US College Data Analysis

2023-08-18

The US College data set is available in ISLR2 package in R . This is an old dataset but is great to practice some data analysis .

In this analysis , we want to find out

1)If colleges with larger full time enrollments have lower grad rates ?

2)Is it diff for public/private institutions?

```
library(tidyverse)
library(ggplot2)
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.2.3
```

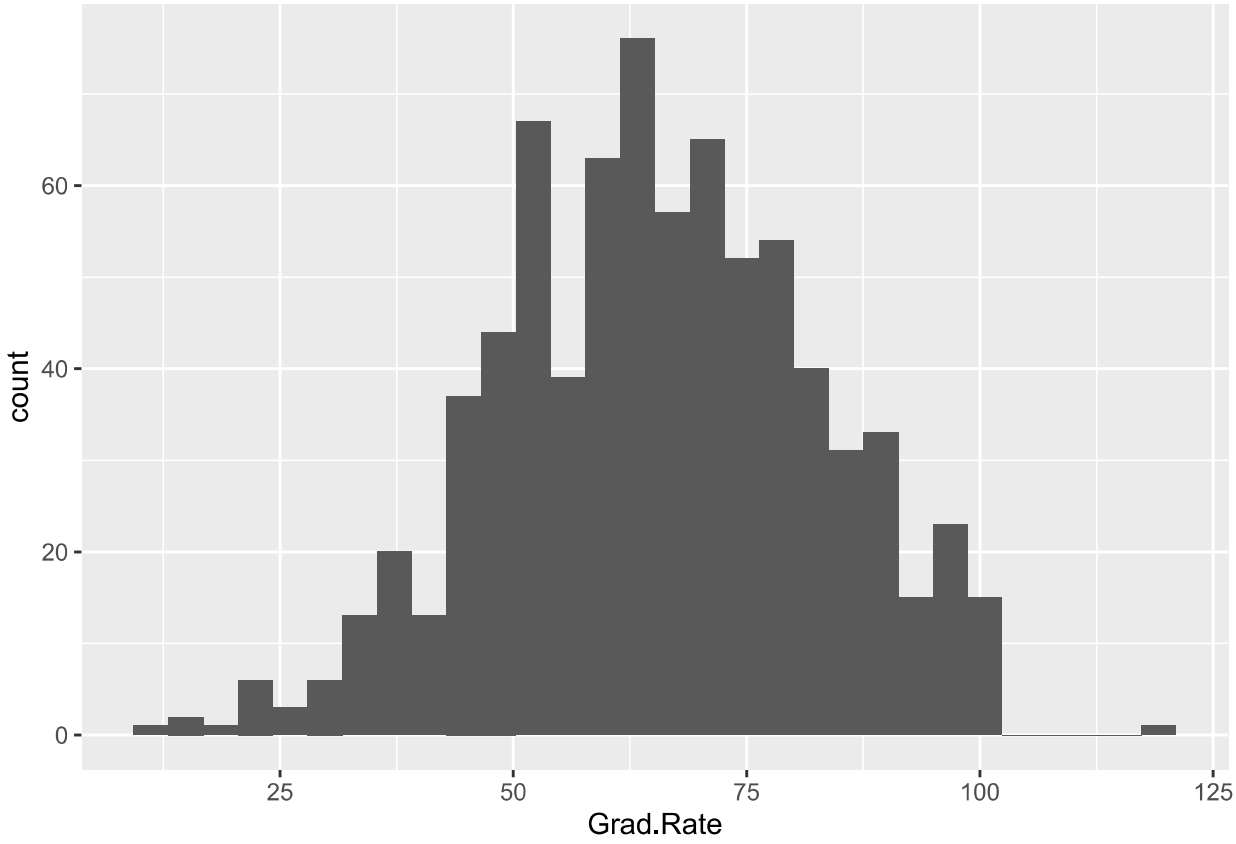
```
glimpse(College)
```

```
## Rows: 777
## Columns: 18
## $ Private     <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes~
## $ Apps        <dbl> 1660, 2186, 1428, 417, 193, 587, 353, 1899, 1038, 582, 173~
## $ Accept      <dbl> 1232, 1924, 1097, 349, 146, 479, 340, 1720, 839, 498, 1425~
## $ Enroll      <dbl> 721, 512, 336, 137, 55, 158, 103, 489, 227, 172, 472, 484,~
## $ Top10perc   <dbl> 23, 16, 22, 60, 16, 38, 17, 37, 30, 21, 37, 44, 38, 44, 23~
## $ Top25perc   <dbl> 52, 29, 50, 89, 44, 62, 45, 68, 63, 44, 75, 77, 64, 73, 46~
## $ F.Undergrad <dbl> 2885, 2683, 1036, 510, 249, 678, 416, 1594, 973, 799, 1830~
## $ P.Undergrad <dbl> 537, 1227, 99, 63, 869, 41, 230, 32, 306, 78, 110, 44, 638~
## $ Outstate    <dbl> 7440, 12280, 11250, 12960, 7560, 13500, 13290, 13868, 1559~
## $ Room.Board  <dbl> 3300, 6450, 3750, 5450, 4120, 3335, 5720, 4826, 4400, 3380~
## $ Books       <dbl> 450, 750, 400, 450, 800, 500, 500, 450, 300, 660, 500, 400~
## $ Personal    <dbl> 2200, 1500, 1165, 875, 1500, 675, 1500, 850, 500, 1800, 60~
## $ PhD         <dbl> 70, 29, 53, 92, 76, 67, 90, 89, 79, 40, 82, 73, 60, 79, 36~
## $ Terminal    <dbl> 78, 30, 66, 97, 72, 73, 93, 100, 84, 41, 88, 91, 84, 87, 6~
## $ S.F.Ratio   <dbl> 18.1, 12.2, 12.9, 7.7, 11.9, 9.4, 11.5, 13.7, 11.3, 11.5, ~
## $ perc.alumni <dbl> 12, 16, 30, 37, 2, 11, 26, 37, 23, 15, 31, 41, 21, 32, 26,~
## $ Expend      <dbl> 7041, 10527, 8735, 19016, 10922, 9727, 8861, 11487, 11644,~
## $ Grad.Rate   <dbl> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74, 68, 55~
```

Exploratory viz

```
ggplot(College, aes(x=Grad.Rate)) + geom_histogram()
```

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



We can see some outlier in the Grad.Rate as the graduation rate ideally should not be exceeding 100%. We will see what the outlier is .

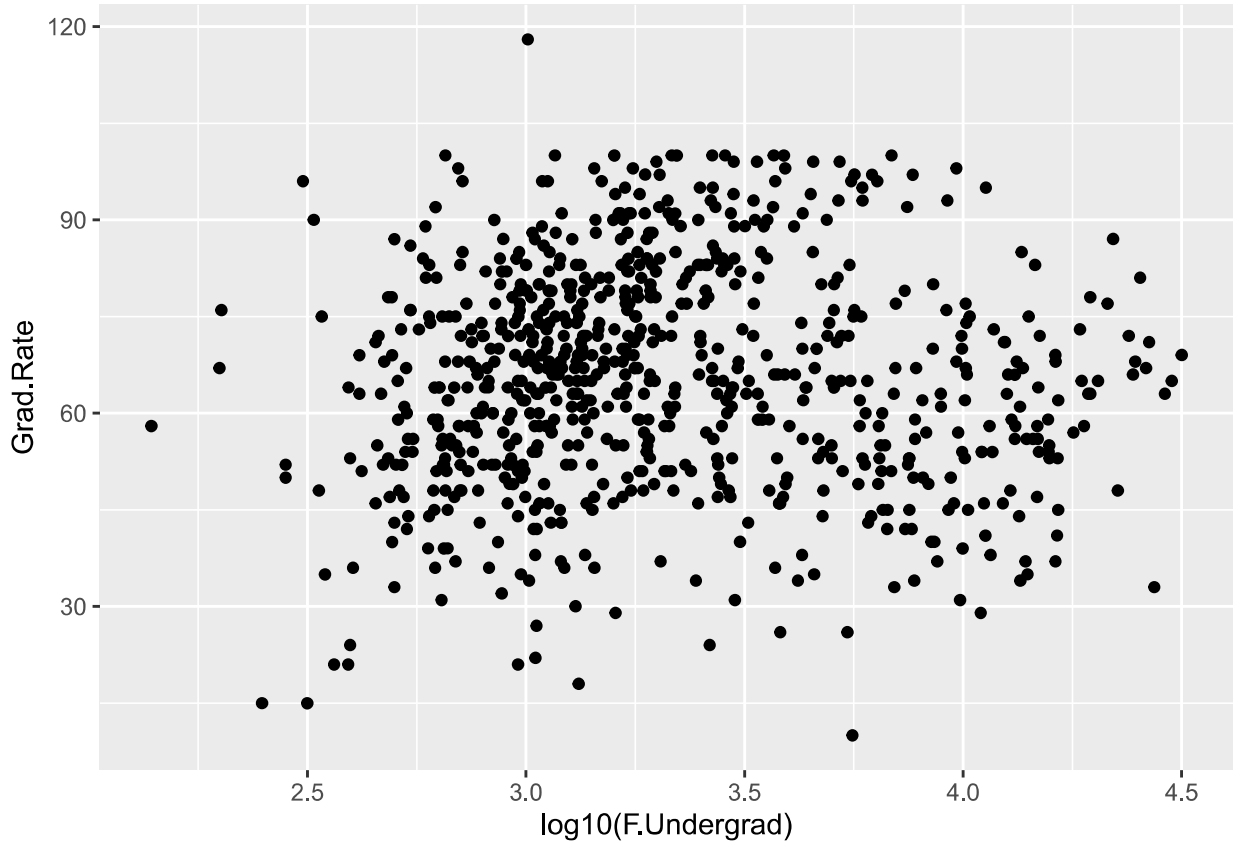
```
outlier <- College |>
  filter(Grad.Rate >=100)

view(outlier)
```

We can see that the Cazenovia College has 118% Graduate Rate . This could be a potential Outlier

Lets plot a scattered graph between full undergrad and graduation rate

```
ggplot(College, aes(x=log10(F.Undergrad),
                    y=Grad.Rate)) +
  geom_point()
```




```
college_final <- College |>
  mutate(log_fulltime = log10(F.Undergrad)) |>
  select(Private,
    Grad.Rate,
    log_fulltime)

glimpse(college_final)
```

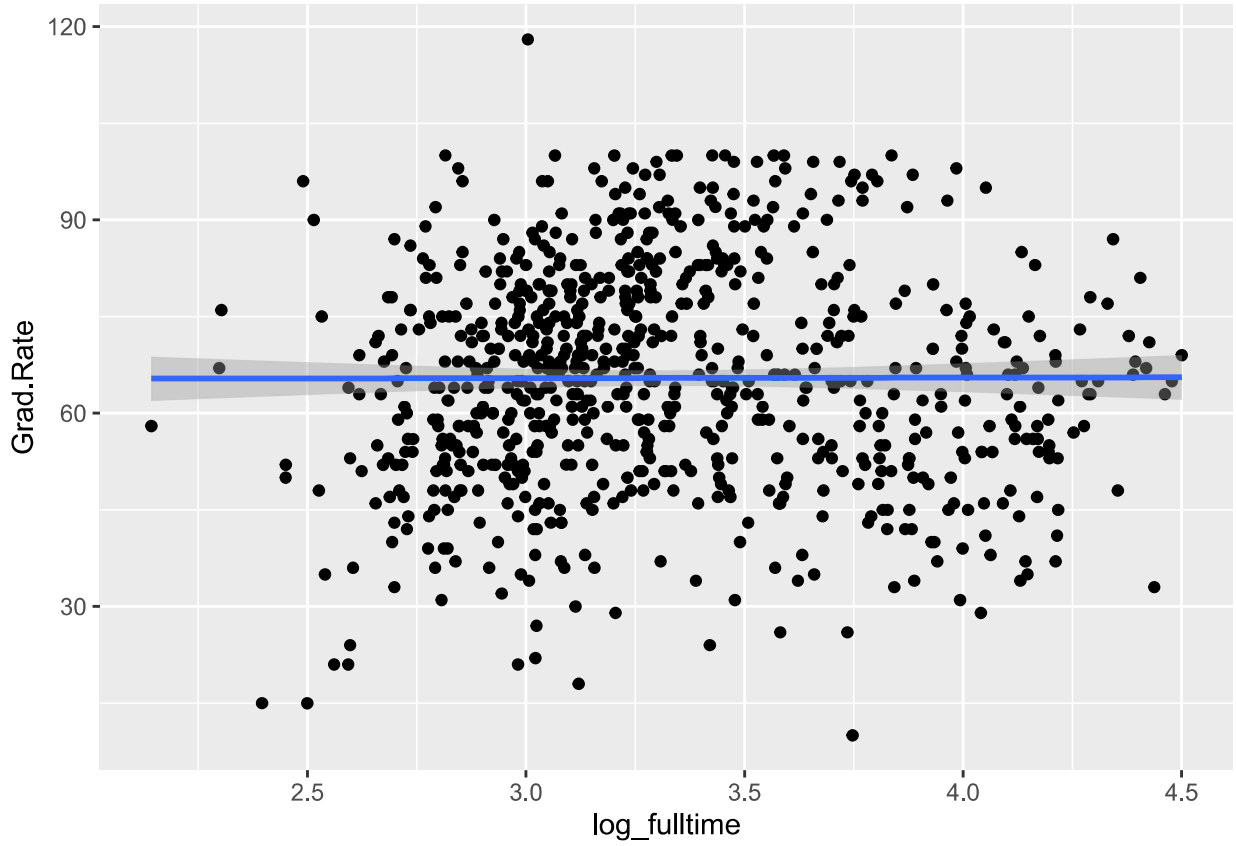
```
## Rows: 777
## Columns: 3
## $ Private      <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Ye~
## $ Grad.Rate    <dbl> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74, 68, 5~
## $ log_fulltime <dbl> 3.460146, 3.428621, 3.015360, 2.707570, 2.396199, 2.83123~
```

Modelling

Model1

```
ggplot(college_final, aes(x = log_fulltime , y = Grad.Rate)) +
  geom_point()+
  geom_smooth(method ='lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
model_undergrad <- lm(Grad.Rate ~ log_fulltime,
                      data = college_final)

summary(model_undergrad)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ log_fulltime, data = college_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.501 -12.433  -0.434  12.539  52.564
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.17524    4.62722  14.085   <2e-16 ***
## log_fulltime  0.08688    1.38301   0.063     0.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.19 on 775 degrees of freedom
## Multiple R-squared:  5.092e-06,  Adjusted R-squared:  -0.001285
## F-statistic: 0.003946 on 1 and 775 DF,  p-value: 0.9499
```

The p-value for the coefficient of log_fulltime is 0.95, which is greater than 0.05. This suggests that the

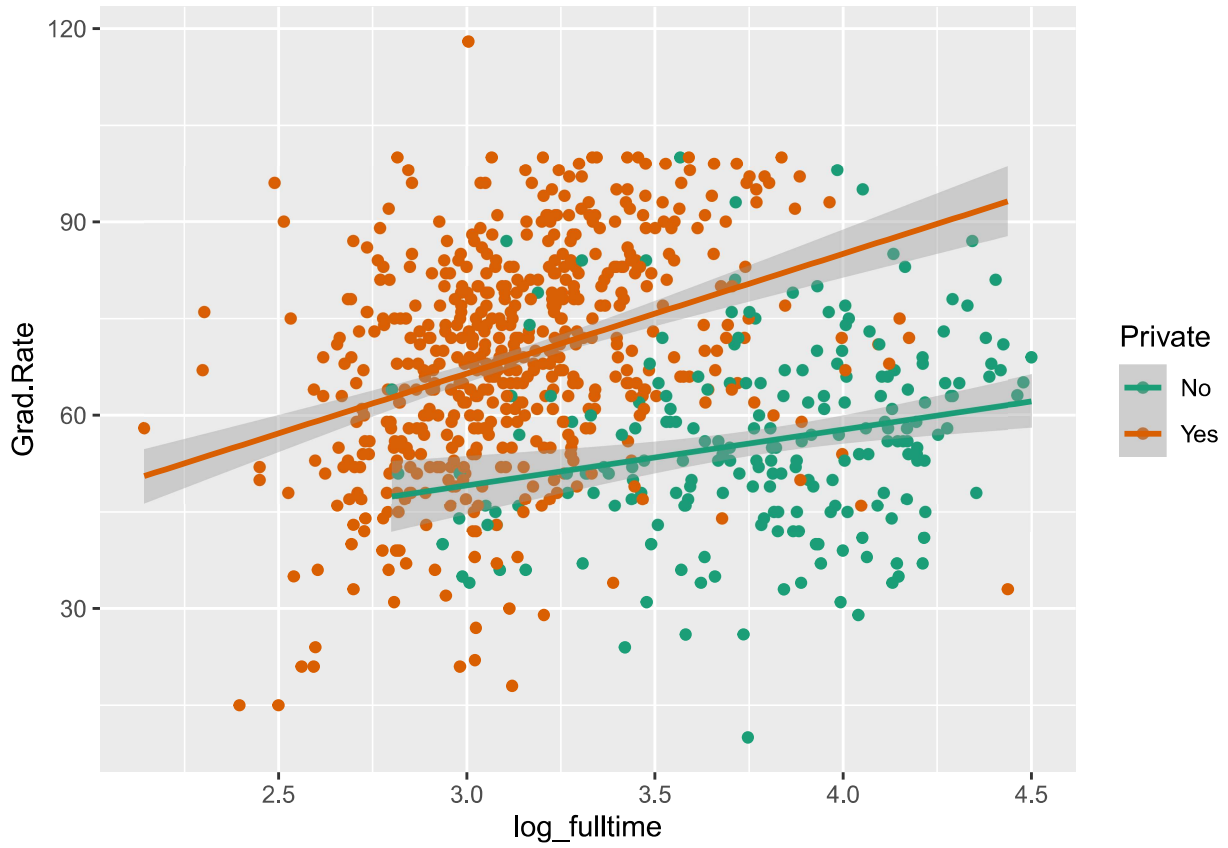
log_fulltime variable is not statistically significant in predicting the Graduation Rate, as its coefficient is not significantly different from zero.

Model 2

Lets plot a graph between fulltime Undergard and Grad Rate with the introduction of another variable that says if a college is Private or not.

```
ggplot(college_final, aes(x=log_fulltime, y = Grad.Rate, color =Private)) +
  geom_point() + geom_smooth(method = 'lm')+
  scale_color_brewer(palette = "Dark2")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
model_private <- lm(Grad.Rate ~ Private + log_fulltime,
                    data = college_final)
summary(model_private)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + log_fulltime, data = college_final)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -55.826  -9.581  -0.128  10.752  51.007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.784      6.330  -0.282    0.778
## PrivateYes     23.007      1.646  13.978   <2e-16 ***
## log_fulltime   15.235      1.644   9.266   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.37 on 774 degrees of freedom
## Multiple R-squared:  0.2016, Adjusted R-squared:  0.1995
## F-statistic:  97.7 on 2 and 774 DF,  p-value: < 2.2e-16
```

Both PrivateYes and log_fulltime have extremely low p-values (p < 2.2e-16), indicated by the '***' next to their estimates. This suggests that both variables are highly statistically significant in predicting the Graduation Rate.

```
model_private_interactive <- lm(Grad.Rate ~
                                  Private * log_fulltime ,
                                data = college_final)
summary(model_private_interactive)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private * log_fulltime, data = college_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.179  -9.488  -0.285  10.789  51.447
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                23.204     10.755   2.158  0.03127 *
## PrivateYes                -12.468     12.482  -0.999  0.31815
## log_fulltime                8.652      2.820   3.068  0.00223 **
## PrivateYes:log_fulltime     9.928      3.463   2.867  0.00426 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 773 degrees of freedom
## Multiple R-squared:   0.21,  Adjusted R-squared:  0.2069
## F-statistic: 68.48 on 3 and 773 DF,  p-value: < 2.2e-16
```

PrivateYes:log_fulltime: This is the combined effect of being a private college and the log_fulltime variable.s

In summary, the model suggests that the interaction between being a private college and the logarithm of the number of full-time undergraduates has a statistically significant effect on predicting the graduation rate. while the effect of PrivateYes on its own does not appear to be statistically significant.

so 1) Do colleges with larger full time enrollments have lower grad rates ? - No

2)Is it diff for public/private institutions? private college and the logarithm of the number of full-time undergraduates has a statistically significant effect on predicting the graduation rate.