PREDICTING ALUMNI GIVING RATE

Topic

- Alumni Giving - To investigate the factors influencing alumni donation.

Goal

- Predict the alumni giving rate using the independent variables.
- Show the impact of the above variables on the alumni donation rate.

Variables

Response Variable

- Alumni Giving Rate – Percentage of alumni who donated to the university.

Predictors

- Graduation Rate
- % of Classes Under 20
- Student-Faculty Ratio

Dataset

	A	B	C	D	E	F
1	University	State	Graduation Rate	% of Classes Under 20	Student-Faculty Ratio	Alumni Giving Rate
2	Boston College	MA	85	39	13	25
3	Brandeis University	MA	79	68	8	33
4	Brown University	RI	93	60	8	40
5	California Institute of Technology	CA	85	65	3	46
6	Carnegie Mellon University	PA	75	67	10	28
7	Case Western Reserve Univ.	OH	72	52	8	31
8	College of William and Mary	VA	89	45	12	27
9	Columbia University	NY	90	69	7	31
10	Cornell University	NY	91	72	13	35
11	Dartmouth College	NH	94	61	10	53
12	Duke University	NC	92	68	8	45
13	Emory University	GA	84	65	7	37
14	Georgetown University	DC	91	54	10	29
15	Harvard University	MA	97	73	8	46
16	Johns Hopkins University	MD	89	64	9	27
17	Lehigh University	PA	81	55	11	40
18	Massachusetts Inst. of Technology	MA	92	65	6	44
19	New York University	NY	72	63	13	13
20	Northwestern University	IL	90	66	8	30
21	Pennsylvania State Univ.	PA	80	32	19	21
22	Princeton University	NJ	95	68	5	67
23	Rice University	TX	92	62	8	40
24	Stanford University	CA	92	69	7	34
25	Tufts University	MA	87	67	9	29
26	Tulane University	LA	72	56	12	17
27	U. of California–Berkeley	CA	83	58	17	18
28	U. of California–Davis	CA	74	32	19	7
29	U. of California–Irvine	CA	74	42	20	9
30	U. of California–Los Angeles	CA	78	41	18	13
31	U. of California–San Diego	CA	80	48	19	8
32	U. of California–Santa Barbara	CA	70	45	20	12
33	U. of Chicago	IL	84	65	4	36
34	U. of Florida	FL	67	31	23	19
35	U. of Illinois–Urbana Champaign	IL	77	29	15	23
36	U. of Michigan–Ann Arbor	MI	83	51	15	13
37	U. of North Carolina–Chapel Hill	NC	82	40	16	26
38	U. of Notre Dame	IN	94	53	13	49
39	U. of Pennsylvania	PA	90	65	7	41

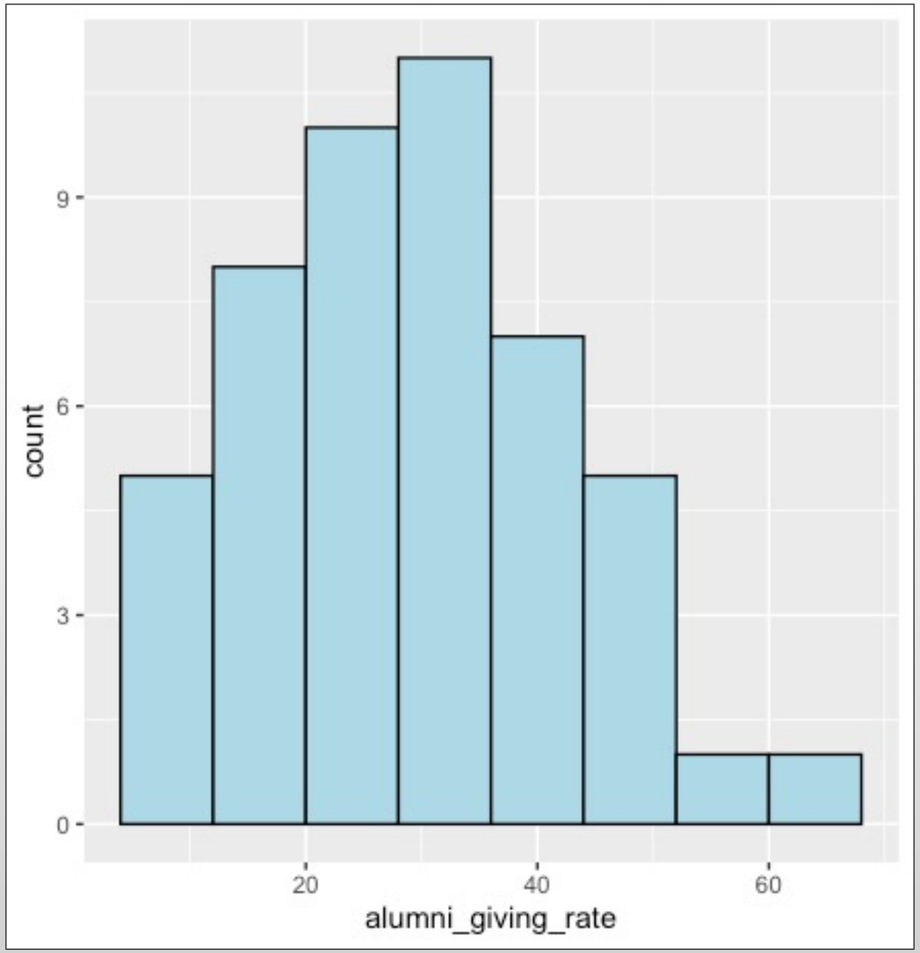
Data summary

- Summary of all the variables
- Alumni giving rate can vary from 7% to 67%

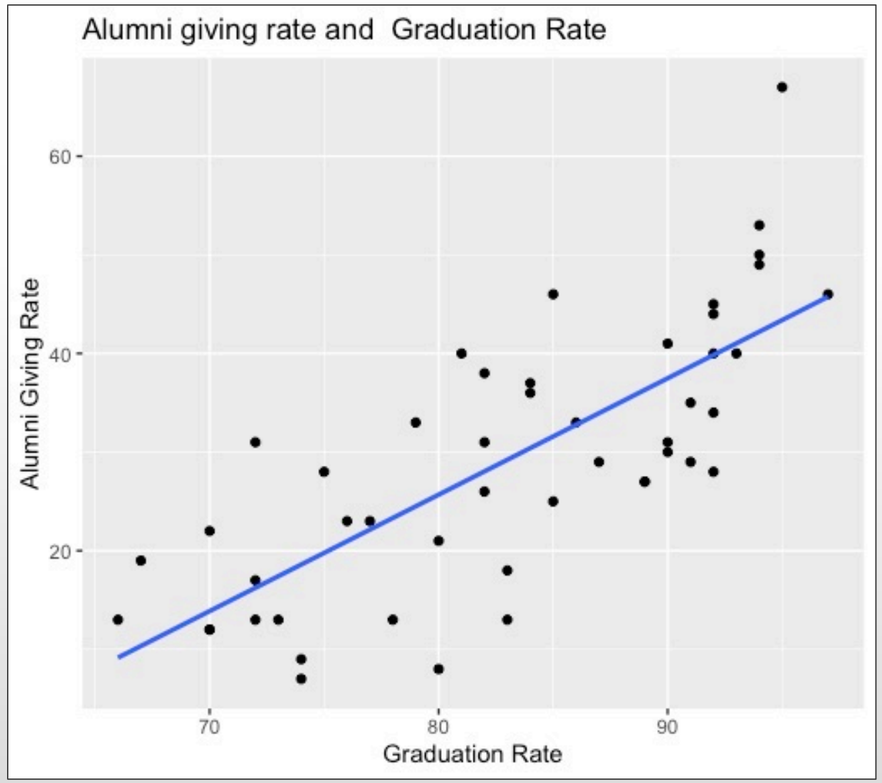
```
 Graduation_rate percent_of_classes_under_20 student_faculty_ratio alumni_giving_rate
 Min.   :66.00   Min.   :29.00               Min.   : 3.00          Min.   : 7.00
 1st Qu.:75.75   1st Qu.:44.75               1st Qu.: 8.00          1st Qu.:18.75
 Median :83.50   Median :59.50               Median :10.50          Median :29.00
 Mean   :83.04   Mean   :55.73               Mean   :11.54          Mean   :29.27
 3rd Qu.:91.00   3rd Qu.:66.25               3rd Qu.:13.50          3rd Qu.:38.50
 Max.   :97.00   Max.   :77.00               Max.   :23.00          Max.   :67.00
```

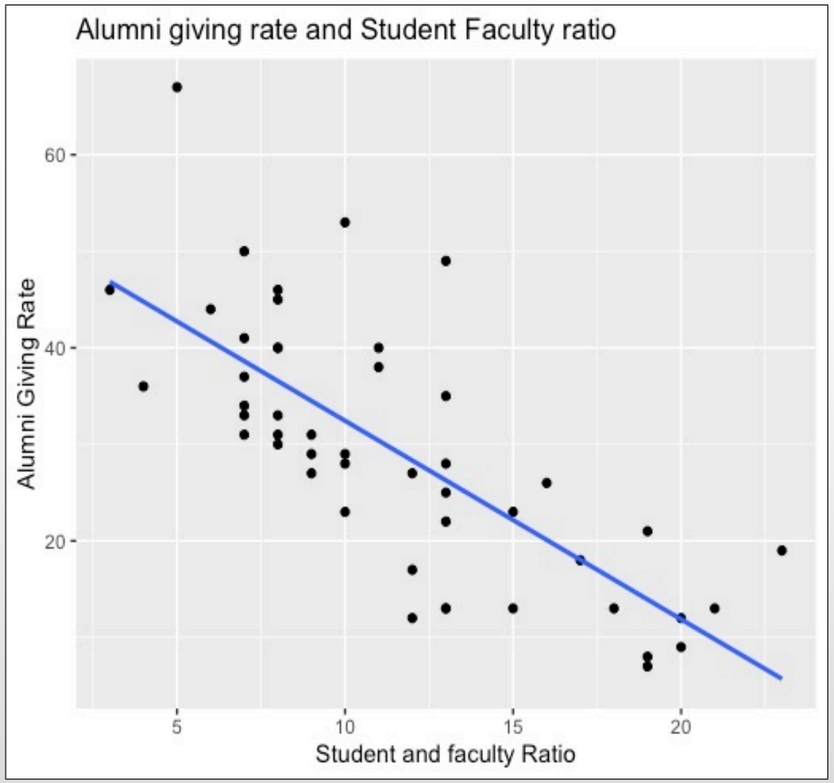
Exploring data

- Slightly right skewed



Exploring data





Correlation

- Positive relationship between graduation rate and Alumni giving rate
- Negative relation with student faculty ratio

```
> alumni_reg %>%
+   select(alumni_giving_rate,Graduation_rate, percent_of_classes_under_20,student_faculty_ratio)%>%
+   cor()
                          alumni_giving_rate Graduation_rate percent_of_classes_under_20 student_faculty_ratio
alumni_giving_rate                 1.0000000       0.7559436                  0.6456504            -0.7423975
Graduation_rate                    0.7559436       1.0000000                  0.5827884            -0.6049379
percent_of_classes_under_20        0.6456504       0.5827884                  1.0000000            -0.7855593
student_faculty_ratio             -0.7423975      -0.6049379                 -0.7855593             1.0000000
```

```
Call:
lm(formula = AlumniGivingRate ~ GraduatRate + NumClassesU20 +
    StdntFactyRatio, data = fulldata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9800  -5.9024  -0.6273   3.7644  20.6281

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -20.72013   17.52137  -1.183  0.24333
GraduatRate      0.74818    0.16596   4.508  4.8e-05 ***
NumClassesU20    0.02904    0.13932   0.208  0.83584
StdntFactyRatio -1.19201    0.38672  -3.082  0.00354 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.61 on 44 degrees of freedom
Multiple R-squared:  0.6999,    Adjusted R-squared:  0.6795
F-statistic: 34.21 on 3 and 44 DF,  p-value: 1.432e-11
```

	p-values	
GraduatRate	4.80E-05	***
NumClassesU20	0.83	
StdntFavtyRatio	0.00354	**

The independent variable “% of Classes Under 20” is removed from the model.

```
Call:
lm(formula = AlumniGivingRate ~ GraduatRate + StdntFactyRatio,
    data = fulldata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9304  -6.1594  -0.5521   3.5910  20.5412

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -19.1063    15.5501  -1.229    0.226
GraduatRate       0.7557     0.1602   4.717 2.35e-05 ***
StdntFactyRatio  -1.2460     0.2843  -4.382 6.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.528 on 45 degrees of freedom
Multiple R-squared:  0.6996,    Adjusted R-squared:  0.6863
F-statistic: 52.41 on 2 and 45 DF,  p-value: 1.765e-12
```

	p-values	
GraduatRate	2.35E-05	***
StdntFavtyRatio	6.95E-05	***

The regression equation is Y = 0.7557 x1 - 1.2460 x2 – 19.1063.

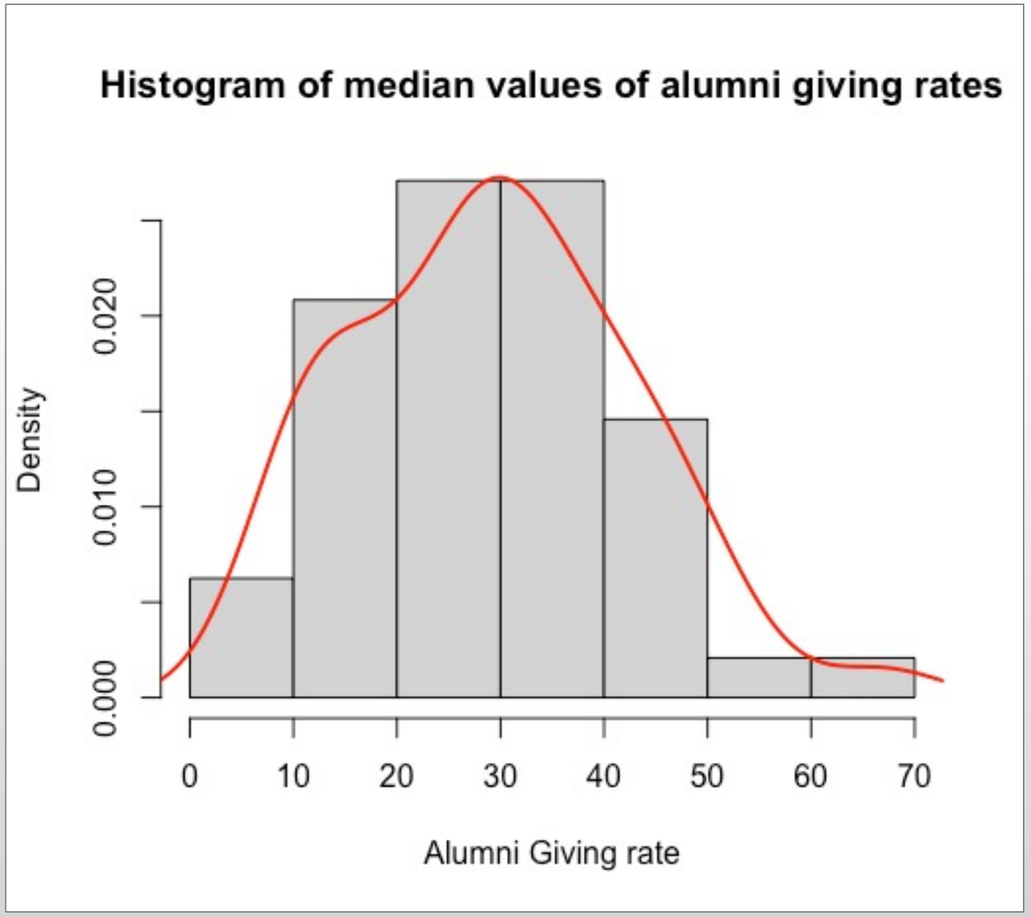
Model 1 and Model 2

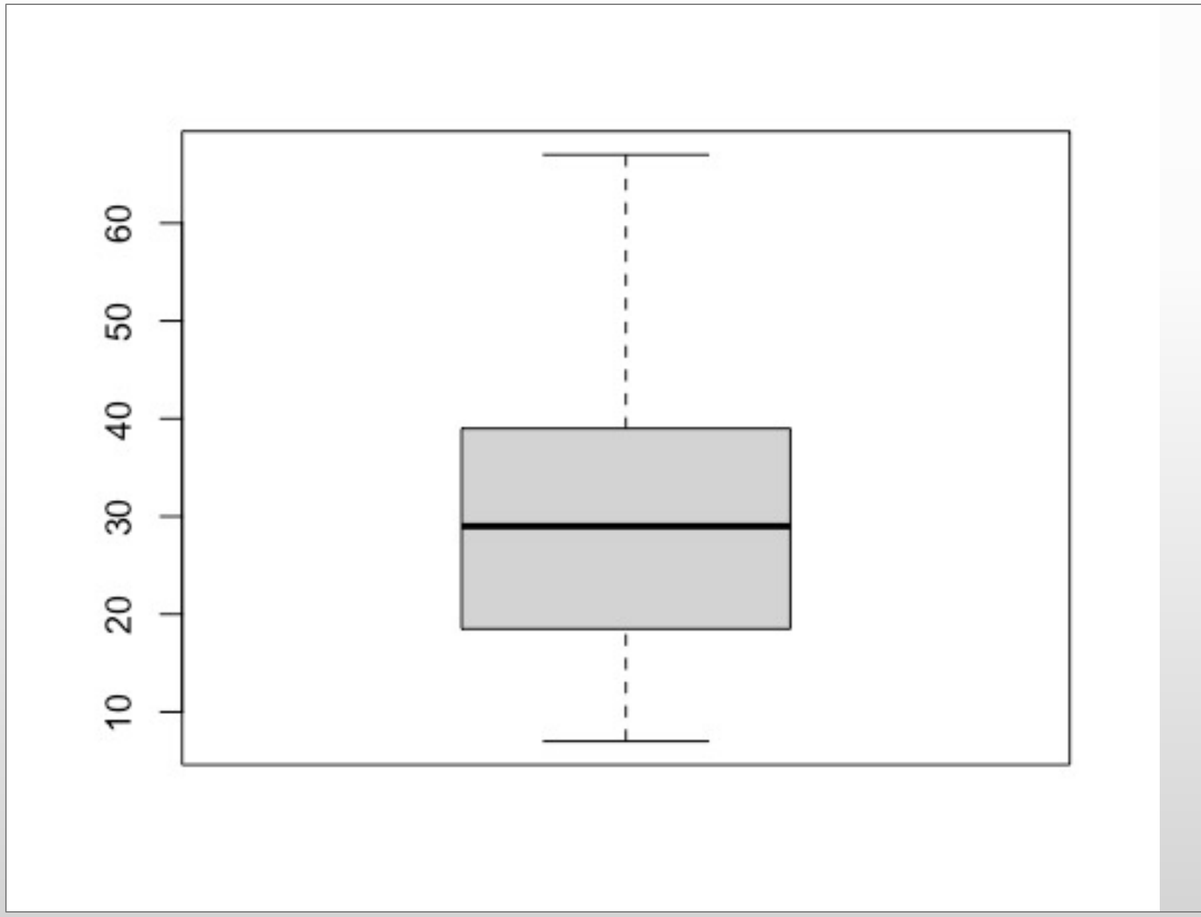
```
> #Fitting linear model for model 1 (including all the 3 predictors)
> mod1<-lm(alumni_giving_rate~.,data=train.df)
> summary(mod1)

Call:
lm(formula = alumni_giving_rate ~ ., data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0634  -6.0319  -0.7689   3.4742  19.9522

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -20.84655   23.37824  -0.892 0.378817
Graduation_rate              0.79618    0.20933   3.803 0.000567 ***
percent_of_classes_under_20 -0.01997    0.17399  -0.115 0.909284
student_faculty_ratio       -1.27698    0.48613  -2.627 0.012833 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.146 on 34 degrees of freedom
Multiple R-squared:  0.7021,    Adjusted R-squared:  0.6758
F-statistic:  26.7 on 3 and 34 DF,  p-value: 4.631e-09

> preds.mod1 <- predict(mod1, newdata = test.df)
> MSE1 <- mean((preds.mod1 - test.df$alumni_giving_rate)^2)
> RMSE1 <- sqrt(MSE1)
> print(RMSE1)
[1] 5.510862
```

```
> mod2<-lm(alumni_giving_rate~Graduation_rate+student_faculty_ratio,data=train.df)
> summary(mod2)

Call:
lm(formula = alumni_giving_rate ~ Graduation_rate + student_faculty_ratio,
    data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0841  -5.9503  -0.8179   3.3338  19.9974

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -22.1257    20.2600  -1.092 0.282256
Graduation_rate        0.7929     0.2045   3.878 0.000444 ***
student_faculty_ratio  -1.2403     0.3613  -3.433 0.001551 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.031 on 35 degrees of freedom
Multiple R-squared:  0.7019,    Adjusted R-squared:  0.6849
F-statistic: 41.21 on 2 and 35 DF,  p-value: 6.316e-10

> preds.mod2 <- predict(mod2, newdata = test.df[,c("Graduation_rate","student_faculty_ratio")]
> MSE2 <- mean((preds.mod2 - test.df$alumni_giving_rate)^2)
> RMSE2 <- sqrt(MSE2)
> print(RMSE2)
[1] 5.463596
```

Exploratory Data Analysis

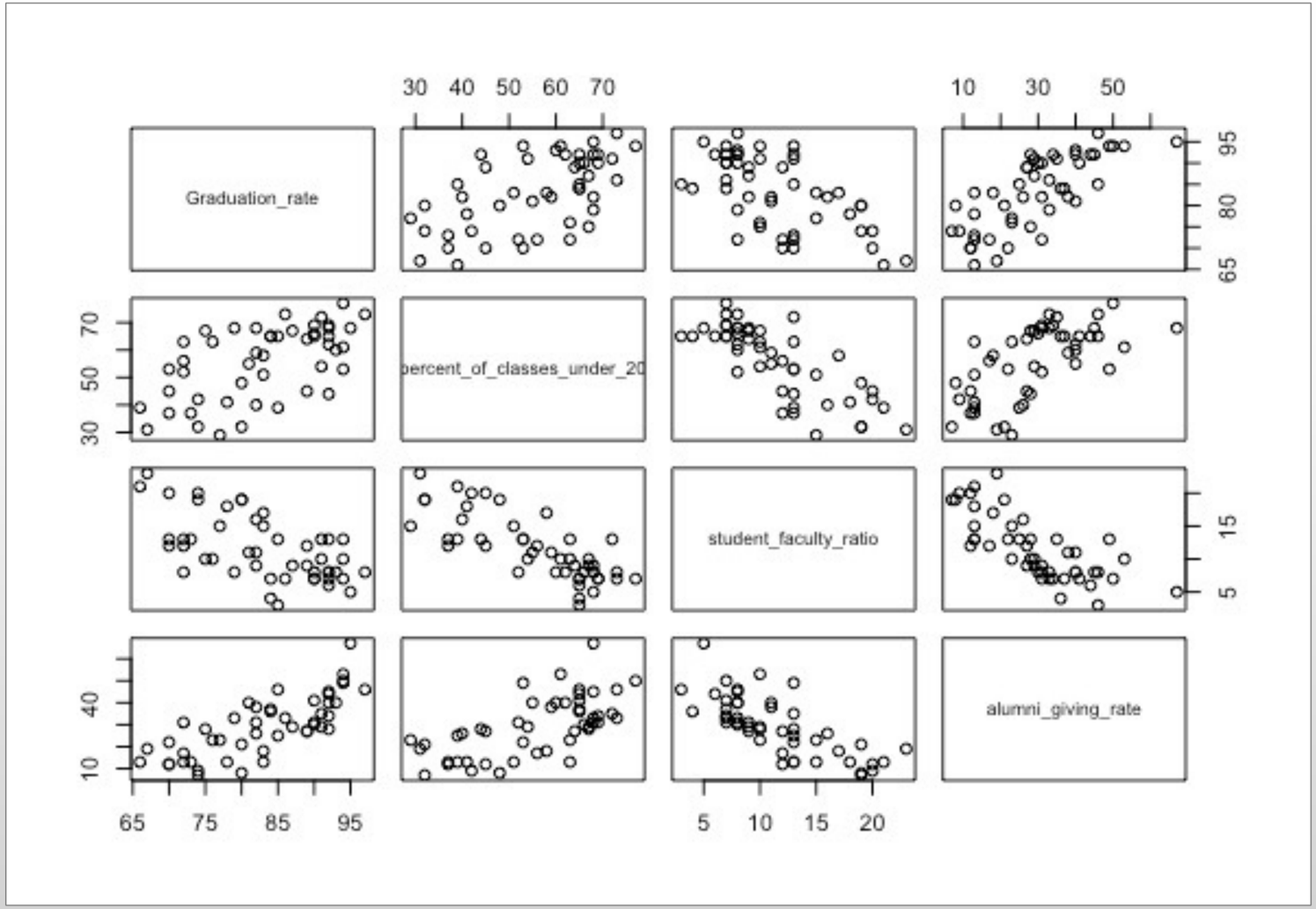
- Alumni giving rate





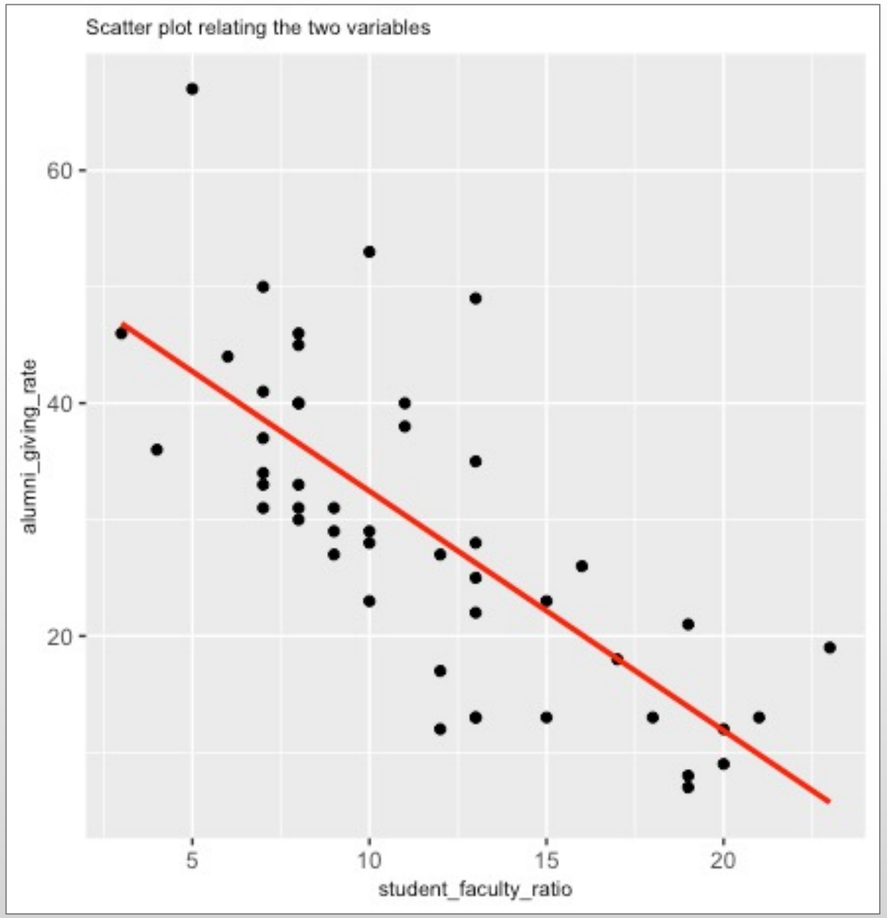
EDA (Continued)

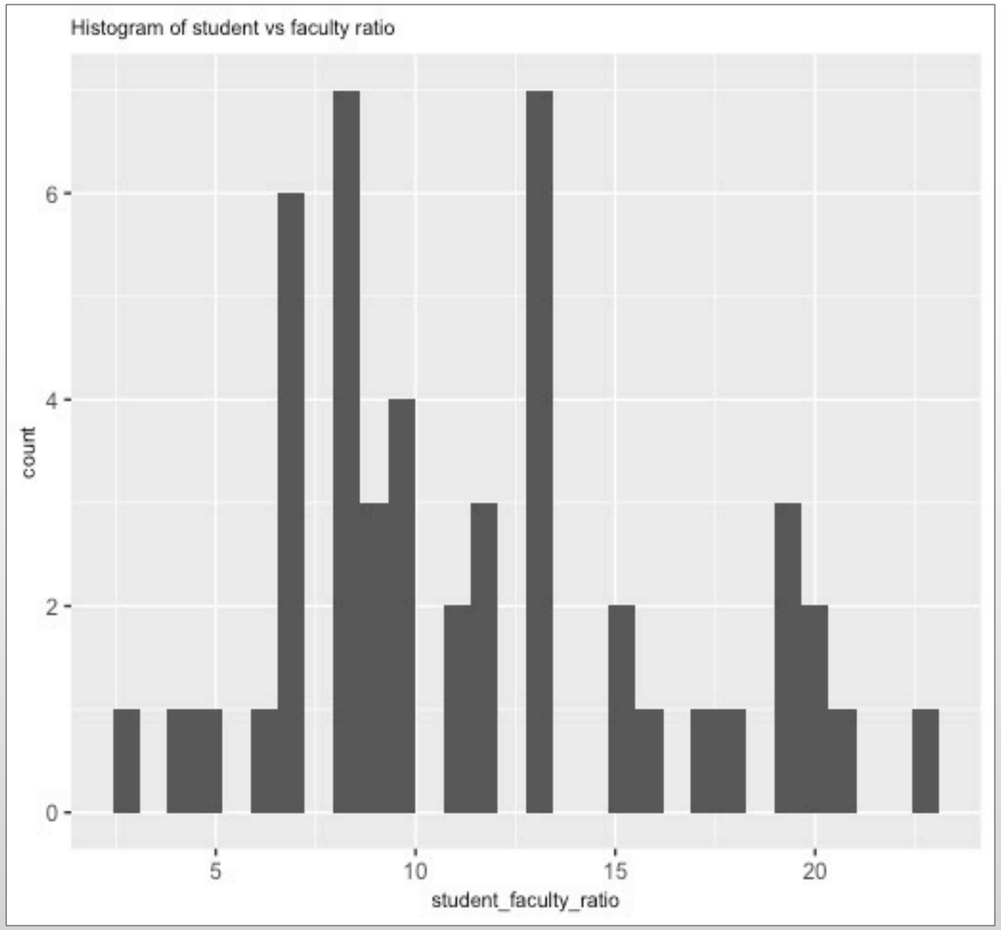
- Side by side plot



EDA (Continued)

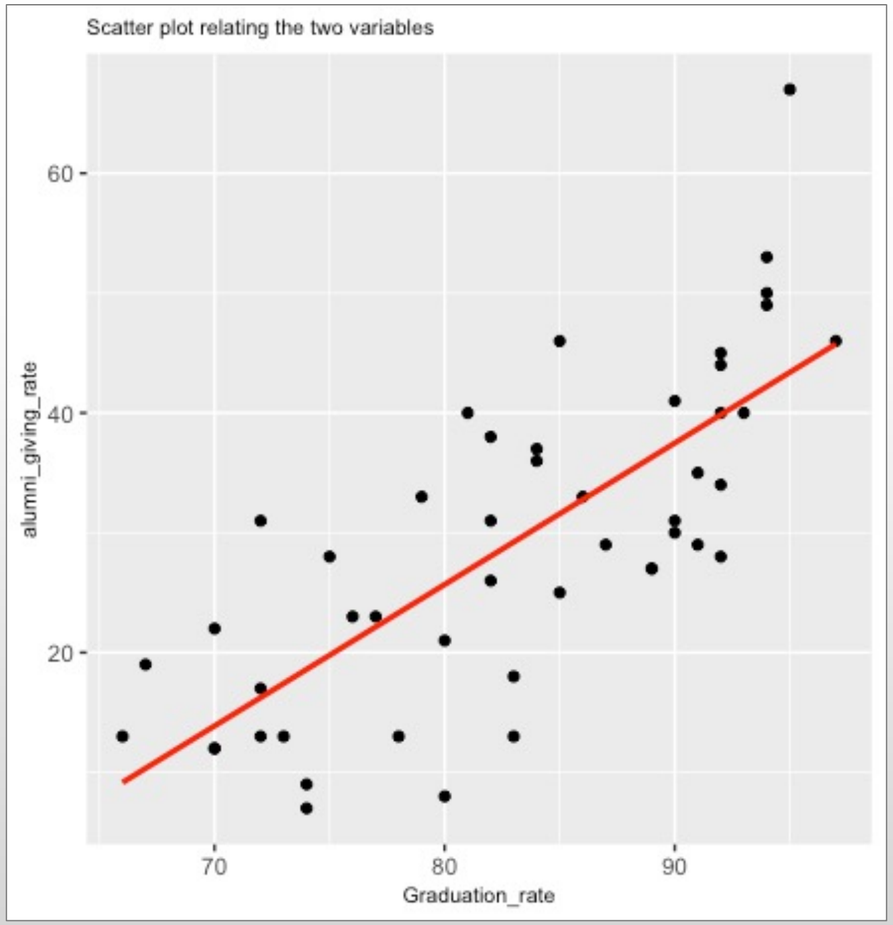
- Distribution between alumni giving rate and student faculty ratio

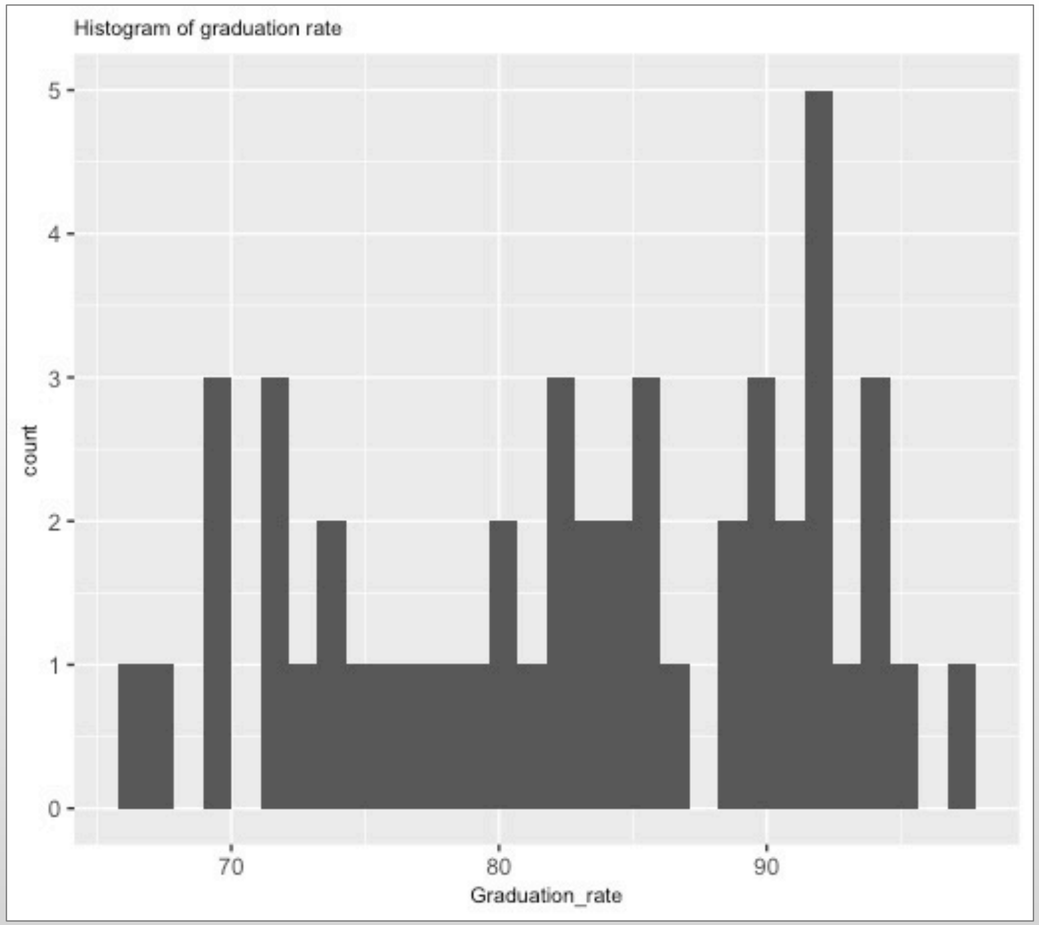




EDA (Continued)

- Distribution between alumni giving rate and graduation rate





Comparing further Models : Model 3

- EDA inference result : Applying squared transformation on student-faculty ratio.
- Root mean squared value increased
- Not a better model

```
> alumni_reg$student_faculty_ratio_new=(alumni_reg$student_faculty_ratio)^2
> n = nrow(alumni_reg)
> p = ncol(alumni_reg)
> set.seed(123)
> train.index <- sample(row.names(alumni_reg), floor(0.8*n))
> test.index <- setdiff(row.names(alumni_reg), train.index)
> train.df <- alumni_reg[train.index,]
> test.df <- alumni_reg[test.index,]
> mod3<-lm(alumni_giving_rate~Graduation_rate+student_faculty_ratio_new, data=train.df)
> preds.mod3 <- predict(mod3, newdata = test.df[,c("Graduation_rate","student_faculty_ratio_new")])
> MSE3 <- mean((preds.mod3 - test.df$alumni_giving_rate)^2)
> RMSE3 <- sqrt(MSE3)
> print(RMSE3)
[1] 5.713069
>
```

Comparing Further Models : Model 4

- Applying squared transformation on graduation rate
- Root mean squared value decreased
- Prediction increased by 15 percent
- Better model

```
alumni_reg$Graduation_rate_new=(alumni_reg$Graduation_rate)^2
n = nrow(alumni_reg)
p = ncol(alumni_reg)
set.seed(123)
train.index <- sample(row.names(alumni_reg), floor(0.8*n))
test.index <- setdiff(row.names(alumni_reg), train.index)
train.df <- alumni_reg[train.index,]
test.df <- alumni_reg[test.index,]
mod4<-lm(alumni_giving_rate~Graduation_rate_new+student_faculty_ratio, data=train.df)
preds.mod4 <- predict(mod4, newdata = test.df[,c("Graduation_rate_new","student_faculty_ratio")])
MSE4 <- mean((preds.mod4 - test.df$alumni_giving_rate)^2)
RMSE4 <- sqrt(MSE4)
print(RMSE4)
1] 5.409025
```

Comparing Further Models : Model 5

- To further increase prediction accuracy: Applying quadratic transformation on student-faculty ratio
- RMSE value further increased

```
> alumni_reg$student_faculty_ratio_new=(alumni_reg$student_faculty_ratio)^2
> n = nrow(alumni_reg)
> p = ncol(alumni_reg)
> set.seed(123)
> train.index <- sample(row.names(alumni_reg), floor(0.8*n))
> test.index <- setdiff(row.names(alumni_reg), train.index)
> train.df <- alumni_reg[train.index,]
> test.df <- alumni_reg[test.index,]
> mod5<-lm(alumni_giving_rate~ Graduation_rate+ student_faculty_ratio+student_faculty_ratio_new,
+ data=train.df)
> preds.mod5 <- predict(mod5, newdata = test.df[,c("Graduation_rate",
+ "student_faculty_ratio","student_faculty_ratio_new")])
> MSE5 <- mean((preds.mod5 - test.df$alumni_giving_rate)^2)
> RMSE5 <- sqrt(MSE5)
> print(RMSE5)
[1] 6.162483
```

Final Model : Model 6

- To further increase prediction accuracy: Applying quadratic transformation on graduation rate.
- The root mean squared value was the least among 6 models.

```
> alumni_reg$Graduation_rate_new=(alumni_reg$Graduation_rate)^2
> mod6<-lm(alumni_giving_rate~ Graduation_rate+student_faculty_ratio+Graduation_rate_new,data=train.df)
> preds.mod6 <- predict(mod6, newdata = test.df[,c("Graduation_rate",
+                                                  "student_faculty_ratio","Graduation_rate_new")])
> MSE6 <- mean((preds.mod6 - test.df$alumni_giving_rate)^2)
> RMSE6 <- sqrt(MSE6)
> print(RMSE6)
[1] 5.228536
```

```
>  print(c(MSE1, MSE2, MSE3, MSE4, MSE5, MSE6))
[1] 30.36960 29.85088 32.63915 29.25755 37.97620 27.33758
>  print(c(RMSE1, RMSE2, RMSE3, RMSE4, RMSE5, RMSE6))
[1] 5.510862 5.463596 5.713069 5.409025 6.162483 5.228536
```

Final Linear Regression Model

alumni_giving_rate = 294.331 + 0.047* Graduation_rate^2 – 6.92 * Graduation_rate - 1.348 * student_faculty_ratio

```
Call:
lm(formula = alumni_giving_rate ~ Graduation_rate_new + Graduation_rate +
    student_faculty_ratio, data = full.df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3867  -5.1909  -0.6283   3.8944  15.7019

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           294.33105  104.69891   2.811  0.00734 **
Graduation_rate_new     0.04666    0.01544   3.022  0.00417 **
Graduation_rate        -6.91997    2.54416  -2.720  0.00931 **
student_faculty_ratio  -1.34835    0.26383  -5.111 6.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.928 on 44 degrees of freedom
Multiple R-squared:  0.7513,    Adjusted R-squared:  0.7343
F-statistic:  44.3 on 3 and 44 DF,  p-value: 2.387e-13
```

Prediction Result

University	State	Graduation Rate	% of Classes Under 20	Student-Faculty Ratio	Alumni Giving Rate	Model6	Percentage difference
Boston College	MA	85	39	13	25	25.72355	3%
Massachusetts Inst. of Technology	MA	92	65	6	44	44.53395	1%
Rice University	TX	92	62	8	40	41.83725	5%
U. of California–Berkeley	CA	83	58	17	18	18.49233	3%
U. of California–Santa Barbara	CA	70	45	20	12	11.60015	-3%
U. of Chicago	IL	84	65	4	36	36.89313	2%
U. of Pennsylvania	PA	90	65	7	41	40.0413	-2%
U. of Rochester	NY	76	63	10	23	24.43799	6%
U. of Southern California	CA	70	53	13	22	21.0386	-4%
U. of Texas–Austin	TX	66	39	21	13	12.54864	-3%

Conclusion & Recommendation

- The alumni giving rate can be maximized by
  - Reducing student-faculty ratio
  - Increased graduation rate in most cases
- Possible recommendation
  - Hire more staff
  - Increasing university facilities and resources