Problem Set 2

Data Visualisation for Social Scientists

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`)

Load in data and clean itby selecting and making data structure clear with factors.

```
# DM 1: Load Data #
NCSS <- read_csv("~/GitHub/DataViz_2026/datasets/NCSS_v1.csv")
NCSS <- NCSS %>%
  select(CASEID, YEAR, GDREGION, NUMOFFMBR, TRAD6, TRAD12,INCOME) %>% #
    select certain vars
  mutate(
    GDREGION = as.factor(GDREGION), #turn character into factors
    TRAD6 = as.factor(TRAD6),
    TRAD12 = as.factor(TRAD12),
    YEAR = as.integer(YEAR)
    ) %>%
  rename( #rename for preferred names
    'id' = 'CASEID',
    'year' = 'YEAR',
    'region' = 'GDREGION',
    'n_members' = 'NUMOFFMBR',
    'trad6' = 'TRAD6',
    'trad12' = 'TRAD12',
    'income' = 'INCOME'
  )
```

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

```
# DM 2:  Filter by religious classifications
unique(NCSS$trad6) #to get exact spellings and see types
NCSS <- NCSS %>% #create
  filter(trad6 %in% c('Chrétiennes', 'Juives', 'Musulmanes')) #filter out
    specific categories into new dataset
```

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. Create a new variable that creates category summaries based on grouping the dataframe by greater religious classification and year.

```
# DM 3: Compute Number of congregations, mean and median income by
    religion by year
n_congr_year <- NCSS %>% #trad 6 all religions
  group_by(trad6, year) %>% #for each year within each
  summarise(
    total_congregations = n(),
    mean_income = mean(income, na.rm = TRUE),
    median_income = median(income, na.rm = TRUE),
    .groups = "drop"
  )

```

```r
xtable(n_congr_year) #print table
```

	trad6	year	total_congregations	mean_income	median_income
1	Chrétiennes	2009	802	539942.35	200000.00
2	Chrétiennes	2022	1172	474600.50	201000.00
3	Juives	2009	18	330908.73	200000.00
4	Juives	2022	13	2332500.00	115000.00
5	Musulmanes	2009	64	62238.16	25000.00
6	Musulmanes	2022	42	77941.18	42500.00

Can see from this that clearly Christian is the biggest religious classification making up the majority of congregations, followed by Muslim and Jewish congregations (not the same as members). The number of Christian congregations grew from 2009 to 2022 (either due to sampling or real growth) whilst there were less reported congregations in both Jewish and Muslim communities. Considering income, mean is higher than the median for all the religions in both year, indicating a positive skew, with a few congregations with exceptionally high incomes compared to the majority.

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.

   The idea here to create a new column for each congregation that is a binary indicator if they are above or below the average income for that year. By grouping the data set by year can use mutate to first create an average income value for each year added as a column before creating the binary variable. This uses casewhen to for each congregation check the income level against the year average and have a value of 1 or 0 depending on whether it is greater or lower. Printing the year average we can see that for 2009 rows the comparison is against 507530 Euros compared to 475588 Euros in 2022.

```r
# DM 4:  Create binary categorical variable avg_income
NCSS <- NCSS %>%
  group_by(year) %>%
  mutate(
    year_avg_inc = mean(income, na.rm = TRUE), #find mean once each group
    avg_income = case_when(
        income >= year_avg_inc ~ 1L, #when income greater than mean, code as 1
        income < year_avg_inc  ~ 0L #less than 0
        )
  ) %>%
  ungroup() #remove the groupings
#Print the average incomes of each year
aggregate(year_avg_inc ~ year, data = NCSS, FUN = mean)
```

```
  year year_avg_inc
1 2009     507530.1
2 2022     475588.1
```

Data Visualization

Create base theme to start with for visualisations so that thematic elements can be more consistent.

```
theme_ellen <- function(...) {
  theme_minimal(base_size = 12, base_family = "serif") +
    theme(
      #Backgrounds
      # Grid
      panel.background = element_rect(fill = "white", colour = NA),
      panel.border = element_rect(fill = NA, colour = "grey20"),
      panel.grid = element_line(colour = "grey92"),
      panel.grid.minor = element_line(size = rel(0.5)),

      #Axis
      axis.title = element_text(family = "serif", size = 12), #title font
      axis.line.x = element_line(size = 0.4), #axis lines
      axis.line.y = element_line(size = 0.4),
      axis.text = element_text(colour = "black", size = 10), #axis text change
      axis.ticks.x = element_line(colour = "black", size = 0.75),

      #Legend
      legend.title = element_text(family = 'serif', size = 12, colour = 'black
    '), #legend
      legend.text =  element_text(family = 'serif', size = 10, colour = 'black
    '),# font change
      legend.key.height = unit(1, "lines"),#spacing
      legend.key.width = unit(1, "lines"),
      legend.spacing.y = unit(1, "lines"),
      legend.background = element_rect(
        fill = "white",
        colour = "grey30", #make box aound legend to align it
        size = 0.5,
        linetype = "solid")
    )
}
```

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`.

   First step is to check income. Can see there are NA's and some outliers with very high income levels compared to the rest. As we have already created a binary variable for average income between 0 and 1 the mean of the variable is the same as the proportion above the average price (i.e. a mean of 0.25 indicates 25% of people are above average and 75% are below average). Therefore we need to create a proportion statistic for each religious classification (trad12) for each year of how many congregations are above and below the yearly average.

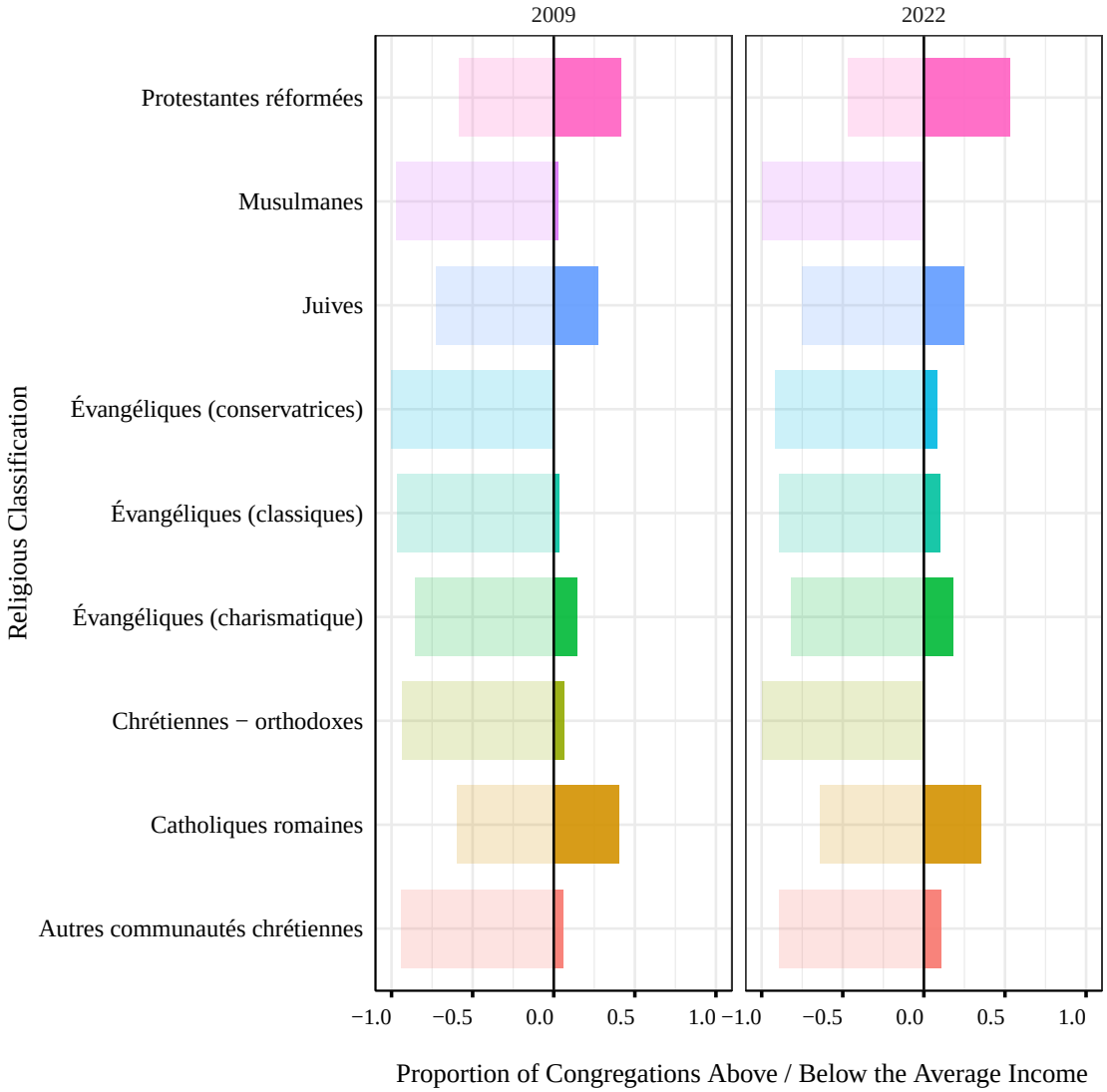
```
# DV 1: Bar Plot Av income Proportion by Trad12
summary(NCSS$income) # lots of NA's and skewed by some very large incomes
    (mean and quartiles )


#Create proportion stat for each religious classification for each year
avg_inc_trad12<- NCSS %>%
  drop_na(income) %>% #remove na values
  group_by(year, trad12) %>% #group by year then by specific category
  summarise(
    prop_abv_avg = mean(avg_income, na.rm = TRUE), #proportion as mean of
    avg income
    prop_blw_avg = 1 - prop_abv_avg, #as range is 1 the remainder from 1
    is the prop below
    .group = "drop"
  )
```

   To visualise the proportion above and below for each group and how it changes over the two years can use a proportion diverging graph. We can visualise the proportion of the congregation with each bar being 1 unit long and the position on the x axis indicating the proportion above/below average. Clearly the majority of groups have a higher proportion of the congregations below average, with Protestant and Roman Catholic congregations having the highest proportion above average both years. This is likely due to the skewed income distribution of congregations seen earlier that increases the average value away from where the majority of incomes lie. Comparing the subplots we can see that the overall distribution of proportions doesn't change drastically from 2009 to 2022, but looking at specific groups we can determine differences. For example with the Catholic orthodox congregations we can see the bar shifts lefts in 2022 from 2009, indicating a reduction in the number of congregations with an above average income. The opposite occurred for Protestants where there were more congregations in 2022 with above average income than in 2009.

```
pdf("PS02_DV1.pdf")
ggplot(avg_inc_trad12, aes(x = trad12))+ #plotting classification on
    bottom
```

```
  geom_col(aes(y = prop_abv_avg, fill = trad12, alpha = "above"), width =
    0.75)+ #bar above
  geom_col(aes(y = -prop_blw_avg, fill = trad12, alpha = "below"), width
   = 0.75)+ # below axis
  coord_flip() + #reverse x and y facts from above to see time
   progression
  facet_wrap(~ year, ncol = 2) + #group by year
  scale_y_continuous(limits = c(-1, 1)) + #clear limits
  scale_alpha_manual(values = c("above" = 0.9, "below" = 0.2),
                     guide = "none") + #transparency different for above
   and below
  geom_hline(yintercept = 0, color = "black") + #strengthen the line of
   the "average"
  guides(fill = "none") + #remove excess legend
  labs(y = "\n Proportion of Congregations Above / Below the Average
   Income",
       x = "Religious Classification") +
  theme_ellen() +
  theme(
    axis.text.x = element_text(hjust = 1, size = 10),
    axis.text.y = element_text(size = 11),
  )
dev.off()
```

Figure 1: Proportion of Congregations Above and Below the Average Income by Religious Classification and 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"`.

   Group the data set into only 2022 and remove congregations with NA member numbers.

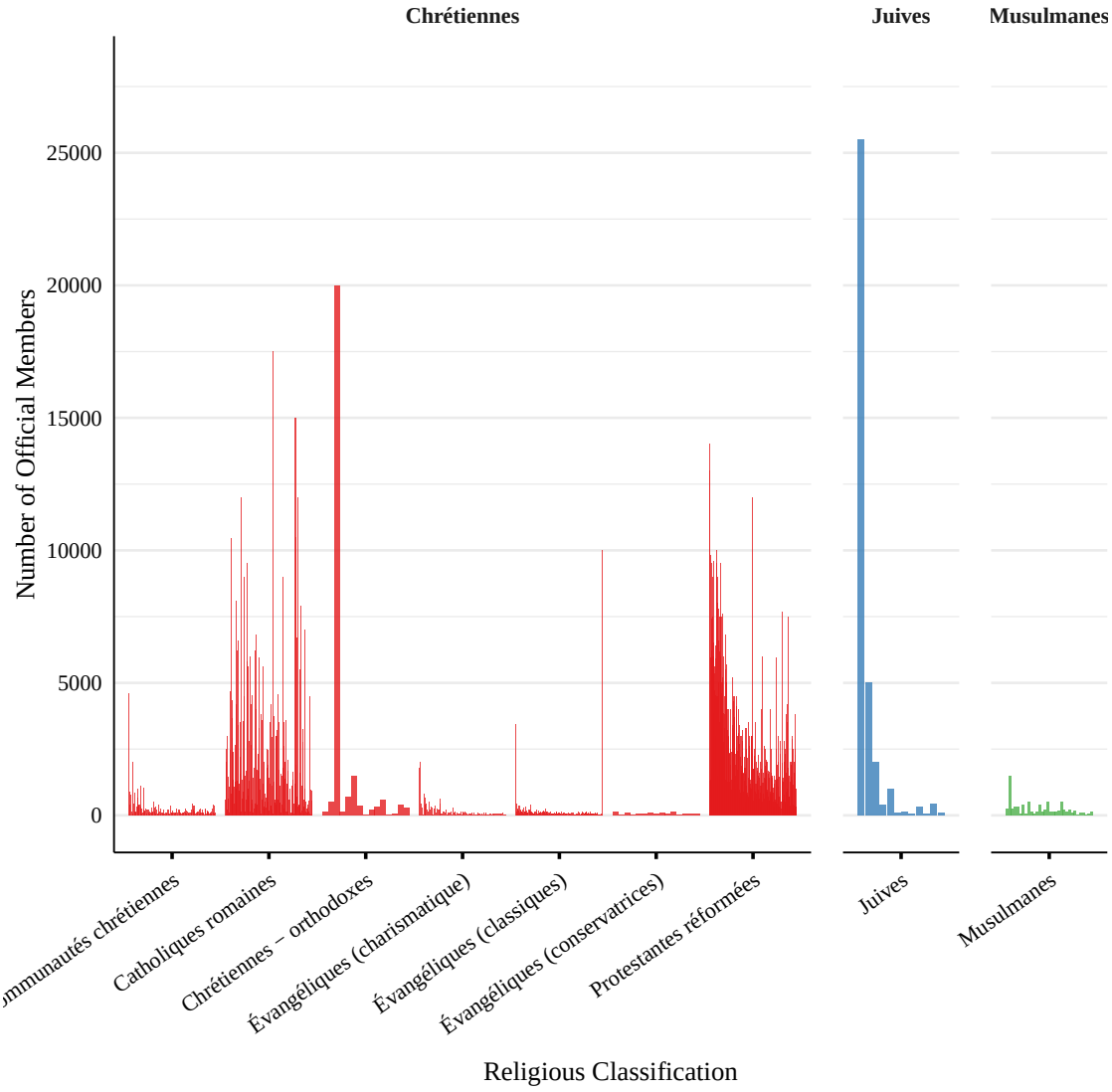
```r
# DV 2: Histogram # 161
summary(NCSS$n_members) #can see NA's

NCSS_groups <- NCSS %>%
  filter(!is.na(n_members), year == "2022") #filter na and for the
    specific year
```

   Make a histogram of the number of members separated by religious category. Using dodge2 as position of geom col bars can see histogram like thin bars of the congregation variance within the specific religious group. By faceting by larger religious categories we are able to distinguish Christian groups from Jewish and Islamic congregations more broadly. By making the scales free, only the relevant groups are plotted on the x-axis for each facet and by editing the appearance it looks like one separated bar graph rather than completely separate plots (allowing only one y axis scale and letting the Christian category subgroups appear more clear). Whilst the scale of the y-axis had to be reduced (otherwise unreadable the rest of the bars) which does eliminate some of the outliers the general pattern and trend of distribution can be seen as well as the scale of the members.

```r
pdf("PS02_DV2.pdf")
ggplot(NCSS_groups,
       aes(x = trad12, y = n_members, fill = trad6)) + #colour by trad 6
  geom_col(position= position_dodge2(width= 0.5), alpha = 0.8) + #dodge2
    for histogram look
  facet_grid(. ~ trad6, scales = "free_x", space = "free_x") + #free x
    scales makes facets only have cateogories in it
  scale_y_continuous(breaks = seq(0,25000,5000),limits = c(0,28000)) + #
    limit outliers out to see distributions
  guides(fill = "none") + #looked cluttered and groups easily read
  labs(y = "Number of Official Members",
       x = "Religious Classification") +
  theme_ellen() +
  theme(
    axis.text.x = element_text(size = 10, angle = 35, hjust = 1), #tilt
    text for legibility
    panel.background = element_blank(),
    panel.border = element_blank(), #remove box
    panel.spacing = unit(0, "lines") ,#remove visual seperation
```

```
    panel.grid.major.x = element_blank(),#keep only y so looks continuous
    panel.grid.minor.x = element_blank(),
    panel.spacing.x = unit(1, "lines"), #add space between facet subplots
    strip.text = element_text(size = 10, face = "bold") #facet title font
  ) +
  scale_fill_brewer(palette = "Set1") #colour
dev.off()
```

Figure 2: 2022 Number of Offical Members in Swiss Congregations by Religious Classification



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

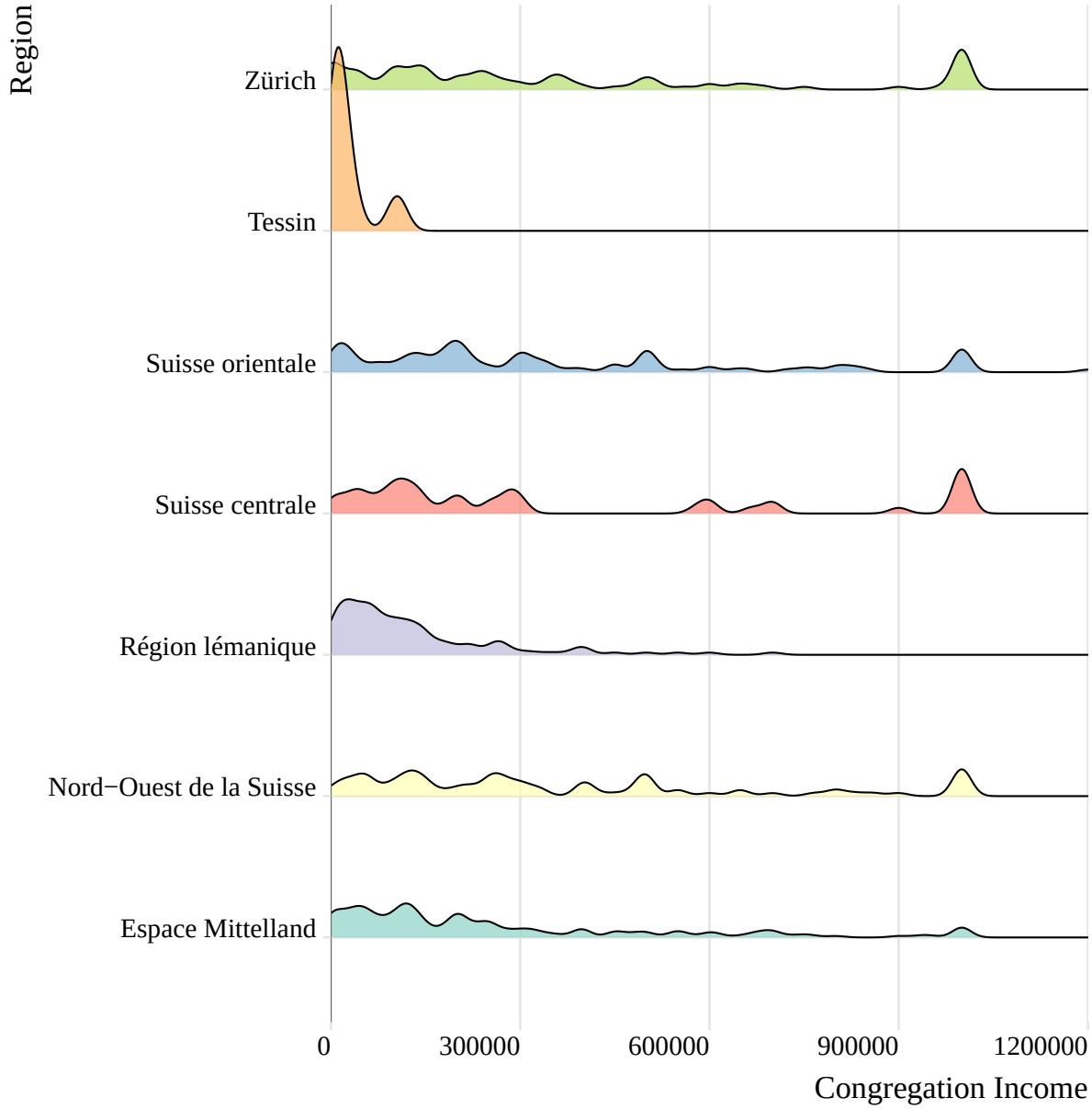
To use ridge plots to demonstrate the differences in income distribution between regions in 2022 first need to filter by year and remove missing values, as well as reduce outliers by the very top quartile, as they were lessening the legibility of the graph and trends.

```
# DV3:  Yearly income distribution
summary(NCSS$region) #one NA region
summary(NCSS$income) #clearly large outlier

NCSS_region <- NCSS %>%
  filter(!is.na(region), !is.na(income), year == "2022") %>% #filter out
   NA's and specify year
  filter(income < quantile(income, 0.95)) #remove extreme values
```

Using GGridges to create the ridgeplot, specifying the y axis as region and the x axis as income we can see the density plot of each region. By setting the bandwidth to 1500 the aim was to see the general trend of the distribution but not so general as to overlook irregularities or peaks in the distribution. The range of the income was decided to include the peak at 1100000, but zoomed in enough to clearly see the spread. Tessin stands out at having a high density of low income congregations compared to the other regions.

```
pdf("PS02_DV3.pdf")
ggplot(NCSS_region, aes(x= income , y = region , fill= region)) +
  geom_density_ridges(alpha = 0.7, #transparency
                      bandwidth = 15000, #smoothness of the plot
                      linewidth = 0.4, #thick line
                      scale = 1.3) + #scale height of factors
  scale_x_continuous(limits = c(0, 1200000), expand = c(0,0))+ #limit x
   axis for visbilty
  geom_vline(xintercept = 0, colour = "grey40", size = 0.3)+ #strengthen
   y-axis
  guides(fill = "none")+ #remove redundant legend
  labs(y = "Region",  x = "Congregation Income", fill = "Region") +
  theme_ridges() + #try theme meant specifically for ridges
  scale_fill_brewer(palette="Set3") +
  theme(
    text = element_text(family = "serif"), #change back font
    axis.text.y = element_text(size = 12),
    axis.text.x = element_text(size = 12, angle = 0, hjust = 1),
    legend.position = "none"
  )
dev.off()
```

Figure 3: 2022 Regional distribution of congregation income in Switzerland



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`. Again filtering for 2022 and removing NA values.

```
# DV 4: Boxplot Members 2022 by classification and region
NCSS_22 <- NCSS %>%
  filter(year == "2022", !is.na(region), !is.na(n_members)) #2022 only
   and rmv missing

# Check outliers
summary(NCSS_22$n_members) #check outliers may distort
```

Boxplot of the number of members by classification and region, using group as x-axis category and regional subplots within them. The most difficult aspect to consider was the scale of the y axis due to the number of extreme values, the boxplots were minimised and it became very difficult to interpret. The decision to exclude these values but at 10000, where outliers are highlighted (giving indication of same trend) but allowing more the box details to be evident. We can see that Christian congregations have generally more members in each congregation, with greater variability of size including many outliers. Muslim congregations report a very low number of members in all regions whilst the there are high member Jewish congregations condensed within Zurich and the Region Lemanique regions.

```
pdf("PS02_DV4.pdf")
ggplot(NCSS_22, aes(x = trad6, y = n_members, fill = trad6)) + #use
   filtered set,
  geom_boxplot(outlier.shape = 1, outlier.size = 1, alpha = 0.75) + #
   specify outlier shape/size
  facet_wrap(~ region, ncol = 4) + #4 columns (7 doesn't divide nicely)
  coord_cartesian(ylim = c(0, 15000)) + #limit the scale to increase box
   readability
  labs(y = "Number of Congregation Members",
       x = "Religious Classification",
       fill = "Religion") +
  theme_ellen() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 7), #angle
   to read without overlap
    legend.position = "top", #at top to give as much width as possibe
    strip.text = element_text(size = 8, face = "bold")
  ) +
  scale_fill_brewer(palette="Set1") #colour scale
dev.off()
```

Figure 4: 2022 Number of Congregation Members by Religion and Region