MGS 616 PREDICTIVE ANALYTICS
ASSIGNMENT 3: DATA PREPARATION

GROUP MEMBERS:

- NEERAJ KUMAR
- KEDAR IMAAMDAR
- SHRUTI BHATNAGAR
- POOJA AGRAWAL
- UDITA DWIVEDI

Data Preparation and Cleaning

NA VALUES

- For the purpose of cleaning data we tried to find out NA values and we got the following result

Filter

	Columnnames	Navalues
1	CollegeName	0
2	State	0
3	ispublicorprivate	0
4	applicationRecieved	10
5	applicationAccepted	11
6	newStudentEnrolled	5
7	top10%NewStudents	235
8	top25%NewStudents	202
9	fullTimeUndergrad	3
10	parttimeUndergrad	32
11	instateTutionFees	30
12	outstateTutionFees	20
13	roomcost	321
14	board	498
15	additionalFees	274
16	estimatedBookcost	48
17	estimatedPersonalCost	181
18	percentteachwithPhd	0
19	studentFacultyRatio	2
20	graduationRate	98

Analysis on output of NA values

Total data with NA:

datawithNA	1302 obs. of 20 variables

As per the observation from above table there are too many na values if we completely remove the na values from the data then we got

Total data without NA:

Data	
datawithoutNA	471 obs. of 20 variables

i.e. almost 850 rows got deleted which is a very high chunk of data.

So we cannot opt this method as we will not get the accurate prediction with only 25% percent of original data.

Missing Data can be classified into 3 categories:

MCAR : Missing Completely at random, which means there is nothing systematic about which observations are missing values.

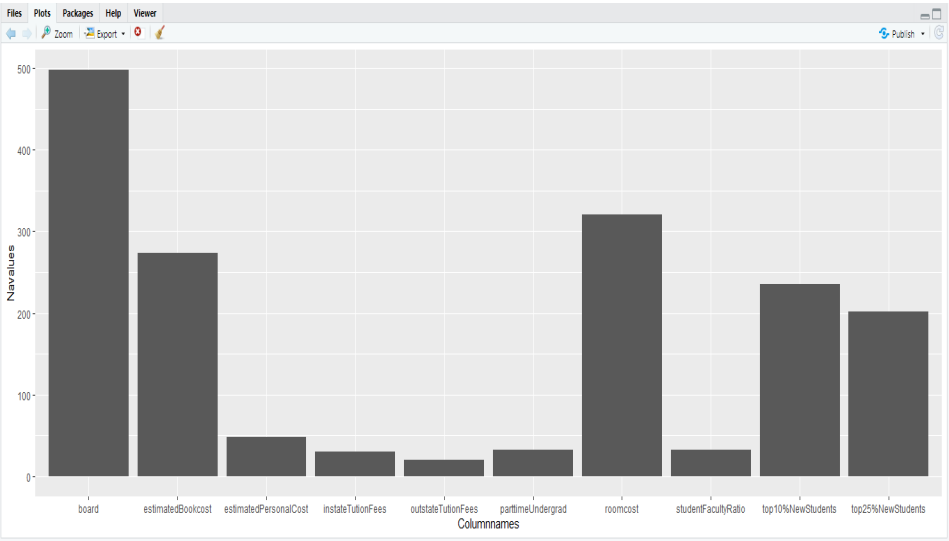
MAR : In this case, the missingness may still be random but may be due entirely to observed variables.

MNAR: In this case, the data is missing 'Not At Random' , which points to cases of missingness due to values of data. For example, censored data.

Missing values in all fields

There are NA fields in 15 out of 20 fields of the dataset.

Graphical representation of number of missing values vs fields



The bottom 3 fields with missing values are Application received, full time undergrad and stud faculty ratio with missing values less than 10, while the top 3 fields are room, board, additional fee with missing values more than 250. While costs of room, board and additional fee are subject to a university's internal administration and hence may be harder to ascertain, there are other fields like top10%NewStudents and top25%NewStudents that have missing values in the range of 200-230. It may be concluded that as a whole this is an instance of MCAR.

Methods of dealing with missing data

Educated Guessing: Since the number of observations are 1000 and 15 out of 20 fields have NA,this is not an appropriate way to go about treatment of the missing data.

Average imputation: This suggests replacing missing values in a field by the average value of that field. Since 15 out of 20 fields have missing values, this would introduce a reduction in the variance of the whole dataset, making it less suitable to work with.

Multiple Imputation: This is the most sophisticated out of all methods. The software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions[1]. [1]

Method used to clean data:

We have used the Mice library in R, which performs the imputation via chained equations. We compute the imputation for the data using "predictive mean matching" method and analyze the set of field values obtained for the dataset.

```
data.imp = mice(data, m = 5, method = 'pmm')
```

Considering all imputations for a certain field:

```
data.imp$imp$applicationsrecieved
```

```
> data.imp$imp$applicationrecieved
        1    2    3     4    5
229   540 1600 1499   983  440
316  6548 6334 9643 12289 6334
354  1499  571  263   894 1386
542   314   77  384   384  260
765  2837 4219 2967  3294 4255
918  4772 6233 6756  3055 4800
936  1380 2643 2174   404 1732
946   774  354  360   272  394
1151  177  492  125   861  342
1278 1256 2807 2056   823 1827
> |
```

Final substitution of the imputed values into the dataset

```
clean_data = Complete(data.imp, 3)
```

Checking for NA values in every field:

```
sum(is.na(clean_data$applicationrecieved))
```

1 http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html

```
> sum(is.na(clean_data$applicationrecieved))
[1] 0
>
```

```
> write_csv(clean_data, path = "cleanuniversities.csv")
> sum(is.na(Data$applicationrecieved))
[1] 10
> sum(is.na(Data$applicationaccepted))
[1] 11
> sum(is.na(Data$newstudentsenrolled))
[1] 5
> sum(is.na(Data$newstudentsfromtop10))
[1] 235
> sum(is.na(Data$newstudentsfromtop25))
[1] 202
> sum(is.na(Data$ftundergrad))
[1] 3
> sum(is.na(Data$ptundergrad))
[1] 32
> sum(is.na(Data$instatetuition))
[1] 30
> sum(is.na(Data$outofstatetuition))
[1] 20
> sum(is.na(Data$room))
[1] 321
> sum(is.na(Data$additionalfees))
[1] 274
> sum(is.na(Data$estimbookcosts))
[1] 48
> sum(is.na(Data$facwithpHD))
[1] 32
> sum(is.na(Data$studfacratio))
[1] 2
> sum(is.na(Data$gradrate))
[1] 98
> plot(wt, mpg, main="Scatterplot Example",
+      xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
Error in plot(wt, mpg, main = "Scatterplot Example", xlab = "Car Weight
  object 'wt' not found
> sum(is.na(clean_data))
[1] 0
>
```

Detection of outliers

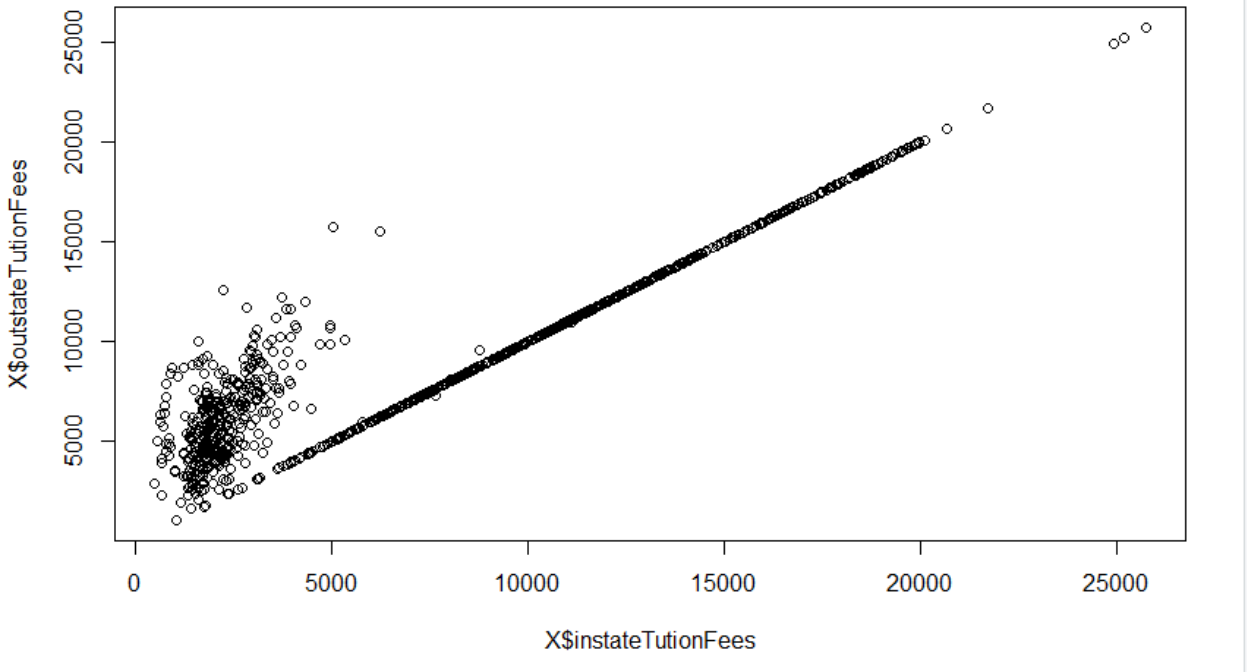
We have assumed following variables as Important variables in our data:-

- application received
- public or private
- tuition fees
- fac with phd
- total cost(in state)or outstate

Outlier Detection

To detect outliers in the fees we have plotted a scatter plot of instate or outstate fees variable

[2]



From the above plot we can see that some outliers may present when instate fees is greater than 15000 and outstate fees is greater than 20000 so tried to find outliers by fixing this range and we got the following result.

2 https://measuringu.com/handle-missing-data/

	X.instateTutionFees	X.outstateTutionFees	X.outlier
456	20100	20100	TRUE
485	20655	20655	TRUE
512	25180	25180	TRUE
976	24940	24940	TRUE
1226	21700	21700	TRUE
1234	25750	25750	TRUE

So out of 1302 observations 6 records are coming out as outliers for the instate and outstate fees variable.

We have removed the outliers from the data and our final clean data set is

https://docs.google.com/spreadsheets/d/1hWLC1swRjsWzr2MtXad-4Dyiel0ELOHoVKV86W7N9sI/edit#gid=452641201

GITHUB Code Link

https://github.com/poojaagr21/Prediction-of-student-university-selection-Public-private-/blob/main/detaprep.R

References

1. http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html
2. https://www.programmingr.com/r-error-messages/subscript-out-of-bounds-r/
3. https://stackoverflow.com/questions/15031338/subscript-out-of-bounds-general-definition-and-solution/15031603
4. http://www.sthda.com/english/wiki/exporting-data-from-r