TEST2

Olivia Mahony

11/7/2021

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.4 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 2.0.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)

#1) Load and output the relig\_income data set. How many rows and columns does it have ?

18 x 11

relig\_income -> relig  
relig

## # A tibble: 18 × 11  
## religion `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Agnostic 27 34 60 81 76 137 122  
## 2 Atheist 12 27 37 52 35 70 73  
## 3 Buddhist 27 21 30 34 33 58 62  
## 4 Catholic 418 617 732 670 638 1116 949  
## 5 Don’t k… 15 14 15 11 10 35 21  
## 6 Evangel… 575 869 1064 982 881 1486 949  
## 7 Hindu 1 9 7 9 11 34 47  
## 8 Histori… 228 244 236 238 197 223 131  
## 9 Jehovah… 20 27 24 24 21 30 15  
## 10 Jewish 19 19 25 25 30 95 69  
## 11 Mainlin… 289 495 619 655 651 1107 939  
## 12 Mormon 29 40 48 51 56 112 85  
## 13 Muslim 6 7 9 10 9 23 16  
## 14 Orthodox 13 17 23 32 32 47 38  
## 15 Other C… 9 7 11 13 13 14 18  
## 16 Other F… 20 33 40 46 49 63 46  
## 17 Other W… 5 2 3 4 2 7 3  
## 18 Unaffil… 217 299 374 365 341 528 407  
## # … with 3 more variables: $100-150k <dbl>, >150k <dbl>,  
## # Don't know/refused <dbl>

#2) The relig\_income data set has column names that are not variables and the data set is much to wide. Use R code and specifically tidying code that will produce a longer table with variable names for the columns. A partial table is given below. Your output should have at least ten rows.

relig%>%  
 select(religion, everything())%>%  
 pivot\_longer(col = 2:11, names\_to = "income", values\_to = "frequency") -> relig1  
relig1

## # A tibble: 180 × 3  
## religion income frequency  
## <chr> <chr> <dbl>  
## 1 Agnostic <$10k 27  
## 2 Agnostic $10-20k 34  
## 3 Agnostic $20-30k 60  
## 4 Agnostic $30-40k 81  
## 5 Agnostic $40-50k 76  
## 6 Agnostic $50-75k 137  
## 7 Agnostic $75-100k 122  
## 8 Agnostic $100-150k 109  
## 9 Agnostic >150k 84  
## 10 Agnostic Don't know/refused 96  
## # … with 170 more rows

#3) Now use R code to produce a data set that only reflects observations for income levels $100-150k. A partial table is given below. Your output should show all 18 rows. A partial table is given below.

relig1%>%  
 select(religion, income, frequency)%>%  
 filter(income == "$100-150k") -> relig2  
relig2

## # A tibble: 18 × 3  
## religion income frequency  
## <chr> <chr> <dbl>  
## 1 Agnostic $100-150k 109  
## 2 Atheist $100-150k 59  
## 3 Buddhist $100-150k 39  
## 4 Catholic $100-150k 792  
## 5 Don’t know/refused $100-150k 17  
## 6 Evangelical Prot $100-150k 723  
## 7 Hindu $100-150k 48  
## 8 Historically Black Prot $100-150k 81  
## 9 Jehovah's Witness $100-150k 11  
## 10 Jewish $100-150k 87  
## 11 Mainline Prot $100-150k 753  
## 12 Mormon $100-150k 49  
## 13 Muslim $100-150k 8  
## 14 Orthodox $100-150k 42  
## 15 Other Christian $100-150k 14  
## 16 Other Faiths $100-150k 40  
## 17 Other World Religions $100-150k 4  
## 18 Unaffiliated $100-150k 321

#4) Use and show R code that will calculate the sum for all frequency values. What is the sum?

sum(relig2$frequency) -> x  
x

## [1] 3197

#5) Using your sum for the frequency column, Use and show R code to create another column that produces proportion values for each observation. A partial table is given below. Your output should show 18 rows.

relig2%>%  
 select(religion, income, frequency)%>%  
 mutate(propotionoftotal = (frequency/3197)) -> relig3  
relig3

## # A tibble: 18 × 4  
## religion income frequency propotionoftotal  
## <chr> <chr> <dbl> <dbl>  
## 1 Agnostic $100-150k 109 0.0341   
## 2 Atheist $100-150k 59 0.0185   
## 3 Buddhist $100-150k 39 0.0122   
## 4 Catholic $100-150k 792 0.248   
## 5 Don’t know/refused $100-150k 17 0.00532  
## 6 Evangelical Prot $100-150k 723 0.226   
## 7 Hindu $100-150k 48 0.0150   
## 8 Historically Black Prot $100-150k 81 0.0253   
## 9 Jehovah's Witness $100-150k 11 0.00344  
## 10 Jewish $100-150k 87 0.0272   
## 11 Mainline Prot $100-150k 753 0.236   
## 12 Mormon $100-150k 49 0.0153   
## 13 Muslim $100-150k 8 0.00250  
## 14 Orthodox $100-150k 42 0.0131   
## 15 Other Christian $100-150k 14 0.00438  
## 16 Other Faiths $100-150k 40 0.0125   
## 17 Other World Religions $100-150k 4 0.00125  
## 18 Unaffiliated $100-150k 321 0.100

#6) Now use R code to make the entries in the income column more descriptive and precise. Change $100 – 150k to $100k - $150k . A partial table is given below. Your table should show all 18 rows.

mutate(relig3, income = "$100k-$150k")

## # A tibble: 18 × 4  
## religion income frequency propotionoftotal  
## <chr> <chr> <dbl> <dbl>  
## 1 Agnostic $100k-$150k 109 0.0341   
## 2 Atheist $100k-$150k 59 0.0185   
## 3 Buddhist $100k-$150k 39 0.0122   
## 4 Catholic $100k-$150k 792 0.248   
## 5 Don’t know/refused $100k-$150k 17 0.00532  
## 6 Evangelical Prot $100k-$150k 723 0.226   
## 7 Hindu $100k-$150k 48 0.0150   
## 8 Historically Black Prot $100k-$150k 81 0.0253   
## 9 Jehovah's Witness $100k-$150k 11 0.00344  
## 10 Jewish $100k-$150k 87 0.0272   
## 11 Mainline Prot $100k-$150k 753 0.236   
## 12 Mormon $100k-$150k 49 0.0153   
## 13 Muslim $100k-$150k 8 0.00250  
## 14 Orthodox $100k-$150k 42 0.0131   
## 15 Other Christian $100k-$150k 14 0.00438  
## 16 Other Faiths $100k-$150k 40 0.0125   
## 17 Other World Religions $100k-$150k 4 0.00125  
## 18 Unaffiliated $100k-$150k 321 0.100

#7) Using the two data tables below: #a) Use and show R code to produce tribbles for the tables superheroes and publishers.

superheros<- tribble(   
 ~name, ~alignment, ~gender, ~publisher,   
 "Magneto", "bad", "male", "Marvel",  
 "Storm", "good", "female", "Marvel",  
 "Mystique", "bad", "female", "Marvel",  
 "Batman", "good", "male", "DC",  
 "Joker", "bad", "male", "DC",  
 "Catwoman", "bad", "female", "DC",  
 "Hellboy", "good", "male", "Dark Horse Comics"  
)  
superheros

## # A tibble: 7 × 4  
## name alignment gender publisher   
## <chr> <chr> <chr> <chr>   
## 1 Magneto bad male Marvel   
## 2 Storm good female Marvel   
## 3 Mystique bad female Marvel   
## 4 Batman good male DC   
## 5 Joker bad male DC   
## 6 Catwoman bad female DC   
## 7 Hellboy good male Dark Horse Comics

publishers <- tribble(  
 ~publisher, ~yr\_founded,  
 "DC", "1934",  
 "Marvel", "1939",   
 "Image", "1992"  
)  
publishers

## # A tibble: 3 × 2  
## publisher yr\_founded  
## <chr> <chr>   
## 1 DC 1934   
## 2 Marvel 1939   
## 3 Image 1992

#b) Use and show R code that establishes name as a key for the data table superheroes

superheros%>%  
 count(name)%>%  
 filter(n>1)

## # A tibble: 0 × 2  
## # … with 2 variables: name <chr>, n <int>

#c) Use and show R code that produces an inner join for the tables. Does the inner join have any missing information? If so explain why.

data missing here is everything about hellboy, since he did not have publisher in common like others

superheros%>%  
 inner\_join(publishers, by = "publisher") -> inner1  
inner1

## # A tibble: 6 × 5  
## name alignment gender publisher yr\_founded  
## <chr> <chr> <chr> <chr> <chr>   
## 1 Magneto bad male Marvel 1939   
## 2 Storm good female Marvel 1939   
## 3 Mystique bad female Marvel 1939   
## 4 Batman good male DC 1934   
## 5 Joker bad male DC 1934   
## 6 Catwoman bad female DC 1934