Challenge 3

```
data <- read_csv("Z:\\Classes\\Unsupervised Machine Learning\\Challenge 3\\data\\anes_2016.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   version = col_character(),
##   pid2d = col_character(),
##   pid2r = col_character(),
##   other10_open = col_character(),
##   race_other = col_character(),
##   employ_t = col_character(),
##   religpew_t = col_character(),
##   disc_fed_disc_police_rnd = col_character(),
##   white_sections_rnd = col_character(),
##   lazy_violent_rnd = col_character(),
##   FEELING_THERMOMETER_rnd = col_character(),
##   meet_rnd = col_character(),
##   givefut_rnd = col_character(),
##   info_rnd = col_character(),
##   ISSUES_OC14_rnd = col_character(),
##   disc_selfsex_rnd = col_character(),
##   lazy_col_rnd = col_character(),
##   lazy_row_rnd = col_character(),
##   violent_col_rnd = col_character(),
##   violent_row_rnd = col_character()
##   # ... with 9 more columns
## )
## i Use `spec()` for the full column specifications.
```

1. (10 points) Create a dichotomous feature for each identity question: amer_ident and race_ident (consider ifelse()). The question wording is, How important to your identity is [being American/your race]? For each of these, the response categories are: 1 Extremely important, 2 Very important, 3 Moderately important, 4 A little important, 5 Not important at all. So your first task is to create a new feature (e.g., strong_amer_ident), where 1 = strong identity (amer_ident = 1 or 2) and 0 = weak identity (amer_ident = 3, 4, or 5). Of note, this is an imperfect measure to be sure, but it gets to the substantive bottom line of this simple case.

I scaled all the columns, then rescaled them to be in the range [0,1]. I did this because even after normalizing the variables using the scale function, some variables had larger ranges than others. But each of these variables (and their scales) are based on different likert scales. I didn't want one opinion to be weighted more than another just because it used a different likert scale. So I rescaled the variables so they would all fit in the range [0,1].

```
issues <- data %>%
    dplyr::select(vaccine, autism, birthright_b, forceblack, forcewhite, stopblack, stopwhite, freetrad
  scale(center = TRUE, scale = TRUE) %>%
  as_tibble()%>%
  mutate_all(rescale, to = c(0,1))%>%
  as.matrix()


anes <- data %>%
  as_tibble()%>%
  mutate(strong_amer_ident = factor(if_else(amer_ident %in% c(1,2), 1, 0)),
         strong_race_ident = factor(if_else(race_ident  %in% c(1,2), 1, 0)))
```

2. (20 points) Build a self-organizing map based on only the above-listed social questions. Note: think carefully about the scale of the responses. The grid can be specified however you'd like, and hyperparameters can be tuned however you'd like. Just make decisions and justify them as you go. Hint: You might consider the kohonen package in R (though there are many others), or the minisom package in Python.

I chose a 13x13 grid, applying the formula

$$M \approx 5\sqrt{N}$$

which I found recommended in this thread: https://www.researchgate.net/post/How-many-nodes-for-self-organizing-maps, due to a paper by Tian et al. (2014).

```
set.seed(11235)

search_grid <- somgrid(xdim = 13,
                       ydim = 13,
                       topo = "rectangular",
                       neighbourhood.fct = "gaussian")

som_fit <- som(issues,
               grid = search_grid,
               radius = 1,
               rlen = 300,
               dist.fcts = "euclidean",
               mode = "batch")
```

3. (20 points) Diagnose the output of your self-organizing map visually. Hint: consider looking at the help documentation for plotting options (e.g., counts, learning rate, etc.). Discuss the model in a few sentences.

Based on the learning rate plot, The model takes about 300 iterations to get to its minimal mean distance to a node, which is about 0.00375.

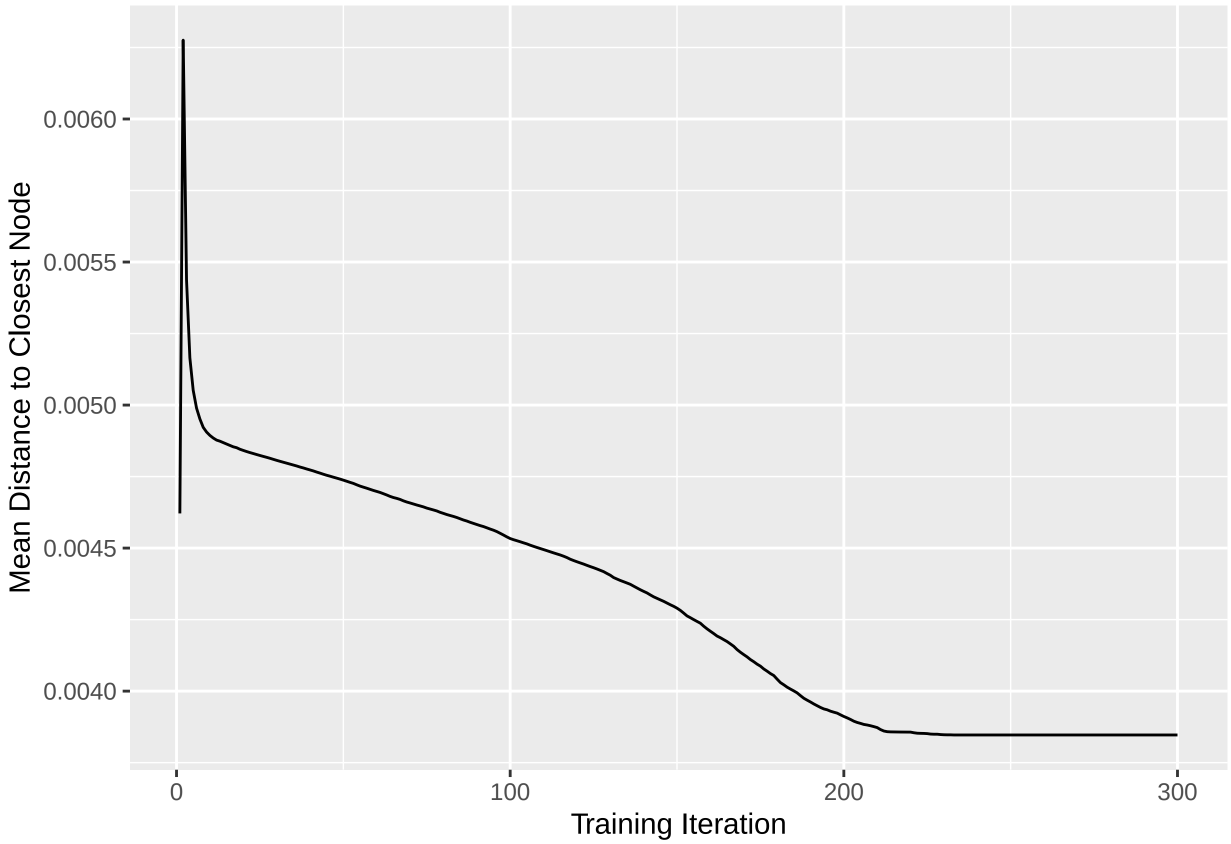
The counts plot tells us that there is a high concentration of observations in the corners, with the highest concentration being in the bottom left.

The Neighbor Distances graph backs this up, showing a low neighbor distances in the far bottom left corner. It also reveals some low niehbor distances in the center of the map.

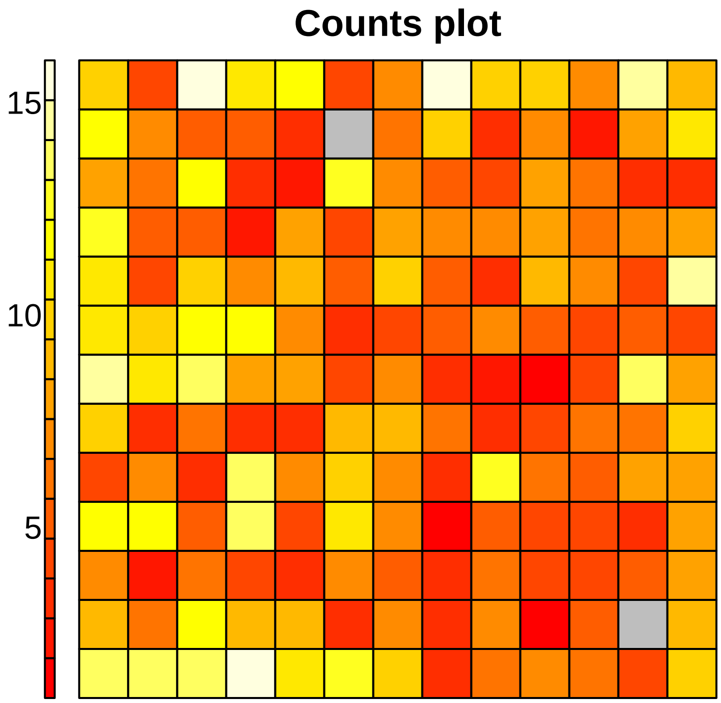
The heat maps of individual inputs are also interesting. For instance, the autism and vaccine graphs appear to have a strong negative relationship. Childcare, healthspend, minwage, and warmdo all have visually similar graphs, suggesting that these social polciy issues have support from the same group and are opposed by the same group. Forceblack and stopblack are also highly related, which makes sense as both are related to police discrimination against Black people.

```
som_fit$changes %>%
  as_tibble()%>%
  dplyr::mutate(changes = V1,
                iteration = seq(1:length(changes))) %>%
  ggplot(aes(iteration, changes)) +
  geom_line() +
  labs(x = "Training Iteration",
       y = "Mean Distance to Closest Node")
```

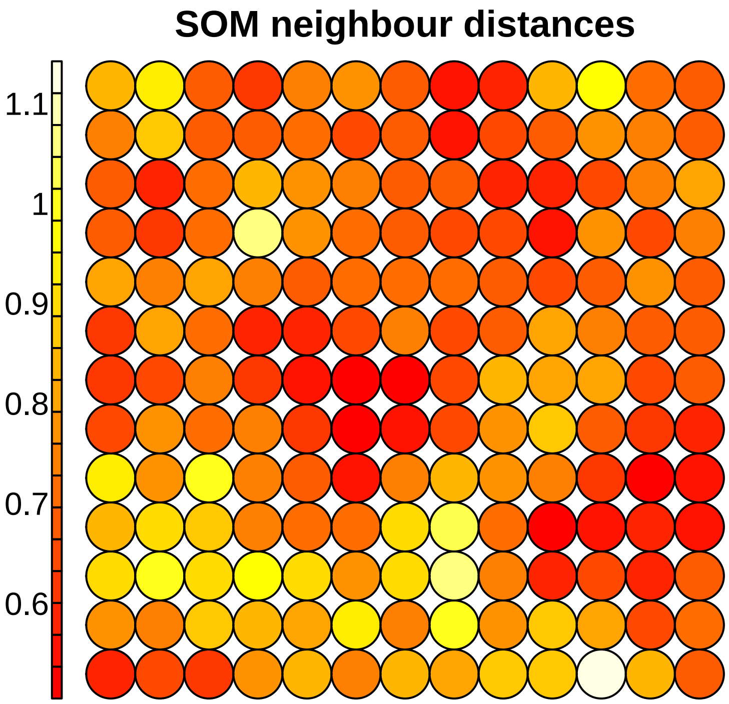
```
## Warning: The ‘x‘ argument of ‘as_tibble.matrix()‘ must have unique column names if ‘.name_repair‘ is
## Using compatibility ‘.name_repair‘.
```



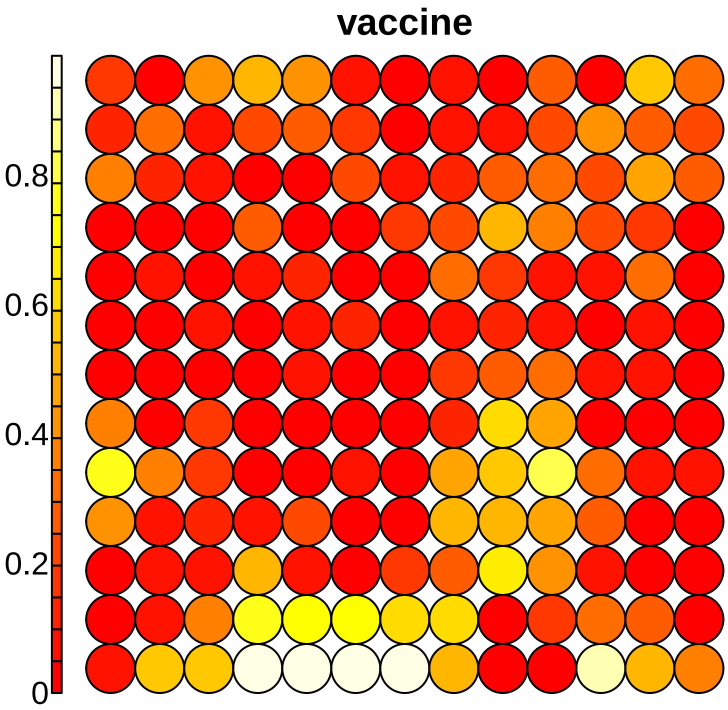
```
plot(som_fit, type="counts", shape = "straight")
```



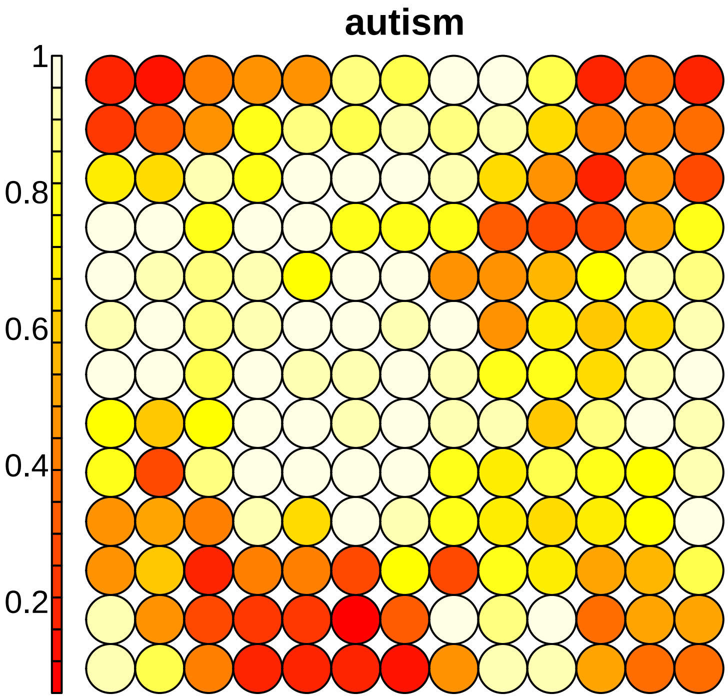
```
plot(som_fit, type="dist.neighbours", main = "SOM neighbour distances")
```



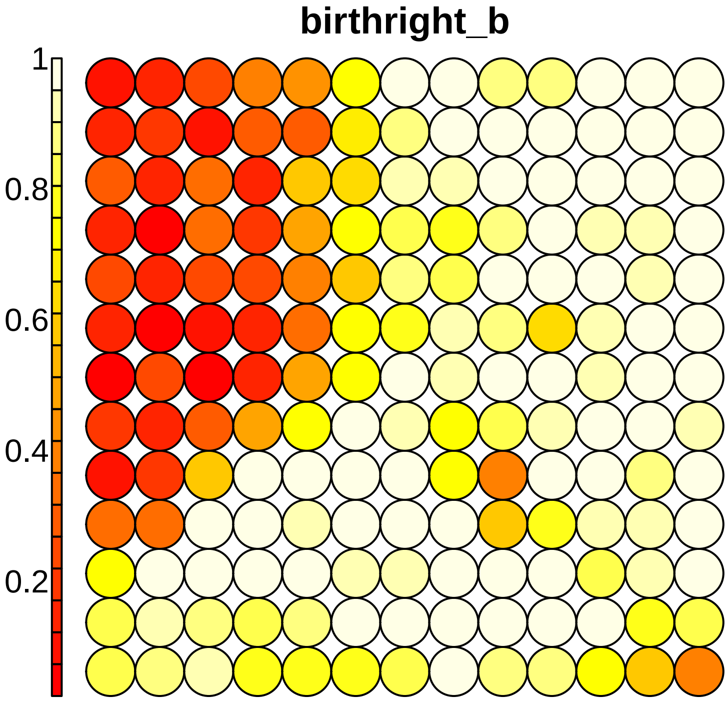
```
#https://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/
for(i in seq(ncol(getCodes(som_fit)))){
  print(
    plot(som_fit, type = "property",
       property = getCodes(som_fit)[,i],
       main = colnames(getCodes(som_fit))[i])
  )
}
```



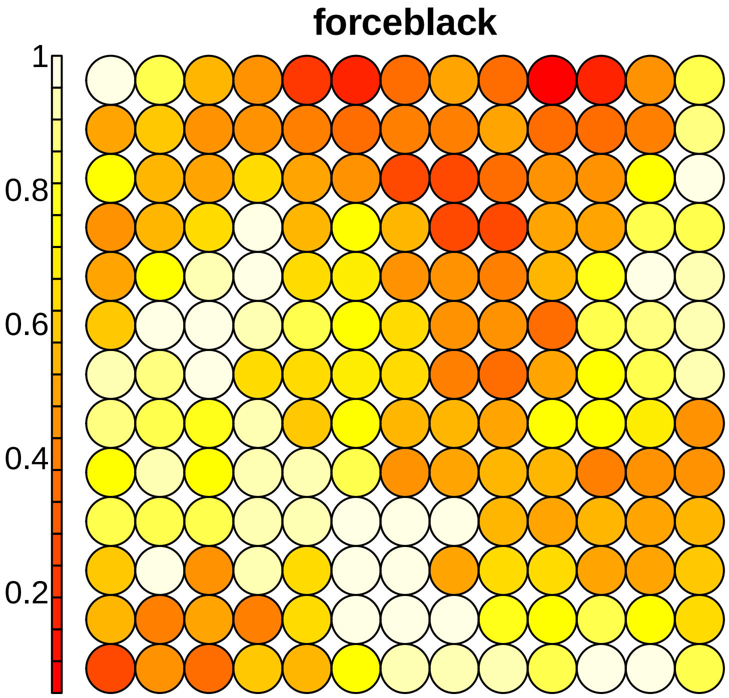
```
## NULL
```



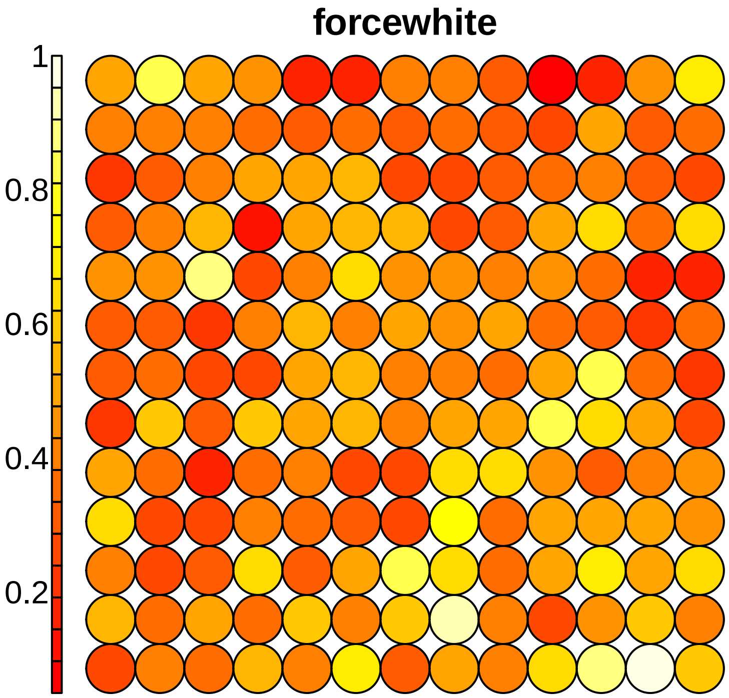
```
## NULL
```



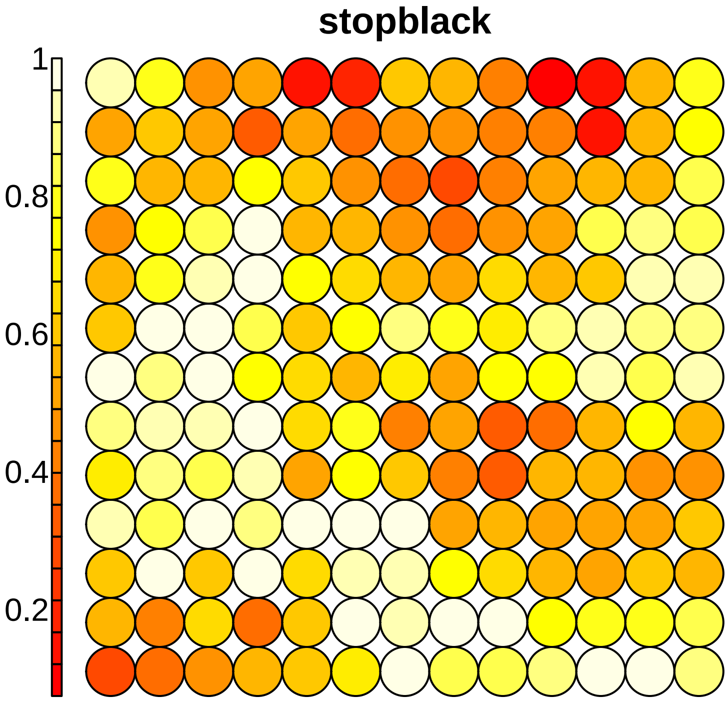
```
## NULL
```



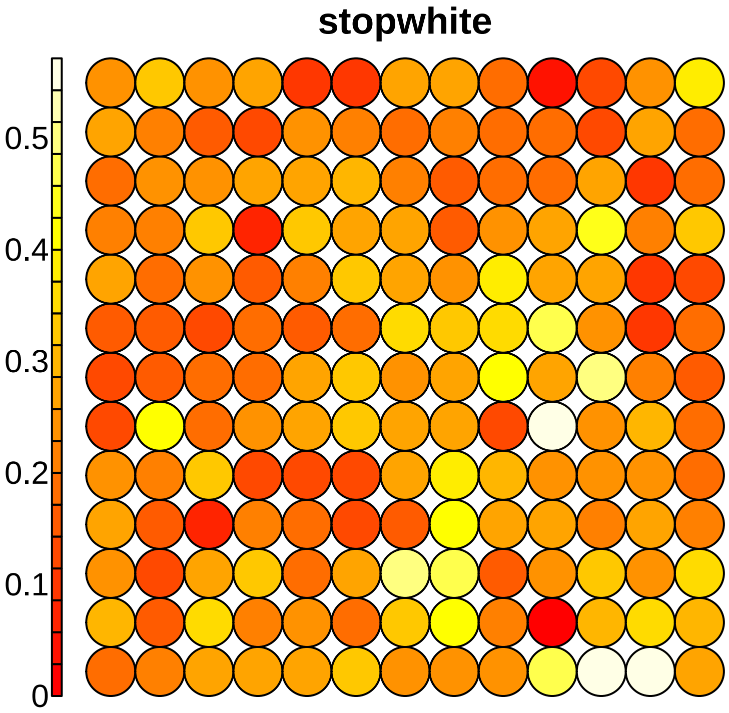
```
## NULL
```



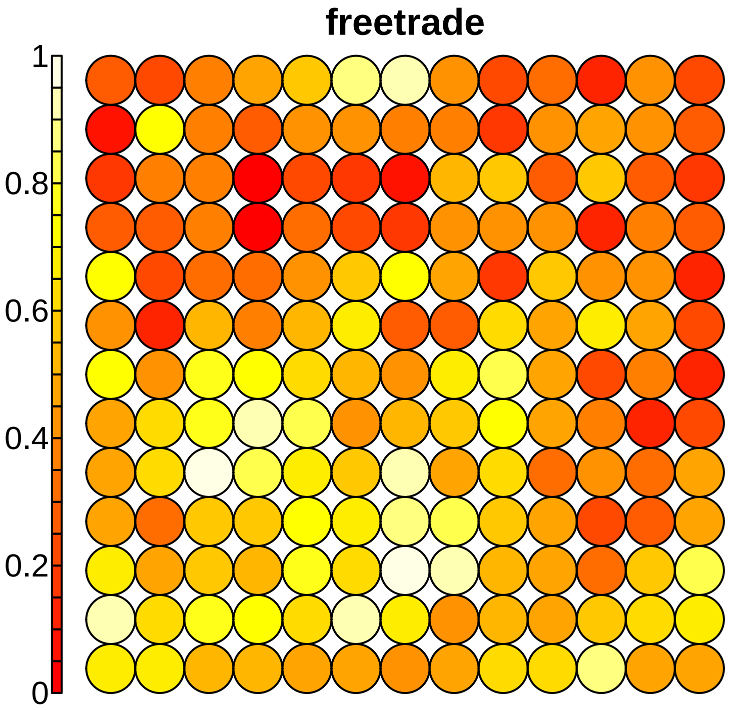
```
## NULL
```



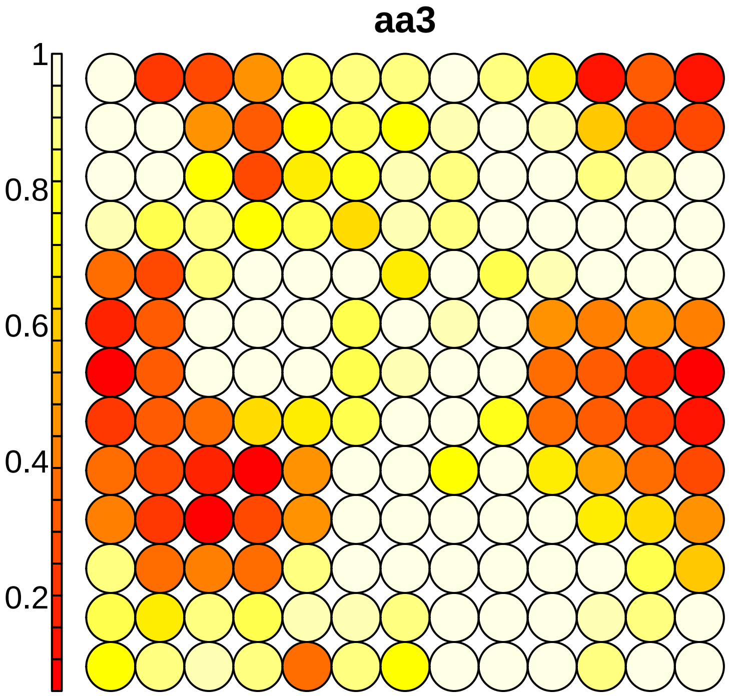
```
## NULL
```



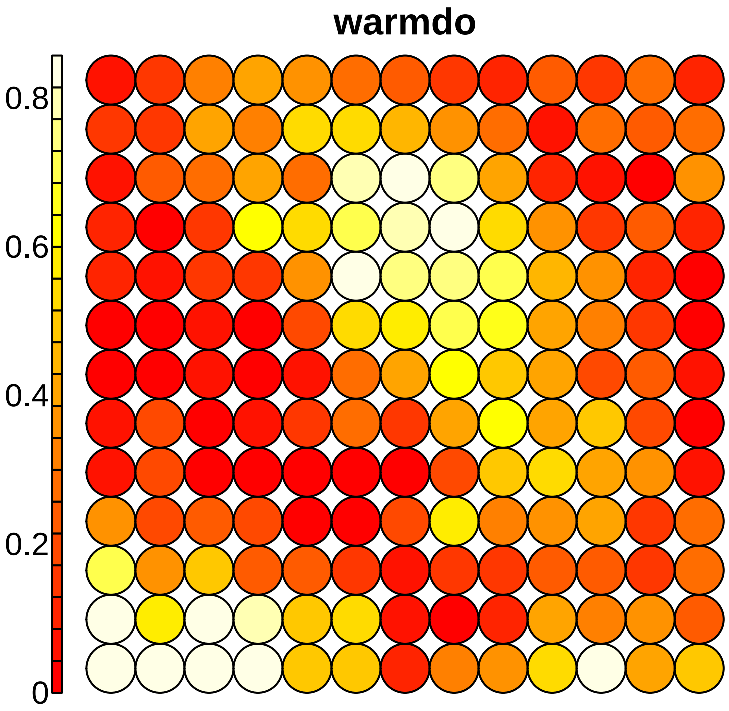
```
## NULL
```



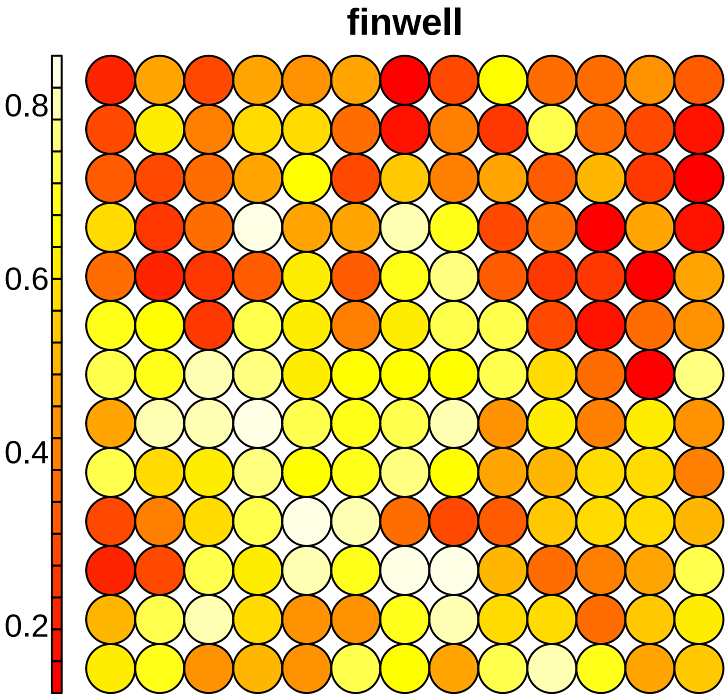
```
## NULL
```



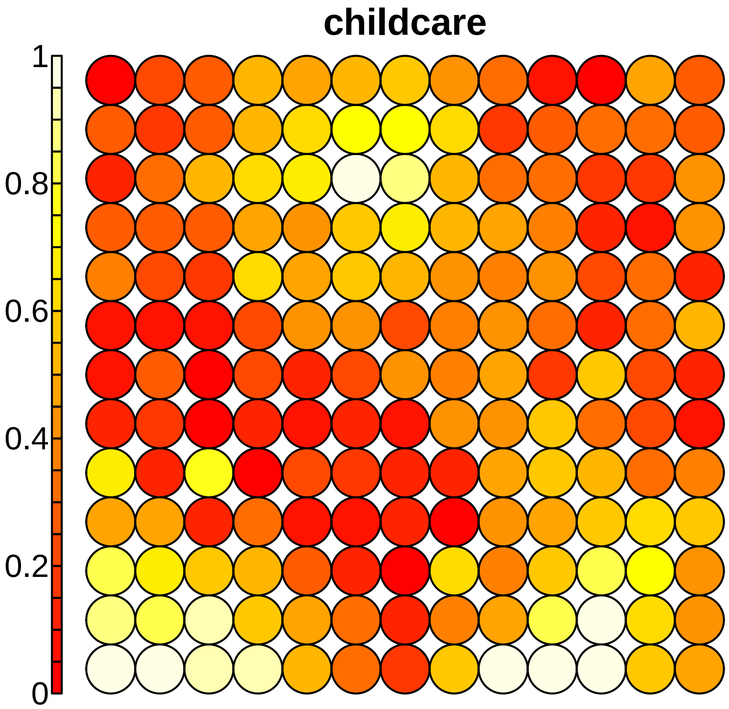
```
## NULL
```



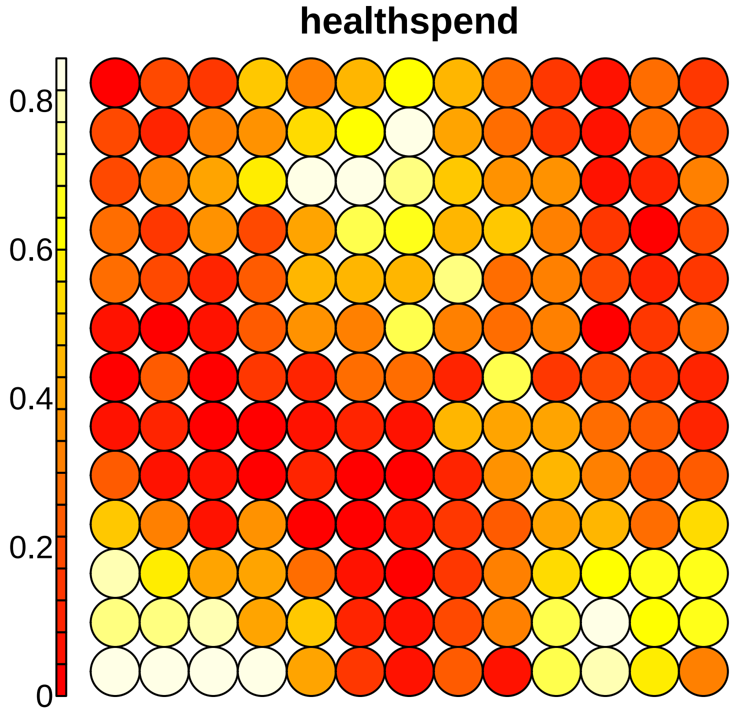
```
## NULL
```



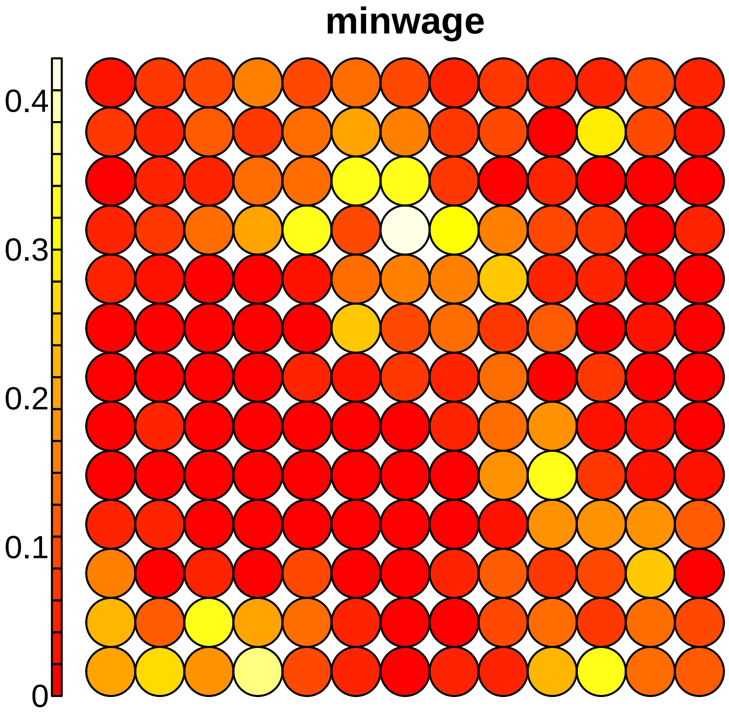
```
## NULL
```



```
## NULL
```



```
## NULL
```



```
## NULL
```

4. (20 points) Color the output layer ("mapping") by your dichotomous feature for weak/strong American identity. Discuss the output in a few sentences. For example, do you see grouping along respondents with similar senses of American identity? Why or why not do you think? Any surprising patterns? And so on.

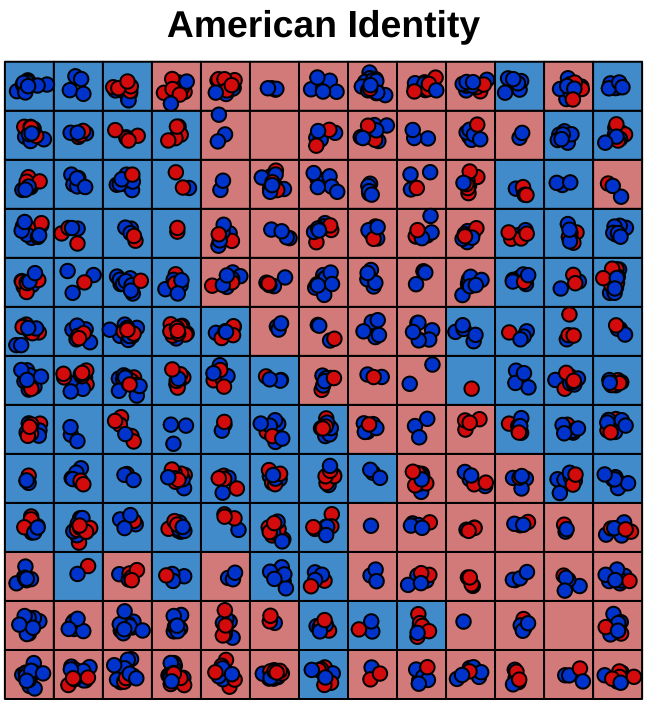
Prior to building the map, I expected to see clear clustering, based on people with high levels of American identity favoring conservative political issues.

Below, blue dots represent strong feelings of American identity being, and red dots represent weak feelings of American identity. I have plotted the SOM two ways: with neurons colored by a kmeans cluster and with no coloring. The k-means did identify clustering, but it wasn't as obvious to my eyes when I removed the kmeans colors.

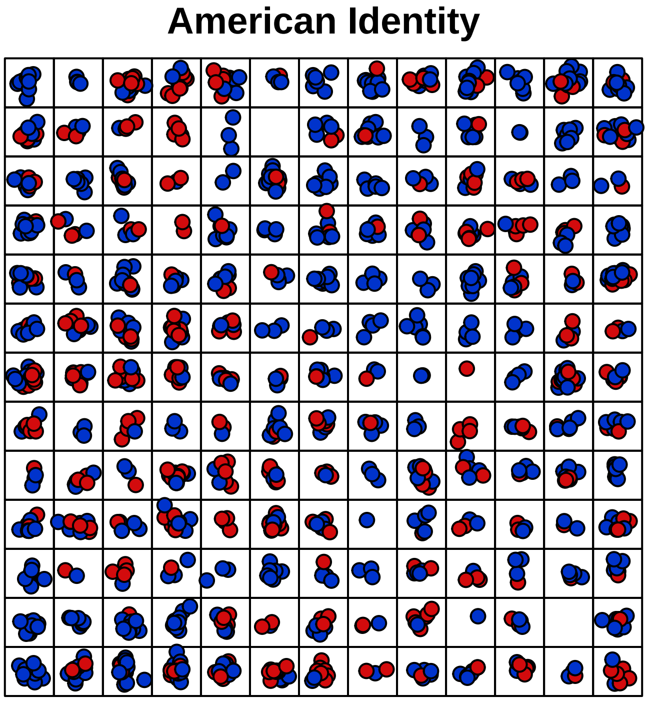
I can pick out some clusters: a blue cluster in the bottom left and also in the center, roughly corresponding to the kmeans. But it's a much more mixed picture than I was expecting. This suggests that strong feelings of american identity are not a great predictor of opinion on social issues, although they do have some degree of influence.

```r
point_colors <- c(amerika_palettes$Republican[2],
                  amerika_palettes$Democrat[2])

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

```
kmeans_clusters <- som_fit$codes[[1]] %>%
  kmeans(., centers = 2)

class_assign_km <- map_dbl(kmeans_clusters$cluster, ~{
  if(. == 1) 2
  else 1
}
)

plot(som_fit,
     type = "mapping",
     pch = 21,
     bg = point_colors[anes$strong_amer_ident],
     bgcol = neuron_colors[as.integer(class_assign_km)],
     shape = "straight",
     main = "American Identity")
```



```
plot(som_fit,
     type = "mapping",
     pch = 21,
     bg = point_colors[anes$strong_amer_ident],
     bgcol = "00000",
     shape = "straight",
     main = "American Identity")
```

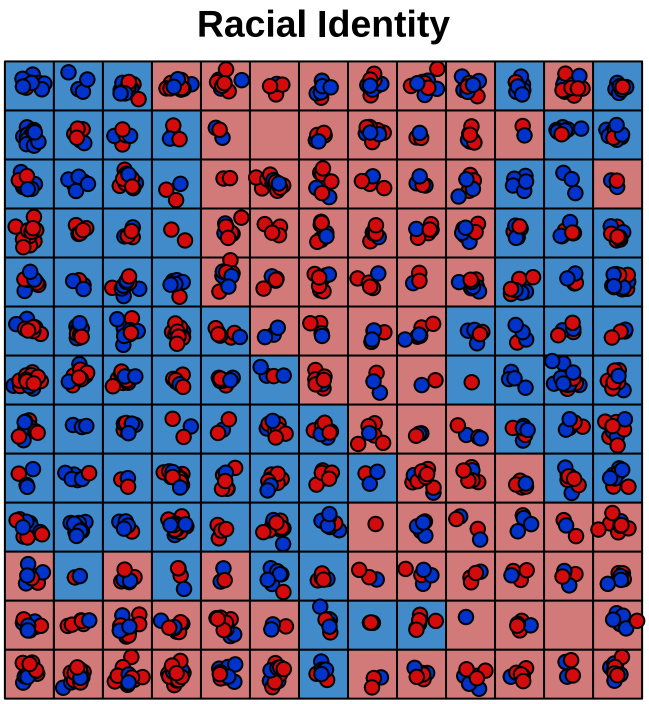


5. (20 points) Color the output layer ("mapping") by your dichotomous feature for weak/strong race identity. Discuss the output in a few sentences. For example, do you see grouping along respondents with similar senses of racial identity? Why or why not do you think? Any surprising patterns? And so on.

Below, blue dots represent strong feelings of racial identity being important, and red dots represent weak feelings of racial identity being important.

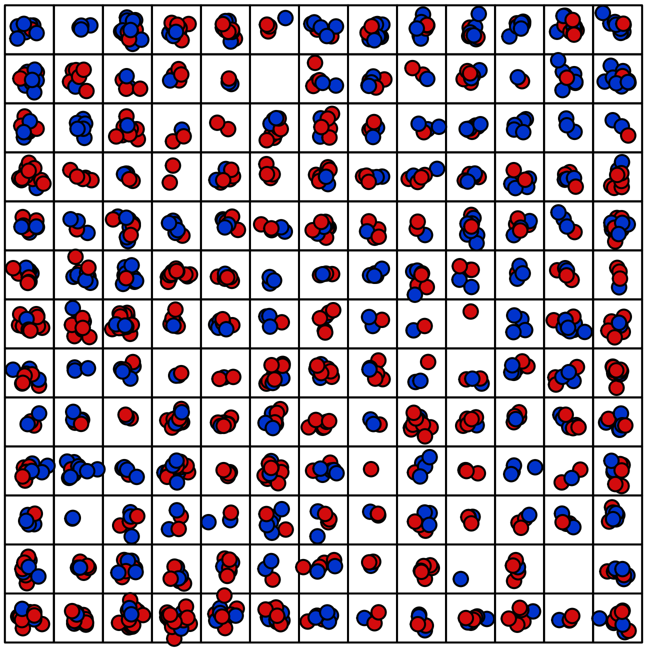
I expected to see stronger clustering with the racial feature, but instead it actually appears to be even weaker than for American identity. Without viewing the k-means clustering, it's pretty difficult to pick out any clear patterns at all. I think this signals that most of the questions in my dataset aren't inflenced in a clear way by the degree to which a person is proud of their racial identity. Upon further reflection, maybe that makes sense: what does it mean to find one's race important? It could signal that the person holds very liberal views but thinks racial justice is important. Or it could signal that the person is a white nationalist. Both would say race is "very important" to their identities even though their politics could not be more different.

```
kmeans_clusters <- som_fit$codes[[1]] %>%
  kmeans(., centers = 2)

class_assign_km <- map_dbl(kmeans_clusters$cluster, ~{
  if(. == 1) 2
  else 1
}
)

plot(som_fit,
     type = "mapping",
     pch = 21,
     bg = point_colors[anes$strong_race_ident],
     bgcol = neuron_colors[as.integer(class_assign_km)],
     shape = "straight",
     main = "Racial Identity")
```



```
plot(som_fit,
     type = "mapping",
     pch = 21,
     bg = point_colors[anes$strong_race_ident],
     bgcol = "00000",
     shape = "straight",
     main = "Racial Identity")
```

Racial Identity



6. (10 points) Offer a few concluding thoughts comparing the patterns in responses to social questions across these two conceptualizations of identity. Talk about similarities and/or differences, what we can learn, how these patterns comport with your loose expectations, and so on. Just a few sentences will suffice.

This analysis was surprising because it shows that the degree to which one feels their identity as an American is important to them and the degree to which one feels their racial identity is important to them really don't have clear-cut impacts on their opinion on a whole range of social issues. It's not that there isn't any clustering by social issues; comparing the issue-by-issue heatmaps illustrates that there are certain issues that seem to "go together." Rather, the issue is that the degree to which people think race and nationality are important to their identities doesn't decisively influence adoption of these opinions. That casts doubt on explanations of American politics that tend to reduce opinion down to these types of identitarian classifications. For instance, it's common to hear strong nationalism associated with conservative political beliefs. I did not find clear evidence of that.