MATH 118: Notes H

forcats

The R package forcats is designed to make working with categorical variables easier and more efficient. It provides a set of functions that allow you to manipulate and analyze categorical data with ease. In this lesson, we'll cover the basics of the `forcats` package and some of its most useful functions.

Categorical Variables

Let's review what categorical data is. Categorical data is a type of data that consists of categories or labels.

Examples of categorical data include:

- Colors (red, blue, green, etc.)
- Types of vehicles (sedan, SUV, truck)
- Educational degrees (high school, college, graduate school)

Categorical data can be further divided into two types: *nominal* and *ordinal*. Nominal data consists of categories that have no inherent order, while ordinal data consists of categories that have a natural order. For example, educational degrees are ordinal data because they can be ordered from least to most advanced.

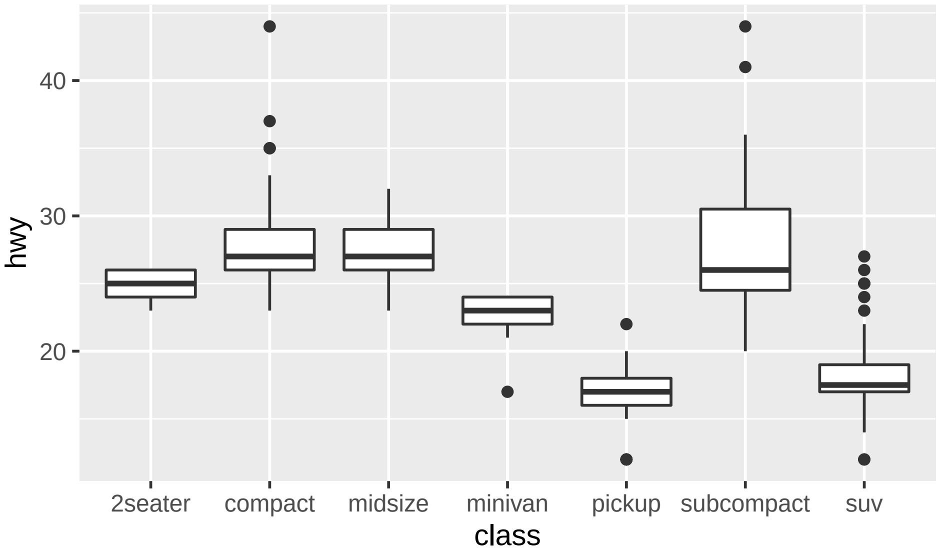
`mpg` Data

We will play with different functions in the `forcats` packages using the `mpg` dataset from earlier in the semester.

```
library(forcats)
library(tidyverse)
data("mpg")
```

Recall our side-by-side boxplot:

```
mpg %>%
  ggplot(aes(x=class, y=hwy)) +
  geom_boxplot()
```



## Reordering Factor Levels

One of the most useful functions is fct_relevel(), which allows you to reorder the levels of a factor. This can be useful when you want to change the default ordering of the levels or when you want to group certain levels together.

Is `class` a factor?

```
mpg$class %>%
  is.factor()
```

```
## [1] FALSE
```

Let's make it a factor!

```
mpg$class <- mpg$class %>%
  as.factor()

mpg$class %>%
  is.factor()
```

```
## [1] TRUE
```

Let's check the levels and their current ordering!

```
mpg$class %>%
  levels()
```

```
## [1] "2seater"    "compact"    "midsize"    "minivan"    "pickup"
## [6] "subcompact" "suv"
```

To reorder the levels:

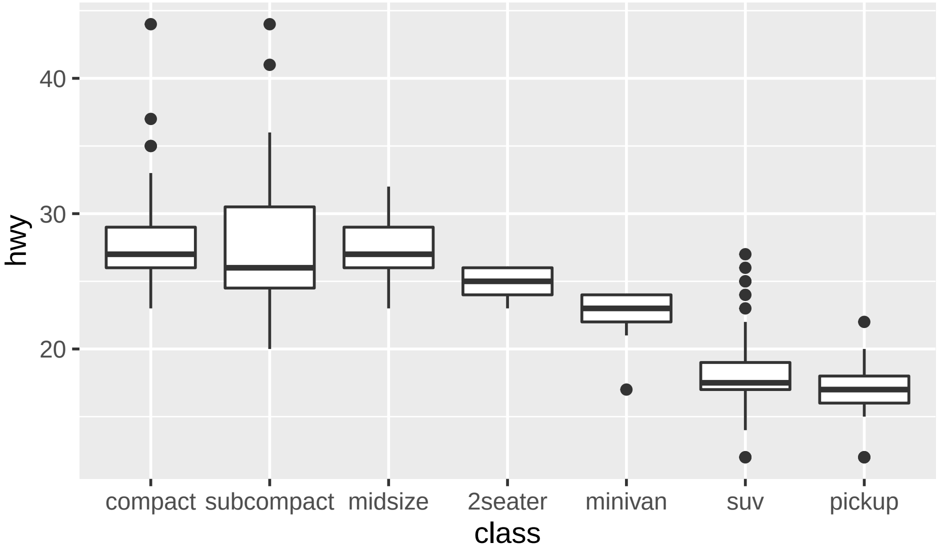
```
mpg$class <- mpg$class  %>%
  fct_relevel("compact","subcompact","midsize","2seater","minivan","suv","pickup")

mpg$class %>%
  levels()
```

```
## [1] "compact"    "subcompact" "midsize"    "2seater"    "minivan"
## [6] "suv"        "pickup"
```

Let's recreate our side-by-side boxplot now:

```
mpg %>%
  ggplot(aes(x=class, y=hwy)) +
  geom_boxplot()
```



Rather than reordering them manually by typing the order, you could also re-level by some numeric criteria. For example:

```
mpg$class <- mpg$class %>%
  fct_reorder(mpg$cty, median)

mpg$class %>%
  levels()
```

```
## [1] "suv"        "pickup"     "2seater"    "minivan"    "midsize"
## [6] "subcompact" "compact"
```

Renaming Factor levels

Sometimes you might not like the way the levels are named.

```
mpg$class <- mpg$class  %>%
  fct_recode("two-seater" = "2seater")

## NEW NAME = OLD NAME

mpg$class %>%
  levels()
```

```
## [1] "suv"        "pickup"     "two-seater" "minivan"    "midsize"
## [6] "subcompact" "compact"
```

```
#Check out the change in the mpg dataset
```

Factor Collapsing

Let’s say we wanted to create only two categories – cars and larger vehicles.

```
mpg$class_two <- mpg$class %>%
  fct_collapse(cars = c("compact", "subcompact", "midsize", "two-seater"),
               big = c("pickup", "suv", "minivan"))

mpg$class_two %>%
  levels()
```

```
## [1] "big"  "cars"
```

Lumping into an other category

- fct_lump_min(): lumps levels that appear fewer than min times.
- fct_lump_prop(): lumps levels that appear in fewer than (or equal to) prop * n times.
- fct_lump_n() lumps all levels except for the n most frequent (or least frequent if n < 0)

```
table(mpg$manufacturer)
```

```
## 
##       audi  chevrolet      dodge       ford      honda    hyundai       jeep 
##         18         19         37         25          9         14          8 
## land rover    lincoln    mercury     nissan    pontiac     subaru     toyota 
##          4          3          4         13          5         14         34 
## volkswagen 
##         27
```

Let's say we wanted only the manufacturers with at least 15 cars produced. Everything else we want to just be other:

```
mpg$manufacturer <- mpg$manufacturer %>%   fct_lump_min(15)

mpg$manufacturer %>%
  levels()
```

```
## [1] "audi"       "chevrolet"  "dodge"      "ford"       "toyota"    
## [6] "volkswagen" "Other"
```