Homework 5 - Unsupervised

Sicily Xie #54385315

Due 6 December at 11pm

0.0.1 Instructions

- Use the space inside of

```
::: {.solbox data-latex=""}


:::
```

to answer the following questions.

- Do not move this file or copy it and work elsewhere. Work in the same directory.
- Use a new branch named whatever you want. Create it now! Can't come up with something, try here. Make a small change, say by adding your name. Commit and push now!
- Try to Knit your file now. Are there errors? Fix them now, not at 11pm on the due date.
- There MUST be some text between `::: {.solbox}` and the next `:::` or this will fail to Knit.
- If your code or figures run off the .pdf, you'll lose 2 points automatically.

Introduction

Throughout this assignment, we will be looking at a data set of Whisky flavour profiles. David Wishart was at one point, Chief Statistician at the Scottish Office of the Civil Service. When he retired, he focused on Whisky and wrote a book called "Whisky Classified". For the book, he collected tasting notes published about 86 different Scotch Whisky distilleries on a number of aspects and "distilled" them down to 12 flavor categories. Then each distillery's representative whisky was given a score on each category from 0-4, 0 meaning that that flavor is not represented in that whisky, 4 meaning that it is strongly represented.

The data set was later expanded to include more distilleries and crowd-sourced tasting notes, but, this data seems to be kept only in a for-profit Windows software which no longer exists. Dr. Wishart passed away in 2020, and there seems to be no way to access the larger data set.

You have here, the version of the data from the first edition of Dr. Wishart's book. An article describing some of his analyses with 185 single malts was published in 2009 in Significance.

We will undertake some similar analyses in this homework assignment. A snapshot of the data is shown below.

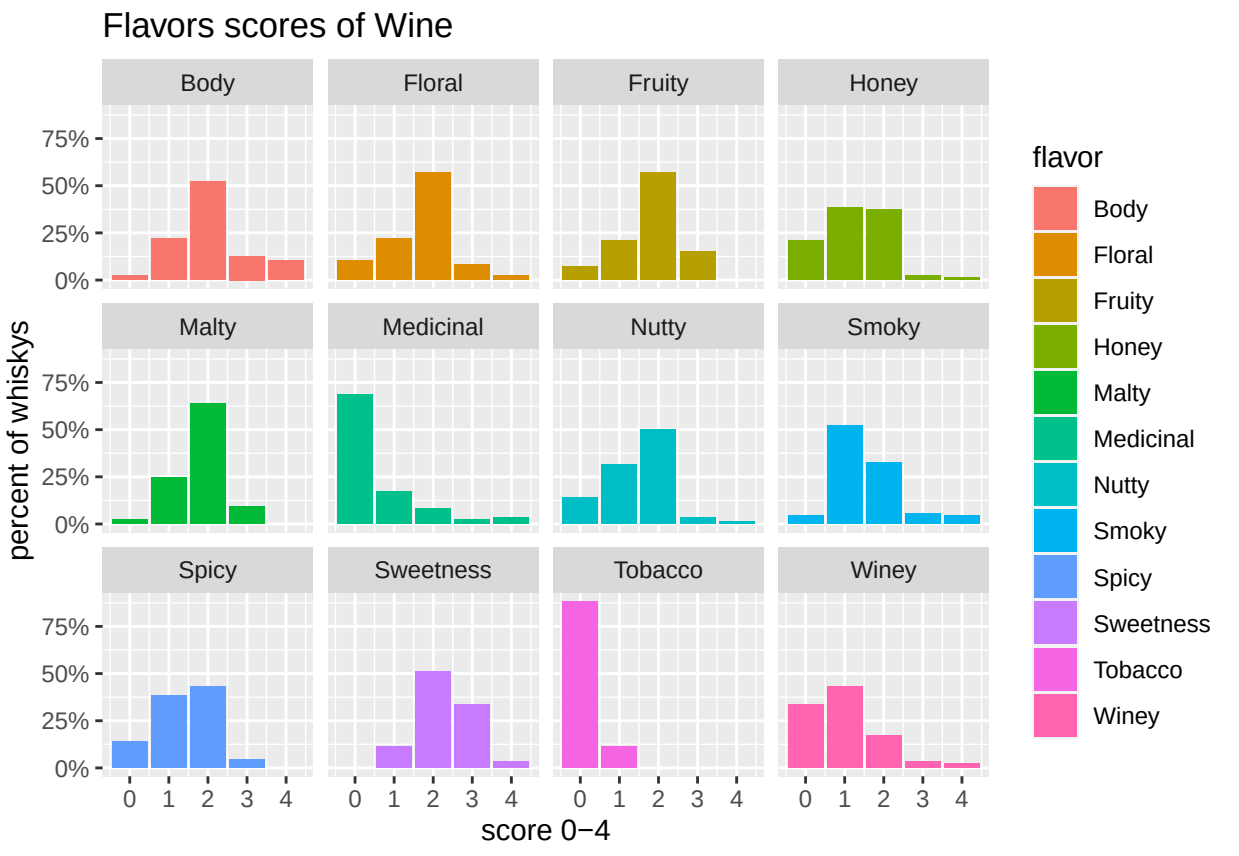
```
## tibble [86 x 16] (S3: tbl_df/tbl/data.frame)
##  $ Distillery: chr [1:86] "Aberfeldy" "Aberlour" "AnCnoc" "Ardbeg" ...
##  $ Body      : num [1:86] 2 3 1 4 2 2 0 2 2 2 ...
##  $ Sweetness : num [1:86] 2 3 3 1 2 3 2 3 2 3 ...
```

```
##  $ Smoky     : num [1:86] 2 1 2 4 2 1 0 1 1 2 ...
##  $ Medicinal : num [1:86] 0 0 0 4 0 1 0 0 0 1 ...
##  $ Tobacco   : num [1:86] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Honey     : num [1:86] 2 4 2 0 1 1 1 2 1 0 ...
##  $ Spicy     : num [1:86] 1 3 0 2 1 1 1 1 0 2 ...
##  $ Winey     : num [1:86] 2 2 0 0 1 1 0 2 0 0 ...
##  $ Nutty     : num [1:86] 2 2 2 1 2 0 2 2 2 2 ...
##  $ Malty     : num [1:86] 2 3 2 2 3 1 2 2 2 1 ...
##  $ Fruity    : num [1:86] 2 3 3 1 1 1 3 2 2 2 ...
##  $ Floral    : num [1:86] 2 2 2 0 1 2 3 1 2 1 ...
##  $ Postcode  : chr [1:86] "PH15 2EB" "AB38 9PJ" "AB5 5LI" "PA42 7EB" ...
##  $ Longitude : num [1:86] -3.85 -3.23 -2.79 -6.11 -2.74 ...
##  $ Latitude  : num [1:86] 56.6 57.5 57.4 55.6 57.4 ...
```

1 Dimension reduction (6 points)

1. (0.5 points) Create a bar chart that shows 12 panels, one for each of the twelve flavor profiles. In each panel, the heights of the bars should be the percent of whiskys with each score 0-4 (out of 86 total) (the y-axis is a percent between 0 and 100). Make sure that your graphic fits on 1 page.

```
# some code
panels_dat <- whisky %>% select(Body:Floral) %>% pivot_longer(cols = everything(),
names_to = "flavor", values_to = "score")
panels_dat %>%
ggplot(aes(score, y = after_stat(count)/sum(after_stat(count)) * 12, fill = flavor))+
facet_wrap(~ flavor) + geom_bar() +
labs(title = "Flavors scores of Wine",
x = "score 0-4",
y = "percent of whiskys") +
scale_y_continuous(labels = scales::percent_format())
```

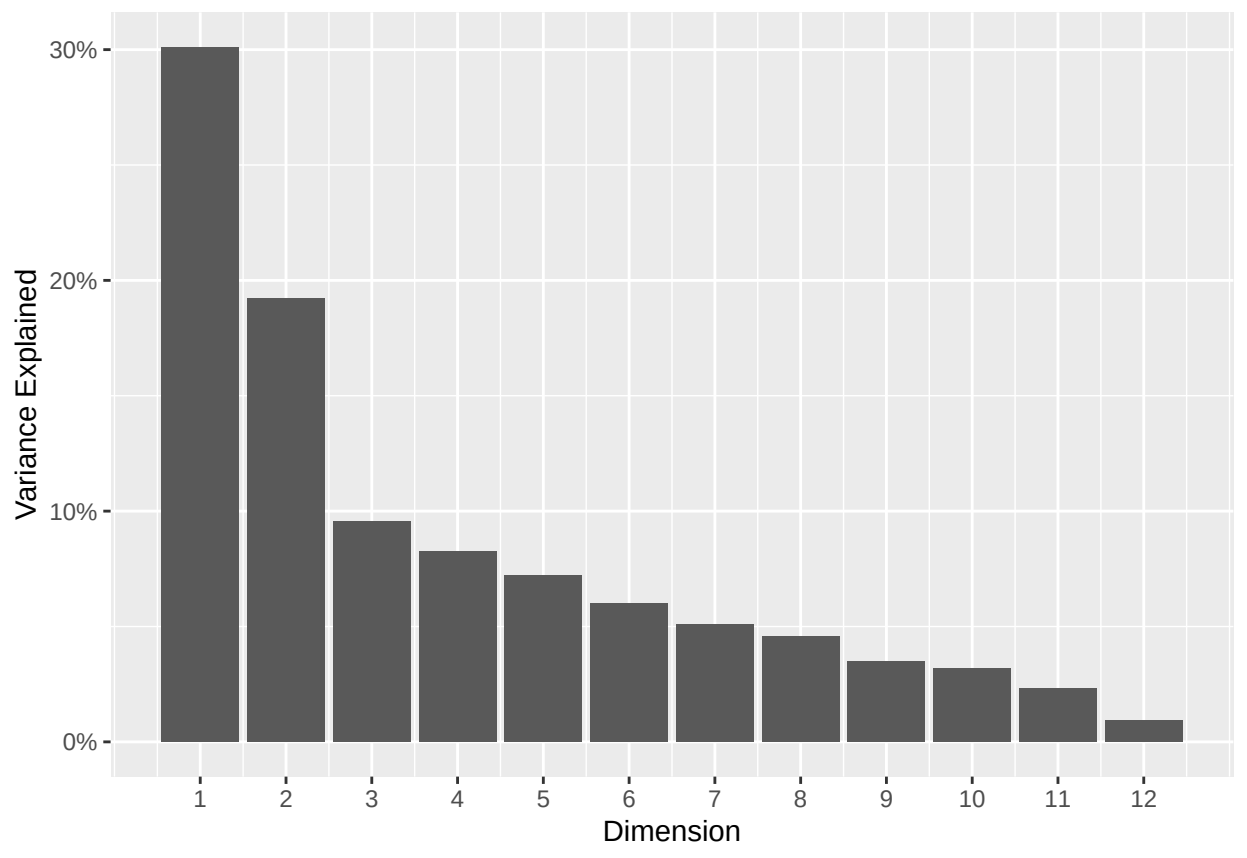


2. (0.5 points) Charts should always be described in any analysis or else they are useless. Examining your chart, describe any patterns you see in a few sentences (at most 3). There are many right answers. Wrong answers are things like "there are 12 panels" or "each panel has 5 bars" or "I don't see anything".

We can notice most wines miss the "Tobacco" flavor, with very few at a low percentage and most taste fruity, floral, and a bit malty as well.

3. (1 point) Perform PCA on the 12 flavor categories. Produce a scree plot with the y-axis displaying the percent of total variance explained by each component. Based on your plot, is 2 a reasonable number of dimensions to use? Why or why not?

```
# some code
X <- whisky %>% select(Body:Floral)
pca <- prcomp(X)
pca_dat <- data.frame(num_pred = seq(1, 12),
var_exp = (pca$sdev)^2) %>% mutate(var_exp = var_exp / sum(var_exp))
ggplot(pca_dat, aes(x = num_pred, y = var_exp)) +
geom_bar(stat = "identity") +
scale_x_continuous(breaks = seq(1, 12, 1)) +
scale_y_continuous(labels = scales::percent_format()) +
labels(x = "Dimension", y = "Variance Explained")
```



Yes. Two clusters still seem reasonable according to our plot since the variance explained seems to decrease more flatly after 2.

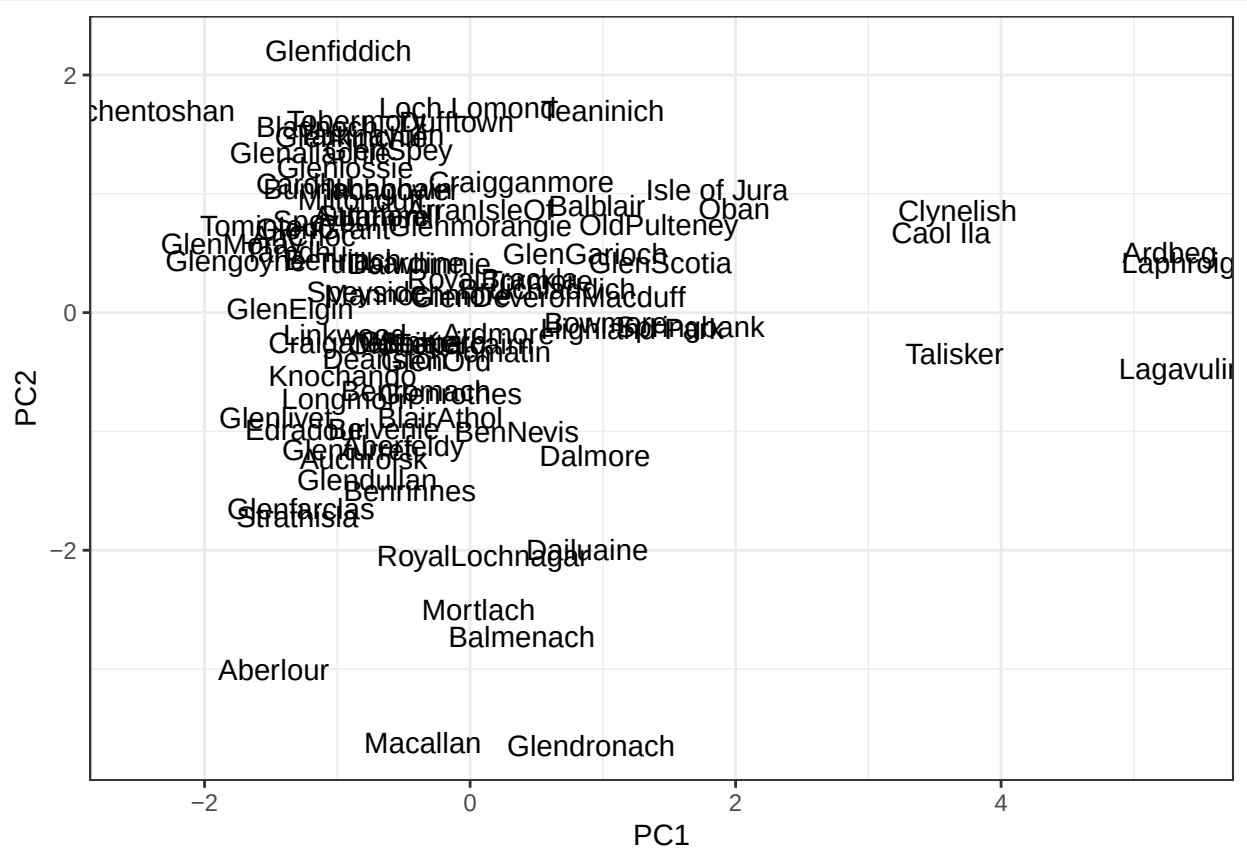
4. (1 point) Produce a plot of PC1 vs PC2. Instead of a dot for each distillery, show the name of the distillery in the figure. (You'll have lots of words, some of which may be hard to read). Here's example code to get the idea.

```
plot(1:5,1:5, pch = "")
text(1:5,1:5, labels = letters[1:5])
```

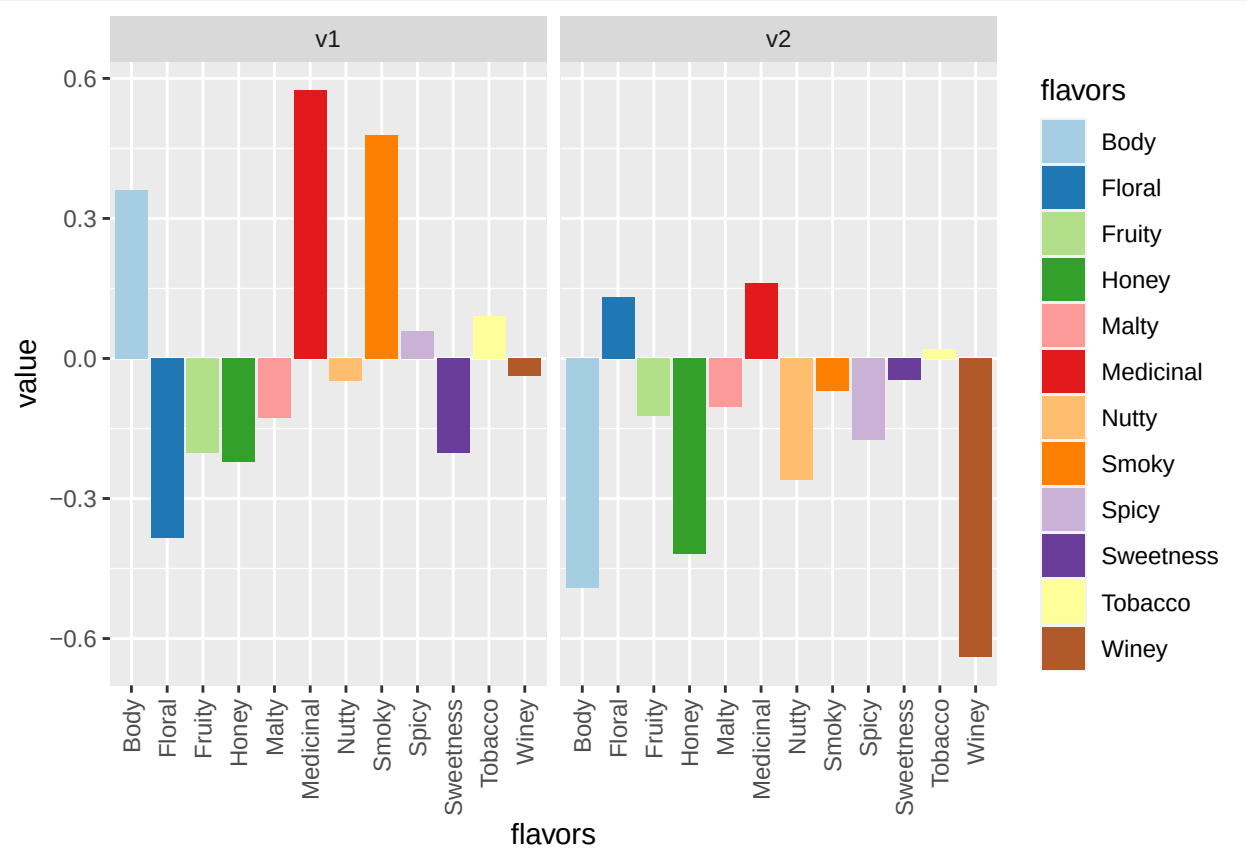
In `ggplot`, you would use `geom_text()` or `geom_label()` rather than `geom_point()`.

Also produce a plot of the Loadings for the first two PCs. Which qualities seem important (positive or negative) for each PC (describe this in 1-2 sentences)? Do any Distilleries stick out, if so, what qualities do they have in common?

```
# Code to plot pc1 against pc2
two_pcs <- data.frame(pca$x[, 1:2]) %>% mutate(Distillery = whisky$Distillery)
two_pcs %>% ggplot(aes(PC1, PC2, label = Distillery)) +
theme_bw() + geom_text()
```



```r
# code to plot the weights
weights <- data.frame(v1 = pca$rotation[,1], v2 = pca$rotation[,2], flavors = rownames(pca$rotation))
weights %>% pivot_longer(-flavors) %>%
ggplot(aes(y = value, x = flavors, fill = flavors)) + facet_wrap(~name) +
geom_bar(stat = "identity") + scale_fill_brewer(palette = "Paired") +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



In v1, Body, Medicinal, and Smoky seemed to be important positively. Floral is more important negatively, while Fruity and Honey Sweetness are less critical. In v2, most are negative, like Body, Honey, Nutty, and Winey. Then some distilleries that stick out are Macallan, Glendronach, Lagavulin, Ardbeg, and Glenfiddich, mostly smoky and winey.

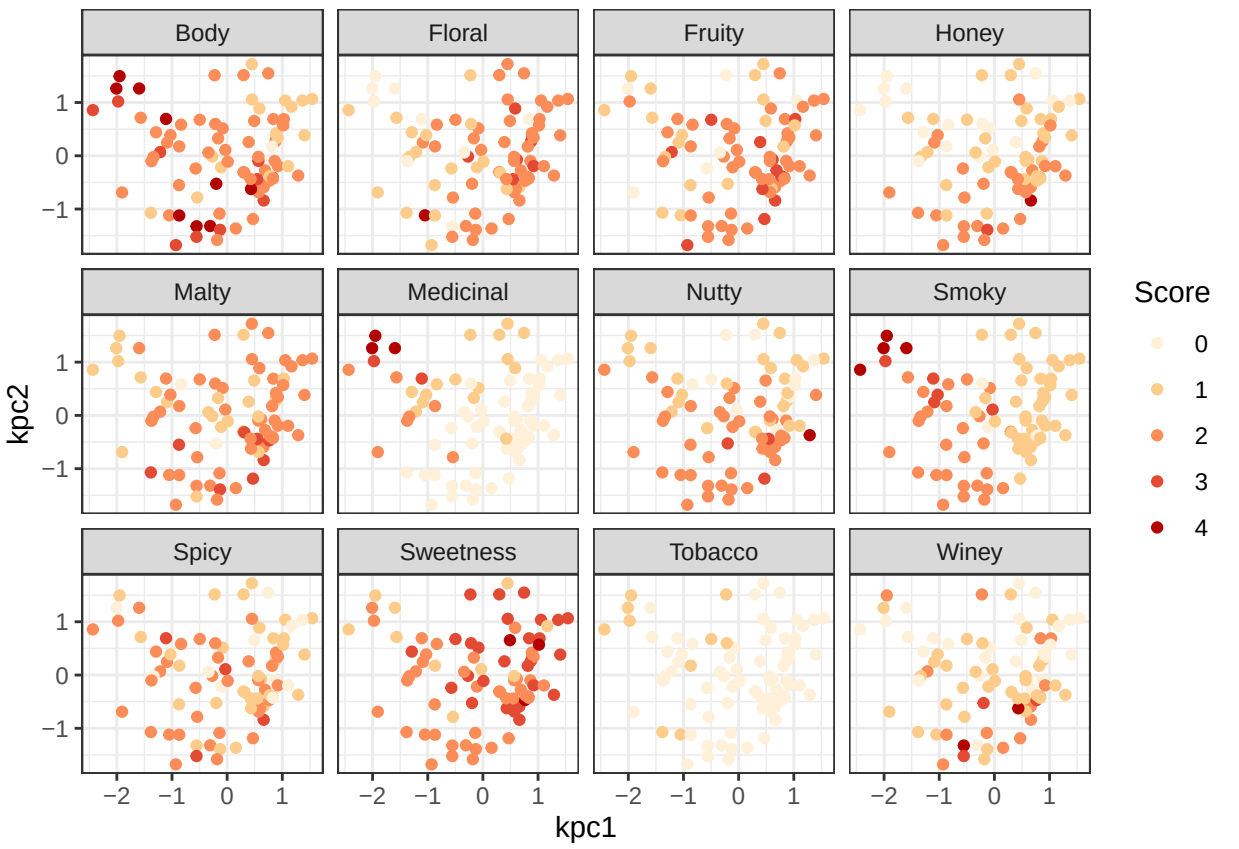
5. (2 points) Complete the function below. It has 3 arguments, (a) a symmetric, non-negative definite Kernel matrix `K`, (b) a target embedding dimension `M`, (c) a tolerance `tol` to test for negative definiteness. Your function must return a matrix of dimension `n` x `M` where `n` is the number of rows of `K`. Do not change the function signature.

   You should make your function produces errors or warnings if invalid inputs are passed. This is easiest way to do with `stop()` or `stopifnot()` or `warning()`. Try examining the documentation for those functions. Looking at the tests should help you determine what sorts of checks to perform and what warnings to throw.

```
kpca <- function(K, M = 2, tol = -1e-8) {
# validating inputs
stopifnot(tol < 0)
stopifnot(nrow(K) == ncol(K))
stopifnot(all(K == t(K)))
if (floor(M) < 1) {
M <- 1
warning("reset to 1")
}
n <- nrow(K)
P <- diag(1, n, n) - 1 / n
K <- P %*% K %*% P
e <- eigen(K)
U <- e$vectors[, 1:M]
D <- diag(sqrt(e$values[1:M]), M, M)
U %*% D
}
```

6. (1 point) Use your `kpca()` function to estimate kernel PCA. Form your `K` by first calculating the distance matrix between rows in your data using the `canberra` distance. The easiest way is with the `dist()` function. You have to convert the result to a matrix with `as.matrix()`. Then set `K` to be `1 -` this result. Pass that in to your function. Use `M = 2`.
   Now the hard part, produce a 12 panel plot as in 1. The x-axis should be the first component, the y-axis should be the second component. Color the points based on the flavour score (0-4). Describe any patterns you notice in 2-3 sentences.

```
# Some code
dist_mat <- dist(scale(X, center = TRUE, scale = TRUE), method = "canberra") %>% as.matrix()
ks <- kpca(K = 1 - dist_mat, M = 2)
est_kpca <- tibble(kpc1 = ks[, 1],
kpc2 = ks[, 2], whisky[,2:13]) %>% pivot_longer(3:14,
names_to = "flavor",values_to = "val") %>% mutate(val = factor(val))
est_kpca %>% ggplot(aes(kpc1, kpc2, color = val)) +
geom_point() + facet_wrap(~flavor) + scale_color_brewer(palette = "OrRd") +
theme_bw() + labs(color = "Score")
```

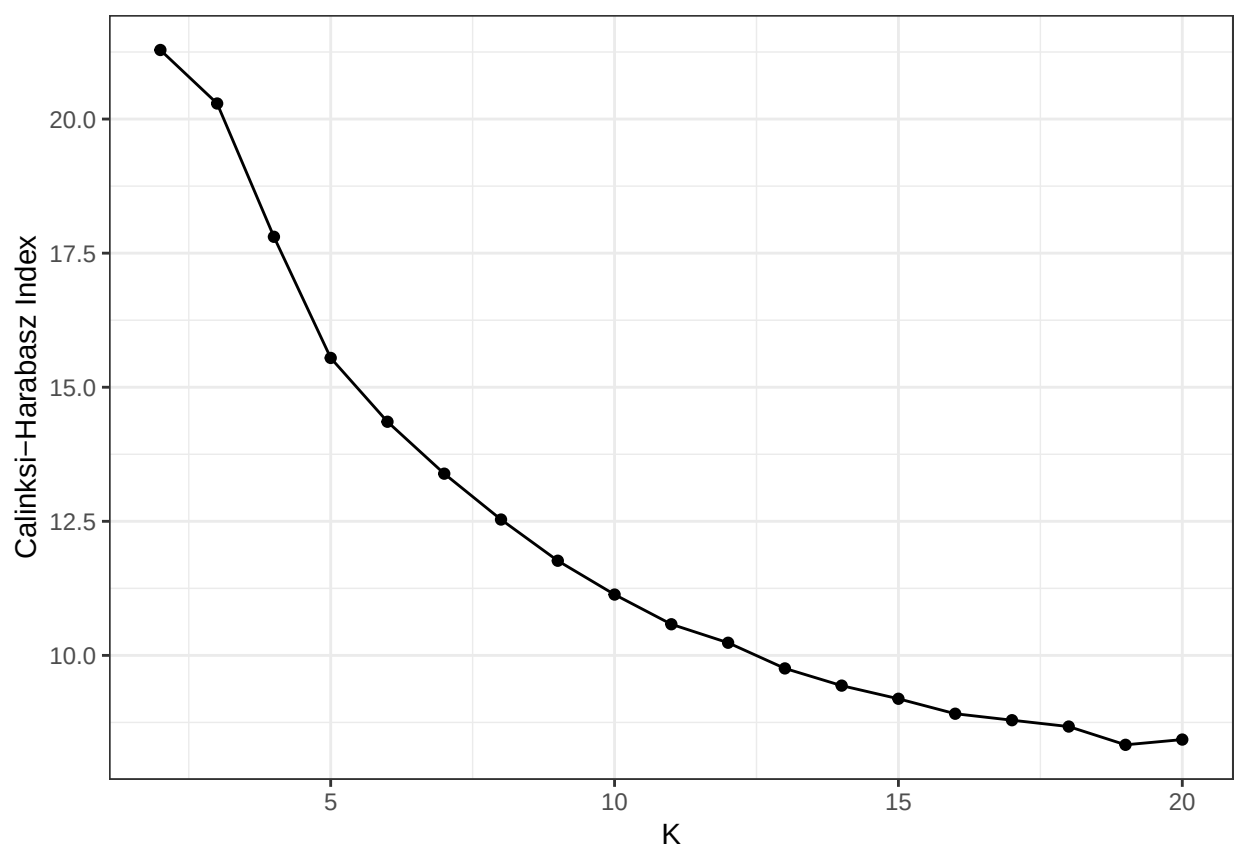


Body whiskys have relatively high scores on kpc2. Honey have low scores on kpc1. Tobacco is low on kpc1 and kpc2. Sweetness is high on both kpc1 and kpc2.

2 Clustering (4 points)

1. (1 point) Use `kmeans()` to cluster whiskys using the 12 flavour metrics. Try K from 2 to 20. Produce a plot of the CH Index against K. Use 20 restarts.

```
set.seed(406406406) # don't change me
K <- 2:20
ch_idx <- double(19L)
for (k in K) {
km <- kmeans(x = X, centers = k, nstart = 20)
B <- km$betweenss
W = km$tot.withinss
ch <- (B/ (k -1)) /(W / (N - k))
ch_idx[k-1] <- ch
}
clus_q1 <- tibble(K = K, ch_idx = ch_idx)
clus_q1 %>% ggplot(aes(K, ch_idx)) + geom_point() + geom_line() +
ylab("Calinksi-Harabasz Index") + theme_bw()
```

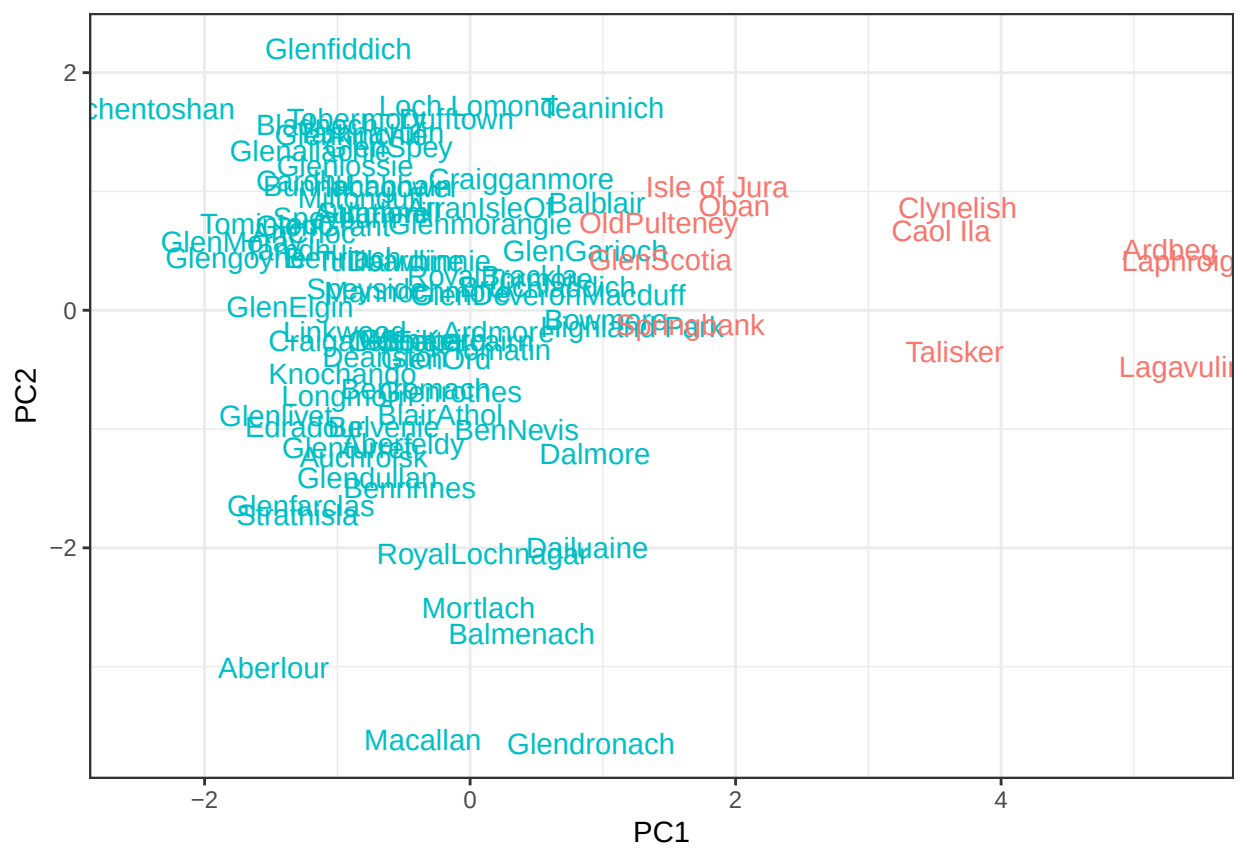


2. (.5 points) Dr. Wishart used 10 categories to describe these whiskys. Does this seem like a justifiable choice given the CH index? Why or why not?

The max CH index occurs at K = 2 with a score of 21.29, dropping drastically after this K value. Therefore, Dr. Wishart's K value has only a CH index of 11.13; this is nearly less than half of the optimum CH index. Consequently, it does not seem like a justifiable choice.

3. (1 point) Make a plot of PC1 vs PC2 as in problem 1.5 above, but colour the distillery names by cluster assignment. (You should be able to copy the code from above and make some minor changes.) Use the number of clusters that maximizes the CH Index. Describe any patterns you see.

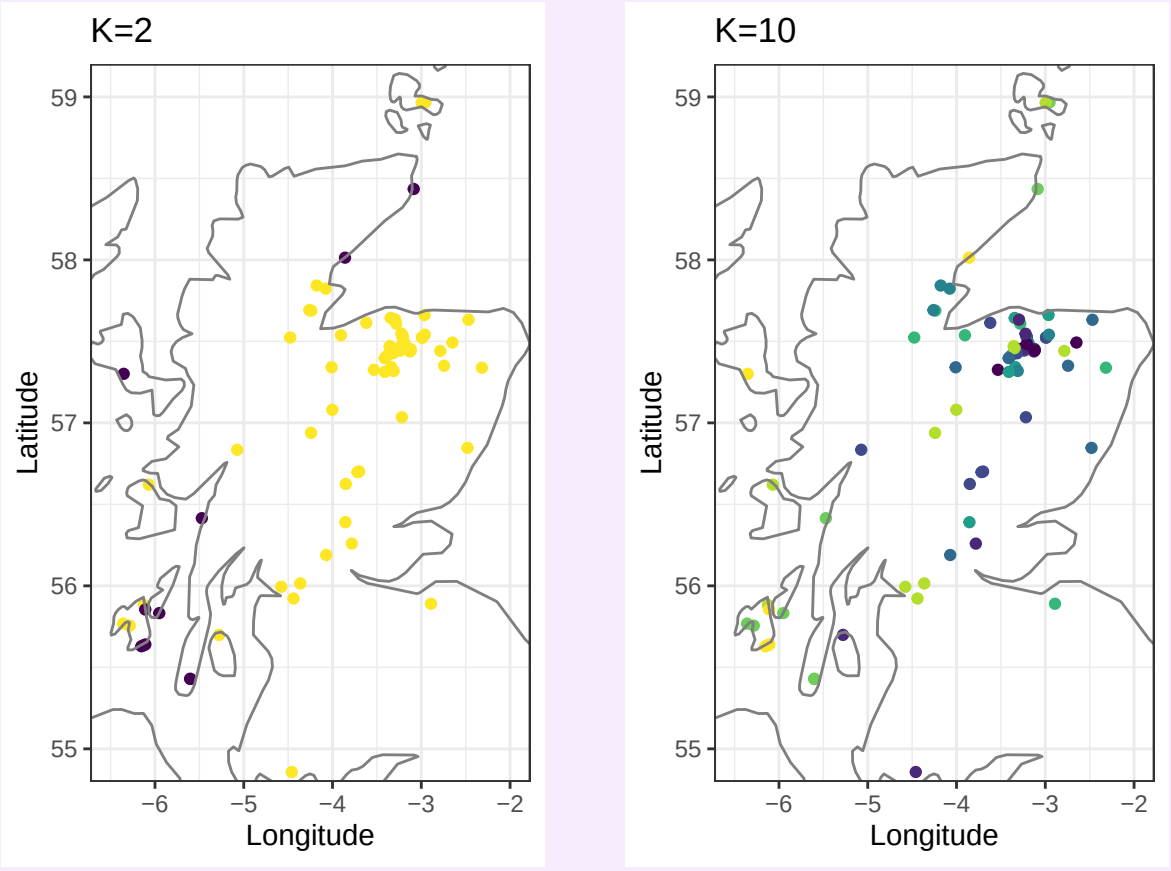
```
# Some code
opt_km <- kmeans(X, centers = 2, nstart = 20)$cluster
cbind(two_pcs, cluster = opt_km) %>% ggplot(aes(PC1, PC2, label = Distillery)) +
geom_text(aes(color = factor(cluster))) + theme_bw() +
theme(legend.position = "none")
```



4. (0.5 points) Use the function below to create a map of Scotland with the distilleries plotted as points coloured by their cluster assignment (the argument `cl` should be a vector of cluster assignments). Do this for the best K (as determined by CH index) and for K = 10. Describe the patterns you see.

```
scotland_cluster_plot <- function(cl) {
  stopifnot(is.vector(cl), length(cl) == N)
  ggplot(whisky %>% bind_cols(cluster = cl), aes(Longitude, Latitude)) +
    geom_point(aes(color = factor(cluster))) +
    coord_quickmap(ylim = c(55,59), xlim=c(-6.5,-2)) +
    borders(regions = "UK") +
    scale_colour_viridis_d() +
    theme_bw() + theme(legend.position = "none")
}
```

```
# Some code
km_10 <- kmeans(X, centers = 10, nstart = 20)$cluster
s1 <- scotland_cluster_plot(opt_km) +
ggtitle("K=2")
s2 <- scotland_cluster_plot(km_10) + ggtitle("K=10")
cowplot::plot_grid(s1,s2, nrow = 1)
```



5. (1 point) Ron Swanson (of *Parks and Recreation*, see the image below) famously loves Lagavulin (as do I). In light of your analyses in this assignment, are there other whiskys you would recommend? What flavours does Ron enjoy? Produce any figures that can be used to back up your claims.

```
knitr::include_graphics("ron.pdf")
```

The flavours highlighted in Lagavulin and enjoyed by Ron are: Body, Smoky, Medical.

```
# Code as needed.
wis <- select(whisky, -c("Postcode","Longitude","Latitude","Distillery"))
temp <- kmeans(wis, centers = 10, nstart = 20)
assignments <- double(nrow(whisky))
used_cl <- temp$cluster
recommendation = c()
for (i in 1:length(used_cl)) {
if(used_cl[i] == used_cl[58]) {
recommendation = append(recommendation,whisky$Distillery[i])
assignments[i] = 1
}}
recommendation
```

```
## [1] "Ardbeg"    "Caol Ila"  "Clynelish" "Lagavulin" "Laphroig"  "Talisker"
```

Based on the Dr.Wishart's number of categories for whiskys, the following others are recommended:"Ardbeg","Caol Ila","Clynelish","Lagavulin","Laphroig","Talisker".