Part2:

The project aims at performing analysis on a real time database called EDGAR. EDGAR is a database which maintains all the financial fillings of US companies. It is easily available to public for analysis.

The team has performed analysis in R programming language.

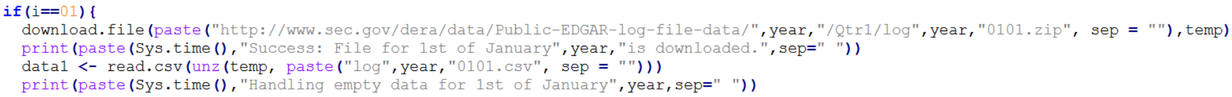
The project aimed at performing the following tasks:

1) WebScrapping
2) Anomaly detection
3) Missing data handling
4) Summarizing data
5) Recording the logs

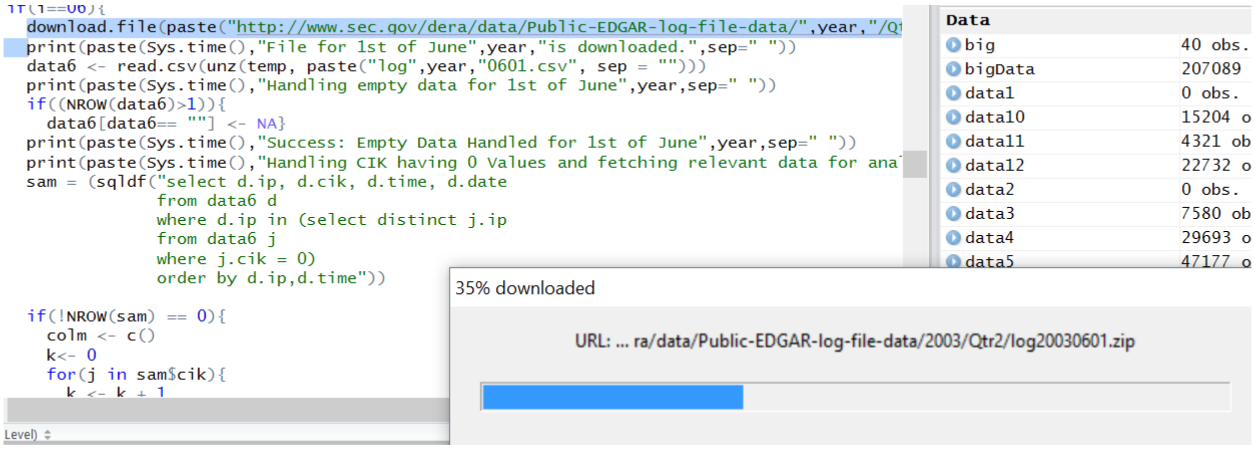
We have performed the following operations on the data available on edgar database.

Step1) A config file is to be configured by a YEAR so that the entire R program runs for the particular year. The program then performs all the operations on all 12 months of that year.

Step2) The file for all 12 months gets downloaded automatically by programmatically generated URLs.



```
if(i==01){
  download.file(paste("http://www.sec.gov/dera/data/Public-EDGAR-log-file-data/",year,"/Qtr1/log",year,"0101.zip", sep = ""),temp)
  print(paste(Sys.time(),"Success: File for 1st of January",year,"is downloaded.",sep=" "))
  data1 <- read.csv(unz(temp, paste("log",year,"0101.csv", sep = "")))
  print(paste(Sys.time(),"Handling empty data for 1st of January",year,sep=" "))
```



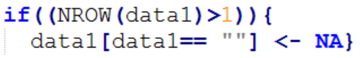
```
if(i==06){
  download.file(paste("http://www.sec.gov/dera/data/Public-EDGAR-log-file-data/",year,"/Qt
  print(paste(Sys.time(),"File for 1st of June",year,"is downloaded.",sep=" "))
  data6 <- read.csv(unz(temp, paste("log",year,"0601.csv", sep = "")))
  print(paste(Sys.time(),"Handling empty data for 1st of June",year,sep=" "))
  if((NROW(data6)>1)){
    data6[data6== ""] <- NA}
  print(paste(Sys.time(),"Success: Empty Data Handled for 1st of June",year,sep=" "))
  print(paste(Sys.time(),"Handling CIK having 0 Values and fetching relevant data for ana
  sam = (sqldf("select d.ip, d.cik, d.time, d.date
               from data6 d
               where d.ip in (select distinct j.ip
               from data6 j
               where j.cik = 0)
               order by d.ip,d.time"))

  if(!NROW(sam) == 0){
    colm <- c()
    k<- 0
    for(j in sam$cik){
      k <- k + 1
```

Data	
big	40 obs.
bigData	207089
data1	0 obs.
data10	15204 o
data11	4321 ob
data12	22732 o
data2	0 obs.
data3	7580 ob
data4	29693 o
data5	47177 o

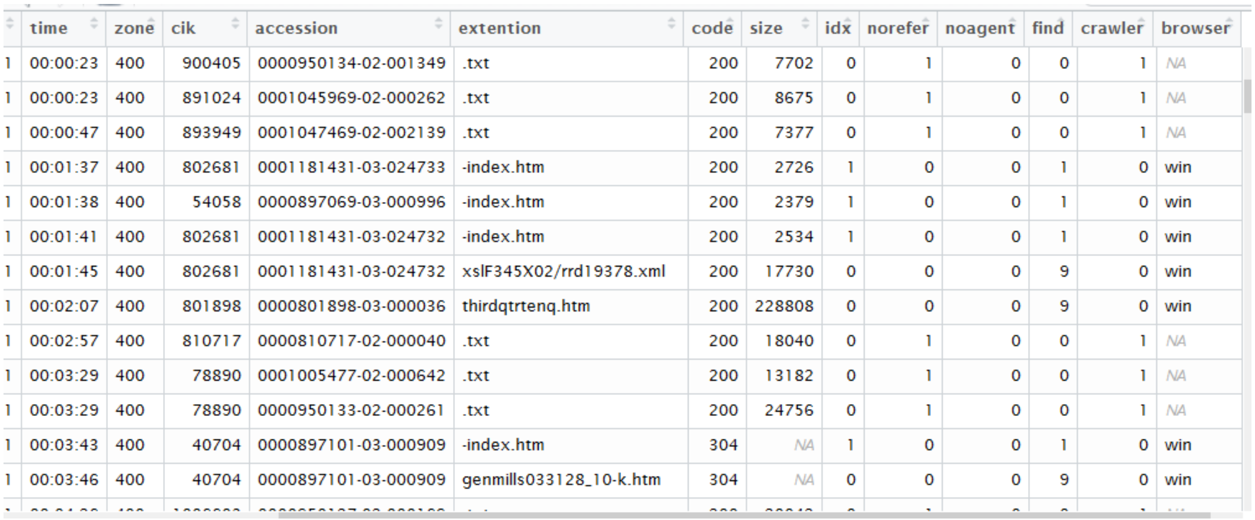
35% downloaded

URL: ... ra/data/Public-EDGAR-log-file-data/2003/Qtr2/log20030601.zip

Step 3) All empty data is handled by inserting NA. Data is handled only if data exists in that particular month.



```
if((NROW(data1)>1)){
  data1[data1== ""] <- NA}
```

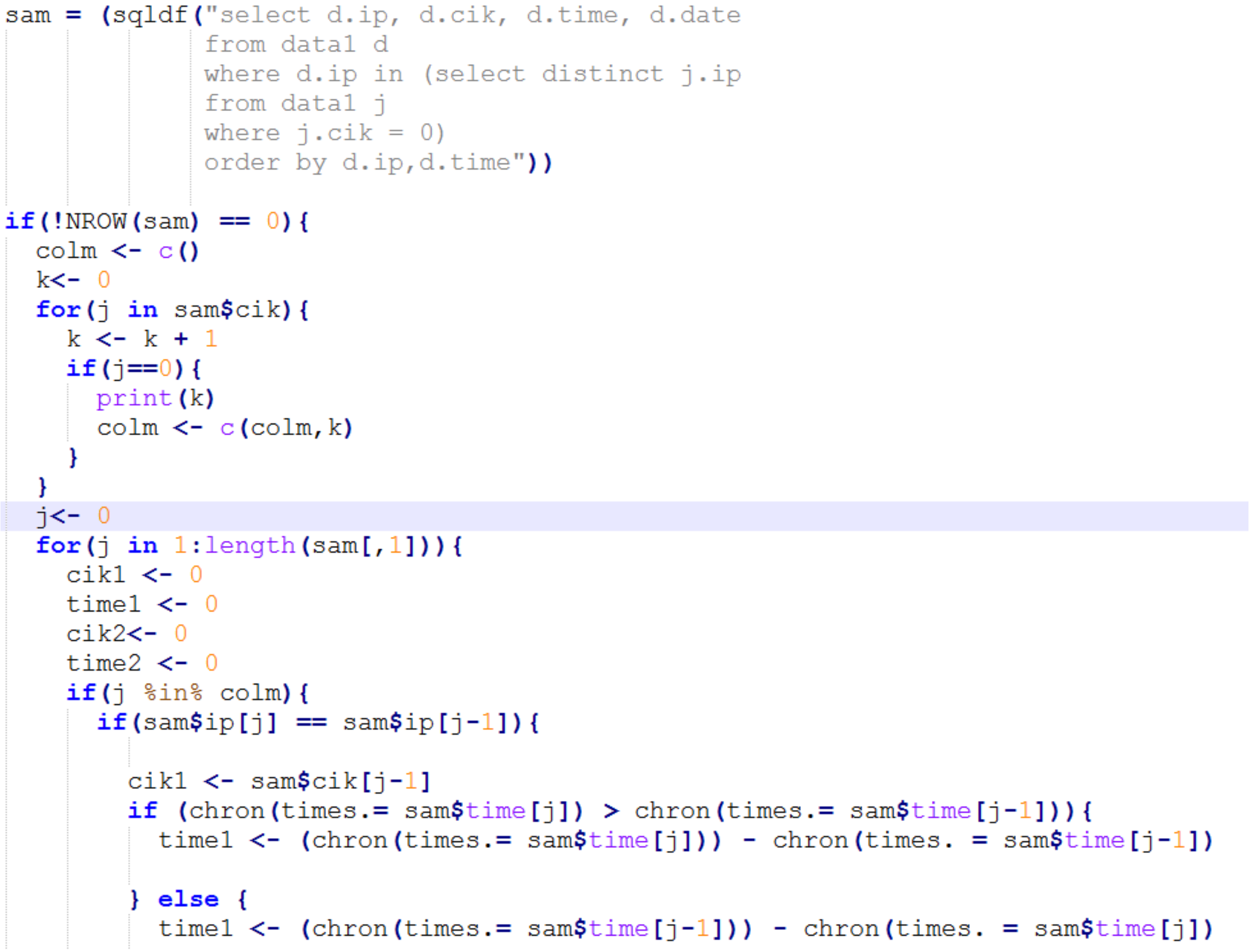


	time	zone	cik	accession	extention	code	size	idx	norefer	noagent	find	crawler	browser
1	00:00:23	400	900405	0000950134-02-001349	.txt	200	7702	0	1	0	0	1	NA
1	00:00:23	400	891024	0001045969-02-000262	.txt	200	8675	0	1	0	0	1	NA
1	00:00:47	400	893949	0001047469-02-002139	.txt	200	7377	0	1	0	0	1	NA
1	00:01:37	400	802681	0001181431-03-024733	-index.htm	200	2726	1	0	0	1	0	win
1	00:01:38	400	54058	0000897069-03-000996	-index.htm	200	2379	1	0	0	1	0	win
1	00:01:41	400	802681	0001181431-03-024732	-index.htm	200	2534	1	0	0	1	0	win
1	00:01:45	400	802681	0001181431-03-024732	xslF345X02/rrd19378.xml	200	17730	0	0	0	9	0	win
1	00:02:07	400	801898	0000801898-03-000036	thirdqtrtenq.htm	200	228808	0	0	0	9	0	win
1	00:02:57	400	810717	0000810717-02-000040	.txt	200	18040	0	1	0	0	1	NA
1	00:03:29	400	78890	0001005477-02-000642	.txt	200	13182	0	1	0	0	1	NA
1	00:03:29	400	78890	0000950133-02-000261	.txt	200	24756	0	1	0	0	1	NA
1	00:03:43	400	40704	0000897101-03-000909	-index.htm	304	NA	1	0	0	1	0	win
1	00:03:46	400	40704	0000897101-03-000909	genmills033128_10-k.htm	304	NA	0	0	0	9	0	win

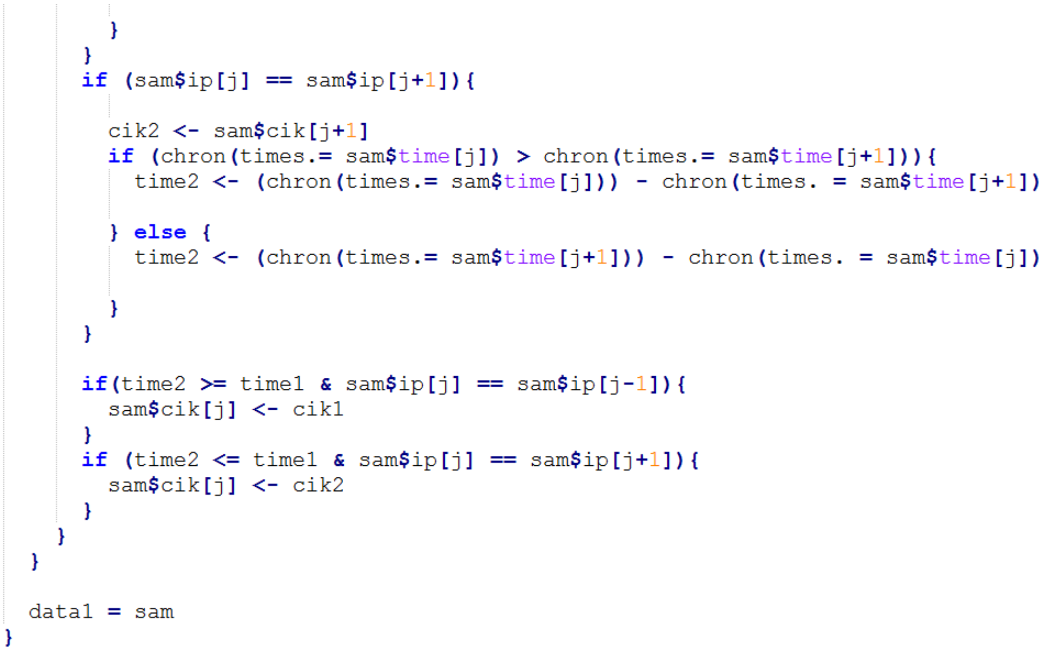
Step 4) On performing Analysis on data for year 2016, the team found out that there were many invalid CIKs. There were many error codes associated with the records. One of the error code was 404 which meant file not found error. This error has probably occurred because either the CIK is not valid or because of network issue.

We found out that there were CIKs with 0 values. 0 is a clear invalid CIK which has been handled programmatically by using the timestamp and IP. The nearest timestamp' CIK obviously has to be the value for 0 CIK.

Thus Handling Missing Data.

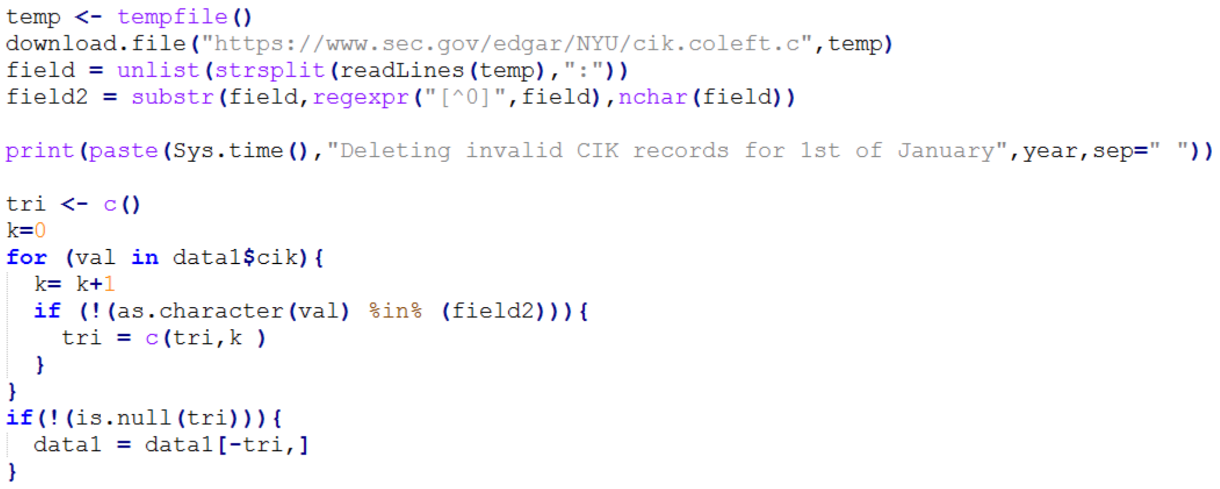


```
sam = (sqldf("select d.ip, d.cik, d.time, d.date
              from data1 d
              where d.ip in (select distinct j.ip
              from data1 j
              where j.cik = 0)
              order by d.ip,d.time"))

if(!NROW(sam) == 0){
  colm <- c()
  k<- 0
  for(j in sam$cik){
    k <- k + 1
    if(j==0){
      print(k)
      colm <- c(colm,k)
    }
  }
  j<- 0
  for(j in 1:length(sam[,1])){
    cik1 <- 0
    time1 <- 0
    cik2<- 0
    time2 <- 0
    if(j %in% colm){
      if(sam$ip[j] == sam$ip[j-1]){

        cik1 <- sam$cik[j-1]
        if (chron(times.= sam$time[j]) > chron(times.= sam$time[j-1])){
          time1 <- (chron(times.= sam$time[j])) - chron(times. = sam$time[j-1])

        } else {
          time1 <- (chron(times.= sam$time[j-1])) - chron(times. = sam$time[j])
```



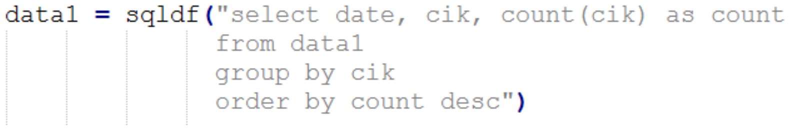
```
        }
      }
      if (sam$ip[j] == sam$ip[j+1]){

        cik2 <- sam$cik[j+1]
        if (chron(times.= sam$time[j]) > chron(times.= sam$time[j+1])){
          time2 <- (chron(times.= sam$time[j])) - chron(times. = sam$time[j+1])

        } else {
          time2 <- (chron(times.= sam$time[j+1])) - chron(times. = sam$time[j])

        }
      }

      if(time2 >= time1 & sam$ip[j] == sam$ip[j-1]){
        sam$cik[j] <- cik1
      }
      if (time2 <= time1 & sam$ip[j] == sam$ip[j+1]){
        sam$cik[j] <- cik2
      }
    }
  }

  data1 = sam
}
```

Step 5) Removing Invalid CIKs was achieved by making using of the CIK master list available on the website. The CIKs for each of the months were compared with the CIK master list and the invalid CIK records were deleted. Thus remaining with valid CIK records.



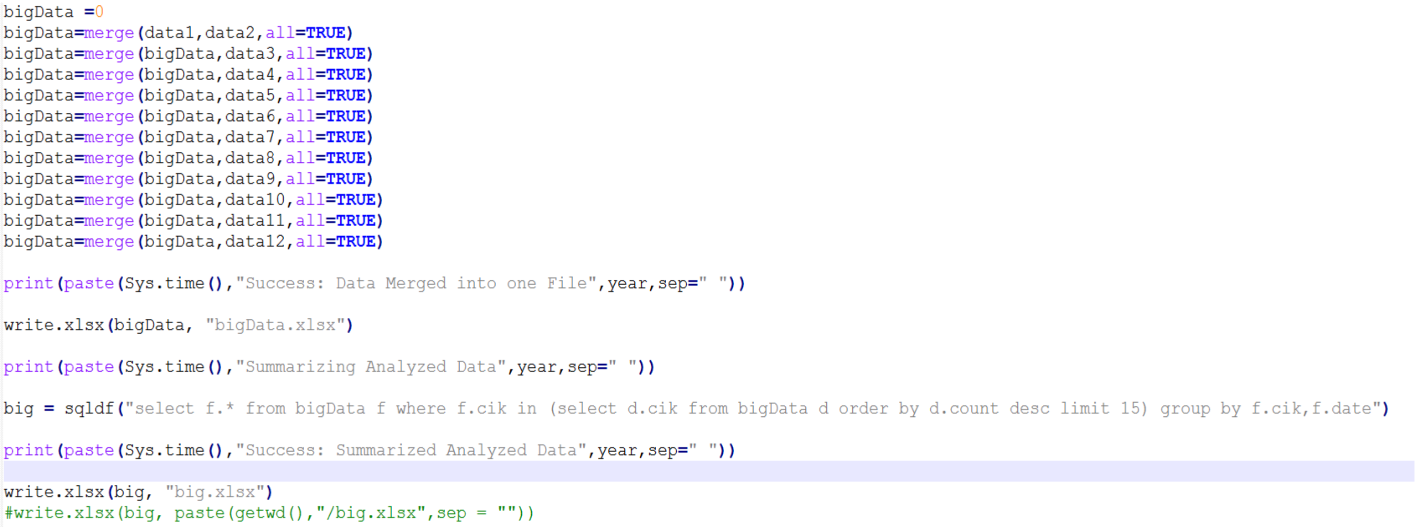
```
temp <- tempfile()
download.file("https://www.sec.gov/edgar/NYU/cik.coleft.c",temp)
field = unlist(strsplit(readLines(temp),":"))
field2 = substr(field,regexpr("[^0]",field),nchar(field))

print(paste(Sys.time(),"Deleting invalid CIK records for 1st of January",year,sep=" "))

tri <- c()
k=0
for (val in data1$cik){
  k= k+1
  if (!(as.character(val) %in% (field2))){
    tri = c(tri,k )
  }
}
if(!(is.null(tri))){
  data1 = data1[-tri,]
}
```

Step 6) Data for each of the months were summarized on basis of count of each of the CIKs. The records were ordered on basis of count in descending order.



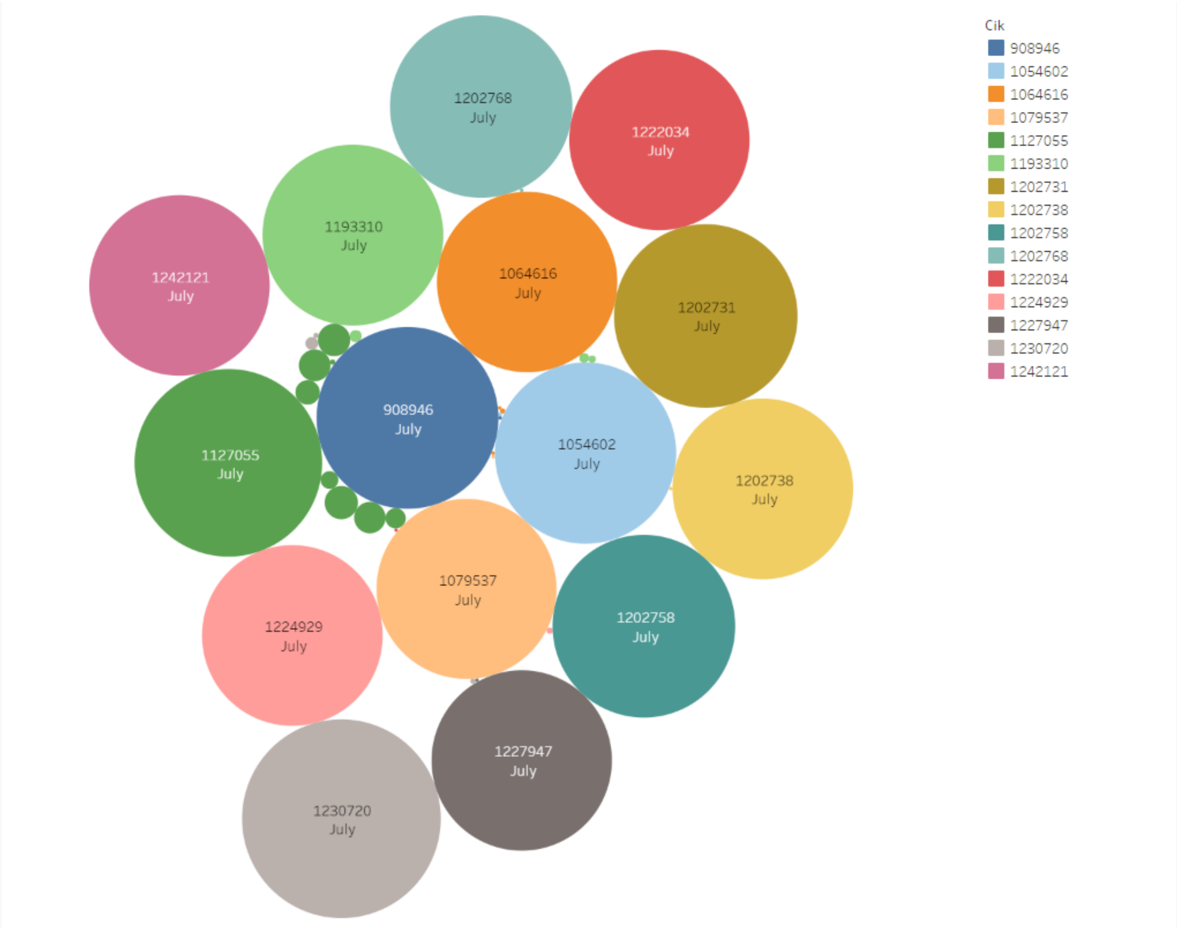
```
data1 = sqldf("select date, cik, count(cik) as count
               from data1
               group by cik
               order by count desc")
```

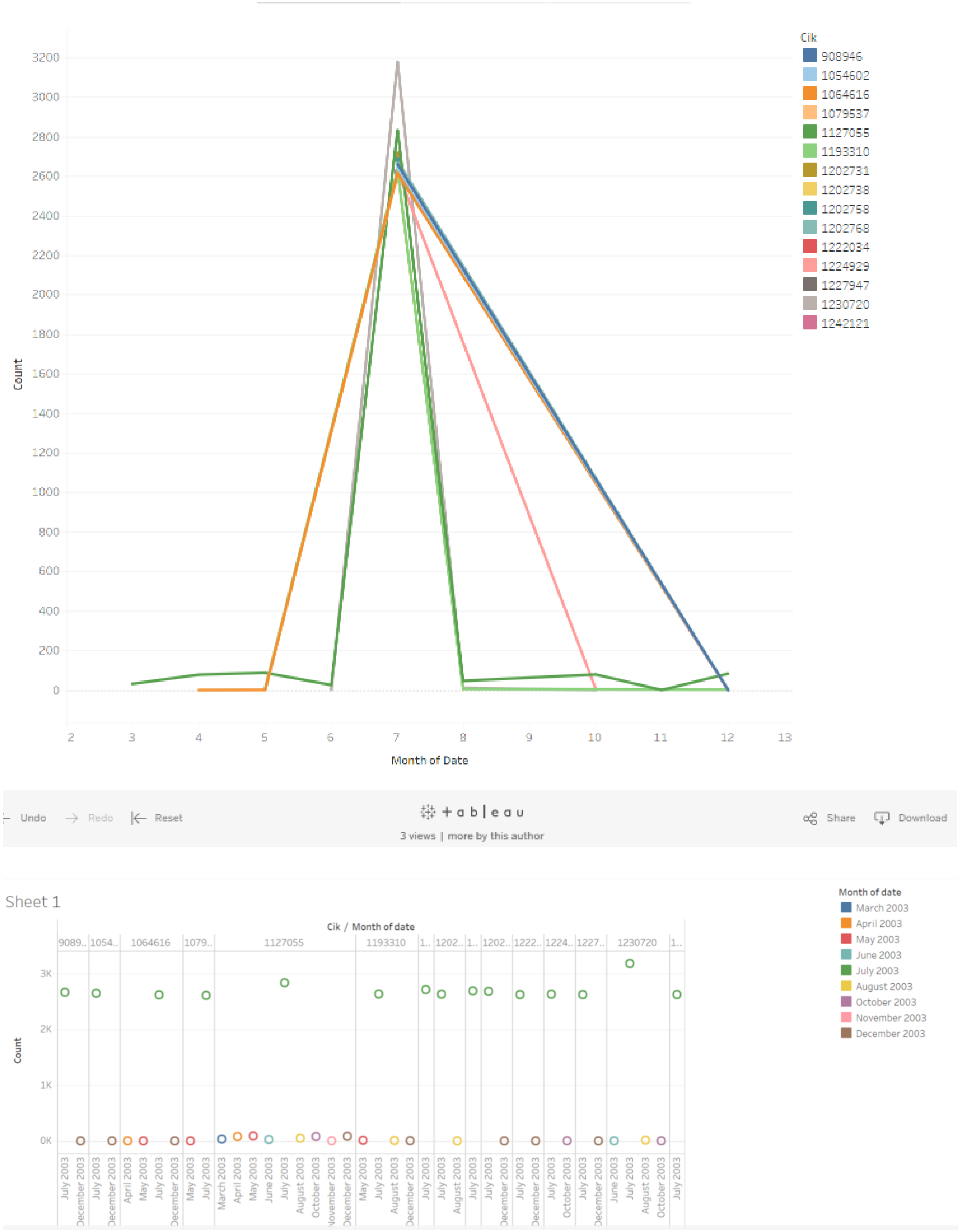
Step 7) Finally all the 12 files are merged together.



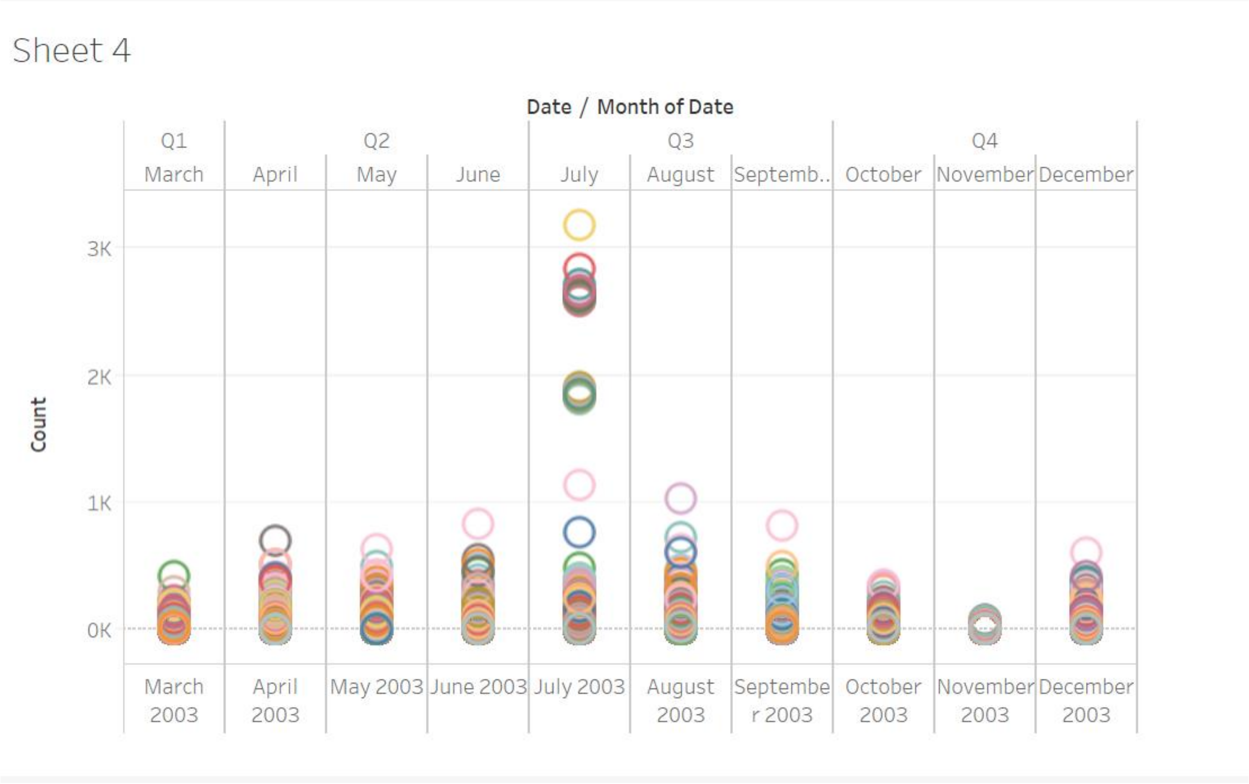
```
bigData =0
bigData=merge(data1,data2,all=TRUE)
bigData=merge(bigData,data3,all=TRUE)
bigData=merge(bigData,data4,all=TRUE)
bigData=merge(bigData,data5,all=TRUE)
bigData=merge(bigData,data6,all=TRUE)
bigData=merge(bigData,data7,all=TRUE)
bigData=merge(bigData,data8,all=TRUE)
bigData=merge(bigData,data9,all=TRUE)
bigData=merge(bigData,data10,all=TRUE)
bigData=merge(bigData,data11,all=TRUE)
bigData=merge(bigData,data12,all=TRUE)

print(paste(Sys.time(),"Success: Data Merged into one File",year,sep=" "))

write.xlsx(bigData, "bigData.xlsx")

print(paste(Sys.time(),"Summarizing Analyzed Data",year,sep=" "))

big = sqldf("select f.* from bigData f where f.cik in (select d.cik from bigData d order by d.count desc limit 15) group by f.cik,f.date")

print(paste(Sys.time(),"Success: Summarized Analyzed Data",year,sep=" "))

write.xlsx(big, "big.xlsx")
#write.xlsx(big, paste(getwd(),"/big.xlsx",sep = ""))
```

Step 8) The team has performed analysis over the companies which had maximum counts for any of the months over the year and has fetched data for those CIKs in order to understand the company's progress over the year.

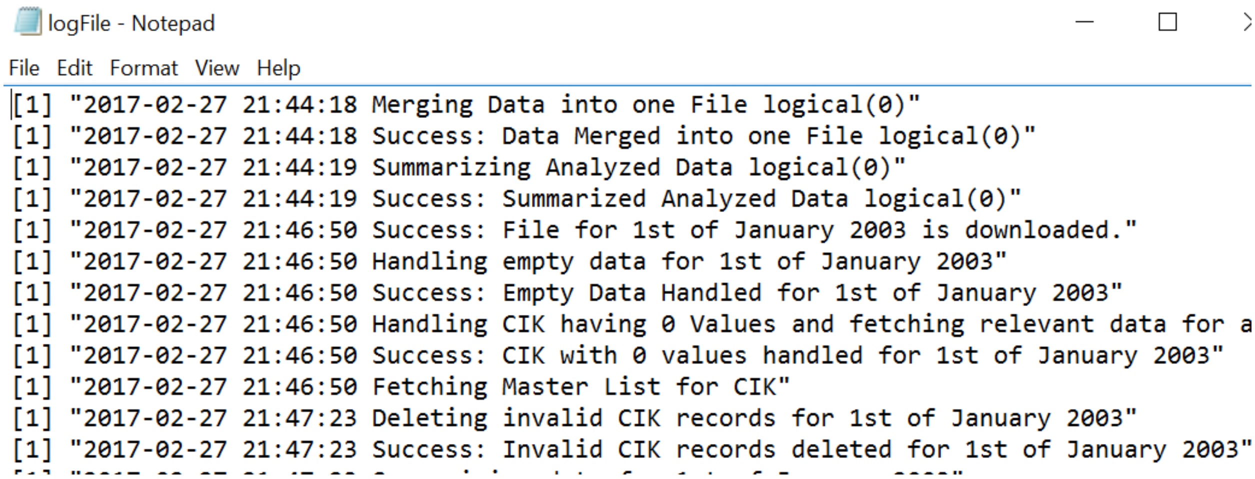




Step 9) The analysis has also been performed for all the CIKs performance.



Step 10) the log file is generated for entire part 2 session with timestamp.



```
[1] "2017-02-27 21:44:18 Merging Data into one File logical(0)"
[1] "2017-02-27 21:44:18 Success: Data Merged into one File logical(0)"
[1] "2017-02-27 21:44:19 Summarizing Analyzed Data logical(0)"
[1] "2017-02-27 21:44:19 Success: Summarized Analyzed Data logical(0)"
[1] "2017-02-27 21:46:50 Success: File for 1st of January 2003 is downloaded."
[1] "2017-02-27 21:46:50 Handling empty data for 1st of January 2003"
[1] "2017-02-27 21:46:50 Success: Empty Data Handled for 1st of January 2003"
[1] "2017-02-27 21:46:50 Handling CIK having 0 Values and fetching relevant data for a
[1] "2017-02-27 21:46:50 Success: CIK with 0 values handled for 1st of January 2003"
[1] "2017-02-27 21:46:50 Fetching Master List for CIK"
[1] "2017-02-27 21:47:23 Deleting invalid CIK records for 1st of January 2003"
[1] "2017-02-27 21:47:23 Success: Invalid CIK records deleted for 1st of January 2003"
```

The CSV output file and log file is uploaded to S3 bucket.

LINKS for Tableau:

https://public.tableau.com/profile/pratik3174#!/vizhome/Analysis3_6/Story3

https://public.tableau.com/profile/publish/AnalysisFull/Sheet4#!/publish-confirm