Snooker Matches Data Exploration

Justin Thomas, Big Data Analytics, L00173759

***Abstract*— Snooker is a data rich game where frame scores decide the winner of a frame and then a match. The data-set I used has matches the whole way back to 1982. I used data visualisations to explore the data and see how the sport has changed throughout the years and who the major players of the sport are. Data visualisation along with description statistics are one of the first steps when conducting an analysis. A lot of the time whoever is conducting the analysis wouldn't have the business knowledge required. Various graphs can give a better understanding of what the data contains and how certain variables relate to each other and may lead to ideas of what to predict or how to go about it. Pandas, Matplotlib and Seaborn are three useful packages which enable relative beginners to carry out very powerful analysis with very few lines of code.**

***Index Terms*— Snooker, Professional, Frames, Billiards, Visualisations, Data Exploration**

I. Introduction

Snooker is a popular billiard sport played on a rectangular table covered with a green cloth called baize, with six pockets; one at each corner and one in the middle of each long side. Using a cue stick, the individual players take turns to strike the white cue ball to pot the other twenty-one snooker balls in the correct sequence, accumulating points for each pot. An individual frame of snooker is won by the player who has scored the most points. A snooker match ends with one of the players having won a predetermined number of frames.

An exploratory data analysis was conducted on data from 38 years of matches. The data was also examined for difference in quality between the professional and non-professional tour.

II. Data-set Description

4 csv files were obtained from Kaggle.com, 1 is a file of the tournament information, including the year, country, season, competition name, the level of competition and the prize pool in British Pounds. Another csv is the players names and where they come from. Another csv is the matches file, containing the names of the players involved, the date of the match, what stage in the competition it was, how many frames it was best off and the score. The last csv is the scores file which is at frame level. It has the score of the frame, which match it was, and lists out all the breaks over 50 points.

III. Methods

The data exploration was conducted in Google Colab. The data-sets downloaded firstly from Kaggle were then uploaded to the storage session and then read them in through the popular Pandas package.

A. Data Preparation

The matches csv is the main data-set i worked with. First of all, the features that were not going to be important or practical to work with were dropped such as URL's and the frame score in a string format. A different data-set has scores which was easier to work with. The tournament data-set was then merged in to get more information on the tournament each match belonged too, columns include name, category, status, prize, city and country. Then i merged in the players data-set to figure out the country represented by each player.

	match_id	date	stage	best_of	player1_name	country_y	player2_name	score1	score2	country	year	name	status	category	prize	country_x	city
0	82716	NaN	Final	31	Terry Griffiths	Wales	Alex Higgins	16	15	Northern Ireland	1982	UK Championship	Professional	Non-ranking	47000.0	England	Preston
1	82805	NaN	Last 16	9	Terry Griffiths	Wales	Alex Higgins	5	2	Northern Ireland	1982	International Open	Professional	Ranking	75500.0	England	Derby
2	83575	NaN	Semi-final	11	Terry Griffiths	Wales	Alex Higgins	6	4	Northern Ireland	1982	Masters	Professional	Invitational	27000.0	England	London
3	83757	NaN	Semi-final	9	Terry Griffiths	Wales	Alex Higgins	5	1	Northern Ireland	1982	Classic	Professional	Invitational	15000.0	England	Oldham
4	90533	NaN	Last 16	25	Terry Griffiths	Wales	Alex Higgins	13	7	Northern Ireland	1985	World Championship	Professional	Ranking	300600.0	England	Sheffield
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
192579	207378	2019-11-17	Last 80	5	Wun Soen Tung	Singapore	Brandon Ng Chuan Wei	3	1	Singapore	2020	Singapore-Amateur Championship	Amateur	National Championship	0.0	Singapore	Singapore
192580	207374	2019-11-17	Last 80	5	Tan Thiam Heng	Singapore	Vincent Lim Hock Kia	3	0	Singapore	2020	Singapore-Amateur Championship	Amateur	National Championship	0.0	Singapore	Singapore
192581	207370	2019-11-16	Last 80	5	Sunny Wang	Singapore	Gareth Keigan Lee	3	1	Singapore	2020	Singapore-Amateur Championship	Amateur	National Championship	0.0	Singapore	Singapore
192582	207376	2019-11-17	Last 80	5	Cheah Weng Keong	Singapore	Andreas Ngiow	3	1	Singapore	2020	Singapore-Amateur Championship	Amateur	National Championship	0.0	Singapore	Singapore
192583	207373	2019-11-17	Last 80	5	Ng Yong Peng	Singapore	Zhang Peihao	3	0	Singapore	2020	Singapore-Amateur Championship	Amateur	National Championship	0.0	Singapore	Singapore

192584 rows × 17 columns

Fig 1: Merged data-frame of matches, tournaments and players

B. Exploration and Visualisations

To begin with the number of matches over the years was plotted. A peak in the early 90's and then a drop but on a steady rise since around 2006 as the sport becomes more global.

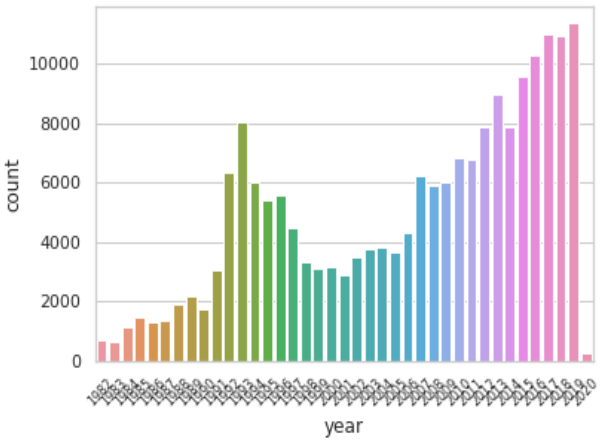

Fig 2: Number of Matches by year

Then a box-plot of the prize-pools was plotted to see the their spread

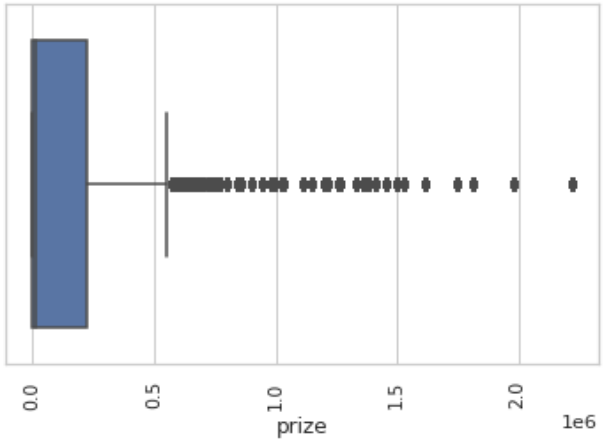

Fig 3: Prize-pools by millions in British Pounds

Then the popularity of snooker in different countries was then investigated. The number of matches held in country was then plotted. As expected England comes out on top as many of the longest running and popular major tournaments are all held in England. Only in recent years has snooker grown in popularity in other countries in particular China. Germany is in number two which I found quite surprising

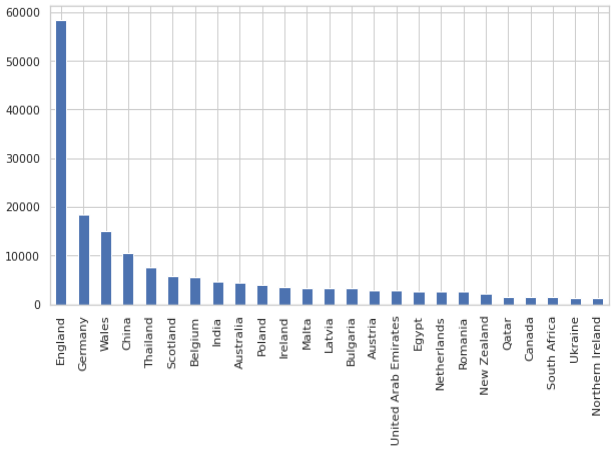

Fig 4: Number of matches by Country

Next I wanted to see what players appear in the data-set the most amount of times. There are a number of players who have stayed near the top of the game for several decades. Jimmy White comes out number one here. Jimmy turned professional at an early age like most of the players and at almost 60 years of age now, he still competes on the tour today. It's also easy to read from the graph that he has lost more matches than any other player. The top 5 players still actually are current tour players

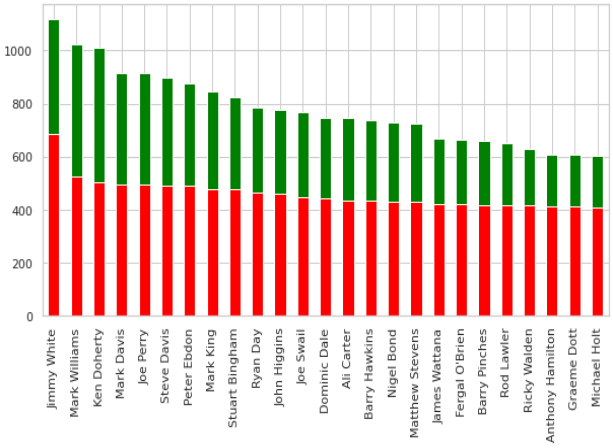

Fig 5: Players that have played the most matches

Then i had a look to see what were the most common matches in the data-set. This turned out to be Stephen Hendry and Steve Davis who were big rivals in the 90's and contested many finals together. Then comes Jimmy White's matches against the two of these who was at his peak around the same time.

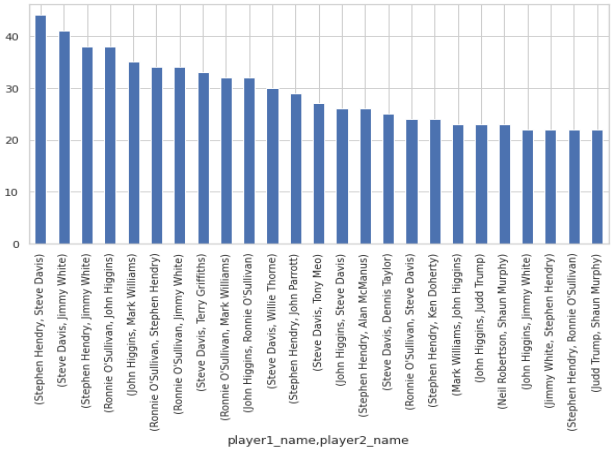

Fig 6: Players that have played the most matches

Then I created a heat-map that looks at any correlations between the numerical variables. 1.0 being fully correlated, zero being not correlated at all, -1 being fully negatively correlated. Score1 correlated to best-of (0.89) but that is expected because the score1 number should be the winners score which can be calculated by (best-of + 1) / 2 so the best of 9 is first to win 5. Best-off is correlated somewhat to prize

(0.54), this would be expected as the bigger tournaments have longer matches in them. For example, the World Championship final is first to 17 frames.

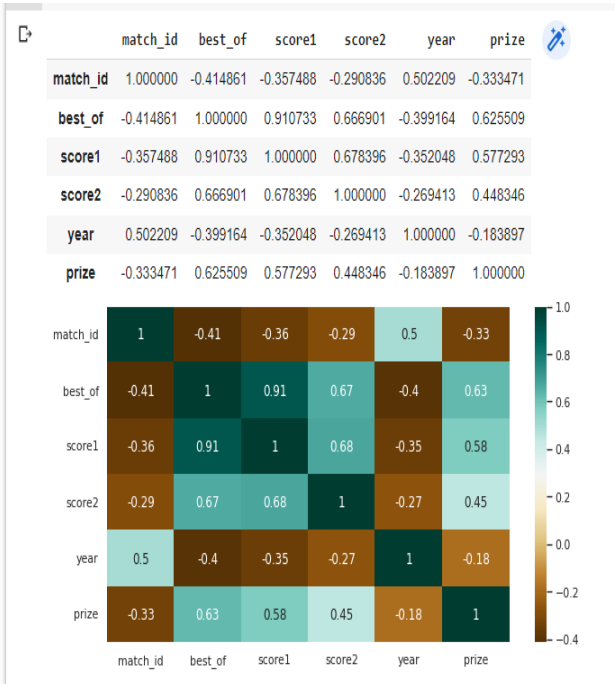

Fig 7: Heat-map of the snooker matches data-frame

Then i plotted score1 versus score2. The match winner's score goes to score1 so we would expect these always to be higher than score2. This indeed is the case in the triangle shaped graph. We can pick out an outlier in the graph and can also see that there are no matches in the whole data-set that are first to 13 or 14 frames

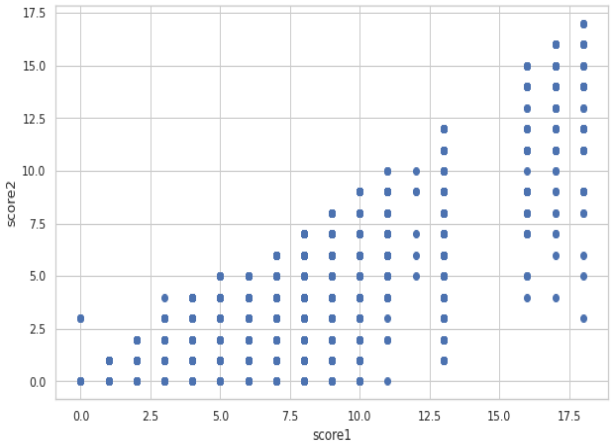

Fig 8: Scatter-plot of score1 versus score 2

I had come across a package called Sweetviz which i then ran to get an automated visualisation of the data-set. It generated some similar graphs that we had seen already. A breakdown of the status and category variables are quite interesting. The majority of the matches are professional matches and 'Ranking' is the main category

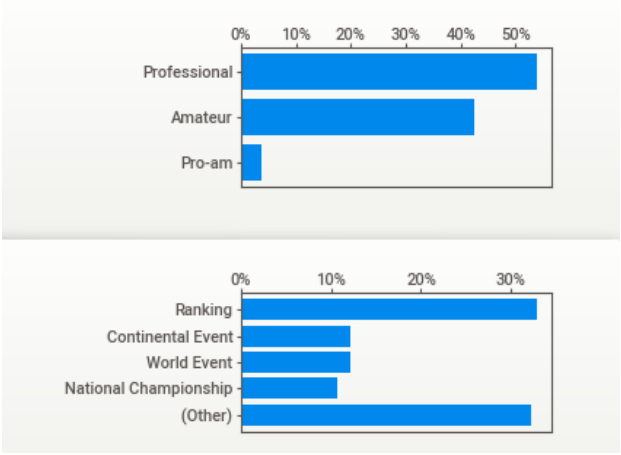

Fig 9: Sweetviz plots of Status and Category splits

Using the scores data-set, i manipulated this to include match information and to have score1 and score 2 on the same row for every frame. I then grouped by year and plotted the average score1 and score2 to see how they have changed throughout the years and see if any patterns emerge. This was restricted to ranking events as colab had struggled to deal with the full data-set

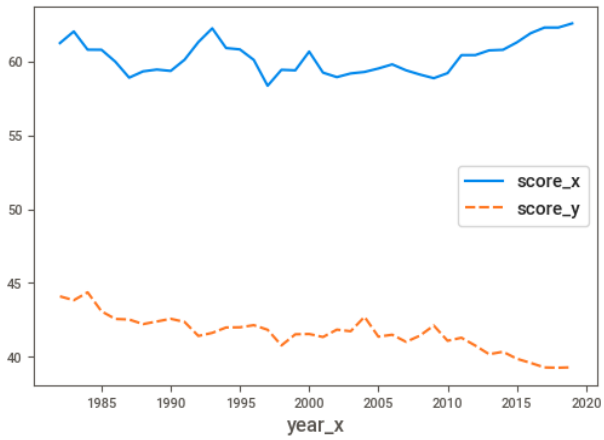

Fig 10: Score1 and Score2 averages by Year

Similarly, I plotted the frame-totals by year.

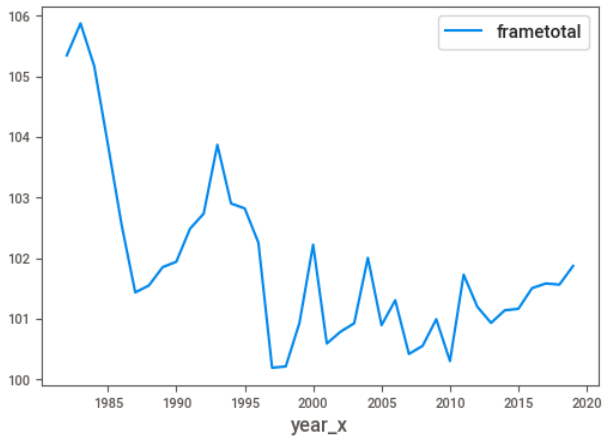

Fig 11: Frame-totals averages by Year

Then I separated them by stage of a tournament. However, the graph contained too many lines and was difficult to read so i filtered down to the later stages of a snooker tournament which is the last 32, last 16, quarter-finals, semi-finals and the final. It looks by examining the graph that finals have slightly higher frame-totals than other matches. I confirmed this afterwards in a summary table.

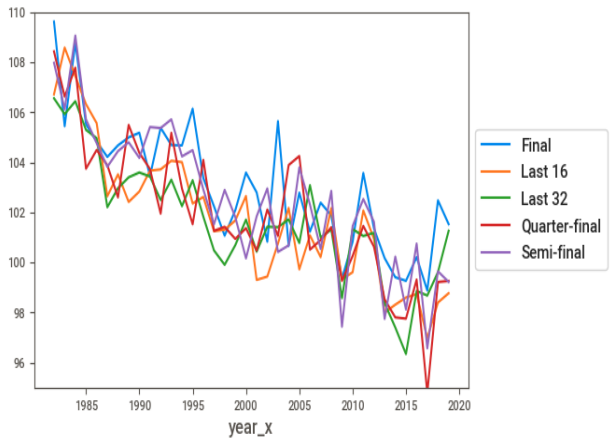

Fig 12: Frame-totals averages by Year split by the later stages of a tournament

One of the questions i asked in the proposal was to investigate if professional snooker is higher quality than amateur snooker. We can clearly see a big different in the frame-totals of the split. Professional matches average close to 100.6 points while Amateur are closer to 92 points. Pro-Am which is a mixture of both players predictably are in-between at 97.1 points. The total points in a frame is not the sole answer of determining the standard of snooker but in general you would expect better players to get bigger breaks and hence have higher scoring frames. Weaker players would miss more shots and have smaller breaks.

status_x	match_id	frame	score_x	score_y	year_x	prize_x	frametotal
Amateur	138985.062443	3.103526	58.737859	33.244948	2014.778296	175.571276	91.982807
Pro-am	175285.478292	3.247088	61.318743	35.792093	2014.622309	427.329333	97.110837
Professional	78546.729223	5.189280	60.622702	40.025827	2005.041710	406081.608555	100.648529

Fig 13: Scores and frame-totals averages by status to compare Professional v Amateur.

IV. Linear Regression Model

I then attempted to create a linear regression model to estimate the frame-total of a frame given the other variables which are frame number, player1 name, player2 name, year, status, stage, category, prize and city. The data-set is filtered to just ranking matches. The categorical variables like players, status, category and cities were all converted to dummy variables so we can perform linear regression on the data-set

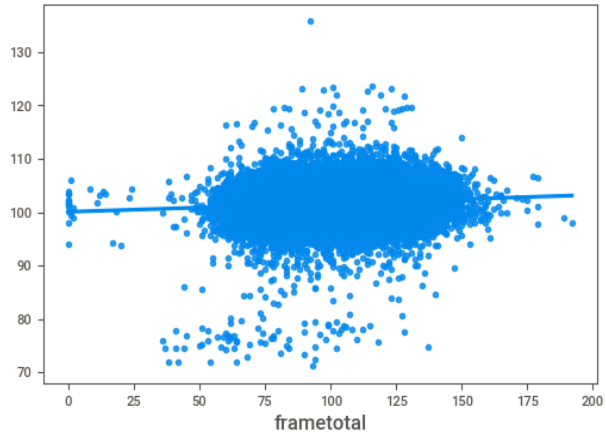

Fig 14: Plot of the predictions

The linear regression results were not great but perhaps trying to predict an actual frame-total which has quite a large range usually between 80 and 120 is quite difficult and not the best approach. The model only accounted for 1.6 percent of the variation in the test data-set. The p-value is less than 0.01 so that indicates the model is somewhat useful. The first few coefficients are quite interesting; the frame coefficient is -0.2 which indicates that earlier frames in a match are higher scoring then later frames. Year coefficent is -0.1 which also indicates that the bigger the year, the lower scoring frame-total tends to be. We noticed that in fig 11 earlier. The individual players, stages and cities all have a wide ranging of coefficients which can impact the frame-total estimate

```
import statsmodels.api as sm
X_train_Sm= sm.add_constant(X_train)
X_train_Sm= sm.add_constant(X_train)
ls=sm.OLS(y_train,X_train_Sm).fit()
print(ls.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            frametotal   R-squared:                       0.023
Model:                         OLS   Adj. R-squared:                  0.016
Method:              Least Squares   F-statistic:                     3.249
Date:             Fri, 18 Feb 2022   Prob (F-statistic):              0.00
Time:                     22:32:20   Log-Likelihood:            -8.9960e+05
No. Observations:           205556   AIC:                         1.802e+06
Df Residuals:               204055   BIC:                         1.818e+06
Df Model:                     1500
Covariance Type:         nonrobust
==============================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                        303.8063     29.736     10.217      0.000     245.524     362.089
frame                         -0.2148      0.012    -17.805      0.000      -0.238      -0.191
year_x                        -0.0989      0.015     -6.697      0.000      -0.128      -0.070
prize_x                    1.051e-08   2.71e-07      0.039      0.969    -5.2e-07    5.41e-07
player1_name_x_Adam Davies    -9.4920      7.419     -1.279      0.201     -24.033       5.049
player1_name_x_Adam Duffy     -0.5900      1.170     -0.504      0.614      -2.883       1.703
player1_name_x_Adam Stefanow   2.3175      2.547      0.910      0.363      -2.675       7.310
player1_name_x_Adam Wicheard  -3.4218      2.407     -1.421      0.155      -8.140       1.297
player1_name_x_Aditya Mehta    3.1736      0.961      3.301      0.001       1.289       5.058
player1_name_x_Adrian Gunnell  3.2229      1.066      3.022      0.003       1.133       5.313
player1_name_x_Adrian Ridley  -2.8069      7.354     -0.382      0.703     -17.220      11.606
player1_name_x_Adrian Rosa    10.3179      8.697      1.186      0.235      -6.728      27.364
player1_name_x_Ahmed Saif     15.7738     11.152      1.414      0.157      -6.084      37.632
player1_name_x_Akani Songsermsawad  1.4298  1.155     1.238      0.216      -0.834       3.694
player1_name_x_Alain Robidoux  0.1072      0.751      0.143      0.886      -1.365       1.579
player1_name_x_Alan Burnett    7.0013      7.895      0.887      0.375      -8.473      22.476
player1_name_x_Alan McManus    0.4879      0.450      1.083      0.279      -0.395       1.370
```

Fig 15: Summary stats of the regression model

V. Random Forrest Model

I then changed my approach to predict a Boolean value, whether the frame-total is over 101.5 or not. This is a common under/over used by bookmakers for top level snooker matches so i picked that one. I used 100 decision trees which is then used to take the average of predictions. Colab struggles to run this algorithm with such a big data-set. The final time i

ran this i got 51.08 percent which is just slightly better than a pure guess. This model would need a lot of improvement to be somewhat useful

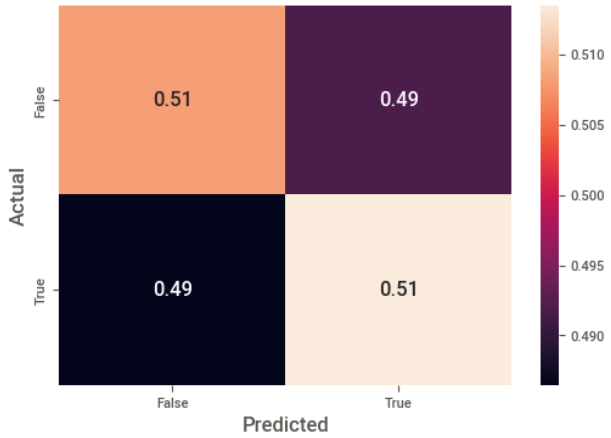

Fig 16: Random Forrest Classifier Confusion Matrix. Accuracy is 51.08 Percent

VI. Results and Conclusions

My approach ended up being a data visualisation approach to pick out interesting findings and trends within the combined data-set. Colab actually struggled to cope with some of the merges and machine learning commands and crashed numerous times. I trimmed down the data-set each time to cope but possibly i should have used Dask or Spark to do some of the work. There was findings that i would have expected from my own personal snooker interest and knowledge such as the common countries and the players with the most matches. The frametotal analysis produced results i wouldn't have expected. The machine learning modelling was unsucessful and needed a bit more thought or data brought in.

VII. Future Work

The original website cuetracker.net now contains average shot time which gives an idea of a players 'tempo' for each match. This would be an interesting variable to investigate and see how it affects players chances of success or the frame-totals in a frame or match. Further work would need to be carried out in estimating the frame-total in a frame by machine learning.

VIII. References

[1] https://en.wikipedia.org/wiki/Snooker
[2] https://www.kaggle.com/rusiano/snooker-data-19822020
[3] https://cuetracker.net