LLO 8200 Data Science Final Report - Kelley Brundage

Kelley Brundage

8/4/2019

Final Report: World University Rankings

Introduction

Ranking universities is very challenging and comes with a variety of political and controversial practices. Throughout the world, there are hundreds of different national and international university ranking systems, many that disagree with each other.

Fortunately, there are a series of public resources available that provide ranking data of this nature. Specifically, I have chosen the World University Ranking dataset provided on Kaggle.com as these files contain three global university rankings from various places throughout the world.

Having the ability to identify and understand how hundreds of institutions throughout the world compare to each other is vital to ensuring accuracy and acceptability. Nevertheless, ranking systems continue to be famous for what they have been doing over the decades, highlighting who is the best of the best in the global context.

Problem and Approach

I intend to compare the three global ranking systems to the amount of publications/research as well as highly cited Researchers/Citations at each institution per ranking system by approaching each dataset with an analytical and statistical viewpoint.

I am analyzing the dataset-specific to the area of research/academic and if common challenges that exist with all ranking systems exist — analyzing variables such as the number of publications and citations produced based on world rank or national rank.

Data

How was the data acquired?

Kaggle.com Dataset file: World University Rankings website:
https://www.kaggle.com/mylesoneill/world-university-rankings

Format of Data

There are a total of three files (.csv) that make up this data set containing three ranking systems: The Center for World University Rankings (CWUR); Academic Ranking of World Universities (Shanghai Ranking) and Times Higher Education World University Ranking (times).

University Ranking Data

The *Center for World University Rankings* is a less well know listing that comes from Saudi Arabia, founded in 2012.

1. How do these rankings compare to each other?
2. Are the various criticisms levied against these rankings fair or not?
3. How does your alma mater fare against the world?

The *Academic Ranking of World Universities*, also known as the *Shanghai Ranking*, is an equally influential ranking. It was founded in China in 2003 and has been criticized for focusing on raw research power and for undermining humanities and quality of instruction.

The *Times Higher Education World University Ranking* is widely regarded as one of the most influential and widely observed university measures. Founded in the United Kingdom in 2010, it has been criticized for its commercialization and for undermining non-English-instructing institutions.

Describe Data/Variables

Center for World University Rankings Methodology

Publishes the only global university ranking that measures the quality of education and training of students as well as the prestige of the faculty members and the quality of their research without relying on surveys and university data submissions.

CWUR uses seven objective and robust indicators to rank the worlds top 1000 universities:

1. Quality of Education, measured by the number of a university's alumni who have won major international awards, prizes, and medals relative to the university's size (15%)
2. Alumni Employment, measured by the number of a university's alumni who have held CEO positions at the world's top companies relative to the university's size (15%)
3. Quality of Faculty, measured by the number of academics who have won major international awards, prizes, and medals (15%)
4. Research Output, measured by the total number of research papers (15%)
5. Quality Publications, measured by the number of research papers appearing in top-tier journals (15%)
6. Influence, measured by the number of research papers appearing in highly-influential journals (15%)
7. Citations, measured by the number of highly-cited research papers (10%)

ARWU/Shanghai Methodology

ARWU considers every university that has any Nobel Laureates, Fields Medalists, Highly Cited Researchers, or papers published in Nature or Science. Also included are universities with a significant amount of articles indexed by Science Citation Index-Expanded (SCIE) and Social Science Citation Index (SSCI). In total, more than 1200 universities are ranked in this dataset, and the 500 best results in being published on the web.

Universities are ranked by several indicators of academic or research performance, including alumni and staff winning Nobel Prizes and Fields Medals, highly cited researchers, papers published in Nature and Science, papers indexed in major citation indices, and the per capita academic performance of an institution. For each indicator, the highest scoring institution is assigned a score of 100, and other institutions are calculated as a percentage of the top score.

The distribution of data for each indicator is examined for any significant distorting effect; standard statistical techniques are used to adjust the indicator if necessary. Scores for each indicator are weighted as shown below to arrive at a final overall score for an institution. The highest scoring institution is

assigned a score of 100, and other institutions are calculated as a percentage of the top score. An institution's rank reflects the number of institutions that sit above it.

Times Dataset Methodology

Only global performance tables that judge research-intensive universities across all their core missions: teaching, research, knowledge transfer and international outlook. The dataset uses 13 carefully calibrated performance indicators to provide the most comprehensive and balanced comparisons, trusted by students, academics, university leaders, industry, and even governments. The basic methodology for this year's rankings is similar to that employed since the 2011-12 tables, but there have been significant changes made to the underlying data (World University Rankings | Times Higher Education (THE)).

The performance indicators are grouped into five areas:

Teaching (the learning environment)
Research (volume, income, and reputation)
Citations (research influence)
International outlook (staff, students, and research)
Industry income (knowledge transfer).

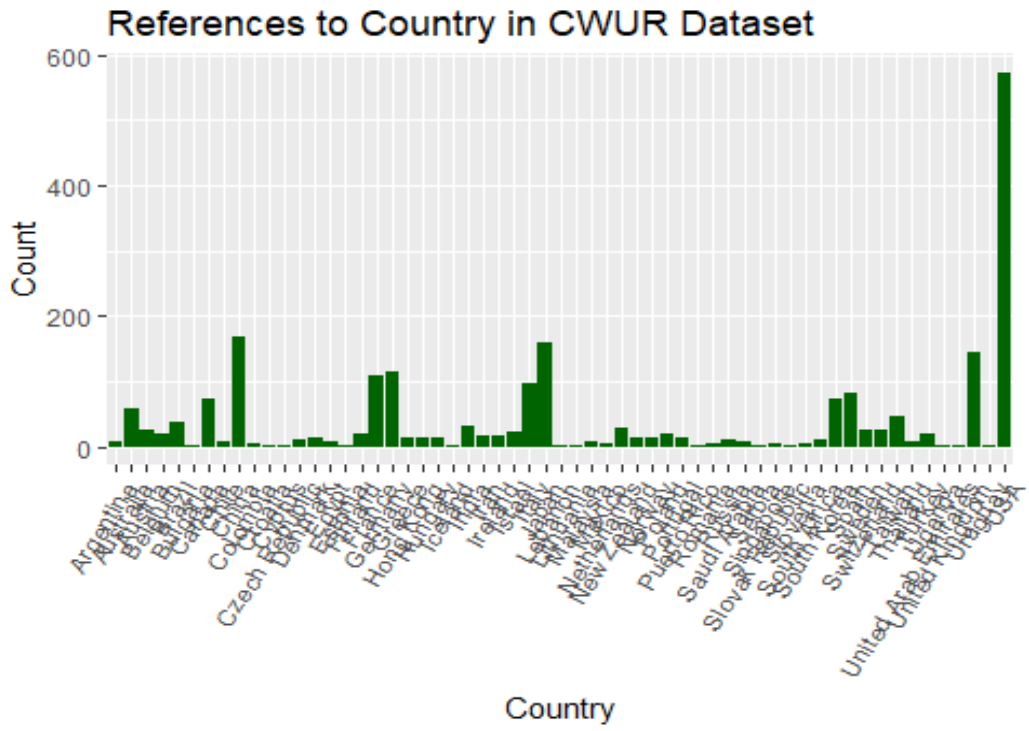
Supporting Displays/Visualizations

CWUR Dataset

CWUR Histogram(s):

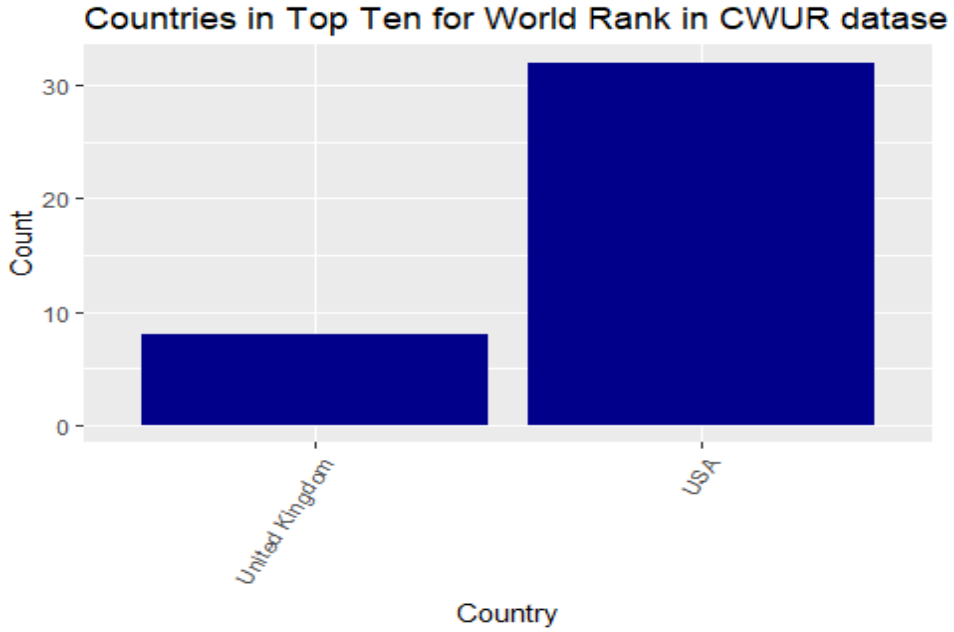
Graph #1:

Reflects all of the countries referenced in the CWUR Dataset by World Rank. There are fifty-nine (59) countries referenced in the CWUR dataset.



Graph #2:

Reflects the countries referenced in the CWUR Top Ten by World Rank Dataset. Only the United Kingdom and the United States of America (USA) fall into the top ten by world rank.

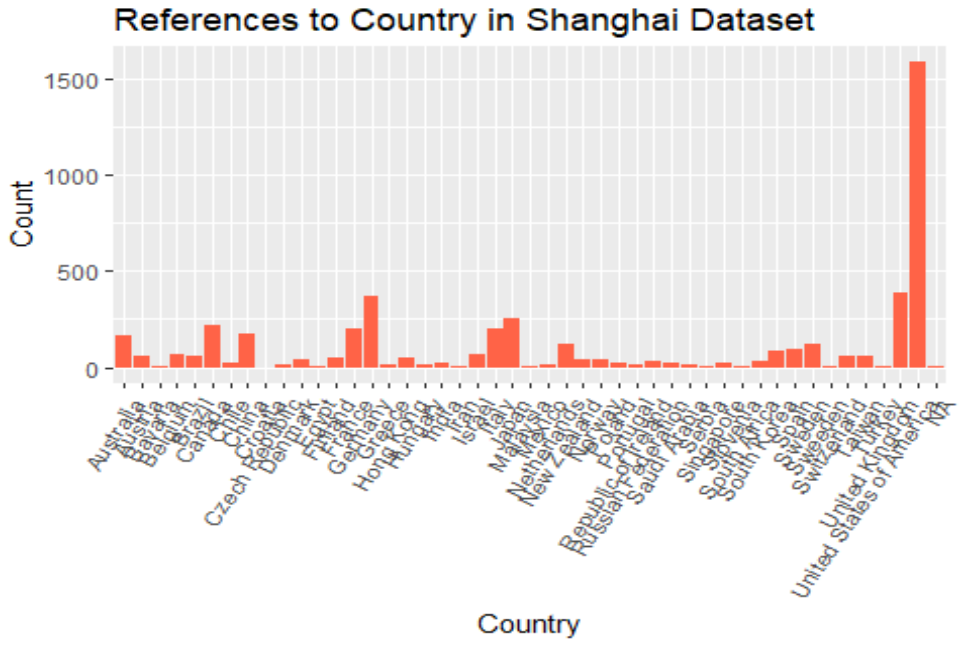


Shanghai Dataset

Shanghai Histogram(s):

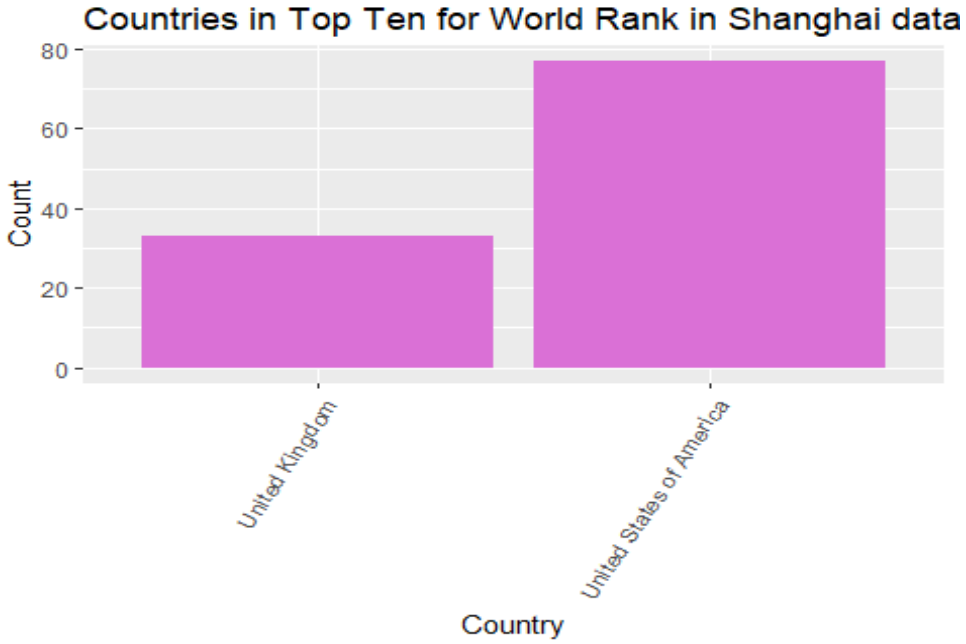
Graph #3:

Reflects all of the countries referenced in the Shanghai Dataset by World Rank. There are forty-seven (47) countries referenced in the Shanghai dataset.



Graph #4:

Reflects the countries referenced in the Shanghai Top Ten by World Rank Dataset. Only the United Kingdom and the United States of America (USA) fall into the top ten by world rank.

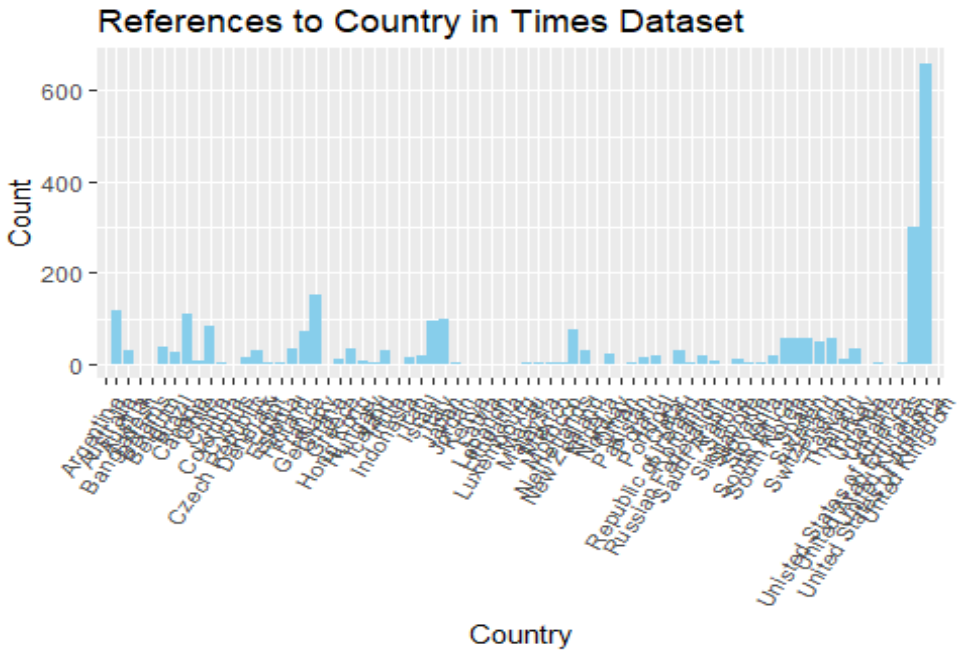


Times Dataset

Times Histogram(s):

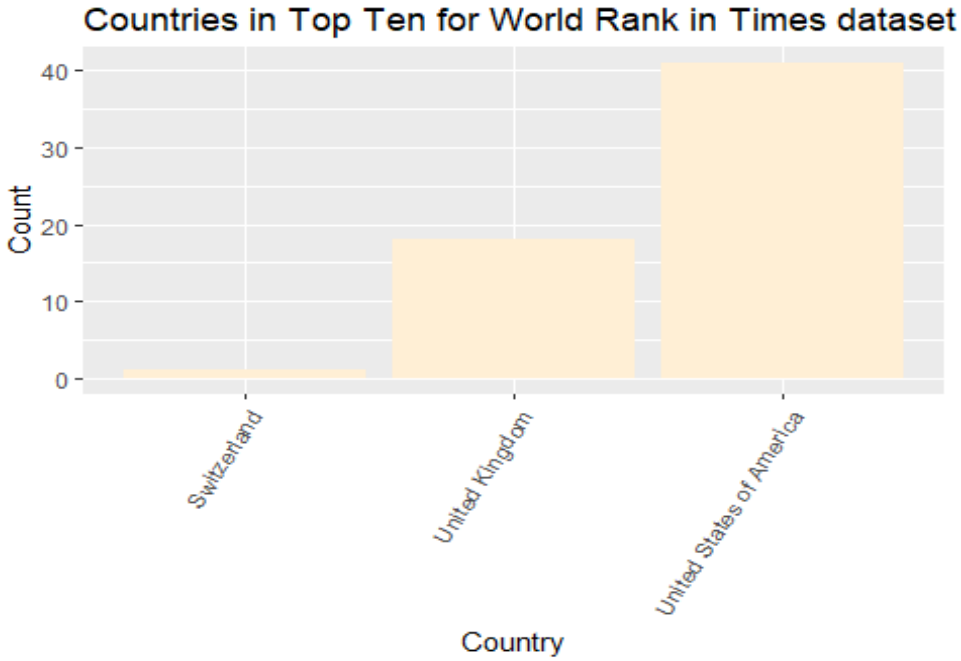
Graph #5:

Reflects all of the countries referenced in the Times Dataset by World Rank. There are seventy-two (72) countries referenced in the Times dataset.



Graph #6:

Reflects the countries referenced in the Times Top Ten by World Rank Dataset. Only Switzerland, the United Kingdom and the United States of America (USA) fall into the top ten by world rank.



Exploratory Data Analysis

The dataset fields listed below were used to compare and analyze each ranking system

CWUR Dataset Fields:
world_rank; institution; country; national_rank; publications; citations

Publications (measured by # of papers in top-tier journals - 15%)
Citations

Shanghai Dataset Fields:
word_rank; university_name; national_rank; pub; hici

Pub (Publications)
HICI (Highly Cited Researchers)

Merged School and Country data into this file by adding a new column after University Name that indicates the Cou ntry.

Times Dataset Fields:
world_rank; university; country; research; citations

Research (volume, income, and reputation)
Citations (research influence)

Extensive Investigation of Dataset

Investigate data: distribution of data, correlations, associations, and predictive potential to solve the proposed problem:

Continued review and analysis of the datasets have led to identifying a series of standard fields within the three primary ranking system datasets: CWUR, Shanghai, and Times. All three hold standard columns such as the World Rank, Institution/University Name, Publications/Research, and Citations. The continued analysis will expand to look at both the research and citations between the three ranking systems as well as to compare the countries that appear in each dataset depending on the references within the datasets.

Data Analysis

Below are the means of the overall publications/citations for the three global ranking systems. For all three datasets; CWUR, Shanghai and Times, an average 50 publications and citations are produced per institution.

Regression Data for CWUR - Publications & Citations

Publications:

Mean of Publications from CWUR Dataset

mean(Publications, na.rm = T)
Min. :459.9
1st Qu.:459.9
Median :459.9
Mean :459.9
3rd Qu.:459.9
Max. :459.9

Summary: Mean Publications Percent within the CWUR Dataset

mean(Publications, na.rm = T)
Min. :459.9
1st Qu.:459.9
Median :459.9
Mean :459.9
3rd Qu.:459.9
Max. :459.9

CWUR Publications: Country with the maximum and minimum publications within the CWUR dataset

Country	mean_pub	sd_pub	max_pub	min_pub
1 Puerto Rico	99.8	0.0643	99.9	99.8

Country	mean_pub	sd_pub	max_pub	min_pub
1 Singapore	11.5	6.54	19.0	6.00

Country	mean_pub	sd_pub	max_pub	min_pub
Length:59	Min. :11.52	Min. : 0.06431	Min. :19.01	Min. : 0.000
Class :character	1st Qu.:46.29	1st Qu.: 7.30438	1st Qu.:83.29	1st Qu.: 9.709
Mode :character	Median :62.40	Median :17.32889	Median :93.82	Median :32.015
NA	Mean :60.88	Mean :15.71551	Mean :86.76	Mean :37.983
NA	3rd Qu.:73.58	3rd Qu.:23.86361	3rd Qu.:98.14	3rd Qu.:60.482
NA	Max. :99.82	Max. :30.25227	Max. :99.95	Max. :99.773

Citations:

Mean of Citations from CWUR Dataset

mean(cwurten$Citations, na.rm = T)
Min. :413.4
1st Qu.:413.4
Median :413.4
Mean :413.4
3rd Qu.:413.4
Max. :413.4

Summary: Mean Citations Percent within the CWUR Dataset

mean(citations_p, na.rm = T)
Min. :47.85
1st Qu.:47.85
Median :47.85
Mean :47.85
3rd Qu.:47.85

Max. :47.85

CWUR Citations: Country with the maximum and minimum citations within the CWUR dataset

Country	mean_cit	sd_cit	max_cit	min_cit
1 Lebanon	806	8.49	812	800
2 United Arab Emirates	806	8.49	812	800
3 Uruguay	806	8.49	812	800

Country	mean_cit	sd_cit	max_cit	min_cit
1 Singapore	135.	77.6	220	50

Country	mean_cit	sd_cit	max_cit	min_cit
Length:59	Min. :134.8	Min. : 7.778	Min. :220.0	Min. : 1.0
Class :character	1st Qu.:378.8	1st Qu.: 90.826	1st Qu.:645.0	1st Qu.: 68.5
Mode :character	Median :502.0	Median :165.363	Median :812.0	Median :220.0
NA	Mean :488.7	Mean :147.426	Mean :711.4	Mean :276.0
NA	3rd Qu.:597.0	3rd Qu.:220.743	3rd Qu.:812.0	3rd Qu.:406.0
NA	Max. :806.0	Max. :258.465	Max. :812.0	Max. :800.0

Regression Data for Shanghai - Publications & Highly Cited Researchers (HICI)

Publications:

Mean of Publications from Shanghai Dataset

mean(shangten$Publications, na.rm = T)
Min. :38.25
1st Qu.:38.25
Median :38.25
Mean :38.25
3rd Qu.:38.25
Max. :38.25

Summary: Mean Publications Percent within the Shanghai Dataset

mean(publications_p, na.rm = T)
Min. :49.88
1st Qu.:49.88
Median :49.88
Mean :49.88
3rd Qu.:49.88
Max. :49.88

Shanghai Publications: Country with the maximum and minimum publications within the Shanghai dataset

Country	mean_spub	sd_spub	max_spub	min_spub
Length:47	Min. :25.00	Min. : 4.739	Min. : 27.90	Min. : 7.30
Class :character	1st Qu.:30.23	1st Qu.: 6.626	1st Qu.: 42.00	1st Qu.:10.40
Mode :character	Median :33.00	Median : 9.002	Median : 52.40	Median :17.10
NA	Mean :34.10	Mean : 9.035	Mean : 54.97	Mean :17.27
NA	3rd Qu.:37.41	3rd Qu.:11.162	3rd Qu.: 64.20	3rd Qu.:22.00
NA	Max. :44.81	Max. :15.070	Max. :100.00	Max. :35.40
NA	NA's :2	NA's :3	NA's :2	NA's :2

HICI/Citations:

Mean of HICI/Citations from Shanghai Dataset

mean(shangten$Highly Cited Researchers, na.rm = T)
Min. :16.22
1st Qu.:16.22
Median :16.22
Mean :16.22
3rd Qu.:16.22
Max. :16.22

Summary: Mean HICI/Citations Percent within the Shanghai Dataset

mean(hici_p, na.rm = T)
Min. :48.48
1st Qu.:48.48
Median :48.48
Mean :48.48
3rd Qu.:48.48
Max. :48.48

Shanghai HICI/Citations: Country with the maximum and minimum citations within the Shanghai dataset

Country	mean_spub	sd_spub	max_spub	min_spub
Length:47	Min. : 2.067	Min. : 3.272	Min. : 5.10	Min. :0.0000
Class :character	1st Qu.: 6.985	1st Qu.: 5.584	1st Qu.: 15.40	1st Qu.:0.0000
Mode :character	Median : 9.130	Median : 6.800	Median : 26.60	Median :0.0000
NA	Mean :10.555	Mean : 7.644	Mean : 28.46	Mean :0.4644
NA	3rd Qu.:14.320	3rd Qu.: 9.024	3rd Qu.: 33.20	3rd Qu.:0.0000
NA	Max. :22.520	Max. :18.188	Max. :100.00	Max. :7.2000
NA	NA's :2	NA's :3	NA's :2	NA's :2

Regression Data for Times - Research/Publications & Citations

Research/Publications:

Mean of Research/Publications from Times Dataset

mean(Research, na.rm = T)
Min. :35.91
1st Qu.:35.91
Median :35.91
Mean :35.91
3rd Qu.:35.91
Max. :35.91

Summary: Mean Research/Publications Percent within the Times Dataset

mean(research_p, na.rm = T)
Min. :49.92
1st Qu.:49.92
Median :49.92
Mean :49.92
3rd Qu.:49.92
Max. :49.92

Times Research/Publicatons: Country with the maximum and minimum citations within the Times dataset

	Country	mean_tpub	sd_tpub	max_tpub	min_tpub
1	Singapore	68.1	13.6	87.2	47.8

	Country	mean_tpub	sd_tpub	max_tpub	min_tpub
1	Morocco	6.4	0.141	6.5	6.3

Country	mean_tpub	sd_tpub	max_tpub	min_tpub
Length:72	Min. : 6.40	Min. : 0.1414	Min. : 6.50	Min. : 2.90
Class :character	1st Qu.:11.21	1st Qu.: 4.4516	1st Qu.:13.60	1st Qu.: 8.15
Mode :character	Median :20.21	Median : 8.4545	Median :28.00	Median :10.35
NA	Mean :22.70	Mean : 9.9026	Mean :38.73	Mean :11.77
NA	3rd Qu.:30.62	3rd Qu.:14.5987	3rd Qu.:60.48	3rd Qu.:12.18
NA	Max. :68.09	Max. :25.1347	Max. :99.40	Max. :47.80
NA	NA	NA's :18	NA	NA

Citations:

Mean of Citations from Times Dataset

mean(Citations, na.rm = T)
Min. :60.92
1st Qu.:60.92
Median :60.92

Mean :60.92

3rd Qu.:60.92

Max. :60.92

Summary: Mean Citations Percent within the Times Dataset

mean(citations_p, na.rm = T)
Min. :49.93
1st Qu.:49.93
Median :49.93
Mean :49.93
3rd Qu.:49.93
Max. :49.93

Times Citations: Country with the maximum and minimum citations within the Times dataset

Country	mean_tcit	sd_tcit	max_tcit	min_tcit
1 Luxembourg	84.8	NaN	84.8	84.8
Country	mean_tcit	sd_tcit	max_tcit	min_tcit
1 Ukraine	2.95	1.77	4.2	1.7

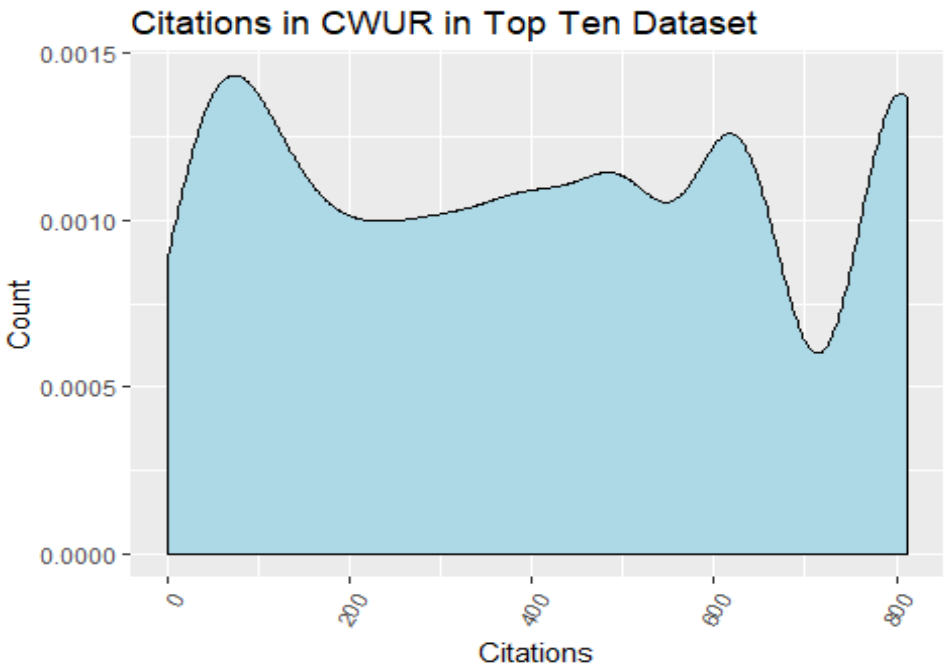
Country	mean_tcit	sd_tcit	max_tcit	min_tcit
Length:72	Min. : 2.95	Min. : 1.768	Min. : 4.20	Min. : 1.20
Class :character	1st Qu.:22.29	1st Qu.:11.765	1st Qu.: 32.05	1st Qu.:10.20
Mode :character	Median :43.94	Median :14.655	Median : 74.25	Median :18.10
NA	Mean :42.89	Mean :16.717	Mean : 63.58	Mean :21.99
NA	3rd Qu.:59.35	3rd Qu.:20.365	3rd Qu.: 91.55	3rd Qu.:27.25
NA	Max. :84.80	Max. :50.134	Max. :100.00	Max. :84.80
NA	NA	NA's :18	NA	NA

Visuals: Density Plots & Scatterplots

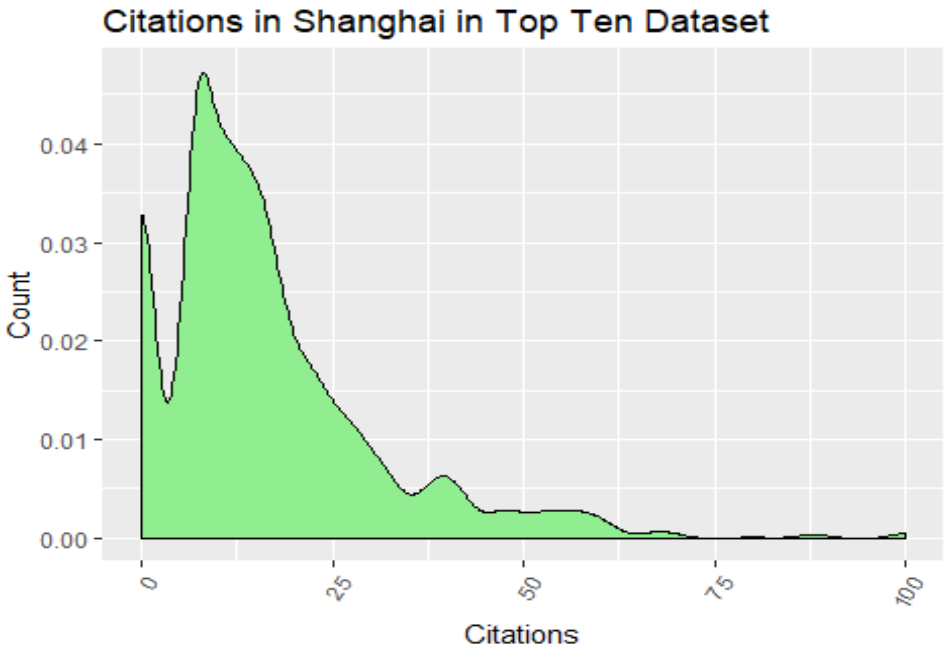
Density plots show the regression data for overall publications and citations within the three world ranking datasets.

Density Plot #1: Citations in CWUR Dataset
Level of citations referenced within the CWUR dataset. Reflects that the mid-way of 400 citations appears

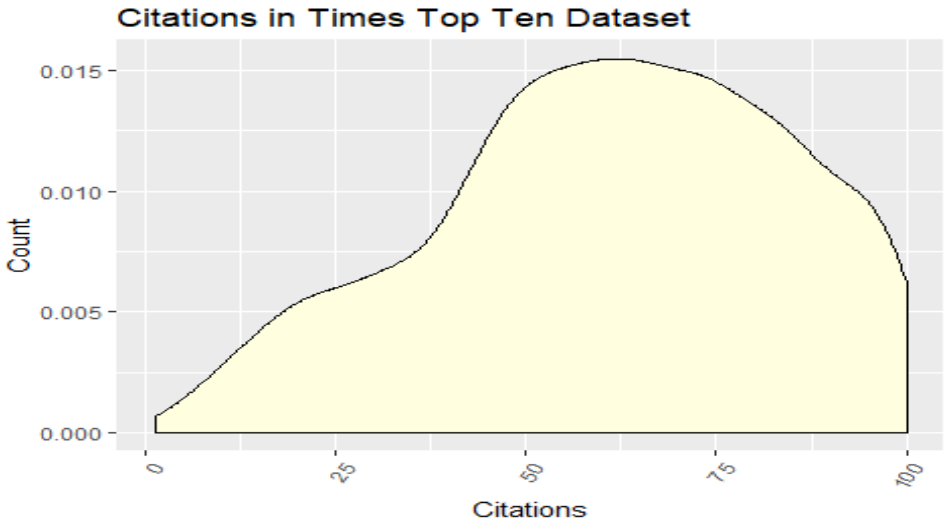
to be the average with two outlier areas of a low point between 0-100 citations and a high point between 700-800 citations.



Density Plot #2: Citations in Shanghai Dataset
Level of citations referenced within the Shanghai dataset. The peak of citations within this dataset is on average 10-12 or a little under the mid-way mark between 0-25 citations.



Density Plot #3: Citations in Times Dataset
The plot above shows the level of citations referenced within the Times dataset. Notice that the average or peak for the citations is at the middle mark between 50-75 citations, approximately 62.5 citations.



World Rankings Dataset Predictions

The data and charts below show the publication/citation predictions based on the three world datasets.

CWUR Dataset Prediction

Publications

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.68	6.709	9.79	3.502e-22
Citations	0.9536	0.01367	69.74	0

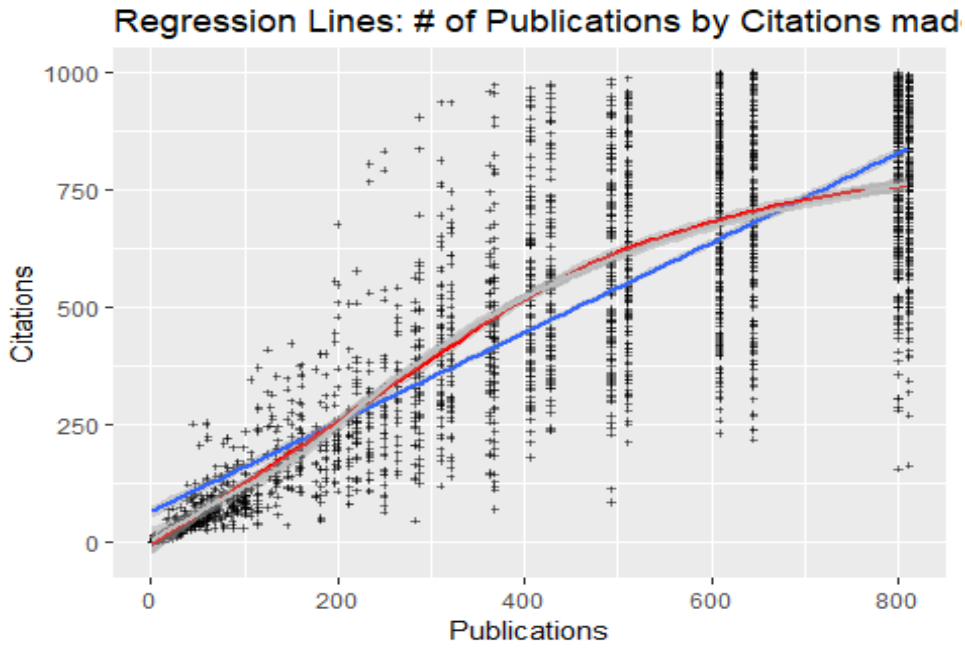
Fitting linear model: Publications ~ Citations

Observations	Residual Std. Error	R^2	Adjusted R^2
2200	169.5	0.6888	0.6886

Citations

Graph 7: CWUR Regression Lines

The Linear and LOESS regression lines are showing a positive relationship between the number of Citations and the number of Publications produced in the CWUR Top Ten dataset.



RMSE for CWUR Publications

[1] 169.4278

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
169.4	169.4	169.4	169.4	169.4	169.4

Coefficient Data for CWUR Publications

	2.5 %	97.5 %
(Intercept)	52.5259090	78.8396802
Citations	0.9267652	0.9803919

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.68	6.709	9.79	3.502e-22
Citations	0.9536	0.01367	69.74	0

Fitting linear model: Publications ~ Citations

Observations	Residual Std. Error	R^2	Adjusted R^2
2200	169.5	0.6888	0.6886

Shanghai Dataset Prediction

Publications

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.33	0.2077	136.4	0

Highly Cited Researchers	0.6117	0.009582	63.83	0

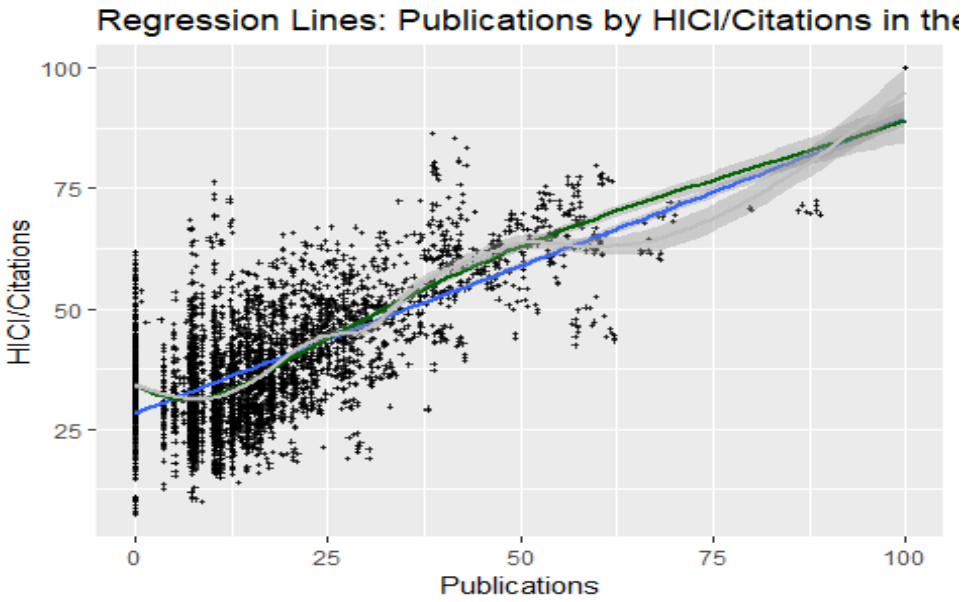
Fitting linear model: Publications ~ Highly Cited Researchers

Observations	Residual Std. Error	R^2	Adjusted R^2
4895	9.641	0.4544	0.4543

HICI/Citations

Graph 8: Shanghai Regression Lines

The scatterplot above with the Linear and LOESS regression lines is showing a positive relationship between the number of HICI/Citations and the number of Research/Publications produced in the Shanghai Top Ten dataset.



RMSE for Shanghai Publications

[1] 9.639154

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.639	9.639	9.639	9.639	9.639	9.639

Coefficient Data for Shanghai Publications

	2.5 %	97.5 %
(Intercept)	27.9254871	28.7399466
Highly Cited Researchers	0.5928685	0.6304384

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.33	0.2077	136.4	0
Highly Cited Researchers	0.6117	0.009582	63.83	0

Fitting linear model: Publications ~ Highly Cited Researchers

Observations	Residual Std. Error	R^2	Adjusted R^2
4895	9.641	0.4544	0.4543

Times Dataset Prediction

Research/Publications

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.432	1.001	6.424	1.571e-10
Citations	0.4839	0.01537	31.48	1.381e-184

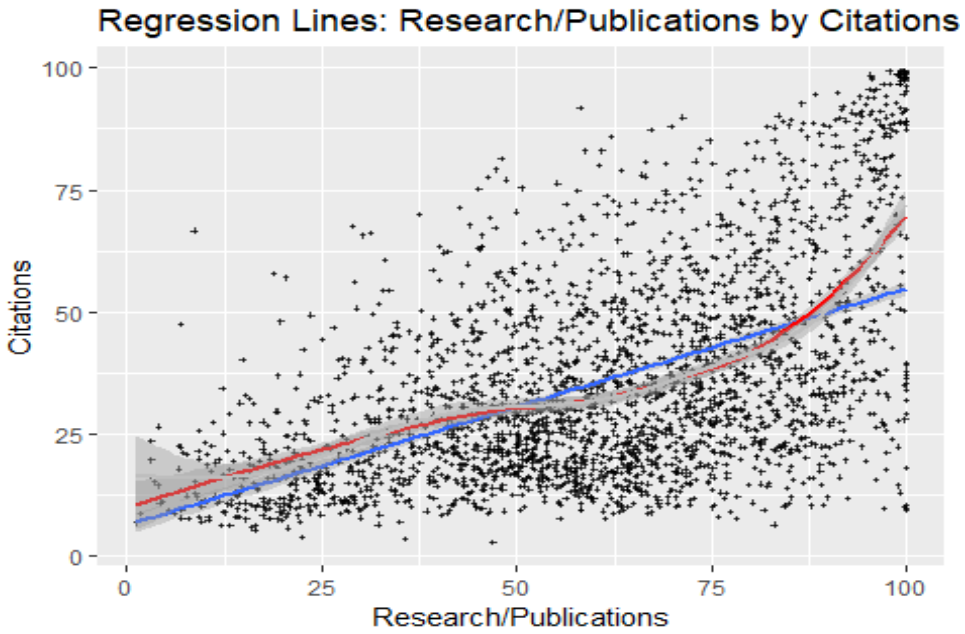
Fitting linear model: Research ~ Citations

Observations	Residual Std. Error	R^2	Adjusted R^2
2603	18.09	0.2759	0.2756

Citations

Graph 9: Times Regression Lines

The scatterplot above with the Linear and LOESS regression lines is showing a positive relationship between the number of Citations and the number of Research/Publications produced in the Times Top Ten dataset.



RMSE for Times Publications

[1] 18.08308

Coefficient Data for Times Publications

	2.5 %	97.5 %
(Intercept)	4.4689367	8.3956180
Citations	0.4537283	0.5140062

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.432	1.001	6.424	1.571e-10
Citations	0.4839	0.01537	31.48	1.381e-184

Fitting linear model: Research ~ Citations

Observations	Residual Std. Error	R^2	Adjusted R^2

2603	18.09	0.2759	0.2756

Summary RMSE data for World Ranking Dataset Publications

CWUR Prediction

169.4

Shanghai Prediction

9.639

Times Prediction

18.08

Comparing the root mean squared error for each of the three global ranking datasets the margin of error appears to be higher with the CWUR dataset at 169 errors versus the Shanghai dataset at ten errors and the Times dataset at 18 errors. Telling us that there is a higher possibility of errors that one may receive within the CWUR dataset when comparing Publications versus Citations.

Models and Methods

Implement Classifiers, Models, Predictors, etc. to solve data science problems. Investigate the learned model and support with visualizations. Report the accuracy and reliability of results with relevant supporting visuals.

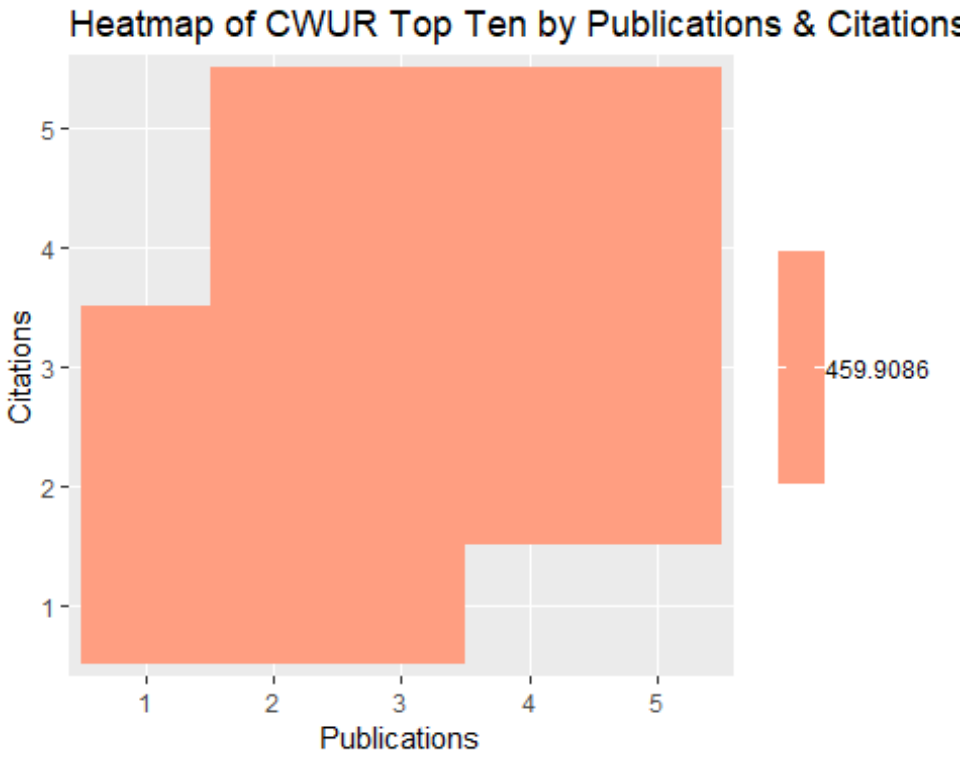
CWUR Top Ten Proportions, Plots & Heat Maps

CWUR top ten list by World Rank (WR) with the number affiliated with the rank 1-10. Compared to the total number of publications produced by the world rank top ten lists reflecting the number of publications and the percentage of those publications by world rank.

Top Ten by World Rank from the CWUR dataset along with the publication probabilities by Country and World Rank.

Country	cwurtt$World Rank	prob_pub
United Kingdom	3	17.32
United Kingdom	4	17.32
United Kingdom	5	17.32
United Kingdom	7	17.32
USA	1	17.32
USA	2	17.32
USA	3	17.32
USA	4	17.32
USA	5	17.32
USA	6	17.32
USA	7	17.32
USA	8	17.32
USA	9	17.32
USA	10	17.32

The heatmap below pulls from the CWUR top ten list by publications and citations. As the heatmap reflects, there is little variance to the gradient related to the probabilities around publications and citations.



Shanghai Top Ten Proportions, Plots & Heat Maps

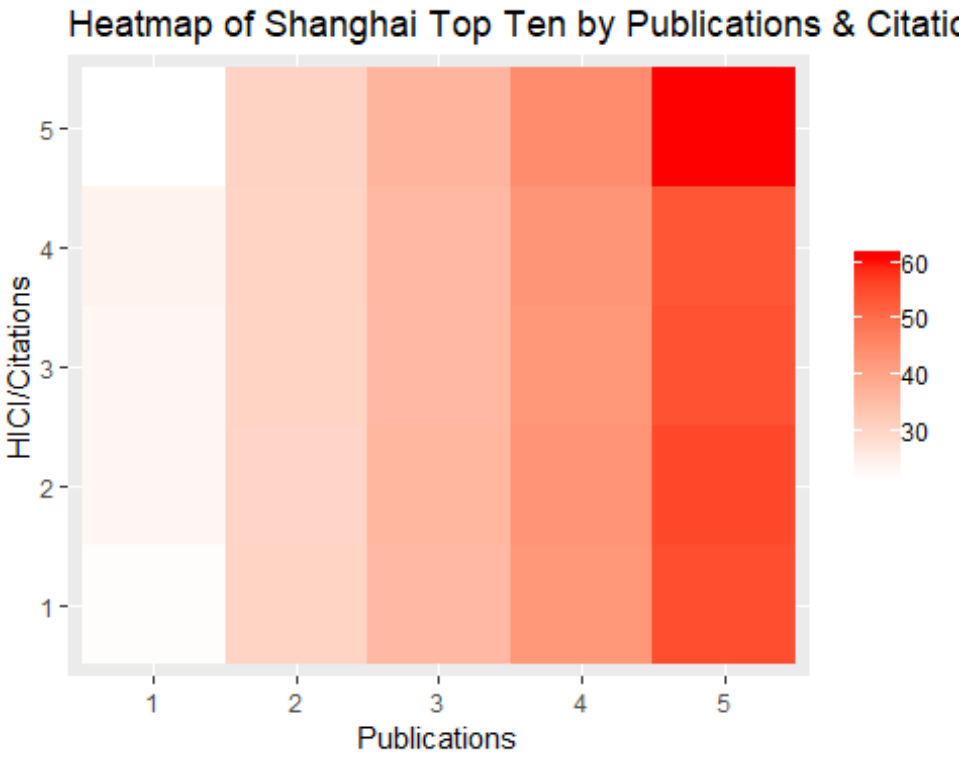
Shanghai top ten list by World Rank (WR) with the number affiliated with the rank 1-10. Compared to the total number of publications produced by the world rank top ten lists reflecting the number of publications and the percentage of those publications by world rank.

Top Ten by World Rank from the Shanghai dataset along with the publication probabilities by Country and World Rank.

Country	shangtt$World Rank	prob_pub
United Kingdom	2	70.75
United Kingdom	4	65.37
United Kingdom	5	65.97
United Kingdom	6	48.34
United Kingdom	7	69.65
United Kingdom	9	49.8
United Kingdom	10	67.96
United States of America	1	100
United States of America	2	69.99
United States of America	3	66.77
United States of America	4	67.24
United States of America	5	62.36
United States of America	6	43.3

United States of America	7	52.76
United States of America	8	58.17
United States of America	9	53.78

The heatmap below pulls from the Shanghai top ten lists by publications and citations. As the heatmap reflects, there is a correlation between the probabilities of more publications having a higher impact on HICI/Citations.



Times Top Ten Proportions, Plots & Heat Maps

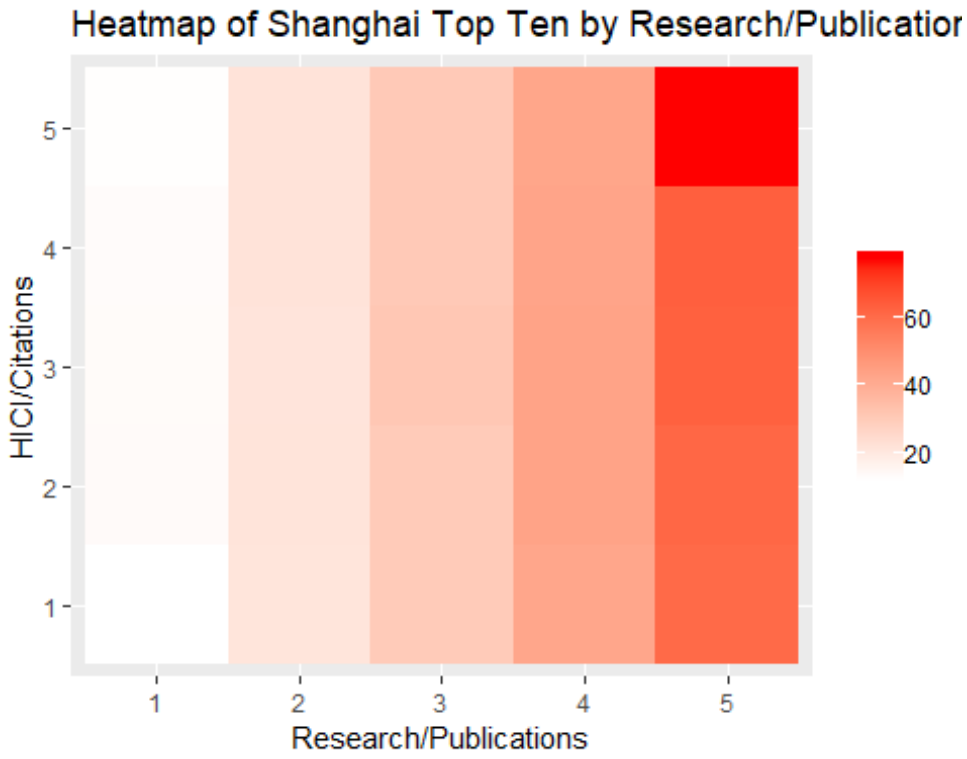
Times top ten list by World Rank (WR) with the number being affiliated with the rank 1-10. Compared to the total number of research/publications produced by the world rank top ten lists reflecting the number of publications and the percentage of those publications by world rank.

Top Ten by World Rank from the Times dataset along with the publication probabilities by Country and World Rank.

Country	World Rank	prob_pub
Switzerland	9	95
United Kingdom	2	98.5
United Kingdom	3	97.7
United Kingdom	4	96.65
United Kingdom	5	95.6
United Kingdom	6	94.07
United Kingdom	7	95.45
United Kingdom	8	89.37
United Kingdom	9	91.4

United Kingdom	10	88.1
United States of America	1	98.37
United States of America	2	98.37
United States of America	3	93.8
United States of America	4	97.55
United States of America	5	92.26
United States of America	6	96.05
United States of America	7	91.33
United States of America	8	97.83
United States of America	9	92.28
United States of America	10	92.73

The Times top ten list by research/publications and citations heatmap reflects a correlation between the probabilities of more publications having a higher impact on HICI/Citations.



In the comparison of the three global ranking systems specifically to the categories of World Rank, Publications and Citations, the results produced two standard-looking heatmaps - the Shanghai and Times datasets. A clear gradient from white to red shows when running the probabilities of research/publications and hici/citations for the top ten world rank in these datasets.

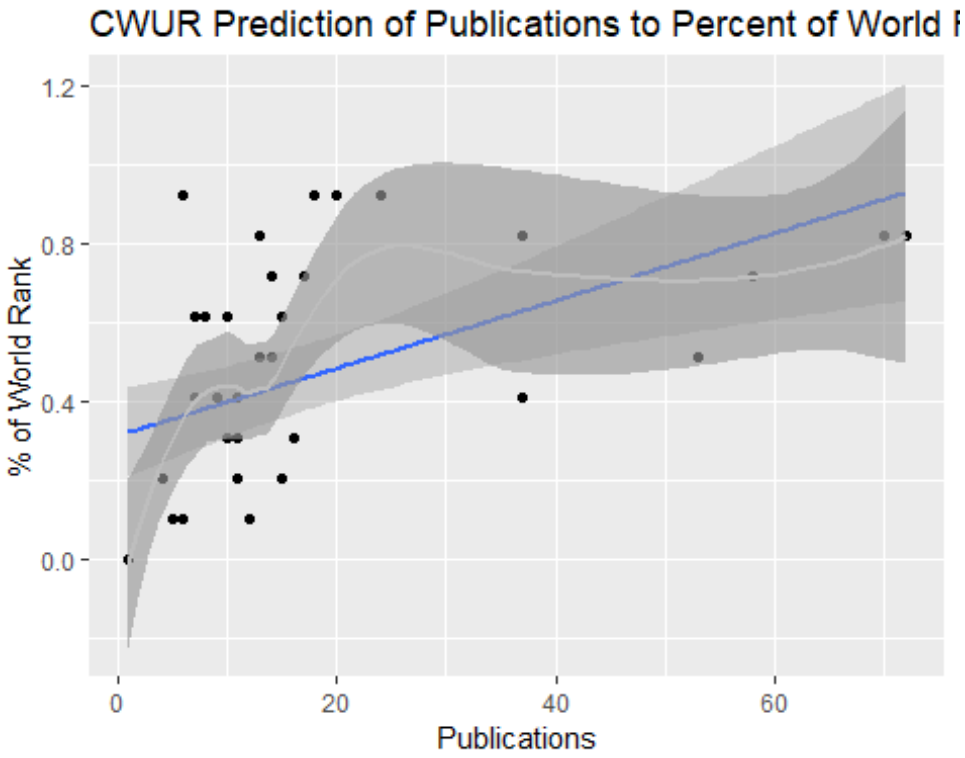
The remaining dataset, CWUR, does not show a successful standard heatmap with a gradient from white to red. There is a series of cross over with one solid color - more orange than red with no white. The results are showing little correlation to the probabilities of publications and citations.

Cross-Validation Models

CWUR Cross-Validation and Prediction

Graph 10: CWUR Prediction

The results above reflect a small linear pattern to the data, and most of the data points fall outside of the regression line. The LOESS model shows a curve to the regression as a better fit of the pattern.



CWUR Cross-Validation with Kfolds

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.096	4.294	1.652	0.09858
cwur$Publications	0.6829	0.01336	51.1	0
cwur$Citations	0.3348	0.01535	21.81	1.357e-95

Fitting linear model: cmod_formula

Observations	Residual Std. Error	R^2	Adjusted R^2
2200	106.2	0.8783	0.8782

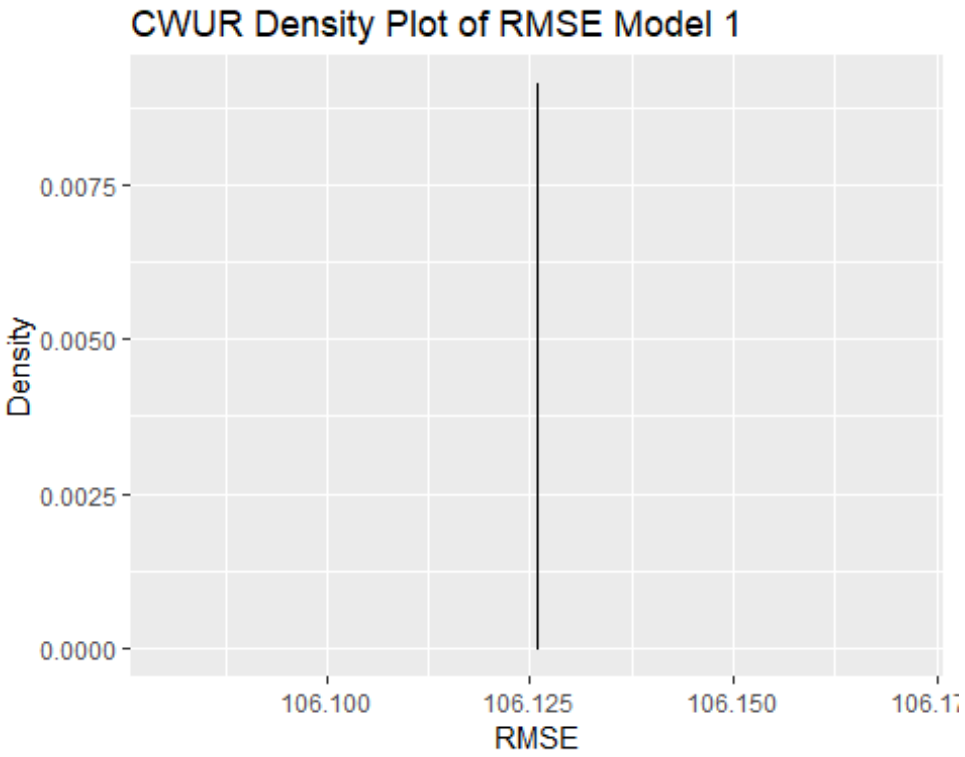
Conversion of the CWUR dataset into Tibbles and apply the model to test dataset to obtain the RMSE.

The resulting dataset includes the id for the cross-validation and the RMSE. We can summarize and plot this new data frame to see what our likely range of RMSE happens to be.

Graph 11: CWUR Density Plot for RMSE Model 1

Within this density plot, no significant results are coming from the cross-validation. The RMSE is the same for all categories within the model.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
106.1	106.1	106.1	106.1	106.1	106.1



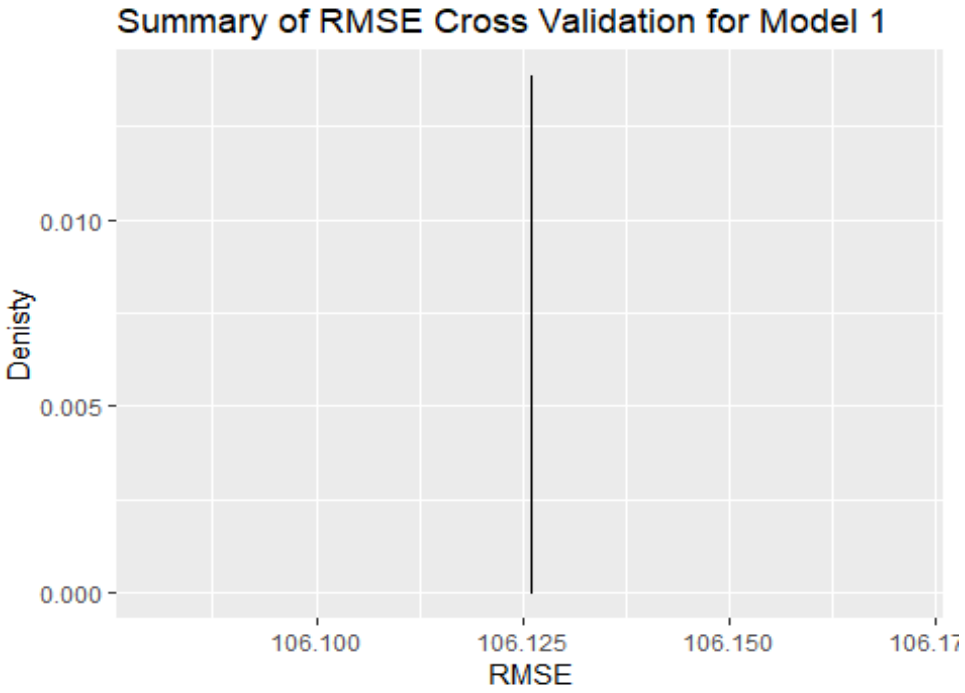
Full Cross-Validation: Random Partition

CWUR RMSE Cross-Validation for Model 1

Graph 12: CWUR Density Plot for Cross-Validation of RMSE Model 1

Within this density plot, no significant results are coming from the cross-validation. The RMSE is the same for all categories within the model.

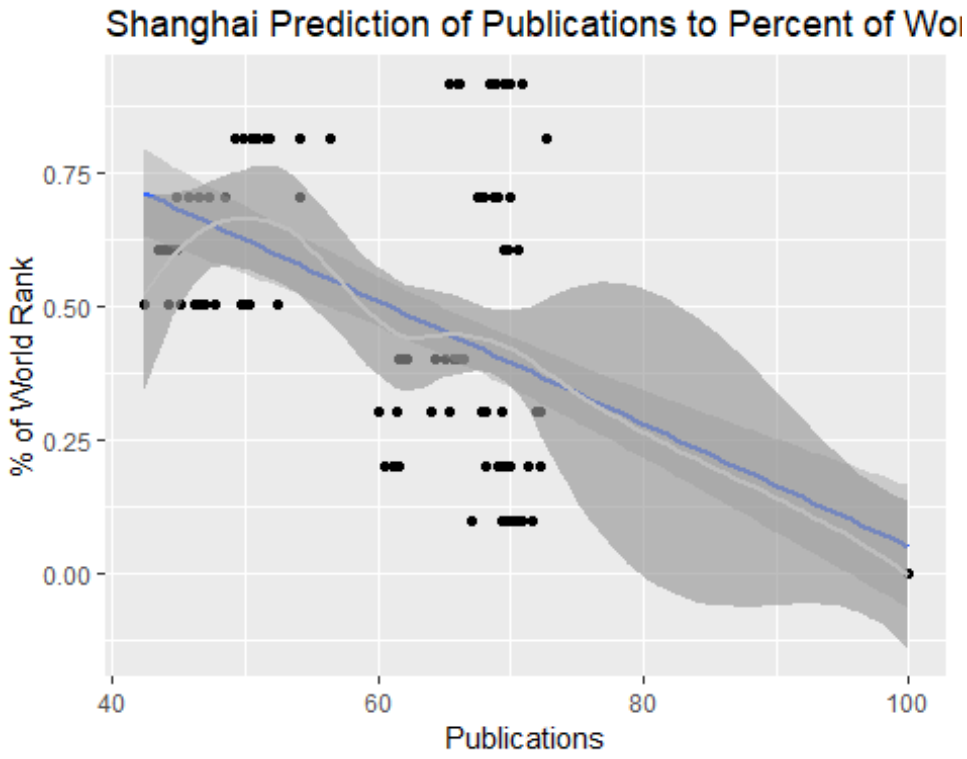
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
106.1	106.1	106.1	106.1	106.1	106.1



Shanghai Cross-Validation and Prediction

Graph 13: Shanghai Prediction

The results above reflect a clustered negative linear pattern to the data, and most of the data points fall outside of the regression line. The LOESS model shows a curve to the regression as a better fit of the pattern.



Shanghai Cross-Validation with Kfolds

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	112.8	2.391	47.19	1.932e-266
Publications	-0.3496	0.05649	-6.19	8.479e-10
Highly Cited Researchers	-1.235	0.04307	-28.69	1.606e-135

Fitting linear model: smod_formula

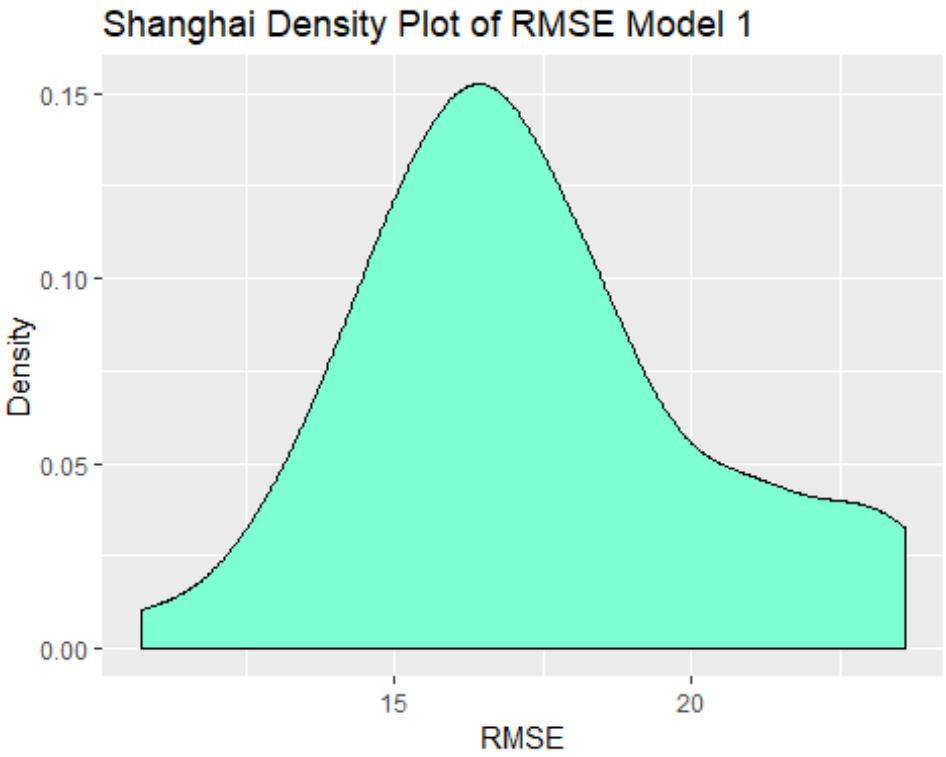
Observations	Residual Std. Error	R^2	Adjusted R^2
1101	17.3	0.6384	0.6377

Shanghai training dataset to Tibbles and then apply the model to test dataset to obtain the RMSE.

The resulting dataset includes the id for the cross-validation and the RMSE. We can summarize and plot this new data frame to see what our likely range of RMSE happens to be.

Graph 14: Shanghai Density Plot for RMSE Model 1

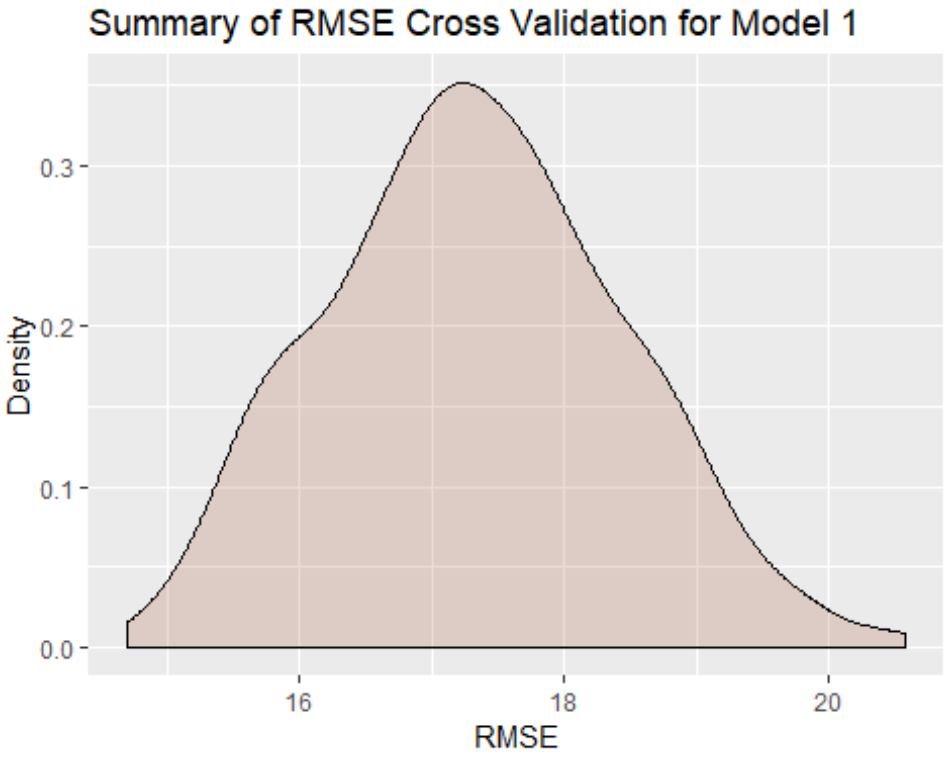
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.74	15.27	16.78	17.27	18.60	23.61



Full Cross-Validation: Random Partition

Shanghai RMSE Cross-Validation of Model 1

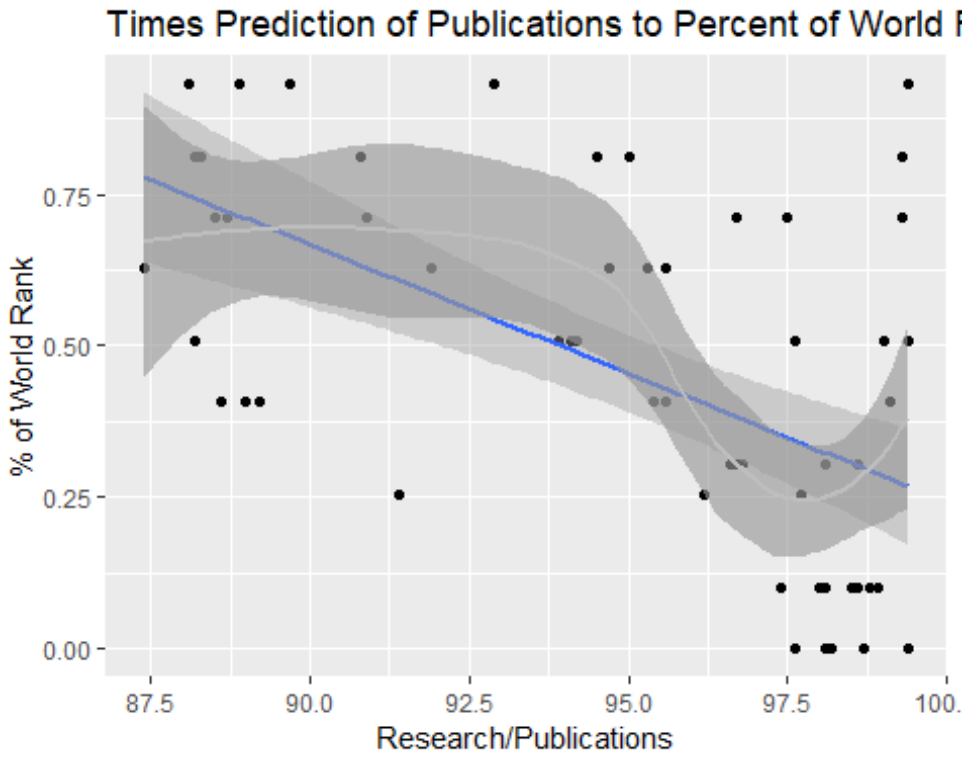
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.7	16.55	17.29	17.32	18.08	20.57



Times Cross-Validation and Prediction

Graph 15: Times Prediction

The results above reflect a small linear pattern to the data, and most of the data points fall outside of the regression line. The LOESS model shows a curve to the regression as a better fit of the pattern.



Times Cross-Validation with Kfolds

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	322.6	3.74	86.24	0
times$Research	-2.23	0.03625	-61.51	0
times$Citations	-1.375	0.04588	-29.98	8.477e-148

Fitting linear model: tmod_formula

Observations	Residual Std. Error	R^2	Adjusted R^2
1201	24.24	0.8234	0.8231

Times training dataset converted to Tibbles and then applying the model to the test dataset to obtain the RMSE.

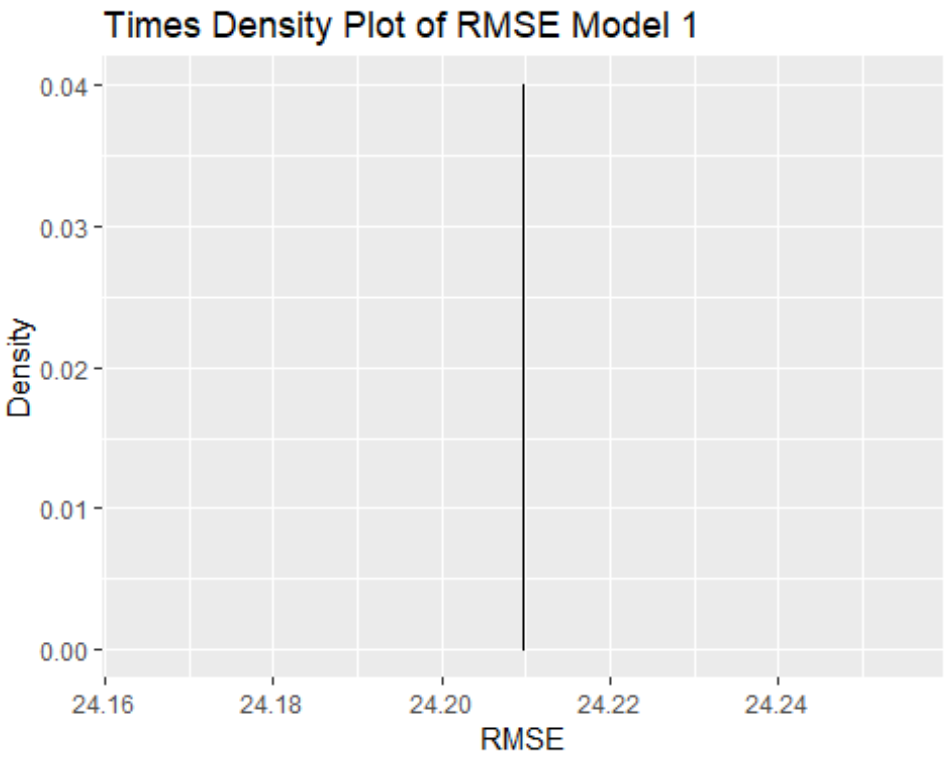
The resulting dataset includes the id for the cross-validation and the RMSE. We can summarize and plot this new data frame to see what our likely range of RMSE happens to be.

Graph 16: Times Density Plot for RMSE Model 1

Within this density plot, no significant results are coming from the cross-validation. The RMSE is the same for all categories within the model.

Min. 1st Qu. Median Mean 3rd Qu. Max.

24.21 24.21 24.21 24.21 24.21 24.21

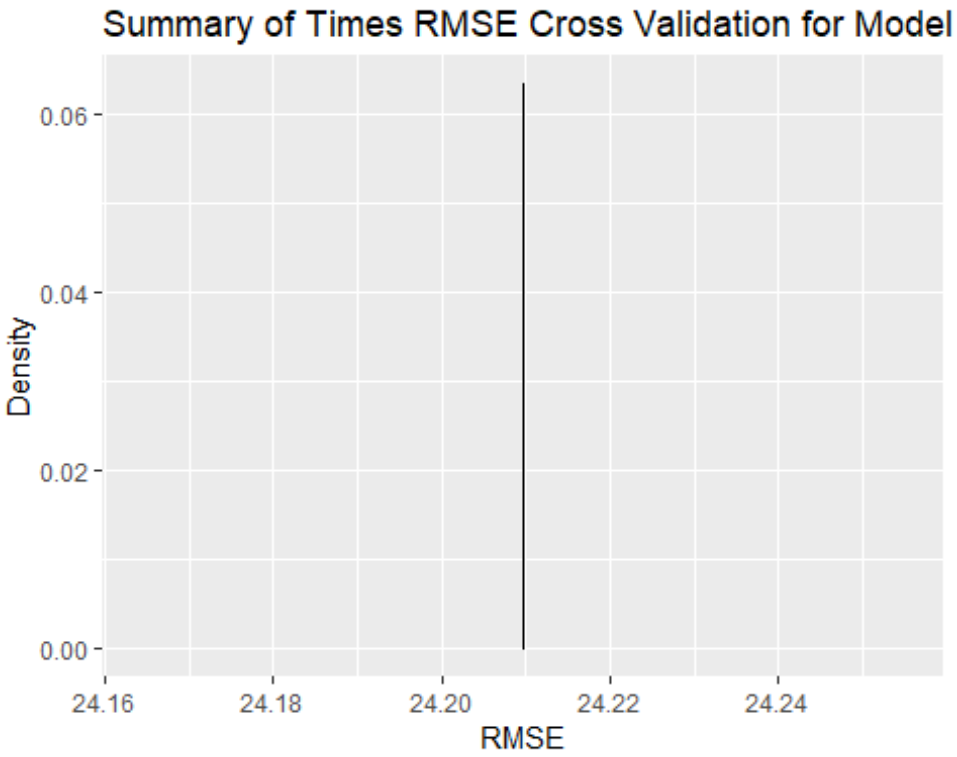


Full Cross-Validation: Random Partition

Times RMSE Cross-Validation for Model 1

Within this density plot, no significant results are coming from the cross-validation. The RMSE is the same for all categories within the model.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
24.21	24.21	24.21	24.21	24.21	24.21



Concluding Remarks:

Based on this analysis, we see the following findings:
- The United Kingdom and United States of America are in the top ten rankings for all three world ranking datasets. Switzerland is also reflected in the Times dataset only.

Regression data for overall publications and citations shows the following:
- Citations referenced in the CWUR dataset reflect a mid-way of 400 citations as the average.
- Citations referenced in the Shanghai dataset reflect a mid-way or peak in the 10-12 average.
- Citations referenced in the Times dataset reflect a mid-way 50-75 citations, approximately 62.5 citations.

- When running a Regression Line analysis looking at the Linear and LOESS regression lines, all three datasets (CWUR, Shanghai and Times) reflect a positive relationship between the number of Citations and the number of Publications produced.

- When performing a Heat map analysis on all three datasets the CWUR dataset reflected little variance in the gradient related to the probabilities around publications and citations. However, both the Shanghai and Times datasets reflected a correlation between the probabilities of more publications having a higher impact on HICI/Citations.

Cross-Validation & Predictions:
- CWUR Prediction reflected a small linear pattern to the data with most of the data points falling outside the regression line. The pattern also reflected a positive relationship. - Shanghai Prediction reflected a clustered negative linear pattern to the data and most of the data points fell outside the regression line.
- Times Prediction reflected a small linear pattern to the data with most of the data points falling outside the regression line. The patter also reflected a negative relationship.

In summary when running various statistical models, linear regression, cross-validation and prediction, etc… we see a small correlation between the number of publications and citations produced within the top ten institutions within the World Ranking datasets. The Regression data and charts appeared to show

the best visual representation of the data and its correlation. This analysis was limited to only two categories versus three to five additional categorical options that were reflected in each dataset.

As a result, additional analysis can be performed pulling in additional datafields to assist with determining additional causes which could influence prediction . Areas such as employment, awards and recognition (Nobel Laureates, Field Medalists) to teaching/quality of faculty, industry income levels and types of journals may affect the results reflected in this paper.

References:

Grolemund, G., & Wickham, H. (2017). R for Data Science (1st ed.). Sebastopol, CA: O'Rielly Media, Inc. Retrieved from https://r4ds.had.co.nz/

Methodology | CWUR | Center for World University Rankings. (2012). Retrieved May 20, 2019, from https://cwur.org/methodology/world-university-rankings.php

Ranking Methodology of Academic Ranking of World Universities. (2015). Retrieved May 20, 2019, from http://www.shanghairanking.com/ARWU-Methodology-2015.html

World University Rankings 2015-2016 methodology. (2015). Retrieved May 20, 2019, from https://www.timeshighereducation.com/news/ranking-methodology-2016

World University Rankings DataSet. (2016). Retrieved May 20, 2019, from https://www.kaggle.com/mylesoneill/world-university-rankings