Big Data Analytics of Socio-Economic Impact on UK University Graduates

Jack Morgan  
BSc Computer Science  
University of WolverhamptonWulfruna Street, Wolverhampton, WV1 1LY  
J.M.Morgan@wlv.ac.uk Kieron Ransley  
BSc Computing and Information Technology  
University of WolverhamptonWulfruna Steet, Wolverhampton, WV1 1LY  
K.D.Ransley@wlv.ac.uk

*Abstract*—The following report details how the socio-economic status of major UK towns and cities affects its universities and the students who are graduating from them. University is the biggest challenge that students will face in their academic careers, and for some, it is also a major financial decision. Students coming from more difficult backgrounds, or a lower financial standing may struggle with the challenges they face in this academic environment.

Utilizing datasets from various sources, which were subsequently read, explored, and visualized using a machine learning model, and then further cleaned and analyzed using a predictive model, demonstrates that students who study in an area with lower socio-economic standing tend to perform at a lower level. This argues that not enough support is given to students or the University institutions themselves by local councils due to insufficient funding requirements or substandard facilities surrounding campuses. Further concerns need to be raised with local councils or the wider government to ensure more awareness and support is raised.

Keywords— socio-economic status, graduating, financial standing, academic environment, machine learning model, predictive model, awareness and support.

# Introduction

## Opening

A big step in any academic career is attending a university. As of 2023, there are currently 166 registered Universities in the United Kingdom, with eleven of which occupying a space on the Times Higher Education[1] top one hundred University rankings. University is frequently a significant personal and financial decision for many students. Given that only a select few institutions exist across the country, students often find themselves having to relocate to attend. This relocation, coupled with course fees, can result in a substantial cost. Some students can often find this a struggle and end up with concerns about finances throughout their degree. While there are options such as the Student Loans Company and other government assistance, concerns can still arise due to the substantial costs associated not only with attending university but also with living and thriving in the nearby campus area. While these costs are increased for the nearly 700,000 international students[2] (data accurate as of 2022), It continues to pose challenges for domestic students, which will be the focus of this report.

## Problem Statement

Approximately one million students in the UK earn some form of an undergraduate degree annually. With a further 400,000 more obtaining a postgraduate degree. However, almost 1 in 10 students who attend University drop out, with one-third of those being first-generation students leaving in the very first year[3]. One might attribute a significant contribution to these statistics to the financial struggles faced by university students. In January 2023, a report stated that 41% of students were considering dropping out of university due to surmounting financial worries[4]. University life is stressful as is, with assignments, deadlines, and working towards a career. Adding financial stress on top of that can set students on the edge and prove to be disruptive to their academia. However, how much do these financial issues disrupt students, and are they self-inflicted, or is it a product of the environment and institution?

## Aims and Objectives

This report aims to research, investigate, and comprehend data and statistics related to UK university graduates, as well as the socio-economic status of each university institution’s geographical location. One will attempt to understand if these standings indeed influence graduation rates and the performance of student’s post-graduation.

The report will assess this hypothesis by acquiring appropriate datasets from various sources and importing said data into a deep learning model. The model will then be manipulated to clean the data, and potentially predict any missing data from the datasets. Conclusions will be provided by visualizing the cleaned data in a suitable format.

## Contribution Statement

### Jack Morgan

### Kieron Ransley

## Report Outline

[I. Introduction 1](#_Toc164672789)

[A. Opening 1](#_Toc164672790)

[B. Problem Statement 1](#_Toc164672791)

[C. Aims and Objectives 1](#_Toc164672792)

[D. Contribution Statement 1](#_Toc164672793)

[1) Jack Morgan 1](#_Toc164672794)

[2) Kieron Ransley 1](#_Toc164672795)

[E. Report Outline 1](#_Toc164672796)

[II. Background Study 2](#_Toc164672797)

[A. Literature Review 2](#_Toc164672798)

[B. Summary 2](#_Toc164672799)

[III. Methodology 3](#_Toc164672800)

[A. Methodology Block Phase Diagram 3](#_Toc164672801)

[B. Gathering Data 3](#_Toc164672802)

[C. Setting Up Google Colab Notebook 4](#_Toc164672803)

[D. Error Checking 4](#_Toc164672804)

[E. Data Conversion 4](#_Toc164672805)

[F. Data Summary 4](#_Toc164672806)

[G. Cleaning Data 4](#_Toc164672807)

[H. Visualization 5](#_Toc164672808)

[I. Choosing Learning Model 5](#_Toc164672809)

[IV. Results 5](#_Toc164672810)

[V. Conclusions 5](#_Toc164672811)

[VI. References 5](#_Toc164672812)

# Background Study

## Literature Review

After reviewing various sources of literature, the consensus amongst all writers is that students that come from a lower socio-economic background and standing tend to perform much worse at university than those who come from a higher socio-economic background, and post-education life is of a much lower value as well. An article from Claire Crawford[5] gives one some statistics to backup this information, stating that those from lower socio-economic backgrounds are 3.4% more likely to drop out, 5.3% less likely to graduate, and 3.7% less likely to graduate with a Second Class Degree or better. The literature also states that more students from lower-economic backgrounds are attending Universities, having an increased percentage of members of the public coming from an area of a lower socio-economic standing. In an article by Hazel McCafferty[6], the author explains that even within similar institutions, those from disadvantaged backgrounds fall shorter than their colleagues from better advantaged backgrounds, suggesting that those from higher socio-economic backgrounds find it better and quicker to establish themselves within University, possibly due to better financial funds which can contribute to travel (and thus attendance), living situations and life costs, as well as equipment.

Some sources of literature do not just look at the socio-economic status of university students, but some literature also looks at the performance rankings of university institutions and the socio-economic status of the location of each University campus. In an article by Jack Britton[7] *et al,* it was found that students who attended higher class University would perform better. By looking at average wages and wage premiums of university graduates, those who attended private schools and higher-class Universities earn on average a 10% higher salary. From a previously referenced article by Claire Crawford[8], she finds that students from lower ranking and worse performing Universities are less likely to drop out regardless of socio-economic status. Suggesting that those University institutions offer much easier to pass courses, or understanding their poor performances tend to receive and accept applications from students that come from a disadvantaged background. One would suggest that these University institutions are based in areas with a worse socio-economic standing, but the current articles do nothing to suggest that.

Other articles continue to explore potential other factors and data points to determine what affects student performance. While most authors look at other social factors like sex, ethnicity, or academic factors like course and institution attended, others take a less data driven approach. Jones[9] *et al* find links between non-academic profiles and academic performance, as well as looking at other sources that may display poor socio-economic standing such as if the student qualifies for free school meals. Another article by Claire Crawford and Laura van der Erve[10] looks at family background, understanding that if a student is a first generation does that affect University performance and graduation opportunities. This is another factor that could potentially show socio-economic standing, as those from a lower standing produce the least number of university students and thus graduation, being a first-generation student may be an indicator of this. One especially important article that must be considering as well is understanding why students look to University, and how socio-economic standing can affect outlooks and attitudes towards obtaining degrees. Steve Cook[11] *et al* explore the value of attending University, and often find dissatisfaction in potential jobs post education, rising costs in attending and then lack of financial support both during and post gradation. The article concludes that these factors are most prevalent in lower socio-economic groups and add to division in performance and acceptance in University students across different socio-economic groups. The article once again proves that those from lower socio-economic backgrounds prove harder to succeed at Universities due to factors that typically stack the odds against them.

## Summary

The literature that has been reviewed all tend to acquire their data and draw their conclusions via the same methods. All articles use a mixture of developing graphs for their data inferencing other data from various tables and other figures that have been found via online sources. For example, in Crawford’s *Socio-Economic Differences In University Outcomes in the UK,* the authors uses the HESA database to plot a dataset into a line graph that depicts the drop-out, completion, and graduation rates of higher education students, but in the next chapter plots data of education systems and the time frame of students attended in a table format for better readability.

The authors also do their own share of cleaning the data, removing groups that they find necessary. A lot of the data found seems to include students from courses that have no relevance to the studies, vocational students, or non-domiciled students. A similar process will be taken in our own methodology, cleaning the data to ensure that only UK home students are included in the data as we will be focused on the socio-economic standing of UK towns and cities only. Most articles as well only focus on full-time students and courses, one way in which our report will differ is that one wants to include those that are part-time, as we have the hypothesis that students from a lower socio-economic background would prefer part-time courses due to better financial stability in securing additional funds through part-time work.

Another way one’s studies will differ from those previously completed is by using data surrounding the socio-economic standings of UK towns and cities using local council data or nation-wide government data. One of the original hypotheses for the report was suggesting that socio-economic standing of the area that a university institution resides in could impact student performance. One believes that lack of funding, lack of acceptable facilities, and overall perceptions and environmental influence of the surrounding area could be one of many reasons for deficient performance amongst students. While similar ideas are touched on in some articles, there is no data presented to support this.

None of the articles reference any predictive learning algorithms deployed either. Most articles were able to strictly employ machine learning models to clean and visualize the data at hand. One believes this because that all datasets found for the subject area contain completed and efficient data, so no additional data is needed to be predicted to complete the report. One could additionally employ predictive algorithms to predict trends in later years based on current data supplied, this way the current report will differentiate from others by seeing if current trends continue, attempting to preemptively conclude how socio-economic standings affect future student performance.

# Methodology

## Methodology Block Phase Diagram

A diagram of a company

Description automatically generated

Figure I: Methodology Block Diagram

## Gathering Data

The first phase of the project was to gather and download various datasets that were available online that are relevant to the project at hand, which is finding any data relating to socio-economic status of UK areas, be it by county, city/town, or even local council constituencies, and also data relating to graduate rates, employability rates of graduates, or any data relating to graduates for both during and post study. The datasets we collected were from reputable sources, ones such as Nomis Labour and Census Statistics, National Office for Statistics, and any other sources from sites similar, like Kaggle.

There were some assumptions made before the beginning of the project and before the gathering of our data. The first assumption was that every different University institution in the same area would be similar in its status, funding, facilities, etc., and therefore will produce comparable results. As we are checking socio-economic standing, we are assuming that the socio-economic standing will affect each school in the area equally, and thus each University will product similar graduate data. The second assumption is that most people that the data captures are people with university degrees, and that anyone who did any form of higher education did so at a university and received an undergraduate level degree. Paths like apprenticeships, degree schemes, HND, or other college level degree’s may be present in the data as receiving a degree, and not presented seperately or omitted altogether. In some cases, this might simply be a case of cleaning out the dataset using tools, or simply keeping the data but being wary of its presence when providing conclusions from the datasets.

Once datasets were selected, they were downloaded and exported to a suitable format, for this project we opted for a CSV format. The datasets were quickly screened and checked, checking to see if the datasets were indeed visualizing data we were looking for, as well as testing for any issues that may affect any of the process, such as additional text in some cells, bad record titles, other errors in the data presentation. One error that was common was placing a comma amongst larger numbers to separate higher numeric columns. This proved difficult when converting data types and accurately reporting values, as the tools did not see those records as numeric value, and often would block conversion into an integer or a float value.

## Setting Up Google Colab Notebook

The next phase of the project was setting up an environment in which to evaluate, modify, clean, and visualize our dataset. The environment chosen was Google Colab, which is a free, online, resource with collaborative capabilities, and links to sites such as GitHub for version control and sharing. First, it was necessary to import necessary Python libraries so that one could commence work on the data, such as NumPy and Pandas for analysis, Matplotlib and Seaborn for visualization, and some of Google’s own libraries to enable connection between our devices and the online virtual environment Colab is hosted on. Next, we were able to import our CSV files and save them within the environment in their own tables for further use. This is where Google’s Python libraries are necessary, allowing for the ability to upload files from a local machine to the online environment. This was essential to the project as it was one of a collaborative effort, and thus each of us would have our own files, each on our own separate machines.

## Error Checking

The next phase of the project was to check the imported datasets for any errors that may have occurred during the importing process. This process may cause issues such as changes in the way data is presented, such as format or layout, or issues in the data itself, missing values or corruption could occur so that the data we receive after importing could be incorrect. One way in which we completed error checking is by using a function to display all record columns in each dataset. From this, we were able to see that at some point during the process of gathering the data and altering the raw CSV files ready for processing, additional columns were added that were empty and unrecognized by the environment and thus had to be dropped. Once one was happy there were no more errors, further phases commenced.

## Data Conversion

The fourth phase of the project is then converting the data types of all the records in the imported CSV files into ones that are appropriate for what the data is conveying. During the import process, all values are automatically created as an “object”, and thus will need to be converted into a data type that is more fitting. This is especially significant, as when it comes to visualizing the data, and especially later when selecting a learning model, the data will need to be a certain type to effectively execute any functions. To start out, the entire tables are converted immediately into “strings”, this is because objects are most compatible with strings, and trying to convert an object into any other data type can cause problems to occur. String is also an appropriate data type for a fair amount of our records, such as place names, course names, qualification types, etc. Especially when thinking later about visualization, strings are particularly beneficial for creating labels out of the data. Further conversions are then converting some strings to either “integers” or “floats”. Numeric values such as amounts or scores will need to be presented in this data type, as some numeric operations will need to be conducted such as summary or averaging to find totals or patterns in the data. Again, when it comes to visualization, numeric values will be preferable for some axis, especially if we are to create bar charts or plot diagrams.

## Data Summary

The next phase of the project was to conduct a brief check on the available data, this serves as a summary of the data that is presented to refer to at any moment of the next phases, as well as a secondary check for any errors that may have slipped past the initial check for errors. To do this, one, simply used functions from the Pandas library, .describe() and .head(). These two functions gives one a successful overview of the data, from information of the data frame itself, such as the number of the non-null values and maximum, minimum, or average values, but also a look at the top 5 records in the data frame for sufficient reference on what each dataset is representing, and a decent look at how the data is presented in each set which will be quite necessary when it comes to cleaning the data.

## Cleaning Data

The sixth phase of the project was cleaning the data. This section is a bit different from the error checking or data conversion sections, in that rather than removing data that is incorrect, or corrupted or wrong, data cleaning is more for removing data that in unneeded, unwanted, or reformatting the datasets into a format that is much more sufficient and would provide better and clearer presentation during the visualization section of the project. The first technique that was employed for cleaning the data was the .drop() function, which allows one to remove specific columns from the data frame. During the summary section, one was able to highlight certain columns that we could deem as unnecessary, which contain data that is not relevant to the project goals and will not be needed to create our results. Therefore, they may as well be removed as then the data frames are smaller, contain only relevant data, and thus will make any further experiments conducted on each data frame much quicker as there is less data to process. An example of using technique is on the dataset that provides data on graduate outcomes, which contained columns on data such as the sex of the graduate, the country of provider, and the academic year. We dropped the sex columns and the sex of the graduate was not an area of contention for our research, only socio-economic data, and we dropped the country and year columns, because during the stage in which we researched and gathered datasets, we already knew this dataset contained only information from the 2021/22 academic year, for countries in the UK. Meaning, these columns represented data that we were already aware of, and presented no value, so it is a valid argument to simply remove it.

Another technique that was employed to clean the data was by removing records that again, presented no value to use and had no relevance to the aims of the project. Similarly to the reason for dropping some of the columns in the data frames, certain records contained data that has relevance, or contained data that was represented already and thus was unnecessary and could be removed for better efficiency. An example of this was again in the dataset that represented data on graduate outcomes, in which there were records that detailed the total of each different outcome. However, included in the data frame already was a column that included this number, so these records were unnecessary, and could be removed. The way in which this technique is achieved is by making a mask, which is a variable that stores the results of a query. The query itself is a search query of the records that is to be removed from the dataset. Then, to remove the records the data frame is then resaved to the environment, but using the operation symbol “~”, the data frame is saved without the contents of the mask.

Often in the datasets, certain columns have various unique records that end up meaning the same and can often be grouped together. For example, displaying “full time employment” and “part time employment” seperately, or separating “full time education” and “part time education”. For the purposes of the project, and the eventual conclusions that will be drawn, this is not necessary. Instead, it will be preferable to just have employment and education, rather than distinguish between the types of employment and education. For this, we then changed all examples of this split to just display the same text. To achieve this, we employed the .replace() function and a combine technique. Using the for loop, one looped continuously through the data frames, and then when the loop encounters a certain record, we employ the .replace to change that into a different record. For example, in the dataset that contained socio-economic data of areas in the UK, there was a column that displayed the occupation of a person. Especially when it came to employment, the dataset had various terms for how a person was employed. As we were not concerned with these terms, instead we replaced them all simply with the term “Employment”. This becomes much easier for one when visualizing the data, as rather than having separate points or bars for all the various terms, they are grouped together, and we achieve one point/bar just for the overall group.

The last technique one employed to clean the dataset was by reformatting the table entirely. One dataset we acquired gave us a data frame that detailed all the different qualification types for areas of the UK. However, the dataset in its original state has a unique column for each different type of qualification. This is unnecessary, and like other issues, makes it less preferable for visualization. Therefore, we aimed to completely change the layout of the data frame. To achieve this, we used the .stack() function. To purpose of the .stack() function is to covert columns into rows by reformatting the data frame into a long format, rather than a wide format. So, each different qualification type is now a unique record in a column titled “Qualification Type” rather than its own individual row. Pairing this function with another function called .to\_frame(), then loops through the data frame, and adds each number in each record for the various columns together into a new column titled “Total”. This now represents the data frame, so each qualification type in a new row per UK area, rather than a unique column.

## Visualization

The second to last phase of the process was to then visualize all the cleaned and finalized data frames. The process of visualization provides a much clearer overview of all the data, and thus will help to draw conclusions from the data as it is presented in a more readable and efficient format. Using the two visualization libraries that were set up in a previous section, this process was easily achieved. To produce some of the visualizations, further changes had to be made to the data frames as and when it was obvious it would be required. Most examples involved grouping data or renaming some columns to provide a much clearer and descriptive graphic. Other times, visualization involved conducting certain queries on one data frame, and the creating a graphic on another data frame that reflects this data. An example of exactly this will be described later in this very report.

## Choosing Learning Model

The last phase of the methodology was employing and selecting a machine learning model to predict unknown values of data. Fortunately, the datasets that were found and used for the data process contained no null values, which we checked for as part of the importing process. However, there were many records that contained zero values. The assumption that we made here was that there was simply no data in these records, as there was not a sufficient response from whichever organization created the datasets, rather than the values being a zero value as zero would be a valid representation of the data. Nonetheless, one tested various MLM’s to assess which would most accurately predict these zero values. To achieve this, one first mapped numeric values to certain string values. For example, for a column in one data frame that recorded different activities of graduate’s post-graduation, each activity was mapped to a number (i.e.: employed became 4, further study became 3, etc.). The reasoning behind this is because learning models work more efficiently with numerical data than text data, and one would receive a more accurate rating using integers rather than strings. After mapping, one would create a separate version of the data frame specifically for processing with the machine learning model. Here, we would split the data frame into its target values, and the values in which the model will use to predict the data. At the same time, we will also decide the size of the test, and the randomness of the data that will be chosen for the model. Finally, a library was imported for each of the chosen learning models, the data was run through each model seperately and compared to evaluate their accuracy. The scores were then printed out in a table, with each record containing the accuracy percentage and the name of the model. This table is created to report which models were most accurate and allow one to quickly select the model that was most accurate.

# Results

From the datasets given, we can concur that a higher concentration of students go into employment post-graduation than any other activity. However, there is also a higher concentration of graduates that become unemployed and go back into education. This potentially suggests that some students may find it hard to enter the job market post-graduation, and the lack of further study could be due to the costly value of completing a post-graduate study. This initial finding gives us a baseline for our hypotheses, giving us the knowledge that there is a high percentage of students already going into employment that any other post-graduation activity. Going forward, we can begin to assess where the lack of employment comes from and evaluate our other hypotheses.

A graph of different colored squares

Description automatically generated with medium confidence

Figure 2: Total Number of Graduate Outcomes by Activity

Another set of results also confirmed another one of our hypotheses, detailing the concentration of different group types throughout differently ranked areas on a socio-economic scale. The below graphic excellently details two varied factors, concentration and minimum and maximum values. From the graph, one can confirm that there is a higher concentration of the population going into employment in higher ranked socio-economic areas of the UK, whereas there is a high concentration of the population in the lower ranked socio-economic areas of the UK ending up in unemployment. This is evidenced with the thicker portion of the graph for each activity spiking along the bottom axis which displays the ranking total associated with various area of the UK. As clearly shown, for the employment activity, the thickest portion is from around 0.25, and on the other end the thickest portion for the unemployment activity is around a rating of 0.05. The graph also shows the minimum and maximum records found in the data frame; from this we can also confirm our hypothesis. The graph confirms our original theory that higher socio-economic areas produce better graduate outcomes and better value for graduates, as one can see that for the unemployment activity, records cease post 0.3 on the SE ranking scale. This tells us that those from a higher ranked socio-economic area tend to find better opportunities post-graduation, and thus do not fail to find employment opportunities. We can assume that this is due to higher socio-economic areas providing better resources, better conditions, and better-quality university opportunities for the population, and offer better support for graduates.

A diagram of a green and orange leaf

Description automatically generated with medium confidence

Figure 3: Population Groups Ranked on a Socio-Economic Scale

# Conclusions

# References

1. World University Rankings (2023) Times Higher Education (THE). Available at: https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking (Accessed: 26th February 2024).
2. Gent (2023) International Student Statistics in UK 2023, Study in UK. Available at: https://www.studying-in-uk.org/international-student-statistics-in-uk/#:~:text=Statistics%20show%20that%20the%20total,outside%20of%20the%20European%20Union. (Accessed: 26th February 2024).
3. Tagliaferro, L. (2023) *Graduation statistics and Fun Facts in 2024*, *Future Fit*. Available at: https://www.futurefit.co.uk/blog/graduation-statistics-and-facts/#:~:text=Key%20Uni%20graduation%20statistics%20in%20the%20UK&text=22.6%25%20of%20British%20adults%20attained,University%20within%20the%20first%20year. (Accessed: 04 March 2024).
4. *41% of students consider dropping out due to money worries as new term starts* (2023) *FE News*. Available at: https://www.fenews.co.uk/education/students-consider-dropping-out-due-to-money-worries-as-new-term-starts/ (Accessed: 04 March 2024).
5. Crawford, C. (2014). *Socio-Economic Differences in University Outcomes in the UK*. Institure for Fiscal Studies. DOI: 10.1920/wp.ifs.2014.1431.
6. McCafferty, H. (2022). *A Systematic Review of the Employability Perceptions of UK Undergraduates from Disadvantaged Socio-Economic Backgrounds*. Research in Post-Compulsary Education. Volume 27. Pages 570-593. DOI: 10.1080/13596748.2022.2110774.
7. Britton, J., et al. (2016). *How English Domiciled Graduate Earnings Vary With Gender, Institution Attended, Subject, and Socio-Economic Background.* Institute for Fiscal Studies. DOI: 10.1920/wp.ifs.2016.1606.
8. Crawford, C. (2014). *Socio-Economic Differences in University Outcomes in the UK*. Institure for Fiscal Studies. DOI: 10.1920/wp.ifs.2014.1431.
9. Jones, S., et al. (2017). *Contextualising Degree-Level Achievement.* Research in Post-Compulsary Education. Volume 22, Issue 4. Pages 455-476. DOI: 10.1080/13596748.2017.1381287.
10. Crawford, C., Van der Erve, L. (2015). *Does Higher Education Level the Playing Field?* Education Sciences. Volume 5, Issue 4. Pages 380-412. DOI: 10.3390/educsci5040380.
11. Cook, S., Watson, D., Webb, R. (2018). *Investigating Perceptions of the Value in Attending University.* Studies in Higher Education. Volume 44, Issue 7. Pages 1256-1267. DOI: 10.1080/03075079.2018.1434616.
12. (No date) *Methodology section for research papers*. Available at: https://www.sjsu.edu/writingcenter/docs/handouts/Methodology.pdf (Accessed: 17 April 2024).
13. *Research guides: Organizing your Social Sciences Research Paper: 6. the methodology* (no date) 6. The Methodology - Organizing Your Social Sciences Research Paper - Research Guides at University of Southern California. Available at: https://libguides.usc.edu/writingguide/methodology (Accessed: 17 April 2024).