*, *Zone* and *Price Tier* were retained for the analysis.

3.2 Data Management

As previously noted, the DFF database is comprised of 60 separate files. While the data in each file can be analyzed individually, there is also much appeal in meaningfully analyzing the data as a whole to obtain a big-picture overview across the stores and their associated demographics. However given the full database contains over 98 million observations, the practicalities of how to analyze the data as a whole becomes a matter of data manipulation and aggregation. The initial pre-processing described above is the first step towards achieving this goal, as it provides flexible data structures for manipulation, while the management process for creating and/or extracting data for visualization forms the second step.

An attraction of using Python and R is that both languages allow the manipulation of data frames in the same manner as a database. We continued to work in Python during the data management phase for reasons cited above, namely the fast implementation of Python structures for data manipulation although the functionality

discussed below applies to both Python and R. While not all database functionality is implemented, key operations made available include:

- **concat:** appends columns or rows from one data frame to another. There is no requirement for a common variable in the data frames.
- **merge:** combines data frames by using columns in each dataset that contain common variables.
- **groupby:** provide a means to easily generate data summaries over a specified characteristic.

The database-style operation *concat* concatenates two data frames by adding rows and/or columns, the latter occurring when data frames are merged and each structure contains no variables in common. Python will automatically insert missing values into the new data frame when a particular row/column combination has not been recorded to pad out the concatenated structure. Thus care needs to be taken when treating missing values in a concatenated data frame - a simple call to *drop.na()* can at times lead to an empty data frame. In this instance only those variables immediately of interest for visualization should be treated for missing data.

The *merge* operation joins two or more data frames on the basis of at least one common variable — called a *join key* — in the data frames [18]. For example, the data frames *ccount* and *demo* both contain the numerical variable *Store*, which captures each Store ID. Merging the *demo* and *ccount* data frames would yield an expanded data frame in which the observations for each store form a row in the data frame while the variables for each store form the columns.

There is some flexibility as to how to join data frames, namely *inner* and *outer* joins. An *inner* join will merge only those records which correspond to the same value of the join key in the data frame. For example, while *Store* appears in *ccount* and *demo*, not every unique store ID necessarily appears in both data frames. An inner join on these data frames would merge only those records for which the store ID appears in both *ccount* and *demo*. Other inner join operations include merging data by retaining all information in one data frame and extending it by adding data from the second data frame, based on common values of the join key. For example, the *demo* data frame can be extended by adding columns of variables from *ccount* that do not already appear in *demo*, for all store IDs common to both data frames. Python also offers a full join, in which a Cartesian combination of data frames is produced. Such structures can grow quite rapidly in size and given the prohibitive nature of Big Data, we opted to use an inner join to reduce computational overhead.

Data summarization can take place via the *groupby* functionality, on either the original or merged/contatenated data frames. For example, if it is of interest to compute the total product profits per store, a *groupby* operation will efficiently perform this aggregation and calculation, thereby producing a much smaller data structure which can then be visualized. The *groupby* operation also has the flexibility to group at multiple levels simultaneously. For example, it might be of interest to group by the socioeconomic status of the store location, and then for each store in each of the socioeconomic groups, compute store profits. Providing the data used to define the levels of aggregation can be treated as categorical, *groupby* can perform any of

these types of aggregation procedures in a single calculation. As *groupby* is defined over the Python data frame structure, this operation is performed quickly over large amounts of data.

It should be noted that SAS also provides database-style support for data manipulation through Structured Query Language (SQL), which is a widely-used language for retrieving and updating data in tables and/or views of those tables. PROC SQL is the SQL implementation within the SAS system. Prior to the availability of PROC SQL in Version 6.0 of the SAS System, DATA step logic and several utility procedures were the only tools available for creating, joining, sub-setting, transforming and sorting data. Both non-SQL base SAS techniques or PROC SQL can be utilized for the purposes of creating new data sets, accessing relational databases, sorting, joining, concatenating and match-merging data, as well as creating new and summarizing existing variables. The choice of approach — PROC SQL or DATA step — depends on the nature of the task at hand and could be also accomplished via the so-called Query Builder, one of the most powerful 'one stop shop' components of the SAS® Enterprise Guide user interface.

4 Big Data Visualization

The challenge of effective data visualization is not new. From as early as the 10^{th} century data visualization techniques have been recorded, many of which are still in use in the current day, including time series plots, bar charts and filled-area plots [23]. However in comparatively more recent years, the perception of effective data visualization as being not only a science, but also an art form, was reflected in the seminal work on graph semiotics [1], through to later adaptations in data visualization [4, 3, 17]. Further influential work on statistical data displays was explored in [23] with an emphasis on avoiding data distortion through visualization, to more recent approaches [14] in which a tabular guide of 100 effective data visualization displays, based on data type, has been presented.

4.1 Visualization semiotics

Data visualization embodies at least two distinct purposes: (1) to communicate information meaningfully; and (2) for the visualization to *"solve a problem"* [1]. It is defensible to suggest that 'solving a problem' in the current context is to answer and/or postulate questions about (big) data from visualization, as was the approach adopted in the originating work [1]. It is thus in the same spirit we adopt graphic semiotics and reference the fundamental data display principles, in the visualization that follows.

At the crux of the works on visualization and graphic semiotics are the retinal variables identified in [1]. These variables are manipulated to encode information

from data for effective communication via visualization (Table 3) with application to data type as indicated [13].

Table 3 Retinal variables for the effective communication of data visualization [1, 13]

Variable	Description	Best for Data Type
Position	Position of graphing symbol relative to axes	Quantitative, Qualitative
Size	Space occupied by graphing symbol	Quantitative, Qualitative
Color Value	Varied to depict weight/size of observation	Quantitative Differences
Texture	Fill pattern within the data symbol	Qualitative, Quantitative Differences
Color Hue	Graduated RGB color to highlight differences	Qualitative Differences
Orientation	Used to imply direction	Quantitative
Shape	Graphic symbol representing data	Quantitative

The usefulness of the retinal variables was experimentally verified in subsequent research [7]. The authors focused solely on the accurate perception of visualization of quantitative data and developed a ranking system indicating the accuracy with which these variables were perceived. The variables *Position* and *Size* were the most accurately understood in data visualizations, whilst *Shape* and *Color* were the least accurate, with area-based shapes somewhat more accurate than volume-based shapes [7]. This work was later extended to include qualitative data in the heavily cited research of [17], in which further distinction was made between the visualization of ordinal and nominal categorical variables and is an approach which we adopt here. The revised ordering, including an extended list of retinal variables and their associated accuracy, is depicted in Fig. 1. The extended list centers around the original retinal variables introduced in [1] – for example *Shape* was extended to consider area- and volume-based representations while *Color* was considered in terms of saturation and hue.

The retinal variables are typically related to the components of the data that are to be visualized. Even from the smaller set of the event retinal variables in Table 3, there is a large choice of possible graphical constructions, with the judicious selection of several retinal variables to highlight data characteristics being perceived as more effective than use of the full set [1].

A motivation for forming data visualizations is the construction and/or answering of questions about the data itself. It was suggested in [1] that any question about data can be defined firstly by its *type* and secondly by its *level*. In terms of question type, the suggestion is that there are at least as many *types* of questions as physical dimensions used to construct the graphic in the first place. However [1] derived these conclusions on the basis of temporal data only, while [12] demonstrated that for spatio-temporal data, the distinction between question types can be independently applied to both the temporal and spatial dimensions of the data.

Questions about the data can be defined at an *elementary-*, *intermediate-* or *overall-level* [1]. From Fig. 2 it can be seen that the answer to elementary-level questions results in a single item of the data (*e.g.* product sales on a given day), answers to intermediate-level questions typically involve at least several items (*e.g.*

	Quantitative	Ordinal	Nominal
More Accurate	Position	Position	Position
	Length	Density	Color Hue
	Angle	Color Saturation	Texture
	Slope	Color Hue	Connection
	Area	Texture	Containment
	Volume	Connection	Density
	Density	Containment	Color Saturation
	Color Saturation	Length	Shape
	Color Hue	Angle	Length
	Texture	Slope	Angle
	Connection	Area	Slope
	Containment	Volume	Area
Less Accurate	*Shape*	*Shape*	Volume

Fig. 1 Accuracy of the perception of the retinal variables by data type [17]. *Position* is the most accurate for all data types, whilst items in gray are not relevant to the specified data type

product sales over the past 3 days) while overall-level questions are answered in terms of the entire data set (*e.g.* what was the trend of the product sales over the entire period?). Questions combining spatio-temporal scales could be phrased as, *e.g.*, *What is the trend of sales for Store 2?*, which is elementary-level with regards to the spatial component, but an overall-level question with regards to the temporal component. We will further elucidate in what follows in which we present visualization of the DFF Database by question level as defined in [1] and indicate combinations between spatio- and temporal-scales as appropriate.

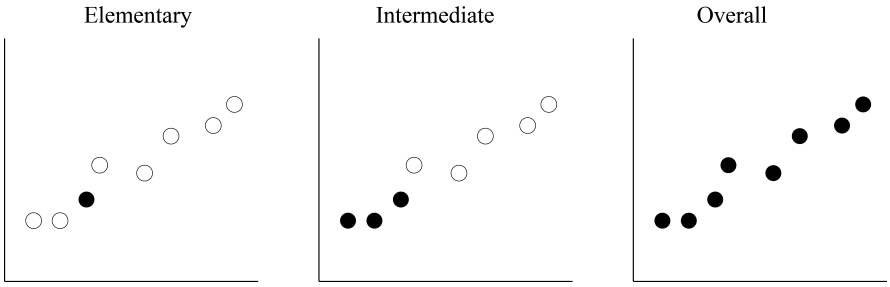

Fig. 2 Elementary-, Intermediate- and Overall-Level questions, based on data [1]. The colored circles indicates the number of data points involved in the answer to each question type

4.2 Visualization of the DFF Database

To achieve data visualization in practical terms, all three software systems considered here (SAS, Python, R) allow for graphical representation of data, however while Python is the preferred choice for data wrangling, we have selected R and SAS for data visualization, as they offer a comprehensive suite of graphical display options that are straightforward to implement. In contrast, data visualization in Python is typically not as straightforward and in practice it has been the case that, *e.g.* several lines of code in R are reproduced by over tens of lines of code in Python. Although Python generally boasts faster processing speeds due to the data frame structure, once data has been pre-processed as described in Section 3, the resulting data set is typically much smaller than the original, thus R and SAS are able to produce graphics with efficiency.

The introduction of Statistical Graphics (SG) Procedures and Graph Template Language (GTL) as part of the ODS Graphics system in SAS® 9.2 has been a great leap forward for data presentation using SAS. The SG procedures provide an easy to use, yet flexible syntax to create most commonly-used graphs for data analysis and presentation in many domains, with visualization options including the SGPLOT, SGSCATTER and SGPANEL procedures. In subsequent SAS releases more features were added that make it easy to customize graphs, including setting of group attributes, splitting axis values, jittering, *etc.* for SAS users of all skill levels. GTL allows the creation of complex multi-cell graphs using a structured syntax, and thus provides highly flexible ways to define graphs that are beyond the abilities of the SG procedures. Alternatively, SAS now offers SAS® Visual Analytics, which is an autocharting solution with in-memory processing for accelerated computations, aimed at business analysts and non-technical users. In this chapter, SG procedures and GTL were used in SAS® Enterprise Guide to generate selected visualizations.

To transfer data between Python and R for visualization, the options provided by Python include: (1) saving the Python data frame structures to file, which is then read into R in an identical data frame structure; and (2) direct communication with R from within Python via the *rpy2* interface. While the latter approach is more elegant and reduces computational overhead, the *rpy2* library is poorly supported across computing platforms and for this reason we have opted for the former approach. However it should be noted that as the merged data frames are the results of data wrangling and management rather than the raw data, these files are typically smaller and thus manageable for file input/output operations. It is also worth noting that R has the facility to read SAS files, and the latest releases of SAS include the facility to process R code, for added flexibility.

There are four primary graphical systems in R: *base, grid, lattice* and *ggplot2*, each of which offers different options for data visualization:

1. **base:** produces a basic range of graphics that are customizable;
2. **grid**: offers a lower-level alternative to the base graphics system to create arbitrary rectangular regions. Does not provide functions for producing statistical graphics or complete plots;

3. **lattice:** implements trellis graphics to provide high-level data visualization to highlight meaningful parts of the data. Useful for visualizing data that can be naturally grouped; and
4. **ggplot2:** creates graphics using a *layering* approach, in which elements of the graph such as points, shape, color *etc.* are layered on top of one another. Highly customizable and flexible.

Of the four graphic systems, only the library *ggplot2* needs to be explicitly installed in R, however this is readily achieved through the in-built installation manager in R with installation a once-for-all-time operation, excepting package updates. In this work we have predominantly used the *ggplot2* libraries to produce data visualizations and the *lattice* library where indicated. The lattice package allows for easy representation of time series data, while the layering approach of ggplot2 naturally corresponds with the retinal variables for visualization [1, 17].

Next we present a small selection of elementary-level questions to highlight the use of the retinal variables [1, 17], as these types of questions are the most straightforward to ask and resolve, even with regards to Big Data. The bulk of the visualization following elementary-level questions will be on the most challenging aspects with regards to visualization; intermediate- and overall-level questions.

4.2.1 Elementary-Level Question Visualizations

To produce an elementary-level question from a visualization, the focus on the data itself needs to be specific and quite narrow. In this regard, it is straightforward to produce one summary value of interest, per category.

For example, Fig. 3 shows a dot plot summarizing the total dollar sales of Beer in a selection of 30 stores in week 103 of the database. For this seemingly straightforward graphic, there are a number of retinal variables at play, including Position, Color and Length. We recall there are at least as many types of questions to be asked as physical dimensions used to construct the plot [1] and as the temporal quantity is fixed (week 103), we can adopt this approach over the spatial domain [12]. Thus sample questions could be: *What is the total dollar sales of beer in Store 103?* or *Which store has the maximum total dollar sales of beer?* The former is a very specific question requiring some level of knowledge of the database whereas the latter is purely exploratory in spirit and would be a typical question of interest.

In Fig. 4, a rainfall plot of the beer sales data is shown as a natural extension to the dot plot of Fig. 3, using the retinal variables Position and Length [2]. Color is now used for aesthetic purposes and not to highlight information, as was the case in Fig. 3. All stores have been included in Fig. 4 for all weeks of the data set, however the introduction of the qualitative variable Zone also increases the opportunity to pose informative, elementary questions, *e.g. Which zone has the largest variability in beer sales?* or *Which zone sells the largest volume of beer?*

On the other hand, Fig. 5 depicts a somewhat more intricate 100% stacked bar chart of the average beer price in each store over time, with the chart separated to show two types of categorical information: the year and the Price Tier of all stores.

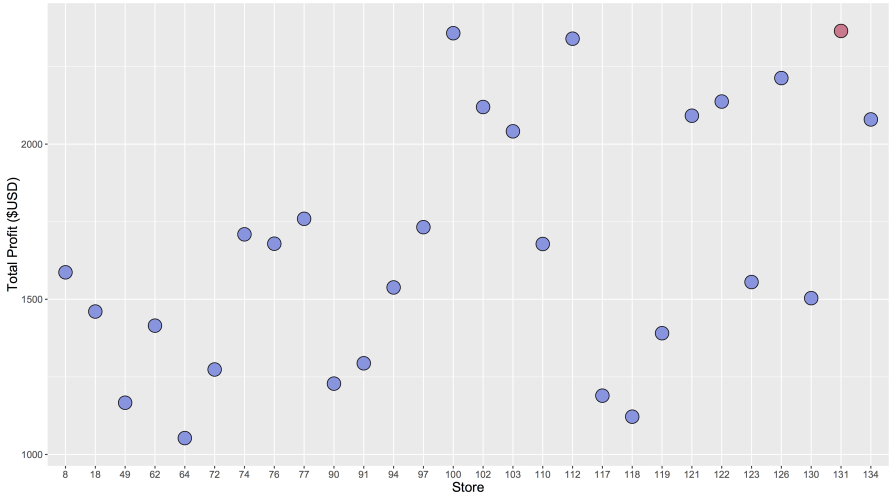

Fig. 3 Dot plot of Beer Sales by Store

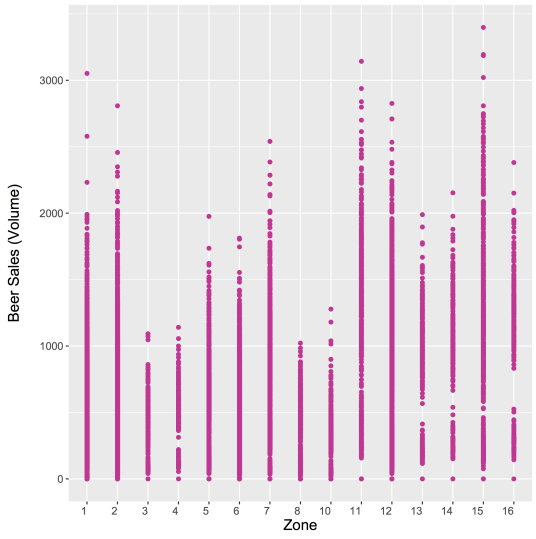

Fig. 4 Rainfall plot for elementary questions regarding Beer Sales by Zone

It is now also possible to observe missing data in the series, represented as vertical gray bars, with four weeks starting from the second week of September in 1994 and the last two weeks of February in 1995. In our exploration of the database we noted that the same gaps appear in records for other products, suggesting a systematic reason for reduced or no trading at Dominick's stores during those weeks.

The retinal variables used in Fig. 5 include Position, Color and Length. The graph varies over both the temporal and spatial dimensions and besides conveying information about a quantitative variable (average beer price), two types of qualitative

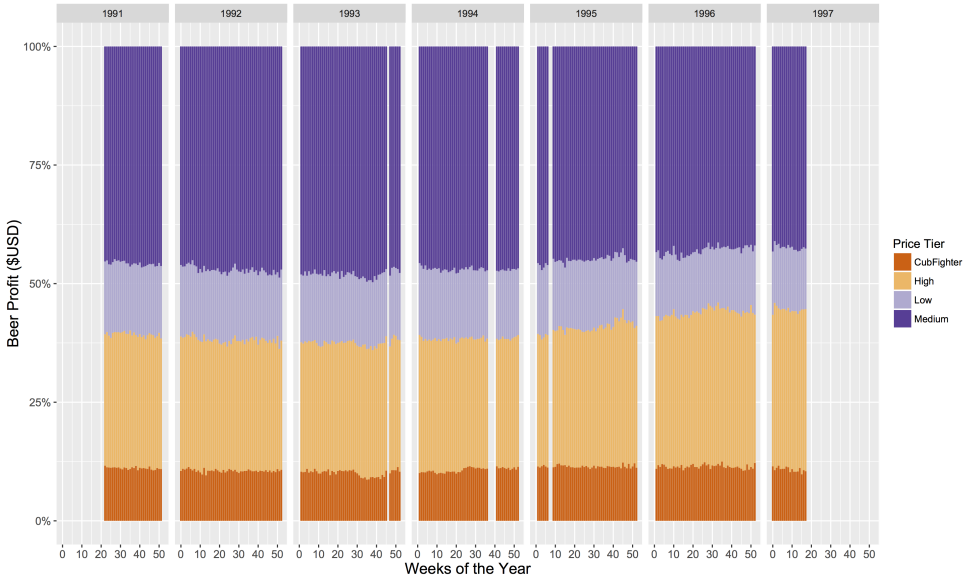

Fig. 5 A 100% Stacked Bar Chart of beer profit by Price Tier between 1991-1997. Missing values appear as vertical gray bars (*e.g.* near Week 40 in 1994)

variables (ordinal and nominal) are included as well. Thus questions can be formulated over either or both of these dimensions, for instance: *Which Price Tier sets the highest average beer price? In 1992, in which Price Tier do stores set the highest average price for beer?* and *In which year did stores in the "High" Price Tier set the lowest average price for beer?*

4.2.2 Intermediate-Level Question Visualizations

While elementary-level questions are useful to provide quick, focused insights, intermediate-level questions can be used to reveal a larger amount of detail about the database even when it is unclear at the outset what insights are being sought. Given data visualizations take advantage of graph semiotics to capture a considerable amount of information in a single graphic, it is reasonable to expect that it would be possible to extract similarly large amounts of information to help deepen an understanding of the database. This is particularly valuable as Big Data is prohibitive in size and viewing the database as a whole is not an option. While visualizations supporting intermediate-level questions do not capture the entire database, they do capture a considerable amount of pertinent information about the database.

Fig. 6 depicts total beer sales for all stores across geographic zones. This graphic resembles a box plot however is somewhat more nuanced, with the inclusion of 'notches', clearly indicating the location of the median value. Thus not only does

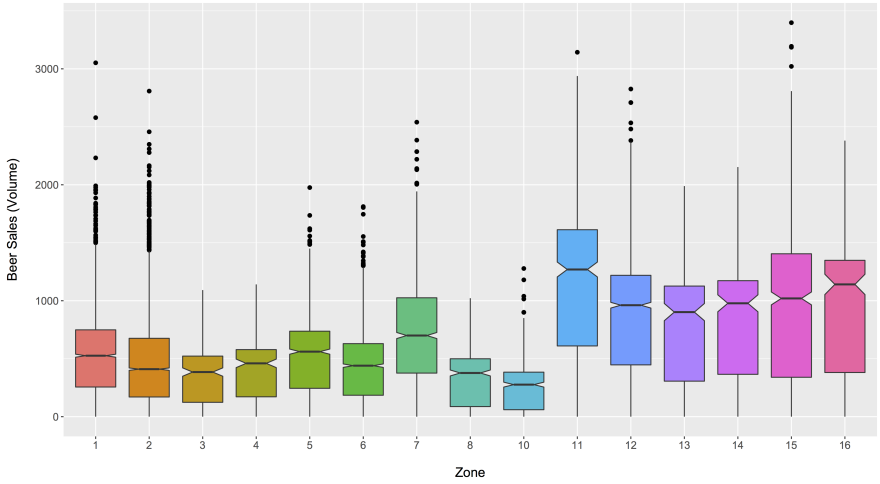

Fig. 6 Notched Boxplot for elementary questions regarding Beer Sales by Zone

Fig. 6 capture the raw data, it provides a second level of detail by including descriptive summary information, namely the median, the interquartile range and outliers. Much information is captured using the retinal variables Position, Color, Shape, Length and Size, meaning that more detailed questions taking advantage of the descriptive statistics can now be posed. For example, *Which zones experience the lowest and highest average beer sales, respectively? Which zones show the most and least variability in beer sales?* and *Which zones show unusually high or low beer sales?* Due to the shape of the notched boxes, comparisons between zones are enabled as well, *e.g. How do stores in Zones 12 and 15 compare in terms of beer sales?*

Adjusting this visualization slightly leads to the violin plot of Fig. 7, which is created for the same data as in Fig. 6. However while a similar set of retinal variables are used in this case, Fig. 7 captures complementary information to that captured by Fig. 6. In particular, the retinal variable *Shape* has been adjusted to reflect the distribution of the data, while notches still indicate the location of the average (median) quantity of beer sales. What is gained, however, in terms of data distribution, is lost in terms of detailed information about the stores themselves, namely the exact outliers captured in Fig. 6 versus the more generic outliers, depicted in Fig. 7.

Fig. 8 merges the best of both of the notched and violin plots to produce an RDI (Raw data/Description/Inference) plot, with Fig. 8(a) representing the same data as in Figs. 6 and 7. However, due to the breadth and depth of data representation, it is possible to easily capture other pertinent information about the database, including maximum beer sales, beer price and beer profit ((b)-(d), respectively), allowing easy comparison between multiple quantitative variables in the database, on the basis of identical qualitative information (Price Tier, Zone). Now more detailed intermediate-level questions can be posed, *e.g. How does beer price and profit com-*

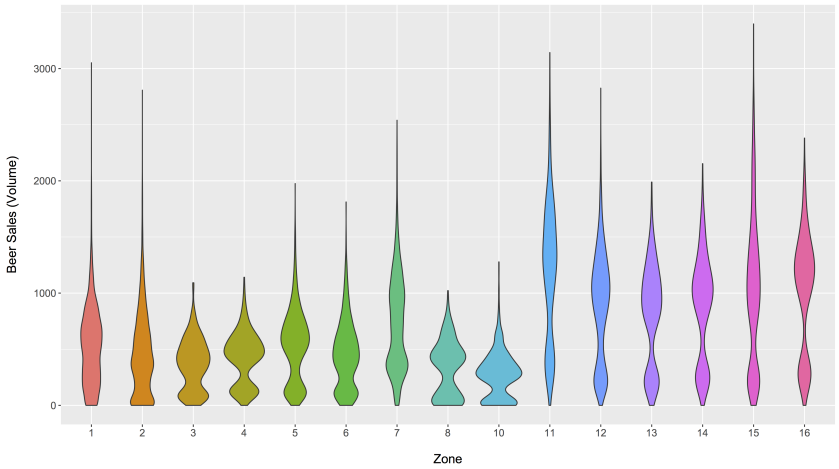

Fig. 7 Violin plot for elementary questions regarding Beer Sales by Zone

pare across stores in the Medium price tier? Generic questions can be now asked of the data as well, such as *How do beer prices in Low price tier stores compare with stores in other tiers?* or *Does any price tier consistently show the most variability across the beer variables Sales, Maximum Sales, Price and Profit?*

On a cautionary note, alternative RDI plots are shown in Fig. 9(a) and (b), in which it can be seen the retinal variable Density has been added to give an indication of the spread of the data points. However it is worth observing the construction of the vertical borders of the plot for each price tier. This is an example in which visual aesthetics have been promoted over data validity and observations that did not fall within the vertical bounds, were instead plotted on the vertical boundaries, resulting in the thicker walls of each bar. The true spread of the data is shown in Fig. 8 (b) which is highly unreadable. Thus while the raw data can be viewed and a descriptive statistic can be obtained from each price tier, interpretation of this type of plot is necessarily more restrained than for RDI plots such as those in Fig. 8.

Thus far, questions have been formulated about characteristics of single variables, however it is also often of interest to determine the association between two variables. Fig. 10 depicts a bubble plot, in which the retinal variables Position, Color, Shape, Density and Size are used to convey information about two quantitative variables (beer profit and price), summarized over a qualitative variable (Price Tier). In this case questions focused on the relationship between the quantitative variables can be asked, *e.g. Are beer prices and profits related? Is the relationship between beer price and profit consistent for lower and higher prices?* Or including the qualitative variable: *e.g. Are beer prices and profits related across price tiers? Which price tier(s) dominate the relationship between high prices and profits?*

A complementary focus for intermediate-level questions is the interplay between specific and general information. Fig. 11 depicts this juxtaposition via a bubble plot, however now focusing on two beer brands and their weekly sales across the

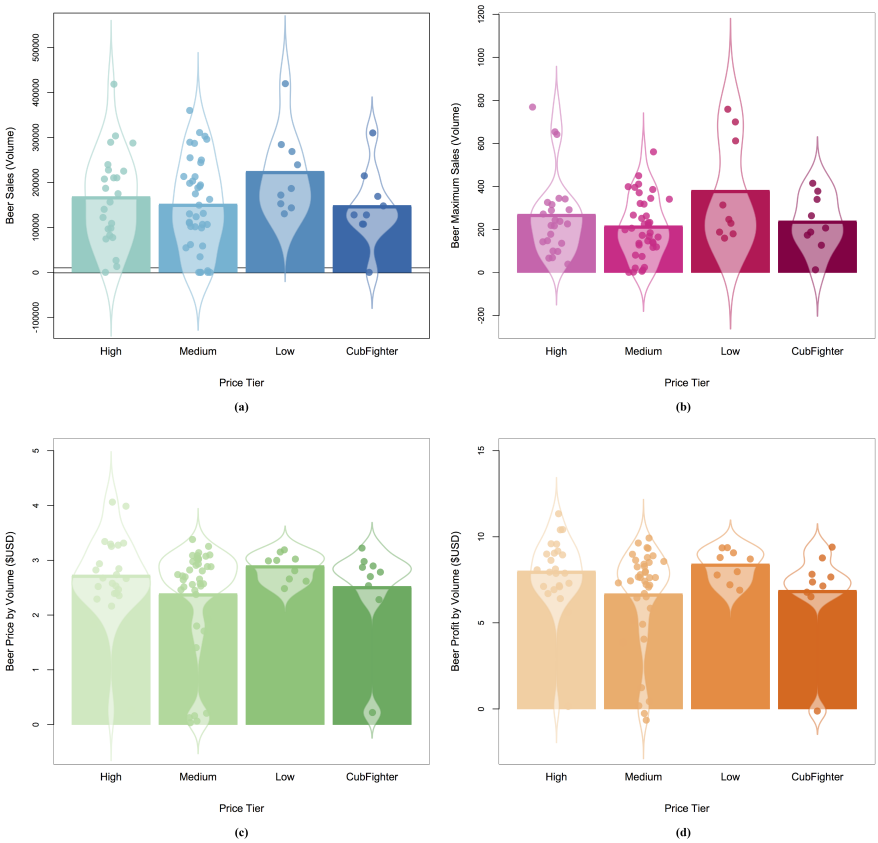

Fig. 8 RDI (Raw Data/Description/Inference) Plots of Beer Price, Profit and Movement

four price tiers, using the retinal variables Color, Shape, Size, Position and Length. Questions that can be posed include *e.g. Amongst which price tiers is Miller the most popular? Is Budweiser the most popular beer in any price tier?* Alternatively, the focus can instead be on the second qualitative variable (price tier) *e.g. Which is the most popular beer in the Low price tier?*

As a second cautionary tale, we note that there is no one graph type that is infallible, and often the selection of variables of interest will determine the usefulness of a visualization. For example, Figs. 12 and 13 both display the same data, namely beer sales over the four price tiers, for two popular brands. However while the box plot of Fig. 12 is an RDI plot in that it captures a great deal of detail about each variable, much of the data features are obscured by the differences in scale of sales for the two selected beer brands. On the other hand, the butterfly plot (Fig. 13) offers a variation on a simple bar chart by utilizing an extra retinal variable — Orientation — and in doing so provides a more informative comparison of the average sales levels.

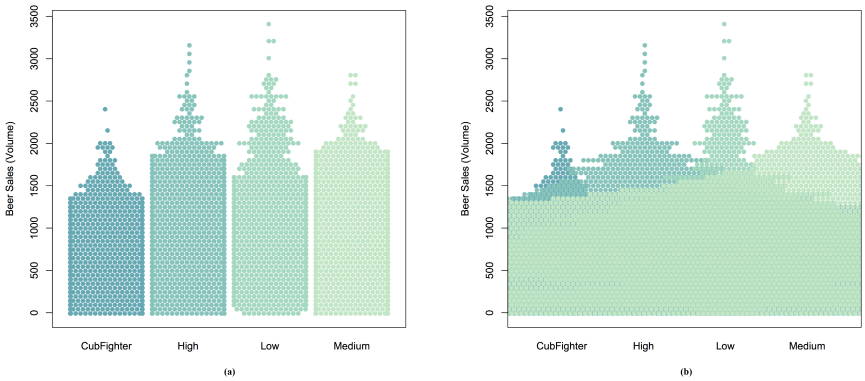

Fig. 9 A swarm plot of beer sales by price tier. The plot in (a) has been altered for visually aesthetic purposes, distorting the true nature of the data while the plot in (b) shows the true data, however is not easily readable

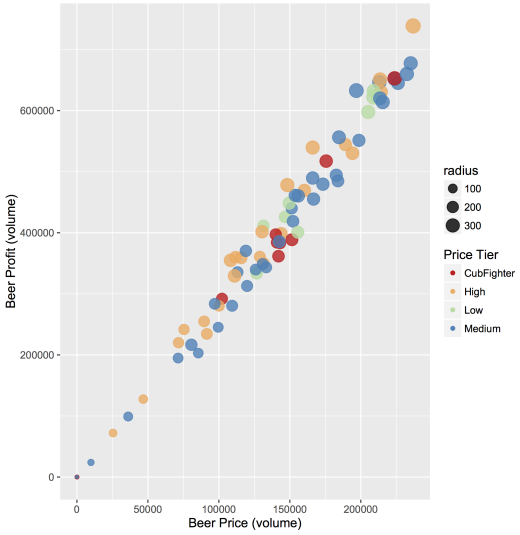

Fig. 10 Bubble Plot of Beer Price vs Profit, relative to product sales

A useful extension to intermediate-level questions comes from using the retinal variables Color Hue and Orientation, to produce a ternary plot [2]. In Fig. 14 the color palette used to represent Hue is shown at the base of a ternary heat map, with the interplay between the three quantitative variables Beer movement (sales), price and profit being captured using shades from this palette.

Intermediate-level questions can now focus on the combination of quantitative variables and can either be specific *e.g. For beer sold in the top 20% of prices, what percentage profit and movement (sale) are they experiencing?* or generic, so as to

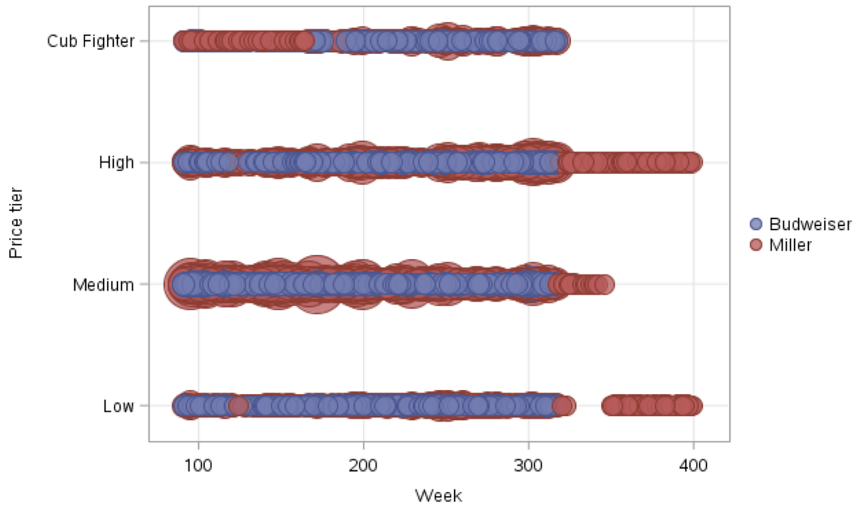

Fig. 11 Bubble plot of beer sales of two popular brands over four store price tiers

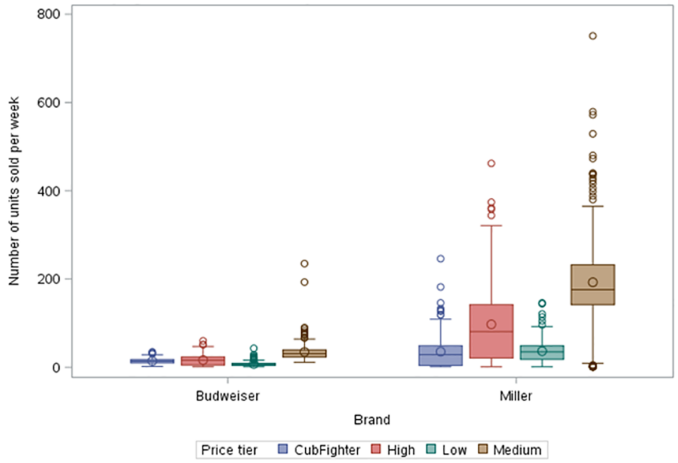

Fig. 12 A box plot of beer sales of two popular brands, over the four price tiers

uncover information, *e.g. do stores that sell beer at high prices make a high profit? Is it worth selling beer at low prices?* The power of question formulation based solely on the judicious selection of retinal variables, makes extracting insights from Big Data less formidable than original appearances may suggest, even when combined with standard graphical representations.

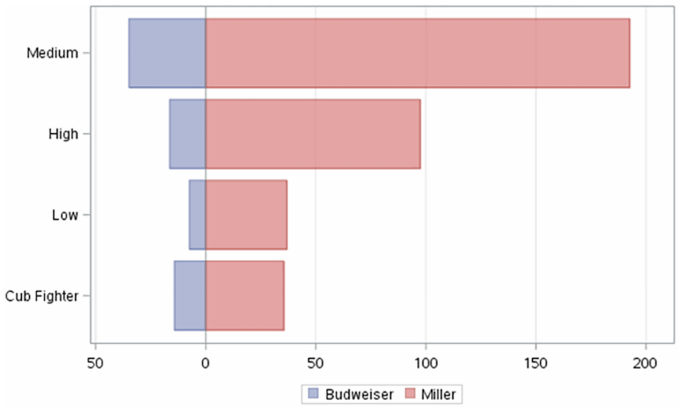

Fig. 13 A butterfly of beer sales of two popular brands, over the four price tiers

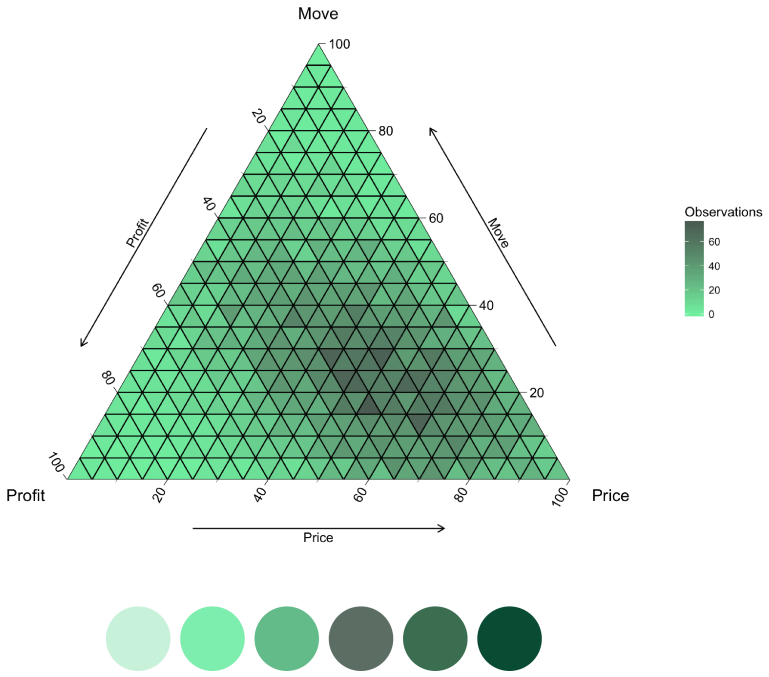

Fig. 14 Ternary plot for intermediate questions about beer over three quantitative variables: Price, Profit and Move (Sales). The color hue palette is shown at the base of the plot and is reflected in the plot and legend

4.2.3 Overall-Level Question Visualizations

Questions at the overall level focus on producing responses that cover the data as a whole, with an emphasis on general trends [1]. Time series plots are useful in

this regard, however traditional time series plots which plots all data points over the entire data collection usually renders very little information and is often difficult to read, unless specialized visualisations of the time series are shown, such as an inset (Fig. 15). Although it is possible to glean very general trends in this case, there is still an opportunity loss in that retinal variables are not being properly utilized to convey more subtle information.

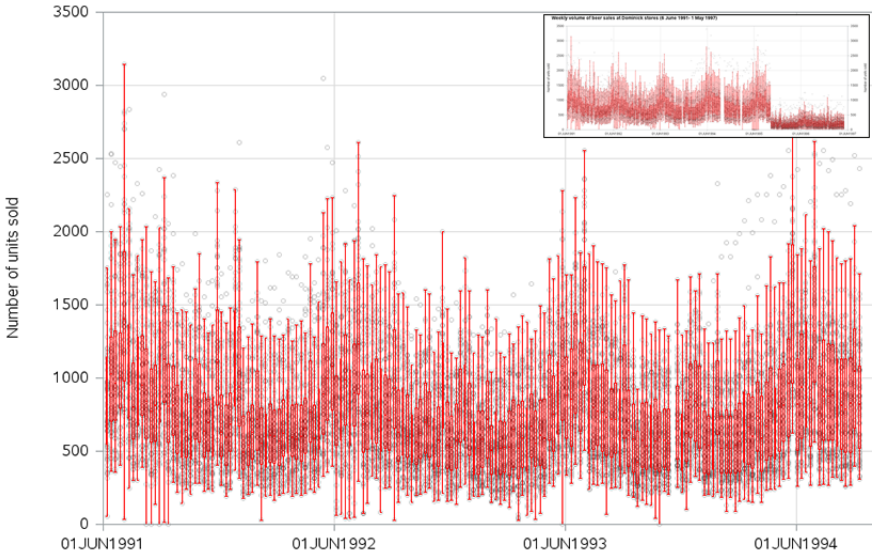

Fig. 15 Time series plot of the distribution of weekly beer sales in each store

Horizon plots are a variation on the standard time series plot by deconstructing a data set into smaller subsets of time series that are visualized by stacking each series one upon another. Fig. 16 demonstrates this principle using beer profit data for all stores between the years 1992-1996. In this case the shorter time interval was chosen to present easily comparable series that each span a single year. Even so, similar questions to those posed for Fig. 15 apply in this case. The attraction of the horizon plot lies however, in the quick comparison between years and months which is not facilitated by the layout of Fig. 15.

Interpretation of the beer data is improved again when the overall time series is treated as an RDI plot (Fig. 17) and information about each store can be clearly ascertained over the entire time period, with the focus now on spatial patterns in the data, rather than temporal.

In contrast, the same information is depicted in Fig. 18 however through the addition of the retinal variable Colour Hue, the data presentation allows for easier interpretation and insight. In this case questions such as *When were beer profits at a low and when were they at a high? What is the general trend of beer profits between 1991 and 1997?* can be asked. Similar questions can also be asked of the data in

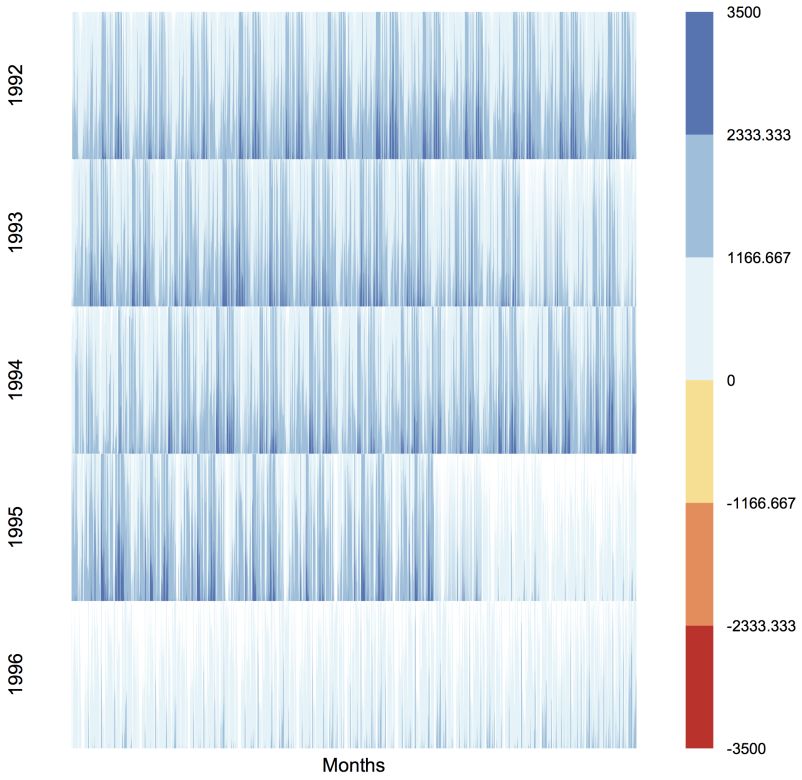

Fig. 16 Horizon plot of Beer Profit over all stores each week between 1992-1996

the time series RDI plot (Fig. 17) however the answer will necessarily involve the stores, providing alternative insight to the same question.

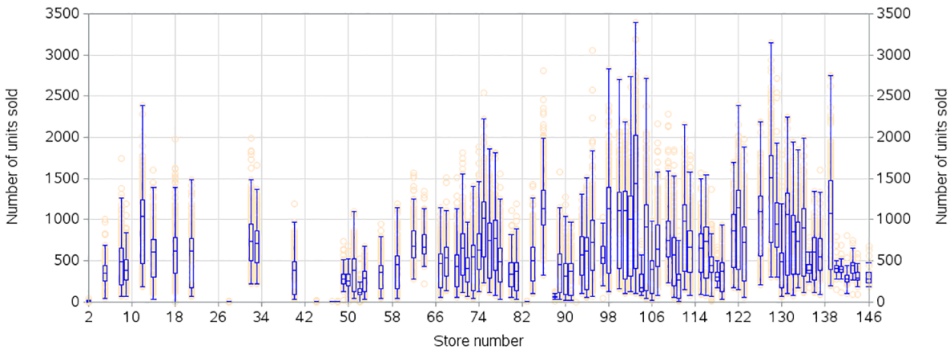

Fig. 17 Time series RDI plot of weekly beer sales in each store

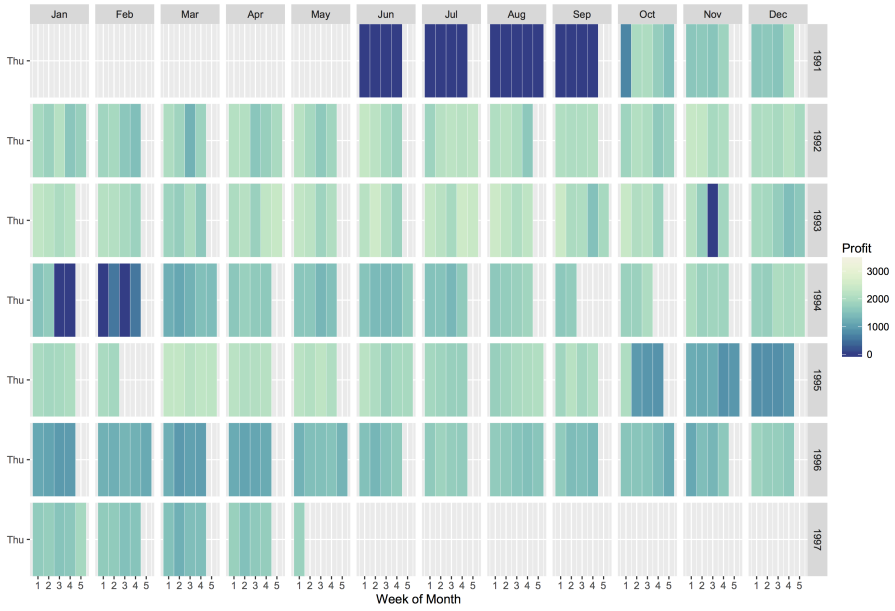

Fig. 18 Calendar Heat Map of Beer Profit over all stores in each week.

An alternative representation of overall trends in the data come from a treemap visualization, which aims to reflect any inherent hierarchy in the data. Treemaps are flexible as they can be used to not only capture time series data, but also separate data by a qualitative variable of interest, relying on the retinal variables Size and Colour Hue to indicate differences between and within each grouping. The advantage of such a display is the easy intake of general patterns that would otherwise be obfuscated by data volume. For this reason, treemaps are often used to visualize stock market behavior [11].

To create a treemap two qualitative and two quantitative variables are required. The item of interest (qualitative) is used to form the individual rectangles or 'tiles', while the group to which the item belongs (qualitative) is used to create separate areas in the map. A quantitative variable to scale the size of each rectangle is required and a second quantitative variable assigns the depth of color to each tile. Figure 19 displays a treemap of the beer data using the variables Store, Price Tier, Beer Price and Beer Profit. Each store corresponds to a single tile in the map while Price Tier is used to divide the map into four separate areas. The size of each tile corresponds to the price of beer at a given store, while the color hue represents the profit made by each store, with the minimum and maximum values indicated by the heat map legend. Questions postulated from a treemap include *Which stores are generating high profits? What is the relationship between beer price and profit? Which price tiers make the largest profit by selling beer? Do stores within a price tier set the price of beer consistently against one another?*

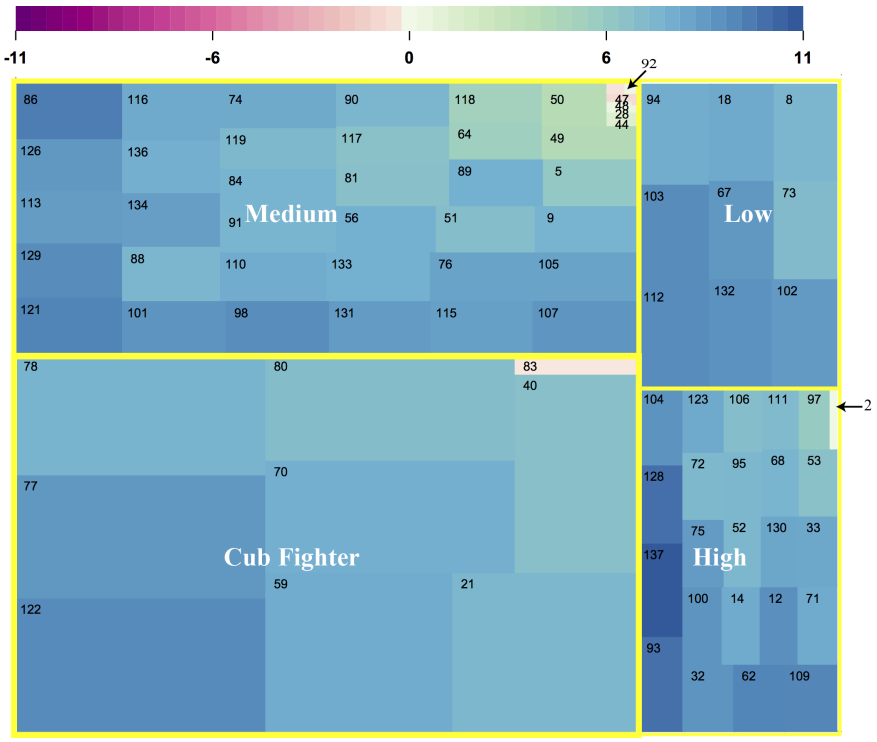

Fig. 19 A treemap showing the relationship between beer price and profit across price tiers and at each individual store

The last graph presented here is another variation of a time series plot, however with the modification of added functionality to depict the behavior of multiple groups simultaneously. Fig. 20 shows a streamgraph of beer profit made in every single store over the entire time period represented in the data set. In this case the retinal variables Length, Orientation and Color Hue are being used to combine quantitative information (beer profit) with qualitative groups (stores) to give an overall view of the general trend. The streamgraph in R however has an added feature that is a modernization of the retinal variable Color Saturation. The R streamgraph is interactive, with a drop-down menu to select a particular store of interest (labeled *Ticker* in Fig. 20). The streamgraph is also sensitive to cursor movements running over the graph, and will indicate in real time over which store the cursor is hovering. Fig. 21 demonstrates the selection of Store 103 and how the modernization of an 'old' technique further enhances the types of insights that can be drawn from this graph.

Thus questions that can be asked about the data based on streamgraphs include *What is the overall trend of beer prices? Does beer price behaviour change over time? Are there repetitive patterns to beer price behaviour?* Then, coupled with Color Saturation to select a store of interest, *What is the overall trend of beer prices*

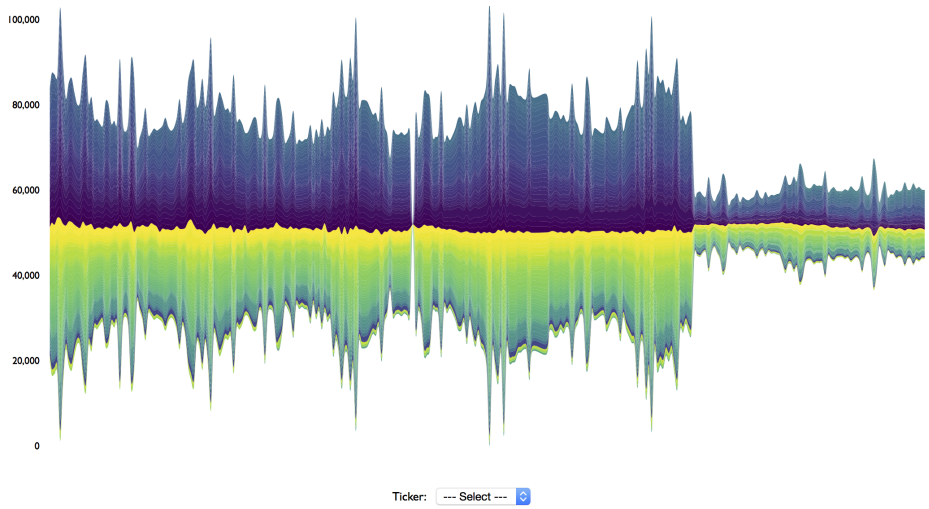

Fig. 20 Streamgraph of Beer Profit over all stores in each week.

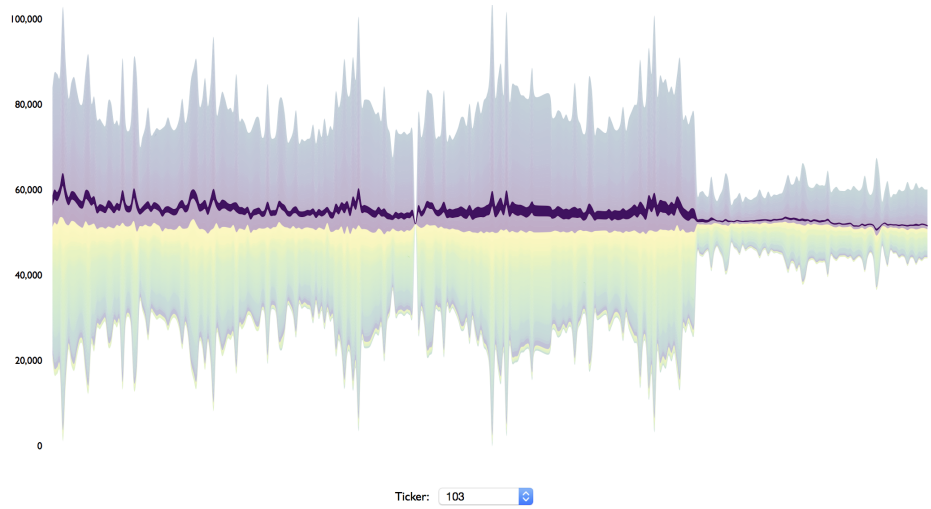

Fig. 21 Streamgraph of Beer Profit for store 103 in each week.

at a particular store? Does the trend of this store behave similarly to the overall pattern? and so forth, allowing for overall-level questions that compare specific item behavior (e.g. a store) with the overall trend in the data.

5 Conclusions

Visualization of data is not a new topic — for centuries there has been a need to summarize information graphically for succinct and informative presentation. However recent advances have challenged the concept of data visualization, through the collection of 'Big Data', that is data characterized by its variety, velocity and volume and is typically stored within databases that run to petabytes in size.

In this chapter, we postulated that while there have been advances in data collection, it is not necessarily the case that entirely new methods of visualization are required to cope. Rather, we suggested that tried-and-tested visualization techniques can be adopted for the representation of Big Data, with a focus on visualization as a key component to drive the formulation of meaningful research questions.

We discussed the use of three popular software platforms for data processing and visualization, namely SAS, R and Python and how they can be used to manage and manipulate data. We then presented the seminal work of [1] in the use of graph semiotics to depict multiple characteristics of data. In particular, we focused on a set of retinal variables that can be used to represent and perceive information captured by visualization, which we complemented with a discussion of the three types of questions that can be formulated from such graphics, namely elementary-, intermediate- and overall-level questions.

We demonstrated application of these techniques using a case study based on Dominick's Finer Foods, a scanner database containing approximately 98 million observations across 60 relational files. From this database, we demonstrated the derivation of insights from Big Data, using commonly known visualizations and also presented cautionary tales as a means to navigate graphic representation of large data structures. Finally, we also showcased modern graphics designed for Big Data, however with foundations still traceable to the retinal variables of [1], in support of the view that in terms of data visualization, everything old is new again.

References

1. Bertin, J. (1967). Semiology of Graphics: Diagrams, Networks, Maps. Madison, Wisconsin: The University of Wisconsin Press, 712 pages.
2. Breckon, C.J. (1975). Presenting Statistical Diagrams. Pitman Australia (Carlton, Victoria), 232 pages.
3. Card, S.K., Mackinlay, J. D., Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
4. Card, S. (2009). Information Visualisation. Sears, A., Jacko, J. A. (eds). Human-computer interaction handbook, pp. 181-215. CRC Press, Boca Raton.
5. Chen, Y., Yang, S. (2007). Estimating Disaggregate Models Using Aggregate Data Through Augmentation of Individual Choice. *Journal Marketing Research*, 44(4), 613-621.
6. Chintagunta, P. K., Vishal, J-P. D. S. (2003). Balancing Profitability and Customer Welfare in a Supermarket Chain. *Quantitative Marketing and Economics*, 1, 111-147.
7. Cleveland, W. S., McGill, R. (1984). Graphical perception: Theory, experimentation and application to the development of graphical methods. *J. Am. Stat. Assoc.*, 79(387):531-554.

8. Eichenbaum, M., Jaimovich, N., Rebelo, S. (2011). Reference Prices, Costs, and Nominal Rigidities. *American Economic Review* 101(1): 234-62.
9. Gelper, S., Wilms, I., Croux, C. (2015). Identifying Demand Effects in a Large Network of Product Categories. *Journal of Retailing*, 92(1):25-39.
10. Huang, T., Fildes, R. and D. Soopramanien (2014), The Value of Competitive Information in Forecasting FMCG Retail Product Sales and the Variable Selection Problem. *European Journal of Operational Research*, 237(2), 738-748.
11. Jungmeister, W-A. (1992). Adapting Treemaps To Stock Portfolio Visualization. *Technical Report UMCP-CSD CS-TR-2996,* College Park, Maryland 20742, U.S.A., 1992.
12. Koussoulakou, A., Kraak, M.J. (1995). Spatio-temporal maps and cartographic communication. *The Cartographic Journal,* 29, 101-108.
13. Krygier, J., Wood, D. (2011). Making maps a visual guide to map design for GIS. 2nd edn, Guilford Publications, New York, 256 pages.
14. Lengler, R., Eppler, M.J. (2007). Towards a Periodic Table of Visualization Methods of Management. In *P*roceedings of Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA, ACTA Press. pp. 1-6.
15. Levy D., Lee D., Chen H.A., Kauffman R.J., Bergen, M. (2011). Price Points and Price Rigidity. *The Review of Economics and Statistics*, 93(4): 1417-1431.
16. McAfee, A., Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review* 90(10), 60-68 .
17. Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5(2), 110-141.
18. McKinney, W. (2012). Python for Data Analysis. O'Reilly Media, 466 pages.
19. Nevo, A., Wolfram, C. (2002). Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. *RAND Journal of Economics*, 33(2), 319-339.
20. Oancea, B., Dragoecsu, R.M. (2014). Integrating R and Hadoop for Big Data Analysis. *Revista Romana de Statistica* 2(62), 83-94.
21. SAS. *Data Visualization Techniques: From Basics to Big Data with SAS Visual Analytics.* SAS: White Paper (2014).
22. Toro-González, D., McCluskey, J.J., Mittelhammer, R.C. (2014). Beer Snobs do Exist: Estimation of Beer Demand by Type. *Journal of Agricultural and Resource Economics*, 39(2), 1-14.
23. Tufte, E.R. (1983). The visual display of quantitative information. Graphics Press, Cheshire, Connecticut, 197 pages.