

HELPrct (Data Verbs) Activity

Code ▾

Mohil Patel

March 2, 2022

Set Up:

Hide

```
# load packages
library(tidyverse) # for data verbs like `group-by()` & `summarise()`
library(mosaicData) # includes `HELPrct` data set
library(mosaic) # includes `mplot()` interactive plot builder
library(esquisse) # includes `esquisser()` interactive plot builder

# Load the HELPrct data from `mosaicData` into our RStudio environment
data("HELPrct", package = "mosaicData")
```

Data Description

Hide

glimpse(HELPrct)	
Rows: 453 Columns: 30	
\$ age	<int> 37, 37, 26, 39, 32, 47, 49, 28, 50, 39, 34, 58, 58, 60, 36, 28, 35, 29, 27, 27, 41, 3-
\$ anysubstatus	<int> 1, 1, 1, 1, 1, 1, 1, NA, 1, 1, 1, NA, 0, 1, 1, 1, 1, 1, 0, 0, 1, NA, 1, NA, 1, 1, 1, 1, ~
\$ anysub	<fct> yes, yes, yes, yes, yes, yes, yes, NA, yes, yes, yes, NA, no, yes, yes, yes, yes, yes, no,-
\$ cesd	<int> 49, 30, 39, 15, 39, 6, 52, 32, 50, 46, 46, 49, 22, 36, 43, 35, 19, 40, 52, 37, 35, 18-
\$ di	<int> 3, 22, 0, 2, 12, 1, 14, 1, 14, 4, 0, 3, 5, 10, 2, 6, 1, 2, 0, 1, 1, 1, 4, 2, 4, 1, 0,-
\$ daysanysub	<int> 177, 2, 3, 189, 2, 31, NA, 47, 31, 115, NA, 192, 6, 6, 0, 27, 2, 220, 198, 52, NA, 12-
\$ dayslink	<int> 225, NA, 365, 343, 57, 365, 334, 365, 365, 382, 365, 365, 365, 22, 443, 41, 405, 449,-
\$ drugrisk	<int> 0, 0, 20, 0, 0, 0, 0, 7, 18, 20, 8, 0, 0, 0, 0, 0, 0, 10, 0, 12, 0, 0, 0, 0, 0, 1,-
\$ e2b	<int> NA, NA, NA, 1, 1, NA, 1, 8, 7, 3, NA, NA, NA, 1, NA, 2, NA, 1, 4, NA, 1, NA, 2, 3, NA-
\$ female	<int> 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0-
\$ sex	<fct> male, male, male, female, male, female, female, male, female, male, female, female, m-
\$ g1b	<fct> yes, yes, no, no, no, no, yes, yes, no, no, no, no, no, no, no, yes, no, yes, yes, no-
\$ homeless	<fct> housed, homeless, housed, housed, homeless, housed, housed, homeless, homeless, homel-
\$ i1	<int> 13, 56, 0, 5, 10, 4, 13, 12, 71, 20, 0, 13, 20, 13, 51, 0, 0, 1, 9, 23, 26, 0, 34, 4,-
\$ i2	<int> 26, 62, 0, 5, 13, 4, 20, 24, 129, 27, 0, 13, 31, 20, 51, 0, 0, 1, 24, 23, 26, 0, 34, -
\$ id	<int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25-
\$ indtot	<int> 39, 43, 41, 28, 38, 29, 38, 44, 44, 44, 34, 11, 40, 41, 38, 26, 17, 40, 37, 37, 36, 2-
\$ linkstatus	<int> 1, NA, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, -
\$ link	<fct> yes, NA, no, no, yes, no, no, no, no, no, no, no, no, yes, no, yes, no, no, yes, no, -
\$ mcs	<dbl> 25.111990, 26.670307, 6.762923, 43.967880, 21.675755, 55.508991, 21.793024, 9.160530,-
\$ pcs	<dbl> 58.41369, 36.03694, 74.80633, 61.93168, 37.34558, 46.47521, 24.51504, 65.13801, 38.27-
\$ pss_fr	<int> 0, 1, 13, 11, 10, 5, 1, 4, 5, 0, 0, 13, 13, 1, 1, 7, 9, 1, 13, 11, 8, 14, 10, 6, 6, 3-
\$ racegrp	<fct> black, white, black, white, black, black, black, white, white, white, white, black, w-
\$ satreat	<fct> no, no, no, yes, no, no, yes, yes, no, yes, no, yes, yes, no, no, yes, no, no, ye-
\$ sexrisk	<int> 4, 7, 2, 4, 6, 5, 8, 6, 8, 0, 2, 0, 1, 4, 8, 3, 4, 4, 3, 7, 4, 4, 6, 9, 7, 5, 4, 5, 2-
\$ substance	<fct> cocaine, alcohol, heroin, heroin, cocaine, cocaine, cocaine, alcohol, alcohol, heroin-
\$ treat	<fct> yes, yes, no, no, no, no, yes, no, yes, no, yes, yes, no, no, yes, yes, yes, no, no, ye-
\$ avg_drinks	<int> 13, 56, 0, 5, 10, 4, 13, 12, 71, 20, 0, 13, 20, 13, 51, 0, 0, 1, 9, 23, 26, 0, 34, 4,-
\$ max_drinks	<int> 26, 62, 0, 5, 13, 4, 20, 24, 129, 27, 0, 13, 31, 20, 51, 0, 0, 1, 24, 23, 26, 0, 34, -
\$ hospitalizations	<int> 3, 22, 0, 2, 12, 1, 14, 1, 14, 4, 0, 3, 5, 10, 2, 6, 1, 2, 0, 1, 1, 1, 4, 2, 4, 1, 0,-

TASK 1: Write a couple of sentences describing the `HELPrct` data—What is the setting for the data? What are they about?

The data is looking at individuals who have struggled with drug abuse. A case is a person who struggles with drug abuse. The variables help describe each individuals drug abuse situation as well as their life situation

Basic Summaries with `summarise()`

TASK 2: Produce an expression involving `summarize()` and `HELPrct` that will produce the following.

- number of people (cases) in `HELPrct` study
- combined total number of times in the past 6 months all the people in `HELPrct` entered a detox program (measured at baseline)
- mean time (in days) to first use of any substance post-detox for all the people in `HELPrct`

Hide

HELPrct %>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
total_people <int>	total_detox <int>
453	536
use_mean <dbl>	75.30738
1 row	

Hide

HELPrct %>% summarise(use_mean = mean(daysanysub,na.rm=TRUE))	
use_mean <dbl>	75.30738
1 row	

Hide

NA
NA

Subgrouping with `group_by()`

TASK 3: Repeat task 2 above, but calculate the results group-by-group. Show your code and write a sentence or two about what you observe in the results for each of the following:

- males versus females

Hide

HELPrct %>% group_by(female)%>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
female <int>	total_people <int>
0	346
1	107
total_detox <int>	447
89	
use_mean <dbl>	72.72727
83.77193	
2 rows	

- homeless or not

Hide

HELPrct %>% group_by(homeless)%>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
homeless <fctr>	total_people <int>
homeless	209
housed	244
total_detox <int>	379
157	
use_mean <dbl>	74.33051
76.22222	
2 rows	

- substance

Hide

HELPrct %>% group_by(substance)%>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
substance <fctr>	total_people <int>
alcohol	177
cocaine	152
heroin	124
total_detox <int>	237
96	
use_mean <dbl>	79.36170
85.86517	
53.65574	
3 rows	

- break down the homeless versus housed further, by sex

Hide

HELPrct %>% group_by(female,homeless) %>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
female <int>	homeless <fctr>
0	homeless
0	housed
1	homeless
1	housed
total_people <int>	169
177	
40	
67	
total_detox <int>	338
109	
41	
48	
use_mean <dbl>	72.46875
73.00000	
82.45455	
84.60000	
4 rows	

- break down the homeless versus housed further, by substance

Hide

HELPrct %>% group_by(homeless,substance)%>% summarise(total_people = n(),total_detox= sum(e2b,na.rm=TRUE),use_mean = mean(daysanysub,na.rm=TRUE))	
homeless <fctr>	substance <fctr>
homeless	alcohol
homeless	cocaine
homeless	heroin
housed	alcohol
housed	cocaine
housed	heroin
total_people <int>	103
59	
47	
74	
93	
77	
total_detox <int>	203
60	
116	
34	
36	
87	
use_mean <dbl>	61.96429
94.86486	
71.64000	
105.00000	
79.46154	
41.16667	
6 rows	

TASK 4:

Include one or more interesting plots of the data involving at least 3 variables per plot. Write a few sentences to explain the story that your plot tells about these data. You can use one of the relationships that you studied in Task 2, or you can explore a different group of variables in the `HELPrct` that show something interesting.

The graph below shows the relationship between the age and drug risk for each of the substances. As you can see that heroin is a very risky drug to take at any stage as compared to alcohol and cocaine

Hide

```
ggplot(data = HELPrct, aes(x = age, y = drugrisk)) + geom_point() + facet_wrap(~substance, ncol = 4) + labs(title = " ")
```

Warning: Removed 1 rows containing missing values (geom_point).

