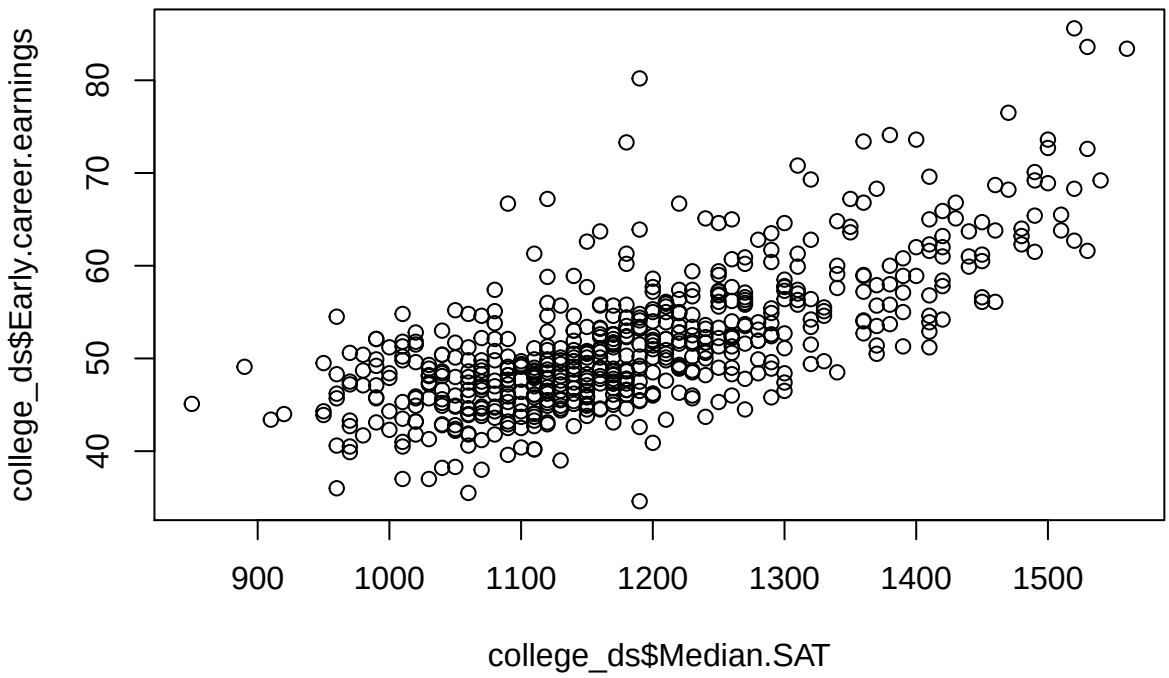
```r
# See the dataset of the best and worst 10 rankings of colleges
head(vital_college_ds, 10)
```

```
##    Median.SAT Median.ACT Early.career.earnings
## 1        1260         29                  57.7
## 2        1270         NA                  57.1
## 3        1500         33                  72.7
## 4        1340         32                  60.0
## 5        1250         29                  59.4
## 6        1470         34                  76.5
## 7        1530         34                  83.6
## 8        1420         32                  62.0
## 9        1310         30                  61.3
## 10       1410         31                  62.3
```

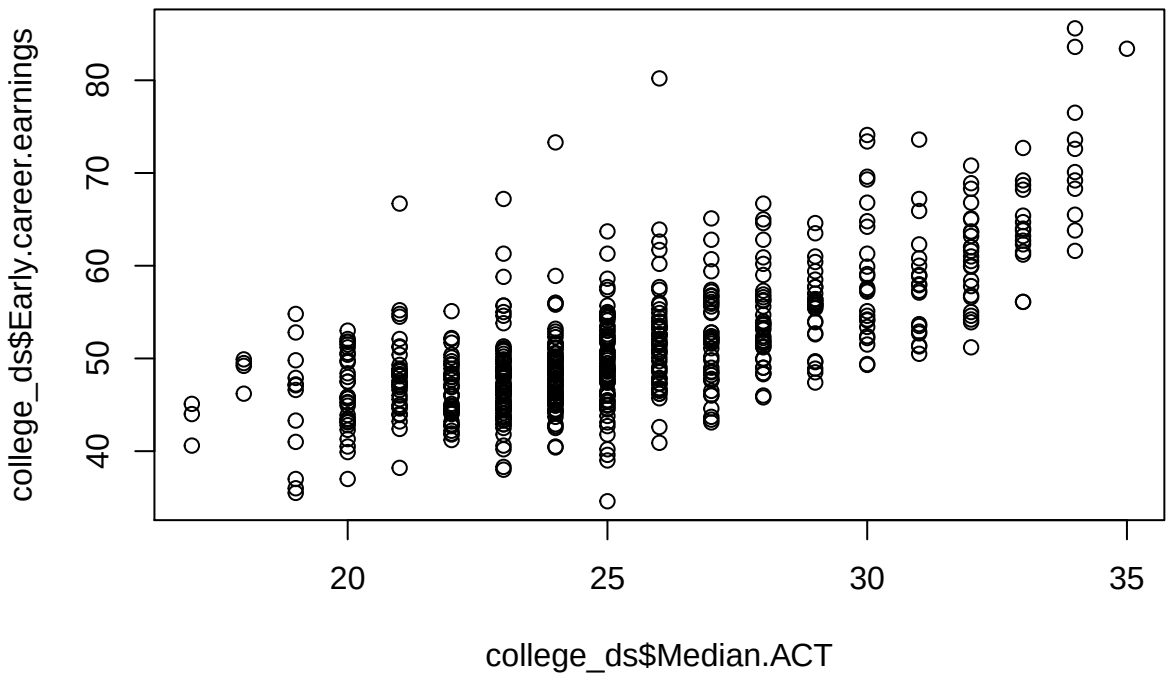
```r
tail(vital_college_ds, 10)
```

```
##     Median.SAT Median.ACT Early.career.earnings
## 735       1150         24                  47.1
## 736        990         20                  45.8
## 737         NA         NA                  43.9
## 738       1130         24                  45.5
## 739        850         17                  45.1
## 740         NA         NA                  38.2
## 741         NA         NA                  45.1
## 742       1000         20                  42.3
## 743       1060         23                  40.6
## 744         NA         NA                  46.2
```

```r
# Plot the dataset of median SAT and median ACT per college into the reported early career earnings in
# SAT score has a much higher range than ACT, so the graph looks more "continuous"
plot(college_ds$Median.SAT, college_ds$Early.career.earnings)
```

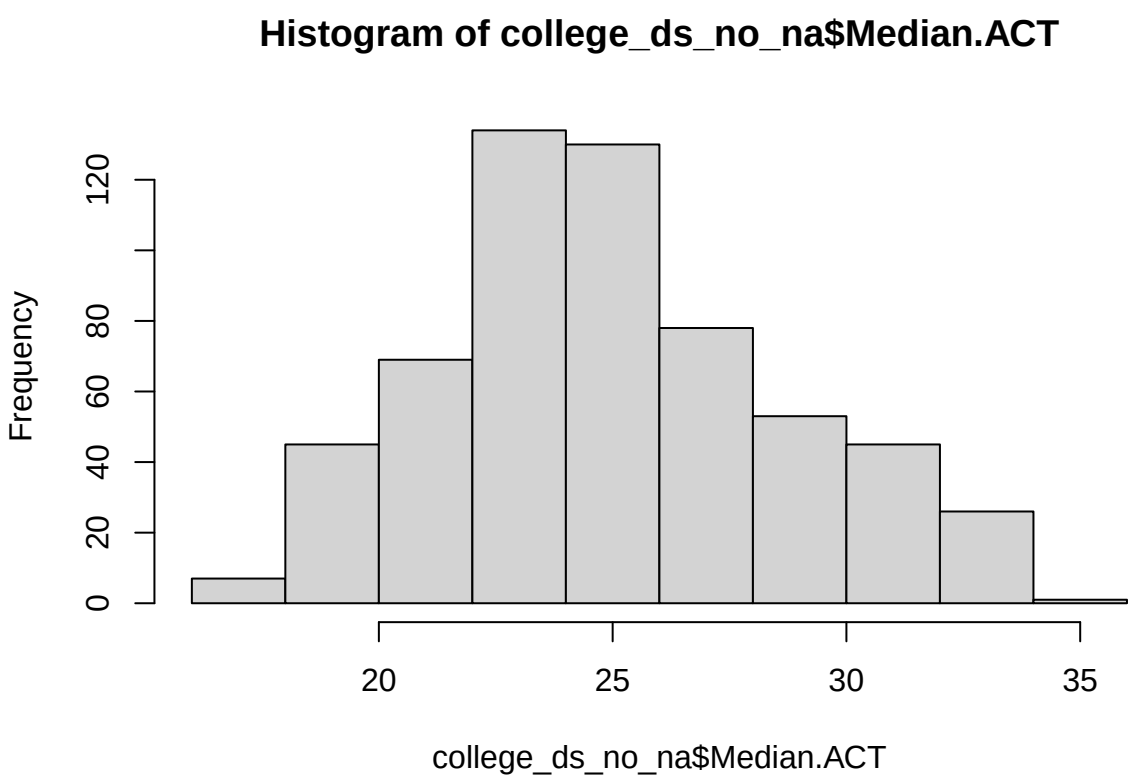



```
# ACT score has less range, so the graph looks "discrete"
plot(college_ds$Median.ACT, college_ds$Early.career.earnings)
```




```
# Contains absolutely no NA
college_ds_no_na = subset(college_ds, !is.na(Median.SAT) & !is.na(Median.ACT))
# Contains only one column of NA per row (so that we can predict missing Median ACT)
college_ds_predict <- subset(college_ds, !is.na(Median.SAT) | !is.na(Median.ACT))
```

```
# summary(college_ds_no_na)
hist(college_ds_no_na$Median.ACT)
```

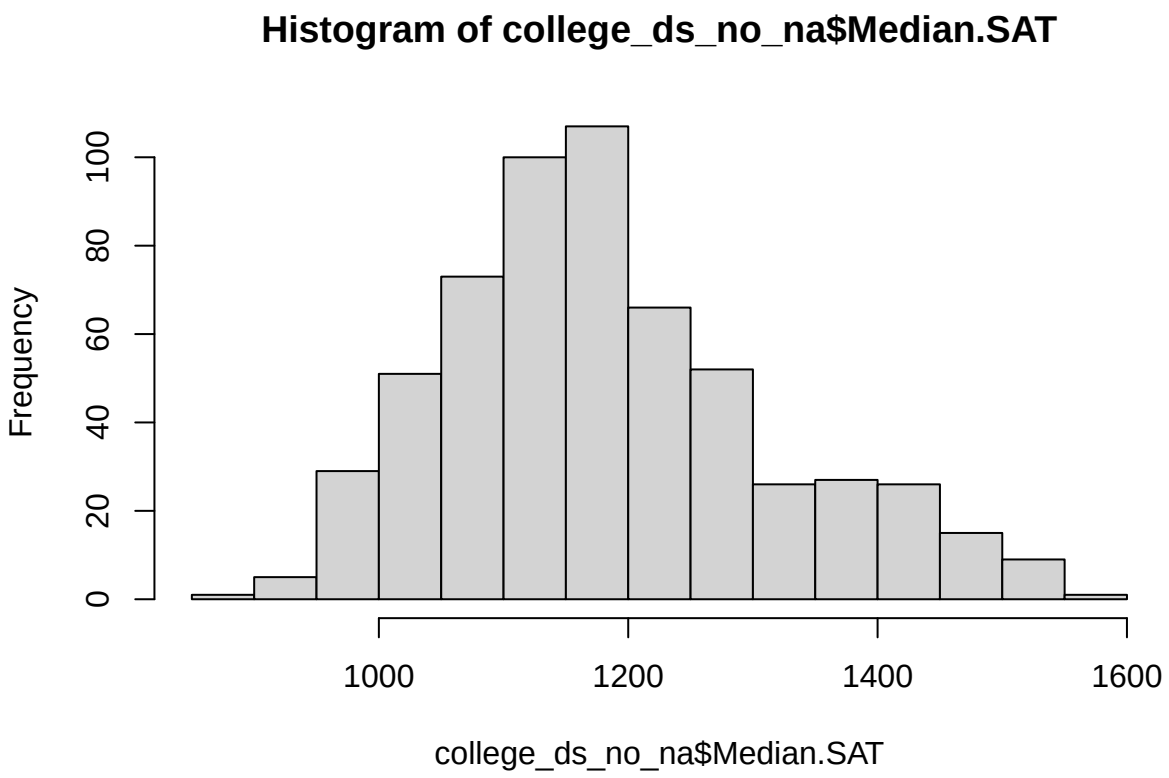



```
summary(college_ds_no_na$Median.ACT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   23.00   25.00   25.48   28.00   35.00
```

The median ACT score for the colleges in the database has the mean at around 25. The histogram of the frequency suggests that colleges tend to accept students with ACT scores slightly above the median (mean > median). The right skewness of the histogram suggests that colleges with median ACT score acceptance tend to be more specific on the requirements as the frequency tends to spread out as the score goes above the mean (25.5).

```
hist(college_ds_no_na$Median.SAT)
```

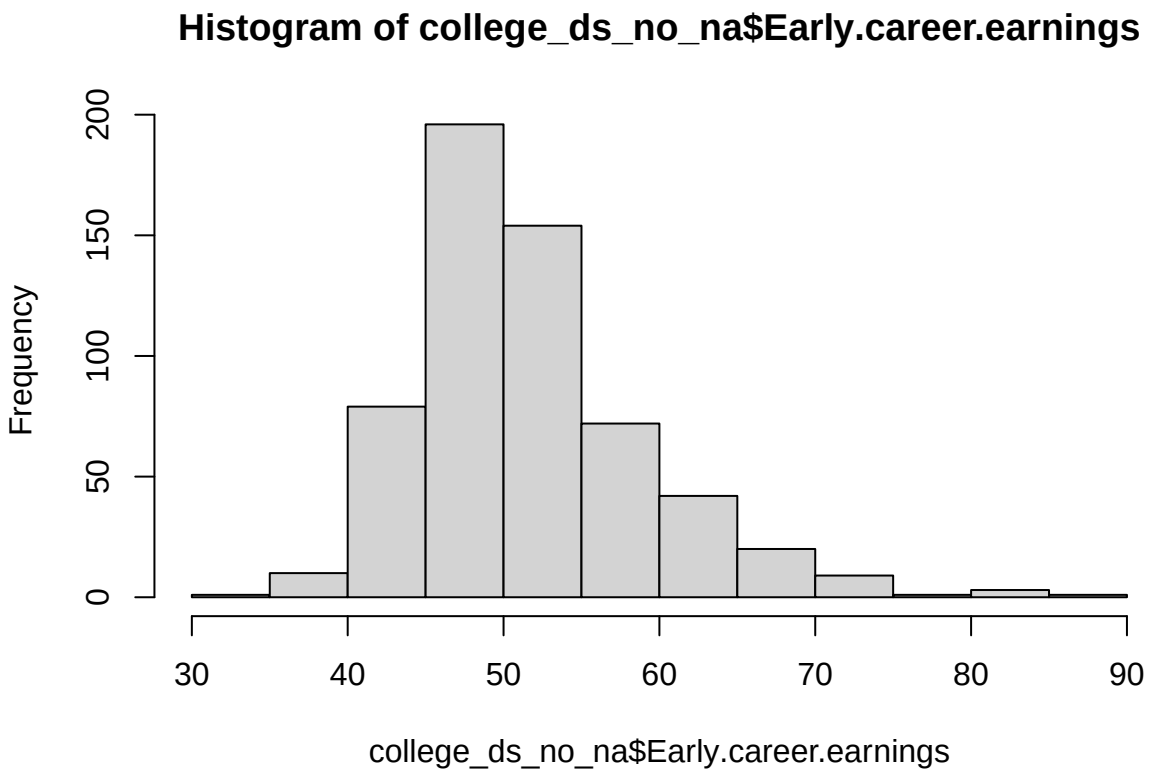



```
summary(college_ds_no_na$Median.SAT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     850    1098    1170    1187    1260    1560
```

The shape of the histogram in median SAT resembles that of the median ACT. The mean (1187) also surpasses the median (1170), and the graph is also right-skewed.

```
hist(college_ds_no_na$Early.career.earnings)
```




```
summary(college_ds_no_na$Early.career.earnings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.60   46.58   50.20   51.67   55.10   85.60
```

The shape of the histogram of early career earnings also resembles the two above histograms. The mean earning ($51,670) also surpasses the median ($50,200). Overall, this histogram spreads out a lot less than the two above histograms, but it still remains right-skewed. This suggests that most people graduating from college will tend to have a specific amount of earnings in their early careers. However, there are cases where people may earn just less than twice the common amount of earning.

In our project, we will investigate whether SAT/ACT contributes to the early career earning. We have plotted out 2 plots of median SAT vs early career earning and median ACT vs early career earning. We found out that there is a relationship between median ACT/SAT vs early career earning since both plots are increasing. We will get into the detail about how the 3 variables interact with each other through linear regression model and may predict missing ACT scores based on the given SAT score if we have sufficient time.