626Midterm2

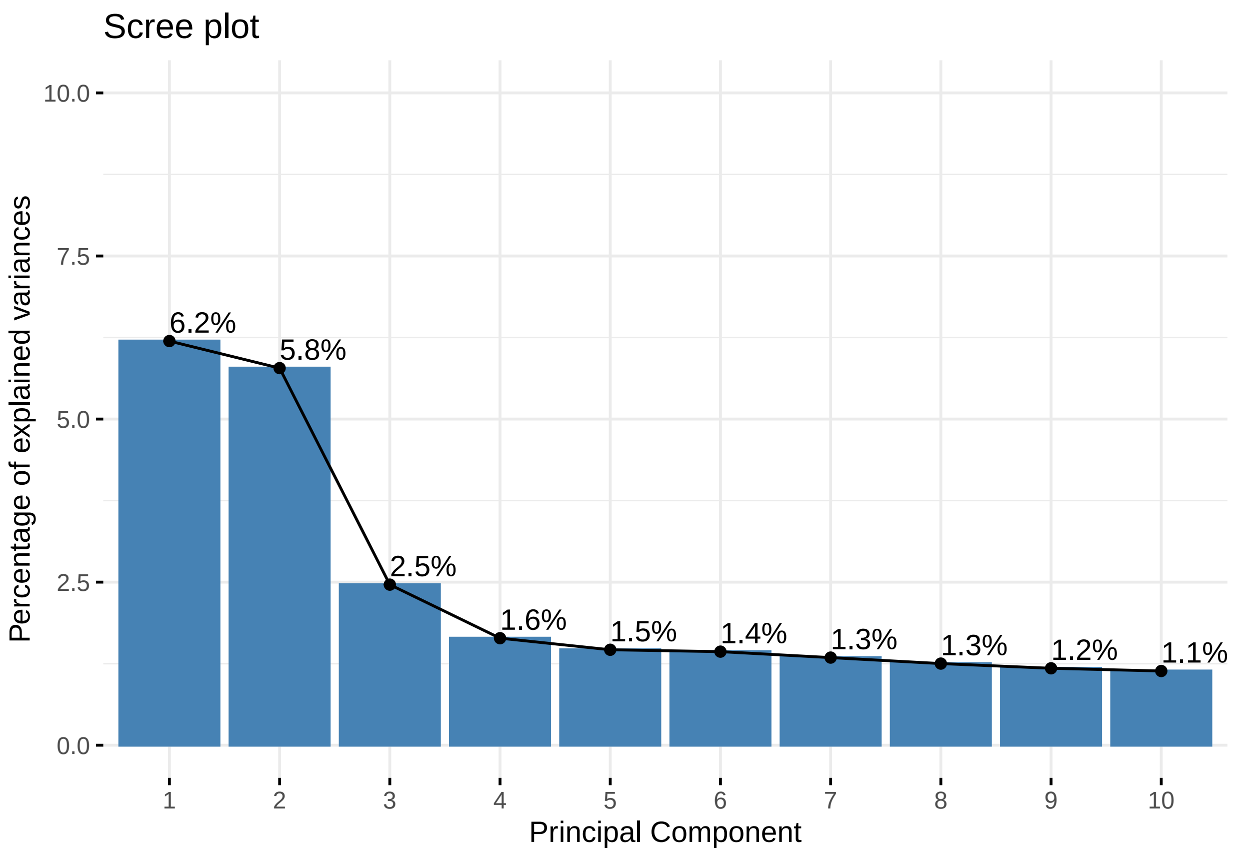
Wenjing Li

2023-04-19

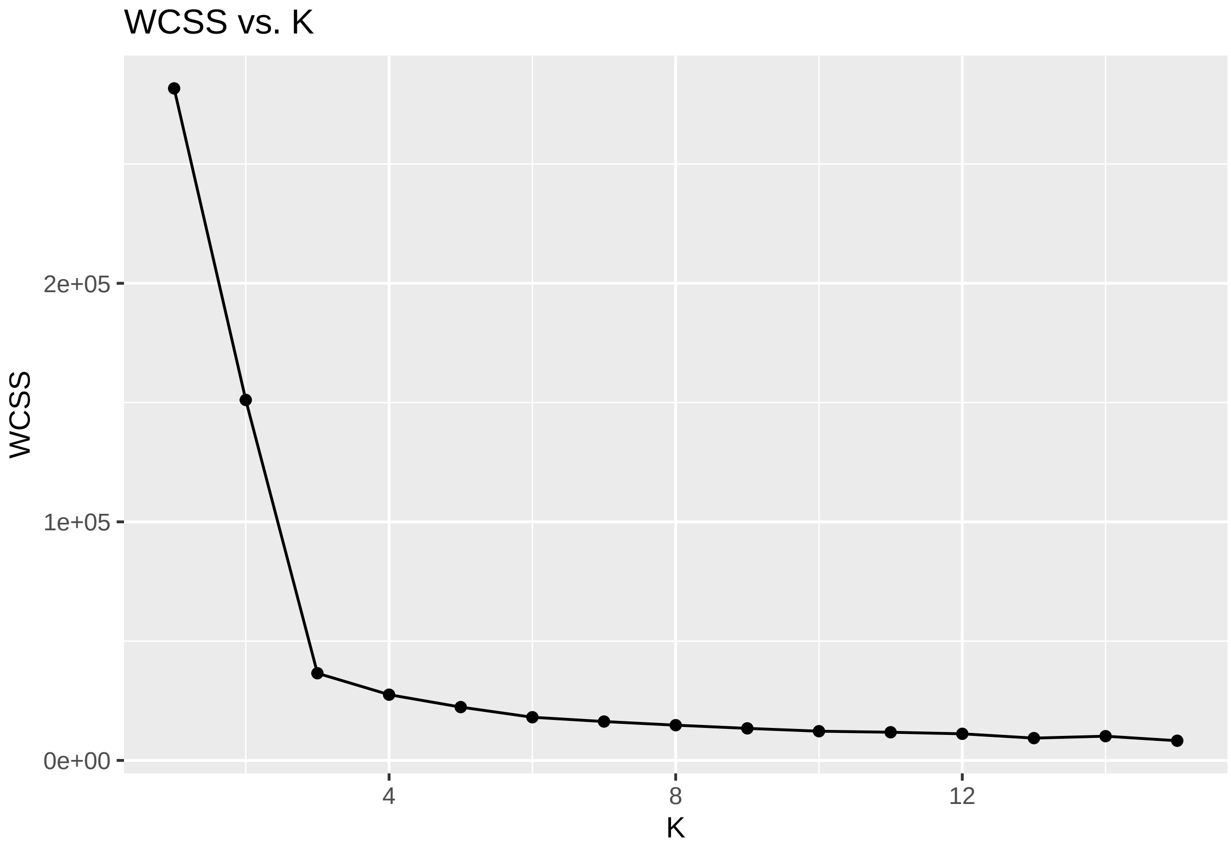
Problem 1

Use PCA followed by clustering algorithms to explore the population structure with only the genotype data (i.e., ignore sampling location and continent information).

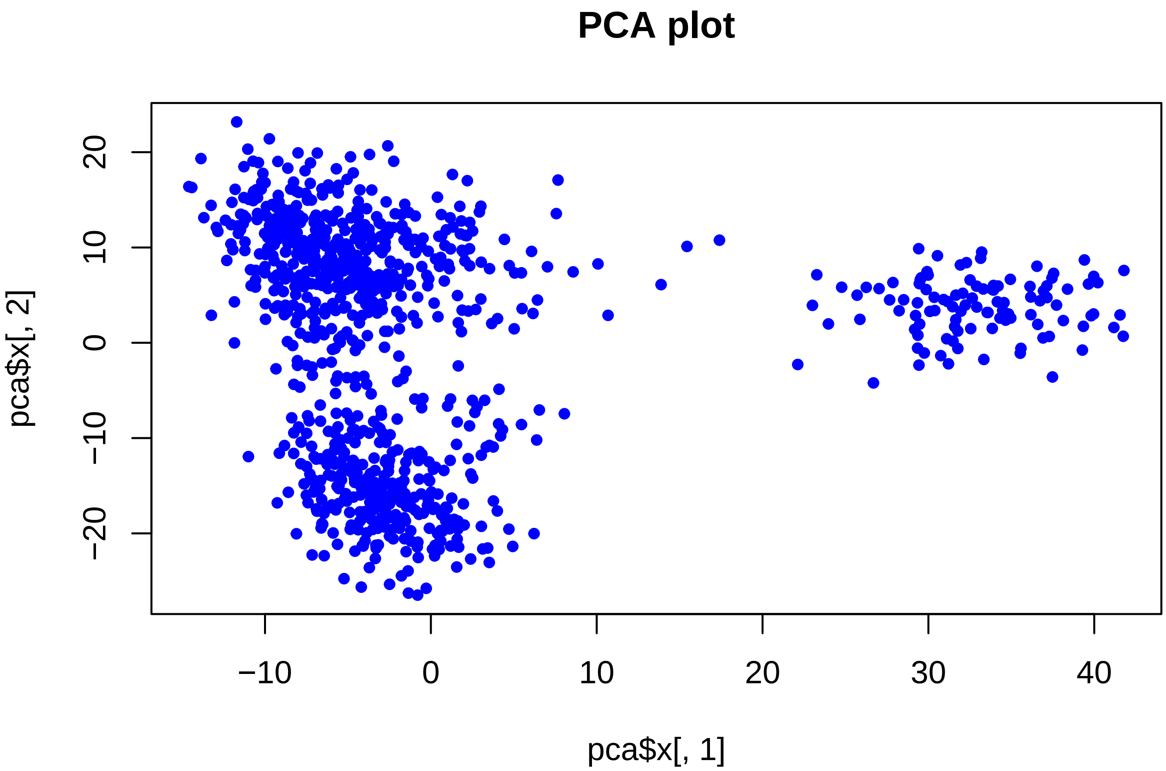
```
Df <- read.table("~/Downloads/hgdp.txt")
colnames(Df)[1:3] <- c("Individual_ID","location","continent")
genotypes <- Df %>% select(starts_with("V"))
convert_genotypes <- function(genotype_data) {

freq_table <- table( c(substr(genotype_data,1,1),substr(genotype_data,2,2)))
most_freq <- names(freq_table)[which.max(freq_table)]
ref_char <- most_freq
encoded_vector<- ifelse(substr(genotype_data,1,1) == ref_char, 1,0) + ifelse(substr(genotype_data,2,2) =
}
encoded_df <- apply(genotypes, 2, convert_genotypes )

# Perform PCA
pca <- prcomp(encoded_df, scale = TRUE)

#Plot the scree plot to visualize the amount of variation explained by each component.
fviz_eig(pca, xlab = "Principal Component" ,addlabels = TRUE, ylim = c(0, 10))
```



```r
#Choose the K
pca_scores <- pca$x[, 1:2]
set.seed(123)
wcss =  matrix(cbind(c(1:15),rep(0,15)),ncol = 2)
for (i in 1:15){
  wcss[i,2] <- kmeans(pca_scores, i, 10)$tot.withinss
}
df_wcss <- as.data.frame(wcss)

ggplot(df_wcss, aes(x = V1, y = V2)) + geom_line() + geom_point() +
  labs(x = "K", y = "WCSS") +
  ggtitle("WCSS vs. K")
```



```
# Plot the first two principal components
plot(pca$x[,1], pca$x[,2], pch = 20, col = "blue", main = "PCA plot")
```

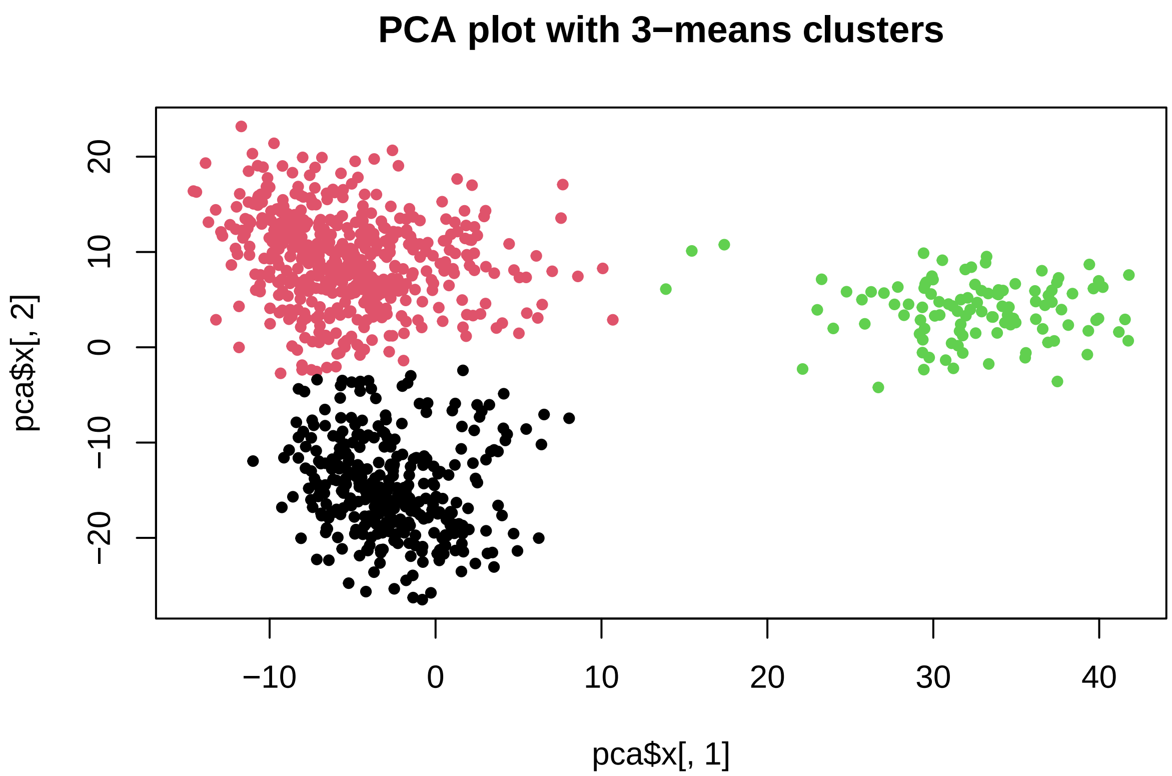


The result of precentage of variation explained by each element, as from this plot, the top two principle components can explain 6.2% and 5.8% of variance relatively, and the explained ratio significantly higher than the other components. For clustering, K-means method is used. We can use elbow method to choose the K, which works by finding WCSS (Within-Cluster Sum of Square) and identifing where the curve in WCSS plot happened. Or we can directly look into the pattern of plot.Here We choose 3 as the number of K.

Problem 2

Visualize the cluster structures identified from PCA and clustering analysis, color each sample point using its continental information. (you may wish to plot multiple pairs of PC scores)

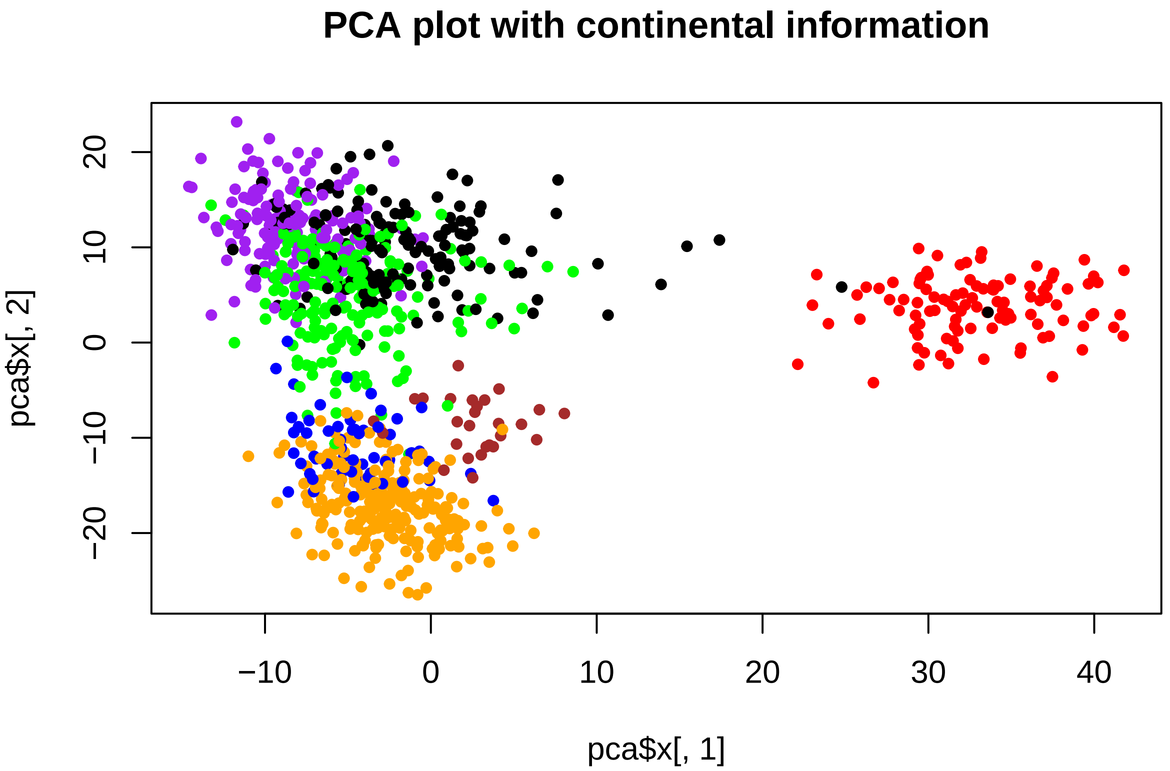
```
# Plot the PCA plot with 3-means cluster assignments
# Perform K-means clustering on the PCA scores
kmeans_res <- kmeans(pca$x[,1:2], centers = 3)
plot(pca$x[,1], pca$x[,2], pch = 20, col = kmeans_res$cluster, main = "PCA plot with 3-means clusters")
```



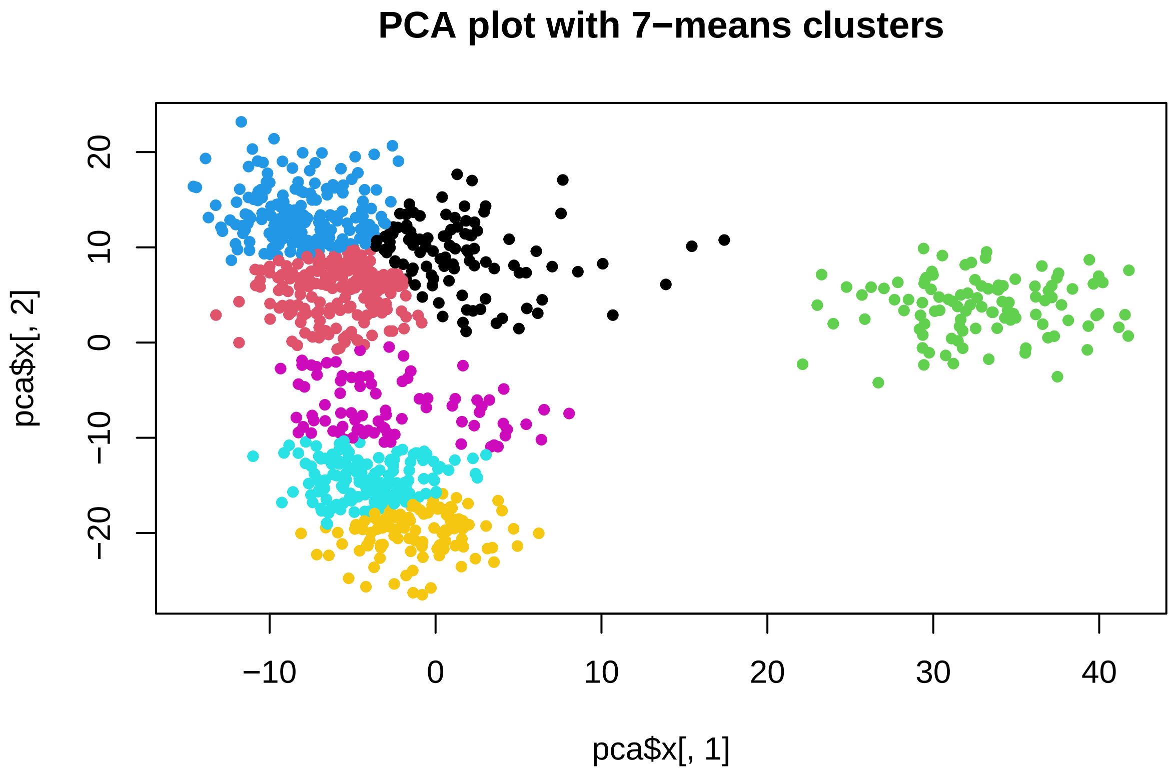
```
color <- as.numeric(factor(Df$continent))
colors <- c("red", "blue", "green", "orange", "purple", "black","brown")
color_vector <- colors[color]
color_table <- matrix(c(names(table(Df$continent)),colors),ncol = 2)
color_table
```

```
##      [,1]                [,2]
## [1,] "AFRICA"            "red"
## [2,] "AMERICA"           "blue"
## [3,] "CENTRAL_SOUTH_ASIA" "green"
## [4,] "EAST_ASIA"         "orange"
## [5,] "EUROPE"            "purple"
## [6,] "MIDDLE_EAST"       "black"
## [7,] "OCEANIA"           "brown"
```

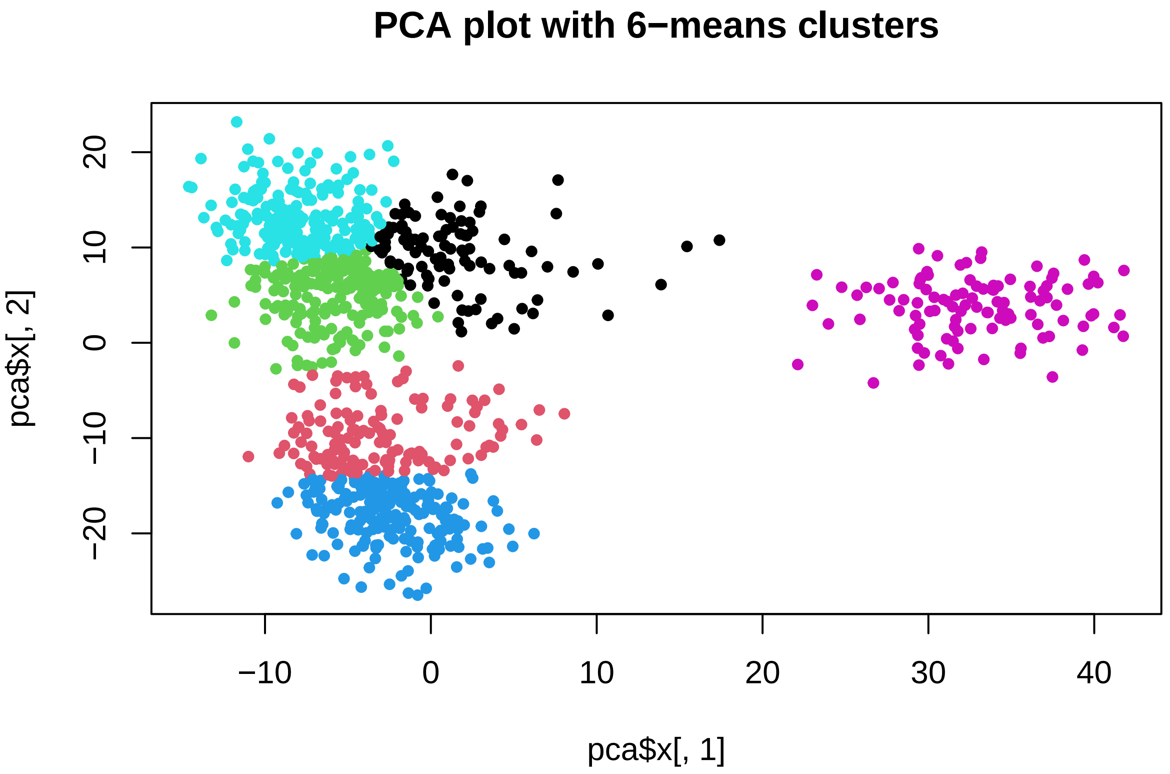
```
# Plot the first two principal components with continental information
plot(pca$x[,1], pca$x[,2], pch = 20, col = color_vector, main = "PCA plot with continental information"
```



```r
# Perform 7-means clustering on the PCA scores
kmeans_res2 <- kmeans(pca$x[,1:2], centers = 7)


# Plot the PCA plot with 7-means cluster assignments
plot(pca$x[,1], pca$x[,2], pch = 20, col = kmeans_res2$cluster, main = "PCA plot with 7-means clusters")
```



```
# Perform 6-means clustering on the PCA scores
kmeans_res3 <- kmeans(pca$x[,1:2], centers = 6)


# Plot the PCA plot with 7-means cluster assignments
plot(pca$x[,1], pca$x[,2], pch = 20, col = kmeans_res3$cluster, main = "PCA plot with 6-means clusters")
```



Problem 3

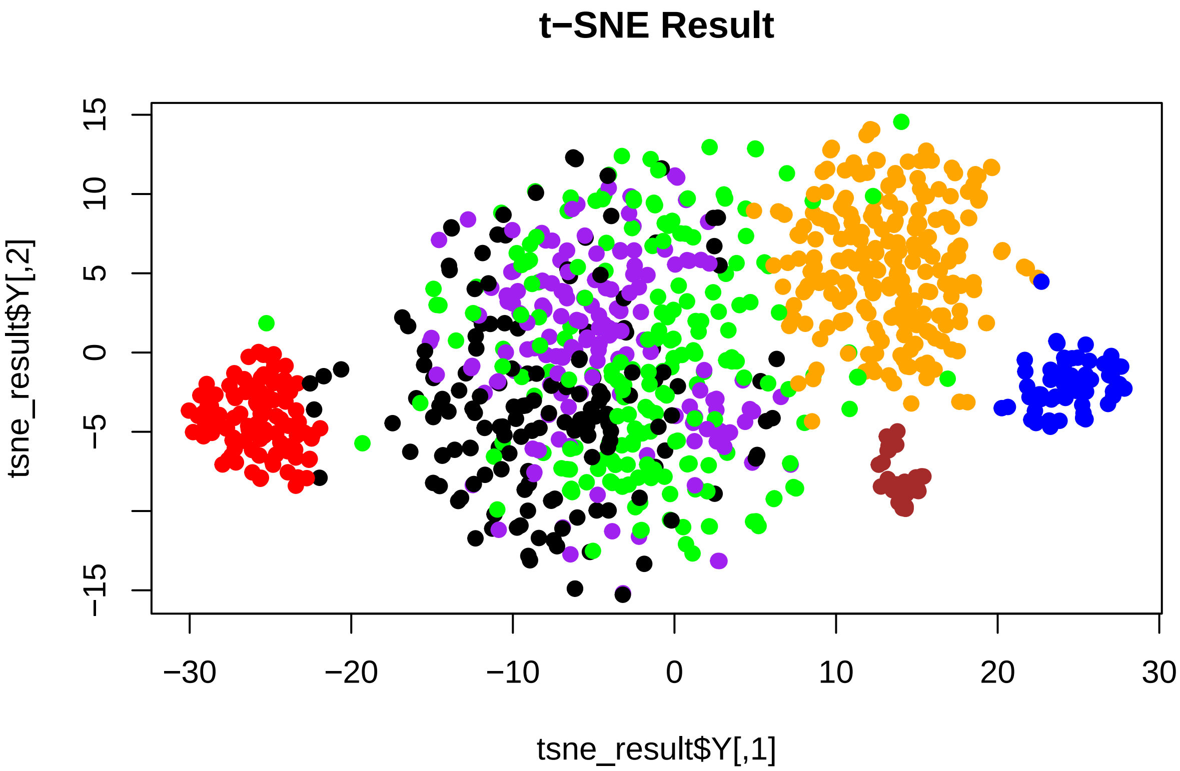
Comment on the cluster structures identified from the analysis.

Based on the pattern, the cluster can be clearly classified into 3 partitions. At the same time, the clustering trends of each continent are also very obvious. Therefore, I attempted to divide the clusters into 6 and 7 partitions respectively to compare the differences between k-means clustering on the PCA scores and clustering based on continental information. The clustering of 6 partitions and clustering based on continental information have some similarity in their plots. This indicates that the combination of many informative SNPs can provide a strong pattern of population clustering, giving us an indication of genetic differences between different continents.

Problem 4

Find necessary resources to study the emerging technique known as "t-distributed stochastic neighbor embedding", or, t-SNE, apply it to the data set. Compare the t-SNE results to the PCA results.

```
set.seed(1234)
tsne_result <- Rtsne(genotypes, dims=2,verbose= FALSE)
plot(tsne_result$Y, col=color_vector, pch=19, main="t-SNE Result")
```



This t-SNE plot shows some similarity with the PCA result, but now exactly the same, they both have good perfermance in classified Africa and East Asia but t-SNE does better on Oceania and America, none of them can classify Europe, Middle East and Central South Asia properly. The difference may because that t-SNE tends to be better at preserving local structure and capturing non-linear relationships between variables, while PCA is better at capturing global patterns and linear relationships.