Problem Set 1

Data Visualisation for Social Scientists

Matilda Tomatis

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)
2. Briefly describe (2–3 sentences each) the unit of analysis and key variables in each of these two datasets.
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 ($V_1 \dots V_n$). Tidy the voting data such that each row/observation is a single vote for a single MEP.
- Create a summary table of counts of decision categories (e.g. Yes/No/Abstain/Present but did not vote/Absent) across all 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.

5. Compute, for each EP group in EP1:

- The mean rate of Yes votes (Yes over Yes+No+Abstain) across all roll calls.
- The mean abstention rate.
- The mean vote preferences along the two contested dimensions (NOM-D1 and NOM-D2).

Answer 1

Both dataset are loaded in the global environment.

```
rcv_ep1 <- read_csv("rcv_ep1.txt")|>
  mutate(MEPID = as.character(MEPID))

mep_info_26Jul11 <- read_excel("mep_info_26Jul11.xls", sheet = "EP1") |>
  rename(MEPID = `MEP id`)|>
  mutate(MEPID = as.character(MEPID))
```

Answer 2

In `mep_info_26Jul11` the units of analysis are MEPs; their main characteristics are present: such as their ID number (`"MEP id"`), their EP group (`"EP Group"`), the country of origin, (`"Member State"`), and their national party of affiliation (`"National Party"`). Moreover, each MEP is given nominate coordinates on two directions.
In `rcv_ep1` are present MEP who partecipated in at least one roll-call vote (the number of observation is smaller than ythe one in mep_info_26Jun11 plausably for this reason). In this dataset are present some basic personal information (such as the MEP id number and name, the country of origin, the EP and national political affiliation), plus wheter MEP were present at each of the EP1 roll-call votes and the vote they casted.

Answer 3

Columns with the name starting with the letter ”V” are those in which vote decisions are stored. The data is tidied from a wide to a long format, as such all the voting decision will be stored in on single column called `vote_decision`.

```
rcv_ep1 <- rcv_ep1 |>
  pivot_longer(
    cols = starts_with("V"),
    names_to = "vote_number",
    values_to = "vote_decision"
  )
```

For clarity in interpretation and subsequent analysis, the `vote_decision` variable is recoded from numeric values into descriptive categorical labels (e.g. Yes, No, Absent) and converted into a factor, with levels corresponding to each voting outcome.

```
rcv_ep1$vote_decision <- recode_factor(rcv_ep1$vote_decision,
                                        "0"= "Absent",
                                       "1"="Yes",
                                       "2"="No",
                                       "3"="Abstain",
                                       "4"="Present_no_vote",
                                       "5"="Not_MEP")
```

Here follows a summary table of both the count and the share of each possible decision category across all roll-call votes and MPEs. The highest percentage are 21.3% and 22.5%, corresponding to `Present but did not vote` and `Not an MEP`

```
summary_vote <- rcv_ep1 |>
  group_by(vote_decision) |>
  summarise(
    total_votes = n()
  )|>
  mutate(
    vote_share = total_votes / sum(total_votes)
  )
```

Table 1: Summary count and share of MEPs vote decisions

vote_decision	total_votes	vote_share
Absent	99753	0.205
Yes	88185	0.182
No	75171	0.155
Abstain	9577	0.020
Present_no_vote	109224	0.225
Not_MEP	103618	0.213

Answer 4

The two datasets are merged by the IDs of the MEPs, subsequently columns which were reduntant in informations were eliminated from the `merged_mep` dataset. Only the information in reagrds to EP1 members are kepts in this joined dataset.

```
mep_merged <- rcv_ep1 |>
  left_join(mep_info_26Jul11, by = "MEPID") |>
  select(!c(8:11))|>
  mutate( ‘NOM-D1‘ = as.numeric(‘NOM-D1‘),
          ‘NOM-D2‘ = as.numeric(‘NOM-D2‘))
```

There are some cases of missingness. More in particular, for some of the MEPs nominate coordinates information is not present.

```
> any(is.na(mep_merged))
[1] TRUE
> colSums(is.na(mep_merged))
MEPID       MEPNAME             MS             NP            EPG
0              0               0              0              0
vote_number vote_decision         NOM-D1         NOM-D2
0              0          42528          42528
```

The data is aggregated in relation to unique IDs to understand how this missigness translates into MEPs.

```
mep_missing <- mep_merged |>
  group_by(MEPID) |>
  summarise(has_missing_coordinates = any(is.na(‘NOM-D1‘)))

MEPs <- length(unique(mep_merged$MEPID))
MEP_not_coord <- sum(mep_missing$has_missing_coordinates)
```

```
> paste("Out of a total of ", MEPs,", ", MEP_not_coord, " have some
missing coordinate information" )
[1] "Out of a total of  548 ,  48  have some missing coordinate information"
```

Answer 5

The following tables present the relevant statistics by EP group. The first table shows the rate of Yes votes, calculated over the total of Yes, No, and Abstain votes. The second table reports the mean abstention rates, using the same denominator as the previous statistic. Finally, the third table presents the mean coordinates for each EP group.
In order to properly calculate these statistics, a wrongly coded entry was corrected to a NA value.

```
mep_merged <- mep_merged |>
  mutate(EPG = na_if(EPG, ”0”))

vote_rate_epg <- mep_merged |>
  filter(vote_decision %in% c(”Yes”, ”No”, ”Abstain”)) |>
  group_by(EPG) |>
  summarise(
    yes = sum(vote_decision == ”Yes”),
    no = sum(vote_decision == ”No”),
    abstain = sum(vote_decision == ”Abstain”),
    yes_rate = yes / (yes + no + abstain),
    abstain_rate = abstain / (yes + no + abstain))

 yes_share <- vote_rate_epg |>
    select(EPG, yes_rate)

abstain_share <- vote_rate_epg |>
    select(EPG, abstain_rate)

mep_merged <- mep_merged |>
  mutate(
    ‘NOM-D1‘ = as.numeric(gsub(”,”, ”.”, ‘NOM-D1‘)),
    ‘NOM-D2‘ = as.numeric(gsub(”,”, ”.”, ‘NOM-D2‘))
  )
```

```

info_EPG_mean_coord <- mep_merged |>
  group_by(EPG) |>
  summarise(
    mean_coord1 = mean(`NOM-D1`, na.rm = TRUE),
    mean_coord2 = mean(`NOM-D2`, na.rm = TRUE),
  )
```

Table 2: The mean rate of Yes votes - by EP group

EPG	yes_rate
C	0.415
E	0.509
G	0.512
L	0.486
M	0.528
N	0.581
R	0.457
S	0.576

The maximum mean rate is found for the `N European Parliament Group`, with a value equal to 58.1%; the lowest rate stands at 41.5%, being the mean rate of yes vote for the `C European Parliament Group`.

Table 3: The mean abstention rate - by EP group

EPG	abstain_rate
C	0.075
E	0.021
G	0.070
L	0.063
M	0.080
N	0.056
R	0.265
S	0.057

The abstention rates tend to be on the lower side, with the `R EP group` standing as an exception, having a rate of 26.5%.

Table 4: The mean of each coordinate dimension - by EP group

EPG	mean_coord1	mean_coord2
C	0.811	0.530
E	0.512	-0.277
G	0.280	-0.818
L	0.409	-0.324
M	-0.357	-0.201
N	0.250	-0.386
R	-0.586	-0.042
S	-0.098	0.261

The mean values for the nominate coordinates vary greatly between different parliamentary groups.

Data Visualization

1. Plot the distribution of the first NOMINATE dimension by EP group, and explain any trends you see.
2. Make a scatterplot of *nomdim1* (x-axis) and *nomdim2* (y-axis), with one point per MEP and color by EP group.
3. Produce a boxplot of the proportion voting *Yes* by EP group to visualize cohesion.
4. Display the proportion voting Yes by national party using a bar plot.

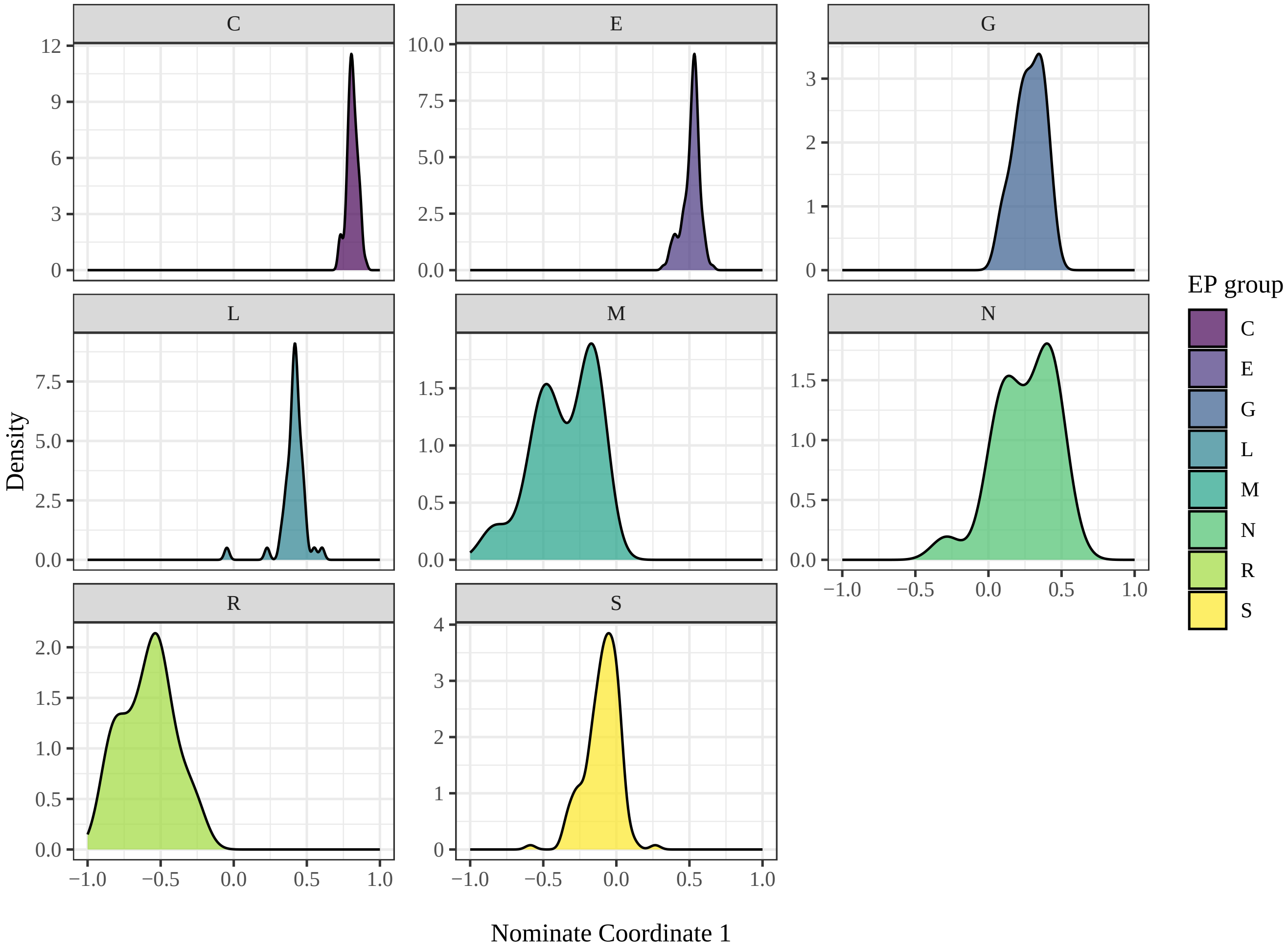
Data Preparation for visualization

In order to be able to plot the data for item 1 and 2, hence to have only one observation for each MEP with the related coordinates, the data was wrangled as follows.

```
mep_merged_viz <- mep_merged |>
  select(MEPID, EPG, ‘NOM-D1‘, ‘NOM-D2‘) |>
  unique() |>
  na.omit()
```

Answer 1

Figure 1:
Coordinate 1 density distribution − by European Parliament Group



```
mep_merged_viz|>
ggplot( aes(x = `NOM-D1`, fill = EPG)) +
  geom_density(alpha = 0.7) +
  facet_wrap(vars(EPG), scales = "free_y") +
  scale_x_continuous(limits = c(-1, 1)) +
  scale_fill_viridis_d(name = "EP group") +
  labs(
    title = "Coordinate 1 density distribution - by European Parliament Group"
    ,
    x = "\nNominate Coordinate 1",
    y = "\nDensity"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5))
```

As shown in Figure 1, the European Parliament groups present different density distributions, with some showcasing a more cohesive pattern along the first coordinate dimension and others being more dispersed.

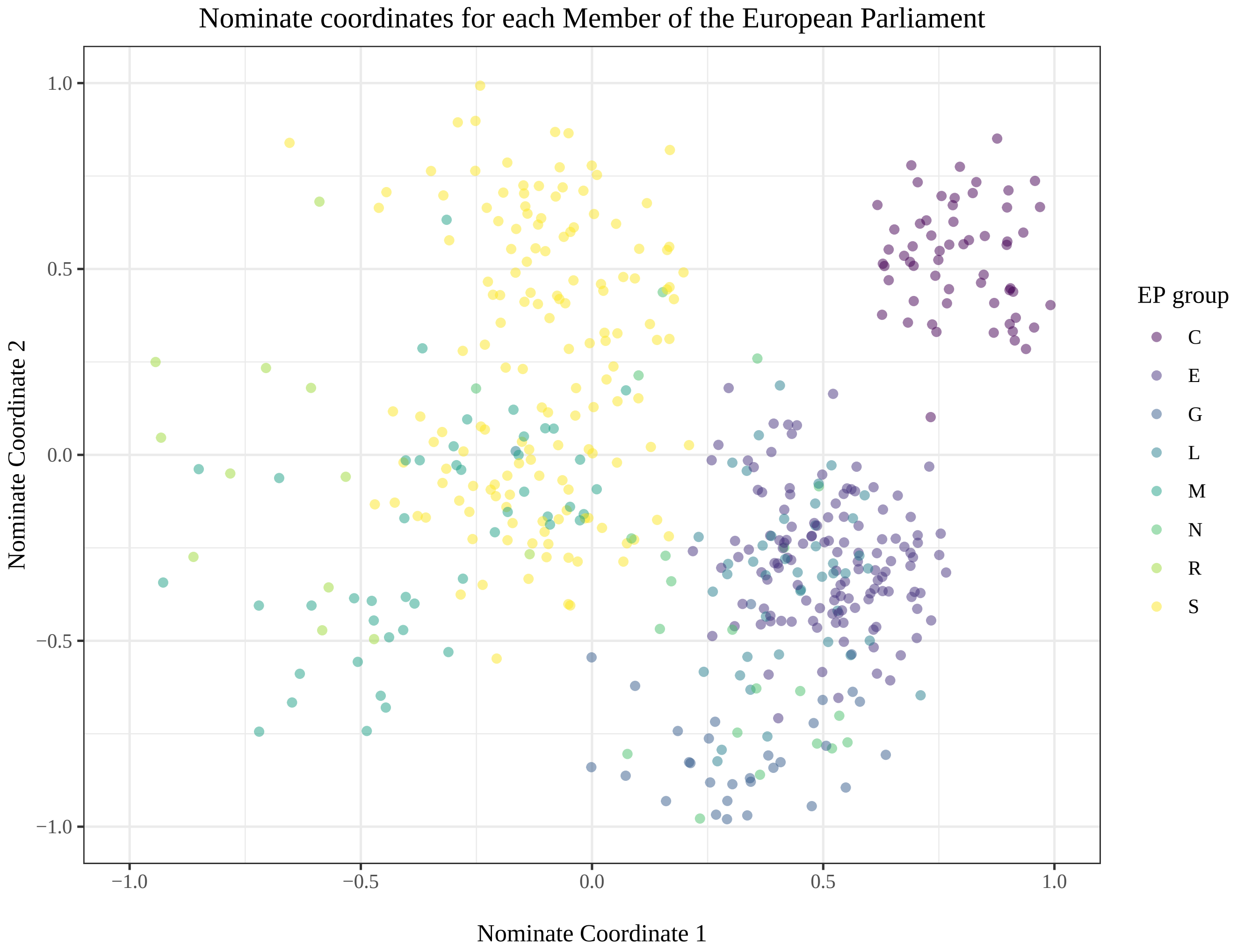
In particular, groups C, E, and L present a high spike, indicating a high probability of their members being positioned on the positive side of the first coordinate spectrum. Group G also displays an entirely positive density distribution, but with a wider spread than the previous groups.

Group S exhibits a center-left density distribution, peaking around 0, but with a relatively broader spread.

Finally, groups M, N, and R show more dispersed density distributions, reflecting a higher degree of intra-group variation along the first coordinate dimension. In particular, groups M and N both show distinguishable peaks overlapping around 0, with M predominantly on the left side and N on the right side. Group R, on the other hand, has its entire density distribution on the negative side.

Answer 2

Figure 2:



```
mep_merged_viz |>
  ggplot(aes(x = ‘NOM−D1‘, y = ‘NOM−D2‘, color = EPG)) +
  geom_jitter(position = position_jitter(width = 0.2, height = 0.2), alpha =
    0.5) +
  scale_x_continuous(limits = c(−1, 1)) +
  scale_y_continuous(limits = c(−1, 1)) +
  scale_color_viridis_d(name = "EP group") +
  labs(
    title = "Nominate coordinates for each Member of the European Parliament",
    x = "\nNominate Coordinate 1",
    y = "\nNominate Coordinate 2"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
```

```r
    plot.title = element_text(hjust = 0.5)
  )
```

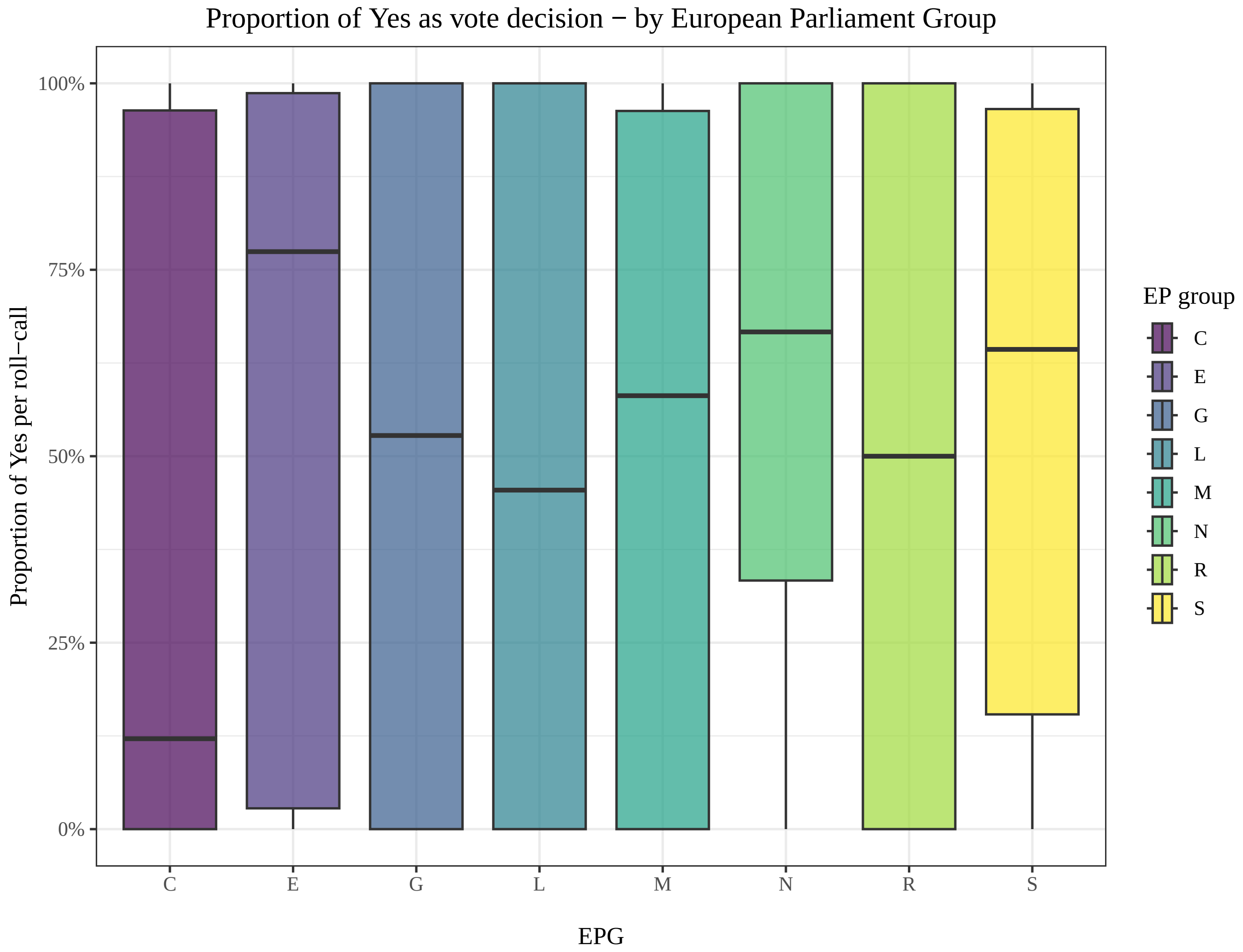
Figure 2 presents the positioning of each Member of the European Parliament along Nominate Coordinates 1 and 2. Some groups form clusters that are easily identifiable, while others show a more scattered distribution, consistent with the previously discussed density patterns along Coordinate 1.

Answer 3

```r
yes_rate_viz <- mep_merged |>
  filter(vote_decision %in% c("Yes", "No", "Abstain")) |>
  group_by(EPG, vote_number) |>
  summarise(
    yes = sum(vote_decision == "Yes"),
    no = sum(vote_decision == "No"),
    abstain = sum(vote_decision == "Abstain"),
    yes_rate = yes / (yes + no + abstain)) |>
  ungroup()
```

To present Figure 3, the merged data were grouped by EP group and roll-call number. For each combination, the proportion of `Yes` votes was computed. From this summarized data, the following boxplots were created.

Figure 3:



```r
yes_rate_viz |>
  ggplot(aes(x = yes_rate, y = EPG, fill = EPG))+
  geom_boxplot(alpha = 0.7) +
  coord_flip() +
  scale_x_continuous(labels = percent) +
  scale_fill_viridis_d(name = "EP group") +
  labs(
    title = "Proportion of Yes as vote decision - by European Parliament Group
  ",
    x = "\nProportion of Yes per roll-call",
    y = "\nEPG"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5)
  )
```

Figure 3 illustrates that, for the majority of parliamentary groups, there is high variability: the groups do not consistently vote cohesively across all policy issues. Groups `G`, `L`, `M`, and `R` especially tend to split in their voting behavior, with all of them having an average proportion of `Yes` votes around 50%.

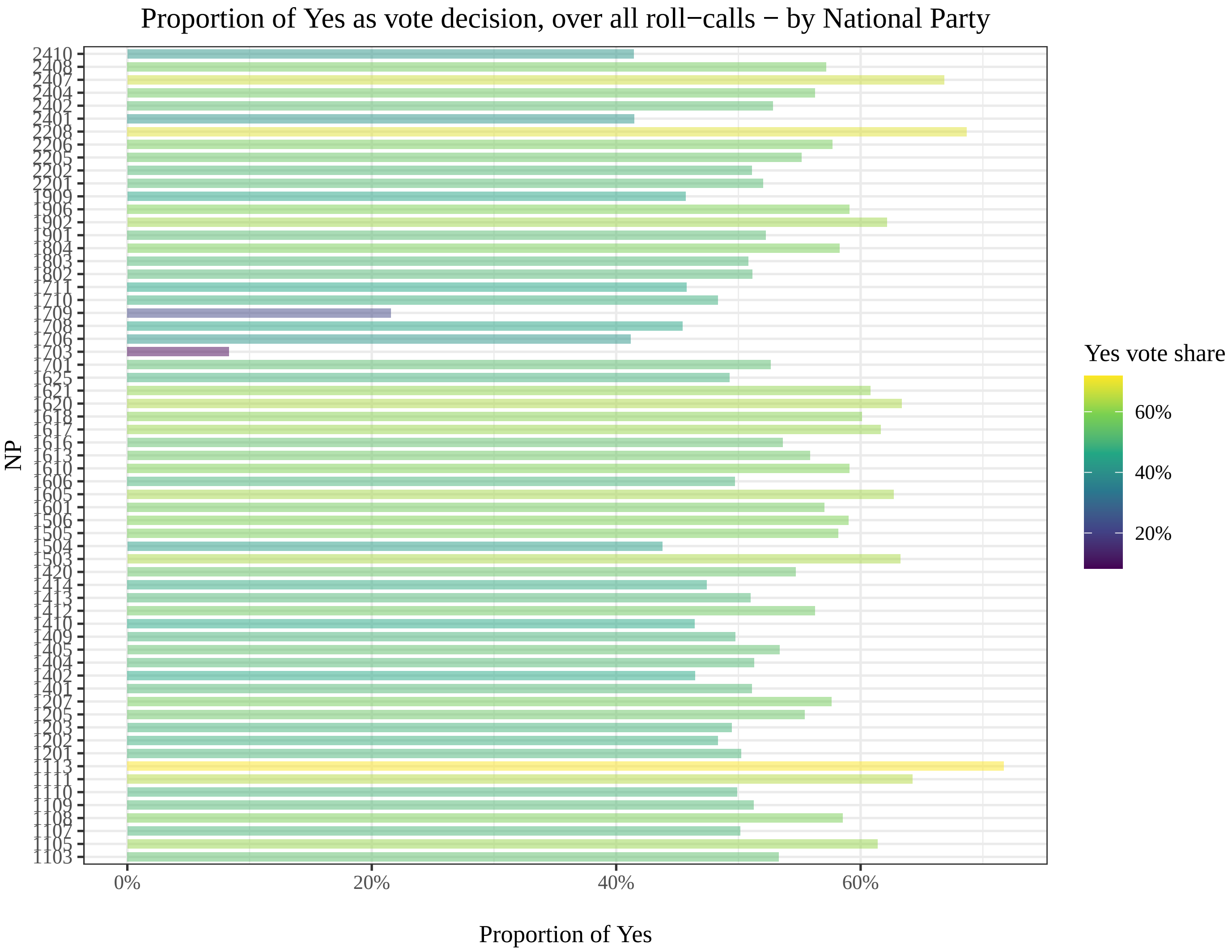
Groups `N` and `S` show narrower boxplots, indicating greater cohesion, with `N` having an average of approximately 60% and a lower quartile around 40%. Groups `C` and `E` are noteworthy as well, displaying average proportions under 12.5% and over 75%, respectively.

Answer 4

```
yes_rate_viz_ng <- mep_merged |>
  filter(vote_decision %in% c("Yes", "No", "Abstain")) |>
  mutate( NP = as.factor(NP)) |>
  group_by(NP) |>
  summarise(
    yes = sum(vote_decision == "Yes"),
    no = sum(vote_decision == "No"),
    abstain = sum(vote_decision == "Abstain"),
    yes_rate_ng = yes / (yes + no + abstain)) |>
  ungroup()
```

In order to obtain the `Yes` proportions needed for the creation of Figure 4, the merged data were grouped by national party. For each subgroup, the sums of `Yes`, `No`, and `Abstain` votes across all roll-call votes were calculated, from these figures the share of `Yes` votes was computed.

Figure 4:



```r
yes_rate_viz_ng |>
ggplot(aes(x = NP, y = yes_rate_ng, fill = yes_rate_ng)) +
  geom_col(alpha = 0.5, width = 0.7) +
  coord_flip() +
  scale_fill_viridis_c(name = "Yes vote share",
                       labels = percent) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Proportion of Yes as vote decision, over all roll-calls - by National Party",
    y= "\nProportion of Yes",
    x = "\nNP"
  ) +
  theme_bw() +
  theme(
    text = element_text(family = "Times"),
    plot.title = element_text(hjust = 0.5)
  )
```

For graphical interpretation, the bars were filled according to each national party's percentage of `Yes` votes.
The national parties `1703` and `1709` have notably lower proportions of `Yes` votes, with the former being below 10%. On the opposite end, the national parties `1113`, `2208`, and `2407` have quite high shares of `Yes` votes, with the first exceeding 70%.
Most of the other national parties, however, have shares ranging between 40% and 60%.