# IT 497 OSEMN Assignment

Done by, Lokender Reddy Sheelam

## TOPIC

Are there more GitHub R repositories created in February than January?

## INTRODUCTION

Let's start the discussion by explaining what is GitHub and what we mean by GitHub repository.

Wikipedia defines Github as -

"GitHub is a web-based Git repository hosting service, which offers all of the distributed revision control and source code management (SCM) functionality of Git as well as adding its own features."

In simple terms, GitHub provides us to store our work remotely and to make changes to work while also giving us the opportunity to do version control. It is one of the most widely used platform for doing development by programmers as it enables them to collaborate and provides an efficient work to do open source distribution.

Using GitHub, we upload our work into data structures called repositories. A repository refers to a central place where data is stored and maintained. A GitHub repository can be classified according to the repository language being used. GitHub automatically understands what code we are pushing to GitHub and classifies it under different languages by reading the syntax being followed.

My task was to find out number of GitHub repositories created using R programming language in the months of February and January.

**CODE**

\documentclass{article}

\title{OSEMN}

\author {Lokender}

\begin{document}

\SweaveOpts{concordance=TRUE}

<<OSEMN,fig=TRUE>>=

library(jsonlite)

library (plyr)

library (ggplot2)

url <- 'https://api.github.com/'

# the GitHub API

path <- 'search/repositories'

# Repositories path

search1 <- '?q=created%3A%222015-01-01+..+2015-01-31%22'

# Created in January of 2015

search2 <- '?q=created%3A%222015-02-01+..+2015-02-28%22'

# Created in February of 2015

searchR <- '+language:r'

# And language = R

pageNo <- '&page=1'

**O**BTAIN THE DATA

# We can specify a page number

pageSize <- '&per\_page=1'

# 100 results per page is the max

URLJan <-paste0(url, path, search1, searchR, pageNo, pageSize)

#Here is what URLJan looks like all together

#URL <-'URL <- 'https://api.github.com/search/repositories

#?q=created%3A%222015-01-01+..+2015-01-31%22+language:r&page=1&per\_page=1'

**O**BTAIN THE DATA

URLFeb <-paste0(url, path, search2, searchR, pageNo, pageSize)

#Here is what URLFeb looks like all together

#URL <-'URL <- 'https://api.github.com/search/repositories

#?q=created%3A%222015-02-01+..+2015-02-28%22+language:r&page=1&per\_page=1'

# The above URL will create a page in json format

# Read the json and convert it to a list using the jsonlite package

l = jsonlite::fromJSON(URLJan)

# Get number of Jan Repositories by capturing the total count

repos.Jan <-l$total\_count

**S**CRUB THE DATA

STARTING FROM HERE

# The above URLFeb will create a page in json format

# Read the json and convert it to a list using the jsonlite package

j = jsonlite::fromJSON(URLFeb)

# Get the number of Feb Repositories by capturing the total count

repos.Feb <-j$total\_count

# Create a dataframe and use ggplot2 to create a bar graph

# Dataframe will have 2 columns -- Months and repositaries

months <-c("January", "February")

repos <-c(repos.Jan, repos.Feb)

#We use the data.frame function to bind the 2 columns together and create df

df = data.frame(months, repos)

#Finally, we create a bar chart

ggplot(data=df, aes(x=months, y=repos, fill=months)) + geom\_bar(stat="identity")+ggtitle("Number of GitHub R repositories created")

ggplot(data=df, aes(x=months, y=repos, group=1)) +

geom\_line()+ggtitle("Factor of GitHub R repositories created")

pie(repos, labels = months, main="Pie Chart of Months")

#To get month with More GitHub R repositaries

HighestMonth<- months[which.max(repos)]

#summary() provides summary of data like min, max, mean, median.

Summary <- summary(df)

**E**XPLORE THE DATA

#Str() provides great info about the structure of object.

str(df)

#class() provides info about what is the data type

#stored like integer, number, data frame etc.

dfclass <- class(df)

monthclass <- class(months)

reposclass <- class(repos)

#view() is used to see the table in grid

view(df)

@

"Are there more GitHub R repositories created in February than January?

Repositories created in February are -\textbf{ \Sexpr{repos.Feb}}

Repositories created in January are - \textbf{\Sexpr{repos.Jan}}

\textbf{\Sexpr{HighestMonth} }has more repositaries created.

"

References:

1. RBlogger

2. Discussion Board setup on ReggieNet -Forums / IT497 OSEMN Assignment Help / Reading JSON in R / Data gathering on Github

3. Class Notes

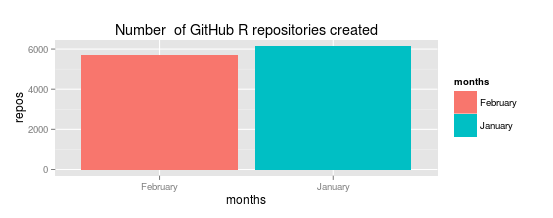
4. R resources in ReggieNet under Resources and Material

% $

\end{document}

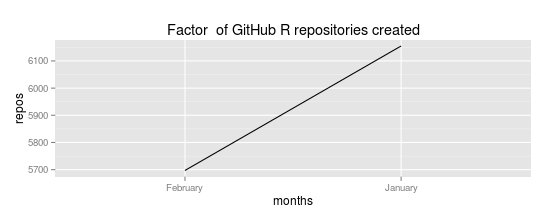
**RESULTS**

GRAPHI**N**G



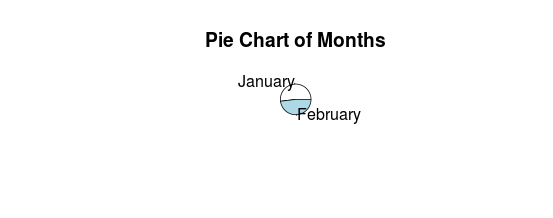
Figure

Figure 1 displays with the help of ggplot R library, a simple bar chart showing the number of repositories created in the month of January compared to month of February. As it is evident in the chart the number of repos(repositories) is depicted by y axis and x axis gives the months. Also it is clear that number of repos created in January are more than ones created in February.



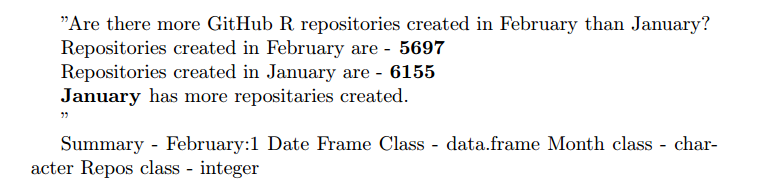
Figure

Figure 2 displays with the help of ggplot R library , a line chart which shows clearly how number of repositories created in January decreased in month of February.



Figure

Figure 3 displays with the help of pie chart that month of January had slightly more than half of the pie which suggests that January had more number of repos created.



Figure

Figure 4 displays the result using \Sexpr command to give us the exact count of repos created in each month. Also it gives us the month which had most repos created.

Following code below was used to find more useful details about the data under investigation.

**M**ORE EXPLORING

> summary(df)

months repos

February:1 Min. :5697

January :1 1st Qu.:5812

Median :5926

Mean :5926

3rd Qu.:6040

Max. :6155

> str(df)

'data.frame': 2 obs. of 2 variables:

$ months: Factor w/ 2 levels "February","January": 2 1

$ repos : int 6155 5697

> class(df)

[1] "data.frame"

> class(months)

[1] "character"

> class(repos)

[1] "integer"