Lecture 17: Advanced summarizing

Taylor Arnold

Advanced summarizing

I wrote the function `group_summarize` because I found that students struggled using the raw summarizing commands early in the semester.

You may find that you need to do some time of summarization that we did not cover, so here are some notes on how to do it.

group_by

We have to use the function group_by and the function summarize on the dataset. The first tells R which variables to summarize by, but the second tells it which new variables to create:

```
summarize(group_by(bikes, season),
          min_temp = min(temp), max_temp = max(temp))
```

```
## # A tibble: 4 x 3
##   season   min_temp max_temp
##    <int>      <dbl>    <dbl>
## 1      1 -12.097394 21.78500
## 2      2   0.700838 37.34998
## 3      3  14.965022 40.87002
## 4      4  -1.425022 27.39500
```

summarize

Each of the new variables, however, must be described explicitly. Here we are able to compute the minimum and maximum

grouped mutate

Group by can also be combined with the `mutate` function to append summary statistics to a group of variables. For example, if we wanted to add the average temperature of each season to every row of the dataset, we would do this:

```
mutate(group_by(bikes, season), avg_temp = mean(temp))
```

grouped top_n

The `top_n` function that we saw last time can also be used with the `group_by` function. Here we find the hottest day in each season:

```
top_n(group_by(bikes, season), n = 1, temp)
```

```
## # A tibble: 5 x 3
## # Groups:   season [4]
##   season  year     temp
##    <int> <int>    <dbl>
## 1      2     0 37.34998
## 2      1     1 21.78500
## 3      3     1 40.87002
## 4      4     1 27.39500
## 5      4     1 27.39500
```