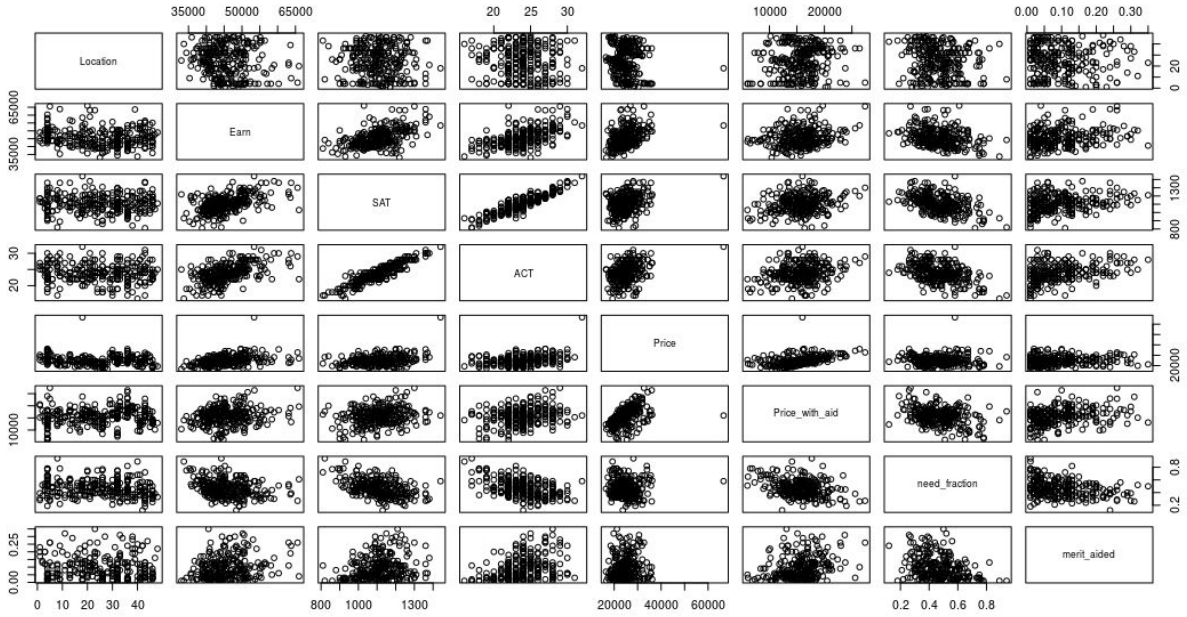
Preliminary Analysis:

Preprocessing and Plotting the Data:

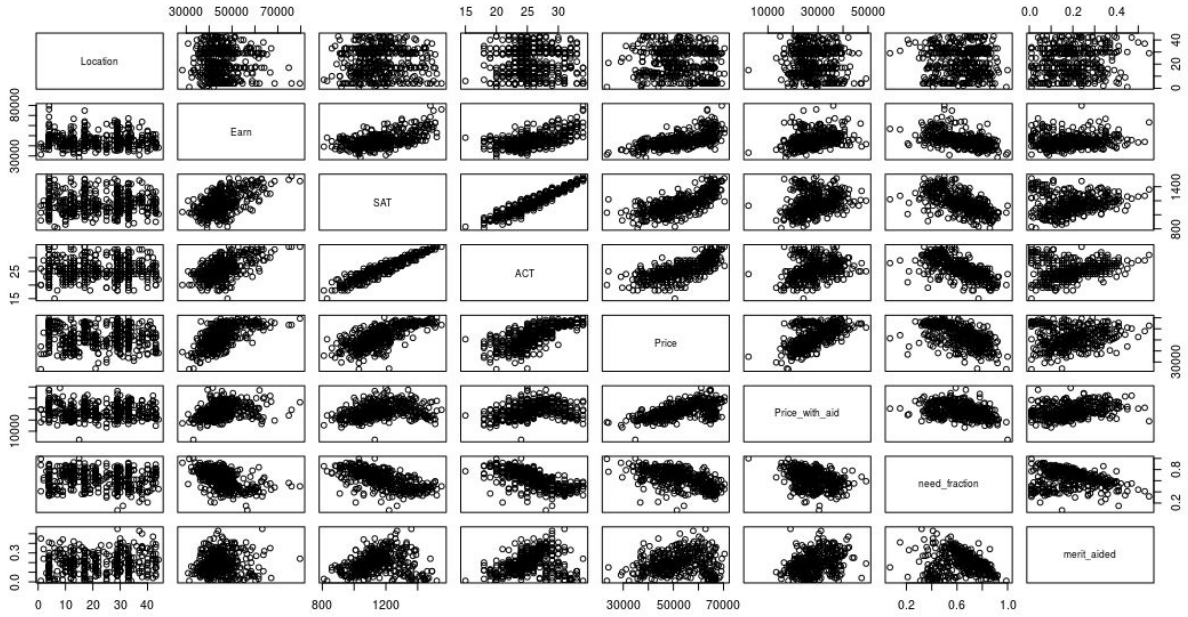
```
> data.raw <- read.csv("GradEarn.csv")
> data.raw[data.raw==""] <- NA
```

Segregating Public from Private Data and Comparing Plots:

```
> data.public <- subset(data.raw, Public == "1")
> data.public <- subset(data.public, select=c(4,5,6,7,8,9,10,11))
> plot(data.public)
```

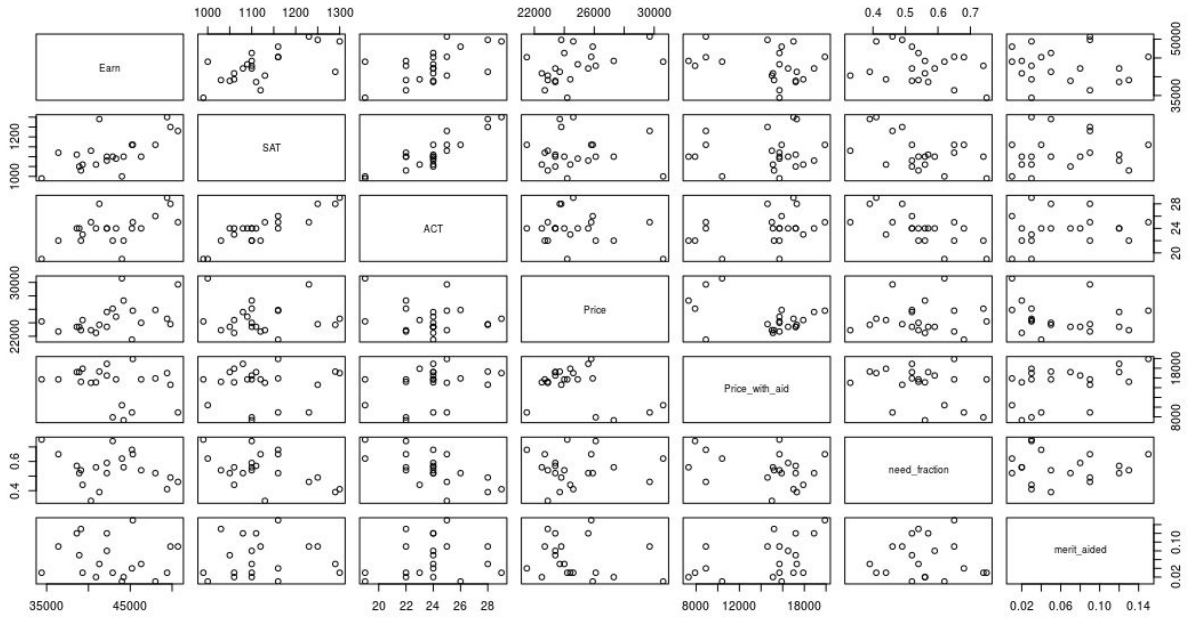



```
> data.private <- subset(data.raw, Public == "0")
> data.private <- subset(data.private, select=c(4,5,6,7,8,9,10,11))
> plot(data.private)
```

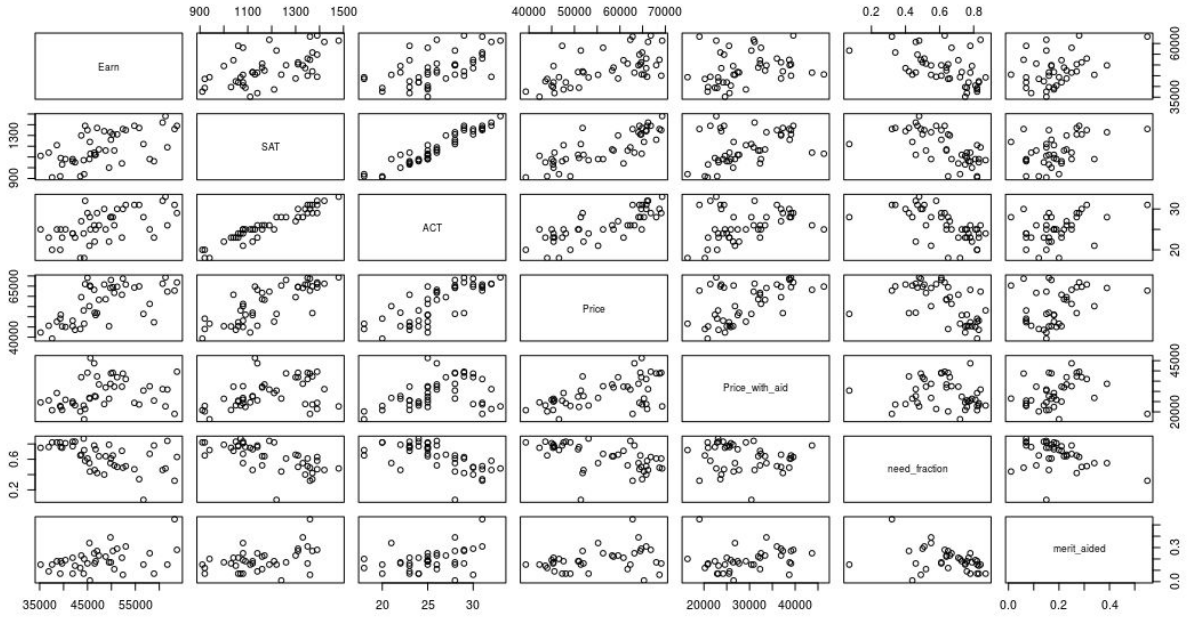


Extracting and Plotting Data for NY Public Schools:

```
> data.public.ny <- subset(data.public, Location == "NY")
> data.public.ny <- subset(data.public.ny, select=c(2,3,4,5,6,7,8))
> plot(data.public.ny)
```




```
> data.private.ny <- subset(data.private, Location == "NY")
> data.private.ny <- subset(data.private.ny, select=c(2,3,4,5,6,7,8))
> plot(data.private.ny)
```



Simple Linear Regression: Public Schools in NY: **need_fraction ~ ACT**

```
> simple.regression <- lm(need_fraction ~ ACT, data = data.public.ny)
> summary(simple.regression)
```

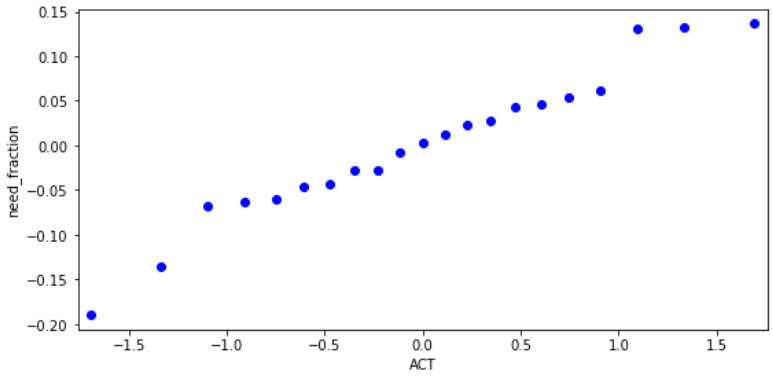
```
Call:
lm(formula = need_fraction ~ ACT, data = data.public.ny)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189680 -0.045582  0.002451  0.046222  0.136222

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.220498   0.179008   6.818 1.65e-06 ***
ACT         -0.028033   0.007433  -3.771  0.00129 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08506 on 19 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4281,    Adjusted R-squared:  0.398
F-statistic: 14.22 on 1 and 19 DF,  p-value: 0.001292
```

QQ Plot:



Simple Linear Regression: Private Schools in NY: **need_fraction ~ ACT**

```
> simple.regression <- lm(need_fraction ~ ACT, data = data.private.ny)
> summary(simple.regression)
```

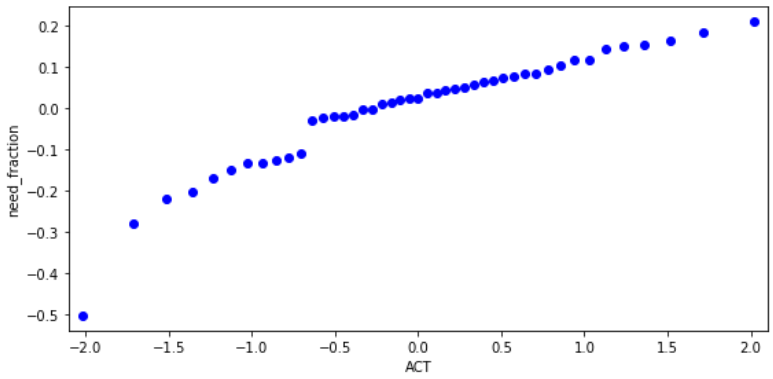
```
Call:
lm(formula = need_fraction ~ ACT, data = data.private.ny)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50351 -0.02777  0.02357  0.08436  0.21076

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.353753   0.143842   9.411 5.27e-12 ***
ACT         -0.027866   0.005466  -5.098 7.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1381 on 43 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3767,    Adjusted R-squared:  0.3622
F-statistic: 25.99 on 1 and 43 DF,  p-value: 7.341e-06
```

QQ Plot:



Multiple Linear Regression: Public Schools in NY: We model Graduate Earnings as a linear combination of other input features.

```
> multiple.regression <- lm(Earn ~ SAT + ACT + Price + Price_with_aid +
need_fraction + merit_aided, data = data.public.ny)
> summary(multiple.regression)
```

```
Call:
lm(formula = Earn ~ SAT + ACT + Price + Price_with_aid + need_fraction +
    merit_aided, data = data.public.ny)

Residuals:
    Min      1Q  Median      3Q     Max
-3270.2 -1017.9   362.9  1122.1  4211.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.779e+04  1.550e+04  -1.148  0.27169
SAT            -2.569e+01  1.651e+01  -1.556  0.14371
ACT             2.601e+03  6.751e+02   3.853  0.00200 **
Price           1.133e+00  2.797e-01   4.053  0.00137 **
Price_with_aid -5.276e-01  2.001e-01  -2.637  0.02050 *
need_fraction   1.238e+04  8.116e+03   1.525  0.15117
merit_aided     2.296e+03  1.350e+04   0.170  0.86753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2232 on 13 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8322,    Adjusted R-squared:  0.7548
F-statistic: 10.75 on 6 and 13 DF,  p-value: 0.0002129
```

From this result, it is evident that modelling with all these features isn’t significantly better than using just ACT, Price and Price_with_aid, and this claim of ours can be proven by the following results:

```
> multiple.regression <- lm(Earn ~ SAT + need_fraction + merit_aided, data
= data.public.ny)
> summary(multiple.regression)
```

```
Call:
lm(formula = Earn ~ SAT + need_fraction + merit_aided, data = data.public.ny)

Residuals:
    Min      1Q  Median      3Q     Max
-7599.2 -1778.0   357.8  2653.7  4479.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     7721.62   16343.96   0.472  0.64299
SAT               32.92      11.27   2.921  0.00999 **
need_fraction  -1366.55    9973.36  -0.137  0.89272
merit_aided   -15186.42   19409.20  -0.782  0.44539
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3649 on 16 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4482,    Adjusted R-squared:  0.3447
F-statistic: 4.331 on 3 and 16 DF,  p-value: 0.02047
```

Notice the change in the Adjusted R-squared value and the p-value.

```
> multiple.regression <- lm(Earn ~ need_fraction + merit_aided, data =
data.public.ny)
> summary(multiple.regression)
```

```
Call:
lm(formula = Earn ~ need_fraction + merit_aided, data = data.public.ny)

Residuals:
    Min      1Q  Median      3Q     Max
-6061.2 -3860.4   367.8  3593.8  6306.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     53238       5933   8.973 7.39e-08 ***
need_fraction  -17193      10060  -1.709    0.106
merit_aided    -10394      23234  -0.447    0.660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4384 on 17 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1538,    Adjusted R-squared:  0.05422
F-statistic: 1.545 on 2 and 17 DF,  p-value: 0.2419
```

Multiple Linear Regression: Private Schools in NY: We repeat the above steps with the data.private.ny dataframe.

```
> multiple.regression <- lm(Earn ~ SAT + ACT + Price + Price_with_aid +
need_fraction + merit_aided, data = data.private.ny)
> summary(multiple.regression)
```

Results:

```
Call:
lm(formula = Earn ~ SAT + ACT + Price + Price_with_aid + need_fraction +
    merit_aided, data = data.private.ny)

Residuals:
    Min      1Q  Median      3Q     Max
-6679.9 -3299.8  -707.3  1173.7 17115.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    36276.6778 13936.7488   2.603   0.0139 *
SAT              -20.0709    22.0172  -0.912   0.3688
ACT              569.1929   810.8717   0.702   0.4878
Price              0.4595     0.1941   2.368   0.0241 *
Price_with_aid    -0.0524     0.2337  -0.224   0.8240
need_fraction  -9907.5467  7184.5644  -1.379   0.1775
merit_aided    14579.2467 10686.3040   1.364   0.1820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5896 on 32 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.4444,    Adjusted R-squared:  0.3402
F-statistic: 4.266 on 6 and 32 DF,  p-value: 0.002883
```

```
> multiple.regression <- lm(Earn ~ SAT + ACT + Price_with_aid +
need_fraction + merit_aided, data = data.private.ny)
> summary(multiple.regression)
```

```
Call:
lm(formula = Earn ~ SAT + ACT + Price_with_aid + need_fraction +
    merit_aided, data = data.private.ny)

Residuals:
    Min      1Q  Median      3Q     Max
-9838.6 -4315.1  -560.6  1978.6 16738.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.220e+04  1.464e+04   2.883  0.00688 **
SAT            -1.427e+01  2.336e+01  -0.611  0.54544
ACT             7.996e+02  8.594e+02   0.930  0.35888
Price_with_aid  2.143e-01  2.186e-01   0.980  0.33411
need_fraction  -1.205e+04  7.608e+03  -1.584  0.12268
merit_aided     1.637e+04  1.138e+04   1.439  0.15959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6294 on 33 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.347,     Adjusted R-squared:  0.2481
F-statistic: 3.508 on 5 and 33 DF,  p-value: 0.01186
```

Conclusion: To predict Graduate Earnings for Public Schools in NY, ACT, Price and Price_with_aid are valuable features that produce significant results. For Private Schools in NY too, the Price variable plays a pivotal role in generating significant results. Thus, it can be concluded that Graduate Earnings for the state of NY greatly depend on the cost of education.

Effect of Missing Values: