Drug consumption:
A study on hard/soft drug use and its predictors

Dr. Musal

Alexander Hernandez, Milo Paciorek, Ryan Rogers

Executive Summary:

With our data on drug consumption for individuals recorded with certain independent variables such as age, personality scores, education, etc., we wanted to test the hypothesis that certain individuals are more likely to try hard or soft drugs based on their other traits. In doing so we find some interesting results that allow us to make generalizable statements about the population. Our goal is to target certain individuals who have a higher risk of trying drugs within the past year to effectively target anti-drug programs and communicate the risks of drug use and addiction to a target audience that makes sense. With our data, we built two different models, one that predicts if an individual has used soft drugs in the past year and another that predicts if they have used hard drugs in the past year.

Our soft drug model has an accuracy of 82% with a 90% chance of correctly identifying if someone used soft drugs. The hard drug model has a 78% accuracy, with a 73% chance of correctly identifying if someone has used hard drugs.[1] A few variables in the model display a strong p-value indicating the rejection of the null hypothesis that that variable has no coefficient effect for the probability of drug use, so we can use those variables to make a few statements. Our first major conclusion is soft drugs and hard drugs have a strong positive correlation, and on average an individual doing one is much more likely to be doing the other. With this, we can communicate the possibility that softer drugs can act as a gateway into harder drugs and try to prevent people from using soft drugs. Also, as the age group increases the coefficient effect on probability decreases, so we can aim our campaigns to younger crowds. Another thing to consider under broad terms is males are more likely to partake in soft and hard drugs given the same traits as a female. The higher one's education level, the lower the probability of doing soft drugs is also important to note, that there is not enough evidence to show that there is an effect on hard drugs, so we must not exclude individuals with higher education from guidance on drug use.

Our models give us very good information concerning the personality scores, which we cleaned and ordered to be able to make better sense of. We are not surprised that the more open a person is to new things based on O-score, the more likely they are to do soft drugs, but their

[1] Appendix I

openness is not reflected on hard drugs. What is seen on hard drugs rather, is that the higher a person's N-score (Neuroticism, a person's measure of how moody and emotional they are, higher neuroticism makes a person more likely to experience anxiety, anger, frustration, depression, loneliness, and fear.) would on average lead to a higher probability of doing hard drugs in the past year but has an unlikely effect on the use of soft drugs. Based on the interpretation of our variables, our best practice would be to target young adults with sensitive mental health, mental health issues, and/or experience with soft drugs to explain the dangers of hard drugs and addictions. Doing so would maximize the probability of the audience being high-risk and needing guidance.

Problem and Data Description:

In this study, we used a dataset from Kaggle with several different variables with that of most interest being the list of drugs. The whole dataset included 1,885 individuals. Along with basic info like age, country, education level, and ethnicity, there were values for the big 5 personality traits (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness) along with impulsivity, and sensation seeking. Also included were 19 drugs and values from CL0-CL6 ranging between, never having tried the drug- tried on the last day. We divided the drugs present into 2 binary categories. The first is whether the individual has or hasn't tried SOFT drugs in the past year. The second is whether the individual has tried HARD drugs in the past year.

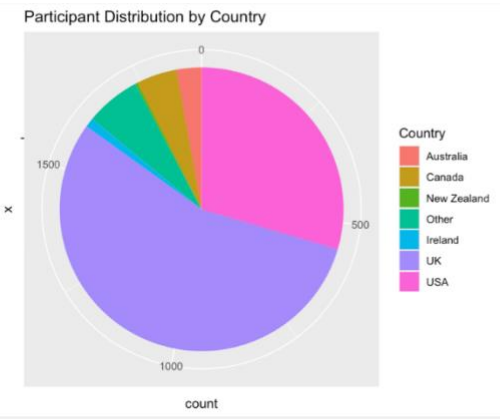
The cleaning involved changing the values from virtually every category into labels that could be easily interpreted. For example, the value for males was "-0.48246" and for female was "0.48246" so we needed to change them to factors with their respective names.[2] After that was completed, we needed to adjust the CL" " values for the time that the drug was last tried and assign them 1 if tried in the last year and 0 if not. This required assigning values CL0-CL2 to be 0 and assigning CL3-CL6 to 1. This turned our target data into binary in order to run a logistical regression model. Note that the drugs in this process aren't all the drugs in the dataset, we

[2] Appendix B

removed the ones that we found to be erroneous in our study. It's also separated above between what we classified as hard and soft drugs. Hard drugs are crack, meth, heroin, coke, ketamine, amphetamine, and benzos (high risk and addiction). Soft drugs are cannabis, nicotine, mushrooms, and LSD (hallucinogens/ ones with no to very small overdose risk) Next, we separated the hard and soft drug use into their own 2 variables in the data frame. Lastly, as far as data cleansing, we used mapping to adjust the "Big 5" personality scores to values more easily interpreted. This involved assigning each value present to ascending values starting at 1-however many scores were present.

Describing and visualizing Data:

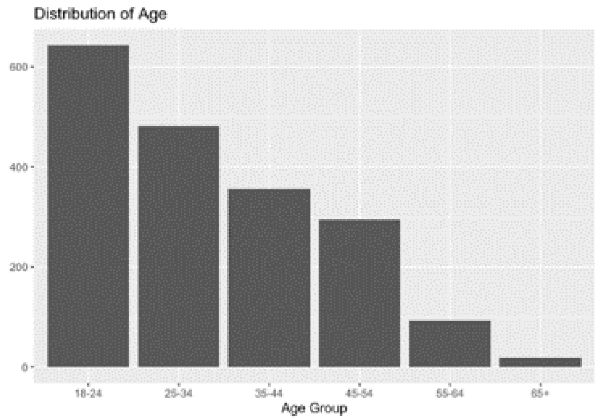
To better understand the data before our analysis we made a few simple charts and graphs to describe the distribution between the variables we plan to use. This will also give us visual representations to refer to when analyzing our regressions and how the models will influence the general public.



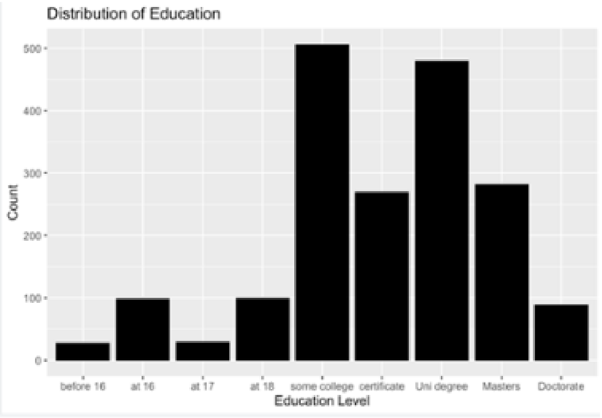
[3]This pie chart shows the distribution of people in each of the countries which assists in preliminary analysis in which areas we have data for and where to target our efforts.

Below we have bar charts just to visualize the distribution of the age groups and the education levels of people in the dataset. This helps us further narrow what groups we might already have good contact with to adjust where our efforts might need to be focused on after our analyses.

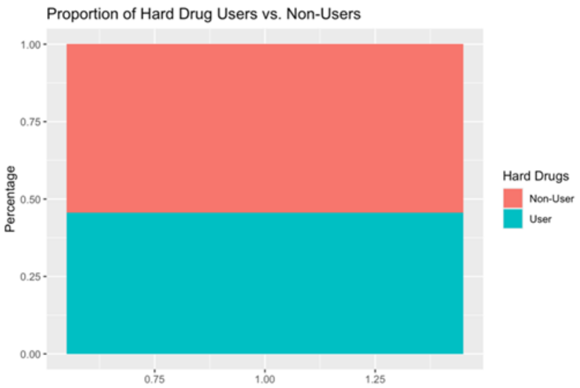
[3] Appendix J



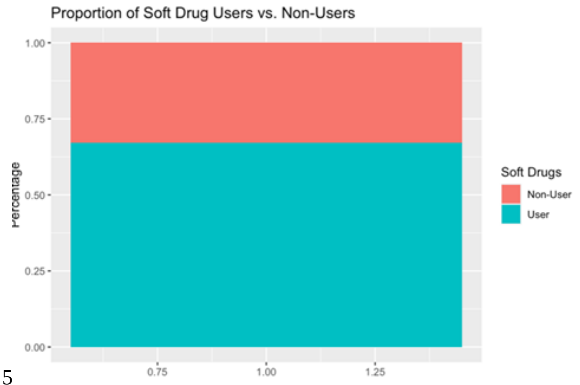
[4]



Lastly for the basic info we have these two stacked bar charts to show that there are much more soft drug users compared to the hard drug users which could contribute to understanding the magnitude of the problem we'd like to solve.

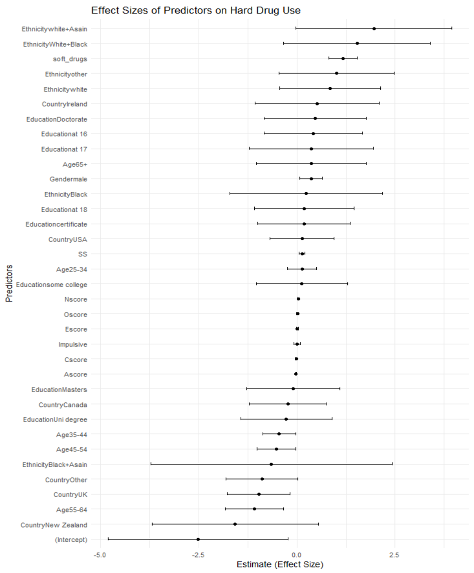


[5]



[4] Appendix J

[5] Appendix J



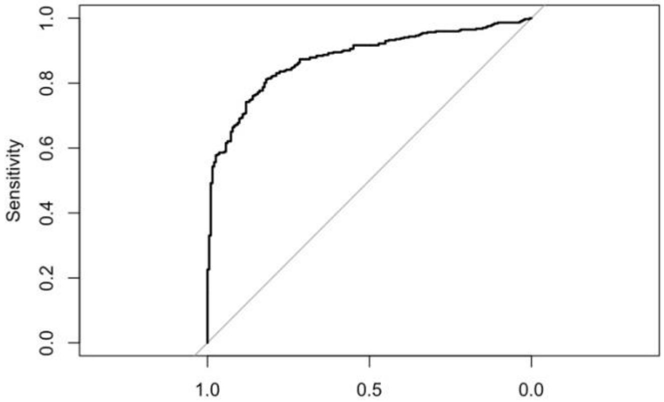
[6]After analysis we've crafted these visuals to help us understand the causes related to hard and soft drug use as referring to the size of their effects. The following coefficient plot visualizes the effect size of each predictor in our logistic regression model. It shows the estimates and its corresponding confidence interval, highlighting significant predictors.

Analysis:

To begin our analysis, we partitioned our data with a 70/30 split.[7] Our training set consisted of 565 observations and our testing had 1320. We used the step-AIC strategy to effectively find which variables should be included in each model. The step AIC ran in both directions to run through the process of dropping and re-adding variables while testing their AIC (Akaike Information Criterion) to see whether it decreased, leaving the model with the lowest one.[8] Note that the variables it kept are different, soft drugs kept education and a few personality scores that hard drugs didn't.[9] After using the AIC to craft our model based on the training data, we received our models and their coefficients. To make more generalizable statements, we ran the training model on the unseen testing data to see a realistic accuracy on data that the model was not trained on.

[6] Appendix L
[7] Appendix D
[8] Appendix J
[9] Appendix J

However, before we compare accuracy, it is important that we use effective cutoff points on what we consider predicted to use drugs in the past year versus not based on their probability. For soft drugs, we found Youden's index helpful, as it gave us the best accuracy, sensitivity, and specificity. However, Youden's index for hard drugs was not very helpful, giving a low cutoff point that showed an undesired sensitivity and low accuracy. Going with the classic cutoff of 0.5 for hard drugs resulted in more useful predictions with better accuracy and sensitivity.[10]



Youden's index visual for soft drugs

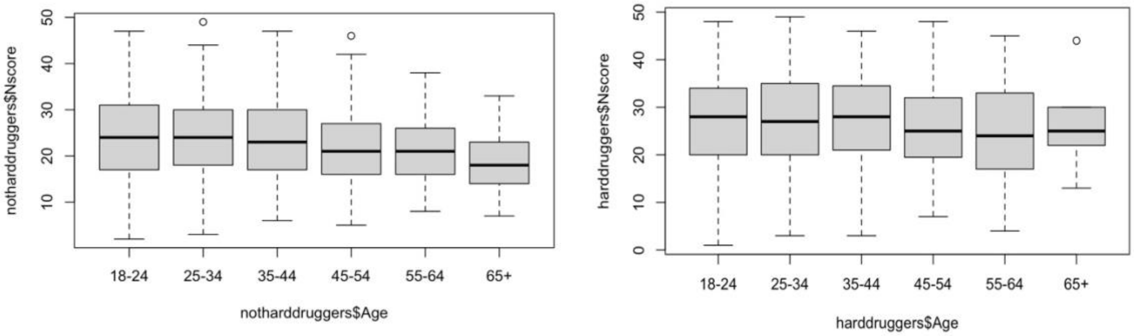
When interpreting the coefficients that are displayed in the result, it is important to use the right interpretation of them when considering their effects. For example, the coefficient for soft drugs on hard drugs is (1.275387), that is a factor as a yes or no and is added in the logit(p) function. However, the coefficient for a personality score like N-score (0.035977), would be multiplied by the value of the N-score that is observed when put in the logit(P) function.[11]

By developing logistic regression models tailored to different types of drugs, we were able to achieve insights into the factors that influence substance use behaviors. Additionally, it holds the potential to offer insights into future educational attainment levels. In exploring the interplay between personality traits and drug use, the personality framework offers valuable insights. Traits like Openness may predispose individuals to experiment with drugs due to their inherent curiosity and desire for new experiences, while high Conscientiousness often correlates with restraint and a lower likelihood of drug use. Individuals with high Neuroticism might use

[10] Appendix I

[11] Appendix H

drugs as a coping mechanism for emotional instability. Developing this correlation between Education and the personality traits of individuals with how drug consumption impacts an individual's performance academically.



N-Score vs Age Boxplot,
Hard Drug user’s average is higher across.

Summary

Overall, this data of 1885 individuals spans 12 attributes, encompassing personality traits, demographics, and substance use history. To refine our analysis, we adjusted the dataset, eliminating ambiguous variables and aligning it with our research objectives. Our primary goal was to develop logistic regression models tailored to differentiate between hard drugs like cocaine and soft drugs like cannabis. This allowed us to explore the intricate relationship between substance use and educational attainment. We employed the 70/30 split to validate our predictive model. This split ensures our model was able to be effective.

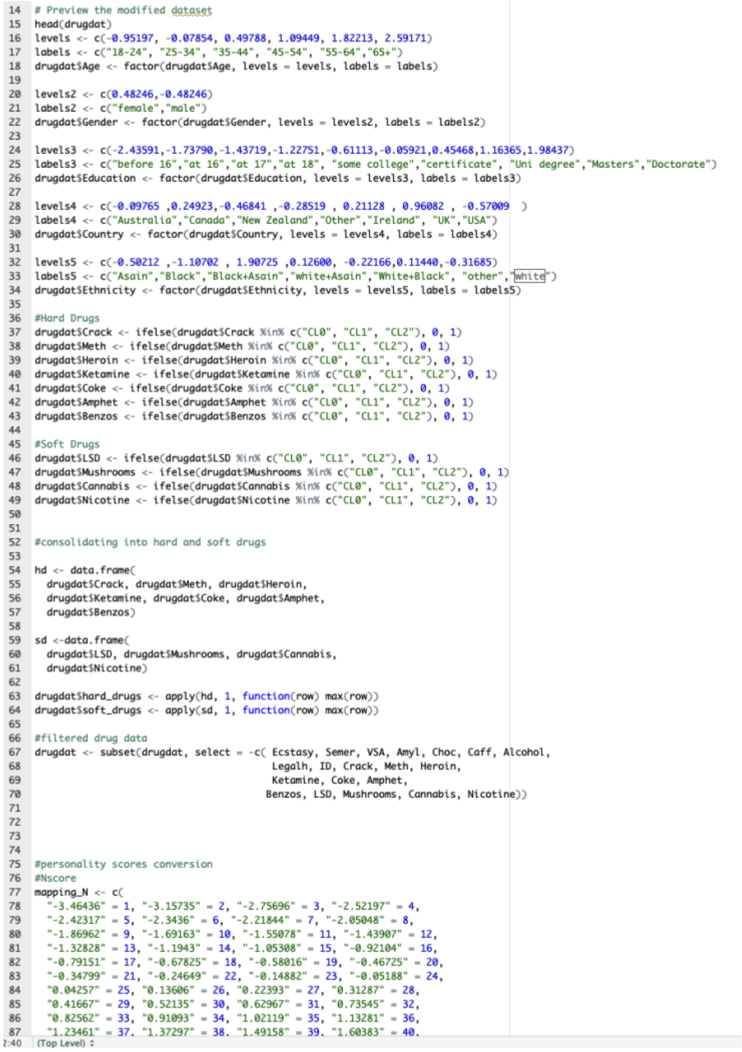
Understanding drug usage patterns is very important for informing public health initiatives and shaping policy decisions. Through logistic regression modeling, we achieved valuable insights into the factors influencing substance use behaviors and their potential correlations with educational attainment levels. In this analysis, we were able to determine that openness, education, and age turned out to be the significant factors in the soft drug model, younger adults in the 18-24 age group with lower levels of education appeared to be more open to using soft drugs. This trend reveals how peer pressure, and other behavioral tendencies can influence overall how well individuals perform in their academics.

Our investigation into the interplay between personality traits and drug use showed crucial findings. For instance, traits like Neuroticism and the use of soft s and were observed to play significant roles in hard drug experimentation and coping mechanisms. Given concerns for mental health, our study also aimed to investigate how drug consumption impacts academic performance in conjunction with individuals' personality traits and educational backgrounds. We tackled some issues with health and drug use, with a focus on how it influences the academic environment. Where schools can be informed to act as a supportive foundation and assist students in recovery from drug abuse in turn increases the levels of education students can succeed at.

Appendix

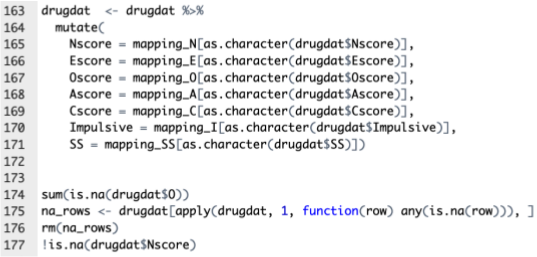
A. Mexwell. (n.d.). Drug consumption classification. Kaggle. Retrieved April 2nd, 2024, from https://www.kaggle.com/datasets/mexwell/drug-consumption-classification

B. Cleaning Data



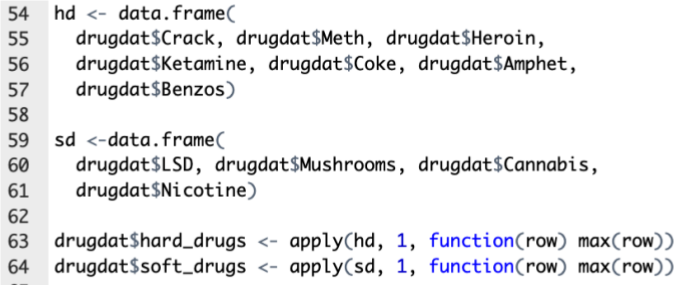
```
# Preview the modified dataset
head(drugdat)
levels <- c(-0.95197, -0.07854, 0.49788, 1.09449, 1.82213, 2.59171)
labels <- c("18-24", "25-34", "35-44", "45-54", "55-64","65+")
drugdat$Age <- factor(drugdat$Age, levels = levels, labels = labels)

levels2 <- c(0.48246,-0.48246)
labels2 <- c("female","male")
drugdat$Gender <- factor(drugdat$Gender, levels = levels2, labels = labels2)

levels3 <- c(-2.43591,-1.73790,-1.43719,-1.22751,-0.61113,-0.05921,0.45468,1.16365,1.98437)
labels3 <- c("before 16","at 16","at 17","at 18", "some college","certificate", "Uni degree","Masters","Doctorate")
drugdat$Education <- factor(drugdat$Education, levels = levels3, labels = labels3)

levels4 <- c(-0.09765 ,0.24923,-0.46841 ,-0.28519 , 0.21128 , 0.96082 , -0.57009  )
labels4 <- c("Australia","Canada","New Zealand","Other","Ireland", "UK","USA")
drugdat$Country <- factor(drugdat$Country, levels = levels4, labels = labels4)

levels5 <- c(-0.50212 ,-1.10702 , 1.90725 ,0.12600, -0.22166,0.11440,-0.31685)
labels5 <- c("Asain","Black","Black+Asain","white+Asain","White+Black", "other","white")
drugdat$Ethnicity <- factor(drugdat$Ethnicity, levels = levels5, labels = labels5)

#Hard Drugs
drugdat$Crack <- ifelse(drugdat$Crack %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Meth <- ifelse(drugdat$Meth %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Heroin <- ifelse(drugdat$Heroin %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Ketamine <- ifelse(drugdat$Ketamine %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Coke <- ifelse(drugdat$Coke %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Amphet <- ifelse(drugdat$Amphet %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Benzos <- ifelse(drugdat$Benzos %in% c("CL0", "CL1", "CL2"), 0, 1)

#Soft Drugs
drugdat$LSD <- ifelse(drugdat$LSD %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Mushrooms <- ifelse(drugdat$Mushrooms %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Cannabis <- ifelse(drugdat$Cannabis %in% c("CL0", "CL1", "CL2"), 0, 1)
drugdat$Nicotine <- ifelse(drugdat$Nicotine %in% c("CL0", "CL1", "CL2"), 0, 1)


#consolidating into hard and soft drugs

hd <- data.frame(
  drugdat$Crack, drugdat$Meth, drugdat$Heroin,
  drugdat$Ketamine, drugdat$Coke, drugdat$Amphet,
  drugdat$Benzos)

sd <-data.frame(
  drugdat$LSD, drugdat$Mushrooms, drugdat$Cannabis,
  drugdat$Nicotine)

drugdat$hard_drugs <- apply(hd, 1, function(row) max(row))
drugdat$soft_drugs <- apply(sd, 1, function(row) max(row))

#filtered drug data
drugdat <- subset(drugdat, select = -c( Ecstasy, Semer, VSA, Amyl, Choc, Caff, Alcohol,
                                        Legalh, ID, Crack, Meth, Heroin,
                                        Ketamine, Coke, Amphet,
                                       Benzos, LSD, Mushrooms, Cannabis, Nicotine))




#personality scores conversion
#Nscore
mapping_N <- c(
  "-3.46436" = 1, "-3.15735" = 2, "-2.75696" = 3, "-2.52197" = 4,
  "-2.42317" = 5, "-2.3436" = 6, "-2.21844" = 7, "-2.05048" = 8,
  "-1.86962" = 9, "-1.69163" = 10, "-1.55078" = 11, "-1.43907" = 12,
  "-1.32828" = 13, "-1.1943" = 14, "-1.05308" = 15, "-0.92104" = 16,
  "-0.79151" = 17, "-0.67825" = 18, "-0.58016" = 19, "-0.46725" = 20,
  "-0.34799" = 21, "-0.24649" = 22, "-0.14882" = 23, "-0.05188" = 24,
  "0.04257" = 25, "0.13606" = 26, "0.22393" = 27, "0.31287" = 28,
  "0.41667" = 29, "0.52135" = 30, "0.62967" = 31, "0.73545" = 32,
  "0.82562" = 33, "0.91093" = 34, "1.02119" = 35, "1.13281" = 36,
  "1.23461" = 37, "1.37297" = 38, "1.49158" = 39, "1.60383" = 40,
```

B – 1. Mutation



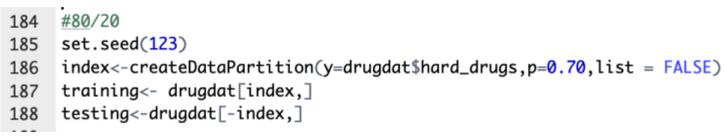
```
drugdat  <- drugdat %>%
  mutate(
    Nscore = mapping_N[as.character(drugdat$Nscore)],
    Escore = mapping_E[as.character(drugdat$Escore)],
    Oscore = mapping_O[as.character(drugdat$Oscore)],
    Ascore = mapping_A[as.character(drugdat$Ascore)],
    Cscore = mapping_C[as.character(drugdat$Cscore)],
    Impulsive = mapping_I[as.character(drugdat$Impulsive)],
    SS = mapping_SS[as.character(drugdat$SS)])


sum(is.na(drugdat$O))
na_rows <- drugdat[apply(drugdat, 1, function(row) any(is.na(row))), ]
rm(na_rows)
!is.na(drugdat$Nscore)
```

C. Consolidation between Hard and Soft Drugs



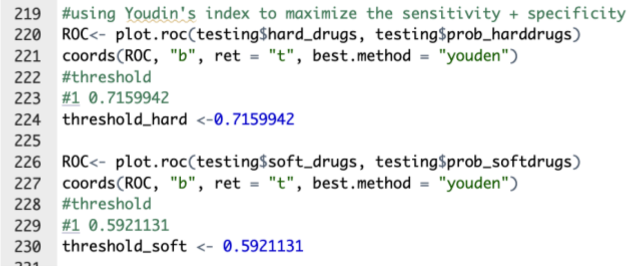
```
hd <- data.frame(
  drugdat$Crack, drugdat$Meth, drugdat$Heroin,
  drugdat$Ketamine, drugdat$Coke, drugdat$Amphet,
  drugdat$Benzos)

sd <-data.frame(
  drugdat$LSD, drugdat$Mushrooms, drugdat$Cannabis,
  drugdat$Nicotine)

drugdat$hard_drugs <- apply(hd, 1, function(row) max(row))
drugdat$soft_drugs <- apply(sd, 1, function(row) max(row))
```

D. Partitioning

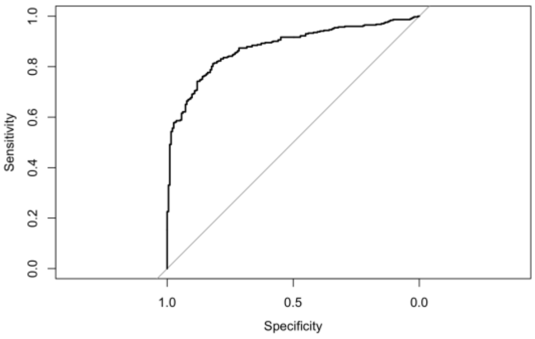


```
#80/20
set.seed(123)
index<-createDataPartition(y=drugdat$hard_drugs,p=0.70,list = FALSE)
training<- drugdat[index,]
testing<-drugdat[-index,]
```

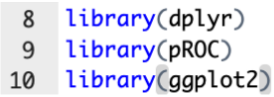
E. Youdin’s Index



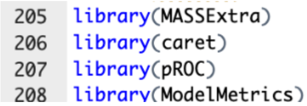
```
#using Youdin's index to maximize the sensitivity + specificity
ROC<- plot.roc(testing$hard_drugs, testing$prob_harddrugs)
coords(ROC, "b", ret = "t", best.method = "youden")
#threshold
#1 0.7159942
threshold_hard <-0.7159942

ROC<- plot.roc(testing$soft_drugs, testing$prob_softdrugs)
coords(ROC, "b", ret = "t", best.method = "youden")
#threshold
#1 0.5921131
threshold_soft <- 0.5921131
```



F. Packages

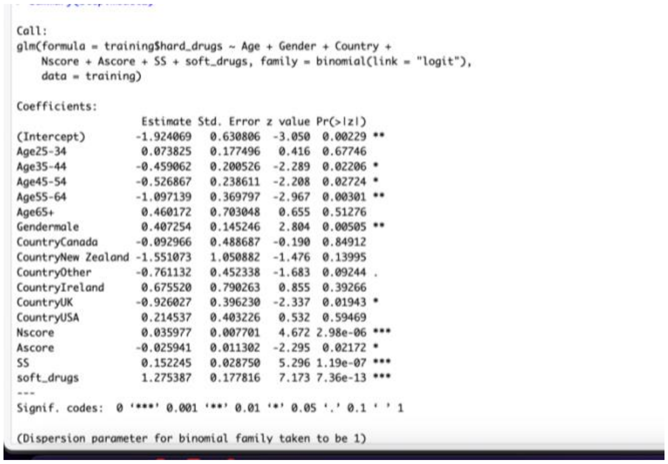


```
library(dplyr)
library(pROC)
library(ggplot2)
```



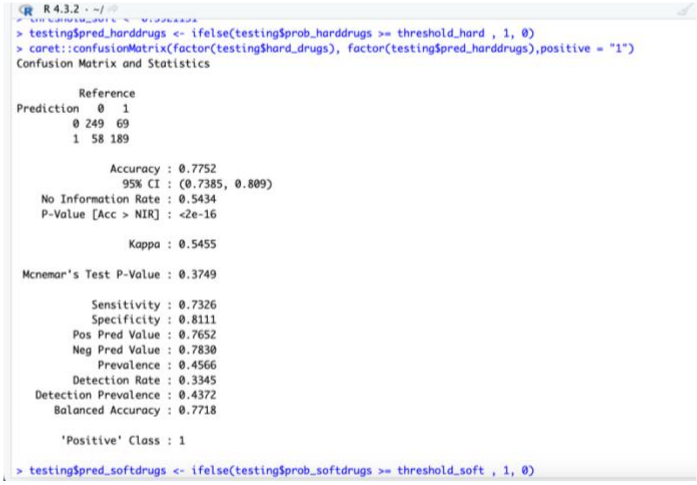
```
library(MASSExtra)
library(caret)
library(pROC)
library(ModelMetrics)
```

H . Co-efficient’s



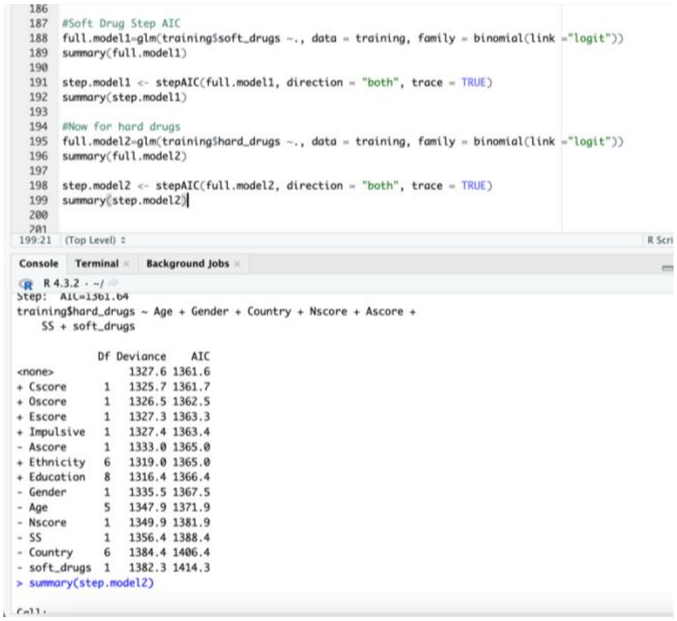
```
Call:
glm(formula = training$hard_drugs ~ Age + Gender + Country +
    Nscore + Ascore + SS + soft_drugs, family = binomial(link = "logit"),
    data = training)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.924069   0.630806  -3.050  0.00229 **
Age25-34            0.073825   0.177496   0.416  0.67746
Age35-44           -0.459062   0.200526  -2.289  0.02206 *
Age45-54           -0.526867   0.238611  -2.208  0.02724 *
Age55-64           -1.097139   0.369797  -2.967  0.00301 **
Age65+              0.460172   0.703048   0.655  0.51276
Gendermale          0.407254   0.145246   2.804  0.00505 **
CountryCanada      -0.092966   0.488687  -0.190  0.84912
CountryNew Zealand -1.551073   1.050882  -1.476  0.13995
CountryOther       -0.761132   0.452338  -1.683  0.09244 .
CountryIreland      0.675520   0.790263   0.855  0.39266
CountryUK          -0.926027   0.396230  -2.337  0.01943 *
CountryUSA          0.214537   0.403226   0.532  0.59469
Nscore              0.035977   0.007701   4.672 2.98e-06 ***
Ascore             -0.025941   0.011302  -2.295  0.02172 *
SS                  0.152245   0.028750   5.296 1.19e-07 ***
soft_drugs          1.275387   0.177816   7.173 7.36e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

I. Confusion Matrix, Hard Drug



```
> testing$pred_harddrugs <- ifelse(testing$prob_harddrugs >= threshold_hard , 1, 0)
> caret::confusionMatrix(factor(testing$hard_drugs), factor(testing$pred_harddrugs),positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 249  69
         1  58 189

               Accuracy : 0.7752
                 95% CI : (0.7385, 0.809)
    No Information Rate : 0.5434
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5455

 Mcnemar's Test P-Value : 0.3749

            Sensitivity : 0.7326
            Specificity : 0.8111
         Pos Pred Value : 0.7652
         Neg Pred Value : 0.7830
             Prevalence : 0.4566
         Detection Rate : 0.3345
   Detection Prevalence : 0.4372
      Balanced Accuracy : 0.7718

       'Positive' Class : 1

> testing$pred_softdrugs <- ifelse(testing$prob_softdrugs >= threshold_soft , 1, 0)
```

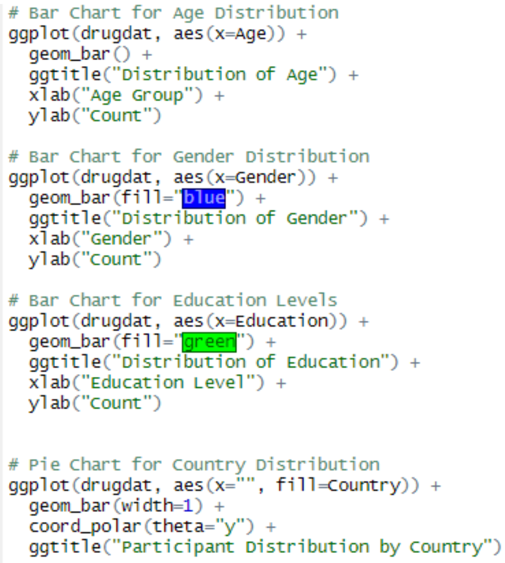
J : Step AIC



```
#Soft Drug Step AIC
full.model1=glm(training$soft_drugs ~., data = training, family = binomial(link ="logit"))
summary(full.model1)

step.model1 <- stepAIC(full.model1, direction = "both", trace = TRUE)
summary(step.model1)

#Now for hard drugs
full.model2=glm(training$hard_drugs ~., data = training, family = binomial(link ="logit"))
summary(full.model2)

step.model2 <- stepAIC(full.model2, direction = "both", trace = TRUE)
summary(step.model2)
```

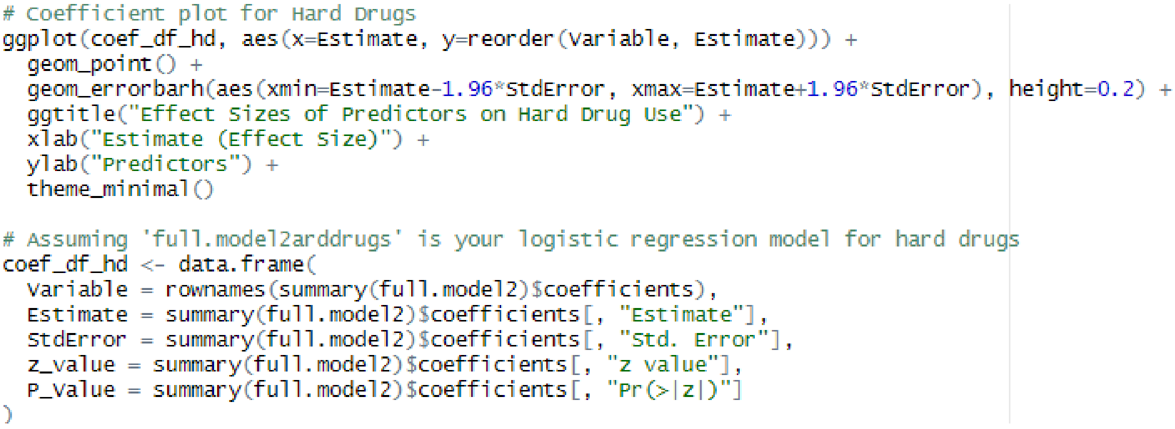
```
Step:  AIC=1361.64
training$hard_drugs ~ Age + Gender + Country + Nscore + Ascore +
    SS + soft_drugs

            Df Deviance    AIC
<none>           1327.6 1361.6
+ Cscore     1   1325.7 1361.7
+ Oscore     1   1326.5 1362.5
+ Escore     1   1327.3 1363.3
+ Impulsive  1   1327.4 1363.4
- Ascore     1   1333.0 1365.0
+ Ethnicity  6   1319.0 1365.0
+ Education  8   1316.4 1366.4
- Gender     1   1335.5 1367.5
- Age        5   1347.9 1371.9
- Nscore     1   1349.9 1381.9
- SS         1   1356.4 1388.4
- Country    6   1384.4 1406.4
- soft_drugs 1   1382.3 1414.3
> summary(step.model2)

Call:
```

K Pie charts Bar charts



```
# Bar Chart for Age Distribution
ggplot(drugdat, aes(x=Age)) +
  geom_bar() +
  ggtitle("Distribution of Age") +
  xlab("Age Group") +
  ylab("Count")

# Bar Chart for Gender Distribution
ggplot(drugdat, aes(x=Gender)) +
  geom_bar(fill="blue") +
  ggtitle("Distribution of Gender") +
  xlab("Gender") +
  ylab("Count")

# Bar Chart for Education Levels
ggplot(drugdat, aes(x=Education)) +
  geom_bar(fill="green") +
  ggtitle("Distribution of Education") +
  xlab("Education Level") +
  ylab("Count")


# Pie Chart for Country Distribution
ggplot(drugdat, aes(x="", fill=Country)) +
  geom_bar(width=1) +
  coord_polar(theta="y") +
  ggtitle("Participant Distribution by Country")
```

L Coefficient plot



```
# Coefficient plot for Hard Drugs
ggplot(coef_df_hd, aes(x=Estimate, y=reorder(Variable, Estimate))) +
  geom_point() +
  geom_errorbarh(aes(xmin=Estimate-1.96*StdError, xmax=Estimate+1.96*StdError), height=0.2) +
  ggtitle("Effect Sizes of Predictors on Hard Drug Use") +
  xlab("Estimate (Effect Size)") +
  ylab("Predictors") +
  theme_minimal()

# Assuming 'full.model2arddrugs' is your logistic regression model for hard drugs
coef_df_hd <- data.frame(
  Variable = rownames(summary(full.model2)$coefficients),
  Estimate = summary(full.model2)$coefficients[, "Estimate"],
  StdError = summary(full.model2)$coefficients[, "Std. Error"],
  z_value = summary(full.model2)$coefficients[, "z value"],
  P_Value = summary(full.model2)$coefficients[, "Pr(>|z|)"]
)
```

THE END