# NSF_KBG_Data_Selection

Nicole Sanchez Flores, Kimberly By Gotia

- Group Members: The names of all individuals in the project group.
- Data: Describe the data that you plan to use and the place/site that you sourced it from.

We received the data from HELP

Health Evaluation and Linkage to Primary Care (HELP): The HELP study was a clinical trial for adults undergoing in-patient detoxication from alcohol, heroin, or cocaine. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking patients to primary medical care. The primary outcome of interest was attendance at a primary care appointment within 12 months, as well as the time until that linkage occurred. A rich collection of patient information (including demographic information; substance use history and frequency; number of prior hospitalizations; perceptions of physical and metal health; and medical literacy) was also collected during the baseline assessment.

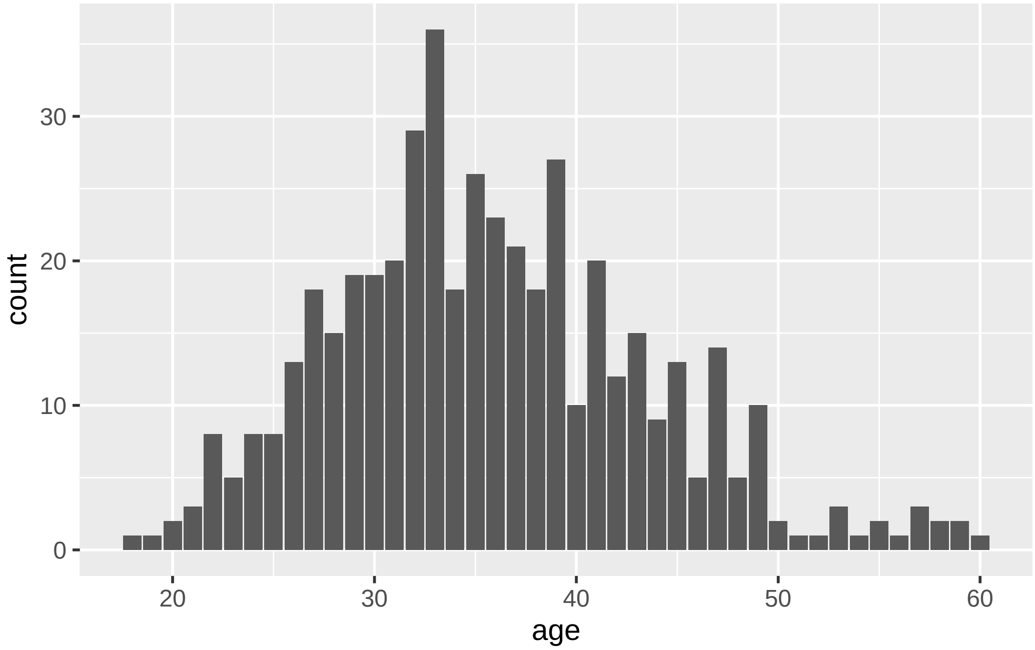
- Research Question/Purpose:

What specific research question(s) do you hope to answer/what specifi- cally do you hope to learn from analyzing these data? You should clearly identify the disease outcome and the exposure(s) of interest, as well as which variables in the data you will use to measure them.

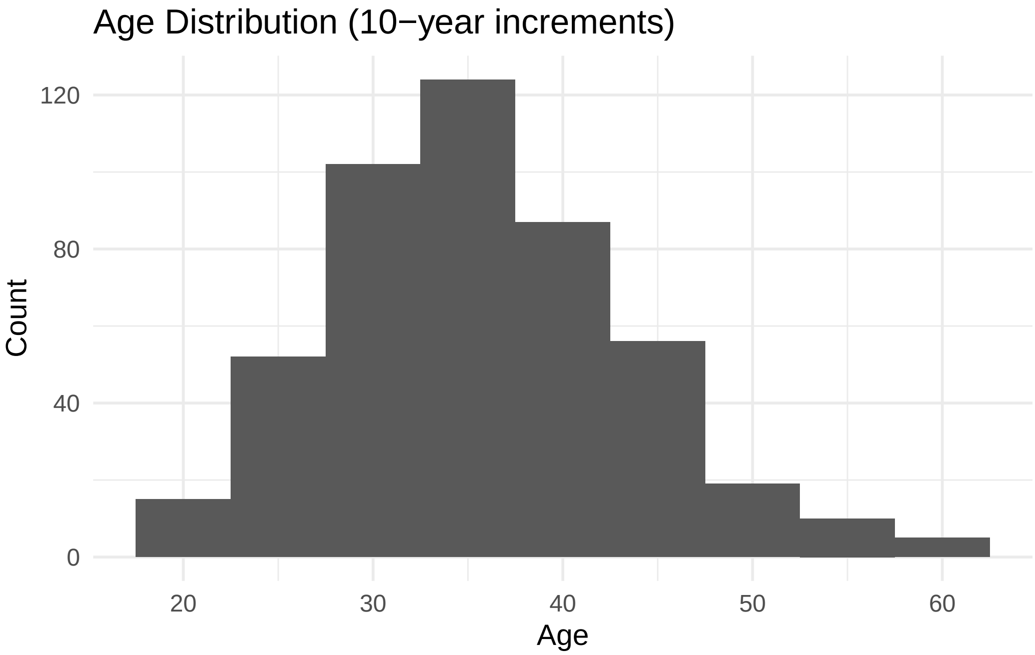
- Exploratory Data Analysis:

To support your argument that you can use this data to address your research question (and as proof that you can read in the data!), conduct a brief exploratory analysis of the dataset. You might consider reporting relevant summary statistics (such as the sample size, measures of center and spread for relevant continuous exposures, and contingency tables for categorical exposures and disease outcomes), examining possible missing data, or making visualizations of the data.

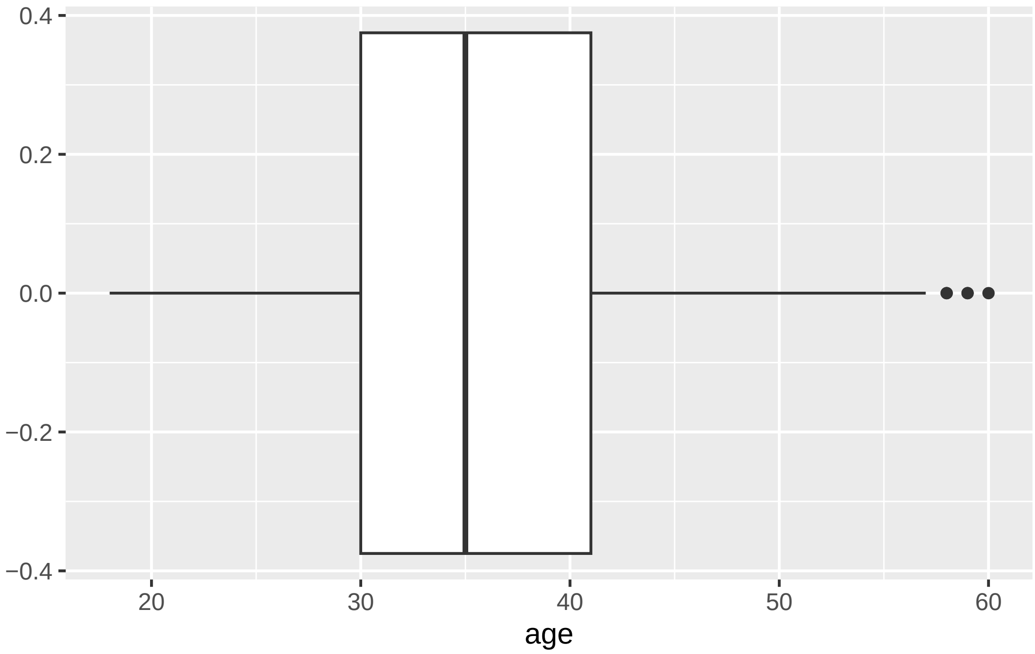
```r
age_bar_original <-
  ggplot(aes(x=age), data = rehab) +
  geom_bar()

age_bar_original
```



```
ggplot(rehab, aes(x = age)) +
  geom_histogram(binwidth = 5) +
  theme_minimal() +
  labs(title = "Age Distribution (10-year increments)",
       x = "Age", y = "Count")
```



```
ggplot(rehab, aes(x = age)) +
  geom_boxplot()
```



```r
summary(rehab$age)
```

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

Age seems to be slightly right skewed. The mean is 35 and the median is also 35. Though its slightly right skewed age seems to be more

- Outline of the Methodological Extension:

You should describe how you plan to divide the labor of learning a new topic: who will do what? You should identify at least two resources that you will use to learn about your topic, as well as the R package(s) or function(s) that you feel will be relevant to applying this method. You will also provide a short (two to four sentence) description of what your extension is and how it builds on ideas from class and/or relates to the data you intend to analyze.