HW Mnist 資料集視覺化

高嘉妤、柯堯珹、吳承恩、趙友誠

2024-10-31

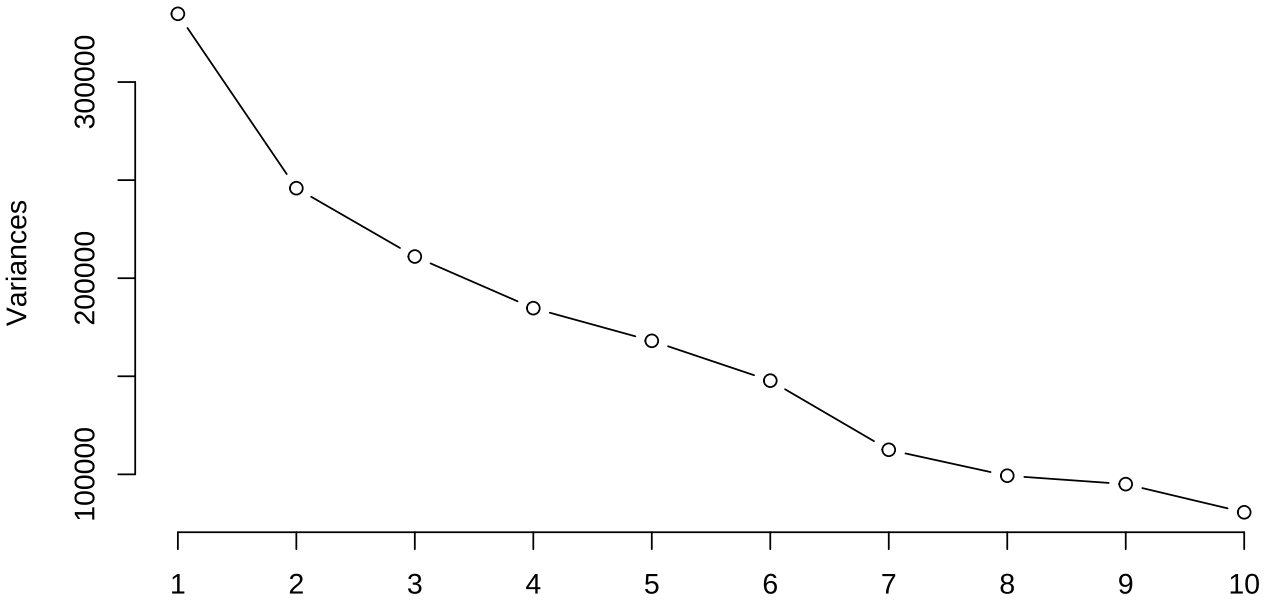
Table of contents


```
mnist <- data.table::fread("MNIST_train.csv")
```

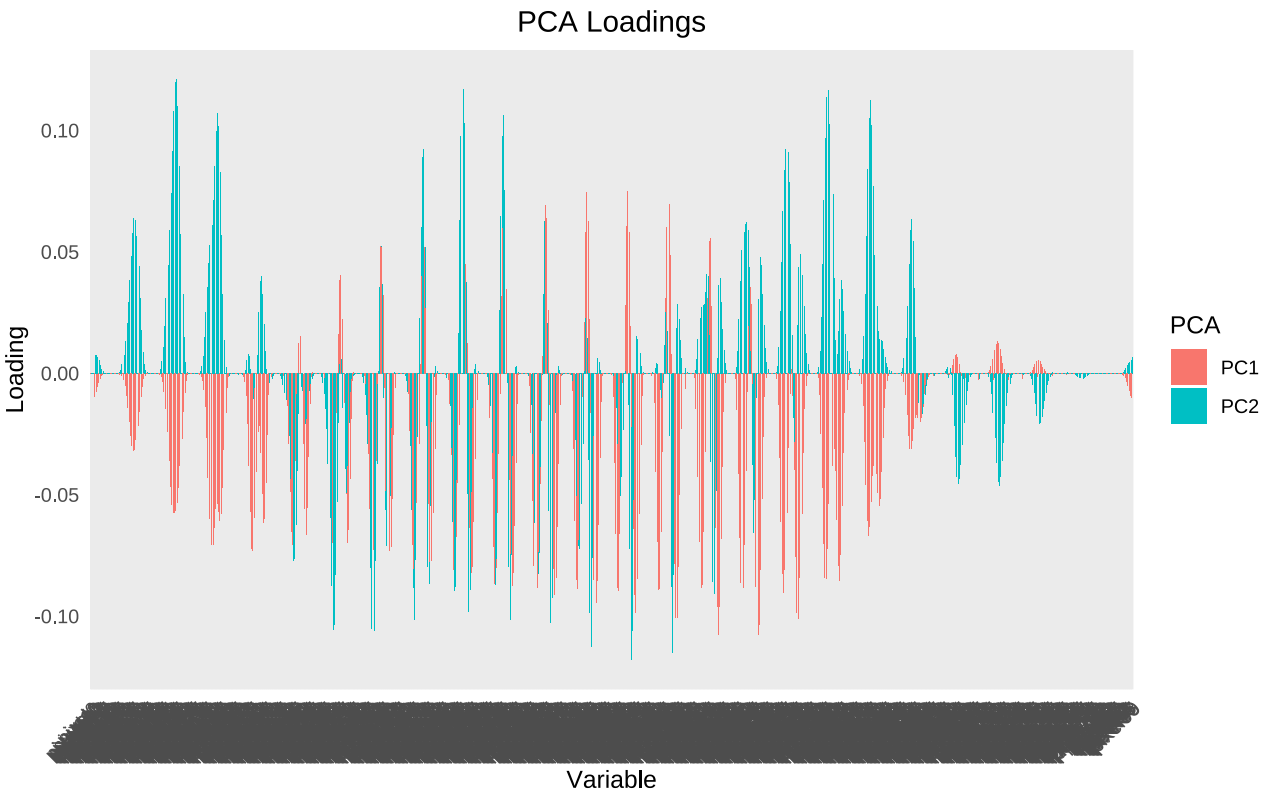
Visualization by PCA

```
library(ggplot2)
library(dplyr)
library(showtext)
showtext_auto()
pca<-prcomp(mnist[,2:785],center = T)
# 做 Scree plot 可以發現 1 到 2 的斜率是最大的，因此我取到 pc2
screeplot(pca,type = 'line', main= "Fig 1: Scree plot")
```

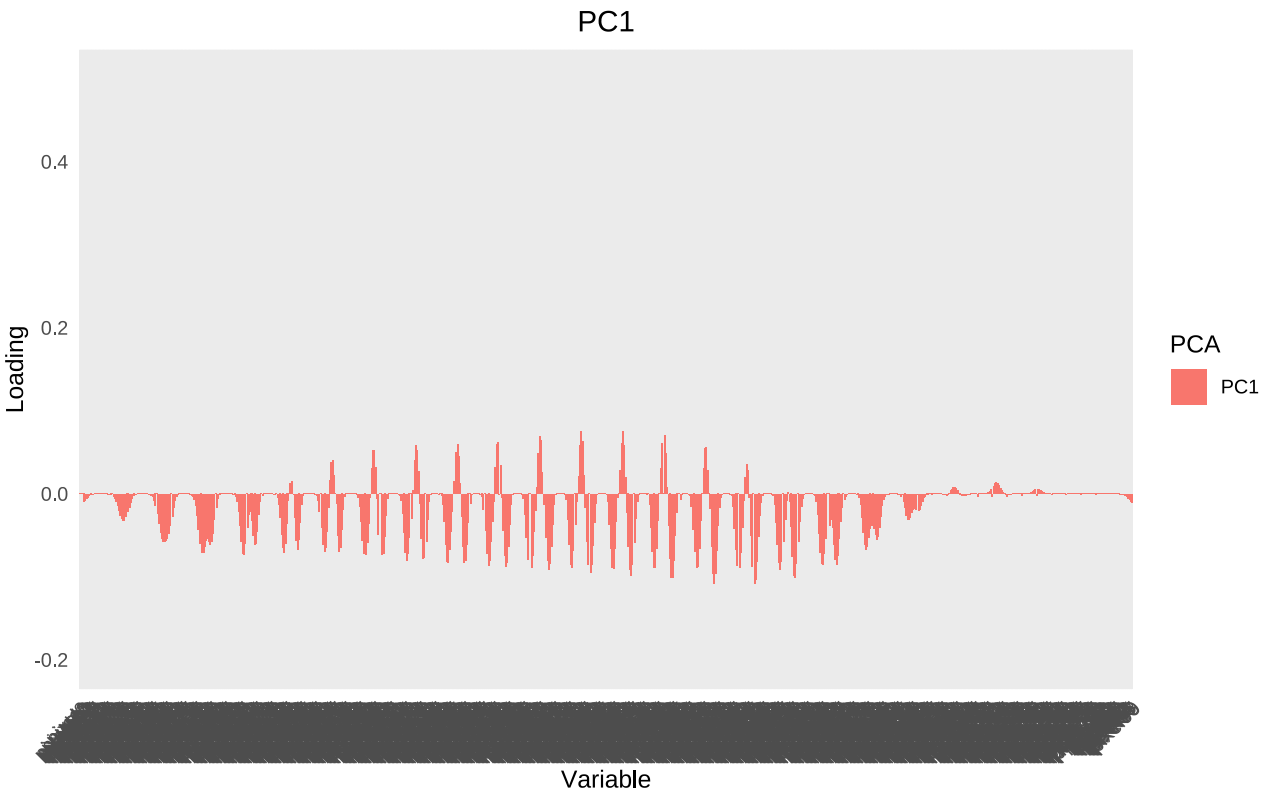
Fig 1: Scree plot



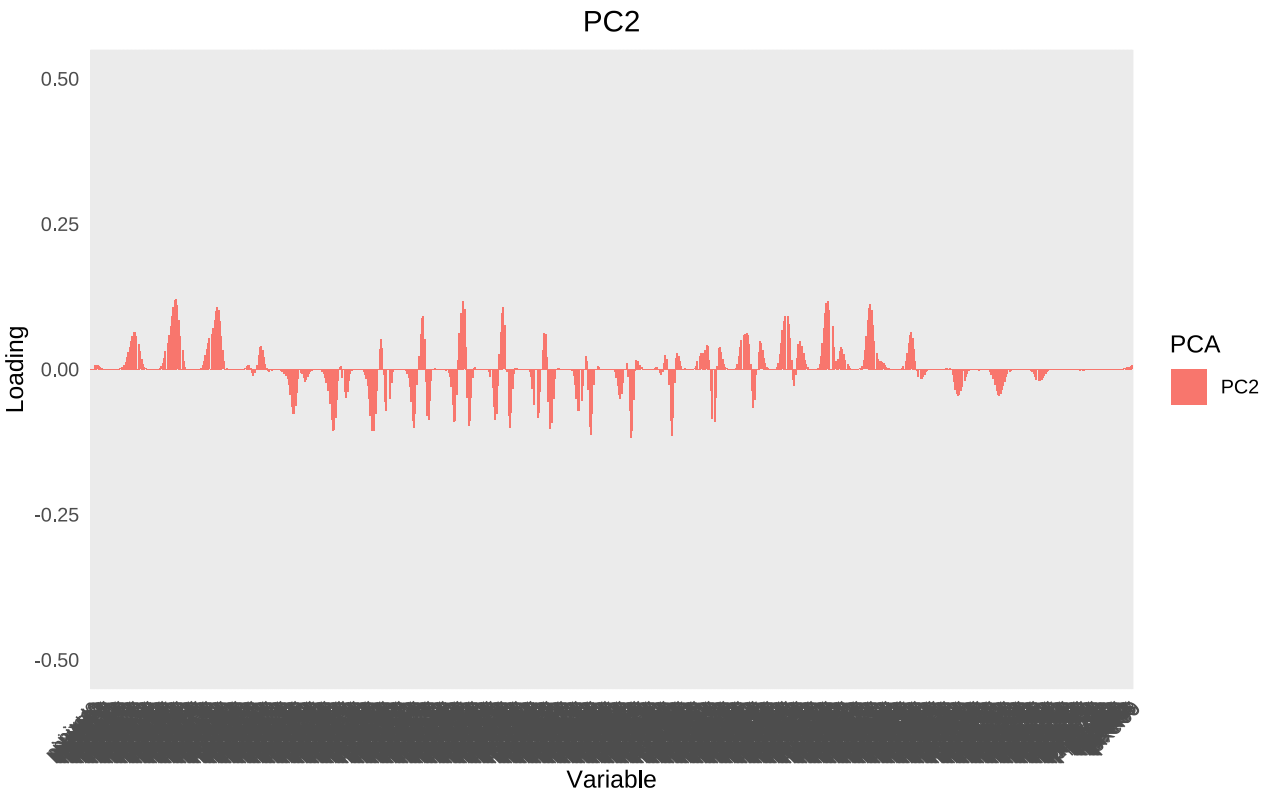

```r
# 取 pc1 跟 pc2
rec_pc <- as.data.frame(pca$x[,1:2])
# 將數字的 label 加進去 pc1 跟 pc2 的 dataset
rec_pc$label<-mnist$label

# 轉成 character 以方便後續做圖
rec_pc$label <- as.character(rec_pc$label)

#pc1 pc2 的 loading (但是變數太多看不出什麼東西)
pca_rotation<-pca$rotation
pca_rotation_df <- data.frame(Variable = rownames(pca_rotation),
                              PC1 = pca_rotation[, 1],
                              PC2 = pca_rotation[, 2])
pca_rotation_longs <- tidyr::pivot_longer(pca_rotation_df,
                                          cols = -Variable,
                                          names_to = "PCA",
                                          values_to = "Rotation")
#loadings data
pca_rotation1<-pca_rotation_longs%>%
  filter(PCA=='PC1')
pca_rotation2<-pca_rotation_longs%>%
  filter(PCA=='PC2')

#loading 圖 pc1 跟 pc2 一起的
ggplot(pca_rotation_longs, aes(x = Variable, y = Rotation, fill = PCA)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "PCA Loadings", x = "Variable", y = "Loading") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



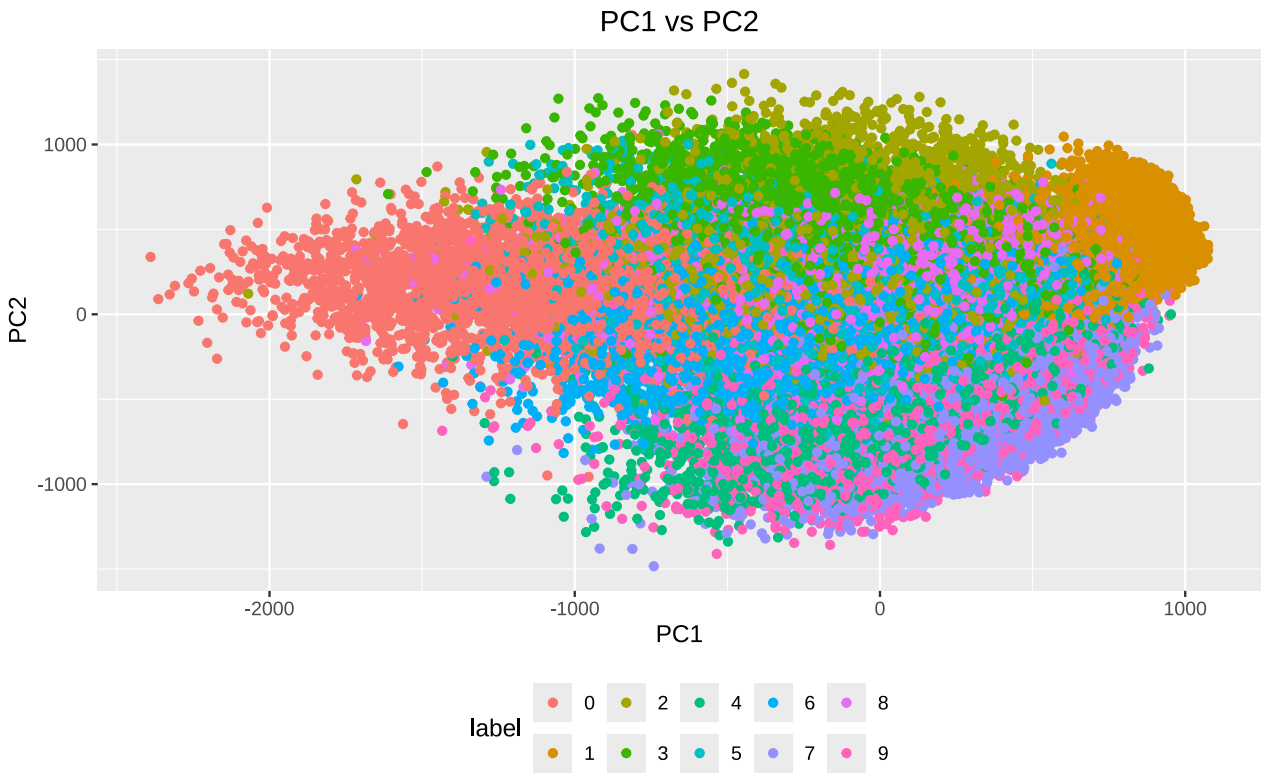

```
# 單獨 pc1 loading 的圖
ggplot(pca_rotation1, aes(x = Variable, y = Rotation, fill = PCA)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "PC1", x = "Variable", y = "Loading") +
  theme_minimal() +
  scale_y_continuous(limits = c(-0.2,0.5))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
# 單獨 pc2 loading 的圖
ggplot(pca_rotation2, aes(x = Variable, y = Rotation, fill = PCA)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "PC2", x = "Variable", y = "Loading") +
  theme_minimal() +
  scale_y_continuous(limits = c(-0.5,0.5))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
# 製作 pc1 對 pc2 的圖，也可以看出大部份同一個 label 都還是有 cluster 形成一個 group
ggplot(rec_pc,aes(x = PC1, y = PC2,colour = label))+
  geom_point()+ labs(title = 'PC1 vs PC2')+
  theme(plot.title = element_text(hjust = 0.5), legend.position = 'bottom')
```



Visualization by MDS

Visualization by t-SNE

```
library(Rtsne)
set.seed(123)
tsne_result <- Rtsne(mnist[,-1], dims = 2)
tsne_df <- as.data.frame(tsne_result$Y)
centroids <- tsne_df %>%
  mutate(label = mnist$label) %>%  # Add labels to the dataframe
  group_by(label) %>%
  summarize(V1 = mean(V1), V2 = mean(V2), .groups = 'drop')

ggplot(data = tsne_df)+
  geom_point(aes(x=V1, y=V2, color  = as.factor(mnist$label)))+
  labs(title = "Fig 3: t-SNE Visualization of MNIST Data",
       x = "V1",
       y = "V2",
       color = "Number") +
  geom_text(data = centroids, aes(x = V1, y = V2, label = label),
            vjust = -1, size = 4, color = "black") +
  theme_gray()
```

Fig 3: t-SNE Visualization of MNIST Data