Tidyverse Assignment

Priya Shaji

4/22/2019

## Drug-use-by-age Dataset

## Using dplyr and ggplot packages of tidyverse to explore Drug-use-by-age Dataset

### Dataset Description

It covers 13 drugs across 17 age groups.

Header Definition

alcohol-use Percentage of those in an age group who used alcohol in the past 12 months

alcohol-frequency Median number of times a user in an age group used alcohol in the past 12  
months

marijuana-use Percentage of those in an age group who used marijuana in the past 12 months

marijuana-frequency Median number of times a user in an age group used marijuana in the past 12 months

cocaine-use Percentage of those in an age group who used cocaine in the past 12 months

cocaine-frequency Median number of times a user in an age group used cocaine in the past 12 months

crack-use Percentage of those in an age group who used crack in the past 12 months

rack-frequency Median number of times a user in an age group used crack in the past 12 months

heroin-use Percentage of those in an age group who used heroin in the past 12 months

heroin-frequency Median number of times a user in an age group used heroin in the past 12 months

hallucinogen-use Percentage of those in an age group who used hallucinogens in the past 12 months

hallucinogen-frequency Median number of times a user in an age group used hallucinogens in the past 12 months

inhalant-use Percentage of those in an age group who used inhalants in the past 12 months

inhalant-frequency Median number of times a user in an age group used inhalants in the past 12 months

pain-releiver-use Percentage of those in an age group who used pain relievers in the past 12 months

pain-releiver-frequency Median number of times a user in an age group used pain relievers in the past 12 months

oxycontin-use Percentage of those in an age group who used oxycontin in the past 12 months

oxycontin-frequency Median number of times a user in an age group used oxycontin in the past 12 months

tranquilizer-use Percentage of those in an age group who used tranquilizer in the past 12 months

tranquilizer-frequency Median number of times a user in an age group used tranquilizer in the past 12 months

stimulant-use Percentage of those in an age group who used stimulants in the past 12 months

stimulant-frequency Median number of times a user in an age group used stimulants in the past 12 months

meth-use Percentage of those in an age group who used meth in the past 12 months

meth-frequency Median number of times a user in an age group used meth in the past 12 months

sedative-use Percentage of those in an age group who used sedatives in the past 12 months

sedative-frequency Median number of times a user in an age group used sedatives in the past 12 months

### Load the dataset

1. Install tidyverse package and load the dataset

#install.packages("tidyverse")  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.1 ✔ purrr 0.3.2   
## ✔ tibble 2.1.1 ✔ dplyr 0.8.0.1  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0   
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

## Using readr to read csv  
df\_drug <- read\_csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/drug-use-by-age/drug-use-by-age.csv')

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## age = col\_character(),  
## `cocaine-frequency` = col\_character(),  
## `crack-frequency` = col\_character(),  
## `heroin-frequency` = col\_character(),  
## `inhalant-frequency` = col\_character(),  
## `oxycontin-frequency` = col\_character(),  
## `meth-frequency` = col\_character()  
## )

## See spec(...) for full column specifications.

── Attaching packages ─────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ── ✔ ggplot2 3.1.0 ✔ purrr 0.3.2  
✔ tibble 2.1.1 ✔ dplyr 0.8.0.1 ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
✔ readr 1.3.1 ✔ forcats 0.4.0  
── Conflicts ────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ── ✖ dplyr::filter() masks stats::filter() ✖ dplyr::lag() masks stats::lag() Parsed with column specification: cols( .default = col\_double(), age = col\_character(), cocaine-frequency = col\_character(), crack-frequency = col\_character(), heroin-frequency = col\_character(), inhalant-frequency = col\_character(), oxycontin-frequency = col\_character(), meth-frequency = col\_character() ) See spec(…) for full column specifications.

### Data Transformations

1. Let’s explore the data type of all the variables of the dataset using dplyr::glimpse

#View(df\_drug)  
dplyr::glimpse(df\_drug)

## Observations: 17  
## Variables: 28  
## $ age <chr> "12", "13", "14", "15", "16", "17", "1…  
## $ n <dbl> 2798, 2757, 2792, 2956, 3058, 3038, 24…  
## $ `alcohol-use` <dbl> 3.9, 8.5, 18.1, 29.2, 40.1, 49.3, 58.7…  
## $ `alcohol-frequency` <dbl> 3, 6, 5, 6, 10, 13, 24, 36, 48, 52, 52…  
## $ `marijuana-use` <dbl> 1.1, 3.4, 8.7, 14.5, 22.5, 28.0, 33.7,…  
## $ `marijuana-frequency` <dbl> 4, 15, 24, 25, 30, 36, 52, 60, 60, 52,…  
## $ `cocaine-use` <dbl> 0.1, 0.1, 0.1, 0.5, 1.0, 2.0, 3.2, 4.1…  
## $ `cocaine-frequency` <chr> "5.0", "1.0", "5.5", "4.0", "7.0", "5.…  
## $ `crack-use` <dbl> 0.0, 0.0, 0.0, 0.1, 0.0, 0.1, 0.4, 0.5…  
## $ `crack-frequency` <chr> "-", "3.0", "-", "9.5", "1.0", "21.0",…  
## $ `heroin-use` <dbl> 0.1, 0.0, 0.1, 0.2, 0.1, 0.1, 0.4, 0.5…  
## $ `heroin-frequency` <chr> "35.5", "-", "2.0", "1.0", "66.5", "64…  
## $ `hallucinogen-use` <dbl> 0.2, 0.6, 1.6, 2.1, 3.4, 4.8, 7.0, 8.6…  
## $ `hallucinogen-frequency` <dbl> 52, 6, 3, 4, 3, 3, 4, 3, 2, 4, 3, 2, 3…  
## $ `inhalant-use` <dbl> 1.6, 2.5, 2.6, 2.5, 3.0, 2.0, 1.8, 1.4…  
## $ `inhalant-frequency` <chr> "19.0", "12.0", "5.0", "5.5", "3.0", "…  
## $ `pain-releiver-use` <dbl> 2.0, 2.4, 3.9, 5.5, 6.2, 8.5, 9.2, 9.4…  
## $ `pain-releiver-frequency` <dbl> 36, 14, 12, 10, 7, 9, 12, 12, 10, 15, …  
## $ `oxycontin-use` <dbl> 0.1, 0.1, 0.4, 0.8, 1.1, 1.4, 1.7, 1.5…  
## $ `oxycontin-frequency` <chr> "24.5", "41.0", "4.5", "3.0", "4.0", "…  
## $ `tranquilizer-use` <dbl> 0.2, 0.3, 0.9, 2.0, 2.4, 3.5, 4.9, 4.2…  
## $ `tranquilizer-frequency` <dbl> 52.0, 25.5, 5.0, 4.5, 11.0, 7.0, 12.0,…  
## $ `stimulant-use` <dbl> 0.2, 0.3, 0.8, 1.5, 1.8, 2.8, 3.0, 3.3…  
## $ `stimulant-frequency` <dbl> 2.0, 4.0, 12.0, 6.0, 9.5, 9.0, 8.0, 6.…  
## $ `meth-use` <dbl> 0.0, 0.1, 0.1, 0.3, 0.3, 0.6, 0.5, 0.4…  
## $ `meth-frequency` <chr> "-", "5.0", "24.0", "10.5", "36.0", "4…  
## $ `sedative-use` <dbl> 0.2, 0.1, 0.2, 0.4, 0.2, 0.5, 0.4, 0.3…  
## $ `sedative-frequency` <dbl> 13.0, 19.0, 16.5, 30.0, 3.0, 6.5, 10.0…

1. Using the functions provided by the dplyr package, select the columns which ends with use

drug\_use <- df\_drug %>%  
 select(age,n,ends\_with("use"))

1. View the dataset drug-use

head(drug\_use)

## # A tibble: 6 x 15  
## age n `alcohol-use` `marijuana-use` `cocaine-use` `crack-use`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 12 2798 3.9 1.1 0.1 0   
## 2 13 2757 8.5 3.4 0.1 0   
## 3 14 2792 18.1 8.7 0.1 0   
## 4 15 2956 29.2 14.5 0.5 0.1  
## 5 16 3058 40.1 22.5 1 0   
## 6 17 3038 49.3 28 2 0.1  
## # … with 9 more variables: `heroin-use` <dbl>, `hallucinogen-use` <dbl>,  
## # `inhalant-use` <dbl>, `pain-releiver-use` <dbl>,  
## # `oxycontin-use` <dbl>, `tranquilizer-use` <dbl>,  
## # `stimulant-use` <dbl>, `meth-use` <dbl>, `sedative-use` <dbl>

1. Now gather the column names as values for a new column drugUse\_name

#drug\_use  
drug\_use<-drug\_use%>%  
 gather(-age,-n,key = "drugUse\_name",value = "drugUse",`alcohol-use`,  
`marijuana-use`,  
`cocaine-use`,  
`crack-use`,  
`heroin-use`,  
`hallucinogen-use`,  
`inhalant-use`,  
`pain-releiver-use`,  
`oxycontin-use`,  
`tranquilizer-use`,  
`stimulant-use`,  
`meth-use`,  
`sedative-use`  
)

1. Using the functions provided by the dplyr package, select the columns which ends with frequency

drug\_freq <- df\_drug %>%  
 select(age,n,ends\_with("frequency"))  
head(drug\_freq)

## # A tibble: 6 x 15  
## age n `alcohol-freque… `marijuana-freq… `cocaine-freque…  
## <chr> <dbl> <dbl> <dbl> <chr>   
## 1 12 2798 3 4 5.0   
## 2 13 2757 6 15 1.0   
## 3 14 2792 5 24 5.5   
## 4 15 2956 6 25 4.0   
## 5 16 3058 10 30 7.0   
## 6 17 3038 13 36 5.0   
## # … with 10 more variables: `crack-frequency` <chr>,  
## # `heroin-frequency` <chr>, `hallucinogen-frequency` <dbl>,  
## # `inhalant-frequency` <chr>, `pain-releiver-frequency` <dbl>,  
## # `oxycontin-frequency` <chr>, `tranquilizer-frequency` <dbl>,  
## # `stimulant-frequency` <dbl>, `meth-frequency` <chr>,  
## # `sedative-frequency` <dbl>

1. Now gather the column names as values for a new column drugFreq\_name

drug\_freq<-drug\_freq%>%  
 gather(-age,-n,key = "drugFreq\_name",value = "drugFreq",`alcohol-frequency`,  
`marijuana-frequency`,  
`cocaine-frequency`,  
`crack-frequency`,  
`heroin-frequency`,  
`hallucinogen-frequency`,  
`inhalant-frequency`,  
`pain-releiver-frequency`,  
`oxycontin-frequency`,  
`tranquilizer-frequency`,  
`stimulant-frequency`,  
`meth-frequency`,  
`sedative-frequency`  
)

1. Merge the two datasets drug\_use and drug\_freq in a single dataframe as tidy\_drug\_data using full\_join() as a function provided by the dplyr package

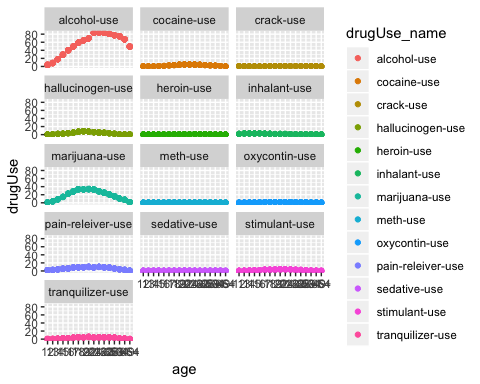
tidy\_drug\_data <- full\_join(drug\_use,drug\_freq,by=c("age","n"))  
  
head(tidy\_drug\_data)

## # A tibble: 6 x 6  
## age n drugUse\_name drugUse drugFreq\_name drugFreq  
## <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 12 2798 alcohol-use 3.9 alcohol-frequency 3   
## 2 12 2798 alcohol-use 3.9 marijuana-frequency 4   
## 3 12 2798 alcohol-use 3.9 cocaine-frequency 5.0   
## 4 12 2798 alcohol-use 3.9 crack-frequency -   
## 5 12 2798 alcohol-use 3.9 heroin-frequency 35.5   
## 6 12 2798 alcohol-use 3.9 hallucinogen-frequency 52

1. Use ggplot() along with facet\_wrap to individually plot the variation of drugs with age.

drugUse\_plot <- ggplot(tidy\_drug\_data,aes(x = age, y = drugUse,color=drugUse\_name)) +  
  
 geom\_point() +  
   
 facet\_wrap(~ drugUse\_name, nrow = 5) +   
 geom\_smooth(color = "black")  
  
drugUse\_plot

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

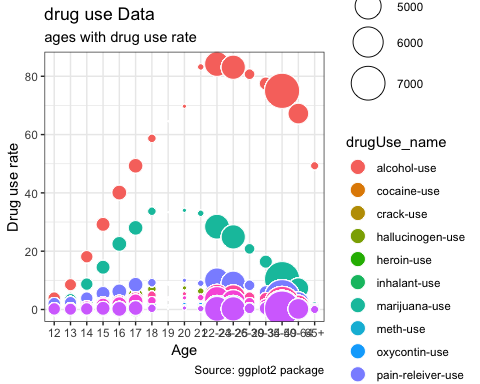


By plotting the graphs individually for each drug, it gives us a clear picture of which drug usage is higher at particular ages compared to other drugs.

1. Here ggplot() is used to plot the drug use rate varying with age. As we see in the data description,n is the number of people survyed for a particular drug.

ggplot(data = tidy\_drug\_data,   
 mapping = aes(x = age, y = drugUse)) +   
 geom\_point(aes(fill = drugUse\_name, size = n), shape = 21, color = "white") +   
 geom\_smooth(aes(x = age, y = drugUse)) +  
 labs(  
 x = "Age",   
 y = "Drug use rate",   
 title = "drug use Data",  
 subtitle = "ages with drug use rate",  
 caption = "Source: ggplot2 package") +   
 scale\_color\_brewer(palette = "Set1") +   
 scale\_size(range = c(0, 12)) +  
 guides(size = guide\_legend(override.aes = list(col = "black")),   
 fill = guide\_legend(override.aes = list(size = 5))) +  
 theme\_bw()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



1. Using filter() function of dplyr package of tidyverse to group drugs whose usage is greater than 60%

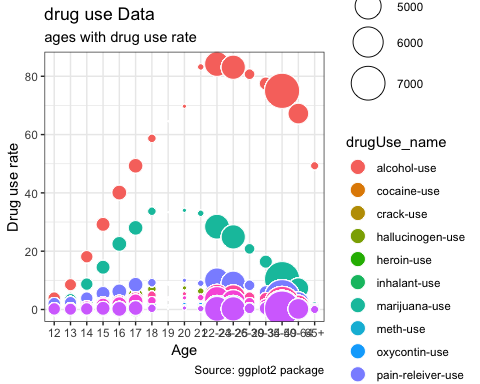
head(tidy\_drug\_data %>%   
 filter(`drugUse` > 60 ))

## # A tibble: 6 x 6  
## age n drugUse\_name drugUse drugFreq\_name drugFreq  
## <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 19 2223 alcohol-use 64.6 alcohol-frequency 36   
## 2 19 2223 alcohol-use 64.6 marijuana-frequency 60   
## 3 19 2223 alcohol-use 64.6 cocaine-frequency 5.5   
## 4 19 2223 alcohol-use 64.6 crack-frequency 2.0   
## 5 19 2223 alcohol-use 64.6 heroin-frequency 180.0   
## 6 19 2223 alcohol-use 64.6 hallucinogen-frequency 3

### Conclusion

ggplot(data = tidy\_drug\_data,   
 mapping = aes(x = age, y = drugUse)) +   
 geom\_point(aes(fill = drugUse\_name, size = n), shape = 21, color = "white") +   
 geom\_smooth(aes(x = age, y = drugUse)) +  
 labs(  
 x = "Age",   
 y = "Drug use rate",   
 title = "drug use Data",  
 subtitle = "ages with drug use rate",  
 caption = "Source: ggplot2 package") +   
 scale\_color\_brewer(palette = "Set1") +   
 scale\_size(range = c(0, 12)) +  
 guides(size = guide\_legend(override.aes = list(col = "black")),   
 fill = guide\_legend(override.aes = list(size = 5))) +  
 theme\_bw()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



As we can see, from the graphs and also from filtering data :

Alcohol is the most abused drug among the age group of 22-23.

More than 80% Percentage of those in an age group of 22-23 who used alcohol in the past 12 months.