Data tidying and visualizing

DATA 202 21FA

This stuff isn't easy.

Q&A

> Why isn't all data already long-format?

Usually: data meant for *human* consumption.

> What are other uses of pivoting? `pivot_wider` and `pivot_longer` together??

See `vignette("pivot", package = "tidyr")`.

> Performance?

Usually a `join` is the tighter bottleneck.

Reminders

- First project milestone due Friday
- Quiz closes today
- Test *next* Friday (in lab)

Micro-Pivot

```
x
```

```
# A tibble: 1 × 3
  id    col1  col2
  <chr> <chr> <chr>
1 A     A1    A2
```

Micro-Pivot

```
x
```

```
# A tibble: 1 × 3
  id    col1  col2
  <chr> <chr> <chr>
1 A     A1    A2
```

```
x %>%
  pivot_longer(
    cols = starts_with("col"),
    names_to = "column_name", values_to = "value_was")
```

```
# A tibble: 2 × 3
  id    column_name value_was
  <chr> <chr>       <chr>
1 A     col1        A1
2 A     col2        A2
```

Micro-Pivot

```
x
```

```
# A tibble: 2 × 3
  id    col1  col2
  <chr> <chr> <chr>
1 A     A1    A2
2 B     B1    B2
```

Micro-Pivot

```
x
```

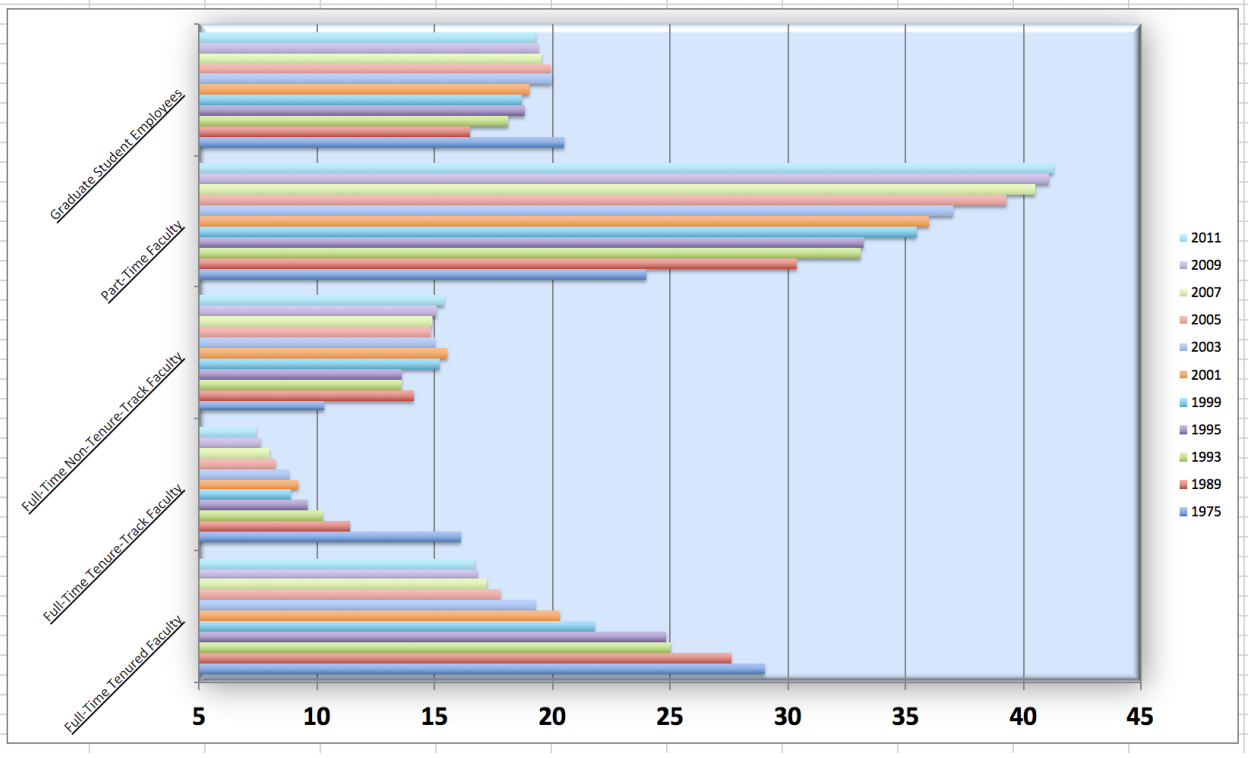
```
# A tibble: 2 × 3
  id    col1  col2
  <chr> <chr> <chr>
1 A     A1    A2
2 B     B1    B2
```

```
x %>%
  pivot_longer(cols = starts_with("col"), names_to = "column_nam
```

```
# A tibble: 4 × 3
  id    column_name value_was
  <chr> <chr>       <chr>
1 A     col1        A1
2 A     col2        A2
3 B     col1        B1
4 B     col2        B2
```

Instructional staff employment trends

- American Association of University Professors (AAUP)
- nonprofit association of academic professionals
- 2013 report: trends in instructional staff employees between 1975 and 2011
- contains an image very similar to this:



Data

Each row in this dataset represents a faculty type, and the columns are the years for which we have data. The values are percentage of hires of that type of faculty for each year.

```
staff <- read_csv("data/instructional-staff.csv")
staff
```

```
# A tibble: 5 × 12
  faculty_type   `1975` `1989` `1993` `1995` `1999` `2001` `2003`
  <chr>           <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Full-Time Ten…   29     27.6   25     24.8   21.8   20.3   19.3
2 Full-Time Ten…   16.1   11.4   10.2    9.6    8.9    9.2    8.8
3 Full-Time Non…   10.3   14.1   13.6   13.6   15.2   15.5   15
4 Part-Time Fac…   24     30.4   33.1   33.2   35.5   36     37
5 Graduate Stud…   20.5   16.5   18.1   18.8   18.7   19     20
# … with 4 more variables: 2005 <dbl>, 2007 <dbl>, 2009 <dbl>,
#   2011 <dbl>
```

Recreate the visualization

To recreate this visualization we need to first reshape the data to have one variable for faculty type and one variable for year. In other words, we will convert the data from *wide format* to *long format*.

But before we do so...

> If the long data will have a row for each year/faculty type combination, and there are 5 faculty types and 11 years of data, how many rows will the data have?


PIVOT!!!

Write the pivot.

```
staff
```

```
# A tibble: 5 × 12
  faculty_type   `1975` `1989` `1993` `1995` `1999` `2001` `2003`
  <chr>           <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Full-Time Ten…   29     27.6   25     24.8   21.8   20.3   19.3
2 Full-Time Ten…   16.1   11.4   10.2    9.6    8.9    9.2    8.8
3 Full-Time Non…   10.3   14.1   13.6   13.6   15.2   15.5   15
4 Part-Time Fac…   24     30.4   33.1   33.2   35.5   36     37
5 Graduate Stud…   20.5   16.5   18.1   18.8   18.7   19     20
# … with 4 more variables: 2005 <dbl>, 2007 <dbl>, 2009 <dbl>,
#   2011 <dbl>
```

Pivot staff data

```
staff %>%
  pivot_longer(
    cols = -faculty_type,
    names_to = "year",
    values_to = "percentage"
    )
```

```
# A tibble: 55 × 3
  faculty_type                year  percentage
  <chr>                       <chr>      <dbl>
1 Full-Time Tenured Faculty 1975        29
2 Full-Time Tenured Faculty 1989        27.6
3 Full-Time Tenured Faculty 1993        25
4 Full-Time Tenured Faculty 1995        24.8
5 Full-Time Tenured Faculty 1999        21.8
6 Full-Time Tenured Faculty 2001        20.3
# … with 49 more rows
```

Pivot staff data, and save result

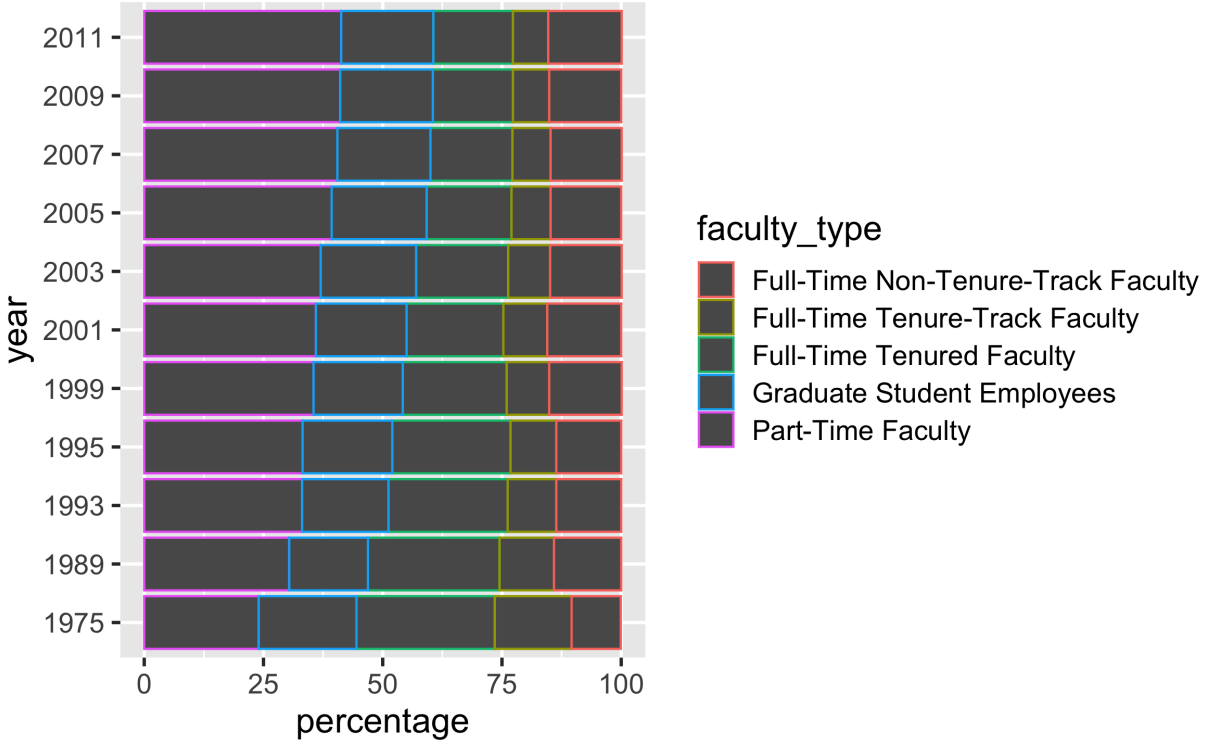
```
staff_long <- staff %>%
  pivot_longer(
    cols = -faculty_type,
    names_to = "year",
    values_to = "percentage"
    )

staff_long
```

```
# A tibble: 55 × 3
  faculty_type                year  percentage
  <chr>                       <chr>      <dbl>
1 Full-Time Tenured Faculty 1975        29
2 Full-Time Tenured Faculty 1989        27.6
3 Full-Time Tenured Faculty 1993        25
4 Full-Time Tenured Faculty 1995        24.8
5 Full-Time Tenured Faculty 1999        21.8
6 Full-Time Tenured Faculty 2001        20.3
# … with 49 more rows
```

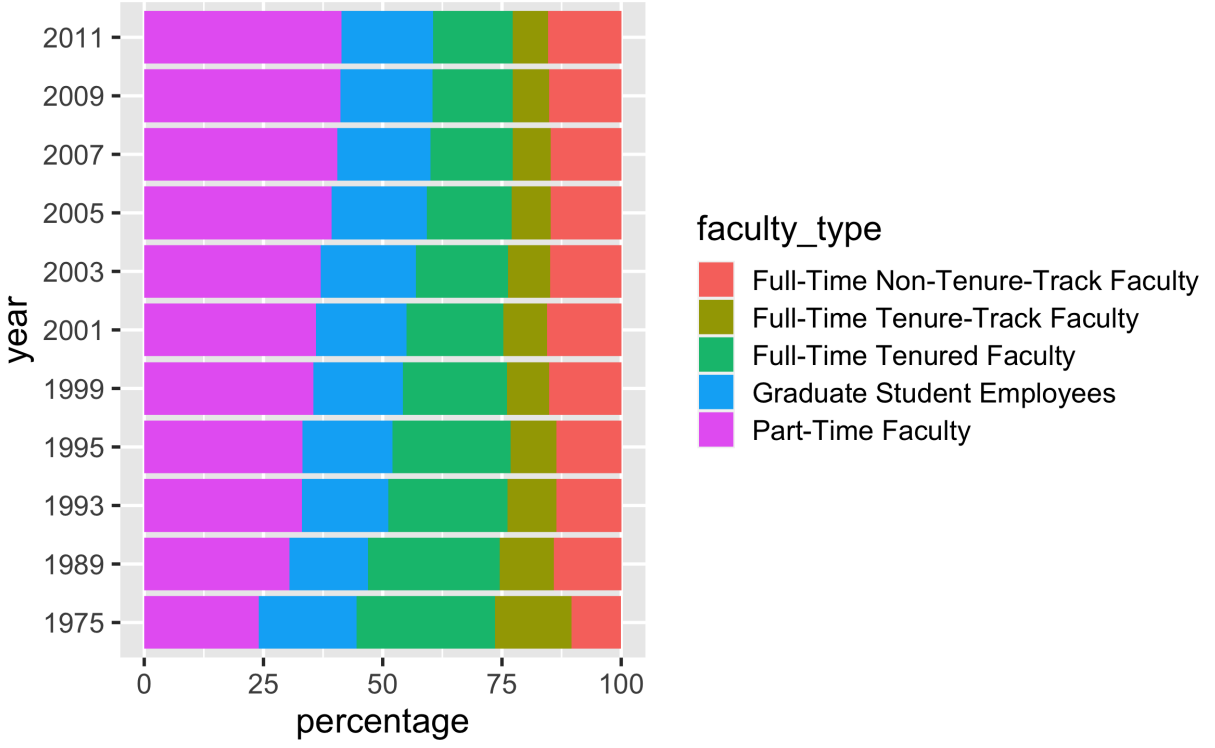
This doesn't look quite right, how would you fix it?

```
ggplot(staff_long, aes(x = percentage, y = year, color = faculty
  geom_col()
```



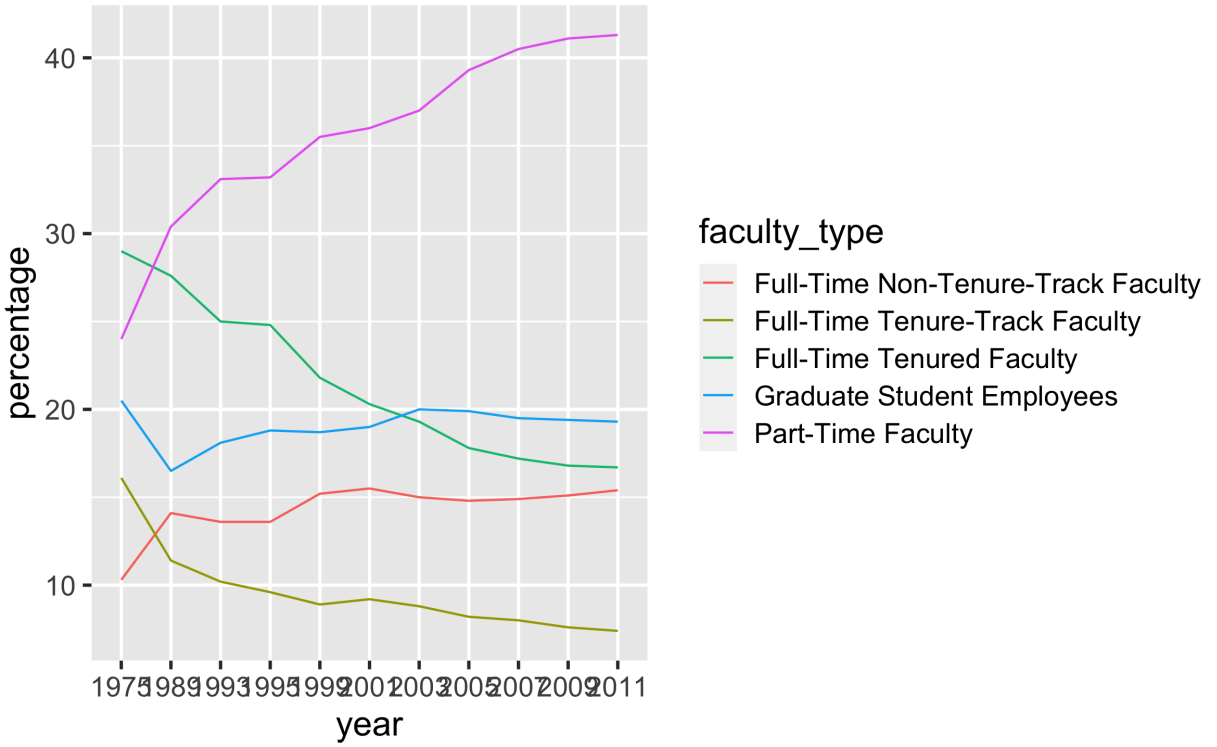
Some improvement...

```
staff_long %>%
  ggplot(aes(x = percentage, y = year, fill = faculty_type)) +
  geom_col()
```

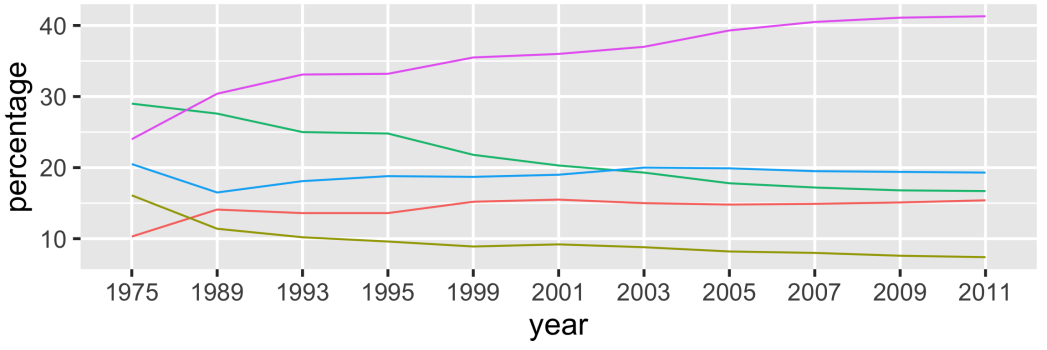


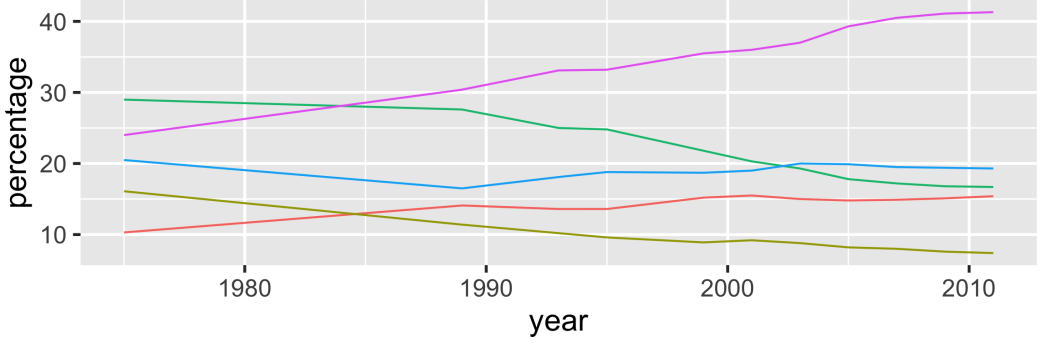
More improvement

```
staff_long %>%
  ggplot(aes(x = year, y = percentage, group = faculty_type, col
  geom_line()
```



What is the difference between these two plots?

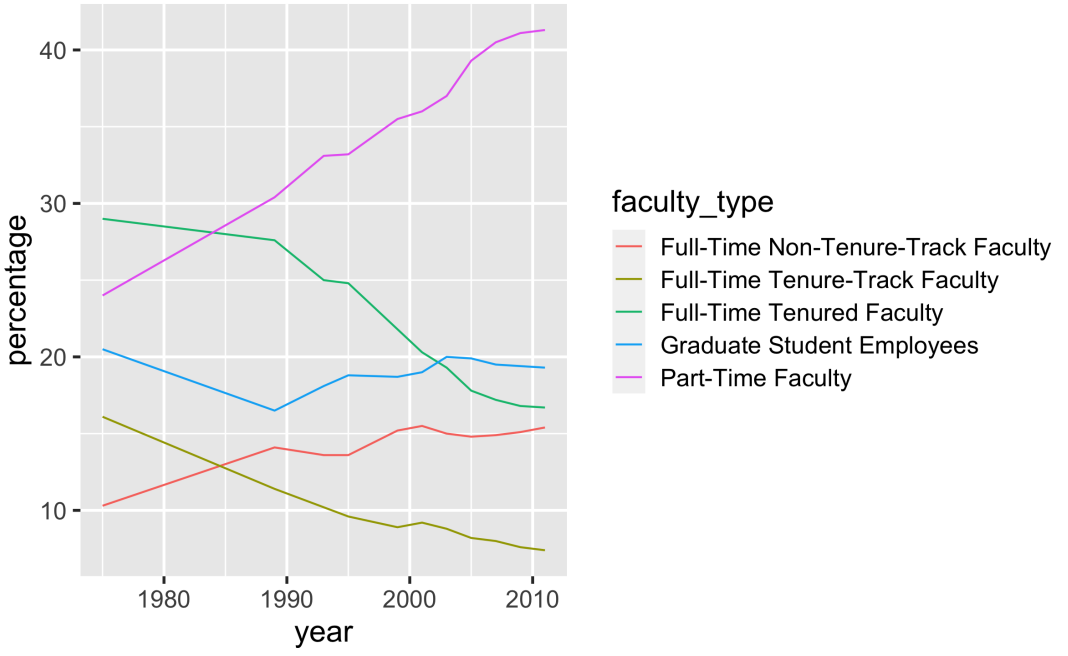




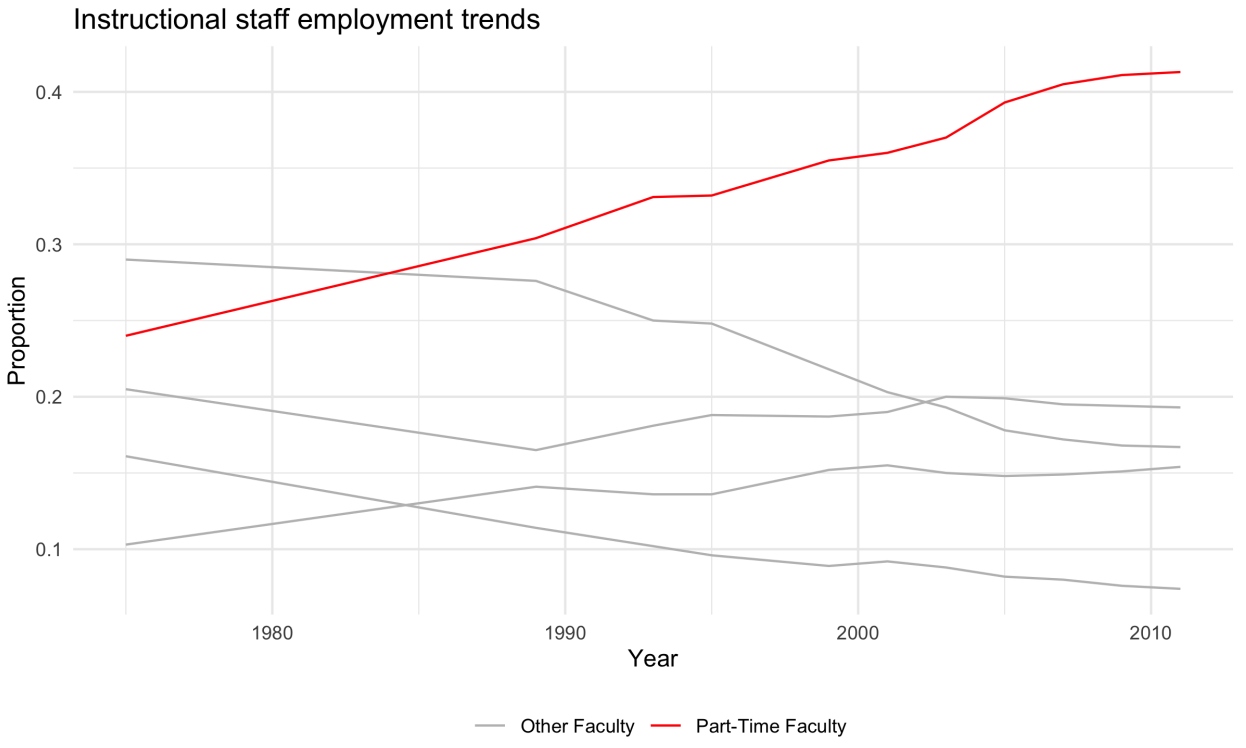
Make year numeric again!

```
staff_long <- staff_long %>%
  mutate(year = as.numeric(year))

staff_long %>%
  ggplot(aes(x = year, y = percentage, group = faculty_type, color = faculty_type)) +
  geom_line()
```

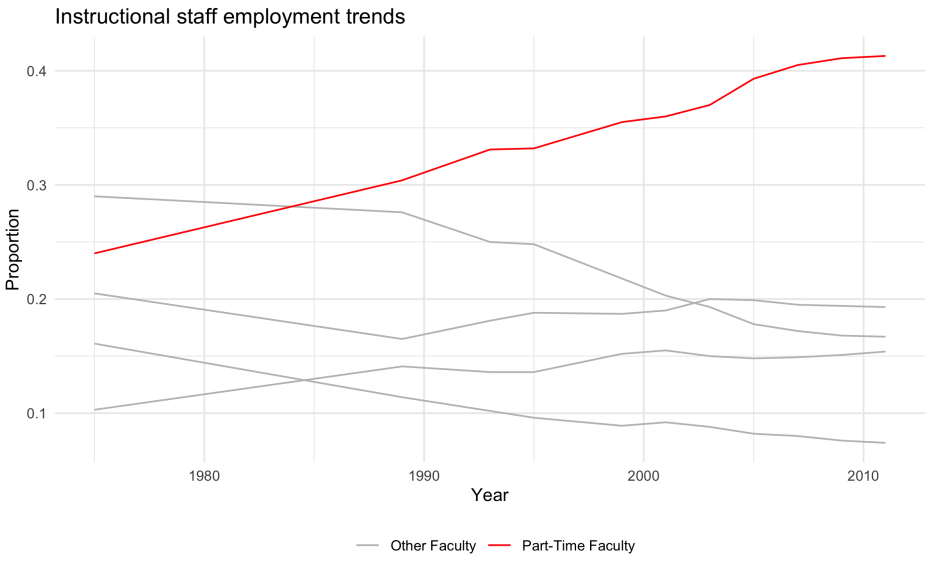


How would you go about creating the following plot?

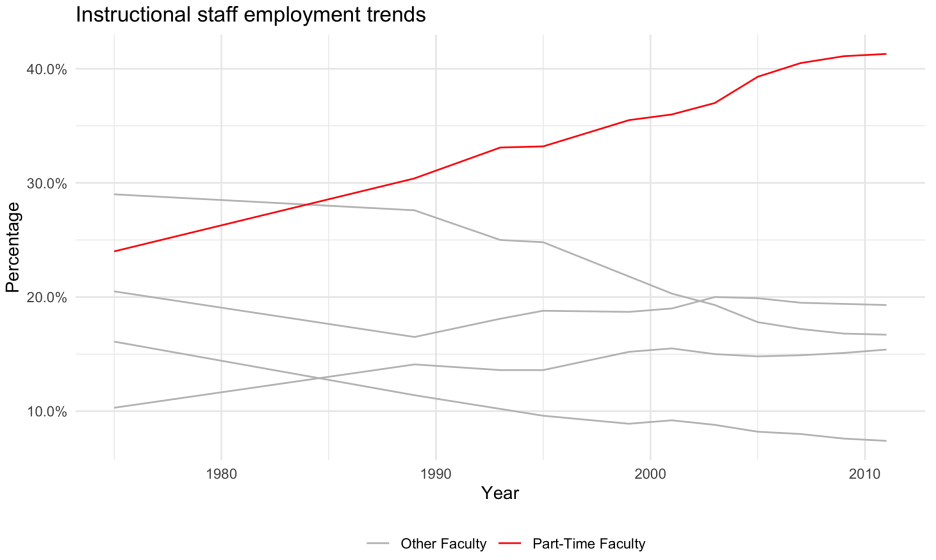


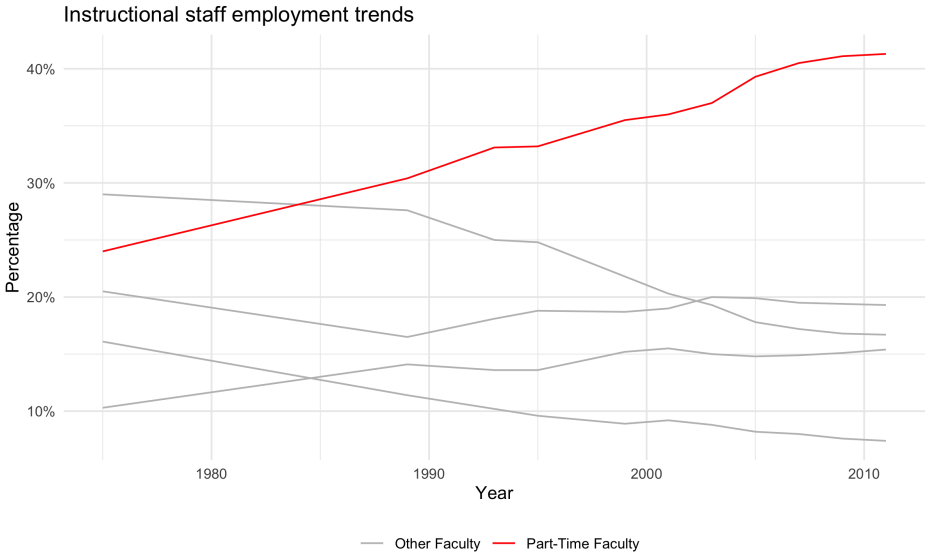
```
staff_long %>%
  mutate(part_time = if_else(faculty_type == "Part-Time Faculty"
                             "Part-Time Faculty",
                             "Other Faculty")) %>%
  ggplot(aes(x = year, y = percentage/100,
             group = faculty_type,
             color = part_time)) +
  geom_line() +
  scale_color_manual(values = c("gray", "red")) +
  theme_minimal() +
  labs(
    title = "Instructional staff employment trends",
    x = "Year",
    y = "Proportion",
    color = ""
  ) +
  theme(legend.position = "bottom")
```

```
staff_long %>%
  mutate(part_time = if_else(faculty_type == "Part-Time Faculty"
                             "Part-Time Faculty",
                             "Other Faculty")) %>%
  ggplot(aes(x = year, y = percentage/100,
             group = faculty_type,
             color = part_time)) +
  geom_line() +
  scale_color_manual(values = c("gray", "red")) +
  theme_minimal() +
  labs(
    title = "Instructional staff employment trends",
    x = "Year",
    y = "Proportion",
    color = ""
  ) +
  theme(legend.position = "bottom")
```


Instructional staff employment trends
0.4
0.3
0.2
0.1
Proportion
1980
1990
2000
2010
Year
Other Faculty
Part-Time Faculty

```r
library(scales)
staff_long %>%
  mutate(part_time =
           if_else(faculty_type == "Part-Time Faculty",
                   "Part-Time Faculty", "Other Faculty")) %>%
  ggplot(aes(x = year, y = percentage/100, group = faculty_type,
             color = part_time)) +
  geom_line() +
  scale_color_manual(values = c("gray", "red")) +
  scale_y_continuous(labels = percent) +
  theme_minimal() +
  labs(
    title = "Instructional staff employment trends",
    x = "Year",
    y = "Percentage",
    color = ""
  ) +
  theme(legend.position = "bottom")
```


Instructional staff employment trends
40.0%
30.0%
20.0%
10.0%
Percentage
1980
1990
2000
2010
Year
Other Faculty
Part-Time Faculty


Instructional staff employment trends
40%
30%
20%
10%
Percentage
1980
1990
2000
2010
Year
Other Faculty
Part-Time Faculty

```r
library(scales)
staff_long %>%
  mutate(part_time =
           if_else(faculty_type == "Part-Time Faculty",
                   "Part-Time Faculty", "Other Faculty")) %>%
  ggplot(aes(x = year, y = percentage/100, group = faculty_type,
             color = part_time)) +
  geom_line() +
  scale_color_manual(values = c("gray", "red")) +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  theme_minimal() +
  labs(
    title = "Instructional staff employment trends",
    x = "Year",
    y = "Percentage",
    color = ""
  ) +
  theme(legend.position = "bottom")
```

Other common tidying moves

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") # directly from
```

```
# A tibble: 12 × 6
  `Religious tradition`         `Less than $30,… `$30,000-$49,99…
  <chr>                                    <dbl>            <dbl>
1 Buddhist                                  0.36             0.18
2 Catholic                                  0.36             0.19
3 Evangelical Protestant                    0.35             0.22
4 Hindu                                     0.17             0.13
5 Historically Black Protestant             0.53             0.22
6 Jehovah's Witness                         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 × 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 Evangelical…             0.35             0.22              0.28
4 Hindu                    0.17             0.13              0.34
5 Historicall…             0.53             0.22              0.17
6 Jehovah's W…             0.48             0.25              0.22
# … with 6 more rows, and 2 more variables:
#   $100,000 or more <dbl>, n <dbl>
```

Rename columns

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

```
# A tibble: 12 × 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 Evangelical…             0.35             0.22              0.28
4 Hindu                    0.17             0.13              0.34
5 Historicall…             0.53             0.22              0.17
6 Jehovah's W…             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

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 × 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))
```

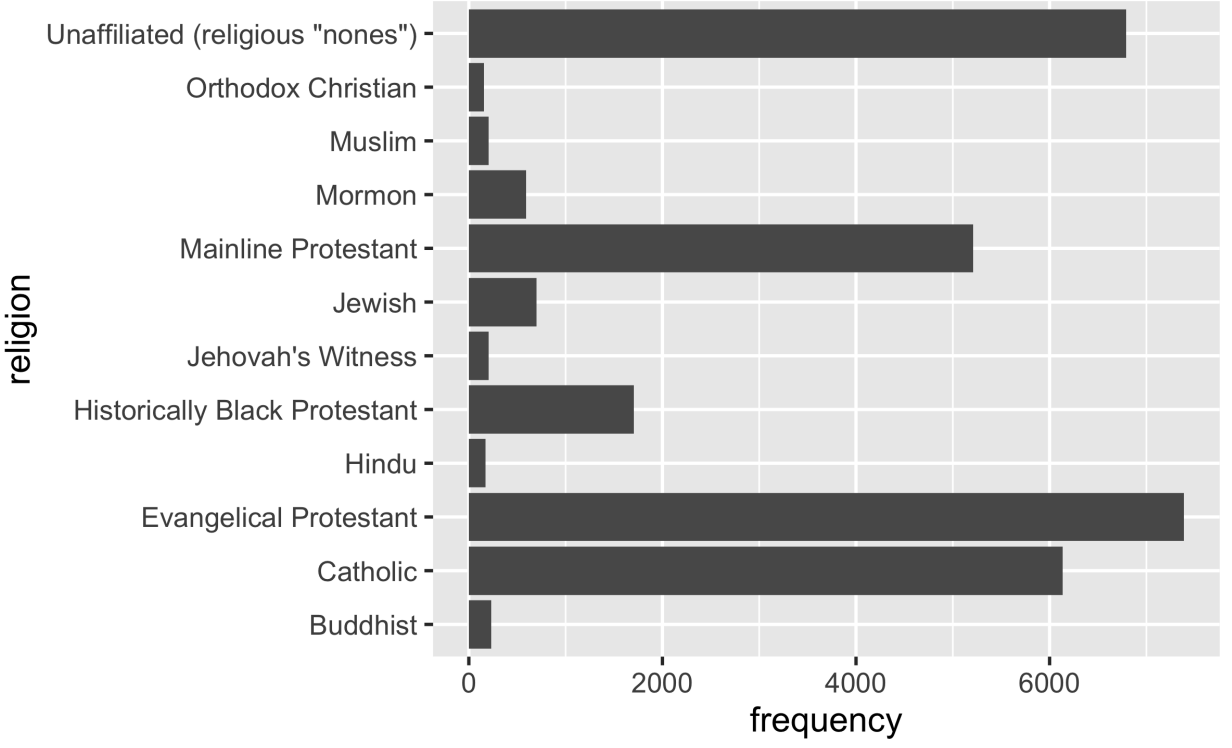
```
# A tibble: 48 × 5
  religion     n income              proportion frequency
  <chr>    <dbl> <chr>                    <dbl>     <dbl>
1 Buddhist   233 Less than $30,000         0.36        84
2 Buddhist   233 $30,000-$49,999           0.18        42
3 Buddhist   233 $50,000-$99,999           0.32        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))
```

Religion

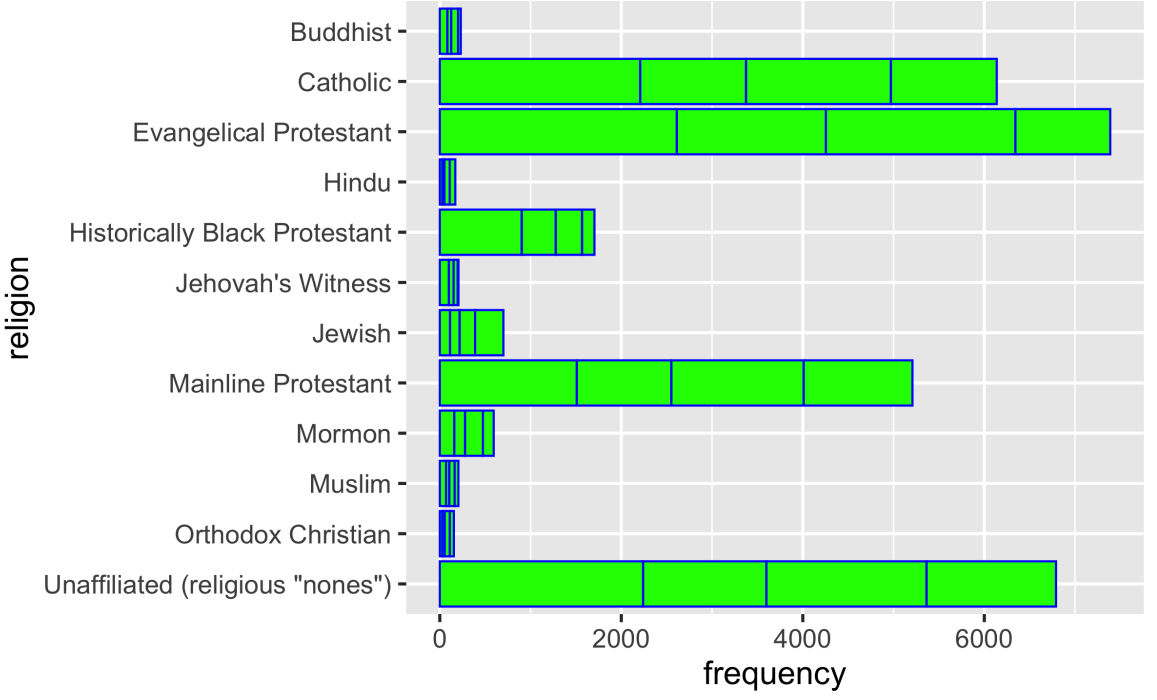
```
ggplot(rel_inc_long, aes(y = religion, x = frequency)) +
  geom_col()
```



Reverse religion order

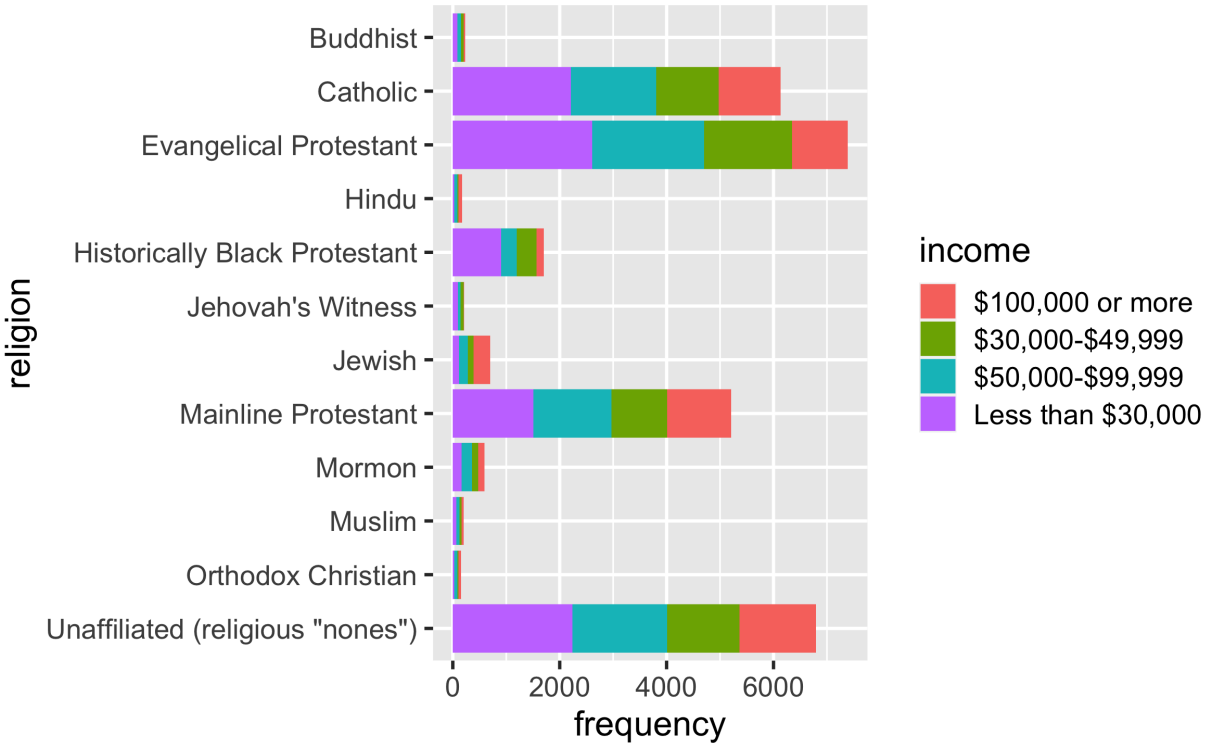
```
rel_inc_long <- rel_inc_long %>%
  mutate(religion = fct_rev(religion))

ggplot(rel_inc_long, aes(y = religion, x = frequency)) +
  geom_col(color = "blue", fill = "green")
```



Add income

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

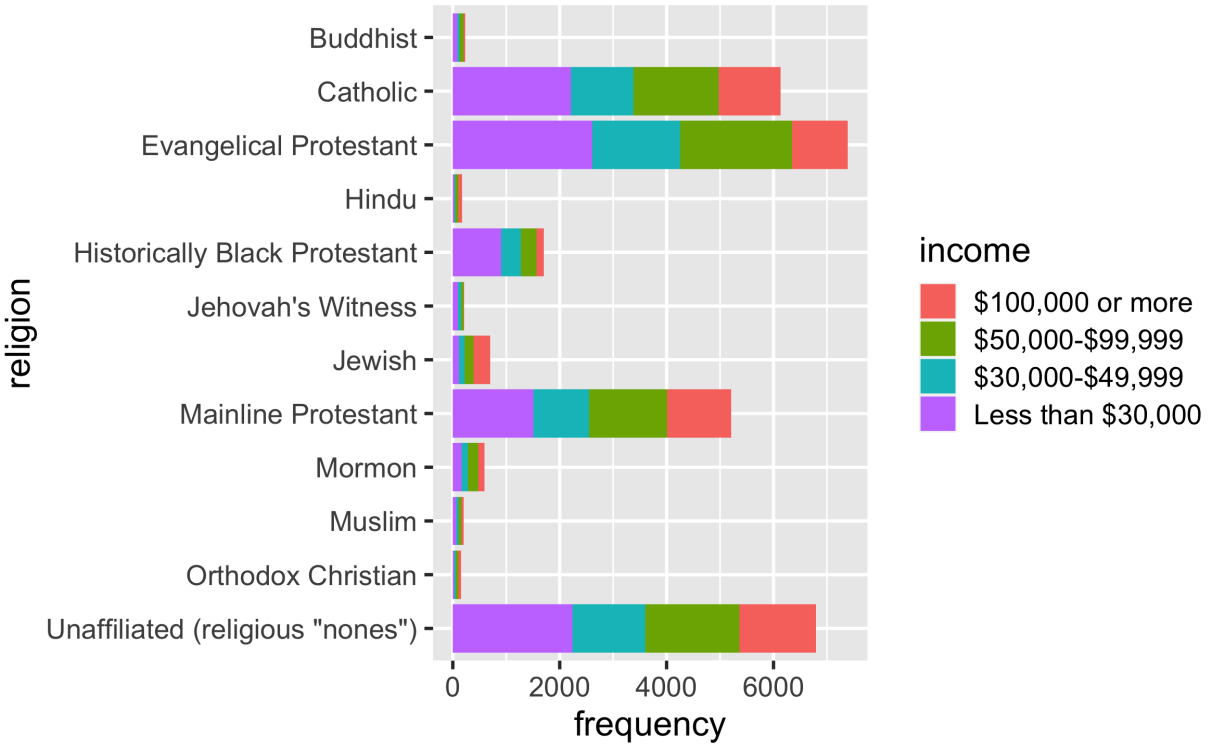


Fix income level ordering

```
rel_inc_long <- rel_inc_long %>%
  mutate(
    income = fct_relevel(income, "$100,000 or more",
                         "$50,000-$99,999", "$30,000-$49,999",
                         "Less than $30,000")
    )
```

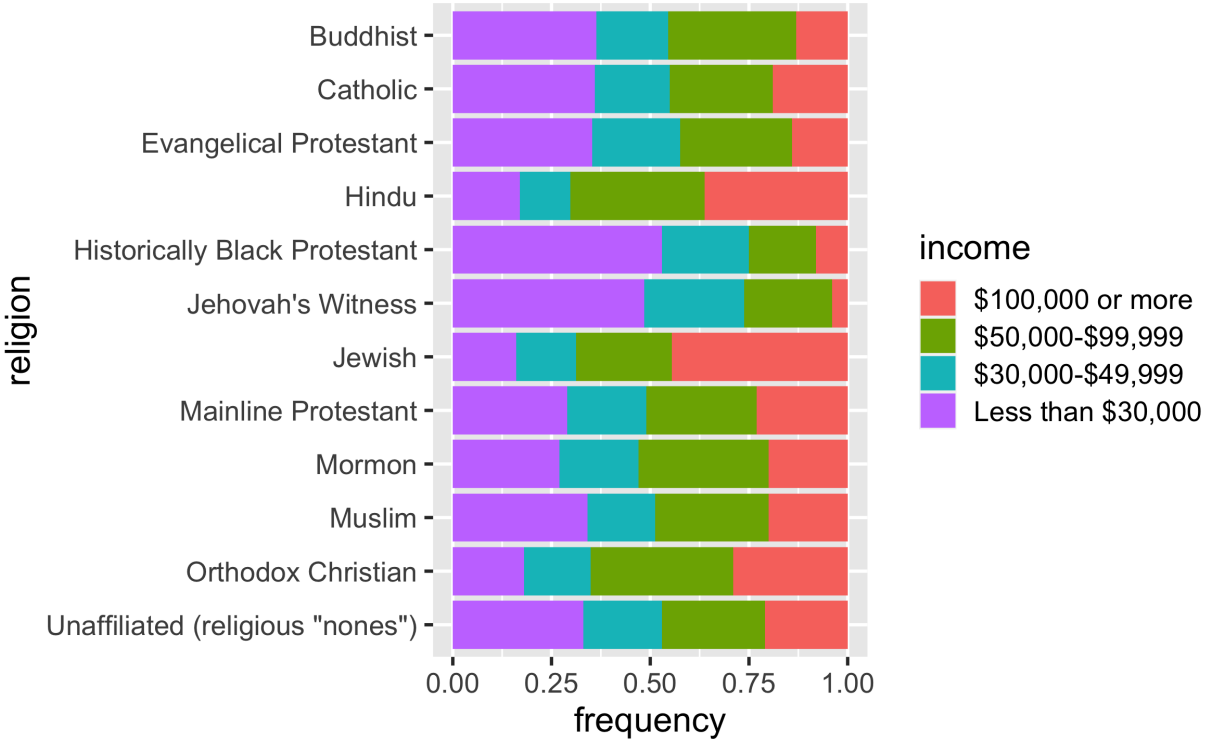
Plot again

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



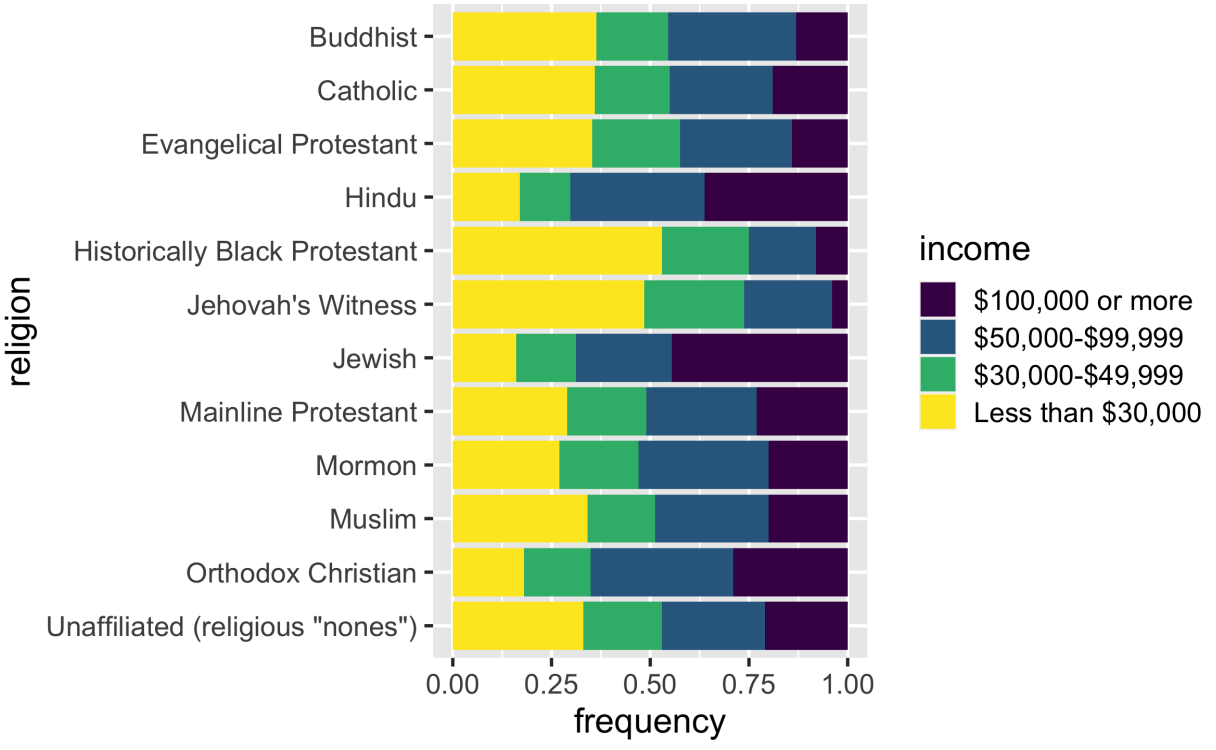
Fill bars

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



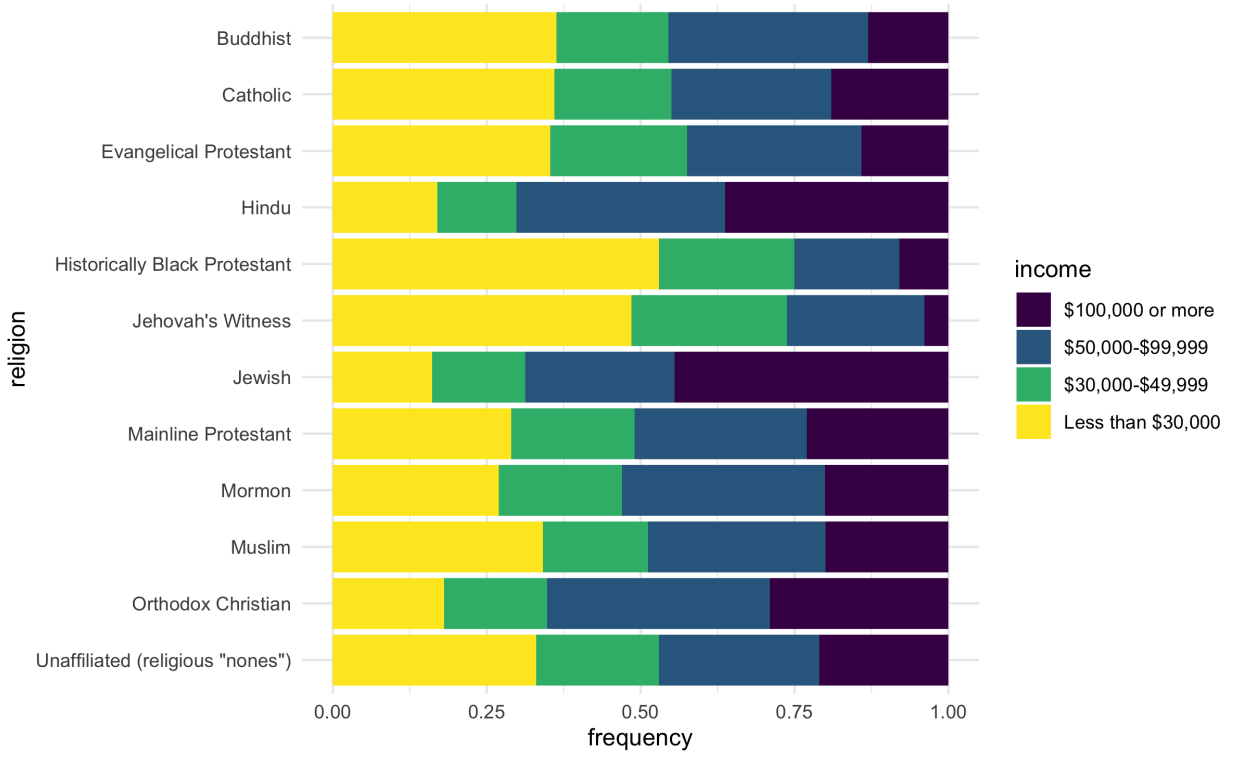
Change colors

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



Change theme

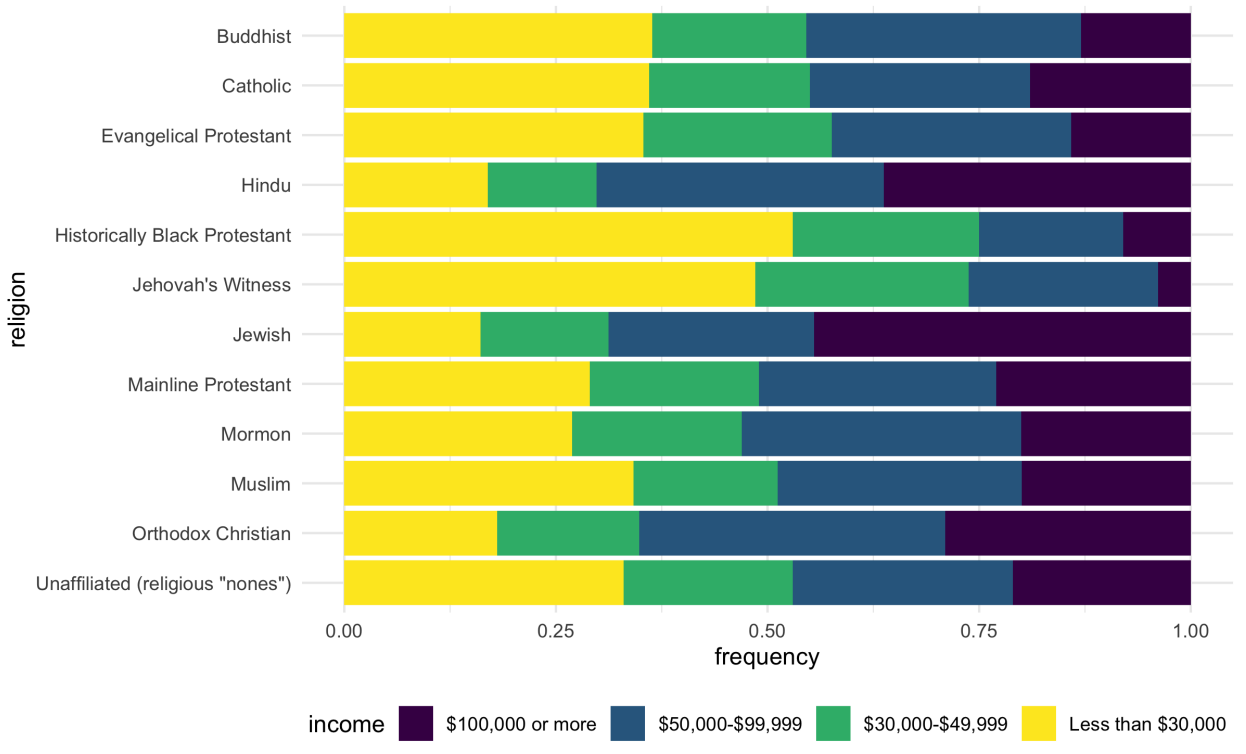
```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = income)) +
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal()
```



Move legend to the bottom

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

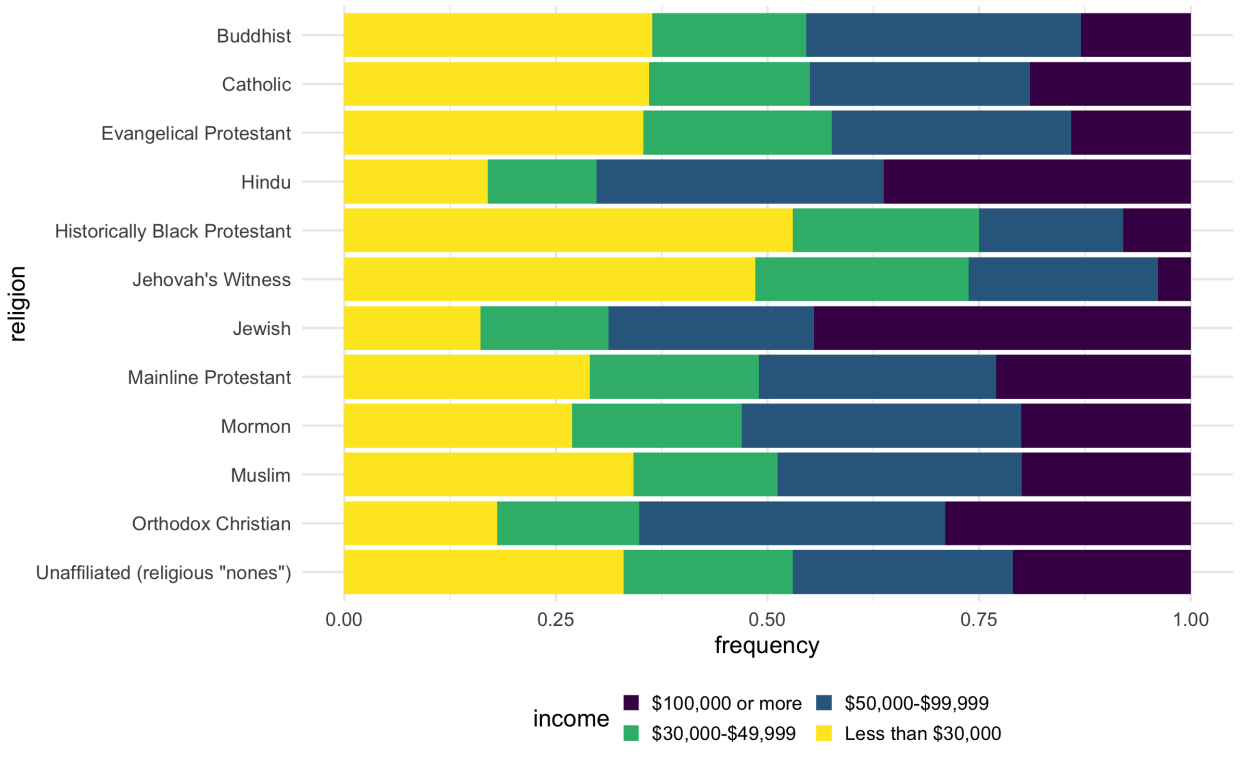
Move legend to the bottom (plot)



Legend adjustments

```
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.key.size = unit(0.3, "cm"),
    legend.box.margin = margin(t = 0, r = 0, b = 0, l = 0, unit = "pt")
    ) +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE))
```

Legend adjustments (plot)



Fix labels

```
ggplot(rel_inc_long, aes(y = religion, x = frequency, fill = inc
  geom_col(position = "fill") +
  scale_fill_viridis_d() +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.key.size = unit(0.3, "cm"),
    legend.box.margin = margin(t = 0, r = 0, b = 0, l = 0, unit
    ) +
  guides(fill = guide_legend(nrow = 2, byrow = TRUE)) +
  labs(
    x = "Frequency", y = "",
    title = "Income distribution by religious group",
    subtitle = "Source: Pew Research Center, Religious Landscape
    fill = "Income"
    )
```

Fix labels (plot)