Data Analytics Coursework 1
Data Exploration

Stefan Hristov
40284739

Edinburgh Napier University, Edinburgh, UK
40284739@live.napier.ac.uk

Abstract. The aim of this coursework was to use visualization to explore a data set and extrapolate seven relationships. The data set consisted of three categorical variables and five discreet numerical variables. The categorical variables were age, location and gender. The discrete numerical variables were the score for four types of sport (sprint, distance, jump and throw) and an overall score. The plotting was done with the help of Rstudio and ggplot library. From the seven relationships found, three were visualized using a box-plot while the remaining four were visualized with a scatter-plot. For the fifth relationship, the help of a regression line and a regression coefficient was used to demonstrate the correlation. The first relationship showed sportsman from location D to have higher performance in the overall score. The second relationship demonstrated participants with age over thirty to perform better at distance score. The third relationship suggested that females jump score increases with age. The fourth relationship described a negative linear correlation between sprint score and throws score for athletes from locations A and C. The fifth relationships demonstrated that the throws score had no impact on the overall score for females from locations B, D and E. The sixth relationship described a quadratic-like power relation between distance score and jumps score for males from location B. The seventh relationship shows a positive linear correlation between sprint score and jumps score for teenage females from location E.

Keywords: Rstudio · Coursework · GGPLOT · Data Relationship · Visualization.

1 Description of the relationships

First Approach The first approach for finding relationships was to plot one numerical variable on the Y-axis and one categorical variable on the X-axis. To avoid cluttering of data, a box plot was used. That way tendencies and patterns can be noticed through the score distribution, rather than looking at all entries. As literature suggests, distributions are best displayed via the use of box plots or histograms [2]. Further, it is suggested that optimal representation for a range of values, a box-plot is the best choice [1]. To make sure the data is thoroughly examined, when a categorical variable is used, the other two categorical variables from the data-set were used to colour and separate the plot.

1.1 Relationship One

While plotting location vs overall score, location D stood out with having its 25th percentile higher than the 75th percentile of the rest of the locations. There were no signs of impact from age or gender. Therefore it was concluded the relationship is entirely based on location and overall score (see Fig. 1).

1.2 Relationship Two

With age versus distance, participants with age over thirty displayed better performance. Their lowest scores were, almost completely across the board, higher than the 75th percentile from the rest of the age groups. For all genders and locations, the athletes showed similar results (see Fig. 2).

1.3 Relationship Three

With age against jumps, the plot suggested a positive correlation for only females. The relationship was not affected by location (see Fig. 7).

Second Approach After all combinations of categorical variables were run against a single numerical variable, two numerical variables were used. As some literature suggests, two numerical variables are best represented by a point plot [1]. Furthermore, the literature suggests that an effective method for finding a correlation is by using point (scatter) plot combined with a regression line [1, 2]. For each combination of variables, the data was subset-ted for females and males. That way the amount of data to be displayed was reduced. Further, the plots were coloured by location and separated via facet_wrap by age.

1.4 Relationship Four

With sprints on the X-axis and throws on the Y-axis, there was a perfect negative linear correlation between athletes from locations A and C. Since gender and age had no effect, the plot was segregated by location rather than age (see Fig. 3).

1.5 Relationship Five

The fifth discovered relationship was between throws score and overall score. It only applied to females from locations B, D and E. It seemed that females overall score was severely unaffected by throws score. In location A and C the line of best fit was almost flat with a low regression coefficient for both genders. That can be explained by the previous relationship since it is almost a perfect negative linear correlation involving throws and sprints. However, for locations B, D and E the line of best fit for females continued to be extremely close to horizontal with coefficients below 0.01. That is unusual since increased performance in one of the sports should contribute to a higher overall score. Age had no significant effect (see Fig. 4).

1.6 Relationship Six

The next relationship found was between distance score and jumps score for only males from location B. The plot looked like a quadratic function but it could have been another power function. Age had no impact on the shape of the curve, it was only a factor in the distribution. That can be explained by one of the previous relationships which involve distance score and age (see Fig. 5).

1.7 Relationship seven

The last relationship found was a positive linear correlation between sprints score and jumps score for females from location E from age group T (see Fig. 6).

2 Most interesting relationship

The third relationship was chosen to be optimized. The X-axis was rearranged so the age groups are in ascending order. This helps with the natural feeling of age progression from left to right. The labels on both axes were re-written. That contributes to a quicker understanding of the plot. A title was not included and the legend was removed. Since there was a legend to the graph and the axis contains notation for the age, both the title and the legend are wasted data-ink [3] and also helps increase the data density [3]. The colour was changed to draw attention to the data [2]. Since there was no direct correlation between age and colour, a tableau with non-bright colours was chosen as optimal [2, 4]. The background colour was changed to white and the grid lines were removed. That helps the colours to be more vivid [2].

References

1. Few, S.: Effectively Communicating Numbers, Selecting the Best Means and Manner of Display. Principal, Perceptual Edge, November 2005
2. Visual Analysis Guidebook, http://www.dataplusscience.com/files/visual-analysis-guidebook.pdf Last accessed 24 Oct 2018
3. The semiology of graphics, Landscape Representation II, March 2015
4. Tableau custom colours, http://www.onlinehelp.tableau.com/current/pro/desktop/en-us/formatting_create_custom_colors.html. Last accessed 24 Oct 2018

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

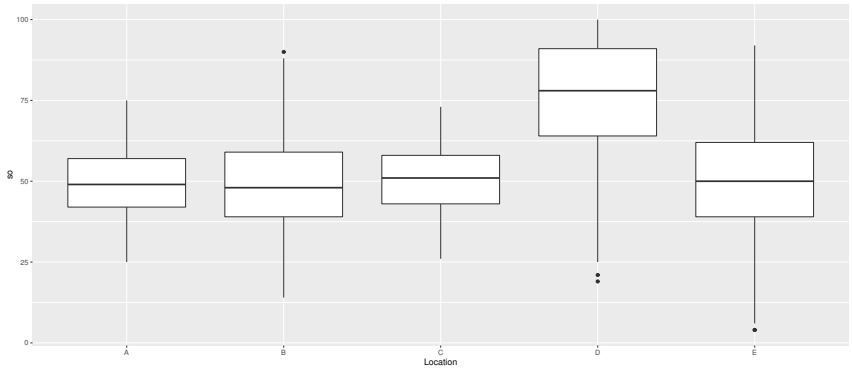

Fig. 1. Location vs Overall Score

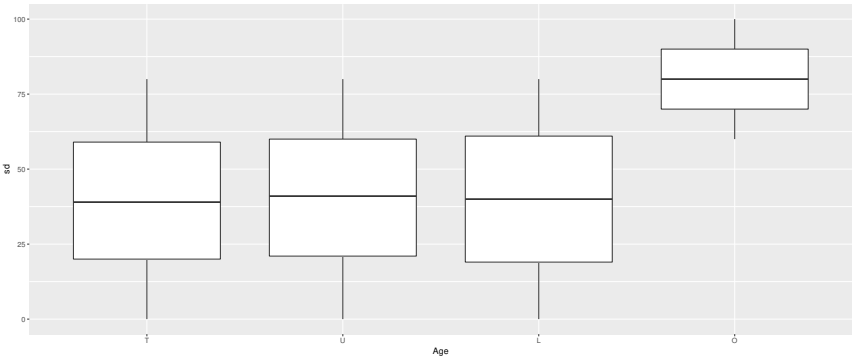

Fig. 2. Age vs Distance Score

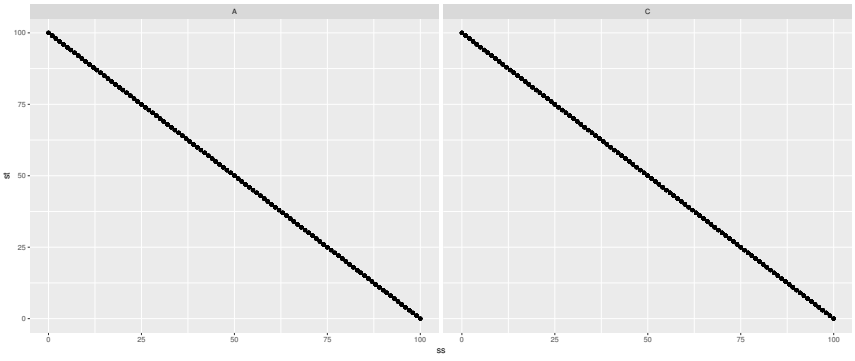

Fig. 3. Sprint Score vs Throws Score for locations A and C

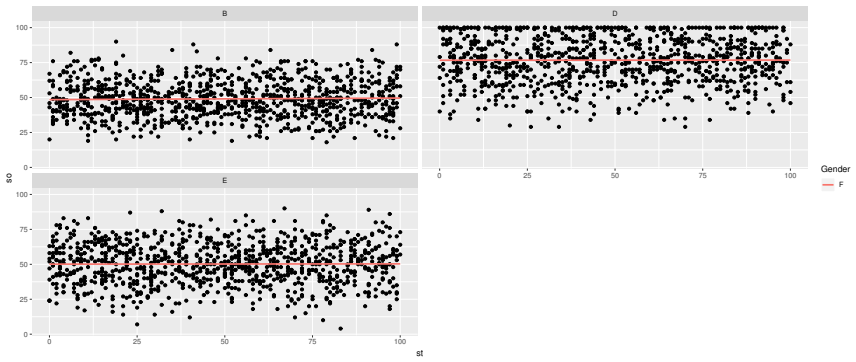

Fig. 4. Throws Score vs Overall Score for females from locations B, D and E

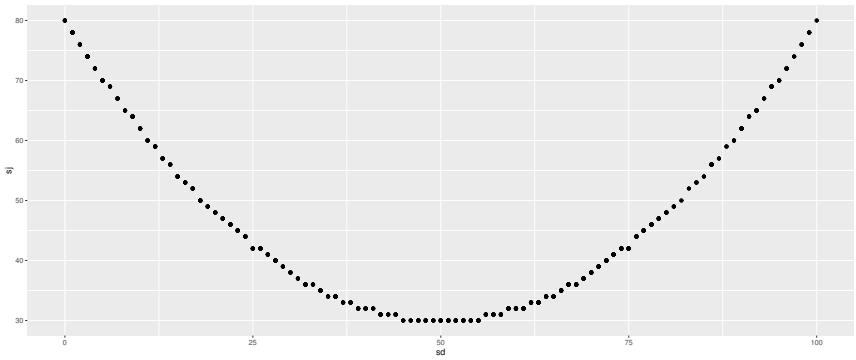

Fig. 5. Distance Score vs Jumps Score for males from location B

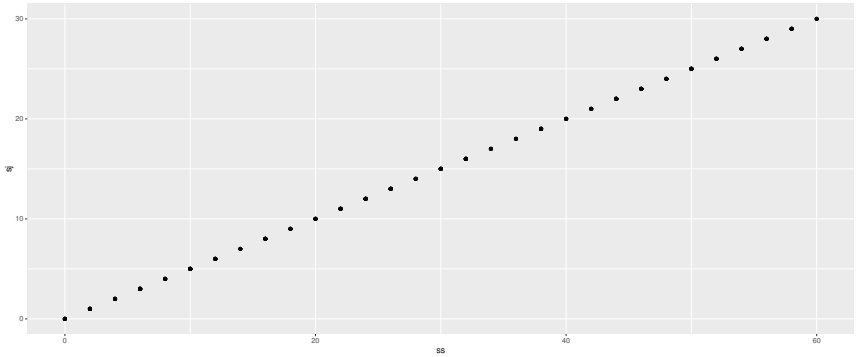

Fig. 6. Sprints Score vs Jumps Score for females from location E and age group T

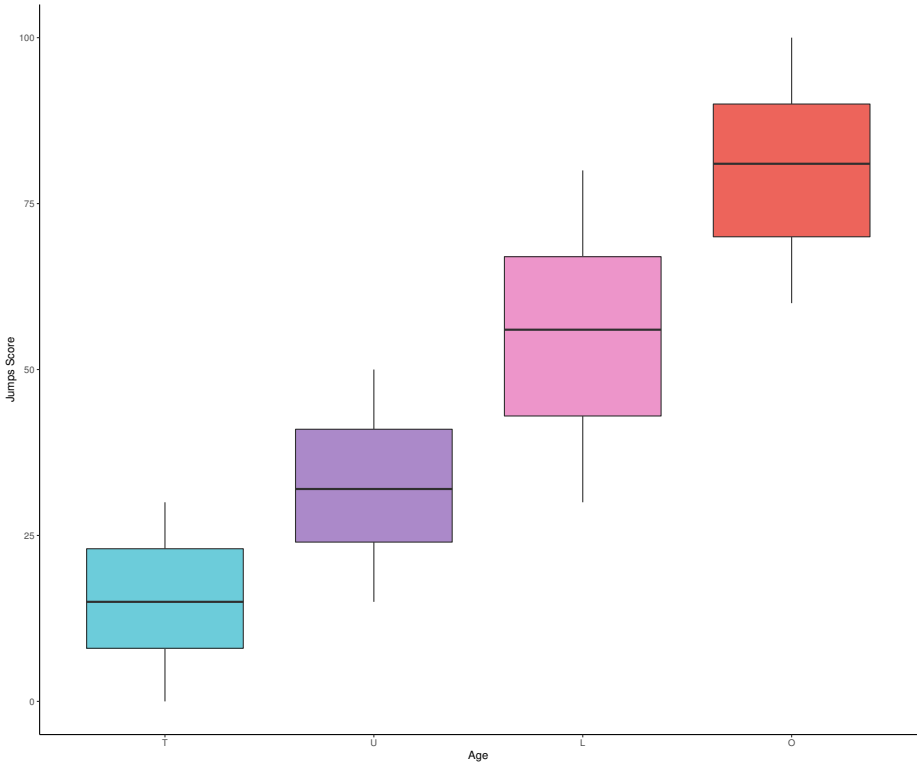

Fig. 7. Age vs Jumps Score for females