# Project Sandbox

Hung Tran, Linh Ta

6/17/2020

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

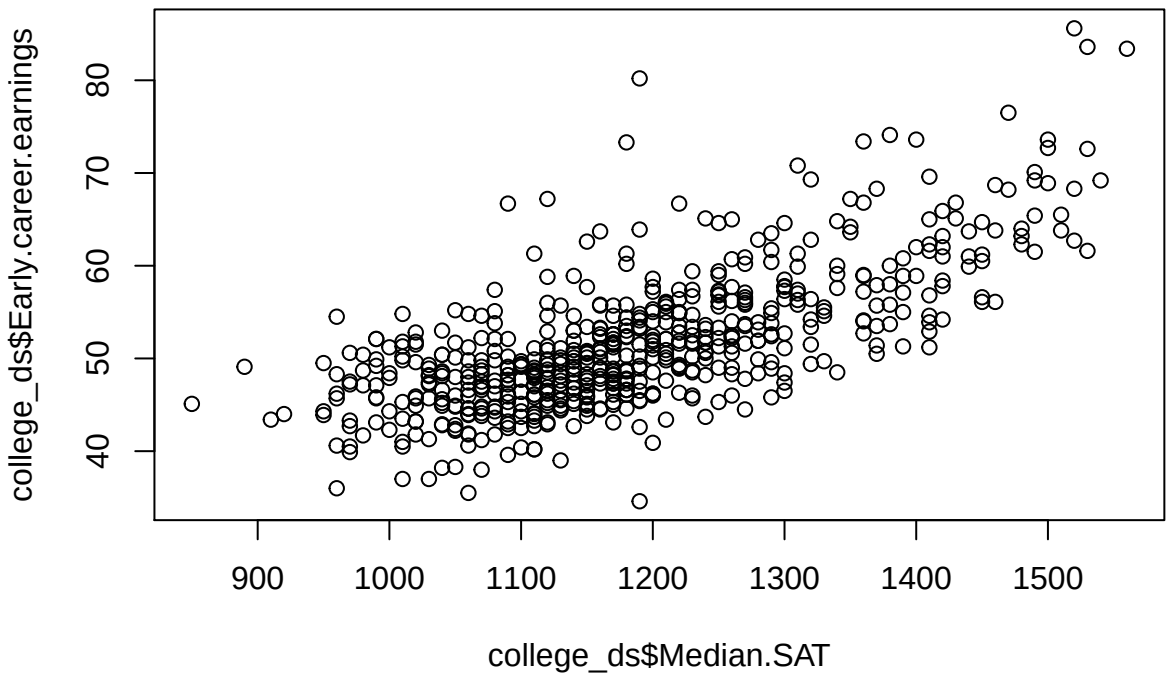
```
##    Rank                          College.Name            Location Median.SAT
## 1     1       University of California-Irvine          Irvine. CA       1260
## 2     2        CUNY Bernard M. Baruch College        New York. NY       1270
## 3     3                  Princeton University       Princeton. NJ       1500
## 4     4  University of California-Los Angeles     Los Angeles. CA       1340
## 5     5        University of California-Davis           Davis. CA       1250
## 6     6                   Stanford University        Stanford. CA       1470
## 7     7 Massachusetts Institute of Technology       Cambridge. MA       1530
## 8     8      University of Michigan-Ann Arbor       Ann Arbor. MI       1420
## 9     9    University of California-San Diego        La Jolla. CA       1310
## 10   10                University of Virginia Charlottesville. VA       1410
##    Median.ACT Est..price.2019.20.without.aid Est..price.2019.20.with.avg..grant
## 1          29                           35.4                                14.9
## 2          NA                           33.2                                 4.9
## 3          33                           69.1                                17.4
## 4          32                           35.4                                15.8
## 5          29                           37.3                                17.6
## 6          34                           72.2                                17.7
## 7          34                           70.4                                23.7
## 8          32                           30.8                                17.5
## 9          30                           33.6                                15.6
## 10         31                           33.6                                17.7
##    X..of.students.who.get.any.grants Average.student.debt Early.career.earnings
## 1                               0.66                16.50                   57.7
## 2                               0.55                10.72                   57.1
## 3                               0.58                 7.50                   72.7
## 4                               0.59                15.00                   60.0
## 5                               0.67                14.00                   59.4
## 6                               0.60                11.45                   76.5
## 7                               0.68                17.13                   83.6
## 8                               0.52                19.15                   62.0
## 9                               0.58                17.50                   61.3
## 10                              0.42                19.00                   62.3
```

```
tail(college_ds, 10)
```

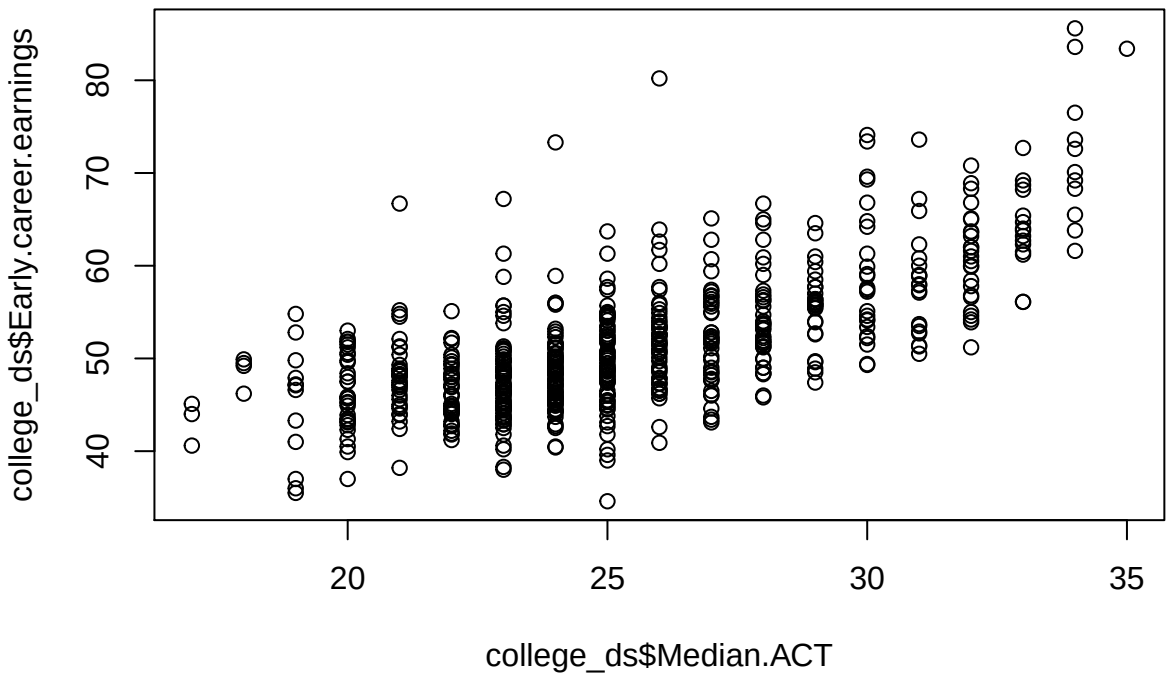
```
##     Rank                        College.Name           Location Median.SAT
## 735  735              High Point University     High Point. NC       1150
## 736  736           Clark Atlanta University        Atlanta. GA        990
## 737  737                New England College       Henniker. NH         NA
```

```
## 738  738 Savannah College of Art and Design       Savannah. GA       1130
## 739  739         Johnson C Smith University      Charlotte. NC        850
## 740  740            Glenville State College      Glenville. WV         NA
## 741  741               Davenport University   Grand Rapids. MI         NA
## 742  742                 Oakwood University     Huntsville. AL       1000
## 743  743                  Regent University Virginia Beach. VA       1060
## 744  744         The University of the Arts   Philadelphia. PA         NA
##     Median.ACT Est..price.2019.20.without.aid
## 735         24                           53.4
## 736         20                           39.0
## 737         NA                           56.2
## 738         24                           58.9
## 739         17                           33.6
## 740         NA                           24.2
## 741         NA                           32.0
## 742         20                           37.3
## 743         23                           30.5
## 744         NA                           68.2
##     Est..price.2019.20.with.avg..grant X..of.students.who.get.any.grants
## 735                               41.7                              0.74
## 736                               28.6                              0.83
## 737                               29.3                              0.79
## 738                               44.6                              0.91
## 739                               19.3                              0.94
## 740                               13.4                              0.80
## 741                               17.9                              0.79
## 742                               27.3                              0.90
## 743                               19.0                              0.87
## 744                               42.0                              0.97
##     Average.student.debt Early.career.earnings
## 735                25.00                  47.1
## 736                31.00                  45.8
## 737                26.00                  43.9
## 738                26.37                  45.5
## 739                32.50                  45.1
## 740                23.91                  38.2
## 741                26.91                  45.1
## 742                31.00                  42.3
## 743                25.00                  40.6
## 744                27.00                  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)
```

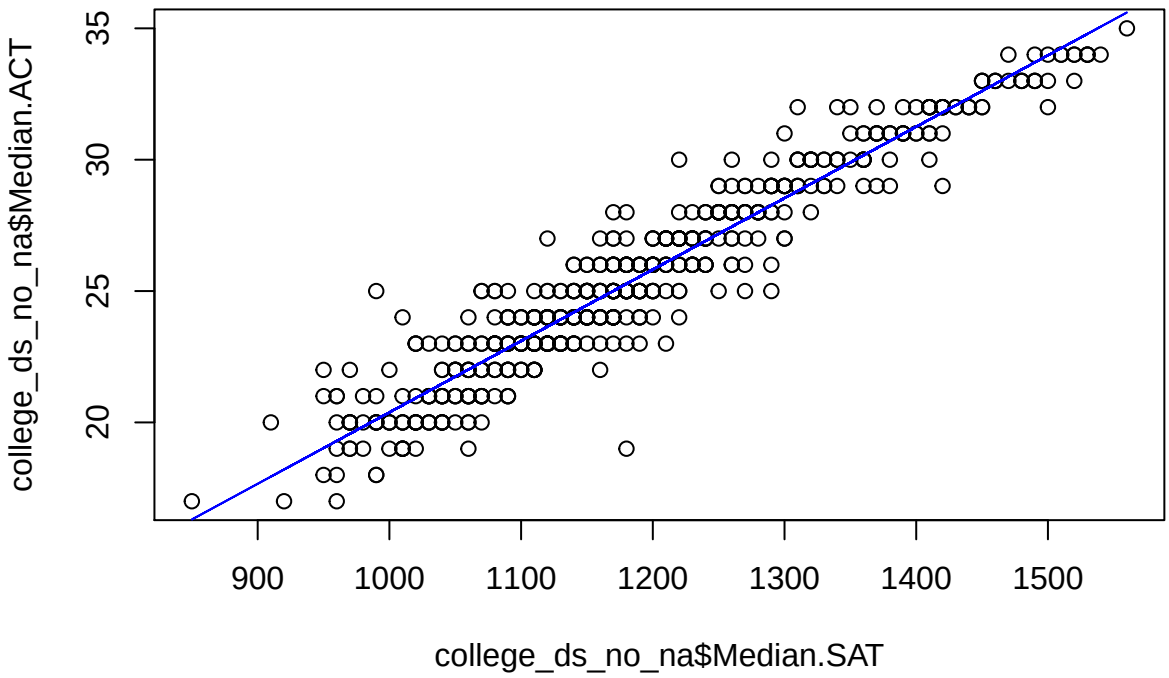



```
# ACT score has less range, so the graph looks "discrete"
plot(college_ds$Median.ACT, college_ds$Early.career.earnings)
```



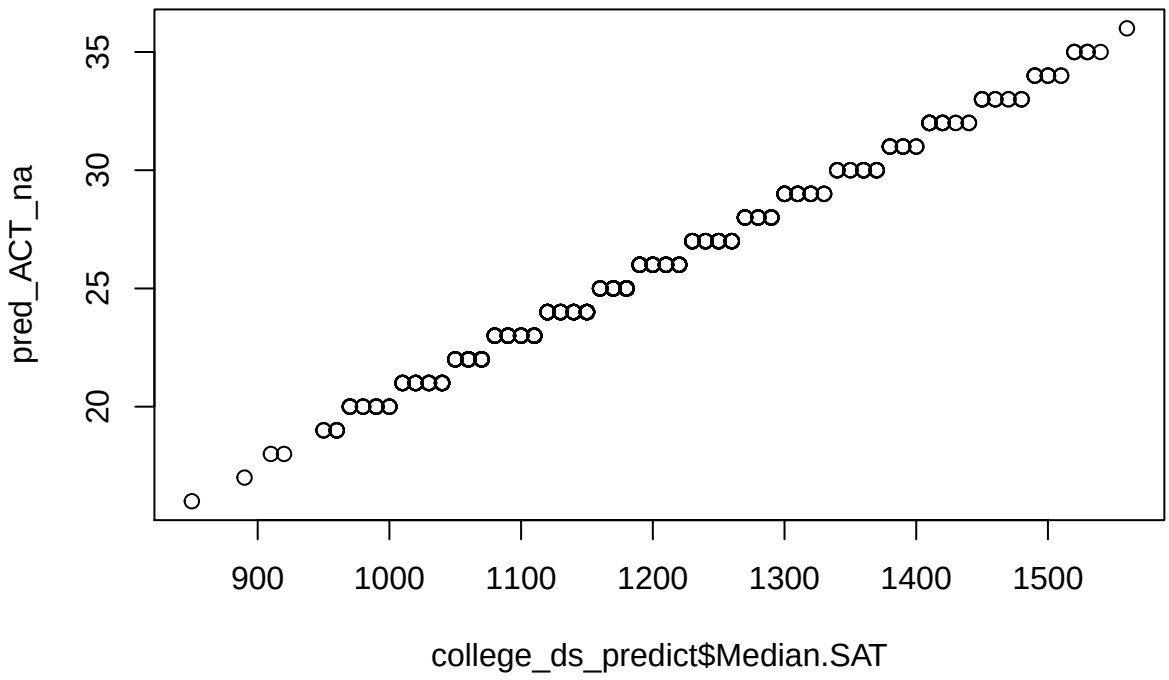

```r
# Get rows that have no NA, or only one of SAT, ACT is NA. Omit if both is NA.
college_ds_predict = subset(college_ds, !is.na(Median.SAT) | !is.na(Median.ACT))
# college_ds_no_na = subset(college_ds, !(is.na(Median.SAT) & is.na(Median.ACT)))
# Omit the row where ANY column of (SAT, ACT) is NA.
college_ds_no_na = subset(college_ds, !is.na(Median.SAT) & !is.na(Median.ACT))
corr_c_ds_act = cor(college_ds_no_na$Median.ACT, college_ds_no_na$Early.career.earnings)
corr_c_ds_sat = cor(college_ds_no_na$Median.SAT, college_ds_no_na$Early.career.earnings)
```

```r
plot(college_ds_no_na$Median.SAT, college_ds_no_na$Median.ACT)
ACT_bounds = c(1, 36)
clamp <- function(x, bounds) {
  pmax(pmin(x, bounds[2]), bounds[1])
}
predict_ACT_lm = lm(Median.ACT ~ Median.SAT, data = college_ds_no_na)

# <not necessary>
act_pred = clamp(predict(predict_ACT_lm),ACT_bounds)
par(new = TRUE)
# Plot prediction on the trained dataset
points(college_ds_no_na$Median.SAT, act_pred, type = "l", col = "blue")
```




```
# </not necessary>
```

```
# Plot ACT prediction from SAT from dataset with NAs
pred_ACT_na = clamp(round(predict(predict_ACT_lm, newdata = college_ds_predict)), ACT_bounds)
plot(college_ds_predict$Median.SAT, pred_ACT_na)
```




```
names(pred_ACT_na) <- NULL

# Intersection, but only take values that are NA in college_ds_predict
# How many NA?
sum(is.na(college_ds_predict$Median.ACT))
```

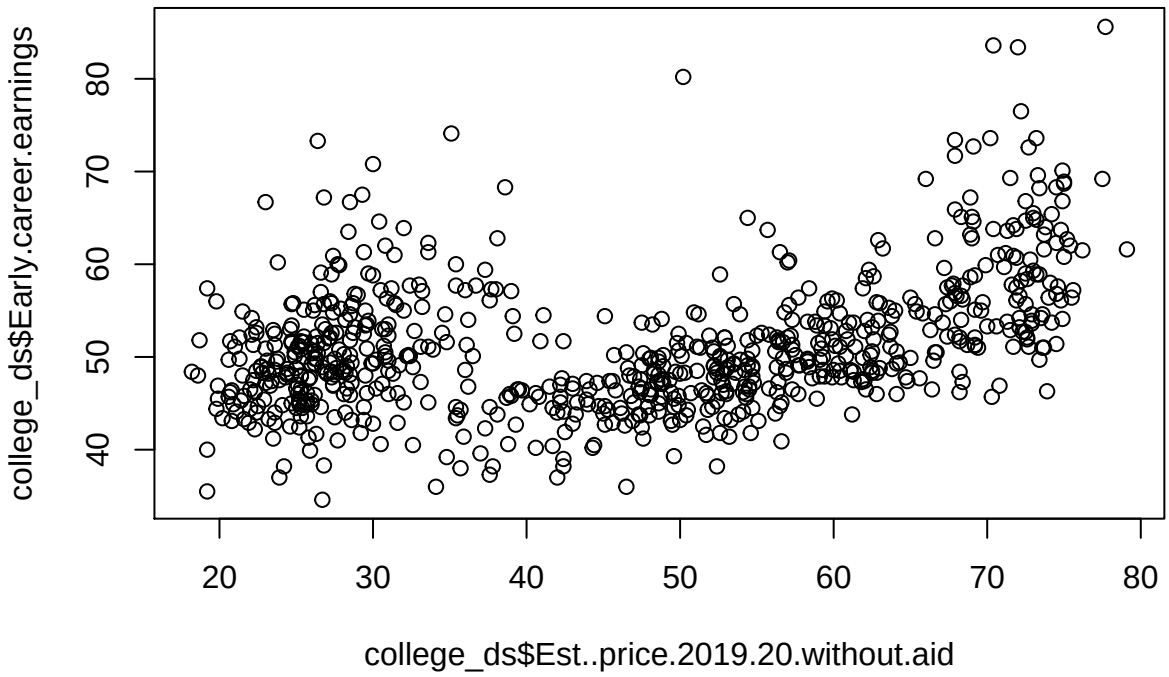
```
## [1] 14
```

```
college_ds_predict$MedACT_pred = pred_ACT_na
college_ds_predict<- within(college_ds_predict, Median.ACT[is.na(Median.ACT)] <- MedACT_pred[is.na(Media
college_ds_predict$MedACT_pred <- NULL
# How many NA?
sum(is.na(college_ds_predict$Median.ACT))
```

```
## [1] 0
```

```
# Predict SAT based on ACT
predict_SAT_lm = lm(Median.SAT ~ Median.ACT, data = college_ds_no_na)
# Use the trained model to predict the datsset with NA
SAT_bounds = c(0,1600)
# predict, round to the nearest tenth, then clamp to the possible score
pred_SAT_na = clamp(round(predict(predict_SAT_lm, newdata= college_ds_predict)/10)*10, SAT_bounds)
college_ds_predict$MedSAT_pred = pred_SAT_na
# Intersection
college_ds_predict<- within(college_ds_predict, Median.SAT[is.na(Median.SAT)] <- MedSAT_pred[is.na(Media
college_ds_predict$MedSAT_pred <- NULL
```

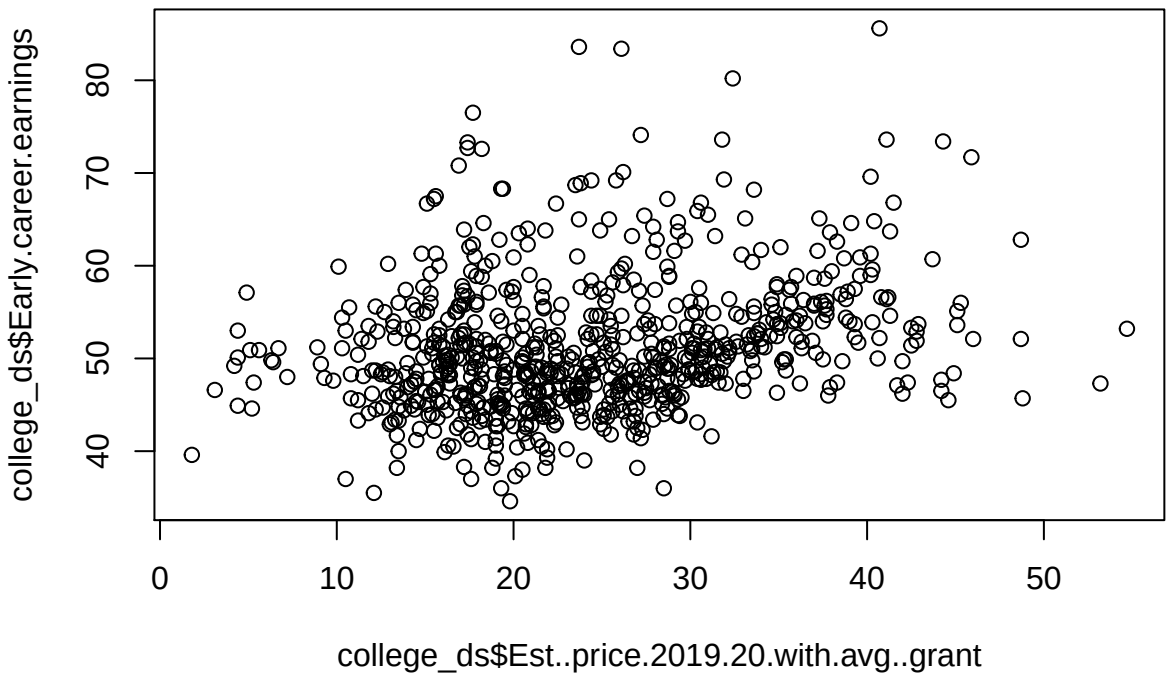
```r
plot(college_ds$Est..price.2019.20.without.aid, college_ds$Early.career.earnings)
```




```r
cor(college_ds$Est..price.2019.20.without.aid, college_ds$Early.career.earnings)
```

```
## [1] 0.3797211
```

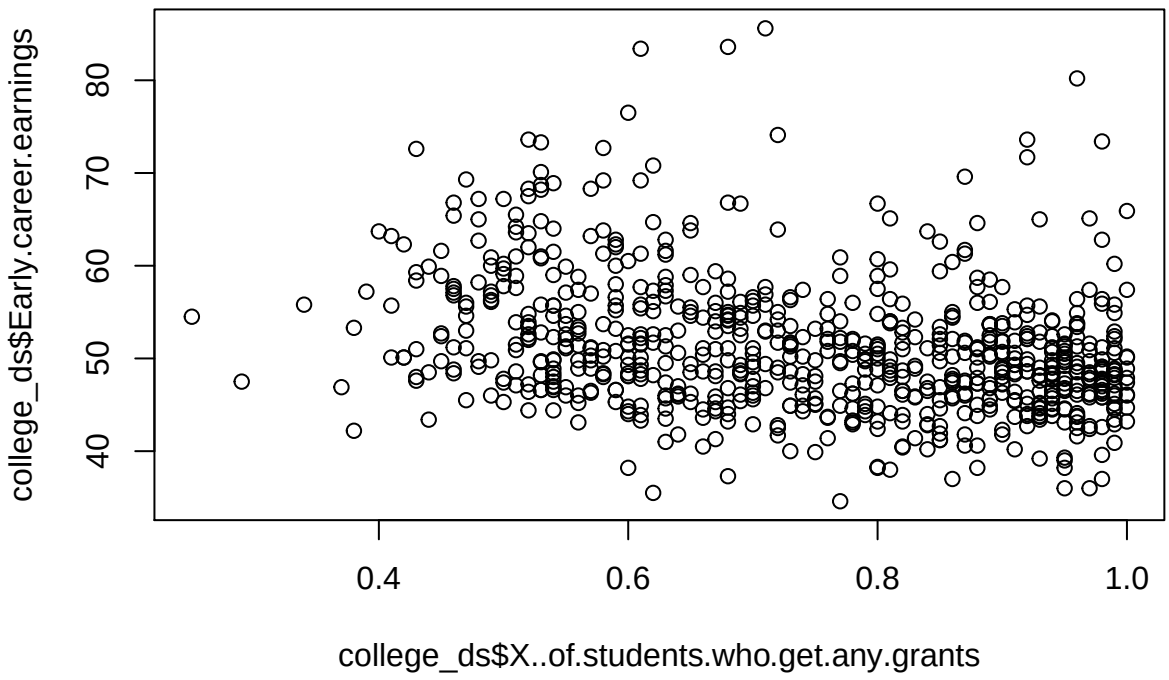
```r
plot(college_ds$Est..price.2019.20.with.avg..grant, college_ds$Early.career.earnings)
```




```
cor(college_ds$Est..price.2019.20.with.avg..grant, college_ds$Early.career.earnings)
```

```
## [1] 0.2639509
```

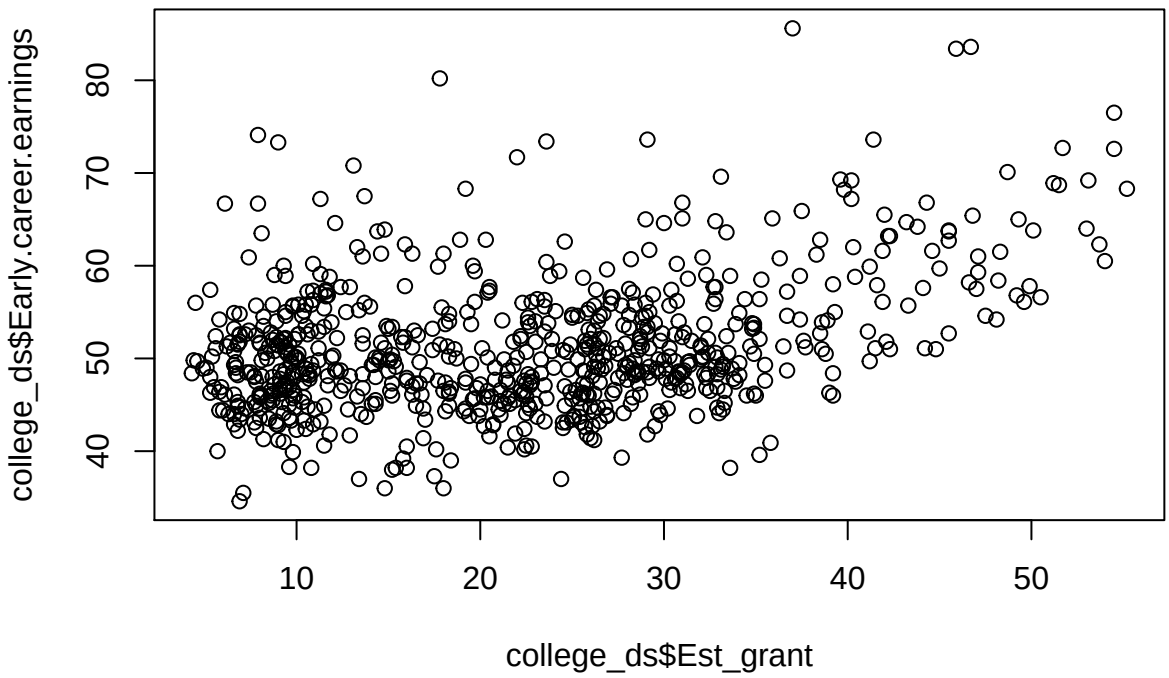
```
plot(college_ds$X..of.students.who.get.any.grants, college_ds$Early.career.earnings)
```




```
cor(college_ds$X..of.students.who.get.any.grants, college_ds$Early.career.earnings)
```

```
## [1] -0.3147469
```

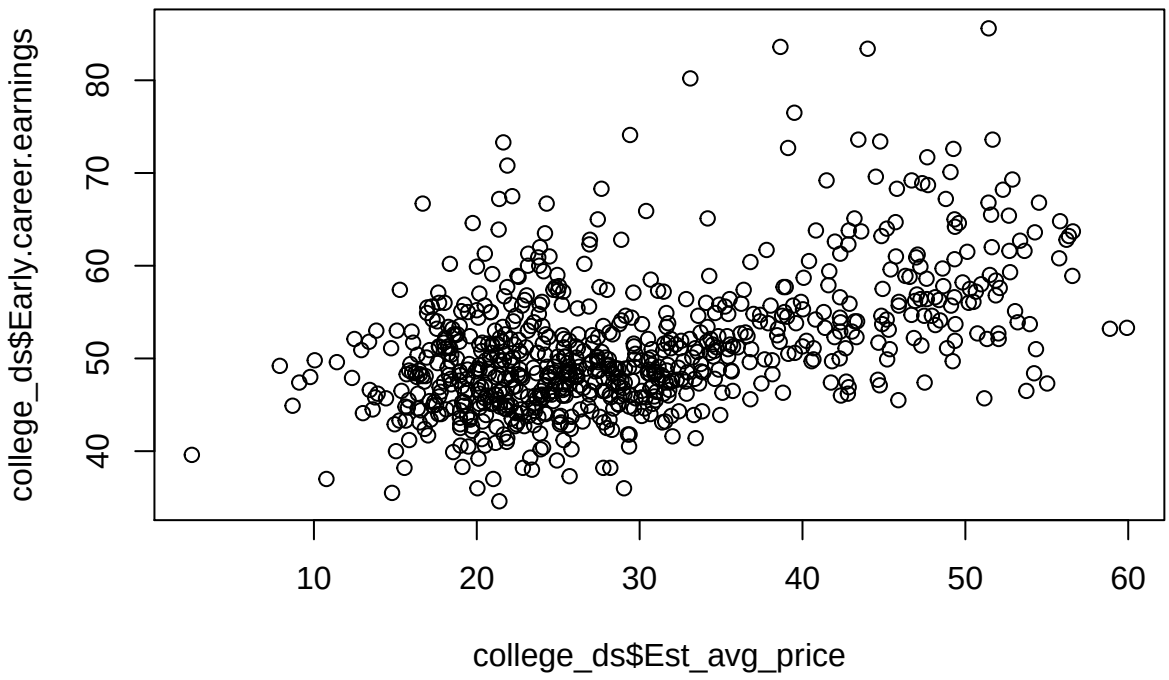
```
# Assuming price w/ avg grant is price given that the student has any grant minus mean grant in that co
college_ds$Est_grant = college_ds$Est..price.2019.20.without.aid - college_ds$Est..price.2019.20.with.a

# Est_avg_price = Est_avg_price_no_aid * (1-prob_any_aid) + Est_avg_price_aid * (prob_any_aid)
college_ds$Est_avg_price = college_ds$Est..price.2019.20.with.avg..grant * college_ds$X..of.students.wh
  college_ds$Est..price.2019.20.without.aid * (1-college_ds$X..of.students.who.get.any.grants)
plot(college_ds$Est_grant, college_ds$Early.career.earnings)
```




```
cor(college_ds$Est_grant, college_ds$Early.career.earnings)
```

```
## [1] 0.3689321
```

```
plot(college_ds$Est_avg_price, college_ds$Early.career.earnings)
```




```
cor(college_ds$Est_avg_price, college_ds$Early.career.earnings)
```

```
## [1] 0.4871582
```