Data analysis using R Programming

Amanpreet Singh

2022-04-09

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(readr)
Ted_Talk <- read_csv("~/Desktop/Assignments /Semester 2/Intro to analytics R programming/Ted Talk Data/
```

```
## Rows: 5440 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): title, author, date, link
## dbl (2): views, likes
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
View(Ted_Talk)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ’’/usr/bin/otool’ -L ’/Library/Frameworks/R.framework/Resources/
## modules/R_de.so’’ had status 1
```

```
attach(Ted_Talk)
str(Ted_Talk) #printing the structure of the dataset
```

```
## spec_tbl_df [5,440 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ title : chr [1:5440] "Climate action needs new frontline leadership" "The dark history of the over
##  $ author: chr [1:5440] "Ozawa Bineshi Albert" "Sydney Iaukea" "Martin Reeves" "James K. Thornton" .
##  $ date  : chr [1:5440] "December 2021" "February 2022" "September 2021" "October 2021" ...
##  $ views : num [1:5440] 404000 214000 412000 427000 2400 422000 412000 455000 66000 584000 ...
##  $ likes : num [1:5440] 12000 6400 12000 12000 72 12000 12000 13000 1900 17000 ...
##  $ link  : chr [1:5440] "https://ted.com/talks/ozawa_bineshi_albert_climate_action_needs_new_frontli
##  - attr(*, "spec")=
##   .. cols(
##   ..   title = col_character(),
##   ..   author = col_character(),
##   ..   date = col_character(),
##   ..   views = col_double(),
##   ..   likes = col_double(),
##   ..   link = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
ls(Ted_Talk) #listing the variables of the datasets
```

```
## [1] "author" "date"   "likes"  "link"   "title"  "views"
```

```
summary(Ted_Talk) #listing the variables of the datasets
```

```
##     title              author              date               views         
##  Length:5440        Length:5440        Length:5440        Min.   :     532  
##  Class :character   Class :character   Class :character   1st Qu.:  670750  
##  Mode  :character   Mode  :character   Mode  :character   Median : 1300000  
##                                                           Mean   : 2061576  
##                                                           3rd Qu.: 2100000  
##                                                           Max.   :72000000  
##      likes             link          
##  Min.   :     15   Length:5440       
##  1st Qu.:  20000   Class :character  
##  Median :  40500   Mode  :character  
##  Mean   :  62608                     
##  3rd Qu.:  65000                     
##  Max.   :2100000                     
```

```
head(Ted_Talk, 15) # printing the top 15 rows of the data sets
```

```
## # A tibble: 15 x 6
##    title                                         author date   views likes link 
##    <chr>                                         <chr>  <chr>  <dbl> <dbl> <chr>
##  1 "Climate action needs new frontline leadersh~ Ozawa~ Dece~ 404000 12000 http~
##  2 "The dark history of the overthrow of Hawaii" Sydne~ Febr~ 214000  6400 http~
##  3 "How play can spark new ideas for your busin~ Marti~ Sept~ 412000 12000 http~
##  4 "Why is China appointing judges to combat cl~ James~ Octo~ 427000 12000 http~
##  5 "Cement's carbon problem - and 2 ways to fix~ Mahen~ Octo~   2400    72 http~
##  6 "The tragedy of air pollution - and an urgen~ Rosam~ Octo~ 422000 12000 http~
##  7 "The myth of Narcissus and Echo"              Iseul~ Febr~ 412000 12000 http~
##  8 "You deserve the right to repair your stuff"  Gay G~ Augu~ 455000 13000 http~
##  9 "What nature can teach us about sustainable ~ Erin ~ Febr~  66000  1900 http~
## 10 "The origins of blackface and Black stereoty~ Dwan ~ Marc~ 584000 17000 http~
## 11 "A sex therapist's secret to rediscovering y~ Ian K~ Augu~  87000  2600 http~
## 12 "How do jetpacks work? And why don't we all ~ Richa~ Febr~ 213000  6400 http~
## 13 "What regret can teach you about living a go~ Danie~ Janu~ 622000 18000 http~
## 14 "How to fix the \"bugs\" in the net-zero cod~ Lucas~ Octo~ 526000 15000 http~
## 15 "\"Big Yellow Taxi\" / \"Song for Sunshine\"" Belle~ Octo~  23000   690 http~
```

```
author<- function(){
print("there are many authors who published the books")
}
author
```

```
## function(){
## print("there are many authors who published the books")
## }
```

```
# creating User-defined function using exsiting varibale in the DataSets
```

```
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
```

```
manipulation_tech=filter(Ted_Talk,views!='NA'& likes!='NA')
#manipulating data and filtering rowsbased on logical criteria by removing NA values from the datset
```

```
library("tidyr")
reshaping_columns <- Ted_Talk %>%
  gather(variable,value ,-c(views,likes))
#identifyed independent and dependent variables and reshaped them
```

```
Clean_dataSets <- na.omit(Ted_Talk) #removing missing values from the dataSets
```

```
missing_values <- complete.cases(Ted_Talk) #identifying and removing duplicate values from the data set
duplicate_data <- sum(duplicated(Ted_Talk))
```

```
distinct_value <- Ted_Talk %>% distinct()  #to find distinct values in the dataSet
```

```
drop_duplicates_likes <- distinct(Ted_Talk,`likes`, .keep_all= TRUE)
drop_duplicates_views <- distinct(Ted_Talk,`views`, .keep_all= TRUE)
# to drop all the duplicates in Views and Likes from the dataSet
```

```
desc_order <- Ted_Talk[order(-views,-likes), ]
#reordering Views and Likes in descending order
```

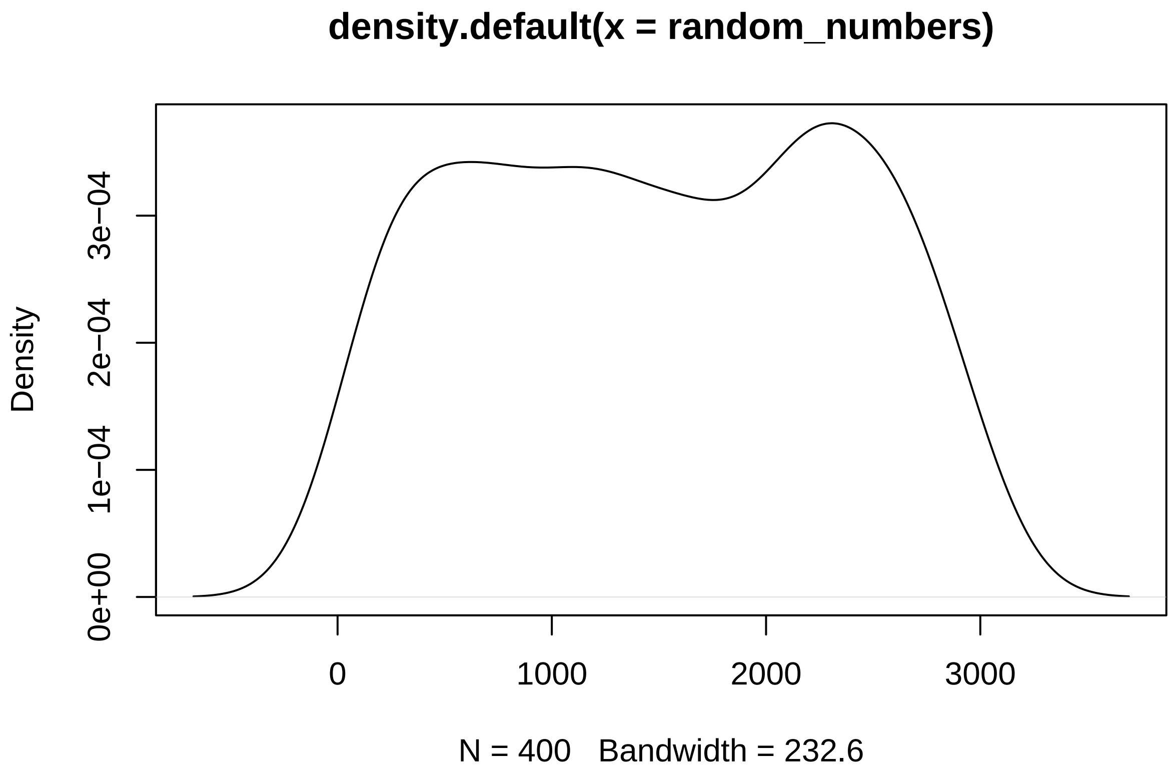
```
rename_columns <- Ted_Talk %>%
  rename(
   AUTHOR = author,
   TITLE = title
   )
View(rename_columns)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## modules/R_de.so'' had status 1
```

```
#Renaming 'author' and title colums to 'Author' and 'Title in the dataSet
```

```
Ted_Talk$Increase <- Ted_Talk$likes + 10
#Adding new Variable as a column in the dataSet by adding 10 to 'Likes'
```

```
set.seed(10)
random_numbers <- runif(400, min = 1, max = 3000)
plot(density(random_numbers))
```



```
summary(Ted_Talk)      #Summary stats for all the column of the data sets
```

```
##     title              author              date               views         
##  Length:5440        Length:5440        Length:5440        Min.   :     532  
##  Class :character   Class :character   Class :character   1st Qu.:  670750  
##  Mode  :character   Mode  :character   Mode  :character   Median : 1300000  
##                                                           Mean   : 2061576  
##                                                           3rd Qu.: 2100000  
##                                                           Max.   :72000000  
##      likes             link              Increase      
##  Min.   :     15   Length:5440        Min.   :     25  
##  1st Qu.:  20000   Class :character   1st Qu.:  20010  
##  Median :  40500   Mode  :character   Median :  40510  
##  Mean   :  62608                      Mean   :  62618  
##  3rd Qu.:  65000                      3rd Qu.:  65010  
##  Max.   :2100000                      Max.   :2100010
```

```
summary(Ted_Talk$`likes`)  #Summary stats for any specific column of the data set
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      15   20000   40500   62608   65000 2100000
```

```
mean(Ted_Talk$likes, na.rm = TRUE)  # mean of the dataSet
```

```
## [1] 62607.62
```

```
median(Ted_Talk$likes, na.rm = TRUE) #median of the dataSet
```

```
## [1] 40500
```

```
mode(Ted_Talk$likes) #mode of the dataSet
```

```
## [1] "numeric"
```

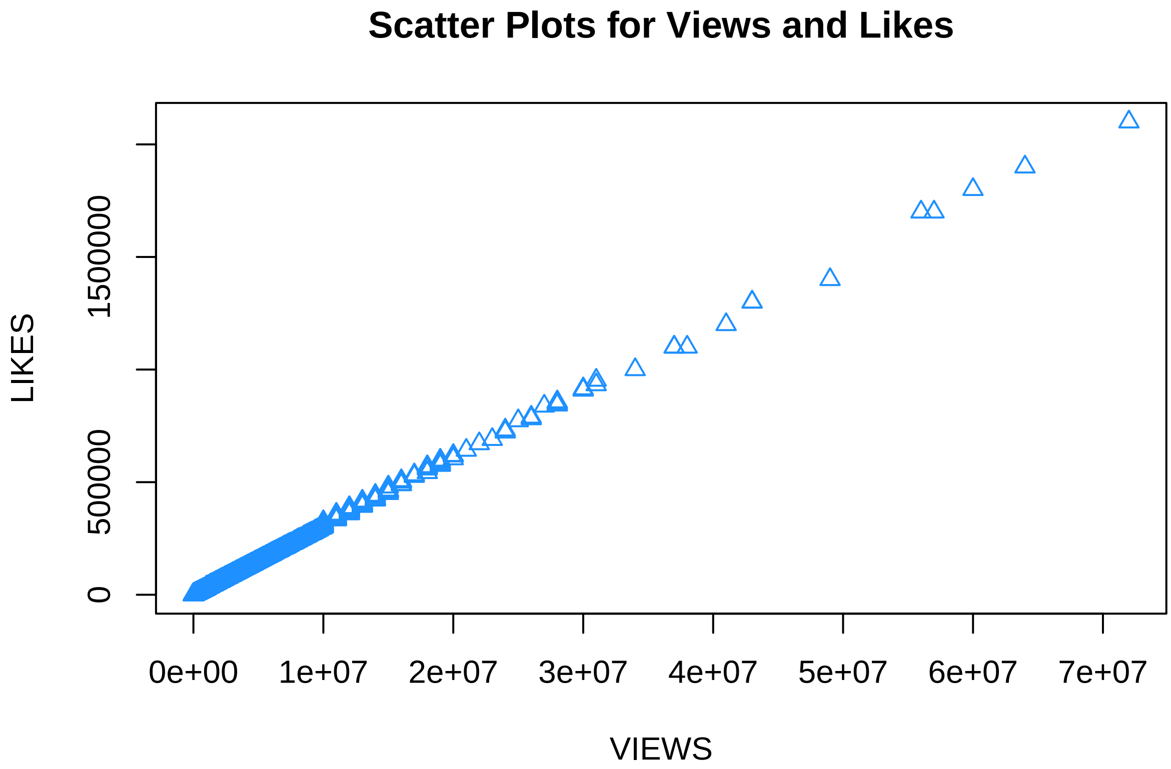
```
range(Ted_Talk$likes, na.rm = TRUE) #range of the dataSet
```

```
## [1]      15 2100000
```

```
sd(Ted_Talk$likes, na.rm = TRUE) #standad dDeviation of the dataSet
```

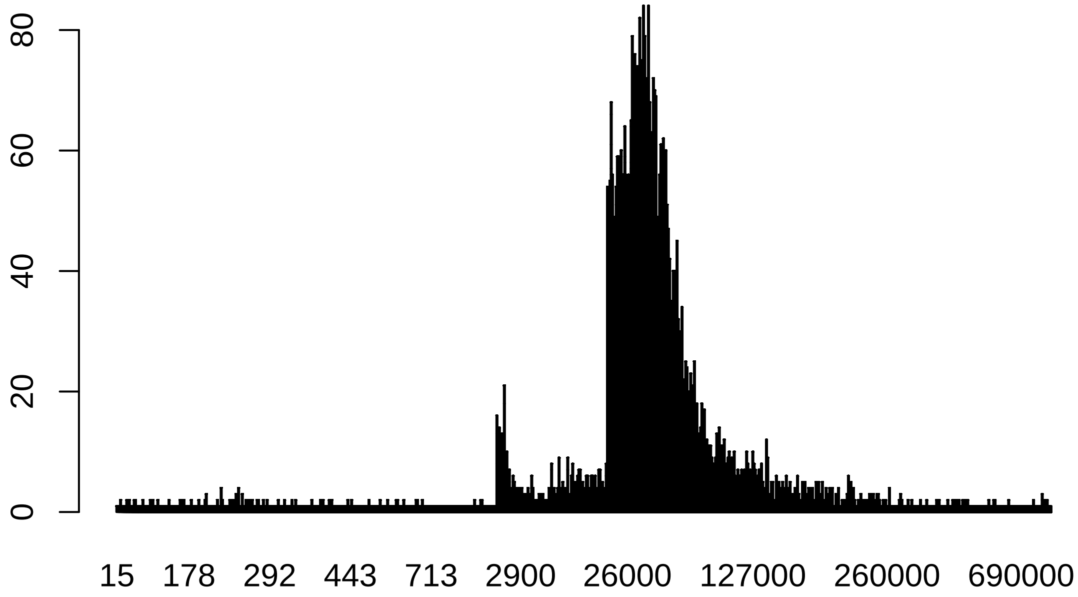
```
## [1] 107646.8
```

```
plot(views, likes, main = "Scatter Plots for Views and Likes", xlab = "VIEWS", ylab="LIKES", pch=24, co
```



```
#Scatter plot for Views and Likes
```

```
bplot <- table(Ted_Talk$views,Ted_Talk$likes)
barplot(bplot)
```



```
#Bar Plot for Views and Likes
```

```
cor(views,likes)
```

```
## [1] 0.999661
```

```
#3 Correlation between Views and Likes
```