EDA

Overview of Data

```
colnames(GDELT)
```

```
##  [1] "id"                   "originating_namespace"
##  [3] "originating_id"       "group_id"
##  [5] "actor1"               "actor2"
##  [7] "action"               "occurred"
##  [9] "source"               "created_at"
## [11] "modified_at"
```

```
head(GDELT)
```

```
##                id originating_namespace originating_id group_id
## 1 gdelt-861910158                 gdelt      861910158       NA
## 2 gdelt-861910190                 gdelt      861910190       NA
## 3 gdelt-861910221                 gdelt      861910221       NA
## 4 gdelt-861910245                 gdelt      861910245       NA
## 5 gdelt-861910281                 gdelt      861910281       NA
## 6 gdelt-861910285                 gdelt      861910285       NA
##
## 1
## 2 {'code': 'AFR', 'name': 'AFRICA', 'location': {'lat': -33.3042, 'lon': 26.5328, 'name': 'Grahamstou
## 3                        {'code': 'BLR', 'name': 'BELARUSIAN', 'location': {'lat': 41.707542, 'lon':
## 4                {'code': 'BUS', 'name': 'COMPANY', 'location': {'lat': 42.1497, 'lon': -74.9384, 'na
## 5      {'code': 'CANJUD', 'name': 'CANADA', 'location': {'lat': 48.3833, 'lon': -123.733, 'name': 'So
## 6                 {'code': 'CHR', 'name': 'CHRISTIAN', 'location': {'lat': 12.4904, 'lon': 11.4944,
##
## 1 {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 34.0983, 'lon': -118.327, 'name': 'Hol
## 2
## 3           {'code': 'UZB', 'name': 'UZBEKISTAN', 'location': {'lat': 53.9, 'lon': 27.5667, 'name':
## 4                {'code': 'GOV', 'name': 'AUTHORITIES', 'location': {'lat': 42.1497, 'lon': -74.9384
## 5           {'code': 'MIL', 'name': 'MILITARY', 'location': {'lat': 48.3833, 'lon': -123.733, 'name'
## 6                     {'code': 'GOV', 'name': 'PRESIDENT', 'location': {'lat': 12.4904, 'lon': 11.494
##
## 1  {'code': '019', 'location': {'lat': 34.0983, 'lon': -118.327, 'name': 'Hollywood, California, Unit
## 2 {'code': '036', 'location': {'lat': -33.3042, 'lon': 26.5328, 'name': 'Grahamstown, Eastern Cape, S
## 3                             {'code': '084', 'location': {'lat': 41.707542, 'lon': 63.84911, 'name':
## 4                {'code': '051', 'location': {'lat': 42.1497, 'lon': -74.9384, 'name': 'New York, Unit
## 5        {'code': '112', 'location': {'lat': 48.3833, 'lon': -123.733, 'name': 'Sooke, British Columb
## 6                {'code': '0253', 'location': {'lat': 12.4904, 'lon': 11.4944, 'name': 'Dapchi, Yobe
##              occurred
## 1 2019-07-26 17:30:00
## 2 2019-07-26 17:30:00
## 3 2019-07-26 17:30:00
## 4 2019-07-26 17:30:00
## 5 2019-07-26 17:30:00
## 6 2019-07-26 17:30:00
##
## 1                              https://www.latintimes.com/angelina-jolie-spill-ugly-details-about-brac
```

```
## 2                                                    https://www.grocotts.co.za/2019/07/26/kc-concert-
## 3      https://eng.belta.by/society/view/belarusian-encyclopedia-publishing-house-releases-about-war-
## 4 https://www.dailymail.co.uk/news/article-7290563/Jeffrey-Epsteins-famous-guests-Lolita-Express-expo
## 5                   https://www.sookenewsmirror.com/news/supreme-court-of-canada-says-militarys-no-ju
## 6                                                      https://dailypost.ng/2019/07/26/leah-sharibu
##            created_at         modified_at
## 1 2019-07-29 20:21:18 2019-07-29 20:21:18
## 2 2019-07-29 20:21:18 2019-07-29 20:21:18
## 3 2019-07-29 20:21:18 2019-07-29 20:21:18
## 4 2019-07-29 20:21:18 2019-07-29 20:21:18
## 5 2019-07-29 20:21:19 2019-07-29 20:21:19
## 6 2019-07-29 20:21:19 2019-07-29 20:21:19
```

```
summary(GDELT)
```

```
##                id         originating_namespace originating_id
##  gdelt-861910068:     1   gdelt:500000          Min.   :861910068
##  gdelt-861910071:     1                         1st Qu.:863329951
##  gdelt-861910074:     1                         Median :864774108
##  gdelt-861910110:     1                         Mean   :864774628
##  gdelt-861910114:     1                         3rd Qu.:866198918
##  gdelt-861910118:     1                         Max.   :867661171
##  (Other)        :499994
##  group_id
##  Mode:logical
##  NA's:500000
##
##
##
##
##
##
##
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 38.8951, 'lon': -77.0364, 'name': 'Wash
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 39.828175, 'lon': -98.5795, 'name': 'Un
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 42.1497, 'lon': -74.9384, 'name': 'New
##  {'code': 'GBR', 'name': 'UNITED KINGDOM', 'location': {'lat': 54, 'lon': -4, 'name': 'United Kingdom
##  {'code': 'CHN', 'name': 'CHINA', 'location': {'lat': 35, 'lon': 105, 'name': 'China', 'country': 'CH
##  (Other)
##
##
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 38.8951, 'lon': -77.0364, 'name': 'Wash
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 39.828175, 'lon': -98.5795, 'name': 'Un
##  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 42.1497, 'lon': -74.9384, 'name': 'New
##  {'code': 'CHN', 'name': 'CHINA', 'location': {'lat': 35, 'lon': 105, 'name': 'China', 'country': 'CH
##  {'code': 'GBR', 'name': 'UNITED KINGDOM', 'location': {'lat': 54, 'lon': -4, 'name': 'United Kingdom
##  (Other)
##
##  {'code': '010', 'location': {'lat': None, 'lon': None, 'name': '', 'country': ''}, 'base_code': '010
##  {'code': '020', 'location': {'lat': None, 'lon': None, 'name': '', 'country': ''}, 'base_code': '020
##  {'code': '042', 'location': {'lat': None, 'lon': None, 'name': '', 'country': ''}, 'base_code': '042
##  {'code': '043', 'location': {'lat': None, 'lon': None, 'name': '', 'country': ''}, 'base_code': '043
##  {'code': '193', 'location': {'lat': 31.7587, 'lon': -106.487, 'name': 'El Paso, Texas, United States
##  {'code': '010', 'location': {'lat': 39.828175, 'lon': -98.5795, 'name': 'United States', 'country':
##  (Other)
```

```
##     occurred
##  Min.   :2019-07-26 17:30:00
##  1st Qu.:2019-08-02 09:15:00
##  Median :2019-08-09 03:15:00
##  Mean   :2019-08-09 13:26:13
##  3rd Qu.:2019-08-16 06:15:00
##  Max.   :2019-08-22 21:00:00
##
##
##  https://www.trtworld.com/middle-east/us-issues-warrant-to-seize-grace-1-gulf-tensions-28056
##  https://www.trtworld.com/middle-east/gibraltar-rejects-us-pressure-to-hold-iranian-oil-tanker-gulf-
##  https://www.trtworld.com/middle-east/iranian-tanker-heads-to-unknown-destination-gulf-tensions-2805
##  https://www.trtworld.com/middle-east/gibraltar-allows-iranian-tanker-to-leave-gulf-tensions-gulf-te
##  https://www.trtworld.com/middle-east/greece-says-it-won-t-assist-iranian-tanker-sought-by-us-gulf-t
##  https://www.trtworld.com/middle-east/greece-says-no-request-from-iran-tanker-to-dock-gulf-tensions-
##  (Other)
##    created_at                  modified_at
##  Min.   :2019-07-29 20:21:18   Min.   :2019-07-29 20:21:18
##  1st Qu.:2019-08-02 09:10:24   1st Qu.:2019-08-02 09:10:24
##  Median :2019-08-09 03:10:06   Median :2019-08-09 03:10:06
##  Mean   :2019-08-09 16:56:10   Mean   :2019-08-09 16:56:10
##  3rd Qu.:2019-08-16 06:09:13   3rd Qu.:2019-08-16 06:09:13
##  Max.   :2019-08-22 20:58:47   Max.   :2019-08-22 20:58:47
##
```

```
# Foreign sources (no '.com', '.org')
nrow(GDELT[!grepl(".com", GDELT$source),])
```

```
## [1] 119296
```

```
foreign <- GDELT[!grepl(".com|.org", GDELT$source),]
head(foreign$source, 10)
```

```
##  [1] https://www.grocotts.co.za/2019/07/26/kc-concert-band-gears-up-for-western-cape-tour/
##  [2] https://eng.belta.by/society/view/belarusian-encyclopedia-publishing-house-releases-about-war-h
##  [3] https://www.dailymail.co.uk/news/article-7290563/Jeffrey-Epsteins-famous-guests-Lolita-Express-
##  [4] https://dailypost.ng/2019/07/26/leah-sharibu-can-breaks-silence-rumoured-death/
##  [5] https://www.independent.ie/breaking-news/irish-news/taoiseach-nodeal-brexit-could-lead-to-suppo
##  [6] http://dunyanews.tv/en/Pakistan/502254-PM-Imran-Turkish-President-Erdogan-speak-over-phone
##  [7] http://dunyanews.tv/en/Pakistan/502254-PM-Imran-Turkish-President-Erdogan-speak-over-phone
##  [8] https://www.independent.ie/breaking-news/irish-news/taoiseach-nodeal-brexit-could-lead-to-suppo
##  [9] https://www.independent.ie/breaking-news/irish-news/taoiseach-nodeal-brexit-could-lead-to-suppo
## [10] https://www.dailymail.co.uk/news/article-7290563/Jeffrey-Epsteins-famous-guests-Lolita-Express-
## 323194 Levels: http://1037thegame.com/stipe-miocic-takes-down-daniel-cormier-to-take-back-ufc-title/
```

```
# Checking whether action contains the information from the actor columns (code, name, location)
select(GDELT, actor1, actor2, action) %>% head(10)
```

```
##
## 1
## 2       {'code': 'AFR', 'name': 'AFRICA', 'location': {'lat': -33.3042, 'lon': 26.5328, 'name': 'Gra
## 3                             {'code': 'BLR', 'name': 'BELARUSIAN', 'location': {'lat': 41.707542,
## 4                       {'code': 'BUS', 'name': 'COMPANY', 'location': {'lat': 42.1497, 'lon': -74.93
## 5            {'code': 'CANJUD', 'name': 'CANADA', 'location': {'lat': 48.3833, 'lon': -123.733, 'nam
## 6                        {'code': 'CHR', 'name': 'CHRISTIAN', 'location': {'lat': 12.4904, 'lon': 11.
## 7                               {'code': 'COP', 'name': 'POLICE', 'location': {'lat': 39.828175, 'lo
```

```
## 8                  {'code': 'CVL', 'name': 'RESIDENTS', 'location': {'lat': 42.8617, 'lon': 75.1391
## 9                                  {'code': 'GBR', 'name': 'UNITED KINGDOM', 'location': {'lat': 54
## 10 {'code': 'GBR', 'name': 'UNITED KINGDOM', 'location': {'lat': 54.5, 'lon': -6.5, 'name': 'Northern
##
## 1  {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 34.0983, 'lon': -118.327, 'name': 'H
## 2
## 3           {'code': 'UZB', 'name': 'UZBEKISTAN', 'location': {'lat': 53.9, 'lon': 27.5667, 'name':
## 4               {'code': 'GOV', 'name': 'AUTHORITIES', 'location': {'lat': 42.1497, 'lon': -74.9384
## 5           {'code': 'MIL', 'name': 'MILITARY', 'location': {'lat': 48.3833, 'lon': -123.733, 'name
## 6                    {'code': 'GOV', 'name': 'PRESIDENT', 'location': {'lat': 12.4904, 'lon': 11.49
## 7                     {'code': 'USA', 'name': 'UNITED STATES', 'location': {'lat': 39.828175, 'lon
## 8                   {'code': 'KGZ', 'name': 'KYRGYZ', 'location': {'lat': 42.8617, 'lon': 75.1391,
## 9
## 10
##
## 1   {'code': '019', 'location': {'lat': 34.0983, 'lon': -118.327, 'name': 'Hollywood, California, Uni
## 2  {'code': '036', 'location': {'lat': -33.3042, 'lon': 26.5328, 'name': 'Grahamstown, Eastern Cape,
## 3                           {'code': '084', 'location': {'lat': 41.707542, 'lon': 63.84911, 'name'
## 4               {'code': '051', 'location': {'lat': 42.1497, 'lon': -74.9384, 'name': 'New York, Uni
## 5        {'code': '112', 'location': {'lat': 48.3833, 'lon': -123.733, 'name': 'Sooke, British Colum
## 6                {'code': '0253', 'location': {'lat': 12.4904, 'lon': 11.4944, 'name': 'Dapchi, Yobe
## 7                       {'code': '180', 'location': {'lat': 39.828175, 'lon': -98.5795, 'name': 'Un
## 8             {'code': '192', 'location': {'lat': 42.8617, 'lon': 75.1391, 'name': 'Cholpon, Chü3, K
## 9              {'code': '012', 'location': {'lat': 54.65, 'lon': -8.11667, 'name': 'Donegal, Donegal
## 10      {'code': '044', 'location': {'lat': 54.5833, 'lon': -5.93333, 'name': 'Belfast, Belfast, Unit
```

```
total <- 0
for (row in GDELT) {
    ifelse (
        ((str_extract(GDELT$actor1[row], "'lat': .{7}") == str_extract(GDELT$action[row], "'lat': .{7}"
        (str_extract(GDELT$actor1[row], "'lon': .{7}") == str_extract(GDELT$action[row], "'lon': .{7}")
        ((str_extract(GDELT$actor2[row], "'lat': .{7}") == str_extract(GDELT$action[row], "'lat': .{7}"
        (str_extract(GDELT$actor2[row], "'lon': .{7}") == str_extract(GDELT$action[row], "'lon': .{7}")
        total <- total,
        total <- total+1
        )
}
# Number of rows where actors location does not match action location
total
```

```
## [1] 5
```

```
# Example
GDELT[181089,5:7]
```

```
##
## 181089 {'code': 'GOV', 'name': 'PRESIDENT', 'location': {'lat': 39.7589, 'lon': -84.1916, 'name': 'Da
##
## 181089 {'code': 'USAGOV', 'name': 'THE WHITE HOUSE', 'location': {'lat': 38.8951, 'lon': -77.0364, 'n
##
## 181089 {'code': '193', 'location': {'lat': 40.3736, 'lon': -82.7755, 'name': 'Ohio, United States',
```

```
colnames(articles1)
```

```
##  [1] "X1"          "id"          "title"       "publication" "author"
##  [6] "date"        "year"        "month"       "url"         "content"
```

```
head(articles1)
```

```
## # A tibble: 6 x 10
##      X1    id title publication author date        year month url   conte~
##   <int> <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr> <chr>
## 1     0 17283 Hous~ New York T~ Carl ~ 2016-12-31  2016    12 <NA>  WASHI~
## 2     1 17284 Rift~ New York T~ Benja~ 2017-06-19  2017     6 <NA>  After~
## 3     2 17285 Tyru~ New York T~ Marga~ 2017-01-06  2017     1 <NA>  When ~
## 4     3 17286 Amon~ New York T~ Willi~ 2017-04-10  2017     4 <NA>  Death~
## 5     4 17287 Kim ~ New York T~ Choe ~ 2017-01-02  2017     1 <NA>  SEOUL~
## 6     5 17288 Sick~ New York T~ Sewel~ 2017-01-02  2017     1 <NA>  LONDO~
```

```
summary(articles1)
```

```
##        X1              id           title           publication
##  Min.   :    0   Min.   :17283   Length:50000       Length:50000
##  1st Qu.:12501   1st Qu.:31237   Class :character   Class :character
##  Median :25004   Median :43758   Mode  :character   Mode  :character
##  Mean   :25694   Mean   :44432
##  3rd Qu.:38630   3rd Qu.:57479
##  Max.   :53291   Max.   :73469
##     author               date                year          month
##  Length:50000       Min.   :2011-11-22   Min.   :2011   Min.   : 1.000
##  Class :character   1st Qu.:2016-05-29   1st Qu.:2016   1st Qu.: 3.000
##  Mode  :character   Median :2016-10-10   Median :2016   Median : 5.000
##                     Mean   :2016-09-08   Mean   :2016   Mean   : 5.509
##                     3rd Qu.:2017-02-15   3rd Qu.:2017   3rd Qu.: 8.000
##                     Max.   :2017-06-21   Max.   :2017   Max.   :12.000
##      url              content
##  Length:50000       Length:50000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
colnames(articles2)
```

```
##  [1] "X1"          "id"          "title"       "publication" "author"
##  [6] "date"        "year"        "month"       "url"         "content"
```

```
head(articles2)
```

```
## # A tibble: 6 x 10
##      X1    id title publication author date        year month url   conte~
##   <int> <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr> <chr>
## 1 53293 73471 Patr~ Atlantic    David~ 2017-01-11  2017     1 <NA>  Patri~
## 2 53294 73472 A Br~ Atlantic    Ed Yo~ 2017-01-11  2017     1 <NA>  In No~
## 3 53295 73474 Obam~ Atlantic    Spenc~ 2017-01-11  2017     1 <NA>  "If o~
## 4 53296 73475 Dona~ Atlantic    David~ 2017-01-11  2017     1 <NA>  Updat~
## 5 53297 73476 Trum~ Atlantic    Kaveh~ 2017-01-11  2017     1 <NA>  Updat~
## 6 53298 73477 Seth~ Atlantic    Megan~ 2017-01-11  2017     1 <NA>  Here ~
```

```
summary(articles2)
```

```
##        X1              id            title           publication
##  Min.   : 53293   Min.   : 73471   Length:49999       Length:49999
```

```
##  1st Qu.: 65820   1st Qu.: 90423   Class :character   Class :character
##  Median : 78450   Median :119047   Mode  :character   Mode  :character
##  Mean   : 78408   Mean   :114517
##  3rd Qu.: 90958   3rd Qu.:135805
##  Max.   :103457   Max.   :151906
##
##     author              date                 year          month
##  Length:49999       Min.   :2003-06-14   Min.   :2003   Min.   : 1.000
##  Class :character   1st Qu.:2016-05-19   1st Qu.:2016   1st Qu.: 3.000
##  Mode  :character   Median :2016-10-21   Median :2016   Median : 5.000
##                     Mean   :2016-10-09   Mean   :2016   Mean   : 5.242
##                     3rd Qu.:2017-03-01   3rd Qu.:2017   3rd Qu.: 8.000
##                     Max.   :2017-07-01   Max.   :2017   Max.   :12.000
##                     NA's   :2626         NA's   :2626   NA's   :2626
##      url                content
##  Length:49999       Length:49999
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
colnames(articles3)
```

```
##  [1] "X1"          "id"          "title"       "publication" "author"
##  [6] "date"        "year"        "month"       "url"         "content"
```

```
head(articles3)
```

```
## # A tibble: 6 x 10
##       X1     id title publication author date        year month url
##    <int>  <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr>
## 1 103459 151908 Alto~ Guardian    Jessi~ 2016-07-13  2016     7 http~
## 2 103460 151909 Shak~ Guardian    <NA>   2016-05-25  2016     5 http~
## 3 103461 151910 My g~ Guardian    Rober~ 2016-10-31  2016    10 http~
## 4 103462 151911 I fe~ Guardian    Bradf~ 2016-11-26  2016    11 http~
## 5 103463 151912 Texa~ Guardian    <NA>   2016-08-20  2016     8 http~
## 6 103464 151914 My d~ Guardian    Steve~ 2016-11-28  2016    11 http~
## # ... with 1 more variable: content <chr>
```

```
summary(articles3)
```

```
##        X1               id            title           publication
##  Min.   :103459   Min.   :151908   Length:42571       Length:42571
##  1st Qu.:114104   1st Qu.:168960   Class :character   Class :character
##  Median :124747   Median :186322   Mode  :character   Mode  :character
##  Mean   :124746   Mean   :186227
##  3rd Qu.:135390   3rd Qu.:204620
##  Max.   :146032   Max.   :218082
##
##     author              date                 year          month
##  Length:42571       Min.   :2000-05-15   Min.   :2000   Min.   : 1.000
##  Class :character   1st Qu.:2016-06-08   1st Qu.:2016   1st Qu.: 3.000
##  Mode  :character   Median :2016-10-21   Median :2016   Median : 5.000
##                     Mean   :2016-10-07   Mean   :2016   Mean   : 5.807
```

```
##                     3rd Qu.:2017-02-16   3rd Qu.:2017   3rd Qu.: 9.000
##                     Max.   :2017-07-06   Max.   :2017   Max.   :12.000
##                     NA's   :4962         NA's   :15     NA's   :15
##      url                content
##  Length:42571       Length:42571
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
# Looking at the data in articles3 where title is "Premier League: 10 things to look out for this weeke
articles3[grepl("Premier League: 10 things to look out for this weekend", articles3$title),]
```

```
## # A tibble: 3 x 10
##       X1     id title publication author date        year month url
##    <int>  <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr>
## 1 104653 153464 Prem~ Guardian    Paul ~ 2016-12-09  2016    12 http~
## 2 104877 153765 Prem~ Guardian    Jacob~ 2016-09-30  2016     9 http~
## 3 105199 154191 Prem~ Guardian    Paul ~ 2016-11-04  2016    11 http~
## # ... with 1 more variable: content <chr>
```

```
# articles are different despit having same title

dups <- articles3[duplicated(articles3$title) | duplicated(articles3$title, fromLast = T),]
dups[order(dups$title),] # some content is the same, some different
```

```
## # A tibble: 32 x 10
##        X1     id title publication author date        year month url
##     <int>  <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr>
##  1 130048 195554 Gold~ Vox         Alex ~ NA          2017     1 http~
##  2 134733 203663 Gold~ Vox         Alex ~ NA          2016    12 http~
##  3 110122 164219 Heav~ NPR         <NA>   2017-04-29  2017     4 http~
##  4 114899 169957 Heav~ NPR         <NA>   2016-06-14  2016     6 http~
##  5 115646 170895 Heav~ NPR         <NA>   2016-07-20  2016     7 http~
##  6 117127 172743 Heav~ NPR         <NA>   2016-09-24  2016     9 http~
##  7 117931 173717 Heav~ NPR         <NA>   2016-10-29  2016    10 http~
##  8 118681 174670 Heav~ NPR         <NA>   2016-12-04  2016    12 http~
##  9 137059 206686 Hunt~ Washington~ DeNee~ 2017-04-11  2017     4 "htt~
## 10 137446 207179 Hunt~ Washington~ DeNee~ 2017-05-01  2017     5 "htt~
## # ... with 22 more rows, and 1 more variable: content <chr>
```

```
# Creating new dataframe combining articles dataframes with columns of interest
dates1 <- articles1$date %>% as.data.frame()
dates1$publication <- articles1$publication
dates1$title <- articles1$title
dates1$from <- "articles1"
dates2 <- articles2$date %>% as.data.frame()
dates2$publication <- articles2$publication
dates2$title <- articles2$title
dates2$from <- "articles2"
dates3 <- articles3$date %>% as.data.frame()
dates3$publication <- articles3$publication
dates3$title <- articles3$title
dates3$from <- "articles3"
```

```r
dates <- rbind(dates1, dates2, dates3)
colnames(dates)[1] <- "date"
# Set 'date' column to date format
dates$date <- as.Date(dates$date)
```

```r
# All duplicated titles
dups <- dates[duplicated(dates$title) | duplicated(dates$title, fromLast = T),]
dups <- dups[order(dups$title),] # sort by title for easier comparison
head(dups) # dates vary, likely some of the content does too looking at above for articles3
```

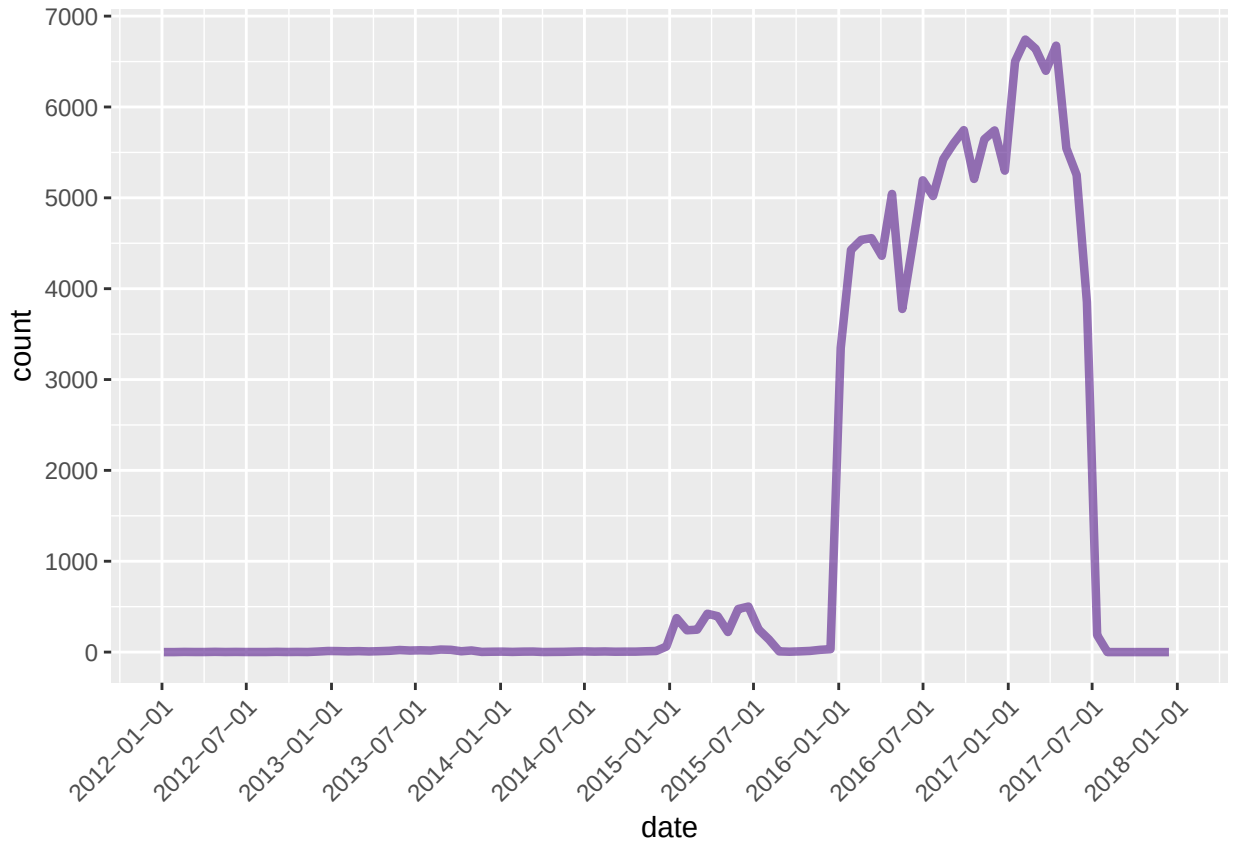
```
##             date      publication
## 44243 2017-04-03 Business Insider
## 48821 2016-10-09 Business Insider
## 49137 2016-11-03 Business Insider
## 66863 2017-03-31    Buzzfeed News
## 66903 2017-04-01    Buzzfeed News
## 69972 2016-08-12    Buzzfeed News
##                                                                                                  title
## 44243                                                                                             \n\n
## 48821                                                                                             \n\n
## 49137                                                                                             \n\n
## 66863 "Never Give Up": Trans People Share Messages Of Love And Support For Trans Day Of Visibility
## 66903 "Never Give Up": Trans People Share Messages Of Love And Support For Trans Day Of Visibility
## 69972                "A Honeypot For Assholes": Inside Twitter's 10-Year Failure To Stop Harassment
##            from
## 44243 articles1
## 48821 articles1
## 49137 articles1
## 66863 articles2
## 66903 articles2
## 69972 articles2
```

Visualizations

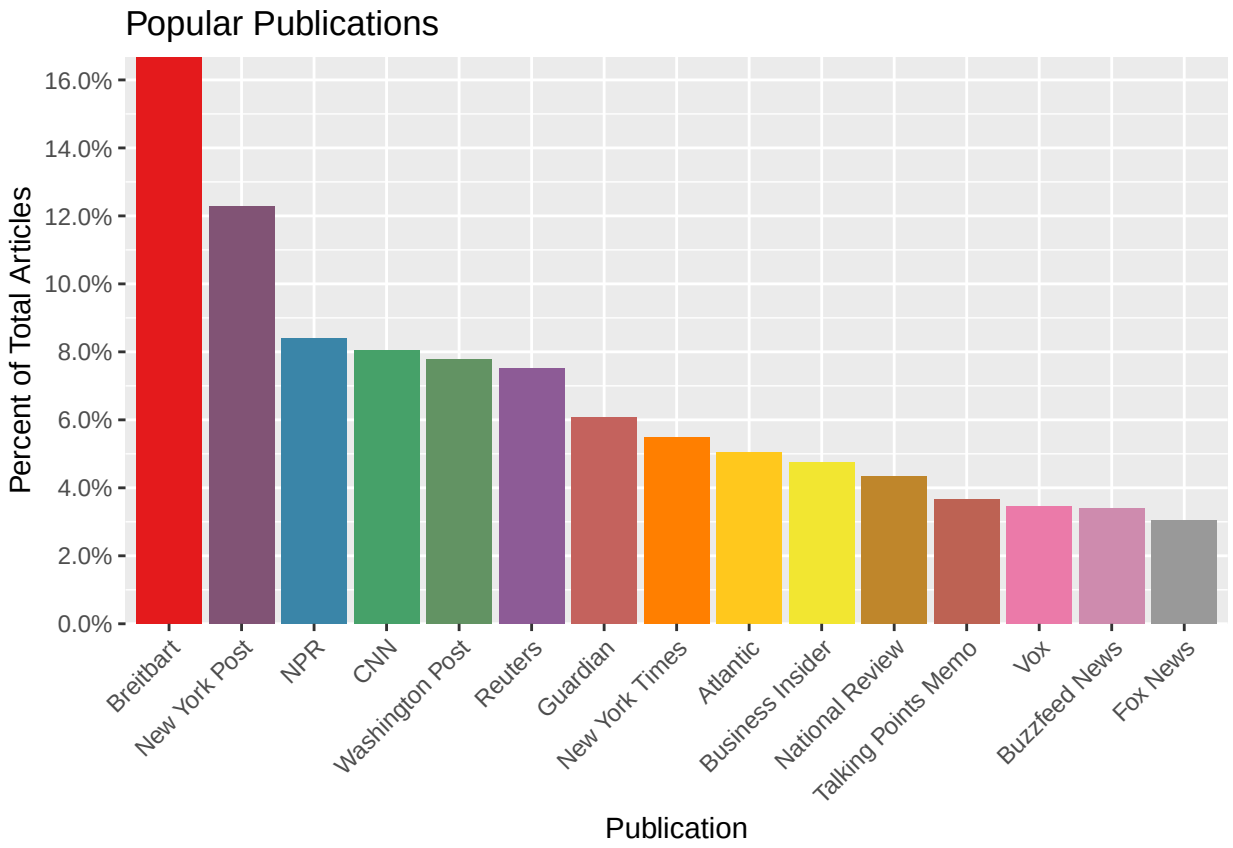
```r
# Looking at the articles published over time from 2012 to 2018
ggplot(dates, aes(x=date)) +
    geom_freqpoly(bins=100, size=1.5, alpha = I(.6), color="purple4") +
    scale_x_date(limits = as.Date(c("2012-01-01","2018-01-01")), breaks = function(x) seq.Date(from = as
    scale_y_continuous(breaks = pretty_breaks(10)) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 7616 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 3 rows containing missing values (geom_path).
```



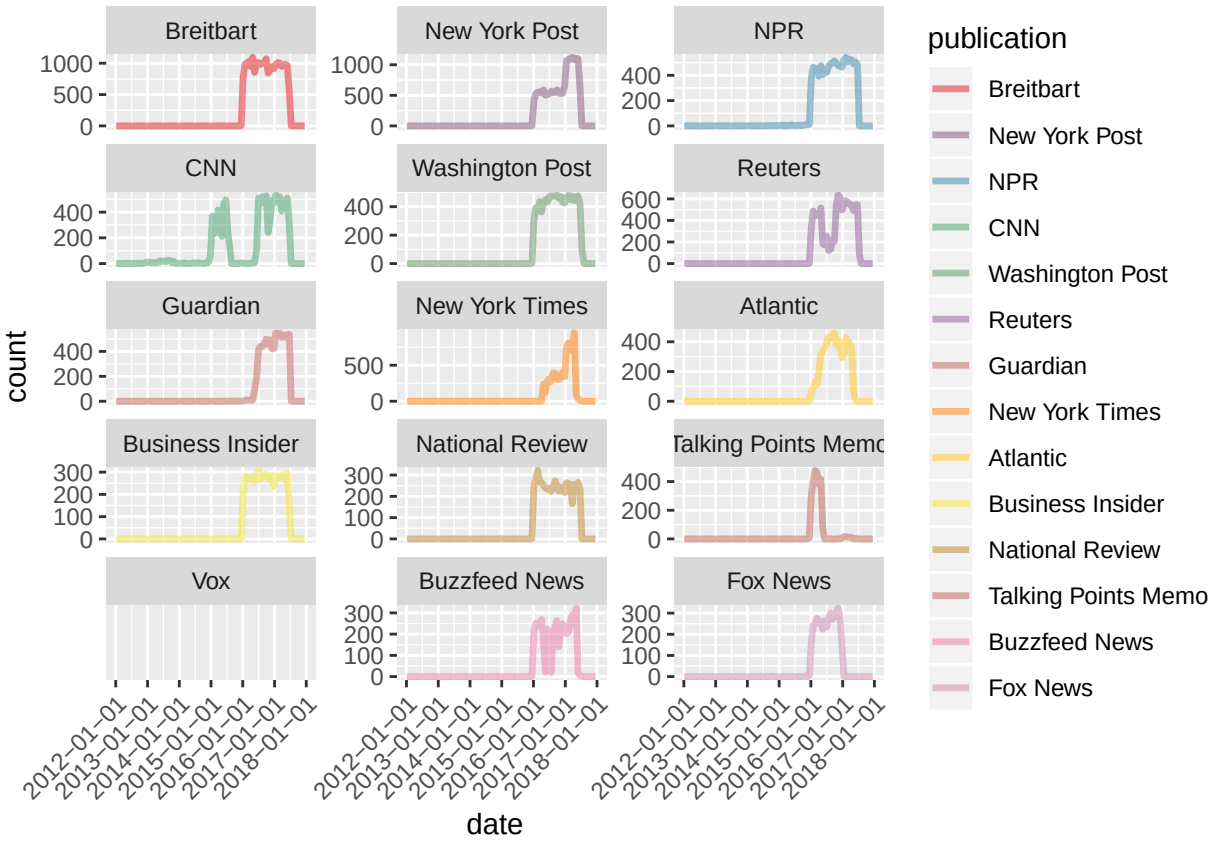

```
# Increase colors in palette
Palette <- colorRampPalette(brewer.pal(9, "Set1"))
# Order data for plot
dates <- within(dates,
                publication <- factor(publication, levels=names(sort(table(publication), decreasing=TRU

# Percent of articles from each publisher
ggplot(dates, aes(x=factor(publication))) +
    geom_bar(aes(y = (..count..)/sum(..count..)), stat="count", fill=Palette(15)) +
    labs(title="Popular Publications", x="Publication", y = "Percent of Total Articles") +
    scale_y_continuous(breaks = pretty_breaks(10), expand = c(0,0), labels = percent) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Number of articles published by publisher over time
ggplot(dates, aes(x=date, color=publication)) +
    geom_freqpoly(bins=100, size=1.2, alpha = I(.5)) +
    scale_x_date(limits = as.Date(c("2012-01-01","2018-01-01")), breaks = function(x) seq.Date(from = as
    scale_y_continuous(breaks = pretty_breaks(3)) +
    scale_color_manual(values = Palette(15)) +
    facet_wrap( ~ publication, ncol=3, scales = "free_y", shrink = T) + #y-axis varies by publisher
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 7616 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 42 rows containing missing values (geom_path).
```

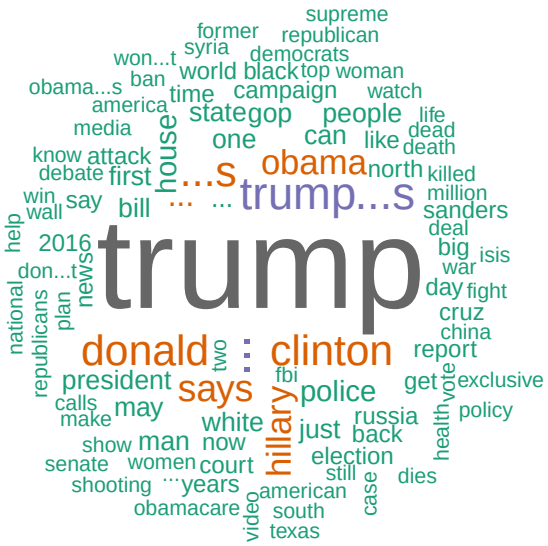



```
articles3[articles3$publication=="Vox",]
```

```
## # A tibble: 4,947 x 10
##        X1     id title publication author date        year month url
##     <int>  <int> <chr> <chr>       <chr>  <date>     <dbl> <dbl> <chr>
##  1 129972 195429 Why ~ Vox         Timot~ NA          2016    12 http~
##  2 129973 195430 Pres~ Vox         Zeesh~ NA          2016    12 http~
##  3 129974 195431 The ~ Vox         Jeff ~ NA          2016    12 http~
##  4 129975 195432 A si~ Vox         Brad ~ NA          2016    12 http~
##  5 129976 195433 I qu~ Vox         Emers~ NA          2016    12 http~
##  6 129977 195434 At w~ Vox         Zacha~ NA          2016    12 http~
##  7 129978 195435 In 2~ Vox         Carol~ NA          2016    12 http~
##  8 129979 195436 The ~ Vox         Yochi~ NA          2016    12 http~
##  9 129980 195437 The ~ Vox         Zeesh~ NA          2016    12 http~
## 10 129981 195438 9 qu~ Vox         Jenni~ NA          2016    12 http~
## # ... with 4,937 more rows, and 1 more variable: content <chr>
```

```
# Vox not showing due to issues with date formats, will fix later
```

```
# Make all words lowercase and remove certain words like "the"
titles <- tolower(dates$title) %>% removeWords(c(stopwords("en"),"breitbart", "new" ,"yorktimes", "times
# Remove possessive "'s"
titles <- gsub("â€s", " ", titles)
# Create wordcloud
suppressWarnings(wordcloud(titles, max.words = 100, random.order = FALSE, colors = brewer.pal(8, "Dark2
```


supreme
former
republican
syria
democrats
won...t
world
black
top
woman
obama...s
ban
time
campaign
watch
america
state
gop
people
life
media
house
one
can
like
dead
death
know
attack
...s
obama
north
killed
debate
first
...
...
trump...s
million
win
say
bill
wall
sanders
help
deal
2016
trump
big
isis
don...t
news
war
national
plan
day
fight
cruz
republicans
china
donald
two
clinton
report
president
says
fbi
hillary
police
get
vote
exclusive
calls
may
make
white
just
russia
health
policy
show
man
now
back
election
senate
women
court
still
dies
shooting
years
case
obamacare
american
video
south
texas