Problem Set 1

Hanyu Li(ID:25346841)

January 27, 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 January 28, 2026. No late assignments will be accepted.

Roll Call Votes in the European Parliament

Data Manipulation

First, you need to download data from the first six elected European Parliaments on each MEP and how they voted in each recorded roll-call vote.

1. Load these datasets into your global environment:
   - `mep_info_26Jul11.xls` (MEP characteristics, EP1–EP5)
   - `rcv_ep1.txt` (EP1 roll-call votes)

2. Briefly describe (2–3 sentences each) the unit of analysis and key variables in each of these two datasets.

   Dataset 1: For the `mep_info_26Jul11` dataset, the unit of analysis is the individual Member of the European Parliament (MEP). This dataset contains the background characteristics of MEPs, covering the first to fifth parliaments (EP1-EP5). The key

variables include *MEPID* as a unique identifier, categorical variables for party affiliation (National Party based on member state and EP Group based on political ideology), and two continuous variables (*nomdim1* and *nomdim2*) representing the MEP's NOMINATE coordinates on two key ideological dimensions. Specifically, *nomdim1* measures the MEP's position on the traditional Left/Right ideological spectrum, while *nomdim2* captures their attitude towards European integration (Pro-EU vs. Eurosceptic), with both scores typically ranging from -1 to $+1$.

Dataset 2: The unit of analysis for the `rcv_ep1` dataset is the individual MEP, showcasing their specific voting behaviors across all recorded sessions in EP1. The dataset records individual decisions on specific roll-call votes. The key variables are the vote columns ($V_1 \dots V_{886}$), which are categorically coded as: 1 = Yes, 2 = No, 3 = Abstain, 4 = Present but did not vote, 0 = Absent, and 5 = Not an MEP.

3. The `rcv_ep1` data are in a wide format, with $V_1, V_2, \dots, V_n$ as separate vote columns.

   - Identify which columns are ID/metadata (*MEPID, MEPNAME, MS, NP, EPG*) and which columns are vote decisions ($V_1 \dots V_n$). Tidy the voting data such that each row/observation is a single vote for a single MEP.
     The codes are as below:

```r
# Data Manipulation Q3
library(tidyverse)
library(readxl)
library(scales)
library(ggplot2)

# Import and sort out rcv
rcv <- read_csv("rcv_ep1.txt")
rcv_clean <- rcv |>
  mutate(
    MEPID = as.character(MEPID),
    NP = as.factor(NP),
    EPG = as.factor(EPG))

# Pivot using the cleaned data
rcv_ep1 <- rcv_clean |>
  pivot_longer(
    cols = starts_with("V"),
    names_to = "Vote_item",
    values_to = "Decision")
head(rcv_ep1)
```

     Top 6 rows in the tidy table are shown as below: with the first 5 columns showing MEP information and the last 2 rows representing vote items and decisions.

```
  MEPID MEPNAME      MS    NP    EPG   Vote_item Decision
  <chr> <chr>        <chr> <fct> <fct> <chr>     <fct>
```

```
1 2   ABENS Victor L     1804  S     V1        2
2 2   ABENS Victor L     1804  S     V2        2
3 2   ABENS Victor L     1804  S     V3        4
4 2   ABENS Victor L     1804  S     V4        0
5 2   ABENS Victor L     1804  S     V5        0
6 2   ABENS Victor L     1804  S     V6        0
```

- Create a summary table of counts of decision categories (e.g. Yes/No/Abstain/Present but did not vote/Absent) across all votes.
  The codes are as below:

```
# Create a summary table
rcv_ep1 <- mutate(rcv_ep1,
                  Decision = factor(Decision,
                                    levels = c(1, 2, 3, 4, 0, 5),
                                    labels = c("Yes", "No", "
    Abstain",
                                               "Present but did
    not vote",
                                               "Absent", "Not an
    MEP")))

vote_sum <- count(rcv_ep1, Decision)
vote_sum
```

  The summary table is as follows:

```
Decision                n
<fct>                   <int>
1 Yes                   88185
2 No                    75171
3 Abstain               9577
4 Present but did not vote 109224
5 Absent                99753
6 Not an MEP            103618
```

4. Construct a new dataset that combines MEP-level information with their vote decisions from EP1 in long format (from part 3). Check for missingness.

   The codes and a preview of combined dataset are as below:

```
# Data Manipulation Q4
# Import mepinfo and set MEPID for combination
mepinfo_ep1 <- read_excel("mep_info_26Jul11.xls", sheet = "EP1") |>
  mutate(MEPID = as.character(`MEP id`)) |>
# Drop repeated varibale like Name,MS,NP,EPG
```

```
  select(MEPID, `NOM-D1`, `NOM-D2`)

# Combine the 2 datasets by MEPID
com_ep1 <- rcv_ep1 |>
  left_join(mepinfo_ep1, by = "MEPID") |>
  mutate(
    `NOM-D1` = as.numeric(`NOM-D1`),
    `NOM-D2` = as.numeric(`NOM-D2`))
head(com_ep1)
```

```
    MEPID MEPNAME      MS    NP    EPG   Vote_item Decision                 `NOM-D1` `NOM-D2`
    <chr> <chr>        <chr> <fct> <fct> <chr>     <fct>                    <dbl>    <dbl>
    1 2   ABENS Victor L     1804  S     V1        No                       -0.021    0.245
    2 2   ABENS Victor L     1804  S     V2        No                       -0.021    0.245
    3 2   ABENS Victor L     1804  S     V3        Present but did not vote -0.021    0.245
    4 2   ABENS Victor L     1804  S     V4        Absent                   -0.021    0.245
    5 2   ABENS Victor L     1804  S     V5        Absent                   -0.021    0.245
    6 2   ABENS Victor L     1804  S     V6        Absent                   -0.021    0.245
```

Then we check and report the missing values. We can find that 42528 missing values in *NOM-D1* and *NOM-D2*. Given each row shows a single vote for a single MEP, this indicates that background information is missing for $\frac{42528}{886} = 48$ specific MEPs.

```
# Check for missing values
missing_repo <- colSums(is.na(com_ep1))
missing_repo
```

```
missing_repo
MEPID  MEPNAME  MS  NP  EPG Vote_item Decision  NOM-D1  NOM-D2
0      0        0   0   0   0         0         42528   42528
```

5. Compute, for each EP group in EP1:

   - The mean rate of Yes votes (Yes over Yes+No+Abstain) across all roll calls. We can see the mean rate of Yes votes for each EP group in EP1 ranging from 41.5% to 58.1%.

```
# Data Manipulation Q5
# The mean rate of Yes votes across all roll calls
yes_rate <- com_ep1 |>
  filter(Decision %in% c("Yes", "No", "Abstain")) |>
  group_by(EPG) |>
  summarise(Yes_rate = mean(Decision == "Yes"))
yes_rate
```

```
  EPG    Yes_rate
  <fct>  <dbl>
  1 C    0.415
  2 E    0.509
```

```
3 G    0.512
4 L    0.486
5 M    0.528
6 N    0.581
7 R    0.457
8 S    0.576
```

- The mean abstention rate.
  We can see the mean rate of Abstention for each EP group in EP1 ranging from 2.15% to 26.5%.

```r
# The mean abstention rate
abs_rate <- com_ep1 |>
  filter(Decision %in% c("Yes", "No", "Abstain")) |>
  group_by(EPG) |>
  summarise(Abstention_rate = mean(Decision == "Abstain"))
abs_rate
```

```
EPG    Abstention_rate
<fct>           <dbl>
1 C             0.0752
2 E             0.0215
3 G             0.0697
4 L             0.0632
5 M             0.0800
6 N             0.0562
7 R             0.265
8 S             0.0574
```

- The mean vote preferences along the two contested dimensions (*NOM-D1* and *NOM-D2*).
  Since dataset '`rcv_ep1`' records one MEP who was not categorized to any EP Group, so here I drop this case when computing the mean vote preferences along the two contested dimensions.

```r
# The mean vote preferences along the two contested dimensions
which(rcv$EPG == "0") # one MEP don't have EP Group information
dimension_pre <- com_ep1 |>
  filter(EPG != "0") |>  # drop the case without EP Group
   category
  group_by(EPG) |>
  summarise(
    D1_mean = mean(`NOM-D1`, na.rm = TRUE),
    D2_mean = mean(`NOM-D2`, na.rm = TRUE))
dimension_pre
```

```
EPG    D1_mean D2_mean
<fct>  <dbl>   <dbl>
1 C    0.811   0.530
2 E    0.512  -0.277
3 G    0.280  -0.818
4 L    0.409  -0.324
5 M   -0.357  -0.201
6 N    0.250  -0.386
7 R   -0.586  -0.0419
8 S   -0.0980  0.261
```

Data Visualization

1. Plot the distribution of the first NOMINATE dimension by EP group, and explain any trends you see.

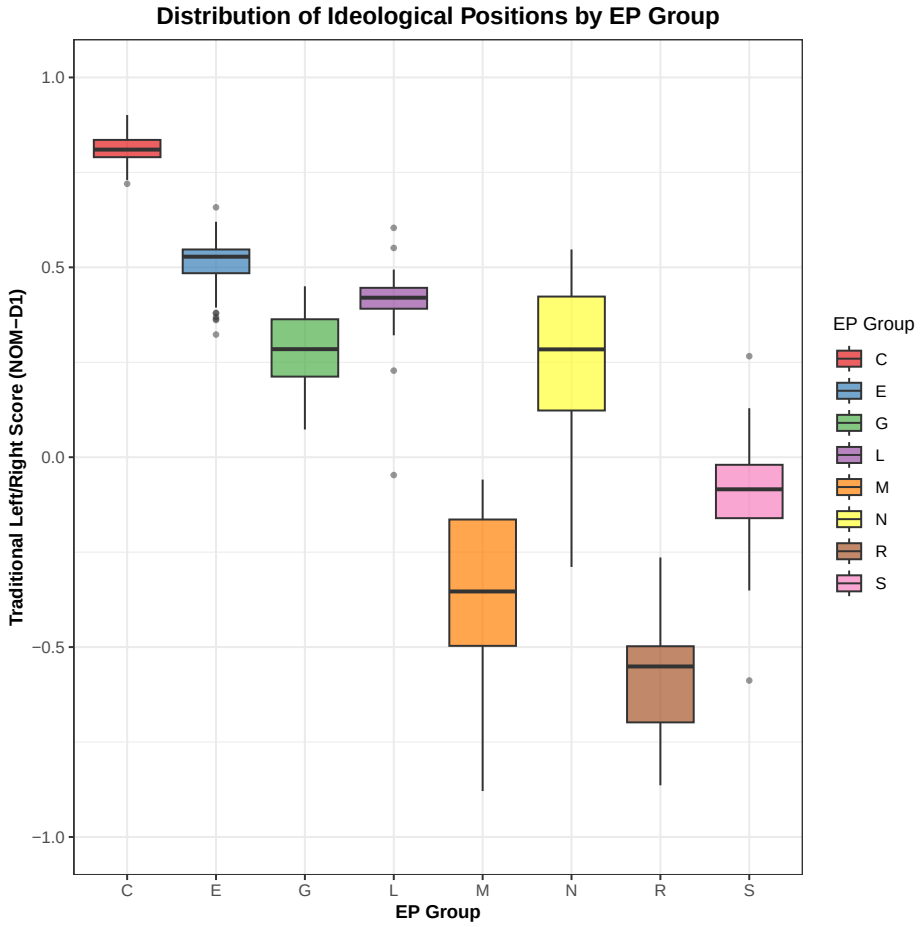
   As each MEP is observed repeatedly across votes, duplicate observations of the first NOMINATE dimension were removed prior to analysis. Since we want to explore the data distribution, here I first chose boxplot to do visualization. Moreover, as EP group categories recur across later figures, their color coding was standardized. All grouping variables used in subsequent analyses were standardized, with categorical variables treated as characters and temporal variables (e.g., year) treated as integers.

```r
# Q1.Plot the distribution of the first NOMINATE dimension by EP
   group
# Data Deduplication
data1 <- com_ep1 |>
  filter(EPG != "0") |>
  select(MEPID, EPG, `NOM-D1`, `NOM-D2`) |>
  unique() |>
  na.omit()

# Plot data1 in boxplot by EP group
pdf("plot1.pdf")
plot1 <- ggplot(data1, aes(x = EPG, y = `NOM-D1`, fill = EPG)) +
  geom_boxplot(alpha = 0.7,outlier.size = 1, outlier.alpha = 0.5) +
  scale_fill_brewer(palette = "Set1", name = "EP Group") +
  scale_y_continuous(limits = c(-1, 1)) +
  labs(
       title = "Distribution of Ideological Positions by EP Group",
       x = "EP Group",
       y = "Traditional Left/Right Score (NOM-D1)") +
  theme_bw() +
  theme(
        plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title.x = element_text(size = 10, face = "bold"),
        axis.title.y = element_text(size = 10, face = "bold"),
        legend.title = element_text(size = 10),
```

```
        legend.position = 'right')
print(plot1)
```

Figure 1: Distribution of Ideological Positions by EP Group



As shown in Figure 1, first, the data reveals a sharp ideological polarization between the EP groups along the traditional Left/Right dimension. The groups are clearly divided: Groups C, E, G, L, and N are consistently positioned on the right (scores > 0), with Group C exhibiting the most pronounced conservative stance (median ≈ 0.8). Conversely, Groups M, R, and S occupy the left-wing spectrum (scores < 0), where Group R represents the furthest left position with a median of approximately -0.5.

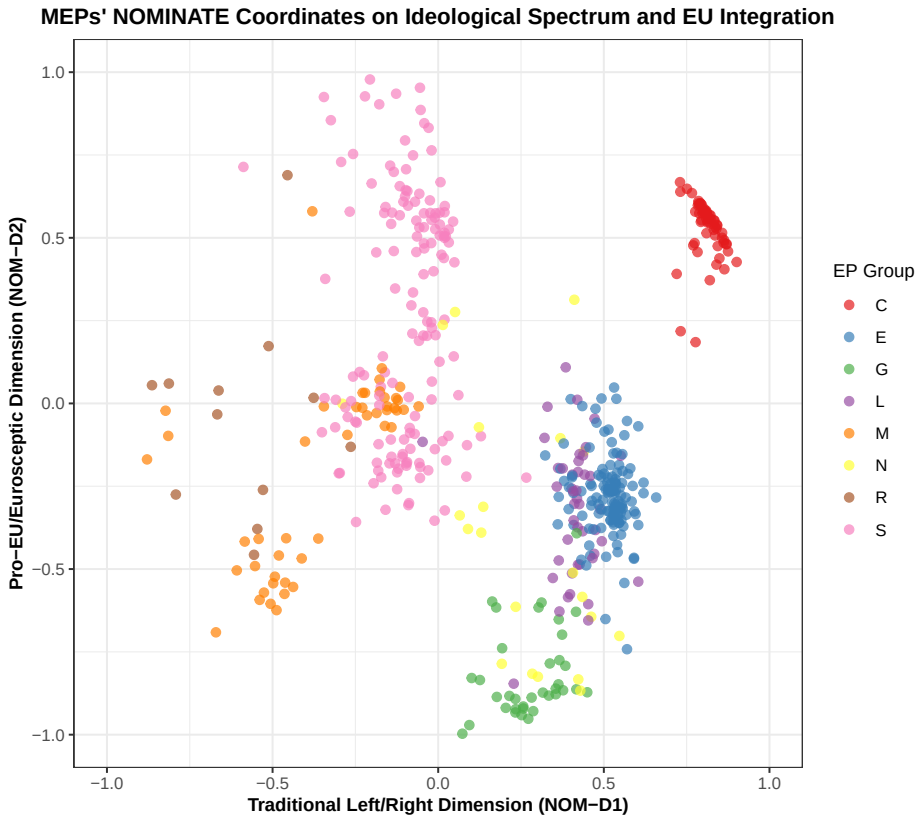
Second, the groups display significant variation in internal cohesion. The right-leaning Groups C, E, and L demonstrate high cohesion, evidenced by their compressed box plots and concentrated score distributions. In contrast, the left-leaning Group M and the right-leaning Group N show lower internal unity; their elongated box plots indicate a more dispersed distribution of preferences and greater ideological heterogeneity

within these groups.

2. Make a scatterplot of *nomdim1* (x-axis) and *nomdim2* (y-axis), with one point per MEP and color by EP group.

```
# Q2. Make a scatterplot of nomdim1 (x-axis) and nomdim2 (y-axis)
pdf("plot2.pdf")
plot2 <- ggplot(data1, aes(x = `NOM-D1`, y = `NOM-D2`, color = EPG))
   +
  geom_point(alpha = 0.7, size = 2) +
  coord_fixed(xlim = c(-1, 1), ylim = c(-1, 1)) +
  scale_color_brewer(palette = "Set1", name = "EP Group") +
  labs(
    title = "MEPs' NOMINATE Coordinates on Ideological Spectrum and
   EU Integration",
    x = "Traditional Left/Right Dimension (NOM-D1)",
    y = "Pro-EU/Eurosceptic Dimension (NOM-D2)") +
  theme_bw() +
  theme(
    plot.title = element_text(size = 12,hjust = 0.5, face = "bold"),
    axis.title.x = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10, face = "bold"),
    legend.title = element_text(size = 10),
    legend.position = 'right')
print(plot2)
```

Figure 2: MEP Ideological Alignment



We can see from Figure 2 that these two dimensions effectively separate MEPs into distinct clusters based on their EP group affiliation. Notably, Group C forms a clearly defined cluster, exhibiting a strong right-wing and pro-European integration stance.

3. Produce a boxplot of the proportion voting *Yes* by EP group to visualize cohesion.

```r
# Q3.Produce a boxplot of the proportion voting Yes by EP group
# Compute the distribution of yes proportion for each roll call
vote_cal <- com_ep1 |>
  filter(Decision %in% c("Yes", "No", "Abstain") & EPG != "0") |>
  group_by(EPG, Vote_item) |>
  summarise(yes_prop = mean(Decision == "Yes"), .groups = "drop")
head(vote_cal)

# Plot the boxplot
pdf("plot3.pdf")
plot3 <- ggplot(vote_cal, aes(x = EPG, y = yes_prop, fill = EPG)) +
```

```
  geom_boxplot(alpha = 0.7, outlier.size = 1, outlier.alpha = 0.5) +
  scale_fill_brewer(palette = "Set1", name = "EP Group") +
  scale_y_continuous(limits = c(0, 1),
                     labels = percent_format()) +
  labs(
    title = "Proportion of Voting Yes by EP Group (per Roll Call)",
    x = "EP Group",
    y = "Proportion of Yes Votes(EP1) ") +
  theme_bw() +
  theme(
    plot.title = element_text(size = 14,hjust = 0.5, face = "bold"),
    axis.title.x = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10, face = "bold"),
    legend.title = element_text(size = 10),
    legend.position = 'right')
print(plot3)
```

Figure 3: Proportion of Yes Votes by EP Group

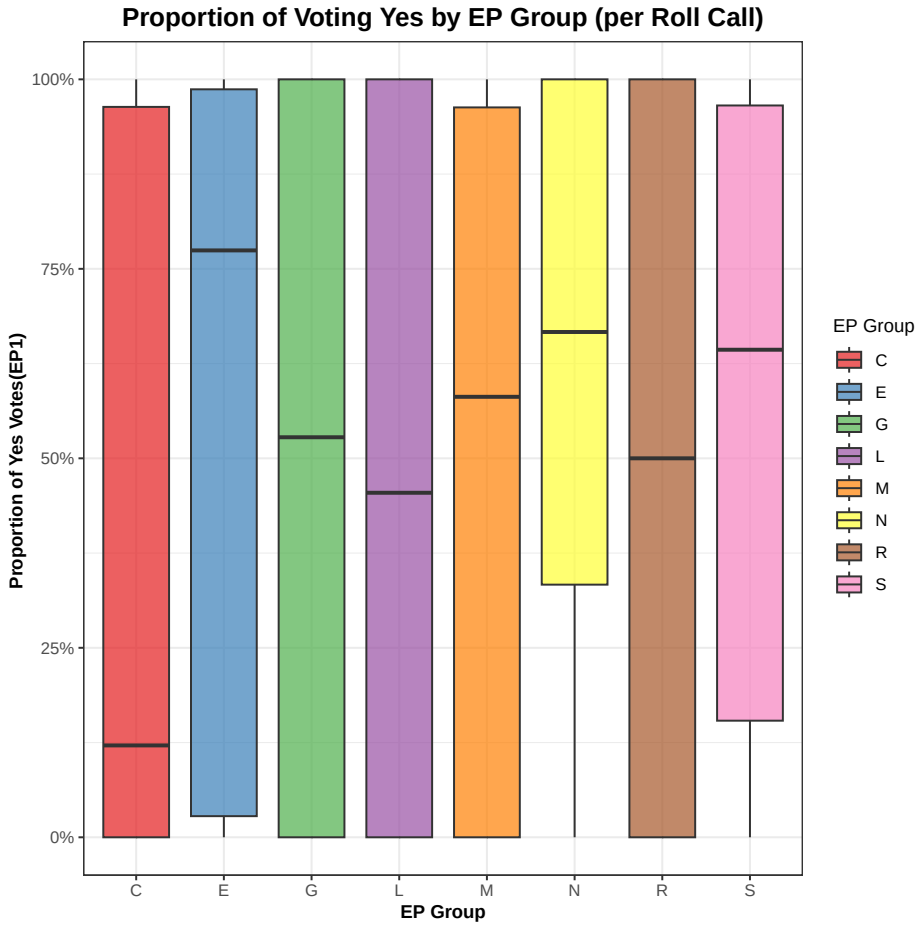


Figure 3 reveals a broad distribution of Yes vote shares across roll calls for all EP

groups, indicating significant variability in voting outcomes. Specifically, in terms of median values, Group E displays the highest rate of affirmative votes, whereas Group C exhibits the lowest.

4. Display the proportion voting *Yes* per year by national party using a bar plot.

   Since the combined data set lacks year information for each call roll, I integrated dataset `vote_info_Jun2010.xls` including the number and date of the vote to get complete data for plotting later.

   Code lines for integration and the final data set are as below:

```
# Q4.Display the proportion voting Yes per year by national party
# Merge the year information for each roll call to combined dataset
library(lubridate)
library(janitor)
vote_info <- read_xls("vote_info_Jun2010.xls", sheet = "EP1")
head(vote_info)
vote_dates <- vote_info |>
  mutate(
  Vote_item = paste0("V", `Vote No. in RCV_EP1 file`),
  Year = year(excel_numeric_to_date(as.numeric(Date)))
  ) |>
  select(Vote_item, Year)

com_ep1 <- com_ep1 |>
  left_join(vote_dates, by = "Vote_item")
head(com_ep1)
```

```
    > head(com_ep1)
    # A tibble: 6 x 10
    MEPID MEPNAME      MS    NP    EPG   Vote_item Decision             `NOM-D1` `NOM-D2` Year
    <chr> <chr>        <chr> <fct> <fct> <chr>     <fct>                <dbl>    <dbl>    <int>
    1 2   ABENS Victor L     1804  S     V1        No                   -0.021    0.245    1979
    2 2   ABENS Victor L     1804  S     V2        No                   -0.021    0.245    1979
    3 2   ABENS Victor L     1804  S     V3        Present but did not vote -0.021 0.245    1979
    4 2   ABENS Victor L     1804  S     V4        Absent               -0.021    0.245    1979
    5 2   ABENS Victor L     1804  S     V5        Absent               -0.021    0.245    1979
    6 2   ABENS Victor L     1804  S     V6        Absent               -0.021    0.245    1979
```

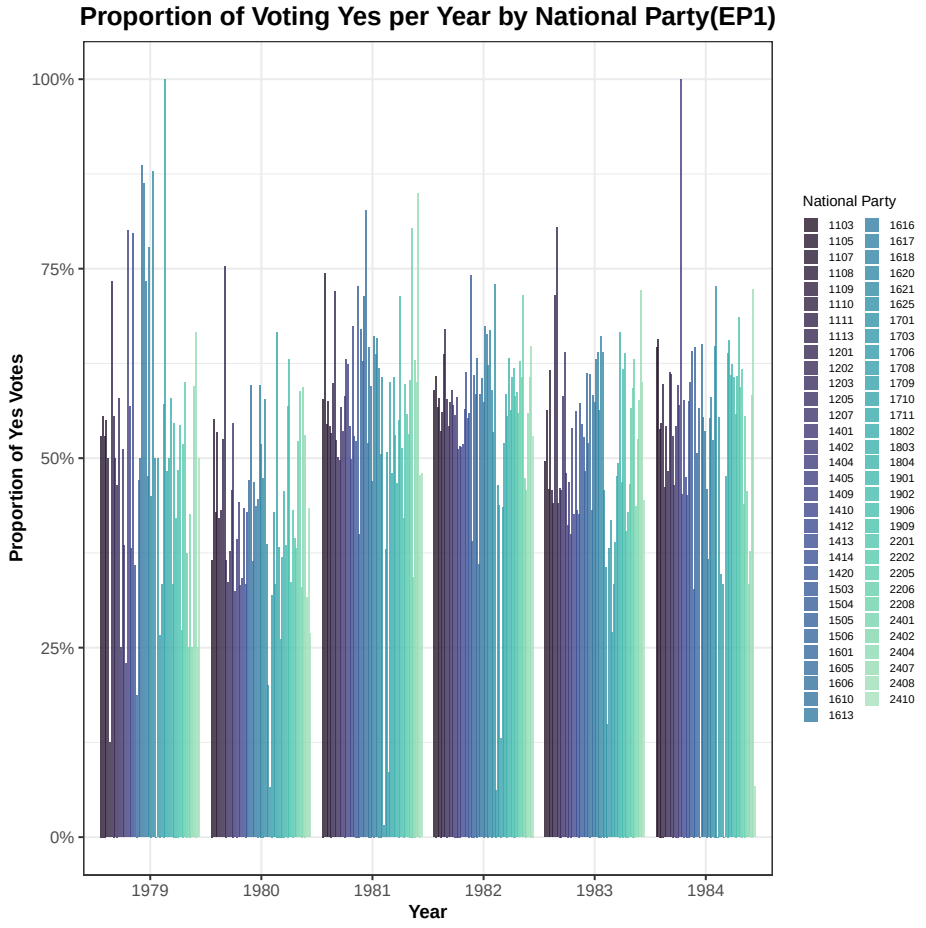
   Codes for data wrangling and plotting are as follow:

```
# Compute the proportion voting Yes per year by national party
# Q4. Display the proportion voting Yes per year by national party
np_votes <- com_ep1 |>
  filter(Decision %in% c("Yes", "No", "Abstain") & !is.na(Year)) |>
  mutate(Year = as.integer(Year)) |>
  group_by(NP, Year) |>
  summarise(np_yes = mean(Decision == "Yes"), .groups = "drop")

# Plot the bar plot
pdf("plot4.pdf")
plot4 <- ggplot(np_votes, aes(x = as.factor(Year), y = np_yes, fill =
    NP)) +
```

```
  geom_col(position = "dodge", alpha = 0.8, color = NA) +
  scale_fill_viridis_d(option = "mako", begin = 0.1, end = 0.9)  +
  scale_y_continuous(limits = c(0, 1), labels = percent_format()) +
  labs(
    title = "Proportion of Voting Yes per Year by National Party(EP1)
   ",
    x = "Year",
    y = "Proportion of Yes Votes",
    fill = "National Party") +
  theme_bw() +
  theme(
    plot.title = element_text(size = 14, hjust = 0.5, face = "bold"),
    axis.title.x = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10, face = "bold"),
    legend.text = element_text(size = 6),
    legend.title = element_text(size = 8),
    legend.key.size = unit(0.3, "cm"),
    legend.position = "right") +
  guides(fill = guide_legend(ncol = 2))
print(plot4)
```

Figure 4: Proportion of Yes Votes per Year by National Party



5. For each EP group, calculate the average *Yes* share per year and plot a line graph. Codes for data wrangling and plotting are as follow:

```
#Q5.Calculate the average Yes share per year and plot a line graph by
    EPG
# Compute the average Yes share per year by EP group
epg_votes <- com_ep1 |>
  filter(Decision %in% c("Yes", "No", "Abstain") & EPG != "0" & !is.
    na(Year)) |>
  mutate(Year = as.integer(Year)) |>
  group_by(EPG, Year) |>
  summarise(mean_yes = mean(Decision == "Yes"), .groups = "drop")
head(epg_votes)

# Plot the line graph
pdf("plot5.pdf")
plot5 <- ggplot(epg_votes, aes(x = Year, y = mean_yes, color = EPG,
    group = EPG)) +
```

```
  geom_line(size = 0.8, alpha = 0.8) +
  geom_point(size = 0.8, alpha = 0.8) +
  scale_color_brewer(palette = "Set1", name = "EP Group") +
  scale_y_continuous(limits = c(0, 1), labels = percent_format()) +
  labs(
    title = "Average Yes Share per Year by EP Group(EP1)",
    x = "Year",
    y = "Average Yes Share",
    color = "EP Group"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 14, hjust = 0.5, face = "bold"),
    axis.title.x = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10, face = "bold"),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 9),
    legend.position = "right")
print(plot5)
```

Figure 5: Average Yes Share per Year by EP Group