Snappy title goes here

Charities, funding, and Twitter

|  |  |  |
| --- | --- | --- |
| Natalie Polack University of Stirling npp00002@students.stir.ac.uk | Tom Wallace University of Stirling tom.wallace@stir.ac.uk | Vikki Richardson University of Stirling vip0001@students.stir.ac.uk |

ABSTRACT

T

KEYWORDS

Charities, Twitter, funding, data.

1 INTRODUCTION

This research paper explores factors that predict charity use of Twitter. Twitter is a social networking platform that is commonly used by charities but remains understudied academically (Obar, Zube, & Lampe, 2012). Charities have many uses for Twitter, including advertising charitable campaigns, interacting with their users or supporters, and even communicating directly with other charitable organizations. Twitter has a relatively low cost of use, only requiring staff time. This may present an opportunity for smaller charities, who are often limited by costs, but larger charities are more likely to have the capacity to dedicate staff to social media and so are still, likely, at an advantage. Our study attempts to predict various aspects of charity Twitter usage (such as overall use, popularity, and networking activity) based on publically available data on charities features (such as funding source, overall income, who the charity helps, and number of staff). This leads to the following research questions:

**RQ1**: Is funding source, the extent to which a charity is funded by the government or the public, related to the use of Twitter?

**RQ2**: Are charities which seek to help the public more popular on Twitter?

**RQ3**: Does number of staff, rather than size, determine a charity’s active use of Twitter?

This study is important because there is a lack of published research on the factors which relate to successful use of social media by charities. Twitter has huge potential for charitable organisations and knowing which factors contribute to success or failure on the platform could help implement measures that lead to more of the former and less of the latter.

The next part of this paper give a brief summation the research questions and data management tasks. Section 2 then describes pervious literature and research in this area. Sections 3 and 4 detail the data processing and analysis methods used before the results are discussed in Section 5 and the paper is then concluded.

1.1  Data sources and management tasks

Three data sources were required to undertake this research. To obtain data on the outcome, Twitter use by charities, we downloaded data directly through the Twitter API (2018) for the charities in our sample which had a handle. For data on charity features, we scraped the UK Charity Commission website (2018) which holds a record for every registered charity in England and Wales (Scotland and Northern Ireland are covered by separate regulators). Given these two data sources, we could sample an arbitrary number of charities, right up to the full population of ~168,000 that are registered (The Charity Commission, 2018). However, the Charity Commission website does not record detailed financial information for charities which is important for RQ1. Therefore, we make use of data collected by Alcock and Mohan of the Third Sector Research Centre (2017). This dataset is based on surveying a random sample of registered charities and contains detailed financial information, such as funding source. As this dataset is the smallest, it forms our sample and we use it as a ‘mask’ to collect data from the Charity Commission and Twitter on this subset of charities. The Alcock and Mohan data gives us a sample of 12,150. Each of the three data sources needed to be interacted with in different ways and presented different data management challenges which are detailed in Section 3.

2 RELATED WORK

2.1 Literature and related work

Twitter use by charities is common. The Lloyds Digital Index found that 44% of the, more than one hundred thousand, charities they surveyed in 2016 were using social media in some form (Lloyds Bank, 2016). Of these platforms, Twitter is by far the most used platform for charitable organisations (Guo & Saxton, 2014). But what do charities use Twitter for? A common use case is broadcasting one-to-many messages to followers to convey information or solicit donations (Phethean, Tiropanis, & Harris, 2015; Waters & Jamal, 2011). Another use of Twitter is networking activity between charities, either because they share a common purpose or for support. Infrastructure organisations, which are charities established to help support other charities, can use Twitter to share links to information, resources, and training (Dayson & Sanderson, 2014).

Overall, the literature, which has been published so far, seems to concur that social media use helps make a charity successful (Lloyds Bank, 2016; McCabe & Phillimore, 2012). Therefore, it is important to understand what factors make a charity an active user of social media and this has largely been neglected by the literature thus far.

3  DATA PROCESSING & ANALYSIS METHODS

3.1 The Alcock and Mohan data

The Alcock and Mohan data was sourced from the UK Data Archive. It contains detailed financial information for charities in England and Wales collected via a survey. The data is longitudinal and contains information from the financial years 2006/07 up to 2013/14. For this project, it was decided to use data from one financial year. The year selected was 2011/12, as the information for the more recent years was less complete and would have provided a smaller sample.

The original data, ‘CharityCharacteristics.csv’, was imported into Python as a Pandas data frame. Extraneous years were then dropped from this dataframe to leave only data from 2011/12 which resulted in a sample of 12,150. Charity number was used as the index for this dataframe and is used throughout this project as an ID number for linkage as it uniquely identified each charity.

The financial details in the Alcock and Mohan data are split between a large number of variables and the names used in the raw data are not particularly informative. Therefore, income was renamed as ‘Income2011-2012’ to distinguish it from up-to-date information collected via a web scrape. A ‘government funding’ variable was then created by combing the following individual variables: ‘funds government sector’, ‘funds central government’, ‘funds local government’, ‘funds regional government’, ‘funds EU government’, ‘funds international government agencies’, ‘funds foreign governments’, ‘funds devolved government’. ‘General public funding’ was contained within a single variable and did not need to be combined.

Two new variables were then created to assess the proportion of each charity’s funding that came from the government or general pubic. This was achieved by dividing each charities government and public income by their total income.

After creation and renaming of the variables of interest, the data was filtered to create a smaller dataframe containing the 12,150 charities and 6 variables: ‘Charity number, 'Income2011-2012', ‘government funding', General public funding', 'Proportional government funding', and 'Proportional general public funding'.

3.2 The Charity Commission data

It was desirable to join the Alcock and Mohan data to data scraped from the Charity Commission website for two reasons. Firstly, it contains updated income data for 2018, which allows growth measures to be calculated, and secondly, it contains variables not found in the Alcock and Mohan data. The new variables found on the Charity Commission website include ‘staff’, ‘who the charity helps’, and ‘website’. ‘Staff’ is a simple integer variable which is found in the Alcock and Mohan data but is unreliable, the version in the Charity Commission data represents fulltime equivalent employees and is collected from account returns, so is more robust. ‘Who the charity helps’ is a categorical variable which charities specify when they register from a dropdown list (it is not free-text). Example categories include ‘Children/young people’, ‘Other charities’, and ‘The general public/mankind’. We are interested in the latter, which is formatted as a binary variable in the final data. ‘Website’, where provided, is an URL to the charities own website. This was used to obtain Twitter handles as detailed in Section 3.3. Several other variables were collected during the web scrape which were not used in this research but could be employed by wider projects. These include ‘Number of volunteers’, ‘Number of trustees’, ‘Expenditure’ and ‘Company number’ which could be used to link to Companies House data in the future.

The web scrape was performed using the BeautifulSoup (**REF**) package in Python. The Charity Commission’s website is structured so each charity’s record is identified by their unique charity number. This made it easy to loop over each charity in the Alcock and Mohan data as this contained charity numbers. On each loop data was collected on all of the above variables for a single charity and these were combined into a dataframe which was saved as a JSON after all of the charities had been scraped. Error handling was built into the scraper to make it robust to missing data.

3.3 The Twitter data

Gathering Twitter metrics for the sample involved a two stage process, first collecting handles for charities which have them via a web scrape, and then gathering metrics for these handles through the Twitter API.

Obtaining Twitter handles for charities

In order to obtain the Twitter handles for each charity it was necessary to scrape each charity's website using the Python library BeautifulSoup to parse HTML. The JSON file which resulted from the web scrape in 3.2 contained charity’s websites (where listed) and was imported as a Pandas dataframe. Each of the URLs was then looped over with BeautifulSoup attempting to find hyperlinks on the charities homepage to Twitter.com

These hyperlinks formed a list which was, in turn, iterated over to find relevant Twitter links for the charity. Some websites contain links to twitter.com for reasons other than linking to their own Twitter account and this should be accounted for. If the link contained certain keywords such as 'search', 'home', 'archives', 'tweet', 'share', 'intent' or 'Twitter' this was likely to point to something other than the charity's Twitter handle and was dropped from the list. Once a relevant link to the charity Twitter handle was found the iteration was escaped and the Twitter handle extracted from the link using regular expressions. If no relevant Twitter handle was found for a charity it was given a missing value. This was assigned to a new variable: 'Twitter Handle'.

Twitter metrics

Once all the Twitter handles were found the Python Tweepy library (developer.twitter.com) was used to interact with the Twitter API and extract metrics.

Similarly to the Twitter handle collection method, a dataframe was created and the Twitter Handle list was iterated over and passed to API. Tweepy handles rate limiting internally which proved fortuitous as the Twitter API penalises users who ignore the limits. Getting the metrics was a two stage process; for each Twitter handle found Tweepy attempted to identify the handle. Once this was found, metrics could be easily obtained with inbuilt commands. These were returned as Pandas series with missing place holders added for any failed users or where there were instances of having no Twitter handle to check for. By the end of the iteration four variables had been created ‘Twitter Handle’, ‘Has Twitter’, ‘Twitter followers’, ‘Twitter following’, ‘Number of tweets in total’.

Table 1. Charity handle ownership

|  |  |  |
| --- | --- | --- |
|  | **Has Twitter** | **Does not have Twitter** |
| **n** | 4875 | 7275 |
| **%** | 40% | 60% |

We tested our scraping methods against a manually collected sample of 155 Twitter accounts. This sample was collected by individually searching Twitter and Google for each of the 155 charity’s Twitter accounts and is 100% accurate. We compared the results of our scraper with this list of accounts and found that our method correctly identified 67 of the 155 accounts or 43%. While this figure is somewhat low it should be noted that these are all false negatives and our method did not return any false positives, which would introduce bias. Given we sample thousands of charities the relatively low accuracy of the scraper should not be problematic for our analysis as charities are expected to be missing at random.

3.4 Combined data

Once the Alcock and Mohan, Charity Commission, and Twitter data had been collected, managed, and stored as JSON files, the final task preceding analysis was to combine these different sets of data into one file. This task was eased by the inclusion of charity number in each set; data was only collected from the Charity Commission and Twitter for charities in the Alcock and Mohan data and charity number was carried through all datasets as a unique identifier. Use of missing markers also mean that every variable within each dataset was the same length. This meant that combining the data was simply a matter of merging using charity number. There were no duplicates and no fuzzy matching was necessary.

With the data combined, some final data management tasks were performed before analysis. Firstly, absolute and ratio funding growth variables were created. The former is simply income from the updated Charity Commission data minus the income in the Alcock and Mohan data, for each charity. This results in the change in funding for charities between 2011-2012 and 2018, for those present in both datasets. The funding ratio was created by dividing income from Charity Commission data by income from the Alcock and Mohan data.

With these new variables created, the final task was to deal with outliers. Funding and Twitter data are both commonly outlier heavy and this could unduly bias our analysis. A function was created which removed data points which were more than 1.5 times outside the interquartile range for each variable. The data removed was set to missing. Testing revealed that this method of outlier removal was more appropriate than using a standard deviation based method.

4  ANALYSIS METHODS

This section briefly details the statistical methods we applied to our data. Starting with basic descriptive methods and culminating in a discussion of our modelling techniques.

4.1 Univariate methods

The focus of our analysis is regression modelling, therefore we only use brief univariate methods to describe the data before analysis. These methods include statistical summaries for metric variables, which show the distribution, mean, median, and extreme points of the data. Histograms are also used in this endeavor, visually showing the distributions of the metric variables such as income, number of Tweets, etc. For our categorical data, we use one-way tables, such as Table 1 already shown.

4.2 Bivariate methods

Bivariate methods fill a similar role to the descriptives, they describe relationships between key variables before they are fully modelled with controls. These methods are not particularly robust, in a causal sense, and are therefore used as primers for the modelling rather than fully inferential components of analysis. As most of our data is metric, we mostly use Pearson’s correlations for this task with pairwise deletion of missing values.

4.3 Multivariate methods

The weight of our analysis rests on regression modeling. This is because models not only provide more detailed output than simpler methods, but also allow for more than one independent variable to be tested against the dependent simultaneously, with the effects controlling for each other. This ability to control for other effects is critical to one of our research questions which seeks to investigate the effect of staff numbers net of income.

We use two types of model, depending on the format of the dependent variable. For metric outcomes, we employ ordinary least squares which is the most conventional regression model, predicting the conditional mean. For binary categorical outcomes, we use logit which predicts the probability of being in the affirmative category of the outcome rather than variation within a metric scale. These two types of model have similar outputs and are interpreted similarly. Where they do differ in interpretation, we note this in the results.

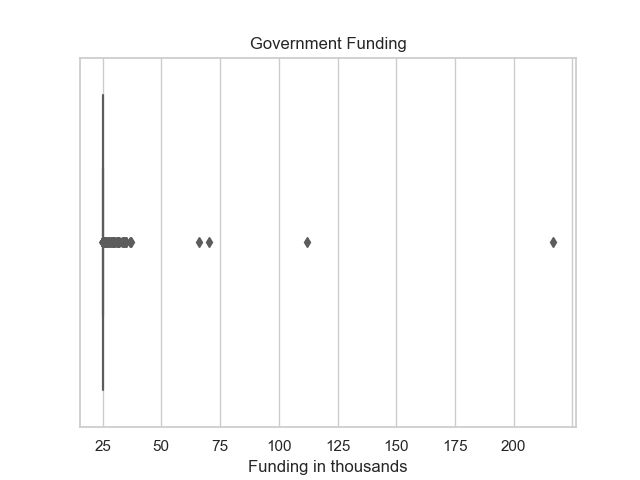
5 RESULTS & DISCUSSION

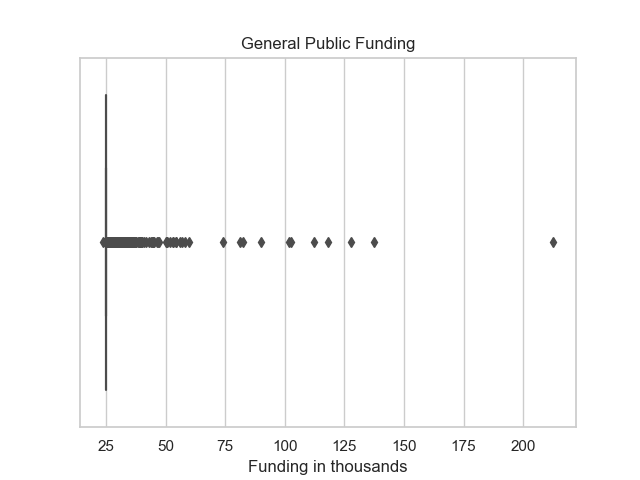
5.1 Question 1

How is source of funding related to charity use of Twitter?

The aim of this research question was to discover if government funding or general public funding had an influence on the likelihood of the charity having and using twitter. Therefore only those charities which were government or public funded were retained for analysis. It seems plausible that charities relying more heavily on public funding would be more inclined to have and use twitter as a way of raising funds. This question will assess the extent to which this is true.

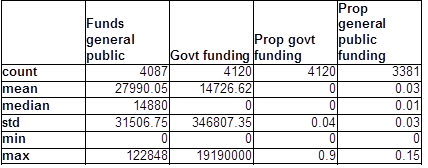
Having a twitter handle

The funding sources are highly positively skewed with extreme outliers as shown by the boxplots for government funding and general public funding.



Outliers were dropped using a interquartile method for the entire dataset.. 2153 charities were dropped from government funding and 2008 charities were dropped from general public funding.

After outlier dropping the table of the summary statistics for government funding and general public funding in this subset of data is displayed below.



The funding variables are still highly positively skewed, therefore using funding as a proportion of the charities income is more appropriate. The data was further made categorical by assessing whether the charities had received any funding from either source. 'CatGeneralFunding' and 'CatGovtFunding' variables were also created. 98% of these charities have general public funding and 5% of these charities have government funding. This data set is, therefore, highly biased towards charities funded by the general public.

Investigating the proportions of each funding category to measure having twitter shows, at a glance, that government funded charities are more like to have twitter.



Logistic regression with the proportion of funding variables

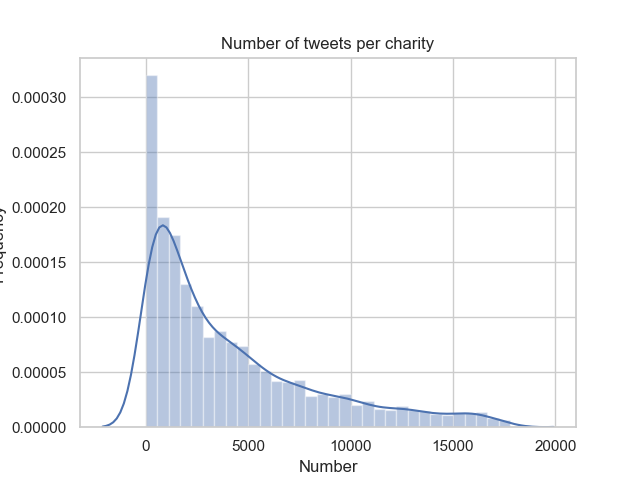
The model analysed proportions of funding using income as a size control with the outcome variable being the binary 'Has Twitter'. Whilst there is a positive effect of having general public funding it is not statistically significant and likewise with government funding whilst there is a negative effect it is not statistically significant.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent: Has Twitter** | **Coefficient** | **Std Error** | **P>|z|** |
| constant | -0.52 | 0.06 | <0.001 |
| Prop\_government\_funding | -0.35 | 1.07 | 0.74 |
| Prop\_general\_public\_funding | 0.95 | 1.05 | 0.36 |
| Income 2018 | 0 | 0 | <.0.001 |

Pseudo R Squared:0.032 AIC: 3726 BIC: 3750

Use of Twitter

Charity interaction with Twitter can also be measured in terms of use, rather than simply having an account. ‘Use of twitter’ is the number of tweets the charity has published in total. This is a metric variable which acts as a proxy for use.



For this question only the charities that were public or government funded and had a twitter handle were used. All of the variables are positively skewed indicating that most charities have a low proportion of public funding, even lower proportion of government funding and most charities don't have a high number of tweets; there are a small number of outliers in each case.

Inspecting the correlation of funding variables, shows no particular correlation between government funding and number of tweets with only a minimal 4% correlation between public funding and number of tweets.

Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent:** Number of Tweets in total | **Coefficient** | **Std Error** | **P>|z|** |
| Prop\_government\_funding | 3824 | 2964 | 0.19 |
| Prop\_general\_public\_funding | 8145 | 2953 | 0.01 |
| Income 2018 | 0 | 0 | < 0.001 |
| constant | 2453 | 193 | < 0.001 |

R-squared: 0.051 Prob (F-statistic): 0.000

According to this model general public funding has a significant positive effect on the total number of tweets a charity will make. A charity with a higher proportion of public funding will make, on average, more tweets than one with less income generated from the public. Based on these results there we cannot make any claim about government funding as it relates to activity on Twitter.

5.2 Question 2: Are charities which seek to help the public more popular on Twitter?

This question investigates the link between public-facing charities and popularity on Twitter. We could assume that charities which seek to help the general public will be more heavily incentivised to engage with Twitter to interact with the public. This would result in them being more popular on Twitter than non-public-facing charities but to what extent does the data substantiate this assumption?

Univariate descriptives

|  |  |  |
| --- | --- | --- |
|  | **Helps the general public** | **Does not help general public** |
| n | 5,528 | 5,272 |
| % | 51.2 | 48.8 |

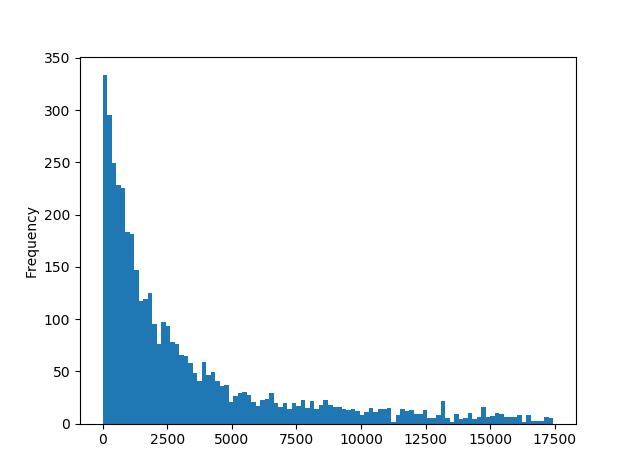
The table above summarises the binary ‘Helps the general public’ variable which records if a charity seeks to help the general public or a more focused group (such as other charities, children, animals, etc.). Charities such as educational trusts are also usually excluded from this group. Roughly half of the sample is in each category which should provide good variation for modelling.

|  |  |  |  |
| --- | --- | --- | --- |
| **Does not** **help general public** | **Income 2018** | **Proportion of general public funding** | **Twitter followers** |
| **Count** | 4,612 | 5747 | 1,976 |
| **Mean** | 1,407,509 | 0.014 | 2,735 |
| **Std** | 1,687,574 | 0.029 | 3,358 |
| **Median** | 758,100 | 0 | 1,508 |
| **Min** | 0 | -0.046 | 0 |
| **Max** | 7,400,000 | 0.151 | 17,324 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Helps general public** | **Income 2018** | **Proportion of general public funding** | **Twitter followers** |
| **n** | 4,846 | 5,544 | 2,218 |
| **Mean** | 1,297,389 | 0.018 | 3788 |
| **Std** | 1,622,773 | 0.034 | 4,118 |
| **Median** | 664,300 | 0 | 2,134 |
| **Min** | 0 | -0.024 | 0 |
| **Max** | 7,400,000 | 0.151 | 17,446 |

The two summary tables above are split by the binary ‘Helps the general public’ variable to show how these groups differ in terms of funding and popularity on Twitter. As shown, charities which help the general public tend to have slightly lower incomes, but gain more of their income from public funding. Most importantly, they appear to have more followers on Twitter, but this will be fully explored in the modelling.

Figure 1.



Twitter followers

This histogram shows the distribution of the dependent variable, Twitter followers, for all charities (after it had been treated for outliers). The distribution appears half-normal and is not overly afflicted with outliers and should be suitable for modelling. Followers is a good proxy for popularity because it records how many other accounts have chosen to actively subscribe to a given charity’s content on Twitter.

Multivariate modelling

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent**: Twitter followers | **Coef.** | **Std error** | **P>|t|** |
| Helps general public (binary) | 1,193 | 121 | <0.001 |
| Income 2018 | <0.001 | <0.001 | <0.001 |
| Constant | 1,632 | 110 | <0.001 |

R-squared 0.064 Prob = <0.001

This is an ordinary least squares model which predicts number of Twitter followers (popularity) based on whether the charity helps the general public. Income is included as a control for size which usually has a large effect on Twitter popularity (larger charities tend to be more popular on Twitter). Both the primary independent and the control are significant in this model. The primary measure is binary so the result 1193 means that charities which help the general public, on average, have 1193 more Twitter followers than those who do not – controlling for size. This suggests public facing charities are more popular on Twitter and positively affirms research question 2. However, the R-squared for this model is small which suggests there are many other factors (or simply random variation) which affect charity popularity on Twitter.

5.3 Question 3: Does number of staff, rather than size, determine a charity’s active use of Twitter?

Univariate descriptives

In order to determine whether the number of staff employed by a charity, rather than the size of the charity, plays a role in a charity’s successful use of Twitter, three variables were selected from the final data set: ‘Staff’, ‘Twitter following’, and ‘Income2018’. ‘Twitter following’ is the number of accounts the charity follows and is used as a measure of active use of Twitter, as some charities may choose not to Tweet very often but may still follow other accounts which shows a level of use.

**Table 1: Summary statistics for Staff, Twitter following and Income2018**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Staff** | **Twitter following** | **Income2018** |
| **Count** | 10,695 | 4,454 | 9,458 |
| **Mean** | 14.73 | 1,000.06 | 1,351,087 |
| **Standard deviation** | 23.89 | 885.69 | 1,655,516 |
| **Minimum** | 0 | 0 | 0 |
| **25%** | 0 | 289 | 138,250 |
| **50%** | 0 | 758.5 | 706,700 |
| **75%** | 22 | 1,505.5 | 1,900,000 |

Bivariate Correlation

The correlation between the number of Staff and the number Twitter following was calculated and the result was positive but weak at 0.07. This suggests there may be some form of association but this needs to be assessed in a more robust modelling framework.

Multivariate modelling

A linear regression was specified using ‘Staff’ and ‘Income2018’ as the independent variables and ‘Twitter following’ as the dependent variable. The results of this model are shown in Table 2.

**Table 2: Results of the Linear Regression Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent:**  Twitter following | **Coef.** | **Std error** | **P>|t|** |
| **Staff** | -0.992 | 0.756 | 0.189 |
| **Income 2018** | <0.001 | <0.001 | <0.001 |
| **Constant** | 867 | 21 | <0.001 |

R-squared = 0.017 Prob = <0.001

The coefficient for ‘staff’ is negative but this is insignificant so we cannot claim to have found any evidence of an association between staff numbers and Twitter popularity. Income has a very small positive and significant result which suggests that larger (income is a common proxy for size) charities are more active on Twitter. However, this variable was a control and out main effect was insignificant.

6 CONCLUSIONS

6.1 Summary

The

6.2 Limitations

6.3 Future work

The

REFERENCES