DataAppendix

Zainab, Rutendo and Margaret

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 Co
##  $ 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")
```

Structure and Names

There are 10 variables we are interested in in this dataset and these are:

```
#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:

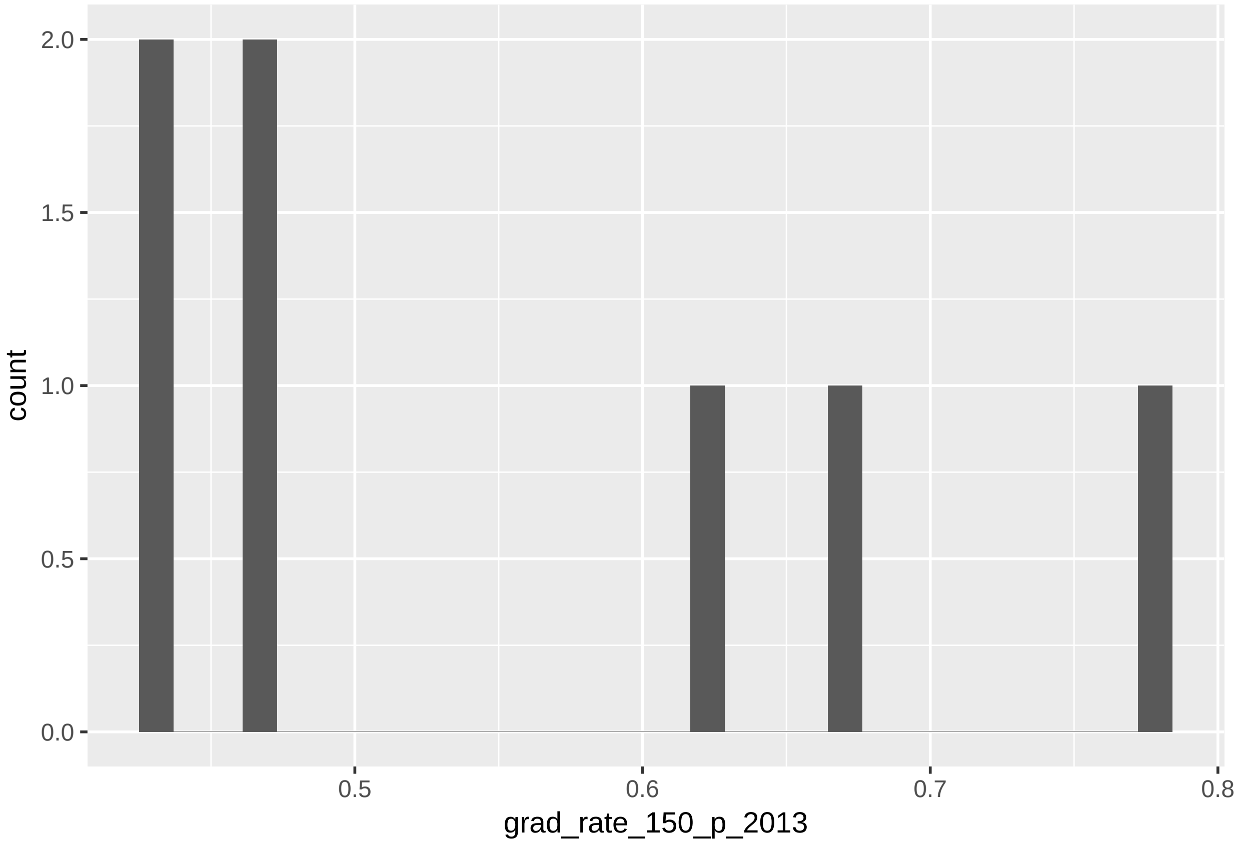
```
newData = newData %>%
  filter(tier > 11)
favstats(~grad_rate_150_p_2013, data=newData)
```

```
##        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
```

```
c <- ggplot(newData, aes(grad_rate_150_p_2013))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 39 rows containing non-finite values (stat_bin).
```



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.

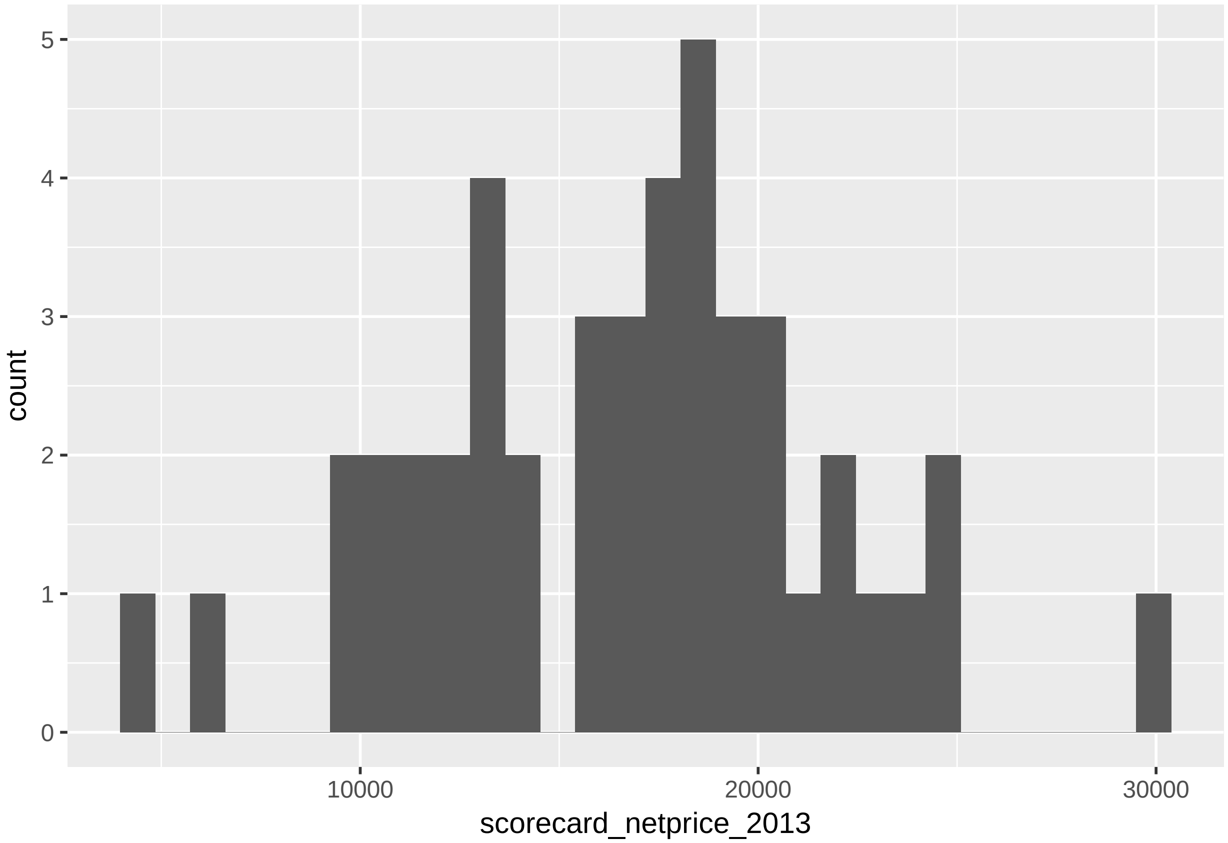
```
favstats(~scorecard_netprice_2013, data=newData)
```

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

```
c <- ggplot(newData, aes(scorecard_netprice_2013))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



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.

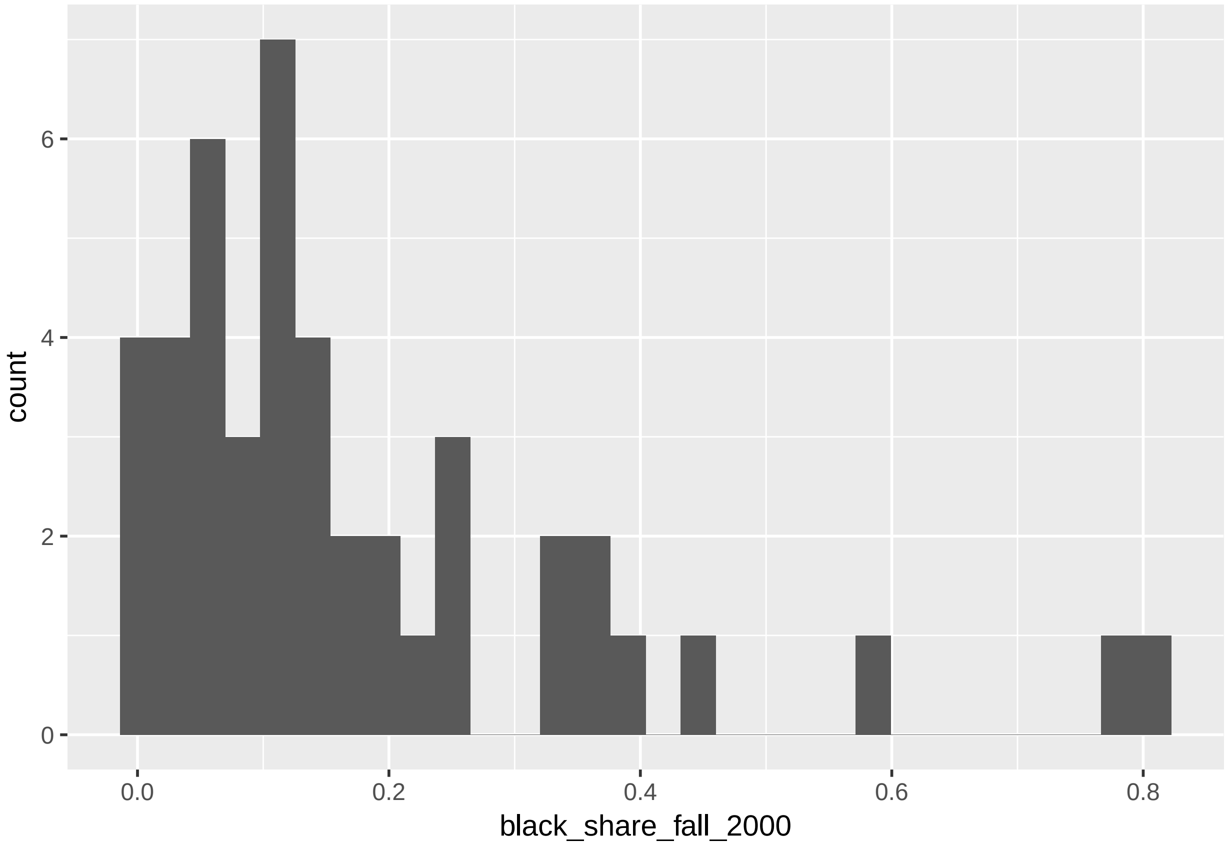
```
favstats(~black_share_fall_2000, data=newData)
```

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

```
c <- ggplot(newData, aes(black_share_fall_2000))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



The black__share__fall__200 variable shows us the share of Black undergraduate students in the fall 2000. There is 1 missing observation that could draw concern. Other than that the max, .8082192, seems a bit high but not impossible and the min is 0 which seems unreasonable.

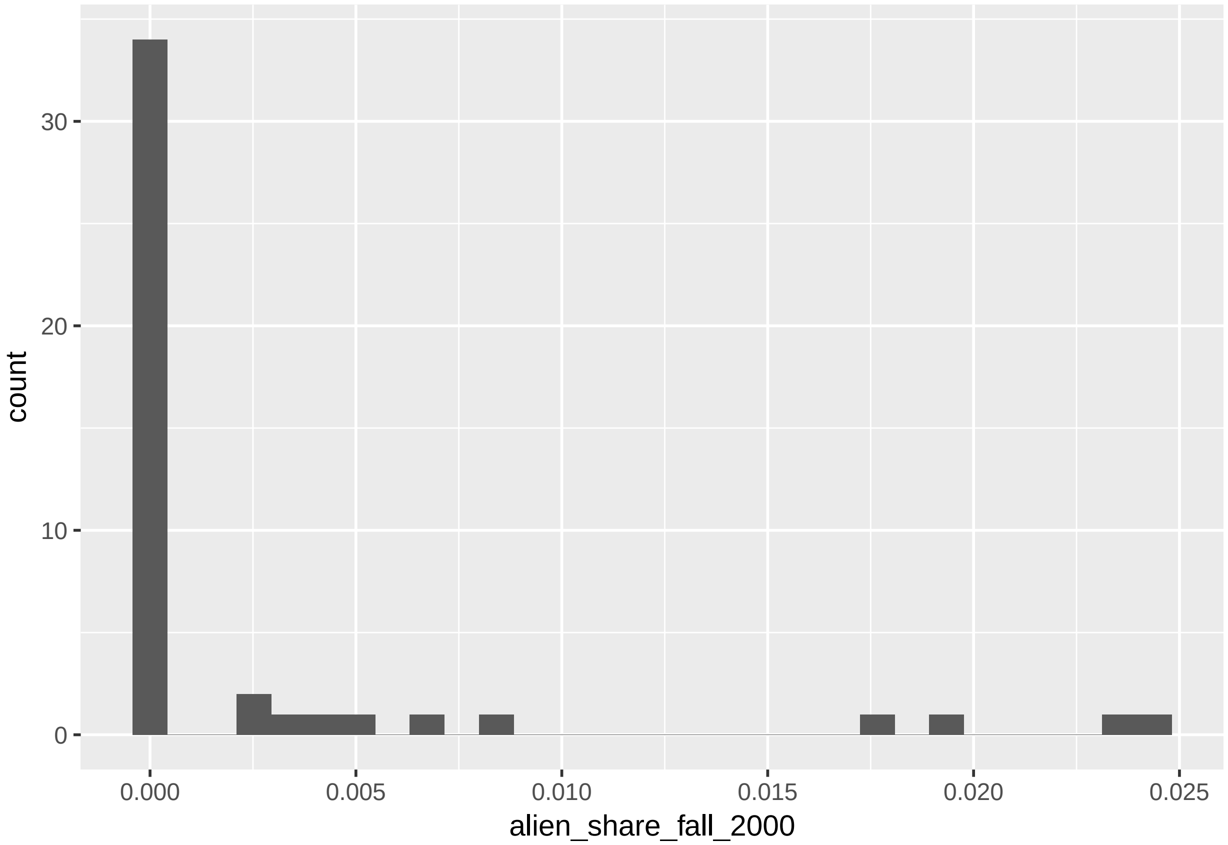
```
favstats(~alien_share_fall_2000, data=newData)
```

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

```
c <- ggplot(newData, aes(alien_share_fall_2000))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



The explanatory variable of alien__share__fall__2000 represents share of non-resident alien undergraduate students in the fall 2000. Again there is 1 missing observation that will have to be managed later. The minimum for the data set is 0 students and the maximum is around 0.02439 share of stduents, both of these are reasonable for the variable.

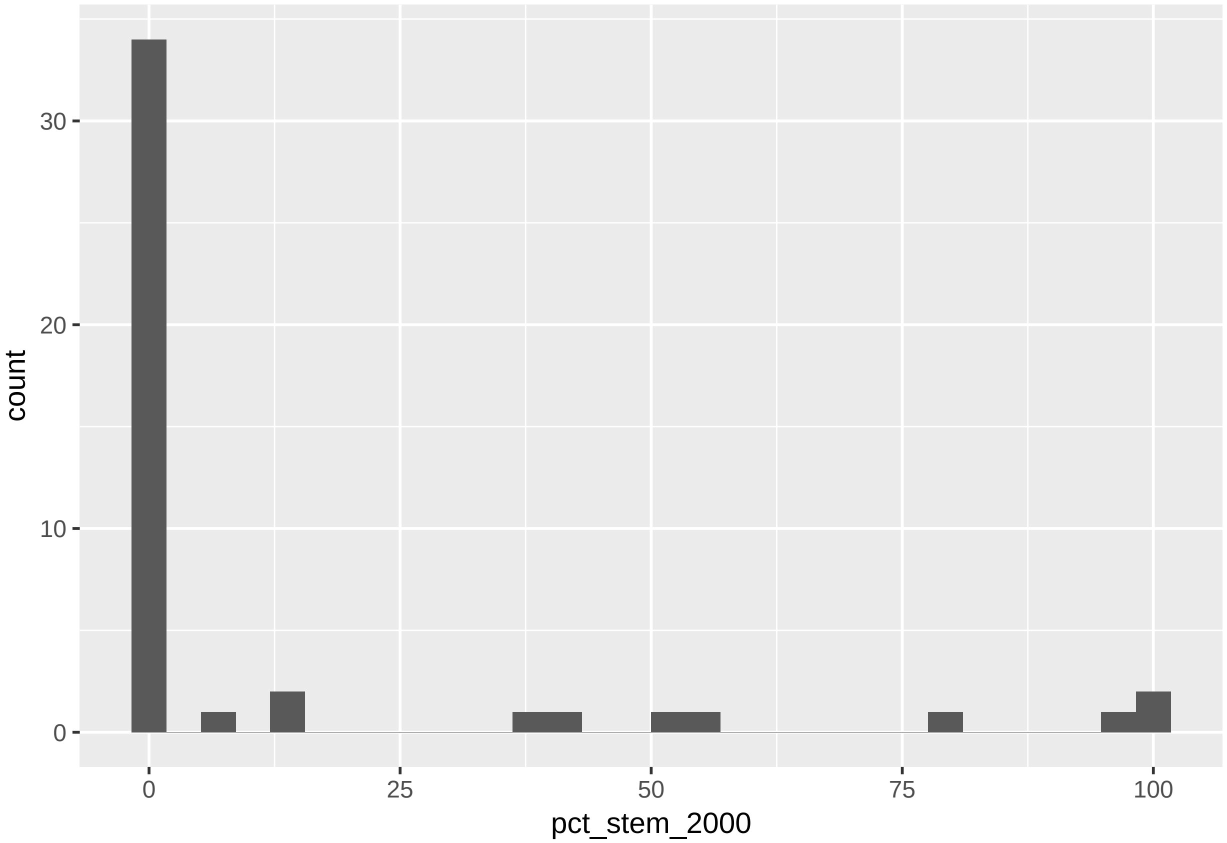
```
favstats(~pct_stem_2000, data=newData)
```

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

```
c <- ggplot(newData, aes(pct_stem_2000))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



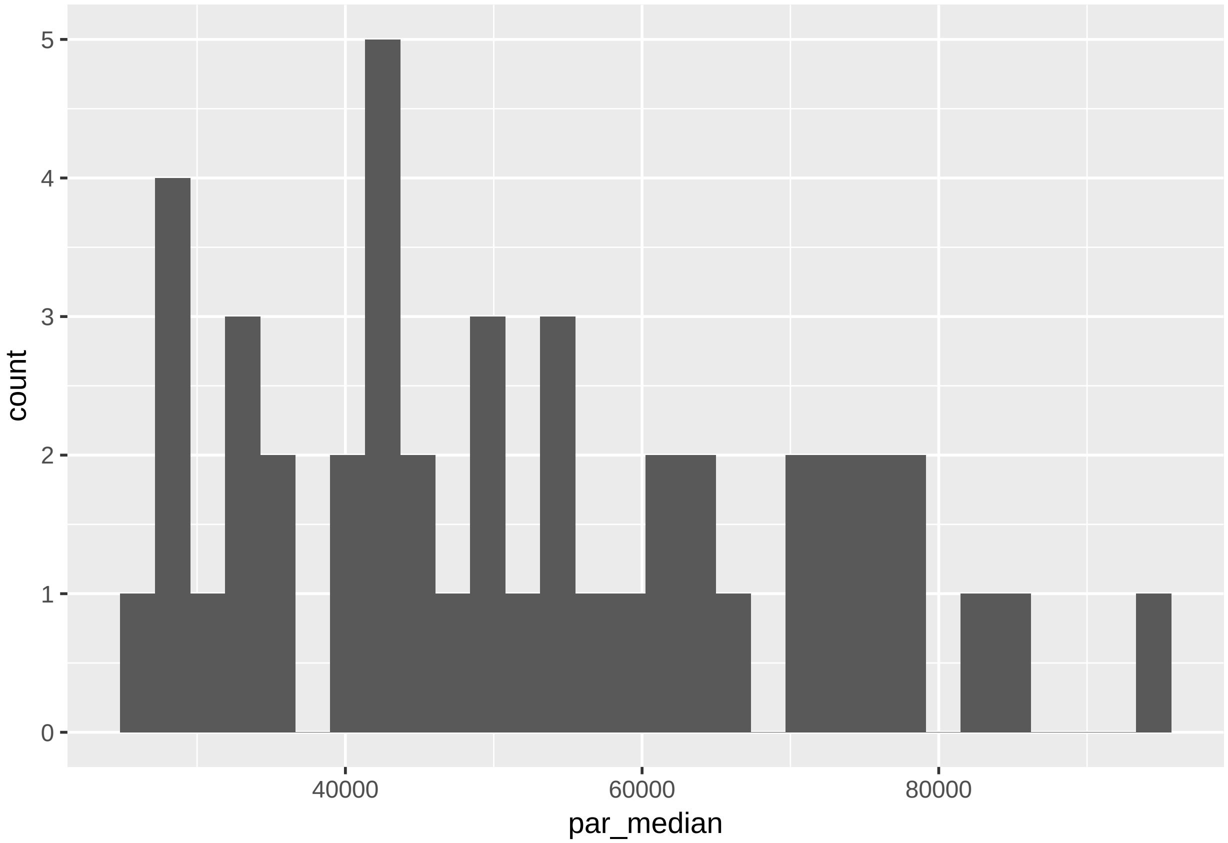
This variable prvides us with the information for the STEM Major Share in 2000. We are still missing one observation. What is odd about this varible is that the median and the min are both zero and the max is a hundred which is very odd and concerning, as it does not fit how the other shares are shown in that they are whole numbers not decimals.

```
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
```

```
c <- ggplot(newData, aes(par_median))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The par_median variable provides us information on the median parent household income for a college. Thankfully there are no missing varibles and both the max and min appear resonable.

```
favstats(~endowment_pc_2000, data=newData)
```

```
##       min       Q1   median       Q3      max     mean sd n missing
##  71.86159 71.86159 71.86159 71.86159 71.86159 71.86159 NA 1      45
```

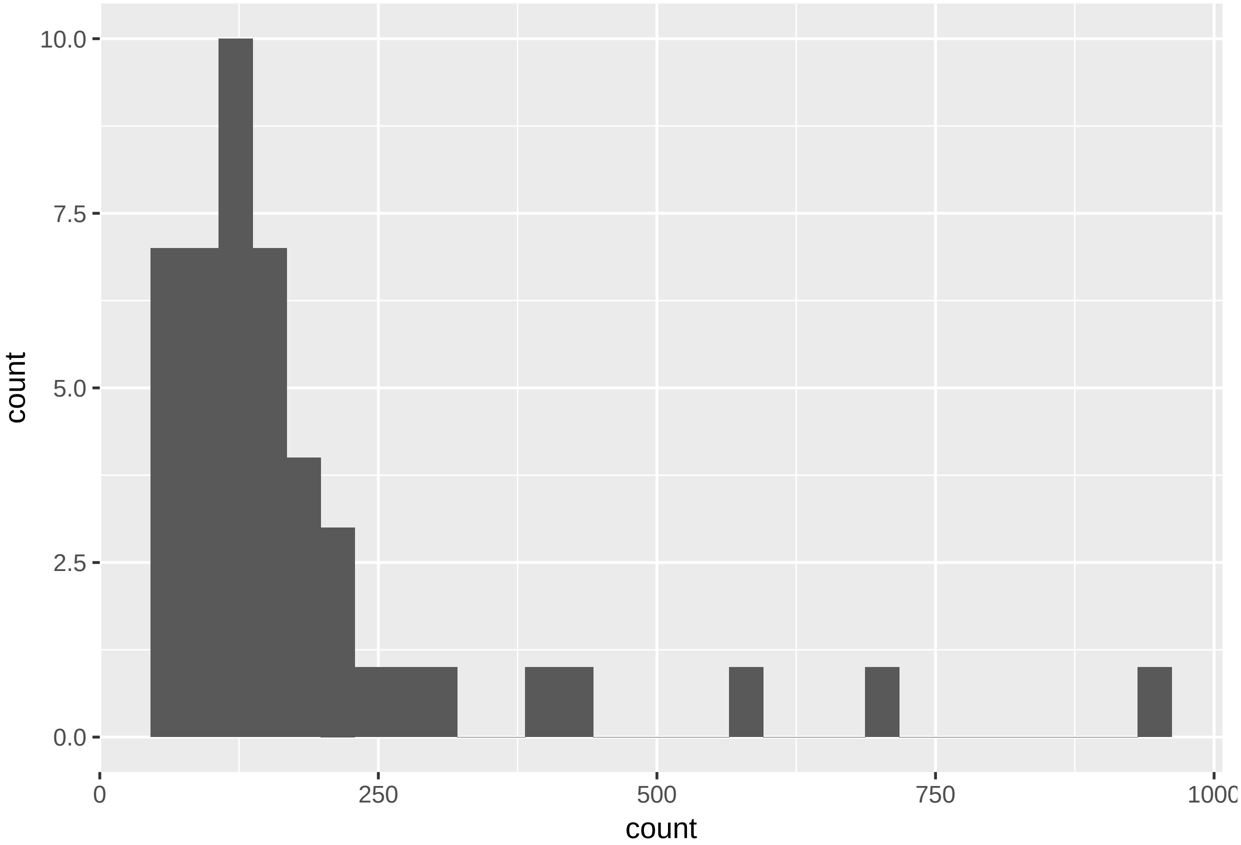
This variable represents the endowment assets per student in 2000. The minimum number of endowments is .634848, and the maximum is 71.86159. There are 45 missing variables, and this may cause a problem in our dataset. A way to get through this is to probably use data imputation methods to account for missing data.

```
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
```

```
c <- ggplot(newData, aes(count))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The count variable is the average number of students per cohort. The minimum of 56 seems too small, but it could be that there is a lower graduation rate at those particular colleges. The 941.5 maximum is reasonable as it represents the very large colleges. There are no missing observations in this case.

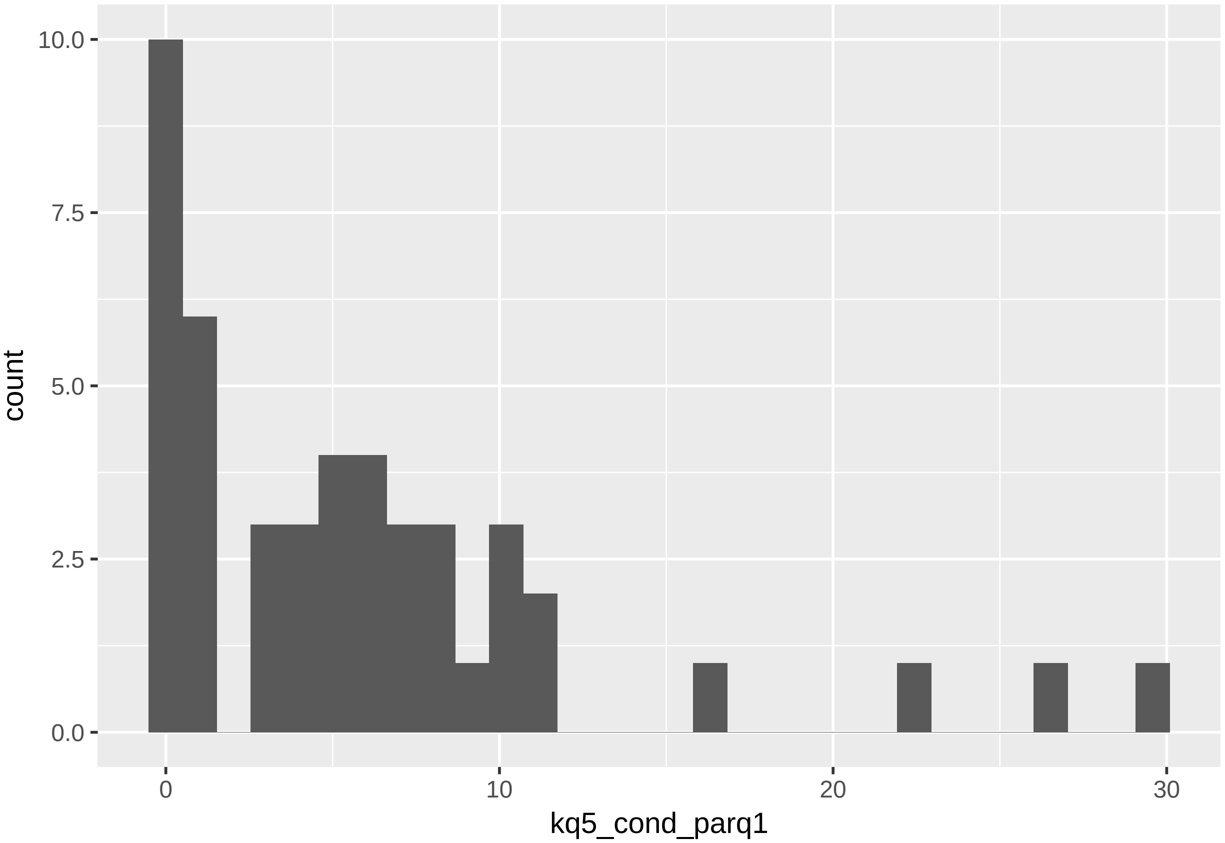
Y variable:

```
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
```

```
c <- ggplot(newData, aes(kq5_cond_parq1))
c+ geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This is our response variable. The kq5__cond__parq1 represents the percentage of children who reach the Top 20% of the income distribution among children with parents in the bottom 20% of the income distribution. The minimum value of 0% make sense, and the maximum value of 29.58% is also reasonable even though it is pretty low.

Most Pressing Data CLeaning

We are worried about the missing values in our variables especially in our graduation rate and endowments variables. Later on in our analysis, we are thinking about either using data imputation methods, or filtering out our missing variables using na.rm = TRUE , !is.na. Before filtering off the missing values, we will first check to see if any differences exist among them and the values we have and shall report this as part of our analysis.

Our next steps are to look at the relationships between our response variable, and the other explanatory variable. We did not include this feature in our data appendix because it is more of an analysis rather than a descriptive statistic.