PreLab 5: Using PCA for Analysis of Time Series Data

Jared Gridley

2/22/2021

Mouse cerebral cortex analysis

** This prelab will help you prepare for your group mini-project that will be created for you and has been posted to LMS prior. Please be sure to come to the lab sessions assigned to your team. There are 5 groups per Lab. All the coding has been done for you. You just have to read 'Prelab5.pdf' and do the quiz. Please answer the quiz using "~/MATP-4400/Lab5/Prelab5.pdf" or equivalently 'Prelab5.pdf' version uploaded to LMS. This will make sure your clusters match the clusters used to prepare the lab. **

Scott the Scientist did an analysis of `RNA-Seq` data from the development of the mouse cortex at days `-8`, `-4`, `0` , `1`, `7`, `16`, `21`, and `26` taken from the Allen Brain Atlas Developing Mouse Portal http://developingmouse.brain-map.org/. This type of data is know as time series data since the features are taken through time. Here is his preliminary report; your job is to help Scott understand the results!

Preparation of the Mouse Homologs Data

Scott begins by reading in the dataset and preparing the data frame. The columns are days at which the samples are collected. The entries in the columns are the amount of RNA for each gene detected on that day in the mouse embryo cerebral cortex. We can use `summary()` and see that the column mean in not 0. This data has already been scaled, so each **row** has *mean 0 and sd 1*. This is different from prior labs where we scaled by **column**. **Row Scaling** has been done so that the analysis can focus on the shape of the time series rather than specific magnitudes. Scott confirmed the scaling was successful by calculating the row means and making sure their norm was near 0.

Here is the mapping of the columns to their actual meanings:

- `DayNeg8` = 8 days before birth
- `DayNeg4` = 4 days before birth
- `Day0` = day of birth
- `DayPos1` = 1 day after birth
- `DayPos7` = 7 days after birth
- `DayPos16` = 16 days after birth
- `DayPos21` = 21 days after birth
- `DayPos28` = 28 days after birth

```
# Read in the data and create a dataframe
# We read in the csv indicating that we have row names.
Mouse.df <-read.csv("~/MATP-4400/data/MouseHomologData.csv", row.names = 1)

# Use shorter column names
colnames(Mouse.df)<-c("-8","-4","0","1","7","16","21","28")

# Create a matrix for our analysis
Mouse.matrix <- as.matrix(Mouse.df)
```

```
# Summarize; note the scaling
summary(Mouse.df)
```

```
##        -8                -4                0                 1
##  Min.   :-2.2436   Min.   :-2.3882   Min.   :-2.3330   Min.   :-2.02584
##  1st Qu.:-0.9138   1st Qu.:-1.0365   1st Qu.:-0.5185   1st Qu.:-0.52867
##  Median :-0.1530   Median :-0.5435   Median : 0.1894   Median :-0.07383
##  Mean   : 0.2260   Mean   :-0.3335   Mean   : 0.3118   Mean   : 0.13705
##  3rd Qu.: 1.4918   3rd Qu.: 0.3735   3rd Qu.: 1.0578   3rd Qu.: 0.76547
##  Max.   : 2.4749   Max.   : 2.4710   Max.   : 2.4739   Max.   : 2.47368
##        7                 16                21                28
##  Min.   :-1.58200   Min.   :-1.7886   Min.   :-1.7495   Min.   :-2.04906
##  1st Qu.:-0.44628   1st Qu.:-0.7918   1st Qu.:-0.7839   1st Qu.:-0.75671
##  Median :-0.04786   Median :-0.4784   Median :-0.4840   Median :-0.44015
##  Mean   : 0.18491   Mean   :-0.2320   Mean   :-0.1980   Mean   :-0.09629
##  3rd Qu.: 0.71298   3rd Qu.: 0.3390   3rd Qu.: 0.4180   3rd Qu.: 0.51383
##  Max.   : 2.45886   Max.   : 2.4749   Max.   : 2.4749   Max.   : 2.47487
```

```
# Demonstrate the scaling by viewing the norm
# norm(rowMeans(Mouse.matrix))
```

Cluster and PCA Analysis

We used Kmeans to create five clusters based on domain knowledge: biologists believe there are five stages of brain development, so we select five clusters. Examining the plot of the kmean objective by cluster size using the "elbow test" suggests that a smaller number of clusters *could* be used; you can verify this on your own. But as we will see, five clusters proves to be an appropriate number for this analysis.

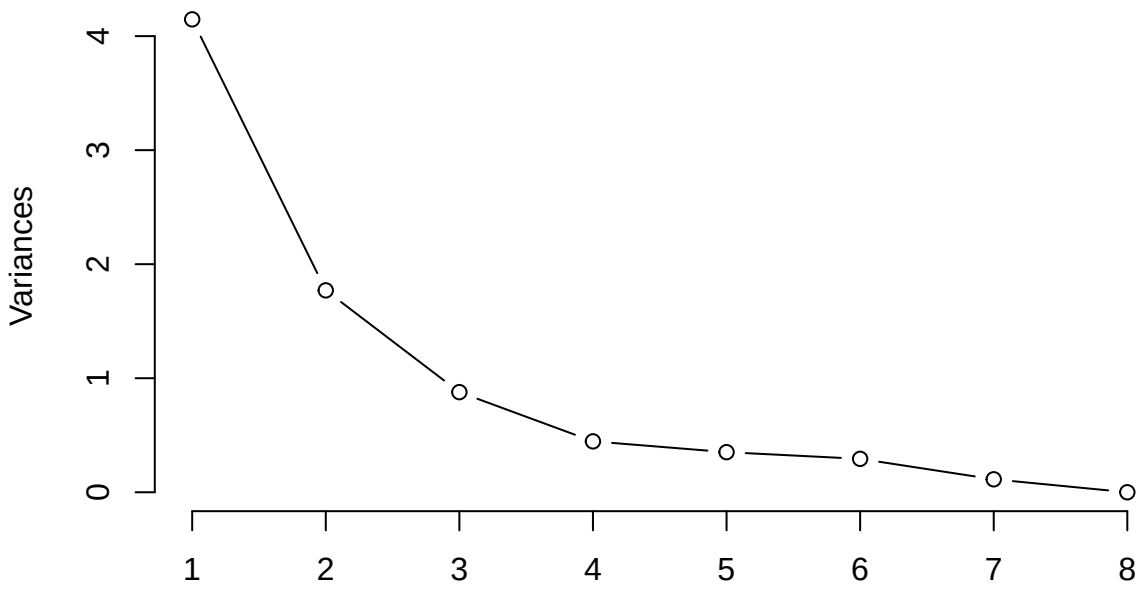
```
set.seed(300)
km <-kmeans(Mouse.matrix, 5)
```

We visualize the cluster using a biplot of two components generated by PCA which explain 74% of the variance. The scree plot suggests that PC3 might also contain significant variance.

```
# Calculate the PCA
my.pca <- prcomp(Mouse.matrix, retx=TRUE, center=TRUE, scale=TRUE)
# Summarize, to see the complete PCA result
summary(my.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0364 1.3306 0.9370 0.66801 0.59296 0.54132 0.33712
## Proportion of Variance 0.5184 0.2213 0.1098 0.05578 0.04395 0.03663 0.01421
## Cumulative Proportion  0.5184 0.7397 0.8494 0.90521 0.94917 0.98579 1.00000
##                           PC8
## Standard deviation     3.306e-11
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

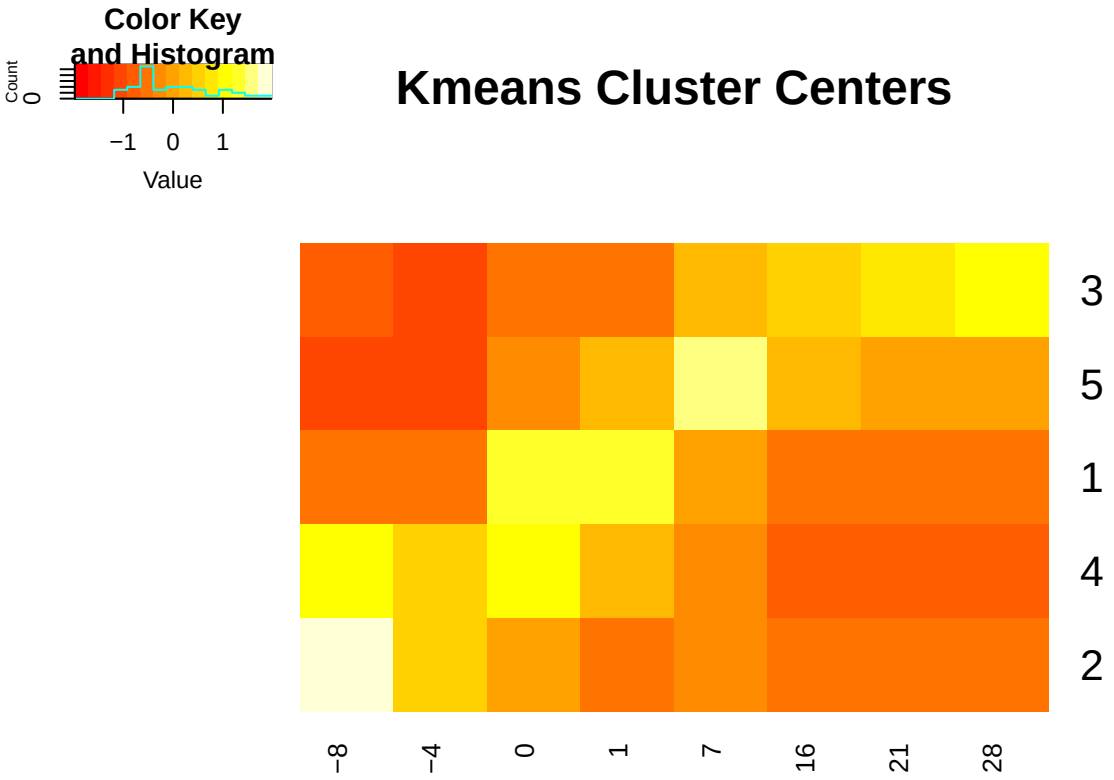
```
# Generate a scree plot
screeplot(my.pca, type = "lines",
          main = 'Explained Variance of Mouse Genes')
```



Examining the heatmap of the Kmeans cluster centers, we can see that each cluster corresponds to different average peaks of gene expressions. A line plot is also a very effective way to view the means as time-series.

```
set.seed(300)
heatmap.2(km$centers,
          scale = "none",
          dendrogram = "none",
          Colv=FALSE,
          cexCol=1.0,
          alpha =0,
          main = "Kmeans Cluster Centers",
          trace ="none")
```

```
## Warning in plot.window(...): "alpha" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "alpha" is not a graphical parameter
## Warning in title(...): "alpha" is not a graphical parameter
```



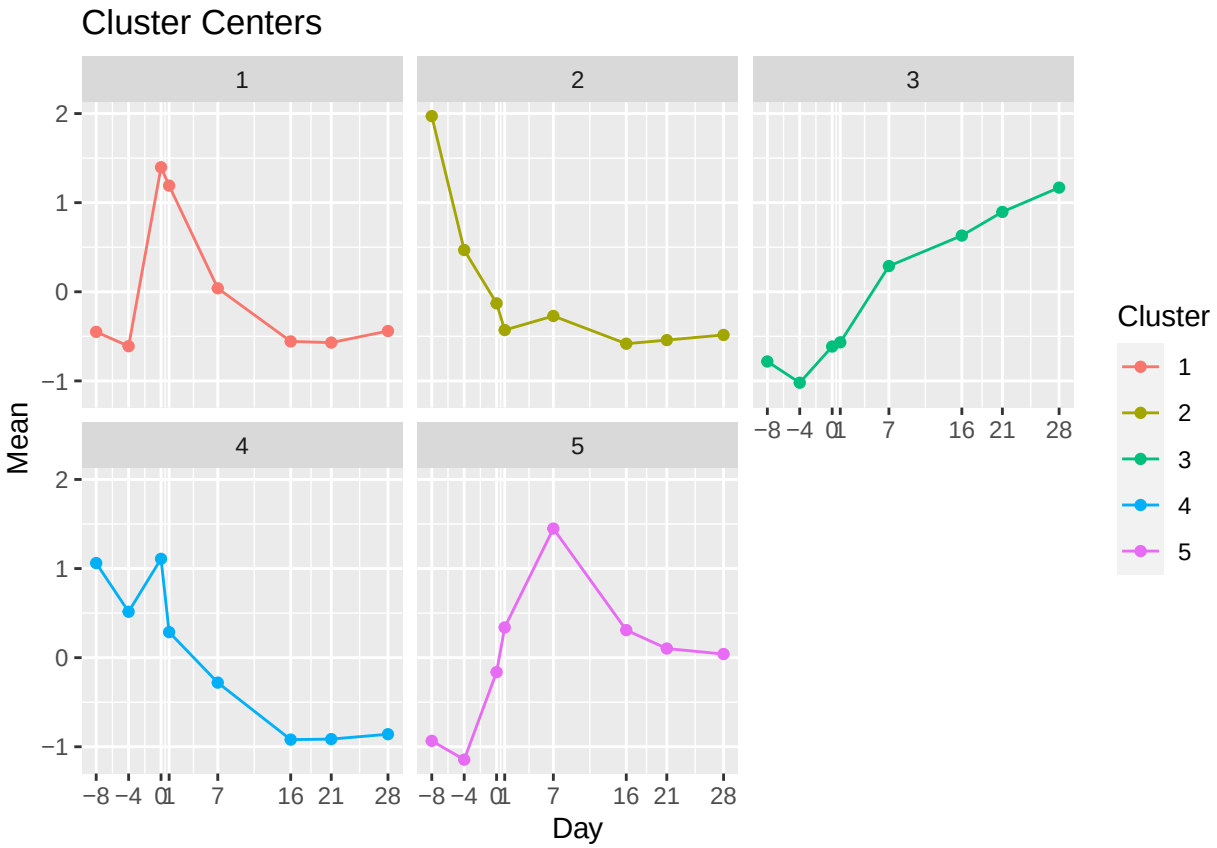
We can also see the time trends in the clusters means by plotting each cluster mean as a line. The cluster means have to be reformatted into a data frame with columns 'Cluster', 'Day' and 'Mean'. This is done using the 'dplyr' package 'gather' command. The factors are also recoded to look nice on the plot.

We use `ggplot` with `geom_line` to make the plots. Note how the 'facet_grid()

```
tics<-c(-8,-4,0,1,7,16,21,28) # x-axis "tics" (for the plot)
clustermean<-km$centers # Extract the cluster means from km

# Set up ggplot-based line plot
# We must "pivot" at dataframe version of clustermean (ie convert from "wide" to "long")
# Rows are our groups
clustermean.df <- as.data.frame(clustermean, row.names = c("1","2","3","4","5"))

# Tidyverse pipeline
# This is making a data frame of the  form
# Cluster Day Mean

clustermeanlong.df <- clustermean.df %>%
  rownames_to_column("Cluster") %>% #  Make a new column called Cluster
  gather(key="Day",value="Mean", -Cluster) %>% # Make a skinny data frame
  #Recode the factors to have short names
 # mutate(Day=recode(Day,"DayNeg8"="-8","DayNeg4"="-4","Day0"="0", "DayPos1"= "1",
 #        "DayPos7"= "7", "DayPos16"= "16","DayPos21"= "21", "DayPos28"= "28"))%>%
  # convert Day to an integer
 convert(int(Day))
# see what data frame looks like.,
kable(head(clustermeanlong.df))
```

Cluster	Day	Mean
1	-8	-0.4496261
2	-8	1.9712955
3	-8	-0.7822509
4	-8	1.0614302
5	-8	-0.9336692
1	-4	-0.6106791

```
#Plot the mean of each cluster in a separate graph":
ggplot(clustermeanlong.df,aes(x=Day, y=Mean, col=Cluster)) +
geom_line() +
geom_point() +
scale_x_continuous(breaks=tics) +
labs(title="Cluster Centers") +
# Use facet_wrap to make a separate plot for each cluster
facet_wrap(Cluster ~.)
```



Take a look, what patterns do you see in the clusters?

Exercise 1

For each cluster mean, identify the peak, i.e. the day with the highest average value. Cluster 5 has been completed for you.

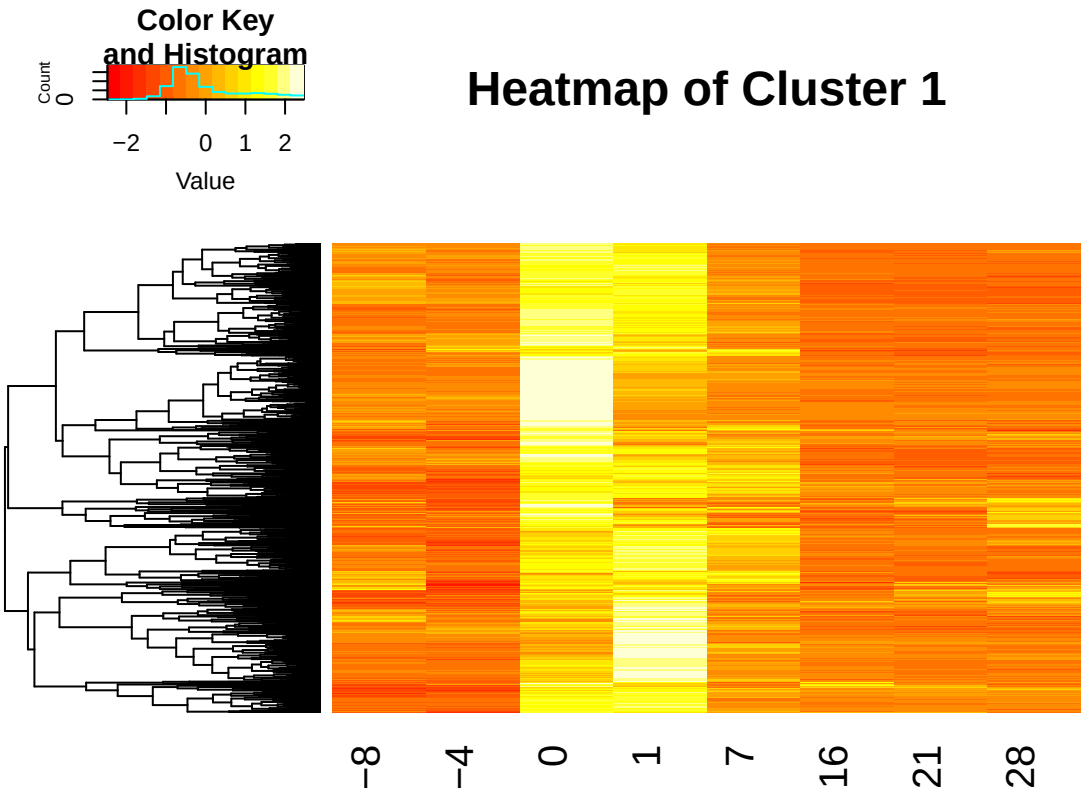
- Cluster 1 = `Day 0`
- Cluster 2 = `Day -8`
- Cluster 3 = `Day 28`

- Cluster 4 = `Day 0`
- Cluster 5 = `Day 7`

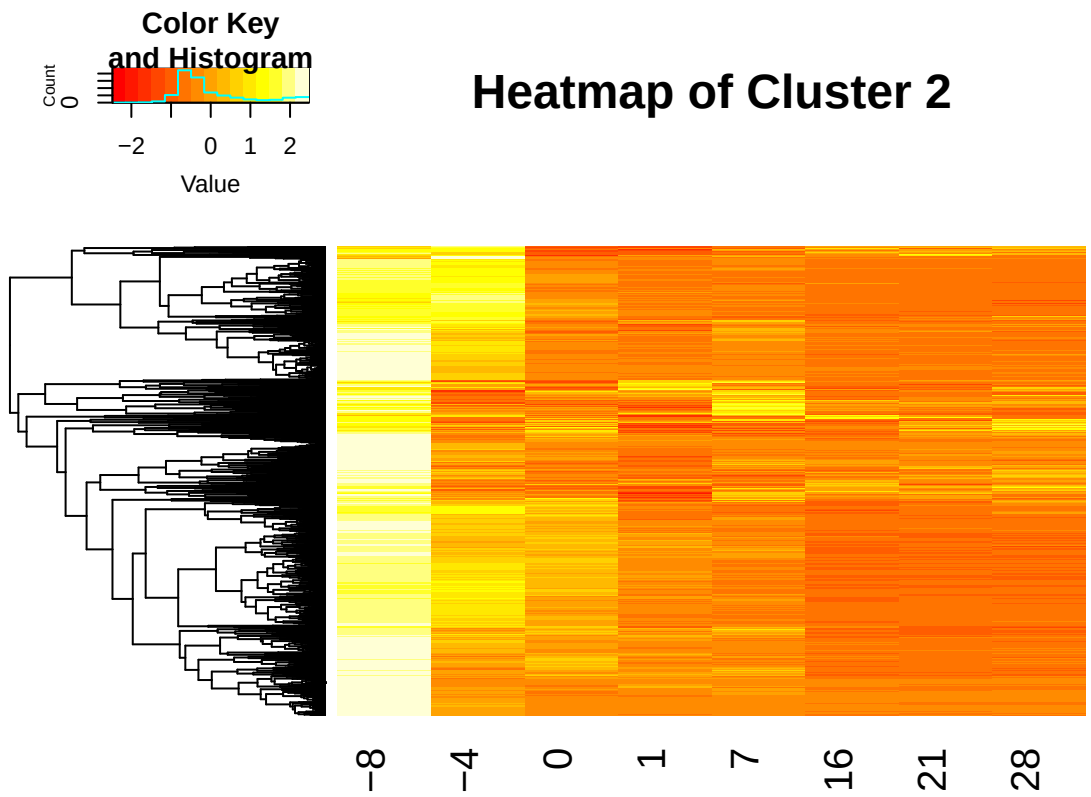
Exercise 2

Make a separate heat map for each of the five clusters. Plot the cluster heatmaps in the order they occur in development. **Do not scale the heatmaps.** Cluster the genes but not the days. An example of how to plot Cluster 5 using heatmap is provided. How do the heatmaps match your results for exercise 1?

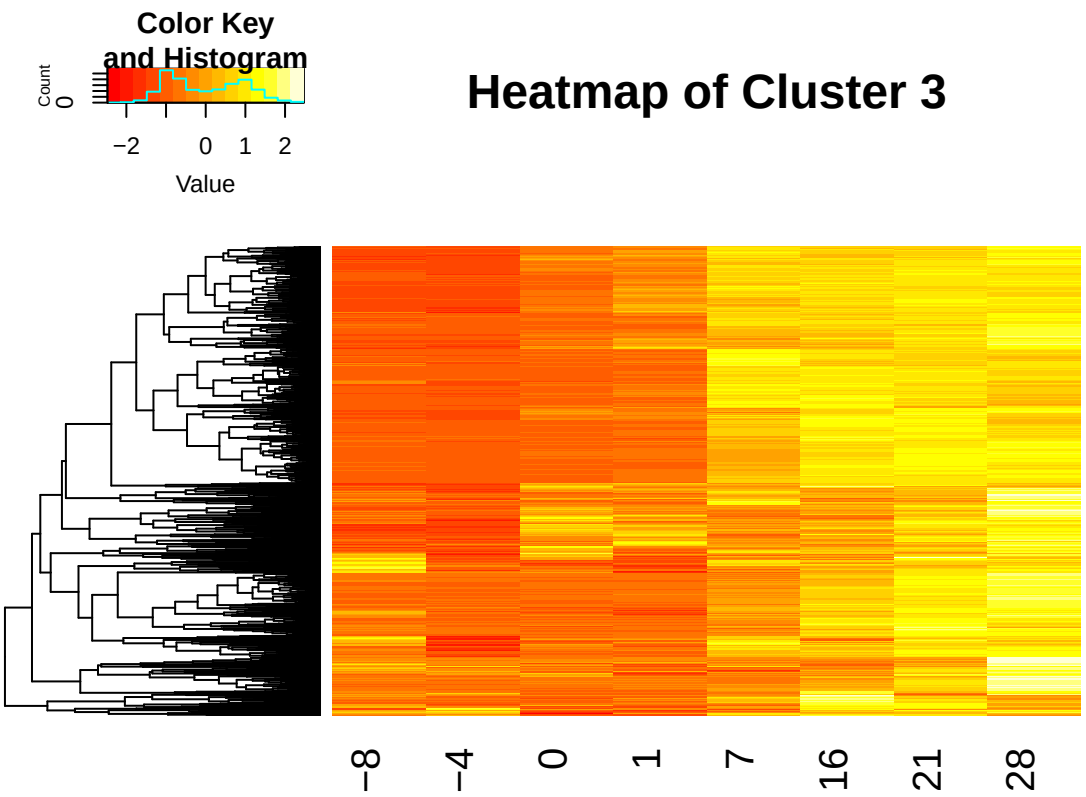
```
set.seed(300)
heatmap.2(Mouse.matrix[km$cluster==1,],
          scale = "none",
          dendrogram = "row",
          Colv=FALSE,
          main = "Heatmap of Cluster 1",
          cexCol=1.5,
          labRow= NA,
          trace ="none")
```



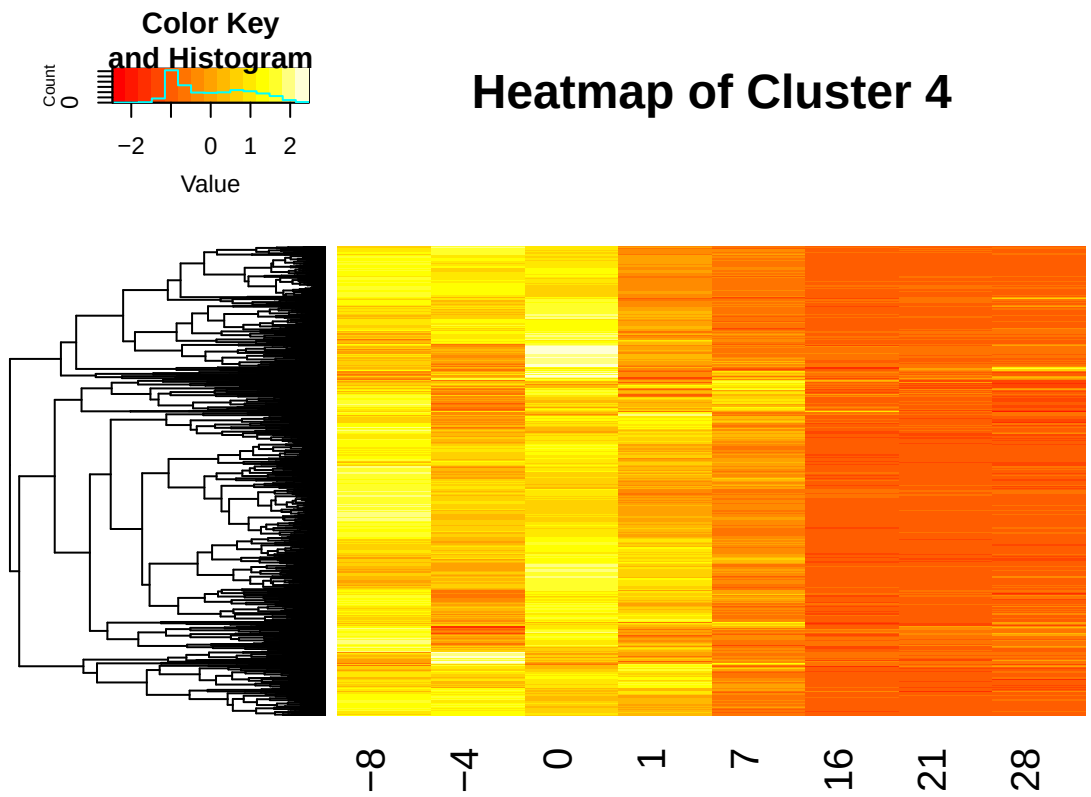

```
heatmap.2(Mouse.matrix[km$cluster==2,],
          scale = "none",
          dendrogram = "row",
          Colv=FALSE,
          main = "Heatmap of Cluster 2",
          cexCol=1.5,
          labRow= NA,
          trace ="none")
```



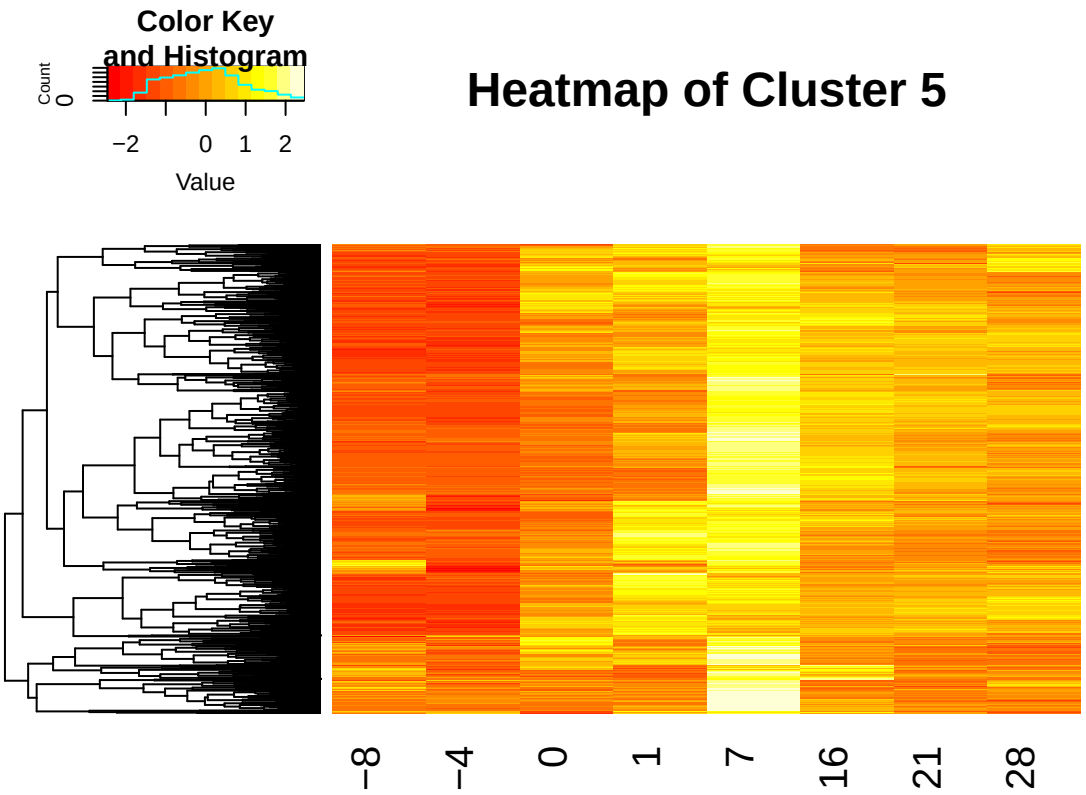

```
heatmap.2(Mouse.matrix[km$cluster==3,],
          scale = "none",
          dendrogram = "row",
          Colv=FALSE,
          main = "Heatmap of Cluster 3",
          cexCol=1.5,
          labRow= NA,
          trace ="none")
```




```
heatmap.2(Mouse.matrix[km$cluster==4,],
          scale = "none",
          dendrogram = "row",
          Colv=FALSE,
          main = "Heatmap of Cluster 4",
          cexCol=1.5,
          labRow= NA,
          trace ="none")
```



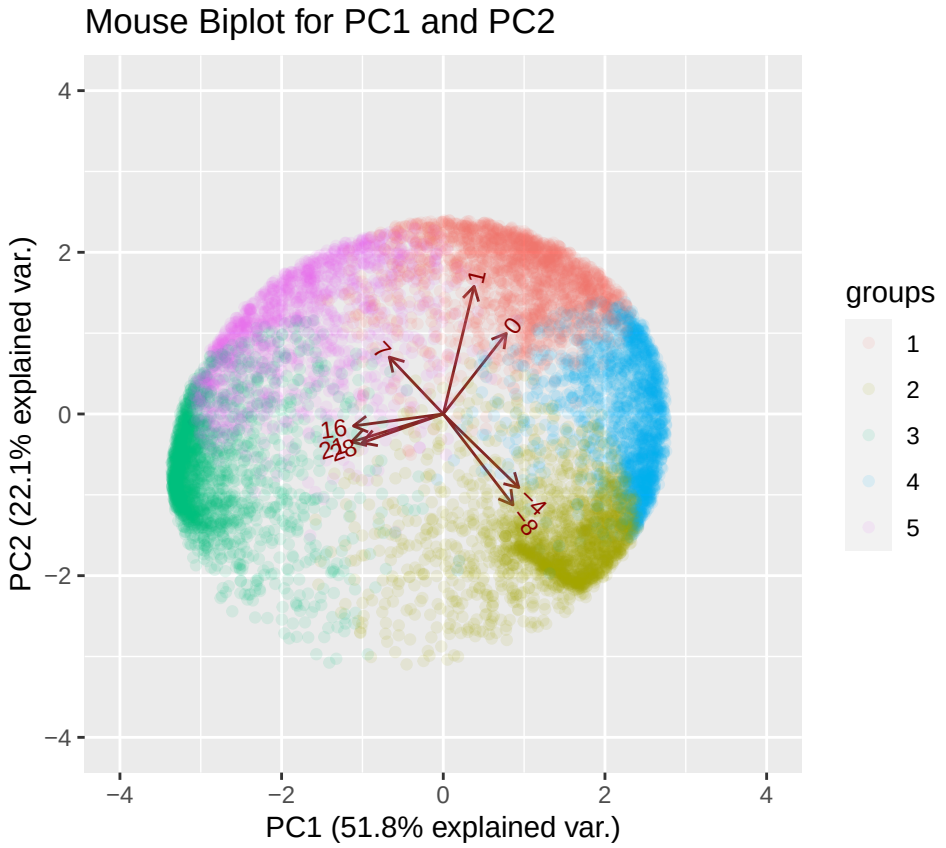

```
heatmap.2(Mouse.matrix[km$cluster==5,],
          scale = "none",
          dendrogram = "row",
          Colv=FALSE,
          main = "Heatmap of Cluster 5",
          cexCol=1.5,
          labRow= NA,
          trace ="none")
```



Visualization by biplot

We display the points in a biplot. The projection of the points makes an interesting disc-type shape which is less dense in the middle, like a donut. We can see that the clusters are arranged in time order around the disc. **Why do we see this?**

```
# Calculate x and y scale limits for the biplot
t<-1.2*max(abs(my.pca$x[,1:2]))
# Generate the biplot using ggbiplot
p <- ggbiplot(my.pca,
              choices=c(1,2),  # Use PC1, PC2
              alpha=.1,        # Make dots transparent
              varname.adjust=1.5,  # Move variables names out a bit
              scale =0,       # Don't rescale data
              groups=as.factor(km$cluster))
p + ggtitle('Mouse Biplot for PC1 and PC2') + xlim(-t,t) + ylim(-t,t) # title plot and make square
```



Exercise 3

We hypothesize that each cluster of genes represents one of the five "stages" of brain development labeled A,B,C,D,E. Assign each cluster to its corresponding stage using the biplot. For example, Cluster 5 represents Stage D of development since it peaks at the first at the point `Day7`.

- Stage A = Cluster 2
- Stage B = Cluster 1 (Can also be 4 if not allowing for duplicates, but matches 1 better)
- Stage C = Cluster 1
- Stage D = Cluster 5
- Stage E = Cluster 3

Exercise 4

Examine the scalar projections of the coordinate axes in the biplot, for Days `-8`, `-4`, `0`, `1`, `7`, `16`, `21` and `28`. Notice that the coordinate vectors act as hours on a "developmental time clock" that starts at `Day -8`. *Does time on this development clock run clockwise or counterclockwise?* Counterclockwise

You've now completed in class Prelab5! Go to LMS and complete the online quiz.

‘‘‘