Twitter Streaming 2

Set the packages. Vital information was hidden as well as warning and output.

```
library(twitteR)

setup_twitter_oauth(key,secret,access,access_secret)

Sys.setenv(JAVA_HOME="/usr/lib/jvm/default-java")
Sys.setenv(HADOOP_CMD="/home/bcarancibia/workspace/cuny_msda_is622/hadoop-2.7.1/bin/hadoop")
Sys.setenv(HADOOP_STREAMING="/home/bcarancibia/workspace/cuny_msda_is622/hadoop-2.7.1/share/hadoop/tools

Sys.setenv(SPARK_HOME = "/home/bcarancibia/workspace/cuny_msda_is622/spark-1.4.1-bin-hadoop2.6")
.libPaths(c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib"), .libPaths()))
library(SparkR)

sc <- sparkR.init(master="local")
sqlContext <- sparkRSQL.init(sc)
```

I am going to look at Twitter data, specifically looking at the hashtag #datarevolution. This is an important hashtag in the international development space because of the recent increase in desire for countries, companies, and aid organizations to integrate analytics into their everyday workflows. I am going to collect hashtags and then count the top ten screen names that use the hashtag #datarevolution.

```
tweets <- searchTwitter("#datarevolution",n=9999)
```

```
## Warning in doRppAPICall("search/tweets", n, params = params,
## retryOnRateLimit = retryOnRateLimit, : 9999 tweets were requested but the
## API can only return 1648
```

```
x <- twListToDF(tweets)

sparkdf <- createDataFrame(sqlContext, x)

group <- agg(group_by(sparkdf, sparkdf$screenName), sum_of_screenname=(count(sparkdf$screenName)))
head(group)
```

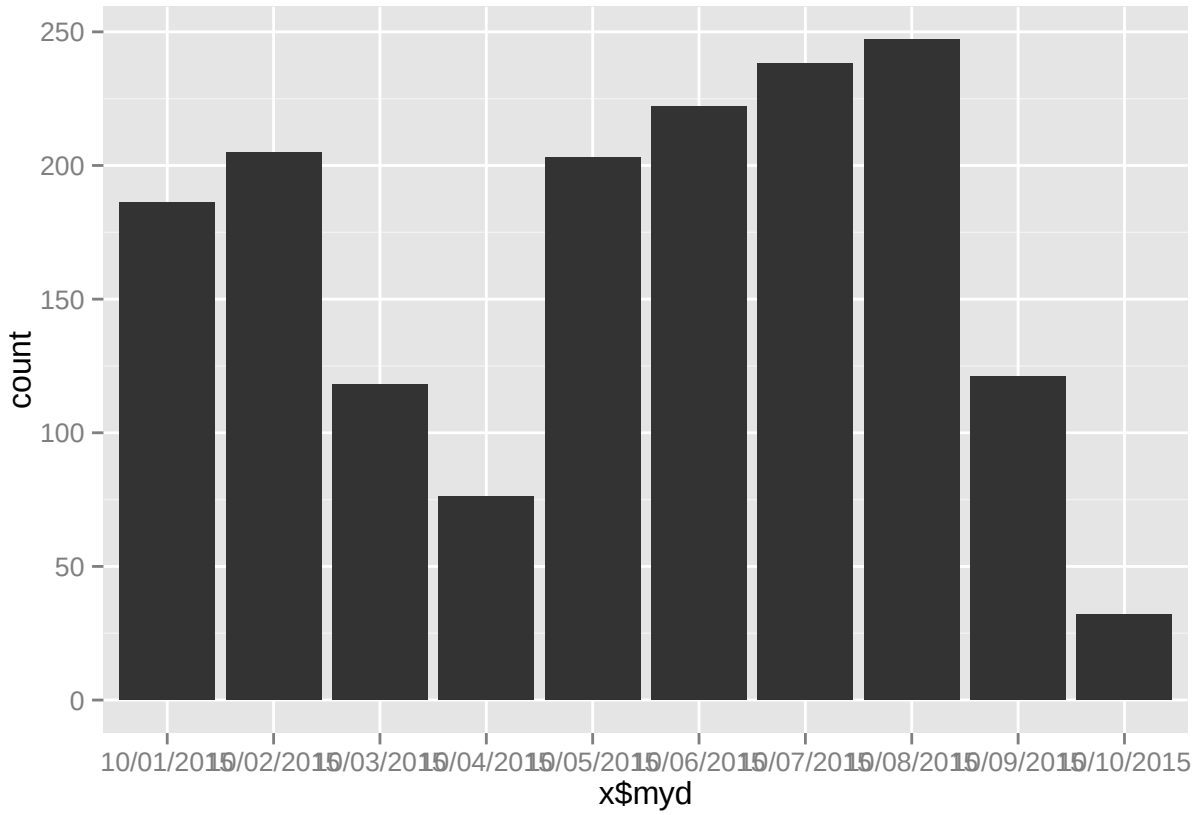
```
##        screenName sum_of_screenname
## 1      blondelena                 7
## 2           fpgil                 1
## 3        keyram10                 3
## 4      SadiQBichi                 1
## 5        alabriqu                 1
## 6 OpenDataService                 2
```

The next step is to parse out dates and then quickly plot the data to get an idea of distribution of the data.

```
library(ggplot2)
library(lubridate)
library(forecast)
```

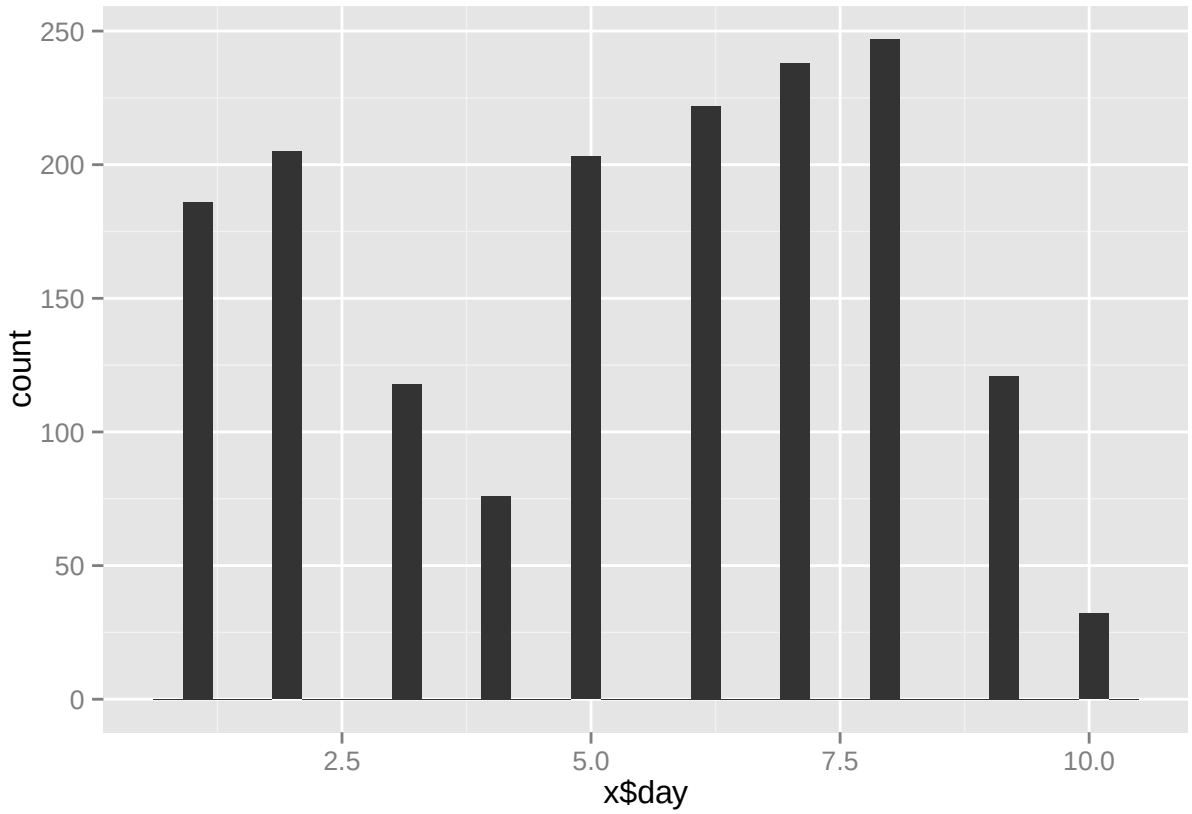
```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: timeDate
## This is forecast 6.1
```

```
x$created <- parse_date_time(x$created, "%Y%m%d %H%M%S", truncated = 3)

x$day <- day(x$created)
x$month <- month(x$created)
x$year <- year(x$created)
x$hour <- hour(x$created)
x$minute <- minute(x$created)
x$time <- sprintf('%02d:%02d', x$hour, x$minute)
x$myd <- sprintf('%02d/%02d/%02d', x$month, x$day, x$year)

#Tweets by Month, Day, Year
qplot(x$myd, data = x, geom="histogram")
```




```
qplot(x$day, data=x, geom="histogram")
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



One thing to notice is that the R package to scrape tweets, only seems to take into account the past week or so. Based on the twitteR package, there could be Twitter API restrictions

Group by Screenname and sum the retweets per screen. This can be used in the future to do a social network analysis.

```
group <- agg(group_by(sparkdf, sparkdf$screenName), sum_of_retweets=(sum(sparkdf$retweetCount)))

head(group,10)
```

```
##         screenName sum_of_retweets
## 1       blondelena             134
## 2            fpgil               7
## 3          keyram10              14
## 4       SadiQBichi              25
## 5          alabriqu               6
## 6  OpenDataService              21
## 7     bracken10011              10
## 8        Cath_Cand               3
## 9        EvarMburu               1
## 10     writeosahon               4
```