Problem Set 2

Data Visualisation for Social Scientists

Matilda Tomatis

Due: February 4, 2026

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Wednesday February 4, 2026. No late assignments will be accepted.

Study of Religious Congregations in Switzerland

The data for this problem set come from the National Congregations Study Switzerland (NCSS), which was conducted in 2008–2009 and 2022–2023. The data provide information on organisational structure, staffing, finances, worship practices, youth and educational activities, social composition, external engagement, and inclusion norms. The data were collected using stratified random samples of congregations drawn from comprehensive censuses, with interviews completed by a single knowledgeable key informant in each congregation, most often the spiritual leader.

Data Manipulation

1. Load the NCSS .csv file from GitHub into your global environment. Use the select() function to keep these variables in your dataframe:

   - Congregation ID (`CASEID`)
   - Year (`YEAR`)
   - Region (`GDREGION`)
   - Number of official members (`NUMOFFMBR`)
   - 6-level religious classification (`TRAD6`)
   - 12-level religious classification (`TRAD12`)
   - Total income in last fiscal year (`INCOME`)

2. Filter the dataset so that you only include Christian, Jewish, and Muslim congregations (Chrétiennes, Juives, Musulmanes) using the `TRAD6` variable.

3. Compute for the number of congregations by religious classification (`TRAD6`) in each year, as well as the mean and median total income in last fiscal year (`INCOME`) by religious classification and year.

4. Create a categorical variable for called `AVG_INCOME` that is binary in which $1 =$ "Above average or average income" and $0 =$ "Below average income", which indicates if a congregation is $\geq$ average income or $<$ average income among congregations that year.

Answer 1

The dataset is loaded in the global environment and the desired variables are subselected.

```r
NCSS_raw <- read.csv("NCSS_v1.csv")
NCSS <- NCSS_raw |>
  select(CASEID, YEAR, GDREGION, NUMOFFMBR, TRAD6,
         TRAD12, INCOME)
```

Answer 2

The `NCSS` dataset only the observations of congregations of Christians, Muslims and Jews are kept.

```r
NCSS <- NCSS |>
  filter(TRAD6 %in% c("Chrétiennes", "Juives", "Musulmanes"))
```

Answer 3

The data is grouped by combinations of year and religious affiliation, afterwards the count of the observations, the mean and the median are computed.

```r
NCSS_count <- NCSS |>
  group_by(YEAR, TRAD6) |>
  summarise(
    count = n(),
    mean_income   = mean(INCOME, na.rm = TRUE),
    median_income = median(INCOME, na.rm = TRUE),
    .groups = "drop"
  )
```

Table 1: Count & summary statistics - by year and religious classification

YEAR	TRAD6	count	mean_income	median_income
2009	Chrétiennes	802	539942.35	200000
2009	Juives	18	330908.73	200000
2009	Musulmanes	64	62238.16	25000
2022	Chrétiennes	1172	474600.50	201000
2022	Juives	13	2332500.00	115000
2022	Musulmanes	42	77941.18	42500

As Table 1 reports, the number of Christian congregations is abundantly bigger thann that of Jew and Muslim congregations. Particularly fascinating is the case of Jews in 2022, where there is a large difference between the mean and the median income of the Jewish congregations, suggesting the presence of big outliers among the 13 observations in the category.

Answer 4

In order to create the binary assignment the data was grouped by year and religious affiliation, for each combination the mean income was calculated. For each congregation it was checked wheter the income of the year was equal to NA, higher than average (coded as a 1) or lower than the mean (coded as 0).

```r
NCSS <- NCSS |>
  group_by(YEAR, TRAD6) |>
```

```
mutate(
  mean_income = mean(INCOME, na.rm = TRUE),
  AVG_INCOME = case_when(
    is.na(INCOME) ~ NA_integer_,
    INCOME >= mean_income~ 1L,
    TRUE ~ 0L
  )
) |>
ungroup() |>
select(-mean_income)
```

Data Visualization

1. Create a bar plot visualizing the proportion of congregations above and below the average income (`AVG_INCOME`) in each year by 12-level religious classification (`TRAD12`). Hint: Use `facet()` for `YEAR`.

2. Make a histogram using `geom_col()` detailing the number of official members using the 12-level religious classification (`TRAD12`) distinguishing between the 6-level religious classification (`TRAD6`) in 2022. Hint: Use `facet()` for `TRAD6`, with `TRAD12` on the x-axis in addition to group/fill with the `position="dodge"`.

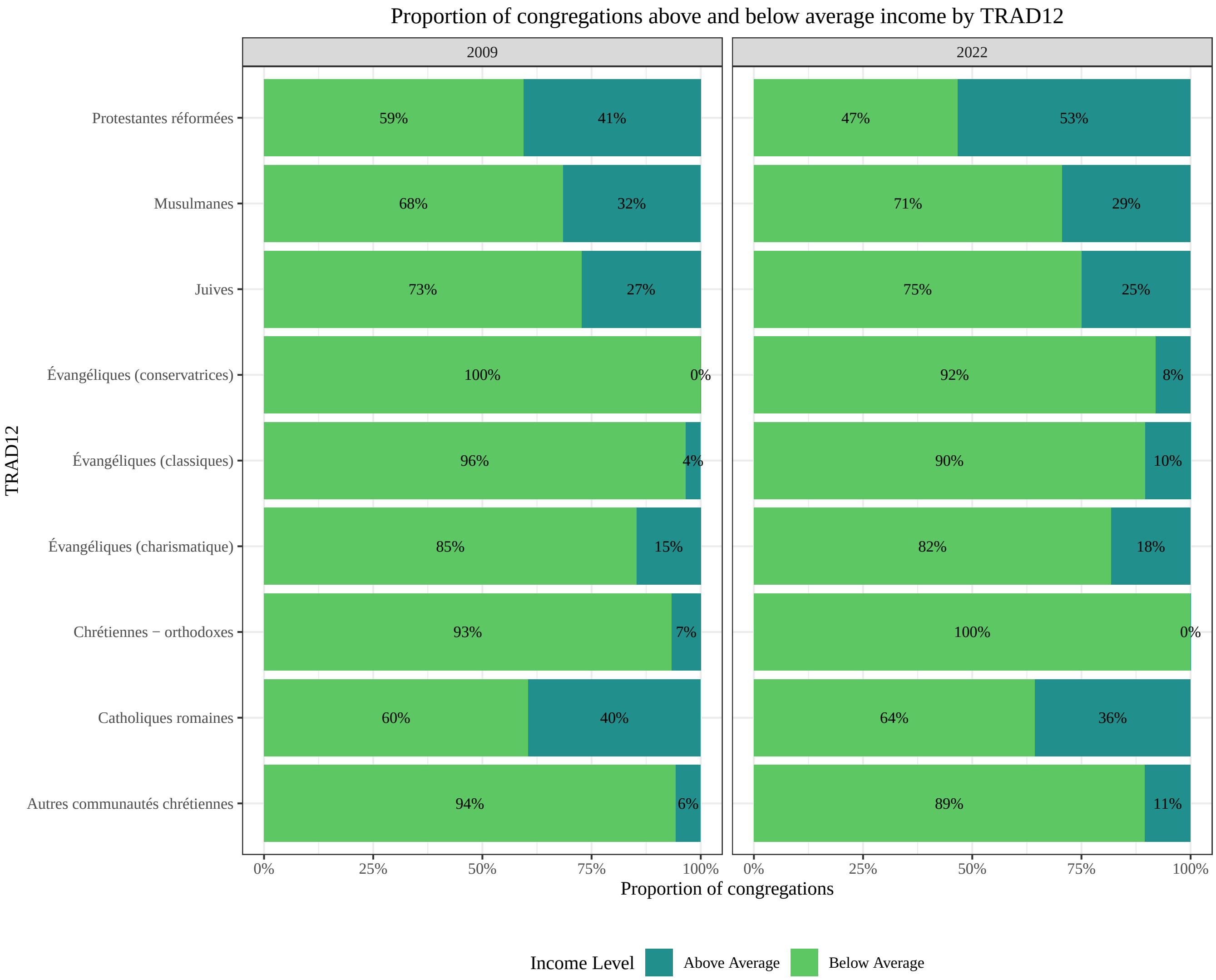
3. Display the distribution of yearly income (`INCOME`) in 2022 for congregations in each region (`GDREGION`) using ridge plots.

4. Create a boxplot of the number of official members per congregation in 2022 by religious classification (`TRAD6`) and region (`GDREGION`). Hint: Use `facet()` for `GDREGION`.

Answer 1

To compute the proportion of congregations with above-average income for each year and 12-level religious classification, the data was further wrangled. For each combination, the mean of the `AVGINCOME` binary variable was calculated, yielding the share of congregations with a value of 1. The proportion reported is out of the congregations reporting their annual income level, thus removing those having NA values for the `INCOME` variable. The proportion of below average income levels congregations was computed as 1 - %above average.
The data was then pivoted so that for each year the proportion of above and below income level congregations would be in the same column, `IncomeLevel`.

```r
NCSS_viz1 <- NCSS |>
  group_by(YEAR, TRAD12) |>
  summarise(prop_above = mean(AVG_INCOME, na.rm = TRUE), .groups = "drop") |>
  mutate(prop_below = 1 - prop_above) |>
  pivot_longer(cols = c(prop_above, prop_below),
               names_to = "IncomeLevel", values_to = "Proportion") |>
  mutate(IncomeLevel = recode(IncomeLevel,
                              prop_above = "Above Average",
                              prop_below = "Below Average"))
```



As shown in Figure 1, the TRAD-12 subgroup with the higher share of congregation having an income above the average is Protestant Reformers. The subgroup had a share equal to 40.6% in 2009 and 53.4% in 2022.

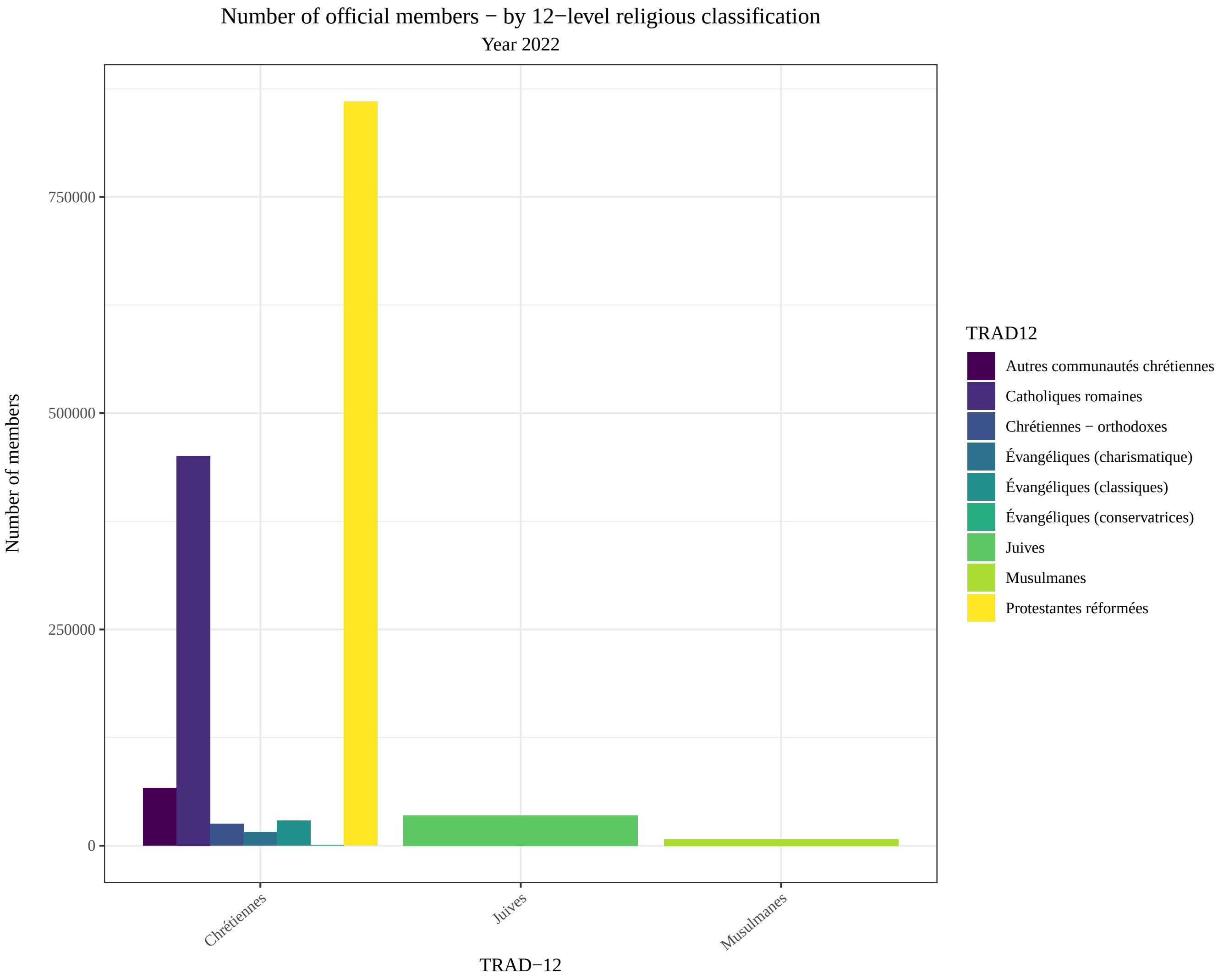
```
ggplot(NCSS_viz1, aes(x = TRAD12, y = Proportion, fill = IncomeLevel)) +
  geom_col(position = "stack") +
  geom_text(aes(label = scales::percent(Proportion, accuracy = 1)),
             position = position_stack(vjust = 0.5),
             size = 3) +
  facet_wrap(~ YEAR) +
  coord_flip() +
  scale_y_continuous(labels = percent_format()) +
  scale_fill_manual(values = my_colors, name = "Income Level") +
  labs(
```

```r
    title = "Proportion of congregations above and below average income by TRAD12",
    x = "\nTRAD12",
    y = "Proportion of congregations\n"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5),
    legend.position = "bottom"
  )
```

Answer 2

The number of official member for each TRAD-12 group in the year 2022 was computed by summing the number of official members in each of its congregations.

```r
NCSS_viz2 <- NCSS |>
  filter(YEAR == 2022) |>
  group_by(TRAD6, TRAD12) |>
  summarise(
    total_members = sum(NUMOFFMBR, na.rm = TRUE),
    .groups = "drop"
  ) |>
  filter(total_members > 0)
```

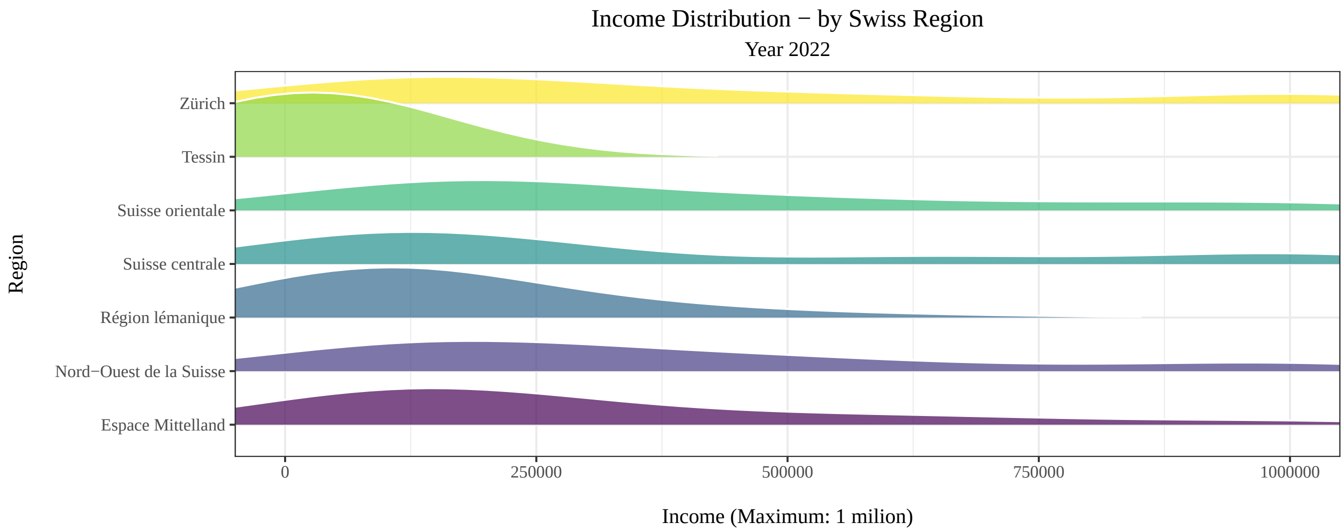


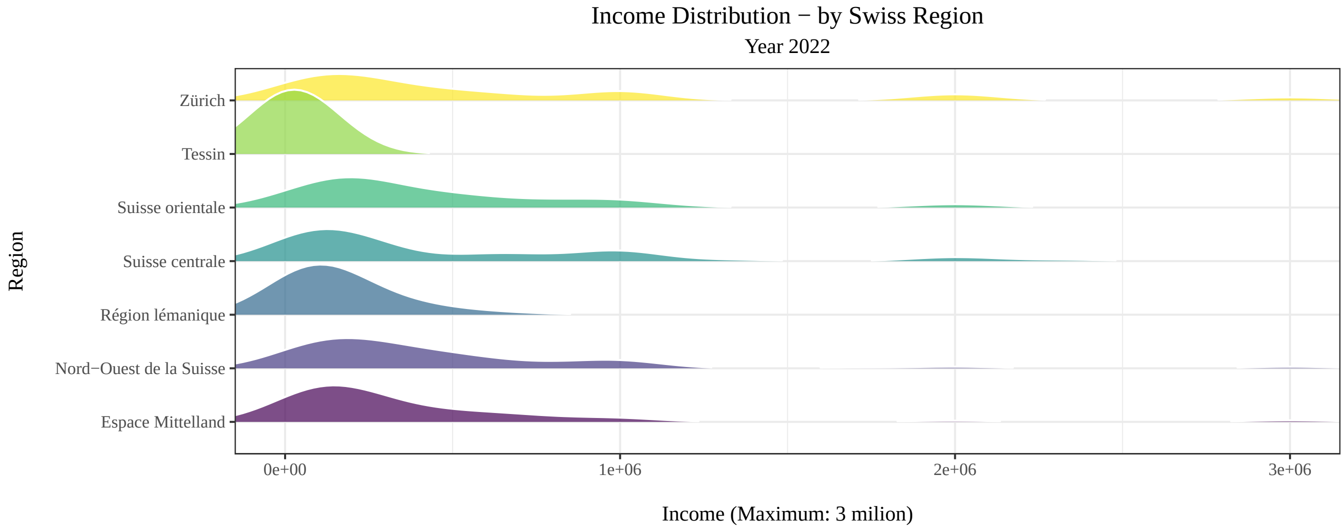
The Protestant Reformers are the group with the highest number of official members in 2022, followed by the Roman Catholics.

```
NCSS_viz2 |>
  ggplot(aes(x = TRAD6, y = total_members, fill = TRAD12)) +
  geom_col(position = "dodge")+
  scale_fill_viridis_d(name = "TRAD12") +
  labs(
    title = "Number of official members - by 12-level religious classification",
    subtitle = "Year 2022",
    x = "TRAD-12\n",
    y = "\nNumber of members\n"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 40, hjust = 1, vjust = 1)
  )
```

Answer 3

For the following visualization, the data were filtered to include only observations from the year 2022. Since the observations span a wide range of income levels, two ridge plots were created: in one, the maximum x-value is set to 1 million, and in the other, it is set to 3 million. This approach allows for a clearer view of the distributions both in detail and overall.





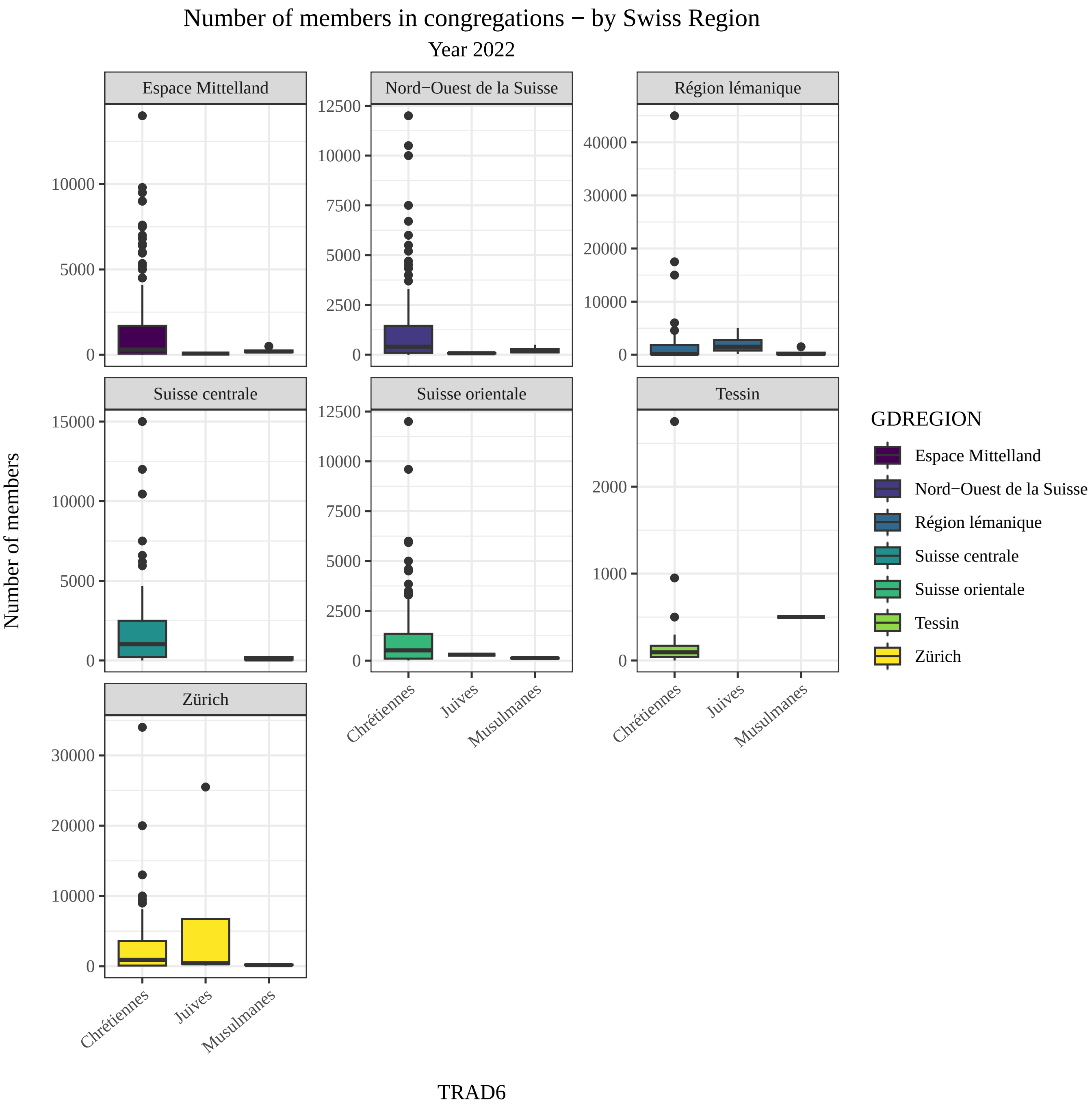
By observing both ridge plots, it is evident that congregations in the Tessin region have an income distribution that is highly left-skewed. Similarly, congregations in the Région lémanique do not exhibit a long tail, although their distribution is less steep than that of Tessin.

```
smaller_viz3 <- NCSS |>
  filter(YEAR == 2022) |>
  na.omit()|>
ggplot(aes(x = INCOME, y = GDREGION, fill = GDREGION)) +
  geom_density_ridges(alpha = 0.7, color = "white", scale = 1.2,
                      rel_min_height = 0.01)  +
  coord_cartesian(xlim = c(0, 1000000))+
  scale_fill_viridis_d(name = "GDREGION") +
```

```r
  labs(title = "Income Distribution - by Swiss Region",
       subtitle = "Year 2022",
        x = "\nIncome (Maximum: 1 milion)",
       y = "Region\n",
       fill = "Region") +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "none"
  )
bigger_viz3 <- NCSS |>
  filter(YEAR == 2022) |>
  na.omit()|>
  ggplot(aes(x = INCOME, y = GDREGION, fill = GDREGION)) +
  geom_density_ridges(alpha = 0.7, color = "white", scale = 1.2,
                      rel_min_height = 0.01)  +
  coord_cartesian(xlim = c(0, 3000000))+
  scale_fill_viridis_d(name = "GDREGION") +
  labs(title = "Income Distribution - by Swiss Region",
       subtitle = "Year 2022",
       x = "\nIncome (Maximum: 3 milion)",
       y = "Region\n",
       fill = "Region") +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "none"
  )
```

Answer 4

For the creation of the boxplots the data was filter so to only mantain observation from the year 2022. The figure allow for the observation of the distribution of the number of members for congregation by combinations of TRAD-6 religious classification and Swiss Region.



In most regions, Christian congregations have a higher median number of members and a more dispersed distribution. Exceptions include the Région lémanique, where Jewish congregations have a higher median, and the Tessin region, where Muslim congregations show a relatively compact but high distribution.

```
NCSS|>
  filter(YEAR == 2022) |>
  ggplot(aes(x = TRAD6, y = NUMOFFMBR, fill = GDREGION)) +
  geom_boxplot() +
  facet_wrap(~ GDREGION, scales = "free_y")+
  scale_fill_viridis_d(name = "GDREGION") +
  labs(title = "Number of members in congregations - by Swiss Region",
       subtitle = "Year 2022",
       x = "\nTRAD6",
       y = "Number of members\n",
       fill = "Region") +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 40, hjust = 1, vjust = 1)
  )
```