

Problem Set 2

Hanyu Li(ID:25346841)

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)

```

1 # 1. Import and sort out data
2 raw_data <- read_csv("NCSS_v1.csv")
3 data <- raw_data |>
4   select(CASEID, YEAR, GDREGION, NUMOFFMBR, TRAD6, TRAD12, INCOME)

```

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

```

1 # 2. Filter dataset using TRAD6
2 data_fil <- data |>
3   filter(TRAD6 %in% c("Chrétiennes", "Juives", "Musulmanes"))

```

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

```

1 # 3. Compute for the number of congregations by religious classification
2 stats <- data_fil |>
3   group_by(TRAD6, YEAR) |>
4     summarise(
5       count = n(),
6       income_avg = mean(INCOME, na.rm = TRUE),
7       income_med = median(INCOME, na.rm = TRUE),
8       .groups = "drop"

```

The outcome is shown in Table 1.

Table 1: Average and Median Income by Religion and Year

Religion	Year	Count	Mean Income	Median Income
Chrétiennes	2009	802	539,942	200,000
Chrétiennes	2022	1,172	474,601	201,000
Juives	2009	18	330,909	200,000
Juives	2022	13	2,332,500	115,000
Musulmanes	2009	64	62,238	25,000
Musulmanes	2022	42	77,941	42,500

4. Create a categorical variable called AVG_INCOME that is binary, where 1 indicates average or above-average income and 0 indicates below-average income within each year. To prepare for subsequent visualization tasks, AVG_INCOME was transformed into a factor at the beginning of the workflow.

```

1 # 4. Create binary variable AVG_INCOME
2 income_data <- data_fil |>
3   filter(!is.na(INCOME)) |>
4   group_by(YEAR) |>
5   mutate(
6     year_avg = mean(INCOME, na.rm = TRUE),
7     AVG_INCOME = ifelse(INCOME >= year_avg, 1, 0),
8     Income_Level = ifelse(AVG_INCOME == 1, "Above Average", "Below
9     Average"),
10    Income_Level = factor(Income_Level, levels = c("Below Average", "
11      Above Average")))
10  ) |>
11  ungroup()

```

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). To improve the readability of the congregation category labels, the x- and y-axes were flipped.

```

1 # DV1. Create a bar plot
2 pdf("plot1.pdf")
3 plot1 <- ggplot(income_data, aes(x = TRAD12, fill = Income_Level)) +
4   geom_bar(position = "fill", width = 0.6) +
5   facet_wrap(~YEAR) +
6   coord_flip() +
7   scale_y_continuous(labels = scales::percent) +
8   scale_fill_manual(values = c("Below Average" = "grey", "Above Average"
9     = "skyblue")) +
10  labs(
11    title = "Proportion of Congregations Above/Below Average Income",
12    x = NULL,
13    y = "\nProportion",
14    fill = "Income Group") +
15  theme_minimal() +
16  theme(
17    plot.title = element_text(size = 12, hjust = 0.5, face = "bold"),
18    axis.title.x = element_text(size = 10, face = "bold"),
19    legend.text = element_text(size = 10),
20    legend.title = element_text(size = 10),
21    legend.position = "top",
22    panel.grid.major.y = element_blank())

```

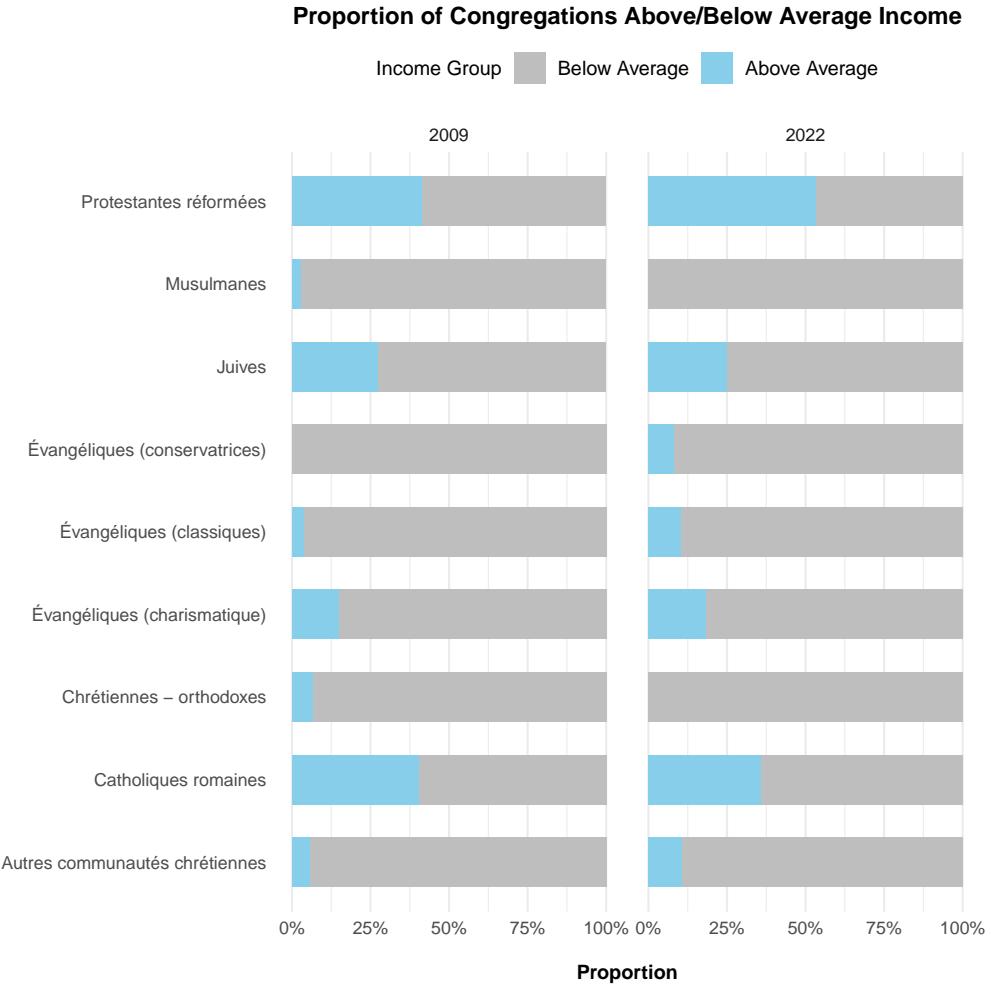


Figure 1: Proportion of Congregations Above and Below Average Income

2. Make a bar plot using `geom_col()` showing the number of official members by 12-level religious classification (`TRAD12`), distinguishing between 6-level religious classifications (`TRAD6`) in 2022. The same color scheme was used to maintain visual consistency.

```

1 data2 <- data.fil |>
2   filter(!is.na(NUMOFFMBR)) |>
3   filter(YEAR == 2022)
4
5 pdf("plot2.pdf")
6 plot2 <- ggplot(data2, aes(x = TRAD12, y = NUMOFFMBR, fill = TRAD6)) +
7   geom_col(width = 0.6) +
8   coord_flip() +
9   scale_fill_manual(values = c("Chr tiennes" = "cadetblue2",
10                      "Juives" = "burlywood2",
11                      "Musulmanes" = "cornsilk3")) +
12   labs(

```

```

13   title = "Official Members by Religion Groups",
14   x = NULL,
15   y = "\nNumber of Official Members",
16   fill = "Religon Groups (TRAD6)") +
17   theme_minimal() +
18   theme(
19     plot.title = element_text(size = 12, hjust = 0.5, face = "bold"),
20     axis.title.x = element_text(size = 10, face = "bold"),
21     axis.title.y = element_text(size = 10, face = "bold"),
22     legend.text = element_text(size = 10),
23     legend.title = element_text(size = 10),
24     legend.position = "top",
25     panel.grid.major.y = element_blank())

```

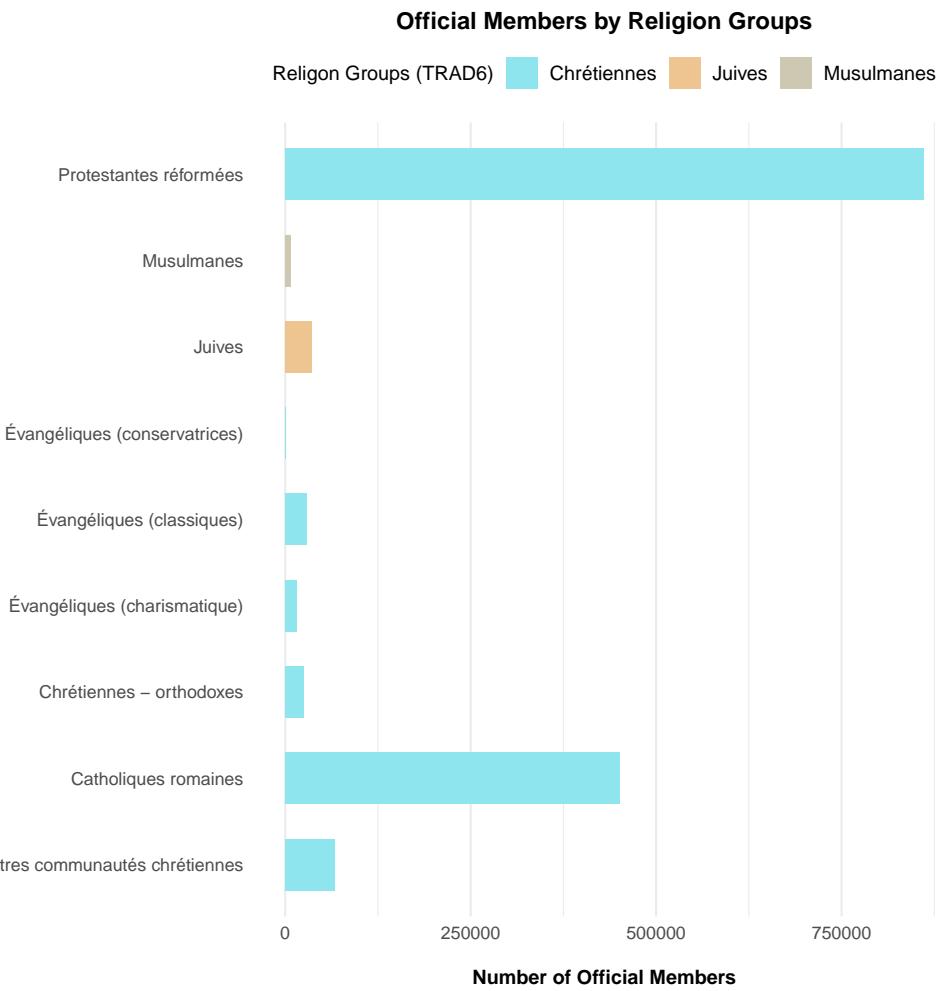


Figure 2: Number of Official Members by Religious Classification

3. Display the distribution of average yearly income (INCOME) for congregations in 2022

across regions (GDREGION) using ridge plots. Regional income means are highlighted with orange points.

```

1 # DV3. Display the distribution of average yearly income (INCOME) for
2   congregations
3 data3 <- data_fil |>
4   filter(YEAR == 2022) |>
5   filter(!is.na(INCOME)) |>
6   filter(!is.na(GDREGION))
7 summary(data3) # get the income range to decide x-axis range
8
9 region_means <- data3 |>
10  group_by(GDREGION) |>
11  summarise(mean_income = mean(INCOME))
12
13 library(ggribbles)
14 pdf("plot3.pdf")
15 plot3 <- ggplot() +
16   geom_density_ridges(data = data3,
17     aes(x = INCOME, y = GDREGION, fill = GDREGION),
18     alpha = 0.6, scale = 1.8, rel_min_height = 0.01) +
19   geom_point(data = region_means,
20     aes(x = mean_income, y = GDREGION, color = "Average Mean"),
21     size = 2) +
22   scale_x_continuous(
23     limits = c(0, 9000000)) +
24   scale_fill_viridis_d(option = "mako", name = "Region") +
25   scale_color_manual(name = NULL,
26     values = c("Average Mean" = "orange")) +
27   labs(
28     title = "Income Distribution by Region (2022)\n",
29     x = "\nTotal Income",
30     y = NULL) +
31   theme_minimal() +
32   theme(
33     plot.title = element_text(size = 12, hjust = 0.5, face = "bold"),
34     axis.title.x = element_text(size = 10, face = "bold"),
35     axis.title.y = element_text(size = 10, face = "bold"),
36     legend.text = element_text(size = 8),
37     legend.title = element_text(size = 10, face = "bold"),
38     panel.grid.minor = element_blank(),
39     legend.position = "right")

```

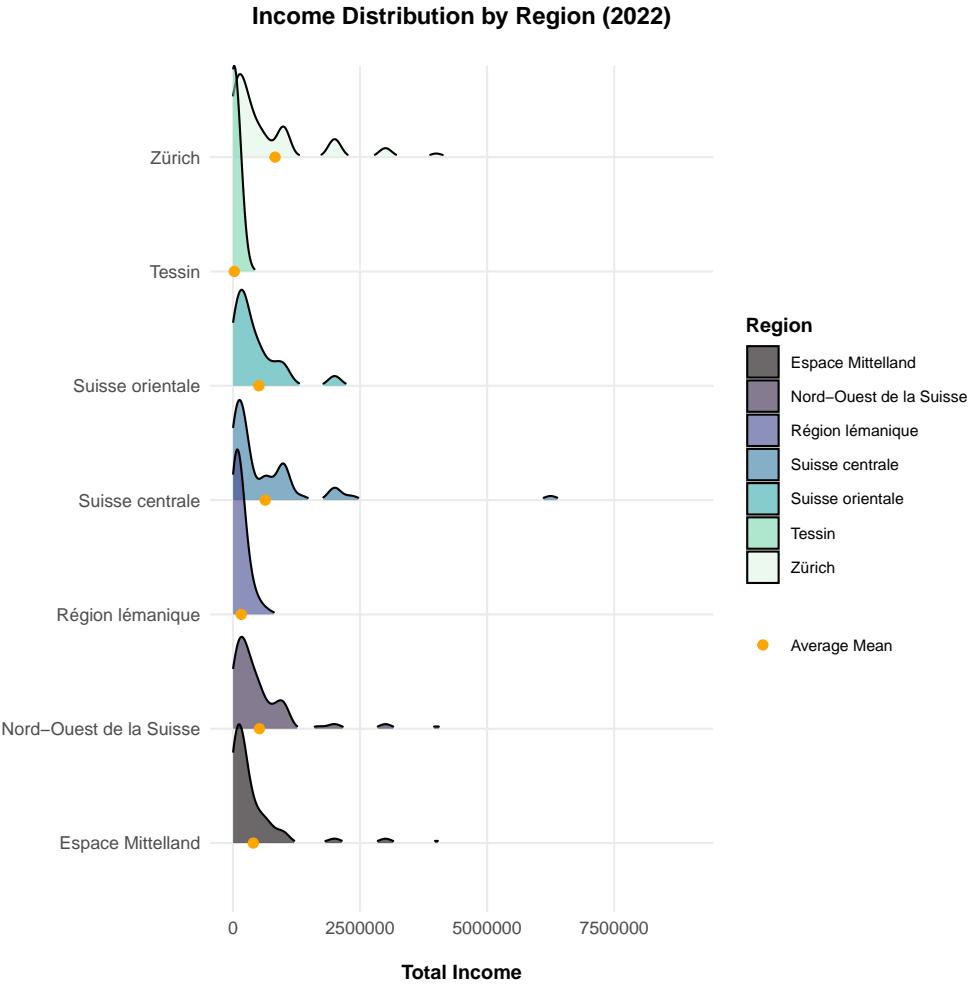


Figure 3: Distribution of Average Income by Region

4. Create a boxplot of the number of official members per congregation in 2022 by religious classification (TRAD6) and region (GDREGION). Because the maximum number of official members is very large, the y-axis was restricted to the range 0–5,000 to improve visual comparability across religious groups.

```

1 # DV4. Create a boxplot of official members by TRAD6 and GDREGION
2 data4 <- data.fil |>
3   filter(YEAR == 2022) |>
4   filter(!is.na(NUMOFFMBR)) |>
5   filter(!is.na(GDREGION))
6
7 pdf("plot4.pdf")
8 plot4 <- ggplot(data4, aes(x = TRAD6, y = NUMOFFMBR, fill = TRAD6)) +
9   geom_boxplot(outlier.size = 1, alpha = 0.8) +
10  facet_wrap(~ GDREGION, scales = "free_x") +
11  scale_y_continuous(limits = c(0, 5000)) +

```

```

12 scale_fill_manual(values = c("Chr tiennes" = "cadetblue2",
13                     "Juives" = "burlywood2",
14                     "Musulmanes" = "cornsilk3")) +
15 labs(
16   title = "Distribution of Official Members by Region and Religion
17 (2022)" ,
18   subtitle = "excluding outliers > 5,000",
19   x = NULL,
20   y = "Number of Official Members\n",
21   fill = "Religon Groups (TRAD6)") +
22 theme_bw() +
23 theme(
24   plot.title = element_text(size = 12, hjust = 0.5, face = "bold"),
25   plot.subtitle = element_text(size = 10, hjust = 0.5, face = "italic")
26   ,
27   axis.title.x = element_text(size = 10, face = "bold"),
28   axis.title.y = element_text(size = 10, face = "bold"),
29   legend.text = element_text(size = 10),
30   legend.title = element_text(size = 10),
31   legend.position = "top",
32   panel.grid.major.y = element_blank())

```

Distribution of Official Members by Region and Religion (2022)

excluding outliers > 5,000

Religion Groups (TRAD6) Chrétiennes Juives Musulmanes

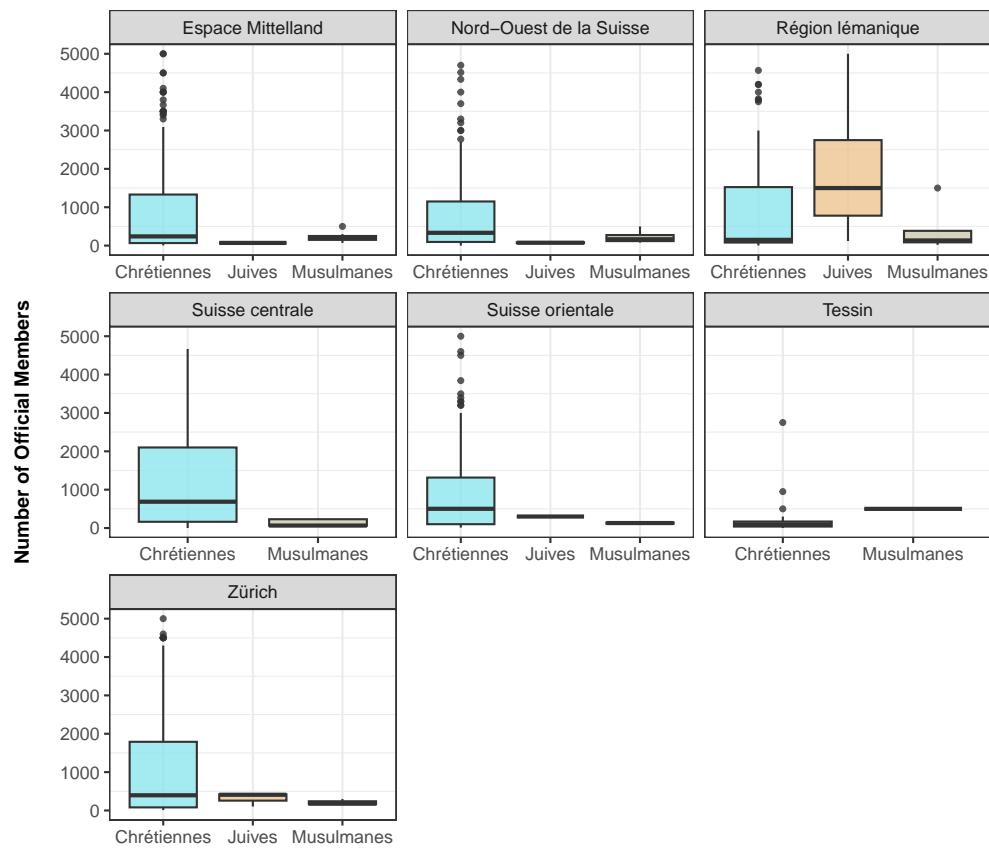


Figure 4: Number of Official Members by Religion and Region