Introduction

The United Kingdom (UK) Government invests millions of pounds each year to improve and

maintain the quality of higher education in the UK, in fact, based on HEFCE (Higher Education

Funding for England) \Guide to funding 2017-18" published report [HEFCE, 2017], a total of

3,536 million pounds has been allocated for academic year 2017 { 18. This allocated money

supports funding for teaching, research, capital grant, knowledge exchange and national facilities

and initiatives, with 45.1% of the funding, a major chunk amounting to 1,595 million pounds,

goes to research. These amounts will then be distributed into HEFCE-funded higher education

institutions (HEIs).

It is therefore imperative to monitor and measure the e\_ectiveness of research grants al-

located to HEIs. To achieve this, an overarching framework for both assessment and funding

of research has been developed, called The Research Excellence Framework (REF). The REF

was created following the UK Government's announcement in December 2006 [Eastwood, 2007]

to replace previous framework, the Research Assessment Exercises (RAE) which was last used

in 2008. In short, REF is a system designed to analyse and evaluate the quality of research

at various HEIs based in the UK and allocate funding based on which HEIs display the most

outstanding quality of work [REF, 2014a].

The REF is conducted every seven to eight years, with the last REF having been conducted

in 2014, where a total of 191,150 datasets had been submitted by HEIs [REF, 2014a]. The

submission period generally opens 14 months prior to the assessment exercise, where the REF

committee will invite HEIs to submit their datasets into 36 di\_erent disciplines segregated

by units of assessments (UOAs). For the 2014 REF, after the submission period had ended,

the expert review process was then conducted by the four main panels throughout 2014, from

January to November, and concluded with publishing of REF results in December 2014. This

process in whole took 26 months to complete.

Project Aim

In view of the lengthy processes and estimated high man-hours required to conduct REF assess-

ment exercises by HEIs and REF committee, a team of \_ve has been tasked to systematically

analyse REF 2014 data. By knowing what factors do or do not contribute to getting a out-

standing REF rating score for HEIs, a more focused application of man-hours on factors that

increase the REF score could be taken by the HEIs.

In summary, this project aims to discover potential features that could a\_ect the scores given

by the REF, based on REF 2014 data, by using exploratory data analysis [NIST SEMATECH,

2013]. This would allow the team to not only \_nd which features a\_ect the REF score, but also

which features might not have had any signi\_cant impact on the overall rating score.

Literature Review

Existing Policy and Study Done on Selective Staff Submission Process

Existing policy has been developed by REF on Equality & Diversity (E&D) to ensure staff selection was done on equitable manner (<http://www.ref.ac.uk/2014/equality/>). In fact, a part of the submission guideline was dedicated to explicitly mentioned the principles of staff selection, and HEIs was requested to produce code of practice and be transparent in the process. (<http://www.ref.ac.uk/2014/equality/codesofpractice/>)

On top of the existing policy and submission guideline, an investigation by Higher Education Funding Council of England (HEFCE), which is one of the funding bodies behind REF, where they analysed biases in terms of quantitative measures, such as disability, age, sex, ethnicity, and nationality from processes of staff selection. However, the report does not include scope of investigating whether there were biases in terms of research quality from this process alone. (<http://www.hefce.ac.uk/pubs/year/2015/201517/>).

Methodology

Feature 6 – Selective Submission of Output

Higher Number of Outputs Per Institution - Initial Hypothesis

Feature 6 was selected based on outcome of the initial checking of linear relationship between the number of outputs submitted per institution versus percentage of outputs with four stars ranking for UOA 11 (Computer Science and Informatics). The initial analysis intent was to see if there’s any impact of higher submissions by institutions to the outcome of the ranking.

However, the original plot done in R Studio revealed an institution which perceived to be good ranking was having low rating of four stars (between 0 to 10%), although the number of submissions was high (> 400 outputs), whereas there were another two institutions with lower submissions count (between 200 to 300 outputs), but have higher four stars rating (between 40 to 50%).

Though the initial analysis was off, this triggered a hypothesis of institutions’ practice of being selective on the outputs, and whether this could lead to a higher output ranking upon weird observation of the plot.

Selective Submission of Output – Current Hypothesis

Checking for this hypothesis required additional dataset outside of existing REF 2014 data as the current submitted Full-Time Equivalent (FTE) data for Category A was not sufficient to prove this hypothesis. Instead, the data used was from HESA, which was given to REF committee during REF 2014 assessment exercise. The calculated FTE data by HESA using internal records approximated the eligible number of staff that were supposed to be returned by HEIs. The calculation was further scaled due to the differences of definitions between HESA and REF. (HESA, 2014).

The analysis was first done by checking the distributions of the submitted FTE of Total Category A by HEIs and Scaled FTE provided by HESA, using Shapiro-Wilk normality test in R. Upon checking the distribution, since there was not enough evidence that both were from normal distribution, then the test proceeded with Spearman correlation method, instead of Pearson. Since three variables were involved, a 3D plot was attempted between “Eligible FTE”, “Four Star Output”, and “Number of Submissions Per University” to see how these two variables could have affected four stars ranking. Then, a multiple correlation coefficient was calculated.

The same processes were replicated for all UOAs, with minor differences on pre-processing part for certain UOAs, where the outlier was removed to check if it will improve the correlation. However, five UOA's were not able to be pre-processed due the some of the output submitted by HEIs were marked as two submissions, instead of one. Therefore, this introduce two different ranking instead of one.

Prior to this current hypothesis analysis, the initial hypothesis was again checked and written properly in R Notebook with different scripts produced for data wrangling and data analysis. This was again confirmed by having similar checking on three stars ranking and average scores of three and four stars ranking.

Higher Number of Outputs Per Institution – Future Hypothesis

However, upon re-checking the result of data produced by higher number of outputs versus four stars ranking produced by R, and compared against similar data produced by Python code written in Jupyter Notebook, the output was stark different although the methodology was similar. This need to be investigated further.

**Results**

Feature 6 – Selective Submission of Output

[plot from original analysis]

Figure xx was the original plot using R Studio that driven the hypothesis to be the current hypothesis. It showed the populations of the data between number of outputs submitted versus four stars ranking for UOA 11.

[Plot from four stars and three star and average of three \& four stars]

Figure yy was the amended plot using R Notebook for re-verification of the original analysis done in Figure xx. It showed the populations of the data between number of outputs submitted versus four stars ranking for UOA 11.

[Table – correlation results from different UOAs]

Table zz was the correlation results for five UOAs for the initial higher number of outputs analysis that driven the current hypothesis. These results confirmed that the number of outputs were not

However, upon detail checking of the data consistency between the two spreadsheets, there were discrepancy between the plot and the actual values in the tables.

[include tables]

For example, institution that has 401

UKPRN that has 401 submission outputs

**Discussion**

Feature 6 – Selective Submission of Output

# confirming the hypothesis