

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)

```

1 dataset_backup <- read_csv("../..../datasets/NCSS_v1.csv")
2 data <- dataset_backup
3
4 data <- data %>%
5   select(
6     CASEID,
7     YEAR,
8     GDREGION,
9     NUMOFFMBR,
10    TRAD6,
11    TRAD12,
12    INCOME
13  )

```

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

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

```

1 data <- data %>%
2   filter(TRAD6 == c("Chretiennes",
3                     "Juives",
4                     "Musulmanes"))
5
6 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.

```

1 congregations_by_year <- data %>%
2   group_by(TRAD6, YEAR) %>%
3   summarize(
4     N = n()
5   )
6
7 view(congregations_by_year)
8
9 income_by_year <- data %>%
10  group_by(TRAD6, YEAR) %>%
11  summarize(
12    mean_income = mean(INCOME, na.rm = TRUE),

```

```

13   median_income = median(INCOME, na.rm = TRUE)
14 )
15
16 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.

```

1 mean_2009 <- mean(data$INCOME[data$YEAR == 2009], na.rm = TRUE)
2 mean_2022 <- mean(data$INCOME[data$YEAR == 2022], na.rm = TRUE)
3
4 data <- data %>%
5   mutate(
6     AVG_INCOME = case_when(
7       (YEAR == 2009 & INCOME >= mean_2009) ~ 1,
8       (YEAR == 2009 & INCOME < mean_2009) ~ 0,
9       (YEAR == 2022 & INCOME >= mean_2022) ~ 1,
10      (YEAR == 2022 & INCOME < mean_2022) ~ 0
11    )
12  )
13
14 data$AVG_INCOME <- factor(data$AVG_INCOME,
15                           levels = c(0, 1),
16                           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.

```

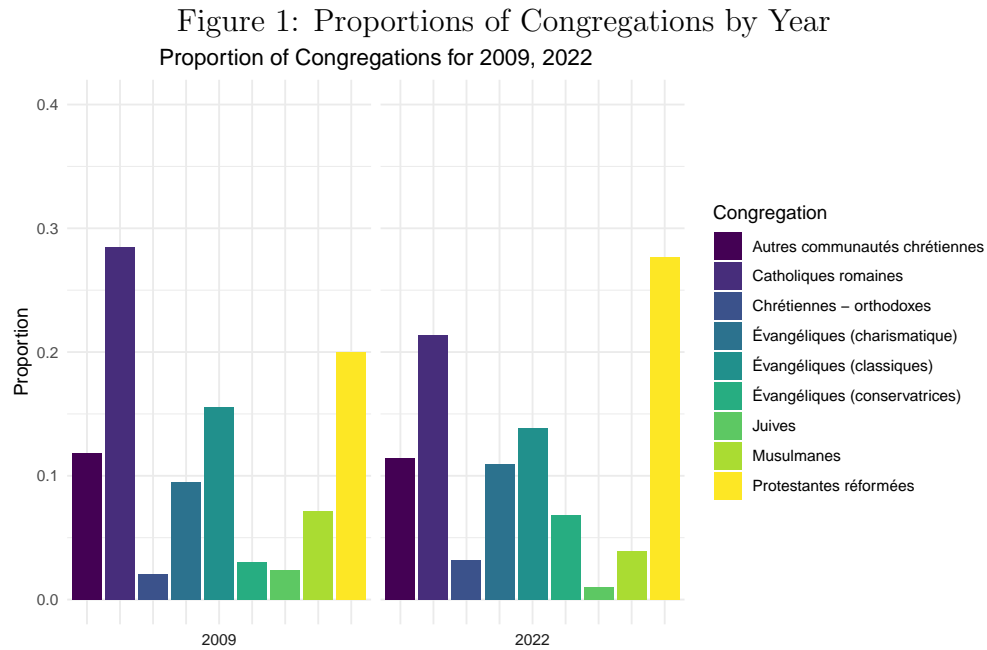
1 prop_12 <- data %>%
2   group_by(TRAD12, YEAR) %>%
3   summarize(N = n()) %>%
4   group_by(YEAR) %>%
5   mutate(PROPORTION = N/sum(N))
6
7 sum(prop_12$PROPORTION[prop_12$YEAR == 2009]) # Checking 2009
8 sum(prop_12$PROPORTION[prop_12$YEAR == 2022]) # Checking 2022
9
10 bp_1.1 <- ggplot(prop_12, aes(
11   x = TRAD12,
12   y = PROPORTION,
13   fill = TRAD12)) +
14   ylim(0, .4) +
15   geom_col() +
16   facet_wrap(prop_12$YEAR, strip.position = "bottom") +
17   scale_fill_viridis_d() +
18   theme(axis.text.x = element_blank(),
19         axis.title.x = element_blank(),
20         plot.title = element_text(hjust = 0.5)) +

```

```

21 labs(
22   title = "Proportion of Congregations for 2009, 2022",
23   y = "Proportion",
24   fill = "Congregation"
25 )

```



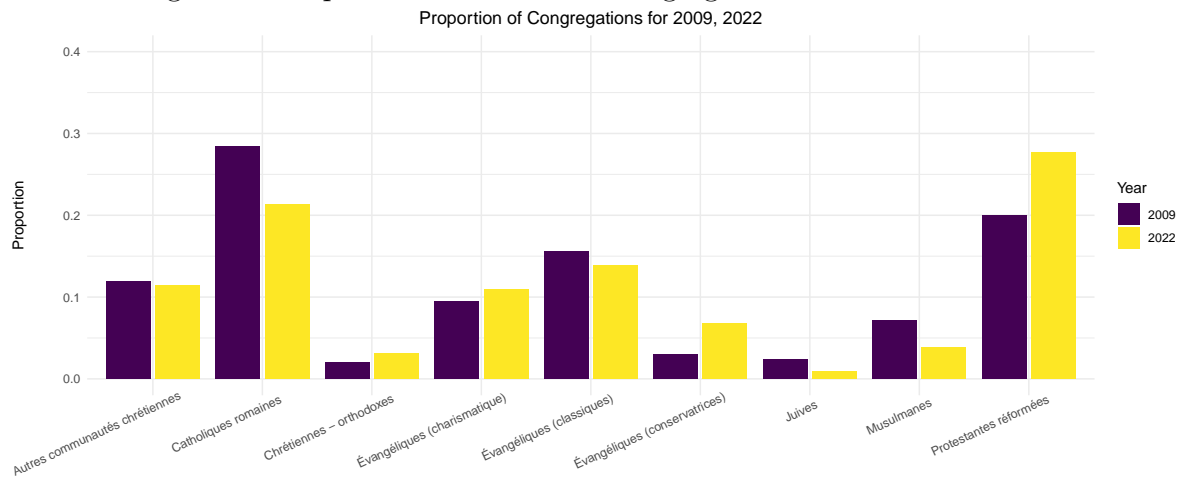
In Figure 1 we have the proportions of congregations wrapped by year (2009 & 2022), but I thought it'd be helpful also to see the years side-by-side for each congregation so we can visualize their proportional growth (see Figure 2).

```

1 bp_1.2 <- ggplot(prop_12, aes(
2   x = TRAD12,
3   y = PROPORTION,
4   fill = factor(YEAR)) +
5   ylim(0, .4) +
6   geom_col(position = "dodge2") +
7   scale_fill_viridis_d() +
8   theme(plot.title = element_text(hjust = 0.5),
9         axis.title.x = element_blank(),
10        axis.text.x = element_text(angle = 25, hjust = 0.85)) +
11   labs(
12     title = "Proportion of Congregations for 2009, 2022",
13     y = "Proportion\n\n",
14     fill = "Year"
15   )

```

Figure 2: Proportional Growth of Congregations from 2009 to 2022



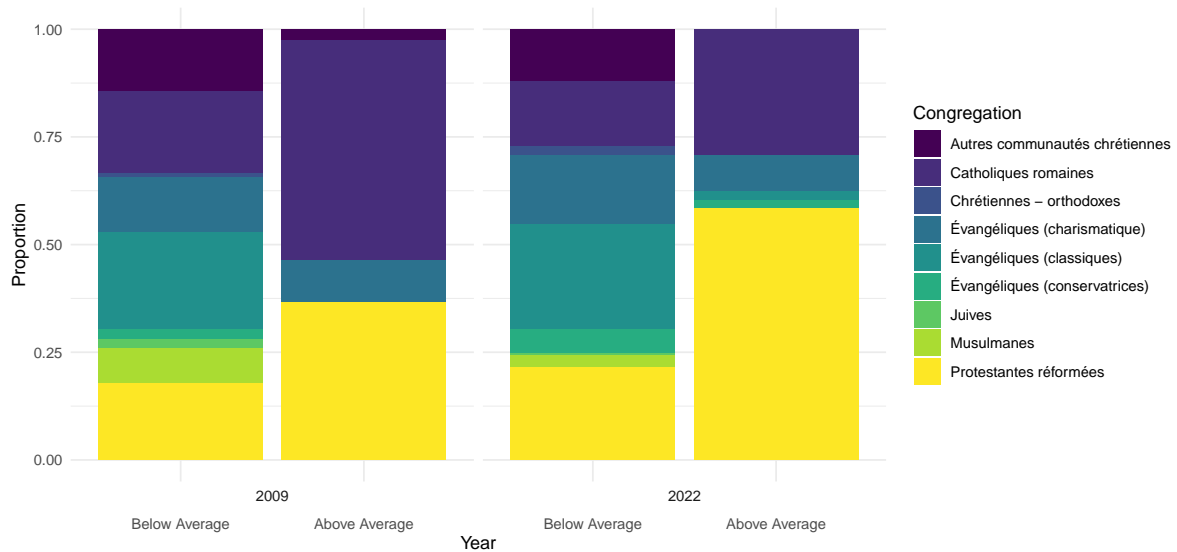
Professor Ziegler, in the updated version of the problem set, requested a proportional bar plot that distinguishes those above and below average income. See Figure 3 for my version.

```

1 prop_bar <- ggplot(data %>% filter(!is.na(AVG_INCOME)),
2   aes(
3     x = factor(AVG_INCOME),
4     fill = TRAD12)) +
5   geom_bar(position = "fill") +
6   facet_wrap(~ YEAR, strip.position = "bottom") +
7   theme(plot.title = element_text(hjust = 0.5)) +
8   labs(x = "Year",
9     y = "Proportion",
10    title = "Proportions Above and Below Average Income",
11    fill = "Congregation") +
12   scale_fill_viridis_d()

```

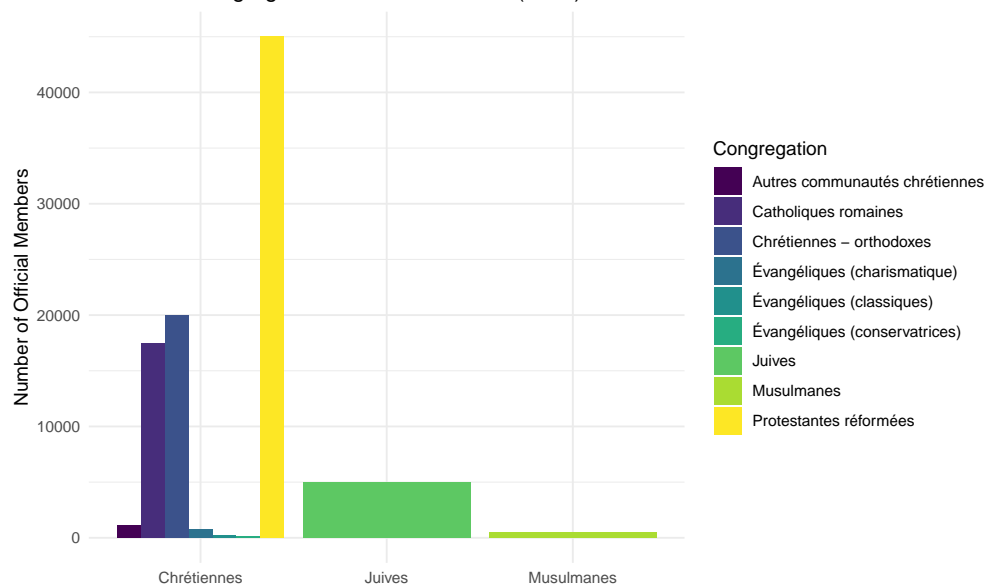
Figure 3: Proportion of Congregations Above and Below Average Income
Proportions Above and Below Average Income



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"`.

```
1 bp_2.1 <- ggplot(data, aes(x = TRAD6, y = NUMOFFMBR, fill = TRAD12)) +
2   geom_col(position = "dodge") +
3   scale_fill_viridis_d() +
4   theme(axis.title.x = element_blank(),
5         plot.title = element_text(hjust = 0.5)) +
6   labs(
7     title = "Congregational Member Counts (2022)",
8     y = "Number of Official Members",
9     fill = "Congregation"
10  )
```

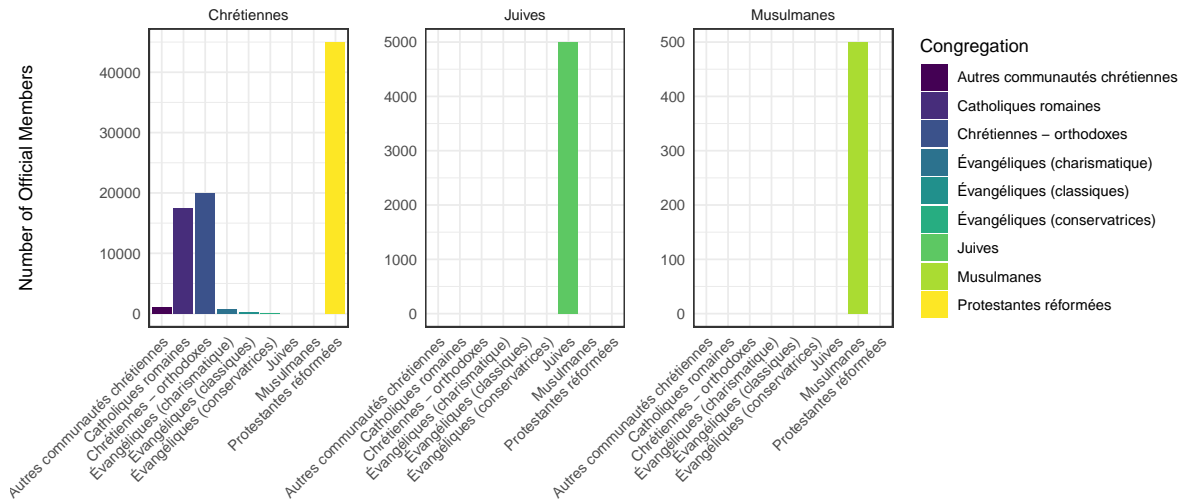
Figure 4: Number of Members by Congregation
Congregational Member Counts (2022)



For Figure 4 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 5 shows much the same information, but wrapped by overarching religious affiliation (`TRAD12`) with differing y-axis scales.

```
1 bp_2.2 <- ggplot(data %>% filter(YEAR == 2022),
2                   aes(x = TRAD12, y = NUMOFFMBR, fill = TRAD12)) +
3   geom_col(position = "dodge") +
4   facet_wrap(~ TRAD6, scales = "free_y") +
5   scale_fill_viridis_d() +
6   theme_minimal() +
7   theme(
8     axis.text.x = element_text(angle = 45, hjust = 1),
9     plot.title = element_text(hjust = 0.5),
10    panel.spacing = unit(1.5, "lines"),
11    panel.border = element_rect(color = "grey20", fill = NA, linewidth =
12      0.8)
13  ) +
14  labs(
15    title = "Number of Official Members by Religion (2022)\n",
16    x = "",
17    y = "Number of Official Members\n\n",
18    fill = "Congregation"
19  )
```

Figure 5: Number of Congregation Members by Religion
Number of Official Members by Religion (2022)



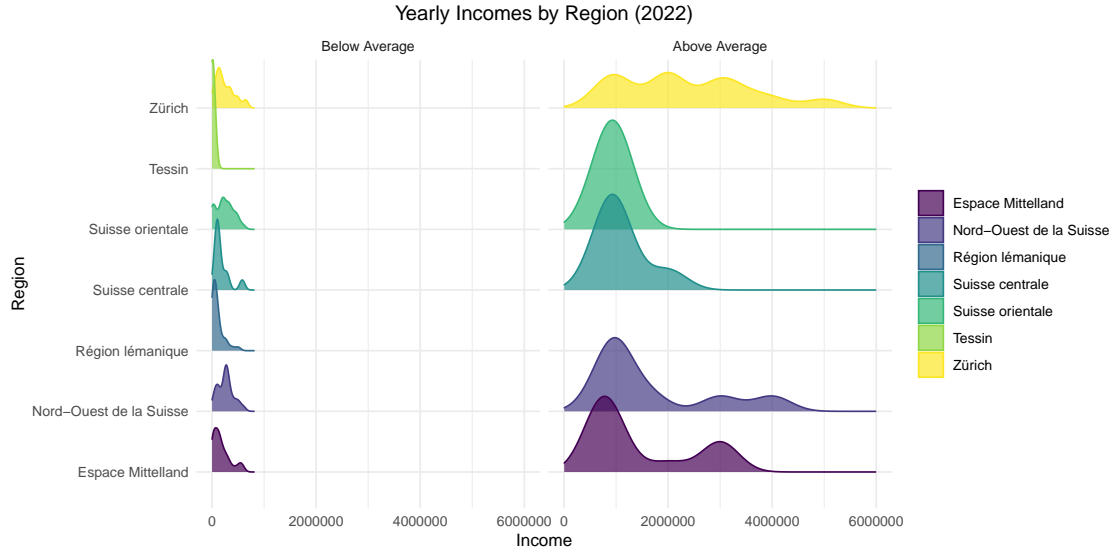
3. Display the distribution of congregations in 2022 above and below the average yearly income (AVG_INCOME) in each region using ridge plots.

```

1 ridges_together <- ggplot(data %>% filter(!is.na(AVG_INCOME), YEAR ==
2   2022),
3     aes(
4       x = INCOME,
5       y = GDREGION,
6       fill = GDREGION,
7       color = GDREGION)) +
8   geom_density_ridges(alpha = 0.7) +
9   facet_wrap(~ AVG_INCOME, labeller = as_labeller(data$AVG_INCOME)) +
10  theme(legend.title = element_blank(),
11        plot.title = element_text(hjust = 0.5)) +
12  labs(x = "Income",
13        y = "Region",
14        title = "Yearly Incomes by Region (2022)") +
15  xlim(0, 6000000) +
16  scale_fill_viridis_d() +
17  scale_color_viridis_d()

```


Figure 6: 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 7 and 8.

```
1 ridges_above <- data %>% filter(!is.na(AVG_INCOME), YEAR == 2022,
2                               AVG_INCOME == 'Above Average')
3 ridges_above_plot <- ggplot(ridges_above, aes(
4                               x = INCOME,
5                               y = GDREGION,
6                               fill = GDREGION,
7                               color = GDREGION)) +
8   geom_density_ridges(alpha = 0.7) +
9   theme(plot.title = element_text(hjust = 0.5)) +
10  labs(x = "Income",
11        y = "Region",
12        title = "Yearly Incomes by Region (2022, Above Average)") +
13  scale_fill_viridis_d() +
14  scale_color_viridis_d()
```

```
1 ridges_below <- data %>% filter(!is.na(AVG_INCOME), YEAR == 2022,
2                               AVG_INCOME == 'Below Average')
3 ridges_below_plot <- ggplot(ridges_below, aes(
4   x = INCOME,
5   y = GDREGION,
6   fill = GDREGION,
7   color = GDREGION)) +
8   geom_density_ridges(alpha = 0.7) +
9   theme(plot.title = element_text(hjust = 0.5)) +
10  labs(x = "Income",
11        y = "Region",
12        title = "Yearly Incomes by Region (2022, Below Average)") +
```

```

13 scale_fill_viridis_d() +
14 scale_color_viridis_d()

```

Figure 7: Congregations Above Average Yearly Income
Yearly Incomes by Region (2022, Above Average)

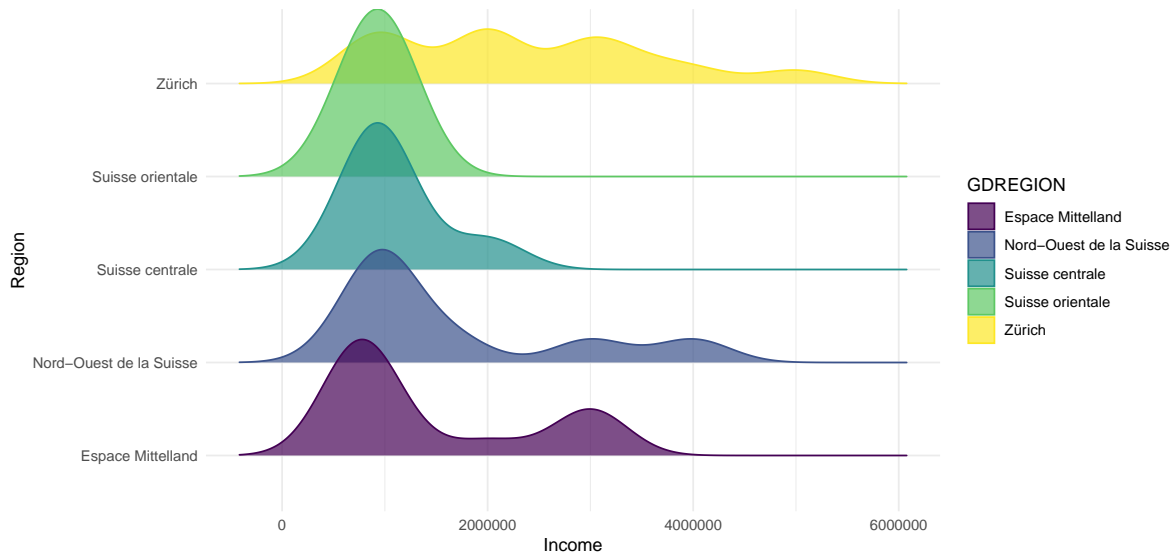
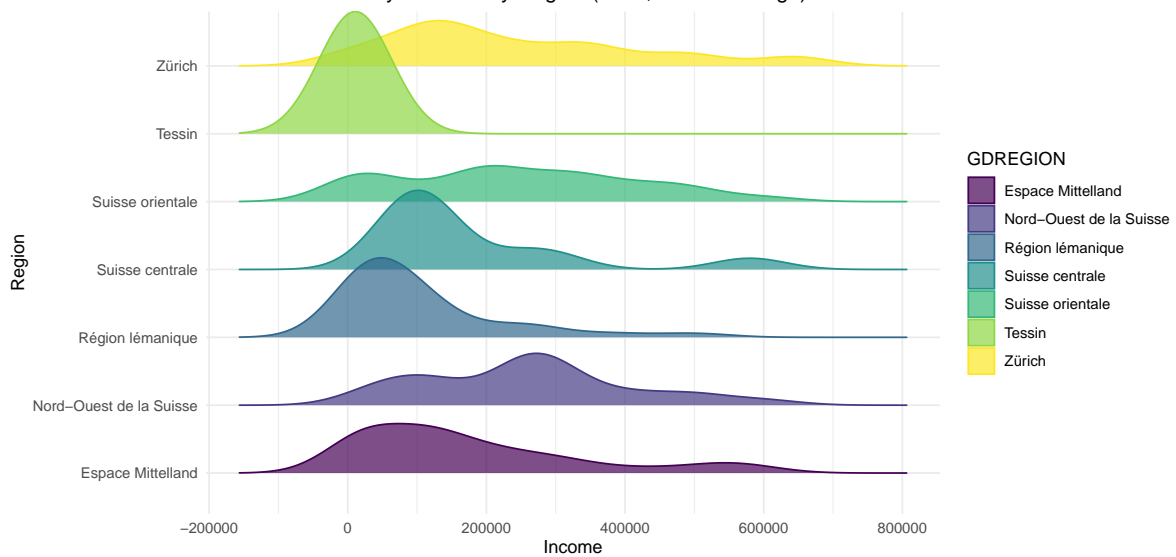
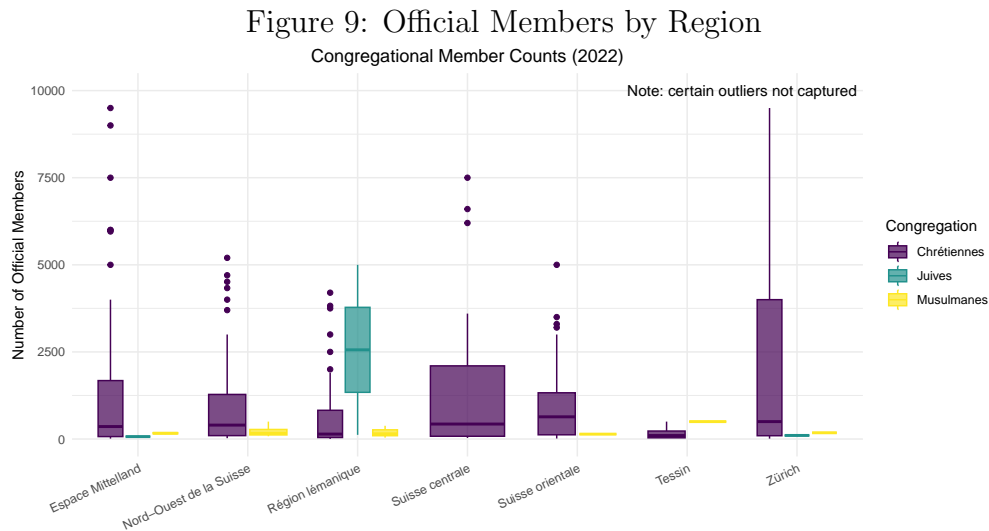


Figure 8: Congregations Below Average Yearly Income
Yearly Incomes by Region (2022, Below Average)



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

```
1 box_plot <- ggplot(data, aes(x = GDREGION, y = NUMOFFMBR, fill = TRAD6,
2   color = TRAD6)) +
3   geom_boxplot() +
4   scale_fill_viridis_d(alpha = 0.7) +
5   theme(axis.title.x = element_blank(),
6     plot.title = element_text(hjust = 0.5)) +
7   labs(
8     title = "Congregational Member Counts (2022)",
9     y = "Number of Official Members",
10    fill = "Congregation"
11  ) +
12  scale_color_manual(name = "Congregation",
13    labels = c("Chretiennes", "Juives", "Musulmanes"),
14    values = c("#440154", "#21918c", "#fde725")) +
15  ylim(0, 10000) +
```



Again, given the drastic difference in scales among religious affiliation membership, I decided to wrap the plot according to affiliation. See Figure 10.

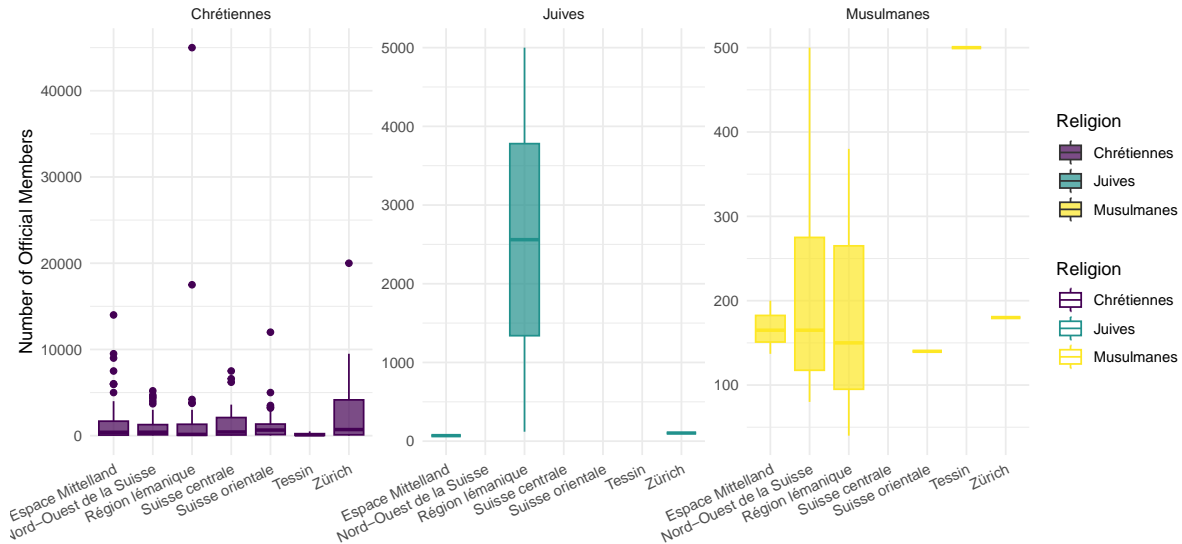
```
1 box_plot_wrapped <- ggplot(data, aes(x = GDREGION, y = NUMOFFMBR, fill =
2   TRAD6, color = TRAD6)) +
3   geom_boxplot() +
4   scale_fill_viridis_d(alpha = 0.7) +
5   theme(axis.title.x = element_blank(),
6     plot.title = element_text(hjust = 0.5)) +
7   facet_wrap(~TRAD6, scales = "free_y") +
8   scale_color_manual(name = "Religion",
9     labels = c("Chretiennes", "Juives", "Musulmanes"),
10    values = c("#440154", "#21918c", "#fde725")) +
11  labs(
12    title = "Congregational Member Counts (2022)",
```

```

12   y = "Number of Official Members",
13   fill = "Religion"
14 ) +
15 theme(axis.text.x = element_text(angle = 25, hjust = 1))

```

Figure 10: Regional Membership by Religious Affiliation
Congregational Member Counts (2022)



Finally, Professor Ziegler suggested in the updated form of the problem set that we wrap by region. Figure 11 shows what I did for that.

```

1 box_plot_region <- ggplot(data, aes(x = TRAD6, y = NUMOFFMBR, fill =
2   TRAD6, color = TRAD6)) +
3   geom_boxplot() +
4   scale_fill_viridis_d(alpha = 0.7) +
5   theme(axis.title.x = element_blank(),
6         axis.text.x = element_blank(),
7         plot.title = element_text(hjust = 0.5)) +
8   facet_wrap(~GDREGION, scales = "free_y", strip.position = "bottom") +
9   scale_color_manual(name = "Religion",
10                     labels = c("Chretiennes", "Juives", "Musulmanes"),
11                     values = c("#440154", "#21918c", "#fde725")) +
12   labs(
13     title = "Congregational Member Counts (2022)",
14     y = "Number of Official Members",
15     fill = "Religion"
16   )

```

Figure 11: Regional Membership by Religious Affiliation

