Perform Naïve Bayes classification on "Monitoring the Future" Study

Project Stage 3

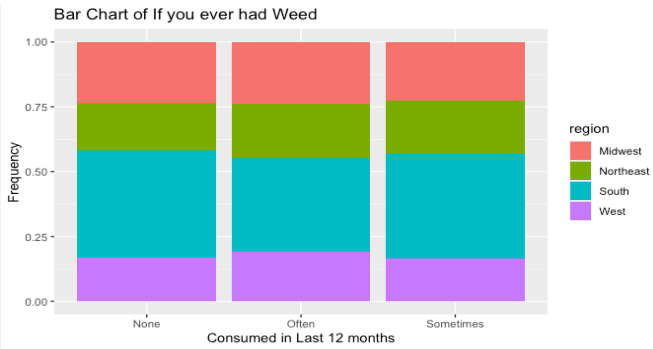
Aretha Kassegnin,

MAT 343 – Explorations In Data Analytics

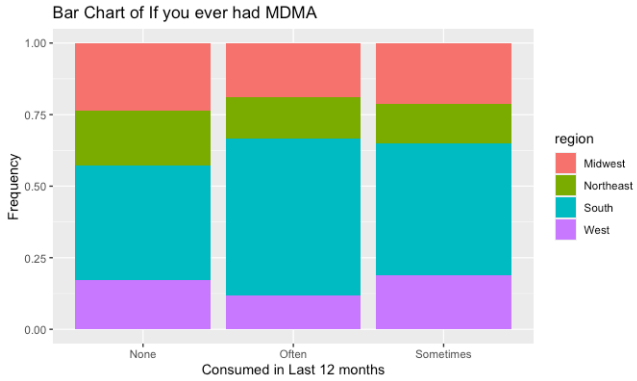
1) Note: This is the same as what we did on Project Stage 2, we just removed all the NA data by using proportions and all our data is categorical.

Below is our EDA

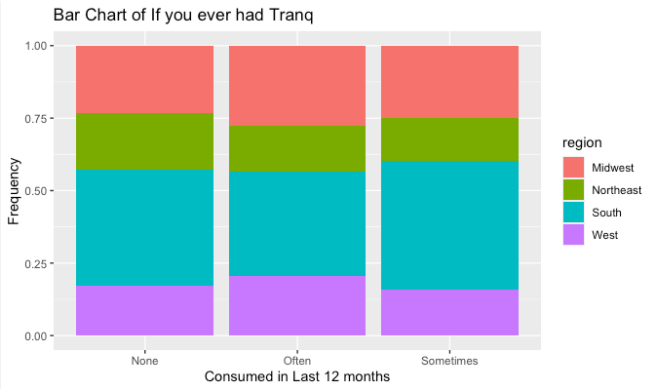
As we can see from this overlay bar graph is that for "None" most students where from the South and this also shows that most students in the Northeast consumed alcohol is the last 12 months is often.



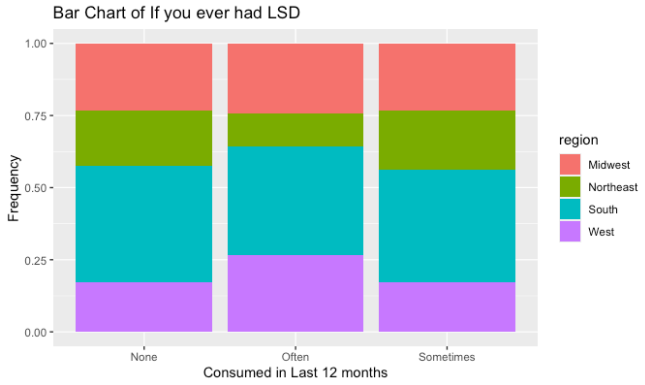
This shows Consumed region vs Weed in last 12 months, we can see most students in the South are the majority in this bar graph.



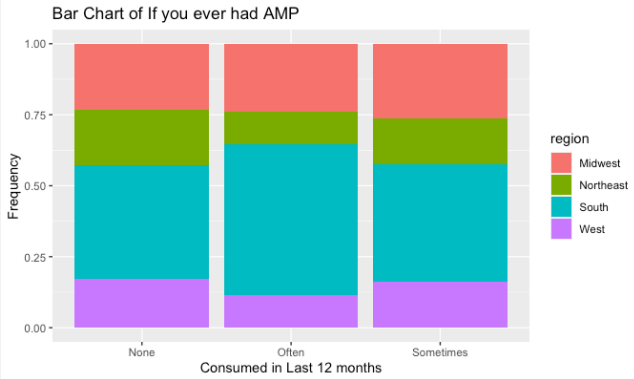
This bar graph is of Students that consumed MDMA in the last 12 months. Its obvious that the majority of students in the South often consume MDMA.



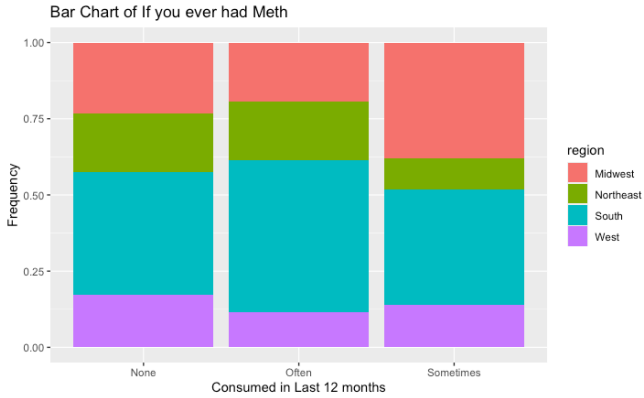
Next, we have a bar graph that shows the relationship between Tranq in the last 12 months and Region and as can see like the others South has majority in all categories with Midwest being the 2nd with the region that often uses this drug.



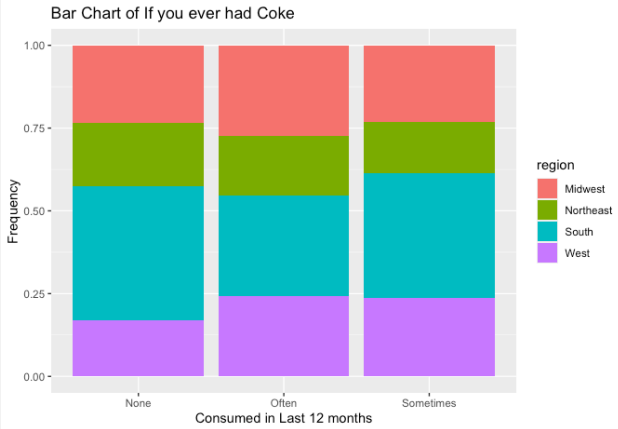
Another graph shows Region VS LSD usage in the last 12 months. It's the same as the other variable the South is dominate in all categories but we can see in "Often" the South and West and close in the amount of students.



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".

2) Create some tables.

```
A-priori probabilities:
Y
  Midwest Northeast     South      West
0.2342303 0.1911325 0.4025377 0.1720995

Conditional probabilities:
           alcohol_12
Y                None      Often  Sometimes
  Midwest   0.65317559 0.08125778 0.26556663
  Northeast 0.64212133 0.09004197 0.26783670
  South     0.67246377 0.07554348 0.25199275
  West      0.65805085 0.07076271 0.27118644

           weed_12
Y                None     Often Sometimes
  Midwest   0.6348070 0.1718555 0.1933375
  Northeast 0.5997711 0.1835177 0.2167112
  South     0.6489130 0.1500000 0.2010870
  West      0.6169492 0.1885593 0.1944915

           LSD_12
Y                 None       Often   Sometimes
  Midwest   0.962951432 0.003424658 0.033623910
  Northeast 0.961465090 0.001907669 0.036627242
  South     0.964130435 0.003079710 0.032789855
  West      0.961016949 0.005084746 0.033898305

           MDMA_12
Y                 None       Often   Sometimes
  Midwest   0.979763387 0.002490660 0.017745953
  Northeast 0.983594048 0.002289203 0.014116749
  South     0.973550725 0.004166667 0.022282609
  West      0.976694915 0.002118644 0.021186441
```

```
           coke_12
Y                 None       Often   Sometimes
  Midwest   0.979140722 0.005603985 0.015255293
  Northeast 0.982830981 0.004578405 0.012590614
  South     0.981884058 0.003623188 0.014492754
  West      0.972033898 0.006779661 0.021186441

           amp_12
Y                 None       Often   Sometimes
  Midwest   0.945205479 0.010585305 0.044209215
  Northeast 0.961083556 0.006104540 0.032811904
  South     0.946195652 0.013586957 0.040217391
  West      0.956355932 0.006779661 0.036864407

           tranq_12
Y                 None       Often   Sometimes
  Midwest   0.962328767 0.007160648 0.030510585
  Northeast 0.972911103 0.004959939 0.022128958
  South     0.962681159 0.005434783 0.031884058
  West      0.966525424 0.007203390 0.026271186

           meth_12
Y                 None       Often   Sometimes
  Midwest   0.991594022 0.001556663 0.006849315
  Northeast 0.995803129 0.001907669 0.002289203
  South     0.993659420 0.002355072 0.003985507
  West      0.995338983 0.001271186 0.003389831
```

Based on what we saw in question 1. All these tables make sense for all the target variable. South would have majority but we should see at least some in at least in Midwest as well because its 2nd but that's what we saw when we tried to find the baseline but we ran into an issue.

```
          test.pred
           Midwest Northeast South West
Midwest        17         0  3194    1
Northeast       2         0  2615    4
South          13         0  5506    1
West            8         0  2347    5
```

As you can see Northeast is all 0 which doesn't make any sense because there are some variables where Northeast isn't strong but its not zero. Below we can see that West as the least amount so it doesn't make logical sense that Northeast is zero.

```
Midwest Northeast     South      West
   3212      2621      5520      2360
```

So we tried another way where we keep South the same and combining all other categories (Midwest, Northeast and West). This was the result.

```
Other South
 8193  5520
```

When we recreate the tables, we get

```
A-priori probabilities:
Y
    Other     South
0.5974623 0.4025377

Conditional probabilities:
       alcohol_12
Y             None      Often  Sometimes
  Other 0.65104357 0.08104479 0.26791163
  South 0.67246377 0.07554348 0.25199275

       weed_12
Y             None     Often Sometimes
  Other 0.6184548 0.1803979 0.2011473
  South 0.6489130 0.1500000 0.2010870

       LSD_12
Y              None       Often   Sometimes
  Other 0.961918711 0.003417552 0.034663737
  South 0.964130435 0.003079710 0.032789855

       MDMA_12
Y              None       Often   Sometimes
  Other 0.980104968 0.002319053 0.017575979
  South 0.973550725 0.004166667 0.022282609

       coke_12
Y              None       Often   Sometimes
  Other 0.978274136 0.005614549 0.016111315
  South 0.981884058 0.003623188 0.014492754

       amp_12
Y              None       Often   Sometimes
  Other 0.953496888 0.008055657 0.038447455
  South 0.946195652 0.013586957 0.040217391

       tranq_12
Y              None       Often   Sometimes
  Other 0.966922983 0.006468937 0.026608080
  South 0.962681159 0.005434783 0.031884058

       meth_12
Y              None       Often   Sometimes
  Other 0.994019285 0.001586720 0.004393995
  South 0.993659420 0.002355072 0.003985507
```

As we can see everything looks okay but when we make the table we get something unexpected happens

```
        test2.pred
         Other South
Other  8138    55
South  5454    66
```

The table kind of flips and South was less predicted values which do not make any sense. We concluded that performing Naïve Bayes classification on our dataset is not possible because the data is skewed and even though the calculations are correct. Something is pulling the data and because we cannot pinpoint was it is, we are not able to get an accurate baseline or accuracy.

Appendix

```
library(e1071)

data$SouthBinary <- as.factor(ifelse(data$region == "South", "South", "Other"))
head(data)
names(data)
test <- naiveBayes(formula = region ~ alcohol_12 + weed_12 +
                   LSD_12 + MDMA_12 + coke_12 + amp_12 +tranq_12 +
                   meth_12, data = data)
test
test.pred <- predict(object = test, newdata = data)
test.pred

table(data$region, test.pred)
table(data$region)
table(data$SouthBinary)

test2 <- naiveBayes(formula = SouthBinary ~ alcohol_12 + weed_12 +
                   LSD_12 + MDMA_12 + coke_12 + amp_12 +tranq_12 +
                   meth_12, data = data)
test2.pred <- predict(test2, newdata = data)
table(data$SouthBinary, test2.pred)

library(caret)
set.seed(325)


predict(object = test2,
      newdata = data("region" = “South” ))
```