HW1\_Data Visualization

Lina Lee

February 10, 2019

library(ggplot2)  
music <- read.csv("https://kmaurer.github.io/documents/data/shuffled\_playlist.csv")

# 1.

# a. How many songs are in the data frame? A:178

str(music)

## 'data.frame': 178 obs. of 13 variables:  
## $ artist : Factor w/ 14 levels "Adele","Childish+Gambino",..: 12 1 11 2 6 8 12 5 7 4 ...  
## $ album : Factor w/ 14 levels "\"Awaken, My Love!\"",..: 6 4 12 1 2 13 6 8 11 10 ...  
## $ genre : Factor w/ 4 levels "classic rock",..: 4 3 1 2 3 2 4 1 1 2 ...  
## $ name : Factor w/ 178 levels "5 out of 6","A Better Place, a Better Time",..: 4 167 132 12 97 175 43 101 105 52 ...  
## $ danceability : num 0.609 0.596 0.592 0.529 0.68 0.501 0.648 0.649 0.38 0.612 ...  
## $ energy : num 0.918 0.838 0.0196 0.403 0.852 0.646 0.896 0.425 0.484 0.738 ...  
## $ loudness : num -6 -6.52 -33.35 -7.1 -4.35 ...  
## $ speechiness : num 0.0894 0.0704 0.0358 0.0349 0.0349 0.231 0.0825 0.0267 0.0455 0.303 ...  
## $ acousticness : num 0.0087 0.0189 0.362 0.123 0.117 0.482 0.0426 0.378 0.395 0.055 ...  
## $ instrumentalness: num 0.00 1.54e-05 8.54e-01 4.96e-01 0.00 1.43e-05 2.15e-06 1.30e-03 1.12e-02 0.00 ...  
## $ liveness : num 0.294 0.108 0.108 0.0647 0.0866 0.274 0.0675 0.108 0.687 0.125 ...  
## $ valence : num 0.541 0.479 0.0322 0.723 0.852 0.648 0.794 0.651 0.165 0.564 ...  
## $ tempo : num 90 95 120 200 102 ...

# b. hat song has the highest “valence” score?

# A: For You Blue - Remastered

music[which.max(music$valence),]

## artist album genre  
## 159 The+Beatles Let It Be (Remastered) classic rock  
## name danceability energy loudness speechiness  
## 159 For You Blue - Remastered 0.88 0.556 -10.773 0.0855  
## acousticness instrumentalness liveness valence tempo  
## 159 0.24 0.0483 0.24 0.958 128.542

# c. What is the average energy for hip hop songs? A:0.6567

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.1 --

## v tibble 1.4.2 v purrr 0.2.5  
## v tidyr 0.8.2 v dplyr 0.7.7  
## v readr 1.1.1 v stringr 1.3.1  
## v tibble 1.4.2 v forcats 0.3.0

## Warning: package 'tidyr' was built under R version 3.5.2

## -- Conflicts -------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

music %>%   
 group\_by(genre) %>%   
 summarise(average = mean(energy))

## # A tibble: 4 x 2  
## genre average  
## <fct> <dbl>  
## 1 classic rock 0.533  
## 2 hip hop 0.657  
## 3 pop 0.596  
## 4 ska 0.886

# d. What percentage of songs the data are by the artist Dessa? A:6.2%

myTable <- data.frame( table(music$artist))   
myTable

## Var1 Freq  
## 1 Adele 11  
## 2 Childish+Gambino 11  
## 3 Dessa 11  
## 4 Doomtree 12  
## 5 Eagles 9  
## 6 Ed+Sheeran 16  
## 7 Fleetwood+Mac 20  
## 8 Kendrick+Lamar 14  
## 9 Less+Than+Jake 14  
## 10 No+Doubt 13  
## 11 Pink Floyd 10  
## 12 Streetlight+Manifesto 12  
## 13 Taylor+Swift 13  
## 14 The+Beatles 12

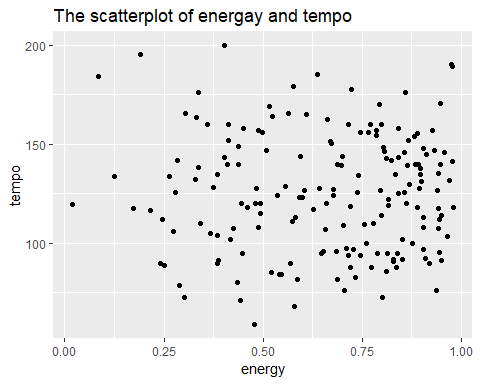
myTable$Prop<-prop.table( myTable$Freq )   
myTable

## Var1 Freq Prop  
## 1 Adele 11 0.06179775  
## 2 Childish+Gambino 11 0.06179775  
## 3 Dessa 11 0.06179775  
## 4 Doomtree 12 0.06741573  
## 5 Eagles 9 0.05056180  
## 6 Ed+Sheeran 16 0.08988764  
## 7 Fleetwood+Mac 20 0.11235955  
## 8 Kendrick+Lamar 14 0.07865169  
## 9 Less+Than+Jake 14 0.07865169  
## 10 No+Doubt 13 0.07303371  
## 11 Pink Floyd 10 0.05617978  
## 12 Streetlight+Manifesto 12 0.06741573  
## 13 Taylor+Swift 13 0.07303371  
## 14 The+Beatles 12 0.06741573

# 

# 2. Create a scatterplot of the relationship between the energy and tempo of songs.Describe in a few sentences what this display tells us about the relationship between the variables.

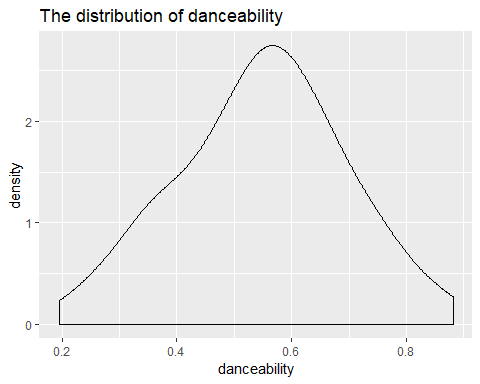
ggplot()+geom\_point(aes(x=energy,y=tempo), data=music)+labs(x="energy",y="tempo",   
 title="The scatterplot of energay and tempo")



# A: The scatterplot does not show the explicit relationship between tempo and energy. Observations spread randomly across all range of tempo (0 to 200) and energy (0 to 1).

# 3. Create and describe a plot that tells us about the distribution of danceability in this playlist.

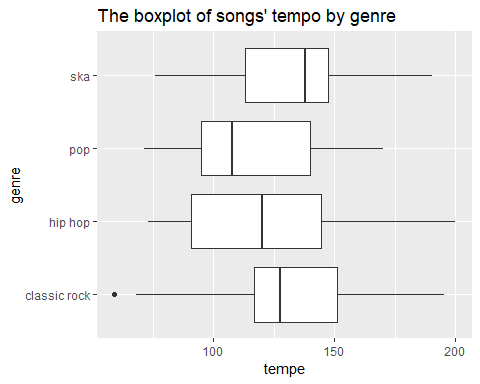
ggplot()+geom\_density(aes(x=danceability),adjust=1,data=music)+ labs(x="danceability",   
 title="The distribution of danceability")



# A: The density(density of frequency of songs) on danceability between 0.5 and 0.6 is the higest, and the density decrease as the value of danceability is farther from the range (0.5,0,6). Therefore, the distribution of danceability is smoothed triangular shape.

# 4. How do the tempo values differ by genre? Use a side-by-side style of plot to help tell a story of how they differ. Describe in a few sentences what this display tells us about the relationship between the variables. Why did you choose this type of display?

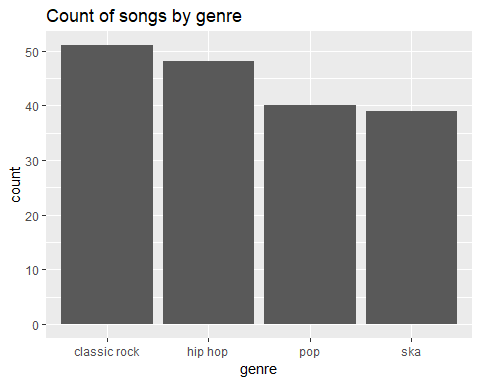
ggplot()+  
 geom\_boxplot(aes(x=genre, y=tempo), data=music) +  
 coord\_flip()+  
 labs(x="genre",   
 y="tempe",  
 title="The boxplot of songs' tempo by genre")



# A: I used the side by side boxplot to compare the distribution of tempo values by genre. The range of the distribution of classic rock’s tempo is on the highest, ska has the second highest range. Moreover, these two distributions has smaller(narrower) range than pop and hiphop. Hiphop has the widest range. Median of ska is the largest among all of them, classic rock has the second largest median, and hiphop has third largest median. Hiphop has the smalllest minimum value.

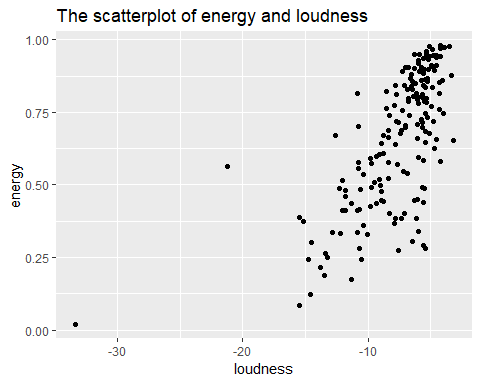
# 5.Create an additional plot of your choosing that tells an interesting story about the data. Describe this plot in a few sentences.

ggplot()+geom\_bar(aes(x=genre),data=music,stat="count")+ labs(x="genre",   
 y="count",  
 title="Count of songs by genre")



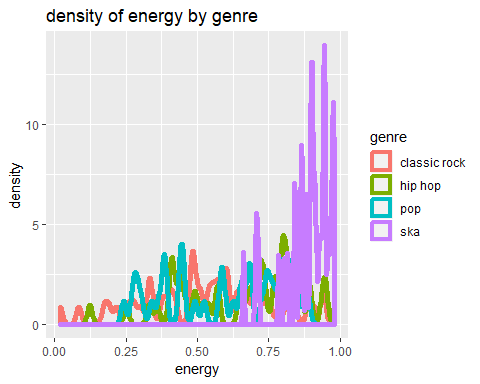
# classic rock has the largest number of songs. The genre which has the second largest number of the songs is hiphop, and ska is the one having the smallest number of the songs.

ggplot()+geom\_point(aes(x=loudness,y=energy), data=music)+labs(x="loudness",   
 y="energy",  
 title="The scatterplot of energy and loudness")



# The scatter plot implies the linear relationship between energy and loudness. That is, more loud a song is, more energy a song has( This is not causal relationship. This implies only correlation).

ggplot()+  
 geom\_density(aes(x=energy,color=genre),size=2,  
 adjust=.1,data=music)+labs(x="energy",   
 y="density",  
 title="density of energy by genre")



# Most of the songs of ska have a tendency to have higher energy than the others. The plot does not show explicit differences in the distribution of density on energy by genre among the other three.

Reference:

I referred the code for number1 from

<https://stackoverflow.com/questions/19449615/how-to-extract-the-row-with-min-or-max-values>

<http://r.789695.n4.nabble.com/Frequencies-proportions-amp-cumulative-proportions-td906623.html>