Visual Analytics: Analysis of the 2011 CrossFit Open Data

Daniel Dixey, *MSc Data Science, City University London*

**Index Terms**—Visual analytics, city university, crossfit open 2011, multidimensional dataset

Introduction

* 1. **Overview of the dataset**

CrossFit is a rigorous fitness methodology that attempts to unite multiple domains of fitness into one programme; weightlifting, High Intensity Interval training and gymnastics, where the primary aim is to prepare all participants to be ready for any physical challenge that may arise. The dataset that has been obtained and used in this report has been taken from the CrossFit Open leader board on the CrossFit Games website [1] from the year 2011. The Open is a worldwide competition that aims to bring to the community of CrossFit together via a series of five weekly physical workouts.

**1.2 Aim**

The identification of early talent in all industries, especially sport, can be argued as one of the most difficult challenges facing all organisations. Although the ease at which information can be mined and collected has improved vastly in recent years because of the development in computational techniques, managing talent in the sport industry proves challenging when it is on a large scale as it is difficult to understand which athletes should be leveraged in the hope that they become the greatest talent.

This report aims to see if the interaction between humans (Analysts/Data Scientists) and the use of visualisation techniques can be used to exploit the challenge in identifying talent. This will be presented firstly through a research questions and then intermediate questions, the scope of study will be discussed prior to the dataset being presented. Following this, the analysis of tasks and analysis of methodology will be presented and the implementation techniques and analysis of processing outlined in detail before the results and conclusion finalize the paper. First to be presented are the research questions

.

**1.5** **Data**

Link to Dataset: [***http://xfit2011.blogspot.co.uk/2012/02/crossfit-open-2011-dataset.html***](http://xfit2011.blogspot.co.uk/2012/02/crossfit-open-2011-dataset.html)

**1.5.1 Dataset source**

The dataset has been acquired using a web scraping technique; the details of the methodology of this scraping have not been disclosed were not disclosed by the source. Typically, the most widely adopted method of extracting the data is through the use of a web crawling application; a web crawler would navigate a website recursively according to a set of defined user parameters to extract the underlying web script of each page. Once the collection has been completed the data is parsed so that the desired features, in this case the features of the dataset, are obtained and output to a tabular format (CSV format), which is then ready for further processing and eventually exploratory data analysis.

* Daniel Dixey, City University. E-mail: ackf415@city.ac.uk

**1.5.2 Data properties**

An outline of the properties of the data has been captured in the table below:



* Class of Data - Object Referenced Data
* Completeness – For all the attributes there exists some a degree of incompletenIn order to use a true representation of the sample of data available, during the analysis any missing values will not be inferred or interpolated but will be ignored at the expense of introducing bias results.
* Uncertainty – The data that has been collected is based on the values that have been supplied by the individuals who have registered. The credibility and integrity of the data could be of some concern, spurious and obviously false (age >100 years and weight <20kg) information will be removed where deemed necessary during the analysis.

**1.5.4 Basic Statistics of the Data**

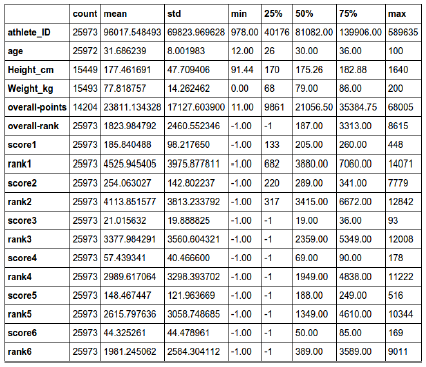
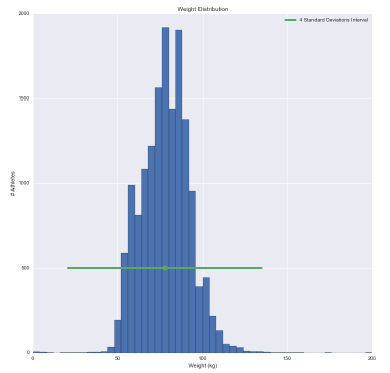


Fig. 1. The table statistics represents the whole dataset prior to pre-processing of the data.

 Fig. 2. Histogram plot of the Weight Feature, the snapshot shown is prior to removing outliers from dataset. The green point shown in the centre indicates the mean; at either end of the green line the maximum point represents four standard deviations.

**1.5.5 Assumption - Population v Sample**

The dataset that has been used for this analysis is only assumed to be representative of the population of competitive athletes that are eligible to compete in the CrossFit Games. As discussed in the article, 209,585: Rise of the Open [2], the growth of the CrossFit Open “tracks nicely” with the number of affiliates, gyms, globally that are opening. Accurate numbers of the number of athletes and affiliates are not disclosed; therefore any conclusions drawn from the sample are also then an approximate representation of the population.

**1.5.6 Limitations – Integrity of Data Collection**

As mentioned earlier in section 1.5.2, the data collection method is somewhat questionable as there may be spurious data when an athlete has populated their profile, i.e. his/her age, weight or height, with an incorrect or inaccurate value. The uncertainty of the data is therefore a limitation in the precision of the results derived from analysis.

**2. Analysis Tasks**

**2.1 Research Questions / Synoptic Task**

The research question I will be addressing in this report is “Compare the variations in athlete scores and profiles to identify different cohorts of performance?”

**2.1.1 Intermediate Questions**

* Who are the top performing individuals, both globally and by region? Are they comparative and therefore exhibit the same types of characteristics?
* How does the decrease in participation from the first Open event to the final event distort the comparison of athletes?
* What key traits do the top athletes demonstrate that are not visible in the lease successful population?
* How do gender, height and weight relate to the performance of athletes in the CrossFit Open competition?

It is expected that during the analysis further questions (elementary and intermediate) will arise as a result of these pre-determined questions.

**2.2 Work Addressed in Literature Review**

As part of the Literature Review, dimensionality reduction (DR) was discussed. The focus of the literature review was deliberately limited to this area, the number of considerations and techniques available in the area of multidimensional analysis are vast but DR was specifically identified as very useful in assisting with the visualisation of multidimensional datasets. When a dataset include an excessive number of dimensions, an Analyst trying to understand the data model will find it difficult to interpret and make meaningful deductions from the data. When this issue is evident the Analyst can use a technique called Principal Component Analysis (PCA) which enables the reduction of a number of features without the loss of information. No loss of information is achieved as the eigenvalues generated are multiplied by the normalised data and will return the original dataset. This is an extremely valuable property as it means any interesting patterns can be explored and will not be lost as a result of implementing a DR method.

**2.2.1 Extension of Techniques**

In order to assist and supplement the analysis using Dimensionality Reduction (DR), additional techniques will also be used. An overview of the additional techniques utilised are briefly outlined below.

**2.2.2 Contingency Tables and Aggregation**

Contingency tables are a type of table which display a consolidated overview of the interaction between different dimensions, these can be used to compare group distributions and also display summarised values, statistics and ratios before undertaking any unsupervised learning techniques. Using this method enables the analyst to gain an understanding of the basic characteristics in the data. This is a very advantageous method, particularly to gain an insight to help decide if the data needs to have a transformation applied to it and also, in some cases, can aid with understanding the context of the visual. Visualising within a chart is often the preferred choice as it is considered easier to interpret the characteristics.

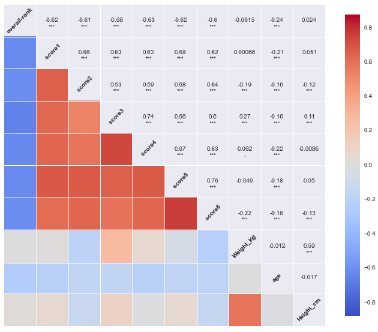


Fig. 3. Post cleaning, this table represents the correlation matrix between each of the features. The main points from this matrix are that there are strong relationships between each of the scores and that as expected there is an inverse relationship between the scores and overall rank.

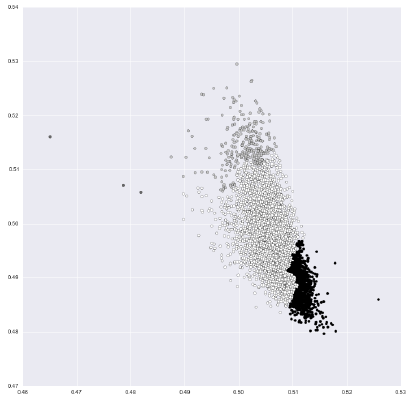


Fig. 4. Dimensionality Reduction using an Autoencoder has reduced the Score features (6) to two components. The clusters identified above have been obtained from using an Agglomerative Hierarchical Clustering algorithm, this plot shows 3 clusters that have been identified. During the analysis many variations of the number of clusters were tested.

**2.2.3 Clustering**

Clustering can be characterised as a branch of unsupervised learning which attempt to identify any hidden characteristics of a dataset. These are usually distance or similarity based methods and clustering can also be subdivided into four main areas; distribution-based, density-based, centroid-based and hierarchical-based clustering methods. Benefits of using clustering are that it is able to identify and expose concealed relationships within the dataset.

**2.3 Relation to Tasks in the Report**

The work undertaken in the Literature review [3][4][5] covered systematic approaches to ensure that interesting subspaces (patterns) were not lost as result of the application of Dimensionality Reduction. A similar approach to that described in those papers will be followed; an overview of this approach is listed below:

* Import the dataset
* Dimension Modification (normalisation)
* Dimension Cluster Selection/Creation
* Generation – relating multiple attributes in a systematic and logic method
* Selection – understanding and interpretation of clustering
* Iteration – refinement of clustering and the perception that can be derived from the visual

The iteration phase involves changing the parameters of the computations to seek improvement and understanding about a particular feature and or aspect of the dataset.

* Data Projection and Visualisation
* Parallel coordinates plots
* Scatter plots

Scatter plots were used well in the Geoffrey Hinton [5] paper to show the effectiveness of using an Artificial Neural Network (ANN) specifically an Autoencoder for DR. Within the paper there were two plots of the same dataset where the dataset was reduced to two dimensions; one visual showed the projection of a PCA output and the other was an Autoencoder output. It can be clearly seen that the Autoencoder representations are a lot easier to interpret due to the separation of the instances and the colouring used, these can be observed in figure [3a] of the paper.

**3. Analysis Methodology**

**3.2.1 Pre-processing and Cleaning**

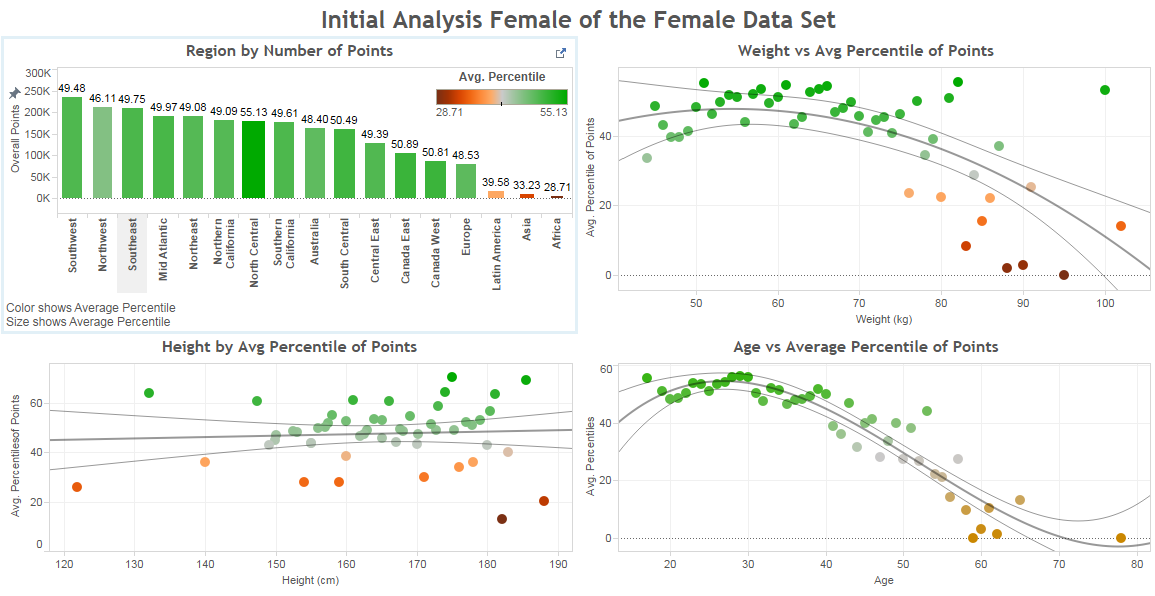
As a result of the web crawling the Open data is untidy and the data will require cleaning to ensure that it is in usable format. The types of issues that are prevalent throughout the dataset are; missing values/data, the units (kg/lb/cm) used for some variables are inconsistent (erroneous values) and finally the names will require delimitation to ensure they are in a usable format.

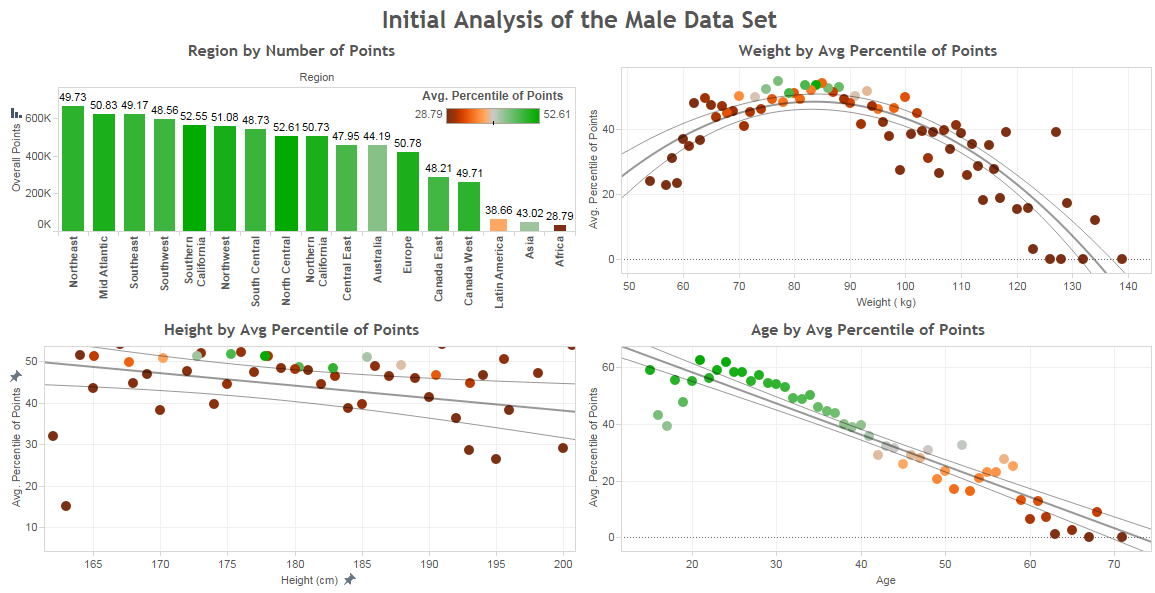
**3.2.2 Transformations**

An unsupervised learning technique called Agglomerative Hierarchical Clustering (AHC) will be used to understand and gain a number of insights into to the number of clusters that exist within a dataset. The output of this method This method of clustering offers a more robust (statistical) use over k-means clustering as in k-means clustering the user is expected to enter the number of clusters they expect to see and finally that the clusters are created at the individual instances level. The reason for this is that an analyst’s domain knowledge could inhibit the outcome of information where they are required to make decisions about what they expect to see, AHC clustering removes the need for a pre-determined outcome as well as this can be determined by setting a threshold at which the algorithm should stop calculating. My own perceived interpretation of the data should not influence the decision of the number of athlete groups in the dataset and this is an advantage of using AHC over k-means.

**3.3 Division of analytical labour**

A five phase breakdown of how both the division of human and computational work will be adopted is described below:



Fig. 5. 6. The dashboards above have been developed using Tableau. There is a feature within Tableau that enables interaction between each of the plots using a dashboard. When passing the cursor across any element, the surrounding visuals update according to the selection. This interactivity supports the user when analysing the data as the updating dimensions enable better understanding of the variations in the data.

**Understanding of the characteristics of the attributes**

The first stage of the analysis is to get an understanding of each of the distributions and basics statistics about every attribute. This will be achieved through the use of histograms, this type of graphic offers the most accessible means of understanding the distribution of the data. When a histogram is used in conjunction with a table of basic metrics about the attribute it will be possible to interpret the magnitude of the values in the attribute which should support human reasoning. Information expected to be discovered by the end of this phase are; the type of distribution (Gaussian, Poisson), whether outliers exist, limitations of the attributes (i.e. missing values) and finally if the attribute requires a transformation at this stage as normalisation cannot be applied at a later stage. With this information I will be able to use data mining and visualisation techniques based on the restrictions of each attribute.

**Understanding the relationships between the attributes**

It is initially thought that this stage will involve looking at a combination of histograms and scatter plots to understand the relationships between attributes depending on the type of feature being displayed. This will be achieved through the computation of visual encodings: shape, size, colour and labelling, to aid the understanding of one or many attributes and whether they are related and/or correlated to one another.

**Describing the relationships**

At this stage a good understanding of the attributes is expected. This stage will involve explaining the relationships that have been found in phases one and two. One computational method will be utilised at this stage; Agglomerative Hierarchical Clustering. This algorithm takes an alternative approach to other unsupervised methods as it works in a bottom-up method such that each cluster has many sub-clusters. This is a computationally expensive method however it will provide ordering and also the potential for exploring small clusters.

As the use of an Autoencoder was reviewed in as part of the literature review [3], a number of models will be generated to see if there can be benefit in using one for this dataset to aid the identification of interesting patterns. The advantage of this method is that if a non-linear relationship exists between features in the dataset then, in theory, the Autoencoder would be able to identify and “learn” these non-linear relationships. This is possible is due to the type of activation function used in each layer, if for example a hyperbolic tangent function [6] is used then it should be possible for the ANN to learn it.

However, due to the nature of ANN it is difficult to determine the most appropriate initial starting conditions for a network, finding “optimised” parameters requires a large amount of time and so therefore will be a big drawback of using this type of method. Dimensionality Reduction will be utilised as a means of reducing the number of features such that they can be well represented when plotting.

**Exploring the findings iteratively through interaction**

It is expected that navigating many times between phases 1-3 in a cyclic approach will be necessary. The reasoning for this is that it is likely the data and analysis will provoke more questions about the data which will need to be considered. Clustering and dimensionality reduction are the main computation techniques used in conjunction with visual encodings.

**Final presentation**

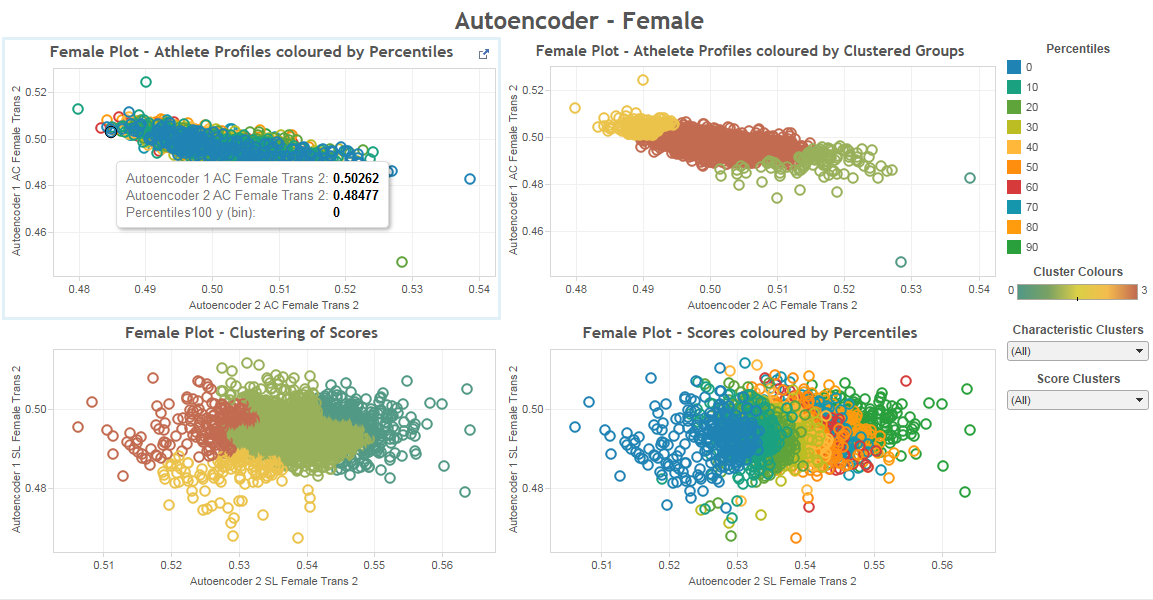
This the concluding phase where the emphasis of on improvement of the visualisations, this will involve using a number of different modules and applications. Specifically it will use Bokeh [7] and Tableau for visualisations aided by interactivity for the user exploring the data.

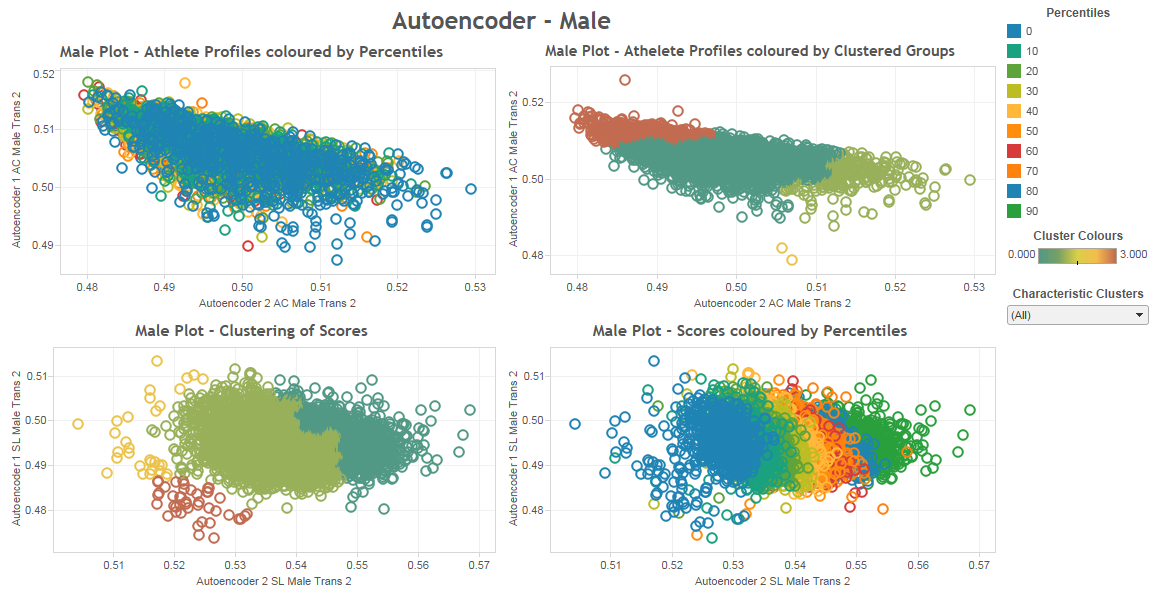
**4. Implementation**

**4.1 Software used**

All of the initial pre-processing tasks of the dataset that are mentioned in Section 3.1.2 will be undertaken in Libre Office Calc (Linux Excel equivalent). The argument for the use of Calc is to recover the highest amount of data possible, as opposed to applying generic rules in a scripting language like Python. The advantage of doing it this way is that if there are any expected or erroneous values then these can be accounted for at a more granular level. For the second phase of the implementation the data is required to be in a CSV format such that it can be read into Python.

Once the pre-processing phase has been completed the data will be loaded in Python and for the majority of the exploratory analysis and visualisation the intention will be to use the module called Orange (python-orange) [8]. Orange is data mining module that supports visualisation and analysis through a user interface called Canvas. Orange Canvas, as shown in figure X, supports the use of visual work flows that enable tracking of analysis and the exploratory journey through the dataset. There are a number of benefits with this type of interface, it cleverly allows the building and testing the users various ideas and visualisations whilst also keeping track through the use of the widgets and work-flows. By design this also supports the iterative nature of visual analytics where the use of computational and visual methods are maximised in potential in order to extract knowledge from the data. The main advantage of using Canvas over a predominately scripting approach means more of the data can be explored and analysis can be done in a much shorter time frame. A shortfall of this type of approach as it is restricted by the limited visual options that the module offers.



Fig. 7. 8. The dashboards, both male and female versions, represent the output of the analysis using dimensionality reduction and clustering to try to understand if clusters of interest exist. It can be observed in the figures that the clusters identified when compared to the percentiles (rainbow colouring) do in fact suggest instances of interest. The colour scale used for the percentiles was picked specifically to enable easy visibility of the different clusters within the data. The characteristics of the scores shown on the lower half of the dashboard are usable as there is clear distinction between the different clusters, whereas the plots on the upper half are unfortunately not usable since no clear distinction exists to explain what the clusters may be indicating.

Although the Orange module was identified as a good tool with the capability of building work-flows, it was established early on that it was not feasible to conduct the analysis using the tool due to the number of errors and random crashing of the application. As a revised workaround IPython was utilised instead. This module offers an organised work flow through the use of cells and the same functionality with regards to algorithms can be utilised in this application, with the additional benefit that the plots are highly customisable.

The expectation is that all the work would have been conducted in Orange Canvas, as the number of widgets available is vast. The widgets were categorised into nine types of process, the most relevant to this analysis are; Data, Visualize, Classify, Regression and Unsupervised. Orange Canvas supported the use of Machine Learning, both unsupervised and supervised algorithms as well the more traditional data mining techniques like Principal Component Analysis (PCA). These methods are imported from the Scikit-Learn module in Python and luckily can be imported into IPython; therefore the same methods could be exercised in IPython once Orange Canvas was seemed unusable.

**5. Analysis process**

The process of analysis followed the logical steps outlined in section 1. The carefully completed analysis was extremely detailed in order that the analytics produced comprehensively representative of the dataset. The figures distributed throughout the document depicted the Visual Analysis process that was undertaken. The author would encourage the reader to review the IPython HTML document and the tableau dashboards for better holistic view understanding of the process.

**6. Results and Conclusion**

**6.1 Evaluations of Results**

It is thought that the tasks outlined in Section 1 of this report have been fulfilled and answered successfully. It can be seen in Figures [7, 8] that that clusters which share similarities also correlated well with the Overall point’s percentiles that the athletes were scored on. The use of Tableau aided with discovery and the synergy between the computer and human, the interactivity of the dashboard helped to decide on the type of clustering algorithm to use and also which DR method was most effective.

Key findings from this analysis are that Weight and Age are good performance indicators of the athletes in the Competition. A drop off in performance can be seen in the figure by the curves that have been added to the visuals. Therefore, there is evidence to suggest that there does exist clusters which exhibit shared similarities with regards to performance in the dataset and these can be observed in Figures [5, 6]. In the context of the overall research the questions, further work would be needed however it is felt that that this best assessed by comparing this dataset to another from another time period. This could be confirm through the use of the Autoencoder, this neural network would have ‘learnt’ the key features that can be represented, by two neurons in the hidden layer, to feed the new unseen data to see if it is capable of producing the same representation as show in Figure [7, 8]. If the results did convey that the performance can be isolated into clusters of similar performance then the use of an ANN is a viable algorithm and also demonstrated it as use case for identifying talented athletes.

**6.2 Complexity Limitations**

**Demands**

In order to successfully complete this report there was a large dependency on the ability to use Data Visualisation methods to uncover interesting patterns which show relationships between the characteristics of the data. As described in Illuminating the Path [9], Thomas and Cook define visual analytics as the “science of analytical reasoning facilitated by interactive visual interfaces”. Although there was a significant amount of data available there was a requirement to find a balance between the most effective interaction between the user and the computer to identify and represent the most beneficial and relevant patterns.

**Lessons Learned**

The initial limitations of the Orange module were exposed early on in the analysis therefore it was thought that this only had a minor impact on the overall analysis. However, perhaps more research could have been completed before commencing with Orange which may have unearthed the issues to be faced and therefore the decision to use an alternative module could have been taken earlier which would have a saved a lot of time and effort. In conjunction with this, the added complexity of scripting in Orange meant that the visuals would have suffered given that it is reliant on a user’s ability and creativity, this should also have been considered more seriously before identifying a method as the negative impact this would have had on time and effort required was only considered once the module had been changed. Secondly, the preparation and prior thought given to the analysis method and research questions addressed helped ensure that throughout the course of the analysis there was structure and purpose.

Fig. 9. This screenshot depicts the type of analysis possible when using Orange-Canvas. The branches that direct upwards were used to understand if the import had been successful; the branches on the lower half are the attempts at removing outliers and erroneous instances. It was at this stage of the analysis that the memory issues and import modules issues arose meaning no further work was completed in Orange-Canvas.

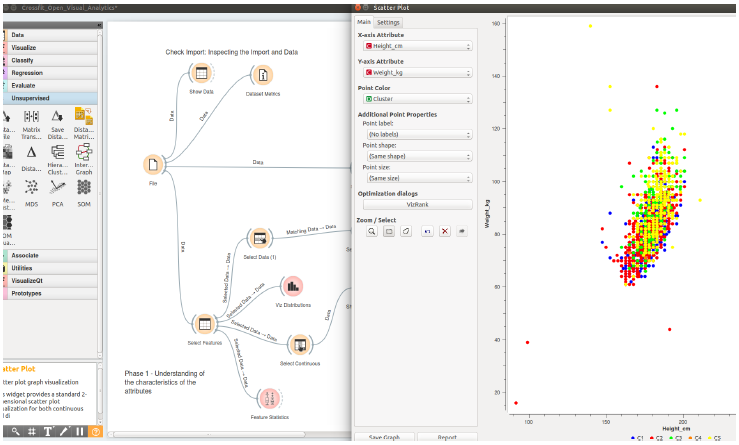
**6.4 Future Work**

* Future developments of this paper could include:
* Analysis of the different yearly datasets to ascertain if the same characteristics are comparable in top performers year-on-year as the sport of CrossFit matures
* To understand if there is an annual increase on performance or alternatively if it is starting to plateau.
* Obtain more data about the Athletes to see if the are other factors that can attribute to their performance in the CrossFit Open
* What the impact would be of adding weight divisions to the Men and Women’s categories to the competition?

**Acknowledgements**

I would graciously like to thank Prof Gennady Andrienko, Prof Natalia Andrienko and Dr Aidan Slingsby for the delivery of the Visual Analytics course; Fernando and Granger, Brian E. for the development of IPython [10] and the many contributes of the Scikit-Learn module which I have used on many occasions to support the development of this report [11].

References

1. CrossFit Games | The Fittest on Earth. 2015. CrossFit Games | The Fittest on Earth. [ONLINE] Available at: http://games.crossfit.com/. [Accessed 23 April 2015].
2. 209,585: Rise of the Open | CrossFit Games. 2015. 209,585: Rise of the Open | CrossFit Games. [ONLINE] Available at: http://games.crossfit.com/article/209585-rise-open. [Accessed 23 April 2015].
3. George E. Hinton and Rusian R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, Science, 313, 504507, 2006.
4. Jing Yang, Matthew O. Ward, Elke A. Rundensteiner and Shiping Huang, "Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets", VisSym, 2003.
5. Bhusan K. Kuntal, Tarini Shankar Ghosh and Sharmila S. Mande, “IglooPlot: A tool for visualization of multidimensional datasets”, Genomics, Volume 103, Issue 1, Pages 1120, January 2014.
6. Bengio, Y. 2012. Practical recommendations for gradient-based training of deep architectures. In Neural Networks: Tricks of the Trade, pp. 437-478. Springer: Berlin, Heidelberg.
7. Welcome to Bokeh — Bokeh 0.8.2 documentation. 2015. Welcome to Bokeh — Bokeh 0.8.2 documentation. [ONLINE] Available at: http://bokeh.pydata.org/en/latest/. [Accessed 23 April 2015].
8. Demšar, J., Curk, T., & Erjavec, A. Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14(Aug):2349−2353, 2013.
9. J.J. Thomas and K.A. Cook (Eds.), Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Press, 2005.
10. Fernando Pérez, Brian E. Granger, IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53. URL: http://ipython.org
11. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.