joshjarvey_final

Josh Jarvey

4/17/2020

1. Load in the necessary libraries.

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggformula)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

## Loading required package: ggstance

##
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh

##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

2. Load in the data sets.

```
  #read in the tweet data, removing neutral tweets. Of 5837 tweets collected, 4171 will be considered for this analysis.
tweets_no_neutral =
read_csv("C:/Users/joshj/Documents/GitHub/ds710spring2020finalproject/combined_dataset.csv") %>% filter(Polarity != 0)
```

```
## Parsed with column specification:
## cols(
##   Polarity = col_double(),
##   Tweet = col_character(),
##   isPositive = col_double()
## )
```

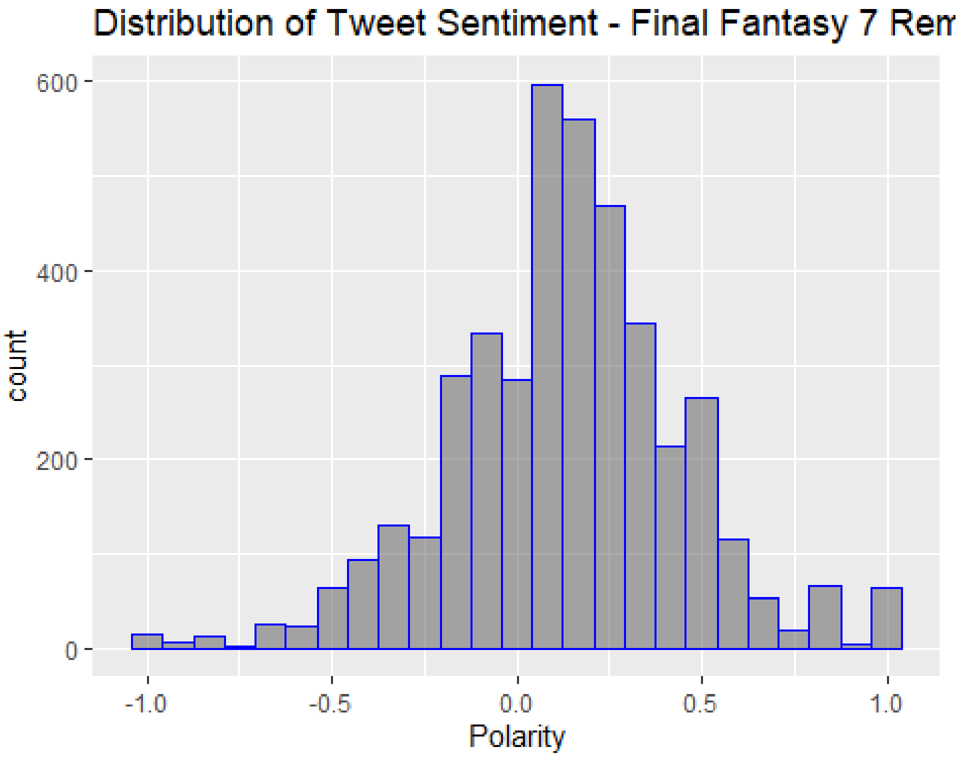
```
  #read in the character name count totals.
character_counts =
read_csv("C:/Users/joshj/Documents/GitHub/ds710spring2020finalproject/charact
er_frequency.csv", col_names = c("Character", "Frequency"))
```

```
## Parsed with column specification:
## cols(
##   Character = col_character(),
##   Frequency = col_double()
## )
```

3. Review data distribution.

Generate a histogram of the twitter sentiment data to review its shape. We expect to see a normal shape, so we can proceed forward with our statistical tests.

```
  #data appears to generally follow a normal distribution.
  #evidence of right-skew toward the positive end.
gf_histogram(~ Polarity, data = tweets_no_neutral, color = "blue") %>%
            gf_labs(title="Distribution of Tweet Sentiment - Final Fantasy 7
Remake")
```



4a. Question 1: Is the game reciving a general positive reception upon its release?

Now that we have confirmed the data generally follows a normal distribution, lets set up our first hypothesis test, a 1-sample proportion test on sentiment.

- Assumptions:
  - Data was collected via SRS process using a pull of twitter data at random for the given search tag.
  - Observations are greater than 10. Sum of positive tweets = 2920, total tweets = 4171.
  - Data generally follows a normal distribution.
- Significance level:
  - alpha = 0.05
- Hypothesis:
  - p1 = The proportion of tweets that are graded as positive sentiment about Final Fantasy 7 remake
  - H_0: p1 <= 0.50
  - H_a: p1 > 0.50

```
  #complete the 1-prop test, checking greater than
prop.test(sum(tweets_no_neutral$isPositive), nrow(tweets_no_neutral), p=0.50,
alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(tweets_no_neutral$isPositive) out of nrow(tweets_no_neutral),
null probability 0.5
## X-squared = 667.04, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.6881537 1.0000000
## sample estimates:
##         p
## 0.7000719
```

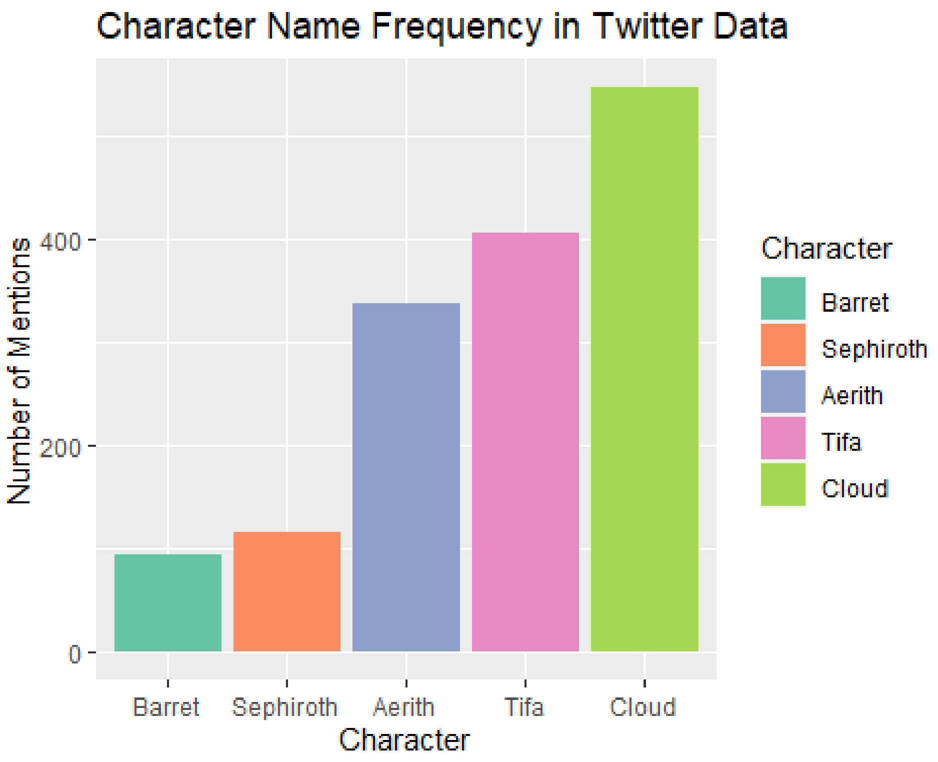
4b. Conclusion:

At a 0.05 significance level, there is enough evidence to support the alternative hypothesis that the majority of twitter user's sentiment about Final Fantasy 7 Remake is positive.

5a. Character frequency:

In this section we look at character name frequency in our sample of tweets. First we generate a bar graph using the character counts data set and review the distribution.

```
  #first re-create the character's name column based on order of frequency
from least to greatest, then create the bar graph.
character_counts %>%
  mutate(Character = reorder(Character, Frequency)) %>%
  gf_col(Frequency ~ Character, fill =~ Character) %>%
  gf_refine(scale_fill_brewer(palette = "Set2")) %>%
  gf_labs(title="Character Name Frequency in Twitter Data",
  y="Number of Mentions")
```



5b. Question 2: Are the game's main characters being equally discussed statistically, or is there evidence to suggest a "favorite"?

We will treat each "character" as its own category, and perform a chi-squared goodness of fit test to test distribution.

- Assumptions:
  - Data was collected via SRS process using a pull of twitter data at random for the given search tag.
  - Total Observations of characters are 1500. Expected counts for each category are greater than 5.
- Significance level:
  - alpha = 0.05
- Hypothesis:
  - p1 = The proportion of tweets that mention the character "Cloud"
  - p2 = The proportion of tweets that mention the character "Tifa"
  - p3 = The proportion of tweets that mention the character "Aerith"
  - p4 = The proportion of tweets that mention the character "Sephiroth"
  - p5 = The proportion of tweets that mention the character "Barret"
  - H_0: p1 = p2 = p3 = p4 = p5
  - H_a: p1 <> p2 <> p3 <> p4 <> p5

```
  #calculating equal distribution. There are 5 characters, so 1/5 is 0.20 (or 20%)
```

```
character_counts = character_counts %>% mutate(equaldistribution =
1/nrow(character_counts))

  #checking expected counts for greater than 5. These equally result in
expected values of 300
character_counts$equaldistribution * sum(character_counts$Frequency)

## [1] 300 300 300 300 300

  #performing a chi-squared goodness of fit test. Null hypothesis is that
each characters frequency is equal
chisq.test(character_counts$Frequency, p =
character_counts$equaldistribution)

##
##  Chi-squared test for given probabilities
##
## data:  character_counts$Frequency
## X-squared = 499.69, df = 4, p-value < 2.2e-16
```

5c. Conclusion:

At a 0.05 significance level, there is enough evidence to support the alternative hypothesis that the frequency of character names being mentioned in tweets are not equally distributed. Thus some characters are being discussed more than others, which may indicate to them being more “popular” or “favorited”.

6a. Question 3: Since character discussion is unevenly distributed, is the character with the most counts the majority favorite?

We will use a 1-proportion z-test to check if the character with the most mentions, Cloud (who is arguably the main character), is who the majority of twitter users are discussing.

- Assumptions:
  - Data was collected via SRS process using a pull of twitter data at random for the given search tag.
  - Observations are greater than 10. Total character name mentions is 1500.
- Significance level:
  - alpha = 0.05
- Hypothesis:
  - p1 = The proportion of the most discussed character, Cloud.
  - H_0: p1 <= 0.50
  - H_a: p1 > 0.50

```
  #testing if the most mentioned character in my sample of tweets is being
mentioned the majority of the time over the other characters.
prop.test(max(character_counts$Frequency), sum(character_counts$Frequency), p
= 0.50, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  max(character_counts$Frequency) out of
sum(character_counts$Frequency), null probability 0.5
## X-squared = 109.35, df = 1, p-value = 1
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.3441562 1.0000000
## sample estimates:
##         p
## 0.3646667
```

6b. Conclusion:

At a 0.05 significance level, there is not enough evidence to support that the character “Cloud” is who the majority of twitter users are discussing in their posts about Final Fantasy 7 remake. While we discovered there is an unequal distribution in #5 above, there doesnt appear to be a single most popular character that’s being mentioned.