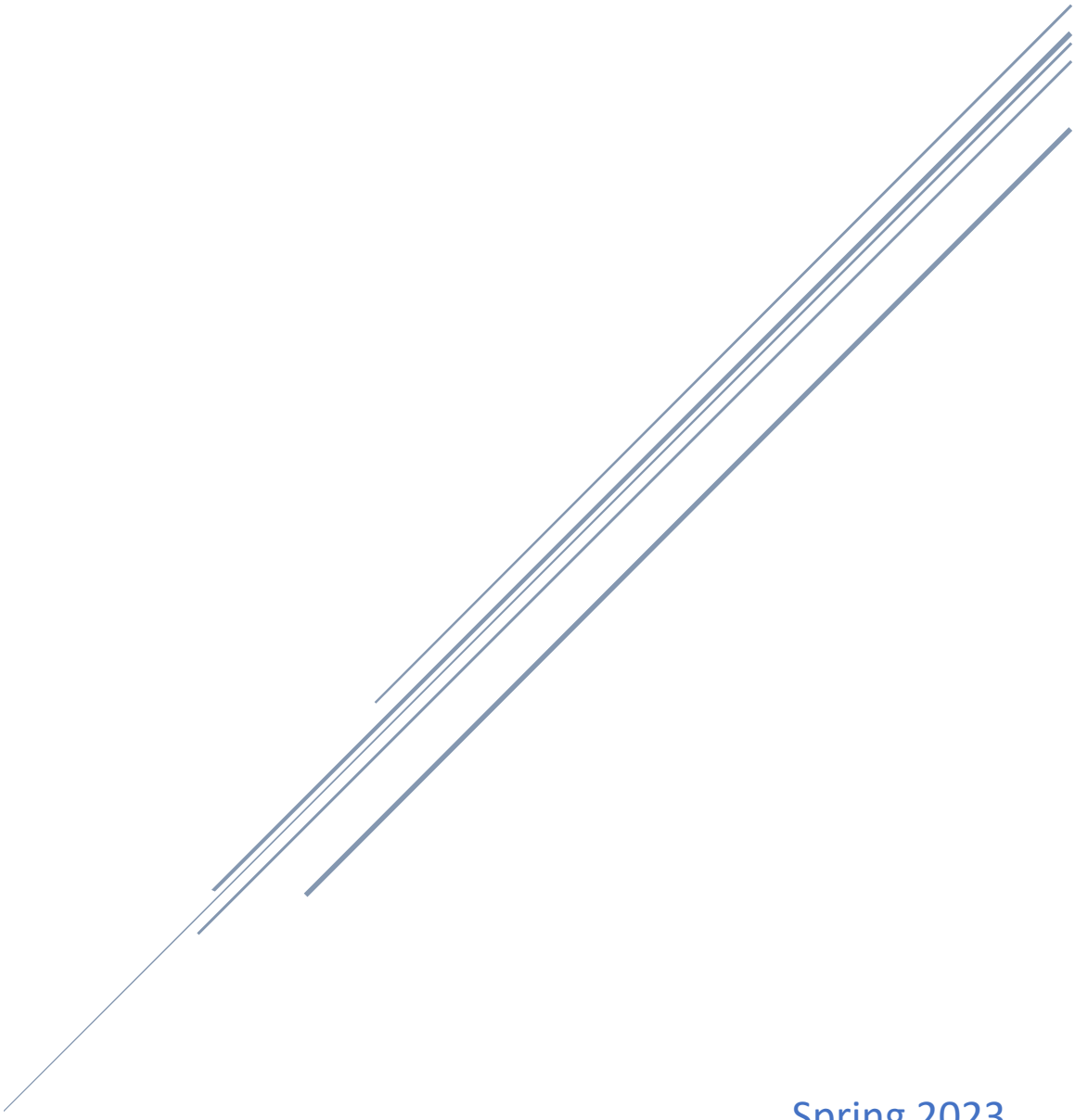
MONITORING THE FUTURE: REGION VS SUBSTANCE USE

Aretha Kassegnin & Brendan Cubberly



Spring 2023

Explorations In Data Analytics

Table of Contents

1) Abstract:

This technical report will utilize several variables that were collected in the Monitoring the Future survey, to identify whether where you live (region) play a role with different types of substances used by 12th grade high school students. Using prior knowledge of what we know to be common drugs, we have identified several variables (different types of substances) which we have hypothesized have a relationship with at least one region.

2) Introduction:

Substance use in America is a very controversial topic. With little to no education about drug use in the school systems, in my experience it was quickly brought up in Health class, with the teacher telling you never do or accept drugs. For some reason substance use conversations were always paired with peer pressure. We believe that there is more than "peer pressure" that would lead a student to consume substances. This report focuses on 7 questions from the Monitoring the Future (MTF) study in hopes to explain their relationships between region a student attends school with likelihood they would consume a certain substance in the last 12 months in 2019.

With more states beginning to legalize the use of marijuana and more research being done on psychedelic drugs, more and more people are using drugs other than alcohol. There is not enough research being done on how different types of substances influence the human body. According to *Drug Use,* published by Out World in Data, found that "Over 350,000 die from overdoses (alcohol and illicit drug use disorders) each year" (Ritchie and Roser). The fact the 12th grades are using strong substances that can lead to an overdose and with fentanyl being on the rise, we believe this study is important to show the public what realistically happens using data. A USA

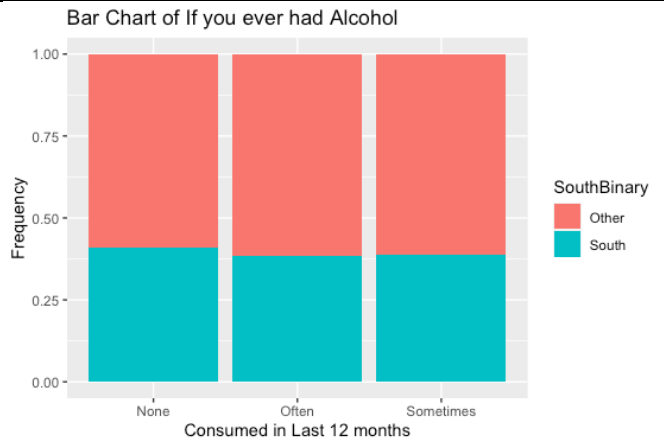
Today article, *100 bags of fentanyl found in bedroom of Connecticut teen who overdosed at school,* police say, “After a 13-year-old boy died from a fentanyl overdose at a Connecticut middle school, Hartford Police said Tuesday they found an extra 100 bags of fentanyl in the teenager's room… The U.S. Drug Enforcement Administration says a fatal dose of fentanyl is small enough to fit on the tip of a pencil.” (Mendoza). Today we are trying to answer the question, what substances are done in what region, and if there is a relationship between a region and that drug.

3) Data Cleaning and Prep

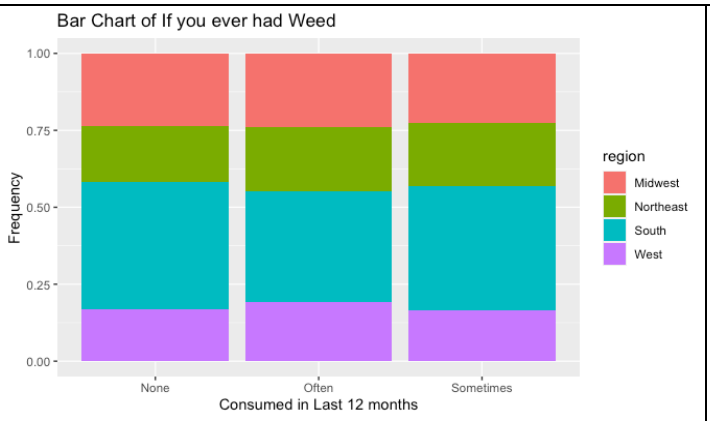
The original data set has 13,713 data points with no missing variables. We have eight predictor variables and seven classes that the students choose from when doing this survey. The classes for the predictor variables were: None, 1-2x, 2-5x, 6-9x, 10-19x, 20- 39x, and 40x. We binned all the variables for by keeping “None” as “None”, next 1-2x, 2-5x and 6-9x were labeled “Sometimes” and everything else was “Often”. For the target variable region, we had four choices: Northeast, Midwest, South and West. For this report we decided to bin the target variable as well, to make it binary, we kept South the same but binned Northeast, Midwest, and West as “Other”.

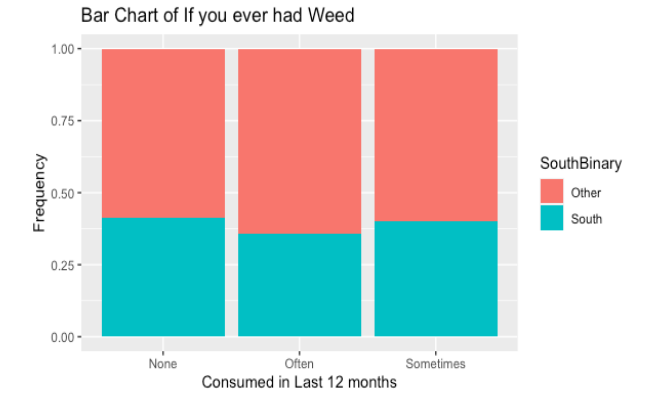
4) Exploratory Data Analysis (EDA).

For EDA we made sure to include the graphs of South and Other so you can see the difference between the two.

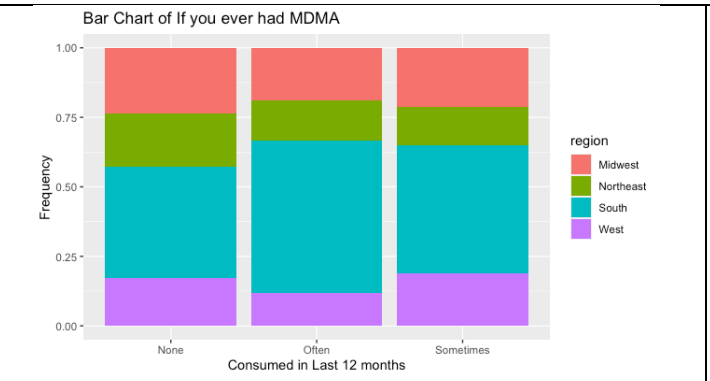


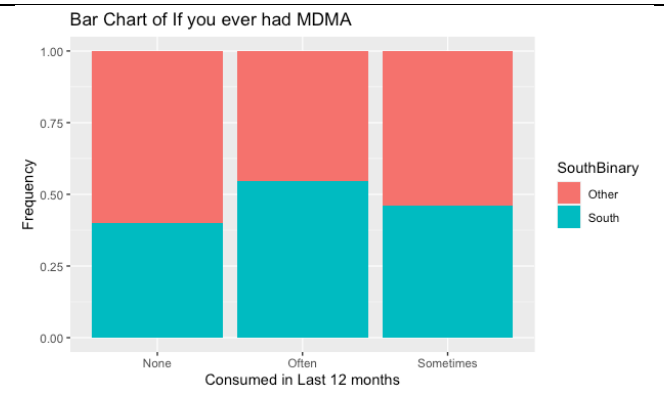
As we can see from this overlay bar graph is that for “None” most students where from the South and this also shows that most students in the Northeast consumed alcohol is the last 12 months is often.



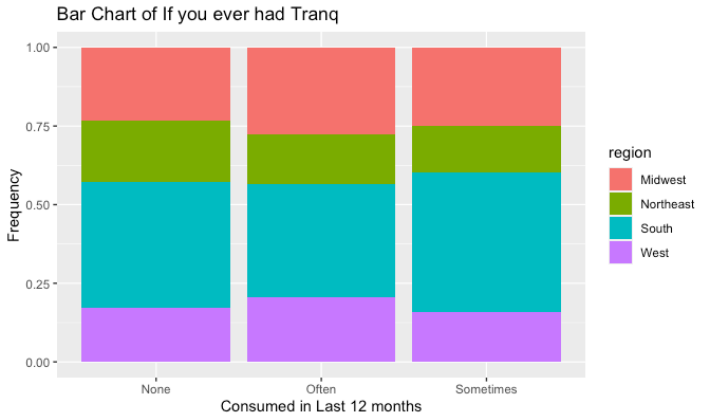


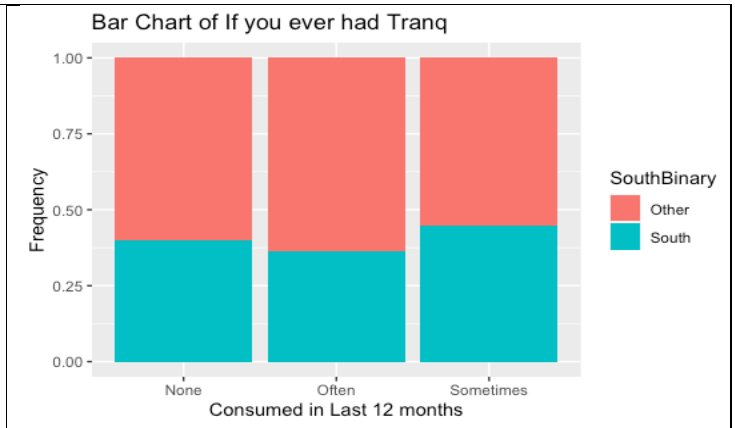
This shows Consumed region vs Weed in last 12 months, we can see most students in the South are the majority in this bar graph.



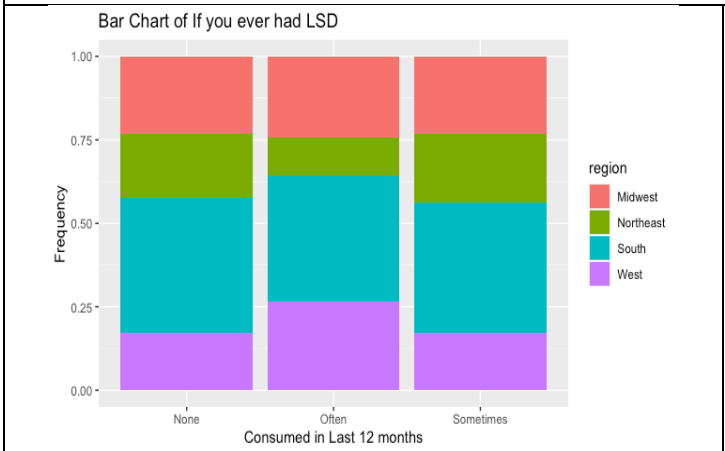


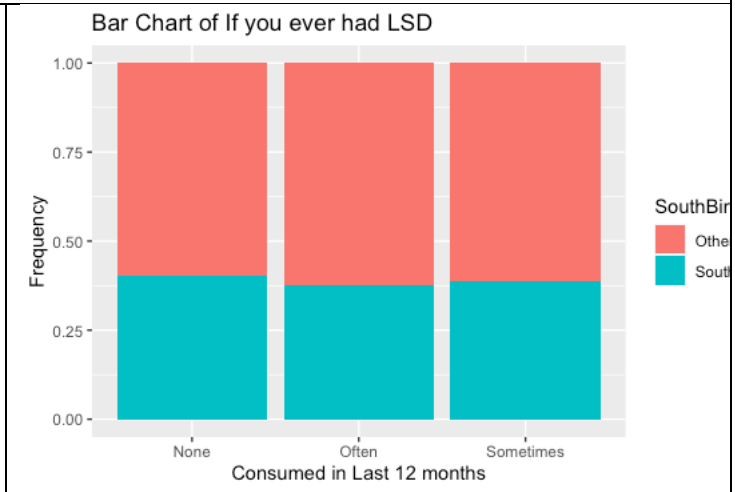
This bar graph is of Students that consumed MDMA in the last 12 months. Its obvious that the majority of students in the South often consume MDMA.



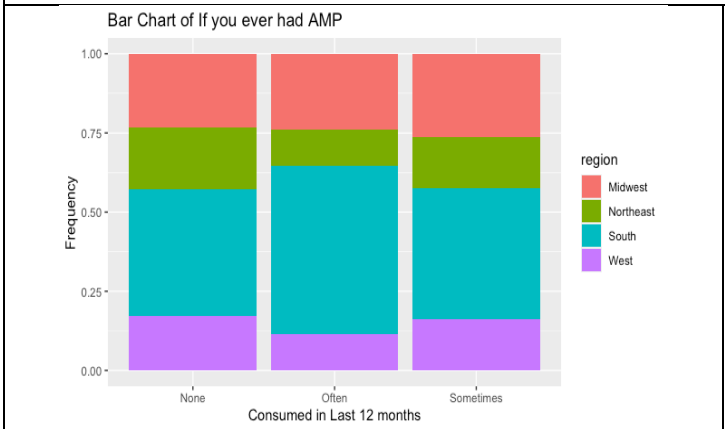


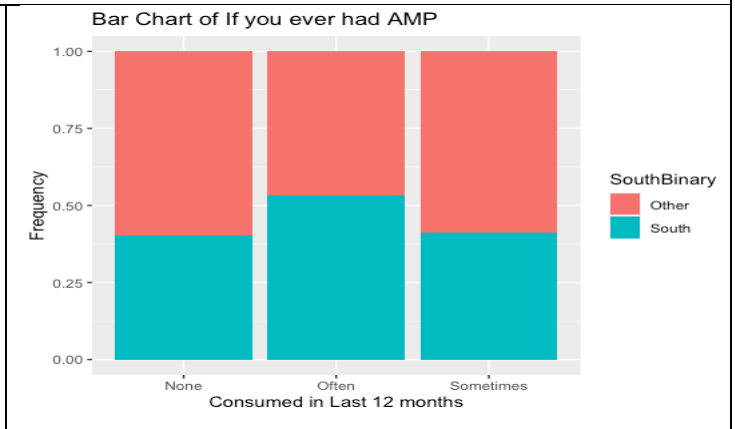
This graph shows Tranq VS Region and as you can see South has dominated this graph as well.



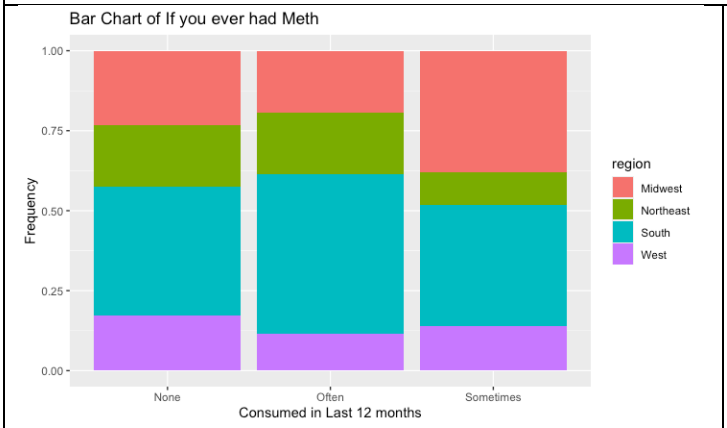


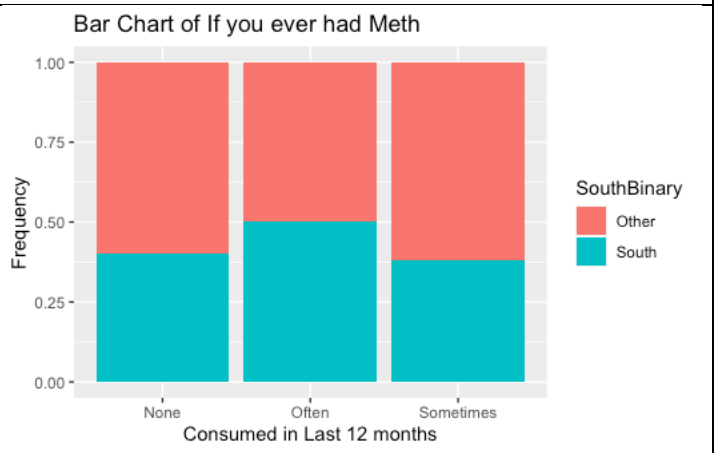
Another graph shows Region VS LSD usage in the last 12 months. It's the same as the other variable the South is dominate in all categories but we can see in "Often" the South and West and close in the amount of students.



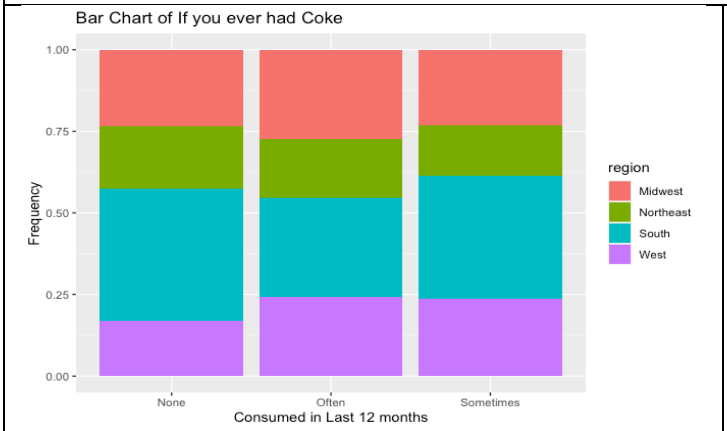


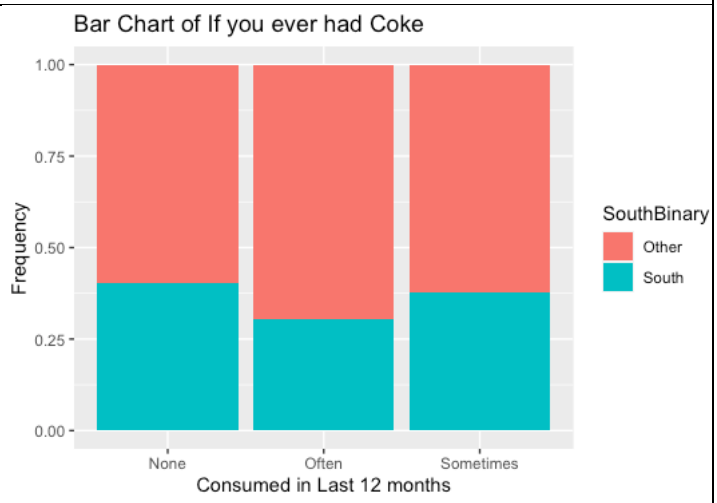
This bar graph shows a AMP consumption in the last 12 months and as we can see again the South is majority in all sections but it's also clear that the northeast consumes the least amount of a AMP “often” in the last 12 months.





This bar graph is showing the relationship between Meth use in the last 12 months and Region. As we can see again the South is really majority of all 3 choices but most students in the South picked “often”.





Lasty, this is the bar graph that shows the relationship between Coke in the last 12 months and Region and again most categories are majority but this time it looks like most students in the South picked “Sometimes”.

Target Variable: Region

```
Other South
 8193  5520
```

Figure 4-1 Other and South

Midwest	Northeast	South	West
3212	2621	5520	2360

Figure 4-2 South, West, Midwest, and Northeast

As you can see from Figure 4-1 and 4-2 from why Other is greater than South but keep in mind that Other are 3 classes together to make it easier to analyze the data.

5) Data Partition and Validation

To prepare the data to be used for the models. This means using the original dataset to create a testing and training dataset. Partitioned the data with 70% for training and 30% for testing dataset. Because we have categorical data, we cannot use the use boxplots to compare between the testing and training datasets or use the kruskal.test to get the p-value. Our dataset has more than two classes for our predictor variables we must use test for the homogeneity of proportions.

The steps for preforming this test are…

1. Make a table of the variables using test and training data.
2. Find the p-value by using the Chi-squared test.
3. Make sure the p-value is greater then .05.

For the target variable because it's just two classes we can use two-sample Z-test for the difference in proportions, to make sure the p-value greater then .05. The p-value we got was 0.9916.

	Other	South
Train	6555	4416
Test	1638	1104

Figure 5-1 Table of Target Variable Training and Testing

Below is a table of all the predictor variables and their tables of testing and training with p-value on the last column.

Variable	P-Value	Table
Alcohol	0.74	None Often Sometime Train 7251 867 2853 Test 1795 214 733
Weed	0.73	None Sometimes Often Train 6905 1845 2221 Test 1744 461 537
LSD	0.73	None Sometimes Often Train 10557 37 377 Test 2646 8 88
MDMA	0.33	None Sometimes Often Train 10716 37 218 Test 2688 5 49
COKE	0.31	None Often Sometime Train 2691 16 35 Test 10744 50 177
AMP	0.76	None Often Sometime Train 2607 31 104 Test 10428 110 433
TRANQ	0.13	None Often Sometime Train 2652 22 68 Test 10584 61 326
METH	0.07	None Often Sometime Train 2728 8 6 Test 10901 18 52

6) Application of Models to Dataset

Model Evaluation on Neural Network

We wanted to see how well our neural network model did by evaluation because we know that our data is skewed to South when compared to our target variable. We only compared it using South Binary because we used it for the other section as well.

	Pre:Midwest	Pre:Northeast	Pre:South	Pre: West	Total
A:Midwest	12	1	790	0	803
A:Northeast	5	0	650	0	655
A: South	9	1	1368	2	1380
A: West	9	3	578	0	590
Total	35	5	3386	2	3428

Figure 6-1 Matrix with all the predictor values; Pre = predicted, A =actual

	Predicted: Other	Predicted: South	Total
Actual: Other	2028	20	2048
Actual: South	1367	13	1380
Total	3395	33	3428

Figure 6-2 Matrix with predictor variable South and Other as Midwest, West and Northeast combined

The accuracy that our model will predict the region that a 12th grader is going a particular drug is about 60%. This means that about 40% of our model will be incorrect (error rate) when predicting what region, a 12th grader is a drug. The sensitivity of the model is 9% which seems correct because we know the model isn't really getting accurate results. Which we know is wrong because the data is skewed. Specificity measures the ability to classify a record negatively with our target variable which is 99% so we know that 99% of the data is wrong. Precision with our model is 39% using region as a target variable which isn't great. F1 = 4%, F2 =2% and F0.5=8%. Please see Figure 6-3.

```
Accuracy:  0.5953909
Error rate:  0.4046091
Sensitivity:  0.00942029
Specificity:  0.9902344
Precision:  0.3939394
F1 score:  0.01840057
F2 score:  0.01170538
F0.5 score:  0.04298942
```

Figure 6-3 Table of Model Evaluation

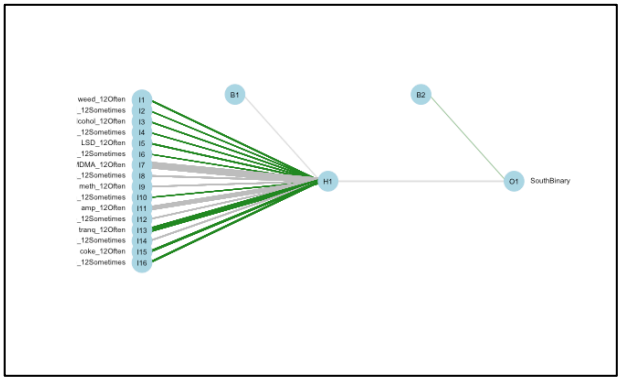


Figure 6-4

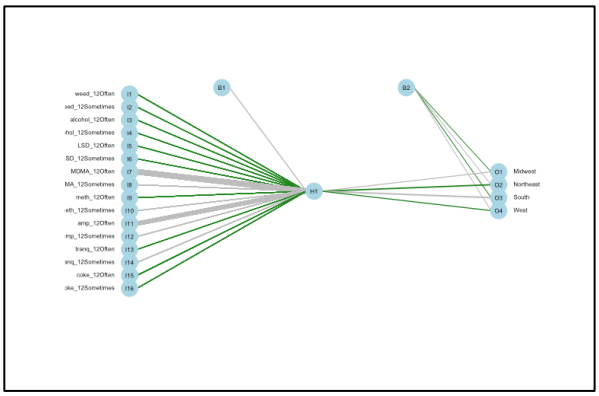


Figure 6-5

Association Rules

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 as “Other”. Below are the tables based on just a few of the more commonly used drugs from the data.

Alcohol:

	None	Often	Sometimes
Other	5334	664	2195
South	3712	417	1391

	None	Often	Sometimes
Other	38.897397	4.842121	16.006709
South	27.069204	3.040910	10.143659

Cannabis/Weed:

	None	Often	Sometimes
Other	5067	1478	1648
South	3582	828	1110

	None	Often	Sometimes
Other	36.950339	10.778094	12.017793
South	26.121199	6.038066	8.094509

LSD:

	None	Often	Sometimes
Other	7881	28	284
South	5322	17	181

	None	Often	Sometimes
Other	57.4710129	0.2041858	2.0710275
South	38.8098884	0.1239700	1.3199154

MDMA:

	None	Often	Sometimes
Other	8030	19	144
South	5374	23	123

	None	Often	Sometimes
Other	58.5575731	0.1385547	1.0500984
South	39.1890906	0.1677241	0.8969591

Cocaine:

	None	Often	Sometimes
Other	8015	46	132
South	5420	20	80

	None	Often	Sometimes
Other	58.4481879	0.3354481	0.9625902
South	39.5245388	0.1458470	0.5833880

From these tables, we can see that most respondents had claimed they never used the drugs. However, for those that did, the majority were those who lived in the South. Below are 2 tables using association rules to help justify the region of the data, which corresponds to the South region as well as the others.

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.

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

Figure 6-6

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 “sometimes” for alcohol use.

7) Next Steps / Conclusion

Working with this data, we know that any model we make will not show us accurate information. We should find data that is current and with a mix of numerical and categorical data, so we are able to use the data to do more evaluations and have better models to compare it with. We don’t want to make any conclusion with the data because it’s skewed to South but after more models, we are figure out why and what the relationship is. We can even use other years to see if there is a connection between the South and a particular drug.

8) Work Citied

Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2019

Tebor, Celina. “100 Bags of Fentanyl Found in Bedroom of Connecticut Teen Who Overdosed at School, Police Say.” *USA TODAY*, www.usatoday.com/story/news/nation/2022/01/26/fentanyl-overdose-connecticut-teenager-bedroom/9233769002/. Accessed 11 May 2023.

Ritchie, Hannah, and Max Roser. “Drug Use.” *Our World in Data*, Dec. 2019, ourworldindata.org/drug-use.

9) Appendix

```
library(nnet)
library(NeuralNetTools)
nnet03 <-  nnet(region ~ weed_12+ alcohol_12+ LSD_12+ MDMA_12 + meth_12
            + amp_12+ tranq_12 +coke_12 , data = data.train, size = 1)

nnet01 <-  nnet(SouthBinary ~ weed_12+ alcohol_12+ LSD_12+ MDMA_12 + meth_12
            + amp_12+ tranq_12 +coke_12 , data = data.train, size = 1)

nnet02 <-  nnet(SouthBinary ~ weed_12+ alcohol_12+ LSD_12+ MDMA_12 + meth_12
            + amp_12+ tranq_12 +coke_12 , data = data.train, size= 4 )
temp = data.train[,-c(11)]
temp1 = data.test[,-c(11)]
nnet03 <- nnet(region~ ., data = temp, size = 1)
pred3 <-  predict(nnet03, newdata = temp1, type = "class")
table(temp1$region, pred3)
plotnet(nnet03)
plotnet(nnet01, cex = 0.5, circle_cex = 4)
plotnet(nnet01, cex = 0.5, circle_cex = 3)
plotnet(nnet03, cex = 0.5, circle_cex = 3, pos_col = "forestgreen",)
plotnet(nnet01, cex = 0.5, circle_cex = 3, neg_col = "maroon")
plotnet(nnet01, cex = 0.5, circle_cex = 3, pos_col = "forestgreen", alpha_val = 0.5)
plotnet(nnet01, cex = 0.5, circle_cex = 3, pos_col = "forestgreen", alpha_val = 0.25)

nnet03$wts
neuralweights(nnet01)
pred1 <- predict(nnet01, newdata = data.test, type = "class")
table(data.test$SouthBinary, pred1)
head(pred1)




#alcohol
table(data.train$alcohol_12);table(data.test$alcohol_12)
x <- matrix(c(7251, 1795, 867,214, 2853, 733), ncol=3)
rownames(x) <- c('Train', 'Test')
colnames(x) <- c("None", "Often", "Sometime")
x <- as.table(x)
y <- chisq.test(x)
y$p.value
##.7399
#region
table(data.train$SouthBinary);table(data.test$SouthBinary)
q<- matrix(c(6555,1638,4416,1104), ncol=2)
rownames(q) <- c('Train', 'Test')
colnames (q) <- c('Other','South')
q <- as.table(q)
```

```
prop.test(q,  correct = FALSE)
### .9916
#weed
table(data.train$weed_12);table(data.test$weed_12)
w <- matrix(c(6905,1744,1845,461,2221,537), ncol=3)
rownames(w) <- c('Train', 'Test')
colnames(w) <- c('None', 'Often', 'Sometime')
w <- as.table(w)
e <- chisq.test(w)
e$p.value
####.730
#LSD
table(data.train$LSD_12);table(data.test$LSD_12)
r <- matrix(c(10557,2646,37,8,377,88), ncol=3)
rownames(r) <- c('Train', 'Test')
colnames(r) <- c("None", "Often", "Sometime")
r <- as.table(r)
u <- chisq.test(r)
u$p.value
##.7832
#MDMA
table(data.test$MDMA_12);table(data.train$MDMA_12)
i <- matrix(c(10716,2688,37,5,218,49), ncol=3)
rownames(i) <- c('Train', 'Test')
colnames(i) <- c("None", "Often", "Sometime")
i <- as.table(i)
o <-chisq.test(i)
o$p.value
#.33333
#COKE
table(data.test$coke_12)
table(data.train$coke_12)
p <- matrix(c(2691,10744,16,50,35,177), ncol=3)
rownames(p) <- c('Train', 'Test')
colnames(p) <- c("None", "Often", "Sometime")
p <- as.table(p)
a <- chisq.test(p)
a$p.value
#.3065

table(data.test$amp_12);table(data.train$amp_12)
d <- matrix(c(2607,10428,31,110,104,433), ncol =3)
rownames(d) <- c('Train', 'Test')
colnames(d) <- c("None", "Often", "Sometime")
d <- as.table(d)
f <- chisq.test(d)
f$p.value
#.7858
 table(data.test$tranq_12);table(data.train$tranq_12)
g <- matrix(c(2652,10584,22,61,68,326), ncol=3)
rownames(g) <- c('Train', 'Test')
```

```
colnames(g) <- c("None", "Often", "Sometime")
g <- as.table(g)
h <- chisq.test(g)
h$p.value
##.1315

table(data.test$meth_12); table(data.train$meth_12)
j <- matrix(c(2728,10901,8,18,6,52), ncol = 3)
rownames(j) <- c('Train', 'Test')
colnames(j) <- c("None", "Often", "Sometime")
j <- as.table(j)
k <- chisq.test(j)
k$p.value
##.0718

library(ggplot2)

ggplot(data, aes(alcohol_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had Alcohol")

ggplot(data, aes(weed_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had Weed")

ggplot(data, aes(MDMA_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had MDMA")

ggplot(data, aes(tranq_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had Tranq")

ggplot(data, aes(LSD_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had LSD")

ggplot(data, aes(amp_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
```

```
  ggtitle("Bar Chart of If you ever had AMP")

ggplot(data, aes(meth_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had Meth")

ggplot(data, aes(coke_12)) +
  geom_bar(aes(fill = SouthBinary), position = "fill") +
  xlab("Consumed in Last 12 months") +
  ylab("Frequency") +
  ggtitle("Bar Chart of If you ever had Coke")



table(data$SouthBinary)
table(data$region)
```