# CMDA3654PoliticalPropaganda

# 1. Description of Dataset:

The dataset contains social media data, specifically Facebook and Twitter data, that the Permanent Select Committee on Intelligence of the US House of Representatives deemed to be connected to a well-known Russian “troll” farm called the Internet Research Agency (IRA). The Facebook data provided advertisement text, data on ad social engagement, amount of money spent per advertisement, and other advertisement related data.

# 2. Interesting Features:

The most interesting features we have identified in the dataset are advertisements’ text, clicks, impressions, location, and the amount of money spent per advertisement.

# 3. Interesting Statistics:

## Money Spent

| Minimum | 1st Quarter | Median | Mean | 3rd Quarter | Maximum |
|---|---|---|---|---|---|
| 0.0 | 132.5 | 300.0 | 1985.8 | 680.2 3 | 31675.8 |

## Clicks

| Minimum | 1st Quarter | Median | Mean | 3rd Quarter | Maximum |
|---|---|---|---|---|---|
| 0.0 | 0.0 | 75.0 | 1079.6 | 863.5 | 73063.0 |

## Impressions

| Minimum | 1st Quarter | Median | Mean | 3rd Quarter | Maximum |
|---|---|---|---|---|---|
| 0.0 | 1.0 | 1090.0 | 11679.0 | 8544.0 | 1334544.0 |

# 4. Interesting Plot:

```
## Loading required package: NLP
```

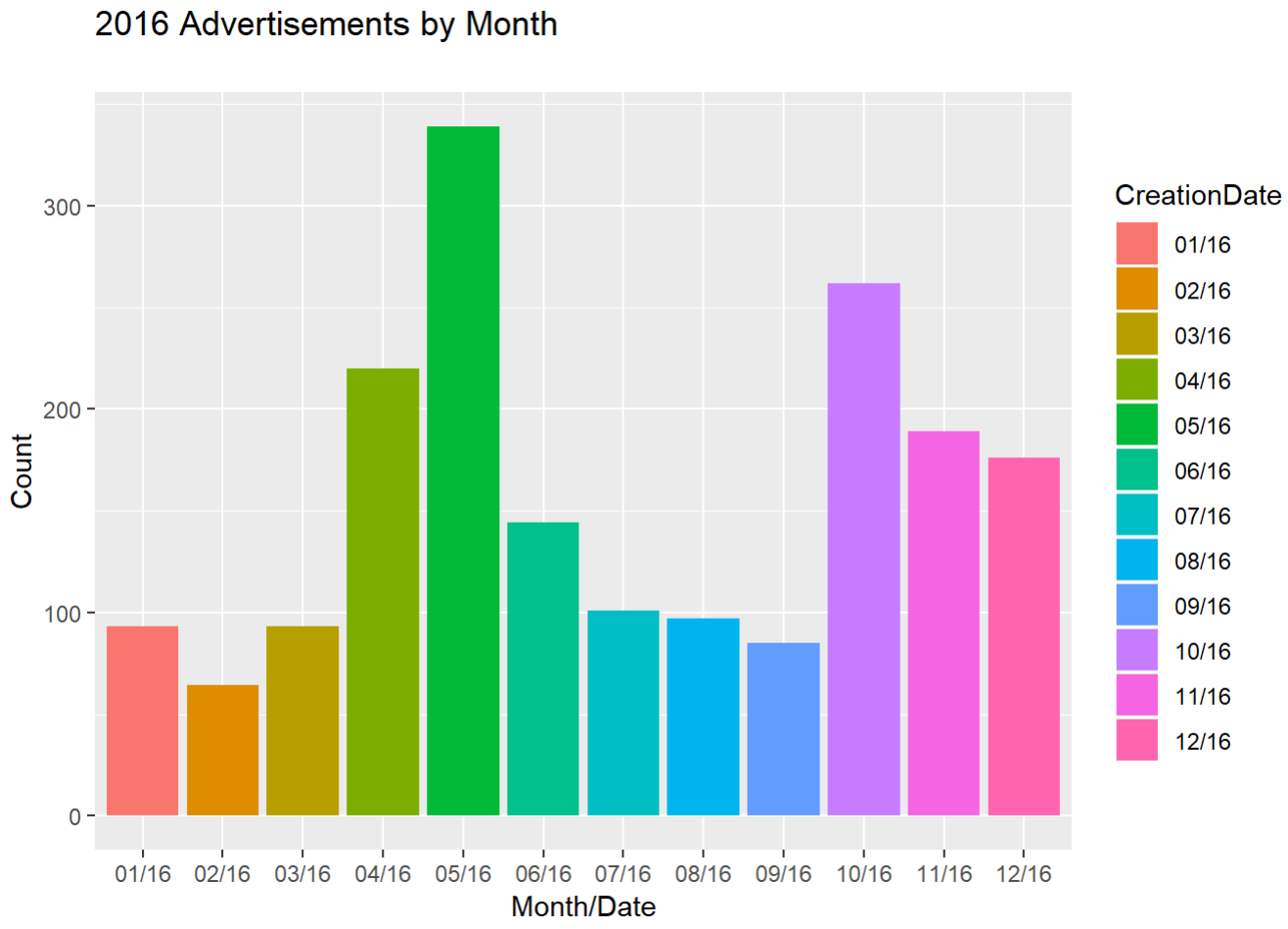
```
## Loading required package: RColorBrewer
```

```
##
## Attaching package: 'SentimentAnalysis'
```

```
## The following object is masked from 'package:base':
##
##     write
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```




This plot shows us how many advertisements were created per month during election year. There seems to have been a ramp up early in the year, then a dip, and another ramp up closer to and after the election. There is a lot of evidence that the advertisements were sponsored by foreign entities, so the continued creation of advertisements after the election may have been done to stir up trouble and create more political polarization. Plots for 2015 and 2017 are provided in the Appendix section at the end of the report.

# 5. Initial Approach:

- Create a new feature on states (extracting the state from the location feature)
- Creating a word cloud on advertisement text
- Clustering using K-means to determine the structure of Russian propaganda
- Scoring each of the words
- Scoring each of the ads
- Plot on a sentiment graphs
- Cluster the sentiment graphs
- Market basket on words and phrases that make people click on an advertisement
- Frequency bar chart of what time of day an advertisement was created and relating that to Moscow time
- Analyzing advertising spending habits and costs per ad as it relates to the election
- For ad targeting, analyzing interests and landing pages fo
- Analyzing landing pages of the top 25 social engaged ads to determine political affiliation of the landing page

# 6. Work Distribution:

- Chris Blair - Sentiment Analysis development, trying to take the adtext and do sentiment analysis on the data
- Numan Khan - Analyzing interests and landing pages to determine political affiliation of the ad i.e. which group is the ad targeting
- Sourav Panth - Market basket on words and phrases that make people click on an advertisement
- Caleb Notheis - Working on sentiment analysis and clustering with Chris
- Daniel - Analyzing the creation dates for the advertisements. This includes checking the number of advertisements created each month to see how they impacted elections and checking the time of day an advertisement was created to determine if the entities outside of US created these advertisements.
- Matt Dorris - Creating scatterplots that visually display the cost of each ad versus the number of clicks the ad got per year.Rmarkdown

# Appendix

# Word Cloud Ad Text

```
# Ad Text ----------------------------------------------------------------



corpus1 = VCorpus(VectorSource(HighestImpressions$AdText))

corpus1
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 100
```

```
wordcloud(corpus1, max.words=100, random.order = FALSE, colors="blue")
```


```
corpus1 <- tm_map(corpus1, removeWords, stopwords("english"))

wordcloud(corpus1, max.words=250, random.order = FALSE, colors="blue")
```

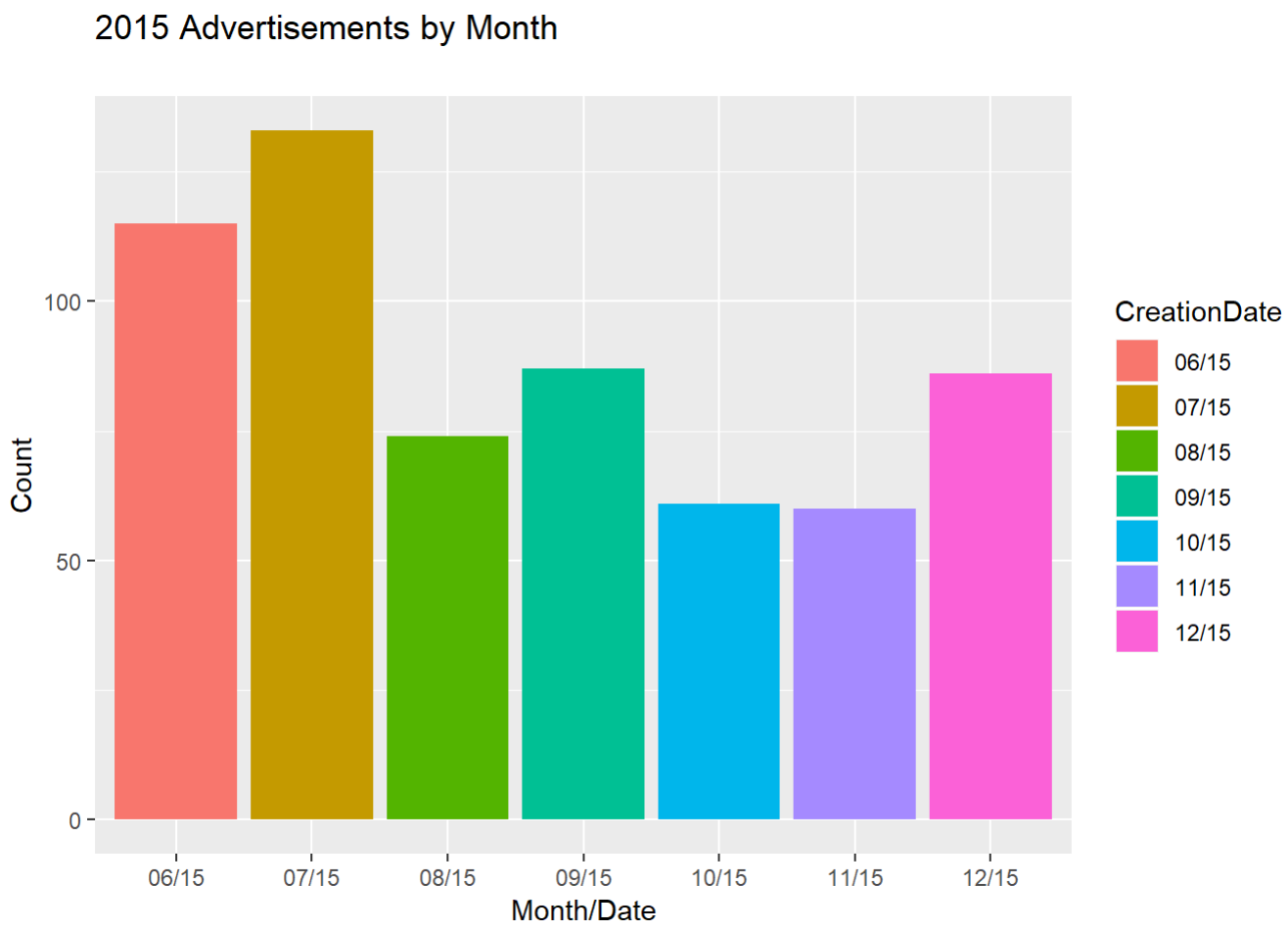

#

Location

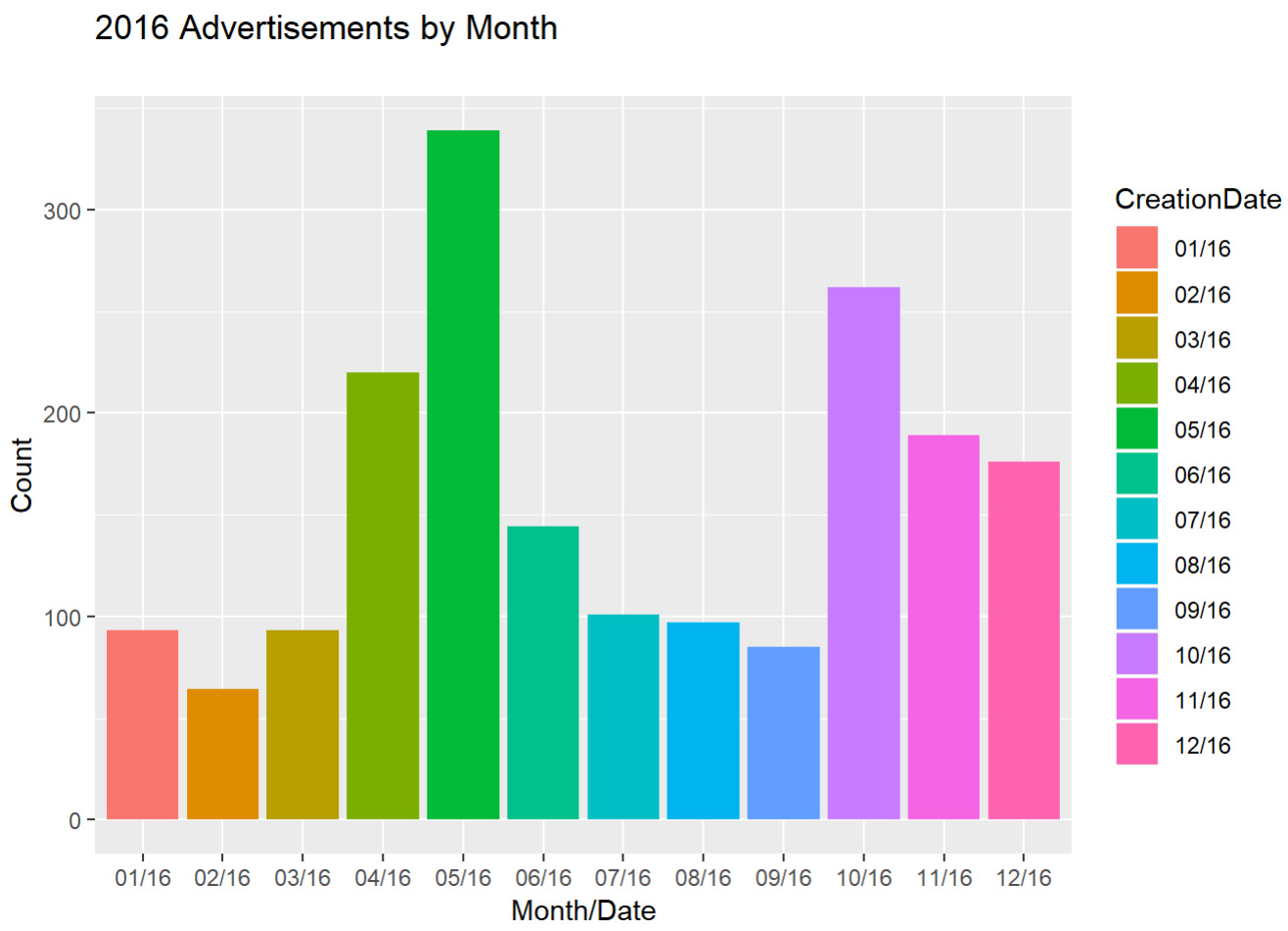
```
# Location ------------------------------------------------------------------



corpus2 = VCorpus(VectorSource(HighestImpressions$Location))

corpus1 <- tm_map(corpus1, removeWords, c("United", "States", "United States", "50",
                                          "20", "mi", "km", "living", "in", "living in"))

wordcloud(corpus2, max.words=100, random.order = FALSE, colors="blue")
```


# Advertisements Per Month by Year

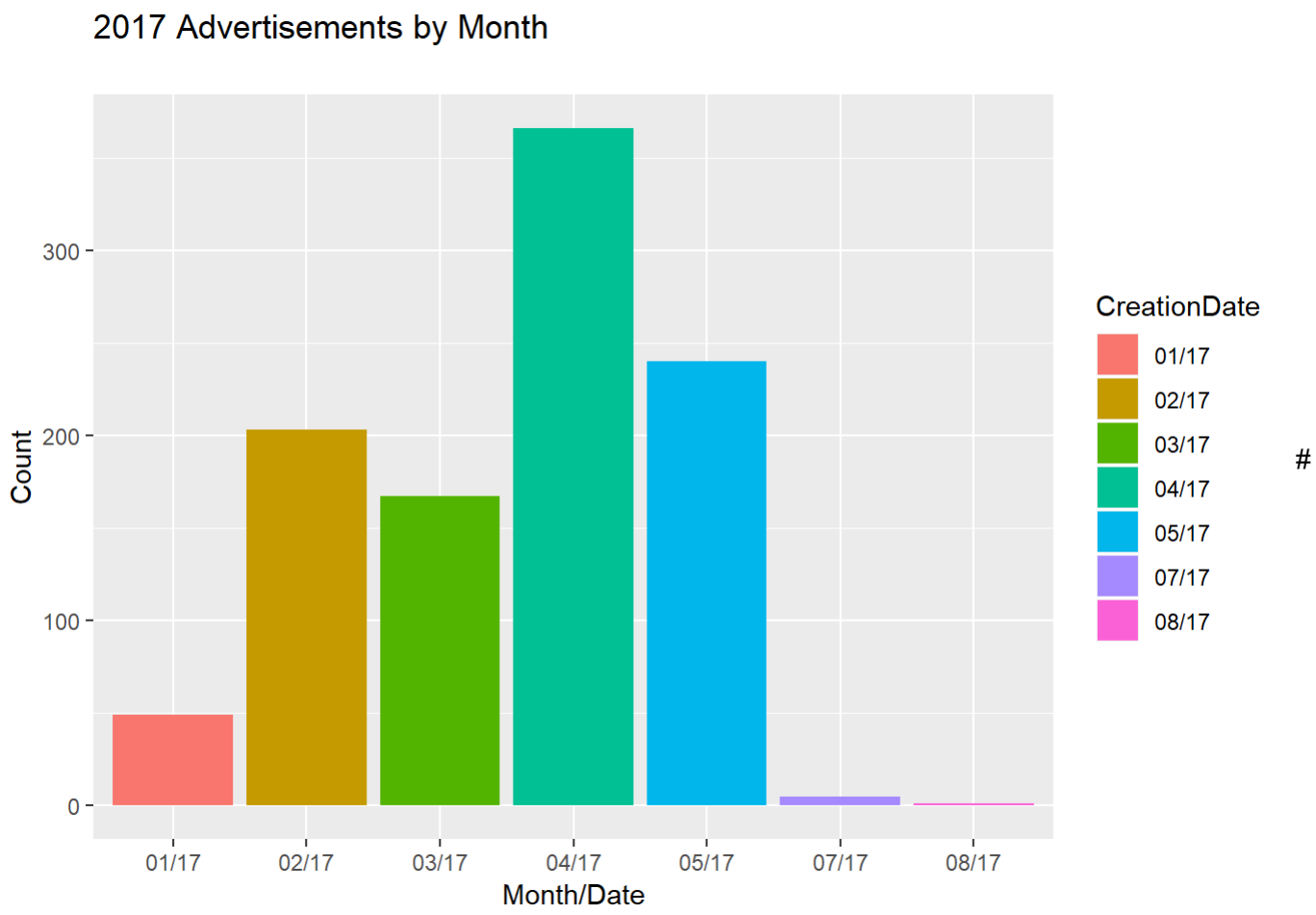
```
# Advertisements Per Month by Year----------------------------------------

shortDate <- dataset
shortDate$CreationDate <- paste(substr(shortDate$CreationDate, 1, 2), substr(shortDate$CreationD
ate, 7, 8), sep="/")

ads2015 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "15",]
ads2016 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "16",]
ads2017 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "17",]

ggplot(data = ads2015) +
  geom_bar(aes(x = CreationDate, fill = CreationDate)) +
  labs(title = "2015 Advertisements by Month\n", x = "Month/Date", y = "Count")
```




```
ggplot(data = ads2016) +
  geom_bar(aes(x = CreationDate, fill = CreationDate)) +
  labs(title = "2016 Advertisements by Month\n", x = "Month/Date", y = "Count")
```

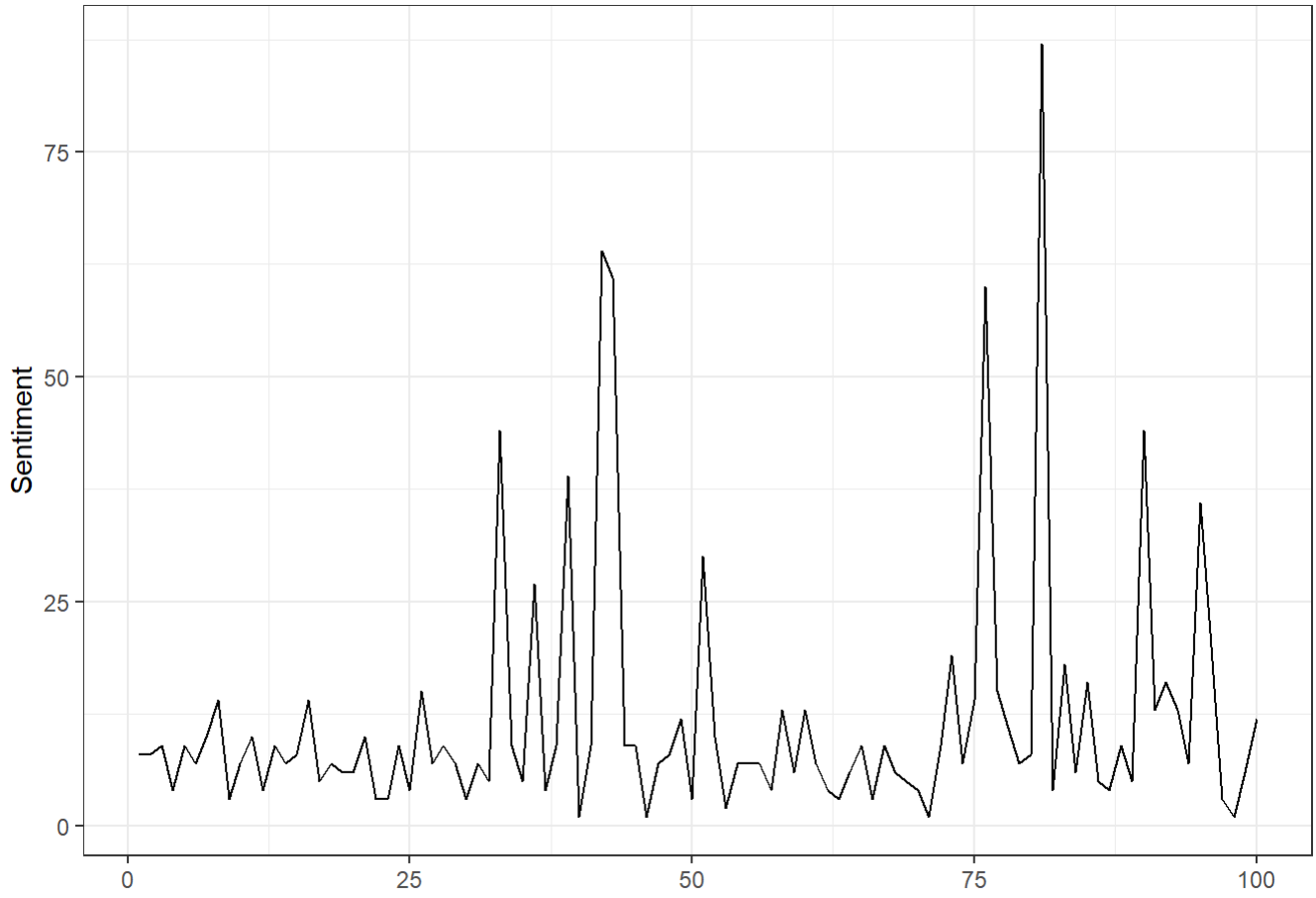



```
ggplot(data = ads2017) +
  geom_bar(aes(x = CreationDate, fill = CreationDate)) +
  labs(title = "2017 Advertisements by Month\n", x = "Month/Date", y = "Count")
```




Sentiment Analysis

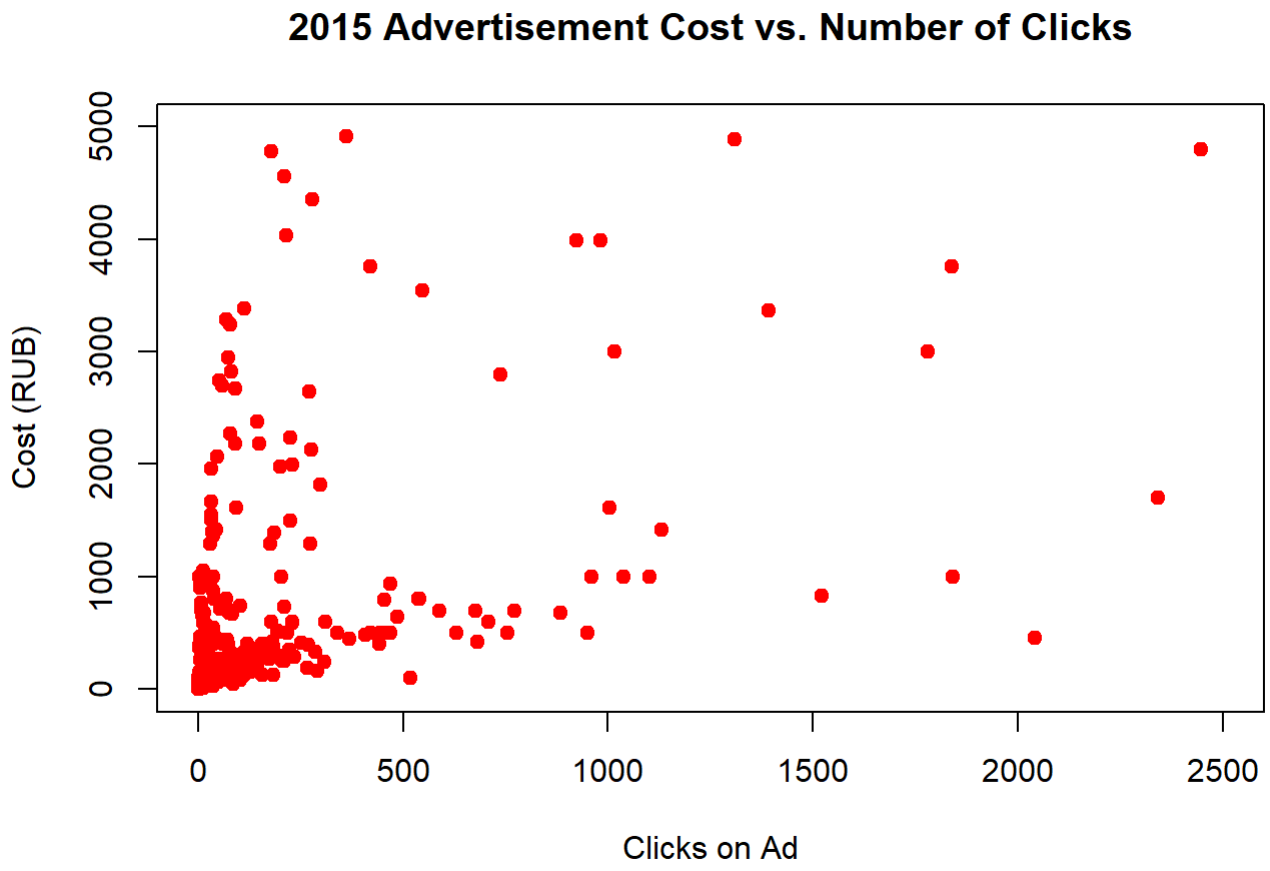
```
# Sentiment Analysis ------------------------------------------------------

corpus3 = VCorpus(VectorSource(HighestImpressions$AdText))


sentiment <- analyzeSentiment(corpus3, language = "english", aggregate = NULL,
                              removeStopwords = TRUE, stemming = TRUE)

plotSentiment(sentiment, x = NULL, cumsum = FALSE, xlab = "",
              ylab = "Sentiment")
```



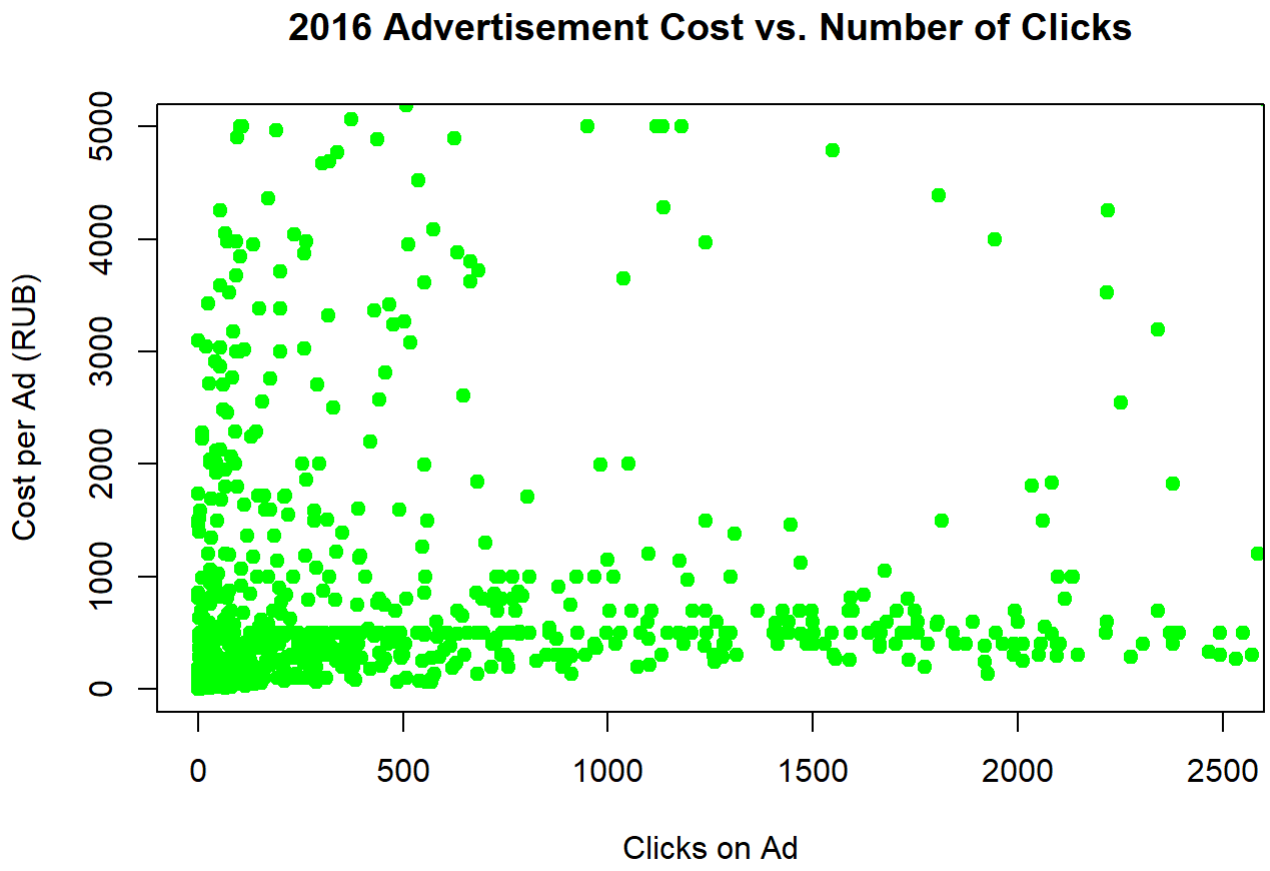

```
#plotSentimentResponse(sentiment, response, smoothing = "gam", xlab = "Sentiment", ylab = "Respo
nse")
```

# Ad Cost vs Clicks

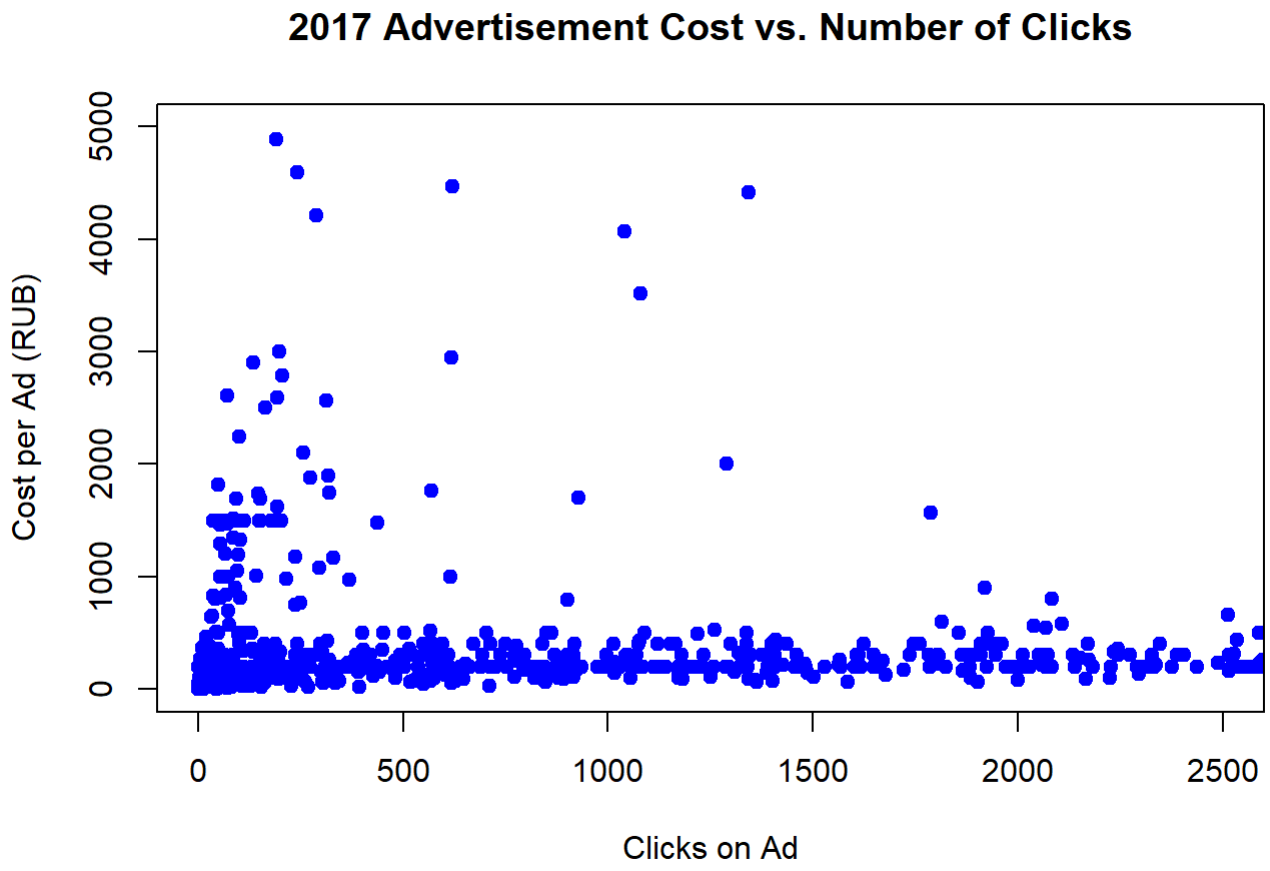
```
# Ad Cost vs Clicks ----------------------------------------

shortDate <- dataset
shortDate$CreationDate <- paste(substr(shortDate$CreationDate, 1, 2), substr(shortDate$CreationD
ate, 7, 8), sep="/")

ads2015 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "15",]
ads2016 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "16",]
ads2017 <- shortDate[substr(shortDate$CreationDate, 4, 5) == "17",]


plot(ads2015$Clicks, ads2015$AdSpend, col = "red", xlim=c(0, 2500),  ylim=c(0,5000), main="2015
 Advertisement Cost vs. Number of Clicks",
     xlab="Clicks on Ad", ylab="Cost (RUB)", pch=19)
```




```
plot(ads2016$Clicks, ads2016$AdSpend, col = "green", xlim=c(0, 2500),  ylim=c(0,5000), main="201
6 Advertisement Cost vs. Number of Clicks",
     xlab="Clicks on Ad", ylab="Cost per Ad (RUB)", pch=19)
```




```
plot(ads2017$Clicks, ads2017$AdSpend, col = "blue", xlim=c(0, 2500),  ylim=c(0,5000), main="2017
 Advertisement Cost vs. Number of Clicks",
     xlab="Clicks on Ad", ylab="Cost per Ad (RUB)", pch=19)
```


2017 Advertisement Cost vs. Number of Clicks
Cost per Ad (RUB)
0
1000
2000
3000
4000
5000
0
500
1000
1500
2000
2500
Clicks on Ad