Project Report for Data Exploration and Analysis of MOOC Dataset

Sanchit Bhatnagar

03/12/2020

R Markdown

BUSINESS OBJECTIVES

We will be analysing a fast growing business application area called Learning Analytics. It is an application for Data Science, defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which it occurs".

In the existing scenario of recording student presence through their attendance, student's engagement outside of the classroom as in their home, libraries,Virtual viewing etc is not reflected. The objective of the Learning Analytics organization is to analyse more student Data based on various sources through which they access course work.

Several different Sources of data are periodically collected about the organizations learners like their use of on-campus facilities, Virtual Learning Environment (VLE) and ReCap access, and student wellbeing referrals.

The main objective of our organization 'Learning Analytics' is to aggregate all the important sources of data to derive insights, provide the impactful measures of engagement.

From analysing the above, there is scope of getting extremely useful information like: -how to better the design of learning -inform intervention processes for at-risk students -improving student attainment.

The Success Criteria is the efficient access of data from all the sources without any loss, discrepancy, damage and quality of the student's data. Also, gathering the maximum data possible in all situations.

For the entire task, we need to efficiently explore, analyse, assess, interpret, judge and find insights from the datasets that have been provided to us.

RESOURCES AVAILABLE

We have been supplied with Data for the Cyber security course from future learn course. Given Datasets of 7 batches of students, which enrolled in a cyber security Future Learn course available online. The data covers various types of information such as: -The enrollments of students with student specific ID and the dates of enrolling and leaving -The types of students (by background), their surveys -The various countries from which students attended the course -The Video statistics based on the number of views of the available videos, number of downloads, region wise viewing, percentage of completion etc. -Dates of completing different steps from the online course.

There is another set of data that displays the user interface of the course. It is same for all 7 cohorts and shows us the all the steps that cover the whole course with a step number and one-line description, that enrolled students require to complete.

RESOURCES REQUIRED

For our end of the analysis, we require following available resources: 1. A computer/laptop with R installed 2. Software Tools: Rstudio with all the required packages and libraries. 3. An official GIT hub account for acting as our project's repository. 4. The Datasets in appropriate format (eg. csv files). 5. The Interface summarizing the videos

ASSUMPTIONS

A set of assumptions based on our Application area and its Datasets as summarized below:

1. The visual display of the interface for the cyber security course is totally same as the actual one and it shows all of the steps that are required for course completion accurately.
2. The Steps of the course are consistent throughout for all the 7 cohorts
3. Some of the data sets are completely blank, we will assume that they fall under incomplete information provided, hence, can be discarded.
4. We are assuming that all of the information is accurate, with minumum discrepancy and gathered from reliable resources.

RISKS AND CONTINGENCIES

Due to safety purposes, we pre-define possible risks :

1. The data provided to us is from 2016 to 2018, there is a possibility of it not being reliable based on the analysis we will carry out on it that will be used for future improvement of the course.

2.Technological risks such as file corruption, data deletion, power outage and entire computer corruption

3.Data quality can be extremely poor and render unreliable, the name of the files determine the sequence number of the cohort but the data within the files could be mismatched.

DATA MINING OBJECTIVE

The data we have been given is in its raw form. We are supposed to analyse and interpret it to find useful information from it. Our main objectives of data mining from the given sets are:

*Determine the number of enrollments in each cohorts and compare them with each other to get the basic idea of the average enrollments in a particular batch.

*To find out the number of enrollments across each cohort for every country, this will help us determine the popular places where our course is in demand and we would need to allocate more resources and bandwidth for those places, as well as, focusing more on advertisements and publicizing in the places with higher potential of clients.

*The files containing video statistics help us to find the correlation between the types of videos that each cohort watches the most, with their length of duration, total number of views and their downloads, can determine what if there is a relation among those variables, that can be used for analysis further to classify types of videos, make improvements, alter the length of duration, improve downloading methods.

*The Video statistics file consists of the information of the percentage of views across various devices, namely: console, desktop, mobile, television, tablet, unknown devices. We will use this data to determine the most widely used device, the least used device and if any devices faced issues, hence, planning our future courses to have more compatibility in certain devices and access to more devices if required.

PROJECT PLAN

An initial project plan is laid out as follows:

-We will use Data analysis and documentation tools in our systems to first select the required data sheets, successfully load them and then perform exploratory analysis on them using R. It will be feasibly carried out using different R packages for reproducibility and also by using a code repository on several different systems.

*We need to determine the number of enrollments across all of the 7 cohorts and get a basic estimate of the number,the number of enrollments across every country whose citizens have enrolled in it, draw inferences from the video statistics on what video is most played, downloaded etc and the various devices used for playing the videos. The above tasks will be performed using data analysis in R using several Packages for this purpose like plyr,countrycode,forcats,testthat,ggpubr, tidyverse,ggplot2 etc.

*Our 40 percent of the time and effort is expected for Data Preparation Phase and 30 percent for the Data Understanding Phase,25 percent in each of the Modeling, Evaluation, and Business Understanding Phases and 5 percent in the the Deployment Phase.

*Our decision will be based on the graphs and plots that we generate for our tasks and the mean and correlations between various video sets will provide us with analytical information to build decisions.

*We need to constantly update our code repository (we are using GIT hub for this) after updating any files or codes in our project.

TECHNOLOGIES USED

All our tools used are free open-source, making them the best choice.

We will work on the following technologies based on our requirements:

All our tools used are free open-source, making them the best choice. #### 1)Microsoft Excel It is best for when the first time we receive the data, we can get the sense of how much of data is provided to us and what is it basically about. The files we have been provided are in .csv format, for which Excel does a really great job in separating into columns and rows.

2)R

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. For our proper analysis of the data, a tool for stats and graphics will be best suited for us, also its convenient syntax, fast computations, several in-built functions will be supportive throughout our analysis.

3)Rstudio

After having R in our system, we can use a good IDE (Integrated development environment) where R can be used even more efficiently and conveniently. Rstudio provides a wide number of benefits like making it easy to write scripts,set a directory and use files on the system, access objects more easily and graphics are easily accessible for a non technical person.

4)Project Template

We are following a CRISP-DM approach for this project, the best way will be to use the 'projectTemplate' template that can be integrated with Rstudio. The project standardization and code reproducibility is a necessity for our project, we need it to run across several systems.

5)GIT Repository

Git is a distributed version-control system for tracking changes in any set of files. Its goals include speed, data integrity, and support for distributed, non-linear workflows. Our Project Template has a version control feature that lets us create an entire project associated to our account in GIT. The code repository can be updated through our RStudio for any changes we make in our code and files in our Project. It is important for us to use as we want a reproducible project and the repository is a safe zone for our code in case of any system risks.

6)R Packages

We will be using several R packages/Libraries namely: reshape2, plyr,countrycode,forcats,testthat,ggpubr, tidyverse, stringr, lubridate, tidyverse, dplyr, ggplot2. We need to filter our datasets, sanity test, create graphics/plots/charts, determine countries from country codes. The named packages above will provide us with all the required functionality.

7)Rmarkdown

Our Project Report is made on another feature in Rstudio i.e. Rmarkdown. It creates an interactive project report and we can directly get the output, code etc from our scripts and add use them as input in our Rmarkdown file. For our requirement of creating a reproducible project, this is one of the best approach as we can run the project code several times and the Rmarkdown report file gets updated. For example, changing the type of plot in our script will be reflected in our new Rmarkdown file.

8)Microsoft PowerPoint

To make a final presentation for our client to summarize what we have majorly done and how it can be reproduced efficiently, we will use Microsoft Powerpoint for creating slides as a means of explanation

9) Adobe Acrobat

We have 7 datasets that show the interface of the online learning and all the steps of the course are summarized in it. It is in PDF format, hence, best way to open them is using Acrobat.

DATA UNDERSTANDING

Collect initial data There is a total of 61 Files provided to us. We examined each file by manually opening them and doing our first bit of analysis. The data is for students who had enrolled in the cybersecurity program from the date 2016-08-10 till 2018-11-01. There had been 7 cohorts who enrolled for the same Course. Hence, we have datasets of 7 cohorts with different information. There are around 8 files for each cohort, named like: Enrolments archetype- survey responses weekly-sentiment-survey-responses team-members step-activity question-response video.stats leaving-survey-responses names for each cohort

For our analysis, we have identified the Enrolments files and video.stats files to be used. As defined earlier, our goal for counting number of enrollments across 7 cohorts and identifying the countries, we will need

the Enrolments file for each cohort. For our task of assessing videos viewership, we will use 1 of the video stats file i.e. for the 3rd cohort We manually observed that the Video statistics for the first 2 cohorts are unavailable, and as for the remaining 5, the data is nearly the same for all cohorts. Hence, on a random basis, we will stick to using the video.stats file for the 3rd Cohort.

Using copy-paste, we load our 7 enrolment files and 1 cyber.security.3_video-stats by saving them in the 'Data' Folder of our Project Template.

Checking the columns of one of the enrolment files as below:

```
colnames(cyber.security.1_enrolments)
```

```
##  [1] "learner_id"              "enrolled_at"
##  [3] "unenrolled_at"           "role"
##  [5] "fully_participated_at"   "purchased_statement_at"
##  [7] "gender"                  "country"
##  [9] "age_range"               "highest_education_level"
## [11] "employment_status"       "employment_area"
## [13] "detected_country"
```

There are 13 columns but our main focus would be on the countries. We will print some rows for detected_country column just to check how the values look like.

```
head(cyber.security.1_enrolments$country)
```

```
## [1] "Unknown" "PE"      "Unknown" "Unknown" "Unknown" "Unknown"
```

We notice that they are in a country code format, to later modify them as part of data preparation. Most of the remaining columns like gender, age range, employment_status have most values as unknown, hence we drop them from our analysis. The enrollment dates and un-enrollment dates are present, but our analysis is not focusing on it as the the duration of the course is fixed for all cohorts.

Checking the columns of video.stats file

```
colnames(cyber.security.3_video.stats)
```

```
##  [1] "step_position"                  "title"
##  [3] "video_duration"                 "total_views"
##  [5] "total_downloads"                "total_caption_views"
##  [7] "total_transcript_views"         "viewed_hd"
##  [9] "viewed_five_percent"            "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"      "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"     "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"      "console_device_percentage"
## [17] "desktop_device_percentage"      "mobile_device_percentage"
## [19] "tv_device_percentage"           "tablet_device_percentage"
## [21] "unknown_device_percentage"      "europe_views_percentage"
## [23] "oceania_views_percentage"       "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"        "antarctica_views_percentage"
```

There are 28 columns detailing the information of different types of videos.

Our interest is in the relation of the types of video and their lengths compared to their number of views, downloads, transcript views, caption views. The other set we will be using is of the columns that give the percentage of viewership across different sets of devices(television, laptop etc). We drop the remaining columns deemed unnecessary for our analysis.

EXPLORE DATA For doing some basic exploration for our data, we calculate the number of rows in all our enrollment files.

```
c=c(nrow(cyber.security.1_enrolments),nrow(cyber.security.2_enrolments),nrow(cyber.security.3_enrolments
c
```

```
## [1] 14394  6488  3361  3992  3544  3175  2342
```

The first cohort has a large number of enrollments, whereas the second has around 6000 enrollments, but the remaining 5 have even half of that. The enrollments are not consistent for all the 7 cohorts.

Calculating the Rows for the video stats file

```
nrow(cyber.security.3_video.stats)
```

```
## [1] 13
```

There are 13 rows for each column. This tells us that there are 13 different videos at different steps throughout the complete course for our data.

###DATA QUALITY For the Data files selected, we want to check the quality of the data and verify that it is appropriate for the type of analysis to perform. Checking the type of the data in each columns for any of the enrolments file:

```
##  [1] "tbl_df"     "tbl"        "data.frame" "tbl_df"     "tbl"
##  [6] "data.frame" "tbl_df"     "tbl"        "data.frame" "tbl_df"
## [11] "tbl"        "data.frame" "tbl_df"     "tbl"        "data.frame"
## [16] "tbl_df"     "tbl"        "data.frame" "tbl_df"     "tbl"
## [21] "data.frame" "tbl_df"     "tbl"        "data.frame" "tbl_df"
## [26] "tbl"        "data.frame" "tbl_df"     "tbl"        "data.frame"
## [31] "tbl_df"     "tbl"        "data.frame" "tbl_df"     "tbl"
## [36] "data.frame" "tbl_df"     "tbl"        "data.frame"
```

From our output, we observe that we have dataframes, tbl format. They can be used to convert into apprpriate tables and used conveniently for our analysis. As we know, the cell values would be in various formats, like characters, dates, integers etc.

Checking the type of data for cell values in video.sets for the 1st row of data

```
##  [1] "numeric"   "character" "integer"   "integer"   "integer"   "integer"
##  [7] "integer"   "integer"   "numeric"   "numeric"   "numeric"   "numeric"
## [13] "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
## [19] "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
## [25] "numeric"   "numeric"   "numeric"   "numeric"
```

From the above, we conclude that the column values are in appropriate format,i.e integer or numeric, to carry out our analysis.Also, the video name is in character format, as expected.

We check for any null values in the video sets, and in the detected_country column for any one of the enrolment set.

```
## [1] 0
```

```
## [1] "There is no NULL in video sets"
```

```
## [1] 11
```

```
## [1] "There are 11 NULL in enrolment data"
```

We remove the NULL in the countries in data preparation.

The data quality for our datasets is good for our analysis.

DATA PREPARATION

We have decided the Data files to use out of all the data provided to us. From the selected files, we are going to subset data, clean data, make some assumptions and transform the data. Further, modified the data so as to perform our analysis efficiently provided the tools we will use, and to get the most from our analysis.

- PREPARATION FOR ENROLMENT SET (7 files):
- Selecting Data - To analyse the number of enrollments across all the 7 Cohorts. We require the distinct learner IDs in each of the 7 data sheets. Present in the first column, we extract all the learner ids from the 7 cohorts' data.
- Constructing Data - Replace the value of all distinct learner ids with the value 'course set i' , where i will be the respective cohort of the learnerId.
- Integrating Data - Merge the total number of all the Learner IDs(whose original value has been replaced with 'course set i) into one data frame. Our set will contain the sum of all rows from all 7 datasets with values like course set 1, course set 2 etc. This is done to be used to create a bar chart.
- PREPARATION FOR ENROLMENT SET for detected_country column (7 files):
- Selecting Data - To analyse the number of countries across all the 7 Cohorts. We select the detected_country column from all of our 7 Data sets.
- Clean Data - As observed in one of our dataset, NULL values in some cells are present, we remove all those cells.
- Formatting Data - The detected_country column has values in the country code format (eg. UK for United Kingdom). We use the coutrycode package and convert the codes into the full country name. Note: Observed error : 'In countrycode(country_new, "ecb", "country.name") Some values were not matched unambiguously: PS'. The data is formatted and this country code is not found, hence, ignored.
- Filtering Data - We observed a large number of countries and each of those have some viewership. Hence, to focus on the countries with greater viewership, we extract those with number of enrollments greater than 20.
- PREPARATION FOR Video.set data set for number of views analysis(1 file):
- Selecting Data - To analyse the different videos and their number of views, downloads etc, we extracted the 3rd to 7th column and the 1st column for indexing

```
## [1] "step_position"          "video_duration"         "total_views"
## [4] "total_downloads"        "total_caption_views"    "total_transcript_views"
```

- PREPARATION FOR Video.set dataset for device views analysis(1 file):
  - Selecting Data - To analyse the the percentage views across different devices, we extracted the 16th to 21st row.

MODELING

Modeling Techniques The different modeling techniques for our 4 types of analysis are as follows: 1) Distribution of Enrollments count across Cohorts: Using modified data from the 7 enrolment dataset, generate a barchart to visualize the count.

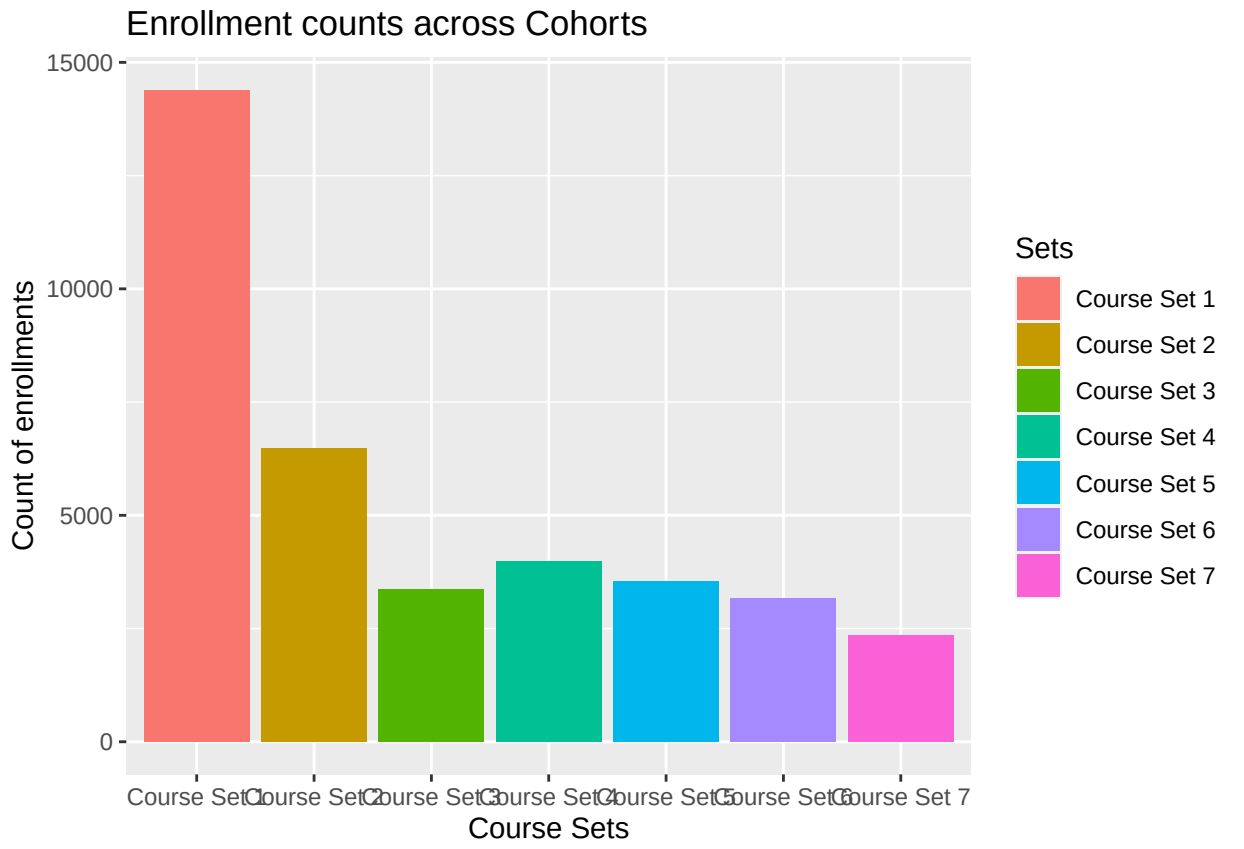
2) Distribution of enrollments based on countries: Using modified data from the 7 enrolment dataset, generate a barchart to visualize the maximum viewerships in different countries.

3) Count of Videos viewed, downloaded, viewed transcripts based on the step videos: Using modified data from video stats 2 data sheet, Generate pair wise plot to check correlation between several variables and a scatter plot for total views vs total downloads.

4) Percentage of videos viewed across several devices: Using column means to detect the different percentages

MODEL1

BAR chart for our analysis



The model above is a simple bar chart for our 7 cohorts with heights as the count of the enrollment per cohort and the X axis is the Cohort Number. Different colors are used to create distinction.

This shows us the number of enrollments in the first cohort were extremely massive (almost 15000) as compared to the other cohorts. The second had around 6000, whereas the other 5 cohorts had a similar enrollment of about 3000 students.

This model satisfies our data mining task to compare the various strengths in the cohorts.

MODEL2

Verified the number of distinct countries in each cohort that have only those countries that are enrollments of more than 20 learners.

```
##    n
## 1 67
```

```
##    n
## 1 44
```

```
##    n
## 1 27
```

```
##    n
## 1 34
```

```
##    n
## 1 26
```

```
##    n
## 1 21
```

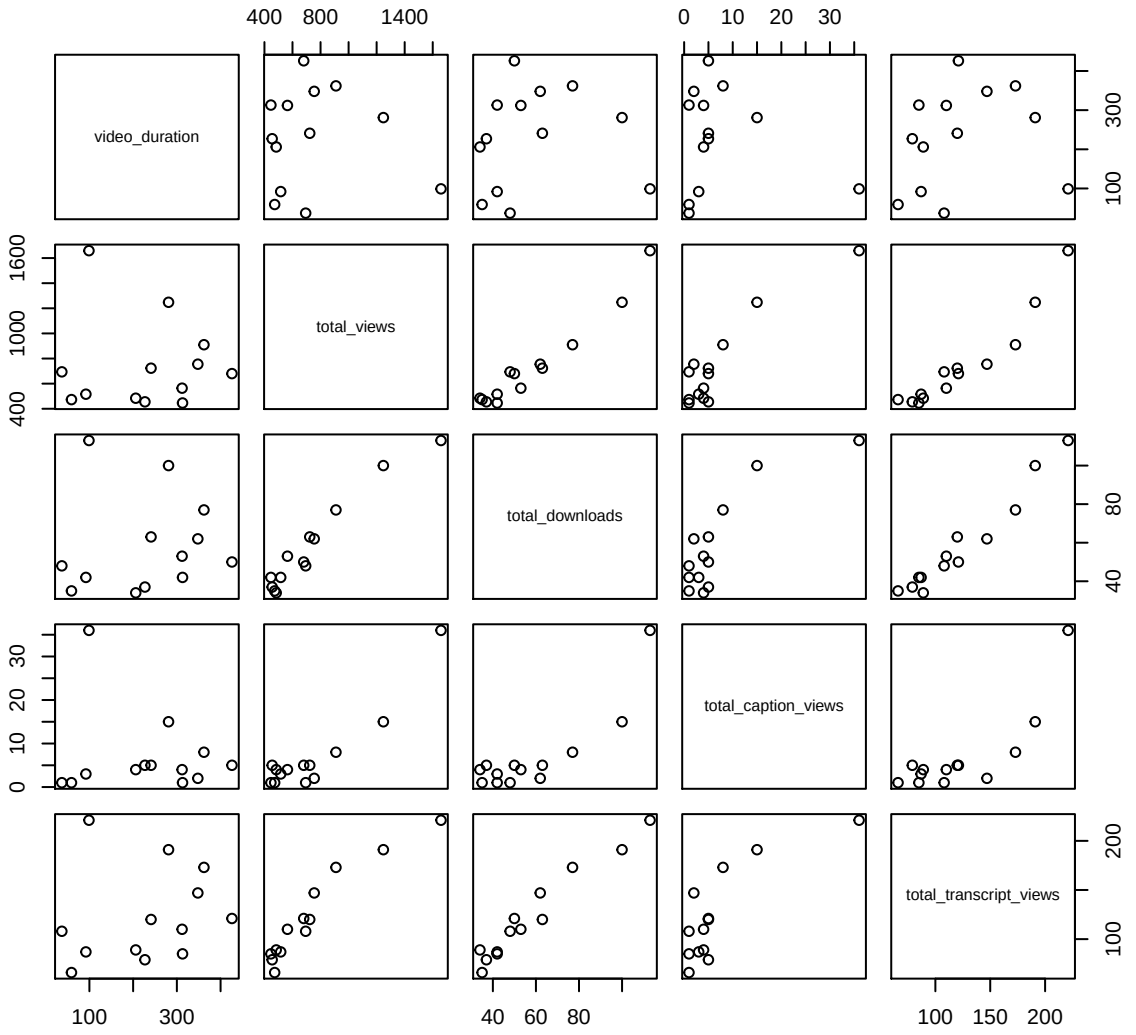
```
##    n
## 1 17
```

The number of countries for each is similar, we will visualize using the list of countries from the first 3 cohorts. ### Inverted BAR chart

This model displays us the range of various number of enrollment based on the country. Using colors, the distinction is quite easily visible even for the countries with similar enrollment number. It is an essential model based on demographics that can be analytically used for making decisions.

MODEL3

Pair wise plot of all the variables for the 13 video steps



A pairs plot is generated, it shows the relation between our variables video duration, total views, total downloads, total caption views and total transcript views. As we expect, there is a linear relation between many variables such as total downloads and total views, total transcript views and total downloads.

Analysing the same set of data. We find the Means of all the variables for each video, but without the video lengths.

```
##           total_views        total_downloads    total_caption_views
##            739.000000              58.153846               6.923077
## total_transcript_views
##            122.846154
```

Generating a correlation matrix:

```
##                         total_views total_downloads total_caption_views
## total_views               1.0000000       0.9720183           0.9191463
## total_downloads           0.9720183       1.0000000           0.8530263
## total_caption_views       0.9191463       0.8530263           1.0000000
## total_transcript_views    0.9519238       0.9735810           0.8084851
##                         total_transcript_views
## total_views                          0.9519238
## total_downloads                      0.9735810
## total_caption_views                  0.8084851
## total_transcript_views               1.0000000
```

The above models give us enough information on the variables and their relationship to each other and the different types of videos.

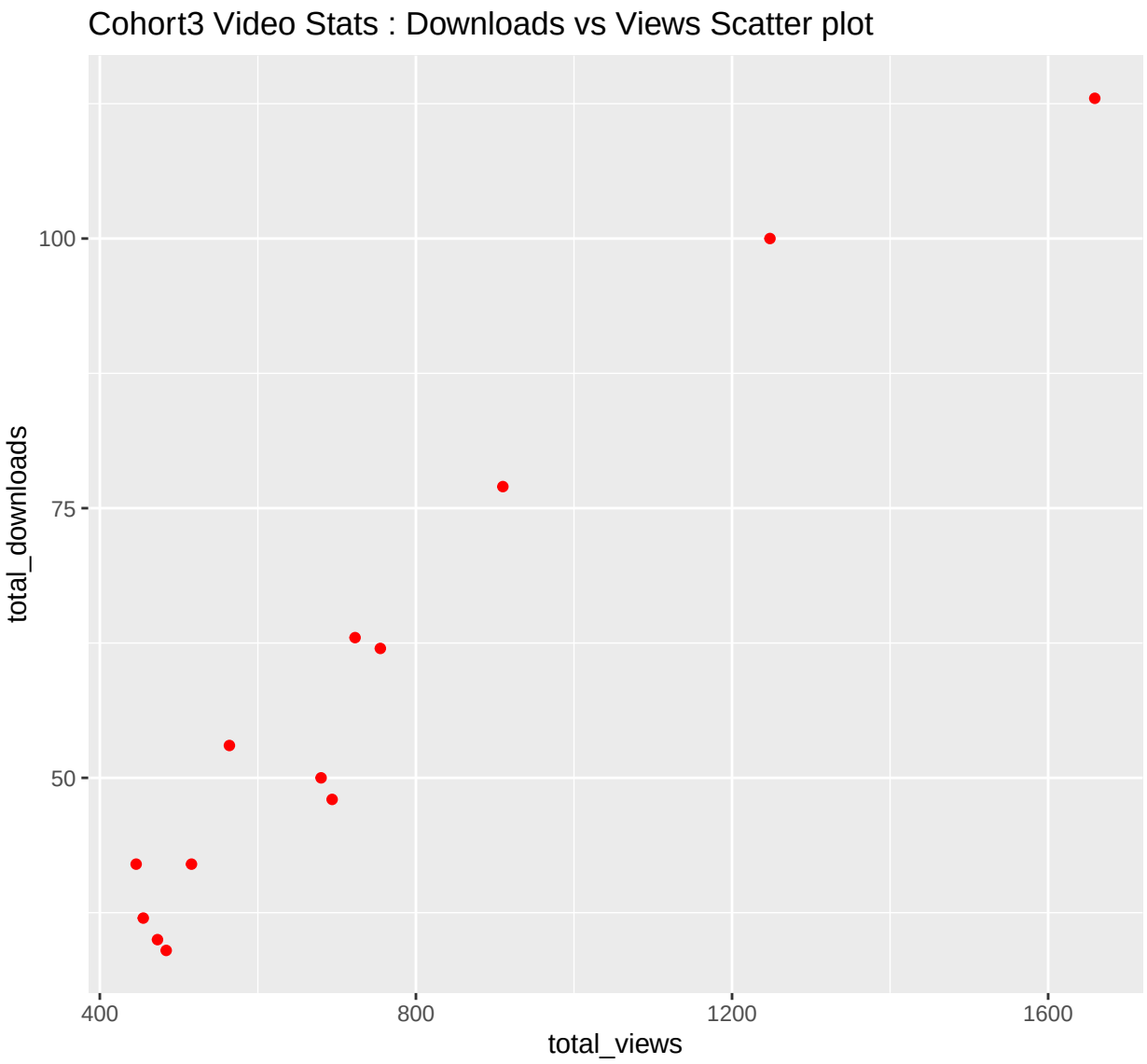
Further, finding the video with maximum views and minimum views by their title names.

```
## [1] "Exploring security: biometric authentication"
```

```
## [1] "Exploring security: biometric authentication"
```

The above results will help us analyse what types of videos are most preferred by students

Scatter-Plot graph of Total views vs Total Downloads



The above plot is to reflect on the linear relationship of the number of views and number of downloads for the videos from each step.

MODEL4

Our video stats file will be used for analysing the viewership across different devices namely

```
## [1] "step_position"             "console_device_percentage"
## [3] "desktop_device_percentage" "mobile_device_percentage"
## [5] "tv_device_percentage"      "tablet_device_percentage"
## [7] "unknown_device_percentage"
```

We find out the means for each of the device

```
means.devices
```

```
## console_device_percentage desktop_device_percentage  mobile_device_percentage
##               0.150769231              80.057692308               8.790769231
##      tv_device_percentage  tablet_device_percentage unknown_device_percentage
##               0.004615385              10.516923077               0.000000000
```

We use these as a measure to determine which of the devices should be focused on and which ones can be completely neglected.

We can draw a conclusion that the models we have used above are all accurate and good fit for our analysis.

EVALUATION