# Creation of a JSON object from the original CSV file

The data set chosen as a starting point for this project was from the UK Data Archive and 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.

The ‘CharityCharacteristics.csv’ file was imported into Python and a pandas data frame was created. Then df.reset\_index followed by df.set\_index was used to set a two level index using Charity Number (‘ccnum’) and Financial Year (‘financial\_year’). Next df.drop was used to drop a list of all the financial years apart from 2011/12. The two level index was no longer needed at this point, so the index was reset to Charity Number and the pandas data frame was saved as a JSON object: ‘charity\_oneyear.json’.

# Creation of new variables and a subset of the 2011/12 Charities data

In the ‘CharityCharacteristics.csv’ and the ‘charity\_oneyear.json’ files, the data for the source of funding is split into a large number of variables and the names used are not particularly informative. A new variable was created in the data frame and the addition operator was used to combine the source of funding data from different columns:

* **Government funding** ('government\_funding')   
  – combination of ig100, ig110, ig121, ig125, ig161, ig162, ig163 and ig180

Two new variables were created to assess the proportion of each charity’s funding that came from Government Funding and General Public Funding:

* **Government Funding Proportion** (‘Prop\_government\_funding’)   
  = Government funding / total income (‘itotal’)
* **Public Funding Proportion** ('Prop\_general\_public\_funding')  
  = ig600 / total income (‘itotal’)

The data types of the variables used in these calculations were checked to ensure they were compatible using df.dtypes and they were found to be integers (‘int64’). Lastly, the variable for each charity’s total income (‘itotal’) was renamed as **'Income2011-2012'** to distinguish it from up-to-date information collected via a web scrape and the variable for funds generated from the general public (‘ig600’) was renamed as **'Funds\_general\_public'**.

After creation and renaming of the variables of interest, the filter () function was applied to the data frame to create a new smaller data frame with data for 12 150 charities and only 6 variables:

* **'ccnum'** – the Charity Number which was kept as the index
* **'Government\_funding'** – total funding from the Government
* **'Prop\_government\_funding'** – proportion of funding coming from the Government
* **'Prop\_general\_public\_funding'** – proportion of funding coming from the general public
* **'Income2011-2012'** – Income from 2011 - 2012
* **'Funds\_general\_public'** – funds generated from the general public

# Research Question 4: Does number of staff, rather than size, determine a charity’s successful use of Twitter?

## Summary statistics of the variables

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: ‘Final\_analysis\_file.json’.

* **‘Staff’** - Number of staff employed by the charity, measured as full-time equivalent (FTE)
* **‘Twitter following’** - Number of people / organisations the charity follows on Twitter is used as a measure of active use of Twitter, as some charities may choose not to Tweet very often
* **‘Income2018’** – Total income from the web scrape carried out in 2018

The df.describe() command was used to produce statistics for these three variables and the results are shown in Table 1. The number of NaN present for all variables in the data frame was counted using df.isnull().sum(): Staff: 1455; Twitter following: 7696; and Income 2018: 2692 but Python automatically excludes these when calculating summary statistics.

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Staff** | **Twitter following** | **Income2018** |
| **Data type** | float64 | float64 | float64 |
| **Count** | 10695.000000 | 4454.000000 | 9.458000e+03 |
| **Mean** | 14.729874 | 1000.057252 | 1.351087e+06 |
| **Standard deviation** | 23.886958 | 885.686228 | 1.655516e+06 |
| **Minimum** | 0.000000 | 0.000000 | 0.000000e+00 |
| **25%** | 0.000000 | 289.000000 | 1.382500e+05 |
| **50%** | 0.000000 | 758.500000 | 7.067000e+05 |
| **75%** | 22.000000 | 1505.500000 | 1.900000e+06 |
| **Maximum** | 102.000000 | 3874.000000 | 7.400000e+06 |

## Correlation between Staff and the number ‘Following’ on Twitter

The correlation between the number of Staff and the number ‘Following’ on Twitter was calculated using df['Staff'].corr(df['Twitter following'] and a weak positive correlation of 0.07 was measured.

## Linear Regression of Staff predicting the number ‘Following’ on Twitter controlling for Income

A linear regression was carried out (y = m1x1 + m2x2) using ‘Staff’ and ‘Income2018’ as the independent variables: x1 and x2, and setting ‘Twitter following’ as the dependent variable: y. The statmodels.api module was used and the results are shown in Table 2.

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

1. OLS Regression Results
2. ==============================================================================
3. Dep. Variable:      Twitter following   R-squared:                       0.017
4. Model:                            OLS   Adj. R-squared:                  0.016
5. Method:                 Least Squares   F-statistic:                     28.26
6. Date:                Tue, 13 Nov 2018   Prob (F-statistic):           6.74e-13
7. Time:                        22:05:34   Log-Likelihood:                -27575.
8. No. Observations:                3369   AIC:                         5.516e+04
9. Df Residuals:                    3366   BIC:                         5.517e+04
10. Df Model:                           2
11. Covariance Type:            nonrobust
12. ==============================================================================
13. coef    std err          t      P>|t|      [0.025      0.975]
14. ------------------------------------------------------------------------------
15. const        867.4814     21.347     40.637      0.000     825.627     909.336
16. Staff         -0.9916      0.756     -1.313      0.189      -2.473       0.490
17. Income2018  8.346e-05   1.29e-05      6.469      0.000    5.82e-05       0.000
18. ==============================================================================
19. Omnibus:                      487.549   Durbin-Watson:                   1.954
20. Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.926
21. Skew:                           1.092   Prob(JB):                    7.71e-157
22. Kurtosis:                       3.595   Cond. No.                     3.05e+06

The Adjusted R-squared value takes into account the number of independent variables used in the model and in this case it is 0.016. The constant coefficient is the intercept with the y axis and is 867.4814. The standard error indicates the accuracy of the constant, ‘Staff’ and ‘Income2018 coefficients with a low value indicating a higher accuracy, and the confidence interval shows the range that the coefficients will probably fall within.

The p-value of 0.189 for ‘Staff’ indicates that changes in this variable are not associated with changes in ‘Twitter following’ – the y variable. However, the p-value of 0 for ‘Income2018’ is statistically significant, so the null hypothesis that size of a charity does not affect the number of people / organisations that the charity follows on Twitter can be rejected. Note: Size of the charity is measured by the Income in this case.