Data-Visual-Analysis-on-Smoking.R

AakashSharma

2021-02-14

# Import needed libraries  
library(ggplot2)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ tibble 3.0.6 ✓ dplyr 1.0.4  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1  
## ✓ purrr 0.3.4

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

library(gridExtra)

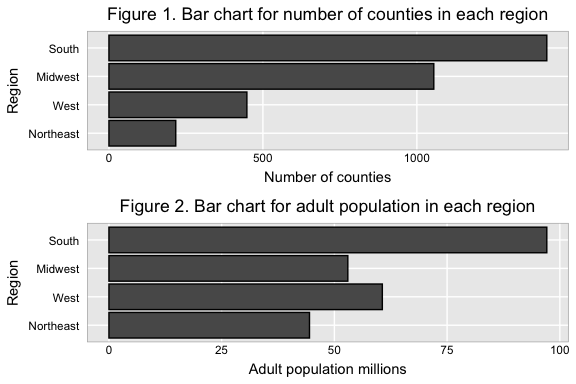
##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

# Import source sheet that has code for the gray theme  
source("hw.R") # calls to theme\_gray() and theme()  
# Reads in the counties smoking data set  
counties\_smoking <- read.csv(file = "counties\_smoking.csv", as.is = FALSE)  
# Reads in the states smoking data set  
states\_smoking <- read.csv(file = "states\_smoking.csv", as.is = FALSE)  
# Reads in the US smoking data set  
us\_smoking <- read.csv(file = "US\_smoking.csv", as.is = FALSE)  
  
# Reorder the regions so the legends & graphs appear in order of the region  
counties\_smoking2 <- counties\_smoking  
counties\_smoking2$region\_name <- factor(counties\_smoking2$region\_name,   
 levels = c('Northeast','West','Midwest','South'), ordered = TRUE)  
# Fig 1

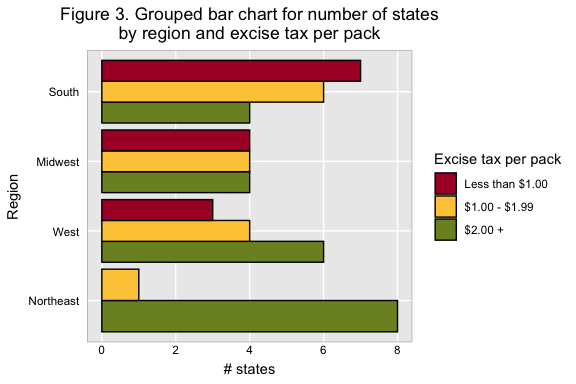
fig1 <- ggplot(data = counties\_smoking2, aes(y = region\_name), fill = "grey30") +  
 geom\_bar(color = "Black") +  
 labs(x = "Number of counties",  
 y = "Region",  
 title = "Figure 1. Bar chart for number of counties in each region") + hw +   
 theme(plot.caption = element\_text(hjust=0))   
  
# Fig 2  
# Reorder the regions so the legends & graphs appear in order of the region  
states\_smoking2 <- states\_smoking  
states\_smoking2$region\_name <- factor(states\_smoking2$region\_name, c('Northeast','West','Midwest','South'))

fig2 <- ggplot(data = states\_smoking2, aes(y= region\_name, weight = adult\_population/1000000),   
 fill = "grey30") +  
 geom\_bar(color = "Black") +  
 labs(x = "Adult population millions",  
 y = "Region",  
 title = "Figure 2. Bar chart for adult population in each region") + hw +   
 theme(plot.caption = element\_text(hjust=0))  
  
# Combine both fig1 and fig2 into a singular figure  
grid.arrange(fig1, fig2, nrow = 2)



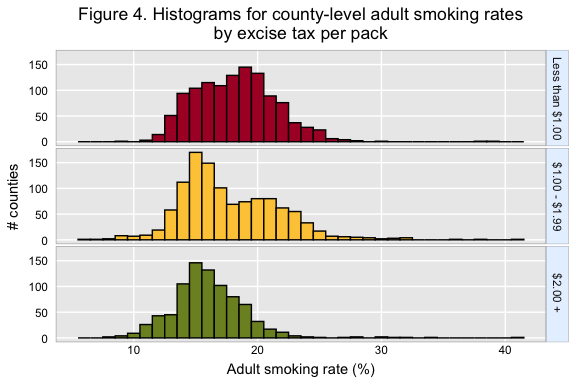
# Fig 3  
states\_smoking2$pack\_tax\_range <- factor(states\_smoking2$pack\_tax\_range,   
 levels = c('Less than $1.00','$1.00 - $1.99','$2.00 + '), ordered = TRUE)

vert <- ggplot(data = states\_smoking2, aes(region\_name, fill = fct\_rev(pack\_tax\_range))) +  
 geom\_bar(position = "dodge", color = "black") +  
 labs(y = "# states",  
 x = "Region",  
 title = paste("Figure 3. Grouped bar chart for number of states", "by region and excise tax per pack",   
 sep="\n"),  
 fill = 'Excise tax per pack') +  
 scale\_fill\_manual(values = c("#ab0730","#fdca44","#7d8f29"),   
 breaks = c('Less than $1.00','$1.00 - $1.99','$2.00 + ')) + hw +   
 theme(plot.caption = element\_text(hjust=0))  
vert + coord\_flip()



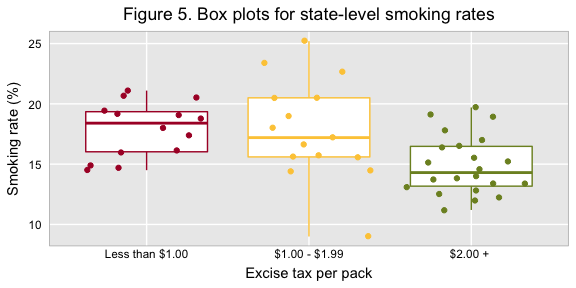
# Fig 4  
# Reorder the pack tax range so the legends & graphs appear in order of the packs  
counties\_smoking2$pack\_tax\_range <- factor(counties\_smoking2$pack\_tax\_range,   
 levels = c('Less than $1.00','$1.00 - $1.99','$2.00 + '), ordered = TRUE)

ggplot(data = counties\_smoking2, aes(x = adult\_smoking\_rate, fill = pack\_tax\_range)) +  
 geom\_histogram(binwidth = 1, color = "black") +   
 scale\_fill\_manual(values=c("#ab0730","#fdca44","#7d8f29")) +  
 facet\_grid(rows = vars(pack\_tax\_range)) +  
 labs(x = "Adult smoking rate (%)",  
 y = "# counties",  
 title = paste("Figure 4. Histograms for county-level adult smoking rates", "by excise tax per pack",   
 sep="\n")) + hw +   
 theme(plot.caption = element\_text(hjust=0), legend.position= "none")



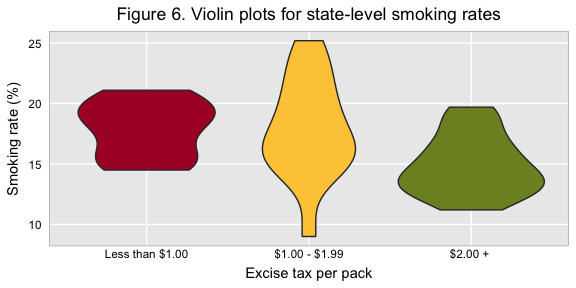
# Fig 5  
# Reorder the pack tax range so the legends & graphs appear in order of the packs  
states\_smoking2$pack\_tax\_range <- factor(states\_smoking2$pack\_tax\_range,   
 levels = c('Less than $1.00','$1.00 - $1.99','$2.00 + '), ordered = TRUE)  
# Create a custome colored pallette and assign the hex values  
coloredpal <- c("#ab0730","#fdca44","#7d8f29")  
set.seed(4977)

fig5 <- ggplot(data = states\_smoking2, aes(x = pack\_tax\_range, y = adult\_smoking\_rate, color = pack\_tax\_range)) +   
 geom\_boxplot(position = "dodge") +   
 geom\_jitter(position = position\_jitter(set.seed(4977))) +  
 labs(y = "Smoking rate (%)",  
 x = "Excise tax per pack",  
 title = "Figure 5. Box plots for state-level smoking rates") + hw +  
 theme(plot.caption = element\_text(hjust=0), legend.position = "none")  
fig5 + scale\_color\_manual(values = coloredpal)



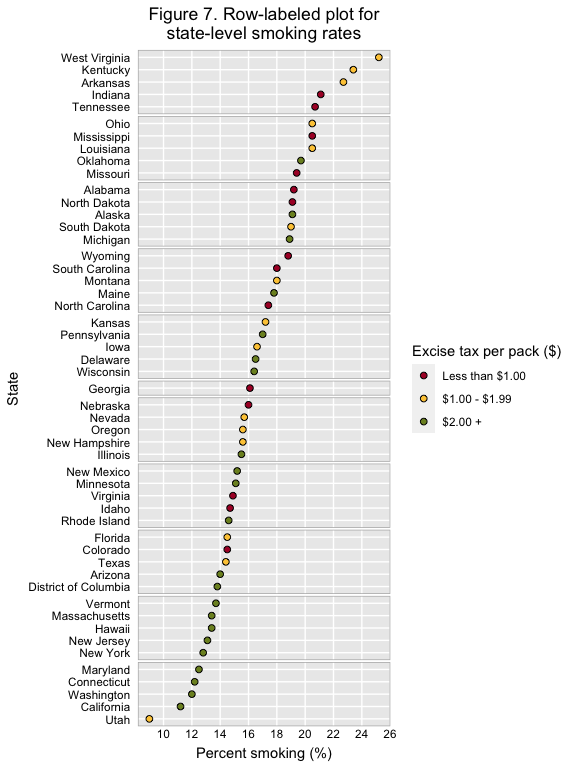
# Fig 6

ggplot(data = states\_smoking2, aes(x = pack\_tax\_range, y = adult\_smoking\_rate, fill = pack\_tax\_range)) +  
 geom\_violin() +  
 scale\_fill\_manual(values=c("#ab0730","#fdca44","#7d8f29")) +  
 labs(y = "Smoking rate (%)",  
 x = "Excise tax per pack",  
 title = "Figure 6. Violin plots for state-level smoking rates") + hw +   
 theme(plot.caption = element\_text(hjust=0), legend.position = "none")



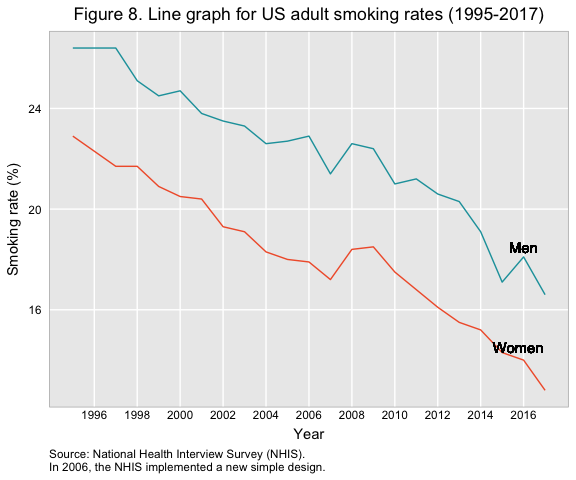
# Fig 7  
# Create the panel num for the facet  
panel\_num <- c(as.integer((5:29)/5), 6L, 6 + as.integer((5:29)/5))  
  
# Before adding panel numbers to the data frame, reorder it so states w/ smallest percent   
# change are at the beginning. These states will be in the first panel.   
states\_smoking\_2 <- states\_smoking[order(-states\_smoking$adult\_smoking\_rate),]  
  
# Add the panel num variable to our new states smoking dataframe  
states\_smoking\_2$panel\_num <- panel\_num  
  
# Because states are a factor, ggplot() will put the states names along the y axis in   
# increasing level order starting at the BOTTOM of the graph. Need to reorder the factor levels  
# of states, so states w/ the largest changes have the lowest level numbers  
states\_smoking\_2$state <- reorder(states\_smoking\_2$state, states\_smoking\_2$adult\_smoking\_rate)  
  
# Must factor the pack tax range for the fill  
states\_smoking\_2$pack\_tax\_range <- factor(states\_smoking\_2$pack\_tax\_range,   
 levels = c('Less than $1.00','$1.00 - $1.99','$2.00 + '), ordered = TRUE)

ggplot(data = states\_smoking\_2, aes(x = adult\_smoking\_rate, y = state, fill = pack\_tax\_range)) +   
 geom\_point(shape = 21, size = 2) +  
 scale\_fill\_manual(values = c("#ab0730","#fdca44","#7d8f29")) +  
 scale\_x\_continuous(breaks = seq(10, 26, by = 2)) +   
 labs(x ="Percent smoking (%)",   
 y = "State",  
 title = paste("Figure 7. Row-labeled plot for", "state-level smoking rates",   
 sep="\n"),  
 fill = 'Excise tax per pack ($)') + hw + guides(fill = guide\_legend(reverse = FALSE)) +  
 facet\_grid(vars(panel\_num), space = "free", scales = "free") +  
 theme(strip.text=element\_blank())



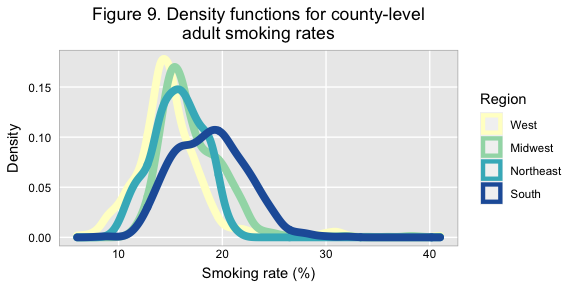
# Fig 8

ggplot(data = us\_smoking, aes(x = year)) +   
 geom\_line(aes(y = male\_smokers), color = "#19a0aa") +   
 geom\_line(aes(y = female\_smokers), color = "#f15f36") +  
 scale\_x\_continuous(breaks= seq(1996, 2016, by = 2)) +   
 scale\_y\_continuous(breaks= seq(16, 24, by = 4)) +   
 geom\_text(aes(x = 2016, y = 18.5, label = "Men")) +   
 geom\_text(aes(x = 2015.75, y = 14.5, label = "Women")) +  
 labs(x = "Year",  
 y = "Smoking rate (%)",  
 title = "Figure 8. Line graph for US adult smoking rates (1995-2017)",  
 caption = paste("Source: National Health Interview Survey (NHIS).",   
 "In 2006, the NHIS implemented a new simple design.",   
 sep="\n")) + hw +  
 theme(plot.caption = element\_text(hjust=0), legend.position = "none")



# Fig 9  
# Reorder the regions so the legends & graphs appear in order of the region  
counties\_smoking$region\_name <- factor(counties\_smoking$region\_name,   
 levels = c('West','Midwest','Northeast', 'South'), ordered = TRUE)  
# Set up our color palette from color brewer, 4-class YlGnBu  
color\_brewer\_palette <- c("#ffffcc","#a1dab4","#41b6c4","#225ea8")

fig1 <- ggplot(data = counties\_smoking, aes(x = adult\_smoking\_rate, color = region\_name)) +   
 geom\_density(alpha = 0.3, size = 2.75) +  
 labs(x = "Smoking rate (%)",  
 y = "Density",  
 title = paste("Figure 9. Density functions for county-level",  
 "adult smoking rates", sep = "\n"),  
 color = "Region") + hw  
fig1 + scale\_color\_manual(values = color\_brewer\_palette)



# Clean Up, reset memory as well clear our env pane  
rm(list=ls())  
gc(verbose = TRUE, reset = TRUE)

## used (Mb) gc trigger (Mb) max used (Mb)  
## Ncells 1205902 64.5 2116267 113.1 1205902 64.5  
## Vcells 2159234 16.5 8388608 64.0 2159234 16.5

format(memory.size(), units = "MB")

## Warning: 'memory.size()' is Windows-specific

## [1] "Inf"