210468180_Report

Parth Umeshkumar Shah

30/11/2021

Introduction

With the rise in specialization in the cyber security,there is an increase in interest of learning it for career growth in candidates. With the more day in and day out usage of Internet and due to unprecended times most of the Universities have decided to make the course available via online sessions.As the course is delivered via online mode, there might be some difficulties, some issues which could be faced by both the learners and the professors who are handling the modules.In order to make the online sessions more comfortable and to make the students get used to it,like they can view the course materials whenever they want etc. These online learnings are recorded in order to improve the course handling. The online activities of enrollments, survey response, leaving survey response, question response, step activity and statistics of viewing the videos are recorded and stored in the Future Learn MOOC dataset.

The observations made on these data are developed based on different simulations on a particular period of time. This report consists of a analysis that is made on the statistics of viewing of the videos. The data consists of various responses of different course modules like video duration, total views of the courses,views based on the hardware devices,views based on the regions and much more. With these responses an exploratory data analysis is prepared here to get a better insight of the statistics of viewing of the videos using different numerical and graphical analogies.

Analysis

Before proceeding with the analysis on data it 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.

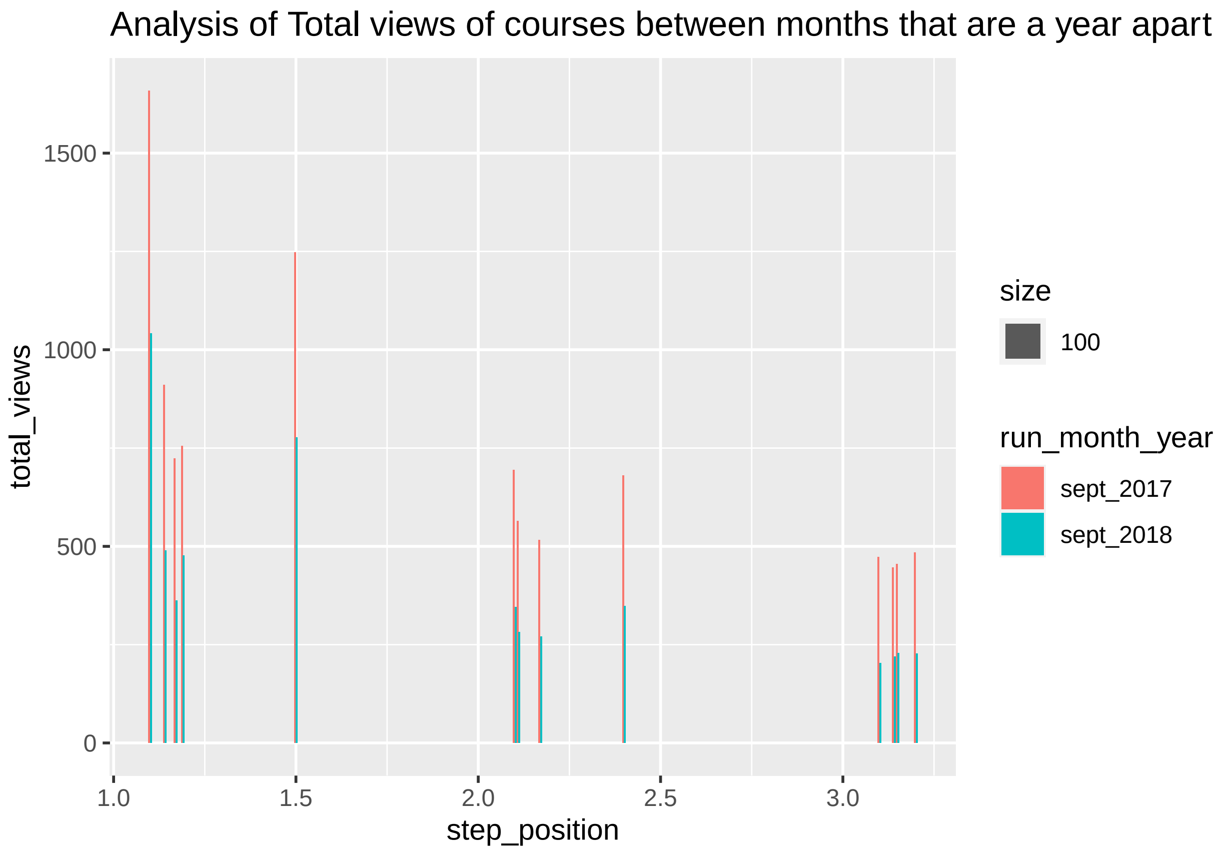
Here for analysis three data sets are explored which are runned in september-2017,february-2018 and september-2018 respectively. This data sets are chosen such that analysis of data at time interval of one year apart and half year apart can be seen and explored,through such distinct datasets a relation can be explored in upgrade or downgrade of the online sessions in a year duration whuch can be essential in business modelling in future for the company.Hence,three datafiles have been used in the analysis process which had captured and developed datasets which were half year and a year apart,2 new files are combination of datasets with one having merge data of september-2017,february-2018 and another consisting of a year apart data i.e. september-2017 and september-2018,files in here are stored as a dataframe.

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.Also datasets are cleaned with removing the columns having null values,further datafiles are arranged as per the exploratory questions demanded.

Whole dataset mainly consists of values encoded as quantitative variables. The first approach of the exploratory analysis is to calculate the average drop in viewership from sept,2017 in respect to sept,2018 surveyed dataset. According to CRISP-DM methodology another cycle of analysis is done on checking whether the yearly__drop from september,2017 to september,2018 has any correlation with mid__year__drop from september,2017 to february,2018 by Finding the average drop/rise in viewership from sept,2017 compared to feb,2018. The datafiles are processed,analysed graphically and with help numerical analogies and numerical equations solving tools using the "dplyr" "ggplot" and "ProjectTemplate" R-packages.

QUESTION__01-Finding the average drop in viewership.

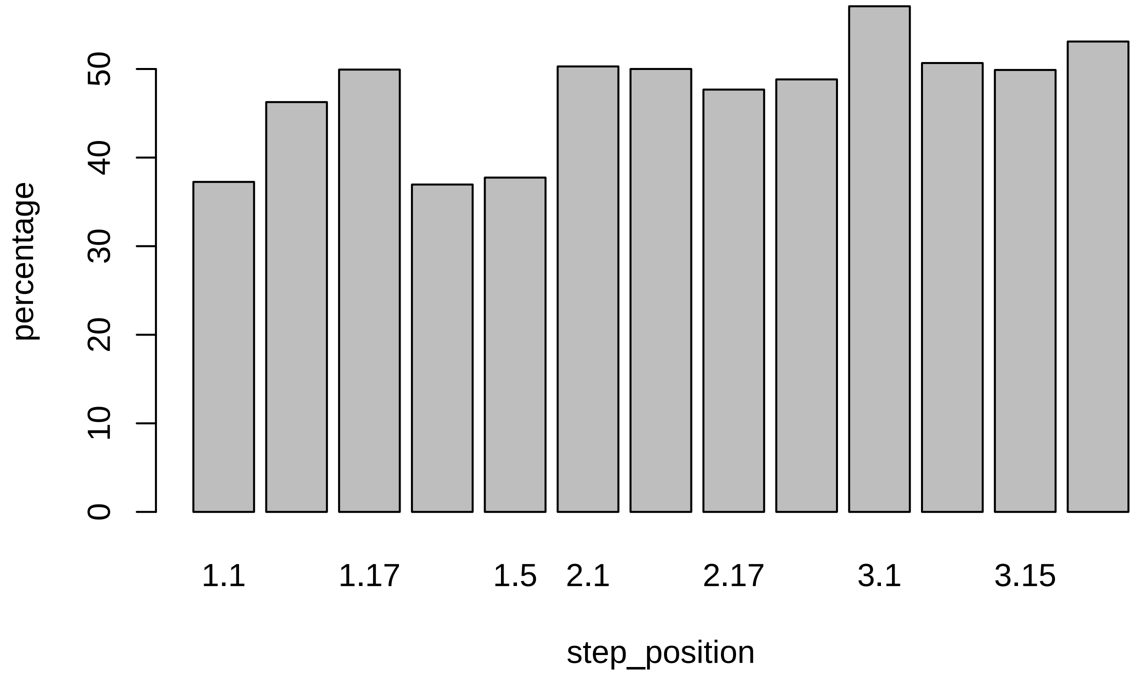
```
#cycle-01
plot_j1 = ggplot(Video_stats, aes(step_position, total_views, fill = run_month_year, size=100))+
geom_col(position = 'dodge') +
ggtitle("Analysis of Total views of courses between months that are a year apart")
plot_j1
```



```
drop_percentage = as.data.frame(matrix(nrow = 0, ncol = 2));
for(i in 1:26){
  if(i %% 2 != 0){
row <- c(Video_stats$step_position[i],
((Video_stats$total_views[i] - Video_stats$total_views[i + 1]) / Video_stats$total_views[i]
             ) * 100)
    drop_percentage <- rbind(drop_percentage,row)
     }
}
colnames(drop_percentage) <- c("step_position", "percentage")
drop_percentage
```

```
##    step_position percentage
## 1           1.10   37.25136
## 2           1.14   46.26374
## 3           1.17   49.93084
## 4           1.19   36.95364
## 5           1.50   37.74038
## 6           2.10   50.28818
## 7           2.11   50.00000
## 8           2.17   47.67442
## 9           2.40   48.82353
## 10          3.10   57.08245
## 11          3.14   50.67265
## 12          3.15   49.89011
## 13          3.20   53.09917
```

```
average_drop = (sum(drop_percentage[,2])/13)
dropplot_1 = barplot(drop_percentage$percentage ~ drop_percentage$step_position,
                     xlab = "step_position", ylab = "percentage")
```

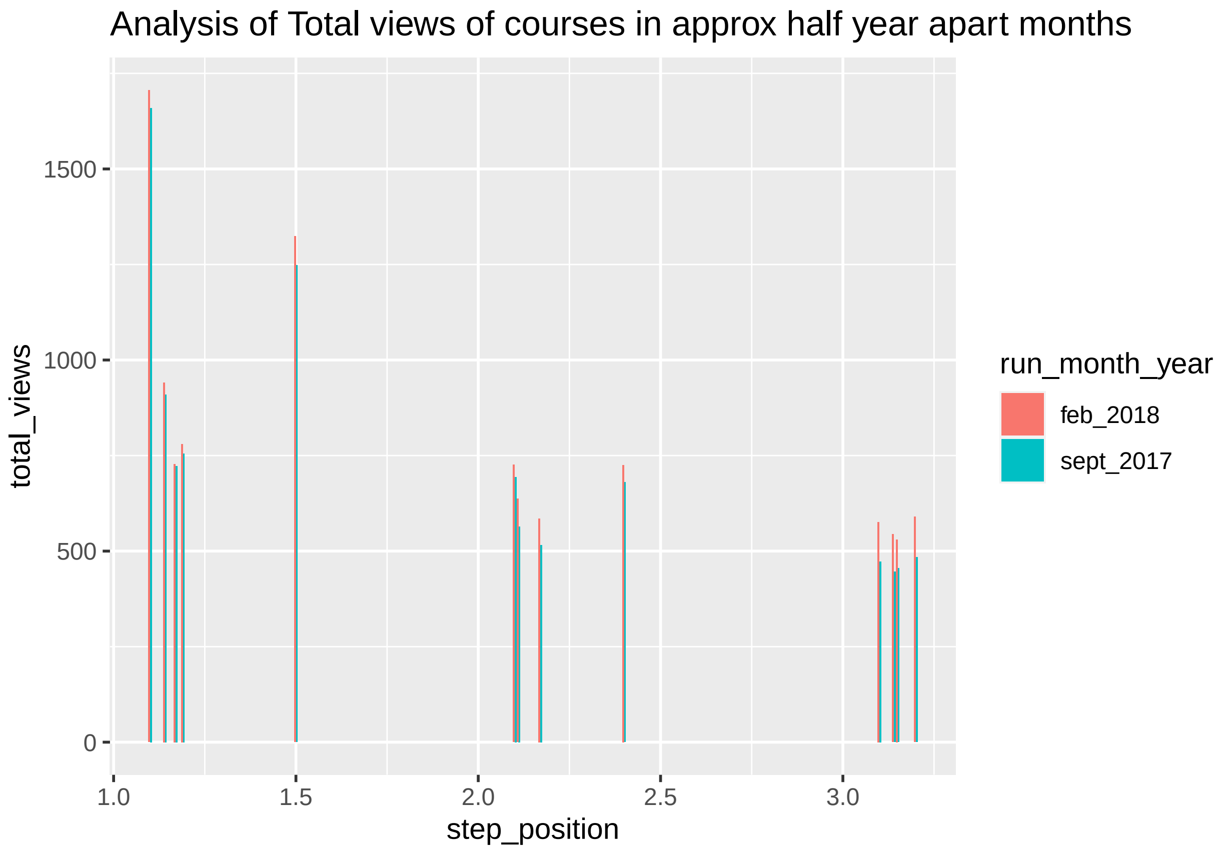


Cycle 01__Que1-Finding the average drop in viewership from sept,2017 in relation to sept,2018

- The plot_j1 represents of total views based on courses for 2 different months with a gap of a year was inserted for visualization purpose.
- Created a data frame named drop_percentage and initialized a for-loop for calculation of drop-percentage such that the data of it is stored in a 2 column data frame consisting of drop percentage for each step-position.

- Through the use of drop percentage,average drop is calculated for the explored data file and was viewed that the viewership is dropped 47.35927% in september,2018 in comparision to september,2017

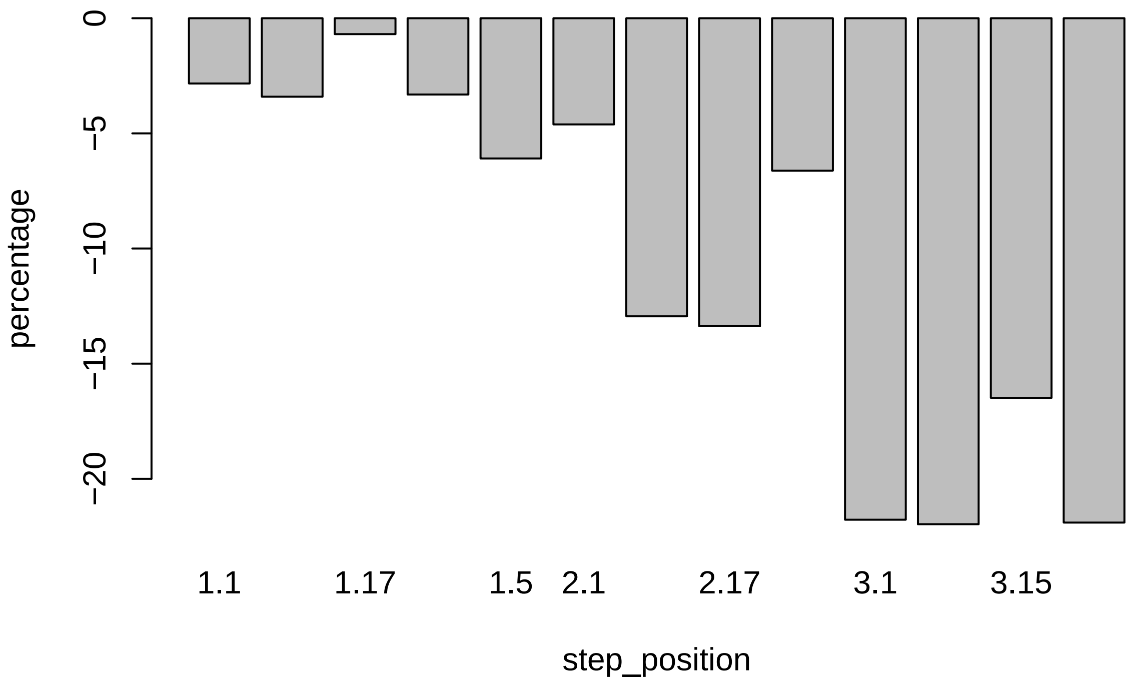
```
#Q1-Cycle 02
plot_j2 = ggplot(Video_stats_Q1_Cycle02, aes(step_position, total_views, fill = run_month_year))+
geom_col(position = 'dodge') +
ggtitle("Analysis of Total views of courses in approx half year apart months")
plot_j2
```



```
drop_percentage_mid_yearly = as.data.frame(matrix(nrow = 0, ncol = 2));
for(i in 1:26){
if(i %% 2 != 0){
row <- c(Video_stats_Q1_Cycle02$step_position[i],
((Video_stats_Q1_Cycle02$total_views[i] - Video_stats_Q1_Cycle02$total_views[i + 1]) / Video_stats_Q1_C
  drop_percentage_mid_yearly <- rbind(drop_percentage_mid_yearly,row)
  }
}
colnames(drop_percentage_mid_yearly) <- c("step_position", "percentage")
drop_percentage_mid_yearly
```

```
##    step_position  percentage
## 1           1.10  -2.8330319
## 2           1.14  -3.4065934
## 3           1.17  -0.6915629
## 4           1.19  -3.3112583
## 5          1.50  -6.0897436
## 6          2.10  -4.6109510
## 7          2.11 -12.9432624
## 8          2.17 -13.3720930
## 9          2.40  -6.6176471
## 10         3.10 -21.7758985
## 11         3.14 -21.9730942
## 12         3.15 -16.4835165
## 13         3.20 -21.9008264
```

```
average_drope_mid_yearly = (sum(drop_percentage_mid_yearly[,2])/13)
dropplot_2 = barplot(drop_percentage_mid_yearly$percentage ~ drop_percentage_mid_yearly$step_position,
                     xlab = "step_position", ylab = "percentage")
```



Cycle 02__Que1- Finding the average drop in viewership from sept,2017 to feb,2018

Checking whether the pattern of yearly_drop from september,2017 to september,2018 has some relation with mid_year_drop from september,2017 to february,2018

- The plot_j2 represents of total views based on courses for 2 different months with a gap of a 5 months(approx = half year period) was inserted for visualization purpose.
- Created a data frame initialized a for-loop for calculation of drop-percentage for files half year apart such that the data of it is stored in a 2 column data frame consisting of drop percentage for each step-position.
- Through the use of drop percentage,average drop is calculated for the explored data file and was viewed that the viewership was rised 10.46227 in February,2018 in comparision to september,2017

- From this two analyzation it can be assumed that in comparision to Sept,2017 the viewership increased in feb,2018 which shows the increasing interest of people in Cyber Security field and the drop in sept,2018 in comparision to sept,2017 shows that maybe people who want to percieve further in Cyber Security has already enrolled and completed the course or also it can be assumed that the content available has been mkoved to a saturated zone where people who has interest in this filed are already known to this much of information.

From the view of Business perspective: 1) It can be assumed and can be said that more additional step-units if added or more marketing is done may we see a rise in future runs. OR 2)It can be assessed and assumed that majority of the courses,if good,rise in the first half and then as usual the decrement is seen,as seen in here,which leads to thinking that company should move on positively and make more online videos on different topics and hence the business continues, also this fine way of online-teaching sustains and prevails with people can company both being benifitted.

Q2.DEFINITE USE OF CORRELATION TO EXPLORE THE DATA and TO FIND CONCLUSIONS ON SEVERAL QUESTIONS.

```
#Q2_CYCLE-01

summary(Video_stats[,09:15])
```

```
##  viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
##  Min.   :68.97       Min.   :65.91      Min.   :61.76            
##  1st Qu.:72.55       1st Qu.:71.52      1st Qu.:67.65            
##  Median :73.67       Median :72.99      Median :69.94            
##  Mean   :74.46       Mean   :72.90      Mean   :70.17            
##  3rd Qu.:76.83       3rd Qu.:75.01      3rd Qu.:73.86            
##  Max.   :81.77       Max.   :79.63      Max.   :76.18            
##  viewed_fifty_percent viewed_seventyfive_percent viewed_ninetyfive_percent
##  Min.   :57.56        Min.   :55.46              Min.   :53.15            
##  1st Qu.:64.92        1st Qu.:61.68              1st Qu.:60.03            
##  Median :67.34        Median :65.88              Median :62.74            
##  Mean   :67.38        Mean   :65.45              Mean   :63.15            
##  3rd Qu.:70.36        3rd Qu.:68.43              3rd Qu.:66.86            
##  Max.   :74.78        Max.   :73.62              Max.   :72.09            
##  viewed_onehundred_percent
##  Min.   :34.09            
##  1st Qu.:49.48            
##  Median :57.09            
##  Mean   :56.00            
##  3rd Qu.:63.55            
##  Max.   :71.01
```

```
cor(cyber.security.3_video.stats[,3], cyber.security.3_video.stats[,9:15])
```

```
##                viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## video_duration           -0.585404         -0.6149269                 -0.830404
##                viewed_fifty_percent viewed_seventyfive_percent
## video_duration           -0.8744593                  -0.852296
##                viewed_ninetyfive_percent viewed_onehundred_percent
## video_duration                -0.8068828                -0.6221957
```

```
cor(cyber.security.5_video.stats[,3], cyber.security.5_video.stats[,9:15])
```

```
##                viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## video_duration          -0.5102708         -0.6518072                 -0.713544
##                viewed_fifty_percent viewed_seventyfive_percent
## video_duration           -0.7358931                 -0.7427564
##                viewed_ninetyfive_percent viewed_onehundred_percent
## video_duration                -0.6827021                -0.5522303
```

```
cor(cyber.security.7_video.stats[,3], cyber.security.7_video.stats[,9:15])
```

```
##                viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## video_duration          -0.5617774         -0.7128997                -0.7862203
##                viewed_fifty_percent viewed_seventyfive_percent
## video_duration           -0.8569559                 -0.8577135
##                viewed_ninetyfive_percent viewed_onehundred_percent
## video_duration                -0.8331848                -0.6383265
```

```
cor(Video_stats_Q2_Cycle01[,3], Video_stats_Q2_Cycle01[,9:15])
```

```
##                viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## video_duration          -0.5360397         -0.6515607                -0.7579282
##                viewed_fifty_percent viewed_seventyfive_percent
## video_duration           -0.8253775                 -0.8205328
##                viewed_ninetyfive_percent viewed_onehundred_percent
## video_duration                -0.7976317                 -0.629897
```

```
cor(Video_stats_Q1_Cycle02[,3], Video_stats_Q1_Cycle02[,9:15])
```

```
##                viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## video_duration          -0.3809696         -0.4596426                -0.6511659
##                viewed_fifty_percent viewed_seventyfive_percent
## video_duration           -0.7079647                 -0.7146213
##                viewed_ninetyfive_percent viewed_onehundred_percent
## video_duration                 -0.669229                -0.5722971
```

This approach to analysis of the data based on the numerical process is to summarize the dataset to explore whether **"More people left at less percentage because of the length of the VIDEO that are availabe in the course?"** Summary() function is called to explore on the data sets of % of people have left after a definite % of video is watched inorder to get the central tendency.By calling the summary() function over the datasets the central tendency is calculated for set of responses which would be easy for analysing.

Q2_CYCLE-01 analysis- Have more people left at less percentage because of the length of the VIDEO?

- The central tendancy is calculated for the responses that are recorded based on it, a general trend of decrease in number of viewers of each video is seen,which proves that approx nearby 40% of people skips videos till they reach 100% length of it.
- As the above statement is an assumption seen we correlated the data of videdo viewed percentage wise with video duration across all the three runs of sept-2017,feb-2018 and sept,2018 and we got to see a pattern as follows :

– All the three data files of different runs shows a negative Co-relation between the video_duration and the video-viewed percentage i.e. (viewed_five_percent, viewed_ten_percent, viewed_twentyfive_percent, viewed_fifty_percent, viewed_seventyfive_percent, viewed_ninetyfive_percent, viewed_onehundred_percent) files.

–This negative Co-relation stats that as the video duration increases the video-viewed percentage is dropped and it proves that yes the video duration is a major consideration across all files

For business purposes,also explored taking the merged files of data across one year apart runned file and half year apart runned files,the same pattern of negative co_relation is seen.

- Hence it can be assumed that if we make more creative short individual topic videos,it migh help people to continue with the course as many people who have started have never reached the end of particular video.

```
#Q2_CYCLE-01 analysis

cor(cyber.security.3_video.stats[,1], cyber.security.3_video.stats[,9:15])
```

```
##               viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## step_position          -0.1434446        -0.02876844                 0.2267197
##               viewed_fifty_percent viewed_seventyfive_percent
## step_position             0.300446                  0.3306305
##               viewed_ninetyfive_percent viewed_onehundred_percent
## step_position                  0.291831                 0.1536037
```

```
cor(cyber.security.5_video.stats[,1], cyber.security.5_video.stats[,9:15])
```

```
##               viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## step_position          -0.6255863         -0.4625301                -0.2373934
##               viewed_fifty_percent viewed_seventyfive_percent
## step_position          -0.02523698                  0.0342674
##               viewed_ninetyfive_percent viewed_onehundred_percent
## step_position               -0.07480335                0.02810366
```

```
cor(cyber.security.7_video.stats[,1], cyber.security.7_video.stats[,9:15])
```

```
##               viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## step_position           -0.402063         -0.3129076                -0.1265947
##               viewed_fifty_percent viewed_seventyfive_percent
## step_position           0.09604537                  0.1895778
##               viewed_ninetyfive_percent viewed_onehundred_percent
## step_position                 0.1512171               -0.05657736
```

```
cor(Video_stats[,1], Video_stats[,9:15])
```

```
##               viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## step_position          -0.2997476         -0.2033912               0.006043731
##               viewed_fifty_percent viewed_seventyfive_percent
## step_position            0.1711968                  0.2385402
##               viewed_ninetyfive_percent viewed_onehundred_percent
## step_position                 0.2076122                0.04720115
```

```
cor(Video_stats_Q1_Cycle02[,1], Video_stats_Q1_Cycle02[,9:15])
```

```
##               viewed_five_percent viewed_ten_percent viewed_twentyfive_percent
## step_position          -0.2854048         -0.1786109               -0.02692023
##               viewed_fifty_percent viewed_seventyfive_percent
## step_position            0.1145564                  0.1614905
##               viewed_ninetyfive_percent viewed_onehundred_percent
## step_position                0.09011239                0.08689115
```

Q2_CYCLE-02

With use of correlation process here analysis is done on step_position and video_viewed_percentage wise for conclusions.

- As analysed running the above specified code across the 3 loaded data files we can draw attention on the trend that with increase in step_position the video_viewed_percentagewise of less percent data decreases for example according to data -with increase in step_position the 5% viewed data decreases,hence generally people who are interested in watching watches more and people who are less intersted leave in the early step_positioned videos only.
- This assumption grows strong by seeing the correlation data trends of increase in step_position leads to increase in the video_viewed_percentagewise of more percent data column. for example according to data -with increase in step_position the 50%,75%,95%,100% viewed data generally found increasing,hence the assumption that generally people who are interested in watching watches more and people who are less interested leave in the early step_positioned videos onlygrows strong.

*From business perspective: it can be stated that Company providing online sessions have a good content and keeping this type of content level will benifit the people who are deeply interested in learnings like Cyber.security and simultaneously the company will also grow.

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

```
##                 total_transcript_views
## total_downloads               0.868636
```

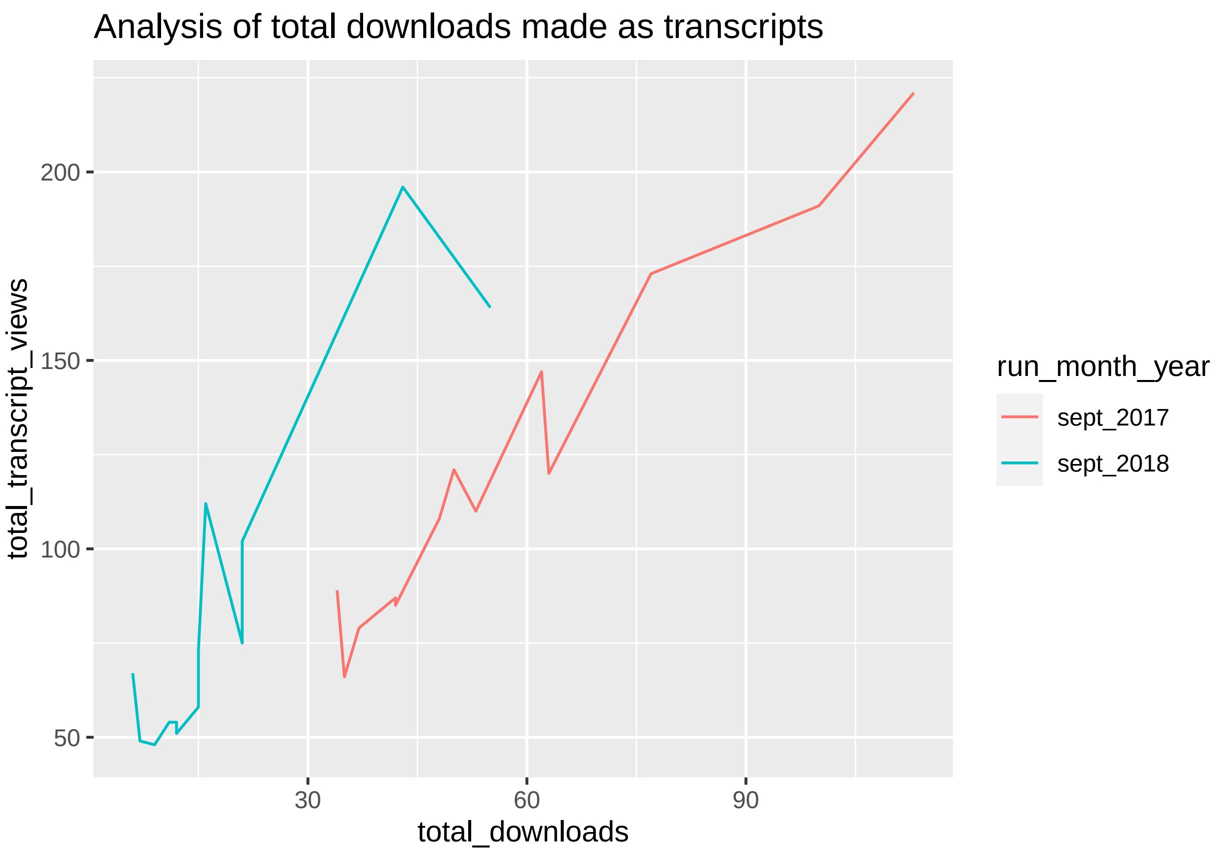
```
cor(Video_stats_Q1_Cycle02[,5], Video_stats_Q1_Cycle02[,7])
```

```
##                 total_transcript_views
## total_downloads              0.9410766
```

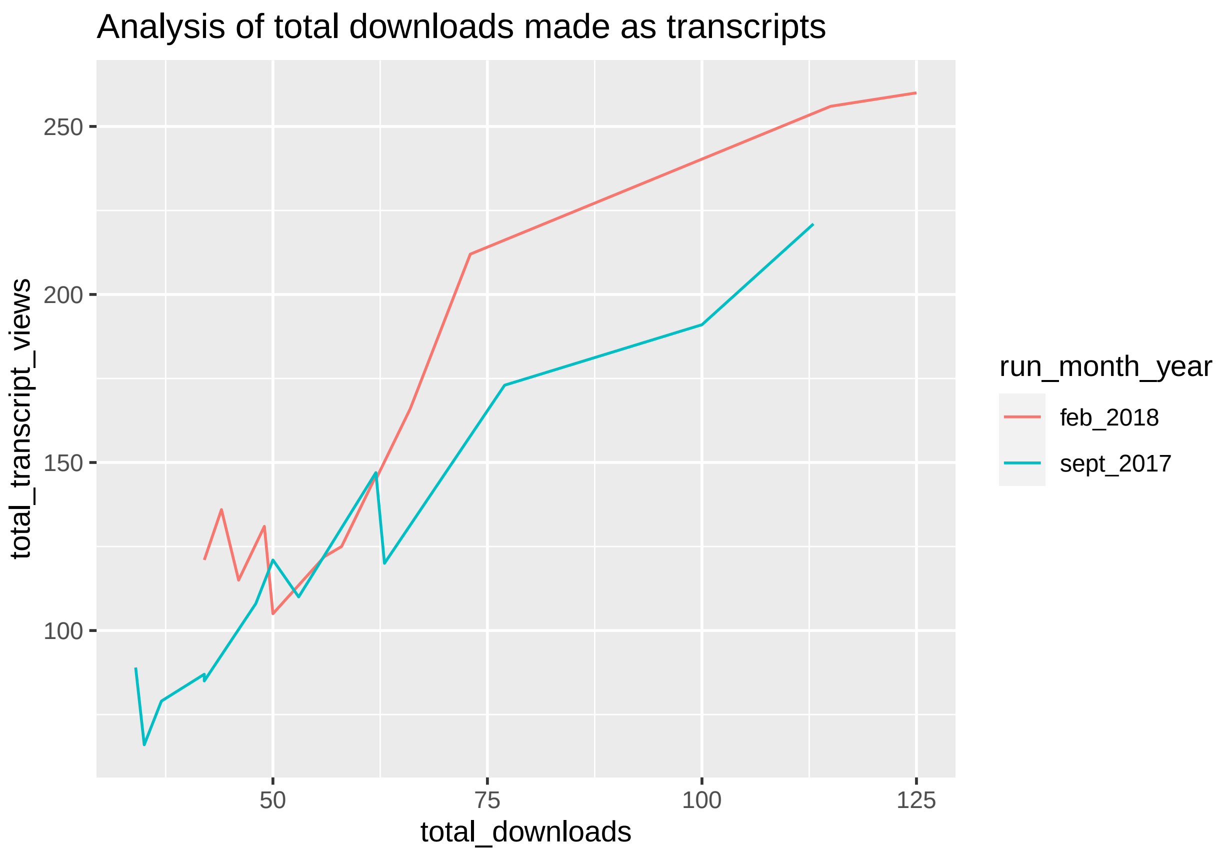
~ **More Analysis using Correlation** * The index 5 and 7 from the code represents the two responses total_downloads and total_transcript_views respectively. * From this, its applicable to represent that the courses that are downloaded are mostly as a transcript type in all the files i.e. files runned in sept-2017,feb-2018 and sept,2018.

With this numerical summaries, the data can be further analysed *graphically* also as:

```
plot_A = ggplot(Video_stats, aes(total_downloads, total_transcript_views, color = run_month_year)) + ge
ggtitle("Analysis of total downloads made as transcripts")
plot_A
```



```
plot_B = ggplot(Video_stats_Q1_Cycle02, aes(total_downloads, total_transcript_views, color = run_month_y
ggtitle("Analysis of total downloads made as transcripts")
plot_B
```

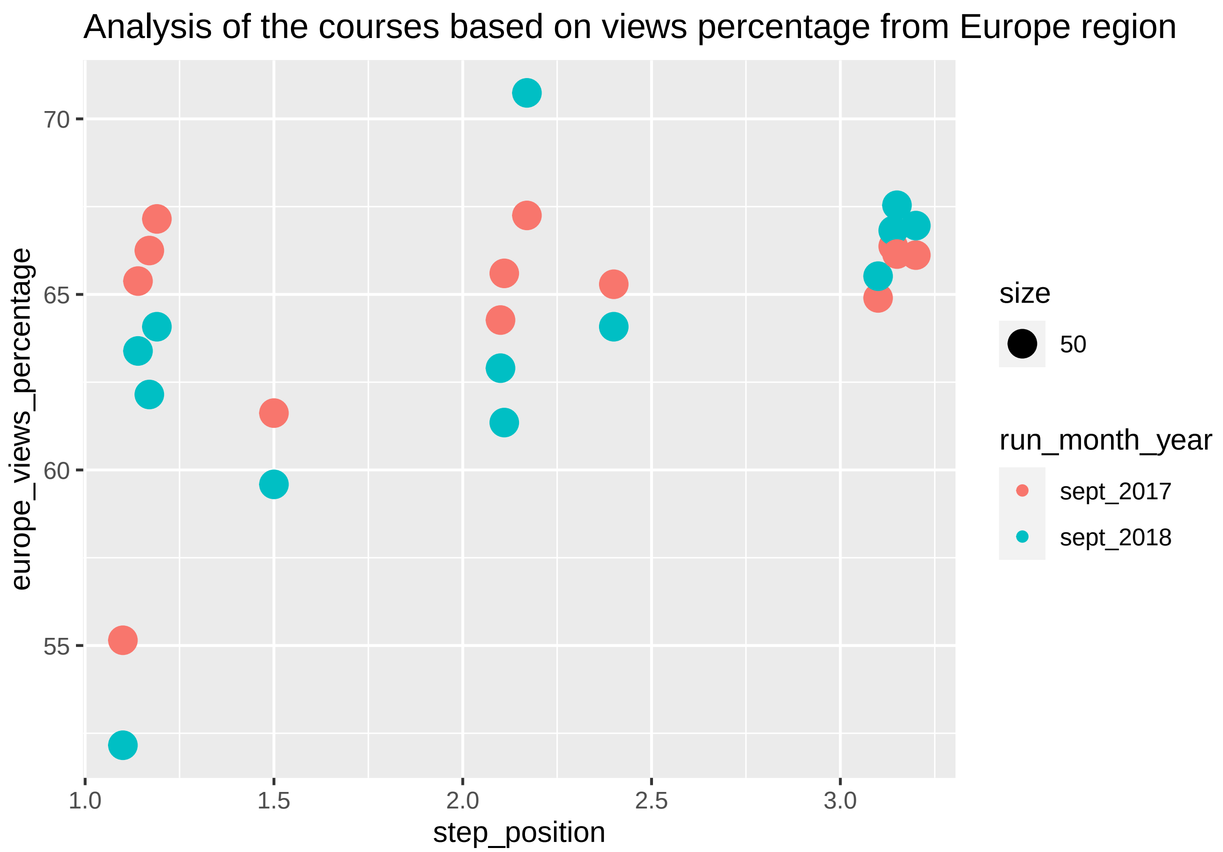


EDA using graphical tool for analysis and TO FIND some conclusions.

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.

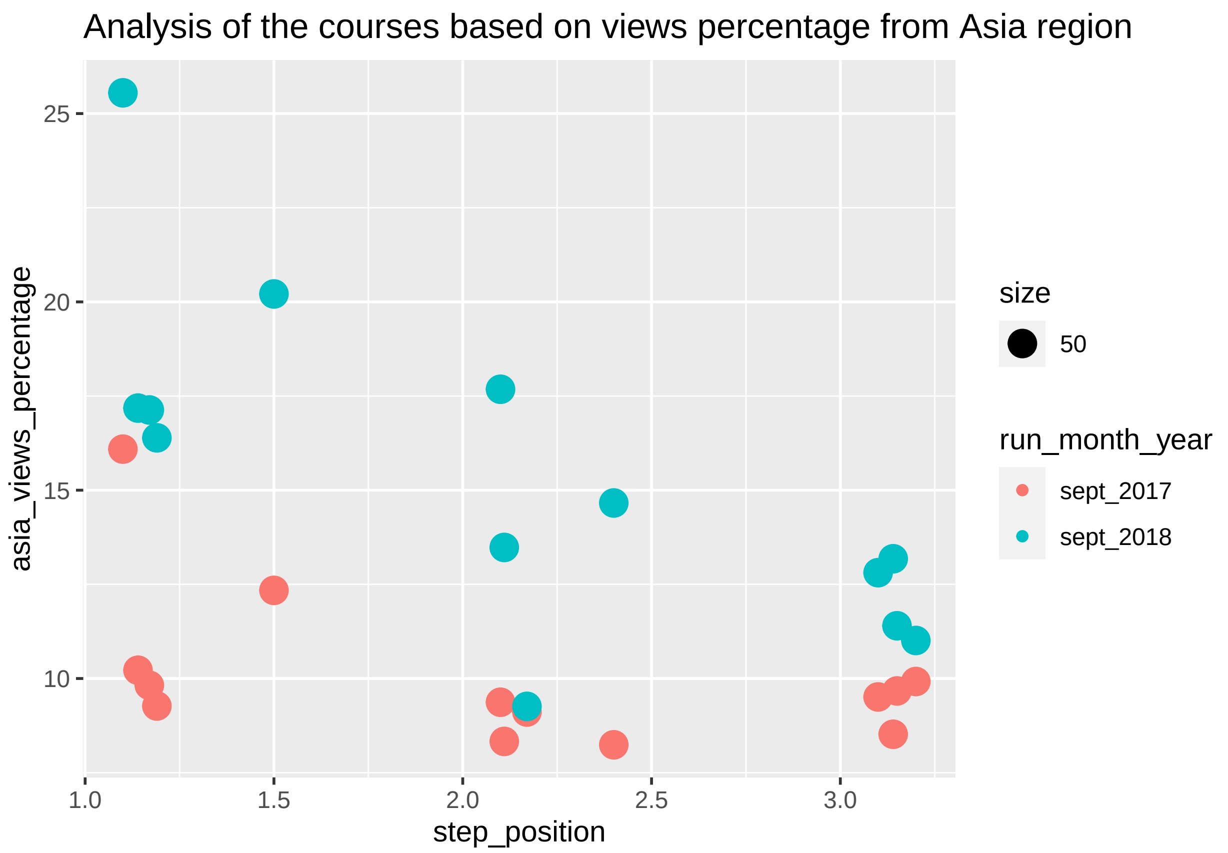
cycle__01 Analysis of the courses based on views percentage from major regions using graphical tool for summary-taking data__files of one year apart for exploration.

```
plot_G1 = ggplot(Video_stats, aes(step_position, europe_views_percentage, color = run_month_year, size =
  ggtitle("Analysis of the courses based on views percentage from Europe region",)
plot_G1
```



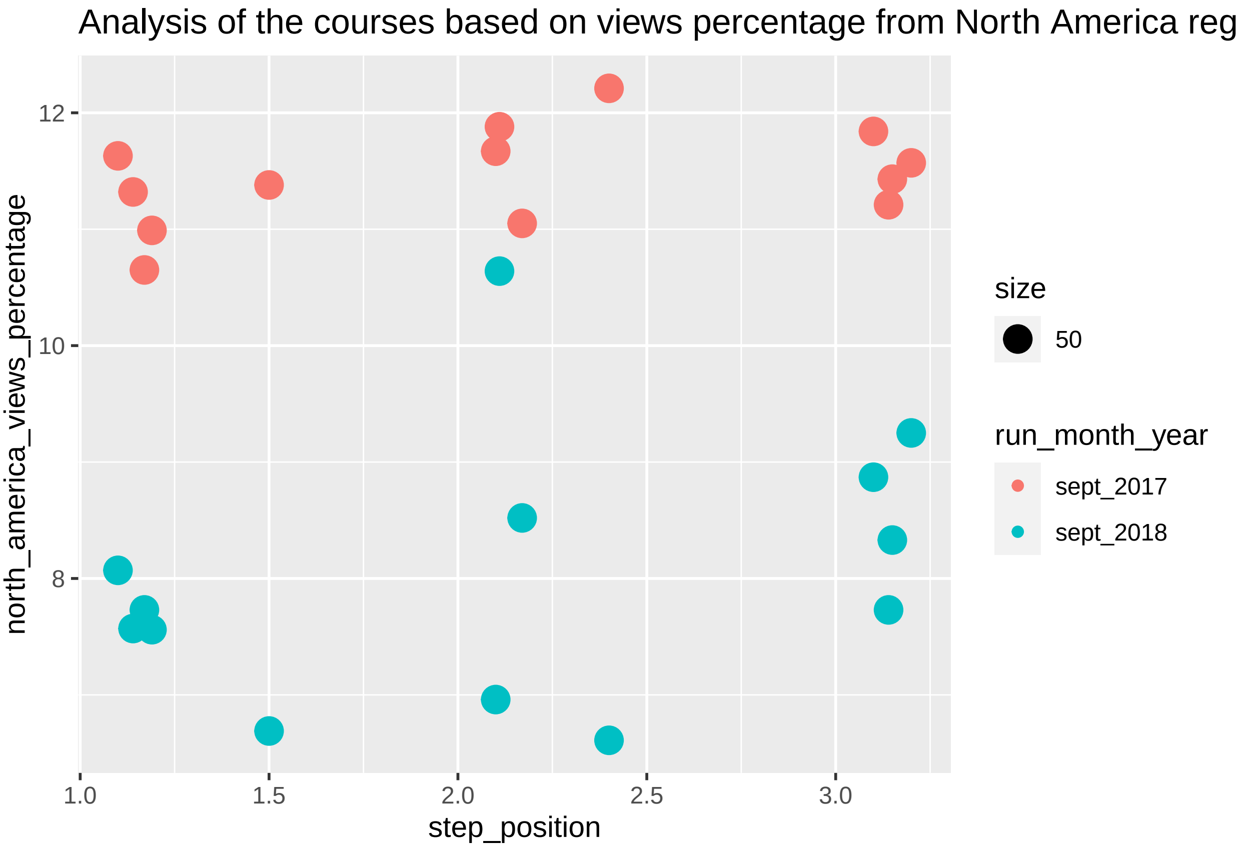
* From the plot, its significantly visible that the module did not made a much impact in the Europe region learners as the view percentage critically dropped during the september,2018 month compared to the month of September,2017. * Also it can be seen that percentage viewership in european region is quite healthy i.e. greater than 50%.

```
plot_G2 = ggplot(Video_stats, aes(step_position, asia_views_percentage, color = run_month_year, size =
  ggtitle("Analysis of the courses based on views percentage from Asia region",)
plot_G2
```



- 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 **50%** of the learners have viewed most of the modules completely.
- And below **40%** of the learners have viewed three or four modules completely. This suggests that the learners are not interested in viewing those modules as it maybe out of the scope.
- Approximately around **50% - 60%** of the learners are completing the modules which they find it covers their necessity.

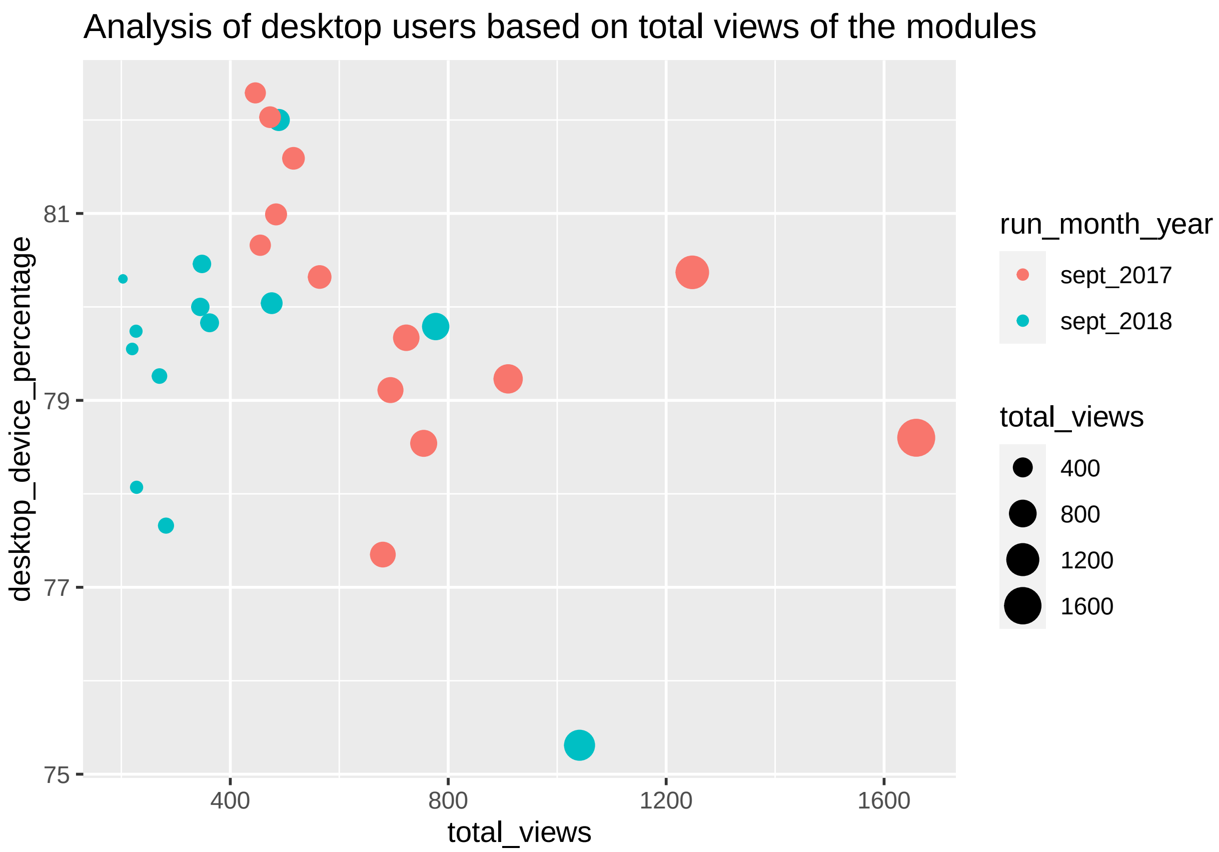
```
plot_G3 = ggplot(Video_stats, aes(step_position, north_america_views_percentage, color = run_month_year
  ggtitle("Analysis of the courses based on views percentage from North America region",)
plot_G3
```



* The above bar plot represents the analysis made based on the modeules that are viewed from the North American region. * This plot completely suggest that the percentage amount of total viewers of the course in the north american region has significantly decreased in sept,2018 compared to sept,2017 * Also the percentage viewership of the course in sept,2017 had a same range across the step_positioned videos,ranging between 10%-13%.

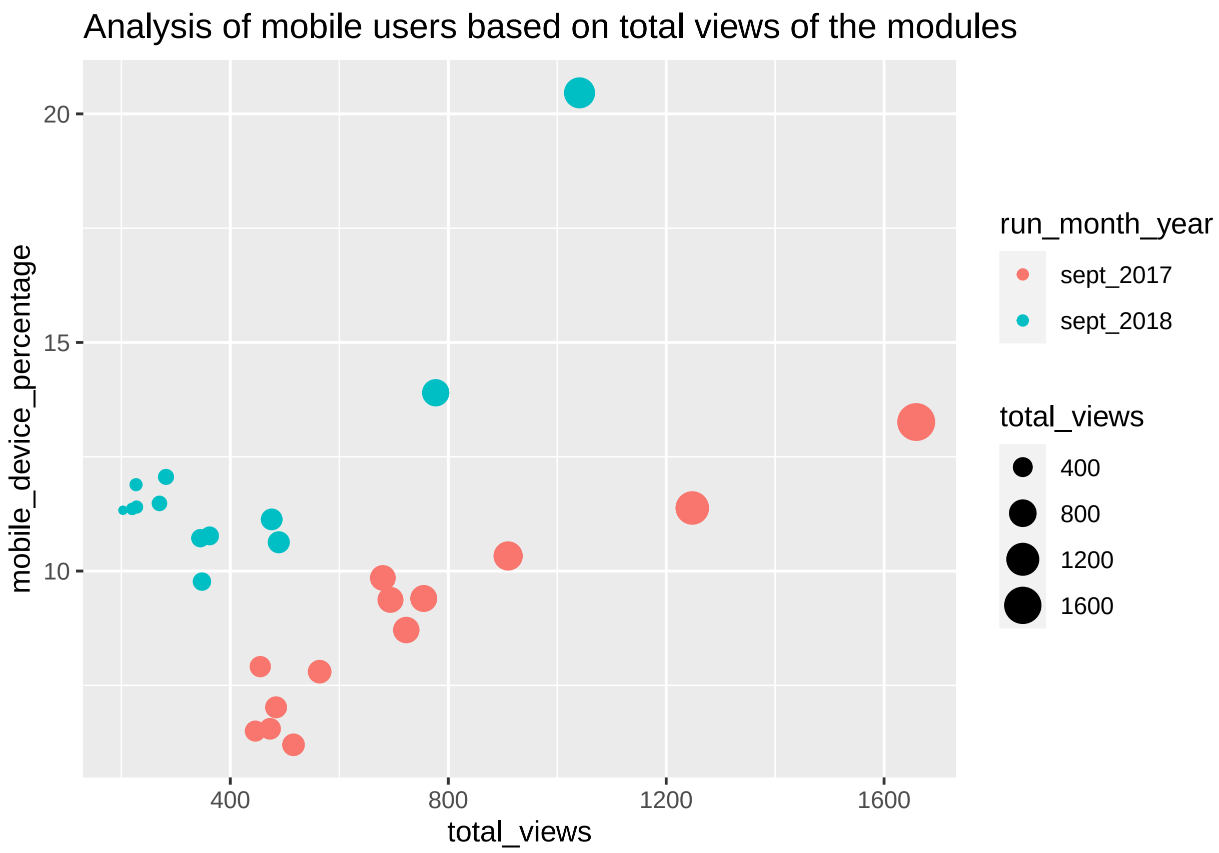
cycle__02 Analysis of the courses based on views percentage from major devices used using graphical tool for summary-taking data_files of one year apart for exploration

```
plot_dv1 = ggplot(Video_stats, aes(total_views, desktop_device_percentage, color = run_month_year, size
  ggtitle("Analysis of desktop users based on total views of the modules")
plot_dv1
```



* The representation is based on 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.

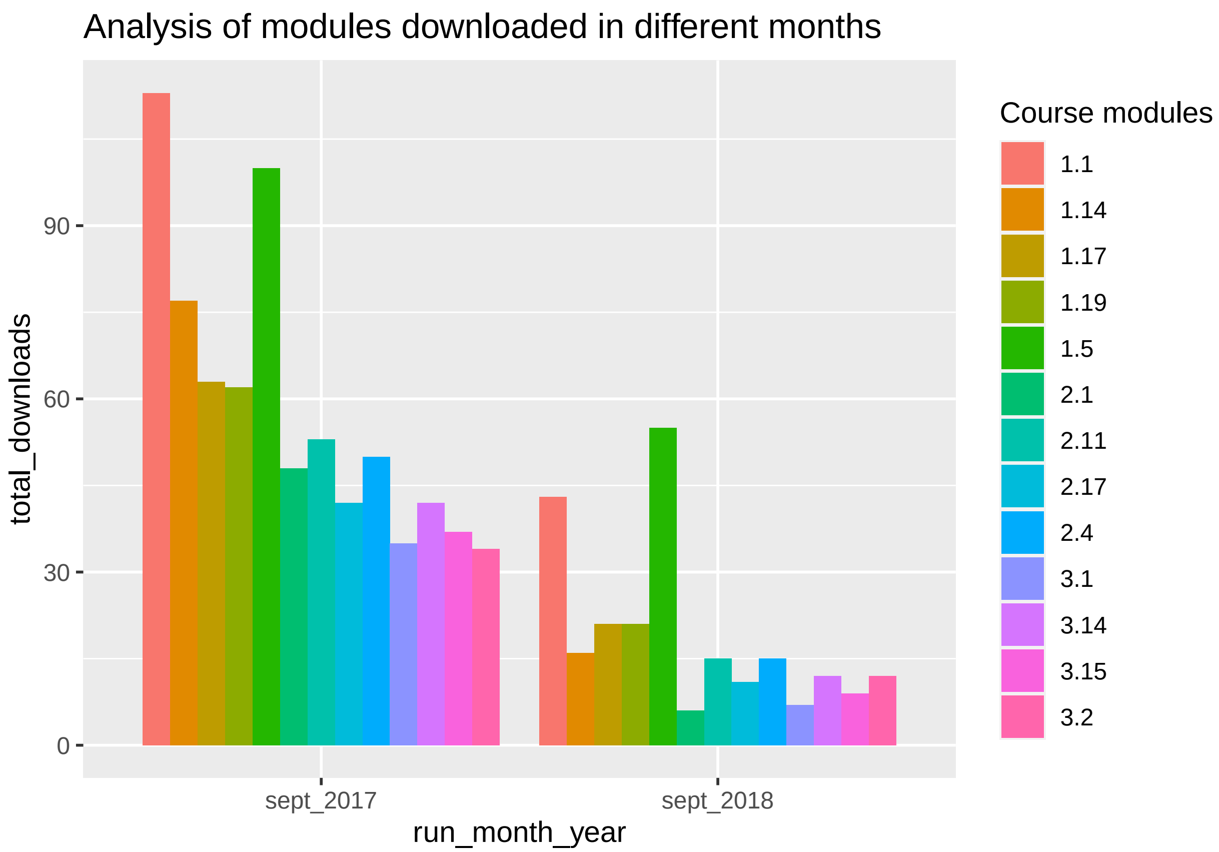
```
plot_dv2 = ggplot(Video_stats, aes(total_views, mobile_device_percentage, color = run_month_year, size
  ggtitle("Analysis of mobile users based on total views of the modules")
plot_dv2
```



* The representation is based on the analysis made on the modules that are viewed using the mobile devices. From the plot 2 things can be analysed that a)Percenatge of mobile device usage has significantly increased in september 2018 in comparision to seotember 2017 b)The viewership has decreased in the following year compared to initial year-2017.

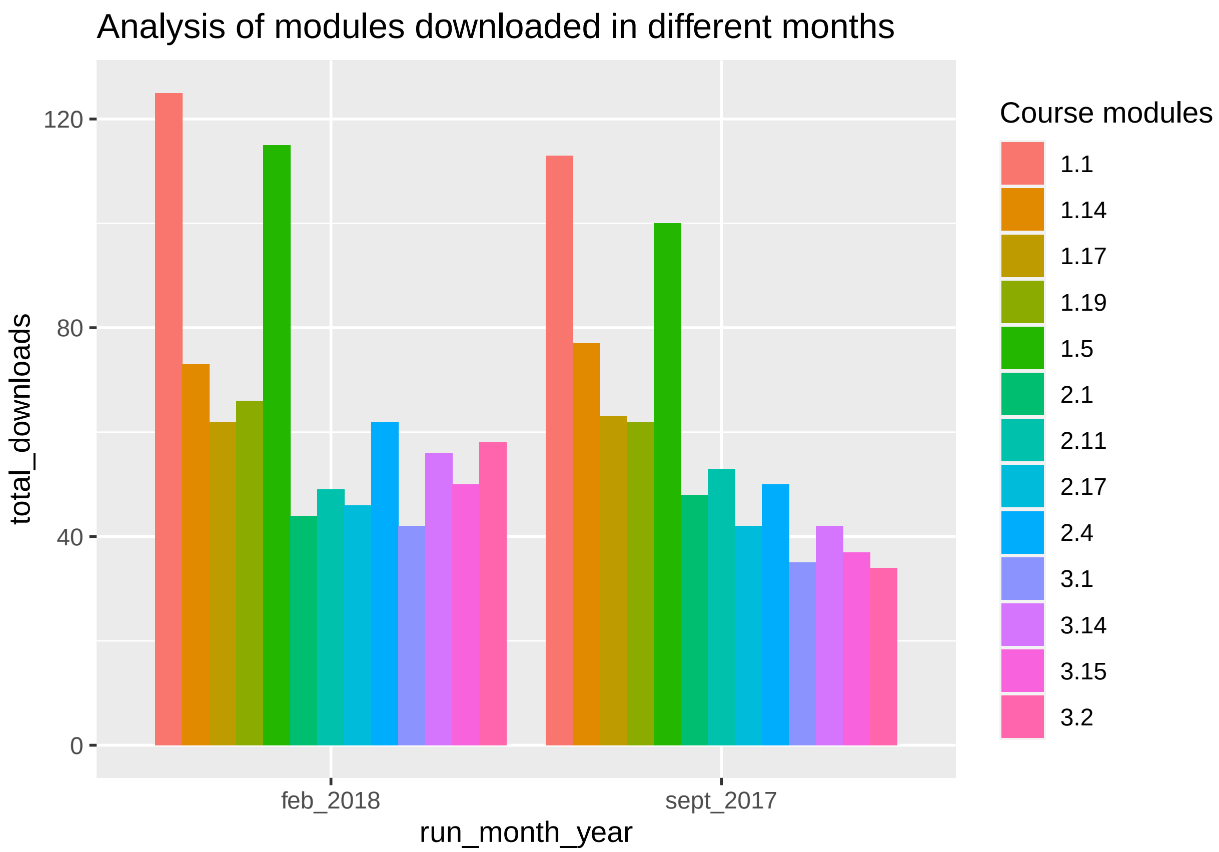
Some more graphical analysis,to explore and understand data more by business perspective

```
plot_x = ggplot(Video_stats) + geom_col(aes( run_month_year, total_downloads, fill=factor(step_position)
  ggtitle("Analysis of modules downloaded in different months")
plot_x
```



* This bar chart clearly shows that there is a decrement in course downloads in sept,2018 month compared to that of the sept,2017 month. * This clearly represents that the learners who wanted to perceive this course have already enrolled and completed the course before sept,2018 run or else learners are less interested in the contents of the course as compared to period of sept,2017 * There was depletion in every topic-course that was being downloaded.

```
plot_y = ggplot(Video_stats_Q1_Cycle02) + geom_col(aes( run_month_year, total_downloads, fill=factor(st
  ggtitle("Analysis of modules downloaded in different months")
plot_y
```



* This bar chart clearly shows that there is a relative increase in course downloads in feb,2018 month compared to that of the sept,2017 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 seen a depletion in some topic-course that was being downloaded compared to the previous month but overall the course reach had a positive growth rate.

Reproducibility

- Project Template is used to pre-process the data, compute the necessary functions mentioned and then generate the report through the Rmarkdown.
- If there is a new file which is required to generate the report, it should be inserted in the data directory and it should contain the same variable names as mentioned in the report to avoid further errors.
- For the purpose of analysis the datasets are combined as a single dataframe and hence if there is any news files which should be reported, it needs to be manually update the pre-process file to combine the data which would be in the munge sub directory.
- The analysis of the report is not completely reproducible as some manual interventions are required to change the variable names, binding up of the datasets as a single dataframe when there appears a new datafile.
- Apart from this, the report will be automatically generated when it is made to run from the R markdown report file which is in the Report sub directory.

Conclusion

After a brief analysis and explroation of the data sets available, it can be said that the course modules are quite engaging for the learners considering the number of downloads of the course materials and other aspects. The downloads depicts that the learners want to use it for reference later or share with other people. The length of video however, at times makes the content lose it's appeal on the viewers and they tend to not finish the course till the very end. More information about the leaving responses of such learner or enrollement of the learners from time to time can be gained through the leaving response surveys and enrollment surveys conducted which is for now out of the scope of this particular analysis.