PROJECT REPORT

TOP MBA COLLEGES OF INDIA

AN ANALYTICAL STUDY

Ameesha Mittal

(BTech, first year, BITS Pilani)

MBA

Data Cleaning | Predictive Modelling | Data Clustering | 2015

GOANALYTICS

Table of Contents

Problem statement

Collect details of MBA courses in colleges in the state:

1. Ranking, Number of students, Fees, Number of programs, number of faculties, number of boys, number of girls, address, trimester or semester, placement status and other useful information

2. Find if available rankings are accurate or not

3. Create predictive models to suggest quality of program and value for money

4. Create clusters of colleges and explain differences between them.

Note:

Mention source of data. If there are multiple sources, or you have collected data by primary survey,
mention that.

Provide both unclean and clean data

Interpretation of problem statement

The problem statement requires us to collect data for top management colleges in India along with their rankings and then carry out data analysis to check the accuracy of given data. This can be done by various analytical models. Thus we need to express our understanding of data as well as data analytical modelling by applying the best fit model for the analysis.

Next, having collected the data for various parameters a new predictive model has to be set up to introduce new parameters of 'quality' of programs and 'value for money (vfm)' and further evaluate the same. Since, there is no universally accepted model to do so, we need to intellectually create our own points for quality of the colleges and then test or observe our present points with this model.

Lastly, we need to perform data clustering by means of various data clustering techniques and compare and explain the results obtained by each.

There is no set solution for the given problem statement. Thus this provides us with immense flexibility and an opportunity to apply our own ideas to data analysis. Thus, through this project we present our complete understanding and implementation of data analysis techniques under the following heads:

1) Data collection
2) Data cleaning
3) Regression analysis
4) Predictive modelling
5) Data clustering

Approach adopted towards solving this problem statement

6) As mentioned earlier there is no well defined approach towards data interpretation and analysis. Thus various methods were used by hit and trial to manipulate with obtained variables and understand their importance to the obtained data frame.

Data extraction:
Various sources had been searched and studied to shortlist two of them which have been finally used to extract data.

Data Cleaning:
Initial data set had been created in excel manually and then read in R and thus contained several null spaces and outliers. Various approaches were applied to impute missing values in the data table. Finally a missing value in fees was filled by the mean of the column while those in 'CAT_Cutoff" were evaluated through package 'MICE'

Regression Analysis:
Suitable parameters were chosen for multiple regression analysis of given rankings and the obtained p value was used to represent the accuracy of given ranking

Predictive Analysis:
Various predictive models were applied to predict the quality of programs and value for money. Out of those most suited was adopted and same has been given here in the project report

Clustering:
Various clustering techniques are available to forn data clusters. However not all can be applied to all kinds of data.
All the techniques were read about and two apparently suitable ones were then applied to the obtained data set.

Data Extraction

Data was extracted mainly from the following sources:

http://bschools.businesstoday.in/
http://www.shiksha.com/top-mba-colleges-in-india-rankingpage-2-2-0-0-0

Uncleaned dataset was created in excel by taking values from the above sources:

Uncleaned dataset looked like:

College Name	Business Today 2014	CAT Cutoff	Average Salary (Annual)(in lacs)	Total Course Fee (INR) (in lacs)	no. of MDP	faculty(permanent)	Campus area	Students	boys	girls	No. of companies	Placement %	international faculty	students exchange	guest speakers
Indian Institute of Management, IIM A	1	96.36,450	18.81	18.5	6	92	104	380	298	82	125	100	9	27	68
Indian Institute of Management,Calcutta	2	99.69	17.13	16.2	158	90	135	458	349	109	159	99	2	116	487
Xavier Labour Relations Institute	3		16.25	16.8	6	69	40	300	229	71	90	100	2	19	34
Faculty of Management Studies (FMS)	4	98.56	16.18	0.67	130	31	2.5	220	165	55	91	94.5	0	6	67
S.P. Jain Institute of Management and Research	5	92	17.2	14	75	49	45	233	140	93	79	100	14	20	113
Management Development Institute, Gurgaon	6	95.7	14.7	14.9	143	74	37	24	201	39	153	100	13	78	90
Indian Institute of Management, Kozhikode	7	85	12.31	13	104	63	99	357	232	125	146	100	0	21	24
Indian Institute of Management, Indore	8	87.35	12.13	13	50	64	193	459	319	140	152	100	0	80	7
International Management Institute, Delhi	9	92	9.92	13.5	61	64	5.5	180	104	76	57	100	12	22	172
Indian Institute of Foreign Trade	10		13.06	12.8	23	46	6.5	222	189	33	71	100	0	28	109
NMIMS, mumabi	11		15.3		86	30	14.69	321	249	72	124	88.7	4	12	116
National Institute of Industrial Engineering	12	97	11.32	8.1	32	55	64	224	194	30	117	100	0	0	54
Xavier Institute of Management, Bhubaneswar	13	94	12.08	15	52	56	20	240	175	65	52	97	0	4	64
Symbiosis Institute of Business Management	14		11.77	10.5	39	20	300	213	149	64	93	92.5	7	4	82
Department of Management Studies, IITD	15	98.28	12.52	4	1	14	325	64	54	10	27	96	0	0	35
Jamnalal Bajaj Institute of Management Studies	16	90	16.18	2	3	5	1.2	119	105	14	89	100	0	0	37
Vinod Gupta School of Management (IIT Kharagpur)	17	97	13.5	6.2	21	23	15	58	51	7	34	97	3	0	41
T. A. Pai Management Institute	18	80	8.5	13.5	17	45	42	384	261	123	79	100	1	3	82
Shailesh J. Mehta School of Management	19	98.67	14.26	8.2	16	23	565	79	71	8	68	98.2	2	0	70
Birla Institute of Management Technology	20	80	6.8	10	90	63	10	216	135	81	75	94.7	0	15	38

Data for top 20 MBA colleges was collected. Parameters read were:

1) Name
2) Rank
3) CAT-Cutoff
4) Average salary offered to students
5) Total course fees
6) Number of management development programs
7) Number of permanent faculty
8) Campus area
9) Number of total students in a batch
10) Number of girls
11) Number of boys
12) Placement percentage
13) Number of companies visiting the campus
14) Number of international faculty
15) Number of guest speakers
16) Number of student exchange programs

However, there were several null values in our data which could have resulted in ambiguity in our results while analysis. Thus the next step was cleaning the data and filling all these null spaces.

Data Cleaning:

Data cleaning, or data preparation is an essential part of statistical analysis

For data cleaning first the number of NAs needed to be noted.

```
> a<-subset(m,select=c("Ranking","CAT_Cutoff","AvgSal","companies","Placement","Fees"))
> summary(a)
```

This displayed 5 NAs in CAT_Cutoff and 1 in fees. The nullity in fees was filled by average fees:

```
>mba$Fees[is.na(mba$Fees)]<-round(mean(mba$Fees,na.rm=TRUE))
```

To evaluate suitable values for null spaces in CAT_Cutoff various techniques were applied using VIM, mvnmle etc.

Finally package MICE was used.

The R package mice imputes incomplete multivariate data by chained equations.

Multiple imputation using Fully Conditional Specification (FCS) implemented by the MICE algorithm. Each variable has its own imputation model. Built-in imputation models are provided for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds). MICE can also impute continuous two-level data (normal model, pan, second-level variables). Passive imputation can be used to maintain consistency between variables. Various diagnostic plots are available to inspect the quality of the imputations.

```
a<-subset(mba,select=c("Ranking","CAT_Cutoff","AvgSal","companies","Placement"))
summary(a)
cov(a)
install.packages("mice")
library(mice)
imputed_data = complete( mice( a ))
mba$CAT_Cutoff=imputed_data$CAT_Cutoff
View(mba)
```

Further, since the number of students or number of boys or girls cannot serve as good evaluating parameters, a new field of sexratio was added.

```
mba$sexratio<-mba$boys/mba$girls
```

Final Cleaned data:

Name	Ranking	CAT_Cutoff	AvgSal	Fees	MDPs	faculty	Campus_.area	Students	boys	girls	companies	Placement	internationalfaculty	studentexchange	guestspeakers	sexratio
Indian Institute of Management, IIM A	1	90	18.81	18.5	6	92	104	380	298	82	125	100	9	27	68	3.634146
Indian Institute of Management,Calcutta	2	99.69	17.13	16.2	158	90	135	458	349	109	159	99	2	116	487	3.201835
Xavier Labour Relations Institute	3	97	16.25	16.8	6	69	40	300	229	71	90	100	2	19	34	3.225352
Faculty of Management Studies (FMS)	4	98.56	16.18	0.67	130	31	2.5	220	165	55	91	94.5	0	6	67	3
S.P. Jain Institute of Management and Research	5	92	17.2	14	75	49	45	233	140	93	79	100	14	20	113	1.505376
Management Development Institute, Gurgaon	6	95.7	14.7	14.9	143	74	37	24	201	39	153	100	13	78	90	5.153846
Indian Institute of Management, Kozhikode	7	85	12.31	13	104	63	99	357	232	125	146	100	0	21	24	1.856
Indian Institute of Management, Indore	8	87.35	12.13	13	50	64	193	459	319	140	152	100	0	80	7	2.278571
International Management Institute, Delhi	9	92	9.92	13.5	61	64	5.5	180	104	76	57	100	12	22	172	1.368421
Indian Institute of Foreign Trade	10	92	13.06	12.8	23	46	6.5	222	189	33	71	100	0	28	109	5.727273
NMIMS, mumabi	11	97	15.3	11	86	30	14.69	321	249	72	124	88.7	4	12	116	3.458333
National Institute of Industrial Engineering	12	97	11.32	8.1	32	55	64	224	194	30	117	100	0	0	54	6.466667
Xavier Institute of Management, Bhubaneswar	13	94	12.08	15	52	56	20	240	175	65	52	97	0	4	64	2.692308
Symbiosis Institute of Business Management	14	90	11.77	10.5	39	20	300	213	149	64	93	92.5	7	4	82	2.328125
Department of Management Studies, IITD	15	98.28	12.52	4	1	14	325	64	54	10	27	96	0	0	35	5.4
Jamnalal Bajaj Institute of Management Studies	16	90	16.18	2	3	5	1.2	119	105	14	89	100	0	0	37	7.5
Vinod Gupta School of Management (IIT Kharagpur)	17	97	13.5	6.2	21	23	15	58	51	7	34	97	3	0	41	7.285714
T. A. Pai Management Institute	18	80	8.5	13.5	17	45	42	384	261	123	79	100	1	3	82	2.121951
Shailesh J. Mehta School of Management	19	98.67	14.26	8.2	16	23	565	79	71	8	68	98.2	2	0	70	8.875
Birla Institute of Management Technology	20	80	6.8	10	90	63	10	216	135	81	75	94.7	0	15	38	1.666667

After cleaning of the data following parameters were obtained for each college(data as per 2013-14):

Ranking
Cat cutoff
Avg Salary of the students who got placed(in lacs)
Total course fees(in lacs)
Campus Area(in acre)
No. of permanent faculty
No. of MDPs(management development programs)
Total no. of students in a batch
No. of boys
No. of girls
Sex ratio
No. of international faculty
No. of guest speakers
No. of student exchange programs
No. of companies who offered placement
Placement percentage

Regression Analysis

Out of the obtained parameters following were decided to be used for regression analysis:

- Ranking: our dependent variable
- Cat cutoff: since most of these colleges take admissions through CAT, CAT Cutoff seems o be a relevant parameters to judge the college's reputation in the society
- Avg Salary: An important parameter since final aim of the student taking admission is to earn a decent package at the end to the course
- Fees: Inverse relation to the ranking. Low fees (with good facilities) makes a room for more people belonging to financially weaker section of the society
- No. of permanent faculty: More the faculty, more the guidance and diversity.
- No. of MDPs: Good colleges shall provide more number of courses
- Sex ratio: Since the admission procedure for all colleges is same sex ratio will help us determine which colleges are preferred for girls i.e. whch college maintains the best co-ed atmosphere
- No. of international faculty and no. of guest speakers: These parameters project the reputation of the colleges amongst other scholars of the world.
- No. of student exchange programs: Shows the amount of exposure given to students
- No. of companies: More companies shall visit the college that produces better qualified students
- Placement percentage: Students take admission with a purpose of getting placed in a good company at the end of the course. Thus this parameters shows how much that purpose is fulfilled.

Note: number of students is not used because more number of students allowed in a batch does not imply better teaching facilities. Good college wont be the one who takes more admissions. It will be the one who train those it takes well.

Exploring relationships among different parameters:

Before fitting a regression model o data, pairwise correlation has been calculated.

```
>cor(mba[c("Ranking","CAT_Cutoff","AvgSal","Fees","MDPs","faculty","companies","Placement","internationalfaculty","studentexchange","guestspeakers","sexratio")])
```

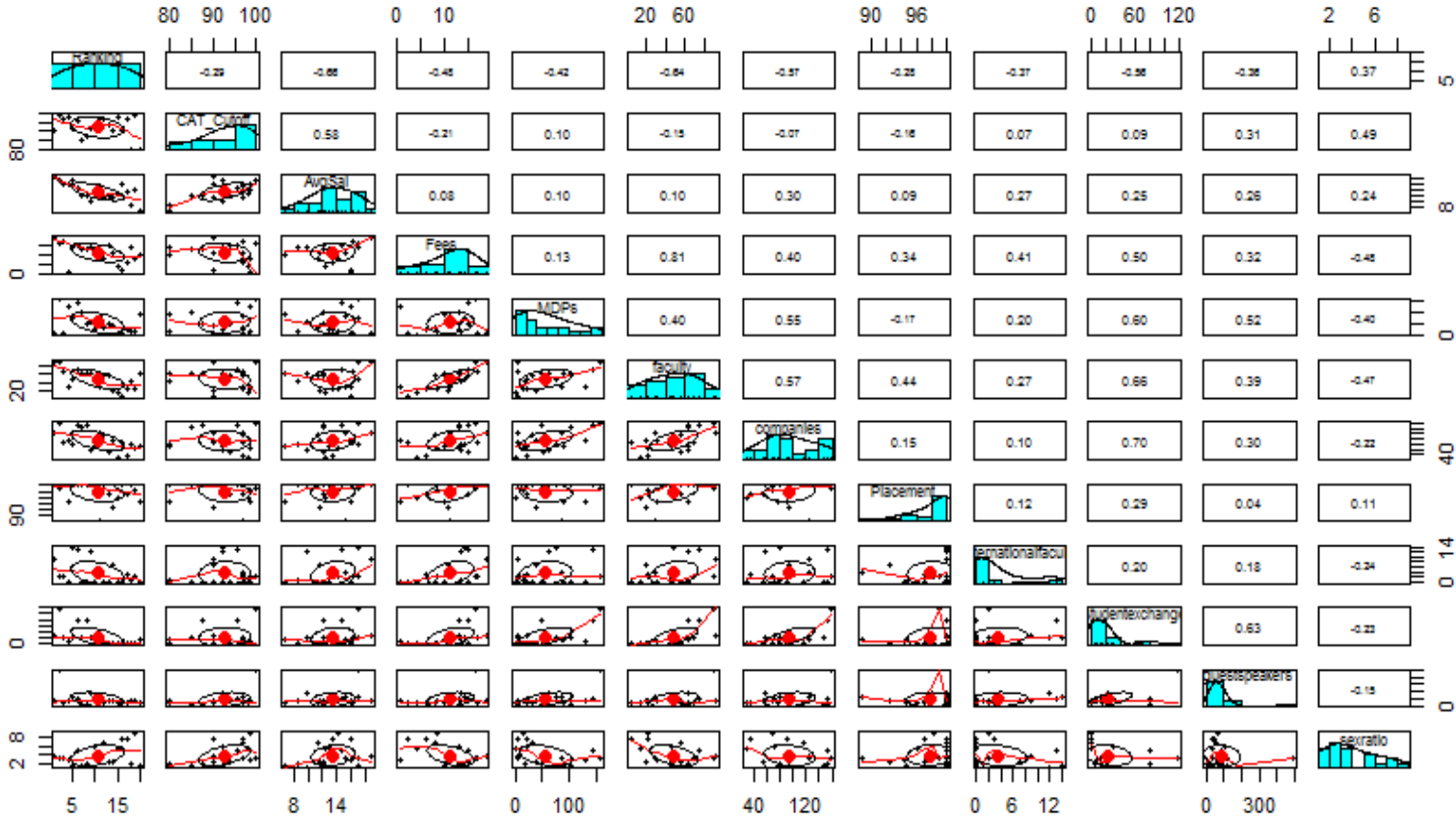
	Ranking	CAT_Cutoff	AvgSal	Fees	MDPs	faculty	companies
Ranking	1.0000000	-0.29491902	-0.66126599	-0.48084544	-0.41950832	-0.64276522	-0.57223475
CAT_Cutoff	-0.2949190	1.00000000	0.57948895	-0.20941749	0.10395069	-0.14763150	-0.07010402
AvgSal	-0.6612660	0.57948895	1.00000000	0.08452388	0.09751273	0.09740114	0.30389652
Fees	-0.4808454	-0.20941749	0.08452388	1.00000000	0.13329264	0.81122302	0.40066728
MDPs	-0.4195083	0.10395069	0.09751273	0.13329264	1.00000000	0.40279683	0.55227342
faculty	-0.6427652	-0.14763150	0.09740114	0.81122302	0.40279683	1.00000000	0.56708777
companies	-0.5722348	-0.07010402	0.30389652	0.40066728	0.55227342	0.56708777	1.00000000
Placement	-0.2760863	-0.15676765	0.08883471	0.33868225	-0.17222276	0.43903160	0.14501607
internationalfaculty	-0.3711043	0.07203207	0.27413184	0.41206188	0.20437766	0.27310245	0.10472509
studentexchange	-0.5621617	0.09207665	0.25320357	0.50186974	0.59858986	0.66113735	0.70140770
guestspeakers	-0.3631085	0.31000093	0.26159592	0.31905406	0.52142883	0.38641952	0.30270695
sexratio	0.3657731	0.48889319	0.23799915	-0.47590952	-0.39729113	-0.46914781	-0.22062137

	Placement	internationalfaculty	studentexchange	guestspeakers	sexratio
Ranking	-0.27608626	-0.37110434	-0.56216172	-0.36310849	0.3657731
CAT_Cutoff	-0.15676765	0.07203207	0.09207665	0.31000093	0.4888932
AvgSal	0.08883471	0.27413184	0.25320357	0.26159592	0.2379991
Fees	0.33868225	0.41206188	0.50186974	0.31905406	-0.4759095
MDPs	-0.17222276	0.20437766	0.59858986	0.52142883	-0.3972911
faculty	0.43903160	0.27310245	0.66113735	0.38641952	-0.4691478
companies	0.14501607	0.10472509	0.70140770	0.30270695	-0.2206214
Placement	1.00000000	0.11746756	0.29167103	0.03793053	0.1061551
internationalfaculty	0.11746756	1.00000000	0.19969888	0.18144371	-0.2424517
studentexchange	0.29167103	0.19969888	1.00000000	0.63114485	-0.2251577
guestspeakers	0.03793053	0.18144371	0.63114485	1.00000000	-0.1478802
sexratio	0.10615515	-0.24245166	-0.22515772	-0.14788022	1.0000000

We see ranking having strong correlation with all the other parameters. All other parameters too show certain degree of correlation which was quite expected since a college with a better name and ranking will have better facilities as well.

Next we try to visualize these correlations:

```
> library(psych)
>
pairs.panels(mba[c("Ranking","CAT_Cutoff","AvgSal","Fees","MDPs","faculty","companies","Placement","internationalfaculty","studentexchange","guestspeakers","sexratio")])
```



Observations:

Rankings hold strong correlation with other parameters as seen by better ellipses in the plot. The distribution of indivisual parameters however seem a bit irregular.

```
> analysis<-
lm(Ranking~CAT_Cutoff+AvgSal+Fees+MDPs+faculty+companies+Placement+internationalfaculty+studentexchange+guestspeakers+sexratio,data=mba)
> analysis
```

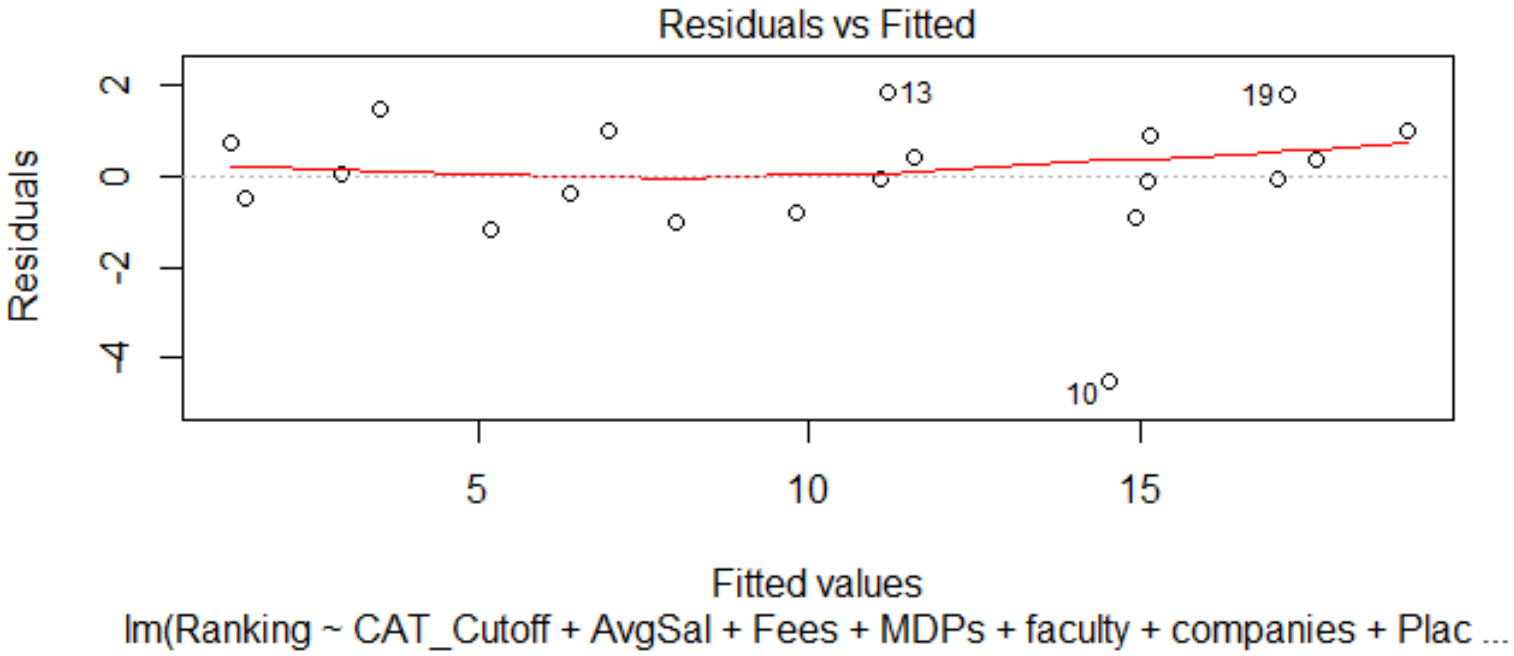
```
Call:
lm(formula = Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty +
    companies + Placement + internationalfaculty + studentexchange +
    guestspeakers + sexratio, data = mba)

Coefficients:
         (Intercept)            CAT_Cutoff                AvgSal                  Fees
           102.284779            -0.424098            -0.935884              0.227985
                 MDPs               faculty             companies             Placement
             0.008785            -0.083590            -0.034490            -0.435141
internationalfaculty       studentexchange       guestspeakers              sexratio
            -0.024947            -0.000334             0.006331              1.657229
```

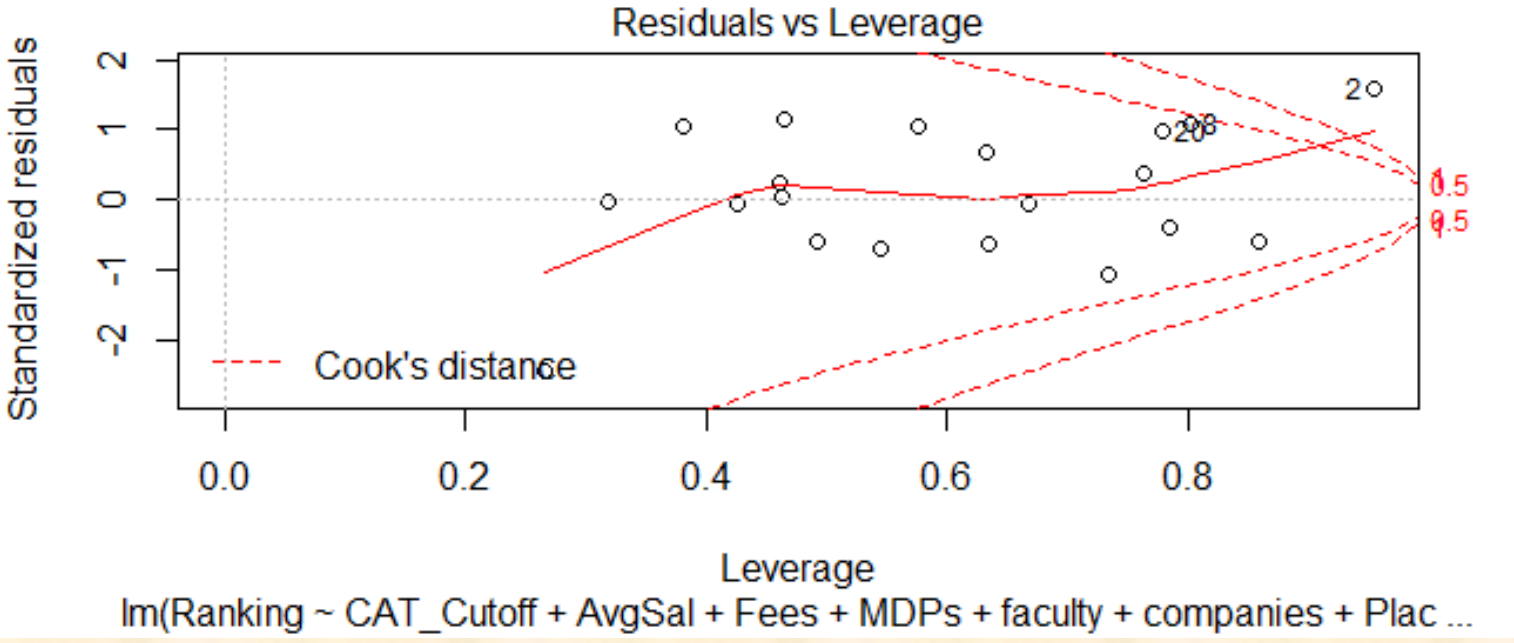
We observe lower ranks are associated with better CAT cutoffs, better salaries, lower fees, more faculty, more companies, better placements, more student exchange programs. However certain factors seem to have really less and opposite effect on ranking like MDPs and guest speakers. We also see better colleges tend to have lower sex ratio that is more girls as compared to other colleges.

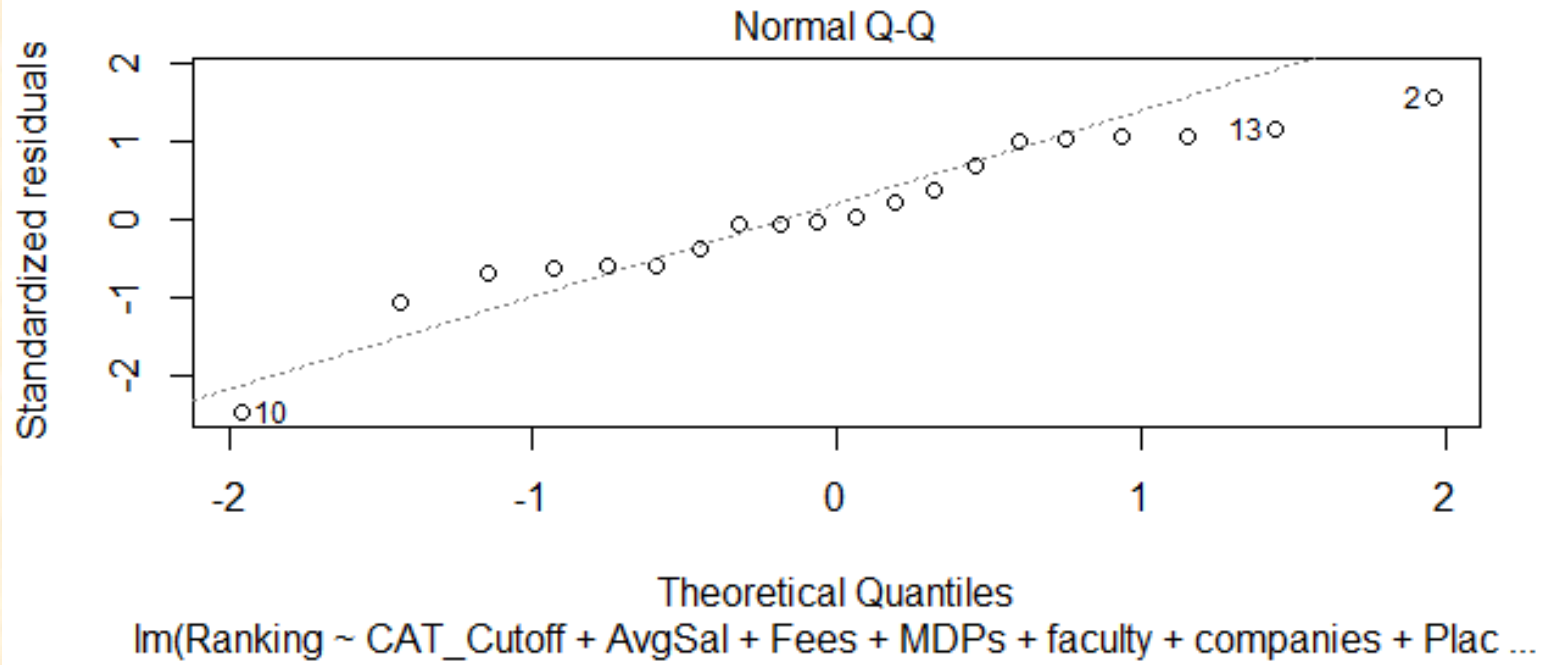
```
> summary(analysis)

Call:
lm(formula = Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty +
    companies + Placement + internationalfaculty + studentexchange +
    guestspeakers + sexratio, data = mba)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5296 -0.5693 -0.0031  0.9117  1.8268

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          102.284779  28.246916   3.621  0.00677 **
CAT_Cutoff            -0.424098   0.146229  -2.900  0.01988 *
AvgSal                -0.935884   0.243736  -3.840  0.00495 **
Fees                   0.227985   0.231389   0.985  0.35334
MDPs                   0.008785   0.019899   0.441  0.67056
faculty               -0.083590   0.049859  -1.677  0.13216
companies             -0.034490   0.023108  -1.493  0.17390
Placement             -0.435141   0.229903  -1.893  0.09503 .
internationalfaculty  -0.024947   0.130020  -0.192  0.85263
studentexchange       -0.000334   0.031888  -0.010  0.99190
guestspeakers          0.006331   0.007391   0.857  0.41653
sexratio               1.657229   0.374181   4.429  0.00220 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.155 on 8 degrees of freedom
Multiple R-squared:  0.9441, Adjusted R-squared:  0.8674
F-statistic: 12.29 on 11 and 8 DF,  p-value: 0.0007528
```

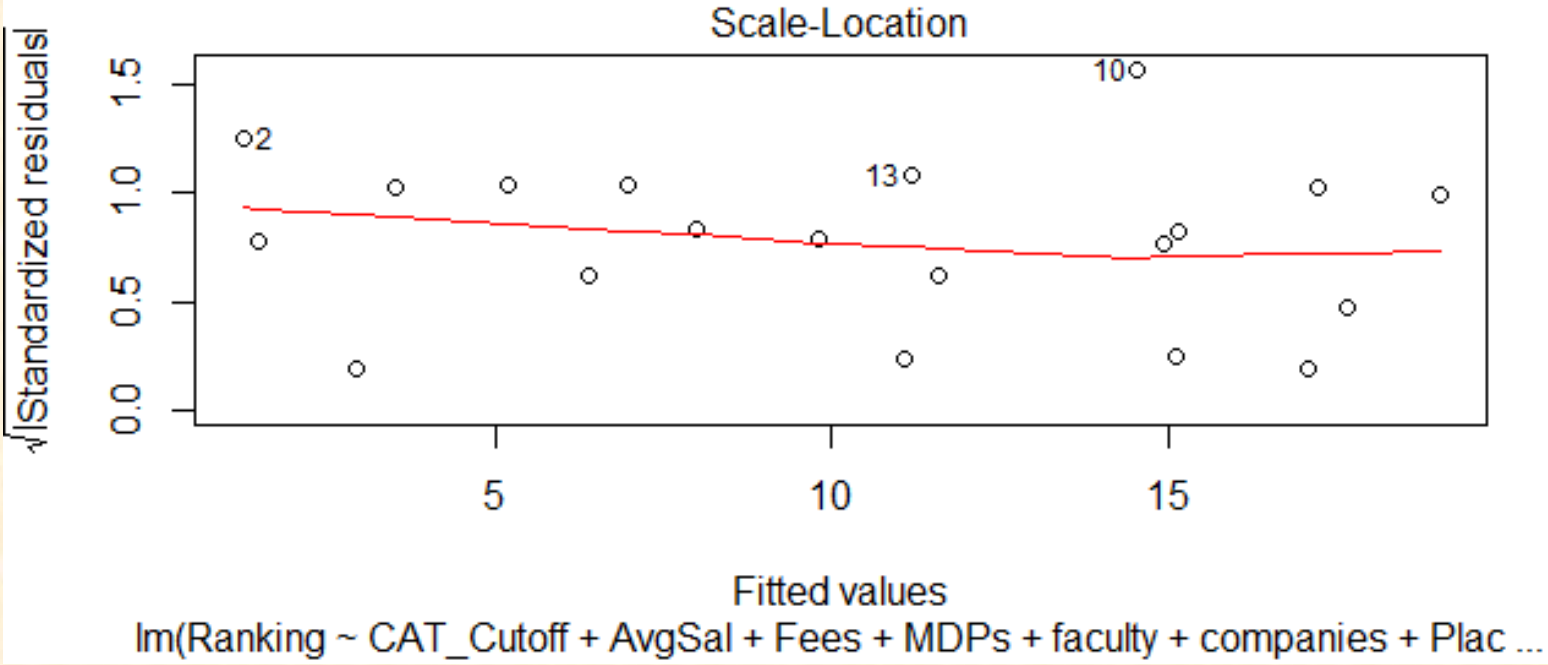
We see the p value is really small implying that our variable i.e ranking is very relevant.

```
> plot(analysis)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty + companies + Plac ...
13
19
10
5
10
15
2
0
-2
-4


Residuals vs Leverage
Standardized residuals
Leverage
lm(Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty + companies + Plac ...
Cook's distance
0.5
0.5
0.0
0.2
0.4
0.6
0.8
2
1
0
-1
-2


Normal Q-Q
Standardized residuals
-2
-1
0
1
2
10
13
2
Theoretical Quantiles
lm(Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty + companies + Plac ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
2
13
10
Fitted values
lm(Ranking ~ CAT_Cutoff + AvgSal + Fees + MDPs + faculty + companies + Plac ...

Predictive analysis of Quality and Value for money

Now, to find the quality of programs an approach of normalization, categorization and weighted mean has been adopted.

Since fees and ranking did not seem to be suitable parameters for quality evaluation, a new subset has been created consisting of remaining parameters. All the values have been then scaled on a scale of 0 to 1 for each parameter. These parameters have been categorized into four categories by an approach of weighted mean:

1) Placements and salary: Three parameters have been used to determine this value. These are avg salary provided, placement percentage and number of companies visiting the campus. The first one has been given twice the weight as compared to other two since more number of companies and 100 percent placement is useless if the candidates are not offered good salaries.

2) Personality development: This is an important aspect to determine the quality of a program. Out of the available parameters two have been found to be suitable to determine the value for this parameter. One of them is the number of student exchange programs and second lesser important is the sex-ratio. The former has been given thrice the weight as compared to the latter

3) International exposure: In this era of globalization this forms an important aspect. Three parameters have been used: student exchange programs, number of international faculty and number of guest speakers. Weights given to them are 3, 2 and 1 respectively.

4) Academics: Last and the most important parameter. Three parameters with equal weights have been used to determine this: Number of Management Development Programs, Number of permanent faculty and number of international faculty.

Finally to evaluate the quality of program in each college the values of these 4 categories have been summed up.

```
> View(mba)
> mba2<-
subset(mba,select=c("CAT_Cutoff","AvgSal","MDPs","faculty","companies","Placement","internation
alfaculty","studentexchange","guestspeakers","sexratio"))
> View(mba2)
> mins <- apply(mba2, 2, min)
> maxs <- apply(mba2, 2, max)
> scaled.mba <- scale(mba2, center = mins, scale = maxs - mins)
> View(scaled.mba)
> categorize<-data.frame(sno=1:20)
> weight<-c(2,1,1)
> categorize$placements<-apply(subset(scaled.mba,select=c("AvgSal","companies","Placement")),
1, function(d) weighted.mean(d, weight))
> weight<-c(3,1)
```

```
> categorize$personality_development<-
apply(subset(scaled.mba,select=c("studentexchange","sexratio")), 1, function(d)
weighted.mean(d, weight))
> View(categorize)
> weight<-c(3,2,1)
> categorize$exposure<-
apply(subset(scaled.mba,select=c("studentexchange","internationalfaculty","guestspeakers")), 1,
function(d) weighted.mean(d, weight))
> weight<-c(1,1,1)
> categorize$academics<-
apply(subset(scaled.mba,select=c("MDPs","faculty","internationalfaculty")), 1, function(d)
weighted.mean(d, weight))
> categorize<-subset(categorize,select=-sno)
> categorize$quality<-rowSums(categorize)
> View(categorize)
```

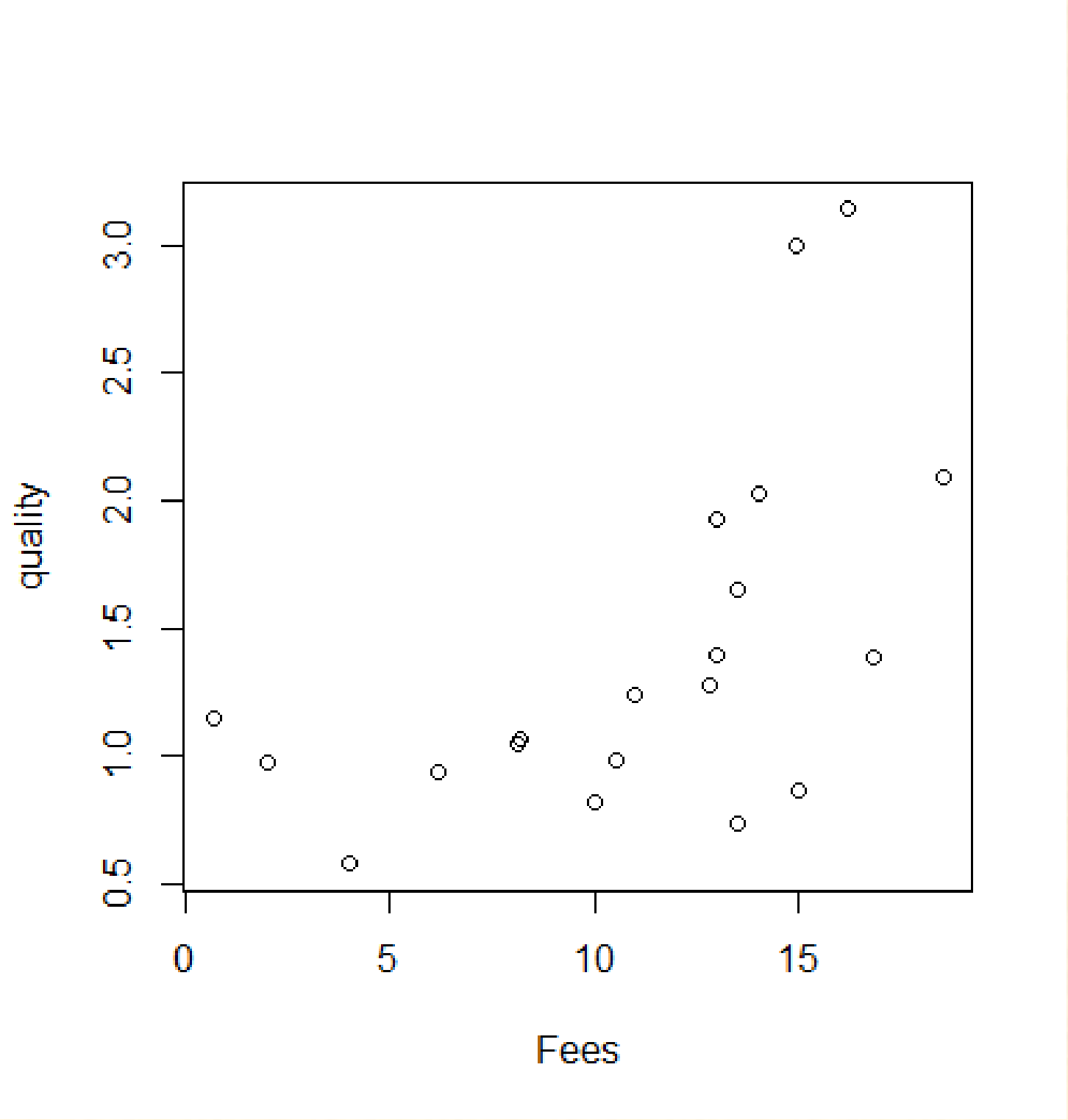
Observations:

It has been observer that Indian Institute of management, Calcutta (ranked 2) provide best quality of education while Indian Institute of management (ranked 1) stands third in terms of quality that has been evaluated here. Whole result can be seen in the files attached

Value for money(VFM)

To begin with correlation coefficient has been found for quality and fees which came out to be 0.56 showing moderate correlation between the two. Further, VFM has been calculated by applying Fees/ quality since it scales down the Fees taken for the program for quality value of one.

```
> categorize<-data.frame(mba$Name,categorize,mba$Fees)
> attach(categorize)
> plot(Fees,quality)
> cov(Fees,quality)
[1] 1.94829
> cor(quality,Fees)
[1] 0.5629342
> categorize$VFM<-Fees/quality
```



We observe that Faculty of management, Delhi provides best value for money in spite of being 11th in terms of quality due to its low fees.

Cluster Analysis

1 Clustering Techniques

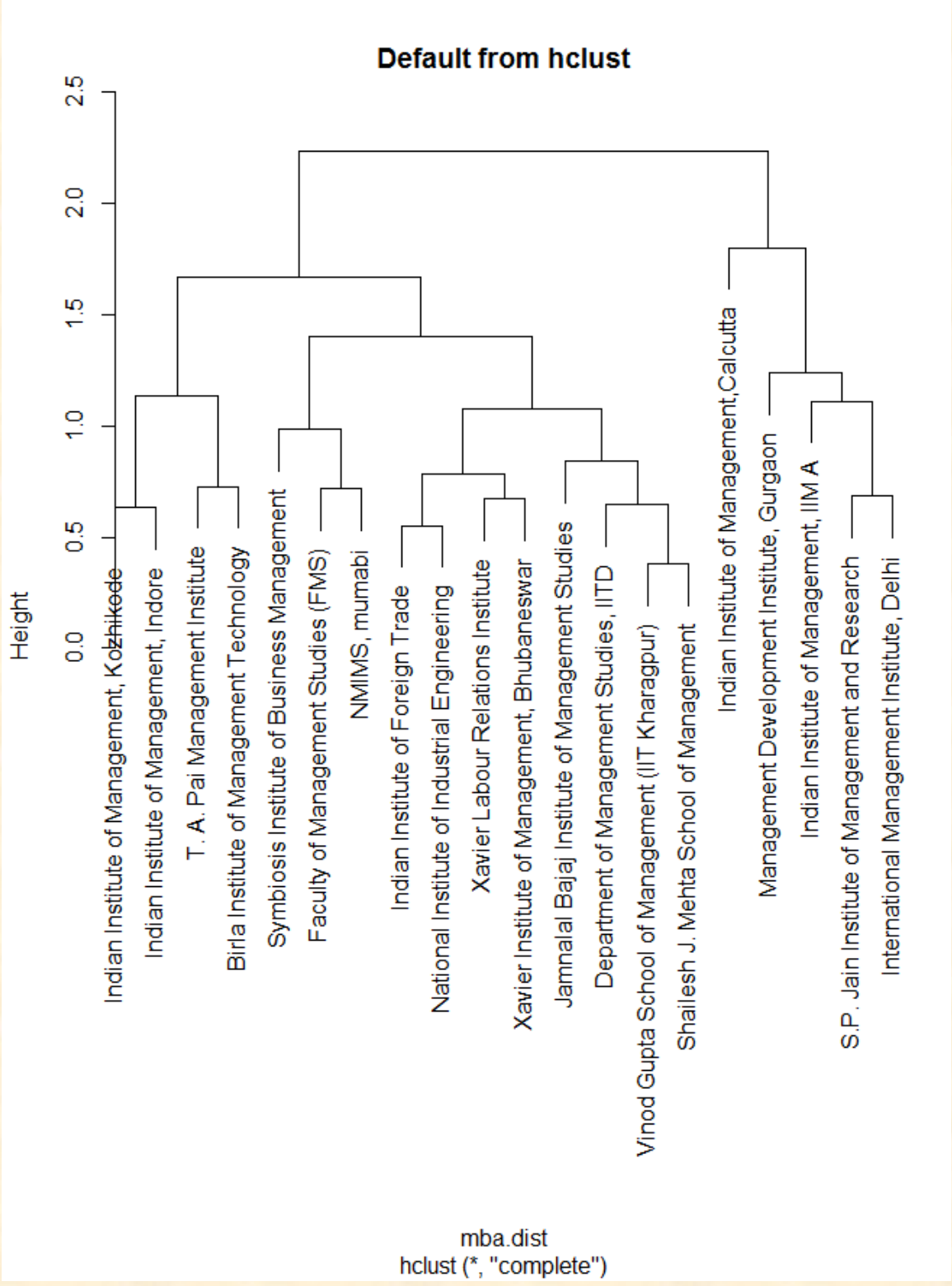
Much of the history of cluster analysis is concerned with developing algorithms that were not too computer intensive, since early computers were not nearly as powerful as they are today. Accordingly, computational shortcuts have traditionally been used in many cluster analysis algorithms. These algorithms have proven to be very useful, and can be found in most computer software.

2 Hierarchial Clustering

For the hierarchial clustering methods, the dendogram is the main graphical tool for getting insight into a cluster solution. When we use hclust or agnes to perform a cluster analysis, we can see the dendogram by passing the result of the clustering to the plot function.

```
> library(cluster)
> mba.dist = dist(scaled.mba)
> mba.dist
          1         2         3         4         5         6         7         8         9        10
2  1.6963909
3  0.7581279 1.7001789
4  1.4521664 1.6159241 1.0471113
5  0.8986495 1.7626823 1.0667895 1.2618641
6  1.1696930 1.2693292 1.4032864 1.4412034 1.0755081
7  1.1678631 1.6135372 1.0525365 1.0951275 1.2831407 1.3156854
8  1.1055398 1.5070174 0.9881033 1.3128763 1.3862977 1.2763001 0.6309410
9  1.1116409 1.8016887 1.0888184 1.3681263 0.6848610 1.2370469 1.2368483 1.3153190
10 1.0938757 1.7447253 0.6318983 1.0527000 1.2468112 1.4775259 1.0288059 0.9838073 1.1281700
11 1.4608854 1.7469246 1.2599190 0.7174457 1.3568417 1.4945404 1.3372197 1.4358631 1.4580655 1.2835726
12 1.1567054 1.7863415 0.7064579 1.0621404 1.3958815 1.4339081 1.0291729 1.0655445 1.2735279 0.5495437
13 1.2070899 1.8141477 0.6718467 0.7986433 1.1776092 1.5714364 0.9715457 1.1015873 0.9832724 0.6087429
14 1.2895023 2.0073285 1.1119909 0.9831645 1.0561249 1.5320306 1.1721629 1.2742494 1.0136529 1.0444463
15 1.5684234 2.2313577 0.9927800 1.0809313 1.4977925 1.9258341 1.5482377 1.5527042 1.3937237 0.7609322
16 1.3636468 2.1665422 1.0195840 1.2392492 1.4771231 1.7839580 1.3460682 1.3683940 1.5665630 0.6864777
17 1.4083335 2.1347262 0.9572396 1.0767705 1.3476790 1.7024844 1.5052155 1.5331409 1.2985715 0.6725214
18 1.3504965 2.1569051 1.1425945 1.4515360 1.3890302 1.8132538 0.9042505 1.0483255 1.1085249 0.8979092
19 1.4049760 2.0681117 0.9850471 1.1552136 1.4743876 1.6670550 1.5313867 1.5319150 1.4741874 0.6954241
20 1.5859824 2.0661094 1.3975871 1.3247359 1.5504667 1.7754767 0.8924012 1.1341990 1.2367669 1.1880091
          11        12        13        14        15        16        17        18        19
2
3
4
5
6
7
8
9
10
11
12 1.2582261
13 1.0848392 0.7816528
14 0.7498059 1.1505431 0.8537213
15 1.2368515 0.9419262 0.7522107 1.0090841
16 1.4047896 0.8483823 1.0781732 1.1803578 0.8395047
17 1.2267493 0.8352005 0.8139402 1.0228043 0.3970360 0.7232082
18 1.5791620 1.1180095 0.8858708 1.0346724 1.2640593 1.1915471 1.2929133
19 1.2913997 0.7074012 1.0058604 1.1999371 0.6446853 0.6152047 0.3786977 1.4286726
20 1.4057960 1.3341449 0.9307353 1.0538733 1.4618082 1.5720286 1.4906699 0.7265169 1.6675998
```

```
> mba.hclust = hclust(mba.dist)
> plot(mba.hclust,labels=mba$Name,main='Default from hclust')
```



The given dendogram was observed.

If we choose any height along the y-axis of the dendogram, and move across the dendogram counting the number of lines that we cross, each line represents a group that was identified when objects were joined together into clusters. The observations in that group are represented by the branches of the dendogram that spread out below the line. For example, if we look at a height of 2, and move across the x-axis at that height, we'll cross two lines. That defines a two-cluster solution; by following the line down through all its branches, we can see the names of the mba colleges that are included in these two clusters. Since the y-axis represents how close together observations were when they were merged into clusters, clusters whose branches are very close together (in terms of the heights at which they were merged) probably aren't very reliable. But if there's a big difference along the y-axis between the last merged cluster and the currently merged one, that indicates that the clusters formed are probably doing a good job in showing us the structure of the data. Looking at the dendogram for the colleges data, there are clearly two very distinct groups; the left hand group seems to consist of two more distinct cluster, while most of the observations in the right hand group are clustering together at about the same height For this data set, it looks like either two or three groups might be an interesting place to start investigating. This is not to imply that looking at solutions with more clusters would be meaningless, but the data seems to suggest that two or three clusters might be a good start. For a problem of this size, we can see the names of the colleges, so we could start interpreting the results immediately from the dendogram, but when there are larger numbers of observations, this won't be possible.

One of the first things we can look at is how many mba colleges are in each of the groups. We'd like to do this for both the two cluster and three cluster solutions. We can create a vector showing the cluster membership of each observation by using the cutree function. Since the object returned by a hierarchical cluster analysis contains information about solutions with different numbers of clusters, we pass thecutree function the cluster object and the number of clusters we're interested in. So to get cluster memberships for two to six cluster solution, we could use:

```
> counts = sapply(2:6,function(ncl)table(cutree(mba.hclust,ncl)))
> names(counts) = 2:6
> counts

$`2`

 1  2
 5 15

$`3`

 1  2  3
 4  1 15

$`4`

 1  2  3  4
 4  1 11  4

$`5`

1 2 3 4 5
4 1 8 3 4

$`6`

1 2 3 4 5 6
3 1 8 3 1 4
```

We observe size 5 to be most optimum by the number of members in each group

```
> groups.5 = cutree(mba.hclust,5)
```

To see the names of colleges in first group:

```
> mba$Name[groups.5==1]
[1] Indian Institute of Management, IIM A
[2] S.P. Jain Institute of Management and Research
[3] Management Development Institute, Gurgaon
[4] International Management Institute, Delhi
20 Levels: Birla Institute of Management Technology ...
```

As usual, if we want to do the same thing for all the groups at once, we can use sapply:

```
> sapply(unique(groups.5),function(g)mba$Name[groups.5 == g]
+ )
[[1]]
[1] Indian Institute of Management, IIM A          S.P. Jain Institute of Management and
Research
[3] Management Development Institute, Gurgaon      International Management Institute, Delhi
20 Levels: Birla Institute of Management Technology ... Xavier Labour Relations Institute

[[2]]
[1] Indian Institute of Management,Calcutta
20 Levels: Birla Institute of Management Technology ... Xavier Labour Relations Institute

[[3]]
[1] Xavier Labour Relations Institute                Indian Institute of Foreign Trade
[3] National Institute of Industrial Engineering     Xavier Institute of Management,
Bhubaneswar
[5] Department of Management Studies, IITD           Jamnalal Bajaj Institute of Management
Studies
[7] Vinod Gupta School of Management (IIT Kharagpur) Shailesh J. Mehta School of Management
20 Levels: Birla Institute of Management Technology ... Xavier Labour Relations Institute

[[4]]
[1] Faculty of Management Studies (FMS)        NMIMS, mumabi
[3] Symbiosis Institute of Business Management
20 Levels: Birla Institute of Management Technology ... Xavier Labour Relations Institute

[[5]]
[1] Indian Institute of Management, Kozhikode Indian Institute of Management, Indore
[3] T. A. Pai Management Institute            Birla Institute of Management Technology
20 Levels: Birla Institute of Management Technology ... Xavier Labour Relations Institute
```

A very useful method for characterizing clusters is to look at some sort of summary statistic, like the median, of the variables that were used to perform the cluster analysis, broken down by the groups that the cluster analysis identified. The aggregate function is well suited for this task, since it will perform summaries on many variables simultaneously. Let's look at the median values for the variables we've used in the cluster analysis, broken up by the cluster groups. One oddity of the aggregate function is that it demands that the variable(s) used to divide up the data are passed to it in a list, even if there's only one variable:

```
> aggregate(scaled.mba,list(groups.5),median)

  Group.1 CAT_Cutoff    AvgSal      MDPs   faculty companies Placement internationalfaculty
1       1  0.6094464 0.7618651 0.4267516 0.7356322 0.5681818 1.0000000             0.8928571
2       2  1.0000000 0.8601166 1.0000000 0.9770115 1.0000000 0.9115044             0.1428571
3       3  0.8633824 0.5395504 0.1114650 0.3390805 0.3219697 0.9203540             0.0000000
4       4  0.8633824 0.7077435 0.5414013 0.2873563 0.5000000 0.3362832             0.2857143
5       5  0.1269680 0.2926728 0.4394904 0.6666667 0.6477273 1.0000000             0.0000000

  studentexchange guestspeakers   sexratio
1      0.21120690      0.196875 0.16003832
2      1.00000000      1.000000 0.24424093
3      0.00000000      0.084375 0.62992059
```

```
4      0.05172414      0.156250 0.21735320
5      0.15517241      0.050000 0.08266809
```

```
> aggregate(mba[,-c(1,2)],list(groups.5),median)

  Group.1 CAT_Cutoff AvgSal  Fees  MDPs faculty Campus_.area Students  boys girls companies
Placement
1       1      92.00 15.950 14.45  68.0    69.0        41.00    206.5 170.5    79     102.0
100.0
2       2      99.69 17.130 16.20 158.0    90.0       135.00    458.0 349.0   109     159.0
99.0
3       3      97.00 13.280  8.15  18.5    34.5        30.00    170.5 140.0    22      69.5
99.1
4       4      97.00 15.300 10.50  86.0    30.0        14.69    220.0 165.0    64      93.0
92.5
5       5      82.50 10.315 13.00  70.0    63.0        70.50    370.5 246.5   124     112.5
100.0

  internationalfaculty studentexchange guestspeakers sexratio
1                12.5            24.5         101.5 2.569761
2                 2.0           116.0         487.0 3.201835
3                 0.0             0.0          47.5 6.096970
4                 4.0             6.0          82.0 3.000000
5                 0.0            18.0          31.0 1.988976
```

3 PAM: Partitioning Around Medoids

Unlike the hierarchical clustering methods, techniques like k-means cluster analysis (available through the kmeans function) or partitioning around mediods (avaiable through the pam function in the clusterlibrary) require that we specify the number of clusters that will be formed in advance. pam offers some additional diagnostic information about a clustering solution, and provides a nice example of an alternative technique to hierarchical clustering. To use pam, we must first load the cluster library. We can pass pam a data frame or a distance matrix; since we've already formed the distance matrix, we'll use that. pam also needs the number of clusters we wish to form. Let's look at the five cluster solution produced by pam:

```
>mba.pam = pam(mba.dist,5)
```

We can use table to compare the results of the hclust and pam solutions:

```
> table(groups.5,mba.pam$clustering)

groups.5 1 2 3 4 5
       1 4 0 0 0 0
       2 0 1 0 0 0
       3 0 0 8 0 0
       4 0 0 0 3 0
       5 0 0 1 0 3
```

The solutions seem to agree, except for 1 observations that hclust put in group 5 and pam put in group 3. Which observations was it?

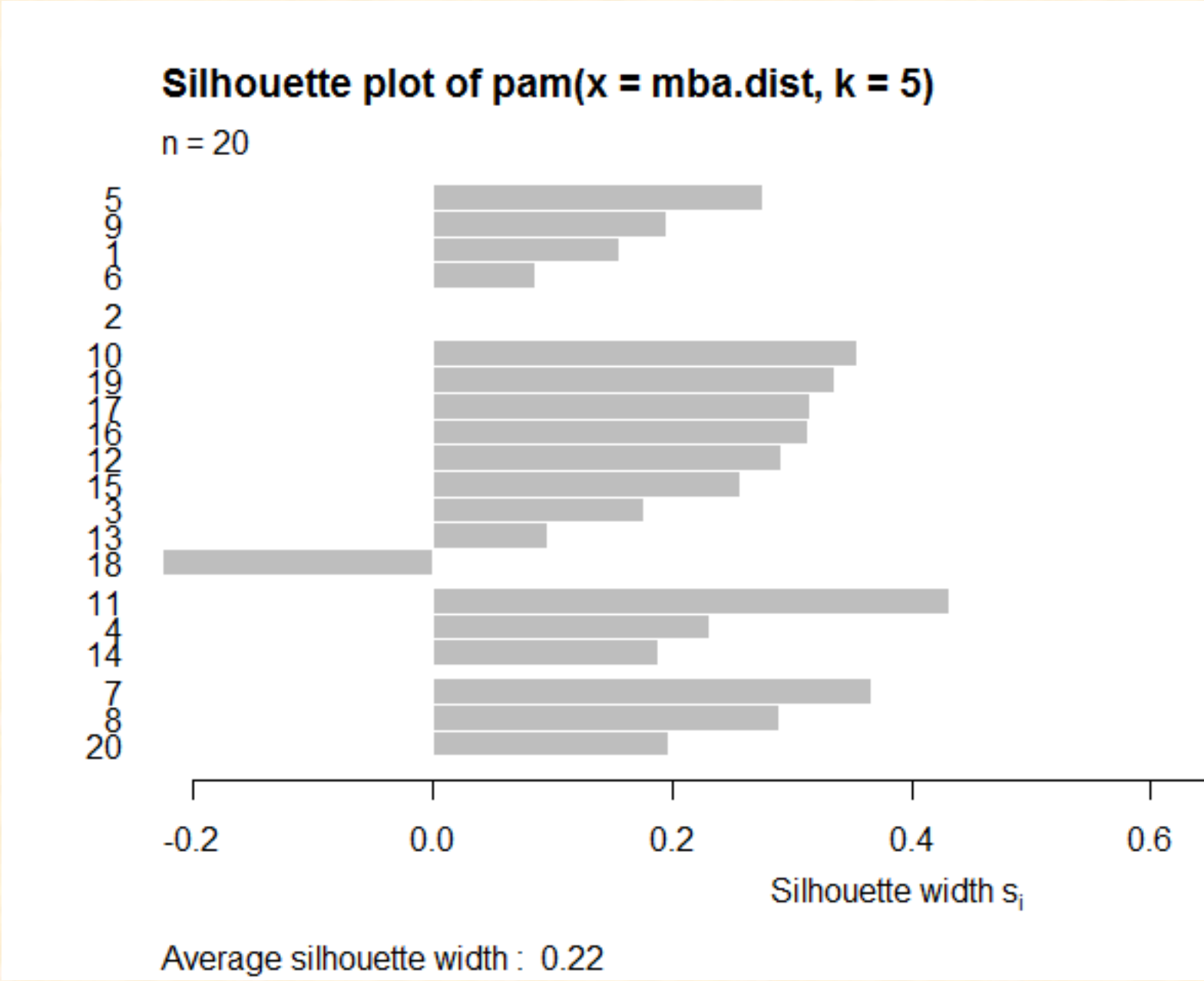
```
> mba$Name[groups.5 != mba.pam$clustering]
[1] T. A. Pai Management Institute
```

One novel feature of pam is that it finds observations from the original data that are typical of each cluster in the sense that they are closest to the center of the cluster. The indexes of the medoids are stored in the id.med component of the pam object, so we can use that component as a subscript into the vector of mba colleges names to see which ones were selected:

```
> mba$Name[mba.pam$id.med]
[1] S.P. Jain Institute of Management and Research
[2] Indian Institute of Management,Calcutta
[3] Indian Institute of Foreign Trade
[4] NMIMS, mumabi
```

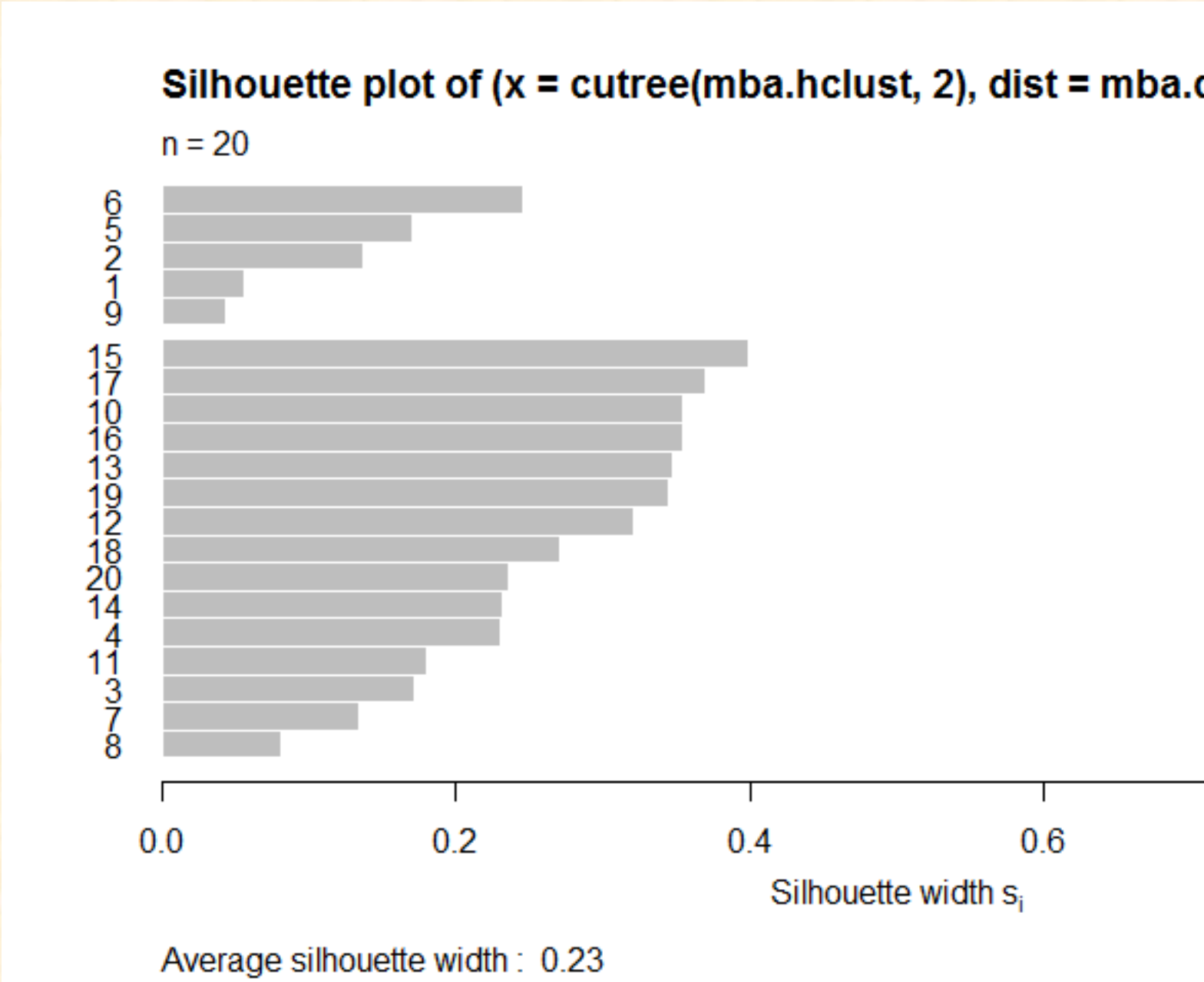
[5] Indian Institute of Management, Kozhikode

sihouette plot is very useful in locating groups in a cluster analysis that may not be doing a good job; in turn this information can be used to help select the proper number of clusters. For the current example, here's the silhouette plot for the five cluster pam solution, produced by the command
> plot(mba.pam)



The plot indicates that there is a good structure to the clusters, with most observations seeming to belong to the cluster that they're in

```
>plot(silhouette(cutree(mba.hclust,2),mba.dist))
```



Above graph shows that, as suggested by hierarchical clustering, cluster of two is well suited too.(much better than 5)

Thus the overall structure seems to take more and more scattered shape as the number of clusters is increased. If each indivisual is categorized as a cluster we'll get the most scattered plot of random lines with each line representing a cluster. Cluster analysis is an important technique for data analysis since it helps us to segregate large amount data data into clusters thus easing further analysis.

CONCLUSIONS

MBA is one of the most prestigious courses in India. Various management colleges have been setup in the country to provide this course. Thus it is important for the candidates to know which one is better than others. Various rankings are available online to help them with this important decision. Here we studied the parameters collected from various such sources and drew various conclusions based on the same.

The overall study of top MBA colleges of India provided us splendid opportunity to understand and apply various strategies of data analysis. According to the statistical study presented various conclusions can be drawn.

1) Ranking of colleges is highly dependent on various parameters like number of faculty, number of courses offered, average salary offered during placements and many more. As per the results of regression analysis the given rankings of colleges is quite accurate with IIM Ahmadabad as the best MBA college in the state followed by IIM Calcutta and Xaviers Labour Institute.

2) To find the quality of programs an approach of normalization, categorization and weighted mean has been adopted. This predictive modelling suggests that the ranking of courses is closely related to quality calculated with few differences. Indian Institute of management, Calcutta (ranked 2) provide best quality of education while Indian Institute of management (ranked 1) stands third in terms of quality that has been evaluated here.

3) Value for money calculation represents the almost direct relationship between fees and quality of program with few exceptions. Faculty of management, Delhi takes extremely low fees which take it up the ladder to the top of “value for money” queue in spite of it being 11th in terms of quality due to its low fees.

4) For clustering, two techniques were adopted both suggesting that the colleges can be clustered mainly in two groups first one comprising of IIM A, S.P. Jain Institute of Management and Research, Management Development Institute, Gurgaon and International Management Institute, Delhi and IIM Calcutta and the second containing all others. This clusters can further be divided with nest suitable clustering coming up at 5 clusters. As we increase the the number of clusters our graphs become more and more structurally diversified.

5) There is no certified solution to the problem statement. The more we play with our variables and draw careful conclusions the more we tend towards better analytical models.


Thank you!