Data Visualisation Report (SET09120)

Valeri Vladimirov

School of Computing, Edinburgh Napier University, Edinburgh

40399682@napier.live.ac.uk

1 Introduction

This coursework report is reviewing a dataset about participants taking part in a throwing sport. There are two outliers and seven relationships hidden in the data, which must be found, visualised and explained. In addition to that someone else's graph must be evaluated and a better solution must be drawn. Data consists of 20000 measurements, which include the skill group, gender, age, wind direction, height, distance, angle, offset, score 1 and score 2 of the participants. This information will be visualised and analysed using ggplot2, RStudio and R language.

2 Outliers

In this section I am going to discuss the outliers I have found, why I think they are outliers and why I have removed one of them.

2.1 Outlier in the height data

Histograms are a great way for finding a big number of outliers. After creating a histogram for the height data (see Fig. 1) we can see some continuous data representing the height of each participants throw and the count of participants which have achieved relatively similar height. On the left we can see that the count drops to 0 at around 82 metres of height, which would usually mean that this is the end of the continuous data, however there is another small peak on the right. This can be classified as an outlier, because if things have dropped to the minimum its very odd to have another peak and it needs to be investigated. As it can be seen on the first histogram (see Fig. 1) there is a lot of space, which is not showing any data and it leads to poor density. This is because of the outlier on the right, which is skewing our data, therefore I have removed it in the second histogram (see Fig. 3), in order to improve the data density of the histogram.

2.2 Outlier in the offset data

Boxplots are a good way for showing small numbers of outliers as in the current situation (see Fig. 2). For the very experienced participants, who have thrown East, North and West, outliers have been found. Since they are a small number, they have been shown as dots in a boxplot. These outliers do not lead to poor density, meaning that they will not be removed.

3 Relationships

This section is about the relationships I have found within the data and information about them.

3.1 First relationship

According to Fig. 4, which is a boxplot showing what offset participants have based on their gender and skill group, there is a relationship between the offset for the various genders and skill groups. It can be seen that male participants have a very consistent offset for their throw in all the skill groups. Whereas female

participants have had a different offset for each skill group, meaning that they have been inconsistent. This is the code for the pattern.

```
ggplot(data, aes(x=Gender, y=Offset, fill=Group)) + geom_boxplot()
```

3.2 Second relationship

Fig. 5 is another boxplot showing what score.1 have participants achieved based on their skill group. As it can be seen participants from the very experienced group have achieved the highest score.1, compared to novice, experienced and professional participants, which have relatively similar and lower score.1. Here is the code used for the pattern.

```
ggplot(data, aes(x=Group, y=Score.1, fill=Group)) + geom_boxplot()
```

3.3 Third relationship

Fig. 6 is showing that almost all participants have had nearly the same throw height based on the wind direction, meaning that they have been consistent. However, this cannot be said for professional male participants, which have had a different height when throwing North and South. They have thrown higher when throwing North and lower when throwing South. Also, it can be said that professional male participants have had a wider variety for the height of their throw than others. In the chart the height outlier can also be seen because the image has been taken before removing it. Here is the code used for the relationship.

```
ggplot(data) + aes(x = Group, y = Height, fill=Wind.Direction) + geom_boxplot() + facet_wrap(~Gender)
```

3.4 Fourth relationship

Fig. 7 is a chart showing what score 2 have participants achieved based on their age and the wind direction. It can be seen that people who have thrown North and South have achieved various kinds of results no matter the age, which is quite chaotic. However, data for participants throwing East and West is ordered. People who have thrown East, their score 2 increases based on their age, meaning that older participants have achieved a higher score 2, compared to younger ones. Opposed to that, those who have thrown West, their score 2 decreases based on their age, having younger participants with a higher score 2 than older ones. Here is the code used for the pattern.

```
ggplot(data) + aes(Age, Score.2, color=Wind.Direction) + geom_point()
```

3.5 Fifth relationship

According to Fig. 8, which is a chart showing the distance of a throw based on the angle, there is a relationship between these two measurements, which is shown by a perfect curve. Participants who have thrown the farthest, which is near 90 metres, have had an angle of 45 degrees. Everyone else who has thrown less than 90 metres has a smaller or bigger angle. Here is the code for the pattern.

```
ggplot(data, aes(x = Angle, y=Distance)) + geom_point()
```

4 Visualisation Evaluation

In this section I will discuss the example visualisation (see Fig. 9) provided and how I have improved upon it with my sketch (see Fig. 10). I have decided to use a bar chart, because it is a good chart when it comes to comparative data, as the height of something.

4.1 Reducing chartjunk

As it can be seen in the example visualisation (see Fig. 9) the person creating this chart has used icons of people instead of normal bars, which leads to chartjunk, because it doesn’t add up anything to the data.

Also, the icons are overlapping and there is no space between them. For example, Australia's icon is hard to see and Latvia's icon goes out of bonds of the y axis. I have replaced the icons with bars in my sketch (see Fig. 10) and added space between the bars, which improves on these bad aspects of the example chart.

4.2 Poor data-ink ratio

The title colour on the example visualisation (see Fig. 9) is one of the things that contributes to the poor data-ink ratio of the chart, because ink should be mainly used for the data. Another bad aspect is how colour has been used on the icons. For example, Latvia and South Africa have the same purple colour used on their icons and a person can think that there is something similar between these elements. The human eye can build a relationship between items that have the same colour, which is called law of similarity. Both these bad aspects have been improved on my sketch (see Fig. 10), imagining the bars on it are coloured in black or grey, which I did not have. On the other hand, one good aspect when talking about data-ink ratio is that the background is not coloured.

4.3 Order

Another good aspect of the example chart (see Fig. 9) is that the bars are ordered based on the height. Starting from the highest females in Latvia to the shortest in India. And it can be said that data is ordinal if we are considering the y axis.

5 Appendix

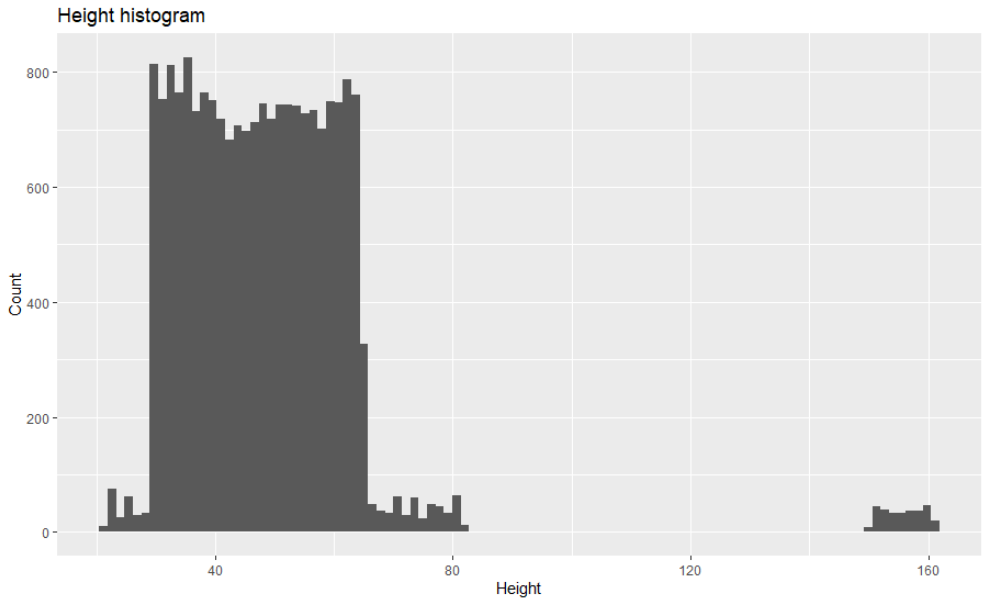

Fig. 1. A histogram showing an outlier in the height data.

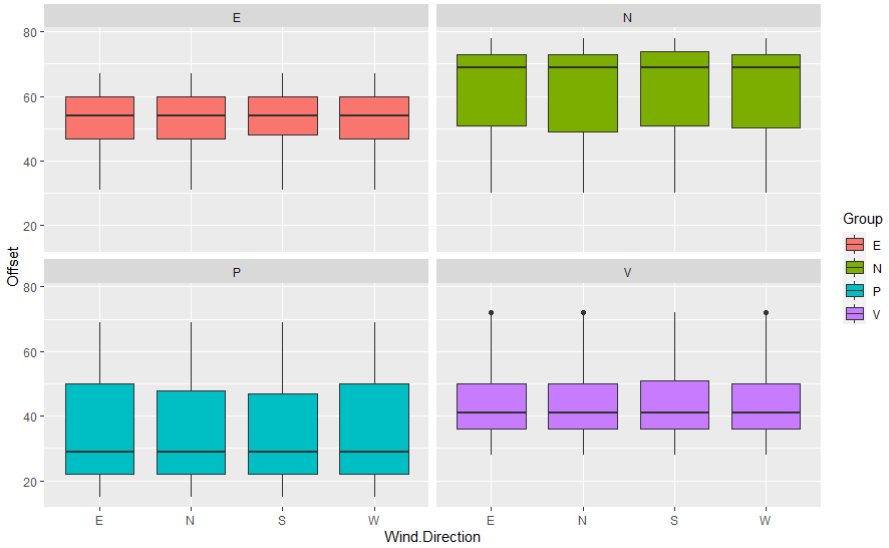

Fig. 2. A boxplot with outliers for the "Very Experienced" participants.

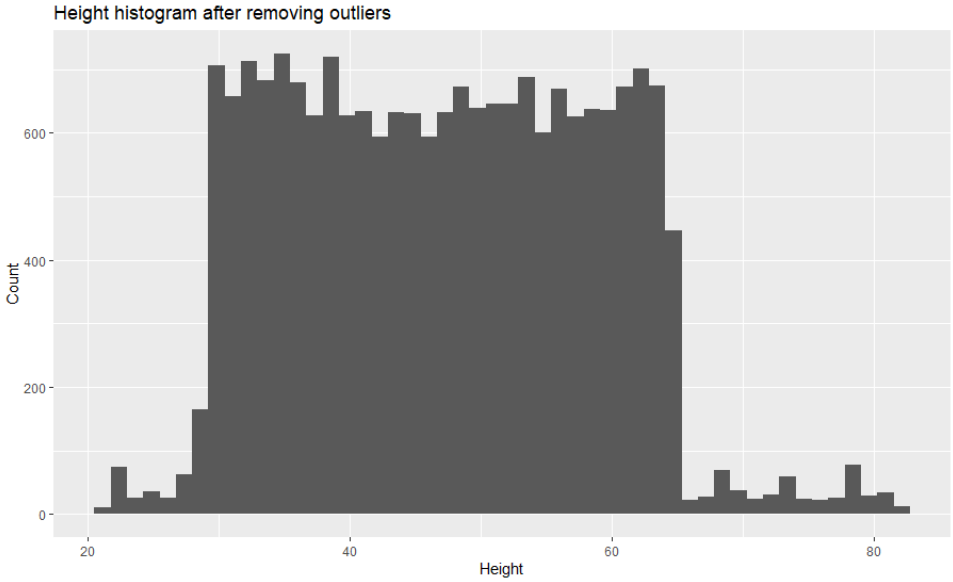

Fig. 3. The height histogram after removing the outliers.

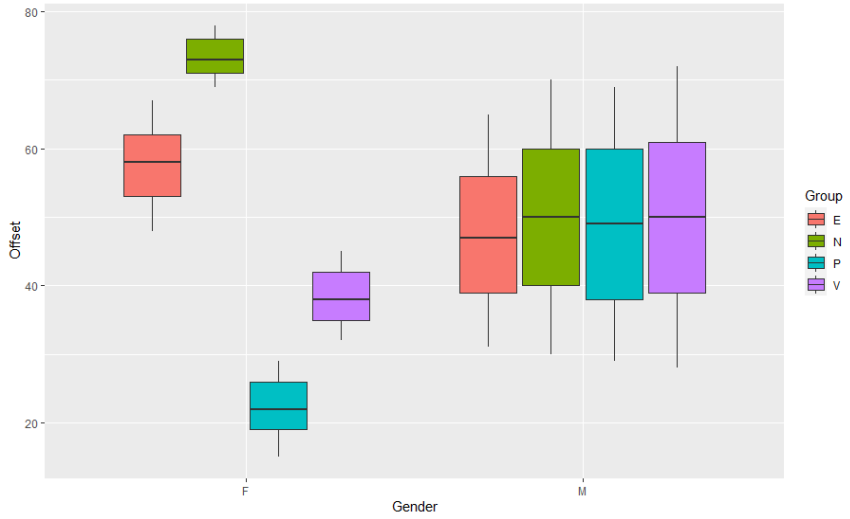

Fig. 4. Showing that females have an inconsistent offset compared to male participants.

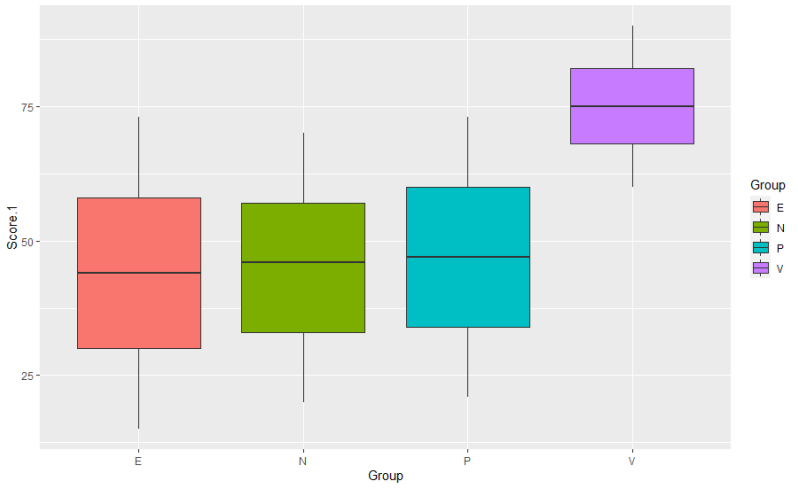

Fig. 5. Showing that very experienced participants have a higher score.1 than the other skill groups.

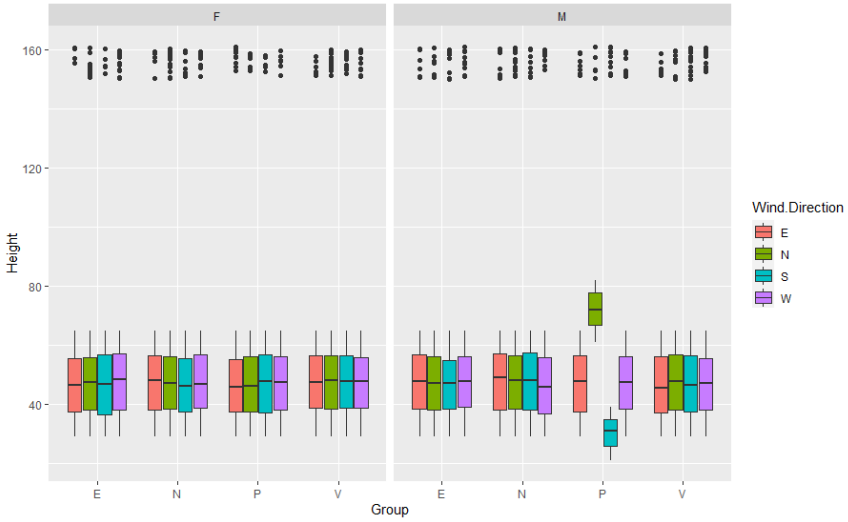

Fig. 6. Showing that professional male participants have had a different height when throwing based on the wind direction.

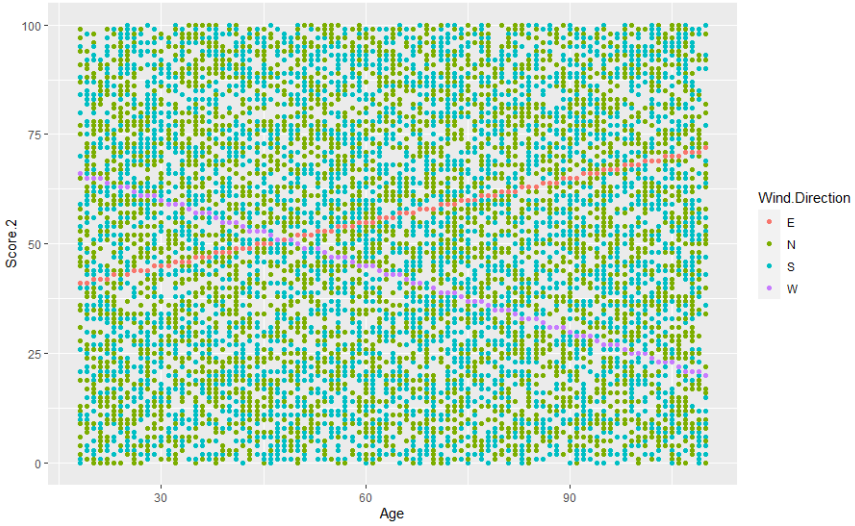

Fig. 7. Showing what score 2 have participants achieved based on their age and wind direction when throwing.

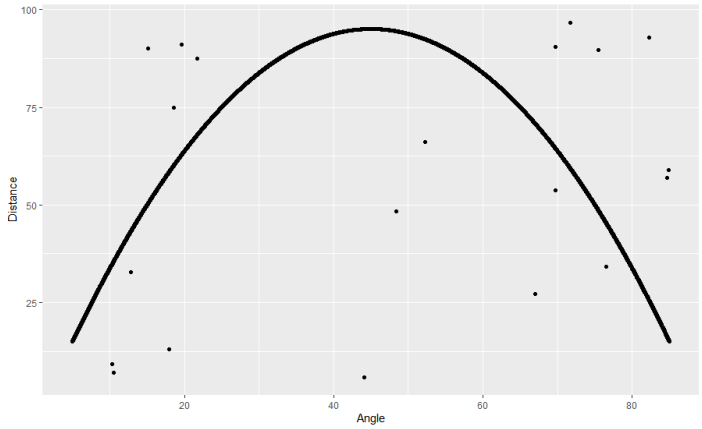

Fig. 8. Showing that the farthest throw distance has been achieved when the angle has been 45 degrees.

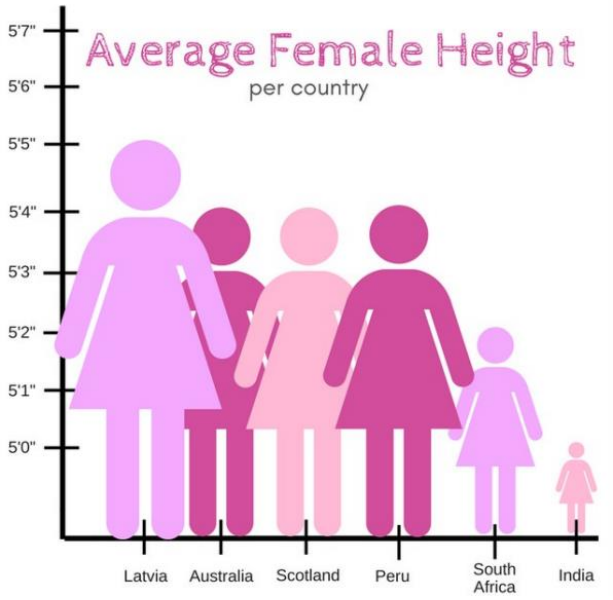

Fig. 9. Example visualisation, which must be discussed and sketched in a better way.

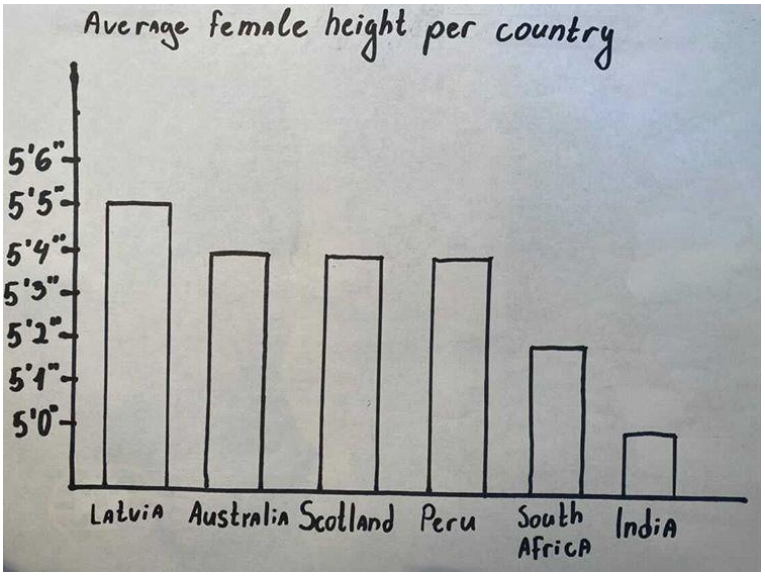

Fig. 10. Sketch of the example visualisation after fixing the bad aspects.