Consumer Interaction Effects on Media Scores

By: Izzy Reeser

Let me introduce the data.

Introduction to the Data

```
head(politics)
```

```
## # A tibble: 6 × 6
##    ...1 Title                          Date       Time     Score `Number of Comme…
##   <dbl> <chr>                          <date>     <time>   <dbl>              <dbl>
## 1     0 Megathread: Sean Spicer WWI… 2017-04-12 00:51:25 14938               6461
## 2     1 Manafort Firm Received Ukra… 2017-04-12 15:12:43 11993                809
## 3     2 Its not too late to get rid… 2017-04-12 17:42:17  3119                748
## 4     3 Donald Trump, who doesnt re… 2017-04-12 14:14:59  5758                518
## 5     4 Did he or didnt he? Trump c… 2017-04-12 16:52:49  2993                256
## 6     5 Trump lists Carter Page amo… 2017-04-12 17:14:42  2018                 58
```

```
head(sports)
```

```
## # A tibble: 6 × 6
##    ...1 Title                          Date       Time     Score `Number of Comme…
##   <dbl> <chr>                          <date>     <time>   <dbl>              <dbl>
## 1     0 Raiders now only team in NF… 2017-04-12 03:55:43 16673                541
## 2     1 Mike Tyson goes full matrix… 2017-04-11 18:11:06  9072                932
## 3     2 Blake Griffins custom Air J… 2017-04-12 14:05:03   399                 10
## 4     3 How would the Premier Leagu… 2017-04-12 16:57:22    13                  8
## 5     4 Fernando Alonso, McLaren dr… 2017-04-12 16:00:34    13                  4
## 6     5 Islamic extremists reported… 2017-04-12 19:57:28     5                  1
```

- 4 data sets
- Each have the same columns, but different values

```
head(worldnews)
```

```
## # A tibble: 6 × 6
##    ...1 Title                          Date       Time     Score `Number of Comme…
##   <dbl> <chr>                          <date>     <time>   <dbl>              <dbl>
## 1     0 Israels Holocaust museum in… 2017-04-12 16:24:39 19641               1515
## 2     1 China rejects North Korean … 2017-04-12 10:57:03  4442                525
## 3     2 FBI obtained court order to… 2017-04-12 06:04:37 17990               2421
## 4     3 Isis now in control of just… 2017-04-12 07:18:01  5871                431
## 5     4 Japanese warships to join U… 2017-04-12 17:03:14   699                144
## 6     5 Tourist carves names into C… 2017-04-12 13:23:30  1111                326
```

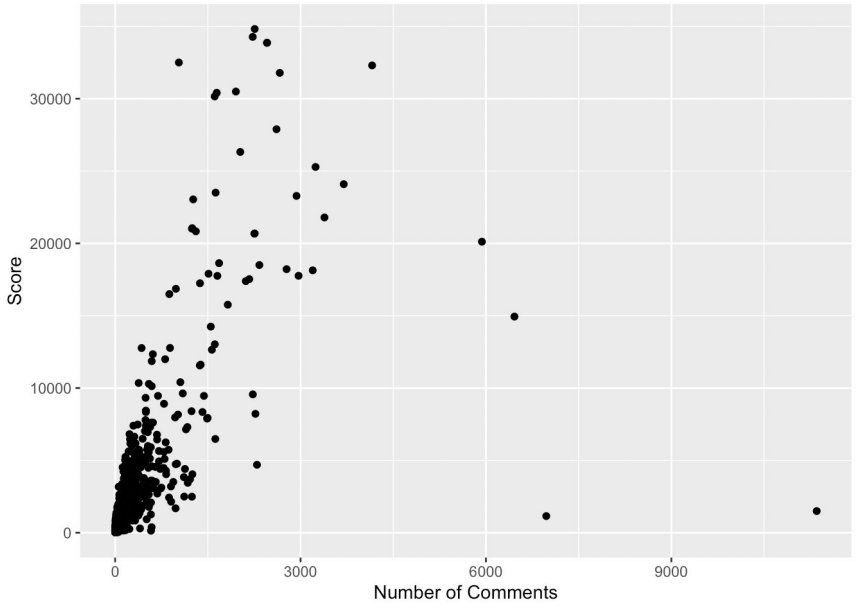
```
head(television)
```

```
## # A tibble: 6 × 6
##    ...1 Title                          Date       Time     Score `Number of Comme…
##   <dbl> <chr>                          <date>     <time>   <dbl>              <dbl>
## 1     0 /r/televisions Whatcha Watc… 2017-04-12 19:08:57     3                 17
## 2     1 Stephen Colbert throws staf… 2017-04-12 01:55:58 26751               1198
## 3     2 38% of Teens Watch Netflix … 2017-04-11 23:54:12 19952               1398
## 4     3 Community - Addicted to Enc… 2017-04-12 16:58:48   271                 26
## 5     4 Danny Pudi when asked about… 2017-04-11 21:10:34  2747                279
## 6     5 Chris Pratt will be honoure… 2017-04-12 01:11:11   981                 82
```

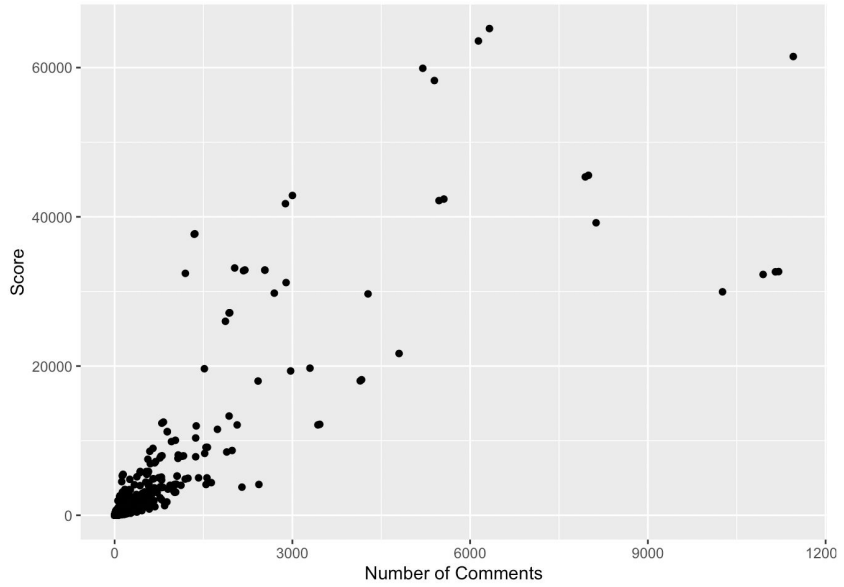
- Comparing the number of comments to score

Introduction to the Data

```
ggplot(politics, aes(x = `Number of Comments`, y = Score)) +
  geom_point()
```



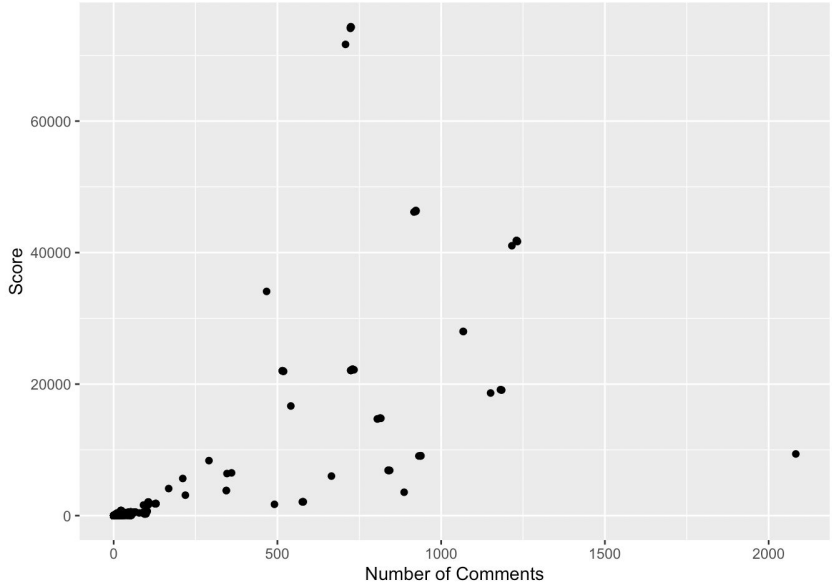

```
ggplot(worldnews, aes(x = `Number of Comments`, y = Score)) +
  geom_point()
```



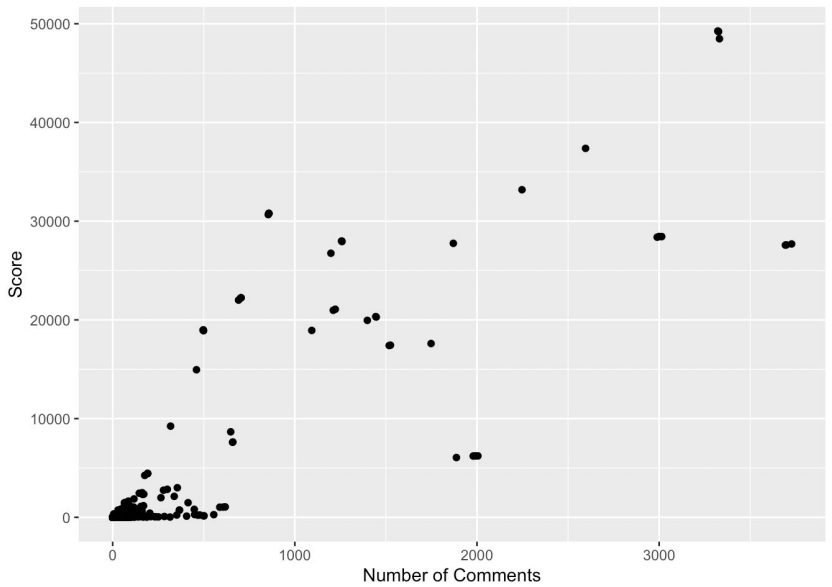
- “Politics” and “worldnews” seem to have steeper slopes.

Introduction to the Data

```
ggplot(sports, aes(x = `Number of Comments`, y = Score)) +
  geom_point()
```




```
ggplot(television, aes(x = `Number of Comments`, y = Score)) +
  geom_point()
```



- “Sports” and “television” do not seem to have steeper slopes.

First, the data needs to be cleaned.

Clean the Data

```
pol_news <- politics %>%
  arrange(Date, Time)
```

```
wrld_news <- worldnews %>%
  arrange(Date, Time)
```

```
sp_news <- sports %>%
  arrange(Date, Time)
```

```
tv_news <- television %>%
  arrange(Date, Time)
```

```
## # A tibble: 6 × 6
##    ...1 Title                          Date       Time     Score `Number of Comme…
##   <dbl> <chr>                          <date>     <time>   <dbl>              <dbl>
## 1    92 Climate Change Is A Nationa… 2017-04-11 17:22:37 10129                592
## 2    97 Attendees chant you lie at … 2017-04-11 18:21:47  6442                679
## 3    89 3 of Alabamas most powerful… 2017-04-11 18:57:40  6177                330
## 4    68 Schumer: If Trump doesnt re… 2017-04-11 19:57:05 12772                888
## 5    90 Sorry America, Your Taxes A… 2017-04-11 20:21:21  4406               1129
## 6    87 Donald Trumps White House c… 2017-04-11 20:36:14  4406                729
```

- Arrange the date and time columns in the data frame so that they align better visually

Clean the Data

```
pol_news <- pol_news %>%
  rename(Comments = `Number of Comments`) %>%
  select(-...1)
head(pol_news)
```

```
sp_news <- sp_news %>%
  rename(Comments = `Number of Comments`) %>%
  select(-...1)
head(sp_news)
```

```
wrld_news <- wrld_news %>%
  rename(Comments = `Number of Comments`) %>%
  select(-...1)
head(wrld_news)
```

```
tv_news <- tv_news %>%
  rename(Comments = `Number of Comments`) %>%
  select(-...1)
head(tv_news)
```

```
## # A tibble: 6 × 5
##   Title                                    Date       Time     Score Comments
##   <chr>                                    <date>     <time>   <dbl>    <dbl>
## 1 Climate Change Is A National Security Issu… 2017-04-11 17:22:37 10129      592
## 2 Attendees chant you lie at U.S. Rep. Joe W… 2017-04-11 18:21:47  6442      679
## 3 3 of Alabamas most powerful Republicans fo… 2017-04-11 18:57:40  6177      330
## 4 Schumer: If Trump doesnt release his tax r… 2017-04-11 19:57:05 12772      888
## 5 Sorry America, Your Taxes Arent High       2017-04-11 20:21:21  4406     1129
## 6 Donald Trumps White House cant even organi… 2017-04-11 20:36:14  4406      729
```

- Rename ‘Number of Comments’ to Comments and take out unnecessary columns.

Next, I created linear models to find the increase of score per comment.

Modeling- Political News

```
pol.lm <- lm(Score ~ Comments, data = pol_news)
summary(pol.lm)
```

```
##
## Call:
## lm(formula = Score ~ Comments, data = pol_news)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52629   -991   -781     28  26634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.4058   104.4025   10.01   <2e-16 ***
## Comments       4.6758     0.1546   30.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3394 on 1198 degrees of freedom
## Multiple R-squared:  0.433,  Adjusted R-squared:  0.4325
## F-statistic: 914.9 on 1 and 1198 DF,  p-value: < 2.2e-16
```

- Suggests that every comment is associated with a 4.7 increase in score
- P-value of 2e-16, which is very close to zero
- 4.7 is highly significant

Modeling- Worldnews

```
wrld.lm <- lm(Score ~ Comments, data = wrld_news)
summary(wrld.lm)
```

```
##
## Call:
## lm(formula = Score ~ Comments, data = wrld_news)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30202.1   -406.7   -307.9   -281.2  30574.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 296.5958   108.9957   2.721   0.0066 **
## Comments      5.5828     0.1042  53.590   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3642 on 1198 degrees of freedom
## Multiple R-squared:  0.7056, Adjusted R-squared:  0.7054
## F-statistic:  2872 on 1 and 1198 DF,  p-value: < 2.2e-16
```

- Suggests that every comment is associated with a 5.6 increase in score
- P-value of 2e-16, which is very close to zero
- 5.6 is highly significant

Modeling- Sports News

```
sp.lm <- lm(Score ~ Comments, data = sp_news)
summary(sp.lm)
```

```
##
## Call:
## lm(formula = Score ~ Comments, data = sp_news)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50449   -123     33     87  53621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -99.1759   149.7841  -0.662    0.508
## Comments     28.7707     0.6875  41.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4986 on 1198 degrees of freedom
## Multiple R-squared:  0.5938, Adjusted R-squared:  0.5935
## F-statistic:  1751 on 1 and 1198 DF,  p-value: < 2.2e-16
```

- Suggests that every comment is associated with a 28.8 increase in score
- P-value of 2e-16, which is very close to zero
- 28.8 is highly significant

Modeling- Television News

```
tv.lm <- lm(Score ~ Comments, data = tv_news)
summary(tv.lm)
```

```
##
## Call:
## lm(formula = Score ~ Comments, data = tv_news)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17119.4   -131.6     19.5     79.2  20874.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -86.9135    72.9543  -1.191    0.234
## Comments     11.6849     0.1877  62.243   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2451 on 1198 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7636
## F-statistic:  3874 on 1 and 1198 DF,  p-value: < 2.2e-16
```

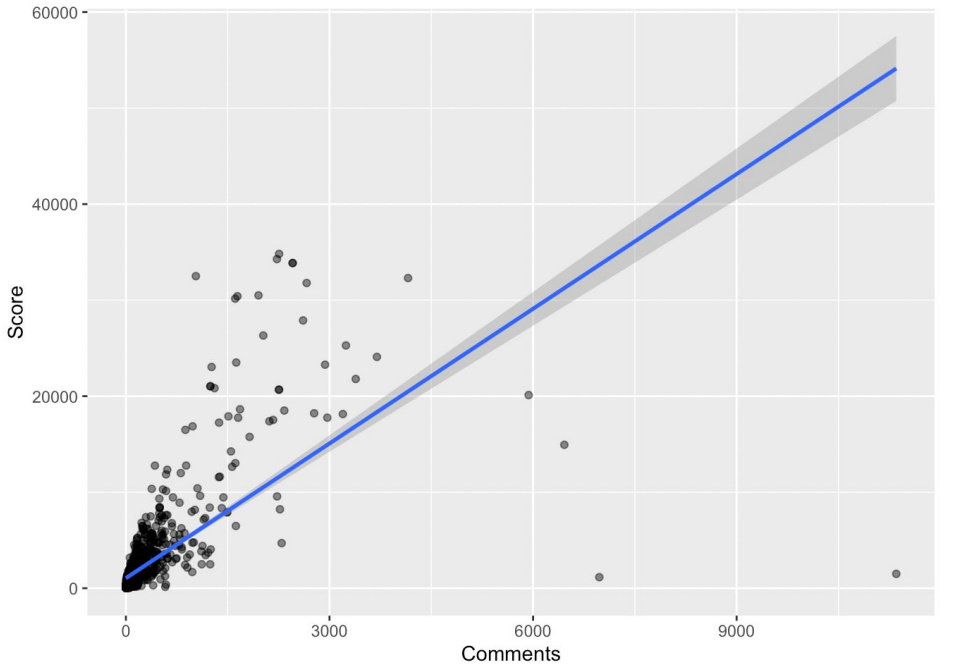
- Suggests that every comment is associated with a 11.7 increase in score
- P-value of 2e-16, which is very close to zero
- 11.7 is highly significant

Then, I graphed the data.

Graphing- Political News

```
ggplot(pol_news, aes(Comments, Score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

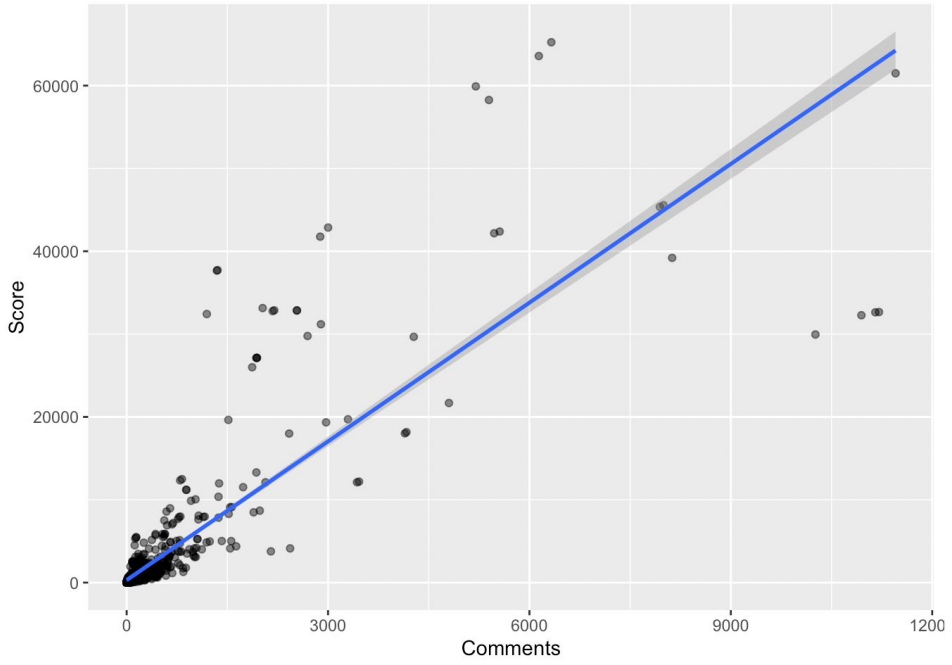


- This graph has outliers past 6,000 comments with low scores.

Graphing- Worldnews

```
ggplot(wrld_news, aes(Comments, Score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

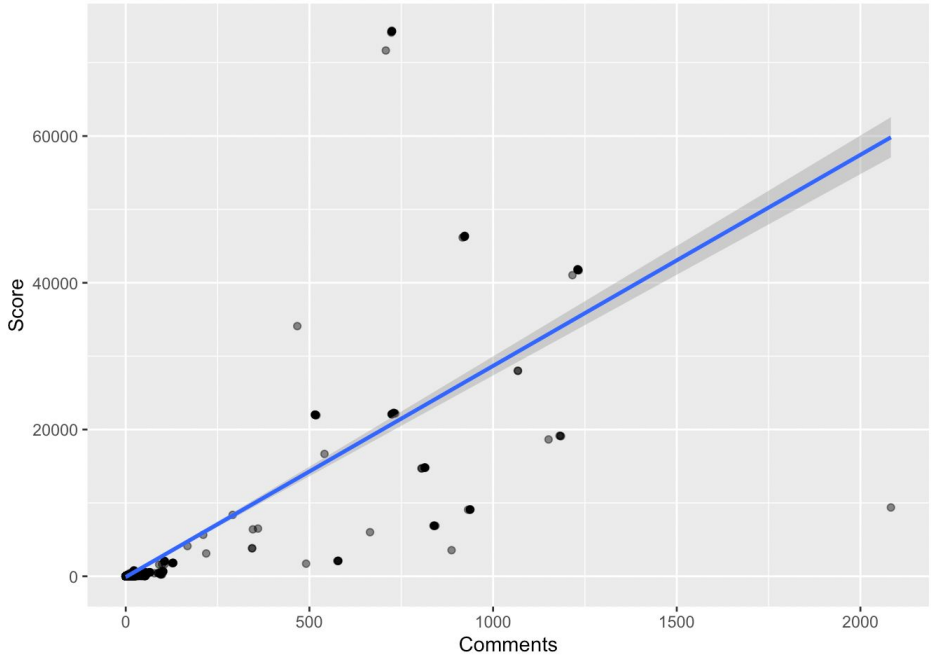


- This graph has outliers past 9,000 comments with low scores.

Graphing- Sports News

```
ggplot(sp_news, aes(Comments, Score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

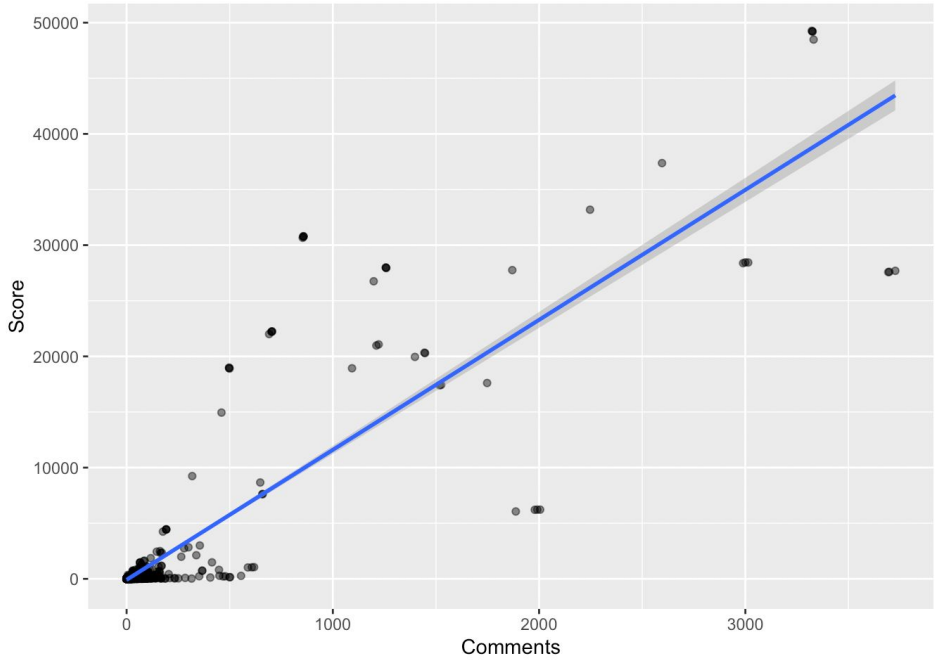


- This graph has outliers past 1,000 comments with low scores.

Graphing- Television News

```
ggplot(tv_news, aes(Comments, Score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- This graph has outliers past 2,000 comments with low scores.

The random outliers do not make sense. Score increases with more interaction.

Cleaning Up Outliers- Political News

```
reduced_pol <- pol_news %>%
  filter(Comments < 4500)

nrow(reduced_pol)
```

```
## [1] 1196
```

- Taking 1200-1196, I figured out that political news only has 4 rows with comment values more than 4,500.

Cleaning Up Outliers- Worldnews

```
reduced_wrld <- wrld_news %>%
  filter(Comments < 4500)

nrow(reduced_wrld)
```

```
## [1] 1185
```

- Taking 1200-1185, I figured out that worldnews only has 15 rows with comment values more than 4,500.

Cleaning Up Outliers- Sports News

```
reduced_sp <- sp_news %>%
  filter(Comments < 1000)

nrow(reduced_sp)
```

```
## [1] 1185
```

- Taking 1200-1185, I figured out that sports news only has 15 rows with comment values more than 1,000.

Cleaning Up Outliers- Television News

```
reduced_tv <- tv_news %>%
  filter(Comments < 2000)

nrow(reduced_tv)
```

```
## [1] 1187
```

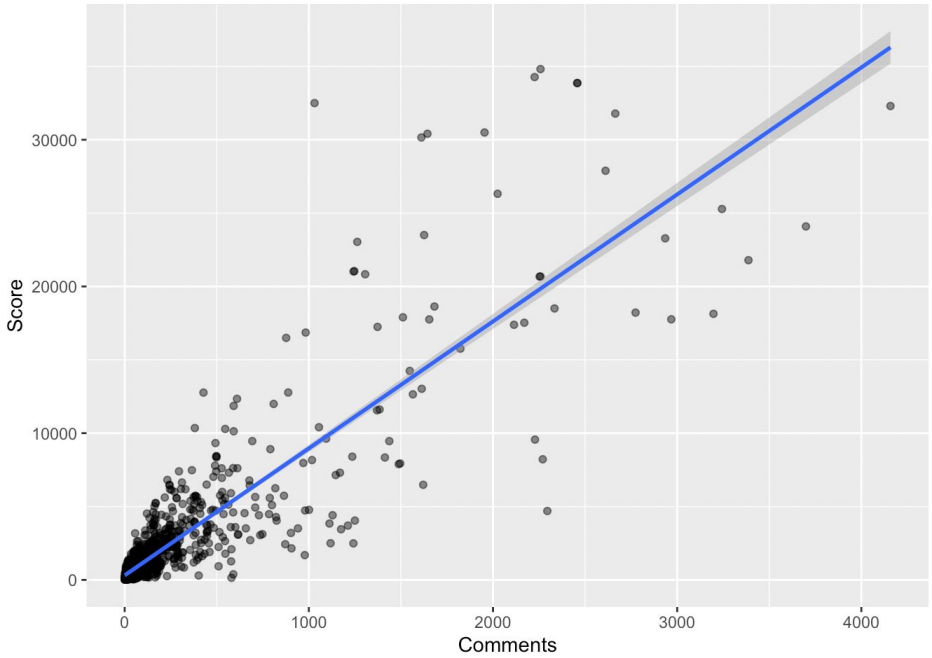
- Taking 1200-1187, I figured out that sports news only has 13 rows with comment values more than 2,000.

I graphed the reduced data.

Reduced Graph- Political News

```
ggplot(reduced_pol, aes(Comments, Score)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- The graph is better, but the comments and source look highly skewed.

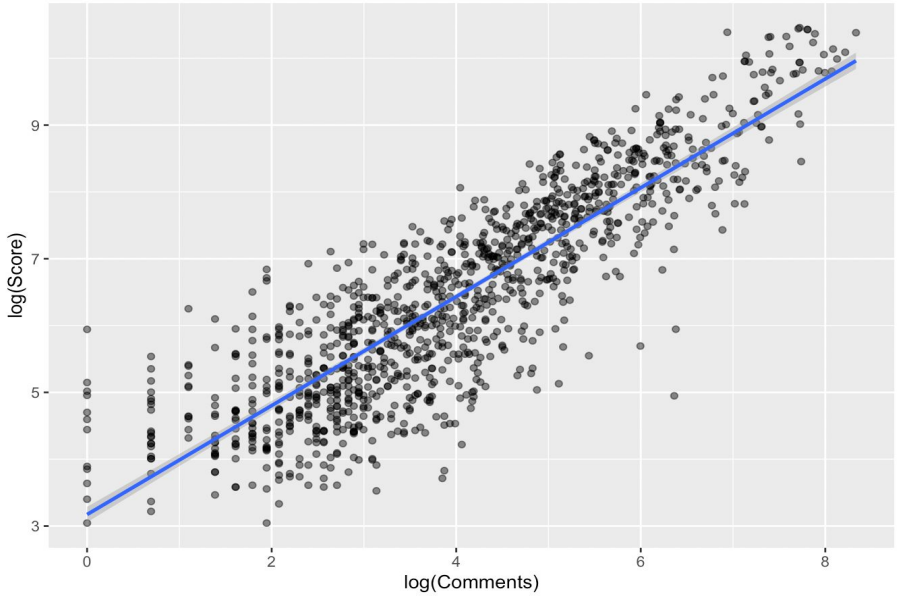
I applied log to create a percentage change.

Log Graph- Political News and Worldnews

- The points look consistent with the line of best fit

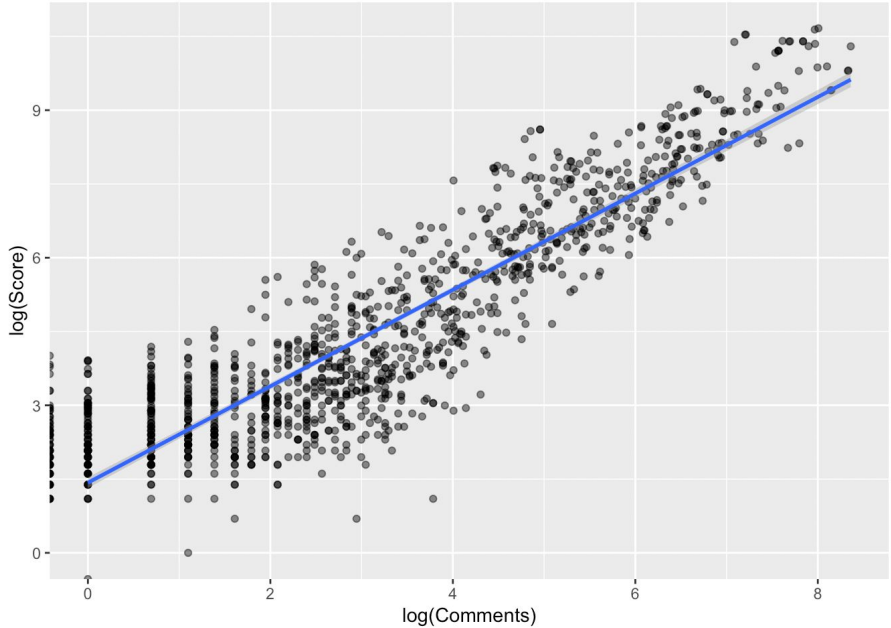
```
pol.log.lm <- lm(log(Score) ~ log(Comments), data = reduced_pol)

ggplot(reduced_pol, aes(log(Comments), log(Score))) +
geom_point(alpha = 0.5) +
geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(reduced_wrld, aes(log(Comments), log(Score))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

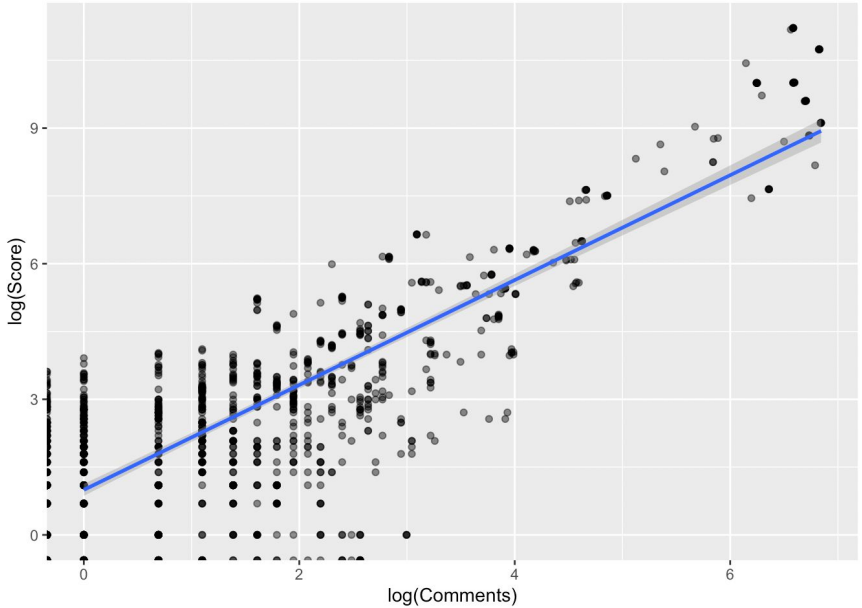


Log Graph- Sports News and Television News

- The points look a little spotty in places for these two graphs

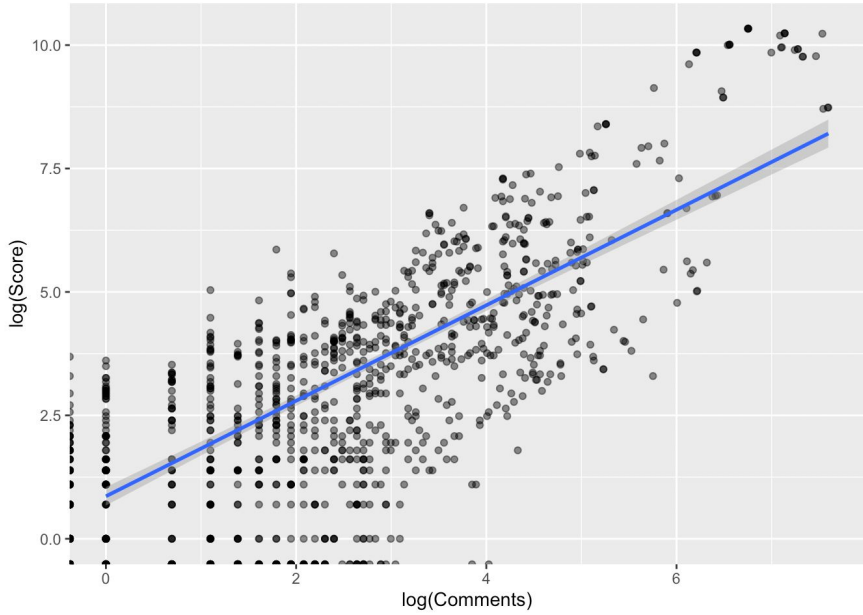
```
ggplot(reduced_sp, aes(log(Comments), log(Score))) +
geom_point(alpha = 0.5) +
geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(reduced_tv, aes(log(Comments), log(Score))) +
geom_point(alpha = 0.5) +
geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Now that we have that information, I am asking does the type of news affect the interaction of consumers and the score?

First, I combined the data frames into one.

Combined Dataframe

```
all <- bind_rows(list(politics = politics,worldnews = worldnews,sports = sports,television = television), .id =
"types") %>%
  rename(Comments = 'Number of Comments') %>%
  select(-...1)
all
```

```
## # A tibble: 4,800 × 6
##    types    Title                                Date       Time     Score Comments
##    <chr>    <chr>                                <date>     <time>   <dbl>    <dbl>
##  1 politics Megathread: Sean Spicer WWII gaf… 2017-04-12 00:51:25 14938     6461
##  2 politics Manafort Firm Received Ukraine L… 2017-04-12 15:12:43 11993      809
##  3 politics Its not too late to get rid of F… 2017-04-12 17:42:17  3119      748
##  4 politics Donald Trump, who doesnt read bo… 2017-04-12 14:14:59  5758      518
##  5 politics Did he or didnt he? Trump contra… 2017-04-12 16:52:49  2993      256
##  6 politics Trump lists Carter Page among hi… 2017-04-12 17:14:42  2018       58
##  7 politics Is Stephen Bannon getting pushed… 2017-04-12 18:17:28  1227      129
##  8 politics Silicon Valley is beginning to f… 2017-04-12 15:55:31  1823       84
##  9 politics Theres proof that ex-Trump campa… 2017-04-12 16:57:03  1370       20
## 10 politics TMZ catches Paul Ryan vacationin… 2017-04-12 17:50:41  1090       99
## # … with 4,790 more rows
```

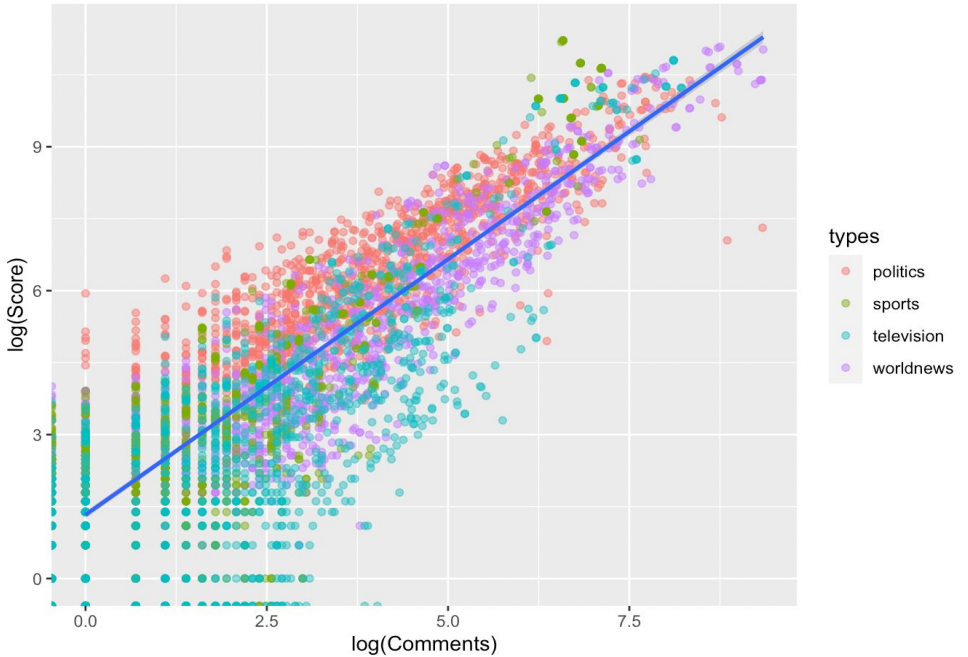
- I used the bind_rows function to create a new column called “types” and identify each value by their type of news.

I graphed the combined data frame.

Combined Graph

```
ggplot(all, aes(log(Comments), log(Score))) +
geom_point(alpha = 0.5, aes(color = types)) +
geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

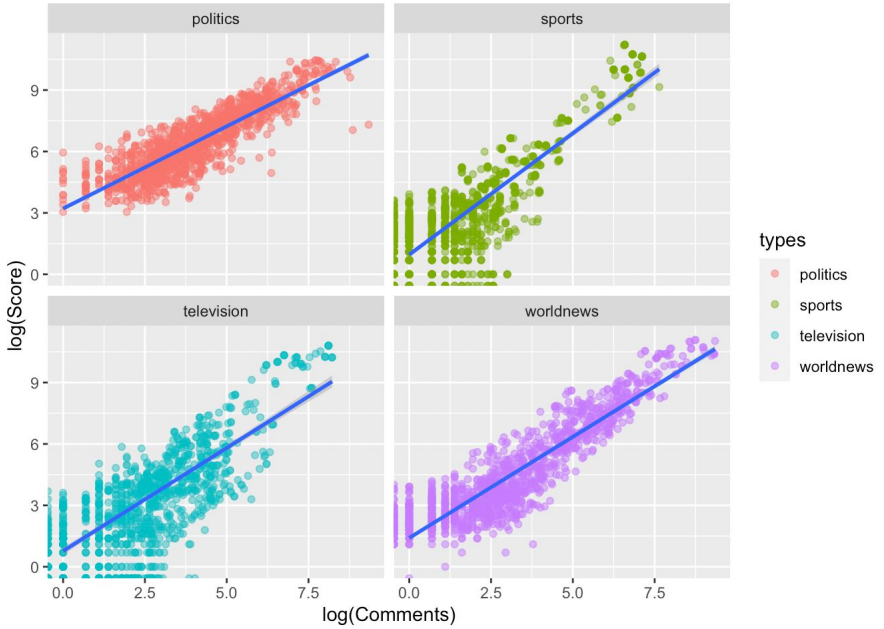


- This first graph shows all of the types combined and separated by color equal to types.
- I can already tell that politics and worldnews have more interaction than sports and television.

Combined Graph

```
ggplot(all, aes(log(Comments), log(Score))) +
geom_point(alpha = 0.5, aes(color = types)) +
geom_smooth(method = "lm") +
  facet_wrap(vars(types))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- Here I faceted the graph to see the values together more clearly.
- Again, politics and worldnews have more values and more consistency.
- This is probably because topics in politics and worldnews make people angrier and more expressive with their opinion. This data is based on solely comments. The data may be different if likes, shares, etc. were involved

Putting it all together

Creating the Data Frame

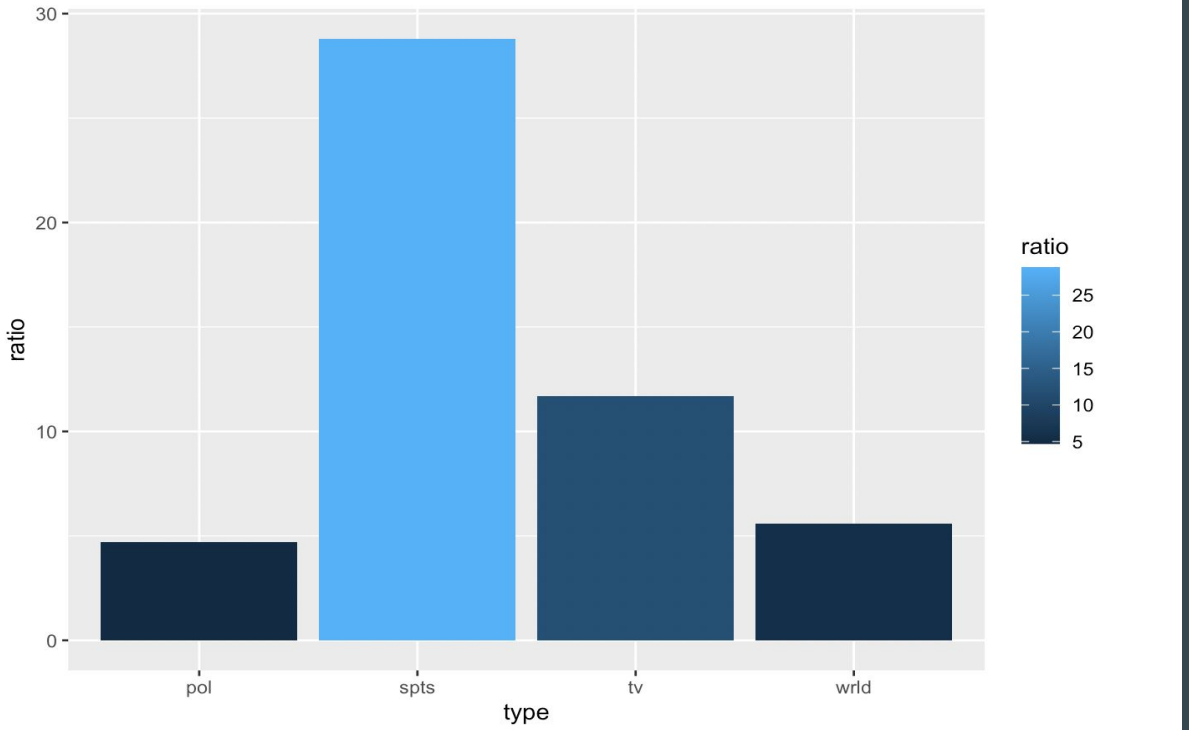
```
avg <- data.frame(type = c("pol", "wrld", "spts", "tv"), ratio = c(4.7, 5.6, 28.8, 11.7))
avg
```

```
##   type ratio
## 1  pol   4.7
## 2 wrld   5.6
## 3 spts  28.8
## 4   tv  11.7
```

- I created a data frame that just contained the types of data and their ratio values of comments to increase of score.

Graphing the Data Frame

```
ggplot(avg, aes(x = type, y = ratio)) +
  geom_bar(stat = "identity", aes(fill = ratio))
```



- From the graph, we can conclude that a smaller comments to increase in score ratio is equivalent to more interaction.