Problem Set 2

Data Visualization
Shelly Veal-Upham
25337422

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`)

```r
dataset_backup <- read_csv("../../../datasets/NCSS_v1.csv")
data <- dataset_backup

data <- data %>%
  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.

   * Note here that the r code reflects ”Chretiennes” whereas the dataset uses Chrétiennes – I did actually filter with the use of Chrétiennes, but for compiling purposes it is reflected here as ”Chretiennes.”

```r
data <- data %>%
  filter(TRAD6 == c("Chretiennes",
                    "Juives",
                    "Musulmanes"))

unique(data$TRAD6) # Checking to ensure filtered correctly
```

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.

```r
congregations_by_year <- data %>%
  group_by(TRAD6, YEAR) %>%
  summarize(
    N = n()
  )

view(congregations_by_year)

income_by_year <- data %>%
  group_by(TRAD6, YEAR) %>%
  summarize(
    mean_income = mean(INCOME, na.rm = TRUE),
    median_income = median(INCOME, na.rm = TRUE)
  )

view(income_by_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.

```
mean_2009 <- mean(data$INCOME[data$YEAR == 2009], na.rm = TRUE)
mean_2022 <- mean(data$INCOME[data$YEAR == 2022], na.rm = TRUE)

data <- data %>%
  mutate(
    AVG_INCOME = case_when(
      (YEAR = 2009 & INCOME >= mean_2009) ~ 1,
      (YEAR = 2009 & INCOME < mean_2009) ~ 0,
      (YEAR = 2022 & INCOME >= mean_2022) ~ 1,
      (YEAR = 2022 & INCOME < mean_2022) ~ 0
    )
  )

data$AVG_INCOME <- factor(data$AVG_INCOME,
                          levels = c(0, 1),
                          labels = c("Below Average", "Above Average"))
```

Data Visualization

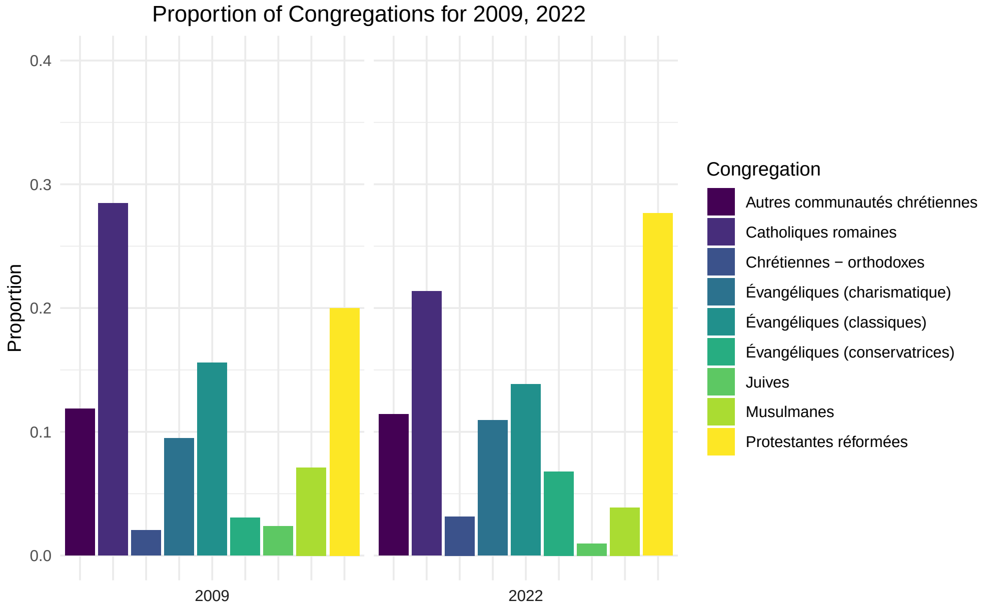
1. Create a bar plot visualizing the proportion of congregations by 12-level religious classification (`TRAD12`) in each year.

```
prop_12 <- data %>%
  group_by(TRAD12, YEAR) %>%
  summarize(N = n()) %>%
  group_by(YEAR) %>%
  mutate(PROPORTION = N/sum(N))

sum(prop_12$PROPORTION[prop_12$YEAR == 2009]) # Checking 2009
sum(prop_12$PROPORTION[prop_12$YEAR == 2022]) # Checking 2022

bp_1.1 <- ggplot(prop_12, aes(
    x = TRAD12,
    y = PROPORTION,
    fill = TRAD12)) +
  ylim(0, .4) +
  geom_col() +
  facet_wrap(prop_12$YEAR, strip.position = "bottom") +
  scale_fill_viridis_d() +
  theme(axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(
    title = "Proportion of Congregations for 2009, 2022",
    y = "Proportion",
    fill = "Congregation"
  )
```

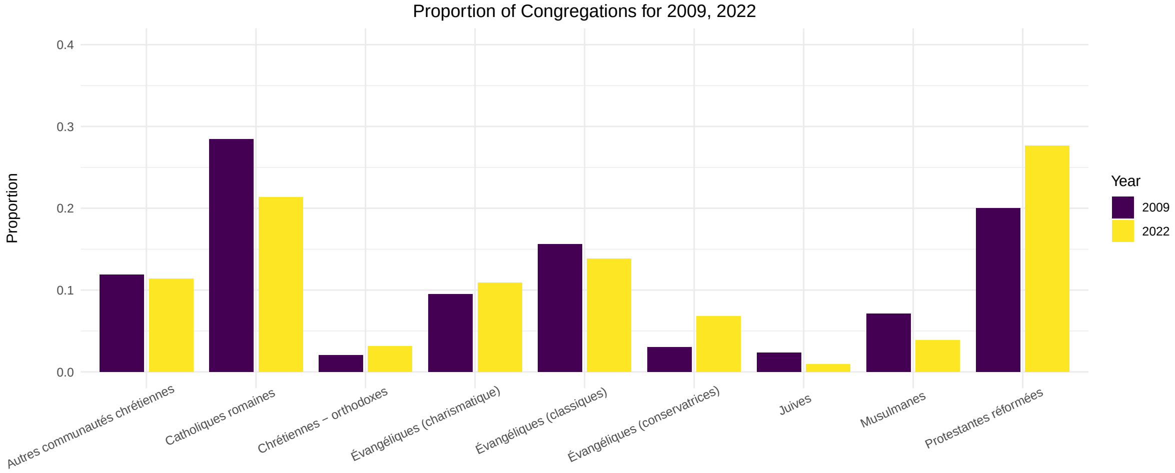
Figure 1: Proportions of Congregations by Year



In Figure 1 we have the proportions of congregations wrapped by year (2009 & 2022), but I thought it'd be helpful also to see the years side-by-side for each congregation so we can visualize their proportional growth (see Figure 2).

```
bp_1.2 <- ggplot(prop_12, aes(
  x = TRAD12,
  y = PROPORTION,
  fill = factor(YEAR))) +
  ylim(0, .4) +
  geom_col(position = "dodge2") +
  scale_fill_viridis_d() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_blank(),
        axis.text.x = element_text(angle = 25, hjust = 0.85)) +
  labs(
    title = "Proportion of Congregations for 2009, 2022",
    y = "Proportion\n\n",
    fill = "Year"
  )
```

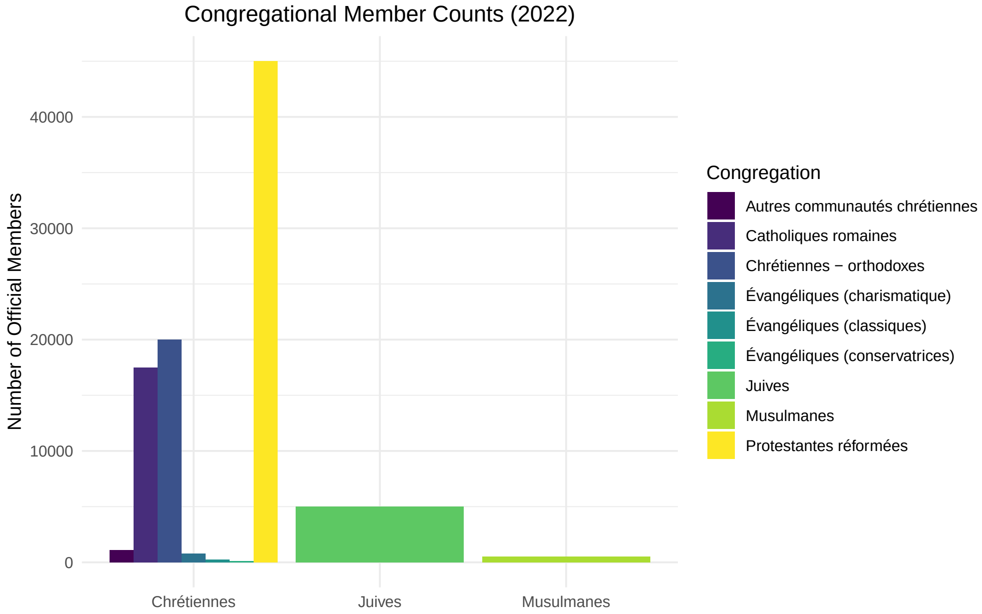
Figure 2: Proportional Growth of Congregations from 2009 to 2022



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

```r
bp_2.1 <- ggplot(data, aes(x = TRAD6, y = NUMOFFMBR, fill = TRAD12)) +
  geom_col(position = "dodge") +
  scale_fill_viridis_d() +
  theme(axis.title.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(
    title = "Congregational Member Counts (2022)",
    y = "Number of Official Members",
    fill = "Congregation"
  )
```

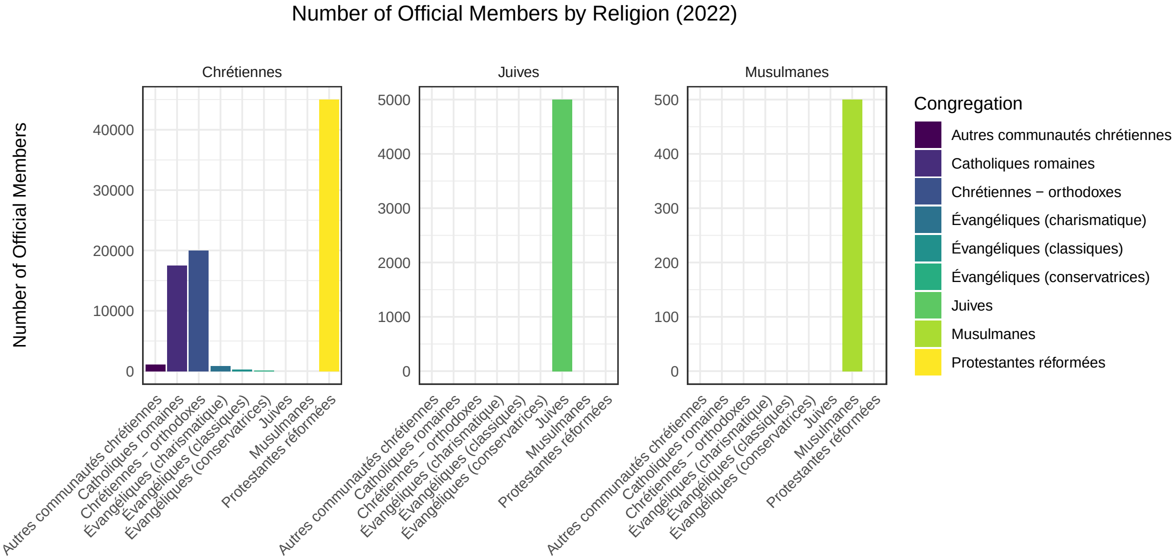
Figure 3: Number of Members by Congregation



For Figure 3 I have used `geom_col()` to compile the total numbers of official members in each congregation and grouped the x-axis by the variable `TRAD6`. Figure 4 shows much the same information, but wrapped by overarching religious affiliation (`TRAD12`) with differing y-axis scales.

```
bp_2.2 <- ggplot(data %>% filter(YEAR == 2022),
                    aes(x = TRAD12, y = NUMOFFMBR, fill = TRAD12)) +
  geom_col(position = "dodge") +
  facet_wrap(~ TRAD6, scales = "free_y") +
  scale_fill_viridis_d() +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5),
    panel.spacing = unit(1.5, "lines"),
    panel.border = element_rect(color = "grey20", fill = NA, linewidth =
   0.8)
  ) +
  labs(
    title = "Number of Official Members by Religion (2022)\n",
    x = "",
    y = "Number of Official Members\n\n\n",
    fill = "Congregation"
  )
```

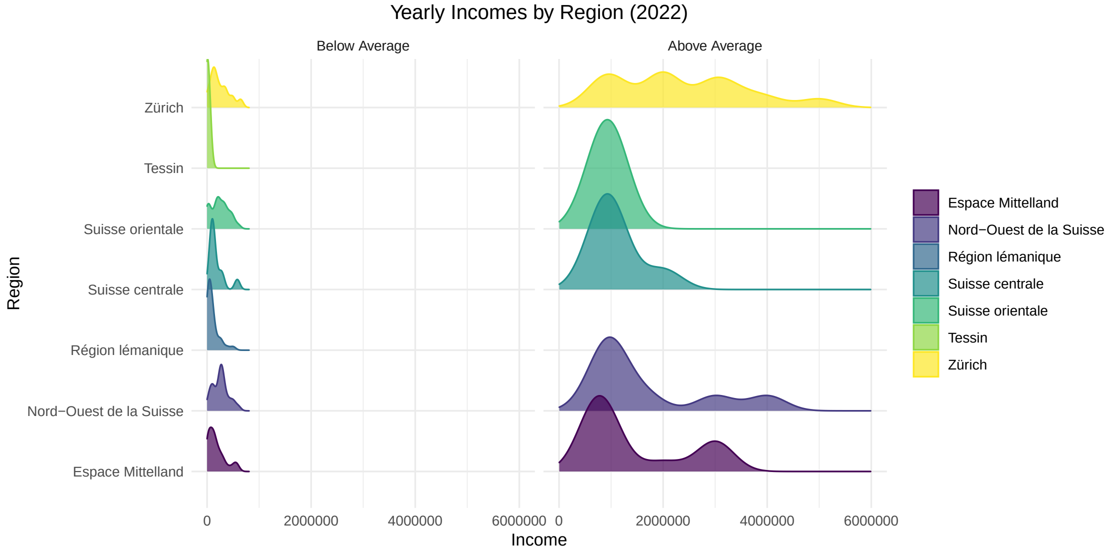
Figure 4: Number of Congregation Members by Religion



3. Display the distribution of congregations in 2022 above and below the average yearly income (`AVG_INCOME`) in each region using ridge plots.

```r
ridges_together <- ggplot(data %>% filter(!is.na(AVG_INCOME), YEAR ==
    2022),
        aes(
          x = INCOME,
          y = GDREGION,
          fill = GDREGION,
          color = GDREGION)) +
  geom_density_ridges(alpha = 0.7) +
  facet_wrap(~ AVG_INCOME, labeller = as_labeller(data$AVG_INCOME)) +
  theme(legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(x = "Income",
       y = "Region",
       title = "Yearly Incomes by Region (2022)") +
  xlim(0, 6000000) +
  scale_fill_viridis_d() +
  scale_color_viridis_d()
```

Figure 5: Congregations Above and Below Average Yearly Income



Because the scale of income is so different for those congregations with above and below average yearly incomes, I decided also to separate the two groups into individual plots for a closer look. See Figures 6 and 7.

```r
ridges_above <- data %>% filter(!is.na(AVG_INCOME), YEAR == 2022,
                                AVG_INCOME == 'Above Average')
ridges_above_plot <- ggplot(ridges_above, aes(
                            x = INCOME,
                            y = GDREGION,
                            fill = GDREGION,
                            color = GDREGION)) +
  geom_density_ridges(alpha = 0.7) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Income",
       y = "Region",
       title = "Yearly Incomes by Region (2022, Above Average)") +
  scale_fill_viridis_d() +
  scale_color_viridis_d()
```

```r
ridges_below <- data %>% filter(!is.na(AVG_INCOME), YEAR == 2022,
                                AVG_INCOME == 'Below Average')
ridges_below_plot <- ggplot(ridges_below, aes(
  x = INCOME,
  y = GDREGION,
  fill = GDREGION,
  color = GDREGION)) +
  geom_density_ridges(alpha = 0.7) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Income",
       y = "Region",
       title = "Yearly Incomes by Region (2022, Below Average)") +
  scale_fill_viridis_d() +
  scale_color_viridis_d()
```

Figure 6: Congregations Above Average Yearly Income

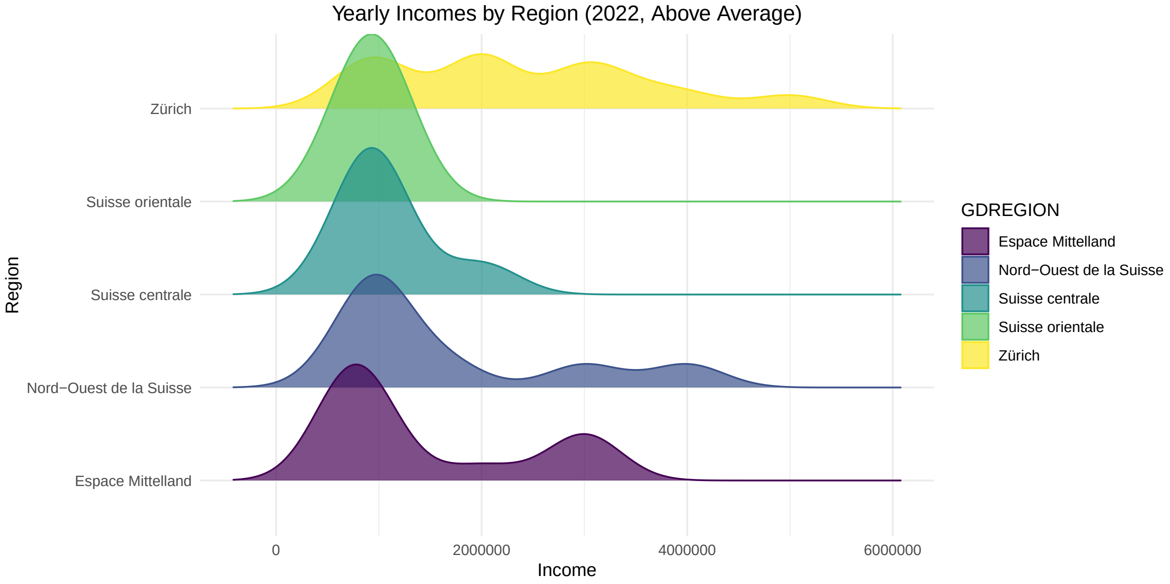
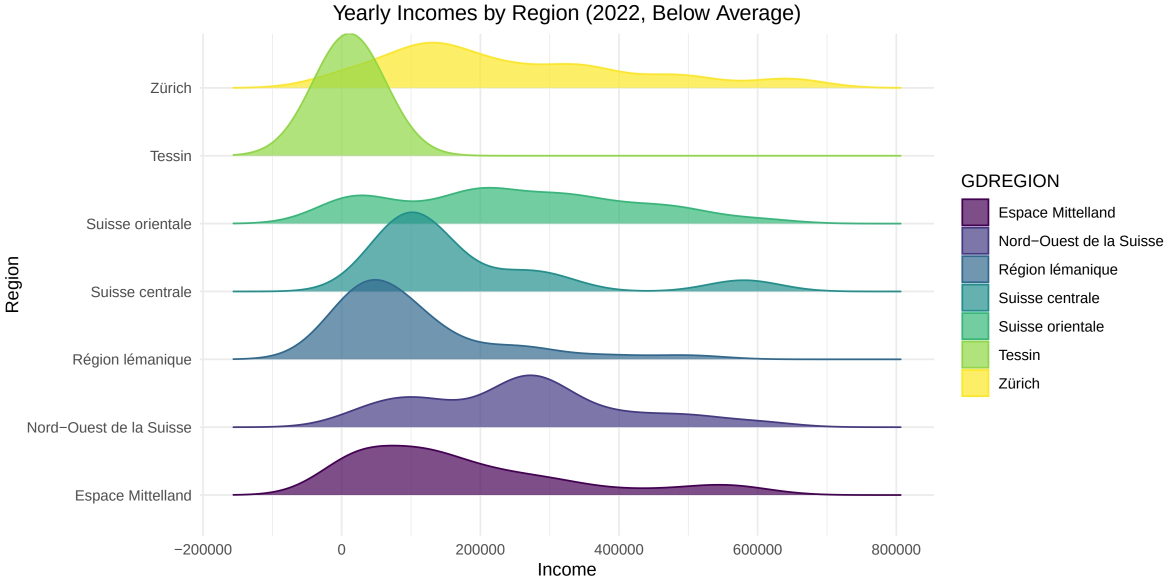


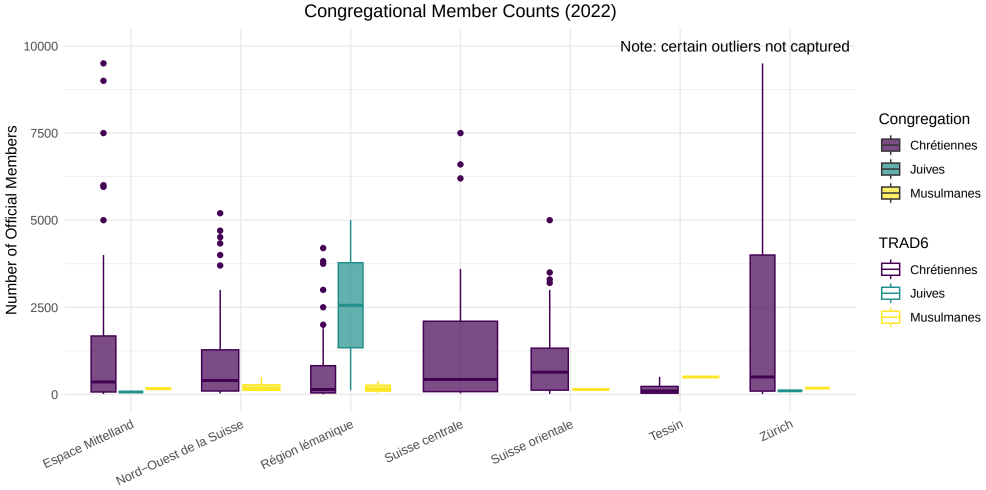
Figure 7: Congregations Below Average Yearly Income



4. Create a boxplot of the number of official members by year and region.

```r
box_plot <- ggplot(data, aes(x = GDREGION, y = NUMOFFMBR, fill = TRAD6, color = TRAD6)) +
  geom_boxplot() +
  scale_fill_viridis_d(alpha = 0.7) +
  scale_color_viridis_d() +
  theme(axis.title.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(
    title = "Congregational Member Counts (2022)",
    y = "Number of Official Members",
    fill = "Congregation"
  ) +
  ylim(0, 10000) +
  annotate("text", label = "Note: certain outliers not captured", x = 6.5, y = 10000) +
  theme(axis.text.x = element_text(angle = 25, hjust = 1))
```

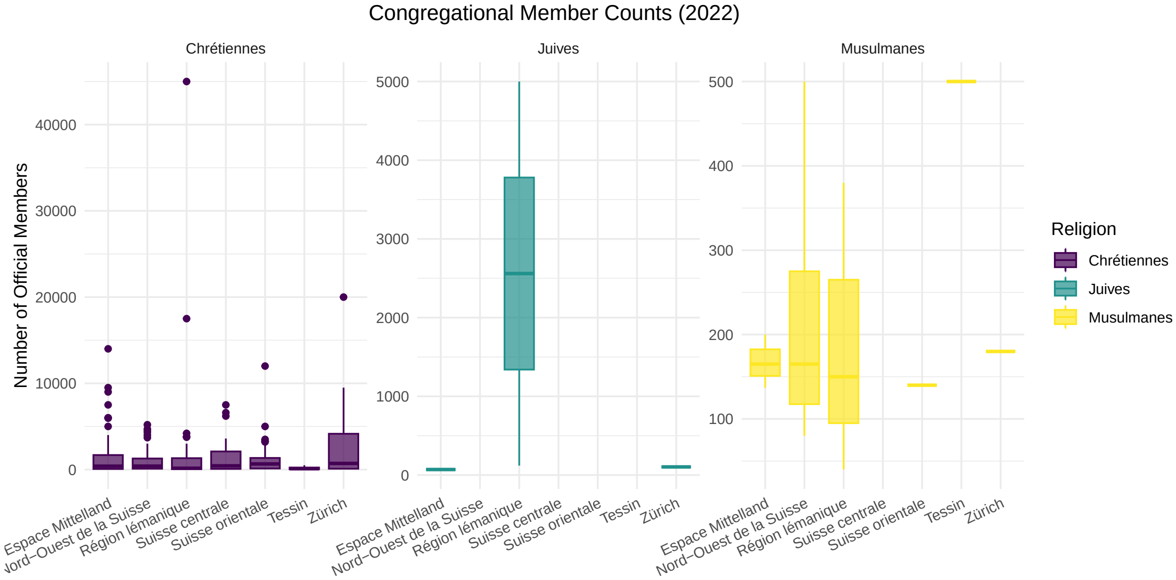
Figure 8: Official Members by Region



Again, given the drastic difference in scales among religious affiliation membership, I decided to wrap the plot according to affiliation. See Figure 9.

```r
box_plot_wrapped <- ggplot(data, aes(x = GDREGION, y = NUMOFFMBR, fill = TRAD6, color = TRAD6)) +
  geom_boxplot() +
  scale_fill_viridis_d(alpha = 0.7) +
  theme(axis.title.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  facet_wrap(~TRAD6, scales = "free_y") +
  scale_color_manual(name = "Religion",
                     labels = c("Chretiennes", "Juives", "Musulmanes"),
                     values = c("#440154", "#21918c", "#fde725")) +
  labs(
    title = ”Congregational Member Counts (2022)”,
    y = ”Number of Official Members”,
    fill = ”Religion”
  ) +
  theme(axis.text.x = element_text(angle = 25, hjust = 1))
```

Figure 9: Regional Membership by Religious Affiliation


Congregational Member Counts (2022)
Chrétiennes
Juives
Musulmanes
Number of Official Members
Espace Mittelland
Nord−Ouest de la Suisse
Région lémanique
Suisse centrale
Suisse orientale
Tessin
Zürich
Religion
Chrétiennes
Juives
Musulmanes