Progress Report 3

Kelley Brundage

July 10, 2019

CWUR Dataset

The table below reflects the total number of times each country in the CWUR dataset is referenced.

Times Dataset

The table below reflects the total number of times each country in the Times dataset is referenced.

Shanghai Dataset

The table below reflects the total number of times each country in the Shanghai dataset is referenced.

Progress Report 3 - Models and Methods

World University Rankings

Synopsis of Problem & Approach from Progress Report 1 & 2

Compare the three global ranking systems to the amount of faculty, publications/research at each institution per ranking system by approaching each dataset with an analytical and statistical viewpoint.

Analyze the dataset specific to the area of research/academic and if common challenges that exist with all ranking systems exist. Problems such as making the correction for institutional size, differences between average and extreme, defining the institutions, measurement of time frame, credit allocation, excellency factors as well as adjustment for scientific fields or types of research.

Columns in each dataset that will be used to compare and analysze the 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 indica
```

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 your proposed problem:

Continued review and analysis of the datasets has led me to identify a series of common fields within the three primary ranking system datasets: CWUR, Shanghai and Times. All three hold common columns such as the World Rank, Institution/University Name, Publications/Research and Citations. My 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.

Models and Methods

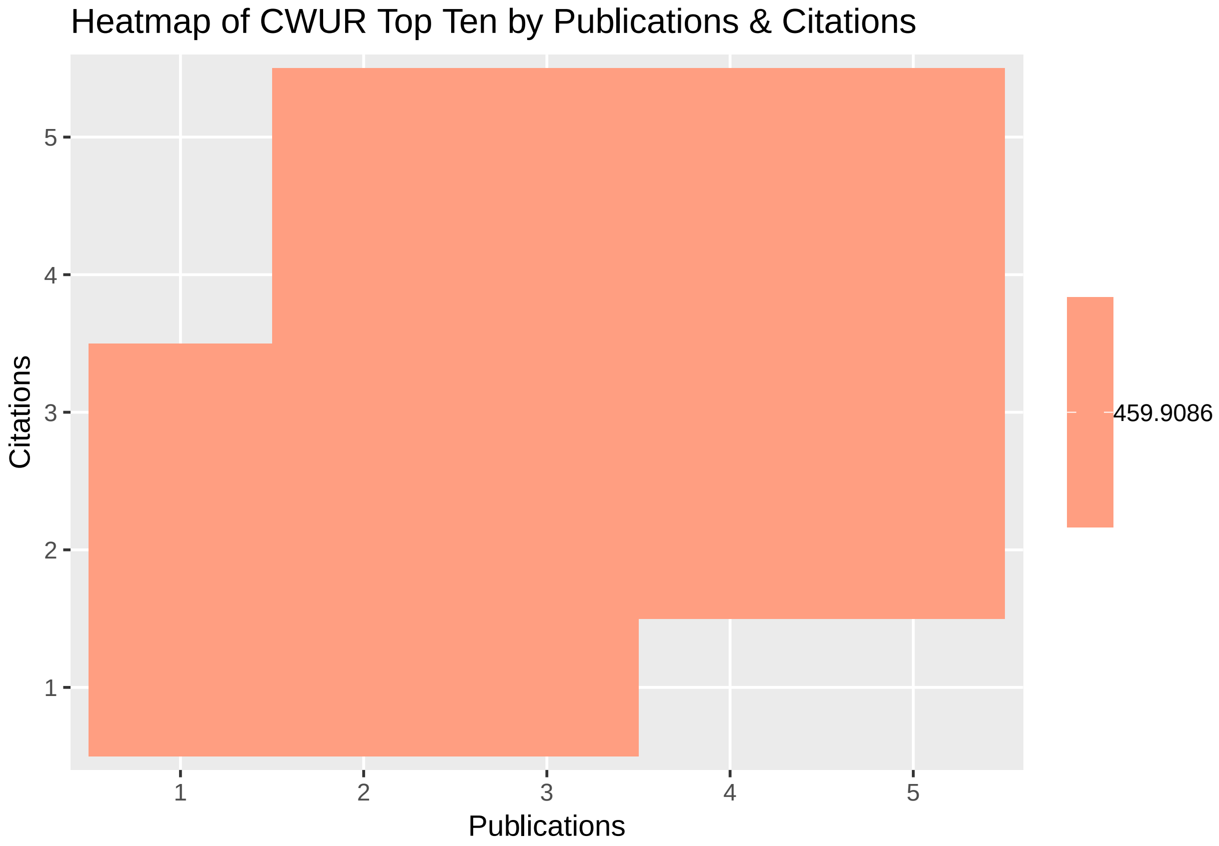
##Implement Classifiers, Models, Predictors, etc. to solve data science problems. Investigate the learned model and support with visualizations. Report the accuracy and reliability of results with relevant supporting visuals.

CWUR Top Ten Proportions, Plots & Heat Maps

The dataset below reflects the CWUR top ten list by World Rank (WR) with the number being affiliated with the rank 1-10. This is compared to the total number of publications produced by the world rank top ten list.

This chart reflects the Top Ten by World Rank from the CWUR dataset along with the publicatiom probabilities by Country and World Rank.

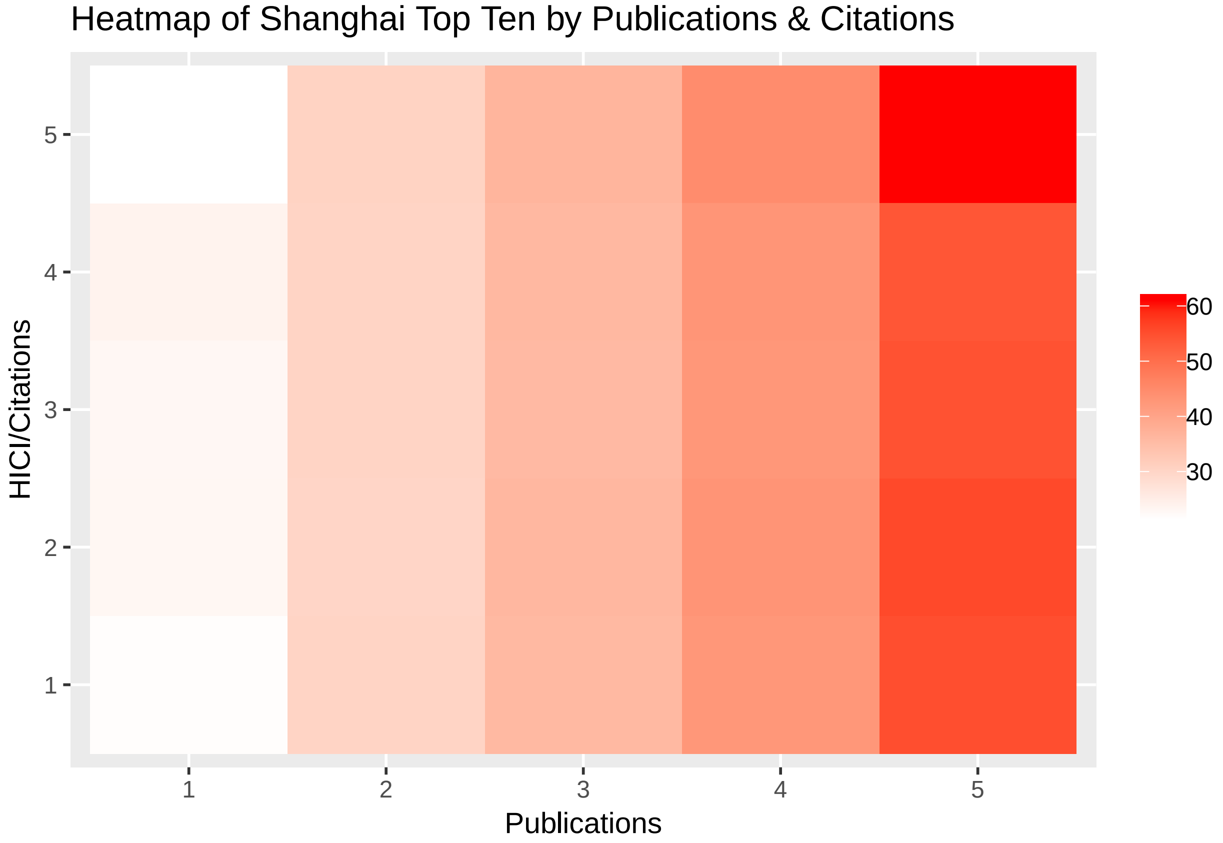
The heatmap below pulls from the CWUR top ten list by publications and citations. As the heatmap reflects there is little variance to the gradiant related to the probabilities around publications and citations.



Shanghai Top Ten Proportions, Plots & Heat Maps

This chart reflects the Top Ten by World Rank from the Shanghai dataset along with the publicatiom probabilities by Country and World Rank.

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

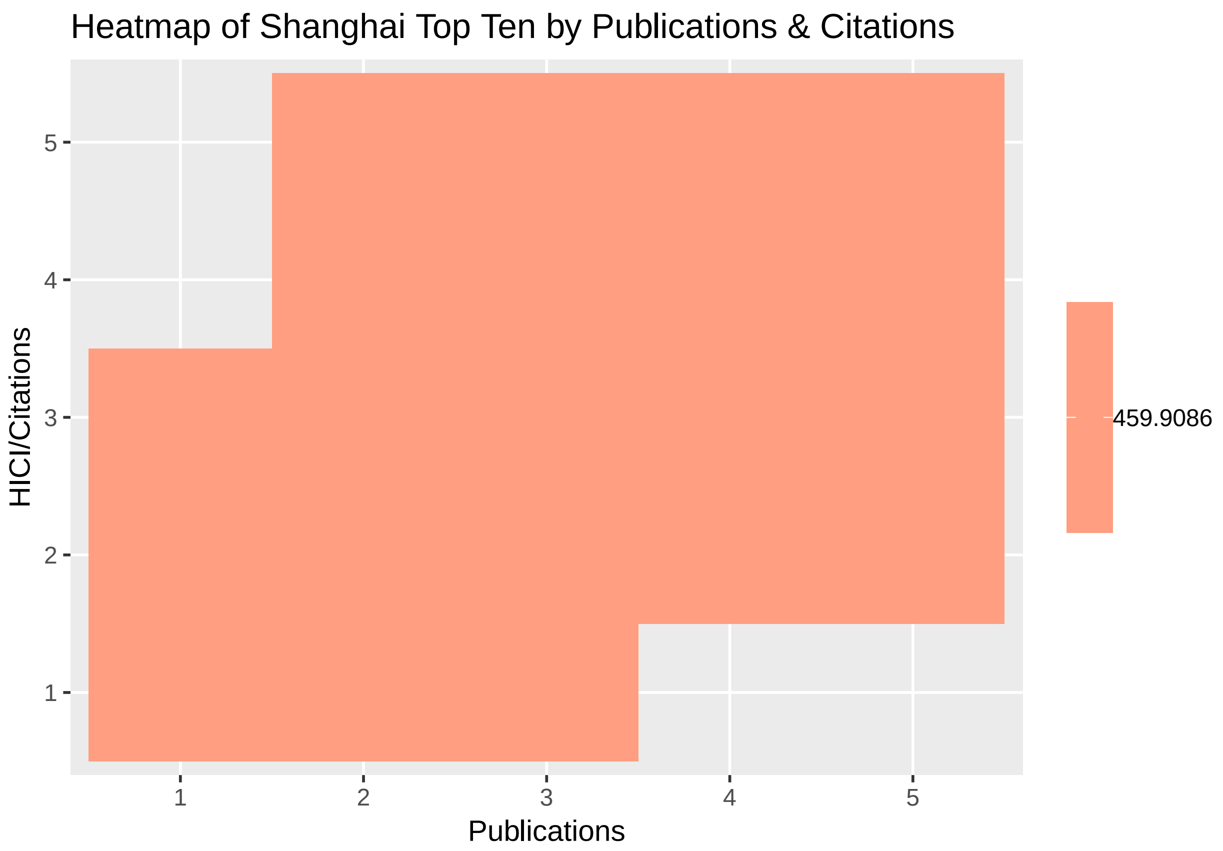


Times Top Ten Proportions, Plots & Heat Maps

This chart reflects the Top Ten by World Rank from the Times dataset along with the publicatiom probabilities by Country and World Rank.

Then we'll create a summary dataset that shows the probabilitie of the outcome across all of the combined categories of the two independent variables.

The heatmap below pulls from the Times dataset top ten list by publications and citations. As the heatmap reflects there is little correlation between the probabilities of publications and citations.



Summary

As I continue to compare the three global ranking systems specifically to the categories of World Rank, Publications and Citations there have been unique challenges to the probability analysis. The one dataset that ran through the probabilities and produced a standard looking heatmap was the Shanghai dataset. There is a clear gradient from white to red when running the probabilities of publications and citations for the top ten world rank in this dataset.

The remaining two datasets, CWUR and Times do not show a successful standard heatmap with a gradient from white to red. Instead, both show a series of cross over with one solid color - more orange than red with no white. These results are showing little correlation to the probabilities of publications and citations.