Applying Neural Network to the 2019 "Monitoring the Future" Study

Project Stage 02

24 February 2023

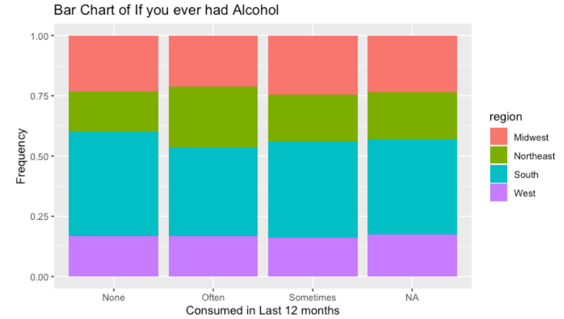
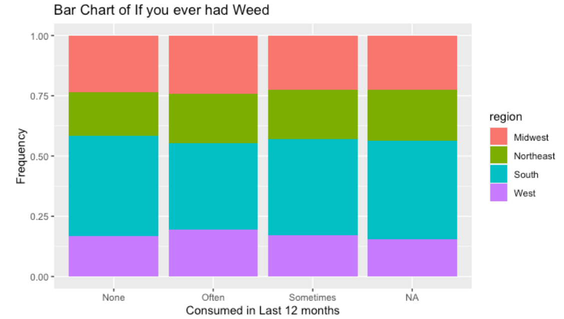
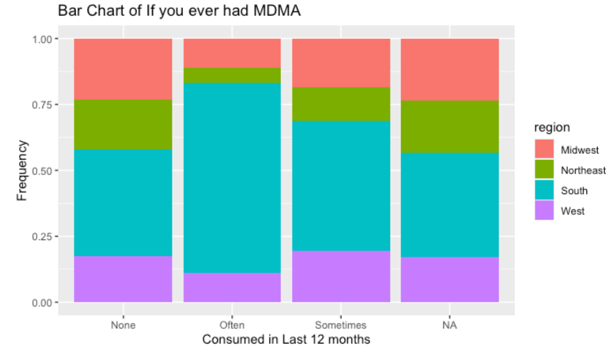
Aretha Kassegnin, [redacted]

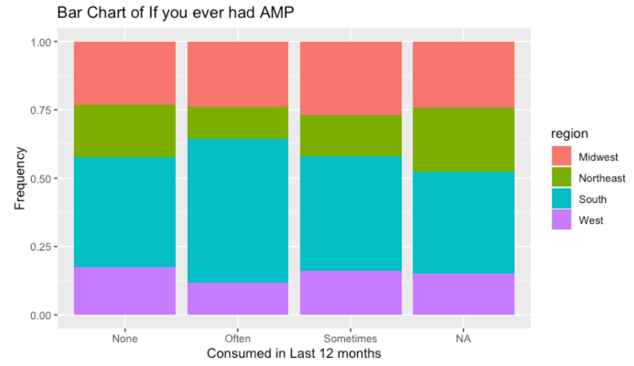
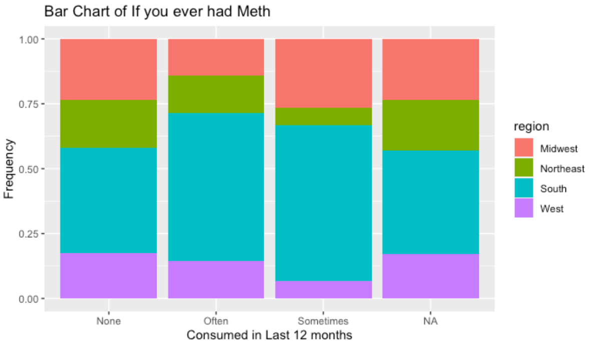
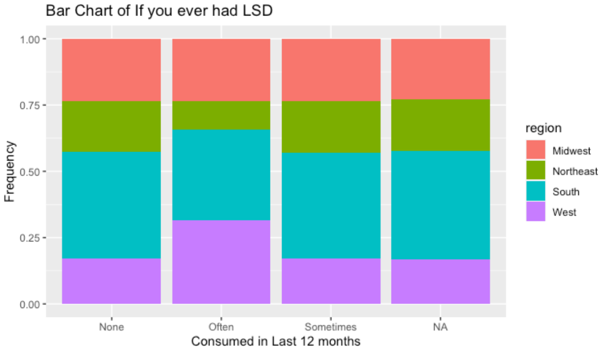
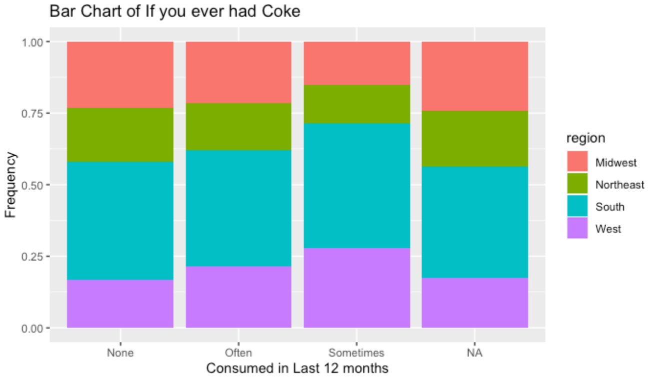
MAT 343 – Explorations In Data Analytics

1. Use exploratory data analysis (EDA) (e.g. overlaid histograms) to estimate the impact of your predictor variables on the target variable.

By creating overlaid histograms about the number of different cases in different regions, we can observe that:

1. Alcohol abuse appears more in northeast region.
2. Weed abuse appears more in northeast and west region.
3. Cocaine abuse appears slightly more in west region.
4. LSD and MDMA abuse cases spread evenly among all regions, we cannot make sure they are affected by regions.

 Bar Chart of If you ever had Alcohol	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.
 Bar Chart of If you ever had Weed	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.
 Bar Chart of If you ever had MDMA	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.

 Bar Chart of If you ever had AMP 	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.
 Bar Chart of If you ever had Meth 	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".
 Bar Chart of If you ever had LSD 	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.
 Bar Chart of If you ever had Coke 	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".

Please note: NA means blank or questions that were not answered in each region. We have not found a way to remove that in the dataset.

2.

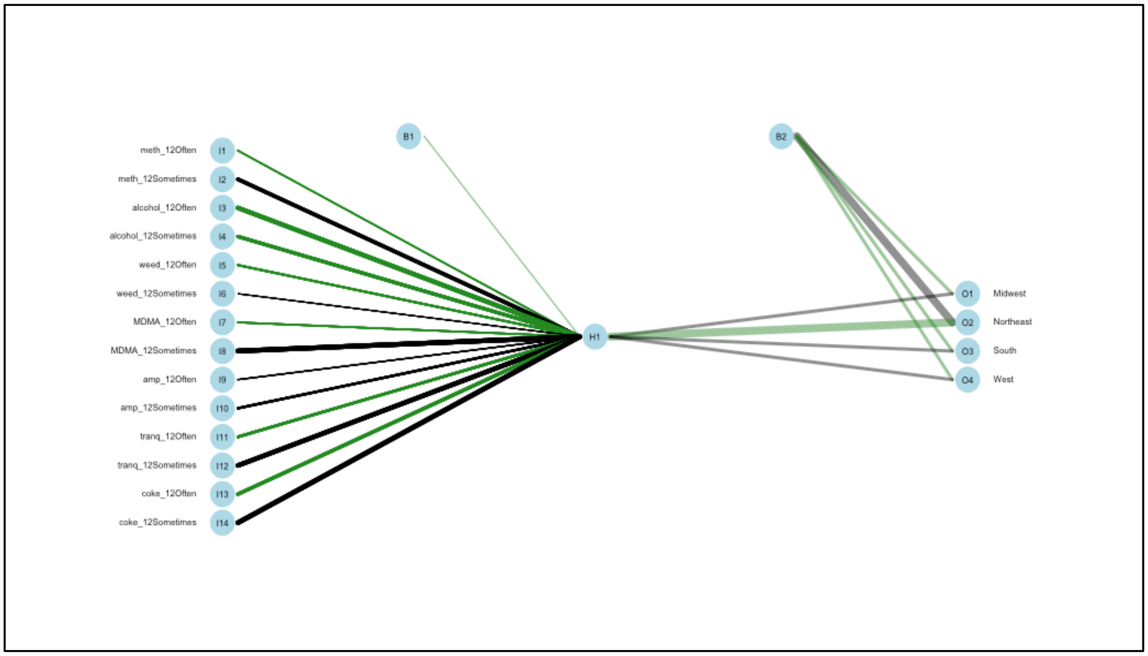
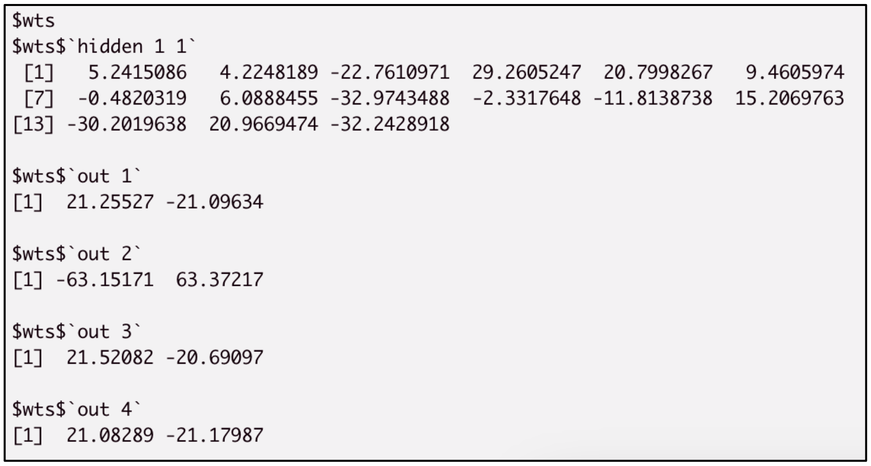


Figure 1: -This shows a Neural Network of our data green is positive and black is negative

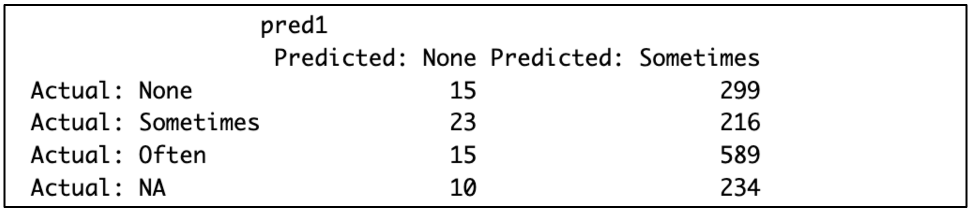


```
$wts
$wts$`hidden 1 1`
 [1]   5.2415086   4.2248189 -22.7610971  29.2605247  20.7998267   9.4605974
 [7]  -0.4820319   6.0888455 -32.9743488  -2.3317648 -11.8138738  15.2069763
[13] -30.2019638  20.9669474 -32.2428918

$wts$`out 1`
[1]  21.25527 -21.09634

$wts$`out 2`
[1] -63.15171  63.37217

$wts$`out 3`
[1]  21.52082 -20.69097

$wts$`out 4`
[1]  21.08289 -21.17987
```

Figure -1 Weights for Figure 1

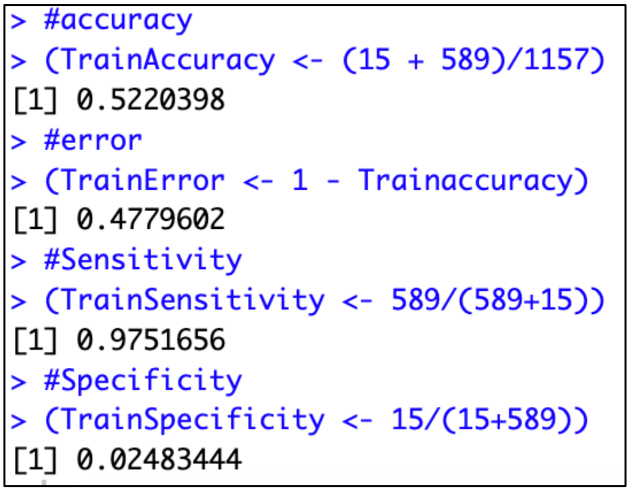
The nodes with the most weights cocaine (sometimes) with -32 and tranquilizer (sometimes) with -30. I believe that it is because they do not play a big part in the data. As we saw, the South has a lot of students that answered.

3.



```
                    pred1
                     Predicted: None Predicted: Sometimes
Actual: None                      15                  299
Actual: Sometimes                 23                  216
Actual: Often                     15                  589
Actual: NA                        10                  234
```

Figure 3: Table of Actual vs. Predicted Values



```
> #accuracy
> (TrainAccuracy <- (15 + 589)/1157)
[1] 0.5220398
> #error
> (TrainError <- 1 - Trainaccuracy)
[1] 0.4779602
> #Sensitivity
> (TrainSensitivity <- 589/(589+15))
[1] 0.9751656
> #Specificity
> (TrainSpecificity <- 15/(15+589))
[1] 0.02483444
```

Figure 4: Calculations of Accuracy, Error, Sensitivity, Specificity

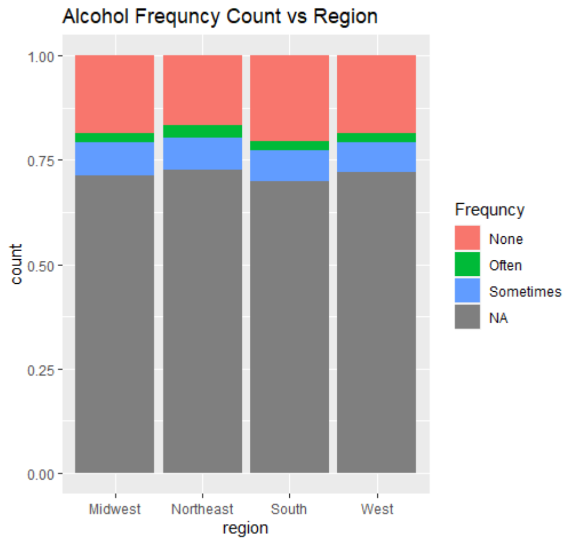
4. Baseline for this model was that 40% of substance use would come from the South and our model is predicting 52% of the data correctly which isn't great. This model will not work for this data set.

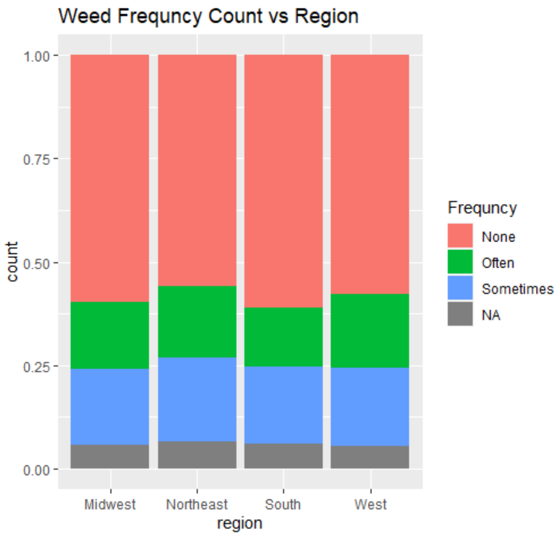
Appendix

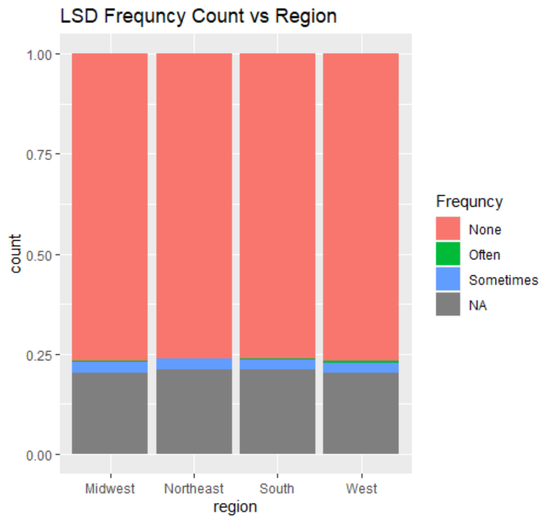
```
#EDA with ggplot
library(caret)
mdata <- MTFData

ggplot(mdata,aes(x=region,fill=factor(alcohol_12)))+
  geom_bar(position = "fill")+
  ggtitle("Alcohol Frequncy Count vs Region")+
  guides(fill=guide_legend(title="Frequncy"))

ggplot(mdata,aes(x=region,fill=factor(weed_12)))+
  geom_bar(position = "fill")+
  ggtitle("Weed Frequncy Count vs Region")+
  guides(fill=guide_legend(title="Frequncy"))
```

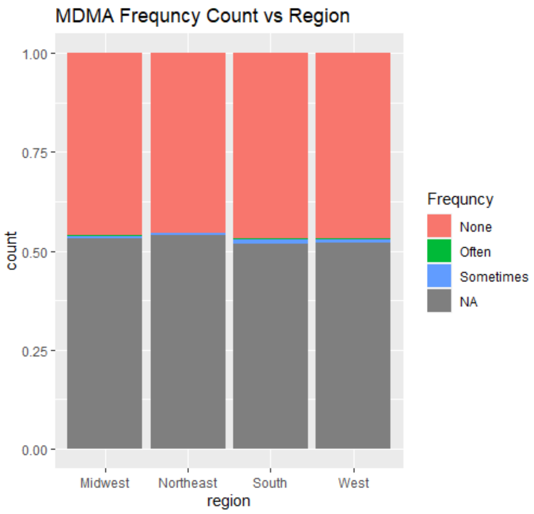
```
ggplot(mdata,aes(x=region,fill=factor(LSD_12)))+
  geom_bar(position = "fill")+
  ggtitle("LSD Frequncy Count vs Region")+
  guides(fill=guide_legend(title="Frequncy"))

ggplot(mdata,aes(x=region,fill=factor(MDMA_12)))+
  geom_bar(position = "fill")+
  ggtitle("MDMA Frequncy Count vs Region")+
  guides(fill=guide_legend(title="Frequncy"))

ggplot(mdata,aes(x=region,fill=factor(coke_12)))+
  geom_bar(position = "fill")+
  ggtitle("Cocaine Frequncy Count vs Region")+
  guides(fill=guide_legend(title="Frequncy"))

library(caret)
set.seed(25)
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)

boxplot(data.all$region ~ (trainortest),
        data = data.all)
boxplot(data.all$coke_12 ~ (trainortest),
        data = data.all)
boxplot(data.all$weed_12 ~ (trainortest),
        data = data.all)
```

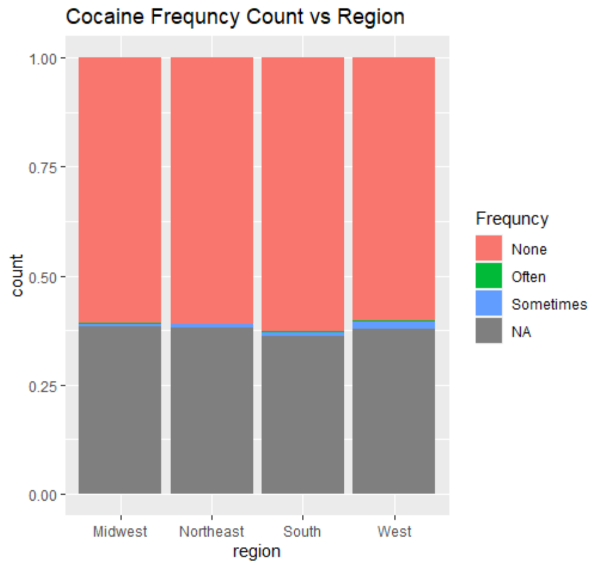
```
boxplot(data.all$LSD_12 ~ (trainortest),
      data = data.all)
boxplot(data.all$alcohol_12 ~ (trainortest),
      data = data.all)
boxplot(data.all$MDMA_12 ~ (trainortest),
      data = data.all)
boxplot(data.all$amp_12 ~ (trainortest),
      data = data.all)

library(nnet)
library(NeuralNetTools)
#everything is the same I
nnet01 <-  nnet(region ~ meth_12+alcohol_12+weed_12+MDMA_12
            + amp_12+ tranq_12 +coke_12 , data = data.train, size = 1)
#plot model
plotnet(nnet01, cex = 0.5, circle_cex = 3, pos_col = "forestgreen", alpha_val = 0.5,
      neg_col = "black")
#obtain weights
nnet01$wts
neuralweights(nnet01)
```


Alcohol Frequncy Count vs Region
count
1.00
0.75
0.50
0.25
0.00
Midwest
Northeast
South
West
region
Frequncy
None
Often
Sometimes
NA


Weed Frequncy Count vs Region
count
1.00
0.75
0.50
0.25
0.00
Midwest
Northeast
South
West
region
Frequncy
None
Often
Sometimes
NA


LSD Frequncy Count vs Region
count
1.00
0.75
0.50
0.25
0.00
Midwest
Northeast
South
West
region
Frequncy
None
Often
Sometimes
NA


MDMA Frequncy Count vs Region
count
1.00
0.75
0.50
0.25
0.00
Midwest
Northeast
South
West
region
Frequncy
None
Often
Sometimes
NA


Cocaine Frequncy Count vs Region
count
1.00
0.75
0.50
0.25
0.00
Midwest
Northeast
South
West
region
Frequncy
None
Often
Sometimes
NA