# Predicting Drug Consumption By Personality and Demographics

HarvardX PH125.9x Data Science Capstone

Eric Richardson

6-1-2020

TABLE OF CONTENTS

## Summary

## Introduction

Drug use is a behavior that constitutes an important factor linked to poor health, including early mortality and which presents significant adverse consequences for society with respect to criminality and families staying together. Early detection of an individual's predisposition to drug consumption offers healthcare professionals an opportunity to short-circuit the onset of addiction.

The present study is based on a dataset that includes demographic and psychological information related to the consumption of 18 legal and illegal drugs by 1885 participants. This study will focus on the data analysis and modeling on the use of marijuana.

## Project Goal

The goal of this project is to assess whether an individual's consumption of marijuana can be predicted by using a combination of demographic and personality data.
We will build and assess the effectiveness of six machine learning modules and review the results obtained from each of the data results.

## Dataset Description

The dataset used was found on the UCI machine learning repository. It is based the research paper by E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv, 2015. The data was collected from 1885 English-speaking participants over 18 years of age between March 2011 and March 2012.

The dataset includes answers to questions related to the use of alcohol, caffeine, benzodiazepines, amphetamines, amyl nitrite, LSD, marijuana, cocaine, crack, ecstasy, heroin, ketamine, legal highs, methadone, chocolate, magic mushrooms, nicotine and volatile substance abuse (VSA)) and one fictitious drug (Semeron). For the purposes of this project I will be concentrating on the use of marijuana consumption. Many feel this is a drug that should be made available to the public because of the health benefits. The data in this project will be divided into two groups; never used and used.

The data will be divided in two groups of centered and pre-normalized predictors.

1. Five demographic predictors: Age, Gender, Level of education, Ethnicity, and Country of origin.
2. The results from seven scored tests administered to assess personality, specifically:

Neuroticism (a long-term tendency to experience negative emotions such as nervousness, tension, anxiety, and depression).
- Extraversion (manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics)
- Openness to experience (a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests)
- Agreeableness (a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion, and cooperativeness)
- Conscientiousness (a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient)
- Impulsiveness
- Thrill-seeking.
- The dataset consists of one Class (Marijuana consumption) and twelve predictors (five demographic and seven personality-related).

## Process Steps

Create a training subset (80% of data) from the dataset for the purpose of training the model, and use the remaining 20% of the data as a control group for comparison.

The analysis consists of two sections:
1. Perform minor data engineering (A), explore, bin, and analyze the dataset (B).
2. Create the modeling (C).
• After a 3-step preprocessing consisting of examining correlation among predictors, seeking low-variance
factors and applying a Recursive Feature Elimination algorithm to seek and potentially discard predictors that do not contribute significantly to the outcome, build models based on the following six popular
machine learning methods:

- Generalized linear model (glm)
- Generalized linear model with penalized maximum likelihood (GLMnet)
- Neural network (nnet)
- Stochastic gradient boosting (gbm)
- Decision tree (rpart)
- Random forest (rf)

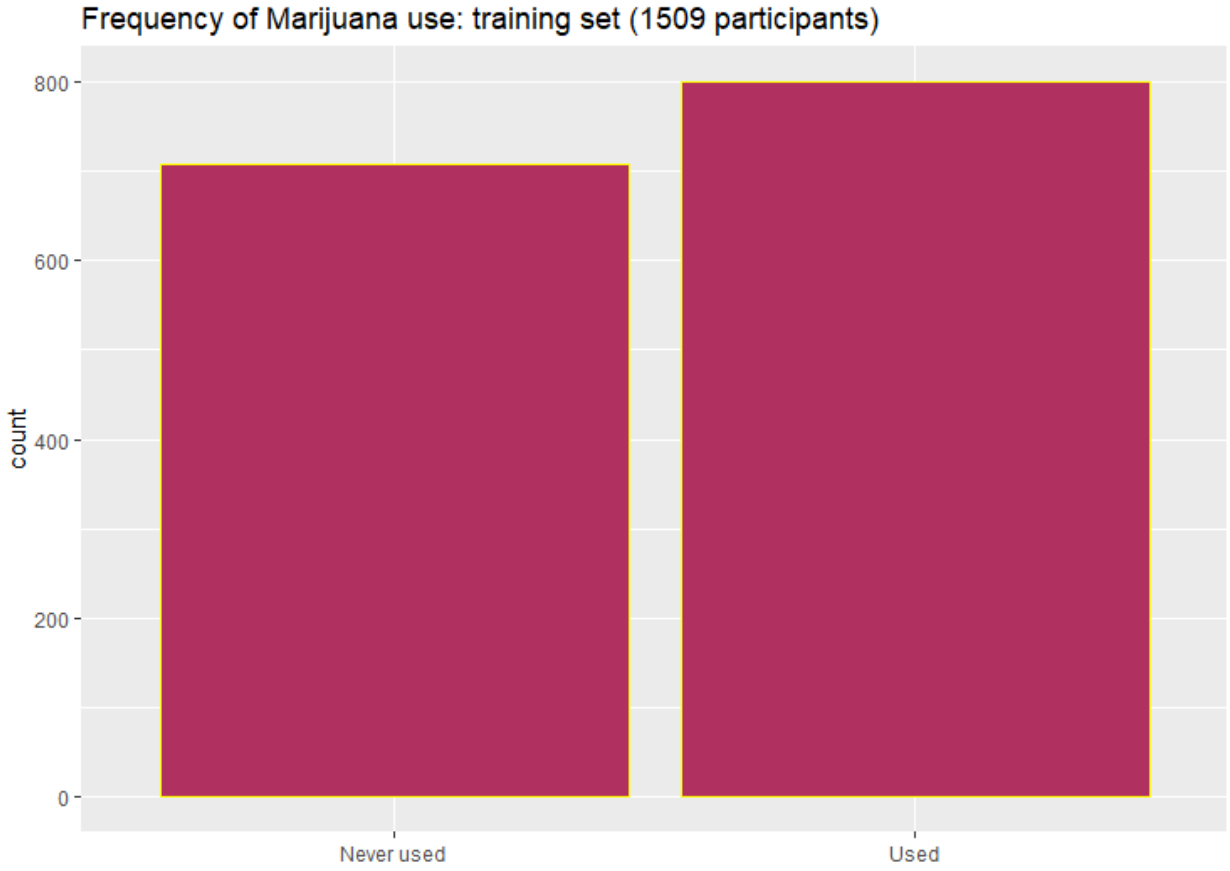
The modeling outputs will be compared to one another to determine if the data is accurate. Keeping in mind that the data is only a small set of test subjects.

## Analysis

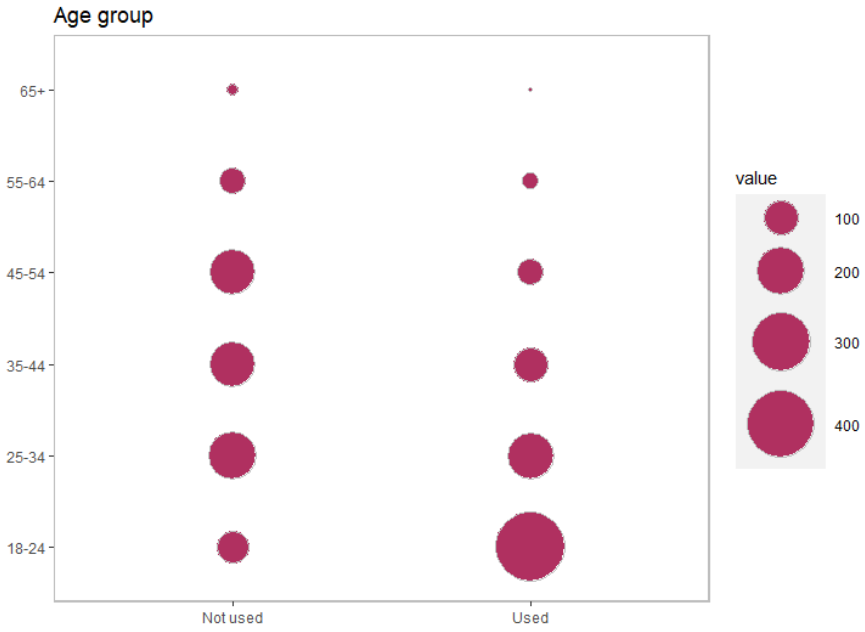
## R Data Setup

- The data in the dataset was normalized.
- There was no 0 data in the dataset so the use of NAs was not needed
- Two classes were created. Used and Never Used
- 80% of the data was trained leaving 20% as the control group

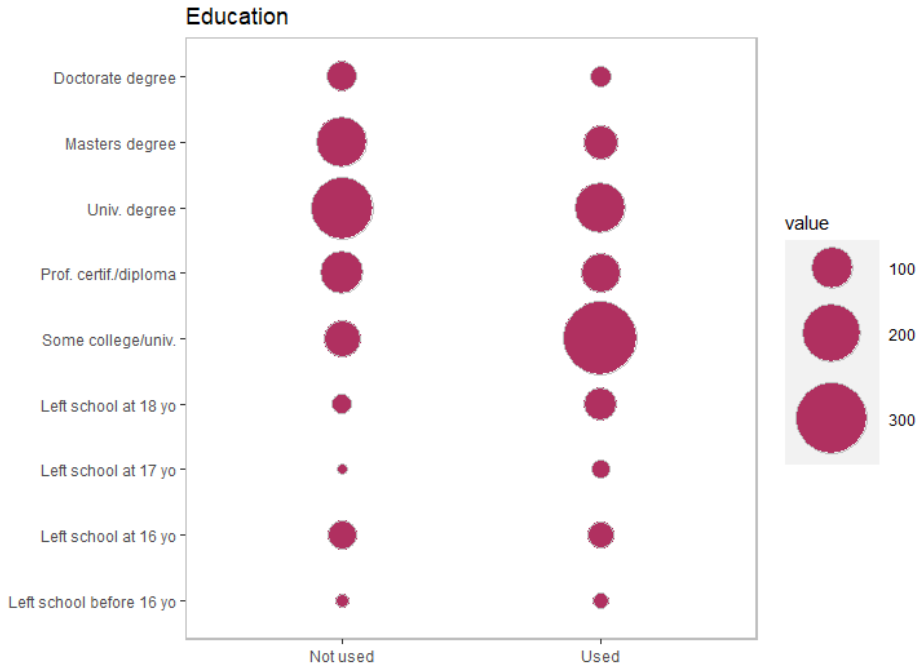
## Data Review (Plot findings)

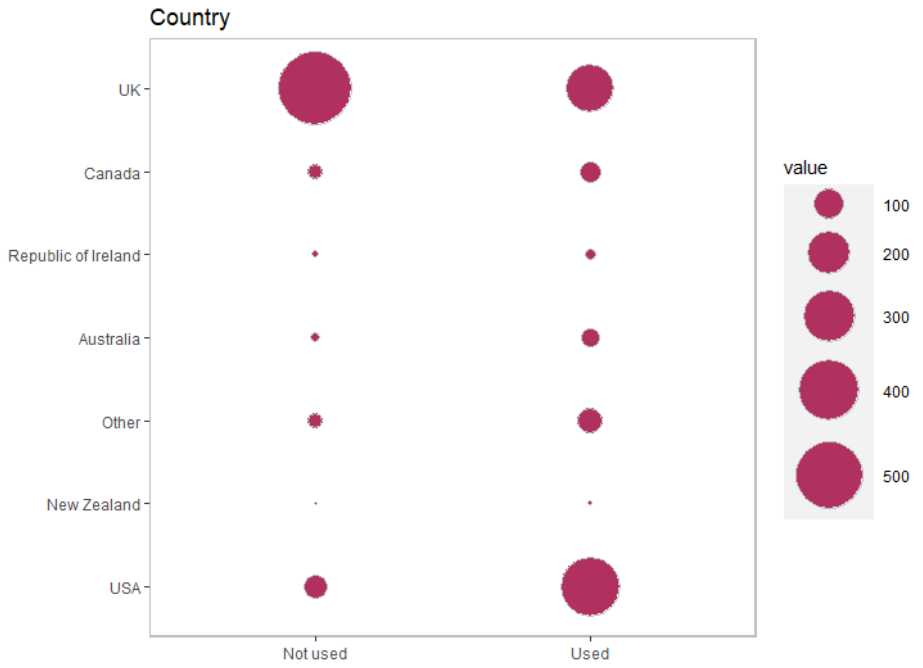


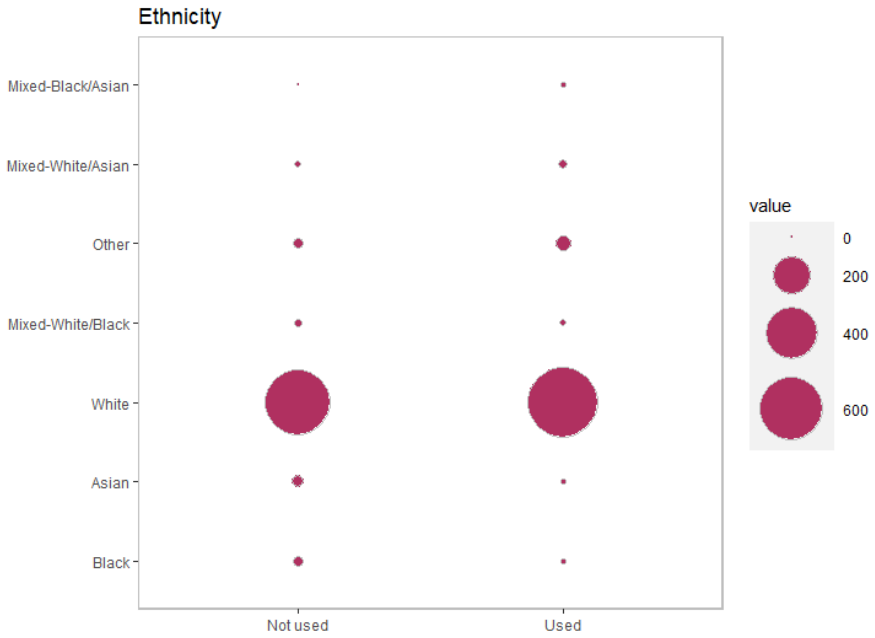

The training set of 1509 participants consists of 800 participants (53.0%) having used cannabis and 709 who never have (47.0%), for a user-to-non-user ratio of 1:1.1.


Age group
65+
55-64
45-54
35-44
25-34
18-24
Not used
Used
value
100
200
300
400


Gender
female
male
Not used
Used
value
250
300
350
400
450

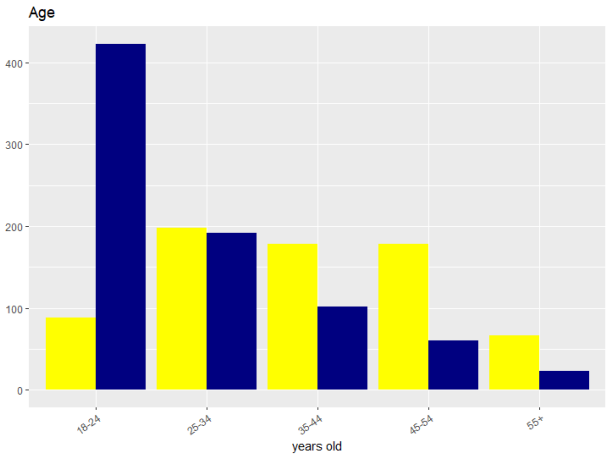

Education
Doctorate degree
Masters degree
Univ. degree
Prof. certif./diploma
Some college/univ.
Left school at 18 yo
Left school at 17 yo
Left school at 16 yo
Left school before 16 yo
Not used
Used
value
100
200
300

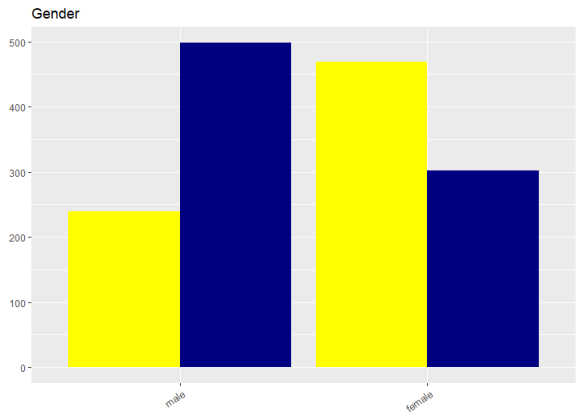

Country
UK
Canada
Republic of Ireland
Australia
Other
New Zealand
USA
Not used
Used
value
100
200
300
400
500

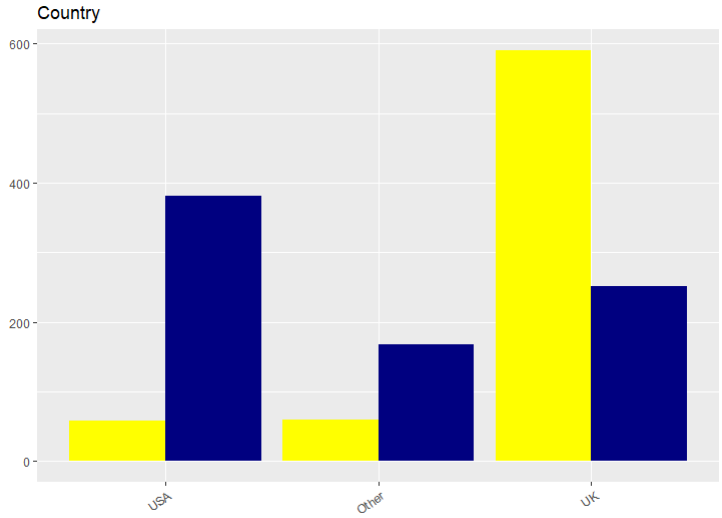


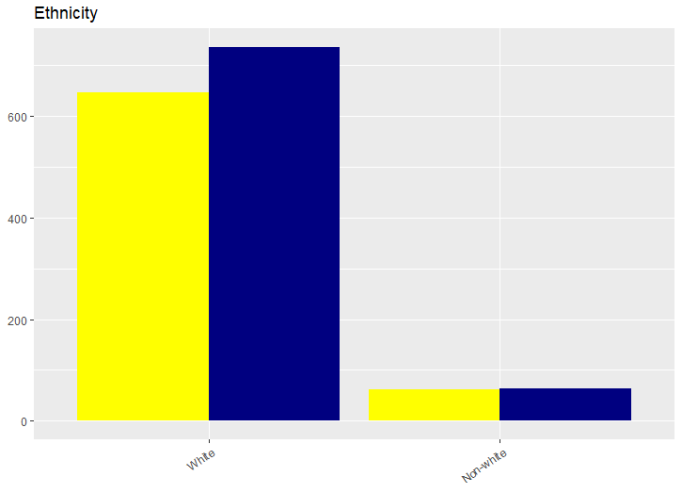

The data is mostly contains educated white British and American males and females.

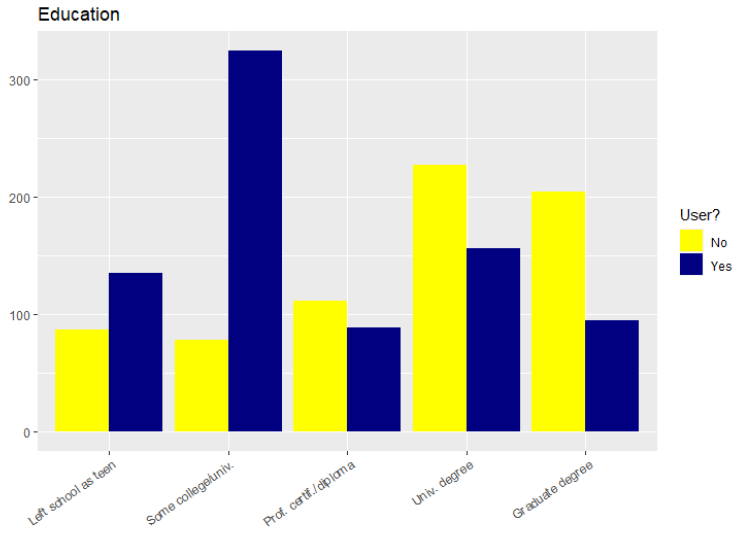
## Demographics data analysis



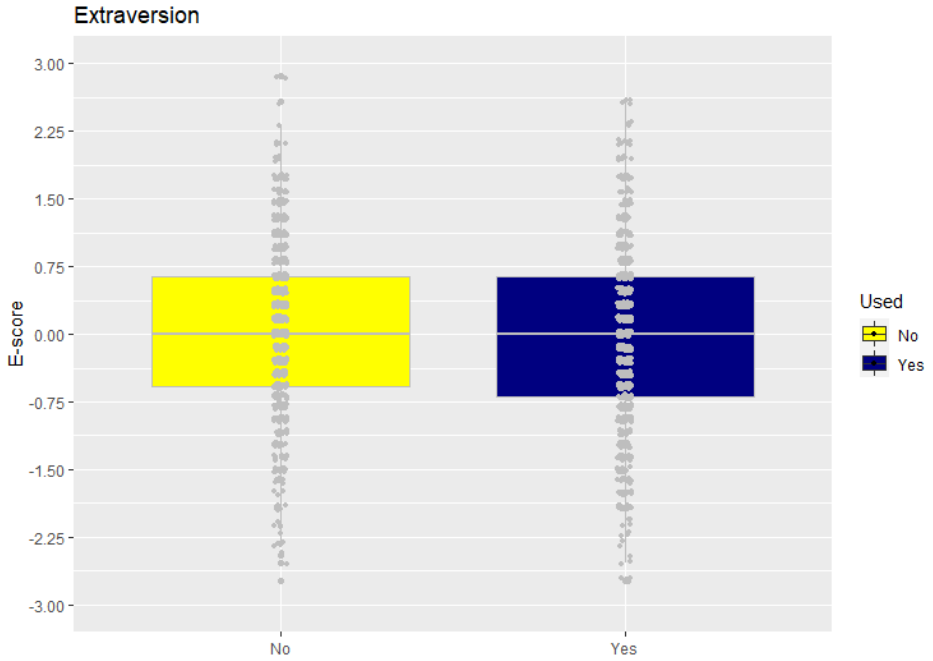


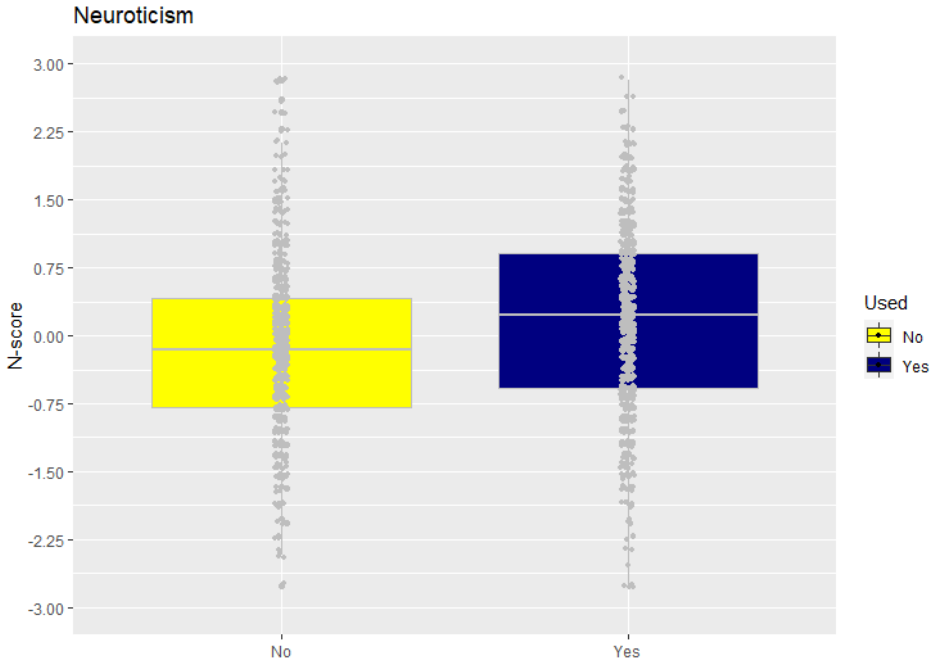



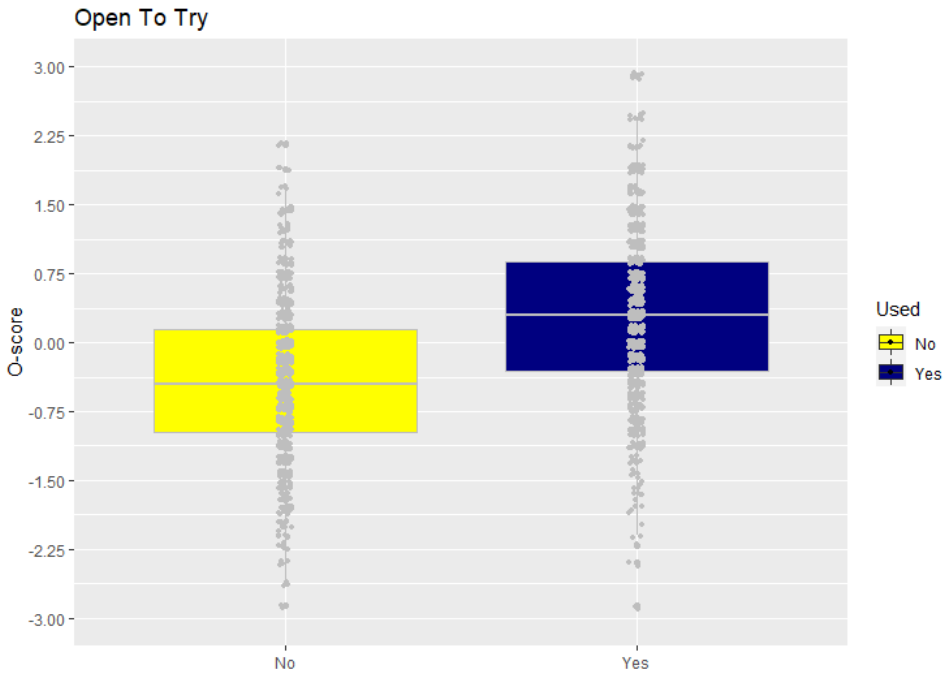



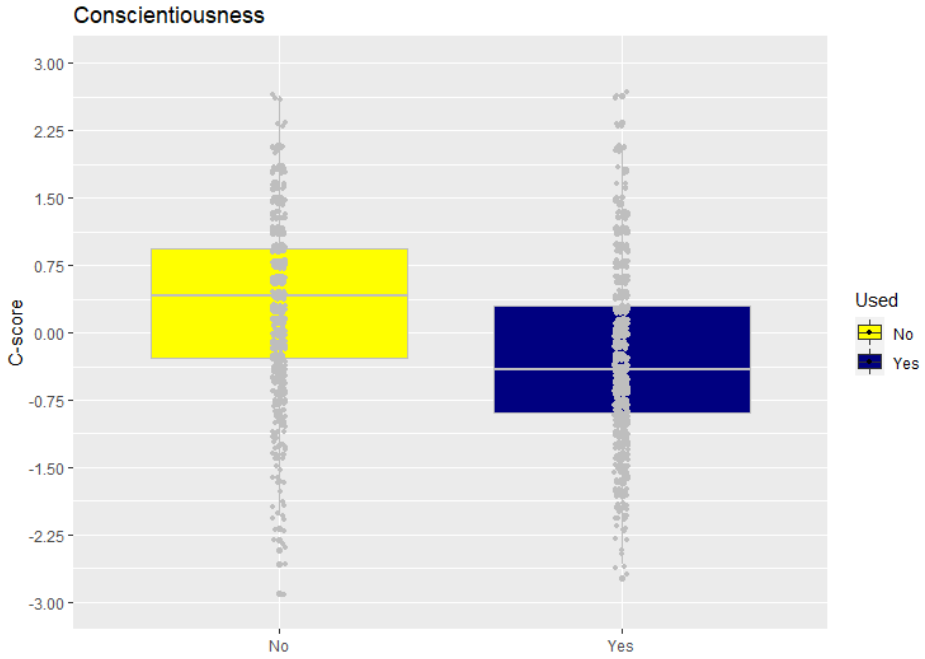

From the graphs above it is logical to determine that users make up a much larger group than non-users. This is especially true for US men between the ages of 18 and 24 that did not obtain a college degree.

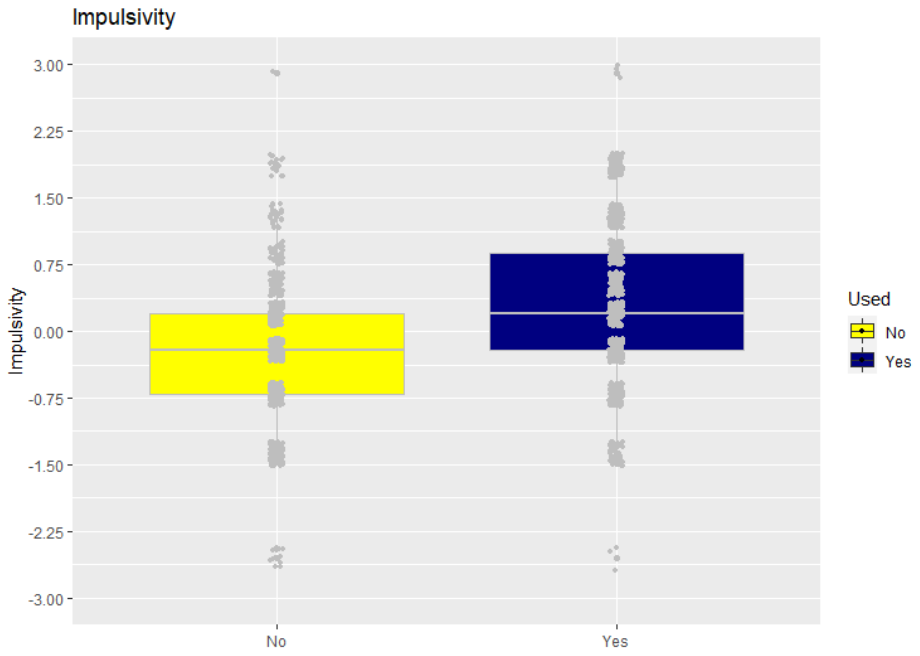
## Personality Results

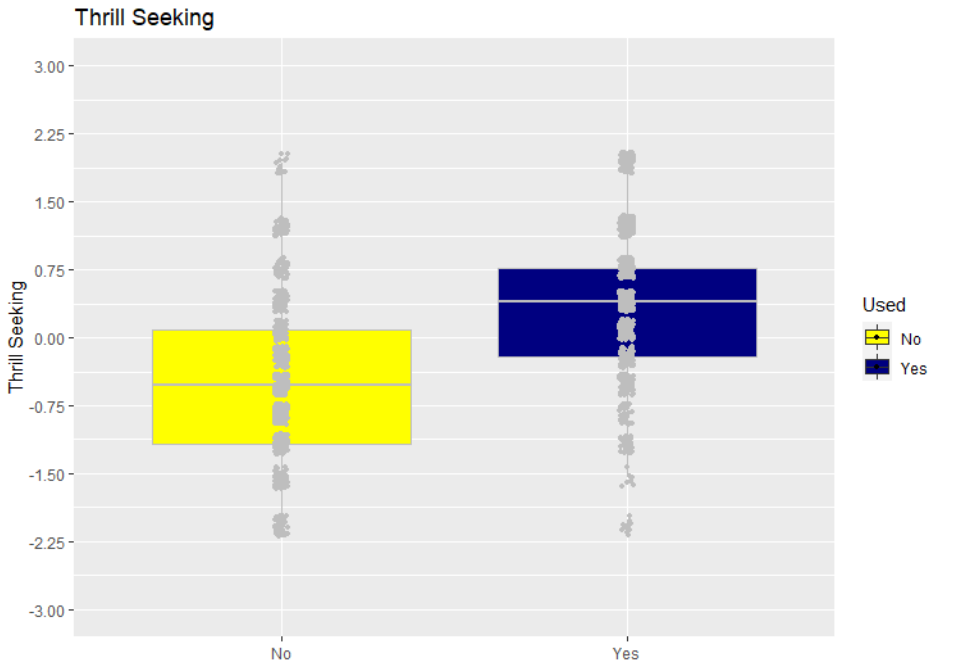


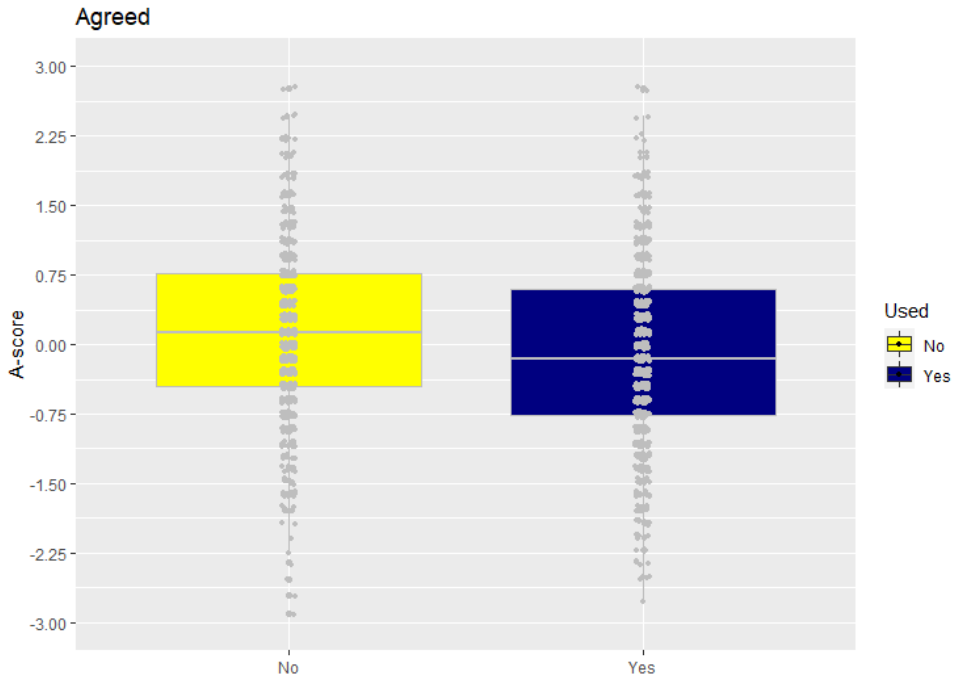




Open To Try
O-score
3.00
2.25
1.50
0.75
0.00
-0.75
-1.50
-2.25
-3.00
No
Yes
Used
No
Yes


Conscientiousness
C-score
3.00
2.25
1.50
0.75
0.00
-0.75
-1.50
-2.25
-3.00
No
Yes
Used
No
Yes


Impulsivity
Impulsivity
3.00
2.25
1.50
0.75
0.00
-0.75
-1.50
-2.25
-3.00
No
Yes
Used
No
Yes



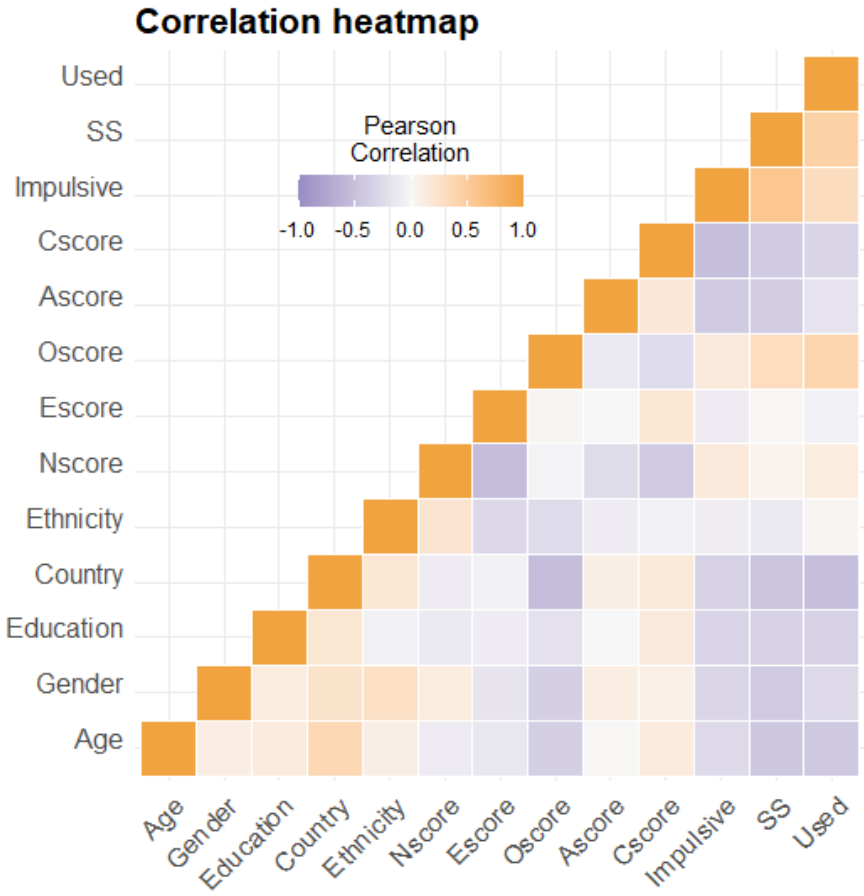

The data shows that individuals with certain personality traits may be more open to smoke marijuana than others. But with this small number of volunteers it would be hard to make a definite conclusion.

## Modeling

The goal was to create a model that would show an improvement on the ratio of users against the population.

### Pre-processing

The data was analyzed with most contributing cells to the total Chi-square score redundancies among the 12 predictors and with the Used class by examining the Pearson _2 residuals.




The maximum correlation of 0.57 was not enough to correlate amongst one another. This still made for some interesting results.

## Results

None of the modeling approaches used provided an improvement over generalized linear regression (84.8% accuracy). Besides offering the highest accuracy, the importance plot related to the glm model is by and large consistent with results from the data exploration:

- Country of origin, Age, Openness to experiment and Education are the factors contributing most to the accuracy.
- Education and Thrill-seeking have a significant importance
- Ethnicity, N-scores and E-scores and do not contribute much to the metric.
- The low impact of Impulsivity is at variance with the observations from data exploration.

The confusion matrix for the GLM model has a sensitivity of (0.82) and specificity of (0.87):

```
## Confusion Matrix and Statistics
##
## Reference
## Prediction 0 1
## 0 146 26
## 1 31 173
##
## Accuracy : 0.8645
## 95% CI : (0.8081, 0.8831)
## No Information Rate : 0.5297
## P-Value [Acc > NIR] : <2e-16
##
## Kappa : 0.6955
##
## Mcnemar's Test P-Value : 0.5298
##
## Sensitivity : 0.8249
## Specificity : 0.8693
## Pos Pred Value : 0.8488
## Neg Pred Value : 0.8480
## Prevalence : 0.4707
## Detection Rate : 0.3883
## Detection Prevalence : 0.4574
## Balanced Accuracy : 0.8471
##
## 'Positive' Class : .7
##
```

The next best model (84.0% accuracy) is obtained with a neural network. It suggests that demographic factors have more impact on marijuana use than personality (top three importance factors: country of origin, ethnicity, and age). Given the statistical

similarity between the two ethnic groups discussed in the personality analysis section, this model is somewhat limited in the results to make a definitive conclusion.

The confusion matrix for the neural network model is :
## Confusion Matrix and Statistics
##
## Reference
## Prediction 0 1
## 0 143 26
## 1 34 173
##
## Accuracy : 0.8402
21
## 95% CI : (0.7994, 0.876)
## No Information Rate : 0.5293
## P-Value [Acc > NIR] : <2e-16
##
## Kappa : 0.679
##
## Mcnemar's Test P-Value : 0.3662
##
## Sensitivity : 0.8079
## Specificity : 0.8693
## Pos Pred Value : 0.8462
## Neg Pred Value : 0.8357
## Prevalence : 0.4707
## Detection Rate : 0.3803
## Detection Prevalence : 0..4498
## Balanced Accuracy : 0.8386
##
## 'Positive' Class : 0
##

With 0 (Non-user) as the ‘positive’ class, the 3-point decrease in sensitivity (0.81) indicates a drop in this model’s ability to predict Non-users.
While less accurate, the random forest model (83.5% accuracy) gives a preponderant importance to country of origin, sensation-seeking trait, age and openness to experiment). It also gives no importance to ethnicity and E-score and less than expected to gender.

The confusion matrix for the random forest model is :
## Confusion Matrix and Statistics
##
## Reference
## Prediction 0 1
## 0 143 28

```
## 1 34 171
##
## Accuracy : 0.8351
## 95% CI : (0.7937, 0.8712)
## No Information Rate : 0.5293
## P-Value [Acc > NIR] : <2e-16
##
## Kappa : 0.6685
##
## Mcnemar's Test P-Value : 0.5254
##
## Sensitivity : 0.8079
## Specificity : 0.8593
## Pos Pred Value : 0.8363
## Neg Pred Value : 0.8341
## Prevalence : 0.4707
## Detection Rate : 0.3803
## Detection Prevalence : 0.4548
## Balanced Accuracy : 0.8336
##
## 'Positive' Class : 0
##
```

With RF also, the decrease in sensitivity (0.81) indicates a drop in this model's ability to predict Non-users. For this dataset, the optimized GLM offers the weakest modeling technique (55.9% accuracy), performing less well than the naive approach (53.0% accuracy).

## Conclusion

That data returned some interesting results. I did struggle with trying come to some conclusions from my results. It would seem that demographics plays more of a role in the use of marijuana than any other factors. The initial though was age and lack of completing higher education would be key factors but the results told a different story.