Topic 4 Task answers

R. S. Verspoor

2024-07-19

Task 1 answers

Produce a table of estimates for the mean and variance of both sepal lengths and widths, within each species.

```
iris %>%
    group_by(Species) %>%
    summarise(mnL = mean(Sepal.Length), varL = var(Sepal.Length),
              mnW = mean(Sepal.Width), varW = var(Sepal.Width))
```

```
## # A tibble: 3 × 5
##   Species      mnL  varL   mnW   varW
##   <fct>      <dbl> <dbl> <dbl>  <dbl>
## 1 setosa      5.01 0.124  3.43 0.144
## 2 versicolor  5.94 0.266  2.77 0.0985
## 3 virginica   6.59 0.404  2.97 0.104
```

Task 2 answers

2.1 Can you produce a mean GDP for each country, averaging over years.

```
gp_income %>%
    group_by(country) %>%
    summarise(mn = mean(gdp))
```

```
## # A tibble: 203 × 2
##    country                 mn
##    <chr>                <dbl>
##  1 Afghanistan          1221.
##  2 Albania              6549.
##  3 Algeria             11081.
##  4 Andorra             35455.
##  5 Angola               4888.
##  6 Antigua and Barbuda 20017.
##  7 Argentina           12909.
##  8 Armenia              4626.
##  9 Aruba               38928.
## 10 Australia           36583.
## # ℹ 193 more rows
```

2.2 Now try to produce the mean GDP for each year, averaged across country.

```
gp_income %>%
    group_by(year) %>%
    summarise(mn = mean(gdp))
```

```
## # A tibble: 25 × 2
##     year     mn
##    <dbl>  <dbl>
##  1  1991 12557.
##  2  1992 12623.
##  3  1993 12656.
##  4  1994 12886.
##  5  1995 13172.
##  6  1996 13470.
##  7  1997 13949.
##  8  1998 14221.
##  9  1999 14442.
## 10  2000 14905.
## # ℹ 15 more rows
```

2.3 Produce a tidy data set called gp_hiv using the tools in tidyverse that we introduced above. The dataset needs to run from 1991 onwards, and we want to end up with columns country, year and prevalence.

```
#note depending on your system (mac or pc) you may need to
#give the file name as "indicator hiv estimated prevalence% 15-49.csv"
#or "indicator hiv estimated prevalence_ 15-49.csv"
#use dir() to check the file name
gp_hiv <- read_csv("indicator hiv estimated prevalence_ 15-49.csv") %>%
            rename(country = `Estimated HIV Prevalence% - (Ages 15-49)`) %>%
            gather(year, prevalence, -country) %>%
            mutate(year = as.numeric(year)) %>%
            filter(!is.na(country)) %>%
            filter(!is.na(prevalence)) %>%
            filter(year > 1990) %>%
            mutate(prevalence = as.numeric(prevalence))
```

```
## Rows: 275 Columns: 34
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (1): Estimated HIV Prevalence% - (Ages 15-49)
## dbl (31): 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1990, 1991, ...
## lgl  (2): 1988, 1989
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```