Module 8 Math 365

James Harrington

10/14/2021

#Exercise 10.13  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

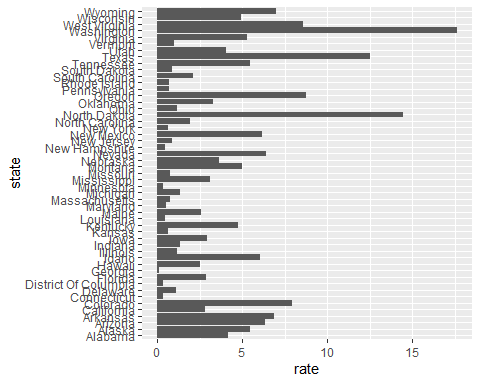
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(dslabs)  
data(us\_contagious\_diseases)  
  
  
#1) Pie charts are appropriate:  
#d. Never. Barplots and tables are always better.

#2) What is the problem with the plot below:  
#b)The axis does not start at 0. Judging by the length, it appears Trump received 3 times as many votes when, in fact, it was about 30% more.

#3) Take a look at the following two plots. They show the same information: 1928 rates of measles across the 50 states.  
  
#c)The plot on the right is better because alphabetical order has nothing to do with the disease and by ordering according to actual rate, we quickly see the states with most and least rates.

#4) To make the plot on the left, we have to reorder the levels of the states’ variables.  
dat <-us\_contagious\_diseases %>%   
 filter(year== 1967 & disease == "Measles" & !is.na(population)) %>%   
 mutate(rate= count/population\*10000\*52/ weeks\_reporting)  
  
#Note what happens when we make a barplot:  
  
dat %>% ggplot(aes(state, rate))+  
 geom\_bar(stat = "identity")+  
 coord\_flip()



state<-dat$state  
  
rate<- dat$count/dat$population\*10000\*52/dat$weeks\_reporting  
  
state<- reorder(state, rate)   
levels(state)

## [1] "Georgia" "District Of Columbia" "Connecticut"   
## [4] "Minnesota" "Louisiana" "New Hampshire"   
## [7] "Maryland" "Kansas" "New York"   
## [10] "Pennsylvania" "Rhode Island" "Massachusetts"   
## [13] "Missouri" "New Jersey" "South Dakota"   
## [16] "Vermont" "Delaware" "Ohio"   
## [19] "Illinois" "Michigan" "Indiana"   
## [22] "North Carolina" "South Carolina" "Hawaii"   
## [25] "Maine" "California" "Florida"   
## [28] "Iowa" "Mississippi" "Oklahoma"   
## [31] "Nebraska" "Utah" "Alabama"   
## [34] "Kentucky" "Wisconsin" "Montana"   
## [37] "Virginia" "Alaska" "Tennessee"   
## [40] "Idaho" "New Mexico" "Arizona"   
## [43] "Nevada" "Arkansas" "Wyoming"   
## [46] "Colorado" "West Virginia" "Oregon"   
## [49] "Texas" "North Dakota" "Washington"

?reorder()#make sure the reorder function is used

## starting httpd help server ... done

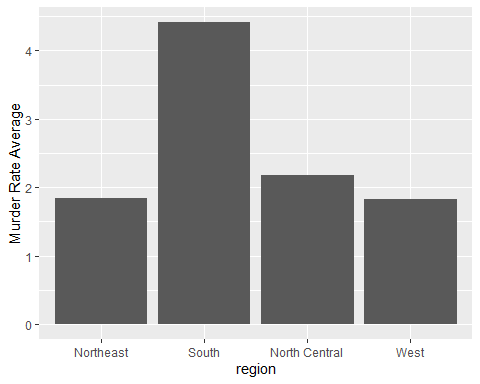
#5) Now with one line of code, define the dat table as done above, but change the use mutate to create a rate variable and re-order the state variable so that the levels are re-ordered by this variable. Then make a barplot using the code above, but for this new dat.  
  
  
  
dat<- us\_contagious\_diseases %>% mutate(rate=count/population\*10000\*52/ weeks\_reporting)  
  
state <- reorder(state, dat)

## Error in tapply(X = X, INDEX = x, FUN = FUN, ...): arguments must have same length

levels(state)

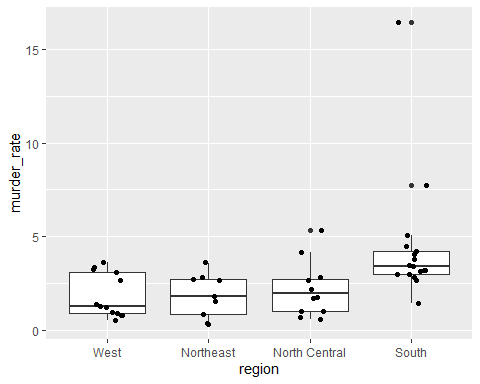
## [1] "Georgia" "District Of Columbia" "Connecticut"   
## [4] "Minnesota" "Louisiana" "New Hampshire"   
## [7] "Maryland" "Kansas" "New York"   
## [10] "Pennsylvania" "Rhode Island" "Massachusetts"   
## [13] "Missouri" "New Jersey" "South Dakota"   
## [16] "Vermont" "Delaware" "Ohio"   
## [19] "Illinois" "Michigan" "Indiana"   
## [22] "North Carolina" "South Carolina" "Hawaii"   
## [25] "Maine" "California" "Florida"   
## [28] "Iowa" "Mississippi" "Oklahoma"   
## [31] "Nebraska" "Utah" "Alabama"   
## [34] "Kentucky" "Wisconsin" "Montana"   
## [37] "Virginia" "Alaska" "Tennessee"   
## [40] "Idaho" "New Mexico" "Arizona"   
## [43] "Nevada" "Arkansas" "Wyoming"   
## [46] "Colorado" "West Virginia" "Oregon"   
## [49] "Texas" "North Dakota" "Washington"

#6)Say we are interested in comparing gun homicide rates across regions of the US. We see this plot:  
  
library(dslabs)  
data("murders")  
murders %>% mutate(rate = total/population\*100000) %>%   
group\_by(region) %>%   
summarize(avg=mean(rate)) %>% # the bars will be measured by their average rate  
mutate(region=factor(region)) %>%   
ggplot(aes(region,avg))+ #region is on the x and avg is on the y   
geom\_bar(stat= "identity")+  
ylab("Murder Rate Average")



#and decide ti movr to a state in the wastern reion. What is the main problem eith this interpretation?  
  
#c)It does not show all the data. We do not see the variability within a region and it’s possible that the safest states are not in the West.

#7) Make a boxplot of the murder rates defined as  
data("murders")  
murders %>% mutate(murder\_rate=total/population\*100000) %>%   
mutate(region= reorder(region,murder\_rate, FUN = median)) %>%   
ggplot(aes(region,murder\_rate)) +   
geom\_boxplot()+  
geom\_jitter(width= .15)



#8) The plots below show three continuous variables.  
#The line   
#x=2  
#appears to separate the points. But it is actually not the case, which we can see by plotting the data in a couple of two-dimensional points.  
#Why is this happening?  
#a)Humans are not good at reading pseudo-3D plots.