Statistical Computing 1, Stat 590 | Homework 1a | Name: _____
Fall 2015 | Prof. Erik B. Erhardt

Part I. (30 points) Do all calculations in LaTeX + R + knitr. Insert computer text output and graphics to support what you are saying. For this assignment, all R code should well commented and be visible (`echo=TRUE`) in the document where you have written it.

(30pts) **1. Improvement of a data visualization**

(a) (10 pts) Find a plot online that can be improved, include the link to the original source, include the graphic in this report. Explain what is good about the display, and what is deficient about the display.

(b) (5 pts) Obtain or `http://www.datathief.org/` the data from the plot.

(c) (10 pts) Improve the plot, either by doodling the improvement on pen/paper and scan the image to be included, or by writing code to produce the improved plot.

(d) (5 pts) Describe why your changes improve the display.

An example

(a) Plot

Original version from 2010, `http://statacumen.com/pub/blog/BrianSanderoff/Party_by_age_line_chart_-_2008-10.pdf`, is shown in Figure 1.

Updated version: `http://watchdog.org/wp-content/blogs.dir/1/files/2014/07/nm-voter-registration-by-age.jpg` at `http://www.capitolreportnewmexico.com/2014/07/an-opening-for-republicans-in-nm-or-are-independents-flexing-t`

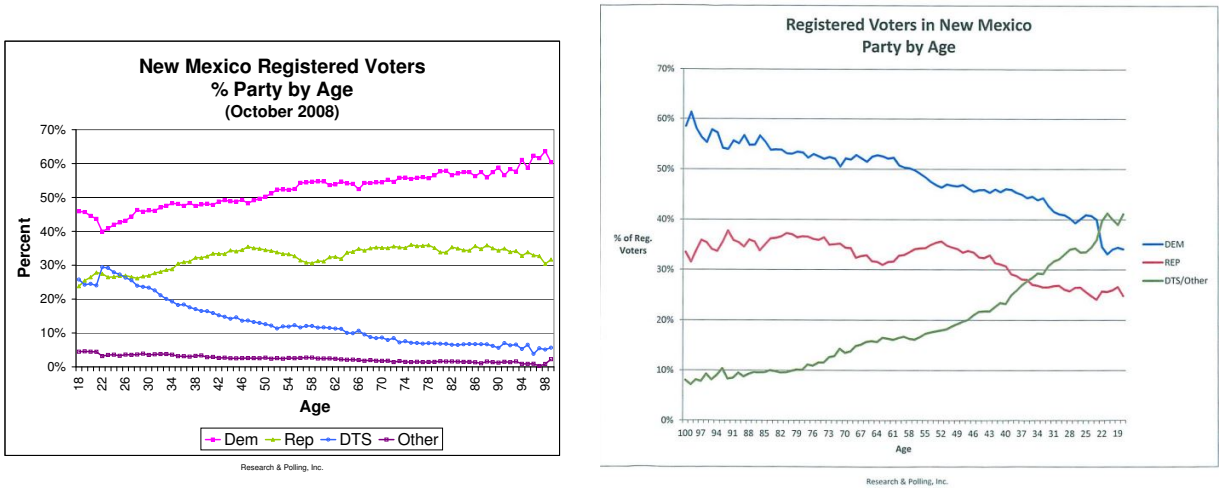


Figure 1: Original party by age line chart

Because population size decreases over time, the older people appear to have more political power than they actually do. Evidence that the very old and very young populations are smaller is indicated by the increased variability of the lines in those regions.

(b) Data

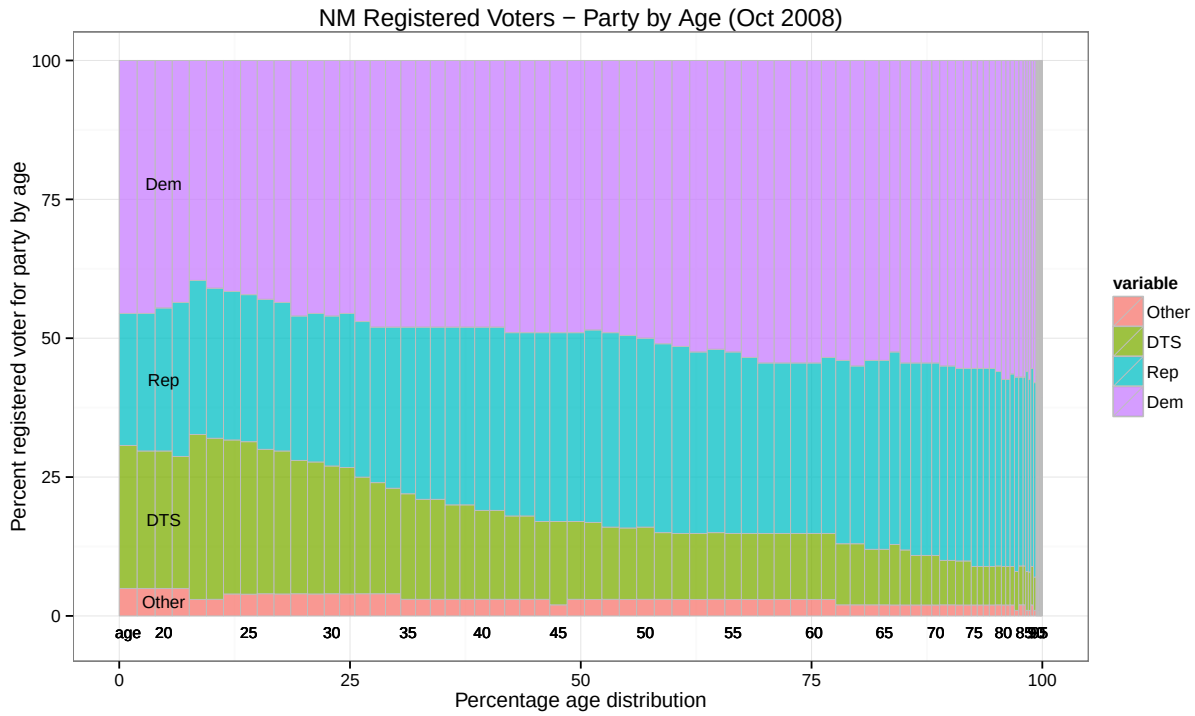
I used `http://www.datathief.org/` to extract the data from the first plot. I also obtained NM population numbers from the 2010 census.

(c) Improvement

```
# Erik B. Erhardt
# 4/28/2012

# Recreating this plot as a Marimekko mosaic chart
# NM Registered Voters - Party by Age Line Chart (Oct 2008)
# http://rpinc.com/wb/media/reports/Party%20by%20age%20line%20chart%20-%202008-10.pdf

# Census population sizes
# NM population numbers
# http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_SF1_QTP1&prodType=table

ages <- c("15-19", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54",
          "55-59", "60-64", "65-69", "70-74", "75-79", "80-84", "85-89", "90-99")
pop.ages <- c(149861,142370,139678,127567,123303,125220,144839,147170,
              136799,120137, 87890, 65904, 50230, 36238, 21622, 10371)

age <- seq(18,99)
pop <- c(rep(pop.ages/5,each=5)[c(4:75)], rep(pop.ages[length(pop.ages)]/10,10))
pop.prop <- pop/sum(pop)

# datathief http://rpinc.com/wb/media/reports/Party%20by%20age%20line%20chart%20-%202008-10.pdf

dem <- c( 0.46,0.46,0.45,0.44,0.40,0.41,0.42,0.43,0.43,0.44,0.46
```

```
         ,0.46,0.46,0.46,0.47,0.48,0.48,0.48,0.48,0.48,0.48,0.48
         ,0.48,0.48,0.49,0.49,0.49,0.49,0.49,0.49,0.49,0.50,0.50
         ,0.51,0.52,0.53,0.52,0.53,0.54,0.55,0.55,0.55,0.55,0.54
         ,0.54,0.55,0.54,0.54,0.53,0.55,0.55,0.55,0.55,0.55,0.55
         ,0.56,0.56,0.56,0.56,0.56,0.56,0.56,0.58,0.58,0.57,0.57
         ,0.57,0.57,0.56,0.58,0.56,0.58,0.58,0.58,0.59,0.59,0.61
         ,0.59,0.62,0.61,0.62,0.60)

rep <- c( 0.24,0.25,0.26,0.28,0.28,0.27,0.27,0.27,0.27,0.27,0.26
         ,0.27,0.27,0.28,0.28,0.28,0.29,0.30,0.31,0.31,0.32,0.32
         ,0.33,0.33,0.33,0.33,0.34,0.34,0.34,0.35,0.35,0.35,0.34
         ,0.34,0.34,0.33,0.33,0.33,0.32,0.31,0.31,0.31,0.31,0.32
         ,0.33,0.32,0.34,0.34,0.35,0.34,0.35,0.35,0.35
         ,0.35,0.35,0.35,0.35,0.36,0.36,0.36,0.36,0.35,0.34,0.34
         ,0.35,0.35,0.34,0.34,0.36,0.35,0.36,0.35,0.34,0.35,0.34
         ,0.34,0.33,0.34,0.33,0.32,0.30,0.31)

dts  <- c( 0.26,0.25,0.25,0.24,0.30,0.29,0.28
          ,0.28,0.26,0.26,0.24,0.24,0.23,0.23,0.21,0.20,0.19,0.19
          ,0.18,0.18,0.17,0.17,0.16,0.16,0.15,0.15,0.14,0.15,0.14
          ,0.14,0.13,0.13,0.13,0.12,0.12,0.12,0.12,0.12,0.12,0.12
          ,0.12,0.12,0.12,0.12,0.11,0.11,0.10,0.10,0.11,0.10,0.09
          ,0.09,0.09,0.08,0.08,0.08,0.08,0.07,0.07,0.07,0.07,0.07
          ,0.07,0.07,0.07,0.07,0.07,0.07,0.07,0.07,0.07,0.06,0.06
          ,0.07,0.07,0.07,0.06,0.06,0.04,0.06,0.05,0.06)

other <- c( 0.05,0.05,0.05,0.05,0.03,0.03,0.04,0.04,0.04,0.04,0.04
           ,0.04,0.04,0.04,0.04,0.04,0.04,0.03,0.03,0.03,0.03,0.03
           ,0.03,0.03,0.03,0.03,0.03,0.02,0.03,0.03,0.03,0.03,0.03
           ,0.03,0.03,0.03,0.03,0.03,0.03,0.03,0.03,0.03,0.03,0.03
           ,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02
           ,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.02,0.01
           ,0.02,0.02,0.01,0.01,0.02,0.01,0.01,0.02,0.01,0.02,0.01
           ,0.01,0.01,0.00,0.01,0.02)

all <- data.frame(dem, rep, dts, other)
rowSums(all)
##  [1] 1.01 1.01 1.01 1.01 1.01 1.00 1.01 1.02 1.00 1.01 1.00 1.01 1.00
## [14] 1.01 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00
## [27] 1.00 1.00 1.00 1.01 1.00 1.01 1.00 1.00 1.01 1.01 1.00 1.01 1.01
## [40] 1.01 1.01 1.01 1.01 1.01 1.00 1.00 1.00 1.00 1.01 1.01 1.01 1.01
## [53] 1.01 1.00 1.00 1.01 1.01 1.01 1.01 1.01 1.01 1.00 1.01 1.01 1.01
## [66] 1.00 1.00 1.00 1.00 1.01 1.01 1.00 0.99 1.02 1.01 1.02 1.01 1.00
## [79] 1.00 0.99 0.98 0.99
# correct rounding errors from datathief
for (i in 1:length(age)) {
  all[i,] <- all[i,]/sum(all[i,]);
}
rowSums(all)
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [34] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [67] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## getting data list above
# x <- scan()
# [datathief numbers]
#
# round(matrix(x,ncol=2,byrow=TRUE)[,2],2)
# plot(round(matrix(x,ncol=2,byrow=TRUE)[,1],0))

# following example from http://learnr.wordpress.com/2009/03/29/ggplot2_marimekko_mosaic_chart/

################################################################################
df <- data.frame(
  segment = age
  , segpct = pop.prop * 100
  , Other = all$other * 100
```

```
  , DTS  = all$dts    * 100
  , Rep = all$rep     * 100
  , Dem = all$dem     * 100
)

df$xmax <- cumsum(df$segpct)
df$xmin <- df$xmax - df$segpct
df$segpct <- NULL

library(ggplot2)
library(reshape)
library(plyr)

dfm <- melt(df, id = c("segment", "xmin", "xmax"))

dfm1 <- ddply(dfm , .(segment), transform, ymax = cumsum(value))
dfm1 <- ddply(dfm1, .(segment), transform, ymin = ymax - value)

dfm1$xtext <- with(dfm1, xmin + (xmax - xmin)/2)
dfm1$ytext <- with(dfm1, ymin + (ymax - ymin)/2)

dfm1$segmentlabel <- rep("",length(dfm1$segment))
ss <- ((dfm1$segment %% 5)==0); # every 5 years, display age
dfm1$segmentlabel[ss] <- dfm1$segment[ss]
dfm1$segmentlabel[(dfm1$segment==18)] <- "age"

p <- ggplot(dfm1, aes(ymin = ymin, ymax = ymax, xmin = xmin, xmax = xmax, fill = variable))

p <- p + geom_rect(colour = I("grey"), alpha=0.75, size=.01) +
  xlab("Percentage age distribution") +
  ylab("Percent registered voter for party by age") +
  labs(title="NM Registered Voters - Party by Age (Oct 2008)")

p <- p + geom_text(aes(x = xtext, y = ytext,
                       label = ifelse(segment == 20, paste(variable), " ")), size = 3.5)

# This removes all legends
p <- p + theme(legend.position="none")
p <- p + theme_bw()

p <- p + geom_text(aes(x = xtext, y = -3, label = paste(dfm1$segmentlabel)), size = 3)
print(p)
```



(d) Justification

Now area is proportional to political power over all age groups. However, older people still vote in higher proportions than younger people, so older people still exercise more political power than shown here.

Labels are in the area they represent, rather than having a key which taxes the short-term memory to remember which color is associated with which party.

The axis is labelled with both percentage and the age of people. Now it's clear that 60 years and older account for only 25% of the voters, before it seemed close to 50%.