FinalprojectDS710_Aditya_Nanduri

Aditya

5/4/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
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
```

```r
library(readr)
library(ggformula)
```

```
## Loading required package: ggplot2
```

```
## 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")
```

```r
# Read CSV files into R
trump_df = read_csv("Trumprelatedtweets05052020_2.csv")
```

```
## Parsed with column specification:
## cols(
##   Index = col_double(),
##   Text = col_character(),
##   screen_name = col_character(),
##   created_at = col_datetime(format = ""),
##   Is_retweeted = col_logical(),
##   retweet_count = col_double(),
##   favorite_count = col_double(),
##   country = col_character(),
##   Sentiment = col_character(),
##   polarity = col_double(),
##   subjectivity = col_double(),
##   attitude = col_character(),
##   candidate = col_character()
## )
```

```
biden_df = read_csv("Bidenrelatedtweets05052020_2.csv")
```

```
## Parsed with column specification:
## cols(
##   Index = col_double(),
##   Text = col_character(),
##   screen_name = col_character(),
##   created_at = col_datetime(format = ""),
##   Is_retweeted = col_logical(),
##   retweet_count = col_double(),
##   favorite_count = col_double(),
##   country = col_character(),
##   Sentiment = col_character(),
##   polarity = col_double(),
##   subjectivity = col_double(),
##   attitude = col_character(),
##   candidate = col_character()
## )
```

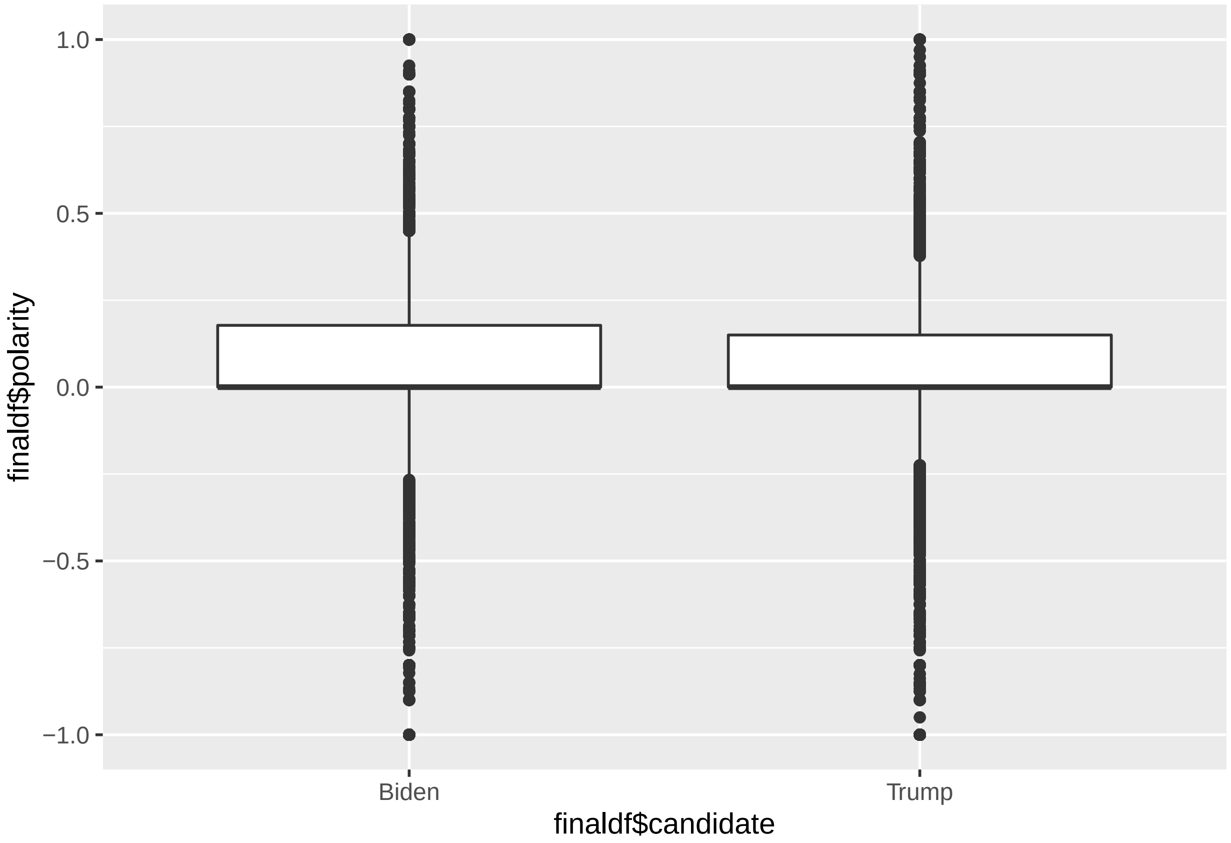
```
# get summary of Trump tweets
summary(trump_df)
```

```
##      Index           Text           screen_name
##  Min.   :    0   Length:14009       Length:14009
##  1st Qu.: 3502   Class :character   Class :character
##  Median : 7004   Mode  :character   Mode  :character
##  Mean   : 7004
##  3rd Qu.:10506
##  Max.   :14008
##    created_at                    Is_retweeted    retweet_count
##  Min.   :2020-05-05 03:44:14   Mode :logical   Min.   :   0.000
##  1st Qu.:2020-05-05 13:03:07   FALSE:14009     1st Qu.:   0.000
##  Median :2020-05-05 16:33:17                   Median :   0.000
##  Mean   :2020-05-05 15:57:12                   Mean   :   1.465
##  3rd Qu.:2020-05-05 19:57:51                   3rd Qu.:   0.000
##  Max.   :2020-05-05 23:05:22                   Max.   :1543.000
##  favorite_count       country            Sentiment            polarity
##  Min.   :   0.000   Length:14009       Length:14009       Min.   :-1.00000
##  1st Qu.:   0.000   Class :character   Class :character   1st Qu.: 0.00000
##  Median :   0.000   Mode  :character   Mode  :character   Median : 0.00000
##  Mean   :   4.693                                         Mean   : 0.04073
##  3rd Qu.:   1.000                                         3rd Qu.: 0.15000
##  Max.   :8186.000                                         Max.   : 1.00000
##   subjectivity      attitude          candidate
##  Min.   :0.0000   Length:14009       Length:14009
##  1st Qu.:0.0000   Class :character   Class :character
##  Median :0.2500   Mode  :character   Mode  :character
##  Mean   :0.3218
##  3rd Qu.:0.5833
##  Max.   :1.0000
```

```
# get summary of Bdien tweets
summary(biden_df)
```

```
##      Index           Text           screen_name
##  Min.   :    0   Length:14033       Length:14033
##  1st Qu.: 3508   Class :character   Class :character
##  Median : 7016   Mode  :character   Mode  :character
##  Mean   : 7016
##  3rd Qu.:10524
##  Max.   :14032
##    created_at                   Is_retweeted    retweet_count
##  Min.   :2020-05-03 18:27:08   Mode :logical   Min.   :   0.000
##  1st Qu.:2020-05-04 04:10:37   FALSE:14033     1st Qu.:   0.000
##  Median :2020-05-04 20:37:54                   Median :   0.000
##  Mean   :2020-05-04 21:04:33                   Mean   :   2.129
##  3rd Qu.:2020-05-05 14:02:33                   3rd Qu.:   0.000
##  Max.   :2020-05-05 23:53:08                   Max.   :5111.000
##  favorite_count       country           Sentiment            polarity
##  Min.   :    0.00   Length:14033       Length:14033       Min.   :-1.00000
##  1st Qu.:    0.00   Class :character   Class :character   1st Qu.: 0.00000
##  Median :    0.00   Mode  :character   Mode  :character   Median : 0.00000
##  Mean   :    7.63                                         Mean   : 0.04811
##  3rd Qu.:    1.00                                         3rd Qu.: 0.17778
##  Max.   :32613.00                                         Max.   : 1.00000
##   subjectivity      attitude          candidate
##  Min.   :0.0000   Length:14033       Length:14033
##  1st Qu.:0.0000   Class :character   Class :character
##  Median :0.2679   Mode  :character   Mode  :character
##  Mean   :0.3310
##  3rd Qu.:0.6000
##  Max.   :1.0000
```

```
# Plot Sentiment scores of Biden vs Trump - This will show graphically who has better scores
finaldf <- rbind(trump_df,biden_df)
gf_boxplot(finaldf$polarity ~ finaldf$candidate, data = finaldf)
```



****HYPOTHESIS****

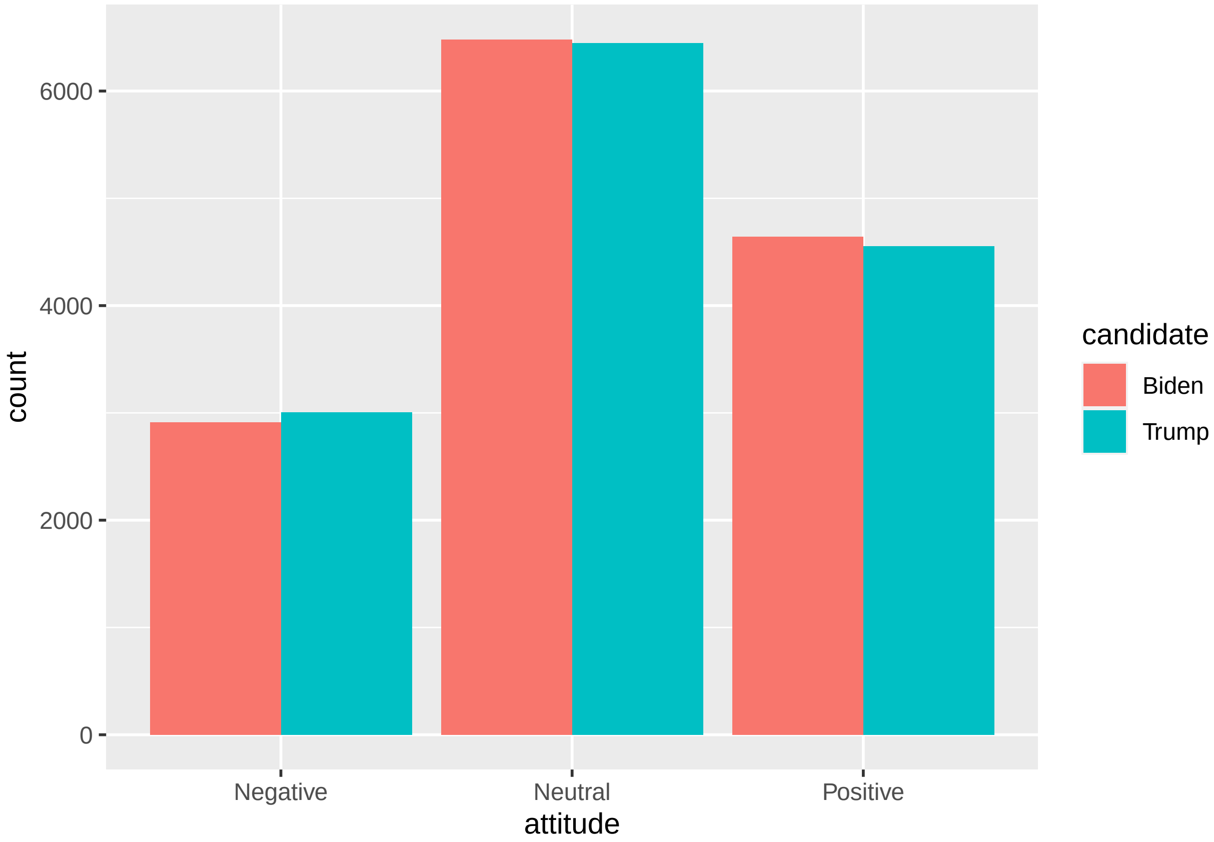
-The Null Hypothesis H0: Mean(sentiment score of trump) <= Mean(sentiment score of Biden) - The alternate Hypothesis Ha: Mean(sentiment score of trump) > Mean(sentiment score of Biden)

```
# Get individual scores of Trump and Biden to run a T-Test
trump_scores = trump_df$polarity
biden_scores = biden_df$polarity
```

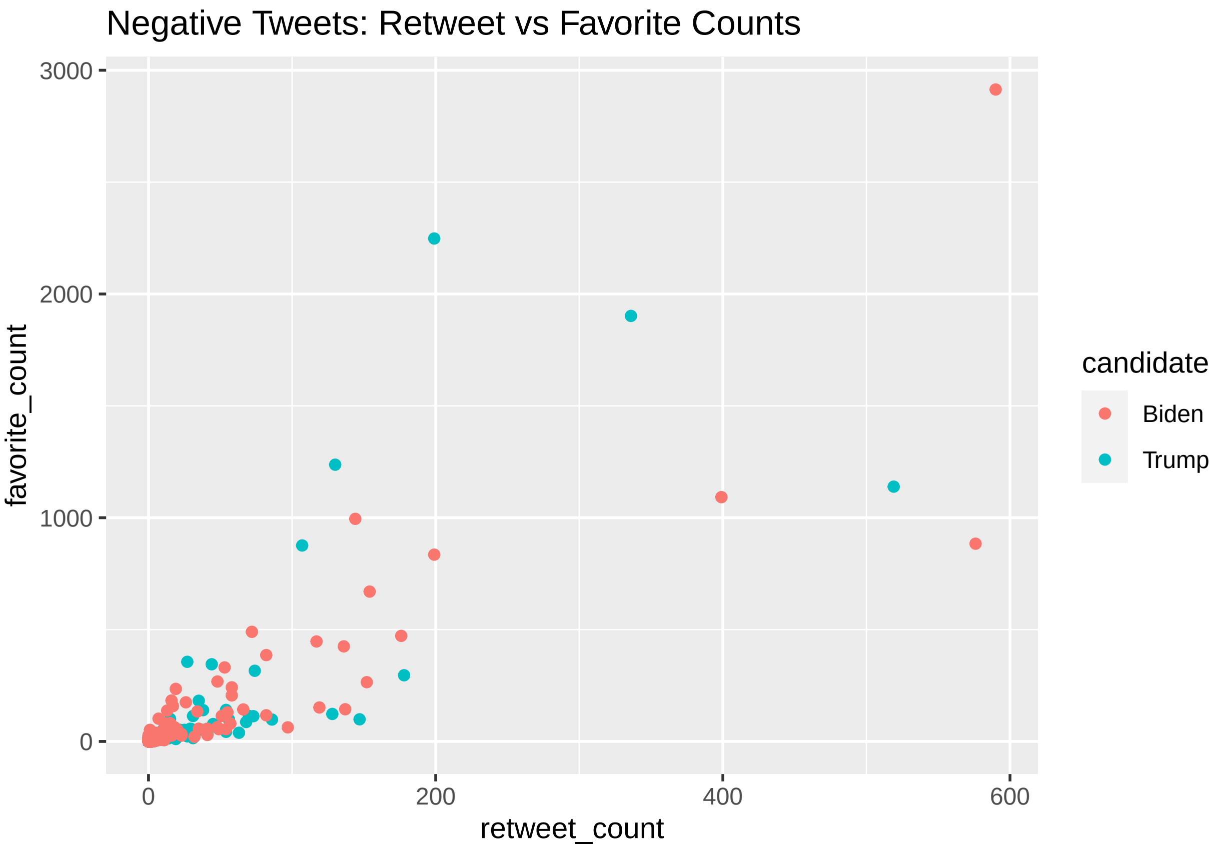
```
t.test(trump_scores,biden_scores, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  trump_scores and biden_scores
## t = -2.0157, df = 28037, p-value = 0.9781
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.01339094         Inf
## sample estimates:
##  mean of x  mean of y
## 0.04073274 0.04810649
```

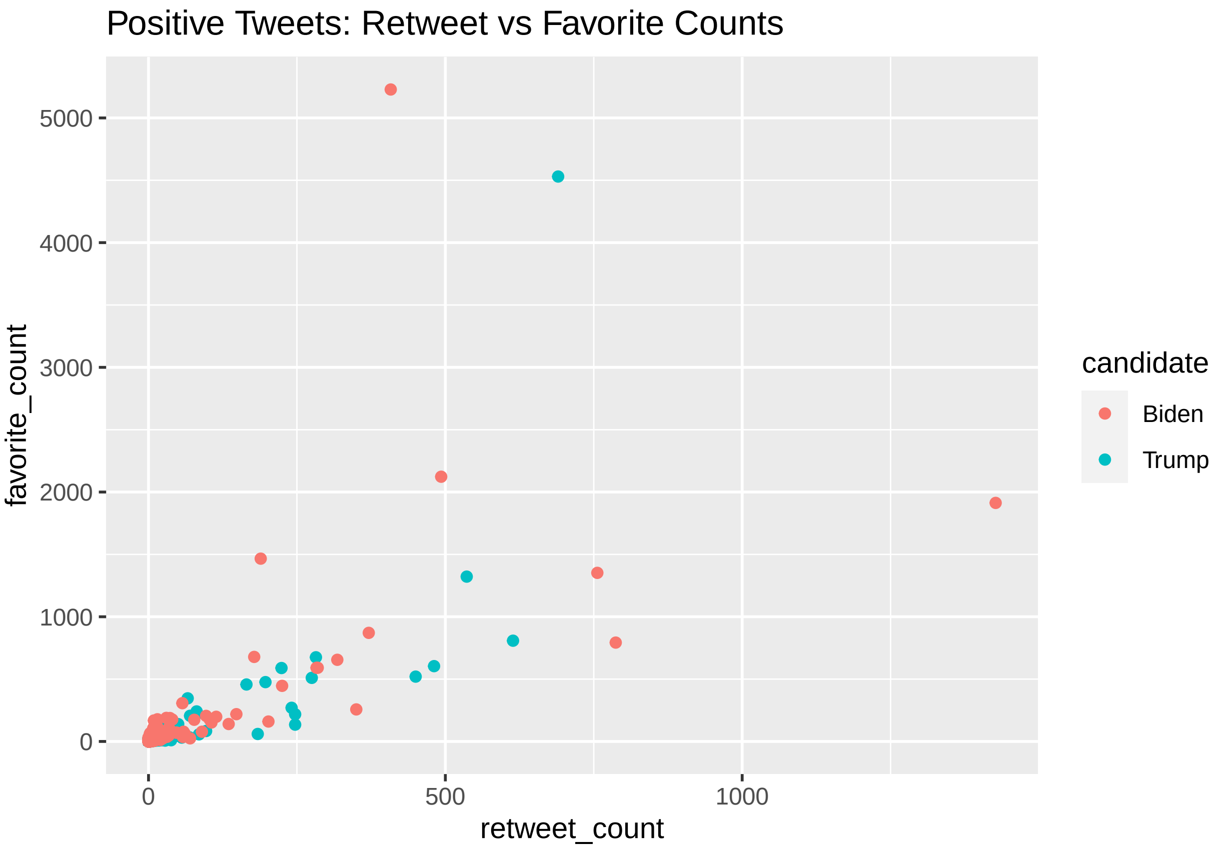
```
# Bar Charts of Number of Positive vs Negative vs Neutral tweets for both the candidates
gf_bar(~ attitude,fill =~candidate,position = position_dodge(),data = finaldf )
```



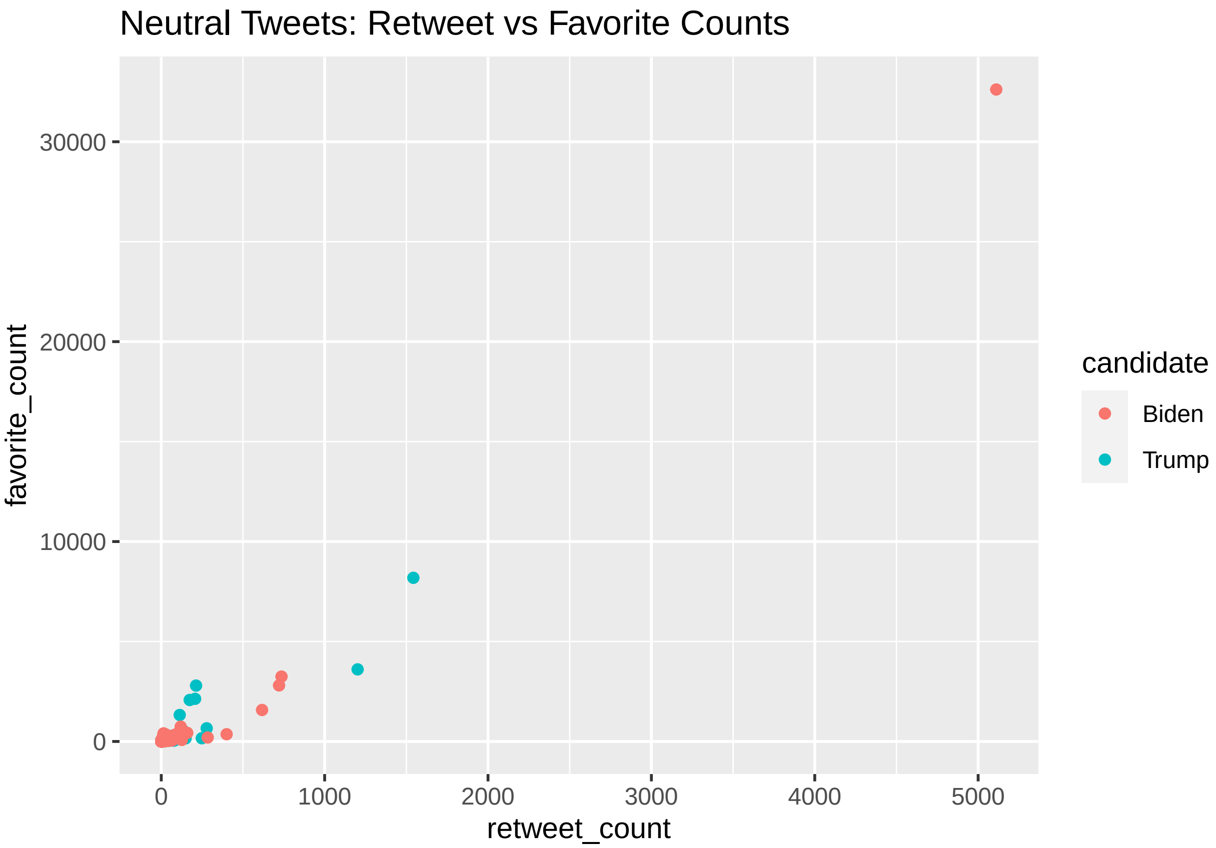
```
# scatter plots of Retweets vs Favorite counts of all Negative Tweets
finaldf_negative <- subset(finaldf,finaldf$attitude == "Negative")
negativeplot <- ggplot(finaldf_negative, aes(x = retweet_count, y = favorite_count)) + geom_point(aes(c
print(negativeplot + ggtitle("Negative Tweets: Retweet vs Favorite Counts"))
```



```
# scatter plots of Retweets vs Favorite counts of all Positive Tweets
finaldf_positive <- subset(finaldf,finaldf$attitude == "Positive")
positiveplot <- ggplot(finaldf_positive, aes(x = retweet_count, y = favorite_count)) + geom_point(aes(co
print(positiveplot + ggtitle("Positive Tweets: Retweet vs Favorite Counts"))
```

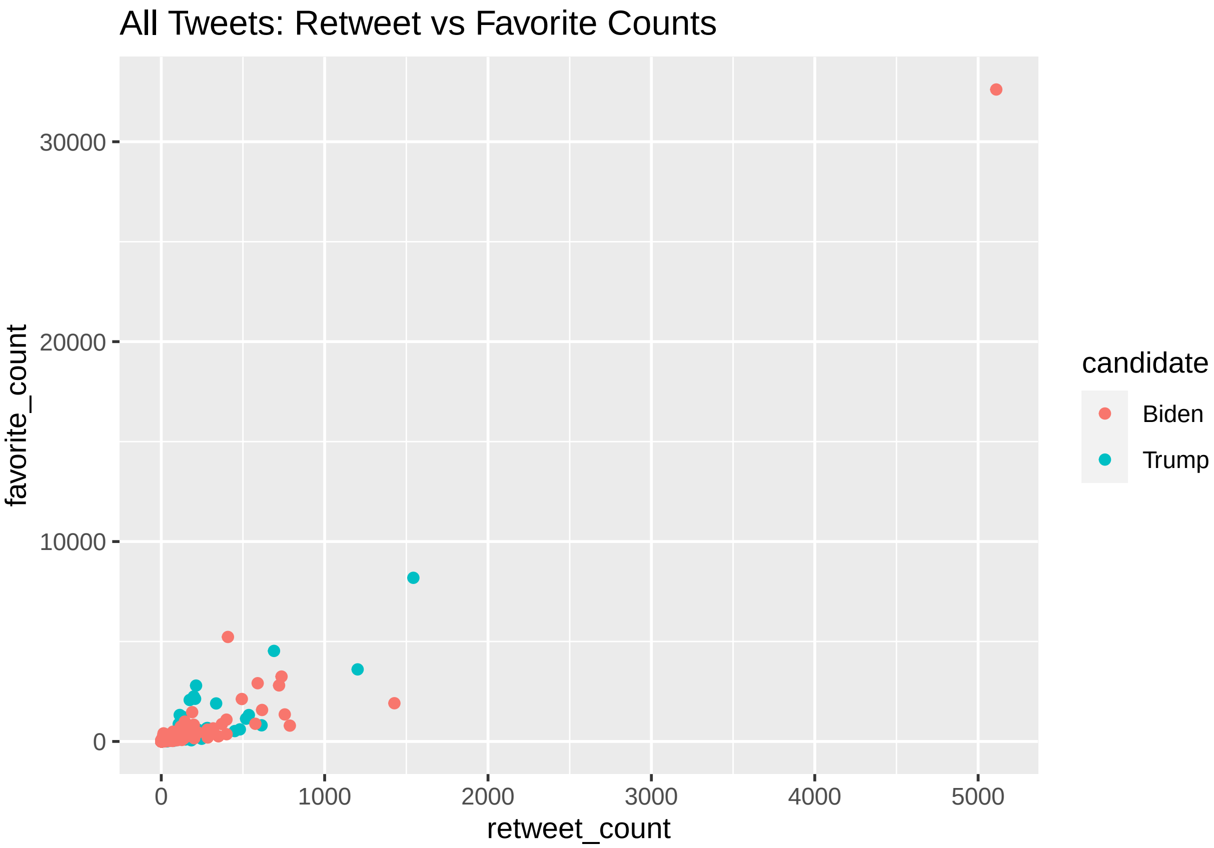


```
# scatter plots of Retweets vs Favorite counts of all Neutral Tweets
finaldf_Neutral <- subset(finaldf,finaldf$attitude == "Neutral")
Neutralplot <- ggplot(finaldf_Neutral, aes(x = retweet_count, y = favorite_count)) + geom_point(aes(col
print(Neutralplot + ggtitle("Neutral Tweets: Retweet vs Favorite Counts"))
```



You can also embed plots, for example:

```
# scatter plots of Retweets vs Favorite counts of all  Tweets
Allplot <- ggplot(finaldf, aes(x = retweet_count, y = favorite_count)) + geom_point(aes(color = candidat
print(Allplot + ggtitle("All Tweets: Retweet vs Favorite Counts"))
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.