Goal of analysis

In this report, we have analysed the U.S. News & World Report's Guide to America's Best Colleges (U.S. News) dataset which contains information on tuition, room & boarding costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for 1300+ colleges and universities. We have linked this with American Association of University Professors (AAUP) to draw in information on average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor ranks.

The goal of this analysis was to determine which of the characteristics of a university tended to have the greatest impact on the average SAT entrance mark. We broke this question down into two parts:

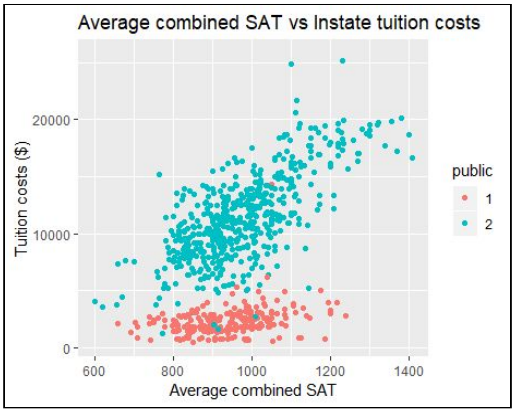
1) Which socio-economic variables affect the average SAT score at a state level?
2) What features of the universities affect the average SAT score?

Data collection and exploratory analysis

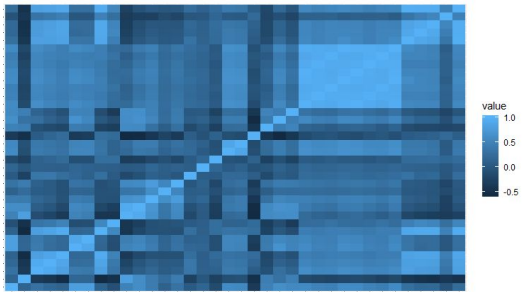
University-specific influences on SAT scores

The U.S. News and AAUP datasets were provided both published by reputable organisations in publicly available newspapers and journals. The dataset contained a number of quality issues which we have dealt with in the following ways:

- Graduation rate being 118% for Cazenovia College - this was normalised back to 100%
- Missing values (NA) for the response variable average SAT for 523 colleges. These colleges were excluded from the analysis because they did not provide statistical insight into determining which of the variables affected SAT scores
- Missing values (NA) for the other predictor variables. We performed 3 sets of analysis based on mean, median and k-nearest neighbour imputation.

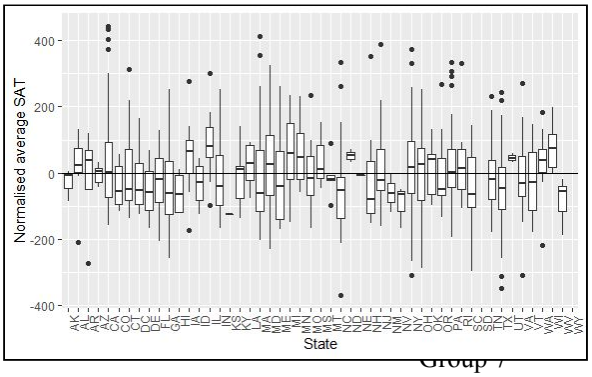


Initial scatter plots helped us narrow down the set of variables that were likely to be appropriate predictors of the average SAT score of a college. In particular, we noticed that the relevance of certain variables tended to depend on whether the college was publicly or privately funded. For instance, the average combined SAT was seemingly uniform across tuition cost for publicly funded colleges, however there was an apparently linear trend with in-state tuition cost and privately funded colleges.



Furthermore, we noticed significant amounts of collinearity between salary and staffing in formation as seen by the lightest shade of blue on the right. This lead to our decision to use only the average salary across all ranks.

On the right, a histogram showing the absolute difference of state SAT scores from the national average is shown. It can be reasonably inferred at a 90% level that not many states have a average that is significantly different from the national average.

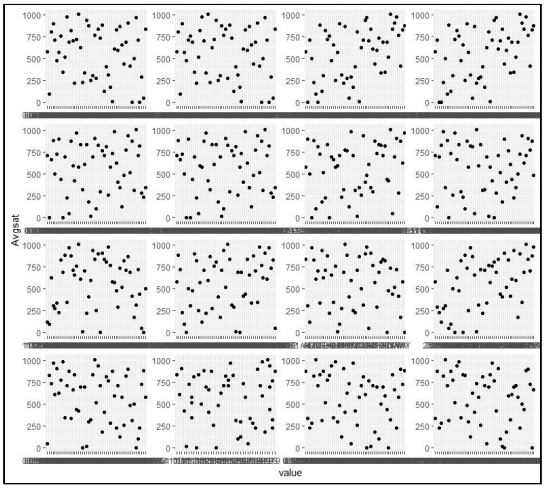


Socio-economic influences on SAT scores

Socio-economic information from 1995 was collected at a state level from various government websites. The data was of a high quality, with no errors aside from the state of Washington D.C. missing a crime index. A number of normalisations we performed to make the data comparable, as described in the table below.

Economic variable	**Source**	**Description**
Median income	Census (1995)	The median individual income was used.
Crime rate	Bureau of Justice Statistics	The crime index which is based on the violence, property damage, murder, rape, robbery, assault, burglary, etc. rates per person in a state was used.
Ethnicity	Census (1995)	The ethnic composition of a state broken down in percentages for White, Black, American-Indian and Asian Males and Females was used. An additional variable which states the proportion of Hispanics was included.
Poverty	Census (1995)	The estimated percentage of people in poverty was used.
Government education taxation and spending	Census (1995)	State education revenue and expenditure per person was used.

The average college SAT was aggregated up to a state level, weighting each college SAT by the proportion of total state students attending a particular college. A scatter plot displaying the average state SAT against the socio-economic variables displays that there seems to be no apparent trend relationship.



Model choice

University-specific model

From the initial exploratory analysis of the variables, we inferred that most of the variables either have a linear relationship with SAT scores or no relationship at all. To estimate how the features of the university that affects the average SAT score of their students, we chose a linear model of average SAT score against different university characteristics.

$$\textbf{Average College SAT} = \mathbf{Xb} + \varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 I)$$

To model the "academic quality" of the students, we chose the average combined SAT score as the response variable. We assumed that the error term ε_i for each observation was independent and had normal distribution with mean 0 and constant variance. We narrowed the possible predictors to the following 19 variables which may affect the type of students enrolled in a university:

- Public or Private school indicator
- Number of received applications
- Number of accepted applications
- Number of students newly enrolled
- Number of full-time undergraduates
- Number of part-time undergraduates
- Instate tuition fees
- Out-state tuition fees
- Room and board costs
- Additional fees
- Book costs
- Personal Spending
- Percentage of faculty with PhDs
- Percentage of faculty with Terminal Degrees
- Student to faculty ratio
- Percentage of alumni who donate
- Instructional expenditure per student
- Graduation rate
- Average salary of staff (All ranks)

Socio-economic model

As a means of checking that none of the socio-economic variables were significant, all of the variables were included in the linear model. The same assumptions about the error term were made.

Model fitting

University-specific model

A linear model was fit using the chosen variables and significance of the predictors was evaluated with R. Below is the initial model with 19 predictors. We used backwards selection to eliminate predictors that were statistically insignificant and confirmed our results with the step() function. We found that choosing a model based on lowest AIC (as opposed to significance of individual predictors) was more effective in producing consistent residuals and quantiles.

```
Call:
lm(formula = averagecombinesat ~ ., data = satall[4:23])

Residuals:
     Min       1Q   Median       3Q      Max
-219.943  -41.977   -0.339   42.280  258.844

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.951e+02  3.884e+01  15.323  < 2e-16 ***
public                    -3.414e+01  1.549e+01  -2.204 0.027910 *
appnumrec                  7.595e-03  2.272e-03   3.343 0.000878 ***
accepted                  -1.307e-02  4.031e-03  -3.243 0.001244 **
newenrolled               -1.507e-02  8.998e-03  -1.675 0.094368 .
fulltime                   6.567e-03  2.011e-03   3.266 0.001148 **
parttime                  -4.273e-03  1.911e-03  -2.236 0.025723 *
instate                    5.395e-03  2.844e-03   1.897 0.058310 .
outstate                   2.336e-03  2.816e-03   0.829 0.407148
roomnboard                -1.139e-02  3.472e-03  -3.281 0.001090 **
additional                -7.333e-03  6.279e-03  -1.168 0.243287
book                       5.381e-02  1.720e-02   3.128 0.001841 **
personal                   1.069e-03  4.589e-03   0.233 0.815833
phd                        1.340e+00  3.378e-01   3.967 8.08e-05 ***
terminal                   4.591e-01  3.544e-01   1.296 0.195598
studentratio              -6.600e-01  8.213e-01  -0.804 0.421929
alumnidonate               1.534e+00  2.915e-01   5.264 1.92e-07 ***
instructional              3.087e-03  8.376e-04   3.685 0.000248 ***
gradrate                   1.408e+00  2.035e-01   6.923 1.08e-11 ***
`Average salary-all ranks`  2.550e-01  5.258e-02   4.850 1.55e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.28 on 643 degrees of freedom
Multiple R-squared:  0.7048,    Adjusted R-squared:  0.696
F-statistic: 80.79 on 19 and 643 DF,  p-value: < 2.2e-16
```

The submodel is again verified using ANOVA, as shown in the output below. It can be seen that the p-value for the F statistic is greater than the alpha value of 0.05. Since there is insufficient evidence against the null hypothesis (i.e. the removed variables are all 0), we prefer the submodel (null hypothesis) over the alternative.

```
Analysis of Variance Table

Model 1: averagecombinesat ~ public + appnumrec + accepted + newenrolled +
    fulltime + parttime + instate + roomnboard + book + phd +
    alumnidonate + instructional + gradrate + `Average salary-all ranks`
Model 2: averagecombinesat ~ public + appnumrec + accepted + newenrolled +
    fulltime + parttime + instate + outstate + roomnboard + additional +
    book + personal + phd + terminal + studentratio + alumnidonate +
    instructional + gradrate + `Average salary-all ranks`
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1    648 2932113
2    643 2910589  5     21524 0.951 0.4473
```

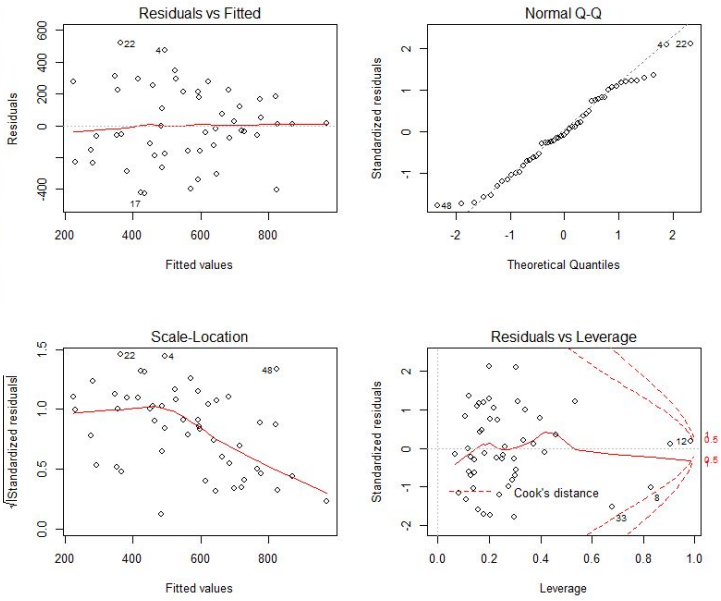
```
> summary(finalboth.lm)

Call:
lm(formula = averagecombinesat ~ public + appnumrec + accepted +
    newenrolled + fulltime + parttime + instate + roomnboard +
     book + phd + alumnidonate + instructional + gradrate + `Average
salary-all ranks`,
    data = satmodel.both[4:23])

Residuals:
     Min       1Q   Median       3Q      Max
-219.102  -41.092   -0.095   41.881  268.490

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.051e+02  2.915e+01  20.760  < 2e-16 ***
public                    -4.058e+01  1.272e+01  -3.192 0.001484 **
appnumrec                  7.180e-03  2.241e-03   3.204 0.001422 **
accepted                  -1.298e-02  3.906e-03  -3.322 0.000943 ***
newenrolled               -1.496e-02  8.961e-03  -1.670 0.095457 .
fulltime                   6.928e-03  1.982e-03   3.495 0.000507 ***
parttime                  -3.772e-03  1.888e-03  -1.998 0.046171 *
instate                    7.863e-03  1.338e-03   5.876 6.71e-09 ***
roomnboard                -1.119e-02  3.412e-03  -3.279 0.001098 **
book                       5.491e-02  1.682e-02   3.264 0.001158 **
phd                        1.673e+00  2.231e-01   7.498 2.14e-13 ***
alumnidonate               1.587e+00  2.870e-01   5.529 4.69e-08 ***
instructional              3.324e-03  7.683e-04   4.326 1.76e-05 ***
gradrate                   1.420e+00  2.019e-01   7.031 5.23e-12 ***
`Average salary-all ranks` 2.492e-01  5.166e-02   4.824 1.75e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.27 on 648 degrees of freedom
Multiple R-squared:  0.7026,    Adjusted R-squared:  0.6962
F-statistic: 109.3 on 14 and 648 DF,  p-value: < 2.2e-16
```

The new model has Multiple R-squared: 0.7026, Adjusted R-squared: 0.6962, which is a strong indication that the predictors are good at determining the average SAT in a university.

Socio-economic model

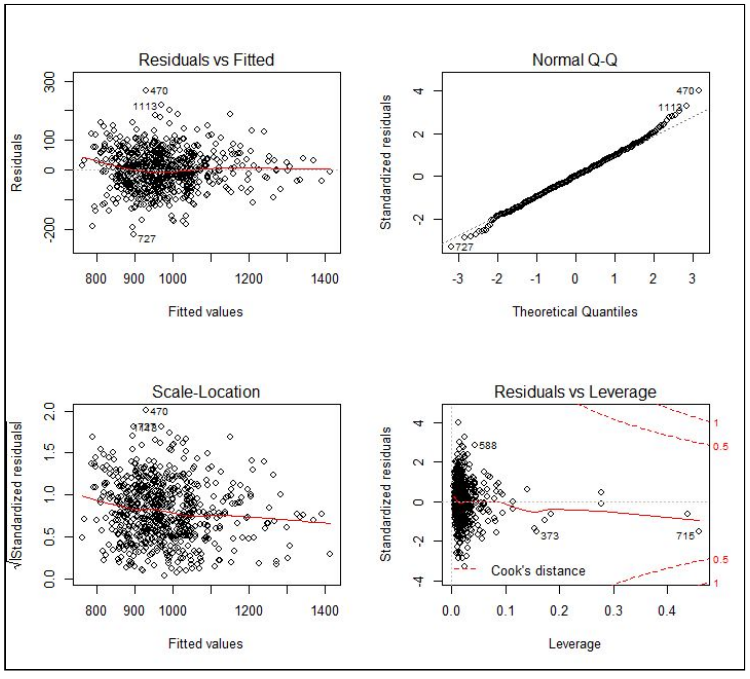
None of the predictors were statistically significant from 0 at a 5% level based on the t-test. This indicates that the socio-economic were not good predictors for the average SAT score in the state. The diagnostic plots also agree with this conclusion as the residuals do not conform to a normal distribution on the QQ plot and there seems to be a change in the variance of residuals.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.779e+04  8.530e+04   0.795    0.432
Crimeindex                    4.699e-02  4.746e-02   0.990    0.329
White.M                      -7.609e+04  8.854e+04  -0.859    0.396
White.F                      -6.056e+04  8.206e+04  -0.738    0.465
Black.M                      -6.558e+04  8.915e+04  -0.736    0.467
Black.F                      -7.063e+04  8.810e+04  -0.802    0.428
American.Indian.M             1.582e+05  1.517e+05   1.043    0.304
American.Indian.F            -2.901e+05  1.785e+05  -1.625    0.113
Asian.M                      -1.387e+05  1.733e+05  -0.800    0.429
Not.hispanic                 -5.835e+02  9.762e+02  -0.598    0.554
Percentage.people.in.povery   9.940e+00  3.375e+01   0.295    0.770
Median.household.income       3.123e-02  2.567e-02   1.217    0.231
Education.revenue.per.thousand 8.400e-01  7.860e-01   1.069    0.292
Education.spending.per.thousand -8.743e-02 7.951e-02  -1.100    0.279
```



Diagnostics

Diagnostic plots for overall SAT to university features model.



Examining the diagnostic plots for residuals, we can see that the residuals are relatively linear and are approximately normally distributed. As mentioned in the previous section, the R-Squared values for this model are Multiple R-Squared = 0.7092, and Adjusted R-squared = 0.6962.

Initially, the model had been based on a smaller set of observations. Although dataset that was used initially had 622 entries with SAT scores, we were limited to 400+ observations after omitting NA values for the 19 predictor variables. A number of imputation methods were used to include all 622 observations. For the model constructed in this report, the kNN method was used to impute missing values as it proved to produce the best R-squared and residuals.

Replacement with column mean: The resulting model had Adjusted R-squared: 0.6889 and 16 predictors.
Replacement with column median: The resulting model had Adjusted R-squared: 0.69 with 16 predictors.

Replacement with k-NN imputation: Based on weighted values of the columns where the observation is NA, it identifies k-nearest neighbour observations and computes the weighted average of the neighbours for the missing value. The resulting model has 14 predictors with Adjusted R-squared: 0.6962. The difference between this model and the original is that the original "Additional spending" predictor is replaced with "Number of full time students" predictor here.

Akaike Information Criterion (AIC)

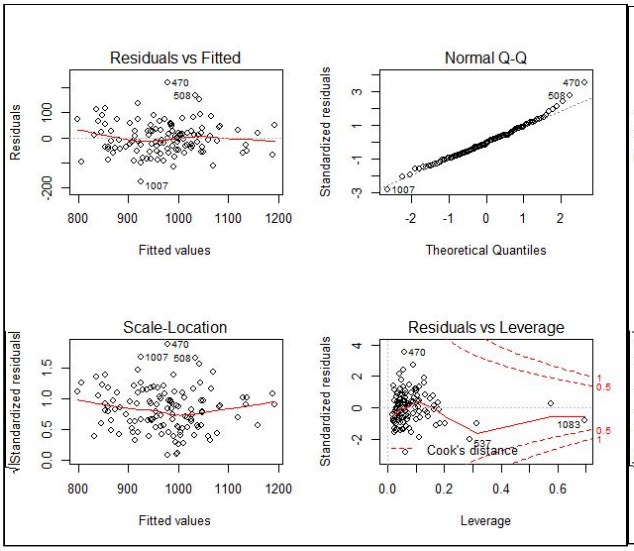
The AIC estimates the relative loss of information for a model, which relates to the quality of the model. The lower the AIC, the more preferable the model is. It rewards the model by goodness of it has a penalty related to the number of estimated parameters, which discourages overfitting. Our model was matched against R's default step() function which uses AIC as criterion, given the same initial 19 variables, and the resulting model was the same as our model.

```
Step:  AIC=5595.53
averagecombinesat ~ public + appnumrec + accepted + newenrolled +
    fulltime + parttime + instate + roomnboard + book + phd +
    alumnidonate + instructional + gradrate + `Average salary-all ranks`
```

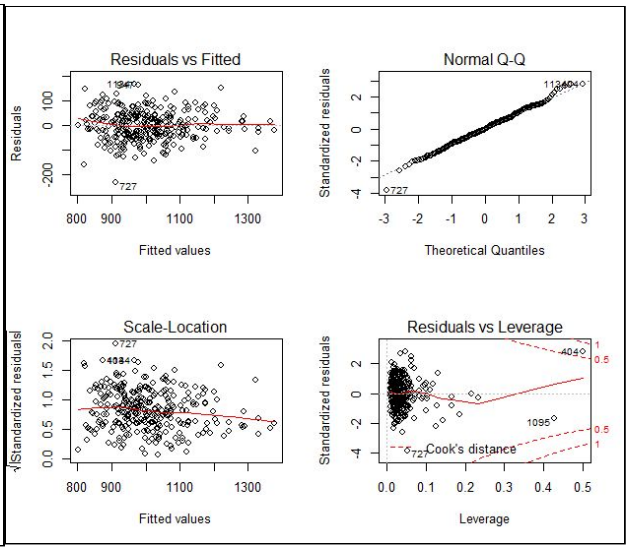
Normality of Residuals

Using two different tests which are Shapiro Wilk test and Anderson-Darling test to test the normality of residuals, both tests have the same null hypothesis, that is, the residuals belong to a normal distribution. The corresponding p-value are 0.1245 and 0.3912, so under 5% level, we cannot reject null hypothesis that residuals are normal distribution.

Public and Private Schools model



Diagnostic plots for public schools



Diagnostic plots for private schools

From the initial exploratory analysis and scatter plots, we inferred that there were different trends for public schools and private schools, and thus modelled them separately. Public schools did not fit well to the linear models we attempted to construct, with our final model only having an Adjusted R-Squared of 0.5924. The final private schools model however, had an Adjusted R-Squared of 0.7484 which explained the average combined SAT scores well.

Public schools model summary	Private schools model summary

Coefficients:

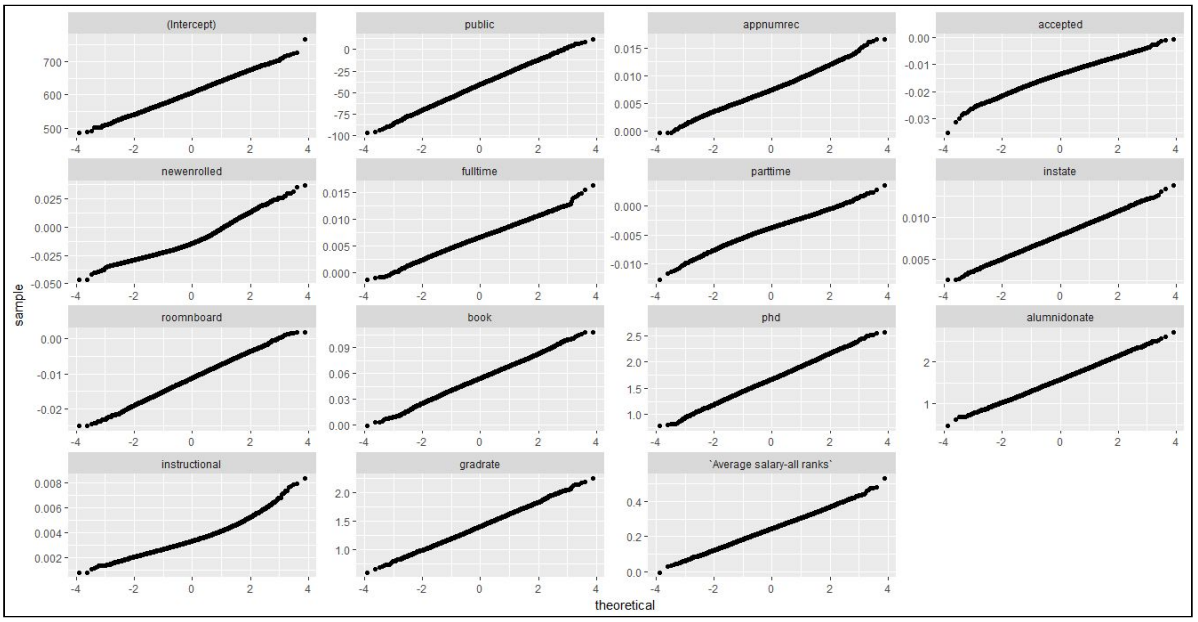
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	497.186911	36.529825	13.610	< 2e-16	***
appnumrec	0.009915	0.002861	3.466	0.000636	***
accepted	-0.016455	0.004869	-3.380	0.000859	***
newenrolled	-0.014381	0.009463	-1.520	0.130009	
fulltime	0.005201	0.002038	2.552	0.011385	*
book	0.067315	0.034629	1.944	0.053189	.
personal	0.011838	0.007977	1.484	0.139267	
phd	2.632831	0.434656	6.057	5.97e-09	***
alumnidonate	2.126158	0.607573	3.499	0.000565	***
instructional	0.002310	0.001557	1.484	0.139306	
gradrate	2.565705	0.345505	7.426	2.46e-12	***

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.700e+02	2.558e+01	22.280	< 2e-16	***
accepted	-2.212e-02	7.427e-03	-2.979	0.003060	**
newenrolled	9.041e-02	2.348e-02	3.850	0.000136	***
parttime	-1.014e-02	5.022e-03	-2.019	0.044097	*
instate	9.302e-03	1.496e-03	6.216	1.24e-09	***
roomnboard	-1.385e-02	3.848e-03	-3.599	0.000358	***
additional	-4.226e-02	1.440e-02	-2.934	0.003527	**
book	4.087e-02	1.862e-02	2.195	0.028719	*
phd	1.154e+00	2.499e-01	4.617	5.18e-06	***
studentratio	-1.320e+00	9.166e-01	-1.440	0.150625	
alumnidonate	1.419e+00	3.238e-01	4.382	1.49e-05	***
instructional	3.381e-03	8.653e-04	3.908	0.000109	***
gradrate	1.057e+00	2.305e-01	4.586	5.97e-06	***
`Average salary-all ranks`	3.623e-01	6.880e-02	5.266	2.23e-07	***

Bootstrap validation

To validate our model assumptions, we performed bootstrapping on the linear regression coefficients. In large part, the coefficients were similar, with the distribution for the model coefficients appearing normal. The 95% bootstrap confidence interval implies that at a 5% level, we cannot reject the null hypothesis that the coefficient is null.



Variable	Linear model estimate	Bootstrap estimate	Lower bound (bootstrap)	Upper bound (bootstrap)
(Intercept)	563.9839	565.9556	518.4918	614.5723
public2	-39.7074	-39.6552	-68.6475	-12.2154
appnumrec	0.0072	0.0076	0.0036	0.0119
accepted	-0.0130	-0.0137	-0.0211	-0.0072
newenrolled	-0.0151	-0.0126	-0.0287	0.0128
fulltime	0.0070	0.0067	0.0025	0.0107
parttime	-0.0038	-0.0039	-0.0077	-0.0008
instate	0.0078	0.0079	0.0052	0.0105
roomnboard	-0.0112	-0.0113	-0.0188	-0.0041
book	0.0548	0.0539	0.0250	0.0815

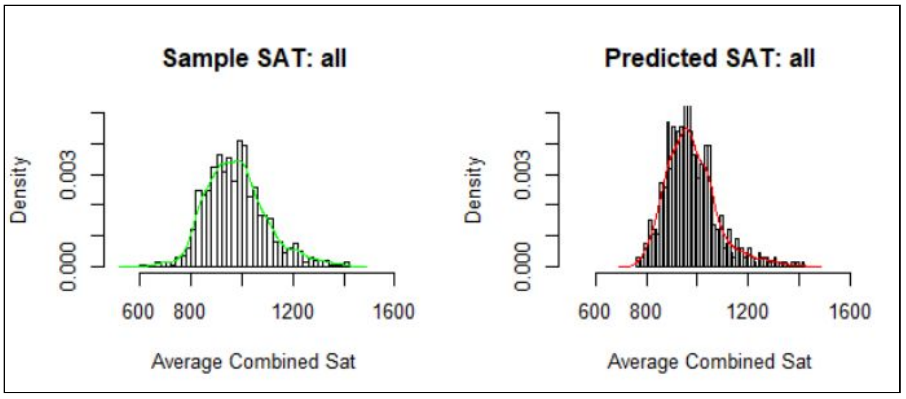
phd	1.6756	1.6799	1.2148	2.1622
alumnidonate	1.5939	1.5876	1.0379	2.1451
instructional	0.0033	0.0034	0.0020	0.0051
gradrate	1.4103	1.4003	0.9820	1.8182
`Average salary-all ranks`	0.2505	0.2469	0.1308	0.3655

Model assessment

The final overall model for SAT scores against university features is as follows:

```
lm(formula = averagecombinesat ~ public + appnumrec + accepted +
    newenrolled + fulltime + parttime + instate + roomnboard +
    book + phd + alumnidonate + instructional + gradrate + `Average salary-all ranks`,
    data = satmodel.both[4:23])
#
# Coefficients omitted
#
Residual standard error: 67.27 on 648 degrees of freedom
Multiple R-squared:  0.7026,    Adjusted R-squared:  0.6962
F-statistic: 109.3 on 14 and 648 DF,  p-value: < 2.2e-16
```

By comparing two histograms of overall SAT score model, the model we fitted is acceptable. From the graphs we see that the predicted values falls in the middle range more than the sample.



The usefulness of this model can be tested on the average SAT scores for 1996, since we do no expect the predictors to change in distribution drastically from year to year.

There are several limitations of model, which are listed as follows:

- **Lack of data**: Of the 1000+ observations in the dataset, there were only 622 entries with SAT scores, and as such we were limited to 400+ observations after omitting NA values for the 19 variables. While kNN imputation was used to replace the missing values, it may incorrectly represent the true values of that population and thus cause more error.
- **Heteroscedasticity:** By using Shapiro Wilk test and Anderson-Darling test, the model failed normality of residuals, due to the non-constant variance. However, both tests are sensitive to outliers and are influenced by sample size. For smaller samples, non-normality is less likely to be detected, for larger samples (i.e. more than one hundred), the normality tests are overly conservative and the assumption of normality might be rejected too easily. The Breusch–Pagan test for non-constant variance had a p-value of 0.0022145 implying that the non-constant variance assumption was indeed violated. To address this issue, a number of transformations can be performed to normalise variance.
  Box-cox transformation is a technique to transform non-normal dependent variables into a more normal shape, so that tests and confidence limits that require normality can be appropriately used. It has mathematical form $Y = (X + \delta)^{\lambda}$ where δ is a shift amount that is added when X is zero or negative and λ is exponent. Therefore, we can use box-cox

transformation to solve issue that residuals in our model have non-constant variance, and improve our model.

Conclusion

Our study has shown that the average SAT score of students enrolled in a university can be modelled by the following variables:

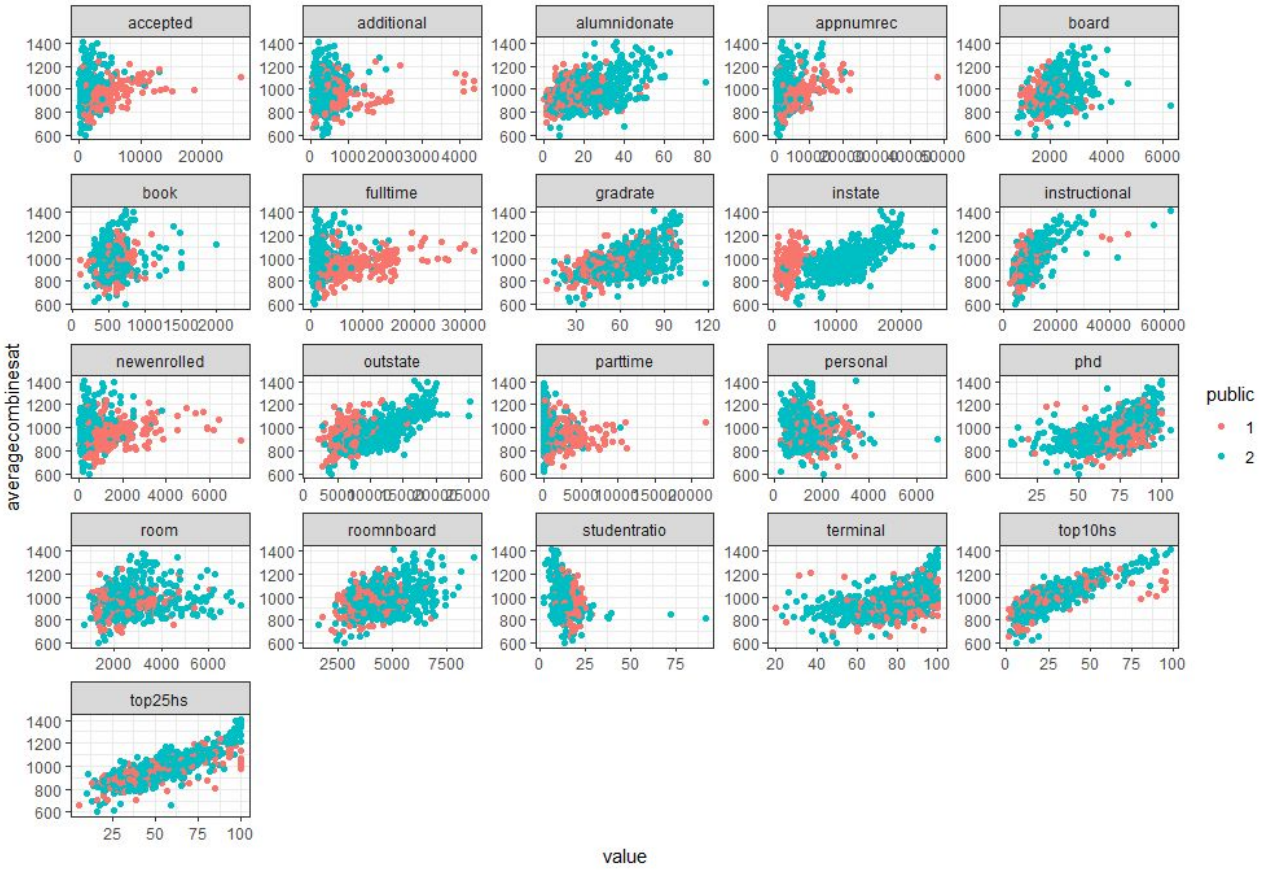
- Public or Private school indicator
- Number of received applications
- Number of accepted applications
- Number of students newly enrolled
- Number of full-time undergraduates
- Number of part-time undergraduates
- Instate tuition fees
- Room and board costs
- Book costs
- Percentage of faculty with PhDs
- Percentage of alumni who donate
- Instructional expenditure per student
- Graduation rate
- Average salary of staff (All ranks)

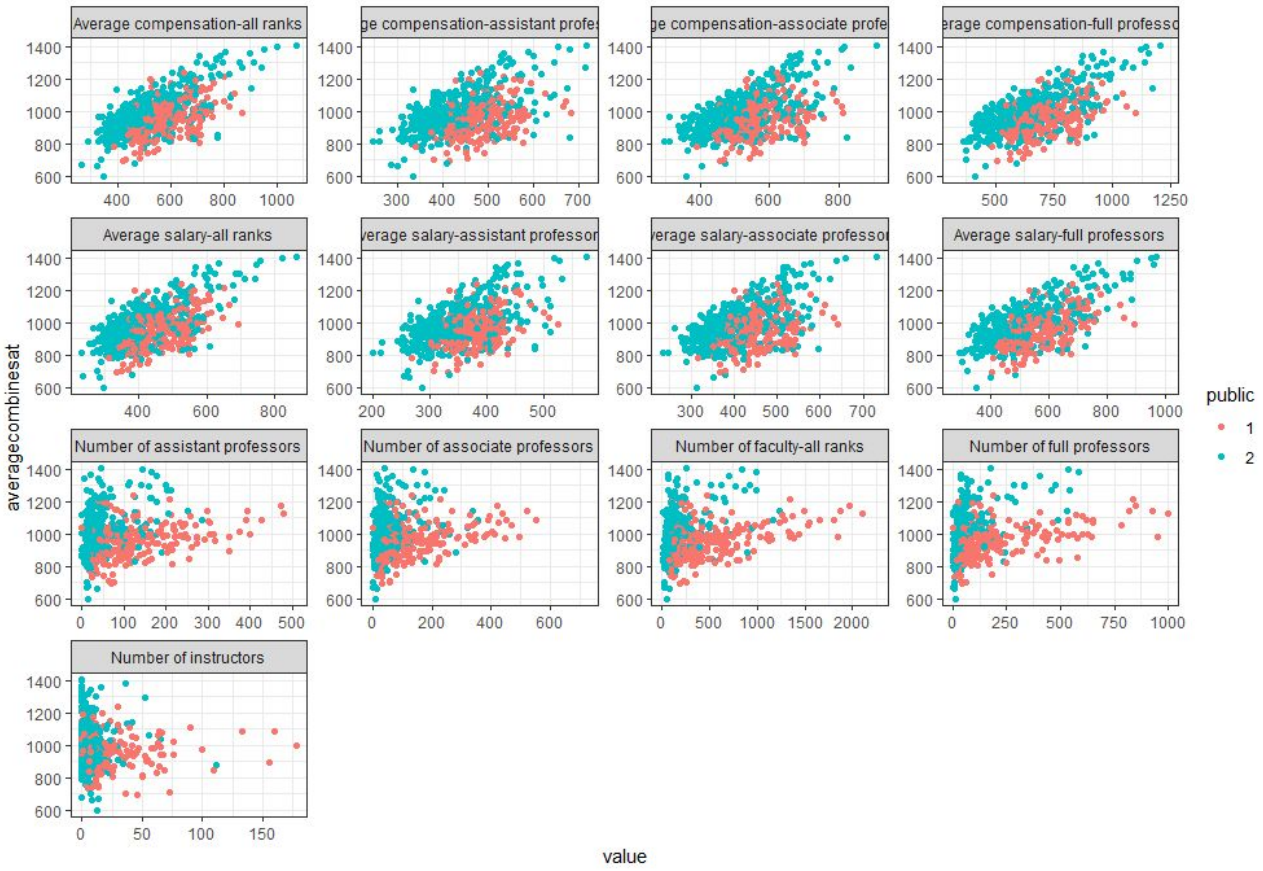
For this study, we considered the average SAT score of students as an indicator of the 'academic excellence' of students enrolled in that university, and to some extent the quality of a university. The percentage of faculty with PhD and graduation rates closely predicted the quality of students and could be interpreted as another proxy for the academic quality of a university. The average salary of staff and donating alumni also have relatively large influence, as they tend to show the financial capabilities of a university, and in turn provide better facilities. Public and private universities have different significant predictors for the SAT and thus we modelled them separately. Due to the lack of data for public universities (100+ observations), the model was ill-fitted compared to the private universities which had a lot more data.

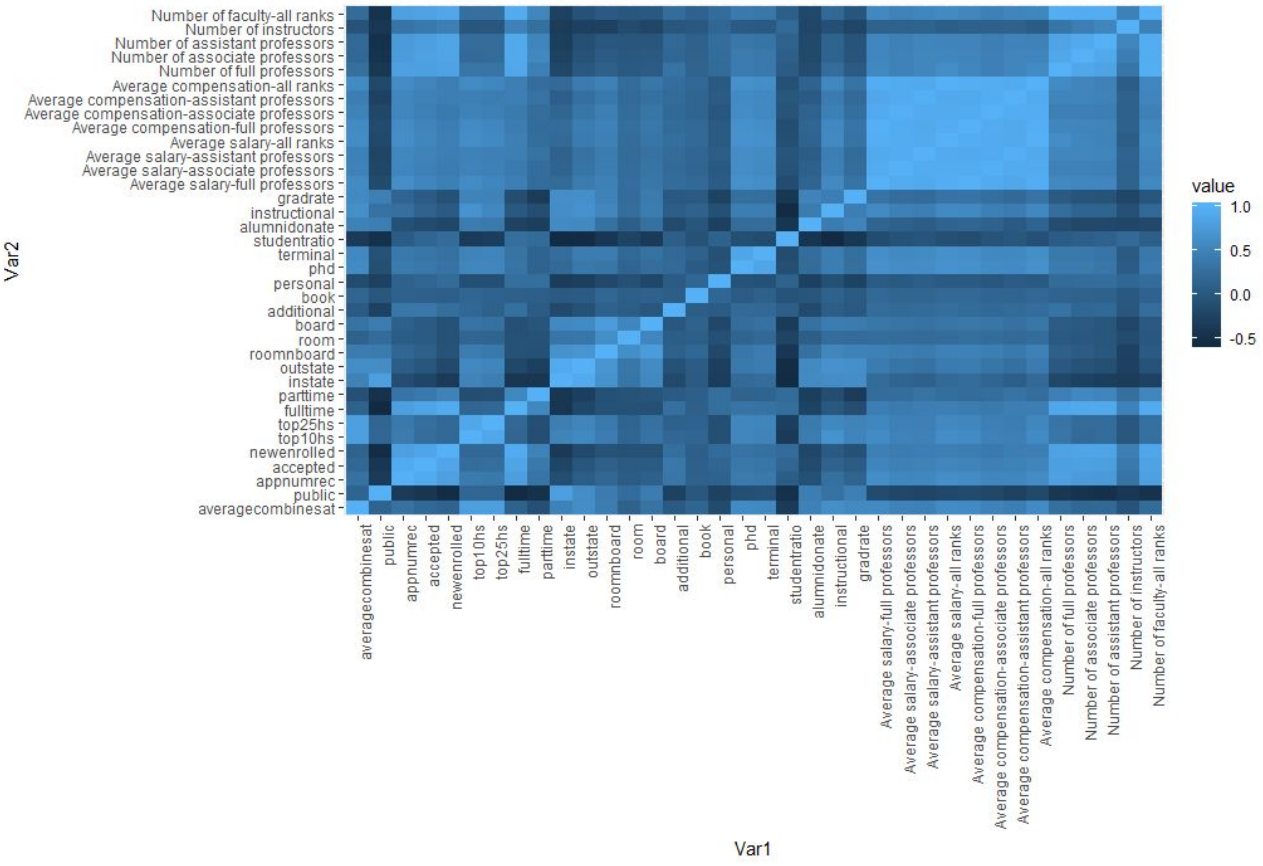
Socioeconomic factors appear to have little amount of influence on the universities. This may be due to the fact that the data was evaluated by states where the universities were located in, and a large proportion of students tend to come from out of state. The data would thus not accurately represent the environments that students were brought up in. There are also different social and financial levels in a state level, so the data is too generalized to directly infer from.

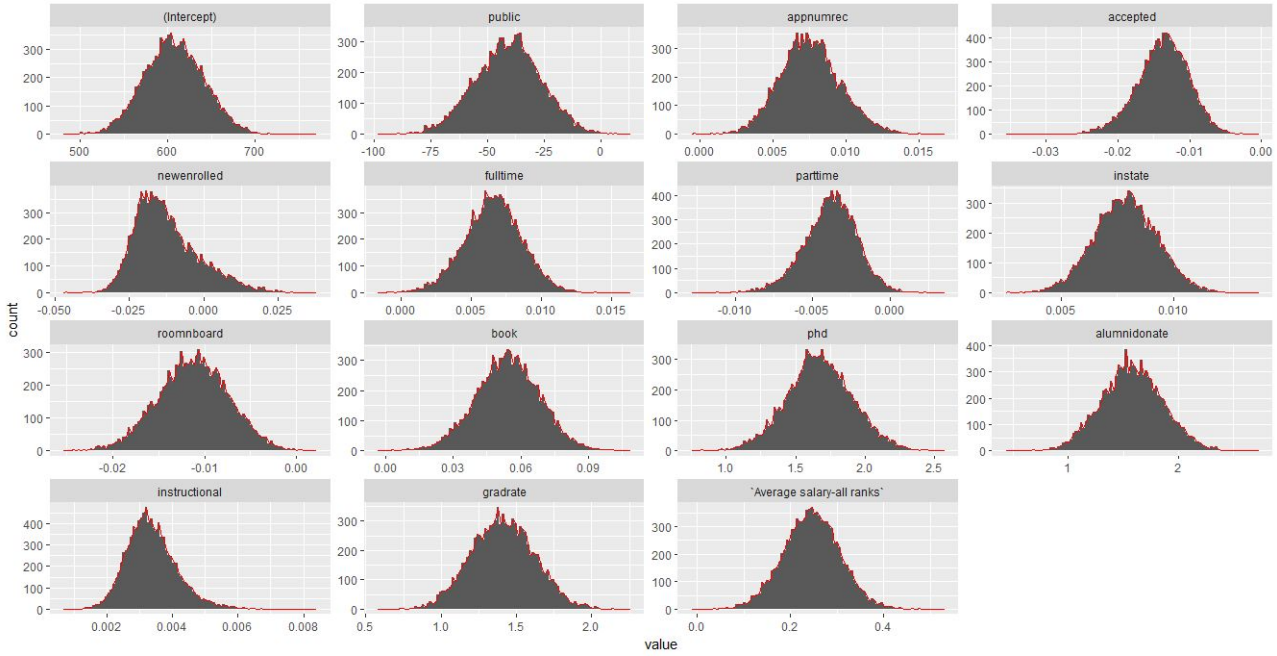
Appendix

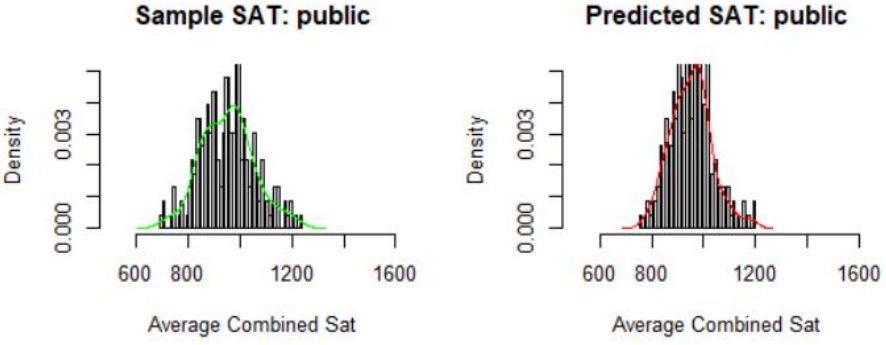
Appendix: Graphs



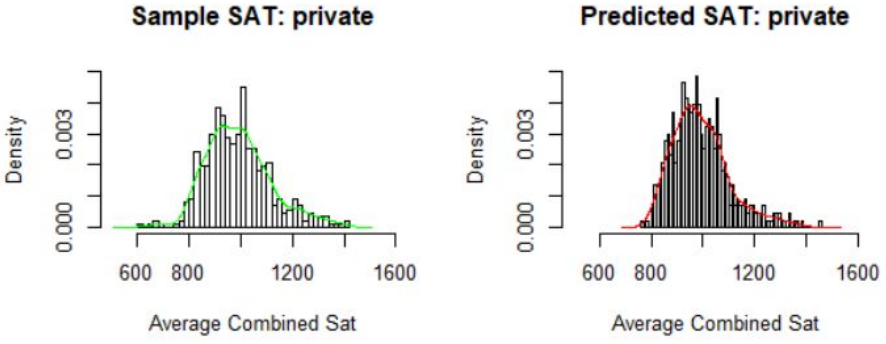

Average compensation-all ranks
ge compensation-assistant profes
ge compensation-associate profe
erage compensation-full professo
Average salary-all ranks
erage salary-assistant professor
erage salary-associate professor
Average salary-full professors
Number of assistant professors
Number of associate professors
Number of faculty-all ranks
Number of full professors
Number of instructors
averagecombinesat
value
public
1
2


Number of faculty-all ranks
Number of instructors
Number of assistant professors
Number of associate professors
Number of full professors
Average compensation-all ranks
Average compensation-assistant professors
Average compensation-associate professors
Average compensation-full professors
Average salary-all ranks
Average salary-assistant professors
Average salary-associate professors
Average salary-full professors
gradrate
instructional
alumnidonate
studentratio
terminal
phd
personal
book
additional
board
room
roomnboard
outstate
instate
parttime
fulltime
top25hs
top10hs
newenrolled
accepted
appnumrec
public
averagecombinesat
Var2
Var1
value
1.0
0.5
0.0
-0.5


(Intercept)
public
appnumrec
accepted
newenrolled
fulltime
parttime
instate
roomnboard
book
phd
alumnidonate
instructional
gradrate
`Average salary-all ranks`
count
value


Sample SAT: public
Density
Average Combined Sat

Predicted SAT: public
Density
Average Combined Sat


Sample SAT: private
Density
Average Combined Sat

Predicted SAT: private
Density
Average Combined Sat

Appendix: R Code

Links to download datasets:
https://1drv.ms/f/s!AvSEHNyiUFvFkYZSX81j6nTxi55Saw

```
############ Preliminary data loading and cleaning ############
library(reshape2)
library(ggplot2)
library(DMwR)
library(tidyr)
library(boot)
library(nortest)
library(SignifReg)
library(car)
#Load in datasets and replace asterisks with NAs
aaup<-read.csv("aaup.data",header=FALSE)
colnames.aaup<-c("FICE (Federal ID number)","College name","State (postal code)",
                 "Type  (I, IIA, or IIB)","Average salary-full professors",
                  "Average salary-associate professors","Average salary-assistant professors",
                 "Average salary-all ranks","Average compensation-full professors",
                 "Average compensation-associate professors","Average compensation-assistant professors",
                 "Average compensation-all ranks","Number of full professors",
                 "Number of associate professors","Number of assistant professors",
                 "Number of instructors","Number of faculty-all ranks")
length(colnames.aaup)
aaup<-data.frame(aaup)
colnames(aaup)<-colnames.aaup
aaup[ aaup == "*" ] <- NA
for (i in seq(5,17)){
        aaup[,i]<-as.numeric(as.character(aaup[,i]))
}


news<-read.csv("usnews.data",header=FALSE)
colnames.news<-c("FICE        (Federal        ID        number)","College        name","State        (postal
code)","public","avemathsat","aveverbalsat","averagecombinesat","aveact",

"mathq1sat","mathq3sat","verbalq1sat","verbalq3sat","actq1","actq3","appnumrec","accepted","newenrolled",

"top10hs","top25hs","fulltime","parttime","instate","outstate","roomnboard","room","board",

"additional","book","personal","phd","terminal","studentratio","alumnidonate","instructional",
                 "gradrate")
length(colnames.news)
news<-data.frame(news)
colnames(news)<-colnames.news
news[ news == "*" ] <- NA
for (i in seq(5,35)){
        news[,i]<-as.numeric(as.character(news[,i]))
}
news$public<-as.factor(news$public)
news$`State (postal code)`<-as.factor(news$`State (postal code)`)

newdata<-merge(news,aaup,by = c("FICE (Federal ID number)"))
```

```
satbystate<-read.csv("SAT-by-state-data.csv")
for (i in seq(2,17)){
        satbystate[,i]<-as.numeric(as.character(satbystate[,i]))
}
head(satbystate)

############ Exploratory analysis of AAUP and NEWS datasets ############

#Scatter plots of average combined SAT against other variables
news.keep.cols<-c("averagecombinesat","public",colnames(news)[15:35])
news[news.keep.cols] %>%
        gather(-averagecombinesat, -public,
        key = "var", value = "value") %>%
        ggplot(aes(x = value, y = averagecombinesat, color = public)) +
        geom_point() +
        facet_wrap(~ var, scales = "free") +
        theme_bw()

newdata.keep.cols<-c("averagecombinesat","public",colnames(aaup)[5:17])
newdata[newdata.keep.cols] %>%
        gather(-averagecombinesat, -public,
        key = "var", value = "value") %>%
        ggplot(aes(x = value, y = averagecombinesat, color = public)) +
        geom_point() +
        facet_wrap(~ var, scales = "free") +
        theme_bw()


#correlation matrix
newdata.keep.cols.news<-c("averagecombinesat","public",colnames(news)[15:35],colnames(aaup)[5:17])
numeric.mat.newdata<-newdata[newdata.keep.cols.news]
numeric.mat.newdata$public<-as.numeric(numeric.mat.newdata$public)
numeric.mat.newdata[is.na(numeric.mat.newdata)]<-NA
cormat<-cor(numeric.mat.newdata, use="pairwise.complete.obs")
melted_cormat <- melt(cormat)
head(melted_cormat)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
        geom_tile()+ theme(axis.text.x = element_text(angle = 90, hjust = 1))

#public vs private school SAT scores
ggplot(news, aes(x=public, y=averagecombinesat)) +
        geom_boxplot() +ggtitle("Boxplot of average combined SAT split by public and private colleges")+
        xlab("Funding (1 for public, 2 for private)") + ylab("Average combined SAT")

#Sat scores by State
ggplot(news, aes(x=`State (postal code)`, y=averagecombinesat)) +
        geom_boxplot()

#Sat scores normalised by state
temp<-data.frame(news,normalisedsat=news$averagecombinesat-mean(news$averagecombinesat,na.rm = T))
ggplot(temp, aes(x=State..postal.code., y=normalisedsat)) +
        geom_boxplot() +geom_hline(yintercept = 0)+ theme(axis.text.x = element_text(angle = 90, hjust =
1))+
        xlab("State") + ylab("Normalised average SAT")
```

```
############ SAT at a state level analysis ############
#replace missing crime rate for DC with the mean
satbystate[8,3]<-mean(satbystate$Crimeindex,na.rm=TRUE)

satbystate%>%
        gather(-Avgsat,
        key = "var", value = "value") %>%
        ggplot(aes(x = value, y = Avgsat)) +
        geom_point() +
        facet_wrap(~ var, scales = "free") +
        theme(strip.text.x = element_blank())

state.lm<-lm(Avgsat~.,data=satbystate[-c(1,11,13)])
summary(state.lm)
plot(state.lm)

############ SAT AAUP and NEWS analysis ############

#seperate the data into public and private
trainingset<-newdata[!is.na(newdata$averagecombinesat),]
trainingset<-trainingset[trainingset$averagecombinesat!="NA",]

trainingset<-trainingset[1:42]
trainingset<-trainingset[-c(5,6,8:14,18,19,25,26,36:41)]
trainingset <- knnImputation(trainingset)
publicdata<-trainingset[trainingset$public==1,]
privatedata<-trainingset[trainingset$public==2,]
par(mfrow=c(2,2))

# combined public & private model
satmodel.both<-trainingset
satboth.lm<-lm(averagecombinesat~.,data=satmodel.both[4:23])
finalboth.lm<-step(satboth.lm,direction="backward")
summary(finalboth.lm)
both.pred<-predict(finalboth.lm)
plot(finalboth.lm)

#install.packages("caret")
library(caret)
boxcoxSAT <- BoxCoxTrans(satmodel.both$averagecombinesat)
satmodel.BC <- cbind(satmodel.both, dist_new=predict(boxcoxSAT, satmodel.both$averagecombinesat))
satboth.BC.lm<-lm(dist_new~.,data=satmodel.BC[5:24])
finalboth.BC.lm<-step(satboth.BC.lm,direction="backward")
plot(finalboth.BC.lm)

# manual attempt
satall<-trainingset
a<-lm(averagecombinesat~.,data=satall[4:23])
summary(a)

b<-lm(averagecombinesat~.,data=satall[-c(1:3,9,11,16,18,19)])
anova(b,a)
summary(b)
plot(b)
hist(predict(b),xlim=c(500,1400),ylim=c(0,0.004),breaks=50,freq=FALSE)
lines(density(predict(b)),col="red")
hist(trainingset$averagecombinesat,xlim=c(500,1400),ylim=c(0,0.004),breaks=50,freq=FALSE)
```

```
lines(density(trainingset$averagecombinesat),col="green")


# public model
satmodel.public<-publicdata
satpub.lm<-lm(averagecombinesat~.,data=satmodel.public[5:23])
summary(satpub.lm)

finalpub.lm<-step(satpub.lm,direction="backward")
summary(finalpub.lm)
pub.pred<-predict(finalpub.lm)
plot(finalpub.lm)

# private model
satmodel.private<-privatedata
satpri.lm<-lm(averagecombinesat~.,data=satmodel.private[5:23])
finalpri.lm<-step(satpri.lm,direction="backward")
summary(finalpri.lm)
plot(finalpri.lm)
pri.pred<-predict(finalpri.lm)


# Compare histograms
hist(trainingset$averagecombinesat,xlim=c(500,1600),ylim=c(0,0.005),breaks=50,main="Sample SAT: all",xlab=
       "Average Combined Sat",freq=FALSE)
lines(density(trainingset$averagecombinesat),col="green")
hist(both.pred,xlim=c(500,1600),ylim=c(0,0.005),main="Predicted SAT: all",xlab=
       "Average Combined Sat",breaks=50,freq=FALSE)
lines(density(both.pred),col="red")

hist(trainingset$averagecombinesat[trainingset$public==1],xlim=c(500,1600),ylim=c(0,0.005),breaks=50,
     main="Sample SAT: public",xlab="Average Combined Sat",freq=FALSE)
lines(density(trainingset$averagecombinesat[trainingset$public==1]),col="green")
hist(pub.pred,xlim=c(500,1600),ylim=c(0,0.005),main="Predicted SAT: public",xlab=
       "Average Combined Sat", breaks=50,freq=FALSE)
lines(density(pub.pred),col="red")

hist(trainingset$averagecombinesat[trainingset$public==2],xlim=c(500,1600),ylim=c(0,0.005),breaks=50,
     main="Sample SAT: private",xlab="Average Combined Sat",freq=FALSE)
lines(density(trainingset$averagecombinesat[trainingset$public==2]),col="green")
hist(pri.pred,xlim=c(500,1600),ylim=c(0,0.005),main="Predicted SAT: private",xlab=
       "Average Combined Sat",breaks=50,freq=FALSE)
lines(density(pri.pred),col="red")


# Test the normality of residuals
res<-residuals(satmodel.both)
shapiro.test(res)
ad.test(res)

ncvTest(finalboth.lm)

############ Bootstrap ############

#define data with only the significant columns
get.coeffic=function(data,indices){
        data=data[indices,]
```

```
        lm.out=lm(averagecombinesat ~ public + appnumrec + accepted +
                newenrolled + fulltime + parttime + instate + roomnboard +
                book + phd + alumnidonate + instructional + gradrate + `Average salary-all
ranks`,data=data)
        return(lm.out$coefficients)
}

boot.out=boot(satmodel.both[4:23],get.coeffic,R=10000)

#get bootstrap estimates
coefficient_comparison<-data.frame(bootstrapest=apply(boot.out$t,2,mean),linearmodest=finalboth.lm$coeffic
ients)

#hypothesis testing using bootstrap --> look to see if the 95% confidence interval contains 0
lowerbound<-c()
upperbound<-c()
for(i in 1:length(finalboth.lm$coefficients)){
        lowerbound<-c(lowerbound,boot.ci(boot.out,index=i,type="perc",conf=0.95)[[4]][4])
        upperbound<-c(upperbound,boot.ci(boot.out,index=i,type="perc",conf=0.95)[[4]][5])
}

bootstrap.ci<-data.frame(lowerbound=lowerbound,upperbound=upperbound)
rownames(bootstrap.ci)<-rownames(coefficient_comparison)

bootstrap.data<-data.frame(boot.out$t)

colnames(bootstrap.data)<-rownames(coefficient_comparison)

d <- melt(bootstrap.data)

ggplot(d,aes(x = value)) + geom_histogram(bins = 100)+geom_freqpoly(bins=100,color = "red")+
        facet_wrap(~variable,scales = "free")

ggplot(d,aes(sample = value)) + stat_qq()+
        facet_wrap(~variable,scales = "free")
```