Data science basics

Recoding data (Case: religion and income)

Prof. Dr. Jan Kirenz

The following content is based on Mine Çetinkaya-Rundel's excellent book Data Science in a Box

Case study: Religion and income

Income distribution by religious group

% of adults who have a household income of...

Chart | Table | Share | Save Image

Religious tradition	Less than $30,000	$30,000-$49,999	$50,000-$99,999	$100,000 or more	Sample Size
Buddhist	36%	18%	32%	13%	233
Catholic	36%	19%	26%	19%	6,137
Evangelical Protestant	35%	22%	28%	14%	7,462
Hindu	17%	13%	34%	36%	172
Historically Black Protestant	53%	22%	17%	8%	1,704
Jehovah's Witness	48%	25%	22%	4%	208
Jewish	16%	15%	24%	44%	708
Mainline Protestant	29%	20%	28%	23%	5,208
Mormon	27%	20%	33%	20%	594
Muslim	34%	17%	29%	20%	205
Orthodox Christian	18%	17%	36%	29%	155
Unaffiliated (religious "nones")	33%	20%	26%	21%	6,790

Sample sizes and margins of error vary from subgroup to subgroup, from year to year and from state to state. You can see the sample size for the estimates in this chart on rollover or in the last column of the table. And visit this table to see approximate margins of error for a group of a given size. Readers should always bear in mind the approximate margin of error for the group they are examining when making comparisons with other groups or assessing the significance of trends over time. For full question wording, see the survey questionnaire.

Source: pewforum.org/religious-landscape-study/income-distribution, Retrieved 14 April, 2020

Read data

```
library(readxl)
rel_inc <- read_excel("data/relig-income.xlsx")
```

```
## # A tibble: 12 x 6
##   `Religious trad… `Less than $30,… `$30,000-$49,99…
##   <chr>                       <dbl>            <dbl>
## 1 Buddhist                     0.36             0.18
## 2 Catholic                     0.36             0.19
## 3 Evangelical Pro…             0.35             0.22
## 4 Hindu                        0.17             0.13
## 5 Historically Bl…             0.53             0.22
## 6 Jehovah's Witne…             0.48             0.25
## # … with 6 more rows, and 3 more variables:
## #   `$50,000-$99,999` <dbl>, `$100,000 or more` <dbl>, `Sample
## #   Size` <dbl>
```

Rename columns

```
rel_inc %>%
  rename(
    religion = `Religious tradition`,
    n = `Sample Size`
  )
```

```
## # A tibble: 12 x 6
##   religion `Less than $30,… `$30,000-$49,99… `$50,000-$99,99…
##   <chr>                <dbl>            <dbl>             <dbl>
## 1 Buddhist              0.36             0.18              0.32
## 2 Catholic              0.36             0.19              0.26
## 3 Evangel…              0.35             0.22              0.28
## 4 Hindu                 0.17             0.13              0.34
## 5 Histori…              0.53             0.22              0.17
## 6 Jehovah…              0.48             0.25              0.22
## # … with 6 more rows, and 2 more variables: `$100,000 or
## #   more` <dbl>, n <dbl>
```

If we want a new variable called `income` with levels such as "Less than $30,000", "$30,000-$49,999", ... etc. which function should we use?

```
## # A tibble: 48 x 4
##    religion                     n income            proportion
##    <chr>                    <dbl> <chr>                  <dbl>
##  1 Buddhist                   233 Less than $30,000       0.36
##  2 Buddhist                   233 $30,000-$49,999         0.18
##  3 Buddhist                   233 $50,000-$99,999         0.32
##  4 Buddhist                   233 $100,000 or more        0.13
##  5 Catholic                  6137 Less than $30,000       0.36
##  6 Catholic                  6137 $30,000-$49,999         0.19
##  7 Catholic                  6137 $50,000-$99,999         0.26
##  8 Catholic                  6137 $100,000 or more        0.19
##  9 Evangelical Protestant    7462 Less than $30,000       0.35
## 10 Evangelical Protestant    7462 $30,000-$49,999         0.22
## 11 Evangelical Protestant    7462 $50,000-$99,999         0.28
## 12 Evangelical Protestant    7462 $100,000 or more        0.14
## 13 Hindu                      172 Less than $30,000       0.17
## 14 Hindu                      172 $30,000-$49,999         0.13
## 15 Hindu                      172 $50,000-$99,999         0.34
## # … with 33 more rows
```

Pivot longer

```
rel_inc %>%
  rename(
    religion = `Religious tradition`,
    n = `Sample Size`
  ) %>%
  pivot_longer(
    cols = -c(religion, n),   # all but religion and n
    names_to = "income",
    values_to = "proportion"
  )
```

```
## # A tibble: 48 x 4
##   religion     n income            proportion
##   <chr>    <dbl> <chr>                  <dbl>
## 1 Buddhist   233 Less than $30,000       0.36
## 2 Buddhist   233 $30,000-$49,999         0.18
## 3 Buddhist   233 $50,000-$99,999         0.32
## 4 Buddhist   233 $100,000 or more        0.13
## 5 Catholic  6137 Less than $30,000       0.36
## 6 Catholic  6137 $30,000-$49,999         0.19
## # … with 42 more rows
```

Calculate frequencies

```
rel_inc %>%
  rename(
    religion = `Religious tradition`,
    n = `Sample Size`
  ) %>%
  pivot_longer(
    cols = -c(religion, n),
    names_to = "income",
    values_to = "proportion"
  ) %>%
  mutate(frequency = round(proportion * n))
```

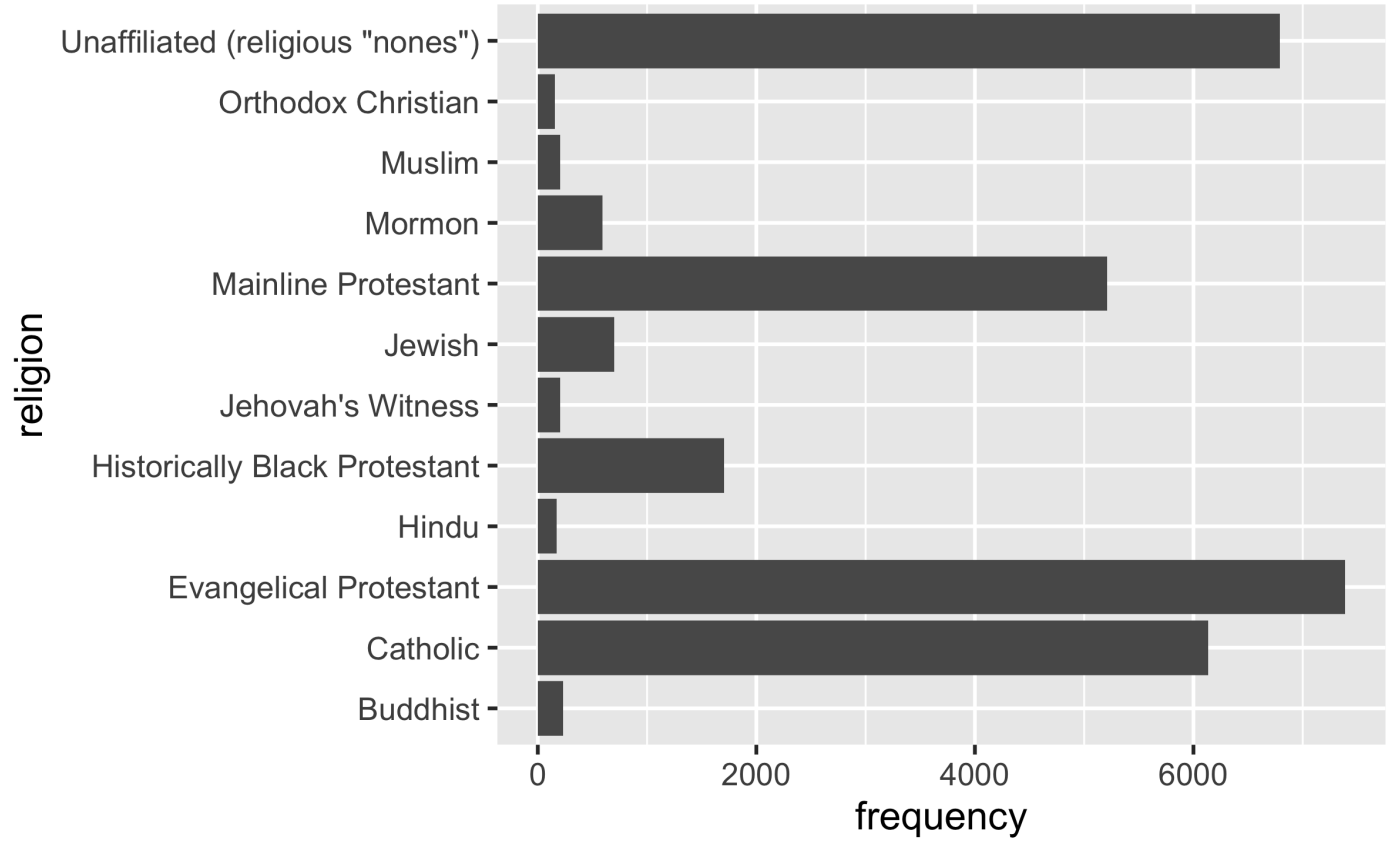
```
## # A tibble: 48 x 5
##   religion     n income            proportion frequency
##   <chr>    <dbl> <chr>                  <dbl>     <dbl>
## 1 Buddhist   233 Less than $30,000       0.36        84
## 2 Buddhist   233 $30,000-$49,999         0.18        42
## 3 Buddhist   233 $50,000-$99,999         0.32        75
## 4 Buddhist   233 $100,000 or more        0.13        30
## 5 Catholic  6137 Less than $30,000       0.36      2209
## 6 Catholic  6137 $30,000-$49,999         0.19      1166
## # … with 42 more rows
```

Save data

```
rel_inc_long <- rel_inc %>%
  rename(
    religion = `Religious tradition`,
    n = `Sample Size`
  ) %>%
  pivot_longer(
    cols = -c(religion, n),
    names_to = "income",
    values_to = "proportion"
  ) %>%
  mutate(frequency = round(proportion * n))
```

Barplot

```
ggplot(rel_inc_long, aes(y = religion, x = frequency)) +
  geom_col()
```



Recode religion

Recode Plot

```
rel_inc_long <- rel_inc_long %>%
  mutate(religion = case_when(
    religion == "Evangelical Protestant"           ~ "Ev. Protestant",
    religion == "Historically Black Protestant"    ~ "Hist. Black Protestant",
    religion == 'Unaffiliated (religious "nones")' ~ "Unaffiliated",
    TRUE                                           ~ religion
  ))
```

Reverse religion order

Recode | Plot

```
rel_inc_long <- rel_inc_long %>%
  mutate(religion = fct_rev(religion))
```

Add income

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col()
```

Fill bars

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill")
```

Change colors

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d()
```

Change theme

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal()
```

Move legend to the bottom

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal() +
  theme(legend.position = "bottom")
```

Legend adjustments

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal() +
  theme(legend.position = "bottom") +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE))
```

Fix labels

Plot Code

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal() +
  theme(legend.position = "bottom") +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE)) +
  labs(
    x = "Proportion", y = "",
    title = "Income distribution by religious group",
    subtitle = "Source: Pew Research Center, Religious Landscape Study",
    fill = "Income"
    )
```