Quantifying and Analyzing the Influence of Energy Researchers in Articles

University of Ottawa

Faculty of Social Science

Department of Economics

Instructor: Benjamin Poirier, Jonas Becker

Student Name: Wenxi Ma

Student number 300051443

Start Date: September 21st, 2021

Submission Date: December 21st,2021

Outline:

INTRODUCTION

1. Summary:

Canada is acting to address climate change. Even though our geographical location is close to the arctic, warming faster than other parts of the planet, articles and research by Canadian energy researchers will influence the strategies of Canadian policymakers implementing the energy transition. These policies will eventually affect the everyday habits of work and life in Canada. Therefore, Canadian energy researchers must influence their peers. Policymakers collect information or opinions from scholarly articles that influence energy research. This assignment is to understand the impact of Canadian scholarship and research. This work will be submitted in the form of a final report.

1.1How to consider the influence of researchers?

Our first task is to quantify the impact of researchers and make their impact deterministic (like an algorithm or a formula). In this way, we can apply this method to any researcher. First, the main task of researchers is to do research, publish articles, and share the results of their analyses. To understand the breadth of their impact and quantify it, we will start by dividing the categories of influence on researchers. Secondly, we divide the power of researchers into articles into three categories: 1. The number of publications 2. The number of clicks 3. Article citation rate. In addition, researchers who attend to the quantities of projects will consider the influence. These three categories as consideration standards are divided into different proportions to evaluate the impact of these researchers. Finally, we will integrate these data for data analysis. In the end, we will get a formula similar to linear regression. For example: if the other two variables remain the same, how much influence will change if the author publishes one more article, or how much his influence will change if his citation rate increases by one time.

2. Visualization of Framework

A framework diagram is always helpful when the research is still in the beginning. In figure 1, I divided this measuring problem into two main parts, which have a significant weight on the influence of Canadian researchers on literature. Publications are the most efficient and easiest way to spread their influence among peers.

2.1 How do the publications make an impact?

The primary purpose of the researchers’ work is to explore the problems, analyze them and give their conclusion based on their research. Therefore, publication has become the primary way to convey their ideas to the outside world and colleagues. The number of clicks indicates how interested others are in the article published by the researcher. The more people who read it proves that the researcher’s article is popular. The number of citations represents the research quality of this researcher. Although this researcher has published many articles and has many page views, his articles have been cited very few times. This can explain that the research results of this researcher have not been accepted and adopted by his colleagues. Therefore, if the researcher’s article has been viewed many times and the citation rate of his article is relatively high, it proves that his peers recognize the researcher’s scientific research results. Over the long term, his influence will continue to increase. In addition, the projects that researchers participate in can also prove their influence. In application practice, people invite more famous colleagues to improve this project’s efficiency and success rate.

MEASURING

3. Collect the different types of data

Based on the modeler and model users listed in the energy modeling initiative. I divide them into six categories based on their organization type: academia, consultant, government, industry, NGO, and utility. Based on everyone's basic information, I collected as much as possible the articles published by them, the number of clicks, and the number of citations. (Although some of the data is incomplete) The purpose of this is to compare and analyze the influence of different types of researchers in the next section. It will allow us to find out the possible reasons for the discrepancies more effectively. After collecting the data, I stored the data in a data set for cleaning and analysis.

3.1 Exclusion and re-organization of the variables

Because of the incompleteness of some data, we will remove some useless data to ensure the algorithm's effectiveness. If the number of citations of the researcher is missing, This data will be removed. If the number of times the researcher's participation in projects is missing, then this condition will be ignored. The influence of researchers will only be affected by the citation rate.

3.2 Measuring Tools

In this project, search for the term by searching the names of modeler and model user. The data sources mainly come from google scholar, research gate, OurWorldinData, and JSTOR. These websites mainly include the number of articles published, the number of clicks, and the citation amount. This is the primary variable for our analysis of its influence. However, among these search engines, google scholar does not have several clicks sources. Classify the data into two types based on their origins. If the modeler/model user cannot be found or the data is not enough to be a complete piece of data, it is deemed to have no influence. And collect all the incomplete data together and classify them to analyze which types of institutions have low impact. In figure 2, Daniel Schrag is a top researcher in the environmental field[1]. He has 365 articles published and 30662 numbers cited by others. Based on these data, a researcher's influence is closely related to their number of citations and publications. Figure 3 shows the numbers of publications, clicks,, and citations of Eric Miller[2], who is highly influential in the transportation field from researchgate. In addition to the number of articles published, reading is also an important variable that affects the citation rate. After the researcher's article has been widely read, other researchers may cite it. It will help to increase the influence of this researcher in this research field.

ANALYSIS AND COMPARISON

4.1 DIKW Model

The full name of the DIKW model is Data, Information, Knowledge, and Wisdom. In this research, the leading framework was built by this model. This model gives us a straightforward way to discover, solve and analyze problems by data analysis. In **figure[14]**, the Data is at the bottom of the pyramid. Data means the symbol(basic details) or signals of every object by their facts. There is no rule to follow. By organizing and processing data in a certain way and analyzing the relationship between the data, the Data has meaning, which is information. So information can also be seen as understood news. If Data is a collection of facts, conclusions about attributes can be drawn. Then knowledge is the collection of information, which makes helpful information. Knowledge is the application of information, judging, and confirming information. Wisdom can be summarized as the ability to make correct judgments and decisions, including the best use of knowledge.

4.2 Data to Information

After we confirmed the problem and collected data from various sources, we need to convert these dates to valuable resources to help us in future steps, as we saw in the DIKW model[3]. (This information can answer some simple questions, such as Who? What? Where? When?)How could we convert this dataset to helpful information? Based on researchers signed to this program, we could separate them. In **figure[4]**, this figure has shown that using a groupby method to roughly classify these data into small groups based on their working institution.. (Academic, Government, Utilities, NGO, Company, Consultant), and count the number of researchers in each category. Academic has the most significant number of people (#54), and the smallest is Company (#6). We have 158 people attending this program. However, it is hard to find some researchers' data on websites. It is essential to clean these missing data to prevent inaccurate analysis. Since we focus on researchers' citations, publications, and the number of reads is our primary key to evaluating their influence. We have to drop the data which their citations are missing(shown in **figure[7]**). After we cleaned the data, there were only 91 data that contained the primary attribute in this dataset. In **figure[5] and figure[6]**, this diagram shows the percentage of available data for each type and the details of missing data. (count the numbers of each attribute[4] group by their classes). Regardless of sample sizes(we will discuss sample size in the next section), the Academic style has the most available data, 91% in this dataset. Consultants, Government, NGOs, and companies have around 50% of available data. Utilities have the lowest available data(42%). Now, We have a brief view of our collection data. Since we convert these data by group, these data have been transferred to information. We can get higher-level knowledge from information.

4.2 Information to Knowledge and Comparison

This section will discuss the meaning of our information and how that information will apply to our problem. (Knowledge can answer the question of "how?" and can help us model and simulate.) In **figure[5] and figure[6],** we already know the percentage of available data for each type. Although, NGOs and companies have 50% available data. However, only six rows belong to Company, and 12 rows belong to NGO, which means a few researchers are from both institutions. Also, the percentage of available data for both institutions is 58% and 50%. The poor rate of general data and sample size will make the analysis meaningless. On the other hand, we can see that the influence of researchers from these two institutions is minimal unless we get more samples so we can re-evaluate since we are focusing on researchers' impact. We need a quick view based on existing information. As mentioned before, citations are the most valuable and measurable variables as the primary key to measuring their influence. Researchers' articles and the number of reads of their articles are related to citations. Created linear regression model for the relationship between publications and citations(**figure[8]**) to compare with the multiple linear regression, which includes publications, citations, and number of reads(**figure[9]**). Because we collect the data from two different types of sources(Google Scholar and Researchgate(**figure[10]**)) hence, it is necessary to discuss them separately. In **figure[8]**, the figure shows that most of the observation points are close to the regression line, which means it is more accurate. In **figure[11]**, this diagram shows the linear regression model of Citations and Numbers of Reads. The graphs represent the relationship between those three significant values to evaluate the influence of Canadian researchers who signed into this program. The following section will discuss the result from the knowledge, which is also the final part of our DIKW model.

4.3 Knowledge to Wisdom

After processing our data to knowledge, the models and tables give us all information related to this analysis problem. (Wisdom can answer the "why" question.) From now on, we need to quantify the influence of a function that could help to evaluate any researchers who have all the required information. In **figure[12],** this figure represents the OLS regression summary.

We can get the multiple linear regression function:
Citations=62.431+31.0645*Publications-0.0469*Reads.

There is multiple information required to be noticed. R-squared is 0.485, which means that this model only applied to 48.5% of samples in this program. It is an inaccurate model by theorem. Secondly, P>|t| values for each variable are 0.872,0.000 and 0.313. Due to the intercept(0.872) and reads(0.313) being greater than 0.005, which is not a significant value. On the other hand, if researchers have no publications, which mean he/they can not gain influence in articles. Hence, there will not be an initial influence on one researcher.They both similarly, we will not influence the quantity of reading if there are no publications posted. Therefore, variable Publications is the only significant to this model. Based on this model, we can know if one researcher posted one publication in public. They expected to get 31 citations. As we mentioned before, researchers have more citations that will have a greater influence on articles. After modifying the function:

Citations=31.0645*Publications

4.3.1 Depth classification

Some researchers with tenThey publications posted have a similar quantity of citations to others with 50 publications, which might cause the model not to be precise. Hence, a decision tree regressor is required to classify the level of the influence of researchers. Figure [13] represents the estimated quantity of citations with various publications. In this tree, the numbers of citations have been converted to log(citations) (small cap), which will help to lower the MSE. For example, the first node shows that if publications are small or equal to 18.5, the value is 5.186. Then, the actual citation number is e^5.186=179.(All numbers are around integers. Due to the publications and citations being counted by an integer) This graph will classify their influence into three levels: Low Impact Level, Medium Impact Level, and High Impact Level. If researchers posted, the number of publications is between 5 to 19, and the number of citations of this researcher is around 90 to 178. We will consider this researcher to have a low impact level in articles. Secondly, if publications are between 20 to 52, this researcher's citations are around 179 to 1761. This researcher has a medium impact level in articles. Finally, if researchers have posted the number of publications between 53 to 162, the number of citations around 1762 to 3587. These researchers will be considered to have a high-level impact on articles. The function of this depth classification will be:

Suppose X is the number of articles published by researchers

C is the number of citations of this researcher.

if X<=5 AND C<=90 → , No Impact

if 5<=X<=19 AND 90<=C<=178 → Low Impact Level

if 20<=X<=52 AND 179<=C<=1761 → Medium Impact Level

if 53<=X<=162 AND 1762<=C<=3587 → High Impact Level

if X>162 AND C>3587 → Top Researcher

Furthermore, if this researcher has enough publications, their citation numbers are lower than this level. These researchers will transfer to the lower level if their citation numbers are higher than the level where they were placed. Then, this researcher will turn to a higher level. Because the numbers of citations are the most important thing to evaluate a researcher's influence level in articles. If one researcher has much more sources, peers widely accept their research.

The function are shown below:

if 5<=X <=19 AND C<90 → No Impact

if 5<=X <=19 AND C>178 → Medium Impact Level

if 20<=X<=52 AND C>1761 → High Impact Level

if 20<=X<=52 AND C<179 → Low Impact Level

if 53<=X<=162 AND C<1762 → Medium Impact Level

if 53<=X<=162 AND C>3587 → Top Researcher

4.5 Verify and Examining

Verifying is always an important thing to support our opinion every time. Figure **[2] and figure[3]** show two top researchers well known in the environment and civil engineering field. To verify our method, we will estimate the numbers of citations by their publication numbers. Daniel Schrag has posted 365 articles as known in public. Eric Miller has 245 articles published in public. According to our linear regression model, their estimated citations will be 11,338 and 7610, close to the accurate citations numbers 11498,6937 for both researchers(reference Citations numbers since 2016). After getting their citation numbers, we can evaluate their influence level in articles. First, Daniel has 365 articles more significant than 162, and his citations are greater than 3587. Secondly, Eric has 245 articles, and his citation numbers are more important than 3587. Both of them achieved the top researcher level, which is the same as the real world.

CONCLUSION

This research report discusses the thinking and method to quantify and analyze the influence of researchers in articles. DIKW model leads the way to depth classify the data to use functions to help us evaluate energy researchers' influence in writings. However, this model's applied percentage is only 48.5%(R squared is 0.485). Too small sample sizes might cause them both to have a more precise analysis. There are too many missing data that limit our analysis processing even though there are too many limitations. But, we are still finding a way to evaluate and analyze the influence by using collected data, cleaning data, creating models and analyses. Their article numbers can estimate the influence of researchers in articles. After getting the estimated quantity of citations, we can classify their impact level by decision tree regressor. Therefore, we can have a clear understanding of the influence of this researcher. The model will be more accurate if more data is added to this program. According to the above analysis, we have a framework to evaluate the researcher's influence on the article. Some low-impact researchers will deliberately increase their scores.no publications are. For example, the system on LinkedIn's homepage will prompt users the number of times they have appeared on the recruiter page. If the user is searched too low, the system will improve the search probability. At this point, if the user wants to be seen by more recruiters, his resume. We need to follow the system prompts to complete the information. Although measuring popularity is essential for guiding and disseminating new ideas from the government or academia. (Bandwagon effect: People tend to believe what influential people say and are affected by subtle influence.) The impact evaluation report should not be made public (only internal system evaluation) to avoid deliberately increasing their influence. When the government wants to quickly popularize and implement a new policy, it can consider the influence of these researchers to popularize their policy.

Appendix:

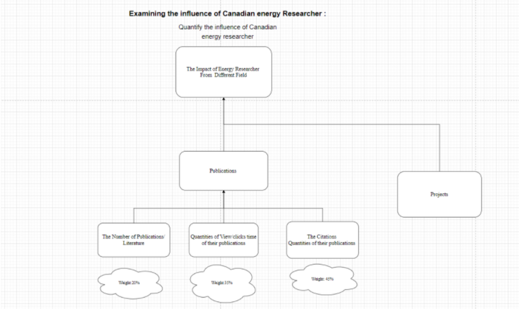
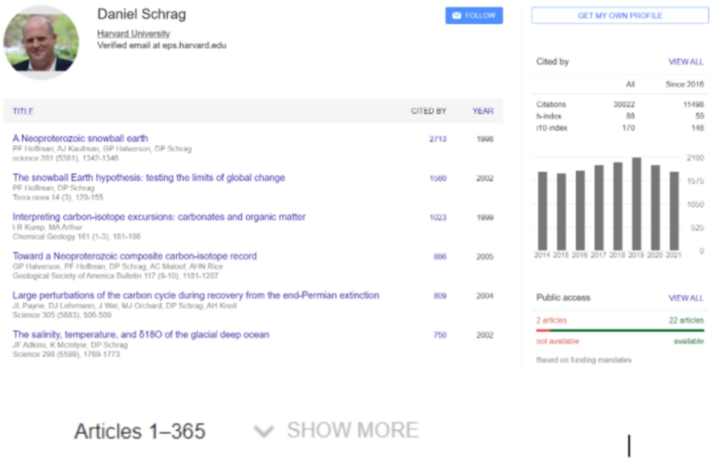

Examining the influence of Canadian energy Researcher :
Quantify the influence of Canadian energy researcher
The Impact of Energy Researcher From Different Field
Publications
Projects
The Number of Publications/ Literature
Quantities of View/clicks time of their publications
The Citations Quantities of their publications
Weight:20%
Weight:35%
Weight: 45%

Figure 1: The Framework of the examining influence of Canadian researchers on literature.
Note: Separate the measuring into two different parts. Give a straightforward diagram to acknowledge what should do next in this report.



Daniel Schrag
Harvard University
Verified email at eps.harvard.edu

FOLLOW

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
A Neoproterozoic snowball earth PF Hoffman, AJ Kaufman, GP Halverson, DP Schrag science 281 (5381), 1342-1346	2713	1998
The snowball Earth hypothesis: testing the limits of global change PF Hoffman, DP Schrag Terra nova 14 (3), 129-155	1580	2002
Interpreting carbon-isotope excursions: carbonates and organic matter LR Kump, MA Arthur Chemical Geology 161 (1-3), 181-198	1023	1999
Toward a Neoproterozoic composite carbon-isotope record GP Halverson, PF Hoffman, DP Schrag, AC Maloof, AHN Rice Geological Society of America Bulletin 117 (9-10), 1181-1207	886	2005
Large perturbations of the carbon cycle during recovery from the end-Permian extinction JL Payne, DJ Lehrmann, J Wei, MJ Orchard, DP Schrag, AH Knoll Science 305 (5683), 506-509	809	2004
The salinity, temperature, and δ18O of the glacial deep ocean JF Adkins, K McIntyre, DP Schrag Science 298 (5599), 1769-1773	750	2002

Cited by VIEW ALL

	All	Since 2016
Citations	30022	11498
h-index	88	59
i10-index	170	148

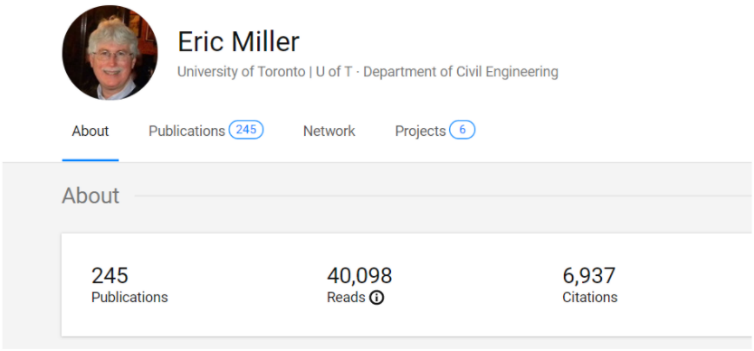
Public access VIEW ALL

2 articles not available
22 articles available
Based on funding mandates

Articles 1–365 SHOW MORE

Figure2: Variables in research engine

Note: This figure shows Daniel Schrag who is a top researcher in the environmental field.



Eric Miller
University of Toronto | U of T · Department of Civil Engineering

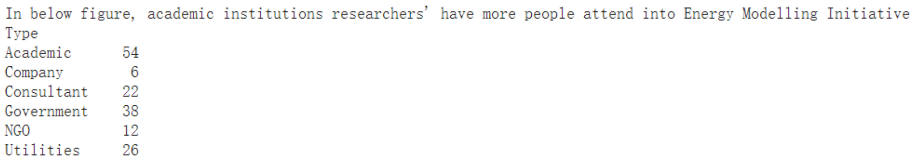
About Publications 245 Network Projects 6

About

245	40,098	6,937
Publications	Reads	Citations

Figure3: Variables in research engine

Note: This figure shows Eric Miller who is a top researcher in the civil engineering field.



```
In below figure, academic institutions researchers' have more people attend into Energy Modelling Initiative
Type
Academic      54
Company        6
Consultant    22
Government    38
NGO           12
Utilities     26
```

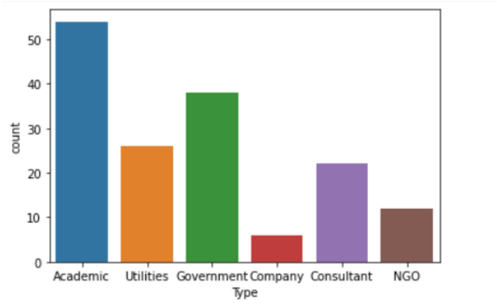
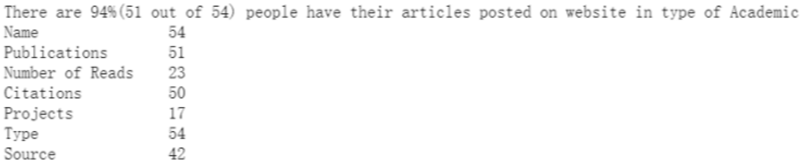
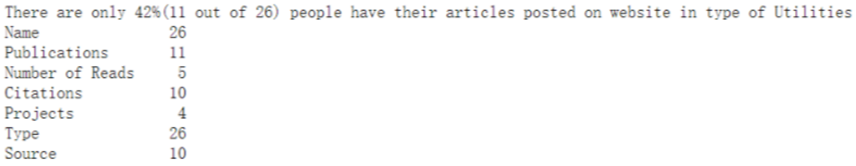


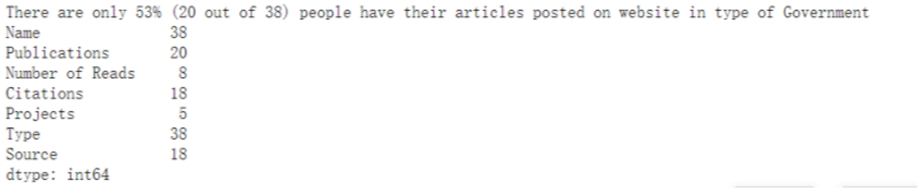
Figure 4: Classified the different groups of researchers who attend into this program
Note: Classified their identity based on their working institution, count numbers of researchers by their group.



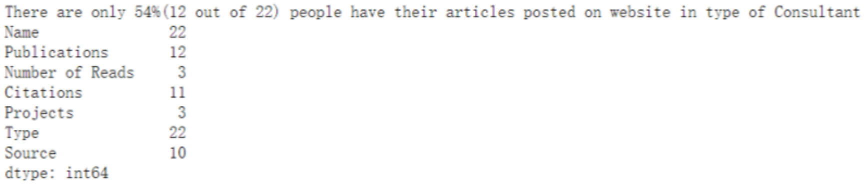
```
There are 94%(51 out of 54) people have their articles posted on website in type of Academic
Name               54
Publications       51
Number of Reads    23
Citations          50
Projects           17
Type               54
Source             42
```



```
There are only 42%(11 out of 26) people have their articles posted on website in type of Utilities
Name               26
Publications       11
Number of Reads     5
Citations          10
Projects            4
Type               26
Source             10
```



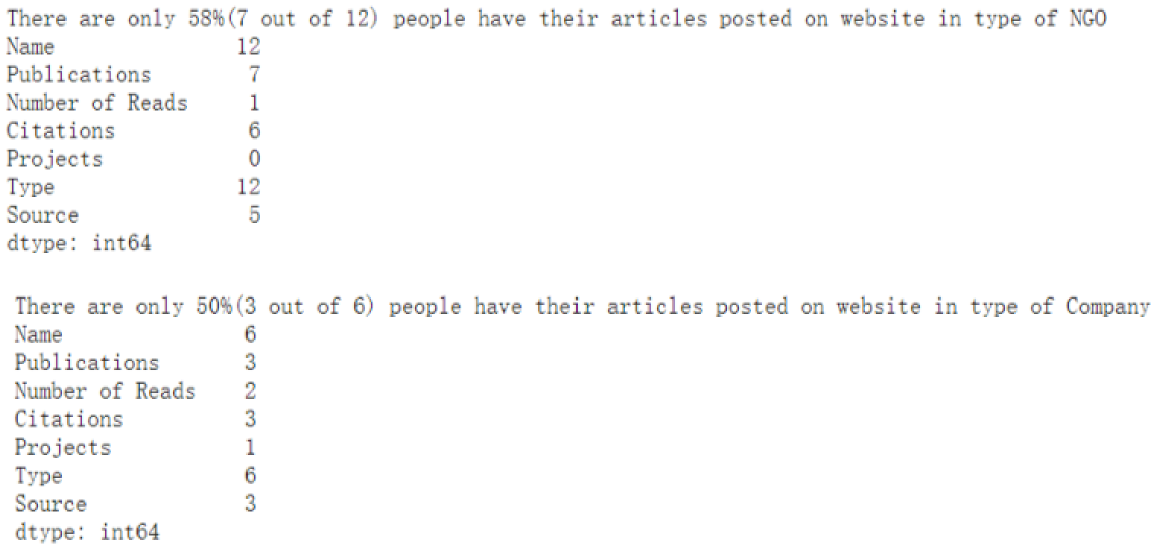
```
There are only 53% (20 out of 38) people have their articles posted on website in type of Government
Name               38
Publications       20
Number of Reads     8
Citations          18
Projects            5
Type               38
Source             18
dtype: int64
```



```
There are only 54%(12 out of 22) people have their articles posted on website in type of Consultant
Name               22
Publications       12
Number of Reads     3
Citations          11
Projects            3
Type               22
Source             10
dtype: int64
```

Figure 5: Count the missing data on each type

Note: Counting the percentage of available data for Academic, Utilities, Government and Consultant.

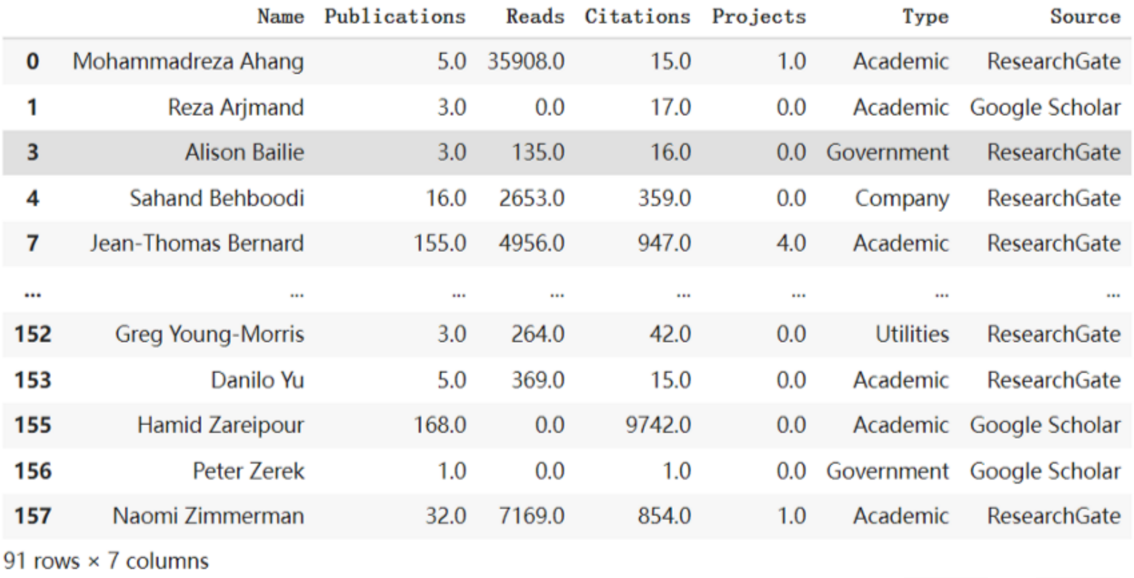


```
There are only 58%(7 out of 12) people have their articles posted on website in type of NGO
Name                12
Publications         7
Number of Reads      1
Citations            6
Projects             0
Type                12
Source               5
dtype: int64

There are only 50%(3 out of 6) people have their articles posted on website in type of Company
Name                6
Publications        3
Number of Reads     2
Citations           3
Projects            1
Type                6
Source              3
dtype: int64
```

Figure 6: Count the missing data on each type

Note: Counting the percentage of available data for NGO, Company



	Name	Publications	Reads	Citations	Projects	Type	Source
0	Mohammadreza Ahang	5.0	35908.0	15.0	1.0	Academic	ResearchGate
1	Reza Arjmand	3.0	0.0	17.0	0.0	Academic	Google Scholar
3	Alison Bailie	3.0	135.0	16.0	0.0	Government	ResearchGate
4	Sahand Behboodi	16.0	2653.0	359.0	0.0	Company	ResearchGate
7	Jean-Thomas Bernard	155.0	4956.0	947.0	4.0	Academic	ResearchGate
...	...	...	...	...	...	...	...
152	Greg Young-Morris	3.0	264.0	42.0	0.0	Utilities	ResearchGate
153	Danilo Yu	5.0	369.0	15.0	0.0	Academic	ResearchGate
155	Hamid Zareipour	168.0	0.0	9742.0	0.0	Academic	Google Scholar
156	Peter Zerek	1.0	0.0	1.0	0.0	Government	Google Scholar
157	Naomi Zimmerman	32.0	7169.0	854.0	1.0	Academic	ResearchGate

91 rows × 7 columns

Figure 7: Cleaning data
Note: Delete data which their citations are missing to ensure the validity of future analysis

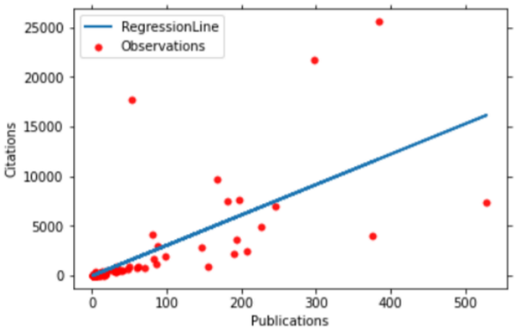


Figure8: Linear Regression Model of Citations ~ Publications
Note: Using this regression line to estimate and predict the relation between citations and publications.

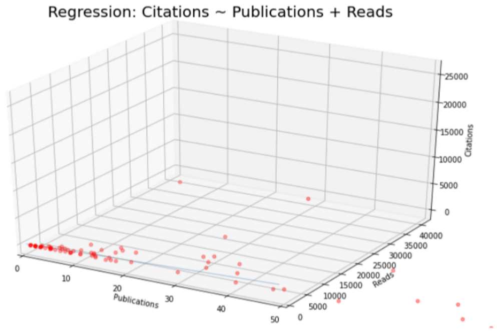
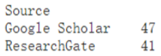


Figure 9: Multiple Linear Regression of Citations ~ Publications + Number of Reads
Note: The multiple linear regression of Citations ~ Publications + Number of Reads in 3D graph.



```
Source
Google Scholar    47
ResearchGate      41
```

Figure 10: Count the Number of Sources
Note: Count the quantity of data collected from different sources

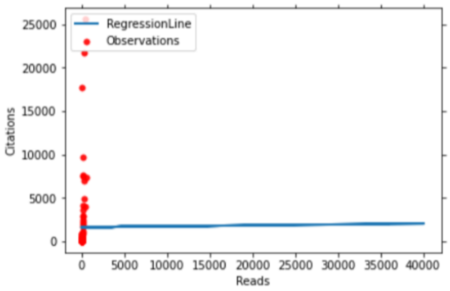
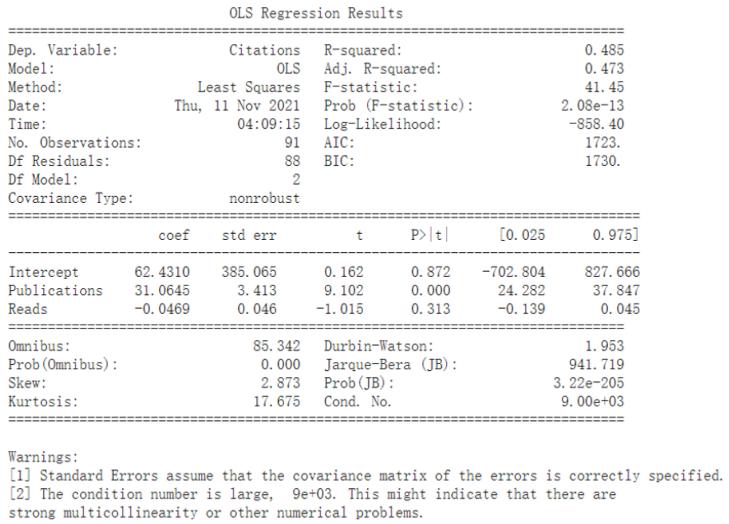


Figure 11: Linear Regression Model of Citations ~ Number of Reads
Note: Using this regression line to estimate and predict the relation between citations and Number of Reads.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Citations   R-squared:                       0.485
Model:                            OLS   Adj. R-squared:                  0.473
Method:                 Least Squares   F-statistic:                     41.45
Date:                Thu, 11 Nov 2021   Prob (F-statistic):           2.08e-13
Time:                        04:09:15   Log-Likelihood:                -858.40
No. Observations:                  91   AIC:                             1723.
Df Residuals:                      88   BIC:                             1730.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       62.4310    385.065      0.162      0.872    -702.804     827.666
Publications    31.0645      3.413      9.102      0.000      24.282      37.847
Reads           -0.0469      0.046     -1.015      0.313      -0.139       0.045
==============================================================================
Omnibus:                       85.342   Durbin-Watson:                   1.953
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              941.719
Skew:                           2.873   Prob(JB):                    3.22e-205
Kurtosis:                      17.675   Cond. No.                     9.00e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large,  9e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 12: OLS Regression Results of Citations ~ Publications + Reads

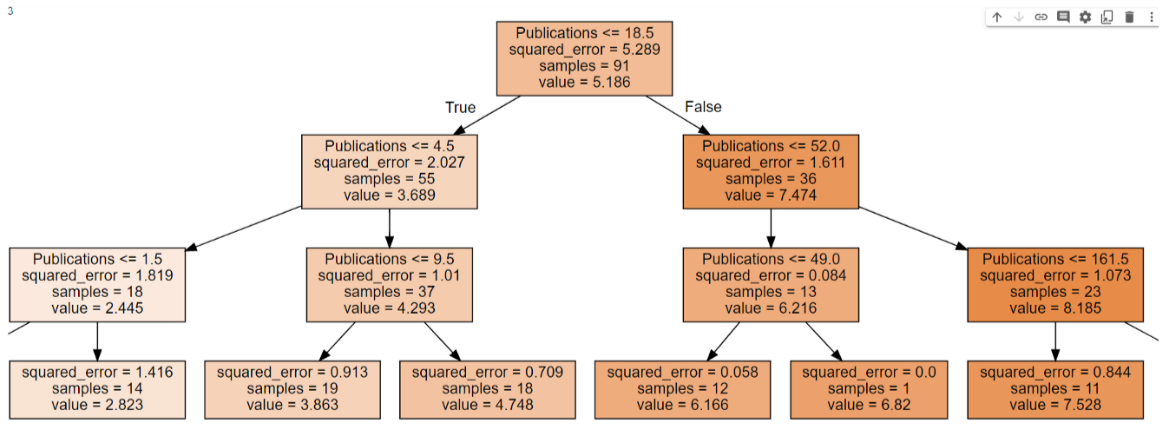


Figure 13: Decision Tree Regressor
Note: The depth of this tree is 3.

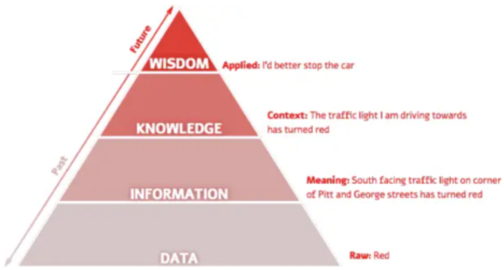

Future
Past
WISDOM
KNOWLEDGE
INFORMATION
DATA
Applied: I'd better stop the car
Context: The traffic light I am driving towards has turned red
Meaning: South facing traffic light on corner of Pitt and George streets has turned red
Raw: Red

Figure 14: DIKW Model

Reference:

[1]:https://www.gov.uk/government/news/uk-china-scientists-deepen-understanding-of-climate-change-risks

[2]: https://civmin.utoronto.ca/home/about-us/directory/professors/eric-miller/

[3] https://www.i-scoop.eu/big-data-action-value-context/dikw-model/

[4] An attribute value is a value used to describe a specific member.