Association Rules on MFT Study

Aretha Kassegnin,

1. From our previous project stages and even the EDA all our variables had every high usage in the South more than any other region. Which is why we wanted to use a binary target variable just because of the power this dataset must skew our results. This way South has its own category and West, Northeast and Midwest in the same category.
    a. If you are not in the South, there is a less chance that you don't use weed.
    b. 12th graders are most likely to use alcohol in the South.
    c. There are low amounts of substance use in the South for drugs that are unknown or considered highly dangerous.

It is important to mention that there is logical reasoning behind these findings as well. For starters, alcohol is the most easily accessible drug in America, as the legal drinking age is 21 and there are not nearly as many restrictions to this drug as others. For example, alcohol is more accessible and socially acceptable in comparison to drugs like MDMA and cocaine. Another note to mention is that, while at the time this data was collected, weed was not completely legal in some areas, but it is overall more common than some of the other drugs in the data. There is also scientific research being done on cannabis use, which could be a potential factor in adolescents using the substance. With the medical research, there may be more of a chance that young people will choose a "safer" substance that is growing in acceptability based on research, in comparison to those that have been confirmed to be dangerous and deadly, just like cocaine. To further expand on this idea, considering the study was conducted on high schoolers, many of them may not have known what some of the substances are. For example, MDMA has different names such as "ecstasy" and "molly", which could be a factor in usage of the substance if the students did not recognize that the drug can be discussed by different terms. It is also

important to remember that some of the drugs in the data are scientifically proven to be highly dangerous, which is not necessarily the case with substances like weed and alcohol.

2.

LHS	RHS	support	confidence	coverage	lift	count
All	All	All	All	All	All	All
{tranq_12=Sometimes}	{SouthBinary=South}	0.013	0.453	0.028	1.125	130.000
{weed_12=Often}	{SouthBinary=Other}	0.108	0.639	0.168	1.069	1,106.000
{LSD_12=Sometimes}	{SouthBinary=Other}	0.021	0.621	0.033	1.039	213.000
{coke_12=Sometimes}	{SouthBinary=Other}	0.010	0.620	0.017	1.038	106.000
{amp_12=Sometimes}	{SouthBinary=South}	0.015	0.414	0.037	1.029	159.000

The table above shows the results our predictions was wrong because students in the other region are 1.069 times often to choose “often” for weed. Our second prediction was also incorrect as South and alcohol are not on the table. Lastly South region is 1.125 times more likely to use tranq sometimes and 1.029 times more likely to use amp .

3.

LHS	RHS	support	confidence	coverage	lift	count
All	["{SouthBinary=South}","{SouthB	All	All	All	All	All
{weed_12=Often}	{SouthBinary=South}	0.061	0.361	0.168	0.898	626.000
{alcohol_12=Sometimes}	{SouthBinary=South}	0.102	0.387	0.263	0.962	1,049.000
{weed_12=None}	{SouthBinary=Other}	0.371	0.587	0.631	0.983	3,814.000
{alcohol_12=None}	{SouthBinary=Other}	0.389	0.590	0.659	0.987	3,996.000
{amp_12=None}	{SouthBinary=South}	0.382	0.401	0.953	0.995	3,925.000

For question 3 we had support = 0.05 and confidence = 0.2. Here, we can see that for regions other than the South, there was a higher number of those that used weed based on the count. The data gave us a count of 1,225 for sometimes, but for none there was a count of 3,814. In the South, those who often used weed only had a count of 626. Compared to our predictions we are correct in our first prediction South is

.898 times more often to use weed then other students. Next our second prediction was also correct South has a .962 time more likely to pick “3sometimes” for alcohol use. Yes, our last prediction was correct South is .995 times more likely to not use amp.

4. We need more time to analyze our clustering since we were not able to use K-means because we had all categorical data.

Appendix

```
#Assoication Rule
library(arules)
library(arulesViz)
library(caret)
set.seed(25)
data$SouthBinary <- as.factor(ifelse(data$region == "South", "South", "Other"))
inTrain <- createDataPartition(y = data$region,
                               p = .75,
                               list = FALSE)

data.train <- data[ inTrain ,  ]
dim(data.train)[1]
dim(data)[1]
```

```
dim(data.train)[1]/dim(data)[1] #.75 so we are good
#testing data
data.test <- data[ -inTrain ,  ]
dim(data.test)[1]/dim(data)[1] #.249 so good


#bind everything together
data.train$trainortest <-
  rep("train", nrow(data.train))
names(data.train)
data.test$trainortest <-
  rep("test", nrow(data.test))
names(data.test)
data.all <- rbind(data.train, data.test)
#LOOKING FOR PATTERENS
table(data$alcohol_12)
table(data$weed_12)
table(data$LSD_12)
table(data$MDMA_12)
table(data$region)
table(data$SouthBinary)
#using SouthBinary
X <-  subset(data.train, select = c("alcohol_12", "weed_12","coke_12","LSD_12",
                                    "amp_12","tranq_12","meth_12", "MDMA_12", "SouthBinary"))
```

```
Y <- subset(data.train, select =c("alcohol_12", "weed_12","coke_12","LSD_12",
                                  "amp_12","tranq_12","meth_12", "MDMA_12", "region"))
X_train <- apriori(data = X,
                   parameter = list(support = 0.01,
                                    confidence = 0.4),
                   maxlen = 2)
P2 <- apriori(data = X,
                  parameter = list(support = 0.05,
                                   confidence = 0.2),
                  maxlen = 2)
inspectDT(X_train)
inspectDT(P2)
```