An Exploratory Study of the US Collegiate System through Data

RK

August 20, 2016

This is a cursory look at a data-set using statistical study and data visualization techniques. The data-set in question is the *May 2016* version of the **U.S College Scorecard dataset** : http://catalog.data.gov/dataset/college-scorecard.

Before we generate any figures or tables, the data-set is cleaned and filtered to include only complete observations within the Continental United States using the `dplyr` package within the following code chunk:

```
require(readr)
require(dplyr)
Colleges <- as.data.frame(read_csv("colleges.csv",col_names = T))
Colleges$STABBR <- factor(Colleges$STABBR)

###Establish filtered Dataset
col <- c(as.character(unique(Colleges$STABBR)))
col <- c(col[52:59],"AK","HI")
Colleges <- tbl_df(Colleges)
Colleges2 <- select(Colleges,INSTNM:STABBR,UGDS:UGDS_NRA,SAT_AVG) %>% mut
ate(STABBR = as.character(STABBR))

for (i in 1:length(col)) {
  Colleges2 <- Colleges2 %>% filter(STABBR != col[i]) %>% filter(UGDS !=
"NULL")
}
Colleges2[,4:13] <- sapply(Colleges2[,4:13],as.numeric)
Colleges2 <- filter(Colleges2, UGDS >= 2500)
```

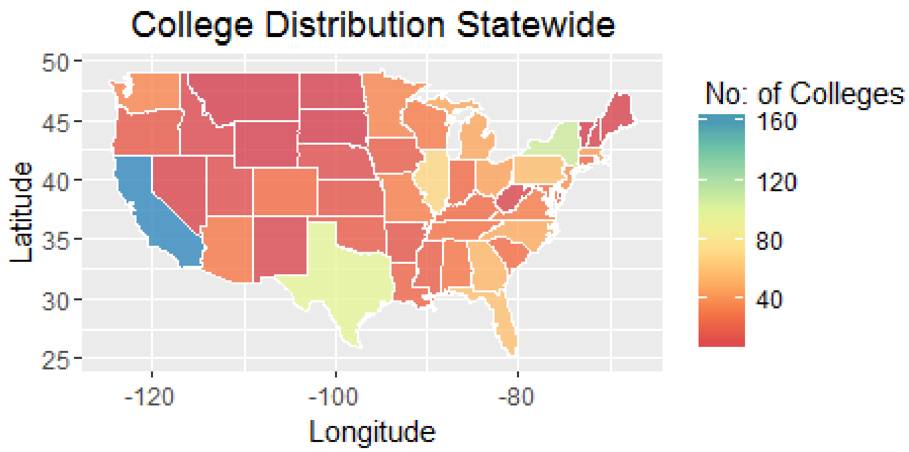
Only Universities containing above **2500** students are considered for analysis, to allow for reasonable averages and reliable results.The following constraints are examined to build the data frame used for State-wide analysis : **Overall Student Population (NStu)**, **Number of Colleges (No.)**, **White Population (Nwhi)**, and % of White **Population or % Diversity(pwhi)**

```
require(ggplot2)
require(mapdata)
require(tidyr)
require(RColorBrewer)
require(knitr)
require(scales)
colx <- group_by(Colleges2,STABBR) %>% mutate(whi = (UGDS * UGDS_WHITE))
%>% summarise(n_coll <- n(),Nwhi<- sum(whi),N_UGDS <- sum(UGDS))
names(colx)<- c("region","No.","Nwhi","NStu")
colz <- read_csv("colx.csv")
snames <- colz[,2]
colx <- mutate(colx,pwhi = Nwhi/NStu)
colsnames <- colz[,2]
colx <- mutate(colx, region = snames$regionname)
#
```

Here a Thermal Map of Collegiate Density by State is prepared with the following code, observations are included below the figure and table:

```
states <- map_data("state")
states <-tbl_df(states)
states <- left_join(states,colx, by = "region")

#No. of instututions with over 2500 students
Stupal <- brewer.pal(n = 8, name = "Spectral")
splot <- ggplot(data = states) +
  geom_polygon(aes(x = long, y = lat,fill= No., group = group, label = No
.), color = "white", alpha = 0.8) +
  labs(list(title = "College Distribution Statewide", x = "Longitude", y
= "Latitude")) + scale_fill_gradientn(name = " No: of Colleges",colours =
 Stupal ) + coord_fixed(1.3)
splot
```



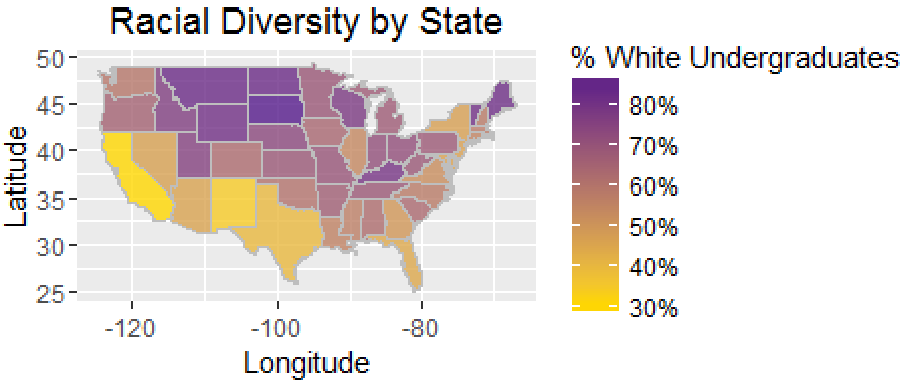
```
colx <-  arrange(colx, desc(No.))
kable(colx[1:11,],col.names = c("State","College Count","White Population
", "Total Population","% Diversity"), digits = c(2,0,0,0))
```

State	College Count	White Population	Total Population	% Diversity
California	168	606129	2036580	0.30
New York	107	390702	839066	0.47
Texas	97	444729	1136085	0.39
Illinois	69	281050	520027	0.54
Florida	59	367302	830486	0.44
Pennsylvania	57	301399	439723	0.69
Georgia	54	188240	379929	0.50
North Carolina	49	211731	363405	0.58
Michigan	47	300442	441409	0.68
Ohio	45	291856	404869	0.72
Massachusetts	43	163717	280698	0.58

The results arent different from expected. Population Centers like *CA,NY and TX* are among the nations top states for the number of institutions located in them. But another trend can be observed these states, **Low values of White Population(% Diversity) appear to be correlated with the number of institutions in the state**.

Now another figure and table will be constructed to plot density by state, with the following code:

```
splot2 <- ggplot(data = states) +
  geom_polygon(aes(x = long, y = lat,fill= pwhi, group = group, label = N
o.), color = "grey", alpha = 0.8) +
  labs(list(title = "Racial Diversity by State", x = "Longitude", y = "La
titude")) + scale_fill_gradient(low = "gold", high = "purple4", name = "%
 White Undergraduates", label = scales::percent_format()) +
  coord_fixed(1.3)
splot2
```



```
colx <- arrange(colx, desc(pwhi))
kable(colx[1:11,],col.names = c("State","College Count","White Population
", "Total Population","% Diversity"), digits = c(2,0,0,0))
```

State	College Count	White Population	Total Population	% Diversity
South Dakota	3	17407	19653	0.89
Montana	3	24192	28919	0.84

North Dakota	4	24025	28737	0.84
Maine	5	22718	27493	0.83
Wyoming	4	15626	19049	0.82
Vermont	3	14044	17222	0.82
Kentucky	24	134335	166654	0.81
Wisconsin	31	193589	245329	0.79
Idaho	8	58736	76084	0.77
Utah	10	144533	189333	0.76
Nebraska	10	53686	72135	0.74

The results of the data visualization and data clustering confirm the aforementioned assumption that **more pouplated states have more diverse student communities**. This can be inferred from the second table shows that the least populated states primarily in the Mountain West are also least diverse.

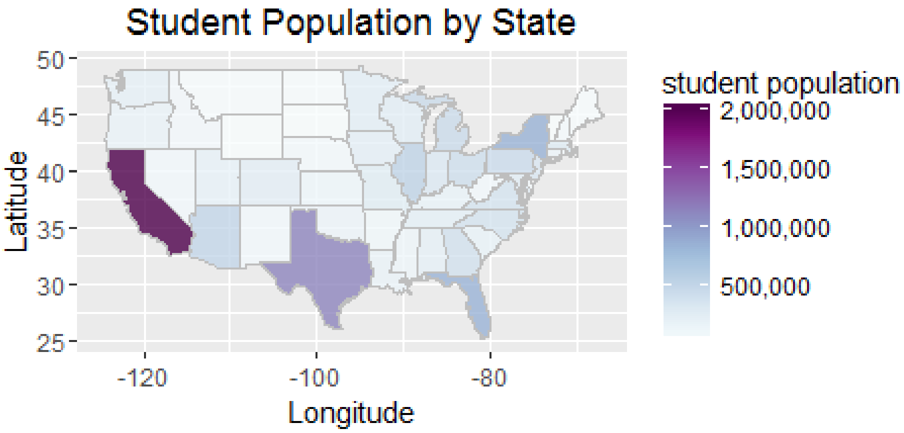
The next figure and Table is very similar to the first set, It looks at No: of Undergraduate Students per state instead at the institution level.

```
Popn <- brewer.pal(n = 10, name = "BuPu")
splot3 <- ggplot(data = states,aes(x = long, y = lat)) +
  geom_polygon(aes(fill= NStu, group = group, label = No.), color = "grey
", alpha = 0.8) +
  labs(list(title = "Student Population by State", x = "Longitude", y = "
Latitude")) +
  scale_fill_gradientn(colours = Popn, name = "student population", label
 = scales::comma) +
  coord_fixed(1.3)
colx <- arrange(colx, desc(NStu))
kable(colx[1:11,],col.names = c("State","College Count","White Population
", "Total Population","% Diversity"), digits = c(2,0,0,0))
```

State	College Count	White Population	Total Population	% Diversity
California	168	606129	2036580	0.30
Texas	97	444729	1136085	0.39
New York	107	390702	839066	0.47
Florida	59	367302	830486	0.44
Illinois	69	281050	520027	0.54
Arizona	30	225007	491966	0.46
Michigan	47	300442	441409	0.68

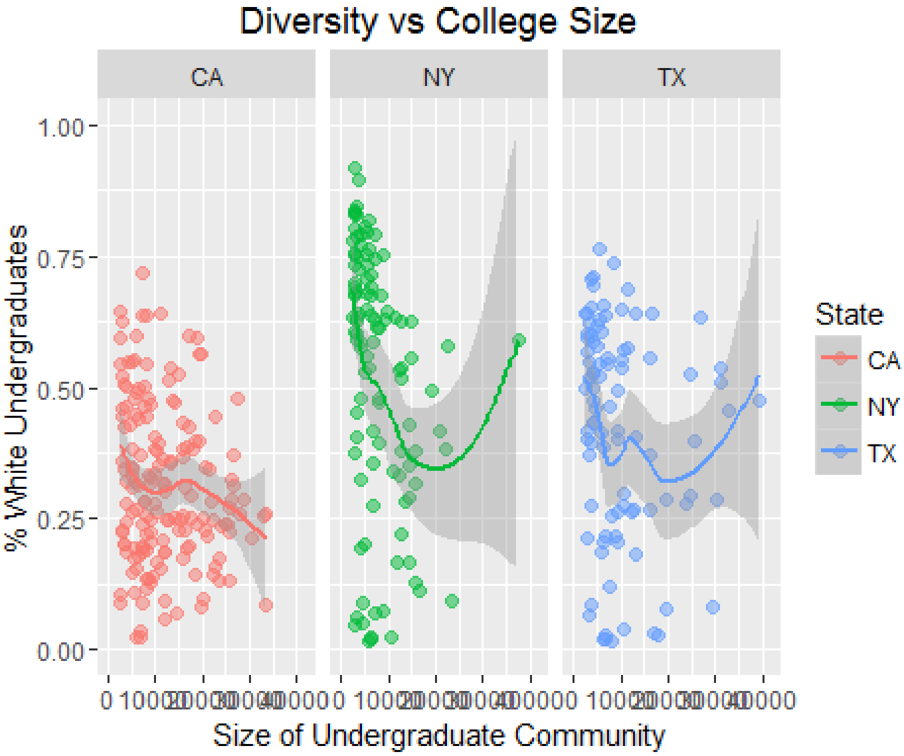
Pennsylvania	57	301399	439723	0.69
Ohio	45	291856	404869	0.72
Georgia	54	188240	379929	0.50
North Carolina	49	211731	363405	0.58

```
splot3
```



This table isn't very different from the first table generated that looked at Number of institutions. Finally, in an attempt to draw some conclusions, scatter plots will be constructed with University Size in the *X Axis* and *% White Undergraduates* in the Y axis.

```
# Racial Breakup Plot
coly = filter(Colleges2,STABBR == "CA"|STABBR == "TX"|STABBR == "NY") %>%
 mutate(STABBR = factor(STABBR))
ggplot(coly,aes(x = UGDS, y = UGDS_WHITE, col = STABBR) ) +
  geom_jitter(alpha = 0.5, size = 2)  + geom_smooth(aes(col = STABBR )) +
  labs(list(title = "Diversity vs College Size", x = "Size of Undergradua
te Community", y = "% White Undergraduates", col = "State")) +
  xlim(0,42000) + ylim(0,1) +
  facet_wrap(~STABBR, nrow = 1, ncol = 3)
```



The graphs indicate a tendency amongst **medium sized institutions(20,000 - 30,000)** to have more diverse student Communities, Leaving the opportunity to explore this trend further, with different data.