**VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI 590018**

****

Report on

**“ANALYSIS OF TWEETS ON DEMONETIZATION”**

By,

AAKASH VERMA (1BM14CS001)

ASHWIN UDAY BHATKAL (1BM14CS024)

DIVIJ ASIJA (1BM14CS037)

Under the Guidance of

**Dr KAVITHA SOODA**

Associate Professor,

Department of CSE

BMS College of Engineering

Work carried out at

****

Department of Computer Science and Engineering

BMS College of Engineering

(Autonomous college under VTU)

P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019

2017-2018

**BMS College of EngineerinG**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

****

***CERTIFICATE***

This is to certify that the BIG DATA and ANALYTICS mini Project titled “**ANALYSIS OF TWEETS ON DEMONETIZATION”** has been carried out by AAKASH VERMA (1BM14CS001), ASHWIN UDAY BHATKAL (1BM14CS024) and DIVIJ ASIJA (1BM14CS037) during the academic year 2017-2018.

Signature of the guide

**Dr KAVITHA SOODA**

Associate Professor,

Department of Computer Science and Engineering

BMS College of Engineering, Bangalore

**BMS College of EngineerinG**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

****

***DECLARATION***

We, AAKASH VERMA (1BM14CS001), ASHWIN UDAY BHATKAL (1BM14CS024), DIVIJ ASIJA (1BM14CS037), students of 7th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this assignment work entitled "**ANALYSIS OF TWEETS ON DEMONETIZATION**" has been carried out by us under the guidance of **Dr.KAVITHA SOODA,** Associate Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Aug- Dec 2017. We also declare that to the best of our knowledge and belief, the assignment reported here is not from any part of any other report by any other students.

**Signature of the Candidates**

AAKASH VERMA (1BM14CS001) .

ASHWIN UDAY BHATKAL (1BM14CS024) .

DIVIJ ASIJA (1BM14CS037) .

Objectives

To perform various kinds of analyses namely sentiment, word cloud and time series analyses on as many tweets, related to **#demonetization,** as we can get.  
To be able to view the results of such analyses using such visual and tabular representations as may be deemed necessary.

The questions to be answered are as under:

1. Generate the Word Cloud from the textual dataset

2. What percentage of tweets are negative, positive or neutral?

3. What are the most famous/re-tweeted tweets?

4. Plot the following:

* Number of Tweets per hour
* Number of re-tweets per hour
* Number of re-tweets per minute

Introduction

The assignment consists of four major analyses –

* Word Cloud – A vibrant pictorial representation of a collection of buzzwords among tweets on demonetization in India.
* Sentiment Analysis – A positive-vs-negative lexicon analysis of sentiments over demonetization in tweets.
* Popularity – A simple sum of retweets and favourite count as a measure of popularity.
* Time – series analyses - Comparisons drawn over tweets and retweets with time.

The data to be analysed is always pulled fresh from Twitter via the Internet post direct OAuth authentication. The main data object is the tweet s4 object that we operate upon. We were able to scavenge a minimum of 1500 tweets for each of these analyses.

Lists and data frames were the two simple and rich data structures that we have used to store into and process intermediate results from. The data underwent a series of relatable transformations where finally, the data variables containing the end results were fed to graph or chart plotting methods with suitable parameters.

We have used R as our programming and analytics tool for whole-body of the process.

Design

*Requirements*: Data set of tweets on demonetization, R libraries such as twitteR, ROAuth, tm, plyr, deplyr, ggplot2, stringr and wordcloud and RStudio.

*Techniques used from BDA*: Data ETL and storage, graph plotting.

**Word Cloud Analysis**

Here, we have produce a word cloud of 100 most frequently occurring words in tweets that we retrieved. For this, the main operation on our side was to cleanse and filter the tweets’ texts to allow us only relevant words in our intermediate result set. The wordcloud() from the wordcloud library did the rest of the work for us by taking in the intermediate result set along with suitable parameters to output the wordcloud as an image.

Standardization of data through cleaning and filtering

Creation of word cloud using the transformed data

Extraction of text from tweet object

API call to Twitter to retrieve tweets

**Sentiment Analysis**

Similar to previous exercise, we fetch the tweets and prepare for sentiment analysis. We have two lexicons of positive and negative words that we use to sum up the corresponding values of words present in an individual tweet and hence, get the “sentiment score”.

We segregate the tweets into Positive, Neutral and Negative classes and plot their numbers against the date in the same graph, when they were retrieved or pulled by the program.

The program stores the tweets cumulatively so that we have a history of tweets and therefore, the graph is able to show the trend over time and not just over a constant dataset.

Plotting a line graph showing no of tweets of different emotions vs time

Application of sentiment score function to each tweet Updating csv file storing the tweets with fresh ones

tweet

Segregation of tweets into their categories on the basis of sentiment score

Updating csv file storing the tweets with fresh ones

ndardization of data through cleaning and filtering

API call to Twitter to retrieve tweets

Extraction of text from tweet object

**Popularity**

The popularity of a tweet is relative and is assumed to be the sum of the number of retweets and the favorited count.

The tweets are examined individually and their retweet and favorited counts are added up to produce a data frame with the tweet id and corresponding popularity score present next to each other.

The tweet IDs are then used to arrange tweets into descending order.

Arrange the data frame in a decreasing order as per popularity score

Store tweet ID and corresponding popularity score in the data frame

Create another list of tweets pulled out from the original according to sorted tweet id column

Create a data frame with columns ‘tweet.id’ and ‘popularity’

API call to Twitter to retrieve tweets

Summation of retweet and favorited count of individual tweets

**Time Series Analyses**

Here, the tweets are examined for their ‘created’ field from which we extracted the hour or minute of their creation and are put to the qplot function to return us a histogram of **no of tweets or retweets** against **the hour of a day or the minute of an hour**.

API call to Twitter to retrieve tweets

Conversion of double value of ‘created’ field to a POSIX timestamp

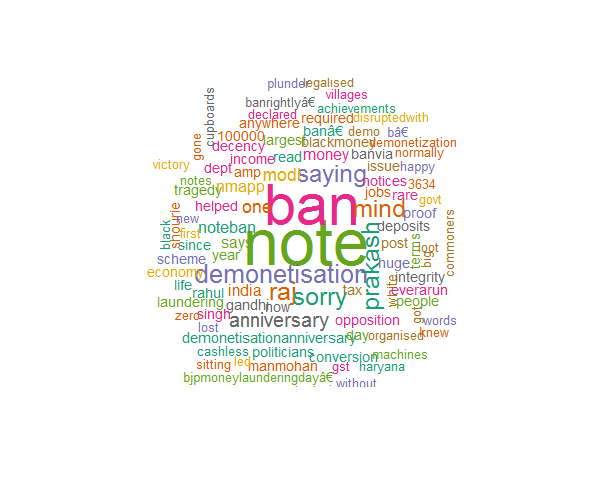
Extraction of hour or minute from the timestamp.

Plotting different graphs with the list data and other suitable parameters.

Appending the extracted value to a list

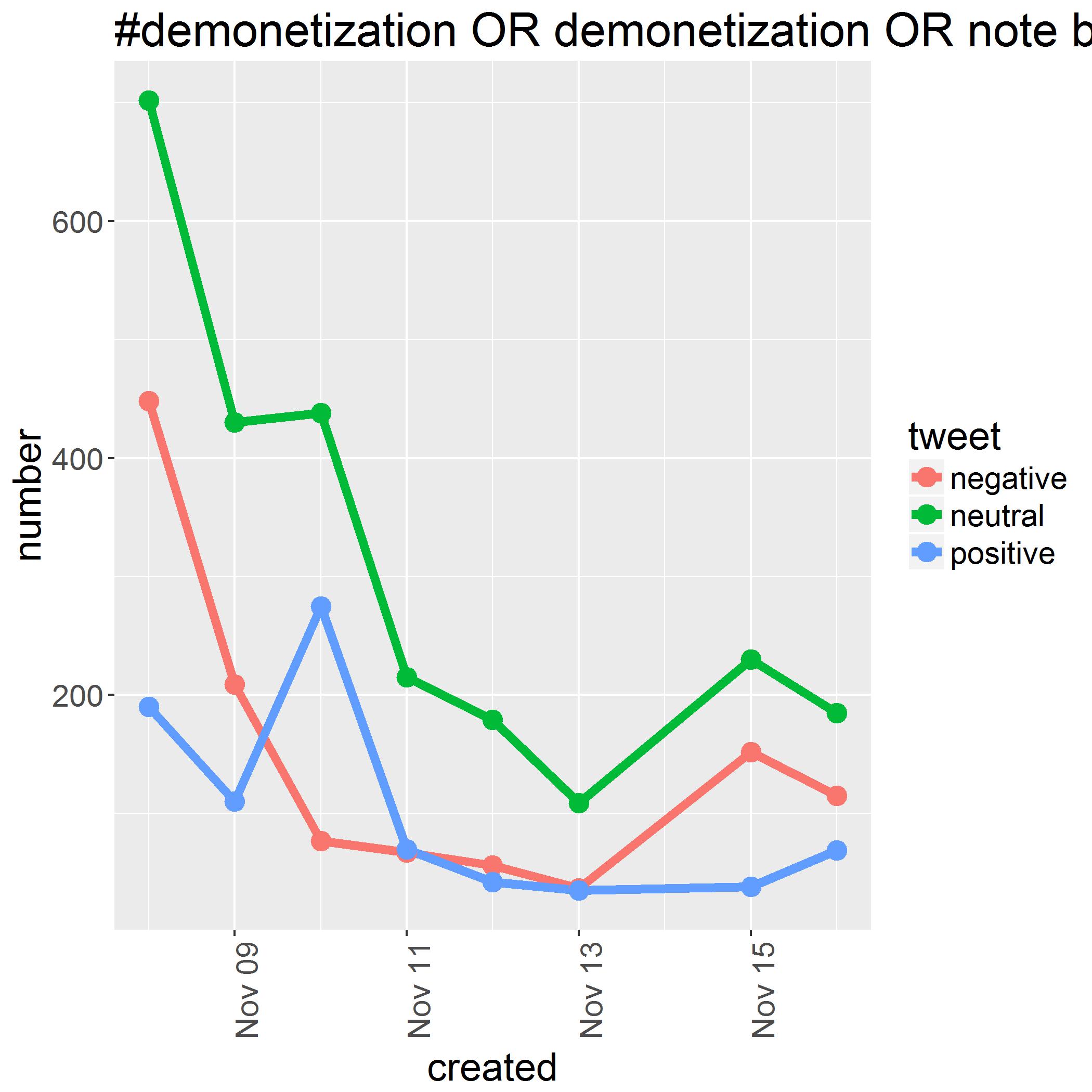
Screenshots

**Word Cloud Analysis**

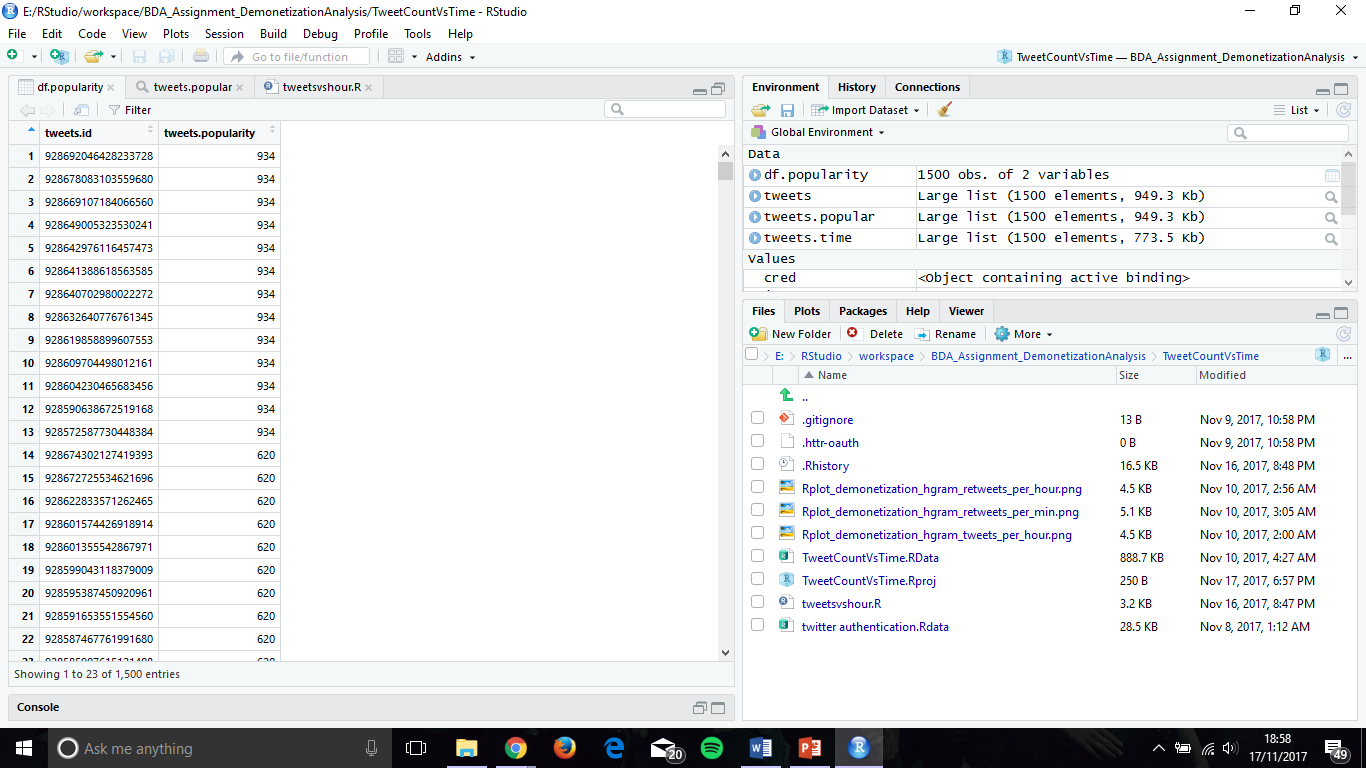


The first of the objectives is the construction of a word cloud which basically decorates a predefined number of popular words from among the tweets on #demonetization in the form of a “cloud”.

**Sentiment Analysis**

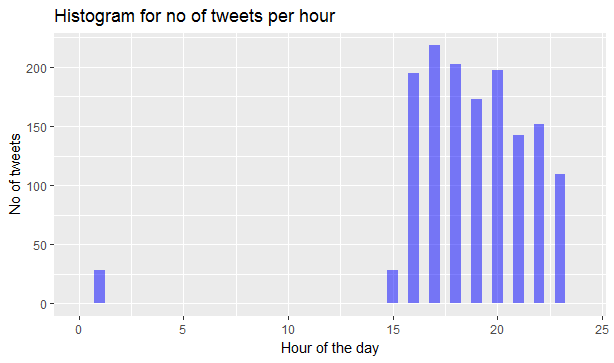
****A plot of a number of neutral, positive and negative tweets from a cumulative dataset based on new everyday findings (made possible by running the function on a daily basis). [1]

**Popularity Analysis**

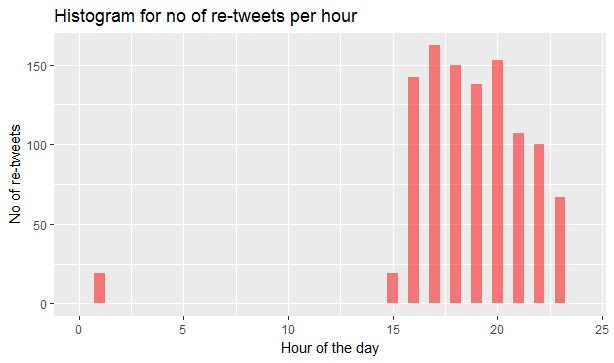


As you can see on your left-hand side, we have a data frame consisting of the tweet IDs of the most popular tweets along with their corresponding popularity score, ordered by the popularity score in a descending fashion. [2] [3]

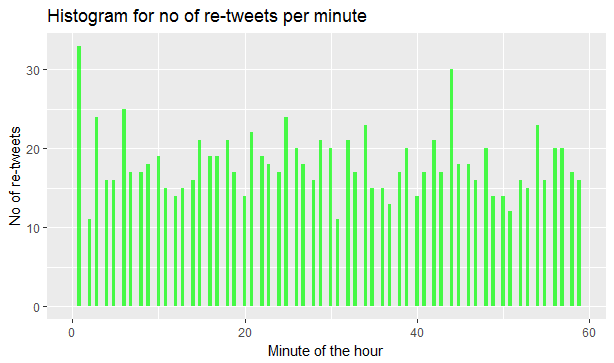
**Time Series Analyses**



This basically depicts the density of tweets on different hours of a day over a period of a number of days. This is parsed from the ‘created’ properties of the tweets and the hour from the resulting timestamp is individually appended to a list. [3]



This represents a similar drawing as the previous one except for the fact that only retweets ( a type of tweet) are filtered down and considered.



A plot of retweets vs minute of the hour, again a parameter extractible from the timestamp of every tweet object.

New Learnings

The project turned to be sweeter than the experience we had anticipated for. Right from learning the very mechanics of the language R to the applied practice of precise sequencing of ETL of the retrieved data from the REST API call helped us grow towards perhaps, what would be a very much required skill set in the not so distant future.

We got to know and explore the different data tools and libraries, required for data handling and graph plotting, which are basic and seem a must for the new generation computer science students.

The meat of the analyses being data transformation and extraction provided us with a compact and applied experience.

To be niche, the right conception of checkpoints through the scripts for storage of intermediate results in appropriate data structures and rigorous check on the maintenance of homogeneity of data is a subtle and slick skill we could discover which makes everything beautiful!

Future Enhancements

To be specific and clear about our thoughts and plans of improvising are as under:

1. A better and complex sentiment analysis algorithm is obvious and much needed for a better outlook and more precise analysis.
2. Stacking up of tweets in a greater way to provide us eventually with GBs of data to operate on for lesser extrapolation.
3. Creation or modification to a generic system that handles not only tweets on demonetization but also other media resources on other topics.
4. To come up with different sketches and plots to give us deeper and different insights.

The additional questions that can be answered by the dataset are:

1. Demographic and geological structuring of tweets can be easily performed.
2. Development of emotions from the demonetization can be tracked through time and factors for the same could be pointed out for the possible movement or trend.

References

Web help:

* [1] How to create a Word Cloud  
  <http://analyticstraining.com/2014/how-to-create-a-word-cloud-in-r/>
* [2] Sentiment Analysis of tweets  
  <https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/>
* [3] Stack Overflow  
  <https://stackoverflow.com>

To list, the tools used for the analyses are:

* R and RStudio
* Twitter Application Management

Libraries used:

* twitteR
* ROAuth
* tm
* dplyr
* ggplot2
* stringr