**DATA COLLECTION**

Go to the WhatsApp Group you want to analyze and import the Data.

WhatsApp ==>> Open the Group whose Data you want to analyze

==>Click on the 3 dotted vertical lined at the top right corner on WhatsApp

==>> Click More

==>> Export data (Choose whether you wish to export the Data with or without Media)

==>> Save the data and note the path you saved it to.

#####Please note that with media usually takes more time and space)####

==>>Select the desired destination you wish to export the Data to.

Open your R- Studio

==>> Click on Packages

==>> Search for readr (To know if the script (readr) for importing Data is installed or not)

In my case, it wasn't installed.

==>>`Click on install to download the readr script package

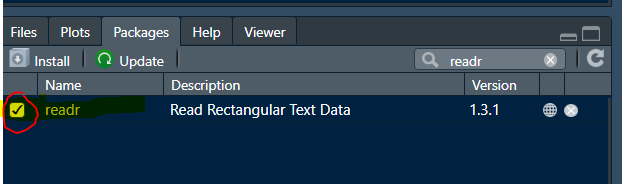
==>> When this is completed, you proceed to install

==>> Type: install.packages ("readr") to unpack the script package.

When this is completed

==>> Type library (readr) to install the script.

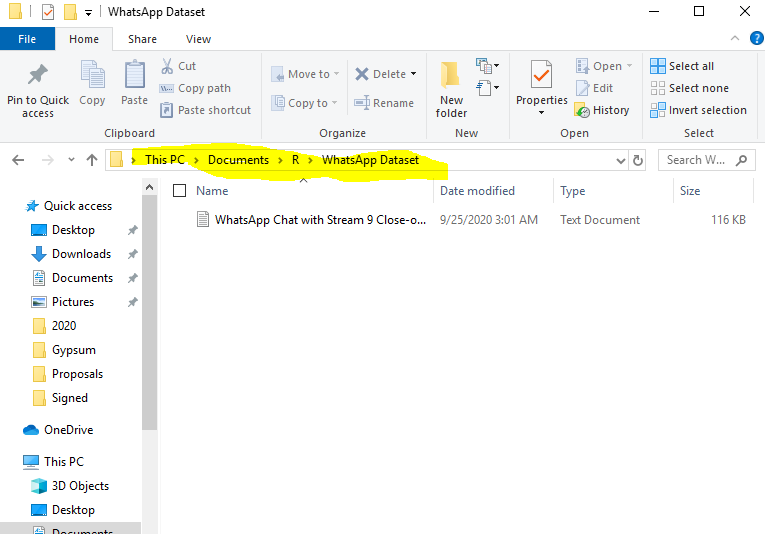
When this completed, you will notice that the readr script button will be ticked good



**IMPORTING THE DATA TO R-STUDIO**

###First set a working directory

==>> Copy the path your data is saved



==>> Type set working directory command

==>> setwd (C:\Users\user\Documents\R\WhatsApp Dataset)

==>> Put the quotation marks at the beginning and ending of the path

Setwd (“C:\Users\user\Documents\R\WhatsApp Dataset”)

== >> Change the slashes to the reverse direction, that is remove the backward slashes and replace with forward slashes.

== >> Setwd (“C:/Users/user/Documents/R/WhatsApp Dataset”)

== > Also put the forward slash at the last letter before the inverted comma

== >> Setwd (“C:/Users/user/Documents/R/WhatsApp Dataset/”)

== >> Save by typing ctrl + enter

###To check if you have saved it, type getwd()###

== >>Ctrl + Enter to save

**IMPORTING OUR DATA TO R, ALTERNATE METHOD**

library(rwhatsapp)

library(lubridate)

library(tidyverse)

library(tidytext)

library(kableExtra)

library(RColorBrewer)

library(rlang)

library(knitr)

setwd("C:/Users/user/Documents/R/WhatsApp Dataset2/")

ICTEduHub <- rwa\_read('ICTEduHub\_1.txt)

ICTEduHub <- ICTEduHub %>%

#########################################################################

library("tidyverse")

library(plotly)

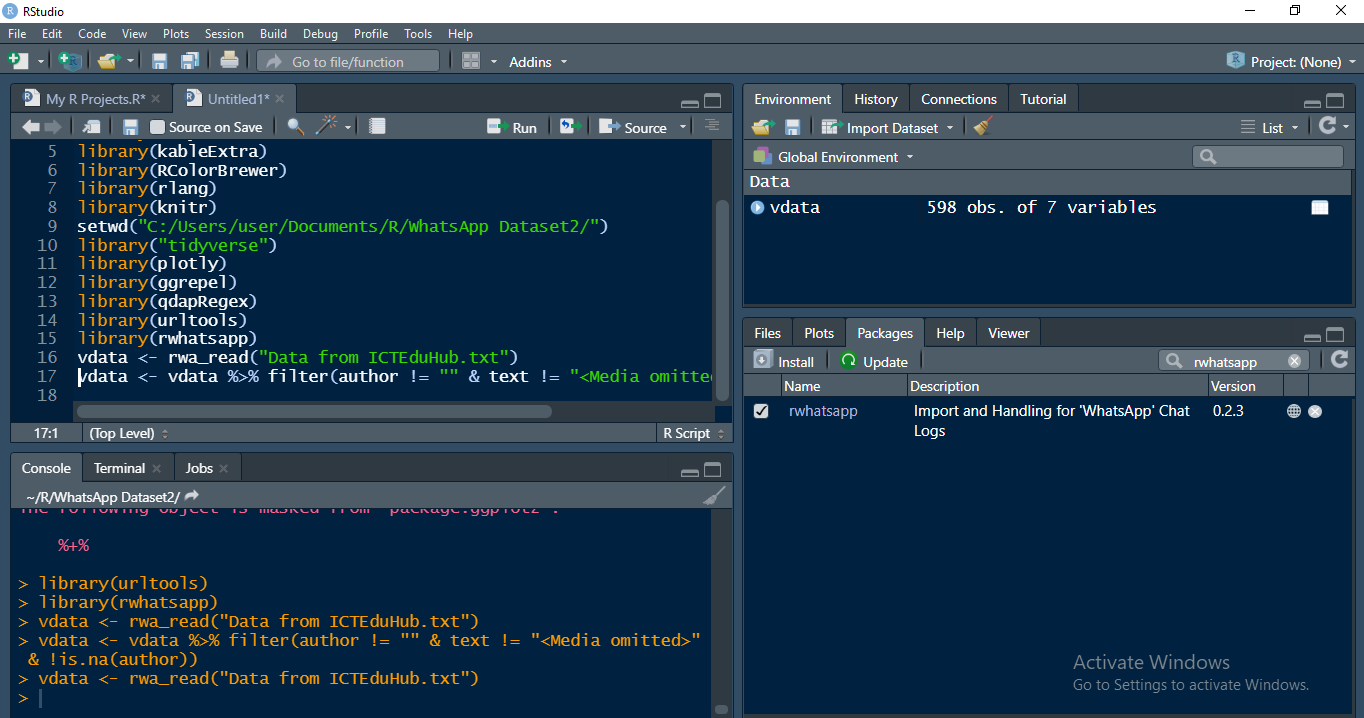
library(ggrepel)#geom\_label with no overlaps

library(qdapRegex)#exctract links

library(urltools)

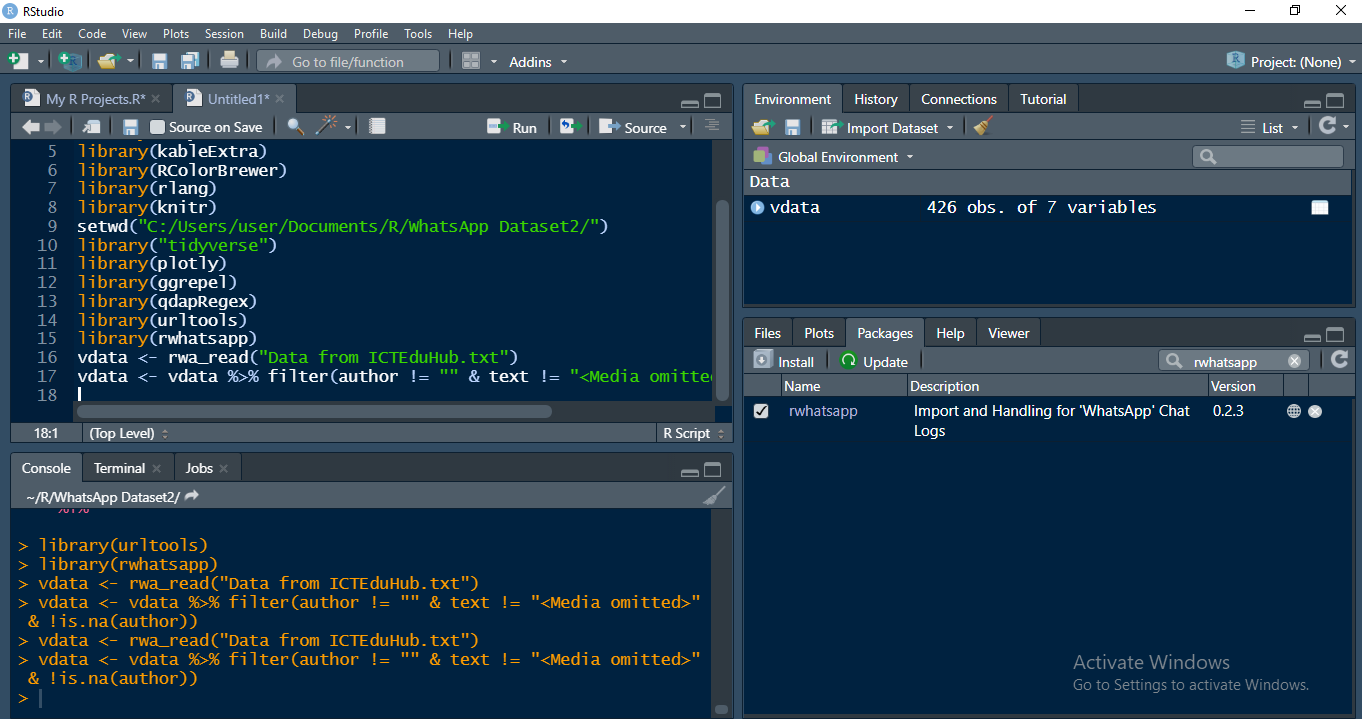
library(rwhatsapp)

vdata <- rwa\_read("Data from ICTEduHub.txt")



# Filter rows of <Media omitted> and messages that don’t have authors

vdata <- vdata %>% filter(author != "" & text != "<Media omitted>" & !is.na(author))



What understanding do we intend to gather from our data?

I begin with the 5 W and 1 H questions to draw insights from our data and they can always vary depending on ones' preference.

For this analysis, some questions I worked with include:

1. What days of the week were texts mostly sent?

2. What is the Monthly Activity and who was the most active member?

3. How many words were used in total?

4. What is the average texts exchanged per day of week?

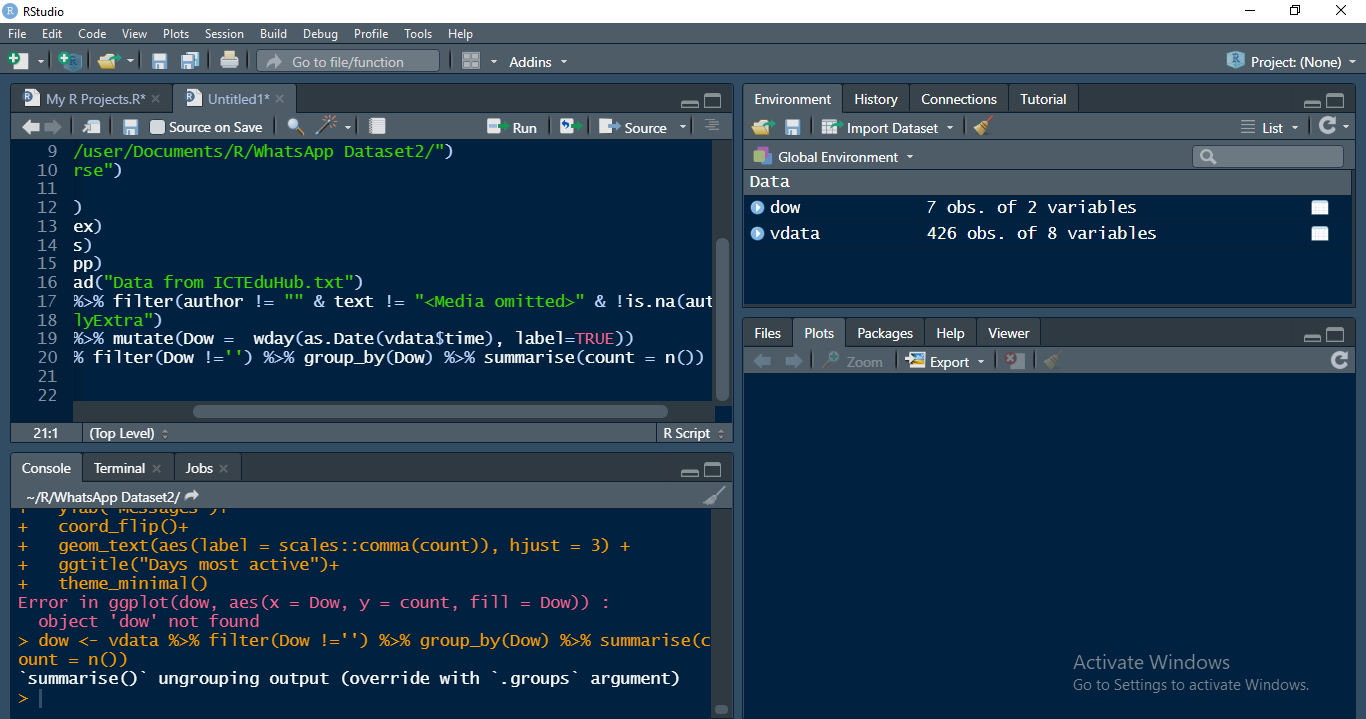
And goes the in-exhaustive list, which could be longer, more complex and more diverse all aimed at showing from our data and how it can speak loudly.

1. **What days of the week were texts mostly sent?**

# Mutate to add a column Dow ( Day of week )

vdata <- vdata %>% mutate(Dow = wday(as.Date(vdata$time), label=TRUE))

dow <- vdata %>% filter(Dow !='') %>% group\_by(Dow) %>% summarise(count = n())



**ggplot(dow,aes(x=Dow,y = count, fill = Dow))+**

**geom\_bar(stat = "identity")+**

**xlab("Days of the week")+**

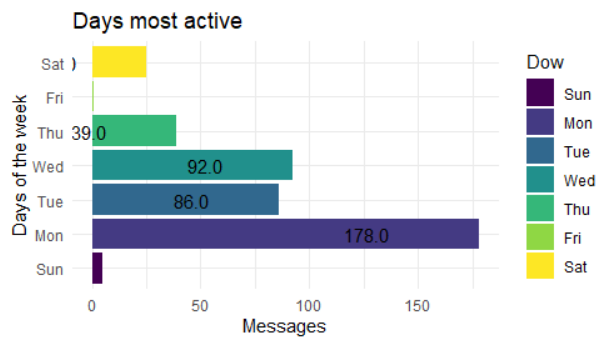
**ylab("Messages")+**

**coord\_flip()+**

**geom\_text(aes(label = scales::comma(count)), hjust = 3) +**

**ggtitle("Days most active")+**

**theme\_minimal()**

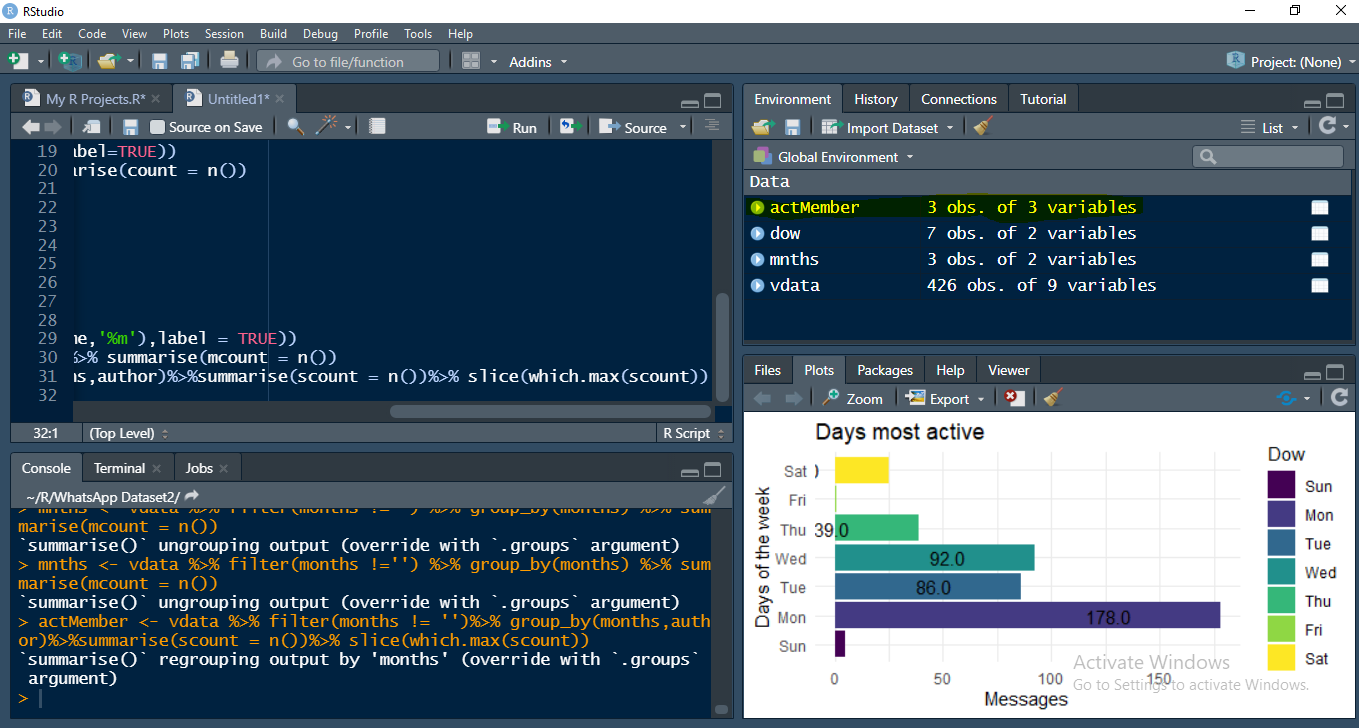


**2. What is the Monthly Activity and who was the most active member?**

vdata <- vdata %>% mutate(months = month(as.POSIXct(vdata$time,'%m'),label = TRUE))

mnths <- vdata %>% filter(months !='') %>% group\_by(months) %>% summarise(mcount = n())

actMember <- vdata %>% filter(months != '')%>% group\_by(months,author)%>%summarise(scount = n())%>% slice(which.max(scount))



mnthsactMember <- merge(mnths, actMember,by="months")

ggplot(mnthsactMember)+

geom\_bar(aes(x=months,y = mcount, fill = months),stat = "identity",width = 1)+

geom\_point(aes(x=months,y = scount,color = author),

size = 4, alpha = 0.5,

stat = "identity",

)+

# geom\_text(aes(x=months,y = scount,label = Name), vjust = 0.5,hjust = -1,color ="white")+

geom\_label(aes(x=months,y = scount,label = paste0(author," (",scount,")")),

fill = 'black', vjust = 0.5,hjust = -0.4,color ="white",alpha = 0.5,size = 3.5

)+

xlab("Months")+

ylab("Messages")+

coord\_flip()+

facet\_wrap(~author, ncol = 2, scales = "free\_y") +

ggtitle("MONTH ACTIVITY AND MOST ACTIVE MEMBER EACH MONTH")+

theme\_minimal(base\_size = 10)

**2b. WHO WAS THE MOST ACTIVE MEMBER PER MONTH?**

ggplot(mnthsactMember)+

geom\_bar(aes(x=months,y = mcount, fill = months),stat = "identity",width = 1)+

geom\_point(aes(x=months,y = scount,color = author),

size = 4, alpha = 0.5,

stat = "identity",

)+

# geom\_text(aes(x=months,y = scount,label = Name), vjust = 0.5,hjust = -1,color ="white")+

geom\_label(aes(x=months,y = scount,label = paste0(author," (",scount,")")),

fill = 'black', vjust = 0.5,hjust = -0.4,color ="white",alpha = 0.5,size = 3.5

)+

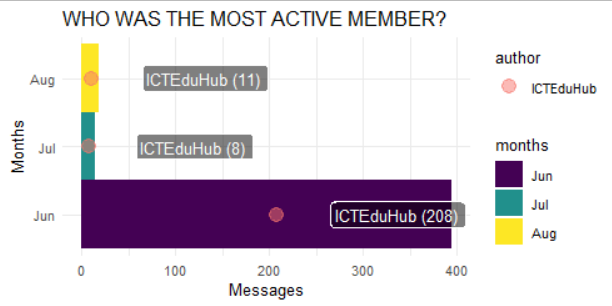
xlab("Months")+

ylab("Messages")+

coord\_flip()+

ggtitle("MONTH ACTIVITY AND MOST ACTIVE MEMBER EACH MONTH")+

theme\_minimal(base\_size = 10)



1. **HOW MANY WORDS WERE USED IN TOTAL?**

vdata = vdata %>% mutate(word\_count = sapply(strsplit(vdata$text, " "), length))

words <- vdata %>% group\_by(author)%>% summarise(count= sum(word\_count))

ggplot(words,aes(x=author,y = count, fill = author))+

geom\_bar(stat = "identity")+

xlab("Days of the week")+

ylab("Messages")+

geom\_text(size = 3,aes(label = paste0(scales::comma(count), " (",round(count/sum(count)\*100,0) ,"%)")), vjust = -1) +

ggtitle("WRITTEN WORDS")+

theme\_minimal()



1. **WHAT IS THE AVERAGE TEXTS EXCHANGED PER DAY OF WEEK?**

vdata <- vdata %>% mutate(Tarehe = format(as.Date(vdata$time,format="%Y-%m-%d %H:%M:%S"), format = "%d"))

txtfreq <- vdata %>% group\_by(Dow,Tarehe) %>% summarise(count =n())

ggplot(txtfreq, aes(Tarehe, Dow)) +

geom\_tile(aes(fill = count), colour = "white") +

scale\_fill\_distiller(palette = "Spectral", direction = -1) +

xlab("Days of month")+

ylab("Week days")+

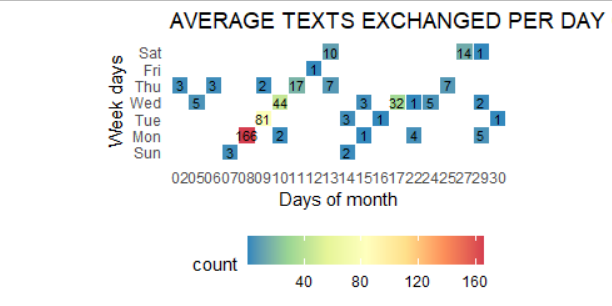
ggtitle("AVERAGE MESSAGES PER DAY OF WEEK ")+

theme\_minimal() +

theme(legend.position = "bottom", legend.key.width = unit(1, "cm"), panel.grid = element\_blank()) +

coord\_equal()+

geom\_text(aes(label=count),size = 3,color = "black")



<https://medium.com/@kalunazor/whatsapp-analysis-using-r-model-1ae5024451bc?sk=6b0dcdfcd3405bace14cc66102a872e9>

<https://medium.com/@kalunazor/whatsapp-analysis-using-r-model-1ae5024451bc>

<https://medium.com/@kalunazor>