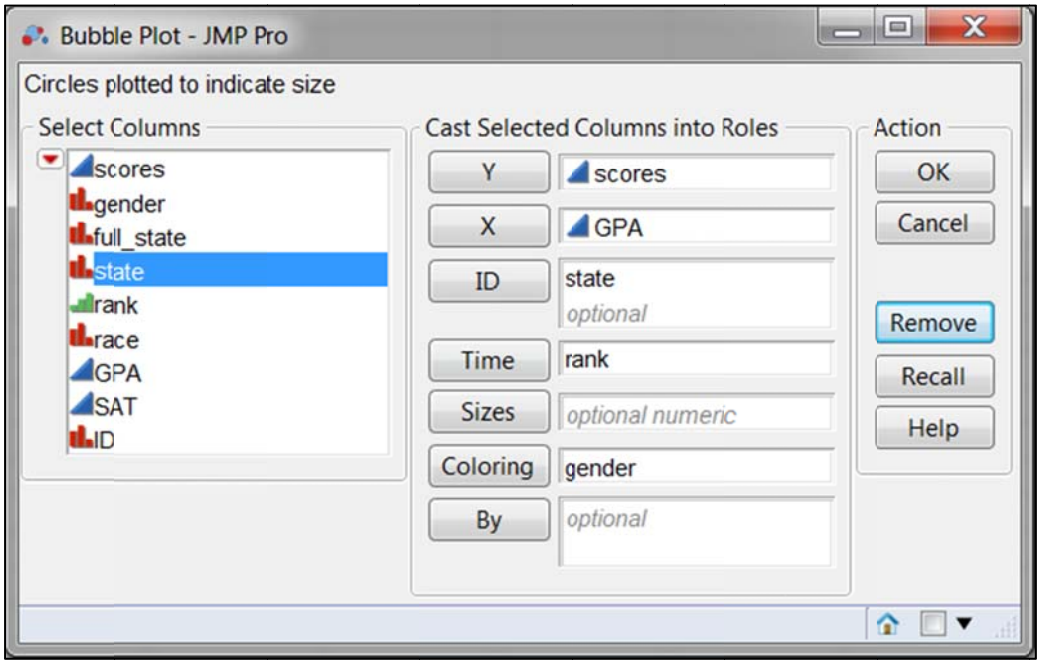
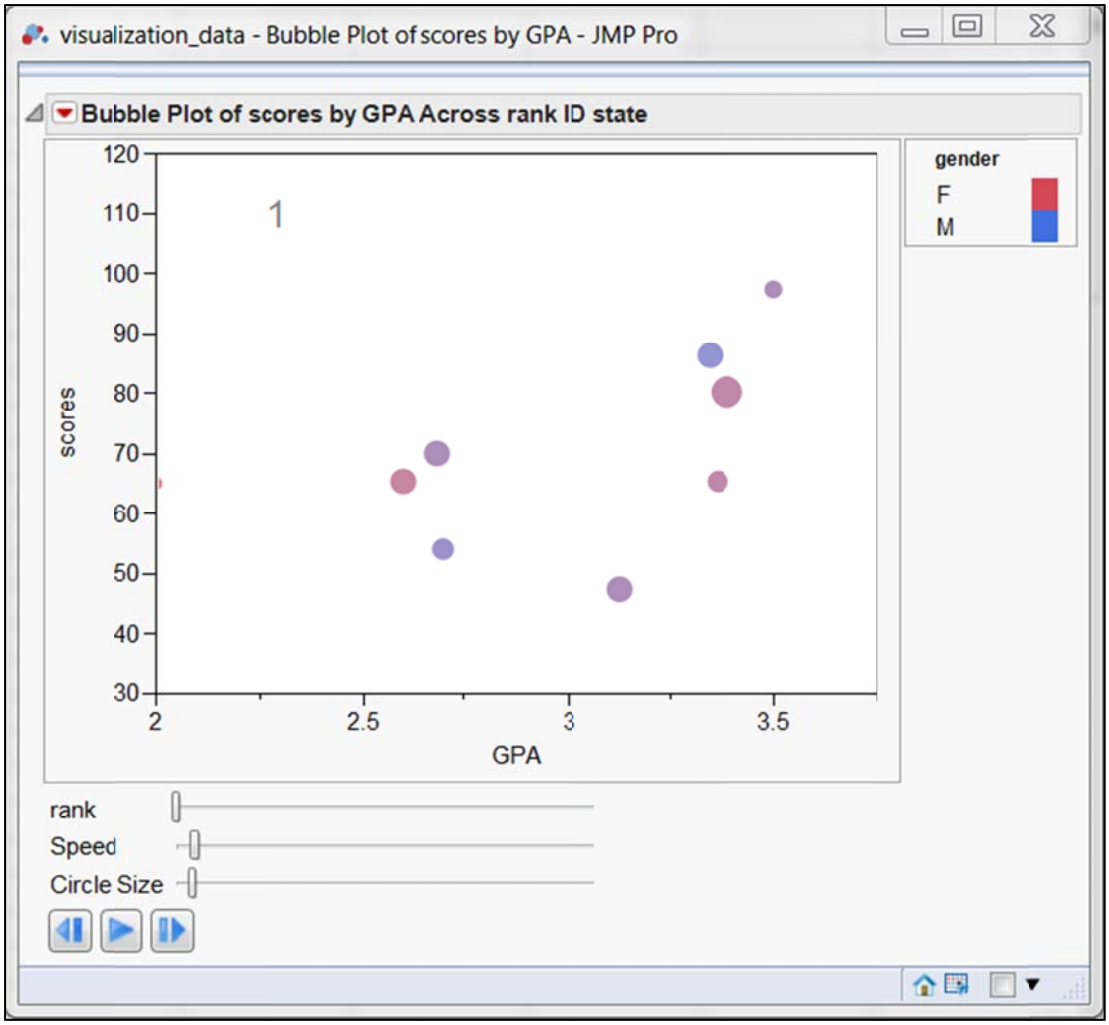
Examining the Bubble Plot frame by frame for multi-dimensional data visualization

Chong Ho Yu, Ph.D. (2013)
Azusa Pacific University
chonghoyu@gamil.com
http://www.creative-wisdom.com/computer/sas/sas.html

Depicting multiple variables in multiple dimensions, a problem known as "the curse of dimensionality," is always a challenge to data visualizers. Many software packages offer animated-based solutions, in which the hidden dimension is treated as the temporal dimension. The Bubble plot in JMP is a typical example. However, the author found that although an eye-catching movie can impress your audience, it is difficult to understand the inter-relationships among many variables when the movie flies through your eyes within a few seconds. In response to this problem, in the following example we will walk through the Bubble plot frame by frame instead of watching a movie.

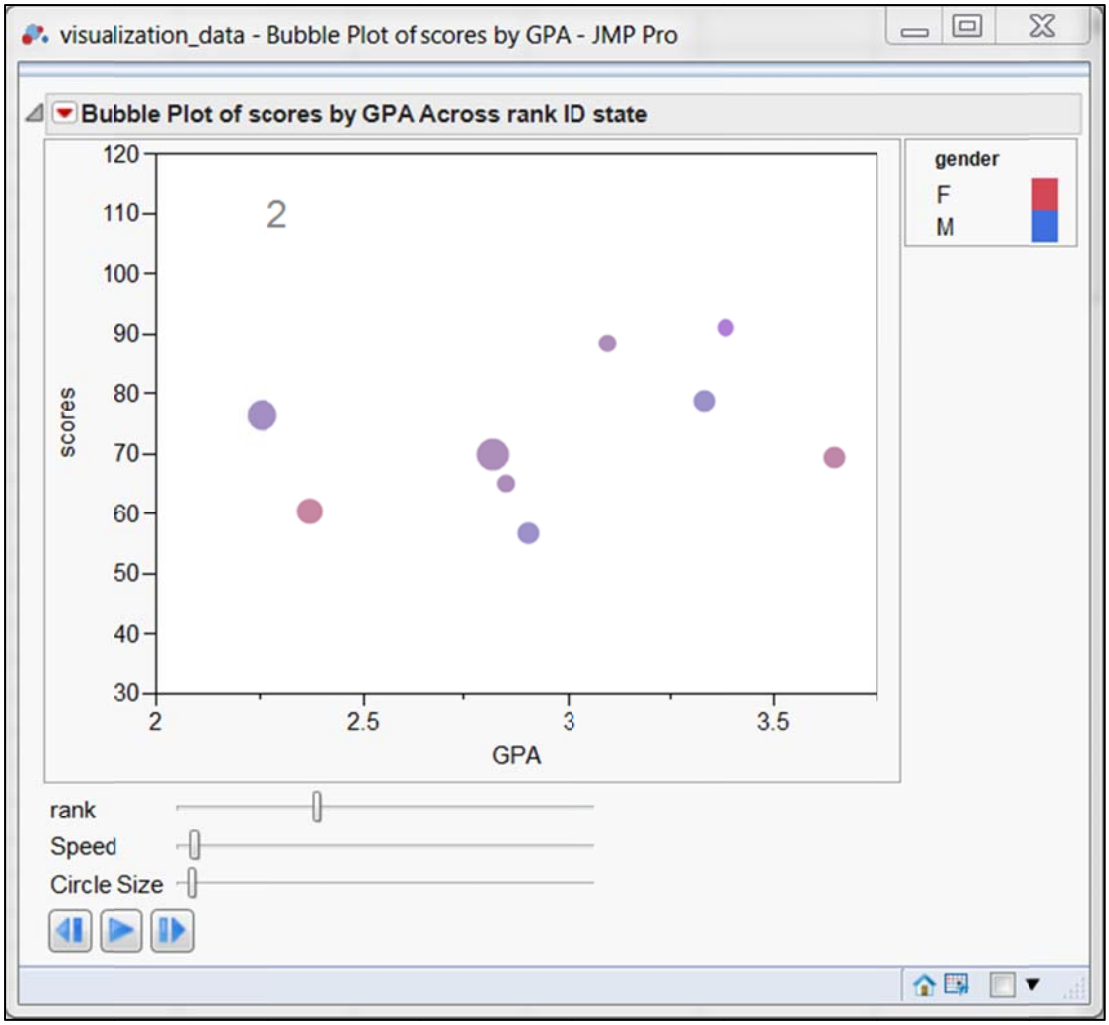


This data set carries academic records and demographic information of 90 students. The author would like to investigate the relationship between their high school GPA and their college test scores. The researcher also wants to understand how the preceding relationship is moderated by various variables, such as their origin (home-state), academic rank (freshman, sophomore, junior, and senior), and gender. In JMP, the author chooses **Bubble Plot** from **Graph**, and then places the variables into the proper boxes as shown above. Potentially the Bubble Plot can display seven dimensions, but it would be very confusing. Thus, only five variables are selected at once for this illustration.

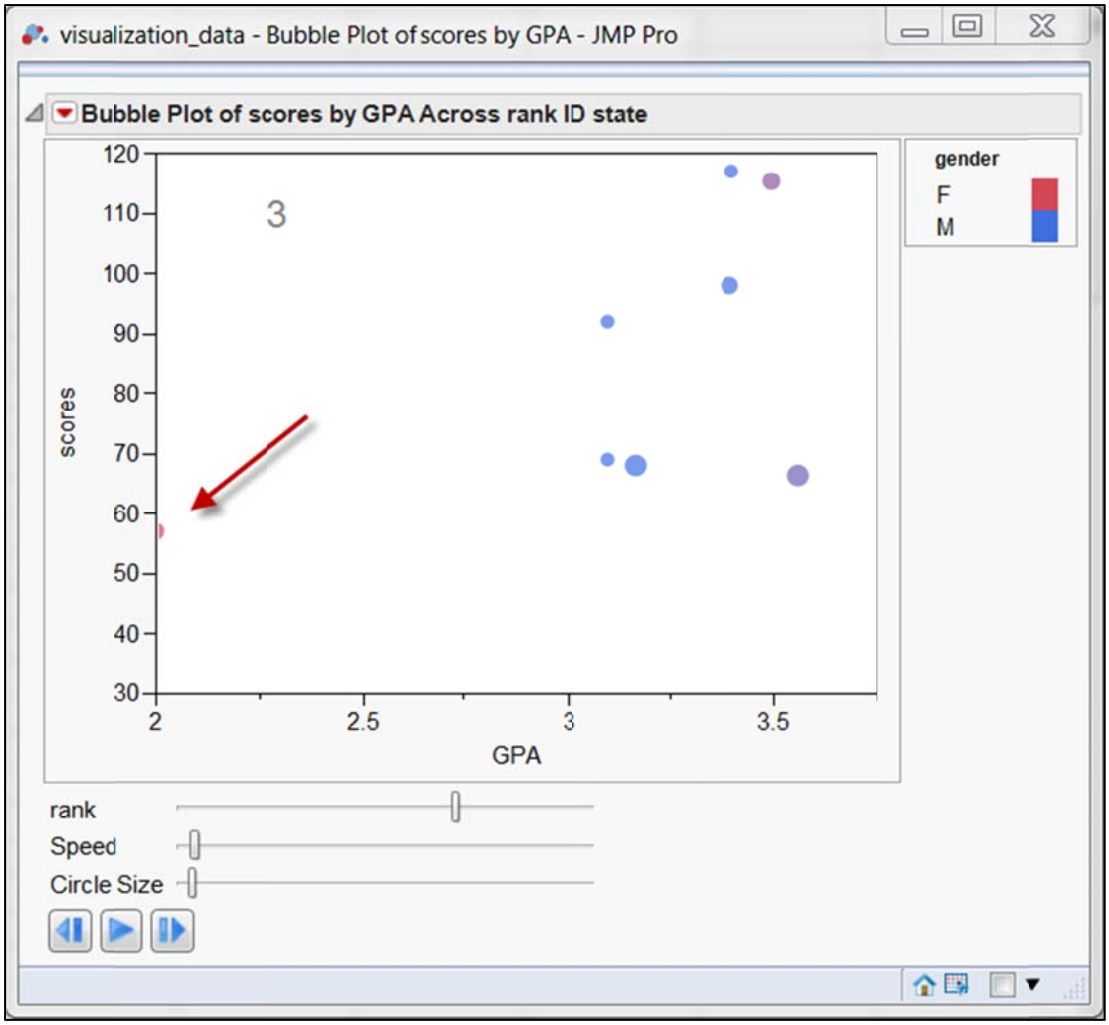


A panel pops up after the OK button is pressed. By default JMP plays the movie in a loop. But it is better to stop the movie and examine the graph frame by frame. In the first frame the number “1” can be seen at the upper left corner. In this example the temporal dimension (time) is “rank” and thus the first frame of the movie shows the relationship between GPA and scores of freshmen.

In the legend females are depicted as red whereas men are portrayed as blue. But in the panel we can see only pink or purple dots rather than red and blue. Why? The eight dots represent the eight home states of the students. When there are more girls than boys in that state, the dot tends to be pink or magenta. Conversely, when men outnumber women, the point turns into purple or cyan. The different sizes of the dots express the sample sizes. Needless to say, a bigger circle means a larger sample size. Even though there is no regression line, the scatterplot suggests that there is a positive and linear association between GPA and scores, and the gender composition is fairly even.

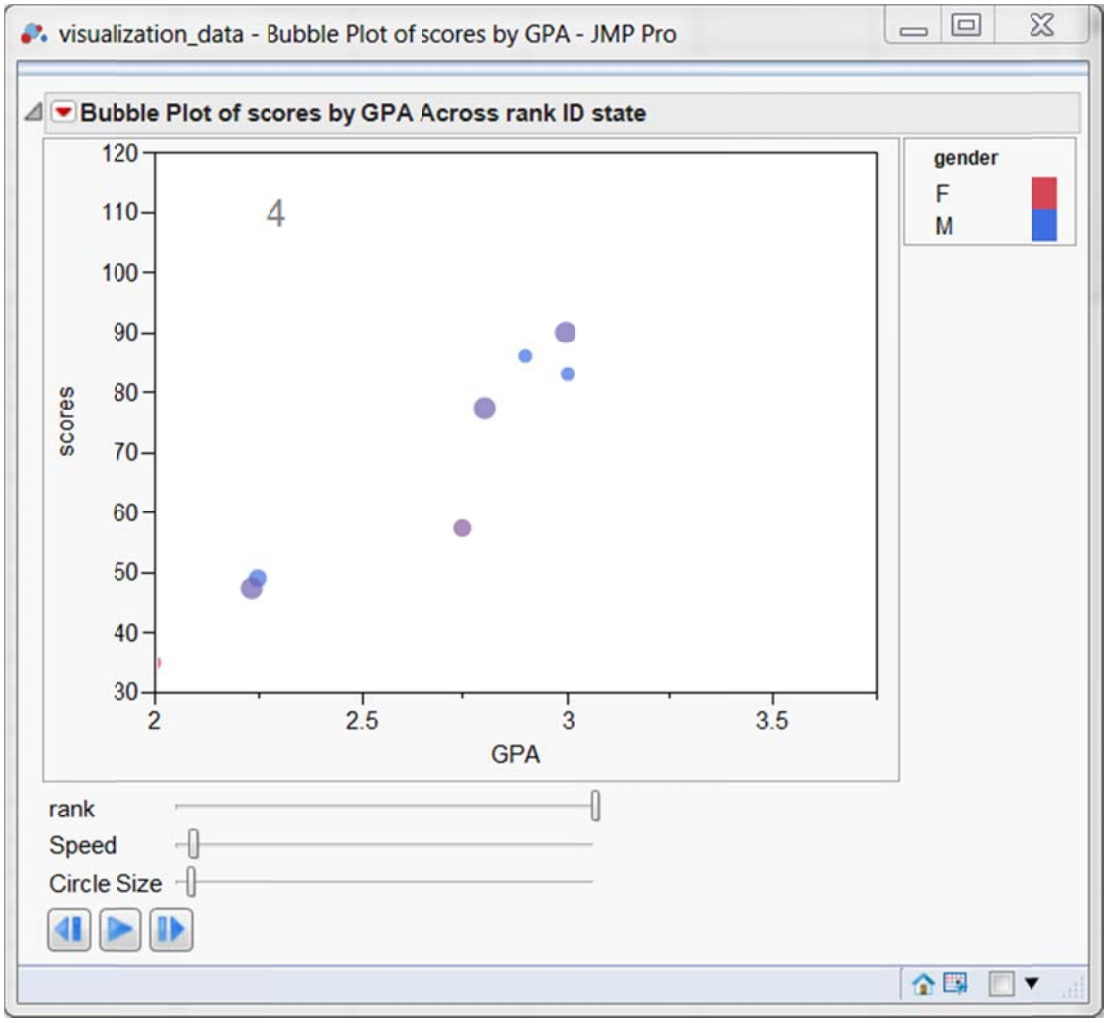


However, if you press the forward button to advance the movie to the next level (2=sophomore), a different story emerges. Suddenly the linear trend disappears, meaning that high school GPA is no longer tied to college test performance. One possible explanation is that Freshmen carry over their work ethics while facing new challenges in their first year. But after they adapted themselves into the environment, they sit back and relax during the second year.



When you go forward to view the junior data, another interesting phenomenon appears. If the outlier is included (see the red arrow), a positive regression line can be fitted into the data. But if the outlier is excluded, it is obvious that there is no significant association between high school GPA and college performance.

There are two more note-worthy points. First, the data points concentrate around high GPA (>3). Second, the color of the observations shifts to the bottom of the spectrum (blue), meaning that there are more boys than girls in the junior level.



Last but not least, in the senior level, the significant relationship between GPA and test performance comes back. One plausible explanation is that seniors are facing the pressure of job-hunting and applying for graduate study, and thus they restore their work ethics. Like the last panel, in this frame the color spectrum is closer to the blue end, implying that the majority of the sample consists of men.

The above illustration shows that the researcher can obtain insightful findings about the data without running a single statistical test. The Bubble Plot is just one of many visualization tools for multi-dimensional data. When you start your own exploration, you will be surprised.