Classification
Kyle Chan, Ryan Banafshay
For classification, we used a dataset (https://www.kaggle.com/datasets/purumalgi/music-genre-classification) that
contains information on 17,996 different songs.

The target variable for this dataset is Class, which represents genres of music.
0 = Folk,
1 = Alt,
2 = Blues,
3 = Bollywood,
4 = Country,
5 = Hip Hop,
6 = Indie,
7 = Instrumental,
8 = Metal,
9 = Pop.

Unlike linear regression models, the target variables in classification are qualitative. Logistic regression models predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, we will try to predict genre of music based on the independent variables of danceability, energy, and popularity. This is an important tool in the study of machine learning as logistic regression as it allows algorithms to predict a dependent data variable by analyzing the relationship between one or more existing independent variables. Naive Bayes models are based on conditional probability and assume strong, or naive, independence between attributes of data points.

Loading in the data
```{r}
MusicGenre <- read.csv("~/Desktop/MusicGenre.csv", header=TRUE)
str(MusicGenre)
```
```{r}
summary(MusicGenre)
```
Data Cleaning

Cleaning the data to only focus on the popularity, danceability, and energy columns. We will try to see how significant these elements are in determining the genre of music.
```{r}
MusicGenre <- MusicGenre[,c(3,4,5,17)]
MusicGenre$Class <- factor(MusicGenre$Class)
head(MusicGenre)
```
Dividing our data into training and testing sets
```{r}
set.seed(1234)
i <-sample(1:nrow(MusicGenre), .80*nrow(MusicGenre), replace=FALSE)
train <- MusicGenre[i,]
test <- MusicGenre[-i,]
```
Data exploration
Exploring the different variables we will use for our model
```{r}
summary(train$danceability)
```
```{r}
summary(train$energy)
```
```{r}
summary(train$Popularity)
```
```{r}
```

range(train$danceability)
```
```{r}
range(train$energy)
```

```{r}
par(mfrow=c(1,1))
plot(train$Popularity~train$danceability, xlab= "Popularity", ylab= "Danceability",
pch=25, bg=c("aquamarine1"))
abline(lm(train$Popularity~train$danceability), col = "red")
```

```{r}
par(mfrow=c(1,2))
dance_den <- density(train$danceability, na.rm = TRUE)
plot(dance_den, main = "Danceability Density", xlab = "Danceability")
polygon(dance_den, col ="wheat")
Popularity_den <- density(train$Popularity, na.rm = TRUE)
plot(Popularity_den, main = "Popularity Density", xlab = "Popularity")
polygon(Popularity_den, col ="slategrey")
```

Logistic Regression Model
```{r}
glm1 <- glm(Class~., data=train, family="binomial")
summary(glm1)
```
From this logistic regression model, we can determine that the popularity of the track has a significant impact the genre of music. This makes sense in some regards, if a music is classified under the genre of "pop" then we can probably guess it will be much more popular than a song under the genre of "folk". Energy seems to also be contribute to the genre of music, while on the other hand danceability seems to have little to no impact.

Logistic regression model for just energy
```{r}
glm2 <- glm(Class~energy, data=train, family="binomial")
summary(glm2)
```

Naïve Bayes
```{r}
library(e1071)
nb1 <- naiveBayes(Class~., data=train)
nb1
```

Evaluate Naïve Bayes
```{r}
p2 <- predict(nb1, newdata=test, type="class")
table(p2, test$Class)
```

```{r}
mean(p2==test$Class)
```
Based on this mean result, it is hard to rely too much on the naive bayes analysis as it shows to be not as accurate as the logistic model for this data.

Strengths and weaknesses of Logistic vs Naive Bayes
Both logistic regression and Naive Bayes have similarities, as they are both linear classifiers and are both used for classification. A strength of logistic regression is that it is typically low bias, meaning it incorporates fewer assumptions about the target
```

function. Lower bias models tend to closely match the training data set. But on the flip side they tend to have a higher variance. This is the opposite for Naive Bayes models, as they tend to have higher bias but lower variance. So if the data set follows the bias then Naive Bayes will be a better classifier. Another benefit of Naive Bayes is that results are easier to predict with less variables and less data. Logisitic regression is better for multinomial classification problems, such as the one we did in this assignment.

Benefits and drawbacks
As we used a large dataset with multinomial classifications (more than two possible discrete outcomes rather than the binary 0 and 1), I felt that the results from logistic regression was far more beneficial for drawing conclusions on the data. The classification methods here are incredibely general though, and I felt some of the variables were a bit arbitrary. I don't understand how the model determined that energy is a determining factor of classification, but not dancability. This very well could be due to how the data was collected.