MSBA_64060_Assignment 1_Atshaya Suresh

2023-09-06

```
knitr::opts_chunk$set(echo = TRUE)
```

```
tinytex::install_tinytex(force = TRUE)
```

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))
```

importing the csv file source: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023

```
#getting the working directory
getwd()
```

```
## [1] "C:/Users/Atshaya Suresh/Documents"
```

```
#importing the file
YouTube <- read.csv("Global YouTube Statistics.csv")

#To read the variables/characteristics
names(YouTube)
```

```
##  [1] "rank"
##  [2] "Youtuber"
##  [3] "subscribers"
##  [4] "video.views"
##  [5] "category"
##  [6] "Title"
##  [7] "uploads"
##  [8] "Country"
##  [9] "Abbreviation"
## [10] "channel_type"
## [11] "video_views_rank"
## [12] "country_rank"
## [13] "channel_type_rank"
## [14] "video_views_for_the_last_30_days"
## [15] "lowest_monthly_earnings"
## [16] "highest_monthly_earnings"
## [17] "lowest_yearly_earnings"
## [18] "highest_yearly_earnings"
## [19] "subscribers_for_last_30_days"
## [20] "created_year"
## [21] "created_month"
## [22] "created_date"
```

```
## [23] "Gross.tertiary.education.enrollment...."
## [24] "Population"
## [25] "Unemployment.rate"
## [26] "Urban_population"
## [27] "Latitude"
## [28] "Longitude"
```

printing the Descriptive Statistics of Quantitative Variables

```
summary(YouTube[, c("subscribers","video.views")])
```

```
##   subscribers         video.views
##  Min.   : 12300000   Min.   :0.000e+00
##  1st Qu.: 14500000   1st Qu.:4.288e+09
##  Median : 17700000   Median :7.761e+09
##  Mean   : 22982412   Mean   :1.104e+10
##  3rd Qu.: 24600000   3rd Qu.:1.355e+10
##  Max.   :245000000   Max.   :2.280e+11
```

printing the Descriptive Statistics of Categorical Variables (Since this does not communicate any valuable information, we will get other details for actionable insights)

```
summary(YouTube[, c("category","Country")])
```

```
##    category           Country
##  Length:995         Length:995
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

printing the Descriptive Statistics of Categorical Variables

```
# Tabulate the counts/Frequency
table(YouTube$category)
```

```
##
##      Autos & Vehicles                Comedy             Education
##                     2                    69                    45
##         Entertainment      Film & Animation                Gaming
##                   241                    46                    94
##         Howto & Style                Movies                 Music
##                    40                     2                   202
##                   nan       News & Politics Nonprofits & Activism
##                    46                    26                     2
##        People & Blogs        Pets & Animals  Science & Technology
##                   132                     4                    17
##                 Shows                Sports              Trailers
##                    13                    11                     2
##       Travel & Events
##                     1
```

```r
# Tabulate the proportions
prop.table(table(YouTube$category))
```

```
##
##      Autos & Vehicles                Comedy             Education
##           0.002010050           0.069346734           0.045226131
##         Entertainment      Film & Animation                Gaming
##           0.242211055           0.046231156           0.094472362
##         Howto & Style                Movies                 Music
##           0.040201005           0.002010050           0.203015075
##                  nan       News & Politics Nonprofits & Activism
##           0.046231156           0.026130653           0.002010050
##        People & Blogs        Pets & Animals  Science & Technology
##           0.132663317           0.004020101           0.017085427
##                Shows                Sports              Trailers
##           0.013065327           0.011055276           0.002010050
##       Travel & Events
##           0.001005025
```

Transforming the data (a) Z-Score Normalization (By creating a function and using it)

```r
normalize_z_score <- function(x) {
  return ((x - mean(x)) / sd(x))
}

normalized_YouTube_Subscribers <- normalize_z_score(YouTube$subscribers)

#After normalization, the mean of the Subscribers is Zero (in summary output)
summary(normalized_YouTube_Subscribers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6095 -0.4840 -0.3014  0.0000  0.0923 12.6678
```

Installing 'caret' package

```r
install.packages("caret")
```

```
## Installing package into 'C:/Users/Atshaya Suresh/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'caret' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Atshaya Suresh\AppData\Local\Temp\Rtmp8iwWwx\downloaded_packages
```

```r
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

Another method for Z-Score Normalization (Note: Mean becomes 0)

```
Subscribers_df <- as.data.frame(YouTube$subscribers)
norm_model_1<-preProcess(Subscribers_df, method = c("center","scale"))
Default_normalized1<-predict(norm_model_1,Subscribers_df)
summary(Default_normalized1)
```

```
##  YouTube$subscribers
##  Min.   :-0.6095
##  1st Qu.:-0.4840
##  Median :-0.3014
##  Mean   : 0.0000
##  3rd Qu.: 0.0923
##  Max.   :12.6678
```

(b) Transforming the data (Min-Max Normalization)

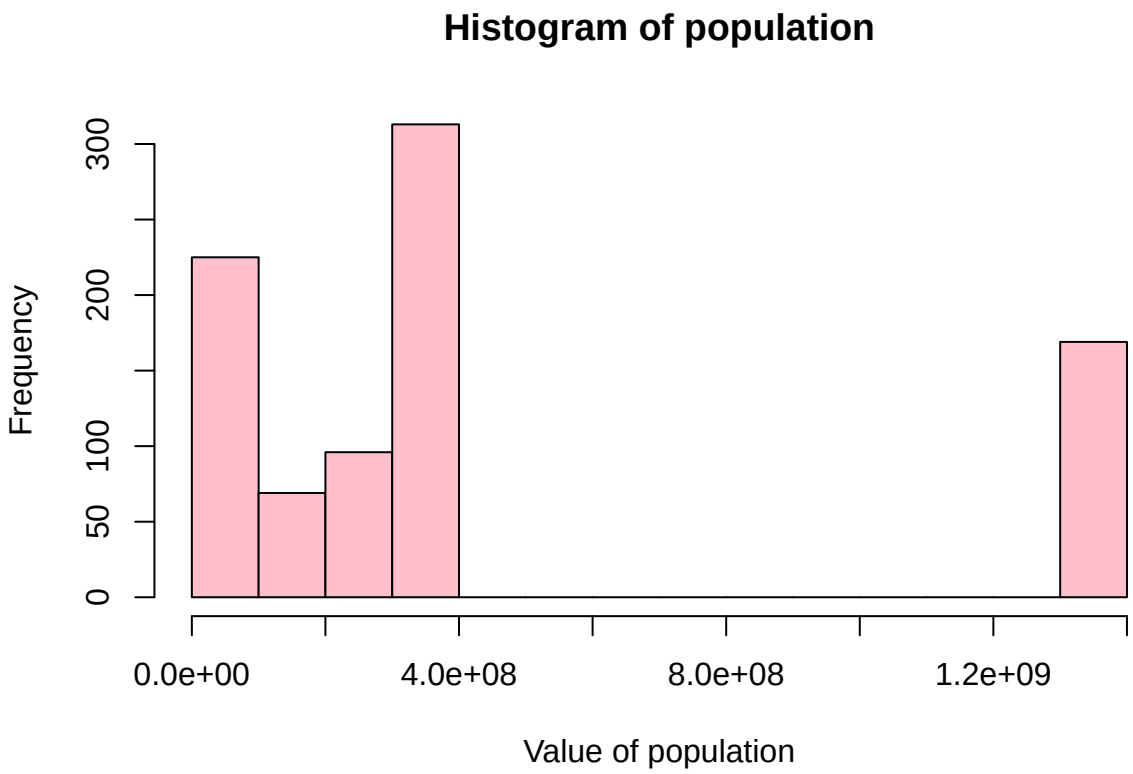
```
norm_model<-preProcess(Subscribers_df, method = c('range'))
Default_normalized<-predict(norm_model,Subscribers_df)

#After normalization, Min is 0 and Max is 1
summary(Default_normalized)
```

```
##  YouTube$subscribers
##  Min.   :0.000000
##  1st Qu.:0.009454
##  Median :0.023206
##  Mean   :0.045906
##  3rd Qu.:0.052858
##  Max.   :1.000000
```

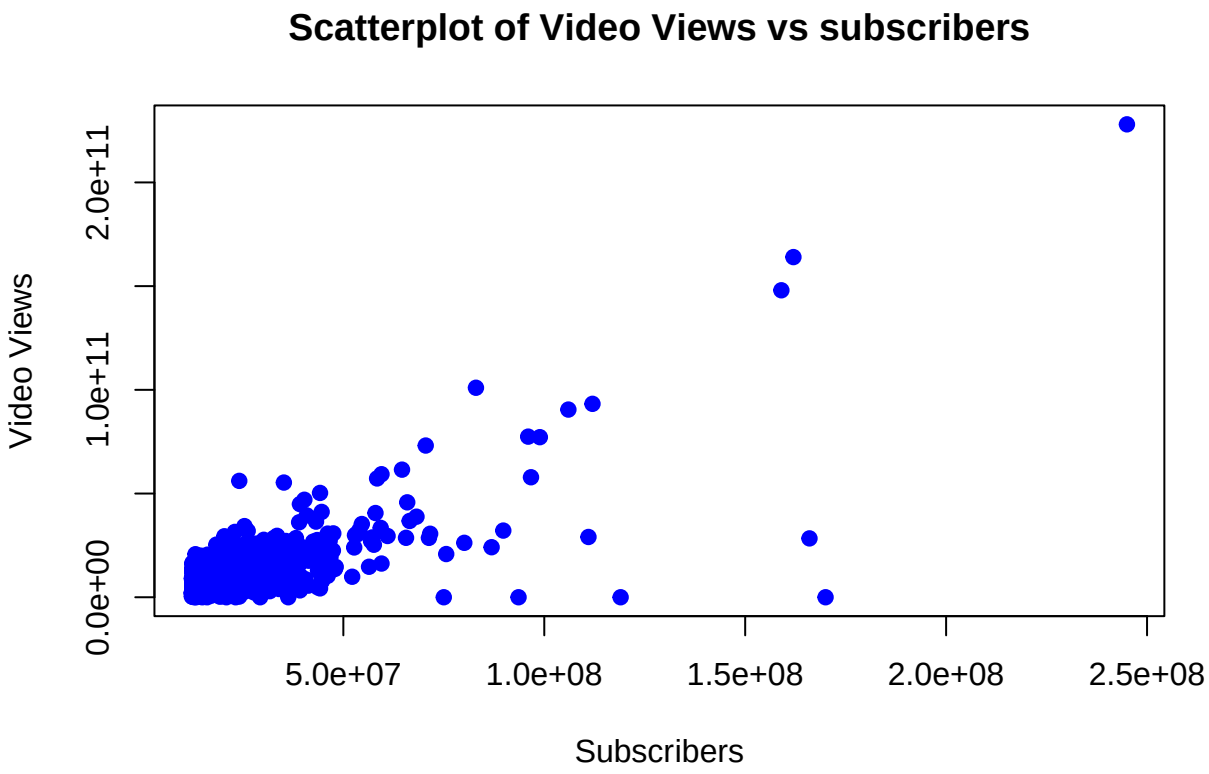
Plotting a Histogram of the Population column in YouTube

```
Hist_1 <- hist(YouTube$Population, main="Histogram of population", xlab="Value of population",
               ylab="Frequency", col="pink", border="black")
```



Plotting a Scatter plot of video_views and Subscribers in YouTube

```
plot(YouTube$subscribers,YouTube$video.views, main="Scatterplot of Video Views vs subscribers", xlab="Su
     ylab="Video Views", pch=19, col="blue")
```



Finding the correlation coefficient to understand the relationship

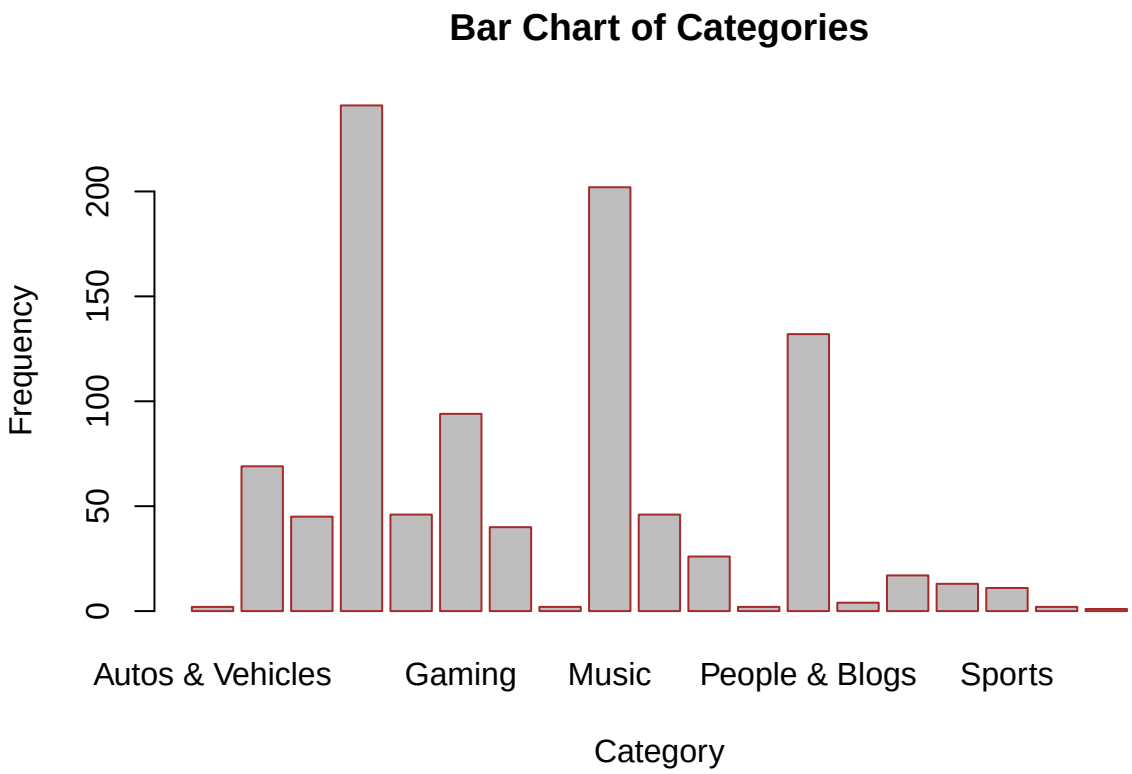
```
Correl_1 <- cor(YouTube$subscribers,YouTube$video.views)
Correl_1
```

```
## [1] 0.7509576
```

From the Correlation Coefficient we infer that, the number of subscribers and video views are positively correlated.

Bar Charts can be used to understand the Frequency of specific categories (Categorical Variables)

```
freq_table <- table(YouTube$category)
bp <- barplot(freq_table, main = "Bar Chart of Categories",
              xlab = "Category",
              ylab = "Frequency", col = "grey",
              border = "Brown")
```


Bar Chart of Categories
Frequency
0
50
100
150
200
Autos & Vehicles
Gaming
Music
People & Blogs
Sports
Category