

Question 3

(a)

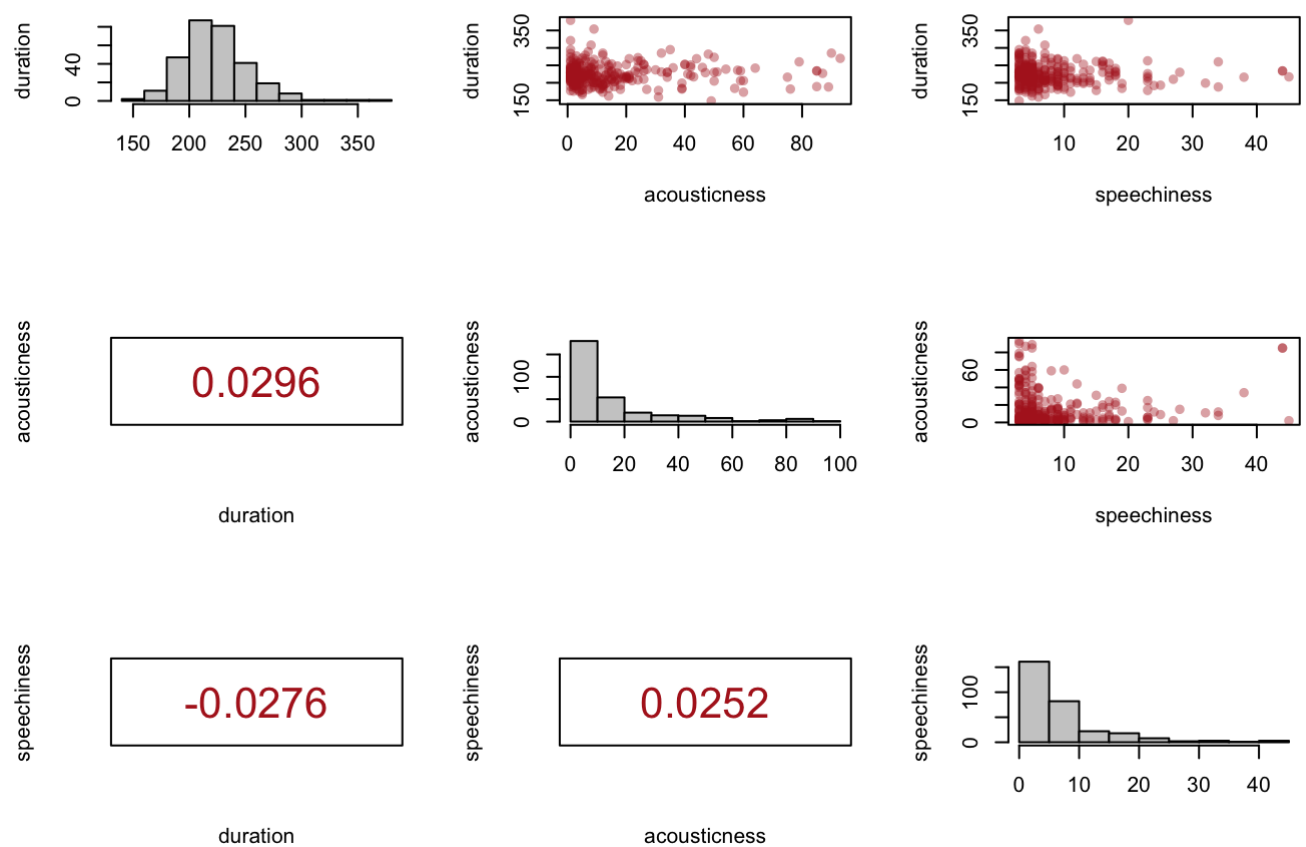
```
# Using Spotify.csv data
# Using columns 11-13 for duration, acousticness, speechiness
data <- read.csv("spotify.csv", header = TRUE)
df <- data[,11:13]
# Plot Correlation coefficients between x,y
plot_corr <- function(x, y, x_lab, y_lab) {
  plot(0, 0,
       main = "",
       xlab = x_lab,
       ylab = y_lab,
       xaxt = "n",
       yaxt = "n",
       col = "white")
  text(x = 0, y = 0,
       labels = paste(round(cor(x, y), 4)),
       col = "firebrick",
       cex = 2)
}

# Plot Scatter Plot between x,y variates
plot_scatt <- function(x,y,x_lab,y_lab) {
  plot(x, y,
       main = "",
       xlab = x_lab,
       ylab = y_lab,
       pch = 16,
       col = adjustcolor(col = "firebrick", alpha.f = 0.4))
}

# Plot Histogram between x,y variates
plot_hist <- function (x,lab) {
  hist(x,
       main = "",
       xlab = "",
       ylab = lab,
       col = "gray80")
}

# Matrix Plot
matrix.plot <- function(df) {
  m <- ncol(df)
  n <- nrow(df)
  # m*m columns of plots
  par(mfcol = c(m, m))
  for (i in 1:m) {
    for (j in 1:m) {
      if (i==j) {
        plot_hist(df[,i],colnames(df)[i])
      }
      else if (i>j) {
        plot_scatt(df[,i],df[,j],colnames(df)[i],colnames(df)[j])
      }
      else {
        plot_corr(df[,i], df[,j], colnames(df)[i], colnames(df)[j])
      }
    }
  }
}

matrix.plot(df)
```



(b)

From below, we can clearly observe that **loudness**, **energy** and **danceability** are most strongly correlated with valence.

Loudness: Positive correlation with valence

Energy: Positive correlation with valence

Danceability: Positive correlation with valence



(c)

Top 5 songs with Highest Valence

```
order_valence <- order(data$valence, decreasing=TRUE)
# Table
kable_styling(kable(data[order_valence, ][1:5, c("title", "artist", "year","valence")]))
```

	title	artist	year	valence
149	Mmm Yeah (feat. Pitbull)	Austin Mahone	2014	98
245	There's Nothing Holdin' Me Back	Shawn Mendes	2018	97
124	Happy - From "Despicable Me 2"	Pharrell Williams	2014	96
284	Sucker	Jonas Brothers	2019	95
126	Shake It Off	Taylor Swift	2014	94

(d)

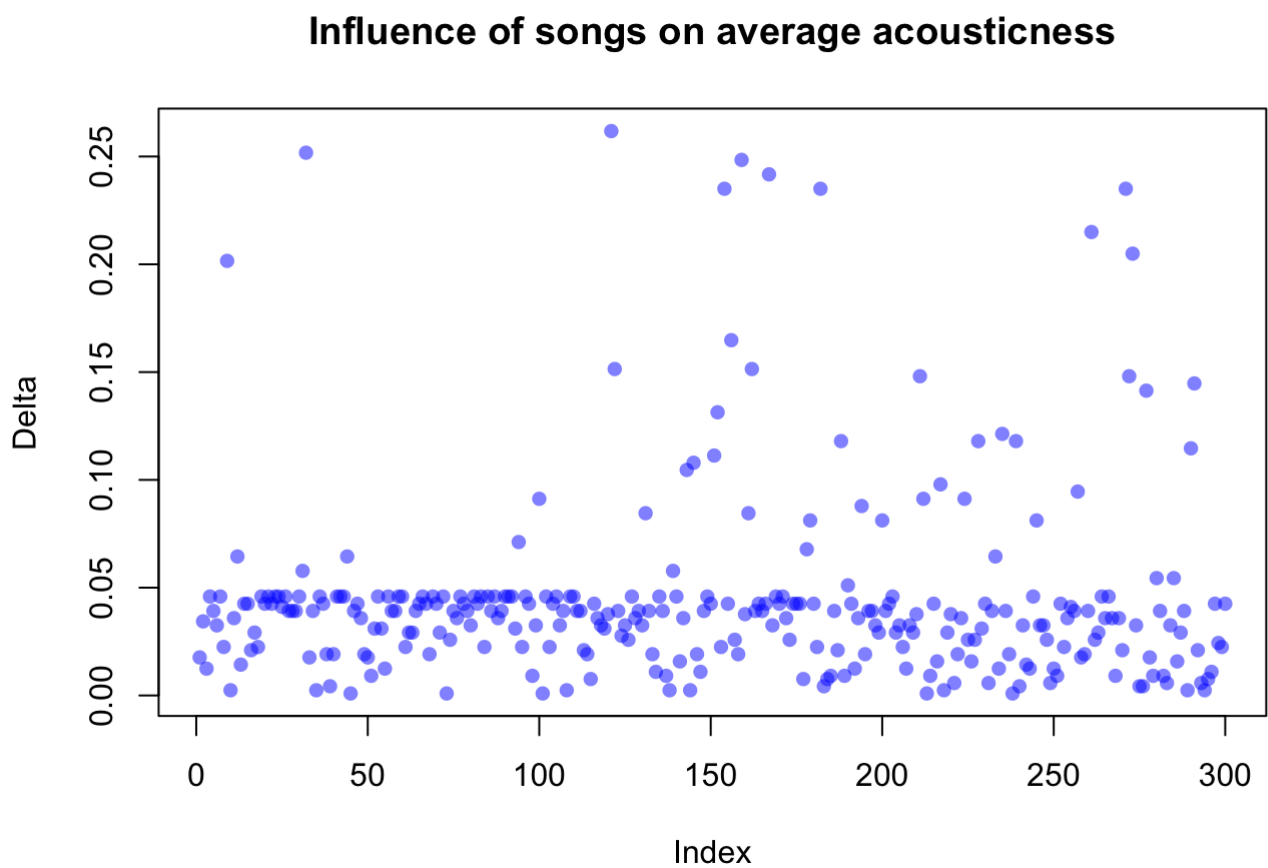
Determining which song has the highest and lowest danceability

```
highest_danceability <- which.max(data$danceability)
lowest_danceability <- which.min(data$danceability)
# Table
kable_styling(kable(data[c(lowest_danceability,highest_danceability), ][c("title","artist","year","danceability")]))
```

	title	artist	year	danceability
160	Love Me Like You Do - From "Fifty Shades Of Grey"	Ellie Goulding	2015	26
207	Me Too	Meghan Trainor	2016	93

(e)

```
par(mfcol = c(1,1))
diff <- abs(data$acousticness - mean(data$acousticness))
Delta_Ac <- diff/(length(data$acousticness)-1)
plot(Delta_Ac,
     main = "Influence of songs on average acousticness",
     ylab = "Delta",
     pch = 16,
     col = adjustcolor("blue", alpha.f=0.5))
```



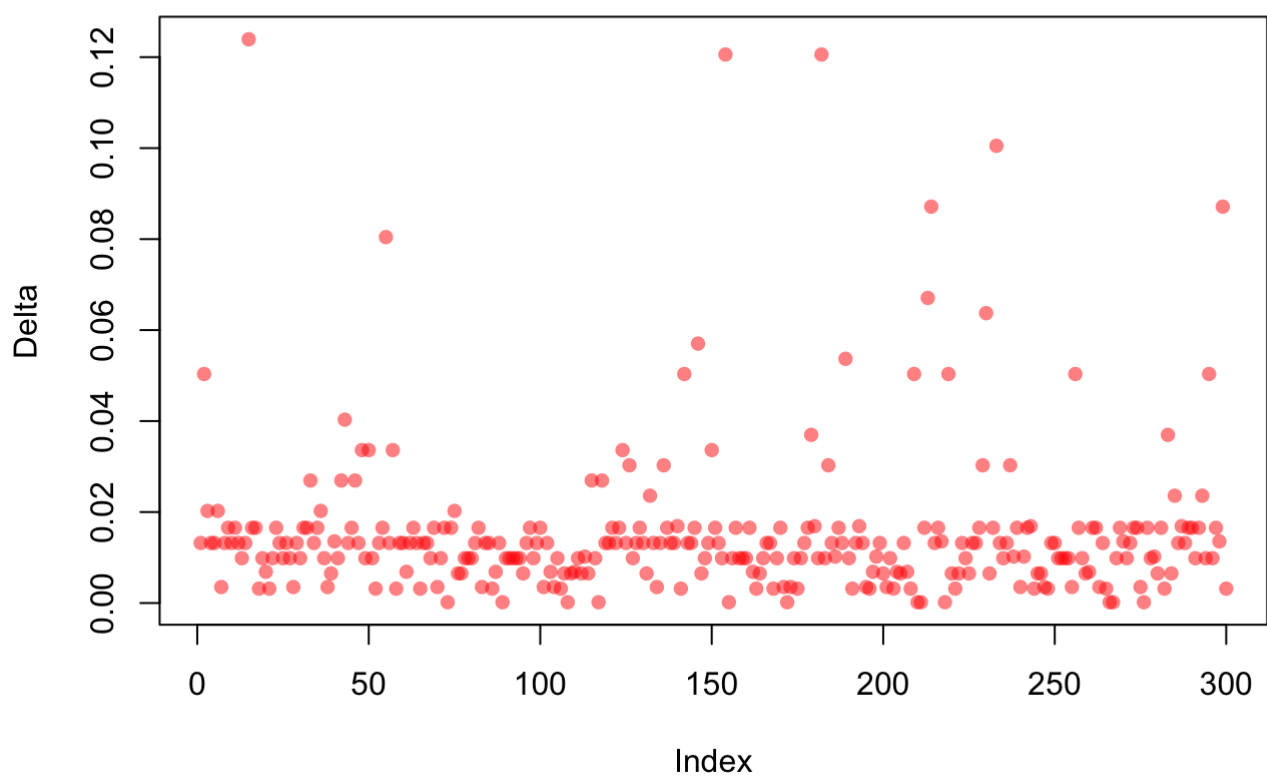
```
# Maximum influence on acousticness
kable_styling(kable(data[c(which.max(Delta_Ac), which.max(data$acousticness)), c("title",
, "artist", "acousticness")]))
```

	title	artist	acousticness
121	All of Me	John Legend	93
121.1	All of Me	John Legend	93

(f)

```
par(mfcol = c(1,1))
diff <- abs(data$speechiness - mean(data$speechiness))
Delta_Sp <- diff/(length(data$speechiness)-1)
plot(Delta_Sp,
     main = "Influence of songs on average speechiness",
     ylab = "Delta",
     pch = 16,
     col = adjustcolor("red", alpha.f=0.5))
```

Influence of songs on average speechiness



```
# Maximum influence on speechiness
kable_styling(kable(data[c(which.max(Delta_Sp), which.max(data$speechiness)), c("title",
"artist", "speechiness")]))
```

	title	artist	speechiness
15	Like A G6	Far East Movement	45
15.1	Like A G6	Far East Movement	45

(g)

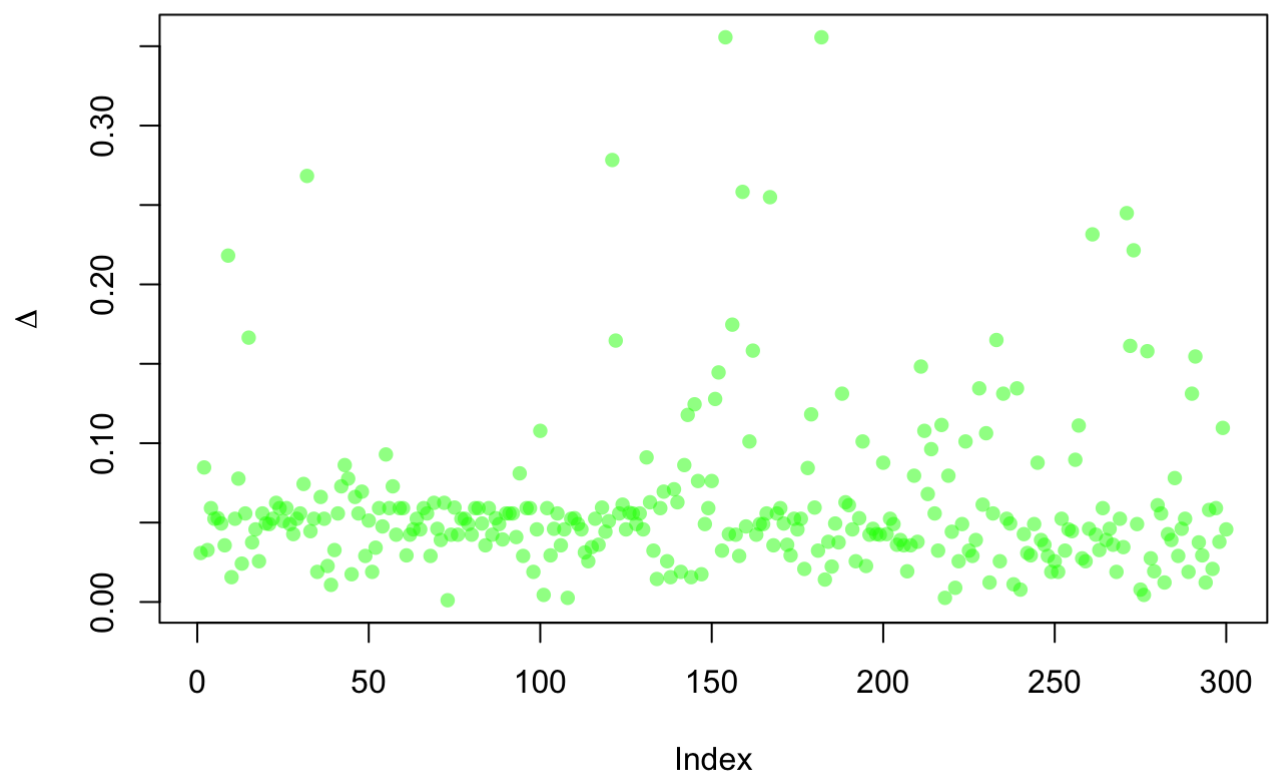
```
par(mfcol = c(1,1))

diff_Sp <- abs(data$speechiness - mean(data$speechiness))
Delta_Sp <- diff_Sp/(length(data$speechiness)-1)

diff <- abs(data$acousticness - mean(data$acousticness))
Delta_Ac <- diff/(length(data$acousticness)-1)

Delta <- Delta_Sp + Delta_Ac
plot(Delta,
      main="Adding 2 measures of influence",
      ylab = bquote(Delta),
      pch = 16,
      col = adjustcolor("green", alpha.f=0.5))
```

Adding 2 measures of influence



Song that has the most influence when the above measures of influence are added:

```
# Maximum Influence
kable_styling(kable(data[c(which.max(Delta), which.max(data$acousticness+data$speechiness)),c("title", "artist", "acousticness", "speechiness")]))
```

	title	artist	acousticness	speechiness
154	Love Yourself	Justin Bieber	85	44
154.1	Love Yourself	Justin Bieber	85	44

(h)

```
# Artists that have appeared in the Billboard Top 30 Four Times
top_artists <- names(which(table(data$artist) == 4))
num_appears <- rep(0, length(top_artists))
artist_pop <- rep(0, length(top_artists))
for (i in 1:length(top_artists)) {
  artist_match <- data$artist == top_artists[i]
  song_pop <- data[artist_match, ]$valence
  num_appears[i] <- length(song_pop)
  artist_pop[i] <- round(mean(song_pop))
}
```