Christina De Cesaris University of California, Davis
STA 223 Project 2

Analysis of Characteristics Influencing Nicotine Usage

Abstract

Nicotine is a highly addictive substance known to cause cancer and other health problems *(Saha et al, 2007)*. To better understand the personality and demographic characteristics which influence nicotine usage, a survey dataset containing responses from 1885 participants regarding their drug usage history and personality scores was processed and fit to a baseline-odds logistic regression model with three outcomes: Never Used, Recent User, and Past User. Model inference revealed that men, people with high Openness Scores, lower Conscientiousness Scores, and without university degrees were more likely to be Recent or Past Users. No lack of fit was detected in the model.

Background

Nicotine is one of the most addictive substances legal to purchase on the market. Smoking is a common form of nicotine consumption, and while many smokers desire to quit the habit, only about 3 percent are successful each year ***(****Benowitz, 2010)*. The adverse effects of smoking are numerous, ranging from cancer to cardiovascular disease *(Saha et al, 2007)*. Despite public health efforts to reduce public nicotine usage by taxing cigarettes, nicotine usage remains prevalent though the purchase of e-cigarettes and vaporizers among younger populations. This has led to an increased interest in the demographic and personal characteristics which influence nicotine usage *(Saha et al, 2007)*. Using a dataset containing the demographics, personality assessment, and drug usage history of 1885 participants, this project seeks to identify the characteristics which influence nicotine usage *(Fehrman et al, 2017)*. Personality characteristics were quantified using the Big Five personality traits model and alternative models relying on impulsivity and sensation-seeking behaviors from phycological research *(Fehrman et al, 2017)*.

The drug consumption dataset was collected through an anonymous online survey and contains 5 categorical variables, 7 numerical score variables, and a series of categorical responses indicating the participant's drug usage history for 18 legal and illegal substances. In this study, Nicotine usage is the primary outcome of interest. Table 1A, which displays all the variables and associated descriptions used in this project, does not show the other 17 drugs in the survey *(Fehrman et al, 2017)*.

Methods

Data Preprocessing

The raw data was presented as numeric with all possible values of both categorical and numeric predictors represented by a coded value. The response was coded as character values from CL0-CL6 (Table 1). An excerpt of the raw data is shown in Table 2A and the processed data in Table 3A, and the associated keys for categorical and numeric variables are shown in Table A4, Table A5, and Table A6.

Usage Code	Usage Label	Simplified Labels
CL0	Never Used	Never Used
CL1	Used over a Decade Ago	Past User
CL2	Used in Last Decade	Past User
CL3	Used in Last Year	Past User
CL4	Used in Last Month	Recent User
CL5	Used in Last Week	Recent User
CL6	Used in Last Day	Recent User

Table 1: Response codes, keys, and simplified label

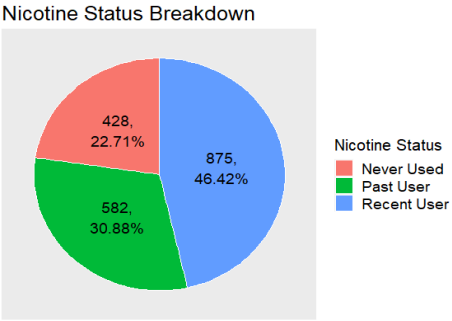

Figure 1: Percentage breakdown of response

The outcome variable was simplified to three levels: Never Used, Past User, and Recent User shown in Table 1. The simplified outcome was plotted by proportion to ensure there was no severe imbalance among the categories (Figure 1). Most study participants belonged to the Recent User category, followed by Past User, and lastly Never Used.

Numeric personality scores were plotted by nicotine usage status, and there appeared to be a notable difference in the Conscientiousness and Openness Scores between Recent Users and Never Users (Figure 2). In general, Never Users tended toward higher Conscientiousness Scores and lower Openness scores when compared to their counterparts. Boxplots for all numeric scores by outcome can be found in Figure A1 and Figure A2. Categorical variables were plotted by proportion by outcome (Figures A3 -A6). Plots for Ethnicity and Country revealed most participants, regardless of nicotine status, were white and from either the USA or UK. Recent Users were predominately between the ages of 18-24, and most of the survey participants were under 50 (Figure A5). Around 59% of Recent Users were male and 64% of never users were female (Figure 3), and the majority of Past and Never Users held university degrees (Figure A3).

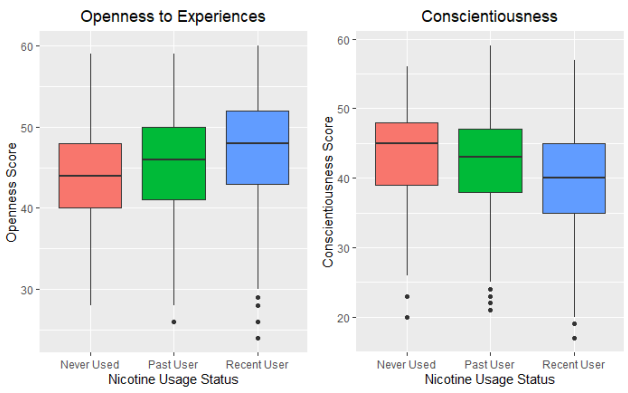

Figure 2: Openness and Conscientiousness scores by outcome

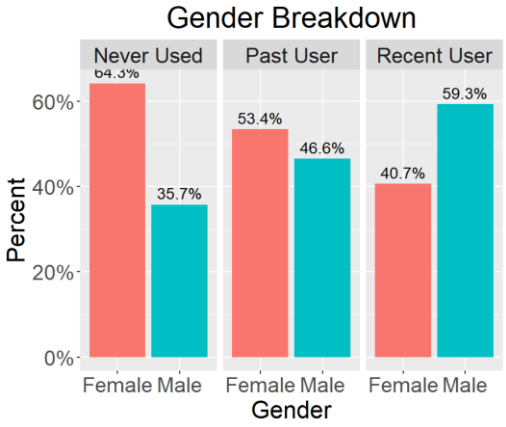

Figure 3: Percentage breakdown of response

Modeling Methods

With three potential categorical outcomes, a multinomial regression model is suitable for this analysis. The three categories of interest: Never Used, Past User, and Recent User are assumed to be unordered because there is no quantified level which differentiates the severity of nicotine usage within the categories. Therefore, a baseline odds model using Never Used as the base level against the Past User and Recent User outcomes in respective submodels was constructed *(Faraway, 2016).*

Baseline Odds Model

Assume the response $Y_i = (y_{i1}, y_{i2}, y_{i3})^T$ follows a multinomial distribution $Y_i \sim Multi(\pi_i, 1)$, with $\pi_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})^T$ response probabilities for the $j = 1 \dots m$ response categories where here $m = 3$. If we define the response category 1 as the baseline and construct a $m - 1 = 2$ baseline-category logistic regression models with logit link functions, we form the following two logistic repression submodels. $log\left(\frac{\pi_{i2}}{\pi_{i1}}\right) = x_i^T\beta_2\ ,\ log\left(\frac{\pi_{i3}}{\pi_{i1}}\right) = x_i^T\beta_3$

Where $x_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$ for p explanatory predictors and $\beta_j = (\beta_{0j}, \beta_{1j}, \beta_{2j} \dots . \beta_{pj})^T, j \neq 1(baseline)$ associated coefficients.

Recall that for each submodel $log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = logit(\pi_{ij}) = \eta_j$.

Model selection was performed by first fitting the full model with the X potential predictors. Backwards, stepwise selection was performed on the full model using BIC criteria to determine a model with suitable predictors for each submodel. The resulting model contained two categorical predictors and three numerical predictors. Consider $j = 1 = Never\ Used,\ j = 2 = Past\ User,$ $and\ j = 3 = Recent\ User$ outcomes.

Assessment of Fit

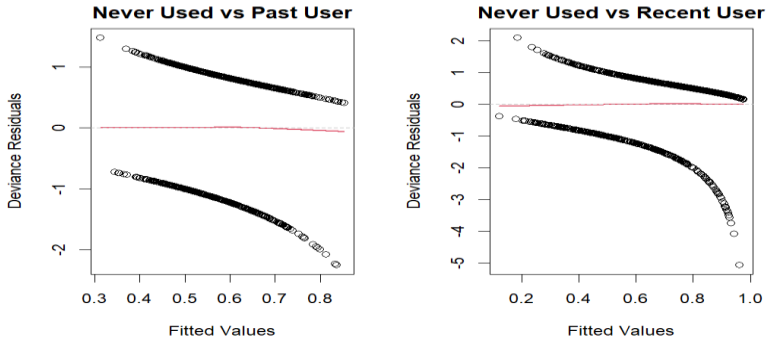

Figure 4: Diagnostic residuals vs. fitted plots for both submodels

There is no indication of a lack of fit from the deviance residuals vs fitted value plots for both models (Figure 4). Between both submodels, the Past User model could be considered a better fit.

Results

Final Model

$$\eta_2 = log\left(\frac{P(PastUser)}{P(NeverUsed)}\right) = \beta_{20} + \beta_{21}gender_{male} + \beta_{22}edu_{highschool} + \beta_{23}edu_{somecollege} + \beta_{24}edu_{somehighschool} + \beta_{25}edu_{university} + \beta_{26}openess + \beta_{27}conscientiousness + \beta_{28}impulsivity$$

$$\eta_3 = log\left(\frac{P(RecentUser)}{P(NeverUsed)}\right) = \beta_{30} + \beta_{31}gender_{male} + \beta_{32}edu_{highschool} + \beta_{33}edu_{somecollege} + \beta_{34}edu_{somehighschool} + \beta_{35}edu_{university} + \beta_{36}openess + \beta_{37}conscientiousness + \beta_{38}impulsivity$$

	Never Used = 0, Past User = 1				***Never Used = 0, Recent User = 1***			
	Estimate	***P Value***	***CI 5%***	***CI 95%***	***Estimate***	***P Value***	***CI 5%***	***CI 95%***
Intercept	*-0.905*	*-0.176*	*-2.004*	*0.195*	*-0.007*	*0.992*	*-1.064*	*1.050*
Male	*0.304*	*0.025 **	*0.080*	*0.527*	*0.558*	*2.142e-05 ****	*0.342*	*0.775*
High School	*0.925*	*0.019 **	*0.279*	*1.571*	*0.861*	*0.0224 **	*0.241*	*1.481*
Some College	*0.419*	*0.073*	*0.034*	*0.804*	*0.511*	*0.018 **	*0.155*	*0.866*
Some High School	*0.229*	*0.442*	*-0.261*	*0.718*	*0.542*	*0.045 **	*0.096*	*0.987*
University	*0.084*	*0.661*	*-0.231*	*0.399*	*-0.499*	*0.006 ***	*-0.800*	*-0.198*
Openness Score	*0.036*	*0.0005 ****	*0.019*	*0.053*	*0 .071*	*3.858e-12 ****	*0.054*	*0.087*
Conscientiousness Score	*-0.016*	*0.1262*	*-0.034*	*0.001*	*-0.047*	*3.108e-06 ****	*-0.064*	*-0.031*
Impulsivity Score	*1.499e-06*	*0.999*	*-0.002*	*0.002*	*-0.003*	*0.001 ***	*-0.004*	*-0.001*
Null deviance: 3980.072		***Residual Deviance: 3693.059***			***AIC: 3729.059***			
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*								

Table 2: Final model estimates, pvalues and associated 95% confidence intervals

Outlier Removal

Observations 809, 1103, and 1004 were identified as influential points with large cook distance (Figures A7 & A8). The points all belonged to Never Users with high school education levels and Conscientiousness scores below the average for all outcome groups (Table A7). These points were removed and the model refit. There was no significant improvement in the lack of fit or estimated values (Table A8).

Regression Effect

$$H_0: \beta_{j1} = \beta_{j2} = \cdots = \beta_{jp}$$

$$H_a: At\ least\ one\ \beta_{\mathrm{jk}} \neq 0, where\ \mathrm{k} \in [1, p]$$

$$p_{\mathrm{value,j=2}} = 8.552\ x\ 10^{-06}, \qquad p_{\mathrm{value,j=3}} = 2.2\ x\ 10^{-16}$$

For both submodels, the p_{value} is less than the test level $\alpha = 0.05$, allowing us to reject the null hypothesis (Table 3). Therefore, we can conclude that there is at least one predictor in the model which has a significant effect on the outcome of nicotine usage status.

Discussion

Between both submodels, Gender, Education, Conscientiousness, Openness, and Impulsivity scores were found to be significant to the level of at least $\alpha = 0.05$ (Table 2). The Openness Score was significant to the level $\alpha = 0.001$ in both models and increases in Openness Score was associated with an increased probability of being a Past User or Recent User of Nicotine. The Conscientiousness Score and Impulsivity Score were significant for outcome of Recent User but were found to be insignificant for the model containing Past User. An increase in Conscientiousness Score and Impulsivity Score was associated with a decreased the probability of belonging to the Recent User category. Gender was significant to the level $\alpha = 0.001$ for the Recent User model and $\alpha = 0.05$ for the Past User model. In both models, males had a higher probability of being a Recent or Past User of nicotine when compared to females. Education was less significant in the Past User model where High School graduates were found to be significantly more likely to be Past Smokers, but little else could be determined as no other Education levels were found significant. In the case of the Recent User model, Education levels of high school, some high school, and some colleges were significantly associated with increased probability of being a Recent User while the university Education level ($\alpha = 0.001$) decreased the probability of belonging to the Recent Smoker category.

There is possible bias in the Education category because majority of Recent Users were between the ages of 18-24 and may have not yet completed or started university compared to the Never Users and Past Users who had a more evenly distributed age level. That is, the age of participants could have a confounding effect with the level of education they have achieved. As well, this projects findings regarding Gender are consistent with recent studies which determined that women tend to smoke for reasons not related to nicotine exposure, such as weight loss or stress response, while men smoke to maintain nicotine levels resulting from addiction (Allen et al, 2016). Finally, the highly significant positive association between nicotine usage and Openness Scores may be due to nicotine's prevalence in counterculture and nonconventional spaces.

References

Allen, A. M., Scheuermann, T. S., Nollen, N., Hatsukami, D., & Ahluwalia, J. S. (2016). Gender Differences in Smoking Behavior and Dependence Motives Among Daily and Nondaily Smokers. *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, *18*(6), 1408–1413. https://doi.org/10.1093/ntr/ntv138

Benowitz N. L. (2010). Nicotine addiction. The New England journal of medicine, 362(24), 2295–2303. https://doi.org/10.1056/NEJMra0809890

Faraway, J. (2016). Extending the linear model with R. Second Edition, Chapman and Hall. ISBN 9781498720960

Fehrman, Elaine & Muhammad, Awaz & Mirkes, Evgeny & Egan, Vincent & Gorban, Alexander. (2017). The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. 10.1007/978-3-319-55723-6_18

Saha, S. P., Bhalla, D. K., Whayne, T. F., Jr, & Gairola, C. (2007). Cigarette smoke and adverse health effects: An overview of research trends and future needs. The International journal of angiology : official publication of the International College of Angiology, Inc, 16(3), 77–83. https://doi.org/10.1055/s-0031-1278254

Dataset link: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

Appendix A

Additional Tables and Figures

Variable Name	Variable Description
ID	Observation ID
Age	Age Range Group
Gender	Gender
Education	Education Level
Country	Country of Orgin
Ethnicity	Ethnicity of Person
Nscore	Neuroticism Score: the long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression
Escore	Extraversion Score: manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics
Oscore	Openness Score: a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests
Ascore	Agreeableness Score: a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness
Cscore	Conscientiousness Score: a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient
Impulsive	Impulsivity Score: the tendincy to act on impulse
SS	Sensation Seeking Score: the tendency to pursue new and different sensations, feelings, and experiences
Nicotine	Nicotine Usage Status

Table A1: List of all relevant variables and descriptions

ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Nicotine
1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	CL2
2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	CL4
3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	CL0

Table A2: First three rows of raw data

ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Nicotine	NicotineL
1	35-44	Female	Certificate/Trade Degree	UK	Mixed	39	36	42	37	42	355	132	Used in Last Decade	Past User
2	25-34	Male	University	UK	White	29	52	55	48	41	307	223	Used in Last Month	Recent User
3	35-44	Male	Certificate/Trade Degree	UK	White	31	45	40	32	34	276	249	Never Used	Never Used

Table A3: First three rows of processed data

Country	Country Code	Education	Education Code	Age Group	Age Code	Ethnicity	Ethnicity Code	Gender	Gender Code
Australia	-0.09765	Some HS (Left before 16, Left at 16, Left at 17)	-2.43591, -1.73790, -1.43719	18-24	-0.95197	Mixed (Mixed-Black/Asian, Mixed-White/Asian, Mixed-White/Black)	1.90725, 0.12600, -0.22166	Male	-0.482
Canada	0.24923	HS Grad	-1.22751	25-34	-0.07854	Asian	-0.50212	Female	0.482
New Zealand	-0.46841	Some College	-0.61113	35-44	0.49788	Black	-1.10702		
Other	-0.28519	Certificate/Trade Degree	-0.05921	45-54	1.09449	White	-0.31685		
Ireland	0.21128	Bach	0.45468	55-64	1.82213	Other	0.11440		
UK	0.96082	>Bach (Masters degree, Doctorate degree)	1.16365, 1.98437	65+	2.59171				
USA	-0.57009								

Table A4: Coded values and labels for categorical variables

Neuroticism Score	Neuroticism Code	Extraversion Score	Extraversion Code	Agreeableness Score	Agreeableness Code	Conscientiousness Score	Conscientiousness Code	Openness to Experience Score	Openness to Experience Code
12	-3.46436	16	-3.27393	12	-3.46436	17	-3.46436	24	-3.27393
13	-3.15735	18	-3.00537	16	-3.15735	19	-3.15735	26	-2.8595
14	-2.75696	19	-2.72827	18	-3.00537	20	-2.90161	28	-2.63199
15	-2.52197	20	-2.5383	23	-2.90161	21	-2.72827	29	-2.39883
16	-2.42317	21	-2.44904	24	-2.78793	22	-2.57309	30	-2.21069
17	-2.34360	22	-2.32338	25	-2.70172	23	-2.42317	31	-2.09015
18	-2.21844	23	-2.21069	26	-2.5383	24	-2.30408	32	-1.97495
19	-2.05048	24	-2.11437	27	-2.35413	25	-2.18109	33	-1.82919
20	-1.86962	25	-2.03972	28	-2.21844	26	-2.04506	34	-1.68062
21	-1.69163	26	-1.92173	29	-2.07848	27	-1.92173	35	-1.55521
22	-1.55078	27	-1.7625	30	-1.92595	28	-1.78169	36	-1.42424
23	-1.43907	28	-1.6334	31	-1.772	29	-1.64101	37	-1.27553
24	-1.32828	29	-1.50796	32	-1.6209	30	-1.5184	38	-1.11902
25	-1.19430	30	-1.37639	33	-1.47955	31	-1.38502	39	-0.97631
26	-1.05308	31	-1.23177	34	-1.34289	32	-1.25773	40	-0.84732
27	-0.92104	32	-1.09207	35	-1.21213	33	-1.13788	41	-0.71727
28	-0.79151	33	-0.94779	36	-1.07533	34	-1.0145	42	-0.58331
29	-0.67825	34	-0.80615	37	-0.91699	35	-0.89891	43	-0.45174
30	-0.58016	35	-0.69509	38	-0.76096	36	-0.78155	44	-0.31776
31	-0.46725	36	-0.57545	39	-0.60633	37	-0.65253	45	-0.17779
32	-0.34799	37	-0.43999	40	-0.45321	38	-0.52745	46	-0.01928
33	-0.24649	38	-0.30033	41	-0.30172	39	-0.40581	47	0.14143
34	-0.14882	39	-0.15487	42	-0.15487	40	-0.27607	48	0.29338
35	-0.05188	40	0.00332	43	-0.01729	41	-0.14277	49	0.44585
36	0.04257	41	0.16767	44	0.13136	42	-0.00665	50	0.58331
37	0.13606	42	0.32197	45	0.28783	43	0.12331	51	0.7233
38	0.22393	43	0.47617	46	0.43852	44	0.25953	52	0.88309
39	0.31287	44	0.63779	47	0.59042	45	0.41594	53	1.06238
40	0.41667	45	0.80523	48	0.76096	46	0.58489	54	1.24033
41	0.52135	46	0.96248	49	0.94156	47	0.7583	55	1.43533
42	0.62967	47	1.11406	50	1.11406	48	0.93949	56	1.65653
43	0.73545	48	1.2861	51	1.2861	49	1.13407	57	1.88511
44	0.82562	49	1.45421	52	1.45039	50	1.30612	58	2.15324
45	0.91093	50	1.58487	53	1.61108	51	1.46191	59	2.44904
46	1.02119	51	1.74091	54	1.81866	52	1.63088	60	2.90161
47	1.13281	52	1.93886	55	2.03972	53	1.81175		
48	1.23461	53	2.127	56	2.23427	54	2.04506		
49	1.37297	54	2.32338	57	2.46262	55	2.33337		
50	1.49158	55	2.57309	58	2.75696	56	2.63199		
51	1.60383	56	2.8595	59	3.15735	57	3.00537		
52	1.72012	58	3.00537	60	3.46436	59	3.46436		
53	1.83990	59	3.27393						
54	1.98437								
55	2.12700								
56	2.28554								
57	2.46262								
58	2.61139								
59	2.82196								
60	3.27393								

Table A5: Coded values and associated scores for Neuroticism, Extraversion, Agreeableness, Conscientious, and Openness variables

Sensation Seeking Score	Sensation Seeking Code	Impulsivity Score	Impulsivity Code
71	-2.07848	20	-2.55524
87	-1.54858	276	-1.37983
132	-1.18084	307	-0.71126
169	-0.84637	355	-0.21712
211	-0.52593	257	0.19268
223	-0.21575	216	0.52975
219	0.07987	195	0.88113
249	0.40148	148	1.29221
211	0.76540	104	1.86203
210	1.22470	7	2.90161
103	1.92173		

Table A6: Coded values and associated scores for impulsivity and sensation seeking variables

	ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Nicotine	NicotineL
809	812	18-24	Male	HS Grad	USA	White	21	45	53	37	39	307	219	Never Used	Never Used
1103	1106	18-24	Female	HS Grad	USA	White	26	42	57	37	28	257	211	Never Used	Never Used
1004	1007	18-24	Female	HS Grad	USA	White	53	33	52	44	35	307	211	Never Used	Never Used

Table A7: Identified outlier observations

	Never Used = 0, Past User = 1		***Never Used = 0, Recent User = 1***	
	Estimate	***P Value***	***Estimate***	***P Value***
Intercept	*-0.962*	*1.512e-01*	*-0.077*	*9.048e-01*
Male	*0.292*	*3.232e-02**	*0.544*	*3.834e-05****
High School	*1.268*	*3.521e-03***	*1.223*	*3.692e-03***
Some College	*0.404*	*8.460e-02 .*	*0.495*	*2.227e-02**
Some High School	*0.232*	*4.366e-01*	*0.545*	*4.451e-02**
University	*0.074*	*6.980e-01*	*-0.510*	*5.410e-03***
Openness Score	*0.039*	*1.671e-04****	*0.074*	*5.201e-13****
Conscientiousness Score	*-0.018*	*8.267e-02 .*	*-0.049*	*1.122e-06****
Impulsivity Score	*0.0001*	*8.990e-01*	*-0.003*	*2.070e-03***
Null deviance: 3971.161		***Residual Deviance: 3675.02***	***AIC: 3711.02***	
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*				

Table A8: Model refit after outlier removal, no significant changes in coefficient significance

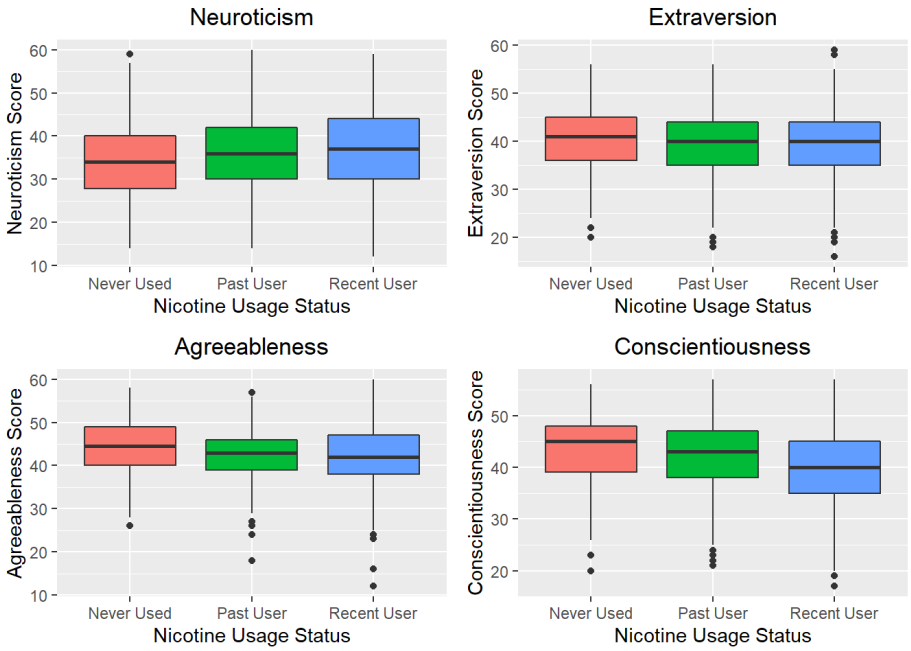

Figure A1: Boxplot distribution of for Neuroticism, Extraversion, Agreeableness, and Conscientious scores by response type

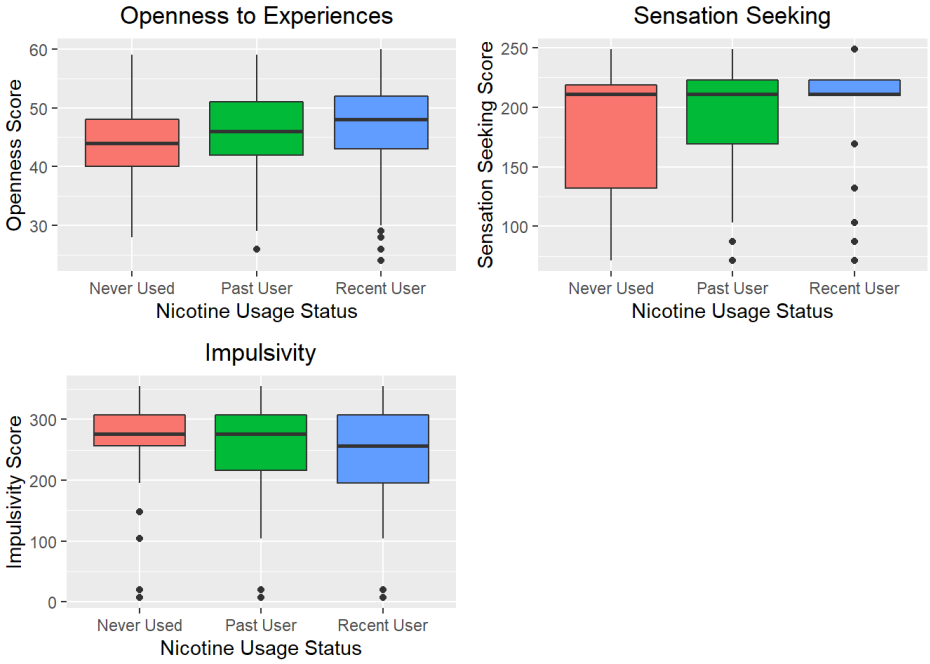

Figure A2: Boxplot distribution of for Openness, Sensation Seeking, and Impulsivity scores by response type

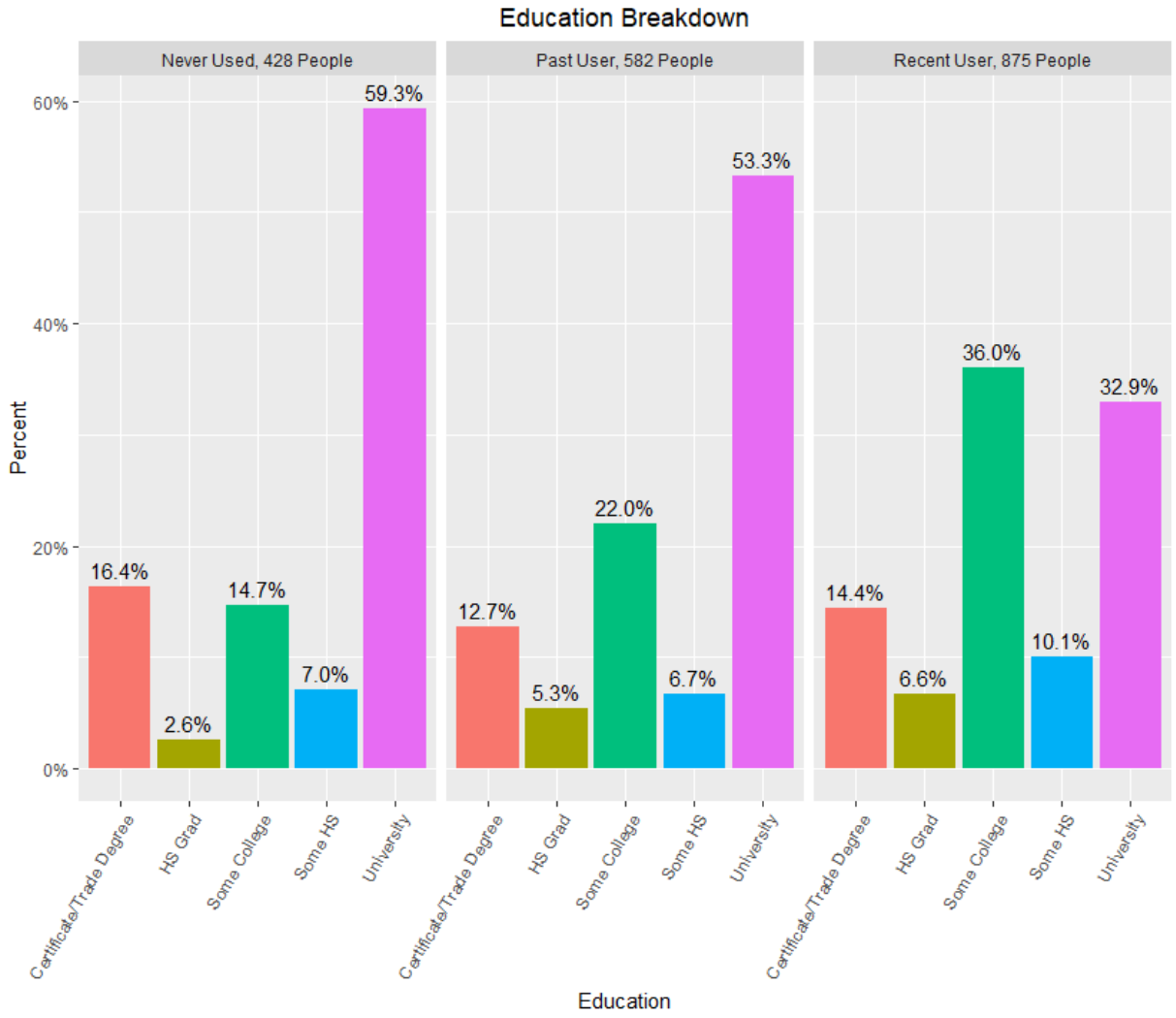

Figure A3: Percentage break down of different education levels

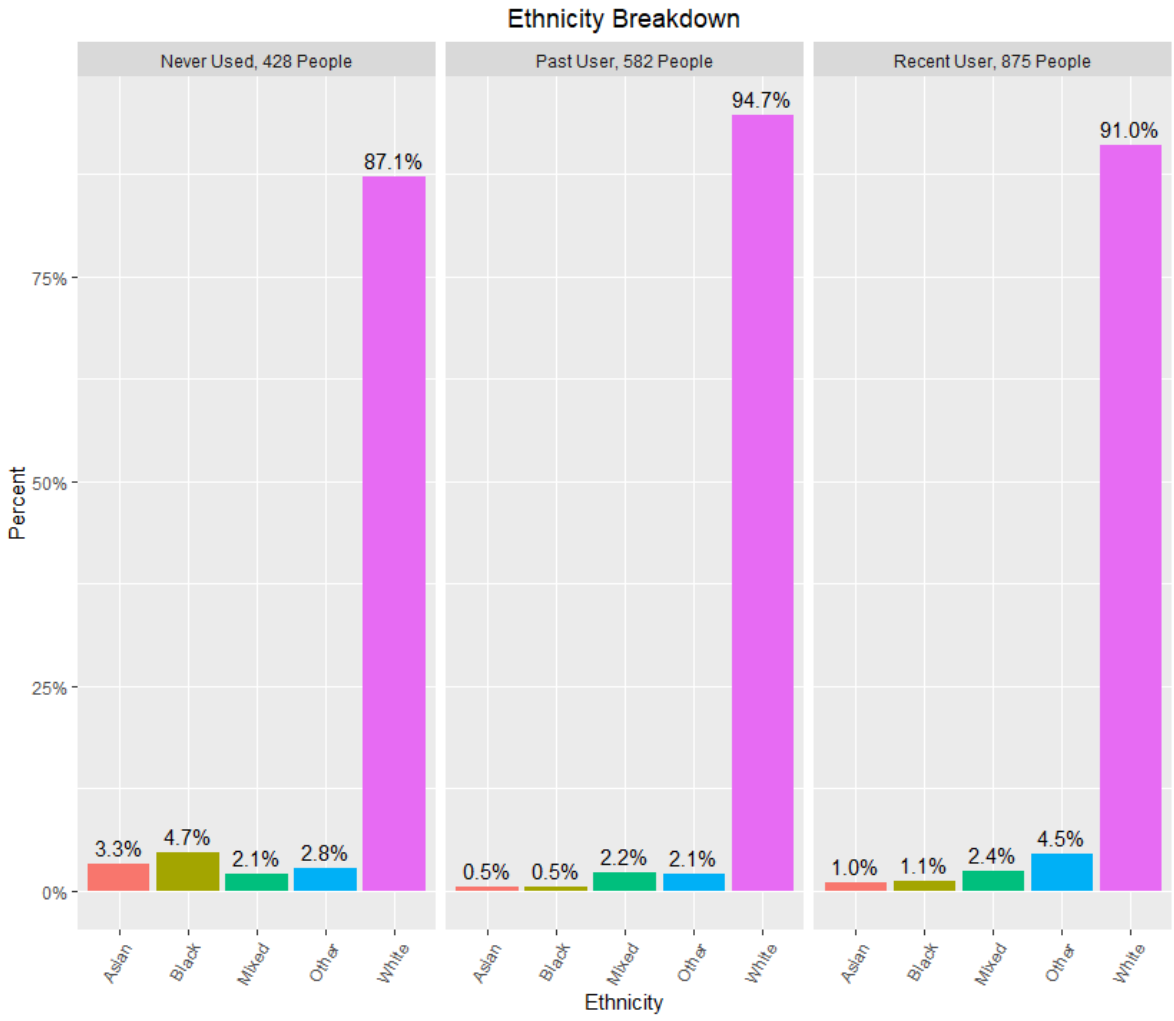

Figure A4: Percentage break down of different ethnicities by nicotine status

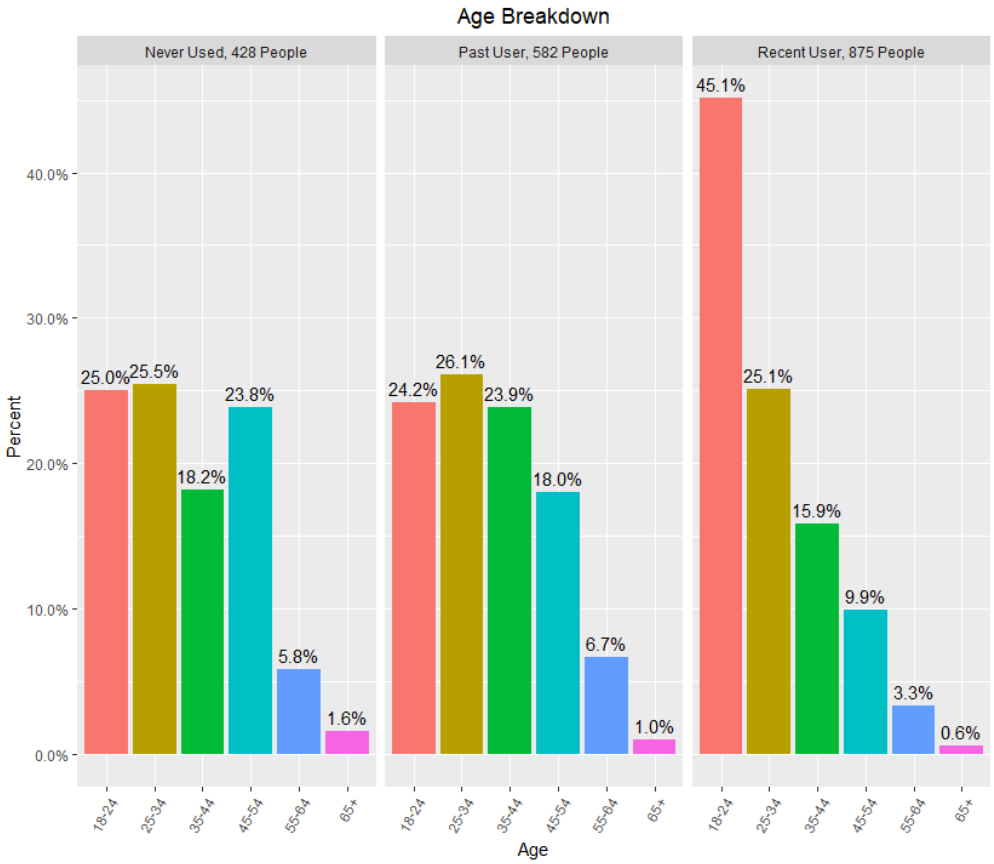

Figure A5: Percentage break down of different age groups by nicotine status

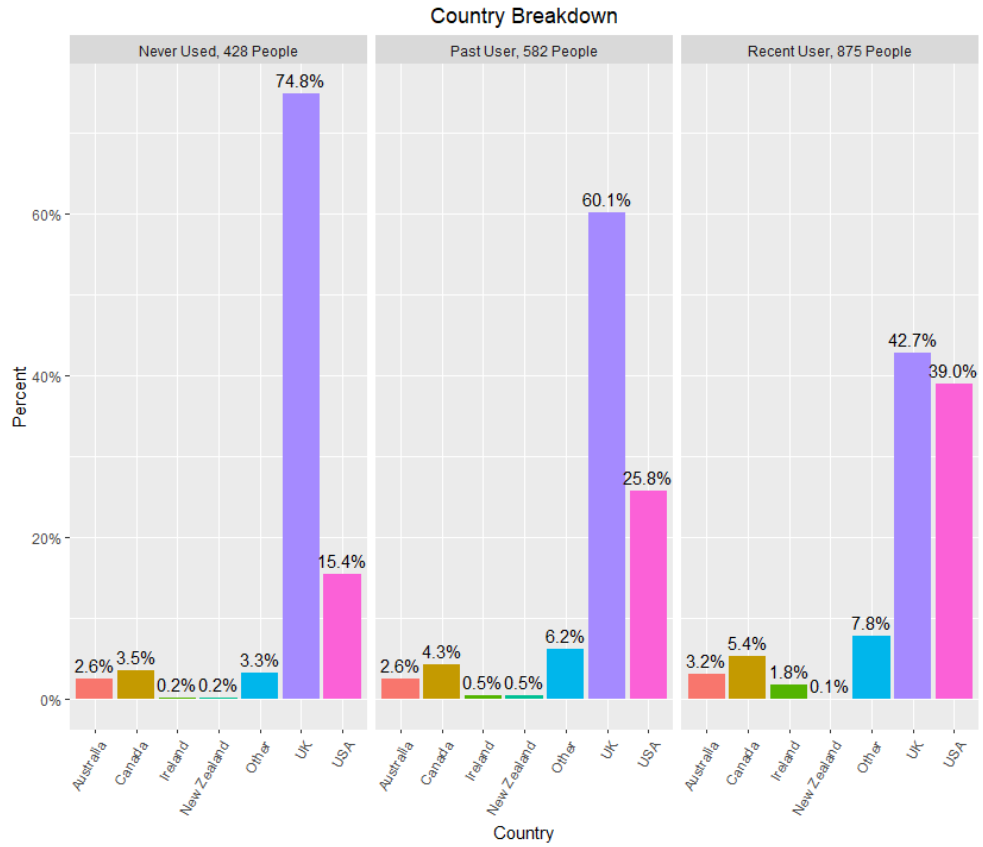

Figure A6: Percentage break down of different countries by nicotine status

Cook's Distance Plots

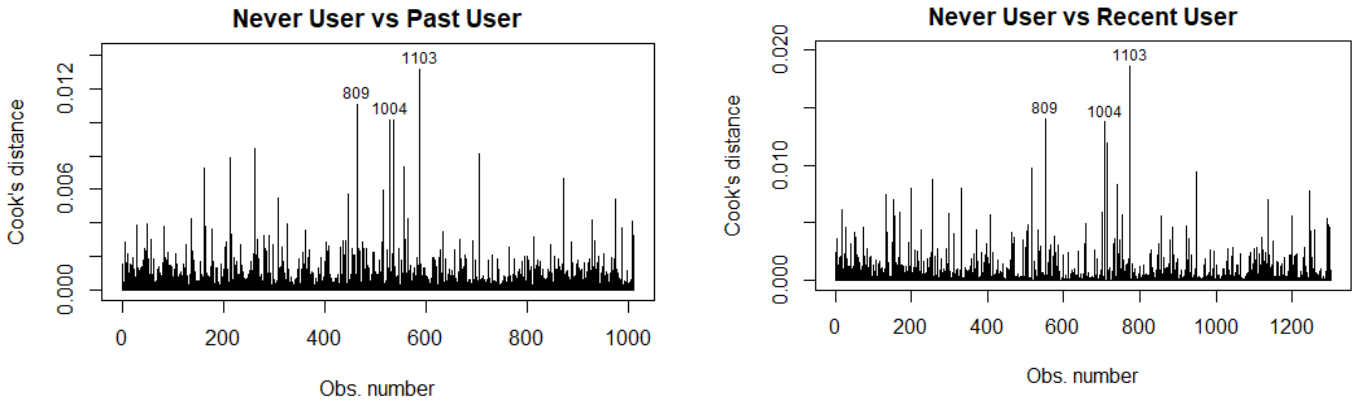

Figure A7: Cook's distance plots for both submodels, points 809, 1103, and 1004 identified in final model

Leverage Plots

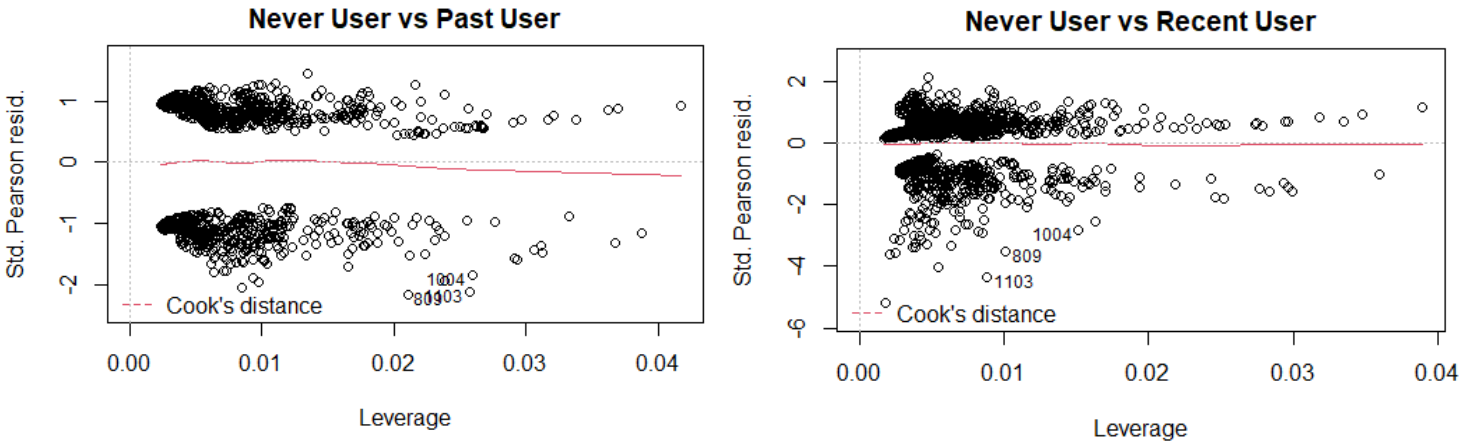

Figure A8: Leverage plots for both submodels, points 809, 1103, and 1004 identified in final model

Appendix B

R Code

```
library(MASS)
library(tidyverse)

setwd("C:/Users/chris/OneDrive/Desktop/STA 223/Project2/NicotineUsage/")

drug =
read.csv("data/drug_consumption.csv",header=F, col.names =
c("ID","Age","Gender","Education","Country","Ethnicity","Nscore","Escore","Oscore","Ascore","Cscore","Impulsive","SS","Alcohol","Amphet","Amyl","Benzos","Caff","Cannabis","Choc","Coke","Crack","Ecstasy","Heroin","Ketamine","Legalh","LSD","Meth","Mushrooms","Nicotine","Semer","VSA"))


#processing the numeric values into their represented data

#Gender
drug = drug%>%
  mutate(Gender = case_when(Gender < 0 ~"Male",
                            Gender > 0 ~ "Female"))
#Age
drug = drug%>%
  mutate(Age = case_when(Age ==  -0.95197 ~"18-24",
                         Age == -0.07854 ~ "25-34",
                         Age == 0.49788  ~ "35-44",
                         Age == 1.09449  ~ "45-54",
                         Age == 1.82213  ~ "55-64",
                         Age == 2.59171  ~ "65+"))

#Education
drug = drug%>%
  mutate(Education = case_when(Education %in% c(-2.43591,-1.73790,-1.43719) ~ "Some HS",
                               Education == -1.22751 ~ "HS Grad",
                               Education == -0.61113  ~ "Some College",
                               Education == -0.05921  ~ "Certificate/Trade Degree",
                               Education %in% c(1.16365,1.98437,0.45468) ~ "University"))


#Ethnicity
drug = drug%>%
  mutate(Ethnicity = case_when(Ethnicity %in% c(1.90725,0.12600,-0.22166) ~ "Mixed",
                               Ethnicity == -0.50212 ~ "Asian",
                               Ethnicity == -1.10702  ~ "Black",
                               Ethnicity == -0.31685  ~ "White",
                               Ethnicity == 0.11440  ~ "Other"))

#Country

drug = drug%>%
  mutate(Country = case_when(Country == -0.09765~ "Australia",
                             Country == 0.24923 ~ "Canada",
                             Country == -0.46841  ~ "New Zealand",
                             Country == -0.28519  ~ "Other",
                             Country == 0.21128  ~ "Ireland",
                             Country == 0.96082  ~ "UK",
                             Country == -0.57009  ~ "USA"))

nscore_key=as.data.frame(cbind(c(-3.46436,
                                 -3.15735,-2.75696,-2.52197,-2.42317,-2.34360,-2.21844,-2.05048,-1.86962,-1.69163,-1.55078,-
```

```
1.43907,-1.32828,-1.19430,-1.05308,-
0.92104,-0.79151,-0.67825,-0.58016,-
0.46725,-0.34799,-0.24649,-0.14882,-
0.05188,
                    0.04257 ,0.13606 ,0.22393
,0.31287 ,0.41667 ,0.52135 ,0.62967
,0.73545,0.82562,0.91093,1.02119,1.13281,1.
23461,1.37297,1.49158,1.60383,1.72012,1.8
3990,1.98437,2.12700,2.28554,

2.46262,2.61139,2.82196,3.27393),seq(12,60
,1)))
```

```
for(val in seq(1,dim(nscore_key)[1],1)){

drug[drug$Nscore==nscore_key$V1[val],]$Ns
core =  nscore_key$V2[val]}
```

```
escore_key=as.data.frame(cbind(c(-3.27393,-
3.00537,-2.72827,-2.53830,-2.44904,-
2.32338,-2.21069,-2.11437,-2.03972,-
1.92173,-1.76250,-1.63340,-1.50796,-
1.37639,-1.23177,-1.09207,-0.94779,-
0.80615,-0.69509,-0.57545,-0.43999,-
0.30033,-
0.15487,0.00332,0.16767,0.32197,0.47617,0.
63779,0.80523,0.96248,1.11406,1.28610,1.4
5421,1.58487,1.74091,1.93886,2.12700,2.32
338,2.57309,2.85950,3.00537,3.27393),c(16,
18,19,20,21,22,23,24,25,26,27,28,29,30,31 ,

32 ,33 ,34 ,35 ,36 ,37 ,38 ,39 , 40, 41, 42, 43,
44, 45, 46 ,47 ,

48 ,49 ,50 ,51 , 52, 53, 54, 55, 56, 58, 59)))
```

```
for(val in seq(1,dim(escore_key)[1],1)){

drug[drug$Escore==escore_key$V1[val],]$Es
core =  escore_key$V2[val]}
```

```
oscore_key=as.data.frame(cbind(c(-3.27393
,-2.85950  ,-2.63199  , -2.39883,-2.21069  ,-
2.09015  , -1.97495, -1.82919, -1.68062, -
1.55521, -1.42424, -1.27553,-1.11902,-
0.97631,-0.84732,-0.71727,-0.58331,-
0.45174,-0.31776,-0.17779,-0.01928,0.14143
,0.29338
,0.44585,0.58331,0.72330,0.88309,1.06238,1.
24033,1.43533,1.65653,1.88511,2.15324,2.4
4904,2.90161),c(24,26,28,29,30,31,32,33,34,
35,36,37,38,39,40,41,42,43,44,45,46,47,48,49
,50,51,52,53,54,55,56,57,58,59,60)))
```

```
for(val in seq(1,dim(oscore_key)[1],1)){

if(is_empty(drug[drug$Oscore==oscore_key$
V1[val],])==F){

drug[drug$Oscore==oscore_key$V1[val],]$Os
core =  oscore_key$V2[val]}}
```

```
ascore_key=as.data.frame(cbind(c(-3.46436,-
3.15735,-3.00537,-2.90161,-2.78793,-
2.70172,-2.53830,-2.35413,-2.21844,-
2.07848,-1.92595,-1.77200,-1.62090,-
1.47955,-1.34289,-1.21213,-1.07533,-
0.91699,-0.76096,-0.60633,-0.45321,-
0.30172,-0.15487,-
0.01729,0.13136,0.28783,0.43852,0.59042,0.
76096,0.94156,1.11406,1.2861,1.45039,1.61
108,1.81866,2.03972,2.23427,2.46262,2.756
96,3.15735,3.46436),c(12,16,18,23,24,25,26,
27,28,29,30,31,32,33,34,35,36,37,38,39,40,41
,42,43,44,45,46,47,48,49,50,51,52,53,54,55,5
6,57,58,59,60)))
```

```
for(val in seq(1,dim(ascore_key)[1],1)){

if(is_empty(drug[drug$Ascore==ascore_key$
V1[val],])==F){

drug[drug$Ascore==ascore_key$V1[val],]$As
core =  ascore_key$V2[val]}}
```

```
cscore_key=as.data.frame(cbind(c(-3.46436,-
3.15735,-2.90161,-2.72827,-2.57309,-
2.42317,-2.30408,-2.18109,-2.04506,-
1.92173,-1.78169,-1.64101,-1.51840,-
1.38502,-1.25773,-1.13788,-1.01450,-
0.89891,-0.78155,-0.65253,-0.52745,-
0.40581,-0.27607,-0.14277,-0.00665,
0.12331, 0.25953,
0.41594,0.58489,0.7583,0.93949,1.13407,1.3
0612,1.46191,1.63088,1.81175,2.04506,2.33
337,2.63199,3.00537,3.46436),c(17,19,20,21,
22,23,24,25,26,27,28,29,30,31,32,33,34,35,36
,37,38,39,40,41,42,43,44,45,46,47,48,49,50,5
1,52,53,54,55,56,57,59)))

for(val in seq(1,dim(cscore_key)[1],1)){

if(is_empty(drug[drug$Cscore==cscore_key$
V1[val],])==F){

drug[drug$Cscore==cscore_key$V1[val],]$Csc
ore =  cscore_key$V2[val]}}


iscore_key=as.data.frame(cbind(c(-2.55524,-
1.37983,-0.71126,-
0.21712,0.19268,0.52975,0.88113,1.29221,1.
86203,2.90161),c(20,276,307,355,257,216,1
95,148,104,7)))


for(val in seq(1,dim(iscore_key)[1],1)){

if(is_empty(drug[drug$Impulsive==iscore_ke
y$V1[val],])==F){

drug[drug$Impulsive==iscore_key$V1[val],]$
Impulsive =  iscore_key$V2[val]}}


sscore_key=as.data.frame(cbind(c(-2.07848,-
1.54858,-1.18084,-0.84637,-0.52593,-
0.21575,0.07987,0.40148,0.76540
,1.22470,1.92173),c(71,87,132,169,211,223,2
19,249,211,210,103)))


for(val in seq(1,dim(sscore_key)[1],1)){

if(is_empty(drug[drug$SS==sscore_key$V1[v
al],])==F){
  drug[drug$SS==sscore_key$V1[val],]$SS =
sscore_key$V2[val]}}


#Drug Usage
decode_usage = function(ROW) #use sapply
{switch(ROW,
      "CL0" ="Never Used",
      "CL1"= "Used over a Decade Ago",
      "CL2" ="Used in Last Decade",
      "CL3" ="Used in Last Year",
      "CL4" ="Used in Last Month",
      "CL5" ="Used in Last Week",
      "CL6" ="Used in Last Day")}


decoded=c(0)
for(col in drug[14:32]){
 decoded =
cbind(decoded,(as.data.frame(sapply(col,
decode_usage))))
}

drug[14:32] = decoded[2:20]


drug=drug[c(1:13,30)]
drug$NicotineL = drug$Nicotine
drug = drug%>%
 mutate(NicotineL = case_when(NicotineL
%in% c("Used in Last Month",
                              "Used in Last Week",
                              "Used in Last Day")
~ "Recent User",
                    NicotineL %in% c("Used in
Last Year",
                              "Used in Last
Decade",
                              "Used over a Decade
Ago") ~ "Past User",
                    NicotineL == "Never Used"  ~
"Never Used"))
```

```
drug= drop_na(drug)

#Tables
library(kableExtra)

raw =
read.csv("data/drug_consumption.csv",heade
r=F, col.names =
c("ID","Age","Gender","Education","Country",
"Ethnicity","Nscore","Escore","Oscore","Ascor
e","Cscore","Impulsive","SS","Alcohol","Amph
et","Amyl","Benzos","Caff","Cannabis","Choc",
"Coke","Crack","Ecstasy","Heroin","Ketamine"
,"Legalh","LSD","Meth","Mushrooms","Nicotin
e","Semer","VSA"))



variables =
c("ID","Age","Gender","Education","Country",
"Ethnicity","Nscore","Escore","Oscore","Ascor
e","Cscore","Impulsive","SS","Nicotine")

description = c("Observation ID",
          "Age Range Group",
          "Gender",
          "Education Level",
          "Country of Orgin",
          "Ethnicity of Person",
          "Neuroticism Score: the long-term
tendency to experience negative emotions
such
as nervousness, tension, anxiety and
depression",
          "Extraversion Score: manifested in
outgoing, warm, active, assertive, talkative,
cheerful, and in search of stimulation
characteristics",
          "Openness Score: a general
appreciation for art, unusual ideas, and
imaginative, creative, unconventional, and
wide interests",
          "Agreeableness Score: a dimension of
interpersonal relations, characterized by
altruism, trust, modesty, kindness,
compassion and cooperativeness",
          "Conscientiousness Score: a tendency
to be organized and dependable,
strong-willed, persistent, reliable, and
efficient",
          "Impulsivity Score: the tendincy to
act on impulse",
          "Sensation Seeking Score:  the
tendency to pursue new and different
sensations, feelings, and experiences",
          "Nicotine Usage Status")


cbind(variables,description)
%>%kable(col.names =c("Variable
Name","Variable Description"))%>%
kable_classic(full_width =F,lightable_options
=c("striped","bordered") )%>%
 column_spec (1:2,
          border_left = T,
          border_right = T)

head(raw[c(1:13,30)],3)%>%kable%>%
kable_classic(full_width =F,lightable_options
=c("striped","bordered") )%>%
 column_spec (1:14,
          border_left = T,
          border_right = T)



 usage_codes = cbind(c("CL0","CL1","CL2",
"CL3", "CL4", "CL5", "CL6"),
               c("Never Used","Used over a
Decade Ago","Used in Last Decade",
                "Used in Last Year","Used in Last
Month","Used in Last Week",
                "Used in Last Day"),c("Never
Used","Past User","Past User","Past
User","Recent User","Recent User","Recent
User"))


usage_codes %>%kable(col.names =c("Usage
Code","Usage Label","Simplified
Labels"))%>% kable_classic(full_width
=F,lightable_options
=c("striped","bordered") )%>%
 column_spec (1:3,
          border_left = T,
          border_right = T)
```

```
head(drug,3)%>%kable%>%
kable_classic(full_width =F,lightable_options
=c("striped","bordered") )%>%
  column_spec (1:15,
           border_left = T,
           border_right = T)%>%
  column_spec(c(15), bold = F, color = "black",
background = "gold")


age_key = cbind(c("18-24",
"25-34",
"35-44",
"45-54",
"55-64",
"65+",""),c("-0.95197",
"-0.07854",
"0.49788",
 "1.09449",
 "1.82213",
"2.59171","") )

gender_key =
cbind((c("Male","Female","","","","","")),c("-
0.482", "0.482","","","","",""))

Education_key = cbind(c("Some HS
\n \n
(Left before 16,
                \n  Left at 16,
                \n  Left at 17)",
"HS Grad",
"Some College",
"Certificate/Trade Degree",
"Bach",
">Bach \n \n
(Masters degree,\n

\n Doctorate degree)",""),
c("-2.43591,



-1.73790,

-1.43719","-1.22751",
"-0.61113",
"-0.05921",
"0.45468",
"1.16365,
\n 1.98437",""))

ethnicity_key = cbind(c("Mixed \n \n
                \n (Mixed-Black/Asian,
                \n Mixed-White/Asian,
                \n Mixed-
White/Black)","Asian","Black","White","Other
","","",""),c("\n \n \n

1.90725,\n \n

0.12600,\n -0.22166","-0.50212","-
1.10702","-0.31685","0.11440","","",""))

country_key =
cbind(c("Australia","Canada","New
Zealand","Other","Ireland","UK","USA"),c("-
0.09765", "0.24923","-0.46841","-0.28519",
"0.21128", "0.96082","-0.57009"))


catvars = do.call("cbind",list(country_key,
Education_key,age_key,ethnicity_key,gender_
key))


data.frame(catvars) %>%
  kable(caption = '',
      col.names = c("Country","Country
Code","Education","Education Code","Age
Group","Age Code","Ethnicity","Ethnicity
Code","Gender","Gender Code"))%>%
kable_classic_2(full_width
=F,lightable_options
=c("striped","bordered") )%>%
  column_spec (1:10,
           border_left = T,
           border_right = T)



numvars1 =
do.call("cbind",list(rev(nscore_key),
rev(rbind(escore_key,cbind(c("","","","","","","","
"),c("","","","","","","","")))),



rev(rbind(ascore_key,cbind(c("","","","","","","","
",""),c("","","","","","","","","")))),
```

```
rev(rbind(cscore_key,cbind(c("","","","","","","
",""),c("","","","","","","","","")))),
rev(rbind(oscore_key,cbind(c("","","","","","","
","","","","","","","","","","","","","","
,"","","","")))))

))


numvars2 = cbind(
rev(sscore_key),rev(rbind(iscore_key,cbind("
","")))))




data.frame(numvars1) %>%
  kable(caption = '',
      col.names = c("Neuroticism
Score","Neuroticism Code","Extraversion
Score","Extraversion Code","Agreeableness
Score","Agreeableness
Code","Conscientiousness
Score","Conscientiousness Code","Openness
to Experience Score","Openness to
Experience Code"))%>%
kable_classic_2(full_width
=F,lightable_options
=c("striped","bordered") )%>%
  column_spec (1:10,
           border_left = T,
           border_right = T)



data.frame(numvars2) %>%
  kable(caption = '',
      col.names = c("Sensation Seeking
Score","Sensation Seeking Code","Impulsivity
Score","Impulsivity Code"))%>%
kable_classic_2(full_width
=F,lightable_options
=c("striped","bordered") )%>%
  column_spec (1:4,
           border_left = T,
           border_right = T)
#Graphs


drug %>% ggplot(aes(x=Nicotine, y=Nscore,
fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine, y=SS,
fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine,
y=Impulsive, fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine, y=Escore,
fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine, y=Oscore,
fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine, y=Ascore,
            fill=Nicotine ))+
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

drug %>% ggplot(aes(x=Nicotine, y=Cscore,
            fill=Nicotine ))+
```

```
  geom_boxplot()+
  theme(legend.position = "none",
      axis.text.x = element_text(angle = 15,
                                 hjust = 0.5,
                                 vjust = 1))

#drop missing values, very few in this data
drug= drop_na(drug)

#graph the personality scores by Nicotine
status, couple of patterns present,
concientious between never and recent..etc
couple patterns.

ns = drug %>% ggplot(aes(x=NicotineL,
y=Nscore, fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Neuroticism Score",title="Neuroticism")

ss = drug %>% ggplot(aes(x=NicotineL, y=SS,
fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Sensation Seeking Score",title="Sensation
Seeking")

is = drug %>% ggplot(aes(x=NicotineL,
y=Impulsive, fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Impulsivity Score",title="Impulsivity")

es = drug %>% ggplot(aes(x=NicotineL,
y=Escore, fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Extraversion Score", title="Extraversion")

os = drug %>% ggplot(aes(x=NicotineL,
y=Oscore, fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Openness Score", title="Openness to
Experiences")

as = drug %>% ggplot(aes(x=NicotineL,
y=Ascore,
                fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Agreeableness Score",
title="Agreeableness")

cs = drug %>% ggplot(aes(x=NicotineL,
y=Cscore,
                fill=NicotineL ))+
  geom_boxplot()+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  labs(x = "Nicotine Usage Status", y =
"Conscientiousness Score", title=
"Conscientiousness")

gridExtra::grid.arrange(ns,es,as,cs)

gridExtra::grid.arrange(os,ss,is, ncol=2)

gridExtra::grid.arrange(os,cs, ncol=1)


drug%>%group_by(NicotineL)%>%count
length(which(drug$NicotineL=="Recent
User"))
#get labels, percent age for each value.
Unfortunatly not used too much because
there is a more efficent way in ggplot using
..prop..
generate_labels =  function(COL){
  levels = unique(COL)
  labels = c()
  for (level in levels){


lvln=(count(as.data.frame(which(COL==level
))))
    labels=rbind(labels, c(lvln,
```

```
            (round(lvln/length(COL),
digits=3)*100)))
 }
 labels=(t(labels))
 colnames(labels)= levels
 return(labels)
}
```

```
#mutate labels so we can grasp how many in
each category in each outcome level
drug%>%ggplot(aes(x=Education,fill=Educat
ion))+
 geom_bar(stat = "count")+
 facet_wrap(~NicotineL)+
 theme(axis.text.x = element_text(angle=40))
```

```
drug=drug%>%
 mutate(NicotineL=case_when(NicotineL ==
"Never Used" ~ paste("Never
Used,",428,"People"),
       NicotineL == "Recent User" ~
paste("Recent User,",875,"People"),
       NicotineL == "Past User" ~ paste("Past
User,",582,"People")))
```

```
ggplot(drug, aes(x= Gender,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day", title =
"Gender Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
 theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))
```

```
ggplot(drug, aes(x= Education,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day",
title="Education Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
 theme(legend.position = "none", axis.text.x =
element_text(angle=60, hjust = 1), plot.title =
element_text(hjust = 0.5))
```

```
ggplot(drug, aes(x= Ethnicity,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day", title =
"Ethnicity Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
 theme(legend.position = "none", axis.text.x =
element_text(angle=60, hjust = 1), plot.title =
element_text(hjust = 0.5))
```

```
ggplot(drug, aes(x= Age,  group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day", title = "Age
Breakdown") +
   facet_grid(~NicotineL) +
```

```
   scale_y_continuous(labels =
scales::percent)+
  theme(legend.position = "none", axis.text.x =
element_text(angle=60, hjust = 1), plot.title =
element_text(hjust = 0.5))



ggplot(drug, aes(x= Country,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day", title =
"Country Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
  theme(legend.position = "none", axis.text.x =
element_text(angle=60, hjust = 1), plot.title =
element_text(hjust = 0.5))



ggplot(drug, aes(x= Gender,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
              y= ..prop.. ), stat= "count", vjust = -
.5,size=5) +
   labs(y = "Percent", fill="day", title =
"Gender Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
  theme(legend.position = "none", plot.title =
element_text(hjust = 0.5))+
  theme(
        axis.text = element_text(size=19),
        strip.text.x = element_text(size = 19),
        title = element_text(size = 22),
        legend.text = element_text(size=19),
        legend.title = element_text(size=20))
```

```
#get levels break down with percents

#easier to just make the data for this
leveldat <- data.frame(
  label=c("Never Used","Recent User","Past
User"),
  value=c(428,875,582)
)


leveldat <- leveldat %>%
  arrange(desc(label)) %>%
  mutate(prop = round(value /
sum(leveldat$value) *100,2)) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop )

# Basic piechart
leveldat%>%
  ggplot( aes(x="", y=prop, fill=label)) +
  geom_bar(stat="identity", width=1,
color="white") +
  coord_polar("y", start=0) +

  geom_text(aes(y = ypos, label =
paste(as.character(value),",","\n",prop,'%',se
p="")), color = "black", size=7)+
  labs(fill="Nicotine Status",x="",y="",
title="Nicotine Status Breakdown")+
  theme(axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        strip.text.x = element_text(size = 10),
        title = element_text(size = 22),
        legend.text = element_text(size=19),
        legend.title = element_text(size=20))

# Basic piechart
leveldat%>%
  ggplot(aes(x="", y=value, fill=label)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)

ggplot(drug, aes(x= Country,
group=NicotineL)) +
   geom_bar(aes(y = ..prop.., fill = factor(..x..)),
stat="count") +
```

```
   geom_text(aes( label =
scales::percent(..prop..,accuracy =  0.1L),
             y= ..prop.. ), stat= "count", vjust = -
.5) +
   labs(y = "Percent", fill="day", title =
"Country Breakdown") +
   facet_grid(~NicotineL) +
   scale_y_continuous(labels =
scales::percent)+
  theme(legend.position = "none", axis.text.x =
element_text(angle=60, hjust = 1), plot.title =
element_text(hjust = 0.5))
#Modeling
library(tidyverse)
library(MASS)
library(nnet)
library(lmtest)
library(AER)
source("preprocessing.R")
head(drug)

dim(drug)

drug= drop_na(drug)


drug2 = drug %>%mutate(Nicotine_nu =
case_when(NicotineL=='Never Used'~1,
                          TRUE ~ 0))

drug2 = drug2 %>%mutate(Nicotine_pu =
case_when(NicotineL=='Past User'~1,
                          TRUE ~ 0))

drug2 = drug2 %>%mutate(Nicotine_ru =
case_when(NicotineL=='Recent User'~1,
                          TRUE ~ 0))



drug2=drug2[,-c(14,15)]


empty_bo =
multinom(cbind(Nicotine_nu,Nicotine_pu,Nic
otine_ru) ~ 1, data=drug2)

drug.bo <-
multinom(cbind(Nicotine_nu,Nicotine_pu,Nic
otine_ru) ~ ., data=drug2)
summary(drug.bo)
drug.bo_bic = stepAIC(drug.bo, k =
log(dim(drug)[1]), trace=F)
summary(drug.bo_bic)


drug.lrempty = glm(cbind(Nicotine_ru,
Nicotine_nu)~1, data = drug2,
family=binomial("logit"))
drug.2rempty = glm(cbind(Nicotine_pu,
Nicotine_nu)~1, data = drug2,
family=binomial("logit"))

drug.lr1 = glm(cbind(Nicotine_ru,
Nicotine_nu)~ Gender + Education + Oscore +
Cscore + Impulsive, data = drug2,
family=binomial("logit"))
plot(drug.lr1)

anova(drug.lrempty,drug.lr1,test ="Chi")


drug.lr2 = glm(cbind(Nicotine_pu,
Nicotine_nu)~ Gender + Education + Oscore +
Cscore + Impulsive, data = drug2,
family=binomial("logit"))
plot(drug.lr2)

anova(drug.2rempty,drug.lr2, test="Chi")

anova(drug.bo_bic,empty_bo, test="Chi")

confint(drug.bo_bic, level=0.9)

coeftest(drug.bo_bic)
empty_bo$edf

plot(drug.lr2,which=4,caption = "", main  =
"Never User vs Past User")
plot(drug.lr1,which=4,main
  =  "Never User vs Recent User" ,caption = "")
plot(drug.lr2,which=5,main
  =  "Never User vs Past User" ,caption = "")
plot(drug.lr1,which=5,main
  =  "Never User vs Recent User" ,caption =
"",cook.legendChanges
=NULL)
```

```
library(kableExtra)
#809, 1103, 1004
drug[c(809, 1103, 1004),]%>%kable%>%
kable_classic(full_width =F,lightable_options
=c("striped","bordered") )%>%
 column_spec (1:16,
          border_left = T,
          border_right = T)

#drug3=drug2[-c(809, 1103, 1004),]
#outlier removed

drug.bo_bicE = multinom(formula =
cbind(Nicotine_nu, Nicotine_pu, Nicotine_ru)
~
   1, data = drug3)

drug.bo_bic2 = multinom(formula =
cbind(Nicotine_nu, Nicotine_pu, Nicotine_ru)
~
   Gender + Education + Oscore + Cscore +
Impulsive, data = drug3)
summary(drug.bo_bic2)
coeftest(drug.bo_bic2)
coeftest(drug.bo_bic)
#submodel leverage and cooks


leverage = hatvalues(drug.lr1)

p <- length(coef(drug.lr1))
n <- nrow(drug2)
plot(names(leverage), leverage, xlab="Index",
type="h")
points(names(leverage), leverage, pch=16,
cex=0.6)
infPts <- which(leverage>2*p/n)
susPts <-
as.numeric(names(sort(cooks[infPts],
decreasing=TRUE)[1:3]))
text(susPts, leverage[susPts], susPts, adj=1,
cex=0.7, col=4)


abline(h=2*p/n,col=2,lwd=2,lty=2)


# ** Cook's Distance ----------------
# high Cook's distance => influential
points/outliers
# leverage points with high Cook's distance
=> suspicious influential points & outliers
#                may need to be deleted -> check
scatterplots

cooks = cooks.distance(drug.lr1)

plot(cooks, ylab="Cook's Distance", pch=16,
cex=0.6)

susPts <-
as.numeric(names(sort(cooks[infPts],
decreasing=TRUE)[1:3]))
text(susPts, cooks[susPts], susPts, adj=2,
cex=0.7, col=4)
susPts

cats= c("Gender" , "Education")
generalhoslem::lipsitz.test(drug.po_bic)
generalhoslem::pulkrob.chisq(drug.po_bic,
catvars = cats)

generalhoslem::pulkrob.deviance(drug.po_bic
, catvars = cats)#fails at significane level 0.01



prd_prob_bo = predict(drug.bo_bic, type =
'prob')
prd_prob_bo = fitted(drug.bo_bic)
head(prd_prob_bo)
prd_labl_bo = predict(drug.bo_bic)
head(prd_labl_bo)

prd_prob_bo2 <- fitted(drug.bo_bic)


obslabel <- t(apply(drug2[,14:16], 1,
function(x) {
 res <- numeric(3)
 res[which.max(x)] <- 1
 res
}))
```

```
resP.bo <- sapply(2:ncol(obslabel),
function(m) {
  # baseline is column 1 here
  # otherwise you should replace "1" with the
corresponding index and adjust the range of
"m" accordingly
  obs_m <-
obslabel[rowSums(obslabel[,c(1,m)]) > 0, m]
  fit_m <-
prd_prob_bo2[rowSums(obslabel[,c(1,m)]) >
0,c(1,m)]
  fit_m <- fit_m[,2] / rowSums(fit_m)
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})


# m= 3 says, set u

par(mfrow=c(1,2))
m=2
fit_m <-
prd_prob_bo2[rowSums(obslabel[,c(1,m)]) >
0,c(1,m)]
fit_m <- fit_m[,2] / rowSums(fit_m)

plot(fit_m,resP.bo[[1]], main = "Never Used vs
Past User",ylab='Deviance Residuals',
xlab='Fitted Values')
lines(smooth.spline(fit_m,resP.bo[[1]],
spar=1.3), col=2,lwd =1.5)
abline(h=0, lty=2, col='grey')

m=3
fit_m <-
prd_prob_bo2[rowSums(obslabel[,c(1,m)]) >
0,c(1,m)]
fit_m <- fit_m[,2] / rowSums(fit_m)

plot(fit_m,resP.bo[[2]], main = "Never Used vs
Recent User",ylab='Deviance Residuals',
xlab='Fitted Values')
lines(smooth.spline(fit_m,resP.bo[[2]],
spar=1.3), col=2,lwd = 1.5)
abline(h=0, lty=2, col='grey')


boxplot(resP.bo[[2]])
boxplot(resP.bo[[1]])
```