Analysis of cyber security Video Stats

Siddharthan

02/12/2020

Introduction

This is a project report document. The document consists of the analysis of statistics for Future Learn MOOC data set. The observations of the data are made up on different simulations on a particular period of time. The report represents the data analysis techniques that has been used to visualise the data based on the business requirements.The Future Learn MOOC dataset consists of various datafiles such as enrolments, survey response, leaving survey response, question response, step activity, video stats. The observations made on these data are captured based on different simulations on a particular period of time. This report consists of a analysis that is made on the video statistics datafiles. The data consists of various responses of different course modules like video duration, total views of the courses, views based on the regions, views based on the hardware devices and much more. With these responses an exploratory data analysis is made up with some numerical and graphical summaries.

Analysis

Before proceeding with the analysis, the data is inspected to verify whether the structure of the data is good for analysis and whether the data has unknown values in some of the responses. These unknown values are removed based on the data cleaning process and the structure is also changed for the columns which are required for analysis.

Some assumptions are made which were missing in the dataset such as the year and month at which the responses were recorded. These two columns where inserted into the dataset and they are used in the analysis process. Since there were less observations, two datafiles have been used in the analysis process and these two files where combined and stored as a dataframe.

Numerical Analysis

The data mainly consists of values encoded as quantitative variables. The first approach to analysis the data based on the numerical process is to summarize the dataset inorder to get the central tendancy. It can be done by calling the summary() function over the dataset. Here the central tendancy is calculated for set of responses which would be easy for analysing.

```
summary(video_stats_data[,21:26])
```

```
##  europe_views_percentage oceania_views_percentage asia_views_percentage
##  Min.   :39.79           Min.   :2.240            Min.   : 8.24
##  1st Qu.:56.08           1st Qu.:3.195            1st Qu.: 9.55
##  Median :57.56           Median :3.720            Median :13.40
##  Mean   :59.47           Mean   :3.614            Mean   :12.99
##  3rd Qu.:65.55           3rd Qu.:4.030            3rd Qu.:15.73
##  Max.   :67.25           Max.   :4.710           Max.   :23.76
##  north_america_views_percentage south_america_views_percentage
##  Min.   : 8.490                 Min.   :1.650
##  1st Qu.: 9.428                 1st Qu.:2.203
##  Median :10.380                 Median :2.470
##  Mean   :10.414                 Mean   :2.439
##  3rd Qu.:11.418                 3rd Qu.:2.675
##  Max.   :12.210                 Max.   :3.750
##  africa_views_percentage
##  Min.   : 5.170
##  1st Qu.: 6.215
##  Median :10.805
##  Mean   : 9.823
##  3rd Qu.:12.390
##  Max.   :19.950
```

The central tendancy is calculated for the responses that are recorded based on the regions. While looking at this we can consider that majority of the courses are viewed highly from the Europe region rather than the other region. The maximum view percentage from the Europe region 67.25 where as the courses are viewed less in the South America region.

Further to our analysis we can try to find is there any correlation between the responses recorded. To compute the correlation matrix the cor() function can be used. The correlation between different responses establishes that many responses are correlated to each other. In that, we could find that the total_downloads response is highly correlated with transcript_views response of **0.9409151**

```
cor(video_stats_data[,5], video_stats_data[,7])
```

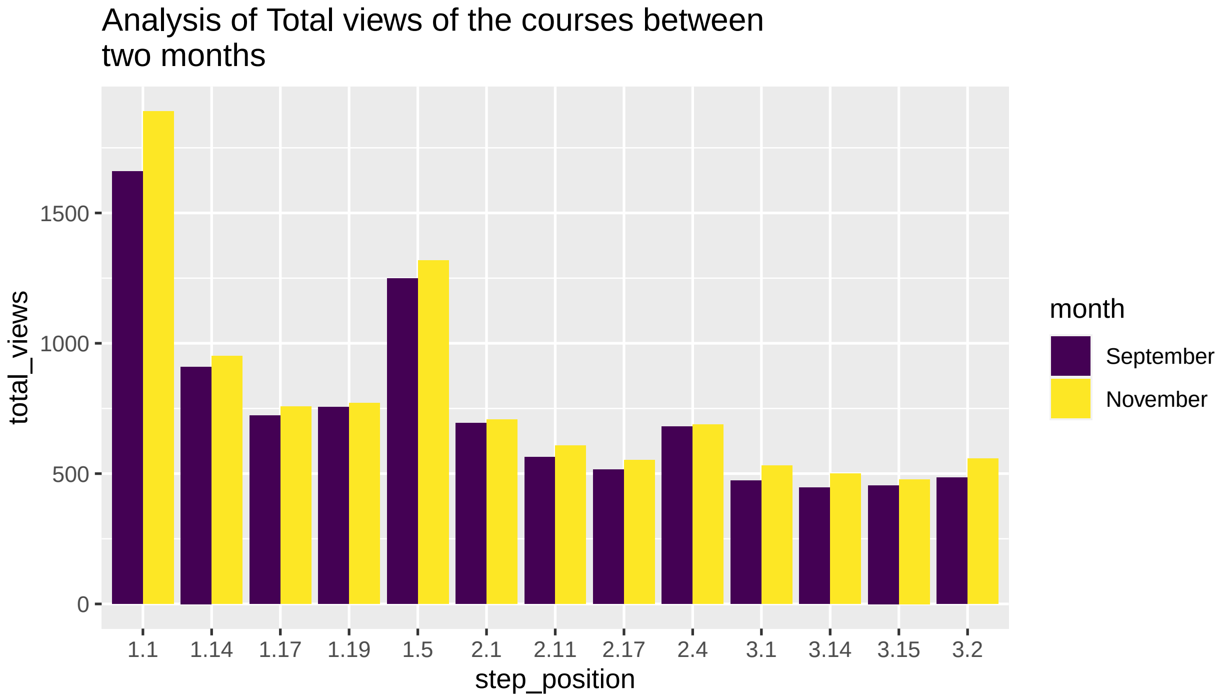
```
## [1] 0.9409151
```

The index 5 and 7 from the code represents the two responses total_downloads abd total_transcript_views respectively. From this its applicable to represent that the courses that are downloaded mostly as a transcript type. With this numerical summaries, the data can be further analysed graphically.

Graphical Analysis

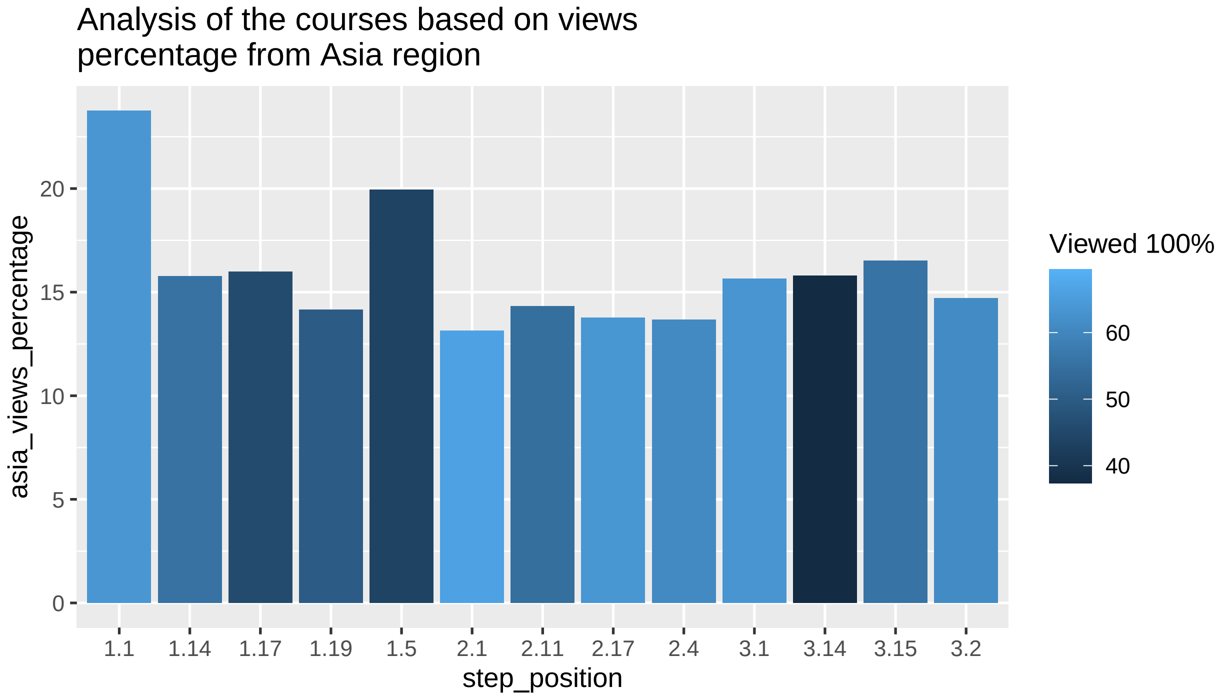
The graphical analysis is made with line graphs, points and bar plots, which helps to visualize and interpret the data. The **ggplot2** library is used for this graphical analysis.

Plot1 :



The plot represents the analysis made with total views of the courses between two months from the dataset. This plot graphically explains that there is a certain amount of increase in views to the latter month which is *November* compared to that of the previous month *September*. The modules has made some significant impact with the leaners and this made a growth in the trend of viewers. The modules which are viewed in less numbers were also made an impact the following month with the increase in the viewers.

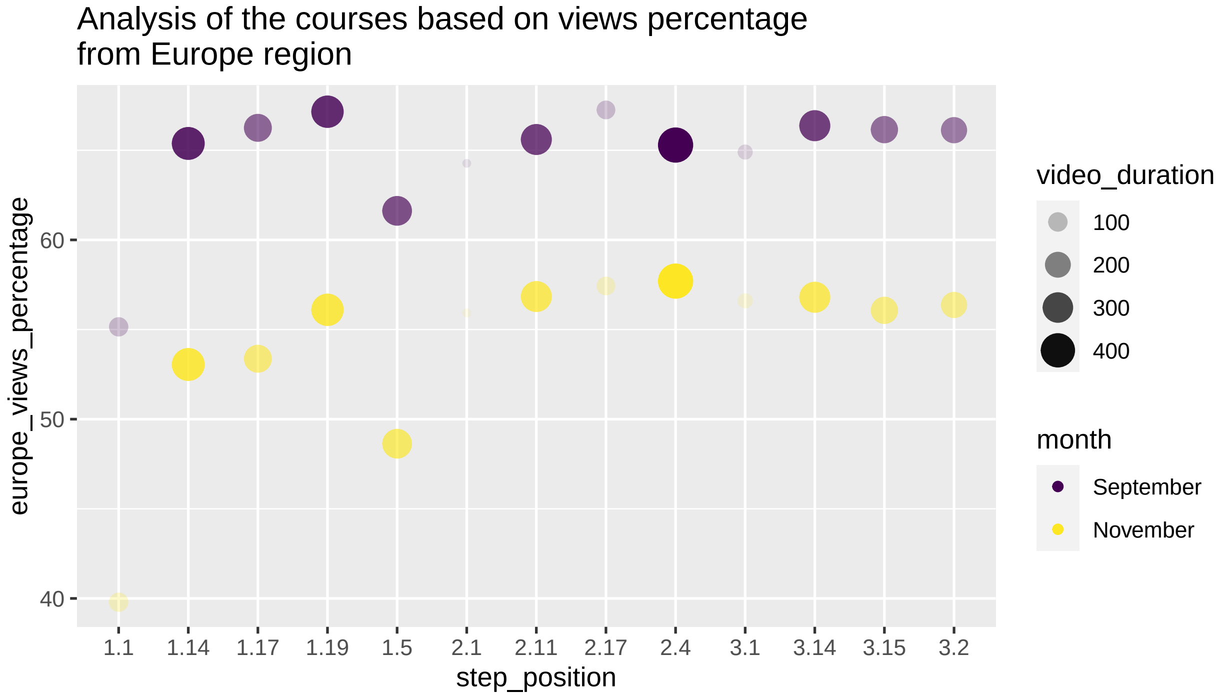
Plot2 :



The above bar plot represents the analysis made based on the modeules that are viewed ***100%*** from the

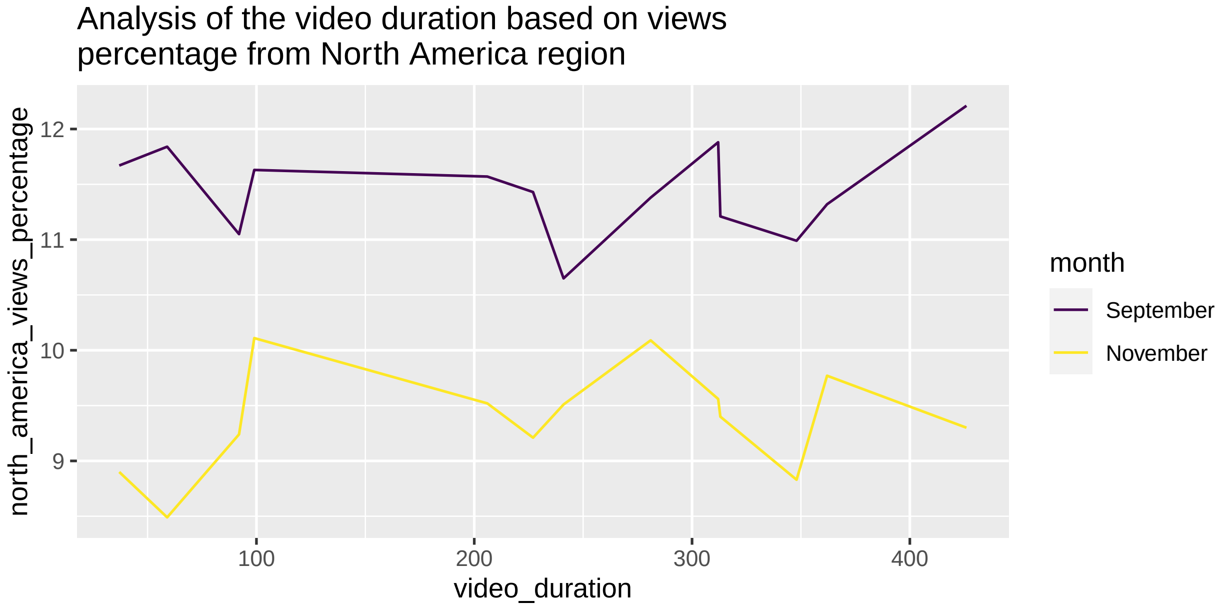
Asia region. From this graph its easy to interpret that above **60%** of the learners have viewed most of the modules completely. And below **40%** of the learners have viewed two or three modules completely. This suggests that the learners are not interested in viewing those modules as it maybe out of the scope.

Plot3 :



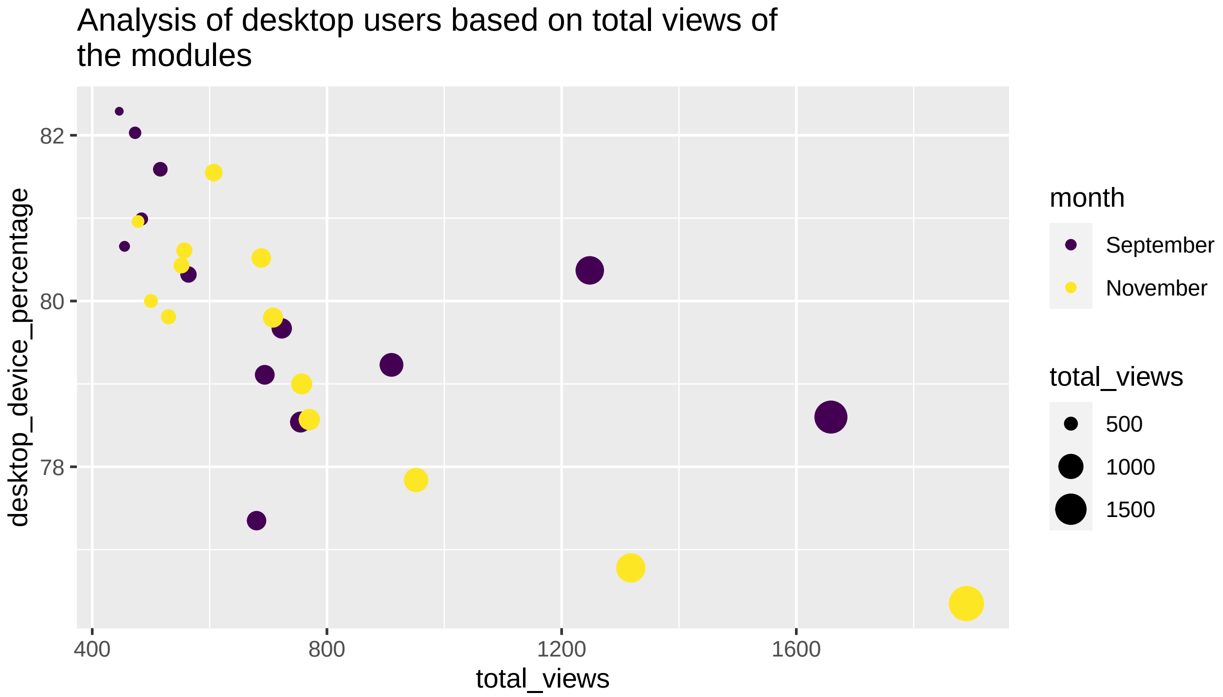
The obtained point plot represents the analysis made on the courses based on the view percentage from the Europe region. From the plot, its significantly visible that the module did not made a much impact in the Europe region learners as the view percentage critically dropped during the **November** month compared to the month of **September**. The lightly shaded points shows that the video duration is less than 100 minutes and the dark shaded points shows the video duration is more than 400 minutes.

Plot4 :



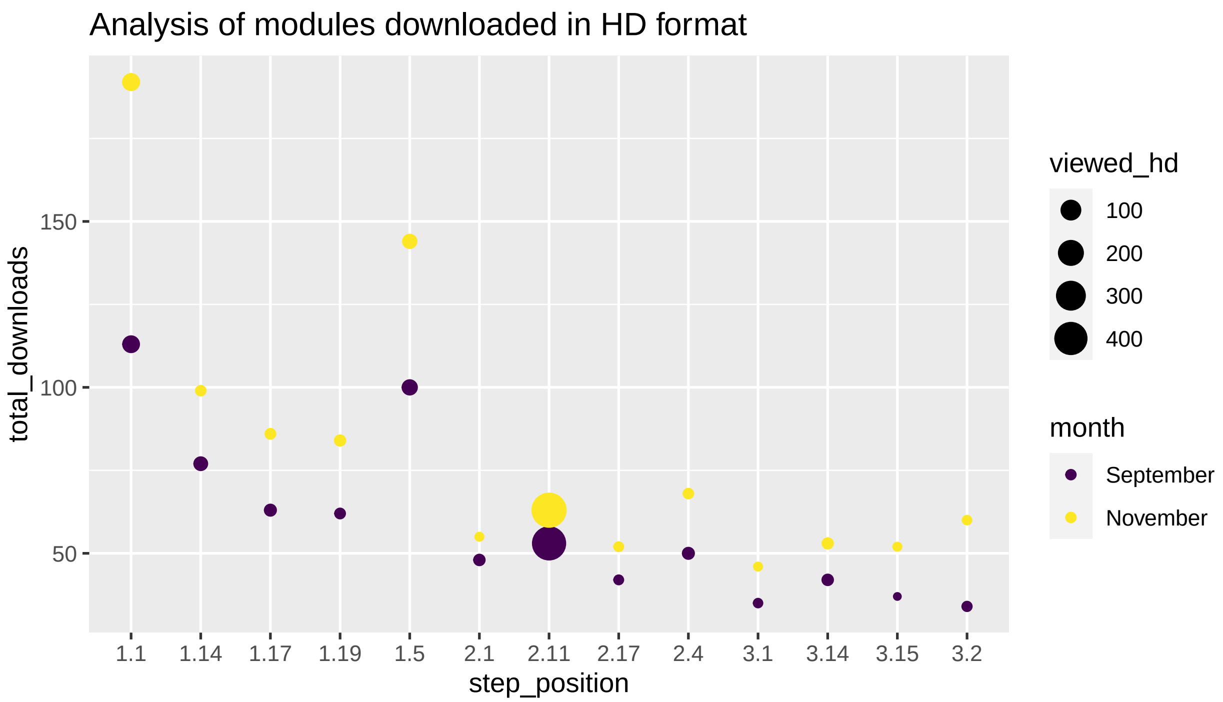
The line graph represents the simple analysis made on the view percentage of the North American region learners with respect to the video duration. It is plotted comparing with the different months. Its obvious that the learners have lost their interest in the modules as it clearly shows a depletion in the view percentage in the **November** month. This sums up that the learners started downloading the contents of the modules rather than viewing the materials online.

Plot5 :



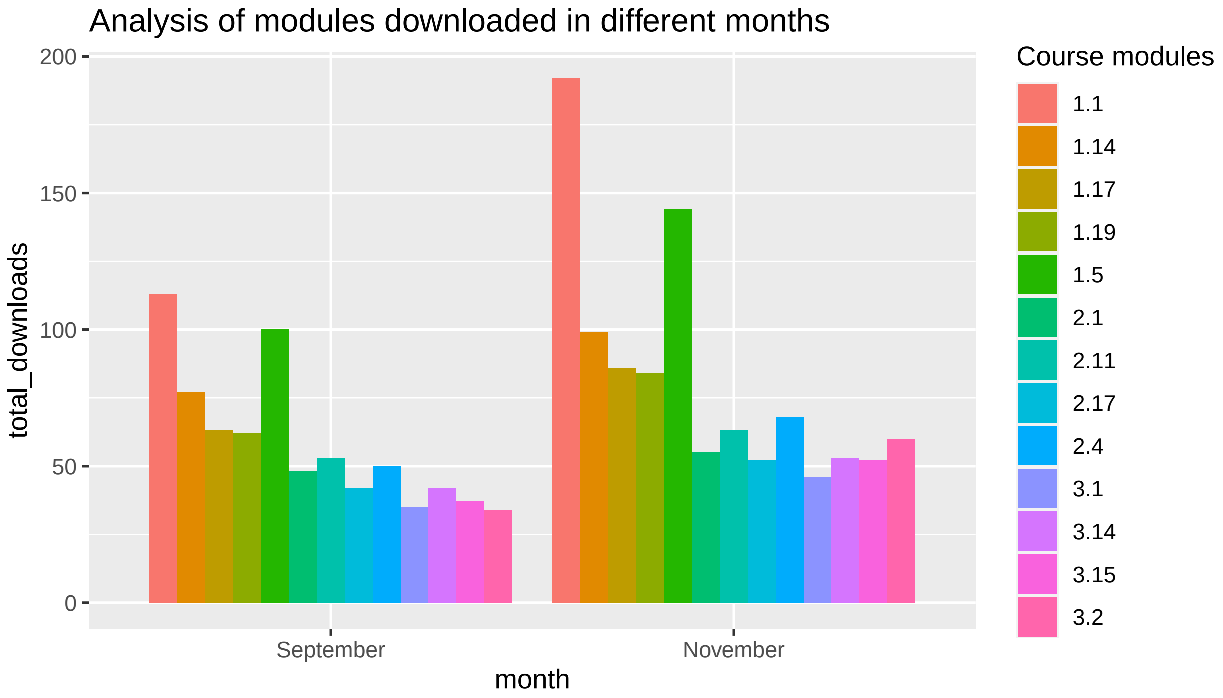
The obtained point plot represents the analysis made on the modules that are viewed using the desktop devices. Based on the total views of the modules, the points represent that the learners viewed through the desktop devices for most important topics, whereas there is a shrinkage in desktop views for the courses which has less viewers.

Plot6 :



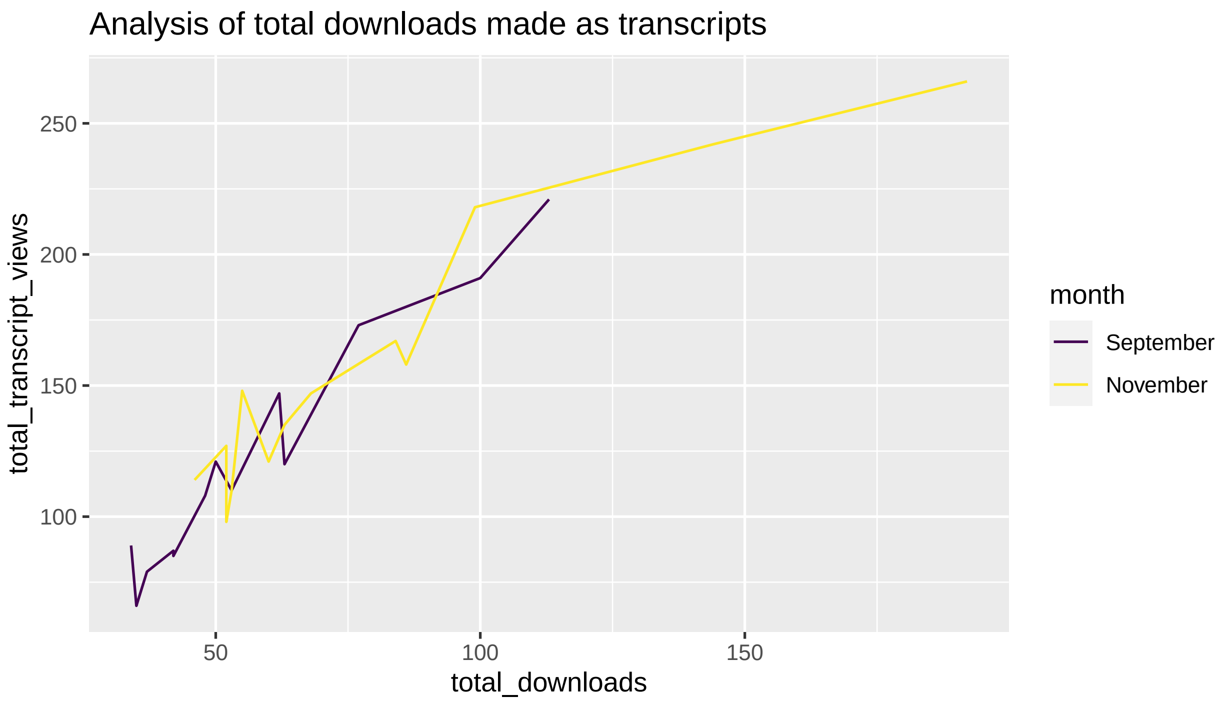
The above obtained plot represents the modules that are downloaded and viewed in the hd format. From this graph, it obvious that only one module is downloaded and viewed in hd format by most of the learners. This suggests that the learners had an impact from that particular module as we can assume it was predominantly used in their work environment.

Plot7 :



This bar chart clearly shows that there is an increase in course downloads in later month compared to that of the previous month. This clearly represents that the learners are interested in the contents of the course and they downloaded the contents for further usage purposes. There was no depletion in any course that was being downloaded compared to the previous month.

Plot8 :



As observed in the numerical summaries that there is huge correlation between the total_downloads and transcription views, this line plot clearly shows that in both months, maximum of the course materials are downloaded as the transcripts. And in the month of **November** the materials that are downloaded as the transcript view has increased to the peak.