Test2

samuel mweni

2022-11-19

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

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

## # A tibble: 18 × 11  
## religion `<$10k` $10-2…¹ $20-3…² $30-4…³ $40-5…⁴ $50-7…⁵ $75-1…⁶ $100-…⁷  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Agnostic 27 34 60 81 76 137 122 109  
## 2 Atheist 12 27 37 52 35 70 73 59  
## 3 Buddhist 27 21 30 34 33 58 62 39  
## 4 Catholic 418 617 732 670 638 1116 949 792  
## 5 Don’t know/r… 15 14 15 11 10 35 21 17  
## 6 Evangelical … 575 869 1064 982 881 1486 949 723  
## 7 Hindu 1 9 7 9 11 34 47 48  
## 8 Historically… 228 244 236 238 197 223 131 81  
## 9 Jehovah's Wi… 20 27 24 24 21 30 15 11  
## 10 Jewish 19 19 25 25 30 95 69 87  
## 11 Mainline Prot 289 495 619 655 651 1107 939 753  
## 12 Mormon 29 40 48 51 56 112 85 49  
## 13 Muslim 6 7 9 10 9 23 16 8  
## 14 Orthodox 13 17 23 32 32 47 38 42  
## 15 Other Christ… 9 7 11 13 13 14 18 14  
## 16 Other Faiths 20 33 40 46 49 63 46 40  
## 17 Other World … 5 2 3 4 2 7 3 4  
## 18 Unaffiliated 217 299 374 365 341 528 407 321  
## # … with 2 more variables: `>150k` <dbl>, `Don't know/refused` <dbl>, and  
## # abbreviated variable names ¹​`$10-20k`, ²​`$20-30k`, ³​`$30-40k`, ⁴​`$40-50k`,  
## # ⁵​`$50-75k`, ⁶​`$75-100k`, ⁷​`$100-150k`

dim(relig\_income)

## [1] 18 11

# the relig\_income dataset has 18 rows and 11 colunm  
  
#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\_income%>%  
 pivot\_longer(col = "<$10k":"Don't know/refused", names\_to = "income",   
 values\_to = "frequency" )-> p  
p

## # 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.  
p%>%  
 filter(income=="$100-150k" )-> g  
g

## # 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(g$frequency)

## [1] 3197

# the sum of the frequency Values is 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.  
#A tibble: 18 x 4  
g%>%  
 mutate(propotionoftotal=frequency/sum(g$frequency))->f  
f

## # 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.  
f%>%  
 mutate(income = recode(income, "$100-150k" = "$100k-$150k"))->t  
t

## # 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.  
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", "DarkHorse Comics")->superheroes  
superheroes

## # 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 DarkHorse Comics

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

## # A tibble: 3 × 2  
## publisher yr\_founded  
## <chr> <dbl>  
## 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  
superheroes%>%  
 count(name)%>%  
 filter(n>1)

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

# name is a key for the superheroes data table for the output shows  
# tibble: 0 X 2 ( The 0 tells us that you do not have more than  
# 1 observation for each name designation,) All name   
# entries are different  
  
#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.  
  
superheroes%>%  
 inner\_join(publishers, by ="publisher")

## # A tibble: 6 × 5  
## name alignment gender publisher yr\_founded  
## <chr> <chr> <chr> <chr> <dbl>  
## 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

#The last row of Superheroes dataset is missing  
# this is because An inner join matches pairs of observations from the tables whenever their keys are equal  
  
#8)Now modify the table that you produced in problem 6, so that it only reflects the religions   
#Jewish, Muslim, Catholic, and Mainline Prot (Mainline Protestant). Use and show all required   
#R code that will produce the table below;  
t%>%  
 filter(religion %in% c("Jewish", "Muslim", "Catholic","Mainline Prot"))->s  
s

## # A tibble: 4 × 4  
## religion income frequency propotionoftotal  
## <chr> <chr> <dbl> <dbl>  
## 1 Catholic $100k-$150k 792 0.248   
## 2 Jewish $100k-$150k 87 0.0272   
## 3 Mainline Prot $100k-$150k 753 0.236   
## 4 Muslim $100k-$150k 8 0.00250

#And finally, use and show R code that will produce a pie chart that is partitioned according to   
#the frequency count for each of the four religions shown in the table.  
  
library(ggplot2)  
bp<- ggplot(s, aes(x="", y=frequency, fill=religion))+  
geom\_bar(width = 1, stat = "identity")  
  
 pie <- bp + coord\_polar("y", start=0)  
pie

