511 Final project

Yilin Yang

2022-12-05

Load the required dataset

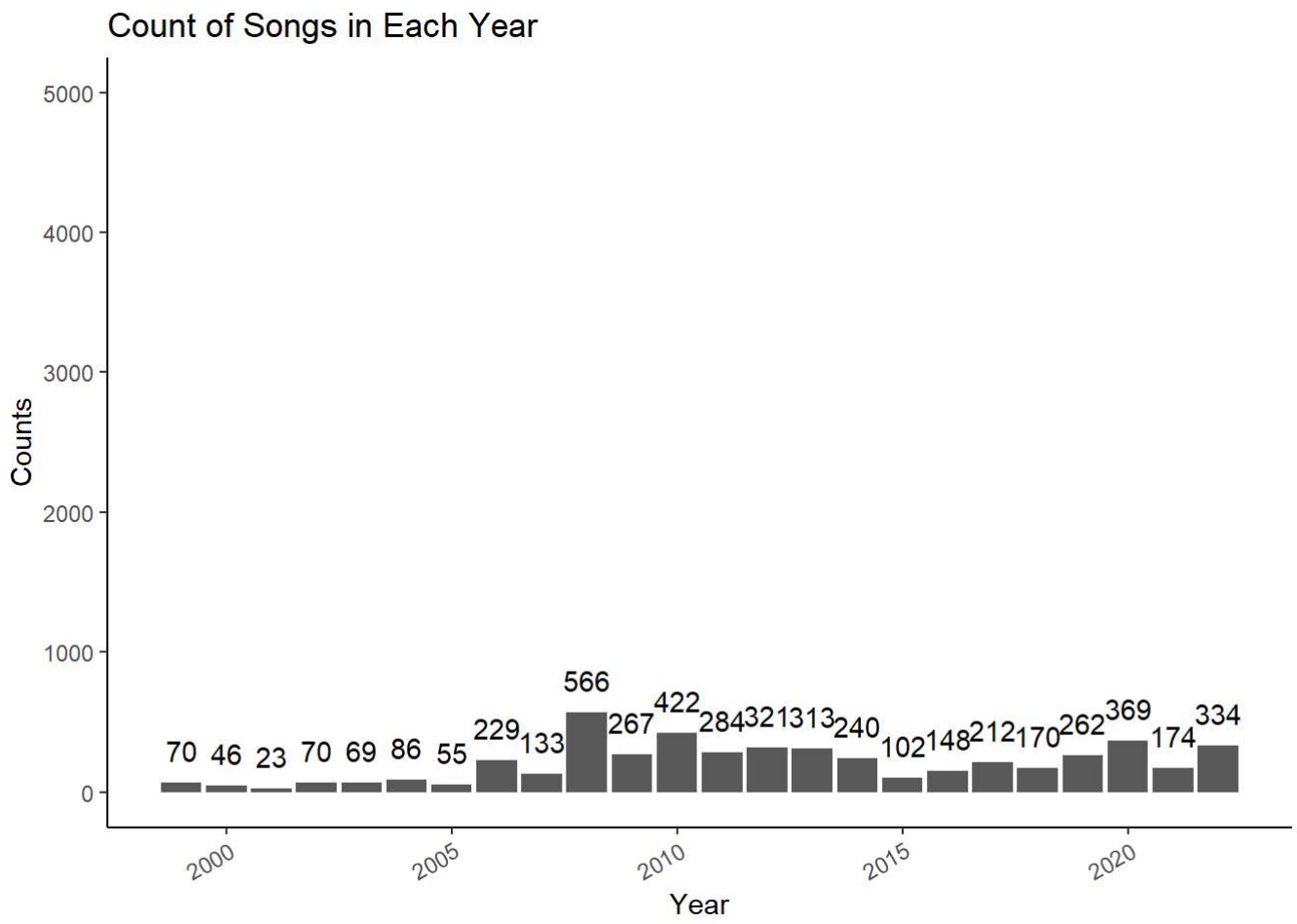
```
artists <- read.csv("./Tracks_Artists.csv")
head(artists)
```

```
##   X  artist_name Valence danceability energy loudness speechiness acousticness
## 1 1 Taylor Swift  0.0984        0.735  0.444  -10.519      0.0684       0.2040
## 2 2 Taylor Swift  0.0382        0.658  0.378   -8.300      0.0379       0.0593
## 3 3 Taylor Swift  0.5190        0.638  0.634   -6.582      0.0457       0.1330
## 4 4 Taylor Swift  0.1540        0.659  0.323  -13.425      0.0436       0.7350
## 5 5 Taylor Swift  0.3760        0.694  0.380  -10.307      0.0614       0.4160
## 6 6 Taylor Swift  0.2300        0.636  0.377  -11.721      0.0708       0.7100
##   liveness   tempo                            track_name
## 1   0.1700  97.038                         Lavender Haze
## 2   0.0976 108.034                                Maroon
## 3   0.1520  96.953                             Anti-Hero
## 4   0.1160 110.007 Snow On The Beach (feat. Lana Del Rey)
## 5   0.1260 120.044               You're On Your Own, Kid
## 6   0.1150 139.966                         Midnight Rain
##               album_name album_release_year
## 1 Midnights (3am Edition)              2022
## 2 Midnights (3am Edition)              2022
## 3 Midnights (3am Edition)              2022
## 4 Midnights (3am Edition)              2022
## 5 Midnights (3am Edition)              2022
## 6 Midnights (3am Edition)              2022
```

```
summary(artists$album_release_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1999    2008    2012    2013    2018    2022
```

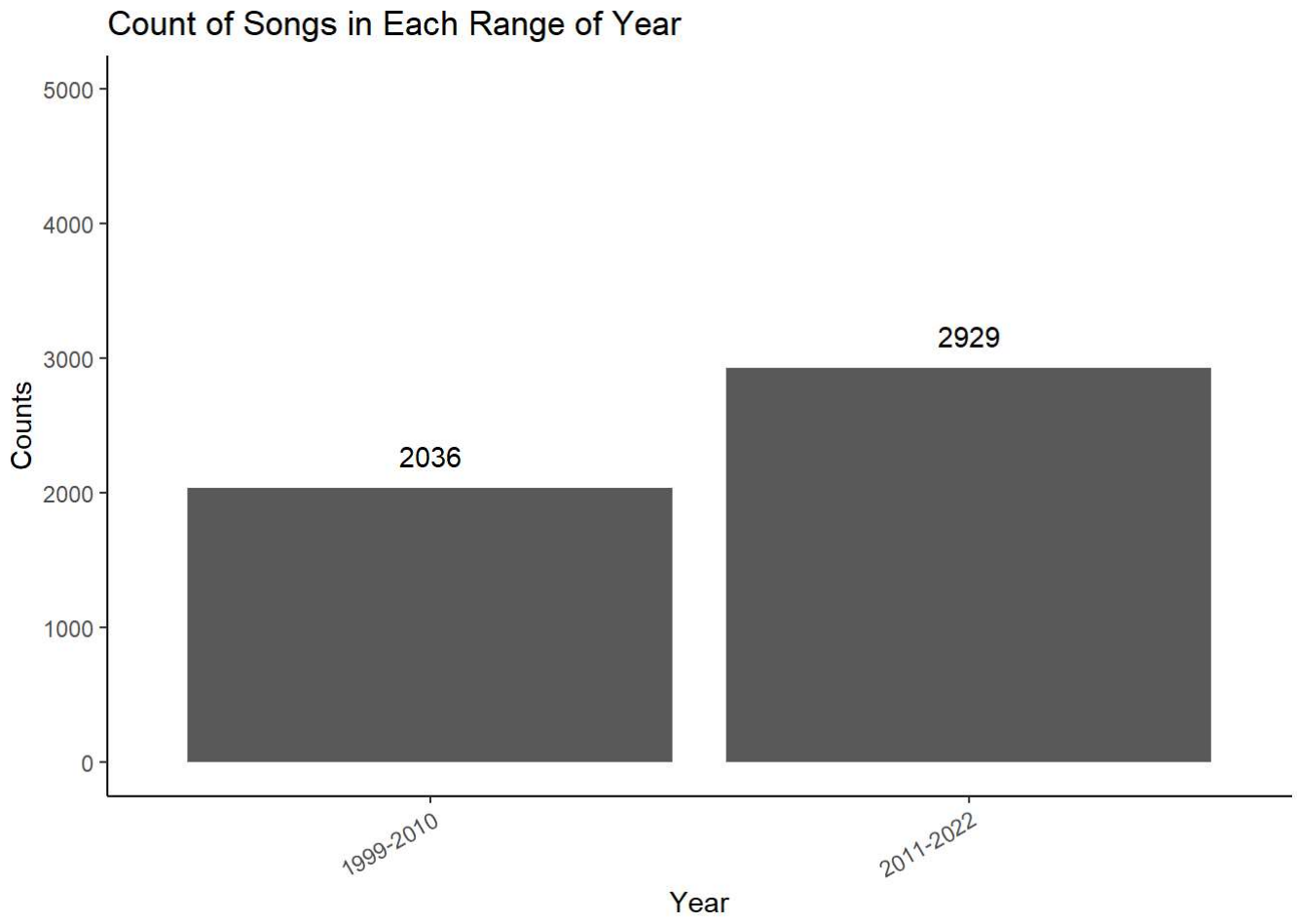
```
library(ggplot2)
ggplot(artists, aes(x=album_release_year)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=-1)+
  ylim(0,5000)+
  theme_classic()+
  theme(axis.text.x=element_text(angle=30,hjust=1))+
  labs(title="Count of Songs in Each Year",
       x ="Year", y = "Counts")
```




```
artists$year_range <- ifelse(artists$album_release_year < 2011, "1999-2010","2011-2022")
head(artists)
```

```
##   X  artist_name Valence danceability energy loudness speechiness acousticness
## 1 1 Taylor Swift  0.0984        0.735  0.444  -10.519      0.0684       0.2040
## 2 2 Taylor Swift  0.0382        0.658  0.378   -8.300      0.0379       0.0593
## 3 3 Taylor Swift  0.5190        0.638  0.634   -6.582      0.0457       0.1330
## 4 4 Taylor Swift  0.1540        0.659  0.323  -13.425      0.0436       0.7350
## 5 5 Taylor Swift  0.3760        0.694  0.380  -10.307      0.0614       0.4160
## 6 6 Taylor Swift  0.2300        0.636  0.377  -11.721      0.0708       0.7100
##   liveness   tempo                              track_name
## 1   0.1700  97.038                           Lavender Haze
## 2   0.0976 108.034                                 Maroon
## 3   0.1520  96.953                              Anti-Hero
## 4   0.1160 110.007 Snow On The Beach (feat. Lana Del Rey)
## 5   0.1260 120.044                You're On Your Own, Kid
## 6   0.1150 139.966                           Midnight Rain
##                album_name album_release_year year_range
## 1 Midnights (3am Edition)               2022  2011-2022
## 2 Midnights (3am Edition)               2022  2011-2022
## 3 Midnights (3am Edition)               2022  2011-2022
## 4 Midnights (3am Edition)               2022  2011-2022
## 5 Midnights (3am Edition)               2022  2011-2022
## 6 Midnights (3am Edition)               2022  2011-2022
```

```
library(ggplot2)
ggplot(artists, aes(x=year_range)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=-1)+
  ylim(0,5000)+
  theme_classic()+
  theme(axis.text.x=element_text(angle=30,hjust=1))+
  labs(title="Count of Songs in Each Range of Year",
       x ="Year", y = "Counts")
```



Hypothesis 2:

Null Hypothesis: I will make null hypothesis as the average Valence of songs in 1999-2010 is higher than songs in 2011-2011.

Alternative Hypothesis: the average Valence of songs in 1999-2010 is lower than songs in 2011-2011.

```
# group the dataset by Yearrange
Year2010 <- subset(artists, artists$year_range == "1999-2010",select = c("Valence"))
Year2022 <- subset(artists, artists$year_range == "2011-2022",select = c("Valence") )
# t.test
t.test(Year2010,Year2022,alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Year2010 and Year2022
## t = 9.9977, df = 4084.5, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.07579864
## sample estimates:
## mean of x mean of y
## 0.4736412 0.4085535
```

```
set.seed(1)
n1 = length(Year2010)
n2 = length(Year2022)
N <- 10000
diff_mean <- numeric(N)

for (i in 1:N)
{
Year2010.sample <- sample(Year2010$Valence, n1, replace = TRUE)
Year2022.sample <- sample(Year2022$Valence, n2, replace = TRUE)
diff_mean[i] <- mean(Year2010.sample) - mean(Year2022.sample)
}

mean(diff_mean)
```

```
## [1] 0.06635376
```

```
quantile(diff_mean, c(.025, .975))
```

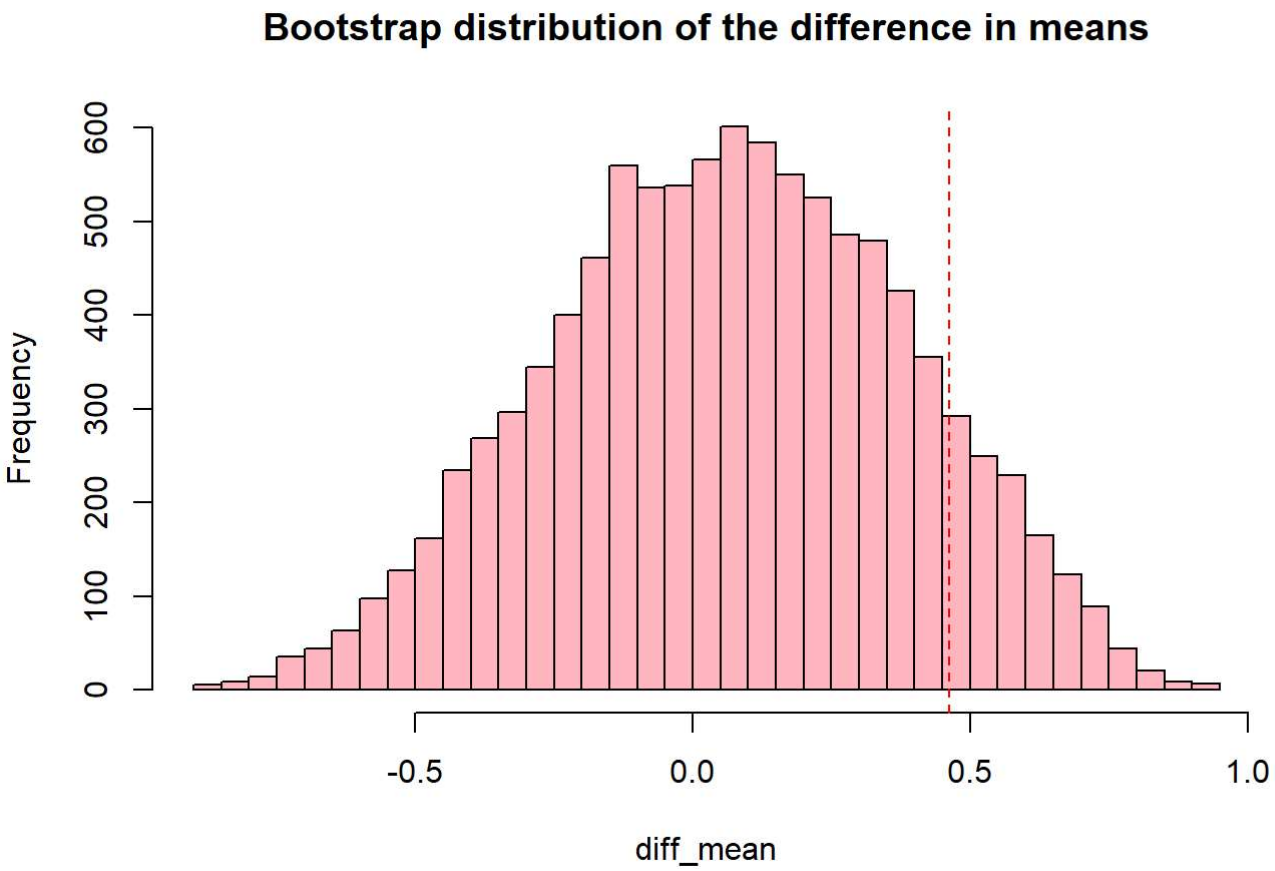
```
##   2.5%  97.5%
## -0.557  0.661
```

```
mydiff = function(mydf){
  index1 = artists$year_range == "1999-2010"
  index2 = artists$year_range == "2011-2022"
  return(mean(artists$Valence[index1]) - mean(artists$Valence[index2]))
}

mydiff(genre.clean) #actual mean difference from the original sample
```

```
## [1] 0.06508774
```

```
hist(diff_mean,breaks=50,main = "Bootstrap distribution of the difference in means",col = 'light
pink')
abline(v = mean(Year2010.sample) - mean(Year2022.sample), col = "red", lty = 2)
```


Bootstrap distribution of the difference in means
Frequency
0
100
200
300
400
500
600
-0.5
0.0
0.5
1.0
diff_mean