CS-BIGS 4(1): 9-22
http://www.bentley.edu/csbigs/documents/wang2.pdf

Dynamic Data Visualization

Chamont Wang
The College of New Jersey, USA

Michele Meisner
The College of New Jersey, USA

In this article, we use four examples to illustrate a variety of techniques for the visualization of complicated data sets. The examples include business data, storm tracking, New Jersey Department of Education records, and classroom observations. The techniques are used to deal with certain geo-spatial patterns and cross-tabulations on the fly. Video clips are referenced throughout to illustrate the interactivity, kinetic actions, and animations of these approaches. The article contains no math and is accessible to all statistics users, including students in high-school AP Stat classes.

Key Words: *Data Visualization, Geographic Patterns, Google Map, Cross-Tabulations*

Introduction

Modern data analysis often involves complicated data structures with multi-years, multi-categories, multi-geographic-regions, and layered cross-tabulations. Moreover, the data may change at an ever-increasing speed. For this kind of situation, traditional tools and code-writing may not be the best way to extract useful information out of a complicated data set.

In recent years, books and software packages have picked up the pace to provide users with new platforms for dynamic data visualization. A Google search on "data visualization" leads to 1,220,000 links. Examples that we like include JMP, RapidNet, Gephi, Perceptual Edge, to name a few.

In this article, we will present examples to illustrate the advantages and limitations of two different visualization technologies and show how to use the two to complement each other. The first is called Tableau and the second is Statistica. The reasons for this choice are as follows:

1. Students or anyone with basic statistical background can start using the tools after a single lab session.
2. The tools can handle complicated data sets rapidly.
3. Both are full of sophisticated techniques to challenge students. There are indeed countless directions to go when the user reaches the Jedi level.
4. Both come with a wide array of sample workbooks with the raw data included.
5. They promote the journey from Data to Story Telling.
6. In spite of their sophistication and advanced features, the guiding philosophy of these technologies is the simplicity of data visualization. This philosophy embodies what Albert Einstein said, "Everything should be made as simple as possible, but not simpler." It also echoes the da Vinci quote: "Simplicity is the ultimate sophistication."

We believe such a philosophy should permeate all phases of data visualization.

For data visualization, Tableau has further advantages:

1. It handles Geographic data with a few clicks of the mouse.
2. In addition, it provides a quick, clever link to Google Earth technology.
3. It is free for academic use.
4. It zooms in on any specific part of the data and then *exports* it for external use with great ease.
5. It uses a Dashboard technology to summarize key findings.

The creator of Tableau is a Stanford professor, Pat Hanrahan, who worked for a Defense Department project aiming at increasing people's ability to analyze information. He is a founding member of Pixar, the studio that made the animated films such as Toy Story and Wall-E (http://www.pixar.com/). His team comprises some of the best minds in the industry.

In our experience, the new technologies sharpen the user's mind on the intricacies of the data rather than taking the user's focus off the data as one often encounters when using traditional tools. In this article, we will present a number of examples to illustrate the power of these tools. The examples, on the other hand, should not be taken as the equivalence of the full power (or even a fraction) of what the new technologies can accomplish.

Example 1 (Super-Store Sales Data)

This data set has 26 columns and 8,400 rows; it is one of Tableau's sample workbooks and free data sets. Their sample workbook provides certain insights of the data; our analysis will venture into a different direction. To begin with, the variables in this data include

- Customer Name
- Customer Segment (Small Business, Corporate, Home Office, etc.)
- Customer State (New York, Ohio, Michigan, etc.)
- Product 1 – Category (Office Supplies, Technology, etc.)
- Sales Volume
- Profit
- Discount
- Others

This data set holds a lot of information about a specific company. Our goal is to dig deeper into some of these variables to decipher how well the company is doing and in which sales categories and geographic locations this company needs to improve.

To proceed, our first question was: What can we do with this data set? A few possibilities are as follows:

- Association Rule which is common in data mining; e.g., ID = Customer id; Target= Product. Companies such as Amazon.com, Walmart.com and countless others use Association Rule to great effect.
- Predictive modeling: Decision Tree, Regression, Neural Network, etc. (e.g., Target = profit, Predictors: sale, discounts, regions, categories, etc.).
- Data Visualization.

In this article, we focus on data visualization. In particular, we will throw a series of questions and then respond with *rapid-fire* answers. The answers below are static; to see them in action, please visit the links below: http://www.youtube.com/watch?v=CdjuKww1zQ8 and http://www.youtube.com/watch?v=BF1WLgBY3K4 for YouTube video clips. The video clips are also posted with this article on the journal web site.

This example is very useful for business applications and will be unfolded in the following manner:

- What is the company's bottom line within each product category and sub-category?
- What kinds of products sell well but are not profitable?
- How can geographic information be used to pinpoint the region where certain products are not profitable?
- Drill down on geographic and calendar information: For states like New Jersey, which year is least profitable? And in which part of New Jersey is the company not doing well?

We now proceed to answer the above questions in tandem:

1. A key issue about how a company is doing would be: what is the sales volume?

   In our video, one can see how in a few seconds, we produced the following chart:

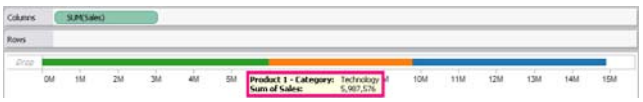


Figure 1.1. Sales Volume.

So the total sales is about $15 million. That is a large amount of money, and a thorough analysis of the data may be worthwhile. For instance, the chart shows that Technology accounts for almost $6 million of the sales, and a more detailed analysis may provide information to help improve sales.

2. Our next question is: what is the sales volume in each product category?

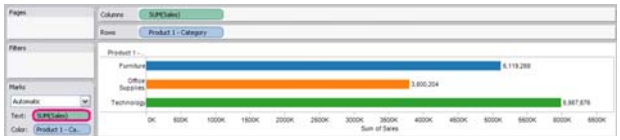


Figure 1.2. Sales Volumes for 3 Different Products.

In Table 1.2, the portion circled in red is called a **shelf** in Tableau. A drag of the variable, Sales, to the **Text** shelf immediately gives the exact numbers of the sales in each category. The separation in this plot makes it easier to see the exact sales volumes of the three categories.

3. Note that high sales volume does not guarantee high profit. Hence our next question is: Which category is most profitable? We see that Furniture sells well, but is not profitable:

Figure 1.3. Profit vs. Sales for 3 Different Products.

4. **Geographic information**: In real estate, the three most important variables are: location, location, and location. This is probably the same with many other business applications. For this study, we can use geographic information to pinpoint the region where furniture is not profitable. East (NY, NJ, …)? West (California, …..)? Central? Or South? Figure 1.4 shows that furniture is losing money in the East.

Note that in Figure 1.4, a Title and Caption have been added to the chart for future reference. These features aid in showing the complete picture and organizing your thoughts.

5. **Drill down**: we now examine the geographic information in more details in an attempt to see which State in the East is least profitable.

Figure 1.4. Profit by Region.

Figure 1.5. Profit by State.

The chart shows that New Jersey loses a lot of money on furniture, and Connecticut is in a similar situation.

Note that we used the **Filters** tab to select only the few States of interest. Again, by point-and-click, this is done in about 15 seconds.

6. **Calendar information**: The data contain information for years 2006, 2007, 2008 and 2009. So our next question is: for states such as New Jersey, which year is least profitable?

Figure 1.6. Profit by Year and by State.

The chart says that New Jersey lost about $9,500 in 2006, lost even more in 2008, but did better in 2009.

7. **Map:** The above analyses used only bar charts. We now add a new dimension by using a map to see which part of New Jersey is not doing well. The answer to this question requires only a few clicks and

drag-and-drops. We double click on Longitude and Latitude to bring up the map.

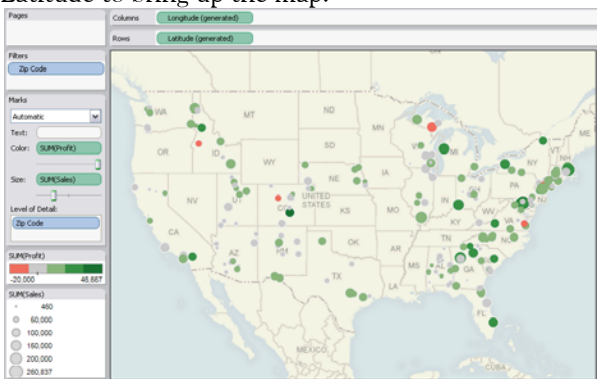


Figure 1.7. Profit Map.

This mapping technique can be used with any data set that has zip code, county information, or numeric values of latitude and longitude variables. The mapping does not require internet access, but the online version of Tableau provides additional map options.

In Figure 1.7, if we zoom in on New Jersey, a big red dot will appear in northen Jersey. In addition we can modify the map to show the *progression through multiple years*. Hovering the mouse on the dot displays the zip code information as shown in the next chart.

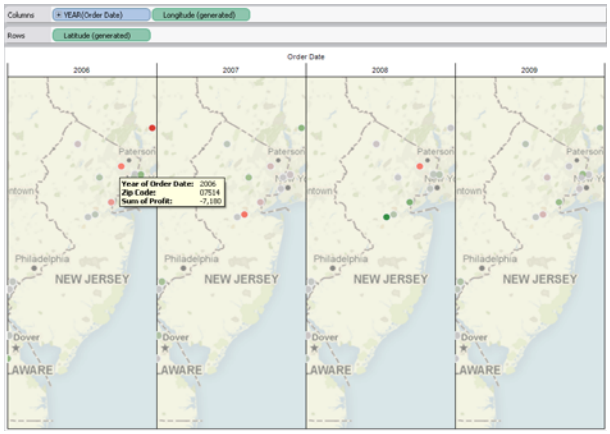


Figure 1.8. Geo-spatial Display of Profit in New Jersey in Different Years.

By moving the mouse over that specific location (zipcode = 07514, which is Paterson, NJ), one can see that the store *lost* about $7,200 in 2006 but broke even in 2009.

8. **A specific question is**: how is the store in Princeton area performing?

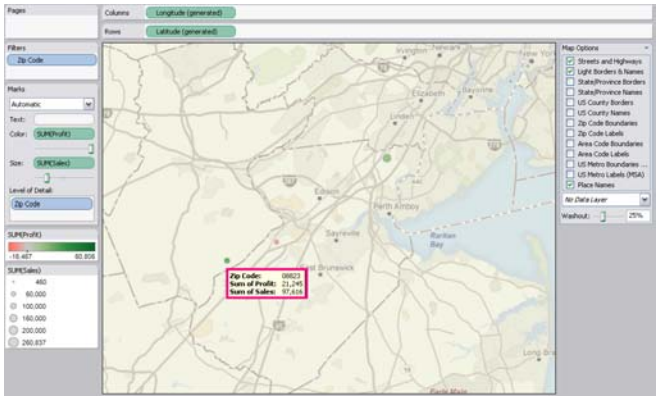


Figure 1.9. Profit Map near Princeton Area.

The chart shows that the Princeton location is doing well and making a profit of $21,245 over the study period.

9. **Drill down-II**: Finally, we want to know what types of furniture are not profitable.

Figure 1.10. Drill Down: Profit of Sub-category.

The chart shows that Bookcases and Tables are money losers. Nevertheless, it may be worthwhile to keep them in stock to help bring in customers for other items.

In conclusion, this case study shows that with the help of modern visualization tools, bar charts *alone* can be used to extract information rapidly from files with a complicated data structure. For this specific data set, one can easily obtain the following information:

- Drill down to specific year, month, region and sub-region.
- Pinpoint the regions that are in the red.
- View the above information in a calender sequence, either one period a time or multiple periods on a dashboard.

In addition to the dynamic use of Bar Charts, modern visualization tools allow the user to view geographic information with only a few clicks of the mouse. This is a leap from a book with words to a map with charts. A

further leap is to add calender information on the map, leading to a geo-spatial display for a broad view of multiple variables on different regions in different time periods.

For this data set, our focus is on sales volumn and profit. Other variables such as *Discount,* can also provide very useful insight of the data. See the following site for a spirited presentation that uses this variable effectively: http://www.tableausoftware.com/products/tour.

Example 2 (Storm Tracking and Animation)

This data set has 16 columns and 572 rows. It was adapted from a Tableau sample workbook. Their static chart led us to modifications and animations in this study. The variables in the data include

- Storm Name (ALEX, BONNIE, DANIELLE, etc.)
- Storm speed (mph)
- Wind speed (kt; 1 knot = 1.852km/hr = 1.151 miles/hr)
- Pressure (mb; 1 millibar = (1/1000) bar; 1 bar corresponds to the atmospheric pressure on earth at sea level)
- Longitude (deg)
- Latitude (deg)
- Date
- Others

In this example, we will investigate the relationships between Wind Speed, Storm speed, and Pressure. Since the data includes Latitude, Longitude, and Date, we can also perform an animation of storm movements. Interestingly, when we performed this animation in class, a student gushed and asked the following question: "Is this what they do on the Weather Channel?"

To perform the storm tracking, we double click on Longitude and Latitude respectively to activate the map. Adding Color (Storm Name), Text (Storm Name), Size (Storm Speed), and Filtering out a few storms make the map easier to read.

To animate the storm tracking, we drag **Date** to the **Pages** shelf, change the **Date** from Year to Day, then click on the **Play** button to see the storms in motion. The speed of the animation can be adjusted if needed. A video of the action is available at http://www.youtube.com/watch?v=-muHR6lHbko (5:17 minutes), and is posted with the article on the journal website.

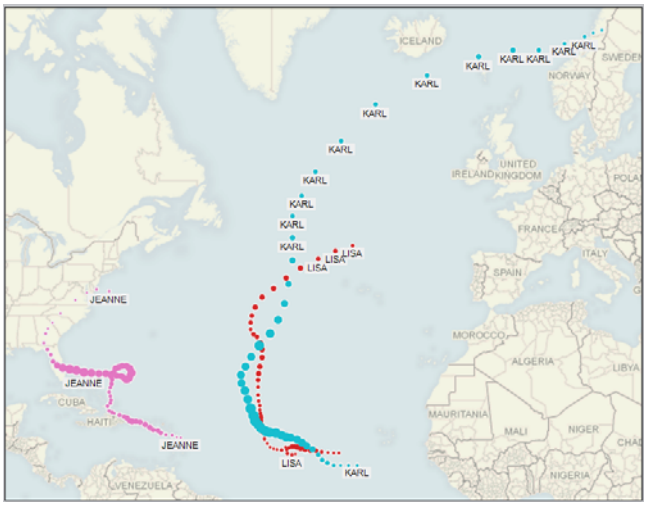


Figure 2.1. Storm Tracking.

Figure 2.1 shows the following:

- During the same length of time, Karl traveled very far, while Jeanne stayed within a more confined area.
- The size of the dots reflects the Wind Speed of the storm. Karl gained a lot of strength early on and then remained a strong force for a very long distance on the sea.
- Jeanne, on the other hand, gained a lot of power in the middle of the course, maintained its strength until hitting Florida, and then weakened substantially on the mainland.
- Karl and Lisa did not threaten people on land, while Jeanne might have caused severe damage to lives and properties.

A potential application of the above technique is the animation of Bubble graphs. A Google image search of Bubble plot yielded 453,000 charts. It may be possible to release some of these bubbles in sequence *if* the time stamp is available in the data.

Next we will examine six (6) variables on a histogram. To begin with, we high-light **Wind Speed** and click on the **Show Me** button to activate the plain histogram:

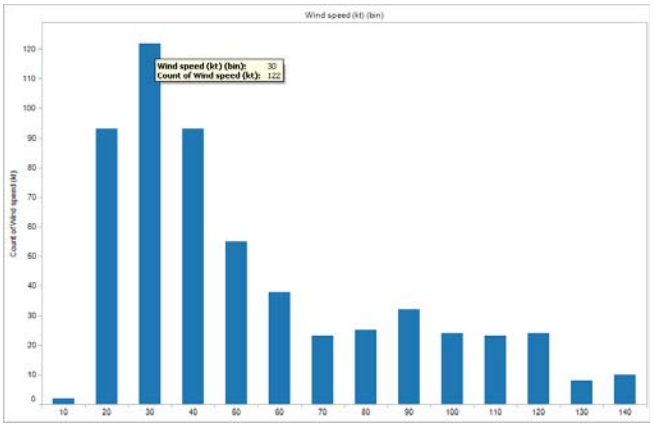


Figure 2.2. Histogram of Wind Speed.

Here we can see that a wind speed of 30 corresponds to the highest count of wind speed, and as the wind speed becomes greater than the 30 peak, the count drops.

To view more variables in this same chart we add **Pressure** to the **Size** shelf, and add **Storm Speed** to the **Color** shelf:

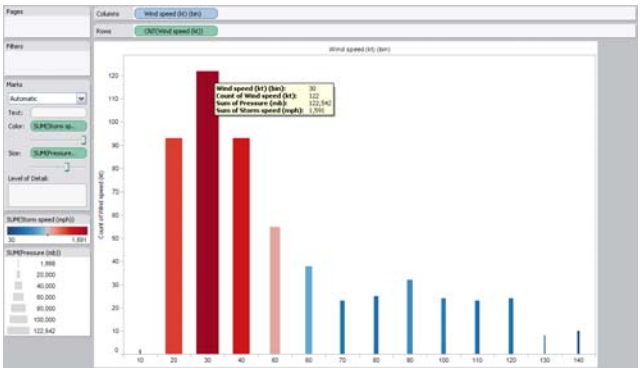


Figure 2.3. Histogram = Wind Speed;
Color = Storm Speed; Size = Pressure.

Here the width of each bar shows the pressure and the color represents the storm speed. The wind speed of 30, which had the highest wind speed count, also corresponds to the greatest pressure (thickest bar) and the highest storm speed (darkest red color). Students who saw this chart tend to conclude that in general, wind speed, storm speed, and pressure are correlated. In the subsequent analysis, we will discuss this issue in more depth.

Another observation is that storms with a wind speed of 10 and 130 seem to have very similar pressures (as seen by the similar widths). The qeustion arises of why they would have the same pressure. Hovering the mouse to the bars would reveal that the pressure is 1998 mb at the left end and 7385 mb when the Wind Speed is 130. So the thin bars are a little misleading. A related issue of the bar size will be discussed after Figure 2.5.

Now we add **Storm Name** to the **Text** shelf:

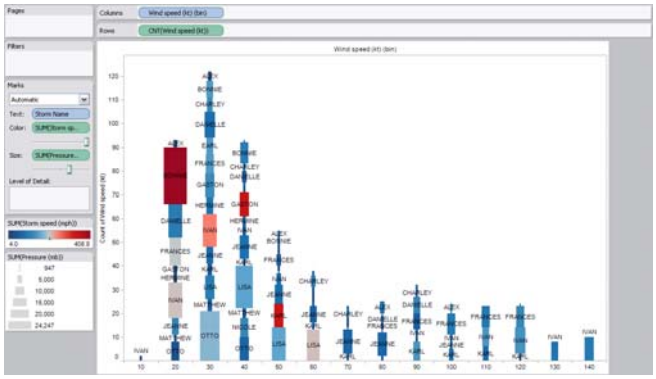


Figure 2.4. Histogram = Wind Speed;
Color = Storm Speed; Size = Pressure; Text = Storm Name.

Figure 2.4 is hard to read on a printing medium such as a piece of paper or a pdf file. As a result, we use the **Filter** option to focus on three storms (Jeanne, Lisa, and Karl) that were shown on Figure 2.1:

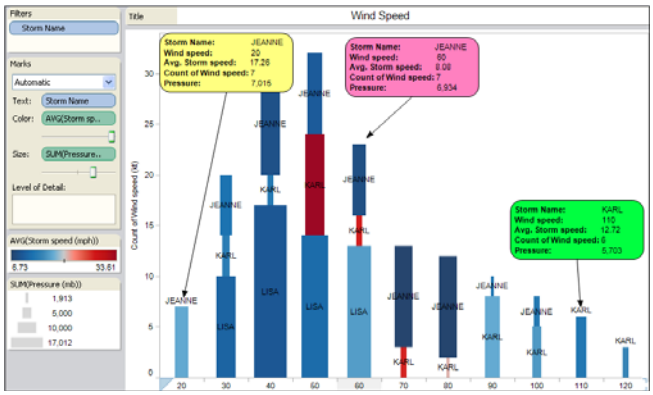


Figure 2.5. Histograms of Jeanne, Lisa, and Karl.

The chart shows that the Wind Speeds (x-axis) of Lisa do not go beyond 60 kt. In contrast, the Wind Speeds of Jeanne and Karl may reach 100 kt or more. One question would be the overall relationships between Wind Speed, Storm Speed and Pressure. For this task, we will use different techniques in the subsequent discussions.

The trade-off between Figures 2.4 and 2.5 is that the first chart provides more information (6 variables for each of the 15 storms) while the other is easier to understand (5 variables for each of the 3 storms). To view the details of each square, we simply hover the mouse to the box. Furthermore, one can add the sixth variable **Date** to the

Pages shelf and flip the pages to see how the Histogram evolves with time.

Note that the legend of the Histogram says that the size of the bar is determined by the Sum of the Pressure. You can easily change the Sum of Pressure to the Average of Pressure (which ranges from 917 to 1008 mb). As a result, the difference of the bar sizes would be so small that the chart would not be as appealing. A purpose of our Figures 2.3-2.5 is to show the versality of the chart, so the more interesting visual was used; the true utility of using the bar size may be realized in other studies.

Next we examine the relationship between **Wind Speed** and **Storm Speed**:

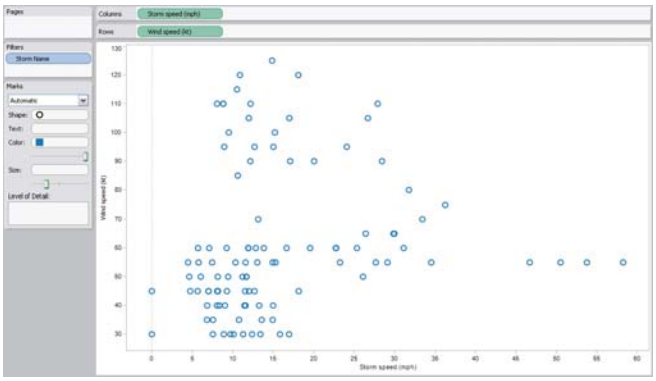


Figure 2.6. Wind Speed vs. Storm Speed: no pattern in the scatterplot.

The scatterplot appears *boring* and does not reveal any relationship between the two variables. To brighten up the dull scatterplot, first we add **Storm Name** to the **Color** and **Text** shelves and then add **Trend Lines** to produce the following chart:

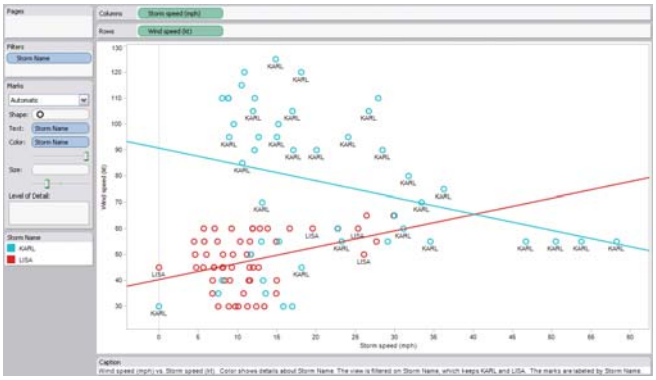


Figure 2.7. Wind Speed vs. Storm Speed; Color = Storm Names.

The chart shows that for Karl, Wind Speed and Storm Speed are negatively correlated, while for Lisa, it is the opposite. This technique has been around for decades, and should be able to yield more insights in countless other studies.

The next chart (Matrix Plot) shows the scatterplots of Wind Speed, Storm Speed, and Pressure. On the diagonal of the matrix, the histograms of the 3 variables show the following:

- Storm Speed is skewed to the right (with few storms moving at high speed),
- Wind Speed is also skewed to the right but not as severely as Storm Speed, and
- Pressure is skewed to the left, contrary to Wind Speed and Storm Speed.

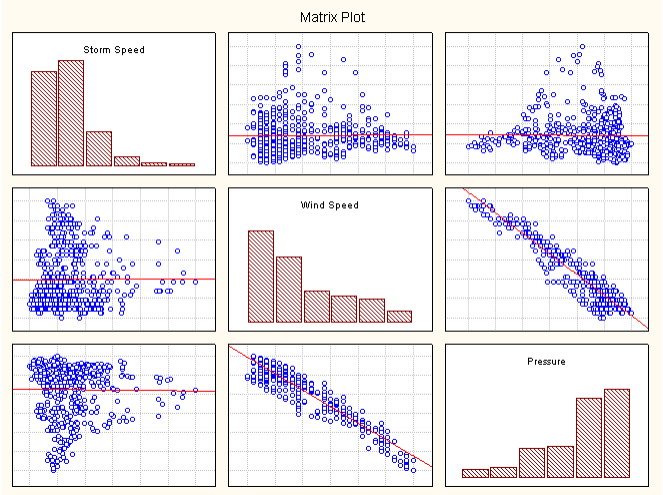


Figure 2.8. Wind Speed, Storm Speed, and Pressure: Histograms and Scatterplots.

The scatterplots, on the other hand, show the following:

- In general, Storm Speed is not correlated to Wind Speed or Pressure.
- Wind Speed and Pressure are negatively correlated. This is consistent with meteorological observations that tropical cyclones generally occur in areas of low atmospheric pressure, with the lowest pressures recorded at the centers of the cyclones.

The raw data is available at the following site; the site also provides other ways to visualize this particular data set: http://www.tableausoftware.com/learning/examples/storm-tracking.

Example 3 (Educational data)

In this example, we used data from the website of New Jersey Department of

Education: http://www.state.nj.us/education/data/. The variables include

- School District
- Budget
- Graduation Rate
- Dropout Rate
- Date
- Others

First, we tried to examine the **Dropout Rate**, and we produced the following chart:

Figure 3.1. Dropout Rates of All Counties.

In Figure 3.1, the horizontal axis is academic year, and the vertical axis is the Dropout Rates of the Counties. The chart is hard to decipher, so we added **District** to the **Pages** shelf and to the **Text** shelf. This is a very important technique.

Now the Pages can be flipped to see the dropout rates for each district. For instance, the dropout rates at Atlantic County went downward, which is an improvement. However dropout rates at Ocean County, after some improvements, went up significantly:

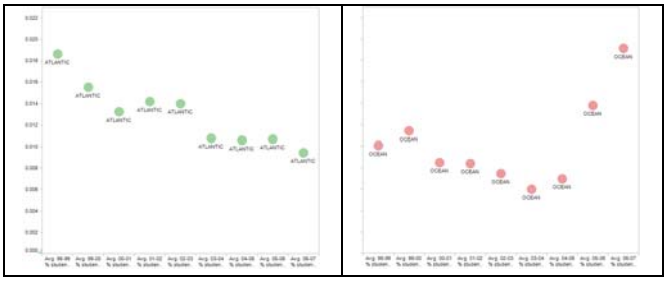


Figure 3.2. Dropout Rates of Atlantic and Ocean Counties.

The scatterplot below examines Dropout Rates against Average Spending per Student for the academic year 06-07:

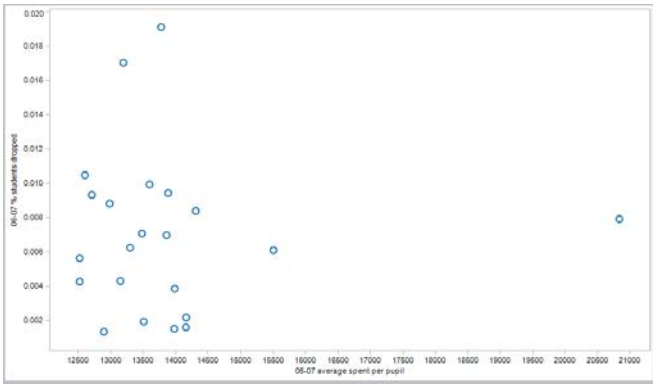


Figure 3.3. Dropout Rates vs. Average Spending per Student.

The conventional way of dealing with scatterplots like the ones in Figure 3.3 is to fit a regression line. But this would not be helpful. Instead, we add **median lines** on both the x-axis and the y-axis (Color = District; Text = District):

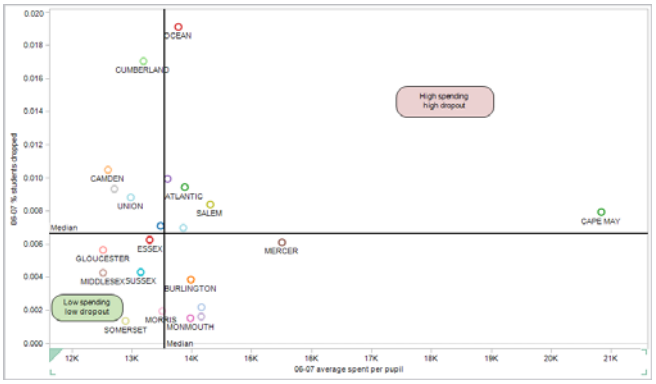


Figure 3.4. Median Splits.

The first quadrant shows the worst (names of the districts with High Dropout Rates and High Spending), while the third quadrant shows the best (names of the districts with Low Dropout Rates and Low Spending). Clearly certain districts have things to learn from others. In short, Figure 3.4 identifies the following: the best districts, the worst, and the outlier, Cape May.

In addition, note that the scatter-plot in Figure 3.4 displays a total of 6 quantities: the two variables on the axes, plus color and County names, plus the medians on the x- and y-axes. Furthermore, we can view all detailed information by moving the mouse to a specific county. This is data visualization on the fly.

Recall that Bar Charts can also display multiple variables (see Figure 2.4 with 5 variables on the chart). But a scatterplot can lock two specific variables on the chart to provide a different way of stratification.

The next chart shows Graduation Rates against Average Spending per Student. The Median Splits are used to decipher which districts have High Spending and Low Graduation rates.

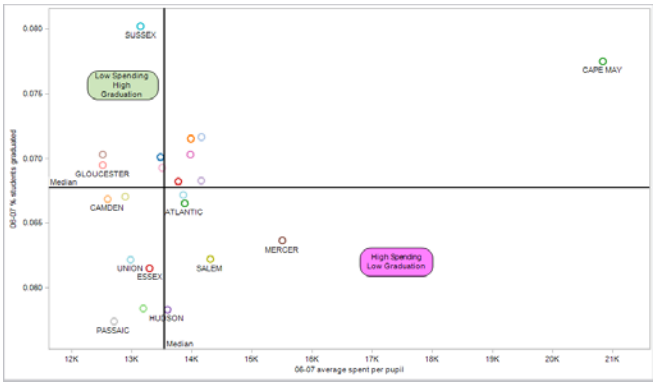


Figure 3.5. Graduation Rates vs. Average Spending per Student.

The second quadrant shows the best (districts with Low Spending and High Graduation), while the fourth quadrant shows the worst (districts with High Spending and Low Graduation rates). The reasons for the Low Graduation Rates, however, are not clear from the data.

On the chart, Cape May is an outlier: high spending and high graduation rate. Citizens in that county probably are committed to good education regardless of the cost.

The next chart shows that the overall **Average Spending per Student** almost doubled over a ten-year period: $8,405 (in 1998) to $15,053 (2009).

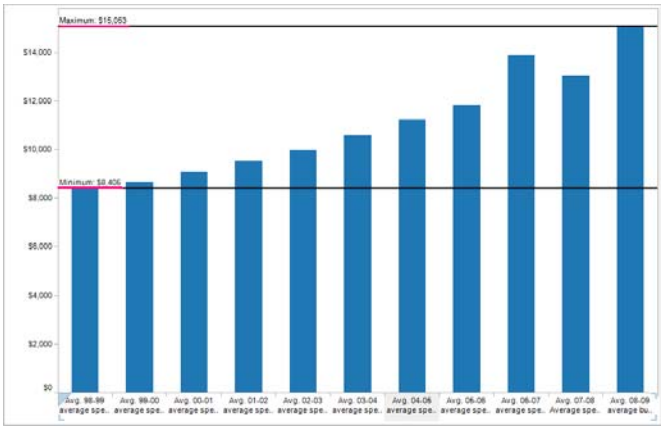


Figure 3.6. Average Spending per Student, 1998-2009.

Figure 3.6 appears to imply the trend of doubling for every county. However, when we added **District** to the **Color** shelf, the chart shows that Mercer county is the big spender in the last year of the study period.

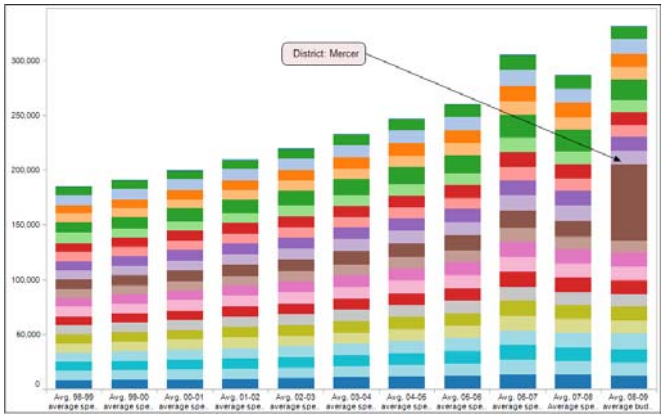


Figure 3.7. Average Spending per Student, by District.

Now that we have two outliers: Cape May (Figure 3.5) and Mercer (Figure 3.7); the next chart (Figure 3.8) compares the two over a period of time from 1998 to 2009.

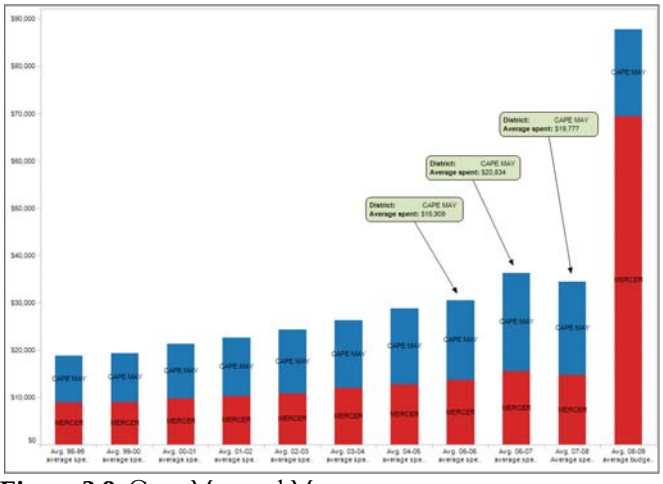


Figure 3.8. Cape May and Mercer.

The chart indicates the following: the spending of Cape May peaked in 06-07, after gradually increasing from 1998, with a 25% jump from 05-06 to 06-07. It overtook Mercer in 06-07 and 07-08, but Mercer went through the roof in 2008-2009. Cape May, on the other hand, cut back to a certain extent after the spending spree.

We now summarize Figures 3.3 to 3.7 on a dashboard:

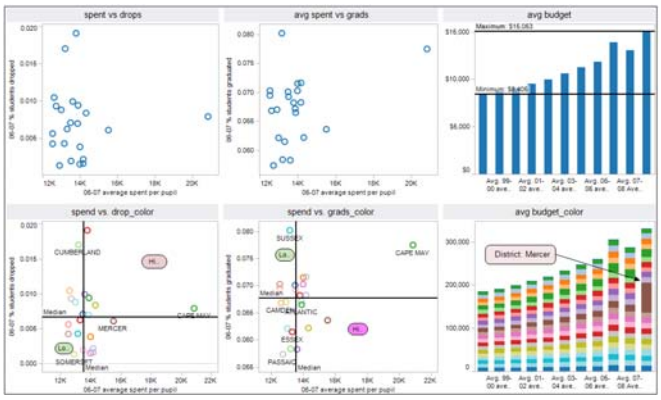


Figure 3.9. Dashboard of the NJ-DoE data.

The top row displays plain graphs. The bottom row demonstrates that with a few clicks, modern visualization techniques can reveal a lot more information. In addition, the bottom row shows dropout rate, graduation rate, and average spending in one glance. The graphs complement each other to provide decision makers a better view of the overall situation.

Finally, we double clicked on the Longitude and Latitude variables to bring up a map. We then put the three variables (**Spending** in the **Color** shelf, **Graduation Rate** in the **Size** shelf, and **Dropout Rate** in the **Text** shelf) on a map and then move the mouse to high-light any case (e.g., Atlantic county in Figure 3.10):

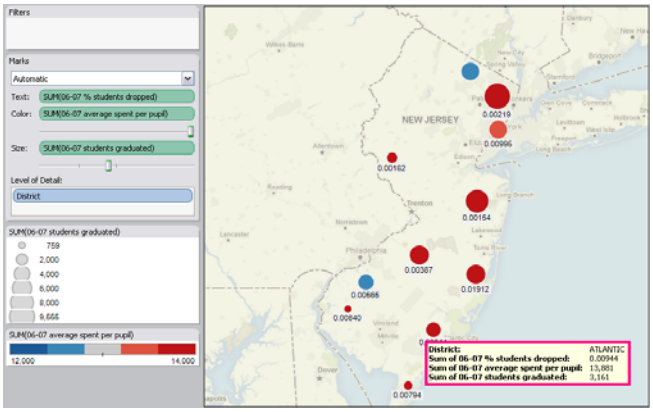


Figure 3.10. Color = Spending, Size = Graduation, Text = Dropout.

The red box on the chart shows the exact Dropout rates, Graduation totals, and Spending at Atlantic county. The numbers may help administration on future planning.

On the map, we right click on any location to find the option of **View Satellite Image** which links Tableau maps to Google maps.

Figure 3.11. Satellite Image.

This technique can be useful in many real-world scenarios. It can help with Campus Space Analysis, which may include the planning of campus construction, the management of campus wireless networks, and perhaps other novel applications. A different issue would be the crime and security concerns on the campuses. The U.S. Department of Education reports that campus crimes are occurring at surprisingly alarming rates. (http://www2.ed.gov/admins/lead/safety/crime/criminaloffenses/index.html). The crimes include Aggravated Assault, Arson, Burglary, Sex Offenses, Motor Vehicle Theft, Murder, Manslaughter, and Robbery. In the category of Robbery alone, there were 11,659 cases in 2001 and 9,367 cases in 2002.

New technologies to fight against crimes are on the rise (see, e.g., Westphal, 2008). Examples in this regard include monitoring crime statistics by type, location, and a number of other criteria. Such applications are very useful in the fast-growing field of Fraud Detection, which includes the uncovering of Construction Fraud, Money Laundering, Alien Smuggling, Social Network Analysis, and Financial Transactions Investigation.

In Fraud Detection conferences that we have attended, quick links to Google Maps have generated considerable enthusiasm (see, e.g., i2 Intelligence-Led Operations Platform, http://www.i2group.com/us/products--services and http://www.i2group.com/us/news--events/events). We do believe that this new technology will be very useful in real-world crime fighting and fraud detection.

Another application of the link to Google Maps is the monitoring of 911 calls for emergency responses to a specific geographic location (medical assistance, assault, or fire related incidents). An example can be found in the Tableau Wow sample workbook.

Two video clips for Example 3 can be found at http://www.youtube.com/watch?v=FQ4S_1KOffE, http://www.youtube.com/watch?v=SHiEnIdUcuE, and are posted with this article on the journal website.

Example 4 (Finding Patterns in the Student Exam Data)

In this example, we will explore a different technology that is more in the mode of traditional statistical analysis. This approach complements the technology in Examples 1-3 of this article. The tools we will use are part of the Statistica package. A rich array of examples in this category can be found at http://statsoft.com/textbook/graphical-analytic-techniques/?button=2.

The Data

The data in the following table contains the exam scores of 19 students in an Introductory Statistics course at a college:

Exam 1	Exam 2	Final
78	72	79
65	92	71
52	48	41
94	106	99
37	31	29
99	90	95
71	79	80
85	67	73
42	49	53
68	74	61
49	41	40
90	91	95
87	94	89
98	60	40
77	78	72
92	87	96
50	48	42
91	80	75
71	90	84

Given the data, one can build regression models to forecast the Final grade. In this article, however, we will focus on the visualization aspects of the data. A total of three video clips are supplied to display the steps of this approach and are posted on the journal website:

a) Pattern Detection in 3-D (1:20 minutes)
http://www.youtube.com/watch?v=Tg0R-dr3kiU
b) Cowboy Lasso and Statistical Lasso ():57 minutes)
http://www.youtube.com/watch?v=ngetV0af_Qc
c) Icon Plots (2:32 minutes)
http://www.youtube.com/watch?v=llfdhurNrs8

To begin with, we try to detect patterns by the use of a 3D rotation graph. The action can be viewed in a short video clip at http://www.youtube.com/watch?v=Tg0R-dr3kiU (1:20 minutes). The default 3D graph and a version of its rotation are in Figure 4.1 below. Examining the graphs, we ask the question: *are there any patterns in the data?*

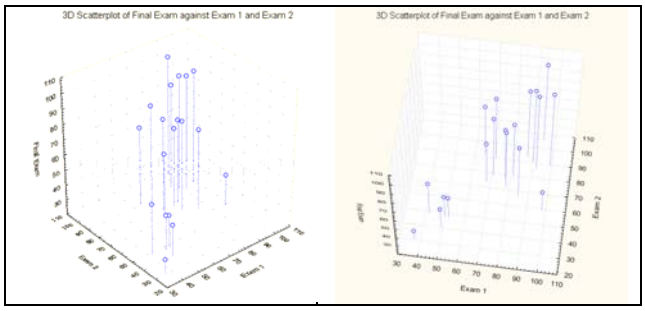


Figure 4.1. 3D rotation graphs of the Exam-Final data.

In our classroom surveys, most students see no pattern in the original depiction. After rotating the plot they notice a linear relationship in the second graph in Figure 4.1 but not much else. A few see the *outlier*. After some explanations, they see that in addition to the outlier, there are *three clusters*: top students, middle students, and failing students.

In the hunt for *failing students*, we use a Statistical Lasso, which was inspired by the traditional Cowboy Lasso:

Figure 4.2. Cowboy Lasso.

For both lassos, the idea of capturing a target for a closer look is the same:

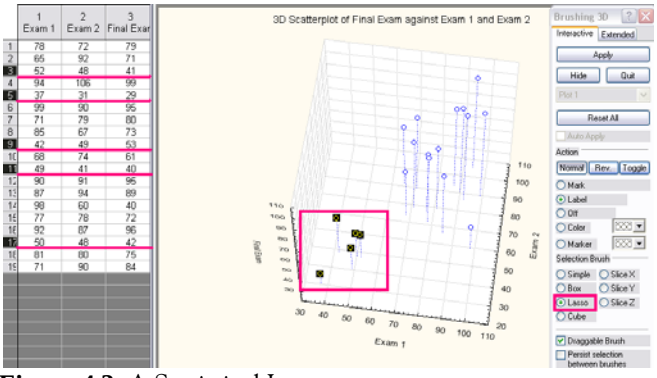
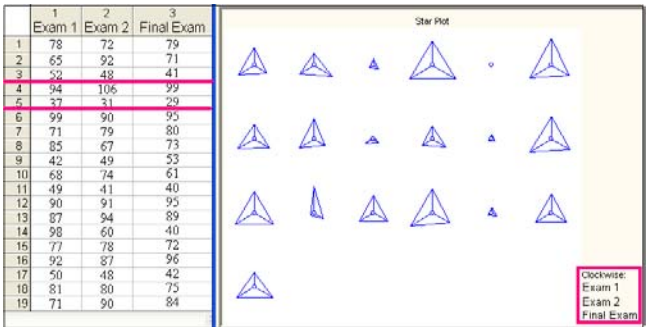


Figure 4.3. A Statistical Lasso.

The hunt reveals that Students #3, 5, 9, 11, and 17 are failing and need immediate attention. A video clip to capture this action is available at http://www.youtube.com/watch?v=ngetV0af_Qc.

Next we use a variety of Icon plots to investigate the students' exam scores. The first of this kind is called **Star plot**:

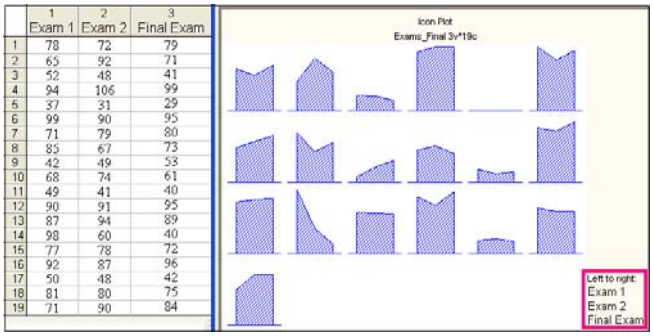


	Exam 1	Exam 2	Final Exam
1	78	72	79
2	65	92	71
3	52	48	41
4	94	106	99
5	37	31	29
6	99	90	95
7	71	79	80
8	85	67	73
9	42	49	53
10	68	74	61
11	49	41	40
12	90	91	95
13	87	94	89
14	98	60	40
15	77	78	72
16	92	87	96
17	50	48	42
18	81	80	75
19	71	90	84

Figure 4.4. Star plot.

In Figure 4.4, a separate star-like icon is plotted for each student; relative values of the test scores for each case are represented (clockwise, starting at 12:00) by the length of the rays in each star. The ends of the rays are finally connected by lines.

The plots indicate that the 3rd student does not do well, and the 4th student is excellent. As for the 5th student, with grades 37, 31, and 29, there is nothing to plot. This is also reflected in the following **Profile plot**:



	Exam 1	Exam 2	Final Exam
1	78	72	79
2	65	92	71
3	52	48	41
4	94	106	99
5	37	31	29
6	99	90	95
7	71	79	80
8	85	67	73
9	42	49	53
10	68	74	61
11	49	41	40
12	90	91	95
13	87	94	89
14	98	60	40
15	77	78	72
16	92	87	96
17	50	48	42
18	81	80	75
19	71	90	84

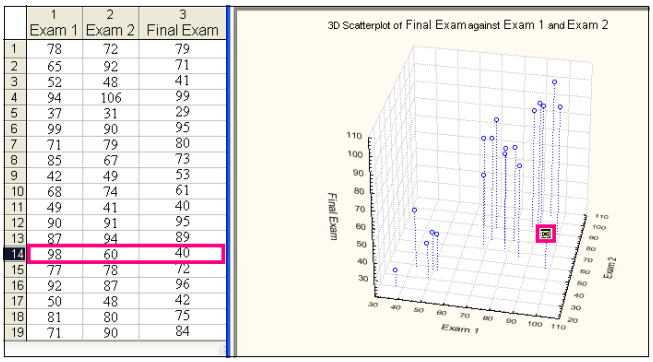
Figure 4.5. Profile plot.

In this diagram each plot represents one student's tests scores, with the height representing each exam. Looking from left to right on the horizontal axis shows the scores in order, from Exam 1 to Final Exam. This plot reveals more information than the Star Plot in Figure 4.4. For example, examining each of the 19 plots, we ask the question: *Which profile is the most interesting?*

The responses we have typically received said that: "#14 (3rd row, 2nd item) is the most interesting".

What happened was that Student #14 had learned some kind of Statistics from certain unknown places and indeed performed very well in Exam-1. Then he cut classes, missed homework assignments, skipped group meetings, and got only 60 on Exam-2. The instructor intervened to no avail. By the time of the Final, the grade of that student went down the drain.

The story of Student #14, as shown in the Profile Plot, stands out from the other students as a straggler. By using the technique of Lasso, we can see again that the outlier on the 3D rotation plot is indeed Student #14:



	Exam 1	Exam 2	Final Exam
1	78	72	79
2	65	92	71
3	52	48	41
4	94	106	99
5	37	31	29
6	99	90	95
7	71	79	80
8	85	67	73
9	42	49	53
10	68	74	61
11	49	41	40
12	90	91	95
13	87	94	89
14	98	60	40
15	77	78	72
16	92	87	96
17	50	48	42
18	81	80	75
19	71	90	84

Figure 4.6. Outlier that is identified by Lasso and by the Profile plots.

Being able to pinpoint this type of student can help intructors find the causes of poor performance to identify specific student needs.

There is also another kind of Icon plot called **Chernoff Faces**. In our classes, we do not show these graphs. Instead, we give hints to students and let them discover the option by themselves. The use of these graphs in introductory statistics courses and advanced data mining classes has proved useful and actually entertaining; the

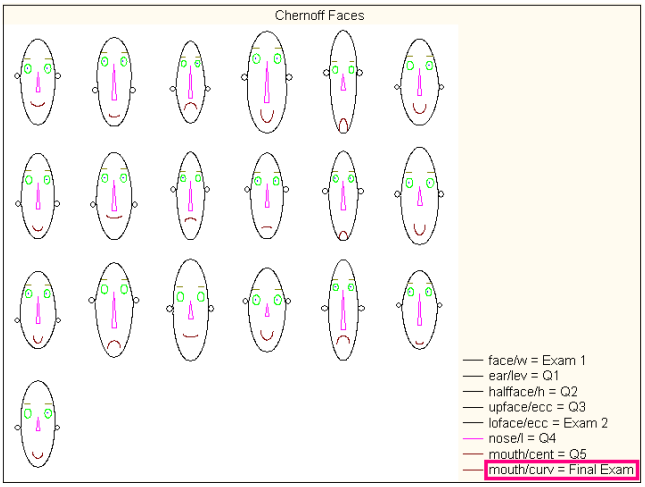


Figure 4.7. Chernoff Faces.

technique should be equally interesting to high-school students.

In Figure 4.7, the plot legend includes 5 quiz grades, Q1-Q5, so that the 8th variable (Final exam) would be represented by the mouth curve: happy faces for high scores, and crying faces otherwise – the sharper the curve, the worse the grade. The plots show that students #3, 5, and a few others were not doing well on the Final.

Concluding Remarks

In this article, we explored a variety of techniques for rapid-fire data visualization. The techniques include map creation and Google satellite images to extract useful information from geo-spatial patterns. This new technology can also handle time elements and calendar information that evolve through multiple years, months, or days. The new tools may help in space analysis, crime fighting, and the management of 911 emergency calls, to name just a few.

In addition, the article shows that with the help of modern visualization tools, traditional bar charts and scatterplots can be used to extract information rapidly from files with complicated data structures. The applications include drill down to specific year, month, region and sub-region, plus the display of multiple charts on a dashboard. In addition, our study indicates that both bar charts and scatterplots can handle multiple quantities of 6 or more variables. For certain applications, regression lines on the scatterplots do not provide any information; instead, the *median splits* of the two specific variables on the chart may be able to provide a different perspective of the study population. Furthermore, one can add color or text to reveal hidden patterns and to examine all detailed information by moving the pointer to a specific category. This is data visualization on the fly.

Finally, we use animation, 3D pattern detection, statistical lasso, and icon plots to help discern clusters and outliers. These techniques can be used to capture a target group and to help pinpoint the causes of a specific problem for individuals within this group. The example we used involves only a small data set with 19 cases and 3 primary variables. For large data sets, the methods may be equally fruitful when the techniques are used in sequential manner to break down the huge quantity of information into small pieces that are more manageable.

In conclusion, as the data sets in modern society continue to grow in size and complexity, analysts are facing up to new challenges and new opportunities. Visualization techniques that are developed by non-statisticians may complement traditional tools in a very fruitful manner. We expect the different approaches to converge in the near future to open up new territories for data exploration.

REFERENCES

Chernoff, H. 1973. The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, 68 no. 342: 361–368.

Few, S. 2009. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press. http://www.perceptualedge.com/library.php#Books.

i2 Intelligence-Led Operations Platform, http://www.i2group.com/us/products--services

StatSoft 2010. Selected Topics in Graphical Analytic Techniques, http://www.statsoft.com/textbook/graphical-analytic-techniques/?button=2.

Tableau 2010. Visual Examples, http://www.tableausoftware.com/learning/examples.

Westphal, C. 2008. *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press. http://www.crcpress.com/product/isbn/9781420067231;jsessionid=1jdutu5JTu2pZytie7mrMA**

Acknowledgments: An early version of the NJ_DoE data was prepared by Amanda Mihalic and Harold Schoonover, who also performed parts of the analyses on the Histogram, Graduation and Dropout rates under the supervision of the first author. Meiyi Zhang did the earlier version of certain video clips which helped in the taping of the later versions. We appreciate the Tableau tech support for their assistance in sorting out a number of issues, especially the activation of the Google map. Finally, we thank the reviewers for their valuable comments, which improved the paper in a very significant manner.

Correspondence: wang@tcnj.edu

APPENDIX: YouTube Videos

Example 1 (Superstore):
http://www.youtube.com/watch?v=CdjuKww1zQ8 (3:03 minutes),
http://www.youtube.com/watch?v=BF1WLgBY3K4 (3:47 minutes).

Example 2 (Storm Tracking)
http://www.youtube.com/watch?v=-muHR6lHbko (5:17 minutes).

Example 3 (NJ Department of Education data)
http://www.youtube.com/watch?v=FQ4S_1KOffE (4:14 minutes)
http://www.youtube.com/watch?v=SHiEnIdUcuE (0:55 minutes, Satellite Image)

Example 4 (Finding out your grade)

- Pattern Detection in 3-D
  http://www.youtube.com/watch?v=Tg0R-dr3kiU (1:20 minutes)
- Cowboy Lasso and Statistical Lasso
  http://www.youtube.com/watch?v=ngetV0af_Qc (0:57 minutes)
- Icon Plots
  http://www.youtube.com/watch?v=llfdhurNrs8 (2:52 minutes)