



Team Name: HackO1

Team Members: Vinal Bagaria

Mohit Badve

Palak Davda

Table Of Contents

[**Problem Definition**](#_n6i8js6qmv1l) **2**

[**Approach Description**](#_xk90sen6tb00) **3**

[Approach Flow](#_qgy0g3irvt5k) 3

[Hypothesis Generation](#_aohwkvp8e1pm) 4

[**Data Exploration**](#_s1qghrpyn777) **5**

[I) Student Data](#_ai2o9uwagqkw) 5

[Data Manipulation](#_hxmvyu8k84ht) 5

[Data Consolidation](#_958vleiccqa) 5

[Data Preparation](#_z6a6lm16ea1a) 6

[Missing value](#_6d4viiefh5iq) 6

[Outliers](#_ndxfhx7ix9zs) 7

[Data Derivation](#_ftwwc4efy1en) 8

[HADS Score](#_4x4ew1ewt8cn) 8

[Age Group](#_xqssf72m3uxy) 9

[BMI index](#_q4okv2iiv6iz) 9

[IQ Score](#_q3efvh6gkbfk) 10

[IsHealthy?](#_zatqnoibzapy) 10

[II) Suicide Victim Data](#_ffxob1j87t4i) 11

[Data Manipulation](#_gndxwqyknjpl) 11

[Data Consolidation](#_6d6cqcfaoqik) 11

[Data Preparation](#_gzv8tdu4mjal) 12

[Missing Value](#_woogeomvxok5) 12

[**Data Analysis**](#_oti7ncb2vs4) **12**

[I) Student Data](#_v6fpxkftcxnu) 12

[1. Gender](#_5u02jk7z9gtt) 12

[2. Peer Pressure](#_46mel1plcdd3) 13

[3. Exercise](#_8ccyo637avco) 14

[4. Living with Family](#_ew2cd4skv0c9) 15

[5. Education](#_ot84enjdoamm) 16

[6. Age](#_oo0gr4uf4x0j) 16

[7. BMI](#_x8k0d9bvbgii) 17

[8.Past History of Depression](#_u7uowfu4nkac) 18

[9. IQ](#_kz73xwu2vp7m) 19

[10. Smoke](#_g8ufu5n9mnlc) 19

[11. Is Healthy?](#_9f3978c551ee) 20

[II) Suicide Victim Data](#_2sebtmulg97a) 20

[1. Yearly Aggregation](#_9kcvq0l7sqm0) 21

[2. Cause](#_p4ai41i23hr5) 21

[3. Gender Bias](#_dvj90c3mvn78) 22

[4. Age Factor](#_xrmhrpqx9p9f) 23

[5. Cause in Yearly Pattern](#_27gbxiowg4jp) 23

[**Modelling**](#_j5a8c8vpvg1s) **25**

[Data Scaling](#_mxf3xiqr9tuk) 25

Modeling and Evaluation 25

[**Results and Conclusions**](#_93jkjggg7q1f) **26**

[**Implications**](#_esixj3jiuoxp) **27**

## **Problem Definition**

1. Identification of Seriousness of the problem

Suicide is a complex public health problem of global dimension. Suicidal behaviour (SB) shows marked differences between genders, age groups, geographic regions and socio-political realities, and variably associated with different risk factors, underscoring likely etiological heterogeneity.

In a recent study published in Depression and Anxiety of more than 67,000 college students from more than 100 institutions, one in five students have had thoughts of suicide, with 9% making an attempt and nearly 20% reporting self-injury. One in four students reported being diagnosed with a mental illness. So it can be directly inferred that sucide may be dependent on the depression.suicide is the second leading cause of death for young people ages 15 to 24.

According to 2015 data from the National Crime Records Bureau, 8934 (6.7% of all suicides) students are committing suicide every year. That's one student every hour. There are several factories contributing to the increasing rate of suicide.

Social media usage being one of the major contributors. Increasing use of social media leads to an increase in depression in the youth. Since the advent of digitalization and high exposure to social media, the problem will become more and more serious.

1. Measurement of students depression

Hospital Anxiety and Depression Scale (HADS) is commonly used by doctors to determine the levels of [anxiety](https://en.wikipedia.org/wiki/Anxiety) and [depression](https://en.wikipedia.org/wiki/Depression_(mood)) that a person is experiencing. The HADS is a fourteen item scale that generates [ordinal data](https://en.wikipedia.org/wiki/Level_of_measurement). Seven of the items relate to anxiety and seven relate to depression. HADS score of these questionnaire can depict the depression level in the student.

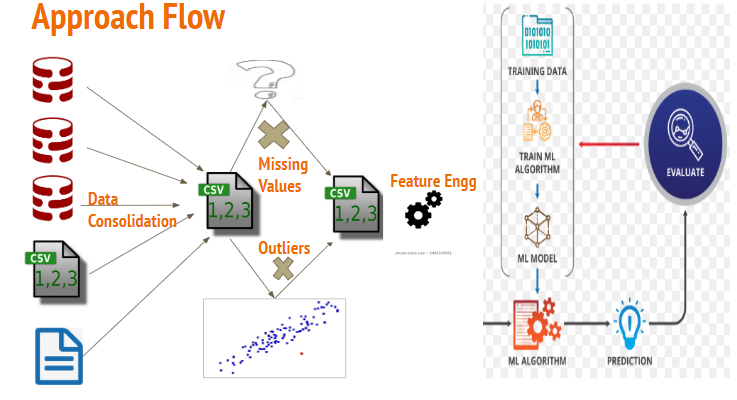
1. Objective

* Analysis of the factors affecting the suicidal
* Prediction of chances to commit suicide by the students, depending on the features that allevate the problem

## **Approach Description**

### Approach Flow

1. Obtain the consolidated data of students from different files for the analysis
2. Treat the missing values, outliers, duplicates if any
3. Get the insights from the data using data visualization
4. The year-wise suicide count data will take us to a broad overview about the different features of the persons such as height, weight, age, professional profile, income, etc. affecting the suicidal count in India.
5. Especially the CAUSE feature from the data can be used to analyze the trend of suicide count over the years with respect to CAUSE feature where CAUSE feature includes the professional profile of the person.



### Hypothesis Generation

Here are some of the hypothesis which could influence the possibility of suicide:

**1.Depression:**

The risk of death by suicide may, in part, be related to the severity of the depression. As depression plays a major role in more than half of the suicide, so higher the depression rate higher the suicidal rate.

**2.Peer Pressure:**

Peer influence may be driving many adolescents to commit suicide. Also more teens are pressured to fit in .

**3.Gender:**

There are number of gender differences this is known as gender paradox of suicide. Males are much more likely to take their lives while women have more suicide thoughts.

**4.Age:**

Younger groups have lower suicide rates than middle aged, and older adults. As with increasing age, larger is the responsibility.

**5.Disregard for personal appearance:**

Obese or extremely skinny people are more prone to bullying or inferiority complex which may leave them highly depressed.

**6.Self-destructive behavior (alcohol/drug misuse, self-injury or mutilation, promiscuity):**

Self critical thoughts can lead to self destructive which inturn on the adverse states will may lead to suicide.

**7.Forced career choices:**

Sometimes in adolescence or in young adults, it may be the case that they are not given the freedom to choose the life they want which leads to frustration.

**8.Loss of a loved one:**

Sometimes, the loss of near and dear ones may indulge the feeling of isolation and fear in the mind which eventually leads to depression.

**9.Social media Influence:**

Internet and social media can influence suicide related behavior. As social media has a large impact on the people's mind, seeing other people happy, the viewers may start demeaning themselves. Also cyberbullying may be the major contributing factor for the depression.

## **Data Exploration**

The data which was scattered in the three folders

**1.HADS data:** It contains the questionnaire and the student’s answer to those questions which helps to calculate the HADS score which indicates the depression level in an individual.

**2.StudentsFactorsData:** It contains the general characteristics of the students which may have some influence on the depression rate for students.

**3. SuicideVictimdata :** It contains the count of the people committing suicide in different states, years , age-groups and sex distributed across the years.

In student data section, HADS data and StudentFactorsData is handled.

In Suicide victim data section, SuicideVictimData folder is handled.

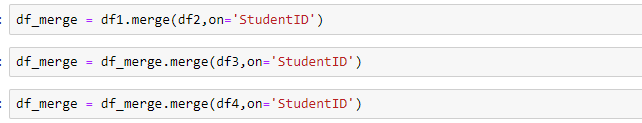
### I) Student Data

#### **Data Manipulation**

The data was in different formats (like .csv, .txt, .xlsx) so it was converted into a standard format that is all the data was converted in .csv format.

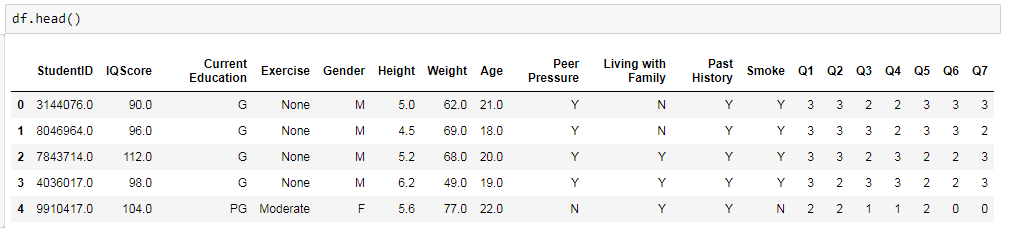
#### **Data Consolidation**

The data was scattered in various folders and files, so it was merged, to generate a single csv file.



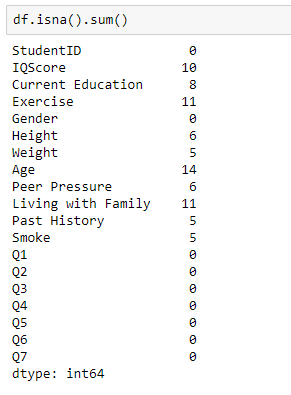
Since, StudentID was the common field in all the files, the merge was conducted on this field.

Post merging the csv file contains-



#### **Data Preparation**

##### **Missing value**



Since, the percentage of missing value of each field is considerably low i.e. IQscore:

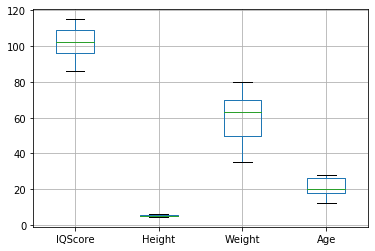
0.02%,Current Education:0.018%, Exercise: 0.025%, Height: 0.0138%, Weight:

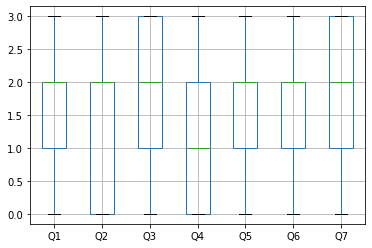
0.0115%, Age: 0.032%, Peer Pressure: %, Living with family: 0.025%. Past History:

0.0115%, Smoke: 0.0115% .

Hence these rows are removed.

##### Outliers

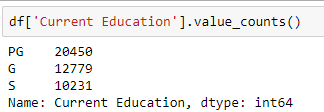
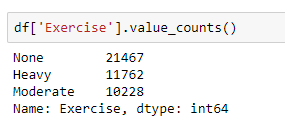


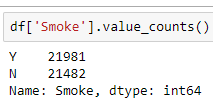
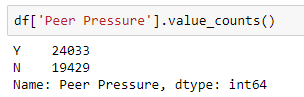


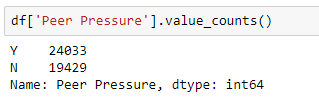
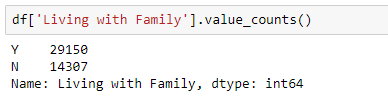
Since boxplot is used to find outliers in numeric data,it can be seen that there are

none of them.

For categorical data,







It can be seen that there are no missing values or any misspelled entries.

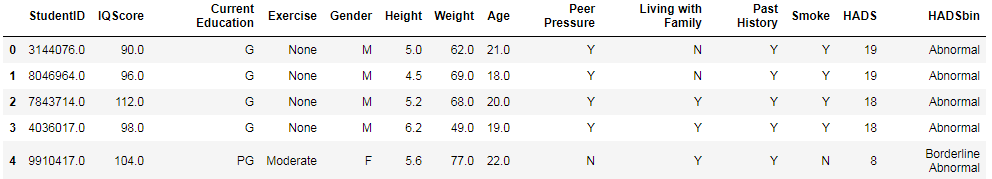


Since there exist no correlation in the data, so no variable is dropped before running the model to avoid extra noise caused by the heavy multicollinearity.

#### **4. Data Derivation**

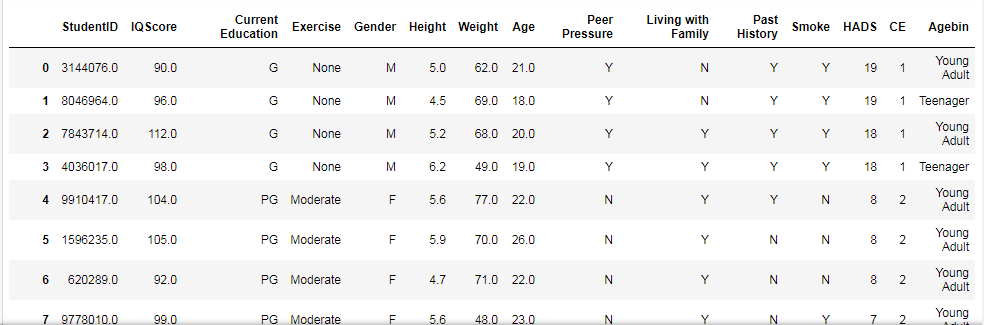
##### **HADS Score**

By calculating the sum of all the questionnaire points, HADS Score was depicted, which was further binned into three categories - Normal(<8), Borderline Abnormal(<11)`, Abnormal(11 and above).



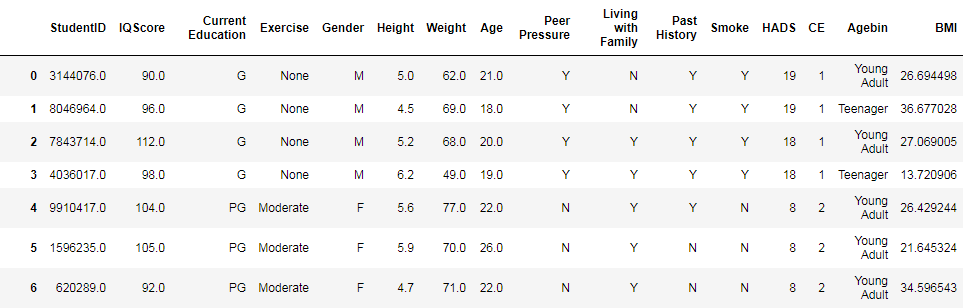
##### **Age Group**

The age filled was binned and categorised into age groups - juniors(<15), teenagers(<20), young adults(<27) and adults(27 and above).

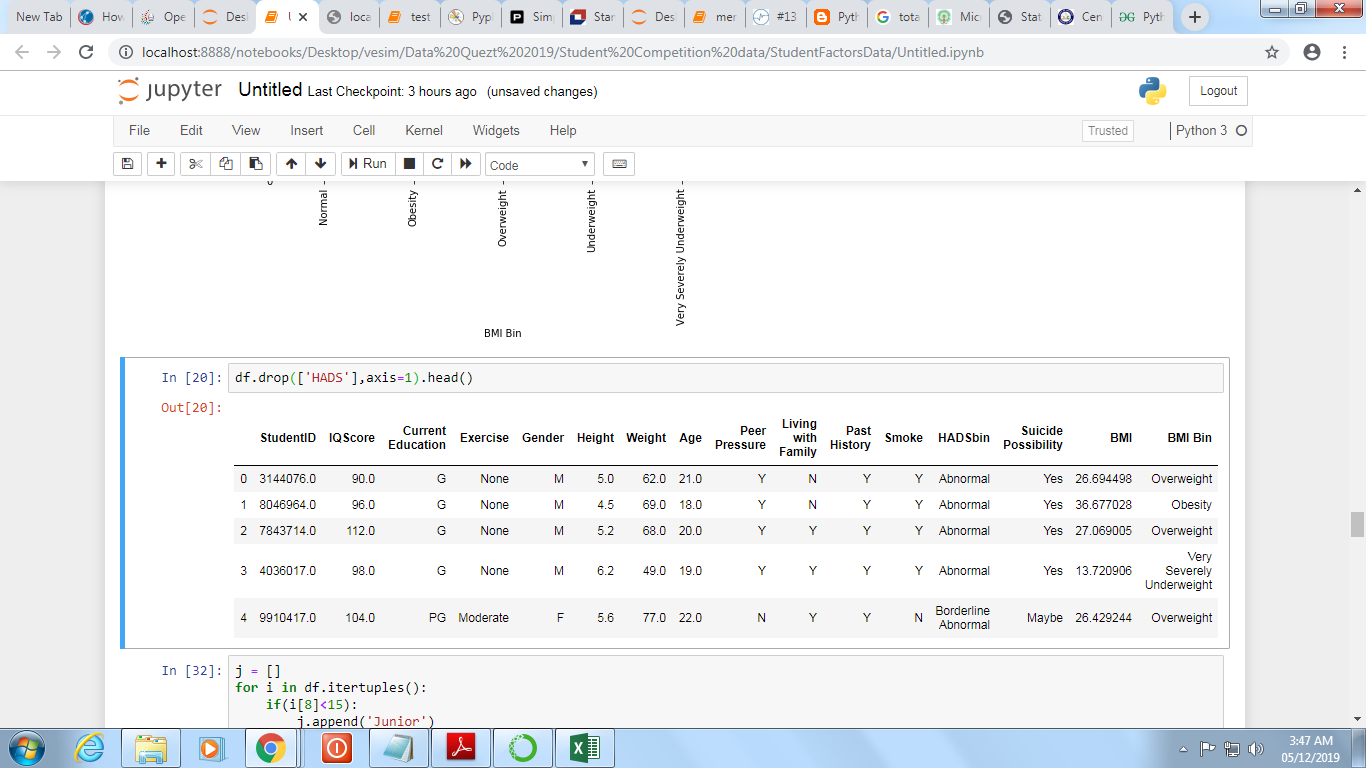


##### **BMI index**

Using height and weight fields, BMI index can be calculated using the formula weight(kg) / (height)2 (meters)

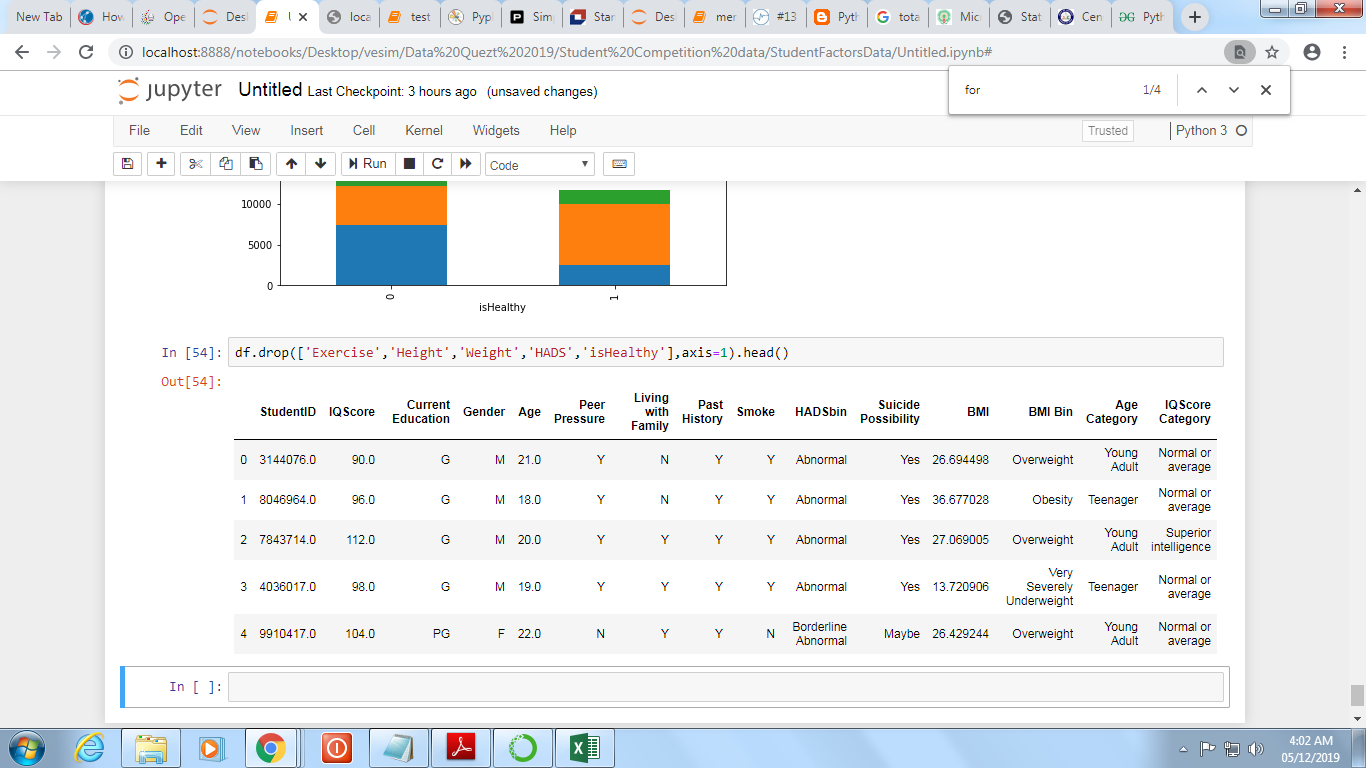


Further, this BMI index field is binned into three categories: Very Severely Underweight(<16), Underweight(<18.5), Normal(<25), Overweight(<30), Obesity(30 and above).



##### **IQ Score**

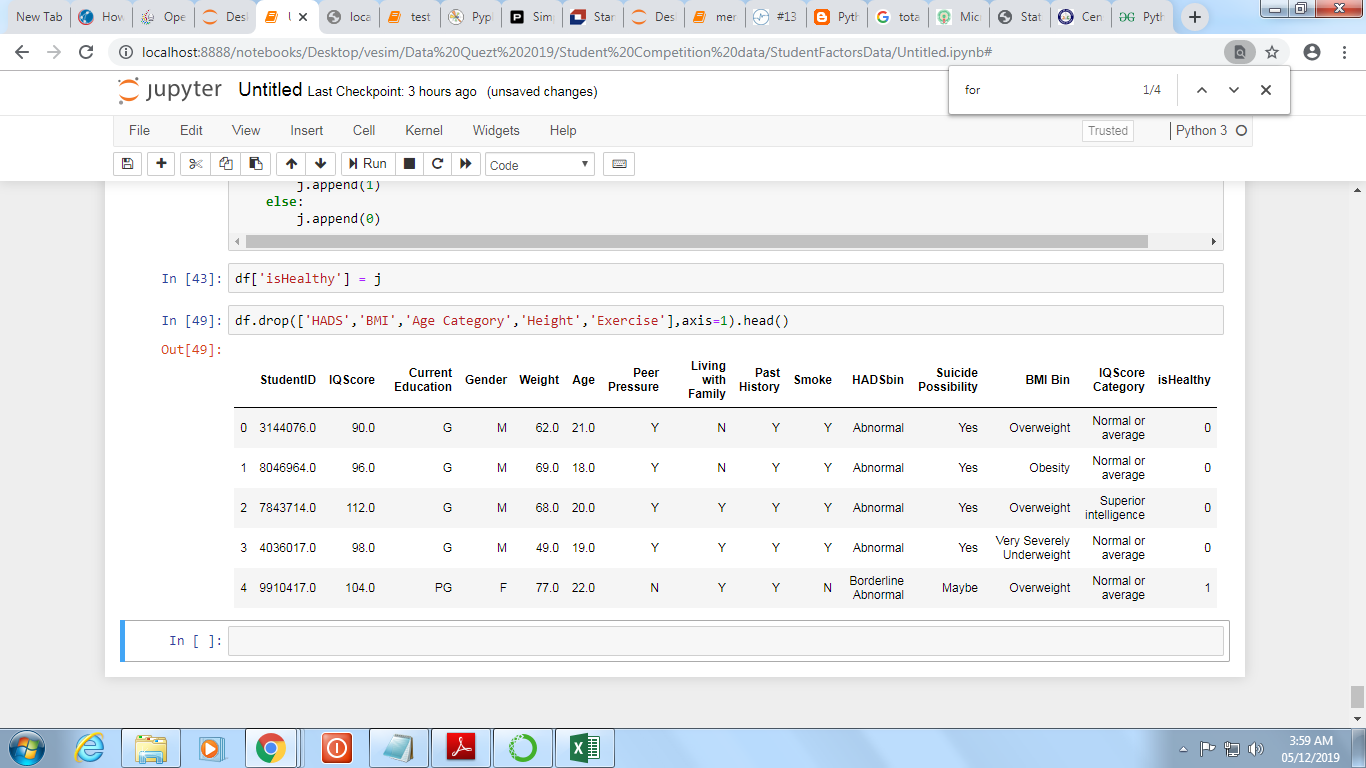
IQscore field is binned into categories: Definite feeble mindedness(<70), Borderline deficiency(<80), Dullness(<90), Normal or average(<110), Superior intelligence(<120), Very superior intelligence(<141), Genius or near genius(141 and above).



##### **IsHealthy?**

Using feature engineering, a new field was added. Depending on the values of other fields like - Smoke, BMI index, Exercise.

If the student doesnot smoke, does atleast some kind of exercise(moderate or heavy) and he/she is not Obese/Severly underweight then isHealthy is true i.e 1 or else false i.e 0.



### II) Suicide Victim Data

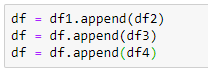
#### **Data Manipulation**

The data was in different formats (like csv, txt, xls) so it was converted into a standard format that is all the data was converted into csv format.

#### **2. Data Consolidation**

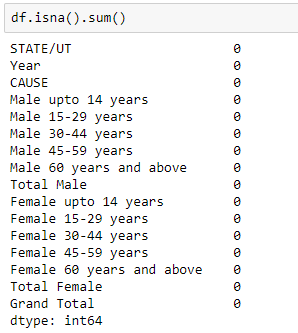
Since the data was segregated year wise, there was no need as such to consolidate it but the rows were appended to extract the yearly patterns.





#### **3.Data Preparation**

##### Missing Value



No missing values are present.

## **Data Analysis**

### I) Student Data

**MOST SUICIDES ARE CONNECTED TO DEPRESSION**

Suicide is the second leading cause of death in young people.A major cause of suicide is mental illness, very commonly depression.

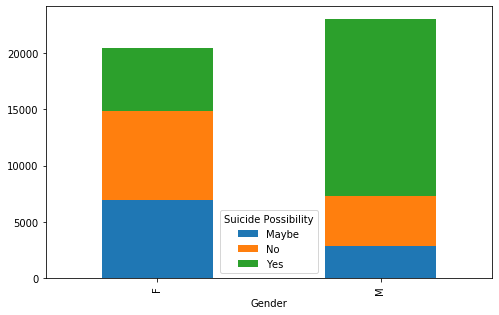
Although the majority of people who have depression do not die by suicide, having major depression does increase suicide risk compared to people without depression. The risk of death by suicide may, in part, be related to the severity of the depression.So keeping this in mind , the HADS score can be considered directly proportional to the possibility of suicide.

Various factors affect the rate of depression. Using exploratory data analysis skills we can find inferences of various factors affecting depression and suicides in turn.

#### 1**. Gender**

There are number of gender differences this is known as gender paradox of suicide. Males are much more likely to take their lives while women have more suicide thoughts.

**Fig 1.1**



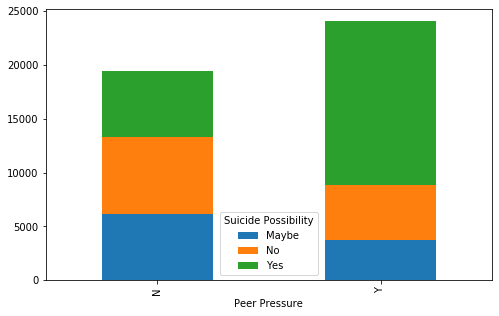
According to the fig 1.1, it can be seen that, more number of males commit suicide

compared to females.

#### **2. Peer Pressure**

When children start school, their parents begin to lose the influence they once had. By the teenage years, peers are the most influential group as teenagers navigate finding an identity and figuring out the roles that they play. Unfortunately, teenagers who choose the wrong peer groups can find themselves getting into a lot of trouble.

**Fig 2.1**



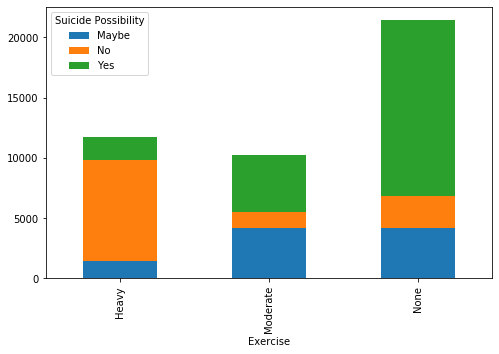
According to the graph, it can be seen that more peer pressure leads to more

possibility of suicide.

#### 3. **Exercise**

Regular exercise significantly reduces both suicidal thoughts and attempts among students who are bullied as exercise is a great way to help improve your overall mental health and ability to cope with mental illness.

**Fig 3.1**

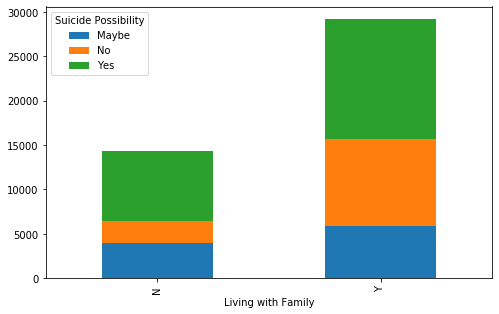


According to fig 3.1, no exercise leads to maximum possibility of suicide.

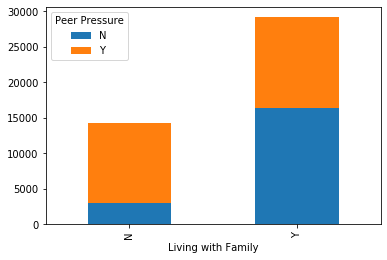
#### **4. Living with Family**

Family time is important because it allows family members feel loved and secured which reduced negative effect of the peers and also lessens the suicidal thoughts.

**Fig 4.1**



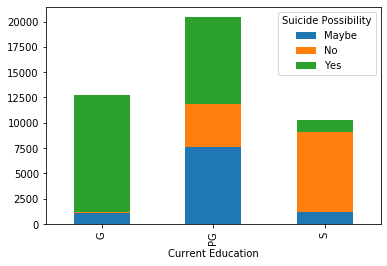
**Fig 4.2**



As seen in figure 4.1 , If the student is living with family suicidal rate is decreased also living with family decreases the peer pressure on the student as seen in fig 4.2

#### **5. Education**

Since the peer pressure is highest in adolescence and young adults , graduates have the highest chances of suicide.

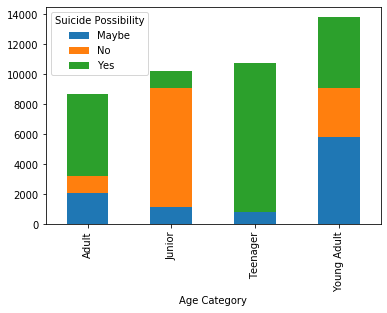


According to the fig , it can be seen that graduates have the highest possibility of suicide followed by post-graduates as they are a little mature and the least possibility is seen in school kids as they have comparitively less pressure and responsibilities.

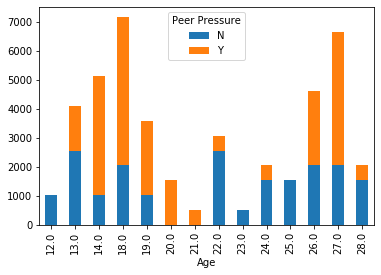
#### **6. Age**

Highest chances of suicides can be found in teenagers as they have higher peer pressure which in turn leads to depression.

**Fig 6.1**



**Fig 6.2**



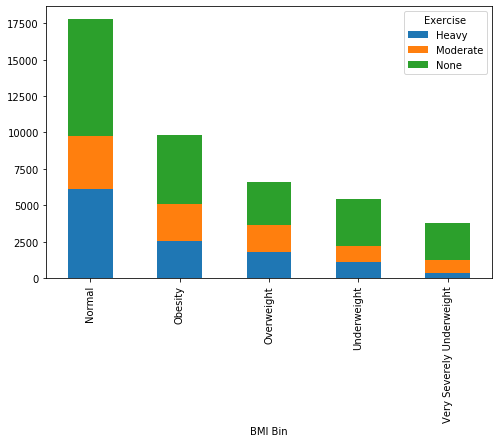
According to the figure 6.1, juniors have the least possibilities, followed by young adults, and then by adults and then the highest is found in teenagers.

From figure 6.2, it is clearly seen that peer pressure is 100% for those age 20 and 21, and comparatively lower in adults.

#### **7. BMI**

Sometimes the physical appearance can be one of the factors contributing to suicidal possibility.

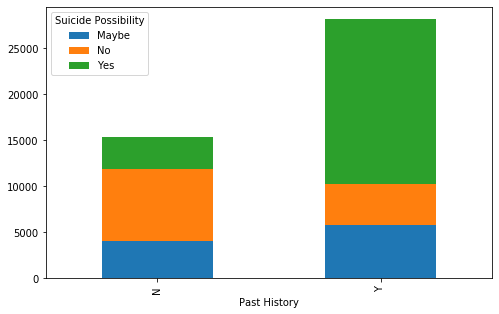
**Fig 7.1**



According to the figure 7.1, obese and very severely underweight student have highest possibility, followed by the underweight and overweight, and then the least is seen in normal.

#### **8. Past History of depression**

**Fig 8.1**

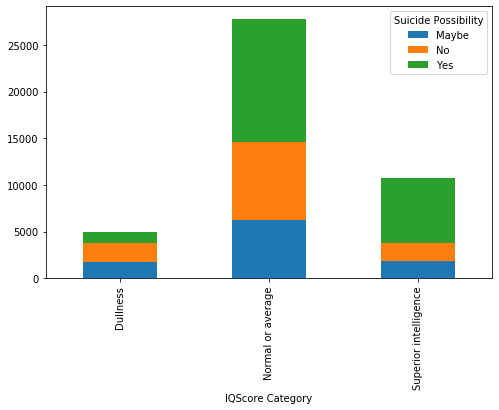


If the student has suffered from any kind of depression in the past, there are high chances for him/her to suffer from it again and then inturn lead to suicide.

#### **9. IQ**

People with sharper brains, have higher chances of commiting suicide, as superior IQs are associated with mental and physical disorders.

**Fig 9.1**

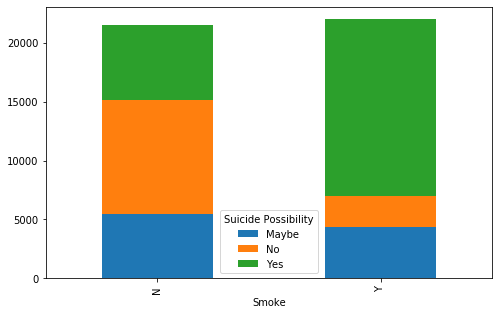


According to the figure, normal and super intelligent people have higher possibility compared to the dull people.

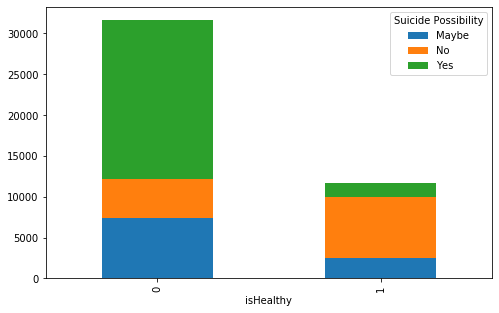
#### **10. Smoke**

Several studies reported that smoking is genrally associated with psychological disorders and high rish behaviors such as substance and alcohol abuse, sexual and physical abuse, which are considered as major cause of suicide.

**Fig 11.1**

  
 According to fig 11.1 it can be seen that if a person smokes more there is a high possibility of suicide.

#### **11. Is Healthy?**

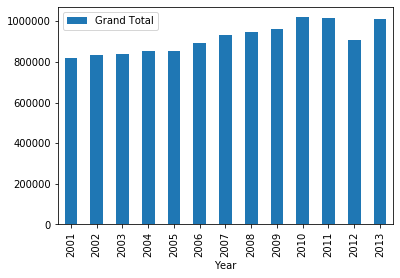


According to figure it can be seen that if a person is healthy i.e. has a good BMI , does not smoke and does enough exercise(moderate and heavy) then the person has a low possibility of suicide.

#### **II) Suicide Victim Data**

#### 1. Yearly aggregation

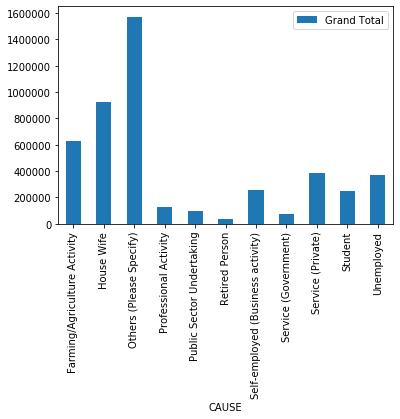
This will generate the yearly pattern of the frequency of suicide.



This graph shows, the increasing frequency of the suicide count. The frequency is linearly related to the year, with a slight decline in the year 2012.

#### **2. Cause**

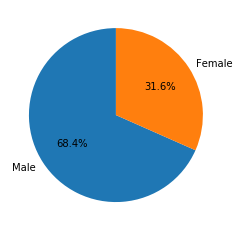
This graph will be used to generate the relationship between the cause (through the years and states) and frequency.



Looking at this graph it can be inferred that Housewives has the largest count and retired person has the lowest count of suicide.

#### **3. Gender Bias**

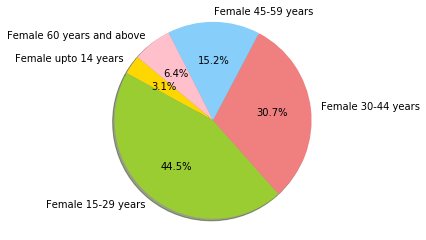
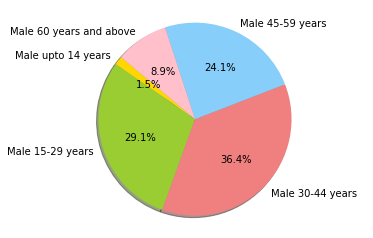
This graph will be used to find gender paradox.

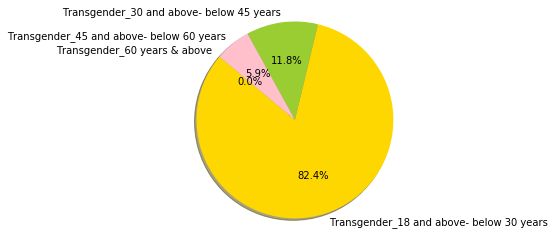


It can be depicted that, the males have almost double the possibility of suicide as that of females.

#### **4. Age factor**

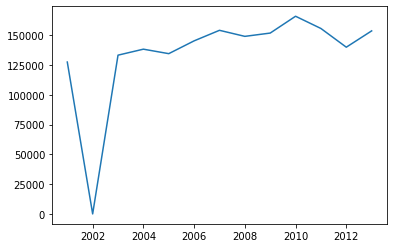
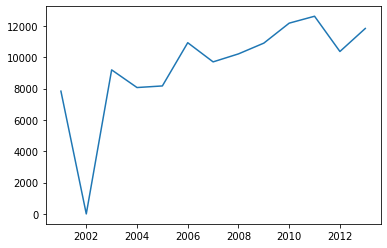
This graph can be used to find the age group of males, females and transgender commiting suicide.



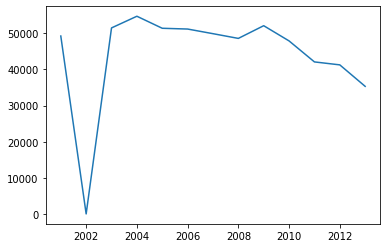
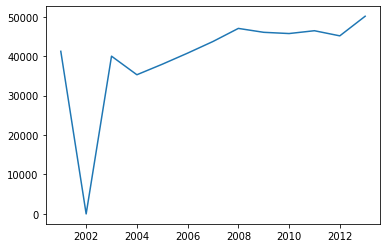


#### **5. Cause in Yearly Pattern**

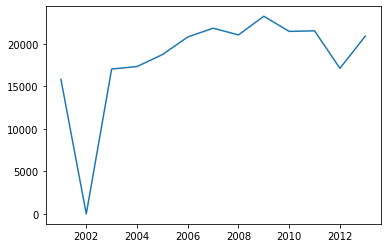
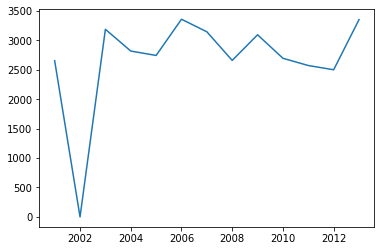
This graph will help us infer the trends in the cause year wise.



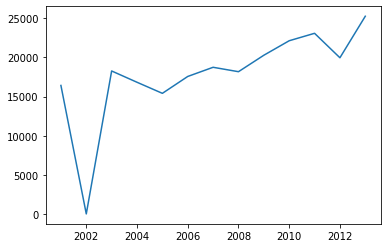
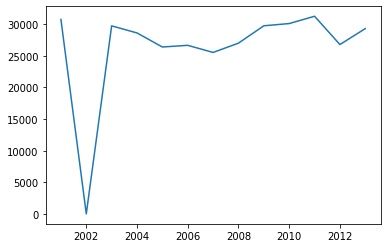
**Professional Activity Total Self employed**



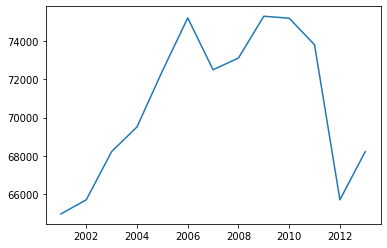
**Total Salaried Farming**

****

**Self-employed (Business activity) Retired Person**



**Unemployed**   **Student**

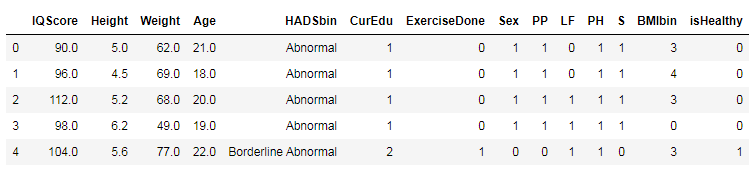


**Housewife**

All the features except Housewife have a similar rising trend, whereas trend of housewife has seen recent decline in the numbers.

## Modelling

This is the dataframe on which the modeling is done.



Here,

HADSbin is the binned HADS Score

CurEdu is corresponding values assigned to education levels as S->0, G->1, PG->2

ExerciseDone is corresponding values assigned to exercise levels as None->0, Moderate->1, Heavy->2

Sex is corresponding values assigned as M->1, F->0

PP is corresponding values assigned as Peer Pressure Y->1, N->0

LFis corresponding values assigned as Living with Family Y->1, N->0

PH is corresponding values assigned as Past History Y->1, N->0

S is corresponding values assigned as Smoke Y->1, N->0

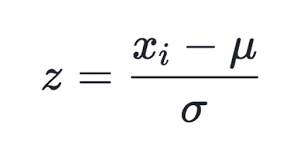
BMIbin is corresponding values assigned to BMI categories as Severely Underweight -> 0, Underweight->1, Normal->2, Overweight->3, Obese->4

(***Missing values, duplicates and outliers were already removed)***

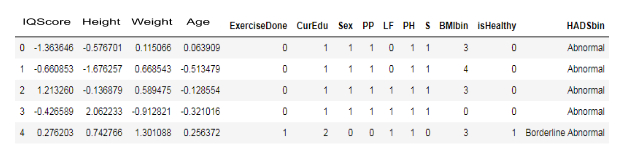
### Data Scaling:

Scaling has to be done on the numeric data which can have greater impact on its contribution to the classification models, due to its higher range of values.

Hence, Standard Scaling was applied on the columns IQScore, Height, Weight, Age.



After scaling data:



Then, the data was split into the train and test data . Following the split, following models were applied:

### Modeling and Evaluation

|  |  |
| --- | --- |
| **Model Name** | **Accuracy Score (%)** |
| Decision Tree | 85.76 |
| Extra Trees (Ensemble Technique) | 85.74 |
| Random Forest | 85.72 |
| Logistic Regression | 74.58 |
| Support Vector Machine (SVM) | 84.75 |
| K Nearest Neighbour (K=3) | 82.71 |
| Naive Bayes | 69.85 |

## Results and Conclusions

Chance of committing suicide is directly proportional to Depression.

1. Males are more probable to suicide compared to females.
2. Higher the peer pressure higher the chances to commit suicide.
3. Teenagers and adolescents i.e. students in graduation are more prone to commit suicide because of the cut-throat competition and career pressure.
4. Healthy habits like exercise and no smoking reduce the chances of depression
5. Depression is relapsing
6. Physical factors like normal height and weight reduce chances of depression
7. The number of suicides were less in the initial years and the number of suicides are increasing in recent years.
8. Housewives have the largest count followed by farmers and retired people have the lowest count of suicide.
9. So we can predict the possibility of suicide using Decision tree which has the highest accuracy for this data.

## Implications

**Society:** India is battling with high suicide rate. Basically the coward one’s take the easy way out. So if we can help such people who are going through a stressful time and has suicidal tendencies, give them some strength, it will be beneficial, Hence some helpline numbers or few centers should be opened to contact such people.

**Students:** Children and young adults are complicated emotional ecosystems that are easily affiliated by behaviour emotional learning and mental disorders.Environmental challenges lead to mental concerns as well religious discrimatination, body shaming ,low self esteem , family , financial problems and even hormonal changes may play a part. Hence it is very important for the parents to provide a healthy environment to the children constantly interrogating them about their problems and taking to the counsellors quarterly

**Government:** Governments play a vital role in this fight of preventing suicide. From creating a national strategy to restricting the most common means of suicide , for example , India and sri lanka has restricted access to pesticides locally. In order to create awareness government can organize anti-depression campaigns.

**Industry and Institutes:** Employees working in industries and institutes sometimes face a lot of work pressure, peer pressure, work abuse , inferiority complex, etc. which may lead to severe depression to prevent this offices must have a regulating body which can provide counselling and establish a friendly and conducive working environment