Homework 7

Applied Mutlivariate Analysis

Emorie Beck

October 22, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(kableExtra)
library(MASS)
library(vegan)
library(smacof)
library(scatterplot3d)
library(ape)
library(ade4)
library(ecodist)
library(lme4)
library(plyr)
library(tidyverse)
```

1.2 data

We know that President Trump is a polarizing figure within the United States, but how is he viewed around the world? In 2017, the Pew Research Center conducted surveys in 38 countries. Each sample was representative of that countrys adult population and sample sizes were typically around 1000 in each country (the details can be found at http://www.pewresearch.org/methodology/international-survey-research/international- methodology/global-attitudes-survey/all-country/2017). Included among the survey items were the following that asked respondents to answer yes or no:

Please tell me whether you think the following describes U.S. President Donald Trump. Do you think of Donald Trump as _____?

1. well-qualified to be president
2. a strong leader
3. dangerous
4. charismatic
5. intolerant
6. caring about ordinary people
7. arrogant

The data are contained in the file, trump.csv. For each country, the percentage of the sample that responded yes is provided.

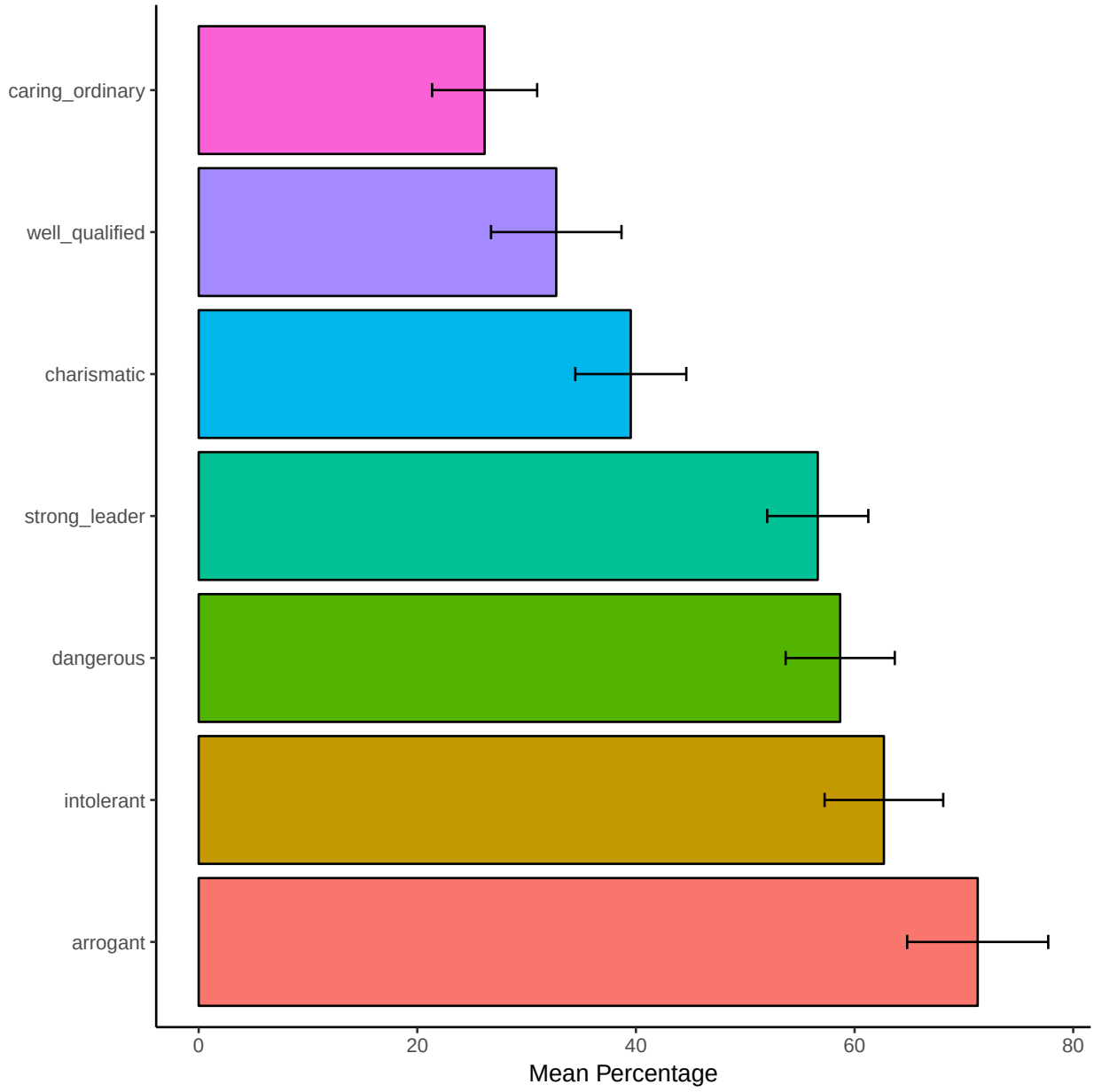
```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework7"

dat <- sprintf("%s/trump.csv", wd) %>%
  read.csv(., stringsAsFactors = F)

head(dat)
```

```
##          Country well_qualified strong_leader intolerant dangerous
## 1 United States              38            52         60        53
## 2        Canada              16            38         78        72
## 3        France              21            54         83        78
## 4       Germany               6            44         81        76
## 5        Greece              33            55         70        55
## 6       Hungary              39            60         50        42
##   charismatic caring_ordinary arrogant
## 1          49              42       81
## 2          37              23       93
## 3          52              18       93
## 4          37              13       91
## 5          35              20       78
## 6          58              33       66
```

2 Question 1

First, provide a profile of the sample. Construct a bar graph that contains the mean percentage for each question along with the 95% confidence interval.

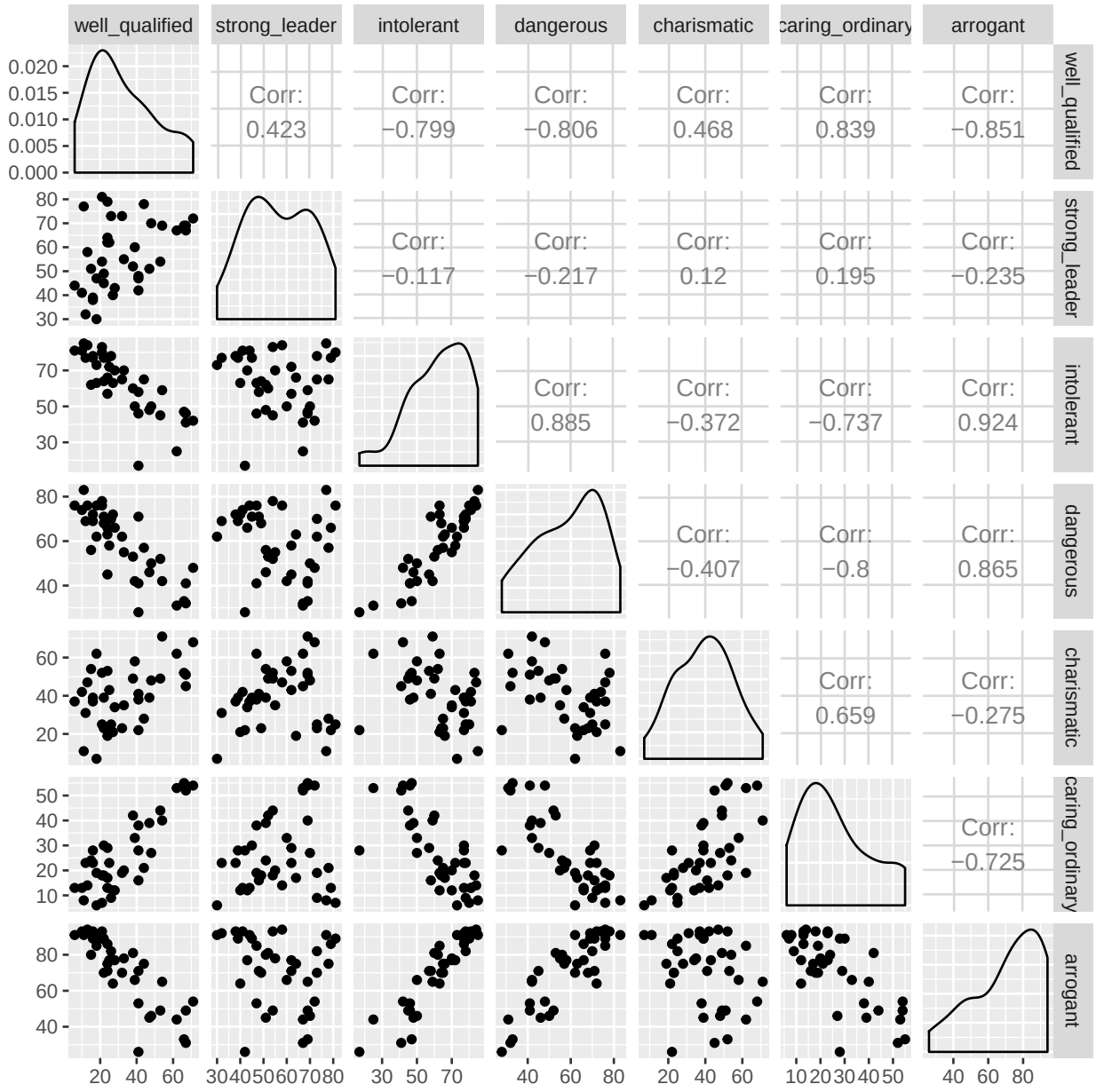
```
dat %>%
  gather(key = item, value = value, -Country) %>%
  Rmisc::summarySE(., measurevar = "value", groupvars = "item") %>%
  mutate(rank = rank(desc(value))) %>%
  arrange(rank) %>%
  mutate(item = factor(item, levels = unique(item))) %>%
  ggplot(aes(x = item, y = value, ymin = value - ci, ymax = value + ci)) +
    geom_bar(aes(fill = item), color = "black", position = "dodge", stat = "identity") +
    geom_errorbar(position = "dodge", width = .1) +
    labs(y = "Mean Percentage", x = NULL) +
    coord_flip() +
    theme_classic() +
    theme(legend.position = "none")
```



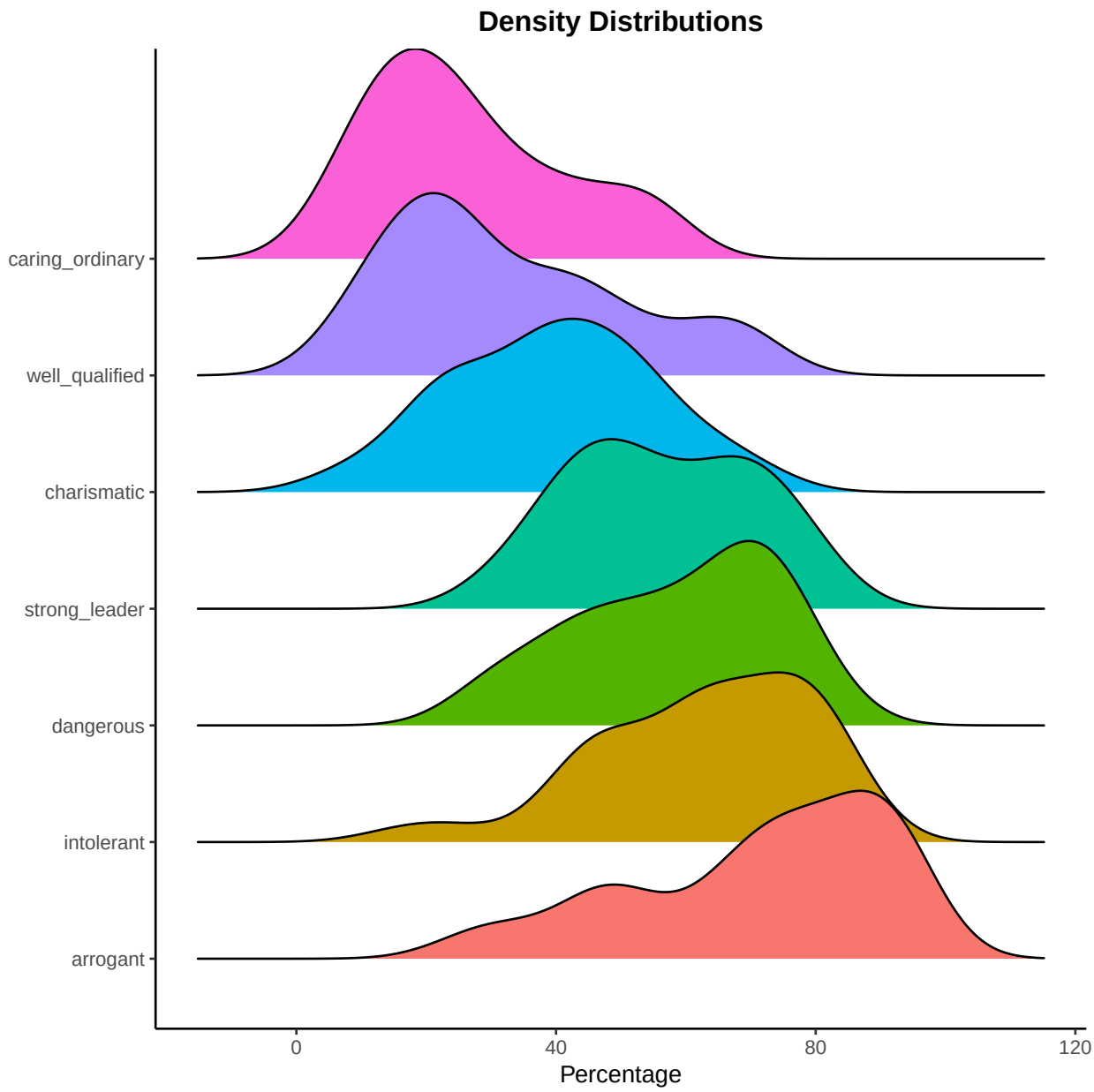
3 Question 2

Second, create two additional data frames representing alternative ways to represent the data.

```
GGally::ggpairs(dat %>% select(-Country))
```




```
dat %>% gather(key = item, value = value, -Country) %>%
  group_by(item) %>%
  mutate(mean = mean(value)) %>%
  ungroup() %>%
  mutate(rank = rank(desc(mean))) %>%
  arrange(rank) %>%
  mutate(item = factor(item, levels = unique(item)))  %>%
  ggplot(aes(x = value, y = item, fill = item)) +
    # geom_boxplot() +
    ggridges::geom_density_ridges() +
    labs(x = "Percentage", y = NULL, title = "Density Distributions") +
    theme_classic() +
    theme(legend.position = "none",
          plot.title = element_text(hjust = .5, face = "bold"))
```



3.1 Part A

The ratings are in a common metric (percentages) but vary in their standard deviations. To give each item equal weight, standardize the ratings.

```
dat <- dat %>%
  gather(key = item, value = raw, -Country) %>%
  group_by(item) %>%
  mutate(z = as.numeric(scale(raw)))
```

3.2 Part B

The percentage scale is bounded and a case might be made that differences at the extremes (e.g., the difference between 90% and 95%) are more important than equal differences in the middle (e.g., the difference between 45% and 50%). One way to provide this unequal emphasis is to convert the percentages to proportions and then to transform them to probits using the qnorm( ) function. This will stretch the scale at the extremes.

```
dat <- dat %>%
  mutate(probit = ecotoxicology::PercentageToProbit(raw/100))
```

4 Question 3

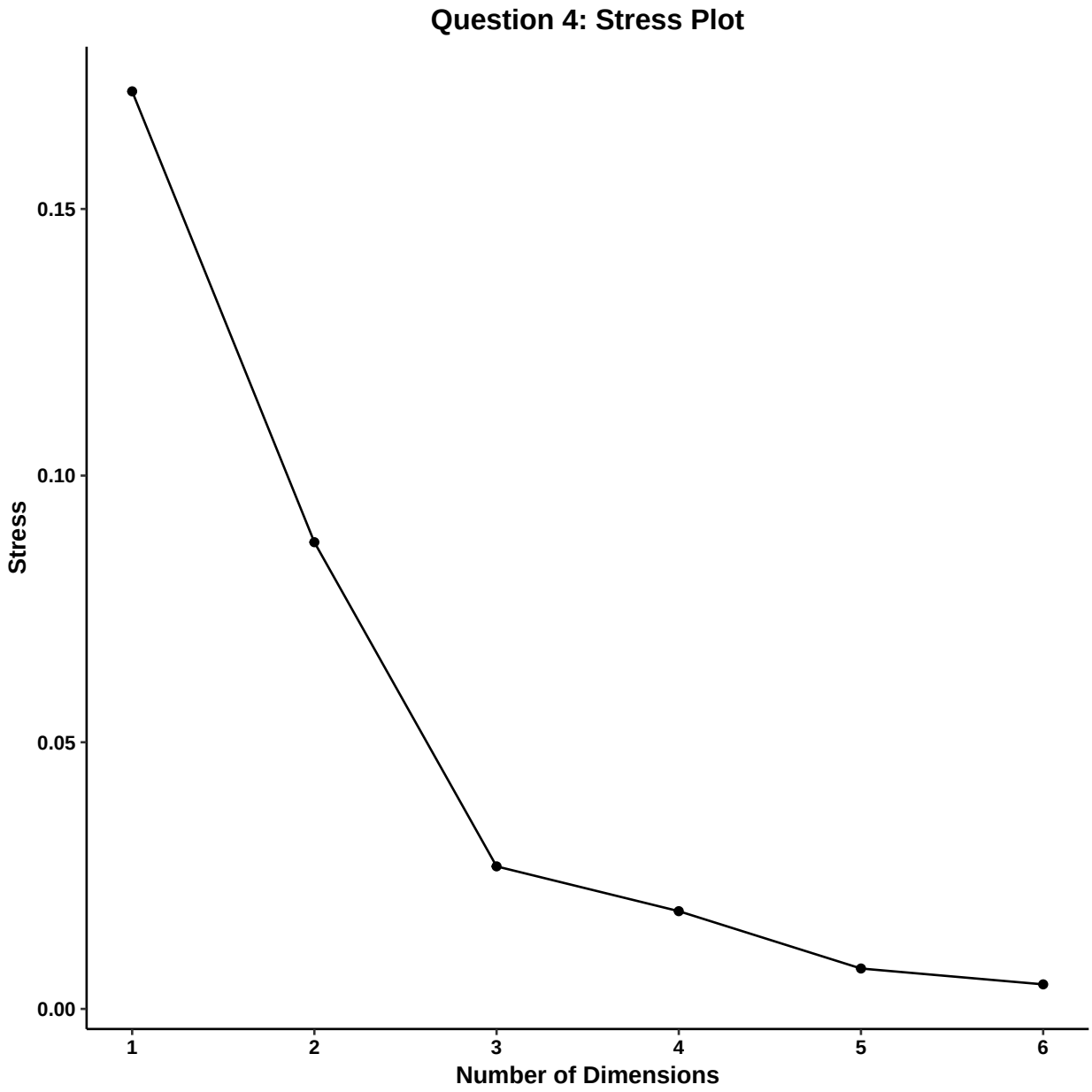
Convert each rating file to a distance matrix using Euclidean distances.

```
nested.dat <- dat %>%
  gather(key = type, value = value, -Country, -item) %>%
  spread(key = item, value = value) %>%
  group_by(type) %>%
  nest() %>%
  mutate(d_mat = map(data, ~dist(.[,-1],method="euclidean",diag=TRUE)))
```

5 Question 4

Using the original ratings data, and metric multidimensional scaling, construct a plot of the stress values (for up to 6 dimensions). How many dimensions are suggested by this plot?

```
nested.dat <- nested.dat %>%
  full_join(crossing(type = nested.dat$type, dim = 1:6)) %>%
  mutate(mds = map2(d_mat, dim, ~smacofSym(.x, ndim = .y,verbose=FALSE,type="ordinal",itmax=1000)),
         stress = map_dbl(mds, ~.$stress))

nested.dat %>%
  filter(type == "raw") %>%
  ggplot(aes(x = dim, y = stress)) +
    geom_line() +
    geom_point() +
    scale_x_continuous(limits = c(1,6), breaks = 1:6) +
    labs(x = "Number of Dimensions", y = "Stress", title = "Question 4: Stress Plot") +
    theme_classic() +
    theme(plot.title = element_text(face = "bold", hjust = .5),
          axis.text = element_text(face = "bold", color = "black"),
          axis.title = element_text(face = "bold"))
```



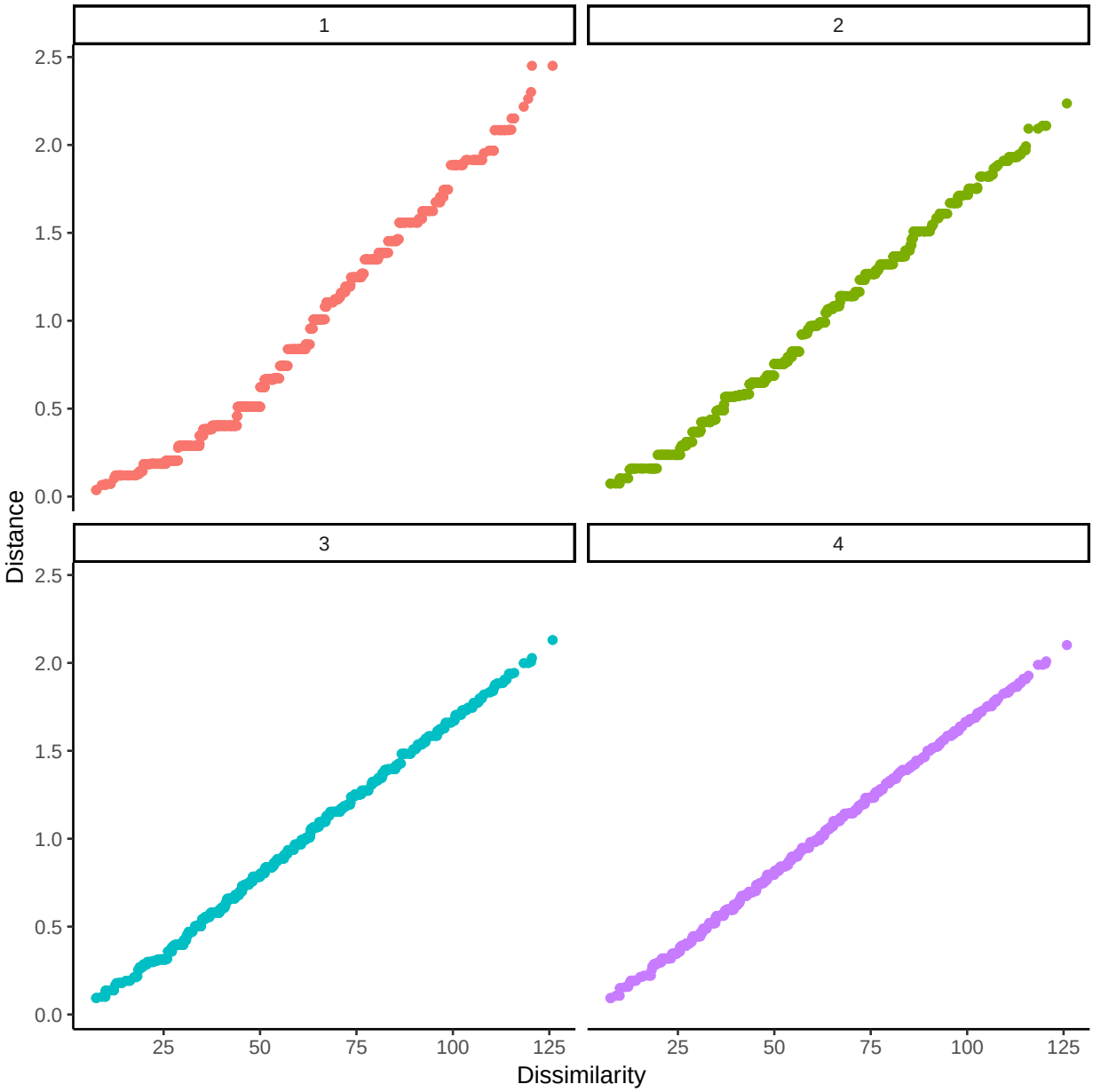
There appear to be 3 dimensions.

6 Question 5

Construct Shepard plots for up to 4 dimensions. Is your decision about the number of dimensions in Question 4 supported by the Shepard plots?

```
shep_extract_fun <- function(shep){
  tibble(Dissimilarity = shep$x, Distance = shep$yf)
}

nested.dat <- nested.dat %>%
  mutate(shep = map2(d_mat, mds, ~Shepard(.x,.y$conf)),
         shep_val = map(shep, shep_extract_fun))
```

```r
nested.dat %>%
  unnest(shep_val) %>%
  filter(dim <= 4 & type == "raw") %>%
  ggplot(aes(x = Dissimilarity, y = Distance, color = factor(dim))) +
    geom_point() +
    facet_wrap(~dim) +
    theme_classic() +
    theme(legend.position = "none")
```

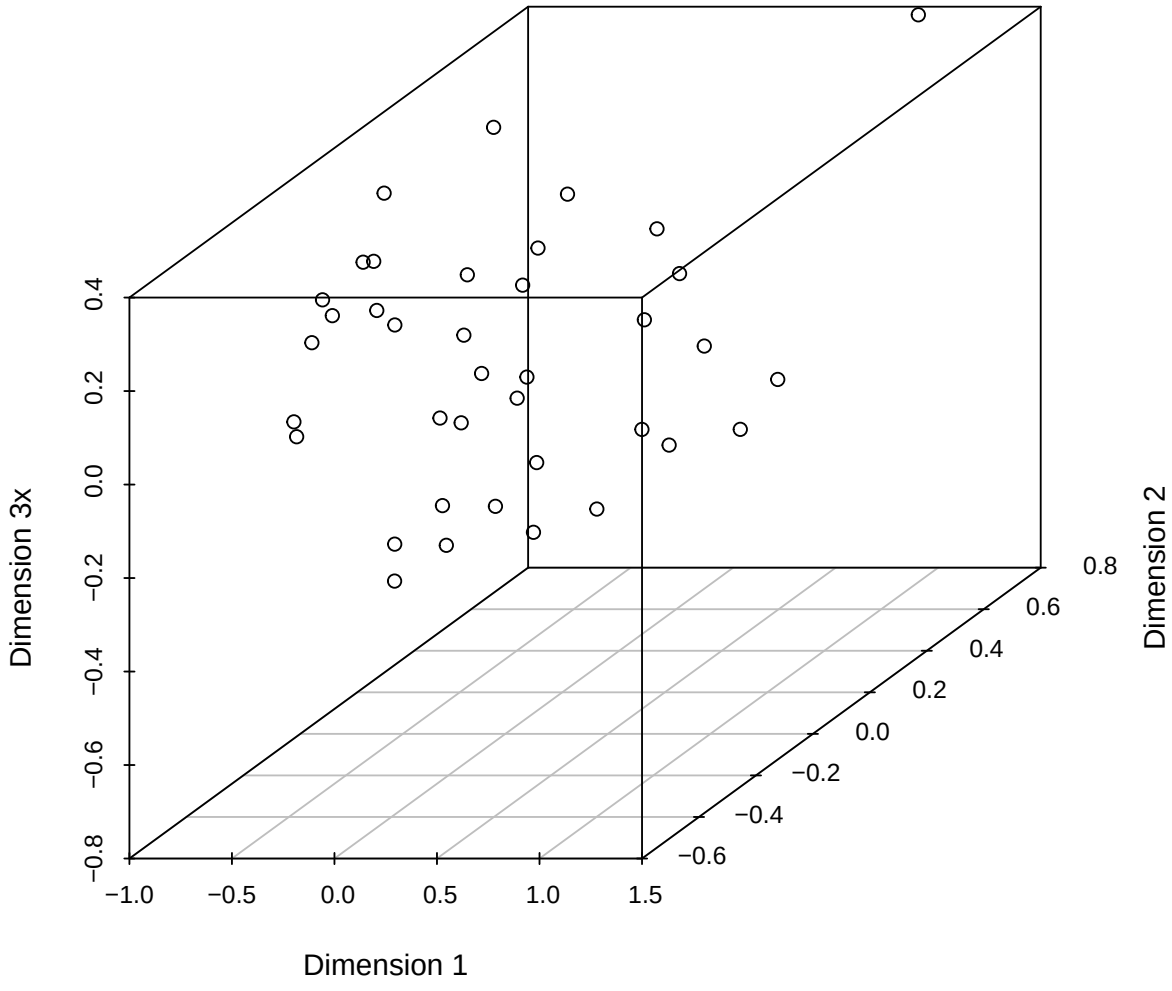


The Shepard plot also suggests 3 dimensions.

7 Question 6

Construct a graphical display of the multidimensional space using the number of dimensions identified in Question 4.

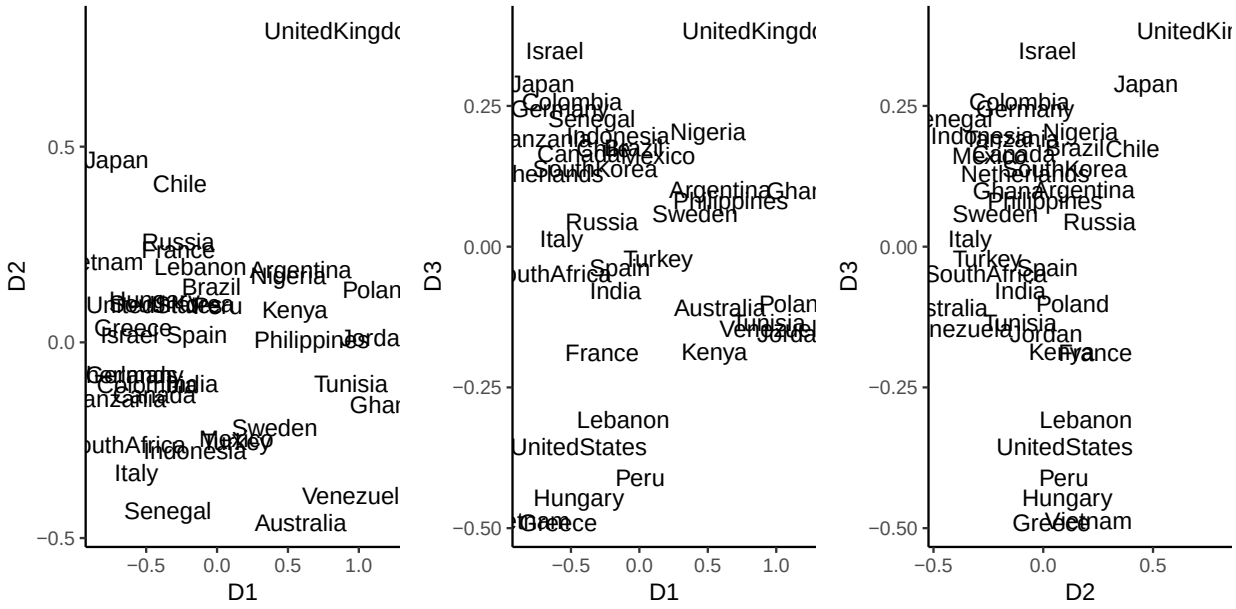
```
mds_3_dat <- (nested.dat %>%
  filter(dim == 3 & type == "raw"))$mds[[1]]$conf %>%
  data.frame

scatterplot3d(x = mds_3_dat$D1, y = mds_3_dat$D2, z = mds_3_dat$D3,
              xlab = "Dimension 1", ylab = "Dimension 2", zlab = "Dimension 3x")
```



8 Question 7

Examine the multidimensional space and provide names for the dimensions.

```
mds_3_dat <- mds_3_dat %>%
  mutate(Country = str_remove_all(unique(dat$Country), " "))

p1 <- mds_3_dat %>%
  ggplot(aes(x = D1, y = D2, label = Country)) +
    geom_text() +
    theme_classic()

p2 <- mds_3_dat %>%
  ggplot(aes(x = D1, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

p3 <- mds_3_dat %>%
  ggplot(aes(x = D2, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

gridExtra::grid.arrange(p1, p2, p3, nrow = 1)
```

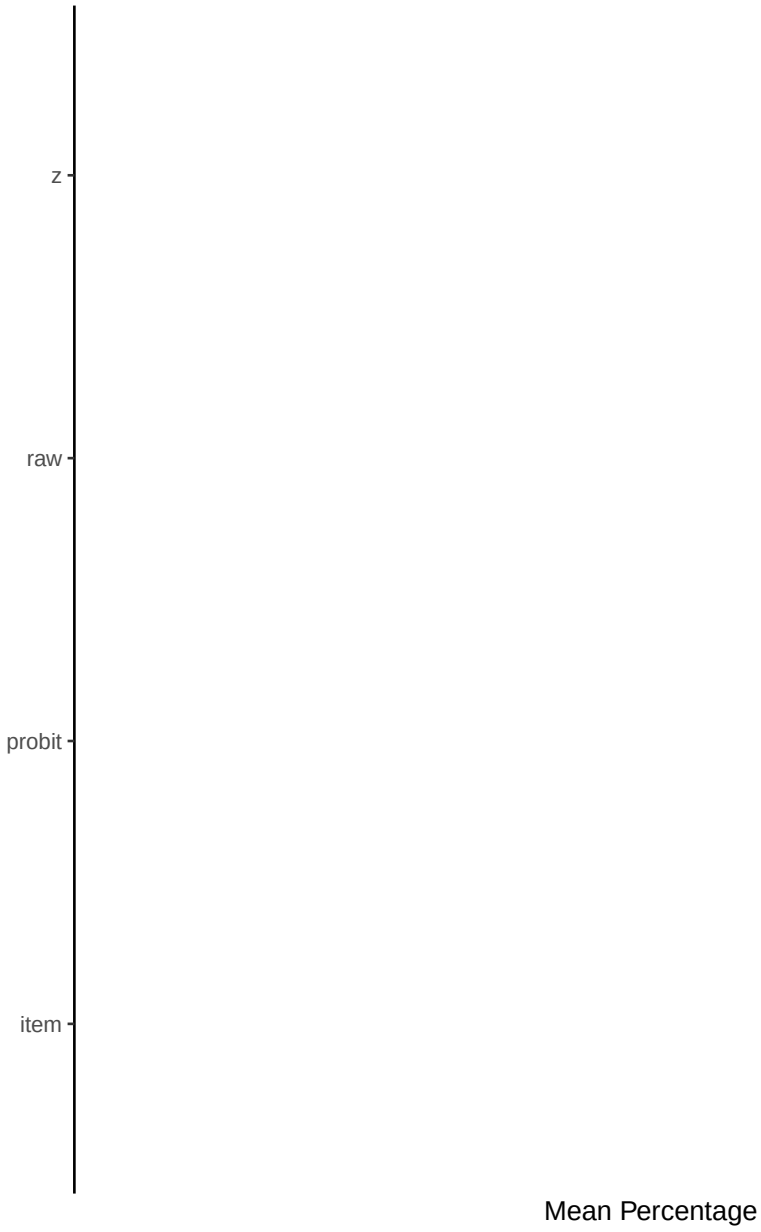


Dimension 1 might be general support for Trump. Dimension 2 might be egalitarianism. Dimension 3 could be Liberalism.

9 Question 8

Are there any clusters of countries that seem to emerge in the multidimensional space? Provide the same profile plot for each cluster as you provided for all countries in Question 1. Columbia, Germany, the Netherlands, Indonesia, Brazil, South Korea, Canada, and Tanzania seem to cluster together across all 3 dimensions.

```r
dat %>%
  mutate(Country = str_remove_all(Country, " "),
         cluster = ifelse(Country %in% c("Colombia", "Germany", "Netherlands",
         "Indonesia", "Brazil", "South Korea", "Canada", "Tanzania"), "Cluster", Country)) %>%
  gather(key = item, value = value, -Country, -cluster) %>%
  Rmisc::summarySE(., measurevar = "value", groupvars = "item") %>%
  mutate(rank = rank(desc(value))) %>%
  arrange(rank) %>%
  mutate(item = factor(item, levels = unique(item))) %>%
  ggplot(aes(x = item, y = value, ymin = value - ci, ymax = value + ci)) +
    geom_bar(aes(fill = item), color = "black", position = "dodge", stat = "identity") +
    geom_errorbar(position = "dodge", width = .1) +
    labs(y = "Mean Percentage", x = NULL) +
    coord_flip() +
    theme_classic() +
    theme(legend.position = "none")
```

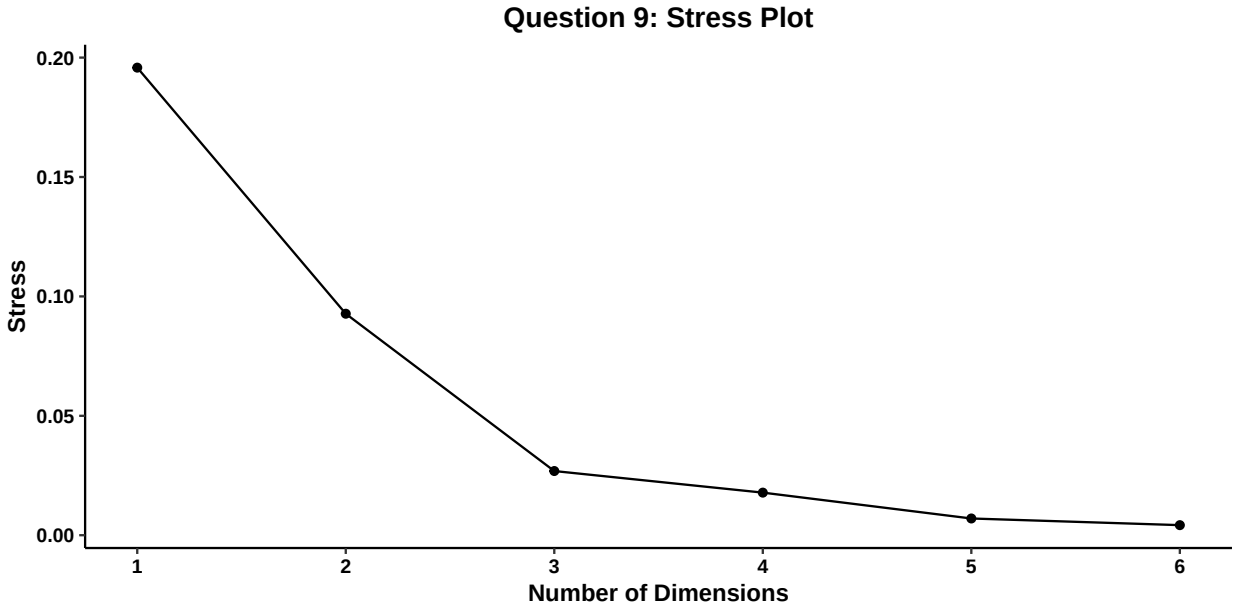


10 Question 9

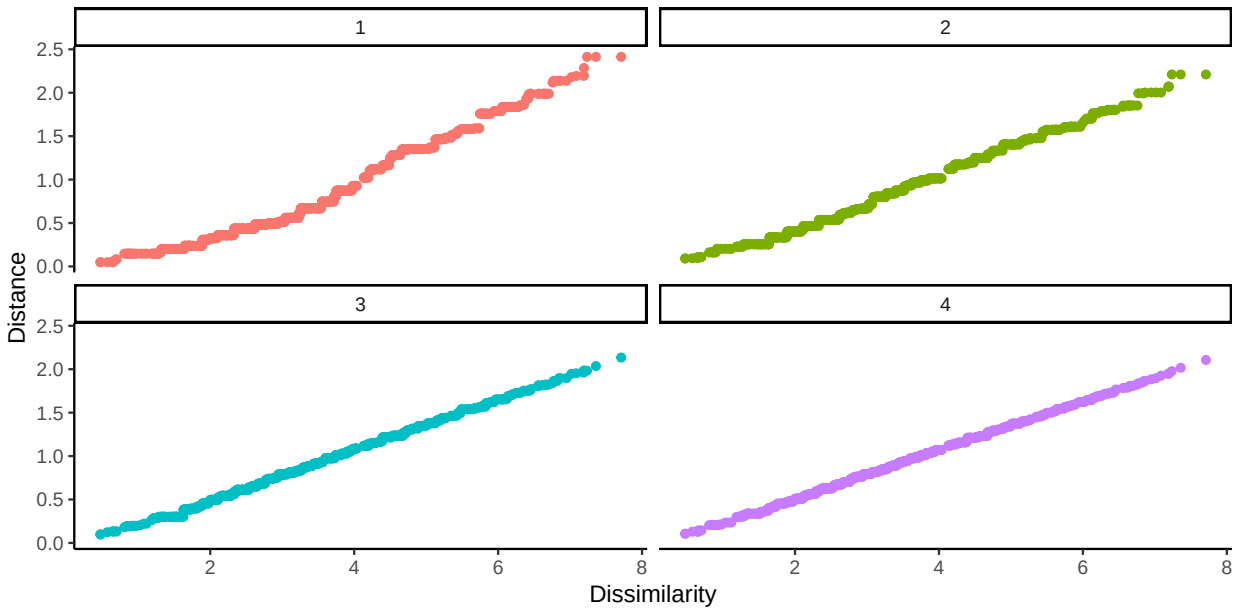
Repeat the analyses using the standardized ratings. Comment on any differences that you observe.

```
# stress plot
nested.dat %>%
  filter(type == "z") %>%
  ggplot(aes(x = dim, y = stress)) +
    geom_line() +
    geom_point() +
    scale_x_continuous(limits = c(1,6), breaks = 1:6) +
    labs(x = "Number of Dimensions", y = "Stress", title = "Question 9: Stress Plot") +
    theme_classic() +
```

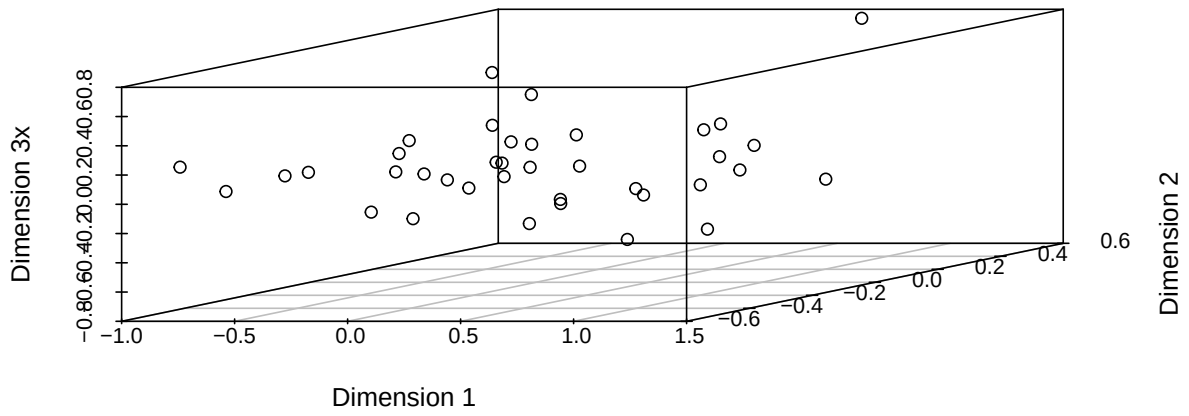
```
    theme(plot.title = element_text(face = "bold", hjust = .5),
          axis.text = element_text(face = "bold", color = "black"),
          axis.title = element_text(face = "bold"))
```



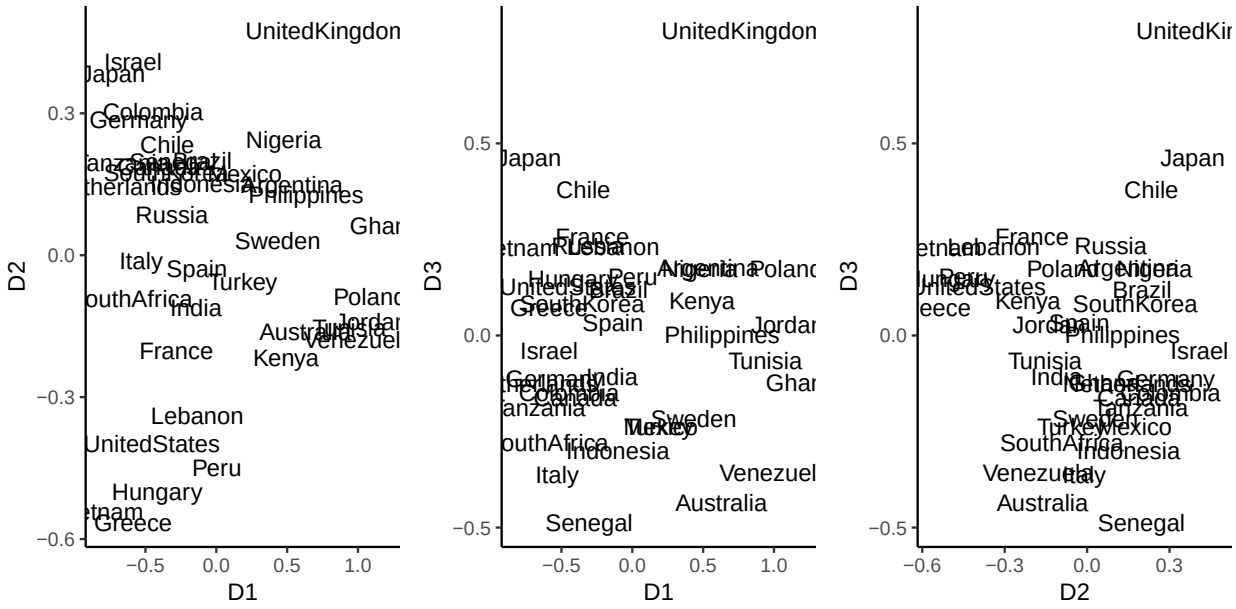

```
# shepard plot
nested.dat %>%
  unnest(shep_val) %>%
  filter(dim <= 4 & type == "z") %>%
  ggplot(aes(x = Dissimilarity, y = Distance, color = factor(dim))) +
    geom_point() +
    facet_wrap(~dim) +
    theme_classic() +
    theme(legend.position = "none")
```



```r
# mds
mds_3_dat_z <- (nested.dat %>%
  filter(dim == 3 & type == "z"))$mds[[1]]$conf %>%
  data.frame

scatterplot3d(x = mds_3_dat_z$D1, y = mds_3_dat_z$D2, z = mds_3_dat_z$D3,
              xlab = "Dimension 1", ylab = "Dimension 2", zlab = "Dimension 3x")
```




```r
mds_3_dat_z <- mds_3_dat_z %>%
  mutate(Country = str_remove_all(unique(dat$Country), " "))

p1 <- mds_3_dat_z %>%
  ggplot(aes(x = D1, y = D2, label = Country)) +
    geom_text() +
    theme_classic()

p2 <- mds_3_dat_z %>%
  ggplot(aes(x = D1, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

p3 <- mds_3_dat_z %>%
  ggplot(aes(x = D2, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

gridExtra::grid.arrange(p1, p2, p3, nrow = 1)
```

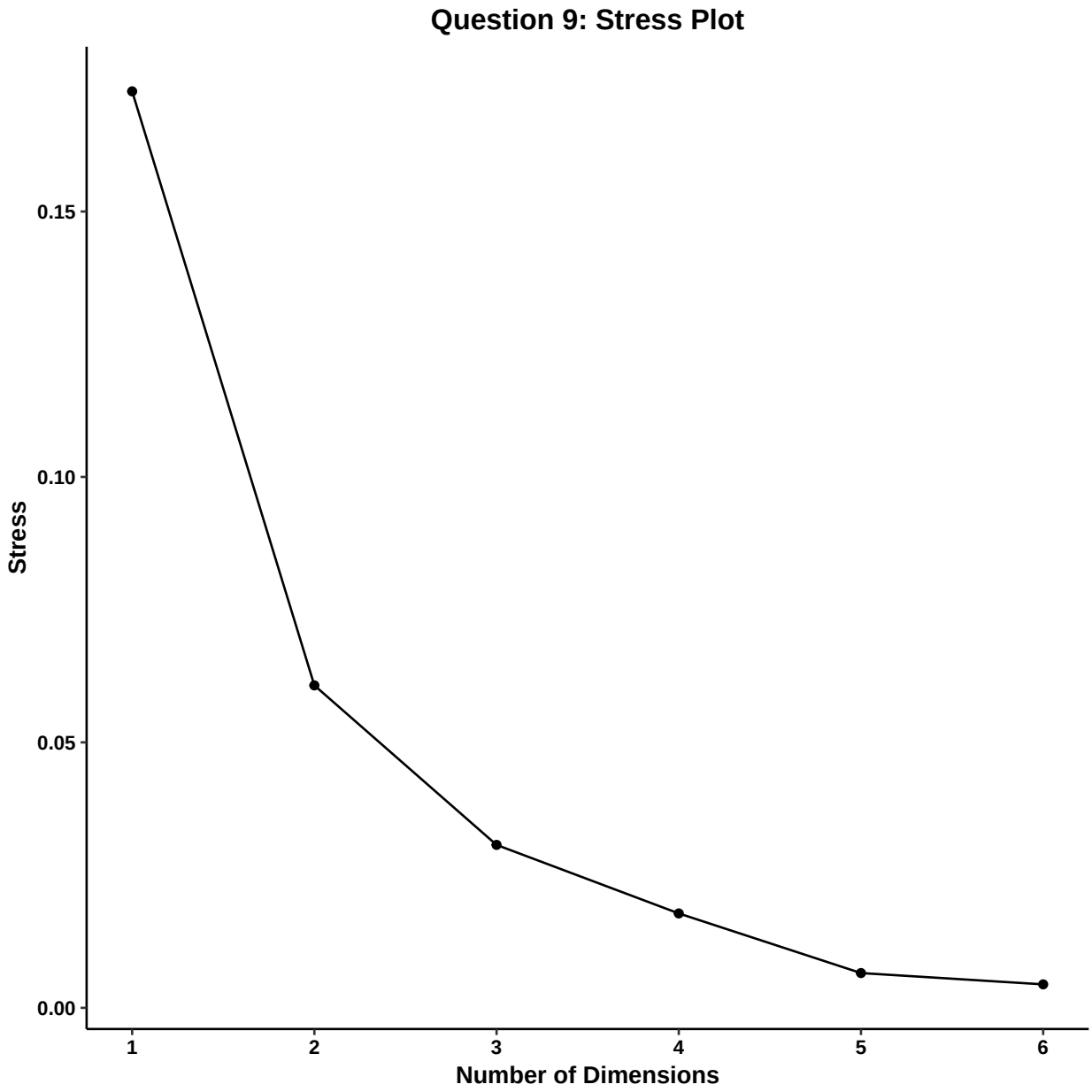


Nothing appears to change when using standardized scores, so the dimension names, etc. would be the same.

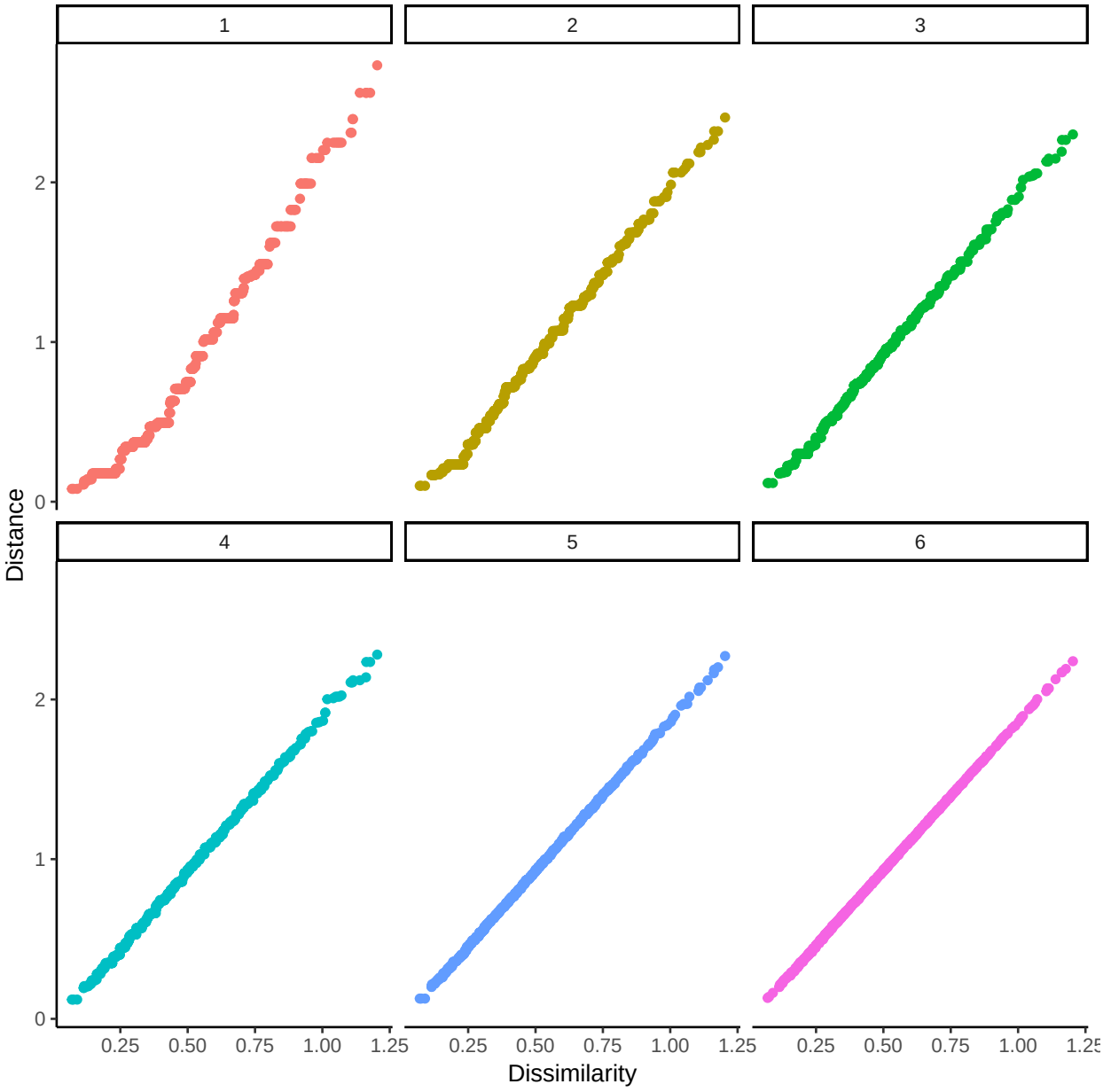
11 Question 10

Repeat the analyses using the probits. Comment on any differences that you observe.

```
# stress plot
nested.dat %>%
  filter(type == "probit") %>%
  ggplot(aes(x = dim, y = stress)) +
    geom_line() +
    geom_point() +
    scale_x_continuous(limits = c(1,6), breaks = 1:6) +
    labs(x = "Number of Dimensions", y = "Stress", title = "Question 9: Stress Plot") +
    theme_classic() +
    theme(plot.title = element_text(face = "bold", hjust = .5),
          axis.text = element_text(face = "bold", color = "black"),
          axis.title = element_text(face = "bold"))
```

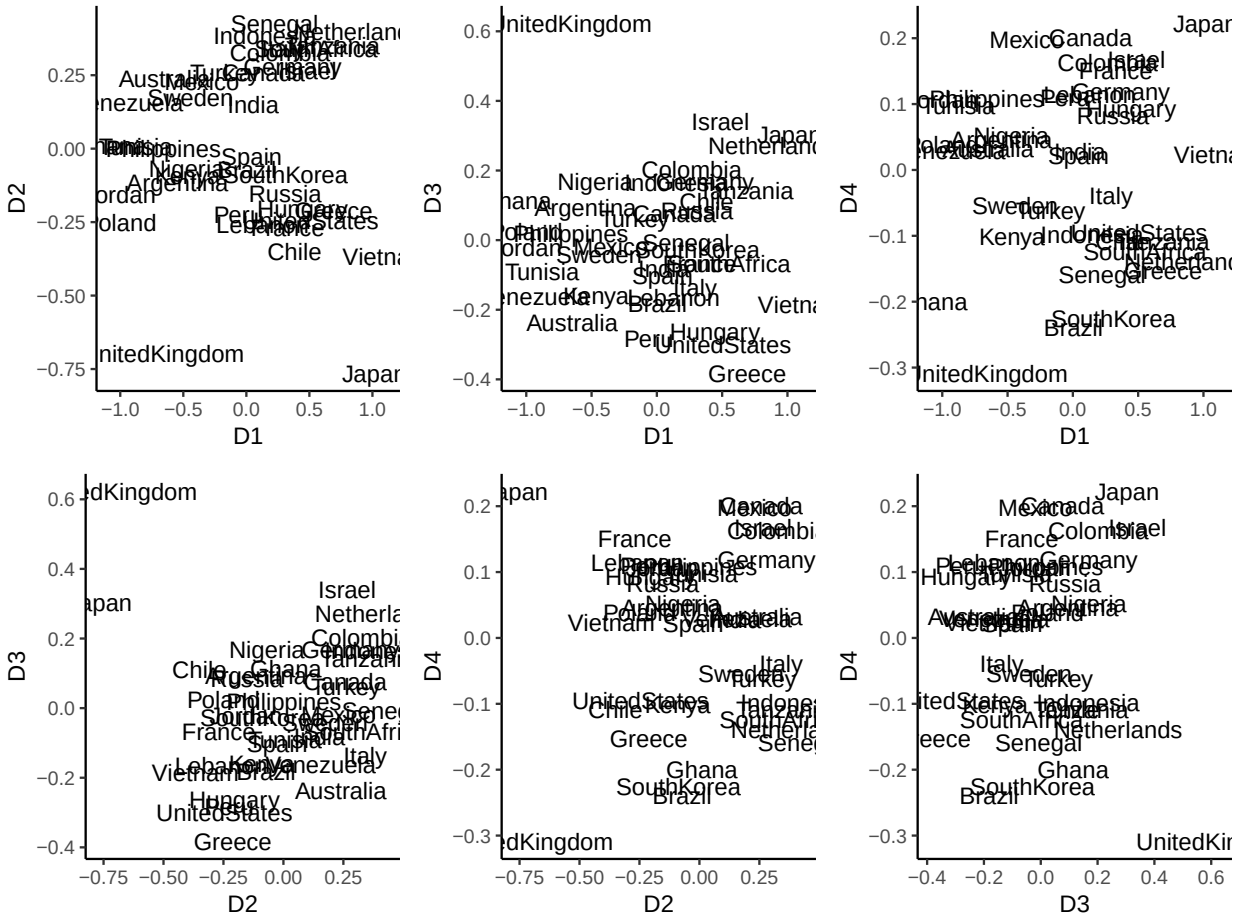



```
# shepard plot
nested.dat %>%
  unnest(shep_val) %>%
  filter(dim <= 6 & type == "probit") %>%
  ggplot(aes(x = Dissimilarity, y = Distance, color = factor(dim))) +
    geom_point() +
    facet_wrap(~dim) +
    theme_classic() +
    theme(legend.position = "none")
```




```
# mds
mds_3_dat_p <- (nested.dat %>%
  filter(dim == 4 & type == "probit"))$mds[[1]]$conf %>%
  data.frame

mds_3_dat_p <- mds_3_dat_p %>%
  mutate(Country = str_remove_all(unique(dat$Country), " "))

p1 <- mds_3_dat_p %>%
  ggplot(aes(x = D1, y = D2, label = Country)) +
    geom_text() +
    theme_classic()

p2 <- mds_3_dat_p %>%
```

```
  ggplot(aes(x = D1, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

p3 <- mds_3_dat_p %>%
  ggplot(aes(x = D1, y = D4, label = Country)) +
    geom_text() +
    theme_classic()

p4 <- mds_3_dat_p %>%
  ggplot(aes(x = D2, y = D3, label = Country)) +
    geom_text() +
    theme_classic()

p5 <- mds_3_dat_p %>%
  ggplot(aes(x = D2, y = D4, label = Country)) +
    geom_text() +
    theme_classic()

p6 <- mds_3_dat_p %>%
  ggplot(aes(x = D3, y = D4, label = Country)) +
    geom_text() +
    theme_classic()

gridExtra::grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2)
```



Using Probit, 4 dimensions appears to be the best solution, both per the Shepard and stress plots. In this soltuion, the UK is finally low on something (Dimensions 1 and 4), suggesting that the fundamental nature of the dimensions has changed. Moreover, on Dimension 4, the countries that clustered together when using raw scores do not seem to cluster together on this dimension.