Music Genre Sentiment Analysis–R Code

Julie Bazalewski

06/20/2020

Overview

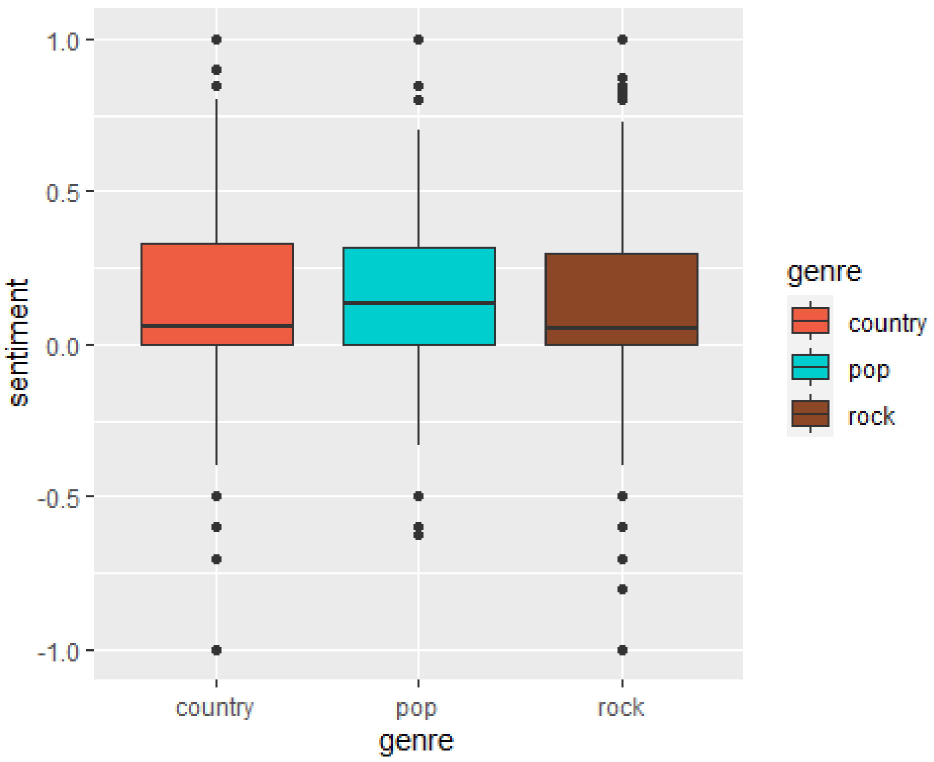
Analyze data from #countrymusic, #rockmusic, and #popmusic Twitter tags to determine if genre affects tweet sentiment.

Import .csv file

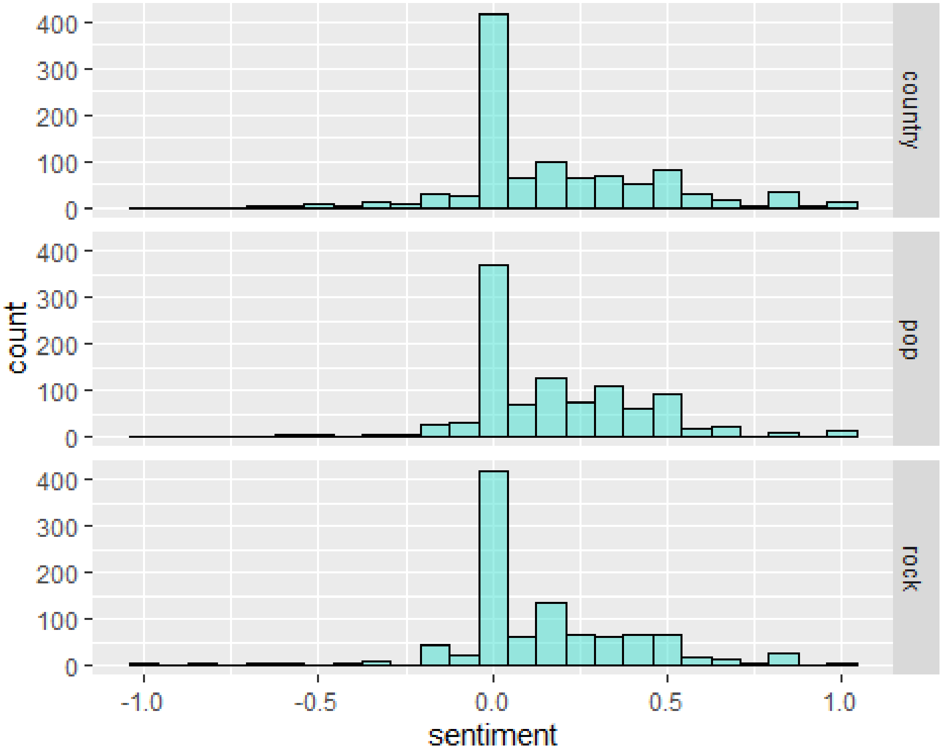
```
tweet_df = read_csv("tweets_3147.csv", n_max = 3147)

## Parsed with column specification:
## cols(
##   text = col_character(),
##   user = col_character(),
##   location = col_character(),
##   genre = col_character(),
##   sentiment = col_double()
## )

summary(tweet_df)

##      text               user             location            genre          
##  Length:3115        Length:3115        Length:3115        Length:3115       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##    sentiment       
##  Min.   :-1.00000  
##  1st Qu.: 0.00000  
##  Median : 0.08333  
##  Mean   : 0.16205  
##  3rd Qu.: 0.31818  
##  Max.   : 1.00000  
```

Group by Genre and create boxplot

```
gf_boxplot(sentiment ~ genre, data = tweet_df, fill=~genre)%>%
  gf_refine(scale_fill_manual(values = c("tomato2","cyan3", "sienna4")))
```



```
#look at histograms
gf_histogram(~sentiment, data=tweet_df, fill="turquoise", color="black") %>%
gf_facet_grid(genre~.)
```



## Calculate mean sentiment of each genre

```
genreSentiment <- tweet_df %>%
  group_by(genre) %>%
  summarise(genre.mean = mean(sentiment)) %>%
  arrange(desc(genre.mean))

## `summarise()` ungrouping output (override with `.groups` argument)

genreSentiment

## # A tibble: 3 x 2
##   genre   genre.mean
##   <chr>        <dbl>
## 1 pop          0.176
## 2 country      0.166
## 3 rock         0.145
```

Perform Chi Squared Test of Independence using sentiment as a categorial variable (negative, neutral, or positive)

H0: There is no relationship between music genre and sentiment type H1: There is a relationship between music genre and sentiment type

```
#turn sentiment into a categorical variable
tweet_df <- tweet_df %>%
  mutate(sentiment_type = case_when(sentiment<0 ~"Negative", sentiment > 0
~"Positive", TRUE ~"Netural"))

#perform chi squared test
chisq.test(tweet_df$genre, tweet_df$sentiment_type)

##
##  Pearson's Chi-squared test
##
## data:  tweet_df$genre and tweet_df$sentiment_type
## X-squared = 13.008, df = 4, p-value = 0.01124
```

Conclusion:

At the alpha = 0.05 significance level there is enough evidence to claim that there is an association between genre and sentiment.

Perform Kruskal-Wallis test as an alternative to ANOVA to check sentiment between groups with sentiment as a non-normal, numerical variable

H0: There is no relationship between music genre and sentiment H1: There is a relationship between music genre and sentiment

```
kruskal.test(sentiment ~ genre, data = tweet_df)

##
##  Kruskal-Wallis rank sum test
##
## data:  sentiment by genre
## Kruskal-Wallis chi-squared = 9.229, df = 2, p-value = 0.009907
```

Conclusion:

At the alpha = 0.01 significance level there is enough evidence to claim that there is an association between genre and sentiment.

Perform T-Tests, the data is not normally distributed but the sample sizes are large. Use numerical sentiment data rather than categorical.

```
country_sentiment = tweet_df$sentiment[which(tweet_df$genre == 'country')]
rock_sentiment = tweet_df$sentiment[which(tweet_df$genre == 'rock')]
pop_sentiment = tweet_df$sentiment[which(tweet_df$genre == 'pop')]

t.test(country_sentiment,rock_sentiment, alternative="less")

##
##  Welch Two Sample t-test
##
## data:  country_sentiment and rock_sentiment
## t = 1.8428, df = 2073.7, p-value = 0.9672
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.04114409
## sample estimates:
## mean of x mean of y
##  0.166305  0.144570
```

```r
t.test(country_sentiment,rock_sentiment, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  country_sentiment and rock_sentiment
## t = 1.8428, df = 2073.7, p-value = 0.03275
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.002326007         Inf
## sample estimates:
## mean of x mean of y
##  0.166305  0.144570
```

```r
t.test(country_sentiment,pop_sentiment, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  country_sentiment and pop_sentiment
## t = -0.81113, df = 2032.4, p-value = 0.2087
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.009514611
## sample estimates:
## mean of x mean of y
## 0.1663050 0.1755536
```

```r
t.test(country_sentiment,pop_sentiment, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  country_sentiment and pop_sentiment
## t = -0.81113, df = 2032.4, p-value = 0.7913
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.02801166         Inf
## sample estimates:
## mean of x mean of y
## 0.1663050 0.1755536
```

```r
t.test(rock_sentiment,pop_sentiment, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  rock_sentiment and pop_sentiment
## t = -2.7963, df = 2070, p-value = 0.002609
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01274991
## sample estimates:
## mean of x mean of y
## 0.1445700 0.1755536
```

```
t.test(rock_sentiment,pop_sentiment, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  rock_sentiment and pop_sentiment
## t = -2.7963, df = 2070, p-value = 0.9974
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.04921723         Inf
## sample estimates:
## mean of x mean of y
## 0.1445700 0.1755536
```

##Conclusions: I looked at one-tail t-tests between each genre in both directions. I found that there is evidence at a significance level of alpha = 0.01 that the true sentiment of the pop tweets is greater than the sentiment of the rock tweets. I also found that there is evidence at a signficiance level of alpha = 0.05 that the true sentiment of the country tweets is greater than the sentiment of the rock tweets. I did not find evidence that the mean of the pop and country tweets differ significantly.

#Summarize location data. Used to filter tweets in python code by finding large numbers of tweets from unexpected locations.

```
country_locationCount <- tweet_df %>%
  filter(genre == "country") %>%
  group_by(location) %>%
  summarise(location.count = n()) %>%
  arrange(desc(location.count))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
rock_locationCount <- tweet_df %>%
  filter(genre == "rock") %>%
  group_by(location) %>%
  summarise(location.count = n()) %>%
  arrange(desc(location.count))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
pop_locationCount <- tweet_df %>%
  filter(genre == "pop") %>%
  group_by(location) %>%
  summarise(location.count = n()) %>%
  arrange(desc(location.count))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```