Answer Sheet - Problem 1 - Nour Kebbi - 23350337

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)

```
#Data Manipulation Part 1
library(tidyverse)
lapply(c("tidyverse", "ggplot2"),  pkgTest)
library(readxl)

#loading datasets
mep_info <- read_excel("/Users/nourkebbi/Documents/GitHub/DataViz_2026/problemSets/
rcv_info <- read.table("/Users/nourkebbi/Documents/GitHub/DataViz_2026/problemSets/
```

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

   My Answer: On the mep info data the key variables are the MEP id, MEP name, national part which is the NP, and their estimated ideological positions NOM-D1 and NOM-D2. The unit analysis for the 'MED id', 'Order in EP5 rcv' and 'National Party' are Integers, while 'NOM-D1' and 'NOM-D2' are floats. The unit analysis for the 'Name', 'Member State' and 'EP Group' are strings. The measure is the individual Member of Euorpean Parliment which is MEP. On the rcv ep1 data the key variables are the MEP id and thee Votes which are in columns V1 to Vn. The unit analysis for the 'MEPNAME', 'MS' and 'EPG' are Strings, while 'MEPID' and all 'V's are Integers. The measure is also the individual Member of European Parliment.

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.

     ```
     #Data Manipulation Part 3A
     rcv_tidy <- rcv_info %>%
       pivot_longer(
         cols = !c(MEPID, MEPNAME, MS, NP, EPG),
         names_to = "vote_id",
         values_to = "decision"
       )
     mep_info <- mep_info %>% #renaming the column so I can join
       rename(MEPID = ‘MEP id‘)
     mep_info <- mep_info %>% #casting it into an INT so I can join
       mutate(MEPID = as.integer(MEPID))
     ```

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

   ```
   #Data Manipulation Part 3B
   summary_table <- rcv_tidy %>%
     group_by(decision) %>%
     summarize(count= n())
   ```

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.

   ```
   #Data Manipulation Part 4
   joint_table <- left_join(rcv_tidy, mep_info, by = "MEPID")
   colSums(is.na(joint_table))
   ```

```
joint_table <- joint_table %>%
  drop_na()
```

5. Compute, for each EP group in EP1:

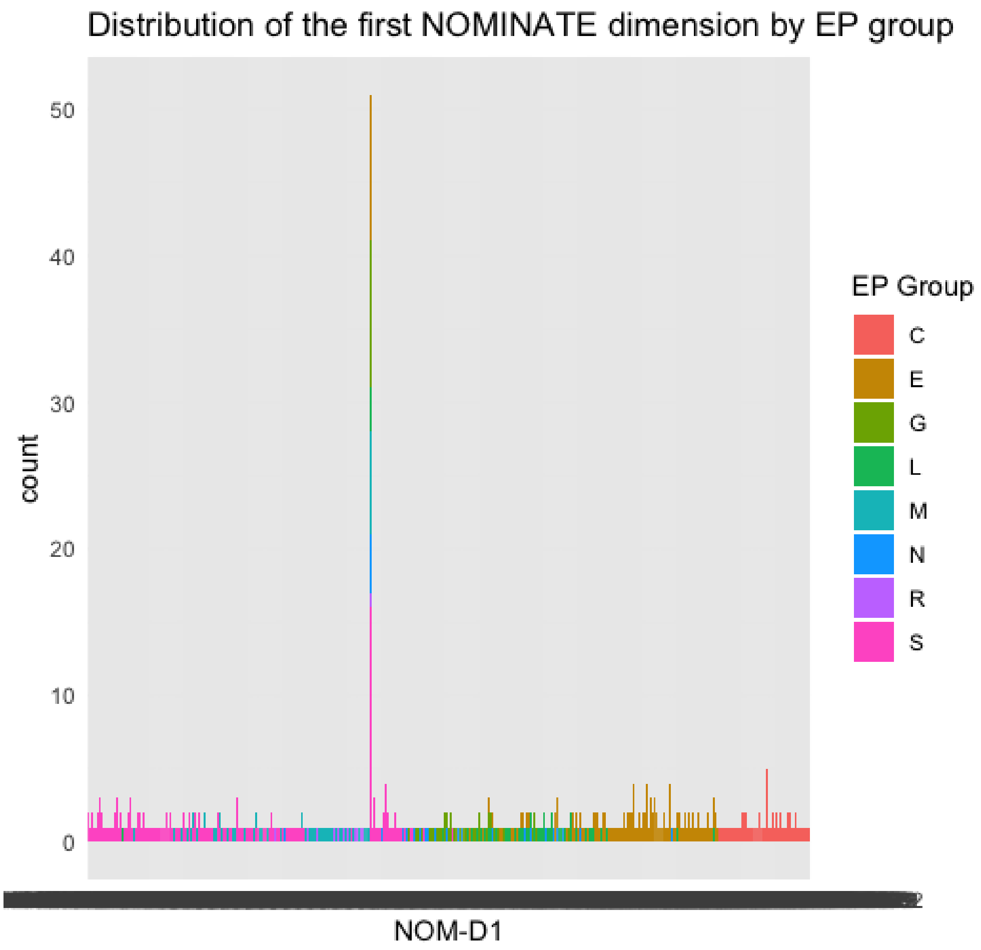
   - The mean rate of Yes votes (Yes over Yes+No+Abstain) across all roll calls.
   - The mean abstention rate.
   - The mean vote preferences along the two contested dimensions (NOM-D1 and NOM-D2).

     ```
     #Data Manipulation Part 5
     avg_table <- joint_table %>%
       group_by(EPG) %>%
       summarize(
         mean_yes_votes = mean(decision == 1)/mean(decision %in% c(1,2,3)),
         mean_abstention = mean(decision == 3),
         mean_nomd1 = mean(‘NOM-D1‘, na.rm = TRUE),
         mean_nomd2 = mean(‘NOM-D2‘, na.rm = TRUE))
     ```

Data Visualization

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

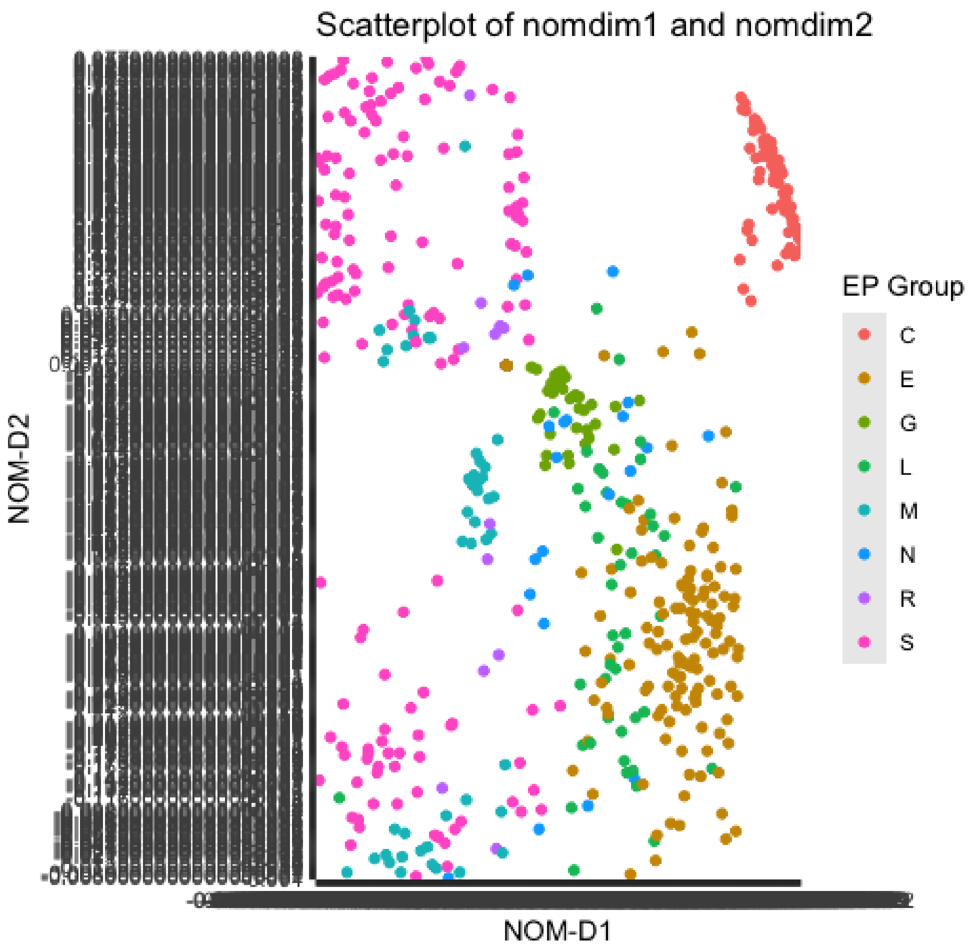
   ```
   #Data Visualization Part 1
   ggplot(mep_info, aes(x = ‘NOM-D1‘, fill = ‘EP Group‘)) +
     geom_bar() +
     labs(title = "Distribution of the first NOMINATE dimension by EP group")
   ```



There is no consistency but the Nominate Dimensions are concentrated in the middle.

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

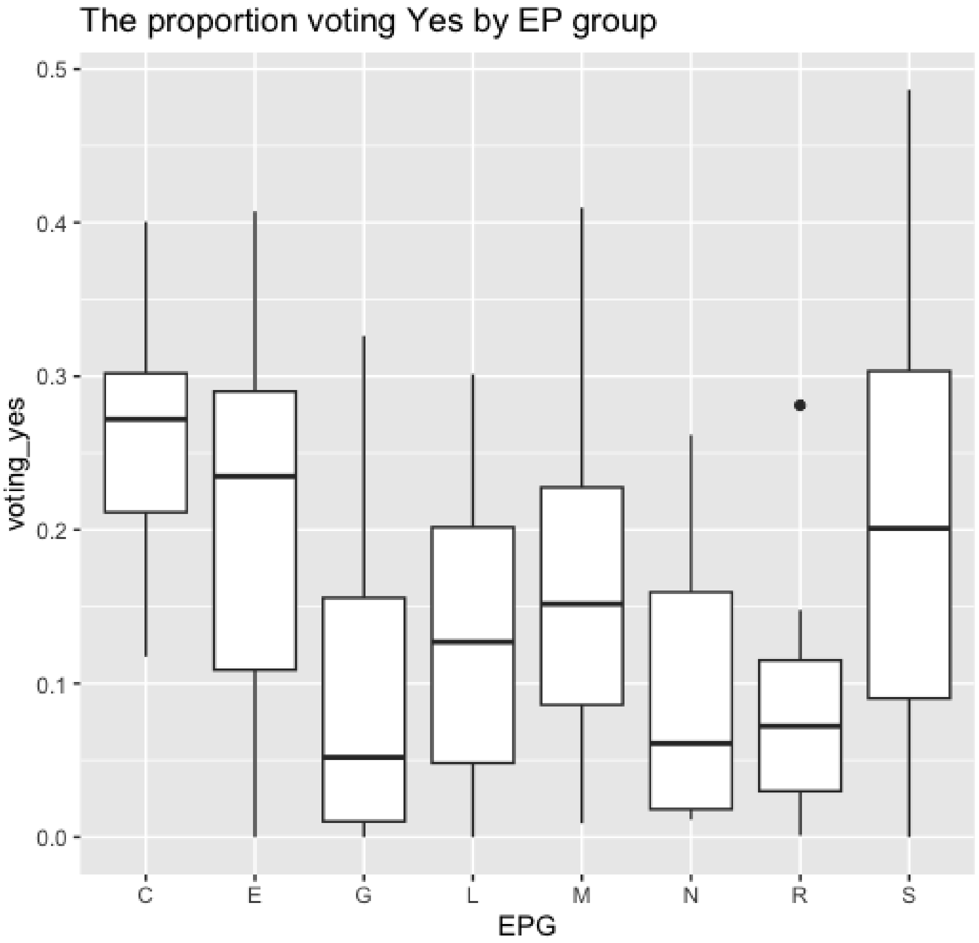
```
#Data Visualization Part 2
ggplot(mep_info, aes(x = ‘NOM-D1‘, y = ‘NOM-D2‘, color = ‘EP Group‘)) +
  geom_point() +
  labs(title = "Scatterplot of nomdim1 and nomdim2")
```



It looks like tthe EP Groups C and E and S are concentrated.

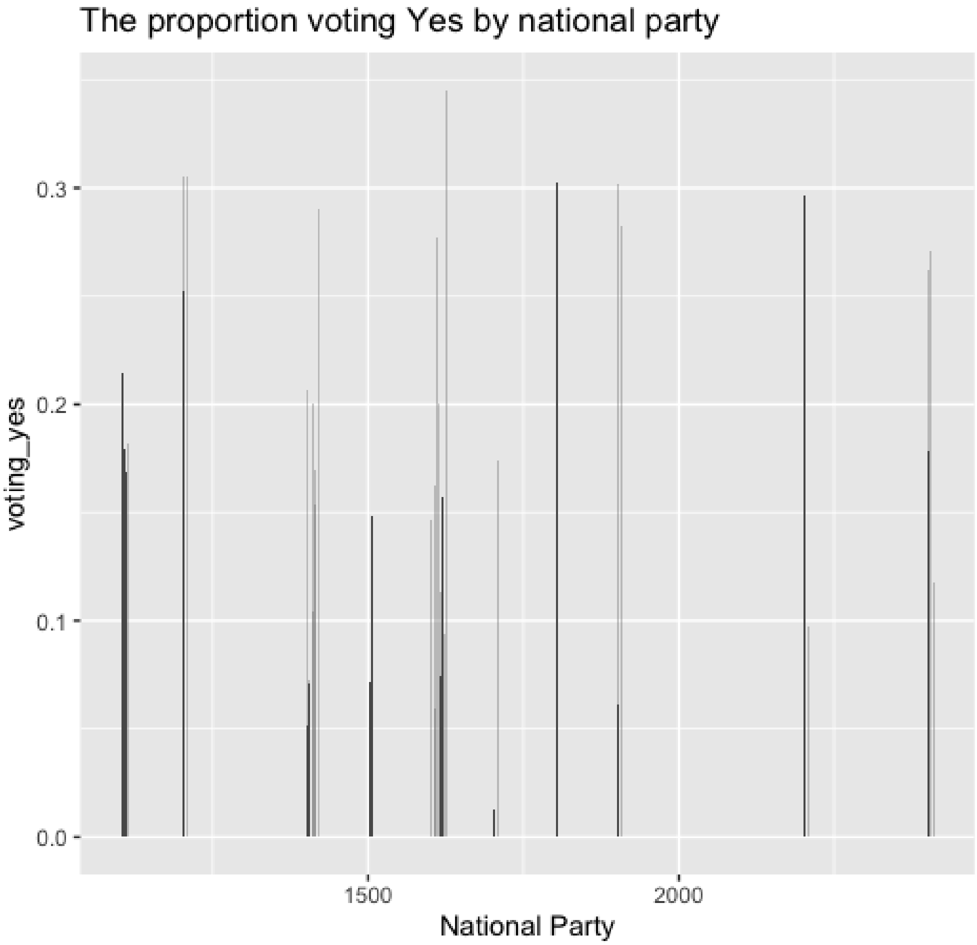
3. Produce a boxplot of the proportion voting *Yes* by EP group to visualize cohesion.

```
#Data Visualization Part 3
dv_part3 <- joint_table %>%
  group_by(MEPID, EPG) %>%
  summarize(voting_yes = mean(decision == 1, na.rm = TRUE))
ggplot(dv_part3, aes(x = EPG, y = voting_yes)) +
  geom_boxplot() +
  labs(title = "The proportion voting Yes by EP group")
```



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

```
#Data Visualization Part 4
dv_part4 <- joint_table %>%
  group_by(‘National Party‘) %>%
  summarize(voting_yes = mean(decision == 1, na.rm = TRUE))
ggplot(dv_part4, aes(x = ‘National Party‘, y = voting_yes)) +
  geom_bar(stat = "identity") +
  labs(title = "The proportion voting Yes by national party")
```



5. For each EP group, calculate the average *Yes* share per year and plot a line graph. I did not do this set