Sentiment Analysis Demo

Max Baskin

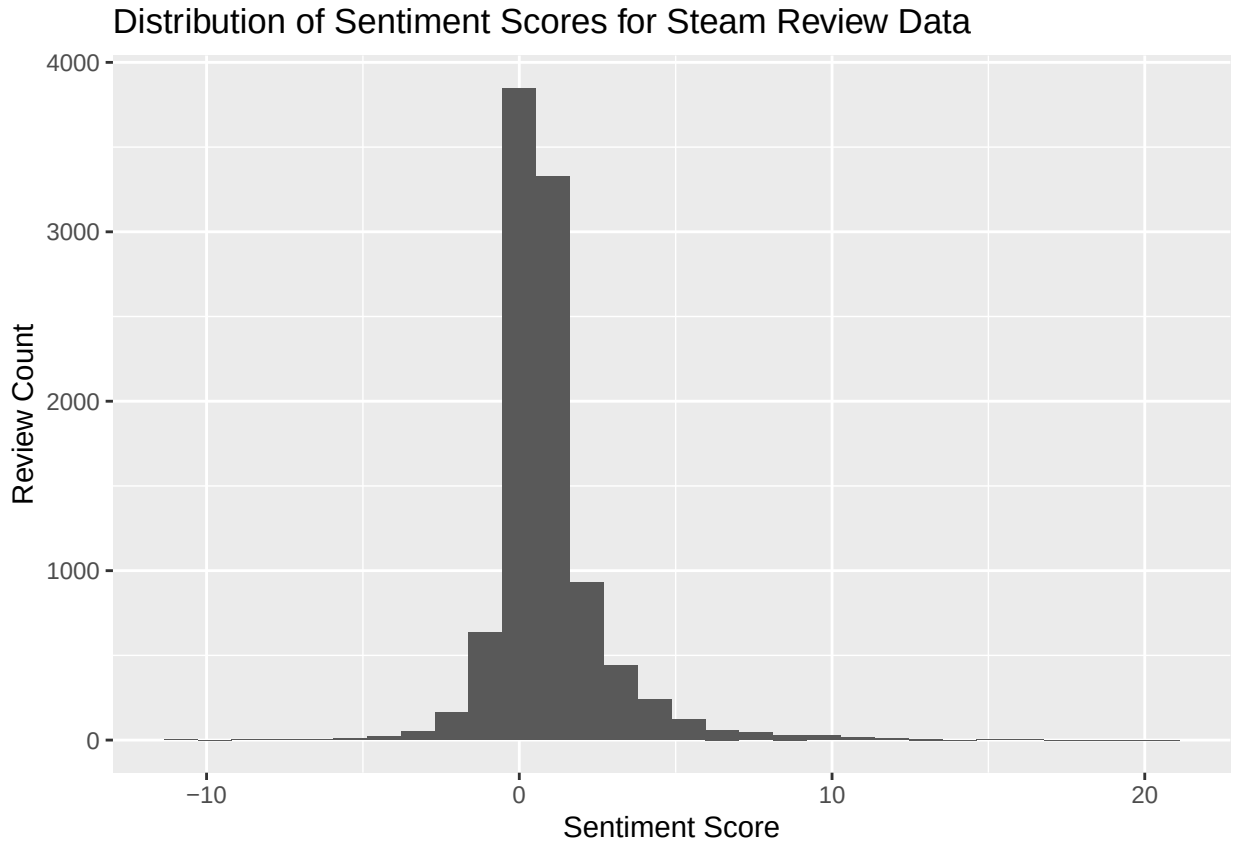
2025-03-10

```
# GUIDE USED: https://www.r-bloggers.com/2021/05/sentiment-analysis-in-r-3/

# data<-read.csv("steam_reviews_english.csv")

# Getting a random sample of 10000 reviews for manageability

# data<- data[sample(nrow(data), size = 10000), ]

# For loading the truncated data from above, for brevity
data<- read.csv("truncated_data.csv")
names(data)
```

```
##  [1] "X"                                "app_id"
##  [3] "app_name"                         "review_id"
##  [5] "language"                         "review"
##  [7] "timestamp_created"                "timestamp_updated"
##  [9] "recommended"                      "votes_helpful"
## [11] "votes_funny"                      "weighted_vote_score"
## [13] "comment_count"                    "steam_purchase"
## [15] "received_for_free"                "written_during_early_access"
## [17] "author.steamid"                   "author.num_games_owned"
## [19] "author.num_reviews"               "author.playtime_forever"
## [21] "author.playtime_last_two_weeks" "author.playtime_at_review"
## [23] "author.last_played"
```

```
# Loading the review texts as a Corpus
library(tm)
corpus1<-iconv(data$review)

# Cleaning the review texts
corpus1<-tolower(corpus1)
corpus1<-removePunctuation(corpus1)
corpus1<-removeNumbers(corpus1)
corpus1<-gsub('\n','', corpus1)

# Perform sentiment analysis
library(syuzhet)

# Get sentiment scores using the NRC method
#(negative is negative sentiment, positive is positive sentiment)
```

```r
data$review_sentiment<-get_sentiment(corpus1)

# Plot distribution of sentiments
library(ggplot2)

sentiments<-ggplot(data)+
  geom_histogram(aes(x = review_sentiment))+
  labs(x = "Sentiment Score", y = "Review Count",
       title = "Distribution of Sentiment Scores for Steam Review Data")
sentiments
```



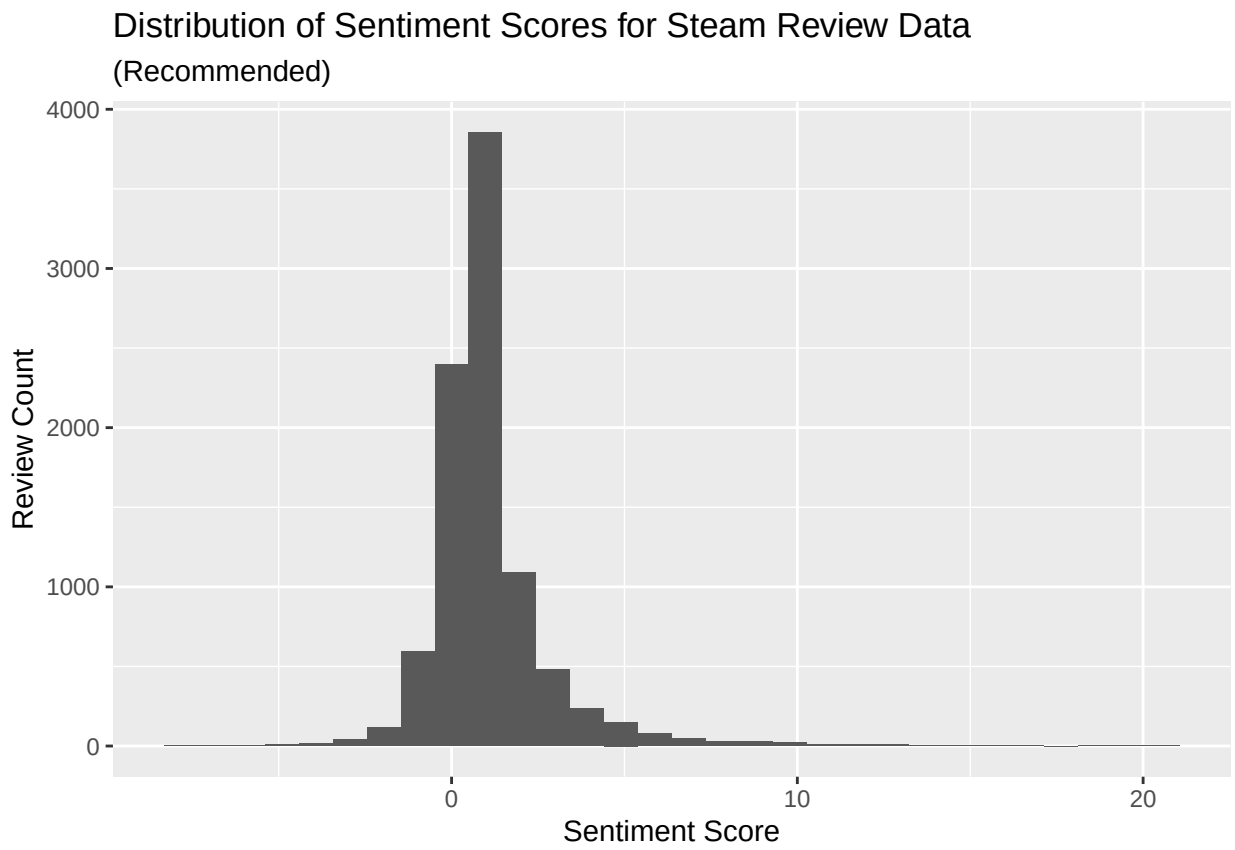

```r
library(dplyr)
data<-data%>%
  mutate(polar = ifelse(review_sentiment > 0, "positive", "negative"))%>%
  mutate(polar = ifelse(review_sentiment == 0, "neutral", polar))

pos<-sum(data$polar == "positive")
neg<-sum(data$polar == "negative")
neu<-sum(data$polar == "neutral")
sentiments_c<-c(pos, neg, neu)

datarec<-ifelse(data$recommended == "True", T, F)
data$recommended<-datarec
rec<-sum(data$recommended)
notrec<-10000-rec
```

```
recommended_c<-c(rec, notrec, NA)

# Compare recommendations to sentiments
library(data.table)
data.frame(recommended_c, sentiments_c,
           row.names = c("Positive","Negative","Neutral"))
```

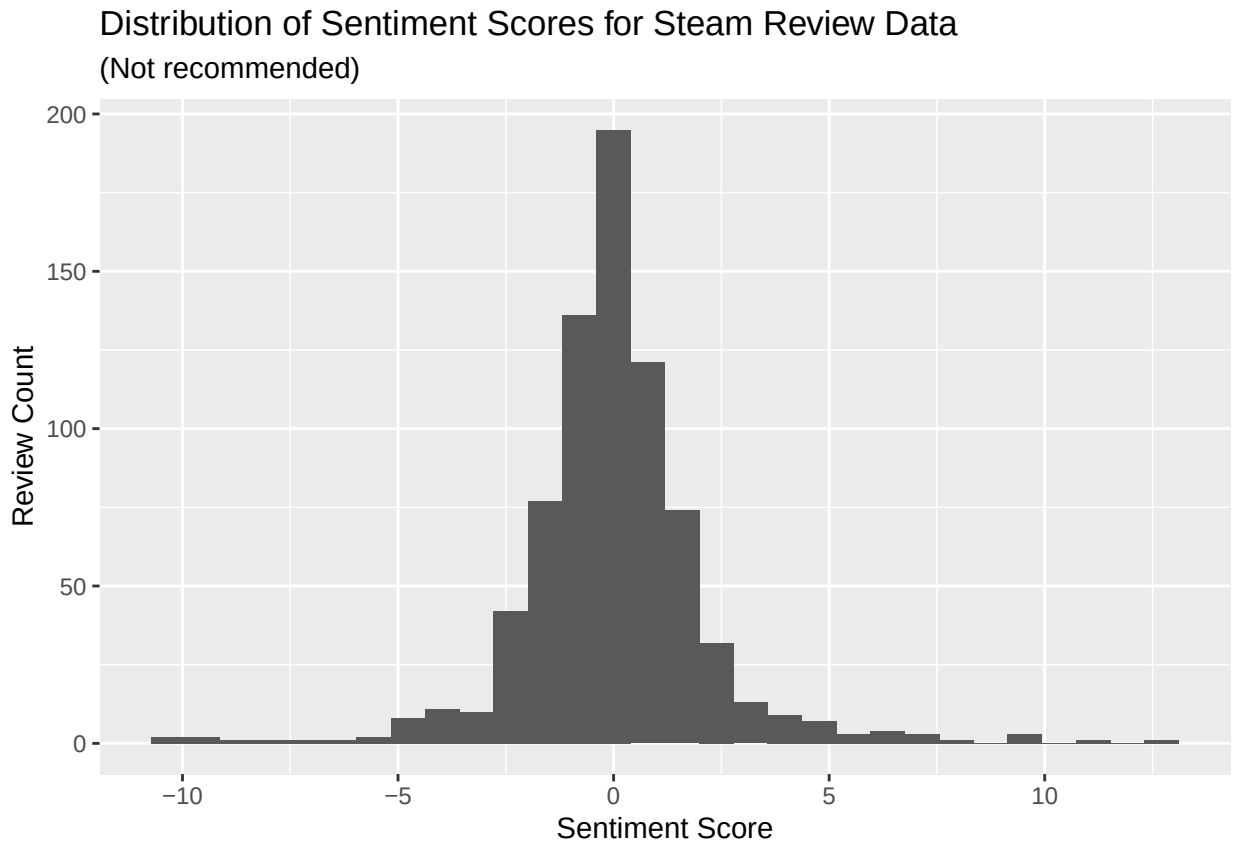
```
##          recommended_c sentiments_c
## Positive          9239         6806
## Negative           761         1484
## Neutral             NA         1710
```

T-tests

```
# Split data into recommended and not recommended games to test hypotheses
recommendedGames<-data[data$recommended == T, ]
sentimentsRecommended<-ggplot(recommendedGames)+
  geom_histogram(aes(x = review_sentiment))+
  labs(x = "Sentiment Score", y = "Review Count",
       title = "Distribution of Sentiment Scores for Steam Review Data",
       subtitle = "(Recommended)")
sentimentsRecommended
```



```r
notRecommendedGames<-data[data$recommended == F, ]
sentimentsNotRecommended<-ggplot(notRecommendedGames)+
  geom_histogram(aes(x = review_sentiment))+
  labs(x = "Sentiment Score", y = "Review Count",
       title = "Distribution of Sentiment Scores for Steam Review Data",
       subtitle = "(Not recommended)")
sentimentsNotRecommended
```




```r
t.test(recommendedGames$review_sentiment, alternative = "greater", mu = 0 )
```

```
##
##  One Sample t-test
##
## data:  recommendedGames$review_sentiment
## t = 54.151, df = 9238, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.9598984       Inf
## sample estimates:
## mean of x
## 0.9899719
```

Appear to solidly reject our null hypothesis (that the mean sentiment for recommended games is negative or neutral)

```r
t.test(notRecommendedGames$review_sentiment, alternative = "less", mu = 0 )
```

```
##
##  One Sample t-test
##
## data:  notRecommendedGames$review_sentiment
## t = -0.056352, df = 760, p-value = 0.4775
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf 0.1242474
## sample estimates:
##    mean of x
## -0.004402102
```

In this case, we cannot reject our null hypothesis (that the mean sentiment for not recommended games is positive or neutral). Upon further inspection, some reviews appear to not recommend the game despite praising certain elements of gameplay, which may be contributing to this result.

RESOURCES USED:

```r
# RESOURCES USED:
# https://www.statology.org/random-sample-in-r/
# Bobbit`t, "How to Select Random Samples in R (With Examples)," Statology, Oct. 22, 2020. https://www.

# https://stackoverflow.com/questions/27044727/removing-characters-from-string-in-r
# ben_aaron, "Removing characters from string in R," Stack Overflow, Nov. 20, 2014. https://stackoverfl

# https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
```