**DATA ANALYTICS IN BUSINESS USING R – MIDTERM REPORT**

**DIGITAL MEDIA INTELLIGENCE**

**1. Dataset Description:**

The dataset includes several months of data on daily trending YouTube videos. The data is captured in the USA, Great Britain, Germany, Canada, France, Russia, South Korea, Mexico, and India. Each country’s data is recorded in a separate CSV file. The Column data includes:

|  |  |
| --- | --- |
| Columns | data type |
| 1. video\_id | String |
| 1. trending\_date | Date |
| 1. title | String |
| 1. channel\_title | String |
| 1. category\_id | Integer |
| 1. publish\_time | Date |
| 1. tags | String |
| 1. views | Integer |
| 1. likes | Integer |
| 1. dislikes | Integer |
| 1. comment\_count | Integer |
| 1. thumbnail\_link | URL |
| 1. comments\_disabled | Boolean |
| 1. ratings\_disabled | Boolean |
| 1. video\_error\_or\_removed | Boolean |
| 1. description | String |

**2. Purpose of the Project:**

Vloggers on YouTube are provided with visual analytics on the content they upload by default, but they do not get the overall visualization of their competitor. Analyzing trending videos may provide publishers with the ability to predict the trend. Thus, they can develop their future content based on the analysis. Our analysis provides market forecast and intelligence which reveals information on the viewership, likes, dislikes. This prediction is based on many parameters and knowledge that our models has acquired by analyzing most trended videos that has been uploaded on YouTube.  So, this can be extremely useful and profitable for YouTube channels who depend on their channels as a source of income.

**3. Intended Audience**

We assume the audience of this report to know the basics of:

1. R programming language
2. Algebra
3. Statistics.

**4. Importing and Tidying the Data:**

The data is being imported using the **read.csv()** function. Although this is slower compared to the read\_csv(), it is important that the integrity and the proper structure of data be maintained. A single function has been constructed which handles the data import and tidying. It can be seen below:

library(tidyverse)

library(dplyr)

# Function to tidy the data sets

tidy\_dataset <- function(file\_name)

{

#---- Reading the data

mydata <- read.csv(file\_name, stringsAsFactors = FALSE)

#---- Tidying the data

#Changing the format of Trending Date to a readable format

mydata$trending\_date <- as.Date(mydata$trending\_date, format = "%y.%d.%m")

#Editing the date to have only the month

mydata$trending\_date <- format(as.Date(mydata$trending\_date), "%m")

#Changing the date to a factor type

mydata$trending\_date <- as.factor(mydata$trending\_date)

#Changing the format of Trending Date to a readable format

mydata$publish\_time <- substr(mydata$publish\_time, 0, 10)

mydata$publish\_time <- as.Date(mydata$publish\_time, format = "%Y-%m-%d")

#Editing the date to have only the month

mydata$publish\_time <- format(as.Date(mydata$publish\_time), "%m")

#Changing the date to a factor type

mydata$publish\_time <- as.factor(mydata$publish\_time)

#Changing column names of trending\_date and publish\_time

names(mydata)[2] <- "trending\_month"

names(mydata)[6] <- "publish\_month"

#Getting rid of unnecessary variables

d\_subset <- mydata[,c(2,3,4,5,6,7,8,9,10,11,13,14,15,16)]

#If the video has no tags or if the video has no description, omitting it

new\_dataset <- d\_subset[d\_subset$tags != '[none]', ]

new\_dataset <- new\_dataset[new\_dataset$description != '', ]

#---- Creating New Dataset

newfile = paste('new',file\_name,sep='')

write.csv(new\_dataset, newfile, row.names = FALSE)

}

The code above shows the function used to import and clean data. After the data is imported, the data is tidied. The process followed for tidying the data is as follows:

1. The date is processed so that it is in a readable format. And, only the required component of the date which is the month is selected.
2. The same process is repeated for both, trending\_date and publish\_time.
3. These variables are changed to a factor type to help in further analysis.
4. Two new variables are created which are basically containing the values of publish\_time and trending\_date. But, now, they are renamed to publish\_month and trending\_month respectively.
5. Then, the videos unnecessary to our project, i.e., the ones with no tags or no description are being omitted for the sake of convenience. This will not have a big impact on the analysis because most of them are outliers and losing them does not affect the analysis greatly.

Then, a new csv file is generated by using the **write.csv()** function. The new file is renamed as the same name as the input file, but with the a ‘new’ appended at the beginning of the filename. So, a new csv file containing the tidied data is generated once the function is called. The input to a function is a file. Eg: tidy\_dataset(CAvideos.csv). So, once this function is called, the CAvideos.csv is tidied and another file called newCAvideos.csv is generated. This holds the tidied data. After the data is tidied, the structure of the data frame is as follows:

|  |  |
| --- | --- |
| Columns | data type |
| 1. trending\_month | Factor |
| 1. title | String |
| 1. channel\_title | String |
| 1. category\_id | Integer |
| 1. publish\_month | Factor |
| 1. tags | String |
| 1. views | Integer |
| 1. likes | Integer |
| 1. dislikes | Integer |
| 1. comment\_count | Integer |
| 1. comments\_disabled | String |
| 1. ratings\_disabled | String |
| 1. video\_error\_or\_removed | String |
| 1. description | String |

**Exploratory Data Analysis**

Exploratory Data Analysis is the critical process of performing initial investigations on the obtained data in order to build hypotheses, expose anomalies present and check assumptions.

As a result of Exploratory Data Analysis, we will know the following:

1. Correlation map for the variables present in the dataset
2. Initial graphical visualizations

```{r}

# Exploratory Data Analysis

# Reading contents from the tidied dataset

tidy\_data\_csv <- read.csv('newUSvideos.csv', stringsAsFactors = F)

tidy\_data\_csv[,c("trending\_month","publish\_month")] <- lapply(

tidy\_data\_csv[,

c(

"trending\_month",

"publish\_month"

)

],as.factor)

# Separating integers type variable from the dataset from the rest of the table

tidy\_data\_integer <- data.frame(tidy\_data\_csv$views,

tidy\_data\_csv$likes,

tidy\_data\_csv$dislikes,

tidy\_data\_csv$comment\_count)

# Renaming the columns created in the previous dataframe.

names(tidy\_data\_integer) <- c("views",

"likes",

"dislikes",

"comment\_count")

```

**Correlation Definition**

Correlation is the bivariate (two variable) analysis that measures the strength of association between two variables. It is also used to determine the direction of the association. The strength of the association is expressed by the correlation coefficient, whose value varies between -1 and +1. A value of +1 indicates a strong positive correlation, a value of -1 indicates a strong negative correlation and a value of 0 indicates there is no correlation. Positive correlation meaning the value of the response variable tends to change in the same direction for every change in the explanatory variable. Negative correlation meaning value of the response variable tends to change in the opposite direction for every change in the explanatory variable. The closer the coefficient value is to zero the weaker the strength of correlation.

**Reasons for performing correlation analysis**

We perform Correlation analysis in order to determine the strength of association between variables in the dataset. By doing so we will be able to know the potential attributes about each variable.

Correlation is strictly used to test association between variables, it does not make any assumptions whether one variable dependent on the other.

**Different types of**

1. Pearson Product Moment Correlation
2. Spearman Rank Correlation
3. Kendall Rank Correlation.

**Pearson Correlation:**

It is a technique used to measure two quantitative and continuous .

**Spearman Rank Correlation:**

It is a technique used when we need to measure correlation between two ranked variables or when we want to measure the correlation between a quantitative variable and a ranked .

**Kendall Rank Correlation**:

It is a technique used when we need to measure the correlation between pairs of bivariate points, the co-ordinates are measured individually to declare each point in the graph as concordant or discordant with respect to the other points on the graph.

Example:

A screenshot of a cell phone

Description automatically generated

Fig 1.0 Description of Kendall Rank

```{r}

# Plotting Pearson Correlation graphical map

correlation\_data\_pearson <- tidy\_data\_integer %>%

cor() %>%

corrplot(

method = "color",

type = "upper", order = "hclust", number.cex = .7,

addCoef.col = "black", # Add coefficient of correlation

tl.col = "black", tl.srt = 90, # Text label color and rotation

# hide correlation coefficient on the principal diagonal

diag = FALSE

)

```

From the Pearson Correlation map, it is clear that,

1. Views and likes
2. Comment count and likes

These are strongly correlated with a coefficient value of 0.85 and 0.85 respectively. Following which,

1. Comment count and dislikes
2. Comment count and views

These pairs moderately correlated with a coefficient value of 0.62 and 0.66 respectively.

While we find the remaining pairs are also positively correlated, we ignore the correlation because the value of the remaining pairs relative to the above four pairs are weakly correlated.

```{r}

#Creating a data frame for calculating Spearman Correlation Map

trended\_month <- as.integer(tidy\_data\_csv$trending\_month)

published\_month <- as.integer(tidy\_data\_csv$publish\_month)

ordinal\_data\_frame <- cbind(trended\_month,published\_month)

#Plotting Spearman Correlation graphical map

correlation\_data\_spearman <- ordinal\_data\_frame %>%

cor(

method = "spearman",

use = "pairwise.complete.obs") %>%

corrplot(

method = "color",

type = "upper", order = "hclust", number.cex = .7,

# Add coefficient of correlation

addCoef.col = "black",

# Text label color and rotation

tl.col = "black", tl.srt = 90,

# hide correlation coefficient on the principal diagonal

diag = FALSE

)

```

From the Spearman Correlation map, it is clear that,

1. Trending month and Published month

are highly correlated, with a Coefficient value of 0.88.

```{r}

# Density Plot

(

# Density Plot for Views

density\_plot\_views <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(views)

)

) +

geom\_density(bw = 1)

)

(

# Density Plot for likes

density\_plot\_likes <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(likes)

)

) +

geom\_density(

bw = 1

)

)

(

# Density Plot for dislikes

density\_plot\_dislikes <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(dislikes)

)

) +

geom\_density(

bw = 1

)

)

(

# Density Plot for comment\_count

density\_plot\_comment <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(comment\_count)

)

) +

geom\_density(

bw = 1

)

)

```

From the Density plots we conclude that all the continuous variables follow the normal distribution.

This brings to the next step for verifying correlation is valid or spurious. We term a correlation is Spurious when the correlation matrix shows a strong correlation coefficient, but the scatterplot does not show any possible linear curve fit.

We perform this step, in order to eliminate spurious correlations. In this step we draw scatter plots to the strong and moderately correlated variable pairs.

```{r}

# Scatter Plot

# Scatter Plot for likes vs. views

(

scatter\_likes\_views <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(views),

y = log(likes)

)

) +

geom\_point()

)

# Scatter Plot for comment\_count vs. likes

(

scatter\_likes\_comment <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(likes),

y = log(comment\_count)

)

) +

geom\_point()

)

# Scatter Plot for comment\_count vs. dislikes

(

scatter\_dis\_com <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(dislikes),

y = log(comment\_count)

)

) +

geom\_point()

)

# Scatter Plot for views vs. comment\_count

(

scatter\_dis\_com <- tidy\_data\_csv %>%

ggplot(

aes(

x = log(views),

y = log(comment\_count)

)

) +

geom\_point()

)

# Scatter Plot for publish\_month vs trending\_month

(

scatter\_dis\_com <- tidy\_data\_csv %>%

ggplot(

aes(

x = publish\_month,

y = trending\_month

)

) +

geom\_point()

)

```

From the Scatter plots we conclude that all the selected variable pairs are not spurious.

Now that we have proven that the correlation is not spurious. We go ahead to assume that the relationship is valid.

Hypotheses:

Hypothesis 1:

Predict the number of likes for a trending video in the dataset

Hypothesis 2:

Predict the number of dislikes for a trending video in the dataset

Hypothesis 3:

Predict the number of comment count for a trending video in the dataset.

Reason for choosing the Hypotheses:

As a result of Exploratory Data Analysis, we noticed that

1. Views and likes
2. Comment count and likes
3. Comment count and dislikes
4. Comment count and views

Are strongly interdependent, however we make **no** **assumption** about the causal effect. The strong interdependence gives us an insight to further investigate to check the causality of this interdependence.

Reason and Selection of Model:

From the Pearson correlation map, we can conclude that there is strong interdependence on many continuous variables of the dataset. This tells us that there is a possibility of using simple linear regression because simple linear regression is useful for finding the relationship between two continuous variables in a dataset.

From the Spearman correlation map. It is evident that Trending date and Publish Date are strongly interdependent, however further investigation is required to help us predict this strong association.

Hypotheses Test

We follow K-Fold cross validation technique, as it is best suited for small .

Challenges

1. Due to the COVID-19, and one of our team members stuck in Boston we will not be able to meet. We need to strategize our communication to adapt to this situation.
2. We need to come up with a model for predicting the causal effect for the strong correlation between the trending month and publish month. We need to still investigate further in order to come up with the suitable hypothesis for the same.

References

1. Kerley, Christa. "What Are the Different Types of Correlations?" *sciencing.com*, https://sciencing.com/different-types-correlations-6979655.html. March 25 2020.
2. University of the West of England, Pearson's Correlation Coefficient.

Retrieved March 25, 2020, from http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442.

1. McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.
2. Clapham, Matthew. E. Jan 30, 2016, 19: Non-parametric correlation, Retrieved March 25, 2020, from https://www.youtube.com/watch?v=bAstMHbytK0
3. SAS (2017). JMP 13 Fitting Linear Models, Second Edition.