Topic and Data Selection

Nicole Sanchez Flores, Kimberly By Gotia

Group Members: Nicole Sanchez Flores, Kimberly By Gotia

Data:

- The data that we will be using comes from the Health Evaluation and Linkage to Primary Care (HELP) study which was a clinical trial for adults receiving in-patient care at a detoxification unit. If patients did not have a primary care physician then they were randomized with the goal of linking to primary medical care. This clinical research data was approved by Institutional Review Board of Boston University Medical Center and is housed in the mosaic RStudio package. For our research focus we will be utilizing the explanatory variables a1 (gender represented by 1 = male and 2 = female), age, homeless (related to homeless status with 0 = no and 1 = yes), and ces_d (center for epidemiologic studies depression measure 0-60). Our chosen variables will help us determine the potential correlation between these factors and the chosen first drug of choice (prim_sub) for an individual in the detoxification unit.

Research Question/Purpose:

Does age, gender, depression score, and homelessness predict primary substance type? Exposures: Age, a1 (gender represented by 1 = male and 2 = female), homeless (related to homeless status with 0 = no and 1 = yes), and ces_d (center for epidemiologic studies depression measure 0-60) Outcomes: The first substance an individual at the detoxification center has taken (prim_sub).

Descriptions of variables:

- Depression levels (ces_d)

  - Depression levels in patients in the detoxification center are measured using the Center for Epidemiological Studies depression measure which ranges from 0-60. Patients with higher scores are deemed to have more depressive symptoms.

- Primary substance (prim_sub)
  - The primary substance in this data set refers to the first drug of choice for the patient in the detoxification center. The drugs are categorized by levels with 0 = none, 1 = alcohol, 2 = cocaine, 3 = heroin, 4 = barbiturates, 5 = benzos, 6 = marijuana, 7 = methadone, and 8 = opiates.

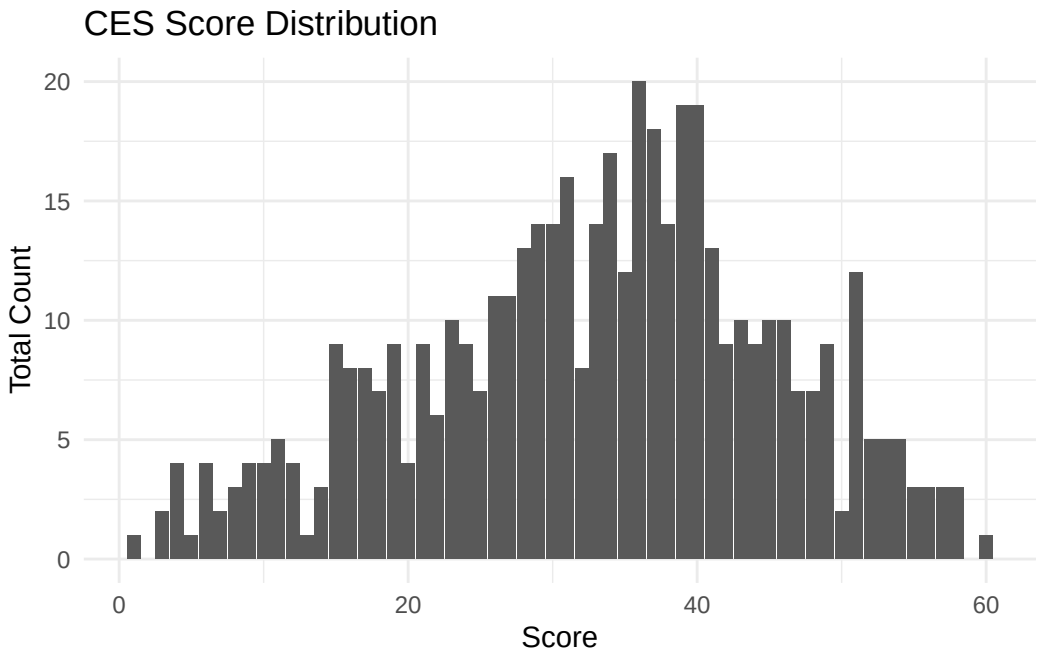
- Age (age)
  - Age is coded as a continuous numeric measure, and it represents the age the participant was at the time of entering the rehab service.

- Gender (a1)
  - Gender is a binary categorical variable in this dataset, where 1 = male and 2 = female. This variable helps us explore whether primary substance choice or depression levels can differ between men and women in the HELP study, among other explorations.

- Homelessness (homeless)
  - Homeless is a binary categorical variable which indicates whether the participant was experiencing homelessness at the time of the study. 0 indicates no homelessness while 1 indicates homelessness

Exploration of Data:

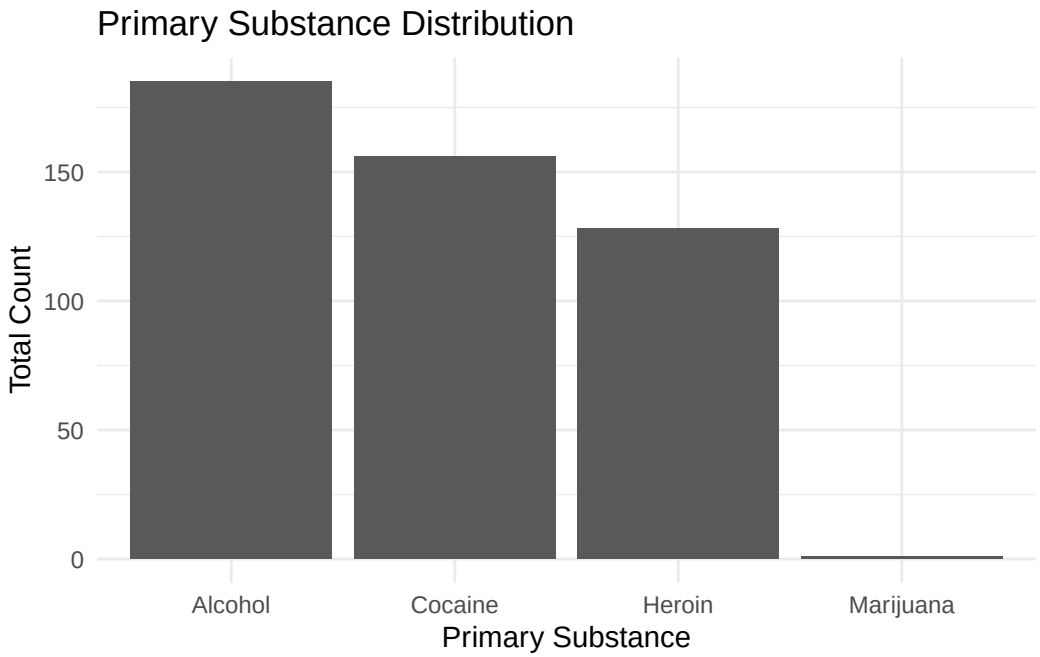
Depression levels (ces_d) general distribution count:




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00   25.00   34.00   32.86   41.00   60.00
```

- Findings: The CES depression scores tend to peak towards the 30-40 score range. This indicates high frequency of mid-range depression symptoms among the population in the HELP data set. This mid-range score finding is also supported by the summary statistics for CES-D scores which indicate a mean of 32.86.

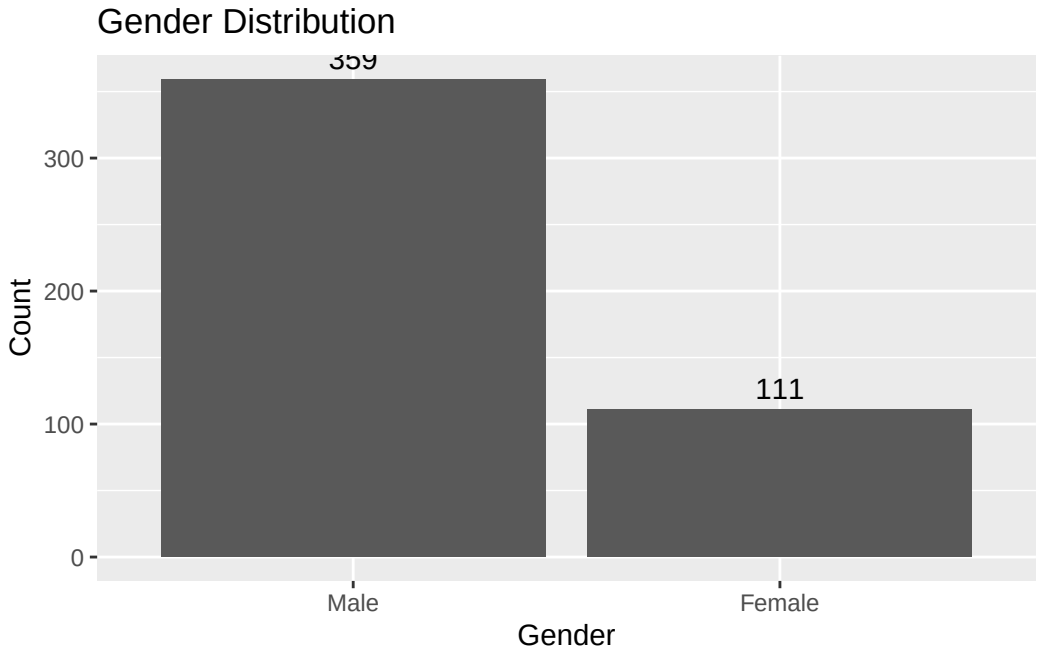
Primary substance (prim_sub) general distribution count:




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   1.887   3.000   6.000
```

```
# A tibble: 4 x 2
# Groups:   prim_sub [4]
  prim_sub     n
     <dbl> <int>
1        1   185
2        2   156
3        3   128
4        6     1
```

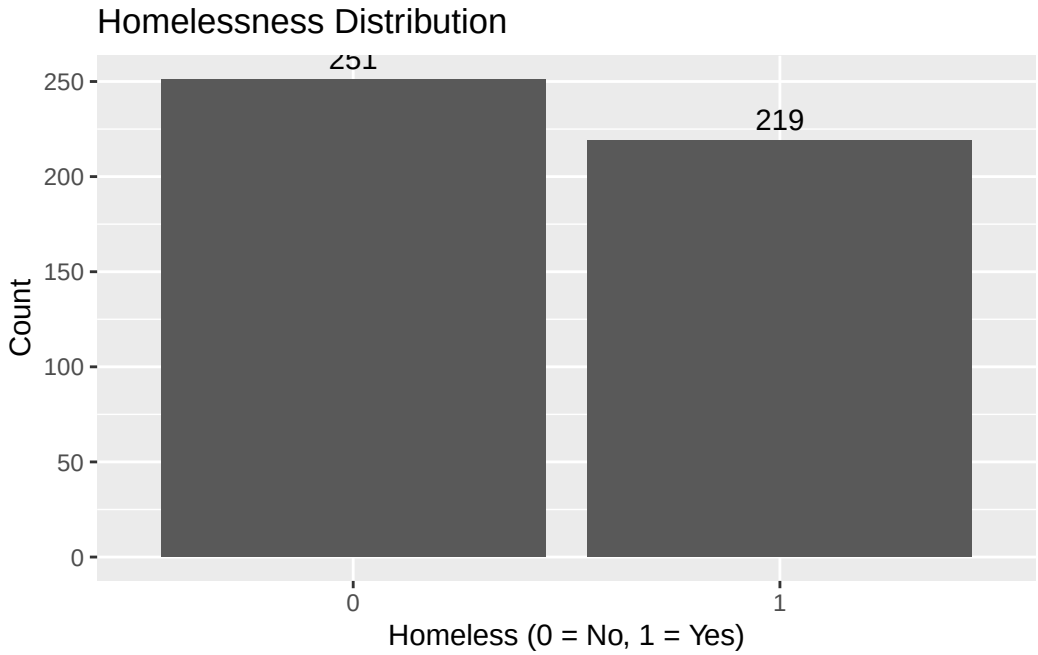
Gender Distribution




```
# A tibble: 2 x 2
# Groups:   gender_label [2]
  gender_label     n
  <fct>        <int>
1 Male           359
2 Female         111
```

- Findings: From the bar graph and from the counts we can see that the majority of the participants are male (n = 359) and a smaller proportion of the particpants are female (n = 111). This implies that men use rehab services more frequently in this sample, which is consistent with larger national trends in admissions to drug treatment programs, as we explored in our research. This is important to note when investigating associations with primary substance type, since gender representation may influence both substance use trends and mental health characteristics in the dataset.

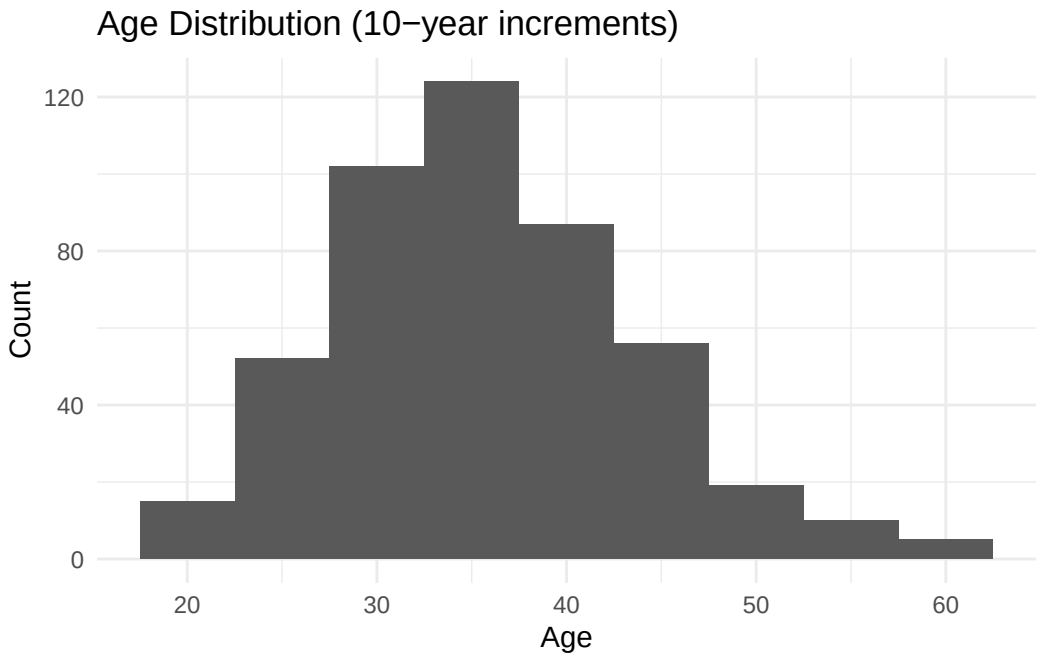
Homelessness

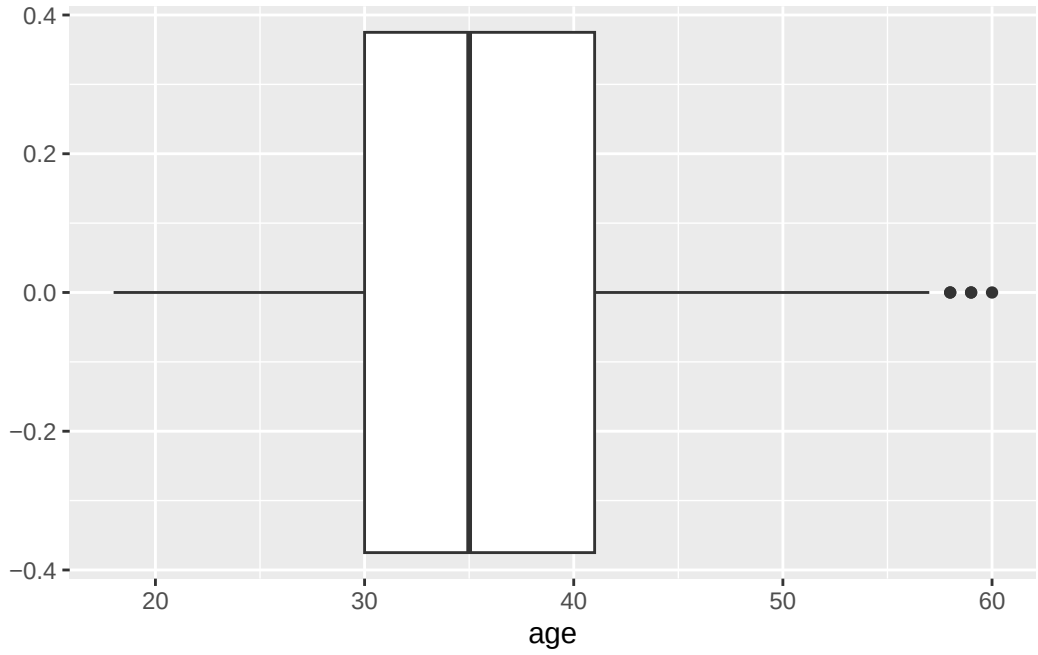



```
# A tibble: 2 x 2
# Groups:   home_label [2]
  home_label     n
  <fct>      <int>
1 No           251
2 Yes          219
```

- Findings: We can see that the homelessness status in the sample seems to be evenly divided, with 251 participants reporting they were not homeless and 219 reporting homelessness. With a relatively high proportion of participants identifying as homeless, it can lead to further investigation for homelessness as an explanatory variable. It may reflect the population's heightened susceptibility to substance use disorders.

Age



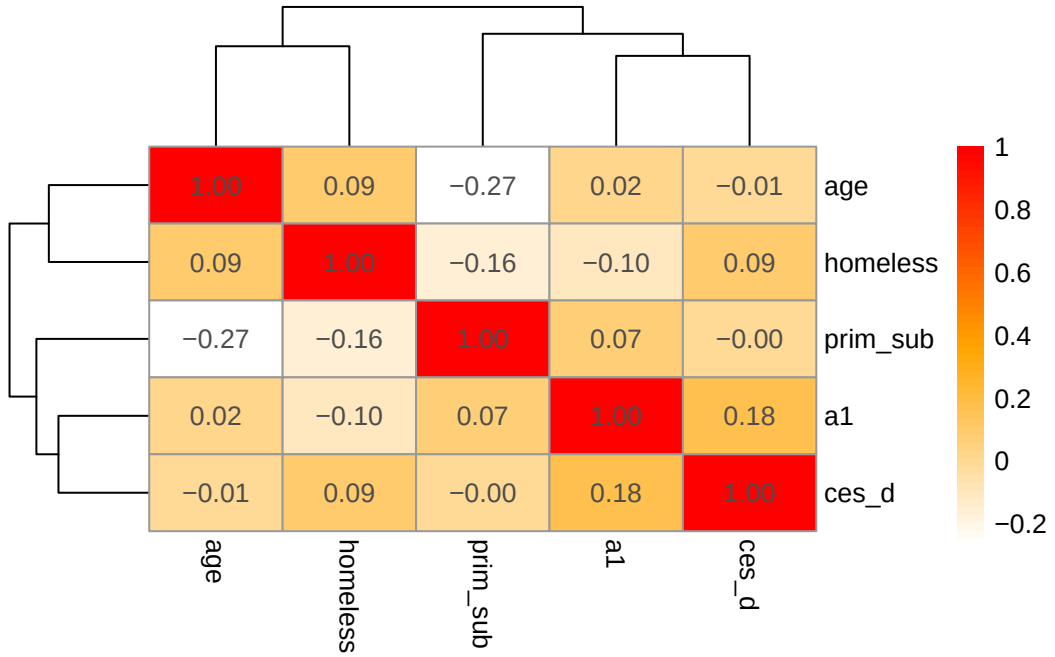



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   30.00   35.00   35.75   41.00   60.00
```

- Findings: The age distribution of participants falls mostly between the ages of 25 and

45, with fewer people at the younger (18-24) and older (55-64) sides of the range. The median age is 35, while the mean age is 35.75, indicating a somewhat right skewed distribution. This shows that the majority of people seeking rehab services are in their early to mid-adult years, which may have an impact on the link between age and primary substance type.

Correlation Matrix



- *Findings:*

- Age and primary substance (r −0.27): Older participants tend to use substances that have lower numeric codes in the data (e.g., more alcoho, since alcohol is coded as 1l vs. cocaine, coded as 2), although this requires more careful interpretation due to coding.
- Gender and CES-D (r 0.18): Females appear to have higher depression scores on average.
- Homelessness and primary substance (r −0.16): Homelessness status shows moderate association with substance type.

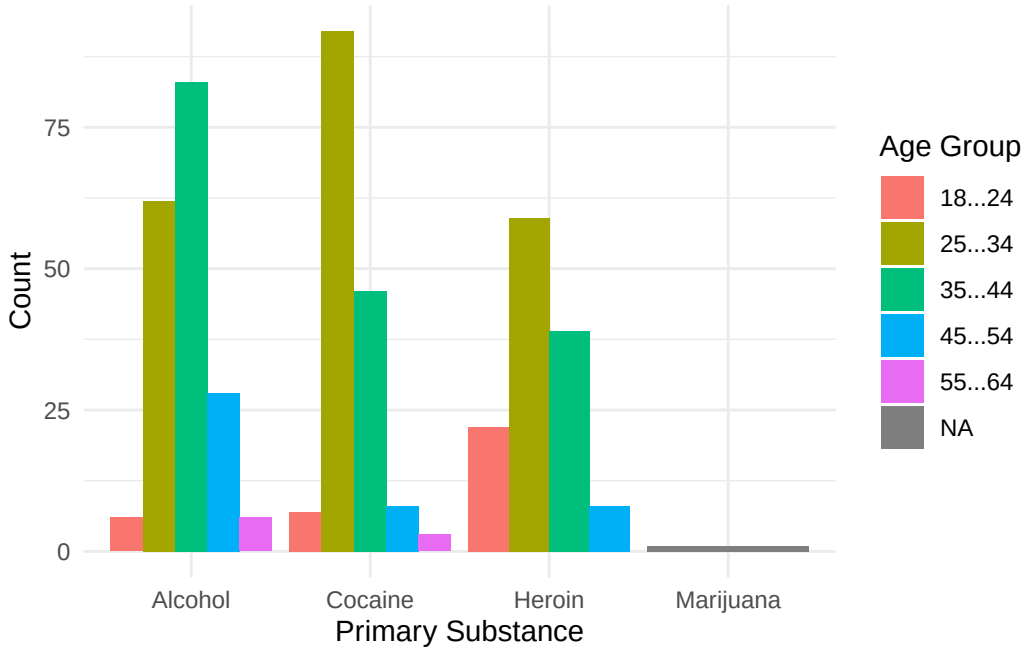
Table 1: Correlation Matrix (r)

	age	a1	prim_sub	homeless	ces_d
age	1.0000000	0.0229102	-0.2731423	0.0862318	-0.0082851
a1	0.0229102	1.0000000	0.0695927	-0.0976233	0.1774121

	age	a1	prim_sub	homeless	ces_d
prim_sub	-0.2731423	0.0695927	1.0000000	-0.1610558	-0.0027732
homeless	0.0862318	-0.0976233	-0.1610558	1.0000000	0.0900843
ces_d	-0.0082851	0.1774121	-0.0027732	0.0900843	1.0000000

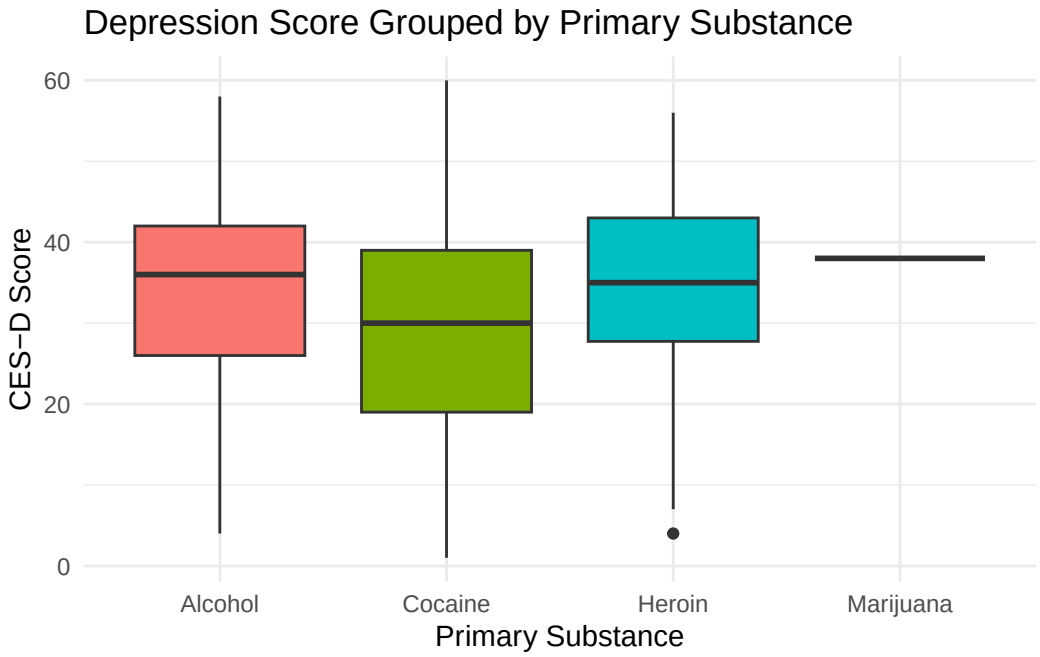
Table 2: P-values

	age	a1	prim_sub	homeless	ces_d
age	NA	0.6203006	0.0000000	0.0617671	0.8578267
a1	0.6203006	NA	0.1319280	0.0343581	0.0001103
prim_sub	0.0000000	0.1319280	NA	0.0004562	0.9521868
homeless	0.0617671	0.0343581	0.0004562	NA	0.0509678
ces_d	0.8578267	0.0001103	0.9521868	0.0509678	NA



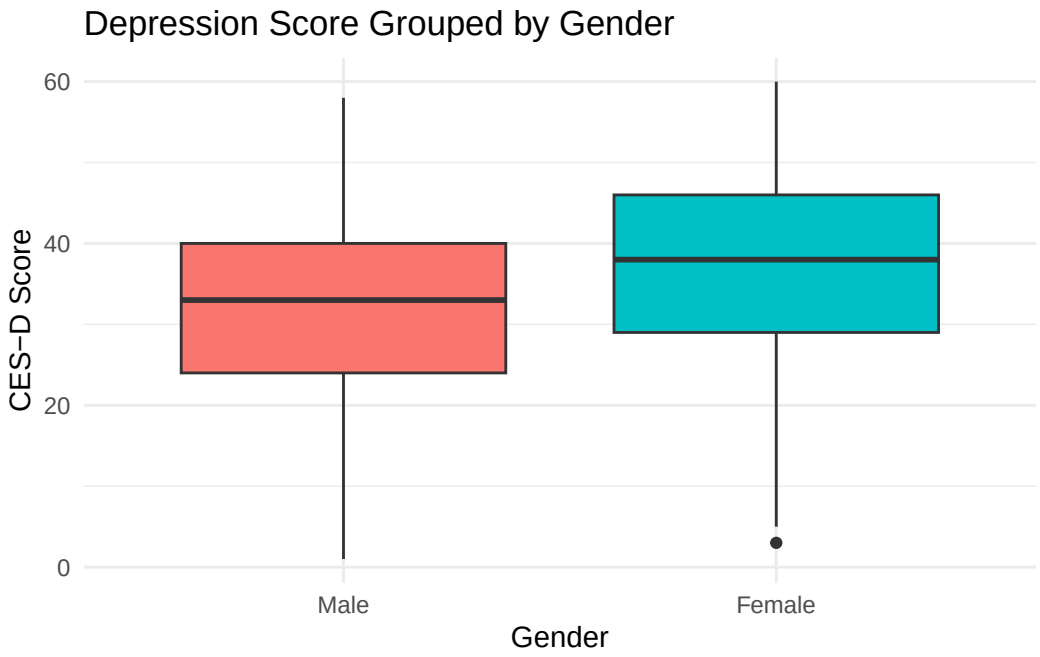
- Findings: Primary substance use patterns vary by age group. Alcohol use is most common in all age groups, but it's particularly prevalent among those aged 35-44 and 45-54. Cocaine and heroin use appear to be more evenly distributed across age groups, but younger groups (18-24 and 25-34) having slightly higher cocaine use. Marijuana use is extremely low in all age groups (there seems to just be one instance of it). These patterns indicate that age may have a significant influence on the primary substance use variable, which reinforces the need to include age as a predictor in our multinomial model.

Depression Score in association with Primary Substance:



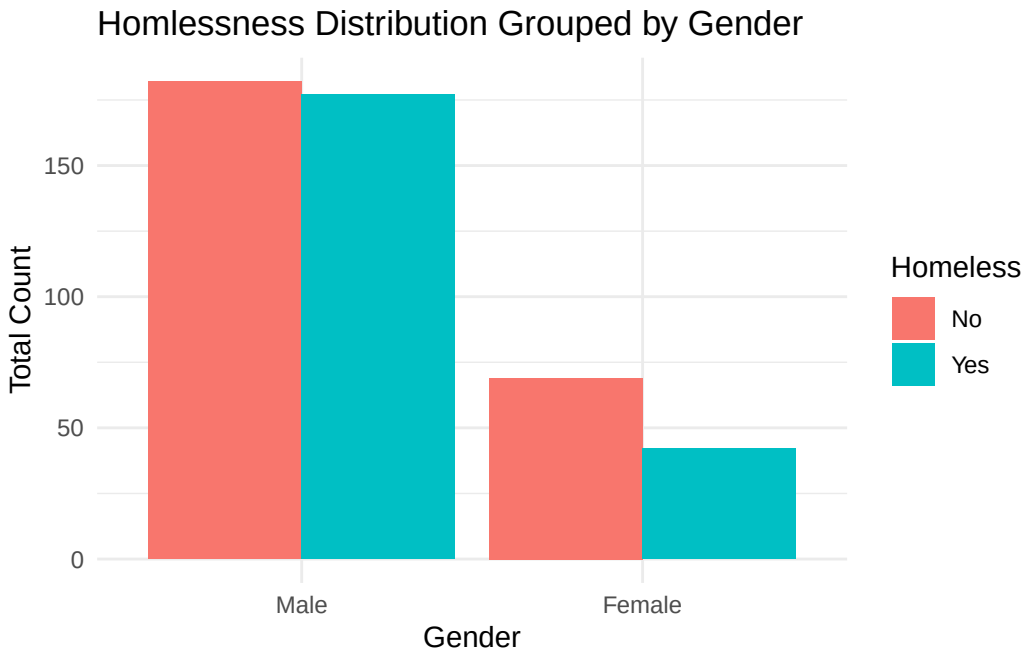
- Findings: After grouping depression score by the different primary substances used by the detoxification patients we found that alcohol tended to have a higher mean depression score (>35) compared to cocaine and heroine. We also saw that in terms of variance cocaine had a wider range of variance extending to the highest score of 60 indicating a high instance of depression symptoms.

Depression Score in association with Gender:



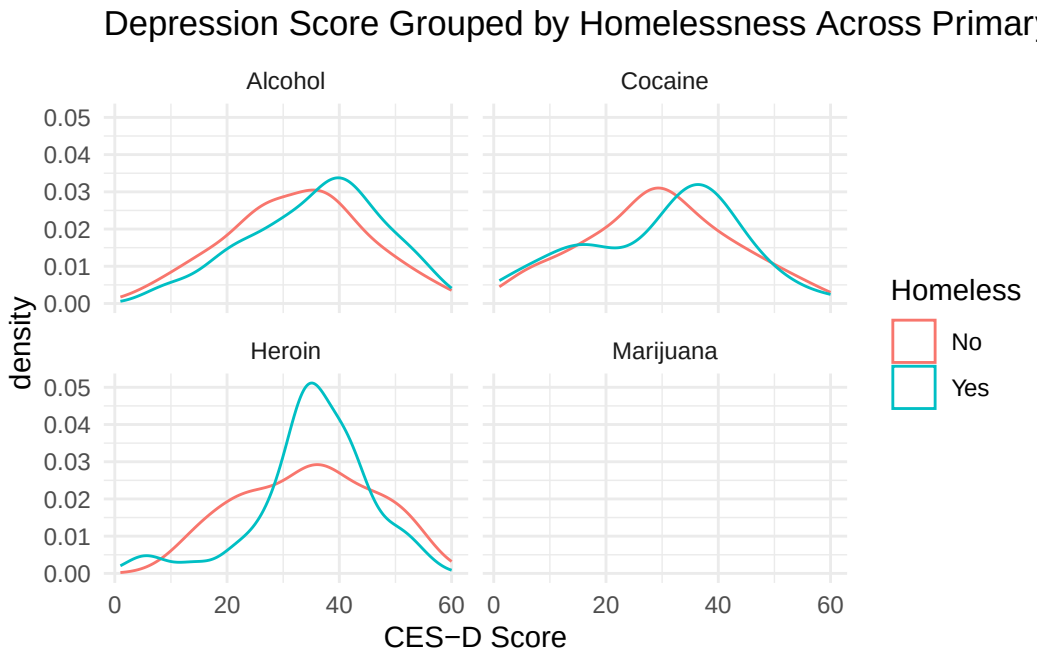
- Findings: When evaluating the depression levels of men versus women in the detoxification center we found that women had a higher depression scores overall. Women had a mean score of >35 compared to men who scored ~30.

Homelessness in association with Gender:



- Findings: From this bar plot we can see that for both categories of homelessness men tend to be more represented overall in the HELP data set. However, within the male population there is a close tie between those that are homeless and those that are not. For women this is not the case. Women who are not homeless are at a higher proportion than those that are in the detoxification center.

Depression Score and Homelessness grouped by Primary Substance:



- Findings: This density chart shows the interaction between homelessness and depression score across different primary substances. From the plot we found an interesting relationship among the heroin drug use group. Detoxification patients who use heroin and are homeless tend to have a higher population representation in scoring a mid-high range depression score (30-40). Meanwhile, those that used heroin and were not homeless were more spread across 20-60.

Outline of the Methodological Extension:

```
# A tibble: 470 x 5
     age    a1 prim_sub homeless ces_d
   <dbl> <dbl>    <dbl>    <dbl> <dbl>
 1    37     1        2        0    49
 2    37     1        1        1    30
 3    26     1        3        0    39
 4    39     2        3        0    15
 5    32     1        2        1    39
 6    47     2        2        0     6
 7    49     2        2        0    52
 8    28     1        1        1    32
 9    50     2        1        1    50
10    39     1        3        1    46
```

```
# i 460 more rows
```

1. Understanding drug use patterns among the homeless population: A systematic review of quantitative studies

    - Summary: Researchers conducted a systematic review of quantitative studies from 2007 and 2020 to assess the trends associated with substance abuse among the global adult homeless population. They found that alcohol was a highly popular primary drug abuse substance, but overtime there was an emergence in psychoactive substance use as well. Substance abuse was also found to be more common among men due to mental health and trauma stressors.

2. Differences in Drug Use among Persons Experiencing Homelessness According to Gender and Nationality

    - Summary: This paper focuses on a cross sectional study observing whether there is a statistical difference in drug consumption between homeless men and women in Spanish shelters. They also assess whether there is higher drug usage present among homeless Spanish nationals or immigrants. After conducting the study they found that there was not a statistically significant difference between men and women in their observed population in terms of their drug use. However, they found that the reason behind initial drug use did differ. Women tended to initially use drugs due to partner influence or lack of family affectivity while men initiated drug use because of personality or social factors. We found the difference in initial drug use between men and women to be very interesting especially as we examine whether factors like depression score influence an individual's primary substance in the HELP data.

    - Additionally, their investigation into whether or not there was a difference in drug risk among homeless Spanish nationals and immigrants found that Spanish nationals were statistically at a higher tendency of drug use compared to immigrants. This section of the research paper seemed compelling for future investigation of whether primary substance differs among homeless populations across the U.S depending on their geographic location. The HELP data does not account for geographic information at this time so for now this is more of a future expansion of our research.

3. Homelessness and Gender: Differences in Characteristics and Comorbidity of Substance Use Disorders at Admission to Services

    1. Summary: Study focuses on examining the associations between homelessness, gender, the severity of substance use, and the presence of mental health comorbidity among individuals entering treatment for SUD. After conducting a logistic regression on the 2017 Treatment Episodes dataset they found that individuals experiencing homelessness that admit into services have a higher usage rate of cocaine and meth, have higher frequency of use, and have higher rates of mental health comorbidity. Among their population they also found that women experiencing

homelessness were highly associated with having mental health comorbidities. This research study helped in answering a question that bubbled up from our reading of, "Differences in Drug Use among Persons Experiencing Homelessness According to Gender and Nationality" where they found different influential factors impact initial drug use between men and women. We believe that mental health could be a factor that influences primary substance use so finding a paper that observed higher mental health comorbidities among homeless women that use substances is helpful to our research.

We will both create different graphs required for understanding the variables in depth from the multinomial results. We will both focus on different statistics for the multinomial logistic regression. Nicole will focus on the multinomial logistic regression model, and Kim will focus on predicted probabilities to further understand the model. Some of the packages we will use include foreign, nnet, ggplot2, and reshape2.

We will use multinomial logistic regression since our outcome in this case is a categorical variable with no real order, such as 1 = Alcohol. This extension will help us to investigate how age, gender, depression levels, and homelessness influence the likelihood of selecting each primary substance.

- nnet: includes the multinom() function, which is what we need for fitting a multinomial logistic regression model. Since our outcome variable (prim_sub) has several unsorted categories (alcohol, cocaine, heroin, etc.), nnet is needed for calculating how age, gender, homelessness, and depression scores influence the probability of each substance category.
- reshape2: We will need reshape2 for predicted probablilities since it elps transform data for preparing tables or probability outputs that we will use from the multinomial model. The predicted probabilities for each substance category in prim_sub may need to be changed before we plot them.
- ggplot2: This package will be needed for visualizations so we can better understand how our model fits, or explore variables before the model. We will use it for bar plots, boxplots, histograms, and predicted-probability graphs.