HBCU Report

Victor Huang

December 9, 2020

Data importing

More details regrading what does each datafile do is to be added.
For now the data sets that are used are:
HD2019: The data of all Universities in the 2019 IPEDS universe.
IC2019: Institution Characteristics for all universities.
C2019_a: A complete list of ratio compositions of universities registered in the IPEDS universe.
f1718_f1a - f1718_f3: Disclosed financial situations of higher institues in the United States. The distinctions are drawn based on accounting principles and purpose of operation (for-profit or public).

```
c2019_a<-read_dta("./C2019_A/dct_C2019_A.dta")
f1718_f1a<-read_dta("./F1718_F1A/dct_F1718_F1A.dta")
f1718_f2<-read_dta("./F1718_F2/dct_F1718_F2.dta")
f1718_f3<-read_dta("./F1718_F3/dct_F1718_F3.dta")
gr2019<-read_dta("./GR2019/dct_efia2019.dta")
gr2019_p<-read_dta("./GR2019_PELL_SSL/dct_efia2019.dta")
hd2019<- read_dta("./HD2019/dct_hd2019.dta")
ic2019<-read_dta("./IC2019/dct_ic2019.dta")
```

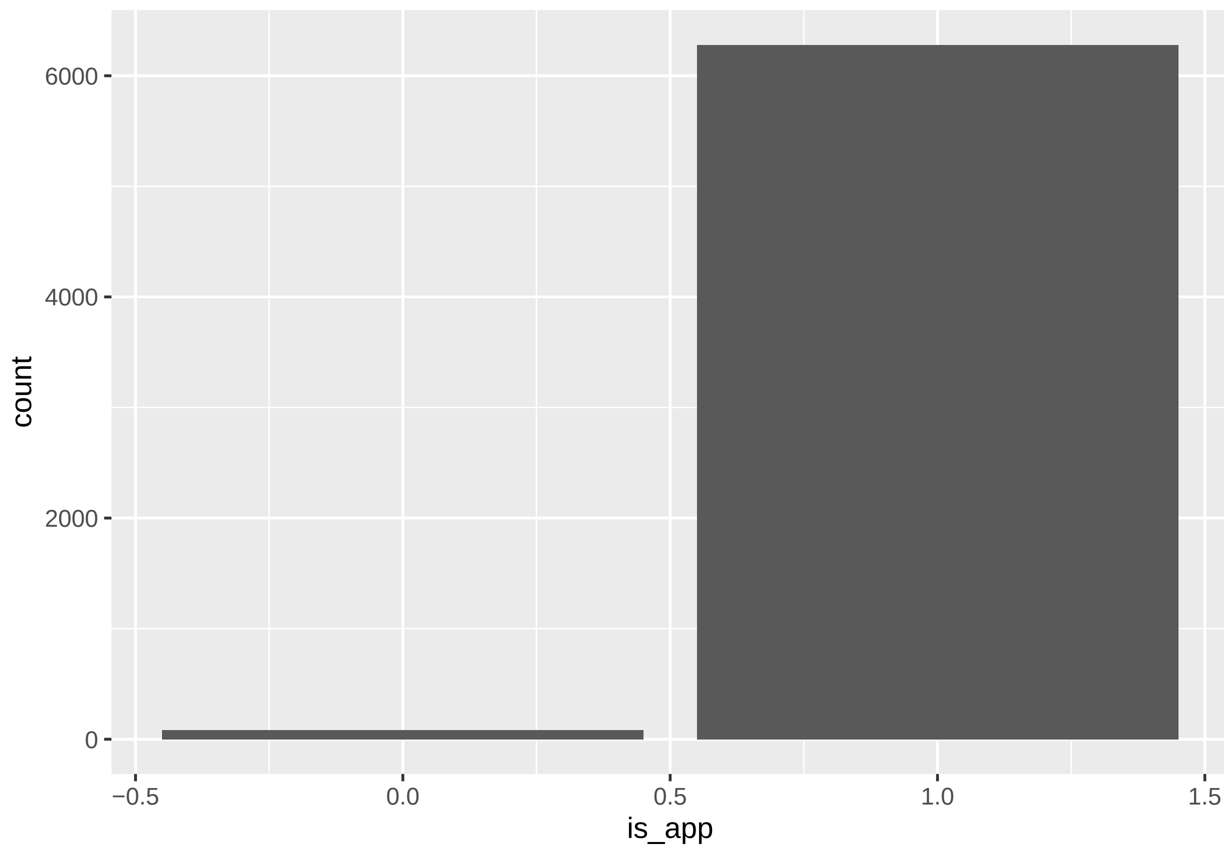
Tibble Generation

The tibble that is studied `joined_1` is created by joining `hd2019` and `ic2019` via unitid, the primary key assigned to each institutions.
The scope of institutions that we are interested in are institutions with four-year or longer programs. Moreover, institutions that did not report remedial services status or to which such reporting mechanism is not applicable are removed from the tibble as well. Since the number of these institutions are small, this removal is reasonable.
Additionally, for the purpose of linear model, I transformed the data in `hbcu` column which had 1 for yes and 2 for no to 1 for yes and 0 for no.

```
joined_1<-left_join(ic2019,hd2019)
```

```
## Joining, by = "unitid"
```

```
ggplot(joined_1 %>% mutate(is_app=(ifelse(stusrv1 %in% c(1,0),1,0)))) + geom_bar(aes(is_app))
```




```
hd2019_1 <- hd2019 %>% select(unitid,iclevel,hbcu)
ic2019_1 <- ic2019 %>% select(unitid,stusrv1)
joined_1<-inner_join(hd2019_1,ic2019_1)
```

```
## Joining, by = "unitid"
```

```
joined_1<-joined_1 %>% filter(iclevel==1) %>% filter(stusrv1 %in% c(0,1))
joined_1$hbcu<--(joined_1$hbcu-2)
```

Observations:

By applying OLS model on `hbcu` and `sturvs1`, the data suggested that on average, 61.37% of non_HBCU schools provide remedial services, while 76.41% of HBCU schools provide it. It is also note-worthy that in grand total, 61.84% of schools provided such service. Suggesting that though service-providing HBCUs are great in percentage, their numbers are relatively small such that the overall percentage is limited.

```
model<-lm(stusrv1~hbcu,data = joined_1)
summary(model)
```

```
##
## Call:
## lm(formula = stusrv1 ~ hbcu, data = joined_1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7640 -0.6137  0.3863  0.3863  0.3863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.613670   0.009277  66.151  < 2e-16 ***
## hbcu        0.150375   0.052265   2.877  0.00404 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4852 on 2823 degrees of freedom
## Multiple R-squared:  0.002924,   Adjusted R-squared:  0.002571
## F-statistic: 8.278 on 1 and 2823 DF,  p-value: 0.004043
```

```
summary(joined_1$stusrv1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.6184  1.0000  1.0000
```

Other Factors

This section is used to illustrate the respective ratio of remedial services in HBCU and non-HBCUs. In the following plot 0 means non_HBCU schools that has no remedial services, 1 stands for non_HBCU schools that has remedial services. While 2 stands for HBCUs that has no remedial services and 3 stands for HBCUs that have them.

It is clear that an exceedingly large portion of HBCUs have remedial services, but their relative smaller number may be source of errors.

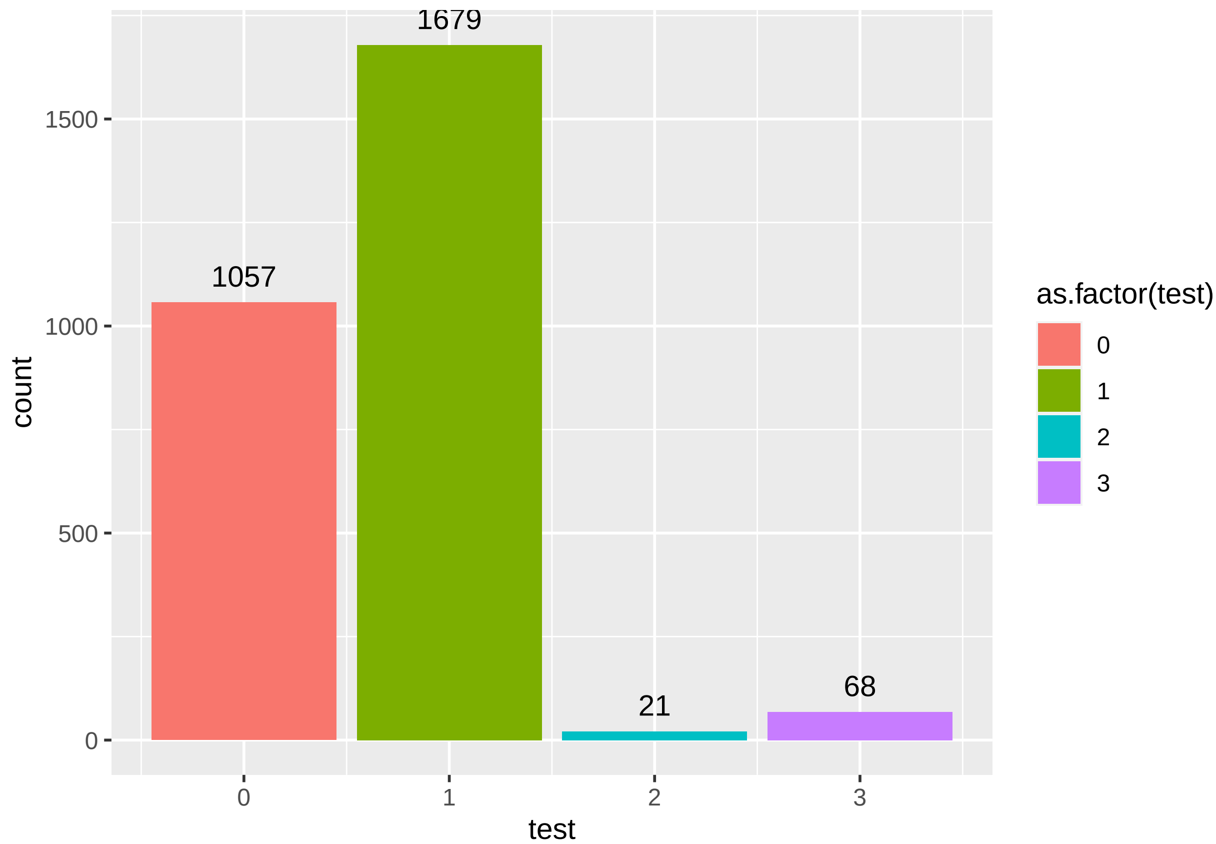
```
joined_1 %>% mutate(test=hbcu*2+stusrv1) %>% group_by(test) %>% summarise(n=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##    test     n
##   <dbl> <int>
## 1     0  1057
## 2     1  1679
## 3     2    21
## 4     3    68
```

```
ggplot(joined_1 %>% mutate(test=hbcu*2+stusrv1)) + geom_bar(aes(test,fill=as.factor(test))) +
  geom_text(data=joined_1 %>% mutate(test=hbcu*2+stusrv1) %>% group_by(test) %>% summarise(n=n()),aes(la
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Remarks:

What are some other issues that should be considered?
1: hedoskadesticity: don't know how to solve yet
2: quasi_experiment bias, `HBCUs` don't just turn into `non-HBCUs`, can be solved, potentially, by using difference in difference methods, although it would require extra data and previous inputs to justify this. (To be done).
3: endogeneity: what if HBCU has correlations with the error term? Find an instrument. (2SLS)
4: multi-variable test, what are some other factors that could help us explain this?

Updated model - Data cleaning and analysis

`Afri_ratio`: a double numeric variable showing the percentage of African students in the respective institutes. However this estimator, by itslef, is skewed and due to its continuous nature makes the estimator difficult to interpret. For ease of computing, I used a categorical variable `high_afri` as a proxy.
`high_afri`: A categorical variable that will be assigned to 1 if and only if the african student percentage in the particular school is above national average. Note: due to the skewedness and clustering of the data, schools that have above-average african students constitutes close to a quarter of the total schools.
`ph_gr` and `high_gr`: The former, like `Afri_ratio` is the numeric variable equal to the phell grant each school receives in millions. As in the case of `Afri_ratio` this variable is highly skewed and showed very extreme outliers. Thus, I chose an empirical value of 10 (ten million dollars) as the cut_off to generate the categorical variable `high_gr`.

```r
c2019_a_1 <- c2019_a %>% group_by(unitid) %>% filter(cipcode==99) %>% summarise_if(is.numeric,sum,na.rm=
  mutate('Afri_ratio'=cbkaat/ctotalt) %>% select(unitid,cbkaat,ctotalt,Afri_ratio)
f1718_f1a_1 <- f1718_f1a%>% select(unitid,'ph_gr'=f1e01)
f1718_f2_1 <- f1718_f2 %>% select(unitid,'ph_gr'=f2c01)
f1718_f3_1 <- f1718_f3 %>% select(unitid,'ph_gr'=f3c01)
f1718_pg<- rbind(f1718_f1a_1,f1718_f2_1,f1718_f3_1)
f1718_pg[is.na(f1718_pg)]<-0
joined_2<-right_join(f1718_pg,c2019_a_1) %>% right_join(joined_1)
```

```
## Joining, by = "unitid"
## Joining, by = "unitid"
```

```r
joined_2$ph_gr<- joined_2$ph_gr/1000000
joined_2<-joined_2 %>% mutate(high_afri=ifelse(Afri_ratio>=0.13424,1,0))
joined_2<-joined_2 %>% mutate(high_gr=ifelse(ph_gr>=10,1,0))
```

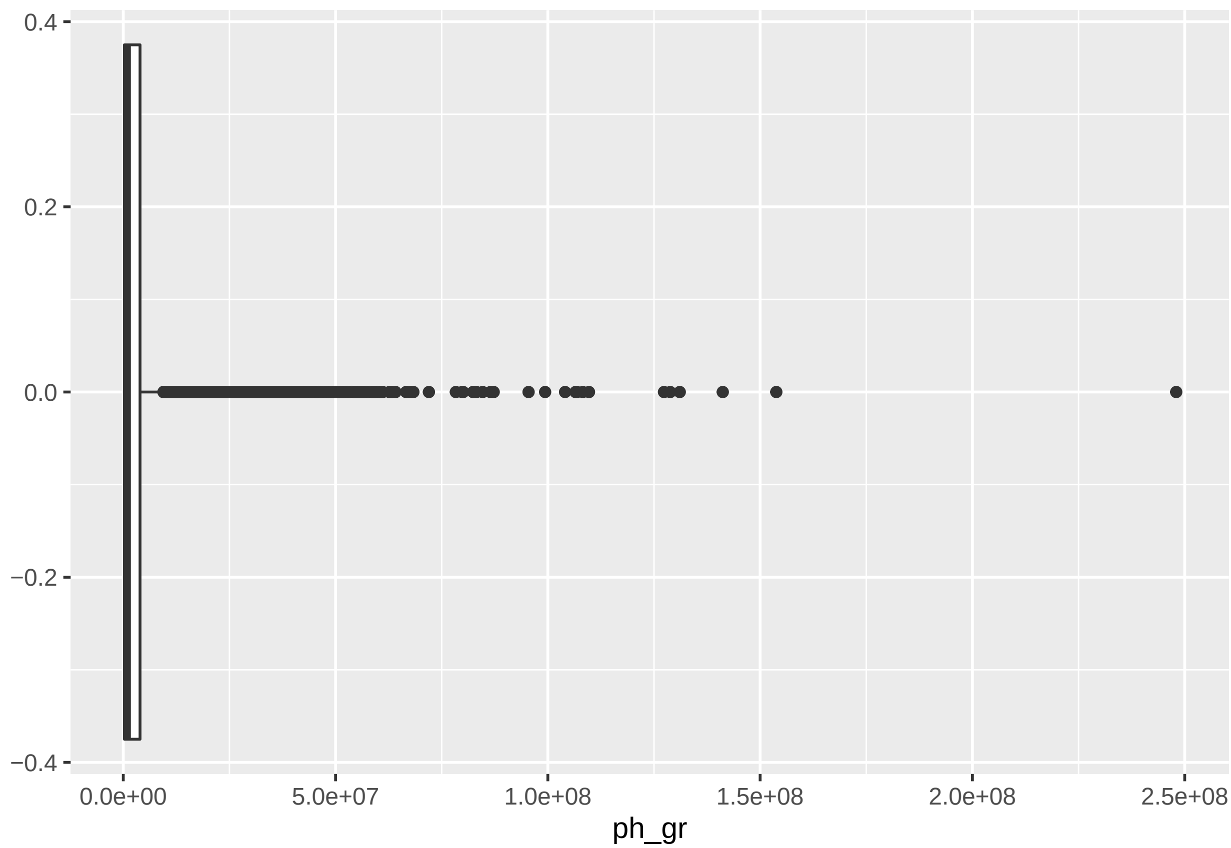
```r
summary(joined_2$Afri_ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.02542 0.06021 0.12518 0.13424 1.00000      28
```

```r
summary(joined_2$ph_gr)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.0000   0.5577   2.2648   6.8749   6.5960 248.0029      153
```

```r
ggplot(f1718_pg) + geom_boxplot(aes(ph_gr))
```



Justification

Both variable showed high skewedness and their mean being inflated by very large outliers, such as schools consists of 100 percent african-american students and schools that received 248 million phell grants. In terms of racial composition, I was interested in seeing if african student concentration above the majory (3rd) quarter of the data. Thus, I chose to mark the cutoff at the mean of the data.
On the other hand, for phell grant, I was interested in seeing how schools with abnormal funds, (i.e. high concentration of students who required federal aid, thus highly possible that could explain remedial service availability). In that case, I was interested in setting the cutoff at `normal` range, in this case, I chose an arbitrary number, 10 million as the cut-off, instead of the mean.

Updated model: Incorporating phell grant and ratio composition

Given the t value, it is largely possible that the `HBCU` status, in itself doe not contribute much to the availability of remedial services in schools. Rather, it is the racial composition and phell grant expenses that offered much of the variablity in the model.

```
model_2<-lm(stusrv1~hbcu+ph_gr+Afri_ratio,data=joined_2)
model_3<-lm(stusrv1~hbcu+high_gr+high_afri,data=joined_2)
summary(model_2)
```

```
##
## Call:
```

```
## lm(formula = stusrv1 ~ hbcu + ph_gr + Afri_ratio, data = joined_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3578 -0.5895  0.3344  0.3956  0.4487
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5758655  0.0126529  45.513  < 2e-16 ***
## hbcu        -0.0504804  0.0702704  -0.718    0.473
## ph_gr        0.0029180  0.0006754   4.320 1.62e-05 ***
## Afri_ratio   0.2658904  0.0673689   3.947 8.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4799 on 2659 degrees of freedom
##   (162 observations deleted due to missingness)
## Multiple R-squared:  0.0154, Adjusted R-squared:  0.01429
## F-statistic: 13.87 on 3 and 2659 DF,  p-value: 5.696e-09
```

```
summary(model_3)
```

```
##
## Call:
## lm(formula = stusrv1 ~ hbcu + high_gr + high_afri, data = joined_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8664 -0.5787  0.2859  0.4213  0.4213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.57874    0.01155  50.121  < 2e-16 ***
## hbcu         0.06418    0.05426   1.183    0.237
## high_gr      0.13534    0.02398   5.645 1.83e-08 ***
## high_afri    0.08813    0.02236   3.941 8.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4786 on 2659 degrees of freedom
##   (162 observations deleted due to missingness)
## Multiple R-squared:  0.02059,    Adjusted R-squared:  0.01948
## F-statistic: 18.63 on 3 and 2659 DF,  p-value: 5.87e-12
```

Ebdigeneity and 2SLS

This might provide further insights into the role of `HBCU` in the availability of remedial services.
Here, I am using `high_afri` as an instrument, note this choice is not necessarylity true and the subsequent model might be very biased as a result. In further research, if I were to find a better instrument I will update this part.

```r
# Evaluate the validity of high_afri as an instrument
coef1<-lm(stusrv1~hbcu+high_gr,data=joined_2)[[1]]%>% as.numeric
joined_2<-joined_2 %>% mutate(error=stusrv1-coef1[1]-coef1[2]*hbcu-coef1[3]*high_gr)
cov1<-sum(joined_2$hbcu*joined_2$high_afri/length(joined_2$hbcu),na.rm = T)/(sd(joined_2$hbcu,na.rm = T)
cov2<-sum(joined_2$error*joined_2$high_afri/length(joined_2$hbcu),na.rm = T)/(sd(joined_2$error,na.rm =
print(c(cov1,cov2))
```

```
## [1] 0.40706411 0.07003777
```

As is shown here, after accounting for phell grant, high_afri is mildly corrolated with the independent variable and relatively uncorrelated with the error term. It can be used as an instrument.

```r
coef<-lm(hbcu~high_afri+high_gr,data=joined_2)[[1]] %>% as.numeric()
joined_2<- joined_2 %>% mutate(hbcu_bar=coef[2]*high_afri+coef[3]*high_gr+coef[1])
model_4<-lm(stusrv1~hbcu_bar+high_gr,data = joined_2)
summary(model_4)
```

```
##
## Call:
## lm(formula = stusrv1 ~ hbcu_bar + high_gr, data = joined_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8110 -0.5787  0.2852  0.4213  0.4213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.57949    0.01147  50.543  < 2e-16 ***
## hbcu_bar     0.76041    0.16815   4.522 6.39e-06 ***
## high_gr      0.12707    0.02412   5.268 1.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4787 on 2660 degrees of freedom
##   (162 observations deleted due to missingness)
## Multiple R-squared:  0.02007,    Adjusted R-squared:  0.01933
## F-statistic: 27.24 on 2 and 2660 DF,  p-value: 1.946e-12
```

In the updated model, hbcu_bar is no longer categorical but is numeric as it is the predicted value from `high_afri` and `high_gr`. Consequently, the sum of all values might exceed one in the case. However, it can be seen from the model that, taken endogeneity into account (racial composition), `hbcu` status plays an important role in the availability of remedial services.

Notes to self and some other questions to consider:

Note to self, remedial services is but a categorical value with have or not have. Try to find ways to account for quasi-experiment problems.
Find the expenses for remedial services for each school (Hopefully.)
Outliers with Phell Grant, see if I can find more about them. Also, the university that has way more phell grant than any other schools is: University of Phoenix-Arizona. Question: Why?