DataAppendix

Zainab

11/7/2017

```
str(collegedata, give.attr = FALSE)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2202 obs. of  15 variables:
##  $ super_opeid       : int  2665 7273 2688 7022 1140 2693 2165 2791 2838 2696 ...
##  $ name              : chr  "Vaughn College Of Aeronautics And Technology" "CUNY Bernard M. Baruch C
##  $ czname            : chr  "New York" "New York" "New York" "New York" ...
##  $ state             : chr  "NY" "NY" "NY" "NY" ...
##  $ par_median        : int  30900 42800 35500 32500 36600 41800 83300 68600 73600 33500 ...
##  $ k_median          : int  53000 57600 48500 40700 43000 45200 112700 60700 60100 37000 ...
##  $ par_q1            : num  36.5 27.6 32.5 36.7 33.1 ...
##  $ par_top1pc        : num  0.12 0.559 0.234 0 0.156 ...
##  $ kq5_cond_parq1    : num  44.8 46.8 36 27.9 29.9 ...
##  $ ktop1pc_cond_parq1: num  1.7666 2.5568 1.4087 0.1896 0.0836 ...
##  $ mr_kq5_pq1        : num  16.36 12.94 11.72 10.24 9.92 ...
##  $ mr_ktop1_pq1      : num  0.6444 0.7065 0.4585 0.0696 0.0277 ...
##  $ trend_parq1       : num  -8 -9.19 -9.8 -5.73 -13.31 ...
##  $ trend_bottom40    : num  -5.75 -12.3 -13.88 -9.07 -14.92 ...
##  $ count             : num  208 1083 582 468 1180 ...
```

```
newData = inner_join(collegeChardata, collegedata, by="name")
# x variables: tier, hbcu, grad_rate_150_p_, sticker_price_, black_share_fall, alien_share_fall, pct_st
# y variable: kq5_cond_parq1
newData%>%
  select(c(tier,hbcu,grad_rate_150_p_2013,sticker_price_2013,black_share_fall_2000,alien_share_fall_200
  cor()
```

```
##                              tier        hbcu grad_rate_150_p_2013
## tier                   1.00000000 -0.09872547                   NA
## hbcu                  -0.09872547  1.00000000                   NA
## grad_rate_150_p_2013           NA          NA                    1
## sticker_price_2013             NA          NA                   NA
## black_share_fall_2000          NA          NA                   NA
## alien_share_fall_2000          NA          NA                   NA
## pct_stem_2000                  NA          NA                   NA
## par_median            -0.65430775 -0.17381911                   NA
## count                 -0.13639892 -0.03638790                   NA
## endowment_pc_2000              NA          NA                   NA
##                       sticker_price_2013 black_share_fall_2000
## tier                                  NA                    NA
## hbcu                                  NA                    NA
## grad_rate_150_p_2013                  NA                    NA
## sticker_price_2013                     1                    NA
## black_share_fall_2000                 NA                     1
## alien_share_fall_2000                 NA                    NA
## pct_stem_2000                         NA                    NA
## par_median                            NA                    NA
## count                                 NA                    NA
## endowment_pc_2000                     NA                    NA
```

```
##                         alien_share_fall_2000 pct_stem_2000 par_median
## tier                                       NA            NA -0.6543077
## hbcu                                       NA            NA -0.1738191
## grad_rate_150_p_2013                       NA            NA         NA
## sticker_price_2013                         NA            NA         NA
## black_share_fall_2000                      NA            NA         NA
## alien_share_fall_2000                       1            NA         NA
## pct_stem_2000                              NA             1         NA
## par_median                                 NA            NA  1.0000000
## count                                      NA            NA  0.1135126
## endowment_pc_2000                          NA            NA         NA
##                              count endowment_pc_2000
## tier                    -0.1363989                NA
## hbcu                    -0.0363879                NA
## grad_rate_150_p_2013            NA                NA
## sticker_price_2013              NA                NA
## black_share_fall_2000           NA                NA
## alien_share_fall_2000           NA                NA
## pct_stem_2000                   NA                NA
## par_median               0.1135126                NA
## count                    1.0000000                NA
## endowment_pc_2000               NA                 1
```

```
str(newData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2199 obs. of  63 variables:
##  $ super_opeid.x                  : int  30955 3537 1541 7531 1345 2666 2860 2234 11484 31275 ...
##  $ name                           : chr  "ASA Institute Of Business & Computer Technology" "Abilene
##  $ region                         : int  1 3 3 4 4 1 1 2 2 3 ...
##  $ state.x                        : chr  "NY" "TX" "GA" "CA" ...
##  $ fips                           : int  36 48 13 6 8 36 36 26 55 51 ...
##  $ cz                             : int  19400 32501 8503 37800 34805 19400 18600 11500 24100 2000
##  $ czname.x                       : chr  "New York" "Abilene" "Valdosta" "San Francisco" ...
##  $ cfips                          : int  36047 48441 13277 6075 8003 36059 36113 26091 55079 51810
##  $ county                         : chr  "Kings" "Taylor" "Tift" "San Francisco" ...
##  $ zip                            : int  11201 79699 31793 94105 81101 11530 12804 49221 53228 2346
##  $ tier                           : int  11 6 7 10 5 6 9 6 11 11 ...
##  $ tier_name                      : chr  "Two-year for-profit" "Selective private" "Nonselective fou
##  $ type                           : int  3 2 1 3 1 2 1 2 3 3 ...
##  $ iclevel                        : int  2 1 1 1 1 1 2 1 2 2 ...
##  $ public                         : int  0 0 1 0 1 0 1 0 0 0 ...
##  $ barrons                        : int  999 3 999 999 4 4 999 4 999 999 ...
##  $ exp_instr_pc_2000              : num  1961 4572 3533 2569 1037 ...
##  $ exp_instr_pc_2013              : num  3643 7046 3301 5040 5939 ...
##  $ multi                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hbcu                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ flagship                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ipeds_enrollment_2013          : int  4711 3727 3394 10508 2284 5040 4230 1646 355 824 ...
##  $ ipeds_enrollment_2000          : int  1432 4739 2630 5995 8188 5908 3151 1081 111 273 ...
##  $ sticker_price_2013             : num  12298 29450 3394 19740 8014 ...
##  $ sticker_price_2000             : num  NA 10910 1664 15090 2186 ...
##  $ grad_rate_150_p_2013           : num  0.316 0.566 NA 0.33 0.261 ...
##  $ grad_rate_150_p_2002           : num  0.592 0.545 0.227 0.322 0.323 ...
##  $ avgfacsal_2013                 : num  4378 5508 4752 7062 5986 ...
##  $ avgfacsal_2001                 : num  NA 48040 43470 50651 44088 ...
```

```
##  $ sat_avg_2013                    : num  NA 1075 925 NA 984 ...
##  $ sat_avg_2001                    : num  NA 1100 NA NA NA ...
##  $ scorecard_netprice_2013         : num  22011 20836 7887 28224 14705 ...
##  $ scorecard_rej_rate_2013         : num  NA 0.511 0.291 NA 0.474 ...
##  $ scorecard_median_earnings_2011  : num  26400 40200 31900 36100 33200 50600 32300 36500 22000 3530
##  $ endowment_pc_2000               : num  NA NA 1594 NA 1027 ...
##  $ exp_instr_2012                  : num  12291059 30223039 9604920 87856929 18314567 ...
##  $ exp_instr_2000                  : num  1417345 20684050 9627084 17809356 8925393 ...
##  $ asian_or_pacific_share_fall_2000: num  0.11802 0.00733 0.00494 0.1564 0.0125 ...
##  $ black_share_fall_2000           : num  0.0817 0.0577 0.157 0.024 0.0351 ...
##  $ hisp_share_fall_2000            : num  0.0733 0.0567 0.0167 0.2059 0.2755 ...
##  $ alien_share_fall_2000           : num  0.0712 0.0399 0.0137 0.2718 0.0024 ...
##  $ pct_arthuman_2000               : num  0 10.8 0 96.1 10.9 ...
##  $ pct_business_2000               : num  6.6 22.5 4.1 0 24 ...
##  $ pct_health_2000                 : num  11.43 5.06 12.07 0 0 ...
##  $ pct_multidisci_2000             : num  0 9.72 47.15 0 33.43 ...
##  $ pct_publicsocial_2000           : num  0 8.79 6.38 0 0 ...
##  $ pct_stem_2000                   : num  82 11.3 29.6 0 13.5 ...
##  $ pct_socialscience_2000          : num  0 31.69 0 3.91 18.18 ...
##  $ pct_tradepersonal_2000          : num  0 0.133 0.683 0 0 ...
##  $ super_opeid.y                   : int  30955 3537 1541 7531 1345 2666 2860 2234 11484 31275 ...
##  $ czname.y                        : chr  "New York" "Abilene" "Valdosta" "San Francisco" ...
##  $ state.y                         : chr  "NY" "TX" "GA" "CA" ...
##  $ par_median                      : int  29000 101000 66000 92300 67200 96300 69900 88500 78100 572
##  $ k_median                        : int  19700 40100 32500 27400 34100 50700 30300 40300 24300 3390
##  $ par_q1                          : num  44.36 5.24 15.46 9.35 12.92 ...
##  $ par_top1pc                      : num  0.0416 2.3049 0.252 3.5681 0.3147 ...
##  $ kq5_cond_parq1                  : num  4.52 27.39 9.61 17.49 14.58 ...
##  $ ktop1pc_cond_parq1              : num  0.00426 3.80796 0.64402 1.45513 0.024 ...
##  $ mr_kq5_pq1                      : num  2 1.44 1.49 1.64 1.88 ...
##  $ mr_ktop1_pq1                    : num  0.00189 0.19969 0.09954 0.13609 0.0031 ...
##  $ trend_parq1                     : num  -4.4635 -1.5127 0.8082 2.8897 0.0724 ...
##  $ trend_bottom40                  : num  -10.206 -5.103 7.144 7.715 0.702 ...
##  $ count                           : num  275 839 679 496 393 ...
```

```
#newData$tier <- as.factor(newData$tier)

favstats(~tier, data=newData)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    1  6      6  9  12 7.101864 2.247553 2199       0
```

This is a categorical variable that encodes the selectivity of a college as following: 1 = Ivy Plus 2 = Other elite schools (public and private) 3 = Highly selective public 4 = Highly selective private 5 = Selective public 6 = Selective private 7 = Nonselective 4-year public 8 = Nonselective 4-year private not-for-profit 9 = Two-year (public and private not-for-profit) 10 = Four-year for-profit 11 = Two-year for-profit 12 = Less than two year schools of any type

From favtstats, we can see that the min and max are 1 and 12 respectively which makes sense. Even though it is a categorical variable, we can make some sense of the mean being close to 7 which stands for Nonselective 4 year public schools. Just out of curiousity, I found out that Smith falls under 4 which is highly selective private schools. This variable has no missing observations which is good.

```
favstats(~hbcu, data=newData)
```

```
##  min Q1 median Q3 max       mean        sd    n missing
```

```
##    0  0      0  0   1 0.02455662 0.1548047 2199       0
```

This is an indicator variable that tells us whether a particular observation belongs to a group of schools known as Historically Black Colleges and Universities (HBCUs). The value is 1 if it is a part of that group and 0 otherwise. The min and max make sense and we have no missing observations.

```
favstats(~grad_rate_150_p_2013, data=newData)
```

```
##         min        Q1    median        Q3 max      mean        sd    n
##  0.02020202 0.2546648 0.4421965 0.6182524   1 0.4493151 0.2259421 2027
##  missing
##      172
```

This explanatory variable represents the percentage of students graduating within 150 percent of normal time in 2013. We have 172 missing values and a majority of them can be explained by the fact that this data is available only for four and two-year institutions (so only tiers 1 through 11). At this point, we have decided that we only want to look at 2-year and 4-year institutions because other variables such as sticker prices and variables about the racial makeup of schools in also not available for tier 12 schools. I will filter out tier 12 and look at this variable again:

```
newData = newData %>%
  filter(tier > 11)

favstats(~grad_rate_150_p_2013, data=newData)
```

```
##        min        Q1    median        Q3       max      mean        sd n
##  0.4255319 0.4491142 0.4709266 0.6491508 0.7727273 0.5522451 0.1371693 7
##  missing
##       39
```

```
# xyplot(kq5_cond_parq1~grad_rate_150_p_2013, data=newData)
```

There are still 39 missing observations and we will look at what to do with them later on in the analysis when he have identified missing values for all variables.

```
favstats(~scorecard_netprice_2013, data=newData)
```

```
##       min       Q1 median    Q3   max     mean       sd  n missing
##  4415.778 13095.47  17631 19793 29950 16697.15 5124.463 45       1
```

This variable represents Net Cost of Attendance for the Bottom 20% Income Quintile in 2013 from College Scorecard in dollars. College Scorecard is the US Department of Education's datastore. We have only one missing observation which is encouraging. The minimum is around 4400 dollars and the maximum is around 29900 dollars which makes sense.

```
favstats(~black_share_fall_2000, data=newData)
```

```
##  min         Q1    median        Q3       max      mean        sd  n
##    0 0.05460751 0.1095406 0.2527716 0.8082192 0.1806284 0.1894258 45
##  missing
##        1
```

```
favstats(~alien_share_fall_2000, data=newData)
```

```
##  min Q1 median Q3        max        mean          sd  n missing
##    0  0      0  0 0.02439024 0.002630138 0.006222074 45       1
```

```
favstats(~pct_stem_2000, data=newData)
```

```
##  min Q1 median       Q3 max     mean       sd  n missing
##    0  0      0 1.025641 100 13.20007 28.84338 45       1
```

```
favstats(~par_median, data=newData)
```

```
##    min    Q1 median    Q3   max     mean       sd  n missing
##  25100 39375  49700 65125 93600 52630.43 17901.67 46       0
```

```
favstats(~count, data=newData)
```

```
##  min     Q1 median      Q3   max     mean       sd  n missing
##   56 94.875 132.75 189.375 941.5 186.8623 174.4908 46       0
```

```
favstats(~exp_instr_2012, data=newData)
```

```
##     min      Q1  median      Q3      max    mean      sd  n missing
##  680873 1555718 2292261 4207983 51893300 4866227 8296887 45       1
```

Y variable:

```
favstats(~kq5_cond_parq1, data=newData)
```

```
##  min       Q1   median       Q3      max     mean       sd  n missing
##    0 1.101693 5.141678 8.269325 29.57651 6.044979 6.737263 46       0
```

This is the explanatory variable.

Most Pressing Data CLeaning

- We are worried about the 39 missing observations in the grad_rate_150_p_2013 variable.