Exploratory data analysis on patterns of psychedelic mushroom usage in relation to health outcomes.

Gurpreet Singh

Assessment 3 Report. Foundations of Data Science

ABSTRACT

The controversy about using psychedelics safely for mental health is ongoing but this report attempts to shed light on some of the patterns of use and its relevance with various health measures. The purpose of this report is to explore various segments of psychedelic usage and how we can use the discoveries to formulate meaningful advances in mental health and alternative medicine. This can also give us a general idea about the foundations of society and human psychology. For this report the data from a survey conducted by Acumen Health Research Institute's (AHRI) established SOPs. This survey was available to public that were 18 years or older. Various data analysis methods were used in RStudio that enable us to clean data, summarise group wise information, visualize patterns, form subsets of data, and more to extrapolate meaningful elements of data. Some of the findings suggested that the major participation was based around people from age group between 21-40 years and were unemployed although this group would fall in under 100k income. A larger percentage of psychedelic users reported no to mild anxiety on the Patient Health Questionnaire (PHQ). The findings of this report may be helpful in informing the public about responsible use of psychedelic mushrooms especially in younger people with developing signs of depression and anxiety.

INTRODUCTION

Our understanding of nature has helped us understand our place in it as a species. This understanding is not an over-night process but millions of years of adaptations to the resources. From surviving a sabretooth attack to withstanding droughts and from great massacres involving mass killings to deadly pandemics, the will of our great ancestors to survive is why humans are a dominant species on this planet. The origins of fungi and its uses are only as old as we could manage our record keeping as the mycelial group was present before the evolution of human species (Blainey, 2005). There are speculations about the "stoned ape theory" which states that consumption of psychedelic mushrooms in their diet by the nomadic Neanderthals could have contributed a lot in the way we changed our cognitive ways from hunter gatherers to homo-erectus to development of language to a modern soon to be inter-planetary species (Mckenna, 1999). Psychedelic mushrooms have been used by the shamanic cultures and tribes since centuries as an alternative medicine. The psychedelic wave in 1960's slowly settled down when most of the usage was considered as illegal, but the therapeutic benefits of psychedelic substances are resurfacing in the fields of mental health, long-term treatment for anxiety, anxiety associated with terminal illness, alternative medicine, psychotherapy, neuroscience, etc. (Tupper, 2015).

This analysis studies the group wise distribution of the psychedelic users and the general categorization of their physical and mental health. A better society starts with better individuals. This study can clarify the common usage patterns of psychedelic substances and bring forth awareness about the potential constructive usage of the substances to help individuals address mental health conditions and create meaning in their lives. The psychedelic substances are still classified as hallucinogens and their availability and usage is greatly affected by the legality issues. This report acknowledges the use of psychedelic substances is currently illegal in Australia, and therefore does not in any way promote or condone their use.

DATA

This study focuses on the dataset acquired from a cross-sectional nationwide survey of adults in the US in 2021. The survey targeted a representative sample of US adults. The rate of psychedelic use, comorbid conditions, mental and physical health, and factors predictive of psychedelic use were assessed (Merlock, 2021). The stats are tracked with compliance to the Code of Practice for Research Data Usage Metrics and can be compared with other COUNTER complaint repositories (Fenner, 2018).

The survey was conducted in accordance with Acumen Health Research Institute's (AHRI) established SOPs. A random stratified sampling framework ensured a community-based sample with a demographic composition representative of the US adult population by region, gender, age, and race, according to the US Census (US Census American Community Survey 5-year estimate, 2011-2015). To participate in the study, respondents were required to be 18 years old or older, residents of United States, and confirm their voluntarily agreement to participate. The survey was open to the general population. Participants were recruited through AHRI's online research panels. Analysis was carried out with SPSS v27.0.1.0. (Merlock, 2021)

The raw data comprises of 59 variables and 7139 entries. Out of these 59 variables the selected variables are discussed in the metadata table. The dataset comprised of only numeric data out of which certain categorical variables were created. These variables were later randomly sampled with at least 10% of the population following various statistical steps to find correlation and similarity. The metadata below consists of some of the derived variables used for data exploration.

Table 1

Variable Name	Type	Range	Description
Age_cat	Derived, Factor/ Categorical	na	Categorizes age in 20 year brackets
Bmi_cat	Derived, Factor/ Categorical	Na	Categorizes bmi into 4 levels
Psy_User	Derived, Numeric	0 - 180	Sum(PSY_USE_YN). Sum of true instances of PSY_USE_YN
PM_User	Derived, Numeric	0 – 80	Sum(PM_USE_YN). Sum of true instances of PM_USE_YN
Income	Derived, Numeric	50k – 110k	Median(INCOME). Median of the income
Comorbid	Derived, Numeric	1 - 7	Mean(C_TOTAL). Mean of the total Comorbid conditions

C_Score	Dervied, Numeric	1 - 4.5	Mean(C_SCORE). C_SCORE was calculated using the sum of all true instances of conditions of anxiety, depression, chronic pain etc.
CCI	Derived, Numeric	0 – 3	Mean(CCI_SCORE). Mean of the Charlson Comorbidity Index score calculated from self-reported conditions
Gen_AD	Dervied, Numeric	2 - 12	Mean(GAD7_SCORE). Mean of the general anxiety disorder scale.
PHQ	Derived, Numeric	0 – 20	Mean(PHQ9_SCORE). Mean of the Patient Health Questionnaire score
Phy_Health	Derived, Numeric	35 – 55	Mean(PCS12). Mean of the Physical health composite score(Veterans-RAND 12 item)
Men_Health	Derived, Numeric	15 - 55	Mean(MCS12). Mean of the Mental health composite score(Veterans-RAND 12 item)

Table 1. Metadata table containing Variables derived from raw data used for data exploration.

METHODS

To start the exploration of data, the main trends regarding the use of psychedelics were unfolded. The data was available in two *.csv* files in which one contained the variable description and other contained the raw data. The raw data was loaded in RStudio version (3) and viewed for integrity. The data was then checked for any missing values/incorrect values in each variable using *sum(is.na())* and *colMeans(is.na()).* The values in the nominal scale variables were -99 which were imputed as a unique number, so they don't conflict with the other column values. Since only frequency distribution calculations were possible on nominal variables the values were imputed to a unique number as per the requirements of the variable column. We first start with separating the participants by their age using *cut()* and organising them in twenty-year age brackets as nominal labels of "Under 20", "21-40", "41-60","61-80", "Above 80". This can help us identify the main data distribution as per age.

After following the age category, the BMI variable was also plotted (Fig. 1a) and was categorised in four different categories. This categorization was done using a function *bmi_func()* created based on index as per government health website (Metro North Hospital and Health Service(2017). The data was then further dissected into sex and employment status of the participants which revealed that majority of the participation was from unemployed group where the female contributions were more than men (Fig. 1a). All the plotting was carried out using *ggplot()* function in *tidyverse* package.

This brings us to the next sectors of the exploratory analysis follows adaptive loading (Idreos, 2015) which is focused on the participants that answered "*yes*" or have value 1 to the previous psychedelic use or have used psychedelic mushrooms. The raw data contained variables that asked the participants to answer if they experienced any depression, anxiety, chronic pain, migraines, insomnia, sleep apnea, and several other ailments. A new variable called C_SCORE was created that summed up the number of positive answers to any one of those conditions. This was created using the *mutate*() function pipelined with *select()* for variables that *starts_with()* the dplyr package in RStudio. C_Score along with other variables like "Psy_User", "PM_User", "Comorbid", "CCI", "Gen_AD", "PHQ", "Phy_Health" and "Men_Health" were chosen as variables of interest in this analysis. The other variables can be used to explore various other aspects of the data which may be outside the scope of this report. The *sum()* values of true instances were used to identify both mushroom and other psychedelic users whereas the *mean()* values were used for the other variables. This data frame helps us summarize the top users grouped by their age and bmi categories. The same data can be plotted using *geom_point()* where physical and mental health scores can be plotted on x and y axis

respectively. The plot can be colour and size coded to depict most of the mushroom usage as per the bmi categories. The linear model between variable "PM_User" and other variables like "Gen_AD" was also plotted to predict if see mushroom usage could be helpful in predicting Patient Health Questionnaire Score (Fig. 2).

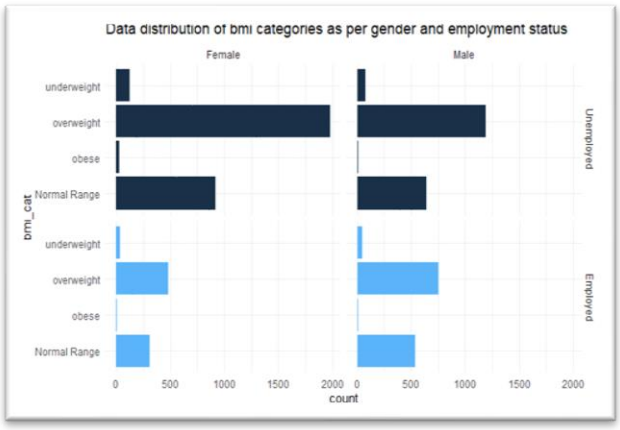
This next exploration can be based on the correlation between variables of interest between two of more age groups. As from previous findings, the top users of psychedelic mushrooms fall in the age group of "21-40" and "41-60". A correlation can be found between the two by creating two data frames by filtering their age using *filter().* The population was checked with *nrow()* and the data frames were then randomly sampled with at least 10% of the population of the larger matrix. The correlation between the matrices were calculated using the *cor()* function which helps us understand if there is a linear relationship between the two quantitative variables (Makowski, 2020). This can tell us about the strength of the correlation.

We can also look at the similarities between the users within the age group using the *heatmap()* function by using the Euclidean method (Reichstetter, 2022). This can enable us to understand hierarchical clustering where the darker clusters show higher associations between participants.

RESULTS AND DISCUSSION

As the data was explored deeper, new discoveries lead to some interesting information about the data. Most of the participation in the survey was from the unemployed group, aged 21 to 40 years and then from 41 to 60 years. Females contributed more than men and were mostly overweight to normal range on the BMI scale as per Figure 1a.

From the participants only the users with any psychedelic usage history were selected. This filtered dataset suggested that 80.34 percent users had answered "yes" to having at-least one medical condition related to mental or physical health. The top users were of the age group of 21- 40 years and the mean income of the group was roughly under 60k. They answered positively to having almost at-least 3 (mean value of 2.89) medical or physical condition within the last year.



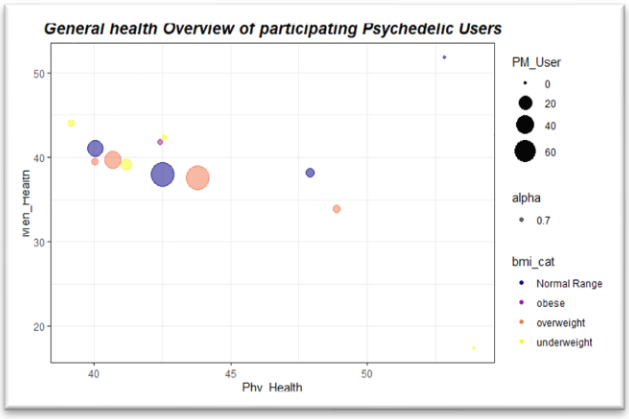


Figure 1a Top Data Distribution

1b Bottom General health Overview

After creating a linear model on the selected data, the mushroom usage and general anxiety scores were plotted. The model represents a positive relationship from the age group of 21-40 as compared to the other groups which suggests that a greater number of mushroom users reported having a high Gen_AD score than other age groups (Fig. 2).

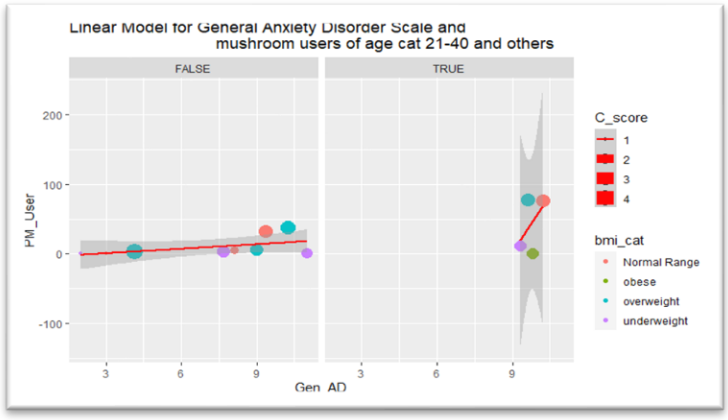


Figure 2. Linear Model for Mushroom Users

The correlation between the two age groups showed minor similarities in variables like Physical health scores, CCI score, Mental health scores and General Anxiety Disorder scores as shown in the darker shaded sections of the heatmap in figure 3. The next discovery was related to the sampled dataset of the age group 21-40 that had used any psychedelics within the last year. We can see some signs of clustering which is indicative of higher similarity between some users of the age group 21-40 (Fig. 3b).

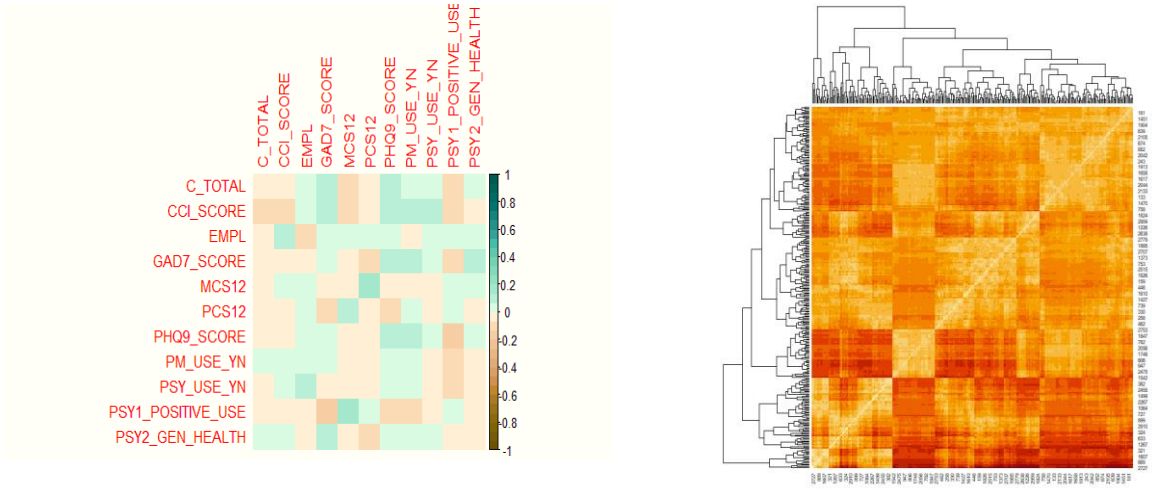


Figure 3a. Corrplot of the correlational matrix using Pearson method (left)
Figure 3b. Heatmap of the similarity matrix using Euclidean method (right)

CONCLUSION

To conclude, the findings of this report may be indicative of a pattern of higher use of psychedelic mushrooms among younger people. There was not enough correlational strength in the two age groups chosen for this analysis, although other variables may lead to a different correlation matrix. People who scored higher on their general anxiety score (Gen _AD) were also high users of psychedelic mushrooms in the survey and if the mushrooms were beneficial to their health remains unclear. There are currently several clinical trials looking at outcomes of psychedelic mushroom use in the treatment of anxiety related mental health conditions (Brian, 2022). These results will help to inform the public and health care industry about the benefits and safety considerations of their use.

REFERENCES

Blainey, M. G. (2005). Combining Social Cohesion Theories with Altered States of Consciousness to Explain the Adaptive Advantages of Spiritual Capacity in Humans. The University of Western Ontario Journal of Anthropology, 13(1).

Brian S. Barnett, Sloane E. Parker, Jeremy Weleff, (2022). *United States National Institutes of Health grant funding for psychedelic-assisted therapy clinical trials from 2006–2020*, International Journal of Drug Policy, Volume 99, 103473, ISSN 0955-3959, https://doi.org/10.1016/j.drugpo.2021.103473.

Fenner M, Lowenberg D, Jones M, Needham P, Vieglais D, Abrams S, Cruse P, Chodacki J. 2018. Code of practice for research data usage metrics release 1. PeerJ Preprints 6:e26505v1 https://doi.org/10.7287/peerj.preprints.26505v1

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015, May). Overview of data exploration techniques. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 277-281).

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Methods and algorithms for correlation analysis in R. Journal of Open-Source Software, 5(51), 2306.

McKenna, T. (1999). Food of the gods: the search for the original tree of knowledge: a radical history of plants, drugs, and human evolution. Random House.

Morlock, Robert. (2021). Data from: Psychedelic mushrooms in the USA: knowledge, patterns of use, and association with health outcomes [Data set]. https://doi.org/10.5061/dryad.bzkh189b6

Reichstetter, M. (2022). Collaborate Sessions. *Topic 2: Proximity Measures, Lecture notes, Foundation of Data Science*. James Cook University, February 2022.

RStudio 2021.09.0+351 "Ghost Orchid" Release (077589bcad3467ae79f318afe8641a1899a51606, 2021-09-20) for Windows Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.8 Chrome/69.0.3497.128 Safari/537.36

Using Body Mass Index – Metro North Hospital and Health Service(2017). https://www.health.qld.gov.au/__data/assets/pdf_file/0031/147937/hphe_usingbmi.pdf

Tupper, K. W., Wood, E., Yensen, R., & Johnson, M. W. (2015). Psychedelic medicine: a re-emerging therapeutic paradigm. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 187*(14), 1054–1059. https://doi.org/10.1503/cmaj.141124

APPENDIX (R Code)

```
# Install packages and load libraries
install.packages("tidyverse")
install.packages("esquisse")
install.packages("cluster")
```

```
install.packages("plotly")
install.packages("viridisLite")
install.packages("class")
library(ggplot2)
library(esquisse)
library(tidyverse)
library(dplyr)
library(cluster)
library(plotly)
library(viridis)
library(class)
library(corrplot)

# Load Data into R
data <- read.csv(file.choose())
View(data)
glimpse(data)
summary(data)

# Check for missing values
sum(is.na(data))
mis_data <- colMeans(is.na(data))
mis_data

# Data imputation (only columns with the non-conflicting numeric values were chosen)
data["PM3_FREQ_POLITICS"][data["PM3_FREQ_POLITICS"] == -99] <- 5
data["PM2_FREQ_COVID"][data["PM2_FREQ_COVID"] == -99] <- 5
data[data == -99] <- 3

# Creating Age, BMI and Income categories
#-------- AGE
age_brackets <- c(1,20,40,60,80,max(data$AGE))
age_labels <- c("Under 20", "21-40", "41-60","61-80", "Above 80")
Age_cat <- cut(data$AGE, breaks = age_brackets, labels = age_labels)

data <- cbind(data, Age_cat)

age_plot <- data %>%
        group_by(Age_cat) %>%
        ggplot() + geom_bar(mapping = aes(x = Age_cat, fill = Age_cat)) +
        labs( title = "Data Distribution for participations based on Age")

# Visualization for Age wise Data Distribution
age_plot

# ------BMI
bmi_func <- function(bmi){
  if(bmi <= 18.5){return("underweight")}
  else if(bmi > 18.5 & bmi < 24.9){return("Normal Range")}
  else if(bmi >= 25){return("overweight")}
  else{return("obese")}
}
```

```
bmi_cat <- sapply(data$BMI,bmi_func)

bmi_cat


data <- cbind(data, bmi_cat)


View(data)


# ------- Income


range(data$INCOME)

income_brackets <- c(1, 20000, 50000, 100000, 150000, 200000, max(data$INCOME))

income_labels <- c("Under 20k", "under 50k", " under 100k","Under 150k", "Under 200k", "Above 200k")

income_cat <- cut(data$INCOME, breaks = income_brackets, labels = income_labels)


data <- cbind(data, income_cat)


# Plot the BMI distribution for the data


library(ggplot2)


label1 <- c(`0` = "Unemployed", `1` = "Employed")

label2 <- c(`0` = "Female", `1` = "Male")


ggplot(data) +

 aes(y = bmi_cat, fill = EMPL) +

 geom_bar(position = position_dodge(0.9)) +

 scale_fill_gradient() + labs(title = "Data distribution of bmi categories as per gender and employment status") +

 theme_minimal() + theme(legend.position = "none") +

 facet_grid(rows = vars(EMPL), cols = vars(SEX),

            labeller = labeller(.rows = label1, .cols = label2))



# Plot based on group based summaries for Users Only

user_hist <- data %>%

  select(starts_with("C_")) %>%

  mutate(C_SCORE = rowSums(user_hist[, -1]))


psy_data <- data %>%

            filter(PSY_USE_YN == 1) %>%

            group_by(Age_cat, bmi_cat) %>%

            summarise(Psy_User = sum(PSY_USE_YN == 1),

                      PM_User = sum(PM_USE_YN == 1),

                      Income = median(INCOME),

                      Comorbid = mean(C_TOTAL),

                      C_score = mean(`user_hist$C_SCORE`),
```

```
                    CCI =
mean(CCI_SCORE),
                    Gen_AD =
mean(GAD7_SCORE),
                    PHQ =
mean(PHQ9_SCORE),
                    Phy_Health =
mean(PCS12),
                    Men_Health =
mean(MCS12), .groups = "drop")

summary(psy_data)

View(psy_data)


top_users <- arrange(psy_data,
desc(PM_User))


# lm model

lin_mod <- lm(PM_User ~ Men_Health,
data = psy_data)

plot(psy_data$Men_Health,
psy_data$PM_User)

abline(lin_mod)


library(ggplot2)


ggplot(data = psy_data, aes(y = PM_User,
x = Gen_AD, col = bmi_cat, size =
C_score)) + geom_point() +

  stat_smooth(method = "lm", col = "red")
+

  labs(title = "Linear Model for General
Anxiety Disorder Scale and

                    mushroom users of age
cat 21-40 and others ") +

  facet_grid(cols = vars((Age_cat=="21-
40")))


# Proportion of Participants with any
history of medical condition.

user_no_hist <-
colSums(user_hist==0)/nrow(user_hist) *
100

user_yes_hist <-
colSums(user_hist>=1)/nrow(user_hist) *
100


qs <- rbind(user_no_hist,user_yes_hist)

data <- cbind(data, user_hist$C_SCORE)


psy_plot <- ggplot(psy_data) + aes( x =
Phy_Health,
                    y = Men_Health,
                    colour = bmi_cat,
                    size = PM_User,
                    alpha = 0.7) +
        geom_point(shape = "circle") +
scale_color_viridis_d(option = "plasma",
direction = 1) +
        scale_size(range = c(1, 10)) +
        labs(title = "General health
Overview of participating Psychedelic
Users") + theme_bw() +
        theme(plot.title = element_text(size
= 15L, face = "bold.italic", hjust = 0.5))


psy_plot


# Subsetting Data and checking for
correlation between two age groups
```

```
mat_2040 <- data %>%
        select(PSY_USE_YN,
PM_USE_YN, C_TOTAL, CCI_SCORE,
GAD7_SCORE,
            PHQ9_SCORE, PCS12,
MCS12, PSY1_POSITIVE_USE,
            PSY2_GEN_HEALTH,
EMPL) %>%
        filter(Age_cat == "21-40")


sam_2040 =
as.matrix(mat_2040[sample(nrow(mat_20
40), 300, replace = FALSE), ])


mat_4060 <- data %>%
        select(PSY_USE_YN,
PM_USE_YN, C_TOTAL, CCI_SCORE,
GAD7_SCORE,
            PHQ9_SCORE, PCS12,
MCS12, PSY1_POSITIVE_USE,
            PSY2_GEN_HEALTH,
EMPL) %>%
        filter(Age_cat == "41-60")


sam_4060 <-
as.matrix(mat_4060[sample(nrow(mat_40
60), 300, replace = FALSE), ])
```

```
cor_mat <- cor(sam_2040, sam_4060,
method = "pearson")


heatmap(cor_mat, Rowv = NA, scale =
"column",symm = T)


corrplot(cor_mat, method = "color", order
= "alphabet", col = COL2('BrBG', 30))


# Creating a distance matrix of the sample
and plotting it for similarities in the
participants.


dist_mat<- as.matrix(dist(sam_2040,
method = "euclidean", diag = T, upper =
T))

heatmap(dist_mat, scale = "column")

image(dist_mat, col = heat.colors(12))



dist_mat2 <- as.matrix(dist(sam_4060,
method = "euclidean", diag = T, upper =
T))

heatmap(dist_mat2, scale = "column")


# ----------
```