Visual Analytics Report

* **Abstract**

To Do.

* **1. Introduction**
  + **1.1 Overview of the dataset and problem**

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

* + **1.2 Aim**

The identification of early talent in Sport, Business and Academia can be argued as one of the most difficult challenge facing all organisations in these domains. Managing talent is a management challenge such that its difficult to access on a large scale if it is possible to profile and understand who in a large number of people that can be leveraged to turn the talent in to greater benefit. In recent years the ease at which information can be mined and collected as improved vastly. This report aim to see if the interaction between humans (Analysts/Data Scientists) and a variety of visualisation techniques can be used to exploit the Talent difficulty.

* + **1.3 Research Questions**

What evidence and knowledge can we ascertain from a data driven assessment is there to suggest that the use of visual and computational methods help to support the identification of early talent?

* + - **1.3.1 Intermediate Questions**

1. Who are the top individuals globally and by region? Are they comparative and exhibit the same types of characteristics?
2. How does the drop of participation of the Open event affect the difficultly in identifying talent?
3. When can the use of unsupervised techniques to utilised and visualised effectively?
4. What key traits do the top athletes have that the less competitive do not?
5. How does the effective of age, weight and height of an athletes performance?
   * **1.4 Scope of Study**
     + **1.4.1 Dimensionality Reduction**

As part of the Literature review an area of the consolidate review was discussed analysing dimensionality reduction. The focus was deliberately limited as the number of options available in the area of multidimensional analysis is vast. Dimensionality reduction is not the only method available however it was identified as very useful technique to assist with the visualisation of multidimensional data. Sections 1.4.2 will give an overview of other techniques that will be used in this report and that were not discussed in the literature review.

* + - **1.4.2 Extension of Techniques**

In order to assist and supplement the analysis in this report additional techniques other than Dimensionality Reduction (DR) will be used. An overview of a series of other methods and accompanying description is discussed below.

* + - * **1.4.2.1 Contingency Tables and Aggregation**

Contingency tables are type of table that displays a consolidated overview of the interaction between different dimensions, these can be used to compare groups distributions and also display summarised values, statistics and ratios. Before undertaking any unsupervised learning techniques creating and understanding the data in order to gain an understanding of some basic characteristics is very useful. This is one of the advantages of why using these tables is very useful to gain information from the data. If discritization or normalisation of the dataset is required as part of transformation, a contingency table can be used (others methods exists like plotting in the form of a histogram) to access what to do.

* + - * **1.4.2.2 Clustering**

Clustering can be characterised as a collection of unsupervised learning algorithms that attempt to identify the hidden underlying characteristics of a dataset usually distance or similarity measure. Clustering can be subset into four main area; Distribution-based, Density-based, Centroid-based and Hierarchical based clustering methods.

* + - * **1.4.2.3 Classification**

Classification algorithms are the supervised variations of the clustering algorithms where the main difference is that the group (target output) is known. This report will limited the number of algorithms used to two algorithms; Decision Trees and Support Vector Machines.

* + **1.5 Data**

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

* + - **1.5.1 Dataset Source**

The dataset has been acquired by a web scraping technique, the details of the methodology of scraping have not been disclosed. The most typical method of extracting the data is through the use of a web spider (web crawling application) that navigates a website recursively according to set a defined user parameters. One the data has been collected the data is parsed such that the desired features from the data is obtained and output to a tabular format, in this case a csv format.

* + - **1.5.2 Data Properties**

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

****

* + - **1.5.4 Basic Statistics of the Data**

****

* **2. Analysis Tasks**
  + **2.1 Multidimensional Data \_ Literature Review**
    - **2.1.1 Dimensionality Reduction**
    - **2.1.2 Other Work Addressed in Literature Review**
* **3. Analysis Methodology**
  + **3.1 Overview of Processing and Investigation Undertaken**
  + **3.2 Description of Methodology**
    - **3.2.1 Preprocessing**
      * **3.2.1.2 Cleaning**
        + **Prior to use**
      * **3.2.1.3 Transformations**
        + **Adding Variables/feature engineering**
      * **3.2.1.4 Assumptions**
        + **Population v Sample**
      * **3.2.1.5 Limitations**
        + **Integrity**
      * **3.2.1.6 Restrictions**
    - **3.2.2 How the analytical labour is divided between the human and computer**
      * **3.2.2.1 Overview**
      * **3.2.2.2 Representation of Steps**
  + **3.3 Computational Methods**
    - **3.3.1 Supervised/Unsupervised Machine Learning Algorithms**
      * **Algo 1**
      * **Algo 2**
    - **3.3.2 Contingency Tables**
  + **3.4 Visual Methods**
    - **3.4.1 Various types of Plotting**
    - **3.4.2 Tables**
    - **3.4.3 Key Statistics**
* **4. Implementation**
  + **4.1 Software used**
    - **4.1.1 Python** 
      * **Outline Methods/Algorithms**
      * **Script Development**
      * **Orange - Canvas**
        + **Visual Development environment**
      * **Scikit Learn**
        + **Machine Learning algorithms**
        + **Algorithms**
        + **Statistics**
    - **4.1.2 Tableau**
      * + **Final Presentation**
* **5. Analysis Processing**
  + **5.1 Visual Limitations of the Data**
  + **5.2 Intermediate Question Presentation and Discussion**
  + **5.3 Discussion**
  + **5.4 Presentation of Results**
* **6. Results and Conclusion**
  + **6.1 Evaluations of Results**
  + **6.2 Importance of Domain Knowledge**
  + **6.3 Complexity**
    - **6.3.1 Demands of Resources**
    - **6.3.2 Findings in context with the Research Question**
    - **6.3.3 Lessons Learned**
    - **6.3.4 Future Work**
  + **6.4 Acknowledgements**

1. **Data [10%].** Description of the data chosen for the analysis: type, structure, size, properties of the components.
2. **Analysis tasks [10%].** Analysis task(s) chosen for your analysis:
   * Do these relate to the tasks you addressed in your literature review (task 1)? If so, how?
   * Specific task formulations for the chosen data and the corresponding generic task types.
3. **Analysis methodology [30%].**Methodology of your analysis.
   * Does this relate to those in your literature review (task 1) and/or a lecture/practical? If do, how? Did you add any modifcations? If so, what and why?
   * Description of your methodology.
     + How the analytical labour is divided between the human and computer.
     + Computational methods involved, types of their inputs and outputs.
     + Data transformations (if any).
     + Visual and interactive techniques involved. How the visualisation supports the human reasoning.
4. **Implementation [30%].** Implementation of the analysis methodology: software used, links between methods (integrated in the same software or data transfer).
5. **Analysis process [10%].** Illustrated description of the analysis process, including all steps and intermediate results. The description must demonstrate fulfilment of the requirements set in section 1. The illustrations must include commented screenshots of the visual displays used in the analysis. The comments must explain what the displays show and how this information was used in the following analysis steps or contributes to the final result.
6. **Results and conclusion [10%].** To what extent the posed task(s) have been fulfilled?