**Final Project Report**

**Student Enrollment and graduation rates in US Universities.**

**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 and Grad rate in a university. 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.

Additionally, we will predict the college type (public or private) for the given data. Though this will not add much value, we are trying to implement Knn model to predict.

**Data and Methodology:**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 |

*Based on the universities data, we chose to explore the factors that play an important role in student enrollment and graduation rate.*

*Some of the variables we studied in Phase 1,*

|  |  |
| --- | --- |
| **Variable Name** | **Variable Type** |
| College Type | Categorical |
| Enroll | Numeric |
| Grad.Rate | Numeric |
| Outstate | Numeric |
| PhD | Numeric |
| Top25perc | Numeric |
| F. Undergrad | Numeric |
| P. Undergrad | Numeric |
| S.F. Ratio | Numeric |

As we don’t have any missing values or inconsistent data we don’t require Preprocessing. Hence, the data is ready for analysis and interpretation. So, we used descriptive statistics for summarizing the data and exploratory data analysis to check relationship between the variables in Phase 1. Further, we found the correlation and chose the model that best suits for predictive analysis.

**Key Exploratory Findings observed in Phase 1:**

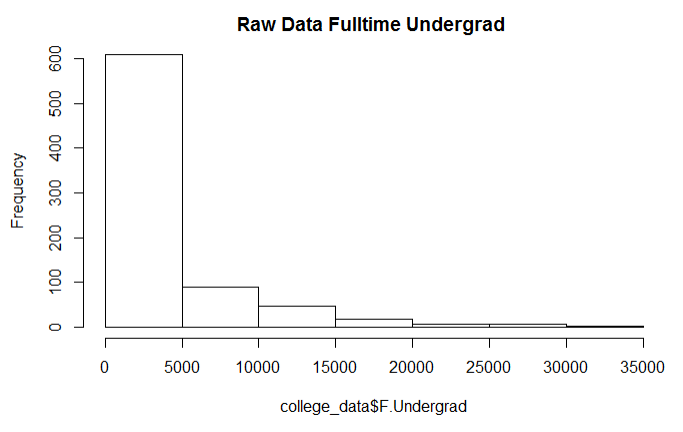
1. Variables F. Undergrad and P. Undergrad are positively skewed.
2. Variables Top25perc, OutState, PhD, S.F. Ratio are symmetric.
3. Using box plot we inferred variables F. Undergrad and P. Undergrad have extreme outliers.
4. Using XY plot we observed that student enrollment increases when no. of PhD and F. Undergrad increases and decreases when outstate tuition fee increases. Also, Grad rate directly depends on No. of top25perc students enroll in a university and decreases when there are more F. Undergrad and P. Undergrad.
5. Variables F. Undergrad and P. Undergrad are not normally distributed.

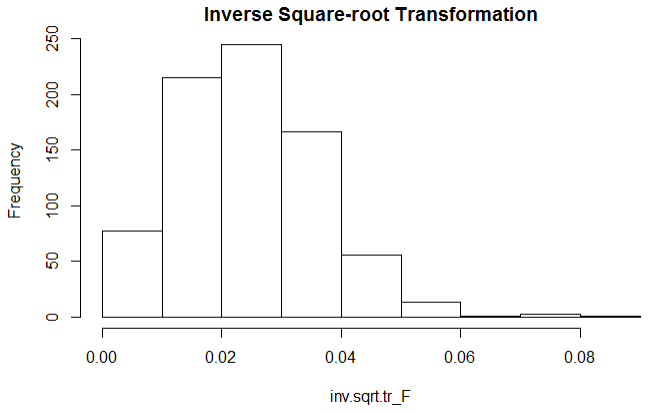
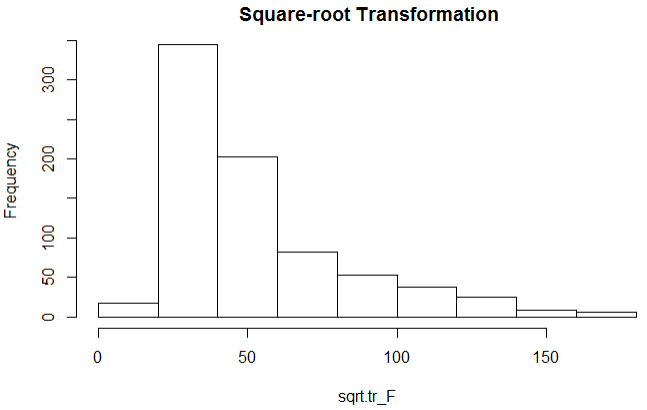
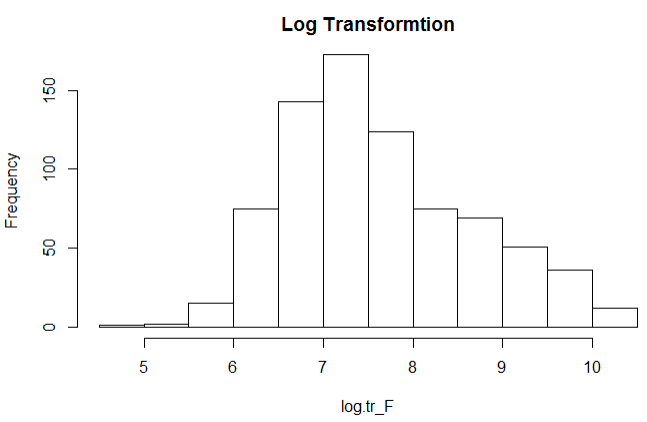
**Analysis Results:**

As concluded in Phase 1, Data mining models expect data to be normally distributed, so there is a need to do transformations to achieve normality for variables F. Undergrad and P. Undergrad.

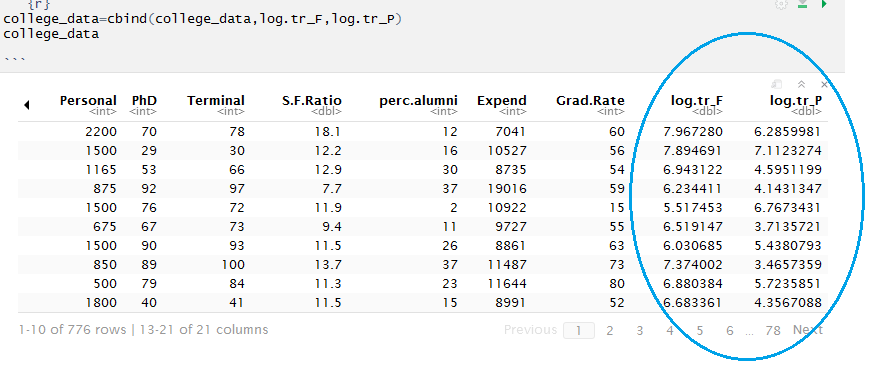
We applied below transformation and observed the normality,

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

****

****

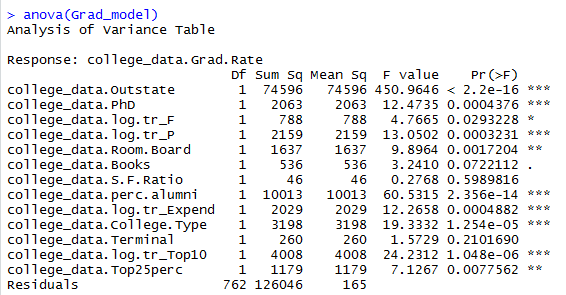
From the histogram it is very clear that Log transformation shows that data is normally distributed. So, we have chosen the transformed (F. Undergrad and P. Undergrad) normality data for further analysis.



**Model Selection:**

Out of different models for prediction, we felt *Multivariate linear regression* would be the best fit for our analysis as our response variable is numeric. Also, we will use classification models knn to predict college Type.

Created Multivariate Liner Regression Model for Grad.Rate below,

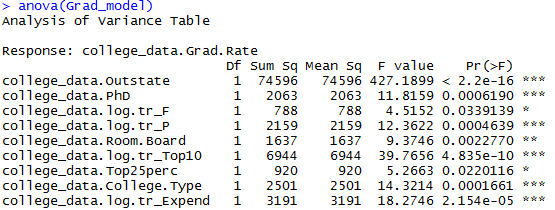


By looking at the p values, the values which are less that 0.05 are significant and they influence response variable. We are ignoring perc.alumni as it doesn’t make any sense.

At the beginning of the analysis we didn’t consider the above variables

* S.F. Ratio(Student:Faculty)
* Room. Board
* Top10perc
* College.Type

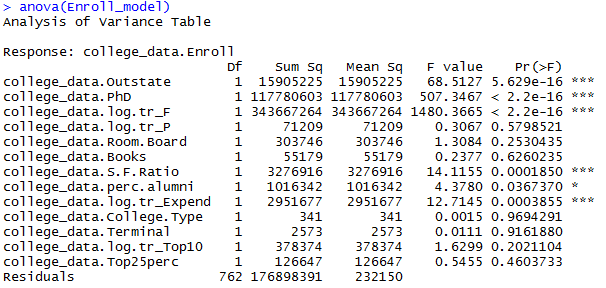
Top10perc and Expend variables are not normally distributed, so they were transformed to normality using Log transformation.



**Equation of Regression Line**

Grad.Rate = 84.931097 + 0.001 \* college\_data.Outstate + 0.022 \*college\_data.PhD +3.15 \*college\_data.log.tr\_F -1.322 \*college\_data.log.tr\_P + 0.0021 \*college\_data.Room.Board +2.194 \* college\_data.Top10perc + 0.1631 \*Top25perc - 6.863\*college\_data.College.TypePublic -8.538\*college\_data.log.tr\_Expend

Enroll: Created Multi variate linear regression model for Enrollment



By Looking at the Variance Table (p-value), we can see that outstate,PhD,F.Undergrad,S.F.Ratio and Expend are significant variables for enrollment with p value of less than 0.05. Ignoring variable alumni as doesn’t make sense.

Regression Line for Enroll\_model

Enroll = -6.966 + 7.672 \* Grad\_data.college\_data.log.tr\_F -8.934 \*college\_data.S.F.Ratio -2.72

**Conclusion:**

1. At the beginning, we expected Grad rate will be much impacted by F.Undergrad, P.Undergrad and Top25 variables. However, our study shows that variables Outstate,PhD,Room.Board, Top10,CollegeType and expenditure also impact.
2. We expected student Faculty ratio will influence grad rate of a college, but the analysis shows that it does not have any impact on Grad rate.
3. Though Linear modeling proves variables Outstate , Room. Board expend have significance in affecting Grad rate which doesn’t make sense.
4. Initially, we expected Enrollment will be explained by College Type , Outstate tuitionfee, PhD, F. Undergrad. Our study shows these variables explains the significance of Enrollment along with variables Expenditure and S.F.Ratio.
5. We have also knn model to predict college type based on college data. The model accuracy was 90%.