**Capstone Final Report**

**Anitha Pissay**

**March 2020**

**Introduction**

The **Youth Risk Behavior Surveillance System (YRBSS)** was developed in 1990 to monitor health behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. These behaviors, often established during childhood and early adolescence, include

* Behaviors that contribute to unintentional injuries and violence.
* Sexual behaviors related to unintended pregnancy and sexually transmitted infections, including HIV infection.
* Unhealthy dietary behaviors.
* Inadequate physical activity.

Use of Tobacco, Alcohol and other Drug significantly affects the youth risk behaviors. Data science utilises digital technologies and computational algorithms to effectively analyse the data in novel ways. Large data set analysis will be able to pick up important trends on the use of TAD and correlation between age groups, sex and the type of abusive substance, which are often incomprehensible by manual analysis. Understanding the trends in use of TAD helps in developing intervention measures such as change in policies, creating guidelines to schools, community awareness programs to reduce the behavioral risks at an early stage.

My area of focus will be based on

* **Tobacco use.**
* **Alcohol and other drug use.**

My objectives for the current project are

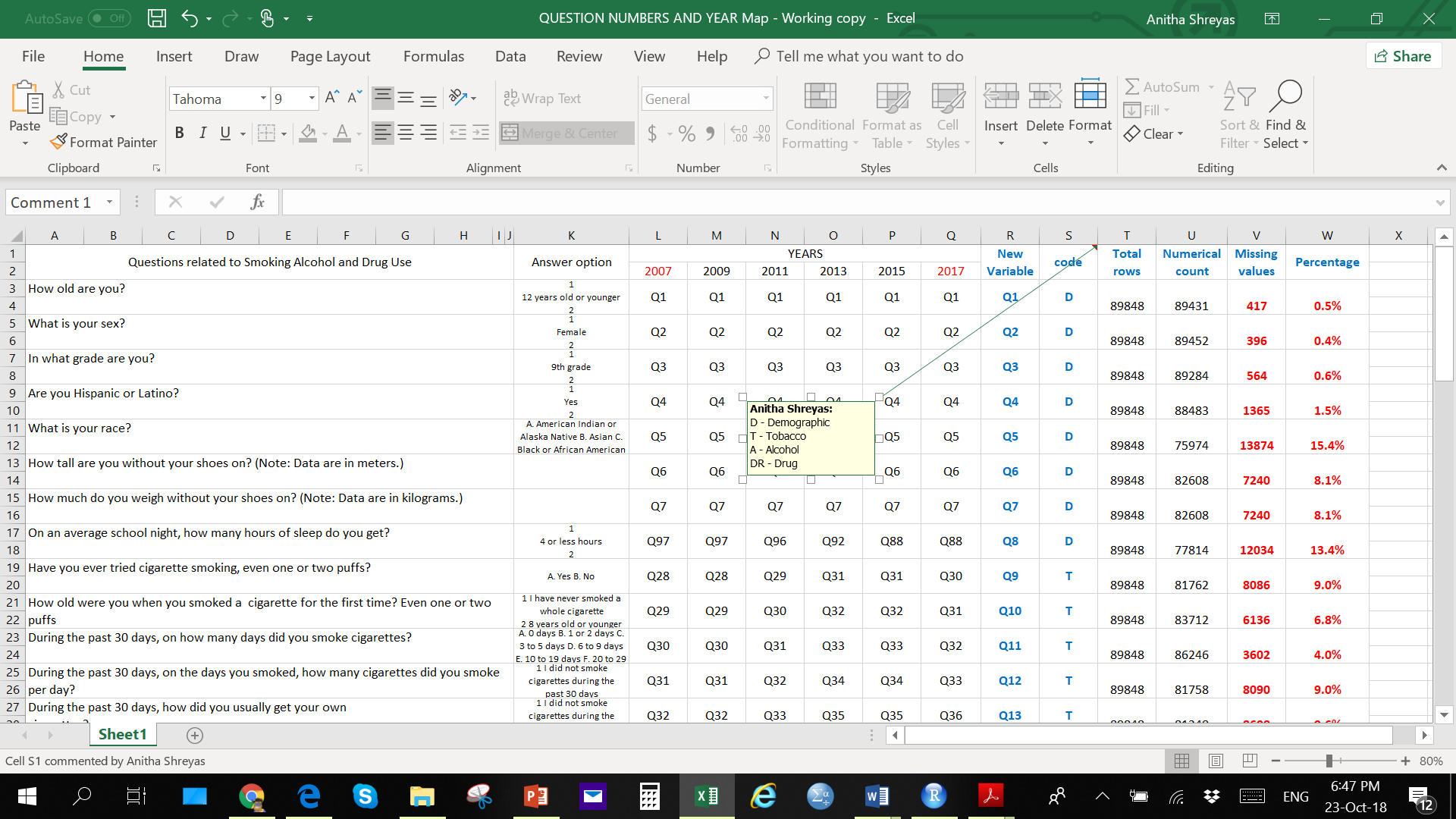
* Large data analysis using **R programming** to evaluate the trend in TAD usage among high-school students over a period of 10 years from 2007-2017.
  + Evaluation of the trends amongst
    - # Smokers vs Non-Smokers;
    - Alcohol Users vs Non-Users;
    - Drug abuse vs Non-Drug use
  + Evaluate the gender bias on the TAD usage
  + Changes in trends of TAD usage amongst
    - 9th to 12th Graders and
    - kids aged 13 years and less
  + Use of E- cigarettes
  + Trend in use of various drug substances
* Draw conclusions based on the results generated.

**Data Collection**

From 1991 through 2017, the YRBSS has collected data from more than 4.4 million high school students in more than 1,900 separate surveys. YRBSS data sources include ongoing surveys as well as one-time national surveys, special-population surveys, and methods studies. The ongoing surveys include school-based national, state, tribal, and large urban school district surveys of representative samples of high school students and, in certain sites, representative state, territorial, and large urban school district surveys of middle school students. My project primarily focuses on **Tobacco, Alcohol and other drug Use** **(TAD)** from 2007 to 2017. The Data sets were downloaded from [“Center for Disease Control and Preventions”](https://www.cdc.gov/healthyyouth/data/yrbs/data.htm) website.

**Data Wrangling**

1. My approach in solving this problem is by collating all the data collected from 2007 to 2017 from the individual years to create a single cumulative project data file with matching variables.
2. As the area of focus was narrowed to only 3 behaviors – TAD, only the responses of the questions related to them were extracted along with the demographic data such as students age, height, weight, etc.
3. The challenge was that questions asked in the survey were *numbered differently* in each data file across different years, therefore posing a greater risk in errors while merging. Hence the questions had to be visually mapped first amongst these files before merging them onto the new datafile. For this purpose, a new set of variables were created, and the responses added as per the map to ensure the accuracy of the data.



**Fig .1** A cropped screen short of the mapping file used to create the datafile

1. An extra variable (YEAR) that did not belong to any of the datafiles was created to identify the year from which the data was being extracted.
2. For ease in data readability and understanding appropriate labels were added for these values during the visualization methods and coding techniques were applied.
3. The variables are labelled **Year** and **Q1 to Q38** within which

* ‘**Year**’ identifies the year from which the responses in each row were extracted from.
* responses to questions related to demographics are in Q1 to Q8
* responses to questions related to Tobacco are in Q9 to Q19
* responses to questions related to Alcohol are in Q20 to Q25
* responses to questions related to drug use are in Q26 to Q38

Q14 and Q15 asks about electronic cigarettes also known as e-cigarette and has data only for years 2015 and 2017. The entire data was merged on an excel file which was later converted to a .csv file.

1. The final datafile contains approximately **88000 rows** with missing values in certain rows and columns for different variables. The responses are recorded as categorical variables.
2. In R studio, the csv file was imported, and data wrangling method was applied starting with replacing the NA (missing values) with ‘0’ as all the responses (except for Demographic responses) were recorded as numeral values starting with 1.
3. For ease in visualization in outputs the values of the variables were converted to labels by applying data wrangling techniques.
4. Once data is wrangled into its desired observation format, it was then concatenated to a cumulative data frame.
5. Further data wrangling techniques was used order to calculate certain statistics, such as ratios, percentages, and trend, and to create data frames for constructing the predictive and classification models.
6. A new variable “drugaddict” was created for to provide further insight into the different drugs usage amongst adolescents.

**Results:**

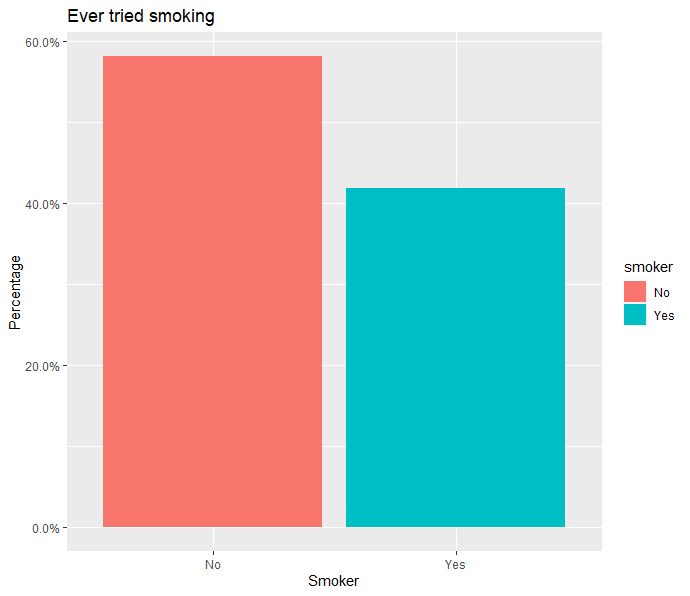
1. **Exploratory Data Analysis on TAD Use**
2. **Tobacco Use :**

**Numbers**

Based on over 80,000 data points collected over a decade, number of Smokers and Non-Smokers were analysed. All the data from 2007 to 2017 was pooled and analysed. Based on the pooled analysis, two in every five high schoolers are smokers (Table 1 and Fig 2). This corresponds to 41.9% of the total.

**Table 1.** Total number of smokers and Non-smokers

|  |  |  |
| --- | --- | --- |
| **SMOKER** | **COUNT** | **%** |
| **No** | 47535 | 58.1 |
| **Yes** | 34227 | 41.9 |



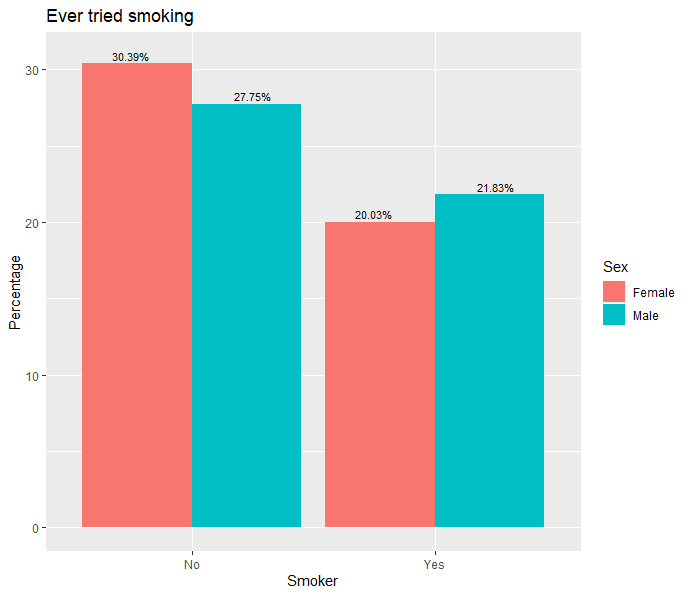
**Fig. 2**. Bar graph representing smoker and non-smoker in high   
school kids over a period of 10 years.

**Tobacco Use: Grouped by Sex**

In order to investigate whether there are differences in tobacco use among different sex, we investigated tobacco use based on gender. There is no significant difference seen in smoking habits withing sex, although number of female Smokers are less (Table 2 and Figure 3). Around 52.2% of male students were smokers.

**Table 2.** Number of male and female smokers

|  |  |  |
| --- | --- | --- |
| **SEX** | **COUNT** | **%** |
| **Female** | 16309 | 47.8 |
| **Male** | 17777 | 52.2 |



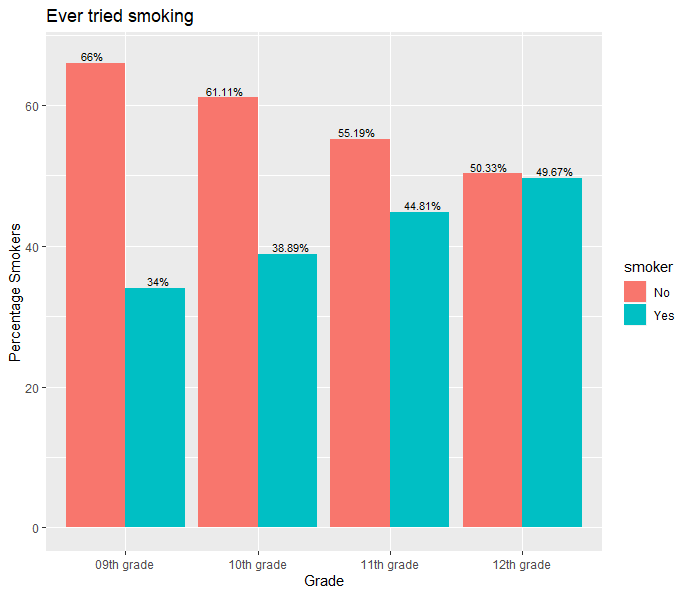
**Fig. 3.** Percentage of male and female smokers and non-smokers.

**Tobacco Use: Grouped by Grade**

Analysis of tobacco users based on the grade students are attending showed that as the grade increases the gap between Smokers and Non-Smokers in each grade is decreasing eventually with almost no difference. The 11th and 12th Graders percentageof Smokers were significantly higher compared to 9th and 10th Graders (Table 3 and Fig. 4). This showed that additional students started smoking as the they progressed to higher grades.

**Table 3.** Number of smokers in each grade.

|  |  |  |
| --- | --- | --- |
| **GRADE** | **COUNT** | **%** |
| 9th | 7045 | 20.8 |
| 10th | 7750 | 22.8 |
| 11th | 9151 | 27.0 |
| 12th | 9991 | 29.4 |



**Fig. 4.** Percentage of Smokers and Non Smokers across Grades

**Tobacco Use: Grouped for 13 Years and below**

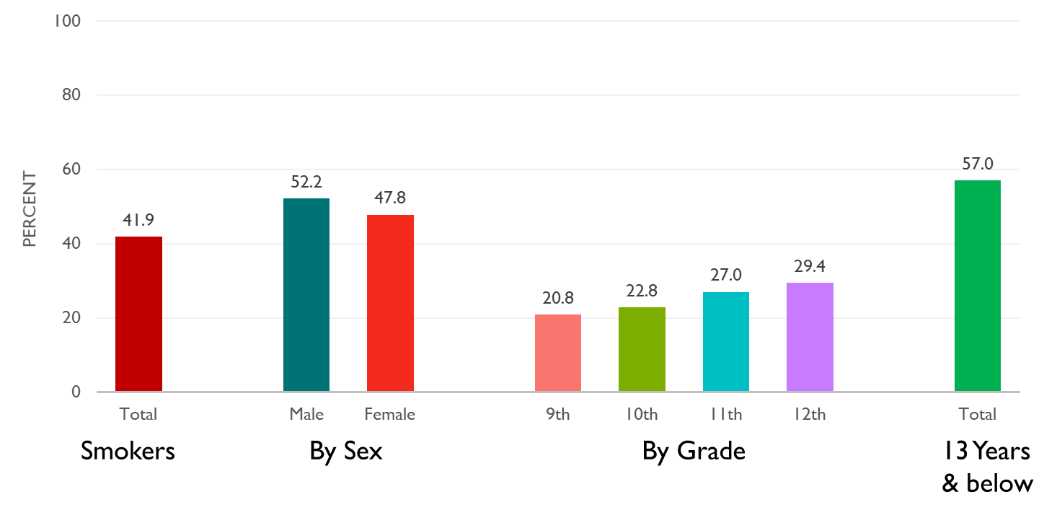
Limited data was available for tobacco use among age group 13 year or below. In this age group, there is an alarmingly higher percentage of Smokers (Table 3 and Fig. 4). Further, data analysis grouped by sex in this age groups showed a higher percentage of male Smokers compared to female Smokers (Fig 6). The number of data collected is low, hence distinct conclusions would not be ideal. Although, more preventive actions by educating the adverse effects should be carried out at the early age.

**Table 4.** Number of Smokers aged 13 years and below

|  |  |  |
| --- | --- | --- |
| **SMOKER** | **COUNT** | **%** |
| No | 39 | 43 |
| Yes | 51 | 57 |

|  |  |
| --- | --- |
|  |  |
| **Fig. 5.** Percentage of Non-Smokers vs Smokers | **Fig. 6.** Percentage of Non-Smokers vs Smokers by sex |

**Overall Graphical summary of tobacco use in High School Students**



**Fig. 7** Summary of Tobacco use, based on pooled samples from 2007-2017.

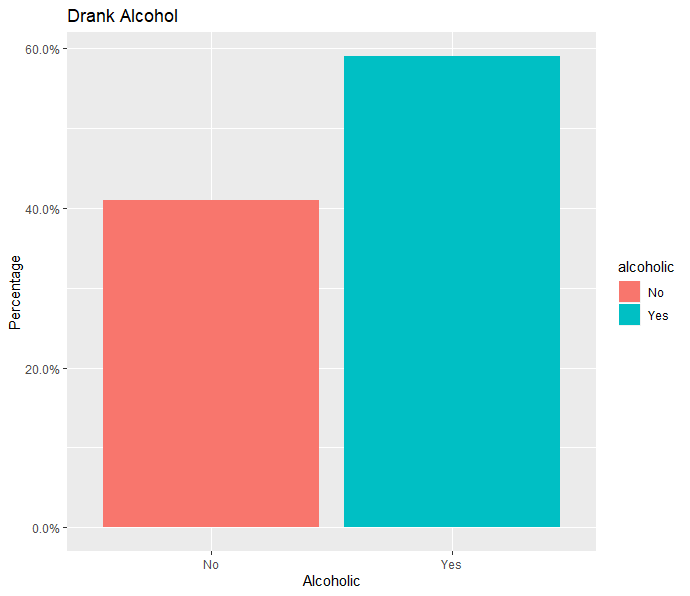
1. **Alcohol Use**

**Numbers**

Pooled analysis for Alcohol use was carried out for the data collected over a decade from 2007-2017, which included over 80,000 data points. Similar to Tobacco use, 2 in 5 students are Alcohol users. Fortyone percent of students analysed had used Alcohol (Table 5 and Fig. 8).

**Table 5.** Total number of Alcoholics vs Non-Alcoholics

|  |  |  |
| --- | --- | --- |
| **ALCOHOLIC** | **COUNT** | **%** |
| No | 50469 | 59 |
| Yes | 35062 | 41 |



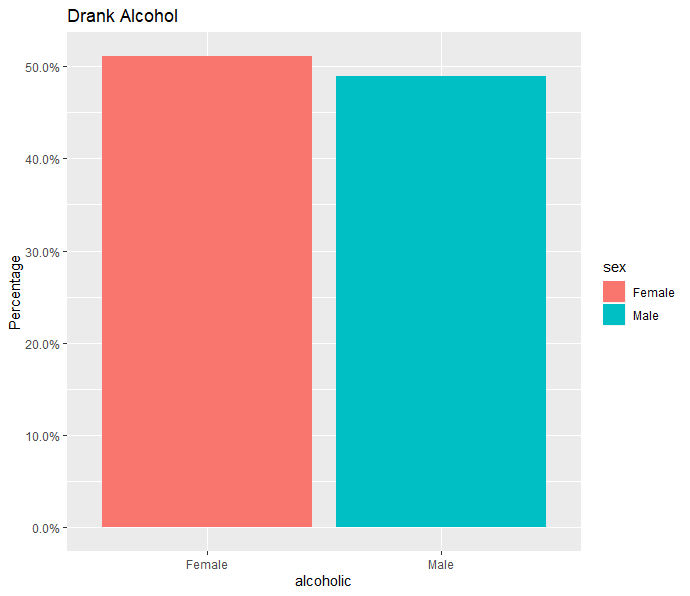
**Fig. 8**. Bar graph representing Alcoholics and Non-Alcoholics in high   
school kids over a period of 10 years.

**Alcohol Use: Grouped by Sex**

There is no significant difference seen in Alcoholic habits within sex, although number of female Alcohol users are marginally higher (Table 6 and Fig. 9).

**Table 6.** Total number of Alcoholics vs Non-Alcoholics

|  |  |  |
| --- | --- | --- |
| **SEX** | **COUNT** | **%** |
| Female | 25684 | 51.1 |
| Male | 24566 | 48.9 |



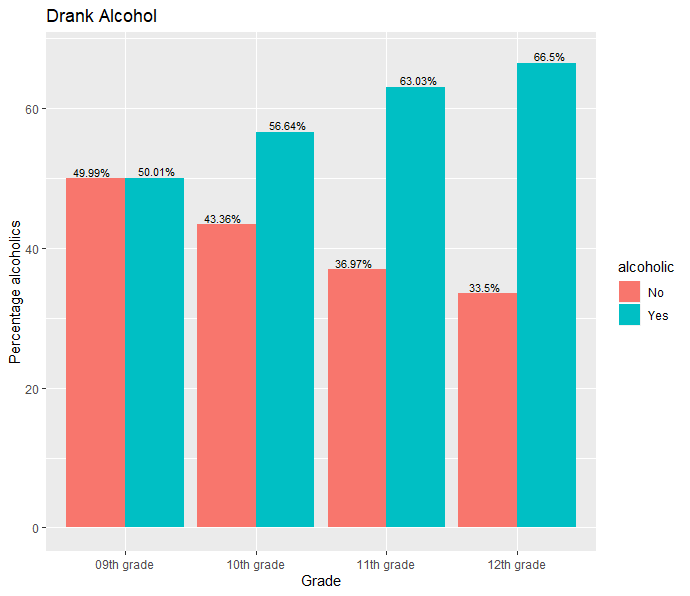
**Fig. 9**. Percentage of male and female Alcoholics

**Alcohol Use: Grouped by Grade**

When we grouped the alcohol users by grde, we noticed that higher the Grade, higher the percentage of Alcohol use. This was similar to Tobacco use. As the grade increases the gap between Alcoholics and Non-Alcoholics in each grade is increasing, suggesting more number of students started using Alcohol, The lowest Alcohol use percentage was 21.8% in grade 8 and the highest being 27.7 % in grade 12 students (Table 7 and Fig. 10).

**Table 7.** Number of Alcoholics in each Grade

|  |  |  |
| --- | --- | --- |
| **GRADE** | **COUNT** | **%** |
| 9th | 10919 | 21.8 |
| 10th | 11785 | 23.5 |
| 11th | 13483 | 26.9 |
| 12th | 13892 | 27.7 |



**Fig. 10.** Percentage of Alcoholics and Non-Alcoholics across Grades

**Alcohol Use: Grouped for 13 years and below**

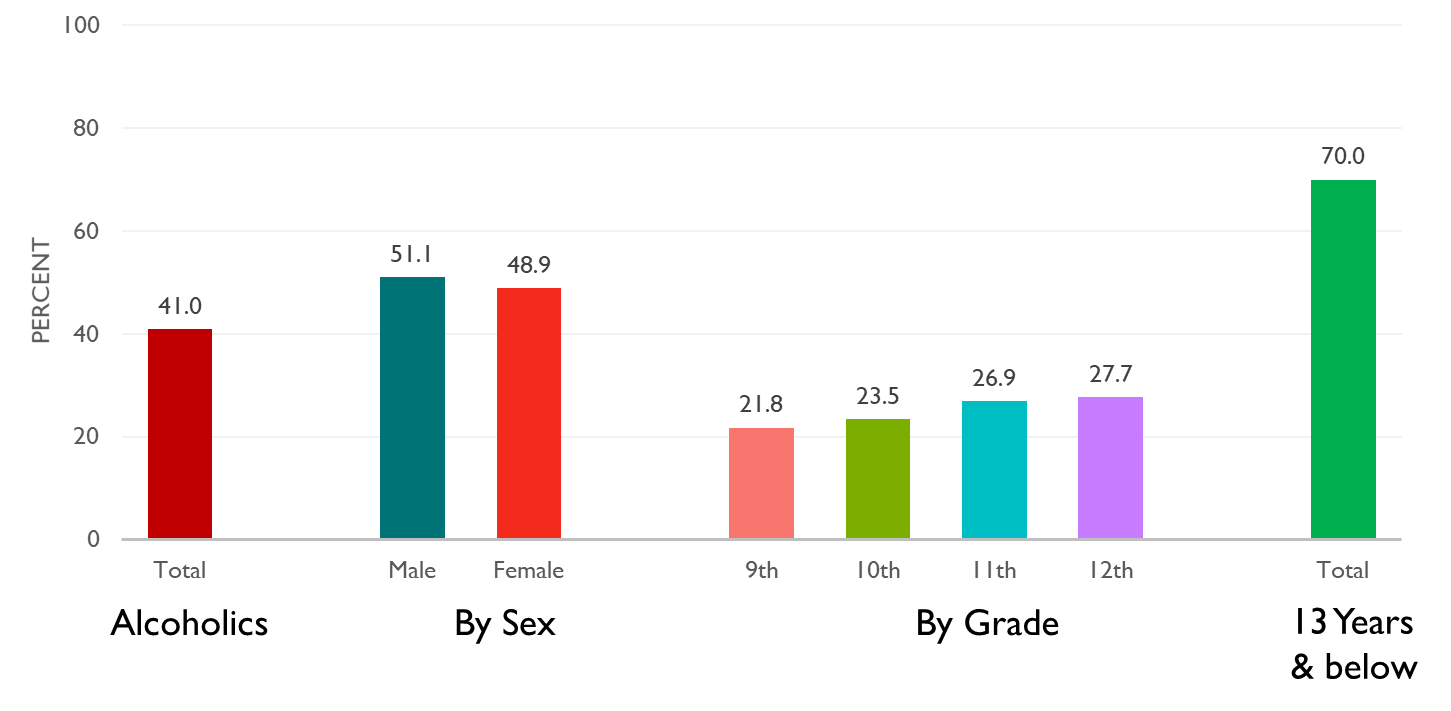
When grouped for Alcoholics and Non-Alcoholics below the age of 13, there is an alarmingly higher percentage of Alcoholics of 70% in this age group (Table 8 and Fig 11). When Alcoholics and Non-Alcoholics at the age of 13 year and below were grouped based on sex, there is a higher percentage of male who use alcohol compared to females (Fig 12). The number of data collected is low, hence distinct conclusions would not be ideal. This behaviour needs to be nipped at the bud to see positive change at a higher age group.

**Table 8.** Number of Alcoholics aged 13 Years and below

|  |  |  |
| --- | --- | --- |
| **ALCOHOLIC** | **COUNT** | **%** |
| No | 38 | 30 |
| Yes | 89 | 70 |

|  |  |
| --- | --- |
|  |  |
| **Fig. 11.** Percentage of Non-Alcoholics vs Alcoholics | **Fig. 12.** Percentage of Non-Alcoholics vs Alcoholics by sex |

**Overall summary of ALCOHOL USE in High School Students**

****

**Fig. 13**. Summary of Alcohol use, based on pooled samples from 2007-2017.

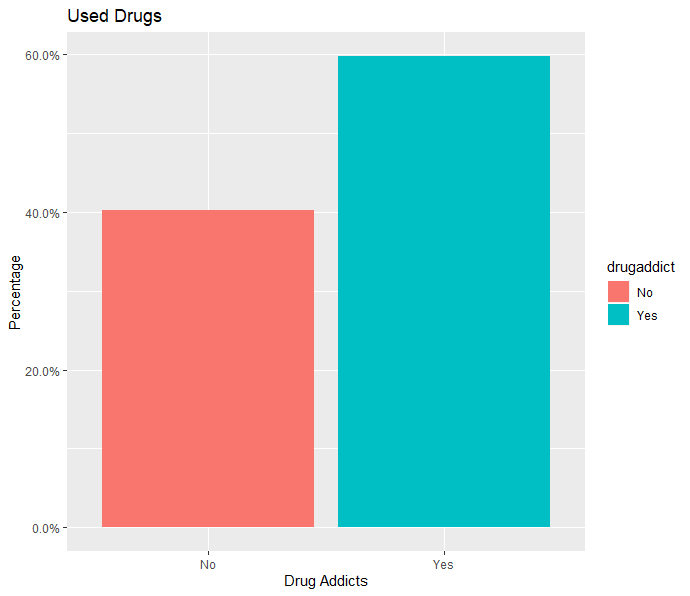
1. **Other Drug Use**

**Numbers**

Other than Tobacco and Alcohol use, data was also collected for other Drug use. Data from 2007 -2017 was polled and analysis was carried out similar to data analysis in Tobacco and Alcohol use. Compared to Tobacco an Alcohol use the number of Drug users were significantly higher. Nearly 60 % of the students, which is three in every five students, are Drug uses (Table 9 and Fig. 14).

**Table 9.** Total number of Drug Users and Non-Drug Users

|  |  |  |
| --- | --- | --- |
| **USED DRUG** | **COUNT** | **%** |
| No | 31116 | 40 |
| Yes | 46215 | 60 |



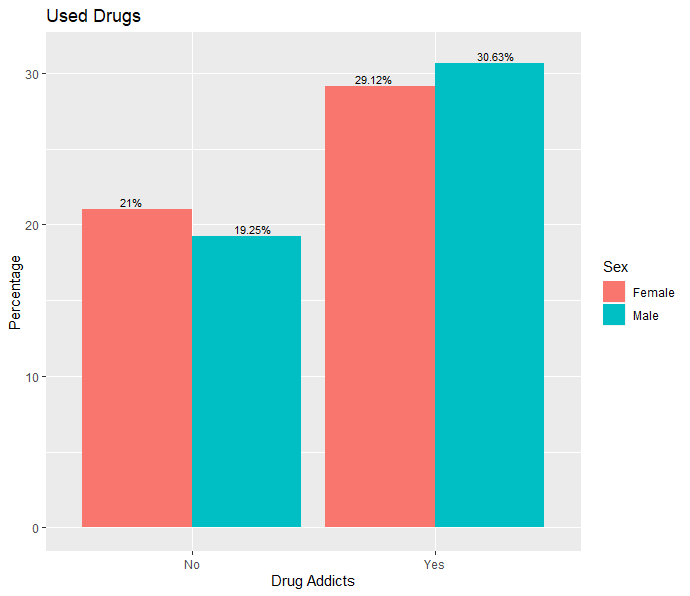
**Fig. 14**. Bar graph representing Drug Users and Non-Drug Users in high   
school kids over a period of 10 years.

**Other Drug Use: Grouped by Sex**

As seen in Tobacco and Alcohol use, no difference in the proportions of male and female Drug users. The percentage of Drug Users are almost the same in both sex (Table 10 and Fig. 15).

**Table 10.** Total number of Drug Users and Non-Drug Users

|  |  |  |
| --- | --- | --- |
| **SEX** | **COUNT** | **%** |
| Female | 22412 | 48.7 |
| Male | 23579 | 51.3 |



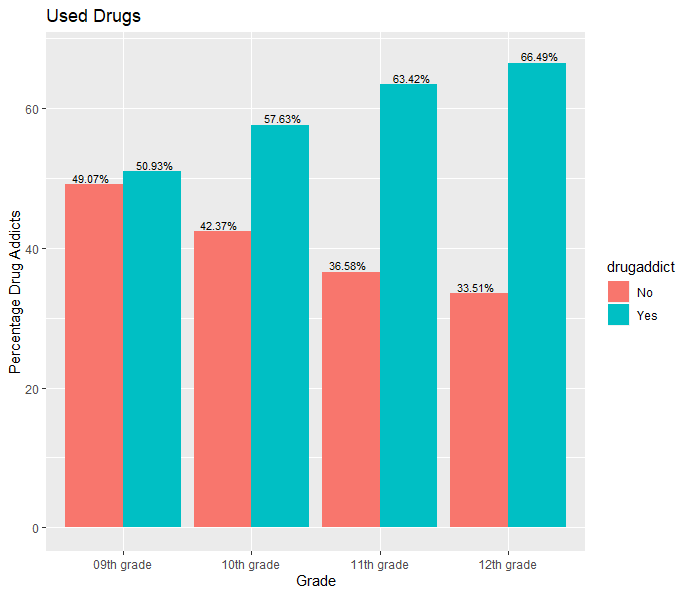
**Fig. 15**. Percentage of male and female Drug and Non-Drug Users

**Other Drug Use: Grouped by Grade**

When we analysed the data for Drug use based on the grades, similar to Tobacco and Alcohol use, as the grade increases the number of students using Drug also increased significantly (Table 11 and Fig. 16). The percentage of Drug users increased from 21.1% in 9th grade to 28.2% in 12th grade students.

**Table 11.** Number of Drug Users in each Grade

|  |  |  |
| --- | --- | --- |
| **GRADE** | **COUNT** | **%** |
| 9th | 9660 | 21.1 |
| 10th | 10783 | 23.5 |
| 11th | 12437 | 27.2 |
| 12th | 12928 | 28.2 |



**Fig. 16**. Percentage of Drug Users vs Non-Drug Users across Grades

**Other Drug Use: Grouped for 13 Years and below**

When grouped for Drug Users of all kind for students aged 13 year and below, there is an alarmingly higher percentage of Drug users (>80%) (Table 12 and Fig. 17). When Drug Users below the age of 13 are grouped for sex, there is a higher percentage of male Drug users as compared to female Drug users (Fig. 18). Once again, as in the case of Tobacco and Alcohol users, the number of data collected for this group is low, hence distinct conclusions would not be ideal. A greater awareness campaign should be implemented to see changes.

**Table 12.** Number of Drug Users aged 13 Years and below

|  |  |  |
| --- | --- | --- |
| **USED DRUG** | **COUNT** | **%** |
| No | 22 | 17 |
| Yes | 107 | 83 |

|  |  |
| --- | --- |
|  |  |
| **Fig. 17.** Percentage of Non-Drug Users vs  Drug Users | **Fig. 18.** Percentage of Non-Drug Users vs  Drug Users by sex |

**Overall summary of Other DRUG USE in High School Students**

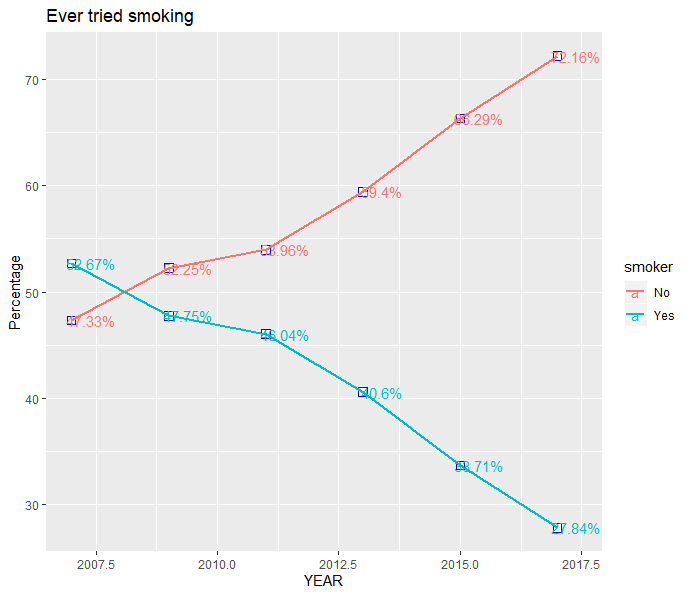


**Fig. 19**. . Summary of Other Drug Use, based on pooled samples from 2007-2017.

**II. Time series analysis from 2007 -2017**

1. **Tobacco Use**

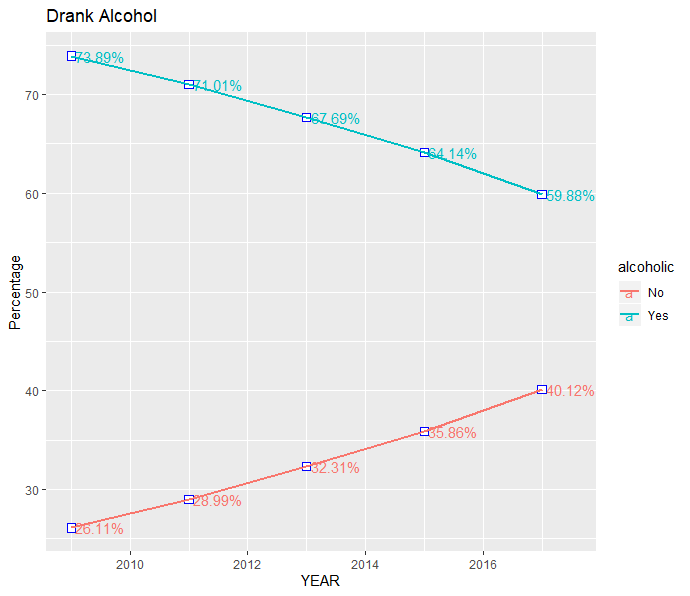
In order to evaluate the patttern of Tobacco users numbers over a period of 10 years, we plotted a time series graph. The smoker pattern over ten years shows a decline in percentage on Smokers from 53% in 2007 to 28% in 2017. The time series trend also show a good increase in percentage of Non-Smokers from 47% in 2007 to 72% in 2017. This can suggest that effective measures and steps taken during this period to bring in awareness about the bad impacts of smoking might be showing a positive effect (Fig. 20).



**Fig. 20**. Time Series plot for Tobacco Use 2007 to 2017

1. **Alcohol Use**

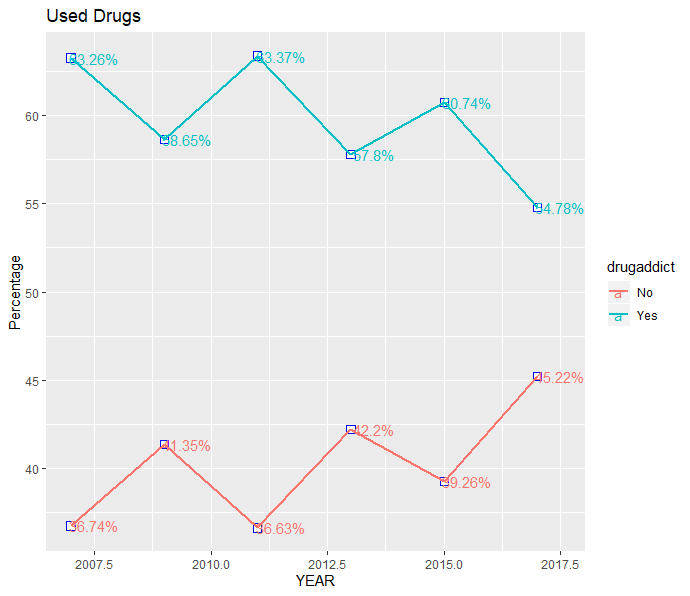
Similar to Tobacco use, the Alcohol use pattern over ten years shows a decline in percentage from 74% in 2007 to 60% in 2017 (Fig. 21). The time series trend also show a good increase in percentage of Non-Alcoholics from 26% in 2007 to 40% in 2017. Though there is a decline in the Alcoholics numbers over the years but there is still a significant high percentage of Alcoholics and we need effective measures to reduce this gap.



**Fig. 21**. Time Series plot for Alcohol Use 2007 to 2017

1. **Other Drug Use**

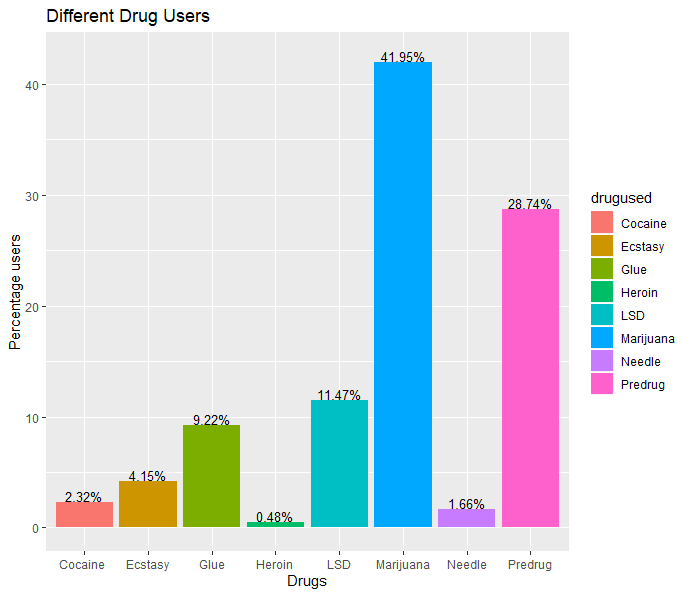
Studying the Other Drug users pattern over the years does not show any direction percentage increase or decrease. But when compared between 2007 and 2017 there is a drop in number of users from 63% to 55% (Fig. 22). Whether this drop in numbers is significant needs to be investigated.



**Fig. 22**. Time Series plot for Other Drug Use 2007 to 2017

**III. Usage of Different drugs types**

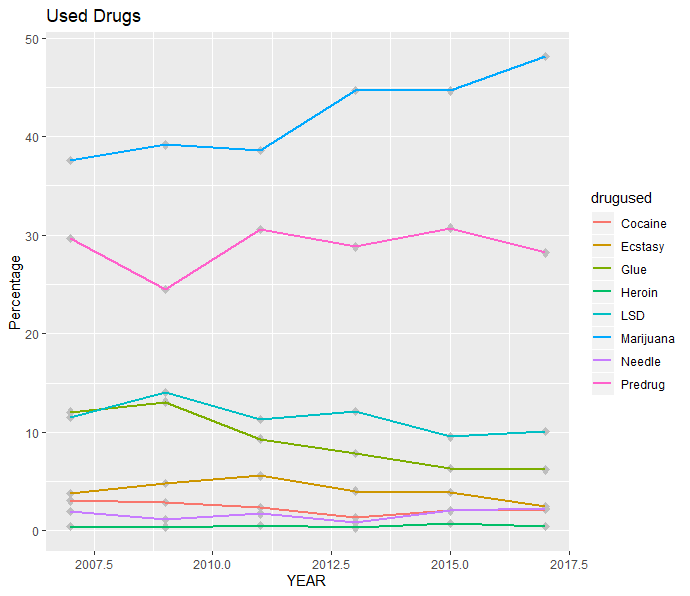
Data was also collected for different drug types used. The graph clearly shows that Marijuana is the most popular kind of substance with 42% amongst the youth followed by prescription drugs not prescribed by the doctors which is 29% (Fig. 23). The least popular substance is Heroin with less than 0.5% users. This may be due to the availability and cost factors.



**Fig. 23**. Percentage of Different types of Drugs used

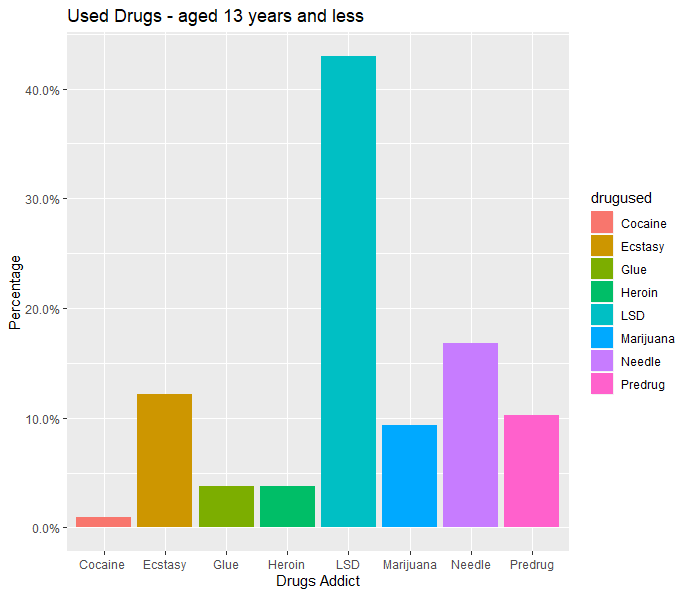
We analysed the pattern of various drug use by carrying out time series plot analysis (Fig. 24). Studying the different drug users pattern over the years we can conclude the following

* The Percentage of Marijuana users have increased over the years from 38% in 2007 to 48% in 2017.
* There is no change in the users of Heroin.
* The percentage of Glue users has decreased from 12% in 2007 to 6% in 2017.
* No significant difference seen in other kinds of drugs over the years.



**Fig. 24**. Time Series plot for different types of drugs used 2007 to 2017

Contrary to the overall popular drug type Marijuana among the high schoolers the 13 year olds and less, finds popularity in LSD [Lysergic acid diethylamide] a hallucinogenic drug with a high percentage of users of 46% (Fig. 25).



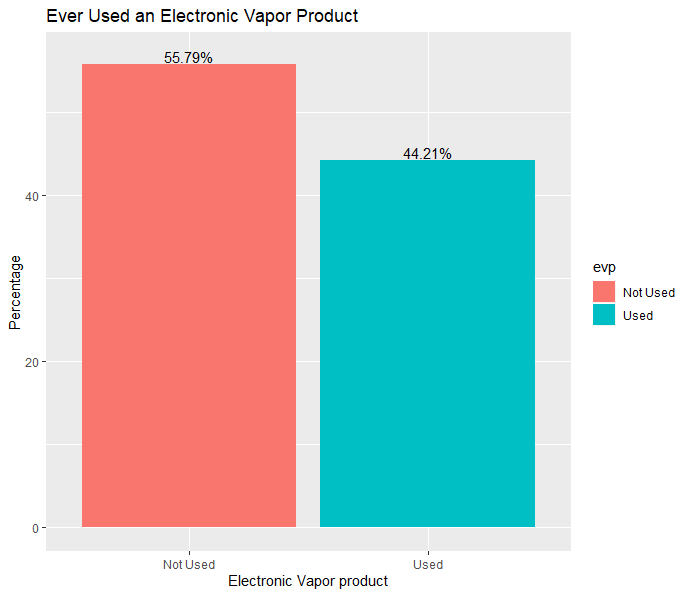
**Fig. 25**. Time Series Popular Drug type in students Aged 13 years and below

**IV. Electronic Vapor product users – E-cigarettes (2015 & 2017)**

The new fad in recent years which is taking over the teenagers is the vaping plague. The data exists only for 2015 and 2017. There is a high percentage of users of this product which is an area of concern (Table 13 and Fig. 26).

**Table 13.** Total number of E-Cigarette Users and Non-Users

|  |  |  |
| --- | --- | --- |
| **E-CIGARETTE** | **COUNT** | **%** |
| NOT USED | 16044 | 55.79 |
| USED | 13360 | 44.21 |

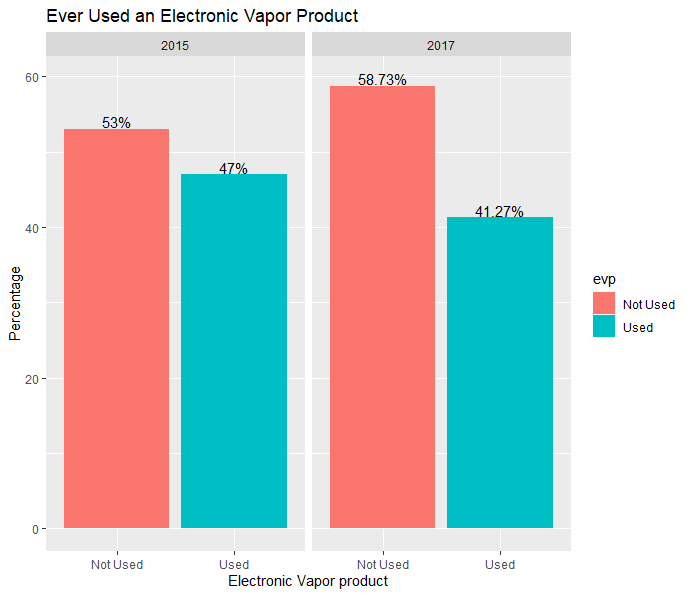


**Fig. 26**. Bar graph representing E-Cigarette Users and Non-Users in high   
school kids from 2015 and 2017

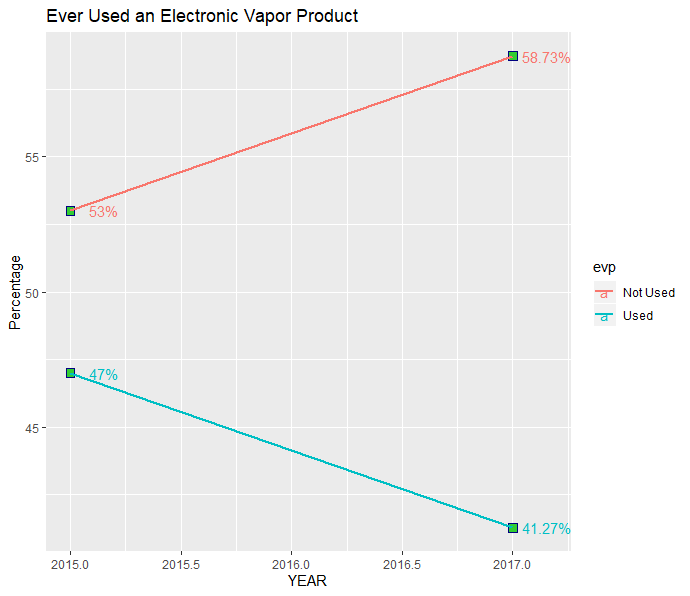
The E-Cigarette users have dropped by 6% between 2015 and 2017 (Table 14 and Fig. 27). The data and also the timeline is insufficient to draw any conclusions. We need to gather more information for further investigation.

**Table 14.** Percentage Users of Electronic Vapor Product in Both 2015 and 2017

|  |  |  |  |
| --- | --- | --- | --- |
| **YEAR** | **E-CIGARETTE** | **COUNT** | **%** |
| 2015 | NOT USED | 7998 | 53 |
| USED | 7093 | 47 |
| 2017 | NOT USED | 8406 | 59 |
| USED | 5907 | 41 |



**Fig. 27**. Bar Graph of percentage users of Electronic Vapor Product in 2015 and 2017



**Fig. 28**. Time Series plot for E-Cigarettes used in 2015 and 2017

**Discussion:**

Based on the data analysis for the data collected over a period 2007-2017, 2 in 5   
high schoolers were smokers and alcoholics. The number of drug abuse was higher with 3 out of 5 students which is signifcantly higher than the Tobacco and Alcohol use. No significant differences in TAD users grouped by Sex, was noticed. In general, as the grade increased the number of TAD users also increased. In Students below 13 years, TAD usage was higher, although the number of data points were low, which suggests insufficient data in order to conclude the impact. Analysis of Time Series showed that there is a significant decrease in TAD users over time. Marijuana was the primary choice of drug users compared to other substances. Whilst, data for E-Cigarettes was only collected from 2015 & 2017, Yet 44% of students used E-cigarettes.

CDC findings for Adult TAD users was used to draw comparisons. Among high schoolers smoking was significantly higher (42%) than adults (14%). Sixteen in every 100 males and twelve in 100 females were smokers, whilst no differences based on sex was noticed in high school students. In general, as the grade increased the number of TAD users also increased among high school students. Whereas in adults, younger (18-24 yr) and older (>65yr) age groups had lower percentage of smokers when compared to middle age (24-65 yr) persons. Two in five students were Alcoholics, compared to one in four in adults. Similar to high-schoolers, adults also showed higher consumption of alcohol until the age of 45(25-35%), it reduces to 22% for age groups of 45-65 years and significantly lower percentage (8%) consumed alcohol at 65. Overall Time Series Analysis (2007 -2017) showed that there is a significant decrease in TAD users over time.

**Recommendations:**

* The Lack of data on E-Cigarettes, limited to just 2 years, showed a declining trend (May be inaccurate) . Hence More data needs to be collected for new trends in E-cigarettes due to its popularity in adolescents in recent years. This will help in better prevention and preparedness methods.
* TAD usage in students under 13 years old, is alarmingly high. This could be due to smaller data points. However, early intervention is necessary to educate on adverse effects of TAD use, to create awareness and inculcate good habits.
* Drug use was significantly higher than the Tobacco and Alcohol use. This is worrisome, given the severity of impact on health due to addiction. Hence, schools and communities should conduct awareness campaigns for not only students but also parents.
* Prevention is better than Cure