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.

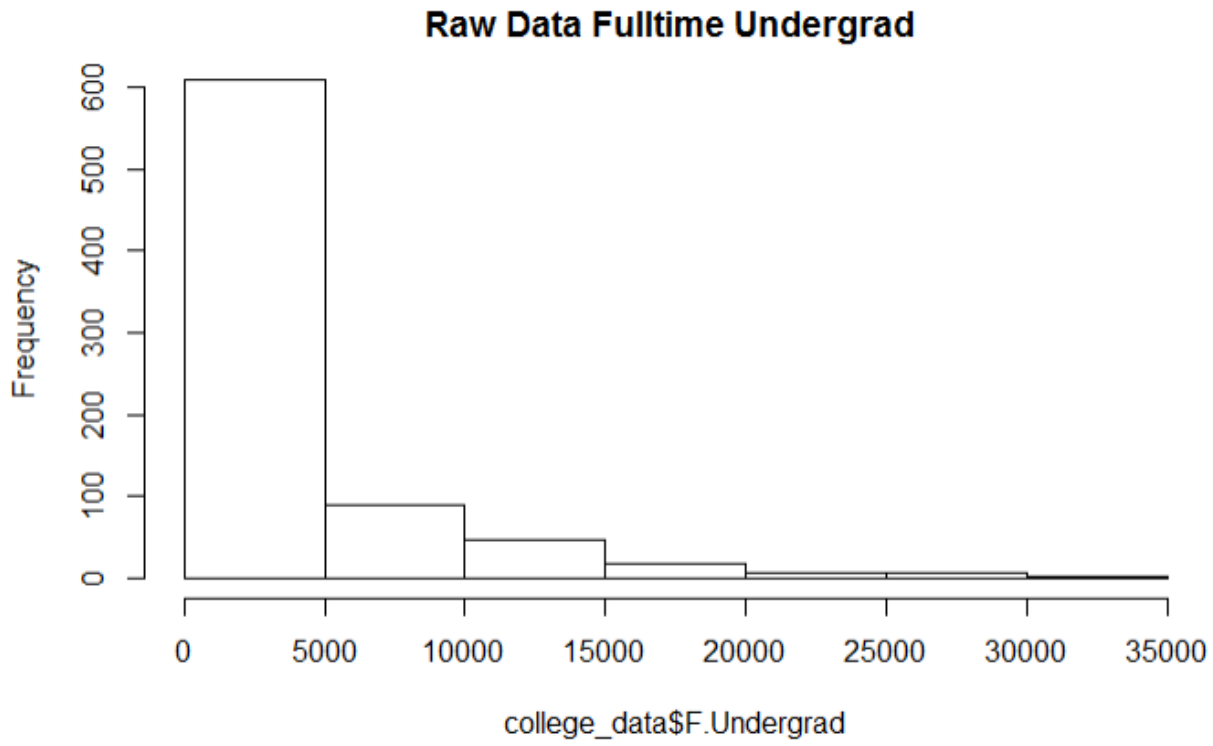
Key Exploratory Findings observed in Phase 1:

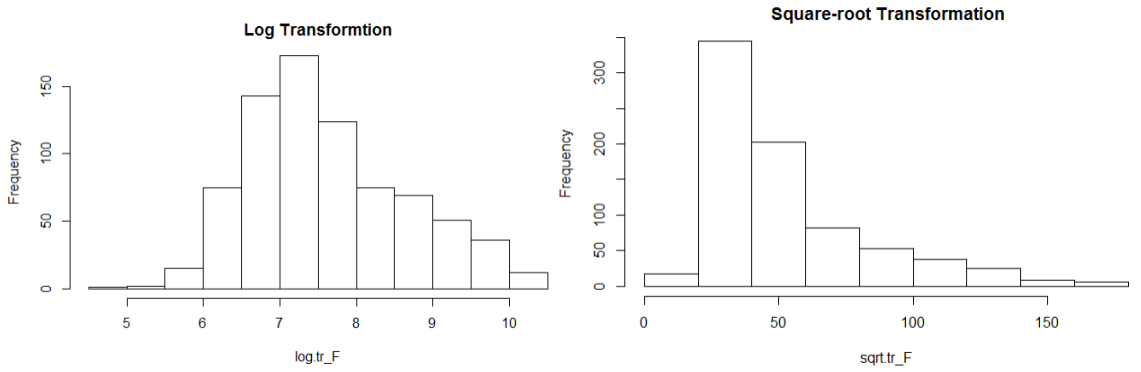
1. Variables F. Undergrad and P. Undergrad are positively skewed.
2. Variables Top25perc, OutState, PhD, S.F. Ratio are symmetric.
3. Using box plot we inferred variables F. Undergrad and P. Undergrad have extreme outliers.
4. Using XY plot we observed that student enrollment increases when no. of PhD 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.

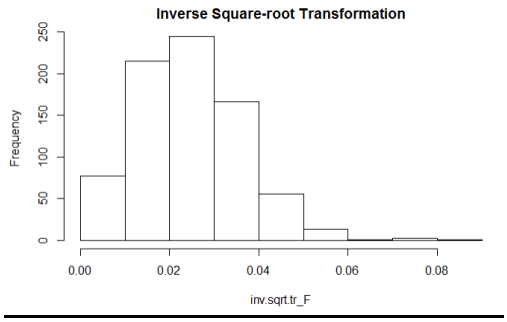
Analysis Results:

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,

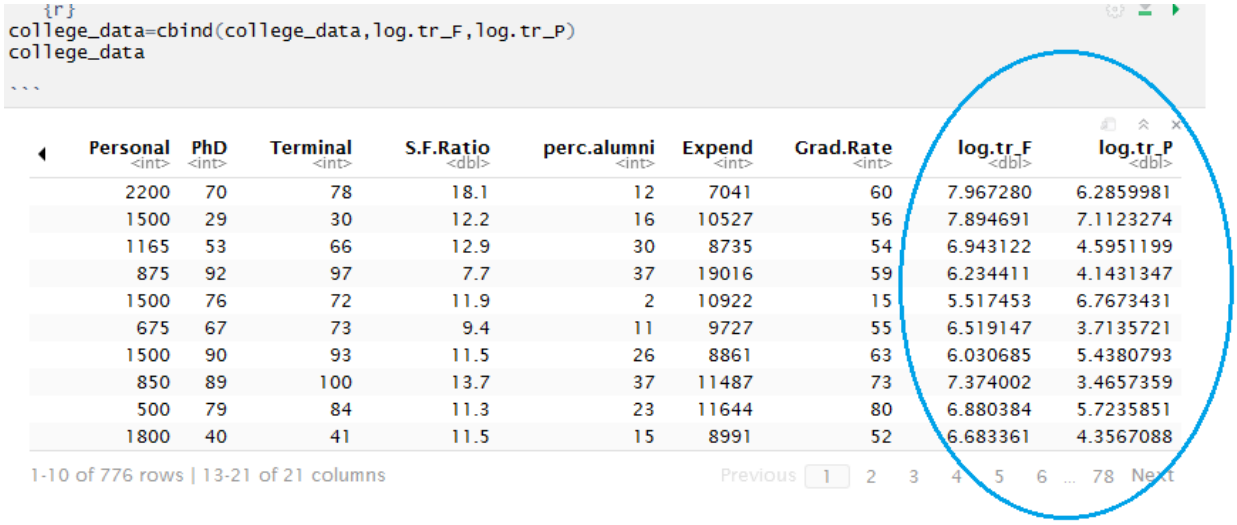
1. Log transformation
2. Square root transformation
3. Reverse square root transformation.







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.



```
{r}
college_data=cbind(college_data,log.tr_F,log.tr_P)
college_data
```

Personal <int>	PhD <int>	Terminal <int>	S.F.Ratio <dbl>	perc.alumni <int>	Expend <int>	Grad.Rate <int>	log.tr_F <dbl>	log.tr_P <dbl>
2200	70	78	18.1	12	7041	60	7.967280	6.2859981
1500	29	30	12.2	16	10527	56	7.894691	7.1123274
1165	53	66	12.9	30	8735	54	6.943122	4.5951199
875	92	97	7.7	37	19016	59	6.234411	4.1431347
1500	76	72	11.9	2	10922	15	5.517453	6.7673431
675	67	73	9.4	11	9727	55	6.519147	3.7135721
1500	90	93	11.5	26	8861	63	6.030685	5.4380793
850	89	100	13.7	37	11487	73	7.374002	3.4657359
500	79	84	11.3	23	11644	80	6.880384	5.7235851
1800	40	41	11.5	15	8991	52	6.683361	4.3567088

1-10 of 776 rows | 13-21 of 21 columns

Previous 1 2 3 4 5 6 ... 78 Next

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,

```
> anova(Grad_model)
Analysis of Variance Table

Response: college_data.Grad.Rate
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
college_data.Outstate      1   74596   74596 450.9646 < 2.2e-16 ***
college_data.PhD           1    2063    2063  12.4735 0.0004376 ***
college_data.log.tr_F      1     788     788   4.7665 0.0293228 *
college_data.log.tr_P      1    2159    2159  13.0502 0.0003231 ***
college_data.Room.Board    1    1637    1637   9.8964 0.0017204 **
college_data.Books         1     536     536   3.2410 0.0722112 .
college_data.S.F.Ratio     1      46      46   0.2768 0.5989816
college_data.perc.alumni   1   10013   10013  60.5315 2.356e-14 ***
college_data.log.tr_Expend 1    2029    2029  12.2658 0.0004882 ***
college_data.College.Type  1    3198    3198  19.3332 1.254e-05 ***
college_data.Terminal      1     260     260   1.5729 0.2101690
college_data.log.tr_Top10  1    4008    4008  24.2312 1.048e-06 ***
college_data.Top25perc     1    1179    1179   7.1267 0.0077562 **
Residuals                762  126046     165
```

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.

```
> anova(Grad_model)
Analysis of Variance Table

Response: college_data.Grad.Rate
                           Df Sum Sq Mean Sq  F value    Pr(>F)
college_data.Outstate       1  74596   74596 427.1899 < 2.2e-16 ***
college_data.PhD            1   2063    2063  11.8159 0.0006190 ***
college_data.log.tr_F       1    788     788   4.5152 0.0339139 *
college_data.log.tr_P       1   2159    2159  12.3622 0.0004639 ***
college_data.Room.Board     1   1637    1637   9.3746 0.0022770 **
college_data.log.tr_Top10   1   6944    6944  39.7656 4.835e-10 ***
college_data.Top25perc      1    920     920   5.2663 0.0220116 *
college_data.College.Type   1   2501    2501  14.3214 0.0001661 ***
college_data.log.tr_Expend  1   3191    3191  18.2746 2.154e-05 ***
```

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

```
> anova(Enroll_model)
Analysis of Variance Table

Response: college_data.Enroll
                           Df     Sum Sq    Mean Sq   F value    Pr(>F)
college_data.Outstate       1  15905225  15905225   68.5127 5.629e-16 ***
college_data.PhD            1 117780603 117780603  507.3467 < 2.2e-16 ***
college_data.log.tr_F       1 343667264 343667264 1480.3665 < 2.2e-16 ***
college_data.log.tr_P       1     71209     71209    0.3067 0.5798521
college_data.Room.Board     1    303746    303746    1.3084 0.2530435
college_data.Books          1     55179     55179    0.2377 0.6260235
college_data.S.F.Ratio      1   3276916   3276916   14.1155 0.0001850 ***
college_data.perc.alumni    1   1016342   1016342    4.3780 0.0367370 *
college_data.log.tr_Expend  1   2951677   2951677   12.7145 0.0003855 ***
college_data.College.Type   1       341       341    0.0015 0.9694291
college_data.Terminal       1      2573      2573    0.0111 0.9161880
college_data.log.tr_Top10   1    378374    378374    1.6299 0.2021104
college_data.Top25perc      1    126647    126647    0.5455 0.4603733
Residuals                 762 176898391    232150
```

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 impacted only by F.Undergrad, P.Undergrad and Top25 variables. However, our study shows that variables Outstate, PhD ,Room. Board, Top10,CollegeType and expenditure also affect Grad rate.
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. Linear modeling proves variables Outstate , Room. Board expend have significance in affecting Grad rate. However, it does not make sense.
4. Initially, we expected Enrollment will be explained by College Type , Outstate tuition fee, PhD, F. Undergrad. Surprisingly, our study shows these variables explains the significance of Enrollment along with variables Expenditure and S.F.Ratio.
5. We have also used knn model to predict college type based on college data. The model accuracy was 90%.