

A Case Study Of Analyzing 2013 Chicago Youth Health Risk Behavior Data By Machine Learning With CRISP-DM Methodology

Submitted as final project fulfillment of Course: ISTE790 Data Analytics for Emerging Technologies

Group: 5

(Saeed Albarhami, Abdullah Hussein)

27 October 2019

Abstract

The objective of this paper is to use data mining techniques to analyze the 2013 Chicago youth risk behavior surveillance survey data, Collected by the centers for Disease Control and Prevention in USA, to investigate interesting relations and patterns in collected data, and provide a recommendation based on findings that been discovered by applying machine learning algorithms such as priori decision tree... etc. following CRISP-DM steps methodology, starting by business understanding, data understanding, data preparation, modeling, and evaluation. R open-source software will be used among all the process steps.

Contents

1	Executive Summary.	2
2	Introduction.	3
2.1	Background.	3
2.2	Business Challenges.	3
3	Business Understanding:	4
3.1	Objective:	4
3.2	Motivation:	4
3.3	Data Description:	4
4	Data Understanding.	5
4.1	Metadata Description.	5
4.2	Loading, Retrieving, Viewing Data.	6
4.3	Exploratory Data Analysis.	6
4.4	Demographic Variables Description	8
5	Data Preprocessing (Demographics)	8
5.1	Creating A Custom R Function To Handle Missing Data For Specific/All Observations.	8
5.2	Data Distribution Overview (Demographics).	9
5.3	Data Overview (Demographics).	9
5.4	Statistical Summary Of Grouped Data (Demographics).	9
5.5	Data Visualzation.	10
6	Data Preparation.	14
6.1	Cleaning Data Using “omitNaForSpecificCols” Custom Function.	14
6.2	Correlation Matrix Of Variables.	15
6.3	Data Aggregation.	17



Figure 1: test

7	Modeling.	19
7.1	Association Rules Algorithms (Apriori & Eclat).	19
7.2	Decision Tree.	22
8	Conclusion.	25
9	References	25
10	Appendices	26

1 Executive Summary.

- The main objective of this project is to examine the associations between victimization, substance use, and suicide attempt among youth using the Youth Health Risk Behavior Survey data for the year 2013.
- The data used consists of information from 1581 respondents, consist of Record Id & 28 variables that are categorized into 4 categories:
 1. Demographic
 2. Victimization
 3. Suicide Attempt
 4. Substance Use
- Each respondent is identified by a record ID and are mostly between the ages of 15-19 studying in grades 9-12.
- Most of the data consists of “Yes” / “No” answers to questions from the survey.
 - Among the victimization variables, most common “Yes” answer was to the question ‘Fought school 1+ times 12 months – around 15% of records
 - Among the substance use variables, most common “Yes” answer was to the question ‘Tried marijuana 1+ times in life’ – around 50% of records
 - Among the suicide attempts variables, most common “Yes” answer was to the question ‘Considered suicide 12 months – around 16% of records
 - There are many NA/missing values in the dataset and after omitting these, we got 934 records that could be used for analysis.
 - In the data preparation stage, the following transformations and modifications are done.
 - Combining variables from each category into one variable each for victimization, substance use and suicide attempt
 - Generating data in transaction format to do association analysis
- In the first part of modeling, we used apriori algorithm to identify most common associations between victimization, suicide attempt and substance use. The major findings are:
 - There is association between suicide attempt and substance use at support of 17% and confidence of around 70%, but the association between suicide attempt and victimization is weaker at support of around 12% and confidence of around 49%.

- The strongest association between individual variables is between qn27 (Considered suicide in last 12 months) and qn47 (Tried marijuana 1+ times in life) at support of around 11% and confidence of around 62%.
- In the second part of modeling, we used decision tree machine learning techniques to automatically segment the class grade and determine how well these derived groupings correspond to victimization and suicide attempt and predict ‘suicide attempt’ using the aggregated variables ‘victimization’ and ‘substance use’ for class ‘grade’.

2 Introduction.

2.1 Background.

Health behaviors and experiences related to sexual behavior, high-risk substance use, violence victimization, mental health and suicide contribute to substantial morbidity for adolescents, including risk for HIV, STDs, and teen pregnancy. The Centers for Disease Control and Prevention in the United States monitors routinely youth health behaviours and experiences by conducting a yearly survey across the country in collaboration with schools to help in prevent future prevention of spread of HIV, drug uses, sexually transmitted diseases, and unintended teen pregnancy in goal to raise awareness and understanding. Collected under the flag of YRBSS system which is developed in 1990 to monitor those risks. From 1991 until 2013, A total 2.6 million high school student data was collected in more than 1,100 separate surveys. In this paper the analysis will be performed on 2013 YRBSS Chicago dataset, downloaded from the Centers for Disease Control and Prevention, CRISP-DM methodology steps will be followed. R open-source software for report generating, analysis, and to communicate findings.

2.1.1 Overview of CRISP-DM:

CRISP-DM was conceived in 1996 and immature data mining market as a standardized process for data mining projects. The methodology provides an overview of the life cycle of a data mining project, In six major steps; Business understanding, Data understanding, Data preparation, Modeling, And Evaluation. According to latest poll (2014) results by <https://www.kdnuggets.com/>, crisp-dm methodology is still the most popular data mining with 43%, Next most popular method is SEMMA by SAS institute, and followed third by KDD process method.

Note:

- As per the course requirement deployment step is not included in process.

2.2 Business Challenges.

This case study examines the associations between victimization, substance use, and suicide attempt among youth in Chicago at 2013, challenges are:

- Are there relations between victimization (fighting, bullying, sexual abuse) and substance use (Tobacco, alcohol and other drug use)?
- Are there relations between victimization (fighting, bullying, sexual abuse) and suicide attempts?
- Are there relations between substance use (Tobacco, alcohol and other substance use) and suicide attempt?

3 Business Understanding:

3.1 Objective:

The main objective of this project is to examine the associations between victimization, substance use, and suicide attempt among youth using the Youth Health Risk Behavior Survey data for the year 2013. We will be using Apriori (an algorithm for frequent item set mining and association rule learning) and Decision Tree. This study has several addressing association of youth mental health with substance use, suicide attempt, and victimization. Data sample includes only youth, and thus focuses on health related outcomes not limited to substance use, suicide, or victimization. Most importantly, by using a representative sample of youth who were not age-cohorted, The data set will allow for identification of a number of potentially valuable insights.

Using the above mentioned algorithms we are going to answer the following questions:

1. Are there relations between victimization (fighting, bullying, sexual abuse) and suicide attempt?
2. Are there relations between substance use (Tabaco, alcohol and other substance use) and suicide attempt?
3. Apply machine learning techniques to automatically segment the class grade into clusters and determine how well these derived groupings correspond to victimization and suicide attempt.

3.2 Motivation:

Youth suicide is a substantial concern for health professionals, educators, lawmakers and society in general. Researchers have estimated that around 11% of all deaths among 12-19 year olds is due to suicide. It is assumed that there is high association between victimization, substance use and suicide attempt. Studying these associations will help in understanding youth behaviors and reducing adverse events. This type of analysis will also help doctors make decisions after taking into account the risk of suicide among their youth patients with history of victimization and/or substance use.

It is also critical to identify high risk groups who may be more associated with suicide attempt so that targeted preventive measures can be taken. For example, the CDC states that historical suicide rates for teens aged 15-19 years in the US differ significantly between genders.

3.3 Data Description:

The Youth Health Risk Behavior Survey is a biannual study undertaken by UNITED STATE CDC that monitors several categories of health-related behaviors among youth. The survey includes adolescents from grades 9-12 in the age group of 14-19 years. In our analysis, we consider behaviors related to victimization (fighting, bullying, sexual abuse, etc.), substance use (tobacco, alcohol, marijuana, etc.) and suicide attempt (considered suicide, attempted suicide, etc.). The responses of the survey questions are initially processed by the CDC to identify logical inconsistencies, convert responses to usable form, create derived variables from responses, etc. We use a subset of the full data and analyze only demographic, victimization, substance use and suicide attempt information.

3.3.1 About the dataset:

This case study is from the Youth Risk Behavior Survey (YRBS) data which are free for use. (Seen from <http://www.cdc.gov/healthyyouth/data/yrbs/data.htm>).

4 Data Understanding.

The dataset used in this project consists of Record ID that serves as a unique identifier, 4 demographic variables, 7 victimization variables, 4 suicide attempt variables and 13 substance use variables. The data is provided in csv format.

4.1 Metadata Description.

Here shows a full description of all dataset variables, with details for all variables.

Variable	Description	Short.Description
record	Record ID of participant	Record
age	Age of participant	Age
sex	Sex of participant	Sex
grade	Grade in which participant was studying	Grade
race4	Race/ethnicity of participant	Race
qn16	Unsafe at school 1 or more times in the past 30 days	Unsafe
qn17	Threatened at school 1 or more times in the past 12 months	Threatened
qn19	Injured at school 1 or more times in the past 12 months	Injured
qn20	Fought at school 1 or more times in the past 12 months	Fought
qn21	Forced to have sex	ForcedSex
qn24	Bullied 1 or more times in the past 12 months	Bullied
qn25	Electronically bullied 1 or more times in the past 12 months	Ebullied
qn27	Considered suicide in the past 12 months	ConsideredSuicide
qn28	Made suicide plan in the past 12 months	MadeSuicidePlan
qn29	Attempted suicide in the past 12 months	AttemptSuicide
qn30	Suicide attempt with or without injury in the past 12 months	SuicideWithOrWithoutInjury
qn33	Smoked 1 or more times in the past 30 days	SmokedMonth
qn37	Smoked daily for 30 days	SmokedDaily
qn43	Had drinks 1 or more times in the past 30 days	Drinks1+
qn45	Had drinks 10 or more times in the past 30 days	Drinks10+
qn47	Tried marijuana 1 or more times in life	Marijuana1+
qn50	Used cocaine 1 or more times in life	Cocaine1+
qn51	Sniffed glue 1 or more times in life	SniffedGlue1+
qn52	Used heroin 1 or more times in life	Heroin1+
qn53	Used meth 1 or more times in life	Meth1+
qn54	Used ecstasy 1 or more times in life	Ecstasy1+
qn55	Took steroids 1 or more times in life	Steroids1+
qn56	Taken prescription drug without prescription 1 or more times in life	PrescriptionDrug
qn57	Injected drugs 1 or more times in life	InjectedDrugs

4.2 Loading, Retrieving, Viewing Data.

Loading the data from the main source and view the 5 occurrences from each variable.

Table 2: First 5 rows of dataset (continued below)

record	age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21	qn24	qn25	qn27	qn28
1115896	NA	NA	NA	2	1	1	2	1	NA	NA	2	1	NA
1115897	NA	NA	4	3	1	2	2	2	2	1	2	1	1
1115898	1	NA	NA	4	1	2	1	1	2	2	2	1	1
1115899	2	NA	2	3	NA	1	1	NA	2	NA	2	2	NA
1115900	3	NA	3	NA	1	1	1	1	NA	2	2	2	2
1115901	4	NA	1	3	NA	NA	NA	NA	1	2	2	NA	NA

qn29	qn30	qn33	qn37	qn43	qn45	qn47	qn50	qn51	qn52	qn53	qn54	qn55	qn56	qn57
NA	NA	NA	NA	NA	NA	1	NA	1	2	2	2	1	2	2
2	2	2	2	1	2	2	2	2	2	2	2	2	1	2
NA	NA	NA	1	NA	NA	1	1	1	1	2	1	1	1	2
1	2	NA	2	1	2	NA	1	1	1	1	1	1	1	1
2	2	NA	2	2	2	1	2	2	2	2	2	2	2	2
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

4.3 Exploratory Data Analysis.

This section introduce an an exploration to dataset by insvestigating all the variables, and observations, missing and completed rows.

Table 4: Sample of observations with Empty values

age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21	qn24	qn25	qn27	qn28	qn29
NA	NA	NA	2	1	1	2	1	NA	NA	2	1	NA	NA
NA	NA	4	3	1	2	2	2	2	1	2	1	1	2

Table 5: Describe basic informations of dataset (continued below)

rows	columns	discrete_columns	continuous_columns	all_missing_columns
1581	29	0	29	0

total_missing_values	complete_rows	total_observations	memory_usage
1841	934	45849	189816

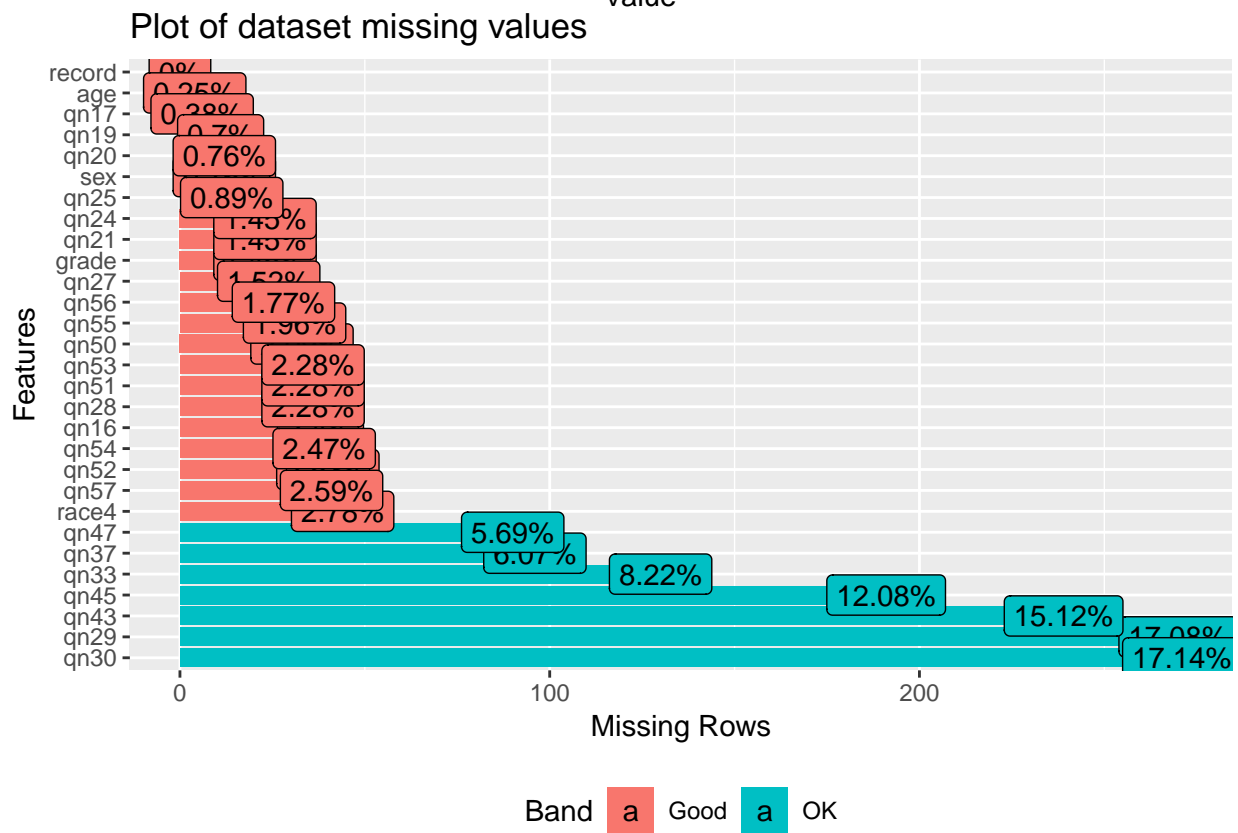
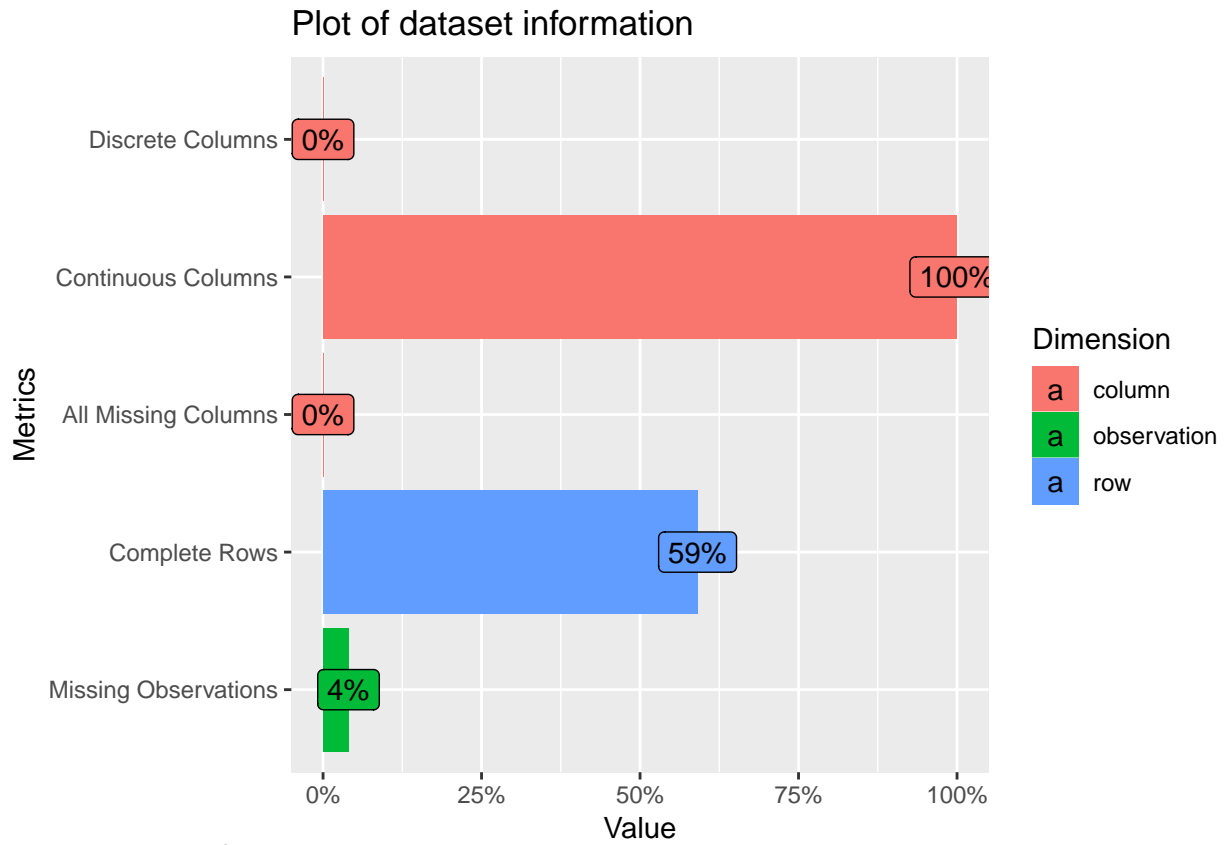


Table 7: Sample of observations with Empty values

age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21	qn24	qn25	qn27	qn28	qn29
NA	NA	NA	2	1	1	2	1	NA	NA	2	1	NA	NA
NA	NA	4	3	1	2	2	2	2	1	2	1	1	2

- Summary:

It is clear that, there are many NA values in the data. that will addressed and preprocessed in the following steps.

4.4 Demographic Variables Description

Description of the factor levels of demographic variables:

In dataset there are 4 demographic variables, grouped as a levels as the below:

- Age: (1= <=12 years old, 2=13 years old, 3=14 years old, 4=15 years old, 5=16 years old, 6=17 years old, 7=18+ years old).
- Sex: (1=Female, 2=Male).
- Race: (1=White, 2=Black/African American, 3=Hispanic/Latino, 4=All other races).
- Grade (1=9th, 2=10th, 3=11th, 4=12th).

Data from the questions are all dichotomous (ordinal) as numerical values with levels “1”, “2”...etc or NA for missing value.

Note:

- Considering “1” corresponds to “Yes” and “2” corresponds to a “No”.

5 Data Preprocessing (Demographics)

This section will implement data cleaning process for all non needed variables.

5.1 Creating A Custom R Function To Handle Missing Data For Specific/All Observations.

5.1.1 Apply On Demographics Variables Only.

This function takes the dataset and the cols to omit the na values

Note:

- This function will used to clean data in two phases, phase (1) will handle the demographic variables, and phase (2) handles the other victimization variables, suicide attempt variables and 13 substance use variables.

```
#This function takes the dataset and the cols to omit the na values
omitNaForSpecificCols <- function(d, desiredCols) {
  completeVec <- complete.cases(d[, desiredCols])
  return(d[completeVec, ])
}
```


5.2 Data Distribution Overview (Demographics).

Here and overview of cleaned data from any Na observation, and statistical summary of the new observations.

- Summary:

New dataset has a 1514 obs. out of 5 demographic variables (age,sex,grade,race), and non empty observations, and levels have been changed into categorical (texts) for analysis in the next steps.

- Note:

NA in this output shows only an empty level eg: “sex variable according to dataset description has only two levels”

Table 8: Demographics data summary

age	sex	grade	race
<=12: 4	Female:866	9th :259	White :132
13 : 0	Male :648	10th:374	Black/African American:549
14 :106	NA	11th:434	Hispanic/Latino :705
15 :279	NA	12th:447	All other races :128
16 :397	NA	NA	NA
17 :427	NA	NA	NA
18+ :301	NA	NA	NA

Table 9: Demographics data structure (continued below)

rows	columns	discrete_columns	continuous_columns	all_missing_columns
1514	4	4	0	0

total_missing_values	complete_rows	total_observations	memory_usage
0	1514	6056	28856

5.3 Data Overview (Demographics).

Here shows a sample of dataset after cleaning, and transforming the variables.

Table 11: New demographic variables

age	sex	grade	race
<=12	Male	10th	Hispanic/Latino
<=12	Male	12th	Hispanic/Latino
<=12	Male	11th	All other races
14	Male	9th	Black/African American
14	Male	9th	Hispanic/Latino

5.4 Statistical Summary Of Grouped Data (Demographics).

Here shows a statistical summary of all grouped demographic variables, this give a meaning how well are our data variables are distributed, and if data set is biased toward a certain variable. A biased dataset in many cases can effect all the anlysis results and give a non relastic insights.

- Summary:

Sex groups variables are normaly distributed.

Race groups variables are normaly distributed.

Age groups variables are normaly distributed.

Grade groups are skwedtowards 12th, and 11th observations.

Table 12: statistical summary of demographics groups

Variable	Valid	Frequency	Percent	CumPercent
race	All other races	25	25.51	25.51
race	Black/African American	26	26.53	52.04
race	Hispanic/Latino	28	28.57	80.61
race	White	19	19.39	100
race	TOTAL	98	NA	NA
sex	Female	46	46.94	46.94
sex	Male	52	53.06	100
sex	TOTAL	98	NA	NA
grade	10th	27	27.55	27.55
grade	11th	26	26.53	54.08
grade	12th	21	21.43	75.51
grade	9th	24	24.49	100
grade	TOTAL	98	NA	NA
age	<=12	4	4.08	4.08
age	14	11	11.22	15.3
age	15	17	17.35	32.65
age	16	25	25.51	58.16
age	17	24	24.49	82.65
age	18+	17	17.35	100
age	TOTAL	98	NA	NA

5.5 Data Visualzation.

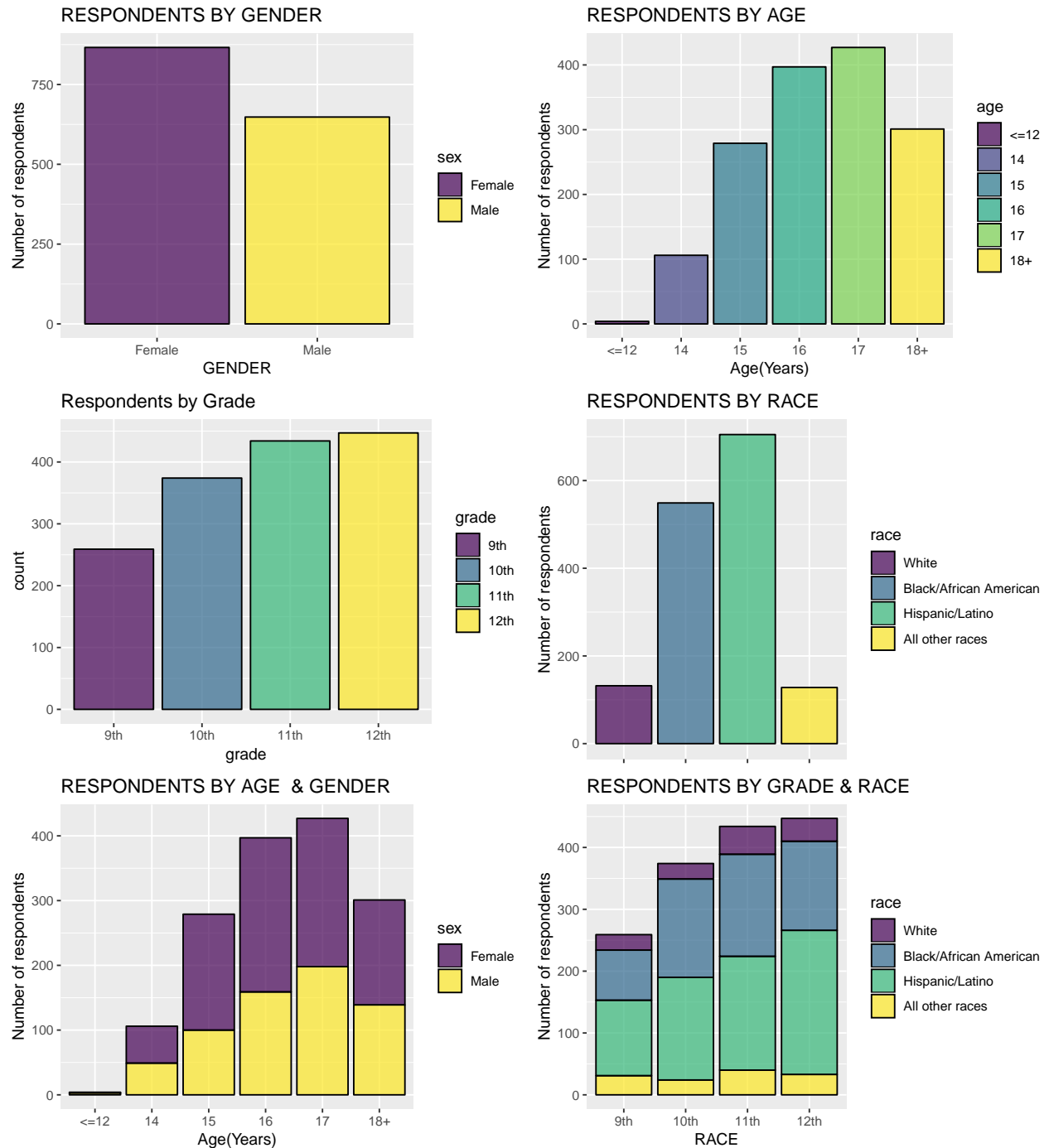
This section invistigate the relations between the diffrent variables of dataset and to visulize intresting insights from data.

5.5.1 Demographic Details (Age,Race,Sex, & Grade) of Respondents.

Barplot measure the distribution of variables over the dataset, it can be usefull to indicate several informative insights such as the skewness of data, how observations are dominant to others.

- Summary:

It can be seen clearly Hispanic/Latinos, and Black/African Americans are dominants when compared to other races, most of data samples are aged above 14 years old, Male and females are distributed normally among all ges in dataset



5.5.2 Victimization Proportion.

This section investigate the victimizations variables in dataset, which are ranged in victimizations groups variables, visulizing thoses variables gave us the following insights.

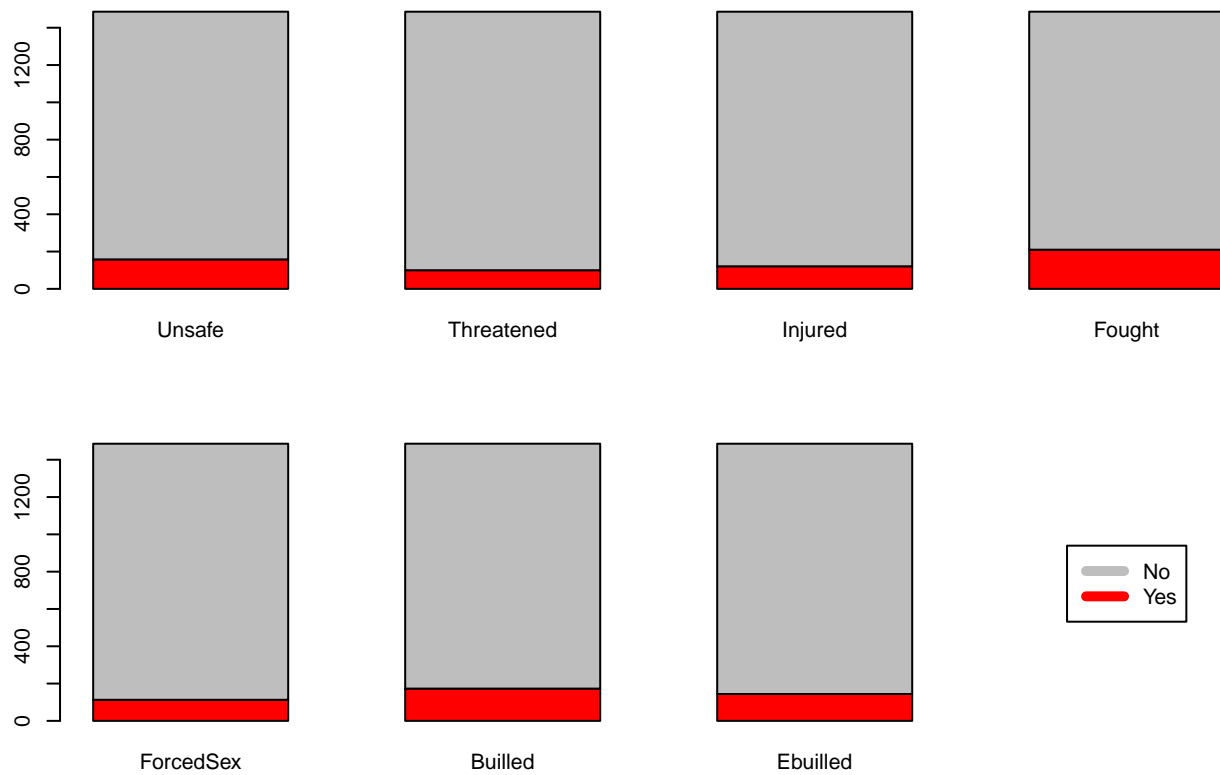
- Summary:

Victimization's highest percentage are:

1. Fought school 1+ times 12 months : around 15%.
2. Missed school b/c unsafe 1+ 30 days : around 12%.

3. Bullied at school 12 months : around 12%.

Split of victimization variables



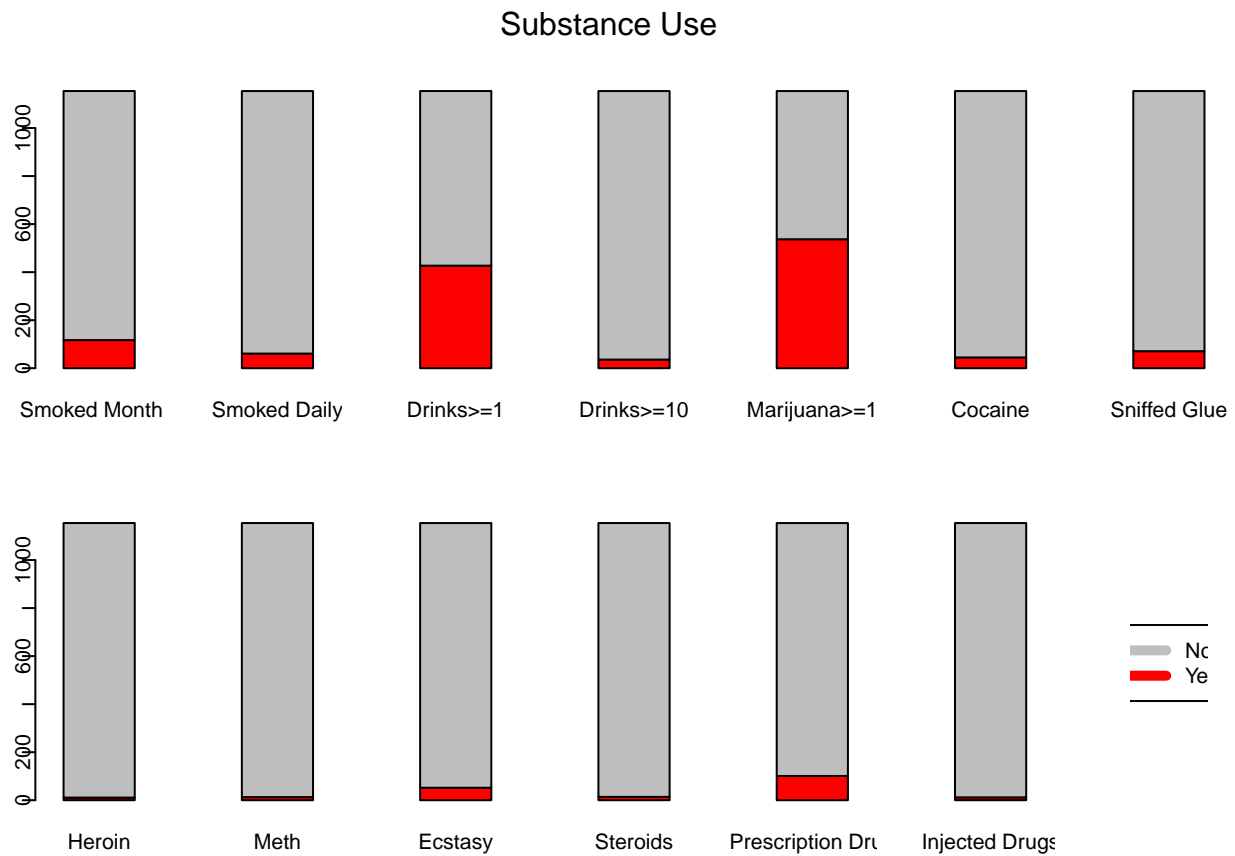
5.5.3 Substance Use Proportion.

This section investigates the substance use variables in dataset, which are ranges in Substance use groups variables, visualizing those variables gave us the following insights.

- Summary:

Substance use highest percentage are:

1. Tried marijuana 1+ times in life : around 50%.
2. Had 1+ drinks past 30 days : around 39%.



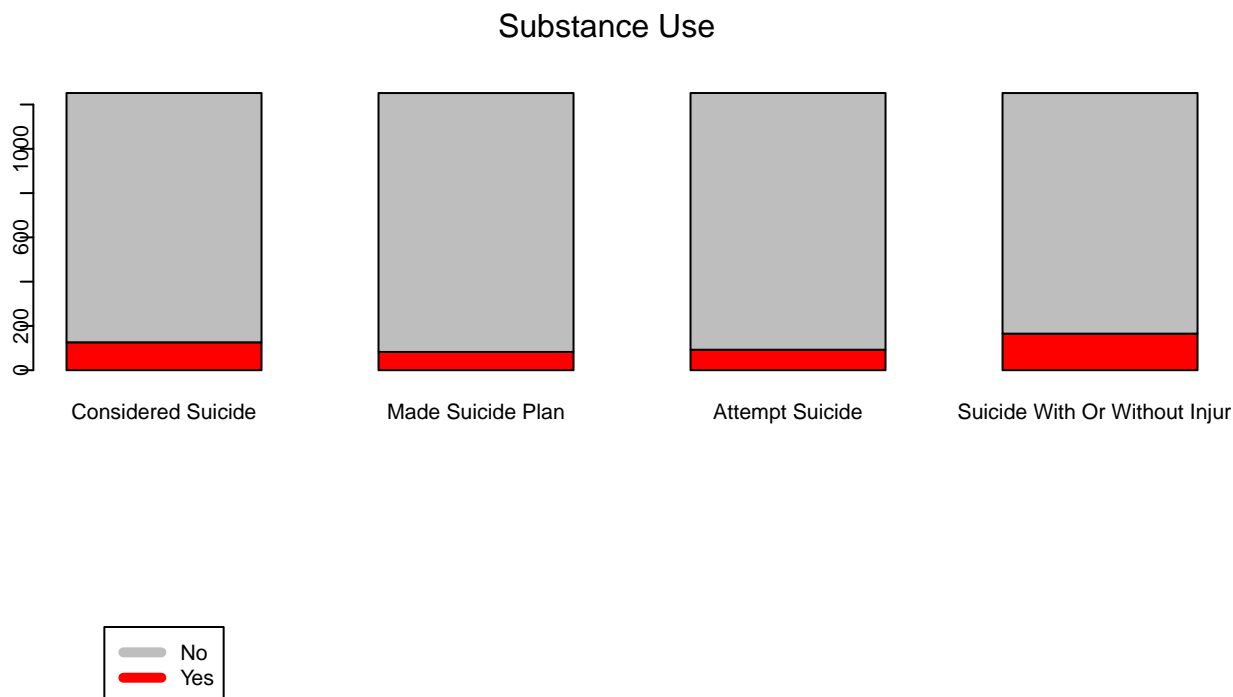
5.5.4 Suicide Attempts Proportion.

This section investigates the substance use variables in dataset, which are ranges in suicide attempts groups variables, visualizing those variables gave us the following insights.

- Summary:

Suicide attempts highest percentage are:

1. Considered suicide 12 month : around 16%.
2. Made suicide plan 12 month : around 14%.



6 Data Preparation.

In the above section, we noticed that there are many NA values in the data. We checked the number of missing values in each field using the following function that generates the following insights.

-Summary:

Question 29, question 30 and question 43 have the most NA values. These are ‘Attempted suicide 1+ times 12 months’, ‘Suicide attempt with injury 12 months and ‘Had 1+ drinks past 30 days’.

Table 13: Table continues below

record	age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21	qn24	qn25	qn27	qn28
0	4	12	23	44	38	6	11	12	23	23	14	24	36

qn29	qn30	qn33	qn37	qn43	qn45	qn47	qn50	qn51	qn52	qn53	qn54	qn55	qn56	qn57
270	271	130	96	239	191	90	33	36	40	36	39	31	28	41

6.1 Cleaning Data Using “omitNaForSpecificCols” Custom Function.

As a first step Here we clean all dataset of any non needed variable, and invistigate if we still have any missing value.

Table 15: Table continues below

record	age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21	qn24	qn25	qn27	qn28
0	0	0	0	0	0	0	0	0	0	0	0	0	0

qn29	qn30	qn33	qn37	qn43	qn45	qn47	qn50	qn51	qn52	qn53	qn54	qn55	qn56	qn57
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 17: sample of clean dataset

	age	sex	grade	race4	qn16	qn17	qn19	qn20	qn21
15	1	2	2	3	2	2	2	2	2
20	3	2	1	3	1	2	1	2	2
22	3	2	1	3	2	2	2	2	2

6.2 Correlation Matrix Of Variables.

Correlation used in EDA to measures the strengths of association between two variables., the value of the correlation coefficient varies between +1 and -1. 1.0 (a perfect positive correlation) and -1.0 (a perfect negative correlation). A 0 indicates no association between ranks. spearman correlation chioce came from the non-linear relation between the variable

Spearman correlation formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where:

d = the pairwise distances of the ranks of the variables xi and yi.

n = the number of samples.

- Summary:

Students Who have been on substance use are high level of suicide thoughts, and attempts.

Unsafe students are on high level to be exposed to drugs substance.

students who have been builled, have an experince or E-builled.

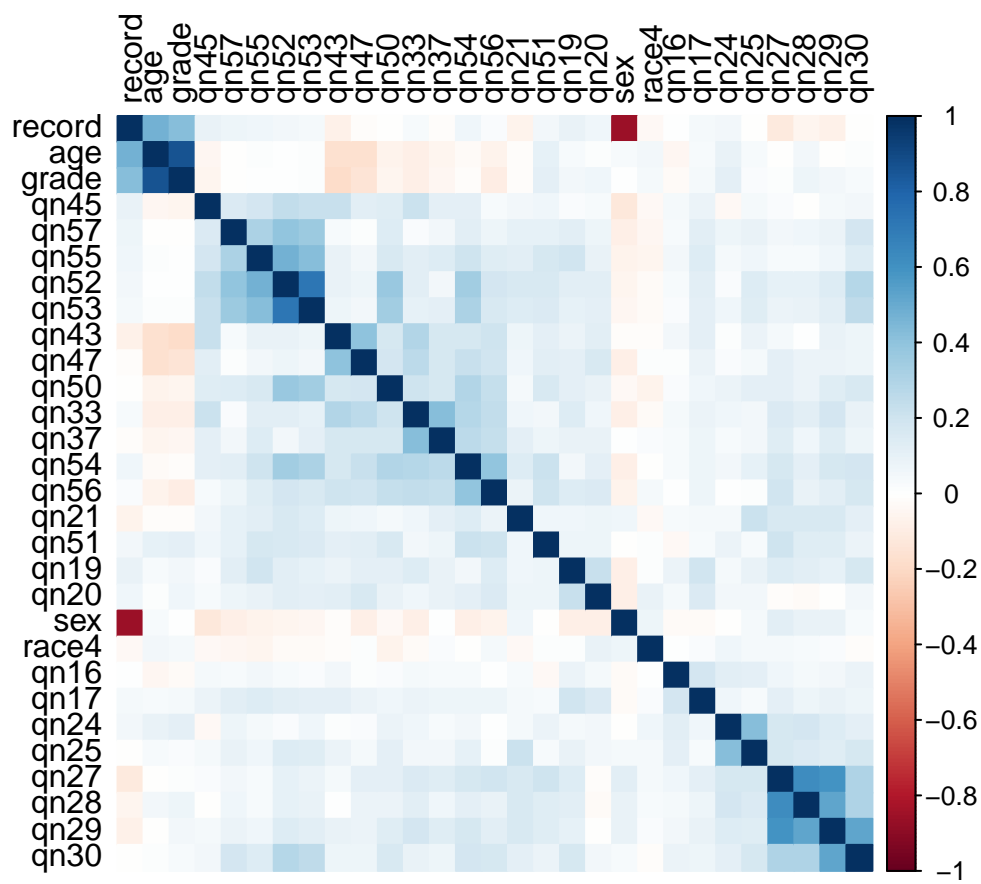


Figure 2: Spearman corrplot

6.3 Data Aggregation.

In this section, we also do some data transformations that will help in analysis and modeling. Specifically, we do the following steps:

1. Combine variables for victimization, suicide and substance use into one variable each.
2. Convert the clean data into transaction format to do association analysis.

6.3.1 Combine variables for victimization, suicide and substance use into one variable each.

Here we combined all the victimization variable and combine them into one binary encoded for each one.

```
## [1] "Unsafe"      "Threatened" "Injured"     "Fought"      "ForcedSex"
## [6] "Bullied"

## [1] "SmokedMonth" "SmokedDaily" "Drinks1+"    "Drinks10+"   "Marijuana1+"
## [6] "Cocaine1+"

## [1] "ConsideredSuicide"      "MadeSuicidePlan"
## [3] "AttemptSuicide"        "SuicideWithOrWithoutInjury"
```

Table 18: Encoded data sample

	victimization	substanceUse	suicideAttempt
15	0	1	0
20	1	0	1
22	0	1	0
23	0	1	0
25	0	0	0

6.3.2 Convert Clean Data Into Transaction Format For Association Analysis (Aggregated columns).

Here a last step before implement the model, aggregated columns herewill be converted into a transactional form to apply apriori rules analysis.

- Summary:

Most frequent transactional items are the highest in SubstanceUse, Then victimization, And Last suicideAttempts.

```
## transactions as itemMatrix in sparse format with
## 672 rows (elements/itemsets/transactions) and
## 3 columns (items) and a density of 0.5128968
##
## most frequent items:
##  substanceUse  victimization  suicideAttempt      (Other)
##           558           312           164           0
##
## element (itemset/transaction) length distribution:
## sizes
##  1  2  3
## 369 244 59
##
##  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 1.000  1.000   1.000   1.539   2.000   3.000
##
```

```
## includes extended item information - examples:
##      labels
## 1  substanceUse
## 2 suicideAttempt
## 3  victimization
##
## includes extended transaction information - examples:
##      transactionID
## 1      1115910
## 2      1115915
## 3      1115917
##
##      substanceUse suicideAttempt victimization
## substanceUse      558          114          226
## suicideAttempt    114          164           81
## victimization    226          81          312
```

6.3.3 Convert Clean Data Into Transaction Format For Association Analysis (non aggregated Columns, except demographics data).

Here will include all variables except demographics one, to see what are most frequentset.

- Summary:

Most frequent items are student who some marijuana, then who drinks more than one time in last 30 days, and last are students who considered suicide.

```
## transactions as itemMatrix in sparse format with
## 672 rows (elements/itemsets/transactions) and
## 24 columns (items) and a density of 0.1229539
##
## most frequent items:
##      Marijuana1+      Drinks1+ ConsideredSuicide      MadeSuicidePlan
##           438           354           120           110
##      Builled      (Other)
##           94           867
##
## element (itemset/transaction) length distribution:
## sizes
##  1  2  3  4  5  6  7  8  9 10 12 13 20 24
## 200 165 119 71 31 34 21 15 9 1 1 3 1 1
##
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      1.000  1.000  2.000  2.951  4.000  24.000
##
## includes extended item information - examples:
##      labels
## 1 AttemptSuicide
## 2      Builled
## 3      Cocaine1+
##
## includes extended transaction information - examples:
##      transactionID
## 1      1115910
## 2      1115915
## 3      1115917
```

7 Modeling.

This section will implement the usage of several algorithms, on our dataset, for mining frequent item sets. Apriori and Eclat algorithms will be used as a model to analyze our encoded transactional dataset as a part of our unsupervised learning section. Apriori and Eclat are one of the most used algorithms to mine such a type of dataset to represent transactions lists and to filter closed and maximal item sets. The second part will go through a supervised learning model of decision tree to understand what influences suicide attempts as a target variable in dataset. Model parameters will be set as the default settings.

7.1 Association Rules Algorithms (Apriori & Eclat).

7.1.1 Generate Association Rules For All Itemsets Without Aggregation (Victimization, Substance & Suicide).

7.1.1.1 Models Parameters.

```
## Eclat
##
## parameter specification:
## tidLists support minlen maxlen          target  ext
##   FALSE      0.1      2      10 frequent itemsets FALSE
##
## algorithmic control:
## sparse sort verbose
##      7    -2    TRUE
##
## Absolute minimum support count: 67
##
## create itemset ...
## set transactions ... [24 item(s), 672 transaction(s)] done [0.00s].
## sorting and recoding items ... [11 item(s)] done [0.00s].
## creating bit matrix ... [11 row(s), 672 column(s)] done [0.00s].
## writing ... [5 set(s)] done [0.00s].
## Creating S4 object ... done [0.00s].

## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##      0.2    0.1    1 none FALSE          TRUE        5    0.1    2
## maxlen target  ext
##      10 rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##    0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 67
##
## set item appearances ... [0 item(s)] done [0.00s].
## set transactions ... [24 item(s), 672 transaction(s)] done [0.00s].
## sorting and recoding items ... [11 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [8 rule(s)] done [0.00s].
## creating S4 object ... done [0.00s].
```

7.1.1.2 Models Results.

7.1.1.2.1 AprioriRules Model Results.

```
##      lhs                      rhs          support  confidence
## [1] {SmokedMonth}             => {Drinks1+}      0.1116071 0.8064516
## [2] {SmokedMonth}             => {Marijuana1+}    0.1205357 0.8709677
## [3] {MadeSuicidePlan}         => {ConsideredSuicide} 0.1145833 0.7000000
## [4] {ConsideredSuicide}       => {Marijuana1+}    0.1101190 0.6166667
## [5] {Drinks1+}                => {Marijuana1+}    0.3839286 0.7288136
##      lift      count
## [1] 1.5308912   75
## [2] 1.3362793   81
## [3] 3.9200000   77
## [4] 0.9461187   74
## [5] 1.1181797  258
```

7.1.1.2.2 Ecalt Model Results.

```
##      items                                support  count
## [1] {Marijuana1+,SmokedMonth}             0.1205357   81
## [2] {Drinks1+,SmokedMonth}                0.1116071   75
## [3] {ConsideredSuicide,MadeSuicidePlan}    0.1145833   77
## [4] {ConsideredSuicide,Marijuana1+}        0.1101190   74
## [5] {Drinks1+,Marijuana1+}                0.3839286  258
```

7.1.2 Generate For All Itemsets With Aggregation (Victimization,Substance & Suicide).

7.1.3 Models Parameters.

```
## Eclat
##
## parameter specification:
## tidLists support minlen maxlen          target  ext
##      FALSE      0.1      2      10 frequent itemsets FALSE
##
## algorithmic control:
## sparse sort verbose
##      7      -2      TRUE
##
## Absolute minimum support count: 67
##
## create itemset ...
## set transactions ...[3 item(s), 672 transaction(s)] done [0.00s].
## sorting and recoding items ... [3 item(s)] done [0.00s].
## creating bit matrix ... [3 row(s), 672 column(s)] done [0.00s].
## writing ... [3 set(s)] done [0.00s].
## Creating S4 object ... done [0.00s].
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##      0.2      0.1      1 none FALSE          TRUE      5      0.1      2
## maxlen target  ext
##      10 rules FALSE
```

```

##
## Algorithmic control:
## filter tree heap memopt load sort verbose
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE
##
## Absolute minimum support count: 67
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[3 item(s), 672 transaction(s)] done [0.00s].
## sorting and recoding items ... [3 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [6 rule(s)] done [0.00s].
## creating S4 object ... done [0.00s].

```

7.1.4 Models Results.

7.1.4.0.1 AprioriRules Model Results.

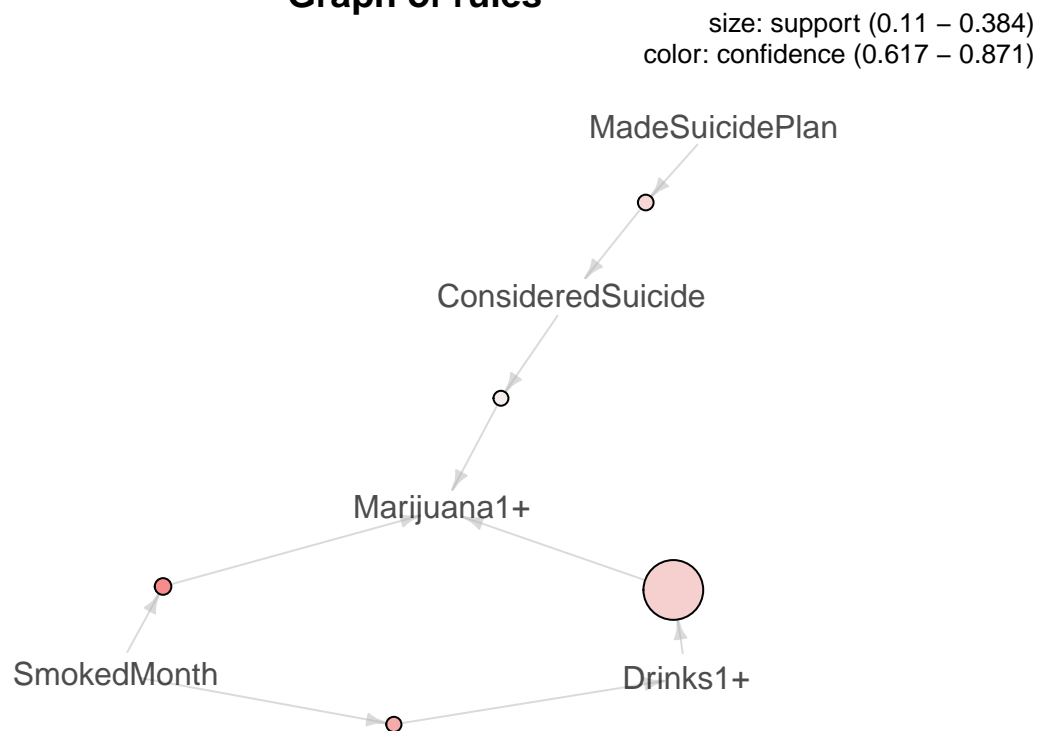
```

## items support count
## [1] {substanceUse,suicideAttempt} 0.1696429 114
## [2] {suicideAttempt,victimization} 0.1205357 81
## [3] {substanceUse,victimization} 0.3363095 226

```

7.1.5 Graphing The Strong Association Rules Without Aggregation(Victimization,Substance & Suicide).

Graph of rules



Graph of Rules

size: support (0.11 – 0.384)
color: confidence (0.617 – 0.871)

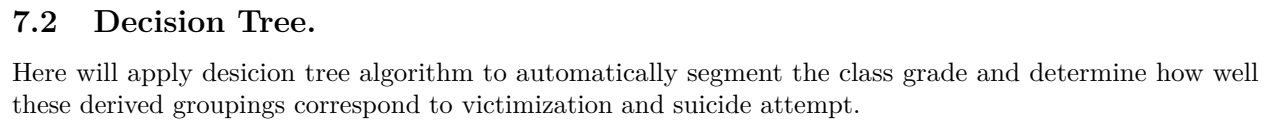


Table 19: Structure of Decision Tree variables

	Grade	substanceUse	victimization	suicideAttempt
15	2	1	0	0
20	1	0	1	1
22	1	1	0	0
23	1	1	0	0
25	1	0	0	0
26	1	0	0	0

	Grade	substanceUse	victimization	suicideAttempt
27	1	0	0	0
28	1	0	0	0
29	1	1	0	0
30	1	1	0	0

7.2.2 Create Train and Test Samples.

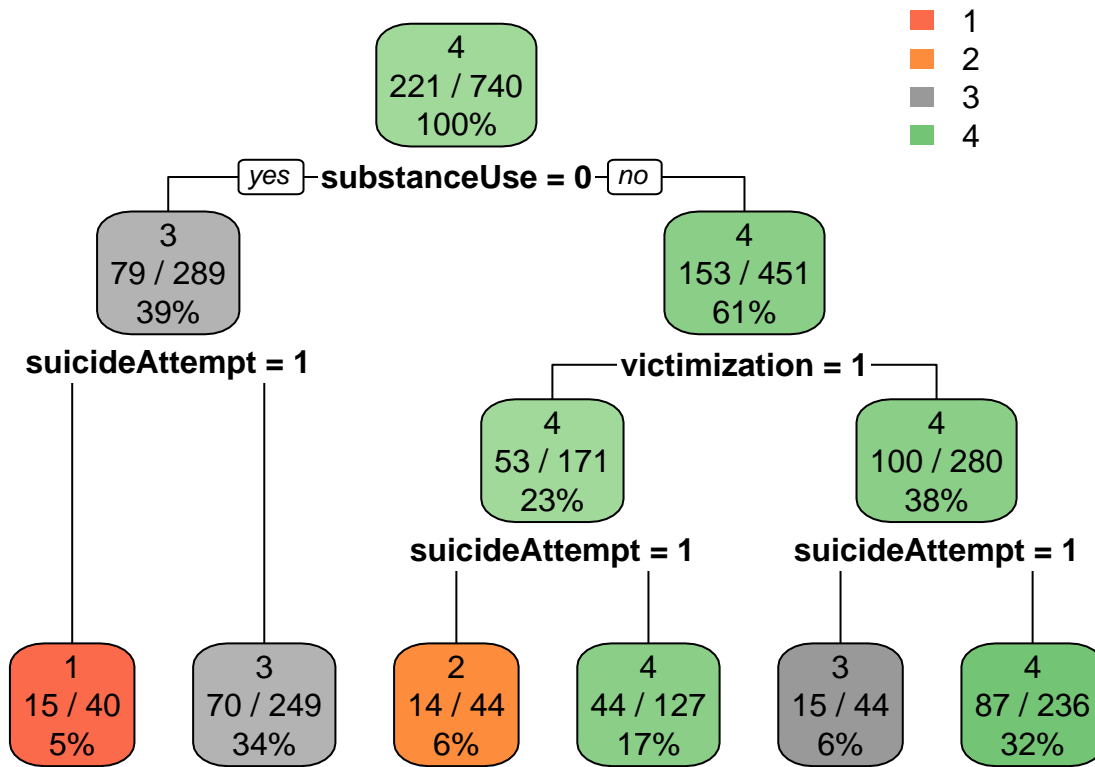
```
## [1] 740 4
```

```
## [1] 194 4
```

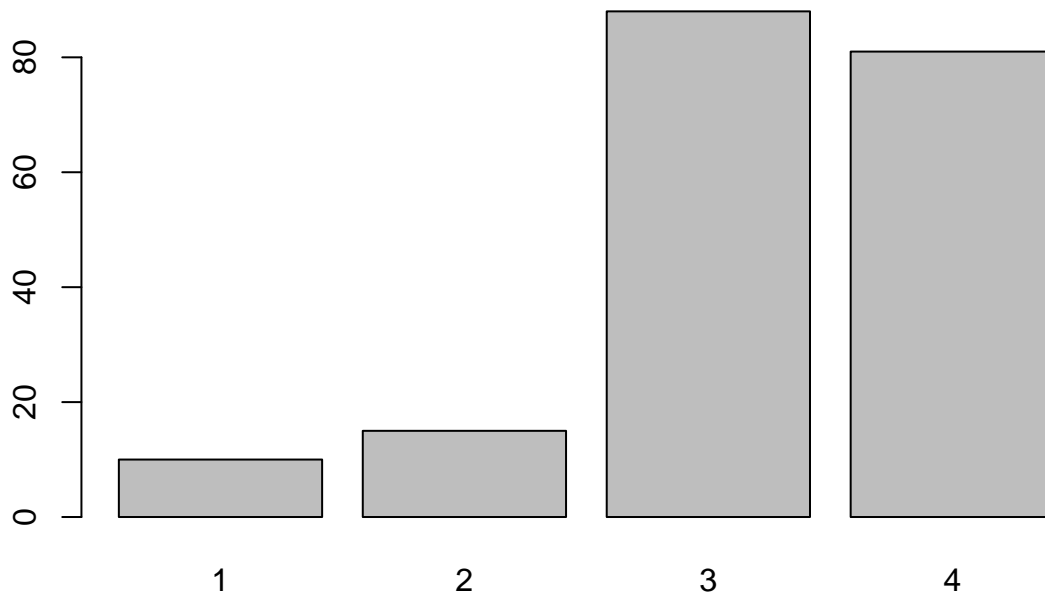
7.2.3 Create Model For Recursive Partitioning and Regression Tree

```
## n= 740
##
## node), split, n, loss, yval, (yprob)
##      * denotes terminal node
##
## 1) root 740 519 4 (0.17432432 0.23648649 0.29054054 0.29864865)
##    2) substanceUse=0 289 210 3 (0.25605536 0.23529412 0.27335640 0.23529412)
##      4) suicideAttempt=1 40 25 1 (0.37500000 0.17500000 0.22500000 0.22500000) *
##      5) suicideAttempt=0 249 179 3 (0.23694779 0.24497992 0.28112450 0.23694779) *
##    3) substanceUse=1 451 298 4 (0.12195122 0.23725055 0.30155211 0.33924612)
##      6) victimization=1 171 118 4 (0.18128655 0.23976608 0.26900585 0.30994152)
##        12) suicideAttempt=1 44 30 2 (0.15909091 0.31818182 0.31818182 0.20454545) *
##        13) suicideAttempt=0 127 83 4 (0.18897638 0.21259843 0.25196850 0.34645669) *
##      7) victimization=0 280 180 4 (0.08571429 0.23571429 0.32142857 0.35714286)
##        14) suicideAttempt=1 44 29 3 (0.11363636 0.25000000 0.34090909 0.29545455) *
##        15) suicideAttempt=0 236 149 4 (0.08050847 0.23305085 0.31779661 0.36864407) *
```

7.2.4 Visualization Model Decision Tree Based On Training Data



7.2.5 Creating Prediction Model



```
##      pModel
##      1  2  3  4
##  0   0  0 77 81
##  1 10 15 11  0
```


7.2.6 Create additional model based oneR

8 Conclusion.

In this project, we studied the Youth Health Risk Behavior using the Observational Data to examine the relations between victimization, substance use and suicide attempt. We used the apriori algorithm to understand strong associations and built a decision tree to understand what influences suicide attempt. The results of this project tells us that adolescents who consider or attempt suicide tend to use substances. By assessing whether an adolescent was victimized and by looking at their sex, it is possible to predict if they are more likely to consider or attempt suicide. This type of analysis is very important from a medical point of view. It provides a data supported backing of what doctors seem to already believe through experience. This also shows the importance of using machine learning techniques to answer key questions and find solutions in society. Overall, we were successful in identifying associations between victimization, substance use and suicide attempt. We can further improve this project by experimenting with other algorithms like logistic regression and random forest and by considering other types of groupings like race.

9 References

- [AS94] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules (1994) Proc. 20th Int. Conf. Very Large Data Bases, VLDB-94. <http://www.vldb.org/conf/1994/P487.PDF>

10 Appendices

10.0.1 Session Information.

Listing Machine that have been used for the project, operating system, R version, And used libraries with their versions for future reproducibility of the project.

R version 3.6.0 (2019-04-26)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

locale: en_US.UTF-8|en_US.UTF-8|en_US.UTF-8|C|en_US.UTF-8|en_US.UTF-8

attached base packages: *stats4*, *grid*, *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *SmartEDA*(v.0.3.2), *DataExplorer*(v.0.8.0), *pander*(v.0.6.3), *corrplot*(v.0.84), *knitr*(v.1.23), *viridisLite*(v.0.3.0), *rpart.plot*(v.3.0.7), *rpart*(v.4.1-15), *party*(v.1.3-3), *strucchange*(v.1.5-1), *sandwich*(v.2.5-1), *zoo*(v.1.8-6), *modeltools*(v.0.2-22), *mvtnorm*(v.1.0-11), *gridExtra*(v.2.3), *ggrepel*(v.0.8.1), *data.table*(v.1.12.2), *arulesViz*(v.1.3-3), *arules*(v.1.6-3), *Matrix*(v.1.2-17), *forcats*(v.0.4.0), *stringr*(v.1.4.0), *dplyr*(v.0.8.1), *purrr*(v.0.3.2), *readr*(v.1.3.1), *tidyr*(v.0.8.3), *tibble*(v.2.1.3), *ggplot2*(v.3.1.1) and *tidyverse*(v.1.2.1)

loaded via a namespace (and not attached): *nlme*(v.3.1-139), *bitops*(v.1.0-6), *matrixStats*(v.0.54.0), *lubridate*(v.1.7.4), *RColorBrewer*(v.1.1-2), *httr*(v.1.4.0), *tools*(v.3.6.0), *backports*(v.1.1.4), *R6*(v.2.4.0), *DT*(v.0.6), *KernSmooth*(v.2.23-15), *lazyeval*(v.0.2.2), *colorspace*(v.1.4-1), *withr*(v.2.1.2), *GGally*(v.1.4.0), *tidyselect*(v.0.2.5), *compiler*(v.3.6.0), *cli*(v.1.1.0), *rvest*(v.0.3.4), *TSP*(v.1.1-7), *xml2*(v.1.2.0), *plotly*(v.4.9.0), *labeling*(v.0.3), *caTools*(v.1.17.1.2), *scales*(v.1.0.0), *lmtest*(v.0.9-37), *digest*(v.0.6.19), *rmarkdown*(v.1.13), *pkgconfig*(v.2.0.2), *htmltools*(v.0.3.6), *highr*(v.0.8), *htmlwidgets*(v.1.3), *rlang*(v.0.4.0), *readxl*(v.1.3.1), *rstudioapi*(v.0.10), *visNetwork*(v.2.0.7), *generics*(v.0.0.2), *jsonlite*(v.1.6), *gtools*(v.3.8.1), *dendextend*(v.1.12.0), *magrittr*(v.1.5), *Rcpp*(v.1.0.1), *munsell*(v.0.5.0), *viridis*(v.0.5.1), *multcomp*(v.1.4-10), *scatterplot3d*(v.0.3-41), *stringi*(v.1.4.3), *yaml*(v.2.2.0), *MASS*(v.7.3-51.4), *gplots*(v.3.0.1.1), *plyr*(v.1.8.4), *parallel*(v.3.6.0), *gdata*(v.2.18.0), *crayon*(v.1.3.4), *lattice*(v.0.20-38), *haven*(v.2.1.0), *splines*(v.3.6.0), *hms*(v.0.4.2), *pillar*(v.1.4.1), *igraph*(v.1.2.4.1), *lpSolve*(v.5.6.13.3), *codetools*(v.0.2-16), *glue*(v.1.3.1), *gclus*(v.1.3.2), *evaluate*(v.0.14), *modelr*(v.0.1.4), *vcd*(v.1.4-4), *foreach*(v.1.4.4), *networkD3*(v.0.4), *cellranger*(v.1.1.0), *gtable*(v.0.3.0), *reshape*(v.0.8.8), *assertthat*(v.0.2.1), *xfun*(v.0.7), *coin*(v.1.3-0), *libcoin*(v.1.0-4), *broom*(v.0.5.2), *survival*(v.2.44-1.1), *seriation*(v.1.2-6), *iterators*(v.1.0.10), *registry*(v.0.5-1), *cluster*(v.2.0.8), *TH.data*(v.1.0-10) and *sampling*(v.2.8)