Advance Statistics Project Report

Analysis and Prediction of Graduate School Admission using R

- Can I get admission in University of my choice?

— Group Members

21BDA38 - Liliya Ponnu Shaji
21BDA56 – Harshitha S A
21BDA59 - Purva Singh

CONTENTS

Introduction

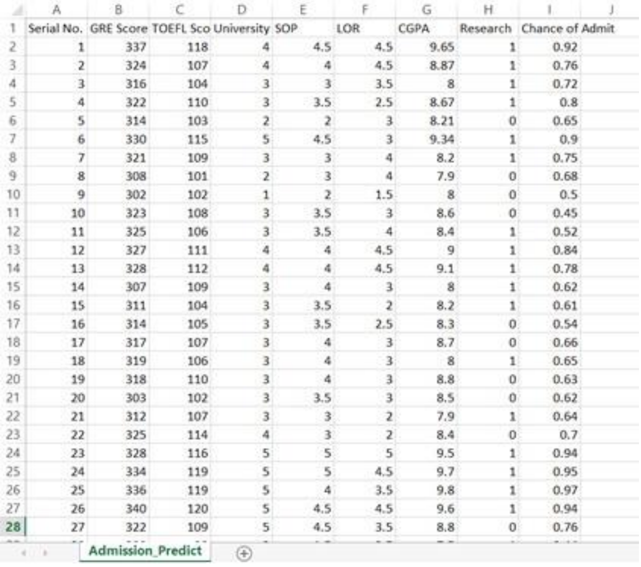
While applying to universities abroad, there are many factors which determine the student selection in respective universities. Some students would have their own choice of university for higher studies. When applying to graduate schools, they are eager to know if they are eligible with their GRE scores, CGPA, Recommendation letter etc. Basicaly the selection is based on the candidate's overall profile.

Problem Statement

Project aims to analyse the secondary data of graduates to estimate the chances of graduate's admission based on several academic performance measurements.

About Dataset

The dataset has been collected from Kaggle. There are 400 rows along with 8 columns.



	A	B	C	D	E	F	G	H	I	J
1	Serial No.	GRE Score	TOEFL Sco	University	SOP	LOR	CGPA	Research	Chance of Admit	
2	1	337	118	4	4.5	4.5	9.65	1	0.92	
3	2	324	107	4	4	4.5	8.87	1	0.76	
4	3	316	104	3	3	3.5	8	1	0.72	
5	4	322	110	3	3.5	2.5	8.67	1	0.8	
6	5	314	103	2	2	3	8.21	0	0.65	
7	6	330	115	5	4.5	3	9.34	1	0.9	
8	7	321	109	3	3	4	8.2	1	0.75	
9	8	308	101	2	3	4	7.9	0	0.68	
10	9	302	102	1	2	1.5	8	0	0.5	
11	10	323	108	3	3.5	3	8.6	0	0.45	
12	11	325	106	3	3.5	4	8.4	1	0.52	
13	12	327	111	4	4	4.5	9	1	0.84	
14	13	328	112	4	4	4.5	9.1	1	0.78	
15	14	307	109	3	4	3	8	1	0.62	
16	15	311	104	3	3.5	2	8.2	1	0.61	
17	16	314	105	3	3.5	2.5	8.3	0	0.54	
18	17	317	107	3	4	3	8.7	0	0.66	
19	18	319	106	3	4	3	8	1	0.65	
20	19	318	110	3	4	3	8.8	0	0.63	
21	20	303	102	3	3.5	3	8.5	0	0.62	
22	21	312	107	3	3	2	7.9	1	0.64	
23	22	325	114	4	3	2	8.4	0	0.7	
24	23	328	116	5	5	5	9.5	1	0.94	
25	24	334	119	5	5	4.5	9.7	1	0.95	
26	25	336	119	5	4	3.5	9.8	1	0.97	
27	26	340	120	5	4.5	4.5	9.6	1	0.94	
28	27	322	109	5	4.5	3.5	8.8	0	0.76	

The dataset contains several parameters which are considered important for Masters Program. The parameters included are:

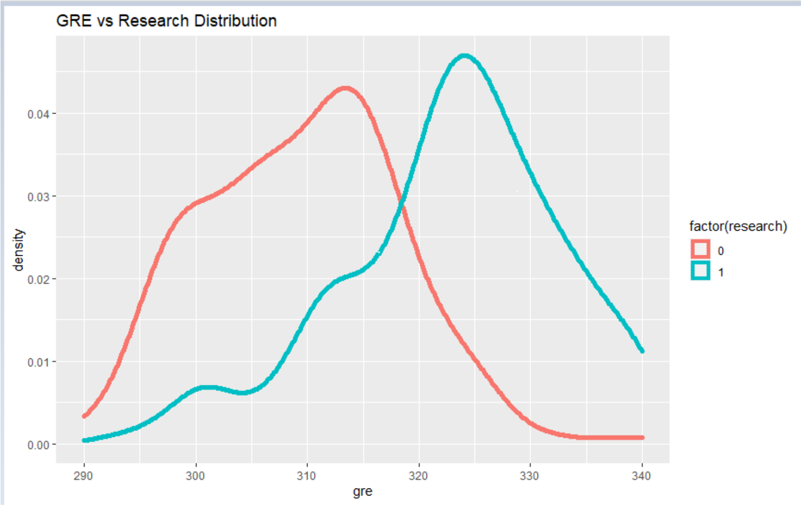
- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose (out of 5)
- Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)
- Research Experience (either 0 or 1)
- Chance of Admit (ranging from 0 to 1)

EXPLORATORY DATA ANALYSIS

We have performed EDA on various variables. Some of our inferences are:

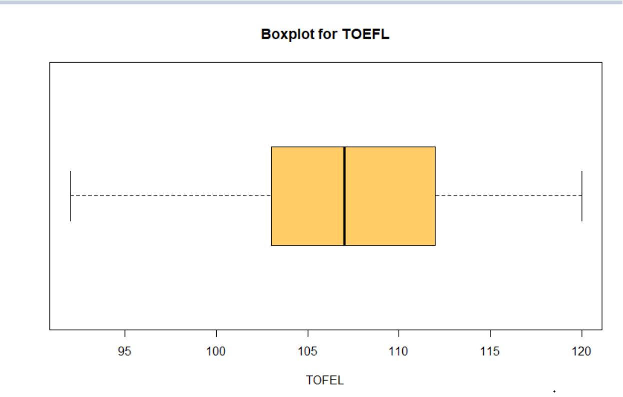
1. **GRE Vs Research Experience-**

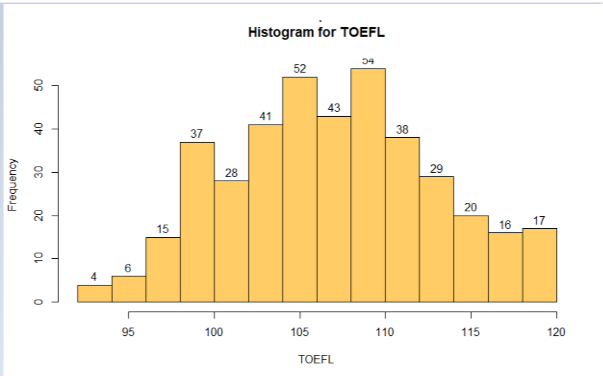
From density plot of GRE Vs Research Experience. We can conclude that students with research experience are more likely to have a higher GRE score.



2) Analysis of TOEFL

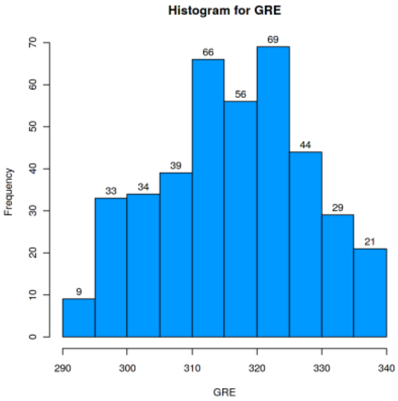
From box plot and histogram of TOEFL, we can say that the median is around 107. Also, one should reach at least 112 if he/she wants to be in the top 25%.

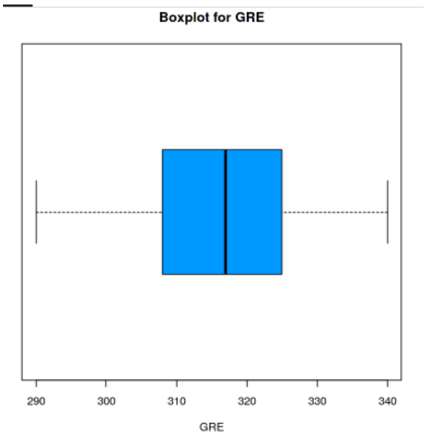




3) Analysis of GRE Score

We can see from the plots that the median of GRE is around 318. Also, one should reach 325 if he/she wants to be in the top 25%.





Hypothesis Testing 1-Sample

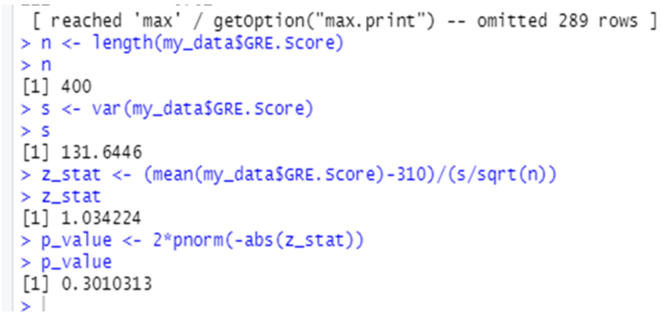
1. **Variable - GRE Score**

Note - The Test Statistic that we are going to use here is z-test because the variance of the dataset is unknown and the sample is more than 30.

We have determined if there is evidence with a significance level of 0.05 to support our claim that the mean GRE Score is different from 310.

Hypothesis statement:

Null hypothesis: μ = 310 (Population mean of GRE Score is equal to 310)
Alternate hypothesis: μ ≠ 310 (Population mean of GRE Score is not equal to/different from 310)
Level of significance = 0.05, z score for 0.05 = 1.96



```
 [ reached 'max' / getOption("max.print") -- omitted 289 rows ]
> n <- length(my_data$GRE.Score)
> n
[1] 400
> s <- var(my_data$GRE.Score)
> s
[1] 131.6446
> z_stat <- (mean(my_data$GRE.Score)-310)/(s/sqrt(n))
> z_stat
[1] 1.034224
> p_value <- 2*pnorm(-abs(z_stat))
> p_value
[1] 0.3010313
> 
```

Inference -

1. Test statistic = 1.034 is greater than -1.96 and less than +1.96. Therefore, we accept the null hypothesis.
2. p-value = 0.301 is greater than level of significance = 0.05. Therefore, we accept the null hypothesis.
3. **We can conclude that Population mean of GRE Score is equal to 310**

Hypothesis Testing 2-Sample

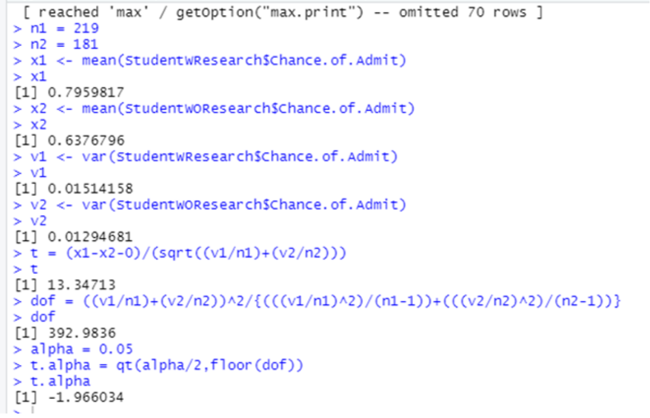
2) **Variable – Research Experience**

Note - The Test Statistic that we are going to use here is t-test because we are comparing the means of two groups.

To determine if there is any difference between the mean of chance of admission of students with research experience and mean of chance of admission of students without research experience. We have split the dataset into 2 groups based on the value of the Research column which is students with research experience and students without research experience. The variances of both groups are assumed to be not equal.

Hypothesis statements -

Null Hypothesis: μ1 = μ2 (sample mean of chance of admission of students with research experience is equal to sample mean of chance of admission of students without research experience)
Alternate Hypothesis: μ1 ≠ μ2 (sample mean of chance of admission of students with research experience differs from sample mean of chance of admission of students without research experience)
Level of Significance – 0.05



```
 [ reached 'max' / getOption("max.print") -- omitted 70 rows ]
> n1 = 219
> n2 = 181
> x1 <- mean(StudentwResearch$Chance.of.Admit)
> x1
[1] 0.7959817
> x2 <- mean(StudentwOResearch$Chance.of.Admit)
> x2
[1] 0.6376796
> v1 <- var(StudentwResearch$Chance.of.Admit)
> v1
[1] 0.01514158
> v2 <- var(StudentwOResearch$Chance.of.Admit)
> v2
[1] 0.01294681
> t = (x1-x2-0)/(sqrt((v1/n1)+(v2/n2)))
> t
[1] 13.34713
> dof = ((v1/n1)+(v2/n2))^2/{(((v1/n1)^2)/(n1-1))+(((v2/n2)^2)/(n2-1))}
> dof
[1] 392.9836
> alpha = 0.05
> t.alpha = qt(alpha/2,floor(dof))
> t.alpha
[1] -1.966034
```

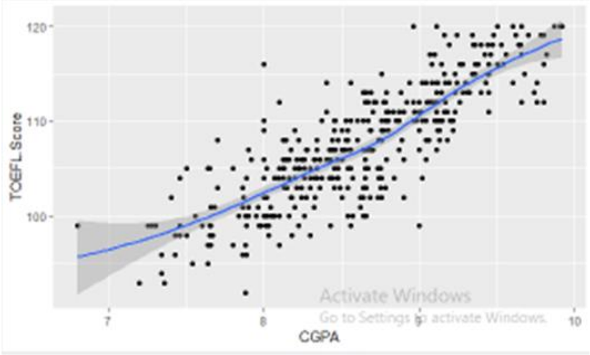
Inferences -

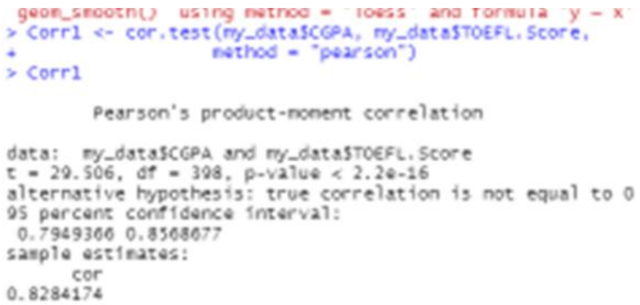
1) Since t = 13.35 is greater than t alpha = 1.96, we reject the null hypothesis.
2) **We can conclude that sample mean of chance of admission of students with research experience differs from sample mean of chance of admission of students without research experience**

Correlation Analysis

The type of correlation coefficients that we've used here is Pearson’s product moment correlation coefficient because we're considering two quantitative variables to see if there's a linear relationship between them

1) **Variables** - **CGPA and TOEFL Score**





```
 geom_smooth() using method = 'loess' and formula 'y ~ x'
> Corr1 <- cor.test(my_data$CGPA, my_data$TOEFL.Score,
+                   method = "pearson")
> Corr1

	Pearson's product-moment correlation

data:  my_data$CGPA and my_data$TOEFL.Score
t = 29.506, df = 398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7949366 0.8568677
sample estimates:
      cor
0.8284174
```

Inference: It can be seen that CGPA increases as the TOEFL Score increases. Since r = 0.828, it shows a strong positive linear relationship. A scatter plot and correlation analysis of the data indicates that there is a relatively strong positive linear association between CGPA and TOEFL Score.

Significance Test for Correlation

To provide more evidence to the fact that there is a linear relationship between TOEFL Score and CGPA, significance test has been conducted at significance level 0.05

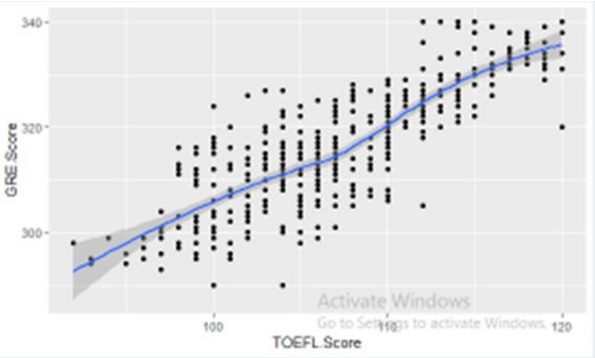
Hypothesis statement

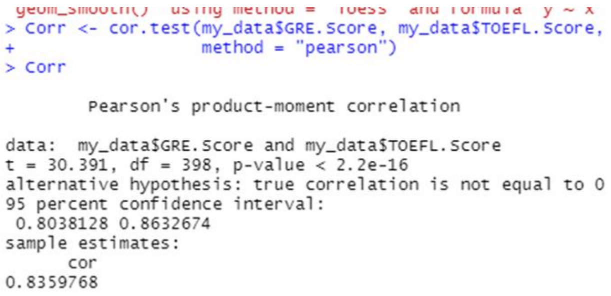
Null hypothesis: ρ = 0 (There is no linear correlation between CGPA and TOEFL Score)
Alternate hypothesis: ρ ≠ 0 (There exists linear correlation between CGPA and TOEFL Score)

Inferences -

1. test statistic, t = 29.506 and p-value = 2.2*10^-16. Since the p-value is less than the significance level of 0.05, the null hypothesis is rejected.
2. **Therefore, we can conclude that there exists linear correlation between CGPA and TOEFL Score.**

2) Variables – TOEFL Score and GRE





```
 geom_smooth() using method = 'loess' and formula 'y ~ x'
> Corr <- cor.test(my_data$GRE.Score, my_data$TOEFL.Score,
+                  method = "pearson")
> Corr

	Pearson's product-moment correlation

data:  my_data$GRE.Score and my_data$TOEFL.Score
t = 30.391, df = 398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8038128 0.8632674
sample estimates:
      cor
0.8359768
```

Inference - It can be seen that TOEFL Score increases as the GRE Score increases. Since r = 0.836, it shows a strong positive linear relationship. A scatter plot and correlation analysis of the data indicates that there is relatively strong positive linear association between TOEFL Score and GRE Score.

Significance Test for Correlation

To provide more evidence in support of the fact that there is a linear relationship between TOEFL and GRE, a significance test is conducted at a significance level of 0.05.

Hypothesis statement

Null hypothesis: ρ = 0 (There is no linear correlation between TOEFL Score and GRE Score)
Alternate hypothesis: ρ ≠ 0 (There exists linear correlation between TOEFL Score and GRE Score)

Inference-

1. test statistic, t = 30.391 and p-value = 2.2*10^-16 the p-value is less than the significance level of 0.05. Therefore the null hypothesis is rejected.

2 There is sufficient evidence of linear relationship between TOEFL Score and GRE Score of students

Conclusion for Correlation Analysis –

From the correlation analyses that we've done, a strong positive correlation between student’s CGPA and student’s GRE Score are found with variable TOEFL Score.

Regression Analysis

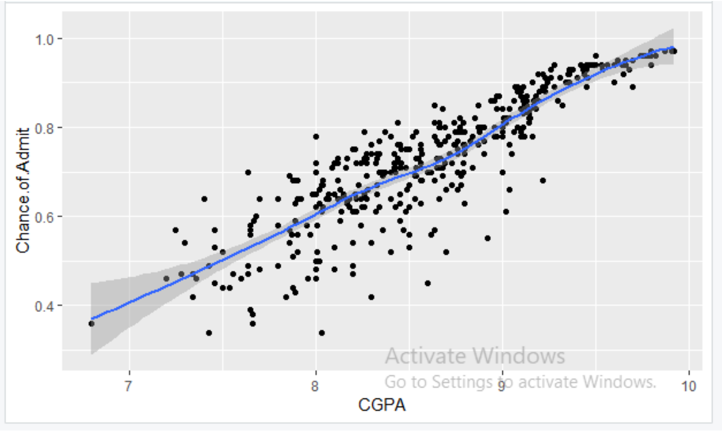
It is important to describe the impact of one independent variable on the other dependent variable, therefore, regression analysis has also been conducted.

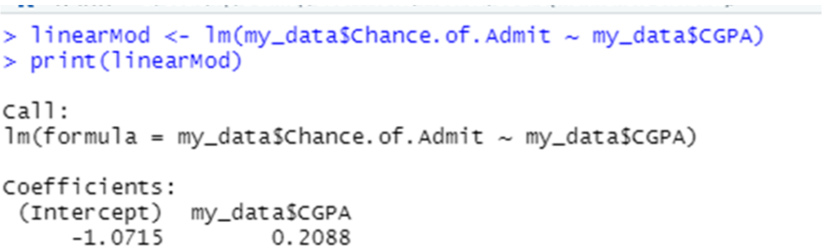
Variables- Chances of Admission and CGPA

Let us consider the dependent variable(y) in this case is the collected variable Chance of Admission and the independent variable(x) is the variable CGPA. This analysis aims to test the existence of a linear relationship between the variable x and y.

1) Based on the graph generated, the regression model involves a single independent variable and it is called simple regression.

2)The regression model is a positive linear relationship with an equation of **y = -1.0715 + 0.2088x**





```
> linearMod <- lm(my_data$Chance.of.Admit ~ my_data$CGPA)
> print(linearMod)

Call:
lm(formula = my_data$Chance.of.Admit ~ my_data$CGPA)

Coefficients:
 (Intercept)  my_data$CGPA  
     -1.0715        0.2088  
```

Interpretation of Intersection Coefficient(b0)

The value of intersection coefficient(b0) = -1.0715 indicates the estimated average value of Chances of Admission when the value of CGPA is 0.

Inference-

Here, as no students would get 0 for CGPA unless they did not participate in any compulsory academic activities, it is safe to say that b0 is just a value of the chance of admit that is not explained by CGPA.

Interpretation of Slope Coefficient(b1)

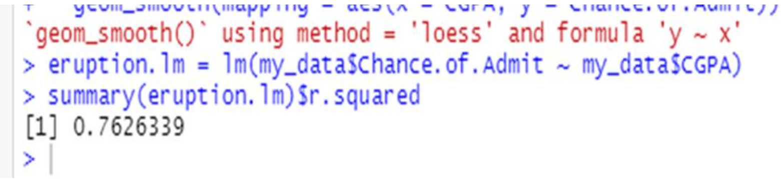
The value of slope coefficient(b1) = +0.2088 indicates the average value of Chance of Admission as a result of a one-unit change in CGPA.

Inference -

In this case, Chances of Admission will increase by 0.2088 on average for each additional one-unit change of CGPA. Therefore, we can say that there is a relationship between the CGPA and Chances of Admission variables based on the calculated value of b1.

Therefore, we can say that there is a relationship between the CGPA and Chances of Admission variables based on calculated value of b1

Coefficient of determination



```
+   geom_smooth(mapping = aes(x = CGPA, y = Chance.of.Admit))
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
> eruption.lm = lm(my_data$Chance.of.Admit ~ my_data$CGPA)
> summary(eruption.lm)$r.squared
[1] 0.7626339
> |
```

Coefficient of determination, R squared = 0.7626339. Since the R square = 0.7626339 which is between 0 and 1, it can be concluded that there exists a weaker linear relationship between CGPA and Chances of Admission

Inference-

Some but not all of the variation in Chances of Admission is explained by variation in CGPA. In this case, around 76% of the variation in Chances of Admission explained by variation in CGPA

Goodness of Fit

Goodness of fit test is used to test the hypothesis that an observed frequency distribution, in this case the students with and without research experience fits some claimed distribution. A claim that the students with and without research experience has equal proportions is tested at 5% level of significance.

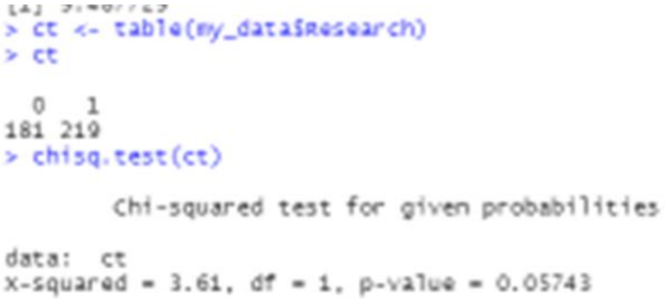
Let's assume P1 as Students with research experience and P2 as Students without research experience.

Hypothesis Statements-

Null Hypothesis: P1 = P2 = 0.5(Frequency of students with and without research experience has the same proportion which is equal to probability of 0.5)

Alternate Hypothesis: At least one of the proportions is different from the others

Level of Significance = 0.05 and Chi-Square for 0.5 Significance Level and 1 degree of freedom = 3.8415



```
> ct <- table(my_data$Research)
> ct

  0   1 
181 219 
> chisq.test(ct)

	Chi-squared test for given probabilities

data:  ct
X-squared = 3.61, df = 1, p-value = 0.05743
```

Conclusions -

1) Since test statistic = 3.61 < critical value = 3.8415, we accept the null hypothesis.

2) Another way to decide whether to reject or fail to reject null hypothesis is by using the p-value method. In this case, since the p-value = 0.05743 > significance level = 0.05, we fail to reject the null hypothesis.

3)Therefore we can conclude that Frequency of students with and without research experience has the same proportion

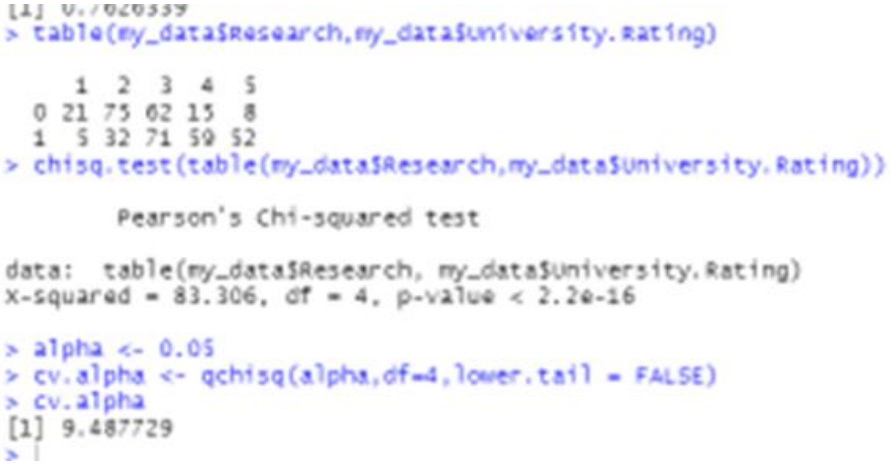
Chi-Square Test for Independence

This analysis is used to find out if there is any relationship between variable Research and variable University Rating.

Hypothesis Statements :-

Null hypothesis = Variables Research and University Rating are independent.

Alternate hypothesis = Variables Research and University Rating are related to each other.



```
[1] 0.7626339
> table(my_data$Research,my_data$University.Rating)

     1  2  3  4  5
  0 21 75 62 15  8
  1  5 32 71 59 52
> chisq.test(table(my_data$Research,my_data$University.Rating))

        Pearson's Chi-squared test

data:  table(my_data$Research, my_data$University.Rating)
X-squared = 83.306, df = 4, p-value < 2.2e-16

> alpha <- 0.05
> cv.alpha <- qchisq(alpha,df=4,lower.tail = FALSE)
> cv.alpha
[1] 9.487729
> |
```

Conclusions-

1) Since the test statistic = 83.306 , is greater than critical value = 9.477, the null hypothesis is rejected at 0.05 significance level.

2) There is significant evidence to conclude that there exists a relationship between the variables Research and University Rating. Thus, we can conclude that a student's research experience and the university rating of student's choice are related to each other.

MODELING

Data modelling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database. One of the goals of data modelling is **to create the most efficient method of storing information while still providing for complete access and reporting**.

For modelling, first we have divided the dataset into test data and train data. 80% of the original data is taken for the training part and the rest 20% is used for testing the data.

For admission prediction, we tried done 3 different models.

1. Random Forest Model
2. Multiple Linear Regression
3. XG Boost

While evaluating any model, there are two most important factors which are taken into consideration. They are-

1) **R SQUARE** - The most common interpretation of the coefficient of determination is how well the regression model fits the observed data. Generally, a higher coefficient indicates a better fit for the model.

2) **RMSE** - It can be interpreted as the standard deviation of the unexplained variance. It has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response.

Random Forest Model

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing *continuous variables* as in the case of regression and *categorical variables* as in the case of classification

Conclusion:

The random forest model got a 0.068 RMSE and a 0.750 R^2 which is pretty good

Multiple Linear Regression Model

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Since we have multiple independent variables which are used for finding the target variable, we thought of using a Multiple linear regression model. It is used to estimate the relationship between two or more independent variables and one dependent variable.

Conclusion:

The multiple linear regression model got a 0.068 RMSE and a 0.786 R^2 which is even better

XG Boost

XG Boost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Conclusion:

The XG Boost model got a 0.042 RMSE and a 0.86 R^2 which is perfect.

Final Conclusions –

1) The number of students with research experience has higher chance of admission to a university in average compared to students without research experience.
2) GRE Score , TOEFL and CGPA are one of the important criteria in student's academic performance to get a better chance of admission to a university.
3) Academic performance measurements has a strong relationship between one another. One academic performance measurement can influence the growth of the other one
4) One academic performance measurement can influence the growth of the other one.
5) From the modelling part, we can conclude that XG Boost is the best model as compared to Random Forest and Multiple Regression.

THANKYOU