Problem Set 1

Data Visualisation for Social Scientists

Due: January 28, 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)

```
install.packages(c("readxl", "readr"))
library(readxl)
library(readr)
library(tidyverse)

setwd("~/Desktop/whiz1/DataViz_2026/problemSets/PS01/my_answers")
mep_info <- read_excel("mep_info_26Jul11.xls", sheet = "EP1", col_names =
    TRUE)
```

```
head(mep_info)
str(mep_info)
rcv_ep1 <- read_table2("rcv_ep1.txt")
head(rcv_ep1)
str(rcv_ep1)
```

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

   (a) **mep_info_26Jul11.xls (MEP characteristics)**
   Unit of analysis: Each row represents an individual Member of the European Parliament (MEP) across the first five European Parliaments (EP1–EP5).
   Key variables: The dataset includes MEP identifiers (e.g., MEP_ID), personal characteristics (such as name), political affiliation (national party and European Parliament group), country of representation, and ideological positioning variables. This dataset describes attributes of MEPs rather than their voting behavior.

   (b) **rcv_ep1.txt (EP1 roll-call votes)**
   Unit of analysis: Each row represents a single vote cast by an individual MEP in a recorded roll-call vote during the first European Parliament (EP1).
   Key variables: Key variables include the MEP identifier, roll-call vote identifier, and the recorded vote decision (e.g., Yes, No, Abstain, Present but did not vote, or Absent). This dataset captures individual voting behavior on legislative decisions.

3. The `rcv_ep1` data are in a wide format, with V1, V2, . . . , Vn as separate vote columns.

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

```
#Question 3
rcv_ep1 <- read_delim("rcv_ep1.txt")
colnames(rcv_ep1)
rcv_ep1_long <- rcv_ep1 %>%
  pivot_longer(
    cols = starts_with("V"),
    names_to = "vote_id",
    values_to = "decision"
  )
```

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

```r
rcv_ep1_long <- rcv_ep1_long %>%
  mutate(
    decision_label = case_when(
      decision == 1 ~ "Yes",
      decision == 2 ~ "No",
      decision == 3 ~ "Abstain",
      decision == 4 ~ "Absent",
      decision == 5 ~ "Present but did not vote",
      decision == 0 ~ "Not a member",
      TRUE ~ "Other"
    )
  )

decision_summary <- rcv_ep1_long %>%
  count(decision_label) %>%
  arrange(desc(n))

decision_summary
```

Decision	Count
Absent	109,224
Present but did not vote	103,618
Not a member	99,753
Yes	88,185
No	75,171
Abstain	9,577

Table 1: Distribution of vote decisions in EP1 roll-call votes

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.

```r
#Question 4
meta_cols <- c("MEPID", "MEPNAME", "MS", "NP", "EPG")
ep1_combined <- rcv_ep1_long %>%
  select(all_of(meta_cols), vote_id, decision, decision_label)
head(ep1_combined)
str(ep1_combined)
missing_summary <- ep1_combined %>%
  summarise(across(everything(), ~sum(is.na(.))))

missing_summary
```

MEPID	MEPNAME	MS	NP	EPG	vote_id	decision	decision_label
0	0	0	0	0	0	0	0

Table 2: Example observation from the combined EP1 voting dataset

5. Compute, for each EP group in EP1:

   - The mean rate of Yes votes (Yes over Yes+No+Abstain) across all roll calls.

```r
#Question 5.1
#Filter only relevant votes (Yes, No, Abstain)
votes_filtered <- ep1_combined %>%
  filter(decision_label %in% c("Yes", "No", "Abstain"))

#Compute Yes rate per roll call per EP group
yes_rate_per_vote <- votes_filtered %>%
  group_by(EPG, vote_id) %>%  # each vote within each EP group
  summarise(
    yes_count = sum(decision_label == "Yes"),
    total_votes = n(),
    yes_rate = yes_count / total_votes,
    .groups = "drop"
  )

#Compute mean Yes rate across all roll calls for each EP group
mean_yes_rate_epg <- yes_rate_per_vote %>%
  group_by(EPG) %>%
  summarise(mean_yes_rate = mean(yes_rate), .groups = "drop") %>%
  arrange(desc(mean_yes_rate))

mean_yes_rate_epg
```

EP Group	Mean Yes Rate
N	0.613
S	0.563
E	0.532
G	0.514
M	0.512
L	0.495
R	0.486
C	0.437

Table 3: Average proportion of Yes votes by EP group in EP1

   - The mean abstention rate.

```r
#Question5.2
#Filter only relevant votes (Yes, No, Abstain)
votes_filtered <- ep1_combined %>%
  filter(decision_label %in% c("Yes", "No", "Abstain"))

#Compute Abstain rate per roll call per EP group
abstain_rate_per_vote <- votes_filtered %>%
  group_by(EPG, vote_id) %>%
  summarise(
    abstain_count = sum(decision_label == "Abstain"),
    total_votes = n(),
    abstain_rate = abstain_count / total_votes,
    .groups = "drop"
  )

#Compute mean Abstain rate across all roll calls for each EP group
mean_abstain_rate_epg <- abstain_rate_per_vote %>%
  group_by(EPG) %>%
  summarise(mean_abstain_rate = mean(abstain_rate), .groups = "drop") %>%
  arrange(desc(mean_abstain_rate))

# View the result
mean_abstain_rate_epg
```

EP Group	Mean Abstention Rate
R	0.2690
M	0.0893
G	0.0806
C	0.0725
L	0.0654
S	0.0598
N	0.0580
E	0.0215

Table 4: Average proportion of abstentions by EP group in EP1

- The mean vote preferences along the two contested dimensions (NOM-D1 and NOM-D2).

```r
#Question5.3
library(dplyr)
mep_info <- read_excel("mep_info_26Jul11.xls", sheet = "EP1",
    col_names = TRUE)

#clean column names
colnames(mep_info) <- colnames(mep_info) %>%
  trimws() %>%
  gsub(" ", ".", .)  # replace spaces with dots
```

```
9
10 #Ensure the key columns exist
11 colnames(mep_info)
12 head(mep_info)
13
14
15 #Fix column types and trim spaces
16 mep_info <- mep_info %>%
17   mutate(
18     MEP.id = trimws(as.character(MEP.id)),   # ensure same type as
     ep1_combined$MEPID
19     `NOM-D1` = as.numeric(`NOM-D1`),
20     `NOM-D2` = as.numeric(`NOM-D2`)
21   )
22
23 ep1_combined <- ep1_combined %>%
24   mutate(MEPID = trimws(as.character(MEPID)))
25
26
27 #Merge NOM-D1 and NOM-D2 into long-format vote dataset
28 ep1_combined_nom <- ep1_combined %>%
29   left_join(
30     mep_info %>% select(MEP.id, `NOM-D1`, `NOM-D2`),
31     by = c("MEPID" = "MEP.id")
32   )
33
34
35 #Compute mean NOM-D1 and NOM-D2 per EP group
36 mean_nominate_epg <- ep1_combined_nom %>%
37   group_by(EPG) %>%
38   summarise(
39     mean_nom_d1 = mean(`NOM-D1`, na.rm = TRUE),
40     mean_nom_d2 = mean(`NOM-D2`, na.rm = TRUE),
41     .groups = "drop"
42   ) %>%
43   arrange(mean_nom_d1)
44
45
46 #View the result
47 mean_nominate_epg
```

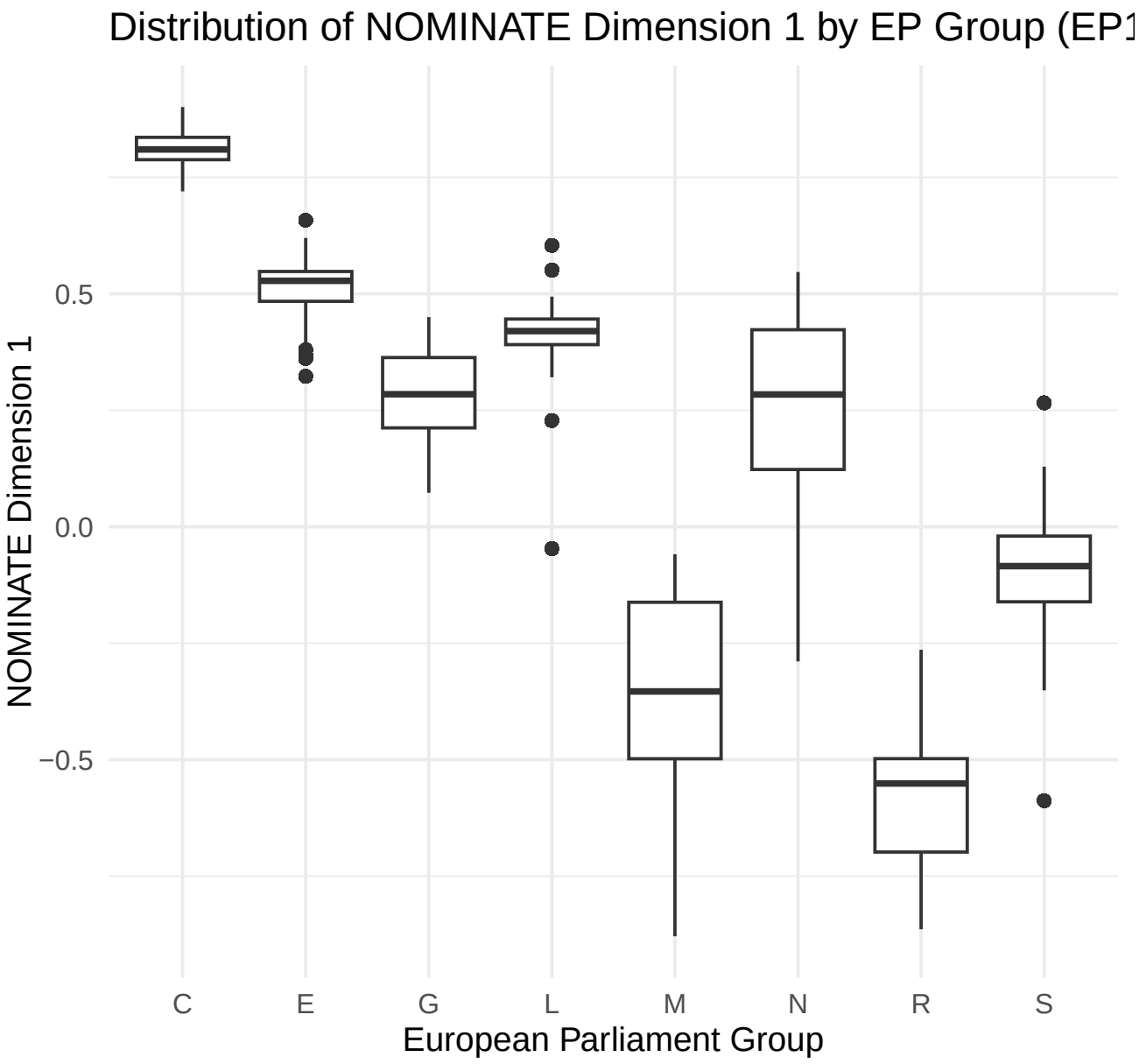
EP Group	Mean NOM-D1	Mean NOM-D2
R	-0.586	-0.0419
M	-0.357	-0.2010
S	-0.0980	0.2610
N	0.250	-0.3860
G	0.280	-0.8180
L	0.409	-0.3240
E	0.512	-0.2770
C	0.811	0.5300
0	NaN	NaN

Table 5: Average NOMINATE coordinates (Dimension 1 and Dimension 2) by EP group in EP1

Data Visualization

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

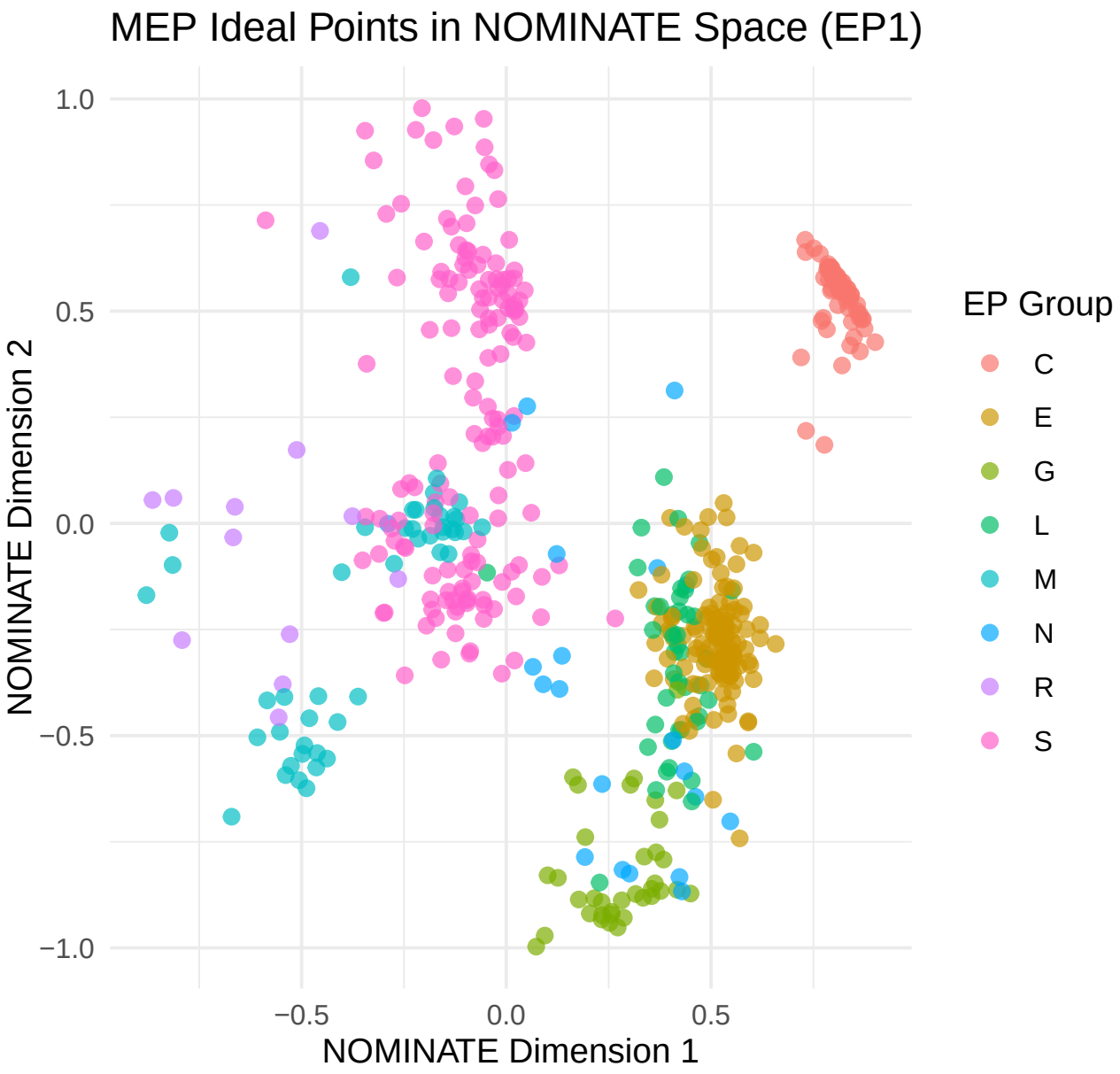
```
#Data Visualization
#Question 1
library(ggplot2)
#Keep only observations with valid NOM-D1 values
plot_data <- ep1_combined_nom %>%
  filter(!is.na(`NOM-D1`))

#Boxplot of NOM-D1 by EP group
ggplot(plot_data, aes(x = EPG, y = `NOM-D1`)) +
  geom_boxplot() +
  labs(
    title = "Distribution of NOMINATE Dimension 1 by EP Group (EP1)",
    x = "European Parliament Group",
    y = "NOMINATE Dimension 1"
  ) +
  theme_minimal()
```


Distribution of NOMINATE Dimension 1 by EP Group (EP1
NOMINATE Dimension 1
0.5
0.0
−0.5
C
E
G
L
M
N
R
S
European Parliament Group

The plot shows clear differences in NOM-D1 across EP groups, with left-leaning groups concentrated at lower values and right-leaning groups at higher values. There is a negative linear relationship shown in the graph. This indicates that the first NOMINATE dimension captures an underlying ideological cleavage in the European Parliament.

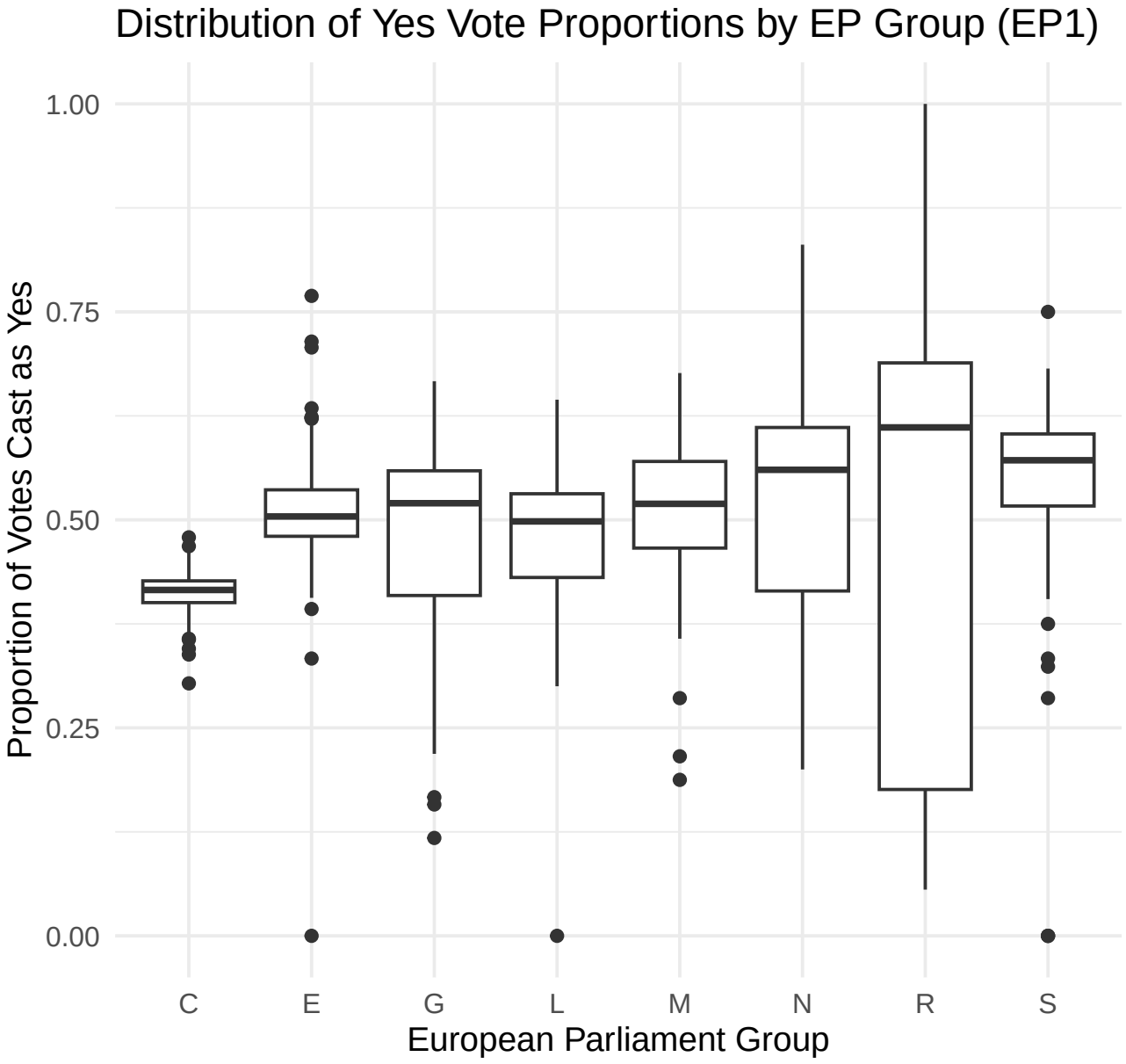
2. Make a scatterplot of *nomdim1* (x-axis) and *nomdim2* (y-axis), with one point per MEP and color by EP group.

```
#Question 2
#Create a unique MEP-level dataset
mep_level <- ep1_combined_nom %>%
  select(MEPID, EPG, `NOM-D1`, `NOM-D2`) %>%
  distinct() %>%                          # one row per MEP
  filter(!is.na(`NOM-D1`), !is.na(`NOM-D2`))

#Scatterplot
ggplot(mep_level, aes(x = `NOM-D1`, y = `NOM-D2`, color = EPG)) +
  geom_point(alpha = 0.7, size = 2) +
  labs(
    title = "MEP Ideal Points in NOMINATE Space (EP1)",
    x = "NOMINATE Dimension 1",
    y = "NOMINATE Dimension 2",
    color = "EP Group"
  ) +
  theme_minimal()
```


MEP Ideal Points in NOMINATE Space (EP1)
NOMINATE Dimension 2
NOMINATE Dimension 1
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
EP Group
C
E
G
L
M
N
R
S

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

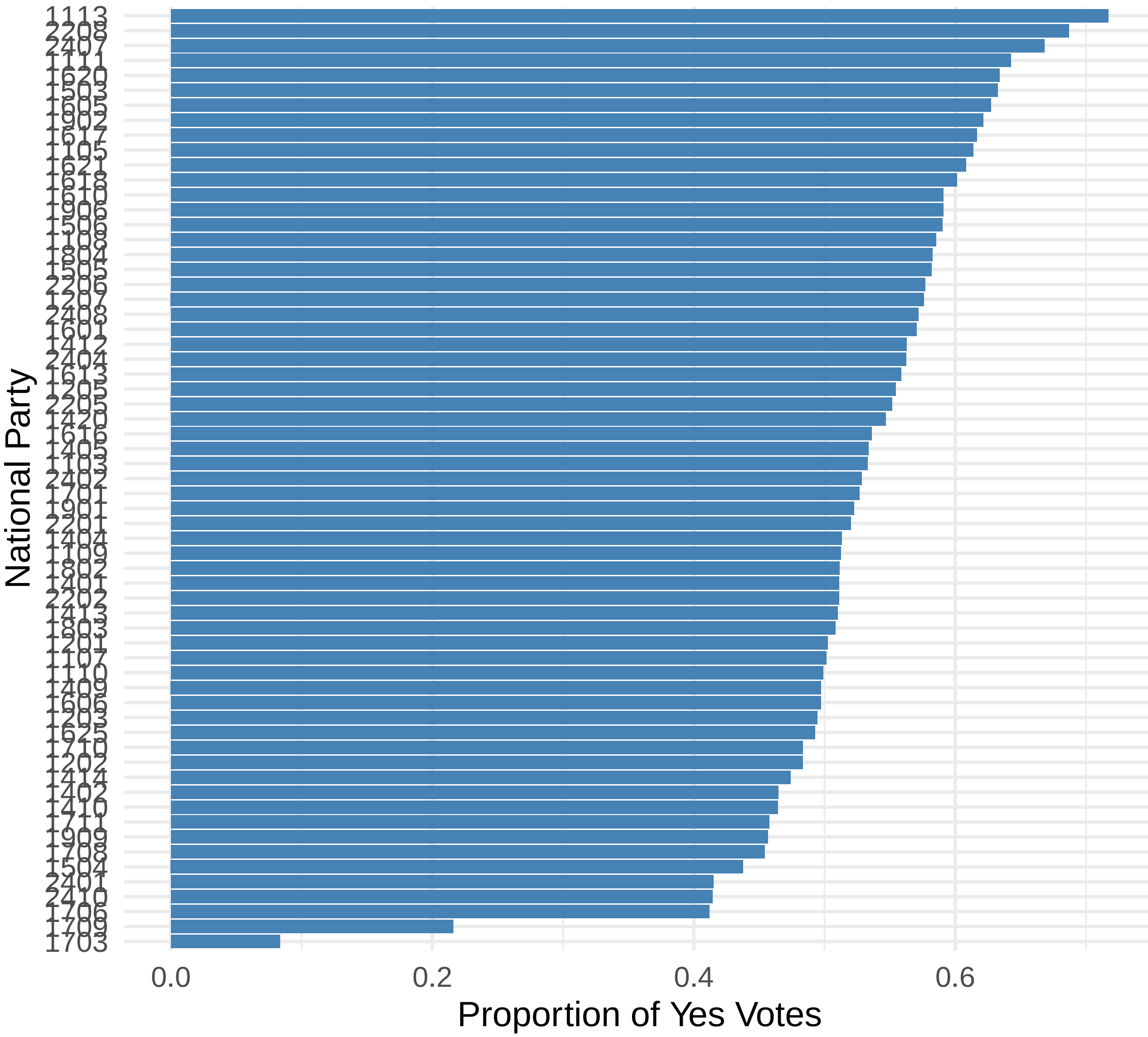
```
#Question 3
# Step 1: Compute proportion of Yes votes per MEP
mep_yes_rate <- ep1_combined %>%
  filter(decision_label %in% c("Yes", "No", "Abstain")) %>% #
    denominator
  group_by(MEPID, EPG) %>%
  summarise(
    yes_rate = mean(decision_label == "Yes"),
    .groups = "drop"
  )

#Boxplot by EP group
ggplot(mep_yes_rate, aes(x = EPG, y = yes_rate)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Yes Vote Proportions by EP Group (EP1)",
    x = "European Parliament Group",
    y = "Proportion of Votes Cast as Yes"
  ) +
  theme_minimal()
```


Distribution of Yes Vote Proportions by EP Group (EP1)
Proportion of Votes Cast as Yes
1.00
0.75
0.50
0.25
0.00
C
E
G
L
M
N
R
S
European Parliament Group

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

```
#Question4
ep1_combined %>%
  filter(decision_label %in% c("Yes", "No", "Abstain")) %>%
  group_by(NP) %>%
  summarise(
    yes_rate = mean(decision_label == "Yes"),
    .groups = "drop"
  ) %>%
  ggplot(aes(x = reorder(factor(NP), yes_rate), y = yes_rate)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Proportion Voting Yes by National Party (EP1)",
    x = "National Party",
    y = "Proportion of Yes Votes"
  ) +
  theme_minimal()
```

Proportion Voting Yes by National Party (EP1)



5. For each EP group, calculate the average *Yes* share per year and plot a line graph.