Homework 5: Self-Organizing Maps

MACS 30100: Perspectives on Computational Modeling
University of Chicago

Boseong Yun

Overview

For each of the following prompts, produce responses *with* code in-line. While you are encouraged to stage and draft your problem set solutions using any files, code, and data you'd like within the private repo for the assignment, *only the final, rendered PDF with responses and code in-line will be graded.*

A Computational Social Science Problem

This final problem set will be a bit different. I am interested in how you address a social science problem with a computational social science workflow given very general guidelines. The goal in this final problem set, then, is to evaluate your ability to take a general set of instructions, and develop a technically correct solution that allows for substantively insightful inferences. This marries the *computational* with the *social science* parts of the program and course, informed by some functional programming skills you have developed.

Your task in this final problem set is to *design and implement a well-rounded self-organizing map analysis to mine public opinion data on 14 questions from the 2016 ANES.* Using these data we've encountered a bit to this point, you will develop your own solution, which requires selection of packages you think are best for completing the task, whether covered in class or not.

You will be evaluated on your ability to accomplish a task using both new and more familiar tools. On the substantive side, you will be mining the ANES data for evidence of whether there are likely to be *partisan* differences in public opinion, where *public opinion* is defined here as responses to 14 survey questions on salient social issues. For simplicity, you may treat "partisan" as three levels, including the two major US political parties and all others (Democrat, Republican, and Other).

As I am giving you only a general set of prompts to guide your process, the following rubric in addition to a technical set of solutions will be used to grade this problem set:

Preprocessing (10 points)	**Exploration** (15 points)	**Modeling** (25 points)	**Validation** (25 points)	**Writing & Programming** (25 points)
Designed and implemented appropriate techniques to get the data in a form that is usable for analysis	Explored the space fully (numeric, viz, etc.) prior to fitting models or training algorithms, resulting in a clear rendering of the space	Fit the correct model with all hyperparameters tuned appropriately, and *all* decisions throughout sufficiently defended	Validated results appropriately, and clearly dug deep and beyond the main model to defend and explain recovered patterns	Proper writing with excellent grammar (spelling, etc.) and thorough responses; elegant and *replicable* code

The Social Issue Questions

Below are the 14 social issue questions along with scales and variable code names to be used in the analysis. Question wording and response categories were copied and pasted from the ANES 2016 Pilot Study Questionnaire.

- `vaccine` - "Do you favor, oppose, or neither favor nor oppose requiring children to be vaccinated in order to attend public schools?" (7 point from favor a great deal (1) to oppose a great deal (7))
- `autism` - "How likely or unlikely is it that vaccines cause autism?" (6 point from Extremely likely (1) to Extremely unlikely (6))
- `birthright_b` - "Do you favor, oppose, or neither favor nor oppose children of unauthorized immigrants automatically getting citizenship if they are born in this country?" (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `forceblack` - "How often do you think police officers use more force than is necessary under the circumstances when dealing with BLACK people?" (5 point from Never (1) to Very often (5))
- `forcewhite` - "How often do you think police officers use more force than is necessary under the circumstances when dealing with WHITE people?" (5 point from Never (1) to Very often (5))
- `stopblack` - "How often do to think police officers stop BLACK people on the street without a good reason?" (5 point from Never (1) to Very often (5))
- `stopwhite` - "How often do to think police officers stop WHITE people on the street without a good reason?" (5 point from Never (1) to Very often (5))
- `freetrade` - "Do you favor, oppose, or neither favor nor oppose the U.S. making free trade agreements with other countries?" (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `aa3` - "Do you favor, oppose, or neither favor nor oppose allowing universities to increase the number of underrepresented minority students studying at their schools by considering race along with other factors when choosing students?" (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `warmdo` - "Do you think the federal government should be doing more about rising temperatures, should be doing less, or is it currently doing the right amount? (7 point from Should be doing a great deal more (1) to Should be doing a great deal less (7))
- `finwell` - "Do you think people's ability to improve their financial well-being is now better, worse, or the same as it was 20 years ago?" (7 point from A great deal better (1) to A great deal worse (7))
- `childcare` - "Do you favor an increase, decrease, or no change in government spending to help working parents pay for CHILD CARE when they can't pay for it all themselves?" (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `healthspend` - "Do you favor an increase, decrease, or no change in government spending to help people pay for HEALTH INSURANCE when they can't pay for it all themselves?" (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `minwage` - "Should the minimum wage be raised, kept the same, lowered but not eliminated, or eliminated altogether?" (4 point from Raised [1], Kept the same [2], Lowered [3], Eliminated [4])

The Task

Here are the prompts to guide your task. Again, **there is no single way this problem set should be executed**. Simply do your best, leveraging all tools and techniques we have covered, and most importantly defend ***all*** choices you make throughout the process so you can at least earn partial credit where appropriate

1. Read in the 2016 ANES data we have been using (`anes_2016.csv`), and create a subset of the data containing *at least* the main 14 questions/features (from above) as these are the core of the analysis.

```r
# Global Knit Options
knitr::opts_chunk$set(message = FALSE, error = FALSE, warning = FALSE)

# Laoding the librareies
library(tidyverse)
library(tidymodels)
library(corrplot)
library(here)
library(skimr)
library(GGally)
library(amerika)
library(tictoc)
library(kohonen)
library(e1071)
library(RColorBrewer)
```

```r
# Reading the data
anes <- read_csv(here("data/anes_2016.csv"))

# Selected Variables
vars <- c("vaccine", "autism", "birthright_b", "forceblack", "forcewhite",
          "stopblack", "stopwhite", "freetrade", "aa3", "warmdo",
          "finwell", "childcare", "healthspend", "minwage", "pid3")

# Glboal Seed
set.seed(77)

# Theme set
theme_set(theme_bw())
```

2. Preprocess and clean the data.

- I have created the party variable with three levels (Republican, Democrat, Others per the instructions) using the pid3 variable. The pid3 variable has 5 values where 1 = Democrat, 2 = Republican, 3 = Indepdent, 4 = Other, and 5 = Not Sure. Since the instruction of the problem set askes us to have three groups comprised of Republicans, Democrats, and Others, I transform the variable in the following way where value greater than 3 are transfomred into Others:
- pid3 == 1 ~ "Democrat"
- pid3 == 2 ~ "Republican"
- pid3 >= 3 ~ "Others"

Afterwards, I filter out missing and unidentified responses in the 14 social questions. Specifically, I filter out 9 for *birthright_b* and *aa3* and filter out 8 for *stopwhite*, *warmdo*, *finwell*, *healthspend*, and *minwage*. In these question, the value of 8 and 9 refer to **Skipped** and **Not Asked** respectively according to the ANES Pilot 2016 Codebook. I have filtered out these value for two primary reasons. First of all, **Skipped** and **Not Asked** does not provide meaningful contribution to the analysis. Secondly, the responses have ordinal scale and thus the value of 8 and 9 can skew the resutls to the right. For instance, we can make inaccurate conclusions that Democrats have strong views on certain social question when in fact they just represent missing value.

Although it is also a good option to impute these value, it has risks of introducing some bias. In the techincal aspect, the variables are categorical variables and thus the mean imputation that results in non-integer values can be inappropriate. Other methods such as median and mode imputation can also be inappropriate because we exepct the partisan differences to manifest in the concentration of certain response (1 or 7) for certain

questions and thus using the median or mode imputation can bias the results. Although other alternatives might be available such as knn or regression imputation, I believe that dropping the values can provide unbaised results and I thus proceed my analysis with data where all inappropriate values have been dropped.

```
# Reading the data

# pid3
anes_cleaned <- anes %>%
  select(vars) %>%
  drop_na() %>%
  mutate(party = ifelse(pid3  == 1, "Democrat",
                        ifelse(pid3 == 2, "Republican", "Others"))) %>%
  mutate(party = factor(party, levels = c("Democrat", "Republican", "Others"))) %>%
  select(party, everything(), -pid3)


# Removing Missing & Unidentified Values:
# birthright_b: 9, aa3: 9, stopwhite: 8, warmdo: 8,
# finwell: 8, healthspend: 8, minwage: 8
anes_cleaned2 <- anes_cleaned %>%
  filter_at(vars(birthright_b, aa3), all_vars(. != 9)) %>%
  filter_at(
    vars(stopwhite, warmdo, finwell, healthspend, minwage),
    all_vars(. != 8)
    )

# Imputation is also an option!
#anes_cleaned2 <- anes_cleaned %>%
#  mutate_at(vars(-party), ~ ifelse(. == 9, NA, .)) %>%
#  mutate_at(vars(-party), ~ ifelse(. == 8, NA, .)) %>%
#  group_by(party) %>%
#  mutate_at(vars(-party), funs(replace(., which(is.na(.)), mode(., na.rm=TRUE)))) %>%
#  ungroup()
```

3. Explore the data using any approach(es) or tool(s) you think best, such as feature-level correlations, boxplots, scatterplots, density plots, etc.

- **Interperation**
- First of all, the skim function provides numerical summary of the variables with the histogram for each variable. It is possible to determine that the value for vaccine and minwage are skewed to the left. That is, many people in the datatset seem to think that the minimum wage must be raised and that children must be vaccinated. Additionally, it noteworthy that forcewhite, stopwhite, and minwage variables have less variation than other variables. The distribution of other variables seem to be more or less equally distributed across many responses.
- Figure 1 further provides infomation about the proportions of responses for each question by party affiiation. This plot shows how people with certain party affiliation respond differently to each question. For Democrats, they are likely to answer 1 healthspend, minwage, and warmdo questions while Republicans are less likely to answer 1 for the same questions. For Republicans, they are likely to answer 7 for aa3, birthright_b, and childcare questions when Democrats are less likely answer for the same questions. For Others, their responses to the questions are more or less equally distributed and do not display noticeable patterns. Overall, it is possible to find out that healthspend, minwage, warmdo, aa3, birthright_b, and childcare questions can provide information about partisan differences. Specifically, we can see that Democrats and Republicans have opposing views on these questions and thus signify potential partisan differences. For questions on free trade and vaccine, the responses seemed uniform

across party affiliation. Linking these responses to the reality, we know that Democrates have strong supporting views on healthspend, minwage, and warmdo (global warmning). Likewise, Republicans have strong opposing views against aa3 (affirmative action), birthright_b(illegal immigration), and child care. That is, we see clear partisan differences along these questions.

- Although some of the information overlap with the previous plots and summary tables, this boxplot visualizes the distribution of the responses for each question by party affiliation. Rather than showing the dominant response by each party affiliation, we can see general trend or leaning of people with certain party affiliation. For instance, it is possible to see that the response of Democrats are skewed to the left for healthspend, minwage, and warmdo questions with less variance than their responses to other questions. In similar contexts, we can see that the response of Republicans are skewed to the right for aa3, birthright_b, and childcare. However, it is possible to see that the variance of their quesitons are realtively wide. These information seem to tell us that Democrats have strong ideas on certain question whereas Republicans seem to have moderately strong ideas on certain questions.
- The final correlation plot shows the relationship among the variables. This can help us determine to what extent the variables are related to each other and thus can influence partisan differences in similar ways. For instance, it is possible to find out that there is a strong degree of positive correlation between healthspend and minwage. It is also possible to find out that healthspend is positively correlated with warmdo. Going back to our previous findings, we can confirm that many of these variables are highly likely to be associated with Democrats. That is, we can find out that a set of questions do show partisan differences! Although the corrleation for variables where we witnesed concentration of republican reponses are less clear, it is possible to see that reponses and partisan differences are related to each other.

```
# Brief Survey of the dataset
skim(anes_cleaned2) %>%
  select(-numeric.p25, -numeric.p75, -complete_rate)
```

Table 2: Data summary

Name	anes_cleaned2
Number of rows	314
Number of columns	15
_____	
Column type frequency:	
factor	1
numeric	14
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	ordered	n_unique	top_counts
party	0	FALSE	3	Dem: 128, Oth: 114, Rep: 72

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p50	p100	hist
vaccine	0	2.40	1.82	1	1.5	7	
autism	0	4.55	1.52	1	5.0	6	
birthright_b	0	4.19	2.19	1	4.0	7	
forceblack	0	3.62	1.03	1	4.0	5	

skim_variable	n_missing	mean	sd	p0	p50	p100	hist
forcewhite	0	2.68	0.87	1	3.0	5	
stopblack	0	3.69	1.06	1	4.0	5	
stopwhite	0	2.62	0.94	1	3.0	5	
freetrade	0	3.91	1.67	1	4.0	7	
aa3	0	4.57	1.95	1	4.0	7	
warmdo	0	3.25	2.11	1	3.0	7	
finwell	0	4.66	1.66	1	5.0	7	
childcare	0	3.43	1.86	1	3.0	7	
healthspend	0	3.30	1.94	1	3.0	7	
minwage	0	1.62	0.92	1	1.0	4	

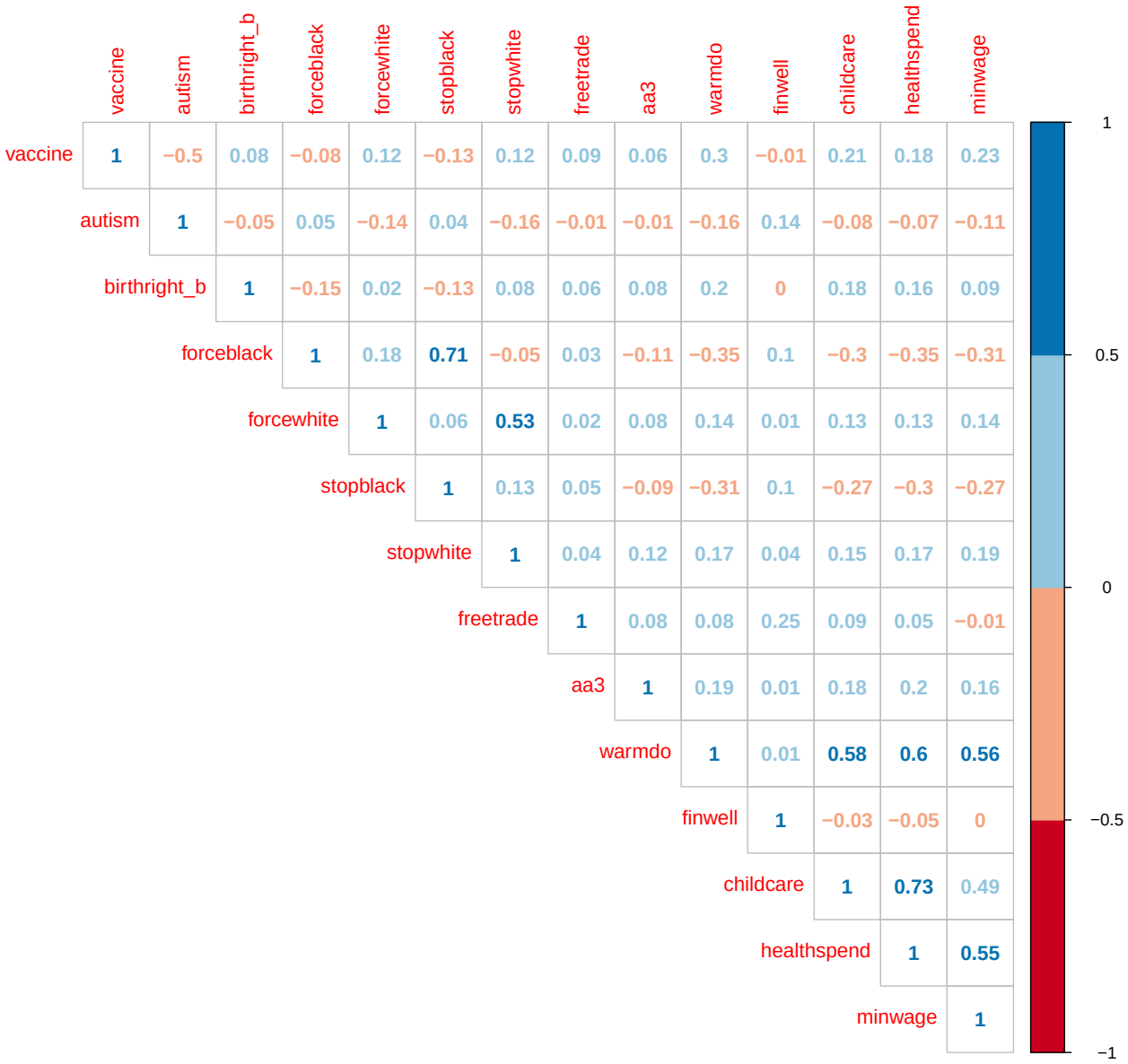
```r
# Exploratory Data Analysis 1: Bar Plot (Show the frequency of values)
anes_cleaned2 %>%
  pivot_longer(-party) %>%
  count(party, name, value) %>%
  group_by(party, name) %>%
  mutate(
    total = sum(n),
    prop = n / sum(n),
    party = factor(party, levels = c("Republican", "Others", "Democrat"))
    ) %>%
  ggplot(aes(x = value, y = prop, fill = party)) +
  geom_col(position = "dodge") +
  scale_x_continuous(breaks = pretty_breaks()) +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~name, scales = "free") +
  labs(
    title = "Figure 1: The Prportions of Reponses for Each Question by Party Affiliation",
    fill = "Party",
    x = "Response",
    y = ""
  )
```

Figure 1: The Prportions of Reponses for Each Question by Party Affiliation


```
# Exploratory Aanalysis 2: Box Plot (Show the distribution of values)
anes_cleaned2 %>%
  pivot_longer(-party) %>%
  mutate(party = factor(party, levels = c("Republican", "Others", "Democrat"))) %>%
  ggplot(aes(x = value, y = party, fill = party)) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(~name, scales = "free") +
  scale_x_continuous(breaks = pretty_breaks()) +
  labs(
    title = "Figure 2: The Distribution of Response for Each Question by Party Affiliation",
    x = "Response",
    y = ""
  )
```

Figure 2: The Distribution of Response for Each Question by Party Affiliation


```
# Exploratory Analysis 3: Correlation Plot (Shows the relationship between variables)
anes_cleaned[, -1] %>%
  cor() %>%
  corrplot(method = "number", type = "upper",
           col = brewer.pal(n = 4, "RdBu"))
```



4. Construct and present a self-organizing map (SOM) of the *question space*. **Think carefully about the scale of response categories, as these vary across questions.** Also, remember to tune the relevant hyperparameters appropriately. You might consider the `kohonen` package in R (though there are many others), or the `minisom` package in Python. A response to this may include creating grids, fitting models, and creating (multiple) visualizations of the results

- Answer: As indicated by the question, my reponse to this include grids, fitting models, and creating (multiple) visualization of the results. I provide substantive discussion of these results in my answer to the question (5).

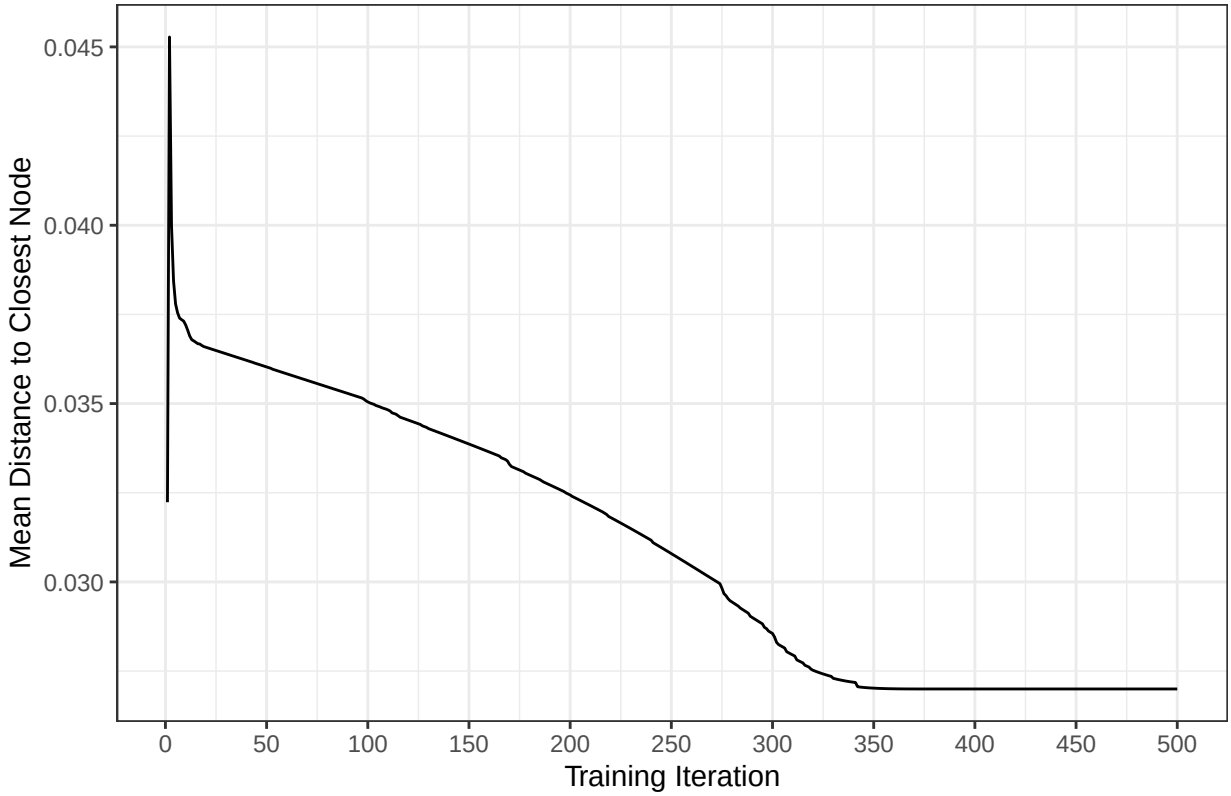
```
# scaling the data
anes_scaled <- anes_cleaned2[, 2:ncol(anes_cleaned)] %>%
  scale()

# create the structure of the output layer
# Notes: changing the grid can help us detect partisan differences in a clear manner
# I have chosen 10 as the grid size because the size of 10 does capture the differences
```

```r
# well!
search_grid <- somgrid(xdim = 10,
                       ydim = 10,
                       topo = "rectangular",
                       neighbourhood.fct = "gaussian")

# Fitting the som
{
  tic()
som_fit <- som(anes_scaled,
               grid = search_grid,
               alpha = c(0.1, 0.001),
               radius = 1,
               rlen = 500,
               dist.fcts = "euclidean",
               mode = "batch")
  toc()
}
```

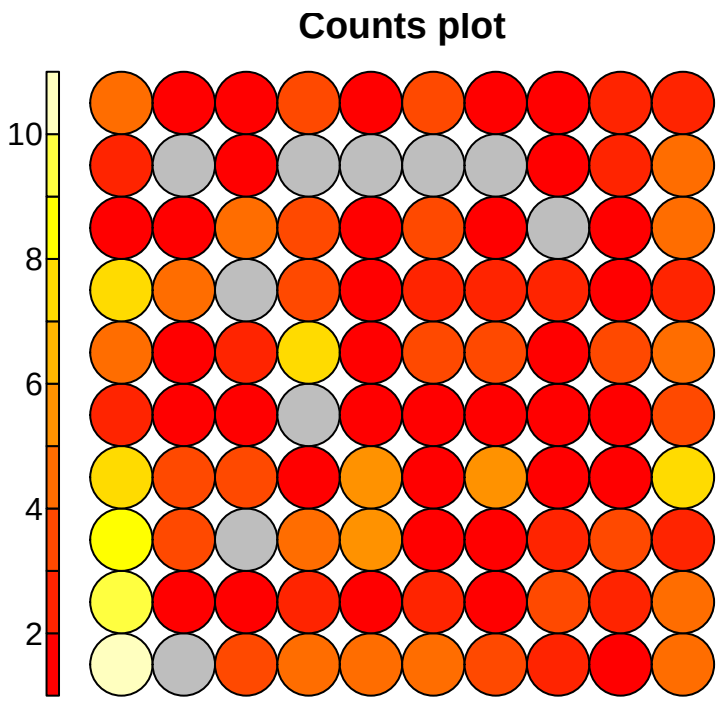
```
## 0.644 sec elapsed
```

```r
# Plot1: The Change in Distance to Closest Mode by Training Iteration
som_fit$changes %>%
  as_tibble() %>%
  mutate(changes = V1,
         iteration = seq(1:length(changes))) %>%
  ggplot(aes(iteration, changes)) +
  geom_line() +
  labs(
    title = "Figure 1: The Change in Distance to Closest Mode by Training Iteration",
    x = "Training Iteration",
    y = "Mean Distance to Closest Node"
    ) +
  scale_x_continuous(breaks = seq(0, 500, by = 50))
```

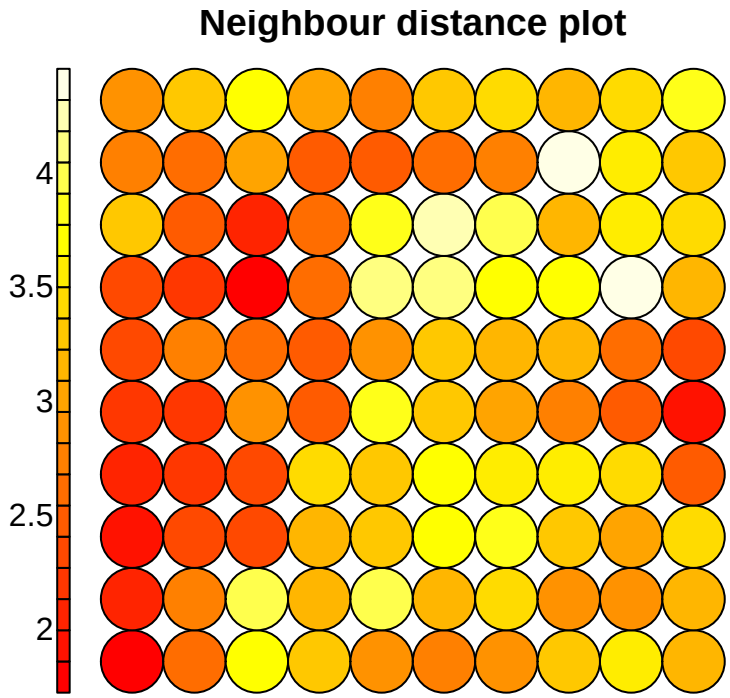
Figure 1: The Change in Distance to Closest Mode by Training Iteration



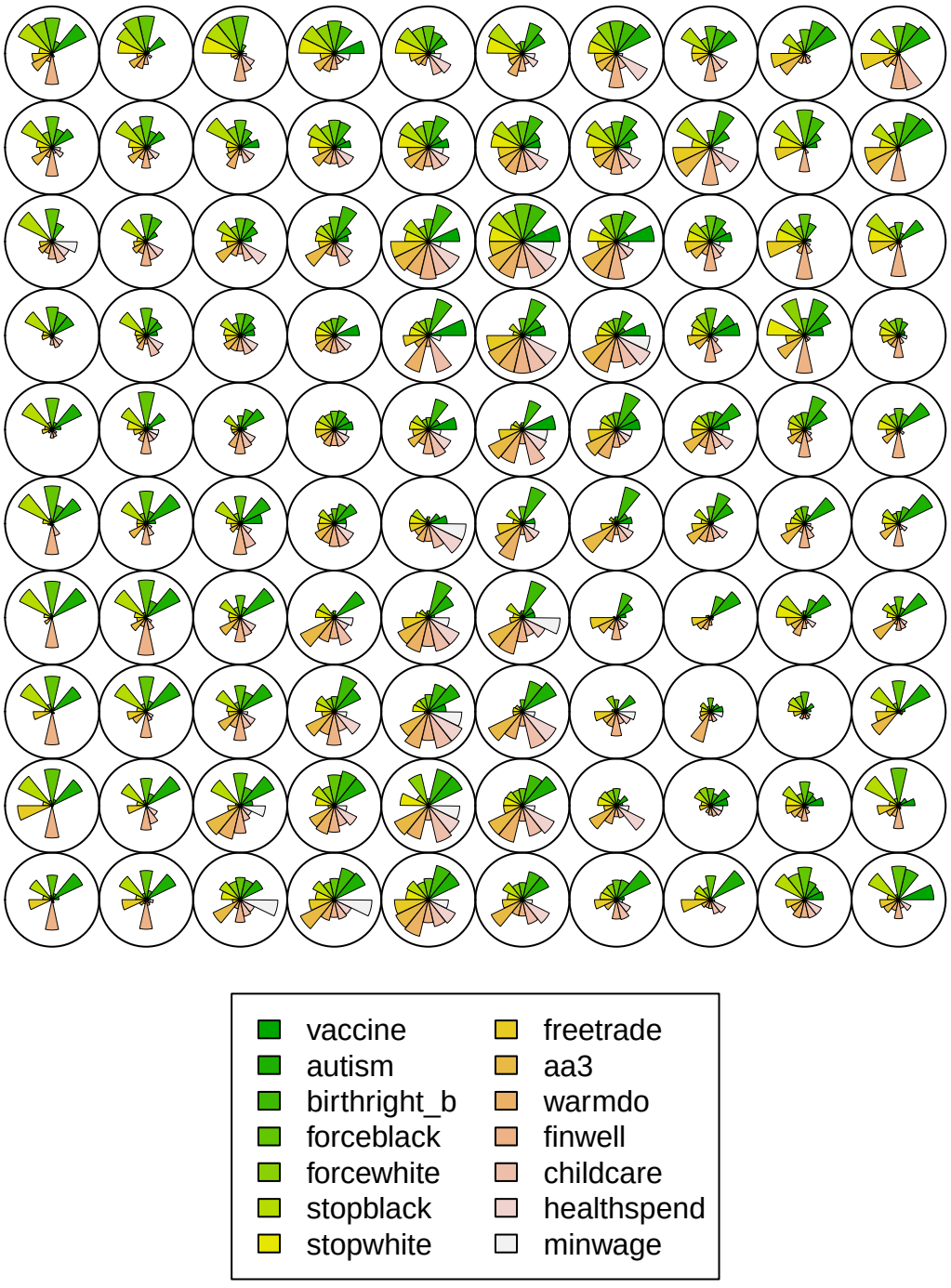

```
# Count Plot
plot(som_fit, type = "counts")
```




```
# Neighbour Distance Plot
plot(som_fit, type = "dist.neighbours")
```

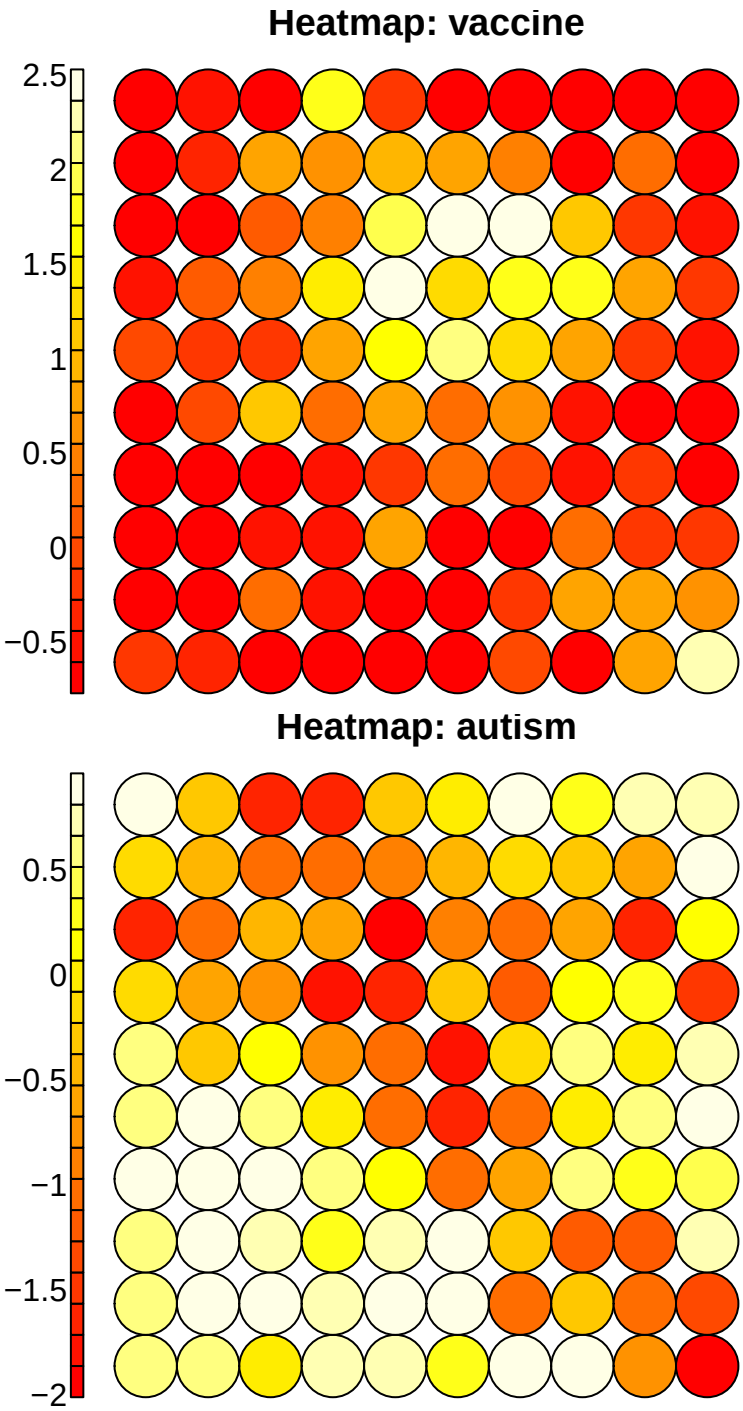


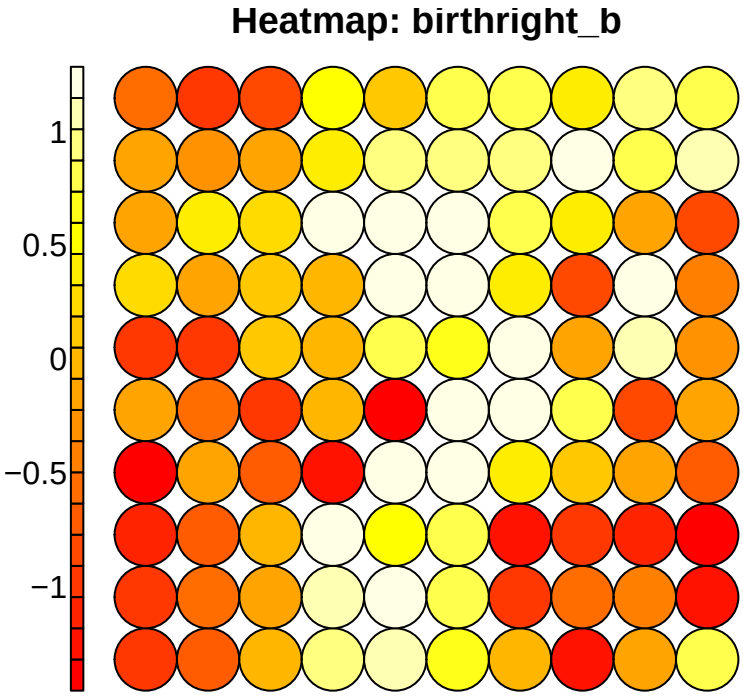

```
plot(som_fit, type = "codes")
```

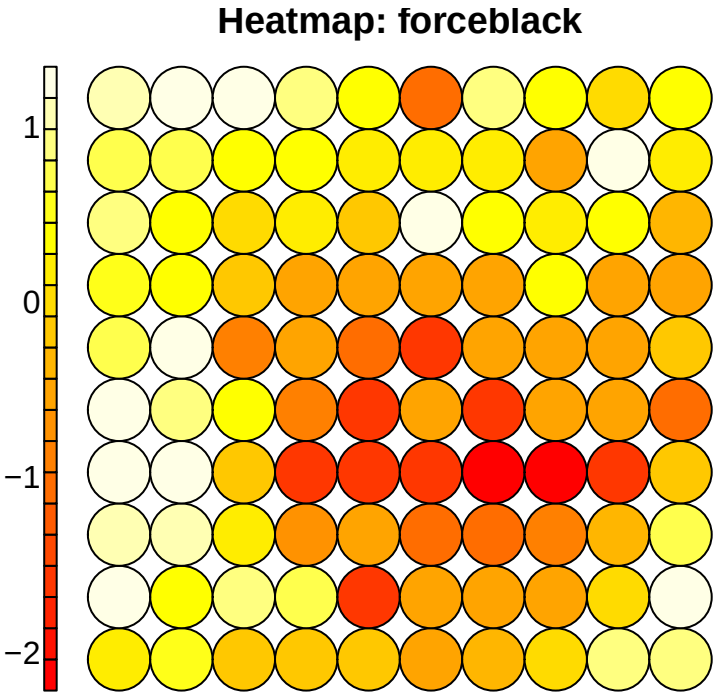


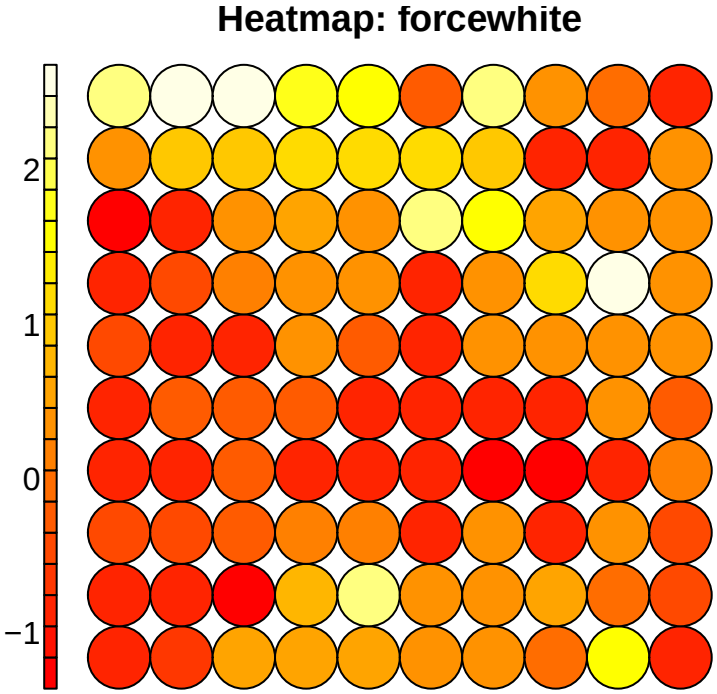

```
# Plot Propert
plot_heat <- function(model) {
  for(i in 1:ncol(anes_scaled)) {
    plot(
      model,
      type = "property",
      property = getCodes(model)[,i],
```

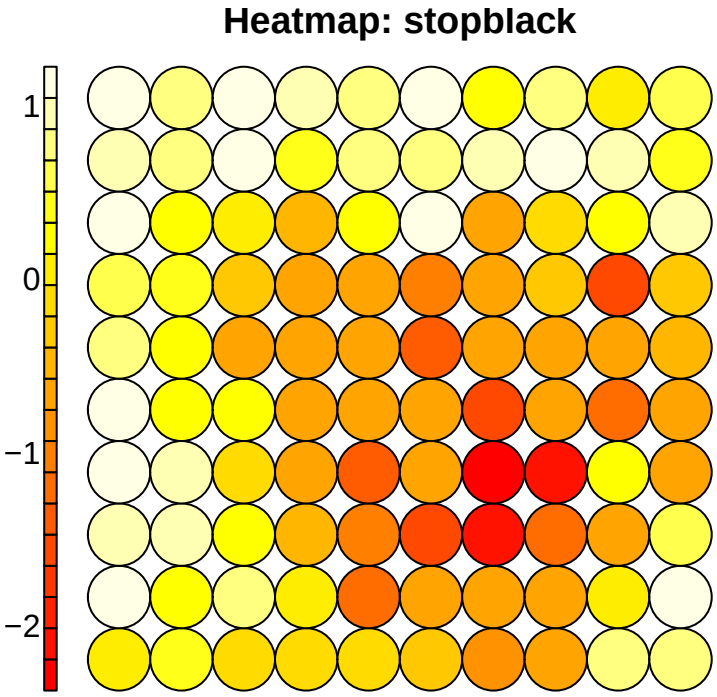
```
      main = paste("Heatmap:", colnames(getCodes(model))[i])
    )
  }
}

#  Plotting the heatmap!
plot_heat(som_fit)
```

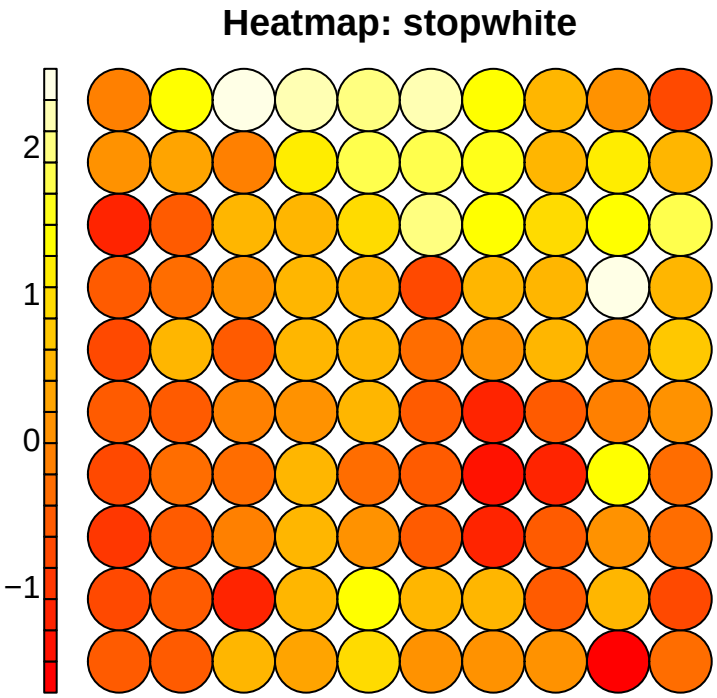


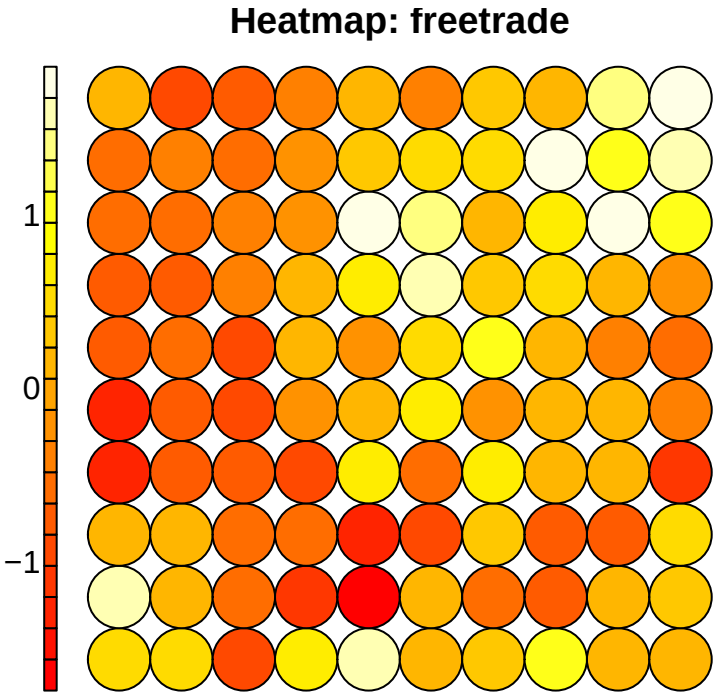

Heatmap: birthright_b
1
0.5
0
−0.5
−1

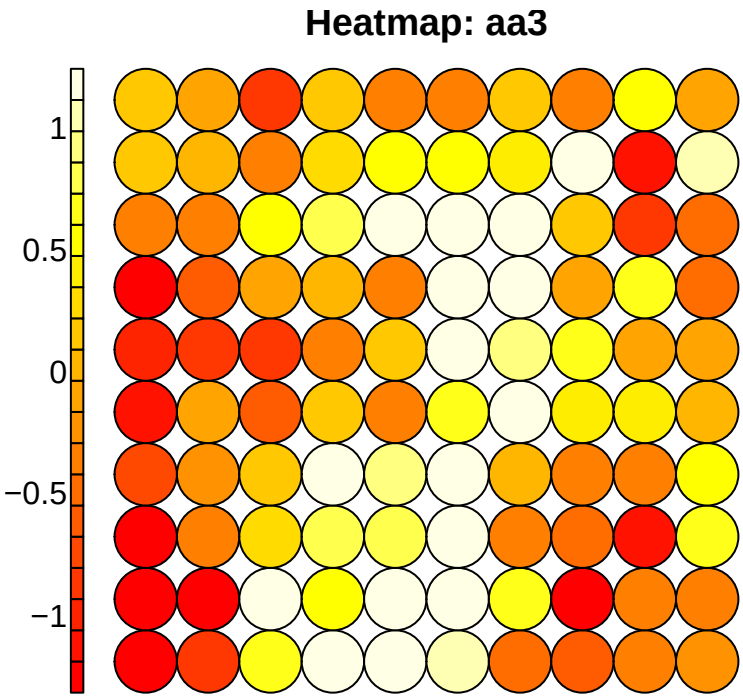

Heatmap: forceblack
1
0
−1
−2

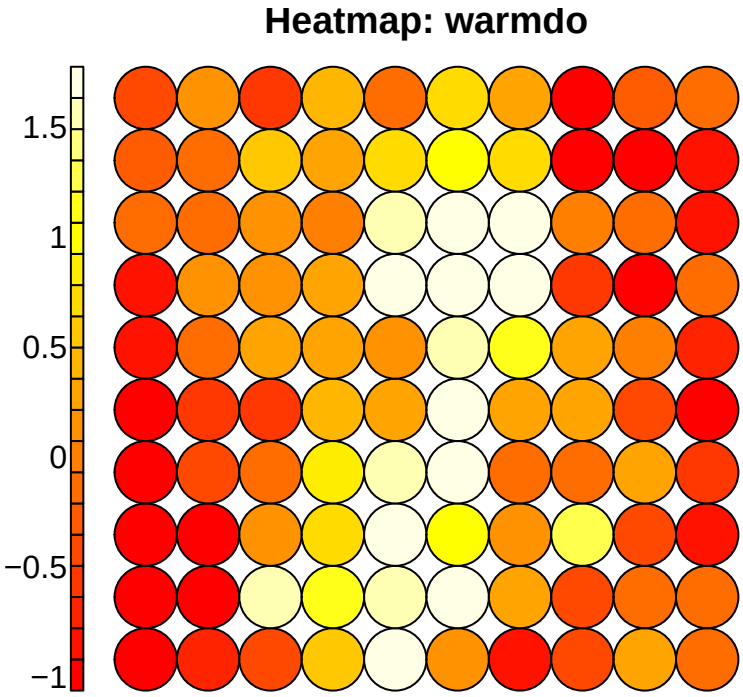

Heatmap: forcewhite
2
1
0
−1

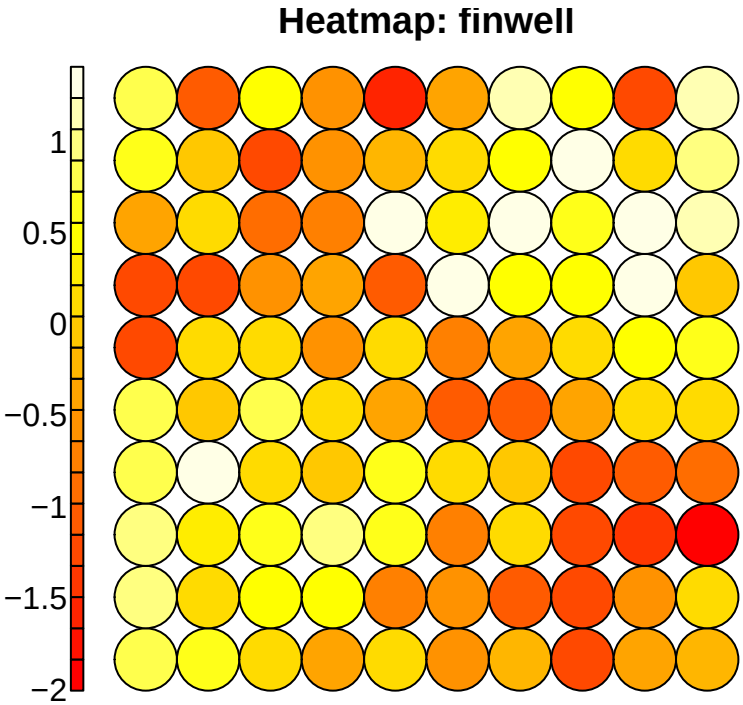

Heatmap: stopblack
1
0
−1
−2

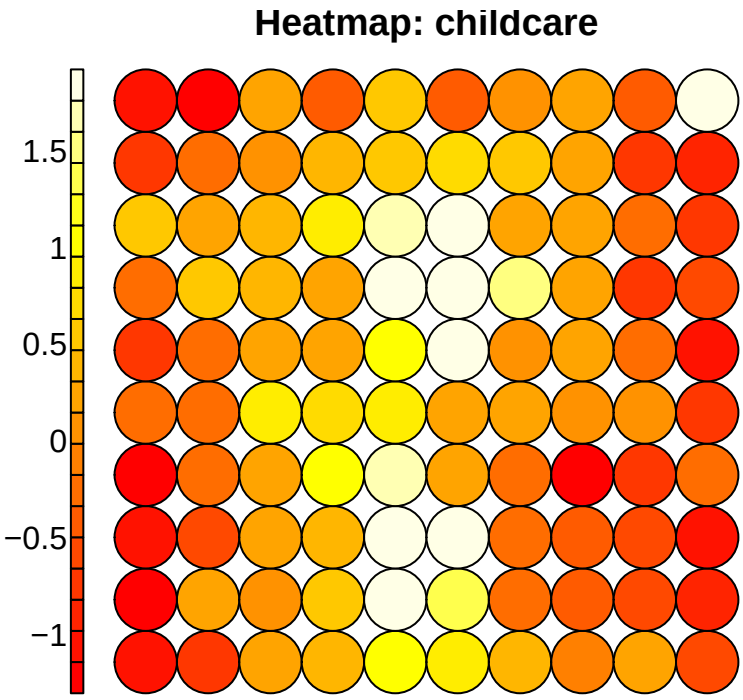

Heatmap: stopwhite
2
1
0
−1

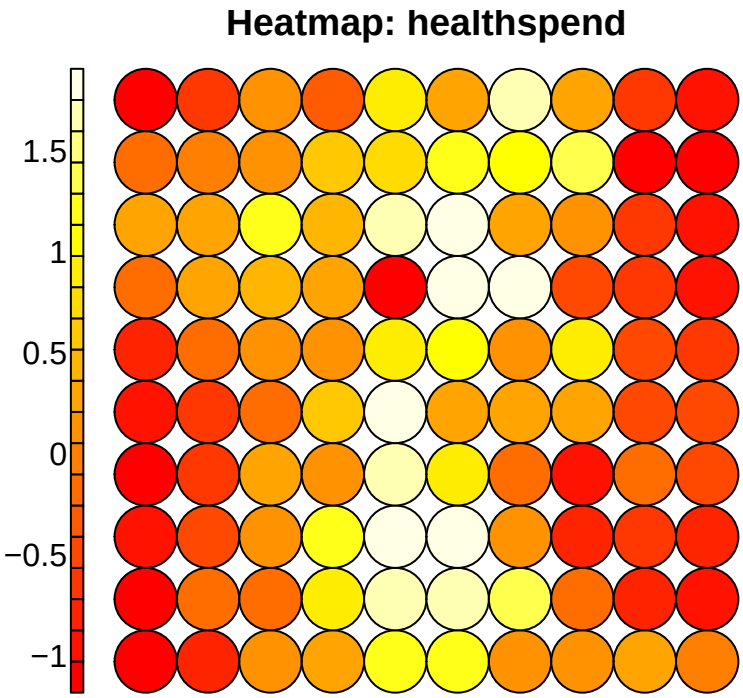

Heatmap: freetrade
1
0
−1

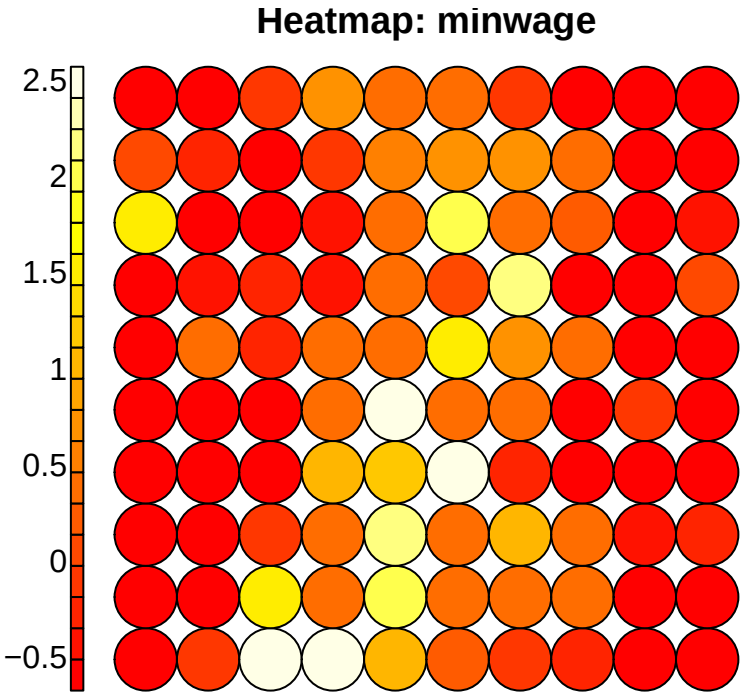

Heatmap: aa3
1
0.5
0
−0.5
−1


Heatmap: warmdo
1.5
1
0.5
0
−0.5
−1


Heatmap: finwell
1
0.5
0
−0.5
−1
−1.5
−2


Heatmap: childcare
1.5
1
0.5
0
−0.5
−1





5. Comment on the results thus far as it relates to the main goal of the task. In other words, did you uncover evidence of partisan differences in the data? Did you not? Regardless, why do you think you got the results you did? What are some other substantive patterns you detected?

- **Answer:**

- Our main goal is to identify the partisan differences based on the answers to the selected questions. I have uncovered some evidence of partisan differences in the data. First of all, the first figure reveals that 350 iterations are appopriate because there is no further reduction in perplexity as the iteration increases. The count plot shows that the samples are not equally distributed to each node. For instance, the nodes on the bottom left show unusually high number of samples compared to other nodes. Since

the number of democrats is larger in our sample, it might be the case that the one on the bottom left are democrats and the nodes on the right are republicans.

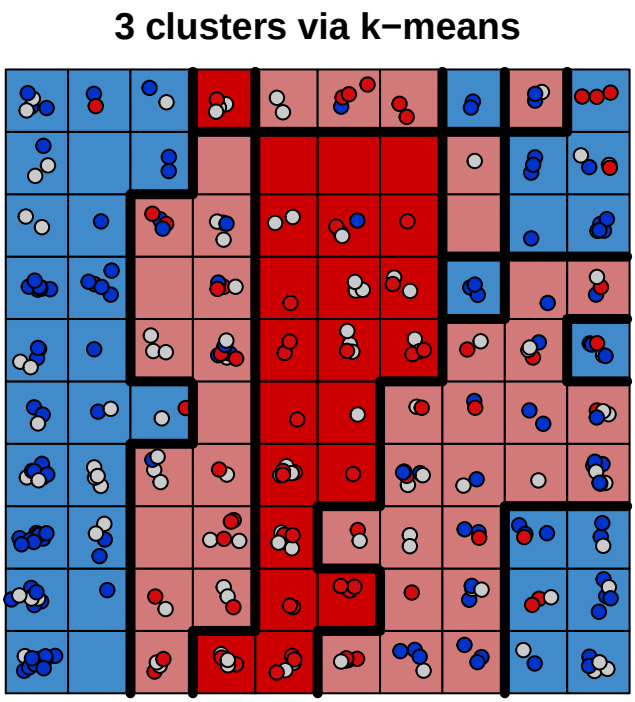
- The neighbour distance plot further shows how close or distant these observations are to each other. That is, less number of distance indicate that the nodes are close to each other. It shows that the nodes of the bottom left are more adjacent to each other than the nodes in the center and the right. This does seem to show that the bottom left nodes have similiar political views. This also reaffirms our previous finding that Democrats have more concentrated reponses to certain questions (less variance) than Republicans who had less concentrated respones to certain questions.
- The codes plot show the representations of the representative vectors where the radius of a wedge shows the magnitude in a particular direction. It is posslbe to determine that the substantive pattern that the bottom left nodes are laregely determined by finwell, stopblack and forceblack without other vectors. Since these questions ask about racial inequality, it is possible to infer that these groups might be represented by Democrats who have strong views on equity in general. As we move to the center, we can see that the number of representative vectors increase as well as the magnitude. For the bottom center nodes, it is possible to see that aa3 and birhtright_b have huge influence. Going back to our previous findings from EDA, it can be said that these groups might represent Republicans. The top center of the nodes have a large number of representative vectors. For those nodes where aa3 have less influence, they are likely to be Others because they showed more or less equal distribution of responses to each question. As we move to the far right, we see similar patterns we noticed in the left. It is likely that they are also democrats for the simliar reasons. I think I got these results because while some people who have strong party affiliations are largely determined by certain questions such as affirmative action, many people have comprehensive view of politics. The SOM algorithm self-organizes smiliar people in nearby nodes and thus see thses patterns. We can further corroborate our finidings with the heatmap plot. Since we have access to the distribution of the answers to each question by party affiliation, it is possible to deduce some information from the heatmap plot for each question. For instance, we have previously discoverd that there were partisan differneces in aa3 and we can see that the heatmap shows distinct clusters of nodes. We can see the same pattern for minwage, another variable where we identified potential partisan differences.

6. To validate the SOM results, fit a k-means algorithm to the data and plot respondents' political party affiliations as well as their cluster assignments from the k-means fit. Discuss the results. E.g., Do you see evidence of partisan differences across the groups? Do you not? How do you know?

- Yes, I see evidence of partisan differences across the groups because those who have simliar political veiws have "self-organized" themselves into clusters. That is, those who have similar political views are grouped together and are close to each other. For instance, those who identifiy themselves as democrats are surrounded by democrats and those who identify themselves as Republicans are surround by Republicans for the most part. With the k-means algorithm, it is also intersting to note that the Republicans are located in the center of the map, buffered by the Others, and finally edged by Democrats. These are in consistent with the observations we found in the SOM results in (5). The SOM, however, does not seem to have done a perfect job in determining partisan differences for the Others group. Nevertheless, this lack of performance still can be reasonable because there was no clear definition of Others in the first space and becuase we have witnessed that the reponsese by the Other groups are more or less equally distirbuted for every question with slightly more visibility in the Others category. Overall, the SOM has done a great job in organizing the partisan differences. I would have witnessed a high degree of mismatch between the results from the K-means algorithm and the SOM results had SOM did a poor job in organizing the partisan differences.

```
# clustering from SOM via k-means
point_colors <- c(amerika_palettes$Democrat[2],
                  amerika_palettes$Republican[2],
                  amerika_palettes$Dem_Ind_Rep3[2]
                  )
```

```r
neuron_colors <- c(amerika_palettes$Democrat[3],
                   amerika_palettes$Republican[3],
                   amerika_palettes$Dem_Ind_Rep3[3]
                   )

## k-means
kmeans_clusters <- som_fit$codes[[1]] %>%
  kmeans(., centers = 3)

# Plot
plot(som_fit,
     type = "mapping",
     pch = 21,
     bg = point_colors[anes_cleaned2$party],
     shape = "straight",
     bgcol = neuron_colors[as.integer(kmeans_clusters$cluster)],
     main = "3 clusters via k-means");
add.cluster.boundaries(x = som_fit, clustering = kmeans_clusters$cluster, lwd = 5, lty = 5)
```



7. Taken with the SOM results, do the k-means results show similar or different patterns? Or is it unclear? Discuss both models in relation to each other for a full, well-rounded validation. *Note*: you are encouraged to think of and implement other ways to validate your results. You are welcome to go back to class notes, Google, etc.

- Answer: Taken with the SOM results, the k-means result show smiliar patterns. Although the results were more clear for Republicans and Democrats and less clear for Others gruops, it was understndable due to the reason explained in 6. We can see that many political affiliations are indeed similar across the SOM reuslts and the K-means results. The cluster assignments from both of the methods told us how Republicans and Democrats had opposing political views as noted by cluster boundaries and the location of nodes. To make well-rounded validation of these results, however, it would be ideal to use other algorithms such as FCM and HAC to see if they produce similar results. By comparing the outputs across different techniques, we can be more or less confident about our results. Additionally,

we can use our domian knowledge to make sense of the results. For instance, we know that the United States experienced quite a substantial politcal divide in the year of 2016 with many ad-hoc campaigns proliferating in the country. With these domain knoweldge in mind, we can confirm that the results do indeed reflect the reality to some extent. We can further validate our results by looking at the potential weaknesses of our models. For instance, our result would have been much more accurate had we access to more sample size and the definition of Others were more clearly defined. Considering the presence of minor american parties and independents, it seems appropriate to give more clear definition of Others to make the analysis more coherenet. In general, however, the results seem to provide simliar and coherent patterns in relation to each other.