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<-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")
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")
```

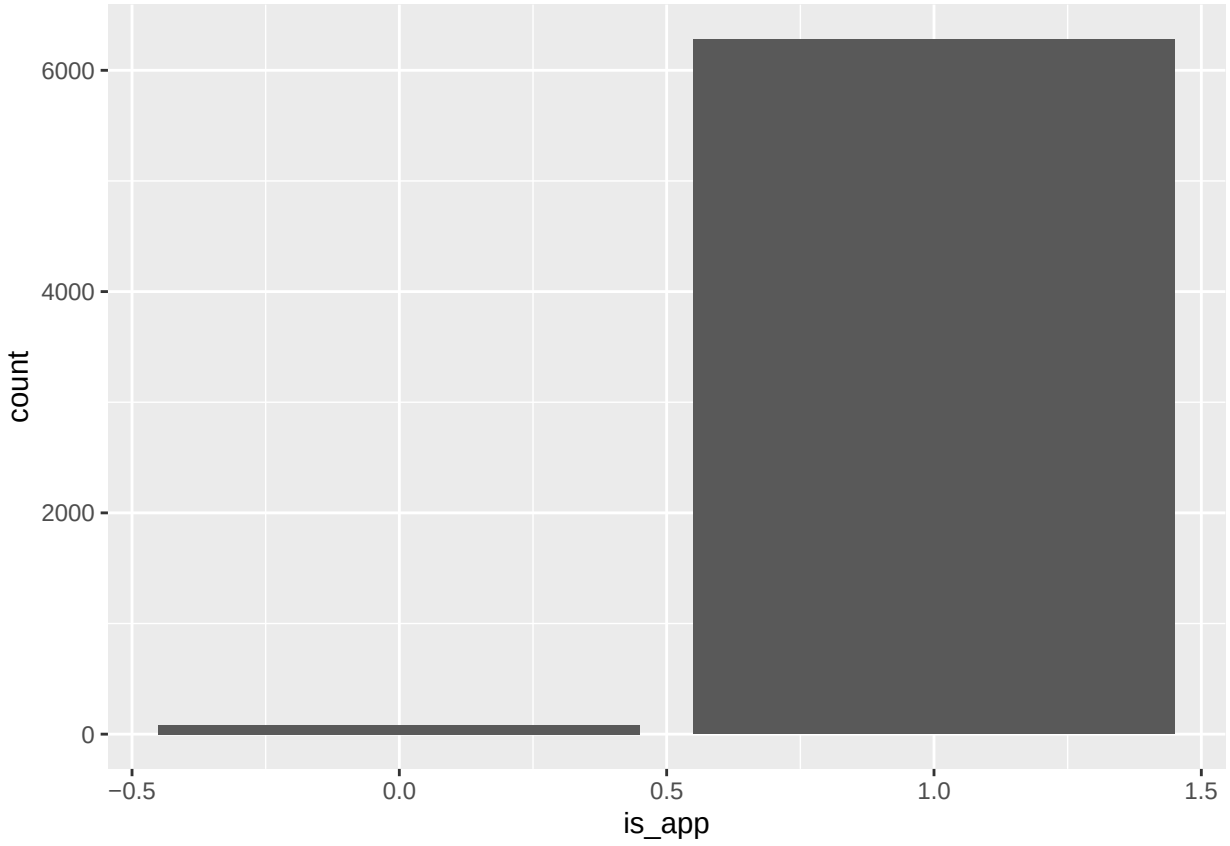
Tibble Generation

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<-inner_join(hd2019,ic2019)
```

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

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




```
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.

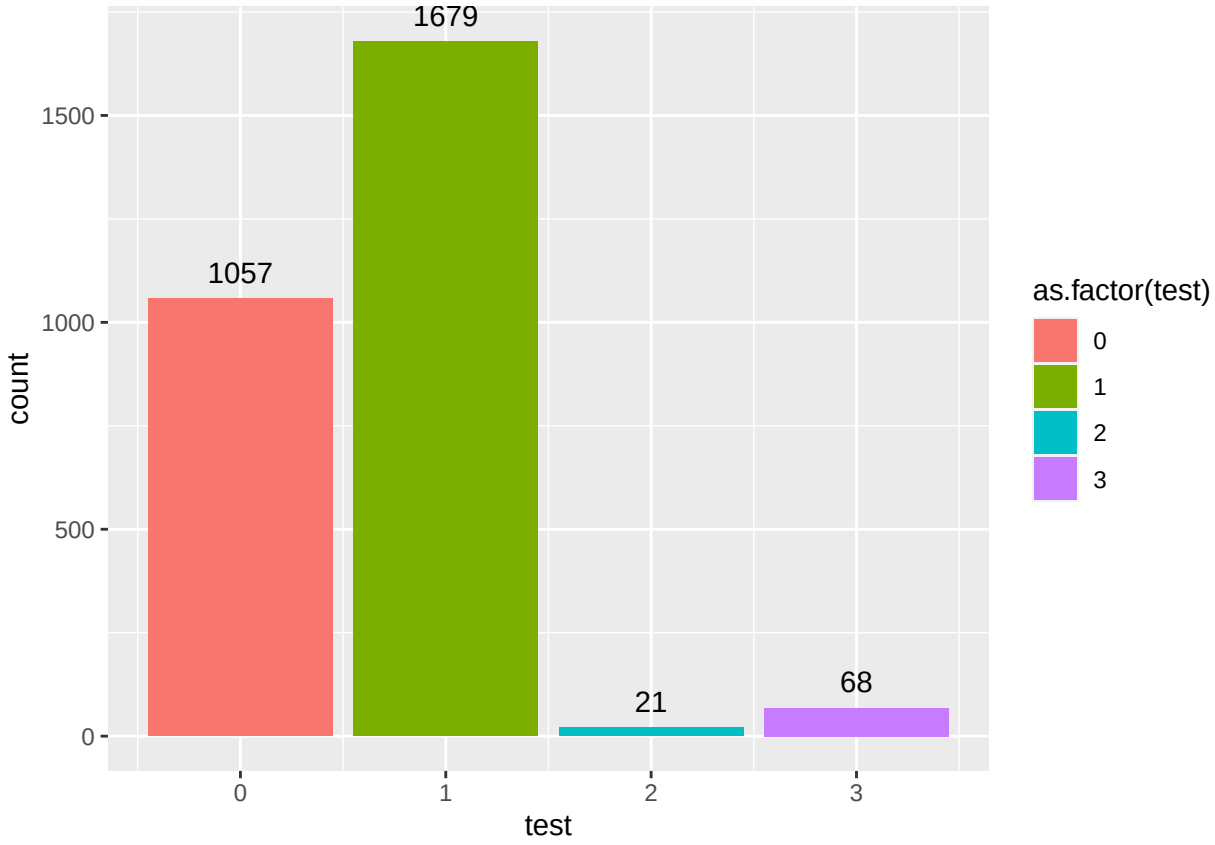
```
joined_1 %>% mutate(test=hbcu*2+stusrv1) %>% group_by(test) %>% summarise(n=n())
```

```
## `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)
```


1679
1057
21
68
1500
1000
500
0
count
0
1
2
3
test
as.factor(test)
0
1
2
3