Factors

Here are some examples from Chapter 15. The examples are related to the General Social Survey from NORC at the University of Chicago.

```r
library(tidyverse)
library(forcats)

# A tibble: 21,483 x 9
#  year marital age race rincome partyid relig denom tvhours
# 1  2000 Never married 26 White $8000 to $9999 - Ind, near - Prot - Southern - 12
# 2  2000 Divorced 48 White $8000 to $9999 - Not strongly - Prot - Baptist - NA
# 3  2000 Widow 67 White Not applicable - Independent - Prot - No denomination - 2
# 4  2000 Never married 39 White Not applicable - Ind, near - Orthodox - Not applicable - 4
# 5  2000 Divorced 25 White Not applicable - Not strongly - None - Not applicable - 1
# 6  2000 Married 25 White $20000 - Strongly - Prot - Southern - NA
# 7  2000 Never married 36 White $25000 or more - Not strongly - Christ - Not applicable - 3
# 8  2000 Divorced 44 White $7000 to $9999 - Ind, near - Prot - Lutheran - NA
# 9  2000 Married 44 White $25000 or more - Not strongly - Prot - Other - 0
# 10 2000 Married 47 White $25000 or more - Strongly - Prot - Southern - 3
# ... with 21,473 more rows

gss_cat %>%
  count(race)
# A tibble: 3 x 2
# race  n
# 1 Other  1959
# 2 Black  3129
# 3 White 16395
```

Factor variables are used to make bar charts. The `geom_bar()` counts the observations in each level of the factor.

```r
ggplot(gss_cat, aes(race)) +
  geom_bar()
```

Forcing NAs.

```r
ggplot(gss_cat, aes(race)) + geom_bar() + scale_x_discrete(drop = FALSE)
```
Modifying the order of a factor.

Examine tv watch time by religion.

```r
relig_summary <- gss_cat %>%
  group_by(relig) %>%
  summarise(
    age = mean(age, na.rm = TRUE),
    tvhours = mean(tvhours, na.rm = TRUE),
    n = n()
  )

relig_summary %>% ggplot(aes(tvhours, relig)) + geom_point()
```
```r
relig_summary %>% ggplot(aes(tvhours, fct_reorder(relig, tvhours))) + geom_point()
```
The `fct_reorder()` function should be used in a mutate statement.

Same as the last code.

```r
relig_summary %>%
  mutate(relig = fct_reorder(relig, tvhours)) %>%
  ggplot(aes(tvhours, relig)) +
  geom_point()
```
Now tv watch time by average age.

```r
rincome_summary <- gss_cat %>%
  group_by(rincome) %>%
  summarise(
    age = mean(age, na.rm = TRUE),
    tvhours = mean(tvhours, na.rm = TRUE),
    n = n()
  )

rincome_summary %>% ggplot(aes(age, fct_reorder(rincome, age))) +
  geom_point()
```
Does this make sense? What is wrong with this plot?

```r
rincome_summary %>% ggplot(aes(age, fct_relevel(rincome, "Not applicable"))) + geom_point()
```
Using `mutate()`

```r
gss_cat %>% ggplot(aes(marital)) + geom_bar()
```
ggplot {gss_cat} %>% mutate(marital = marital) %>%
ggplot(aes(marital)) +
geom_bar()
gss_cat %>% mutate(marital = marital %>% fct_infreq()) %>%
ggplot(aes(marital)) +
  geom_bar()
gss_cat %>% mutate(marital = marital %>% fct_infreq() %>% fct_rev()) %>%
ggplot(aes(marital)) + geom_bar()
Modifying factor levels.

```r
gss_cat %>% count(partyid)
```

```r
## # A tibble: 10 x 2
## #  partyid n
## #   <fct> <int>
## 1 No answer     154
## 2 Don’t know     1
## 3 Other party    393
## 4 Strong republican 2314
## 5 Not str republican 3032
## 6 Ind,near rep   1791
## 7 Independent    4119
## 8 Ind,near dem   2499
## 9 Not str democrat 3690
## 10 Strong democrat 3490
```

Re-coding

```r
gss_cat %>%
  mutate(partyid = fct_recode(partyid,
    "Republican, strong" = "Strong republican",
    "Republican, weak" = "Not str republican",
    "Independent, near rep" = "Ind,near rep",
    "Independent, near dem" = "Ind,near dem",
    "Democrat, weak" = "Not str democrat",
    "Democrat, strong" = "Strong democrat"
  ))
```

12
```r
count(partyid)

## A tibble: 10 x 2
## partyid     n
## <fct>       <int>
## 1 No answer  154
## 2 Don't know  1
## 3 Other party 393
## 4 Republican, strong 2314
## 5 Republican, weak 3032
## 6 Independent, near rep 1791
## 7 Independent 4119
## 8 Independent, near dem 2499
## 9 Democrat, weak 3690
## 10 Democrat, strong 3490

Other category
gss_cat %>%
  mutate(partyid = fct_recode(partyid,
                             "Republican, strong" = "Strong republican",
                             "Republican, weak" = "Not str republican",
                             "Independent, near rep" = "Ind,near rep",
                             "Independent, near dem" = "Ind,near dem",
                             "Democrat, weak" = "Not str democrat",
                             "Democrat, strong" = "Strong democrat",
                             "Other" = "No answer",
                             "Other" = "Don't know",
                             "Other" = "Other party")
) %>%
  count(partyid)

## A tibble: 8 x 2
## partyid     n
## <fct>       <int>
## 1 Other 548
## 2 Republican, strong 2314
## 3 Republican, weak 3032
## 4 Independent, near rep 1791
## 5 Independent 4119
## 6 Independent, near dem 2499
## 7 Democrat, weak 3690
## 8 Democrat, strong 3490

Collapse a factor
gss_cat %>%
  mutate(partyid = fct_collapse(partyid,
                             other = c("No answer", "Don't know", "Other party"),
                             rep = c("Strong republican", "Not str republican"),
                             ind = c("Ind,near rep", "Independent", "Ind,near dem"),
                             dem = c("Not str democrat", "Strong democrat")
) %>%
  count(partyid)

## A tibble: 4 x 2
```

13
## partyid n
## <fct> <int>
## 1 other 548
## 2 rep 5346
## 3 ind 8409
## 4 dem 7180