SIES COLLEGE OF ARTS, SCIENCE AND COMMERCE
(AUTONOMOUS)

DEPARTMENT OF STATISTICS

SION MUMBAI - 400 022

2021-2022

CERTIFICATE

This is to certify that the project "**EDUCATIONAL MAZE**" is carried by a group of five students during the academic year 2021-2022.

The team comprises of :

Aaliya Mohammad (TS2122101)

Bhoomika Patel (TS2122115)

Sharwari Pawar (TS2122117)

Priyanka Mandya (TS2122120)

Saranya Jayakrishnan (TS2122122)

Head of the department

Prof.Pallavi Rege

Project guide

Prof.Priyanka Dangar

EDUCATIONAL MAZE

ACKNOWLEDGEMENT

The success of this project is a result of hard work and support of our peers and teachers. We take immense pleasure in submitting this project on "**EDUCATIONAL MAZE**" . This project has helped us to understand the difficulties that students and young people face when they have to choose their career path. We are extremely thankful to our project guide Prof. Priyanka Dangar for her guidance and support. We are extremely glad to have her as a project guide. We would like to thank our Prof. Pallavi Rege, head of department for extending her help and support. We would also like to thank other professors of the department and the entire staff of department of statistics of SIES College of Arts, Science and Commerce (Autonomous) for providing us with all the infrastructure and cooperating with us at every stage of a project. A special thanks to all the respondents for providing us with the vital information, precious time and the precious support. Finally we must express our heartiest gratitude towards all those who helped us directly or indirectly in completing this project.

INDEX

INTRODUCTION

By the time a student reaches High School there is a frantic search to define, discuss and decide which career path to choose. Students clearing boards face a great dilemma and the biggest question hounding them is "Which career is right for me?" Or " How can I find the best career option for me?". The frantic search for the perfect career option for each of the students is an uphill struggle as the modern information age has opened up uncountable possibilities in all fields. The digital age has brought a paradigm shift in the way an individual perceives a career and work options. From Millennials to Gen Z the fixation for traditional career options like engineers, doctors, CA's, lawyers has fizzled out and unconventional skill oriented jobs and business have captured everyone's imagination. More students are perplexed because they have taken up their career planning quite late.

A typical career path in the past involved almost guaranteeing job security. You obtained a job for life and you could expect steady progress progression up and organisation structure but factors like technological growth and others have changed the world . Career development is now a lifelong process rather than a lifelong job and then was moving laterally upwards within the organisation or between organisations .

These days you are responsible for your own career and you have to continuously reflect on 'where you are' and 'where you want to be' and 'how you will' prepare for the career opportunities . It is important to make sure that your career decisions are suggested by a realistic assessment of the information available rather than relying on examinations that you or others take up. Some of these assumptions may have gained credibility over time especially if they have been reinforced by the media and the people whose opinion you value. The idea of a career for life changes as each person is likely to have several careers in their lifetime. There are no test inventories of career assessment instruments that will tell you your perfect career. These assessment tools will help you to assemble your motivations, interest, skill and value as they are at the point in your life when you complete them.

Career decisions are influenced by a variety of factors which need to come together before you are able to make an effective career decision. You need to think where you are going with your career vision and how you will get to your career goal. We need to gather information and think about our interest value skills, career influences, study options and support we are likely to get from others.

When you choose a career , you are choosing work that you will likely become an expert in over time through education and experience. For that you will probably work in a career field for at least several years . To stop narrowing down and choosing the path to follow is one of the most important decisions you make.

IMPORTANCE OF AMBITION

Ambition is a major driver for personal growth and development. No one can succeed without a healthy dose of ambition. Ambitious people take charge of their destiny and do not expect others to bring about their needs , they have willpower and determination & they know where they are going and what they have to do to get there. Those who wish to be no more to more give more have more and have purpose and powerful internal drive that leads them to dream bigger and go for the. Ambition drives them to advance and accomplish their goals well . Ambition reflects a healthy self esteem and higher power of abstraction and visualisation of the future.

CHANGE IN CHILDHOOD AMBITION WITH TIME

As children we always dreamt of becoming Astronauts, Rock-Star, Police and much more .Our dreams and aspirations used to change as soon as we started developing an interest in one ambition or sometimes because of a relative whom we idolise very much and other times it's because of role models . During childhood there are no set of dreams and ambitions that we could really follow and be called as the final goal .On the other hand, Sometimes the dreams we have are not feasible and with time we get to know what we actually are capable of . Parents guide them from a young age about what should be the ideal job and if things are on the right path from the starting there will be no need to change the childhood dream or ambition. There are multiple factors which influence the change of ambition in a child and this project also focuses on them.

FACTORS INFLUENCING OUR AMBITION

1) family
2) friends
3) finance
4) teachers and counsellors
5) media
6) role models

NEED FOR THE STUDY

Education plays a key role in a person's success.It provides stability, financial-security, equality and self dependency. It also makes our dreams and ambitions come true.

SCOPE OF THE STUDY

A lot of people believe we are actually happier when our mind is occupied with meaningful things. Scope of this study is to analyse why people change or didn't change their ambition and depending on what factors they choose to do so & also toIdentify and analyse what are their expectations for their dream job

OBJECTIVES OF THE STUDY

- Our main objective is to know the current satisfaction of people in their chosen career/stream.
- Why do people change their ambition ?
- What is the most desired ambition among people ?
- What factors influence the most for change in ambition ?
- What are the primary required expectations while accepting the job ?

RESEARCH METHODOLOGY

A research methodology involves specific techniques that are adopted in the research process to collect, assemble and evaluate data. It defines those tools that are used to gather relevant information in a specific research study. The term ‘research’ refers to the systematic method of enunciating the problem, formulating a hypothesis, and collecting the facts or data. Analyzing the facts or data, analyzing the facts and reaching certain conclusions either in the form of solutions toward the concerned problems or in certain generalisation for some theoretical formulation.

RESEARCH INSTRUMENT

The instrument used in the study is a structured questionnaire. A questionnaire is a Google form containing questions related to the topic. The questionnaire is given to the respondents to be filled in.

DATA COLLECTION

Data refers to information or facts. Often researchers understand data as only numerical figures. It also includes descriptive facts, on numerical information, qualitative and quantitative information. Collection of data in an important stage in research. In fact the quality of the data collected determines the quantity of the research. Collection of data is done by 2 methods :-

- Primary Data Collection - It is also known as the data collected for the first time through the field survey. Such data were collected with a set of objectives to assess the current of any variable studied.
- Secondary Data Collection

Our data collection method is **Primary Data Collection.**

TARGET POPULATION

The target population is the total group of individuals from which the sample might be drawn. Target Population :- Students(from the age of 18 to 24) . In Mumbai city, Mumbai suburbs, from Thane to Karjat and Navi-Mumbai to Panvel

SAMPLE SIZE

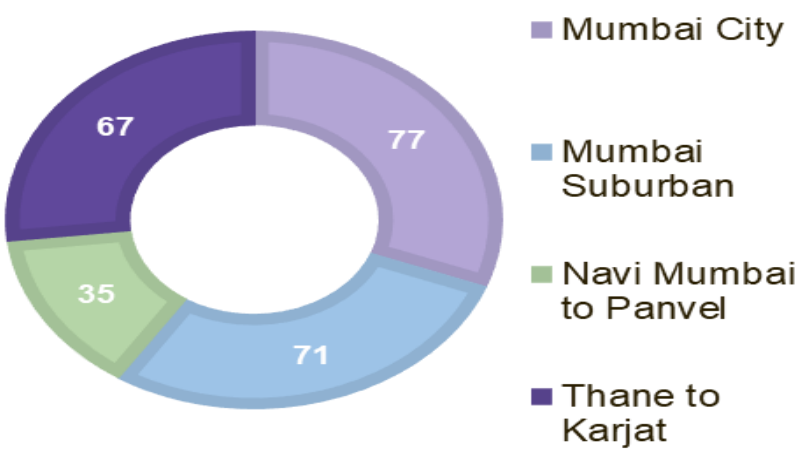
It refers to the number of items selected from the universe to constitute a sample.
Sample Size :- 250 after data cleaning
Our sample size for this project is obtained by Cochran's formula
Cochran's formula : $(Z^2 * p * q)/\alpha^2$
where , $Z=1.96$ $p=0.795$ $q=0.205$ $\alpha=0.05$
which comes out to be approximately 250.

SAMPLING METHOD

CONVENIENCE SAMPLING UNDER SRS is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

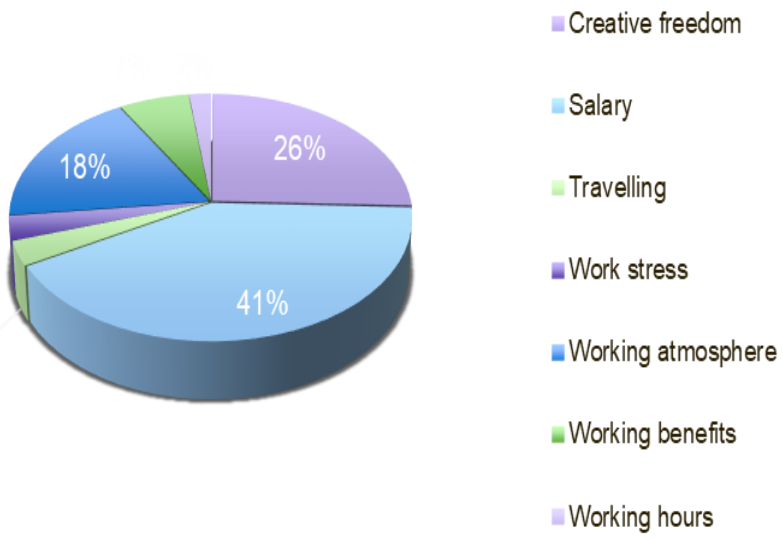
GRAPHICAL REPRESENTATION

1) LOCATION WISE DISTRIBUTION

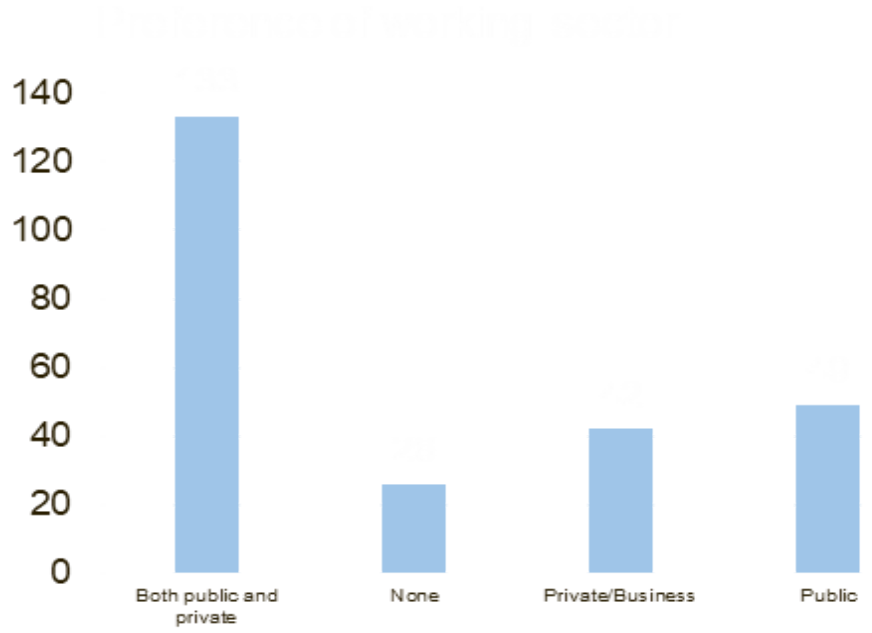


\

2) FIRST JOB EXPECTATION



3) PREFERENCE IN WORKING SECTORS



ONE SAMPLE Z PROPORTION TEST

The One Sample Proportion Test is used to estimate the proportion of a population. It compares the proportion to a target or reference value and also calculates a range of values that is likely to include the population proportion. This is also called the hypothesis of inequality. If the sample size is less then binomial enumeration gives much more accurate results.

ASSUMPTIONS:

- The data are simple random values from the population
- Population follows a binomial distribution
- When both mean (np) and variance(n(1-p)) values are greater than 10, the binomial distribution can be approximated by the normal distribution

Test statistic for one sample Z proportions test

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

- z is test statistic
- $\hat{P}$ is observed proportion
- p_0 is hypothesised probability
- n is sample size

HYPOTHESIS : Ho:p=po v/s H1:p>po

Decision criteria : Reject Ho if the p-value < po

STUDENTS SWITCHING FOR THE FIRST TIME

Ho:p=0.45 V/S H1: p > 0.45

Data: 135 out of 250 p-value = 0.00258

95 % confidence interval: (0.4860055 ,1.0000000)

Decision criteria: since p-value is less than 0.45 we have enough evidence to reject Ho .

CONCLUSION : We can say that more than 45% students change their ambition

Similarly we performed the proportion test on different samples:

Sample	Data	p-value	po	Decision
Science students switching for the 1st time	107 of 200	0.009058	0.45	We can say that more than 45% of science students will change their ambitions further
Non-science students switching for the 1st time	28 of 50	0.01519	0.4	We can say that more than 40% of science students will change their ambitions further

Students from Mumbai city who are switching for the 1st time	40 of 77	0.0215	0.4	We can say that more than 40% of students from Mumbai city will change their ambitions further
Students from Mumbai suburbs who are switching for the 1st time	40 of 71	0.03585	0.45	We can say that more than 45% of students from Mumbai suburbs will change their ambitions further
Students from Navi Mumbai to Panvel who are switching for the 1st time	22 of 35	0.02537	0.45	We can say that more than 45% of students from Navi Mumbai to Panvel will change their ambitions further
Students from Thane to Karjat who are switching for the 1st time	33 of 67	0.01022	0.35	We can say that more than 35% of students from Thane to Karjat will change their ambitions further
Students switching for the 2nd time	63 of 250	0.02405	0.2	We can say that more than 20% students will change their ambitions further
Science students who are switching for the 2nd time	51 of 200	0.03172	0.2	We can say that more than 20% of science students will change their

				ambitions further
Non-science students who are switching for the 2nd time	12 of 50	0.5	0.25	We can say that more than 25% of non-science students will not change their ambitions further
Students from Mumbai city who are switching for the 2nd time	24 of 77	0.01051	0.2	We can say that more than 20% of students from Mumbai city will change their ambitions further
Students from Mumbai suburbs who are switching for the 2nd time	20 of 71	0.05792	0.2	We can say that more than 20% of students from Mumbai suburbs will change their ambitions further
Students from Navi Mumbai to Panvel who are switching for the 2nd time	7 of 35	0.5	0.2	We can say that more than 20% students from Navi Mumbai to Panve will not change their ambitions further
Students from Thane to Karjat who are switching for the 2nd time	12 of 67	0.6083	0.2	We can say that more than 20% students from Thane to Karjat will not change their ambitions further

CORRELATION

Correlation is commonly used statistical methods to investigate the relationship between two or more quantitative variables.A *correlation coefficient* is a measure of how two variables are related. Correlation helps answer the questions "do the variables covary?" and "what is the strength of the relationship between the two variables?".Correlation coefficients range from -1 to 1. When the value of the correlation coefficient is close to +/- 1, we say the two variables are *highly correlated*. As the coefficient approaches 0, the relationship between the two variables is said to be weaker. Correlation does not imply causation. Just because two variables move together does not mean that one is causing the other to do so. Often there are other factors (called "lurking" or "confounding" variables) that strongly affect the correlation.

ASSOCIATION

WHAT IS CROSS TABULATION?

Cross tabulation is a combination of two or more frequency tables arranged such that each cell in the resulting table represents an unique combination of specific values of cross tabulated attributes.

1) CHI-SQUARE TEST OF ASSOCIATION

Chi-Square test of association is used on the data that can be nominal or ordinal.The Pearson's Chi-Square is used to test the independence of the two attributes. It can be used to test the null hypothesis that the two categorical variables under study are independent of each other in which we compare the observed cell frequencies with expected cell frequencies. Expected frequencies are the number of cases that should fall in each cell if two variables under study are independent.
Test Statistic: The test statistic is a Chi-Square random variable (χ^2) defined by the following equation. The formula for Chi-Square is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where: χ^2 is the value for Chi Square. $\sum$ is the sum. O is the observed frequency. E is the expected frequency.
Decision Criteria: Reject H0 if p-value < 0.05

We performed cross tabulation to study the association between the two independent variables.

Working sector encouraged V/s Mother's qualification

H0: There is no association between Working sector encouraged and Mother's Qualification.
v/s
H1: There is an association between Working sector encouraged and Mother's Qualification.

Contingency table for the influence of Mother's qualification on the working section encouraged is as follows:

2 x 2 contingency table

	None	Public	Private	Both
Below bachelor's	72	16	23	32
Bachelors and above	61	10	19	17

Chi Square test was performed on above contingency table which gave the following output:

Chi squared value	P Value	Degrees Of Freedom	Level of Significance
2.1273	0.5464	3	0.05

Level of significance (α) : 0.05
Decision Criteria: we reject H0 if p-value < 0.05.
Since p-value < 0.05, we Reject H0.
CONCLUSION: We reject H0 and conclude that there is significant association between gender and switching careers.
CRAMER VALUE IS 0.1697

Similarly we have checked association between different combinations of variables.

Variable 1	Variable 2	p-value	Decision
Working sector	Mother's	0.5464	Accept Ho

encouraged	qualification		
	Father's qualification	0.3141	Accept Ho
Residential area	Switching in school	0.5725	Accept Ho
	Further switching	0.2429	Accept Ho
	Marks in SSC	0.8572	Accept Ho
	Marks in HSC	0.01862	Reject Ho

2) LINEAR BY LINEAR ASSOCIATION

The linear-by-linear test can be used to test the association among variables in a contingency table with ordered categories . This test or a test with a similar function is sometimes called "ordinal chi-square" test.
In R, the test can be performed by a permutation test with the *coin, survival and rcompanion* package.

An association test can also be performed on a contingency table with one ordered nominal variable and one non-ordered nominal variable.

The null hypothesis for the linear-by-linear test is that there is no association among the variables in the table. A significant *p*-value suggests that there is an association. This is similar to a chi-square test, except that the categories are ordered in nature.

<u>Residential area V/s Influence of SSC/HSC marks of the students who switched their careers on choosing their ambition</u>

H0: There is no association between Residential area and SSC/HSCmarks
v/s
H1: There is an association between Residential area and SSC/HSCmark.

Contingency table for Residential area and SSC/HSCmark of the students who switched their careers or not is as follows:

4 x 2 contingency table

	Least influenced		Neutral	Somewhat influenced	Most influenced
Mumbai City	7	2	12	12	7
Mumbai Suburban	12	6	12	4	6
Navi Mumbai to Panvel	7	5	6	3	1
Thane to Karjat	10	6	11	5	1

Linear by linear test was performed on above contingency table which gave the following output:

Z value	P Value	Level of Significance
-2.762	0.005746	0.05

Level of significance (α) : 0.05
Decision Criteria: we reject H0 if p-value < 0.05.
Since p-value < 0.05, we reject H0.
CONCLUSION: We reject H0 and conclude that there is significant association between residential area and influence of SSC/HSC marks of the students who switched their careers on choosing their ambition
Freeman.theta 0.214 epsilon.squared 0.0739

Similarly we have checked association between different combinations of variables.

VARIABLE 1	VARIABLE 2	P value	DECISION CRITERIA
RESIDENTIAL AREA	Loss of interest	0.9211	Accept Ho
	Decreasing scope of subject	0.3087	Accept Ho
	Family	0.1315	Accept Ho
	Friends	0.5459	Accept Ho
	Financial status	0.4014	Accept Ho
	Mental and physical health	0.9704	Accept Ho
	Teachers and counsellors	0.1676	Accept Ho
	Media platforms	0.04125 F 0.159 E 0.0395	Reject Ho
	Role model	0.2402	Accept Ho
	Language barrier	0.4688	Accept Ho
	Marks in SSC/HSC	0.7858	Accept Ho
	Interest in subject	0.3576	Accept Ho
	Increasing scope of the subject	0.3738	Accept Ho
	Family	0.5144	Accept Ho
	Friends	0.06824	Accept Ho

	Financial status	0.407	Accept Ho
	Mental and physical health	0.1792	Accept Ho
	Teachers and counsellors	0.3248	Accept Ho
	Media platforms	0.7586	Accept Ho
	Role models	0.2497	Accept Ho
STREAM (SCIENCE AND NON SCIENCE)	Marks in SSC/HSC	0.4555	Accept Ho
	Loss of interest	0.04242 F 0.247 E0.0317	Reject Ho
	Decreasing scope of subject	0.07402	Accept Ho
	Family	0.8182	Accept Ho
	Friends	0.08675	Accept Ho
	Financial status	0.1328	Accept Ho
	Mental and physical health	0.538	Accept Ho
	Teachers and counsellors	0.7294	Accept Ho
	Media platforms	0.04593 F 0.248 E 0.0317	Reject Ho
	Role model	0.004453 F 0.33 E 0.0569	Reject Ho
	Language barrier	0.4192	Accept Ho
FOR STUDENTS WHO DID NOT	Marks in SSC/HSC	0.7518	Accept Ho

CHANGE			
	Interest in subject	0.007069 E 0.079 F 0.349	Reject Ho
	Increasing scope of the subject	0.08959	Accept Ho
	Family	0.6671	Accept Ho
	Friends	0.1671	Accept Ho
	Financial status	0.07092	Accept Ho
	Mental and physical health	0.5157	Accept Ho
	Teachers and counsellors	0.885	Accept Ho
	Media platforms	0.8488	Accept Ho
	Role models	0.4453	Accept Ho

3) KENDALL ' S ASSOCIATION

The Kendall rank correlation coefficient, commonly referred to as Kendall's τ coefficient is a statistic used to measure the ordinal association between two ordinal variables . A τ test is a non-parametric hypothesis test for statistical dependence based on the τ coefficient.

It is a measure of rank correlation. Rcompanion

ASSUMPTIONS :

- One ordinal variable and one nominal variable, or two ordinal variables. Usually expressed as a contingency table.
- Experimental units aren't paired.

Influence of SSC/HSC marks V/s interest in the subject of the students who switched their careers while choosing their ambition

H0: There is no association between SSC/HSCmarks and interest in the subject

V/s

H1: There is an association between SSC/HSC marks and interest in the subject.

Linear by linear test was performed gave the following output:

Z value	P Value	Level of Significance	Tau value
1.2461	0.2127	0.05	0.08724147

Level of significance (α) : 0.05
Decision Criteria: we reject H0 if p-value < 0.05.
Since p-value > 0.05, we accept H0.
CONCLUSION: We accept H0 and conclude that there is no significant association between SSC/HSC marks and interest in the subject of the students who switched their careers while choosing their ambition

Similarly we have checked association between different combinations of variables.

VARIABLE 1	VARIABLE 2	P VALUE	TAU VALUE	DECISION
Marks in SSC/HSC	Decreasing scope of subject	0.2415	0.08197	Accept Ho
	Family	0.02767	0.15448	Reject Ho
	Friends	0.004377	0.20345	Reject Ho
	Financial status	0.01965	0.16304	Reject Ho
	Mental and physical health	0.4715	0.05031	Accept Ho
	Teachers and counsellors	0.05564	0.13426	Accept Ho
	Media platforms	0.008055	0.18557	Reject Ho
	Role model	0.009361	0.18252	Reject Ho
	Language	0.001291	0.23262	Reject Ho

	barrier			
Loss of interest	Decreasing scope of subject	9.922e-06	0.30794	Reject Ho
	Family	0.968	-0.00280	Accept Ho
	Friends	0.0008384	0.237428	Reject Ho
	Financial status	0.5843	0.03806903	Accept Ho
	Mental and physical health	0.003631	0.20236	Reject Ho
	Teachers and counsellors	0.00911	0.18217	Reject Ho
	Media platforms	3.279e-07	0.35610	Reject Ho
	Role model	5.668e-07	0.34985	Reject Ho
	Language barrier	2e-04	0.26769	Reject Ho
Decreasing scope of subject	Family	0.4423	0.053657	Accept Ho
	Friends	2.124e-05	0.30214	Reject Ho
	Financial status	0.3986	0.05872	Accept Ho
	Mental and physical health	0.0001998	0.2586927	Reject Ho
	Teachers and counsellors	5.662e-06	0.316954	Reject Ho
	Media platforms	3.287e-07	0.3559736	Reject Ho
	Role model	1.663e-05	0.301066	Reject Ho
	Language barrier	0.001923	0.223207	Reject Ho
Family	Friends	6.245e–6	0.32186	Reject Ho
	Financial status	0.0001302	0.266808	Reject Ho

	Mental and physical health	0.9028	-0.008518	Accept Ho
	Teachers and counsellors	0.0009271	0.23183	Reject Ho
	Media platforms	0.2613	0.078504	Accept Ho
	Role model	0.5993	0.03682	Accept Ho
	Language barrier	0.03371	0.15317	Reject Ho
Friends	Financial status	0.01205	0.17817	Reject Ho
	Mental and physical health	0.00154	0.22468	Reject Ho
	Teachers and counsellors	2.459e-08	0.39723	Reject Ho
	Media platforms	9.135e-08	0.37997	Reject Ho
	Role model	2.244e-08	0.39885	Reject Ho
	Language barrier	5.136e-06	0.33467	Reject Ho
Financial status	Mental and physical health	1.046e-05	0.30609	Reject Ho
	Teachers and counsellors	0.005192	0.19488	Reject Ho
	Media platforms	0.01955	0.16253	Reject Ho
	Role model	0.5413	0.04264	Accept Ho
	Language barrier	0.001472	0.228504	Reject Ho

Mental and physical health	Teachers and counsellors	8.939e-05	0.27314	Reject Ho

	Media platforms	4.16e-06	0.32037	Reject Ho
	Role model	0.00039	0.24711	Reject Ho
	Language barrier	0.000269	0.26171	Reject Ho
Teachers and counsellors	Media platforms	2.044e-07	0.36304	Reject Ho
	Role model	4.061e-10	0.43678	Reject Ho
	Language barrier	0.00143	0.22996	Reject Ho
Media platforms	Role model	9.681e-15	0.54175	Reject Ho
	Language barrier	1.392e-05	0.31286	Reject Ho
Role model	Language barrier	2.259e-05	0.30603	Reject Ho
FOR STUDENTS WHO DID NOT SWITCH THEIR AMBITION				
Marks in SSC/HSC	Interest in subject	0.7099	0.03078	Accept Ho
	Increasing scope of the subject	0.00719	0.21371	Reject Ho
	Family	0.008434	0.20921	Reject Ho
	Friends	0.3874	0.06810	Accept Ho
	Financial status	0.06399	0.14586	Accept Ho
	Mental and physical health	0.07855	0.1382008	Accept Ho
	Teachers and	0.003689	0.227546	Reject Ho

	counsellors			
	Media platforms	0.4039	0.065195	Accept Ho
	Role models	0.6804	-0.032395	Accept Ho
Interest in subject	Increasing scope of the subject	4.696e-06	0.38492	Reject Ho
	Family	0.422	-0.067429	Accept Ho
	Friends	0.1036	-0.1356	Accept Ho

Interest in subject	Financial status	0.04551	-0.16653	Reject Ho
	Mental and physical health	0.6994	0.032071	Accept Ho
	Teachers and counsellors	0.8813	0.012369	Accept Ho
	Media platforms	0.9075	-0.0095926	Accept Ho
	Role models	0.1659	0.11521	Accept Ho
Increasing scope of the subject	Family	0.02817	0.177113	Reject Ho
	Friends	0.1582	0.112964	Accept Ho
	Financial status	0.3176	0.0799669	Accept Ho
	Mental and physical health	0.005793	0.22024	Reject Ho
	Teachers and counsellors	0.009168	0.207468	Reject Ho
	Media platforms	0.4171	0.064396	Accept Ho

	Role models	0.5458	0.04827	Accept Ho
Family	Friends	4.972e-08	0.43593	Reject Ho
	Financial status	3.296e-06	0.37171	Reject Ho
	Mental and physical health	1.961e-07	0.41479	Reject Ho
	Teachers and counsellors	0.0002542	0.29089	Reject Ho
	Media platforms	0.4918	0.05449	Accept Ho
	Role models	0.05607	0.15247	Accept Ho
Friends	Financial status	2.155e-07	0.41109	Reject Ho
	Mental and physical health	1.342e-06	0.38226	Reject Ho

Friends	Teachers and counsellors	5.808e-05	0.31718	Reject Ho
	Media platforms	9.341e-05	0.30725	Reject Ho
	Role models	0.07585	0.14056	Accept Ho
Financial status	Mental and physical health	8.125e-06	0.352703	Reject Ho
	Teachers and counsellors	0.001102	0.257308	Reject Ho
	Media platforms	0.002488	0.23775	Reject Ho
	Role models	0.01718	0.188576	Reject Ho
Mental and physical health	Teachers and counsellors	1.378e-07	0.414406	Reject Ho
	Media platforms	5.611e-05	0.31588	Reject Ho

	Role models	0.006649	0.21425	Reject Ho
Teachers and counsellors	Media platforms	3.328e-05	0.32457	Reject Ho
	Role models	0.0268	0.174393	Reject Ho
Media platforms	Role models	2.841e-07	0.40299	Reject Ho
Marks in SSC	Satisfaction rate	0.9697	0.0017627	Accept Ho
Marks in HSC	Satisfaction rate	0.1206	0.0719615	Accept Ho

NON-PARAMETRIC TESTS

Nonparametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed). Due to this reason, they are sometimes referred to as distribution-free tests

Nonparametric tests serve as an alternative to parametric tests such as T-test that can be employed only if the underlying data satisfies certain criteria and assumptions.

Note that nonparametric tests are used as an alternative method to parametric tests, not as their substitutes. In other words, if the data meets the required assumptions for performing the parametric tests, the relevant parametric test must be applied.

1) Mann-Whitney U Test

The Mann-Whitney U Test is a nonparametric version of the independent samples t-test. The test primarily deals with two independent samples that contain ordinal data.

Ho:Mx=My v/s H1:Mx>My

Mx=Median of satisfaction level of people who change their ambition after school

My: Median of satisfaction level of people who don't want to further switch their career

Decision criteria:p-value= 0.1617

Since p is greater than 0.05 thus we accept Ho at 0.05 l.o.s

CONCLUSION: Median satisfaction level of people who did not change their ambition after school is greater than median satisfaction level of people who don't want to further switch their career.

Similarly , Mann-Whitney U Test is performed between the following variables as follows

VARIABLE 1	VARIABLE 2	HYPOTHESIS	P VALUE	CONCLUSION
Changed first ambition	Want to switch further	Mx>My	0.003169	REJECT Ho
	Didn't change first ambition	Mx<My	3.205*10^-6	REJECT Ho
	Don't want to switch further	Mx<My	5.99*10^-6	REJECT Ho
Didn't change first ambition	Want to switch further	Mx>My	5.99*10^-6	REJECT Ho
Want to switch further	Don't want to switch further	Mx>My	3.352*10^-9	REJECT Ho

2. Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test is a nonparametric counterpart of the paired samples t-test. The test compares two dependent samples with ordinal data.

Although the sign test can be used to test both one-sample and two-sample paired data, the Wilcoxon signed-rank test is more powerful than the sign test for these tasks because it makes use of the magnitudes of the differences rather than just their signs.

The Wilcoxon signed rank test was developed by Frank Wilcoxon1 in 1945. We will illustrate its use using two-sample paired data. Following our checklist from Section 5.2, the basic idea behind the Wilcoxon signed-rank test is:

Form null and alternative hypotheses and choose a degree of confidence. The null hypothesis is that the median of the population of differences between the paired data is zero. The alternative hypothesis is that it is not., that is, without regard to their sign; next, we sum the ranks of the positive and negative differences; finally, we pick the minimum of the sums as our test statistic.Compare the test statistic to a critical value. If the test statistic is less than the critical value, then we reject the null hypothesis

Ho:Medians of Percentages of SSC and HSC are Equal

v/s

H1:Medians of percentages of SSC is less than that of HSC

Decision criteria:We reject Ho if p value is less than 0.05

p-value=2.2*10^-16

Since p value is less than 0.05 we reject Ho

CONCLUSION: Median of percentages of SSC and HSC are not equal

REGRESSION

Regression is a way of describing how one variable, the outcome, is numerically related to predictor variables. A regression analysis expresses this relationship in the form of an equation. The dependent variable is also referred to as Y, dependent or response and is plotted on the vertical axis (ordinate) of a graph. The predictor variable(s) is(are) also referred to as X, independent, or explanatory variables.

Linear regression attempts to identify a line (of the form $y = ax + b$) that is the best fit for the data under analysis. Other forms of regression analysis include polynomial ($y = ax^2 + bx + c$ and higher-order polynomials), logarithmic ($y = \log_b x$), and exponential ($y = b^x$). Manually computing these regressions is tedious and time consuming by hand.

The general process of correlation and regression analysis of a data set follows these steps:

- Identify the independent variable(s) in the data set.
- Identify the dependent variable(s) in the data set.
- Construct a scatter plot for the data.
- Using the scatter plot as a guide fitting curves to the data using various regression analysis models. Visual evaluation of fitted curves superimposed on scatter plots is combined with computing the appropriate correlation coefficients to determine the best numerical model for the data (if one exists)

BINARY LOGISTIC REGRESSION

Binary Logistic Regression is the statistical technique used to predict the relationship between predictors and a predicted variable where the dependent variable is binary. Logistic regression is an extension of simple linear regression. There must be two or more independent variables or predictors for a logistic regression. The predictors can be continuous or categorical. We cannot use Simple Linear Regression where the dependent variable is dichotomous or binary in nature. Use a binary regression model to understand how changes in the predictor values are associated with changes in the probability of an event occurring.

ASSUMPTIONS:-

I. Dependent variable should be binary.

II. One or more independent variables are continuous, ordinal or categorical.

III. There is no multicollinearity.

(MORE)

MULTICOLLINEARITY:-

The chi-square test is a hypothesis test designed to test for a statistically significant relationship between nominal and ordinal variables organised in a bivariate table. In other words, it tells us whether two variables are independent of one another.
VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.
VIF starts at 1 and has no upper limit.
VIF = 1, no correlation between the independent variable and the other variables.
VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.

> vif(LOG1)

	GVIF	Df	GVIF^(1/(2*Df))
SSC	1.229746	1	1.108939
SATISFACTION	1.236206	1	1.111848
Father.s..Qualification	1.362416	5	1.031409

> vif(LOGISTIC1)

	GVIF	Df	GVIF^(1/(2*Df))
O1	1.243691	1	1.115209
O2	1.130135	1	1.063078
O9	1.653393	1	1.285843
010	1.617864	1	1.271953
JOB1	1.121013	6	1.009565

a) BLR1

Our dependent variable is switching your ambitions during school and independent variables are marks, interest , media platforms, role models, and their primary job expectation
We fitted the BLr on train data for the above variables and got the output

```
glm(formula = SWITCH.1 ~ O1 + O2 + O9 + O10 + JOB1, family = "binomial",
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.33433  -0.43214   0.00016   0.40474   2.38827

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    9.6167     1.6502   5.828 5.62e-09 ***
O1            -0.8406     0.2164  -3.885 0.000102 ***
O2            -1.7279     0.2896  -5.966 2.44e-09 ***
O9             0.4357     0.2629   1.658 0.097393 .
O10           -0.5918     0.2449  -2.417 0.015654 *
JOB12          1.2138     0.5570   2.179 0.029322 *
JOB13        -22.8071  2679.0431  -0.009 0.993208
JOB14         17.7339  1951.7803   0.009 0.992750
JOB15          0.2111     0.6237   0.338 0.735028
JOB16         -0.1001     0.8635  -0.116 0.907754
JOB17         18.0282  2113.2900   0.009 0.993193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 275.98  on 199  degrees of freedom
Residual deviance: 130.12  on 189  degrees of freedom
AIC: 152.12

Number of Fisher Scoring iterations: 17
```

Confusion matrix of Train data

	0	1
F	79	16
T	13	92

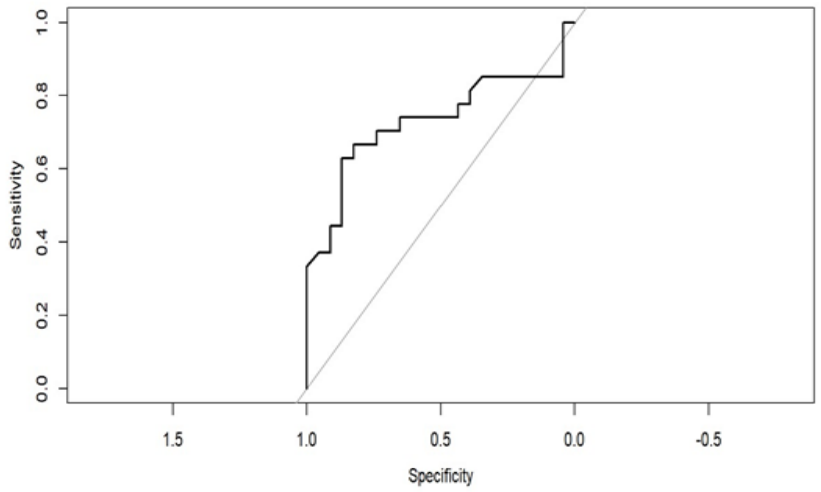
Accuracy : 85.5%

Confusion matrix of Test data

	0	1
F	19	9
T	4	18

Accuracy :74%

ROC Curve



AUC : 0.7311

GLOBAL TEST

It is used to check whether at least one of the independent variables is significant for the regression model.

Ho: No independent variable is significant v/s H1: At least one of the independent variable is significant

Model 1: SWITCH.1 ~ 1
Model 2: SWITCH.1 ~ O1 + O2 + O9 + O10 + JOB1

Analysis of Deviance Table

Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	199	275.98			
2	189	310.12	10	145.86 <	2.2e-16 *

Therefore we reject Ho and conclude that at least one of them is significant .

b) BLR2

Our dependent variable is switching your ambitions further and independent variables are
We fitted the BLR on train data for the above variables and got the output

```
R 4.1.2 · C:/Users/bhoom/Downloads/
Call:
glm(formula = SWITCH2 ~ SSC + SATISFACTION + Father.s..Qualification,
    family = "binomial", data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8273  -0.7060  -0.5001   0.5196   2.2226

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               7.07574    1.78128   3.972 7.12e-05 ***
SSC                      -0.05919    0.01837  -3.223  0.00127 **
SATISFACTION             -0.52470    0.09961  -5.268 1.38e-07 ***
Father.s..Qualification1  0.37366    0.74894   0.499  0.61784
Father.s..Qualification2  0.11825    0.74539   0.159  0.87395
Father.s..Qualification3  0.77144    0.79383   0.972  0.33115
Father.s..Qualification4  0.68557    0.67586   1.014  0.31041
Father.s..Qualification5  1.83833    0.82195   2.237  0.02532 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.69  on 200  degrees of freedom
Residual deviance: 185.61  on 193  degrees of freedom
AIC: 201.61

Number of Fisher Scoring iterations: 5
```

Confusion matrix of train data

	0	1
F	140	34
T	10	17

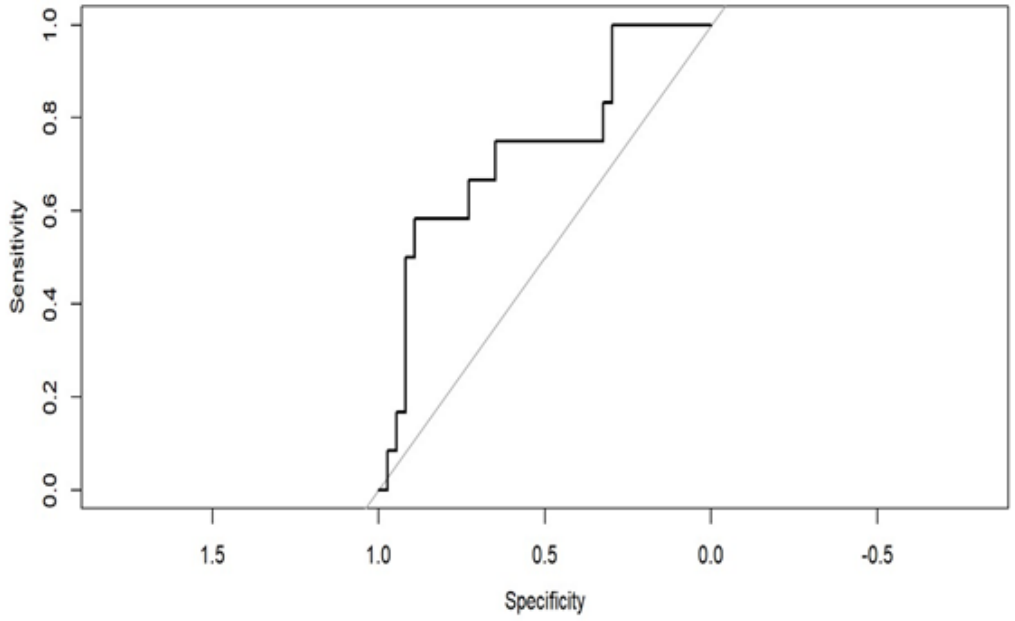
Accuracy=78.11%

Confusion matrix for the test data

	0	1
F	34	10
T	3	2

Accuracy : 73.47%

ROC Curve



AUC : 0.732

GLOBAL TEST

It is used to check whether at least one of the independent variables is significant for the regression model.

Ho: no independent variable is significant

v/s

H1: at least one of the independent variable is significant

Model 1: SWITCH2 ~ 1
Model 2: SWITCH2 ~ SSC + SATISFACTION + Father.s..Qualification

Analysis of Deviance Table

Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	200	227.69			
2	193	185.61	7	42.079	5.021e-07 *

Therefore we reject Ho and conclude that at least one of them is significant .

CHAID

CHAID is a type of decision tree technique, based upon adjusted significance testing (Bonferroni testing). The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic. CHAID can be used for prediction (in a similar fashion to regression analysis, this version of CHAID being originally known as XAID) as well as classification, and for detection of interaction between variables. CHAID stands for Chi-squared Automatic Interaction Detection based upon formal extension of the US AID (Automatic Interaction Detection) and THAID (Theta Automatic Interaction Detection) procedure of the 1960s and 70s which in turn were extensions of earlier research, including that performed in the UK in the 1950s. In practice, CHAID often used in the context of direct marketing to select groups of consumers and predict how their responses to some variables after other variables, although other early applications were in the field of medical and psychiatric research. Like other decision trees, CHAID's advantages are that its output is highly visual and easy to interpret. Because it uses multiway splits by default, it needs rather large sample sizes to work effectively, since with small sample sizes the respondent groups can quickly become too small for reliable analysis.

One important advantage of CHAID over alternatives such as Multiple regression is non-parametric.

- CHAID or Chi-Square Automatic Interaction Detector is an exploratory
- method or more precisely an algorithm to study the relationship
- between a dependent variable and a series of predictor variables.
- CHAID is an useful method of summarizing the data and is analogous to
- stepwise regression.
- CHAID analysis also known as decision tree analysis, can be used for
- prediction as well as classification.
- CHAID identifies discrete groups of variables and seeks to predict the
- impact of each variable on the target variable.
- First select independent variable whose subgroups differ most with
- respect to dependent variables.
- Each subgroup of this variable is further divided into subgroups on
- remaining variables.
- These subgroups are tested for differences on dependent variables.
- Variable with the greatest difference is selected next.
- Continue until subgroups are too small.

FOLLOWING DECISION TREE COMPONENTS ARE
USED IN CHAID ANALYSIS:

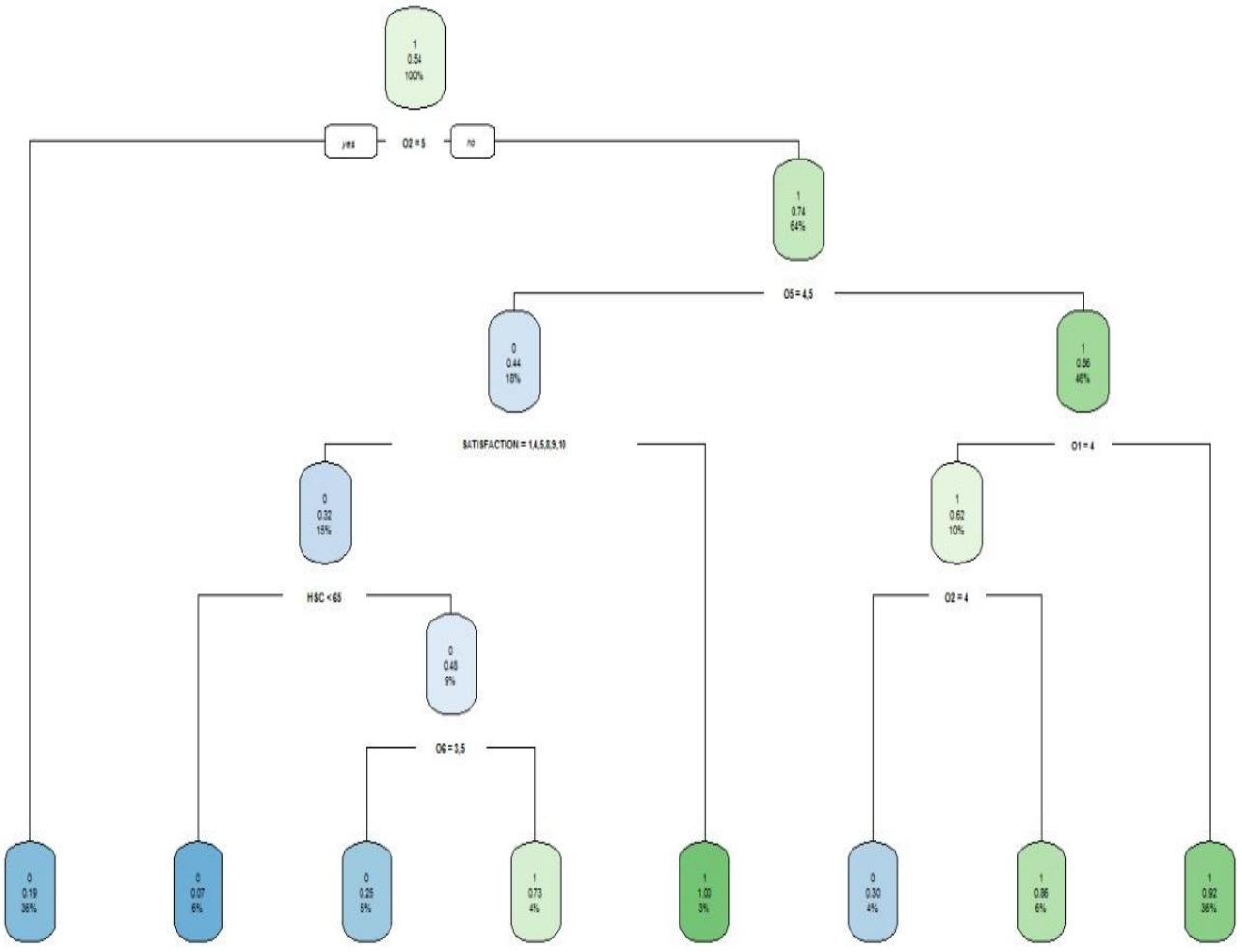
1. ROOT NODE :
Root node contains the dependent, or target, variable. For example,
CHAID is appropriate if a bank wants to predict the credit card risk based upon information like age, income, number of credit cards, etc. In this example, credit card risk is the target variable and the remaining factors are the predictor variables.

2. PARENTS NODE :
The algorithm splits the target variable into two or more categories. These categories are called parent node or initial nodes. For the bank for example: high, medium and low categories are the parent's nodes.

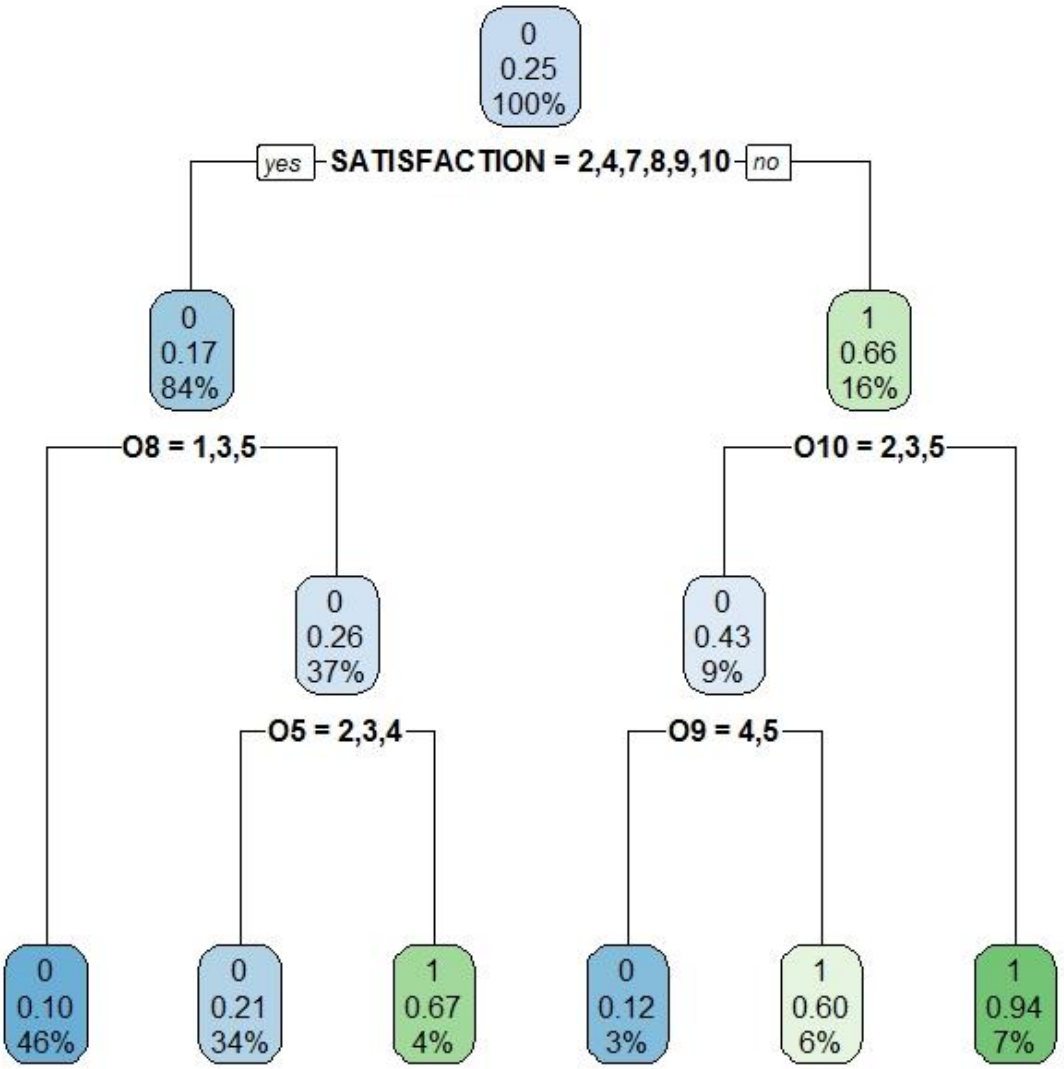
3. CHILD NODE :
Independent variable categories which come below the parent's categories in the CHAID analysis tree are called the child node.

4. TERMINAL NODE ;
The last categories of the CHAID analysis tree are called the terminal node. In the CHAID analysis tree, the category that is a major influence on the dependent variable comes first and the less important category comes last. Thus, it is called the terminal node.

1) CHAID FOR STUDENTS WHO SWITCHED FOR THE FIRST TIME



2) CHAID FOR STUDENTS WHO ARE GOING TO SWITCH FURTHER

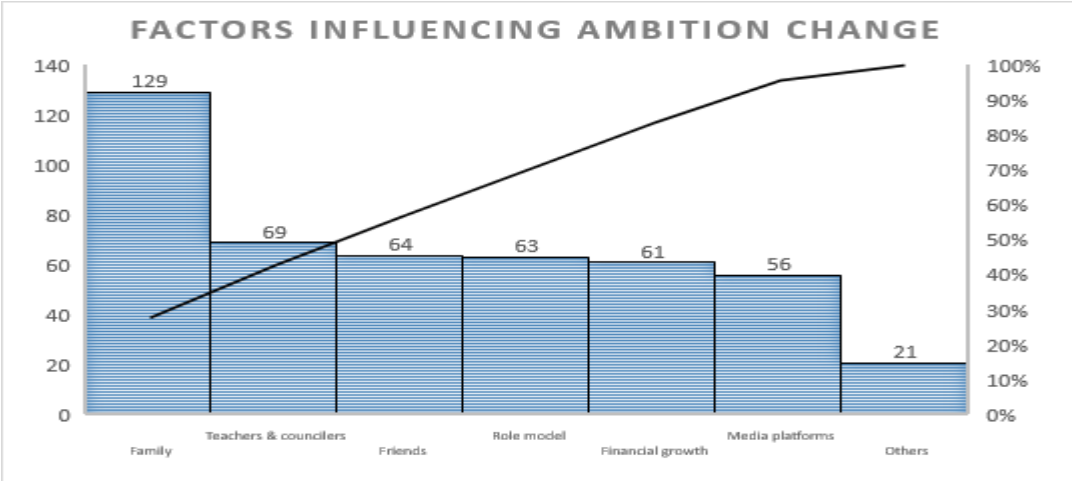


PARETO CHART

A Pareto chart is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line. The chart is named for the Pareto principle, which, in turn, derives its name from Vilfredo Pareto, a noted Italian economist. The left vertical axis is the frequency of occurrence, but it can alternatively represent cost or another important unit of measure. The right vertical axis is the cumulative percentage of the total number of occurrences, total cost, or total of the particular unit of measure. Because the values are in decreasing order, the cumulative function is a concave function. The purpose of the Pareto chart is to highlight the most important among a set of factors. In quality control, it often represents the most common sources of defects, the highest occurring type of defect, or the most frequent reasons for customer complaints, and so on. Wilkinson (2006) devised an algorithm for producing statistically based acceptance limits for each bar in the Pareto chart. These charts can be generated by simple spreadsheet programmes, specialized statistical software tools, and online quality charts generators. The Pareto chart is one of the seven basic tools of quality control. Normally Pareto charts are used for multiple options.

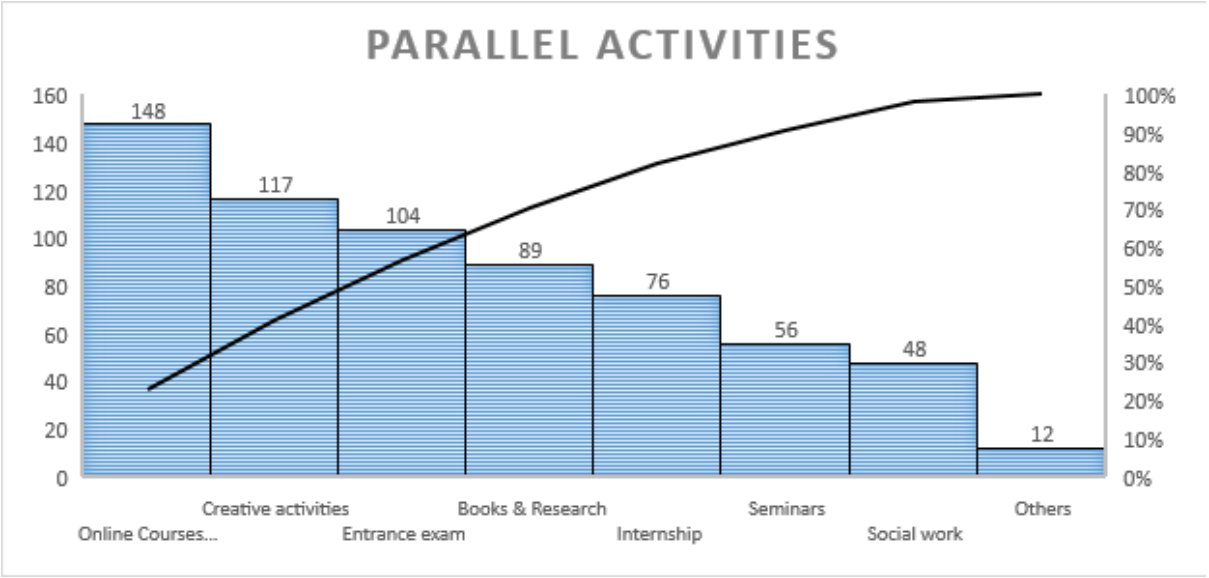
PARETO PRINCIPLE :- Management consultant Joseph M. Juran suggested the principle and named it after Italian economist Vilfredo Pareto, who noted the 80/20 connection while at the University of Lausanne in 1896. In his first work, Pareto showed that approximately 80% of the land in Italy was owned by 20% of the population. The Pareto principle is only tangentially related to Pareto efficiency. Pareto developed both concepts in the context of the distribution of income and wealth among the population. Mathematically, the 80/20 rule is roughly followed by a power law distribution (also known as a Pareto distribution) for a particular set of parameters, and many natural phenomena have been shown empirically to exhibit such a distribution. It is an axiom of business management that 80% of sales come from 20% of clients.

1)



Family influences the most for the change of ambition in a person followed by teachers,counsellors and friends .

2)



Maximum students take online courses as a parallel activity to achieve their ambition followed by creative activities and competitive exams.

WORD CLOUD

A word cloud (tag cloud or wordle or weighted list in visual design) is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualise free form text. Tags are usually single words, and the importance of each tag is shown with font size or colour. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is. This format is useful for quickly perceiving the most important terms to determine its relative prominence. A word cloud is a popular visualization of words typically associated with internet keywords and text data. They are most commonly used to highlight popular or trending terms based on frequency of use and prominence. A word cloud is a beautiful, informative image that communicates much in a single glance. It is used as website navigation aids, the terms are hyperlinked to items associated with the tag. Word clouds are an easy to use and inexpensive option for visualizing text data. One of the challenges of interpreting word clouds is that the display emphasizes frequency of words. Word cloud is used in Marketing: Make a word cloud from all the customer reviews of a product. Helps you target profitable customers. Quantitative Researchers: For communicating qualitative data. Twitter: Scan Twitter tweets and make a word cloud to discover the most tweeted terms and topics. Share this on other social media profiles. Non-profits: As an affordable method to collect and share sentiment on social media sites. Politicians and journalists: Word Clouds to share the latest political trends! Word clouds can also be used to evaluate your brand identity, optimize blogs and marketing content for search engines, and research your competitor's SEO strategy.

Word clouds are a method for visually presenting text data. They are popular for text analysis because they make it easy to spot word frequencies. The more frequent the word is used, the larger and bolder it is displayed. Word clouds add clearity. Word clouds can identify trends and patterns that would otherwise be unclear or difficult to see in a tabular format. Word clouds are impactful. Word clouds are especially impactful when shaped into an image that reflects your topic or theme. You are far more likely to captivate your audience with a word cloud than a table or bar graph. Word clouds are easy to understand. Besides being more visually appealing than a table of data, word clouds are easier to understand. When an appropriate title is used, they are pretty self explanatory. Word clouds are easy to share. Another advantage of using a pictorial representation of your data, is that images are far more likely to be shared than textual data. This is why image - based social media sites are so popular. There are three main types of word cloud applications in social software, distinguished by their meaning rather than appearance. In the first type, there is a tag for the frequency of each item, whereas in the second type, there are global tag clouds where the frequencies are aggregated over all items and users. In the third type, the cloud contains categories, with size indicating number of subcategories. Here we are using the first type of a tag for the frequency of each item. In this type, size represents the number of times that tag has been applied to a single item. This is useful as a means of displaying metadata about an item that has been democratically "voted" on.
We asked one open ended question in the questionnaire, "Desired ambition during school" Many students voted on it. Using the word cloud we can easily identify more important

factors with the help of their font size. Here, if the organisation makes changes to the large number of words that can occur frequently

1)

2)

3)


SoftwareDeveloper
AirForce
TeacherDoctorSocial
Officer
PharmacistDoctor
Athlete
Auditor
Writer
Pilot
IAS
Journalist
Engineer
Detective
Advocate
Architect
AutomobileEngineer
Astronomer
Researcher
Work
MechanicalEngineer
Teacher
Artist
AirForcePilot
SoftwareEngineer
Astrophysicist
Well
AeronauticalEngineer
ISROAstronomer
Physicist
Professor
ArmedForces
Police
Lecturer
Dancer
SportPerson
Lawyer
Mathematician
CA
Scientist
Defence
CivilServices
Chef
job
Actor
Psychologist
Doctor
Physiotherapist
ActuarialScience
Cricketer
Engineering
DataScientist
Accountant
Analyst
worker
settled
Astronaut
Cinematographer
Entrepreneur
DataAnalyst
FashionDesigner
Pharmacist
None
Particle
TheoreticalPhysics
SCHOOL AMBITION OF HIGHLY SATISFIES PEOPLE (8 to 10)

4)


Astronomer
Particle
SoftwareEngineer
Architect
Lawyer
Pilot
None
PharmacistDoctor
FashionDesigner
Athlete
Psychologist
ComputerEngineer
MechanicalEngineer
CA
IAS
teacher
IFS
TeacherDoctorSocial
ArmedForces
MusicProducer
Doctor
Scientist
Chef
Journalist
DoctorTeacher
Astronaut
Collector
Defence
Engineer
Researcher
AutomobileEngineer
Physicist
Mathematician
Astrophysicist
SportPerson
AirForcePilot
Cricketer
CivilServices
Detective
CMA
InteriorDesigner
DataScientist
SoftwareDeveloper
Officer
Physiotherapist
DataAnalyst
Marketing
CHILDHOOD AMBITION

5)


Architect
Astronaut
Doctor
Engineer
Teacher
Pilot
Journalist
IAS
IFS
CA
Cricketer
Professor
Entrepreneur
MechanicalEngineer
Nurse
Aviator
SportPerson
CHILDHOOD AMBITION AFTER SWITCHING (STATISTICS RELATED)

FINDINGS

- The majority of the respondents are of the age 20.
- The majority of the respondents belong to the graduate category.
- The majority of the respondents are female.
- The majority of the respondents are staying in Mumbai.
- Chances of further career switch for science students is more than that of non science students.
- Marks of the students who changed their ambitions are influenced by the media platforms they use whereas not for those who did not switch.
- The mental and physical health of the students are significantly associated with their financial status.
- Working sector encouraged by their parents is not biased towards their qualifications.
- The 10t and 12th marks do not affect their satisfaction in the career now.
- The marks of the students are influenced by the area where they live.
- The stream they choose and their interest in the subjects are associated.
- There is no relation between marks of HSC and SSC and ambition change.
- Family's influence and financial status of the students are also associated with staying and changing the ambition.
- There is a significant association between the Influence of Media platforms and student's role models.
- The satisfaction level of people who did not change their first ambition and those who do not want to further change is significantly the same.
- Family influences the most for the change of ambition in a person followed by teachers,counsellors and friends.
- Maximum students take online courses as a parallel activity to achieve their ambition followed by creative activities and competitive exams.
- Most of the students wanted to be doctors and engineers when they were in school but now since statistics is the most wanted subject, most people want to be something related to our field.

QUESTIONNAIRE

Hello Everyone!We the students of SIES college are conducting a survey on career choices for our Research project purpose. We request you to please spare some time and kindly fill the form.
Information will be used only for educational purpose.We greatly appreciate your responses.
Thank you!

---

* Required

1. Gender *
   Mark only one oval.
   - ☐ Male
   - ☐ Female
   - ☐ Other : _____

2. Age completed (eg.20 , 24) *

   _____

3. Location *
   Mark only one oval.
   - ☐ Mumbai City
   - ☐ Mumbai Suburban
   - ☐ Navi Mumbai to Panvel
   - ☐ Thane to Karjat

4. Latest Degree qualification (ongoing) *
   Mark only one oval.
   - ☐ HSC
   - ☐ Bachelor's
   - ☐ Masters
   - ☐ PHD
   - ☐ Diploma

5. Current stream of Education *
Mark only one oval.
- ☐ Science
- ☐ Arts
- ☐ Commerce
- ☐ Other: _____

6. Father's Educational Qualification *
Mark only one oval.
- ☐ Below SSC
- ☐ SSC
- ☐ HSC
- ☐ Bachelor's
- ☐ Masters
- ☐ PHD
- ☐ Diploma
- ☐ Other:

7. Mother's Educational Qualification *
Mark only one oval.
- ☐ Below SSC
- ☐ SSC
- ☐ HSC
- ☐ Bachelor's
- ☐ Masters
- ☐ PHD
- ☐ Diploma
- ☐ Other:

8. Your percentage in SSC *

_____

9. Your percentage in HSC *

_____

10. Desired ambition during school (eg- doctor, teacher etc) *

_____

11. Which working sector was highly encouraged by your parents *
Mark only one oval.

- ☐ Public
- ☐ Private/Business
- ☐ Both public and private
- ☐ None

12. What factors influenced your decision for choosing that ambition during school (multiple selections are allowed) *
Tick all that apply.

- ☐ Family
- ☐ Friends
- ☐ Financial growth
- ☐ Teachers & counsellors
- ☐ Media platforms
- ☐ Role model
- ☐ Other:

13. Did you change your ambition *
Mark only one oval.

- ☐ Yes Skip to question 14
- ☐ No Skip to question 16

Skip to question 17

DECISION FACTORS

14. What is your desired ambition now ? (Eg Data scientist, Researcher etc) *

_____

15. Main factors that contributed in change of your ambition *
Mark only one oval per row.

	strongly agree	agree	neutral	disagree	strongly disagree
Marks in HSC/SSC	☐	☐	☐	☐	☐
Loss of interest in subject	☐	☐	☐	☐	☐
Decreasing scope of subject/increase of scope in others.	☐	☐	☐	☐	☐
Family	☐	☐	☐	☐	☐
Friends	☐	☐	☐	☐	☐
Financial status	☐	☐	☐	☐	☐
Health(mental & physical)	☐	☐	☐	☐	☐
Teachers & Counsellors	☐	☐	☐	☐	☐
Media platforms	☐	☐	☐	☐	☐
Role model	☐	☐	☐	☐	☐
Language barrier	☐	☐	☐	☐	☐

Skip to question 17

DECISION FACTORS

16. Main factors that contributed in change of your ambition *
Mark only one oval per row.

	strongly agree	agree	neutral	disagree	strongly disagree
Marks in HSC/SSC	☐	☐	☐	☐	☐
Loss of interest in subject	☐	☐	☐	☐	☐
Decreasing scope of subject/increase of scope in others.	☐	☐	☐	☐	☐
Family	☐	☐	☐	☐	☐
Friends	☐	☐	☐	☐	☐
Financial status	☐	☐	☐	☐	☐
Health(mental & physical)	☐	☐	☐	☐	☐
Teachers & Counsellors	☐	☐	☐	☐	☐
Media platforms	☐	☐	☐	☐	☐
Role model	☐	☐	☐	☐	☐

17. What are your top three primary expectations regarding your desired job ? *

	Salary	Creative freedom	Working atmosphere	Travelling	Working benefits	Working hours	Work stress
First	☐	☐	☐	☐	☐	☐	☐
Second	☐	☐	☐	☐	☐	☐	☐
Third	☐	☐	☐	☐	☐	☐	☐

18. In which parallel activity/course did you invest yourself along with your academics ? (Multiple selections are allowed) *

Tick all that apply.

- ☐ Online Courses (academic)
- ☐ Internship
- ☐ Seminars
- ☐ Books & Research
- ☐ Preparation for entrance exam
- ☐ Creative activities(dance, music, fitness etc)
- ☐ Social work
- ☐ Other: _____

19. Are you planning for further career switch? *

Mark only one oval.

- ☐ Yes
- ☐ No

20. Rate your satisfaction level regarding your current career decision (1 being least satisfied and 10 being most satisfied) *

Mark only one oval.

1	2	3	4	5	6	7	8	9	10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

THANK YOU!!

R-codes

- One sample proportion test
  prop.test(63,250,0.2,alternative = "greater",0.95)

- Chi square test
  chisq.test(T3)

- Linear by linear test
  lbl_test(T30)

- Kendall's tau

  cor.test(O$O1,O$O7,method ="kendall")

  freemanTheta(T40)

  epsilonSquare(T40)

- Mann Whitney U test

  wilcox.test(MAZE$percentage_SSC,MAZE$percentage_HSC,paired=TRUE,alternative = "less",conf.level = 0.95)

- Wilcoxon Signed Rank test

  wilcox.test(Sort$Satisfaction_level...3,Sort$Satisfaction_level...6,mu=0, alt="greater", conf.int=T,conf.level=0.95,paired=F,exact=F,correct=T)

- BLR:

  Multicollinearity:-

  library(car)

  vif(LOG1)

  vif(LOGISTIC1)

BLR for 1st time switchers:-

```
LOGISTIC<-glm (SWITCH.1~Location+Father.s..Qualification +Mother.s.Qualification+ SSC +  HSC + SECTOR+O1 +O2 +O3 +O4 +O5 +O6 +O7 +O8 +O9 +O10+ JOB1+Current.stream.of.Education,family = "binomial",,data = train)

LOGISTIC

summary(LOGISTIC)

LOGISTIC1<-glm(SWITCH.1~O1 +O2 +O9 +O10+ JOB1,family = "binomial",data = train)

LOGISTIC1

summary(LOGISTIC1)

predicted.data<-data.frame(probability.of.SWITCH.1=LOGISTIC1$fitted.values,SWITCH.1 =train$SWITCH.1)

predicted.data<-predicted.data[order(predicted.data$probability.of.SWITCH.1,decreasing = FALSE),]

predicted.data$rank<-1:nrow(predicted.data)

length(predicted.data)

library(ggplot2)

dim(predicted.data)

library(cowplot)

ggplot(data=predicted.data,aes(x=rank,y=probability.of.SWITCH.1))+geom_point(aes(color= SWITCH.1),alpha=1,shape=4,stroke=2)+xlab("Index")+ylab("Predicted probabaility of SWITCH.1")

head(predicted.data)

##confusion matrix

a=table(predicted.data$probability.of.SWITCH.1>0.5,predicted.data$SWITCH.1)

a

##CLASSIFICATION RATE

(sum(diag(a)))/sum(a)

library(pROC)
```

```
plot.roc(train$SWITCH.1,predicted.data$probability.of.SWITCH.1)
auc(train$SWITCH.1,predicted.data$probability.of.SWITCH.1)
PRED1<-predict(LOGISTIC1,test,type = "response")
PRED1
##confusion matrix
a=table(PRED1>0.5,test$SWITCH.1)
a
##CLASSIFICATION RATE
(sum(diag(a)))/sum(a)
library(pROC)
plot.roc(test$SWITCH.1,PRED1)
auc(test$SWITCH.1,PRED1)
```

CHAID (students who will switch further)

```
        install.packages("dplyr")
        library(dplyr)
        install.packages("ggplot2")
        library(ggplot2)
        install.packages("rpart")
        library(rpart)
        install.packages("rpart.plot")
        library(rpart.plot)
        install.packages("rattle")
        library(rattle)
        install.packages("ISLR")
        library(ISLR)
        str(REGT)
        factor=as.factor(REGT)
        data.frame=data.frame(REGT)
```

```
str(data.frame)
REGT$Location=as.factor(REGT$Location)
is.factor(REGT$Location)
REGT$`Father's  Qualification`=as.factor(REGT$`Father's  Qualification`)
REGT$`Mother's Qualification`=as.factor(REGT$`Mother's Qualification`)
REGT$SECTOR=as.factor(REGT$SECTOR)
REGT$`SWITCH 1`=as.factor(REGT$`SWITCH 1`)
REGT$O1=as.factor(REGT$O1)
REGT$O2=as.factor(REGT$O2)
REGT$O3=as.factor(REGT$O3)
REGT$O4=as.factor(REGT$O4)
REGT$O5=as.factor(REGT$O5)
REGT$O6=as.factor(REGT$O6)
REGT$O7=as.factor(REGT$O7)
REGT$O8=as.factor(REGT$O8)
REGT$O9=as.factor(REGT$O9)
REGT$O10=as.factor(REGT$O10)
REGT$JOB1=as.factor(REGT$JOB1)
REGT$SWITCH2=as.factor(REGT$SWITCH2)
REGT$SATISFACTION=as.factor(REGT$SATISFACTION)
REGT$`Current stream of Education`=as.factor(REGT$`Current stream of Education`)

analysis=rpart(SWITCH2~
Location+SECTOR+O1+O2+O3+O4+O5+O6+O7+O8+O9+O10+JOB1+
SATISFACTION+`Current stream of Education`, data=REGT,method = "class")
analysis
rpart.plot(analysis)
prp(analysis)
```