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`)

Answer: I loaded the dataset using `read_csv` and selected the required variables using the `select()` function as instructed.

```
# Load the NCSS .csv file
ncss_data <- read_csv("NCSS_v1.csv")

# Use select() to keep specific variables
df <- ncss_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.

   Answer: I filtered the dataframe to include only the specified religious traditions using the `filter()` function. I used Unicode escape sequences to handle special characters safely.

```
# Filter for Christian, Jewish, and Muslim congregations
# Note: Using unicode \u00E9 for 'e' with accent to prevent LaTeX errors
df_filtered <- df %>%
  filter(TRAD6 %in% c("Chr\u00E9tiennes", "Juives", "Musulmanes")) ##
   LaTeX encoding fix
```

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.

   Answer: I grouped the data by year and religious classification to calculate the statistics.

   - **Number of Congregations:** In 2022, there were 1,172 Christian, 13 Jewish, and 42 Muslim congregations.
   - **Income (2022):** Christian congregations had a median income of 201,000, while Muslim congregations had a lower median of 42,500. Jewish congregations showed a very high mean (approx. 2.3M) compared to their median (115,000), indicating significant outliers.

```
# Compute number of congregations by religious classification and year
cong_count <- df_filtered %>%
  count(YEAR, TRAD6)

print(cong_count)
```

```
  YEAR     TRAD6          n
1 2009 Chrétiennes   802
2 2009 Juives         18
3 2009 Musulmanes     64
4 2022 Chrétiennes  1172
5 2022 Juives         13
6 2022 Musulmanes     42
```

```r
# Compute mean and median total income
income_stats <- df_filtered %>%
  group_by(YEAR, TRAD6) %>%
  summarise(
    mean_income = mean(INCOME, na.rm = TRUE),
    median_income = median(INCOME, na.rm = TRUE)
  )

print(income_stats)
```

```
  YEAR TRAD6       mean_income median_income
1 2009 Chrétiennes     539942.        200000
2 2009 Juives          330909.        200000
3 2009 Musulmanes       62238.         25000
4 2022 Chrétiennes     474601.        201000
5 2022 Juives         2332500         115000
6 2022 Musulmanes       77941.         42500
```

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: I calculated the yearly average income and created a binary variable `AVG_INCOME`. I assigned 1 if the income was greater than or equal to the yearly average, and 0 otherwise, converting it to a factor for visualization.

```r
# Create categorical variable AVG_INCOME
# First, calculate average income per year
df_aug <- df_filtered %>%
  group_by(YEAR) %>%
  mutate(yearly_avg = mean(INCOME, na.rm = TRUE)) %>%
  ungroup()

# Create binary variable: 1 if >= average, 0 if < average
df_aug$AVG_INCOME <- ifelse(df_aug$INCOME >= df_aug$yearly_avg, 1, 0)

# Convert to factor for better plotting
```

```r
df_aug$AVG_INCOME <- factor(df_aug$AVG_INCOME, levels = c(0, 1), labels =
    c("Below Avg", "Above/Avg"))
```

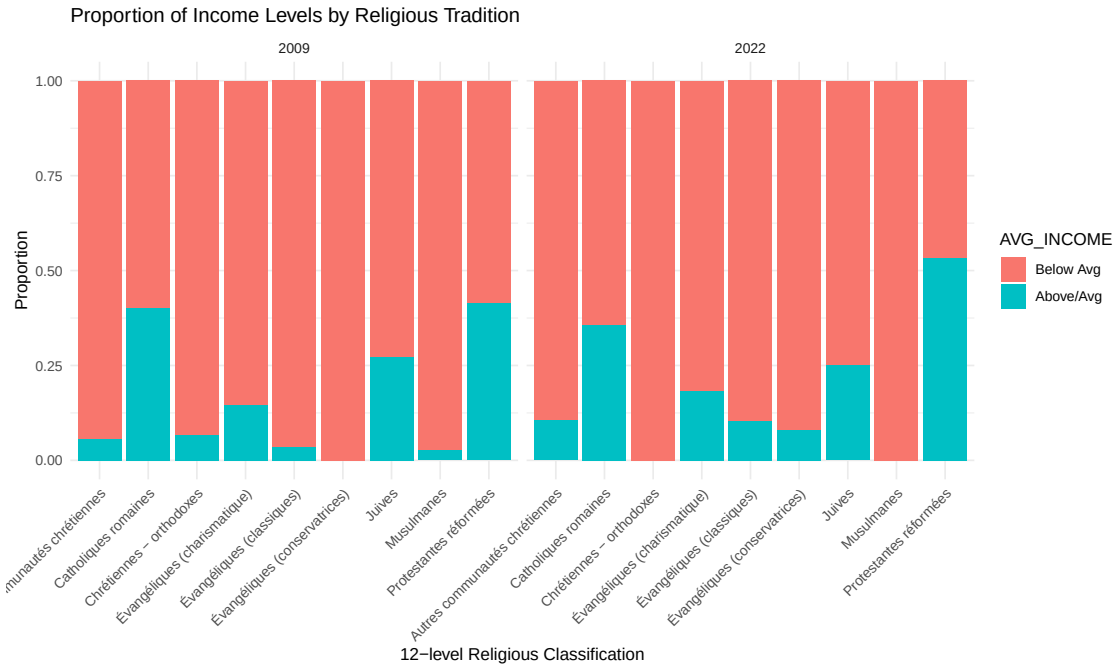
Data Visualization

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`.

   Answer: I visualized the proportion of income levels using a stacked bar chart. We can observe that across most religious traditions, the proportion of congregations with "Above/Avg" income has remained relatively stable or slightly increased between 2009 and 2022, although variations exist within specific evangelical subgroups.

```r
# Bar plot: Proportion of congregations above/below avg income
prop_income <- df_aug %>%
  filter(!is.na(AVG_INCOME)) %>%
  group_by(YEAR, TRAD12, AVG_INCOME) %>%
  summarise(N = n()) %>%
  mutate(prop = N / sum(N))

pdf("Plot1_Income_Proportion.pdf", width = 10, height = 6)
ggplot(prop_income, aes(x = TRAD12, y = prop, fill = AVG_INCOME)) +
  geom_col(position = "fill") +
  facet_wrap(~YEAR) +
  labs(title = "Proportion of Income Levels by Religious Tradition",
       x = "12-level Religious Classification", y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()
```



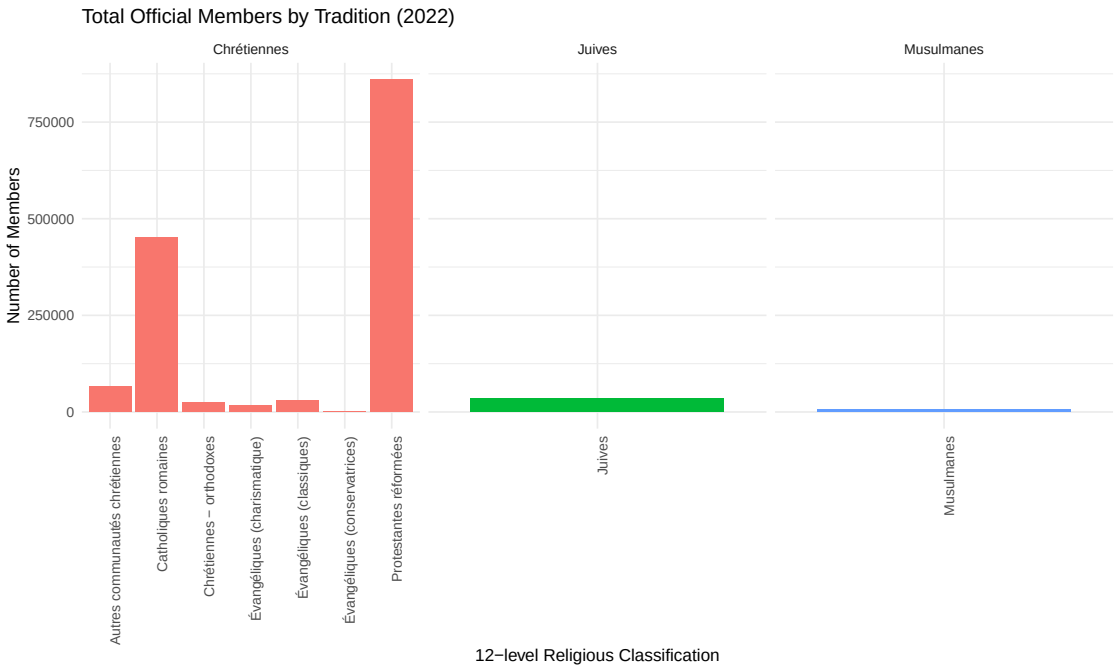
2. Make a histogram using `geom_col()` detailing the number of official members using the 12-level religious classification (`TRAD12`) distinguishing between the 6-level religious classification (`TRAD6`) in 2022. Hint: Use `facet()` for `TRAD6`, with `TRAD12` on the x-axis in addition to group/fill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

   Answer: For the year 2022, I aggregated the total number of official members. I used `geom_col()` with `position="dodge"` to display the counts, faceting by the broader `TRAD6` classification.

```r
df_2022 <- df_aug %>% filter(YEAR == 2022)

members_summary <- df_2022 %>%
  group_by(TRAD6, TRAD12) %>%
  summarise(total_members = sum(NUMOFFMBR, na.rm = TRUE))

pdf("Plot2_Members_Histogram.pdf", width = 10, height = 6)
ggplot(members_summary, aes(x = TRAD12, y = total_members, fill = TRAD6)) +
  geom_col(position = "dodge") +
  facet_wrap(~TRAD6, scales = "free_x") +
  labs(title = "Total Official Members by Tradition (2022)",
       x = "12-level Religious Classification", y = "Number of Members") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none")
```

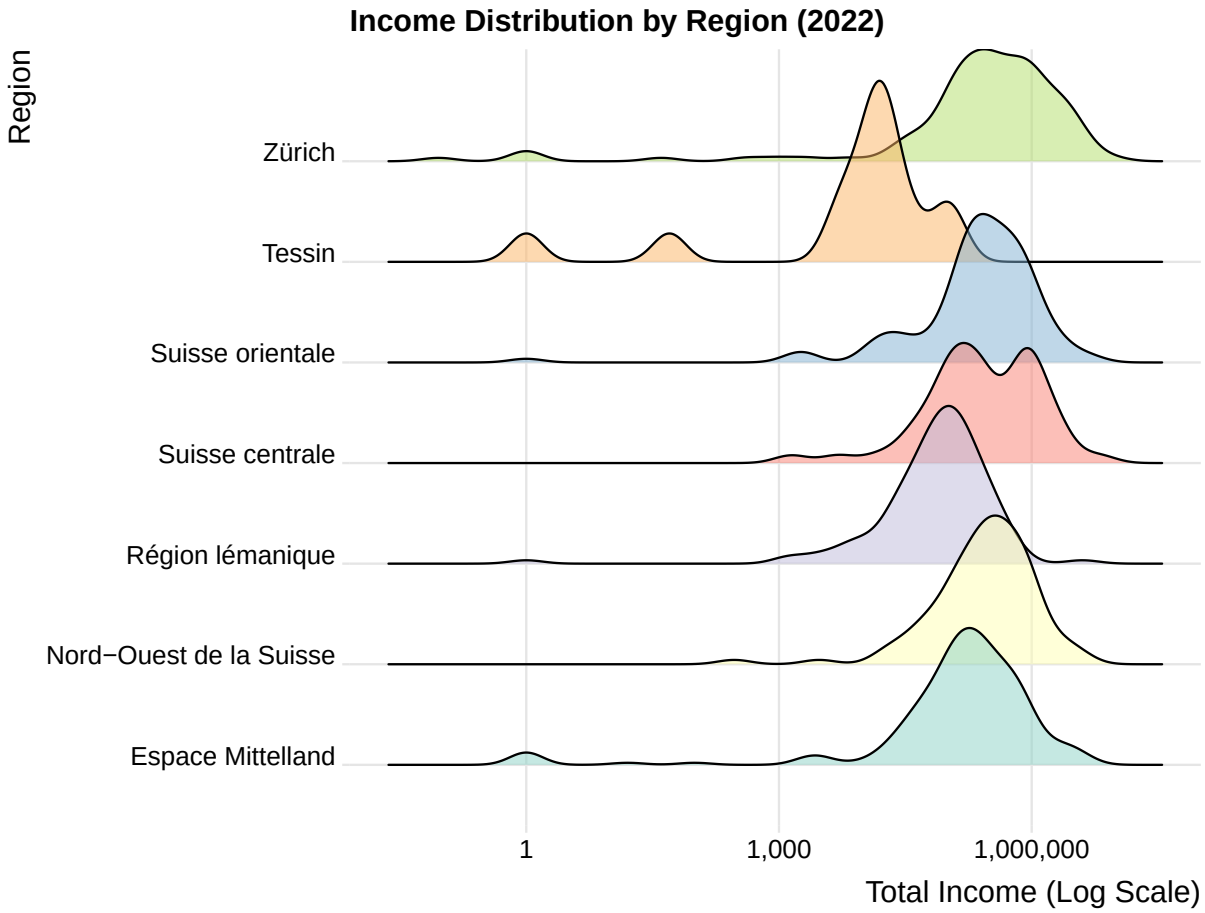


3. Display the distribution of yearly income (`INCOME`) in 2022 for congregations in each region (`GDREGION`) using ridge plots.

   Answer: I used the `ggridges` package to visualize income distribution by region. I applied a log scale (`scale_x_log10`) to the x-axis to better handle the wide range and skewness of the income data.

```
ggplot(df_2022, aes(x = INCOME, y = GDREGION, fill = GDREGION)) +
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Income Distribution by Region (2022)",
       x = "Total Income (Log Scale)", y = "Region") +
  scale_fill_brewer(palette="Set3") +
  theme_ridges() +
  theme(legend.position = "none")
dev.off()

### 4
```



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: I created boxplots to show the distribution of official members by religious classification, faceted by region. I used a log scale on the y-axis to clearer visualize the distribution across different congregation sizes.

```
  facet_wrap(~GDREGION) +
  scale_y_log10() +
  labs(title = "Official Members per Congregation by Region (2022)",
       x = "Religious Classification", y = "Number of Members (Log Scale)
    ") +
  scale_fill_brewer(palette="Pastel1") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()
```