Change mindset: Review the data based on visualization

Author Siye Liu, Roche (China) Holding Ltd., Shanghai, China

ABSTRACT

Data visualization is already a mainstream in many different industries, and we do benefit from the results that the companies achieved with the approach. So, are you still using the traditional way in your daily work? Or change mindset and implementing data visualization to help yourself and support other functions in your team? In this paper, by using SAS and other visualization tools, for example SPOTFIRE, excel. To generate a few examples presenting the value of using visualization to review data compare to the traditional way.

INTRODUCTION

There are many data visualization tools such as: R, SAS, Spotfire, Microsoft Excel, Google Spreadsheets, Tableau Software. Some of them are very easy to use, some of them may need some programming skills. This paper using SAS and Spotfire to generate a few examples presenting the value of visualization in our daily work.
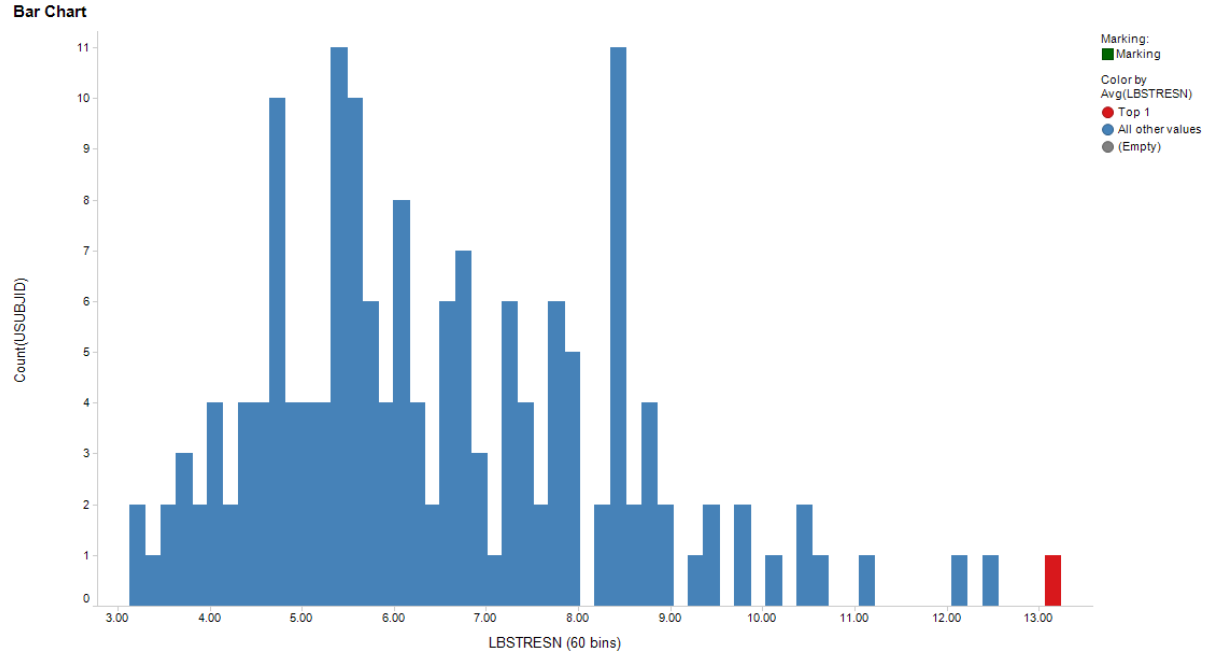
BAR CHART

Let's start from a simple example. There is a request ask to check if the baseline WBC results meet the inclusion criteria (<=13*10^9/L) of protocol or not. Usually we can create a listing to output the baseline lab result by using below SAS code:

```
DATA _temp;
    SET sdtm.lb;
    LENGTH MSG $50.;
    IF visit='BASELINE' and LBSTRESN>13 then MSG='Out of IE range';
RUN;
```

We got a data dump listing of Lab result and potential data issue if MSG column is not blank. It is very straightforward and efficient. Next time if the criteria is changed, we just need to do a minor change in the code and generate another listing. OR we can do a good summary of the lab results by using a Bar Chart.

Using SPOTFIRE as example:

.

Bar Chart

In above bar chart, the value of lab result has been selected as category axis, you can see there are 60 bins has been derived from these values, this means the values have been group into 60 bins in order to form 60 bars, and the value axis has been configured as the count of SUBJECTID in each of these bins. Therefore, we can see how many subjects has the low value of WBC and how many subjects has the high value of WBC. Hopefully, this histogram immediately shows that the distribution of the value of WBC, most of the values are in the range (from 4-9) and a few of the values are far from this range, especially the top one value is greater than 13 which is the potential data issues that we are looking for.
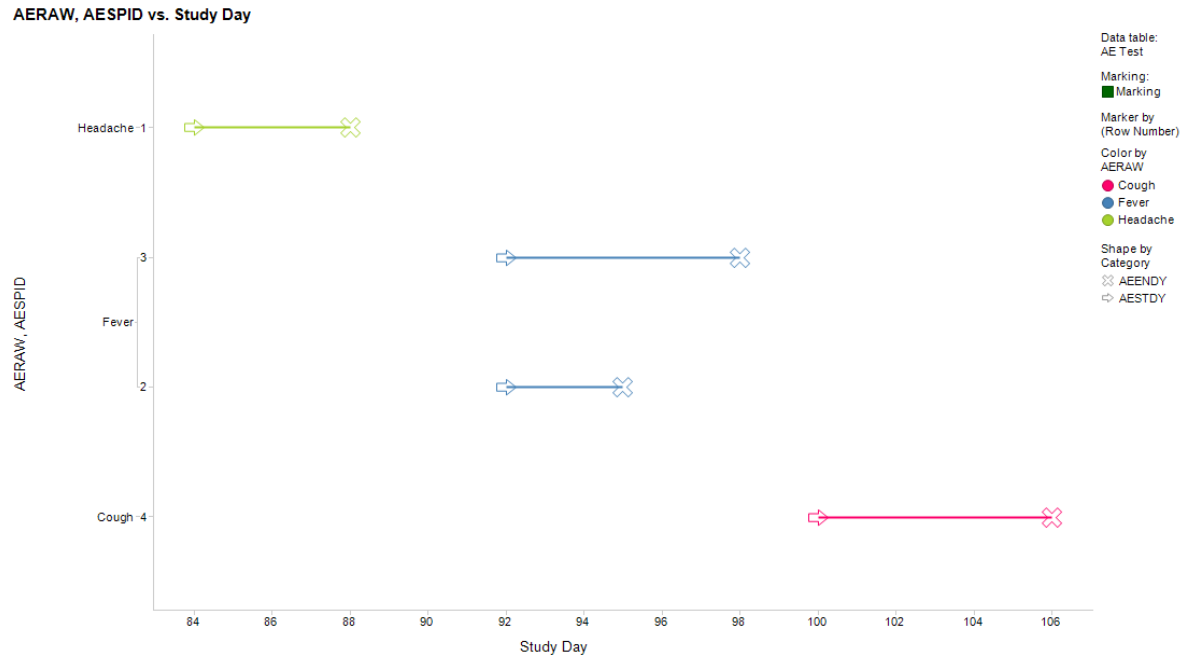
SCATTER PLOT

The first example shows that a visualization can tell the "story" of the data. It was generated from the listing, however, it is more clear to present the data from the summary point of view. Another example is about focus on the details. During the study conduct phase, there is a request, ask to check if there is any overlapping data entered in AE table. Usually we can create a listing to output the overlapping records in AE table by using PROC SQL in SAS.

```
DATA _temp;
    SET sdtm.ae;
RUN;

CREATE TABLE _merge AS
SELECT
    A.SUBJECTID,
    A.AETERM,
    A.AESTDTC,
    A.AEENDTC
FROM _temp AS A
LEFT JOIN sdtm.ae AS B
ON  A.SUBJECTID=B.SUBJECTID
AND A.AETERM=B.AETERM
AND A.AESEQ^=B.AESEQ
AND A.AESTDTC<=B.AEENDTC
AND A.AEENDTC>=B.AESTDTC
;
QUIT;
```

We got a data listing of AE if any AE term are same and sequence number are different and dates are overlapping. Well, by using the visualization tools, we can also create a detail review scatter plot per subject.

Using SPOTFIRE as example:.

In above scatter plot, the study day derived from AESTDY and AEENDY has been selected as category axis, you can see this axis mode is continuous days present the start date and end date of each AE. The value axis has been configured as AE term and separated by AE number. There is a connection line was draw separately per value in AE number, different colors were used per AE term and different sharps to indicate the start day and end day. So it is very clear that there are 4 AEs occurred, headache start on day 84 and end on day 88; Cough start on day 100 and end on day 106, there are two Fever occurred on day 92, and there are several overlapping days which might be the potential data issue.

BUBBLE CHART

As above two examples show that the visualization can present data both for summary review and details review, data listing also can support such kind of needs, however, if we want to see the trends of the data, we'd better choose the visualizations.

Using SPOTFIRE as example:
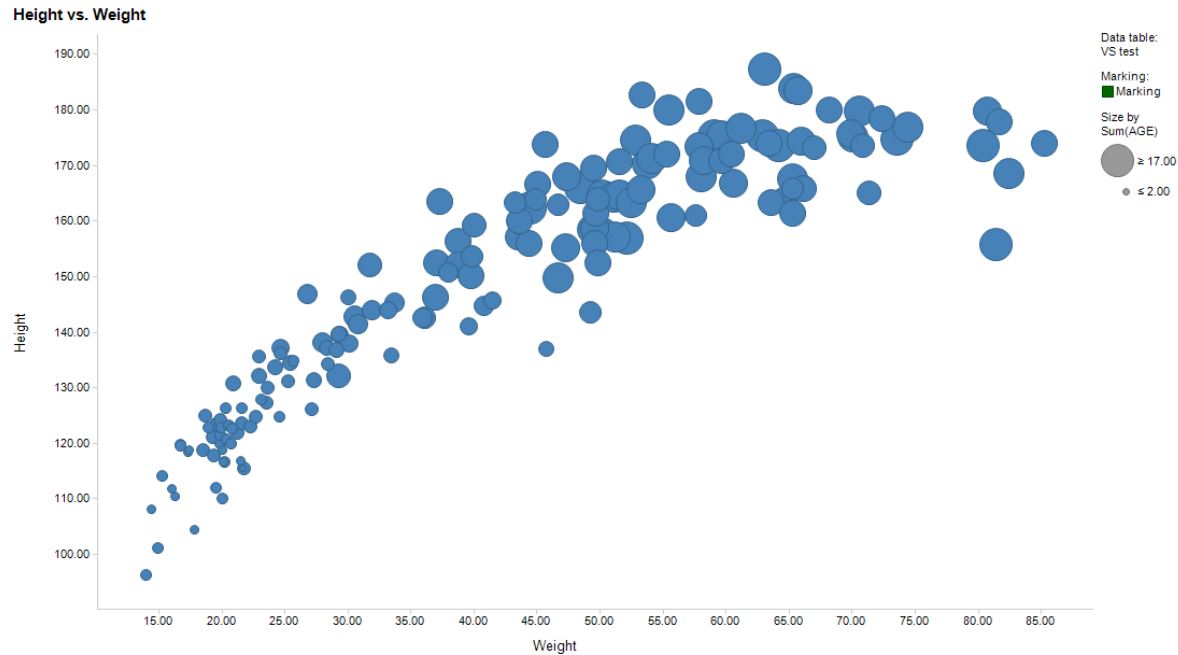
Height vs. Weight

In this bubble Chart, the average of each subject's Weight was selected as x-axis and average of each subject's Height was selected as y-axis. The size of the bubble was configured as Age. So from this chart we can see the big size bubbles were in the right upper corner of the axis and the small bubbles were in the left lower corner, which shows that trends of subject's age, weight and height. If any bubbles were far away from the line of the trends, then we might consider it was a potential data issue.

CONCLUSION

Data visualization is more and more popular among different industry; we can see very fancy graphs every day from news, reports, social media, TV. A lot of software helps us to generate the visualization very easily, However the most important thing is how we can using the visualization to tell the "story" of the data. The more we dig the data and the more we knew the data will helps us to interpreted the data and also helps us to do the predication.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Siye Liu
Enterprise: Roche (China) Holding Ltd.,
Address: Long dong ave 1100. Building 11.
City, State ZIP: 201203
Work Phone: +86 21 2892 2101
E-mail: siye.liu@roche.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.