Data visualization and customization in R

3/1/2019

- What makes good data visualization in your field? Other fields?

- What is figure customization? Why customize plots? What aspects of customization are easy/difficult in R?
- How much do people use R for data visualization/customization alone or with a pdf editing software?

Tools for data visualization and customization in R

R users have a huge variety of options for visualizing their data. Some packages are highly specialized for particular data or analysis types. There are specific packages for text mining, phylogenetic trees, time series, model diagnostics and more.

For general purpose data visualization (bar plots, scatter plots, heat maps…), people tend to either use base graphics or `ggplot2`. I learned base R first and ggplot later. Now, I use a combination of the two depending on what I'm doing.

```
barplot(dat$total_bill,
        names.arg=dat$time,
        col="#AFC0CB",
        border=FALSE,
        main="Average Bill for Two-Person Meal")

ggplot(data=dat, aes(x=time, y=total_bill, fill=time)) +
  geom_bar(colour="black", fill="#DD8888", width=.8, stat="identity") +
  guides(fill=FALSE) +
  xlab("Time of day") + ylab("Total bill") +
  ggtitle("Average bill for 2 people")
```

The base and ggplot code looks visually different each other. Base R may look more familiar. Which plot do you prefer "out of the box"?

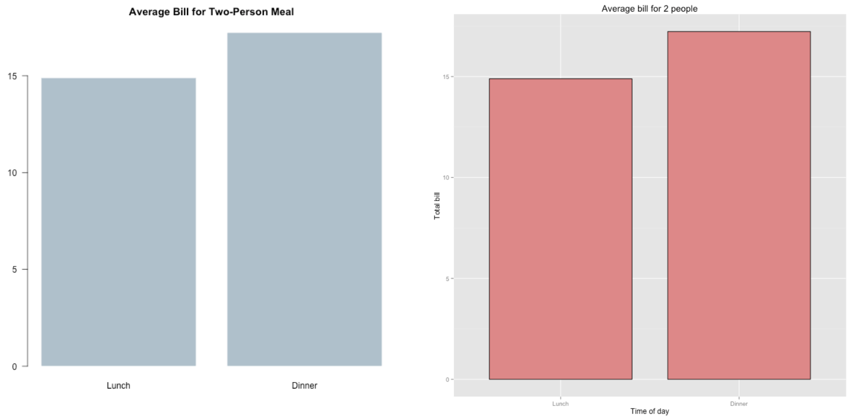


Figure 1: Base R and ggplot2 barplots.

Demonstration with data from FiveThirtyEight

In this demonstration, we'll analyze my attempts to recreate high quality, interpretable, and complex figures as seen on the statistical journalism website, fivethirtyeight.

Data I used were collected as part of a survey by FiveThirtyEight and WNYC Studios to figure out **What Men Think It Means To Be A Man?**.

I've attempted to reproduce a few figures in the article. Note - they've weighted their data by age and race/ethnicity so my plots using raw data will differ slightly.

Customizing plots with ggplot 2

```r
rawDat <- read_csv("masculinity-survey/raw-responses.csv")

head(rawDat[, 24:28])
```

```
## # A tibble: 6 x 5
##   q0008_0001   q0008_0002   q0008_0003            q0008_0004   q0008_0005
##   <chr>        <chr>        <chr>                 <chr>        <chr>
## 1 Not selected Not selected Your hair or hairline Not selected Not selected
## 2 Not selected Your weight  Not selected          Not selected Not selected
## 3 Not selected Not selected Not selected          Not selected Not selected
## 4 Not selected Not selected Not selected          Not selected Not selected
## 5 Not selected Your weight  Not selected          Not selected Not selected
## 6 Not selected Not selected Not selected          Not selected Not selected
```

```r
# data for plotting
dat2 <- rawDat %>% select(24:35) # columns used for figure 2
colnames(dat2) <- c("Your height", "Your weight", "Your hair",
                    "Your physique", "Appearance of genitals",
                    "Your clothing or style", "Sexual performance",
                    "Your mental health", "Your physical health",
                    "Your finances", "Your ability to provide",
                    "None of the above")

# wrangling for ggplot
dat2 %<>% gather("variable", "answer") %>%
  count(variable, answer) %>%
  group_by(variable) %>%
  mutate(prop=n/sum(n)) %>%
  mutate(percent=prop*100) %>%
  mutate(answer=ifelse(answer!="Not selected", "Yes", "No"))

ords <- dat2 %>%
  filter(answer=="Yes") %>%
  group_by(variable) %>%
  arrange(desc(prop))

ords$ord <- 1:nrow(ords)
ords %<>% select(variable, ord)
noa <- ords[8, ]
ords <- ords[-8, ]
ords <- bind_rows(ords, noa)
```

```r
ords$ord <- 1:nrow(ords)
dat2 <- left_join(dat2, ords)

dat2
```

```
## # A tibble: 24 x 6
## # Groups:   variable [?]
##    variable                answer     n   prop percent   ord
##    <chr>                   <chr>  <int>  <dbl>   <dbl> <int>
##  1 Appearance of genitals  Yes      148 0.0916    9.16    10
##  2 Appearance of genitals  No      1467 0.908    90.8     10
##  3 None of the above       Yes      259 0.160    16.0     12
##  4 None of the above       No      1356 0.840    84.0     12
##  5 Sexual performance      No      1261 0.781    78.1      7
##  6 Sexual performance      Yes      354 0.219    21.9      7
##  7 Your ability to provide No      1171 0.725    72.5      5
##  8 Your ability to provide Yes      444 0.275    27.5      5
##  9 Your clothing or style  No      1415 0.876    87.6      9
## 10 Your clothing or style  Yes      200 0.124    12.4      9
## # ... with 14 more rows
```

```r
# visual variables
col2 <- c("#ed713a", "#e1e1e1")

# defining ggplot
p2 <- ggplot(dat2, aes(x = reorder(variable, -ord), y = percent,
                 fill = factor(answer, levels = c("No", "Yes")))) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=rev(col2)) + # real meat of the customization starts here
  coord_flip() +
  guides(fill = FALSE) +
  xlab("") +
  ylab(" ") +
  ggtitle("") +
  theme_fivethirtyeight() +
  theme(axis.text.y = element_text(hjust = 0),
        axis.title.x = element_text(hjust = 1),
        panel.grid = element_blank())

ggdraw(p2) +
  draw_text("What do you worry about on a \n near-daily basis?",
                      x = 0.01, y = 0.98, hjust = 0, vjust = 1)
```

- Is it easy to add percents to the left of the bars?
- Is this “good enough” for the purposes you come across? (e.g., quick analysis, lab presentation, publication)

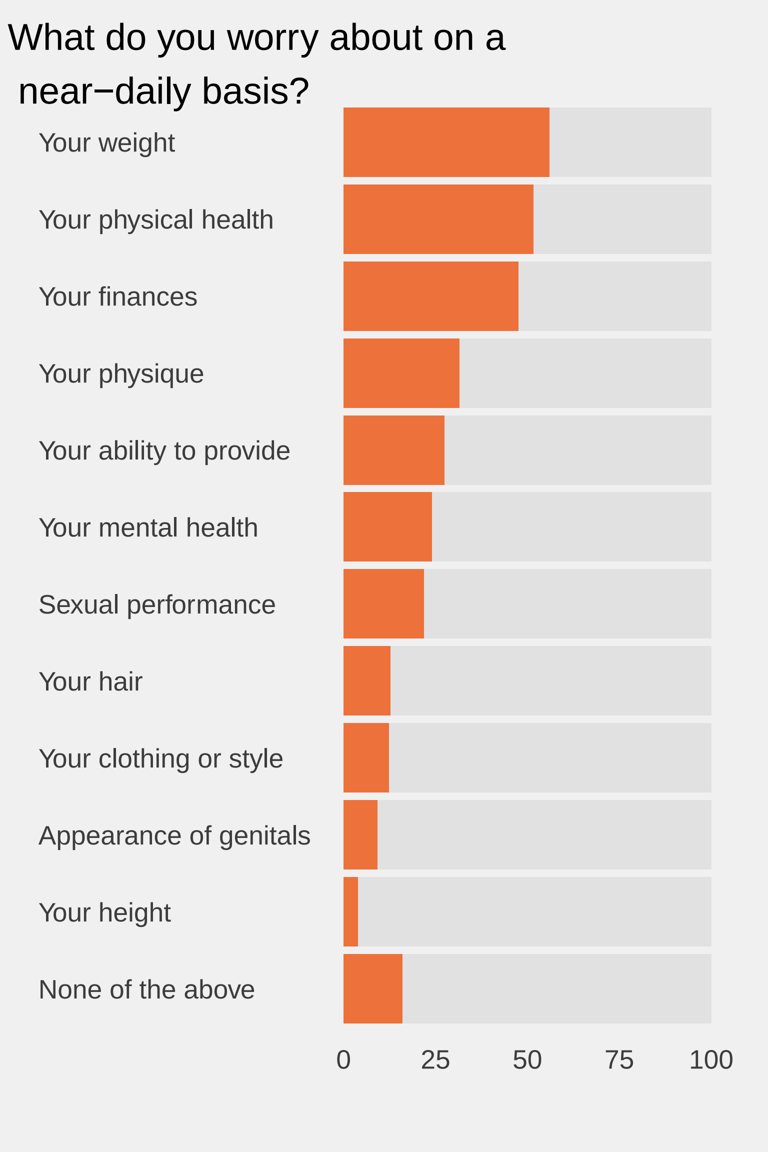


Figure 2: Second figure in article reproduced using ggplot.

Doing the same thing in base R…

```r
# data for plotting
dat2 <- rawDat %>% select(24:35)
colnames(dat2) <- c("Your height", "Your weight", "Your hair",
                    "Your physique", "Appearance of genitals",
                    "Your clothing or style", "Sexual performance",
                    "Your mental health", "Your physical health",
                    "Your finances", "Your ability to provide",
                    "None of the above")

# data wrangling for base R barplot
dat2 %<>% gather("variable", "answer") %>%
  count(variable, answer) %>%
  group_by(variable) %>%
  mutate(prop=n/sum(n)) %>%
  mutate(percent=prop*100) %>%
  mutate(answer=ifelse(answer!="Not selected", "Yes", "No")) %>%
  select(answer, percent) %>% # keep only these columns
  spread(answer, percent) %>% # turn "answer" column
                              # into "yes" and "no"
                              # columns with "percent" as the value
  select(variable,Yes,No) # fix order of columns

# sort
dat2 %<>% arrange(Yes)

dat2 <- rbind(filter(dat2, variable=="None of the above"),
              filter(dat2, variable!="None of the above"))

# make into a regular data frame to remove all grouping from tbl
dat2 <- as.data.frame(dat2)

# make into matrix, transpose
row.names(dat2) <- dat2$variable
dat2 %<>%
  select(-1) %>%
  t()

dat2
```

```
##     None of the above Your height Appearance of genitals
## Yes          16.03715    3.900929               9.164087
## No           83.96285   96.099071              90.835913
##     Your clothing or style Your hair Sexual performance Your mental health
## Yes                12.3839   12.6935            21.9195           24.02477
## No                 87.6161   87.3065            78.0805           75.97523
##     Your ability to provide Your physique Your finances
## Yes                27.49226      31.57895       47.6161
## No                 72.50774      68.42105       52.3839
##     Your physical health Your weight
## Yes             51.70279    55.97523
## No              48.29721    44.02477
```

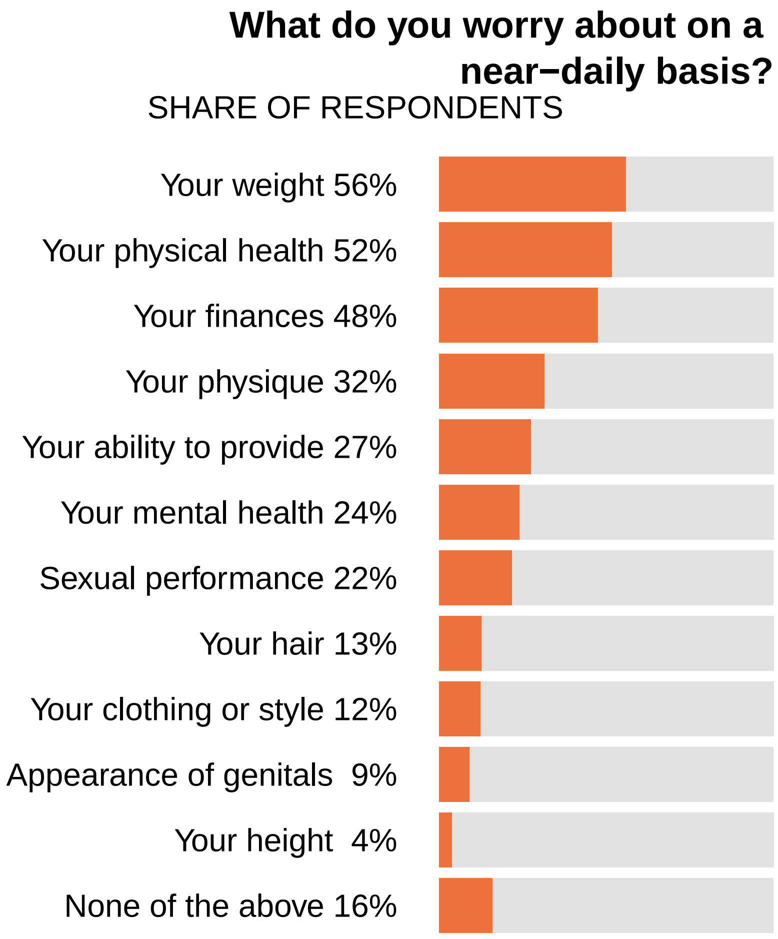


Figure 3: Second figure in article reproduced using base R.

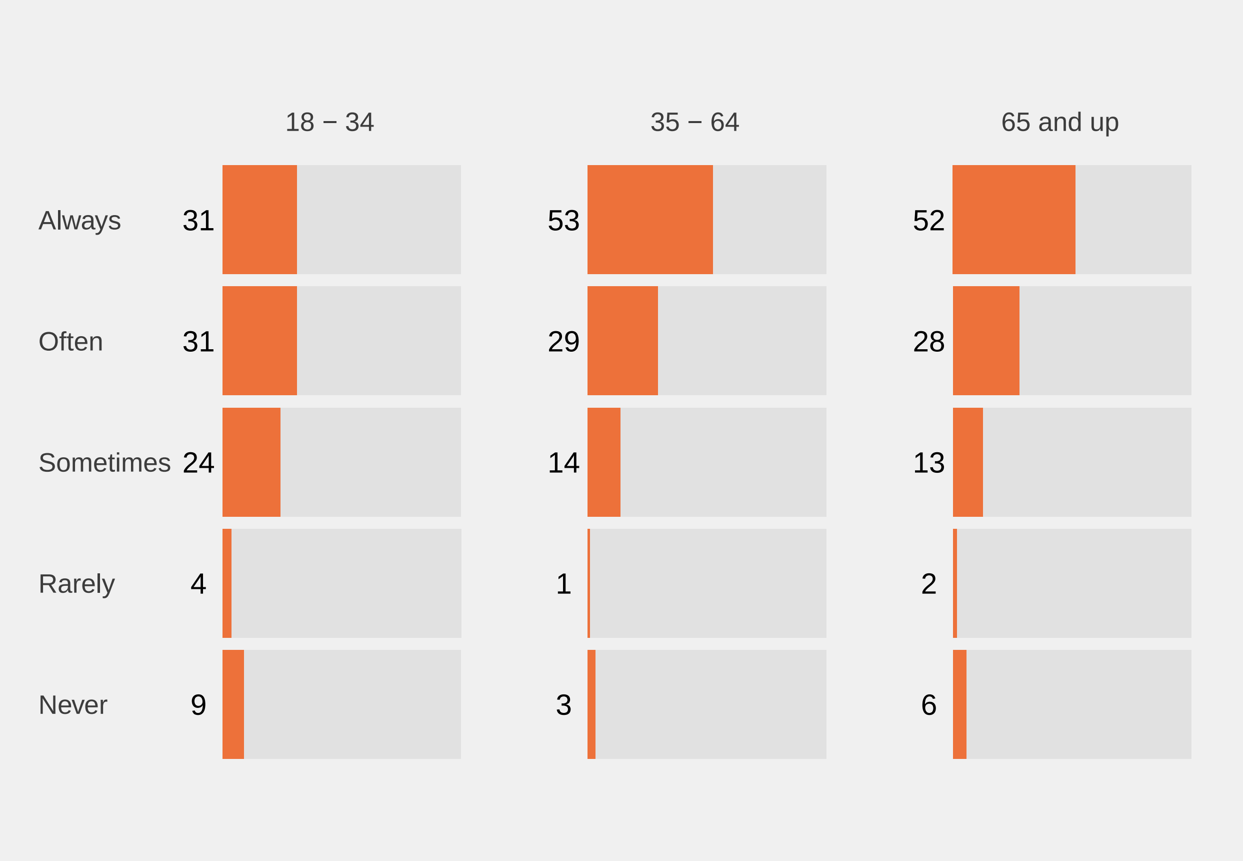
```r
# visual variables
col2 <- c(orange = "#ed713a", grey = "#e1e1e1")
par(mar=c(4,15,4,1)+0.1)

# base R plot
barplot(height = dat2,
        horiz=TRUE,
        axes=FALSE, ann=FALSE, # remove axis lines and tick marks
        adj=1, # right justify
        main = "What do you worry about on a \n near-daily basis?",
        names.arg = paste(
          colnames(dat2),
          paste0(format(dat2['Yes',], digits = 0), "%")
          ),
        las=1, # orient labels horizontally
        border = NA, col = col2 # style the bars
        )
mtext("SHARE OF RESPONDENTS",side=3, at=-25)
```

- Which plot do you prefer? Why?
- Which required more lines of code?

- Is it easier to wrangle data for ggplot or base R?

Faceting with ggplot and adding summary statistics

```r
dat5 <- rawDat %>% select(c(61, 95)) %>% rename(answer=q0018)

dat5 %<>%
  filter(answer!="No answer") %>%
  count(answer, age3) %>%
  group_by(age3) %>%
  mutate(prop=n/sum(n)) %>%
  mutate(yes=prop*100) %>%
  mutate(no=100-yes) %>%
  select(answer, age3, yes, no) %>%
  gather("response", "percent", 3:4)

dat5$age3 <- factor(dat5$age3,
                    levels=c("18 - 34","35 - 64", "65 and up"))
dat5$answer <- factor(dat5$answer,
                      levels=rev(c("Always", "Often",
                                   "Sometimes", "Rarely", "Never")))

col2 <- c("#ed713a", "#e1e1e1")

dat5 %>%
  ggplot(aes(x=answer, y=percent, fill=response)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values=rev(col2)) +
  coord_flip(clip = "off") +
  facet_grid(~ age3) +
  guides(fill = FALSE) +
  xlab("") +
  ylab(" ") +
  ggtitle("") +
  theme_fivethirtyeight() +
  theme(axis.text.x = element_blank(),
        axis.title.x = element_text(hjust = 1, vjust=1),
        axis.text.y = element_text(hjust = 0),
        panel.grid = element_blank(),
        panel.spacing = unit(2, "lines")) -> p5

mylabels <- dat5 %>% filter(response=="yes") %>% group_by(answer, age3) %>%
  summarise(label=round(percent)) %>%
  mutate(percent=-10, response="yes")

p5 + geom_text(data=mylabels, aes(label=label))
```



Combining multiple plots with ggplot

```r
dat1 <- rawDat %>% select(q0005, age3)

fig1a <- dat1 %>%
  group_by(q0005) %>%
  tally() %>% mutate(prop=(n/(sum(n))*100)) %>%
  mutate(question=1)

# color scale with grey
gg_color_hue <- function(n) {
  hues = seq(15, 375, length = n + 1)
  hcl(h = hues, l = 65, c = 100)[1:n]
}
col3 <- gg_color_hue(2)
col3 <- c(col3[1], "grey", col3[2])

fig1a %>%
  ggplot(aes(x=question, y=prop, fill=q0005)) +
  geom_bar(stat="identity") + coord_flip() +
  scale_fill_manual(values=rev(col3)) +
  scale_y_continuous(position = "right") +
  xlab("") + ylab("") + ggtitle("Do you think that society puts presson men in a way that is unhealthy
  theme_fivethirtyeight() +
  theme(legend.position="none",
        axis.text.x  = element_text(size=10),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        plot.margin = margin(0, 0.5, 0, 2, "cm"),
        plot.title = element_text(size=14)) +
  geom_text(x=1, y=10, label="Yes", size=5) +
  geom_text(x=1, y=70, label="No", size=5) -> p1a

fig1b <- dat1 %>%
  group_by(age3, q0005) %>%
  tally() %>% mutate(prop=(n/(sum(n))*100))

fig1b$ord <- rep(c(1,2,3), each=3)

fig1b %>%
  ggplot(aes(x=reorder(age3, -ord), y=prop, fill=q0005)) +
  geom_bar(stat="identity") + coord_flip() +
  #scale_x_discrete(limits = rev(levels(age3))) +
  scale_fill_manual(values=rev(col3)) +
  xlab("") + ylab("") +
  theme_fivethirtyeight() +
  theme(legend.position="none",
        axis.ticks.y=element_blank(),
        plot.margin = margin(1, 0.5, 1.5, 0.5, "cm"))  -> p1b

plot_grid(p1a, p1b, ncol = 1, rel_heights = c(.25, .75))
```

- Spacing out bars?