Audience and User Analysis

By:

Deepak Malpani (17BCE0306)

Udit Singhania (17BCE2060)

Final Report Submitted for

CSE3021 SOCIAL INFORMATION NETWORKS

Slot – B1+TB1

Computer Science and Engineering

NOVEMBER 2020

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

When you share content in an online social network, who is listening? Users have scarce information about who actually sees their content, making their audience seem invisible and difficult to estimate. However, understanding this invisible audience can impact both science and design, since perceived audiences influence content production and self-presentation online. In this paper, we combine survey and large-scale log data to examine how well users' perceptions of their audience match their actual audience on Facebook. We find that social media users consistently underestimate their audience size for their posts, guessing that their audience is just 27%of its true size. Qualitative coding of survey responses reveals folk theories that attempt to reverse-engineer audience size using feedback and friend count, though none of these approaches are particularly accurate. We analyze audience logs for 222,000 Facebook users' posts over the course of one month and find that publicly visible signals — friend count, likes, and comments — vary widely and do not strongly indicate the audience of a single post. Despite the variation, users typically reach 61% of their friends each month. Together, our results begin to reveal the invisible undercurrents of audience attention and behaviour in online social networks.

Keywords: Information Network, Facebook, Descriptive Statistics, Inferential Statistics, Sentimental Analysis, Covid19, Twitter.

Introduction

Everyday millions of people posts their activities on social network sites like Facebook, Twitter, and Instagram etc. The users don't know who all are actually viewing their posts in their friend list. The only confirmation they can get is from number of likes or comments their post's received. The number of comment and likes shows number of people who actually reacted on your post but what about the people who viewed your posts but didn't like or comment? That audience varies from day to day: friends may not log in to the site, may not see the content, or may not reply. Established media producers can estimate their audience through surveys, television ratings and web analytics, social network sites typically do not share audience information. This design decision has privacy benefits which preserve right to privacy but it also means that users may not accurately estimate their invisible audience when they post content. Social media users create a mental model of their imagined audience, then use that model to guide their activities on the site. In this project have analyzed people perception on their audience count and how they misinterpret the actuality. The data was collected by estimating the actual audience size using server logs.

We found the type of people who are active on social networks by considering factors like age and gender, and how they impact on daily activities on Facebook.

We also collected the tweets on the topic "Covid19" For conducting the sentiment analysis of the public regarding this governmental policy, I have collected data in Twitter.

Literature Review Summary Table

Authors and Year (Reference)	Title (Study)	Concept/ Theoretical model/ Framework	Methodology used/ Implementation	Dataset details/ Analysis	Relevant Finding	Limitations/ Future Research/ Gaps identified
Altman, Irwin (1975)	The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding	analysis of the concepts of privacy, crowding, territory, and personal space, with regard to human behavior	an analysis of privacy in terms of meaning, conceptions, mechanisms, and dynamics; both crowding and territory	Social and privacy concerns of SIN	-	-
Samuel D. Gosling, Sam Gaddis, Simine Vazire (2007)	Personality Impressions Based on Facebook Profiles	Examine impression based on 133 FB profiles, comparing them with how the targets See themselves and are seen by close acquaintances	Facebook profiles, Observer ratings, Accuracy criteria, Instrument	Facebook details of people who participated in the experiment	mean ICC(2,1) = .15; e accuracy correlation was .23, mean single-observer accuracy correlation was .13	Fake data is provided on personal profile which deviates from true findings

Kelly Caine, Lorraine G. Kisselburgh, Louise Lareau (2010)	Audience Visualization Influences Disclosures in Online Social Networks	Introduce visualization and numeric audience information as potential interface solutions to the problem of privacy behaviors that are misaligned with privacy preferences	experimentally varied audience presentation using either Text, Numbers, or Visualization, and measured individual responses to disclosing five kinds of information to determine the effect of audience awareness	2330 participants responded to a 5-item survey scale asking them to indicate to whom they intended to share their personal information	A one-way between subject's ANOVA testing whether there was an effect of visualization on number of items Disclosed for non-users was significant F (2, 419) = 3.20, p = .042	audience visualizations used in this study is rudimentary, and drawn for empirical purposes only. Future work can focus on the usefulness of particular design choices for effective audience visualization
Lars Backstrom, Eytan Bakshy, Jon Kleinberg, Thomas M. Lento, Itamar Rosenn (2011)	Center of Attention: How Facebook Users Allocate Attention Across Friends	Analysis of personal networks, based on the way in which an individual divide his or her attention across contacts	compute metrics for a number of different modalities of attention. The modalities were divided into two distinct groups: communication and viewing. Visualization done on mathematical models.	FB Data	The balance of attention is a relatively stable property of an individual over time, and that it displays interesting variation across both different groups of people and different modes of interaction.	Improvement in Mathematical models for better accuracy
			Fielded surveys for three		that patterns or use, perception and attitude sometimes	

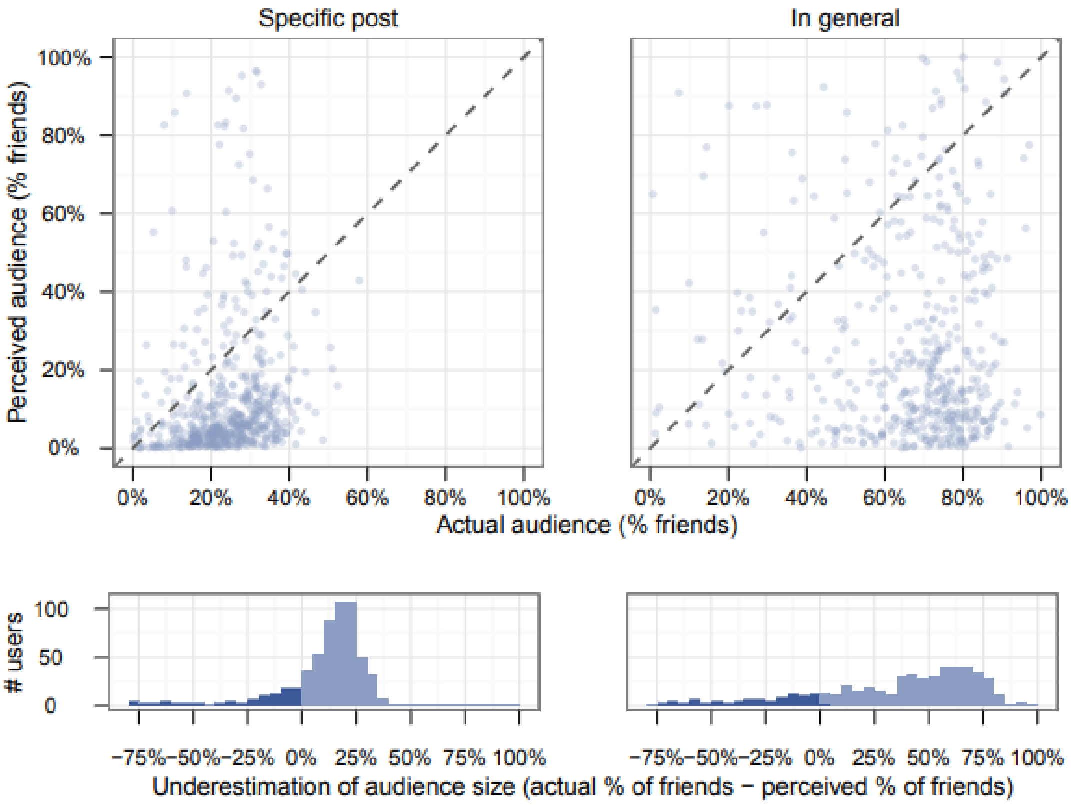
Cliff Lampe, Nicole B. Ellison, Charles Steinfieldr	Changes in Use and Perception of Facebook	Changes in Views of perception of Facebook users from time to time.	consecutive years. Conducted interviews and survey about use and perceptions of Facebook.	three consecutive years of survey data and interviews with a subset of survey respondents	change over time, though rarely drastically.That changes, when they do occur, may result from both changes in the user's social context (such as moving to or from college), and perhaps in response to a major change in features, such as the introduction of the News Feed on FB.	The main limitation of this study is the descriptive nature of the results, which makes it impossible to discern causal relationships among the variables explored.
Geeta, Rajdeep Niyogi (2016)	Demographic analysis of Twitter users	Opinions of different users have been analyzed and then sentiment analysis is performed and at the end demographic analysis is achieved to get the required data	Demographic Analysis	Data are collected from the tweets by users Twitter Data	Result shows the opinions of users in five different countries United States has high percentage of tweets done in Oscar event, India has high percentage of tweets in T20 event, France user tweet more on Paris attack and Australian users tweets high on	Current location of users is not identified. So, it is not clear that user tweets from the real location or not

					formula 1	
Prateek Dewan, Shrey Bagroy, Ponnuranga m Kumaraguru (2016)	Hiding in Plain Sight: Characterizin g and Detecting Malicious Facebook Pages	Bag of words produced sparse vector and this vector used for classification	Supervised learning algorithms. Bag of words Crowdsourcing technique: web of trust (WOT)	Like, comment and share are analyzed, and textual contents was collected from three sources: message, name, and link Facebook pages	Results are based on the different classifier and it is concluded that Neural Network classifier of Trigram feature set has high rate of accuracy of 84.13%	Large group and events were not covered Bag of words is based on limited history of 100 posts Pages can change behavior over time

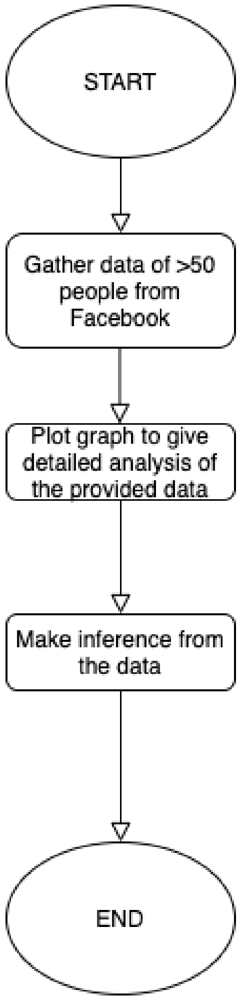
Overview of the Proposed System

The objective of the project was to analyze the audience of Facebook social network. The relation between its users. How activities on Facebook varies according to factors like age, gender, tenure. When any one share content in an online social network, who is listening? Users have scarce information about who actually sees their content, making their audience seem invisible and difficult to estimate. However, understanding this audience can impact both science and design, since perceived audiences influence content production and self-presentation online. Now days normal marketing is shifting to digital marketing. Advertisement agency are advertising through social media and targeting the social media users. As Facebook has one of the biggest social network so it’s necessary to analyze the perceived audience.

This project also deals with the sentimental analysis along with other generalizations we can attain in twitter. Twitter plays an important role in expressing our feelings about an event. The expression of anguish, as well as pleasure, can act as a measure of acceptance or rejection of certain ordinances. We will deal with classifying tweets according to the sentiments expressed in them which can be positive, negative or neutral. The aim of this project is to analyze for accurate and automatic sentiment on tweets based on trending issues like Covid19. This project makes a fair judgment about these government policies (Covid19) by using the concept of sentiment analysis.



Our analysis indicates that social media users underestimate how many friends they reach by a factor of four. Many users who want larger audiences already have much larger audiences than they think. However, the actual audience cannot be predicted in any straightforward way by the user from visible cues such as likes, comments, or friend count. The core result from this analysis is that there is a fundamental mismatch between the sizes of the perceived audience and the actual audience in social network sites. This mismatch may be impacting users' behavior, ranging from the type of content they post, how often they post, and their motivations to share content. The mismatch also reflects the state of social media as a socially translucent rather than socially transparent system. Social media must balance the benefits of complete information with appropriate social cues, privacy and plausible deniability. Alternately, it must allow users to do so themselves via practices such as butler lies. The mismatch between estimated and actual audience size highlights an inconsistency: approximately half of our participants wanted to reach larger audiences, but they already had much larger audiences than they estimated. One interpretation would suggest that if these users saw their actual audience size, they would be satisfied. Or, these users might instead anchor on this new number and still want a larger audience.



Innovation component in the project:

- As analysing audience of small social network has been done already we will try to expand the analysis by analysing the overall Facebook audience which is similar to real life and give overall idea about its audience.

- In twitter analysis what most of the researchers focus on sentimental analysis what we aim to do is to analyse the recent trends on twitter using the most trending topic in world which is CoVID-19.

Work done and implementation

Methodology adapted

We will be using descriptive statistics concept to analyze the data set. Do basic measure and then proceed with the various graphs to analyze the social networks data in order to bring about meaningful inferences or observation.

Hardware and software requirements

Hardware: minimum of 2GB RAM, Windows 7 or higher, Intel Core i3 or equivalent

Software Requirement: R, RStudio

Dataset used:

a) Pseudo Facebook dataset from Udacity.

Tweets scraped from 1st August to 23rd October 2020 on “corona”

Tools Used:

RStudio - is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics is used in this project.

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

We have used GGplot, GGally, GridExtra, Dplyr R packages for analysing and modelling data set.

Python 2&3 - python is a high level programming language for general purpose programming

Jupyter Notebook - JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.

TWITTER ANALYSIS Code for Scraping data:

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> library(twitteR)

Attaching package: 'twitteR'

The following objects are masked from 'package:dplyr':

    id, location

> library(knitr)
> library(ggplot2)
> api_key <- "IHJAq8hoX4pQMiZpEBssmEfRm"
> api_secret <- "KBh4MHayOFAOxkGFnCMjAn2QCBRcAilUHUSJXxkI7LjirMZowq"
> token <- "1123799972921663489-ZV1QtuyBwwZkOGT5ulP4D1lIW93sjS"
> token_secret <- "6fOOfBasTpLNdJwzdIi4ItVFeHo40xBQQgm8Wj2NwUIKm"
> setup_twitter_oauth(consumer_key = api_key, consumer_secret = api_secret, access_token = token, access_secret = token_secret)
[1] "Using direct authentication"
> tweets <- searchTwitter("#corona", n=10000, lang="en", since="2020-08-01", geocode='18.911674,78.558637,1310.280km')
Warning message:
In doRppAPICall("search/tweets", n, params = params, retryOnRateLimit = retryOnRateLimit,  :
  10000 tweets were requested but the API can only return 5061
> tweetsDF <- twListToDF(tweets)
> names(tweetsDF)
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"     "created"       "truncated"     "replyToSID"    "id"            "replyToUID"    "statusSource"
[11] "screenName"    "retweetCount"  "isRetweet"     "retweeted"     "longitude"     "latitude"
```

Output of Scraped data:

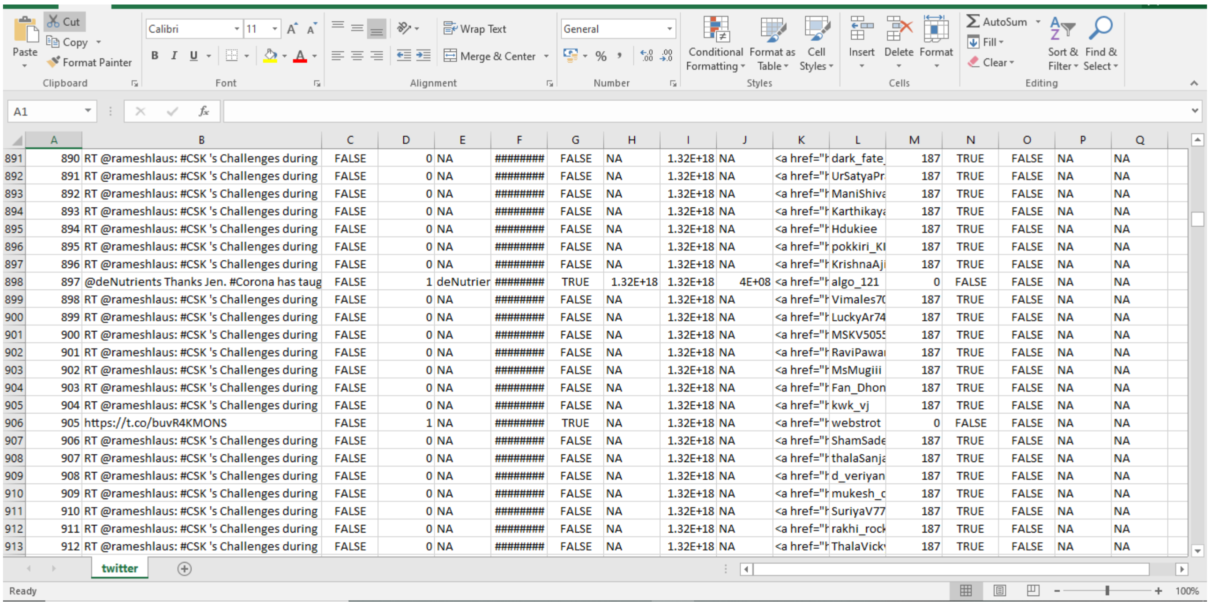
Dataset:

twitter - Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		text	favorited	favoriteCo	replyToSN	created	truncated	replyToSII	id	replyToUI	statusSou	screenNar	retweetCo	isRetweet	retweeted	longitude	latitude
2	1	RT @asho	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	kushal_ge	92	TRUE	FALSE	NA	NA
3	2	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bharathja	4	TRUE	FALSE	NA	NA
4	3	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bharathja	3	TRUE	FALSE	NA	NA
5	4	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bharathja	3	TRUE	FALSE	NA	NA
6	5	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bharathja	2	TRUE	FALSE	NA	NA
7	6	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bloodreqt	3	TRUE	FALSE	NA	NA
8	7	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bloodreqt	3	TRUE	FALSE	NA	NA
9	8	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	bloodreqt	4	TRUE	FALSE	NA	NA
10	9	RT @Anoc	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Adipkr	2	TRUE	FALSE	NA	NA
11	10	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	avinashsa	2	TRUE	FALSE	NA	NA
12	11	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	avinashsa	3	TRUE	FALSE	NA	NA
13	12	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	avinashsa	3	TRUE	FALSE	NA	NA
14	13	#Hyderab	FALSE	0	NA	########	TRUE	NA	1.32E+18	NA	<a href="h	Bloodpoir	2	FALSE	FALSE	NA	NA
15	14	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	n_shekar	24	TRUE	FALSE	NA	NA
16	15	RT @asho	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	SevadalSK	92	TRUE	FALSE	NA	NA
17	16	#Gurugra	FALSE	0	NA	########	TRUE	NA	1.32E+18	NA	<a href="h	Bloodpoir	3	FALSE	FALSE	NA	NA
18	17	RT @Anoc	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	yogeeii	2	TRUE	FALSE	NA	NA
19	18	#Hyderab	FALSE	0	NA	########	TRUE	NA	1.32E+18	NA	<a href="h	Bloodpoir	3	FALSE	FALSE	NA	NA
20	19	@scribe_	FALSE	0	scribe_pra	########	TRUE	1.32E+18	1.32E+18	4.93E+08	<a href="h	SahinVaaz	0	FALSE	FALSE	NA	NA
21	20	@Ahmed	FALSE	0	Ahmedsh	########	TRUE	1.32E+18	1.32E+18	1.01E+08	<a href="h	SahinVaaz	0	FALSE	FALSE	NA	NA
22	21	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	avinashsa	4	TRUE	FALSE	NA	NA
23	22	RT	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Hajibasha	4	TRUE	FALSE	NA	NA

twitter

Ready



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
891	890	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	dark_fate_	187	TRUE	FALSE	NA	NA
892	891	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	UrSatyaPr	187	TRUE	FALSE	NA	NA
893	892	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	ManiShiva	187	TRUE	FALSE	NA	NA
894	893	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Karthikaya	187	TRUE	FALSE	NA	NA
895	894	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Hdukiee	187	TRUE	FALSE	NA	NA
896	895	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	pokkiri_KI	187	TRUE	FALSE	NA	NA
897	896	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	KrishnaAji	187	TRUE	FALSE	NA	NA
898	897	@deNutrients Thanks Jen. #Corona has taug	FALSE	1	deNutrier	########	TRUE	1.32E+18	1.32E+18	4E+08	<a href="h	algo_121	0	FALSE	FALSE	NA	NA
899	898	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Vimales70	187	TRUE	FALSE	NA	NA
900	899	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	LuckyAr74	187	TRUE	FALSE	NA	NA
901	900	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	MSKV5055	187	TRUE	FALSE	NA	NA
902	901	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	RaviPawar	187	TRUE	FALSE	NA	NA
903	902	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	MsMugiii	187	TRUE	FALSE	NA	NA
904	903	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	Fan_Dhon	187	TRUE	FALSE	NA	NA
905	904	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	kwk_vj	187	TRUE	FALSE	NA	NA
906	905	https://t.co/buvR4KMONS	FALSE	1	NA	########	TRUE	NA	1.32E+18	NA	<a href="h	webstrot	0	FALSE	FALSE	NA	NA
907	906	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	ShamSade	187	TRUE	FALSE	NA	NA
908	907	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	thalaSanja	187	TRUE	FALSE	NA	NA
909	908	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	d_veriyan	187	TRUE	FALSE	NA	NA
910	909	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	mukesh_c	187	TRUE	FALSE	NA	NA
911	910	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	SuriyaV77	187	TRUE	FALSE	NA	NA
912	911	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	rakhi_rock	187	TRUE	FALSE	NA	NA
913	912	RT @rameshlaus: #CSK 's Challenges during	FALSE	0	NA	########	FALSE	NA	1.32E+18	NA	<a href="h	ThalaVicky	187	TRUE	FALSE	NA	NA

FACEBOOK DATA:

```
Saving facebook.tsv to facebook.tsv
```

```
df_fb = pd.read_csv('/content/facebook.tsv', sep='\t')
df_fb.head()
```

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes	mobile_likes_received	ww
0	2094382	14	19	1999	11	male	266.0	0	0	0	0	0	0	
1	1192601	14	2	1999	11	female	6.0	0	0	0	0	0	0	
2	2083884	14	16	1999	11	male	13.0	0	0	0	0	0	0	
3	1203168	14	25	1999	12	female	93.0	0	0	0	0	0	0	
4	1733186	14	4	1999	12	male	82.0	0	0	0	0	0	0	

Analysis:

Twitter

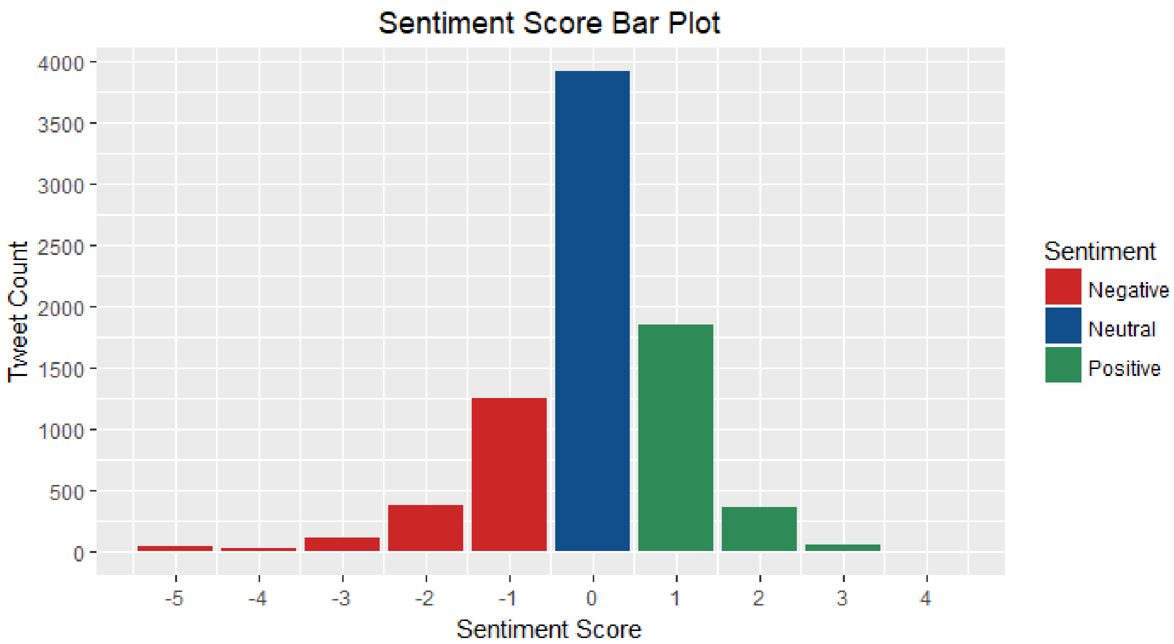
Here, we take a look at wordclouds of the tweets of the users:

Wordclouds: Complete

This is a wordcloud which shows that which word is the most used among all the words in the scraped tweets. As we can observe that 'CoVID19' is the most used word in the scraped tweets among all. After people, 'India' word is used most in these tweets. Like these all the words according to their number of times being used are being shown in thus wordcloud. In the positive polarities, wordcloud, cool masks, combo and check are some words which help preventing Corona and negative wordcloud contains IPL 2020 scenario of CSK team, as the timeline has IPL in it, most of the recent tweets are about CSK.

Positive

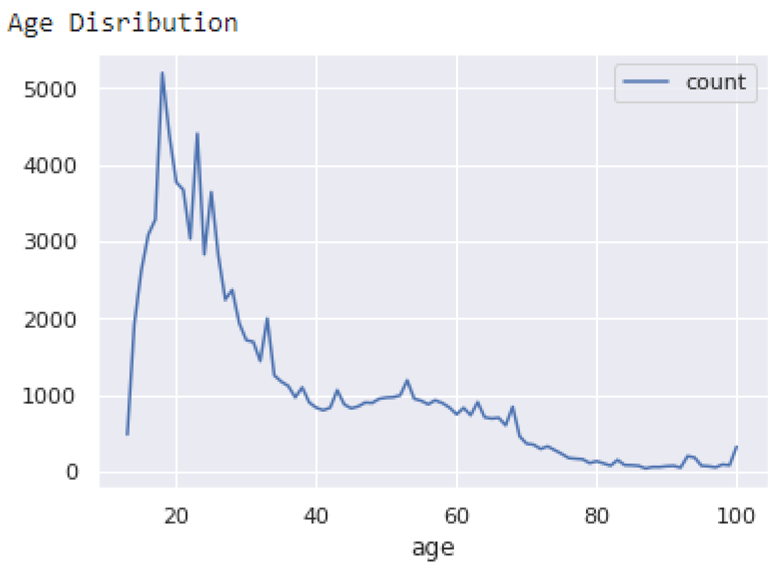
Negative:



This is the graph which shows us all the sentimental analysis part of the scraped tweets. We can see from these observations that the sentiment score bar plot for neutral tweets is much more than that of positive and negative sentiment score bar plot. After this positive sentiment score bar plot is more than that of the negative sentiment score bar plot. From this observation we can conclude that the Covid19 had a neutral effect on people when observed in a given interval of time. And most of the people were positive about the prevention and very few were negative but not about covid but about IPL.

The aim of this project was to analyze the effect of the Covid19 policy implemented by the India by using the concept of sentiment analysis. The result of our analysis shows that most of the people have now accepted the new environment in order to fight with corona. But somewhere and somehow it also had a negative effect on some common people for few days due to lack of effective cure for the disease.

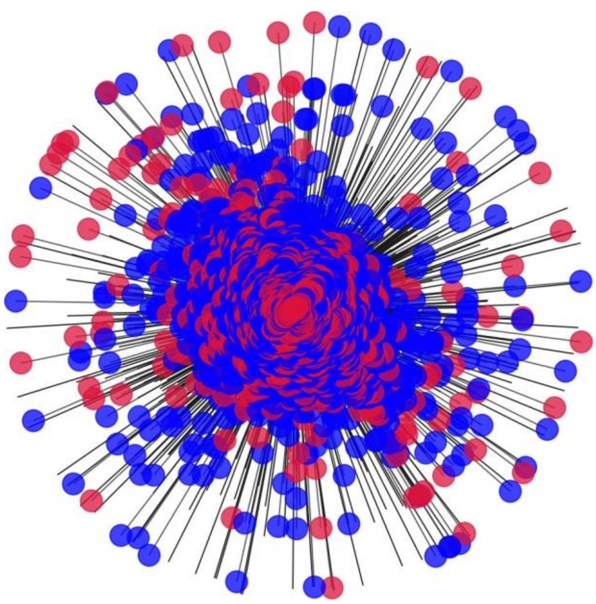
FACEBOOK ANALYSIS



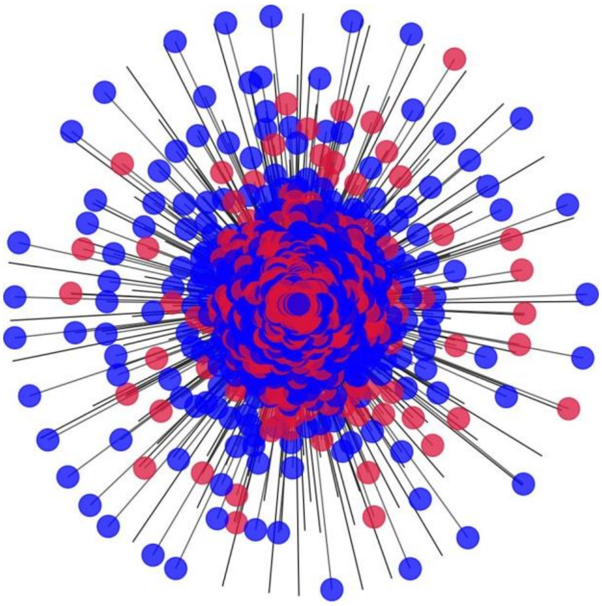
From the above graph, it is evident that youngsters of age 19 to 25 have most number of friends, and likes on their posts.

1. **Age v/s No. of Friends**

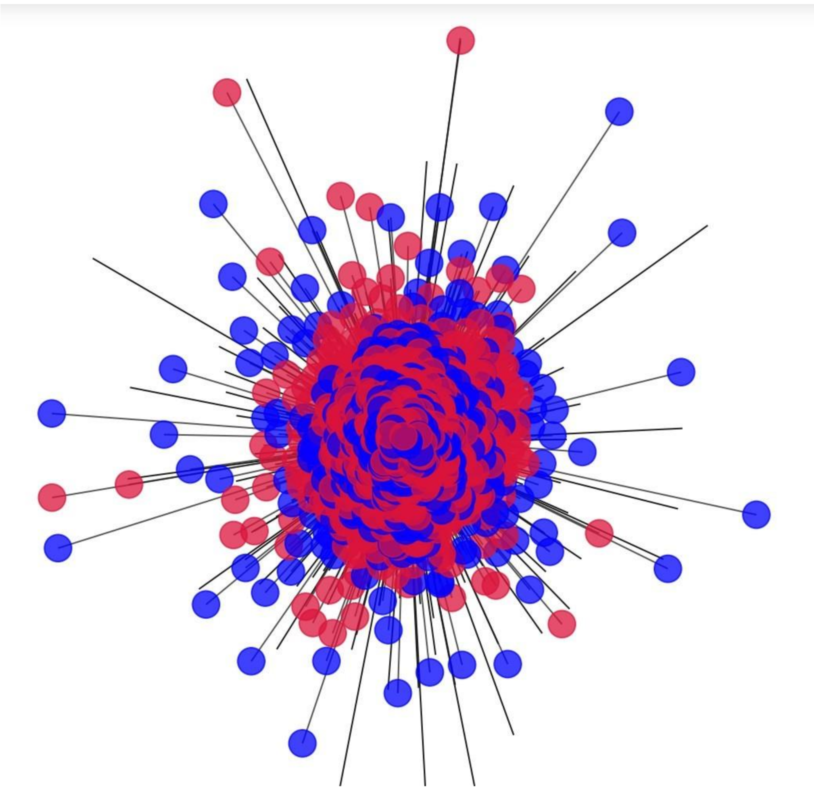
[Male = Blue, Female = Red and Edge length = No. of Friends]



Age 20



Age = 40

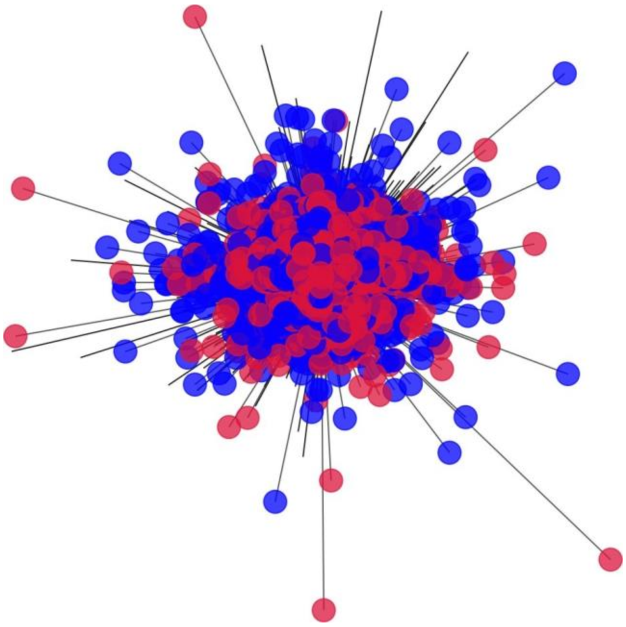


Age = 65

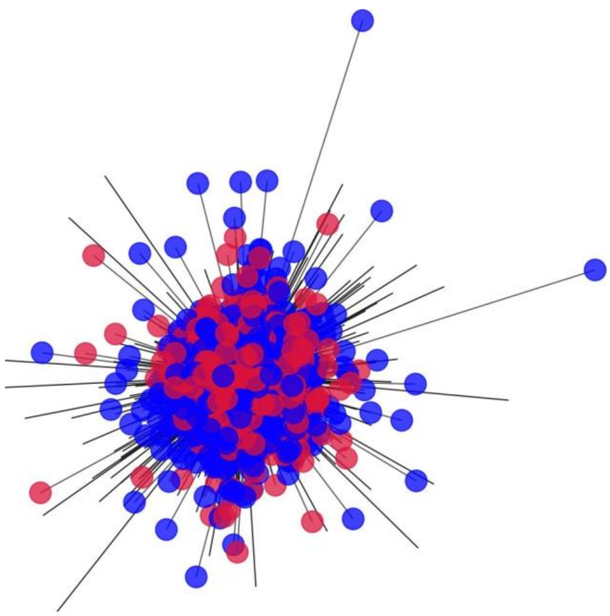
We notice that with age the number of friendships start to decrease.

2. **Age v/s Tenure**

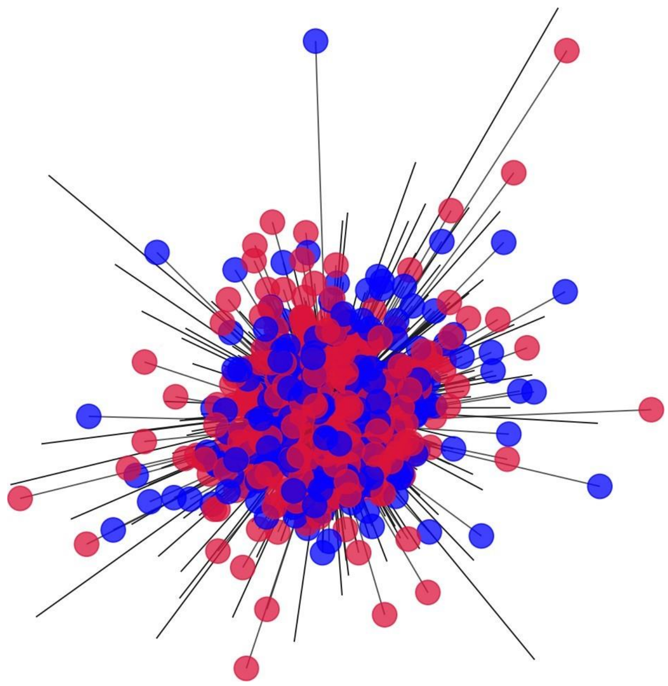
[Male = Blue, Female = Red and Edge length = Tenure]



Age = 20



Age = 40

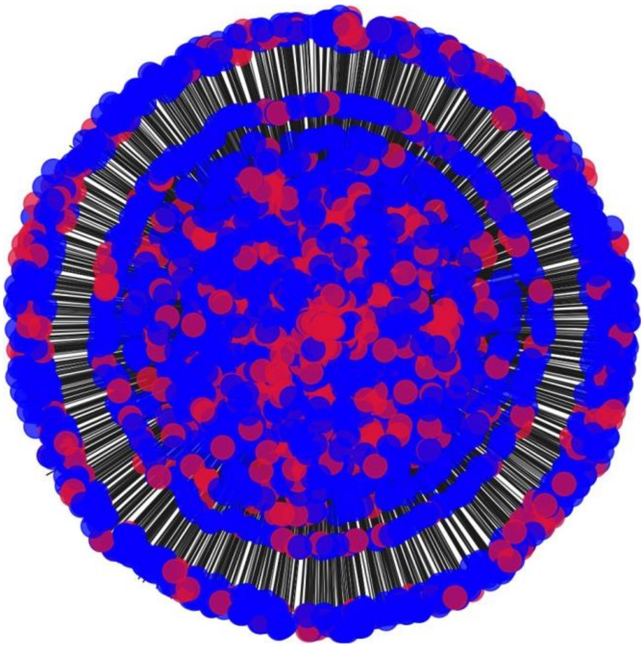


Age = 65

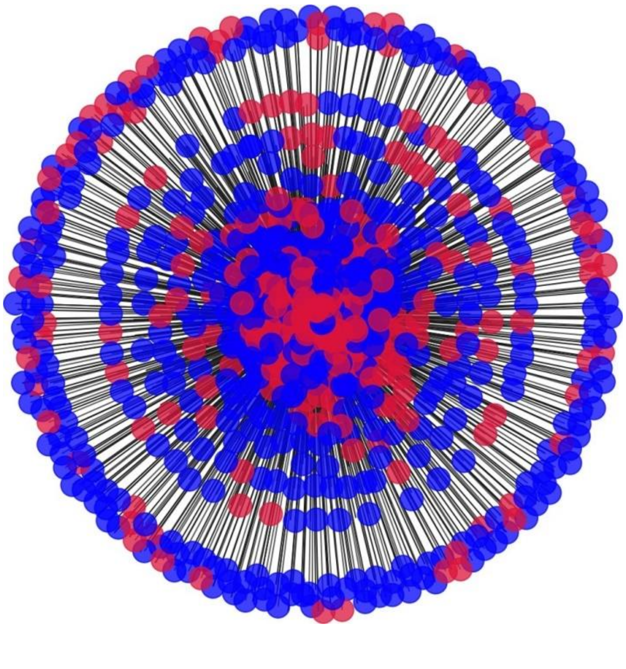
We notice that with age, tenure increases.

3. Age v/s Likes Received

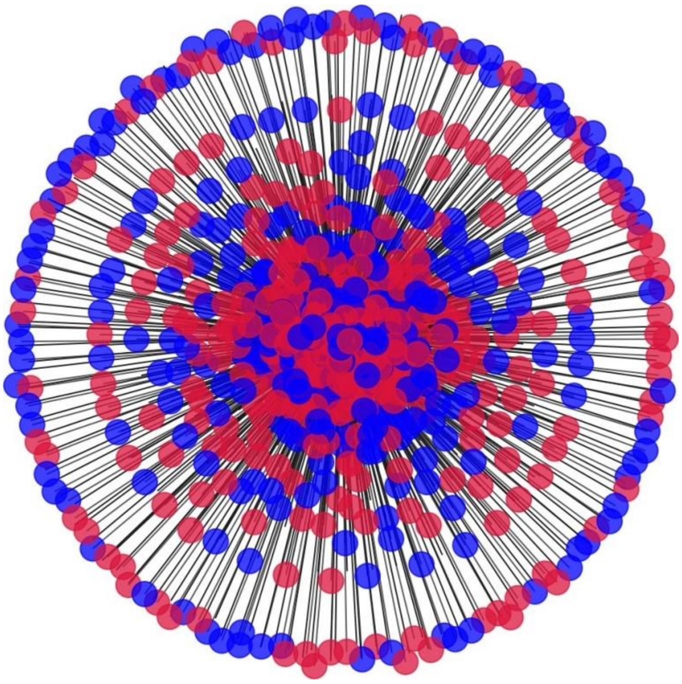
[Male = Blue, Female = Red and Edge length = Likes Received]



Age = 20



Age = 40



Age = 65

From these graphs, we notice that with age the likes received start to form levels thus showing levels of popularity. Another inference can be made is that as age increases the number of men on the social network starts to drop off.

Results and discussion

Facebook Analysis

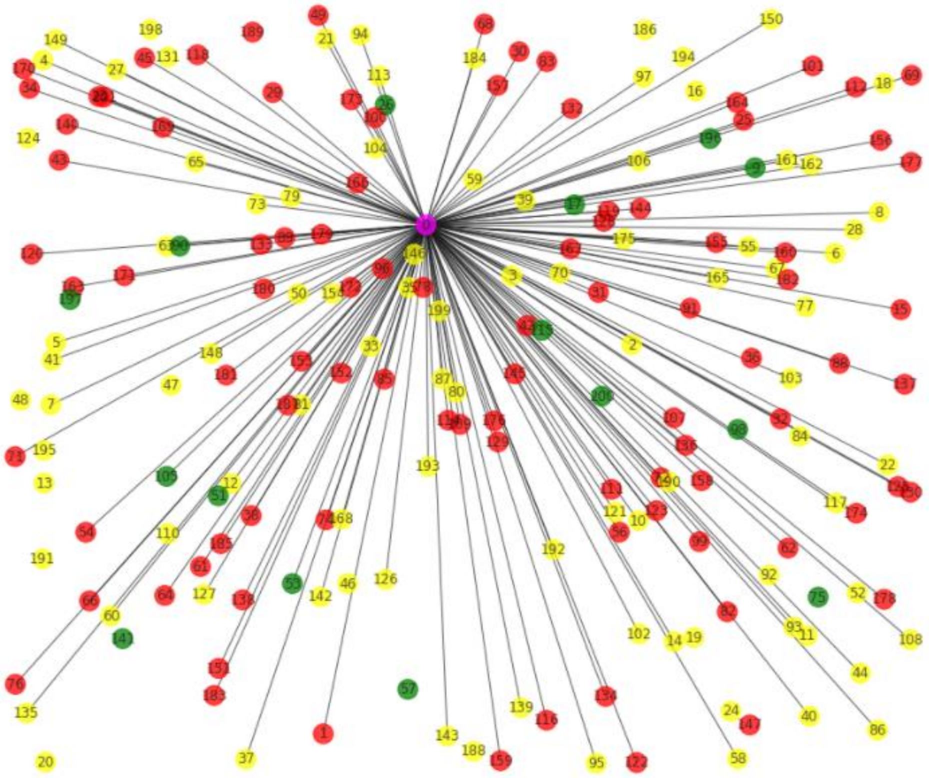
- First quartile range of female count is 37, median 96, mean is 342, third quartile is 244 and maximum count is 4923.
- First quartile range of male count is 27, median 74, mean is 165, third quartile is 182 and maximum count is 4917.
- Mean of female count is more than male count.
- Female has maximum number of friend count.
- Difference between female and male friend count:96-74=22
- Median is better measure than the mean because I have a long tail to the right, the mean would tend to over-estimate the average.
- From histogram of count vs age, we found that majority of users in the 20-30 age group
- There is spike in interest for 50-65 year olds.
- Total female likes were 3507665.
- Total male likes were 1430175.
- From number of likes, under category gender I found that number of female likes are approximately 2x times the sum of number of male likes.
- First quartile ranges of female's friendship initiated is 19, median 49, mean is 113.9, third quartile is 124.8 and maximum count is 3654.
- First quartile ranges of male's friendship initiated is 15, median 44, mean is 103.11, third quartile is 111.0 and maximum count is 4144.
- Sum of female friendship initiated is 4584894.
- Sum of male friendship initiated is 6037023.

- Number of Facebook users who are signing in from mobile are 63947 i.e. 65%. That means people are using mobile application more than web application.
- Young people (< 30years) seem to have people with extremely large number of friends up to 5000.
- Mostly, other age groups have less than 1000 friends.
- Also some spikes for ~68 and again for 100+ age.
- There is quite a bit of outliers even above 90% quartile for younger age groups.
- Correlation between friends and friend count is -0.027.
- Correlation between friends with age<70 and friend count is -0.1717.
- In the 13 to 18 age group, women have 2 to 2.5x more friends than men. But even otherwise, women consistently have more friends than men for age <= 70.
- I see that people who have stayed longer on FB have more friends.
- Grand mean plot basically tells us that much of the data is for the newer cohorts due to its proximity to the 2011-2012 cohort.
- Median friend rate=0.22, Maximum friend rate=417.
- People initiate friendship when they join and rate goes down with tenure as hypothesized.

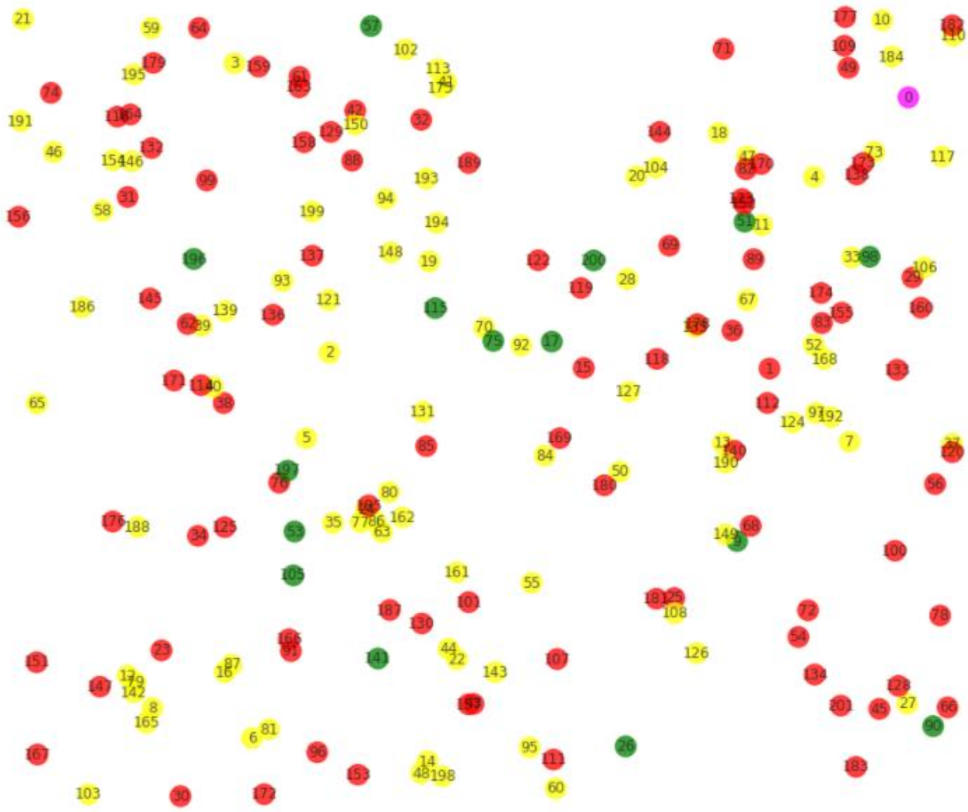
Twitter Analysis

- First, second and third quartile ranges of favorite count are all 0, mean is 3.19, standard deviation is 40.19 and maximum count is 1499.
- First, second and third quartile ranges of retweet count are 1,10 and 45 respectively. Mean is 48.21, standard deviation is 178.81 and maximum count is 4756.
- Minimum counts of both favorite and retweeted are same, 0.
- Polarity is the tweets' score, with minimum being -1 and maximum being 1.
- Mean and standard deviation of polarity are 0.07 and 0.22.
- The lowest longitude and latitude point to Ebolowa Cameroun,Africa, whereas the maximum point to Tibet.

(Node color = Sentiment of user's tweet, Edge Length = No. of Retweets) [Green = Positive, Red = Negative & Yellow = Positive]



From the above graph we infer that, central node acts center and it creates an edge for each retweeted count. If the particular user has is retweet count, its count acts as edge length. For the below, graph it depicts the retweeted and the retweeted count.



Conclusion:

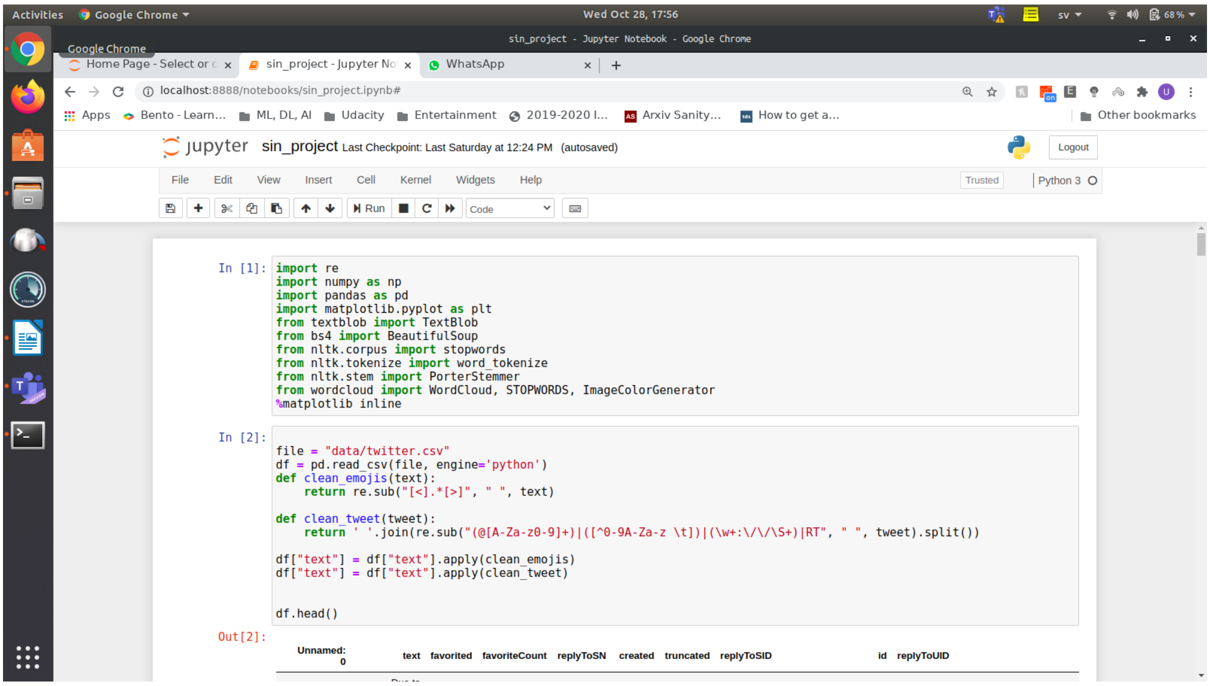
From the twitter data, we conclude that we can see that many of tweets regarding Covid-19 are focused on people asking for plasma donaters. The tweets which had positive sentiment are mostly tweets promoting social distancing and wearing masks. The tweets which had negative sentiment are mostly tweets regarding major event getting cancelled like IPL. Also as we can see from the graphs that the no. of tweets which had positive sentiment received more retweets than which had negative sentiments. This signifies that people promoted the positive sentiment more regarding than negative sentiments.

From the Facebook data, we can see that the age group that is most active belongs to "adult" category. Among the people who are 20 years old, no. of likes received by females are almost 2 times than that of males. Also this age group has the most number of friends. I see that people who have stayed longer on Facebook have more friends. People initiate friendship when they first join but later this number goes down.

References

1. https://eric.ed.gov/?id=ED131515
2. S. Gosling, S. Gaddis, S. Vazire, et al. Personality impressions based on Facebook profiles. In Proc. ICWSM 2007, 2007.
3. K. Caine, L. Kisselburgh, and L. Lareau. Audience visualization influences disclosures in online social networks. In Ext. Abst. CHI 2011, 2011.
4. L. Backstrom, E. Bakshy, J. Kleinberg, T. Lento, and I. Rosenn. Center of attention: How Facebook users allocate attention across friends. In Proc. ICWSM 2011, 2011.
5. C. Lampe, N. Ellison, and C. Steinfield. Changes in use and perception of Facebook. In Proc. SCW 2008, 2008.
6. https://www.kaggle.com/
7. http://msudatascience.com/blog/2016/8/27/quick-analysis-in-r-with-the-iris-dataset
8. https://www.r-bloggers.com/how-to-analyze-a-new-dataset-or-analyzing-supercar-data- part-1/
9. https://blog.datazar.com/exploratory-data-analysis-using-r-part-i-17e4e8e03961
10. https://en.wikipedia.org/wiki/R_(programming_language)

Appendix:

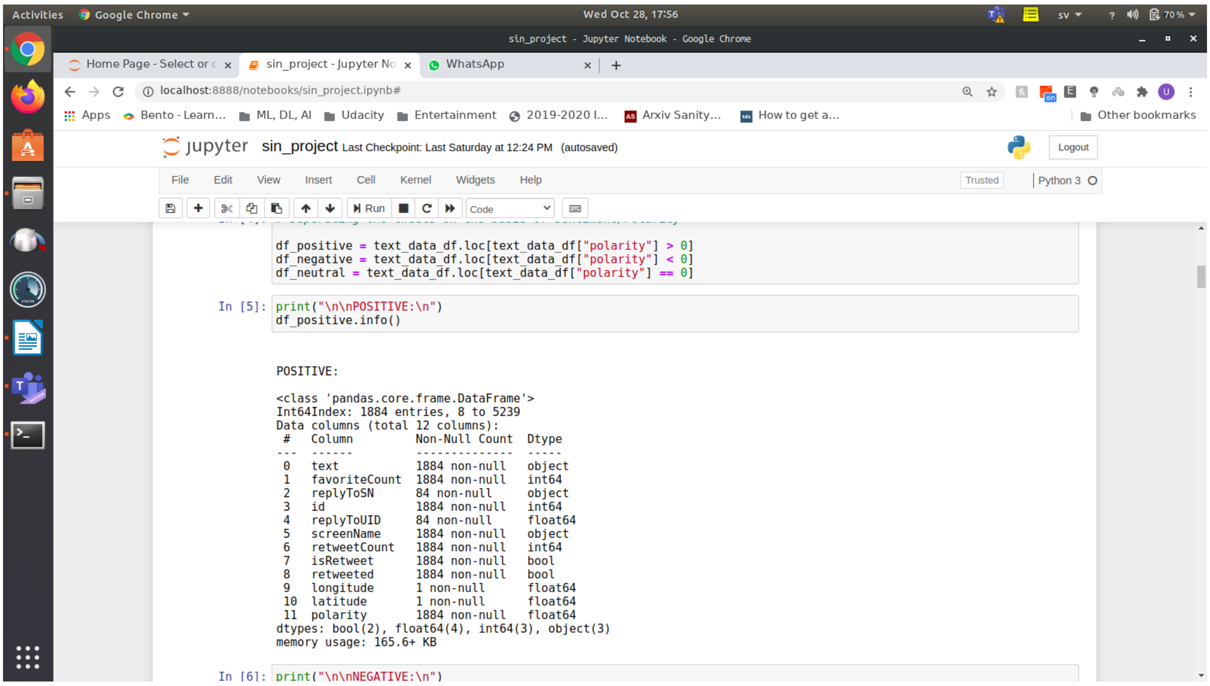


In [1]:
```python
import re
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from textblob import TextBlob
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
%matplotlib inline
```

In [2]:
```python
file = "data/twitter.csv"
df = pd.read_csv(file, engine='python')
def clean_emojis(text):
    return re.sub("[<].*[>]", " ", text)

def clean_tweet(tweet):
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|RT", " ", tweet).split())

df["text"] = df["text"].apply(clean_emojis)
df["text"] = df["text"].apply(clean_tweet)


df.head()
```

Out[2]:

Unnamed: 0	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID

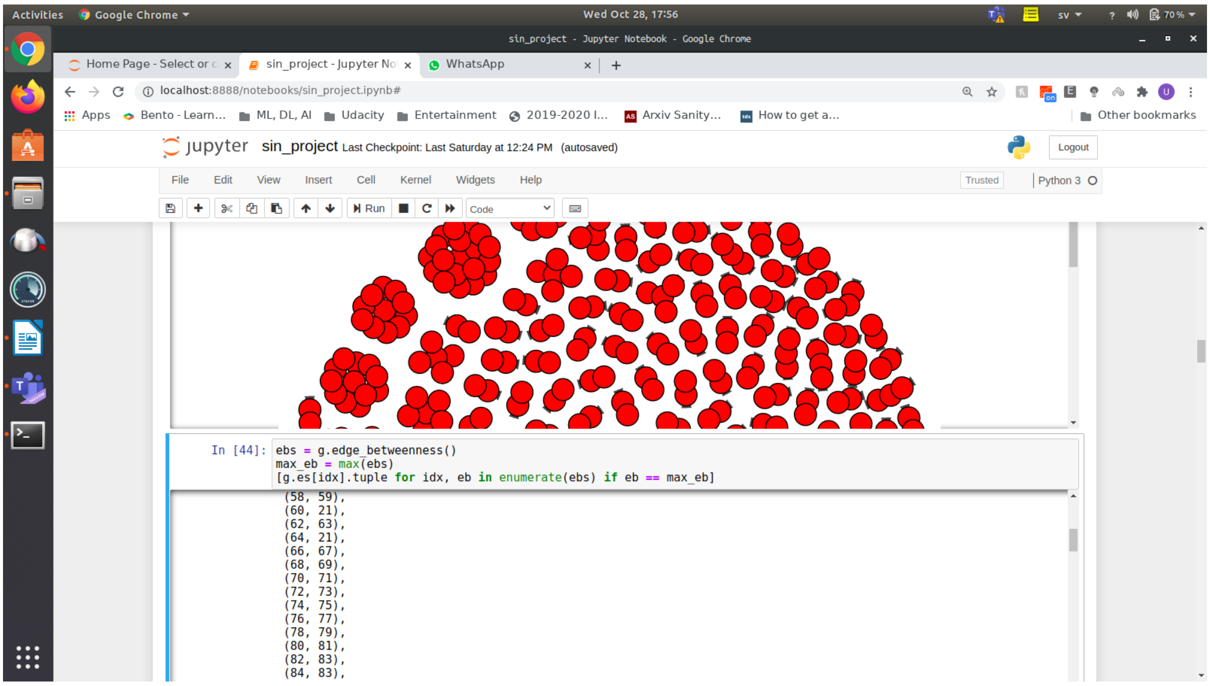


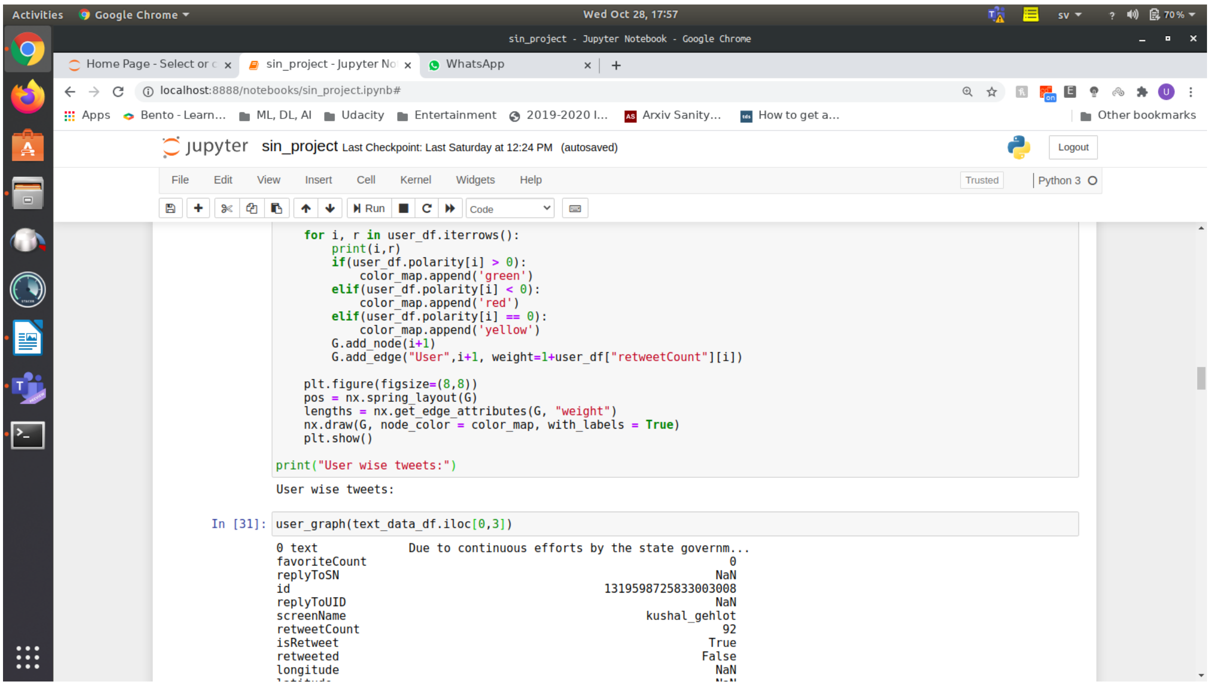
```python
df_positive = text_data_df.loc[text_data_df["polarity"] > 0]
df_negative = text_data_df.loc[text_data_df["polarity"] < 0]
df_neutral = text_data_df.loc[text_data_df["polarity"] == 0]
```

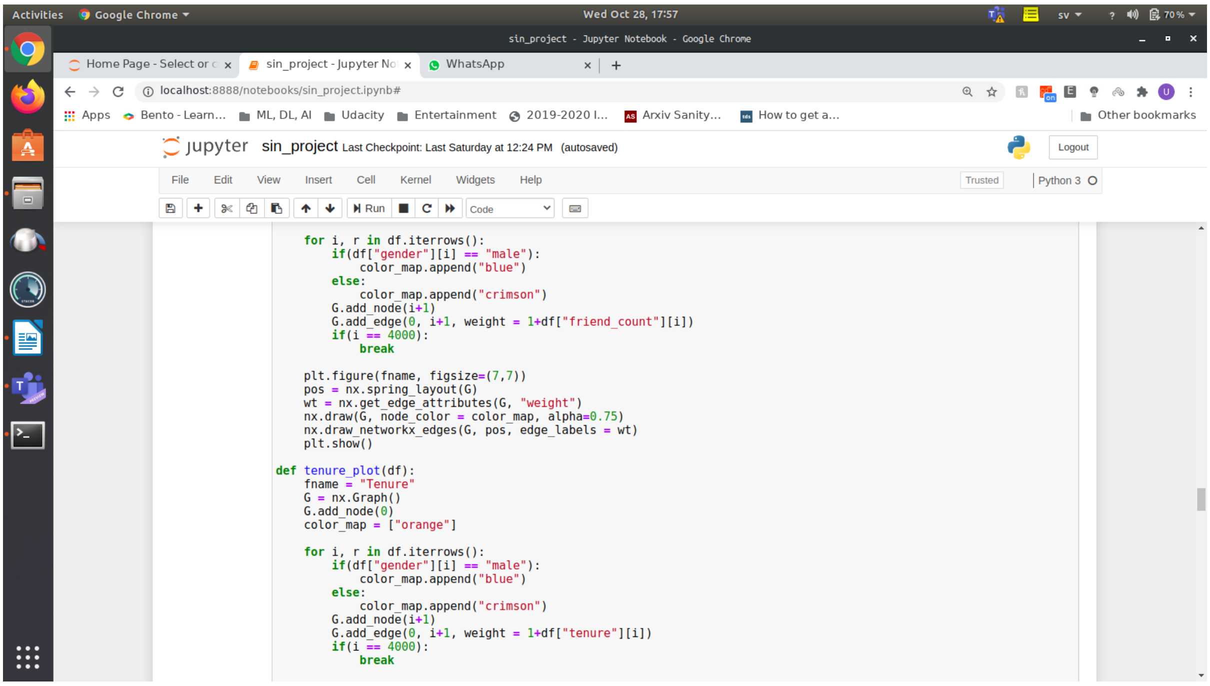
In [5]:
```python
print("\n\nPOSITIVE:\n")
df_positive.info()
```

```
POSITIVE:

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1884 entries, 8 to 5239
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   text           1884 non-null   object
 1   favoriteCount  1884 non-null   int64
 2   replyToSN      84 non-null     object
 3   id             1884 non-null   int64
 4   replyToUID     84 non-null     float64
 5   screenName     1884 non-null   object
 6   retweetCount   1884 non-null   int64
 7   isRetweet      1884 non-null   bool
 8   retweeted      1884 non-null   bool
 9   longitude      1 non-null      float64
 10  latitude       1 non-null      float64
 11  polarity       1884 non-null   float64
dtypes: bool(2), float64(4), int64(3), object(3)
memory usage: 165.6+ KB
```

In [6]:
```python
print("\n\nNEGATIVE:\n")
```


Activities
Google Chrome
Wed Oct 28, 17:56
sin_project - Jupyter Notebook - Google Chrome
Home Page - Select or c
sin_project - Jupyter No
WhatsApp
localhost:8888/notebooks/sin_project.ipynb#
Apps
Bento - Learn...
ML, DL, AI
Udacity
Entertainment
2019-2020 I...
Arxiv Sanity...
How to get a...
Other bookmarks
jupyter
sin_project Last Checkpoint: Last Saturday at 12:24 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [44]:
ebs = g.edge_betweenness()
max_eb = max(ebs)
[g.es[idx].tuple for idx, eb in enumerate(ebs) if eb == max_eb]
(58, 59),
(60, 21),
(62, 63),
(64, 21),
(66, 67),
(68, 69),
(70, 71),
(72, 73),
(74, 75),
(76, 77),
(78, 79),
(80, 81),
(82, 83),
(84, 83),


Activities
Google Chrome
Wed Oct 28, 17:57
sin_project - Jupyter Notebook - Google Chrome
Home Page - Select or c
sin_project - Jupyter No
WhatsApp
localhost:8888/notebooks/sin_project.ipynb#
Apps
Bento - Learn...
ML, DL, AI
Udacity
Entertainment
2019-2020 I...
Arxiv Sanity...
How to get a...
Other bookmarks
jupyter
sin_project Last Checkpoint: Last Saturday at 12:24 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
    for i, r in user_df.iterrows():
        print(i,r)
        if(user_df.polarity[i] > 0):
            color_map.append('green')
        elif(user_df.polarity[i] < 0):
            color_map.append('red')
        elif(user_df.polarity[i] == 0):
            color_map.append('yellow')
        G.add_node(i+1)
        G.add_edge("User",i+1, weight=1+user_df["retweetCount"][i])
    plt.figure(figsize=(8,8))
    pos = nx.spring_layout(G)
    lengths = nx.get_edge_attributes(G, "weight")
    nx.draw(G, node_color = color_map, with_labels = True)
    plt.show()
print("User wise tweets:")
User wise tweets:
In [31]:
user_graph(text_data_df.iloc[0,3])
0 text   Due to continuous efforts by the state governm...
favoriteCount   0
replyToSN   NaN
id   1319598725833003008
replyToUID   NaN
screenName   kushal_gehlot
retweetCount   92
isRetweet   True
retweeted   False
longitude   NaN


Activities
Google Chrome
Wed Oct 28, 17:57
sin_project - Jupyter Notebook - Google Chrome
Home Page - Select or c
sin_project - Jupyter No
WhatsApp
localhost:8888/notebooks/sin_project.ipynb#
Apps
Bento - Learn...
ML, DL, AI
Udacity
Entertainment
2019-2020 I...
Arxiv Sanity...
How to get a...
Other bookmarks
jupyter
sin_project Last Checkpoint: Last Saturday at 12:24 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
    for i, r in df.iterrows():
        if(df["gender"][i] == "male"):
            color_map.append("blue")
        else:
            color_map.append("crimson")
        G.add_node(i+1)
        G.add_edge(0, i+1, weight = 1+df["friend_count"][i])
        if(i == 4000):
            break
    plt.figure(fname, figsize=(7,7))
    pos = nx.spring_layout(G)
    wt = nx.get_edge_attributes(G, "weight")
    nx.draw(G, node_color = color_map, alpha=0.75)
    nx.draw_networkx_edges(G, pos, edge_labels = wt)
    plt.show()
def tenure_plot(df):
    fname = "Tenure"
    G = nx.Graph()
    G.add_node(0)
    color_map = ["orange"]
    for i, r in df.iterrows():
        if(df["gender"][i] == "male"):
            color_map.append("blue")
        else:
            color_map.append("crimson")
        G.add_node(i+1)
        G.add_edge(0, i+1, weight = 1+df["tenure"][i])
        if(i == 4000):
            break