U.S. News and World Report's College Data Analysis
- STORMTROOPERS

INTRODUCTION:

The objective of this case analysis is to assess U.S. News and World Report's College Data and develop a model which will forecast likelihood of enrollment of a student in a university.

Why do we care?

We care about this data because nowadays there are huge number of universities have been establishing with new courses and it's important that they should be aware of the key factors which student mostly consider enrolling. This exploratory analysis will help both the student and universities.

Objectives:

1. **Dataset Discovery**
2. **Methodology**
3. **Data visualization and pattern discovery**
4. **Conclusion**

Overview of dataset:

The dataset we chose contains US universities undergraduate's data. This dataset has been picked from Kaggle website. This data has been collected by StatLib library which is maintained at Carnegie Mellon University since the year 1995.

The data set has 776 observations and 17 variables. The meta data is described below

Variable Name	Meta data
College Type	Indicates whether private or public university
Apps	Number of applications received
Accept	Number of applications accepted
Enroll	Number of new students enrolled
Top10perc	Percent of new students from top 10% of H.S. class
Top25perc	Percent of new students from top 25% of H.S. class
F.Undergrad	Number of fulltime undergraduates
P.Undergrad	Number of part time undergraduates
Outstate	Out-of-state tuition fee
Room.Board	Room and board costs
Books	Estimated book costs
Personal	Estimated personal spending
PhD	Percent of faculty with Ph.D.'s
Terminal	Percent of faculty with terminal degree
perc.alumni	Percent of alumni who donate
Expend	Instructional expenditure per student
Grad.Rate	Graduation rate

	College.Name <fctr>	College.Type <fctr>	Apps <int>	Accept <int>	Enroll <int>	Top10perc <int>	Top25perc <int>
1	Abilene Christian University	Private	1660	1232	721	23	52
2	Adelphi University	Private	2186	1924	512	16	29
3	Adrian College	Private	1428	1097	336	22	50
4	Agnes Scott College	Private	417	349	137	60	89
5	Alaska Pacific University	Private	193	146	55	16	44
6	Albertson College	Private	587	479	158	38	62

6 rows | 1-8 of 18 columns

Some of the variables we intend to study are:

Variable Name	Variable Type
College Type	Categorical
Enroll	Numeric
Grad.Rate	Numeric
Outstate	Numeric
PhD	Numeric
Top25perc	Numeric
F. Undergrad	Numeric
P. Undergrad	Numeric

What factors contributed in universities enrollment (i.e. Enroll) ?

College Type **:** This field has a big effect in the decision to enroll due to the level of tuition cost for the private ones.

Outstate tuition fee: We believe that outstate tuition fee might have an effect on-student enrollment.

PhD: This variable says about Number of professors with PhD degree and we plan to study this variable as this will have an effect in new student enrollment**.**

F. Undergrad: We observed that Number of full time under graduates currently studying in the university may influence student enrollment.

What factor influences the graduation rate (i.e. Grad.Rate) ?

F.Undergrad: We observed that total no. of full-time under graduates will impact the graduation rate.

P.Undergrad: We observed that total no.of part-time under graduates will impact the graduation rate.

Top25perc: We believe Top25 % of students joined in a university will influence the graduation rate.

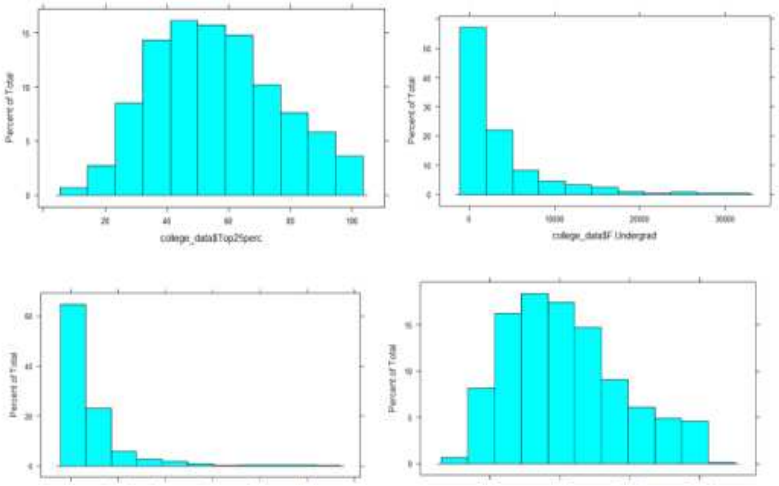
Methodology:

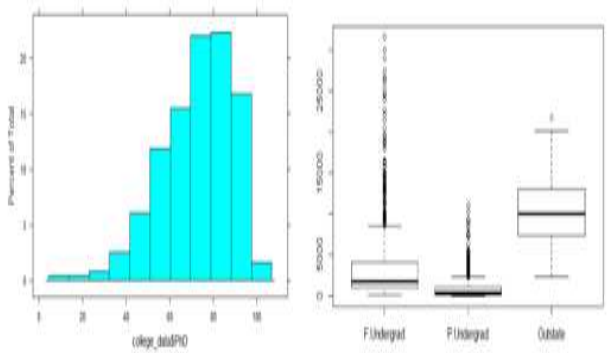
As we don't have any missing values or inconsistent data it doesn't require preprocessing. Hence, the data is ready for analysis and interpretation. So, we will use descriptive statistics for summarizing the data and exploratory data analysis to check relationship between the variables. Further, we will find the correlation and choose the model that best suits for predictive analysis.

Finding the descriptive statistics for the below numerical variables and observing the patterns.

```
   Top25perc      F.Undergrad     P.Undergrad       Outstate        PhD
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :  8.00
 1st Qu.: 41.0   1st Qu.:  991   1st Qu.:   95.0   1st Qu.: 7305   1st Qu.: 62.00
 Median : 54.0   Median : 1707   Median :  352.5   Median : 9990   Median : 75.00
 Mean   : 55.8   Mean   : 3683   Mean   :  828.3   Mean   :10443   Mean   : 72.64
 3rd Qu.: 69.0   3rd Qu.: 3969   3rd Qu.:  964.0   3rd Qu.:12931   3rd Qu.: 85.00
 Max.   :100.0   Max.   :31643   Max.   :10962.0   Max.   :21700   Max.   :103.00
```

```
sd(college_data$Top25perc)      [1] 19.81753
sd(college_data$F.Undergrad)    [1] 4831.686
sd(college_data$P.Undergrad)    [1] 1323.658
sd(college_data$Outstate)       [1] 4025.254
sd(college_data$PhD)            [1] 16.32938
```

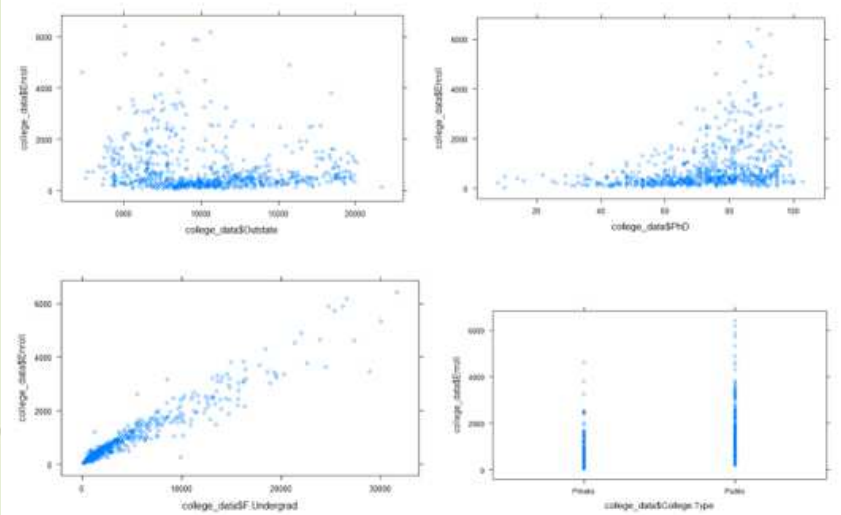


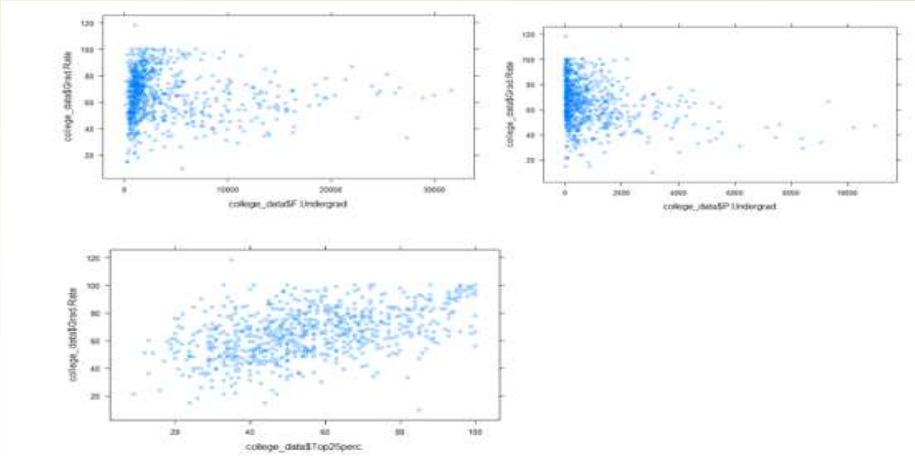


Inferences about the Descriptive statistics:

- Based on summary statistics we can infer mean, median and Range for the variables.
- Top25perc: The skewness is 0.27 and it is fairly symmetric.
- F.Undergrad: The skewness is 1.2 and it is positively skewed.
- P.Undergrad: The skewness is 1.07 and it is positively skewed.
- OutState: The skewness is 0.3 and it is fairly symmetric.
- PhD: The skewness is - 0.4 and it is fairly symmetric.
- Outliers in a data can distort predictions and affect the accuracy if we don't detect and handle them appropriately. By using box plot we can infer that the variables *F.Undergrad* and *P.Undergrad* has extreme outliers.

Finding the Relationship between response and exploratory variable using XY plot:





From the above graphs we can infer the following facts,

Response_Variable	Direct Relationship	Inverse Relationship
Enroll	PhD, F.Undergrad	Outstate
Grad.rate	Top25perc	F.Undergrad, P.Undergrad

Conclusion:

In Exploratory analysis we identified the influencing variables that affect student enrollment and graduation rate. In addition, it is observed that the variables F.Undergrad and P.Undergrad are not normally distributed. Since data mining models expect data to be normally distributed, we need to do transformations to achieve normality, this can be done using any of the following:

1. Log transformation ; 2.Square root transformation ; 3.Reverse square root transformation.

After transforming to normality, we will proceed with establishing correlation and by then build a model for predictive analysis in Phase 2.