The making of champions: An analysis of university esports data

ID 1804798

Abstract

Esports (competitive, organised, videogaming) is a phenomenon that has taken the world by storm over the last couple of years, with the industry now worth over a billion US Dollars.[1] Being such a fresh and flourshing industry, there is a large amount of data surrounding esports waiting to be looked at but with little work done, especially outside of the professional scene. This project aims to look at data on the UK university esports scene to draw conclusions about which universities dominate the UK esports scene, and what gets them to the top.

1 Introduction

University esports has been happening in the UK for more than a decade now, but has been reaching a critical mass of attention and support in recent years. Within the UK, students compete in two rival tournaments, the **NUEL**[2] (National University Esports League) and **NSE**[3] (National Student Esports). We will be looking at data from **NSE**, due to its tournament structure facilitating data analysis better.

Players compete in teams with other students from their institution in a number of different tournaments and games. Each team will be allocated a number of points for how they performed in that tournament and game, and their university will get to add those points to their total on the **BUEC**[4] (British University Esports Championship) league table. This league table works in a similar fashion to how the **BUCS** (British Universities Colleges Sport) table works for universities and "conventional" sports.

At the end of each academic year, the institution at the top of this table gets to claim the title of "Esports University of the Year". Throughout this project we will be analysisng universities' performance in the **BUEC** table, and attempting to draw meaningful conclusions from this data.

2 Data Sets

There are 3 main data sources used in this project: The **BUEC League Table**, The **BUEC Seeding data**, and the **HESA Higher Education Institution Data**

2.1 BUEC League Table

2.1.1 Description

The **BUEC League Table** Comprises 2 sections. The first section is exactly the table described in the introduction, which we will call the **BUEC table** or the **university table**. It details how each university has performed across the year in each different game, and then sums up these game performances to detail a total ranking of **BUEC points**. The second section is a description of how each team performed in each game, with the points gained by each team for each university in each game detailed, which we shall call the **team BUEC table**, or just the **team table**.

The **BUEC Point tables** for the winter and spring terms for the past 3 years can be found at *https://nse.gg/tournaments/buec winter-2021/*, and then clicking the link to show the full standings.

For the most part, we will be using the Winter 2021 Data, as it is the most recent and so will have the most teams/universities participating, although we may compare data from previous years if time permits.

The first step of preprocessing we must do is separating the data into the data about universities, and the data about teams. This can be done relatively easily in excel, or any other data editing software, because the data is formatted with the two data sets separated.

2.1.2 Data Description and Preprocessing (University Table)

The **BUEC table** has 18 Attributes, which we can reduce down into 13 attributes. The Rank attribute can be removed, because it is not useful, and we can merge the different fighting game attributes into one combined fighting

game attribute (just the sum of the previous attributes). Likewise, we can combine the two Smash attributes, as well as the racing games attributes. Lastly, we make sure that the attributes share consistent naming schemes with the ones used in previous years' worth of data (eg, Hearthstone Teams → Hearthstone). The data set is mostly clean, however there are a few missing entries that should be corrected. Using the linux command line command should identify any of the problematic entries: `\textit{grep -e ",," -e",\$" -e"^," BUECData.csv`

2.1.3 Data Description and Preprocessing (Team Table)

The **team table** consists of several independent tables, one for each game, with each table containing the final standings for that game/tournament. For each tournament, we are given the final standings, i.e the position each team came, which university that team belonged to, and how many **BUEC points** that team earnt.

To preprocess this data, we simply need too merge all of these different game tables into one large table (i.e by putting them all "on top" of each other, so that the team, university, points and placement columns/attributes allign). This can be done quickly enough by hand in excel, or similar software.

The next step of preprocessing will be running through all of the team results, and recording for each university how many points they received from teams with x points (i.e how many points did they get from teams that earnt 1 point, how many from 2 point teams... how many from 150 (max) point teams) (The reasoning for doing so will be discussed in the results). It is recommended that you use a programming language (Python was used in this case) for this, although it can be achieved with scripting.

The following two commands may be of use:

```
 sed 's/,,//g' \\"Team Points Winter 2021.csv\\" | sed '/\S/!d' |
cut -d, -f2 | sort -f | uniq -i > Universities.txt}
```

(will generate you a list of all of the relevant universities)

```
sed 's/,,//g' \"Team Points Winter 2021.csv\" | sed '/\S/!d' |
sort -k3 -t ',' -n -r | grep -n -i \"," + uni + ",\" | cut -d, -f3 > temp.txt
```

(cleans the data and gets all of the points for a given university.) Note: if your machine can use linux commands, you can integrate this into your python scripts with os.system(command)

2.2 BUEC Seeding data

2.2.1 Description

The **BUEC seeding data** contains all of the seeding data obtained at the start of each tournament. The seeding data is an estimate of a teams skill, based off of their rank/elo from their respective game's own internal matchmaking system. Tournament organisers will ask for player/team ranks/elos, in order to seed the tournament appropriately. This is similar to how seeding/elo works in other tournament games, like chess[6]. The data contains several independent tables (one for each game), with each having one column for the team name, and another column for the seed.

Whilst the seeding information is public for participants of the tournaments at the beginning of the year, it is not publicly accessible once the tournament has finished. Therefore, unless you are scraping the information off the **NSE** website at the beginning of each tournament, it is advised that you reach out to the good people at **NSE** to see if they would be willing to share this data with you. The seeding data for this project was generously provided by the team at **NSE**.

2.2.2 Preprocessing

There is a fair bit of preprocessing to do with the seeding data. First, you must assign each team the **BUEC points** they would receive if they were to actually finish at their seeded position (we are only given seeds, i.e positions, not points.) The **BUEC point** allocations are calculated in a tremendously complicated manner[7]. Thankfully, these point allocations are actually based on how many teams sign up at the beggining of the tournament, so we do not have to recalculate them ourselves. Instead, we can cross reference the point allocations from the **Team BUEC Table** from before to figure out how many points each seed should get. (Eg: if 1st place got 150 points in CS:GO in the team table, then 1st seed should get 150 points for the seeding table).

Some caution should be taken as some teams share the top seed (due to their game's elo system having too low a ceiling), so points should be averaged among those teams. Also, for some games, a Northern and Southern

League is played, before the top teams get funnelled into a national division. For this, we simply average the points that the national team would get between the 2 regional teams. (Eg: if 1st place national got 150, and 2nd place national got 148, then 1st seed Northern and 1st seed Southern both get 149 points...). This was done by hand in excel, but could also be semi-automated with a python script.

Then, we can simply combine all of the different game tables together into one table with a column for team name, a column for seed and a column for points (as we did before with **Team BUEC Table**). However we may have noticed that unlike with the team table, we don't have the university the team played for this time, so we have to put in some extra effort to get that. A naive approach is to compare with the finished product of our previous preprocessing for the team table, as that table has both a university and a team name column. Therefore, we should be able to cross-reference between that table to extract a team's home institution. However, this approach makes an incorrect assumption: that teams have unique names. Whilst a team's name is unique within a tournament, it is not necessarily unique across all tournaments. (Eg: there is a different "Portsmouth Paladins" team competing in every game). Whilst it is reasonable to assume that most teams with the same name belong to the same university (like with the Porstmouth Paladins), there are a few exceptions to this rule. Therefore, we can't use this approach, and instead have to write a python script that will read the Team Tables data and compare a team against the teams from the tournament it participated in in order to find its home institution.

Given that this data set has never been used before, there were several errors/mistakes in the names that needed cleaning, both through human error, and teams changing name mid-tournament. Therefore, an override was programmed in to the python script that would prompt the user to manually enter the institution, whenever it could not be found. The institutions for each team can be found by searching up the team on **NSE**'s website: https://nse.gg.

Now that all the seeding data is contextualised, we can finally finish our preprocessing to extract some actually useful results. From the seeding data, we create a new data set with 3 new attributes:

- A "**Predicted BUEC Points**" attribute, similar to the **Universities BUEC points** attribute, but instead using what their predicted **BUEC points** would be if they performed exactly as their seeding would suggest
- A "**Predicted BUEC Points Delta**" attribute, showing the difference in predicted **BUEC points** for each university compared to their actual **BUEC points** result
- A "**Relative Predicted BUEC Points Delta**" attribute, showing the percentage difference in predicted **BUEC points** to actual **BUEC points**

We also create a data set predicting how many points each university would receive from teams with x points (as before with the Team Table). (Eg: how many points is a university predicted to receive from teams earning 1 point, how many from teams earning 2 points... how many from teams earning 150 (max) points.)

2.3 HESA Institution Data

The HESA[8] (Higher Education Statistics Agency) is the principle and official organisation/agency for data and statistics on the higher education sector. We will be using their "**students by HE provider**" data set, also called the **HESA Institution Data** set, which contains demographic information about the student bodies at each institution.

There are several attributes of interest we could look at, but for the purpose of this project, we focused on 5 key attributes (apart from the institution itself):

- the total number of students at the institution
- the total number of male students at the institution
- the total number of female students at the institution
- the total number of first year students at the institution
- the total number of non-first year students at the institution

For this project, the data from 2019/20 was used, as it was the most up to date, and it can be found here: https://www.hesa.ac.uk/data-and-analysis/students/where-studyprovider

2.3.1 Preprocessing

Because this data set is coming from an official source, not much cleaning is needed. All that is needed to be done, will be to extract from the data set the relevant attributes/columns to use, by filtering the data appropriately. Then, once each column is extracted, we just have to merge the columns in whichever method we are most comfortable (excel lookup tables was used here). Take note that although the institution name is standardised under the database, **NSE** do not use these standardised names, as they are using the public facing institution names. Therefore, you will need to clean this data set by fixing the university names so that they match, and can be merged with the **NSE** data appropriately. One example would be to remove the "The"'s at the front of institution names (eg: "The University of Warwick" $\rightarrow$ "University of Warwick")

3 Does size matter?

3.1 Hypothesis

A natural first question to ask on the data set is does size matter? That is, to what extent does the size of an institution have an impact on the amount of BUEC points they will receive? We hypothesise that size does indeed matter, and that BUEC point results probably do heavily depend on the size of an institution. We also hypothesise that given that esports is currently heavily male-dominated, institutions that are more male-dominated would also benefit from this and have better BUEC points results.

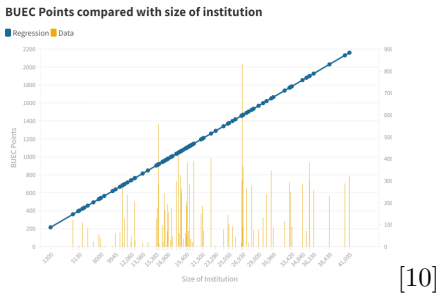
3.2 Results

To go about answering this question, we first merge the **University BUEC points table** with the **HESA Institution Data**, into one big table, with 18 attributes (1 for university, 12 for the games from the **BUEC table** and 5 from the **HESA** data).

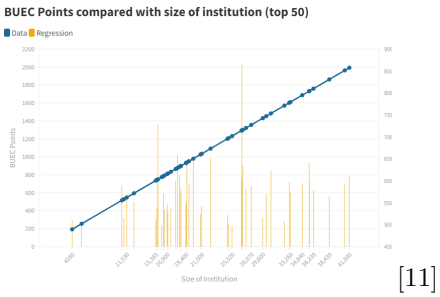
We can then run a simple linear regression between the total points scored for each university, against each of the different attributes from the **HESA data**.

Ran cross the entire data set, we find that the total student number has a correlation coefficient of 0.3907, which is far higher than any other relationship we find across this project. This tells us that there is a quite strong relationship between BUEC Points and institution size. This makes quite a bit of sense, as if an institution has more students, they are likely to have more teams, hence more BUEC points, but also have a greater pool of prospective talent to draw from, and so have better teams as well.

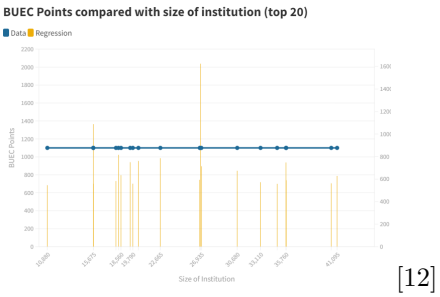
We should note, that there is more to the story. If we restrict the data set to only the top 50 universities, by **BUEC points**, we get that the correlation coefficient drops to 0.2232, suggesting a far smaller relation, and then if we restrict further to the top 20, we drop to 0.0477, suggesting close to no relation. This suggests that although there is a strong relation on the entire data set between institution size and **BUEC point** performance, as you approach the top of the table, this relation disappears. We can interpret this as saying, although institution size can be a good indicator for whether an institution will get lots of **BUEC points**, once we know that it is getting a decent number of **BUEC points**, size ceases to matter, and there are other more impactful factors at play shaping the outcome of the top end of the table.



[10]



[11]



[12]

The above images show the relationship disappearing at the top end of the table

When it came to gender data, our simple linear regressions found that although male students had a higher correlation coefficient (0.0362) than female students (0.0145), these are both very low correlations, and so while more male students does predict better BUEC Points than more female students, neither is of much use, and more students in general is a better indicator. After some thought, this makes sense, because a high or low male/female

student number tells us little about how many students of other genders there may be, and so an institution with a low number of male students may have many female students, and so still be successful, and vice versa.

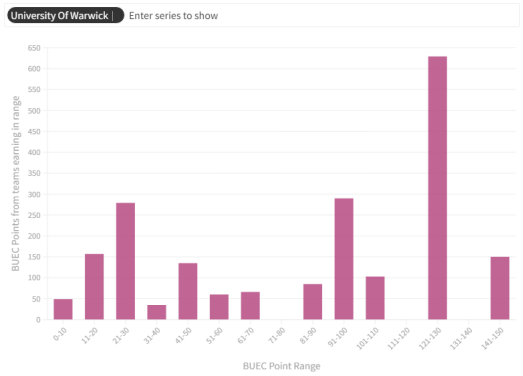
4 Zerg Rush!

4.1 Hypothesis

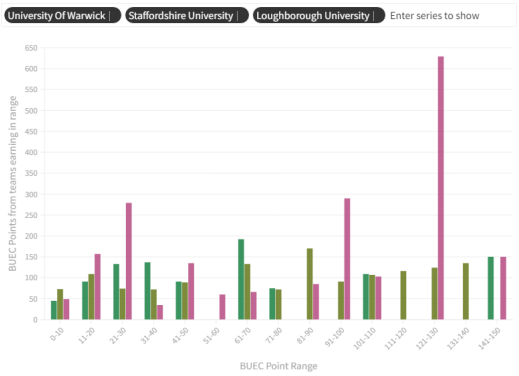
Now that we know that size does matter (at least, sort of..) for predicting **BUEC Points**, an obvious follow-up question is to ask why, and take a deeper look at what is happening. One suggestion, is that universities with more students are simply "Zerg Rushing"[9] the **BUEC Point** leaderboard. That is to say, that universities are simply fielding many many low quality teams to rack up the easy small points and have so many of them that they will easily top the leaderboard, even if the other universities perform better and make it further into tournaments.

4.2 Results

In order to test this hypothesis, we require some manipulation of the data. As described in the pre-processing steps, we need to record how many points each university gets from each team that earns "x" points. So, we must record how many points a university gets from teams that earn 1 point, how many from teams that earn 2,... how many from teams that earn 150 (max). We do this using a custom python program. We can then "Bucket" this data into bins, so that we can visualise it.



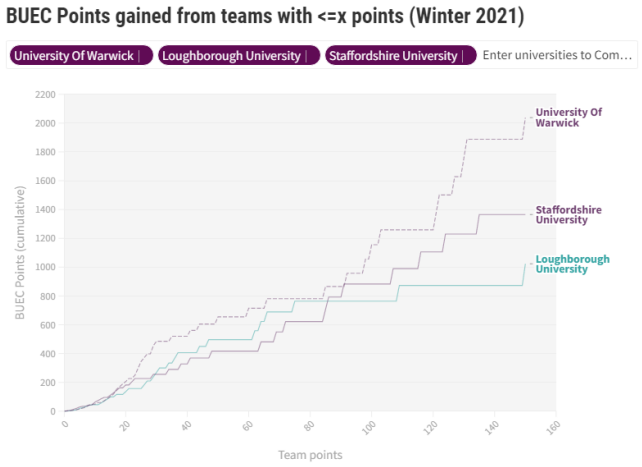
[13]



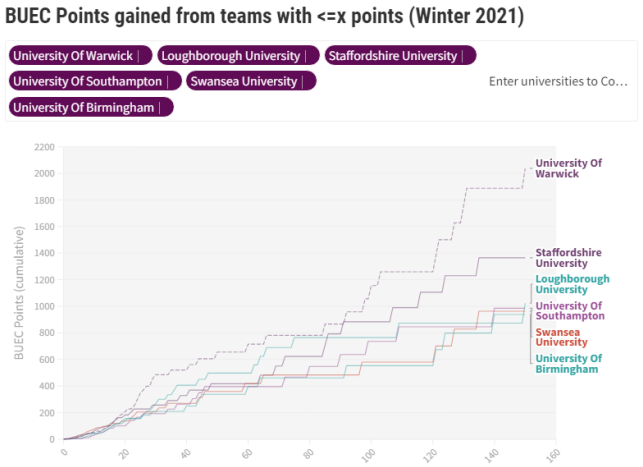
[13]

We can observe that the University of Warwick do exceptionally well in the 121-130 points range.

However, whilst these visualisations are useful for individual analysis, they quickly get confusing and hard to read when making comparisons. So, we use a different type of visualisation for comparisons, where we can plot each university's "race" to the top of the leaderboard, as we slowly include teams with more and more points. We'll refer to these forms of visualisations as "**BUEC Point** races"



[14]



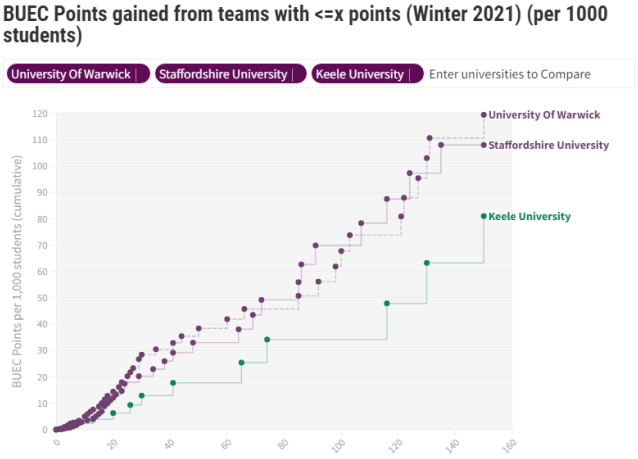
[14]

Most of the top universities don't seem guilty of "Zerg rushing", although Loughborough looks a bit suspicious.

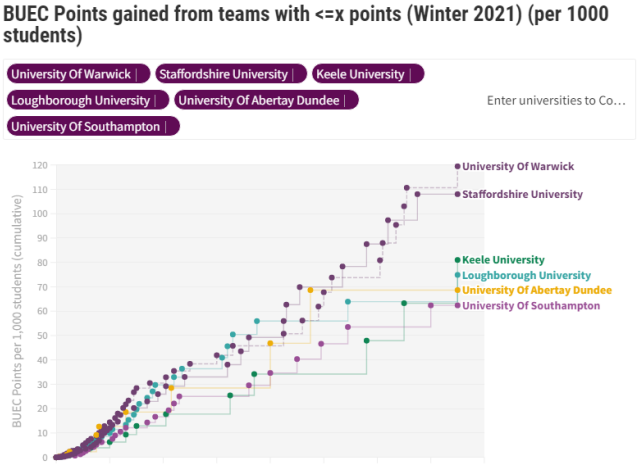
One problem with an accusation of "Zerg Rushing", is that it may not be intentional. A university could just have a very tightly nit and well organised casual gaming community, and so gain a lot of points through them. Alternatively, a university may just be bad, and not actually have any top calibre teams. Nevertheless, it seems

that on large, teams that are ranking at the top of the **BUEC table** aren't getting a disproportionate number of points exclusively from weaker teams (or stronger teams, mind you), and are gaining their points from a mix of both. A notable exception to this rule is the University of Warwick, who gain a similar/slightly number of points from weaker/average teams, but gain an exceptional number of points from stronger teams, compared to other top universities.

We can also adjust these graphs for university size, to get a picture of how that may impact our assessment.



[15]



[15]

When you adjust for size, many of the previous champions fall from the top. Interestingly, Warwick and Staffordshire's graphs look near identical.

None of this is to say that none of the universities at the top don't "deserve" their positions; it is merely to take a look at what got them there. From these size-adjusted results, we can see that for a lot of universities at the top, a significant part of their lead was coming from their size, enough to change the shape of the table quite a bit.

5 The power of institutions

5.1 Hypothesis

Our previous findings have told us that size is an important factor in topping the table, but that institutions also need a combination of many weak/average teams, as well as some superstar teams to carry them over the finish line. So, our final question to ask is what impact an institution can have on the performance of a team. If a team starts the term as the 10th best team in the country, can an institution provide the resources and support to make that team number 1? We hypothesise that a university's ability to do this positively correlates with **BUEC point** totals.

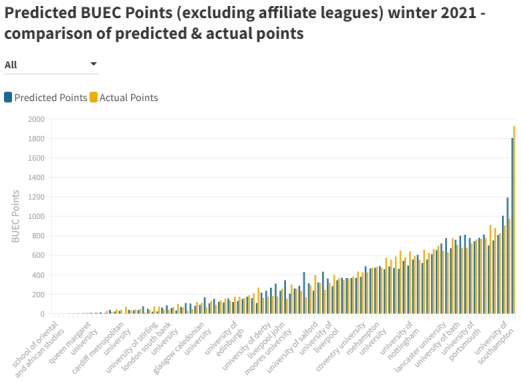
5.2 Results

In order to determine the skill of a team at the beginning of a season, we will be using the **BUEC seeding** data discussed in the data section. We use this seeding data to make predictions for how many **BUEC Points** an institution would receive if their teams performed exactly as their seeding would suggest. (NOTE: we only have seeding data available for the tournaments run by **NSE** themselves, not their affiliate league, so the point total will appear to be less than previous data shown)

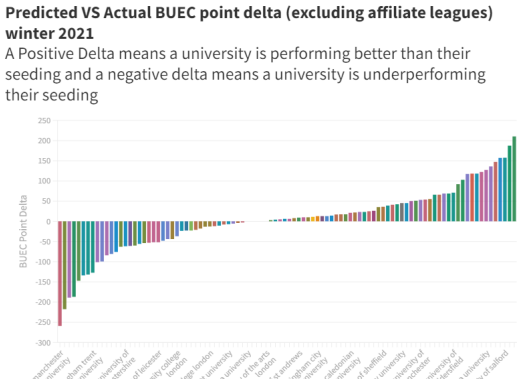
We can then subtract this from the amount of **BUEC** Points the university *actually* received to calculate the **BUEC Points Delta**

There is a fair bit of diversity in **BUEC Points Deltas**, with Staffordshire surprisingly under-performing their seeding by over 200 points, despite coming 2nd overall!

If we run a simple linear regression on the **BUEC Points Delta** relative to total **BUEC points** on the whole data set, we get a modest correlation coefficient of 0.1359. However, again when we restrict to just top 50 this drops to 0.0493 and drops to 0.0334 for top 20. As with before, this tells us that **BUEC Points Delta** does positively correlate with **BUEC Points** on the whole data set, but that this correlation disappears when we are just looking at the top universities.

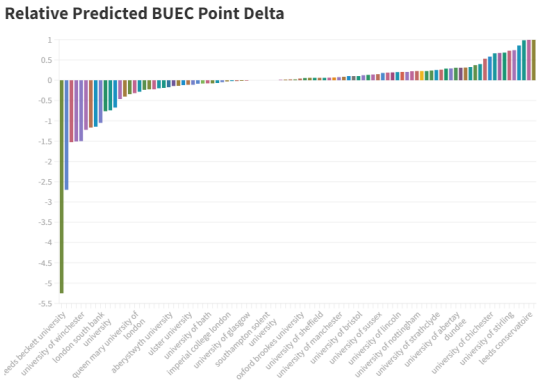


[16]

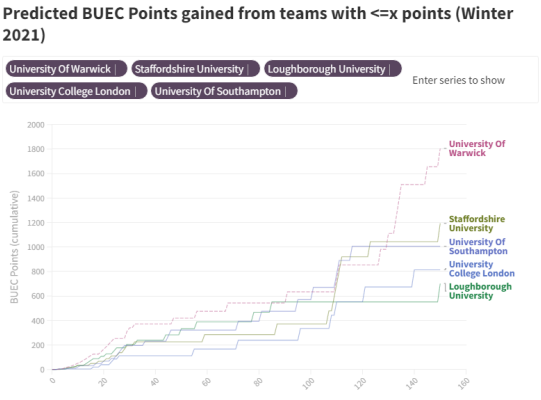


[17]

We can then extend upon this to create **Relative BUEC Point Delta**, which is normalised for how many **BUEC Points** an institution received (so it tells you their percentage increase in **BUEC points**) However, this extension doesn't provide any improvement in correlation. Lastly, with this seeding data, we can create another "**BUEC Point** race", like we did before, but this time with predicted/expected **BUEC Points**.



[18]



[19]

Some universities made an impressive percentage improvement on their predicted points! The predicted points table also looks very different!

The results of this analysis seem to contradict our initial hypothesis. Although the points delta correlated with total points, this wasn't the strongest correlation and more importantly, it disappeared at the top of the table, where we were expecting it to be showcased the most. However this doesn't completely sink the notion of some universities having institutional advantages. It could just be that these advantages kick in before the seeding forms have been filled in, or take longer than one term to kick in, like a year perhaps. This is strongly supported by the fact that over the last 3 years, the university at the top of the table has been Warwick, and the university at number 2 has been Staffordshire. This would suggest that there are some serious institutional factors at play which we simply haven't been able to measure.

6 Conclusions

In conclusion, we have found several relationships between **BUEC Point totals** and a number of other factors, and explored numerous ideas and patterns in the data. To give context to this project, this is by far the largest scale project ever done on UK university esports data, with the only previous work being by William Rice[20], a student from the University of Exeter. Notably, no previous work had been done on data cleaning, and on preprocessing the data sets into more useful representations. In my opinion, that aspect of this project is its strongest aspect, and I hope that this work will enable future, more detailed analysis of the data. The data and preprocessing algorithms can be found here[21].

On the topic of future work, there were several areas found in the project that would benefit from future looking into. By far at the top of that list would be the "institutional advantages" of Warwick an Staffordshire, and what keeps them at the top. It was especially interesting how close their **BUEC Point** graphs looked when you adjusted for institution size. Moreover, examining different classification models for size vs total BUEC points would be of interest, because the relationship appears to be logarithmic. Another area for future work is comparing/finding some way to bring in the previous 3 years of BUEC Data. Lastly, another area to explore would be doing further analysing the data behind the "BUEC Points race" data we used to examine the "Zerg Rush" effect. It is well

worth trying to fit models to those data sets for each university to try and gain some insights.

7 Acknowledgements

I would like to thank **NSE** for being so generous with their data, and their time, in breaking down and explaining any irregularities to me. I would also like to thank William Rice for his feedback on a lot of this analysis, and for his previous related work which inspired this project.

All of the graphics used in this project were made using flourish.com, and the links to all the graphics can be found in the references (These graphics are also very cool and interactive so I suggest you check them out).

References

[1] Global Esports Revenue Reaches More Than 1 Billion As Audience Figures Exceed 433 Million - Forbes

`https://www.forbes.com/sites/jamesayles/2019/12/03/global-esports-revenue-reaches-more-than-1-billion-as-audience-figures-exceed-433-million/?sh=6b0ccbd61329`

[2] The National University Esports League: https://thenuel.com/

[3] National Student Esports: https://nse.gg

[4] British University Esports Championship: https://nse.gg/tournaments/buec-winter-2021/

[5] British Universities Colleges Sport: https://www.bucs.org.uk/

[6] Seeding in chess tournaments: `https://en.wikipedia.org/wiki/Swiss-system_tournament`

[7] BUEC Point allocation:

`https://www.nse.gg/media/15360/points-awarded-british-university-championships-2021_22.pdf`

[8] Higher Education Statistics Agency: https://www.hesa.ac.uk/

[9] Zerg Rush: `https://en.wikipedia.org/wiki/Rush`$(video_games)$

[10] BUEC points vs size (all): https://public.flourish.studio/visualisation/8288965/

[11] BUEC points vs size (top50): https://public.flourish.studio/visualisation/8291323/

[12] BUEC points vs size (top20): https://public.flourish.studio/visualisation/8291290/

[13] BUEC points vs size (top20): https://public.flourish.studio/visualisation/8282459/

[14] BUEC point "race": https://public.flourish.studio/visualisation/8174227/

[15] BUEC point "race" adjusted for institution size: https://public.flourish.studio/visualisation/8194181/

[16] Predicted BUEC Points: https://public.flourish.studio/visualisation/8238442/

[17] Predicted BUEC point Delta https://public.flourish.studio/visualisation/8238730/ (comparable version): https://public.flourish.studio/visualisation/8282364/

[18] Predicted BUEC point Delta https://public.flourish.studio/visualisation/8238920/ (comparable version): https://public.flourish.studio/visualisation/8282422/

[19] Predicted BUEC point Delta https://public.flourish.studio/visualisation/8282430/

[20] Previous work done on this topic: https://nse.gg/news/buec-points-stats-and-maps/ https://nse.gg/news/buec-points-maps-and-stats/ https://nse.gg/news/nse-overwatch-sr-data/

[21] Data sets used in this project: `://drive.google.com/drive/folders/1E6Tc5z`$_t E_x 7ZxG0F4Nej8gM3FlZkqZ5?usp = sharing$