LLO 8200 Data Science Final Report - Kelley Brundage

Kelley Brundage

8/4/2019

```
##This code allows the Knit function to still work even with errors
knitr::opts_chunk$set(echo=TRUE,warning=FALSE,results='hide',include=TRUE,messages=TRUE)

#We always start with a standard set of setup commands by loading the correct libraries.

##Load libraries in order to successfully run the code below - the suppressMessages coding will stop the
install.packages information, etc.. from coming up in the Console and showing you what has run.

suppressMessages(library(caret)) #Misc functions for training and plotting classification and regression models.
suppressMessages(library(dplyr)) #able to select, filter, organize, and manipulate data stored within an R data
frame
suppressMessages(library(evaluate)) #Parsing and Evaluation Tools that Provide More Details than the Default
suppressMessages(library(forcats)) #Tools for Working with Categorical Variables (Factors)
suppressMessages(library(formatR)) #Provides a function tidy_source() to format R source code.
suppressMessages(library(ggplot2)) #A system for 'declaratively' creating graphics, based on "The Grammar of
Graphics".
suppressMessages(library(haven)) #Import foreign statistical formats into R via the embedded 'ReadStat' C
library
suppressMessages(library(knitr))#General-Purpose Package for Dynamic Report Generation in R
suppressMessages(library(lubridate)) #Functions to work with date-times and time-spans: fast and user friendly
parsing of date-time data, extraction and updating of components of a date-time
suppressMessages(library(ModelMetrics)) #Collection of metrics for evaluating models written in C++ using
'Rcpp'.
suppressMessages(library(modelr)) #Functions for modelling that help you seamlessly integrate modelling into a
pipeline of data manipulation and visualisation.
suppressMessages(library(pander))#provide a minimal and easy tool for rendering R objects
 panderOptions('table.style', "multiline")
 panderOptions('table.alignment.default',function(df)ifelse(sapply(as.data.frame(df),
                                    is.numeric),'right','left'))

suppressMessages(library(readxl)) #reads in Excel Files
suppressMessages(library(rvest)) #scraping websites
suppressMessages(library(tibble)) #Provides a 'tbl_df' class (the 'tibble') that provides stricter checking and better
formatting than the traditional data frame.
suppressMessages(library(tidyverse)) #set of packages that work in harmony because they share common data
representations and 'API' design

## Warning: package 'tidyverse' was built under R version 3.6.1

##Define My PDF setup - This code does not show in the final document but will assist with definining the margin
cutoff point and wraps the text to the next line.
knitr::opts_chunk$set(fig.path = "Figs/", results='hide', tidy.opts=list(width.cutoff=60))
my_pdf = function(file,width,height)
 {pdf(file, width=width, height=height,pointsize=12)}

##The code below will upload the World University Ranking Datasets obtained from Kaggle.com
```

```
##The next set of files are the three World Ranking system files
library(readxl)
cwur <- read_excel("~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/cwurData.xlsx")

save(cwur, file = "cwurData.xlsx") #save as excel file name cwurData.xlsx

library(readxl)
shanghai <- read_csv("~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv")

## Parsed with column specification:
## cols(
##   world_rank = col_double(),
##   university_name = col_character(),
##   country = col_character(),
##   national_rank = col_double(),
##   total_score = col_double(),
##   alumni = col_double(),
##   award = col_double(),
##   hici = col_double(),
##   ns = col_double(),
##   pub = col_double(),
##   pcp = col_double(),
##   year = col_double()
## )

## Warning: 5609 parsing failures.
##  row          col             expected actual                                                                                                      file
## 1103 world_rank    no trailing characters   -150 '~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv'
## 1103 national_rank no trailing characters   -69  '~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv'
## 1104 world_rank    no trailing characters   -150 '~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv'
## 1105 world_rank    no trailing characters   -150 '~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv'
## 1106 world_rank    no trailing characters   -150 '~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/shanghaiData.csv'
## .... ............. .................... ...... ...........................................................................................................
## See problems(...) for more details.

save(shanghai, file = "shanghaiData.csv") #save as csv file name shanghaiData.csv

library(readxl)
times <- read_excel("~/School/EdD program - Vanderbilt/01 - Classes/LLO 8200 - Intro to Data Science/Final Project/datafiles/timesData.xlsx")

save(times, file = "timesData.xlsx") #save as excel file name timesData.xlsx

#the code below will show us the type of data coded in each column on each dataset
sapply(cwur,class)

sapply(times,class)

sapply(shanghai,class)
```

```
#the code above showed us an issue with the times and shanghai dataset so the code below will convert the data type for the times dataset

snames <- c(1,4:12)
shanghai[,snames] <- lapply(shanghai[snames], as.numeric)

sname <- c(2:3)
shanghai[,sname] <- lapply(shanghai[sname], as.character)
str(shanghai)

tnames <- c(1,4:14)
times[,tnames] <- lapply(times[tnames], as.numeric)

tname <- c(2:3)
times[,tname] <- lapply(times[tname], as.character)

str(times)

## This code will clean up the column labels to align directly with the variable definitions for the columns as well as rename the columns into a readable format that describes the column data.

names(cwur)<-c("World Rank",
               "Institution Name",
               "Country",
               "National Rank",
               "Quality of Education",
               "Alumni Employment",
               "Quality of Faculty",
               "Publications",
               "Influence",
               "Citations",
               "Broad Impact",
               "Patents",
               "Score",
               "Year")

head(cwur)

## This code will clean up the column labels to align directly with the variable definitions for the columns as well as rename the columns into a readable format that describes the column data.

names(shanghai)<-c("World Rank",
               "University Name",
               "Country",
               "National Rank",
               "Total Score",
               "Alumni",
               "Award",
               "Highly Cited Researchers",
               "Nature & Science Pubs",
               "Publications",
               "Per Capita Performance",
               "Year")
```

```
head(shanghai)

## This code will clean up the column labels to align directly with the variable definitions for the columns as well as rename the columns into a readable format that describes the column data.

names(times)<-c("World Rank",
                "University Name",
                "Country",
                "Teaching",
                "International",
                "Research",
                "Citations",
                "Income",
                "Total Score",
                "Number of Students",
                "Student/Staff Ratio",
                "International Students",
                "Female/Male Ratio",
                "Year")

head(times)
```

CWUR Dataset

```
##Only reflect the top 10 institutions listed by World Rank in the CWUR Dataset

cwurten <- cwur[order(cwur$`World Rank`),]

head(cwurten)

##This code will pull out rows one through 40 from the reorder dataset above which are the institutions in the top ten by world rank

cwurtt <- cwurten[1:40,]

head(cwurten)

##shows the number of times each country is referenced in the cwur dataset

ccount <- cwur%>%
  count(Country)%>%
  arrange(-n)

head(ccount)
```

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

```
colnames(ccount) <- c("Country", "Total References")

head(ccount)
```

Shanghai Dataset

```
##Only reflect the top 10 institutions listed by World Rank in the Shanghai Dataset

shangten <- shanghai[order(shanghai$`World Rank`),]
```

```
head(shangten)

##Code that pulls out rows 1-110 into a new dataset which represents the top ten institutions by world rank in the Shanghai dataset

shangtt <- shangten[1:110,]

head(shangten)
```

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

```
##shows the number of times each country is referenced in the Shanghai dataset

scount <- shanghai%>%
 count(Country)%>%
 arrange(-n)

head(scount)
```

A tibble: 6 x 2

Country	n
1 United States of America	1586
2 United Kingdom	383
3 Germany	374
4 Japan	258
5 Canada	217
6 France	203

Times Dataset

```
##Only reflect the top 10 institutions listed by World Rank in the CWUR Dataset

timesten <- times[order(times$`World Rank`),]

head(timesten)

##Code that pulls out rows 1-60 from the Times dataset reflecting the top ten institutions by world rank

timestt <- timesten[1:60,]

head(timesten)

##shows the number of times each country is referenced in the Times dataset

tcount <- times%>%
 count(Country)%>%
 arrange(-n)

head(tcount)
```

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

```
colnames(tcount) <- c("Country", "Total References")

head(tcount)
```

A tibble: 6 x 2

Country Total References 1 United States of America 659 2 United Kingdom 300 3 Germany 152 4 Australia 117 5 Canada 108 6 Japan 98

```
#The code below will apply some baseline Tidy-Data principles which will allow future use of the clean data.

#1. Each variable forms a column
#2. Each observation forms a row
#3. Each type of observational units forms a table

is.data.frame(cwur)
is.tibble(cwur)
is_tibble(cwur)
typeof(cwur)

is.data.frame(shanghai)
is.tibble(shanghai)
is_tibble(shanghai)
typeof(shanghai)

is.data.frame(times)
is.tibble(times)
is_tibble(times)
typeof(times)
```

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 Rankins 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

```
##Only reflect the top 10 institutions listed by World Rank in the CWUR Dataset

cwurten <- cwur[order(cwur$`World Rank`),]
```

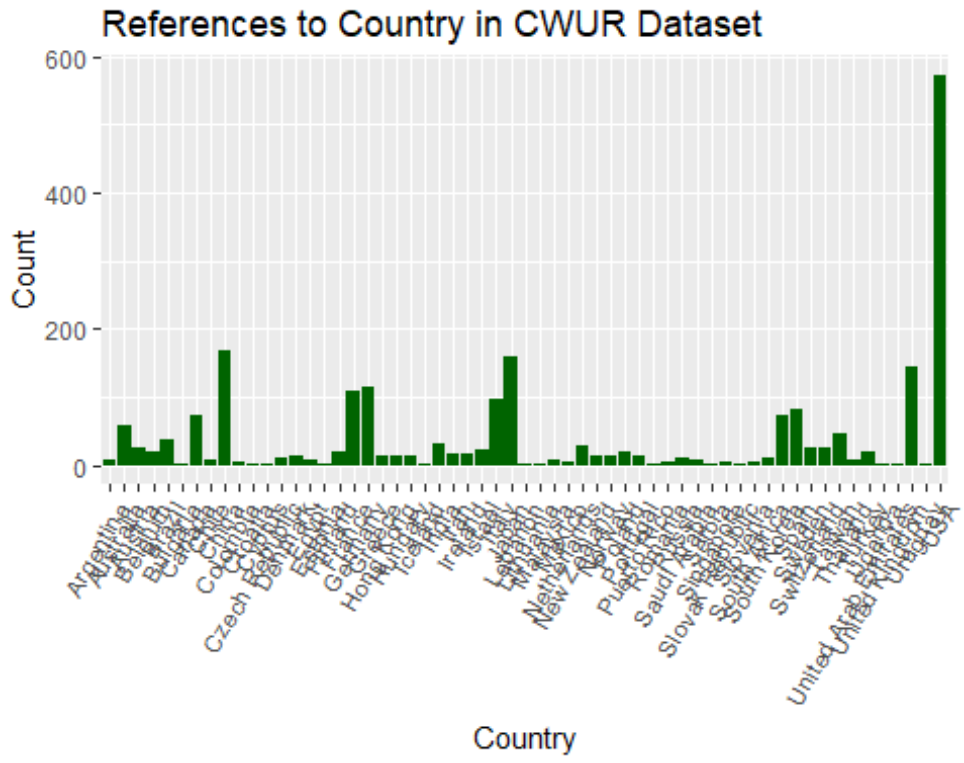
```
head(cwurten)

##this code pulls rows 1-40 out of the cwurten dataset and places it in a new dataset

cwurtt <- cwurten[1:40,]

head(cwurten)

##shows the number of times each country is referenced in the cwur dataset

ccount <- cwur%>%
 count(Country)%>%
 arrange(-n)

head(ccount)
```

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.

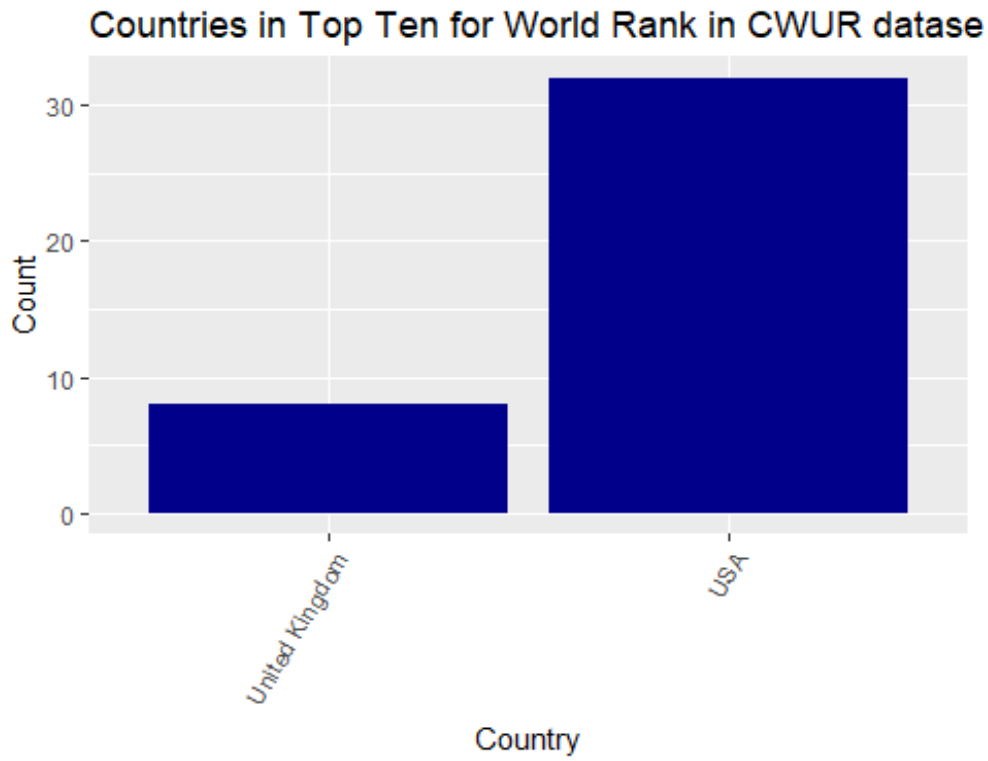
```
ggten<-ggplot(cwurten,aes(x=Country))+
 geom_histogram(stat = "count",binwidth = 20,fill="darkgreen")+
 theme(axis.text.x = element_text(angle = 60, hjust = 1))+
 labs(title ="References to Country in CWUR Dataset", x="Country", y="Count" )
ggten
```



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.

```
ggtt <- ggplot(cwurtt,aes(x=Country))+
 geom_histogram(stat="count",binwidth = 20,fill="darkblue")+
 theme(axis.text.x = element_text(angle = 60, hjust = 1))+
 labs(title="Countries in Top Ten for World Rank in CWUR dataset", x="Country", y="Count")
ggtt
```




```
##code groups and arranges the data in the cwur dataset by World Rank and then country

cwur%>%
 group_by(`World Rank`)%>%
 summarize(Country=n_distinct(Country))%>%
 arrange(desc(Country))
```

Shanghai Dataset

```
##Only reflect the top 10 institutions listed by World Rank in the CWUR Dataset

shangten <- shanghai[order(shanghai$`World Rank`),]

head(shangten)

##this code pulls rows 1-110 out of the cwurten dataset and places it in a new dataset

shangtt <- shangten[1:110,]

head(shangten)
```

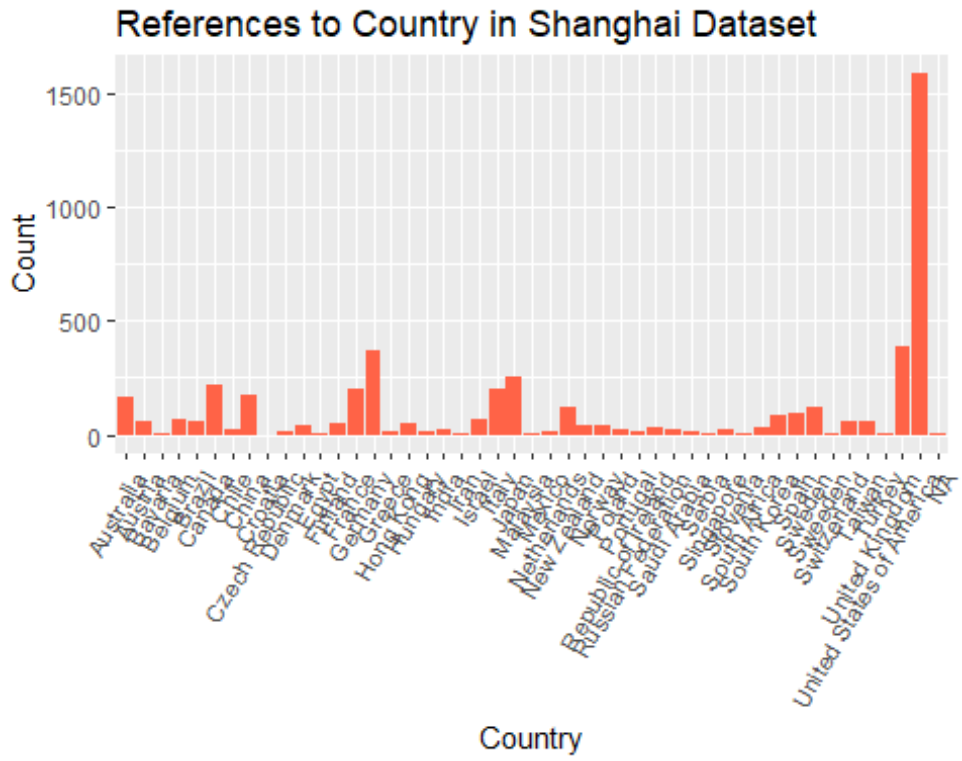
```
##shows the number of times each country is referenced in the cwur dataset

scount <- shanghai%>%
 count(Country)%>%
 arrange(-n)

head(scount)
```

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.

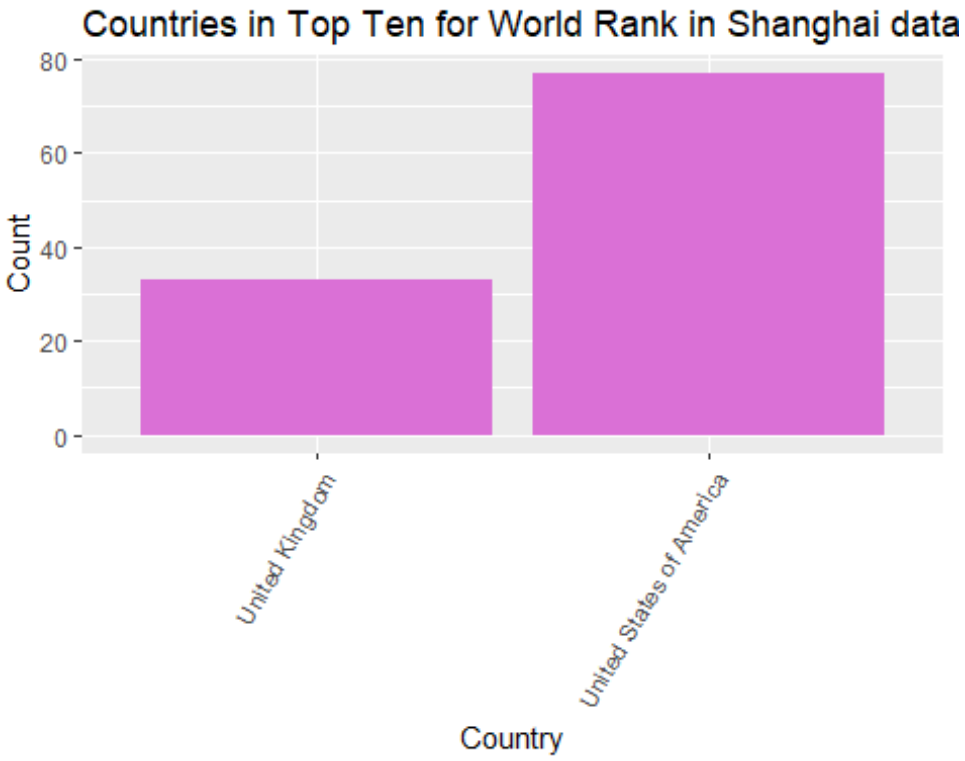
```
ggsten<-ggplot(shangten,aes(x=Country))+
 geom_histogram(stat = "count",binwidth = 20,fill="tomato")+
 theme(axis.text.x = element_text(angle = 60, hjust = 1))+
 labs(title ="References to Country in Shanghai Dataset", x="Country", y="Count" )
ggsten
```



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.

```
ggstt <- ggplot(shangtt,aes(x=Country))+
 geom_histogram(stat="count",binwidth = 20,fill="orchid")+
 theme(axis.text.x = element_text(angle = 60, hjust = 1))+
 labs(title="Countries in Top Ten for World Rank in Shanghai dataset", x="Country", y="Count")
ggstt
```




```
##code groups and arranges the data in the Shanghai dataset by World Rank and then country

shanghai%>%
 group_by(`World Rank`)%>%
 summarize(Country=n_distinct(Country))%>%
 arrange(desc(Country))
```

Times Dataset

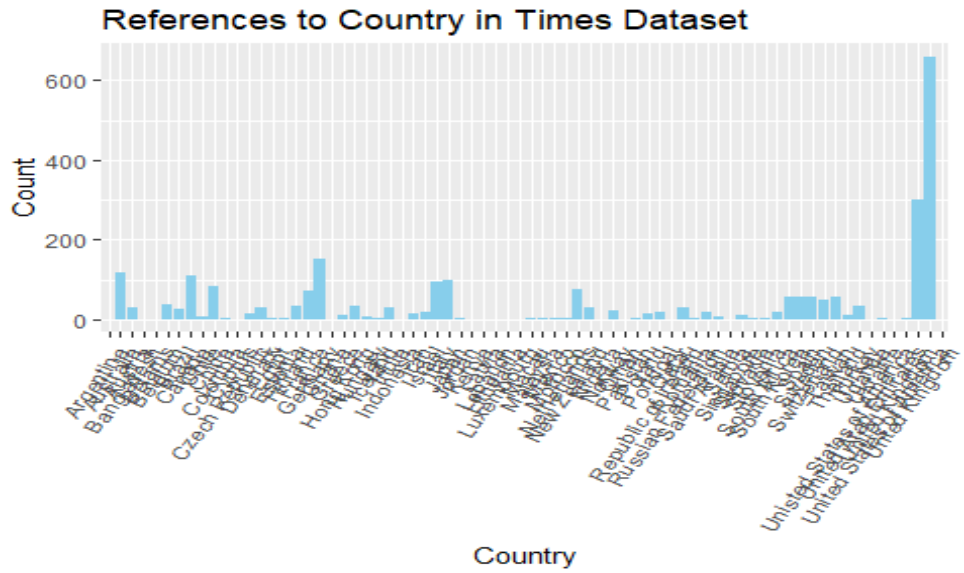
```
##Only reflect the top 10 institutions listed by World Rank in the CWUR Dataset

timesten <- times[order(times$`World Rank`),]


head(timesten)

##this code pulls rows 1-60 out of the cwurten dataset and places it in a new dataset

timestt <- timesten[1:60,]

head(timesten)

##shows the number of times each country is referenced in the cwur dataset

tcount <- times%>%
 count(Country)%>%
 arrange(-n)

head(tcount)
```

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.

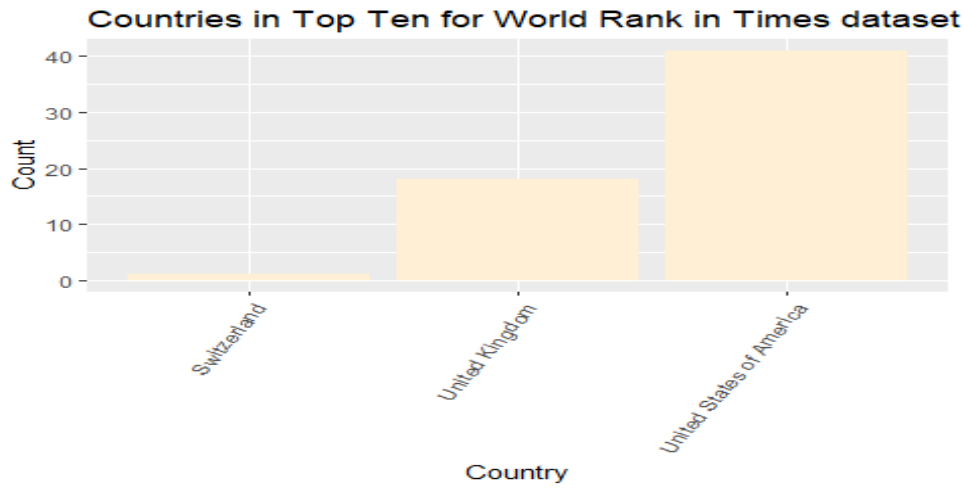
```
ggtten<-ggplot(timesten,aes(x=Country))+
  geom_histogram(stat = "count",binwidth = 20,fill="skyblue")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  labs(title ="References to Country in Times Dataset", x="Country", y="Count" )
ggtten
```



Graph #6:

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

```
ggttt <- ggplot(timestt,aes(x=Country))+
  geom_histogram(stat="count",binwidth = 20,fill="papayawhip")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  labs(title="Countries in Top Ten for World Rank in Times dataset", x="Country", y="Count")
ggttt
```



```
timesten%>%
 group_by(`World Rank`)%>%
 summarize(Country=n_distinct(Country))%>%
 arrange(Country)
```

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 Country.

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

```
##this code calculates the mean of the cwur top ten ordered list for publications

cwurp_mean <- cwurten%>%summarise(mean(Publications,na.rm = T))

kable(summary(cwurp_mean), show_level=T)
```

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

```
 #na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean Publications Percent within the CWUR Dataset

```
cwurten <- cwurten%>%mutate(publications_p=percent_rank(cwurten$Publications)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the publication number in the CWUR dataset

cwurpp_mean <- cwurten%>%summarise(mean(publications_p,na.rm = T))

kable(summary(cwurp_mean))
```

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

```
 #gives the mean of the publications in the CWUR dataset

#The code below shows Publications by Country in the CWUR dataset showing the mean, standard deviation, max and min.

desc_cwur <- cwurten%>%
 group_by(Country)%>%
 summarise(mean_pub = mean(publications_p),
        sd_pub = sd(publications_p),
        max_pub= max(publications_p),
        min_pub= min(publications_p))
```

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

```
#The code below shows which country has the maximum and minimum Publications inthe CWUR dataset

desc_cwur%>%
 filter(max(mean_pub) == mean_pub)
```

A tibble: 1 x 5

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

```
desc_cwur%>%
 filter(min(mean_pub) == mean_pub)
```

A tibble: 1 x 5

Country	mean_pub	sd_pub	max_pub	min_pub
1 Singapore	11.5	6.54	19.0	6.00

```
kable(summary(desc_cwur))
```

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

```
cwurc_mean <- cwurten%>%summarise(mean(cwurten$Citations,na.rm = T))

kable(summary(cwurc_mean))
```

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

```
#na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean Citations Percent within the CWUR Dataset

```
cwurten <- cwurten%>%mutate(citations_p=percent_rank(cwurten$Citations)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the publication number in the CWUR dataset

cwurcp_mean <- cwurten%>%summarise(mean(citations_p,na.rm = T))

kable(summary(cwurcp_mean))
```

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

```
 #gives the mean of the publications in the CWUR dataset

#The code below shows Citations by Country in the CWUR dataset showing the mean, standard deviation, max and min.

desc_cwur <- cwurten%>%
 group_by(Country)%>%
 summarise(mean_cit = mean(Citations),
        sd_cit = sd(Citations),
        max_cit= max(Citations),
        min_cit= min(Citations))
```

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

```
#The code below shows which country has the maximum and minimum Citations inthe CWUR dataset

desc_cwur%>%
 filter(max(mean_cit) == mean_cit)
```

A tibble: 3 x 5

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

```
desc_cwur%>%
 filter(min(mean_cit) == mean_cit)
```

A tibble: 1 x 5

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

```
kable(summary(desc_cwur))
```

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

```
spub_mean <- shangten%>%summarise(mean(shangten$Publications,na.rm = T))

kable(summary(spub_mean))
```

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

```
 #na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean Publications Percent within the Shanghi Dataset

```
shangten <- shangten%>%mutate(publications_p=percent_rank(shangten$Publications)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the publication number in the Shanghai dataset
```

```
spubp_mean <- shangten%>%summarise(mean(publications_p,na.rm = T))

kable(summary(spubp_mean))
```

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

```
 #gives the mean of the publications in the Shanghai dataset

#The code below shows Citations by National Rank in the Shanghai dataset showing the mean, standard deviation, max and min.

desc_shanghai <- shangten%>%
 group_by(Country)%>%
 summarise(mean_spub=mean(Publications),
        sd_spub=sd(Publications),
        max_spub=max(Publications),
        min_spub=min(Publications))
```

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

```
#The code below shows which country has the maximum and minimum Publications in the Shanghai dataset

#Max Publications
desc_shanghai%>%
 filter(max(mean_spub)==mean_spub)
```

A tibble: 0 x 5
... with 5 variables: Country , mean_spub , sd_spub , max_spub , min_spub

```
#Min Publications
desc_shanghai%>%
 filter(min(mean_spub)==mean_spub)
```

A tibble: 0 x 5
... with 5 variables: Country , mean_spub , sd_spub , max_spub , min_spub

```
kable(summary(desc_shanghai))
```

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

```
shici_mean <- shangten%>%summarise(mean(shangten$`Highly Cited Researchers`,na.rm = T))

kable(summary(shici_mean))
```

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

```
 #na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean HICI/Citations Percent within the Shanghai Dataset

```
shangten <- shangten%>%mutate(hici_p=percent_rank(shangten$`Highly Cited Researchers`)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the publication number in the Shanghai dataset

shicip_mean <- shangten%>%summarise(mean(hici_p,na.rm = T))

kable(summary(shicip_mean))
```

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

```
#gives the mean of the publications in the Shanghai dataset

#The code below shows Citations by Country in the Shanghai dataset showing the mean, standard deviation, max and min.

desc_shanghai <- shangten%>%
 group_by(Country)%>%
 summarise(mean_spub=mean(`Highly Cited Researchers`),
        sd_spub=sd(`Highly Cited Researchers`),
        max_spub=max(`Highly Cited Researchers`),
        min_spub=min(`Highly Cited Researchers`))
```

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

```
#The code below shows which country has the maximum and minimum HICI/Citations in the Shanghai dataset

desc_shanghai%>%
 filter(max(mean_spub) == mean_spub)
```

A tibble: 0 x 5
... with 5 variables: Country , mean_spub , sd_spub , max_spub , min_spub

```
desc_shanghai%>%
 filter(min(mean_spub) == mean_spub)
```

A tibble: 0 x 5
... with 5 variables: Country , mean_spub , sd_spub , max_spub , min_spub

```
kable(summary(desc_shanghai))
```

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

```
timesr_mean <- timesten%>%summarise(mean(Research,na.rm = T))

kable(summary(timesr_mean))
```

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

```
 #na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean Research/Publications Percent within the Times Dataset

```
timesten <- timesten%>%mutate(research_p=percent_rank(timesten$Research)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the publication number in the Times dataset

timesrp_mean <- timesten%>%summarise(mean(research_p,na.rm = T))

kable(summary(timesrp_mean))
```

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

```
 #gives the mean of the publications in the Times dataset

#The code below shows Publications/Research by Country in the Times dataset showing the mean, standard deviation, max and min.

desc_times <- timesten%>%
 group_by(Country)%>%
 summarise(mean_tpub = mean(Research),
        sd_tpub = sd(Research),
        max_tpub= max(Research),
        min_tpub= min(Research))
```

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

```
#The code below shows which country has the maximum and minimum Publications/Research in the Times dataset
```

```
desc_times%>%
 filter(max(mean_tpub) == mean_tpub)
```

A tibble: 1 x 5

	Country	mean_tpub	sd_tpub	max_tpub	min_tpub
1	Singapore	68.1	13.6	87.2	47.8

```
desc_times%>%
 filter(min(mean_tpub) == mean_tpub)
```

A tibble: 1 x 5

	Country	mean_tpub	sd_tpub	max_tpub	min_tpub
1	Morocco	6.4	0.141	6.5	6.3

```
kable(summary(desc_times))
```

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

```
timesc_mean <- timesten%>%summarise(mean(Citations,na.rm = T))

kable(summary(timesc_mean))
```

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

```
#na.rm=T specifies what to do with any missing data in the dataset
```

Summary: Mean Citations Percent within the Times Dataset

```
timesten <- timesten%>%mutate(citations_p=percent_rank(Citations)*100)
 #default would give 0-1 so *by 100 and it will make it easier to understand the citations number in the Times dataset

timescp_mean <- timesten%>%summarise(mean(citations_p,na.rm = T))

kable(summary(timescp_mean))
```

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

```
#gives the mean of the publications in the Times dataset

#The code below shows Publications/Research by Country in the Times dataset showing the mean, standard deviation, max and min.

desc_times <- timesten%>%
 group_by(Country)%>%
 summarise(mean_tcit = mean(Citations),
        sd_tcit = sd(Citations),
        max_tcit= max(Citations),
        min_tcit= min(Citations))
```

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

```
#The code below shows which country has the maximum and minimum Publications/Research in the Times dataset

desc_times%>%
 filter(max(mean_tcit) == mean_tcit)
```

A tibble: 1 x 5

Country	mean_tcit	sd_tcit	max_tcit	min_tcit
1 Luxembourg	84.8	NaN	84.8	84.8

```
desc_times%>%
 filter(min(mean_tcit) == mean_tcit)
```

A tibble: 1 x 5

Country	mean_tcit	sd_tcit	max_tcit	min_tcit
1 Ukraine	2.95	1.77	4.2	1.7

```
kable(summary(desc_times))
```

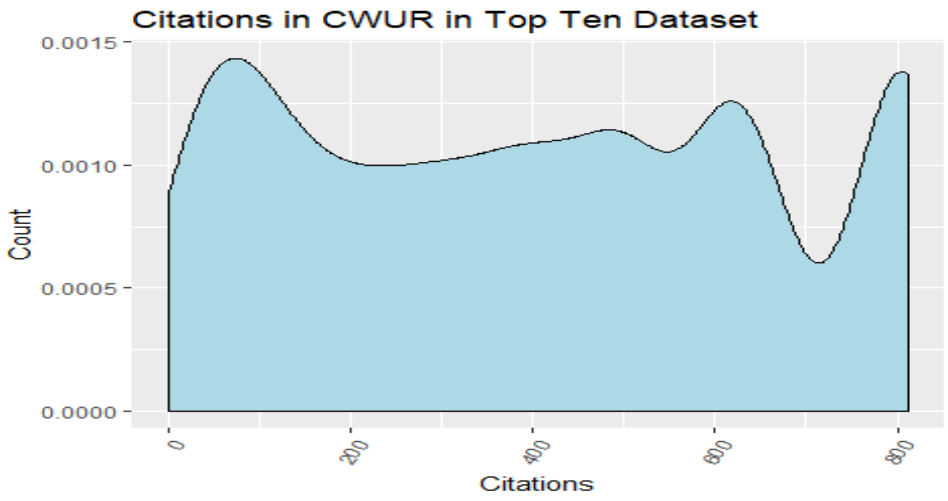
Country	mean_tcit	sd_tcit	max_tcit	min_tcit
Length:72	Min. : 2.95	Min. : 1.768	Min. : 4.20	Min. : 1.20
Class :character	1st Qu.:22.29	1st Qu.:11.765	1st Qu.: 32.05	1st Qu.:10.20
Mode :character	Median :43.94	Median :14.655	Median : 74.25	Median :18.10
NA	Mean :42.89	Mean :16.717	Mean : 63.58	Mean :21.99
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 that 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.

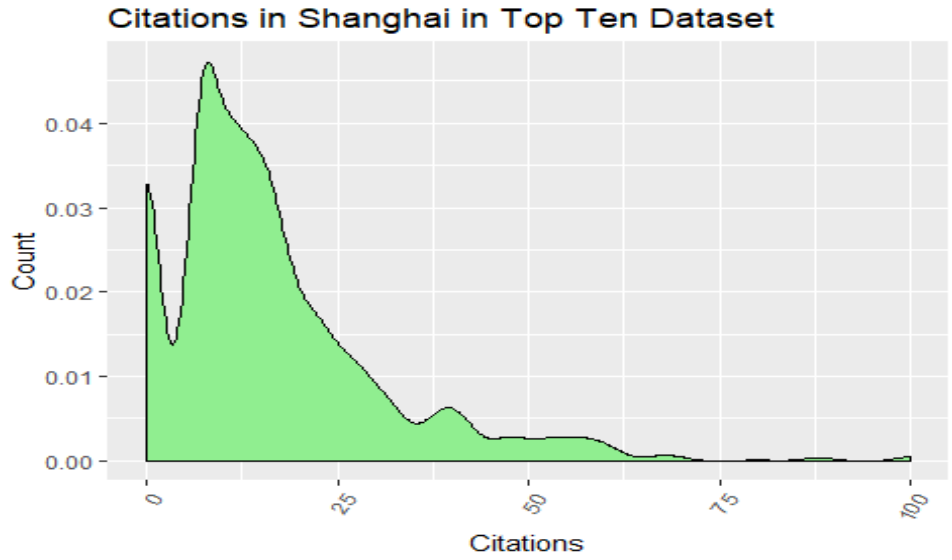
```
gc1 <- ggplot(cwurten,aes(Citations))+
 geom_density(binwidth=1, fill="lightblue")+ #Density is the Chart Shape
 theme(axis.text.x=element_text(angle=60, hjust=1))+
 labs(title="Citations in CWUR in Top Ten Dataset", x="Citations", y="Count") #chart labels

gc1
```



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.

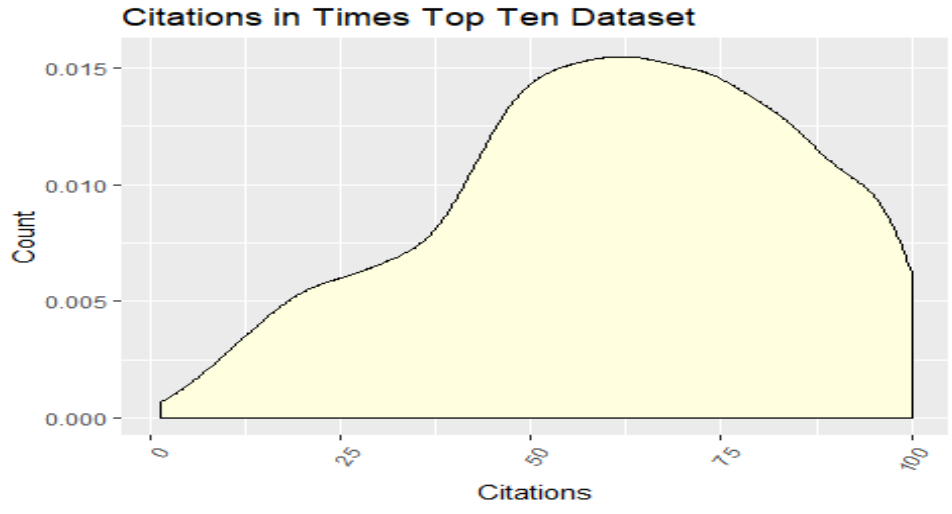
```
gs1 <- ggplot(shangten, aes(x=shangten$`Highly Cited Researchers`))+
 geom_density(binwidth=1, fill="lightgreen")+ #Density is the Chart Shape
 theme(axis.text.x=element_text(angle=60, hjust=1))+
 labs(title="Citations in Shanghai in Top Ten Dataset", x="Citations", y="Count") # chart labels

gs1
```



Density Plot #3: Citations in Times Dataset

The plot above shows the level of citations referenced within the Times dataset. You will notice that the average or peak for the citations is at the middle mark between 50-75 citations, approximately 62.5 citations.

```
gt1 <- ggplot(timesten, aes(x=Citations))+
 geom_density(binwidth=1, fill="lightyellow")+ #Density is the Chart Shape
 theme(axis.text.x=element_text(angle=60, hjust=1))+
 labs(title="Citations in Times Top Ten Dataset", x="Citations", y="Count") #labels

gt1
```



World Rankings Dataset Predictions

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

CWUR Dataset Prediction

Publications

```
cpub1<-lm(Publications~Citations,data=cwurten)
 #publications ~(as a function of) citations, data=dataset(cwurten)

#outcome (publications) on left, predictor (citations) on right

pander::pander(summary(cpub1))#shows the results of the regression
```

	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

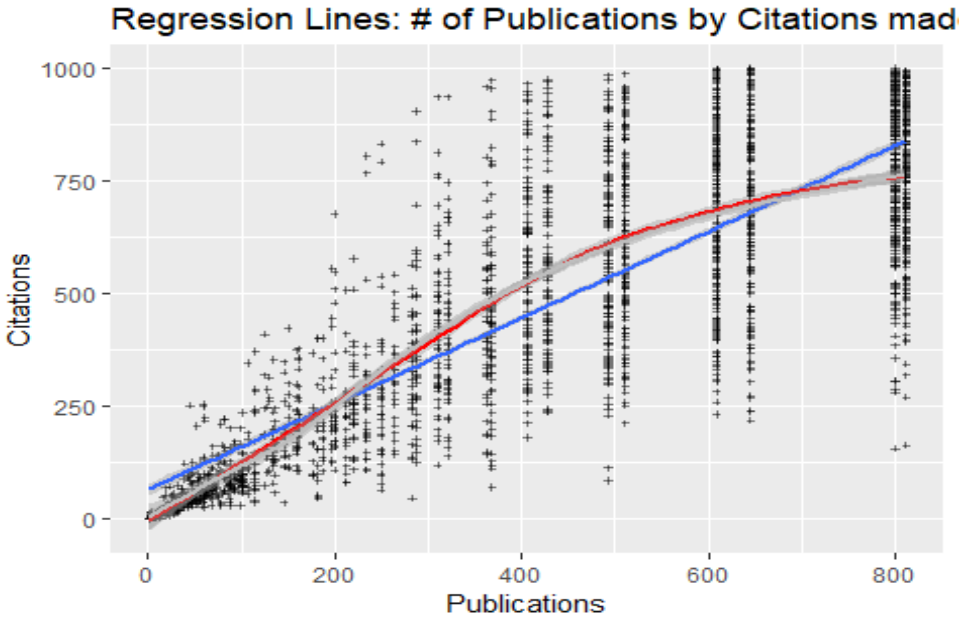
Observations	Residual Std. Error	R^2	Adjusted R^2
2200	169.5	0.6888	0.6886

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.

```
#The graph/plot below is pulling from the top ten dataset - cwurten in order to present the best point plot

gc2 <- ggplot(cwurten, aes(x=Citations, y=Publications))+
 geom_point(shape=3, alpha=.5, size=.5)+
 geom_smooth(method="lm")+
 geom_smooth(method="loess", color="red")+
 geom_smooth(color="grey")+
 labs(title="Regression Lines: # of Publications by Citations made in CWUR Top Ten Dataset", x="Publications",
y="Citations")

gc2

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



##lm puts a straight line through the publication data points and is similiar to a best line fit.
##LOESS fits a curve through the publication data points and is similar to modeling with calculus as it is the weighted sum of squared erros and may accurately account for the range within the dataset.

RMSE for CWUR Publications

```
#The code below runs the root mean squared error number from a validation of the model data above

cwurten<- cwurten%>%add_predictions(cpub1)%>%rename(predc1=pred)
 #predict using data in memory

rmse_cpub1<-modelr::rmse(cpub1,cwurten);rmse_cpub1
```

[1] 169.4278

```
pander::pander(summary(rmse_cpub1))
```

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

#shows the root mean squared average for the CWUR publication dataset prediction

Coefficient Data for CWUR Publications

```
confint.lm(cpub1)
```

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

```
pander::pander(summary(cpub1))
```

	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

```
#This code only shows the coefficient Data
```

Shanghai Dataset Prediction

Publications

```
spub1<-lm(Publications~`Highly Cited Researchers`,data=shangten)
  #publications ~(as a function of) HICI, data=dataset(shangten)

#outcome (publications) on left, predictor (HICI/Citations) on right

pander(summary(spub1))#shows the results of the regression
```

	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

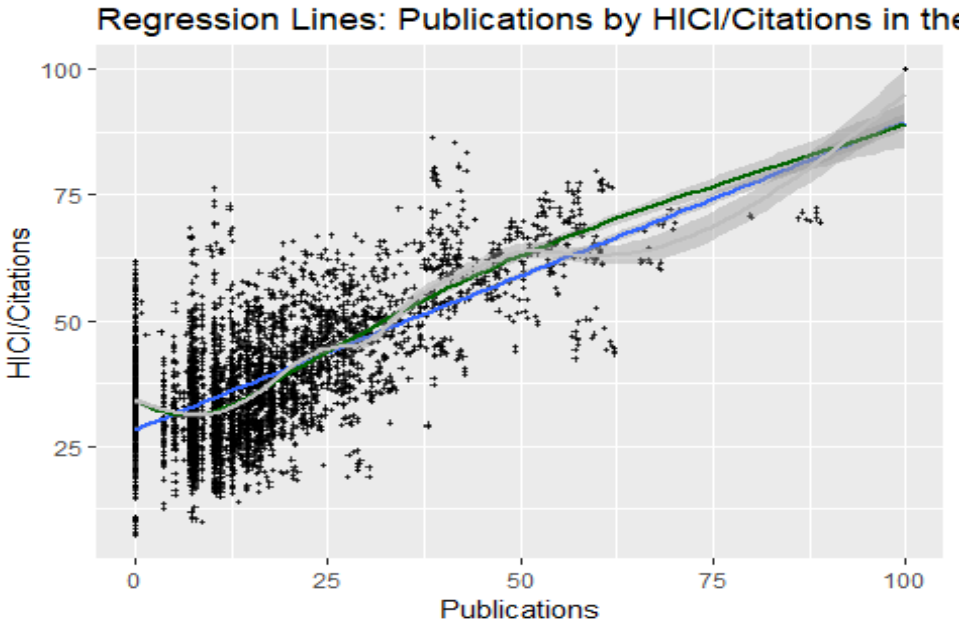
Observations	Residual Std. Error	R^2	Adjusted R^2
4895	9.641	0.4544	0.4543

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.

```
#The graph/plot below is pulling from the top ten dataset - shangten in order to present the best point plot

gs2 <- ggplot(shangten, aes(x=shangten$'Highly Cited Researchers', y=Publications))+
  geom_point(shape=3, alpha=.75, size=.25)+ #specifies the point by shape and size
  geom_smooth(method = "lm")+
  geom_smooth(method = "loess", color="darkgreen")+
  geom_smooth(color="grey")+
  labs(title="Regression Lines: Publications by HICI/Citations in the Shangahi Top Ten Dataset", x="Publications",
y="HICI/Citations")

gs2

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



##lm puts a straight line through the publication data points and is similiar to a best line fit.
##LOESS fits a curve through the publication data points and is similar to modeling with calculus as it is the weighted sum of squared erros and may accurately account for the range within the dataset.

RMSE for Shanghai Publications

```
#The code below runs the root mean squared error number from a validation of the model data above

shangten<- shangten%>%add_predictions(spub1)%>%rename(preds1=pred)
 #predict using data in memory

rmse_spub1<-modelr::rmse(spub1,shangten);rmse_spub1
```

[1] 9.639154

```
pander::pander(summary(rmse_spub1))
```

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

#on average we are off by 9.63 points

Coefficient Data for Shanghai Publications

```
confint.lm(spub1)
```

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

```
pander::pander(summary(spub1))
```

	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

```
#This code only shows the coefficient Data
```

Times Dataset Prediction

Research/Publications

```
tpub1<-lm(Research~Citations,data=timesten)
 #research/pubs ~(as a function of) citations, data=dataset(timesten)

#outcome (research/pubs) on left, predictor (citations) on right

pander::pander(summary(tpub1))#shows the results of the regression
```

	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

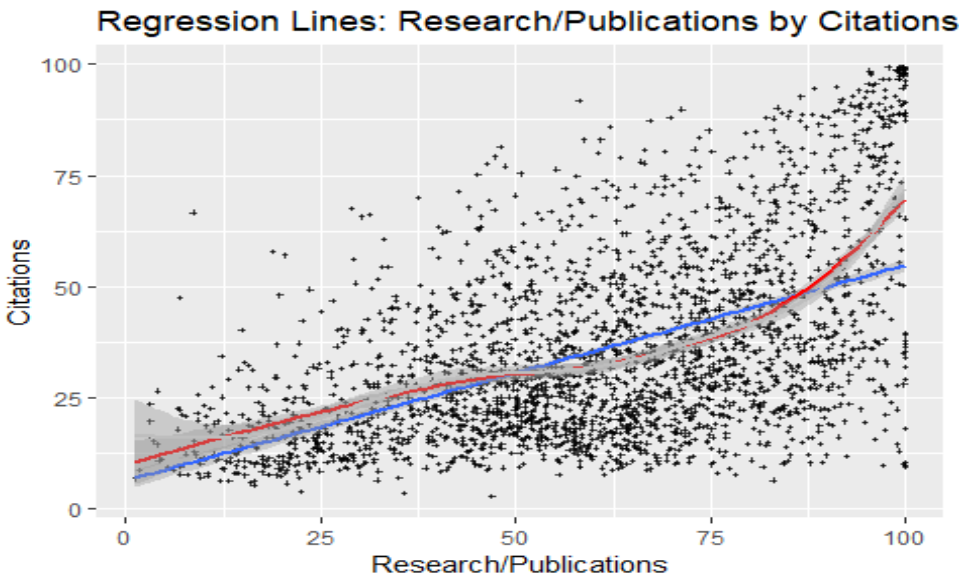
Observations	Residual Std. Error	R^2	Adjusted R^2
2603	18.09	0.2759	0.2756

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.

```
#The graph/plot below is pulling from the top ten dataset - timesten in order to present the best point plot

gt2 <- ggplot(timesten, aes(x=Citations, y=Research))+
 geom_point(shape=3, alpha=.75, size=.25)+ #specifies the points
 geom_smooth(method = "lm")+ #linear model line
 geom_smooth(method = "loess", color="red")+ #LOESS line
 geom_smooth(color="grey")+
 labs(title = "Regression Lines: Research/Publications by Citations in the Times Top Ten Dataset",
x="Research/Publications", y="Citations")

gt2

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



##lm puts a straight line through the publication data points and is similiar to a best line fit.
##LOESS fits a curve through the publication data points and is similar to modeling with calculus as it is the weighted sum of squared erros and may accurately account for the range within the dataset.

RMSE for Times Publications

```
#The code below runs the root mean squared error number from a validation of the model data above

timesten <- timesten%>%add_predictions(tpub1)%>%rename(predt1=pred)
  #predict using data in memory

rmse_tpub1 <- modelr::rmse(tpub1,times);rmse_tpub1

pander::pander(summary(rmse_tpub1))
  #on average we are off by 18 points
```

Coefficient Data for Times Publications

```
confint.lm(tpub1)
```

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

```
pander::pander(summary(tpub1))
```

	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

#This code only shows the coefficient Data

Summary RMSE data for World Ranking Dataset Publications

CWUR Prediction

```
#Prediction from the Publications data under the 3 world ranking datasets

pander::pander(rmse_cpub1)
```

169.4

Shanghai Prediction

```
pander::pander(rmse_spub1)
```

9.639

Times Prediction

```
pander::pander(rmse_tpub1)
```

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

```
#the code below will create the CWUR Top Ten cross-tab within the CWURtt table

tab_cten <- with(cwurtt,table(Publications,cwurtt$`World Rank`))

#with command to make a table that uses a specific set of data

tab_cten

colnames(tab_cten) <- c("WR1","WR2","WR3","WR4","WR5","WR6","WR7","WR8","WR9","WR10")
#the code above names the column headers - WR stands for World Rank and the number is affiliated with the rank 1-10

summary(tab_cten) ##kable command will output the table in a format that is appropriate for markdown
```

CWUR Proportion of Top Ten by World Rank:

```
#In general recommends using proportions instead of counts.

tab_cten_prop <- prop.table(tab_cten, margin=1)  #creates a proportion table

kable(tab_cten_prop)
```

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 reflecting the number of publications and the percentage of those publications by world rank.

```
#code below will change the proportion to a %

print(kable(round(tab_cten_prop*100,2)),
    only.contents=T,
    comment=F,
    sanitize.colnames.function=identity,
    sanitize.rownames.function=identity,
    hline.after=0:10)

#multiply by 100 and rounds to 2 decimal places
#warning to not have more than 2 decimal points and when it does it indicates a false sense of percision that doesn't reflect things like measurement error or other items in the data
```

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

```
#the code below produces the CWUR top ten list by Publication Probability

cten_sum <- cwurtt%>%
 group_by(Country,cwurtt$`World Rank`)%>%
 summarize(prob_pub=mean(cwurtt$Publications,na.rm=TRUE))

pander(cten_sum,
    only.contents=T,
    comment=F,
    sanitize.colnames.function=identity,
    sanitize.rownames.function=identity,
    hline.after=0:3)
```

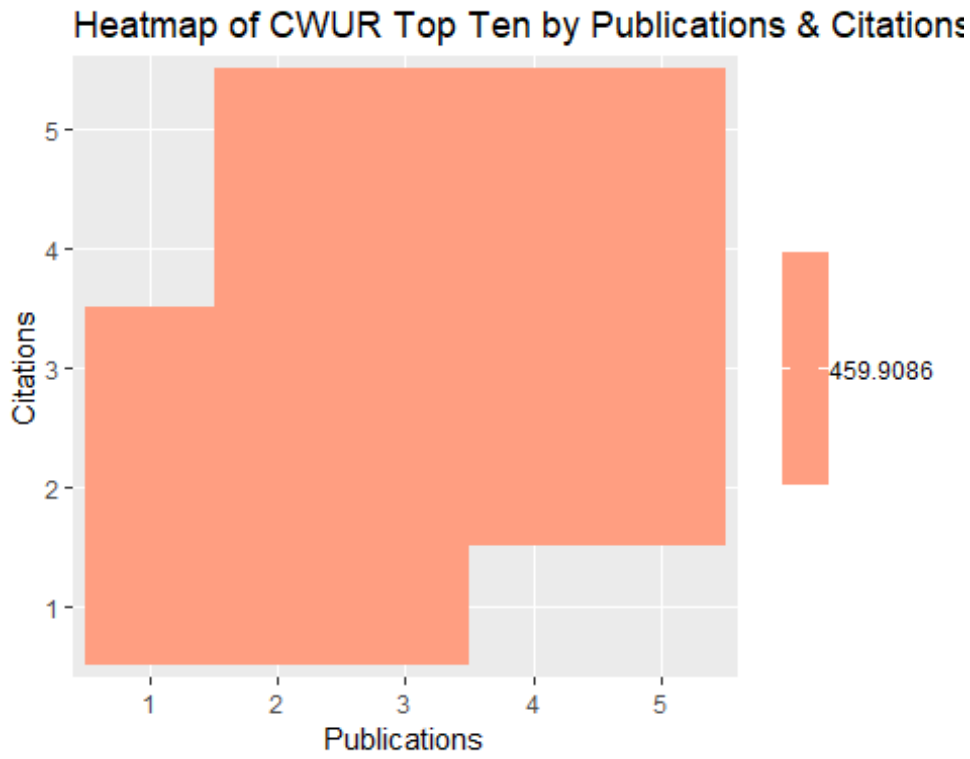
Country	cwurtt$World Rank	prob_pub
United Kingdom	3	17.32
United Kingdom	4	17.32
United Kingdom	5	17.32
United Kingdom	7	17.32
USA	1	17.32
USA	2	17.32
USA	3	17.32
USA	4	17.32
USA	5	17.32
USA	6	17.32
USA	7	17.32
USA	8	17.32
USA	9	17.32
USA	10	17.32

```
#the code below will divide the publication and citation independent variables into quintiles for the CWUR top ten dataset
```

```
cwurten <- cwurten%>%
 mutate(Publications_quintile=ntile(Publications,5),
        Citations_quintile=ntile(Citations,5))

#Create a summary dataset that shows the probabilities of the outcome across all of the combined categories of the two independent variables.
#the code below combines the publication quintile and citation quintile categories

cten1_sum <- cwurten%>%
 group_by(Publications_quintile,Citations_quintile)%>%
 summarize(prob_pub=mean(cwurten$Publications,na.rm=TRUE))%>%
 arrange(-prob_pub)

#Missing data isn't important, so we'll drop it such as n/a's that are in the dataset

cten1_sum <- cten1_sum%>%
 filter(!(is.na(Publications_quintile)),!(is.na(Citations_quintile)))
```

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.

```
gc3 <- ggplot(cten1_sum,
       aes(x=as.factor(Publications_quintile),
           y=as.factor(Citations_quintile),fill=prob_pub))
gc3 <- gc3+geom_tile()+
 scale_fill_gradient(low="white",high="red")+
 labs(title = "Heatmap of CWUR Top Ten by Publications & Citations", x="Publications", y="Citations")+
 theme(legend.title=element_blank())

gc3
```



Shanghai Top Ten Proportions, Plots & Heat Maps

```
#the code below will create the Shanghai Top Ten cross-tab within the Shangtt table

tab_sten <- with(shangtt,table(Publications,`World Rank`))

#with command to make a table that uses a specific set of data

colnames(tab_sten) <- c("WR1","WR2","WR3","WR4","WR5","WR6","WR7","WR8","WR9","WR10")
#the code above names the column headers - WR stands for World Rank and the number is affiliated with the rank 1-10

summary(tab_sten) ##kable command will output the table in a format that is appropriate for markdown

#In general recommends using proportions instead of counts.

tab_sten_prop <- prop.table(tab_sten, margin=1) #creates the proportions table
kable(tab_sten_prop)
```

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

```
#code below will change the proportion to a %

kable(round(tab_sten_prop*100,0))

#multiply by 100 and rounds to 0 decimal places
#warning to not have more than 2 decimal points and when it does it indicates a false sense of percision that doesn't reflect things like measurement error or other items in the data
```

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

```
##Probability of Publications by Top Ten World Rank.

sten_sum <- shangtt%>%
 group_by(Country,shangtt$`World Rank`)%>%
 summarize(prob_pub=mean(Publications,na.rm=TRUE))

pander(sten_sum,
    only.contents=T,
    comment=F,
    sanitize.colnames.function=identity,
    sanitize.rownames.function=identity,
    hline.after=0:10)
```

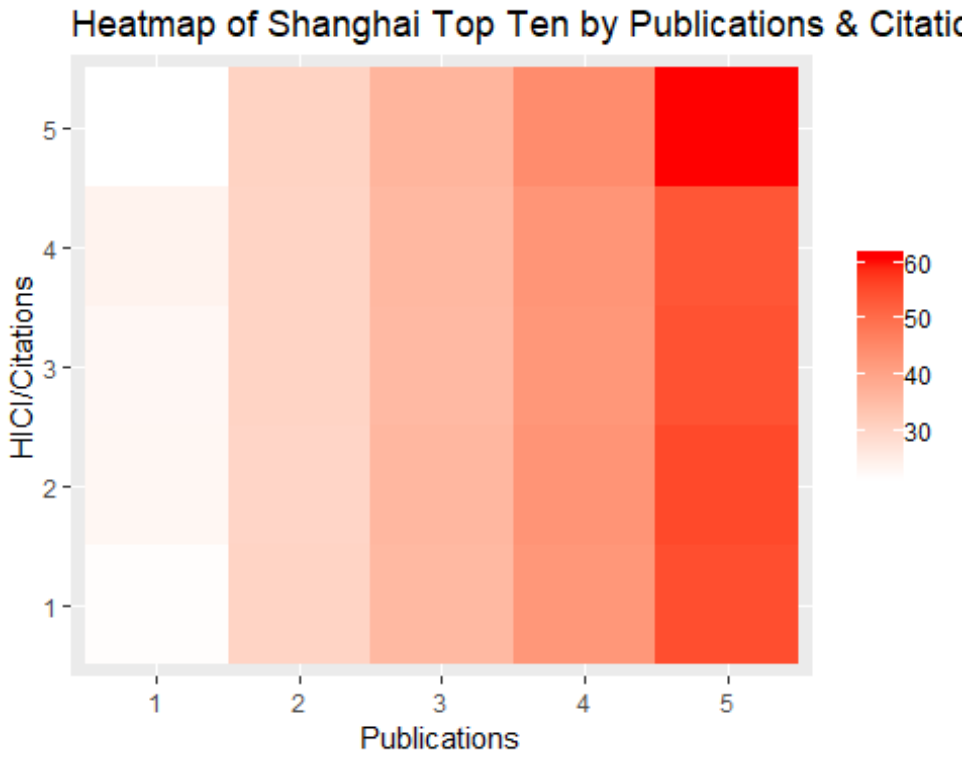
Country	shangtt$World Rank	prob_pub
United Kingdom	2	70.75
United Kingdom	4	65.37
United Kingdom	5	65.97
United Kingdom	6	48.34
United Kingdom	7	69.65
United Kingdom	9	49.8

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 code below will divide the publication and citation independent variables into quintiles for the CWUR top ten dataset

shangten <- shangten%>%
 mutate(Publications_quintile=ntile(Publications,5),
      HICI_quintile=ntile(shangten$`Highly Cited Researchers`,5))

#Create a summary dataset that shows the probabilities of the outcome across all of the combined categories of the two independent variables.
#the code below combines the publication quintile and citation quintile categories

sten1_sum <- shangten%>%
 group_by(Publications_quintile,HICI_quintile)%>%
 summarize(prob_pub=mean(Publications,na.rm=TRUE))%>%
 arrange(-prob_pub)

#Missing data isn't important, so we'll drop it such as n/a's that are in the dataset

sten1_sum <- sten1_sum%>%
 filter(!(is.na(Publications_quintile)),!(is.na(HICI_quintile)))
```

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.

```
gs3 <- ggplot(sten1_sum,
       aes(x=as.factor(Publications_quintile),
          y=as.factor(HICI_quintile),fill=prob_pub))
gs3<-gs3+geom_tile()+
 scale_fill_gradient(low="white",high="red")+
 labs(title = "Heatmap of Shanghai Top Ten by Publications & Citations", x="Publications", y="HICI/Citations")+
 theme(legend.title=element_blank())

gs3
```



Times Top Ten Proportions, Plots & Heat Maps

```
#the code below will create the Times Top Ten cross-tab within the Timestt table

tab_tten <- with(timestt,table(Research,`World Rank`))

#with command to make a table that uses a specific set of data

#In general recommends using proportions instead of counts.

#colnames(tab_tten) <- c("WR1","WR2","WR3","WR4","WR5","WR6","WR7","WR8","WR9","WR10")

tab_tten_prop <- prop.table(tab_tten, margin=1)#creates the proportions table

kable(tab_tten_prop)
```

Times 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 research/publications produced by the world rank top ten list reflecting the number of publications and the percentage of those publications by world rank.

```
#code below will change the proportion to a %

kable(round(tab_tten_prop*100,2))

#multiply by 100 and rounds to 2 decimal places
#warning to not have more than 2 decimal points and when it does it indicates a false sense of percision that doesn't reflect things like measurement error or other items in the data
```

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

```
##Probability of Publications by Top Ten World Rank.
```

```
tten_sum <- timestt%>%
 group_by(Country,`World Rank`)%>%
 summarize(prob_pub=mean(Research,na.rm=TRUE))

pander(tten_sum,
    only.contents=T,
    comment=F,
    sanitize.colnames.function=identity,
    sanitize.rownames.function=identity,
    hline.after=0:10)
```

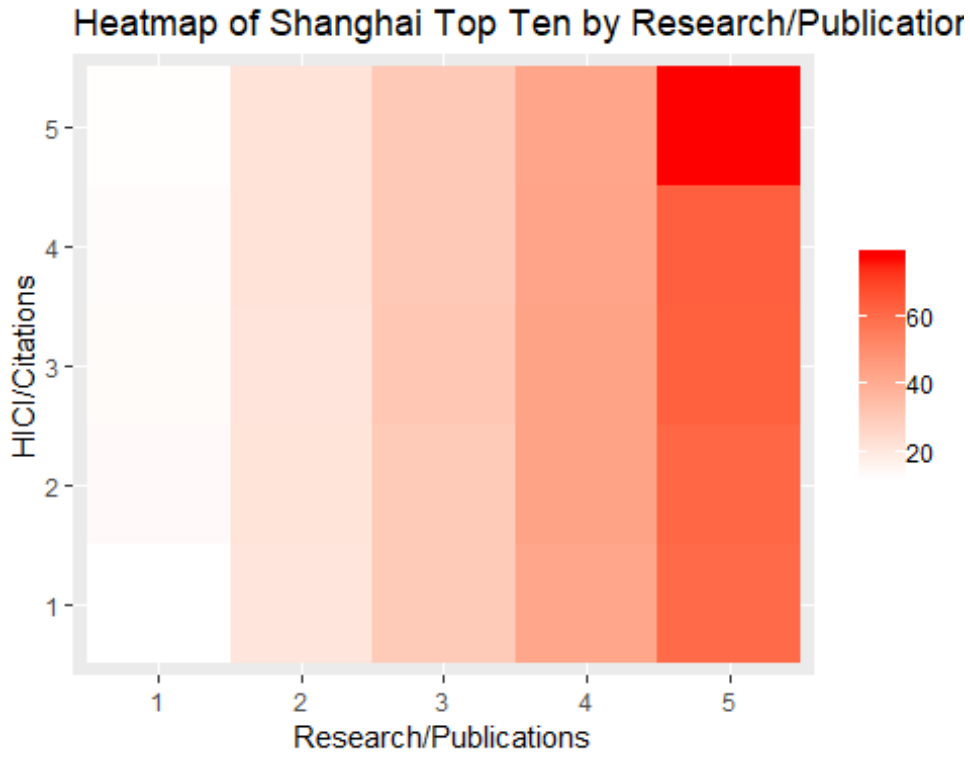
Country	World Rank	prob_pub
Switzerland	9	95
United Kingdom	2	98.5
United Kingdom	3	97.7
United Kingdom	4	96.65
United Kingdom	5	95.6
United Kingdom	6	94.07
United Kingdom	7	95.45
United Kingdom	8	89.37
United Kingdom	9	91.4
United Kingdom	10	88.1
United States of America	1	98.37
United States of America	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 code below will divide the research/publication and citation independent variables into quintiles for the CWUR top ten dataset

timesten <- timesten%>%
 mutate(Research_quintile=ntile(Research,5),
        Citations_quintile=ntile(Citations,5))

#Create a summary dataset that shows the probabilities of the outcome across all of the combined categories of the two independent variables.
#the code below combines the publication quintile and citation quintile categories

tten1_sum <- timesten%>%
 group_by(Research_quintile,Citations_quintile)%>%
 summarize(prob_pub=mean(Research,na.rm=TRUE))%>%
 arrange(-prob_pub)
```

```
#Missing data isn't important, so we'll drop it such as n/a's that are in the dataset

tten1_sum <- tten1_sum%>%
 filter(!(is.na(Research_quintile)),!(is.na(Citations_quintile)))
```

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

```
gt3 <- ggplot(tten1_sum,
       aes(x=as.factor(Research_quintile),
         y=as.factor(Citations_quintile),fill=prob_pub))
gt3 <- gt3+geom_tile()+
 scale_fill_gradient(low="white",high="red")+
 labs(title = "Heatmap of Shanghai Top Ten by Research/Publications & Citations", x="Research/Publications",
y="HICI/Citations")+
 theme(legend.title=element_blank())

gt3
```



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 is shown 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

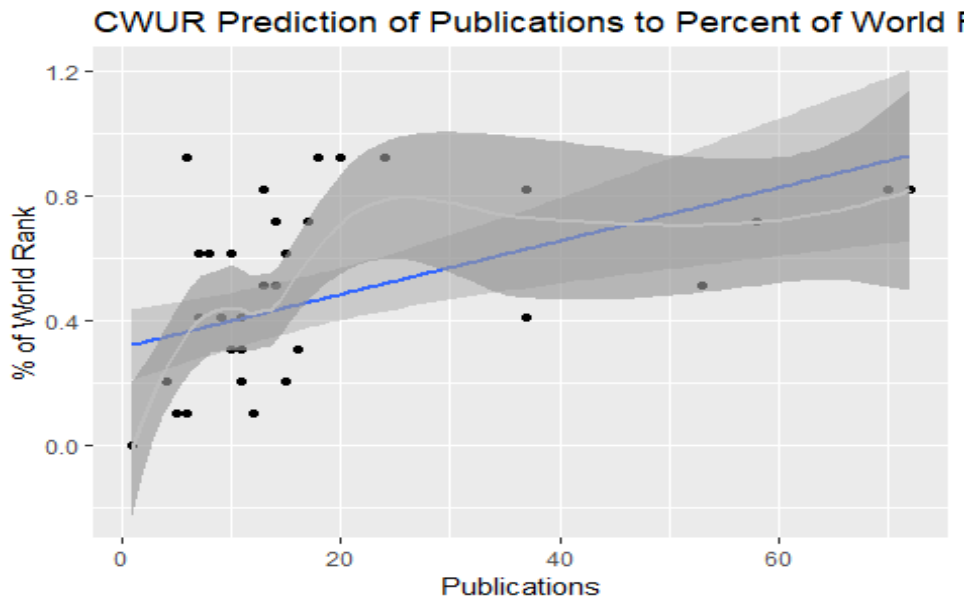
CWUR Cross-Validation and Prediction

```
cwurtt <- cwurtt%>%
 select(Publications, Citations, cwurtt$`National Rank`, cwurtt$`World Rank`)%>%
 mutate_all(funs(as.numeric))%>%
 mutate(world_rank=percent_rank(cwurtt$`World Rank`))%>%
 tbl_df()
```

Graph 10: CWUR Prediction

The results above reflects 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.

```
gcp <- ggplot(cwurtt, aes(x=Publications, y=world_rank))+
 geom_point()+
 geom_smooth(method="lm")+
 geom_smooth(method="loess", color="red")+
 geom_smooth(color="grey")+
 labs(title="CWUR Prediction of Publications to Percent of World Rank", x="Publications", y="% of World Rank")

gcp

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



CWUR Cross-Validation with Kfolds

```
##Define the Model
cmod_formula <- formula(cwur$`World Rank`~cwur$Publications + cwur$Citations)

## Run the model against all of the data
basic.mod <- lm(cmod_formula,
          data=cwur);
pander::pander(summary(basic.mod))
```

	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

```
#uses command kfold and divides dataset into the different folds.  each time you run the code you will get new data as it is randomized

cwur_cf <- cwur%>%
 crossv_kfold(50)

cwur_cf

#Column 1 is training dataset, Column 2 is testing datset, Column 3 is ID - this is a dataset within a dataset (nested dataset).  19 of 20 in training dataset and 20th one is in the testing dataset.

#to see what samples are in each

cwur_cf$test$'1'
cwur_cf$test$'2'
```

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

```
##starts by converting all of the individual training datasets to tibbles. Then the model is run on each training dataset. Then apply the predictions from the model to each testing dataset, and finally pull the rmse from each of the testing datasets.

rmse_cmod1<-cwur_cf %>% ##create new object of rmse for model 1
 mutate(train = map(train, as_tibble)) %>% ## Convert train column/dataset to tibbles
 mutate(model = map(train, ~ lm(cmod_formula,
                    data = .))) %>% ##running model on train dataset
 mutate(rmse = map2_dbl(model, test, rmse)) %>% ## apply model to test dataset, get rmse
 select(.id, rmse) ## pull just id and rmse which is the standard measure of performance

kable(summary(rmse_cmod1))
```

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.

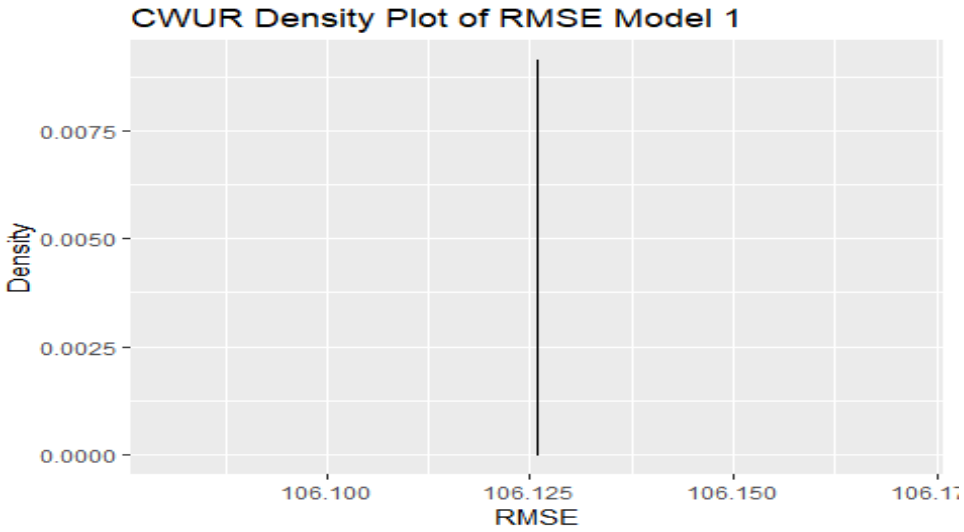
Graph 11: CWUR Density Plot for RMSE Model 1

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

```
##this model generally misses by about .2 (20% points on the home rank scale)

pander(summary(rmse_cmod1$rmse))
```

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

```
gc4 <- ggplot(rmse_cmod1, aes(rmse))+
 geom_density(binwidth=1, fill="azure")+
 labs(title="CWUR Density Plot of RMSE Model 1", x="RMSE", y="Density")

gc4
```



Full Cross-Validation: Random Partition

```
#randomly draws from the dataset

cwur_cv<-cwur%>%
  crossv_mc(n=400,test=.2) ##prop of data to be held out - 20% or 400 rows to test
cwur_cv
```

CWUR RMSE Cross-Validation for Model 1

```
##same approach, but with the MUCH larger qf_cv dataset. This will take a bit of time.

cmod1_rmse_cv<-cwur_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(cmod_formula, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse

kable(summary(cmod1_rmse_cv))
```

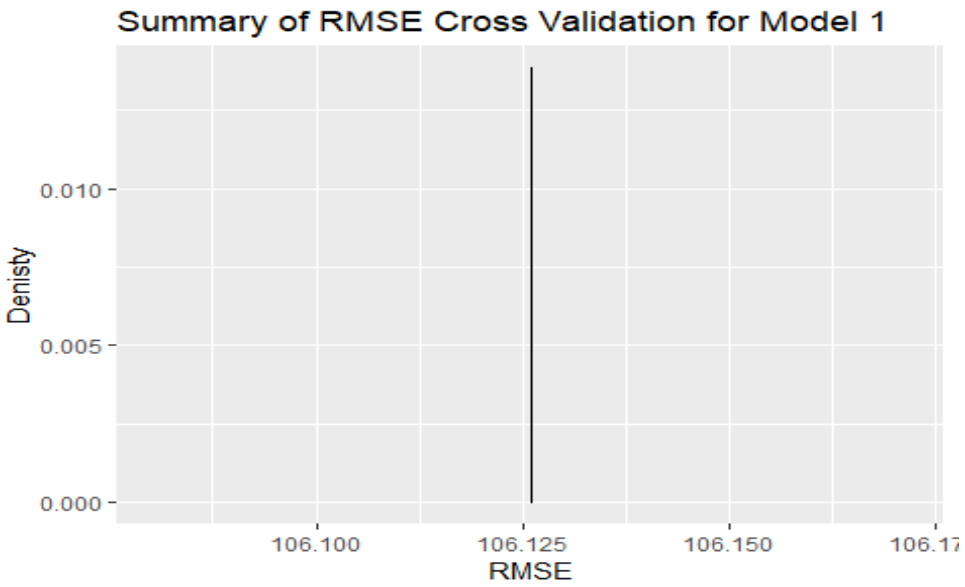
Graph 12: CWUR Density Plot for Cross-Validation of RMSE Model 1

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

```
pander(summary(cmod1_rmse_cv$rmse))
```

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

```
gc5 <- ggplot(cmod1_rmse_cv, aes(rmse))+
  geom_density(bins=50, fill="sienna", alpha=.2)+
  labs(title="Summary of RMSE Cross Validation for Model 1", x="RMSE", y="Denisty")

gc5
```



Shanghai Cross-Validation and Prediction

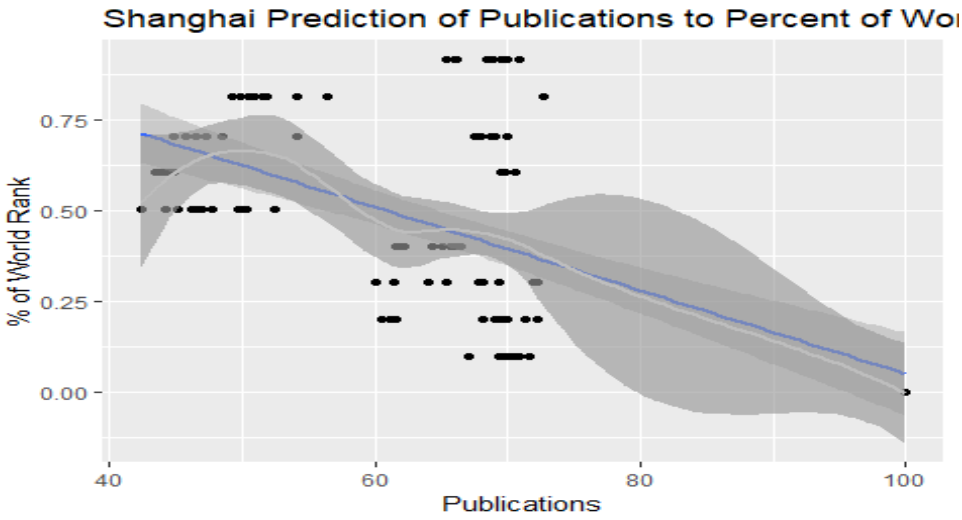
```
shangtt <- shangtt%>%
 select(Publications,`Highly Cited Researchers`,`National Rank`,`World Rank`)%>%
 mutate_all(funs(as.numeric))%>%
 mutate(world_rank=percent_rank(shangtt$`World Rank`))%>%
 tbl_df()
```

Graph 13: Shanghai Prediction

The results above reflects 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.

```
gsp <- ggplot(shangtt, aes(x=Publications, y=world_rank))+
 geom_point()+
 geom_smooth(method="lm")+
 geom_smooth(method="loess", color="red")+
 geom_smooth(color="grey")+
 labs(title="Shanghai Prediction of Publications to Percent of World Rank", x="Publications", y="% of World Rank")

gsp
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Shanghai Cross-Validation with Kfolds

```
##Define the Model
smod_formula <- formula(`World Rank`~Publications+`Highly Cited Researchers`)

## Run the model against all of the data
basic.mod <- lm(smod_formula,
          data=shanghai);
pander::pander(summary(basic.mod))
```

	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

```
#uses command kfold and divides dataset into the different folds.  each time you run the code you will get new data as it is randomized

shang_cf <- shanghai%>%
 crossv_kfold(50)

shang_cf

#Column 1 is training dataset, Column 2 is testing datset, Column 3 is ID - this is a dataset within a dataset (nested dataset).  19 of 20 in training dataset and 20th one is in the testing dataset.

#to see what samples are in each

shang_cf$test$'1'
shang_cf$test$'2'
```

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

```
##starts by converting all of the individual training datasets to tibbles. Then the model is run on each training dataset. Then apply the predictions from the model to each testing dataset, and finally pull the rmse from each of the testing datasets.
```

```
rmse_smod1<-shang_cf %>% ##create new object of rmse for model 1
 mutate(train = map(train, as_tibble)) %>% ## Convert train column/dataset to tibbles
 mutate(model = map(train, ~ lm(smod_formula,
                    data = .))) %>% ##running model on train dataset
 mutate(rmse = map2_dbl(model, test, rmse)) %>% ## apply model to test dataset, get rmse
 select(.id, rmse) ## pull just id and rmse which is the standard measure of performance

kable(summary(rmse_smod1))
```

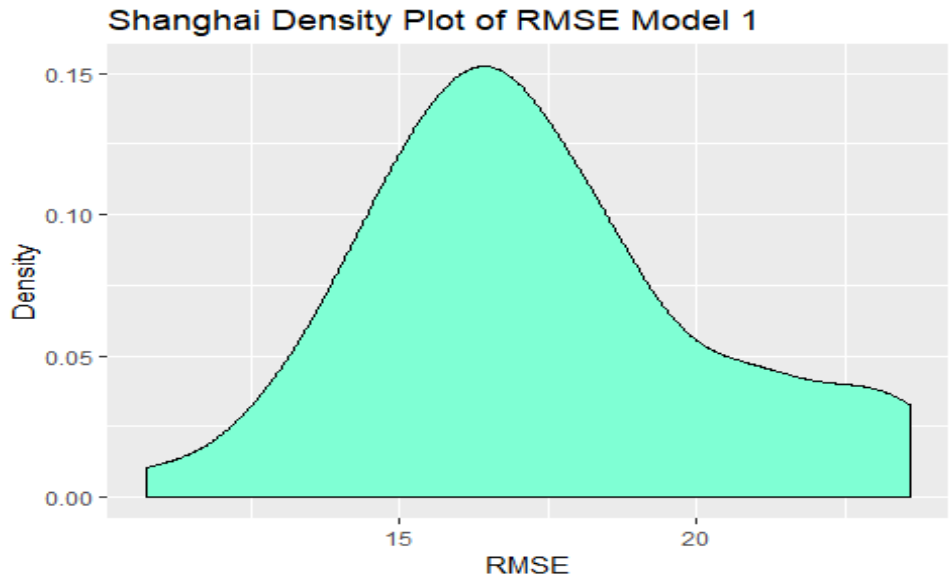
The resulting dataset includes the id for the cross-validation and the RMSE. We can summarize and plot this new data frame to see what our likely range of RMSE happens to be.

Graph 14: Shanghai Density Plot for RMSE Model 1

```
##this model generally misses by about .2 (20% points on the home rank scale)

summary(rmse_smod1$rmse)
```

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

```
gc4 <- ggplot(rmse_smod1, aes(rmse))+
 geom_density(binwidth=1, fill="aquamarine")+ #Density is the Chart Shape
 labs(title="Shanghai Density Plot of RMSE Model 1", x="RMSE", y="Density")

gc4
```



Full Cross-Validation: Random Partition

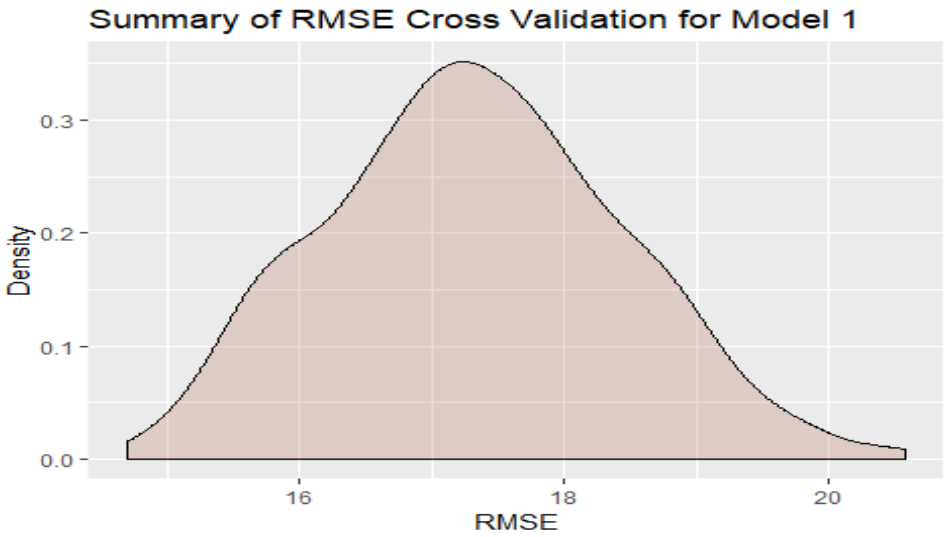
```
#randomly draws from the dataset
shang_cv<-shanghai%>%
 crossv_mc(n=450,test=.1) ##prop of data to be held out -10% or 450 rows to test
shang_cv
```

Shanghai RMSE Cross-Validation of Model 1

```
##same approach, but with the MUCH larger qf_cv dataset. This will take a bit of time.

smod1_rmse_cv<-shang_cv %>%
 mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
 mutate(model = map(train, ~ lm(smod_formula, data = .)))%>%
 mutate(rmse = map2_dbl(model, test, rmse))%>%
 select(.id, rmse) ## pull just id and rmse

kable(summary(smod1_rmse_cv))

pander(summary(smod1_rmse_cv$rmse))
```

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

```
gc5 <- ggplot(smod1_rmse_cv, aes(rmse))+
 geom_density(bins=50, fill="sienna", alpha=.2)+
 labs(title="Summary of RMSE Cross Validation for Model 1", x="RMSE", y="Density")

gc5
```



Times Cross Validation and Prediction

```
timestt <-timestt%>%
 select(Research,Citations,`World Rank`)%>%
 mutate_all(funs(as.numeric))%>%
 mutate(world_rank=percent_rank(timestt$`World Rank`))%>%
 tbl_df()
```

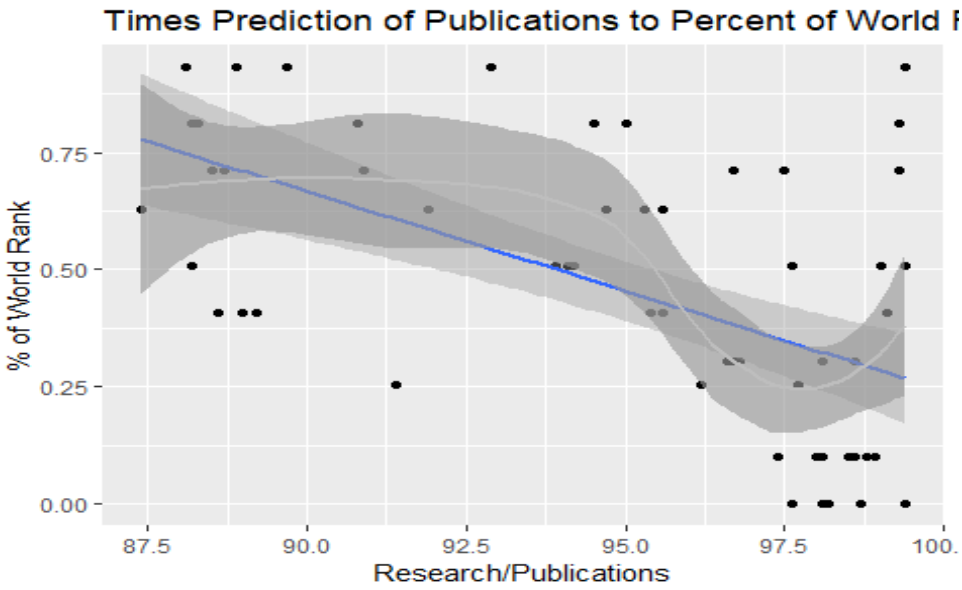
Graph 15: Times Prediction

The results above reflects 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.

```
gcp <- ggplot(timestt, aes(x=Research, y=world_rank))+
 geom_point()+
 geom_smooth(method="lm")+
 geom_smooth(method="loess", color="red")+
 geom_smooth(color="grey")+
 labs(title="Times Prediction of Publications to Percent of World Rank", x="Research/Publications", y="% of
World Rank")
```

gcp

`geom_smooth()` using method = 'loess' and formula 'y ~ x'



Times Cross Validation with Kfolds

```
##Define the Model
tmod_formula <- formula(times$`World Rank`~times$Research + times$Citations)

## Run the model against all of the data
basic.mod <- lm(tmod_formula,
          data=cwur);
pander::pander(summary(basic.mod))
```

	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

```
#uses command kfold and divides dataset into the different folds.  each time you run the code you will get new data as it is randomized

times_cf <- times%>%
 crossv_kfold(50)

times_cf

#Column 1 is training dataset, Column 2 is testing datset, Column 3 is ID - this is a dataset within a dataset (nested dataset).  19 of 20 in training dataset and 20th one is in the testing dataset.

#to see what samples are in each
```

```
times_cf$test$'1'
times_cf$test$'2'
```

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

```
##starts by converting all of the individual training datasets to tibbles. Then the model is run on each training dataset. Then apply the predictions from the model to each testing dataset, and finally pull the rmse from each of the testing datasets.

rmse_tmod1 <- times_cf %>% ##create new object of rmse for model 1
 mutate(train = map(train, as_tibble)) %>% ## Convert train column/dataset to tibbles
 mutate(model = map(train, ~ lm(tmod_formula,
                    data = .))) %>% ##running model on train dataset
 mutate(rmse = map2_dbl(model, test, rmse)) %>% ## apply model to test dataset, get rmse
 select(.id, rmse) ## pull just id and rmse which is the standard measure of performance

kable(summary(rmse_tmod1))
```

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.

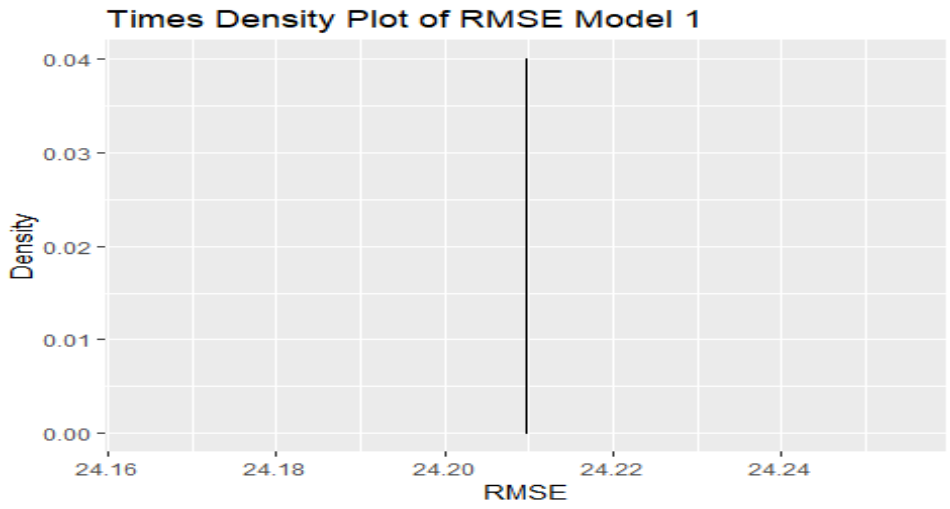
Graph 16: Times Density Plot for RMSE Model 1

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

```
##this model generally misses by about .2 (20% points on the home rank scale)

summary(rmse_tmod1$rmse)
```

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

```
gc4 <- ggplot(rmse_tmod1, aes(rmse))+
 geom_density(binwidth=1, fill="aliceblue")+
 labs(title="Times Density Plot of RMSE Model 1", x="RMSE", y="Density")

gc4
```



Full Cross-Validation: Random Partition

```
#randomly draws from the dataset

times_cv<-times%>%
 crossv_mc(n=500,test=.2) ##prop of data to be held out - 20% or 1,000 rows to test
times_cv
```

Times RMSE Cross Validation for Model 1

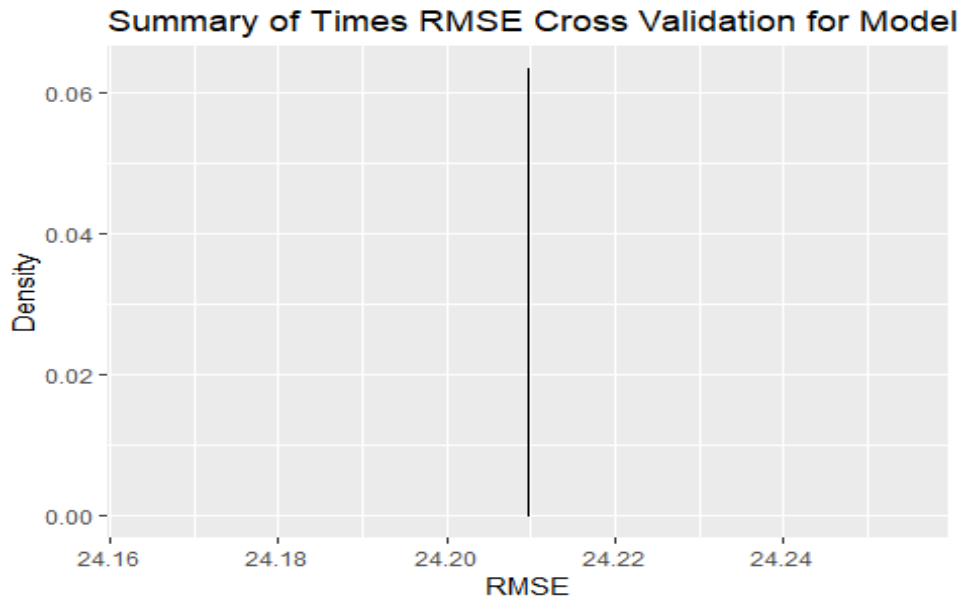
```
##same approach, but with the MUCH larger qf_cv dataset. This will take a bit of time.

tmod1_rmse_cv<-times_cv %>%
 mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
 mutate(model = map(train, ~ lm(tmod_formula, data = .)))%>%
 mutate(rmse = map2_dbl(model, test, rmse))%>%
 select(.id, rmse) ## pull just id and rmse

kable(summary(tmod1_rmse_cv))
```

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

```
pander(summary(tmod1_rmse_cv$rmse))
```

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

```
gc5 <- ggplot(tmod1_rmse_cv, aes(rmse))+
 geom_density(bins=50, fill="sienna", alpha=.2)+
 labs(title="Summary of Times RMSE Cross Validation for Model 1", x="RMSE", y="Density")

gc5
```



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