2019

Effect of Common Drugs on Depression Tendencies

MARCH 15, 2019

AJINKYA SHETH |CHRIS LEE | HARRY XIE | REMYA KOSHY | ROXINE LEE

Abstract

Depression is a pervasive issue that plagues the United States. We chose to explore the relationship and understand the impact that factors like alcohol and drug consumption have on depression tendencies on different groups. In order to explore depression as a function of substance abuse, we took NHANES datasets that combined substance abuse and depressive factors. Our research revealed some statistically significant correlations could be drawn between marijuana and depression levels. However, our research fails to draw correlations between other drugs and alcohol with depression. Nevertheless, we provide our insights into how one might improve our models. The importance of our research outlines potential causes for depression, which can provide people with insights to make better decisions, and furthermore could inform policy discussions about healthy consumption patterns.

Contents

1. Introduction

Depression and drug use have plagued the United States. The Anxiety and Depression Association of America (AADA) stated that over 15 million adults are affected nationwide (2017). Furthermore, according to the World Health Organization (WHO), globally, more than 300 million people of all ages suffer from depression (WHO, 2017). On a different spectrum, in 2017, 30.5 million people aged 12 or older used an illicit drug in the past 30 days (Substance Abuse and Mental Health Services, 2017). Having an understanding of the destructive and pervasive nature of depression and substance abuse coupled with the large number of people affected by substance abuse and depression motivated our research. By exploring the factors of substance abuse on depression we were compelled to shed light on whether substance abuse aggravated depressive tendencies. Thereby helping individuals make better choices that lead to healthier lives.

Research Question: How does drug & alcohol use impact depression levels on non-military citizens in the US?

- Which drug is the strongest indicator of depression? (alcohol, marijuana, meth, cocaine, heroin)
- What is the frequency between the frequency of alcohol consumption and depression?
- Binge drinking and depression
- Frequency of drinking
- Does depression affect genders differently?

2. Methods

2.1 Dataset

The datasets we used were from the National Health and Nutrition Examination Survey (NHANES), which aims to assess the health and nutritional status of adults and children in the United States. In 2015-2016, 15,327 persons were selected for NHANES from 30 different survey locations. Of those selected, 9,971 completed the interview and 9,544 were examined. We decided on this data for our research, and while scoping our research questions, we narrowed in on the following datasets, which contained the data which was most important to answer the research questions we posed:

- Demographics Data
- Questionnaire Data
- Drug Use
- Alcohol Use
- Mental Health - Depression Screener

2.2 Initial Data Processing

Due to the immense size of NHANES data and its large number of variables, we needed to select and clean the data to fit our needs. After formulating and specifying the research question, the NHANES datasets were analyzed, first in a manual way and then using R.

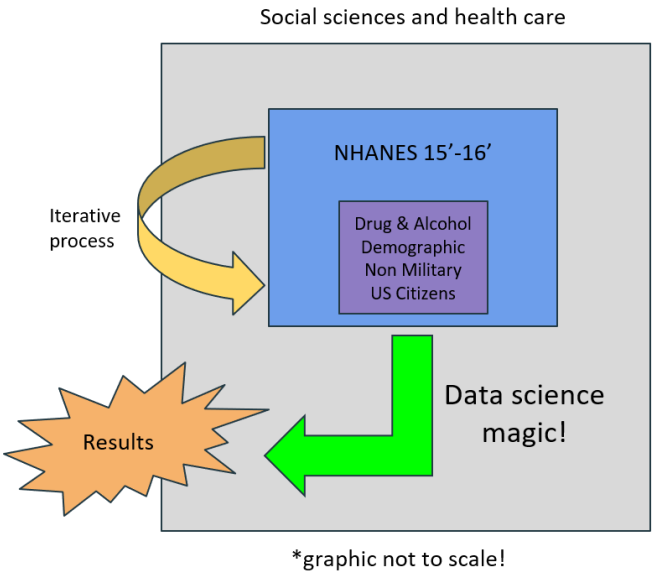


Figure 1 Research process

2.2.1 Drug and alcohol data selection process

For the use of alcohol and drugs, we first selected variables to filter the data, for example "whether heroin had been used", and if an interviewee's response for this question was negative, the corresponding record was not involved in our analytical process related to heroin use. On this basis, in order to unify the units of measurement for all questions, we chose similar survey questions for all types of drugs and alcohol. Specifically, the variables associated with the use of the drug and alcohol by the interviewees over the past one month was used for our analysis. After selecting the specific variables that needed to be analyzed, we cleaned the data. Variables are discussed in detail in their respective parts of the report. We removed all null values. We also deleted the corresponding values of the interviewees who were unwilling to answer the questions.

2.2.2 Demographic data selection process

For demographics data, we selected several items that are directly related to the research questions, including "military service status", "age", "gender", and "citizenship status". We removed non-citizens and military personnel due to the scope of our research question. Additionally, we felt that military people may be influenced by their service and be more prone to depression, making our research biased.

2.2.3 Depression score selection process

In the process of quantifying the degree of depression, we considered the results of the researchers of the NHANES questionnaire on depression, and finally decided to sum the reference values of 9 questions related to depressive symptoms in the questionnaire, resulting in a value range of 0 to 27 depression scores. We used this score in all the next analyses related to depression.

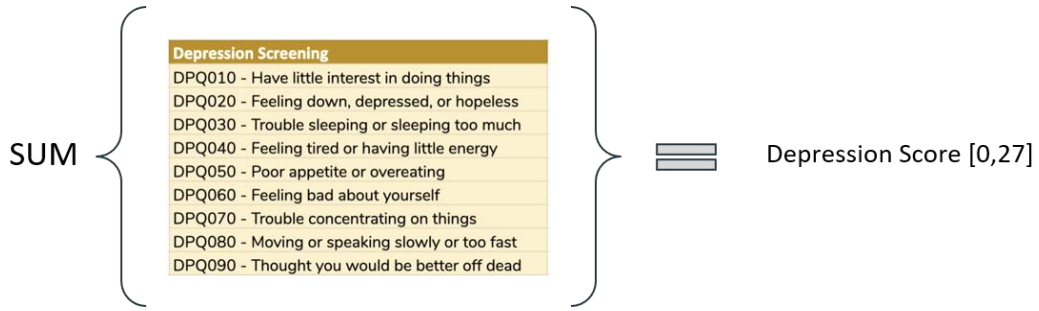


Figure 2 Sum of question values to produce our DepressionScore

2.3 Drug Use vs Depression Score methodology

As mentioned previously, we used the sum of the nine variables (DPQ010 to DPQ090) to calculate the variable, "DepressionScore". The following table illustrates the variables regarding four drugs (i.e., marijuana, cocaine, heroin, and meth) and the other two supplemental variables about injecting illegal drug and rehabilitation programs.

DUQ213 - Age started regularly using marijuana
DUQ217 - How often would you use marijuana?
DUQ219 - How many joints or pipes smoke (marijuana) in a day?
DUQ230 - # days used marijuana/month
DUQ250 - Ever use any form of cocaine
DUQ280 - # of days used cocaine/month
DUQ290 - Ever used heroin
DUQ320 - # of days used heroin/month
DUQ330 - Ever used methamphetamine
DUQ360 - # days used methamphetamine/month
DUQ370 - Ever use a needle to inject illegal drug
DUQ430 - Ever been in rehabilitation program

Figure 3 Selected survey questions on drugs

We divided all the variables into two groups. For those variables with specific values (i.e., DUQ213, DUQ217, DUQ219, DUQ230, DUQ280, DUQ320, and DUQ360), we used simple linear regression to explore its relationship with the depression score. Before starting the linear regression analysis, whether the distribution of the variables meets the assumption of linear regression (especially whether the residuals were nearly normally distributed) was checked with a residual plot and residual histogram

```
modDaysMJ <- lm(daysMJ$DepressionScore ~ daysMJ$DUQ230,
                data = daysMJ)
resDaysMJ <- resid(modDaysMJ)
plot(daysMJ$DUQ230, resDaysMJ)
abline(0, 0)
hist(resDaysMJ)
```

Figure 4 Example code for residual histogram

As for the yes-no questions (i.e., DUQ250, DUQ290, DUQ330, DUQ370, and DUQ430), we adopted Welch Two Sample T-test to check if there is a significant difference between the YES group and the NO

group. Before conducting the T-test, we have performed the F-test to see whether the variance between the two variables was equal or unequal, and thus we can determine the parameter “var.equal” in the t.test() function was TRUE or FALSE.

```
var.test(sumInject_Y, sumInject_N)

##
##  F test to compare two variances
##
## data:  sumInject_Y and sumInject_N
## F = 1.6842, num df = 62, denom df = 2251, p-value = 0.001518
## alternative hypothesis: true ratio of variances is not equal to 1

t.test(sumInject_Y, sumInject_N, var.equal = FALSE)

##
##  Welch Two Sample t-test
##
## data:  sumInject_Y and sumInject_N
## t = 2.2694, df = 64.076, p-value = 0.02662
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1951921 3.0654017
## sample estimates:
## mean of x mean of y
##  6.285714  4.655417
```

Figure 5 Example code for var.test and t.test

2.4 Alcohol vs Depression methodology

All the primary research for depression and alcohol has been derived from NHANES and CDC resources. We used the previously described depression score (2.2.3) to quantify depression. For alcohol, after filtering out the unnecessary fields, the final representation is as follows.

Alcohol Use
ALQ120Q - How often drink alcohol over past 12 mos
ALQ120U - # days drink alcohol per wk, mo, yr
ALQ130 - Avg # alcoholic drinks/day - past 12 mos
ALQ141Q - # days have 4/5 drinks - past 12 mos
ALQ141U - # days per week, month, year?
ALQ151 - Ever have 4/5 or more drinks every day?
ALQ160 - # days have 4/5 or more drinks in 2 hrs

Figure 6 Survey questions used for alcohol

All NHANES Data sets conveniently included a common field SEQN which acted as our Primary Key. We cleaned the data with standard practices such as filtering out the missing and irrelevant data. The result was three tables Alcohol_Clean, Depression_Clean, and Demographics_Clean. Based on definitions from

the CDC, excess consumption of alcohol is defined differently for males and females (2019). Hence, we needed gender information to infer the relationship between alcoholism and depression via gender. To keep things simple, we are joined the Demographic_Clean and Depression_Clean table to Alcohol_Clean table via inner join, to maintain all the information for analysis.

2.4.1 Quantification

Our data is largely qualitative in nature. Therefore, quantification was the most important step in our process, and it was the most time-consuming as well. In order to establish a metric for measuring alcohol consumption, we referred to the CDC's definition of 'drink'. As per the CDC, Drink in the US contains 0.6 ounces of pure alcohol. This amount is present in:

- 12-ounces of beer (5% alcohol content).
- 8-ounces of malt liquor (7% alcohol content).
- 5-ounces of wine (12% alcohol content).
- 1.5-ounces of liquor (40% alcohol content)

In our analysis, we are considering three metrics to measure alcohol consumption

1. Total Alcohol Consumption per year

We are considering the following fields for this calculation:

- ALQ120Q - How often drink alcohol over past 12 mos
- ALQ120U - # days drink alcohol per wk, mo, yr
- ALQ130 - Avg # alcoholic drinks/day - past 12 mos

Total Alcohol Consumption (TAC) = Number of Drinking Days * Number of Drinks per Day

TAC = f (ALQ120Q, ALQ120U) * ALQ130

2. Binge Drinking

As per the CDC, it's a pattern of drinking that brings blood alcohol concentration (BAC) to 0.08 gm percent or more. As per the CDC, this typically happens when men consume 5 or more drinks or women consume 4 or more drinks in about 2 hours or a single session.

We established the degree of Binge Drinking (BD) as Total Alcohol Consumed during binge drinking.

BD = number of binge drinking sessions * alcohol consumed in a typical binge drinking session

We consider two fields:

- ALQ141Q - # days have 4/5 drinks - past 12 mos
- ALQ141U - # days per week, month, year?

BD = f (ALQ141Q, ALQ141U) * (4 or 5)

- 4 for females and 5 for males

3. Recent Heavy Drinking

This metric essentially indicated heavy drinking in the last 30 days. We were curious if heavy drinking in the last 30 days indicated the presence of depression.

We establish this metric using the following field:

- ALQ160 - # days have 4/5 or more drinks in 2 hrs (In last 30 days)

The degree of heavy drinking (HD) = Number of heavy drinking sessions * alcohol consumed in heavy drinking sessions

HD = ALQ160 * (4 or 5)

3. Results

3.1 Drug Use vs. Depression Score

In case of marijuana, the results illustrated that there appeared to be a correlation between marijuana use and depression levels in men. And the degree of depression in women did not show a significant difference in moderate and high marijuana use.

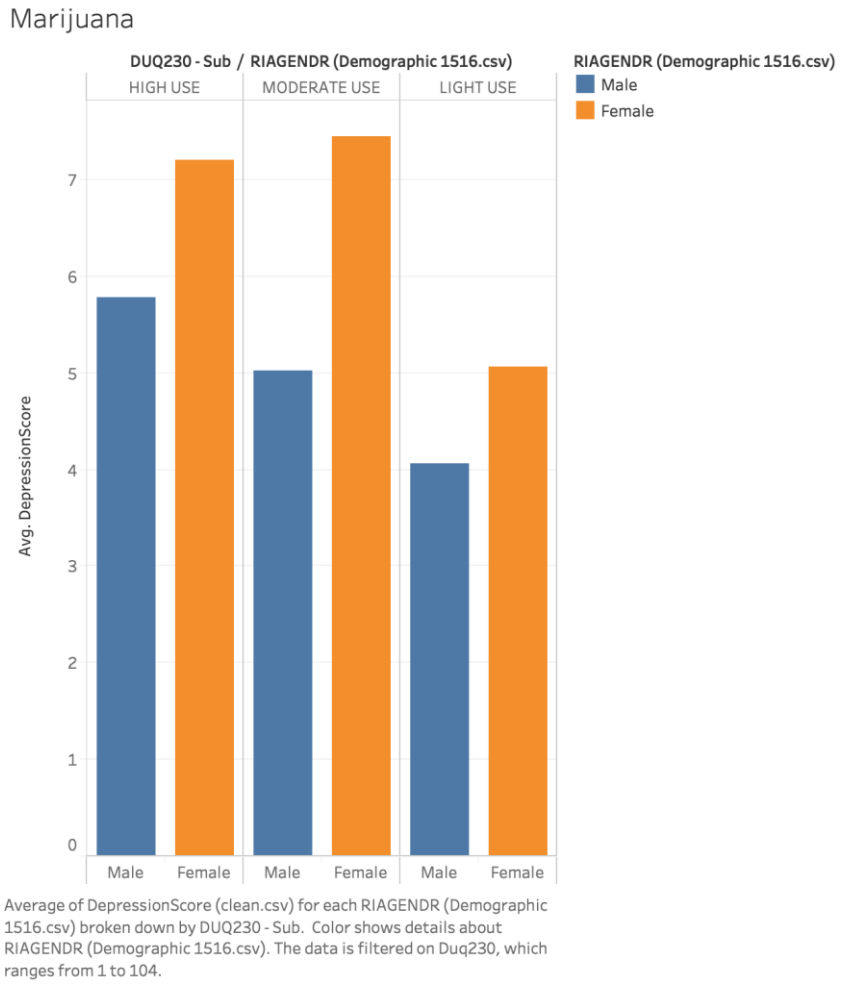


Figure 7 Marijuana use by gender vs Depression

As we mentioned, for those variables with specific values, simple linear regression was adopted. However, due to limitations such as sample size and some biases that would be discussed later, some of the variables did not meet the assumptions of linear regression that their residuals should be nearly normally distributed. Therefore, we just had two variables (i.e., DUQ217 and DUQ230) that produced significant results. That is, "an increase in the frequency of using marijuana (times and days per month) is associated with an expected increase in the depression score". The two linear regression formulas and scatter plots with best-fitting lines are shown as below.

```
DepressionScore = 0.34 * oftenMJ + 4.5 (p-value = 0.043)
DepressionScore = 0.06 * daysMJ + 4.74 (p-value = 0.004)
```

Figure 8 Linear regression formulas for marijuana

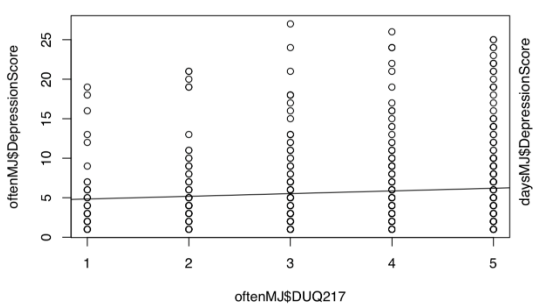


Figure 9 Respective chart for daysMJ

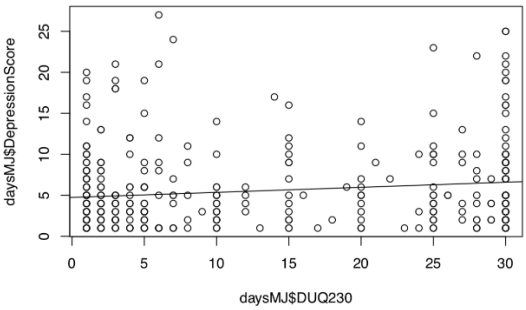


Figure 10 Respective chart for oftenMJ

As for the yes-no questions where we used Welch Two Sample T-test, we had three interesting results that showed significant differences between the depression scores of "those who have ever used heroin vs. those who haven't", "those who have ever used a needle to inject illegal drug vs. those who haven't", and "those who have ever been in a rehabilitation program vs. those who haven't". All the formers were more depressed than the latters. The following graph is an example of the box plot of the distribution of different groups' depression scores.

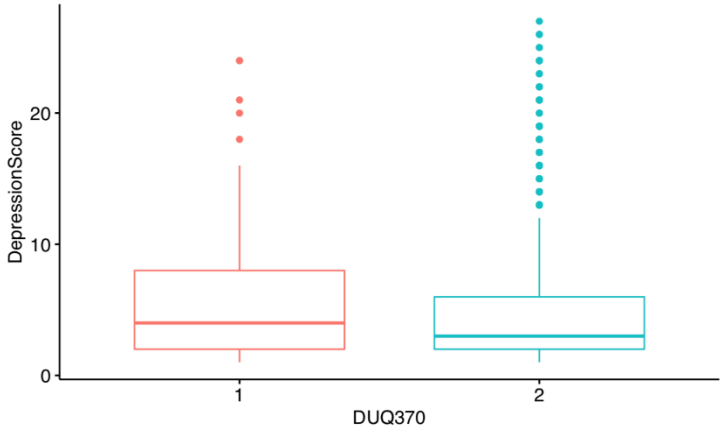


Figure 11 Example box plot of used heroin vs those who haven't and depression

3.2 Alcohol Consumption vs. Depression Score

Based on our knowledge of the background of the study, we believed that the degree of depression may be related to income. In order to understand the data, we selected the use of alcohol, depression, and

income as variables for visualization (Figure 12). In terms of results, most people in NHANES maintained lower alcohol use and lower levels of depression, which is in line with our expectations. The distribution of household income is balanced, showing that the data is not biased. By adjusting the filter, we found that the degree of depression in the relatively wealthy population was not sensitive to alcohol consumption. Specifically, wealthy people have less access to high depression scores than low-income people, even if they have relatively high levels of alcohol use.

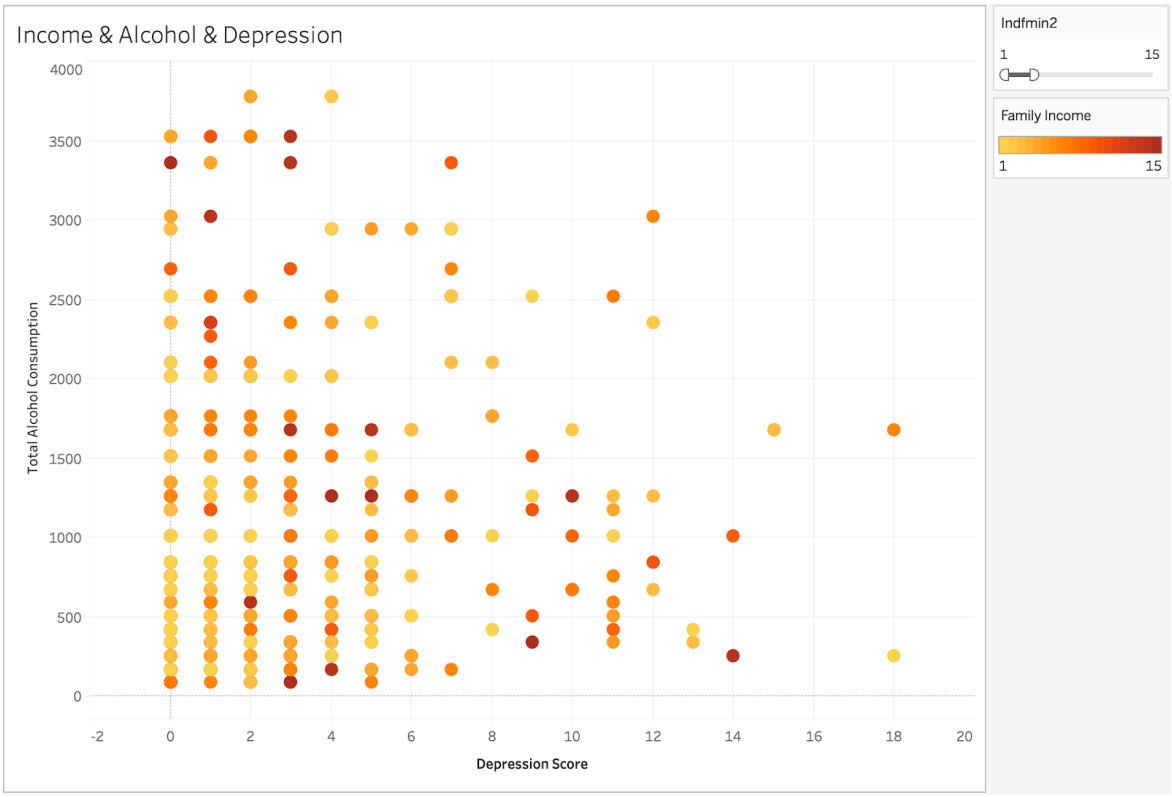


Figure 12 Income & Alcohol Depression

We were also interested in alcohol use, gender and depression scores. In order to get an abstract observation, we divided alcohol use into four categories: mild use (TAC <= 2000), moderate use (2000-4000 TAC) heavy use (4000-8000 TAC) , and crazy use (>8000 TAC). As can be seen from this chart (Figure 13), women have higher depressive scores compared to men in the case of similar alcohol use. This makes us wonder if women are more sensitive to the relationship between alcohol intake and depression levels. In addition, in moderate and severe alcohol use, male do not exhibit significant differences in depression levels, which makes us wonder whether alcohol use and depression have a linear regression relationship for men.

Alcohol

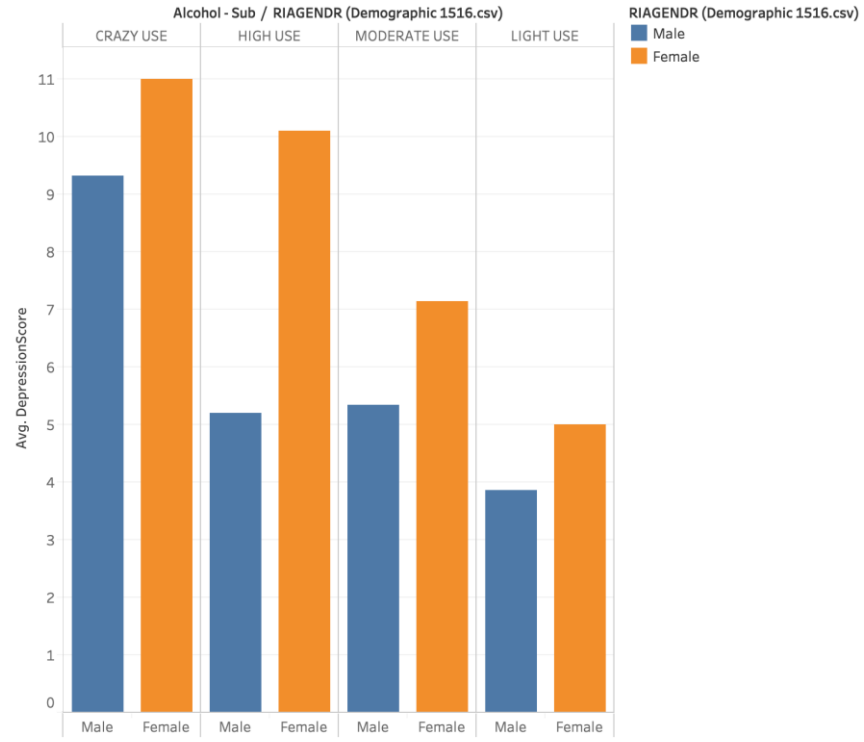


Figure 13 Alcohol vs Depression by gender

To further our results, we employed univariate regression over them. The results of our analysis are as follows:

1. Total Alcohol Consumption and Depression Score

```
lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$TotalConsumption,
    data = AlcoholAnalysis)

Residuals:
   Min     1Q Median     3Q    Max
-4.026 -2.990 -1.983  2.002 25.009

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.981e+00  1.539e-01  32.359   <2e-16 ***
AlcoholAnalysis$TotalConsumption 5.960e-06  7.833e-06   0.761    0.447
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.838 on 998 degrees of freedom
Multiple R-squared:  0.0005798, Adjusted R-squared:  -0.0004216
F-statistic: 0.579 on 1 and 998 DF,  p-value: 0.4469
```

Figure 14 Linear regression model for TAC and Depression Score

Since the p-value is insignificant and the slope is almost zero. We can infer that there is no correlation between Total Alcohol Consumption and Depression Score.

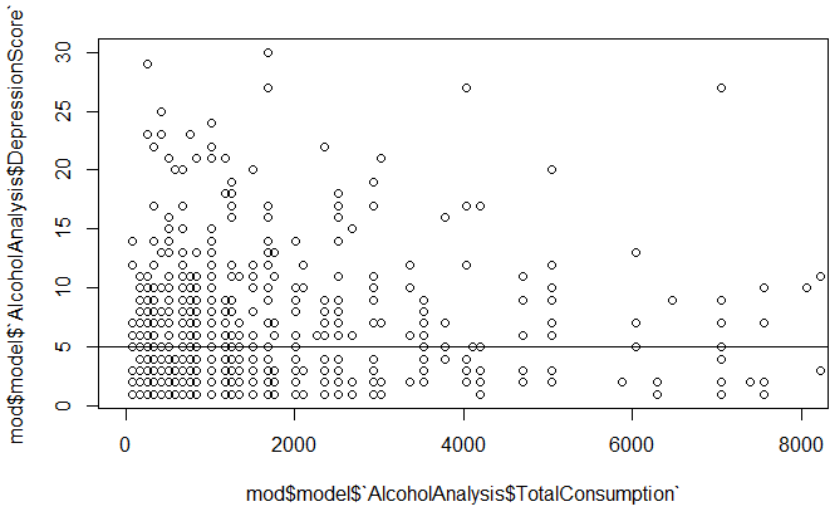


Figure 15 Total Consumption vs Depression Score

2. Depression Score and Binge Drinking

```
lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$BingeConsumption,
    data = AlcoholAnalysis)

Residuals:
   Min     1Q Median     3Q    Max
-4.061 -2.994 -1.992  2.006 24.999

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.992e+00  1.576e-01   31.68   <2e-16 ***
AlcoholAnalysis$BingeConsumption 1.058e-06  1.749e-05    0.06    0.952
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.839 on 998 degrees of freedom
Multiple R-squared:  3.663e-06, Adjusted R-squared:  -0.0009983
F-statistic: 0.003656 on 1 and 998 DF,  p-value: 0.9518
```

Figure 16 Depression Score vs Binge Drinking linear regression

Results for Binge drinking metric is the same as that for total alcohol consumption

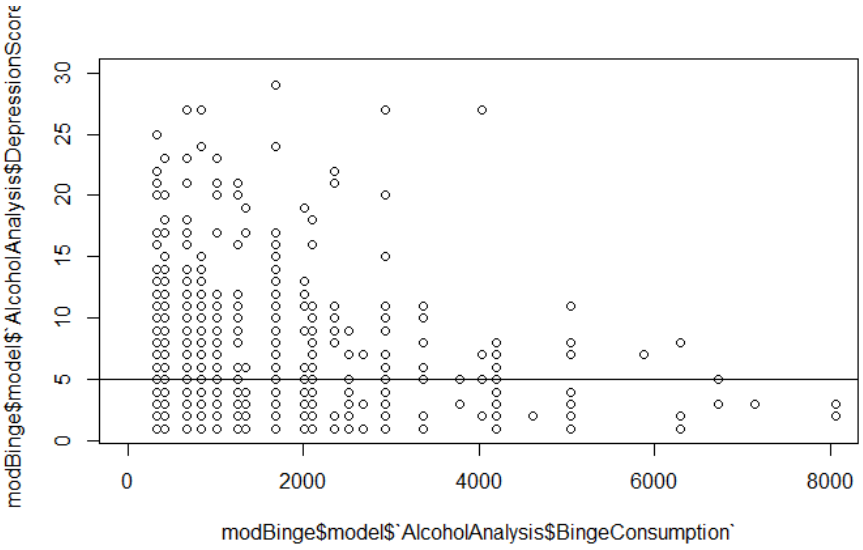


Figure 17 Binge Consumption vs Depression graph

3. Depression Score and Heavy Recent Drinking

```
lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$RecentAddiction,
    data = AlcoholAnalysis)

Residuals:
   Min     1Q Median     3Q    Max
-6.930 -3.092 -1.845  1.668 25.000

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   4.84546    0.16204  29.902  < 2e-16 ***
AlcoholAnalysis$RecentAddiction  0.03084    0.01119   2.755  0.00597 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.827 on 994 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.007579,  Adjusted R-squared:  0.006581
F-statistic: 7.591 on 1 and 994 DF,  p-value: 0.005973
```

Figure 18 Depression score and Heavy Drinking linear regression

For recent heavy drinking, we did observe some correlation as the P-value for this analysis is significant. However, this plot fails the Shapiro test for normal distribution. Hence, our inference is the same as earlier

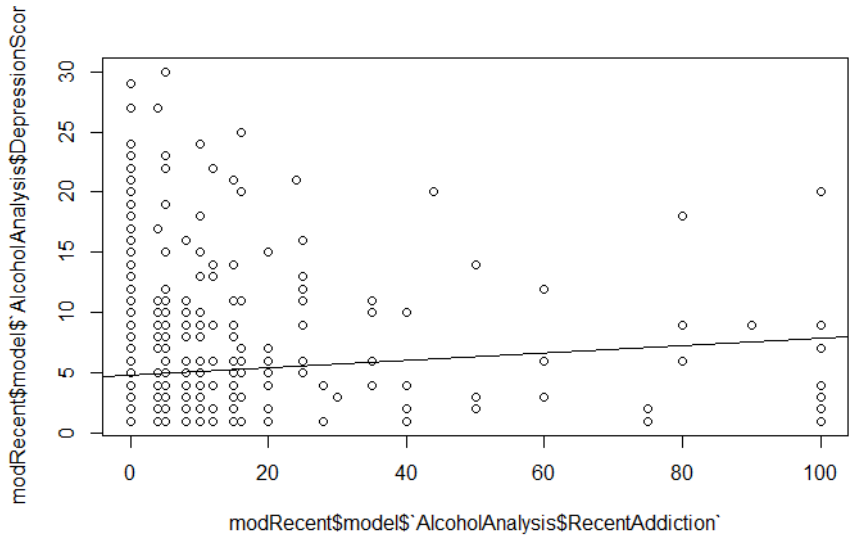


Figure 19 Recent addiction vs Depression Score graph

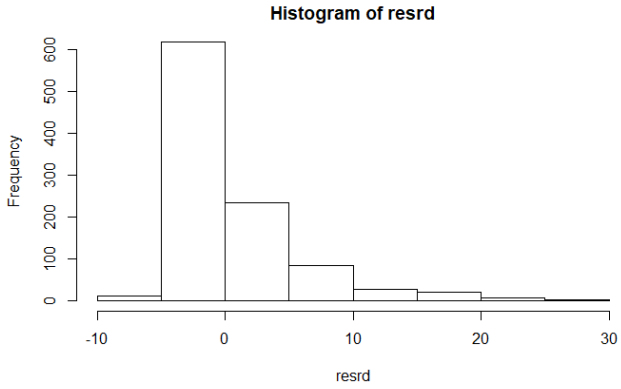


Figure 20Histogram of residuals of Recent Addiction vs Depression Score

3.3 Instrument optimization: A multiple logistic regression model

Besides analyzing the impact of alcohol use and drug use on people's depression tendency, we also tried to come up with ways to improve the nine-question depression screening instrument that we used to quantify people's depression. During our data analysis process, it seems that there must be different predictive abilities in terms of the nine symptoms (DPQ010 to DPQ090, as shown below). We kept thinking about the possibility for us to give weightage to these different symptoms or even build a new model to optimize this screening instrument.

DPQ010 - Have little interest in doing things
DPQ020 - Feeling down, depressed, or hopeless
DPQ030 - Trouble sleeping or sleeping too much
DPQ040 - Feeling tired or having little energy
DPQ050 - Poor appetite or overeating
DPQ060 - Feeling bad about yourself
DPQ070 - Trouble concentrating on things
DPQ080 - Moving or speaking slowly or too fast
DPQ090 - Thought you would be better off dead
DPQ100 - Difficulty these problems have caused

Figure 21 The nine questions used for depression score model

Therefore, we developed a tentative multiple logistic regression model, trying to better predict one's "probability of depression". To begin with, we took a look at the correlation coefficients of the nine items toward Question 10, which is the difficulty caused by these nine symptoms, then picking five items which have the highest correlation coefficients as highlighted below.

```
cor(depression_fn2)[11, ]

##        SEQN      DPQ010      DPQ020      DPQ030      DPQ040      DPQ050
## -0.01474193  0.36164054  0.46526939  0.27308684  0.31958126  0.28855599
##      DPQ060      DPQ070      DPQ080      DPQ090      DPQ100
##  0.48091484  0.42499980  0.39825480  0.43242678  1.00000000
```

Figure 22 Multivariate analysis

Next, we created a new column in our dataset, aiming to make the depression tendency "binary". We set the cutoff as 10 points (i.e., if one's depression score is larger than or equal to 10, he/she is labeled as "depressed"), which was defined by the authors who developed the nine-item instrument. Statistics of the logistic model can be summarized as below.

```
Call:
glm(formula = depressed ~ ., family = "binomial", data = dp_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8026  -0.2135  -0.1174  -0.1174   3.1565

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.97482    0.19849 -25.063  < 2e-16 ***
DPQ020       1.20505    0.11068  10.888  < 2e-16 ***
DPQ060       0.85346    0.10986   7.768 7.95e-15 ***
DPQ070       1.18536    0.09607  12.339  < 2e-16 ***
DPQ080       1.04883    0.11126   9.427  < 2e-16 ***
DPQ090       0.45365    0.22286   2.036   0.0418 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 23 Multivariate analysis results

After running the logistic model, we can try to predict one's probability of depression based on the existing data with new data given. From the two prediction results displayed as below, it is obvious that

even though the sum of these two people's answers to the five items are the same (equal to 7), they actually have quite different probabilities of depression, which we would not perceive with the original nine-item screening instrument.

```
new1 <- data.frame(DPQ020 = 3, DPQ060 = 2, DPQ070 = 1, DPQ080 = 1, DPQ090 = 0)
result1 <- predict(dp_glm, newdata = new1, type = "response")
result1
```

```
##         1
## 0.9296583
```

```
new2 <- data.frame(DPQ020 = 0, DPQ060 = 2, DPQ070 = 0, DPQ080 = 2, DPQ090 = 3)
result2 <- predict(dp_glm, newdata = new2, type = "response")
result2
```

```
##         1
## 0.5475403
```

Figure 24 Selection of two people with same scores

4. Discussion

4.1 Marijuana vs Depression

In the future, if we plan to keep refining the tentative logistic model defined above, we can take a look at its performance, such as calculating its accuracy with the confusion matrix, or check its sensitivity and specificity with the ROC curve and the AUC value. As shown in the graphs below, our current logistic model has a good predictive ability (AUC = 0.95).

```
##         prediction
## reality   0   1
##       0 728  14
##       1  34  48
```

```
cm[4] / sum(cm[, 2])
```

```
## [1] 0.7741935
```

```
cm[1] / sum(cm[, 1])
```

```
## [1] 0.9553806
```

```
accuracy <- sum(diag(cm)) / sum(cm)
accuracy
```

```
## [1] 0.9417476
```

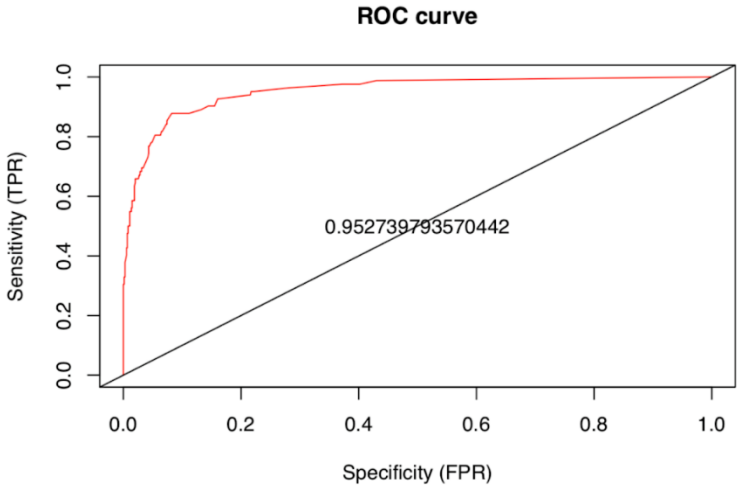


Figure 25 Improving the model

4.2 Alcohol vs Depression

Since we failed to reject the null hypothesis, we conclude that our analysis is insufficient or there is no correlation at all between alcohol consumption or drinking habits and the tendency for depression.

We believe our analysis is insufficient because of our Alcohol Consumption metrics defined earlier are accurate in indicating the alcohol consumption, however, are not precise in indicating how drunk the individual is. CDC defines binge drinking as drinking habit which brings the blood alcohol concentration level to 0.08 gm %. While this definition is scientifically sound, the premise that men require 5 drinks

(and women 4) to reach that concentration level is fallacious. This is because there are other factors like a person's weight and height. We need a more personalized approach for both data collection and for further analysis.

With respect to the survey questionnaire, we believe there is a possibility that respondents may have failed to estimate their own drinking habits. When asking questions which over a broad timeline (for example how many days per week, per month, or per year did you drink?) Inaccurate time perception comes into effect which is best explained by Vierordt's law (Shorter intervals tend to be overestimated while longer intervals tend to be underestimated). More research is needed to understand time perception and find ways to nullify it through reformulating questions. We also recommend NHANES to add a question to establish social or solo drinking habits. Drinking in solitude is often associated with addiction or depression. (Addiction Resource, 2018)

4.3 General

After conducting analysis on the effect of drugs and alcohol on reported depressed behavior, we found out that no statistically significant relation could be drawn between alcohol consumption and depression. It was similar for drugs and depression, where a relationship could only be established for very few factors.

Ethical considerations were a part of the data analysis process from the start. We researched findings from our research and analysis, in order to verify that other researchers had also either reached the same conclusions, and in case there was a wide discrepancy, knew that more work and more data was required. We also have not extrapolated the findings of this research to a wider community, because of the limitations with the data. Some of the limitations were:

1. There was very limited data about drug users with the filters we had scoped to (i.e. non-military personnel and only citizens of the U.S.). We did run tests using this data, but we also understand that these results cannot be used to predict depression in a wider audience.

2. Because the entire test is self-reported, there is a chance that the answers given by people are not accurate. This runs the risk of predictions being made off incorrect data.

3. There is also the chance of mistakes in the data, which again could affect accuracy of predictions.

4. For drug usage, there weren't direct questions about whether they were current drug users, which may have an impact on depressive tendencies. Detailed data was only available for Marijuana usage (and even then, did have answers to whether they were current drug users), and not for any of the other drugs considered in this research.

5. We also postulated that a lot of current or recent drug users could currently be imprisoned, which would mean that data from a vital section of people isn't present in the dataset.

We also found some interesting relationships while analyzing drug use, and one of the most interesting relationships was between a person attending rehabilitation facilities for substance abuse, and depression levels. We found the relationship that a person attending rehab was more likely to be depressed, which isn't the most immediately obvious conclusion. This is something we're interested in looking into further, as rehab is one of the best ways to target and overcome substance abuse.

We understand that in order for our research to be more indicative of the larger community, we need to consider many more data points about the usage, demographics, etc., and the relationships of these data points on each other. There should also be a much larger set of people included in the surveys, in order to test for diverse factors like racial background, socio-economic factors, history of mental illness in the family, medical details of any known markers for depression, etc.

5. Conclusion

Depression is a pervasive illness that affects not only the affected individual, but also close friends and family of the individual. Conducting this type of research is very important, as an individual prone to depression may unwittingly be indulging in behaviors that exacerbate depressive tendencies. We hope that more research along the same lines, with a larger scope, finds definite and strong relationships between various substances and depression. This could then help people make more informed decisions that help them live happier, healthier lives. This benefits not only individuals, but also the community around them, leading to a happier and more prosperous society.

References

Addiction Resource. (2018, December 3). Drinking Alone: Are You in Danger of Becoming an Alcoholic? Retrieved March 15, 2019, from https://addictionresource.com/alcohol/resources/drinking-alone/

"Anxiety and Depression Facts & Statistics." Anxiety and Depression Association of America. https://www.adaa.org/about-adaa/press-room/facts-statistics. Accessed June 2017.

CDC. (n.d.). CDC - Fact Sheets-Alcohol Use And Health - Alcohol. Retrieved March 15, 2019, from https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2003). The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical care, 1284-1292.

World Health Organization . "Depression." WHO, http://www.who.int/mediacentre/factsheets/fs369/en/. Accessed June 2017.

Substance Abuse and Mental Health Services Administration. (n.d.). Retrieved from https://www.samhsa.gov/

Appendix:

Code is attached below this document.

Alcohol Consumption

Ajinkya Sheth

March 9, 2019

```
alcohol = read.csv("./Alcohol Use_1516.csv", header=TRUE)
demographic = read.csv("./Demographic_1516.csv", header=TRUE)
depression = read.csv("./Depression_1516.csv", header=TRUE)
drug = read.csv("./Drug Use_1516.csv", header=TRUE)
```

Variables considered for this analysis - Alcohol table

ALQ120Q - How often drink alcohol over past 12 mos ALQ120U - # days drink alcohol per wk, mo, yr ALQ130 - Avg # alcoholic drinks/day - past 12 mos ALQ141Q - # days have 4/5 drinks - past 12 mos ALQ141U - # days per week, month, year? ALQ151 - Ever have 4/5 or more drinks every day? ALQ160 - # days have 4/5 or more drinks in 2 hrs ***

```
alcohol_clean <- alcohol %>%
  select(SEQN,ALQ120Q,ALQ120U, ALQ130, ALQ141Q, ALQ141U, ALQ151, ALQ160) %>%
  na.omit(cols=seq_along(ALQ120Q, ALQ120U, ALQ130, ALQ141Q, ALQ141U, ALQ151, ALQ160)) %>%
  as.data.frame()

#write.table(alcohol_clean, "./alchohol_clean.csv", sep=",")
```

Demographic: RIAGENDR - Gender RIDAGEYR - Age in years at screening DMQMILIZ - Served active duty in US Armed Forces DMDCITZN - Citizenship status DMDEDUC3 - Education level - Children/Youth 6-19 INDFMIN2 - Annual family income INDFMPIR - Ratio of family income to poverty DMDEDUC2 - Education level - Adults 20+ DMDMARTL - Marital status ***

```
demographc_clean <- demographic %>%
  select(SEQN, RIAGENDR, RIDAGEYR, DMQMILIZ, DMDCITZN, DMDEDUC3, INDFMIN2, INDFMPIR, DMDEDUC2, DMDMARTL) %
>%
  as.data.frame()
```

Depression: DPQ010 - Have little interest in doing things DPQ020 - Feeling down, depressed, or hopeless DPQ030 - Trouble sleeping or sleeping too much DPQ040 - Feeling tired or having little energy DPQ050 - Poor appetite or overeating DPQ060 - Feeling bad about yourself DPQ070 - Trouble concentrating on things DPQ080 - Moving or speaking slowly or too fast DPQ090 - Thought you would be better off dead DPQ100 - Difficulty these problems have caused ***

Quantify depression

We quantify depression as a sum of all the variables in the depression table except DPQ100 Since DPQ100 by nature and by definition is a multiplicative variable

```
depression_clean <- depression %>%
  na.omit(seq_along(DPQ010, DPQ020, DPQ030, DPQ040, DPQ050, DPQ060, DPQ070, DPQ080, DPQ090, DPQ100)) %>%
  as.data.frame()

cols_to_mutate <- c("DPQ010", "DPQ020", "DPQ030", "DPQ040", "DPQ050", "DPQ060", "DPQ070", "DPQ080", "DPQ09
0")

depression_clean %>%
  select(DPQ010:DPQ100) %>%
    mutate_at(cols_to_mutate, function(x) {
    case_when(
      x == 1 ~ 1,
      x == 2 ~ 2,
      x == 3 ~ 3,
      x == 0 ~ 0,
      T ~ as.numeric(NA)
    )
  }) %>%
  rowSums(na.rm=TRUE) -> depression_clean$DepressionScore


# We scraped out the multiplicative variable in our final analysis
normalizeDepressionScore <- function(score, factor) {
  case_when(
      factor == 0 ~ 1*score,
      factor == 1 ~ 2*score,
      factor == 2 ~ 4*score,
      factor == 3 ~ 8*score,
      T ~ as.numeric(NA)
  )
}

#\depression_clean$DepressionScore <- normalizeDepressionScore(depression_clean$DepressionScore,
depression_clean$DPQ100)
```

Join alcohol, demographic and drepression table for univariate analysis

According to CDC definitions, excess consumption of alcohol is defined differently for males and females. Hence, we need gender information to infer relationship between alcoholism and depression precisely. To keep things simple, we are joining demographic and depression table to alcohol table via inner join. The reason of choosing inner join is because: We would need all the information to conduct further analysis

```
depression_trunc <- depression_clean %>% select(SEQN,DepressionScore,DPQ090) %>% as.data.frame()

AlcoholAnalysis <- alcohol_clean %>%
                    inner_join(depression_trunc,
                      by="SEQN",
                      copy=False) %>%
                    inner_join(demographc_clean,
                      by="SEQN",
                      copy=False)
```

Quantify alcoholism

Alcohol in the USA

In the United States, a standard drink contains 0.6 ounces (14.0 grams or 1.2 tablespoons) of pure alcohol. Generally, this amount of pure alcohol is found in 12-ounces of beer (5% alcohol content). 8-ounces of malt liquor (7% alcohol content). 5-ounces of wine (12% alcohol content). 1.5-ounces of 80-proof (40% alcohol content) distilled spirits or liquor (e.g., gin, rum, vodka, whiskey).4 (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm))

1. Drinkers

Definition of Moderate Drinking: The Dietary Guidelines for Americans defines moderate drinking as up to 1 drink per day for women and up to 2 drinks per day for men. In addition, the Dietary Guidelines do not recommend that individuals who do not drink alcohol start drinking for any reason. (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm)) **Scheme of quantification:** Total Alcohol Consumption (TotalConsumption) = Drinks/Day * Drinking days TotalConsumption = Drinking Days[function(ALQ120Q, ALQ120U)] * Avg drinks/day [ALQ130]

2. Binge Drinkers

Definition of Binge Drinking: The National Institute on Alcohol Abuse and AlcoholismExternal defines binge drinking as a pattern of drinking that brings a person's blood alcohol concentration (BAC) to 0.08 grams percent or above. This typically happens when men consume 5 or more drinks or women consume 4 or more drinks in about 2 hours. (https://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm (https://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm)) In our analysis, we will consider consumption in throughout the day as well, **Scheme of quantification:** Degree of binge drinking= Number of bringe drinking sessions*Alcohol consumed in bringe drinking sessions

3. Heavy Drinkers:

Definition of Heavy Drinking: Heavy drinking is defined as consuming For women, 8 or more drinks per week. For men, 15 or more drinks per week. (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm (https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm)) **Scheme of quantification:** Degree of heavy drinking= Number of heavy drinking sessions * alcohol consumed in heavy drinking sessions

4. Recent Excess Consumption:

We consider the field "ALQ160" to establish this metric.

Note: Occasional drinkers or drinkers who do not fit into the above category are not conisdered in this analysis Heavy drinkers are out of scope as well

```
getTotalConsumption <- function(freq, unit, avg_drinksByDays){
      case_when(
          freq<375 | unit == 1 ~ freq*7*12*avg_drinksByDays,
          freq<375 | unit == 2 ~ freq*12*avg_drinksByDays,
          freq<375 | unit == 3 ~ freq*avg_drinksByDays,
          freq<375 | unit == 7 ~ as.numeric(NA),
          freq<375 | unit == 9 ~ as.numeric(NA),
          TRUE ~ as.numeric(NA)
          )
 }

AlcoholAnalysis$TotalConsumption <- mapply(getTotalConsumption,
                                           as.numeric(AlcoholAnalysis$ALQ120Q),
                                           as.numeric(AlcoholAnalysis$ALQ120U),
                                           as.numeric(AlcoholAnalysis$ALQ130))

bingeNumber <- function(gender) {
  case_when (
    gender == 1 ~ 5,
    gender == 2 ~ 4,
    TRUE ~ as.numeric(NA)
  )
}
getBingeConsumption <- function(freq, unit, bingenum){
      case_when(
          freq<375 | unit == 1 ~ freq*7*12*bingenum,
          freq<375 | unit == 2 ~ freq*12*bingenum,
          freq<375 | unit == 3 ~ freq*bingenum,
          freq<375 | unit == 7 ~ as.numeric(NA),
          freq<375 | unit == 9 ~ as.numeric(NA),
          TRUE ~ as.numeric(NA)
          )
}
AlcoholAnalysis$BingeConsumption <- mapply(getBingeConsumption,
                                           AlcoholAnalysis$ALQ141Q,
                                           AlcoholAnalysis$ALQ141U,
                                           bingeNumber(AlcoholAnalysis$RIAGENDR))

getRecentAddiction <- function(freq, bingenum) {
  case_when(
    freq <= 19 ~ freq*bingenum,
    freq == 20 ~ 20*bingenum,
    TRUE ~ as.numeric(NA)
  )
}

AlcoholAnalysis$RecentAddiction <- mapply(getRecentAddiction,
                                          AlcoholAnalysis$ALQ160,
                                          bingeNumber(AlcoholAnalysis$RIAGENDR))
```

Drop unncessary alcohol columns

```
AlcoholAnalysis <- within(AlcoholAnalysis, rm(ALQ120Q,ALQ120U,ALQ130,ALQ141Q,ALQ141U,ALQ151,ALQ160))
```

```
write.csv(AlcoholAnalysis, file = "AlcoholAnalysis.csv")
```

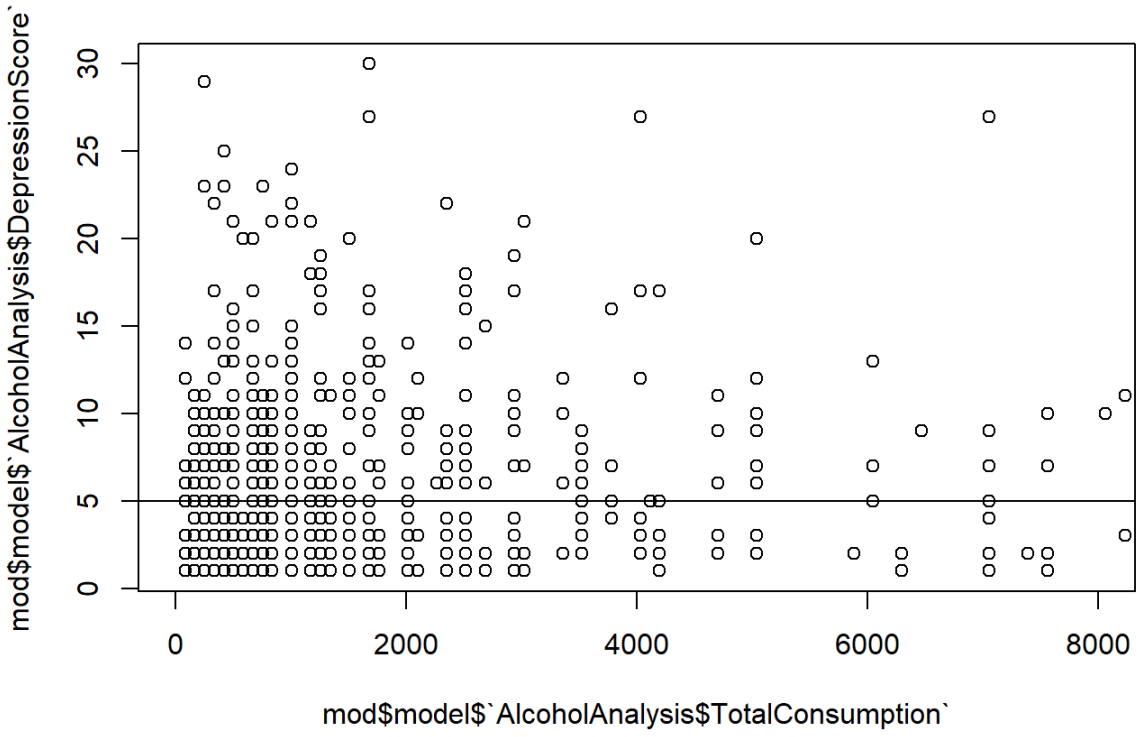
Univariate Linear Regression

Depression Score ~ Total Alcohol Consumption

```
mod <- lm(AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$TotalConsumption, data = AlcoholAnalysis)
#mod <- glm(DPQ090 ~ TotalConsumption,  data = AlcoholAnalysis ) #, family="binomial")
summary(mod)
```

```
##
## Call:
## lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$TotalConsumption,
##     data = AlcoholAnalysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.026 -2.990 -1.983  2.002 25.009
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.981e+00  1.539e-01  32.359   <2e-16 ***
## AlcoholAnalysis$TotalConsumption 5.960e-06  7.833e-06   0.761    0.447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.838 on 998 degrees of freedom
## Multiple R-squared:  0.0005798,  Adjusted R-squared:  -0.0004216
## F-statistic: 0.579 on 1 and 998 DF,  p-value: 0.4469
```

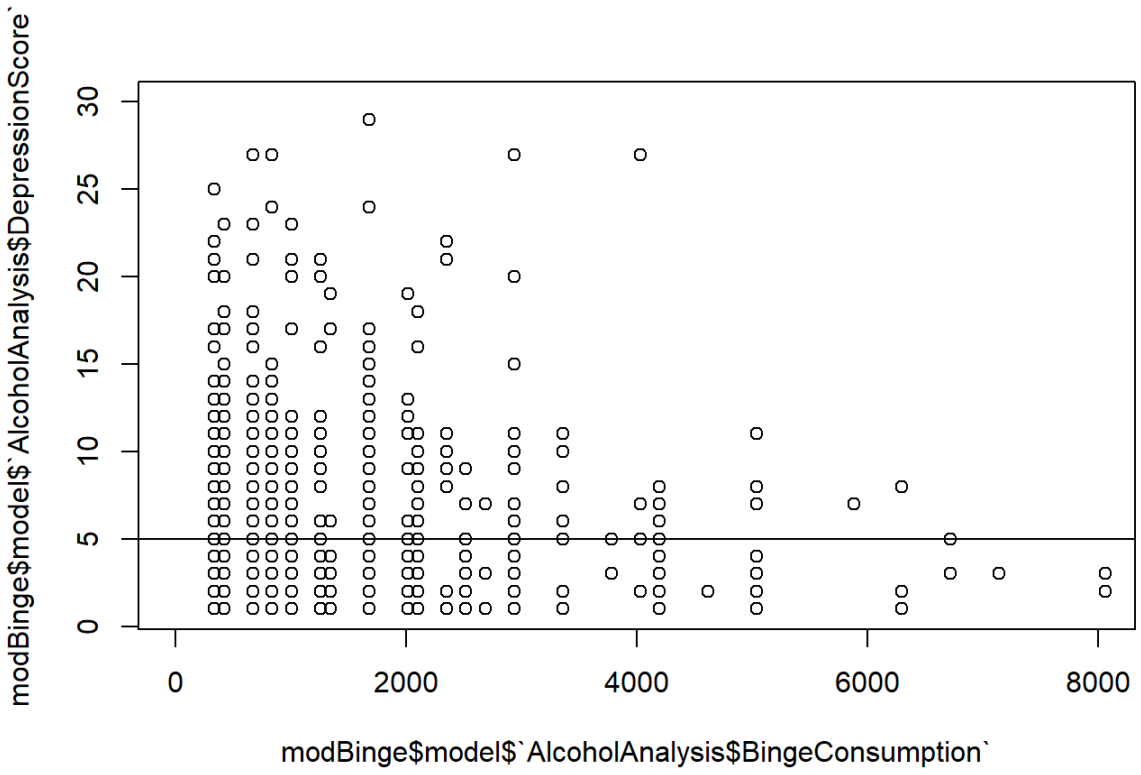
```
plot(mod$model$`AlcoholAnalysis$TotalConsumption`,mod$model$`AlcoholAnalysis$DepressionScore`, xlim=c(0,800
0)) # the limit can be set between (5000 and 20000)
abline(mod)
```




```
modBinge <- lm(AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$BingeConsumption, data = AlcoholAnalysis)
summary(modBinge)
```

```
##
## Call:
## lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$BingeConsumption,
##     data = AlcoholAnalysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.061 -2.994 -1.992  2.006 24.999
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.992e+00  1.576e-01   31.68   <2e-16 ***
## AlcoholAnalysis$BingeConsumption 1.058e-06  1.749e-05    0.06    0.952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.839 on 998 degrees of freedom
## Multiple R-squared:  3.663e-06,  Adjusted R-squared:  -0.0009983
## F-statistic: 0.003656 on 1 and 998 DF,  p-value: 0.9518
```

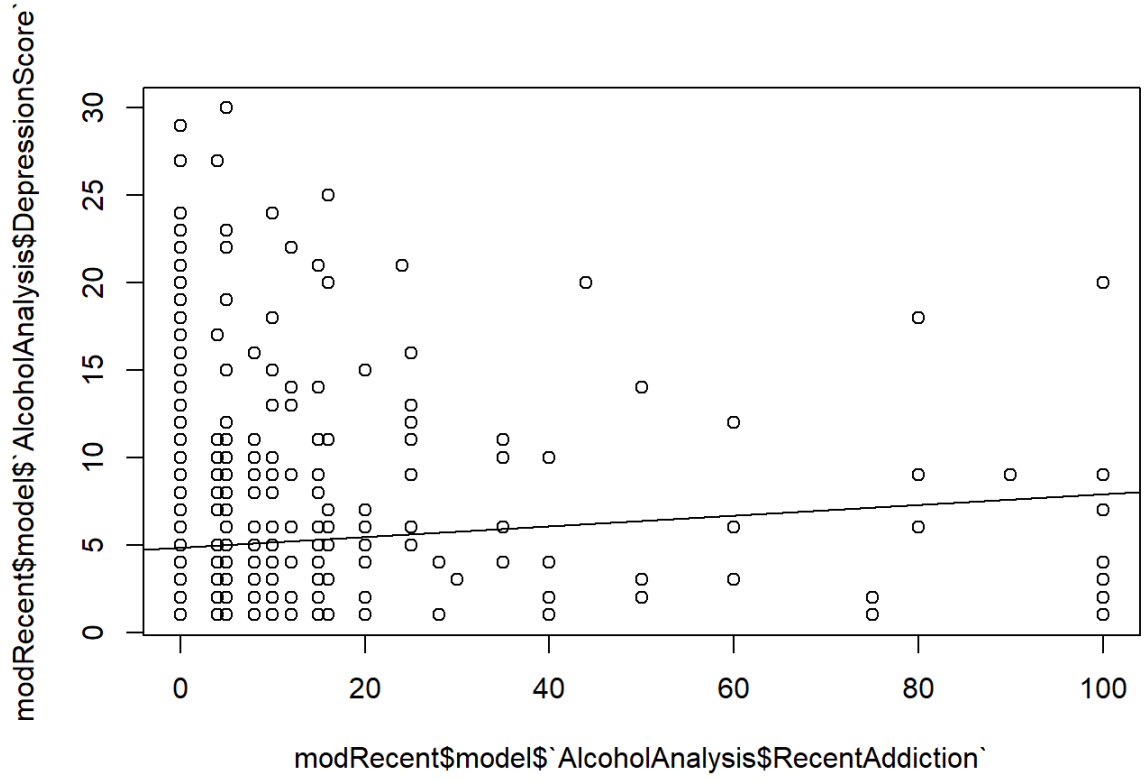
```
plot(modBinge$model$`AlcoholAnalysis$BingeConsumption`,modBinge$model$`AlcoholAnalysis$DepressionScore`, xl
im=c(0,8000)) # the limit can be set between (5000 and 20000)
abline(modBinge)
```



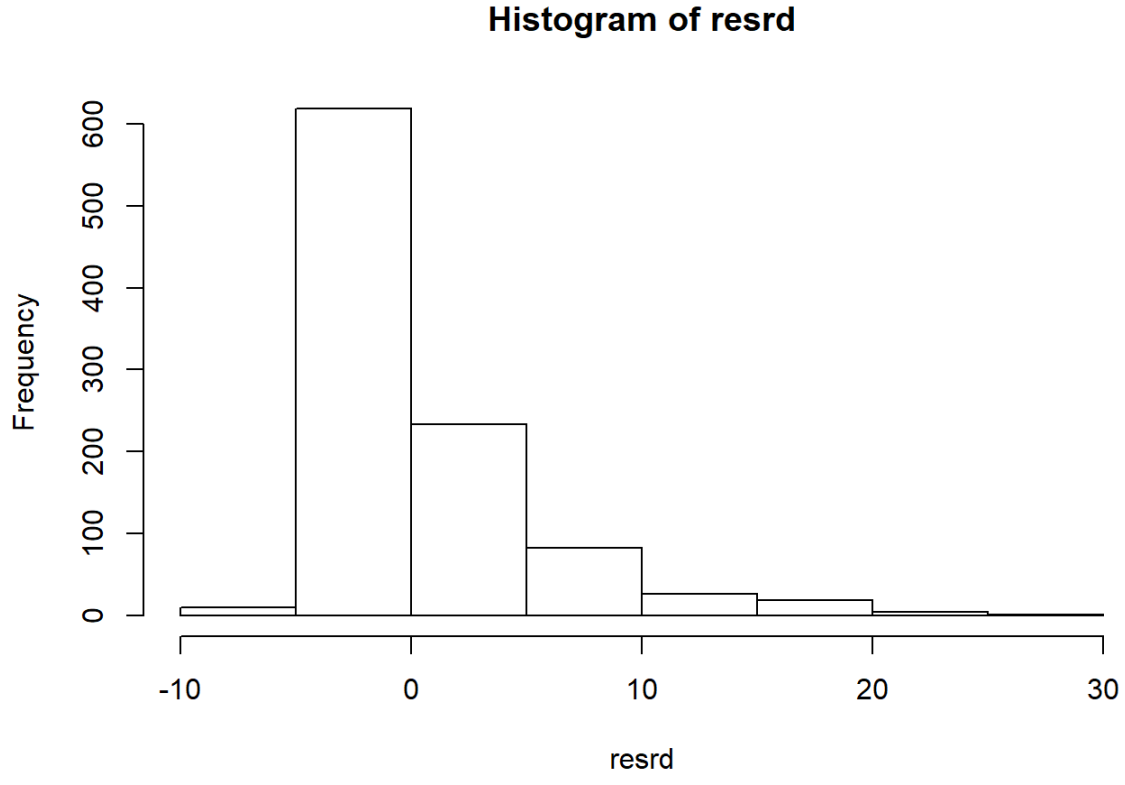

```
modRecent <- lm(AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$RecentAddiction, data = AlcoholAnalysis)
summary(modRecent)
```

```
##
## Call:
## lm(formula = AlcoholAnalysis$DepressionScore ~ AlcoholAnalysis$RecentAddiction,
##     data = AlcoholAnalysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.930 -3.092 -1.845  1.668 25.000
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.84546    0.16204  29.902  < 2e-16 ***
## AlcoholAnalysis$RecentAddiction  0.03084    0.01119   2.755  0.00597 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.827 on 994 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.007579,   Adjusted R-squared:  0.006581
## F-statistic: 7.591 on 1 and 994 DF,  p-value: 0.005973
```

```
plot(modRecent$model$`AlcoholAnalysis$RecentAddiction`,modRecent$model$`AlcoholAnalysis$DepressionScore`, x
lim=c(0,100)) # the limit can be set between (5000 and 20000)
abline(modRecent)
```



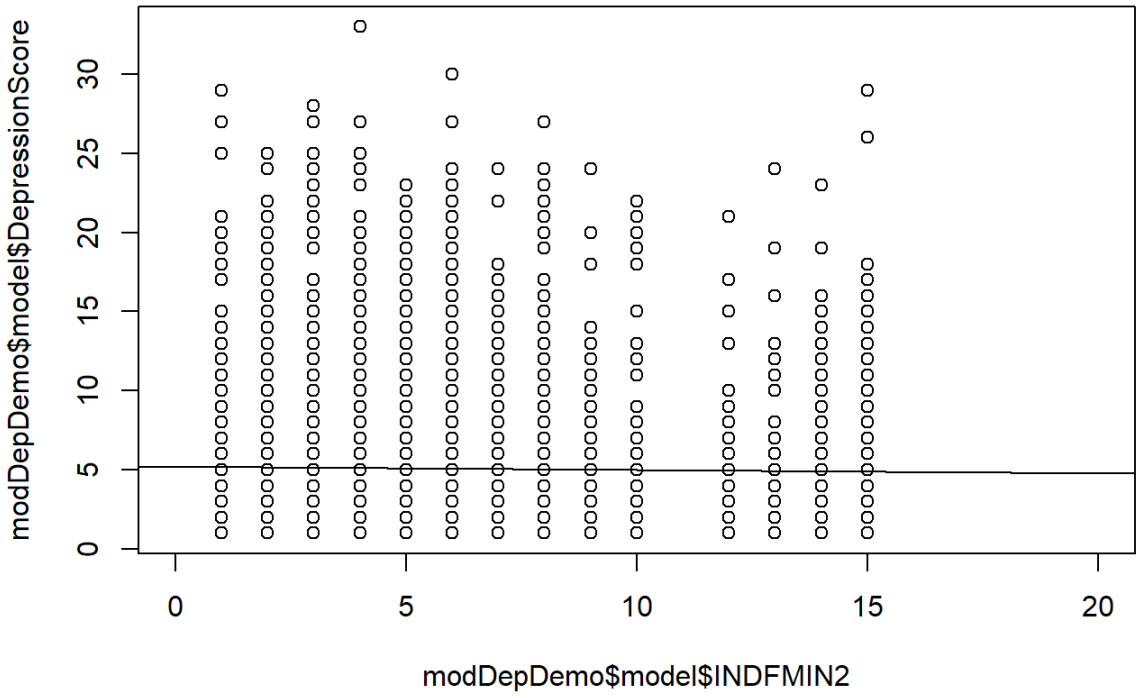

```
resrd <- resid(modRecent)
hist(resrd)
```



Correlation of

Depression score against demographics - age, gender, income

```
#hist(AlcoholAnalysis$DepressionScore)
DepressionDemographics <- demographc_clean %>%
                    inner_join(depression_trunc,
                      by="SEQN",
                      copy=False)
#head(DepressionDemographics)
modDepDemo <- lm(DepressionScore ~ INDFMIN2, data = DepressionDemographics)
plot(modDepDemo$model$INDFMIN2,modDepDemo$model$DepressionScore, xlim=c(0,20))
abline(modDepDemo)
```




```
summary(modDepDemo)
```

```
##
## Call:
## lm(formula = DepressionScore ~ INDFMIN2, data = DepressionDemographics)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.191 -3.089 -1.903  1.829 27.870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.212042   0.099092  52.598  < 2e-16 ***
## INDFMIN2    -0.020569   0.005558  -3.701 0.000218 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 3485 degrees of freedom
##   (92 observations deleted due to missingness)
## Multiple R-squared:  0.003915,   Adjusted R-squared:  0.003629
## F-statistic:  13.7 on 1 and 3485 DF,  p-value: 0.0002182
```

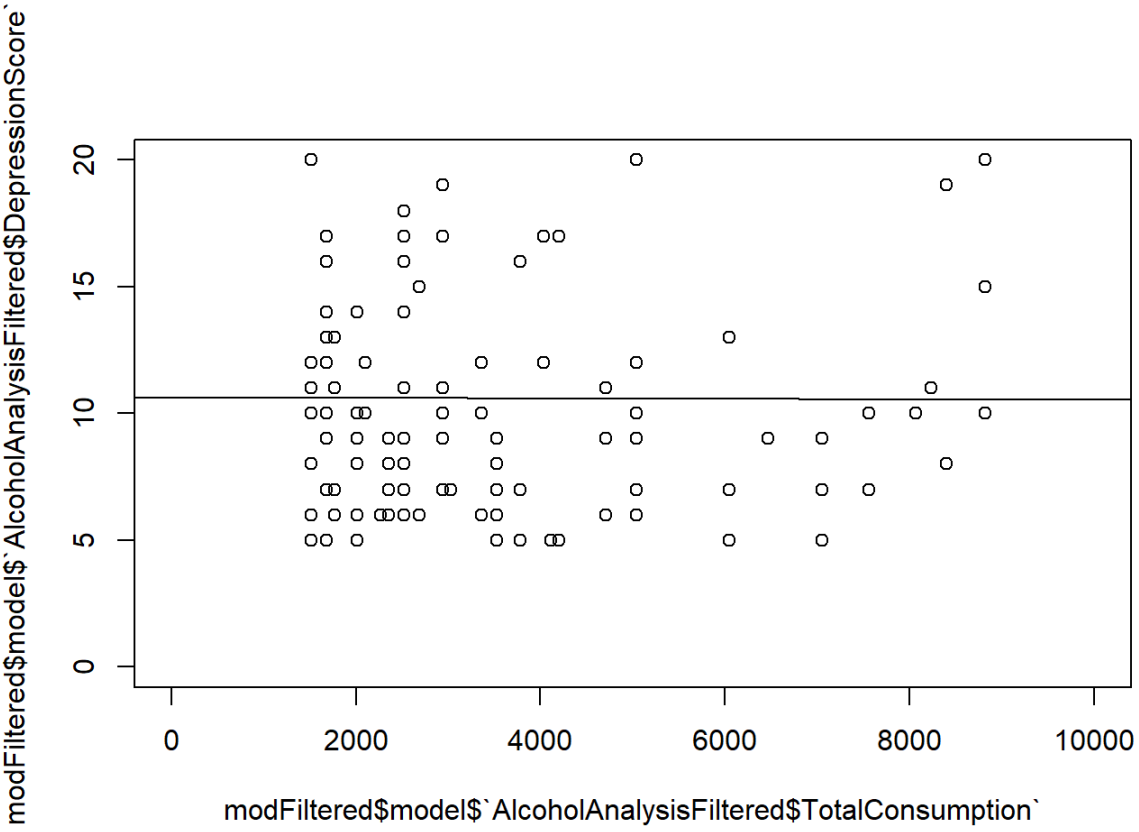
```
ServerelyDepressed <- AlcoholAnalysis %>% filter(TotalConsumption >=1500) %>% filter(DepressionScore >= 5)
AlcoholAnalysisFiltered <- AlcoholAnalysis %>% inner_join(ServerelyDepressed)
```

```
## Joining, by = c("SEQN", "DepressionScore", "DPQ090", "RIAGENDR", "RIDAGEYR", "DMQMILIZ", "DMDCITZN", "DM
DEDUC3", "INDFMIN2", "INDFMPIR", "DMDEDUC2", "DMDMARTL", "TotalConsumption", "BingeConsumption", "RecentAdd
iction")
```

```
modFiltered <- lm(AlcoholAnalysisFiltered$DepressionScore ~ AlcoholAnalysisFiltered$TotalConsumption, data
= AlcoholAnalysis)
#mod <- glm(DPQ090 ~ TotalConsumption,  data = AlcoholAnalysis ) #, family="binomial")
summary(modFiltered)
```

```
##
## Call:
## lm(formula = AlcoholAnalysisFiltered$DepressionScore ~ AlcoholAnalysisFiltered$TotalConsumption,
##     data = AlcoholAnalysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.629 -3.869 -1.619  1.646 19.373
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                              1.064e+01  5.006e-01  21.259
## AlcoholAnalysisFiltered$TotalConsumption -9.364e-06  9.297e-06  -1.007
##                                          Pr(>|t|)
## (Intercept)                                <2e-16 ***
## AlcoholAnalysisFiltered$TotalConsumption    0.316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.66 on 130 degrees of freedom
## Multiple R-squared:  0.007744,   Adjusted R-squared:  0.0001114
## F-statistic: 1.015 on 1 and 130 DF,  p-value: 0.3157
```

```
plot(modFiltered$model$`AlcoholAnalysisFiltered$TotalConsumption`,modFiltered$model$`AlcoholAnalysisFiltere
d$DepressionScore`, ylim=c(0,20),xlim=c(0,10000)) # the limit can be set between (5000 and 20000)
abline(modFiltered)
```



Drug Use & Logistic Model

Roxine Jian-Sin Lee

3/15/2019

Load the libraries and datasets.

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(SASxport)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
depression <- read.csv("Depression_1516.csv")
drugUse <- read.csv("Drug Use_1516.csv")
demographic <- read.csv("Demographic_1516.csv")
```

Quantify the depression tendency.

```
depression_clean <- depression %>%
  na.omit(seq_along(DPQ010, DPQ020, DPQ030, DPQ040, DPQ050, DPQ060, DPQ070, DPQ080, DPQ090, DPQ100)) %>%
  as.data.frame()
```

```
cols_to_mutate <- c("DPQ010", "DPQ020", "DPQ030", "DPQ040", "DPQ050", "DPQ060", "DPQ070", "DPQ080", "DPQ
```

```
depression_clean %>%
  select(DPQ010:DPQ090) %>%
  mutate_at(cols_to_mutate, function(x) {
    case_when(
      x == 1 ~ 1,
      x == 2 ~ 2,
      x == 3 ~ 3,
      x == 0 ~ 0,
      T ~ as.numeric(NA)
    )
  }) %>%
  rowSums(na.rm=TRUE) -> depression_clean$DepressionScore
```

Make the depression tendency "binary" (a new column created).

```
depression_clean$depressed[depression_clean$DepressionScore >= 10] <- 1
depression_clean$depressed[depression_clean$DepressionScore < 10] <- 0
```

Only keep the data of non-military U.S. citizens.

```
demo_c <- c("SEQN", "DMQMILIZ", "DMDCITZN")
demo_nmc <- demographic[, demo_c]
depression_nmc <- merge(x = depression_clean, y = demo_nmc, by = "SEQN", all.x = TRUE)

drug_c <- c("SEQN", "DUQ213", "DUQ217", "DUQ219", "DUQ230", "DUQ250", "DUQ280", "DUQ290", "DUQ320", "DU
drug_key <- drugUse[, drug_c]
depression_haro <- merge(x = depression_nmc, y = drug_key, by = "SEQN", all.x = TRUE)
```

```
depression_fn <- depression_haro[depression_haro$DMQMILIZ == 2, ]
depression_fn <- depression_fn[depression_fn$DMDCITZN == 1, ]
```

Remove null values.

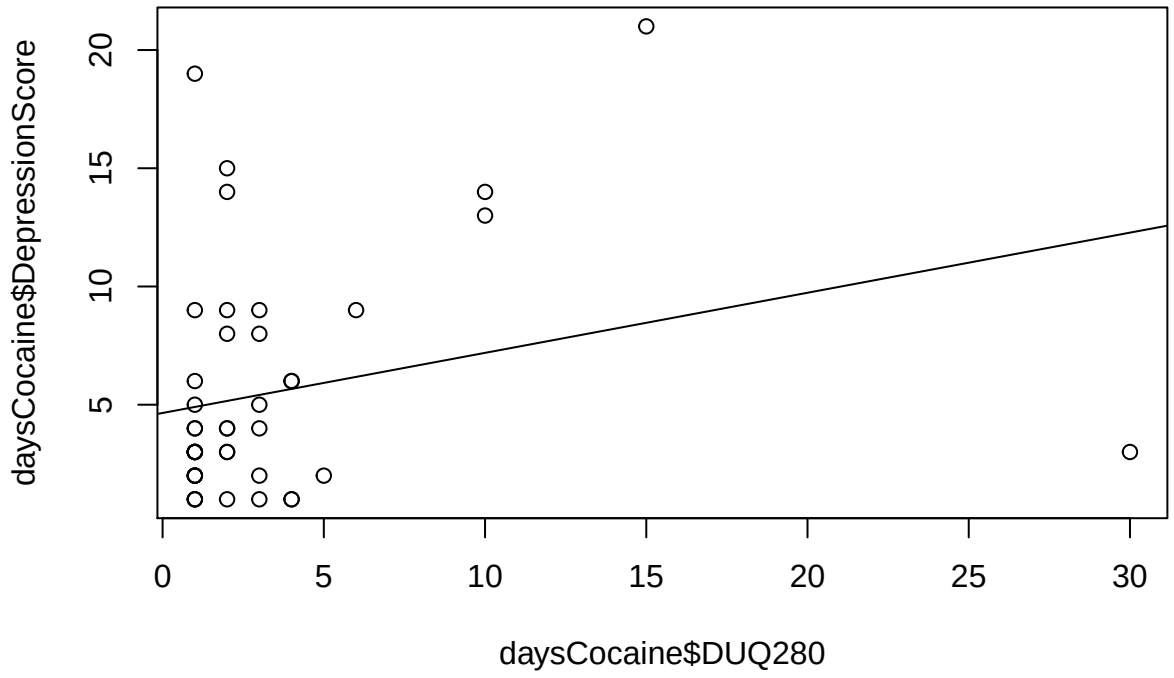
```
daysCocaine <- depression_fn[complete.cases(depression_fn[ , "DUQ280"]), ]
daysHeroin <- depression_fn[complete.cases(depression_fn[ , "DUQ320"]), ]
daysMeth <- depression_fn[complete.cases(depression_fn[ , "DUQ360"]), ]
startMJ <- depression_fn[complete.cases(depression_fn[ , "DUQ213"]), ]
oftenMJ <- depression_fn[complete.cases(depression_fn[ , "DUQ217"]), ]
pipesMJ <- depression_fn[complete.cases(depression_fn[ , "DUQ219"]), ]
daysMJ <- depression_fn[complete.cases(depression_fn[ , "DUQ230"]), ]
```

Conduct simple linear regression for the four drugs (marijuana, cocaine, heroin, and meth) and the other two supplemental variables regarding injecting illegal drug and rehabilitation programs.

```
modCocaine <- lm(daysCocaine$DepressionScore ~ daysCocaine$DUQ280, data = daysCocaine)
summary(modCocaine)
```

```
##
## Call:
## lm(formula = daysCocaine$DepressionScore ~ daysCocaine$DUQ280,
##     data = daysCocaine)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.278 -2.905 -1.659  2.646 14.095
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.6506     0.8843   5.259 4.58e-06 ***
## daysCocaine$DUQ280    0.2542     0.1486   1.711   0.0944 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.855 on 42 degrees of freedom
## Multiple R-squared:  0.06518,    Adjusted R-squared:  0.04292
## F-statistic: 2.928 on 1 and 42 DF,  p-value: 0.09441
```

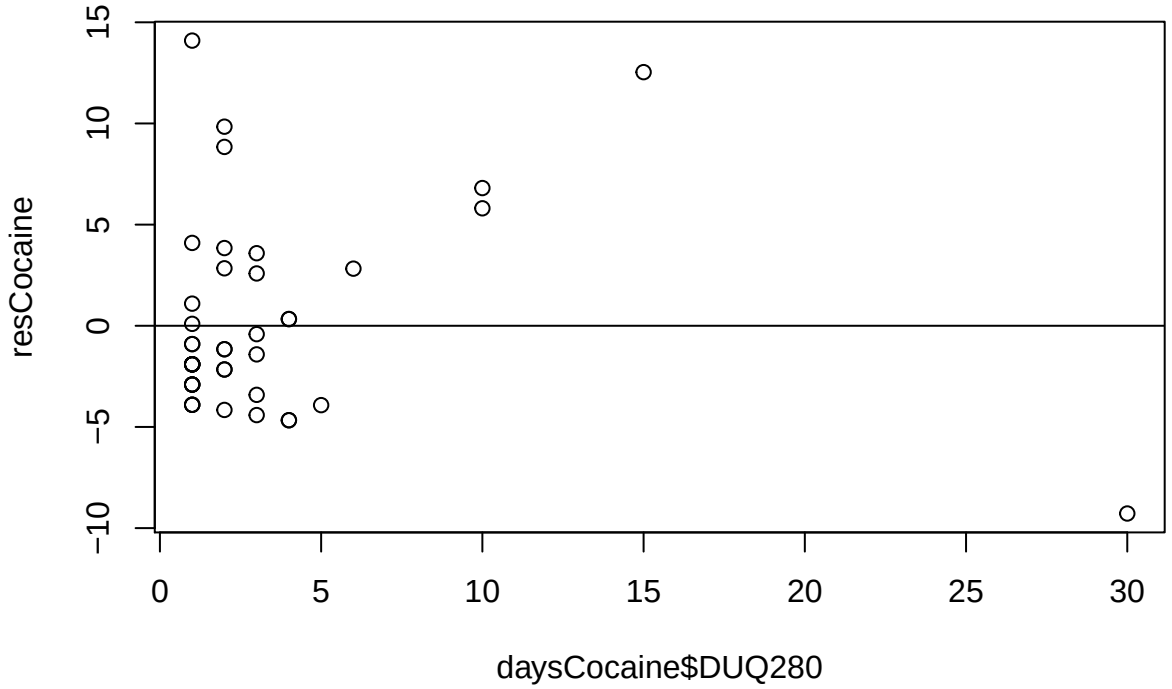
```
plot(daysCocaine$DUQ280, daysCocaine$DepressionScore)
abline(modCocaine)
```



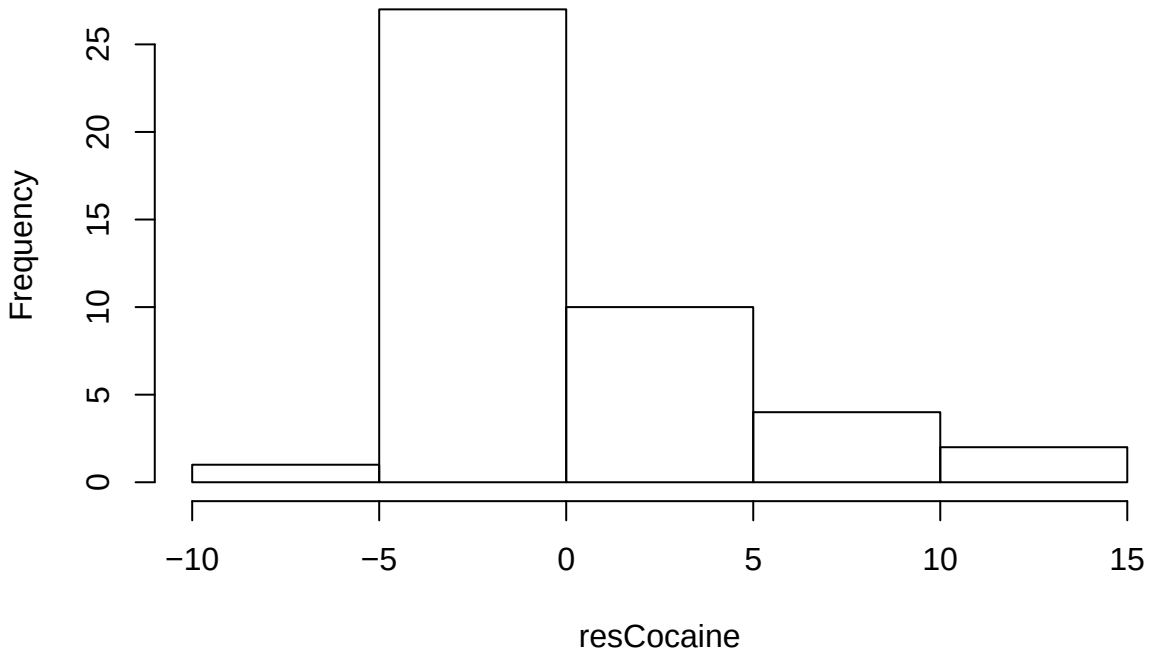

```
resCocaine <- resid(modCocaine)
shapiro.test(resCocaine)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resCocaine
## W = 0.8839, p-value = 0.0003612
```

```r
plot(daysCocaine$DUQ280, resCocaine)
abline(0, 0)
```



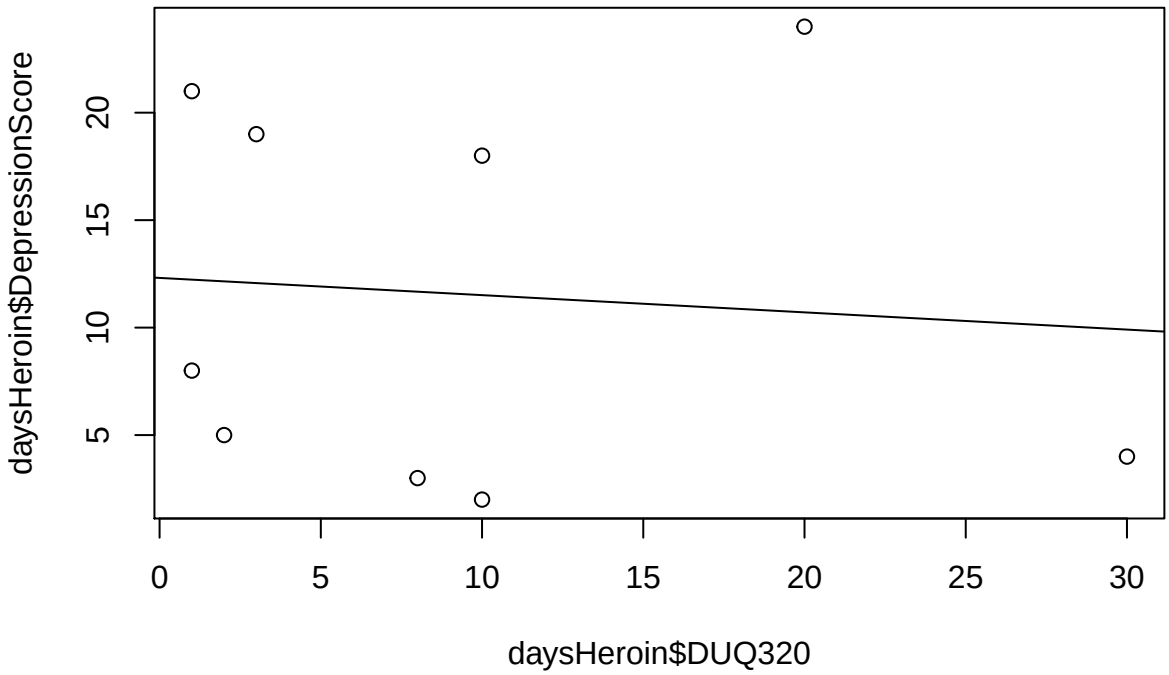

```r
hist(resCocaine)
```




```
modHeroin <- lm(daysHeroin$DepressionScore ~ daysHeroin$DUQ320, data = daysHeroin)
summary(modHeroin)
```

```
##
## Call:
## lm(formula = daysHeroin$DepressionScore ~ daysHeroin$DUQ320,
##     data = daysHeroin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.511 -7.152 -4.232  6.928 13.291
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.31263    4.44981   2.767   0.0278 *
## daysHeroin$DUQ320 -0.08016    0.33595  -0.239   0.8182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.36 on 7 degrees of freedom
## Multiple R-squared:  0.008068,   Adjusted R-squared:  -0.1336
## F-statistic: 0.05693 on 1 and 7 DF,  p-value: 0.8182
```

```
plot(daysHeroin$DUQ320, daysHeroin$DepressionScore)
abline(modHeroin)
```



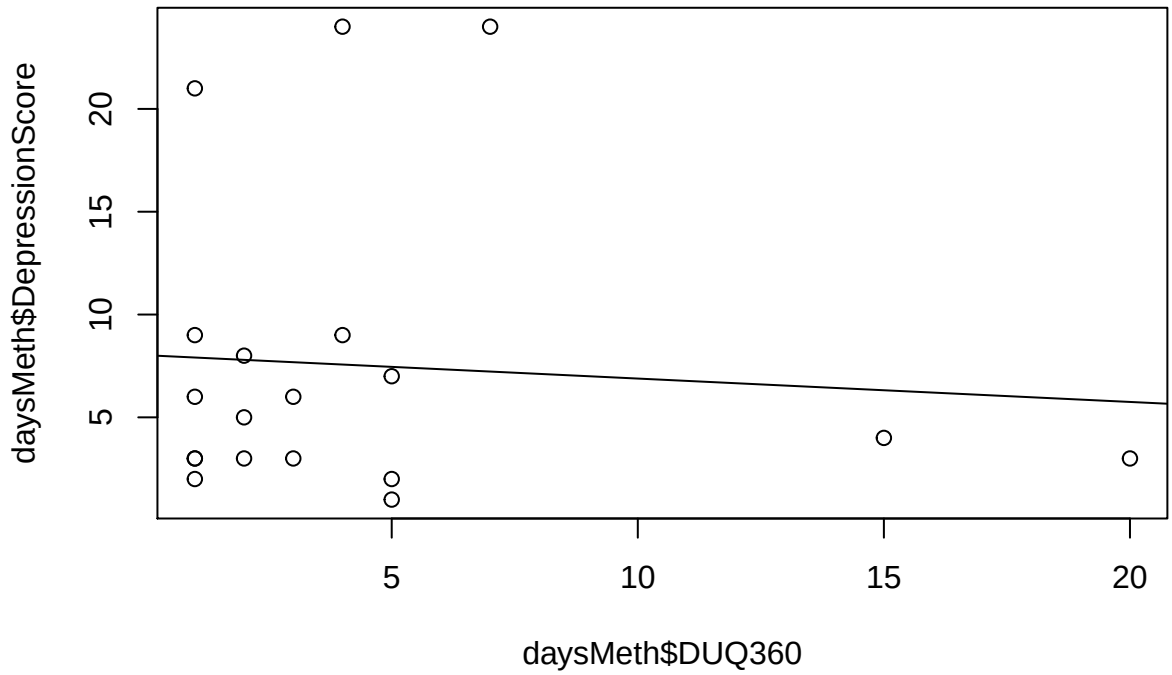

```
resHeroin <- resid(modHeroin)
shapiro.test(resHeroin)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resHeroin
## W = 0.86616, p-value = 0.1118
```

```
modMeth <- lm(daysMeth$DepressionScore ~ daysMeth$DUQ360, data = daysMeth)
summary(modMeth)
```

```
##
## Call:
## lm(formula = daysMeth$DepressionScore ~ daysMeth$DUQ360, data = daysMeth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4545 -4.8526 -2.3171  0.6474 16.7730
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.0232     2.3030   3.484  0.00284 **
## daysMeth$DUQ360  -0.1137     0.3512  -0.324  0.75000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.487 on 17 degrees of freedom
## Multiple R-squared:  0.006131,   Adjusted R-squared:  -0.05233
## F-statistic: 0.1049 on 1 and 17 DF,  p-value: 0.75
```

```
plot(daysMeth$DUQ360, daysMeth$DepressionScore)
abline(modMeth)
```



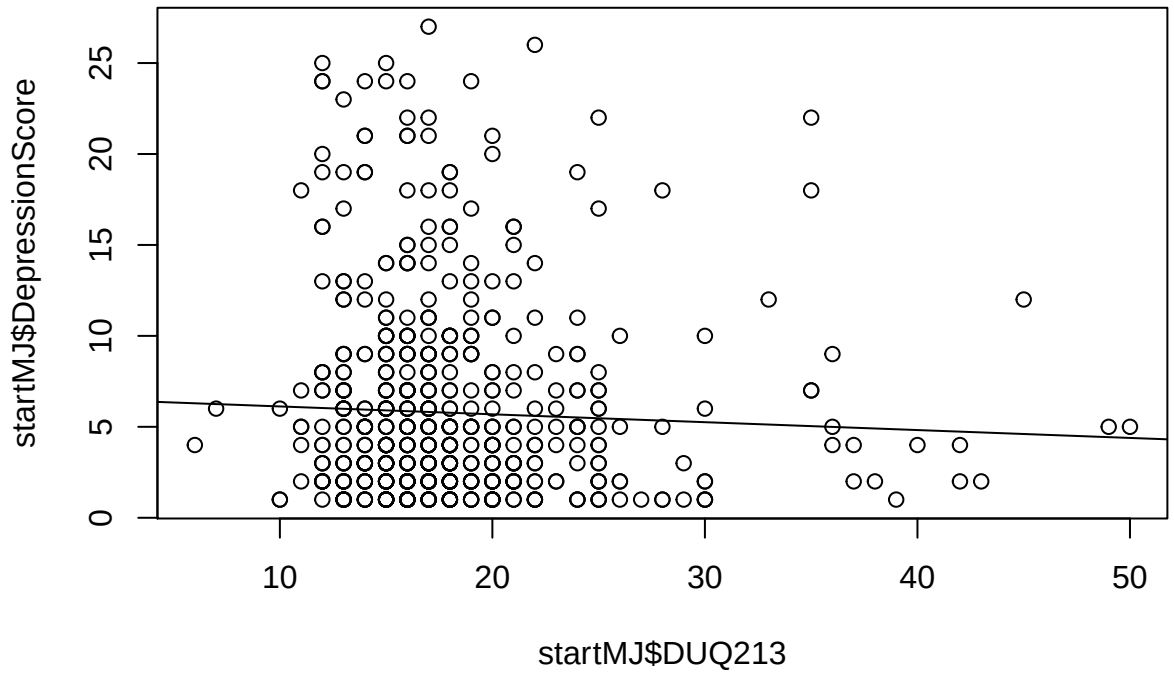

```
resMeth <- resid(modMeth)
shapiro.test(resMeth)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resMeth
## W = 0.73972, p-value = 0.0001721
```

```
modStartMJ <- lm(startMJ$DepressionScore ~ startMJ$DUQ213, data = startMJ)
summary(modStartMJ)
```

```
##
## Call:
## lm(formula = startMJ$DepressionScore ~ startMJ$DUQ213, data = startMJ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.121 -3.775 -1.775  1.961 21.182
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.55433    0.73929   8.866   <2e-16 ***
## startMJ$DUQ213 -0.04329    0.03945  -1.097    0.273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.25 on 628 degrees of freedom
## Multiple R-squared:  0.001914,   Adjusted R-squared:  0.0003246
## F-statistic: 1.204 on 1 and 628 DF,  p-value: 0.2729
```

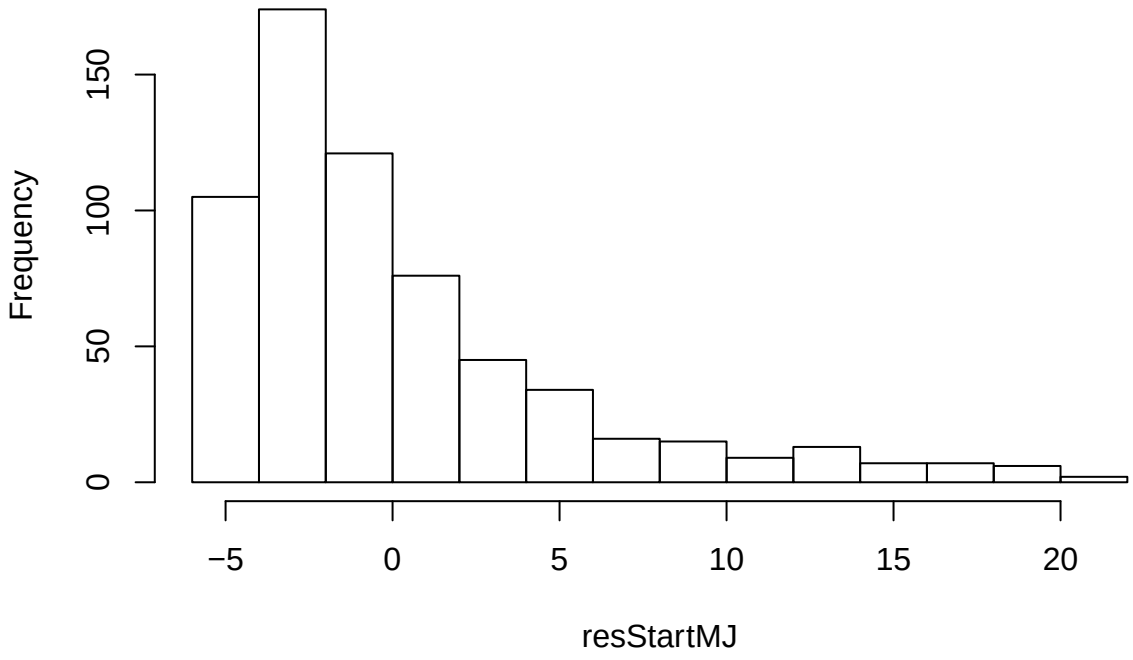
```
plot(startMJ$DUQ213, startMJ$DepressionScore)
abline(modStartMJ)
```




```
resStartMJ <- resid(modStartMJ)
shapiro.test(resStartMJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resStartMJ
## W = 0.80961, p-value < 2.2e-16
```

```
hist(resStartMJ)
```

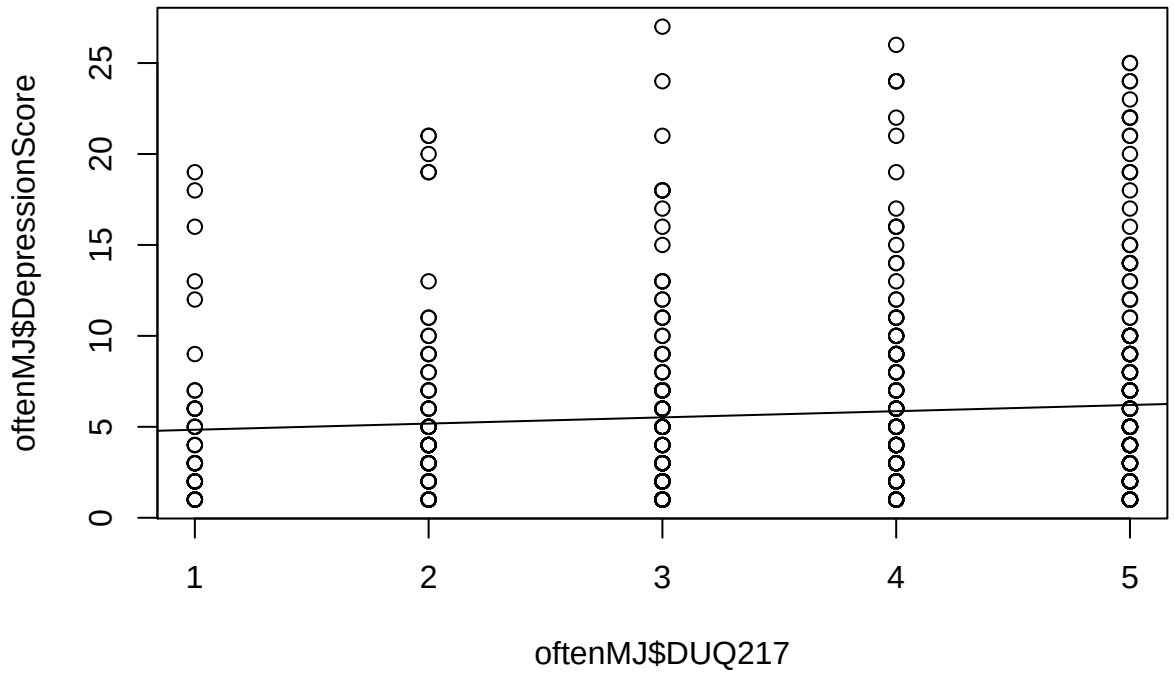



```
oftenMJ <- oftenMJ[oftenMJ$DUQ217 != "7", ]
oftenMJ <- oftenMJ[oftenMJ$DUQ217 != "9", ]

modOftenMJ <- lm(oftenMJ$DepressionScore ~ oftenMJ$DUQ217, data = oftenMJ)
summary(modOftenMJ)
```

```
##
## Call:
## lm(formula = oftenMJ$DepressionScore ~ oftenMJ$DUQ217, data = oftenMJ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.199 -3.838 -1.678  1.801 21.481
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.4977     0.6545   6.872 1.53e-11 ***
## oftenMJ$DUQ217   0.3403     0.1681   2.024   0.0434 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.227 on 626 degrees of freedom
## Multiple R-squared:  0.0065, Adjusted R-squared:  0.004913
## F-statistic: 4.096 on 1 and 626 DF,  p-value: 0.04342
```

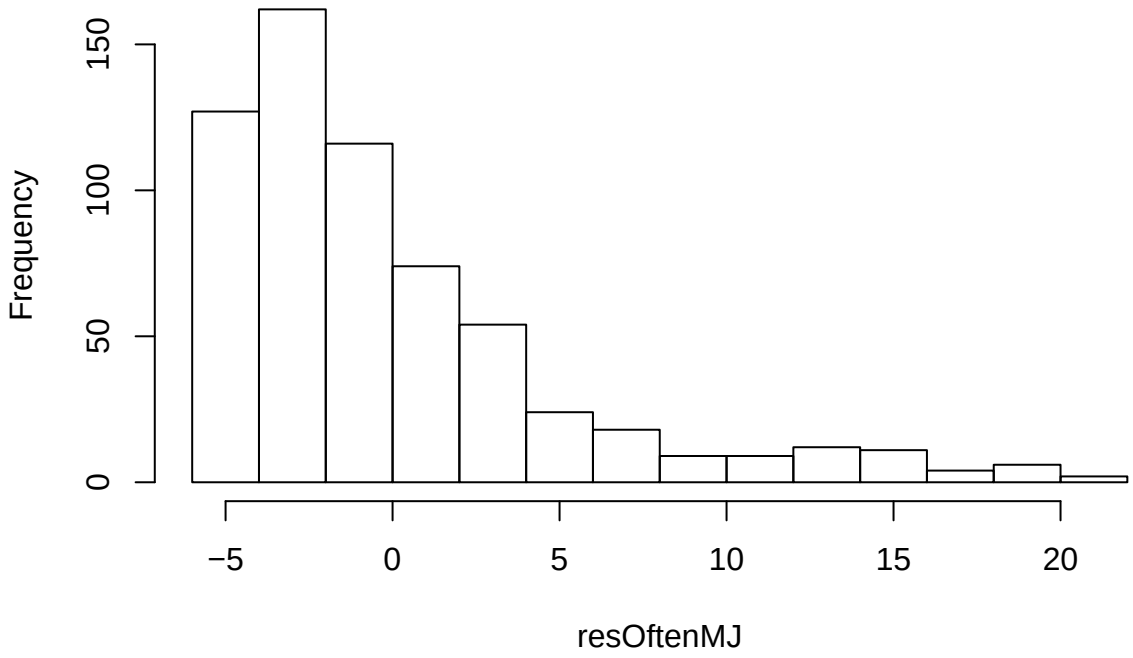
```
plot(oftenMJ$DUQ217, oftenMJ$DepressionScore)
abline(modOftenMJ)
```




```
resOftenMJ <- resid(modOftenMJ)
shapiro.test(resOftenMJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resOftenMJ
## W = 0.81438, p-value < 2.2e-16
```

```
hist(resOftenMJ)
```

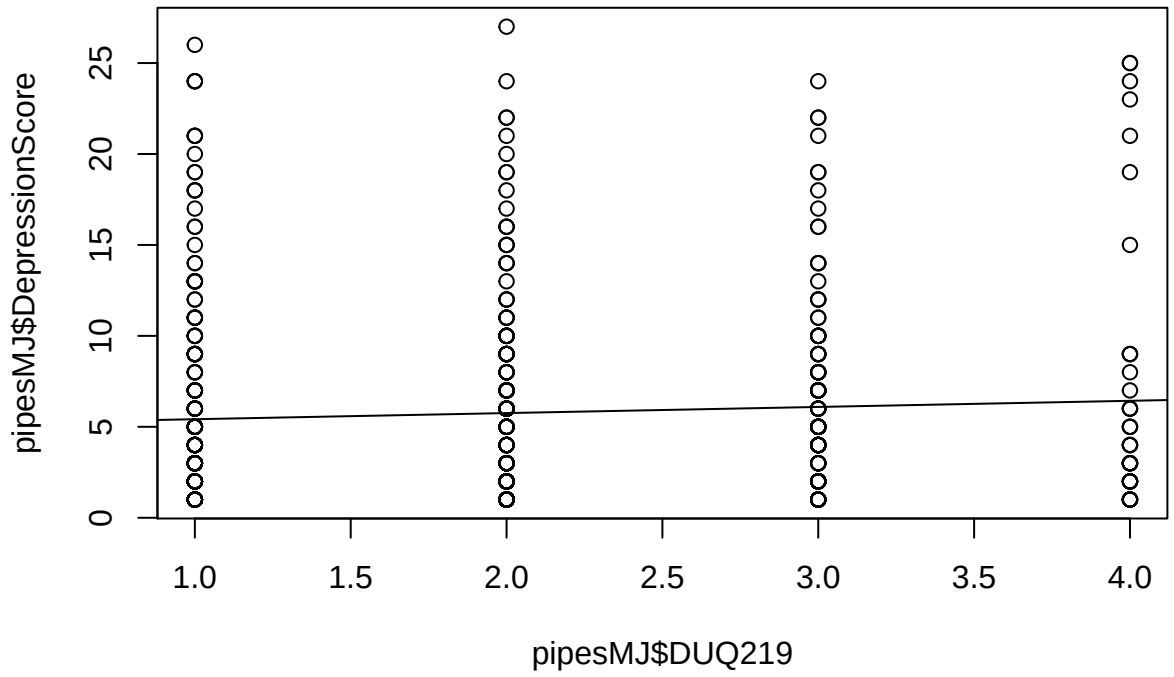



```
pipesMJ <- pipesMJ[pipesMJ$DUQ219 != "9", ]

modPipesMJ <- lm(pipesMJ$DepressionScore ~ pipesMJ$DUQ219, data = pipesMJ)
summary(modPipesMJ)
```

```
##
## Call:
## lm(formula = pipesMJ$DepressionScore ~ pipesMJ$DUQ219, data = pipesMJ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.431 -3.755 -1.755  1.583 21.245
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.0796     0.4964  10.234   <2e-16 ***
## pipesMJ$DUQ219   0.3378     0.2247   1.503    0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.227 on 627 degrees of freedom
## Multiple R-squared:  0.003591,   Adjusted R-squared:  0.002001
## F-statistic: 2.259 on 1 and 627 DF,  p-value: 0.1333
```

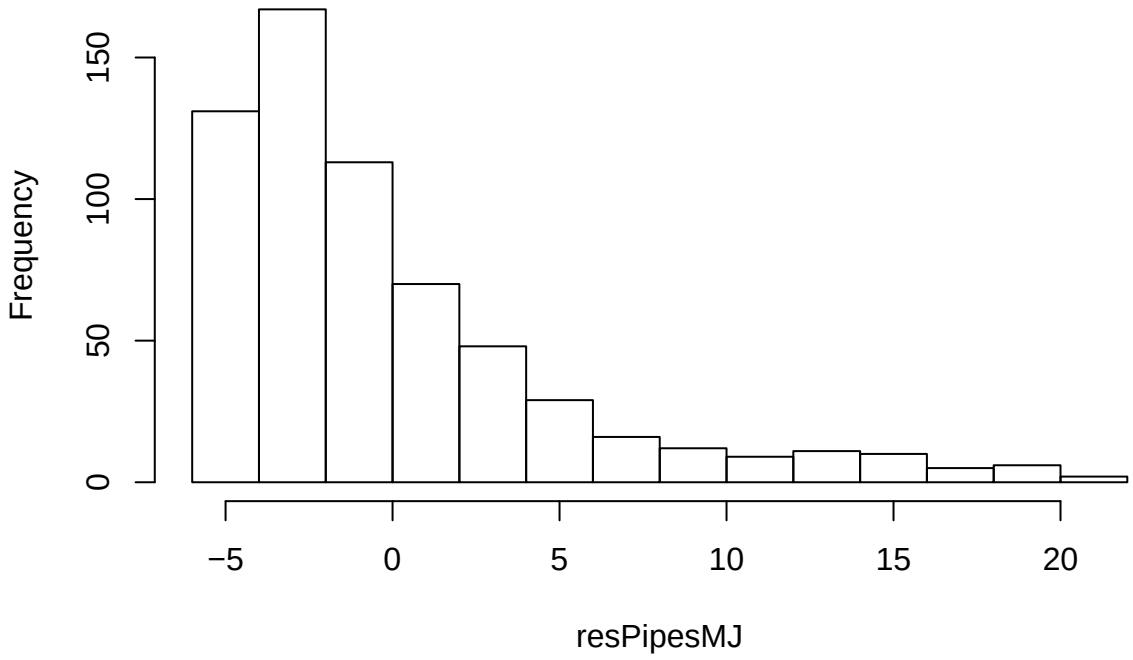
```
plot(pipesMJ$DUQ219, pipesMJ$DepressionScore)
abline(modPipesMJ)
```




```
resPipesMJ <- resid(modPipesMJ)
shapiro.test(resPipesMJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resPipesMJ
## W = 0.81176, p-value < 2.2e-16
```

```
hist(resPipesMJ)
```

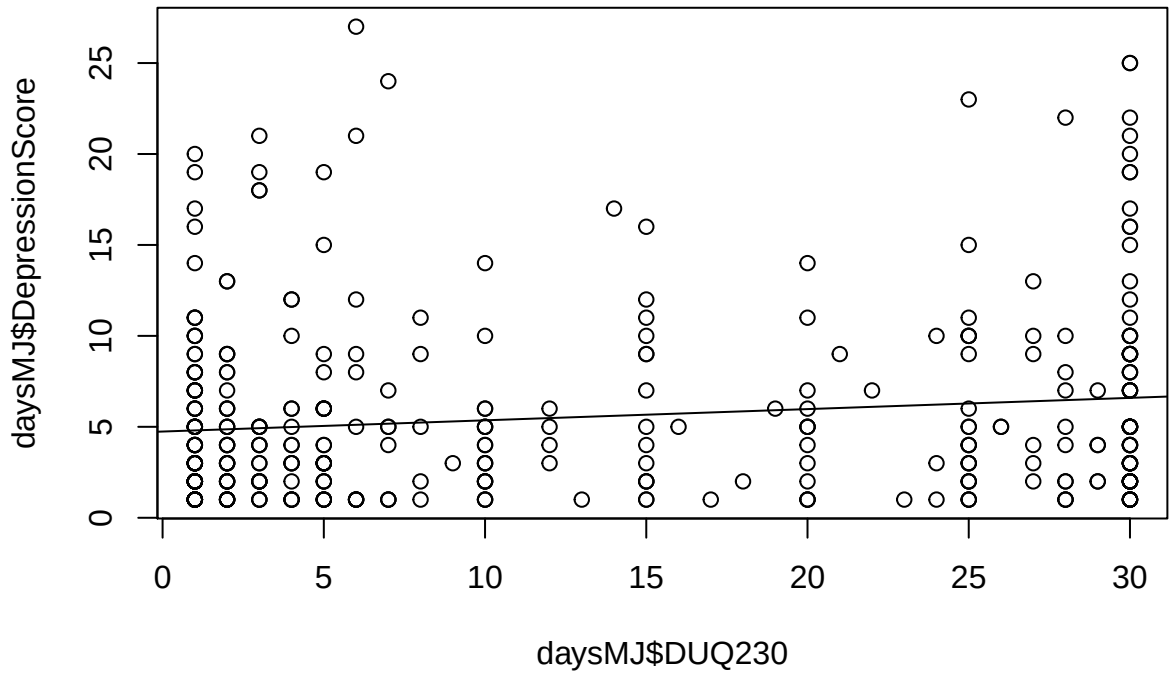



```
daysMJ <- daysMJ[daysMJ$DUQ230 != "999", ]

modDaysMJ <- lm(daysMJ$DepressionScore ~ daysMJ$DUQ230, data = daysMJ)
summary(modDaysMJ)
```

```
##
## Call:
## lm(formula = daysMJ$DepressionScore ~ daysMJ$DUQ230, data = daysMJ)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.591 -3.591 -1.591  2.152 21.890
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.74034    0.37582  12.613  < 2e-16 ***
## daysMJ$DUQ230  0.06170    0.02157   2.861  0.00445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.026 on 390 degrees of freedom
## Multiple R-squared:  0.02056,    Adjusted R-squared:  0.01804
## F-statistic: 8.185 on 1 and 390 DF,  p-value: 0.004451
```

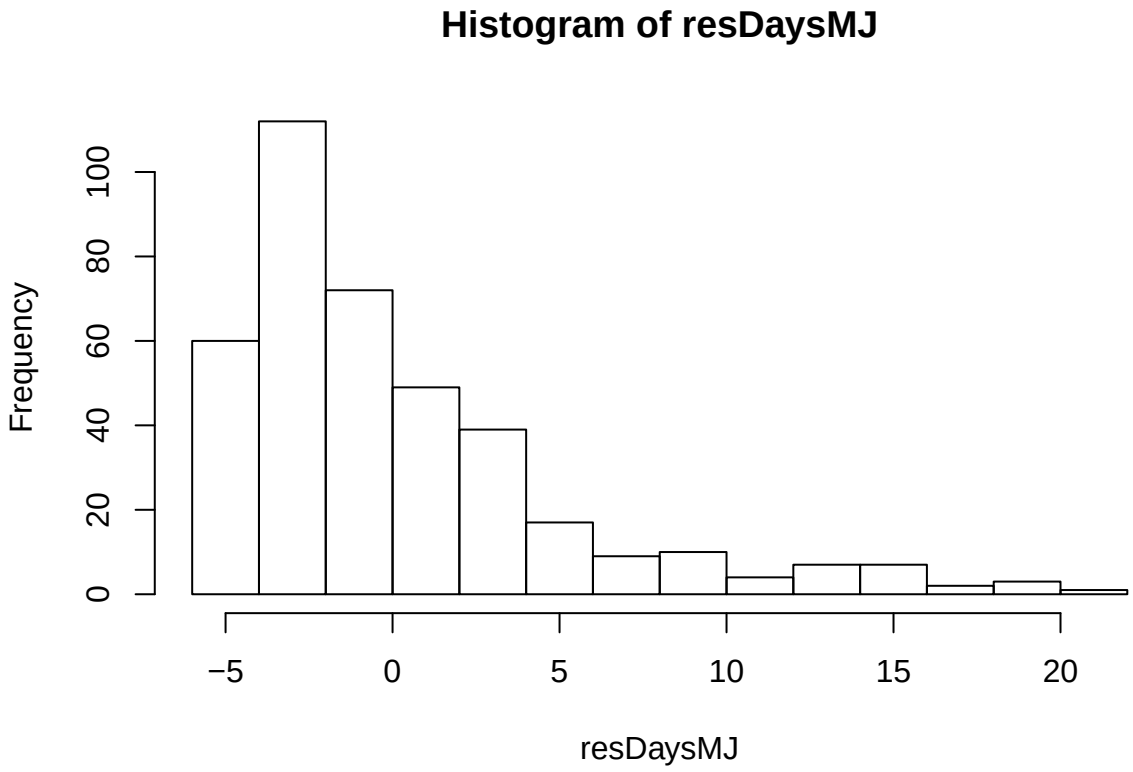
```
plot(daysMJ$DUQ230, daysMJ$DepressionScore)
abline(modDaysMJ)
```




```
resDaysMJ <- resid(modDaysMJ)
shapiro.test(resDaysMJ)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resDaysMJ
## W = 0.82664, p-value < 2.2e-16
```

```
hist(resDaysMJ)
```



Conduct Welch Two Sample T-test for the yes-no questions.

```
everCocaine_tb <- depression_fn[complete.cases(depression_fn[ , "DUQ250"]), ]
everCocaine_tb <- everCocaine_tb[everCocaine_tb$DUQ250 != "7", ]

everCocaine_Y <- everCocaine_tb[everCocaine_tb[, "DUQ250"] == 1, c("SEQN", "DepressionScore", "DUQ250")]
sumCocaine_Y <- c(everCocaine_Y[, "DepressionScore"])

everCocaine_N <- everCocaine_tb[everCocaine_tb[, "DUQ250"] == 2, c("SEQN", "DepressionScore", "DUQ250")]
sumCocaine_N <- c(everCocaine_N[, "DepressionScore"])

var.test(sumCocaine_Y, sumCocaine_N)
```

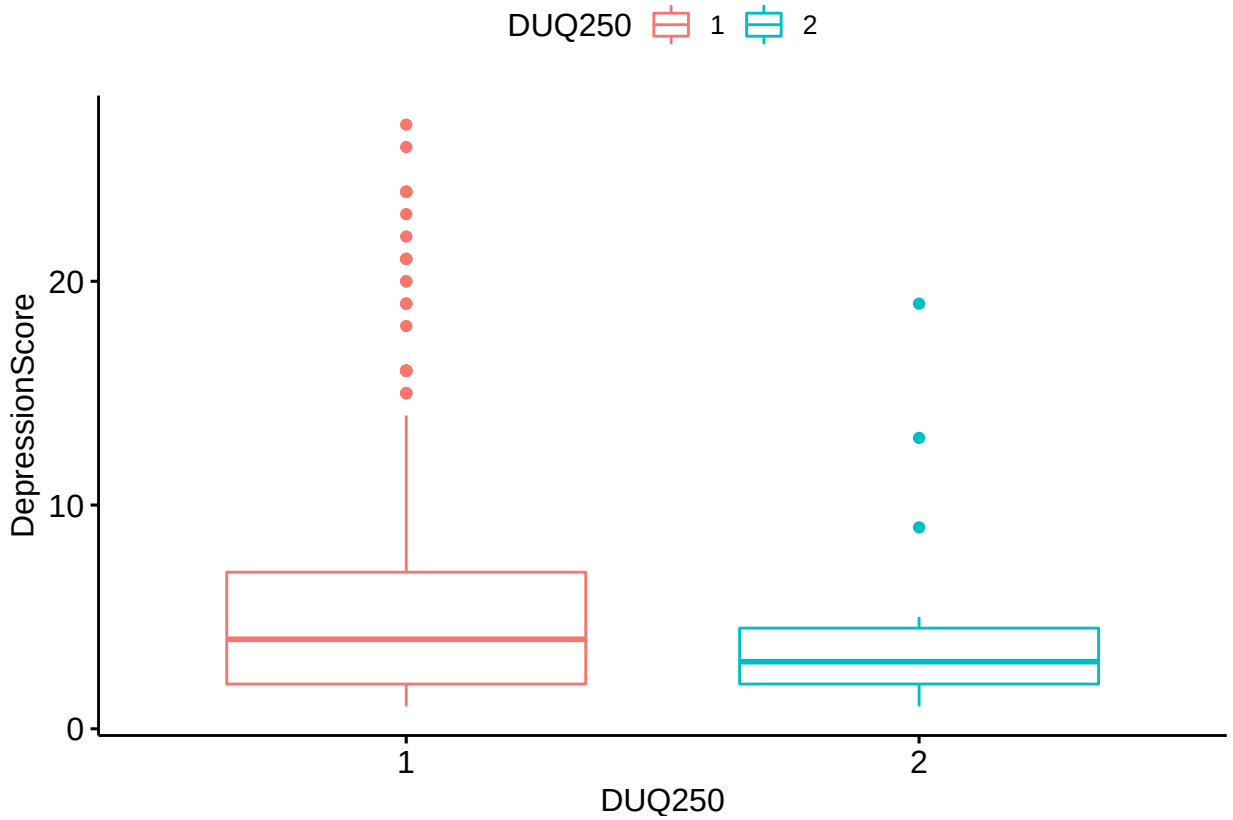
```
##
##  F test to compare two variances
##
## data:  sumCocaine_Y and sumCocaine_N
## F = 0.96934, num df = 426, denom df = 14, p-value = 0.8404
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.386850 1.838552
## sample estimates:
## ratio of variances
##          0.9693417
```

```
t.test(sumCocaine_Y, sumCocaine_N, var.equal = TRUE)
```

```
##
##  Two Sample t-test
```

```
##
## data:  sumCocaine_Y and sumCocaine_N
## t = 0.49801, df = 440, p-value = 0.6187
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.924738  3.231217
## sample estimates:
## mean of x mean of y
##  5.519906  4.866667
```

```
ggboxplot(everCocaine_tb, x = "DUQ250", y = "DepressionScore", color = "DUQ250")
```




```
everHeroin_tb <- depression_fn[complete.cases(depression_fn[ , "DUQ290"]), ]

everHeroin_Y <- everHeroin_tb[everHeroin_tb[, "DUQ290"] == 1, c("SEQN", "DepressionScore", "DUQ290")]
sumHeroin_Y <- c(everHeroin_Y[, "DepressionScore"])

everHeroin_N <- everHeroin_tb[everHeroin_tb[, "DUQ290"] == 2, c("SEQN", "DepressionScore", "DUQ290")]
sumHeroin_N <- c(everHeroin_N[, "DepressionScore"])

var.test(sumHeroin_Y, sumHeroin_N)
```

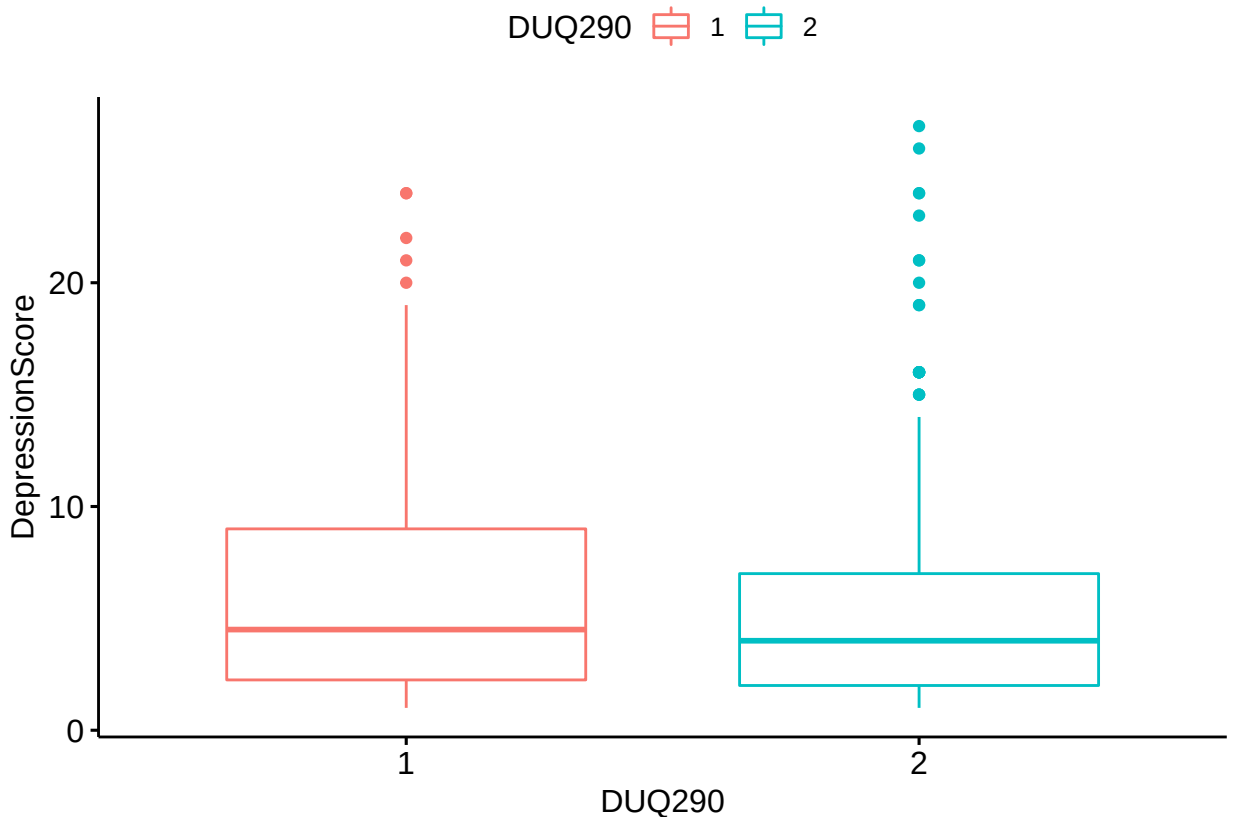
```
##
##  F test to compare two variances
##
## data:  sumHeroin_Y and sumHeroin_N
## F = 1.9218, num df = 57, denom df = 384, p-value = 0.0003633
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```
##  1.331090 2.942424
## sample estimates:
## ratio of variances
##           1.921764
```

```
t.test(sumHeroin_Y, sumHeroin_N, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  sumHeroin_Y and sumHeroin_N
## t = 2.21, df = 66.226, p-value = 0.03057
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1887281 3.7178550
## sample estimates:
## mean of x mean of y
##  7.189655  5.236364
```

```
ggboxplot(everHeroin_tb, x = "DUQ290", y = "DepressionScore", color = "DUQ290")
```




```
everMeth_tb <- depression_fn[complete.cases(depression_fn[ , "DUQ330"]), ]
everMeth_tb <- everMeth_tb[everMeth_tb$DUQ330 != "9", ]

everMeth_Y <- everMeth_tb[everMeth_tb[, "DUQ330"] == 1, c("SEQN", "DepressionScore", "DUQ330")]
sumMeth_Y <- c(everMeth_Y[, "DepressionScore"])

everMeth_N <- everMeth_tb[everMeth_tb[, "DUQ330"] == 2, c("SEQN", "DepressionScore", "DUQ330")]
sumMeth_N <- c(everMeth_N[, "DepressionScore"])
```

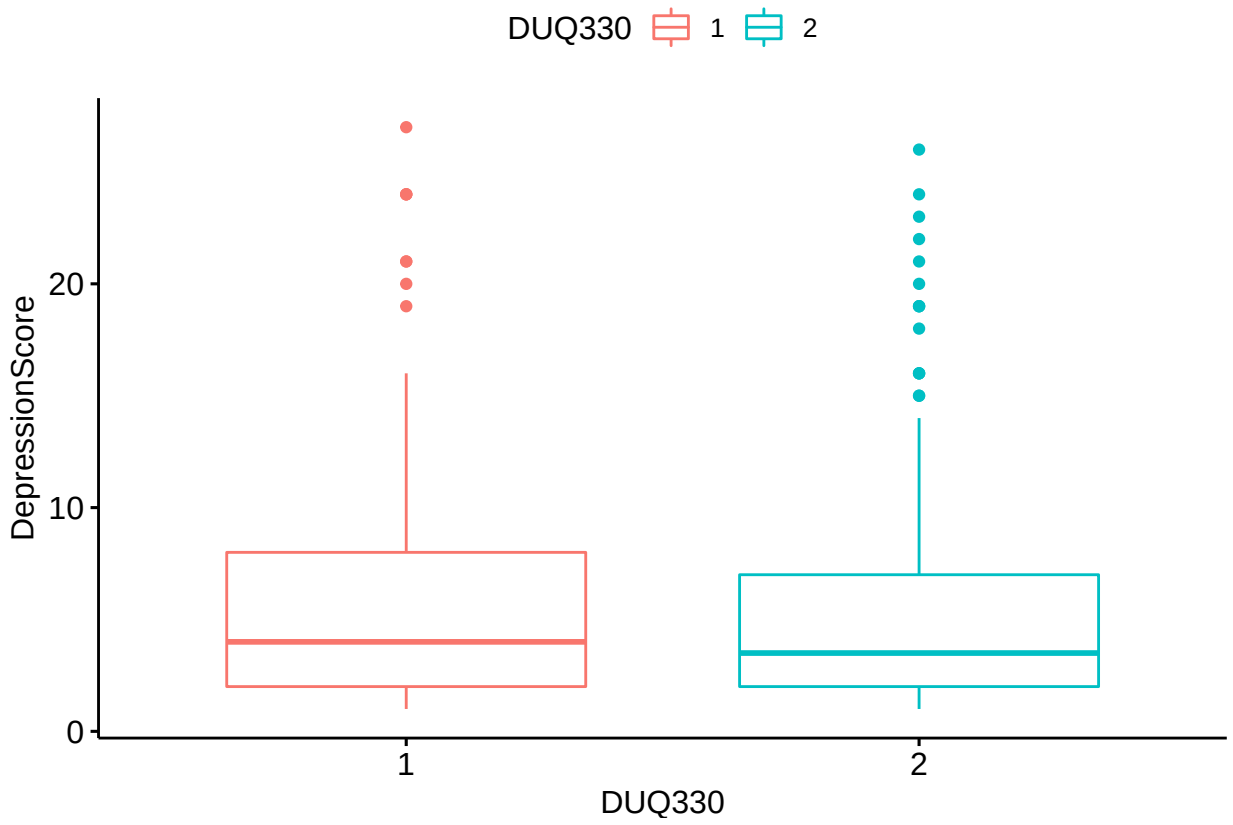
```r
var.test(sumMeth_Y, sumMeth_N)
```

```
##
##  F test to compare two variances
##
## data:  sumMeth_Y and sumMeth_N
## F = 1.2306, num df = 169, denom df = 271, p-value = 0.1293
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.941232 1.623117
## sample estimates:
## ratio of variances
##            1.23064
```

```r
t.test(sumMeth_Y, sumMeth_N, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sumMeth_Y and sumMeth_N
## t = 1.6577, df = 440, p-value = 0.09808
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1494131  1.7597072
## sample estimates:
## mean of x mean of y
##  5.970588  5.165441
```

```r
ggboxplot(everMeth_tb, x = "DUQ330", y = "DepressionScore", color = "DUQ330")
```




```
everInject_tb <- depression_fn[complete.cases(depression_fn[ , "DUQ370"]), ]
everInject_tb <- everInject_tb[everInject_tb$DUQ370 != "7", ]
everInject_tb <- everInject_tb[everInject_tb$DUQ370 != "9", ]

everInject_Y <- everInject_tb[everInject_tb[, "DUQ370"] == 1, c("SEQN", "DepressionScore", "DUQ370")]
sumInject_Y <- c(everInject_Y[, "DepressionScore"])

everInject_N <- everInject_tb[everInject_tb[, "DUQ370"] == 2, c("SEQN", "DepressionScore", "DUQ370")]
sumInject_N <- c(everInject_N[, "DepressionScore"])

var.test(sumInject_Y, sumInject_N)
```

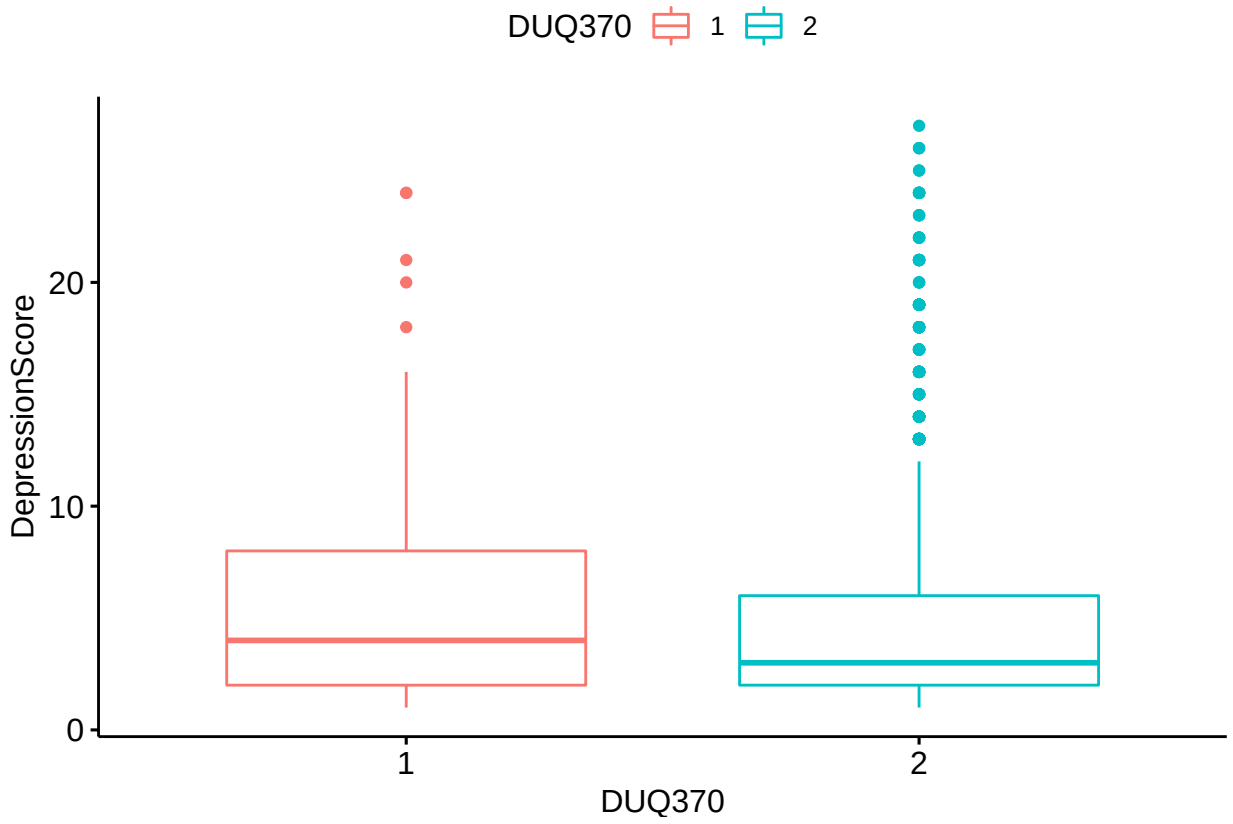
```
##
##  F test to compare two variances
##
## data:  sumInject_Y and sumInject_N
## F = 1.6842, num df = 62, denom df = 2251, p-value = 0.001518
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.21221 2.48855
## sample estimates:
## ratio of variances
##           1.684201
```

```
t.test(sumInject_Y, sumInject_N, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  sumInject_Y and sumInject_N
## t = 2.2694, df = 64.076, p-value = 0.02662
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1951921 3.0654017
## sample estimates:
## mean of x mean of y
##  6.285714  4.655417
```

```
ggboxplot(everInject_tb, x = "DUQ370", y = "DepressionScore", color = "DUQ370")
```




```
everRehabi_tb <- depression_fn[complete.cases(depression_fn[ , "DUQ430"]), ]

everRehabi_Y <- everRehabi_tb[everRehabi_tb[, "DUQ430"] == 1, c("SEQN", "DepressionScore", "DUQ430")]
sumRehabi_Y <- c(everRehabi_Y[, "DepressionScore"])

everRehabi_N <- everRehabi_tb[everRehabi_tb[, "DUQ430"] == 2, c("SEQN", "DepressionScore", "DUQ430")]
sumRehabi_N <- c(everRehabi_N[, "DepressionScore"])

var.test(sumRehabi_Y, sumRehabi_N)
```

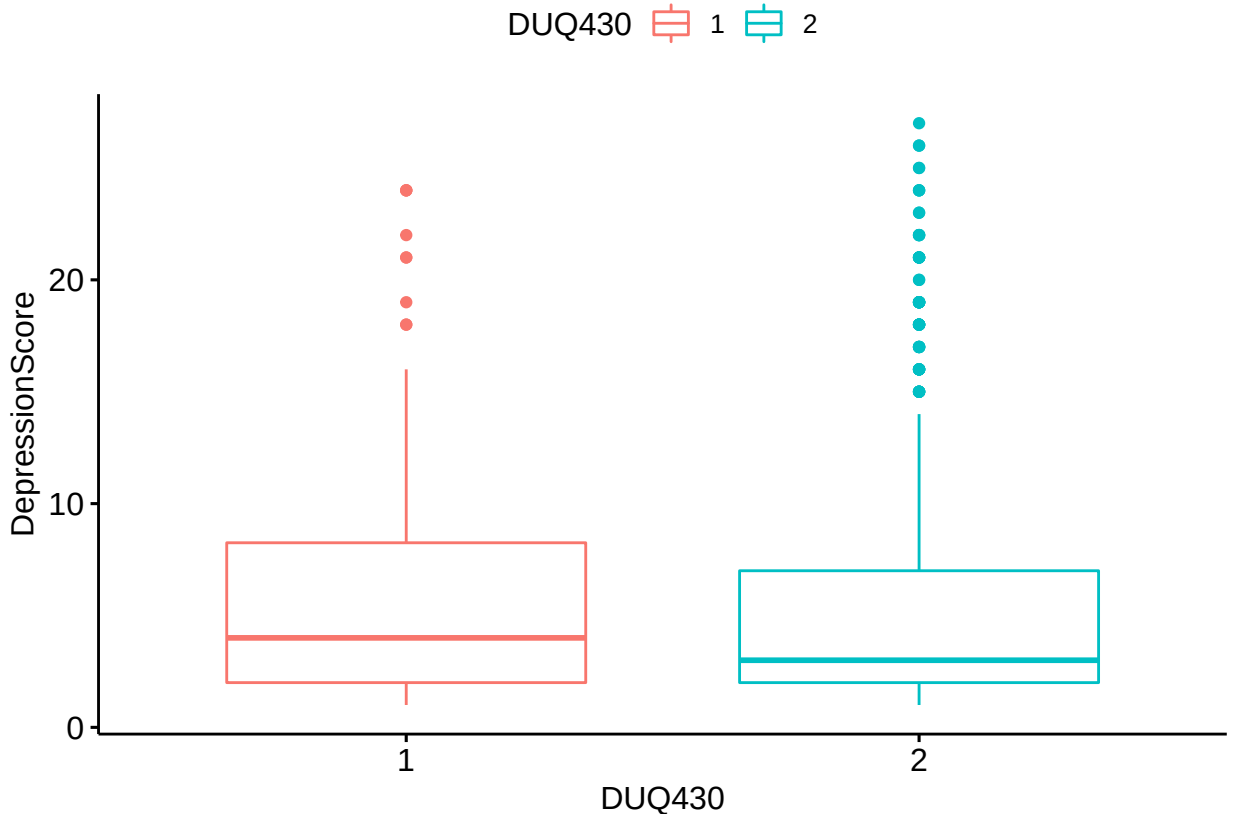
```
##
##  F test to compare two variances
##
## data:  sumRehabi_Y and sumRehabi_N
## F = 1.5017, num df = 115, denom df = 1042, p-value = 0.001782
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.159854 2.004151
```

```
## sample estimates:
## ratio of variances
##            1.50174
```

```
t.test(sumRehabi_Y, sumRehabi_N, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  sumRehabi_Y and sumRehabi_N
## t = 2.764, df = 132.58, p-value = 0.006523
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4344243 2.6210688
## sample estimates:
## mean of x mean of y
##  6.560345  5.032598
```

```
ggboxplot(everRehabi_tb, x = "DUQ430", y = "DepressionScore", color = "DUQ430")
```



Check the correlation coefficients of the nine items toward Question 10.

```
depression_fn2 <- depression_fn[ , c("SEQN", "DPQ010", "DPQ020", "DPQ030", "DPQ040", "DPQ050", "DPQ060"
cor(depression_fn2)[11, ]
```

```
##        SEQN      DPQ010      DPQ020      DPQ030      DPQ040      DPQ050
## -0.01474193  0.36164054  0.46526939  0.27308684  0.31958126  0.28855599
##      DPQ060      DPQ070      DPQ080      DPQ090      DPQ100
```

```
##  0.48091484  0.42499980  0.39825480  0.43242678  1.00000000
```

Develop a multiple logisitc regression model.

```
dp_data <- subset(depression_fn, select = c(depressed, DPQ020, DPQ060, DPQ070, DPQ080, DPQ090))
dp_data <- dp_data[complete.cases(dp_data[, 1]), ]

dp_glm <- glm(formula = depressed ~ ., family = "binomial", data = dp_data)
summary(dp_glm)
```

```
##
## Call:
## glm(formula = depressed ~ ., family = "binomial", data = dp_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8026  -0.2135  -0.1174  -0.1174   3.1565
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.97482    0.19849 -25.063  < 2e-16 ***
## DPQ020       1.20505    0.11068  10.888  < 2e-16 ***
## DPQ060       0.85346    0.10986   7.768 7.95e-15 ***
## DPQ070       1.18536    0.09607  12.339  < 2e-16 ***
## DPQ080       1.04883    0.11126   9.427  < 2e-16 ***
## DPQ090       0.45365    0.22286   2.036   0.0418 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2013.33  on 2746  degrees of freedom
## Residual deviance:  845.71  on 2741  degrees of freedom
## AIC: 857.71
##
## Number of Fisher Scoring iterations: 7
```

```
new1 <- data.frame(DPQ020 = 3, DPQ060 = 2, DPQ070 = 1, DPQ080 = 1, DPQ090 = 0)
result1 <- predict(dp_glm, newdata = new1, type = "response")
result1
```

```
##         1
## 0.9296583
```

```
new2 <- data.frame(DPQ020 = 0, DPQ060 = 2, DPQ070 = 0, DPQ080 = 2, DPQ090 = 3)
result2 <- predict(dp_glm, newdata = new2, type = "response")
result2
```

```
##         1
## 0.5475403
```

Take a look at the model's performance (with confusion matrix, the ROC curve, and the AUC value).

```
n <- nrow(dp_data)
set.seed(2747)
newdp_data <- dp_data[sample(n), ]

t_idx <- sample(seq_len(n), size = round(0.7 * n))
traindata <- newdp_data[t_idx, ]
testdata <- newdp_data[-t_idx, ]

dp_glm2 <- glm(formula = depressed ~ ., family = "binomial", data = traindata)
result3 <- predict(dp_glm2, newdata = testdata, type = "response")
result_dp <- ifelse(result3 > 0.6, 1, 0)
cm <- table(testdata$depressed, result_dp, dnn = c("reality", "prediction"))
cm
```

```
##         prediction
## reality   0   1
##       0 728  14
##       1  34  48
```

```
cm[4] / sum(cm[, 2])
```

```
## [1] 0.7741935
```

```
cm[1] / sum(cm[, 1])
```

```
## [1] 0.9553806
```

```
accuracy <- sum(diag(cm)) / sum(cm)
accuracy
```

```
## [1] 0.9417476
```

```
library(ROCR)
```

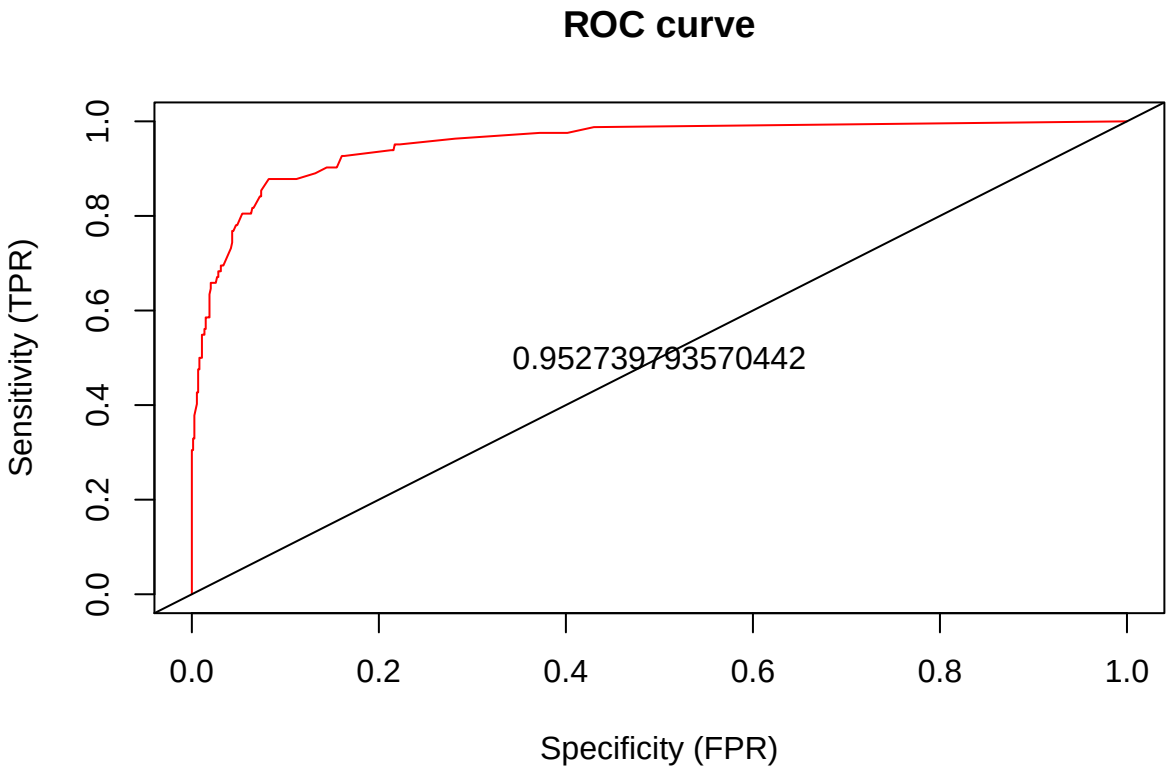
```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
pred <- prediction(result3, testdata$depressed)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, "auc")

plot(perf, col = rainbow(7), main = "ROC curve", xlab = "Specificity (FPR)", ylab = "Sensitivity (TPR)")
abline(0, 1)
text(0.5, 0.5, as.character(auc@y.values[[1]]))
```


ROC curve
Sensitivity (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
0.952739793570442
Specificity (FPR)
0.0
0.2
0.4
0.6
0.8
1.0