Stat235 Final Project

Nils Dahlin, Connor Guyette, Jon Kramer

10/20/2022

Stat235 Final Project: Drug Use Analysis

Nils Dahlin, Connor Guyette, and Jon Kramer

Introduction

Originally this project was going to be a general analysis of drug use and how factors like age, education level, and personality traits can affect a person's chances of using or trying different drugs. Soon into the process however we discovered some discussions about the increase in violent and destructive crime in Burlington and the possibility that it could be linked to an increase in Meth usage in the city which can lead to more expressive and chaotic outbreaks compared to drugs like Morphine or Heroin which normally have calming or sedative effects for the user. Due to this discovery we thought it would be interesting to focus on Meth as well as Heroin due to the ongoing Opioid epidemic in the region. In this project we will explore the relationships between these drugs and the features of Age and Education as well as looking at further complicated Logistic Regression models at the end with the addition of personality traits to explore those relationships a bit more.

Data Description

The data set we will be using is a collection of 1884 data points that track different attributes/features of individuals including age/gender/education level/country/ethnicity/personality trait scores/and drugs used and when.

Here is a quick look at the first 5 entries of our data set.

```
head(drug_data,5)
```

```
##   ID   Age Gender                        Education Country Ethnicity   Nscore
## 1  2 25-34      M                  Doctorate degree      UK     White -0.67825
## 2  3 35-44      M Professional certificate/ diploma      UK     White -0.46725
## 3  4 18-24      F                    Masters degree      UK     White -0.14882
## 4  5 35-44      F                  Doctorate degree      UK     White  0.73545
## 5  6   65+      F           Left school at 18 years  Canada     White -0.67825
##     Escore   Oscore   AScore   Cscore Impulsive       SS Alcohol Amphet Amyl
## 1  1.93886  1.43533  0.76096 -0.14277  -0.71126 -0.21575     CL5    CL2  CL2
## 2  0.80523 -0.84732 -1.62090 -1.01450  -1.37983  0.40148     CL6    CL0  CL0
## 3 -0.80615 -0.01928  0.59042  0.58489  -1.37983 -1.18084     CL4    CL0  CL0
## 4 -1.63340 -0.45174 -0.30172  1.30612  -0.21712 -0.21575     CL4    CL1  CL1
## 5 -0.30033 -1.55521  2.03972  1.63088  -1.37983 -1.54858     CL2    CL0  CL0
##   Benzos Caff Cannabis Choc Coke Crack Ecstasy Heroin Ketamine Legalh LSD Meth
## 1    CL0  CL6      CL4  CL6  CL3   CL0     CL4    CL0      CL2    CL0 CL2  CL3
## 2    CL0  CL6      CL3  CL4  CL0   CL0     CL0    CL0      CL0    CL0 CL0  CL0
## 3    CL3  CL5      CL2  CL4  CL2   CL0     CL0    CL0      CL2    CL0 CL0  CL0
## 4    CL0  CL6      CL3  CL6  CL0   CL0     CL1    CL0      CL0    CL1 CL0  CL0
## 5    CL0  CL6      CL0  CL4  CL0   CL0     CL0    CL0      CL0    CL0 CL0  CL0
##   Mushrooms Nicotine Semer VSA
## 1       CL0      CL4   CL0 CL0
## 2       CL1      CL0   CL0 CL0
## 3       CL0      CL2   CL0 CL0
## 4       CL2      CL2   CL0 CL0
## 5       CL0      CL6   CL0 CL0
```

These are the descriptions for the more nondescript features:

Personality Traits: NScore - Neuroticism EScore - Extroversion OScore - Openness to Experience AScore - Agreeableness CScore - Conscientiousness Impulsiveness SS - Sensation Seeing

The dataset we pulled had altered the personality trait scores for analysis. The scores have been normalized and specific values represent a Z-Score or Standard Deviation. We will be using the scores for Neuroticism, Extroversion, Openness to Experience, Agreeableness, Conscientiousness, and Impulsiveness later on towards the end of this project.

Drug Use Classifications: 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

These classifications will later be reduced to whether an individual has used the drug at some point or never used the drug.

```
# removing features we will not be using
main_df <- select(drug_data, Age, Gender, Education, Country, Nscore, Escore, Oscore, AS
core, Cscore, Impulsive, Heroin, Meth)
sjPlot::view_df(main_df,
 show.frq = T,
 show.prc = T,
 show.na = T,
 show.string.values = T)
```

Data frame: main_df

ID	*Name*	*Label*	*missings*	*Values*	*Value Labels*	*Freq.*	*%*
1	Age		0 (0.00%)		18-24	643	34.13
					25-34	481	25.53
					35-44	355	18.84
					45-54	294	15.61
					55-64	93	4.94
					65+	18	0.96
2	Gender		0 (0.00%)		F	941	49.95
					M	943	50.05
3	Education		0 (0.00%)		Doctorate degree	89	4.72
					Left school at 16 years	99	5.25
					Left school at 17 years	30	1.59
					Left school at 18 years	100	5.31
					Left school before 16 years	28	1.49
					Masters degree	283	15.02
					Professional certificate/ diploma	269	14.28
					Some college or university, no certificate or degree	506	26.86
					University degree	480	25.48
4	Country		0 (0.00%)		Australia	54	2.87
					Canada	87	4.62
					New Zealand	5	0.27
					Other	118	6.26
					Republic of Ireland	20	1.06
					UK	1043	55.36
					USA	557	29.56
5	Nscore		0 (0.00%)	*range: -3.5-3.3*			
6	Escore		0 (0.00%)	*range: -3.3-3.3*			
7	Oscore		0 (0.00%)	*range: -3.3-2.9*			
8	AScore		0 (0.00%)	*range: -3.5-3.5*			
9	Cscore		0 (0.00%)	*range: -3.5-3.5*			
10	Impulsive		0 (0.00%)	*range: -2.6-2.9*			

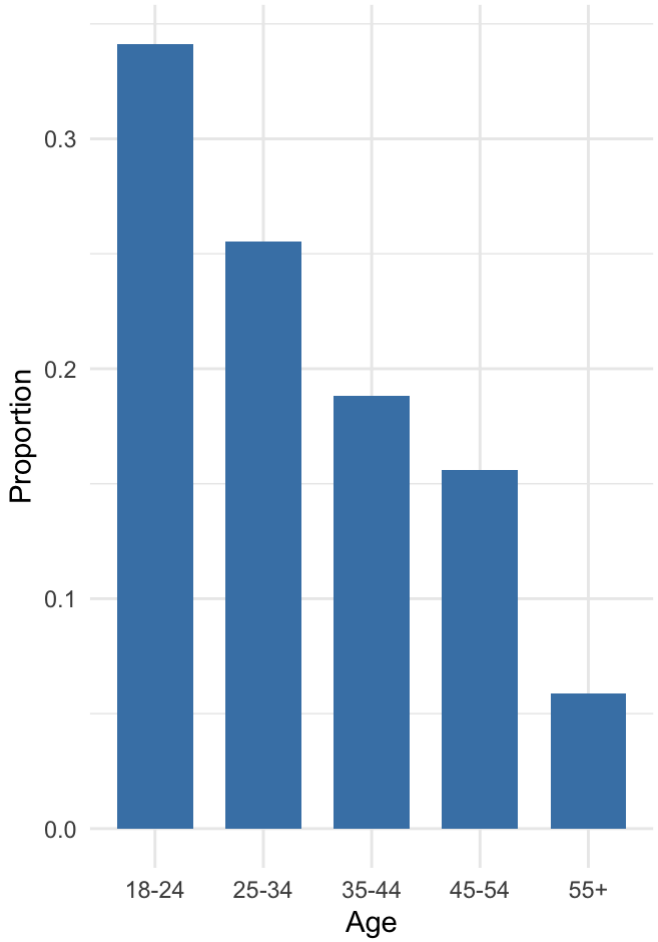
11	Heroin	0 (0.00%)	CL0	1604	85.14
			CL1	68	3.61
			CL2	94	4.99
			CL3	65	3.45
			CL4	24	1.27
			CL5	16	0.85
			CL6	13	0.69
12	Meth	0 (0.00%)	CL0	1428	75.80
			CL1	39	2.07
			CL2	97	5.15
			CL3	149	7.91
			CL4	50	2.65
			CL5	48	2.55
			CL6	73	3.87

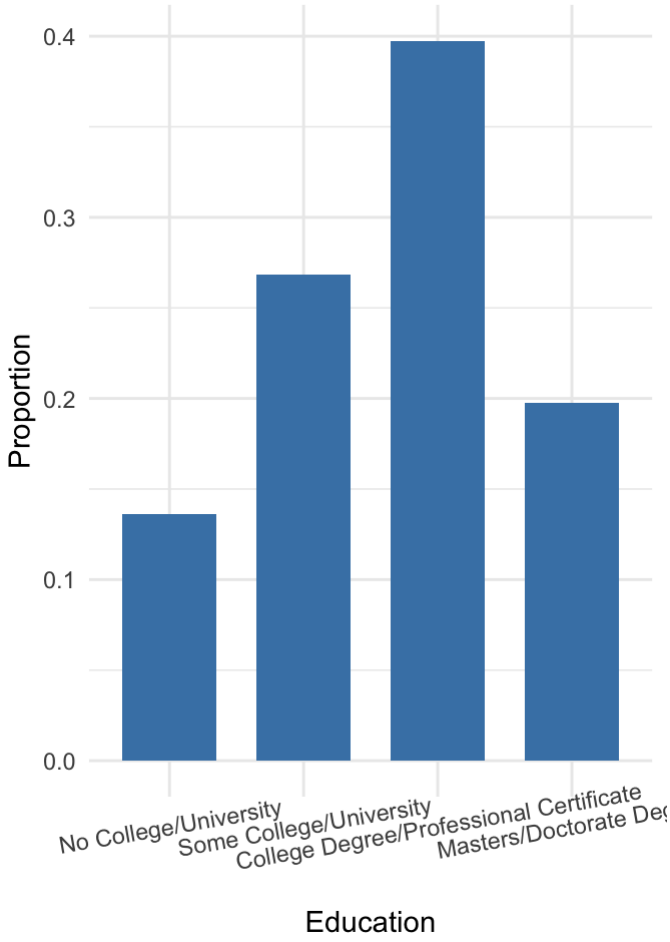
Data Summary

We are interested in initially looking at the association between age and certain types of drug use as well as personality traits and drug use. From there we are interested at looking at the associations between these controlling for education level. These are some of our initial data summaries:

Distribution of Ages and Education

```
#plot props
age_prop <- main_df |>
  group_by(Age) |>
  count(Age) |>
  ungroup() |>
  mutate(prop = n/sum(n))

edu_prop <- main_df |>
  group_by(Education) |>
  count(Education) |>
  ungroup() |>
  mutate(prop = n/sum(n))

age_dist <- ggplot(age_prop, aes(x=Age, y=prop))+
  geom_bar(stat="identity", width=0.7, fill="steelblue")+
  theme_minimal()+
  xlab("Age") + ylab("Proportion")

edu_dist <- ggplot(edu_prop, aes(x=Education, y=prop))+
  geom_bar(stat="identity", width=0.7, fill="steelblue")+
  theme_minimal()+
  xlab("Education") + ylab("Proportion")+
  theme(axis.text.x = element_text(angle = 10, vjust=.9))

grid.arrange(age_dist,edu_dist,nrow=1)
```

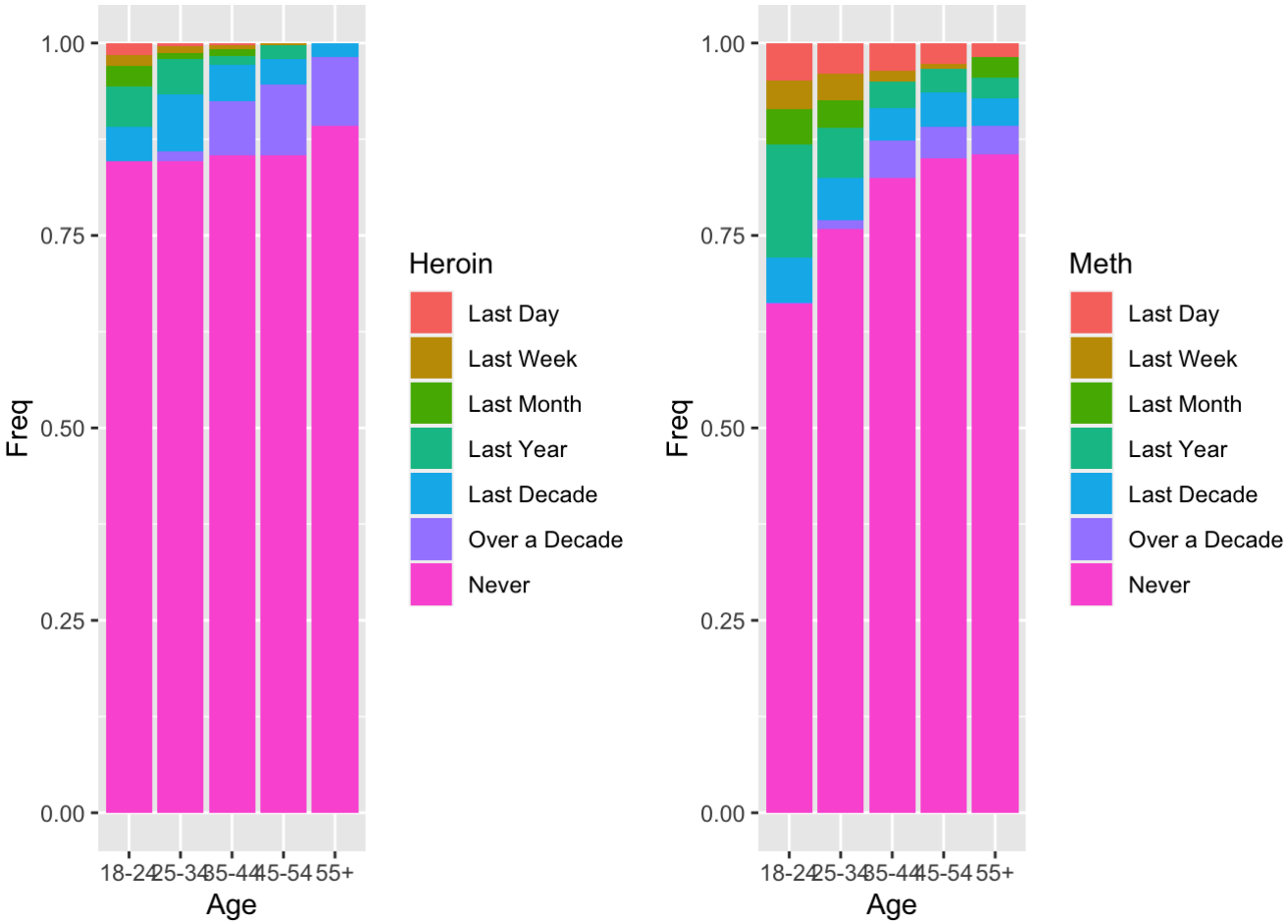




The distribution of ages is heavily skewed right with the majority of the data consisting of those younger than 45. The amount of data in the youngest bracket is the largest with the amount tailing off as age brackets increase.

Usage by Age

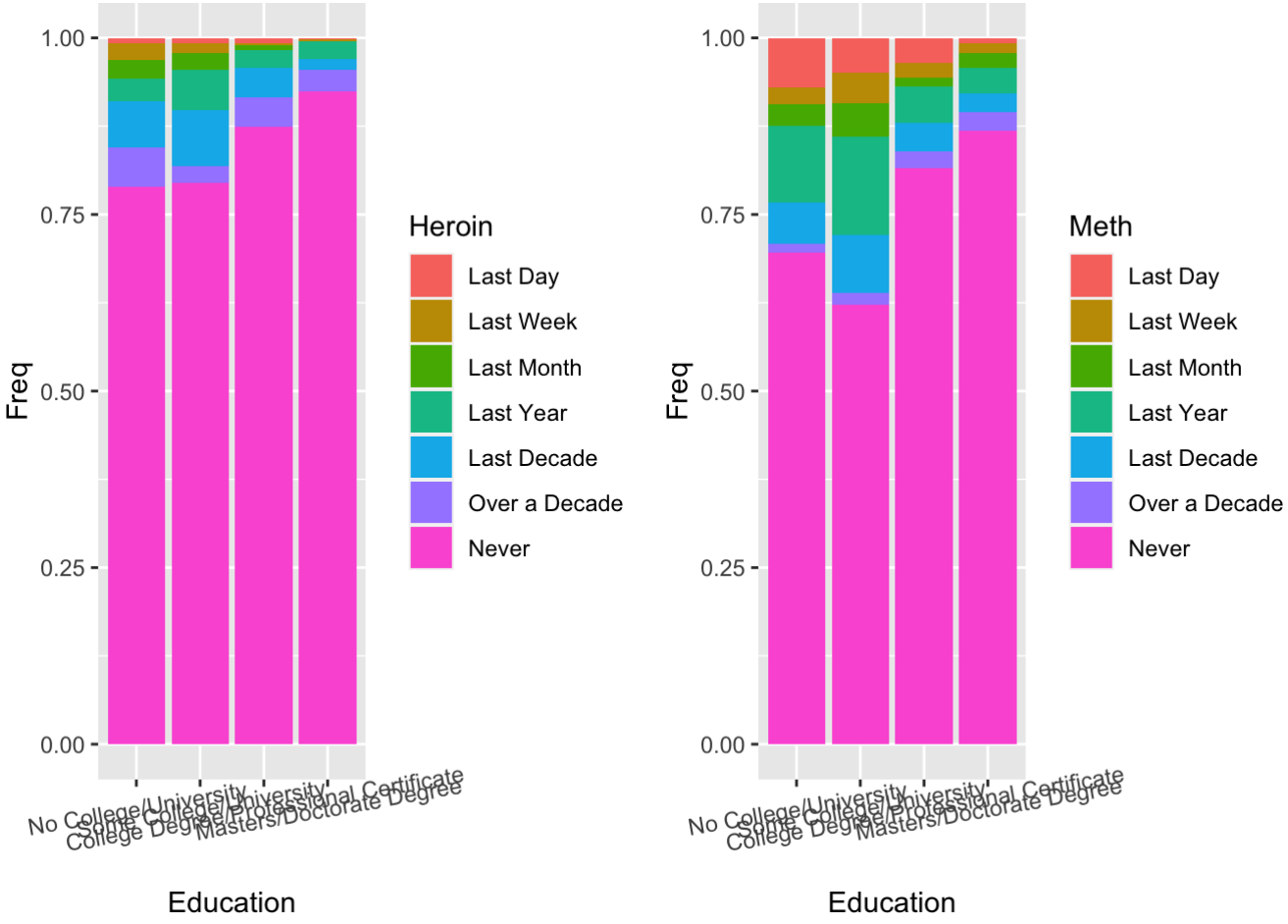
```
#Heroin
Age_Heroin_prop<-data.frame(prop.table(xtabs(data = main_df, ~Heroin+Age),margin="Age"))


age_h_plt <- ggplot(Age_Heroin_prop, aes(x=Age,y=Freq, fill=Heroin))+
  geom_col()


#Methamphetamines
Age_Meth_prop<-data.frame(prop.table(xtabs(data = main_df, ~Meth+Age),margin="Age"))


age_m_plt <- ggplot(Age_Meth_prop, aes(x=Age,y=Freq, fill=Meth))+
  geom_col()

grid.arrange(age_h_plt,age_m_plt,nrow=1)
```



From an initial analysis of the relationship of Age with Usage there seem to be some interesting patterns. In regards to Heroin it looks there is a pretty even distribution across the age ranges under 55, however there is much more recent use, within the last year, among younger age brackets whereas the use among the higher brackets is largely within the last decade or over a decade ago. In regards to Meth we see a different story with there being a larger proportion of those within younger age brackets using or have used compared to those within the older age brackets.

Usage by Education

```
# Heroin
Edu_Heroin_prop<-data.frame(prop.table(xtabs(data = main_df, ~Heroin+Education),margin
="Education"))


edu_h_plt <- ggplot(Edu_Heroin_prop, aes(x=Education,y=Freq, fill=Heroin))+
  geom_col()+
  theme(axis.text.x = element_text(angle = 10, vjust=.9))

# Meth
Edu_Meth_prop<-data.frame(prop.table(xtabs(data = main_df, ~Meth+Education),margin="Educ
ation"))


edu_m_plt <- ggplot(Edu_Meth_prop, aes(x=Education,y=Freq, fill=Meth))+
  geom_col()+
  theme(axis.text.x = element_text(angle = 10, vjust=.9))

grid.arrange(edu_h_plt,edu_m_plt,nrow=1)
```



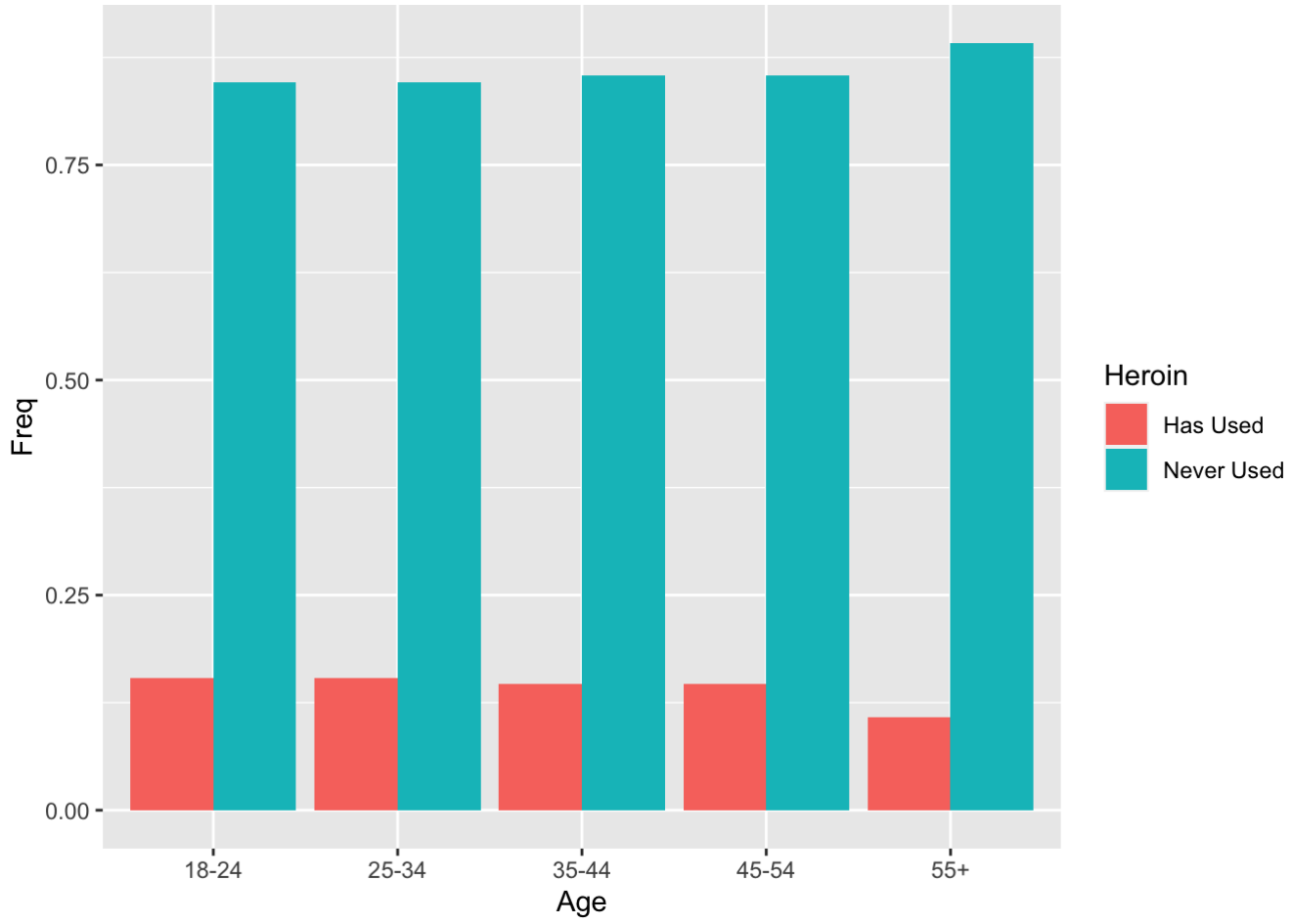
From an initial analysis on Education and Usage we a bit more variation amongst the groups however there is a noticeable trend with the groups of those having not completed college or some form of higher education having a larger proportion of usage.

Examining Initial Relationships (Tests for Two Variables)

We will now examine the statistical relationships between Usage, Age, and Education starting with simple tests of association between the Age and Education in relation to the usage of these drugs.

Heroin by Age

```
#reducing heroin usage to whether an individual has used or not
heroin_used_df <- main_df |>
  select(-Meth) |>
  mutate(Heroin=ifelse(Heroin=='Never','Never Used','Has Used'))

#getting counts for each group of heroin by age
heroin_age_freq <-
  xtabs(formula = ~ Age + Heroin,
        data = heroin_used_df)

#visualization of marginal proportions
ggplot(data.frame(prop.table(heroin_age_freq, margin = "Age")), aes(x=Age,y=Freq, fill=H
eroin))+
  geom_col(position = "dodge")
```



```
#creating df of marginal props for Chi-squared test
heroin_age_freq |>
  # Convert counts to conditional proportions
  prop.table(margin = "Age") |>
  # Display 3 significant digits
  signif(digits = 3) |>
  # Convert to a data frame
  data.frame() |>
  pivot_wider(names_from = "Heroin",
              values_from = "Freq")
```

```
## # A tibble: 5 × 3
##   Age   `Has Used` `Never Used`
##   <fct>      <dbl>        <dbl>
## 1 18-24      0.154        0.846
## 2 25-34      0.154        0.846
## 3 35-44      0.146        0.854
## 4 45-54      0.146        0.854
## 5 55+        0.108        0.892
```

```
#testing association
chisq_test(x=heroin_age_freq)
```

```
## # A tibble: 1 × 6
##       n statistic     p    df method          p.signif
## * <int>     <dbl> <dbl> <int> <chr>           <chr>
## 1  1884      1.71 0.788     4 Chi-square test ns
```

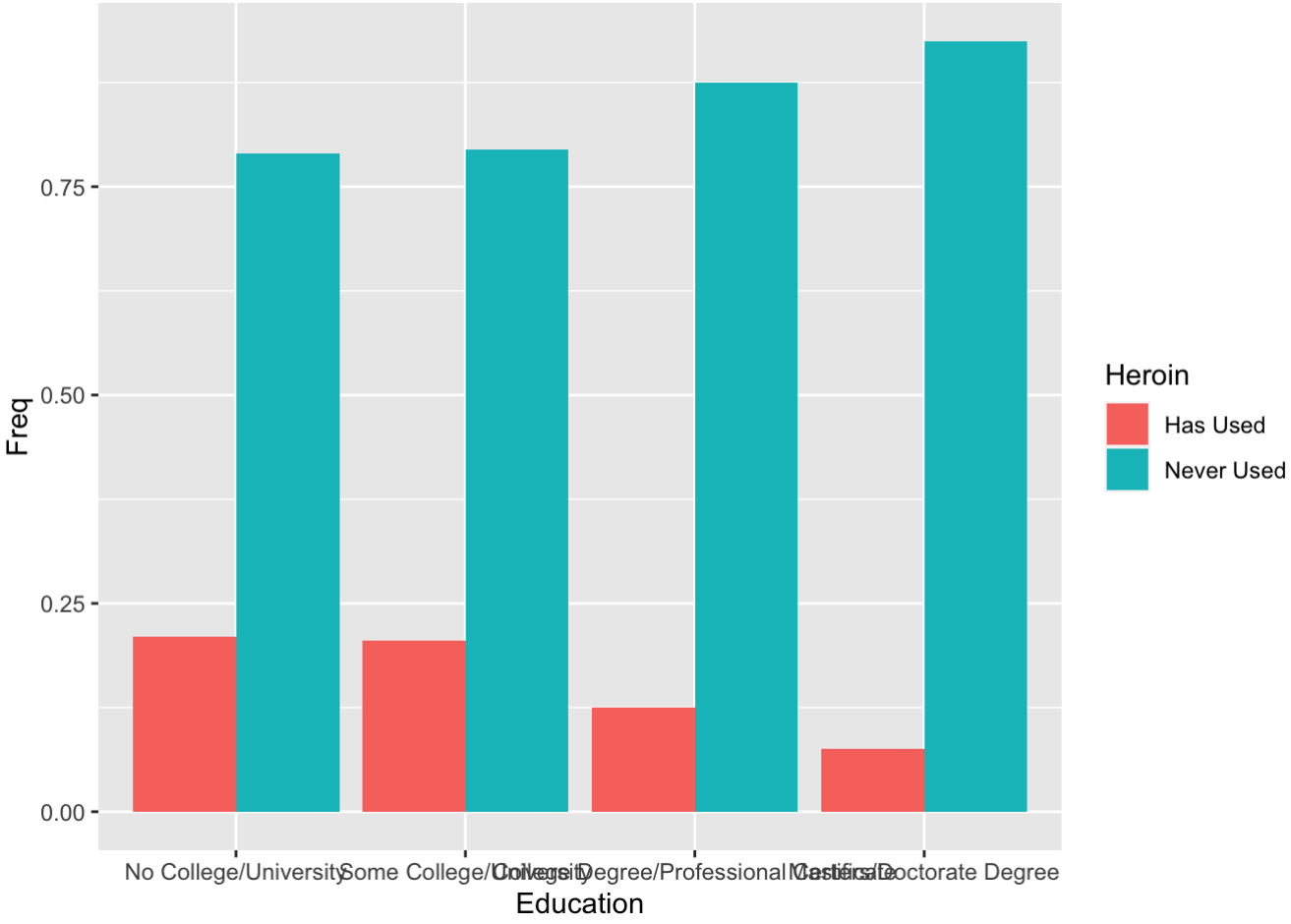
```
cramerV(heroin_age_freq,
        ci = T,
        conf = 0.95)
```

```
##   Cramer.V lower.ci upper.ci
## 1  0.03017  0.01978  0.08701
```

We can see from the initial graph there is not that much difference between the groups when we reduce the used or not. This is isn't unexpected from our initial exploration. Further the results of the Chi-square test returned a p-value of 0.78 which would lead us to failing to reject the null hypothesis of the test and conclude that there is not a significant association between Heroin usage and Age.

Heroin by Education

```
#getting counts for each group of heroin by education
heroin_edu_freq <-
  xtabs(formula = ~ Education + Heroin,
        data = heroin_used_df)

#visualization of marginal proportions
ggplot(data.frame(prop.table(heroin_edu_freq, margin = "Education")), aes(x=Education,y=
Freq, fill=Heroin))+
  geom_col(position = "dodge")
```




```
#creating df of marginal props for Chi-squared test
heroin_edu_freq |>
  # Convert counts to conditional proportions
  prop.table(margin = "Education") |>
  # Display 3 significant digits
  signif(digits = 3) |>
  # Convert to a data frame
  data.frame() |>
  pivot_wider(names_from = "Heroin",
              values_from = "Freq")
```

```
## # A tibble: 4 × 3
##   Education                                `Has Used` `Never Used`
##   <fct>                                         <dbl>        <dbl>
## 1 No College/University                        0.21          0.79
## 2 Some College/University                      0.206         0.794
## 3 College Degree/Professional Certificate      0.126         0.874
## 4 Masters/Doctorate Degree                     0.0753        0.925
```

```
#testing association
chisq_test(x=heroin_edu_freq)
```

```
## # A tibble: 1 × 6
##       n statistic            p    df method          p.signif
## * <int>     <dbl>        <dbl> <int> <chr>           <chr>
## 1  1884      39.6 0.0000000128     3 Chi-square test ****
```

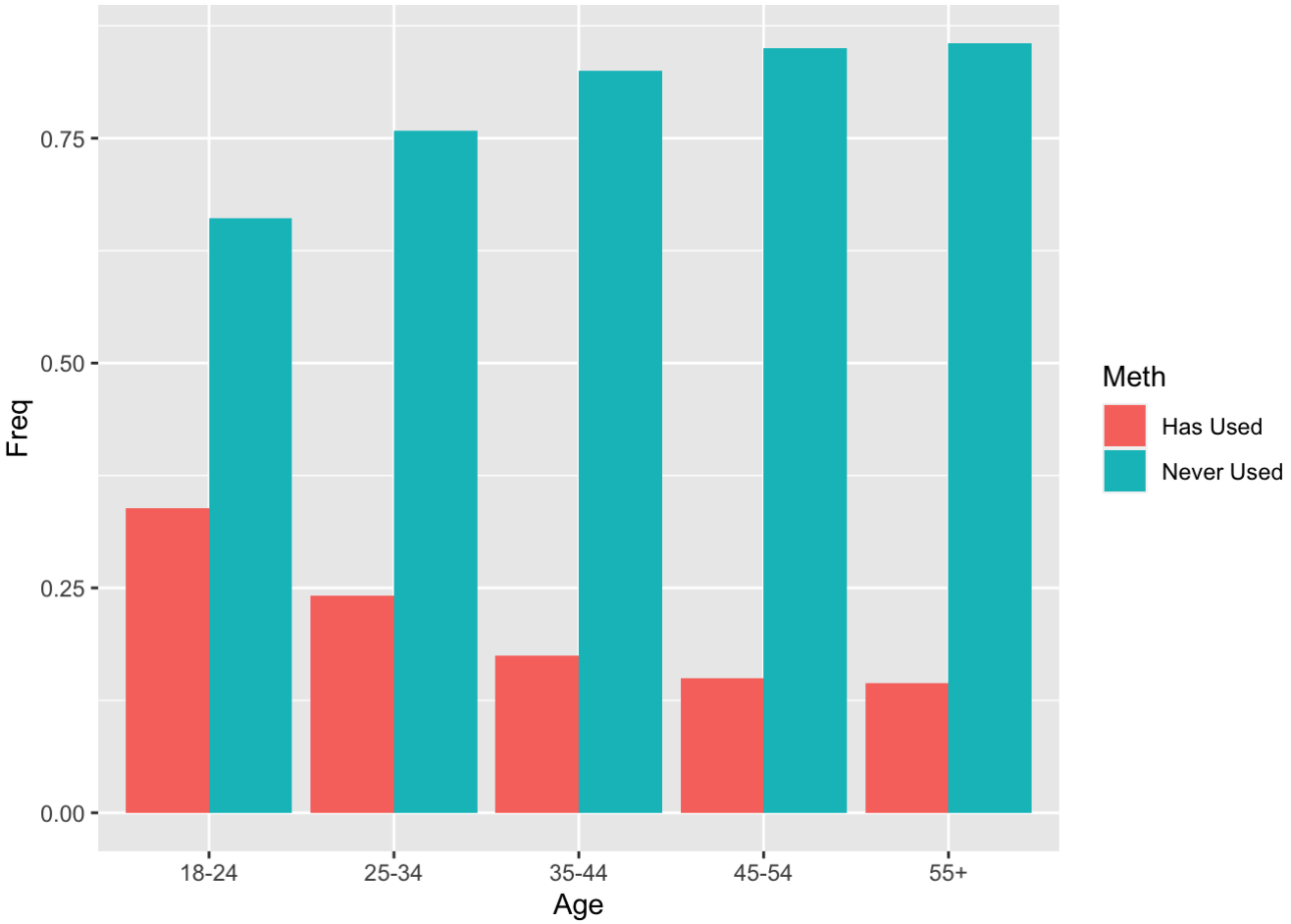
```
cramerV(heroin_edu_freq,
        ci = T,
        conf = 0.95)
```

```
##   Cramer.V lower.ci upper.ci
## 1    0.145   0.1072   0.1922
```

The initial graph for Education and Heroin however shows a bit more variation among the groups. When we run the Chi-square test we can see that the resulting p-value is very small which would lead us to rejecting the null hypothesis and concluding that there is a significant association between Education and Heroin usage.

Meth by Age

```
meth_used_df <- main_df |>
  select(-Heroin) |>
  mutate(Meth=ifelse(Meth=='Never','Never Used','Has Used'))


meth_age_freq <-
  xtabs(formula = ~ Age + Meth,
        data = meth_used_df)

ggplot(data.frame(prop.table(meth_age_freq, margin = "Age")), aes(x=Age,y=Freq, fill=Met
h))+
  geom_col(position = "dodge")
```




```
meth_age_freq |>
  # Convert counts to conditional proportions
  prop.table(margin = "Age") |>
  # Display 3 significant digits
  signif(digits = 3) |>
  # Convert to a data frame
  data.frame() |>

  pivot_wider(names_from = "Meth",
              values_from = "Freq")
```

```
## # A tibble: 5 × 3
##   Age   `Has Used` `Never Used`
##   <fct>      <dbl>        <dbl>
## 1 18-24      0.339        0.661
## 2 25-34      0.241        0.759
## 3 35-44      0.175        0.825
## 4 45-54      0.15         0.85
## 5 55+        0.144        0.856
```

```
chisq_test(x=meth_age_freq)
```

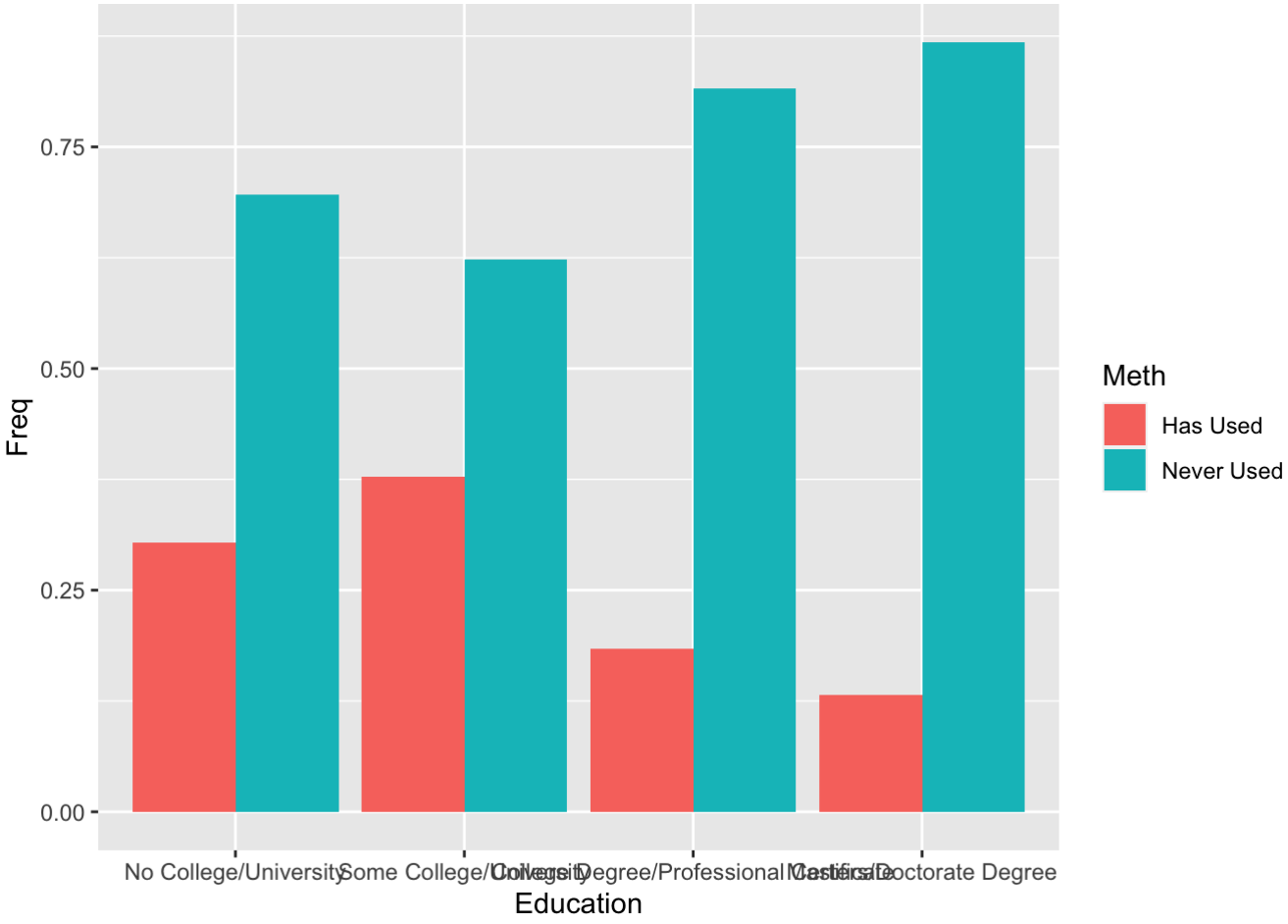
```
## # A tibble: 1 × 6
##       n statistic        p    df method          p.signif
## * <int>     <dbl>    <dbl> <int> <chr>           <chr>
## 1  1884      61.2 1.59e-12     4 Chi-square test ****
```

```
cramerV(meth_age_freq,
        ci = T,
        conf = 0.95)
```

```
##   Cramer.V lower.ci upper.ci
## 1   0.1803     0.14   0.2297
```

Meth by Education

```
meth_edu_freq <-
  xtabs(formula = ~ Education + Meth,
        data = meth_used_df)

ggplot(data.frame(prop.table(meth_edu_freq, margin = "Education")), aes(x=Education,y=Fr
eq, fill=Meth))+
  geom_col(position = "dodge")
```



```
meth_edu_freq |>
  # Convert counts to conditional proportions
  prop.table(margin = "Education") |>
  # Display 3 significant digits
  signif(digits = 3) |>
  # Convert to a data frame
  data.frame() |>

  pivot_wider(names_from = "Meth",
              values_from = "Freq")
```

```
## # A tibble: 4 × 3
##   Education                                `Has Used` `Never Used`
##   <fct>                                         <dbl>        <dbl>
## 1 No College/University                         0.304        0.696
## 2 Some College/University                       0.377        0.623
## 3 College Degree/Professional Certificate       0.184        0.816
## 4 Masters/Doctorate Degree                      0.132        0.868
```

```
chisq_test(x=meth_edu_freq)
```

```
## # A tibble: 1 × 6
##       n statistic        p    df method          p.signif
## * <int>     <dbl>    <dbl> <int> <chr>           <chr>
## 1  1884      94.2 2.75e-20     3 Chi-square test ****
```

```
cramerV(meth_edu_freq,
        ci = T,
        conf = 0.95)
```

```
##   Cramer.V lower.ci upper.ci
## 1   0.2236   0.1802   0.2706
```

Looking at the results for Meth we can see from the initial visualization that there definitely seems to be some noticeable variation between the groups when looking at both Age and Education. In both instances the p-values were small enough to lead us to rejecting the null hypothesis and concluding that there are significant associations between both of these variables and Meth usage.

Examining More Complicated Relationships (Tests for Multiple Variables)

After doing our initial analysis we will now dive a little deeper into the usage of these drugs and examine the more complicated relationships between the combination of Age and Education in relation to Usage instead of looking at them separately. We will be using X = Education, Y = Drug Use, and Z = Age for our analysis.

Heroin

Complete Independence

```
#Getting Counts and Expected Proportions
heroin_sum <-
  heroin_used_df |>
  count(Age, Education, Heroin) |>
  # Calculating the proportions: n/sum(n)
  mutate(FA_prop = n/sum(n))

I <- n_distinct(heroin_used_df$Education)
J <- n_distinct(heroin_used_df$Heroin)
K <- n_distinct(heroin_used_df$Age)

heroin_CI <-
  heroin_sum |>
  group_by(Age) |>
  mutate(age_n = sum(n)) |>
  ungroup() |>

  group_by(Education) |>
  mutate(edu_n = sum(n)) |>
  ungroup() |>

  group_by(Heroin) |>
  mutate(heroin_n = sum(n)) |>
  ungroup() |>

  # Now we can calculate the expected proportion for each outcome assuming complete inde
pendence
  mutate(CI_prop = age_n/sum(n) * edu_n/sum(n) * heroin_n/sum(n)) |>

  # Dropping the count columns because we don't need them in the future:
  select(-age_n, -edu_n, -heroin_n)

#Getting Test Statistics:
CI_FA_test <-
  heroin_CI |>
  # Calculating the individual pieces of our test statistics (chi^2 and G)
  mutate(zi2 = (FA_prop - CI_prop)^2/CI_prop,
         gi = n*log(FA_prop/CI_prop)) |>

  # Adding the individual pieces to get the test statistic:
  summarize(chi2 = sum(n)*sum(zi2),
            lrt_g = 2*sum(gi)) |>

  # Changing the results from being stored in separate columns to the same column
  pivot_longer(cols = chi2:lrt_g,
               names_to = "test",
               values_to = "stat")


# Calculating P-Values:
# The number of unique proportions needed for the FA model
r1 <- I*J*K - 1
```

```
# The number of unique proportions needed for the CI model
r0 <- I + J + K - 3

# The degrees of freedom: r1 - r0
df_CI <- r1 - r0
df_CI
```

```
## [1] 31
```

```
CI_FA_test |>
  mutate(p_val = pchisq(stat, df = df_CI, lower = F))
```

```
## # A tibble: 2 × 3
##   test   stat    p_val
##   <chr> <dbl>    <dbl>
## 1 chi2   430. 9.81e-72
## 2 lrt_g  463. 1.80e-78
```

```
#Checking Sample Size:
# heroin_CI |>
#   mutate(n_CI = sum(n)*CI_prop) |>
#   arrange(n_CI)
```

With a P-value of nearly 0 we reject our null hypothesis that the three variables of Age, Heroin Use, and Education are all independent. We conclude that at least two of the variables are associated. This was to be expected as we saw that education level and heroin use were associated in the initial examination. One thing to note throughout these analyses is the fact that due to the smaller size of the data set and limited time and resources even when combining the higher age brackets we had a few combinations of higher aged groups who expected counts were under 5 which could negatively affect the tests. We wanted to maintain some amount of complexity/difference between the age and education groups and we also felt that due to limited time and the scope of the project it would be ok to maintain what we had and not spend more time manipulating the data and rerunning tests on the new data in hopes of being able to increase expected counts.

Joint Independence

```
heroin_JI <-
  heroin_used_df |>
  mutate(edu_use = interaction(Education, Heroin, sep = ":"))


JI_vs_FA <-
  chisq_test(x = heroin_JI$edu_use,
             y = heroin_JI$Age)

JI_vs_FA
```

```
## # A tibble: 1 × 6
##       n statistic        p    df method          p.signif
## * <int>     <dbl>    <dbl> <int> <chr>           <chr>
## 1  1884      415. 1.68e-70    28 Chi-square test ****
```

```
#Checking Expected Counts:
expected_freq(JI_vs_FA) |>
  round(digits = 1)
```

```
##                                                        y
## x                                                     18-24 25-34 35-44 45-54
##   No College/University:Has Used                       18.4  13.8  10.2   8.4
##   Some College/University:Has Used                     35.5  26.6  19.6  16.2
##   College Degree/Professional Certificate:Has Used     32.1  24.0  17.7  14.7
##   Masters/Doctorate Degree:Has Used                     9.6   7.1   5.3   4.4
##   No College/University:Never Used                     69.3  51.8  38.3  31.7
##   Some College/University:Never Used                  137.2 102.6  75.7  62.7
##   College Degree/Professional Certificate:Never Used 223.5 167.2 123.4 102.2
##   Masters/Doctorate Degree:Never Used                 117.4  87.8  64.8  53.7
##                                                        y
## x                                                       55+
##   No College/University:Has Used                       3.2
##   Some College/University:Has Used                     6.1
##   College Degree/Professional Certificate:Has Used     5.5
##   Masters/Doctorate Degree:Has Used                    1.6
##   No College/University:Never Used                    12.0
##   Some College/University:Never Used                  23.7
##   College Degree/Professional Certificate:Never Used 38.6
##   Masters/Doctorate Degree:Never Used                 20.3
```

With an extremely small P-value we reject our null hypothesis, and conclude that there is strong evidence that either Education or Heroin or both differ between Age ranges. We need to include education level in our Age vs. Heroin use analysis.

Conditional Independence

```
I <- n_distinct(heroin_used_df$Education)
J <- n_distinct(heroin_used_df$Heroin)
K <- n_distinct(heroin_used_df$Age)


partial_chisq_tests_Her <-
  heroin_used_df |>
  # Group by the control variable, Z

  group_by(Age) |>

  # Calculating the test statistic, df, and p-value for each individual partial table
  summarize(test_stat = chisq_test(Heroin, Education)$statistic,
            df = chisq_test(Heroin, Education)$df,
            p_val = chisq_test(Heroin, Education)$p)

cond_ind_test_Her <- partial_chisq_tests_Her$test_stat |> sum()

cond_ind_test_Her
```

```
## [1] 66.90239
```

```
df<- (I*J*K) - 1 -((I + J -1)*K) - 1
df
```

```
## [1] 13
```

```
# P-value:
pchisq(q = cond_ind_test_Her,
       df = df,
       lower = F)
```

```
## [1] 2.968721e-09
```

With a test statistic of 66.9 and 13 degrees of freedom the resulting P-value is (2.9*10^-9). With a P-value < .05 we reject our null hypothesis that the relationship between Heroin use and Education is conditional on Age range. We have strong evidence that the odds ratios for Heroin use by Education at different Age ranges are unequal to 1.

Homogeneous Test

```
#make education variable binary so we can run Homogeneous test
heroin_used_df <- mutate(heroin_used_df, college= ifelse(Education %in% c("Masters/Doctorate Degree","College Degree/Professional Certificate"), "Graduated", "Didn't Graduate"))

#view odds ratios
heroin_used_df |>
  group_by(Age) |>
  summarize(odds_ratio = epitools::oddsratio(table(college, Heroin), rev = "col")$measure[2,1])
```

```
## # A tibble: 5 × 2
##   Age   odds_ratio
##   <chr>      <dbl>
## 1 18-24      0.527
## 2 25-34      0.301
## 3 35-44      0.339
## 4 45-54      0.447
## 5 55+        1.52
```

```
#run Breslow Day test for association between heroin use and age at different levels of education
drug_data_BDtest <-
  xtabs(formula = ~ Heroin + college + Age,
        data = heroin_used_df) |>

  BreslowDayTest()

drug_data_BDtest
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  xtabs(formula = ~Heroin + college + Age, data = heroin_used_df)
## X-squared = 6.8955, df = 4, p-value = 0.1415
```

We fail to reject the null and conclude we do not have strong evidence in favor of the alternative. We do not have strong evidence that the odds ratios for Heroin use by Education at different Age ranges are unequal. When we examine the odds ratios we can see they are all pretty similar expect for the outlier in the 55+ group although we believe this is due to the issues regarding the expected counts.

Meth

Complete Independence

```
#Getting Counts and Expected Values:
meth_sum <-
  meth_used_df |>
  count(Age, Education, Meth) |>
  # Calculating the proportions: n/sum(n)
  mutate(FA_prop = n/sum(n))

I <- n_distinct(meth_used_df$Education)
J <- n_distinct(meth_used_df$Meth)
K <- n_distinct(meth_used_df$Age)

meth_CI <-
  meth_sum |>
  group_by(Age) |>
  mutate(age_n = sum(n)) |>
  ungroup() |>

  group_by(Education) |>
  mutate(edu_n = sum(n)) |>
  ungroup() |>

  group_by(Meth) |>
  mutate(meth_n = sum(n)) |>
  ungroup() |>

  # Now we can calculate the expected proportion for each outcome assuming complete inde
pendence
  mutate(CI_prop = age_n/sum(n) * edu_n/sum(n) * meth_n/sum(n)) |>

  # Dropping the count columns because we don't need them in the future:
  select(-age_n, -edu_n, -meth_n)


#Getting Test Statistics:
CI_FA_test <-
  meth_CI |>
  # Calculating the individual pieces of our test statistics (chi^2 and G)
  mutate(zi2 = (FA_prop - CI_prop)^2/CI_prop,
         gi = n*log(FA_prop/CI_prop)) |>

  # Adding the individual pieces to get the test statistic:
  summarize(chi2 = sum(n)*sum(zi2),
            lrt_g = 2*sum(gi)) |>

  # Changing the results from being stored in separate columns to the same column
  pivot_longer(cols = chi2:lrt_g,
               names_to = "test",
               values_to = "stat")


# Calculating P-Values:
# The number of unique proportions needed for the FA model
```

```
r1 <- I*J*K - 1

# The number of unique proportions needed for the CI model
r0 <- I + J + K - 3

# The degrees of freedom: r1 - r0
df_CI <- r1 - r0
df_CI
```

```
## [1] 31
```

```
CI_FA_test |>
  mutate(p_val = pchisq(stat, df = df_CI, lower = F))
```

```
## # A tibble: 2 × 3
##   test   stat    p_val
##   <chr> <dbl>    <dbl>
## 1 chi2   564. 4.19e-99
## 2 lrt_g  536. 2.12e-93
```

```
#Checking Sample Size:
# meth_CI |>
#   mutate(n_CI = sum(n)*CI_prop) |>
#   arrange(n_CI)
```

We calculated a p-value less than .05, so we reject the null and conclude in favor of the alternative hypothesis. We have strong evidence that at least 2 of our three variables (age, education, meth use) are associated. This is expected given our initial analysis of meth use and these variables and both tests resulting in a conclusion of association.

Joint Independence

```
meth_JI <-
  meth_used_df |>
  mutate(edu_use = interaction(Education, Meth, sep = ":"))

meth_JI |>
  head(n = 10)
```

```
##      Age Gender                                 Education Country   Nscore
## 1  25-34      M               Masters/Doctorate Degree      UK -0.67825
## 2  35-44      M College Degree/Professional Certificate      UK -0.46725
## 3  18-24      F               Masters/Doctorate Degree      UK -0.14882
## 4  35-44      F               Masters/Doctorate Degree      UK  0.73545
## 5    55+      F                  No College/University  Canada -0.67825
## 6  45-54      M               Masters/Doctorate Degree     USA -0.46725
## 7  35-44      M                  No College/University      UK -1.32828
## 8  35-44      F College Degree/Professional Certificate  Canada  0.62967
## 9    55+      M               Masters/Doctorate Degree      UK -0.24649
## 10 25-34      F College Degree/Professional Certificate      UK -1.05308
##      Escore   Oscore   AScore   Cscore Impulsive       Meth
## 1   1.93886  1.43533  0.76096 -0.14277  -0.71126   Has Used
## 2   0.80523 -0.84732 -1.62090 -1.01450  -1.37983 Never Used
## 3  -0.80615 -0.01928  0.59042  0.58489  -1.37983 Never Used
## 4  -1.63340 -0.45174 -0.30172  1.30612  -0.21712 Never Used
## 5  -0.30033 -1.55521  2.03972  1.63088  -1.37983 Never Used
## 6  -1.09207 -0.45174 -0.30172  0.93949  -0.21712 Never Used
## 7   1.93886 -0.84732 -0.30172  1.63088   0.19268 Never Used
## 8   2.57309 -0.97631  0.76096  1.13407  -1.37983 Never Used
## 9   0.00332 -1.42424  0.59042  0.12331  -1.37983 Never Used
## 10  0.80523 -1.11902 -0.76096  1.81175   0.19268 Never Used
##                                              edu_use
## 1                  Masters/Doctorate Degree:Has Used
## 2  College Degree/Professional Certificate:Never Used
## 3                Masters/Doctorate Degree:Never Used
## 4                Masters/Doctorate Degree:Never Used
## 5                   No College/University:Never Used
## 6                Masters/Doctorate Degree:Never Used
## 7                   No College/University:Never Used
## 8  College Degree/Professional Certificate:Never Used
## 9                Masters/Doctorate Degree:Never Used
## 10 College Degree/Professional Certificate:Never Used
```

```
JI_vs_FA <-
  chisq_test(x = meth_JI$edu_use,
             y = meth_JI$Age)

JI_vs_FA
```

```
## # A tibble: 1 × 6
##       n statistic        p    df method          p.signif
## * <int>     <dbl>    <dbl> <int> <chr>           <chr>
## 1  1884      439. 2.62e-75    28 Chi-square test ****
```

```
#Checking Expected Counts:
expected_freq(JI_vs_FA) |>
  round(digits = 1)
```

```
##                                                        y
## x                                                       18-24 25-34 35-44 45-54
##   No College/University:Has Used                         26.6  19.9  14.7  12.2
##   Some College/University:Has Used                       65.2  48.8  36.0  29.8
##   College Degree/Professional Certificate:Has Used       47.1  35.2  26.0  21.5
##   Masters/Doctorate Degree:Has Used                      16.7  12.5   9.2   7.6
##   No College/University:Never Used                       61.1  45.7  33.7  27.9
##   Some College/University:Never Used                    107.5  80.4  59.4  49.2
##   College Degree/Professional Certificate:Never Used 208.5 156.0 115.1  95.3
##   Masters/Doctorate Degree:Never Used                   110.2  82.5  60.9  50.4
##                                                        y
## x                                                       55+
##   No College/University:Has Used                        4.6
##   Some College/University:Has Used                     11.3
##   College Degree/Professional Certificate:Has Used      8.1
##   Masters/Doctorate Degree:Has Used                     2.9
##   No College/University:Never Used                     10.5
##   Some College/University:Never Used                   18.6
##   College Degree/Professional Certificate:Never Used 36.0
##   Masters/Doctorate Degree:Never Used                  19.0
```

With a p-value less than .05, we reject the null and conclude that we have strong evidence in favor of the alternative. We have strong evidence that Education or Meth use or both differ among age range.

####Conditional Independence

```
I <- n_distinct(meth_used_df$Education)
J <- n_distinct(meth_used_df$Meth)
K <- n_distinct(meth_used_df$Age)


partial_chisq_tests_Meth <-
  meth_used_df |>
  # Group by the control variable, Z

  group_by(Age) |>

  # Calculating the test statistic, df, and p-value for each individual partial table
  summarize(test_stat = chisq_test(Meth, Education)$statistic,
            df = chisq_test(Meth, Education)$df,
            p_val = chisq_test(Meth, Education)$p)


cond_ind_test_Meth <- partial_chisq_tests_Meth$test_stat |> sum()
cond_ind_test_Meth
```

```
## [1] 86.08701
```

```
df<-(I*J*K) - 1 -((I + J -1)*K) - 1
df
```

```
## [1] 13
```

```
# P-value:
pchisq(q = cond_ind_test_Meth,
       df = df,
       lower = F)
```

```
## [1] 7.789163e-13
```

With a test statistic of 86.09 and 13 degrees of freedom the resulting P-value is ~0. With a P-value < .05 we reject our null hypothesis that the relationship between Meth use and Education is conditional on Age range. We have strong evidence that the odds ratios for Meth use by Education at different Age ranges are unequal to 1.

Homogeneous Test

```
#make education variable binary so we can run homogeneous test
meth_used_df <- mutate(meth_used_df, college= ifelse(Education %in% c("Masters/Doctorate
Degree","College Degree/Professional Certificate"), "Graduated", "Didn't Graduate"))

#view odds ratios
meth_used_df |>
  group_by(Age) |>
  summarize(odds_ratio = epitools::oddsratio(table(college, Meth), rev = "col")$measure
[2,1])
```

```
## # A tibble: 5 × 2
##   Age   odds_ratio
##   <chr>      <dbl>
## 1 18-24      0.458
## 2 25-34      0.298
## 3 35-44      0.479
## 4 45-54      0.524
## 5 55+        0.802
```

```
#run Breslow Day test for association between drug use and age at different levels of ed
ucation
drug_data_BDtest <-
  xtabs(formula = ~ Meth + college + Age,
        data = meth_used_df) |>

  BreslowDayTest()

drug_data_BDtest
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  xtabs(formula = ~Meth + college + Age, data = meth_used_df)
## X-squared = 4.4953, df = 4, p-value = 0.3431
```

After running the Breslow Day Test, we computed a chi-squared value of 4.49 with df = 4, which resulted in a p-value of .343. We fail to reject the null and conclude we do not have strong evidence in favor of the alternative. We do not have strong evidence that the odds ratios for meth use by education at different levels of age are unequal. Again we some slight variation amongst odds ratios but the biggest gap coming from that 55+ group.

Logistic Regression - Looking at Meth Use

Moving on to the final analysis of the data, we wanted to further explore the relationships by using some logistic regression techniques.

```
# 0 = Never used, 1 = Has used
meth_log_reg <- meth_used_df |>
  mutate(use = ifelse(Meth=="Never Used",0,1)) |>
  select(-Meth)
```

Looking at How Age and Education Affect Meth Use

```
add_model <-
  glm(formula = use ~ Age + Education,
      family = binomial,
      data = meth_log_reg)

int_model <-
  glm(formula = use ~ Age * Education,
      family = binomial,
      data = meth_log_reg)

fit_stats <-
  bind_rows(
    "add" = broom::glance(add_model),
    "int" = broom::glance(int_model),
    .id = "model"
    )

fit_stats
```

```
## # A tibble: 2 × 9
##   model null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##   <chr>         <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1 add           2085.    1883  -983. 1983. 2027.    1967.        1876  1884
## 2 int           2085.    1883  -970. 1980. 2091.    1940.        1864  1884
```

```
c("test stat" = fit_stats$deviance |> diff() |> abs(),
  "p-value" = pchisq(q = fit_stats$deviance |> diff() |> abs(),
                     df = fit_stats$df.residual |> diff() |> abs(),
                     lower = F))
```

```
##    test stat      p-value
## 26.791711367  0.008278565
```

```
age_model <-
  glm(formula = use ~ Age,
      family = binomial,
      data = meth_log_reg)

# Now we can use anova just by giving it multiple models from simplest to the most compl
icated:
anova(age_model, add_model, int_model,
      test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: use ~ Age
## Model 2: use ~ Age + Education
## Model 3: use ~ Age * Education
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1879     2023.6
## 2      1876     1966.5  3   57.053 2.504e-12 ***
## 3      1864     1939.8 12   26.792  0.008279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#since some sample sizes are small test AICc
k_add = 7
AIC_add = 1982.54
AICc_add = AIC_add + (2*(k_add^2 + k_add))/(1884-k_add-1)
k_int = 19
AIC_int = 1979.748
AICc_int = AIC_int + (2*(k_int^2 + k_int))/(1884-k_int-1)
diff = AICc_add - AICc_int

AICc_add
```

```
## [1] 1982.6
```

```
AICc_int
```

```
## [1] 1980.156
```

```
diff
```

```
## [1] 2.443976
```

As we can see from the results of the model comparisons and tests, the interaction term is needed in the model. However when we look at the decrease in AIC it doesn't improve by much and further, due to the lower sample size we decided to also examine the AICc scores which still showed improvement however the decrease/difference in AICc was halved from roughly 4 down to 2.

Stepwise Model Selection

We finally decided we would try stepwise selection on our altered dataset for Meth use and see which features would be selected as the best predictors for usage.

```
min_model <-
  glm(formula = use ~ 1,  # 1 means intercept only
      family = binomial,
      data = meth_log_reg)
max_model <-
  glm(formula = use ~ .,
      family = binomial,
      data = meth_log_reg)

#Forward Selection of features
forward_glm <-
  MASS::stepAIC(object = min_model,
                direction = "forward",
                scope = formula(max_model),
                trace = 0)

#Backward Selection of features
backward_glm <-
  MASS::stepAIC(object = min_model,
                direction = "forward",
                scope = formula(max_model),
                trace = 0)

#Results from forward, backward, and both selection
both_glm <-
  MASS::stepAIC(object = min_model,
                direction = "both",
                scope = formula(max_model),
                trace = 0)

c("forward"  = forward_glm$formula,
  "backward" = backward_glm$formula,
  "both"     = both_glm$formula)
```

```
## $forward
## use ~ Country + Cscore + Gender + Nscore + Oscore + AScore +
##     Education + Impulsive + Escore
##
## $backward
## use ~ Country + Cscore + Gender + Nscore + Oscore + AScore +
##     Education + Impulsive + Escore
##
## $both
## use ~ Country + Cscore + Gender + Nscore + Oscore + AScore +
##     Education + Impulsive + Escore
```

```
#suggested model
sugg_model <- glm(formula = use ~ Country + Cscore + Education + Oscore + AScore + Nscor
e +
    Impulsive + Escore,
      family = binomial,
      data = meth_log_reg)

#summary(sugg_model)
#anova(sugg_model)
tidy(sugg_model)
```

```
## # A tibble: 16 × 5
##    term                                          estim…¹ std.e…² stati…³ p.value
##    <chr>                                           <dbl>   <dbl>   <dbl>   <dbl>
##  1 (Intercept)                                    -0.871  0.354   -2.46  1.38e-2
##  2 CountryCanada                                   0.179  0.415    0.431 6.66e-1
##  3 CountryNew Zealand                              0.637  1.01     0.630 5.29e-1
##  4 CountryOther                                   -0.272  0.405   -0.672 5.01e-1
##  5 CountryRepublic of Ireland                     -0.139  0.663   -0.209 8.34e-1
##  6 CountryUK                                      -0.806  0.349   -2.31  2.08e-2
##  7 CountryUSA                                      1.14   0.345    3.30  9.69e-4
##  8 Cscore                                         -0.155  0.0749  -2.06  3.92e-2
##  9 EducationSome College/University               -0.488  0.197   -2.48  1.32e-2
## 10 EducationCollege Degree/Professional Certifi…  -0.604  0.193   -3.13  1.73e-3
## 11 EducationMasters/Doctorate Degree              -0.721  0.236   -3.05  2.27e-3
## 12 Oscore                                          0.320  0.0720   4.45  8.61e-6
## 13 AScore                                         -0.206  0.0654  -3.15  1.61e-3
## 14 Nscore                                          0.155  0.0726   2.13  3.31e-2
## 15 Impulsive                                       0.151  0.0755   2.00  4.55e-2
## 16 Escore                                         -0.135  0.0736  -1.83  6.67e-2
## # … with abbreviated variable names ¹estimate, ²std.error, ³statistic
```

```
predictions <- predict(sugg_model,
        newdata = meth_log_reg,
        type = "response") |>

  round(digits = 3) #|>

  #data.frame()

pred_df <- meth_log_reg |>
  mutate(pred_used = ifelse(predictions>.5,1,0),
         correct_pred = ifelse(use==pred_used,1,0))


#overall accuracy
model_acc <- sum(pred_df$correct_pred)/length(pred_df$correct_pred)
model_acc
```

```
## [1] 0.7935244
```

```
#accuracy predicting use
# aggregate(correct_pred~use,data=pred_df,FUN = sum)
correct_use_pred = 193
total_num_use = sum(pred_df$use==1)
correct_pred_acc = correct_use_pred/total_num_use
correct_pred_acc
```

```
## [1] 0.4232456
```

```
broom::glance(sugg_model)
```

```
## # A tibble: 1 × 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         2085.    1883  -817. 1667. 1755.    1635.        1868  1884
```

```
anova(add_model, int_model, sugg_model,
      test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: use ~ Age + Education
## Model 2: use ~ Age * Education
## Model 3: use ~ Country + Cscore + Education + Oscore + AScore + Nscore +
##     Impulsive + Escore
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1876     1966.5
## 2      1864     1939.8 12   26.792 0.008279 **
## 3      1868     1634.7 -4  305.095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After trying the suggested model from the stepwise model selection and doing some analysis we can see that that accuracy of the model is ~80% with an accuracy of predicting use being ~42%. The AIC dropped significantly however from looking at the anova results it looks like the difference between the interaction model from before and the suggestion model is not significant enough. We assume this might either be due to potential overfitting or maybe that the model is lacking interaction terms and is only additive of the selected features.

Conclusion

Based on the results we have seen from our analysis in both instances of Heroin and Meth use we ended up rejecting the null hypothesis when comparing more complex models until we reached the homogeneous test. This led us to conclude in these cases that the odds ratios for Drug use (used or not) by Education (graduated college/professional cert or not) were equivalent across Age ranges, which was a somewhat interesting discovery. We found some interesting results towards the end when looking at logistic regression methods, especially the model from the stepwise selection with the personality traits included. From an initial glance it looked like Openness to Experience, Neuroticism, and Impulsiveness increased the chances of being a user while Agreeableness and Extroversion seemed to reduce the chances slightly. This surprised us a bit as we would have expected an increased level of Agreeableness and Extroversion to increase the chances someone might come in contact with a drug like Meth and potentially be willing to try it. Of course it is important to note that due to the size of the dataset and that low amount of elderly people represented, some of the expected counts for categories involving older folks were lower than 5. This could have impacted the results of tests and interpretations but due to certain limitations as well as the way age brackets were arranged we focused on doing the analysis we found appropriate if the conditions were ideal. In terms of next steps, ways to improve the project from here for ourselves, and general ideas, we thought it would be even better and more interesting to look at data focused on Burlington or the state of Vermont. This could potentially be available from state data records or could possibly be obtained from an organization fro the Howard center if they happened to have collected data on that, either way it would be very interesting and intriguing. One more realistic method we discussed with more time would be searching for possible additions to this dataset (if this is only a piece of the whole study/set) or finding other related datasets regarding drug use that have more samples. More data will always help increase the accuracy of tests as well as reduce the overfitting of models. This is an important field to study and collect data on as the issues with drug use in our society increase. In this project we have explored some of the statistical relationships between Meth/Heroin usage, Education, and Age, as well as a few personality traits briefly, and found that there do exists associations between these features beyond simple associations and they should be studied further with more data.