PS01: Roll-Call Votes in the European Parliament (EP1)

1 Data Manipulation

1.1 (1) Load datasets into R

I loaded (i) MEP characteristics from `mep_info_26Jul11.xls` (sheet EP1) and (ii) roll-call votes from `rcv_ep1.txt`. The key code is:

```
library(tidyverse)
library(readxl)
library(janitor)

mep_ep1 <- read_excel("mep_info_26Jul11.xls", sheet = "EP1") %>%
clean_names() %>%
rename(
mepid = mep_id,
epg = ep_group,
ms = member_state,
np = national_party
) %>%
mutate(
nom_d1 = readr::parse_number(as.character(nom_d1)),
nom_d2 = readr::parse_number(as.character(nom_d2))
)

rcv_ep1 <- readr::read_delim("rcv_ep1.txt", delim = ",", show_col_types = FALSE)
    %>%
clean_names()
```

Listing 1: Loading and cleaning datasets (key code)

1.2 (2) Unit of analysis and key variables

MEP characteristics dataset (`mep_info_26Jul11.xls`). The unit of analysis is the individual MEP. Key variables include `mepid`, `ms`, `np`, `epg`, and ideological positions `nom_d1` and `nom_d2`. The NOMINATE dimensions summarize MEPs' revealed preferences inferred from roll-call voting.

Roll-call vote dataset (`rcv_ep1.txt`). The dataset records how each MEP voted in each roll-call vote. In the raw file, the unit of observation is the MEP and the vote outcomes are stored in wide format (`v1` to `v886`). Each cell contains a numeric code for voting behaviour.

1.3 (3) Reshape votes from wide to long format

The ID/metadata columns are `mepid`, `mepname`, `ms`, `np`, and `epg`. All other columns correspond to vote decisions. I reshaped the dataset into long format using `pivot_longer()`.

```
id_cols <- c("mepid", "mepname", "ms", "np", "epg")
vote_cols <- setdiff(names(rcv_ep1), id_cols)

votes_long <- rcv_ep1 %>%
pivot_longer(
cols = all_of(vote_cols),
names_to = "vote_id",
values_to = "decision"
)
```

Listing 2: Wide-to-long reshaping (key code)

1.4 (4) Merge votes with MEP characteristics and recode decisions

I merged the long-format votes with MEP characteristics by `mepid`. Following the assignment convention, I excluded the unlabelled group (`epg = "0"`). Then I recoded vote outcomes into Yes/No/Abstain/Absent-Other, and marked valid votes (Yes/No/Abstain) for denominators.

```
mep_votes_ep1 <- votes_long %>%
left_join(mep_ep1, by = "mepid", suffix = c("", "_mep")) %>%
select(-epg_mep) %>%
filter(epg != "0")

mep_votes_a5 <- mep_votes_ep1 %>%
mutate(
decision_label = case_when(
decision == 1 ~ "Yes",
decision == 2 ~ "No",
decision == 3 ~ "Abstain",
TRUE ~ "Absent/Other"
),
is_valid = if_else(decision_label %in% c("Yes", "No", "Abstain"), 1L, 0L)
)
```

Listing 3: Merge and recode decisions (key code)

1.5 (3) Summary of decision categories across all votes

```
decision_counts <- mep_votes_a5 %>%
count(decision_label, sort = TRUE)
decision_counts
```

Listing 4: Counting decision categories (key code)

Table 1 reports the same counts in formatted form.

Table 1: Counts of Voting Decisions Across All Votes

decision_label	n
Absent/Other	311709
Yes	88185
No	75171
Abstain	9577

1.6 (5) EP-group statistics: Yes rate, abstention rate, and NOMINATE positions

For each EP group, I computed: (i) mean Yes vote rate = Yes / (Yes+No+Abstain), (ii) mean abstention rate, (iii) mean `nom_d1` and `nom_d2` across MEPs in the group, and group size.

```
a5_analysis <- mep_votes_a5 %>%
group_by(epg) %>%
summarise(
total_valid = sum(is_valid, na.rm = TRUE),
mean_yes_rate = if_else(total_valid > 0,
sum(decision == 1, na.rm = TRUE) / total_valid, NA_real_
),
mean_abstention_rate = if_else(total_valid > 0,
sum(decision == 3, na.rm = TRUE) / total_valid, NA_real_
),
.groups = "drop"
)

a5_nominate <- mep_ep1 %>%
filter(epg != "0") %>%
group_by(epg) %>%
summarise(
mean_nom_d1 = mean(nom_d1, na.rm = TRUE),
mean_nom_d2 = mean(nom_d2, na.rm = TRUE),
n_meps = n(),
.groups = "drop"
)

a5_final_table <- left_join(a5_analysis, a5_nominate, by = "epg")
a5_final_table
```

Listing 5: Computing A5 statistics (key code)

Table 2: Summary Statistics by EP Group (EP1)

epg	total_valid	mean_yes_rate	mean_abstention_rate	mean_nom_d1	mean_nom_d2	n_meps
C	34740	0.42	0.08	0.81	0.53	63
E	50916	0.51	0.02	0.51	-0.28	137
G	6710	0.51	0.07	0.28	-0.82	46
L	12605	0.49	0.06	0.41	-0.32	48
M	12735	0.53	0.08	-0.36	-0.20	53
N	3437	0.58	0.06	0.25	-0.39	25
R	2025	0.46	0.27	-0.59	-0.04	13
S	49765	0.58	0.06	-0.10	0.26	166

2 Data Visualisation

2.1 (B1) Distribution of NOMINATE Dimension 1 by EP group

Figure 1 plots the distribution of `nom_d1` by EP group. Figure B1 shows that the distribution of NOMINATE Dimension 1 differs clearly across EP groups. Several groups are concentrated around distinct regions of the ideological spectrum, indicating systematic differences in ideological positioning between EP groups. While some overlap exists, the overall distributions suggest that EP group affiliation is associated with legislators' positions on the primary ideological dimension.

```
plot_data <- mep_ep1 %>% filter(epg != "0")

p_b1 <- ggplot(plot_data, aes(x = nom_d1, fill = epg)) +
geom_density(alpha = 0.35) +
labs(x = "nom_d1", y = "Density", fill = "EP group") +
theme_minimal()
```

Listing 6: B1 density plot (key code)

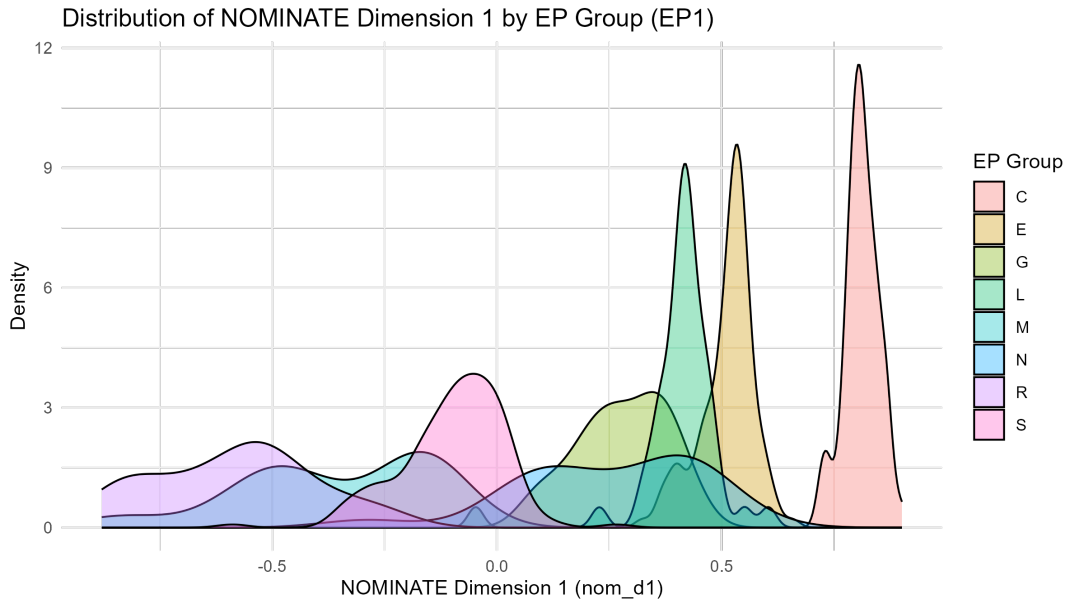

Figure 1: B1: Distribution of NOMINATE Dimension 1 by EP Group (EP1)

2.2 (B2) NOMINATE Dimension 1 vs Dimension 2

Figure 2 shows a scatterplot of `nom_d1` vs `nom_d2` coloured by EP group. Figure B2 plots MEPs' positions on NOMINATE Dimension 1 against Dimension 2. The scatterplot shows visible clustering by EP group, with groups occupying different regions of the two-dimensional ideological space. This suggests that EP group membership is associated not only with positions on the primary dimension, but also with variation along the second ideological dimension.

```
p_b2 <- ggplot(plot_data, aes(x = nom_d1, y = nom_d2, color = epg)) +
geom_point(alpha = 0.7) +
labs(x = "nom_d1", y = "nom_d2", color = "EP group") +
theme_minimal()
```

Listing 7: B2 scatter plot (key code)

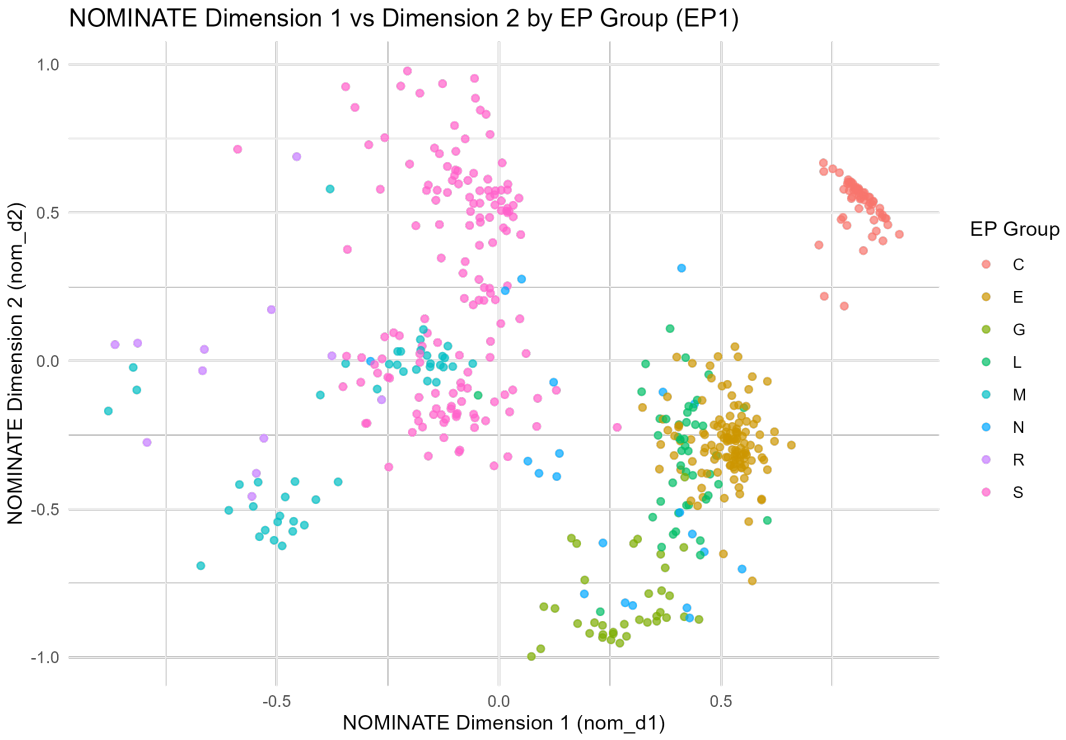

Figure 2: B2: NOMINATE Dimension 1 vs Dimension 2 by EP Group (EP1)

2.3 (B3) Cohesion within EP groups

Cohesion is measured using each MEP's Yes proportion across valid votes. Figure 3 shows boxplots of this proportion by EP group. Figure B3 presents boxplots of the proportion of Yes votes at the MEP level by EP group. Some groups display relatively narrow interquartile ranges, indicating more consistent voting behaviour among their members, while others show wider dispersion. This suggests that levels of internal voting cohesion vary across EP groups.

```
mep_yes_prop <- mep_votes_a5 %>%
filter(decision %in% c(1, 2, 3)) %>%
group_by(mepid, epg) %>%
summarise(
yes_prop = sum(decision == 1, na.rm = TRUE) / n(),
.groups = "drop"
)

p_b3 <- ggplot(mep_yes_prop, aes(x = epg, y = yes_prop)) +
geom_boxplot() +
labs(x = "EP group", y = "Yes proportion") +
theme_minimal()
```

Listing 8: B3 cohesion (key code)

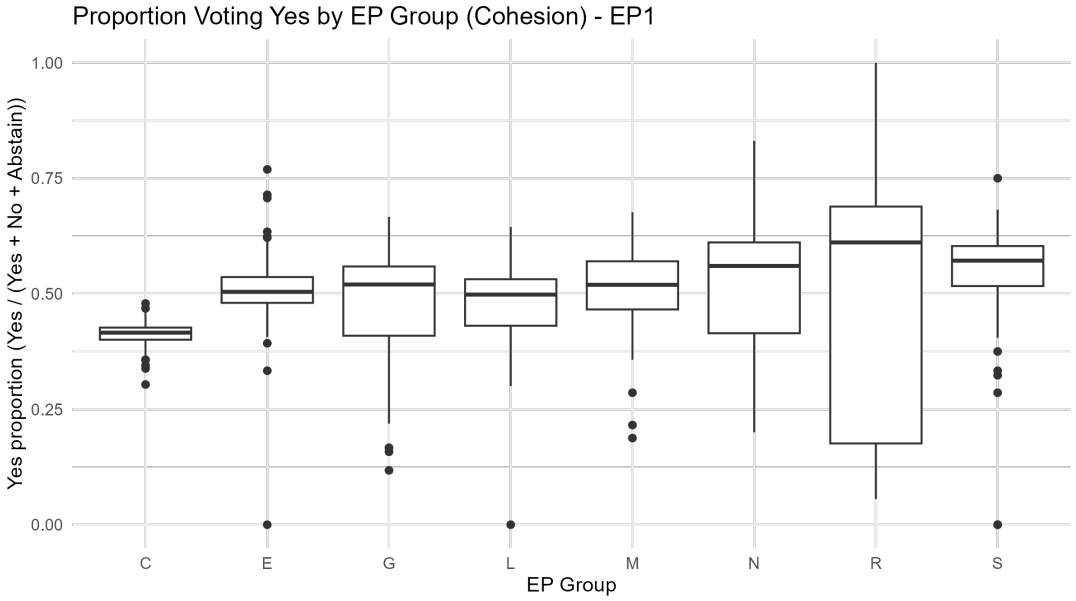

Figure 3: B3: Proportion Voting Yes by EP Group (Cohesion) – EP1

2.4 (B4) Proportion voting Yes per year by national party

I created a lookup table mapping `vote_id` to year using `vote_info_Jun2010.xls` (sheet EP1), then merged year back into votes. The diagnostic output reported 886 rows and years 1979–1984 (with a small number of missing years). Figure B4 shows the proportion of Yes votes over time for major national parties. The figure indicates that Yes voting rates differ across parties and vary across years. While some parties display relatively stable patterns, others show noticeable changes over time, suggesting temporal variation in voting behaviour at the national party level.

```
library(lubridate)
library(stringr)

vote_info_ep1 <- read_excel("vote_info_Jun2010.xls", sheet = "EP1") %>%
clean_names()

date_col <- intersect(names(vote_info_ep1), c("vote_date", "date"))[1]

vote_year_lookup <- vote_info_ep1 %>%
mutate(
vote_date_raw = .data[[date_col]],
vote_date_chr = str_trim(as.character(vote_date_raw)),
vote_date_num = suppressWarnings(as.numeric(vote_date_chr)),
vote_date_from_num = if_else(
!is.na(vote_date_num),
as.Date(vote_date_num, origin = "1899-12-30"),
as.Date(NA)
),
vote_date_from_chr = as.Date(
suppressWarnings(lubridate::parse_date_time(
str_sub(vote_date_chr, 1, 19),
orders = c("Ymd", "Y-m-d", "dmy", "d/m/Y", "m/d/Y", "Y/m/d", "d.m.Y")
))
),
vote_date = coalesce(vote_date_from_num, vote_date_from_chr),
vote_id = paste0("v", row_number()),
year = year(vote_date)
) %>%
select(vote_id, year)

votes_valid_year <- votes_long %>%
left_join(vote_year_lookup, by = "vote_id") %>%
mutate(
decision_label = case_when(
decision == 1 ~ "Yes",
decision == 2 ~ "No",
decision == 3 ~ "Abstain",
TRUE ~ NA_character_
)
) %>%
filter(!is.na(year), decision_label %in% c("Yes", "No", "Abstain"))

top_np <- votes_valid_year %>% count(np, sort = TRUE) %>% slice_head(n = 10) %>%
    pull(np)

b4_data <- votes_valid_year %>%
mutate(np_plot = if_else(as.character(np) %in% as.character(top_np), as.character
    (np), "Other")) %>%
group_by(year, np_plot) %>%
summarise(yes_prop = mean(decision_label == "Yes"), .groups = "drop")
```

Listing 9: B4 year merge and aggregation (key code)

```
p_b4 <- ggplot(b4_data, aes(x = factor(year), y = yes_prop, fill = np_plot)) +
geom_col(position = "dodge") +
labs(x = "Year", y = "Yes proportion", fill = "National party") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Listing 10: B4 plot (key code)

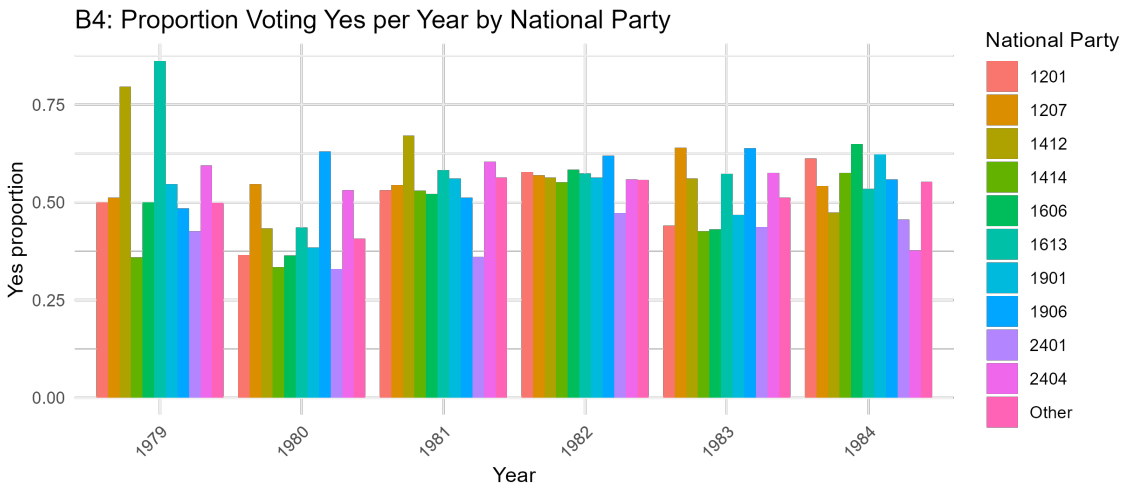

Figure 4: B4: Proportion Voting Yes per Year by National Party

2.5 (B5) Average Yes share per year by EP group

Figure 5 shows the average Yes share per year for each EP group.Figure B5 displays the average Yes share per year for each EP group. The figure shows that EP groups differ in their overall levels of support, with some groups maintaining consistently higher or lower Yes shares across years. In addition, the trajectories indicate that the average Yes share is not constant over time, highlighting temporal variation in group-level voting behaviour.

```
b5_data <- votes_valid_year %>%
filter(epg != "0", !is.na(epg)) %>%
group_by(year, epg) %>%
summarise(avg_yes_share = mean(decision_label == "Yes"), .groups = "drop")

p_b5 <- ggplot(b5_data, aes(x = year, y = avg_yes_share, color = epg, group = epg
    )) +
geom_line(linewidth = 1) +
geom_point() +
labs(x = "Year", y = "Average Yes share", color = "EP group") +
theme_minimal()
```

Listing 11: B5 aggregation and plot (key code)

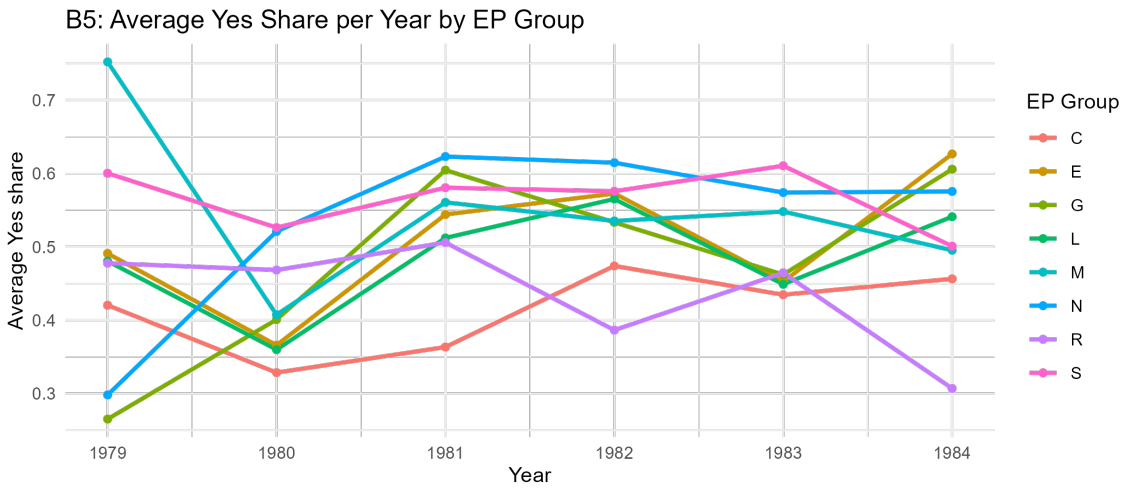

Figure 5: B5: Average Yes Share per Year by EP Group

3 Notes on Warnings (reproducibility)

When running the script, R reported parsing warnings when converting NOMINATE variables to numeric (e.g., `parse_number()` parsing failures) and removed non-finite observations in plotting. These warnings reflect missing or non-numeric entries in the raw files and do not change the overall workflow.