GROUP 1

External students

Members

Stephen Maher

Jinxi Luo

.

Contents

1. Recommendations: Summary

The core findings from the analytical plan and the subsequent interpretation of these findings were:

- This Altmetric dataset is a *heavily biased representation* of the full range of altmetric outcomes and likely only represents the subset of papers, articles, and other scored publications with a high altmetric attention score
- This Altmetric dataset is *strongly aligned* with social media metrics
- This Altmetric dataset is *weakly aligned* with journal impact factors and broader scientometric measures
- The altmetric attention scores in this Altmetric dataset show a *predisposition towards open-access* publications and publications specific fields
- The altmetric attention score is potentially predictable *within the range of the scores* contained in this Altmetric dataset
- It is unknown if the altmetric attentions score is predictable outside of the range of score contained in this Altmetric dataset

Using these finding and the insights captured from the analytical plan, the **recommendations** as to the use of the altmetric attention score "AAS" are as follows:

- To *only use the AAS as a measure for the social attention* that a paper attracts
- To *focus on selected subjects* if promoting publications via social media
- To *focus on open-access* avenues for the distribution of papers
- To *benchmark actual AAS outcomes* against predicted outcomes

2. Introduction

This analysis extends the preliminary work conducted on the altmetric dataset. This work encompassed developing the altmetric dataset, including data review and cleansing, the development of a data dictionary, exploratory analysis and the development of an analytical plan.

The implementation of the analytical plan and the subsequent results are described as follows:

1. The findings are presented in Section 3
2. The interpretation of the findings are presented in Section 4
3. Business recommendations based on the findings and ensuing interpretation are presented in Section 5
4. Directions for future research are discussed in Section 6

The implemented analytical has deviated from that discussed in the initial report. This was expected and is due to a continuously developing understanding of the altmetric dataset as the analysis progressed.

3. Findings

3.1 Findings: Altmetric Data Set 2013 – 2019

Four analytical techniques were selected for the analysis of the Altmetric dataset. These were:

- Clustering: K-prototype
- Association
- Decision Tree
- Multiple Linear Regression

The following variables were available for analysis within the dataset:

Discrete (quantitative)
(Altmetric Attention Score) (Blog Mentions) (F1000 Mentions) (FB Mentions) (Google+ Mentions) (News Mentions) (Patent Mentions) (Policy Mentions) (Reddit Mentions) (Twitter Mentions) (Video Mentions) (Wikipedia Mentions)
Binary (qualitative)
(ArXiv ID) (Dimensions ID) (DOI)(Journal ISSNs) (PubMed ID)
Nominal (qualitative)
(Category) (Journal) (OA)

3.1.1 Cluster Analysis

K- Prototype is a clustering method for both numerical and categorical variables which uses a combination of k-means and k-mode dissimilarity measures. Methodological details are detailed in Appendix 6.1 along with additional figures. The goal of using this analysis to provide an overview of the effect that the different attributes within the dataset have on the AAS.

Three optimal clusters were found using the elbow method. Publication density's AAS per cluster can be seen in Figure 1. Most publications are rated with an AAS in the range 2000 to 5000 as found in cluster 2. However, clusters 0 and 1 have outlier AAS within each cluster. A comparison between the distribution of features across the three clusters revealed that no major differentiation exists between the clusters besides open access status, publication category, twitter mentions and altmetric score. From Figure 1, open access publications show a propensity to be aligned with a higher AAS within all clusters.

Figure 1: Cluster density by open access

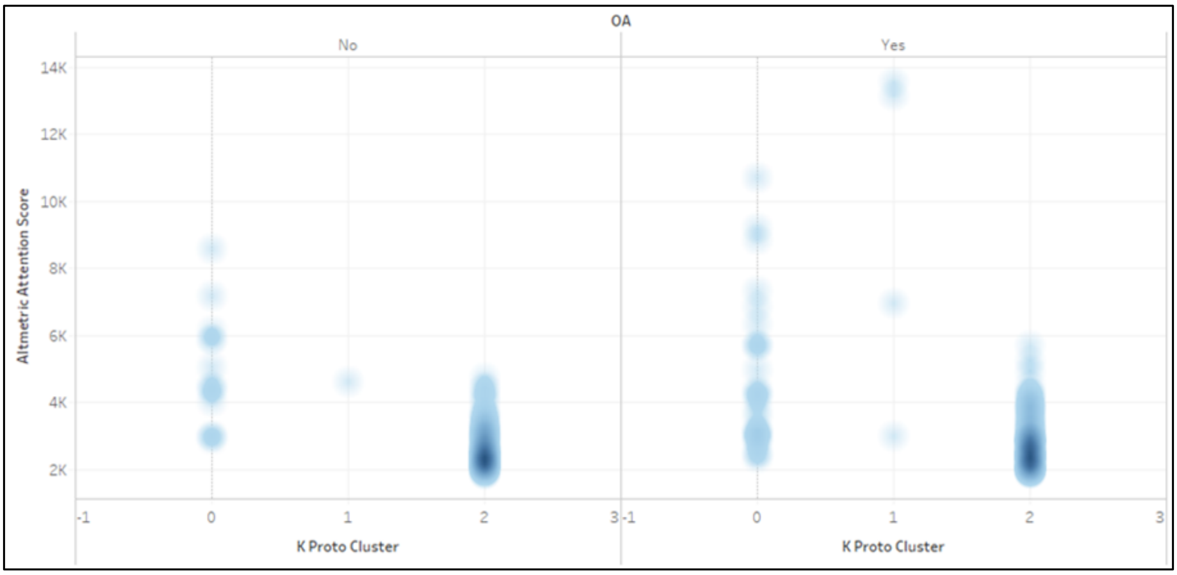
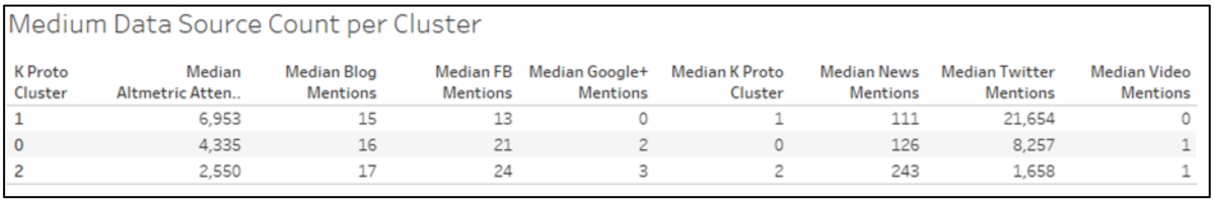


Figure 2 below shows the medium count of altmetric data sources per cluster. Medium is used as Figure 1 shows that the use of mean may skew results in clusters 0 and 1. Note that the F1000, Policy, and Wikipedia attributes had zero in every entry in the dataset and were removed from the Figure 2. Clusters 0 and 1 have the highest twitter counts and AASs. In contrast, cluster 2 had the lowest twitter mentions but did have the highest news mentions in terms of data source variation. This suggests that a high twitter count may be the cause of the high (or unusual) AAS in clusters 0 and 1 (from the exploratory analysis, a relationship between AAS and twitter mentions was expected).

Figure 2: Medium data source count per cluster

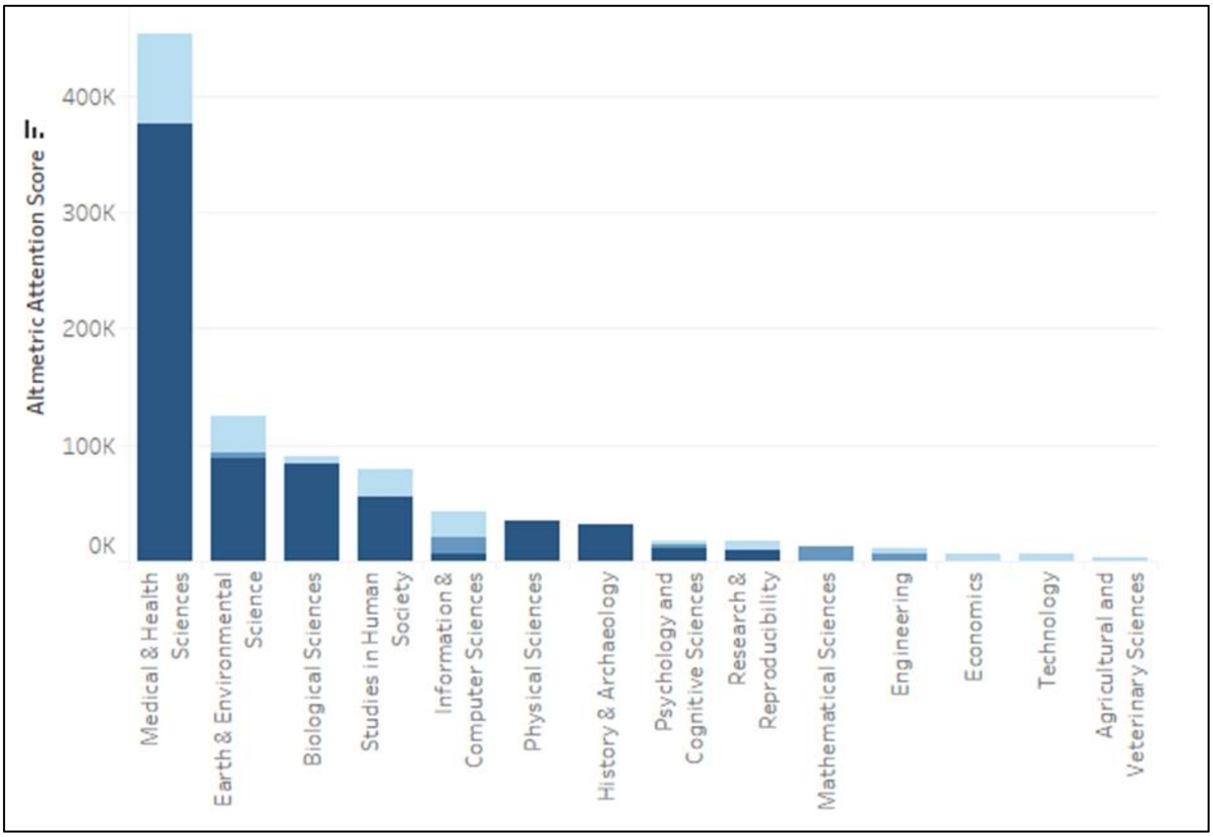


Medium Data Source Count per Cluster

K Proto Cluster	Median Altmetric Atten..	Median Blog Mentions	Median FB Mentions	Median Google+ Mentions	Median K Proto Cluster	Median News Mentions	Median Twitter Mentions	Median Video Mentions
1	6,953	15	13	0	1	111	21,654	0
0	4,335	16	21	2	0	126	8,257	1
2	2,550	17	24	3	2	243	1,658	1

However, it was also found that publication category was a characteristic of cluster assignment. Figure 3 details the contribution to sum altmetric score per category grouped by cluster contribution (Cluster 0: light blue, cluster 1: medium light blue and cluster 2: dark blue. The majority of altmetric attention scores for cluster 1 are in mathematical sciences, information & computer sciences and engineering. The top three categories for cluster 2 are medical & health sciences, earth & environmental science and biological sciences. Cluster 0 is like cluster 2, but with studies in human society replacing biological sciences.

Figure 3: Sum AAS per category by cluster (cluster 0 light blue, cluster 1 medium light blue and cluster 2 dark blue)



3.1.2 Association

Association rule mining with Apriori is used to find common occurrences of items in categorical data. AAS was converted to a categorial feature with five factor levels for association and the following decision tree analysis. The levels were determined based on the quantiles of all AAS from 2013 to 2019 (Table 1).

Table 1: Quintile factor categorisation of AAS

LEVEL	AAS
LOW	498 <= AAS < 1174
LOW2MEDIUM	1174 <= AAS < 2124
MEDIUM	2124 <= AAS < 2873
MEDIUM2HIGH	2873 <= AAS < 5060
HIGH	5060 <= AAS

Figure 4 is a visualisation of the 50 top rules (detailed in Figure 18, Appendix 6.2). When the rules are sorted by lift, the top 30 rules have high lift but low support. Interestingly, support only increases after the 30th rule. This suggests that the first 30 rules are useful and that the LHS (precedent) strongly implies the RHS (consequent) as lift is greater than 1. The balance of the 50 top rules, while not evidencing as significant lift, may be useful as they occur reasonably frequently. The five top rules are listed in Table 2

While the goal of this analysis was to find associations of attributes alongside the AAS, the top rules require some interpretation. Only AAS=low was featured in these rules, which suggests that the presence of these attributes results in a higher probability of an occurrence of a low AAS. However, there is no evidence to suggest that the absence of these attributes would suggest a high AAS.

Figure 4: Visualisation of the 50 top rules by lift and support

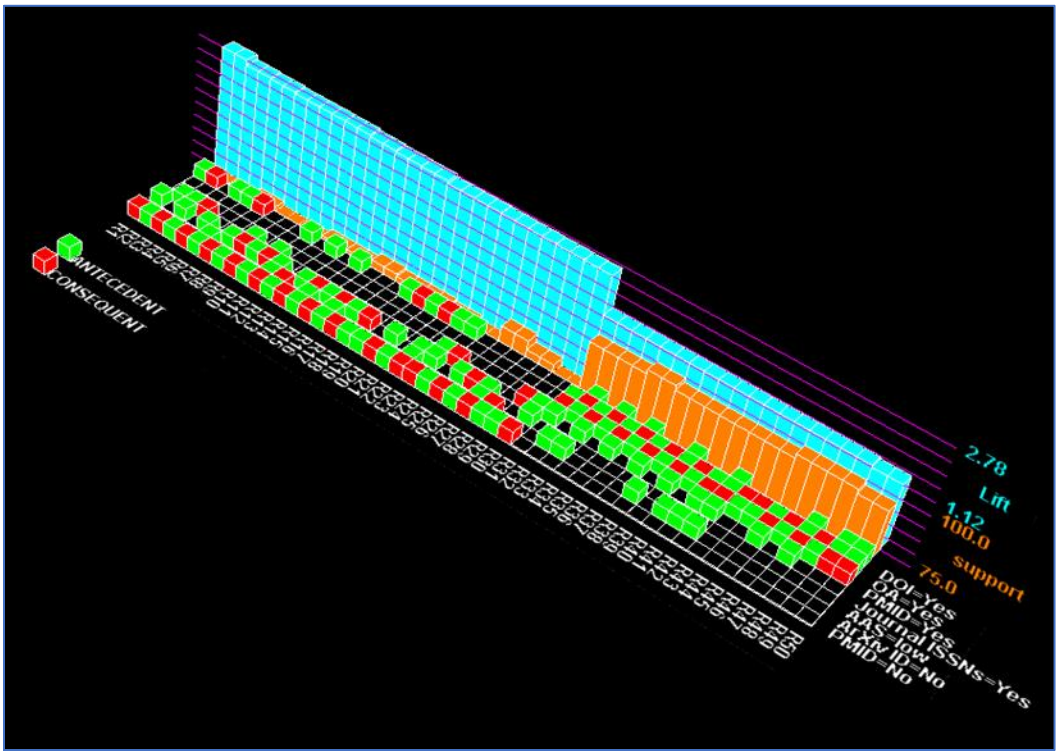


Table 2: Top 5 association analysis rules (ID attributes included in the analysis).

RULE	LHS	RHS	CONFIDENCE	LEVERAGE	LIFT
1	AAS=low DOI=Yes	JournalISSNs=Yes PMID=No	0.93	0.16	2.78
2	JournalISSNs=Yes PMID=No	ASS=low DOI=Yes	0.75	0.16	2.78
3	AAS=low JournalISSNs=Yes	DOI=Yes PMID=No	0.95	0.16	2.64
4	DOI=Yes PMID=No	AAS=low JournalISSNs=Yes	0.69	0.16	2.64
5	AAS=low DOI=Yes	ArXivID=No PMID=No	0.95	0.16	2.62

Beyond suggesting not what to do, this analysis has provided no further useful information. However, dropping the ID related attributes from the association analysis yielded some further useful results. The top three discovered rules (Table 3 and detailed in Figure 17, Appendix 6.2) show that:

- Medical & Health Sciences are associated with a medium2high and medium AAS
- Non open access publications are associated with a low AAS
- Open access publications are associated with a medium2high AAS

Table 3: Subset of top 10 association rules, IDs removed:

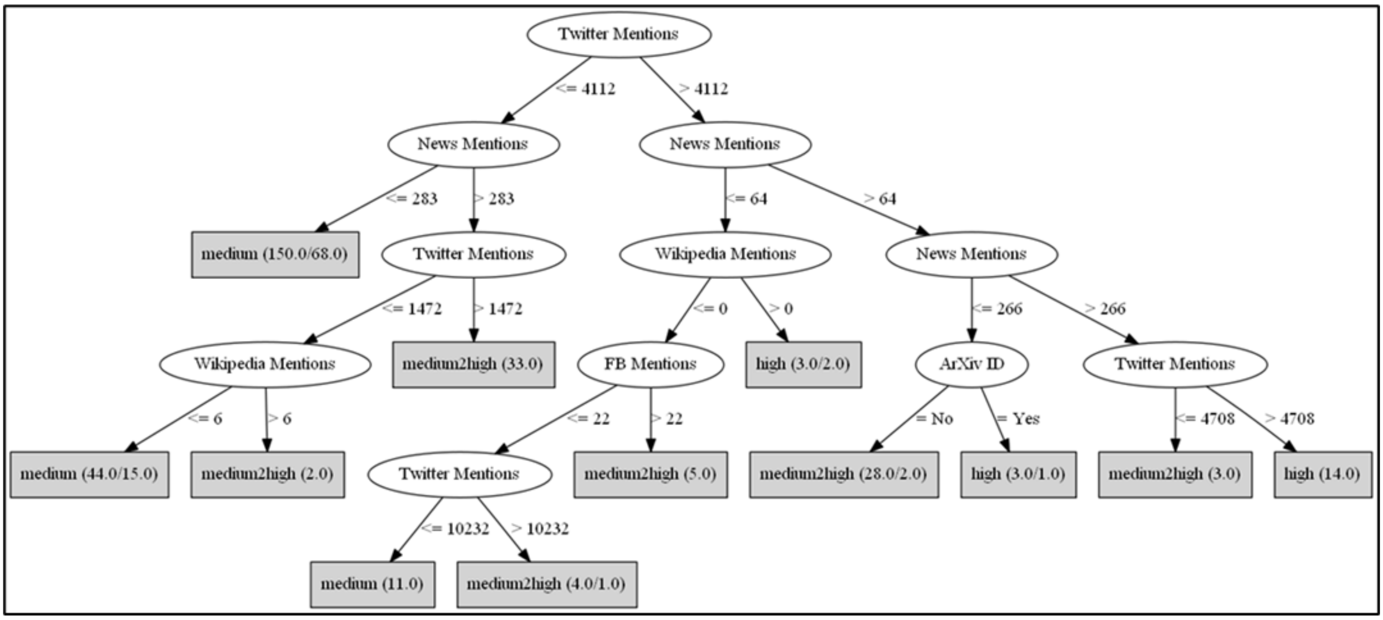
RULE	LHS	RHS	CONFIDENCE	LEVERAGE	LIFT
1	Category=Medical&Health	AAS=medium2high	0.44	0.06	1.3
5	OA=No	ASS=low	0.37	0.04	1.28
8	OA=Yes	AAS=medium2high	0.3	0.02	1.05

3.1.3 Decision Tree

A decision tree is a classifier that uses both categorical and numerical variables. From Section 3.1.2 (Association), there are a range of rules associated with the AAS and a decision tree may support further refinement. Figure 5 details the resultant Decision Tree which has an accuracy of 60%. This shows that social media sources (such as Twitter Mentions, News Mentions and Wikipedia Mentions) have the highest predictive accuracy for AAS (as categorised using the association analysis quantiles). The decision tree was based on data available for 2017, 2018 and 2019, as these subsets of the altmetric dataset provided the most complete coverage of these attributes.

In general, a high data source count is found to be associated with a high AAS. Twitter mentions, news, wikipedia and FB data sources are the best predictors for a higher AAS. Of the non-social media metrics, an ArXiv value (of all the identifier style attributes) also appears to be a good indicator for a higher AAS. The ROC graphs (Appendix 6.3, Figures 21 - 24) show that the information value of predictions for a high AAS is reasonable and that the decision tree may warrant further attention.

Figure 5: Decision tree based on the altmetric datasets for 2017, 2018 and 2019



We can expect factors leading to a high and a medium2high AAS to be a good true positive predictor of AAS. The tree shows these factors to be the altmetric data sources of Twitter, News, Wikipedia, along with the ArXiv ID. Below lists the decision rules for achieving high and medium2high AAS:

- High
  - 4708 < Twitter & 266 < News
  - 4112 < Twitter & 64 < News <= 226 & ArXiv = Yes
  - 4112 < Twitter & News <= 64 & 0 < Wikipedia

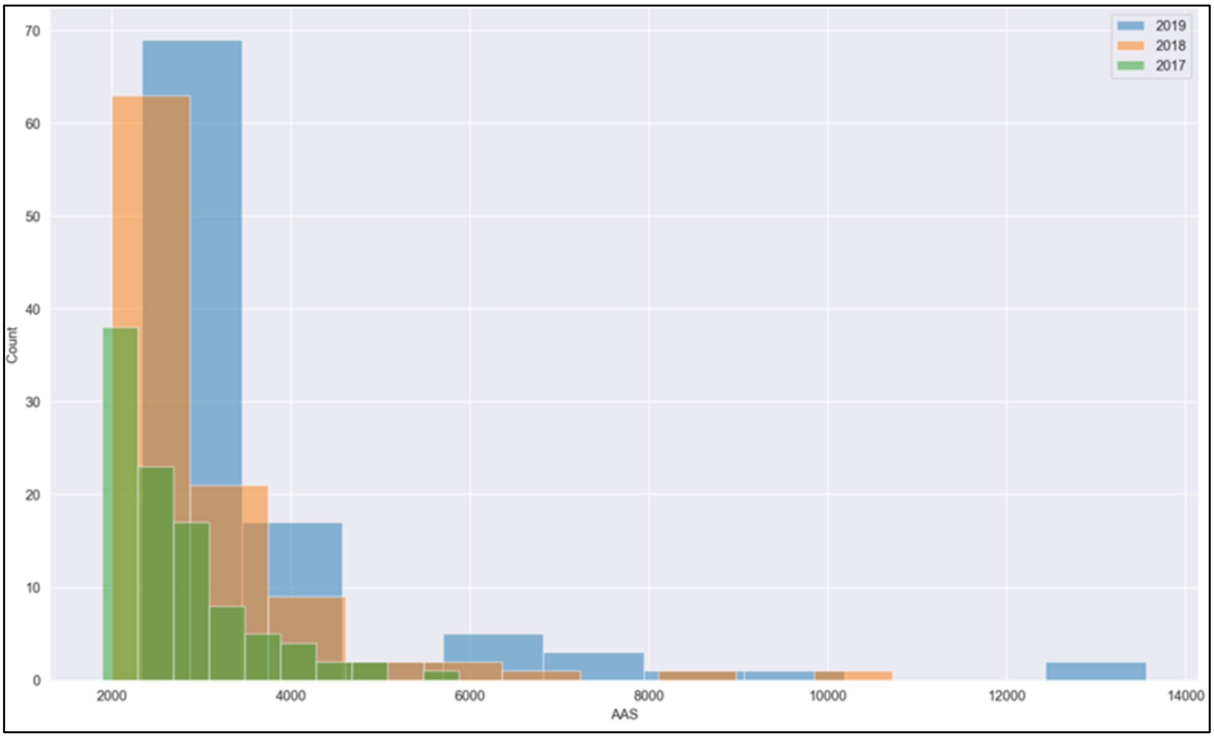
- Medium2High
  - 4708 < Twitter <= 4708 & 266 < News
  - 4112 < Twitter & 64 < News <= 226 & ArXiv = No
  - 4112 < Twitter & News <= 64 & Wikipedia <= 0 & 22 < FB
  - 10232 < Twitter & News <= 64 & Wikipedia <= 0 & FB <= 22
  - 1472 < Twitter <= 4112 & 283 < News
  - Twitter <= 1472 & 283 < News & 6 < Wikipedia

3.1.4 Multiple Linear Regression

Multiple linear regression is used to model the relationship between more than one independent variable and a dependent variable (independent variables are expected to have a linear relationship with the dependent variable).

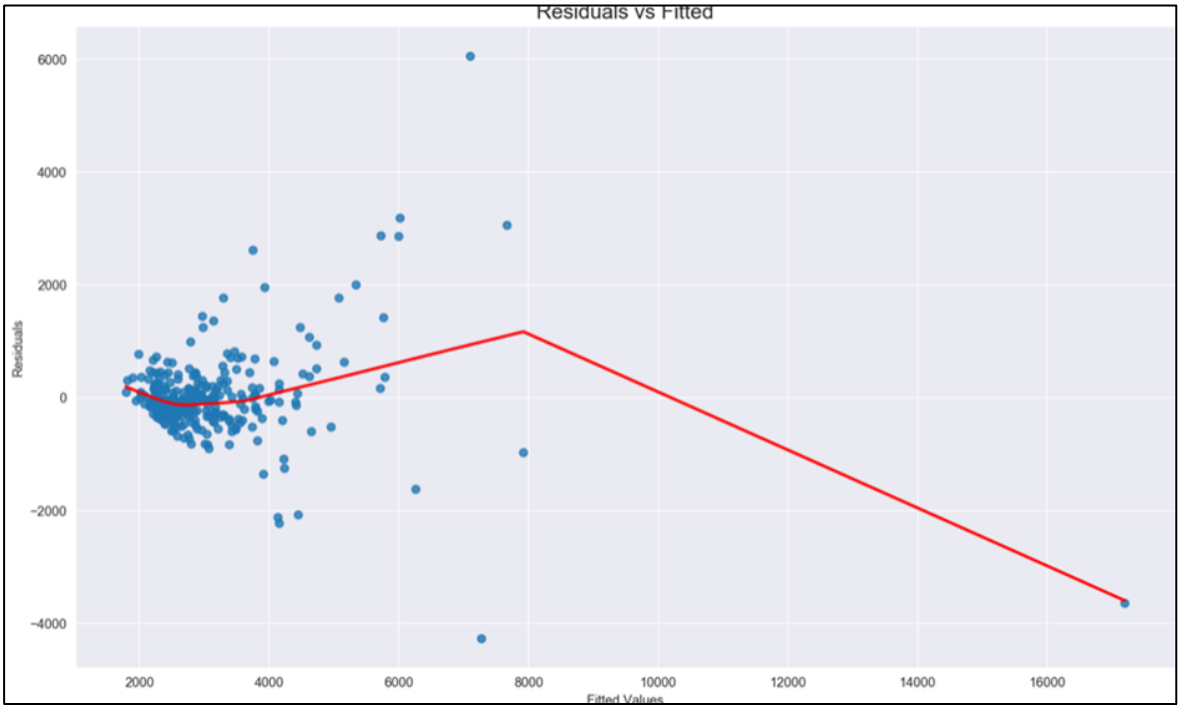
From the decision tree analysis, it was possible to identify the most significant attributes with which to predict the AAS. However, we are limited by the categorical nature of the AAS prediction. Multiple liner regression using these attributes produced a model to predicting AAS numerical values. An additional benefit of this approach is that it may be used to also predict future altmetric attention scores, particularly as that the AAS for the top 100 publications has been increasing through time (Figure 6).

Figure 6: AAS distribution for 2017, 2018 and 2019



The results from the multiple linear regression indicated that the [second] model found does not satisfy the homoscedasticity assumption (refer Appendix 6, Section 6.4 for further detail). Figures 7 shows the residuals plot analysis (refer Figures 25-28, Appendix 6, Section 6.4 for additional results). This may be due to insufficient data and variance associated with the change in AAS factor weightings in 2018 and 2019. In the first run of the regression modelling, Wikipedia mentions were found to not be significant and the model was subsequently reduced to three features [second model].

Figure 7: Multiple linear regression analysis; residual vs. fitted (second model)



The regression equation found supports the results found in the decision tree and we can see that this justifies the use of three features to predict AAS. The equation is as follows:

$$AAS = 4.20(FB\ Mentions) + 5.56(News\ Mentions) + 0.26(Twitter\ Mentions) + 779.43$$

These results contrast with the actual altmetric weightings for news mentions (8), twitter mentions (1) and facebook (0.25). However, there are also additional data source weightings applied for specific news sources, along with twitter and facebook credibility measures for example, which are not reflected directly in our data sources. Finally, this equation can only apply when there are meaningful number of FB Mentions, New Mentions and Twitter Mentions already existing as the AAS equation has a minimum value of 779.43. If this was not the case, then any paper would achieve a substantial AAS with none or limited social mentions.

3.2 Findings: Bornmann & Haunschild

Two analytical techniques were selected for the analysis of the Bornmann & Haunschild[1] dataset. These were:

- Clustering: K-means
- Factor Analysis and Principal Components Analysis

The following variables were available for analysis within the dataset:

Discrete (quantitative)
(pubyear) (wos_cits_3) (sco_cits_3) (tweets) (me_readers) (total_f1000_score)
Continuous (quantitative)
(altmetric_score) (citescore) (item_ijif)

The attribute “pubyear” was dropped early in the analytical process as the results showed no differentiation on this attribute.

3.2.1 Cluster Analysis

The cluster analysis showed some differentiation within the dataset. Two cluster (using K-means) were identified with one cluster containing 87% of the rows in the dataset and the other containing 13%. The key differentiation between the two clusters was the mean value of the attributes within each cluster (Table 4). As can be seen in Table 4, there is a high mean attribute value cluster (cluster 0) and low mean attribute cluster (cluster 1).

The relative distribution is evident in Figures 8 and 9 but is not as clear in Figure 10, which shows a relatively tight correlation between “tweets” and “altmetric_score” with both clusters overlain on each other. Cluster 0 (blue) appears to show some association between higher levels of citation and other measures of quality and social media attention as measured by tweets and the altmetric score. It is worth noting however, that reverse is less evident. That is, a relatively high tweet or altmetric score may not necessarily indicative of higher levels of citation or other quality measures.

Table 4: Bornmann & Haunschild Dataset: K-means cluster analysis, mean attribute value by cluster

	Overall	Cluster 0	Cluster 1	
Row Count	33683	5593	28090	Difference [0-1]
wos_cits_3	29.8576	81.7348	19.5283	62.2065
sco_cits_3	31.0315	83.1534	20.6535	62.4999
tweets	10.1401	33.8874	5.4118	28.4756
me_readers	76.4402	212.6832	49.3129	163.3703
altmetric_score	17.122	55.1847	9.5433	45.6414
citescore	7.2166	14.6243	5.7417	8.8826
item_ijif	11.0997	32.0108	6.9361	25.0747
total_f1000_score	2.0417	3.8731	1.677	2.1961

[1] Bornmann, L & Haunschild, R 2018, ‘Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data.’, PLoS ONE, vol. 13, no. 5, p. E0197133.

Figure 8: Bornmann & Haunschild Dataset: clusters identified on item_ijif (y-axis) and altmetric_score (x-axis) [cluster 0 = blue, cluster 1 = red]

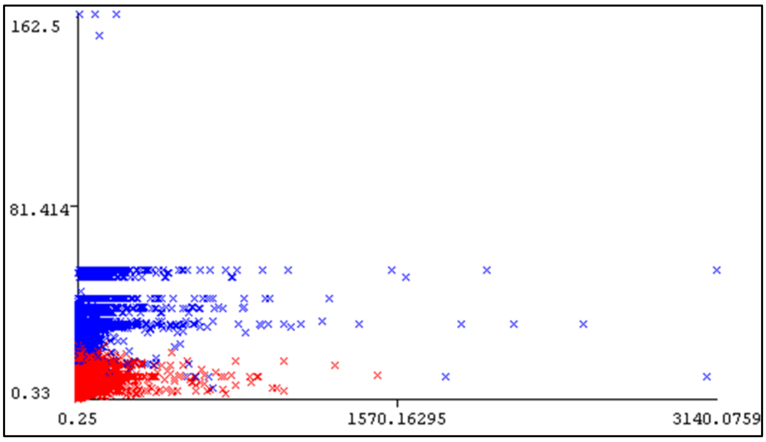


Figure 9: Bornmann & Haunschild Dataset: clusters identified on item_ijif (y-axis) and citescore (x-axis) [cluster 0 = blue, cluster 1 = red]

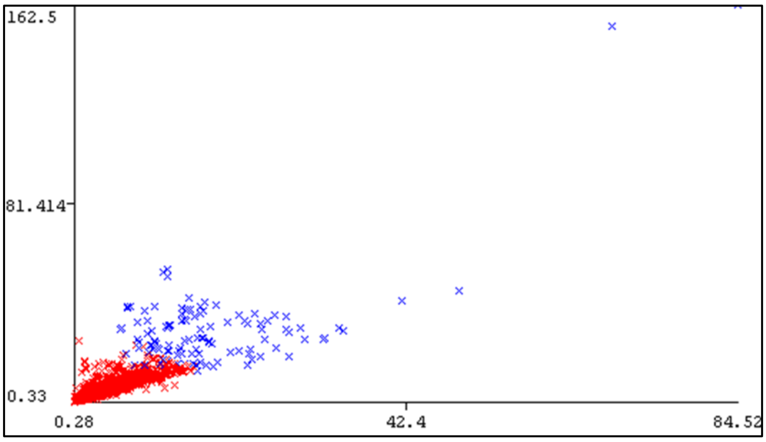
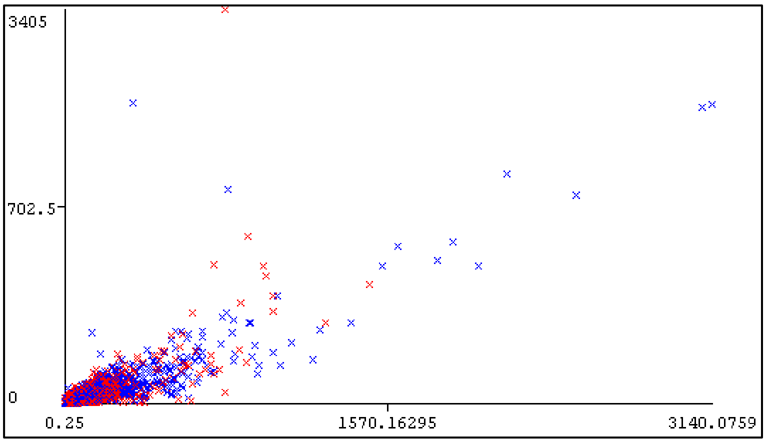


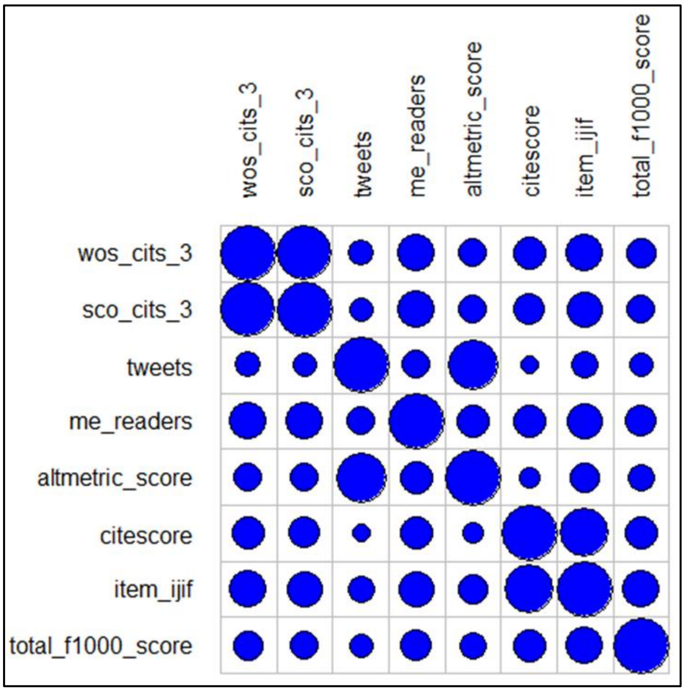
Figure 10: Bornmann & Haunschild Dataset: clusters identified on tweets (y-axis) and altmetric_score (x-axis) [cluster 0 = blue, cluster 1 = red]



3.2.2 Factor Analysis and Principal Components Analysis

The second analysis on the Bornmann and Haunschild data comprised principal components analysis "PCA" and factor analysis "FA". The initial step was to review the correlogram of the data (and excluding pubyear). The correlogram is shown in Figure 11 and feature three primary sets of correlations: 1) wos_cits_3 and sco_cits_3, 2) tweets and altmetric_score, and 3) citescore and item_ijif. This suggests that we may see at least three principal components in the PCA.

Figure 11: Correlogram of Bornman & Haunschild data



Extending the analysis to PCA (after applying standardisation), the relative importance of components is shown in Table 5. Using Kaiser's Criterion, the first three factors are selected as suitable and this was confirmed by a scree plot of the data (refer Appendix 9). Table 6 shows the resultant loadings due to these three components. From Table 6, it is evident that PC2 loads heavily on tweets and altmetric_score. PC3 loads on all factors excluding tweets, altmetric_score and me_readers. Notably for PC3, wos_cits_3 and sco_cits_3 have a negative sign while the remaining factors have a positive loading. While PC2 and PC3 show relatively clear loadings, PC1 loads on all factors.

Table 5: Bornman & Haunschild Data: Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.9013	1.2278	1.0792	0.84435	0.78863	0.44073	0.403	0.14606
Proportion of Variance	0.4519	0.1885	0.1456	0.08912	0.07774	0.02428	0.0203	0.00267
Cumulative Proportion	0.4519	0.6403	0.7859	0.87501	0.95275	0.97703	0.9973	1

Table 6: PCA: component loadings

	PC1	PC2	PC3
wos_cits_3	-0.41443	-0.19297	-0.4995
sco_cits_3	-0.40717	-0.19188	-0.52022
tweets	-0.27094	0.649947	0.035927
me_readers	-0.35547	-0.01647	-0.03661
altmetric_score	-0.31123	0.608903	0.028927
citescore	-0.34356	-0.29702	0.46509
item_ijif	-0.40432	-0.2004	0.393595
total_f1000_score	-0.29037	-0.064	0.324317

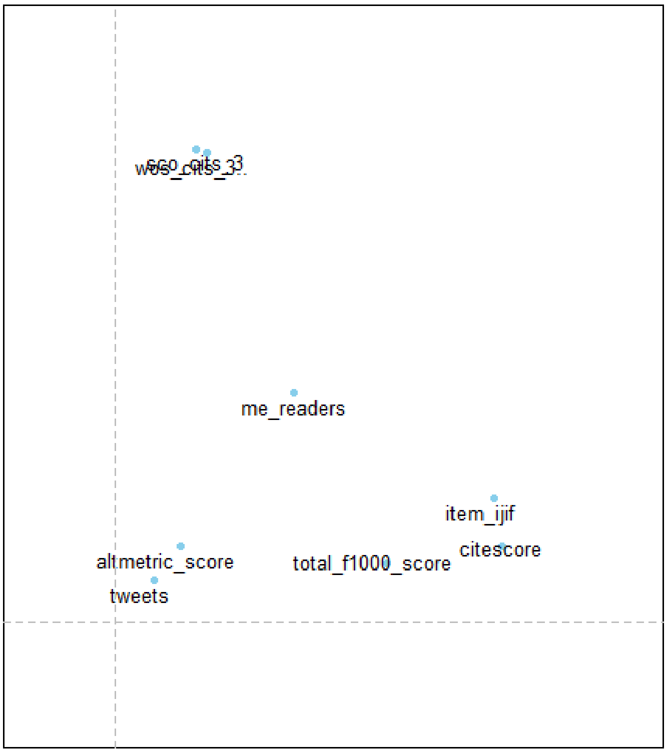
In addition to PCA, exploratory factor analysis was also conducted over the dataset (sample size is nominally excellent for this analysis). Preliminary statistics suggested that the use of EFA would be marginal effective at best. There was evidence of multicollinearity and a KMO test indicated that the common variance was unacceptable for factor analysis.

Table 7: Bornmann and Haunschild dataset: EFA using varimax rotation

	item	RC1	RC3	RC2	h2	u2	com
citescore	6	0.89			0.81	0.188	1.1
item_ijif	7	0.87			0.83	0.168	1.2
total_f1000_score	8	0.62			0.43	0.567	1.2
sco_cits_3	2		0.96		0.97	0.03	1.1
wos_cits_3	1		0.96		0.97	0.032	1.1
me_readers	4	0.41	0.47		0.46	0.541	2.6
tweets	3			0.94	0.9	0.096	1
altmetric_score	5			0.93	0.91	0.09	1.1

Notwithstanding concerns with respect to the application of EFA to the Bornmann and Haunschild dataset, EFA was performed under varimax rotation. Three factors were selected for the analysis and the cumulative variance explained was 79% and communalities are generally high (excluding total_f1000_score and me_readers). The results are shown in Table 7 and the factor plot is shown in Figure 12. The factors for the dataset are clearly aligned along the lines of 1) number of citations, 2) journal impact factor, and 3) social media metrics.

Figure 12: Bornmann and Haunschild EFA Factor Plot under varimax rotation



4. Interpretation

Three data sets were sourced to support the first iteration of this analysis. The principal dataset supporting this analysis comes from Altmetric and comprises data pertaining to the Top 100 ranked papers (by Altmetric score) for the years 2013 through to 2019. Only one of the two additional datasets was also included in the analysis. This came from a dataset supporting an altmetric analysis by Bornmann and Haunschild. The excluded dataset was sourced from Plum Analytics (an Altmetric competitor) and did enhance the finding in this analysis.

Interpreting the findings, three key themes come to the fore:

1. Paper notoriety
2. Paper accessibility
3. Paper subject

These will become evident as the interpretation progresses.

An important observation to make prior to any interpretation is that the Altmetric dataset is likely to be biased in some manner as it only comprises data from the top 100 papers for the 2013 through to 2019. Any interpretation of analysis conducted on this dataset will need this fact for appropriate contextualisation. Figure 13 shows the distribution of the altmetric attention for the Altmetric dataset by year. The plot is truncated around the altmetric attention score of 500. That is, there are no lower values. Figure 14 illustrates this shortcoming more clearly. The boxplot shows the altmetric attention scores from the altmetric dataset alongside the altmetric attention scores from the Bornmann and Haunschild dataset. Notwithstanding an underlying difference in time coverage (Altmetric dataset: 2013 – 2019 [700 datapoints], Bornmann and Haunschild: 2011 -2013 [32,000+ datapoints], and noting the trend to higher top altmetric scores through time (increasing median values), high AASs are likely the exception and not the norm.

This point will be particularly relevant when considering predictive models of AAS (using the Altmetric dataset) as these models (using the Altmetric dataset) will be based on the top scores and not all scores. A full predictive model should be able to generate the full spectrum of altmetric attention scores and not just a subset of the altmetric attention scores. Given this observation, there may be some "chicken and egg" in obtaining a high altmetric attention score. That is, in order to obtain a high altmetric attention a publication must receive considerable attention in the first instance (largely through social media) which may become self-generative (the more a publication gets talked about – the more that publication gets talked about!).

Figure 13: Distribution chart showing count of altmetric attention scores by band and year

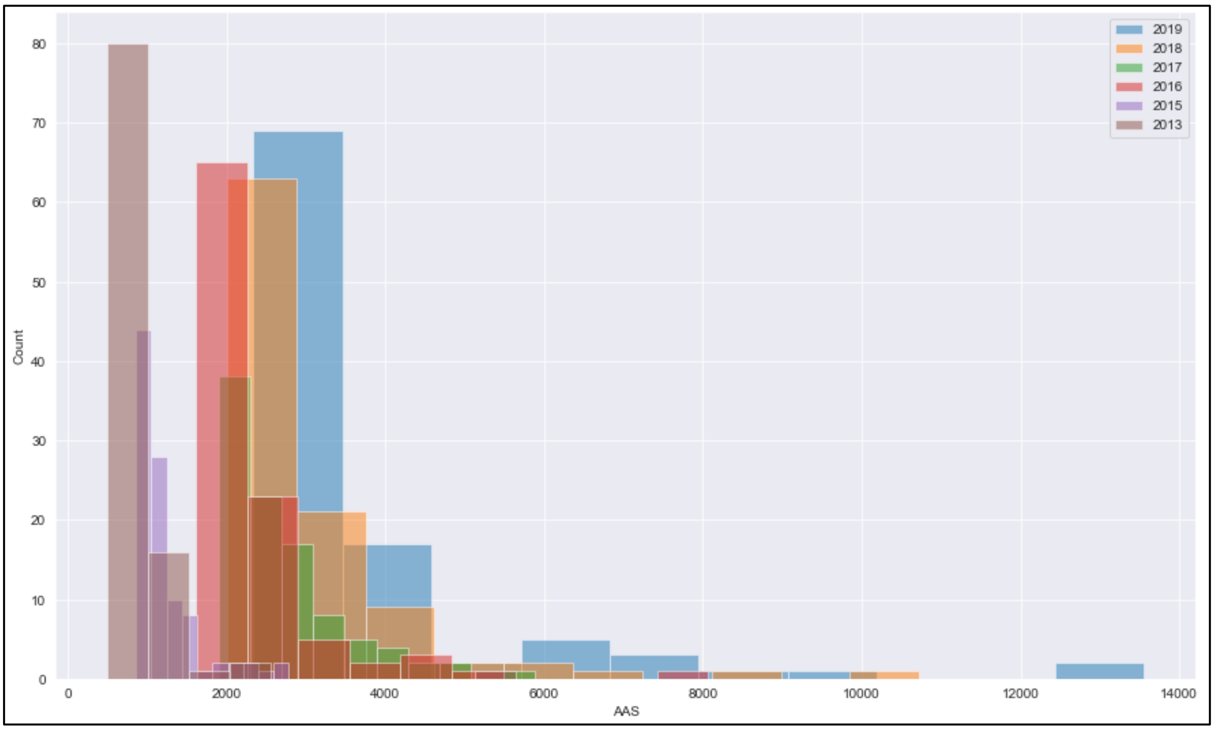
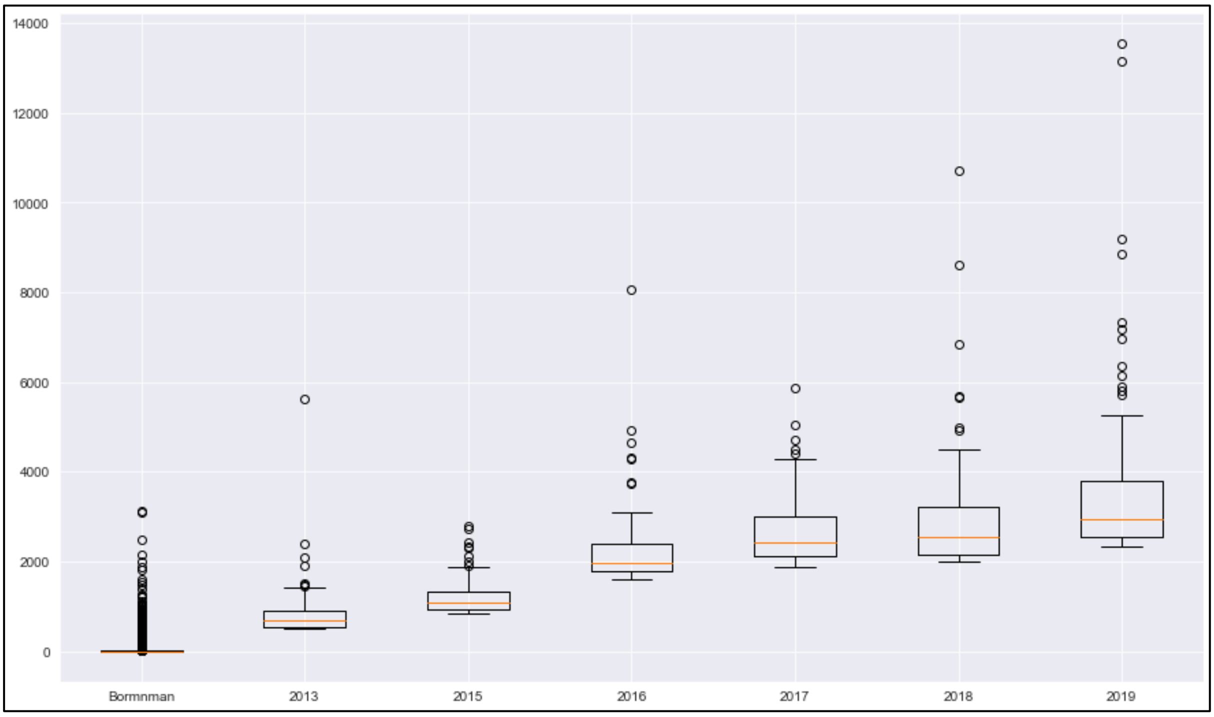


Figure 14: Boxplot of altmetric attention scores for the Bornmann dataset and the altmetric dataset by year



It is apparent from the Altmetric data and the Bornmann and Haunschild data that the AAS is heavily aligned with twitter and other social media mentions (***paper notoriety***). Additionally, it is generally poorly aligned with measures of academic quality (both number of citation numbers and journal impact factors). However, there may be some element of positive reinforcement between both in that a well-recognised paper (either from citation or impact measures may drive some additional attention benefits) noting the smaller cluster in the Bornmann and Haunschild analysis.

Focussing on the Altmetric dataset analysis, the analysis has suggested that within the data available, apart from higher twitter mentions, open access publications and subject, there is little difference between most of the attributes of the Altmetric dataset which could be positively or negatively contributing to AAS. Thus, while the differentiation of clusters exists, this could be attributed to distance measure errors that may be reduced with feature scaling.

The alignment with twitter is reasonably well established (noting twitter weights and relative volume of twitter mentions vis-à-vis other attributes) within the AAS measurement framework. The link to open access is less well established with the potential benefit of Open Access likely derived from the capacity of those who do not have paid access to journals preferring to engage with free sources (***paper accessibility***). Further analysis was conducted to determine if the presence of a paper ID was associated with accessibility (on the assumption that an ID was associated with a paid journal), but the results showed no relationship between accessibility and ID.

Subject matter was another feature of the cluster analysis with medical and health sciences, earth and environmental science, biological sciences featuring prominently within (and between) clusters. There certainly appears to be definite biases in terms of the top-ranking subject categories (***paper subject***).

The association tree analysis reinforced the findings of the cluster analysis. Subject matter (specifically medical and health sciences) and open access showed good uplift in an increased probability of a higher AAS. Interestingly, non-open access appeared to point to a higher probability of a lower AAS.

Decision tree and regression analysis focused on the altmetric data sources which supported predicting AAS. These results show that twitter should be leveraged as much as possible to generate a high AAS for a publication (within this analysis, at least 4708 mentions: ***paper notoriety***). Beyond twitter, the use of news media for supporting a high AAS is also important (again, in this analysis at least 266 news mentions). If news mentions are problematic, there is also some potential to support a high AAS via Wikipedia and it may be worthwhile to also make a publication available on ArXiv.

The regression analysis was derived from the decision tree analysis and inherently supports these findings. Moreover, it shows that there may be an indirect relationship between the data source attributes which was not seen in the broader analysis and which makes facebook mentions a significant variable in the regression model. However, the regression analysis is only applicable where there are substantial social mentions in place for a paper.

From the findings, three key themes have been identified which can support recommendations as to the use of the AAS. These are notoriety, accessibility and subject and each of these appear to have some capacity to influence AAS outcomes. Notoriety reflects the amount of attention that a paper garners through social media. Accessibility reflects the relative ease with which a reader may access a paper (no paywalls). And subject matter reflects the areas that attract the most interest.

Finally, what the AAS does not do is measure the quality of papers in the traditional scientometric sense. It has been reasonably well established that there is nearly no link between social mentions and quality measures. Hence the use of AAS as a measure for a paper must reflect this reality.

5. Recommendations

5.1 Business Objectives

There were a few key findings from the Altmetric and other dataset:

- The AAS is not an academic quality metric.
- The AAS is a social attention metric.
- AAS scores appear biased to open accessibility.
- AAS scores appear biased towards specific subjects.
- AAS scores are modellable, but only within a specific range (as dictated by the Altmetric dataset)

The application of the AAS to business objectives should be aligned with these findings. It would be inappropriate for example, to apply AAS as a quality measure for a paper as AAS reflects social attention and not academic quality. It would be appropriate to use AAS as a measure of attention that a paper achieves. More specifically, if a paper(s) or similar output from an institution are promoted through social media, then the AAS would be an appropriate measure of the relative success of that promotion.

5.2 Recommendations

1. Use AAS as a measure for the social attention that a paper attracts
2. When promoting papers focus on AAS aligned subject matter:
    a. medical & health sciences
    b. earth & environmental science
    c. biological sciences
    d. human society
3. When promoting papers, use open access channels for distribution and arrange for an ID (ArXiv, DOI or PubMed if possible)
4. Promote papers via twitter (moreover, consider encouraging the development of twitter forums or groom existing forums). In addition:
    a. promote through news forums
    b. seek to include on Wikipedia
    c. promote through facebook
5. Benchmark actual AAS outcomes against predicted outcomes to test the efficacy of the predictive model (note that a threshold level of attention applied)

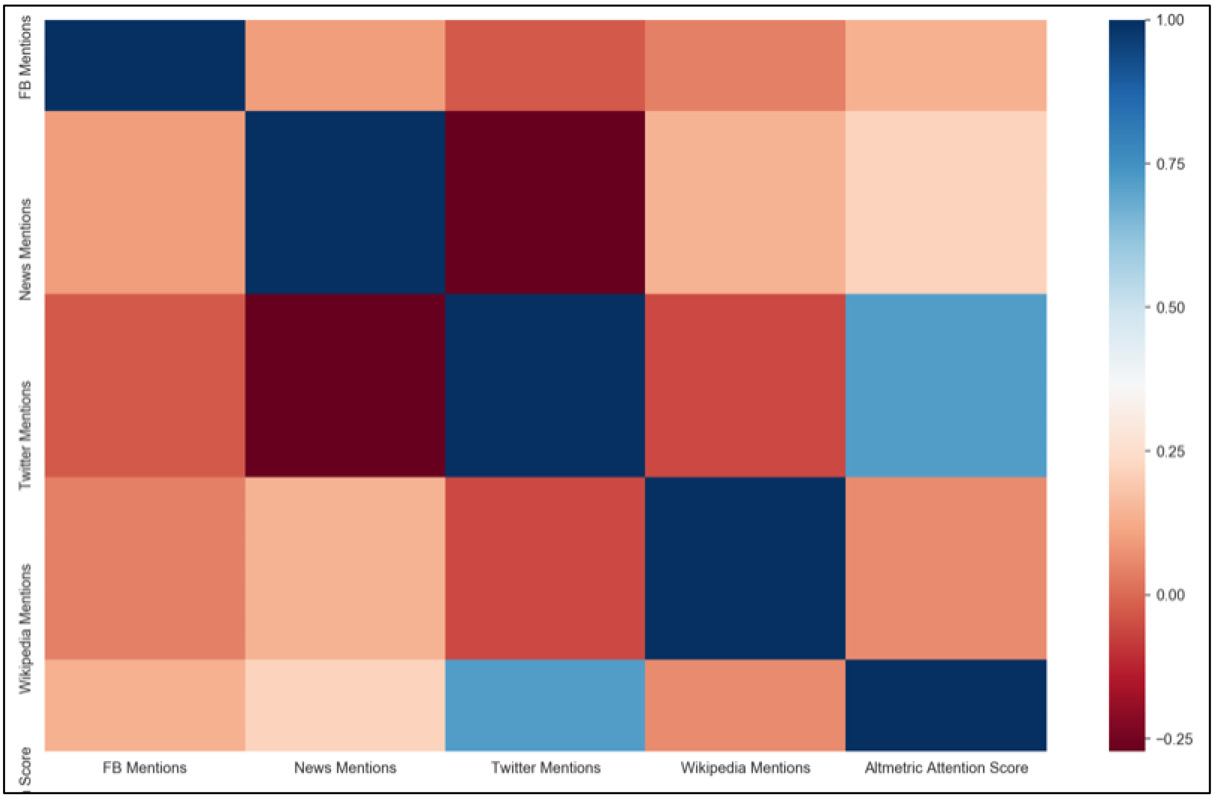
5.3 Future Research

It is important to note that the data used in this report is a biased subset of all available Altmetric data. This means that an unstated, but potentially implicit hypothesis in this analysis is that publications which are in the top 100 should have broadly similar characteristics. However, the dataset is missing attributes between years, and it is missing a full range of AAS outcomes. As a result, the data is likely insufficient to more fully develop these findings. This shortfall in the data was clearly shown in Figure 14.

As the data used is derived from the top 100 publications by AAS from 2013 to 2019, it was difficult to clearly define differences between publication attributes which have a positive effect on AAS. Figure 15 (correlation plot of key altmetric attention score social media metrics) highlights this. Further differentiation by AAS could be due to chance or due to the impact of additional features not available or considered in the dataset.

For future research, it is recommended that more data be sourced. For example, the top 200 or more publications per year or alternatively, a random sampling of all papers (or similar) with an AAS. Either approach would provide a deeper data set to work with. Additionally, the inclusion of all attributes in the dataset may support further resolution in the findings. Finally, more current data should be used to support isolating internet specific effects (noting the that year-on-year traffic and usage is growing significantly).

Figure 15: Correlation plot of key measures in the altmetric attention score (Altmetric dataset)



6. Appendix 1: Methodology and Figures

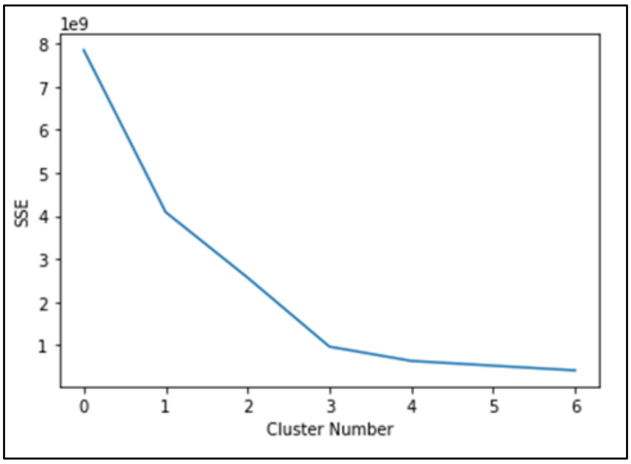
In this appendix, the methodology applied to the Altmetric dataset is detailed.

6.1 Cluster Analysis

K- Prototype is a clustering method for both numerical and categorical variables which uses a combination of k-means and k-mode dissimilarity measures. The goal of using this analysis to provide an overview of the affect, different features available, have on AAS.

Huang's approach of initial prototype selection was used to initialize distinct publication entries to be the starting K clusters. Using the elbow method (Figure 16) of finding optimal clusters, the analysis grouped the available data into three clusters which results in the sum of least squared errors.

Figure 16: Scree plot to identify the optimal number of clusters in the Altmetric dataset



6.2 Association

Association rule mining with Apriori is used to find common occurrences of items in categorical data. Using the K-prototype analysis, a general view was developed of the publication relationships between clusters. However, it was not possible to view potential relationships between publications' attributes. Association mining will determine if useful associations can be found between attributes and an AAS. AAS was converted to a categorial feature with five factor levels for association and the following decision tree analysis. The levels were determined based on the quantiles of all AAS from 2013 to 2019. Figure 17 is a boxplot visualisation of the altmetric attention score data in the Altmetric dataset.

Figure 17: Boxplot of AAS data from the altmetric dataset

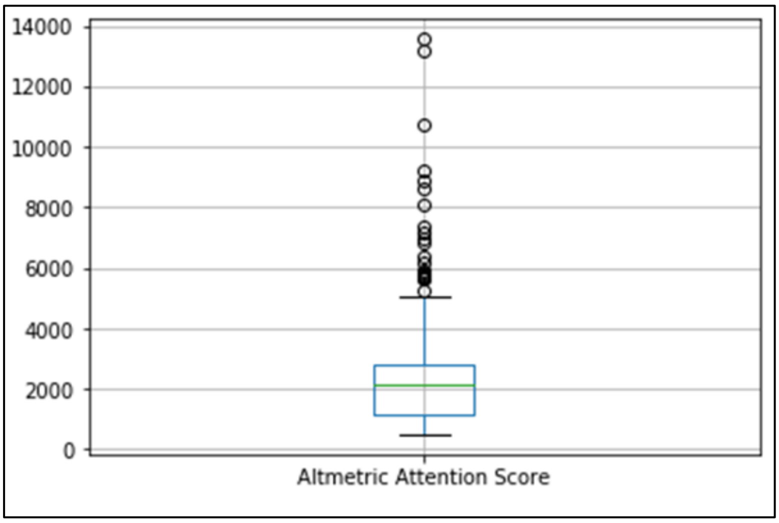
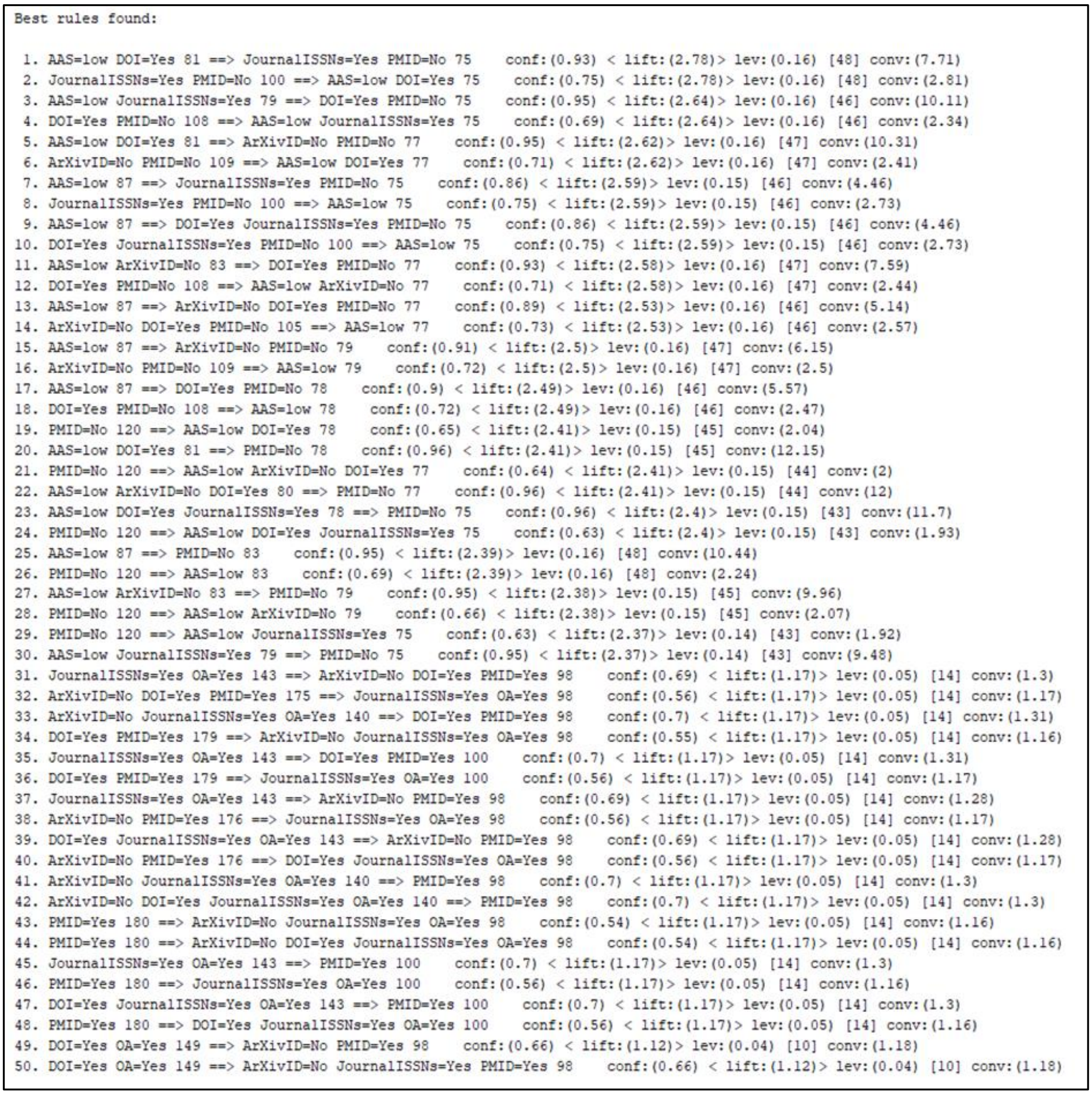
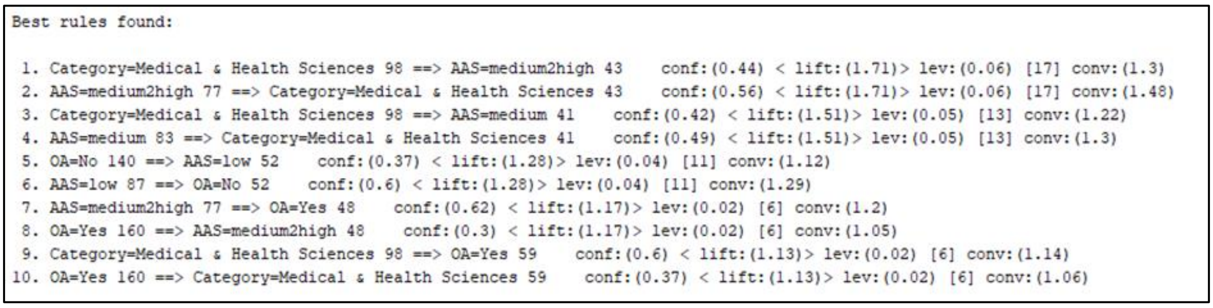


Figure 18: 50 top rules from association analysis (including ID attributes)



```
Best rules found:

 1. AAS=low DOI=Yes 81 ==> JournalISSNs=Yes PMID=No 75    conf:(0.93) < lift:(2.78)> lev:(0.16) [48] conv:(7.71)
 2. JournalISSNs=Yes PMID=No 100 ==> AAS=low DOI=Yes 75    conf:(0.75) < lift:(2.78)> lev:(0.16) [48] conv:(2.81)
 3. AAS=low JournalISSNs=Yes 79 ==> DOI=Yes PMID=No 75    conf:(0.95) < lift:(2.64)> lev:(0.16) [46] conv:(10.11)
 4. DOI=Yes PMID=No 108 ==> AAS=low JournalISSNs=Yes 75    conf:(0.69) < lift:(2.64)> lev:(0.16) [46] conv:(2.34)
 5. AAS=low DOI=Yes 81 ==> ArXivID=No PMID=No 77    conf:(0.95) < lift:(2.62)> lev:(0.16) [47] conv:(10.31)
 6. ArXivID=No PMID=No 109 ==> AAS=low DOI=Yes 77    conf:(0.71) < lift:(2.62)> lev:(0.16) [47] conv:(2.41)
 7. AAS=low 87 ==> JournalISSNs=Yes PMID=No 75    conf:(0.86) < lift:(2.59)> lev:(0.15) [46] conv:(4.46)
 8. JournalISSNs=Yes PMID=No 100 ==> AAS=low 75    conf:(0.75) < lift:(2.59)> lev:(0.15) [46] conv:(2.73)
 9. AAS=low 87 ==> DOI=Yes JournalISSNs=Yes PMID=No 75    conf:(0.86) < lift:(2.59)> lev:(0.15) [46] conv:(4.46)
10. DOI=Yes JournalISSNs=Yes PMID=No 100 ==> AAS=low 75    conf:(0.75) < lift:(2.59)> lev:(0.15) [46] conv:(2.73)
11. AAS=low ArXivID=No 83 ==> DOI=Yes PMID=No 77    conf:(0.93) < lift:(2.58)> lev:(0.16) [47] conv:(7.59)
12. DOI=Yes PMID=No 108 ==> AAS=low ArXivID=No 77    conf:(0.71) < lift:(2.58)> lev:(0.16) [47] conv:(2.44)
13. AAS=low 87 ==> ArXivID=No DOI=Yes PMID=No 77    conf:(0.89) < lift:(2.53)> lev:(0.16) [46] conv:(5.14)
14. ArXivID=No DOI=Yes PMID=No 105 ==> AAS=low 77    conf:(0.73) < lift:(2.53)> lev:(0.16) [46] conv:(2.57)
15. AAS=low 87 ==> ArXivID=No PMID=No 79    conf:(0.91) < lift:(2.5)> lev:(0.16) [47] conv:(6.15)
16. ArXivID=No PMID=No 109 ==> AAS=low 79    conf:(0.72) < lift:(2.5)> lev:(0.16) [47] conv:(2.5)
17. AAS=low 87 ==> DOI=Yes PMID=No 78    conf:(0.9) < lift:(2.49)> lev:(0.16) [46] conv:(5.57)
18. DOI=Yes PMID=No 108 ==> AAS=low 78    conf:(0.72) < lift:(2.49)> lev:(0.16) [46] conv:(2.47)
19. PMID=No 120 ==> AAS=low DOI=Yes 78    conf:(0.65) < lift:(2.41)> lev:(0.15) [45] conv:(2.04)
20. AAS=low DOI=Yes 81 ==> PMID=No 78    conf:(0.96) < lift:(2.41)> lev:(0.15) [45] conv:(12.15)
21. PMID=No 120 ==> AAS=low ArXivID=No DOI=Yes 77    conf:(0.64) < lift:(2.41)> lev:(0.15) [44] conv:(2)
22. AAS=low ArXivID=No DOI=Yes 80 ==> PMID=No 77    conf:(0.96) < lift:(2.41)> lev:(0.15) [44] conv:(12)
23. AAS=low DOI=Yes JournalISSNs=Yes 78 ==> PMID=No 75    conf:(0.96) < lift:(2.4)> lev:(0.15) [43] conv:(11.7)
24. PMID=No 120 ==> AAS=low DOI=Yes JournalISSNs=Yes 75    conf:(0.63) < lift:(2.4)> lev:(0.15) [43] conv:(1.93)
25. AAS=low 87 ==> PMID=No 83    conf:(0.95) < lift:(2.39)> lev:(0.16) [48] conv:(10.44)
26. PMID=No 120 ==> AAS=low 83    conf:(0.69) < lift:(2.39)> lev:(0.16) [48] conv:(2.24)
27. AAS=low ArXivID=No 83 ==> PMID=No 79    conf:(0.95) < lift:(2.38)> lev:(0.15) [45] conv:(9.96)
28. PMID=No 120 ==> AAS=low ArXivID=No 79    conf:(0.66) < lift:(2.38)> lev:(0.15) [45] conv:(2.07)
29. PMID=No 120 ==> AAS=low JournalISSNs=Yes 75    conf:(0.63) < lift:(2.37)> lev:(0.14) [43] conv:(1.92)
30. AAS=low JournalISSNs=Yes 79 ==> PMID=No 75    conf:(0.95) < lift:(2.37)> lev:(0.14) [43] conv:(9.48)
31. JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No DOI=Yes PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
32. ArXivID=No DOI=Yes PMID=Yes 175 ==> JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
33. ArXivID=No JournalISSNs=Yes OA=Yes 140 ==> DOI=Yes PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.31)
34. DOI=Yes PMID=Yes 179 ==> ArXivID=No JournalISSNs=Yes OA=Yes 98    conf:(0.55) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
35. JournalISSNs=Yes OA=Yes 143 ==> DOI=Yes PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.31)
36. DOI=Yes PMID=Yes 179 ==> JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
37. JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.28)
38. ArXivID=No PMID=Yes 176 ==> JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
39. DOI=Yes JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.28)
40. ArXivID=No PMID=Yes 176 ==> DOI=Yes JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
41. ArXivID=No JournalISSNs=Yes OA=Yes 140 ==> PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
42. ArXivID=No DOI=Yes JournalISSNs=Yes OA=Yes 140 ==> PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
43. PMID=Yes 180 ==> ArXivID=No JournalISSNs=Yes OA=Yes 98    conf:(0.54) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
44. PMID=Yes 180 ==> ArXivID=No DOI=Yes JournalISSNs=Yes OA=Yes 98    conf:(0.54) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
45. JournalISSNs=Yes OA=Yes 143 ==> PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
46. PMID=Yes 180 ==> JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
47. DOI=Yes JournalISSNs=Yes OA=Yes 143 ==> PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
48. PMID=Yes 180 ==> DOI=Yes JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
49. DOI=Yes OA=Yes 149 ==> ArXivID=No PMID=Yes 98    conf:(0.66) < lift:(1.12)> lev:(0.04) [10] conv:(1.18)
50. DOI=Yes OA=Yes 149 ==> ArXivID=No JournalISSNs=Yes PMID=Yes 98    conf:(0.66) < lift:(1.12)> lev:(0.04) [10] conv:(1.18)
```

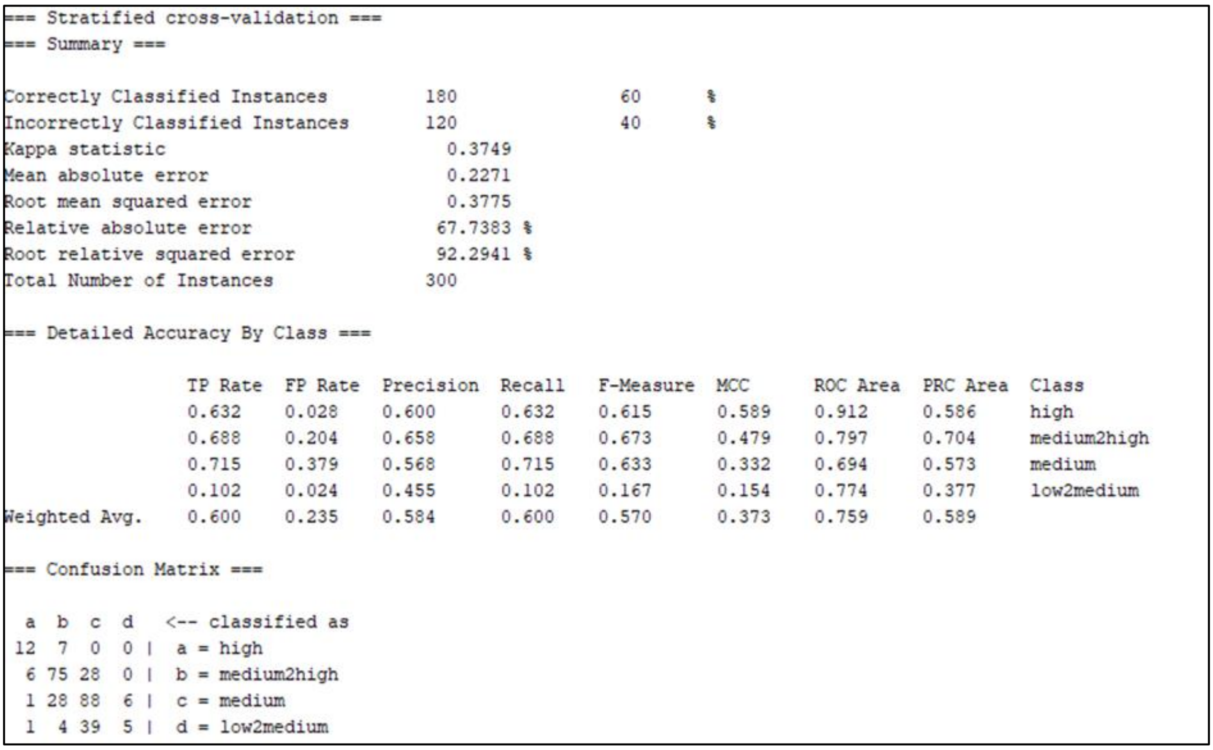
Figure 19: 10 top rules from association analysis (excluding ID attributes)



```
Best rules found:

 1. Category=Medical & Health Sciences 98 ==> AAS=medium2high 43    conf:(0.44) < lift:(1.71)> lev:(0.06) [17] conv:(1.3)
 2. AAS=medium2high 77 ==> Category=Medical & Health Sciences 43    conf:(0.56) < lift:(1.71)> lev:(0.06) [17] conv:(1.48)
 3. Category=Medical & Health Sciences 98 ==> AAS=medium 41    conf:(0.42) < lift:(1.51)> lev:(0.05) [13] conv:(1.22)
 4. AAS=medium 83 ==> Category=Medical & Health Sciences 41    conf:(0.49) < lift:(1.51)> lev:(0.05) [13] conv:(1.3)
 5. OA=No 140 ==> AAS=low 52    conf:(0.37) < lift:(1.28)> lev:(0.04) [11] conv:(1.12)
 6. AAS=low 87 ==> OA=No 52    conf:(0.6) < lift:(1.28)> lev:(0.04) [11] conv:(1.29)
 7. AAS=medium2high 77 ==> OA=Yes 48    conf:(0.62) < lift:(1.17)> lev:(0.02) [6] conv:(1.2)
 8. OA=Yes 160 ==> AAS=medium2high 48    conf:(0.3) < lift:(1.17)> lev:(0.02) [6] conv:(1.05)
 9. Category=Medical & Health Sciences 98 ==> OA=Yes 59    conf:(0.6) < lift:(1.13)> lev:(0.02) [6] conv:(1.14)
10. OA=Yes 160 ==> Category=Medical & Health Sciences 59    conf:(0.37) < lift:(1.13)> lev:(0.02) [6] conv:(1.06)
```

6.3 Decision Tree

Figure 20: Decision tree results



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         180               60      %
Incorrectly Classified Instances       120               40      %
Kappa statistic                          0.3749
Mean absolute error                      0.2271
Root mean squared error                  0.3775
Relative absolute error                 67.7383 %
Root relative squared error             92.2941 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.632    0.028    0.600      0.632    0.615      0.589    0.912     0.586     high
                 0.688    0.204    0.658      0.688    0.673      0.479    0.797     0.704     medium2high
                 0.715    0.379    0.568      0.715    0.633      0.332    0.694     0.573     medium
                 0.102    0.024    0.455      0.102    0.167      0.154    0.774     0.377     low2medium
Weighted Avg.    0.600    0.235    0.584      0.600    0.570      0.373    0.759     0.589

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 12  7  0  0 |  a = high
  6 75 28  0 |  b = medium2high
  1 28 88  6 |  c = medium
  1  4 39  5 |  d = low2medium
```

Figure 21: Decision tree ROC: high

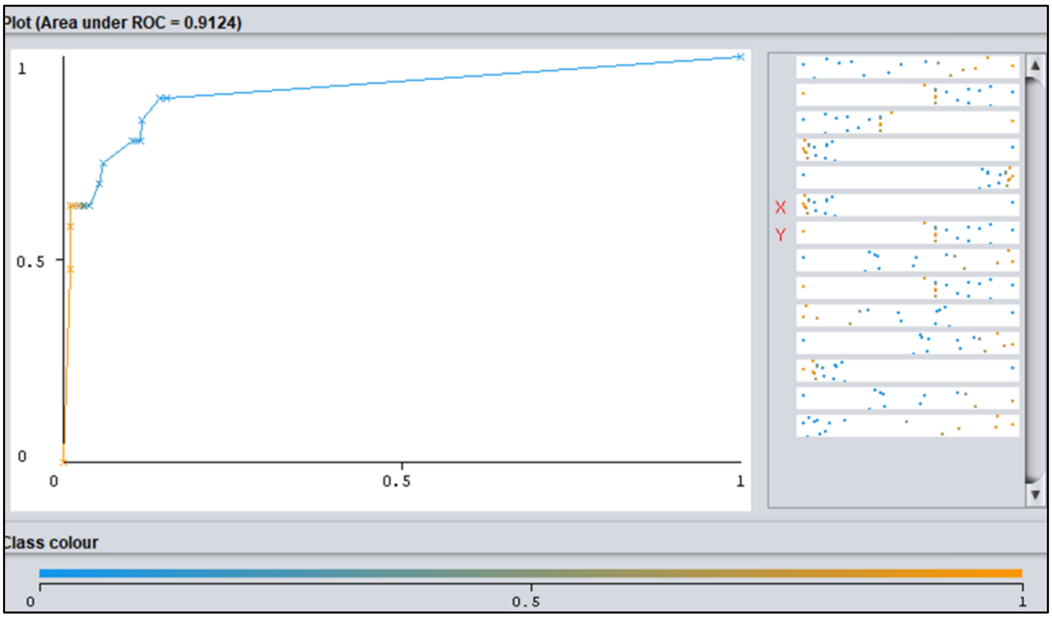


Figure 22: Decision tree ROC: medium2high

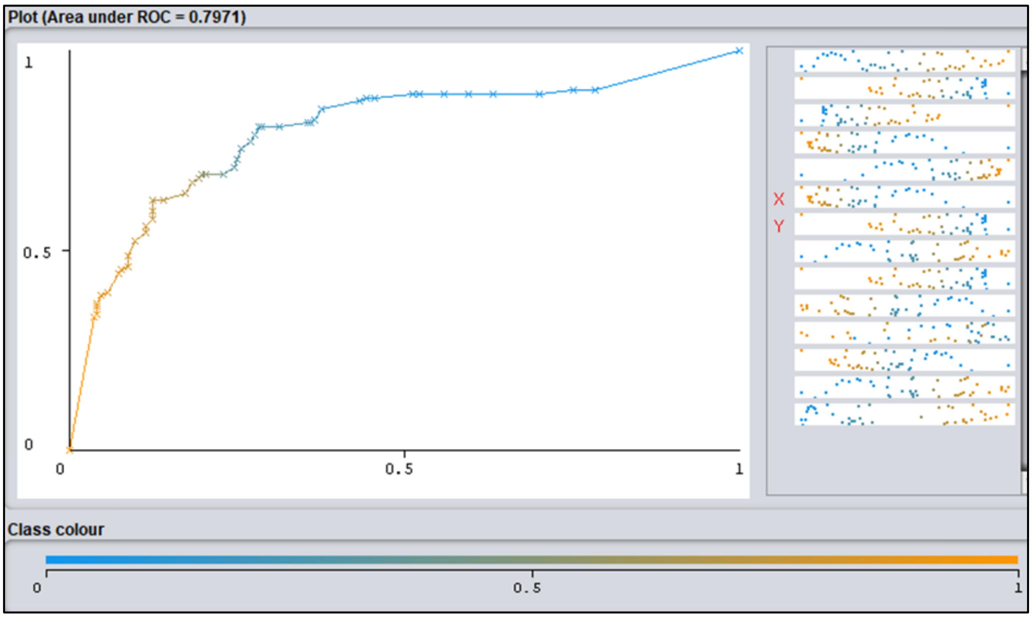


Figure 23: Decision tree ROC: medium

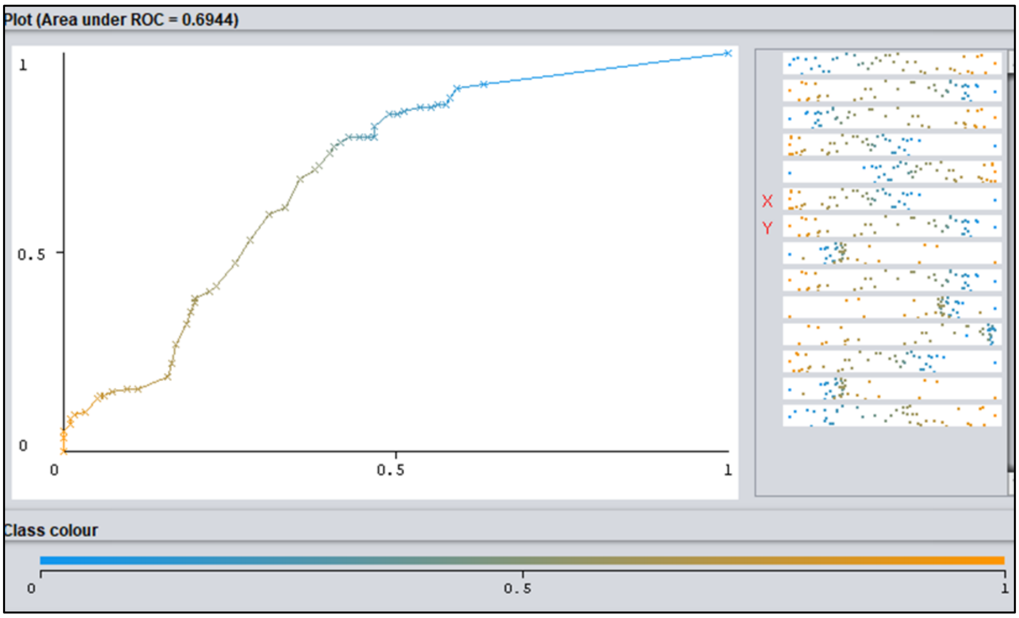
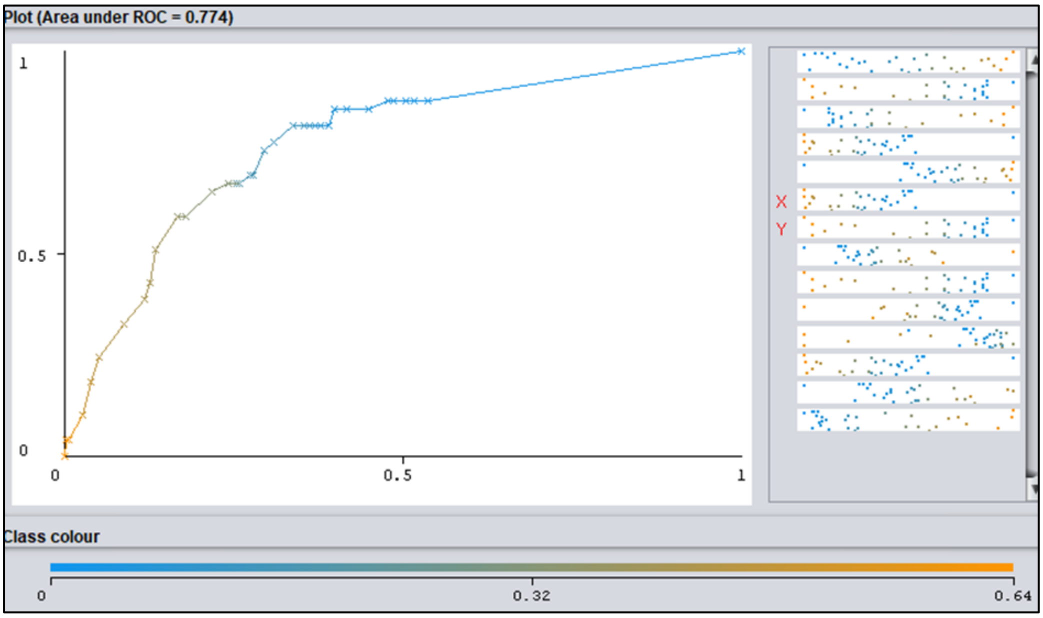


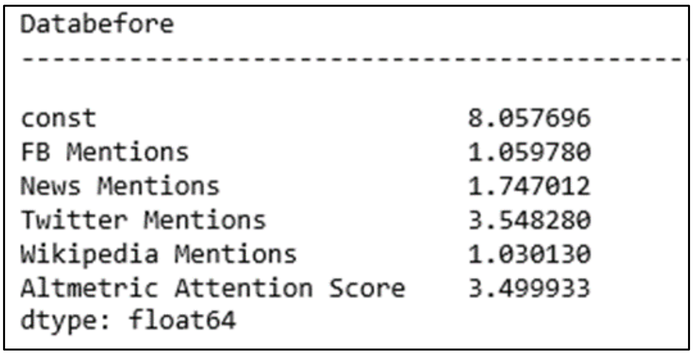
Figure 24: Decision tree ROC: low2medium



6.4 Multiple Linear Regression

The key model assumptions considered were multicollinearity, normally distributed residuals and homoscedasticity. For evaluation, R2 and a 5% significance level were used. Only data from 2017 through to 2019 was considered. The first model used all numerical altmetric attributes (R2=0.757 with 5 features after removing significant features). The second model only used the predictive features resulting from the decision tree (R2 =0.712 with 3 features). The simpler model will be used as it includes only the features from the decision tree (excluding wikipedia mentions).

Figure 25: Multiple linear regression analysis multicollinearity check (second model)



```
Databefore
------------------------------------------

const                         8.057696
FB Mentions                   1.059780
News Mentions                 1.747012
Twitter Mentions              3.548280
Wikipedia Mentions            1.030130
Altmetric Attention Score     3.499933
dtype: float64
```

Figure 26: Multiple linear regression – normally distributed residuals (second model)

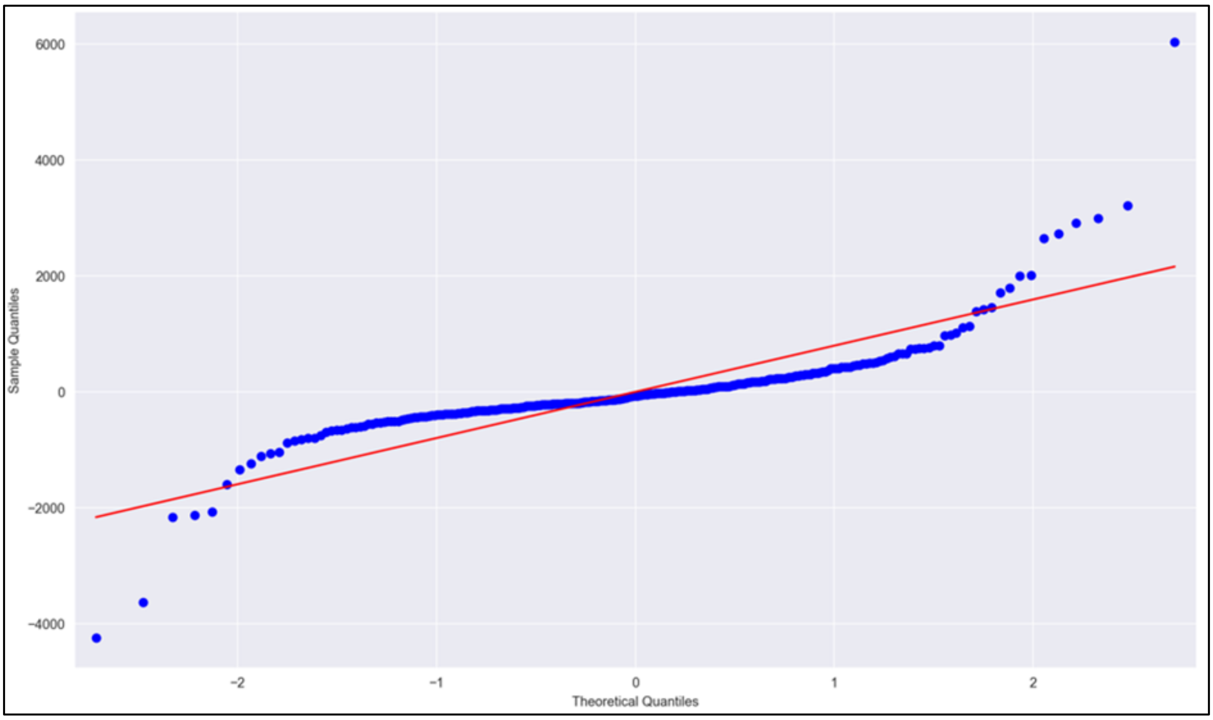
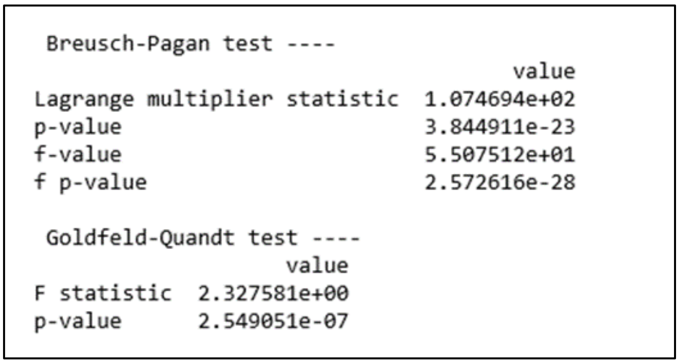
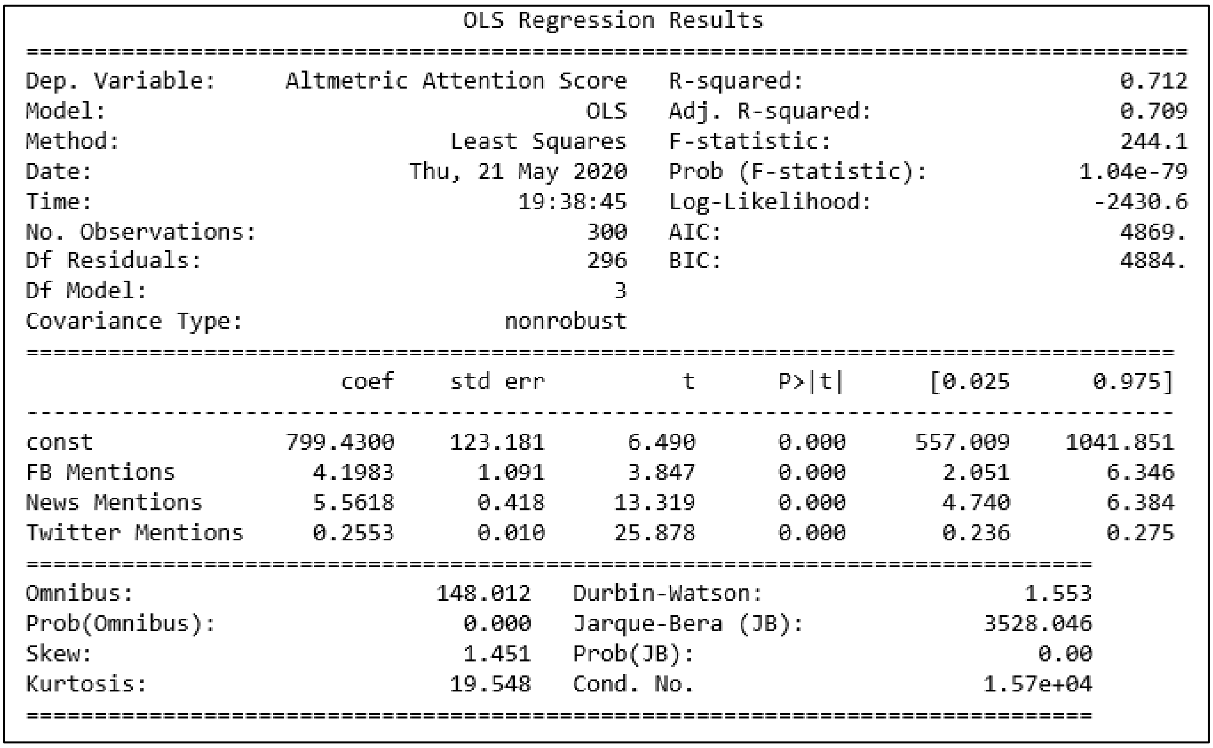


Figure 27: Homoscedasticity tests (second model)



```
 Breusch-Pagan test ----
                                     value
Lagrange multiplier statistic  1.074694e+02
p-value                        3.844911e-23
f-value                        5.507512e+01
f p-value                      2.572616e-28

 Goldfeld-Quandt test ----
                   value
F statistic  2.327581e+00
p-value      2.549051e-07
```

Figure 28: Regression results output (second model)



```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     Altmetric Attention Score   R-squared:                       0.712
Model:                                   OLS   Adj. R-squared:                  0.709
Method:                        Least Squares   F-statistic:                     244.1
Date:                       Thu, 21 May 2020   Prob (F-statistic):           1.04e-79
Time:                               19:38:45   Log-Likelihood:                -2430.6
No. Observations:                        300   AIC:                             4869.
Df Residuals:                            296   BIC:                             4884.
Df Model:                                  3
Covariance Type:                   nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              799.4300    123.181      6.490      0.000     557.009    1041.851
FB Mentions          4.1983      1.091      3.847      0.000       2.051       6.346
News Mentions        5.5618      0.418     13.319      0.000       4.740       6.384
Twitter Mentions     0.2553      0.010     25.878      0.000       0.236       0.275
==============================================================================
Omnibus:                      148.012   Durbin-Watson:                   1.553
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3528.046
Skew:                           1.451   Prob(JB):                         0.00
Kurtosis:                      19.548   Cond. No.                     1.57e+04
==============================================================================
```

7. Appendix 2: Contribution Table

Contribution Table	**Stephen**	**Jinxi**
Introduction	x	x
Findings Altmetric	3.1.1, 3.1.4	x
Findings Bornmann & Haunschild	x	
Interpretation	x	x
Recommendations	x	x
Appendix 1	x	x
Appendix 3		x
Appendix 4	x	

8. Appendix 3: Jinxi Luo

Introduction

Assignment 1 exploratory analysis resulted in elimination of features that have been suggested to have little association with altmetric attention score (AAS). This second study will further focus on the remaining features. Clustering, association and regression analysis will be used to evaluate relationships seen previously with statistical significance.

The available variables used in this report can be found in the table below:

Discrete (quantitative) [purple]
(Altmetric Attention Score) (Blog Mentions) (F1000 Mentions) (FB Mentions) (Google+ Mentions) (News Mentions) (Patent Mentions) (Policy Mentions) (Reddit Mentions) (Twitter Mentions) (Video Mentions) (Wikipedia Mentions)
Binary (qualitative) [green]
(ArXiv ID) (Dimensions ID) (DOI)(Journal ISSNs) (PubMed ID) (OA)
Nominal (qualitative) [red]
(Category) (Journal)

The main goal of this report is to make findings on features that impact altmetric attention score, as features are not complete throughout the years, this longitudinal research will involve sub-setting data for analysis that have the available features to above as much data as possible for better model training. Thus, before we perform analysis it is helpful to view the features available in each dataset 2013-2019 after initial preparation; where grey shows features missing from the 2018 data set that contains the most complete set of features.

2013	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Altmetric Attention Score	Category	Altmetric Attention Score	Altmetric Attention Score	Altmetric Attention Score	Altmetric Attention Score	Altmetric Attention Score
ArXiv ID	DOI	Blog Mentions	Blog Mentions	ArXiv ID	ArXiv ID	ArXiv ID
Blog Mentions	Journal	Category	Category	Blog Mentions	Blog Mentions	Blog Mentions
Category	OA	DOI	DOI	Category	Category	Category
DOI	Altmetric Attention Score	FB Mentions	F1000 Mentions	DOI	DOI	DOI
F1000 Mentions	ArXiv ID	Google+ Mentions	FB Mentions	F1000 Mentions	F1000 Mentions	F1000 Mentions
FB Mentions	Blog Mentions	Journal	Google+ Mentions	FB Mentions	FB Mentions	FB Mentions
Google+ Mentions	F1000 Mentions	News Mentions	Journal	Google+ Mentions	Google+ Mentions	Google+ Mentions
Journal	FB Mentions	OA	News Mentions	Journal	Journal	Journal
Journal ISSNs	Google+ Mentions	Twitter Mentions	OA	News Mentions	Journal ISSNs	Journal ISSNs
News Mentions	Journal ISSNs	Video Mentions	Policy Mentions	OA	News Mentions	News Mentions
OA	News Mentions	Wikipedia Mentions	Reddit Mentions	PMID	OA	OA
PMID	Patent Mentions	ArXiv ID	Twitter Mentions	Policy Mentions	Patent Mentions	Patent Mentions
Reddit Mentions	PMID	F1000 Mentions	Video Mentions	Reddit Mentions	PMID	PMID
Twitter Mentions	Policy Mentions	Journal ISSNs	Wikipedia Mentions	Twitter Mentions	Policy Mentions	Policy Mentions
Patent Mentions	Reddit Mentions	Patent Mentions	ArXiv ID	Video Mentions	Reddit Mentions	Twitter Mentions
Policy Mentions	Twitter Mentions	PMID	Journal ISSNs	Wikipedia Mentions	Twitter Mentions	Video Mentions
Video Mentions	Video Mentions	Policy Mentions	Patent Mentions	Journal ISSNs	Video Mentions	Wikipedia Mentions
Wikipedia Mentions	Wikipedia Mentions	Reddit Mentions	PMID	Patent Mentions	Wikipedia Mentions	Reddit Mentions

K-Prototype

Technique information

- K- Prototype is a clustering method for both numerical and categorical variables; which uses a combination of k-means and k-mode dissimilarity measures.

Goal

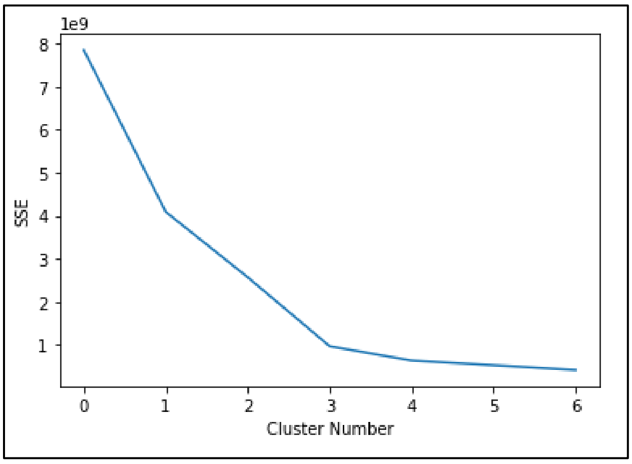
- The use of K-Prototype serves as an overview of the affect features available have on AAS. As previously done in assignment 1, only examination of a few variables alongside their correlation to AAS was done. This did not allow us to specifically see if the combination of all those factors would agree with individual results. Using K-Prototype, this process can be performed backwards, then applying more specific methods to examine peculiarities.

Method and Considerations

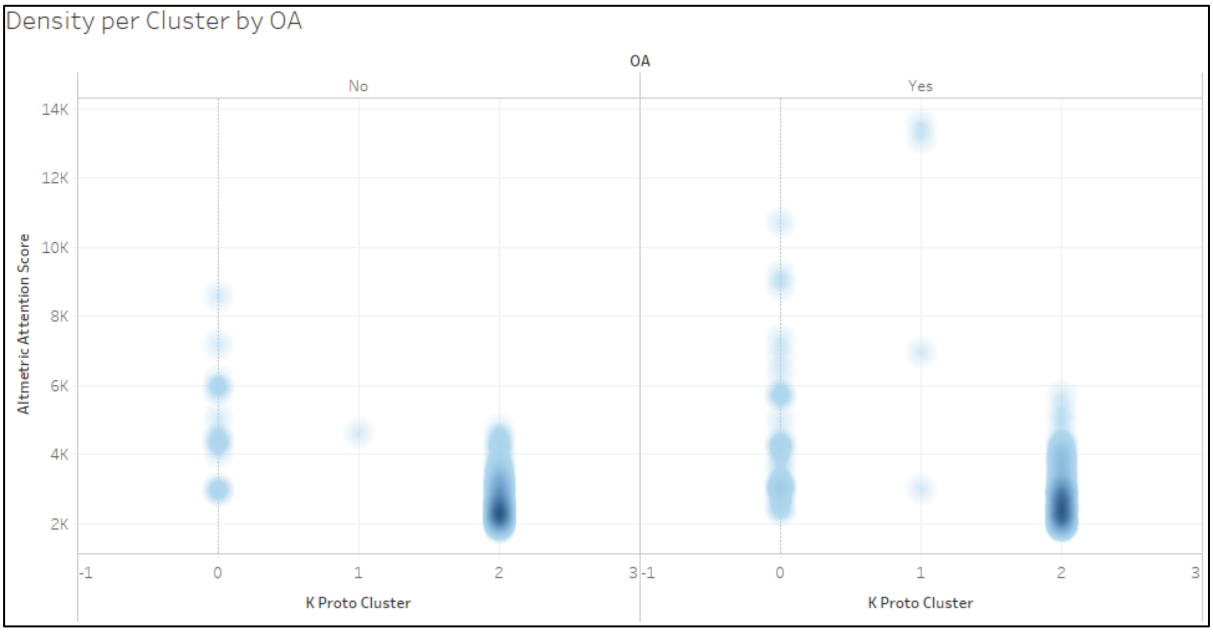
- In order to source as much data as possible from the data set and not introduce new assumptions that might affect the results. Entries from 2017 to 2019 will be combined to create a total of 300 entries.
- Journal ISSNs, patent mentions, Reddit mentions will also be removed from the 2018 and 2019 columns. This is to decrease further assumptions that are added to the data, if we were to try and compensate for the missing columns in the 2017 and 2019 data sets.
- Elbow method will be used to find the optimal clusters
- Feature scaling will not be applied to the data as it results in 1 optimal cluster which does not provide useful information for initial investigation.
- Huang approach of initial prototype selection will be used as initial distinct entries from the data set provides a more randomized selection of clustering in our dataset. And we are interested in more potential clusters as possible which is signified potentially by unique publication attributes.

Results

- Optimal clusters are found to be 3 (same when using Cao), further information about entries in each cluster can be found in this report's Jupyter notebook. Also see Tableau file for more visualizations.

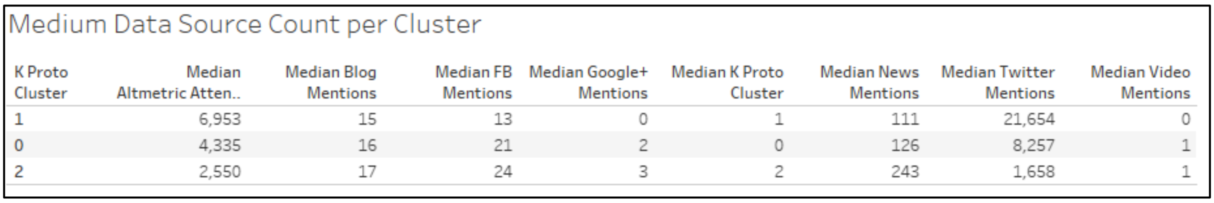


- Publication density’s AAS per cluster can be seen in graph below. It is seen that most publications are rated with AAS score in range 2000 to 5000 as found in cluster 2. While clusters 0 and 1 seems to be have outlier AAS per feature data. It will be of interest to investigate publications in clusters 0 and 1 as they also contain some of the highest AAS. This has a potential to show that factors attributing to a high AAS may be found in these clusters or that they are just outliers and that cluster 2 publications will is more consistent and relevant to our investigation. OA type seems to have a little positive affect on score.



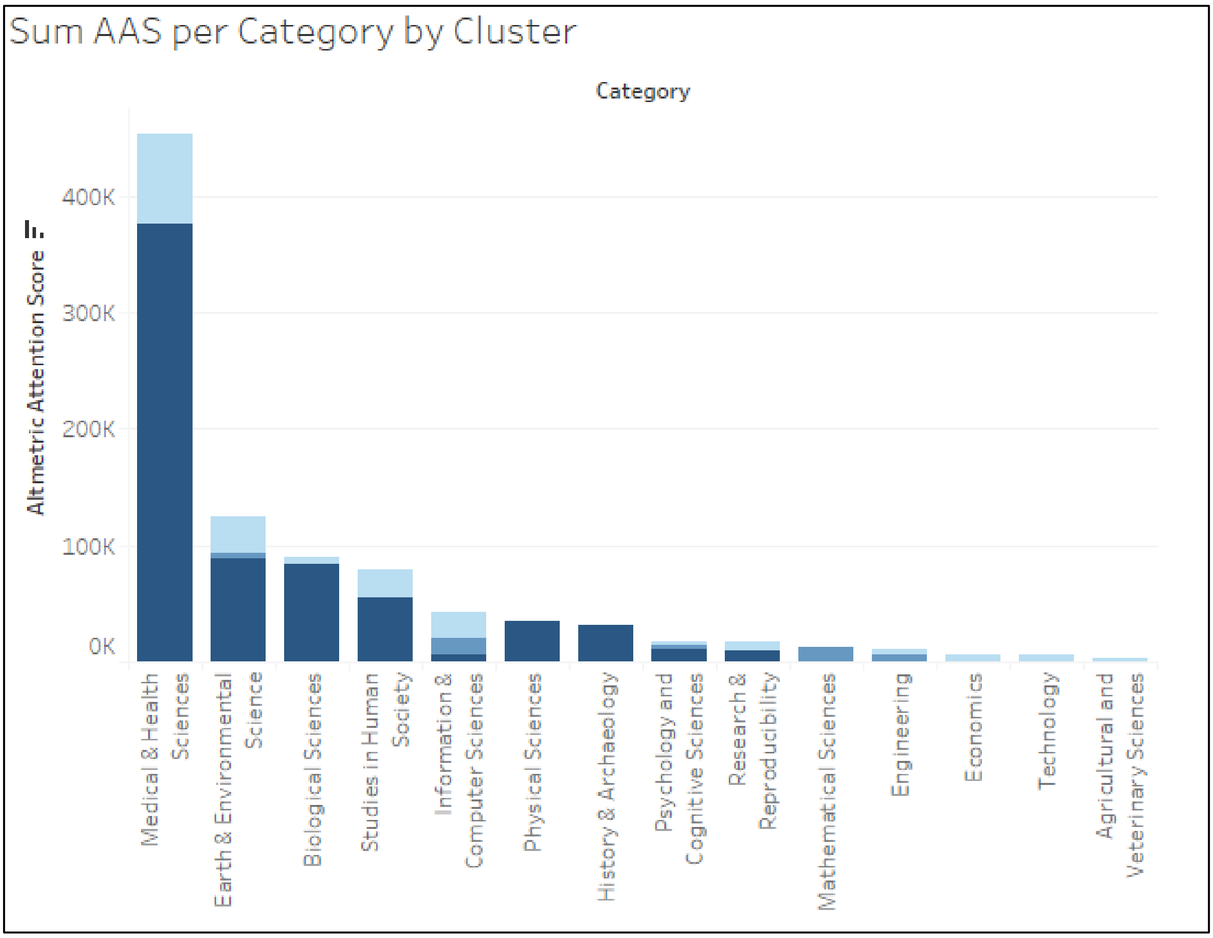
- The table below shows the medium count of altmetric data sources per cluster; medium is used as seen from graph above average may skew results in clusters 0 and 1. Note that, F1000, Policy, Wikipedia had no mentions and were removed from the table.

- For table below it is seen that clusters 0 and 1 had the highest twitter counts which also was the highest AAS. And while cluster 2 had the highest news mentions in terms of data source variation compared to other clusters, it also had the lowest twitter mentions. This suggests that a high twitter mention count is very likely the cause of the high/unusual AAS in clusters 0 and 1 with other data sources constant.

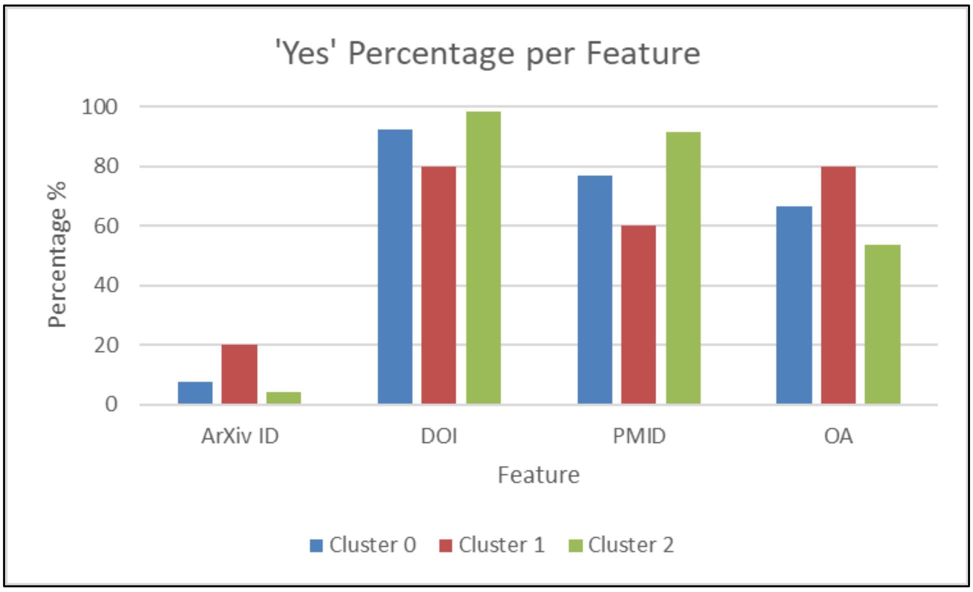

Medium Data Source Count per Cluster

K Proto Cluster	Median Altmetric Atten..	Median Blog Mentions	Median FB Mentions	Median Google+ Mentions	Median K Proto Cluster	Median News Mentions	Median Twitter Mentions	Median Video Mentions
1	6,953	15	13	0	1	111	21,654	0
0	4,335	16	21	2	0	126	8,257	1
2	2,550	17	24	3	2	243	1,658	1

- The next graph examines the contribution to sum altmetric score per category grouped by cluster contribution. With cluster 0 light blue, cluster 1 medium light blue and cluster 2 dark blue. We see that the majority of AAS for cluster 1 is in maths, comp sci and engineering.
- While the top three categories for cluster 2 is in health, earth and biological science, this is like cluster 0 except that studies in human society has higher contribution than in biological science.



- For the binary level features, we look at the percentage of Yes per cluster and there does not appear to be a trend.



Key Take Away

- There does not seem to be any major difference between the clusters. This suggests that the use of feature scaling would have been appropriate also as it helps to reduce clustering errors due to different scales.
- Results in the most common cluster, cluster 2. Is expected as we have seen previously in assignment one.
- The higher AAS in clusters 0 and 1 seems to be attributed to higher twitter counts with everything else consistent with cluster 2. Viewing the outlier AAS in these two clusters suggests that they were classified separate from cluster 2 simply for their high AAS and categories.

- The clustering analysis grouped the available data into three clusters which results in the least sum of square (SSE) error . Comparing the distribution of features across the three clusters suggests that no major differentiation exists between the clusters in general besides publication category, twitter mentions and altmetric score. Investigation into the effect of these three features within the clusters shows that in general a higher twitter mentions was associated with higher altmetric score. This result was expected from explanatory analysis in assignment 1, however it was also found that publication category also had an influence in cluster assignment, but this did not have an obvious difference that may not be attributed to other factors not considered in this analysis.
- In short, the analysis has suggested that within the data available there is little difference between publication features used that could be related to positively or negatively contributing to AAS.

Association

Technique information

- Association rule mining with Apriori is used to find common occurrences of items in categorical data

Goal

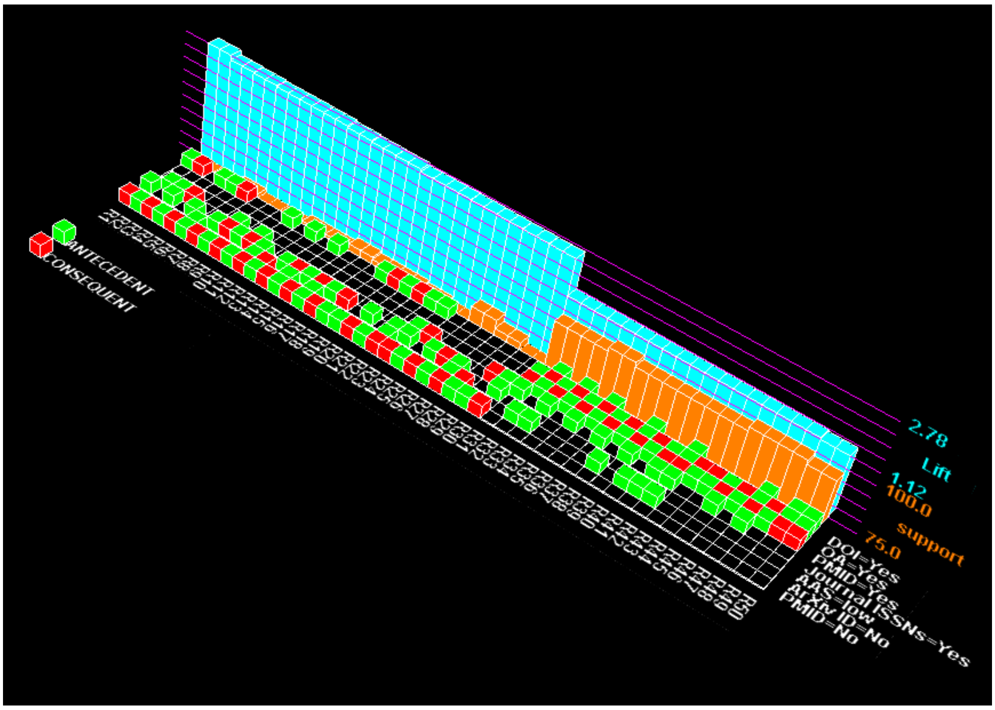
- From K prototype, we were able to get a general view of how each publication related to each other, however with clusters we were not able to see directly the potential relationships between publications' features. With Association mining, it is possible to focus only on those this goal and see if useful associations can be found with them and AAS.

Method and Considerations

- As Apriori uses categorical variables, AAS has been recoded to bins. This was determined by quantiles of all AAS from 2013 to 2019.
- The data used will be that of 2013, 2018 and 2019 as they contain the most complete categorical features.
- Since we have little domain knowledge about confidence, support and lift ranges, they will be kept at default and sorted by lift. This allows the examination of the most common important and common occurrences of items together. A Total of 50 rules will be created.
- For this analysis we are mainly interested when AAS appears in the RHS of a rule.

Results

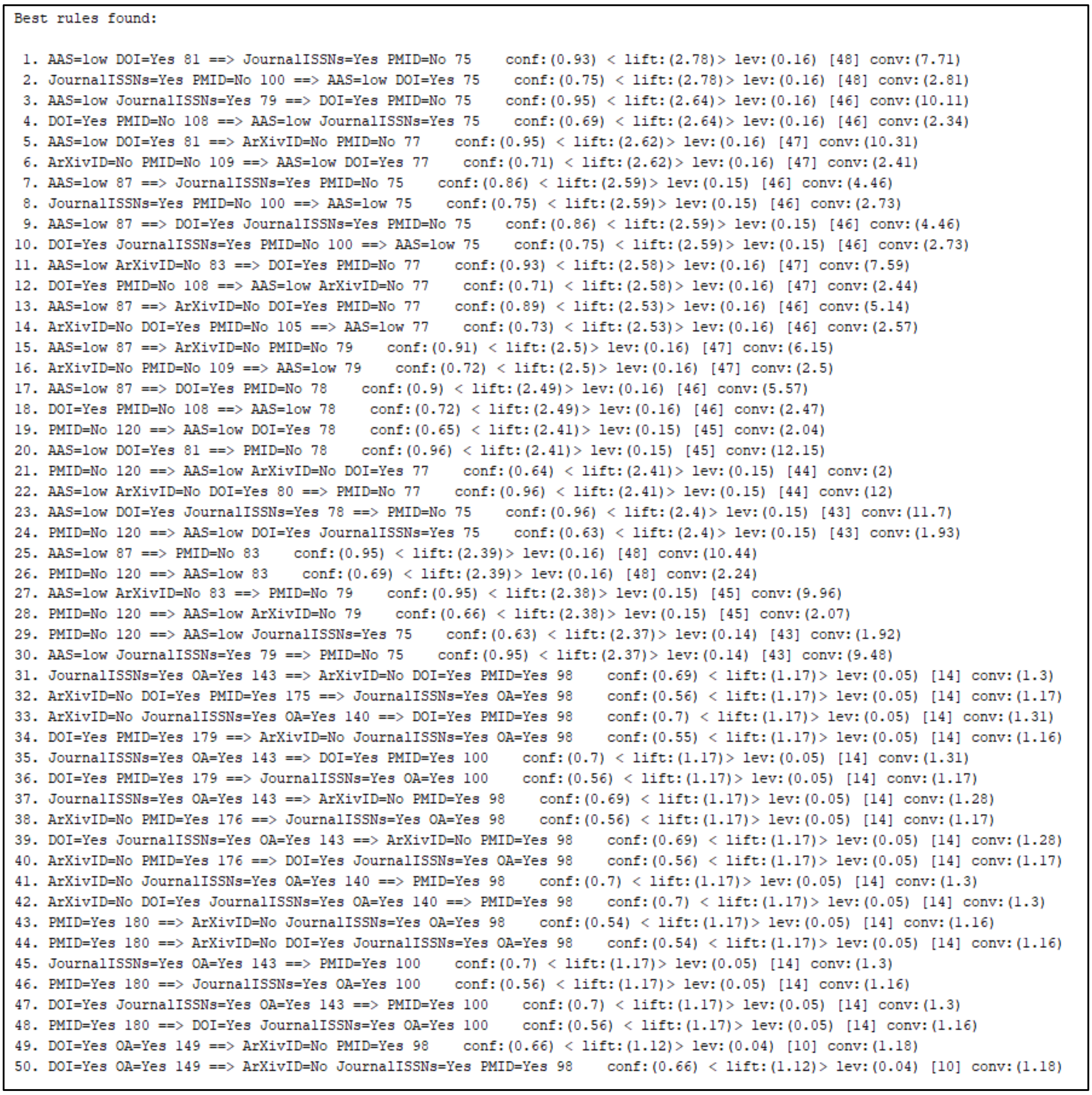
- Below is a visualization of the best 50 rules. As the rules are sorted by lift the top 30 rules have high lift but low support. Interestingly support only increases after the 30th rule.
- This says that the first 30 rules have are good and LHS strongly implies RHS as lift is greater than 1, while the rest are less good but still good in terms of occurring more frequently.



- Below is a table of the top 5 rules from above

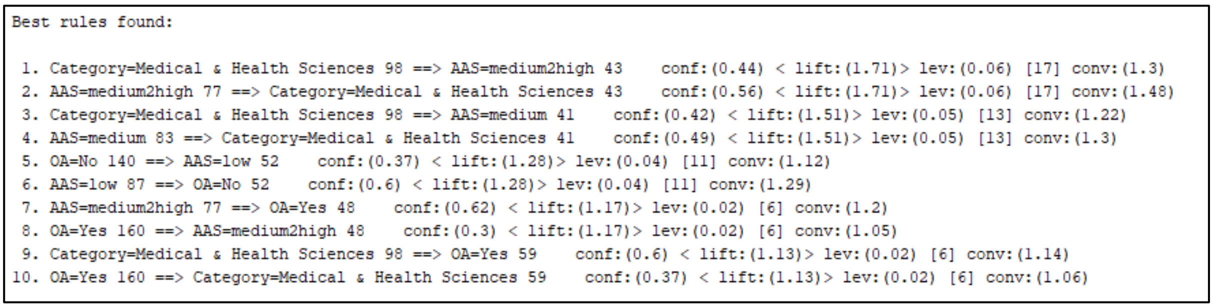
Rule	LHS	RHS	Confidence	Leverage	Lift
1	AAS=low DOI=Yes	JournalISSNs=Yes PMID=No	0.93	0.16	2.78
2	JournalISSNs=Yes PMID=No	ASS=low DOI=Yes	0.75	0.16	2.78
3	AAS=low JournalISSNs=Yes	DOI=Yes PMID=No	0.95	0.16	2.64
4	DOI=Yes PMID=No	AAS=low JournalISSNs=Yes	0.69	0.16	2.64
5	AAS=low DOI=Yes	ArXivID=No PMID=No	0.95	0.16	2.62

- Below are the total 50 rules



```
Best rules found:

 1. AAS=low DOI=Yes 81 ==> JournalISSNs=Yes PMID=No 75    conf:(0.93) < lift:(2.78)> lev:(0.16) [48] conv:(7.71)
 2. JournalISSNs=Yes PMID=No 100 ==> AAS=low DOI=Yes 75    conf:(0.75) < lift:(2.78)> lev:(0.16) [48] conv:(2.81)
 3. AAS=low JournalISSNs=Yes 79 ==> DOI=Yes PMID=No 75    conf:(0.95) < lift:(2.64)> lev:(0.16) [46] conv:(10.11)
 4. DOI=Yes PMID=No 108 ==> AAS=low JournalISSNs=Yes 75    conf:(0.69) < lift:(2.64)> lev:(0.16) [46] conv:(2.34)
 5. AAS=low DOI=Yes 81 ==> ArXivID=No PMID=No 77    conf:(0.95) < lift:(2.62)> lev:(0.16) [47] conv:(10.31)
 6. ArXivID=No PMID=No 109 ==> AAS=low DOI=Yes 77    conf:(0.71) < lift:(2.62)> lev:(0.16) [47] conv:(2.41)
 7. AAS=low 87 ==> JournalISSNs=Yes PMID=No 75    conf:(0.86) < lift:(2.59)> lev:(0.15) [46] conv:(4.46)
 8. JournalISSNs=Yes PMID=No 100 ==> AAS=low 75    conf:(0.75) < lift:(2.59)> lev:(0.15) [46] conv:(2.73)
 9. AAS=low 87 ==> DOI=Yes JournalISSNs=Yes PMID=No 75    conf:(0.86) < lift:(2.59)> lev:(0.15) [46] conv:(4.46)
10. DOI=Yes JournalISSNs=Yes PMID=No 100 ==> AAS=low 75    conf:(0.75) < lift:(2.59)> lev:(0.15) [46] conv:(2.73)
11. AAS=low ArXivID=No 83 ==> DOI=Yes PMID=No 77    conf:(0.93) < lift:(2.58)> lev:(0.16) [47] conv:(7.59)
12. DOI=Yes PMID=No 108 ==> AAS=low ArXivID=No 77    conf:(0.71) < lift:(2.58)> lev:(0.16) [47] conv:(2.44)
13. AAS=low 87 ==> ArXivID=No DOI=Yes PMID=No 77    conf:(0.89) < lift:(2.53)> lev:(0.16) [46] conv:(5.14)
14. ArXivID=No DOI=Yes PMID=No 105 ==> AAS=low 77    conf:(0.73) < lift:(2.53)> lev:(0.16) [46] conv:(2.57)
15. AAS=low 87 ==> ArXivID=No PMID=No 79    conf:(0.91) < lift:(2.5)> lev:(0.16) [47] conv:(6.15)
16. ArXivID=No PMID=No 109 ==> AAS=low 79    conf:(0.72) < lift:(2.5)> lev:(0.16) [47] conv:(2.5)
17. AAS=low 87 ==> DOI=Yes PMID=No 78    conf:(0.9) < lift:(2.49)> lev:(0.16) [46] conv:(5.57)
18. DOI=Yes PMID=No 108 ==> AAS=low 78    conf:(0.72) < lift:(2.49)> lev:(0.16) [46] conv:(2.47)
19. PMID=No 120 ==> AAS=low DOI=Yes 78    conf:(0.65) < lift:(2.41)> lev:(0.15) [45] conv:(2.04)
20. AAS=low DOI=Yes 81 ==> PMID=No 78    conf:(0.96) < lift:(2.41)> lev:(0.15) [45] conv:(12.15)
21. PMID=No 120 ==> AAS=low ArXivID=No DOI=Yes 77    conf:(0.64) < lift:(2.41)> lev:(0.15) [44] conv:(2)
22. AAS=low ArXivID=No DOI=Yes 80 ==> PMID=No 77    conf:(0.96) < lift:(2.41)> lev:(0.15) [44] conv:(12)
23. AAS=low DOI=Yes JournalISSNs=Yes 78 ==> PMID=No 75    conf:(0.96) < lift:(2.4)> lev:(0.15) [43] conv:(11.7)
24. PMID=No 120 ==> AAS=low DOI=Yes JournalISSNs=Yes 75    conf:(0.63) < lift:(2.4)> lev:(0.15) [43] conv:(1.93)
25. AAS=low 87 ==> PMID=No 83    conf:(0.95) < lift:(2.39)> lev:(0.16) [48] conv:(10.44)
26. PMID=No 120 ==> AAS=low 83    conf:(0.69) < lift:(2.39)> lev:(0.16) [48] conv:(2.24)
27. AAS=low ArXivID=No 83 ==> PMID=No 79    conf:(0.95) < lift:(2.38)> lev:(0.15) [45] conv:(9.96)
28. PMID=No 120 ==> AAS=low ArXivID=No 79    conf:(0.66) < lift:(2.38)> lev:(0.15) [45] conv:(2.07)
29. PMID=No 120 ==> AAS=low JournalISSNs=Yes 75    conf:(0.63) < lift:(2.37)> lev:(0.14) [43] conv:(1.92)
30. AAS=low JournalISSNs=Yes 79 ==> PMID=No 75    conf:(0.95) < lift:(2.37)> lev:(0.14) [43] conv:(9.48)
31. JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No DOI=Yes PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
32. ArXivID=No DOI=Yes PMID=Yes 175 ==> JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
33. ArXivID=No JournalISSNs=Yes OA=Yes 140 ==> DOI=Yes PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.31)
34. DOI=Yes PMID=Yes 179 ==> ArXivID=No JournalISSNs=Yes OA=Yes 98    conf:(0.55) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
35. JournalISSNs=Yes OA=Yes 143 ==> DOI=Yes PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.31)
36. DOI=Yes PMID=Yes 179 ==> JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
37. JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.28)
38. ArXivID=No PMID=Yes 176 ==> JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
39. DOI=Yes JournalISSNs=Yes OA=Yes 143 ==> ArXivID=No PMID=Yes 98    conf:(0.69) < lift:(1.17)> lev:(0.05) [14] conv:(1.28)
40. ArXivID=No PMID=Yes 176 ==> DOI=Yes JournalISSNs=Yes OA=Yes 98    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.17)
41. ArXivID=No JournalISSNs=Yes OA=Yes 140 ==> PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
42. ArXivID=No DOI=Yes JournalISSNs=Yes OA=Yes 140 ==> PMID=Yes 98    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
43. PMID=Yes 180 ==> ArXivID=No JournalISSNs=Yes OA=Yes 98    conf:(0.54) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
44. PMID=Yes 180 ==> ArXivID=No DOI=Yes JournalISSNs=Yes OA=Yes 98    conf:(0.54) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
45. JournalISSNs=Yes OA=Yes 143 ==> PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
46. PMID=Yes 180 ==> JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
47. DOI=Yes JournalISSNs=Yes OA=Yes 143 ==> PMID=Yes 100    conf:(0.7) < lift:(1.17)> lev:(0.05) [14] conv:(1.3)
48. PMID=Yes 180 ==> DOI=Yes JournalISSNs=Yes OA=Yes 100    conf:(0.56) < lift:(1.17)> lev:(0.05) [14] conv:(1.16)
49. DOI=Yes OA=Yes 149 ==> ArXivID=No PMID=Yes 98    conf:(0.66) < lift:(1.12)> lev:(0.04) [10] conv:(1.18)
50. DOI=Yes OA=Yes 149 ==> ArXivID=No JournalISSNs=Yes PMID=Yes 98    conf:(0.66) < lift:(1.12)> lev:(0.04) [10] conv:(1.18)
```

- Reperforming analysis removing IDs, only 10 rules could be produced.



```
Best rules found:

 1. Category=Medical & Health Sciences 98 ==> AAS=medium2high 43    conf:(0.44) < lift:(1.71)> lev:(0.06) [17] conv:(1.3)
 2. AAS=medium2high 77 ==> Category=Medical & Health Sciences 43    conf:(0.56) < lift:(1.71)> lev:(0.06) [17] conv:(1.48)
 3. Category=Medical & Health Sciences 98 ==> AAS=medium 41    conf:(0.42) < lift:(1.51)> lev:(0.05) [13] conv:(1.22)
 4. AAS=medium 83 ==> Category=Medical & Health Sciences 41    conf:(0.49) < lift:(1.51)> lev:(0.05) [13] conv:(1.3)
 5. OA=No 140 ==> AAS=low 52    conf:(0.37) < lift:(1.28)> lev:(0.04) [11] conv:(1.12)
 6. AAS=low 87 ==> OA=No 52    conf:(0.6) < lift:(1.28)> lev:(0.04) [11] conv:(1.29)
 7. AAS=medium2high 77 ==> OA=Yes 48    conf:(0.62) < lift:(1.17)> lev:(0.02) [6] conv:(1.2)
 8. OA=Yes 160 ==> AAS=medium2high 48    conf:(0.3) < lift:(1.17)> lev:(0.02) [6] conv:(1.05)
 9. Category=Medical & Health Sciences 98 ==> OA=Yes 59    conf:(0.6) < lift:(1.13)> lev:(0.02) [6] conv:(1.14)
10. OA=Yes 160 ==> Category=Medical & Health Sciences 59    conf:(0.37) < lift:(1.13)> lev:(0.02) [6] conv:(1.06)
```

Key Take Away

- Given the goal of this analysis was to find associations of features alongside altmetric score, the top rules did not provide much value. For instance, while it can be seen from rule 4 that there are associations with AAS except that the features associated are not every informative and there was only the result of AAS=low within the top 50 rules.
- This analysis has provided no useful information. Or any that we can make useful sense of as it had only shown the most common associations of feature factors that had occurred together the most.
- Reperforming the analysis removing ID features which contributed mostly to this issue yield more useful results such as:
  - Medical & Health Sciences is associated with medium2high and medium AAS
  - Non open access publications are associated with low AAS
  - Open access publications are associated with medium2high AAS
- Thus, it is recommended that medical & health sciences publications should be published if an institution is aiming solely for higher AAS responses. While in general they should aim to make publications open access which is found to be associated with higher AAS than non-open access publications.

Decision Tree

Technique information

- A decision tree is a classifier that uses both categorical and numerical variables; and produces a nodded tree (J48).

Goal

- As seen from association, there were some rules with AAS. But it doesn't give the full picture so a decision tree may help.

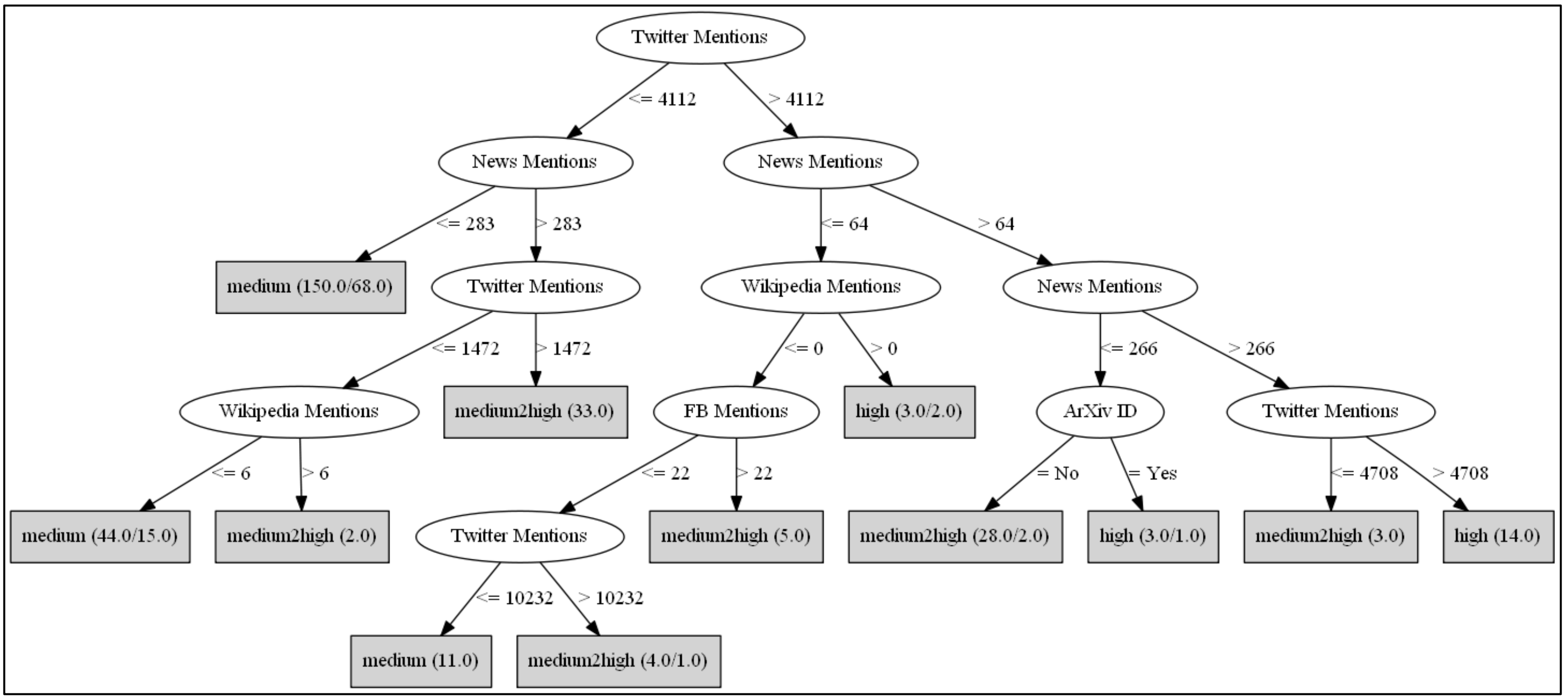
Method and Considerations

- Data is 2017, 2018 and 2019 data
- For pruning the tree, a confidence factor of 0.25 will first be set. Through testing it was found that while a high confidence factor over 40% and up to 45% resulted in more features to evaluate AAS, model accuracy also dropped to near 50%.

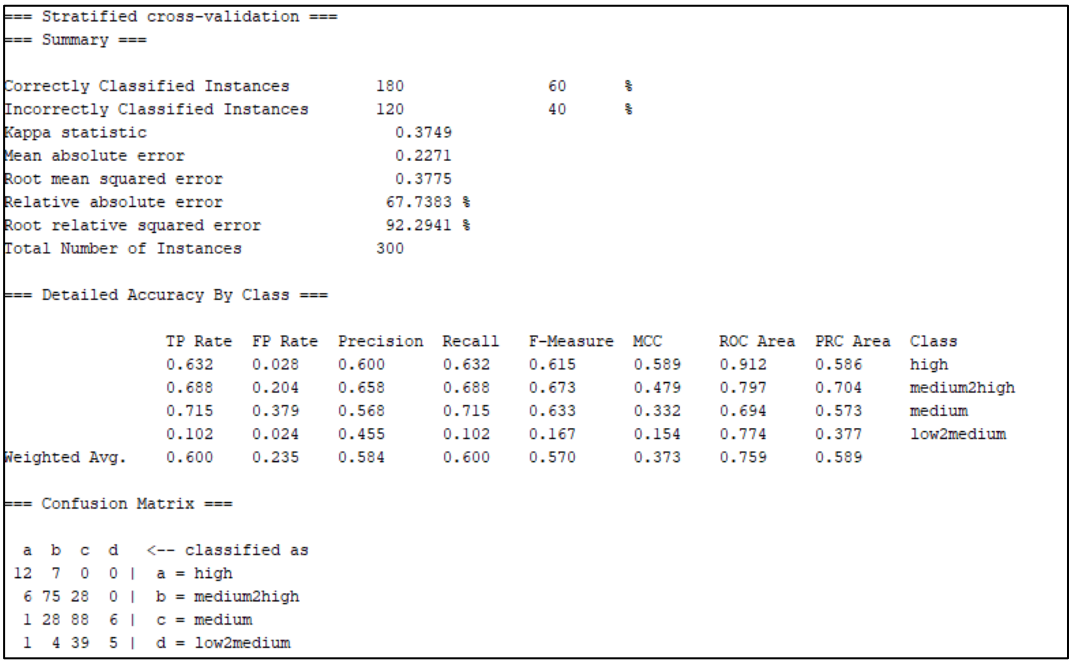
- Cross validation is used to find best model. Other methods of model testing could be used if more data was available.

Results

- The following tree has accuracy of 60% and was the . Altmetric data sources seem to have highest predictive accuracy for AAS.



- Below are the Weka output and ROC curves for the predicted factor levels of AAS



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         180               60      %
Incorrectly Classified Instances       120               40      %
Kappa statistic                          0.3749
Mean absolute error                      0.2271
Root mean squared error                  0.3775
Relative absolute error                 67.7383 %
Root relative squared error             92.2941 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.632    0.028    0.600      0.632    0.615      0.589    0.912     0.586     high
                 0.688    0.204    0.658      0.688    0.673      0.479    0.797     0.704     medium2high
                 0.715    0.379    0.568      0.715    0.633      0.332    0.694     0.573     medium
                 0.102    0.024    0.455      0.102    0.167      0.154    0.774     0.377     low2medium
Weighted Avg.    0.600    0.235    0.584      0.600    0.570      0.373    0.759     0.589

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 12  7  0  0 |  a = high
  6 75 28  0 |  b = medium2high
  1 28 88  6 |  c = medium
  1  4 39  5 |  d = low2medium
```

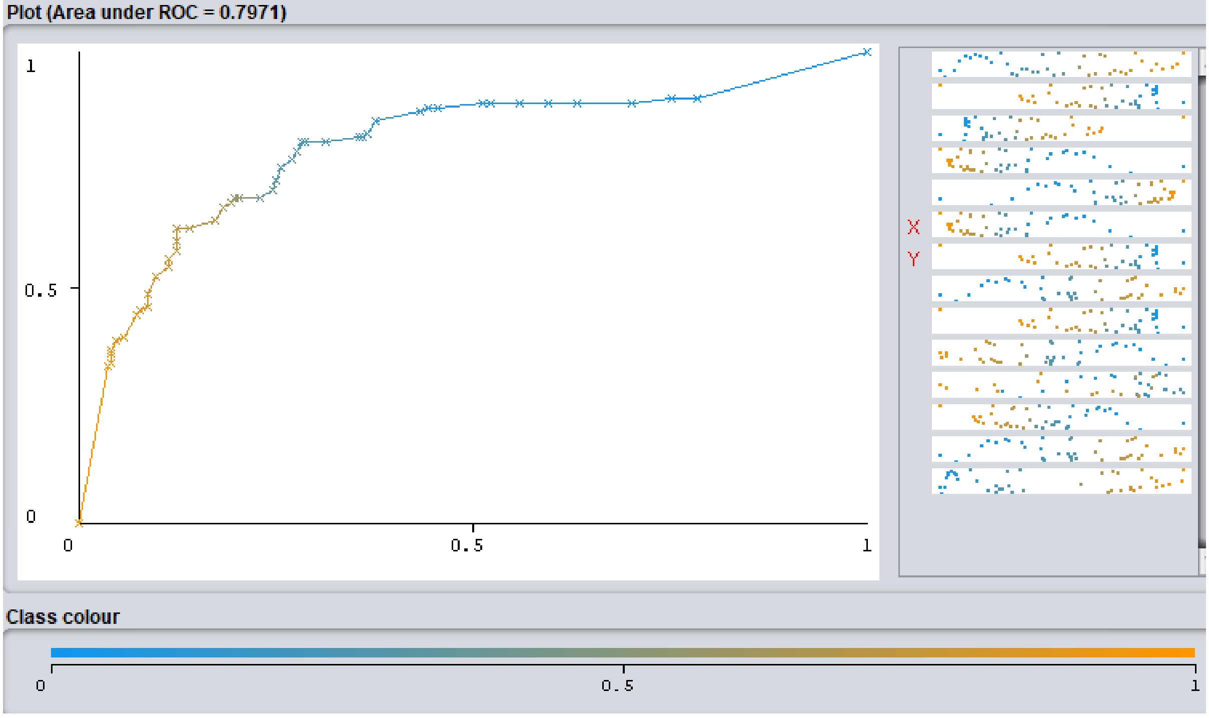


Figure 29med2high

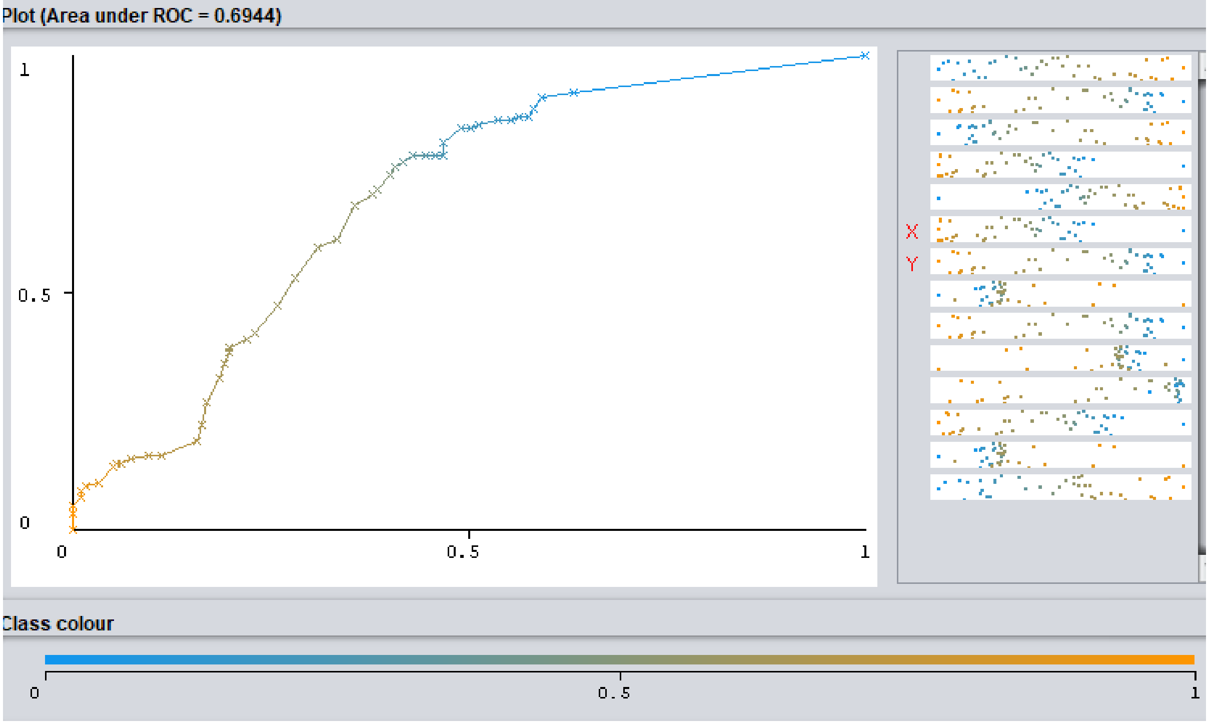


Figure 30medium

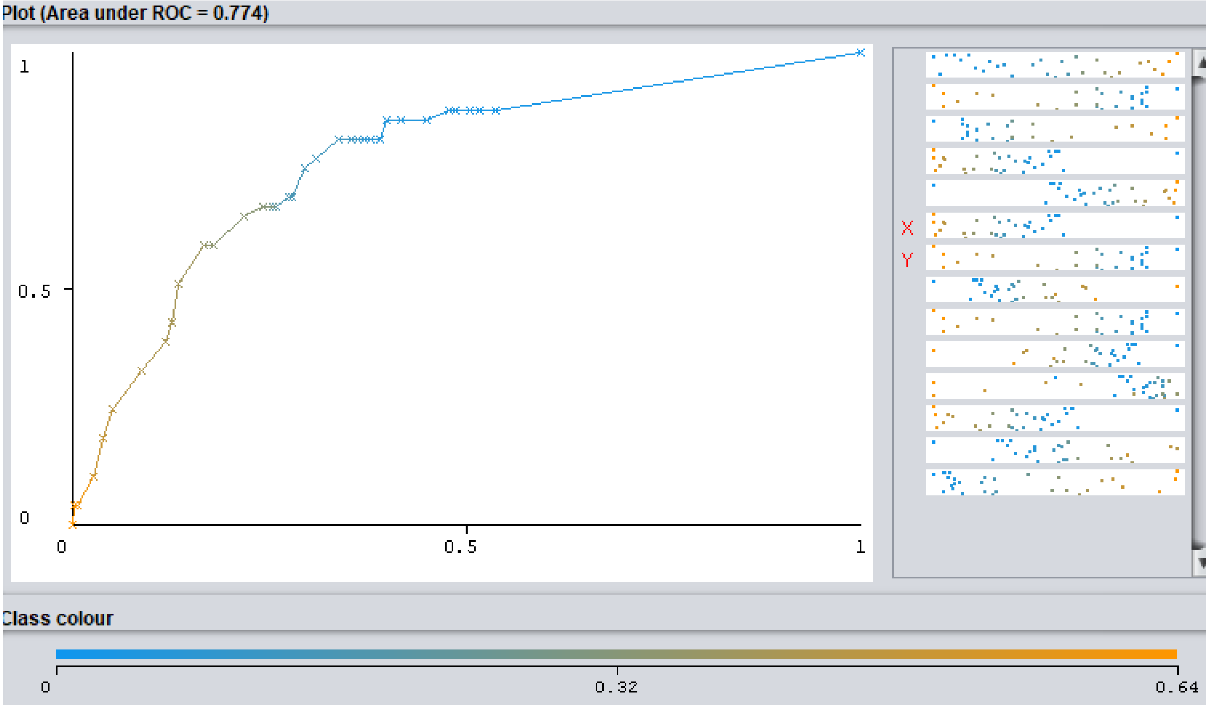


Figure 31low2med

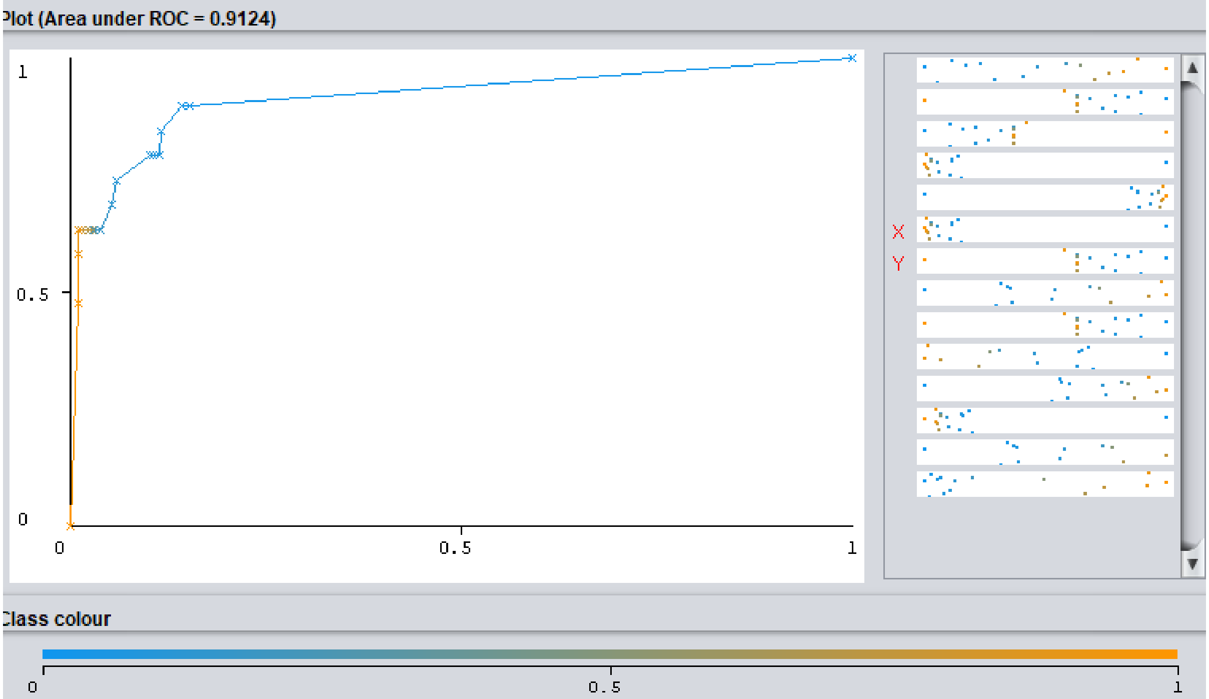


Figure 32high

Key Take Away

- In general, a high data source count is still found to be associated with high AAS, however it seems that Twitter, news, wikipedia and FB data sources seems to best predict AAS. And ArXiv is a good indicator out of all the IDs for AAS.
- From ROC graphs, overall, predictions are good and that it is very good for predicting high AAS, so more attention should be given in that area of the decision tree.

- From the decision tree and ROC values, we can expect factors leading up the consequences of high and medium2high AAS to be a good true positive predictor of AAS. The tree shows these factors to be the altmetric data sources of, Twitter, News and Wikipedia as well as the ArXiv ID. Below lists the decision rules for achieving high and medium2high AAS:
    - <u>High</u>
        - 4708 < Twitter & 266 < News
        - 4112 < Twitter & 64 < News <= 226 & ArXiv = Yes
        - 4112 < Twitter & News <= 64 & 0 < Wikipedia

    - <u>Medium2High</u>
        - 4708 < Twitter <= 4708 & 266 < News
        - 4112 < Twitter & 64 < News <= 226 & ArXiv = No
        - 4112 < Twitter & News <= 64 & Wikipedia <= 0 & 22 < FB
        - 10232 < Twitter & News <= 64 & Wikipedia <= 0 & FB <= 22
        - 1472 < Twitter <= 4112 & 283 < News
        - Twitter <= 1472 & 283 < News & 6 < Wikipedia
- In short, these results show that institutions should in general promote their publication on Twitter as much as possible to achieve at least 4708 mentions. The second media to focus promotion on is the news, where they should aim to achieve over 266 mentions of their publication. If that is not possible, they should also promote on Wikipedia and it may be worthwhile to also make the publication available on ArXiv.

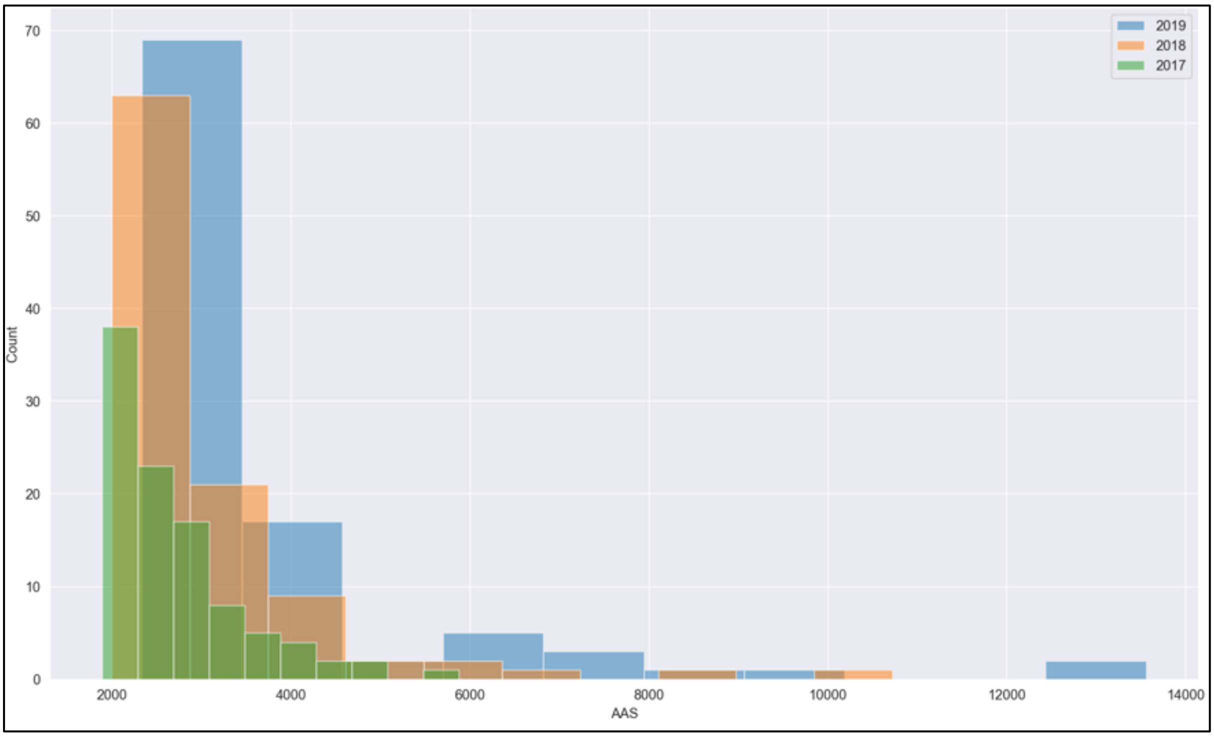
<u>Multiple Linear Regression</u>

Technique information

- To Model the relationship between more than one independent variable and a dependent variable. Where independent variables are expected to have a linear relationship with the dependent variable.

Goal

- From decision tree, it was possible to obtain the most important data source features to predict AAS. However, we are limited by the categorical nature of the AAS prediction. Using regression those data sources will be used to produce a model that can predict AAS to a numerical value.
- Another benefit of regression is that it may be used to also better predict the future trend of AAS, as had been seen in the data that AAS in the top 100 publications has been increasing through the years. For the data used in the analysis this is also evident through a histogram:

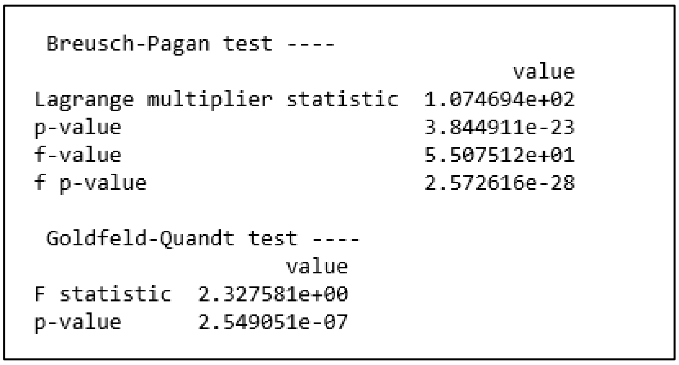


Method and Considerations

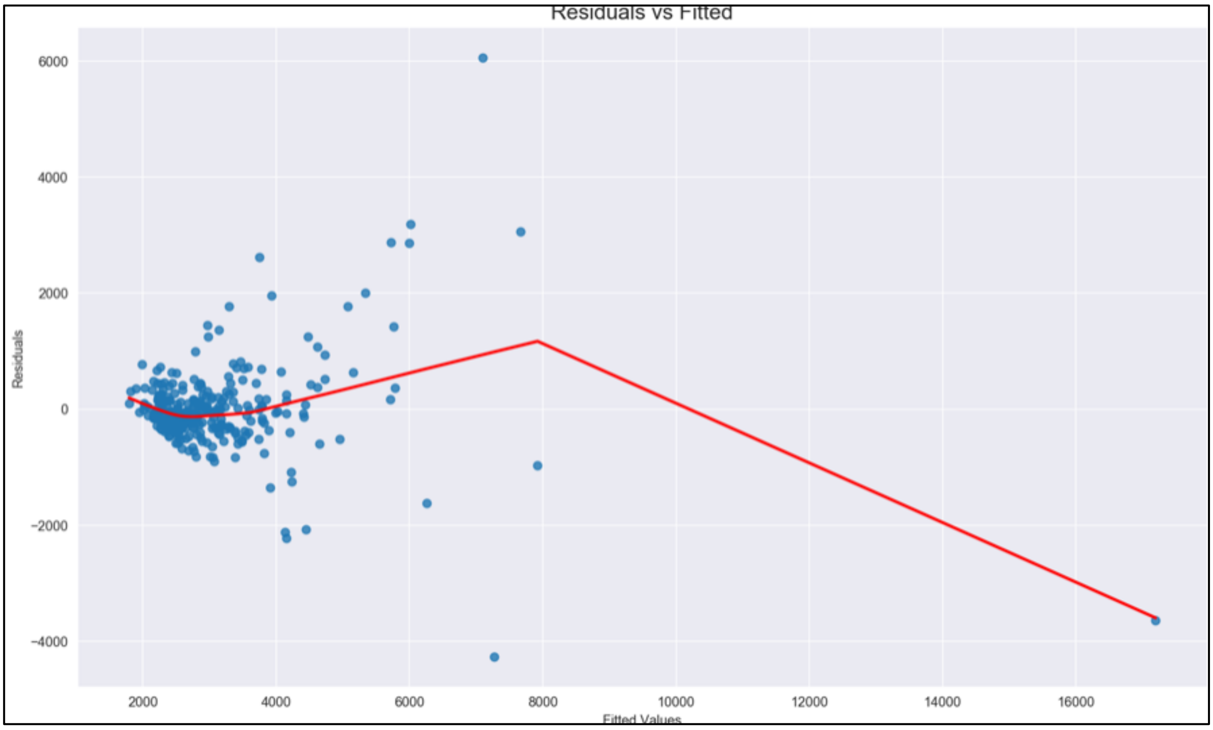
- Data source is 2017 to 2019 without categorical features
- The key model assumptions will be considered: multicollinearity, normally distributed residuals and homoscedasticity.
- 5% significance level and R^2 will be used to evaluate the model
- One model was first made with all altmetric data sources available and gotten R^2 =0.757 with 5 features after removing significant features. Another model was made with only the predictive features from decision tree and gotten R^2 =0.712 but is simpler with 3 features. The simpler model will be used as it also includes only the features from the decision tree; without wikipedia mentions.

Results

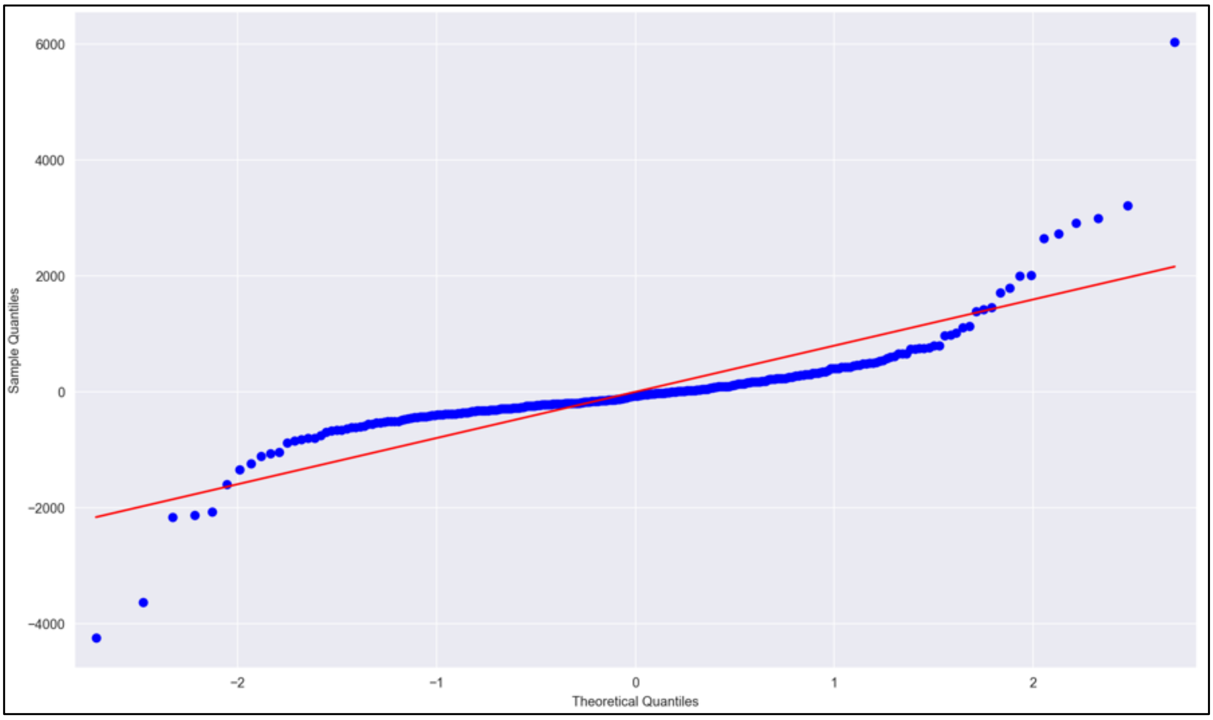
- The model found does not satisfy the homoscedasticity assumption, failing Breusch-Pagan and Goldfeld-Quandt test. However, it was reasoned that this was the cause of insufficient data and variance associated with the change in AAS in 2018 and 2019 compared to 2017 as seen in residuals vs fitted graph.



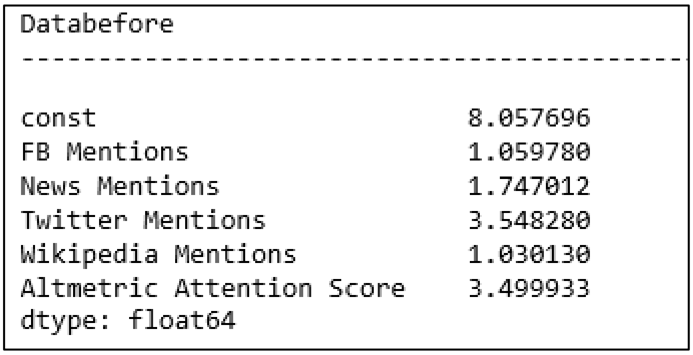
```
 Breusch-Pagan test ----
                                     value
Lagrange multiplier statistic  1.074694e+02
p-value                        3.844911e-23
f-value                        5.507512e+01
f p-value                      2.572616e-28

 Goldfeld-Quandt test ----
                  value
F statistic  2.327581e+00
p-value      2.549051e-07
```



- Linear relationship between independent and dependent variables is not clearly seen in scatter plots, however in theory a linear relationship exists as each AAS is the sum of data source mentions multiplied by a scalar as weighting; a linear combination.
- Residuals are approximately normally distributed



- Shows no multicollinearity using variance inflation factor



```
Databefore
------------------------------------------

const                          8.057696
FB Mentions                    1.059780
News Mentions                  1.747012
Twitter Mentions               3.548280
Wikipedia Mentions             1.030130
Altmetric Attention Score      3.499933
dtype: float64
```

- Wikipedia mentions was found to not be significant and the model was reduced to 3 features with the following equation: AAS=4.2(FB Mentions)+5.6(News Mentions)+0.26(Twitter Mentions)+799.4 to two decimal places

Key Take Away

- AAS=4.2(FB Mentions)+5.6(News Mentions)+0.26(Twitter Mentions)+799.4
- The regression equation found supports the results found in the decision tree and we can see that we may only need three features to predict AAS. These results are quite contradictory to the altmetric weightings for news that are 8, twitter that are 1 and facebook with 0.25. However, as discussed in assignment 1, there are also additional data source weightings applied for specific news source, twitter and facebook credibility and so on that are not reflected in our data sources directly.
- Thus, while the weightings in the equation may not directly mean the importance of a feature. There may be an indirect relationship between the data source features that was not seen which makes face book mention a significant variable in the model. So, it may also be worthwhile for institutions to also focus on promoting publications on facebook as well in addition to twitter and news so that they could have a means to predict their potential AAS to evaluate progress.

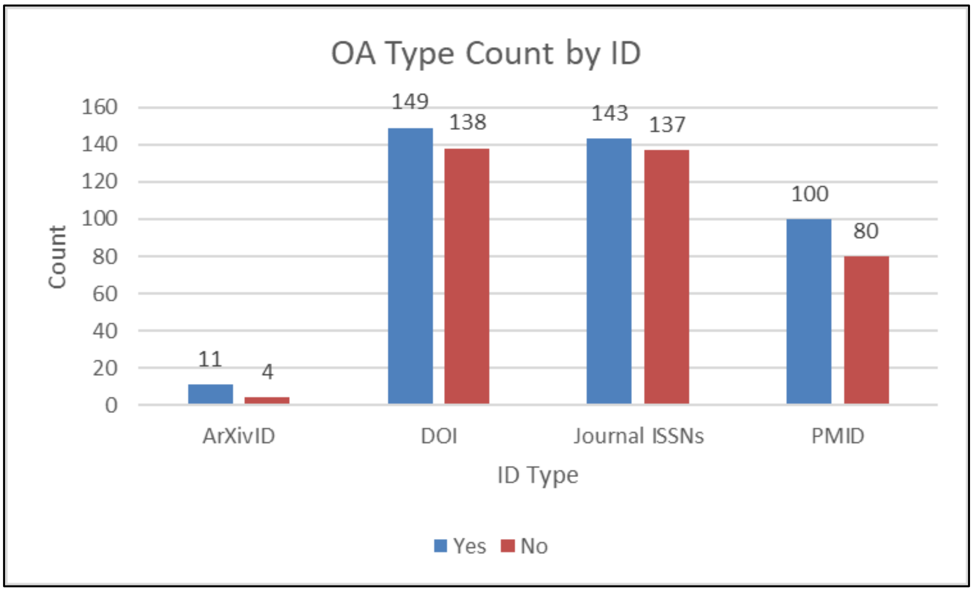
IDs and OA

From association it was found that open access publications are association with AAS in the range medium2high while non open access are associated worth low AAS. This prompts the following analysis which investigates if having certain IDs would imply OA status.

There are two approaches used. The first examines the OA type of a publication specially based on each ID. This is to see if each ID by itself has some relation with OA.

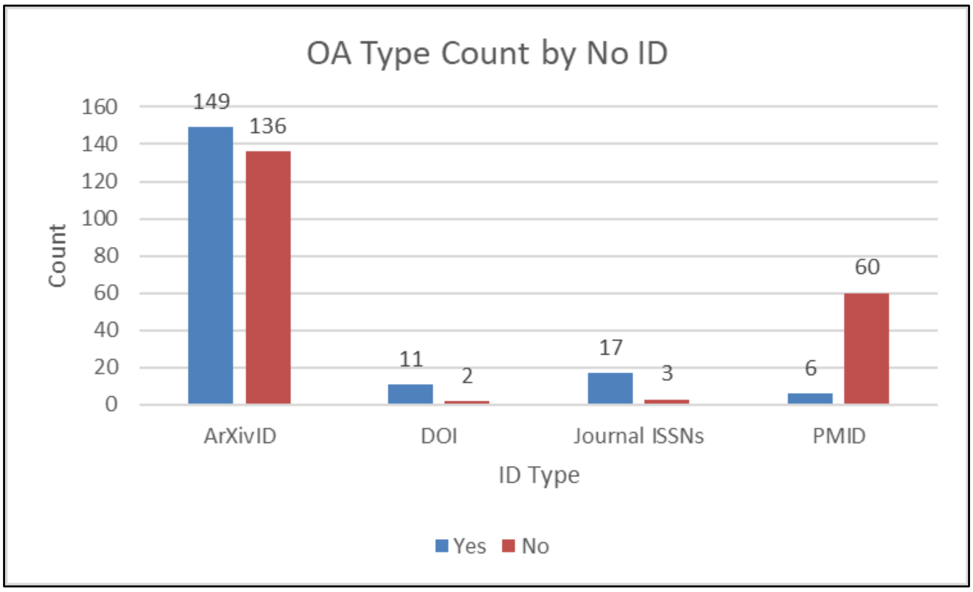
YES ID

OA	ArXivID	DOI	Journal ISSNs	PMID
YES	11	149	143	100
NO	4	138	137	80



NO ID

OA	ArXivID	DOI	Journal ISSNs	PMID
YES	149	11	17	6
NO	136	2	3	60



The tables and graph show that across all ID types having a specific ID or not is almost equally likely for a publication to be open access or not.

The second approach considers each ID type combination within the data set. For each combination we examine if it is open access or not. The following tables below shows this result and we can see that the top 2 combinations that exists in both tables are the same and that they include most of the publications. Thus, we have reason to believe that IDs are not likely to be associated with publication OA type.

So, while institutions may not directly see the benefits of having a publication within a certain data base to be open access; which have been found to positively affect AAS. There may be underlaying factors that they should focus on to make their publication open accessed and more popular within these data bases to achieve higher AAS.

OA = Yes	*ArXiv ID*	*DOI*	*Journal ISSNs*	*PMID*
98	No	Yes	Yes	Yes
42	No	Yes	Yes	No
8	Yes	No	No	No
6	No	Yes	No	No
3	No	No	No	No
2	Yes	Yes	Yes	Yes
1	Yes	Yes	Yes	No

OA = No	*ArXiv ID*	*DOI*	*Journal ISSNs*	*PMID*
77	No	Yes	Yes	Yes
55	No	Yes	Yes	No
2	No	Yes	No	No
2	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	No
1	No	No	Yes	Yes
1	No	No	No	No

OA = Yes	*OA = No*	*ArXiv ID*	*DOI*	*Journal ISSNs*	*PMID*
98	77	No	Yes	Yes	Yes
42	55	No	Yes	Yes	No
8	0	Yes	No	No	No
6	2	No	Yes	No	No
3	1	No	No	No	No
2	2	Yes	Yes	Yes	Yes
1	2	Yes	Yes	Yes	No
0	1	No	No	Yes	Yes
(160)	(140)		SUMS		

Limitations

It is important to address that the data used in this report has mainly been a subset of all available data and that the features considered for analysis may not be enough to fully explain results as not all features available are used.

Another issue to address is on the sampling of data used in the report. As the data used is the top 100 publications by AAS it is difficult to clearly examine differences between publication features that have a positive effect on AAS. It is reasoned that publications that have achieved to be in the top 100 would have very similar characteristics where the differentiation of any of them having a higher AAS could be due to chance or effect of additional features not available or considered in the data set.

Thus, for future research it is recommended that more data be sourced specifically, top 200 or more publications per year. In addition to the examination of additional features that may also influence AAS.

9. Appendix 4: Stephen Maher

Findings: Bornmann & Haunschild

Two analytical techniques were selected for the analysis of the Bornmann & Haunschild dataset. These were:

- Clustering: K-means
- Factor Analysis and Principal Components Analysis

The following variables were available for analysis within the dataset:

Discrete (quantitative)
(pubyear) (wos_cits_3) (sco_cits_3) (tweets) (me_readers) (total_f1000_score)
Continuous (quantitative)
(altmetric_score) (citescore) (item_ijif)

The attribute “pubyear” was dropped early in the analytical process as the results showed no differentiation on this attribute.

Cluster Analysis

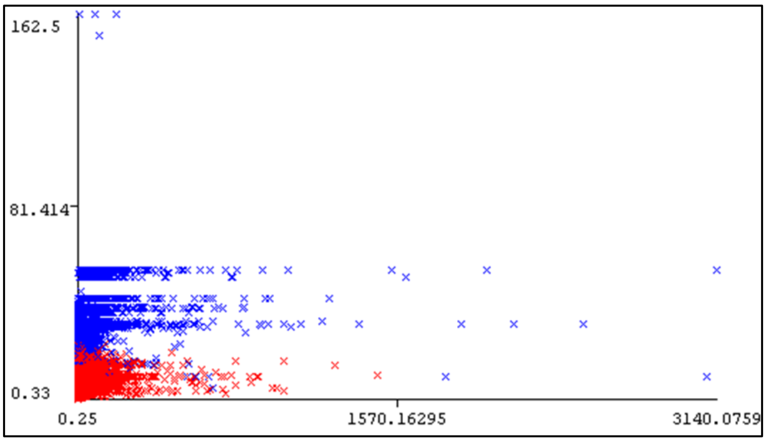
The cluster analysis showed some differentiation within the dataset. Two cluster (using K-means) were identified with one cluster containing 87% of the rows in the dataset and the other containing 13%. The key differentiation between the two clusters was the mean value of the attributes within each cluster (table below). As can be seen in this table, there is a high mean attribute value cluster (cluster 0) and low mean attribute cluster (cluster 1).

The relative distribution is evident in the following two figures but is not as clear in the third figure, which shows a relatively tight correlation between “tweets” and “altmetric_score” with both clusters overlain on each other. Cluster 0 (blue) appears to show some association between higher levels of citation and other measures of quality and social media attention as measured by tweets and the altmetric score. It is worth noting however, that reverse is less evident. That is, a relatively high tweet or altmetric score may not necessarily indicative of higher levels of citation or other quality measures.

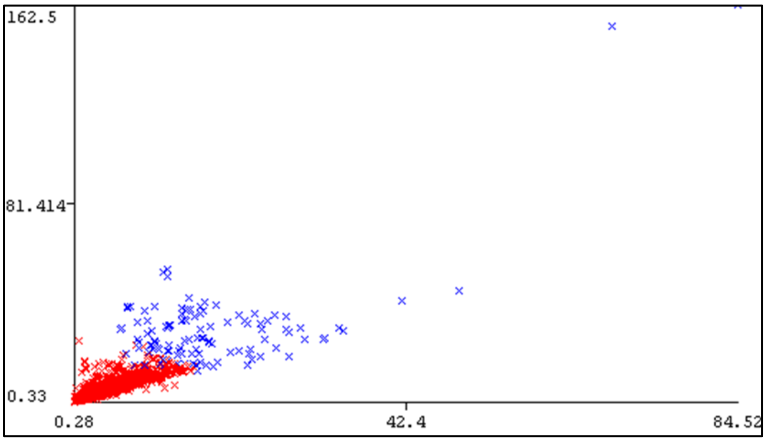
Bornmann & Haunschild Dataset: K-means cluster analysis, mean attribute value by cluster

		CLUSTER 0	CLUSTER 1	
ROW COUNT	33683	5593	28090	Difference
WOS_CITS_3	29.8576	81.7348	19.5283	62.2065
SCO_CITS_3	31.0315	83.1534	20.6535	62.4999
TWEETS	10.1401	33.8874	5.4118	28.4756
ME_READERS	76.4402	212.6832	49.3129	163.3703
ALTMETRIC_SCORE	17.122	55.1847	9.5433	45.6414
CITESCORE	7.2166	14.6243	5.7417	8.8826
ITEM_IJIF	11.0997	32.0108	6.9361	25.0747
TOTAL_F1000_SCORE	2.0417	3.8731	1.677	2.1961

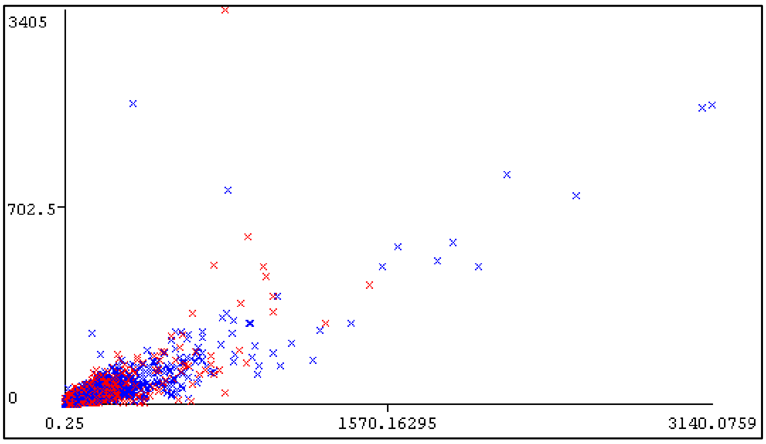
Bornmann & Haunschild Dataset: clusters identified on item_ijif (y-axis) and altmetric_score (x-axis) [cluster 0 = blue, cluster 1 = red]



Bornmann & Haunschild Dataset: clusters identified on item_ijif (y-axis) and citescore (x-axis) [cluster 0 = blue, cluster 1 = red]



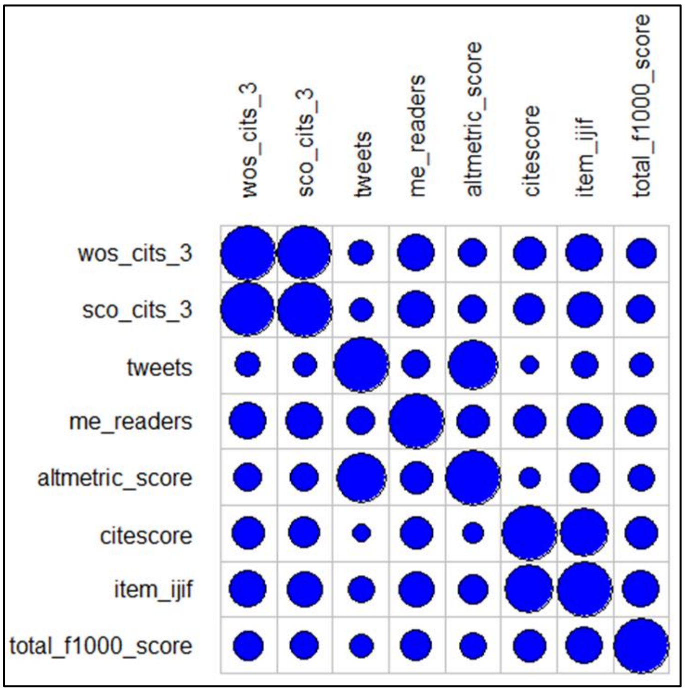
Bornmann & Haunschild Dataset: clusters identified on tweets (y-axis) and altmetric_score (x-axis) [cluster 0 = blue, cluster 1 = red]



Factor Analysis and Principal Components Analysis

The second analysis on the Bornmann and Haunschild data comprised principal components analysis "PCA" and factor analysis "FA". The initial step was to review the correlogram of the data (and excluding pubyear). The correlogram is shown in the figure below and features three primary sets of correlations: 1) wos_cits_3 and sco_cits_3, 2) tweets and altmetric_score, and 3) citescore and item_ijif. This suggests that we may see at least three principal components in the PCA.

Correlogram of Bornman & Haunschild Data

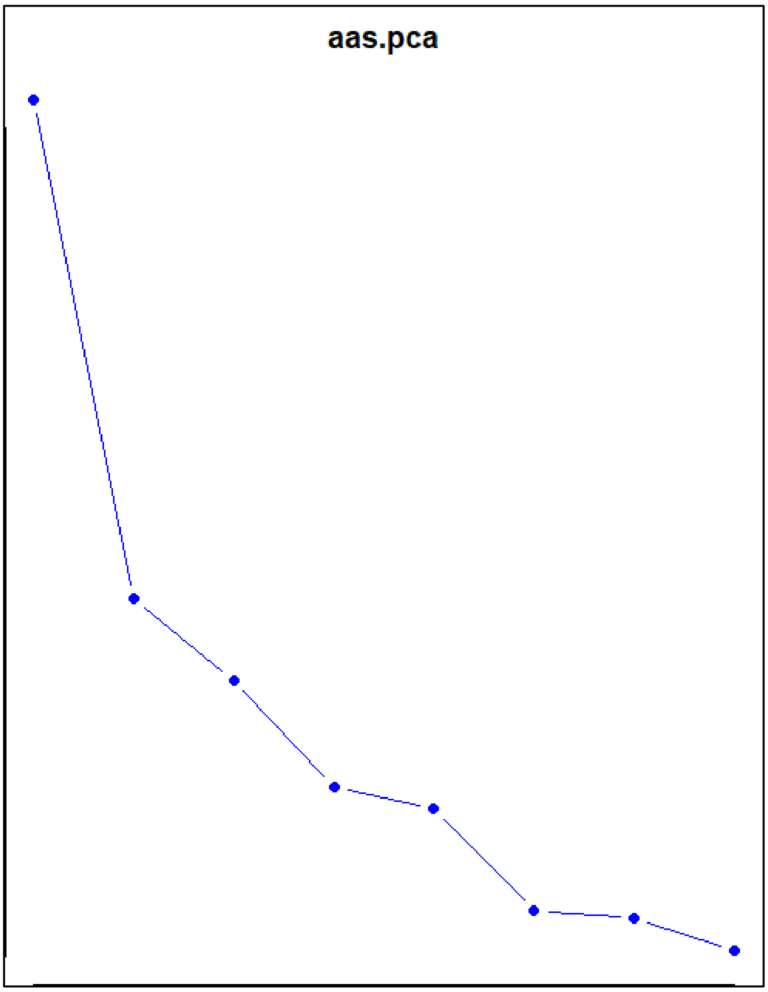


Extending the analysis to PCA (after applying standardisation), the relative importance of components is shown in the following table. Using Kaiser's Criterion, the first three factors are selected as suitable and this was confirmed by a scree plot of the data. The table below shows the resultant loadings due to these three components. From this table, it is evident that PC2 loads heavily on tweets and altmetric_score. PC3 loads on all factors excluding tweets, altmetric_score and me_readers. Notably for PC3, wos_cits_3 and sco_cits_3 have a negative sign while the remaining factors have a positive loading. While PC2 and PC3 show relatively clear loadings, PC1 loads on all factors.

Bornman & Haunschild Data: Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.9013	1.2278	1.0792	0.84435	0.78863	0.44073	0.403	0.14606
Proportion of Variance	0.4519	0.1885	0.1456	0.08912	0.07774	0.02428	0.0203	0.00267
Cumulative Proportion	0.4519	0.6403	0.7859	0.87501	0.95275	0.97703	0.9973	1

Scree plot of principal components



PCA Analysis: Component Loadings

	PC1	PC2	PC3
wos_cits_3	-0.41443	-0.19297	-0.4995
sco_cits_3	-0.40717	-0.19188	-0.52022
tweets	-0.27094	0.649947	0.035927
me_readers	-0.35547	-0.01647	-0.03661
altmetric_score	-0.31123	0.608903	0.028927
citescore	-0.34356	-0.29702	0.46509
item_ijif	-0.40432	-0.2004	0.393595
total_f1000_score	-0.29037	-0.064	0.324317

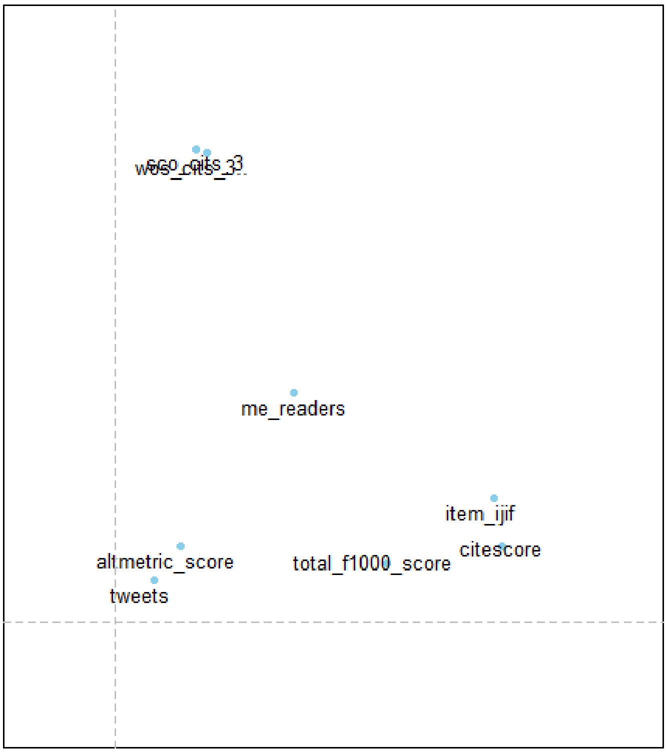
In addition to PCA, exploratory factor analysis was also conducted over the dataset (sample size is nominally excellent for this analysis). Preliminary statistics suggested that the use of EFA would be marginal effective at best. There was evidence of multicollinearity and a KMO test indicated that the common variance was unacceptable for factor analysis.

Bornmann and Haunschild dataset: EFA using Varimax Rotation

	item	RC1	RC3	RC2	h2	u2	com
citescore	6	0.89			0.81	0.188	1.1
item_ijif	7	0.87			0.83	0.168	1.2
total_f1000_score	8	0.62			0.43	0.567	1.2
sco_cits_3	2		0.96		0.97	0.03	1.1
wos_cits_3	1		0.96		0.97	0.032	1.1
me_readers	4	0.41	0.47		0.46	0.541	2.6
tweets	3			0.94	0.9	0.096	1
altmetric_score	5			0.93	0.91	0.09	1.1

Notwithstanding concerns with respect to the application of EFA to the Bornmann and Haunschild dataset, EFA was performed under varimax rotation. The results are shown in the following table and the factor plot is shown after the table. Three factors were selected for the analysis and the cumulative variance explained was 79% and communalities are generally high (excluding total_f1000_score and me_readers). The factors for the dataset are clearly aligned along the lines of 1) number of citations, 2) journal impact factor, and 3) social media metrics.

Bornmann and Haunschild EFA Factor Plot under varimax rotation


sco_cits_3
wos_cits_3..
me_readers
item_ijif
altmetric_score
total_f1000_score
citescore
tweets