Factors

Here are some examples from Chapter 15. The examples are related to the [General Social Survey](http://gss.norc.org/) from NORC at the Unversity of Chicago.

library(tidyverse)
library(forcats)

gss\_cat

## # A tibble: 21,483 x 9
## year marital age race rincome partyid relig denom tvhours
## <int> <fct> <int> <fct> <fct> <fct> <fct> <fct> <int>
## 1 2000 Never ma… 26 White $8000 to… Ind,near … Protes… Southe… 12
## 2 2000 Divorced 48 White $8000 to… Not str r… Protes… Baptis… NA
## 3 2000 Widowed 67 White Not appl… Independe… Protes… No den… 2
## 4 2000 Never ma… 39 White Not appl… Ind,near … Orthod… Not ap… 4
## 5 2000 Divorced 25 White Not appl… Not str d… None Not ap… 1
## 6 2000 Married 25 White $20000 -… Strong de… Protes… Southe… NA
## 7 2000 Never ma… 36 White $25000 o… Not str r… Christ… Not ap… 3
## 8 2000 Divorced 44 White $7000 to… Ind,near … Protes… Luther… NA
## 9 2000 Married 44 White $25000 o… Not str d… Protes… Other 0
## 10 2000 Married 47 White $25000 o… Strong re… Protes… Southe… 3
## # … with 21,473 more rows

gss\_cat %>%
 count(race)

## # A tibble: 3 x 2
## race n
## <fct> <int>
## 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.

ggplot(gss\_cat, aes(race)) +
 geom\_bar()



Forcing NAs.

ggplot(gss\_cat, aes(race)) +
 geom\_bar() +
 scale\_x\_discrete(drop = FALSE)



Modifying the order of a factor.

Examine tv watch time by religion.

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()



relig\_summary %>% ggplot(aes(tvhours, fct\_reorder(relig, tvhours))) +
 geom\_point()



The *fct\_reorder()* functon should be used in a mutate statement.

Same as the last code.

relig\_summary %>%
 mutate(relig = fct\_reorder(relig, tvhours)) %>%
 ggplot(aes(tvhours, relig)) +
 geom\_point()



Now tv watch time by average age.

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?

rincome\_summary %>%ggplot(aes(age, fct\_relevel(rincome, "Not applicable"))) +
 geom\_point()



Using *mutate()*

gss\_cat %>% ggplot(aes(marital)) +
 geom\_bar()



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.

gss\_cat %>% 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 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

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