Lab 2 – SIRCSS – Machine Learning for Social Science (Solutions)

Although not obligatory in any way, I recommend—for good practices and learning purposes—to create a document that includes code, relevant output, as well as brief answers to the questions. I recommend the use of RMarkdown for this purpose. But Word or any software like it is of course fine too.

Given the rather short amount of time available, you may likely not finish all tasks, and that is totally understandable and fine. I advise doing the tasks in order. Some tasks are marked as bonus tasks. At the end of the session, I will upload the solutions. Lastly, refer to the R helpfile for code that might be useful in solving the tasks; it includes code for solving similar types of problems using the R functions referenced in this document.

Part 1: Salary prediction

In the first part of this lab, we will work with a dataset (`adults`) containing information about a sample of US individuals' salaries as well as a handful sociodemographic variables. We will also consider an augmented version of this dataset (`adults_aug`), combining the original data with (simulated) highly nuanced lifecourse information. With this, the objective is to explore how accurately we can predict the salaries of individuals, comparing neural networks with other methods from previous weeks.

1. Begin by importing `adults.rds` and partition the data into a *training* and *test* set. We will use the training set to fit our model, and the test set to assess accuracy and compare across models. The dataset contains cirka 50,000 individuals, and for each we have information about five traditional variables (age, education, hours worked per week, capital gain and capital loss). We are reasonably sure there are no complex interactions or non-linearities. With this as the background, do you believe a neural network model is likely to excel on this dataset? Why/why not?

Based on the provided information — that the data contain few variables, and that domain knowledge suggests their relationship with the outcome is not overly complex — suggest that we may get rather small gains from using (deep) neural networks compared to a standard linear model. The fact that we do have a substantial number of observations relative to the number of variables, though, should allow the model to robustly identify any smaller non-linearities that may exist, which may provide a small improvement.

```
# Import
adults <- readRDS(file = 'data/final/adults.rds')
# Train / test split (80/20)
adults_X <- adults[,c(1:5)]
adults_y <- adults[,6]
n <- nrow(adults_X)
set.seed(1234)
idx <- sample(seq_len(n), size = floor(0.8 * n))
X_train <- adults_X[idx, , drop = FALSE]
y_train <- adults_y[idx]
X_test  <- adults_X[-idx, , drop = FALSE]
```

```
y_test  <- adults_y[-idx]
input_dim <- ncol(X_train)
```

2. Begin with estimating a standard logistic regression model to the training set you just created. Hint: you may use the glm(..., family='binomial') function to estimate a standard logistic regression model. When estimation has finished, use the predict() function to predict the outcome on the test dataset, and calculate (report) the accuracy.

```
# Standard logit model
logit <- glm(formula = y ~ .,
             data = data.table(X_train,y=y_train),
             family = 'binomial')
logit_preds <- predict(object = logit,
                       newdata = as.data.table(X_test))
logit_preds <- ifelse(logit_preds>0.5, 1, 0)
logit_accuracy <- mean(logit_preds==y_test)
print(logit_accuracy)
```

```
## [1] 0.8043812
```

3. Estimate a neural network with 1 *hidden layer* and 5 *hidden units*. In the `compile()` function, set the optimizer to `optimizer_adam()`, the loss function equal to `"binary_crossentropy"` (as the outcome is binary), and metrics to `"accuracy"`. Report the accuracy. Did the result match your expectations formulated in #1?

Adding a hidden layer only provides a marginal improvement. This is in line with expectations.

```
# NN with 1 hidden layer
k_clear_session()
set.seed(1234)
nn1 <- keras_model_sequential() |>
  layer_dense(units = 5, activation = "relu", input_shape = input_dim) |>
  layer_dense(units = 1, activation = "sigmoid")

nn1 |>
    compile(
      optimizer = optimizer_adam(),
      loss      = "binary_crossentropy",
      metrics   = "accuracy")

history_nn1 <- nn1 |>
  fit(
    x = X_train, y = y_train,
    epochs = 50,
    batch_size = 128,
    validation_split = 0.2,
    verbose = 0
  )

# Evaluate on holdout
nn1_preds <- as.numeric(nn1 |> predict(X_test))
```

```
## 306/306 - 0s - 200us/step
```

```
nn1_preds <- ifelse(nn1_preds>0.5, 1, 0)
nn1_accuracy <- mean(nn1_preds==y_test)
print(nn1_accuracy)
```

```
## [1] 0.8179957
```

4. Next, you shall estimate a considerably more complex neural network, containing 4 *hidden layers*. The first hidden layer should have 256 *hidden units*, the second hidden layer 128 hidden units, the third hidden layer 64 hidden units, and the fourth hidden layer 32 hidden units. Use the same settings for `compile()` as in #3. Report the *total number of parameters* and then estimate the model. Do you find that it outperforms #3 meaningfully? Why do you think this is (or is not) the case?

```
# NN with 4 hidden layer
k_clear_session()
set.seed(1234)
nn2 <- keras_model_sequential() |>
  layer_dense(units = 256, activation = "relu", input_shape = input_dim) |>
  layer_dense(units = 128, activation = "relu") |>
  layer_dense(units = 64, activation = "relu") |>
  layer_dense(units = 32, activation = "relu") |>
  layer_dense(units = 1, activation = "sigmoid")

print(nn2)
```

```
## Model: "sequential"
##
##  Layer (type)                        Output Shape                    Param #
##
##  dense (Dense)                       (None, 256)                       1,536
##
##  dense_1 (Dense)                     (None, 128)                      32,896
##
##  dense_2 (Dense)                     (None, 64)                        8,256
##
##  dense_3 (Dense)                     (None, 32)                        2,080
##
##  dense_4 (Dense)                     (None, 1)                            33
##
##  Total params: 44,801 (175.00 KB)
##  Trainable params: 44,801 (175.00 KB)
##  Non-trainable params: 0 (0.00 B)
```

Specified this way, the neural network contains 44,801 parameters to be estimated.

```
nn2 |>
    compile(
      optimizer = optimizer_adam(),
      loss      = "binary_crossentropy",
      metrics   = "accuracy")
```

```
history_nn2 <- nn2 |>
  fit(
    x = X_train, y = y_train,
    epochs = 50,
    batch_size = 128,
    validation_split = 0.2,
    verbose = 0
  )

# Evaluate on holdout
nn2_preds <- as.numeric(nn2 |> predict(X_test))
```

```
## 306/306 - 0s - 288us/step
```

```
nn2_preds <- ifelse(nn2_preds>0.5, 1, 0)
nn2_accuracy <- mean(nn2_preds==y_test)
print(nn2_accuracy)
```

```
## [1] 0.8254683
```

We again see a marginal, but non-neglibile improvement in performance. A neural network with a single hidden layer and 5 hidden units remains a relatively simple model (36 parameters). Expanding it to a 44,801 parameter model is a substantial increase in flexibility, enabling the model to pick up on complex non-linearities and interaction effects. However, because the data does not contain overly complex patterns, the gain is marginal.

5. Suppose now that we retrieve a dataset (`adults_aug`) that expands upon the original 'adults dataset. This dataset contains 6 extra variables; which capture complex aspects of the individuals life courses, with various interdependencies between them and possible non-linearities. Could this expanded data make a difference in terms of more clearly outperforming the standard logit? Investigate this by repeating steps 2-4 on the adults_aug dataset. Does your conclusion about the relevance of more complex network structure differ between the two datasets? Why?

```
# Import
adults_aug <- readRDS(file = 'data/final/adults_aug.rds')
# Train / test split (same idx)
adults_aug_X <- adults_aug[,c(1:11)]
adults_aug_y <- adults_aug[,12]
X_train <- adults_aug_X[idx, , drop = FALSE]
y_train <- adults_aug_y[idx]
X_test  <- adults_aug_X[-idx, , drop = FALSE]
y_test  <- adults_aug_y[-idx]
input_dim <- ncol(X_train)
```

```
# 2: Standard logit model
logit <- glm(formula = y ~ .,
             data = data.table(X_train,y=y_train),
             family = 'binomial')
logit_preds <- predict(object = logit,
                       newdata = as.data.table(X_test))
logit_preds <- ifelse(logit_preds>0.5, 1, 0)
logit_accuracy <- mean(logit_preds==y_test)
print(logit_accuracy)
```

```
## [1] 0.8046883
```

```
# 3: NN with 1 hidden layer
k_clear_session()
set.seed(1234)
nn1 <- keras_model_sequential() |>
  layer_dense(units = 5, activation = "relu", input_shape = input_dim) |>
  layer_dense(units = 1, activation = "sigmoid")

nn1 |>
    compile(
      optimizer = optimizer_adam(),
      loss      = "binary_crossentropy",
      metrics   = "accuracy")

history_nn1 <- nn1 |>
  fit(
    x = X_train, y = y_train,
    epochs = 50,
    batch_size = 128,
    validation_split = 0.2,
    verbose = 0
  )

# Evaluate on holdout
nn1_preds <- as.numeric(nn1 |> predict(X_test))
```

```
## 306/306 - 0s - 200us/step
```

```
nn1_preds <- ifelse(nn1_preds>0.5, 1, 0)
nn1_accuracy <- mean(nn1_preds==y_test)
print(nn1_accuracy)
```

```
## [1] 0.9624322
```

```
# 4: NN with 4 hidden layer
k_clear_session()
set.seed(1234)
nn2 <- keras_model_sequential() |>
  layer_dense(units = 256, activation = "relu", input_shape = input_dim) |>
  layer_dense(units = 128, activation = "relu") |>
  layer_dense(units = 64, activation = "relu") |>
  layer_dense(units = 32, activation = "relu") |>
  layer_dense(units = 1, activation = "sigmoid")

nn2 |>
    compile(
      optimizer = optimizer_adam(),
      loss      = "binary_crossentropy",
      metrics   = "accuracy")

history_nn2 <- nn2 |>
```

```
  fit(
    x = X_train, y = y_train,
    epochs = 50,
    batch_size = 128,
    validation_split = 0.2,
    verbose = 0
  )

# Evaluate on holdout
nn2_preds <- as.numeric(nn2 |> predict(X_test))
```

```
## 306/306 - 0s - 282us/step
```

```
nn2_preds <- ifelse(nn2_preds>0.5, 1, 0)
nn2_accuracy <- mean(nn2_preds==y_test)
print(nn2_accuracy)
```

```
## [1] 0.9969291
```

Here we see substantial improvements — at each step. The addition of a single hidden layer provides a significant boost; and adding even more complexity (36 → 45K parameters) further improves the results. With the most complex model, we are almost perfectly predicting the individuals' salaries (whether they are above/below the median). If this was a real dataset, this could for example indicate that there are intricate path dependencies in the life-course, between life-events and life-outcomes, that are not well captured by a linear model, which seems plausible. So, addressing the question: the utility of a more complex model is indeed different here compared to before. The reason is that the augmented dataset contains within it complex interactions and non-linearities.

Part 2: Image prediction

In the second part of the lab, we'll work with *Fashion MNIST data* (https://www.kaggle.com/datasets/zalando-research/fashionmnist) containing images of 10 different types of fashion items — including t-shirts, coats, bags, and sneakers. All images are 28 pixels in width and 28 pixels in height (784 in total). Each pixel has a single value associated with it; its gray-scale value ranging between 0 and 255.

For the purpose of this lab, we are imagining a fictive scenario in which each image correspond to a consumed good, and that we have (simulated) *basic socioeconomic information* about the *customer* who bought the fahion item, as well as the *year* it was consumed. Our social scientific interest is to examine whether there are any trends in the association between types of fashion goods and the socioeconomic status of the customer. For example, are sneakers more or less associated with low/high economic status, and has this changed over time?

The practical problem that we face, is that we only have labeled data of the images (i.e., identifying which fashion item is in the image) for part of the time-period (in 2016 but not in 2017). Our objective, thus, is to use *deep learning* to estimate a model on the part of the data where we do have labeled image data, and then use the trained model to predict on the yet to be labeled images.

1. Begin by importing the datafiles *"fashion_2016_train.rds"* and *"fashion_2016_test.rds"*.

```r
# Import
train_2016 <- readRDS(paste0(mywd,"data/fashion_2016_train.rds"))
test_2016 <- readRDS(paste0(mywd,"data/fashion_2016_test.rds"))
x_train <- train_2016$images
y_train <- train_2016$labels
x_test <- test_2016$images
y_test <- test_2016$labels

# Normalize pixel values
x_train <- x_train / 255
x_test <- x_test / 255

# Add channel dimension (Fashion MNIST is grayscale)
x_train <- array_reshape(x_train, c(dim(x_train), 1))
x_test <- array_reshape(x_test, c(dim(x_test), 1))

# Dims
n_train <- dim(x_train)[1]
n_test <- dim(x_test)[1]
```

2. Next, estimate a simple convolutional neural network with K=1 convolutional layers and M=1 regular type of (fully connected) hidden layers. Specify the *number of filters* (in the convolutional layer) and the *number of hidden units* (in the fully connected / regular hidden slides) to be 7. Remember also that included after each added convolutional layer, *pooling* (max-pooling, $2 * 2$) should to be applied. Finally, include `layer_dense(units = 10, activation = "softmax")` at the end. When *compiling* the model using the compile() function, set loss = "sparse_categorical_crossentroy". Report the results, and reflect on whether you think improvement can be made by increasing either M, K, or the number of filters or hidden units in the regular hidden layer.

```r
# Simple CNN (M=1; K=1); 7 hidden units in dense layer, 7 filters.
k_clear_session()
set.seed(1234)
cnn1 <- keras_model_sequential() %>%
  # First convolutional block
  layer_conv_2d(filters = 7,
                kernel_size = c(3, 3),
                activation = "relu",
                input_shape = c(28, 28, 1)) %>%
  layer_max_pooling_2d(pool_size = c(2, 2)) %>%
  # Dense layers
  layer_flatten() %>%
  layer_dense(units = 7, activation = "relu") %>%
  layer_dense(units = 10, activation = "softmax")  # 10 fashion categories

# Compile model
cnn1 %>% compile(
  optimizer = optimizer_adam(),
  loss = "sparse_categorical_crossentropy",
  metrics = c("accuracy")
)

# Fit
history_cnn1 <- cnn1 |>
```

```
  fit(x_train,
      y_train,
      validation_split=0.2,
      epochs=50,
      batch_size=128,
      verbose=0)

# Accuracy
cnn1_test_accuracy <- cnn1 %>% evaluate(x_test, y_test, verbose = 0)
print(cnn1_test_accuracy)
```

```
## $accuracy
## [1] 0.8798701
##
## $loss
## [1] 0.3484554
```

Increasing the number of convolutional layers (M) and the number of filters enable the learning of more intricate patters in the images. Increasing the number of regular, fully-connected layers (K) enable learning more complex associations between the detected features of the images and the outcome classes. Our image data is relatively simple (it is easy to think of much more complex images, with more details in them), such that the potential for gains by increasing M or the number of filters is unclear; but worth evaluating. A similar case can be made for the number of regular, fully-connected hidden layers: do we suspect that there are complex associations between the learned image patterns and the object classes? Not so easy to say without knowing the image-features beforehand, of course. But, the complexity of the images and how *easily distinguishable* we think the different outcome classes are can inform this guess.

3. Next, you shall estimate a slightly more complex convolutional neural network, increasing the number of filters to 32 in the first layer, and 64 in the second convolutional layer. Similarly, increase the number of hidden units in the first regular/fully-connected hidden layer to 64, and the second to 32. Report the results. Does this slightly more complicated model improve/degrade upon the simple one in #2? Speculate why in terms of the bias/variance trade-off. Report and contrast also the number of parameters between the two.

```
# More sophisticated CNN (M=2; K=2)
k_clear_session()
set.seed(1234)
cnn2 <- keras_model_sequential() %>%
  # Convolutional layers
  layer_conv_2d(filters = 32,
                kernel_size = c(3, 3),
                activation = "relu",
                input_shape = c(28, 28, 1)) %>%
  layer_max_pooling_2d(pool_size = c(2, 2)) %>%
  layer_conv_2d(filters = 64,
                kernel_size = c(3, 3),
                activation = "relu") %>%
  layer_max_pooling_2d(pool_size = c(2, 2)) %>%
  # Dense layers
  layer_flatten() %>%
  layer_dense(units = 64, activation = "relu") %>%
  layer_dense(units = 32, activation = "relu") %>%
```

```
  layer_dense(units = 10, activation = "softmax")  # 10 fashion categories

# Compile model
cnn2 %>% compile(
  optimizer = optimizer_adam(),
  loss = "sparse_categorical_crossentropy",
  metrics = c("accuracy")
)

# Fit
history_cnn2 <- cnn2 |>
  fit(x_train,
      y_train,
      validation_split=0.2,
      epochs=50,
      batch_size=128,
      verbose=0)

# Accuracy
cnn2_test_accuracy <- cnn2 %>% evaluate(x_test, y_test, verbose = 0)
print(cnn2_test_accuracy)
```

```
## $accuracy
## [1] 0.8977273
##
## $loss
## [1] 0.5997512
```

We see a slight improvement. To the extent that this is a robust difference, it should reflect a decrease in bias — since this is a more complex model compared to the previous one. If we wanted to understand the source of the improvement in more detail, we could estimate additional models where we only add more layers and units for the convolutional layer, keeping the dense hidden layer simple, and vice versa. Note: had this been a real research setting, I would have performed a systematic grid search of the many parameter combinations.

4. (BONUS) In #2-3, we used max-pooling of 2*2. Why is it generally not a good idea to increase the dimension of the pooling to become large relative to the images? Would this increase bias or variance?

When we apply (max) pooling, we keep one (the largest) cell-value within the pooling window. If we use a large pooling window, we risk blurring out the distinctive patterns detected by the filter. In the extreme, if we have a pool window as large as the image itself then each feature map would just contain a single value, completely washing out any pattern detected. In terms of bias/variance, pooling across increasingly larger surfaces smoothens patterns out, and should thus increase bias.

5. Because we are interested in using the predictions from these models for downstream analysis (of trends in consumer polarization), it is extra important to validate the measure. This is what you shall do now, focusing on the CNN model you deemed the best thus far. In particular you should break down the accuracy per item category. Is there meaningful difference in predictability between item classes? Does the ordering make substantive sense? Is it relevant to consider these errors when using predictions from the model for downstream tasks? Hint: to compute accuracy by group, you may use the following code:

```
# Generate predicted probabilities for each class
predictions <- cnn2 %>% predict(x_test)
```

```
## 289/289 - 0s - 1ms/step
```

```
# Convert probabilities to class labels (0-9) by selecting highest probability
predicted_classes <- apply(predictions, 1, which.max) - 1
# Create data.table with actual and predicted class labels
accuracy_by_class_dt <- data.table(actual = y_test, predicted = predicted_classes)
accuracy_by_class_dt[, correct := fifelse(actual == predicted, yes = 1, no = 0)]
# Define human-readable labels for each fashion item category
fashion_labels <- c("T-shirt/top", "Trouser", "Pullover", "Dress", "Coat",
                    "Sandal", "Shirt", "Sneaker", "Bag", "Ankle boot")
fashion_labels_dt <- data.table(label = fashion_labels, number = 0:9)
# Merge to add item labels to the accuracy data
accuracy_by_class_dt <- merge(x = accuracy_by_class_dt, y = fashion_labels_dt,
                              by.x = 'actual', by.y = 'number')
# Calculate and display accuracy for each item category, sorted highest to lowest
accuracy_by_class_dt[, .(accuracy = mean(correct)), by = label][order(accuracy, decreasing = T)]
```

```
##           label  accuracy
##          <char>     <num>
##  1:     Trouser 0.9800664
##  2:         Bag 0.9727749
##  3:      Sandal 0.9682366
##  4:     Sneaker 0.9680511
##  5:  Ankle boot 0.9542982
##  6:       Dress 0.9087003
##  7:    Pullover 0.8779395
##  8: T-shirt/top 0.8561644
##  9:       Shirt 0.7497349
## 10:        Coat 0.7479339
```

The classes the model struggles with are similar clothing items with relatively few distinctive features. By contrast, sandals, trousers, and bags have clearer cues and are easier to predict.

```
# Create confusion matrix (rows = actual, columns = predicted)
conf_matrix <- table(Actual = y_test, Predicted = predicted_classes)
# Convert to proportions (row-normalized: each row sums to 1)
conf_matrix_prop <- prop.table(conf_matrix, margin = 1)
# Melt to long format for analysis/plotting
conf_long <- as.data.table(conf_matrix_prop)
setnames(conf_long, c("actual", "predicted", "proportion"))
conf_long[, actual := as.integer(as.character(actual))]
conf_long[, predicted := as.integer(as.character(predicted))]
# Add item labels
conf_long <- merge(conf_long, fashion_labels_dt, by.x = "actual", by.y = "number")
setnames(conf_long, "label", "actual_label")
conf_long <- merge(conf_long, fashion_labels_dt, by.x = "predicted", by.y = "number")
setnames(conf_long, "label", "predicted_label")
# Order according to most common confusions (excluding correct predictions)
conf_long <- conf_long[actual != predicted][order(-proportion)]
```

```
# Subset on actual_label equal to the most error-prone label
conf_long[actual_label=="shirt"]
```

```
## Empty data.table (0 rows and 5 cols): predicted,actual,proportion,actual_label,predicted_label
```

6. Next, we shall finally address the question we set out to answer: how patterns in consumer behavior changed between 2016 to 2017. To answer this question, please follow these steps:

- Import the `fashion2016_2017_unlabeled.rds` file, which contains all images and the sociodemographic info of the individual associated with the image. Note: the data has already been preprocced (normalized pixel values, etc.).
- Use the predict() function to predict for this dataset based on the CNN model you thought were the best. Note that, when you use the predict() function for a model where you have multiple-category outcome variables, like here, each observation get a probability over the items. To assign the prediction to the maximum value, you may use this code:

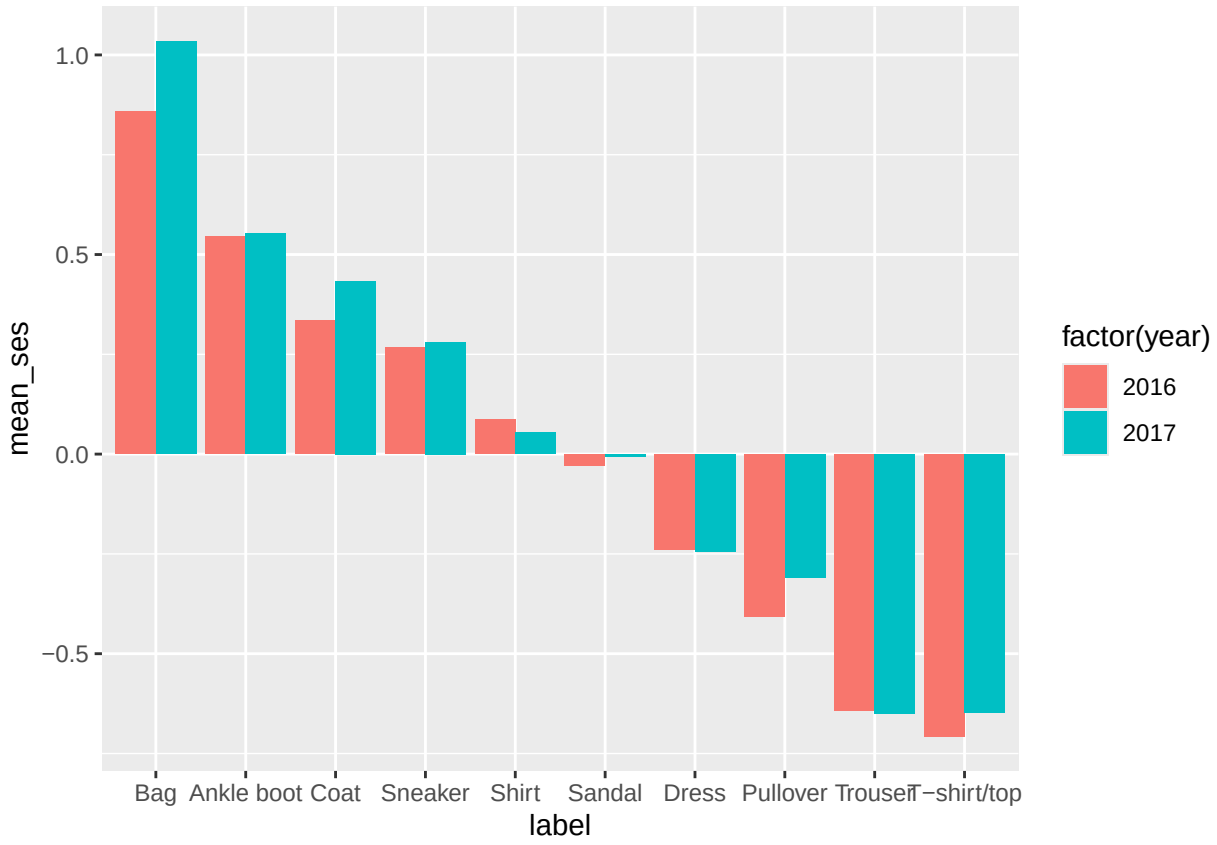
```
apply(preds_2016_2017, 1, which.max) - 1
```

- Then you can simply merge the predicted category with the demographic variable data.frame, and calculate various associations by classical statistical means.
- Report your findings.

```
# i) import
unlabeled_2016_2017 <- readRDS(paste0(mywd,"data/final/fashion_2015_2016_unlabeled.rds"))
x_unlabeled_2016_2017 <- unlabeled_2016_2017$images

# ii) predict
preds_2016_2017 <- cnn2 %>% predict(x_unlabeled_2016_2017)
```

```
## 2188/2188 - 2s - 909us/step
```

```
pred_class_2016_2017 <- apply(preds_2016_2017, 1, which.max) - 1

# iii) merge ses and predicted item
consumption_ses_dt_1617 <- setDT(cbind(unlabeled_2016_2017$demographics,
                                pred_y=pred_class_2016_2017))
consumption_ses_dt_1617 <- merge(x=consumption_ses_dt_1617,
                                y=fashion_labels_dt,
                                by.x='pred_y',by.y='number')

# iv) calculate some basic associations
consumption_ses_dt_1617_2 <- consumption_ses_dt_1617[,.(mean_ses=mean(ses_score)),
                                                      by=.(label,year)]
# Basic plot
lab_order <- consumption_ses_dt_1617_2[
  , .(mx = max(mean_ses)), by = label
][order(-mx), label]
consumption_ses_dt_1617_2[, label := factor(label, levels = lab_order)]
ggplot(consumption_ses_dt_1617_2,aes(x=label,y=mean_ses,fill=factor(year))) +
  geom_bar(stat = 'identity', position='dodge')
```

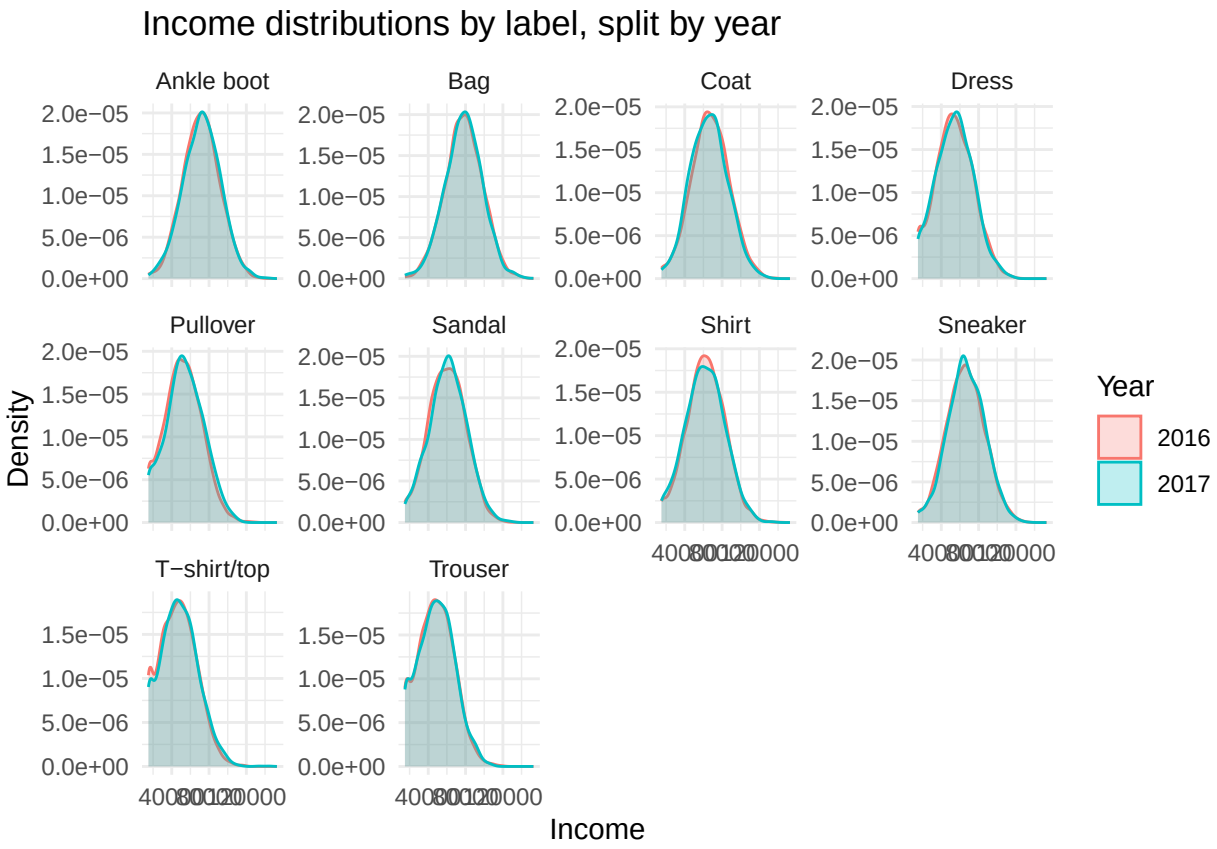



```
# Gini
gini_zero <- function(x, na.rm = TRUE) {
  if (na.rm) x <- x[!is.na(x)]
  x <- as.numeric(x); n <- length(x)
  if (n < 2) return(0)
  m1 <- mean(abs(x))
  if (m1 == 0) return(0)
  sx <- sort(x)
  S  <- sum((2*seq_len(n) - n - 1) * sx)         # sum_{i<j} (x_j - x_i)
  gmd <- 2 * S / (n * (n - 1))                   # average |xi - xj|
  gmd / (2 * m1)
}
consumption_ses_dt_1617_2[,gini_zero(mean_ses),by=year]
```

```
##     year        V1
##    <num>     <num>
## 1:  2016 0.7488974
## 2:  2017 0.7627036
```

```
# Other
DT <- copy(consumption_ses_dt_1617)
setDT(DT)
DT[, `:=`(
  label  = factor(label),
  urban  = factor(urban, levels = c(0,1), labels = c("rural","urban")),
```

```
  region = factor(region),
  year   = factor(year)
)]

# Income density: compare years within each label
ggplot(DT, aes(x = income, fill = year, color = year)) +
  geom_density(alpha = 0.25, adjust = 1.1) +
  facet_wrap(~ label, scales = "free_y") +
  labs(title = "Income distributions by label, split by year",
       x = "Income", y = "Density", fill = "Year", color = "Year") +
  theme_minimal()
```



Part 3: Taste clustering and influence

In you are unfamiliar with unsupervised learning and clustering, and would prefer working on that, this is also totally fine: Whatever you think is most useful to you :) Here, we will consider a (simulated) data set which contains information about a sample of (fictive) individuals' music tastes as well as a measure of their influence on others.

1. Begin by importing the file "`taste_influence.csv`". Report the number of rows and columns of the data set, and the genres contained in it. Create a scatter-plot of two combinations of genres of your choice. Based on this, do you get any indication that the data is clustered along musical tastes?

```
dt <- fread('/Users/marar08/Documents/Teaching/MLSS_HT2025/Labs/W4/taste_influence.csv')
dim(dt)
```

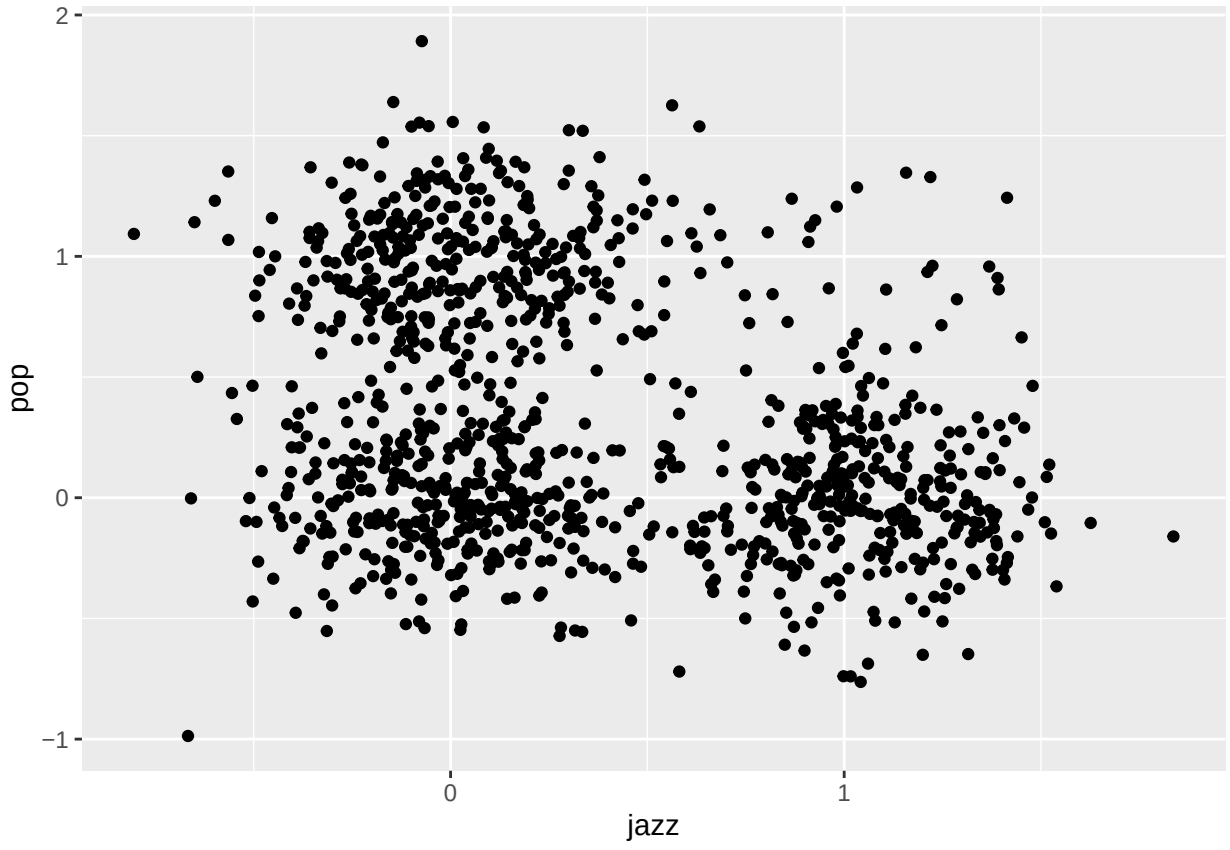
```
## [1] 1075    4
```

```
head(dt)
```

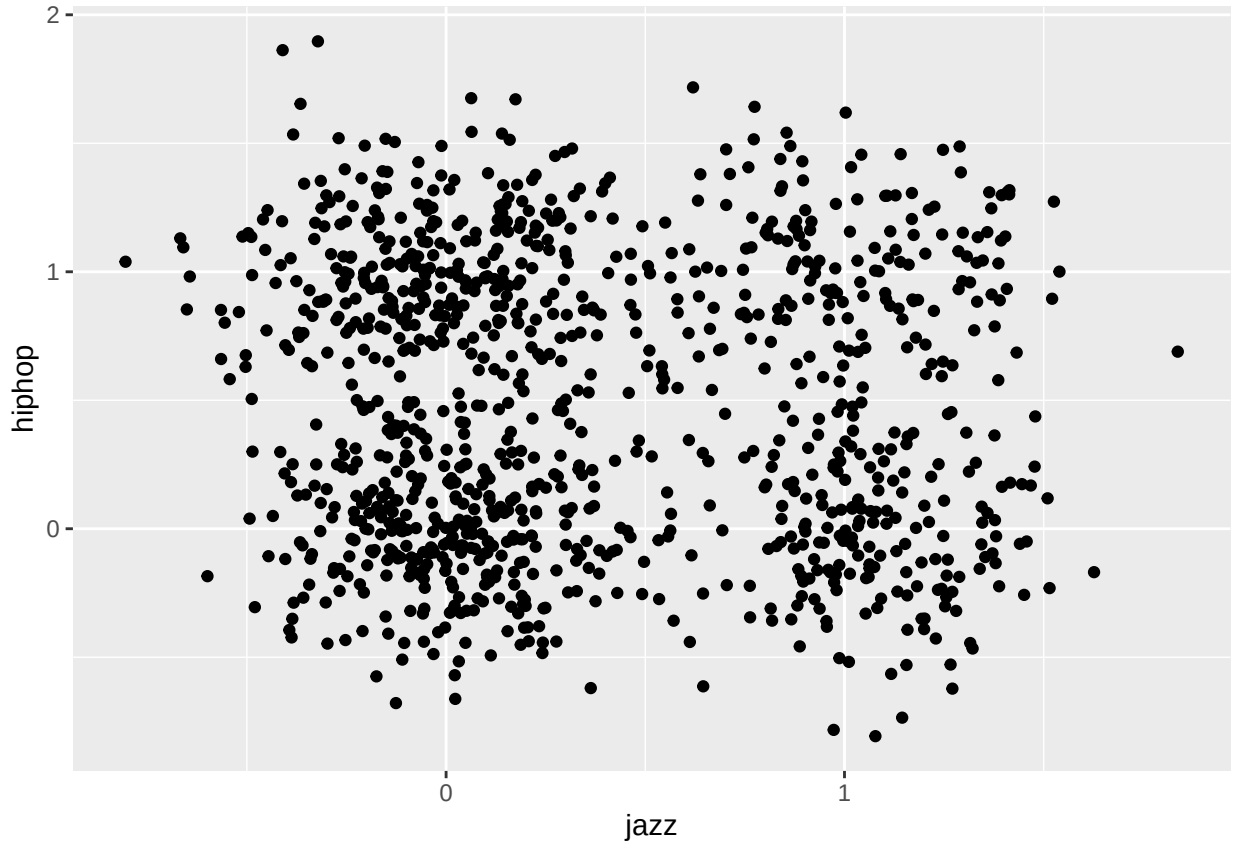
```
##           jazz         pop      hiphop   influence
##          <num>       <num>       <num>       <num>
## 1: -0.03743187 -0.02836584  1.17409185  1.67634621
## 2:  0.55748147  0.15872313 -0.03007249  0.07487850
## 3: -0.08302068 -0.02067557 -0.11942695 -0.09607821
## 4:  0.11750772  1.39635593 -0.06703518  0.36594376
## 5:  0.24060609  1.01726019  0.66100663  1.81186663
## 6:  0.93673898  0.04925761 -0.05323063  1.77677839
```

The data sets contains 1075 rows and 4 columns. Three of the variables/columns describe the users taste (genres: jazz, pop, and hiphop), and one their influence on others.

```
ggplot(dt, aes(x=jazz,y=pop)) + geom_point()
```




```
ggplot(dt, aes(x=jazz,y=hiphop)) + geom_point()
```



Based on these bivariate scatterplots, I would say there are indeed some indications of clustering along the taste dimensions. Although the separation is *not super-clear cut*, there are clear *differences in density* of data points in different regions. Taking jazz–hiphop as an example, I would say there looks like there are four dense regions of data points with centers at [0,0], [0,1], [1,0], and [1,1], and where there is a decline in desity at their respective borders.

2. Now you shall do some clustering. To prepare the data, do the following: (*i*) store/copy the data to a new `R` object, and subset it so that it only contains the three "taste columns"— these are the columns you will cluster based upon, (*ii*) standardize this data table (hint: you can e.g., use `scale()` for this purpose), (*iii*) transform it into a matrix (hint: e.g., by using `as.matrix()`).

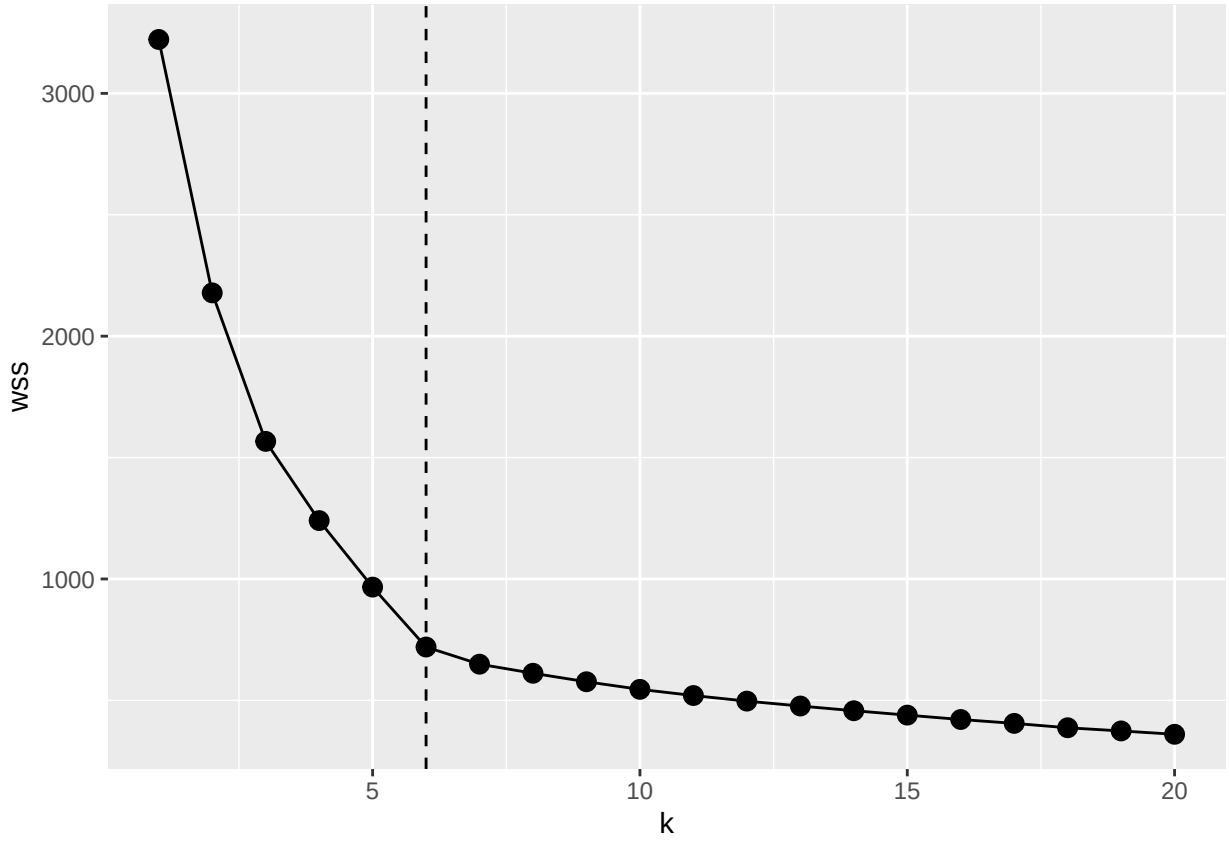
```
# (i)
taste_mat <- copy(dt)
taste_mat$influence <- NULL
# (ii)
taste_mat <- as.matrix(taste_mat)
# (iii)
taste_mat <- scale(taste_mat)
# Inspect
head(taste_mat)
```

```
##           jazz        pop     hiphop
## [1,] -0.7253625 -0.6712253  1.2209825
## [2,]  0.3576840 -0.3239461 -0.9018071
## [3,] -0.8083574 -0.6569504 -1.0593277
## [4,] -0.4432932  1.9733787 -0.9669676
```

```
## [5,] -0.2191912  1.2696917  0.3164781
## [6,]  1.0481267 -0.5271387 -0.9426319
```

3. Having formatted the data according to #2, you shall now use the *kmeans* algorithm to cluster your data. Recall that a requisite for running *kmeans* is that the parameter k has been specified. In practice—and as is the case here—we often do not know the appropriate number of clusters a priori. Therefore, you shall implement a loop that, at every iteration, runs kmeans with a different number of clusters, and extracts the *total within cluster sum of squares* (hint 1: which can be extracted using `$tot.withinss` | hint 2: set the argument `nstart=100` to ensure robustness of the local optima you find). Consider no. clusters ranging from 1 to 20, with an interval of 1. Plot k against `tot.withinss`. Which number of clusters do you find appropriate? Motivate.

```
ks <- 1:20
wss <- c()
for(i in 1:length(ks)){
  temp <- kmeans(x = taste_mat,
                 centers = ks[i],
                 nstart = 100,
                 iter.max = 1000)
  wss[i] <- sum(temp$tot.withinss)
}
ggplot(data.table(k=ks,wss=wss),aes(x=k,y=wss)) +
  geom_point(size=3) +
  geom_line() +
  geom_vline(xintercept = 6, linetype='dashed')
```



The relative gain in terms of additional *total within sum of squares* per extra cluster starts to decline sharply

after $k = 6$. Hence, seeking to find a good balance between variance accounted for and parsimony/complexity, $k = 6$ appears to be a good choice here.

4. For the specification (of k) that you decided on in #3, extract the *centroids* and interpret each cluster in terms of what distinguishes it from the rest. Do the clusters seem meaningfully distinct?

```
set.seed(1234)
finalk <- kmeans(x = taste_mat,
                 centers = 6,
                 nstart = 100,
                 iter.max = 1000)
finalk$centers
```

```
##         jazz        pop     hiphop
## 1 -0.6606565 -0.6264451 -0.8310363
## 2  1.1975400 -0.5153682  0.9374598
## 3 -0.6034402  1.2157422 -0.8781105
## 4 -0.5225678  1.2650179  0.8785221
## 5  1.2617093 -0.6363058 -0.8861036
## 6 -0.7075403 -0.6519389  0.9275712
```

The clusters does indeed seem to capture meaningfully distinct taste-profiles. I would label each cluster as follows (note that scale is a bit funny because we have standardized our data; e.g., +1 = one positive standard deviation from the mean, which is 0):

- Cluster 1: People who do not like either jazz, pop or hiphop
- Cluster 2: Jazz and hiphop fans
- Cluster 3: Pop fans
- Cluster 4: Pop and hiphop fans
- Cluster 5: Jazz fans
- Cluster 6: Hiphop fans.

5. To get a feeling for the role that the choice of k plays, estimate another *kmeans* model but this time with $k = 2$. Inspecting the centroids, how does your clustering change; how does it alter your understanding of the population?

```
set.seed(1234)
k2 <- kmeans(x = taste_mat,
             centers = 2,
             nstart = 100,
             iter.max = 1000)
k2$centers
```

```
##         jazz        pop      hiphop
## 1  1.1107205 -0.6203654 -0.08308994
## 2 -0.6634616  0.3705600  0.04963173
```

Partitioned this way, it would appear we have jazz fans—who are OK with hiphop but does not like pop, on the one hand (cluster 1). And then moderate pop fans, who are OK with hophop but does not like jazz. In other words, this gives a very different picture of the population. By being so coarse, it blends together different tastes, and we only see pooled averages.

6. Clustering provides a tool for discovering underlying structures in our data. Once these structures have been discovered, they can be studied in separate analyses. That is what you shall do now. We want to examine whether different "taste types" have differential degree of influence on others. To do so, (*i*) create a new column in your original data set storing the the retrieved cluster assignments (hint: you find the cluster assignments using `$cluster`). Then (*ii*) estimate a linear regression with the *influence score* (`infuence`) as the outcome variable, and the clusters (formatted as a factor) as predictors. Interpret the results: are there any difference in influence between the clusters?

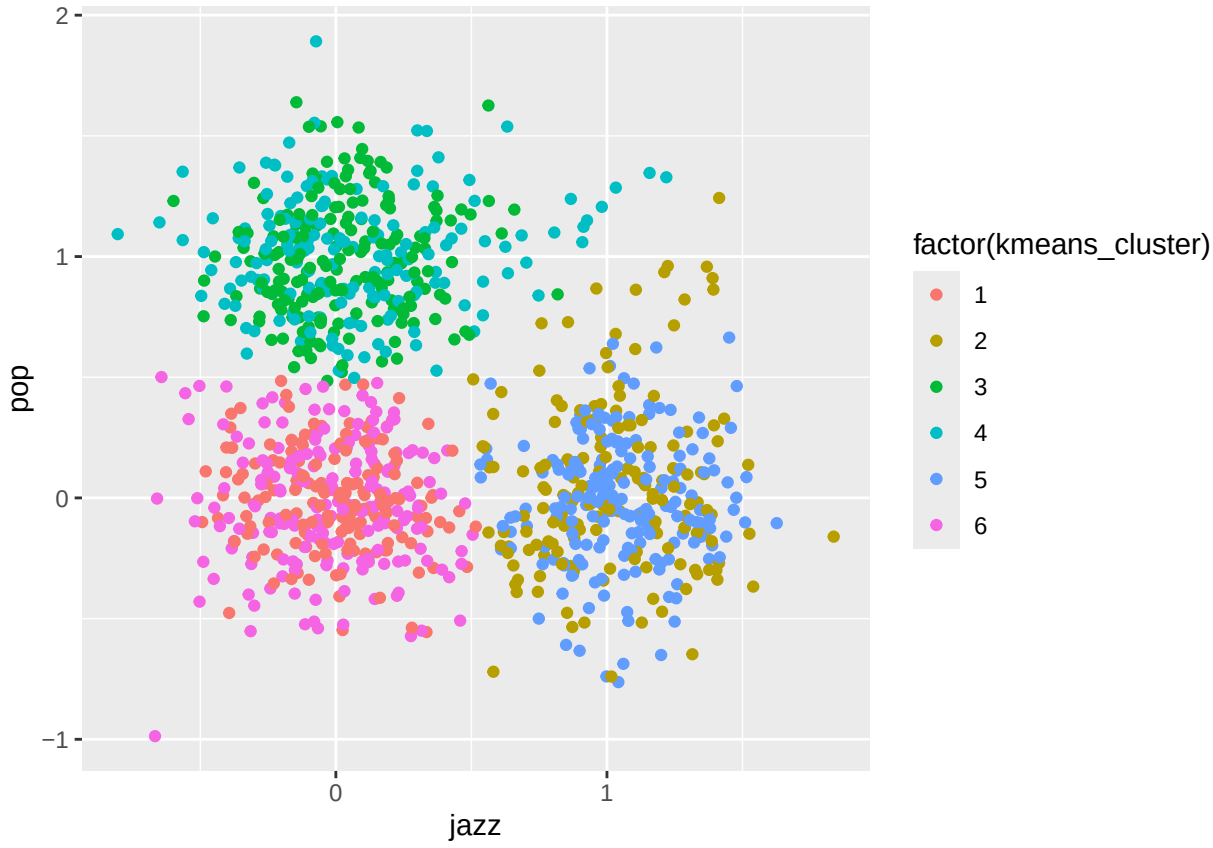
```
dt[,kmeans_cluster := factor(finalk$cluster)]
summary(lm(influence~kmeans_cluster,data=dt))
```

```
##
## Call:
## lm(formula = influence ~ kmeans_cluster, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5574 -0.5409  0.0660  0.6140  3.4105
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.19778    0.07516   2.631  0.00863 **
## kmeans_cluster2  2.42014    0.10662  22.698  < 2e-16 ***
## kmeans_cluster3  0.52539    0.10300   5.101 3.99e-07 ***
## kmeans_cluster4  1.79207    0.10598  16.910  < 2e-16 ***
## kmeans_cluster5  1.57779    0.10166  15.521  < 2e-16 ***
## kmeans_cluster6  1.60424    0.10237  15.671  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9655 on 1069 degrees of freedom
## Multiple R-squared:  0.4019, Adjusted R-squared:  0.3991
## F-statistic: 143.6 on 5 and 1069 DF,  p-value: < 2.2e-16
```

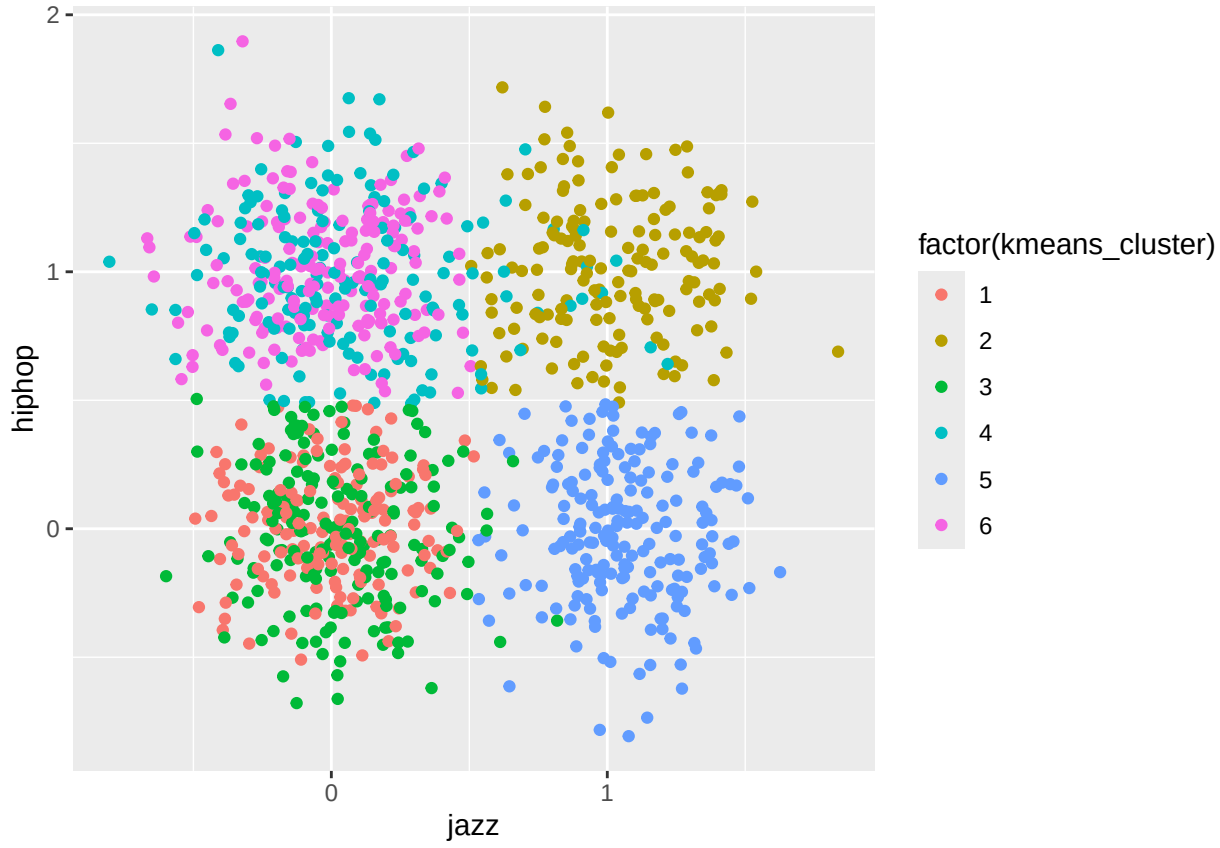
There does indeed seem to be differences influence across the 6 clusters. The most influential is cluster 2, which we labeled as those who like both jazz and hiphop. The least influential is the base category, cluster 1, which reflects individuals who neither likes jazz, hiphop or pop. One might also note that differences in taste explain as much as 40% of the variation in infliuence.

7. Now that you have merged the cluster assignments to the original data, produce the same plots as you did in #1, but now colored by the cluster assignments. Does it look like *kmeans* have picked up on the patterns you observed in #1? Further—what you think of the separation between the clusters? Is there clear spacing betweeen the clusters, or are the borders almost touching each other (note that there will be certain overlap due to plotting the data in 2D)?

```
ggplot(dt,aes(x=jazz,pop,color=factor(kmeans_cluster))) + geom_point()
```




```
ggplot(dt,aes(x=jazz,hiphop,color=factor(kmeans_cluster))) + geom_point()
```



kmeans have indeed picked up on the density differences spotted earlier. Note: the reason for the two-colors-per-cluster pattern is that clusters—as we know from the labeling—are defined by three dimensions, not two.