Macquarie University

Graphics, Multivariate Methods and Data Mining

SGTA Exercises

Week 3

Question 1

The data consists of three variables:

- *Short*: the number of short suspensions
- *Long*: the number of long suspensions
- *Gender*: whether the student identifies as male or female

The *Gender* variable contains binary values where 1 = male and 2 = female to denote the variable being in a binary state i.e., a student can only be either male or female. As digital computers cannot understand words the way humans do, employing binary values allows computers to read and process data in the same manner.

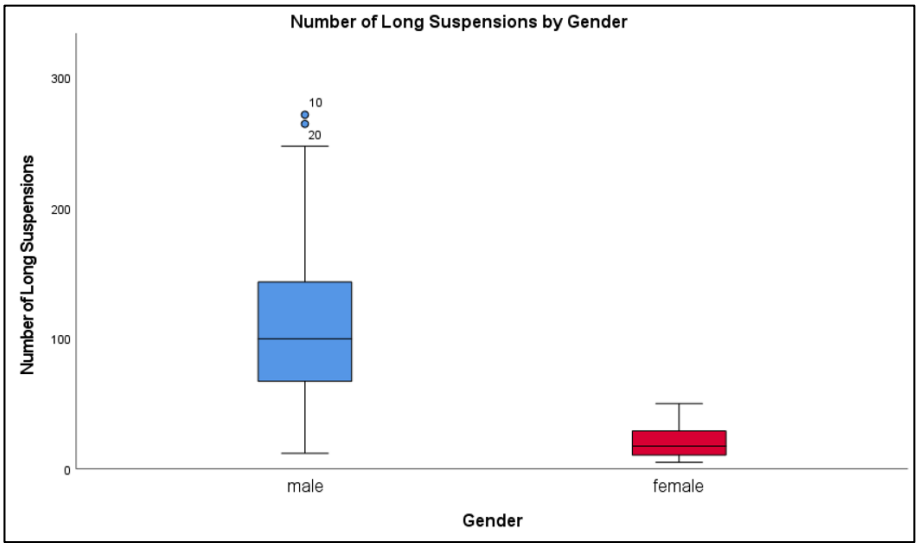
The descriptive statistics for the number of long suspensions, regardless of gender is given:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
long	80	5	271	67.45	67.173
Valid N (listwise)	80				

From the descriptive statistics, we can see that out of 80 students, the minimum number of long suspensions held by a student is 5 and the maximum is 271. On average, students have held 67.45 long suspensions and the standard deviation is 67.173.

By generating a comparative boxplot that assesses the number of long suspensions held by males and females, the following output is produced:



To support the below conclusions, the descriptive statistics for the number of long suspensions, for males and females separately, are obtained:

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
long	40	12	271	114.43	66.814
Valid N (listwise)	40				

a. gender = male

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
long	40	5	50	20.48	12.237
Valid N (listwise)	40				

a. gender = female

From the boxplot, the following conclusions can be made:

1) **Median**: Male students have a higher median of approximately 100 long suspensions compared to female students, which have a median of roughly 20 long suspensions. This suggests that females are less likely to have a large number of long suspensions, as opposed to male students.
    a. The median of the male class reveals that the group has a slightly asymmetric distribution, skewed towards the higher range of suspensions.
    b. The median of the female class is also asymmetrically distributed, conveying a skewed distribution towards the upper quartiles.
2) **Range**: Male students have a greater spread of 259 suspensions, indicating a higher variation in the number of long suspensions, compared to female students, at 45 suspensions. This is supported by the higher standard deviation of 66.814 in male students, compared to 12.237 in female students.
    a. From this, male students also have a greater interquartile range of approximately 80 suspensions compared to females, at roughly 20 suspensions.
    b. Both males and females have greater upper whiskers i.e., "longer tails" indicating a skewed distribution.
3) **Outliers**: Male students have two outliers i.e., two males with a larger number of long suspensions that is outside the range, while female students do not have any outliers.

Question 2

1) Let:

$$\boldsymbol{X} = [\mathbf{0} \quad \mathbf{3} \quad \mathbf{5} \quad \mathbf{6}]$$

$$\boldsymbol{B} = \begin{bmatrix} \mathbf{0} & \mathbf{3} & \mathbf{1} \\ \mathbf{-1} & \mathbf{2} & \mathbf{-2} \\ \mathbf{2} & \mathbf{0} & \mathbf{4} \\ \mathbf{0} & \mathbf{4} & \mathbf{2} \end{bmatrix}$$

I. Write down the transpose vector X^T of X and the transpose matrix B^T of B.

$$X^T = \begin{bmatrix} 0 \\ 3 \\ 5 \\ 6 \end{bmatrix}$$

$$B^T = \begin{bmatrix} 0 & -1 & 2 & 0 \\ 3 & 2 & 0 & 4 \\ 1 & -2 & 4 & 2 \end{bmatrix}$$

II. Write down $X \times B$

$$X \times B = [0 \quad 3 \quad 5 \quad 6] \times \begin{bmatrix} 0 & 3 & 1 \\ -1 & 2 & -2 \\ 2 & 0 & 4 \\ 0 & 4 & 2 \end{bmatrix}$$

$$= [0 \times 0 + 3 \times (-1) + 5 \times 2 + 6 \times 0 \quad 0 \times 3 + 3 \times 2 + 5 \times 0 + 6 \times 4 \quad 0 \times 1 + 3 \times (-2) + 5 \times 4 + 6 \times 2]$$

$$= [-7 \quad 30 \quad 26]$$

2) Let:

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{1} & \mathbf{4} \\ \mathbf{3} & \mathbf{2} \end{bmatrix}, \qquad \boldsymbol{C} = \begin{bmatrix} \mathbf{0} & \mathbf{6} \\ \mathbf{1} & \mathbf{7} \end{bmatrix}, \qquad \boldsymbol{Y} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}$$

I. $\boldsymbol{A + C}$

$$A + C = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 6 \\ 1 & 7 \end{bmatrix}$$

$$= \begin{bmatrix} 1+0 & 4+6 \\ 3+1 & 2+7 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 10 \\ 4 & 9 \end{bmatrix}$$

II. $\boldsymbol{A - C}$

$$A + C = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 6 \\ 1 & 7 \end{bmatrix}$$

$$= \begin{bmatrix} 1-0 & 4-6 \\ 3-1 & 2-7 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & -2 \\ 2 & -5 \end{bmatrix}$$

III. $\boldsymbol{A \times C}$

$$A \times C = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix} \times \begin{bmatrix} 0 & 6 \\ 1 & 7 \end{bmatrix}$$
$$= \begin{bmatrix} 1 \times 0 + 4 \times 1 & 1 \times 6 + 4 \times 7 \\ 3 \times 0 + 2 \times 1 & 3 \times 6 + 2 \times 7 \end{bmatrix}$$
$$= \begin{bmatrix} 4 & 34 \\ 2 & 32 \end{bmatrix}$$

IV. $\boldsymbol{C \times Y}$

$$C \times Y = \begin{bmatrix} 0 & 6 \\ 1 & 7 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 \times 0 + 6 \times 1 \\ 1 \times 0 + 7 \times 1 \end{bmatrix}$$
$$= \begin{bmatrix} 6 \\ 7 \end{bmatrix}$$

V. $\boldsymbol{Y^T \times Y}$

$$Y^T \times Y = \begin{bmatrix} 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
$$= [0 \times 0 + 1 \times 1]$$
$$= 1$$

VI. $\boldsymbol{Y \times Y^T}$

$$Y \times Y^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} 0 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 \times 0 & 0 \times 1 \\ 1 \times 0 & 1 \times 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Question 3

a) What is the subject of Minard's graph?
The subject is Napoleon's 1812 campaign into Russia

b) How many different variables are shown in the graph?
There are six different variables displayed in two dimensions in Minard's graph:

1. The number of Napoleon's troops
2. The distance travelled
3. The temperature
4. The latitude and longitude
5. The direction of travel
6. The location relative to specific dates

c) What features of the graph are good, and what features are not so good?
The features of the graph that is good is that the graph is highly detailed i.e., so many variables are communicated in just a single graph, conveying a huge amount of information. The graph also explicitly shows the relationship between each of the variables. A certain feature that is not so good is the text in the graph and its readability. It is also unclear if the graph pertaining to its geography is accurately scaled to size.

d) What method was used to prepare it?
It is likely that the graph was hand drawn and was prepared in a 'layered' manner.

e) What 'message' does the graph send?
The graph not only shows the conditions these soldiers were in, but the significant loss of approximately 342,000 troops during the campaign. The graph essentially conveys the 'cost' of the campaign.

f) Why do you think the graph got such a good rating from Tufte?
Minard's graph tells a story, it conveys a large amount of information in just a single graph, the graph is overly detailed, and every aspect of the graph is highly relevant. It also shows the interconnectedness of the six different variables and its relationship in the loss of troops.

g) Give your opinion in a sentence or two (e.g., positive and negative) of the 3D version of Minard's map

Assuming the graph is a more modern version of Minard's graph, it's much clearer, allows the user to zoom in and it is easy to read and understand. It is also nice to see each 'layer' individually. The disadvantage of using this graph is that not every user will know how to read a 3D graph, and thus it may be difficult to comprehend.